Paper 2

Dan Birman

Moving Beyond Saliency Maps to Modeling the Full Role of Visual Attention

As we move through the world we attend from moment-to-moment to visual information both by moving our eyes and 'covertly', without saccades. Cognitive neuroscience suggests that visual attention, at the neural level, is a modification of the inputs and outputs of specific neural populations (Pessoa, Kastner, & Ungerleider, 2003). At a higher level of abstraction attention can be understood of as a shift in the representation of different stimulus features. For example, the representation of contrast in visual cortex is a monotonic function of physical stimulus contrast. The effect of attention is to alter the properties of that function: for example, the readout mechanism can be altered by spatial attention (Hara & Gardner, 2014; Pestilli, Carrasco, Heeger, & Gardner, 2011). At the most abstract level, Maar's computational level (Maar, David, 1982), attention is a way to increase the signal to noise ratio of a noisy input. It's clear that the visual system has a mechanism for attention, but how do we optimize the deployment of attention? At the core this is a probabilistic computation: we must continuously decide where to deploy attentional resources, at the cost of other locations and features, based on our expectations about the state of the world. Increases in computing strength have allowed researchers to begin probing potential models of bottom-up and top-down attention (Borji & Itti, 2013). In a recent paper Borji, Sihite, and Itti designed a probabilistic model of top-down attention, designed to predict eye fixations (Borji, Sihite, & Itti, 2012). At its core their method identifies what region of the visual scene will be most informative to attend to. But a major shortcoming of their method is that it doesn't take into account the costs of making eye movements—instead they make the simplifying assumption that all attention is fixation. But saccades and fixations are not necessarily the optimal action when there is ambiguity about scene salience. A complete understanding of attention will likely require a model that takes into account the costs and limits of deploying attention as well as the stimulus salience and current goals.

Most recent algorithms for predicting fixation have focused on bottom-up scene salience which can be calculated from scene statistics alone. But visual salience alone is inadequate for explaining viewing behavior in complex tasks: bottom-up salience alone accounts for only a minority of human fixation behaviors in dynamic tasks like video games (Borji & Itti, 2013). To understand why this is the case we can consider a stereotypical human task. In driving the most salient part of a visual scene may be the background (e.g. driving in Alaska) while the most task-relevant region may be the road. This is also true in artificial scenes such as games, where distracting motion and background visuals may have very high salience while being largely task-irrelevant (Borji et al., 2012). Borji et al. introduce top-down attention to their model by taking into account prior knowledge about each visual task, including where fixations have occurred previously for difference scene gists. Their approach uses three pieces of abstract knowledge: the global 'gist' of the scene (in part this is the bottom-up salience, but it also takes

into account task relevance), recent past eye movements, and motor actions. Their goal is to estimate $P(X|Knowledge)$, the probability of attending to location X given all known information. Their full formula and derivation are explained in more detail in their methods, but it approximates to the following:

$$P\left(X_t \middle| G_{1:t}, X_{1:t-1}, A_{1:t-1}^{j=1:n}\right)$$

Where X denotes the possible fixation locations as (x, y) coordinates. The priors they implement are: G, the gist of the scene, X, the fixations on the previous trials, and the n actions A performed on each of the previous trials. Inverting their formula with Bayes' rule requires computing $P(G_t|X_t)$, the bidirectional mapping of gist and fixation (i.e. given a fixation, what is distribution over scene gist), $P(X_{t-1}|X_t)$, previous fixations, and $P\left(A_{t-1}^{j=1:n}\middle|X_t\right)$ or the distribution of prior actions. They compare this probabilistic model with a battery of more simplistic visual salience models that attempt to take into account the scene gist and previous viewing statistics. They show that their total model outperforms all of the other models compared, which only take into account visual salience. Comparing their model predictions to human fixations shows that at a false positive (FP) rate of ~10% their model achieves a true positive (TP) rate of 50-75%, this is a large (~25%) improvement over prior models.

Borji et al. provide compelling evidence that top-down attention is well modeled as a probabilistic process. Their approach makes a prediction about the most informative region to look at next, given the previous areas looked at, the scene gist, and the overall scene statistics. But their approach stays entirely at the computational level and makes no prediction about what kinds of constraints and requirements an actual visual architecture might have. This is fine for computer science applications of attention (e.g. for A.I. in video games that need to emulate human behavior) but is a poor approximation of the visual system for cognitive science. An interesting and fruitful direction to push this research would be in the line outlined by Griffiths et al. (Griffiths, Lieder, & Goodman, 2015). In essence, constraining the algorithms that are allowed to be implemented by these models according to increasingly rigorous constraints imposed by known constraints in the brain. We might wonder, for example, about how exactly attentional information should be used. Borji et al.'s model matches only ~65% TP at a 10% FP rate—are the other 35% of fixations occurring in other areas because the human brain doesn't continuously deploy overt fixations? This is a plausible explanation: we might imagine that fixations come at some cost, both in terms of actual physical cost but also the opportunity cost of losing visual acuity in regions that aren't fixated. The visual system may have evolved covert attention as a solution to this. In the Borji et al. model fixations tend not to occur twice in the same region because of their strong fixation prior. Their model predicts that observers should change fixations quickly to scan all of the currently salient regions of the visual field. But there are many situations where human observers violate this strategy. While driving we often fixate directly ahead, using covert attention to scan for potential dangers in our peripheral vision. We use short saccades to confirm noisy evidence from covert attention—did we really see a deer, or just the reflection of our headlights on a sign?

The model outlined by Borji et al. could be extended to model covert attention. They use a threshold method to find the current maximum location at which to fixate. But in ambiguous situations where two or more locations demand attention they could allow for fixation to remain fixed while

covert attention is deployed. Again this is a logical solution to ambiguity when a saccade to an un-informative location can be a dangerous error. Adding covert attention to Borji et al.'s model would require an additional covert attention prior to be calculated. Covert attention would also need to be tracked as a second map of information, similar to their existing fixation map. To test this more complex model would require having some knowledge of how humans deploy covert attention, requiring a measurement that might correlate with covert attention. There's some evidence that microsaccades might be a useful way to measure this (Hafed & Clark, 2002), and this is conveniently a measure that exists in the data that Borji et al. already collected.

As models of visual cortex have leveraged increasing computing power it has become difficult to understand their inner workings. There has been remarkable success using convolutional neural networks in object recognition, but we know relatively little about the inner workings of these models (Yamins et al., 2014). As an alternative to building complicated algorithmic models of cortex we can look to approaches like that of Borji et al.: constructing intuitive computational models that might be extended, using logical biological constraints, down to the algorithmic level. Borji et al. make the compelling argument that fixations are deployed as part of a generative probabilistic model that takes into account task demands and previous actions (Borji et al., 2012). The next logical step beyond their approach is to introduce additional biological constraints—such as the direct and indirect costs of saccadic eye movements. Probabilistic models provide a compelling and understandable advance in our knowledge of how and why attention is deployed. Building in biological constraints to these understandable models suggests a roadmap to algorithmic models that are interpretable and make testable predictions.

References

Borji, A., & Itti, L. (2013). State-of-the-Art in Visual Attention Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 185–207. http://doi.org/10.1109/TPAMI.2012.89

Borji, A., Sihite, D. N., & Itti, L. (2012). Probabilistic learning of task-specific visual attention. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (pp. 470–477). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6247710

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational Use of Cognitive Resources: Levels of Analysis Between the Computational and the Algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229. http://doi.org/10.1111/tops.12142

Hafed, Z. M., & Clark, J. J. (2002). Microsaccades as an overt measure of covert attention shifts. *Vision Research*, *42*(22), 2533–2545.

Hara, Y., & Gardner, J. L. (2014). Encoding of graded changes in spatial specificity of prior cues in human visual cortex. *Journal of Neurophysiology*. http://doi.org/10.1152/jn.00729.2013

Maar, David. (1982). Chapter 1: The Philosophy & Approach. In *Vision* (pp. 19–29). Cambridge, MA: MIT Press.

Pessoa, L., Kastner, S., & Ungerleider, L. G. (2003). Neuroimaging studies of attention: from modulation of sensory processing to top-down control. *The Journal of Neuroscience*, *23*(10), 3990–3998.

Pestilli, F., Carrasco, M., Heeger, D. J., & Gardner, J. L. (2011). Attentional enhancement via selection and pooling of early sensory responses in human visual cortex. *Neuron*, *72*(5), 832–46. http://doi.org/10.1016/j.neuron.2011.09.025

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.