Saliency-based Bayesian Modeling of Dynamic Viewing of Static Scenes

Daniel J. Campbell*
Child Study Center
Yale University

Joseph Chang†
Department of Statistics
Yale University

Katarzyna Chawarska‡
Child Study Center
Yale University

Frederick Shic§
Child Study Center
Yale University

Abstract

Most analytic approaches for eye-tracking data focus either on identification of fixations and saccades, or on estimating saliency properties. Analyzing both aspects of visual attention simultaneously provides a more comprehensive view of strategies used to process information. This work presents a method that incorporates both aspects in a unified Bayesian model to jointly estimate dynamic properties of scanpaths and a saliency map. Performance of the model is assessed on simulated data and on eye-tracking data from 15 children with autism spectrum disorder and 13 control children. Saliency differences between ASD and TD groups were found for both social and non-social images, but differences in dynamic gaze features were evident in only a subset of social images. These results are consistent with previous region-based analyses as well as previous fixation parameter models, suggesting that the new approach may provide synthesizing and statistical perspectives on eye-tracking analyses.

CR Categories: J.4 [Computer Applications]: Social and Behavioral Sciences—Psychology

Keywords: Bayesian model, Gibbs sampling, saliency, autism spectrum disorder

1 Introduction

Approaches to analysis of eye-tracking data to understand the processes directing visual attention frequently make use of a saliency map, a topographical representation of a visual field in which regions of high 'salience' or visual interest are assigned higher values [Koch and Ullman 1985]. Saliency maps play an integral part of some computational models explaining observed processes of selective attention [Itti et al. 1998; Itti and Koch 2000; Bruce and Tsotsos 2005]. However, models that investigate saliency often disregard dynamic properties of visual movement by preprocessing the data to extract fixations and using only the cleaned fixation locations to estimate saliency. Features of the eye-movement process, such as duration of fixations, subtle movements within fixations, or distributions of saccade lengths are not considered, and as such the models emphasize 'where' fixations are assigned on the saliency map at the expense of 'how.'

Instead of estimating the entire saliency model, some approaches estimate relative salience of portions of the stimulus through regions of interest (ROIs). This approach involves dividing a stimulus into important regions such as faces, eyes, and mouths of people, the background, and objects, and calculating the proportions of fixation time allocated to each region [Lewkowicz and Hansen-Tift 2012; Johnson et al. 2003; Richardson and Dale 2005]. ROI measures are faster and simpler to estimate than the entire saliency map, but are still based on features of salience and therefore the approach still neglects the dynamic properties of eye movement. The ROI approach also introduces biases in favor of *a priori* hypotheses, with attention towards areas not coded with their own regions impossible to characterize. This top-down assignment of regions by the researcher can be advantageous in testing hypotheses of attention in controlled experiments, but can preclude discovery of interesting but novel regions of interest in naturalistic viewing paradigms.

On the other hand, many analyses take the opposite approach and analyze gaze movements without regard to saliency [Rimey and Brown 1991; Yamato et al. 1992; Salvucci and Goldberg 2000; Feng 2006], including entropy-based methods to estimate variability of eye movements [Shic et al. 2008a; Harris Sr et al. 1986]. Such approaches yield important information regarding the pattern and distribution of gaze shifts in response to visual stimuli, but do not take into account the effect that relative saliency of different stimuli can have on eye movements, and vice versa. In addition, these approaches are complicated by the need to specify fixation patterns using information exterior to the model [Shic et al. 2008c].

While application of different yet separate approaches to the same data can address this dichotomy, doing so neglects the interplay between saliency and dynamic eye movement that can be studied using a single, integrated framework. Recent work has combined dynamic and saliency aspects of visual attention into unified models of attention [Brockmann and Geisel 2000; Renninger et al. 2007; Van Der Lans et al. 2008]. In this vein, we propose a new model for frame-by-frame (rather than fixation-by-fixation) eye movement on an estimated saliency map that does not heavily rely on preprocessing of data into fixations and saccades.

Performance of the model will be assessed using simulated data as well as eye-tracking data collected on a sample of two-year-old children with autism spectrum disorder or typical development, with differences in both saliency and dynamic viewing compared between groups. Individuals with autism spectrum disorder, a developmental disorder characterized by deficits in social communication and the presence of repetitive and restricted interests [APA 2013], frequently manifest altered patterns of visual attention to social stimuli [Chawarska and Shic 2009; Klin et al. 2002; Dalton et al. 2005; Norbury et al. 2009; Simmons et al. 2009, for review]. Analysis of this sample here serves not only to validate the proposed model on a well-characterized sample, but also to replicate previously-reported group differences in salience and in fixation patterns identified using other approaches [Chawarska and Shic 2009; Shic et al. 2008b].

*e-mail:daniel.campbell@yale.edu
†e-mail:joseph.chang@yale.edu
‡e-mail:katarzyna.chawarska@yale.edu
§e-mail:frederick.shic@yale.edu

ETRA 2014, March 26 – 28, 2014, Safety Harbor, Florida, USA.

2 Methods

2.1 Statistical Model

We propose a Bayesian model to generate and describe scanpaths across a static scene that integrates dynamic, frame-by-frame viewing of a scene with global saliency properties. This model contains three components: (1) a saliency map capturing the relative importance of each part of the image, modeled as a bivariate Gaussian mixture model; (2) a random walk component, specified by a bivariate Gaussian component added to the mixture model described above, whose variance is controlled by a bandwidth; and (3) a multiplier, which controls the relative importance of the saliency components and the random walk component.

Table 1: *Descriptions of model components.*

Model Component	Model Parameter	Description
Saliency map	ω	Salience of coordinates in the visual image. Modeled as a bivariate Gaussian mixture model with mixture weights ω.
Bandwidth	β	Variability of movement within fixations.
Multiplier	μ	Frequency of saccades.

Let $\mathbf{z}_{i0}$, $\mathbf{z}_{i1}$, $\mathbf{z}_{i2}$, ..., $\mathbf{z}_{iT}$ denote the point-of-regard coordinates in two-dimensional space for participant i at frames $t = 0, 1, 2, ..., T$ during a scanpath of a static image. Each participant i belongs to a single group g_i, and each group has its own saliency map. Let β_i and μ_i denote the bandwidth and multiplier, respectively, of participant i.

The point of regard of participant i at frame t is modeled as follows:

1. $\mathbf{z}_{i0}$ is drawn randomly from the saliency map for group g_i.
2. For $t > 0$, $\mathbf{z}_{i,t+1}$ is drawn from a bivariate Gaussian distribution (the random walk component) with mean $\mathbf{z}_{it}$ and variance $\beta_i I_2$ with probability $\frac{\mu_i}{\mu_i+1}$. This simulates a fixation step.
3. For $t > 0$, $\mathbf{z}_{i,t+1}$ is drawn from the saliency map for group g_i with probability $\frac{1}{\mu_i+1}$. This simulates a saccade step.

Alternatively, step 1 above can be replace with a draw from a random walk component at a known central fixation point, if called for in the experimental design.

This specification lends itself to two complementary interpretations. First, because the random walk component moves constantly with the point of regard, it can be viewed as a 'lens' that distorts the participant's perception of the underlying saliency map. The bandwidth and multiplier in this case describe the degree of distortion: a small bandwidth and large multiplier would distort the saliency map greatly, restricting the participant's view to a small region centered on the point of regard, while a large bandwidth and small multiplier would leave the saliency map relatively undisturbed, allowing greater perception of the entire visual field. From this perspective, the perceived saliency map is continuously being updated based on the current point of regard, and no distinction is made between fixations and saccades. The second interpretation treats movement across the map as a hidden Markov model with two hidden states: a fixation state, in which the next gaze point is drawn from the random walk component, and a saccade state, in which the next gaze point is instead drawn from any location on the saliency map. The multiplier then determines the transition probabilities between these two states, while the bandwidth only affects movements within the fixation state.

More specifically, the saliency map for each group is modeled as a Gaussian mixture model with K components, and has probability density function

$$f(x; \nu_1, ...\nu_K, \Sigma_1, ...\Sigma_K, \omega) = \sum_{i=1}^{K} \omega_i \phi_{\nu_i, \Sigma_i}(x)$$

where

$$\phi_{\nu,\Sigma}(x) = \frac{1}{2\pi \left|\Sigma\right|^{1/2}} \exp\left\{-\frac{1}{2}(x-\nu)^T \Sigma^{-1}(x-\nu)\right\}$$

is the probability density function of a bivariate Gaussian random variable with mean $\nu \in \mathbf{R}^2$ and 2x2 variance-covariance matrix Σ, and ω is a vector of non-negative mixture weights that sum to one. The mean ν_i and variance Σ_i of each mixture component are fixed; versions of this model in which the means and standard deviations are allowed to vary will be explored in further work. It is not necessary for Σ_i to be a diagonal matrix.

The vector of mixture weights, ω, for each group is assigned a Dirichlet prior, which has probability density function

$$f(\omega; \alpha) = \Gamma\left(\sum_{i=1}^{K} \alpha_i\right) \prod_{j=1}^{K} \frac{\omega_j^{a_j - 1}}{\Gamma(\alpha_j)}$$

with parameter vector α equal to a vector of 1's of length K, and $\Gamma(\cdot)$ is the Gamma function,

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$$

The parameters β_i and μ_i are individual-level random variables, and are assigned prior distributions $\beta_i \sim \text{Exp}\left(\widetilde{\beta}_{g_i}\right)$ and $\mu_i \sim \text{Exp}\left(\widetilde{\mu}_{g_i}\right)$ for each group g_i, where $\text{Exp}(\cdot)$ denotes the exponential distribution with probability density function

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

with rate parameter λ. To describe systematic group differences in the distributions generating these parameters, the mean parameters $\widetilde{\beta}_{g_i}$ and $\widetilde{\mu}_{g_i}$ are in turn assigned Exponential hyperprior distributions with rate parameters 0.01 (i.e. with mean and variance equal to 100).

Estimation of posterior distributions of all parameters on real and simulated data is done by Markov chain Monte Carlo (MCMC) simulation in Just Another Gibbs Sample (JAGS) [Plummer 2003]. For simulated data, two MCMC simulations are performed: in the first, the means and variances supplied to JAGS are the same ones used to simulate the data so that accuracy of individual component weights can be assessed, and in the second, the means and variances are selected by *k*-means [MacQueen et al. 1967; Hartigan and Wong 1979] performed on the set of all $\mathbf{z}_{it}$ points from both groups, and computing the means and variances of the points assigned to each cluster. For real data, means and variances are similarly selected by *k*-means since the true underlying saliency structure is unknown. MCMC results are analyzed in R version 2.14.0 [R Core Team 2013].

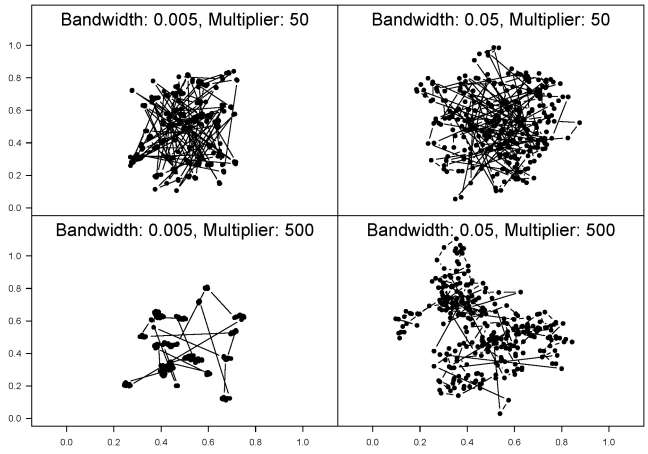

Figure 1: *Examples of simulated scanpaths with different values for bandwidth and multiplier parameters. Each scanpath is 300 frames in length.*

Examples of simulated scanpaths with various bandwidth and multiplier values are shown in Figure 1. The examples illustrate that the bandwidth controls movement within each fixation, with small values allowing little movement but large values permitting more. The multiplier, on the other hand, controls how many saccade steps are taken; small multipliers yield more saccade steps than large values. This is due to the fact that the multiplier controls the probability of taking the next step from the global saliency map, and this is virtually the only situation where a large saccade step may be taken.

This model generates a saccade in one frame regardless of its duration, so each saccade step must be interpreted as the endpoint of a saccade. In practice, due to physical constraints on eye movement, saccades are not taken instantaneously but are instead spread out across multiple consecutive frames, where the number of frames required to complete the saccade is dependent on the saccade length, the velocity of the eye, and the rate of frame capture. To account for this, steps for which the spatial distance between consecutive points of regard $\mathbf{z}_{i,t}$ and $\mathbf{z}_{i,t+1}$ exceeds a threshold can be interpolated with additional, intermediate points of regard. In this work, the intermediate points were assigned along the vector connecting $\mathbf{z}_{i,t}$ and $\mathbf{z}_{i,t+1}$ with velocity increasing linearly until the midpoint and decreasing linearly thereafter. Conversely, to splice out frames interior to saccades when analyzing real data, a preprocessing saccade-identification step was applied using the hysteresis method and saccade frames were removed.

2.2 Participants

The model was applied to scanpaths from a sample of 28 children between 8 and 43 months of age who participated in eye-tracking experiments at a university-based clinic at the Yale Child Study Center. Based on clinical evaluations at the age of 24-36 months, 15 of these children were assigned a best estimate clinical diagnosis of autism spectrum disorder (ASD). Diagnoses were based on standardized test results including the Mullen Scales of Early Learning (MSEL) [Mullen 1995] to measure language and cognitive skills and the Autism Diagnostic Observation Schedule (ADOS) [Lord et al. 2000] to assess the severity of autism symptoms, as well as medical and family history. The remaining 13 children were not diagnosed with any developmental disorders and were considered typically-developing (TD), confirmed through assessment with the MSEL and a parent interview concerning medical and developmental history. Sample characterization for the children is shown in Table 2. The study was approved by the Human Investigations Committee of Yale University, and an informed written consent was obtained from all parents prior to testing.

Table 2: *Sample characterization*

	ASD (n=15)	TD (n=13)	*p*-value
% Male	93.3%	58.3%	.09
% Caucasian	80.0%	84.6%	.86
% Hispanic	10.0%	0.0%	.71
Age (months)	24.0 (7.5)	21.9 (7.7)	.47
MSEL Visual Reception DQ	80.1 (25.1)	104.6 (25.7)	.06
MSEL Fine Motor DQ	85.8 (19.8)	105.5 (18.8)	.04
MSEL Receptive Language DQ	52.4 (34.5)	94.5 (19.8)	.003
MSEL Expressive Language DQ	61.8 (32.0)	85.6 (15.4)	.04
MSEL Nonverbal DQ	82.9 (21.2)	105.1 (19.1)	.03
MSEL Verbal DQ	57.1 (32.2)	90.1 (16.4)	.008
ADOS SA	15.1 (3.1)	-	-
ADOS RRB	5.0 (2.0)	-	-
ADOS Total	20.1 (4.3)	-	-
ADOS severity score	7.8 (1.7)	-	-

Abbreviations: MSEL, Mullen Scales of Early Learning; DQ, developmental quotient; ADOS, Autism Diagnostic Observation Schedule; SA, social affect; RRB, restricted and repetitive behaviors.

2.3 Stimuli and Apparatus

Stimuli consisted of twelve color images, six of faces and six of blocks. Face images were selected from the Karolinska Directed Emotional Faces database [Lundqvist et al. 1998], and depict affectively neutral female faces. Each image was 10.8 x 15.2 cm in size and was viewed on a 20" widescreen LCD monitor from a distance of 75 cm, so that each image subtended 8.2 x 11.6 degrees of visual angle. Each stimulus was presented as a separate trial, with faces and blocks presented in a predetermined but randomly assigned order.

Gaze trajectories were recorded at a sampling rate of 60 Hz using a SensoMotoric Instruments iView X^{TM}RED eye-tracking system [Sen 2005]. Eye-tracking data were pre-processed for data calibration and blink detection using custom software written in Matlab [MathWorks 2010].

2.4 Experimental Protocol

Stimuli were presented in the context of a fixed-level Visual Paired Comparison (VPC) paradigm [Fantz 1964] to assess visual discrimination and recognition of faces. In each trial, the stimulus was presented for 10 seconds in a familiarization phase, followed by side-by-side presentation of the original stimulus and a novel stimulus of similar type in a recognition phase. In the VPC paradigm, recognition is measured by a difference in looking time between the familiar and novel stimuli [Kaplan et al. 1995]. However, for our purposes, only data from the 10 second familiarization phase is

analyzed as our focus is on naturalistic gaze viewing patterns rather than recognition.

During eye-tracking, toddlers were seated in a car seat in a dark, soundproof room. Each eye-tracking session began with a short cartoon video to help the child get settled. A five-point eye-tracking calibration procedure was then initiated, with the calibration repeated if necessary until all five calibration points were successfully identified. Each stimulus image was preceded by a central fixation to refocus the attention of the child, and then the stimulus was displayed for as long as necessary for the child to attend to the image for 10 seconds. Because the presentation of the central fixation attracts attention to the center of screen, this could create an artifact of increased salience of this central region if the initial frames were included in analysis. To limit the effect of these central fixation artifacts on saliency estimation, the first 30 frames (0.5 seconds) of each scanpath were not included in analysis. Scanpaths for whom calibration uncertainty, defined as the average absolute deviation between experiment-wide calibrated scanpaths and a set of scanpaths cablibrated using nearest-neighbor calibration points [Shic 2008], was greater than 2° of visual arc over the whole experiment were excluded from analysis, with average calibration error less than $^3/_4{}^\circ$.

3 Results

3.1 Simulated Data

Bandwidth and multiplier parameters for 40 participants from two groups (20 participants per group) were generated according to the exponential distributions described above, with $\widetilde{\beta}_1 = 0.02$, $\widetilde{\beta}_2 = 0.04$, $\widetilde{\mu}_1 = 20$, and $\widetilde{\mu}_2 = 10$.

Saliency maps for the two groups were a mixture of 25 Gaussian components arranged in a 5x5 grid. For Group 1, components in the top three rows received mixture weights of $^{10}/_{160}$ and components in the bottom two rows received weights of $^1/_{160}$. For Group 2, mixture component weights were $^{10}/_{160}$ for the rightmost three columns, and $^1/_{160}$ for the leftmost two columns. Contour maps of these surfaces are shown in Figure 2. These maps were selected because they share regions of equal saliency at both high and low saliency levels, as well as regions where each group has higher saliency than the other. The saliency maps combined with the bandwidth and multiplier parameters were then used to simulate a single scanpath for each of 40 participants for T=600 frames (10 seconds).

Distributions of the posterior distributions for mixture weights, bandwidths, and multipliers for each image were simulated by 15,000 MCMC samples with every 30th sample retained, after a burn-in period of 5,000 samples. This provided a sample of 500 draws for each parameter's posterior distribution to be used in estimation of posterior means and quantiles. Credible intervals for each parameter were estimated by 2.5% and 97.5% sample quantiles of the posterior samples.

Estimation of group saliency maps via MCMC was quite accurate, regardless of whether the true 25 mixture components or the 30 k-means-determined mixture components (Figure 2) were supplied to the model. Estimation of individual bandwidth and multiplier parameters was also accurate, as shown in Figure 3; true and estimated parameter values were nearly equal, and in only a few cases did a 95% credible interval exclude the true value. Variability of the parameter estimates increased with parameter magnitude due to the use of the Exponential distribution as prior distributions in the model, for which the variance is equal to the mean. Estimates of $\widetilde{\beta}$ and $\widetilde{\mu}$ were also quite accurate (Figure 4). Posterior mean values

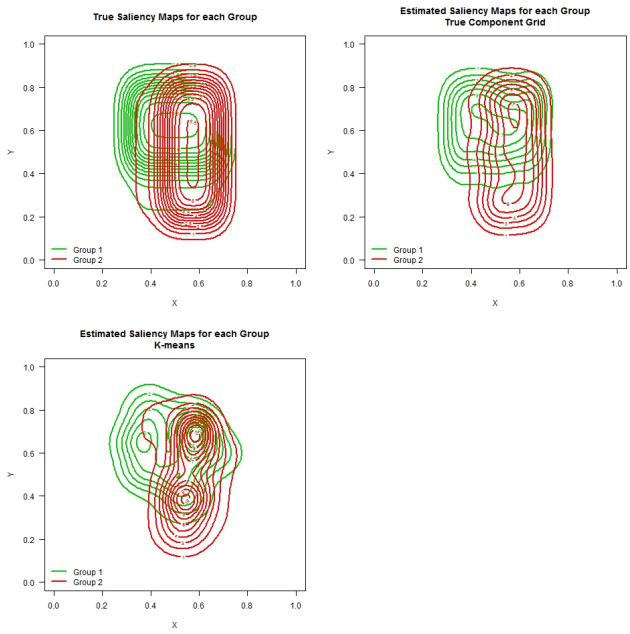

Figure 2: *Contour plots of group saliency maps for simulated data: true surface (top left), estimated via MCMC with true Gaussian mixture components (top right), and estimated via MCMC with Gaussian mixture components initialized by* k*-kmeans (bottom left).*

for $\widetilde{\beta}_1$, $\widetilde{\beta}_2$, $\widetilde{\mu}_1$, and $\widetilde{\mu}_2$ were 0.021, 0.036, 23.01, and 11.98, respectively, which are very close to the true values of 0.02, 0.04, 20, and 10.

3.2 Experimental Data

3.2.1 Saliency Results

For all images, saliency was overwhelmingly placed on inner face regions (eyes, nose, and mouth). Relative saliency between the mouth and the eyes was higher in the TD group compared to the ASD group, as shown in Figure 5. This distinction is especially noticeable in the second image, where TD children looked almost exclusively at the mouth and ASD children looked almost exclusively at the eyes. This diagnostic difference in preference for the eyes in the ASD group but preference for the mouth in the TD group is consistent with previously reported findings using regions of in-

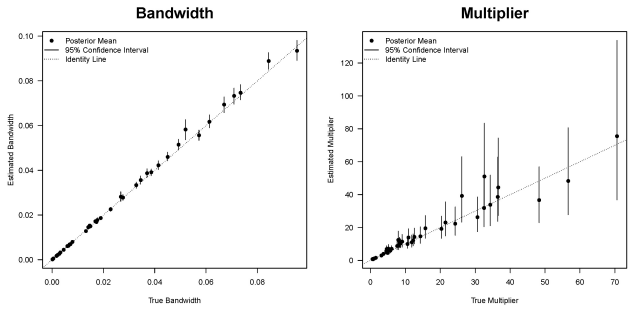

Figure 3: *Estimated vs. actual bandwidth parameters (left) and multiplier parameters (right) for each of the 40 simulated individual scanpaths.*

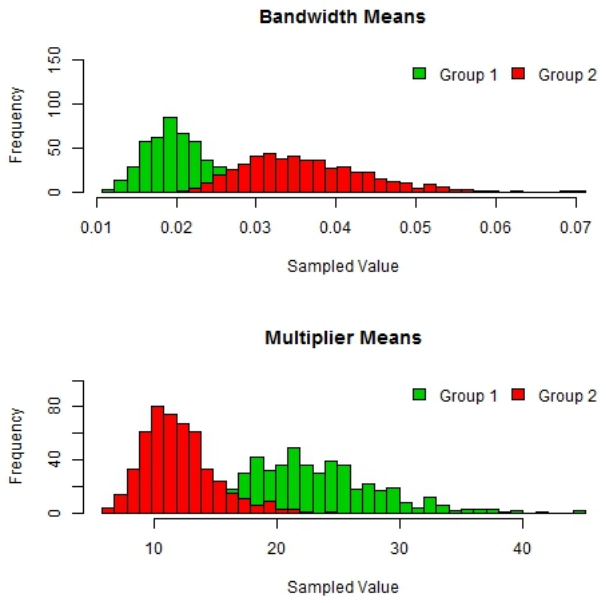

Figure 4: *Samples from the posterior distribution for $\widetilde{\beta}$ and $\widetilde{\mu}$ for both groups. True values are $\widetilde{\beta}_1 = 0.02$, $\widetilde{\beta}_2 = 0.04$, $\widetilde{\mu}_1 = 20$, and $\widetilde{\mu}_2 = 10$.*

terest analysis [Chawarska and Shic 2009].

To verify that these saliency results are indeed convergent with results from previous ROI approaches, we calculated the percentage of time each participant scanned ROIs corresponding to the eyes and the mouth. Attention towards eyes (%Eyes) is higher in ASD than in TD, while attention towards the mouth is lower in ASD relative to TD (Figure 6). The eye-to-mouth ratio (defined as %Eyes/(%Eyes+%Mouth)) is also larger in ASD relative to TD. While these results are not statistically significant (p-values are .46, .16, and .19 for %Eyes, %Mouth, and eye-to-mouth ratio, respectively), this is likely due to low power to detect effects in this sample. Cohen's d effect sizes are moderately large (d = 0.30, 0.56, and 0.51, respectively), and discrepancies in significance between this study and [Chawarska and Shic 2009] likely reflect the fact that the prior study includes data from two visits instead of the single visit analyzed here. However, despite this lack of statistical significance at the 0.05 level, the direction of the differences in group means is consistent with saliency results described above.

Results for a subset of block images are shown in Figure 7. For these images, substantial attention is drawn to the two circles, although attention is not divided equally among them. Preference for one circle over the other does not appear to be determined solely by the circle's color, as half of the images displayed preferences in both groups for the green circle (as in the first row of Figure 7), and the other half evidenced preference for the white circle (as in the second row of Figure 7).

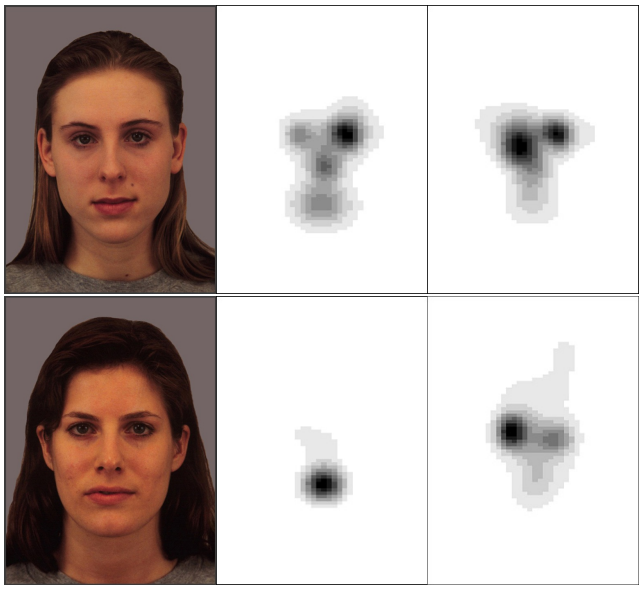

Figure 5: *Heatmaps of estimated saliency for faces: stimuli (left), estimated heatmaps for the TD group (middle), and estimated heatmaps for the ASD group (right).*

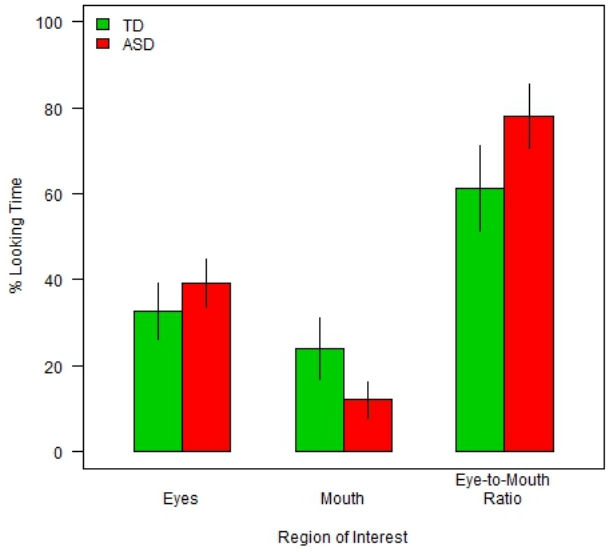

Figure 6: *Means +/- one standard error of TD and ASD groups on the key regions of interest (ROIs) of Eyes, Mouth, and Eye-to-Mouth Ratio, defined as (%Eyes)/(%Eyes + %Mouth).*

Also, certain surprising combinations of blocks in some images attract substantial attention, as the central arrangement of white and yellow blocks in the first row of Figure 7 exemplifies for both groups, but especially so for the ASD group. Even considering the possible explanation of enhanced saliency due to brightness, it is not clear why this region should appear so salient while similar regions in other images (such as the white and yellow blocks in the lower left corner of the second block image in Figure 7) do not attract comparable levels of attention. The fact that such re-

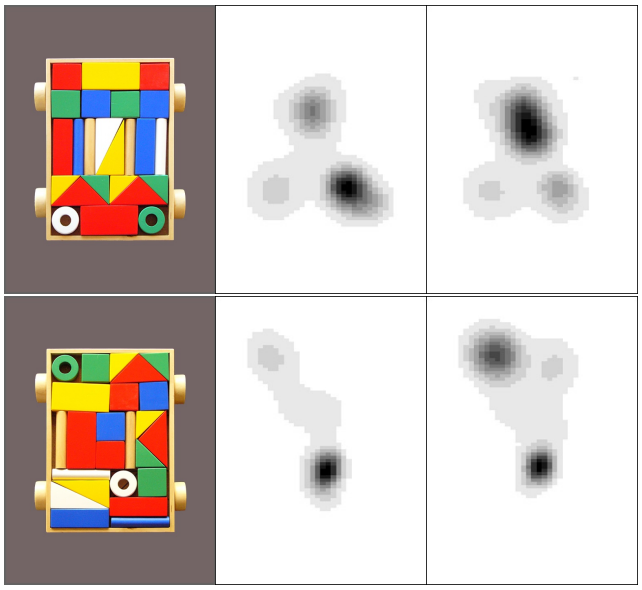

Figure 7: *Heatmaps of estimated saliency for blocks: stimuli (left), estimated heatmaps for the TD group (middle), and estimated heatmaps for the ASD group (right).*

gions are identified by this model as highly salient despite their non-obviousness speaks to the advantage of this model over ROI-based techniques. The very fact that such regions are difficult to identify beforehand as potentially salient means that they would never be selected as regions in ROI analysis, and their relevance to group differences in saliency would not be noticed.

3.2.2 Bandwidth and Multiplier Results

When estimated values of $\widetilde{\beta}$ in TD and ASD groups were compared on faces, differences were quite large for three of the six images, with 95% credible intervals for the differences excluding zero (Figure 8, top). In these three images, the TD group had a larger $\widetilde{\beta}$ value than the ASD group, indicating a more focused gaze during fixations for the ASD group. Cohen's *d* effect sizes for the three images were 2.35, 2.17, and 1.47. Group differences in $\widetilde{\beta}$ for block images were much smaller, with Cohen's *d* effect sizes ranging from 0.21 to 0.53, and all six 95% credible intervals overlapped zero.

The fact that large differences in $\widetilde{\beta}$ were discovered for only three face images may be explained by the existence of extreme values. For these three images, at least one participant from the TD group had an estimated β_i well outside the range of values for the other participants, which had a large effect on the TD group mean. Because these extreme values represented different individuals across images (that is, it was not the same participant producing the extreme values in every image) and because it is not clear whether such large β_i values truly represent outliers or indicate an uncommon but important phenonemon which our small sample cannot adequately reflect, these participants were not removed from analysis. Future work will explore selection of prior specifications on $\widetilde{\beta}$ and $\widetilde{\mu}$ that are less susceptible to outliers than the Exponential distribution.

When estimated values of $\widetilde{\mu}$ were compared, no large group differences were evident for either face or block images (Figure 8, bottom). Cohen's *d* effect sizes ranged from -0.15 to 0.88 for faces, and from -0.21 to 0.86 for blocks, and all 95% credible intervals

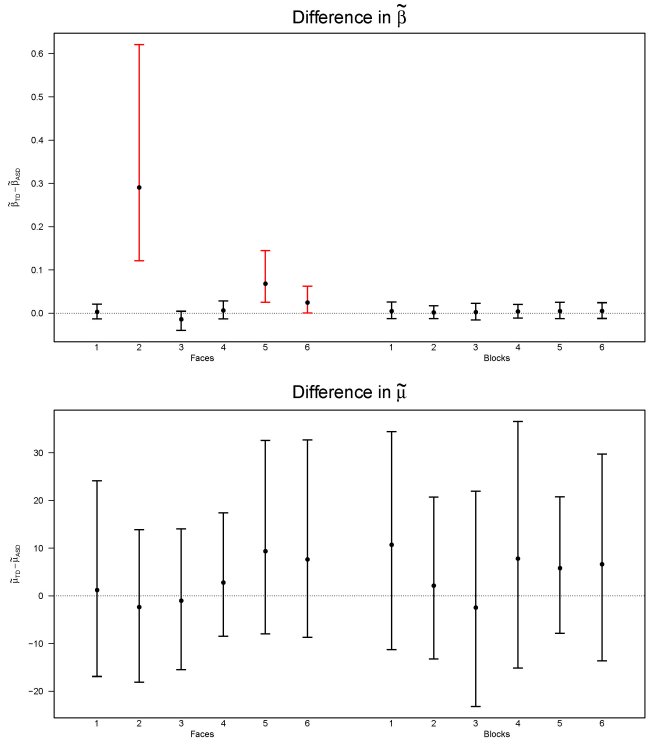

Figure 8: *Means and 95% credible intervals for group differences in bandwidth mean ($\widetilde{\beta}$, top) and multiplier mean ($\widetilde{\mu}$, bottom), for each of the twelve image stimuli. Credible intervals drawn in red indicate exclusion of zero from the interval.*

overlapped zero.

4 Discussion

We have presented here a new statistical model that integrates both saliency and frame-by-frame dynamic gaze properties into a unified framework. It accurately recovers known parameters from simulated data, and detects differences between typically-developing and ASD groups in saliency and dynamic scene processing when applied to real data.

Saliency differences found between ASD and TD groups include a preference for the mouth over the eyes in typical development, but a reversed preference in ASD. This difference was also observed in attention to eye and mouth regions when the same data was analyzed using region of interest analysis, reflecting a convergence of results using different approaches. This finding is consistent with a similar, previously published study [Chawarska and Shic 2009], where the disparity in statistical significance of findings between that study and this one can be attributed to different methodologies (a 2 Group x 2 Visit ANOVA model in the former versus a single visit in the current study).

Saliency differences were also found in response to block stimuli. While attention to certain aspects of the block images, like the two circles contained in each, were expected and indeed observed in both groups, the relative salience of the two circles to each other differed widely by group and by image. Why this should be so is unclear, as the differences in salience do not appear to be driven by the color or position of the circles. Other regions of the block images besides the circles attracted differing levels of salience from

each group as well, and the reasons for the enhanced salience of these regions is likewise unclear. In fact, these regions would have been missed entirely if only the circles had been coded as key regions in an ROI analysis. The fact that the proposed model, being free of assumptions regarding the importance of regions, could identify such regions anyway means that the model does not rely on *a priori* assumptions regarding salience and can estimate saliency properties in a purely bottom-up, data-driven approach. ROI-based techniques are restricted in that they can only compare salience in known regions, not discover new ones.

Differences in dynamic movement in fixations at the group level, represented by a difference in $\widetilde{\beta}$ between groups, were also found in response to face stimuli, with typically developing children exhibiting a significantly larger value of $\widetilde{\beta}$ and hence decreased variability of movement within fixations. This difference in fixational movement was not observed during block stimuli, nor were group differences evident in the frequency of saccades (represented by $\widetilde{\mu}$) during either facial or block stimuli. Although this is an intriguing finding that has not been previously reported, there is a possibility that it may be due to the employment of prior distributions in our model that exhibit some susceptibility to outliers. Robustness of this finding to changes in modeling assumptions will be assessed in future work.

The saliency map in our model is parametrized as a Gaussian mixture model with known components and unknown weights, and is not built up from low-level properties of the scene like color, brightness, and contrast. However, the number, location, size, and orientation of mixture components can be flexibly chosen to reflect knowledge of, and assumptions regarding, the underlying low-level features of the scene, for instance by placing mixture components with small variances on regions of high brightness or contrast. The model can then be used to estimate the relative salience of such regions, or even to estimate the dynamic patterns of eye movement on an entirely specified saliency map. In this work we have selected component means and variances in a data-driven fashion through *k*-means, but the flexibility of the model presented here allows the saliency map to reflect bottom-up image features if desired.

An important advantage of this model is the ability to make statistical inferences about the estimated saliency map. The posterior samples for the mixture component weights were used to calculate posterior means, and thus estimated saliency heatmaps similar to what could be obtained by other saliency-estimation approaches. However, the posterior samples also allow for estimation of the variance of the mixture weights, and thereby quantify the variability of each estimated saliency map. It is possible, for instance, to calculate 95% credible intervals for the difference between the estimated TD and ASD saliency maps at any point of the screen, and to identify regions where this pointwise interval does not overlap zero; these results were not included here due to limited space. These posterior inferences are possible with this model, but not easily made using other heatmap approaches that only compute an estimated saliency heatmap with no measure of variance.

4.1 Limitations and Future Directions

The means and variances of the mixture components making up the saliency map in this model are fixed, and only mixture weights are assigned distributions. Further extensions of the model include allowing the means and variances to vary as well. While this is straightforward in principle by the addition of prior distributions on the means and precisions of the Gaussian components, in practice this can add substantial computational cost to the model as well as complicate convergence of the MCMC to the stationary distribution. Also, the interpolation of single-frame saccades so that they span multiple frames is currently an additional step and not an integrated feature of the model, and incorporation of this step into the model is an continuing area of research. Additional extensions of this work include refinement of the prior distributions on bandwidth and multiplier parameters to be more robust against outliers, investigation of the convergence of the MCMC and sensitivity of model results to changes in prior distributions, and the combination of multiple images into a single hierarchical model rather than separate analyses for each image.

Acknowledgements

This study was supported by Autism Speaks #7614 (D.C.), NIMH grant T32 MH18268 (D.C. & F.S.), R03 MH092618-01A1 (F.S.), CTSA Grant Number UL1 RR024139 (F.S.), Expedition in Computing (award #1139078) (F.S), NICHD grant PO1 HD003008 Project 1 (K.C.), NIMH STAART grant U54 MH66494 (K.C.), NIMH R01 MH087554 (K.C.), and by the Associates of the Child Study Center.

References

AMERICAN PSYCHIATRIC ASSOCIATION. 2013. *Diagnostic and statistical manual of mental disorders*, 5th ed. Washington, DC.

BROCKMANN, D., AND GEISEL, T. 2000. The ecology of gaze shifts. *Neurocomputing 32*, 643–650.

BRUCE, N., AND TSOTSOS, J. 2005. Saliency based on information maximization. In *Advances in neural information processing systems*, 155–162.

CHAWARSKA, K., AND SHIC, F. 2009. Looking but not seeing: Atypical visual scanning and recognition of faces in 2 and 4-year-old children with autism spectrum disorder. *Journal of Autism and Developmental Disorders 39*, 12, 1663–1672.

DALTON, K. M., NACEWICZ, B. M., JOHNSTONE, T., SCHAEFER, H. S., GERNSBACHER, M. A., GOLDSMITH, H., ALEXANDER, A. L., AND DAVIDSON, R. J. 2005. Gaze fixation and the neural circuitry of face processing in autism. *Nature neuroscience 8*, 4, 519–526.

FANTZ, R. L. 1964. Visual experience in infants: Decreased attention to familiar patterns relative to novel ones. *Science 146*, 3644, 668–670.

FENG, G. 2006. Eye movements as time-series random variables: A stochastic model of eye movement control in reading. *Cognitive Systems Research 7*, 1, 70–95.

HARRIS SR, R. L., GLOVER, B. J., AND SPADY JR, A. A. 1986. Analytical techniques of pilot scanning behavior and their application. Tech. rep.

HARTIGAN, J. A., AND WONG, M. A. 1979. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 28*, 1, 100–108.

ITTI, L., AND KOCH, C. 2000. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research 40*, 10, 1489–1506.

ITTI, L., KOCH, C., AND NIEBUR, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 20*, 11, 1254–1259.

JOHNSON, S. P., AMSO, D., AND SLEMMER, J. A. 2003. Development of object concepts in infancy: Evidence for early learning in an eye-tracking paradigm. *Proceedings of the National Academy of Sciences 100*, 18, 10568–10573.

KAPLAN, P. S., GOLDSTEIN, M. H., HUCKEBY, E. R., AND COOPER, R. P. 1995. Habituation, sensitization, and infants' responses to motherse speech. *Developmental Psychobiology 28*, 1, 45–57.

KLIN, A., JONES, W., SCHULTZ, R., VOLKMAR, F., AND COHEN, D. 2002. Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of General Psychiatry 59*, 809–816.

KOCH, C., AND ULLMAN, S. 1985. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology 4*, 4, 219–227.

LEWKOWICZ, D. J., AND HANSEN-TIFT, A. M. 2012. Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences 109*, 5, 1431–1436.

LORD, C., RISI, S., LAMBRECHT, L., COOK JR, E. H., LEVENTHAL, B. L., DILAVORE, P. C., PICKLES, A., AND RUTTER, M. 2000. The autism diagnostic observation schedule-generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of autism and developmental disorders 30*, 3, 205–223.

LUNDQVIST, D., FLYKT, A., AND ÖHMAN, A. 1998. The karolinska directed emotional faces. *Stockholm, Sweden: Karolinska Institute*.

MACQUEEN, J., ET AL. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, California, USA, 14.

MATHWORKS. 2010. *MATLAB 7.0 (Release 14)*. Natick, Massachusetts.

MULLEN, E. 1995. *Mullen Scales of Early Learning (AGS Edition ed.)*. American Guidance Service, Inc., Circle Pines, MN.

NORBURY, C. F., BROCK, J., CRAGG, L., EINAV, S., GRIFFITHS, H., AND NATION, K. 2009. Eye-movement patterns are associated with communicative competence in autistic spectrum disorders. *Journal of Child Psychology and Psychiatry 50*, 7, 834–842.

PLUMMER, M. 2003. Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). March*, 20–22.

R CORE TEAM. 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RENNINGER, L. W., VERGHESE, P., AND COUGHLAN, J. 2007. Where to look next? eye movements reduce local uncertainty. *Journal of Vision 7*, 3.

RICHARDSON, D. C., AND DALE, R. 2005. Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science 29*, 6, 1045–1060.

RIMEY, R. D., AND BROWN, C. M. 1991. Controlling eye movements with hidden markov models. *International Journal of Computer Vision 7*, 1, 47–65.

SALVUCCI, D. D., AND GOLDBERG, J. H. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, ACM, 71–78.

SENSOMOTORIC INSTRUMENTS. 2005. *iView X (TM) RED*.

SHIC, F., CHAWARSKA, K., BRADSHAW, J., AND SCASSELLATI, B. 2008. Autism, eye-tracking, entropy. In *Development and Learning, 2008. ICDL 2008. 7th IEEE International Conference on*, IEEE, 73–78.

SHIC, F., CHAWARSKA, K., AND SCASSELLATI, B. 2008. The amorphous fixation measure revisited: with applications to autism. In *30th Annual Meeting of the Cognitive Science Society, Washington, DC*.

SHIC, F., CHAWARSKA, K., AND SCASSELLATI, B. 2008. The incomplete fixation measure. In *Proceedings of the 2008 Symposium on Eye Tracking Research and Applications, Savannah, Georgia*, 111–114.

SHIC, F. 2008. *Computational methods for eye-tracking analysis: Applications to autism*. PhD thesis, Yale University.

SIMMONS, D. R., ROBERTSON, A. E., MCKAY, L. S., TOAL, E., MCALEER, P., AND POLLICK, F. E. 2009. Vision in autism spectrum disorders. *Vision Research 49*, 22, 2705–2739.

VAN DER LANS, R., PIETERS, R., AND WEDEL, M. 2008. Eye-movement analysis of search effectiveness. *Journal of the American Statistical Association 103*, 482, 452–461.

YAMATO, J., OHYA, J., AND ISHII, K. 1992. Recognizing human action in time-sequential images using hidden markov model. In *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on*, IEEE, 379–385.