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 a vast number of different parts of the visual world. Cognitive science makes a strong and well-founded claim that visual attention, at the neural level, is a process that modulates the strength of signals in visual cortex (Pessoa, Kastner, & Ungerleider, 2003). At a higher level of abstraction from neural activity 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 the physical stimulus contrast. The effect of attention is to alter the properties of that function: for example, the readout mechanism can all be altered by spatial attention (Hara & Gardner, 2014; Pestilli, Carrasco, Heeger, & Gardner, 2011). In tandem, the effects of attention for non-spatial features have been repeatedly observed in behavioral paradigms. For example in dual tasks subjects show significant impairment for ‘unattended’ stimuli such as natural scenes (Cohen, Alvarez, & Nakayama, 2011; Mack, Erol, & Clarke, 2015; Mack, Arien & Rock, Irvin, 1998). 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. Attention is, at the core, a probabilistic computation: deciding where to deploy attentional resources (at the cost of other locations and features) implies some expectation about the state of the world. Understanding how attention is deployed is a fascinating and difficult problem and one that has only begun recently to fall to efficient computational methods (Borji & Itti, 2013). In a recent paper Borji, Sihite, and Itti designed a probabilistic model of visual salience (Borji, Sihite, & Itti, 2012). Their model takes into account both task goals and visual salience and can correctly predict with high accuracy the eye movements of humans playing a variety of visual games (relative to other models). At its core their method identifies what region of the visual scene will be most informative, given prior saccades and knowledge about scene gist and task goals (their design is efficient in the sense that it doesn’t require knowledge of all the objects in a scene). But their method also makes no claim about how the visual system should use this knowledge—should representations be covertly modified without an eye movement to improve their signal to noise ratio? Or should the eyes be moved via a saccade to gain more information about a precise location in the visual scene. Understanding attention will likely require an understanding of both how salience is generated from bottom-up and top-down requirements and an understanding of what constraints on attention exist in the visual system.

Visual salience alone is insufficient to explain our viewing behavior in complex tasks. For example 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. solve the question of where attention should be directed in these complex situations by formulating the problem as a question of probabilistic prior knowledge. Their approach uses three pieces of abstract knowledge: the global ‘gist’ of the scene, recent past eye movements, and motor actions. Using Bayes rule they attempt to estimate P(X|I), 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:

Where X denotes the possible fixation locations as (x, y) coordinates. The priors they implement are: G, the gist of the scene, 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 , the probability that a scene contained a certain gist given that an observer fixated a particular region, , and or the probability of the previous actions occurring prior to the this fixation. (Note that a number of parts of the formula drop out after applying Bayes’ theorem under some assumptions that they make). They compare this probabilistic model with a battery of more simplistic salience models that attempt to take into account the scene gist and the visual statistics. They show that their total model outperforms all of the other models compared, achieving an AUC of ~.8 across their test video games. At a false positive (FP) rate of ~10% their model achieves a true positive (TP) rate of 50-75% of observed human fixations.

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 constraints 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 the actual neural architecture. We might wonder, for example, about how exactly this 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 can deploy both overt fixations and covert attention? This is a plausible explanation, we might imagine that their model, which tends not to fixate twice on the same region, would expect an observer to change fixations rather quickly. But this isn’t true in some situations, for example while driving we often fixate directly ahead, using covert attention to scan for potential dangers in our peripheral vision. Once identified, we often quickly saccade to check—did we really see a deer, or just the reflection of our headlights on a sign? This hypothesis could easily be checked in the data that Borji et al. collected. In addition their model can be easily extended to allow for covert attention. They use a threshold method to determine where fixation should be, but in ambiguous situations where two or more locations demand our attention (like driving) they could allow for fixation to remain fixed while covert attention is deployed to those regions, in preparation for a saccade.

**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

Cohen, M. a, Alvarez, G. a, & Nakayama, K. (2011). Natural-scene perception requires attention. *Psychological Science*, *22*(9), 1165–72. http://doi.org/10.1177/0956797611419168

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

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.

Mack, A., Erol, M., & Clarke, J. (2015). Iconic memory is not a case of attention-free awareness. *Consciousness and Cognition*, *33*, 291–299. http://doi.org/10.1016/j.concog.2014.12.016

Mack, Arien, & Rock, Irvin. (1998). *Inattentional Blindness.* The 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