Top-down Attention Signals in Saliency

WORKS BY VIDHYA NAVALPAKKAM

邓凝旖

2014.11.10

Introduction of Vidhya Navalpakkam

EDUCATION

* Ph.D, Computer Science, Fall 2006, University of Southern California (USC), Los Angeles, CA; Advisor: Dr. Laurent Itti

* B.Tech, Computer Science, Fall 2001, Indian Institute of Technology (IIT), Kharagpur, India

EXPERIENCE

* Google, Research Scientist, May 2012-present

* Yahoo! Research, Research Scientist, Jul 2010-May 2012

* Caltech, Biology, Postdoctoral Research Scholar, Jan 2007-Jul 2010; Advisors: Dr. Pietro Perona, Dr. Christof Koch

* Stanford CS, Visiting Postdoctoral Scholar, Aug 2009-Jul 2010; Host: Dr. Fei-Fei Li

Top-down Attention Selection is Fine Grained

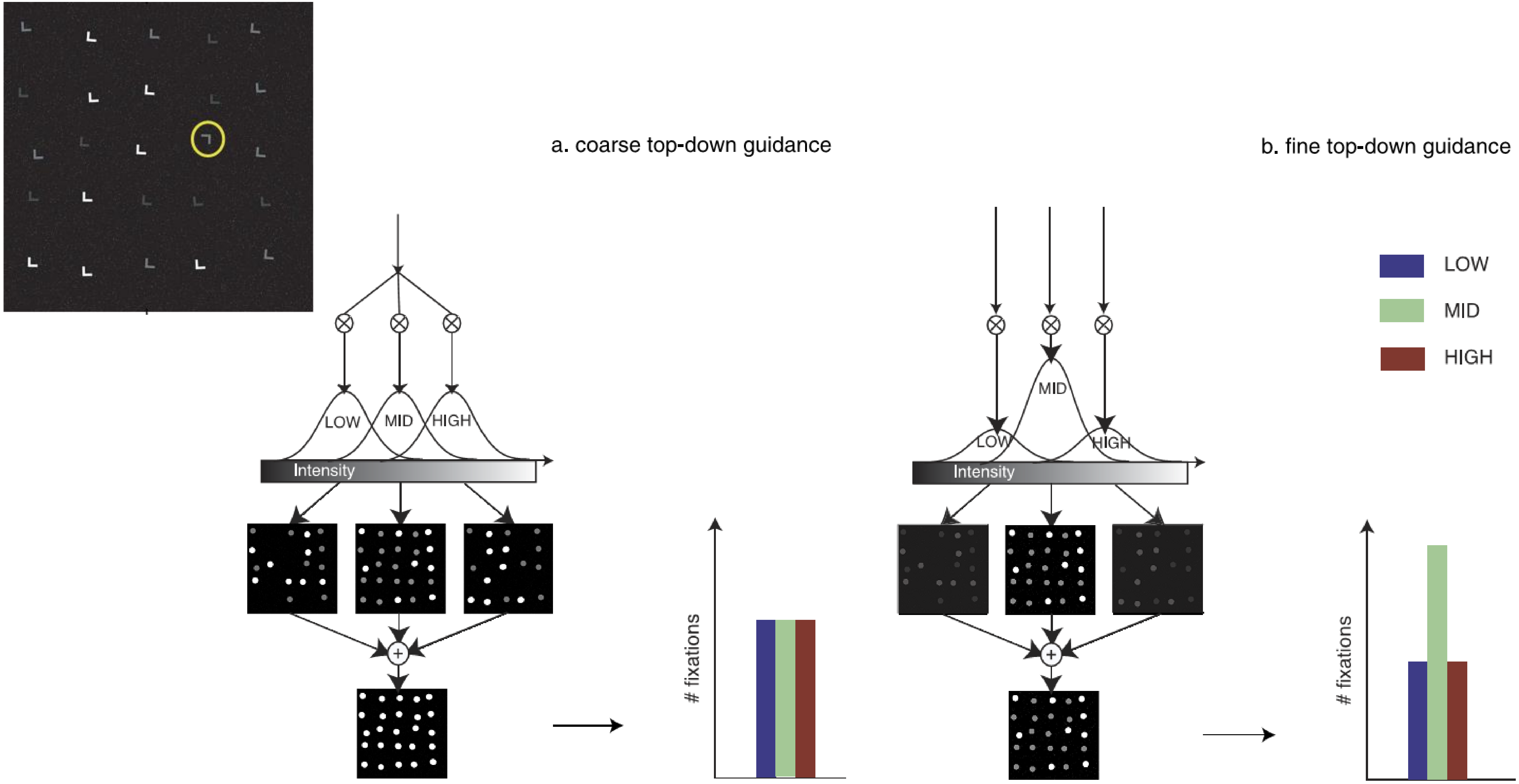
VIDHYA NAVALPAKKAM & LAURENT ITTI

JOURNAL OF VISION 2006

Importance of top-down signals

In natural world, when predators are camouflaged and, hence, visually nonsalient, the prey’s survival depends on whether top-down can guide attention by selecting the fine-grained target feature

© Scott K Macleod REX



a. coarse top-down guidance

b. fine top-down guidance

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

VIDHYA NAVALPAKKAM & LAURENT ITTI

CVPR2006

Background

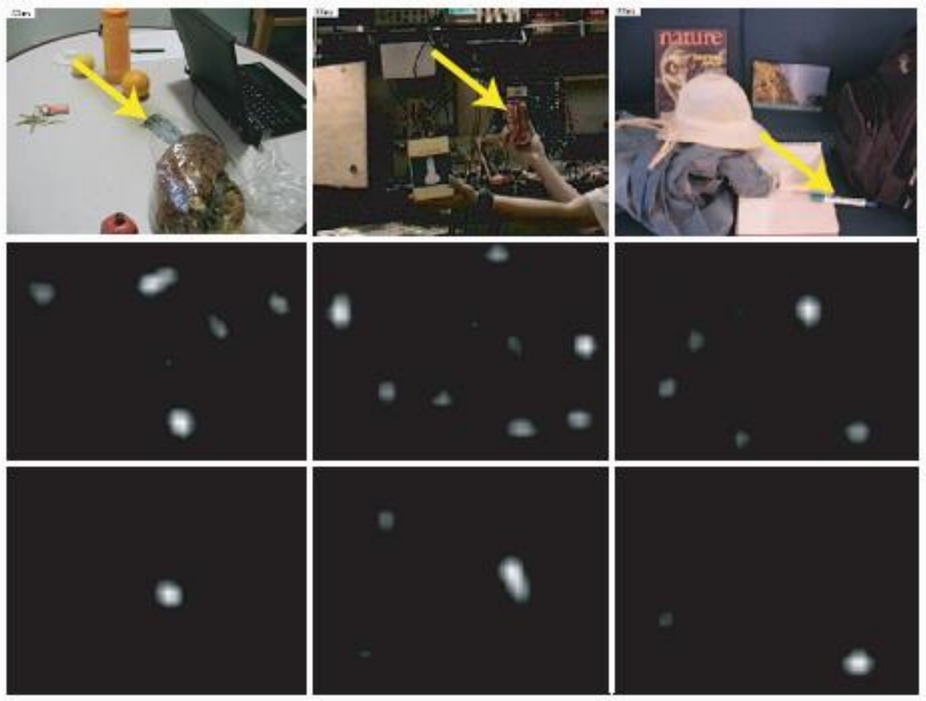
Use attention to accelerate detection speed

Need to integrate top-down and bottom-up attentional influences

Need to consider knowledge of the target and distracting background

Goal

Get the saliency map of target



Approach

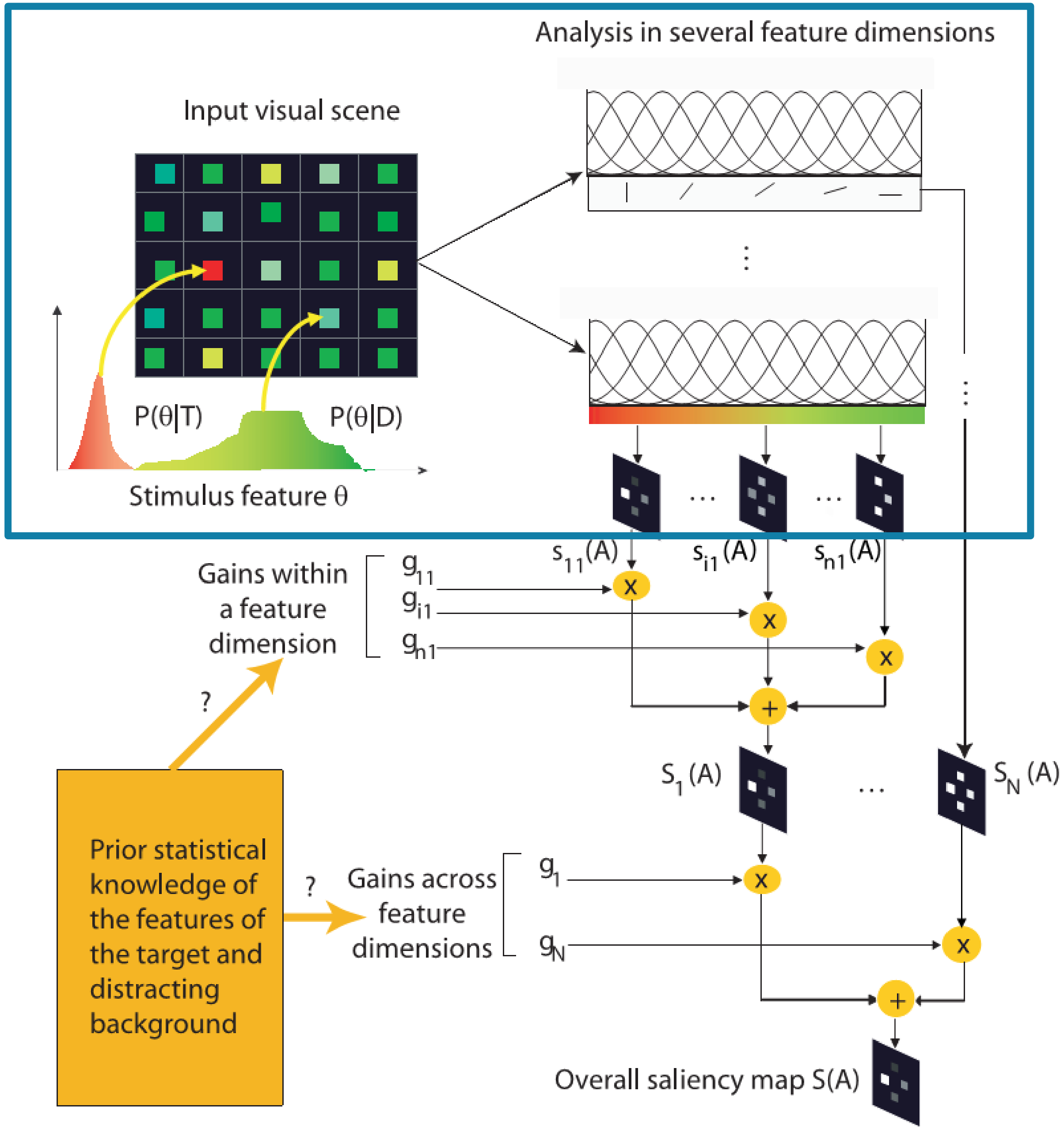
Propose a new model that combines both bottom-up as well as top-down attentional influences

The 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

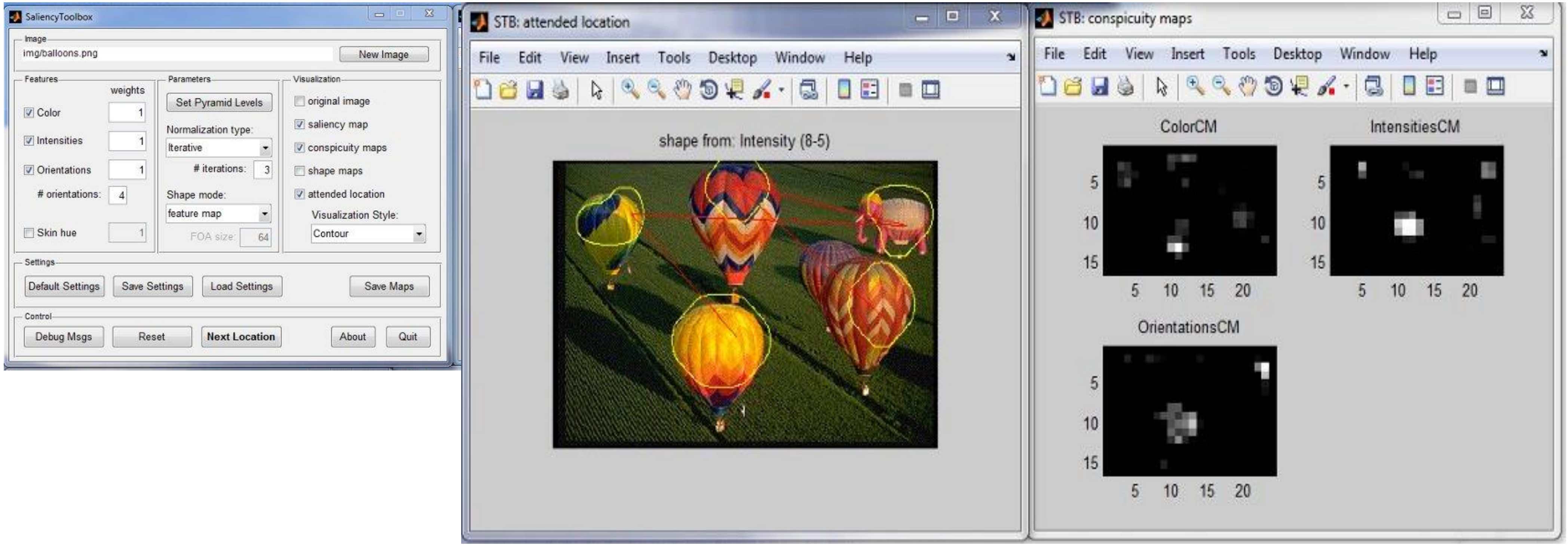
Such optimization renders the target more salient than the distractors, thereby maximizing target detection speed

Bottom-up Saliency Map


Analysis in several feature dimensions
Input visual scene
P(θ|T)
P(θ|D)
Stimulus feature θ
s_{11}(A)
s_{i1}(A)
s_{n1}(A)
Gains within a feature dimension
g_{11}
g_{i1}
g_{n1}
?
S_1(A)
S_N(A)
Prior statistical knowledge of the features of the target and distracting background
?
Gains across feature dimensions
g_1
g_N
x
+
Overall saliency map S(A)

Saliency model by Itti

L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. PAMI 1998.



SaliencyToolbox

Image: img/balloons.png

Features

	weights
☑ Color	1
☑ Intensities	1
☑ Orientations	1
# orientations:	4
□ Skin hue	1

Parameters

Normalization type: Iterative

iterations: 3

Shape mode: feature map

FOA size: 64

Visualization

□ original image

☑ saliency map

☑ conspicuity maps

□ shape maps

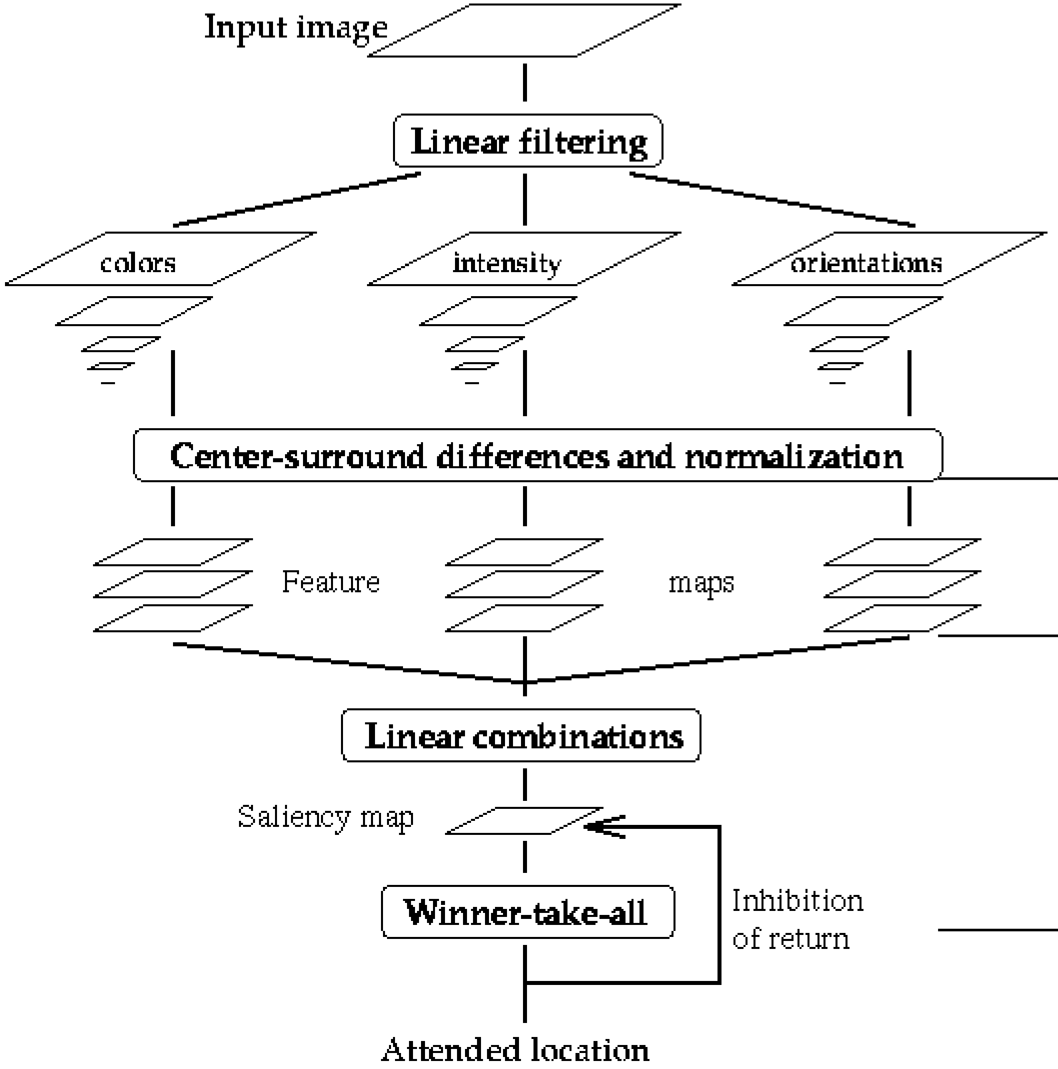
☑ attended location

Visualization Style: Contour

STB: attended location

shape from: Intensity (8-5)

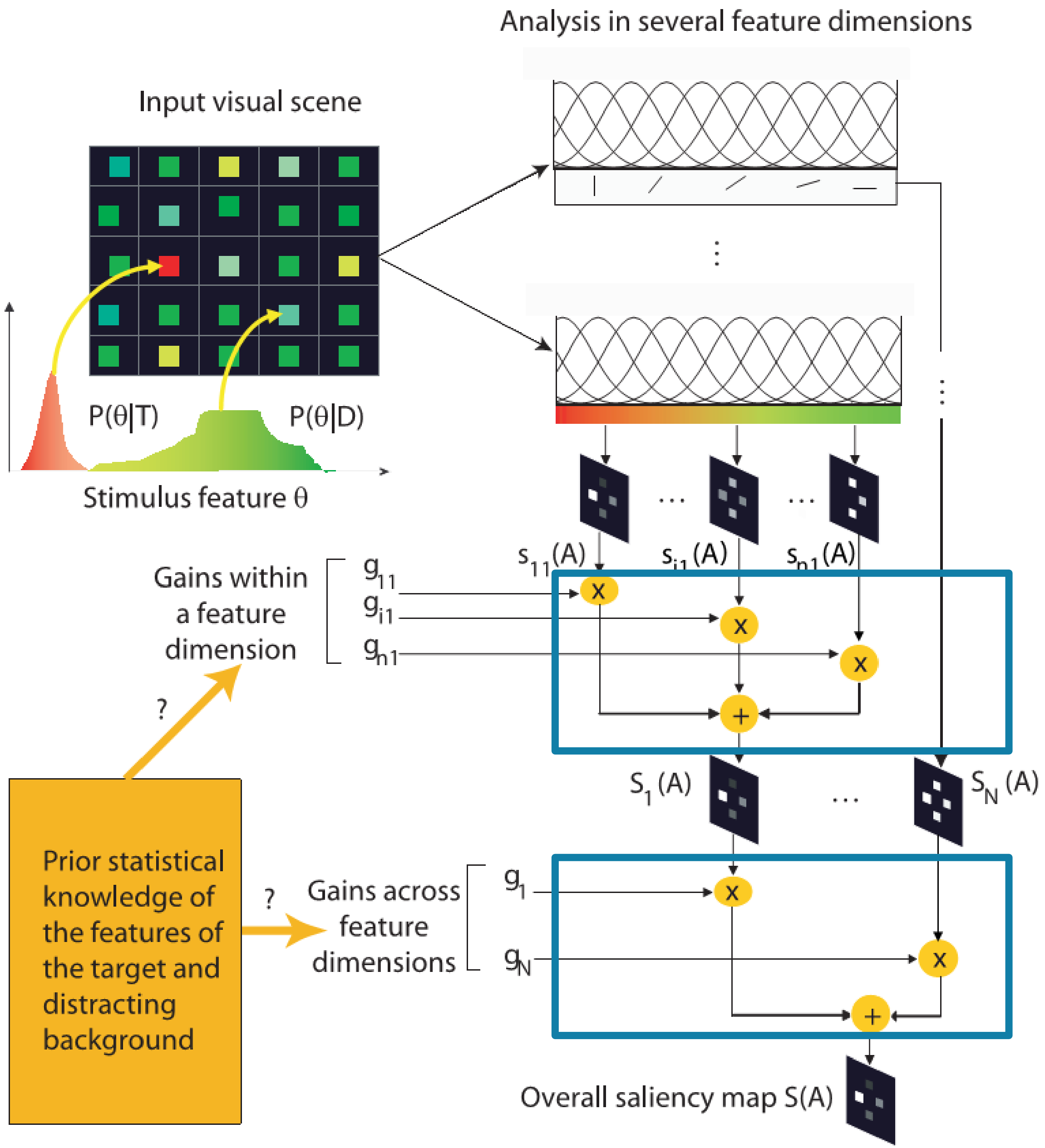
STB: conspicuity maps



Differences between feature pyramid

Normalization

Gauss Convolution


Analysis in several feature dimensions
Input visual scene
P(θ|T)
P(θ|D)
Stimulus feature θ
s_{11}(A)
s_{i1}(A)
s_{n1}(A)
Gains within a feature dimension
g_{11}
g_{i1}
g_{n1}
x
+
?
S_1(A)
S_N(A)
Prior statistical knowledge of the features of the target and distracting background
Gains across feature dimensions
g_1
g_N
Overall saliency map S(A)

Top-down Gains

Relevant objective function to be optimized

SNR

- 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)
- The relevant goal for maximizing object detection speed is to **maximize signal-to-noise ratio** SNR
- ST(A)be a function of the input search arrayA, which is a function of the visual features of the target Θ|T(sampled from probability density functions P(Θ|T)). A is also a function of the relative locations or spatial configuration of the target and distractors (C). Since C and Θ|T are random variables, so is ST(A). ST(A)is also influenced by noise in neural response, η. And the same for the salience of the distractors, SD(A)

Relevant objective function to be optimized

SNR: the ratio of expected salience of the target over distractors

$$\mathcal{SNR} = E_{\Theta|T,C,\eta}[S_T(A)]/E_{\Theta|D,C,\eta}[S_D(A)]$$

Salience within a dimension

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

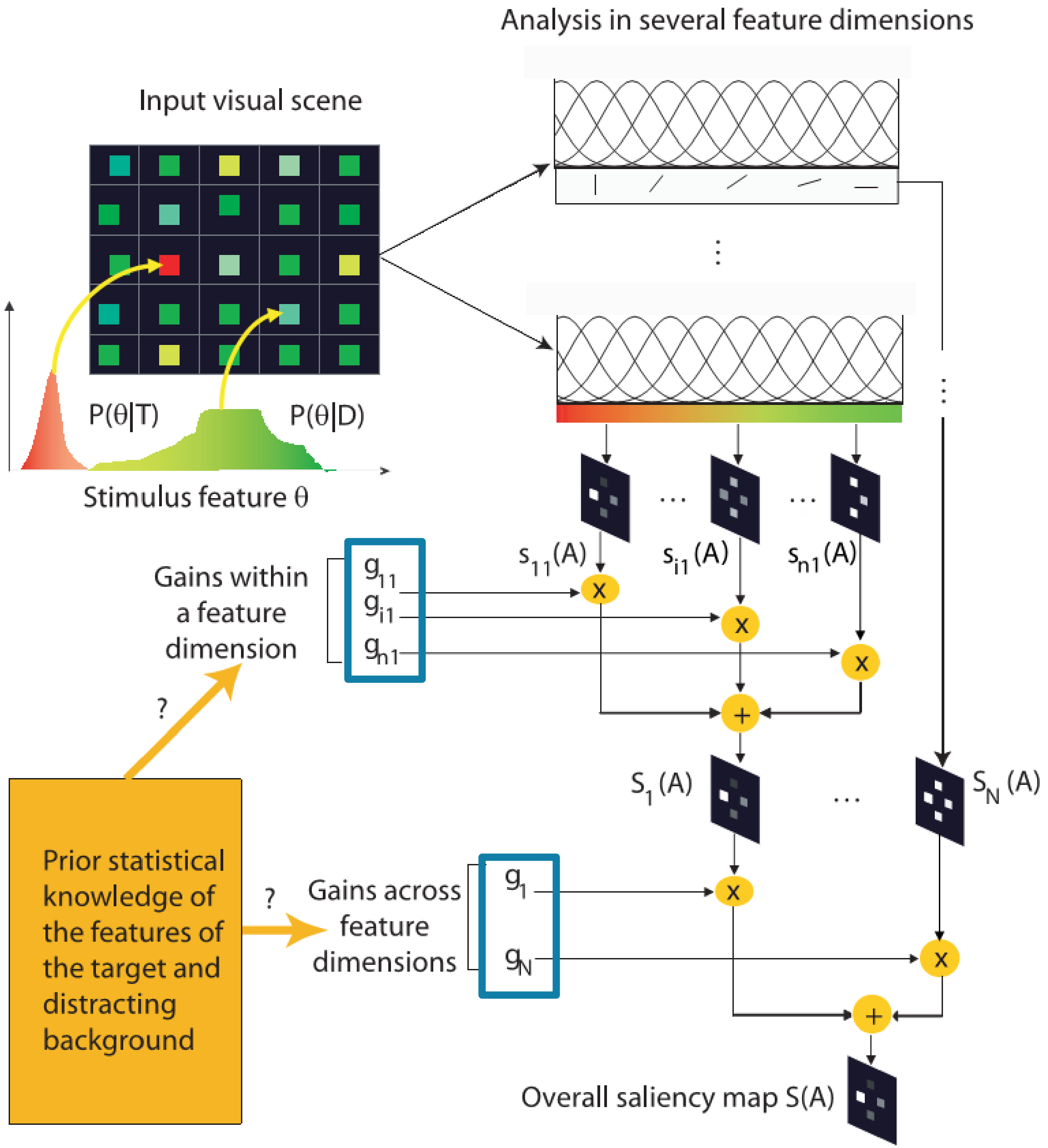
Salience across dimensions

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

The expected salience of the target and distractors

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

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


Analysis in several feature dimensions
Input visual scene
P(θ|T)
P(θ|D)
Stimulus feature θ
$s_{11}(A)$
$s_{i1}(A)$
$s_{n1}(A)$
Gains within a feature dimension
g_{11}
g_{i1}
g_{n1}
x
+
?
$S_1(A)$
$S_N(A)$
Prior statistical knowledge of the features of the target and distracting background
Gains across feature dimensions
g_1
g_N
Overall saliency map S(A)

Learning top-down gains

Maximizing SNR to obtain the optimal gains

$$\frac{\partial}{\partial g_{ij}}\mathcal{SNR} = \frac{\frac{\mathcal{SNR}_{ij}}{\mathcal{SNR}} - 1}{\alpha_{ij}}$$

$$\frac{\partial}{\partial g_{j}}\mathcal{SNR} = \frac{\frac{\mathcal{SNR}_{j}}{\mathcal{SNR}} - 1}{\alpha_{j}}$$

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

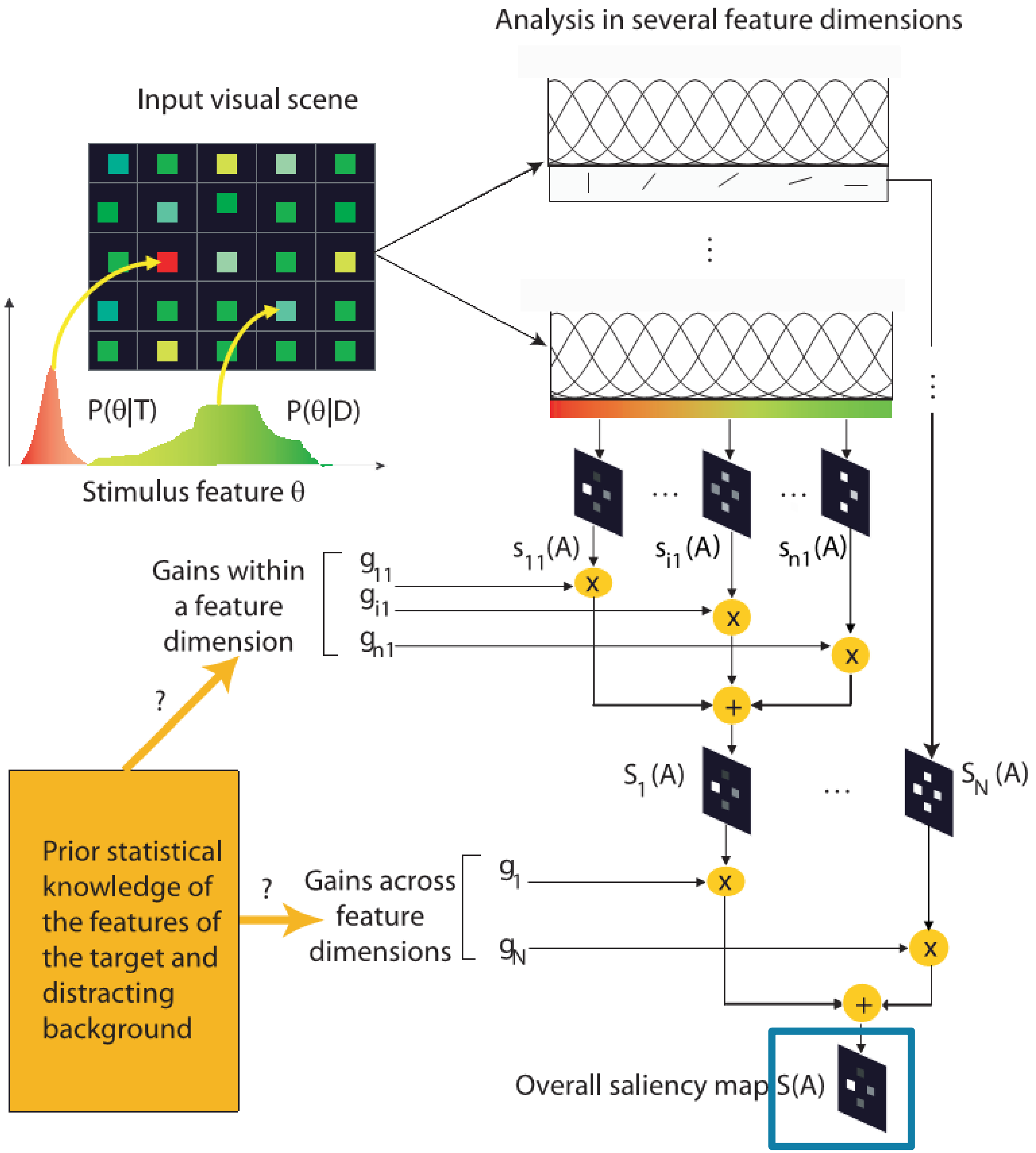
$$\mathcal{SNR}_{j} = \frac{E_{\Theta|T}[E_C[E_\eta[S_{jT}(A)]]]}{E_{\Theta|D}[E_C[E_\eta[S_{jD}(A)]]]}$$

Maximizing SNR to obtain the optimal gains

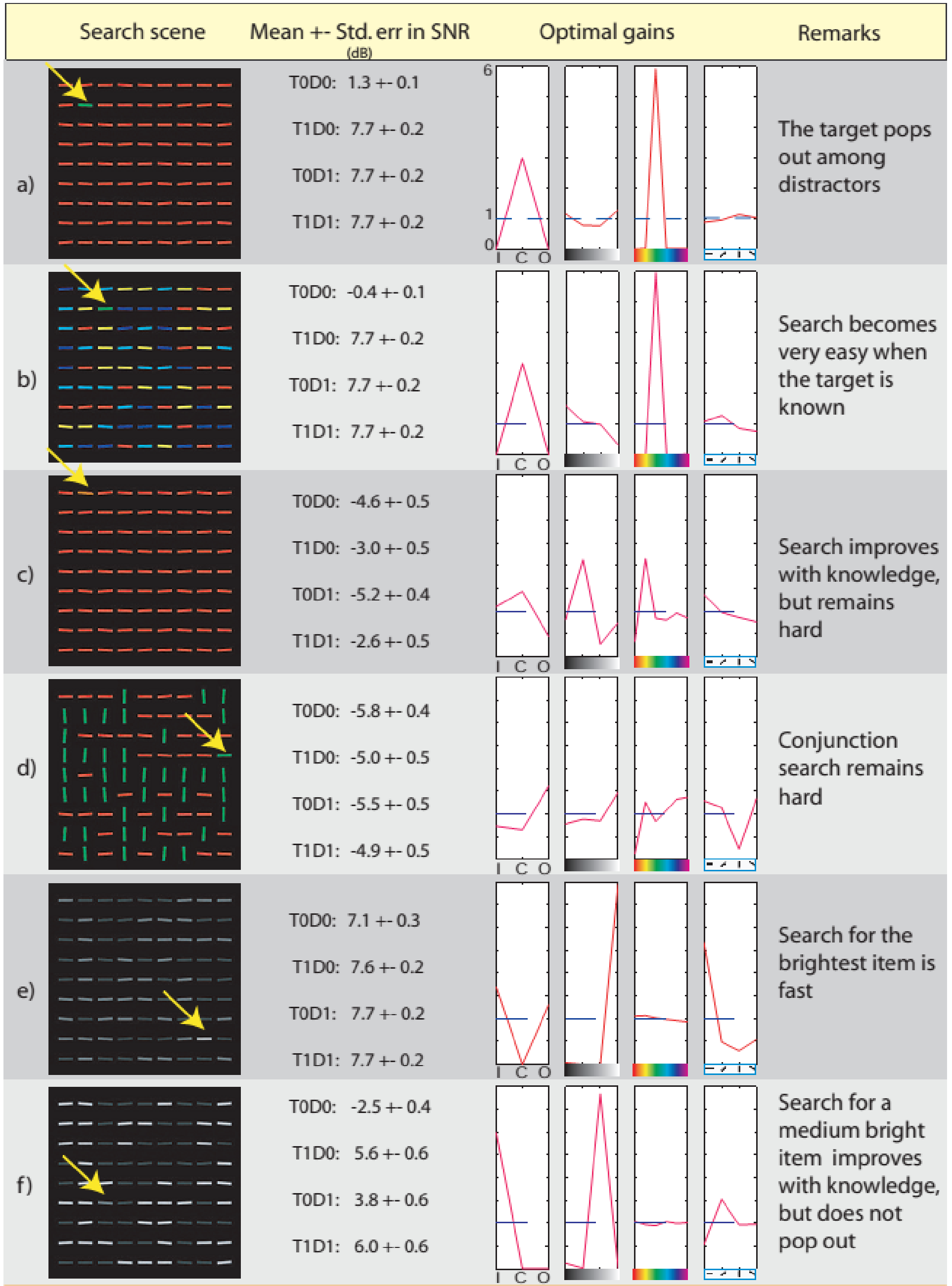
$$\begin{aligned}
\frac{\mathcal{SNR}_{ij}}{\mathcal{SNR}} &< 1 \Rightarrow \left(\frac{\partial}{\partial g_{ij}}\mathcal{SNR}\right)_{g_{ij}=1} < 0 \Rightarrow g_{ij} < 1 \\
&= 1 \Rightarrow \left(\frac{\partial}{\partial g_{ij}}\mathcal{SNR}\right)_{g_{ij}=1} = 0 \Rightarrow g_{ij} = 1 \\
&> 1 \Rightarrow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

$$\begin{aligned}
g_{ij} &= \frac{\mathcal{SNR}_{ij}}{\frac{1}{n}\sum_{k=1}^{n}\mathcal{SNR}_{kj}} \\
g_{j} &= \frac{\mathcal{SNR}_{j}}{\frac{1}{N}\sum_{k=1}^{N}\mathcal{SNR}_{k}}
\end{aligned}$$

$$\sum_{i=1}^{n} g_{ij} = n$$


Analysis in several feature dimensions
Input visual scene
P(θ|T)
P(θ|D)
Stimulus feature θ
$s_{11}(A)$
$s_{i1}(A)$
$s_{n1}(A)$
Gains within a feature dimension
g_{11}
g_{i1}
g_{n1}
?
$S_1(A)$
$S_N(A)$
Prior statistical knowledge of the features of the target and distracting background
?
Gains across feature dimensions
g_1
g_N
Overall saliency map S(A)

Resault



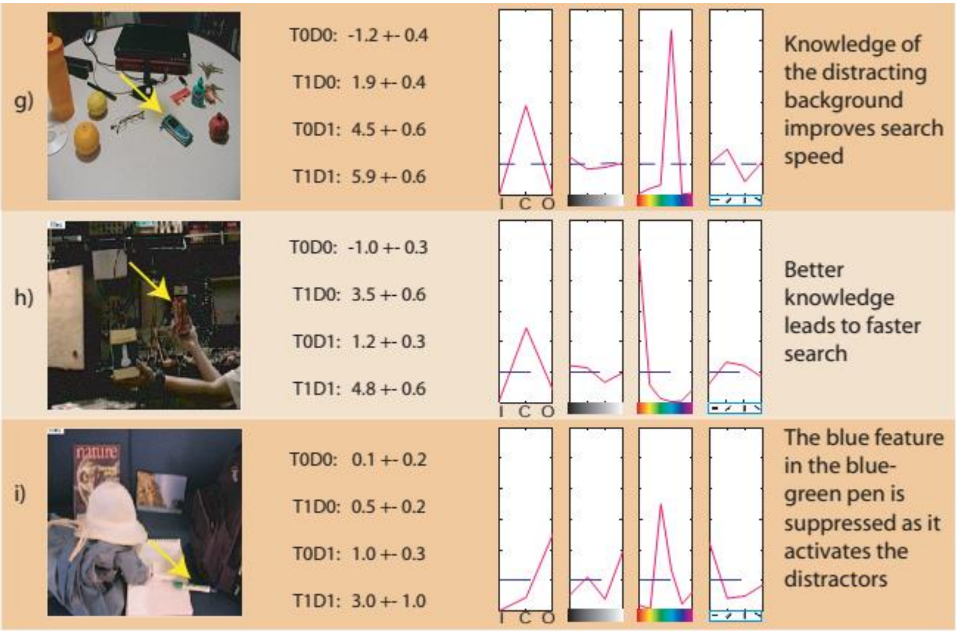
	Search scene	Mean +- Std. err in SNR (dB)	Optimal gains	Remarks
a)		T0D0: 1.3 +- 0.1 T1D0: 7.7 +- 0.2 T0D1: 7.7 +- 0.2 T1D1: 7.7 +- 0.2		The target pops out among distractors
b)		T0D0: -0.4 +- 0.1 T1D0: 7.7 +- 0.2 T0D1: 7.7 +- 0.2 T1D1: 7.7 +- 0.2		Search becomes very easy when the target is known
c)		T0D0: -4.6 +- 0.5 T1D0: -3.0 +- 0.5 T0D1: -5.2 +- 0.4 T1D1: -2.6 +- 0.5		Search improves with knowledge, but remains hard
d)		T0D0: -5.8 +- 0.4 T1D0: -5.0 +- 0.5 T0D1: -5.5 +- 0.5 T1D1: -4.9 +- 0.5		Conjunction search remains hard
e)		T0D0: 7.1 +- 0.3 T1D0: 7.6 +- 0.2 T0D1: 7.7 +- 0.2 T1D1: 7.7 +- 0.2		Search for the brightest item is fast
f)		T0D0: -2.5 +- 0.4 T1D0: 5.6 +- 0.6 T0D1: 3.8 +- 0.6 T1D1: 6.0 +- 0.6		Search for a medium bright item improves with knowledge, but does not pop out

T0D0, the naive, bottom-up model 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 sijT, while ignoring D by considering sijD 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.



g)		T0D0: -1.2 +- 0.4 T1D0: 1.9 +- 0.4 T0D1: 4.5 +- 0.6 T1D1: 5.9 +- 0.6		Knowledge of the distracting background improves search speed
h)		T0D0: -1.0 +- 0.3 T1D0: 3.5 +- 0.6 T0D1: 1.2 +- 0.3 T1D1: 4.8 +- 0.6		Better knowledge leads to faster search
i)		T0D0: 0.1 +- 0.2 T1D0: 0.5 +- 0.2 T0D1: 1.0 +- 0.3 T1D1: 3.0 +- 1.0		The blue feature in the blue-green pen is suppressed as it activates the distractors

Training and test data

For each search condition with the synthetic stimuli, the model learn target belief in salience (SbT, SbD) from 50 training images, computes the mean salience of the target and distractors $(E_{\Theta|T,C,\eta}[S_T^b(A)], E_{\Theta|D,C,\eta}[S_D^b(A)])$

In each of the 100 test image 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 Θ|T,Θ|D. Internal neural noise η is added by the saliency model.

Training and test data