Computational Cognitive Science

Visual Attention

Chris Lucas
(Slides adapted from Frank Keller's)

School of Informatics
University of Edinburgh
clucas2@inf.ed.ac.uk

19 November 2019

Reading: Itti, Koch, and Niebur (1998).

The Visual Processing Pipeline

The rest of the course will deal with *human visual cognition.* We will focus on high-level visual processing (not visual neuroscience):

- *Visual attention:* How do we decide which parts of an image to focus on?
- *Visual search:* How do we search for a target in an image?
- *Object recognition:* How do we identify objects in an image?

We will introduce computational models in all three domains.

The Visual Processing Pipeline

When we view an image, we actually see this:

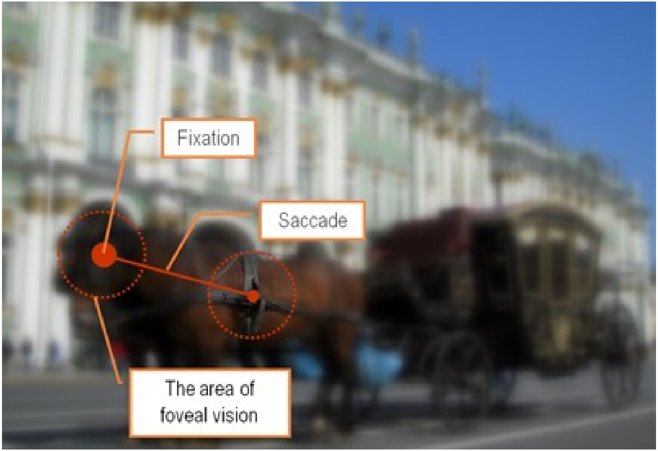


Image from http://eyetracking.me/

Only the *fovea*, a small area in the center of the retina, is in focus.

The Visual Processing Pipeline

In order to take in the whole image, we have to move our eyes: *fixations* (stationary periods) and *saccades* (rapid movements).

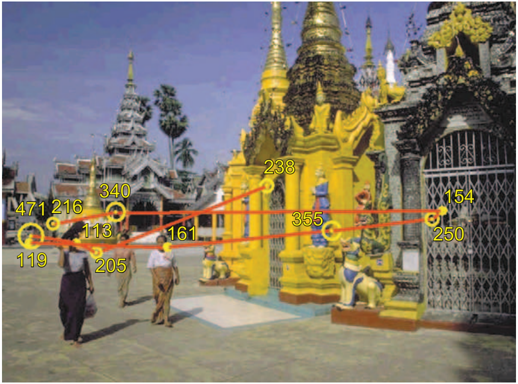
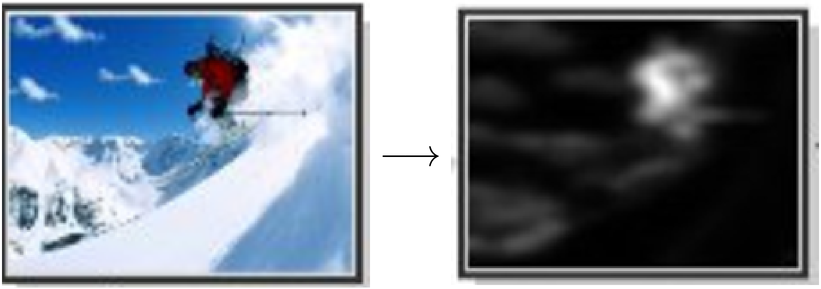


Image from Henderson, 2003

How do we determine *where to look?* We need to work out which area are interesting, i.e., attract *visual attention.*

Visual Saliency

We attend to the areas that are *visually salient.* An area is salient if it stands out, is different from the rest of the image.



The visual system computes a *saliency map* of the image, and then moves the eyes to the most salient regions in turn (Itti et al., 1998).

Visual Features

Saliency can be tested using *visual search experiments:* participants have to find a target item among a number of distractors.

Examples for visual features that can make a target salient:

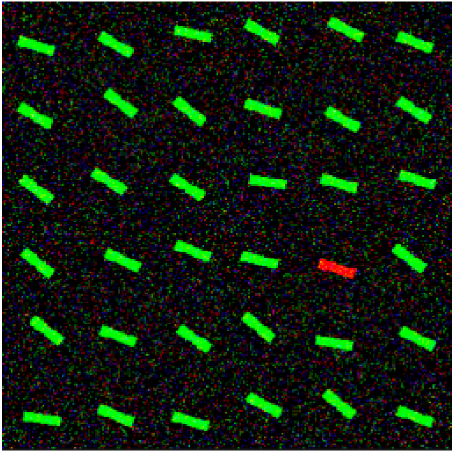
- color;
- orientation;
- intensity.

Saliency can make the target *pop out* from its distractors if it differs in one of this features.

The pop-out effect doesn't occur if the target is different from the distractors in two aspects (conjunction target).

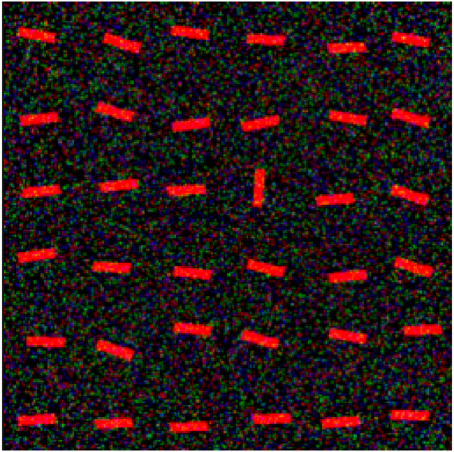
Visual Features

Pop-out because of color (Itti, 2007):



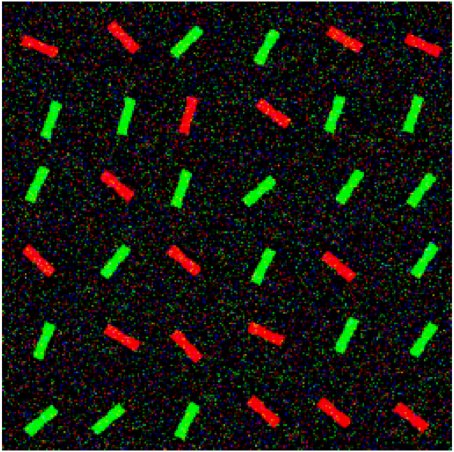
Visual Features

Pop-out because of orientation (Itti, 2007):



Visual Features

No pop-out: conjunction target (Itti, 2007):



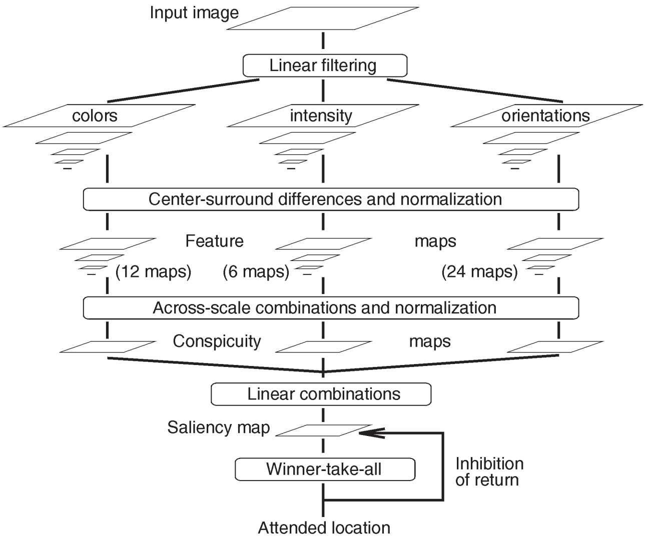
Model Architecture

Itti et al.'s (1998) computational model of saliency:

- compute feature maps for color, intensity, orientation at different scales;
- compute center-surround difference and apply a normalization;
- combine the maps across scales into conspicuity maps;
- saliency map is a linear combination of the conspicuity maps;
- winner-takes-all operator predicts attended locations.

This model mainly works for free viewing. In the next lectures we will talk about models that can account for search data.

Model Architecture



Feature Maps

Feature maps are computed at nine *spatial scales* (1:1 to 1:256) by low-pass filtering (blurring) and subsampling the image.

A *center-surround operator* is used to detect locations that stand out from their surroundings:

- this is implemented as the difference between finer and coarser scales;
- the center is a pixel at scale $c \in \{2, 3, 4\}$;
- the surround is the corresponding pixel at scale $s = c + d$, with $d \in \{3, 4\}$;
- the across-scale difference between two maps is denoted as $\ominus$.

Intensity

At each spatial scale, a set of feature maps are computed based on the red, green, and blue color values (r, g, b) of the pixels.

Intensity map: compute the intensity function $I = (r + g + b)/3$ and then the intensity map using the center surround operator:

$$\mathcal{I}(c, s) = |I(c) \ominus I(s)|$$

with $c \in \{2, 3, 4\}$ and $s = c + d$, with $d \in \{3, 4\}$.

Color

Color maps: compute four color values $R = r - (g + b)/2$ for red, $G = g - (r + b)/2$ for green, $B = b - (r + g)/2$ for blue, and $Y = (r + g)/2 - |r - g|/2 - b$ for yellow.

Then compute color maps again using center-surround:

$$\mathcal{RG}(c, s) = |(R(c) - G(c)) \ominus (G(s) - R(s))|$$

$$\mathcal{BY}(c, s) = |(B(c) - Y(c)) \ominus (Y(s) - B(s))|$$

These are based on color opponencies (exist in the visual cortex).

Orientation

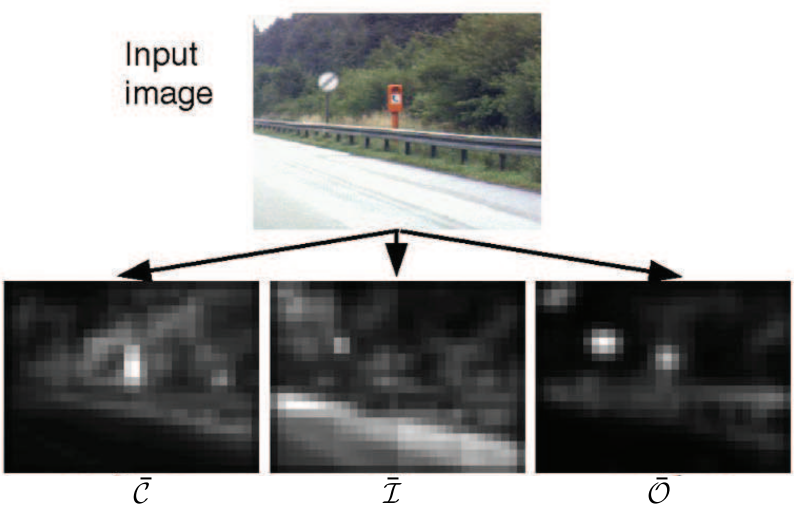
Orientation map: compute Gabor pyramids $O(\sigma, \theta)$ where $\sigma \in [0 \dots 8]$ is the scale and $\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$ is the preferred orientation.

Then compute color maps again using center-surround:

$$\mathcal{O}(c, s, \theta) = |O(c, \theta) \ominus O(s, \theta)|$$

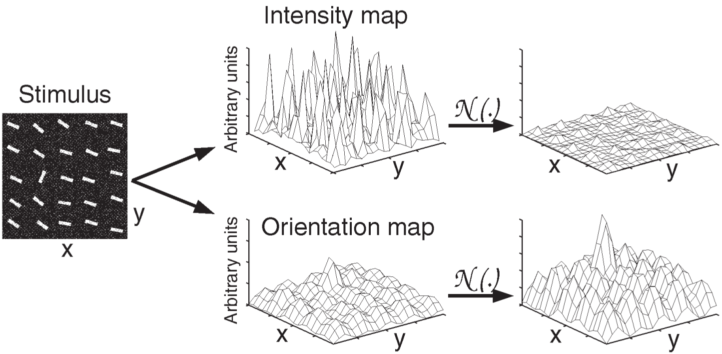
In total, 42 feature maps are computed: six for intensity, 12 for color, and 24 for orientation.

Example



Saliency Map

Before we combine feature maps, a normalization operator $\mathcal{N}(\cdot)$ is applied, which promotes maps with a small number of strong peaks, and suppressed maps with many similar peaks.



Saliency Map

The feature maps are combined into three conspicuity maps for intensity, color, and orientation at the same scale ($\sigma = 4$).

For intensity and color, we get:

$$\bar{\mathcal{I}} = \oplus_{c=2}^{4} \oplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{I}(c,s))$$

$$\bar{\mathcal{C}} = \oplus_{c=2}^{4} \oplus_{s=c+3}^{c+4} [\mathcal{N}(\mathcal{RG}(c,s)) + \mathcal{N}(\mathcal{BY}(c,s))]$$

where the $\oplus$ operator reduces each map to scale 4 and performs point-by-point addition.

Saliency Map

For orientation, we first combine the six feature maps for a given angle and then add them to get a single conspicuity map:

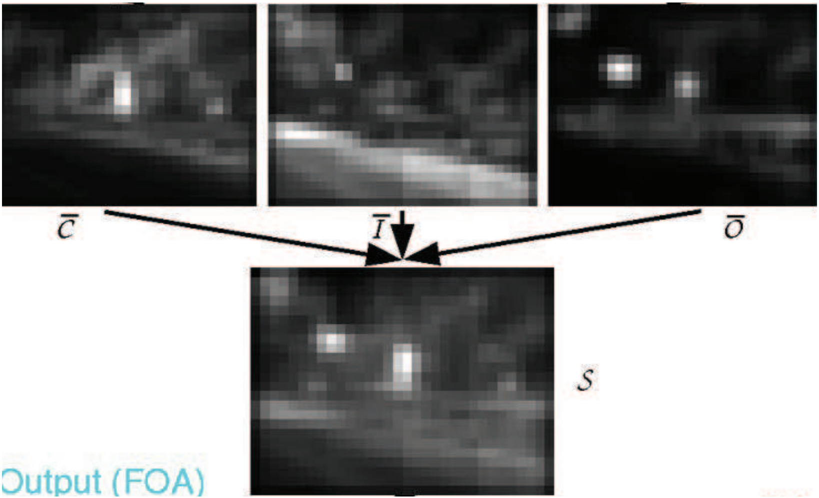
$$\bar{\mathcal{O}} = \sum_{\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}} \mathcal{N}(\oplus_{c=2}^{4} \oplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{O}(c, s, \theta)))$$

The overall saliency map is then computed by normalizing and averaging the three conspicuity maps:

$$\mathcal{S} = \frac{1}{3}(\mathcal{N}(\bar{\mathcal{I}}) + \mathcal{N}(\bar{\mathcal{C}}) + \mathcal{N}(\bar{\mathcal{O}}))$$

Why do we normalize each conspicuity map separately? Similar features compete strongly for saliency, while different ones contribute independently to saliency.

Example



Inhibition of Return

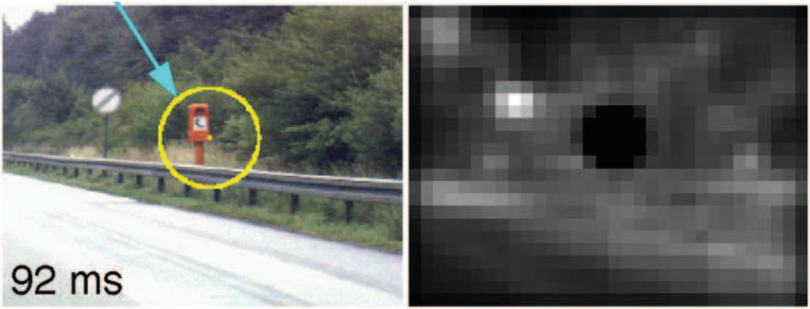
Now we can predict sequences of fixations from a saliency map:

- the maximum of $\mathcal{S}$ is the most salient location, which becomes the focus of attention (FOA);
- all other locations are ignored (inhibited);
- then the saliency around the FOA is reset, so that the second most salient location becomes the new FOA.

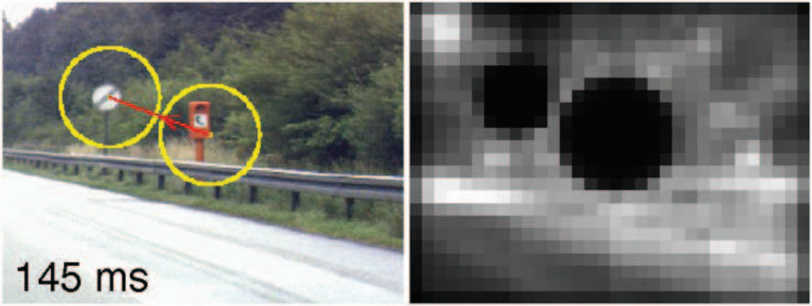
The last property is crucial: it results in *inhibition of return,* so that the FOA doesn't immediate return to the most salient location.

Itti et al. (1998) implement this using a winner-take-all neural network. This allows them to simulate fixation durations.

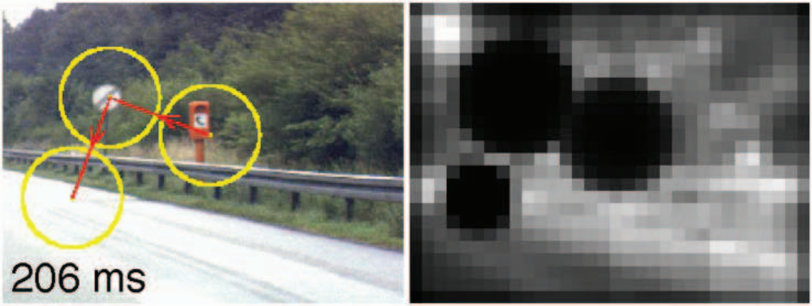
Inhibition of Return



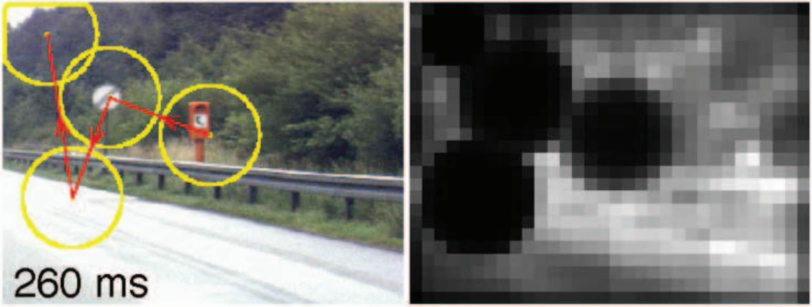
Inhibition of Return



Inhibition of Return

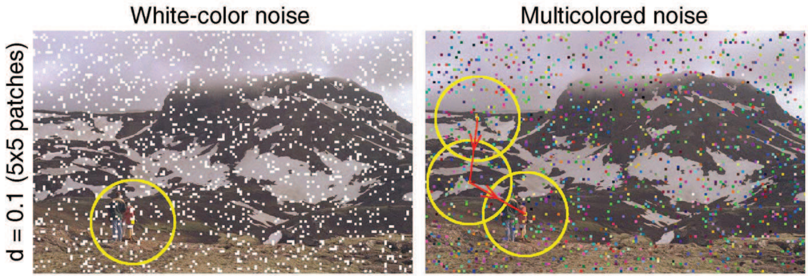


Inhibition of Return



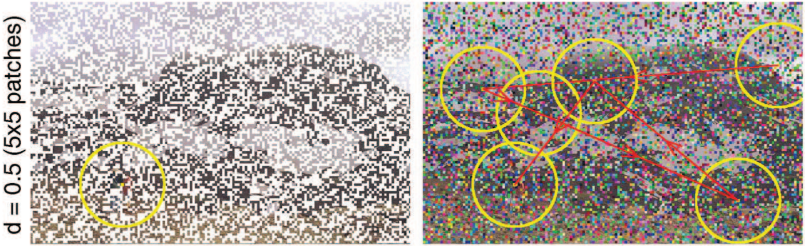
Robustness to Noise

Test the model by adding noise to the image, see if it is still able to pick out salient locations correctly.



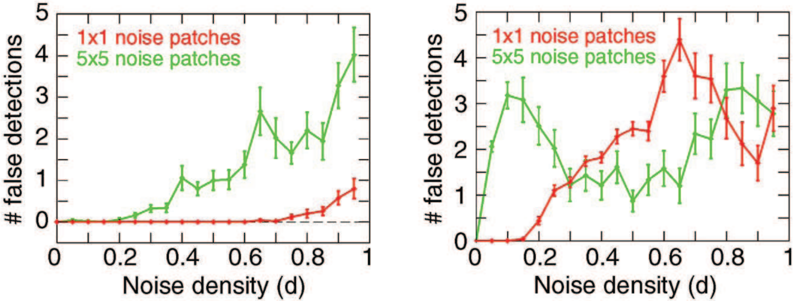
Robustness to Noise

Test the model by adding noise to the image, see if it is still able to pick out salient locations correctly.



Robustness to Noise

Test the model by adding noise to the image, see if it is still able to pick out salient locations correctly.



Evaluation

Evaluation reported by Itti et al. (1998):

- saliency model can reproduce human performance in pop-out tasks (including conjunction target);
- tested also on images of traffic signs, red soda cans, and emergency triangles (though no details given in the paper);
- outperforms spatial frequency models.

No evaluation of saliency against eye-tracking data. However, there is a lot of subsequent work on this topic, such as Borji, Sihite, and Itti (2013).

Strengths and Limitations

Strengths:

- simple feed-forward architectures generates complex behavior;
- massively parallel implementation (biologically plausible);
- very successful as model of early visual processing.

Weaknesses:

- can only detect regions that are salient based on either color, intensity, or orientation;
- other features (e.g., T junctions, line termination) or conjunctions of features are not accounted for in the model;
- motion is important for saliency, but is not modeled;
- the normalization function $\mathcal{N}(\cdot)$ plays a crucial role without being theoretically well-founded;
- no notion of object in the model (saliency is a property of a point); but objectness crucial for human scene perception.

Summary

- Attention selects the part of the visual input which is fixated and processed in detail;
- attention is directed to visually salient areas in an image, i.e., areas that are different from the rest of the image;
- the saliency model is based on color, orientation, intensity maps computed at various spatial scales;
- center-surround differences are applied, and the maps normalized and combined into a single saliency map;
- a winner-takes-all mechanism then predicts attended locations;
- model is robust to noise and models human fixation behavior.

References

Borji, A., Sihite, D. N., & Itti, L. (2013). Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, *22*(1), 55–69.

Henderson, J. (2003). Human gaze control in real-world scene perception. *Trends in Cognitive Sciences*, *7*, 498–504.

Itti, L. (2007). Visual salience. *Scholarpedia*, *2*(9), 3327.

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.