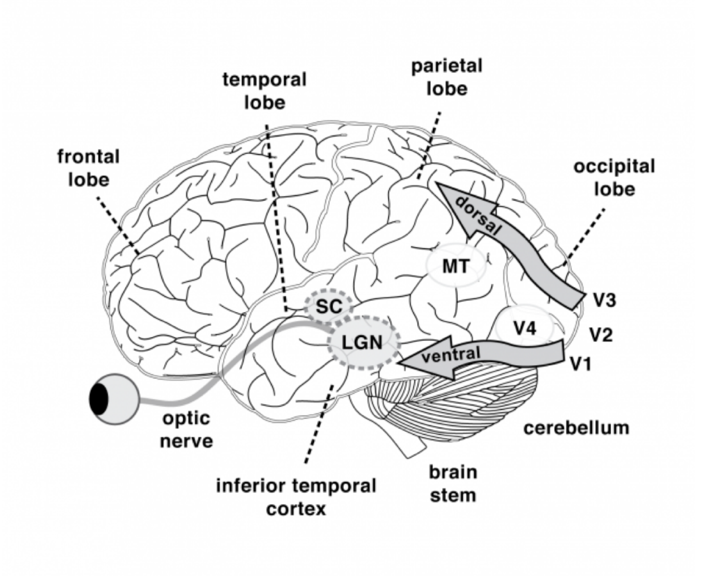
Computational Models of Visual Attention: Bottom-Up and Top-Down

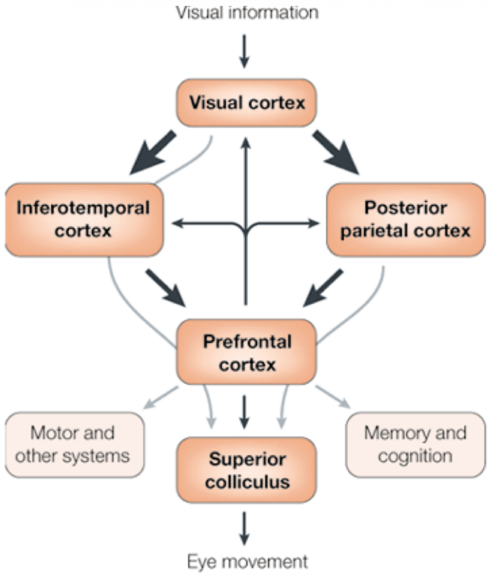
By: Soheil Borhani

Neural Mechanisms for Visual Attention

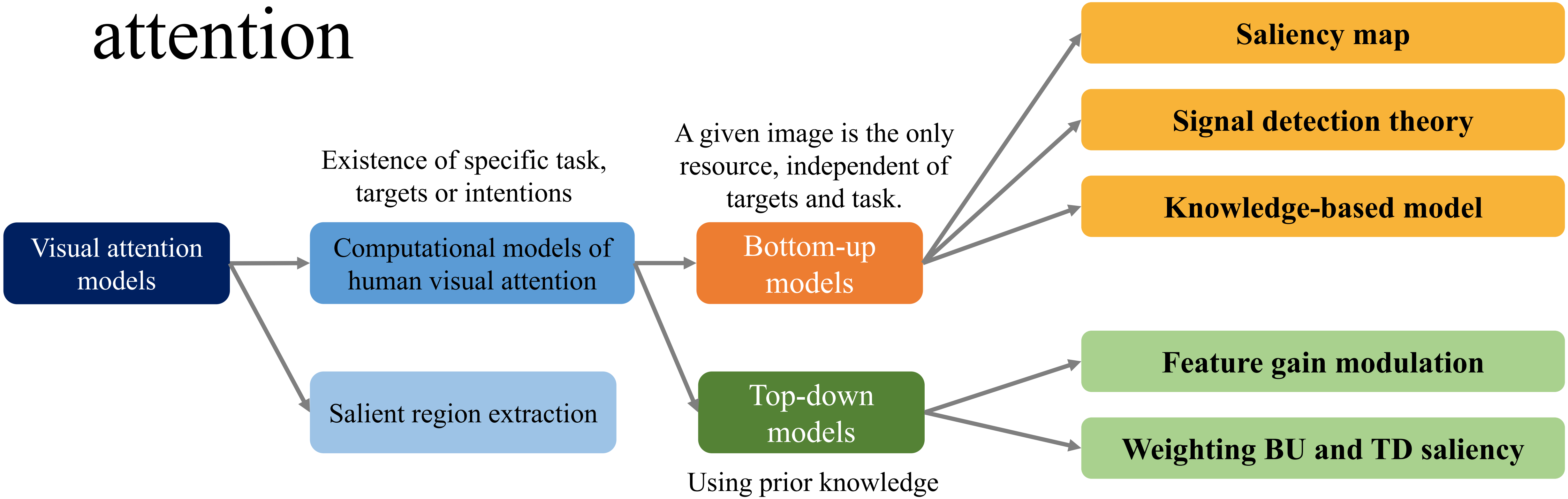
1. Visual information enter the primary visual cortex via lateral geniculate nucleus (LGN).

2. Then, visual information progresses along two parallel streams:
   - **Dorsal stream (PPC):** Mainly involved in spatial localization and directing attention. Control of attentional deployment takes place in this area.
   - **Ventral stream (IT):** Mainly concerned with the recognition and identification of visual stimuli. Although not directly involved in the control of attention, ventral stream receives attentional feedback modulation and involved in the representation of attended locations and objects.





Classifying computational models of visual attention



Visual attention models

Computational models of human visual attention

Existence of specific task, targets or intentions

Salient region extraction

Bottom-up models

A given image is the only resource, independent of targets and task.

Top-down models

Using prior knowledge

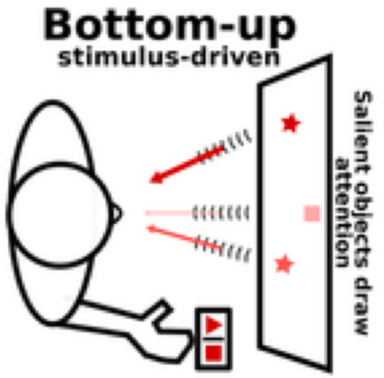
Saliency map

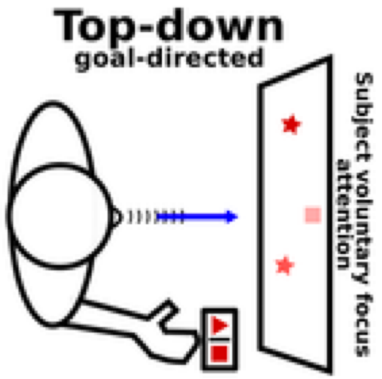
Signal detection theory

Knowledge-based model

Feature gain modulation

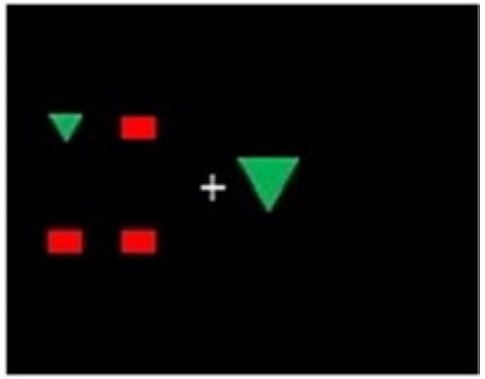
Weighting BU and TD saliency



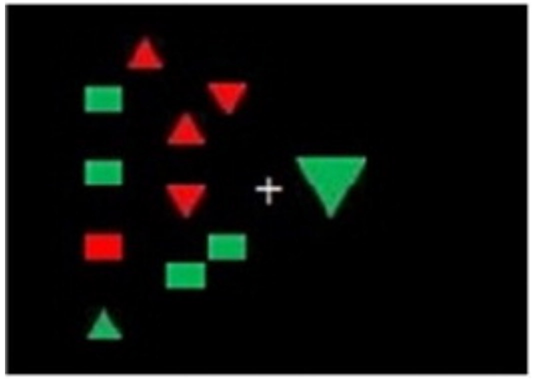


Visual Search

- Feature search requires identification of a pop-out target, defined by a single feature such as intensity, color, edge orientation. (e.g., search for the only triangle among other stimuli)
- Conjunction search requires identification of a target defined by a combination of two or more features (e.g., search for a green triangle among green squares and red triangles & red squares).



Feature search



Conjunction search

- We try to examine visual perception via visual search with various relationship of target and distractors

Feature Integration Theory (FIT)-(Bottom-up)

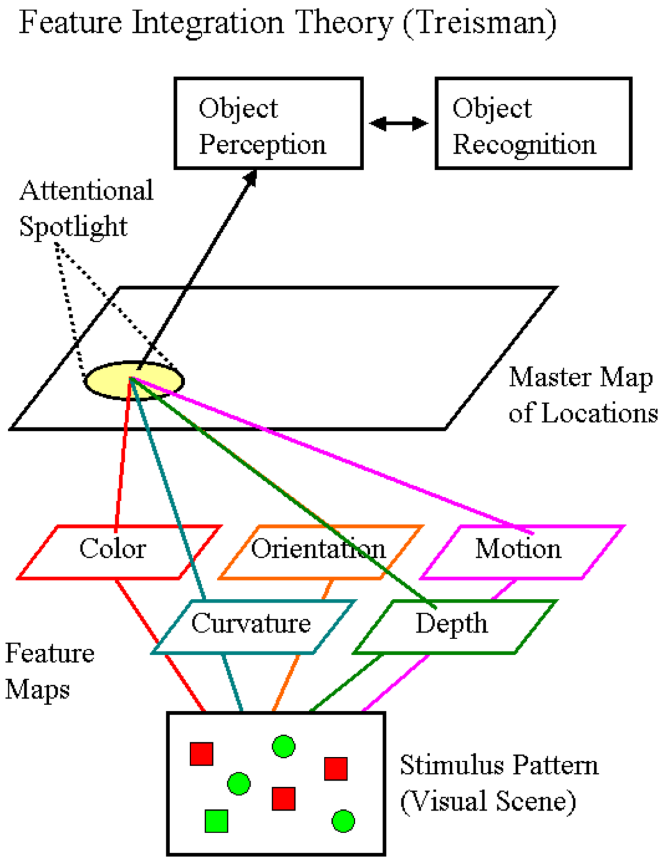
✓Proposed by Treisman [1980], the methodology is to measure **reaction time** of subjects who search for a target among several distractors under the condition of the feature and conjunction searches.

✓She argued that the **feature search** is conducted in **parallel** and the **conjunction search** is conducted **sequentially**.

- **Assumptions:**

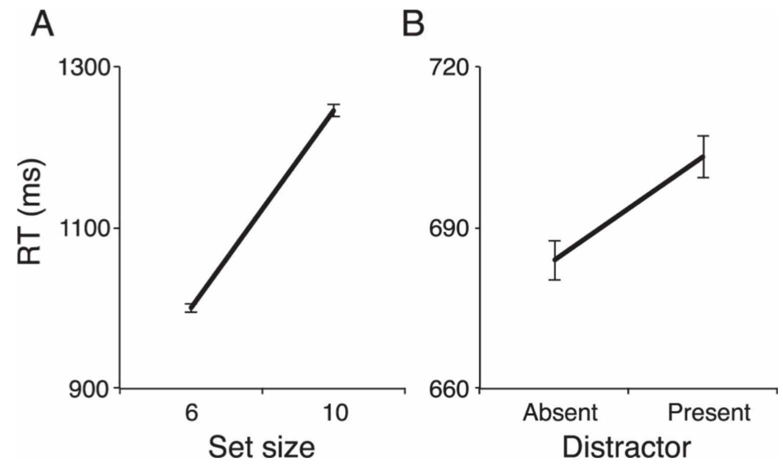
✓A constant reaction time independent of the number of distractors is a sign of parallel search.

✓An increased reaction time with increasing the number of distractors is a sign of sequential search.



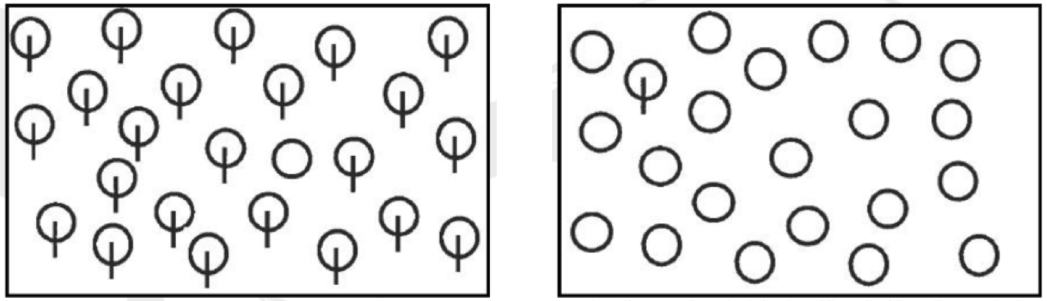
Critiques on FIT

- **Reaction times** can differ greatly according to the given tasks, even with the same sets of visual stimuli. This phenomenon is interpreted in terms of trade-off between the spatial coverage and resolution of attention.



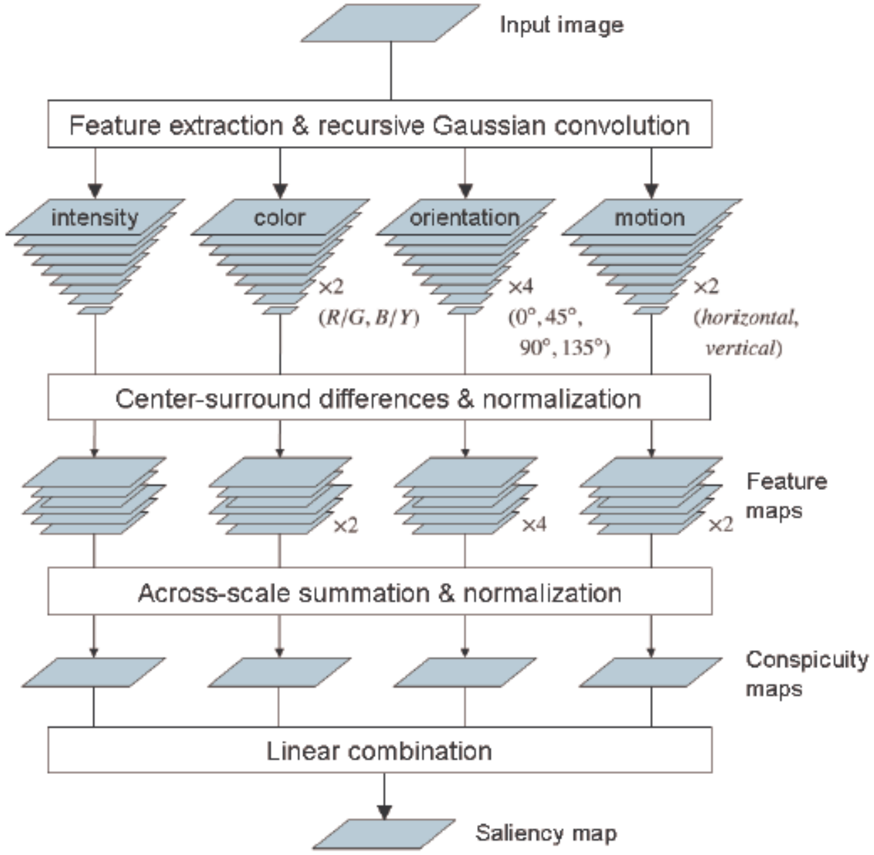
- Several studies have reported the results indicating that both feature and conjunction search have a difficulty depending on the similarity between targets and distractors as well as the similarity between distractors.

- **Asymmetry** between searching result of **A** among **B** stimulus and **B** among **A** stimulus.



Saliency Map (Bottom-Up)

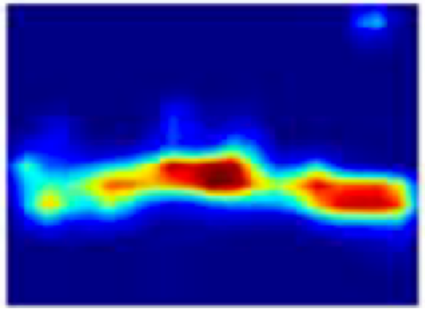
- Saliency map model proposed by Itti, Koch and Niebur [1998]
- The proposed architecture introduces a multi-resolution structure
- The structure computes multi-scale spatial contrasts for “fine” and “coarse” scales.
- The saliency map model introduces a winner takes-all (WTA) mechanism, which selects the location where the pixel value of the saliency map is greater than at any other locations.



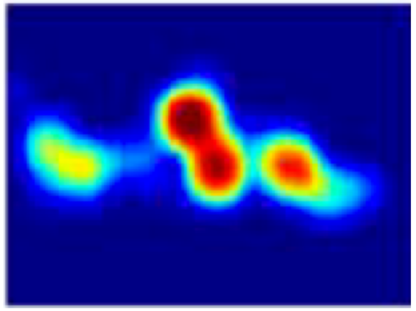
Critique on the Saliency map model

- A **deterministic extraction** of the Saliency map which implies that all humans would focus on the same location for the same input image. However, according to the findings, humans may focus on different locations in the same input image (Caused by top-down intention, knowledge and preferences)

Original scene



Itti saliency map

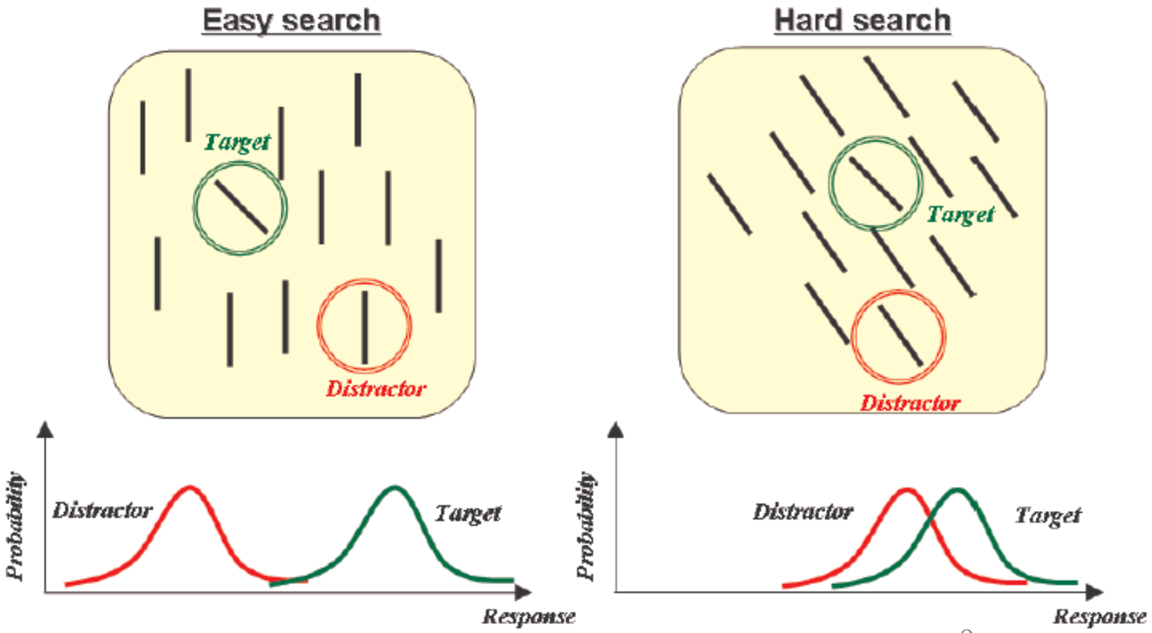


Extracted saliency map from eye tracker on one subject

Signal Detection Theory (SDT)

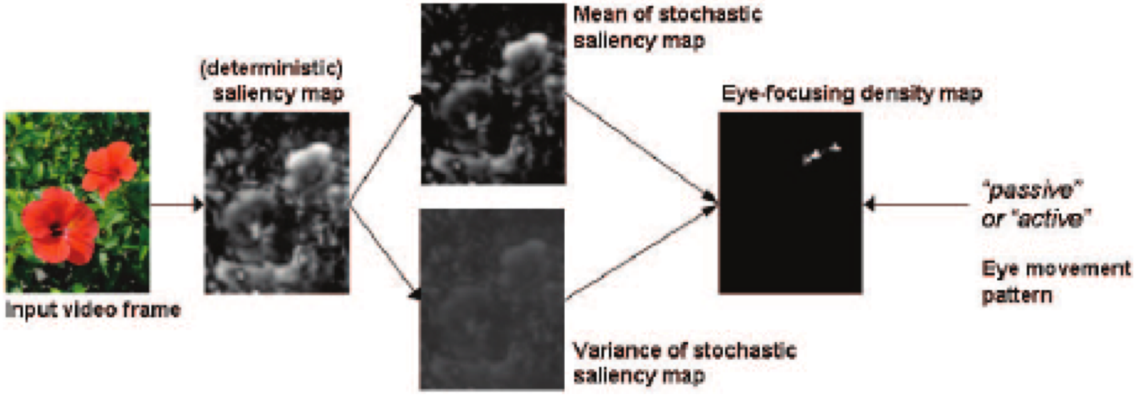
- The theory has been employed in communication theory.
- Signal detection theory described a mechanism that causes **attention ambiguity** using only bottom-up processing.
- Eckstein [2001] applied SDT to model visual attention.

- FIT → Human eyes cannot make a mistake!
- SDT → Assumes distractors might be also recognized as a target



Signal Detection Theory

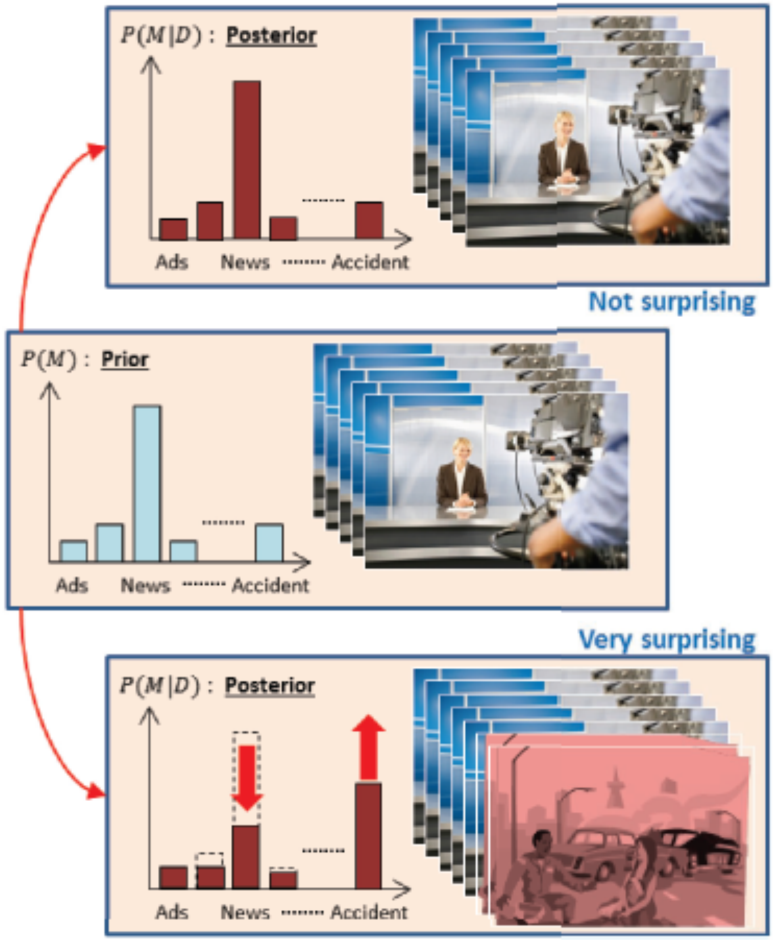
- Takahiko Koike and Jun Saiki [2006] first introduced a stochastic mechanism of human visual attention into a computational model and verified that with psychophysical settings.
- Pang et al. [2008, 2010] extended this model to video inputs, and constructed a dynamic Bayesian network that considered the stochastic ambiguity and temporal smoothness of visual saliency simultaneously
- Miyazato et al. [2009] achieved real-time computing of this model with the full use of parallel processors in GPUs.

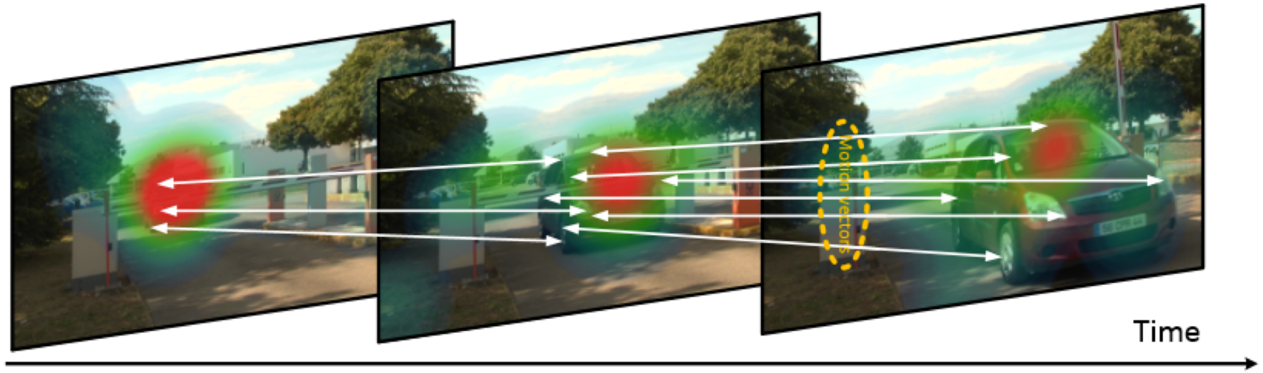

Input video frame
(deterministic) saliency map
Mean of stochastic saliency map
Variance of stochastic saliency map
Eye-focusing density map
"passive" or "active"
Eye movement pattern

Saliency in temporal changes

- Itti and Baldi [2009 & 2010] incorporated the temporal dynamics of image features into computational models of human visual attention, and proposed the Bayesian Surprise model that regards the difference between the visual features that are expected to be obtained and those that are actually obtained as indicating saliency.

- Sophie Marat [2009] utilized the motion features in a computational model of human visual attention.





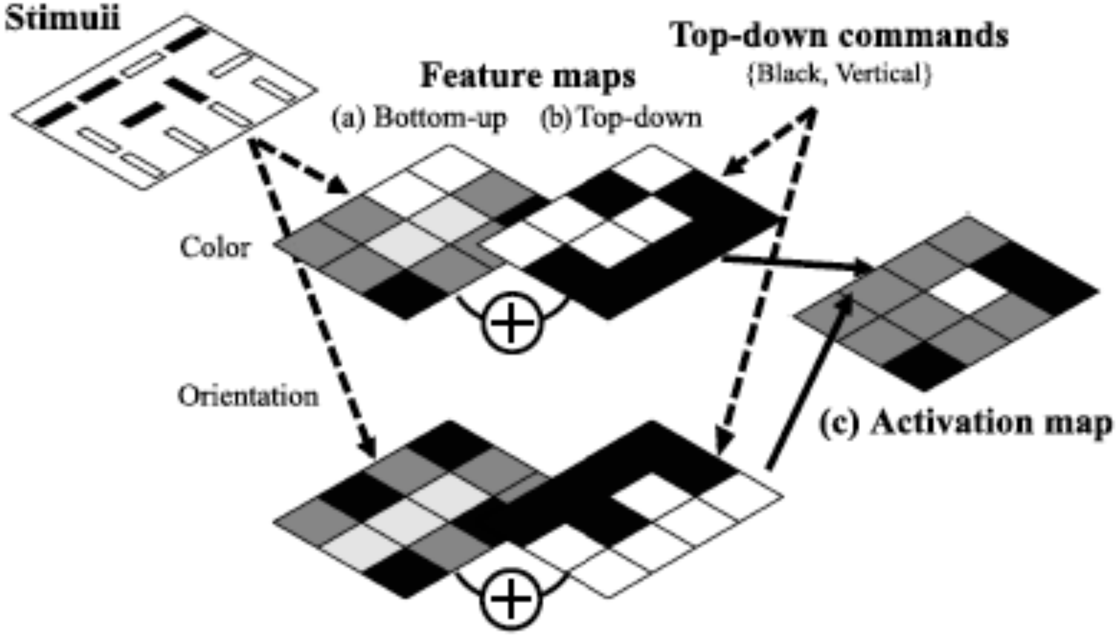
Guided Search Model (Bottom-Up &Top-Down)

- Proposed by Jeremy Wolfe [1989], it has introduced **top-down knowledge** on characteristics of target stimuli in visual search. The model employs both **bottom-up** and **top-down** activation.

- The **bottom-up** features come from saliency in each feature (e.g., color, orientation). The feature maps obtained from feature search.

- The **top-down** activation is obtained based on the correlation between input and target stimuli with regard to each of the features.

- The activation maps is achieved by summing up the top-down and bottom-up activation maps.



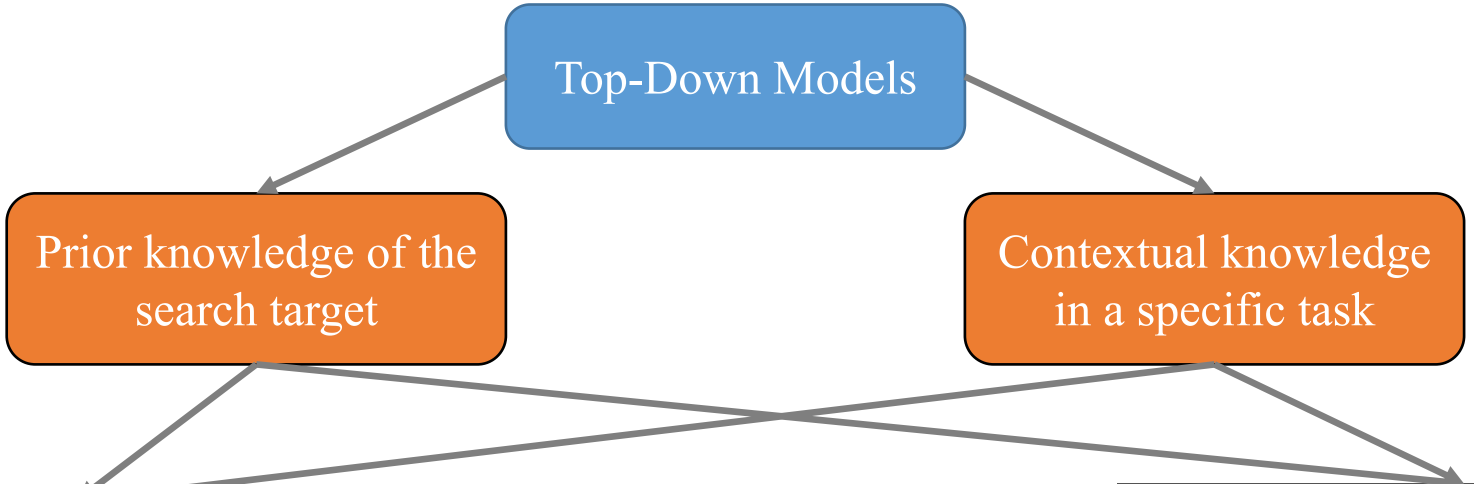
Top-Down Models

- Many researchers proposed top-down models including Wolfe's guided search model
- The top-down models fall into two categories:

1. Computational models based on **prior knowledge of the search target** in a visual search task.

2. Computational models based on **contextual knowledge in a specific task** situation.

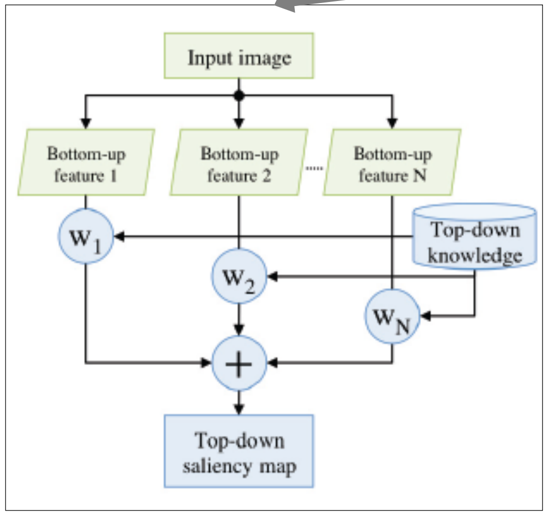
Top-Down Models



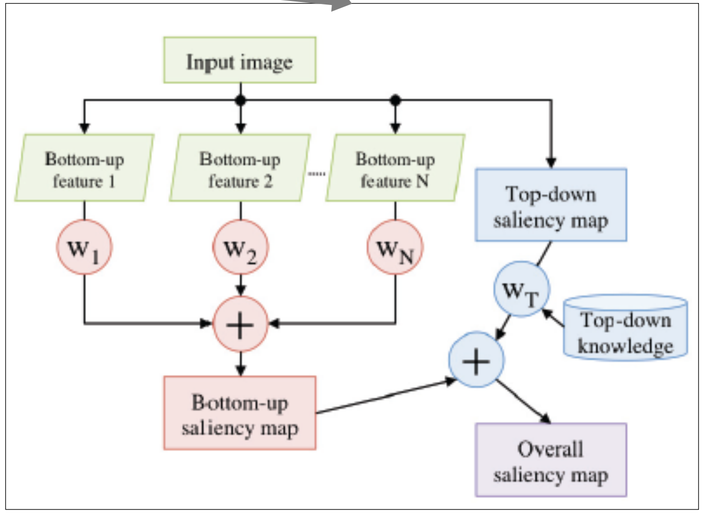
Top-Down Models

Prior knowledge of the search target

Contextual knowledge in a specific task



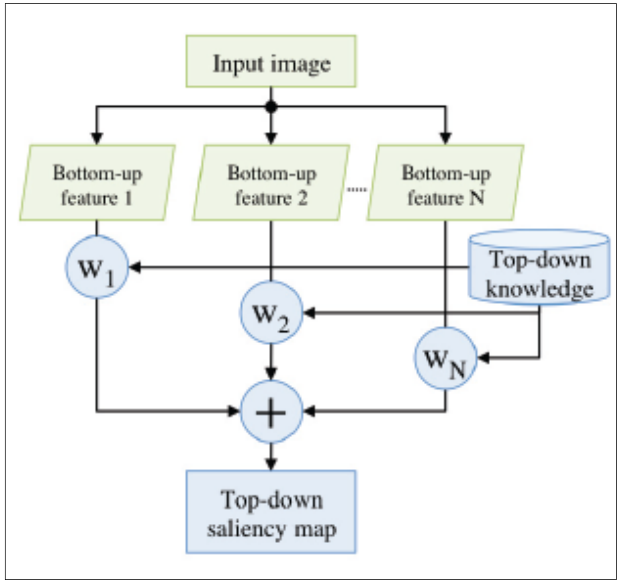
Weight modulation of bottom-up features



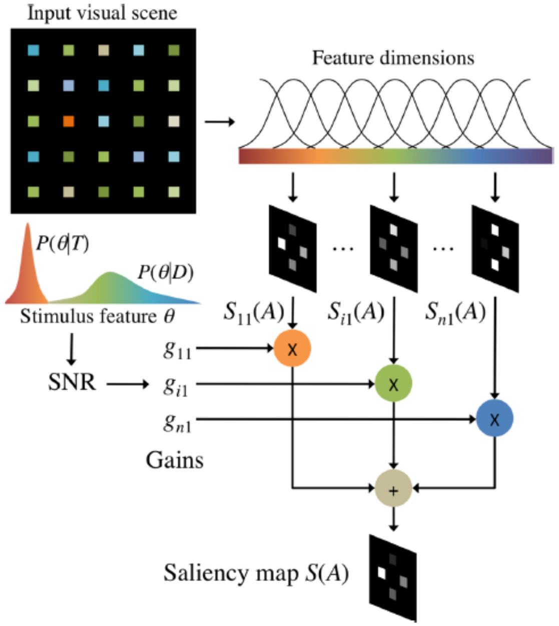
Weighted combination of bottom-up and top-down saliency maps

Weight Modulation of Bottom-Up features using Top-down features

- The weighted response of each channel to the target is compared with its average response to the distractors. The channel with the greatest positive difference is selected to compute the top-down saliency map.

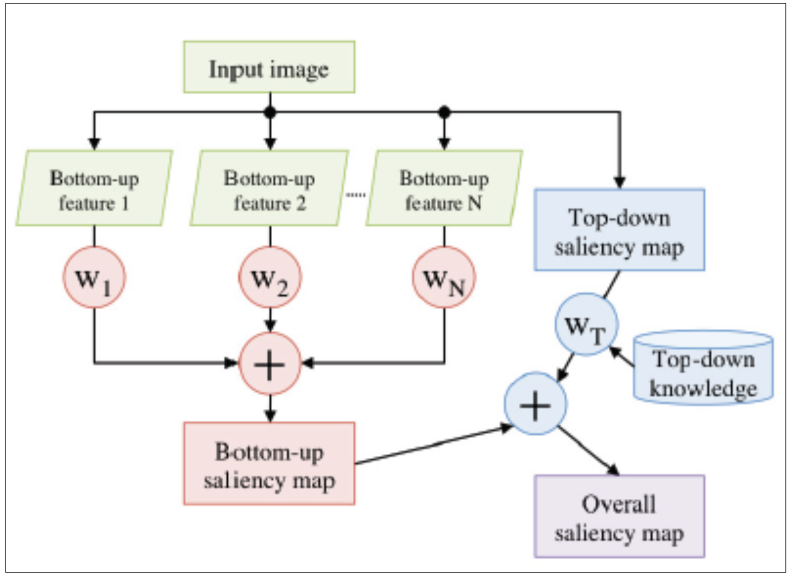


- The signal-to-noise ratio (SNR), i.e., the ratio between target salience and distractor salience, is effective information for controlling the weights of feature channels. (Frintrop et al. [2006])

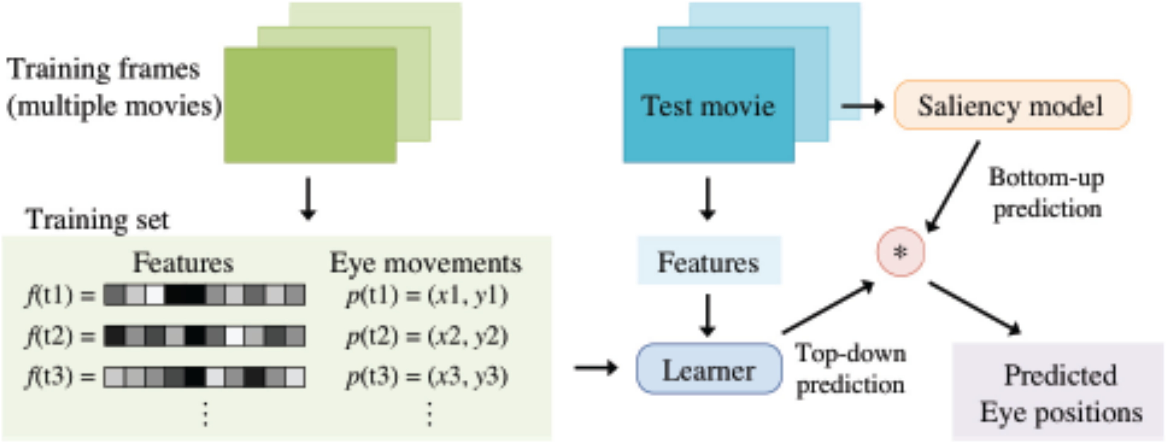


Weighted combination of bottom-up and top-down saliency maps

- In the original guided search model, Bottom-up saliency is combined using equal weight with top-down saliency calculated using the selected feature channels.



- The model has evolved through several versions which give rise to "Weighted combination of bottom-up and top-down". It has an additional top-down mechanism to handle a scene context and the inhibition tagging mechanism to simulate visual attention more accurately.



Thank you!