An Integrated Model of Top-down and Bottom-up Attention for Optimizing Detection Speed

Vidhya Navalpakkam
USC
navalpak@usc.edu

Laurent Itti
USC
itti@usc.edu

Abstract

Integration of goal-driven, top-down attention and image-driven, bottom-up attention is crucial for visual search. Yet, previous research has mostly focused on models that are purely top-down or bottom-up. Here, we propose a new model that combines both. The bottom-up component computes the visual salience of scene locations in different feature maps extracted at multiple spatial scales. The top-down component uses accumulated statistical knowledge of the visual features of the desired search target and background clutter, to optimally tune the bottom-up maps such that target detection speed is maximized. Testing on 750 artificial and natural scenes shows that the model's predictions are consistent with a large body of available literature on human psychophysics of visual search. These results suggest that our model may provide good approximation of how humans combine bottom-up and top-down cues such as to optimize target detection speed.

1. Introduction

State of the art in object detection: Traditional models of object detection use a sliding window across the image and apply a binary classifier at each window to detect the presence or absence of the desired target object [17, 13]. While this approach has been successfully applied to detecting rigid objects such as faces and cars [20, 22, 29] and even pedestrians [17, 30], it is slow and computationally expensive as each classifier (corresponding to every object) is run independently at every window within the image.

Role of attention in accelerating detection speed: Recent models of object detection overcome the speed bottleneck of the sliding window approach by using a generic attention operator to quickly select a few interest points in the image [7, 2]. This area has received much interest recently, with several systems using attention as a front-end to accelerate detection speed [10], to reduce complexity of automated multi-target detection and tracking [31], and to enable automated learning and recognition of objects in cluttered scenes [21]. However, most such models are either purely goal-driven (top-down) [19] or image-driven (bottom-up) [9, 27].

Need to integrate top-down and bottom-up attentional influences: There have been few attempts to integrate both top-down and bottom-up attention [16]. Such integration is crucial for robot navigation, visual surveillance and any realistic visual search. For instance, in visual surveillance, it is important to detect goal-relevant targets like suspects, and to simultaneously notice unexpected visual events like gun shots or sudden explosions. Similarly, robot navigation requires top-down detection of landmarks and road signs, as well as bottom-up detection of unexpected obstacles and accidents. In this paper, we present a new model that combines both top-down and bottom-up influences to guide attention during visual search for a target object in distracting clutter.

Need to consider knowledge of the target and distracting background: One of the central challenges in integrating bottom-up and top-down attention is to find the optimal top-down influence on bottom-up processes such that detection speed is maximized. This is an unsolved challenge as yet, since most models of top-down attention are sub-optimal heuristics and driven by knowledge of the desired target only [32, 19, 13, 10], while ignoring the contribution due to knowing the distracting background. Few top-down models consider the distractors [23], by using global features representing the scene context. But they do not consider the local features of the background, that are known to facilitate search [11, 3].

Current open challenges: Further progress in building fast, next generation target detection systems requires a thorough investigation of how statistical knowledge of the local features of the target and distracting background yields optimal top-down attentional signals that combine

with bottom-up attention to maximize detection speed.

Highlight of our approach: We propose a new model that combines both bottom-up as well as top-down attentional influences. Our proposed model first computes the naive, bottom-up salience of every scene location for different local visual features (e.g., different colors, orientations and intensities) at multiple spatial scales. Next, the top-down component uses learnt statistical knowledge of the local features of the target and distracting clutter, to optimize the relative weights of the bottom-up maps such that the overall salience of the target is maximized relative to the surrounding clutter. Such optimization renders the target more salient than the distractors, thereby maximizing target detection speed [33].

Related work: Previously, Navalpakkam and Itti derived a theory of top-down guidance for simplified stimuli defined within one feature dimension only [15]. Here, we present a new model (theory and implementation) that combines bottom-up and top-down attention and considers complex targets and distracting objects that are defined as a conjunction of different features across multiple feature dimensions. Our model is applicable to natural scenes as well as artificial search arrays. Unlike the former study that assumes an ideal observer with complete prior knowledge of the target and distractors, our model allows realistic observers with different beliefs (ranging from no knowledge to complete knowledge), thereby allowing significantly higher prediction power that captures the performance of a novice to an expert.

Our contribution: In section 2, we formally derive the optimal theory of top-down and bottom-up attention. In section 3, we describe the model's implementation and its results on 750 synthetic search arrays and natural scenes. With little computational cost in the form of multiplicative top-down gains on bottom-up saliency maps, we show that our model can predict many reported bottom-up [25, 18, 5, 14, 24, 32, 6, 1] and top-down effects [5, 34, 28, 8] on human visual search behavior. Systematic evaluation of different models with varying degrees of knowledge reveals that knowledge of the local features of the distracting background, in addition to the target, yields better search performance.

2. Theory

Relevant objective function to be optimized: Consider searching for a fruit in the trees. While a ripe red fruit readily captures our attention due to its high visual salience, an unripe green fruit does not capture our attention due to its low salience relative to the distracting leaves, and is hard to detect. Thus, the detection speed depends on the ratio between the strength of signal detecting the target (i.e., target salience), over that detecting the distracting background (i.e., distractor salience) [33]. Here, we will refer to this ratio as the search's signal-to-noise ratio $\mathcal{SNR}$. The relevant goal for maximizing object detection speed is to maximize $\mathcal{SNR}$.

Formalizing visual search: As shown in figure 1, let the perceived salience of the target, $S_T(A)$ be a function of the input search array A, which is a function of the visual features of the target $\Theta|T$ (sampled from probability density functions $P(\Theta|T)$). A is also a function of the relative locations or spatial configuration of the target and distractors (C). Since C and $\Theta|T$ are random variables, so is $S_T(A)$. $S_T(A)$ is also influenced by noise in neural response, η. Similarly, the salience of the distractors, $S_D(A)$, depends on the distractor features $\Theta|D$, configuration C and internal noise η. Thus, we define $\mathcal{SNR}$ as the ratio of expected salience of the target over distractors, with the expectation taken over random variables $\Theta|T, \Theta|D, C, \eta$. $\mathcal{SNR} = E_{\Theta|T,C,\eta}[S_T(A)]/E_{\Theta|D,C,\eta}[S_D(A)]$.

Computing salience within a dimension: The overall perceived salience (combined top-down and bottom-up salience), S_j, for a feature dimension j is computed as a linear combination of the bottom-up saliences s_{ij} for features (values) within that dimension (figure 1). To simulate human like behavior, we assume that the feature responses are modulated in a top-down manner by multiplicative gain modulation [26, 12].

$$S_j(x, y, A) = \sum_{i=1}^{n} g_{ij} s_{ij}(x, y, A) \quad (1)$$

Combining salience across dimensions: To combine information across N feature dimensions, we integrate linearly across all dimensions to obtain the overall perceived salience S (as suggested by the Guided Search theory, [32]).

$$S(x, y, A) = \sum_{j=1}^{N} g_j S_j(x, y, A) \quad (2)$$

Salience of the target and distractors: The expected salience of the target (S_T) can be computed in terms of its salience $s_{ijT}, i \in \{1...n\},\ j \in \{1...N\}$ in each of the n saliency maps within the N feature dimensions. Further, assuming that η, C, and Θ are independent random variables,

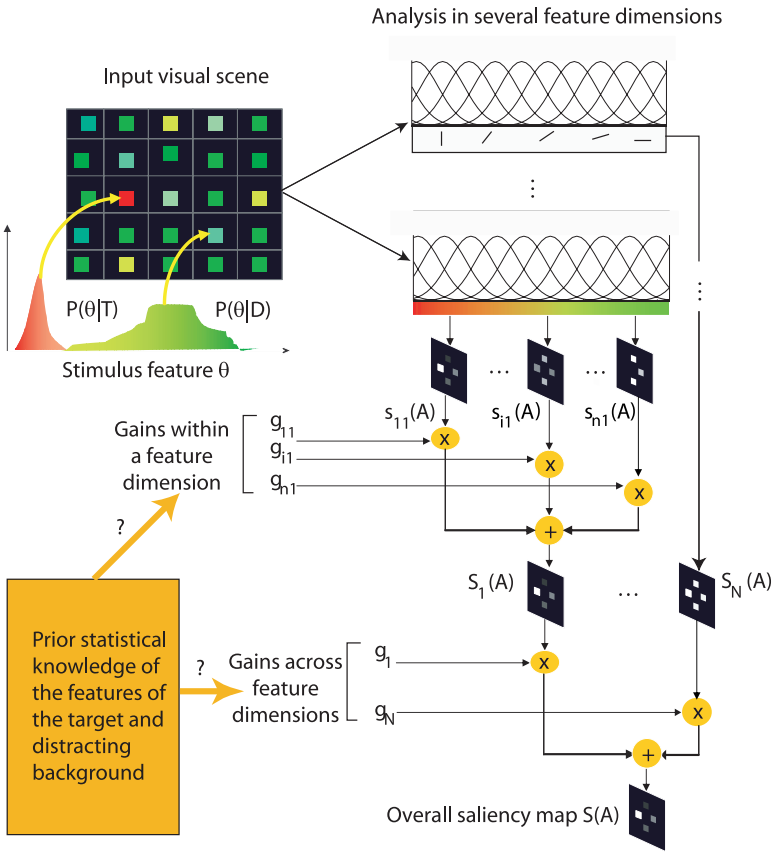

Figure 1. **Overview of our model:** Let the incoming visual scene A contain target and distractors sampled from probability density functions $P(\Theta|T)$ and $P(\Theta|D)$. Our model assumes that the visual input is analyzed in different feature dimensions by a population of neurons with broad and overlapping tuning curves. Bottom-up saliency maps $s_{ij}(A)$ are extracted for the i^{th} feature within the j^{th} dimension, $i \in \{1...n\}$, $j \in \{1...N\}$. Prior knowledge of the target and distractors is used to compute the top-down gains g_{ij} and g_j. The bottom-up maps $s_{ij}(A)$ are then multiplicatively weighted by the top-down gains g_{ij} and are summed to yield $S_j(A)$, the saliency map for the j^{th} dimension. The resulting saliency maps $S_j(A)$ are again weighted by top-down gains g_j and summed across different feature dimensions to form the overall saliency map $S(A)$. The goal here is to choose optimal top-down weights that maximize the target's salience relative to the background, thereby maximizing the speed of detecting the target.

we obtain:

$$\begin{aligned} E[S_T(A)] &= E_{\Theta|T,C,\eta}\left[\sum_{j=1}^{N} g_j S_{jT}(A)\right] \\ &= E_{\Theta|T,C,\eta}\left[\sum_{j=1}^{N} g_j \sum_{i=1}^{n} g_{ij} s_{ijT}(A)\right] \\ &= \sum_{j=1}^{N} g_j \sum_{i=1}^{n} g_{ij} E_{\Theta|T}[E_C[E_\eta[s_{ijT}(A)]]] \end{aligned}$$

Similarly for distractors. Thus, we have,

$$\mathcal{SNR} = \frac{\sum_{j=1}^{N} g_j \sum_{i=1}^{n} g_{ij} E_{\Theta|T}[E_C[E_\eta[s_{ijT}(A)]]]}{\sum_{j=1}^{N} g_j \sum_{i=1}^{n} g_{ij} E_{\Theta|D}[E_C[E_\eta[s_{ijD}(A)]]]} \quad (3)$$

Maximizing SNR to obtain the optimal gains: To maximize $\mathcal{SNR}$, we differentiate it wrt g_{ij} and g_j and obtain the following:

$$\frac{\partial}{\partial g_{ij}}\mathcal{SNR} = \frac{\frac{\mathcal{SNR}_{ij}}{\mathcal{SNR}} - 1}{\alpha_{ij}} \quad (4)$$

$$\frac{\partial}{\partial g_j}\mathcal{SNR} = \frac{\frac{\mathcal{SNR}_j}{\mathcal{SNR}} - 1}{\alpha_j} \quad (5)$$

where α_{ij}, α_j are positive normalization terms and

$$\mathcal{SNR}_{ij} = \frac{E_{\Theta|T}[E_C[E_\eta[s_{ijT}(A)]]]}{E_{\Theta|D}[E_C[E_\eta[s_{ijD}(A)]]]} \quad (6)$$

$$\mathcal{SNR}_j = \frac{E_{\Theta|T}[E_C[E_\eta[S_{jT}(A)]]]}{E_{\Theta|D}[E_C[E_\eta[S_{jD}(A)]]]} \quad (7)$$

The sign of the derivative $\frac{\partial}{\partial g_{ij}}\mathcal{SNR}$ determines whether g_{ij} should increase, decrease or remain at the baseline ($g_{ij} = 1$), in order to maximize $\mathcal{SNR}$. Eqn. 4 yields:

$$\begin{aligned} \frac{\mathcal{SNR}_{ij}}{\mathcal{SNR}} &< 1 \Rightarrow \left(\frac{\partial}{\partial g_{ij}}\mathcal{SNR}\right)_{g_{ij}=1} < 0 \Rightarrow g_{ij} < 1 \\ &= 1 \Rightarrow \left(\frac{\partial}{\partial g_{ij}}\mathcal{SNR}\right)_{g_{ij}=1} = 0 \Rightarrow g_{ij} = 1 \\ &> 1 \Rightarrow \left(\frac{\partial}{\partial g_{ij}}\mathcal{SNR}\right)_{g_{ij}=1} > 0 \Rightarrow g_{ij} > 1 \end{aligned}$$

Thus g_{ij} increases as $\frac{\mathcal{SNR}_{ij}}{\mathcal{SNR}}$ increases. We simplify this monotonic relationship by assuming proportionality. With an added constraint that the gains cannot increase indiscrim-

inately, but must sum to a constant, $\sum_{i=1}^{n} g_{ij} = n$, we get:

$$g_{ij} = \frac{\mathcal{SNR}_{ij}}{\frac{1}{n}\sum_{k=1}^{n} \mathcal{SNR}_{kj}} \quad (8)$$

$$g_j = \frac{\mathcal{SNR}_j}{\frac{1}{N}\sum_{k=1}^{N} \mathcal{SNR}_k} \text{ (similarly)} \quad (9)$$

Interpretation of the result: Thus, the top-down weight on the i^{th} visual feature in the j^{th} feature dimension depends on its signal-to-noise ratio $\mathcal{SNR}_{ij}$, over the mean in that dimension. Similarly, the top-down gain on the j^{th} feature dimension depends on its signal-to-noise ratio $\mathcal{SNR}_j$, over the mean across all dimensions. In other words, a feature is relevant and receives a high weight if it renders the target more salient than the distractors, and is irrelevant otherwise.

Ideal observer vs. real observer: The current analysis considers an ideal observer who knows the true underlying distribution of the target and distractor features ($\theta|T, \theta|D$). But in reality, the observer may possess incomplete knowledge or a different belief ($\theta^b|T, \theta^b|D$). This belief may be learnt through observation of several displays (bottom-up priming [11]), or through explicit verbal instruction such as "find the red object" [34, 28]. Our model captures such top-down influences in the following manner: A forward internal model translates the observer's belief in feature space ($\theta^b|T, \theta^b|D$) into a belief in salience of the target and distractors (S_T^b, S_D^b), which is then used to derive the belief in signal-to-noise ratio ($\mathcal{SNR}^b$). Top-down gains are chosen according to eqns. 8 and 9, thereby optimizing $\mathcal{SNR}^b$. These gains g_{ij} are then applied to the bottom-up saliency maps (s_{ij}) within each feature dimension to compute the biased saliency maps S_j, which are multiplied by the gains g_j to obtain the overall saliency map S. Thus, the bottom-up saliency maps are combined with the optimal top-down gains to yield a saliency map where the target's salience is maximized relative to the distractors. This saliency map is now used to guide attention to likely target locations.

3. Results

In this section, we present a systematic evaluation of the model's predictions for different observer beliefs, and search tasks on artificial search arrays and natural scenes.

Computing salience: For computing bottom-up saliency maps, we use the Itti and Koch saliency model [9]. We use the following set of biologically inspired, low-level visual features: 6 hues within the color dimension, 4 intensities within the luminance dimension, and 4 orientations ($0°, 45°, 90°, 135°$) within the orientation dimension. The input visual scene is analyzed in all feature dimensions in parallel and for each of the above features, feature maps (topographic maps of feature responses at all scene locations) are extracted in 6 different spatial scales (downsized by a factor of 1, 2, 4, 8, 16, and 32). After local center-surround feature contrast operations, and global nonlinear interactions in space, these maps are weighted by the top-down gains (whose baseline is unity) and are linearly combined into a conspicuity map for that feature dimension. The conspicuity maps are also weighted by top-down gains (default weight is 1) and are combined linearly to obtain the overall saliency map. The active locations in this map indicate likely target locations.

Interpreting $\mathcal{SNR}$: The $\mathcal{SNR}$ in the overall saliency map may be high and search may be efficient due to high bottom-up salience of the target relative to the distractors (e.g., a green target pops out among red distractors [25] as $s_{ijT} >> s_{ijD}$ in at least one saliency map i in dimension j), or due to efficient top-down guidance to the target (e.g., a green target among randomly colored distractors becomes easy to find once subjects know that the target is green [4] since $g_{ij} >> 1$ on the green feature), or both.

Comparison of four models: We test the predictions of the above theory by implementing four different models: T0D0, T1D0, T0D1 and T1D1, where T and D refer to the target and distractors. T0D0, the naive, bottom-up model [9] does not know T or D (hence, uses default top-down weights of 1). T1D0 combines bottom-up salience with knowledge of T only. Hence, it computes top-down weights based only on target salience s_{ijT}, while ignoring D by considering s_{ijD} to be some constant. T0D1 combines bottom-up salience with knowledge of D only. T1D1 combines bottom-up salience and top-down knowledge of both T and D. It chooses weights according to eqns. 8 and 9.

Training and test data: We compare the performance of the above models on synthetic search array stimuli used in psychophysics tasks (to study human behavior in a controlled and simplified environment), as well as in real-world natural scenes with complex stimuli. For each search condition with the synthetic stimuli, the model learns the belief in salience (S_T^b, S_D^b) from 50 training images, computes the mean salience of the target and distractors ($E_{\Theta|T,C,\eta}[S_T^b(A)], E_{\Theta|D,C,\eta}[S_D^b(A)]$) and uses it to compute gains (g_{ij}, g_j), that are subsequently applied on 100 new, previously unseen test images. In each of these images, the target and distractors can occur randomly at any cell within the 9x9 grid, and their location within the cells is further jittered by upto 10 pixels (thereby changing C). Noise in stimulus features is also added, in the form of jitter in orientation (upto $5°$), and jitter in color values (upto 20 in R,G and B), thereby changing $\Theta|T, \Theta|D$. Internal neural noise η is added by the saliency model. Results are reported in figure 2a-i.

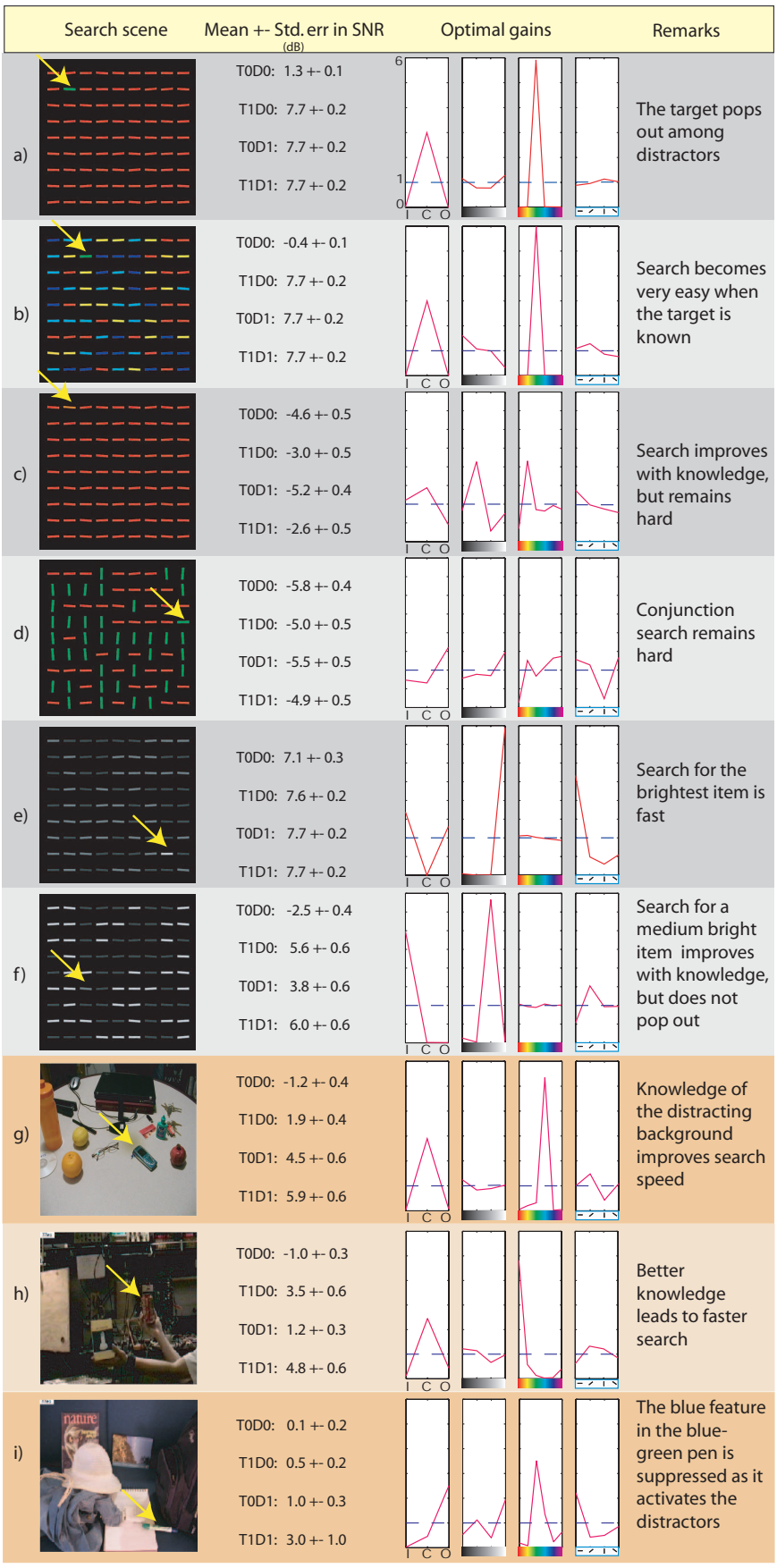

Figure 2. **Simulation results:** This figure shows the results of testing on 750 artificial search arrays and natural scenes. Each row shows a different search task with different targets and distractors. The first column shows a sample test scene. The second column shows the $\mathcal{SNR}$ (in decibels) predicted by four different models described in section 3. The third column shows the distribution of optimal top-down gains derived from knowledge of the target and distractors, as computed by model T1D1. The dotted blue lines are the default gains (1) used by model T0D0. The first plot shows the gains on the intensity (I), color (C) and orientation (O) dimensions. The subsequent plots show the gains within these dimensions (in the order of intensity, color and orientation). The final column shows some remarks. As described in section 3.1, these results are consistent with bottom-up and top-down effects reported in psychophysics experiments. Across all search tasks, model T1D1 performed atleast as good as or better than T1D0, T0D1, which performed better than T0D0. These results suggest that knowledge of both the target and the distracting background plays an important role in improving search speed.

3.1. Artificial search arrays

Pop-out search: Figure 2a shows an example of a pop-out search with a green target among red distractors. The naive bottom-up saliency model T0D0 predicts that $\mathcal{SNR}$ will be reasonably high (1.3±0.1dB), indicating that search will be fast [25]. Consistent with psychophysics experiments on "priming of pop-out" [11], knowledge of the target and distractors allows the relevant features to be primed and hence increases $\mathcal{SNR}$ to 7.7±0.2 dB, leading to a faster search. The distribution of optimal gains shows that the gain on the color dimension increases while suppressing intensity and orientation; and within color, the target's green feature is maximally boosted while the distracting red feature and other irrelevant features are suppressed.

Distractor heterogeneity: Figure 2b shows an example of search for a green target among heterogeneous distractors of different colors. As observed in human visual search behavior [5], the naive model T0D0 predicts a hard search ($\mathcal{SNR}$ is -0.4±0.1 dB). But psychophysics experiments [5] also show that this hard search becomes efficient if the target is known, consistent with the prediction of models T1D0 and T1D1. Note that in both figures 2a and 2b, the target and distractor features are well separated in feature space, hence, the optimal gains reduces to increased gain on target features and suppression of others.

Poor target-distractor discriminability: Figure 2c shows an example of search for an orange target that is less discriminable from the red distractors. The naive model T0D0 correctly predicts a very hard search [18, 5, 14, 24, 32] ($\mathcal{SNR}$ is -4.6±0.5 dB). Simply knowing the target feature is not so helpful, since boosting the target's red feature also activates all the distractors that share that feature. Instead, model T1D1 that knows both the target and distractors performs better as it promotes the yellow feature that selectively activates the target, while suppressing the red feature that activates the distractors.

Conjunction search: Figure 2d shows conjunction search for a green-horizontal target among green-vertical and red-horizontal distractors. The naive model T0D0 correctly predicts a very hard search [25] ($\mathcal{SNR}$ is -5.8±0.4 dB). Extra knowledge allows model T1D1 to slightly improve search by promoting the target's horizontal feature, while suppressing the distractor's red feature. But consistent with psychophysics experiments, search remains hard.

Linearly separable target: Figure 2e shows search for a bright target among medium-bright and dark distractors. Search is easy ($\mathcal{SNR}$ is 7.1±0.3 dB), confirming earlier reports of easy search for a target that can be separated from distractors by a line in feature space [6, 1]. T1D1 suggests higher gain on intensity dimension, and within intensity, higher gain on the high intensity values than others.

Figure 3. Example training data

Non-linearly separable target: Figure 2f shows search for a medium-bright target among dark an bright distractors. The naive model T0D0 predicts hard search ($\mathcal{SNR}$ is -2.5±0.4 dB) confirming the "linear-separability effect" [6, 1] that search for a medium-type target that cannot be linearly separated from distractors is harder than when the target is linearly separable (as shown in 2e). Consistent with previous experiments, there is a top-down effect of knowledge leading to faster search [8]. In this case, model T1D1 suggests increased gain in the medium intensity value and suppression of high and low intensity values (corresponding to boosting the target and suppressing the distractors).

3.2. Natural scenes

Training and testing: To test the model's performance on natural scenes, we train it on 10 images containing different views of the target, appearing at different locations in the scene. Some examples are shown in figure 3. The learned top-down gains are subsequently applied on 50 new test scenes where the target can appear in slightly different backgrounds, different locations, views and sizes.

Search for targets in natural scenes: The results for finding a cell phone on a cluttered desk are shown in figure 2g. While the naive model T0D0 struggles to find the non-salient phone ($\mathcal{SNR}$ is -1.2±0.4 dB), knowledge of the phone and the distracting background (through training) speeds search significantly (model T1D1 yields an $\mathcal{SNR}$ of 5.9±0.6 dB). Inspection of the gains reveal that color is the useful dimension and within color, the target's blue feature discriminates it best from the background. Similar results

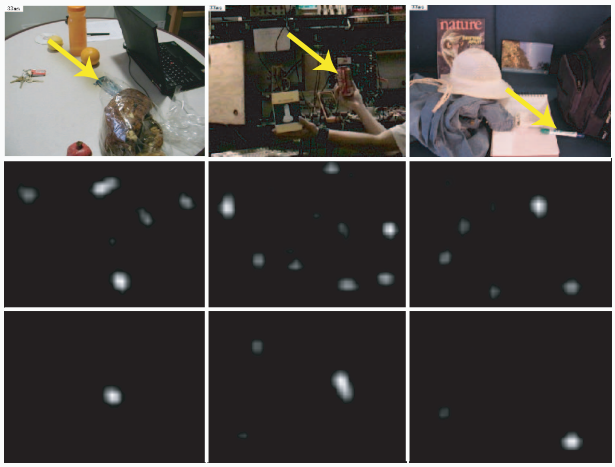

Figure 4. Comparison of saliency maps of the naive bottom-up model T0D0 (second row) vs. T1D1 (third row) are shown during search for a phone on a desk (first column), a coke can in a cluttered scene (second column), and a pen in a distracting background (third column). Although the target is not bottom-up salient, prior knowledge of the target and the distracting background (acquired through training) helps in improving the $\mathcal{SNR}$, thereby rendering the target more salient and suppressing noisy activity due to the distractors.

are shown during search for a pen (figure 2i) and coke can (figure 2h) in distracting backgrounds. Figure 4 shows sample saliency maps for further comparison between the naive model (T0D0) and the combined top-down and bottom-up model (T1D1).

4. Discussion

Summary of results: By integrating top-down, knowledge-driven and bottom-up, image-driven approaches, we account for a large body of visual search literature. All models successfully account for fast search in pop-out tasks (e.g., green target pops out among red distractors) [25], slow search in conjunction tasks (e.g., green vertical target among green horizontal and red vertical distractors) [25], slow search when the target is more similar to the distractors (e.g., orange target among red distractors) [18, 5, 14, 24, 32], and faster search for an extreme feature valued target than a medium valued target (e.g., faster search for a bright target among dark and medium distractors, while slower search for a medium target among dark and bright distractors) [6, 1]. In addition, knowledge based models T1D0 and T1D1 also account for fast search for a known target among heterogeneous distractors, while the naive model T0D0 indicates a slow search (e.g., search for a green target among red, yellow and blue is slow if we dont know that the target is green, and is otherwise fast)[5].

Better knowledge leads to faster search: Systematic comparison of different models reveals that model T1D1 performs significantly better than models T1D0, T0D1, which perform better than T0D0. Thus, we provide a computational correlate for the behavioral effect that better knowledge leads to faster search [34, 28]. The gradual progression of models from T0D0 (no knowledge), to T1D1 (complete knowledge of both target and distractors) allows us to capture the behavior of novices to experts, such as due to priming [11].

Role of knowledge of the distracting background: Contrary to previous target based approaches which assume that knowledge of the target suffices [32, 19], we suggest that knowledge of the distractors is also crucial. Hence, while the former approaches suggest that the target features always be promoted, as shown in figure 2c, our model predicts that the target features may even be suppressed if the distractor activates the same features. Similar examples are also shown in figure 2h, where the blue feature in the blue-green pen is suppressed as it activates the background. Thus, distractors and not just the target, play an important role in priming features so as to maximize target detection speed.

Integration of bottom-up and top-down attentional influences: With little computational cost incurred through multiplicative top-down weights on bottom-up saliency maps, our model combines both stimulus-driven and goal-driven attention, to optimize speed of guidance to likely target locations, while simultaneously being sensitive to unexpected stimulus changes. As mentioned earlier, this is an important ability for robot navigation, visual surveillance and other active vision tasks that operate in unconstrained environments where unexpected visual events such as accidents may occur.

Future extensions: We currently consider simple, low-level visual features such as intensities, hues and orientations at multiple spatial scales. But the theory derived in section 2 is general and can be applied to any feature dimension, such as complex shape features.

References

[1] B. Bauer, P. Jolicoeur, and W. B. Cowan. Visual search for colour targets that are or are not linearly separable from distractors. *Vision Res*, 36(10):1439–1465, May 1996.

[2] G. Bouchard and B. Triggs. Hierarchical part-based visual object categorization. *In Proc. CVPR*, pages 710–715, 2005.

[3] J. J. Braithwaite and G. W. Humphreys. Inhibition and anticipation in visual search: evidence from effects of color

foreknowledge on preview search. *Percept Psychophys*, 65(2):213–237, Feb 2003.

[4] J. Duncan. Boundary conditions on parallel processing in human vision. *Perception*, 18(4):457–469, 1989.

[5] J. Duncan and G. W. Humphreys. Visual search and stimulus similarity. *Psychological Rev*, 96:433–458, 1989.

[6] M. D'Zmura. Color in visual search. *Vision Research 31*, 6:951–966, 1991.

[7] R. Fergus, P. Perona, and A. Zisserman. A sparse object category model for efficient learning and exhaustive recognition. *In Proc. CVPR*, pages 380–387, 2005.

[8] J. Hodsoll and G. W. Humphreys. Driving attention with the top down: the relative contribution of target templates to the linear separability effect in the size dimension. *Percept Psychophys*, 63(5):918–926, Jul 2001.

[9] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12):1489–1506, May 2000.

[10] L. Itti and C. Koch. Feature combination strategies for saliency-based visual attention systems. *Journal of Electronic Imaging*, 10(1):161–169, Jan 2001.

[11] V. Maljkovic and K. Nakayama. Priming of pop-out: I. role of features. *Mem Cognit*, 22(6):657–672, Nov 1994.

[12] J. C. Martinez-Trujillo and S. Treue. Feature-based attention increases the selectivity of population responses in primate visual cortex. *Curr Biol*, 14(9):744–751, May 2004.

[13] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Trans. PAMI*, 19(7):696–710, 1997.

[14] A. L. Nagy and R. R. Sanchez. Critical color differences determined with a visual search task. *Journal of the Optical Society of America A 7*, 7:1209–1217, 1990.

[15] V. Navalpakkam and L. Itti. Optimal cue selection strategy. *NIPS, Vol. 19*, pages 1–8, 2006.

[16] A. Oliva, A. Torralba, M. S. Castelhano, , and J. M. Henderson. Top-down control of visual attention in object detection. *IEEE Proc. ICIP*, pages September 14–17, 2003.

[17] C. Papageorgiou and T. Poggio. A trainable system for object detection. *Int. J. Comput. Vision*, 38(1):15–33, 2000.

[18] H. Pashler. Target-distractor discriminability in visual search. *Percept Psychophys*, 41(4):385–392, Apr 1987.

[19] R. P. Rao, G. Zelinsky, M. Hayhoe, and D. H. Ballard. Eye movements in iconic visual search. *Vision Research*, 42(11):1447–1463, Nov 2002.

[20] H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Trans. PAMI*, 20(1):23–38, 1998.

[21] U. Rutishauser, D. Walther, C. Koch, and P. Perona. Is bottom-up attention useful for object recognition?, 2004.

[22] H. Schneiderman and T. Kanade. A statistical method for 3d object detection applied to faces and cars. pages 746–751, 2000.

[23] A. Torralba and P. Sinha. Statistical context priming for object detection. volume 1, pages 763–770, 2001.

[24] A. Treisman. Search, similarity, and integration of features between and within dimensions. *J Exp Psychol Hum Percept Perform*, 17(3):652–676, Aug 1991.

[25] A. Treisman and G. Gelade. A feature integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

[26] S. Treue and J. C. M. Trujillo. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, 399(6736):575–579, Jun 1999.

[27] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. H. Lai, N. Davis, and F. Nuflo. Modeling visual-attention via selective tuning. *Artificial Intelligence*, 78(1-2):507–45, 1995.

[28] T. J. Vickery, L.-W. King, and Y. Jiang. Setting up the target template in visual search. *J Vis*, 5(1):81–92, Feb 2005.

[29] P. Viola and M. J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57(2):137–154, 2004.

[30] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *Int. J. Comput. Vision*, 63(2):153–161, 2005.

[31] D. Walther, D. Edgington, and C. Koch. Detection and tracking of objects in underwater video., 2004.

[32] J. M. Wolfe. Guided search 2.0: a revised model of visual search. *Psyonomic Bulletin and Review*, 1(2):202–238, 1994.

[33] J. M. Wolfe, S. J. Butcher, and M. Hyle. Changing your mind: On the contributions of top-down and bottom-up guidance in visual search for feature singletons. *J Exp Psychol Hum Percept Perform*, 29(2):483–502, 2003.

[34] J. M. Wolfe, T. S. Horowitz, N. Kenner, M. Hyle, and N. Vasan. How fast can you change your mind? The speed of top-down guidance in visual search. *Vision Res*, 44(12):1411–1426, Jun 2004.