Computational Cognitive Science

Lecture 19: Contextual Guidance of Attention

Chris Lucas
(Slides adapted from Frank Keller's)

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

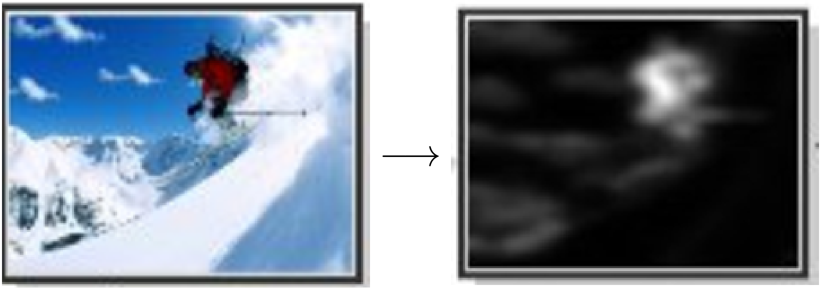
21 November 2019

1. Introduction
   - Visual Saliency
   - Contextual Guidance
2. Contextual Guidance Model
   - Model Architecture
   - Local Features
   - Global Features
   - Context
3. Results
   - Eye-tracking Data
   - Evaluation

Reading: Torralba, Oliva, Castelhano, and Henderson (2006).

Visual Saliency

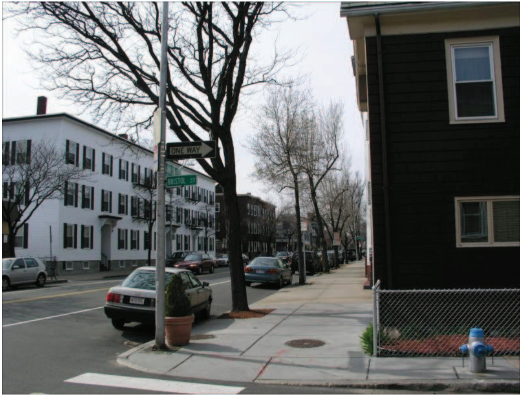
We attend to the areas that are *visually salient.* An area is salient if it stands out, is different from the rest of the image.



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

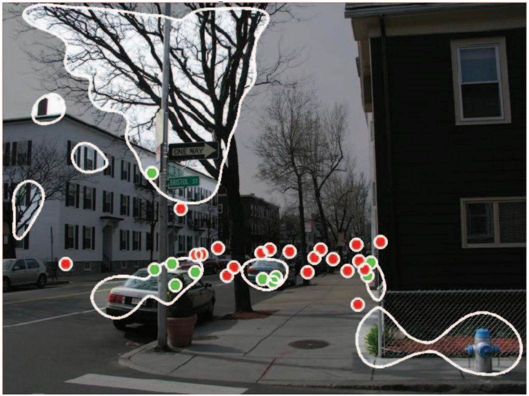
Contextual Guidance

However, visual search in real-world scenes is different from visual search in artificial stimuli (arrays of lines, etc.). Ex: person search:



Contextual Guidance

However, visual search in real-world scenes is different from visual search in artificial stimuli (arrays of lines, etc.). Ex: person search:



Saliency map (white area) is quite different from actual fixations.

Contextual Guidance

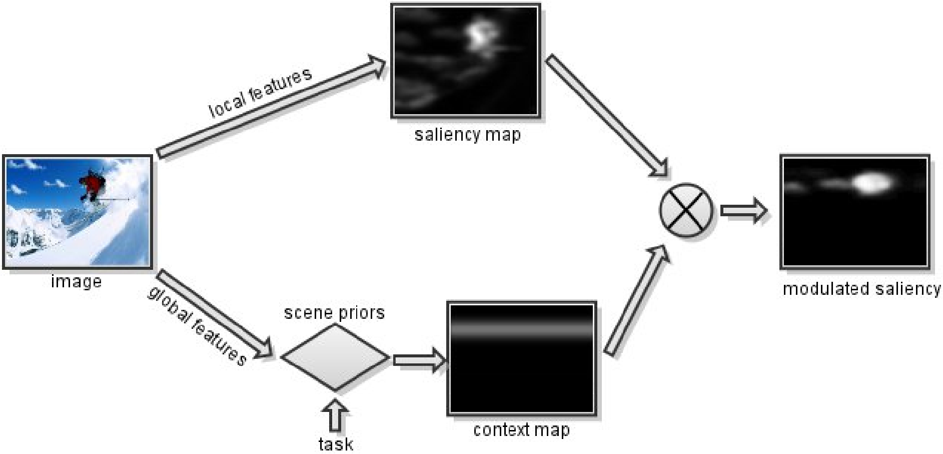
In real-world scenes, factors in addition to saliency guide search, including knowledge of:

- *which* objects occur in given type of scene (e.g., a kitchen typically contains an oven);
- where objects tend to be located (e.g., a person typically is on the ground);
- how objects are related to each other and co-occur (e.g., computer and monitor occur together);

People might also use different strategies for different tasks (e.g., search vs. memorization).

We will look at the *Contextual Guidance Model* (Torralba et al., 2006), which combines contextual knowledge with saliency.

Model Architecture



Model Architecture

The Contextual Guidance Model (CGM) combines saliency with *scene gist* to model visual search.

Intuitively, scene gist represents what type of scene we're dealing with (indoor scene, street scene, landscape, etc.).

Model Architecture

The CGM computes the *probability that target object O is present at point $X = (x, y)$* in the image:

$$\begin{aligned} p(O = 1, X|L, G) &= \frac{1}{p(L|G)} p(L|O = 1, X, G) \cdot \\ & p(X|O = 1, G) p(O = 1|G) \end{aligned} \quad (1)$$

where L is a set of *local image features* at X and G is a set of *global image features* representing gist.

Model Architecture

Components of the CGM in Equation (1):

- $\frac{1}{p(L|G)}$ is a saliency model (implemented differently from Itti et al., but same idea);
- $p(L|O=1,X,G)$ enhances locations with features that are consitent with beliefs about target's appearance;
- $p(X|O=1,G)$ is the contextual prior, provides information about likely target locations;
- $p(O=1|G)$ is the probability that O is present in the scene.

Note that unlike Itti's model, the CGM is fully probabilistic.

Model Architecture

The CGM basically combines two components: a *saliency map* and *a context map.*

In the implementation of the CGM, Equation (1) is simplified to:

$$S(X) = \frac{1}{p(L|G)} p(X|O = 1, G) \tag{2}$$

Contextually modulated saliency $S(X)$ is saliency combined with a prior over target locations, conditioned scene gist.

Local Features

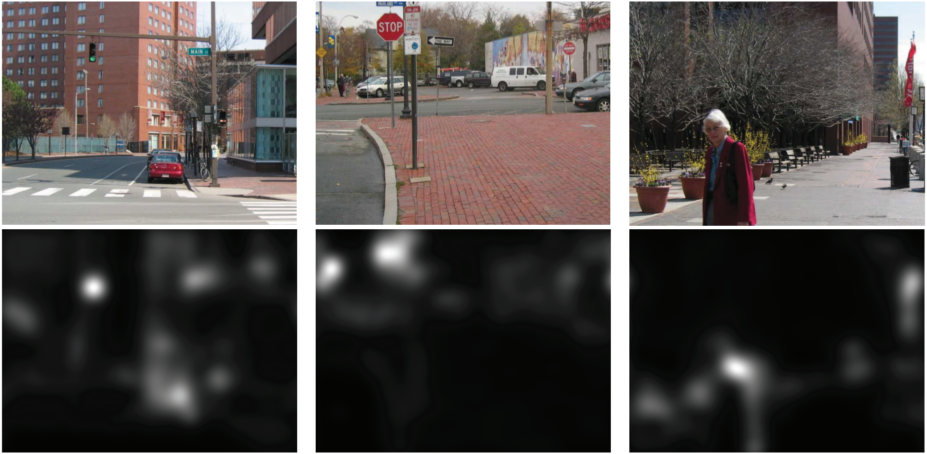
Probabilistic definition of saliency: local image features are salient when they are statistically distinguishable from the background.

Compute local image features (vector L):

- compute orientation features separately for the three color channels, at 6 orientations and 4 scales (Steerable pyramid);
- model the resulting distribution using a multivariate power-exponential (generalization of Gaussian);
- then compute $p(L|G)$, distribution of local features conditioned on global features.

Local Features

Examples for saliency maps:



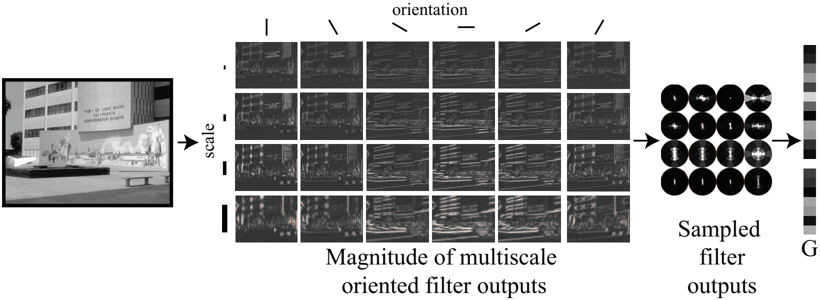
Global Features

Before computing saliency or performing object recognition, the visual system computes a global summary of the image (gist).

This can be simulated by pooling the outputs of local feature detectors across large regions of the visual field:

- compute luminance (intensity) as mean of red, green, blue values;
- compute orientation features for luminance at 6 orientations and 4 scales (Steerable pyramid);
- compute the average of each feature over 4 × 4 non-overlapping windows in the image;
- reduce resulting vector G using principal component analysis.

Global Features



Context

The context component of the CGM associates a scene type (i.e., a set of global features) with likely target locations.

Example: When searching for people in a street scene, locations in the bottom half of the scene are likely.

Assume the expected target location is a *weighted mixture of the target locations* in all scenes:

$$p(X, G|O = 1) = \sum_{n=1}^{N} P(n)p(X|n)p(G|n)$$

where we assume that the scenes are clustered into N *prototypes:* $P(n)$ is the weight, $p(X|n)$ the distribution of target locations, $p(G|n)$ the distribution of global features for prototype n.

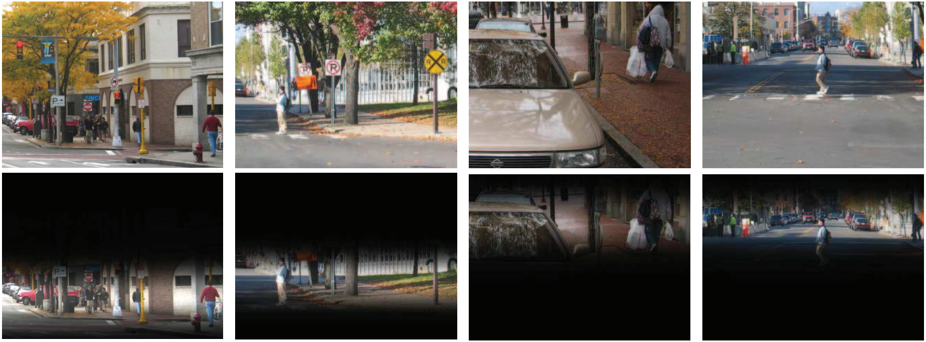
Context

Implementation of context model:

- context only predicts the vertical location of the target object (horizontal location is unconstrained);
- so we can approximate $p(X|O = 1, G)$ as $p(y|O = 1, G)$;
- the model parameters can be estimated using the expectation maximization algorithm (this infers the prototypes);
- train model on images containing three types of target objects: people in street scenes; paintings and mugs in indoor scenes (around 300 images per object type);
- number of prototypes set to $N = 4$;
- then compute modulated saliency map $S(X)$ as weighted product of saliency map and context map.

Context

Context prototypes for people in street scenes:



Eye-tracking Data

The CGM predicts fixation locations, so it can be evaluated against eye-tracking data.

Collect eye-tracking from participants performing visual search:

- task was to count the number of people, mugs, or paintings present in the image;
- a total of 72 images were used with up to six targets each;
- about half the images contained no target;
- participants could take up to 10 s for the task; accuracy was the same for target-present and target-absent conditions.

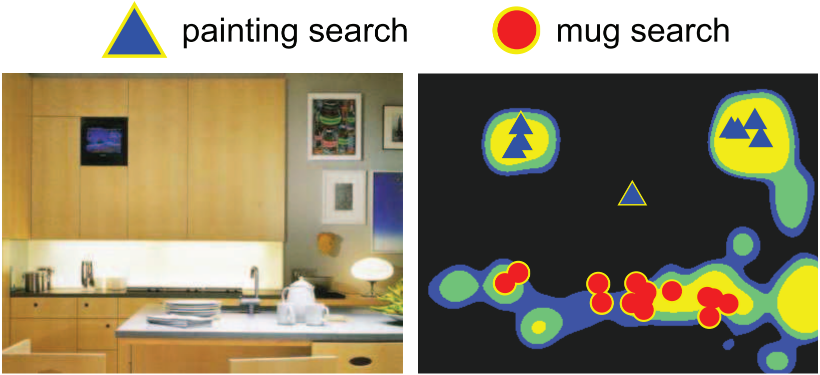
Evaluation

Use eye-tracking data to evaluate the CGM:

- compare how well the CGM predicts fixation locations compared to saliency alone for the three search tasks;
- the model outputs a probability value for each image location;
- apply a threshold to this probability so that the model selects a fixed percentage of the image (here 20%);
- then count how many fixations fall within the selected region;
- chance baseline: 20% correct; upper limit: consistency across participants;
- check also how fixation number influences performance (hypothesis: CGM models early stages of visual search).

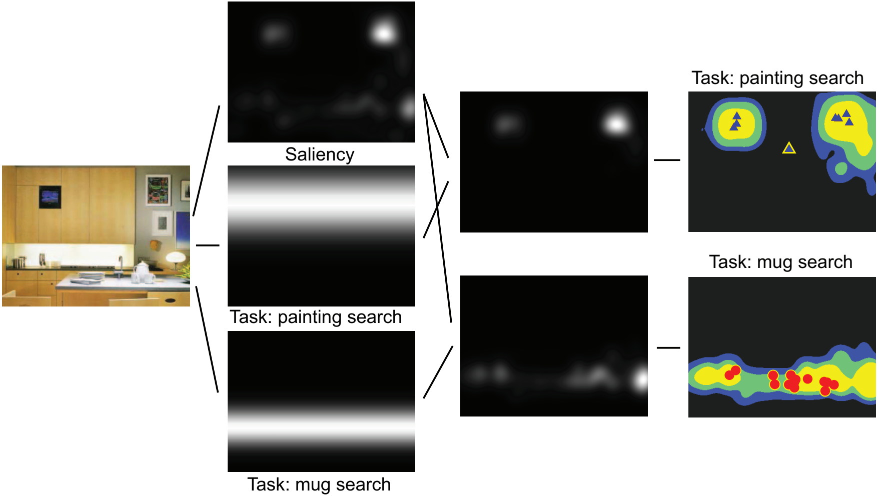
Evaluation

Example: saliency model vs. fixations for painting and mug search:



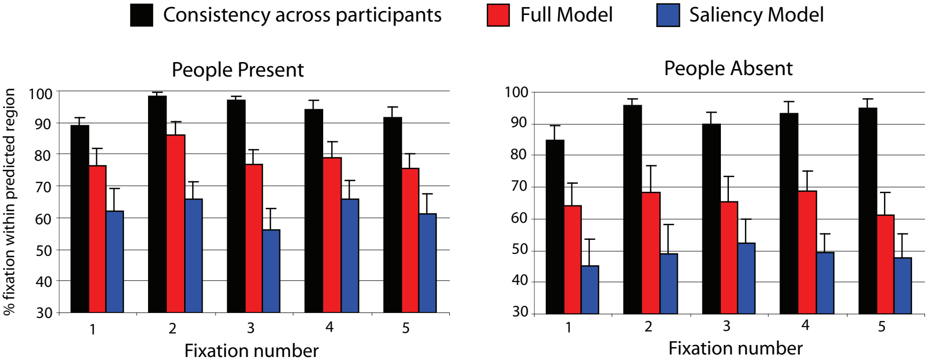
Evaluation

Example: CGM vs. fixations for painting and mug search:



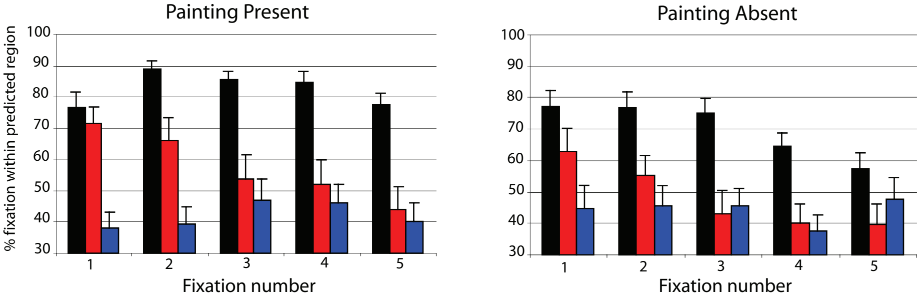
Evaluation

CGM evaluation by fixation number:



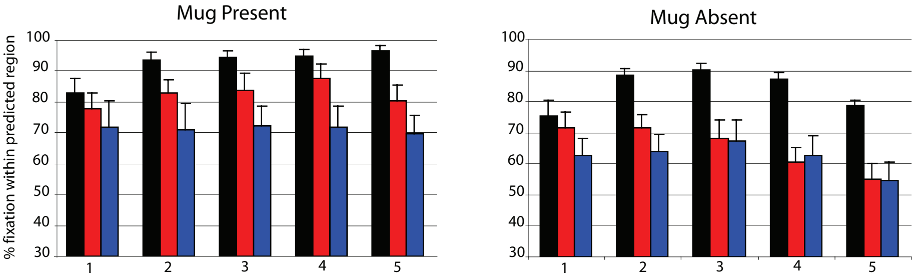
Evaluation

CGM evaluation by fixation number:



Evaluation

CGM evaluation by fixation number:



Summary

- Information about scene context is important for visual search;
- the Contextual Guidance Model combines saliency with context, both conditioned on scene gist, to compute likely fixation locations;
- gist is essentially an orientation/intensity map of the scene at a coarse scale;
- context is modeled a distribution over likely vertical locations of the target object;
- the CGM successfully models eye-tracking data on visual search in photorealistic scenes.

References

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.

Torralba, A., Oliva, A., Castelhano, M., & Henderson, J. M. (2006). Contextual guidance of attention in natural scenes: The role of global features on object search. *Psychological Review*, *113*(4), 766–786.