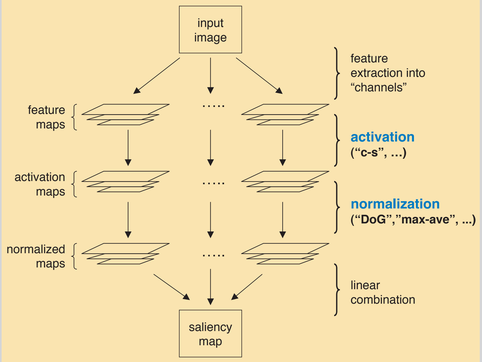
Graph-Based Visual Saliency

Jonathan Harel, Christof Koch, and Pietro Perona
California Institute of Technology, Pasadena, CA

Background

We propose a new "bottom-up" saliency model (GBVS) which exploits the distributed nature of graph algorithms.

But first, we organize the standard approaches into the following steps:



We propose an alternative to the standard activation and normalization schemes.

(See [1], [2] for examples)

Approach

In biology, individual "nodes" (neurons) exist in a connected, retinotopically organized, network (the visual cortex), and communicate with each other (synaptic firing) in a way which gives rise to emergent behavior, viz., rapid scene analysis for salient locations.

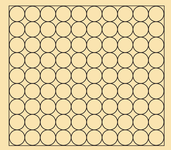
Therefore, we propose a distributed, graph-based solution which uses local computation to obtain a saliency map which is everywhere dependent on global information.

For both activation and normalization, we will construct a directional graph with edge weights given from the input map, treat it as a Markov chain, and compute the equilibrium distribution.

Graph Construction

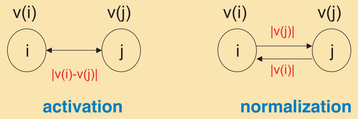
We construct a graph as follows:

1. We instantiate a node for every location in an input map (feature or activation).



2. We introduce directional edges in both directions between every pair of nodes.

3. We assign edge weights as follows:



but we multiply the edge weights by a Gaussian distance penalty, so that nodes which are distant only weakly interact.

This approach is extended to multiple spatial scales by introducing nodes at every location at every scale, and defining the edges and their weights the same as before with an appropriate definition of distance across nodes at different scales.

We we treat nodes as states and edge weights as transition probabilities, and compute the equilibrium distribution of the Markov chain.

If an input map has size NxN, this will have time complexity $O(N^4k)$ where k<<N is some small number of iterations required to meet equilibrium.

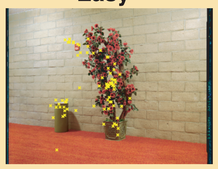
Experiments

For each image in some data corpus, we create a collection of saliency maps by concatenating various activation and normalization procedures subsequent to the exact same feature extraction step. We then compare consistency of each saliency map with fixation data using an ROC score (see [4]).

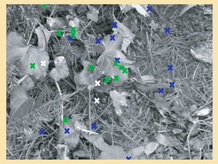
For the main results, a corpus of 750 modifications of 108 grayscale images of nature were used from a recent study [3].

Sample Images

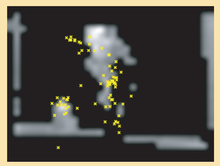
Easy

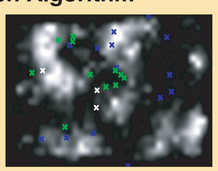


Difficult

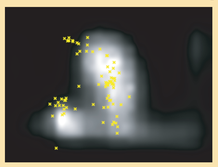


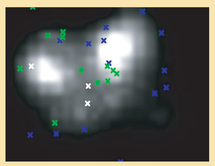
Standard Itti/Koch Algorithm





Our Approach

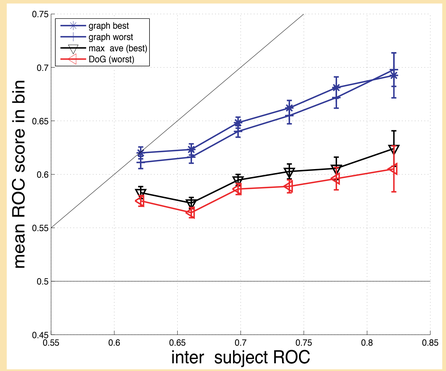




Images, together with the human fixations (on right, different subjects' shown in different colors).
Sometimes, saliency is highly predictive (left), but in the absence of strong bottom-up stimuli, saliency algorithms struggle (right).

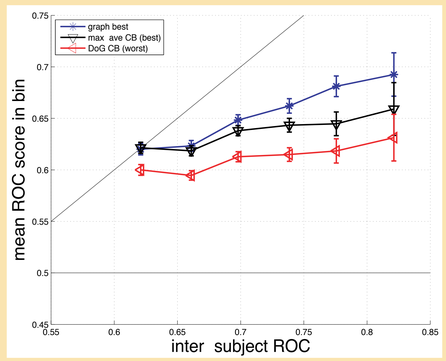
Results

Here, we show the mean ROC score for images in different bins. Each bin corresponds to an "image easiness", parametrized as mean inter-subject ROC score.



The best and worst settings of parameters for the graph-based method are show in blue. The best and worst settings for the standard approach of Itti and Koch [1] are also shown (above).

Below, we see what happens when a center-bias is added to the standard approaches. Their performance improves dramatically, but is still significantly lower in some regimes.



Interpretation

Transitions into center nodes are more probable on average than transitions into any one peripheral node, thus our algorithm results in an emergent center-bias which is favorable for performance.

Also, our algorithm leaves saliency mass away from object borders in a non-trivial way that cannot be mimicked by smoothing alone.

Also, our approach finds saliency values at each location which depend on the entire image plane. This is different than most modern approaches ([1], [2]) which rely on local information.

Conclusions

We propose a new, unified framework for computing bottom-up saliency maps based on a simple, biologically plausible, and distributed computation.

The model shows a strong consistency with the attentional deployment of human subjects on a grayscale natural image corpus.

We believe its superior predictive power stems from its emergent center bias, its ability to pick out salient regions away from borders, and its implicit use of global information.

References

[1] L. Itti, C. Koch, PAMI 1998
[2] N. Bruce, J. Tsotsos, NIPS 2005
[3] Einhaeuser, et al. Vis. Res. 2006
[4] Tatler, et al. Vis. Res. 2005