Models of Bottom-Up Attention and Saliency

Laurent Itti [*]

Abstract: *Visually conspicuous, or so-called salient, stimuli often have the capability of attracting focal visual attention towards their locations. Several computational architectures subserving this bottom-up, stimulus-driven, spatiotemporal deployment of attention are reviewed in this article. The resulting computational models have applications not only to the prediction of visual search psychophysics, but also, in the domain of machine vision, to the rapid selection of regions of interest in complex, cluttered visual environments. We describe an unusal such application, to the objective evaluation of advertising designs.*

One of the most important functions of selective visual attention is to rapidly direct our gaze towards objects of interest in our visual environment. From an evolutionary standpoint, this rapid orienting capability is critical in allowing living systems to quickly become aware of possible preys, mates or predators in their cluttered visual world. It has become clear that attention guides where to look next based on both bottom-up (image-based) and top-down (task-dependent) cues (James, 1890/1981). As such, attention implements an information processing bottleneck, only allowing a small part of the incoming sensory information to reach short-term memory and visual awareness [Linking Attention to Learning, Expectation, Competition and Consciousness]. That is, instead of attempting to fully process the massive sensory input in parallel, nature has devised a serial strategy to achieve near real-time performance despite limited computational capacity: Attention allows us to break down the problem of scene understanding into rapid series of computationally less demanding, localized visual analysis problems.

Developing computational models of how attention is deployed onto complex visual scenes has been a long-standing challenge for fundamental neuroscience, with additional motivation provided by numerous potential applications in artificial vision, for tasks including surveillance, automatic target detection, navigational aids and robotics control. Here we focus on biologically-plausible computational modeling of bottom-up guidance of attention towards salient image locations, while [Attention and scene understanding] casts these models within broader frameworks that combine bottom-up and top-down attention control signals.

1 Preattentive Features and Saliency Map

Development of computational models of attention started with the Feature Integration Theory of Treisman & Gelade (1980), which proposed that only simple visual features are computed in a massively parallel

[*] University of Southern California, Hedco Neuroscience Building HNB-30A, Los Angeles, CA 90089-2520

manner over the entire visual field. Attention is then necessary to bind those early features into a united object representation, and the selected bound representation is the only part of the visual world that passes though the attentional bottleneck. This idea was further specified into a neurally-plausible computational architecture by Koch and Ullman (Koch & Ullman, 1985) and later by Itti *et al.* (Itti *et al.*, 1998) **(Figure 1)**. These models are centered around a "saliency map," that is, an explicit two-dimensional topographic map which encodes for stimulus conspicuity, or salience, at every location in the visual scene. The saliency map receives inputs from early visual processing, and provides an efficient control strategy by which the focus of attention simply scans the saliency map in order of decreasing saliency.

Several important lessons have been learned from these models and the empirical studies that helped better specify them. First, different features contribute with different strengths to perceptual salience (Nothdurft, 2000), and this relative feature weighting can be influenced in a task-dependent manner, top-down (Wolfe, 1994) and through training. Second, at a given visual location, there is little evidence for strong interactions across different visual modalities, such as color and orientation (Nothdurft, 2000). This is not too surprising from a computational standpoint, as one would otherwise expect these interactions to also be trainable and modulable top-down, resulting in the ability to learn to efficiently detect conjunctive targets, which we lack **(Figure 2)**. Third and most importantly, what appears to matter in guiding bottom-up attention is feature contrast, not local absolute feature strength. Indeed, not only are most early visual neurons tuned to some type of local spatial contrast (such as center-surround or oriented edges), but neuronal responses are also strongly modulated by context, in a manner that extends far beyond the range of the classical receptive field (Allman *et al.*, 1985). In a first approximation, the computational consequences of non-classical surround modulation are two-fold: First, a broad inhibitory effect is observed when a neuron is excited with its preferred stimulus but that stimulus extends beyond the neuron's classical receptive field, compared to when the stimulus is restricted to the classical receptive field and the surrounding visual space either is empty or contains non-preferred stimuli (Sillito *et al.*, 1995). Second, long-range excitatory connections in V1 appear to enhance responses of orientation-selective neurons when stimuli extend to form a contour (Gilbert *et al.*, 2000). These interactions are thought to play a critical role in perceptual grouping. The net result is that activity in early cortical areas is surprisingly sparse when monkeys are free-viewing natural scenes, compared to the vigorous responses that can be elicited by small laboratory stimuli presented in isolation **(Figure 2)**.

2 Implemented Architectures

Many successful models for the bottom-up control of attention are architectured around a saliency map. What differentiates the models, then, is the strategy employed to prune the incoming sensory input and extract salience.

One strategy proposed by Wolfe (1994) has been to rely on top-down knowledge to emphasize those features which may distinguish a target item being searched for from surrounding clutter: if you are searching for a red object, emphasize the contribution of color to your saliency map, while toning down the influence of orientation. The FeatureGate model of Cave *et al.* provides a full neural network implementation of a system that combines this idea with bottom-up mechanisms [The FeatureGate Model of Visual Selection], so that a target may attract attention based on combined bottom-up salience and top-down expectations. In a similar vein, Rao *et al.* [Probabilistic Models of Attention based on Iconic Representations and Predictive Coding] have proposed that saliency could be computed from the Euclidean distance between a target feature vector (that is, a target signature, in terms of respective amounts of various colors, orientations, etc present in the target) and feature vectors extracted at every location in the visual input. Extending these ideas in a Bayesian framework, Torralba has proposed that a coarse global analysis of the entire scene's gist may provide contextual guidance cues [Contextual Influences on Saliency]: for example, when searching for cars, rapidly realizing that you are looking at a city scene and having coarsely understood its layout may help you more rapidly focus onto street surfaces, where cars are likely to be.

Studying richer interactions between low-level vision and a saliency map, Tsotsos and colleagues [The Selective Tuning Model of Attention] implemented attentional selection using a combination of a feedforward bottom-up feature extraction hierarchy and a feedback selective tuning of these feature extraction mechanisms. In this model, the target for attention focusing is selected at the top level of the processing hierarchy (the equivalent of a saliency map), based on feedforward activation and on possible additional top-down biasing for certain locations or features. That location is then propagated back through the feature extraction hierarchy, through the activation of a cascade of winner-take-all networks embedded within the bottom-up processing pyramid. Spatial competition for salience is thus refined at each level of processing, as the feedforward paths not contributing to the winning location are pruned (resulting in the feedback propagation of an "inhibitory beam" around the selected target).

In view of the affluence of models based on a saliency map, it is important to note that postulating centralized control based on such map is not the only computational alternative for the bottom-up guidance of attention. For example, Zhaoping [The primary visual cortex creates a bottom-up saliency map] proposed that salience may not necessarily be computed in a separate brain area from the low-level visual features, and that instead it may be expressed as a modulation onto feature responses observed in V1. Similarly, Desimone and Duncan (Desimone & Duncan, 1995) argued that salience is not explicitly represented by specific neurons, but instead is implicitly coded in a distributed modulatory manner across the various feature maps. Attentional selection is then performed based on top-down weighting of the bottom-up feature maps that are relevant to a target of interest. Several models have successfully applied this strategy to synthetic stimuli [How the detection of objects in natural scenes constrains attention in time], [A Neurodynamical Model of Visual Attention].

Although originally a theoretical construct supported by sparse experimental evidence, the idea of a unique, centralized saliency map appears today to be challenged by the multiplicity of candidate neural correlates recently unraveled, including areas in the lateral intraparietal sulcus of the posterior parietal cortex, the frontal eye fields, the inferior and lateral subdivisions of the pulvinar and the superior colliculus (Kustov & Robinson, 1996; Gottlieb *et al.*, 1998). One possible explanation for this multiplicity could be that some of the neurons in all those areas indeed are concerned with the explicit computation of salience, but are found at different stages along the sensory-motor processing stream. For example, other functions have also been assigned to the posterior parietal cortex, such as that of mapping retinotopic to head-centered coordinate systems and of memorizing targets for eye or arm movements (Andersen *et al.*, 1990; Dominey & Arbib, 1992). More detailed experimental studies are thus needed to tell apart possible subtle differences in the functions and representations found in those brain areas. Most probably, the main difference between those brain regions is the balance between their role in perception and action [Dissociation of Selection from Saccade Programming].

3 Attention and recognition

So far, we have reviewed computational modeling and supporting experimental evidence for a basic architecture concerned with the bottom-up control of attention: Early visual features are computed in a set of topographic feature maps; spatial competition for salience prunes the feature responses to only preserve a sparse representation of the few most conspicuous locations; all feature maps are then combined into a unique scalar saliency map; and, finally, the saliency map is scanned by the focus of attention through the interplay between winner-take-all and inhibition-of-return. While such simple computational architecture may accurately describe how attention is deployed within the first few hundreds of milliseconds following the presentation of a new scene, it is obvious that a more complete model of attentional control must include top-down, volitional biasing influences as well. The computational challenge, then, lies in the integration of bottom-up and top-down cues, such as to provide coherent control signals for the focus of attention, and in the interplay between attentional orienting and scene or object recognition.

An very interesting example of integration in a model was recently provided by Schill *et al.* [A Model of Attention and Recognition by Information Maximization]. Their model aims at performing scene (or object) recognition, using attention (or eye movements) to focus on those parts of the scene being analyzed which are most informative in disambiguating its identity. To this end, a hierarchical knowledge tree is trained into the model, in which leaves represent identified objects, intermediary nodes represent more general object classes, and links between nodes contain sensorymotor information used for discrimination between possible objects (i.e., bottom-up feature response to be expected for particular points in the object, and eye movements targeted at those points). During the iterative recognition of an object, the system programs its next fixation towards the location which will maximally increase information about the object being recognized, in that it will best allow the model to discriminate between the various current candidate object classes.

Rybak *et al.* [Attention-Guided Recognition Based on "What" and "Where" Representations: A Behavioral Model] proposed a related model, in which scanpaths (containing motor control directives stored in a "where" memory and locally expected bottom-up features stored in a "what" memory) are learned for each scene or object to be recognized. When presented with a new image, the model starts by selecting candidate scanpaths based on matching bottom-up features in the image to those stored in the "what" memory. For each candidate scanpath, the model deploys attention according to the directives in the "where" memory, and compares the local contents of the "what" memory at each fixation to the local image features. This model has demonstrated strong ability to recognize complex grayscale scenes and faces, in a translation, rotation and scale independent manner.

A more extreme view at the basis of the models just mentioned is the "scanpath theory" of Stark (Noton & Stark, 1971), in which the control of eye movements is almost exclusively under top-down control. The theory proposes that what we see is only remotely related to the patterns of activation in our retinas, as suggested by our permanent illusion of crisp perception over our entire visual environment, although only the central 2° of our foveal vision provide such crisp sampling of the visual world. Rather, the scanpath theory argues that a cognitive model of what we expect to see is the basis for our percept; the sequence of eye movements which we make to analyze a scene, then, is mostly controlled top-down by our cognitive model of that scene. This theory has had a number of successful applications to robotics control, in which an internal model of a robot's working environment was used to restrict the analysis of incoming video sequences to a small number of circumscribed regions important for a given task [Scanpath Theory, Attention and Image Processing Algorithms for Predicting Human Eye Fixations].

One important challenge for combined models of attention and recognition consists of finding suitable neuronal correlates. Despite a biological inspiration in their architectures, the models just reviewed in this section, indeed, do not relate in much detail to biological correlates of object recognition. Although a number of biologically-plausible models have been proposed for object recognition in the ventral "what" stream [Object recognition in cortex: Neural mechanisms, and possible roles for attention], their integration to biological models mostly concerned with attentional control in the dorsal "where" stream remains an open issue. This integration will, in particular, have to account for the increasing experimental support for an object-based rather than purely spatial focus of attention [Neural Evidence for Object Based Attention].

4 Applications

Attention is a desirable mechanism not only for biological organisms, but also for efficient allocation of resources in artificial systems [Saliency in Computer Vision]. Several chapters in the present volume indeed explore how an attentional component may serve to build better sensing machines [Attentive Wide-Field Sensing for Visual Telepresence and Surveillance], [Neuromorphic Selective Attention Systems], [the role of visual attention in the control of locomotion], [Attention Architectures for Machine Vision and Mobile Robots], [Attention for Computer Graphics Rendering].

Here we relate a rather anecdotical but unusual application of attention models, to the quantitative evaluation of advertising designs. In particular, we explore the problem of selecting among several candidate images for a magazine's cover. In creative and advertising design, current evaluation and selection criteria for candidate artwork creations mostly relies on focus groups, where groups of expert and/or naive human subjects compare the qualities and drawbacks of several competing design proposals. A question which arises, then, is whether a more objective metric could be developed to rank competing proposals. One such metric, particularly popular in the 1980s, has been to use eye-tracking devices to monitor eye movements of human subjects viewing the candidate designs. However, the high cost and difficulty of implementation of such evaluation has limited its applicability, and few advertising agencies appear to have developed eye-tracking setups in-house.

Computational models of attention, to the extent to which they may indeed relatively well predict the visual attractiveness of different locations in an artwork piece, may provide a simpler and more cost-effective solution to the problem of developing rapid, objective and unbiased evaluation criteria. The assumption is that locations marked as highly salient by the model ought to attract the gaze of a majority of potential customers. A better design, then, is one where information which matters to the advertiser is conveyed at these locations.

Post-sales market analysis in the domain of magazine sales have revealed that what sells a given magazine issue is not the quality or beauty of the artwork present on the cover, but the catchy text messages which tease casual viewers about the contents of the magazine (source: Peter Walker of McCann-Erickson, a large multinational advertising agency). This allows one to set a fairly simple quantitative criterion for the evaluation of various cover proposals: whichever candidate design has the highest average model-predicted salience over the important text messages wins. An example of such application is shown in **Figure 3**. While in one design most of the text messages are reliably marked as salient by the model, in the other approximately the lower half of the text is ignored, because of its low contrast against a textured background.

This simple application certainly has limitations, in particular because a purely bottom-up model with no object recognition or text reading capability was here used to evaluate the covers, but helps in opening the horizon of potential applications for attention models. In a similar vein, such models could also be used to assist merchandising experts in optimally arranging their products on supermarket shelves, or to assist product placement experts in optimizing the exposure of sponsor products in films or televised sports.

5 Summary and Conclusion

We have reviewed recent work on biologically-plausible computational models of attention, with a particular emphasis on bottom-up control of attentional deployment. While much progress has been made in the past few decades, the field of attention modeling still is very young and appears full of promises. In particular, one of the grand challenges for the coming years will be in further developing system-level models where attention interacts with object recognition, gist analysis, symbolic knowledge and top-down task demands during complex goal-driven scene understanding [Attention and Scene Understanding]. Yet, already today, attention modeling enjoys many exciting applications which also deserve to be further explored, from automatic target detection to advertising design.

References

Allman, J., Miezin, F., & McGuinness, E. 1985. Stimulus specific responses from beyond the classical receptive field: neurophysiological mechanisms for local-global comparisons in visual neurons. *Annu. Rev. Neurosci.*, **8**, 407–430.

Andersen, R. A., Bracewell, R. M., Barash, S., Gnadt, J. W., & Fogassi, L. 1990. Eye position effects on visual, memory, and saccade-related activity in areas LIP and 7a of macaque. *J. Neurosci.*, **10**(4), 1176–96.

Desimone, R, & Duncan, J. 1995. Neural mechanisms of selective visual attention. *Annu Rev Neurosci*, **18**, 193–222.

Dominey, P. F., & Arbib, M. A. 1992. A cortico-subcortical model for generation of spatially accurate sequential saccades. *Cereb. Cortex.*, **2**(2), 153–175.

Gilbert, C., Ito, M., Kapadia, M., & Westheimer, G. 2000. Interactions between attention, context and learning in primary visual cortex. *Vision Res.*, **40**(10-12), 1217–1226.

Gottlieb, J P, Kusunoki, M, & Goldberg, M E. 1998. The representation of visual salience in monkey parietal cortex. *Nature*, **391**(6666), 481–4.

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.

James, W. 1890/1981. *The Principles of Psychology.* Cambridge, MA: Harvard University Press.

Koch, C, & Ullman, S. 1985. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol*, **4**(4), 219–27.

Kustov, A A, & Robinson, D L. 1996. Shared neural control of attentional shifts and eye movements. *Nature*, **384**(6604), 74–7.

Nothdurft, H. 2000. Salience from feature contrast: variations with texture density [In Process Citation]. *Vision Res.*, **40**(23), 3181–3200.

Noton, D, & Stark, L. 1971. Scanpaths in eye movements during pattern perception. *Science*, **171**(968), 308–11.

Sillito, A M, Grieve, K L, Jones, H E, Cudeiro, J, & Davis, J. 1995. Visual cortical mechanisms detecting focal orientation discontinuities. *Nature*, **378**(6556), 492–6.

Treisman, A M, & Gelade, G. 1980. A feature-integration theory of attention. *Cognit Psychol*, **12**(1), 97–136.

Wolfe, J M. 1994. Visual search in continuous, naturalistic stimuli. *Vision Res*, **34**(9), 1187–95.

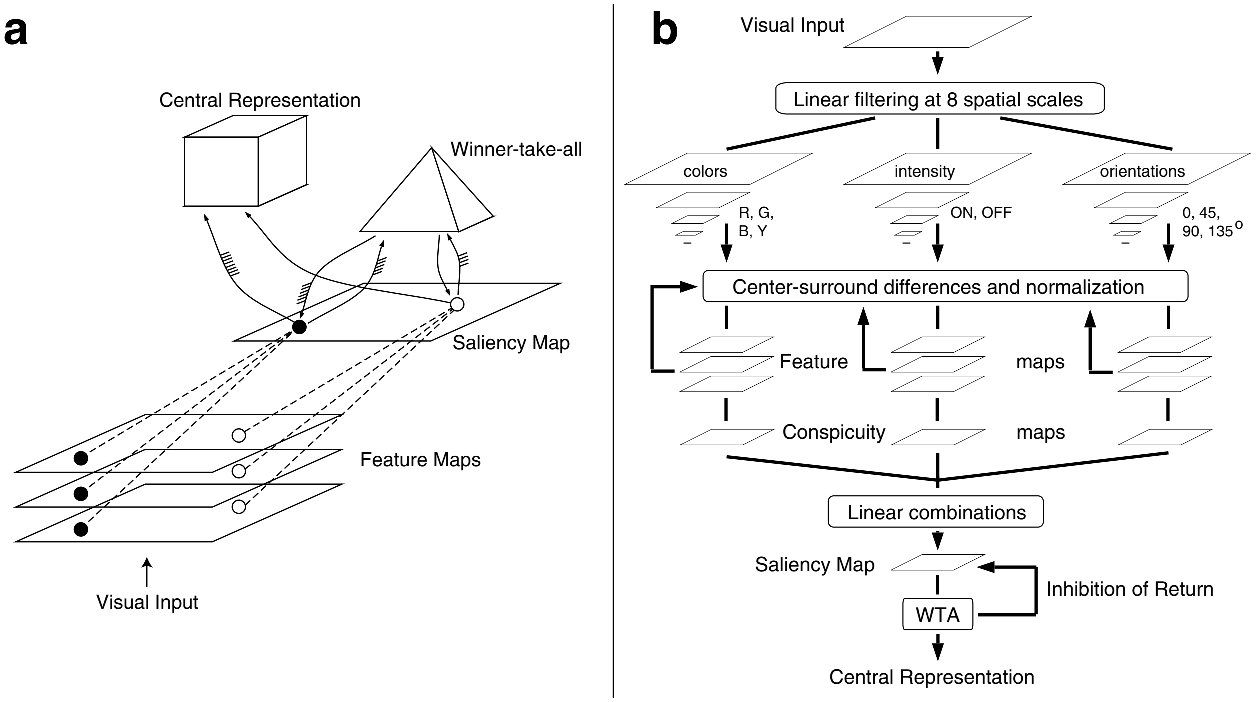

Figure 1: **(a)** Koch and Ullman's (1985) proposal for a computational model of bottom-up attention. Early visual features are computed, in a massively parallel manner, in a set of pre-attentive feature maps receiving retinal input. Activity from all feature maps is combined at each location, giving rise to responses in the topographic saliency map. A winner-take-all network detects the most salient location and directs attention towards it, such that only features from this location are routed towards further analysis and central representation. **(b)** Extended architecture of such scheme proposed by Itti *et al.* (1998). It directly builds on the architecture proposed in **(a)**, but provides a complete implementation of all processing stages. Visual features are computed using linear filtering at eight spatial scales, followed by center-surround differences among six pairs of scales, which compute local spatial contrast in each feature dimension, for a total of 42 maps. An iterative lateral inhibition scheme instantiates non-classical surround competition for salience within each feature map. After competition, feature maps are combined into a single "conspicuity map" for each feature type. The seven conspicuity maps then are summed into the unique topographic saliency map. The saliency map is implemented as a 2D lattice of artificial leaky integrator neurons. The winner-take-all (WTA), also implemented using leaky integrators, detects the most salient location and directs attention towards it. An Inhibition-of-Return mechanism transiently suppresses this location in the saliency map, such that attention is autonomously directed to the next most salient image location.

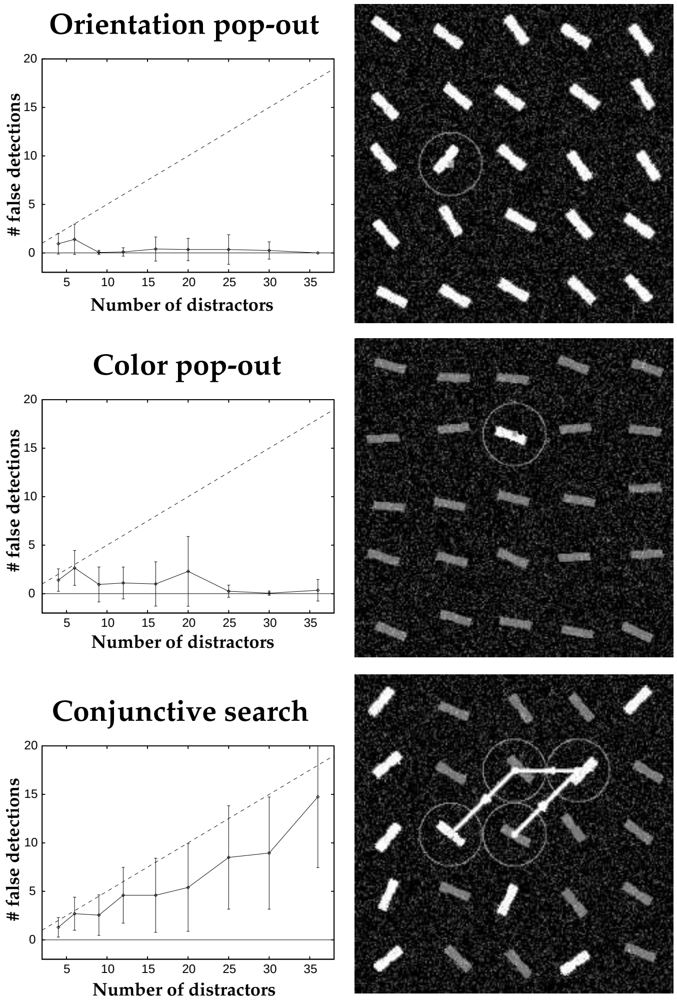

Figure 2: Model performance on noisy versions of pop-out and conjunctive tasks of the type pioneered by Treisman and Gelade (1980). Dark and light bars in the figure represent isoluminant red and green colored bars in the original displays, with strong speckle noise added. Dashed lines show chance value, based on the size of the simulated visual field and the size of the candidate recognition area (corresponds to the performance of an observer who would scan, on average, half of the distractors prior to finding the target). Solid lines (mean±S.D.) show performance of the model. The typical search slopes of human subjects in feature search and conjunction search, respectively, are successfully reproduced by the model: search is easy and slopes are flat in the pop-out cases (when the target differs from the distractors in one feature dimension, e.g., color), while search is more difficult and search time increases with the number of items in the display when the target differs from the distractors by a conjunction of two features (e.g., here the target is the only bar which is both light and tilted counter-clockwise from vertical). Each stimulus was drawn inside a 64×64 pixels box, and the radius of the focus of attention was fixed to 32 pixels. For a fixed number of stimuli, we tested twenty randomly generated images in each task; the saliency map and winner-take-all were initialized to zero (corresponding to a uniformly black visual input) prior to each trial.

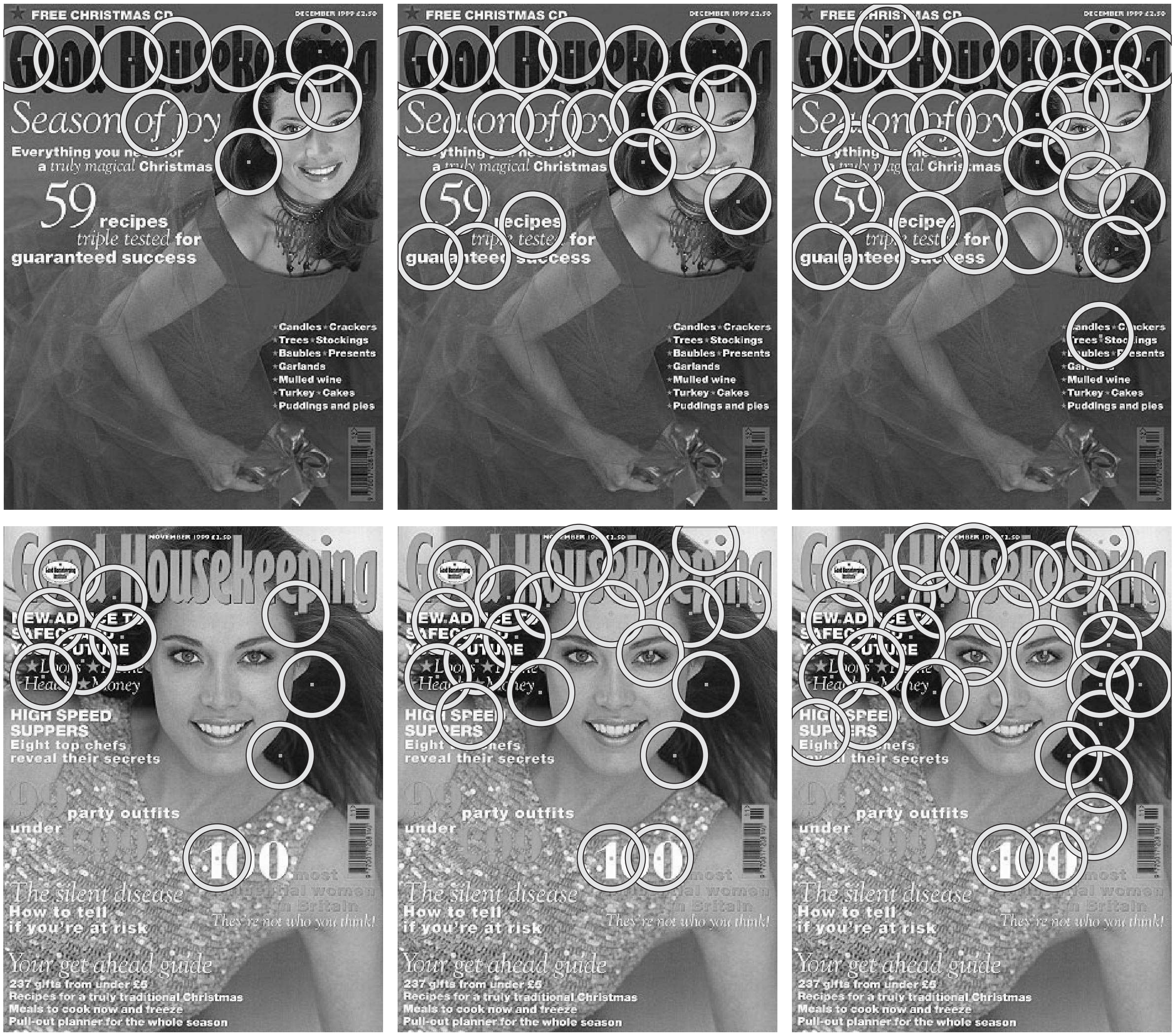

Figure 3: Example application to the quantitative evaluation of advertising designs. Each row shows, from left to right, the first 10, 20 and 30 attention shifts predicted by our model. For the cover shown on the top row, good coverage of all text messages is predicted after 30 shifts, indicating that the text is likely to attract attention and gaze. Under the assumption that these text messages are what sells the magazine, the top cover hence is evaluated as a good design by the model. In contrast, after 30 attention shifts on the cover of the bottom row, large regions of text have remained unvisited by the model. This is not too surprising, as the ignored portions of text are masked by the strongly textured background and are hard to read. Nevertheless, both of the covers shown here have actually been published in two different issues of the British magazine "Good Housekeeping." Hence, current quality control and design selection processes, mostly based on subjective evaluations by focus groups, could benefit from the addition of an objective measure like average saliency of text regions in the images. Photographs courtesy of Peter Walker, McCann-Erickson.