Measuring Context: The Gaze Patterns of Children with Autism Evaluated from the Bottom-Up

Frederick Shic, Brian Scassellati
Yale University - Department of Computer Science
51 Prospect St, New Haven, CT 06511
frederick.shic@yale.edu, scaz@cs.yale.edu

David Lin, Katarzyna Chawarska
Yale Child Study Center
40 Temple St Suite 7I, New Haven, CT 06510
david.lin@yale.edu, katarzyna.chawarska@yale.edu

Abstract –**In this paper we use the mechanisms of a popular bottom-up computational model of visual attention in order to evaluate the gaze patterns of individuals in terms of elementary modalities such as color, orientation, motion, and intensity. We show that children with autism, even when watching naturalistic scenes, use less motion information, extending basic perceptual findings of motion deficits in autism to real-world scenes. In addition, by modifying the context of videos shown to children with and without autism (by changing the video scene, inverting the video, and displaying the video with and without sound) we show that that typical children, as compared to children with autism, are more affected by scene inversion. We discuss these and other results in terms of known sensory and cognitive abnormalities in autism and highlight the advantages and limitations of computational strategies in evaluating the effects of context on perceptual utilization.**

Index Terms – autism, context, visual attention, eye-tracking

I. Introduction

Autism is a pervasive developmental disorder marked by severe deficits in social functioning [1]. It is hypothesized that the social dysfunction evident in autism is the result of an early derailment of the typical experience-dependent social-cognitive developmental process [2]. Evidence for this theory is provided for by the atypical looking patterns of children with autism viewing naturalistic dynamic scenes. In contrast to typical controls, individuals with autism attend preferentially to mouths and bodies of characters rather than eyes [3]. As the eyes of an individual convey a great deal of information about his or her internal mental state [4], not looking at the eyes would necessarily lead to deficits in processing social information.

It has been hypothesized that the established atypical viewing patterns have some neural basis, which is to say that abnormal looking patterns are not the ultimate cause of social dysfunction, but rather an expression of some underlying neurocognitive divergence [5]. Insight into a neurocognitive mechanism is potentially provided for by exploring basic perceptual abnormalities in individuals with autism. These perceptual abnormalities could bias the child with autism away from building the typical scaffolding upon which social skills are built. For example, individuals with autism are known to have preferences and advantages for local visual processing (as compared to global processing) [6,7,8]. This inherent preference may play a role in discrepancies observed during the viewing of inverted faces: whereas typical individuals are disturbed by inversion (likely due to disruption of global configural features), individuals with autism are not [9]. In addition, there is evidence for motion processing deficits in autism [10,11]. The lack of salient attribution to biological motion [12] might lead to the inability to properly associate human contact to social reward (e.g. consider the prominence of the appearance of the mother's face from the viewpoint of a crying infant).

In order to investigate basic perceptual abnormalities, we could construct sensory experiments aimed at testing specific psychophysical effects. However, this approach has the drawback of not testing perception in the natural environment, making it difficult to generalize the role of deficits in a specific task to everyday functioning. If we want to access the usage of basic perceptual features in an ecologically valid way (i.e. in situations closely aligned with real social experience), we need mechanisms by which scenes viewed by individuals can be decomposed into elementary properties. That is, we need a low-level analogue of high-level interpretations of abnormal gaze patterns. To do this we can follow the route of [13,14]: we can employ computational signal-processing tools in the analysis of scene content. Specifically, we can use the feature extraction processes of computational models that emulate the visual attention processes of primates. These computational models of attention typically decompose the visual scene from the bottom-up, defining how "attractive" a spatiotemporal point is based on a particular perceptual modality; they are thus an appropriate fit for evaluating the utilization of basic visual attributes.

Evaluating gaze patterns in terms of elementary features can provide measures for comparing the preferences for low-level modalities in one group against another. However, this only provides a perceptual baseline for a set of cognitive processes affected by multiple aspects. To gain access to higher-level aspects we must take into account context, where context is here operationally defined as those factors not accounted for by the particular computational framework. Manipulating the context of a scene gives us a direct quantitative measure, in terms of effects on basic perceptual properties, of that contextual factor. We are specifically interested in two contextual effects known to impact visual attentional response: scene orientation (e.g. face inversion) and sound (e.g. loud noises causing alarm).

The purpose of this work is to compare the gaze patterns of children with autism to typically-developing controls in terms of elementary scene properties, as computed by the Itti model of visual attention [15]. Furthermore, within this framework we would like to gauge the contribution and impact of the contextual modification of scene orientation and sound. We will show results that are consistent with previous results found in literature and which also provide interesting avenues for future exploration.

II. Computing Elementary Features

Though there is no standard method by which elementary perceptual features can be extracted, many computational models of visual attention accomplish feature decomposition en route to their goal. Computational models of visual attention operate by taking in some representation of the spatiotemporal scene and returning the locations in the scene to which attention should be drawn (Figure 1). Though many models of visual attention exist [15,16,17], here we will rely on the model of Itti et al. [15], as it is perhaps the best known and most used model.

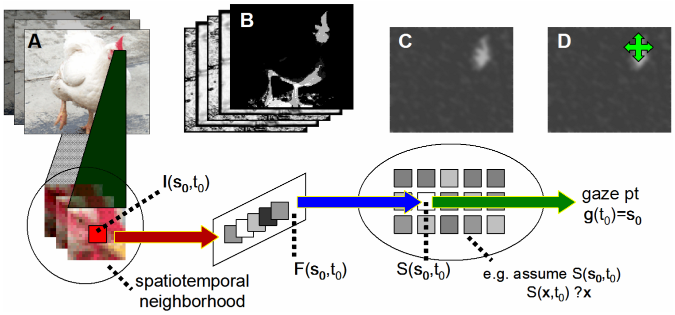

Figure 1: Framework for Computational Models of Visual Attention. The original spatiotemporal scene (A) is decomposed into a set of features (B); these features are in turn combined into a salience map (C). From the salience map the location of gaze determined (cross, D)

We use a custom-written version of the Itti Model of visual attention as defined in [15] and augmented with motion [18]. This model extracts the pre-attentive modalities of color, intensity, orientation, and motion from an image (Table I) and evaluates "conspicuity", or modality-specific salience. For instance, a red stop sign against a background of green forest is highly "conspicuous" in terms of color. Similarly, a car driving in the opposite direction of traffic, when viewed from above, is conspicuous in terms of motion. The conspicuity of each modality is assembled into a conspicuity map which indicates the prominence of that modality at every point in space (e.g. Figure 1B could be a color conspicuity map).

For completeness, we recap the computational model here and note that justification for the techniques employed can be found in the original sources [15,18] (those less interested in computational formalities should skip to section IV). In brief, the modalities of color, intensity, orientation, and motion are assembled into a multiscale representation using Gaussian and Laplacian pyramids [19]. Within each modality, center-surround operators are applied in order to generate multiscale feature maps. An approximation to lateral inhibition is then employed to transform these multiscale feature maps into conspicuity maps. Finally, conspicuity maps are linearly combined to determine the saliency of the scene.

TABLE I
Description of Modalities in The Extended Itti Model

Modality	**description**
Intensity	Contrasts in luminance; e.g. a small bright area on a larger darker background
Orientation	Pop-out effects based on differences in orientation; e.g. a single diagonal bar in a grid of horizontal bars
Color	Pop-out effects based on color contrasts; e.g. a single red object on a background of green.
Motion	Contrasts in motion; e.g. an object moving to the left as many other objects move to the right

We begin by describing the original Itti model, which was intended to operate over static images (motion will be summarized subsequently). The first stage of the Itti model is to create a multiscale representation of each modality. Given an input image with three color channels, red (*r*), green (*g*), and blue (*b*), the Itti Model first computes the associated intensity of the image as $I=(r+g+b)/3$. I is then used to create the Gaussian pyramid $I(\sigma)$ and the Laplacian pyramid $L(\sigma)$, where σ is the pyramid scale, in the following filter-subtract-decimate manner [19]:

$$I^0(n+1) = W * I^0(n) \tag{1}$$

$$L(n) = I^0(n) - I^0(n+1) \tag{2}$$

$$I(n+1) = \text{SUBSAMPLE}[I^0(n+1)] \tag{3}$$

with $I^0(0) = I$, the Gaussian filter $W=W_0W_0^{\mathrm{T}}$, W_0^{T}=[1/16, 1/4, 3/8, 1/4, 1/16], and SUBSAMPLE a function which subsamples the input image by a factor of 2. The scales created are $\sigma\in[0..8]$.

The same filter-subtract-decimate method is applied to the individual color channels, *r*, *g*, and *b*, to obtain a multiscale representation of colors, $r(\sigma)$, $g(\sigma)$, and $b(\sigma)$. Normalized color maps at each scale, $r'(\sigma)$, $g'(\sigma)$, and $b'(\sigma)$, are then computed by point-by-point division of color with intensity (points with intensities in $I(\sigma)$ less than 1/10th the maximum of $I(\sigma)$ are zeroed). These normalized color maps are combined to yield broadly tuned color channels red (R), green (G), blue (B), and yellow (Y) for each scale:

$$R(\sigma) = r'(\sigma) - (g'(\sigma) + b'(\sigma))/2 \tag{4}$$

$$G(\sigma) = g'(\sigma) - (r'(\sigma) + b'(\sigma))/2 \tag{5}$$

$$B(\sigma) = b'(\sigma) - (r'(\sigma) + g'(\sigma))/2 \tag{6}$$

$$Y(\sigma) = (r'(\sigma) + g'(\sigma))/2 - |r'(\sigma) - g'(\sigma)|/2 \tag{7}$$

Orientations at multiple scales are computed by taking the real component of spatial Gabor filtering over levels of the Laplacian pyramid (described in [19] with alternative notation and slight variation):

$$O_c(\sigma,\theta) = F_s(\theta) * L(\sigma) \tag{8}$$

$$O(\sigma,\theta) = \Re e\{O_c(\sigma,\theta)\} \tag{9}$$

with $F_s(\theta)$ the coarse Gabor filter at orientation $\theta=\pi N / 4$, $N\in[0..3]$, defined in 2D for a given point at (x,y):

$$F_{s;x,y}(\theta) = W_{x-x_0,y-y_0} e^{i\frac{\pi}{2}(x\cos\theta + y\sin\theta)} \tag{10}$$

where W is the Gaussian filter used in (1), with x_0 and y_0 chosen to appropriately center W ($x_0 = y_0 = -2$ in our case).

From these multi-scale representations of intensity, color, and orientation, feature maps are derived. Feature maps are created with the aid of a center-surround difference operator Θ. For a given multi-scale modality X, $X(c)\ \Theta\ X(s)$ interpolates the image with lower resolution to the resolution of the finer image, and then subtracts point-by-point. The interpolation is accomplished through the inverse application of equations (1) and (3). For all modalities, the center scales are $c \in \{2,3,4\}$ and the surround scales are $s=c+\delta$, $\delta \in \{3,4\}$.

The intensity feature maps $\mathcal{I}(c,s)$ are straightforward:

$$\mathcal{I}(c,s) = |\ I(c)\ \Theta\ I(s)\ | \qquad (11)$$

The color feature maps are slightly reordered to emulate color double-opponency for red-green $\mathcal{RG}(c,s)$, and blue-yellow $\mathcal{BY}(c,s)$:

$$\mathcal{RG}(c,s) = |\ (\ R(c) - G(\mathrm{c})\ )\ \Theta\ (\ G(\mathrm{s}) - R(s)\ )\ | \qquad (12)$$

$$\mathcal{BY}(c,s) = |\ (\ B(c) - Y(\mathrm{c})\ )\ \Theta\ (\ Y(\mathrm{s}) - B(s)\ )\ | \qquad (13)$$

Finally, the orientation feature maps $\mathcal{O}(c,s,\theta)$ are separately coded for each orientation Θ:

$$\mathcal{O}(c,s,\theta) = |\ O(c,\theta)\ \Theta\ O(s,\theta)\ | \qquad (14)$$

Feature maps for each modality are then combined into conspicuity maps. Conspicuity maps represent the salience of the modality as a whole. This is mediated through a normalization operator, $\mathcal{N}$, and a cross-scale addition operator, ⊕. The normalization operator $\mathcal{N}(\mathcal{M})$ returns a rescaled version of map $\mathcal{M}$, approximating lateral inhibition [20], by first linearly scaling $\mathcal{M}$ into a fixed range $[0,M]$, then multiplying the map by $(M-m)^2$, where m is the average of all local maxima in $\mathcal{M}$ except one point where the value is M. In our work, local maxima were locations with values greater than all eight neighbors. The cross-scale addition operator, ⊕, expands or reduces maps to scale 4 and then adds point-by-point.

The intensity conspicuity map $\bar{I}$, color conspicuity map $\overline{C}$, and orientation conspicuity maps $\overline{O}$ are then defined:

$$\bar{I} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{I}(c,s)) \qquad (15)$$

$$\overline{C} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} [\ \mathcal{N}(\mathcal{RG}(c,s)) + \ \mathcal{N}(\mathcal{BY}(c,s))\ ] \qquad (16)$$

$$\overline{O} = \sum_{\theta \in \{0,\frac{\pi}{4},\pi,\frac{3\pi}{4}\}} \mathcal{N}\left(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{O}(c,s,\theta)) \right) \qquad (17)$$

For motion, we need to extend our previous equations in time [18] (e.g. the intensity modality $I(\sigma)$ becomes $I(t,\sigma)$, the red-green feature map $\mathcal{RG}(c,s)$ becomes $\mathcal{RG}(t,c,s)$, etc.). For N frames $I(t,\sigma)$, $t\in[1..N]$, we obtain motion feature maps in the following manner:

1) Compute the N-th order first temporal derivative, $M_t(t,\sigma)$.
2) Compute the spatial derivative, $M_s(t,\sigma,\theta)$:
$$M_s(t,\sigma,\theta) = \boldsymbol{Im}\{O_c(t,\sigma,\theta)\} \qquad (18)$$
3) Compute the motion feature map $\mathcal{M}(t,\sigma,\theta)$:
$$\mathcal{M}(t,\sigma,\theta) = \mathcal{M}_s(t,\sigma,\theta) \cdot \mathcal{M}_t(t,\sigma,\theta) \qquad (19)$$

To obtain the motion conspicuity maps we:

1) Compute the direction of motion for each orientation to obtain positive and negative directional features. The positive directional feature $\mathcal{M}_+(t,\sigma,\theta)$ is $\sqrt{\mathcal{M}(t,\sigma,\theta)}$ at locations where $\mathcal{M}(t,\sigma,\theta)$ is positive, and is 0 otherwise; the negative directional feature is computed similarly.
2) Compute the directional contribution to motion conspicuity, $\mathcal{M}_d(t,\sigma,\theta)$:
$$\mathcal{M}_d(t,\sigma,\theta) = \mathcal{N}(\mathcal{M}_+(t,\sigma,\theta)) \oplus \mathcal{N}(\mathcal{M}_-(t,\sigma,\theta)) \qquad (20)$$
3) Compute across-scale contribution for each orientation:
$$\mathcal{M}_o(t,\theta) = \bigoplus_{\sigma=0}^{8} \mathcal{N}(\mathcal{M}_d(t,\sigma,\theta)) \qquad (21)$$
4) Finally, we compute the conspicuity map for motion:
$$\overline{M}(t) = \sum_{\theta \in \{0,\frac{\pi}{4},\pi,\frac{3\pi}{4}\}} \mathcal{N}\big(\mathcal{M}_o(t,\ \theta)\big) \qquad (22)$$

All conspicuity maps are combined in order to yield a saliency S for every point:

$$S = \frac{1}{4}\Big(\mathcal{N}(\bar{I}) + \mathcal{N}(\overline{C}) + \mathcal{N}(\overline{O}) + \mathcal{N}(\overline{M})\Big) \qquad (23)$$

III. Matching Gaze Patterns To Features

An individual will look at particular points in space over time; we are interested in matching these points to their associated low-level perceptual interpretations. In order to accomplish this, we employ the internal representations of the Itti model. We begin with a conspicuity map, here generically defined as $\overline{V}(\boldsymbol{s},t)$ for some feature V, and spatial location $\boldsymbol{s}$ and time t. In order to obtain comparability for all time points and all modalities, we first normalize the values of the conspicuity map by rank ordering all the spatial values for a given time, to obtain the rank-ordered conspicuity map $\overline{V}_r(\boldsymbol{s},t)$:

$$r(x,thr) = \begin{cases} 0, & x \geq thr \\ 1, & otherwise \end{cases} \qquad (24)$$

$$\overline{V}_r(\boldsymbol{s},t) = \frac{\sum_{\boldsymbol{s'}\in\boldsymbol{S'}} r(\overline{V}(\boldsymbol{s'},t),\overline{V}(\boldsymbol{s},t))}{|\boldsymbol{S'}|} \qquad (25)$$

Given the gaze patterns of some individual i, $\boldsymbol{g}_i(t)$, we can obtain the perceptual usage of V at time t by i as $v_i(t) = \overline{V}_r(\boldsymbol{g}_i(t),t)$; we can in turn use this time-varying perceptual usage score to compute the aggregate perceptual score $p_{v,i} = \mathrm{median}_t(v_i(t))$.

IV. Experimental Methods

A. Subjects and Data

26 children with autism spectrum disorder (ASD) and 15 typical-developing controls (TD) participated in the study. The age of the TD population was 39.2 (16.5) months; the age of the ASD population was 42.2 (12.1) months. The diagnosis of ASD was determined by expert clinicians as part of a comprehensive clinical examination at the Yale Child Study Center.

Each child was accompanied by his parent into the room where the experiment was conducted. Children with sufficient neck support sat in a car child seat strapped to a chair; younger children were held over their parent's shoulder as the parent sat in a chair. A monitor, centered with the eye-line of the child, was mounted 75 cm from the child's face. The child's gaze patterns were tracked using a commercial eyetracker from SensoMotoric Instruments (iView X RED) at 60hz.

The experiment in this study was embedded in a large run of several different experiments so as to minimize the overall amount of time spent positioning and calibrating the child. The child saw 4 movie clips, with each clip measuring approximately 24x18 (width x height) visual degrees and lasting for 30 seconds. All clips depicted a natural interaction between an adult caregiver and a child (e.g. playing with a toy) (Figure 2, Table II). Each clip was shown in one of four conditions representing the modulation of two variables: orientation (inverted or upright), and sound (mute or sound).

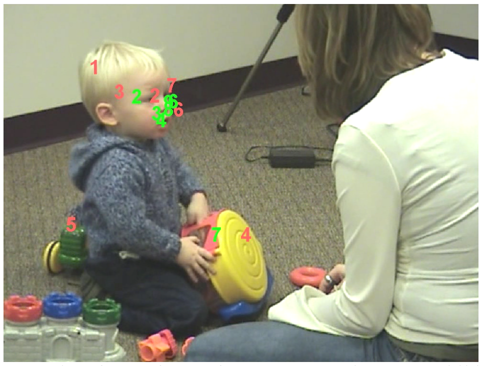

Figure 2: Example of one frame from a scene shown to children (in the upright with sound condition), with the gaze locations of ASD individuals (red) and TD individuals (green) for that frame overlayed.

TABLE II
SCENE DESCRIPTIONS

Scene	Description
1	Child tries to put objects into a colorful container; Caregiver instructs child to insert toys in holes by pointing gestures
2	Depicted in Figure 2; Child lifts head to speak to caregiver; Child opens container and drops in toys
3	Child offers toy to caregiver; Caregiver teaches child name of toy; Child repeats name of toy while continuing to play
4	Child enters scene with Caregiver and a mechanical toy that spits out colored balls; Child plays with back to camera (obstructing toy)

During a single experimental session, the clips were always presented in an order from most contextually disturbed to least in order to minimize irreversible recognition effects: inverted mute, inverted with sound, upright mute, upright with sound. Each clip presented to the child during a single session contained different scene content. However, a child could engage in multiple sessions, with each session conducted on a different day. In cases where children engaged in multiple sessions, they would see the same scene content on different days, but would never see the same scene-condition pairing twice, as the scene content would be rotated amongst the conditions. Clips were rejected from analysis if they contained less than 10 seconds (600 points) of valid eye-tracking data (ASD 22 clips; TD 8 clips); typically this rejection occurred due to the child affect or inattention. In total, the ASD population contributed 157 clip viewings over 49 sessions; the TD population contributed 88 clip viewings over 24 sessions.

B. *Data Processing and Analysis*

Features were extracted from the movie clips in the manner described in section II. For each clip viewing, the gaze patterns of the children were mapped to the associated features 200 ms in the past, as described in section III. This delay was incorporated to account for the fact that saccades and fixations are responses, and not instantaneous correlates, to visual events; 200 ms is a rough estimate of the time required to plan and complete non-anticipatory saccades.

Data processing yielded an aggregate perceptual score for each modality (intensity, orientation, color and motion) for every clip viewing. Each modality was analyzed independently using a univariate analysis of variance with factors: diagnosis (ASD or TD), orientation (upright or inverted), sound (mute or with-sound), and specific scene content (a subset of 4 possible scenes). Age was listed as a covariate as the age range in both ASD and TD populations was large. Because the perceptual score for a particular modality was tightly coupled to the scene (i.e. when analyses were originally conducted, the effect of specific scene was by far the most significant effect), scenes were analyzed together only when the set of scenes together did not register significant between-subject effect. In the event of multiple choices amongst scene combinations, the combination resulting in the greatest number of subjects was retained. The data used in this study are summarized in Table III:

TABLE III
DATA CHARACTERIZATION

Modality	$\mathbf{n_{asd}}$	$\mathbf{n_{td}}$	$\mathbf{n_{upr}}$	$\mathbf{n_{inv}}$	$\mathbf{n_{snd}}$	$\mathbf{n_{mute}}$	$\mathbf{n_{scene}}$
Intensity	122	67	97	92	87	102	3
Orientation	79	45	66	58	49	75	2
Color	43	22	31	34	38	27	1*
Motion	76	42	62	56	60	58	2

Number of clip-viewings: n_{asd} (ASD), n_{td} (TD), n_{upr} (orientation upright), n_{inv} (orientation inverted), n_{snd} (with sound), n_{mute} (no sound); n_{scene} (number of scenes combined for analysis (based on comparable means)); *for color no scenes were comparable to any other.

V. RESULTS

The pattern of aggregate perceptual scores was tightly coupled to the scene content, as shown in Table IV.

TABLE IV
PERCEPTUAL SCORES OF MODALITIES FOR EACH SCENE

scene	**Intensity**			**Orientation**			**Color**			**Motion**		
	μ	**σ**	**N**	**μ**	**σ**	**N**	**μ**	**σ**	**N**	**μ**	**σ**	**N**
1	.35	.10	65	~~.67~~	~~.07~~	~~65~~	.70	.11	65	~~.77~~	~~.08~~	~~65~~
2	.35	.07	62	.55	.05	62	~~.78~~	~~.10~~	~~62~~	.70	.06	62
3	~~.49~~	~~.07~~	~~56~~	~~.67~~	~~.10~~	~~56~~	~~.46~~	~~.08~~	~~56~~	.69	.08	56
4	.35	.04	62	.55	.07	62	~~.65~~	~~.06~~	~~62~~	~~.55~~	~~.03~~	~~62~~

Scenes not comparable to other scenes were removed (crossed-out). Reported means are collapsed across conditions and diagnoses.

After controlling for scene content, no significant effects of diagnosis, scene orientation, sound, or age were found for color and orientation. However, significant differences were detected for intensity and motion. For intensity, there was a main effect of diagnosis (ASD vs TD) (F(1,188) = 5.7, p<0.05) and scene orientation (upright vs inverted) (F(1,188)=11.6, p<0.001), and an interaction for diagnosis x scene orientation (F(2,188)=6.8, p<0.01). The effects of sound (with-sound or mute) and age were not significant.

To examine the nature of the interaction, simple between-group comparisons for each of the four conditions (i.e., inverted-mute, inverted-sound, upright-mute, upright-sound) were conducted. The comparisons indicated that ASD and TD groups differed significantly only in the upright-mute (F(1,41)=11.94, p < .001) and upright-sound (F(1,49)=12.13, p <.001) conditions, but not the inverted-mute (p>.28) or inverted-sound (p>.77) conditions. Furthermore, we compared intensity scores within each group in the upright and inverted conditions. These within-group comparisons indicated that toddlers with ASD were not affected by scene inversion (p>.43), but in TD toddlers the salience of intensity increased significantly when the scenes were inverted position, F(1,66)=18.55, p < .001).

For motion, there was a main effect of diagnosis (F(1,117) =6.3, p<0.05), scene orientation (F(1,)=4.0, p<0.05), and sound (F(1,117)=9.2, p<0.01). There were no significant interactions between the factors, but the effect of age on the salience of motion was significant (F(1,117)=5.6, p<.05). Toddlers with ASD were less sensitive to motion cues, regardless of the condition (i.e., scene orientation or presence/absence of sound). All toddlers tended to be more sensitized to motion in the sound than no sound conditions and when the scenes were inverted as compared to the upright. Results are summarized in Figure 3.

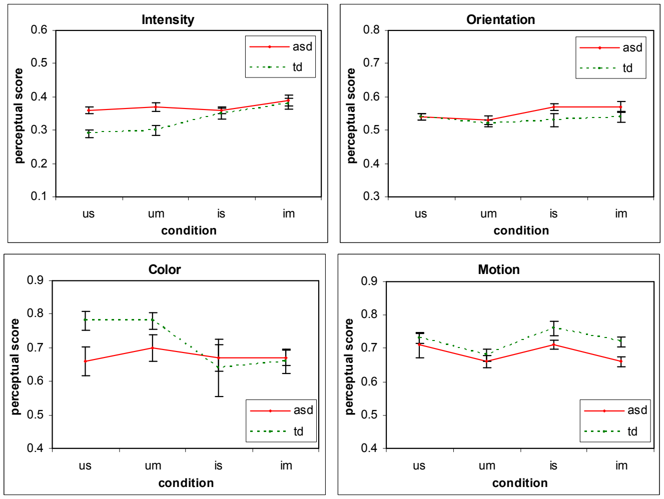

Figure 3: Aggregate perceptual scores by diagnosis for each modality. The x-axis is ordered from least disturbed to most disturbed. The categories are: inverted muted (im), inverted with sound (is), upright muted (um), and upright with sound (us). Only the differences for intensity and motion were significant. Note that the scores for each modality are displayed on different scales but over the same range. Bars are ± one standard error.

V. Discussion

At the baseline, that is, in the least disturbed state, typical controls use less intensity information than children with autism. Upon scene inversion, however, the utilization of intensity by typical controls increases to the point where the ASD and TD populations are indistinguishable. By comparison, for children with autism, no effect of scene inversion is observed. This result is evocative of experiments in [9] where it was found that face inversion decreased face recognition performance for typical individuals, but not individuals with autism. Furthermore, this result is consistent with evidence for local features preference or global processing deficits in autism. If typically developing controls, in the least disturbed condition, are using configural or holistic processing, it is expected that by inverting the scene we would disrupt the use of specialized visual strategies, leading to a commensurate increase in purely perceptual usage. By contrast, if individuals with autism have a preference for local visual processing then we would expect that scene inversion would largely leave the attentional strategy of these individuals intact, as is supported by our data. We note that this possibility is not the only interpretation, for it is possible that a disturbed contextual factor could in turn be replaced by some other contextual factor. Controlling for these effects remains to be explored.

For motion, under all conditions, we note that children with autism use less motion information than their typically developing peers. This result is consistent with results from literature demonstrating motion processing deficits in autism [8,10,12]. In addition, the covarying effect of age on the motion usage in typical children is supported by research showing decreasing thresholds for motion detection with increasing age [21]. Finally, the trend for both typically-developing children and children with autism was for greater motion usage in the presence of sound. In many of the scenes shown, a large amount of motion accompanied large changes in sound. It is possible that the presence of sound increased the urgency of motion.

We also note that the specific scene content under consideration plays a critical role in the reported perceptual usage. In the scenes presented in this experiment, the variations in choice of focus, the different actions that were performed, the various implicit object-people interactions, all could have had dramatic impacts on the expected perceptual values. This dramatic effect of scene content points to the inherent limitations of computational models of visual attention which try to model human gaze explicitly without optimizing for individual biases. This is a lesser problem for the use of computational feature extraction techniques in the *evaluation* of basic perceptual modalities. Nonetheless, the quality of the evaluation of modality usage is only as good as the techniques employed. If our selected computational model could adequately predict attentional salience, we would see not only that context effects contributed only in a minor fashion to perceptual usage, but that the computational model accounts for a majority of an individual's gaze patterns. This is, however, not the case in this study, as the results for perceptual usage based on our computational model are nowhere near the maximum.

The present study is limited in many ways. First and foremost, the complexity of the experimental design necessitates a large sample size, whereas the present study only has a moderate sized set of subjects. For this reason the results in this paper should only be taken as preliminary. With a greater sample size it is possible that interesting effects could be uncovered for color and orientation modalities, both of which yielded no significant results. Secondly, the appropriateness and quality of the Itti model feature calculations was not conducted. It is certainly the case that the Itti model, as a model of *visual attention*, does not match up with human gaze patterns, though the specific saliency values are better than chance [13]. Without optimizing the parameters of the model [18,22], it will necessarily be a poor fit. It is, however, in our opinion, a sufficient construct for serving as an *evaluative* model of visual preference. Further work should examine alternative approaches to feature decomposition.

VI. Conclusions

We have taken a computational model of visual attention and employed its internal mechanisms as a strategy for evaluating gaze patterns in terms of elementary perceptual features. We have adapted this model with a framework for evaluating context by scene manipulation, and used this framework to evaluate the perceptual strategies of individuals with autism as compared with those of typically developing controls. Through these techniques, we have generated several results, framed in terms of perceptual utilization, which are consistent with other results from literature. We find that children with autism use less motion information than their typically-developing peers, consistent with motion deficits shown in autism. We find that children with autism, in terms of the perceptual utilization of intensity, are more resistant to scene inversion, supporting the role of local visual processing preferences in autism. We also find motion effects consistent with developmental trends in age and consistent with known interactions with sound. Finally, we have discussed the advantages and limitations of our computational methods in evaluating the perceptual usage of typical and atypical populations. It is our hope that this work can serve as a bridge between basic research using elementary stimuli in their investigation of the perceptual abnormalities in autism and clinical observations which examine sensory, social, and cognitive deficits during live interactions.

References

[1] American Psychiatric Association, *Diagnostic and statistical manual of mental disorders,* 4th ed. (DSM-IV), Washington, DC: American Psychiatric Association, 1994.

[2] A. Klin, W.Jones, R. Schultz, and F.Volkmar, “The enactive mind ,or from actions to cognition: lessons from autism,” *Phil. Transactions of the Royal Society B: Biological Sciences*, vol. 358, no. 1430, pp. 345-360, February 2003.

[3] A. Klin, W. Jones, R. Shultz, F. Volkmar, D. Cohen, “Visual Fixation Patterns During Viewing of Naturalistic Social Situations as Predictors of Social Competence in Individuals with Autism,” *Arch of General Psychiatry* , vol. 59, no. 9, pp. 809 – 816, September 2002.

[4] S. Baron-Cohen, R. Campbell, A. Karmiloff-Smith, “Are children with autism blind to the mentalistic significance of the eyes,” *British Journ. of Dev. Psych.*, vol. 13, no. 4, pp. 379-398, 1995.

[5] M. Belmonte, G. Allen, A. Beckel-Mitchener, L. Boulanger, R. Carper, and S. Webb, “Autism and Abnormal Development of Brain Connectivity,” *Journal of Neuroscience.*, vol. 24, no. 42, pp. 9228-9231, 2004.

[6] N. Rinehart, J. Bradshaw, S. Moss, A. Brereton, and B. Tonge, “Aytpical Interference of Local Detail on Global Processing in High-function Autism and Asperger’s Disorder,” *Journ. of Child Psychology And Psychiatry and Allied Disciplines*, vol. 41. pp. 769-778. Oct. 2000.

[7] F. Happé, “Autism: cognitive deficit or cognitive style,” *Treds in Cognitive Sciences*, vol. 3, no. 6, June 1999.

[8] U. Frith, Autism Explaining the Enigma. Oxford: Blackwell, 1989.

[9] D. Tantam, et al., “Autistic children’s ability to interpret faces: a research note,” *Journal of Child Psychology and Psychiatry*, vol. 30, pp. 623-630. 1989.

[10] E. Milne, J. Swettenham, P. Hansen, R. Campbell, H. Jeffries, and K. Plaisted, “High motion coherence thresholds in children with autism,” *Journal of Child Psychology and Psychiatry*, vol. 43, no. 2, pp. 255-263, 2002.

[11] S. Dakin and U. Frith, “Vagaries of Visual Perception in Autism,” *Neuron*, vol. 48, pp. 497-507, Nov 2005.

[12] R. Blake, L. Turner, M. Smoski, S. Pozdol, W. Stone, “Visual recognition of biological motion is impaired in children with autism,” *Psychological Science,* vol. 14, no. 2, pp. 151-157, 2003.

[13] D. Parkhurst and E. Niebur, “Scene content selected by active vision,” *Spatial Vision*, vol. 16, no. 2, pp. 125-154, Nov 2004.

[14] D. Neumann, et al., “Looking you in the mouth: abnormal gaze in autism resulting from impaired top-down modulation of visual attention”, *Social Cognitive and Affective Neuroscience*, vol. 1, no. 2, pp. 194-202, 2006.

[15] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254-1259, 1998.

[16] J.M. Wolfe and G. Gancarz, “Guided Search 3.0: A model of visual search catches up with Jay Enoch 40 years later”, In V. Lakshminarayanan (Ed.), *Basic and Clinical Applications of Vision Science*, Dordrecht, Netherlands: Kluwer Academic, 1996.

[17] C. Balkenius, A.P. Eriksson, & K. Astrom, “Learning in Visual Attention,” In *Proceedings of LAVS ’04*. St Catharine’s College, Cambridge, UK, 2004.

[18] F. Shic and B. Scassellati, “A Behavioral Analysis of Computational Models of Visual Attention,” *International Journal of Computer Vision*, vol. 73, no. 2, pp. 159-177, 2007.

[19] P. Burt and E. Adelson, “The Laplacian Pyramid as a Compact Image Code,” *Communications, IEEE Transactions on*, vol. 31, no. 4, pp. 532-540, 1983.

[20] L. Itti and C. Koch, “A Comparison of Feature Combination Strategies for Saliency-Based Visual Attention Systems,” *Proc. SPIE Human Vision and Electronic Imaging IV (HVEI'99)*, San Jose, CA, vol. 3644, pp. 473-82, Jan 1999.

[21] O. Braddick, J. Atkinson, and J. Wattam-Bell, “Normal and anomalous development of visual motion processing: motion coherence and `dorsal-stream vulnerability'”, *Neuropsychologia*, vol. 41, no. 13, pp. 1769-1784, 2003.

[22] F. Shic, W. Jones, A. Klin, and B. Scassellati, “Swimming in the Underlying Stream: Computational Models of Gaze in a Comparative Behavioral Analysis of Autism,” *28th Annual Conference of the Cognitive Science Society*, 2006.