Vision Research 49 (2009) 2992–3000

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Saliency does not account for fixations to eyes within social scenes

Elina Birmingham [a,*], Walter F. Bischof [b], Alan Kingstone [c]

[a] California Institute of Technology, Division of the Humanities and Social Sciences, Pasadena, CA, USA
[b] University of Alberta, Department of Computing Science, Edmonton, Alberta, Canada
[c] University of British Columbia, Department of Psychology, Vancouver, British Columbia, Canada

ARTICLE INFO

Article history:
Received 16 May 2008
Received in revised form 15 September 2009

Keywords:
Visual saliency
Gaze selection
Social Attention
Eye movements
Scene perception

ABSTRACT

We assessed the role of saliency in driving observers to fixate the eyes in social scenes. Saliency maps (Itti & Koch, 2000) were computed for the scenes from three previous studies. Saliency provided a poor account of the data. The saliency values for the first-fixated locations were extremely low and no greater than what would be expected by chance. In addition, the saliency values for the eye regions were low. Furthermore, whereas saliency was no better at predicting early saccades than late saccades, the average latency to fixate social areas of the scene (e.g., the eyes) was very fast (within 200 ms). Thus, visual saliency does not account for observers' bias to select the eyes within complex social scenes, nor does it account for fixation behavior in general. Instead, it appears that observers' fixations are driven largely by their default interest in social information.

1. Introduction

Research has shown that humans have a fundamental bias to attend to the eyes of other people (Birmingham, Bischof, & Kingstone, 2008a, 2008b; Henderson, Williams, & Falk, 2005; Itier, Villate, & Ryan, 2007; Pelphrey et al., 2002; Walker-Smith, Gale, & Findlay, 1977; Yarbus, 1967). For instance, when presented with pictures of faces, observers tend to fixate (look at) the internal features of the face, with a particular focus on the eyes (e.g., Henderson et al., 2005; Pelphrey et al., 2002; Walker-Smith et al., 1977). Furthermore, this bias to fixate the eyes of a face emerges very early on, often within the first fixation (van der Geest, Kemner, Verbaten, & van Engeland, 2002).

Most researchers have interpreted this evidence as demonstrating that humans, from a very early age, have a strong preferential bias to attend to the eyes of others because eyes are extraordinary powerful sources of social information that convey such things as emotional and attentional states (e.g., Baron-Cohen, Wheelright, & Joliffe, 1997; Emery, 2000; Nummenmaa, 1964). Indeed, failure to appreciate the social importance of eyes has been proposed to be a key causal factor in atypical social disorders such as autism (Baron-Cohen, 1994; Dalton et al., 2005; Klin, Jones, Schultz, Volkmar, & Cohen, 2002).

A potentially significant limitation of the studies thus far, however, has been that most investigations have presented observers with a face alone, isolated from its body and surrounding context. In such situations the eyes have relatively high contrast and are visually conspicuous within the face (Kobayashi & Koshima, 1997), and thus it is very possible that observers look at the eyes not because of the high-level social meaning that they convey but because low-level *visually salient* features make the eyes stand out from their surround, drawing attention automatically (Itti & Koch, 2000; Itti, Koch, & Niebur, 1998; Koch & Ullman, 1985). This influence of saliency might be expected to be especially strong for early fixations to the eyes, before top-down influences are thought to come into play (e.g., Henderson, Weeks, & Hollingworth, 1999; Parkhurst, Law, & Niebur, 2002).

One way to address this issue directly is to present the face along with other objects, for instance, within the context of a complex natural scene that contains many different items for possible selection. Recent research has demonstrated that when these complex scenes are presented, observers quickly fixate the eyes of the people in the scenes. This suggests that observers select the eyes because of the social information they provide rather than because they are the most salient visual items in the scenes (e.g., Birmingham et al., 2008a; Smilek, Birmingham, Cameron, Bischof, & Kingstone, 2006). Converging support for this interpretation is that the early bias for eyes is steady across a variety of different scenes and tasks, although the overall interest in the eyes can be enhanced by explicitly asking observers to assess the social attentional states within the scenes (Birmingham et al., 2008b). Currently the working hypothesis is that observers have an early default bias to inspect the eyes of others, not because they are visually salient within the scene, but because they understand them to be socially communicative stimuli that provide important information about a social scene (Birmingham, Bischof, & Kingstone, 2009).

* Corresponding author.
E-mail address: elinab@hss.caltech.edu (E. Birmingham).

0042-6989/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2009.09.014

Note that this hypothesis depends critically on the untested assumption that visual saliency of the eyes contributes little, if anything at all, to observers' preference to look at the eyes of the people in the scenes. Fletcher-Watson, Leekam, Benson, Frank, and Findlay (2009) touched on this issue, and reported that the saliency of various features (e.g., color, orientation) did not adequately differentiate between the scan paths of individuals with autism and typically developing controls who viewed scenes containing one person. However, two aspects of this study preclude it from addressing our central question of whether the saliency of people's eyes in scenes can explain why they attract attention. For instance, Fletcher-Watson et al. did not obtain the high fixation frequencies for eyes that have been reported in previous work (e.g., Birmingham et al., 2008a; Smilek et al., 2006). One cannot obtain a strong test of how the saliency model accounts for fixations to the eyes of others when fixations to the eyes are relatively infrequent. (Note that the failure to find many fixations to the eyes may be due to the scenes that Fletcher-Watson et al. used, i.e., in one of the example scenes presented in Fletcher-Watson et al., the eyes in the scene are not even visible to the observer.)

Similarly, in another recent study by Cerf, Harel, Einhäuser, and Koch (2008) it was reported that visual saliency could not *fully* account for where observers look within social scenes. Cerf et al. showed that the model that best predicted where observers committed fixations within scenes with people was a saliency model combined with a face-detection model. This combined-model outperformed the saliency model alone, which in turn was found to perform slightly above chance. This result provided support for the notion of a specialized "face" channel in the visual system, one that rapidly detects faces for selection (e.g., Viola & Jones, 2001). While this study sheds light on observers' interest in the faces of people, and showed that this interest cannot be fully explained by the standard bottom-up saliency models, it did not quantify the extent to which observers fixated the eyes alone (i.e., there was no 'eye' ROI as this was not the focus of their study). Thus, we do not know whether observers showed an early bias to look at the eyes or if other features of the face attracted attention. Furthermore, if there was a specific bias to look at the eyes, it is not possible to ascertain whether the saliency of the eye region may have contributed to this bias.

These limitations notwithstanding, it is also the case that both Cerf et al. (2008) and Fletcher-Watson et al. (2009) assessed the saliency model across all fixations (across 2- and 3-s of viewing, respectively) – and not for the very first fixations – which is when the saliency model is thought to be most likely to account for performance (Henderson et al., 1999; Parkhurst et al., 2002; but see Tatler, Baddeley, & Gilchrist, 2005 for evidence that this is not the case). In summary, no study has examined explicitly the role that visual saliency plays in the early bias to fixate the eyes of people in complex natural scenes when this fixation placement to the eyes is robust and reliable. The present study set out to address precisely this issue.

1.1. Data analyses

1.1.1. First fixations

In order to assess whether there is a clear early bias to fixate the eyes across a variety of social scenes in our data set, we computed the location of the first fixation made by observers across three experiments. Experiment 1 is Birmingham, Bischof, and Kingstone's (2008a) and Experiment 2 is Birmingham, Bischof, and Kingstone's (2008b). Experiment 3 is Birmingham, Bischof, and Kingstone's (2007) with the data from additional participants and hence more power.

1.1.2. Basic performance of the saliency model

We were interested in how well the classic saliency model of Itti and Koch (2000) accounted for first fixation data across a variety of experimental data. We chose to analyze the first fixation data because saliency is predicted to be most influential early on in scene viewing, and so analyzing later (or all) fixations might underestimate the role of saliency in determining fixation position (e.g., Fletcher-Watson et al., 2009).

The most popular saliency model (Itti & Koch, 2000; Itti et al., 1998; Koch & Ullman, 1985) is based on feature integration theory of visual search (Treisman & Gelade, 1980). The basic assumption of the saliency model is that before attention is focused on any one aspect of the scene, the visual field is processed rapidly and in parallel for basic visual features. The output of this 'preattentive' processing of the scene is the construction of several topographic feature maps, each coding for local differences in a particular feature (e.g., changes in intensity, color (red–green; blue–yellow), and edge orientation). These feature maps are computed across several spatial scales, varying from fine to course, and then combined across scales into three "conspicuity maps", one for intensity, one for color, and one for orientation. The three conspicuity maps are normalized to a fixed range (e.g., 0–1) and combined into the final saliency map, which is a modality-independent map coding for conspicuous (i.e., salient) scene locations. It is this saliency map that guides the deployment of attention[1]. According to the 'winner-take-all' hypothesis, the most salient location in the map 'wins' focal attention (see Fig. 1B of Itti & Koch, 2000). After the winner is attended, attention moves along the remaining salient locations in order of decreasing saliency. Furthermore, an inhibitory mechanism (inhibition of return) is implemented which prevents attention from returning to previously attended (salient) locations.

For each experiment, we computed the average saliency of fixated scene locations and compared this value to two control values. The first control value was the average saliency of random locations sampled uniformly from the image (called "uniform-random"). To control for the known bias to fixate the lower central regions of scenes (see Tatler, 2007 for more information on the central bias), the second control value was the average saliency of random locations sampled from the smoothed probability distribution of all first-fixation locations from participants' eye movement data across all scenes (called "biased-random"). These comparisons allowed us to determine whether the saliency model accounted for first fixation position above what would be expected by chance.

1.1.3. Latency analysis

If saliency does contribute to placement of the first fixation within complex social scenes, it is most likely to do so for fixations resulting from early (faster) saccades than for fixations resulting from later (slower) saccades. This early effect of saliency has been demonstrated in more impoverished displays (e.g., van Zoest, Donk, & Theeuwes, 2004), and yet this analysis has not been conducted for complex social scenes. If the contribution of bottom-up mechanisms is much higher for early saccades than for later ones, one would expect the influence of social information on fixation placement to occur relatively late. That is, one would expect eyes to be fixated quite slowly relative to other, less social but more salient areas of the scene. Thus, we examined: (a) fixated saliency for early versus later first fixations and (b) saccade latencies for fixations landing on each region of interest (e.g., Fletcher-Watson, Findlay, Leekam, & Benson, 2008).[2]

[1] Other models propose early influence of both bottom-up saliency and top-down factors on the control of attention (e.g., Torralba, 2003). All of these more flexible saliency models, however, assume that an initial saliency map is computed, and that top-down 'maps' simply combine with the visual saliency maps to control the allocation of attention.

[2] We thank an anonymous reviewer for suggesting this analysis.

1.1.4. Saliency of regions of interest

We computed the relative saliency of the eyes versus the other regions in the scene. Previous studies have shown that the eyes and heads are both highly likely to be fixated within the first one or two fixations (Birmingham et al., 2008b; Cerf et al., 2008), and more so than any other region (e.g., bodies, foreground objects), except for the background of scenes, which also tends to receive several first fixations (Birmingham et al., 2008b). But how salient are the eyes and heads relative to the rest of the scene? Using saliency maps from Itti & Koch's (2000) model, we compared the saliency of eyes, heads, bodies, foreground objects, and background. This allowed us to determine whether the eyes or heads were more salient than the other regions, which might explain why they attracted initial fixations.

First we will provide a brief summary of the aims from each experiment. Then we will present the methodology shared between all three experiments, followed by a description of methods unique to each experiment. Finally, we present the first fixation and saliency analyses for each experiment.

1.2. Experiment 1

Experiment 1 examined whether where observers look within complex scenes is influenced by social content and activity in a scene (Birmingham et al., 2008a). We monitored observers' eye movements while they freely viewed real-world social scenes that contained either one person or three persons, who were either doing something (e.g. reading a book; *Active scenes*) or were doing nothing (e.g., just sitting on their own; *Inactive* scenes).

1.3. Experiment 2

Are fixations to the eyes affected by the task given to observers? To get at this question we presented participants with 20 complex real-world social scenes (Birmingham et al., 2008b). Participants were given one of three possible tasks. For one group, participants were asked to simply look at the scenes that they were shown (Look task). Participants in a second group were asked to describe the scene (Describe task). Participants in a third group were asked to describe where people in the scene were directing their attention (Social Attention task).

1.4. Experiment 3

In Experiment 3 we were interested in whether observers perceive the eyes to be informative for remembering social scenes (Birmingham et al., 2007). Observers were assigned randomly to two groups. One group was told that they would later be asked to recognize the scenes in a test session (Told group); another group was not informed of the later memory test and simply asked to freely view the images (Not Told group). Both groups were subsequently given a memory test, in which scenes from the pretest session were presented along with scenes that were never seen before.

2. Method

2.1. Participants

All participants had normal or corrected to normal vision, and were naïve to the purpose of the experiment. Each participant received course credit for participation in a 1-h session. Specific details about participants included in each experiment are presented later.

2.2. Apparatus

Eye movements were monitored using an Eyelink II tracking system. The on-line saccade detector of the eye tracker was set to detect saccades with an amplitude of at least 0.5° of visual angle (°), using an acceleration threshold of 9500°/s^2 and a velocity threshold of 30°/s.

2.3. Stimuli

Full color digital photos were taken of different rooms in the UBC Psychology building. Image size was 36.5 × 27.5 (cm) corresponding to 40.1° × 30.8° at the viewing distance of 50 cm, and image resolution was 800 × 600 pixels. Specific details about stimuli for each experiment are presented later.

2.4. Procedure

Participants were seated in a brightly lit room, and were placed in a chin rest so that they sat approximately 50 cm from the computer screen. Before the experiment, a calibration procedure was conducted. Participants were instructed to fixate a central black dot, and to follow this dot as it appeared randomly at nine different places on the screen. This calibration was then validated with a procedure that calculates the difference between the calibrated gaze position and target position and corrects for this error in future gaze position computations. After successful calibration and validation, the scene trials began. Specific procedural details for each experiment are presented below.

2.5. Experiment 1

2.5.1. Participants

Twenty undergraduate students from the University of British Columbia participated in this experiment.

2.5.2. Stimuli

Forty scenes were presented. Each scene contained either one or three persons, who were either doing something (*Active*) or nothing (*Inactive*). All scenes were comparable in terms of their basic layout: each room had a table, chairs, objects, and background items.

2.5.3. Procedure

Participants were told that they would be shown several images, each one appearing for 15 s, and that they were to simply look at these images. At the beginning of each trial, a fixation point was displayed in the center of the computer screen in order to correct for drift in gaze position. Participants were instructed to fixate this point and then press the spacebar to start a trial. The 40 pictures were shown in a random order. Each picture was shown in the center of the screen and remained visible until 15 s had passed, after which the picture was replaced with the drift correction screen. This process repeated until all pictures had been viewed.

2.6. Experiment 2

2.6.1. Participants

Thirty-nine undergraduate students from the University of British Columbia participated.

2.6.2. Stimuli

Participants viewed 20 scenes, each containing either one or three persons, who were either *Active* or *Inactive*.

2.6.3. Procedure

Participants were told that they would be shown several images, each one appearing for 15 s. Each participant was randomly assigned to one of three tasks. The *Look* group was told to simply "look at" each image. The *Describe* group was told to "look at, and then describe" each image. The *Social Attention* group was asked to "describe where people in the picture are directing their attention". The Describe and Social Attention groups were given an answer booklet, with space available for answering their assigned question for each picture in the order presented. Participants were told that they would have to write their answer for any given picture *after* the trial was over, i.e., after the image disappeared, and that they could take as long as they needed to write their answer.

At the beginning of each trial, a fixation point was displayed in the center of the computer screen in order to correct for drift in gaze position. Participants were instructed to fixate this point and then press the spacebar to start a trial. The 20 pictures were shown in a random order. Each picture was shown in the center of the screen and remained visible until 15 s had passed, after which the picture was replaced with the drift correction screen. During this time participants in the *Describe* and *Social Attention* groups wrote an answer using the booklet provided. This process repeated until all pictures had been viewed.

2.7. Experiment 3

2.7.1. Participants

Eighteen undergraduate students from the University of British Columbia participated.

2.7.2. Stimuli

Pretest session (15 scenes: three rooms, five scene types). Twelve of the 15 study scenes were "people scenes". These scenes contained a variety of social situations containing one or three persons, who were either *Active* or *Inactive*. Three of the fifteen pretest scenes were "No people scenes", containing a single object resting on the table.

Test session (56 scenes: eight rooms, seven scene types). Thirty-two of the test scenes were "People scenes" as above (12 old, 20 new). Sixteen of the test scenes were "No people scenes", containing one or three objects resting on the table (three old, 13 new). Eight additional (new) scenes contained one person doing something unusual, such as sitting with a Frisbee on his head. These scenes were included to keep the participants interested, but were not analyzed.

2.7.3. Procedure

Participants were told that they would be shown several images, each one appearing for 10 s. *Pretest session*: Each participant was randomly assigned to one of two instruction groups. The *Told* group was told that they would be shown 15 images, and that they would be asked to recognize each image in a later memory test. The *Not Told* group was told to simply "look at" each image, and was not informed of the later memory test. After the pretest session, a brief questionnaire was given to participants asking them about their impressions of the experiment.

Test session: Both groups (Told, Not Told) were informed that they would be shown 56 images, and that they were to view each one and then decide if the image was OLD (i.e., they had seen it in the pretest session), or NEW (i.e., they had never seen it before). After an image was presented, a response screen appeared asking them to respond with '1' on the keyboard if they thought the image was OLD, and '2' on the keyboard if they thought the image was NEW. Participants had an unlimited amount of time to respond.

At the beginning of each trial, a fixation point was displayed in the center of the screen in order to correct for drift in gaze position. Participants were instructed to fixate this point and then press the spacebar to start a trial. A picture was then shown in the center of the screen until 10 s had passed, after which the picture was replaced with the response screen (test session) or the drift correction screen (pretest session). In the test session, after the participant entered a response, the drift correction screen appeared in preparation for the next trial. This process repeated until all pictures had been viewed.

3. Results

3.1. Eye movement data handling

For each image, an outline was drawn around each region of interest (e.g., "eyes") and each region's coordinates and area were recorded. We defined the following regions in this manner: eyes, heads (excluding eyes), bodies (including arms, torso and legs), foreground objects (e.g., tables, chairs, objects on the table) and background objects (e.g., walls, shelves, items on the walls). See Fig. 1 for examples of these regions. First we will present a very brief description of the results for the overall fixation proportions (the reader can refer to the published articles for more detailed results), followed by a more in-depth analysis of the first fixation data, and then finally the saliency analyses.

3.2. Fixation proportions

3.2.1. Experiment 1

We had hypothesized that observers would demonstrate a preferential bias to fixate the eyes of the people in the scene, although other items would also receive attention. The overall fixation preferences across the entire viewing period showed that observers fixated primarily the eyes of people in the scenes, followed by heads. Furthermore, fixations to the eyes were enhanced when the social content (number of people) and activity in the scenes were high (see Birmingham et al., 2008a for more details).

3.2.2. Experiment 2

The overall preference to look at eyes was enhanced when the task was to report on the Social Attention within a scene. Nevertheless, it was also the case that participants selected the eyes more than any other stimulus, regardless of task instruction (see Birmingham et al., 2008b).

3.2.3. Experiment 3

The overall fixation preferences showed that observers in the Told group fixated the eyes more frequently than observers in the Not Told group, both in the pretest session and the test session,[3] suggesting that the Told group perceived the eyes to be informative for remembering the scenes (see Birmingham et al., 2007, and Footnote 3, for more details).

[3] We reported this finding in Birmingham et al. (2007). The same effects emerged with the addition of more subjects in the present study. Specifically, the data for the people scenes were submitted to a mixed ANOVA with instruction (Told, Not Told) as a between-subjects factor and session (pretest, test) and region (eyes, heads, bodies, foreground objects, background) as within-subjects factors. This analysis revealed a highly significant effect of Region ($F(4, 64) = 9.45$, $p < 0.0001$), reflecting that overall participants preferred to scan the eyes over any other region. An instruction × region interaction ($F(4, 64) = 3.18$, $p < 0.02$) indicated that while both groups fixated the eyes more than any other region, the eyes were fixated more frequently by the Told group than the Not Told group (Fishers LSD $p < 0.05$), and that the heads were fixated more frequently by the Not Told group than by the Told group (Fishers LSD $p < 0.05$). Fixation proportions for the other regions did not differ between the two groups. In short, participants are especially likely to look at the eyes when they are simply asked to encode and remember scenes with people. There were no other interactions, (all $ps > .05$), including the instruction × session × region interaction ($F(4, 64) = 2.14$, $p > 0.05$), indicating that the viewing patterns within each instruction group did not change from study to test sessions.

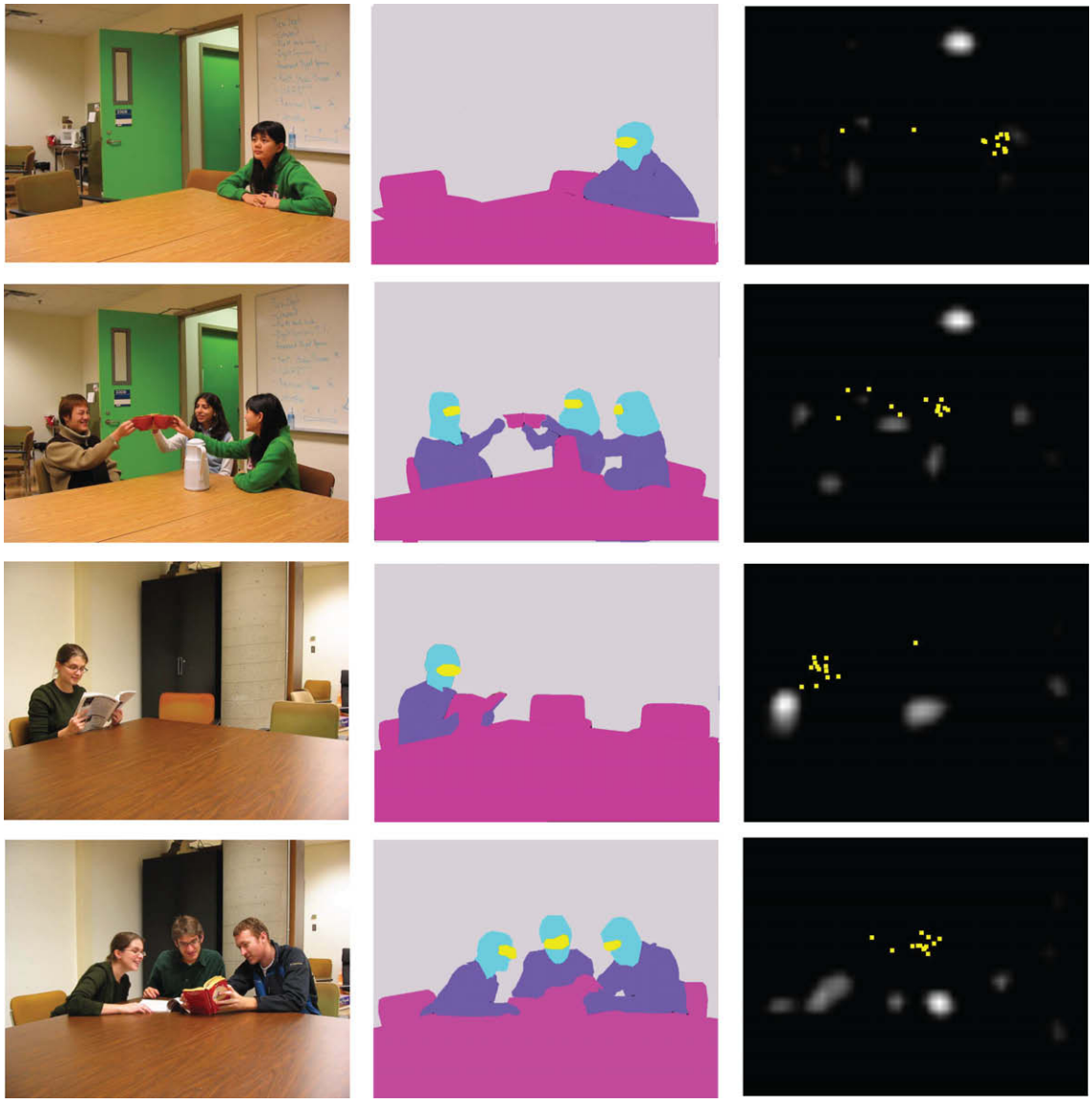

Fig. 1. Examples of the scenes used, the regions we defined (eyes, heads, bodies, foreground objects, and background) and their corresponding saliency maps (Itti & Koch, 2000) overlaid with the first fixations of Experiment 2.

3.3. First fixation

To determine where observers' initial saccades landed in the visual scene, we computed the proportion of first fixations that landed in a region (i.e., the first fixation after the experimenter-determined fixation at center). See Table 1 for the first fixation data broken down as a function of Experiment.

3.3.1. Experiment 1

A one-way repeated measures analysis of variance (ANOVA) revealed an effect of region ($F(4, 76) = 23.48$; $p < 0.0001$). Pair-wise comparisons (Tukey–Kramer, $p < 0.05$) revealed that the heads were highly likely to be fixated first (0.37), followed by background (0.24) and eyes (0.17), bodies (0.16), and foreground objects (0.07). Thus, the early interest in eyes was not obvious in the first fixation data of Experiment 1, with heads getting by far the most first fixations. The second fixation data, however, revealed an emerging interest in eyes. We found a significant effect of region ($F(4, 76) = 13.48$, $p < 0.00001$), this time reflecting that eyes and heads were both most likely to be fixated on the second fixation (eyes: 0.28; heads: 0.29) and more so than the other regions (background: 0.17; foreground objects: 0.15; body: 0.11), (Tukey–Kramer, $p < 0.05$).

3.3.2. Experiment 2

A task (Look, Describe, Social Attention) × region (eyes, head, body, foreground, background) mixed ANOVA revealed a main effect of region ($F(4, 144) = 24.08$, $p < 0.0001$). Pair-wise comparisons (Tukey–Kramer, $p < 0.05$) revealed that eyes (0.24), heads (0.31) and background (0.32) were equally likely to receive the first fixation, and more so than foreground objects (0.04) and bodies (0.08). There was no task × region interaction ($F < 1$), reflecting that the groups were no different in the placement of their first fixation.

3.3.3. Experiment 3

We computed the proportion of first fixations in the People scenes that landed in each region, as a function of instruction (Told, Not Told) and session (pretest, test). A mixed instruction by session by region ANOVA revealed a main effect of region ($F(4, 64) = 29.45$, $p < 0.0001$), reflecting that heads were most likely to be fixated first

Table 1
Proportion of first fixations landing in each region for each experiment.

Experiment	Task	Eyes	Heads	Bodies	Foreground objects	Background
1	Look	0.17	0.37	0.16	0.07	0.24
2	Look	0.21	0.27	0.14	0.05	0.33
	Describe	0.23	0.36	0.05	0.04	0.30
	Social Attention	0.27	0.30	0.06	0.03	0.33
3	Told (pretest)	0.38*	0.31	0.06	0.06	0.19
	Not Told (pretest)	0.20	0.53**	0.07	0.05	0.15
	Told (test)	0.32*	0.42	0.08	0.03	0.14
	Not Told (test)	0.23	0.47	0.09	0.03	0.17

* Significantly higher than the Not Told group at $p < 0.05$.
** Significantly higher than the Told group (pretest) at $p < 0.05$.

than any other region (head: 0.43; eyes: 0.29; background: 0.16; bodies: 0.08; foreground objects: 0.04) (Tukey Kramer, $p < 0.05$). Eyes were the next most likely to be fixated, significantly more so than background, foreground objects, and bodies. There was also a significant instruction × session × region interaction ($F(4, 64) = 3.07$, $p < 0.05$). As can be seen from Table 1, this higher order interaction indicated that the Told group was more likely to fixate the eyes first than the Not Told group, both in the pretest and test sessions (Tukey–Kramer, $p < 0.05$). In contrast, the Told group was *less* likely to fixate the heads first than the Not Told group, but only in the study session (Tukey–Kramer, $p < 0.05$).

3.4. Saliency analyses

Itti and Koch (2000) have developed an algorithm that enables the measurement of the visual saliency of an image by identifying strong changes in intensity, color and local orientation. To compute saliency, we used the Saliency Toolbox (Walther & Koch, 2005, 2006), and to remain as consistent as possible with others using this toolbox, we used all default parameters.[4] The final saliency maps were scaled to the same resolution as the fixation analysis (800 × 600 pixels) using bilinear interpolation. Examples of scenes, their regions, corresponding saliency maps are shown in Fig. 1. Saliency values were normalized to a range of 0 (absolutely non-salient) to 1 (highly salient).

Due to differences in the number of people (1 or 3) and variation in distance between the people and the camera, the eye region in the people scenes varied in area from 659 to 3684 pixels, with an average area of 1779 pixels. Even at the smallest area of 659 pixels, the eye region was large enough to be detected by the saliency model.

3.4.1. Basic performance of saliency model

The saliency at the location of first fixation was compared to the two chance-based estimates described earlier (*uniform-random* and *biased-random*). These data are presented in Table 2, broken down by experiment. Also, see Fig. 1 for fixations overlaid on the saliency maps of each example image. To determine whether the saliency model accounted for first fixation position above what would be expected by chance, non-parametric statistics (Mann–Whitney *U* tests) were performed to compare the medians of *fixated* saliency and *uniform-random* saliency as well as the medians of *fixated* saliency and *biased-random* saliency. Thus, in all subsequent analyses a *p* value less than 0.05 represents a significant difference between medians as indicated by the Mann–Whitney *U* test.

[4] The default normalization parameter in Saliency Toolbox is Iterative Normalization. It should be noted that we have tried multiple variations of the Saliency Toolbox, for example by changing the normalization type (e.g., LocalMax instead of Iterative Normalization). The pattern of results did not change substantially across these variations of the model.

3.4.1.1. Experiment 1. The fixated saliency was very low (0.011), as was uniform-random saliency (in fact, there were identical, 0.011, $p > 0.10$). Fixated saliency was also no different from biased-random saliency (0.017; $p > 0.10$). Thus, the results of Experiment 1 indicate that the saliency at fixated locations was no higher than would be expected by the random models.

3.4.1.2. Experiment 2. The fixated saliency was again very low (0.004), and no different from uniform-random saliency (0.011; $p > 0.10$). Fixated saliency was also no different from biased-random saliency (0.008; $p > 0.10$).

3.4.1.3. Experiment 3. People scenes: The fixated saliency (0.002) was actually significantly *lower* than uniform-random saliency (0.012), $p < 0.01$. Fixated saliency was also significantly lower than biased-random saliency (0.017), $p < 0.0001$. Thus, in Experiment 3 observers fixated regions that were *less* salient than would be expected by chance. The analysis of the basic performance of the saliency model revealed that both fixated and randomly generated saliency values were extremely low (close to zero) and did not differ statistically. Thus, the saliency model failed to predict the location of first fixation across three experiments.

3.4.2. Latency analysis

To determine whether saliency may have had more of an influence for early saccades, we binned saccade latencies into 50 ms intervals, from 0 to 550 ms. Fig. 2 depicts fixated saliency as a function of saccade latency bin. Separate ANOVAs for each experiment, with average fixated saliency as the dependent variable and bin (0–50 ms, 50–100 ms, ..., 500–550 ms) as the independent variable revealed that fixated saliency did not vary across latency bin (Experiment 1: $F < 1$, Experiment 2: $F(9, 359) = 1.19$, $p > 0.29$, Experiment 3: $F(9, 1023) = 1.03$, $p > 0.41$).

Furthermore, we were interested in whether saccade latency differed among the regions, to determine whether observers orient to social stimuli, such as eyes, rapidly or slowly (see Table 3 for a summary of these data). For each experiment we conducted an ANOVA on saccade latency as a function of region fixated (eyes, head, body, foreground objects, background). Saccade latencies to all regions were very fast, falling within 100–250 ms, and did not vary across regions, except that in Experiments 1 and 3 heads were fixated significantly faster than foreground objects (Experiment 1: $F(4, 711) = 3.30$, $p < 0.05$; Experiment 3: $F(4, 1028) = 4.27$, $p < 0.005$, no difference between regions for Experiment 2: $F(4, 364) = 1.95$, $p > 0.10$), confirmed by Tukey–Kramer post hoc comparisons ($p < 0.05$). As can be seen from Table 3, eyes and heads were consistently fixated very quickly.

3.4.3. Saliency of regions

Using saliency maps from Itti and Koch's (2000) model, we compared the saliency of eyes, heads, bodies, foreground objects,

Table 2
Median values for saliency of fixated regions, uniform-random saliency, and biased-random saliency, as a function of experiment. In the first two experiments, the saliency of fixated locations was no different from uniform-random saliency or biased-random saliency (*ps* > 0.10). In Experiment 3 the saliency of fixated locations was significantly *lower* than the random models. Thus, saliency failed to predict the location of the first fixation.

Experiment	Median saliency of fixated location	Uniform-random saliency	Biased-random saliency
1	0.011	0.011	0.017
2	0.004	0.011	0.008
3	0.002	0.012*	0.017*

* Indicates a significant difference with saliency at fixated locations at $p < 0.05$.

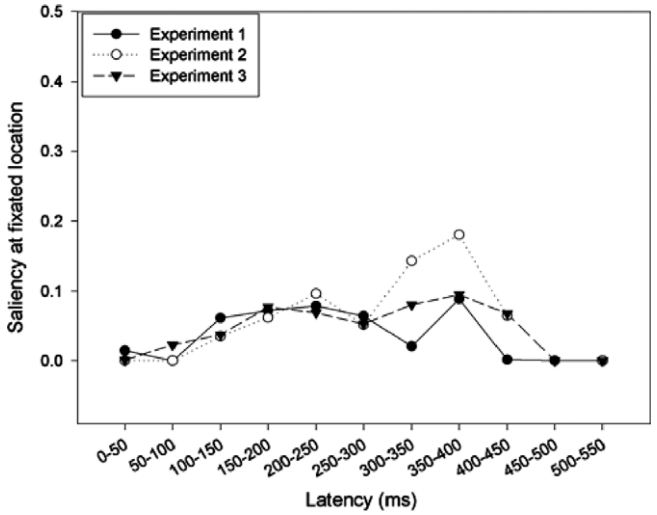

Fig. 2. Fixated saliency as a function of saccade latency bin (ms) and Experiment.

Table 3
Saccade latencies (ms) of first fixations landing in each region for each experiment.

Experiment	Eyes	Heads	Bodies	Foreground objects	Background
1	193	189	194	222*	207
2	178	178	164	203	187
3	184	176	178	203*	190*

* Different from heads at $p < 0.05$.

and background. This allowed us to determine whether the eyes or heads were salient, which might explain why they attracted initial fixations. Although this interpretation is unlikely given that saliency did not account for first fixation behavior, we wanted to be sure that eyes and heads were not more salient than the other regions.

We computed the average saliency of each region for each scene used in the experiments. We immediately noticed that median saliency of eyes (0.012) and heads (0.049) was very low, i.e., these regions were not salient. This was also true of the other regions (foreground objects: 0.011; bodies: 0.024; background: 0.005). Note that some regions were quite large (e.g., background), meaning that even if parts of that region were highly salient, computing saliency over all of the region's pixels would reduce its overall saliency value. However, this was not a problem for eyes or heads because they were relatively small. A Kruskal–Wallis One-Way ANOVA on Ranks revealed that at least two saliency values were different ($p < 0.001$). Pair-wise comparisons (Kruskal–Wallis Multiple-Comparison *Z*-Value Test, corrected for multiple comparisons) confirmed that the 'other' region was less salient than all regions except for foreground objects and eyes. Thus, eyes were among the least salient regions in the scenes.

4. Discussion

It has recently been established that observers often direct their initial fixations to the eyes and heads of other people that are depicted in complex social scenes (Birmingham et al., 2008a, 2008b; Cerf et al., 2008). One account is that people look at the eyes of others because they are a rich source of social information (e.g., Baron-Cohen et al., 1997; Birmingham et al., 2008b; Emery, 2000; Nummenmaa, 1964). An alternative account is that people look at the eyes of others because the eyes are visually salient (e.g., Kobayashi & Koshima, 1997). While some recent studies have sought to address the role that visual saliency plays in the viewing performance of natural social scenes, these studies have not provided a strong test of whether saliency can account for observers' interest in the eyes because they have: (1) not observed a robust and reliable fixation preference for the eyes of others (Fletcher-Watson et al., 2009), (2) not examined fixations to the eyes alone (Cerf et al., 2008), and (3) not examined the saliency model at initial fixations which is when visual saliency is expected to have its greatest impact on fixation behavior (Cerf et al., 2008; Fletcher-Watson et al., 2009). If people look at other people's eyes simply because eyes are visual salient, then this will disconfirm the hypothesis that people look at the eyes of others because they are a rich sources of social information. Alternatively, finding that visual saliency does not account for fixations to the eyes, even when initial fixation performance is analyzed, will disconfirm a saliency explanation and dovetail with the social importance explanation. The present study addressed this issue.

We started by determining, across three experiments, where observers placed their first fixation. We found a preference to fixate the heads that was consistent across experiments, and a significant bias to fixate the eyes that was present within the first fixation in two experiments and emerged by the second fixation in another study (Experiment 1). Next, we used three analyses to assess how well saliency accounts for these data, and found that saliency accounted for virtually nothing. Not only did the saliency model do no better at predicting first fixations than would be expected by chance, i.e., a random model; we also found that saliency at fixated locations was extremely low. In fact, in one experiment (Experiment 3) observers fixated regions that were actually *less* salient than would be expected by chance. In addition, the eyes and heads were generally non-salient (median saliency close to 0) and no more salient than any other region, except that heads were more salient than the 'other' region, which was the least salient. Finally, we discovered that saliency was no more effective at explaining fixation placement for early saccades than for late saccades. In contrast, saccades to the eyes were fast, suggesting a rapid detection of eyes from complex scenes. Thus, visual saliency cannot explain why observers direct their early fixations to the eyes (or heads) of people in the scene.

How can one be certain that when observers made their first fixation to the eyes, it was intentional? One might argue that our results could be accounted for by a face-selective mechanism (e.g., Viola & Jones, 2001) that simply orients early fixations to the geometric centre-of-faces, i.e., the nose or cheek region

depending on head orientation (Bindemann, Scheepers, & Burton, 2009). Fixations to the eyes may have simply occurred due to their proximity to the centre of the face, since the eye-tracker resolution may not have been sufficient to reliably distinguish between saccades directed to the eyes and saccades directed to the nose or cheeks (even though the tracker resolution was sufficient to distinguish fixations to the eyes versus the head, our two regions of interest). This issue, that eyes and centre-of-face are difficult to differentiate, is one that applies to all real-world social scenes in which the faces do not take up a large portion of the image. One of our reviewers even pointed out that with these constraints it might be impossible to determine whether saccades are "intended" to land on the face or the eyes in such scenes. This notion certainly invites future research, and it will be important for investigators to determine if saccades can be distinguished between eyes and centre-of-face given that the location of these target regions is typically confounded.

However, the possibility that the eyes *were* indeed detected and the target of early saccades is consistent with work by Lewis and Edmonds (2003). They showed that masking the eyes, but not other facial features, slows the detection of faces from a complex scene. Lewis and Edmonds suggested that the eyes play a special role in face processing, and that a fast face detection mechanism relies on locating the eyes within a face. This role of the eyes in face detection is also central to the Viola and Jones (2001) face detection algorithm, which relies on filters that detect the contrast between the eyes and cheeks and between the eyes and the bridge of the nose (Viola & Jones, 2001). Thus, the eyes themselves appear to play an important role in the rapid detection of faces within complex social scenes.

However, even if it is impossible to say with certainty that first saccades were intended for the eyes in our study, the finding still stands that if a saccade to the eyes/face cannot be explained by low-level saliency, then it provides evidence that social information overrides saliency in determining fixation placement in complex social scenes. This is an important finding that has implications for both the saliency models and for the social attention literature.

Our finding that visual saliency does a poor job of predicting first fixation performance in natural scenes can be added to a growing list of instances where the saliency model has been found wanting. There is mounting evidence that purely bottom-up saliency models are very poor at predicting human fixations when scene context/layout can be used to guide search (e.g., Torralba, Oliva, Castelhano, & Henderson, 2006) and the task is active, such as when walking down a hallway towards a target (Turano, Geruschat, & Baker, 2003) or searching for items within natural scenes (e.g., Foulsham & Underwood, 2007; Henderson, Brockmole, Castelhano, & Mack, 2007; Zelinsky, Zhang, Yu, Chen, & Samaras, 2006). Indeed, it appears that task instructions can completely override or even reverse the influence of saliency on fixation placement in complex scenes (Einhäuser, Rutishauser, & Koch, 2008; Foulsham & Underwood, 2008; Nyström & Holmqvist, 2008; Rothkopf, Ballard, & Hayhoe, 2007). In fact, the model appears to perform well above chance only when the task is unstructured and the scene is contrived in such a way that there are many highly salient regions to be fixated (e.g., the fractals of Parkhurst et al., 2002), or when the scene contains salient motion cues (Itti, 2005). In the instances where the saliency model fails to predict human attention, it must be argued that either the implementation of the saliency model needs modification, that top-down guidance can largely override bottom-up control of attention, or that saliency simply does not play a substantial role in guiding attention in active tasks involving rich (often real-world) stimuli.

The present study has demonstrated unequivocally that visual saliency does not account for fixations within real-world scenes containing people. Our alternative hypothesis is that observers look to the eyes of others because observers understand eyes to be important social stimuli, e.g., eyes communicate the attentional states and intentions of others. This interpretation is bolstered by our finding that the overall fixation preference for eyes (but not heads) is enhanced by the instruction to study social aspects of the scene, such as the attentional states of people in the scene (Experiment 2: Birmingham et al., 2008b). Note that these instructions never explicitly told people to look at the eyes of people in the scene, and so these results suggest that observers share an understanding that the eyes provide important information about the attentional states of others. Overall fixations to the eyes are also enhanced by the task to study the scenes for a later memory test, suggesting that observers perceive the eyes to be informative for remembering social scenes (Experiment 3; Birmingham et al., 2007). That people show a general bias to select eyes regardless of task is consistent with our understanding that humans are fundamentally social organisms (Adolphs, 2001; Brothers, 1990; Emery, 2000).

This default interest in the social information from other people's eyes may be a hallmark of normal social cognition, one that develops early in life. Indeed, even young infants are highly sensitive to the presence of other people's eyes, and begin to preferentially scan the eyes of a face by about 2 months of age (Haith, Bergman, & Moore, 1977; Maurer & Salapatek, 1976). Baron-Cohen (1994) has even suggested that this preference for the eyes is supported by a specific module, called the Eye Direction Detector (EDD), which both detects the presence of eyes and computes the direction of gaze. It is also thought that a failure to develop an interest in the social information from other people's eyes is an underlying factor in social disorders like autism (Baron-Cohen, 1994; Dawson, Meltzoff, Osterling, Rinaldi, & Brown, 1998; Klin et al., 2002). Thus, there is accruing evidence that, as part of normal social development, humans have a fundamental tendency to rapidly select and process the social information from eyes.

References

Adolphs, R. (2001). The neurobiology of social cognition. *Current Opinion in Neurobiology, 11*, 231–239.

Baron-Cohen, S. (1994). How to build a baby that can read minds: Cognitive mechanisms in mindreading. *Cahiers de Psychologie Cognitive, 13*, 513–552.

Baron-Cohen, S., Wheelright, S., & Joliffe, T. (1997). Is there a "language of the eyes" evidence from normal adults, and adults with autism or Asperger syndrome. *Visual Cognition, 4*(3), 311–331.

Bindemann, M., Scheepers, C., & Burton, A. M. (2009). Viewpoint and center of gravity affect eye movements to human faces. *Journal of Vision, 9*(2), 1–16.

Birmingham, E., Bischof, W. F., & Kingstone, A. (2007). Why do we look at eyes? *Journal of Eye Movement Research, 1*(1), 1–6.

Birmingham, E., Bischof, W. F., & Kingstone, A. (2008a). Social attention and real world scenes: The roles of action, competition, and social content. *Quarterly Journal of Experimental Psychology, 61*(7), 986–998.

Birmingham, E., Bischof, W. F., & Kingstone, A. (2008b). Gaze selection in complex social scenes. *Visual Cognition, 16*(2/3), 341–355.

Birmingham, E., Bischof, W. F., & Kingstone, A. (2009). Get real! Resolving the debate about equivalent social stimuli. *Visual Cognition, 17*(6/7), 904–924.

Brothers, L. (1990). The social brain: A project for integrating primate behavior and neurophysiology in a new domain. *Concepts Neurosci, 1*, 27–51.

Cerf, M., Harel, J., Einhäuser, W., & Koch, C. (2008). Predicting human gaze using low-level saliency combined with face detection. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.). *Advances in neural information processing systems* (Vol. 20). Cambridge, MA: MIT Press.

Dalton, K. M., Nacewicz, B. M., Johnstone, T., Schaefer, H. S., Gernsbacher, M. A., Goldsmith, H. H., et al. (2005). Gaze fixation and the neural circuitry of face processing in autism. *Nature Neuroscience, 8*(4), 519–526.

Dawson, G., Meltzoff, A. N., Osterling, J., Rinaldi, J., & Brown, E. (1998). Children with autism fail to orient to naturally occurring social stimuli. *Journal of Autism and Developmental Disorders, 28*, 479–485.

Einhäuser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision, 8*(2), 1–19.

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

Fletcher-Watson, S., Findlay, J. M., Leekam, S. R., & Benson, V. (2008). Rapid detection of person information in a naturalistic scene. *Perception, 37*, 571–583.

Fletcher-Watson, S., Leekam, S. R., Benson, V., Frank, M. C., & Findlay, J. M. (2009). Eye-movements reveal attention to social information in autism spectrum disorder. *Neuropsychologia, 47*, 248–257.

Foulsham, T., & Underwood, G. (2007). How does the purpose of inspection influence the potency of visual salience in scene perception? *Perception, 36*, 1123–1138.

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision, 8*(2), 1–17.

Haith, M. M., Bergman, T., & Moore, M. J. (1977). Eye contact and face scanning in early infancy. *Science, 198*, 853–855.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In R. van Gompel, M. Fisher, W. Murray, & R. Hill (Eds.), *Eye movement research: Insights into mind and brain*. Elsevier.

Henderson, J. M., Weeks, P. A., Jr., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during scene viewing. *Journal of Experimental Psychology: Human Perception and Performance, 25*(1), 210–228.

Henderson, J. M., Williams, C. C., & Falk, R. (2005). Eye movements are functional during face learning. *Memory & Cognition, 33*(1), 98–106.

Itier, R. J., Villate, C., & Ryan, J. D. (2007). Eyes always attract attention but gaze orienting is task-dependent: Evidence from eye movement monitoring. *Neuropsychologia, 45*, 1019–1028.

Itti, L. (2005). Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition, 12*(6), 1093–1123.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40*, 1489–1506.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*, 1254–1259.

Klin, A., Jones, W., Schultz, R., Volkmar, F., & 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.

Kobayashi, H., & Koshima, S. (1997). Unique morphology of the human eye. *Nature, 387*, 767–768.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, 4*, 219–227.

Lewis, M. B., & Edmonds, A. J. (2003). Face detection: Mapping human performance. *Perception, 32*, 903–920.

Maurer, D., & Salapatek, P. (1976). Developmental changes in the scanning of faces by young infants. *Child Development, 47*, 523–527.

Nummenmaa, T. (1964). *The language of the face (Jyvaskyla studies in education), psychology, and social research*. Finland: Jyvaskyla.

Nyström, M., & Homlqvist, K. (2008). Semantic override of low-level features in image viewing – Both initially and overall. *Journal of Eye Movement Research, 2*(2), 1–11.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42*, 107–123.

Pelphrey, K. A., Sasson, N. J., Reznick, S., Paul, G., Goldman, B. D., & Piven, J. (2002). Visual scanning of faces in autism. *Journal of Autism and Developmental Disorders, 32*(4), 249–261.

Rothkopf, C. A., Ballard, D. H., & Hayhoe, M. M. (2007). Task and context determine where you look. *Journal of Vision, 7*(14), 1–20.

Smilek, D., Birmingham, E., Cameron, D., Bischof, W. F., & Kingstone, A. (2006). Cognitive ethology and exploring attention in real world scenes. *Brain Research, 1080*, 101–119.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motorbiases and image feature distributions. *Journal of Vision*, *7*(14), 4, 1-17.

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research, 45*, 643–659.

Torralba, A. (2003). Modeling global scene factors in attention. *Journal of the Optical Society of America, 20*(7), 1407–1418.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 113*(4), 766–786.

Treisman, A. M., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology, 12*, 97–136.

Turano, K. A., Geruschat, D. R., & Baker, F. H. (2003). Oculomotor strategies for the direction of gaze tested with a real-world activity. *Vision Research, 43*, 333–346.

van der Geest, J. N., Kemner, C., Verbaten, M. N., & van Engeland, H. (2002). Gaze behavior of children with pervasive developmental disorder toward human faces: a fixation time study. *Journal of child psychology and psychiatry, and allied disciplines, 43*(5), 669–678.

van Zoest, W., Donk, M., & Theeuwes, J. (2004). The role of stimulus-driven and goal-driven control in saccadic visual selection. *Journal of Experimental Psychology: Human Perception and Performance, 30*(4), 746–759.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*.

Walker-Smith, G., Gale, A. G., & Findlay, J. M. (1977). Eye movement strategies involved in face perception. *Perception, 6*(3), 313–326.

Walther, D., & Koch, C. (2005). Saliency Toolbox (Version 2.0) (Computer Software). <http://www.saliencytoolbox.net> Retrieved 2.12.05.

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks, 19*, 1395–1407.

Yarbus, A. L. (1967). *Eye movements and vision (B. Haigh, Trans.)*. New York: Plenum Press (Original work published 1965).

Zelinsky, G., Zhang, W., Yu, B., Chen, X., & Samaras, D. (2006). The role of top-down and bottom-up processes in guiding eye movements during visual search. In Y. Weiss, B. Scholkopf, & J. Platt (Eds.). *Advances in neural information processing systems* (Vol. 18, pp. 1569–1576). Cambridge, MA: MIT Press.