Complementary effects of gaze direction and early saliency in guiding fixations during free-viewing

Ali Borji[a], Daniel Parks[b], Laurent Itti[a,b,c]

[a]*Department of Computer Science, University of Southern California, 3641 Watt Way, Los Angeles, CA 90089*
[b]*Neuroscience Graduate Program, University of Southern California, 3641 Watt Way, Los Angeles, CA 90089*
[c]*Department of Psychology, University of Southern California, 3641 Watt Way, Los Angeles, CA 90089*

Abstract

Gaze direction provides an important and ubiquitous communication channel in daily behavior and social interaction of humans and some animals. While several studies have addressed gaze direction in synthesized simple scenes, few have examined how it can bias observer attention and how it might interact with early saliency during free viewing of natural scenes. Experiment 1 used a controlled, staged setting in which an actor was asked to look at two different objects in turn, yielding two images that only differed by the actor's gaze direction, to causally assess the effects of actor gaze direction. Over all scenes, the median probability of following an actor's gaze direction was higher than the median probability of looking towards the single most salient location (0.22 vs. 0.10; sign test, $p = 3.223e-06$), and higher than chance (both uniform, 0.02; $p = 6.750e-17$, and Naive Bayes, 0.06; $p = 6.171e-10$). Experiment 2 confirmed these findings over a larger set of unconstrained scenes collected from the web and containing people looking at objects and/or other people. To further compare the strength of saliency vs. gaze direction cues, we computed gaze maps by drawing a cone in the direction of gaze of the actors present in the images. Gaze maps predicted observers' fixation locations significantly above chance, although below saliency (AUC; gaze map vs. saliency map 0.612 vs. 0.797 in exp 1 and 0.625 vs. 0.789 in exp 2). Finally, to gauge the relative importance of actor face and eye directions in guiding observer's fixations, in experiment 3, observers were asked to guess the gaze direction from only an actor's face region (with the rest of the scene masked), in two conditions: actor eyes visible or masked. Median probability of guessing the true gaze direction within $\pm 9^\circ$ was significantly higher when eyes were visible (0.2 vs. 0.13; sign test, $p = 5.76e-15$), suggesting that the eyes contribute significantly to gaze estimation, in addition to face region. Our results highlight that gaze direction is a strong attentional cue in guiding eye movements, complementing low-level saliency cues, and derived from both face and eyes of actors in the scene. Thus gaze direction should be considered in constructing more predictive visual attention models in the future.

Keywords: visual attention, eye movements, gaze direction, gaze following, line of sight, top-down attention, bottom-up saliency, saliency modeling, fixation prediction, free viewing, gaze cueing, overt attention, human-computer interaction

1. Introduction

A tremendous amount of research has been undertaken to discover and understand cues that influence eyes movements in daily life attentional behavior for tasks such as reading, scene perception, and visual search (See for example: Posner (1980); Land & Hayhoe (2001); Henderson (2003); Schütz et al. (2011); Carrasco (2011); Tatler et al. (2011); Borji & Itti (2012b)). Two categories of cues have been identified: *Bottom-up (BU) cues* such as discontinuities in contrast, color, intensity, brightness, motion, and spatial frequency in the visual stimulus (a.k.a., saliency; Treisman

Email addresses: borji@usc.edu (Ali Borji), danielfp@usc.edu (Daniel Parks), itti@pollux.usc.edu (Laurent Itti)

& Gelade (1980); Koch & Ullman (1985); Milanese et al. (1994); Mannan et al. (1996); Itti et al. (1998); Reinagel & Zador (1999); Krieger et al. (2000); Itti & Koch (2001); Parkhurst et al. (2002); Peters et al. (2005); Borji & Itti (2012a); Borji (2012)) and *Top-down (TD) cues* mediated by factors such as task demands (Yarbus, 1967; Land & Lee, 1994; Ballard et al., 1995; Land & Hayhoe, 2001; Triesch et al., 2003; Einhäuser et al., 2008; Borji et al., 2014; Borji & Itti, 2014), context effects and scene Gist (Torralba et al., 2006), expertise with similar scenes (Underwood et al., 2009), object appearance and spatial priors in visual search (Wolfe, 1998; Wolfe & Horowitz, 2004; Ehinger et al., 2009; Kanan et al., 2009), tendency of observers to look near the center of the displays also know as image center-bias (Tatler, 2007), tendency of observers to look near the center of objects also know as object center-bias (Nuthmann & Henderson, 2010), memory (Droll et al., 2005; Carmi & Itti, 2006), emotion (Ramanathan et al., 2014), gender (Shen & Itti, 2012), and culture (Chua et al., 2005). One additional attentional cue is gaze direction which so far has been overlooked since it is hard to quantitatively measure it in unconstrained natural scenes. A number of electrophysiology and imaging studies (Hoffman & Haxby, 2000; Pelphrey et al., 2004) have shown that several brain areas (e.g., superior temporal sulcus) are highly sensitive to gaze direction. Further, gaze direction has been shown to be important in social interaction and in disease states such as autism (Klin et al., 2009; Fletcher-Watson et al., 2009). Thus, we believe it is important to assess what role this cue plays in predicting eye fixations in natural scenes. Finally, gaze behavior is a function of both BU and TD components. For example in driving depending on the task temporal task demand magnitude of attention may change from freely wandering around when driving in a straight less crowded road to highly focused attention when negotiating a curvy road which is a dangerous.

Gaze direction (Bock et al., 2008) and pointing (Louwerse & Bangerter, 2005) are two well-known cues thought to guide behaviors in human and non-human species. Gaze following is a rapid and effortless socio-cognitive behavior that serves several functions for purposes such as social interactions, social learning, collaboration, coordination, threat assessment, understanding the intentions of others, signaling what is important in the shared visual field to direct the attention of others, and communicating thoughts, judgments, emotions, desires and needs (Emery, 2000). Some studies have postulated that gaze cueing might be influenced by facial cues signaling status, gender, dominance, leadership, familiarity and group membership. For example, monkeys selectively attend more to gaze direction of individuals rated as high in social status (Shepherd et al., 2006). Gaze following is also an important aspect of joint attention (when two or more people are paying attention to the same entity) and allows coupling visual appearances to verbal descriptions (Bakeman & Adamson, 1984; Kobayashi & Kohshima, 1997).

In laboratory experiments, perceived gaze has been shown to facilitate covert attention in the direction of gaze, even when the gaze cue is not informative of the task (Friesen & Kingstone, 1998; Driver et al., 1999). Perceived gaze has also been shown to facilitate overt attention if the gaze is congruent with the task direction, and hamper it if the gaze is not congruent (Ricciardelli et al., 2002). In a set of experiments, the subject had to saccade to either the left or the right target based on a color cue at a central fixation. Placing a face with a gaze directly out of the screen, or congruent with the color direction had about the same error. However, the incongruent direction had higher error, which did not happen when the faces were replaced with arrows. These studies suggest that these attentional cues happen automatically, in spite of top-down knowledge of their lack of usefulness, and beyond the explicit control of the subject. Thus, it seems that to have a complete model of bottom up attention, simple features (orientations, color, intensity) along with face detection is not enough (Cerf et al., 2009), as gaze cueing is also required.

Gaze following helps us understand and interpret intentions and mental states of others. Humans have the ability to model other people's current knowledge of the world as well as their goals, which is frequently referred to as *Theory of Mind* (Premack & Woodruff, 1978). An observer has to infer body pose, head pose and eye postures of other people and from this information extract the effective gaze direction (line of sight) which then in turn allows him to perceive the world from the perspective of others. Autistics have deficits in using gaze to facilitate Theory of Mind (Baron-Cohen et al., 1995). For example, autistics were able to assess whether a cartoon was looking at them as accurately as

normal children. This indicates they are able to see and process eye gaze to a certain extent. However, they failed to use that knowledge to infer internal states about the viewed cartoon. In another experiment four candies were placed around a cartoon face, which gazed at one of the candies. Children were asked which candy the cartoon face wanted. Autistic children were unable to infer that the cartoon face desired the candy that it was gazing towards (See Figure 3, page 386 in Baron-Cohen et al. (1995)). Instead, when asked to pick the candy that the cartoon face likes, they consistently chose the candy that they themselves wanted. This seems to imply that they ignore cues from the other person, and instead model that person with the same goal states as they themselves have.

From a developmental perspective, it has been shown that non-verbal communication cues such as gaze imitation play a central role in social and cognitive development, language acquisition and cognitive learning during early infancy (Hoffman et al., 2006; Okumura et al., 2013; Baldwin, 1995; Tomasello, 2009). This is also known as the *cooperative eye hypothesis* (Kobayashi & Kohshima, 1997). Infants are capable of following the gaze direction of humans (Flom et al., 2007; Gredebäck et al., 2008; Theuring et al., 2007) and non-human agents or robots (Csibra & Gergely, 2006; Meltzoff et al., 2010; Okumura et al., 2013). In humans, gaze direction estimation has been shown to be at least crudely present (capable of discriminating left, right, or straight) as early as 3 months of age (D'Entremont et al., 1997). Infants as young as 6-months are capable of following the direction of gaze. At 12 months, an infant can accurately follow a perceived gaze across an intervening stimulus and to the correct target. Prior to this age, objects that occur between the initial infant fixation and the true target seem to either override the search task, or are presumed to be the target (Butterworth & Jarrett, 1991). One year old infants are able to follow the gaze direction of an adult/caregiver and determine which object he is currently focused on (Butterworth, 1991). It is not until 14 months of age, however, that both eye position and head pose are reliably taken into account when estimating gaze (Caron et al., 2002). By 18 months, the infant can even estimate perceived gaze that is outside their field of view (Butterworth & Jarrett, 1991; Moll & Tomasello, 2004). But it is at the age of 3 years when toddlers are able to explicitly discriminate the gaze direction (Doherty et al., 2009).

Previous literature has investigated the role of subcomponents of gaze following (body, head, face, and eye) in directing attention in natural scenes. It is known that people in scenes capture eye fixations (potency of people (Humphrey & Underwood, 2010; Zwickel & Võ, 2010)). Within the human body, faces provide a wealth of information regarding a person's identity, gender, direction of attention (Cerf et al., 2009), needs, desires, interpersonal relations, and emotion as well as internal mental states (e.g., Baron-Cohen et al. (1997)). Observers also derive complex trait judgments, such as competence (e.g. Ballew & Todorov (2007)) and trustworthiness (Winston et al., 2002; Todorov et al., 2008) from facial appearance, and such judgments are made rapidly and accurately and can help guide decision making during social interactions (c.f. Skarratt et al. (2012)). A large number of studies have shown that observers preferentially attend to faces in free viewing of scenes (e.g., Vuilleumier (2000); Ro et al. (2001); Bindemann et al. (2005); Cerf et al. (2009)). Within faces, it is believed that the eyes convey the most information, and it has been shown that people have a strong preferential bias to attend to the eyes of other people (Yarbus, 1967; Baron-Cohen et al., 1997; Birmingham et al., 2008; Pelphrey et al., 2002; Henderson et al., 2005; Foulsham et al., 2010). The importance of facial features and gaze following is further witnessed by the finding that some dedicated brain regions exist for representing faces (e.g., Kanwisher et al. (1997); Kanwisher & Yovel (2006); Tsao et al. (2006); Moeller et al. (2008)) and gaze directions (e.g., superior temporal sulcus (Hoffman & Haxby, 2000; Pelphrey et al., 2004)).

While influence of gaze direction in perception and the judgment of gaze direction has been given extensive treatment using simple synthetic scenes (e.g., von Griinau & Anston (1995); Driver et al. (1999); Friesen & Kingstone (1998, 2003); Mansfield et al. (2003); Jones et al. (2010); some under the name of gaze cueing by employing standard gaze cueing paradigm of Posner (1980)), less work has inspected the influence of gaze direction in the viewing of natural scenes. Birmingham et al. (2009) showed that saliency does not account for fixations to eyes within social scenes. Instead, it appears that observers' fixations are driven largely by their default interest in social information.

Some studies predict that observers are more likely to attend towards regions (and for longer times) that are being looked at by the central figure in the display (e.g., Fletcher-Watson et al. (2008)). Some other studies have investigated gaze direction in the context of magic tricks (Macknik et al., 2008; Kuhn & Tatler, 2005) where a technique called 'misdirection' is frequently used to distract the attention of the observer (often through gazing at something irrelevant). While the effect of gaze direction is well established, less work has addressed its interaction with other factors such as properties of the gazed-at objects in the scene (e.g., saliency (Itti et al., 1998) or importance (Spain & Perona, 2010)). In other words, gaze following behavior not only does depend on the gaze direction of an actor but it also depends on the entity being looked at. Sometimes, in interpreting a scene, gaze direction can override or modulate bottom-up saliency allowing attention to be directed to less salient objects or locations.

To the best of our knowledge, Castelhano et al. (2007) were the first to study the effect of gaze direction of actors in natural scenes in guiding the eye movements of observers of those scenes. They conducted a study in which subjects viewed a sequence of scenes presented as a slide show that portrayed the story of a janitor (the actor) cleaning an office. They found that: a) the actor's face was highly likely to be fixated (as also later suggested and modeled by Cerf et al. (2009) in 2007) and b) the observer's next saccade was more likely to be toward the object that was the focus of the actor's gaze than in any other direction. Castelhano *et al.*'s study is interesting as it provides the seed to look deeper in the role of gaze direction in free-viewing and fixation prediction. Eventually, gaze direction of actors in a scene could provide an additional source of information for visual attention models, similar to the manner in which human faces and written text have recently been added to saliency models (Cerf et al., 2009; Judd et al., 2009; Wang & Pomplun, 2012). But before gaze direction can be included in saliency models, several central questions need to be answered which have not yet been fully addressed by previous studies. For example, if saliency can already explain all saccades that leave the head in the direction of the gaze, then there is no need to model gaze direction explicitly. These questions are:

1. Does gaze direction causally guide observer attention and eye movements?
2. How might gaze direction cues complement low-level, feature-based saliency cues in predicting fixation locations?
3. Are gaze direction cues modulated by the presence of other faces in the scene, as such presence could be a confounding factor in attracting attention?
4. Does gaze following strength change over viewing time?
5. What is the contribution of eyes beyond what the face already offers in estimation of gaze direction?

In the first two experiments, we test whether viewers, in interpreting real-world scenes, prioritize looking at less salient objects or locations if they happen to be in the direction of the gaze. We also investigate the power of a simple gaze map in predicting fixations. In the third experiment, we quantify the informativeness of head and eye regions in determining gaze direction. Further, we discuss how models can be built that effectively incorporate gaze direction and what are the important problems to address.

2. Experiment one

Here we attempt to test the hypothesis that free-viewing observers follow the gaze direction of people in the scene above chance on a set of controlled stimuli. Since gaze following might be due to saliency of the object at the gaze endpoint, we also account for this confounding factor. In what follows, we will refer to "actors" as persons depicted in images, whose gaze direction might influence the attention and eye movements of external "observers" inspecting those images. Throughout the paper we concentrate on observer saccades that start from actor head regions (unless stated otherwise).

2.1. Methods

2.1.1. Stimuli

We collected two sets of controlled stimuli. Several actors were instructed to look, in turn, at one of two objects in a scene. Fig. 1 shows ten example image pairs along with blurred[1] saccade heat maps (made only from fixations leaving the actor's head area) and saliency maps. We collected 30 pairs of images (60 images in total). See Appendix A for complete set of pairs used in this experiment. The camera position remained fixed across the two shots from the same scene. We then collected another 60 images at random from the web to be used as fillers so that observers would be less likely to realize that the study was about gaze direction following. There were no gaze direction in some of these images while in some others people looked at something or faced the camera. Filler images had much more variety (multiple objects, different locations, etc) than our controlled set. We then created two sets of stimuli by putting together one image from each image pair (30 images) and 30 images chosen from the random set. Thus each set had 60 images. No observer viewed both images in a pair. Images were resized to 1920×1080 pixels, by adding gray margins, while preserving the aspect ratio (See Fig. 1).

2.1.2. Observers

Two groups of observers, comprising 15 subjects each, participated in this experiment. Observers in group 1 (6 male, 9 female; mean age = 19.73, SD = 1.03) viewed images in Set 1. Observers in group 2 (3 male, 12 female; avg age = 19.8, SD = 1.26) viewed images in Set 2. Observers were undergraduate students at the University of Southern California (USC) from these majors: Neuroscience, Psychology, Biology, Sociology, Business, Communication, Music, Biomedical Engineering, Religion, etc. The experimental methods were approved by the USC's Institutional Review Board (IRB). Observers had normal or corrected-to-normal vision and received course credit for participation. They were naïve to the purpose of the experiment and had not previously seen the stimuli. They were instructed to simply watch and enjoy the pictures (free viewing).

2.1.3. Apparatus and Procedure

Observers sat 106 cm away from a 42 inch LCD monitor screen so that scenes subtended approximately $45.5^\circ \times 31^\circ$ visual angle. A chin rest was used to stabilize head movements. Stimuli were presented at 60Hz at a resolution of 1920×1080 pixels. Eye movements were recorded via a non-invasive infrared Eyelink (SR Research, Osgoode, ON, Canada) eye-tracking device at a sample rate of 1000 Hz (spatial resolution less than 0.5°). Each image was shown for 30 seconds followed by a 5 seconds delay (gray screen). The eye tracker was calibrated using 5 points calibration at the beginning of each recording session. Observers viewed images in random order. Saccades were classified as events where eye velocity was greater than 35°/s and eye acceleration exceeded $9500^\circ/s^2$ as recommended by the manufacturer for the Eyelink-1000 device. Faces took up, on average 0.99% of the image, while they accounted for 9.22% of all fixations. The controlled pairs of objects took up 1.12% of the image each, but the gazed at object received 7.75% of all fixations, while the ignored object received only 5.20% of all fixations. The average head size was 143×180 pixels ($\sim 3.6^\circ \times 5.2^\circ$ visual angle). The average object size was 167×174 pixels ($\sim 4.2^\circ \times 5.1^\circ$ visual angle). The average ground truth[2] gaze length was 724 pixels ($\sim 18^\circ$ visual angle).

For saliency computation, we used the Adaptive Whitening Saliency model (AWS) by Garcia-Diaz et al. (2012) which has been shown to outperform other saliency models in recent benchmarks[3] in predicting eye movements of observers in free-viewing of natural scenes (Borji et al., 2012a).

[1] Code for map blurring using Matlab®: *blurredSacMap= conv2(sacMap,fspecial('gaussian',100,20))*.

[2] Ground truth gaze direction is the direction that an actor is looking at. It is known in experiment one by construction but we estimate it for the images of the second experiment.

[3] Note that as we have argued in our other works (Borji et al., 2013a,b), selection of a saliency model is very important and in some cases may alter conclusions of a study. For this reason, here we choose the best performing model in the saliency literature (i.e., the AWS model).

Figure 1: Sample image pairs in experiment one along with their corresponding heat maps and saliency maps from the AWS model (Garcia-Diaz et al., 2012). Heat maps are constructed from those fixations of all observers that start from the head region and end somewhere in the image. In each image, an actor was explicitly instructed to look at one of the two objects. Head and two object regions as well as gaze directions were annotated. The looked-at object is marked with the red polygon (ignored=green). Blue dots represent fixations.

2.2. Analysis and Results

We annotated (manually traced) head regions[4] and the two looked-at objects in all image pairs (object boundaries). Note that the ground truth gaze direction is known here since actors in images were explicitly instructed to look at one of the two objects.

2.2.1. Actor gaze direction influences observer saccade endpoints irrespective of saliency

To evaluate the effect of gaze direction on free viewing behavior, we quantify the extent to which fixations are drawn to the gazed-at object vs. the ignored object. We have two objects in each image: gazed-at and ignored. We compute the fraction of all saccades that start from the head and end inside each of these two objects (overall 3,331 such saccades). This gives us two fractions, one for the gazed-at and one for the ignored object. For all 60 images, we thus obtain two arrays/vectors of size 60 of these fractions (one for the gazed-at and one for the ignored object). We then compare the median of these two vectors. The median of the gazed-at vector is 0.069 which is significantly higher than the median 0.022 of the ignored vector (sign test; $p = 6.132e - 9$). Medians of the normalized saliency map activation for both cases are not significantly different from each other (0.019 vs. 0.020; $p = 0.901$). For 78.3% of cases, the gazed-at object attracted a higher fraction of fixations compared to the ignored object. These results clearly show that low-level saliency does not account for the influence of gaze direction on fixations.

2.2.2. Actor gaze direction predicts observer saccade direction

Beyond analyzing observer saccade endpoints, we here also consider more broadly observer saccade directions, as this yields an analysis that can be extended to Experiment 2 that uses an uncontrolled dataset where it is unknown exactly which objects actors may be looking at.

For each image, we measured probability distributions of angular directions for all saccades initiating somewhere from the head region to some part of the image but not on the same head. We used all fixations starting from the face, whether they end up inside one of the two objects or not, because we were interested in the gaze following effect. We first measured the histogram of angular saccade directions in 20 bins, of 18 degrees each[5], and then converted it to a probability density function (pdf) by dividing this histogram to its sum. A higher pdf value at the ground truth gaze direction hence means stronger gaze following irrespective of the exact endpoint. Fig. 2 shows 12 image pairs along with their distributions of saccade directions for data of all observers. As this figure shows, there is a peak in the direction of the looked-at object in the majority of images.

Here, we show that the value of the saccade distribution (pdf) in the direction of the ground truth gaze is significantly higher than in a random direction chosen uniformly (i.e., chance level). For 60 images, we have a vector of size 60 of pdf values. Similarly, we have another vector of the same size with pdf values read out from random directions. Shown in Fig 3.A, the median of the saccade direction vector is 0.220, which is significantly higher than the median of the saccade vector at uniform random directions 0.023 using sign test[6] (i.e., versus uniform chance level; $p = 6.750e - 17$). The median of the saccade direction vector is also significantly higher than a smart chance level in which directions are sampled randomly from the average ground truth gaze direction pdf shown in Fig 3.B (i.e., Naive Bayes chance level of 0.061; sign test $p = 6.171e - 10$). Hence, observers tended to look significantly more (overall) in the direction of actor gaze than in any other direction.

In another analysis, we assessed the relative strength of the saccade direction pdf values when the actor looked at an object and when he did not (i.e., looked at the other object). Let P_+ be the pdf value in the direction of an object when it was looked at by the actor, and P_- when it was not (because the actor looked at the other object). We define

[4] We do not distinguish between faces that are in-profile or frontal facing.

[5] We chose these parameters to: 1) have enough saccade data in each bin, and 2) be reasonably precise in labeling each gaze direction.

[6] Using the Matlab®ranksum function.

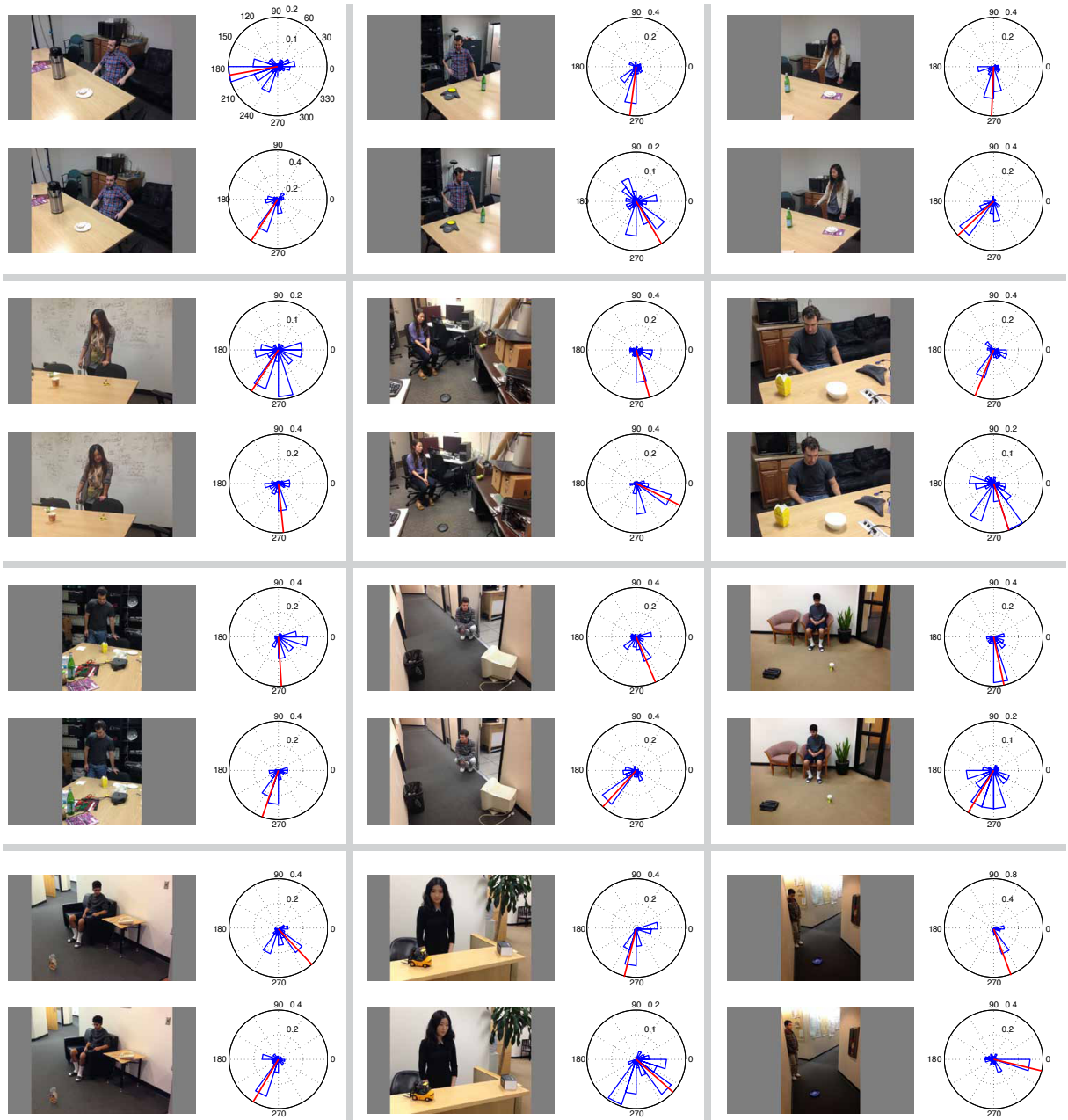

Figure 2: Sample images in the first experiment. In each panel, the two images of the image pair are shown in the left with their corresponding saccade probability distributions (for saccades starting somewhere in the annotated head; See Fig. 1) shown in the right (polar plot). Red lines in polar plots indicate the ground truth gaze direction. Gaze following is strong for some images (e.g., the person looking at the CRT monitor and the trash can) while it is weaker for some others (e.g., the person looking at the yellow food box and the tissue perhaps due to the complexity of the background and saliency of one of the objects).

the relative gaze strength denoted by z to be $P_+/(P_+ + P_-)$. A z value of 1 means that observers always looked at the object that the actor was looking at, i.e., observers always followed the actor's gaze direction, and a z value of zero means they never followed the actor's gaze direction. Averaged over all image pairs we reached the value of $\mu_z = 0.647$ ($\sigma_z = 0.142$) for z which is higher than 50% chance level (t-test[7], $p = 5.03e-11$). This implicitly means that if we were to guess the ground truth gaze direction on each image based on the relative saccade direction pdf value (i.e., decision criterion being the direction of the gazed at object with higher pdf value) then we would have reached the accuracy of 65%. This result shows that actor gaze direction has a significant causal effect onto the directions of observer eye movements.

2.2.3. Gaze direction vs. most salient location direction

To complement the analysis of saliency at saccade endpoints in section 2.2.1, we here also evaluate saliency in particular gaze directions.

One might argue that it was saliency that attracted observers to look in a particular direction and not gaze following (i.e., the effect could be partly due to low-level saliency). To account for this confounding factor, we measured the saccade pdf value in the direction of maximum saliency (not necessarily inside annotated objects) and compared it with the saccade pdf value in the ground truth gaze direction. Median saccade pdf value at maximum saliency direction is 0.10 which is significantly lower than the median pdf value 0.22 at the ground truth gaze direction (sign test, $p = 3.223e-06$). Saccade pdf value at the direction of maximum salient location is significantly higher that the uniform chance level ($p = 6.022e-09$) and Naive Bayes Chance ($p = 1.788e-02$). Thus, while saliency is an important factor in predicting observer gaze direction, it can not fully explain our data. This means that gaze direction and saliency are two complementary sources of information in guiding eye movements.

[7] We empirically verified via Kolmogorov–Smirnov test that variable z is normally distributed.

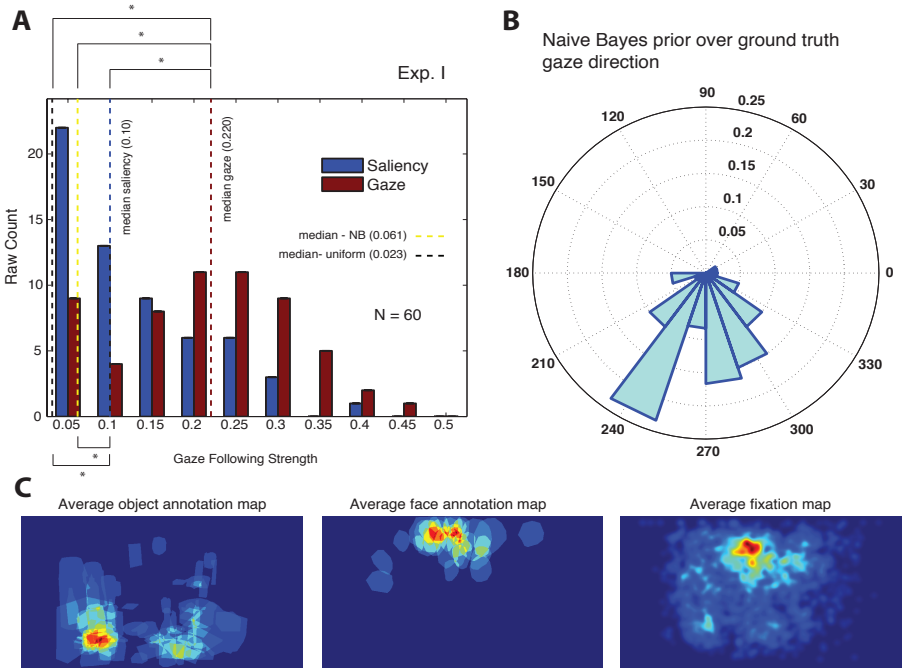

Figure 3: Results of the first experiment. A) the histogram of saccade direction pdf values (gaze following strength) in the ground truth gaze direction and the most salient location direction (saliency maps are normalized to be pdfs). B) the distribution of ground truth gaze directions over all 60 images in experiment one (i.e., prior gaze direction distribution). Prior distribution is used to compute the Naive Bayes chance level. C) average annotation map (average of all object polygons over all images) for faces and objects, as well the mean eye position map over all images for saccades that start from somewhere inside the head region. As it can be seen, there is a high fixation density around faces even for saccades that leave the head region. This can be partly due to uncertainty of observers in landing saccades or eye tracker error.

3. Experiment two

Our aim in this experiment was to explore the generality of gaze following behavior on a wider range of uncontrolled natural scenes including scenes with more cluttered backgrounds and with multiple people interacting with each other as well as multiple other objects. Having several persons in a scene raises an additional challenge that gaze following behavior might not be due to a person looking towards someone, but it might be because the looked-at person is salient and captures attention by itself. Indeed, recent evidence suggests that human faces strongly attract eye movements, often capturing the first fixation on a new scene (Cerf et al., 2009; Judd et al., 2009; Borji, 2012). Here we investigate this challenge by breaking down the analysis into cases where the entity at the gaze endpoint is another person or an object.

3.1. Methods

3.1.1. Stimuli

Stimuli consisted of a set of 200 color photographs collected mostly from the Flickr dataset[8]. Photographs span a variety of topics and locations such as indoor, outdoor, social interactions, object manipulations, instrument playing, athletics, reading, and dancing. As in experiment one, images were resized to 1920 by 1080 pixels (See Figs. 4 and 5 for sample images) while keeping the aspect ratio. We chose images in which at least one of the faces were large enough, visible, and gazing at something visible in the scene (another person or an object). In several of the images, some people are looking into the camera or out of the scene, but these heads were discarded in the gaze following

[8]http://www.flickr.com/. Some images were also borrowed from the AFW dataset (Zhu & Ramanan, 2012)

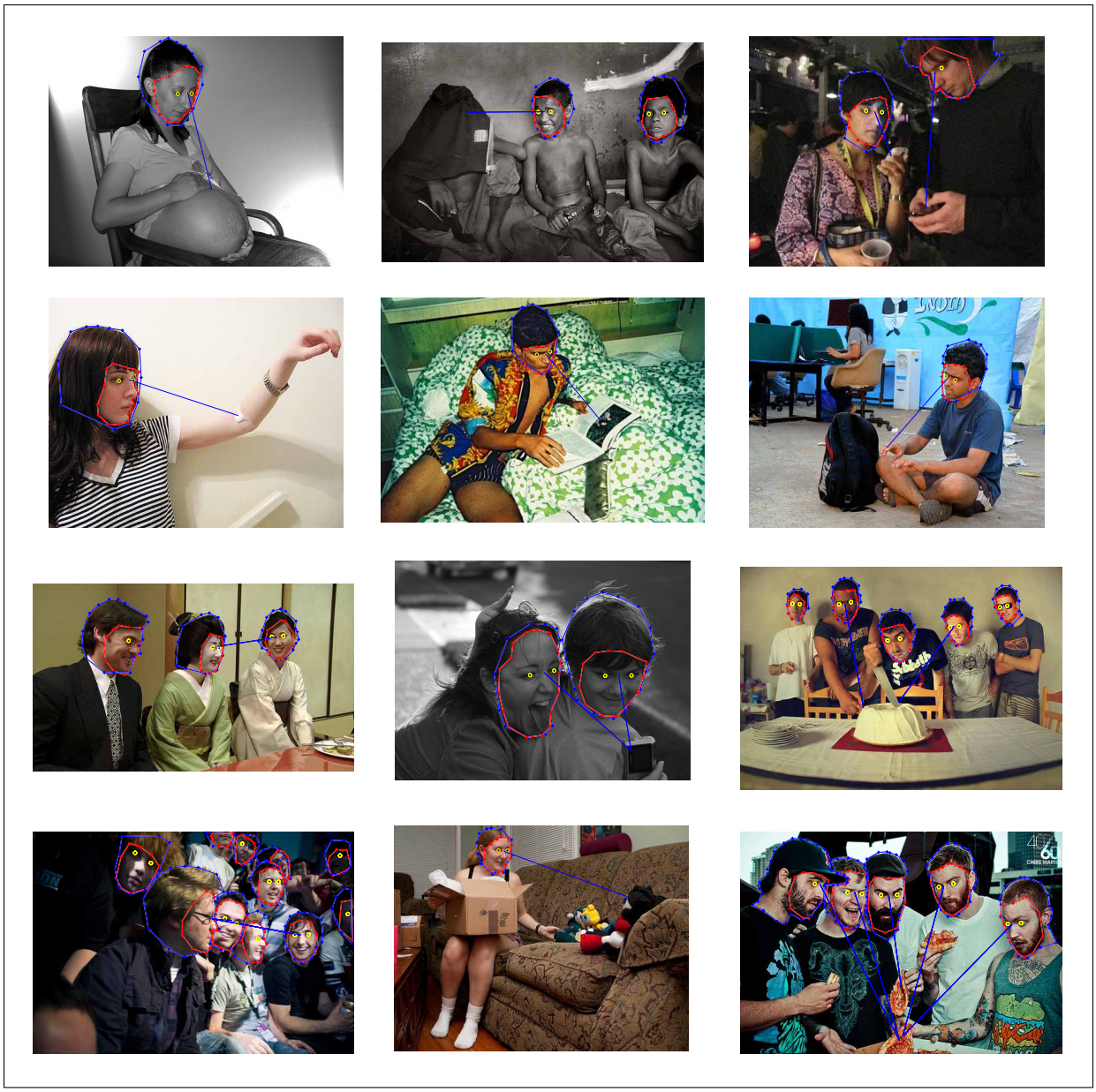

Figure 4: Sample images in the second experiment with heads, faces, eyes, and gaze directions annotated. Blue lines indicate ground truth gaze direction. We use the head region for calculating saccade pdfs. We annotated the gaze directions of the people that looked at something or a person in the scene (and not to the camera or out of the image plane). All faces were annotated.

analysis. Images were shown to observers in two sessions with 100 images each. Observers had 5 minutes break in between two sessions. The eye tracker was re-calibrated before the second session. We annotated heads, faces, eyes, and gaze directions for all 200 images. We annotated the gaze direction from entire image content including the head area and candidate gazed-at objects. When multiple objects are spatially close to each other, the gaze can be ambiguous. In this case, even if we can infer general head direction, it is quite hard to know which object the person is gazing at. Thus, spatial accuracy of gaze following requires a very precise computation of gaze data with a high-resolution eye image, which are not be available in all our images.

The average number of heads (with faces visible or not) in scenes was 2.65 (SD = 2.02, Median = 2). Fifty one

Figure 5: Twelve sample images in the second experiment along with their corresponding eye movement map (2nd rows), fixation map composed from all saccades (third rows), fixation map composed from all saccades that start from one face (fourth rows), and AWS saliency maps (fifth rows). Eye movement data is over all observers. Note that the AWS saliency model does not have an explicit face channel. Try to guess to which face the third row belongs to!

images had only one head in them, 78 had 2 and 71 images had 3 or more. Overall there were 530 heads which out of them only 305 were looking at something in the scene (had their gaze annotated). From these 305 heads, 138 were looking at another head and 167 where looking elsewhere. In every image there was at least one person whose face was visible and who was looking at someone or something visible within the image, which could be used for analysis of gaze following. Although a face occupies 2.68% of the image on average, it contained 15.3% of all fixations. The average head size was 220 $\times$ 270 pixels ($\sim 5.5^\circ \times 7.9^\circ$ visual angle). The average gaze length was 459 pixels ($\sim 11.5^\circ$).

3.1.2. Observers

A total of 30 students (4 male, 26 female) from the University of Southern California (USC) took part in the study (Mean age = 19.46, SD = 0.97). Observers had normal or corrected-to-normal eyesight and were compensated by course credits.

3.1.3. Apparatus and Procedure

Procedure and apparatus were the same as in experiment one, except that here images were shown for 10 seconds with 5 seconds gray screen in between two consecutive images.

3.2. Analysis and Results

Fig. 4 shows sample images from the stimulus set in experiment 2 and their annotations. Fig. 5 shows sample images along with their corresponding fixation maps (from all observers), blurred fixation maps, blurred maps for fixations starting from head, and AWS saliency maps. Fig. 6 shows angular saccade probability distributions for sample images in the second experiment (over all observers). Note how there is a bias toward the upper regions of fixation maps where faces are more likely to occur in our scenes (Figs. 3.C & 7.B). This bias is away from the classical center bias (Tatler, 2007; Borji et al., 2011) but is close to the hotspot present in the head annotation map.

3.2.1. Analysis of gaze following

We repeat the same analysis as in experiment one by reading out and comparing saccade pdf values in the ground truth gaze direction, in the direction of the maximum saliency location (anywhere in the scene), as well as in random directions. Results are shown in Fig. 7. We break-down the analysis (stimulus set) into three cases: 1) *All data*, 2) *Single-Head* where there is only one person in the scene looking at something, and 3) *Face Saliency Control* where a person is gazing at something other than a face, and there are multiple persons in the scene. Case 1 addresses if our results in experiment one generalize to uncontrolled complex natural scenes. Case 2 verifies generality of our results in natural scenes with only one person. Case 3 controls that gaze following is due to the gaze direction and not face saliency. In other words, check whether the gaze following effect is still present when the direction of other faces is incongruent with the actor's gaze direction. For example, consider the image in Fig. 6.A in which there are two men, one looking at a newspaper and the other looking at the first man. The question is whether an observer starting from the left man's head will saccade to the newspaper or to the right man's head. This partitioning of the dataset results in 51 samples for case 2 and 116 samples for case 3. Note that some images in case 3 may contain several faces looking at something resulting in several data points (e.g., Fig. 6.B). The overall number of gaze following data points (case 1) is 305.

Total number of saccades over all observers and images (case 1) that start from annotated heads is 21,737. From this, 4,411 saccades belong to case 2 and 7,776 saccades belong to case 3. Median saccade pdf value at the ground truth gaze direction (case 1) is 0.326 and is significantly above uniform chance level of 0.029 (sign test, $p = 4.051e-62$; over a vector of 305 probability values one for each head, across all observers). Median saccade pdf in the gaze direction in case 2 is 0.378 which is again significantly above uniform chance level of 0.046 (sign test, $p = 1.586e-$

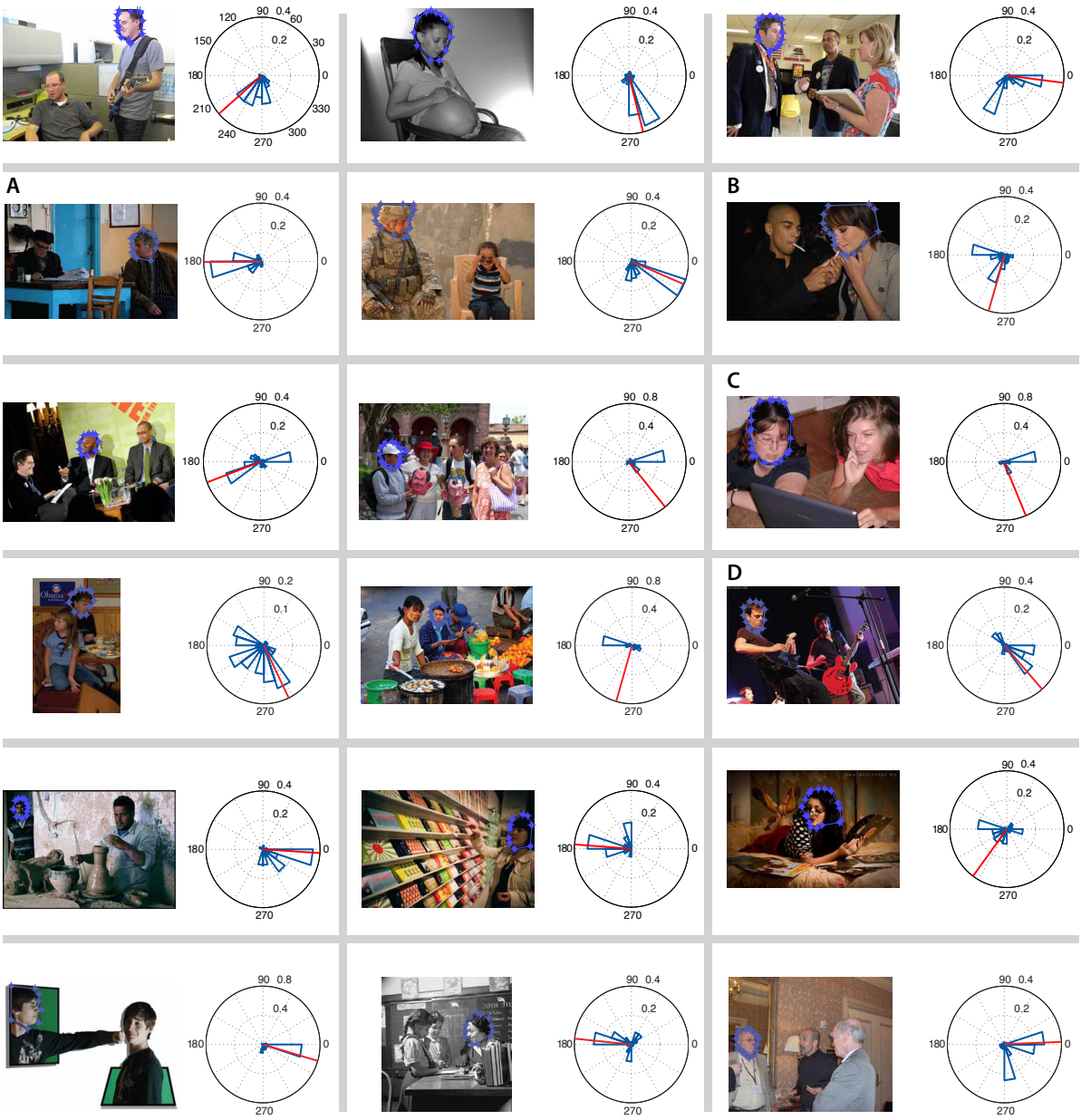

Figure 6: Sample images and their corresponding saccade direction pdfs in the second experiment (case 1). In each panel, the face under investigation in the image (left) is marked with a polygon. The polar plot (right) shows saccade direction pdf for all saccades (case 1; see text). Data is for saccades that start from in the polygon and land somewhere else in the image.

12). This is in alignment with our results in experiment one. Median saccade direction pdf at the gaze direction for case 3 is 0.241 which is above uniform chance level of 0.035 (sign test, $p = 1.167e-18$). Median gaze following strengths in all three cases are significantly above Naive Bayes chance levels using sign test (p-values in order are $3.975e-39$, $1.767e-09$, and $8.656e-10$). Comparing gaze following strengths for cases 2 and 3 shows a significant difference (sign test, $p = 1.220e-05$). This indicates that observers follow the gaze direction less on images with multiple faces suggesting that they were sometimes distracted by face saliency.

Median saccade pdf values in the direction of maximum saliency over all data (case 1) is 0.132 which is significantly above uniform chance level of 0.037 (sign test, $p = 1.102e-19$). It means that observers gazed towards

something salient in the scene from the actor's face significantly more than expected by chance. These values for case 2 and case 3, in order, are: 0.149 (significantly different vs. uniform chance using sign test; $p = 4.436e-05$) and 0.138 (significant vs. uniform chance; $p = 1.428e-07$). Saccade pdf values in the direction of the maximum salient location are significantly above Naive Bayes chance levels in all three cases (p-values in order are $1.955e-03$, $p = 2.324e-02$, and $p = 3.438e-02$).

Median saccade pdf values in ground truth gaze directions are significantly higher than median saccade pdf values in the maximum saliency direction in all three cases (p values in order are $9.696e-30$, $1.220e-05$, and $1.380e-04$). This, in accordance with experiment one, confirms that gaze direction drives saccade direction more than the direction of the most salient location in free-viewing of natural scenes.

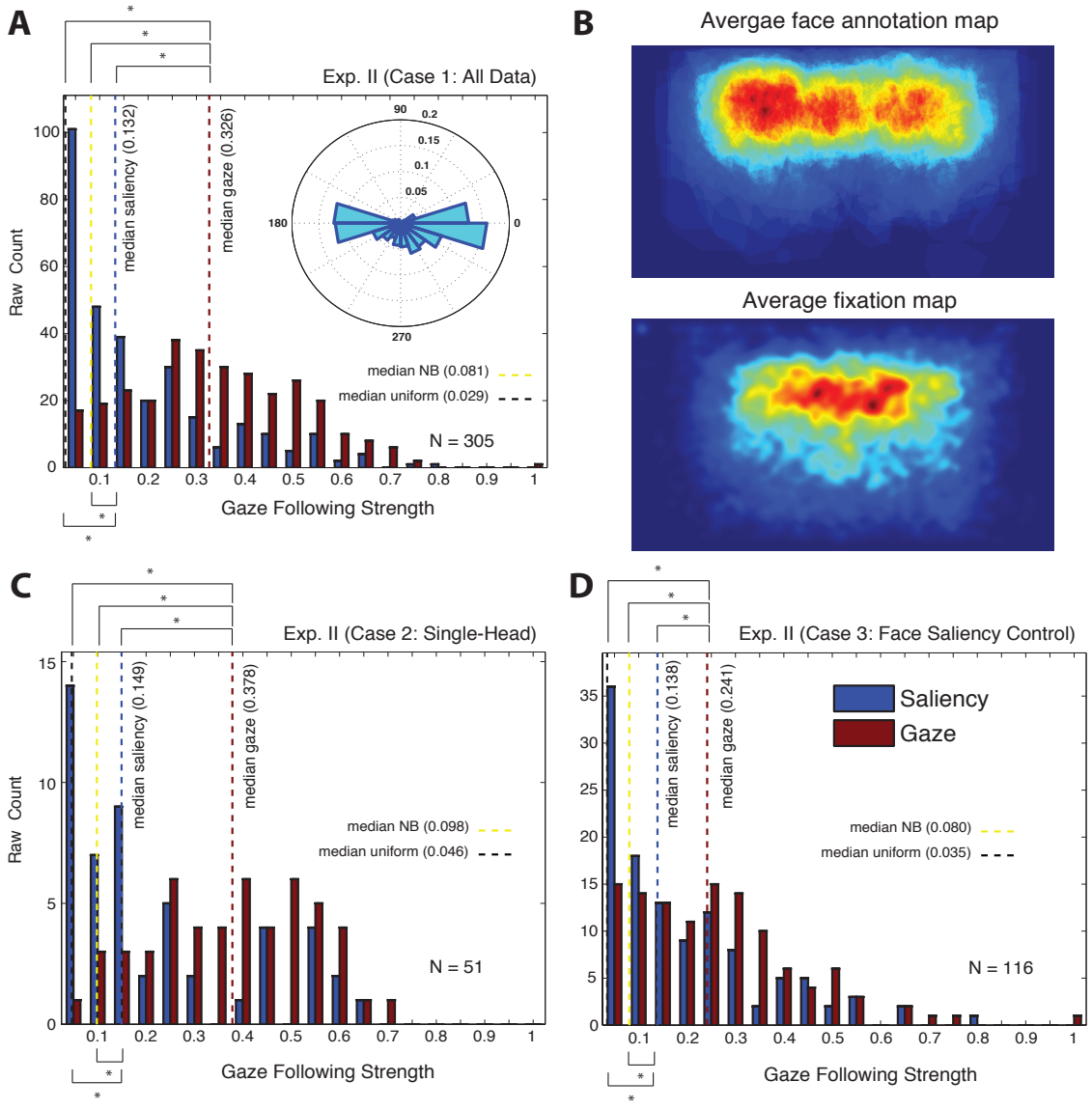

Figure 7: Results of the second experiment 2. A) Distribution of gaze following and saliency strengths over all data (case 1). Inset shows the average distribution of saccades over all data (for saccades starting from a head region). The horizontal bias could be largely caused by location of people and their heads in scenes. This prior is used for calculating Naive Bayes chance level in all three cases. B) Average head annotation (case 3) and average fixation map over all data (for saccades staring from face). Top horizontal biases can be because faces happen around the top of scenes. C&D) Distribution of gaze following and saliency strengths for cases 2&3. Gaze following strength is significantly higher than both chance levels and maximum saliency direction in all three cases.

3.2.2. *Addressing memory confound*

In our analysis so far, we considered all saccades that start from the head region. One confounding factor here is the memory of previously visited locations, which may attract or repel subsequent fixations. Observers might want to look back at the objects because they may find them somehow interesting or important, or they may want preferentially discover new items in the scene. As a consequence, one might for example first explore the scene in many directions, then look at a face, and from there follow the actor's gaze direction accidentally because it points towards a yet unexplored portion of the scene. Here, to make sure the gaze following effect is not because of memory, we only limited our analysis to the first saccades in the scene which also happened to be on the face[9]. The difference between the median of gaze following strength and median saliency strength (in the direction of the maximum location) in experiment two for just the first fixations (n=151) is statistically significant (0.428 vs. 0.000; sign test, $p = 1.142e-08$). Median values for both uniform and Naive Bayes chance levels are zero. Hence, even discounting any possible effect of memory, we still see a significant gaze following effect. Please see Appendix B for histogram of gaze following strength in this analysis.

3.2.3. *Temporal analysis of gaze following strength*

To investigate gaze following over viewing time, we analyzed the effect of saccade order in gaze following strength in experiments 1 and 2. Figure 8 illustrates the results for the first 30 saccades partitioned in bins of 10 saccades (culled from the first 30 saccades over all data, but only saccades that initiated somewhere in the head region were selected). We find that gaze following is a stronger cue during early saccades and drops over time in both experiments (it stays above maximum saliency strength over all 30 saccades). For the first experiment, the first bin of ten saccades had a significantly higher median strength of 0.42 while the second and third bins had medians of 0.28 and 0.25, respectively (Bonferroni corrected (Bland & Altman, 1995) significance value of 0.5/3=0.017; p values for bin comparisons: 1 vs. 2: $2.20e-5$, 1 vs. 3: $2.20e-5$, and 2 vs. 3: $5.96e-1$). This was also true for the second experiment, where the medians were 0.33, 0.26, and 0.27, respectively (p values for bin comparisons: 1 vs. 2: $2.60e-4$, 1 vs. 3: $1.52e-4$, and 2 vs. 3: $9.70e-1$).

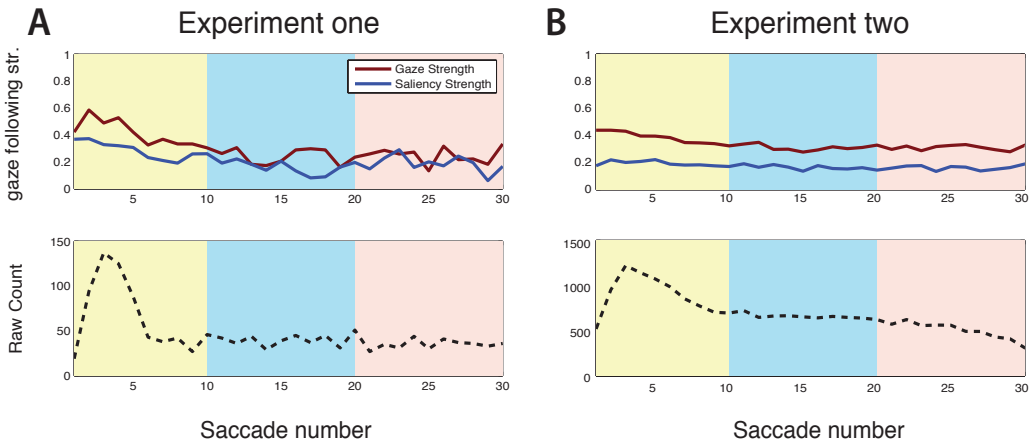

Figure 8: Temporal analysis of gaze following strength for the first 30 saccades. Average is taken for all saccades that start from a face and land somewhere else in the image. For example, an order of second means that all saccades that started from a face and were the second overall saccade for a subject on an image. In both panels, the top part shows the gaze following strength and the bottom panel shows the number of samples in each order. The data was binned into three groups when comparing strength over time: Bin-[Saccade Numbers] 1-[1:10], 2-[11:20], 3-[21:30], which are color coded: yellow, blue, and pink, respectively. A) Experiment one, B) Experiment two.

[9]We perform this analysis on the 2nd experiment since we did not have enough of such fixations in the 1st experiment (n=17).

3.2.4. Predicting fixation locations with a simple gaze map

Having shown that gaze direction influences eye movements in free-viewing, here we explore how gaze direction can be utilized to explain fixation locations. We construct the simplest map, referred to here as the *gaze map*, which has ones inside a cone ($\pm 9^\circ$ starting from a head and centered along the gaze vector) and zeros anywhere else. This cone corresponds to a single bin in the polar histograms shown in Figs. 3 and 7. Gaze maps present the advantage of being directly comparable to saliency maps, so that we can here run a direct quantitative assessment of the relative strengths of gaze following versus saliency in driving observer eye movements. Gaze and saliency maps are converted to a pdf by dividing each of their values to the total sum of their values (See Fig. 10 for example cones).

Fig. 9 shows the histogram of gaze and saliency map values at fixation locations over both experiments. As can be seen, the histogram for gaze map values are bimodal, a large leftmost peak near zero and another peak at higher values. The peak around zero is because many fixations do not fall inside the cone. Note that this a relative effect: although a large portion of fixations fall off the cone, the gaze cone still contains the largest fraction of fixations relative to any other cone of the same size in the image. Saliency map values, on the other hand, only show a left peak near zero. This observation hints toward efficient ways to integrate saliency and gaze maps (See the Discussion section). Although gaze map has higher frequencies at larger values, its median is dominated by zeros which makes its median lower than the saliency median (Dashed vertical lines in Fig. 9).

As a complementary analysis we also calculated the Receiver Operation Characteristics (ROC) curves by thresholding maps and measuring true positive rate (fraction of fixations above threshold) and false positive rate (fraction of uniformly random chosen points above threshold). Results are shown in Fig. 9 insets. The Area Under the ROC Curve (AUC) values in both experiments are significantly above chance which is AUC = 0.5 (t-test over gaze direction cases; 60 in Exp 1, $p = 3.855e-18$; 305 in Exp 2 case 1, $p = 4.79e-61$) but significantly below the AUC values of the saliency map (t-test; Exp 1, $p = 1.18e-21$; Exp 2, case 1, $p = 2.27e-56$). Thus we conclude that both gaze direction and saliency strongly influence observer eye movements, with saliency here being a stronger predictor of saccade endpoint but gaze direction still providing significant prediction performance.

Is there any benefit from gaze direction in prediction of fixation locations on top of early saliency? To answer this question, in Fig. 10 we illustrate scatter plots of gaze and saliency map predictions on all images in both experiments. Gaze map results in above chance accuracy (AUC $>$ 0.5) for 56/60 (93.33%) of images in experiment 1 and for 236/305 (77.38%) of images in experiment 2 (case 1). Corresponding numbers for saliency map are 60/60 (100%) and 299/305 (98.03%). Gaze map outperforms saliency map for 2/60 (3.33%) of images in experiment 1 and for 25/305 (8.20%) of images in experiment 2. Some success and failure cases for both maps are shown in Fig. 10 right side. Images that gaze map has high prediction power (3 in Fig. 10.A and 2 in Fig. 10.B) usually contain low background clutter and few salient objects at the gaze direction. Several reasons may lead to low performance of the gaze map such as high scene clutter, ambiguous gaze angle, and large gaze map area when the cone starts near one image corner and points to the opposite corner (See 1 in Fig. 10.A and 1&7 in Fig. 10.B).

We learn that uniform distribution of activation in the cone is not efficient as this simple cone has no sense of features or objects. In some instances, however, gaze map was able to account for observer fixations that were almost completely missed by saliency, pointing towards the possibility of future synergies (e.g., the belly of the pregnant lady or the woman watching TV in Fig. 10.B). Perhaps the best to combine saliency and gaze maps is to multiply them first and add the result to the saliency map (See Discussion section).

Table. 1 summarizes results of the first two experiments.

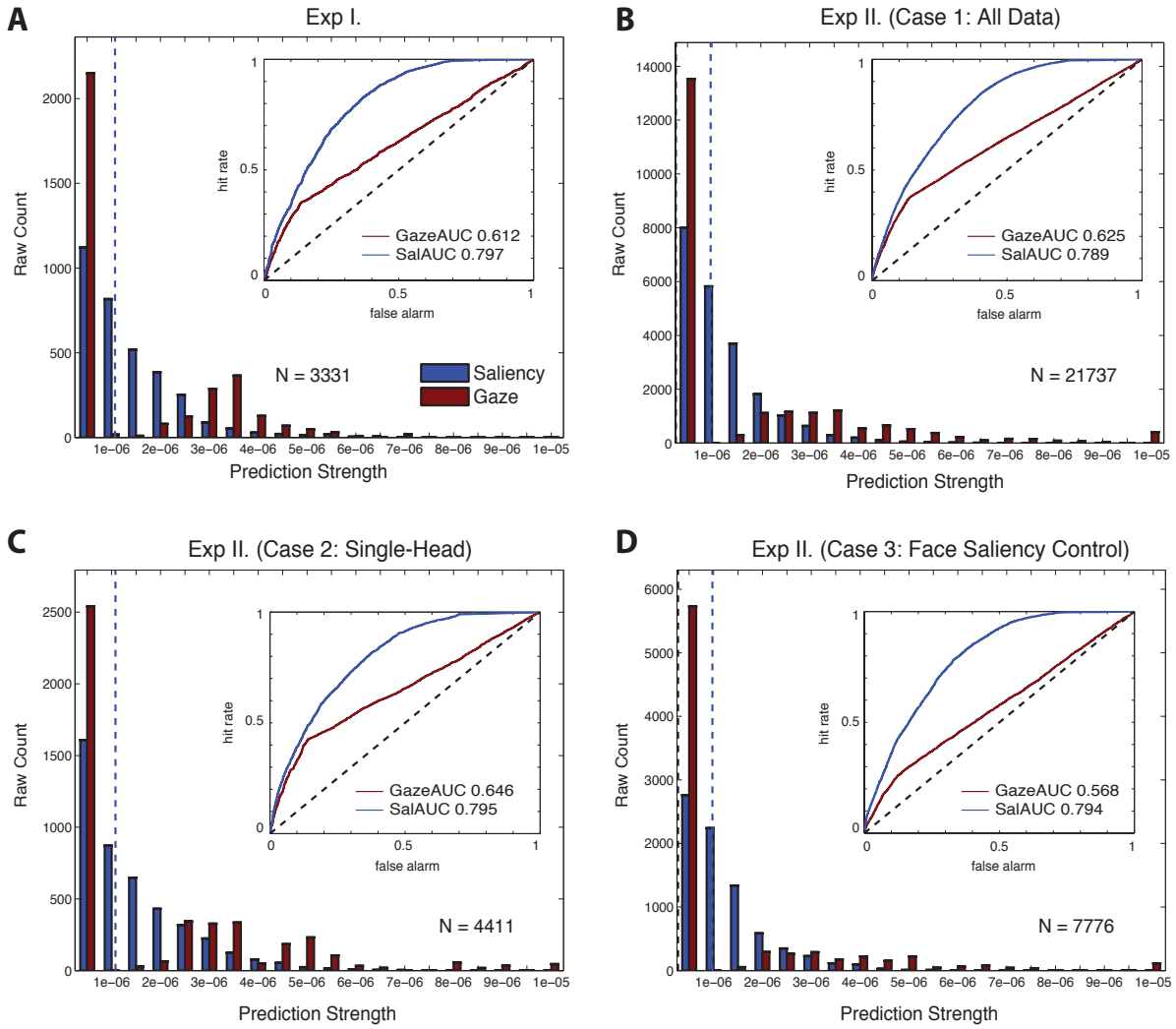

Figure 9: Fixation prediction results for our simple gaze map and the AWS saliency model over both experiments. Both maps are normalized to be pdfs. Histogram of saliency and gaze map values at fixation locations (saccades starting from head regions) for: A) controlled pairs in the first experiment, B) all data in experiment two (case 1), C) when there is only one person in the scene looking at something (Exp2, case 2) , and D) where a person is gazing at something other than a face, and there are multiple persons in the scene (Exp2, case 3). Note that, as expected, in all cases there is a peak at the left around zero for the gaze map because many saccades fell off the gaze map (i.e., misses) mainly because observers did not follow the gaze (but overall followed gaze direction higher than any other direction). Dashed lines represent medians. Insets show the ROC curves. ROC is measured by thresholding all gaze maps and then calculating the ratio of ground truth and random fixations that fall above the threshold (corresponding to true positive rate/hit rate and false positive rate/false alarm).

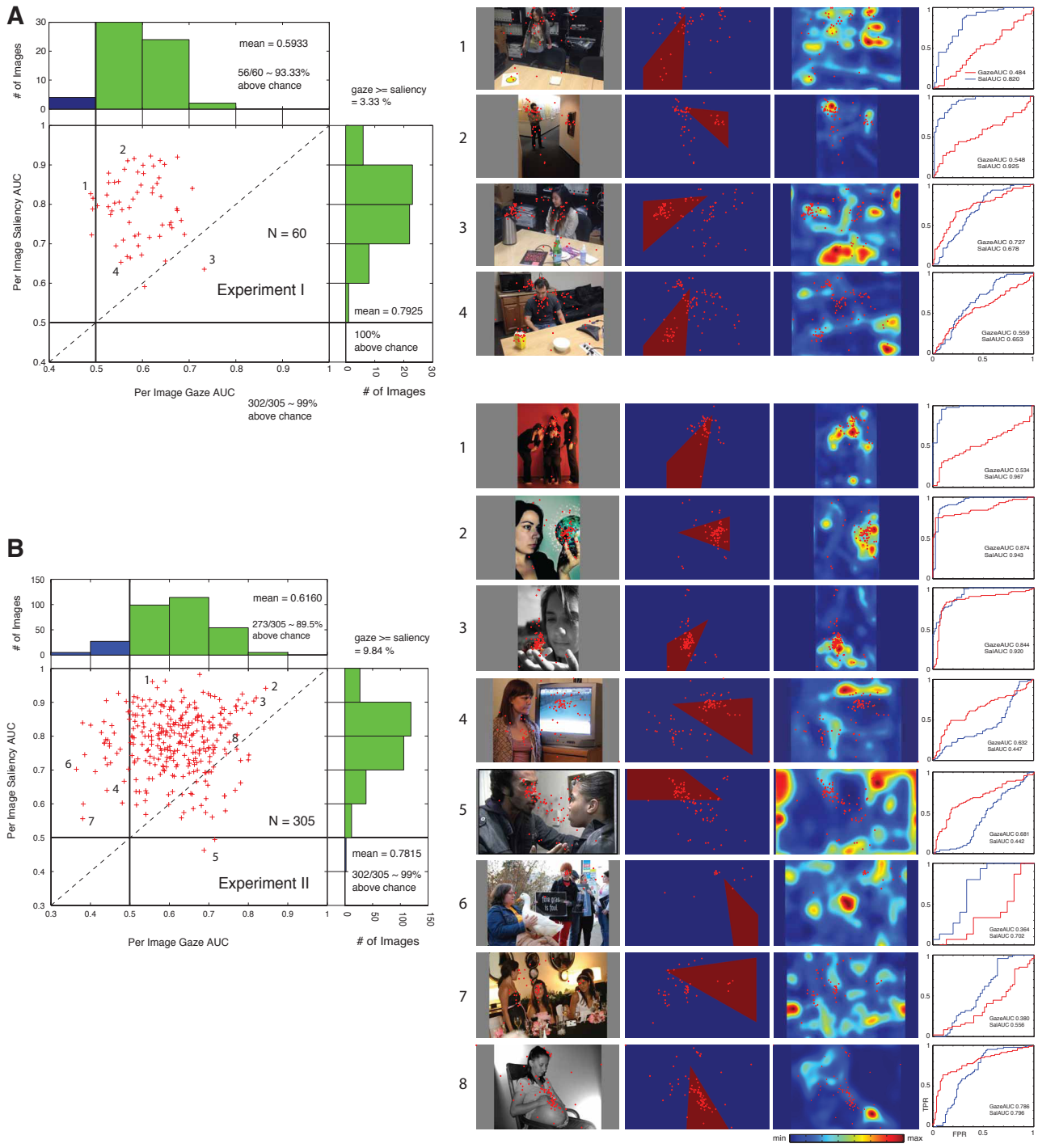

Figure 10: Gaze maps predict fixations in free viewing. A) Area under ROC curve for gaze and saliency map's prediction of fixated locations pooled over all observers in each image. Each data point corresponds to one head gaze (thresholding each map separately). Histogram of AUCs are depicted as marginals (same axes as scatter plot). For points below the diagonal, gaze map's accuracy is better than the saliency map (2 images). The opposite is the case above diagonal (58 images). Gaze map has above chance accuracy for 56 images (60 for saliency map). B) same as A but over experiment 2. Here, 25 images are below diagonal. Gaze map performs better than chance over 236 images (296 for saliency map). Example images where either one or both of maps perform well are shown in right along with their gaze maps, saliency maps, and ROC curves. Overlaid points represent eye movements. Interestingly, in some cases gaze map predicts fixations much better than saliency model (e.g., pregnant lady and the woman watching TV).

	Exp I	Ep II - Case 1	Exp II - Case 2	Exp II - Case 3
Predicting saccade direction (medians)				
Gaze direction	0.220	0.326	0.378	0.241
Most salient direction	0.101	0.132	0.149	0.138
Uniform chance	0.023	0.029	0.046	0.035
Naive Bayes chance	0.061	0.081	0.098	0.080
Predicting fixation locations (AUC values)				
Gaze map	0.612	0.625	0.646	0.568
Saliency map	0.797	0.789	0.795	0.794
Chance	0.5	0.5	0.5	0.5

Table 1: Summary results of experiments 1& 2 for prediction of gaze direction and fixation locations (for fixations that start from the head region). **For saccade direction prediction:** Both gaze and most salient directions are significant predictors of observers' saccade direction ($p < 0.05$ using sign test vs. both uniform and Naive chance levels) in two experiments and in all cases. In both experiments, gaze direction performs significantly higher than maximum saliency direction in predicting observers' saccade directions. **For saccade endpoint prediction:** Both gaze and saliency maps outperform chance significantly above chance (AUC for chance is 0.5 corresponding to a white noise map). AUCs here are calculated by thresholding all maps at a certain threshold level and measuring true positive and false positive rates across all maps. Overall, our simple gaze map explains fixations significantly below the best existing purely bottom-up saliency model, but there are some cases where our gaze map wins over the saliency map.

4. Experiment three

Our dataset provides rich opportunities to measure characteristics of gaze following that may provide important constraints for future modeling efforts. Thus, here we explore inter-observer variability in estimating gaze direction and gaze distance, as well as the relative importance of head pose (inferred from the face region) versus eye orientation in a person's ability to estimate final gaze direction, as these may guide the development of computational models.

Several studies have carefully explored the accuracy of humans observers in eye gaze and head pose estimation under various conditions (e.g., in controlled and virtual environments). Langton et al. (2004) studied the effect of head contour and nose angle on head pose estimation. Kluttz et al. (2009) inspected the effect of head turn on gaze perception. Todorovic (2009) found a face eccentricity effect on gaze perception. Gibson & Pick (1963) showed that proper gaze estimation relies on the relative position of the iris within the eye socket as well as the head pose. For instance, when the eyes are looking right while the head is facing left, they combine to form a straight ahead gaze. Or when the eyes are looking straight ahead along with a left facing head, which combine to produce a left facing gaze (See Figure 2, page 389 in Gibson & Pick (1963)). Wilson et al. (2000) studied head pose estimation accuracy for a limited horizontal range (-30° to 30°). They found a 2° discrimination threshold for 0° - 15° head poses and a threshold of 4.9° for 30° head poses.

Humans' ability of gaze estimation in uncontrolled and unconstrained scenes have so far been overlooked. In this experiment, we aimed to study the accuracy and the degree to which observers can explicitly determine the gaze direction of people on natural scenes. Further, we want to know which item is more informative in determination of final gaze direction: *faces* or *eyes*. Note that gaze direction depends on both. Results of this experiment will be particularly important to enhance the gaze map suggested in the previous section and build general saliency models that utilize both face and eyes. The reason for choosing faces was because they are easier to detect and incorporate in a gaze-augmented saliency model. The same experimental setting, however, can be used to study head regions.

4.1. Methods

4.2. Stimuli

Stimuli were the same as in experiment two (i.e., 200 color photographs). See Fig. 11. Ground-truthing was done by using the entire image while observers only viewed the head region. Thus, the ground truth is only approximate because it was created posthoc from a collection of pictures that have been captured beyond authors' control in a wide range of settings. While this is a limitation, we believe it is outweighed by the strength of using a diverse uncontrolled dataset.

4.3. Observers

We had two groups of observers. Fifteen observers (4 male, 11 female) were in group one with a mean age of 19.53 (SD = 1.50). Thirteen observers (2 male, 11 female) were in the second group with a mean age of 19.46 (SD = 1.26). Observers were undergraduate students from USC and had normal or corrected vision.

4.4. Procedure

Observers were asked to draw a line starting from a point on the face (pre-marked by the experimenter and falling somewhere between the two eyes) to the point where they think the person in the image is looking at. Observers were required to guess both gaze direction and gaze end point. Observers in group one were shown faces with eyes visible while observers in the second group were shown faces with eyes masked. The experiment was self paced and observers had to press any key to proceed to the next trial. Fig. 11 shows distribution of gaze estimations for 12 sample images with eyes-visible and eyes-masked conditions.

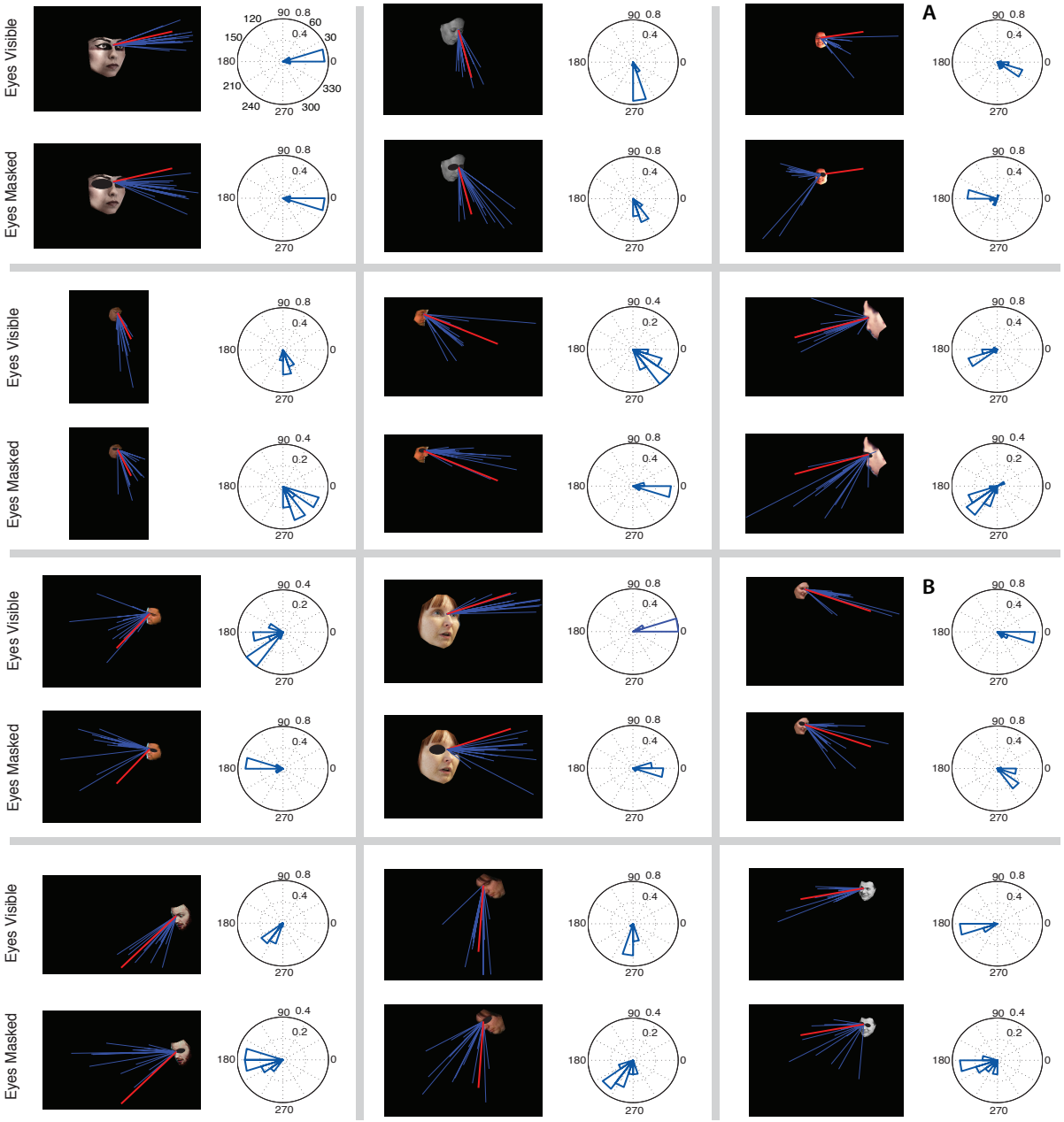

Figure 11: Sample images in the third experiment. Each panel shows the image, the ground truth gaze direction shown by a red line (annotated in the unmasked original scene), and observers' estimations of gaze directions when eyes are visible (top left) and when eyes are masked (bottom left). Top right and bottom right show corresponding polar distributions. The initial point from where observers start their vector has been annotated by the experimenter. For some images, eyes are critical for correctly estimating the gaze direction (e.g., panel A). For some cases, observers were equally good with eyes-visible and with eyes-masked (e.g., panel B) although it seems that in general observers had more variance in their decisions in the eyes masked case.

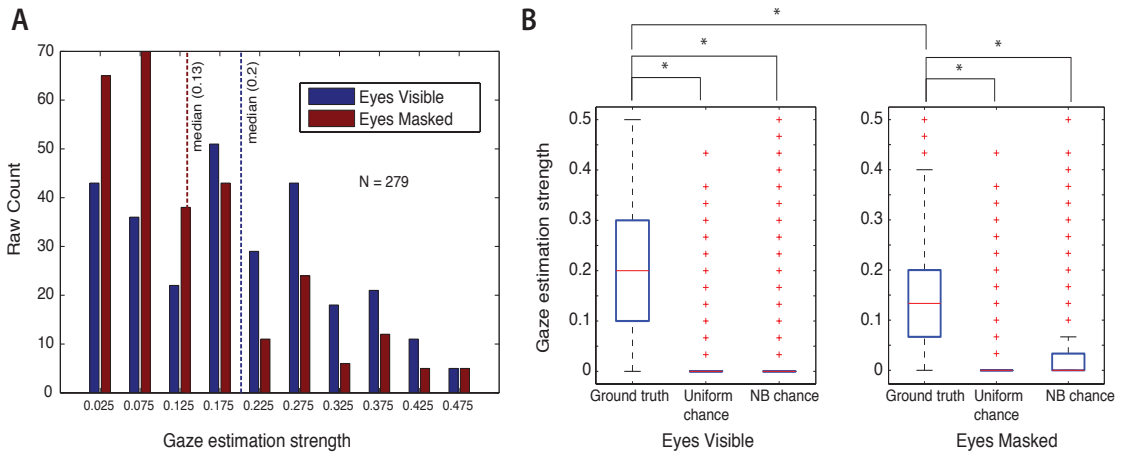

Figure 12: Results of the third experiment. A) histogram of gaze estimation strength by observers for eyes-visible and eyes-masked cases. B) Median gaze strengths for ground truth and two chance levels for two cases as well as statistical test using sign test. Median value for chance in the eyes-visible case is 0 (Mean=0.022; SD = 0.065) also for the eyes-masked case is 0 (Mean =0.021; SD = 0.058). See Fig. 7.A; inset for Naive Bayes (NB) prior.

4.5. Analysis and Results

4.5.1. Analysis of observer gaze estimation accuracy

The analysis was carried out in the same fashion as in the first two experiments except here we replaced the saccade direction pdf with the pdf of observers' estimated gaze direction (See Fig. 12). We read out the estimated pdf values in the ground truth gaze direction and in a uniformly random direction. This way, we generated two real-valued vectors of size 279 (each value in [0 1]). The medians of these vectors are significantly different from each other for eyes-visible (0.2 vs. 0; sign test; $p = 3.60e-129$) and eyes-masked cases (0.13 vs. 0; sign test; $p = 2.05e-107$). We also calibrated a smart (Naive Bayes) chance by sampling from the prior distribution over the ground truth gaze directions (see Fig. 7.A; inset). The median gaze estimation using this chance level for the eyes-visible case is 0 (Mean= 0.035; SD = 0.087) which is still significantly lower than observers' annotations ($p = 3.75e-110$). Naive Bayes median chance level for the eyes-masked case is also 0 (Mean=0.040; SD = 0.089) which is again significantly below the observers' annotations (sign test; $p = 3.00e-76$).

We compared eyes-visible and eyes-masked cases and observed that the difference is statistically significant (sign test; $p = 5.76e-15$). This indicates eye regions convey information regarding estimation of gaze direction, as observers are better at determining gaze direction when eyes are visible (equal to eyes-masked condition; pdf value of 0.13). This result is important from a modeling perspective in that, when constructing a saliency model that incorporates gaze direction, one might want to weigh pose detection more than eye gaze detection.

Fig. 13 shows the normalized gaze length for ground truth annotations (gaze direction vectors) and estimated gaze vectors (by the observers) in experiment three for both conditions. Normalization is performed by dividing the length of the drawn vector to the length of a vector in the same direction extending to the image border. Histograms of ground truth and estimated gaze lengths look similar as both have a central peak around $0.4 - 0.5$. It seems that observers slightly underestimated the gaze length and leaned towards closer distances to the person in the image. This result means that observers are to some extent able to locate the object a person is looking at (by using relative size cues, scene depth, etc). It is also important from the sense that it can be used for spatially biasing a gaze map (e.g., emphasizing more at a distance half way along the cone to the image edge instead of uniformly distributing activations across the entire map).

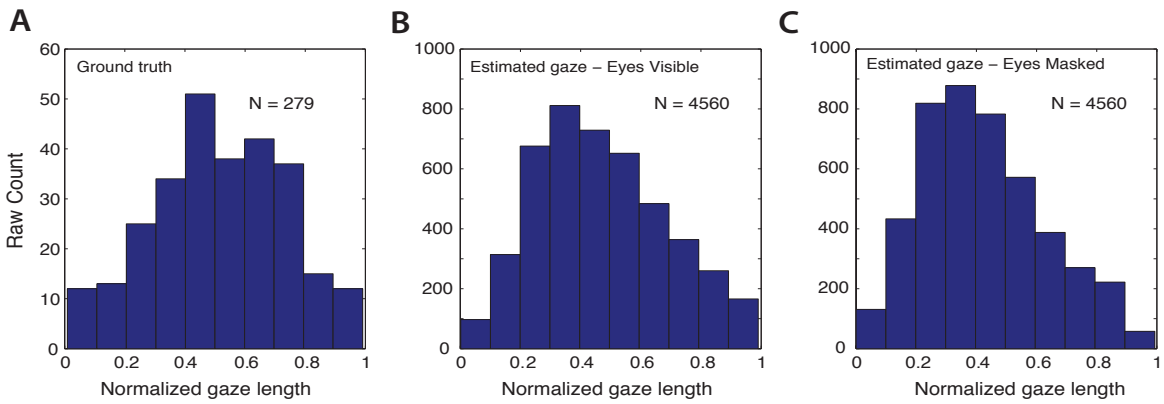

Figure 13: Histogram of normalized saccade length for: A) ground truth gaze annotations (by the experimenter and when the entire scene is visible), B) observers' guesses with eyes of people in the scene visible, and C) observers' guesses with eyes of people in the scene masked. All three histograms show a central peak around $0.4 - 0.5$.

4.5.2. Analysis of observer gaze estimation uncertainty

We also analyzed uncertainty of observers in estimating gaze angle in both conditions. For each ground truth gaze annotated face, we measured the standard deviation of estimated gaze angles (in degrees). Histograms of values are shown in Fig. 14.A. Median standard deviation for eyes-visible is significantly lower than the eyes-visible condition (11.77 vs. 18.73; sign test, $p = 4.092e - 17$) indicating that observers were more consistent when people's eyes were visible in the scenes. This analysis helps understand how to spread angular uncertainty in a gaze map.

Finally, to obtain a sense of the cone width that a gaze map should use, we measured the range of observers' estimated angles. Fig. 14.B shows the histogram of absolute differences between maximum and minimum angles over all annotated faces. As it shows, the median range is significantly smaller when eyes were visible in images versus when they were masked (41.09 vs. 68.29; sign test; $p = 2.475e - 16$). This result suggests that a cone width about 41° might be a good choice for building a gaze map. Here we attempted to study the ability of observers to explicitly report the gaze direction to constrain the parameters of a gaze map. Another alternative would be learning parameters directly from eye movements for fixation prediction.

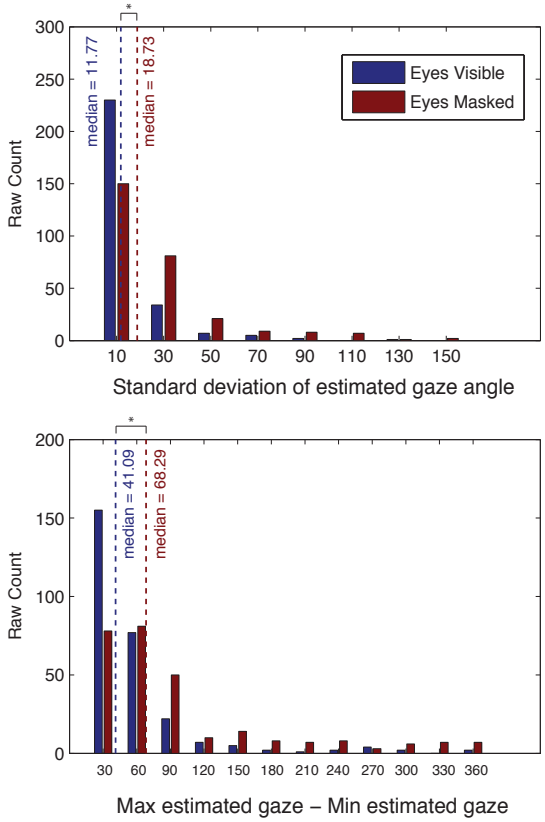

Figure 14: A) Histogram of gaze estimation variance in exp 3 (in degrees), B) Histogram of range of estimated gaze angles (cone width).

5. Discussion

Our quantitative results in the first two experiments indicate that free-viewing observers strongly follow the gaze direction of human actors in natural scenes. In experiment one, we found that the fraction of fixations that start from a head region and land on an object was significantly higher for the attended object compared to the ignored object. While actor gaze was a stronger predictor of saccade direction than saliency or chance in both experiments, it performed worse than saliency in predicting saccade endpoint, although above chance. We also noted that observers follow the gaze direction less on images with multiple faces, suggesting that they were sometimes distracted by face saliency (results of our 2nd experiment; higher gaze following strength and AUC values for case 2 vs. case 3 in Fig. 7 and Fig. 9, respectively). However, it remains to be investigated which cue (face, saliency, or gaze direction) observers prioritize in viewing natural scenes. In addition to our quantitative results, we also noted a few qualitative observations. Our observers sometimes followed the gaze direction of inanimate objects (statues, robots, dolls, masks, etc). We saw a cyclic behavior in fixation patterns in alignment with Yarbus (1967), DeAngelus & Pelz (2009), and Borji & Itti (2014), such that observers look back and forth between the actor and the gazed-at object. Contrary to our expectation, we did not find a correlation between low-level saliency and gaze following strength.

We suggest that four main components are involved in a gaze following behavior: 1) stimulus (scene content, actors and objects at gaze endpoints), 2) observer top-down knowledge and internal states, 3) task, and 4) context or environment in which gaze following happens. Here, we looked at the first component which is the effect of actor gaze direction on observer viewing behavior. Some studies have investigated how mood and emotional states of actors modulate the gaze following effect. For example, it has been shown that gaze following in monkeys is modulated by

observed facial expressions (Goossens et al., 2008). Also it has been suggested that fearful emotions result in stronger gaze cueing effects than happy faces (Kuhn & Tipples, 2011). The complexity of the scene or scene clutter (i.e., images with large field of view which may contain many faces, objects, background clutter, etc) might also influence the magnitude of the gaze following. If the scene is very complex, observers may not follow actor gaze as much, particularly when the presentation time is short (See for example Fig. 6.D). Instead, they may use their time to scan the entire scene (i.e., gaze direction as a function of scene complexity).

The second component is observer's top-down inferential mechanisms which may also interact with the first component. For example, observers may not follow the gaze of an actor who is looking outside the scene (e.g., when a person in the scene is near the right side of the image and is looking toward the right), looking at an open space (a mountain, beach or horizon), or looking at an object which is close to the actor's body (in such case observer may perceive actor and object both at the same time). Another example is a person looking in a book, newspaper or a monitor which is facing toward him but its back is toward the observer (See for example Fig. 6.C).

Task, the third component, also can play a major role in gaze following strength. For example, when asked to "guess what people in the scene are doing", "make a story of the scene", or "guess relationships among people and objects" observers might be more influenced by the gaze direction. Conversely, when observers are asked questions such as "search for a particular object" or "remember positions of objects", one might expect less role for gaze direction. This is similar in spirit to Cerf *et. al*'s finding that salience of faces and text is task dependent (Cerf et al., 2009).

Finally, the fourth component is context, where the current saccade decision is influenced by a more global understanding of the image (e.g., that it is a scene of a beach, and particular regions are more likely to be interesting). Less work has addressed the second and fourth components of gaze following behavior, which warrant more attention in the future. For example, emotional and physical state of the observer or his mental state (e.g., alertness, sleepiness, and confusion) may influence the gaze following strength. This relates to findings that normal human beings from different categories and/or under different conditions might have different viewing behaviors (e.g., depending on gender (Shen & Itti, 2012) or culture (Chua et al., 2005)). Regarding context, viewers may follow the gaze direction more strongly in a risky environment than a safe environment.

In experiment 1 we introduced a novel causality-testing stimulus set. The same paradigm can be used to study other effects in viewing behavior of observers and reduce complexity of the stimuli. In the second experiment, we showed that gaze following happens on arbitrary natural scenes containing humans and objects. It still remains to be discovered what properties of a face modulate gaze following more. For example, do observers look more at the face in the direction of gaze than a control face somewhere in the image (i.e., dyadic and triadic gaze interactions with objects and humans)? More controlled stimuli and tasks are needed to measure the gaze following strength for the following example conditions: a) three faces, two of them looking at each other, the third one facing the camera, b) three faces, two of them looking at an object, the third one facing the camera, c) three faces, two of them looking at an object, the third one looking at another object, and d) four faces, three of them looking at the other face while he is looking at an object.

Note that we found that gaze direction while providing a strong directional cue, overall is a weaker predictor than saliency in predicting the saccade endpoint, and it is important to qualify this statement. First, here we employed the best existing bottom-up saliency model according to a recent comparative benchmark of 35 saliency models (Borji et al., 2012a). This is particularly important since using weaker models can sometimes reverse the conclusions of a study (see for example Borji et al. (2013a)). Yet, like all models, this saliency model is only a coarse approximation to human saliency. Second, most saliency models analyze the image along a number of feature dimensions, some of which can be fairly complex (e.g., face detection, text detection). Because saliency modeling is a mature field and these models are complex, it would have been quite surprising that gaze direction alone (in the form of our simple gaze map) might have surpassed complex and sophisticated saliency models, which integrate many known cues that attract attention.

However, our results point towards a possible future synergy between gaze direction models and saliency models. Our analysis shows an important concept which is that predicting fixations is different from predicting directions. Gaze can be viewed as something like a face or text channel, which alone cannot explain a large number of fixations, but when combined with bottom-up saliency (assuming that they are reliable and make rare mistakes) can enhance the prediction power of the model (i.e., biasing saliency in one direction).

Here, we discuss how a saliency model that takes advantage of gaze direction could be built. Such a model should include three components. *First*, it should be able to detect heads, faces, and eyes. *Second*, it should be able to detect gaze direction of people in the scene (head and gaze directions; here called the effective gaze direction). *Third*, it should have a mechanism to bias the saliency map in the direction of the gaze. The first problem is considered almost solved (at least for faces) as nowadays reliable face detectors such as Viola & Jones (2001) exist, though their performance is mainly limited to frontal-view faces. The second problem has recently been addressed (such as Zhu & Ramanan (2012); Marin-Jimenez et al. (2014); Park et al. (2012); Asteriadis et al. (2013)) ranging from works that estimate body pose (e.g., Ferrari et al. (2008)), eye gaze (e.g., Wang et al. (2003)), head orientation (e.g., Murphy-Chutorian et al. (2007)), or a combination of all (e.g., Weidenbacher et al. (2006); Yücel et al. (2013); Valenti et al. (2012)). However, still reliable gaze direction estimation models for arbitrary 2D angles do not exist especially from only pixel data (also in user-independent calibration-free conditions). Another challenge here is the projection from the 3D world to 2D images which may cause confusion about what object a person in the scene is looking at. Further, eye detection and tracking remains challenging due to the individuality of eyes, occlusion, variability in scale, location, and lighting conditions. Not much effort have been devoted to the third problem mainly because problem two is still unsolved.

In our second experiment, we addressed the third problem from above. We studied the prediction power of a very simple gaze map and offered insights towards building more predictive gaze-augmented saliency models. We discovered that four challenges need to be addressed before making a general model: 1) *how to adjust the parameters of the cone?* Some parameters that constrain a gaze map include width of the cone, angular uncertainty, and effective length from the center to focus energy (attentional span). Ideally, one would want to only bias the object(s) that the actor is looking at in the scene. If one modulates everything along the cone, then background unattended objects falling on the cone will be modulated as well (which is not efficient). In the second experiment, we employed a uniform map. Results of our third experiment can be used to calibrate this simple map. For example, a better than uniformly distributing activation would be to emphasize more at the center line (bisector angle) of the cone and decay it using a Gaussian distribution (e.g., similar to Schauerte et al. (2010) who used cone maps to emphasize pointing direction). It might be also better to emphasize more at the center of the cone map (i.e., somewhere between the origin of the cone and the image border). In the end, it should be studied whether a cone is the best shape to use to modulate a saliency map according to actor gaze direction, 2) *how to combine a low-level saliency map S and a gaze map G?* Some possibilities include addition, multiplication, or a combination of both (i.e., $\alpha S + \beta G + \gamma SG$ and learning free parameters α, β, and γ from eye movement data), 3) *how to detect and handle cases where observers do not follow the gaze direction?* Some examples include images where people look out of plane or look at something outside the image, 4) *how to combine multiple gaze direction cues present in a single image?* For example, when multiple people are jointly looking at an object, how should gaze direction cues be integrated in constructing once cone map?

We believe the result of third experiment is important for two reasons: First, it helps constrain parameters of saliency models that bias predictions in the direction of gaze. For example what would be a good cone width to use in a model given the variability in lines drawn by our observers. This information is not dependent on the accuracy of our ground truth gaze directions. Second, currently we are far away from models that can accurately detect gaze direction in digital images. Some moderately accurate models for head pose detection exist but almost no model yet addresses the final gaze direction since eyes are often small in images and are poorly visible. It is very important to

know to what extent each element carries information regarding the final gaze direction. Gaze estimation in natural scenes can be inherently ambiguous (e.g., when multiple objects overlap with the gaze direction). In some scenarios in daily life, however, it is possible to reliably estimate gaze for example when talking to people at a short distance in a bus. In such scenarios, both eye and head directions are important. Our results from experiment 3 show that eye direction contributes significantly to the final gaze direction.

Gaze direction is a function of both head pose and gaze direction. With eyes masked, observers are still able to report the gaze direction to some extent (we assume that subjects would rely on head pose to estimate the gaze line when eyes are masked). The standard deviation of subject responses with eyes masked was 18.73 degrees vs. 11.77 degrees with the eyes visible, suggesting that subjects were not drawing their lines at random. Further, randomly-drawn lines would lead to much smaller gaze estimation strengths than those reported in Fig. 12 for eyes-masked case. We believe this is an ecologically-relevant task and happens sometimes in real life for example when people wear dark sunglasses. In a further study, one might want to annotate the ground-truth head direction and see whether observers' judgments align better with this measure or with the ground-truth gaze direction.

More accurate gaze direction prediction and saliency models can be useful in a number of scientific and engineering applications: computer vision (e.g., action recognition, scene understanding in videos (Marin-Jimenez et al., 2014), reading intentions of people in scenes (Yun et al., 2013), attentive user interfaces), human-computer and human-robot interaction (e.g., Hoffman et al. (2006); Lungarella et al. (2003); Nagai et al. (2002); Breazeal & Scassellati (2002); Bakeman & Adamson (1984); Borji et al. (2012b)), determining the attention levels of a driver (e.g., Murphy-Chutorian et al. (2007)), and enriching e-learning systems (e.g., Asteriadis et al. (2013)). Also such models can be useful for psychological research to study psychological disorders and diagnose patients with mental illness (e.g., anxiety and depression (Compton, 2003; Horley et al., 2004; Kupfer & Foster, 1972), schizophrenia (Franck et al., 2002; Langton, 2000), autism (Klin et al., 2009; Fletcher-Watson et al., 2009)), and patients suffering from dysfunctions related to joint attention. Specifically, augmented models of low-level saliency and gaze following can be used for constructing signatures for distinguishing patients from healthy normal people. A similar trend has been followed over the years using purely bottom-up saliency models to conduct neuropsychological and clinical research (e.g., Foulsham et al. (2009); Tseng et al. (2012)).

Recent effort has been focused on new means of obtaining users' gaze on portable devices (both hand-held and wearable) such as laptops, tablets, smart phones, and glasses. Gaze tracking on such devices can complement the touch modality. Rapid progress in increasing processing power, memory, screen size, and camera resolution of these devices allows running computationally intensive applications under different conditions in real-time. Some gaze-aware applications on visual wearable devices include evaluating user experience, monitoring and enhancing reading behavior, as well as adapting and displaying out-of-reach content on the screen (e.g., Wood & Bulling (2014)).

6. Conclusion

Gaze is a critical cue for understanding the meaning of complex natural scenes but has been so far overlooked in the vision community, partly because it is hard to accurately estimate it in uncontrolled cluttered natural scenes. In the present study, we investigated behavior of passive external observers viewing complex real world scenes involving humans and objects. We showed that actor gaze direction has a profound effect on the allocation of eye movements of an observer during real-world scene exploration supporting and extending the results by Castelhano et al. (2007). From the first two experiments, we conclude that gaze direction provides a complementary source of attention guidance from the properties of the gazed-at object (e.g., its salience) and memory of visited locations. In experiment two, over a wild dataset of natural scenes, we discounted face saliency and showed that even in the presence of faces, observers follow the gaze direction. In the third experiment we showed that humans are able to report the gaze direction from

the face region and do even better when eyes are visible. This indicates that there is significant additional information provided by the eyes beyond what the head alone provides.

A promising future research direction is constructing more predictive gaze-augmented saliency models by combining saliency, face detection, and gaze direction. Studying the relationships between gaze direction, attention, language, scene and object properties is another interesting area. Some related works in this regard have investigated the connection between gaze direction and language (Houston-Price et al., 2006), attention and language (Itti & Arbib, 2006) (a notion known as the "minimal subscene"), and language and event perception (Papafragou et al., 2008; Macdonald & Tatler, 2013).

Acknowledgments

This work was supported by the National Science Foundation (grant number CMMI-1235539), the Army Research Office (W911NF-11-1-0046 and W911NF-12-1-0433), and U.S. Army (W81XWH-10-2-0076). The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States government or any agency thereof. Thanks to Boris Schauerte for his help collecting the stimulus set for our second experiment. We wish to thank reviewers for valuable comments. The data and stimuli from of this study are publicly available at http://ilab.usc.edu/borji/Resources.html.

Author contributions: Ali Borji and Daniel Parks contributed equally to this paper.
Commercial relationships: none.
Corresponding author: Ali Borji.
Email: borji@usc.edu.
Address: University of Southern California, Hedco Neuroscience Building, 3641 Watt Way, Los Angeles, CA 90089-2520, USA

References

Asteriadis, S., Karpouzis, K., & Kollias, S. (2013). Visual focus of attention in non-calibrated environments using gaze estimation. *International Journal of Computer Vision*, (pp. 1–24).

Bakeman, R., & Adamson, L. B. (1984). Coordinating attention to people and objects in mother–infant and peer–infant interaction. *Child development*, .

Baldwin, D. A. (1995). Understanding the link between joint attention and language., .

Ballard, D., Hayhoe, M., & Pelz, J. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience.*, *7*, 66–80.

Ballew, C., & Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proc Natl Acad Sci U S A.*, *13*, 17948–53.

Baron-Cohen, S., Campbell, R., Karmiloff-Smith, A., Grant, J., & Walker, J. (1995). Are children with autism blind to the mentalistic significance of the eyes? *British Journal of Developmental Psychology*, *13*, 379398.

Baron-Cohen, S., Wheelwright, S., Jolliffe, & Therese (1997). Is there a" language of the eyes"? evidence from normal adults, and adults with autism or asperger syndrome. *Visual Cognition*, *4*, 311–331.

Bindemann, M., Burton, A. M., Hooge, I. T., Jenkins, R., & de Haan, E. H. (2005). Faces retain attention. *Psychonomic Bulletin & Review*, *12*, 1048–1053.

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

Birmingham, E., Bischof, W. F., & Kingstone, A. (2009). Saliency does not account for fixations to eyes within social scenes. *Vision research*, *49*, 2992–3000.

Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: the bonferroni method. *Bmj*, *310*, 170.

Bock, S., Dicke, P., & Thier, P. (2008). How precise is gaze following in humans? *Vision Research.*, *48*, 946–957.

Borji, A. (2012). Boosting bottom-up and top-down visual features for saliency estimation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (pp. 438–445). IEEE.

Borji, A., & Itti, L. (2012a). Exploiting local and global patch rarities for saliency detection. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (pp. 478–485).

Borji, A., & Itti, L. (2012b). State-of-the-art in modeling visual attention. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, *35*, 185–207.

Borji, A., & Itti, L. (2014). Defending yarbus: Eye movements predict observers' task. *Journal of vision*, .

Borji, A., Sihite, D., & Itti, L. (2014). What/where to look next? modeling top-down visual attention in complex interactive environments. *IEEE Transactions on Systems, Man, and Cybernetics, PART A-SYSTEMS AND HUMANS*, .

Borji, A., Sihite, D. N., & Itti, L. (2011). Quantifying the relative influence of photographer bias and viewing strategy on scene viewing. *Journal of Vision*, *11*, 166–166.

Borji, A., Sihite, D. N., & Itti, L. (2012a). Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Trans. Image Processing.*, *22*, 55–69.

Borji, A., Sihite, D. N., & Itti, L. (2012b). Salient object detection: A benchmark. In *Computer Vision–ECCV 2012* (pp. 414–429). Springer.

Borji, A., Sihite, D. N., & Itti, L. (2013a). Objects do not predict fixations better than early saliency: A re-analysis of einhäuser et al.'s data. *Journal of vision*, *13*, 18.

Borji, A., Sihite, D. N., & Itti, L. (2013b). What stands out in a scene? a study of human explicit saliency judgment. *Vision research*, *91*, 62–77.

Breazeal, C., & Scassellati, B. (2002). Challenges in building robots that imitate people. *Imitation in animals and artifacts*, (p. 363).

Butterworth, G. (1991). The ontogeny and phylogeny of joint visual attention., .

Butterworth, G., & Jarrett, N. (1991). What minds have in common is space: Spatial mechanisms serving joint visual attention in infancy. *British Journal of Developmental Psychology*, *9*, 55–72.

Carmi, R., & Itti, L. (2006). The role of memory in guiding attention during natural vision. *Journal of Vision*, *6*, 4.

Caron, A. J., Butler, S., & Brooks, R. (2002). Gaze following at 12 and 14 months: Do the eyes matter? *British Journal of Developmental Psychology*, *20*, 225–240.

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research*, *51*, 1484–1525.

Castelhano, M. S., Wieth, M., & Henderson, J. M. (2007). I see what you see: Eye movements in real-world scenes are affected by perceived direction of gaze. In *Attention in cognitive systems. Theories and systems from an interdisciplinary viewpoint* (pp. 251–262). Springer.

Cerf, M., Frady, E. P., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision*, *9*.

Chua, H. F., Boland, J. E., & Nisbett, R. E. (2005). Cultural variation in eye movements during scene perception. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 12629–12633.

Compton, R. J. (2003). The interface between emotion and attention: A review of evidence from psychology and neuroscience. *Behavioral and cognitive neuroscience reviews*, *2*, 115–129.

Csibra, G., & Gergely, G. (2006). Social learning and social cognition: The case for pedagogy. *Processes of change in brain and cognitive development. Attention and performance XXI*, (pp. 249–274).

DeAngelus, M., & Pelz, J. B. (2009). Top-down control of eye movements: Yarbus revisited. *Visual Cognition*, *17*, 790–811.

D'Entremont, B., Hains, S., & Muir, D. (1997). A demonstration of gaze following in 3- to 6-month-olds. *Infant Behavior and Development*, *20*, 569–572.

Doherty, M. J., Anderson, J. R., & Howieson, L. (2009). The rapid development of explicit gaze judgment ability at 3years. *Journal of experimental child psychology*, *104*, 296–312.

Driver, J., Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E., & Baron-Cohen, S. (1999). Gaze perception triggers reflexive visuospatial orienting. *Visual cognition*, *6*, 509–540.

Droll, J. A., Hayhoe, M. M., Triesch, J., & Sullivan, B. T. (2005). Task Demands Control Acquisition and Storage of Visual Information. *Journal of Experimental Psychology Human Perception and Performance*, *31*, 1416–1438.

Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual cognition*, *17*, 945–978.

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.

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

Ferrari, V., Marin-Jimenez, M., & Zisserman, A. (2008). Progressive search space reduction for human pose estimation. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (pp. 1–8). IEEE.

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

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

Flom, R. E., Lee, K. E., & Muir, D. E. (2007). *Gaze-following: Its development and significance.*. Lawrence Erlbaum Associates Publishers.

Foulsham, T., Cheng, J. T., Tracy, J. L., Henrich, J., & Kingstone, A. (2010). Gaze allocation in a dynamic situation: Effects of social status and speaking. *Cognition*, *117*, 319–331.

Foulsham, T., J.J., B., Kingstone, A., Dewhurst, R., & G., U. (2009). Fixation and saliency during search of natural scenes: the case of visual agnosia. *Neuropsychologia*, *47*, 1994–2003.

Franck, N., Montoute, T., Labruyère, N., Tiberghien, G., Marie-Cardine, M., Daléry, J., d'Amato, T., & Georgieff, N. (2002). Gaze direction determination in schizophrenia. *Schizophrenia research*, *56*, 225–234.

Friesen, C., & Kingstone, A. (1998). The eyes have it! reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin & Review*, *5*, 490–495.

Friesen, C. K., & Kingstone, A. (2003). Abrupt onsets and gaze direction cues trigger independent reflexive attentional effects. *Cognition*, *87*, B1–B10.

Garcia-Diaz, A., Leboran, V., Fdez-Vidal, X. R., & Pardo, X. M. (2012). On the relationship between optical variability, visual saliency, and eye fixations: A computational approach. *Journal of Vision.*, *12*.

Gibson, J., & Pick, A. (1963). Perception of another person's looking behavior. *American Journal of Psychology*, *76*, 386–394.

Goossens, B., Dekleva, M., Reader, S. M., Sterck, E. H., & Bolhuis, J. J. (2008). Gaze following in monkeys is modulated by observed facial expressions. *Animal Behaviour*, *75*, 1673–1681.

Gredebäck, G., Theuring, C., Hauf, P., & Kenward, B. (2008). The microstructure of infants' gaze as they view adult shifts in overt attention. *Infancy*, *13*, 533–543.

von Griinau, M., & Anston, C. (1995). The detection of gaze direction: A stare-in-the-crowd effect. *Perception*, *24*, 1297–1313.

Henderson, J. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences.*, *7*, 498–504.

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

Hoffman, E. A., & Haxby, J. V. (2000). Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nature neuroscience*, *3*, 80–84.

Hoffman, M. W., Grimes, D. B., Shon, A. P., & Rao, R. P. N. (2006). A probabilistic model of gaze imitation and shared attention. *Neural Networks*, *19*, 299–310.

Horley, K., Williams, L. M., Gonsalvez, C., & Gordon, E. (2004). Face to face: visual scanpath evidence for abnormal processing of facial expressions in social phobia. *Psychiatry research*, *127*, 43–53.

Houston-Price, C., Plunkett, K., & Duffy, H. (2006). The use of social and salience cues in early word learning. *Journal of Experimental Child Psychology*, *95*, 27–55.

Humphrey, K., & Underwood, G. (2010). The potency of people in pictures: Evidence from sequences of eye fixations. *Journal of Vision*, *10*.

Itti, L., & Arbib, M. A. (2006). Attention and the minimal subscene. In M. A. Arbib (Ed.), *Action to Language via the Mirror Neuron System* (pp. 289–346). Cambridge, U.K.: Cambridge University Press.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature reviews. Neuroscience*, *2*, 194–203.

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.

Jones, B. C., DeBruine, L. M., Main, J. C., Little, A. C., Welling, L. L., Feinberg, D. R., & Tiddeman, B. P. (2010). Facial cues of dominance modulate the short-term gaze-cuing effect in human observers. *Proceedings of the Royal Society B: Biological Sciences*, *277*, 617–624.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th international conference on* (pp. 2106–2113). IEEE.

Kanan, C., Tong, M. H., Zhang, L., & Cottrell, G. W. (2009). Sun: Top-down saliency using natural statistics. *Visual Cognition*, *17*, 979–1003.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, *17*, 4302–4311.

Kanwisher, N., & Yovel, G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *361*, 2109–2128.

Klin, A., Lin, D., Gorrindo, P., Ramsay, G., & Jones, W. (2009). Two-year-olds with autism orient to non-social contingencies rather than biological motion. *Nature*, *459*.

Kluttz, N. L., Mayes, B. R., West, R. W., & Kerby, D. S. (2009). The effect of head turn on the perception of gaze. *Vision research*, *49*, 1979–1993.

Kobayashi, H., & Kohshima, 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.

Krieger, G., Rentschler, I., Hauske, G., Schill, K., & Zetzsche, C. (2000). Object and scene analysis by saccadic eye-movements: An investigation with higher-order statistics. *Spatial vision*, *13*, 201–214.

Kuhn, G., & Tatler, B. W. (2005). Magic and fixation: now you dont see it, now you do. *Perception.*, *34*, 1155–1161.

Kuhn, G., & Tipples, J. (2011). Increased gaze following for fearful faces. it depends on what youre looking for! *Psychonomic bulletin & review*, *18*, 89–95.

Kupfer, D., & Foster, F. G. (1972). Interval between onset of sleep and rapid-eye-movement sleep as an indicator of depression. *The Lancet*, *300*, 684–686.

Land, M. F., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision research*, *41*, 3559–3565.

Land, M. F., & Lee, D. N. (1994). Where we look when we steer. *Nature*, *369*, 742–744.

Langton, S. R. (2000). The mutual influence of gaze and head orientation in the analysis of social attention direction. *The Quarterly Journal of Experimental Psychology: Section A*, *53*, 825–845.

Langton, S. R., Honeyman, H., & Tessler, E. (2004). The influence of head contour and nose angle on the perception of eye-gaze direction. *Perception & psychophysics*, *66*, 752–771.

Louwerse, M., & Bangerter, A. (2005). Focusing attention with deictic gestures and linguistic expressions. In *Proc. Ann. Conf. Cog. Sci. Soc.*.

Lungarella, M., Metta, G., Pfeifer, R., & Sandini, G. (2003). Developmental robotics: a survey. *Connection Science*, *15*, 151–190.

Macdonald, R. G., & Tatler, B. W. (2013). Do as eye say: Gaze cueing and language in a real-world social interaction. *Journal of vision*, *13*.

Macknik, S. L., King, M., Randi, J., Robbins, A., Teller, Thompson, J., & Martinez-Conde, S. (2008). Attention and awareness in stage magic: turning tricks into research. *Nature Reviews Neuroscience*, *9*, 871–879.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial vision*, *10*, 165–188.

Mansfield, E., Farroni, T., & Johnson, M. (2003). Does gaze perception facilitate overt orienting? *Visual Cognition*, *10*, 7–14.

Marin-Jimenez, M., Zisserman, A., Eichner, M., & Ferrari, V. (2014). Detecting people looking at each other in videos. *International Journal of Computer Vision*, (pp. 1–15).

Meltzoff, A. N., Brooks, R., Shon, A. P., & Rao, R. P. (2010). social robots are psychological agents for infants: A test of gaze following. *Neural Networks*, *23*, 966–972.

Milanese, R., Wechsler, H., Gill, S., Bost, J.-M., & Pun, T. (1994). Integration of bottom-up and top-down cues for visual attention using non-linear relaxation. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on* (pp. 781–785). IEEE.

Moeller, S., Freiwald, W. A., & Tsao, D. Y. (2008). Patches with links: a unified system for processing faces in the macaque temporal lobe. *Science*, *320*, 1355–1359.

Moll, H., & Tomasello, M. (2004). 12-and 18-month-old infants follow gaze to spaces behind barriers. *Developmental science*, *7*, F1–F9.

Murphy-Chutorian, E., Doshi, A., & Trivedi, M. M. (2007). Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation. In *Intelligent Transportation Systems Conference, 2007. ITSC 2007. IEEE* (pp. 709–714). IEEE.

Nagai, Y., Asada, M., & Hosoda, K. (2002). A developmental approach accelerates learning of joint attention. In *Development and Learning, 2002. Proceedings. The 2nd International Conference on* (pp. 277–282). IEEE.

Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing, . *10*.

Okumura, Y., Kanakogi, Y., Kanda, T., Ishiguro, H., & Itakura, S. (2013). The power of human gaze on infant learning. *Cognition*, *128*, 127–133.

Papafragou, A., Hulbert, J., & Trueswell, J. (2008). Does language guide event perception? evidence from eye movements. *Cognition*, *108*, 155–184.

Park, H. S., Jain, E., & Sheikh, Y. (2012). 3d social saliency from head-mounted cameras. In *NIPS* (pp. 431–439).

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, J. S., Paul, G., Goldman, B. D., & Piven, J. (2002). Visual scanning of faces in autism. *Journal of autism and developmental disorders*, *32*, 249–261.

Pelphrey, K. A., Viola, R. J., & McCarthy, G. (2004). When strangers pass processing of mutual and averted social gaze in the superior temporal sulcus. *Psychological Science*, *15*, 598–603.

Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision research*, *45*, 2397–2416.

Posner, M. (1980). Orienting of attention. *Q. J. Exp. Psychol.*, *32*, 3–25.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, *1*, 515–526.

Ramanathan, S., Divya, S., Nicu, S., & David, M. (2014). Emotion modulates eye movement patterns and subsequent memory for the gist and details of movie scenes. *Journal of Vision*, *14(3)*.

Reinagel, P., & Zador, A. (1999). Natural scenes at the center of gaze. *Network.*, *10*, 341–50.

Ricciardelli, P., Bricolo, E., Aglioti, S. M., & Chelazzi, L. (2002). My eyes want to look where your eyes are looking: Exploring the tendency to imitate another individual's gaze. *NeuroReport: For Rapid Communication of Neuroscience Research*, *13*, 2259–2264.

Ro, T., Russell, C., & Lavie, N. (2001). Changing faces: A detection advantage in the flicker paradigm. *Psychological Science*, *12*, 94–99.

Schauerte, B., Richarz, J., & Fink, G. A. (2010). Saliency-based identification and recognition of pointed-at objects. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on* (pp. 4638–4643). IEEE.

Schütz, A., Braun, D., & Gegenfurtner, K. (2011). Eye movements and perception: A selective review. *Journal of Vision.*, *11*, 1–30.

Shen, J., & Itti, L. (2012). Top-down influences on visual attention during listening are modulated by observer sex. *Vision research*, *65*, 62–76.

Shepherd, S. V., Deaner, R. O., & Platt, M. L. (2006). Social status gates social attention in monkeys. *Current Biology*, *16*, R119–R120.

Skarratt, P. A., Cole, G. G., & Kuhn, G. (2012). Visual cognition during real social interaction. *Frontiers in human neuroscience*, *6*.

Spain, M., & Perona, P. (2010). Measuring and predicting object importance. *International Journal of Computer Vision*, .

Tatler, B., Hayhoe, M., Land, M., & Ballard, D. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of vision*, *11*, 1–23.

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

Theuring, C., Gredebäck, G., & Hauf, P. (2007). Object processing during a joint gaze following task. *European Journal of Developmental Psychology*, *4*, 65–79.

Todorov, A., Baron, S. G., & Oosterhof, N. N. (2008). Evaluating face trustworthiness: a model based approach. *Social cognitive and affective neuroscience*, *3*, 119–127.

Todorovic, D. (2009). The effect of face eccentricity on the perception of gaze direction. *Perception*, *38*, 109.

Tomasello, M. (2009). *The cultural origins of human cognition*. Harvard University Press.

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*, 766–786.

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

Triesch, J., Ballard, D., Hayhoe, M., & Sullivan, B. (2003). What you see is what you need. *Journal of Vision.*, *3*, 86–94.

Tsao, D. Y., Freiwald, W. A., Tootell, R. B., & Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. *Science*, *311*, 670–674.

Tseng, P., Cameron, I. G. M., Pari, G., Reynolds, J. N., Munoz, D. P., & Itti, L. (2012). High-throughput classification of clinical populations from natural viewing eye movements. *Journal of Neurology*, .

Underwood, G., Foulsham, T., & Humphrey, K. (2009). Saliency and scan patterns in the inspection of real-world scenes: Eye movements during encoding and recognition. *Visual Cognition*, *17*, 812–834.

Valenti, R., Sebe, N., & Gevers, T. (2012). Combining head pose and eye location information for gaze estimation. *Image Processing, IEEE Transactions on*, *21*, 802–815.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on* (pp. I–511). IEEE volume 1.

Vuilleumier, P. (2000). Faces call for attention: evidence from patients with visual extinction. *Neuropsychologia*, *38*, 693–700.

Wang, H.-C., & Pomplun, M. (2012). The attraction of visual attention to texts in real-world scenes. *Journal of vision*, *12*, 26.

Wang, J., Sung, E., & Venkateswarlu, R. (2003). Eye gaze estimation from a single image of one eye. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on* (pp. 136–143). IEEE.

Weidenbacher, U., Layher, G., Bayerl, P., & Neumann, H. (2006). Detection of head pose and gaze direction for human-computer interaction. In *Perception and Interactive Technologies* (pp. 9–19). Springer.

Wilson, H. R., Wilkinson, F., Lin, L.-M., & Castillo, M. (2000). Perception of head orientation. *Vision Research*, *40*, 459–472.

Winston, J. S., Strange, B. A., O'Doherty, J., & Dolan, R. J. (2002). Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nature neuroscience*, *5*, 277–283.

Wolfe, J. M. (1998). Visual search. *Attention*, *1*, 13–73.

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, *5*, 495–501.

Wood, E., & Bulling, A. (2014). Eyetab: model-based gaze estimation on unmodified tablet computers. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (pp. 207–210). ACM.

Yarbus, A. (1967). *Eye movements and vision.*. New York: Plenum.

Yücel, Z., Salah, A., Meriçli, C., Meriçli, T., Valenti, R., & Gevers, T. (2013). Joint attention by gaze interpolation and saliency. *Cybernetics, IEEE Transactions on*, *43*, 829–842.

Yun, K., Peng, Y., Samaras, D., Zelinsky, G. J., & Berg, T. L. (2013). Exploring the role of gaze behavior and object detection in scene understanding. *Frontiers in Psychology.*, *4*.

Zhu, X., & Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (pp. 2879–2886). IEEE.

Zwickel, J., & Võ, M. L.-H. (2010). How the presence of persons biases eye movements. *Psychonomic bulletin & review*, *17*, 257–262.

7. Appendix A

Figure 15 shows the complete set of scenes used in experiment one (30 pairs of images).

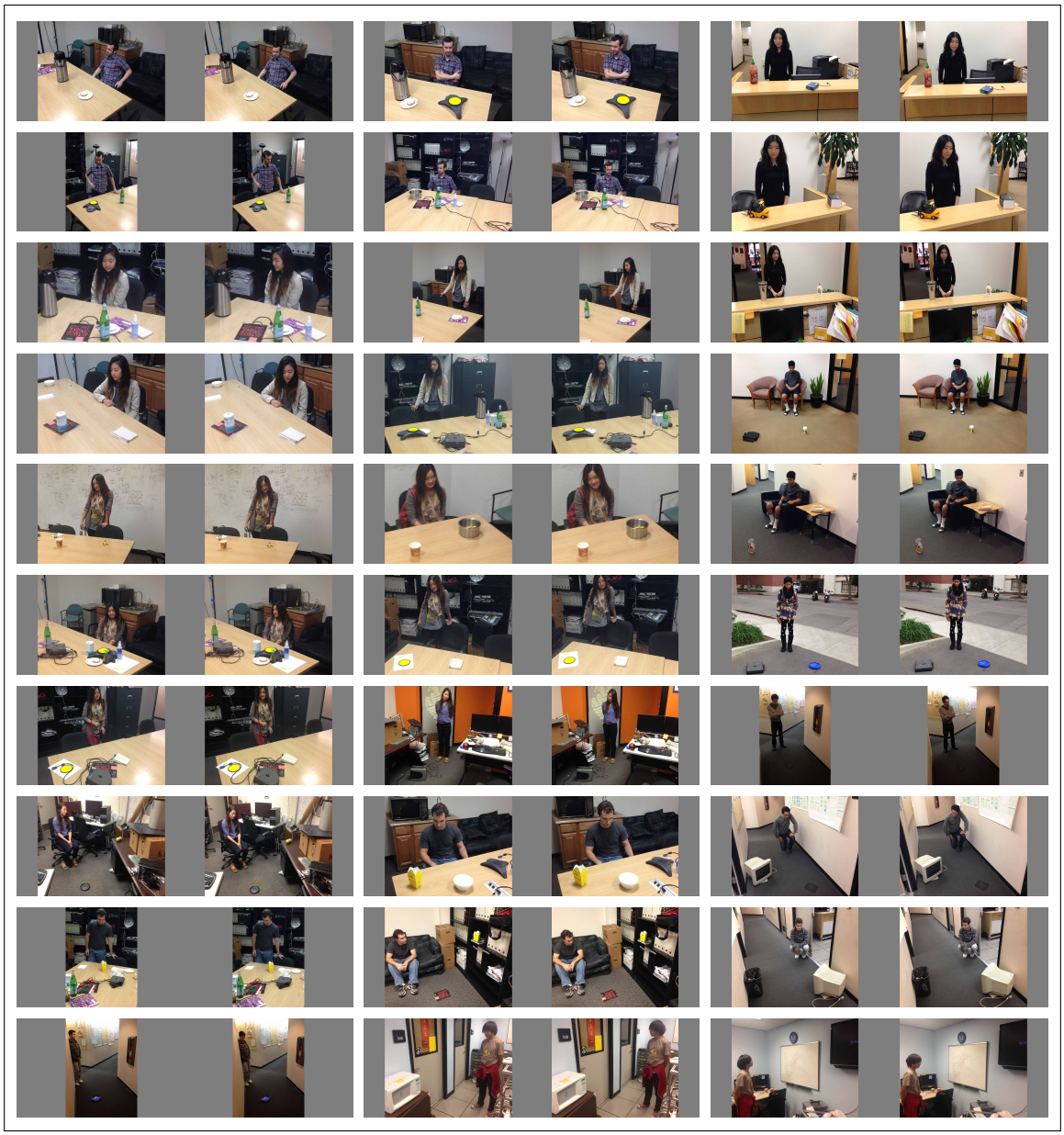

Figure 15: The complete set of scenes used in experiment one (30 pairs of images).

8. Appendix B

Fig. 16 depicts histogram of gaze following strengths over all data in both experiments for first saccades that start from faces. Such saccades are not contaminated by memory effect.

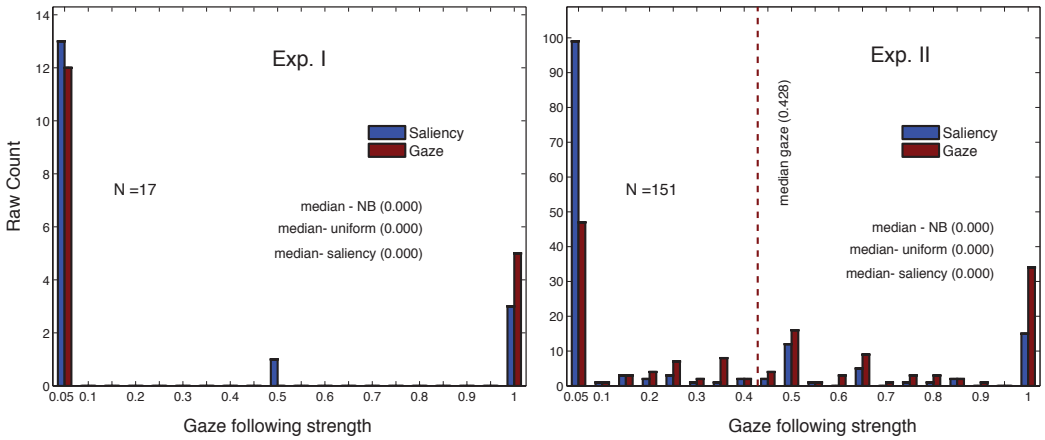

Figure 16: Gaze following strength for experiment one and all data (case 1) in experiment two. Data include only first saccades that also happen to be on the faces.