Assessing a Bayesian account of human gaze perception

Peter C. Pantelis (PCPANTEL@indiana.edu)
Daniel P. Kennedy (DPK@indiana.edu)
Indiana University–Bloomington, Department of Psychological and Brain Sciences
1101 E. 10th Street, Bloomington, IN 47405 USA

Abstract

Although gaze can be directed at any location, different locations in the visual environment vary in terms of how likely they are to draw another person's attention. One could therefore weigh incoming perceptual signals (e.g., eye cues) against this prior knowledge (the relative visual saliency of locations in the scene) in order to infer the true target of another person's gaze. This Bayesian approach to modeling gaze perception has informed computer vision techniques, but we assess whether it is a good model for *human* performance. We present subjects with a "gazer" fixating his eyes on various locations on a 2-dimensional surface, and project an arbitrary photographic image onto that surface. Subjects judge where the gazer is looking in the image. A full Bayesian model, which takes image saliency information into account, fits subjects' gaze judgments better than a reduced model that only considers the perceived direction of the gazer's eyes.

Keywords: gaze; Bayesian modeling; social perception; social attention; visual saliency

Introduction

Another person's gaze direction is a strong cue for where this person may be directing his or her visual attention, and therefore helps one to infer what may be on his or her mind. Additionally, because people (and other animals) tend to direct their visual attention to the informative and behaviorally relevant areas of the environment (Mackworth & Morandi, 1967), the ability to infer attention also provides hints as to the important things that may be happening in the immediate environment (Byrne & Whiten, 1991). Given that the direction of another person's eye fixation is a basic cue for tracking gaze (and therefore, attention), the human visual system has evolved to process this perceptual signal with remarkable accuracy and efficiency (Cline, 1967).

Nevertheless, the perceptual signal extracted from another person's eyes is noisy and ambiguous. As such, other cues like head position (Wallaston, 1824; Ken, 1990; Langton, 2000) also inform the judgment of gaze direction. But additionally, if one has reliable intuitions about where in the visual scene another person is likely to direct his or her gaze—*a priori* of perceiving the signal from his or her eyes—then this prior information could potentially be integrated with the eye cue to improve the inference of gaze direction.

This basic approach—combining perceptual cues from the target person's eyes (or head position, etc.) with the visual saliency of the scene—has been exploited in computer vision to improve machine performance both in the discrimination of gaze direction (Hoffman, Grimes, Shon, & Rao, 2006; Yücel et al., 2013) and in the related task of identifying where another person is pointing (Schauerte, Richarz, & Fink, 2010). And although it has been speculated that human gaze perception may employ a similar mechanism, this remains untested. Thus, we here ask whether a Bayesian model that incorporates a visual saliency map as a prior can account for actual human subjects' performance better than one which ignores this information, and uses only the eye cues.

Our experimental subjects view photographs of another person gazing at various locations on a partially transparent surface situated between him and the camera. The subjects are instructed to indicate where on this surface this other person is looking; we define this task computationally as the inference of the location $[x, y]$ where the photographed individual is **g**azing within the continuous 2-dimensional plane ($G_{x,y}$) given the gaze **d**irectional cue from the eyes of the person (D) and the **i**mage presented in that plane (I). Bayes' rule yields the posterior probability distribution, continuous over the 2-dimensional hypothesis space:

$$p(G_{x,y}|D) \propto p(D|G_{x,y})p(G_{x,y}). \quad (1)$$

In our treatment, the prior—$p(G_{x,y})$—is equivalent to the relative visual saliency of location $[x, y]$ within image I, where saliency is some model of where people are *a priori* likely to direct their visual attention and fixation.

One example of how human gaze perception incorporates prior information under conditions of uncertainty is that people show a prior bias that another person's gaze is directed toward them (Ken, 1990; Mareschal, Calder, & Clifford, 2013). This empirical finding makes sense given basic intuitions about human nature. That is, other people's faces (including one's own) would naturally be regions of interest in a counterpart's visual scene (Yarbus, 1967), and even the most mundane face is surely more interesting than, say, the empty space immediately to the left and right of it.

However, it should be clear that indeed *all* of the locations in the counterpart's visual environment (including one's own face) are salient to varying degrees—that is, *a priori* more or less likely to capture the other person's visual attention. Thus, we predict that that prior considerations with respect to presumed visual saliency should, in the general case, factor into human gaze perception.

Methods

Subjects

23 undergraduates at Indiana University received course credit for their participation in the experiment.

Stimuli

Photographs of the "gazer" We took a set of photographs of a young man (the "gazer") seated behind a glass surface.

Figure 1: After the presentation of a fixation cross for 1400 ms, the scene appeared. After 500 ms, a mouse cursor appeared as a red square at a random location within the projected image (this image was a photograph in block 1, and uniform gray in blocks 2-5). The subject indicated with a mouse click where he or she thought the gazer was looking. After the subject clicked, the next trial began. (*Note:* The fixation crosses and red mouse cursors are enlarged in this figure to be more visible.)

In each photograph, the gazer fixated his eyes on a different location on the glass surface, where a grid of points had been marked (later, these marks were digitally removed from the photographs, leaving no observable trace). Though other cues (such as head position) can also be exploited to infer the target of gaze, for this experiment we aimed only to vary the eye cues among these photographs. Therefore, the gazer maintained minimal head and body movement as he fixated on the various locations on the glass surface.

The height of the origin of this grid of points, the camera lens, and the center point between the gazer's eyes was 125 cm. The glass surface was 115 cm from the gazer's face, and the glass surface was 160 cm from the camera. The gazer's face was lit from above, both from the left and right, so as to avoid casting heavy shadows on his face. The photographs were taken with a Canon EOS Digital Rebel XT camera, a 50 mm lens, 1/125 s exposure time, and no flash. The original resolution of these photographs was 3456×2304 pixels.

Thirty-three photographs were used in the experiment, corresponding to when the gazer had been fixating on 32 respective points in a lattice spread over 7 rows and 9 columns (the first row had 5 dots spaced at even 10 cm intervals, the second row had 4 dots with the same spacing but shifted 5 cm, the third row had 5 dots, and so on), plus the origin (i.e. straight ahead, and directly into the camera).

The experiment was presented on a 2560×1440 pixel display. One of the 33 photographs of the gazer appeared in every trial of the experiment, within a 1200×800 pixel window at the center of the display. The unused, background portion of the display (falling outside of the edges of the 1200×800 pixel window) was made gray.

For every trial, a rectangular gray frame (inner dimensions: 550×733 pixels; outer dimensions: 570×753 pixels) was superimposed on the photograph. When the gazer had been photographed, he had always fixated on locations that would have fallen within this gray frame. Either an image (for block 1) or uniform gray (for blocks 2-5) was presented within the rectangular gray frame in each presented scene, and alpha blended (at alpha = 180, where 0 is fully transparent and 255 is fully opaque) with the background photograph of the gazer (see Fig. 1). For the subject, this created a perceptual effect akin to the subject and gazer being on opposite sides of a partially transparent surface, with the gazer's silhouette faintly visible through it. Only a tight ellipse around the gazer's eyes was fully visible through the image, with the area around the eyes smoothly transitioning to greater opacity. Thus, in either condition (projected image, or uniform gray), the gazer's eyes were made fully visible to the subject, and presented simultaneously with the supposed target of his gaze.

Projected images For the first block of trials, images were projected onto the plane upon which the gazer had fixated. The 165 color images (provided by Judd, Ehinger, Durand, & Torralba, 2009) included a wide range of indoor and outdoor scenes, some containing and some not containing people. These images were originally 768×1024 pixels, but were resized to fit the presented 550×733 frame.

Procedure

The experiment consisted of 5 blocks, each consisting of 165 trials. The subject took a 5 minute break after the 3rd block.

Before the first trial of each block, four photographs were displayed in succession, each for 1 s. In these four photographs, the gazer was fixated on four respective locations (marked with 8 × 8 pixel black squares) near the four respective corners of the gazed-upon glass surface. This was a "cal-

ibration" of sorts for the subject, who could get a sense of how the gazer's eyes were positioned when he had been photographed fixating on the extremes of the glass surface.

Each trial began with a black fixation cross, presented at the center of the screen for 1.4 s against a gray background. The subject was then presented with a static scene. Over the course of each block, these scenes featured each of the 33 photographs of the gazer (fixating on 33 respective locations) 5 times, with these 165 total trials being randomly shuffled.

For the first block, one of 165 color images (from the Judd et al., 2009 set) was randomly assigned to each of these 165 trials and projected into the frame in front of the gazer; thus, the projected image and the direction of the subject's gaze in the photograph were randomly paired. For the 2nd-5th blocks, the frame in front of the gazer was filled with a uniform gray.

500 ms after the presentation of this scene, a 10×10 red square appeared at a random location within the frame, and could be controlled by the mouse. After the time when this red cursor appeared, the subject could indicate with a mouse click where, within the frame, he or she believed that the gazer was looking. There was no enforced time limit for this task, and the entire scene remained on the screen until the subject responded. After the subject clicked, the next trial began. The experimental procedure for each trial is illustrated in Figure 1.

Bayesian Model

The *likelihood*: Using eye cues Computational treatments of the problem of discriminating the target of another person's gaze from eye and head cues (e.g., Kim & Ramakrishna, 1999; Hoffman et al., 2006; Yücel et al., 2013; Gao, Harari, Tenenbaum, & Ullman, 2014) often model gaze as a vector or blurry cone emanating from the gazer's face and intersecting with surfaces in the environment. A complete, self-contained algorithm for judging another person's gaze would employ one of these rigorous computer vision approaches in order to compute what we here define as the likelihood function: $L(G_{x,y}|D)$.

We instead derive the likelihood function empirically from each subject's gaze judgments recorded during blocks 2-5 (These were the trials for which the gazer was presented as viewing a uniform gray surface). We associate each photograph of the gazer—associated with the gazer's eyes being fixated in 1 of 33 directions—with a 2-dimensional likelihood function, which we assume to be elliptical (a bivariate Gaussian distribution). This assumption of an elliptical shape makes sense if one imagines a cone of gaze emanating from the gazer's eyes, because the intersection of this cone with the gazed-upon planar surface would be elliptical in shape (indeed, this the geometric definition of an ellipse, one of the basic types of conic section).

After collecting responses from each subject as he or she cycled 20 times through the complete set of 33 eye directions, we estimated the mean (μ) and 2×2 covariance matrix (Σ) of all 33 Gaussian ellipses comprising a complete set of personalized likelihood functions. Each of these probabilistic 2-dimensional likelihood maps was renormalized to sum to 1. For an example of a likelihood map derived for one subject and one of 33 directional cues from the eyes, see Figure 2.

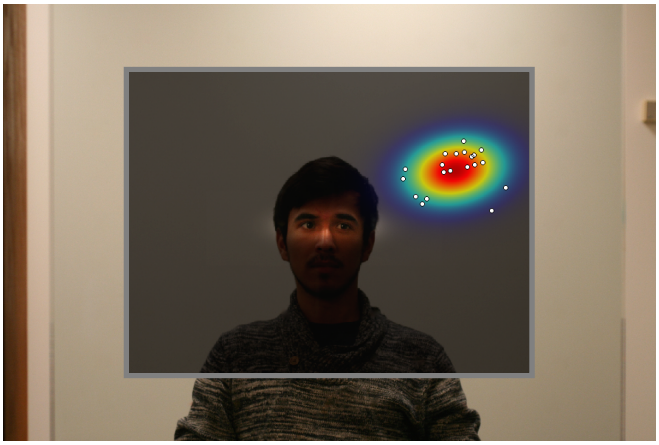

Figure 2: *Likelihood.* Subjects indicated where they thought the person in the photo was looking, within a uniform gray area. The "gazer" was shown fixating on each of 33 target locations within the frame, 20 times per subject. Here, the white dots represent the 20 locations selected by one actual subject (via mouse click) when presented with this same scene. We fit a Gaussian ellipse to these 20 points (superimposed here on the scene), and this ellipse enters into the computational model as the likelihood function with respect to this particular directional cue from the eyes of the gazer.

The *prior*: Using saliency information We hypothesized that it would be expedient for the human visual system to exploit a model of where people are *a priori* likely to look in a scene. Many computer vision models have already been developed to serve precisely this function—predicting where human observers are likely to fixate their visual attention in a given image (e.g., Itti, Koch, & Niebur, 1998; Harel, Koch, & Perona, 2006; Rezazadegan, Rahtu, & Heikkilä, 2011)—and the performance of many of these models has been systematically benchmarked at `saliency.mit.edu`.

We used the saliency model put forth by Judd et al. (2009), because they make freely available a.) MATLAB code for their saliency model, b.) a set of images against which their saliency algorithm has been validated, and against which other algorithms have been tested for comparison, and c.) pre-computed saliency maps corresponding to these images. The Judd et al. algorithm incorporates low-level visual features (e.g., intensity and color contrast), higher-level features (e.g., face detection), and a prior bias toward the center. We selected a subset of 165 images they provide, on the basis that they were all of a consistent size (768×1024 pixels). We incorporate these images' corresponding, pre-computed saliency maps as the *prior* in our Bayesian model of human gaze perception. See Figure 3 for an example of a saliency

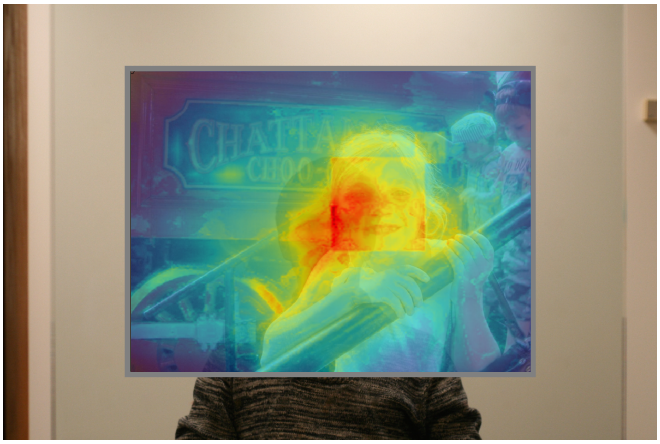

Figure 3: *Prior.* During the first block of the experiment, images were projected into the frame, and subjects indicated where in the picture they thought the person in the photo was looking. Here, we superimpose the saliency map corresponding to this particular image, a continuous 2-dimensional function that enters into the computational model as the prior.

Figure 4: *Posterior.* The posterior probability outputted by the Bayesian model (superimposed here on a screenshot from the experiment) is a multiplication of the likelihood function (given this gaze direction) and prior (given this image). For this particular trial, we present one possible location a subject may have clicked, as a small white bullseye. We assess the model's performance on a given trial as the likelihood of the subject's gaze judgment given the model's posterior prediction map.

map corresponding to one of these 165 images.

The *posterior*: Combining the eye cue with image saliency The posterior distribution outputted by the Bayesian model (given a photograph of the gazer fixating in a particular direction, and a particular gazed-upon image with a corresponding saliency map), is the pixel-by-pixel multiplication of the likelihood map ($p[D|G_{x,y}]$) and saliency map ($p[G_{x,y}]$). After this multiplication, the posterior distribution is renormalized to sum to 1. The resulting map is a hybrid of the two maps giving rise to it. The full Bayesian model indeed exploits the saliency map, but typically only within the neighborhood of locations where the gazer may have plausibly been looking, given the direction of his eyes.

Results

Validation of the likelihood model

Our model of the likelihood function—33 ellipses fit to gaze judgments in blocks 2-5—was first validated for each subject. This approach was cross-validated by fitting ellipses to the subject's responses during three of these blocks, and testing how well they predicted responses on the fourth block. This leave-one-out cross-validation was done each of the four possible ways (leaving each of the four blocks out as the test set). The main diagonals of the covariance matrices Σ of all 33 ellipses were multiplied by one additional parameter, which was optimized for each subject via this same cross-validation procedure. This multiplication procedure increased the variances of the likelihood function ellipses in a manner that increased their predictive power.

The cross-validated performance of the likelihood model was good and remarkably consistent across subjects. For most subjects, multiplying the variances of the fit ellipses by 1.6 proved to be optimal. For only one very atypical subject, we were unable to validate the likelihood model—that is, no parameterization of the likelihood model trained on any three of the subject's blocks was at all predictive of the subject's responses on the remaining test block. We therefore excluded this subject from subsequent analyses.

Model assessment and comparison

We compared the full Bayesian model (outputting a posterior distribution over the image) with a more basic model that only relied on the perceptual signal from the eye cues of the gazer (i.e., the likelihood model, *not* multiplied with the saliency map). We tested the relative performance of these two models in predicting the gaze judgments of subjects during the first block of the experiment (These were the trials in which the gazer was presented with a projected photograph—unlike in blocks 2-5, in which the gazer was presented with a uniform gray surface). Because the likelihood function (a component of both models) was independently validated and optimized for each subject with respect to the four other blocks of the experiment, neither the full Bayesian model nor the reduced model fit any free parameters to the responses of the subjects in the first block.

The relative performance of these two models was first assessed in terms of log likelihood ratio (LLR). For a given trial, the gaze judgment made by the subject had a likelihood given the prediction maps of either model (e.g., as in Fig. 4). Over the subject's 165 trials, the predictions of the two models were compared via their cumulative likelihood ratio. The natural logarithm of this ratio was computed for

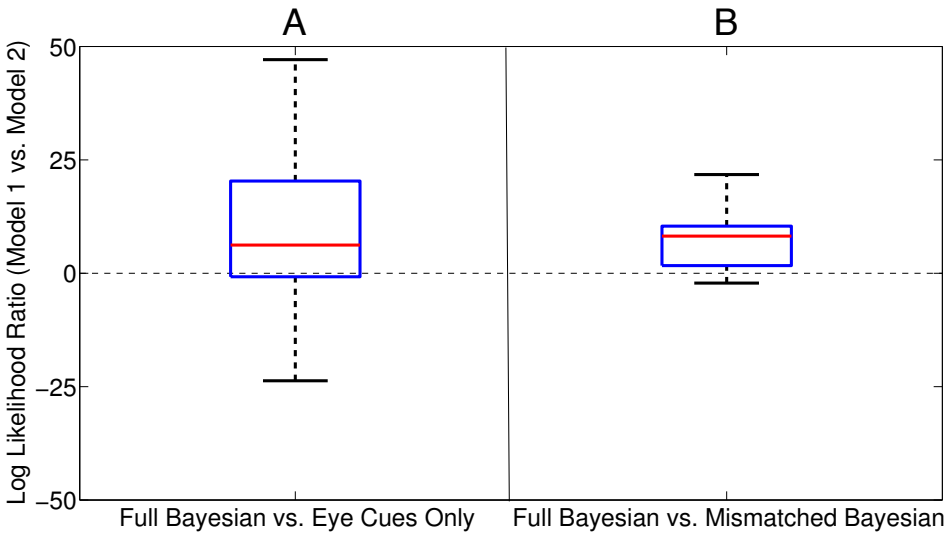

Figure 5: These box plots show the relative performance of the full Bayesian model compared to two other candidate models, assessed for all 22 subjects via log likelihood ratio (LLR). The red line represents the mean LLR of these 22 subjects, the blue box represents the 25th and 75th percentiles, and the black whiskers represent the entire range of LLRs. *Left:* For most subjects, and for the average subject, the full Bayesian model outperforms the reduced model that only relies on eye cues. *Right:* Consistently across subjects, the full Bayesian model outperforms a mismatched Bayesian model, computed with an inappropriate saliency map.

each subject, with positive values favoring the full Bayesian model and negative values favoring the reduced "eye cues only" model. 16 out of 22 subjects' judgments (73%) favored the full Bayesian model (see Fig. 5A), and the cumulative LLR across all subjects very strongly favored the full model (209.5). For each subject, we also calculated the percentage of trials for which the full Bayesian model was a better fit. For 18 out of 22 subjects (82%), the Bayesian model was the better fit on the majority of trials. And for the average subject, the full Bayesian model was preferred on significantly more than half of trials ($M = 61\%$, $SD = 14\%$, $t[21] = 3.69$, $p = .001$).

These data confirmed our hypothesis that subjects would exploit prior information about the relative saliency of locations in the gazed-upon image, *in addition* to using the directional signal computed from the gazer's eye cue.

To provide additional context for the assessment of these two models, we reran the full Bayesian model, but instead of feeding the model the appropriate saliency map corresponding to the gazed-upon image in a given trial, we mismatched each image with a saliency map corresponding to one of the other 164 images in the set. The motivation for the assessment of this mismatched Bayesian model was to examine whether the true Bayesian model had improved the performance of the reduced "eye cues only" model for some superficial reason that was not specific to features of the particular image. If the true Bayesian model were a truly better model of subjects' performance, it would systematically outperform the mismatched Bayesian model. And indeed, the true Bayesian model was a better fit for 18 out of 22 subjects (cumulative LLR = 161.2; see Fig. 5B),

Finally, although most subjects (and the average subject) showed the predicted effect, we by no means wish to gloss over the individual differences we observed (see Fig. 5A). Not all subjects incorporated saliency information into their strategy; indeed, 5 (out of 22) subjects apparently ignored this cue such that the reduced model provided a *far* better fit to their gaze judgments (individual $LLRs = [-23.7, -14.5, -12.8, -11.1, -8.1]$). The full Bayesian model accounted poorly for the performance of these 5 subjects, and fit their judgments hardly any better than a mismatched Bayesian model would have (mean $LLR = 1.5$).

Discussion and Conclusions

In this paper, we developed a Bayesian model for gaze perception, which takes into account both cues from the gazer's eyes and prior saliency information present in the visual scene. Via a quantitative model comparison, we demonstrated that the performance of most subjects is better explained by this full Bayesian model than a reduced model that only takes the eye cues into account. The full Bayesian model also easily outperforms a model that incorporates incorrect saliency information. We consider these data to be strong preliminary support for a Bayesian account of gaze perception, and of closely related social processes like shared attention, gaze following and joint attention.

We emphasize that we do not mean to present this paper as a study of how gaze perception relates to "saliency" (defined

in any one particular way, via a specific algorithm), as a visual feature in itself. Rather, we use computed saliency (according to one algorithmic approach) as a simplified stand-in (that is, a model) for the predictive computation of which locations in a scene are expected to draw another person's visual attention. Most subjects' judgments revealed that they were at least implicitly sensitive to these *a priori* expectations, which were apparently correlated with the output of the saliency model we employed.

The data also appear to indicate that a subset of subjects (20-25%) utilized only the cues from the eyes of the gazer. These individual differences in strategy raise many questions to be addressed in future experiments: Is the tendency to use one strategy over the other relatively stable to the individual? Was this saliency algorithm we used a poor model for where some subjects expect other people will look in the scene? What experimental conditions would favor the use of one strategy over the other?

A Bayesian account of this social perceptual process makes several specific predictions. For example, the noisier the perceptual signal, the more the observer should rely on prior information. This was indeed the result Mareschal, Calder, Dadds, and Clifford (2013) observed in their study of gaze perception; subjects' prior bias toward direct eye contact was modulated by the amount of noise the experimenters added to the observed eyes. We expect that manipulations like this could also be applied to the basic experimental framework presented in this paper, with analogous results. Besides adding noise to the gazer's eyes (e.g., via blurring), one could manipulate the amount of time the observer is given to view the stimulus, the amount of time the observer is given to respond, the size or contrast of the stimulus, or the distance between the gazer and the gazed-upon surface in the scene. The perceptual consequences of each of these manipulations could then be interpreted within the context of this Bayesian treatment, providing additional insight into the nature of human gaze perception.

Acknowledgments

This research was supported by faculty start-up funding at Indiana University. Special thanks to Sebastian Kagemann for his help with the creation of stimuli.

References

Byrne, R., & Whiten, A. (1991). Computation and mindreading in primate tactical deception. In A. White (Ed.), *Natural theories of mind.* Oxford: Basil Blackwell.

Cline, M. G. (1967). The perception of where a person is looking. *The American Journal of Psychology*, *80*(1), 41–50.

Gao, T., Harari, D., Tenenbaum, J., & Ullman, S. (2014). *When computer vision gazes at cognition* (Tech. Rep.). Center for Brains, Minds, & Machines.

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. *Advances in Neural Information Processing Systems*, 545–552.

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.

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*(11), 1254–1259.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *IEEE International Conference on Computer Vision (ICCV).*

Ken, M. (1990). Perception of where a person is looking: Overestimation and underestimation of gaze direction. *Tohoku Psychologica Folia*, *49*, 33–41.

Kim, K., & Ramakrishna, R. S. (1999). Vision-based eye-gaze tracking for human computer interface. In *1999 IEEE international conference on systems, man, and cybernetics* (Vol. 2, pp. 324–329).

Langton, S. R. H. (2000). The mutual influence of gaze and head orientation in the analysis of social attention direction. *The Quarterly Journal of Experimental Psychology*, *53*(3), 825–845.

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception & Psychophysics*, *2*(11), 547–552.

Mareschal, I., Calder, A. J., & Clifford, C. W. G. (2013). Humans have an expectation that gaze is directed toward them. *Current Biology*, *23*, 717–721.

Mareschal, I., Calder, A. J., Dadds, M. R., & Clifford, C. W. G. (2013). Gaze categorization under uncertainty: Psychophysics and modeling. *Journal of Vision*, *13*(5), 1–10.

Rezazadegan, T. H., Rahtu, E., & Heikkilä, J. (2011). Fast and efficient saliency detection using sparse sampling and kernel density estimation. *Image Analysis*, 666-675.

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

Wallaston, W. H. (1824). On the apparent direction of eyes in a portrait. *Philosophical Transactions of the Royal Society of London*, *114*, 247–256.

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

Yücel, Z., Salah, A. A., Meriçli, C., Meriçli, T., Valenti, R., & Gevers, T. (2013). Joint attention by gaze interpolation and saliency. *IEEE Transactions on Cybernetics*, *43*(3), 829–842.