A Survey on the Cognitive Basis of Visual Attention in Real-World Behavior

Sven Bambach

School of Informatics and Computing
Indiana University
Bloomington, IN

sbambach@indiana.edu

September 8, 2013

Abstract

As humans navigate through daily life, their visual systems constantly encounter more stimuli than they can process in real time. To overcome this issue, humans acquire relevant visual information selectively and sequentially using gaze fixations. Understanding the patterns of these fixations is the topic of many cognitive studies on human visual attention. In this paper, we try to give a broad overview about this research, starting with early work on the role of visual conspicuity of attended photographs and show how this research affected follow-up work. We also show more recent gaze studies on non-static, real-world behavior and how they contributed to current research turning away from modeling gaze primarily based on visually conspicuity and towards task-centered models instead.

1 Introduction

The visual world that humans encounter on a daily basis is cluttered with various objects and events which causes their visual systems to be constantly exposed to more stimuli than they can process in real time. During scene perception, high quality visual information is acquired only from a limited spatial region surrounding the center of the gaze, as the visual acuity falls off rapidly from the gaze center into low-resolution visual surroundings. The visual-cognitive system exploits this fact by actively controlling gaze to direct fixations towards important and informative regions. We move our eyes roughly three times per second using rapid movements called "saccades", during which we are effectively blind. Visual information is only acquired during "fixations", i.e. periods of relative gaze stability. Thus, understanding visual attention and scene perception of humans is closely related to understanding gaze control (i.e. eye movements). In a review paper, Henderson [1] listed three main reasons underlining the importance of gaze research: First, "vision is an active process in which the viewer seeks out task-relevant visual information (...) and virtually all animals with developed systems actively control their gaze" [2]. Second, because "eye movements are an overt behavioral manifestation of the allocation of attention in a scene, they serve as a window into the operation of the attentional system." Lastly, "eye movements provide an unobtrusive, sensitive, real time behavioral index of ongoing visual and cognitive processing". Consequently, a lot of the early research on visual attention has been done by studying eye gaze of subjects looking at static images [3].

In this survey paper, we try to give an overview of recent research progress and controversy within the field of studying visual attention. In section 2, we will show that investigators found specific image properties such as high spatial frequency and edge density to be more frequent

around fixated scene patches than unfixated scene patches and how these findings started a trend to build computational models based on these low-level image statistics, coining the term of "visual saliency". In section 3, we report on work criticizing that these models do not generalize well in natural, less-controlled environments. In section 4, we give an overview of recent studies that looked at eye gaze in real-world scenarios such as driving a car or playing table tennis. In section 5, we introduce the idea of task based modeling of gaze allocation and present a recent study that uses a model based on reward maximization and uncertainty reduction in detail. Finally, in section 6 we conclude the paper.

2 Visual Conspicuity and Saliency

2.1 Relation between Gaze and Visual Conspicuity

Based on the idea that early stages of visual processing may exploit the characteristic structure of natural visual stimuli, Reinagel and Zador [4] conducted a study where they recorded eye fixations of human subjects as they viewed static black-and-white images presented on a computer monitor. During the first 4 seconds of image exposure, they extracted fovea-sized square image patches around the subject's center of gaze every 20 ms. They compared these samples with randomly selected image patches in terms of contrast (local standard deviation within the patch) and correlation (spatial frequencies). They found that subjects looked at image regions that had high spatial contrast and regions where intensities of nearby pixels were less correlated with each other than in regions selected at random.

A similar study was conducted by Mannan *et al.* [5] who also looked at the fixations of human observers during a brief (3 s) presentation of natural images. Instead of comparing features directly around the fixations, they computed global maps of contrast, spatial frequency and edge density and compares them to fixation maps in terms of least square errors. They find significant similarities between the locations of eye fixations and those of edge density image features.

2.2 Saliency Models

The findings discussed in the previous section gave rise to the emergence of computational models that aimed to explain attentional capture based on visual conspicuity. The most well-known of these models is that of Itti and Koch [6, 7]. They computed feature maps based on features that they considerd biologically plausible, which are intensity (luminance), color (more precisely specific differences between color channels based on existing chromatic opponencies in the visual cortex [8]) and orientation (based on Gabor Filter responses at different orientations). Each feature is extracted at different spatial scales which are normalized and combined into three conspicuity maps. These maps are combined again into one, global "saliency map" which has the purpose of representing the conspicuity or "saliency" at every location in the visual field by a scalar quantity. They tested their model based on two different visual search tasks, using both synthetic images and real images. The synthetic images, based on classic work by Treisman and Gelade [9], always contained a target (a rectangular bar) and distractors (other bars) that differ from the target in various ways (such as color or orientation). In the natural images, the target was, for example, an army vehicle (which was very small in comparison to the image size) hidden in a forest environment. They compared the total time and number of fixations it took for humans to find the target with their own model. To simulate gaze fixations based on their saliency map, they employed a winner-takes-all neural network in combination with an inhibition of return policy. Basically, the most salient point in the map is chosen as the first fixation, then (after about 30-70 ms) the map is recalculated with the area of the first fixation being oppressed, and so on. They showed that, in the case of the synthetic images, their model can successfully reproduce the search behavior of humans as presented in the work of [9] and is also able to find search targets in natural images.

Other researchers followed up with further empirical evaluations of the saliency model using complex, natural images. Parkhurst *et al.* [10] found a significantly greater correlation between computed stimulus saliency and fixation locations than that expected by chance alone and also noticed that this correlation was greatest for eye movements that immediately followed the stimulus beginning.

Recently, Foulsham and Underwood [11] conducted an

experiment that compared model-generated and experimental eye movements, both in terms of the spatial location of individual fixations and their sequential order (the "scan path"). They had subjects look at the same image twice, once for memorization (encoding) and then again for recognizing. They did this to investigate possible biases resulting from the "scan path theory" [12] which argues that eye movements are generated top-down, in particular in response to a previously seen image. They found that there was a tendency for fixations to target the most salient regions, as selected by the Itti model, which could not be explained by simpler models that were just biased towards central distributions. Also, the connection between saliency and fixations did not vary signficantly with the demands of the task (encoding or recognizing). In terms of scan paths, they found that they were most similar when compared between two viewings of the same image by the same person (memorization and recognition) and more importantly, the saliency model-predicted scanpaths were not highly similar to human scanpaths.

2.3 Evidence against Saliency

While it is clear that the studies mentioned in the previous section provide a proof of principle that the visual system can select fixation targets based on visual conspicuity, many researchers claim that correlations between fixations and saliency alone should not be taken to imply any causal link between features and fixation locations.

In a recent paper, Henderson *et al.* explicitly reject the hypothesis that fixation locations during search tasks in real-world scenes are primarily determined by visual saliency [13]. In their study, they showed participants photographs of real-world outdoor scenes and engaged them in a visual search task by asking them to count the number of people who appeared in each scene and recorded eye movements. They evaluated their results in three ways. First, they found that the Itti model performed poorly in terms of predicting gaze fixations. Second, they examined whether image properties differ on fixated and non-fixated locations and found clear differences in contrast, intensity and edge density, which is consistent with the studies mentioned section 2.1. However, they also compared the "semantic informativeness" between fixated and non-fixated locations. They did this by showing 300 selected patches (random between fixated and non-fixated locations) to a group of seven people and had them rate how well they thought they could determine the overall content of the scene based on the patch. They found that attended locations are more informative than random locations and thus argue that any observed correlations between fixation locations and image statistics may be due to the informativeness of the fixated locations rather than differences in the image statistics themselves.

Einhäuser *et al.* [14] follow the alternative hypothesis that observers attend to "interesting" objects when looking at photographs of natural scenes. In their study, subjects observed various pictures in two different scenarios. In one scenario, they were asked to rate the image on how interesting it was, pretending to be a "judge for an art competition". In the other scenario, they were to decide whether a named search target was in the scene or not. For both scenarios and after being exposed to the picture, subjects were asked to name up to five keywords to describe the scene. They generated an "object map" to compete with the saliency map, where they counted the number of objects (based on the keywords for an image i named by the observer) overlapping with pixel (x, y) in image i, and normalized this over all observers, yielding an object map O_i. They compared object maps and saliency maps in terms of how well they predict eye gaze using ROC curves and found that the object maps beat the saliency maps in a significant fraction of the images (57:36). They also combined both maps and found that performance is indistinguishable from the object map alone, concluding that saliency contributes little to fixation prediction once objects are known. More interestingly, they assigned each object a relative "total object saliency", defined as the sum of saliency map values over the object divided by the sum across the whole image. They found that object saliency does predict how frequently an object is recalled, meaning that the objects mentioned the most when observers were asked to describe the scene where those with the highest object saliency. They concluded that saliency maps might predict fixations indirectly, as objects tend to be more salient than the background, rather than because fixations depend directly on saliency.

3 Generalization to the Real-World

The notion of visual saliency (based on the model of Itti *et al.* [6, 7]) became quite popular, even among studies that look at more complex behavior and turn away from static image viewing. However, there is a lot of discussion about how well saliency can generalize and how much explanatory power it has when looking at visual attention in real world behavior. In a recent review paper, Tatler *et al.* [15] argue that "we need to move away from this class of model and find principles that govern gaze allocation in a broader range of settings". They emphasize that the original intention of Itti was not to make a model that predicts eye movements in complex scenes, but rather explain attentional capture, and they just so happened to (reasonably) evaluate their model in terms of eye fixations. Tatler *et al.* go on to summarize several general issues that need to be considered when generalizing attention models from controlled, static, lab settings to free, real-world behavior, which we describe in the following two sections.

3.1 The Picture-viewing Paradigm

Problems naturally arise from the clear physical differences between photographs and real environments, with photographs having a smaller dynamic range and lacking depth cues (both motion and stereo parallax). Additionally, there is a reliable bias that photographs of scenes tend to have higher saliency towards the center, both caused by the tendency of photographers to center the subjects of interest, but also due to physical qualities (e.g. sky in the upper visual field and ground pane in the lower field). More interestingly, Tatler *et al.* [16] found a strong tendency towards subjects making early fixations near the center of an image irrespectively of the scene's content. As picture-viewing experiments often take the form of a sudden onset of an image, followed by a few seconds of image viewing and then a sudden offset of the image, they argue that this effect my account for a lot of the success attributed to saliency and that, in fact, the sudden onset itself may influence the inspection behavior. In contrast, the key goal of vision in natural behavior may be the extraction of visual information required to complete an involved task and artistic biases or effects caused by the sudden presentation of a scene may not play a role at all.

Another issue is the assumption that saccades precisely target the locations, i.e. information at the center of the gaze contains the intended target of each saccade. Under static lab conditions, this assumption is evidenced by the observation of small, corrective saccades that occur when saccades with a large amplitude originally "miss" the target [17]. However, in the context of more natural tasks, such precision may be unnecessary. For example, Johansson *et al.* [18] conducted a study where they looked at hand-eye coordination during object manipulation, where they had subjects reach for and grasp a bar and then move it past obstacles to press a target switch. They found that, when moving the bar past obstacles, saccades that got the center of vision within 3 degrees of the obstacle were sufficient and were not corrected.

3.2 Videos as an Approximation of Real-World Settings

A growing number of studies are starting to use videos to overcome some of the problems mentioned in the previous section and because video allows the evaluation of dynamic temporal features.

Itti extended his original model [6, 7] with motion cues and conducted a study where he had eight subjects look at a heterogeneous collection of 50 video clips, including television ads, music videos, sport videos and clips from video games [19]. In addition to existing low-level features (color, intensity, orientation), they added temporal flicker features as well as four oriented motion energies. Temporal flicker essentially describes the pixel-wise intensity contrast between consecutive frames and motion energies describe intensity contrasts resulting from shifting the frame one pixel to the left/right/top/bottom. Considering the temporal nature of the data, they measured the saliency around the location of the future endpoint of a saccade at the moment when that saccade began and compared results which uniformly distributed random locations. They found that motion and temporal change were stronger predictors of human saccades than color, intensity and orientations features, with the best predictor being all features combined.

While Itti's work shows that dynamic temporal features can be predictors of eye movements, other more recent work questions how well the tested videos generalize to natural behaviors. Hirose *et al.* looked at the effects of

editorial cuts, as they are common in many movie-like sequences [20]. They created a scene where the camera follows an actor walking through a room and passing a desk with various items on it, followed by a cut showing the desk in more detail. They changed different object categories (color of an item on the desk, spatial arrangement of items, etc.) between the cut and examined eye movement behavior and recognition memory. They found that there are discrepancies between eye movement (oculomotor) behavior and memorial behavior in comparison to normal scene perception, with subjects usually recalling object properties as seen from the most recent camera viewpoint. Furthermore, irregularities in the spatial arrangement of items between cuts were significantly less noticed than changes in object color or type, suggesting that spatial information is represented differently from other object qualities when looking at videos with cuts between different viewpoints.

Dorr *et al.* [21] compared the variability of eye movements when viewing videos of dynamic natural scenes with Hollywood-style movie trailers containing cuts and other artistic influences. They found that gaze patterns while viewing professionally cut Hollywood trailers were very different from natural movies. In particular, eye movements during movie trailers were much more similar among different subjects and contained a much stronger bias for the center of the screen. The authors took this to explicitly highlight the importance of studying vision under naturalistic conditions, as eye movements are presumably optimized to deal with natural scenes.

4 Real-World Behavior Studies

In this section, we present examples of studies that looked at visual attention during real-world scenarios, meaning scenarios that are not constrained to static photo or video viewing, but involve subjects engaged in free head and body movements. These studies are becoming more and more popular, mostly driven by technological advances that make head-mounted eye-tracking gear practicable. This gear (e.g. *Tobii* or *SMI*) usually consists of a lightweight, head-mounted, egocentric camera approximating the visual field of the person wearing it, in combination with a second, infrared camera that keeps track of the person's pupil.

We will first look at studies with tasks specifically designed to observe the coordination of eyes, head and hands and then at studies that document eye gaze in real-world activities.

4.1 Coordination of Eyes, Head and Hands

Pelz *et al.* looked at the temporal coordination of eye, head and hand movements while subjects performed a simple block-moving task, where they were exposed to three *lego* boards [22]. One board contained at target model consisting of eight blocks of different color in a specific spatial configuration. A second board contained a set of 12 "resource blocks". Finally, the third board was empty. The subjects were told to recreate the target model from the first board on the third, empty board as fast as possible, using the blocks from the resource board. Thus, the task involved fixations to gather information about the blocks and visually guided hand movements to manipulate them. They found "rhythmic patterns of eye, head and hand movements in a fixed temporal sequence or coordinative structure." They further observed that hand movements towards a block were delayed until the eyes were available for guidance and that head movements were the most flexible among subjects and frequently diverged from gaze change, appearing instead to be linked to the hand trajectories. They concluded that the coordination of eye, hand, head and gaze changes follows a synergistic binding rather than an obligatory one and that these temporary synergies simplify the coordination problem by reducing the number of control variables.

As briefly mentioned in section 3.1, Johansson *et al.* [18] analyzed the coordination between gaze behavior and fingertip movements while subjects manipulated a test object. They built an apparatus which let them track finger movements and eye gaze patterns while subjects reached for a bar, moved it past obstacles to touch a target switch, and finally moved the bar back towards its original position. They observed that subjects almost exclusively fixated "landmarks" which were critical for the control of the task. Those landmarks involved contact points between fingers and bar, the target switch, as well as the obstacle. However, subjects never fixated on the hand or the moving bar. Instead, gaze was temporally leading hand and bar movements, such that the instant that gaze exited a given landmark coincided with a kinematic event

at the landmark. They concluded that gaze supports hand movement planning by marking key positions to which the hand is subsequently directed and that the saliency (in a more general way) of gaze targets arises from the functional requirements of the task.

4.2 Real-World Activities

In this section, we present examples from studies that looked at eye movements and how they contribute in the organization of real-life activities such as walking, driving, ball sports or common indoor activities like making tea. We also want to point out that an extensive, in-depth review in this area was recently done by Land [23].

4.2.1 Walking

Patler and Vickers studied how far people look ahead when walking across a "difficult" terrain [24]. To do so, they instructed subjects to step on a series of footprints that were regularly and irregularly spaced over a 10 meter distance. They discovered two main types of gaze fixations: footprint fixation; and travel fixation where gaze is stable and thus "traveling" with the speed of the body. They found that when participants fixated on the footprint in front of them, on average, they looked two steps ahead and did so about 800-1,000 ms before stepping on the target area. They concluded that this would allow them sufficient time to successfully modify their gait patterns. They further found that most of the travel time (over 50%) was spent on travel fixations and hypothesized that this behavior facilitates the acquisition of both environmental and self-motion information from the optical flow generated by the self-motion.

Hollands *et al.* looked at the role of eye and head movements when subjects changed direction while walking [25]. They had subjects walk an even path and gave either predefined spots to change directions or gave cues by the onset of a light in the direction of the new path. They found that every turn was accompanied by a saccade to the new direction in combination with a head movement, such that eye-head combination brought head and gaze into line with the new direction as the body turn was being made. They hypothesized that the pre-alignment of the head provides a new reference frame that can be utilized to control the rest of the body.

4.2.2 Driving and Steering

Land and Lee [26] researched where people look when they steer a car by simultaneously recording the driver's gaze direction and the steering-wheel angle. They used a tortuous, one-way street to ensure the need of planned steering while avoiding distractions of other traffic as far as possible. They found that drivers spent most of the time looking at what they call "tangent point", meaning the point inside a bend that causes the driver's line of sight to be tangential to the inner road edge. As a result of this, there is a clear correlation between gaze angle and steering angle, where the peak of the cross-correlation was found at a delay of about 0.8 s between changes in gaze angle and changes in steering angle.

4.2.3 Ball sports

Land and Furneaux looked, among other activities, at the gaze patterns of ordinary people playing table tennis [27]. They found that players generally do keep their "eyes on the ball". However, in crucial moments, they perform anticipatory saccades to where they expect the ball to be. This mainly applies to the ball bouncing off the table on either side, where players roughly fixate the location at where the ball will bounce about 400 ms in advance. They conclude that the reason players anticipate the bounce is that location and timing of the bounce are crucial in the formulation of the return shot. More importantly, they claim that, until the bounce, the trajectory of the ball as seen by the receiver is ambiguous as stereopsis processing may not be fast enough to contribute useful depth signals.

In a somewhat similar study, Land and McLeod looked at cricket [28]. Here, one player (the bowler) throws the ball towards the other player (the batsman) in a way that it bounces off the ground exactly one time before the batsman tries to hit it. They found that the "batsmen's eye movements monitor the moment the ball is released, make a predictive saccade to the place where they expect it to hit the ground, wait for it to bounce, and follow its trajectory for 100-200 ms after the bounce." Similar to table tennis, they concluded that the information the batsman needs to judge where and when the ball will reach his bat is mostly given by time and place of the bounce. Worse players had a higher latency for their predictive saccade and consequently missed balls that were too fast.

4.2.4 Indoor everyday activities

Land *et al.* [29] looked at eye gaze data of three subjects during the task of making tea in a kitchen environment. The task involved picking up the kettle, walking across the kitchen to the sink, filling the kettle with water, walking back and so on. They found a lot of similarities among subjects both in the scan path and the number of fixations dedicated to each subtask. They further conclude that saccades are made almost exclusively towards objects that are directly involved in the subtask despite the presence of other visually salient objects. Also, the eyes deal with one object at a time, which may involve a number of fixations on different parts of the objects but no alternating fixations between objects.

Similar results in terms of almost all fixations targeting task-relevant objects were found by Land and Hayhoe [30], who looked at subjects preparing sandwiches while sitting on a table. They also found that eyes usually reached the next temporally relevant object before the first sign of manipulative action, indicating that eye movements lead motor actions.

5 Modeling the Role of Task in Gaze Control

The results mentioned in the previous section show that visual attention in real-world behavior is strongly task-related and cannot be properly modeled in terms of visual conspicuity clues which are the basis of bottom-up models such as Itti's model of visual saliency [6, 7]. Consequently, a lot of researchers such as Tatler *et al.* explicitly ask for a "reinterpretation of salience" [15]. Indeed, a lot of attempts to "generalize" static saliency models are based on modifying the existing saliency frameworks in a way such that the core bottom-up mode of looking is modified by various high-level constraints. For instance, Torralba *et al.* introduce a model based on a Bayesian framework that combines saliency with global, scene-centered features [31]. Essentially, their model extends the Itti model with semantic priors to improve search task results in static, real world scenes. When looking at a kitchen scene and searching for a mug, the model favors regions around the countertop while searching for a painting favors regions along the walls. However, Tatler *et al.* argue that the assumptions at the heart of such studies are still problematic as the principles that might be expected from picture-viewing studies simply do not match the principles that are found across many instances of natural vision. Taking the table tennis study [27] as a prominent example, it is indubitable that the regions that subjects fixated on in anticipation of the ball have absolutely no difference in visual saliency compared to their surroundings, but are only fixated because of their relevance for the task of hitting the ball. Thus, Tatler, Ballard *et al.* [15, 32] stress the importance of future research on frameworks that model the role of task in the control of gaze.

5.1 Reward-based models

One proposed way of modeling tasks are reward-based systems of reinforcement learning, where visual information acquired during fixations can be thought of as a "secondary reward which can mediate learning of gaze patterns by virtue of its ultimate significance for adaptions and survival" [15]. This idea is also supported by recent neuroscientific findings that show that the brain's internal reward mechanisms are closely linked to the neural machinery controlling eye movements and that saccade-related areas in the cortex exhibit sensitivity to reward [33, 34].

Gaze models that use reward as central components are rare and current research is far from a state where computational models can explain eye movements across multiple instances of natural behavior. However, there are successful studies in this area with a prominent example being the work of Sprague, Ballard and Robinson [35]. For their study, they developed a virtual reality graphics environment, including a simulated human agent named "Walter". Walter successfully learned to allocate gaze to avoid obstacles and control his walking direction. His behavior was based on three microbehaviors (collision avoidance, sidewalk navigation, litter collection), each of which was linked to a visual routine that created the corresponding state information. Litter was signaled by purple objects, which had to be "picked up" (by colliding with them), so potential litter had to be isolated as being of the right color and also nearby. While real humans would use stereo, parallax depth, etc. to obtain depth information, the model directly sampled depth from the scene graph. Sidewalk navigation uses color information to label pixels

that border both sidewalk and grass regions and fit a line to estimate the edge of the sidewalk. Finally, collision detection worked similar to litter collection, with obstacles being blue objects. Each of those three tasks/behaviors is associated with a reward value. Their model assumes that only one location (similar to the human viewpoint) can be attended at a time and that the uncertainty about unattended tasks grows over time. The decision about which task to attend is based on the expected reward of switching attention and is evaluated every 300 ms, approximating 3 saccades/fixations per second. Movements are made to maximize reward by reducing the uncertainty that could arise as a result of suboptimal actions. To evaluate Walter's behavior, they introduced human subjects to the virtual environment by having them wear head-mounted binocular displays that contained eye tracking capabilities. They found similar patterns between Walter and humans in the sense that the relative proportion of fixations on locations relevant to each of the three tasks was the same. One discrepancy was that humans used fewer sidewalk fixations than suggested by the model. The authors explanation for this is that humans, unlike the model, where able to use litter/obstacle fixations for two routines. As both litter and obstacles were only located within the sidewalk, seeing an unobstructed litter implies that you can walk towards it knowing you are still on the sidewalk.

6 Conclusion

We have shown that early investigations of visual attention have been largely driven by studies of static picture viewing. Different image statistics between visually attended patches and control patches lead to the development of models using visual conspicuity to predict human gaze, most notably the visual saliency model of Itti *et al.* [6, 7]. However, a lot of researchers claim that correlation between gaze patterns and image conspicuity does not imply causality and that the predictiveness of saliency, even in static scene-viewing scenarios is very task-dependent (e.g. search versus recall). Some researchers, such as Handerson, even go as far as to say that visual saliency does not account for eye movements at all [13].

We also introduced studies that looked at eye gaze during real-life activities, where subjects were allowed to move freely while following activities such as walking, driving or making tea. All of these studies have in common that subjects seemed to focus largely on locations that contained crucial information for the task at hand with gaze taking a "planning role" for hand manipulation and other body movements. On the other hand, visual saliency did not seem to contribute a lot towards the decision of which visual information to attend to.

Consequently, researchers (Tatler, Hayhoe, Land, Ballard [15, 32]) started criticizing the trend of using models of low level saliency that are modified by high level constraints and argue for future research to focus on task-centered models on the basis of reward maximization and uncertainty reduction.

We largely agree with this idea, while emphasizing that developing gaze allocation models that can generalize across the countless instances of natural behavior is a verity difficult goal. We hypothesize that this problem played a large roll in the fame of bottom-up models, as one thing that all real-world activities have in common is the presence of low level visual stimuli. However, there is overwhelming evidence that visual saliency simply plays little to no role during visual attention in the real world. Recent success shows that gaze behavior in a free-moving, real world task can adequately be modeled with relatively simple, reward based models [35]. Additionally, vision researchers do not necessarily have to reinvent the wheel and may benefit from existing work on reinforcement learning and reward based models that have been intensively studied in the robotics and artificial intelligence community.

References

[1] J. M. Henderson, "Human gaze control during real-world scene perception," *Trends in cognitive sciences*, vol. 7, no. 11, pp. 498–504, 2003.

[2] M. F. Land, "Motion and vision: why animals move their eyes," *Journal of Comparative Physiology A*, vol. 185, no. 4, pp. 341–352, 1999.

[3] A. L. Yarbus, B. Haigh, and L. A. Rigss, *Eye movements and vision*. Plenum press New York, 1967, vol. 2, no. 5.10.

[4] P. Reinagel and A. M. Zador, "Natural scene statistics at the centre of gaze," *Network: Computation in Neural Systems*, vol. 10, no. 4, pp. 341–350, 1999.

[5] S. K. Mannan, K. H. Ruddock, and D. S. Wooding, "The relationship between the locations of spatial features and those of fixations

made during visual examination of briefly presented images," *Spatial vision*, vol. 10, no. 3, pp. 165–188, 1996.

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

[7] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision research*, vol. 40, no. 10, pp. 1489–1506, 2000.

[8] S. Engel, X. Zhang, and B. Wandell, "Colour tuning in human visual cortex measured with functional magnetic resonance imaging," *Nature*, vol. 388, no. 6637, pp. 68–71, 1997.

[9] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[10] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vision research*, vol. 42, no. 1, pp. 107–123, 2002.

[11] T. Foulsham and G. Underwood, "What can saliency models predict about eye movements? spatial and sequential aspects of fixations during encoding and recognition," *Journal of Vision*, vol. 8, no. 2, 2008.

[12] D. Noton and L. Stark, "Scanpaths in saccadic eye movements while viewing and recognizing patterns," *Vision research*, vol. 11, no. 9, pp. 929–IN8, 1971.

[13] J. M. Henderson, J. R. Brockmole, M. S. Castelhano, M. Mack, M. Fischer, W. Murray, and R. Hill, "Visual saliency does not account for eye movements during visual search in real-world scenes," *Eye movements: A window on mind and brain*, pp. 537–562, 2007.

[14] W. Einhäuser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," *Journal of Vision*, vol. 8, no. 14, 2008.

[15] B. W. Tatler, M. M. Hayhoe, M. F. Land, and D. H. Ballard, "Eye guidance in natural vision: Reinterpreting salience," *Journal of vision*, vol. 11, no. 5, 2011.

[16] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: Effects of scale and time," *Vision research*, vol. 45, no. 5, pp. 643–659, 2005.

[17] B. W. Tatler and B. T. Vincent, "Systematic tendencies in scene viewing," *Journal of Eye Movement Research*, vol. 2, no. 2, pp. 1–18, 2008.

[18] R. S. Johansson, G. Westling, A. Bäckström, and J. R. Flanagan, "Eye–hand coordination in object manipulation," *the Journal of Neuroscience*, vol. 21, no. 17, pp. 6917–6932, 2001.

[19] L. Itti, "Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes," *Visual Cognition*, vol. 12, no. 6, pp. 1093–1123, 2005.

[20] Y. Hirose, A. Kennedy, and B. W. Tatler, "Perception and memory across viewpoint changes in moving images," *Journal of Vision*, vol. 10, no. 4, 2010.

[21] M. Dorr, T. Martinetz, K. R. Gegenfurtner, and E. Barth, "Variability of eye movements when viewing dynamic natural scenes," *Journal of vision*, vol. 10, no. 10, 2010.

[22] J. Pelz, M. Hayhoe, and R. Loeber, "The coordination of eye, head, and hand movements in a natural task," *Experimental Brain Research*, vol. 139, no. 3, pp. 266–277, 2001.

[23] M. F. Land, "Eye movements and the control of actions in everyday life," *Progress in retinal and eye research*, vol. 25, no. 3, pp. 296–324, 2006.

[24] A. E. Patla and J. N. Vickers, "How far ahead do we look when required to step on specific locations in the travel path during locomotion?" *Experimental Brain Research*, vol. 148, no. 1, pp. 133–138, 2003.

[25] M. A. Hollands, A. Patla, and J. Vickers, "look where youre going!: gaze behaviour associated with maintaining and changing the direction of locomotion," *Experimental Brain Research*, vol. 143, no. 2, pp. 221–230, 2002.

[26] M. F. Land and D. N. Lee, "Where do we look when we steer." *Nature*, 1994.

[27] M. F. Land and S. Furneaux, "The knowledge base of the oculomotor system," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 352, no. 1358, pp. 1231–1239, 1997.

[28] M. F. Land and P. McLeod, "From eye movements to actions: how batsmen hit the ball," *Nature neuroscience*, vol. 3, no. 12, pp. 1340–1345, 2000.

[29] M. Land, N. Mennie, J. Rusted *et al.*, "The roles of vision and eye movements in the control of activities of daily living," *PERCEPTION-LONDON-*, vol. 28, no. 11, pp. 1311–1328, 1999.

[30] M. F. Land and M. Hayhoe, "In what ways do eye movements contribute to everyday activities?" *Vision research*, vol. 41, no. 25, pp. 3559–3565, 2001.

[31] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search." *Psychological review*, vol. 113, no. 4, p. 766, 2006.

[32] D. H. Ballard and M. M. Hayhoe, "Modelling the role of task in the control of gaze," *Visual Cognition*, vol. 17, no. 6-7, pp. 1185–1204, 2009.

[33] W. Schultz, L. Tremblay, and J. R. Hollerman, "Reward processing in primate orbitofrontal cortex and basal ganglia," *Cerebral Cortex*, vol. 10, no. 3, pp. 272–283, 2000.

[34] M. C. Dorris and P. W. Glimcher, "Activity in posterior parietal cortex is correlated with the relative subjective desirability of action," *Neuron*, vol. 44, no. 2, pp. 365–378, 2004.

[35] N. Sprague, D. Ballard, and A. Robinson, "Modeling embodied visual behaviors," *ACM Transactions on Applied Perception (TAP)*, vol. 4, no. 2, p. 11, 2007.