Human Visual Search as a Deep Reinforcement Learning Solution to a POMDP

Aditya Acharya (axa384@student.bham.ac.uk)
School of Computer Science
University of Birmingham

Xiuli Chen (x.chen.1@bham.ac.uk)
School of Psychology
University of Birmingham

Christopher W. Myers (Christopher.Myers2@wpafb.af.mil)
Air Force Research Laboratory
Performance and Learning Models Branch, USA

Richard L. Lewis (rickl@umich.edu)
Department of Psychology
University of Michigan,Ann Arbor, USA

Andrew Howes (HowesA@bham.ac.uk)
School of Computer Science
University of Birmingham

Abstract

When people search for a target in a novel image they often make use of eye movements to bring the relatively high acuity fovea to bear on areas of interest. The strategies that control these eye movements for visual search have been of substantial scientific interest. In the current article we report a new computational model that shows how strategies for visual search are an emergent consequence of perceptual/motor constraints and approximately optimal strategies. The model solves a Partially Observable Markov Decision Process (POMDP) using deep Q-learning to acquire strategies that optimise the trade-off between speed and accuracy. Results are reported for the Distractor-ratio task.

Keywords: Computational Rationality; Deep Reinforcement Learning; Deep Q-Learning; Visual Attention.

Introduction

One of the many tasks for which people use vision is to search for items in the environment. Visual search might be used to locate a phone on a table, a car in a parking lot or a family member in a crowd. In a typical laboratory *visual search task*, participants are asked to find a visual target amongst distractors. For example, searching for a Gabor patch in a high contrast noisy background (Najemnik & Geisler, 2008), or searching for a red coloured letter O in a display that consists of red Xs and green Os (Shen, Reingold, & Pomplun, 2000). Many, though not all, visual search tasks require a number of fixations and saccades before the target is found.

From a cognitive science perspective, visual search is interesting because data from visual search experiments can be used to inform theories of the underlying constraints on vision (e.g (Geisler, 2011) and also to inform theories of how people adapt eye movement strategies to these constraints (e.g (Najemnik & Geisler, 2005). Human behaviour is a consequence of both the constraints and the adapted strategies and explanations of behaviour require both (Lewis, Howes, & Singh, 2014). In fact, there is a long history of cognitive science research on visual search and there are a number of competing theoretical approaches.

First are the *map-based* approaches described by (Kowler, 2011), such as salience maps (Itti & Koch, 2000) and activation maps (Pomplun, Reingold, & Shen, 2003; Wolfe, 2007), where the perceived visual information is represented as a topological distribution in a graphical map form. The salient area or peaks in the map represent items that significantly differ from their neighbouring items, that may contain attributes of interest. These peaks in the map are then used to guide the eyes through the display using some selection rules, such as a greedy heuristic (Pomplun et al., 2003) or a winner-take-all heuristic (Itti & Koch, 2000). To summarize, the map based approach assumes that saccades are programmed to move the fovea to those areas in the display that stand out from surroundings.

Second are the Bayes *optimal state estimation* approaches (Myers, Lewis, & Howes, 2013; Najemnik & Geisler, 2008), in which it is assumed that visual information is recorded as a Bayesian estimate of the state of the world. On each fixation the estimated state is updated by optimally integrating information (Bayes rule) from the previous state and from the fovea and from the periphery according to its reliability. The eye movements are then made using these states and applying a heuristic decision rule (e.g., 'Maximum A Posteriori' (MAP)) to navigate. This rule generates a behaviour in which attention is directed to areas which have the highest probability of target present. Alternatively, Najemnik and Geisler (2005) observed that the number, and spatial distribution, of saccades could be better explained by a model in which each saccade was directed to an 'ideal' location (i.e., a location that maximises information gained). Their model was sensi-

tive to known human constraints on vision, i.e., the accuracy of perceiving a feature degrades with eccentricity.

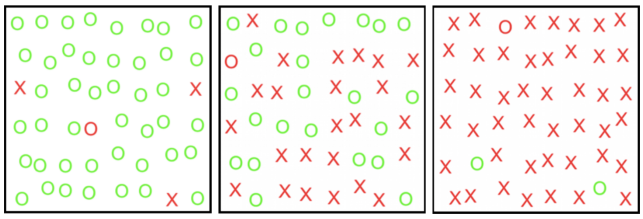
Third are the *optimal control* approaches (Butko & Movellan, 2008; Hayhoe & Ballard, 2014; Nunez-Varela & Wyatt, 2013; Sprague, Ballard, & Robinson, 2007), in which it is assumed that the eye movements are not made to estimate some statistics about the world but rather the goal is to maximize the overall performance utility. The maximum reward/utility an individual can attain throughout the task is bounded by the noisy encoding of the visual information by the human brain. In contrast to map-based and optimal state estimation approaches, where prior assumptions about eye movement decisions are made by heuristic rules, the control strategy emerges as a consequence of bounds imposed by the human visual system. To summarize, the optimal control approach assumes that the saccades are programmed to move the fovea so as to maximise task utility/reward.

In the current article we report a novel (approximately) optimal control model of the distractor ratio task. The purpose of this model is to (1) explain phenomena not previously explained as optimal control, (2) to further elucidate the framing of visual search as a Partially Observable Markov Decision Process (POMDP) (Kaelbling, Littman, & Cassandra, 1998), and (3) to explore the role of deep Q-learning (Mnih et al., 2015) in solving the tractability problems with previous optimal state estimation and optimal control approaches. The model goes beyond the optimal state estimation model of Myers in that it is applied to the full display size used by (Shen, Reingold, & Pomplun, 2003). The model uses deep Q-learning to solve a POMDP. It attempts to maximise a reward signal given constraints imposed by the human visual information processing system. We compare the performance of the optimal control model to a model that uses MAP-like heuristics. We show that the optimal control model offers higher utility and better fits to the human data than the heuristic model. Lastly, we use the model to explain phenomena associated with the distractor ratio paradigm (Bacon & Egeth, 1997; Shen et al., 2000; Zohary & Hochstein, 1989). A phenomena that has previously been explained using the salience-map based approach.

The Distractor Ratio Task

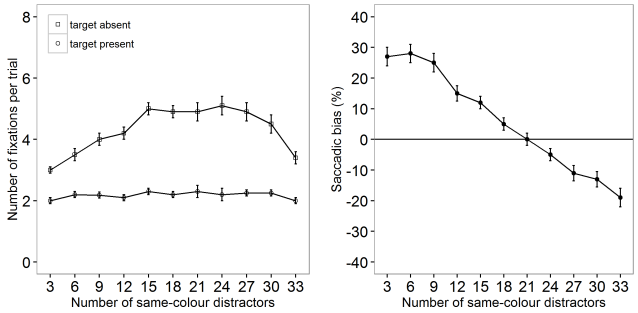
In the distractor ratio task the display consists of a target object, which is randomly positioned amongst distractor objects each of which shares at least one common feature with the target. The goal is to respond whether the target is present or absent. An example display is shown in Figure 1 where the target is a red letter O. The distractors in this display share either a same-colour or same-shape feature with the target.

In a number of studies it has been observed that people respond more quickly, and with fewer eye movements, for extreme ratios of same colour to same shape distractors (Egeth, Virzi, & Garbart, 1984; Shen et al., 2003). In Figure 1, the target – a red letter O – can be located easily in display (a) and (c) with ratios 3:45 and 46:2 respectively as compared to display (b), for which a response takes a relatively long time. The distractor ratio effect reported by Shen et al. (2003) is shown in Figure 2.



(a) (b) (c)

Figure 1: Distractor ratio stimuli with ratio distributions: (a) 3:45, (b) 24:24, (c) 46:2 and target stimuli: red coloured letter O.



(a) (b)

Figure 2: (a) Average number of fixations per trial as a function of the number of distractors sharing colour with the search target in target-absent trials and target-present trials for high discriminability condition. (b) Saccadic bias (the difference between the observed frequency and chance performance) as a function of the number of same-colour distractors in target- absent trials for high discriminability condition (Shen et al., 2003)

In addition to the distractor-ratio effect, Shen et al. (2003) also observed a saccadic selectivity effect. In Figure 2, the frequency of saccades to same-colour distractors is plotted against the number of same-colour distractors. In the plot, the saccade frequencies are higher for rare features (colour or shape) than should be expected by chance (represented by the horizontal line). When the same-colour distractors are rare in the display, the participants were more likely to make eye movements towards them than when they were common. Conversely, when the number of same-colour distractors was high, the participants were more likely to make eye movement towards same-shape distractors.

The Model

In the following sections we describe the individual components of the model for performing a 36-element distractor-ratio task, and provide a walk-through of the model process before presenting the model results.

External Display

In the model, we represent the display by randomly distributing the target and the distractors in a grid, where each cell consists of either a target object, a distractor object with common colour or a distractor object with common shape. In the display, there is only one target object and the number of distractors are determined by randomly sampling a ratio per trail.

The display is represented by two feature vectors, one for colour and one for shape. The presence or absence of a feature in each cell in the model is represented numerically by the number 1 for presence and 0 for absence. The random distribution of these features in the environment was achieved by sampling randomly from the following set of ratios, r = R (3:33, 6:30, 9:27, 12:24, 15:21, 18:18, 21:15, 24:12, 27:9, 30:6, 33:3).

Actions

The action space consists of (1) fixate on a cell, (2) respond present and (3) respond absent. In our study there was a grid of 6x6 coloured shapes and there were therefore a total of 36 possible fixation actions. A trial was terminated by the choice of the present or absent action.

Reward

A reward was given after choosing a present or absent action. The reward distribution was defined as a value 10 for a correct response, a value of -10 for an incorrect response and a value of -1 for each fixation. The penalty on each fixation imposes a speed-accuracy trade-off. More fixations gives greater accuracy but at a cost.

Observation Model

Every time the model fixates, it also makes an observation. The observation obtained by the model is constrained by the noise in the human visual system. Two types of noise are added to the signal: spatial smearing noise and feature noise.

1. **Feature Noise:** The human eye's ability to discriminate and perceive object features degrades with eccentricity according to a hyperbolic function (Strasburger, Rentschler, & Jüttner, 2011). To model this function we added Gaussian white noise with mean 0 and standard deviation as eccentricity, i.e., a function of visual angle 'θ' between the fovea and the given location, and a scalar weight '$w_{featural}$' to scale the effect of distance to the fovea for feature noise. Therefore, the equation for the observation after adding feature noise at location j given that the eye is focused on location k is as follows,

$$\delta_{featural}(S_t, j) = v[s_t] + N(\theta, \sigma_f(\theta_{jk}, w_{featural}))$$

$$\sigma_{featural}(\theta_{jk}, w_{featural}) = \frac{\theta_{jk}}{(w_{featural})} + c$$

where, $v[s_t] = 1$ if the location s_t contains a target feature, else $v[s_t] = 0$, c is a constant with value 10^{-4} to avoid 0 variance in the model, $\sigma_f(\theta, w_f)$ is the variance to simulate the degrading eccentricity and 'θ' is the distance between the fixated cell and location j.

2. **Spatial Smearing:** Another source of uncertainty in the human visual system is the localization error (Levi, 2008), where information in the parafovea may erroneously combine features from one location with adjacent locations. Therefore, for each location in the colour and shape vector a weighted sum is calculated for the location and its adjacent eight locations. For example, If a red X is surrounded by green Os in the parafovea then, as a consequence of spatial smearing, the participant would be uncertain whether they are actually looking at a red X or a green O.

   In the model, spatial smearing is represented by a weighting function (Gaussian kernel) with standard deviation as a function of visual angle 'θ' between the fovea and the given location, and a scalar weight '$w_{spatial}$' to scale the effect of distance to the fovea for spatial noise. The weighting function here is a normalised function. As 'θ' (distance) increases the acuity decreases and the standard deviation of the Gaussian kernel increases, this means that the percept of the item at a given location suffers greater interference from surrounding items. This encoding is done for each location in the display. Thus, the equation for the observation after adding spatial noise at location j given that the target features are at location $S_t \in (1, 2, ..., n)$ and the eye is focused on location k is as follows,

$$\delta_{percept}(S_t, j) = K(s, \sigma_s(\theta_{jk}, w_{spatial})) \times \delta_{featural}(S_t, j)$$

$$\sigma_{spatial}(\theta_{jk}, w_{spatial}) = \frac{\theta_{jk}}{(w_{spatial})} + c$$

where, K is the Gaussian kernel with kernel size $s = 1$, $\sigma_s(\theta_{jk}, w_s)$ is the variance. $\delta_{percept}(S_t, j)$ is calculated separately for both shape and colour feature vectors. c is a constant with value 10^{-4} to avoid 0 variance in the model.

Now each percept ($\delta_{percept}$) (one for colour and one for shape) is represented as a vector of noisy observations for each location. A consequence of introducing the noise is uncertainty in the content of the location.

State Estimation

At each time step t on which a fixation is made the model receives a noisy observation for each location. The values for perceived colour and shape are then combined (*Hadamard product*) for each location $[i, j]$. We refer to these combined values as *relevance scores*, where a higher score in a location signifies high relevance to the task. These scores are then integrated across fixations, using naive Bayesian inference (*Kalman filter*), to get the current state B_t which is a vector of estimated relevance scores across fixations[1].

[1]The integration of information across fixation is a local update for each cell.

Heuristic Control Model

The Heuristic control model makes fixations and observations as described above. In order to decide which fixations to use and when to respond it makes use of two heuristics. The first uses a MAP-like strategy to determine where to fixate next, and the second uses a thresholded stopping rule.

Optimal Control Model

As we have said, at each point in time, the model observes the external environment through a noisy percept with a high resolution fovea and low resolution parafovea and receives an observation o_t. The model then extract the high resolution local information from the environment by taking actions $a_t \in A$ (A is the set of actions) to move the fovea (e.g., choose where to move the fovea). Since the environment is only partially observed the model needs to integrate information over time in order to determine how to act and how to make eye movements most effectively. It does this using the Bayesian state estimator described above.

At each step, the model receives a scalar reward r_t (which depends on the action taken by the agent), and the goal of the agent is to maximize the total sum of such rewards $R = \mathbb{E}[\sum \gamma^{t-1} r^t]$, where $\gamma \in (0,1)$ is the discount factor.

The most important aspect of the Optimal Control model is that rather than using heuristics to choose what to do next, it learns an approximately optimal policy using Deep Q-learning.

Deep Q-learning The Deep Q-learner made use of the following network architecture.

The relevance score estimate B_t (36 element vector) from the state estimator (above) was taken as the input. This input was connected to a fully connected hidden layer consisting of nodes equivalent to number of elements in the display, i.e., 36, with rectifier activation function. This is followed by a second fully connected hidden layer consisting of again nodes equivalent to number of elements in the display, i.e., 36, with sigmoid activation function. Finally, the output layer was a fully connected linear layer of 38 nodes with single output for each action in the task. To avoid over-fitting of the network $l2$ regularization of the weights was applied with value 10^{-5}.

During the training process a fixed size batch of transitions $< s,a,r,s' >$ were sampled from a *replay memory* and used for learning. For each time step (t), the deep Q-network (with parameters θ) is trained to approximate the action-value (Q-value) function from the sampled transitions by minimizing the loss functions $L(\theta_i)$:

$$L(\theta_i) = \mathbb{E}_{s,a\sim\pi_\theta}[(y_i - Q(s,a;\theta))^2]$$

where $y_i = r + \gamma max_{a'} Q(s',a';\theta')$ is the target Q-value estimated from a target Q-network (θ'). The parameters of target Q-network (θ') is copied over from the learned network (θ) after a fixed number of iterations.

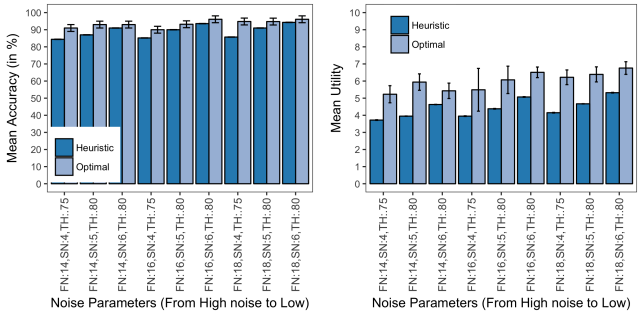
Algorithm 1 Deep Q Network Algorithm

initialize replay memory D, weights of the main network θ and target network θ'.
observe the initial state s.
repeat
  select an action a
    with probability epsilon select a random action.
    otherwise select $a = argmax_{a'} Q(s,a';\theta)$.
  perform the action a.
  observe the reward r and new state s' for action a.
  store transition $< s,a,r,s >$ in the replay memory D.
  sample random transitions $< s,a,r,s >$ from the replay memory D.
  calculate the target value t for each sampled transition.
  if s' is the terminal state **then**
    $t = r$
  else
    $t = r + \gamma Q(s', max_{a'} Q(s',a';\theta);\theta')$
  end if
  update the network using $(t - Q(s,a;\theta))^2$ as the loss.
  $s = s'$
  after every fixed steps $\theta' = \theta$
until terminal state

Model Results

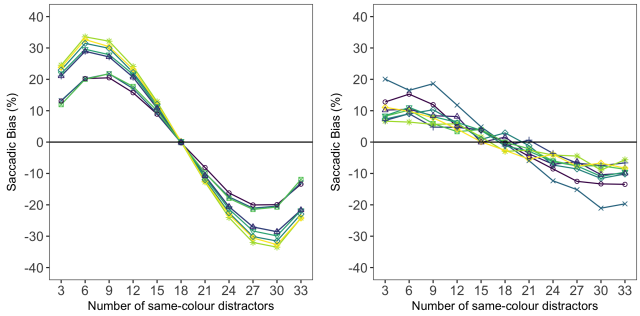
The Heuristic control model was run for 30,000 trials and 10 regression runs to check for consistency. The Optimal control model was run for 20 million trials. We first tested the accuracy of the models. Accuracy is the proportion of trials on which the model correctly responded either present or absent. The best fitting optimal control model achieved an accuracy of 96% in its last 50000 trials. In comparison, human participants achieved 98% accuracy. The accuracy of the best fitting Heuristic control model was 94%. Accuracy and utility of both models is plotted in Figure 3. The plots show a clear advantage of the Optimal control model for all explored parameter settings. In other words, the approximately Optimal control model outperforms the Heuristic control model in all cases.

Plots of fixation frequency versus same colour distractor-ratio at different levels of spatial and feature noise are shown in Figure 5. The results show that both model Heuristic and Optimal control model generate similar distractor ratio curves to humans (Figure 2) for target absent, where more fixations are required for ratios close to 1. While the RMSEs for the Heuristic control model were smaller than for the Optimal control model (Optimal: $RMSE = 0.81$; Heuristic $RMSE = 0.41$), the goodness of fit against Human performance for the Heuristic control model was $R^2 = 0.95$ and for the Optimal control model was $R^2 = 0.98$. A weakness of the Heuristic control model was that it produced DR effects for both target present and target absent. In contrast, the Optimal control model predicted a DR effect in the absent condition



(a) (b)

Figure 3: (a) Mean accuracy achieved by both models plotted against different noise parameter settings. (b) Mean utility gained by both models plotted against different noise parameter settings. Where, FN is Feature noise, SN is Spatial Noise and TH is the threshold set for heuristic control model.



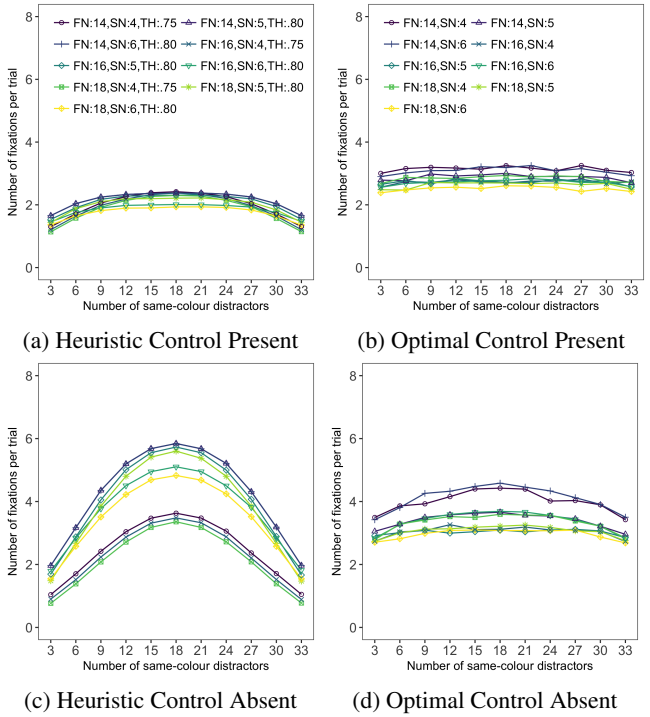
(a) Heuristic Control model (b) Optimal Control model

Figure 4: Saccadic bias as a function of the number of same colour distractors for Target Absent.

only. In terms of the shape of the DR curve and saccadic selectivity curve, the similarity between humans and Optimal control model is greater than the similarity between Heuristic control model and humans (see Figure 2).

The saccadic bias effect is shown in Figure 4. For the explored parameter settings, the Heuristic control model generated higher levels of saccadic bias than generated by the Optimal control model and these levels were nearer to those generated by humans (Optimal: $RMSE = 8.93$; Heuristic $RMSE = 6.93$). However, the Optimal control model explained more of the variance. The goodness of fit of the best fitting Heuristic control model was $R^2 = .94$. In contrast, the best fitting Optimal control model had a goodness of fit of $R^2 = 0.97$. While the Heuristic control model predicts a magnitude of saccadic bias that corresponds to that of humans at extreme levels of same-color (around 30%), it is the Optimal control model that has the better fit. This is likely due to the extreme curvature (sinusoidal) of the saccadic bias for the Heuristic model which is not present in the humans.

One of the effects in the human data that is not captured by either the Optimal or the Heuristic control model is the asymmetric effect of shape and colour (see Figure 2). This is very



(a) Heuristic Control Present (b) Optimal Control Present

(c) Heuristic Control Absent (d) Optimal Control Absent

Figure 5: Number of fixations as a function of same-colour distractors for (a) the Heuristic model with target present, (b) the Control model with target present, (c) the Heuristic model with target absent, (d) the Heuristic model with target present.

likely due to the fact that we used the same noise parameter values for both shape and colour in the model's observation function. Further work is needed to explore the effect of the known differences in acuity functions for shape and colour (Kieras & Hornof, 2014).

Discussion and Conclusion

While the results presented here are preliminary, they offer some evidence that the distractor-ratio effect is the consequence of an approximately optimal adaptation to the constraints imposed by the human visual information processing system. Unlike previous work, including Myers et al. (2013), our results are based on a model that makes approximately optimal control decisions to choose fixation locations rather than a model that uses MAP-like heuristics.

Achieving these results required two contributions to cognitive modeling. The first is the novel application of POMDPs to the framing of the distractor-ratio problem, further extending the work of Butko and Movellan (2008). The POMDP framing is important because it provides a rigorous basis for exploring the *computationally rational* adaptation of human strategies to known information processing constraints (Lewis et al., 2014; Howes, Lewis, & Vera, 2009). It thereby helps make the crucial link between cognitive mechanism and rationality that supports deep explanations of behaviour.

The second contribution is the novel application of Deep

Q-Learning (Mnih et al., 2015) to determine the optimal policy given a theory of human visual information processing capacities. The role of reinforcement learning based algorithm's have previously been proposed as means of explaining human learning processes (Dayan & Daw, 2008) and also, as means of deriving rational analyses of what a person should do in particular task (Chater, 2009). Our work is more aligned with the goals of (Chater, 2009). The purpose of our reinforcement learner was not to model the step-by-step learning process, but rather to model the rational outcome of the learning process – an approximately optimal adaptation to information processing limits.

There is a substantial amount of work to be done. While the best fitting Optimal control model explained 98% of the variance, to be fully confident that it is better than the Heuristic control model, we need to more fully explore the parameter space of both models. For example, for the Heuristic control model, it might be the case that even higher feature noise, and lower spatial noise, might further improve the fit. We also need to find a fit that reduces the RMSE of the Optimal control model.

In conclusion, we have demonstrated that framing the visual search problem as a POMDP and solving this problem with deep Q-learning is a viable approach to explaining effects such as distractor-ratio and saccadic selectivity.

References

Bacon, W. F., & Egeth, H. E. (1997). Goal-directed guidance of attention: evidence from conjunctive visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(4), 948.

Butko, N. J., & Movellan, J. R. (2008). I-pomdp: An infomax model of eye movement. In *Development and learning, 2008. icdl 2008. 7th ieee international conference on* (pp. 139–144).

Chater, N. (2009). Rational and mechanistic perspectives on reinforcement learning. *Cognition*, *113*(3), 350–364.

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(4), 429–453.

Egeth, H. E., Virzi, R. A., & Garbart, H. (1984). Searching for conjunctively defined targets. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(1), 32.

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision research*, *51*(7), 771–781.

Hayhoe, M., & Ballard, D. (2014). Modeling task control of eye movements. *Current Biology*, *24*(13), R622–R628.

Howes, A., Lewis, R. L., & Vera, A. (2009). Rational adaptation under task and processing constraints: implications for testing theories of cognition and action. *Psychological review*, *116*(4), 717.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, *40*(10), 1489–1506.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial intelligence*, *101*(1), 99–134.

Kieras, D. E., & Hornof, A. J. (2014). Towards accurate and practical predictive models of active-vision-based visual search. In *Proceedings of the 32nd annual acm conference on human factors in computing systems* (pp. 3875–3884).

Kowler, E. (2011). Eye movements: The past 25years. *Vision research*, *51*(13), 1457–1483.

Levi, D. M. (2008). Crowdingan essential bottleneck for object recognition: A mini-review. *Vision research*, *48*(5), 635–654.

Lewis, R. L., Howes, A., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in cognitive science*, *6*(2), 279–311.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... others (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533.

Myers, C. W., Lewis, R. L., & Howes, A. (2013). Bounded optimal state estimation and control in visual search: Explaining distractor ratio effects. In *Proc. cogsci.*

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*(7031), 387–391.

Najemnik, J., & Geisler, W. S. (2008). Eye movement statistics in humans are consistent with an optimal search strategy. *Journal of Vision*, *8*(3), 4–4.

Nunez-Varela, J., & Wyatt, J. L. (2013). Models of gaze control for manipulation tasks. *ACM Transactions on Applied Perception (TAP)*, *10*(4), 20.

Pomplun, M., Reingold, E. M., & Shen, J. (2003). Area activation: A computational model of saccadic selectivity in visual search. *Cognitive Science*, *27*(2), 299–312.

Shen, J., Reingold, E. M., & Pomplun, M. (2000). Distractor ratio influences patterns of eye movements during visual search. *Perception*, *29*(2), 241–250.

Shen, J., Reingold, E. M., & Pomplun, M. (2003). Guidance of eye movements during conjunctive visual search: the distractor-ratio effect. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *57*(2), 76.

Sprague, N., Ballard, D., & Robinson, A. (2007). Modeling embodied visual behaviors. *ACM Transactions on Applied Perception (TAP)*, *4*(2), 11.

Strasburger, H., Rentschler, I., & Jüttner, M. (2011). Peripheral vision and pattern recognition: A review. *Journal of vision*, *11*(5), 13–13.

Wolfe, J. M. (2007). Guided search 4.0. *Integrated models of cognitive systems*, 99–119.

Zohary, E., & Hochstein, S. (1989). How serial is serial processing in vision? *Perception*, *18*(2), 191–200.