EECS 4422/5323 Computer Vision

Feature Detection Lecture 4

Calden Wloka

30 September, 2019

Announcments

- Labs this week: panoramic image stitching
- There isn't always one right approach on the assignments - so long as you have read the question carefully and followed the instructions which are specified, you will not be penalized for a different interpretation or approach

Outline

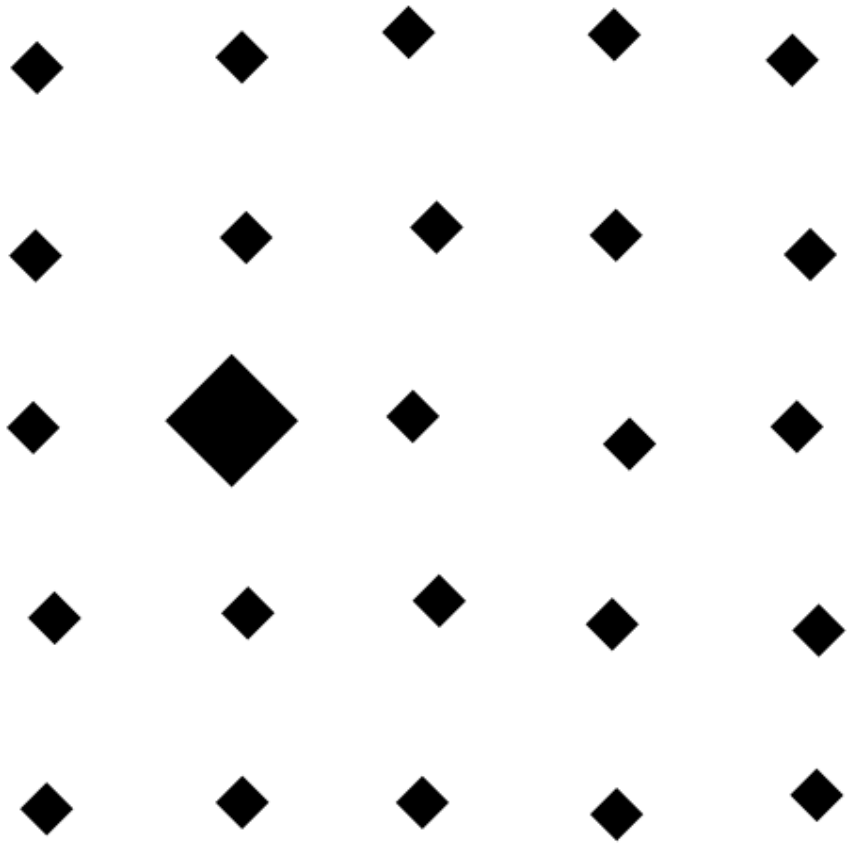
- Definitions of Saliency
  - Proscribed features
  - Information theoretic formulation
  - Learned function
- Applications of Saliency

What is Saliency?

First, it is important to note what is meant by the term *saliency*: the stimulus-driven attentional pull of a visual element.

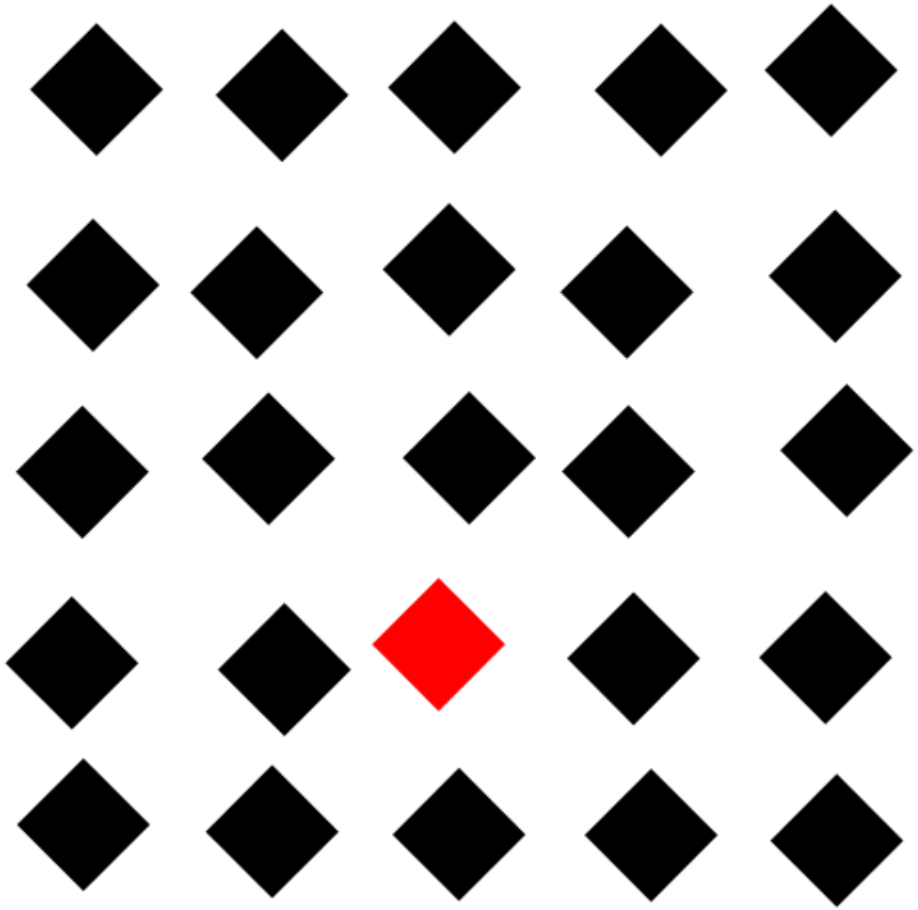
What is Saliency?

First, it is important to note what is meant by the term *saliency*: the stimulus-driven attentional pull of a visual element.



What is Saliency?

First, it is important to note what is meant by the term *saliency*: the stimulus-driven attentional pull of a visual element.



What is Saliency?

First, it is important to note what is meant by the term *saliency*: the stimulus-driven attentional pull of a visual element.

What is Saliency?

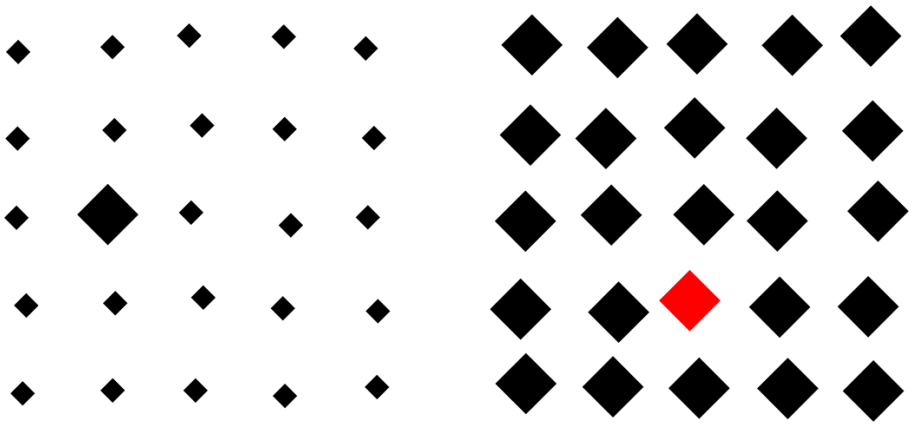
First, it is important to note what is meant by the term *saliency*: the stimulus-driven attentional pull of a visual element.

What is Saliency?

First, it is important to note what is meant by the term *saliency*: the stimulus-driven attentional pull of a visual element.

Saliency

In summary, saliency is stimulus-driven attentional pull based on local contrast (Koch and Ullman, 1985), and not necessarily consistent with the most semantically interesting elements of an image.



Stimulus-Driven Attention

Saliency can in some ways be viewed as attention prior to understanding; it is the "best guess" for what will be most important to process further.

In humans, saliency is often overridden by endogenous (self-driven) control, or by task or world knowledge. When applying saliency to a computer vision system, it is important to think about how to balance the benefits of a saliency module with the potential for it to make a poor decision or assignment of saliency values.

Saliency and Human Vision

In the original framing of a saliency model (Koch and Ullman, 1985), the aim was stated as understanding *selective visual attention*, particularly in the human visual system.

A good understanding of human selective attention provides:

Saliency and Human Vision

In the original framing of a saliency model (Koch and Ullman, 1985), the aim was stated as understanding *selective visual attention*, particularly in the human visual system.

A good understanding of human selective attention provides:

- A strong model of information prioritization in complex environments.

Saliency and Human Vision

In the original framing of a saliency model (Koch and Ullman, 1985), the aim was stated as understanding *selective visual attention*, particularly in the human visual system.

A good understanding of human selective attention provides:

- A strong model of information prioritization in complex environments.
- Insight into human visual cognition.

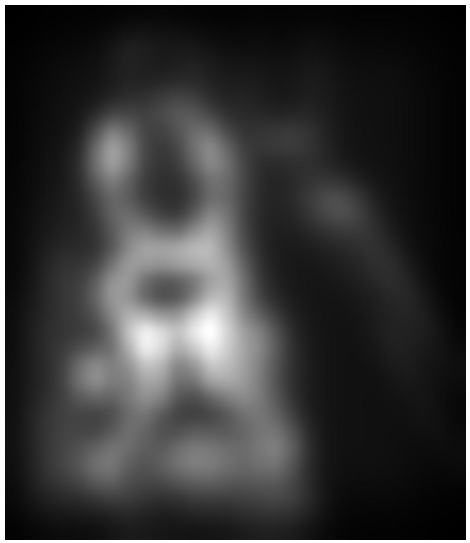
The Saliency Map

The dominant computational model of salience is the *saliency map*, which outputs a pixel-wise assignment of salience to an input image in the form of a heatmap.

The Saliency Map

The dominant computational model of salience is the *saliency map*, which outputs a pixel-wise assignment of salience to an input image in the form of a heatmap.

Input Image

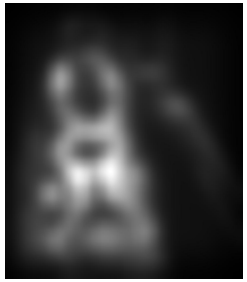


IKN Map

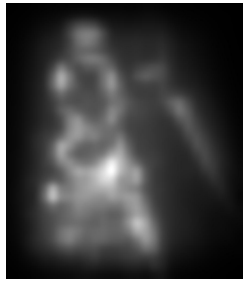
Itti *et al.* (1998) released the first widely available implementation of a saliency model in the form of the Itti-Koch-Niebur (IKN) model.

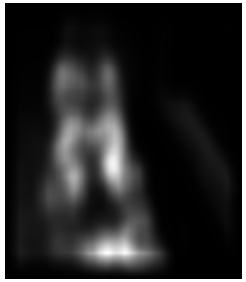
There are a lot of Saliency Map Models...


Input


Itti *et al*. 1998


Bruce & Tsotsos 2006


Harel *et al*. 2006

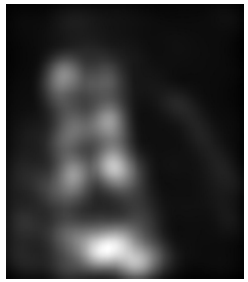

Hou & Zhang 2008


Seo & Milanfar 2009

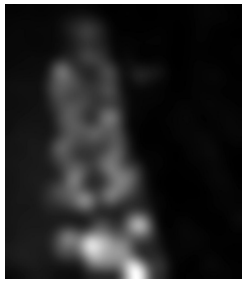

Goferman *et al*. 2010

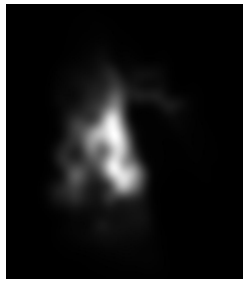

Tavakoli *et al*. 2011

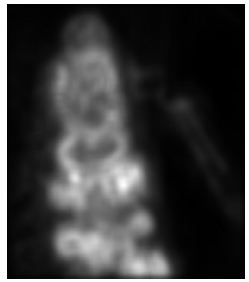

Garcia-Diaz *et al*. 2012

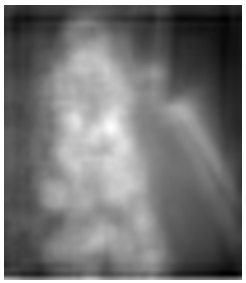

Hou *et al*.2012


Schauerte & Stiefelhagen 2012

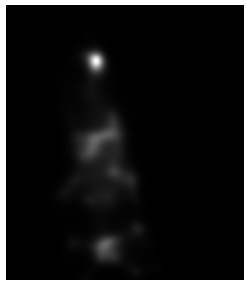

Zhang & Sclaroff 2013

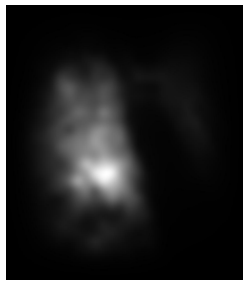

Erdem & Erdem 2013

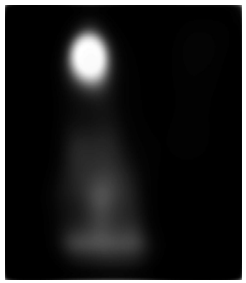

Riche *et al*. 2013

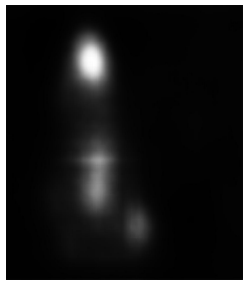

Vig *et al*. 2014

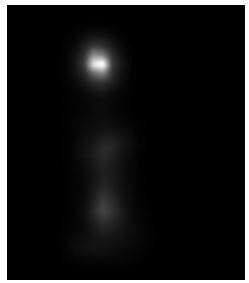

Huang *et al*. 2015


Kümmerer *et al*. 2016


Fang *et al*. 2017


Pan *et al*. 2017


Wang & Shen 2018


Cornia *et al*. 2018

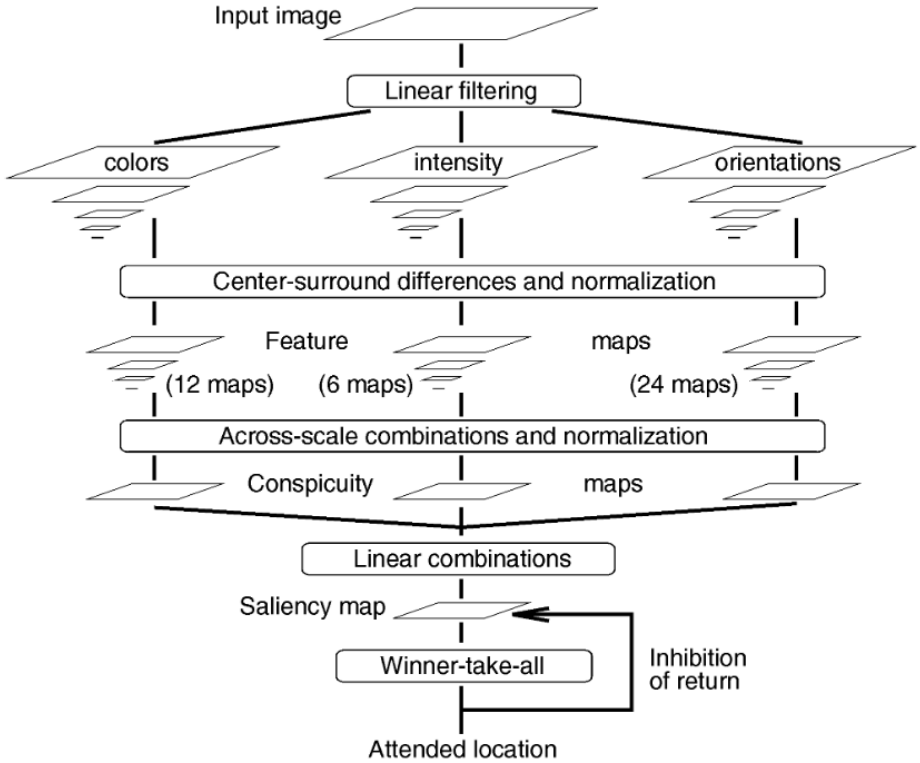
Saliency by Design

A popular class of saliency model is constructed by explicitly picking the features which the designer believes will most effectively detect salient stimuli, and then designing the rule by which those features will be mapped onto a final output map.

We will discuss in detail one such model: the Itti-Koch-Niebur (IKN) model.

Overview of the IKN model

- Multi-scale processing through a Gaussian pyramid
- Features are computed along three feature channels:
  - Colour
  - Intensity
  - Orientation
- Normalization looks for unique peaks of activity



IKN Features

- Intensity features are computed directly from the pyramid as a difference of Gaussians, $I(c, s) = |I(c) \ominus I(s)|$

IKN Features

- Intensity features are computed directly from the pyramid as a difference of Gaussians, $I(c, s) = |I(c) \ominus I(s)|$
- Colour features are computed as a difference of Gaussians over colour opponency channels

IKN Features

- Intensity features are computed directly from the pyramid as a difference of Gaussians, $I(c, s) = |I(c) \ominus I(s)|$
- Colour features are computed as a difference of Gaussians over colour opponency channels
  - $RG(c, s) = |(R(c) - G(c)) \ominus (G(s) - R(s))|$

IKN Features

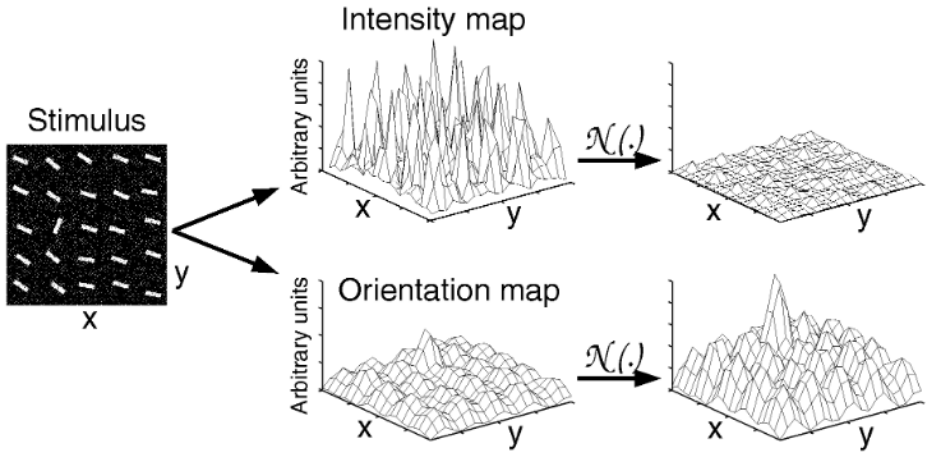
- Intensity features are computed directly from the pyramid as a difference of Gaussians, $I(c,s) = |I(c) \ominus I(s)|$
- Colour features are computed as a difference of Gaussians over colour opponency channels
  - $RG(c,s) = |(R(c) - G(c)) \ominus (G(s) - R(s))|$
  - $BY(c,s) = |(B(c) - Y(c)) \ominus (Y(s) - B(s))|$

IKN Features

- Intensity features are computed directly from the pyramid as a difference of Gaussians, $I(c,s) = |I(c) \ominus I(s)|$
- Colour features are computed as a difference of Gaussians over colour opponency channels
  - $RG(c,s) = |(R(c) - G(c)) \ominus (G(s) - R(s))|$
  - $BY(c,s) = |(B(c) - Y(c)) \ominus (Y(s) - B(s))|$
- Orientation features are computed by the difference over scales of Gabor filters along orientation directions

IKN Normalization

Normalization in the IKN model seeks to amplify unique activity rather than just cast values onto an allowable range.
In addition to mapping values to a proscribed range, for each feature map the average of local maxima, $\bar{m}$, is calculated, and then the map is globally multiplied by $(M - \bar{m})$, where M is the global maximum for that feature map.



Some Notes for IKN

- A number of additional computational tricks and techniques have been added to the model over the years, but aren't well documented (*e.g.* center bias)

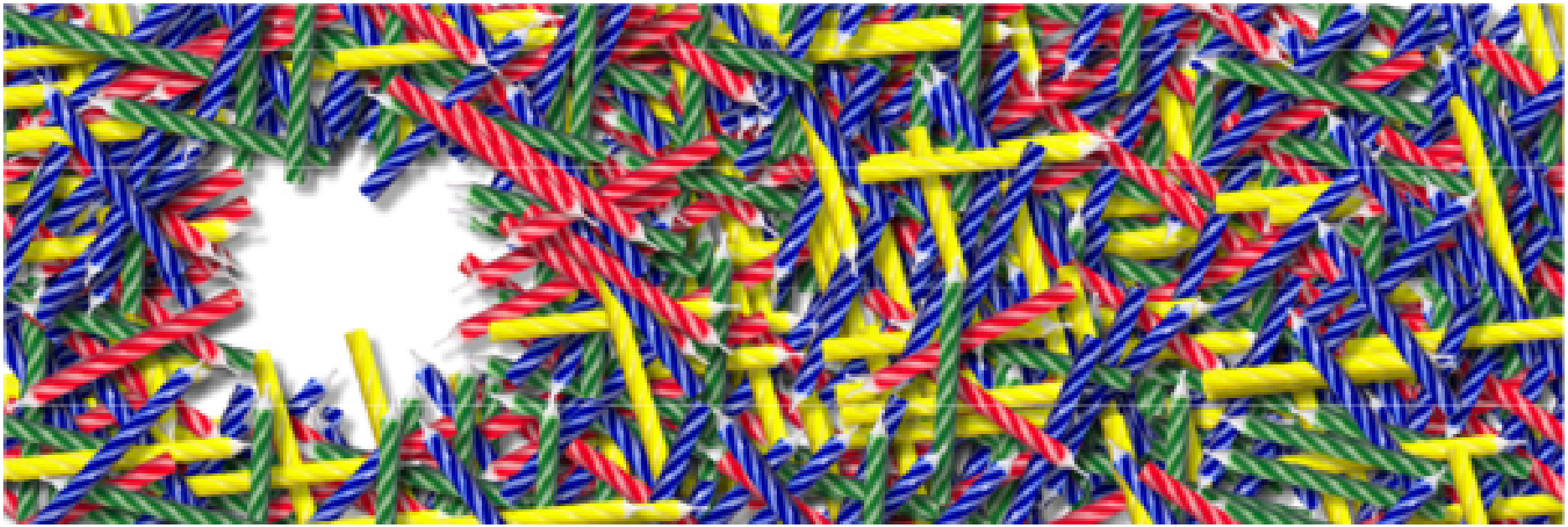
Some Notes for IKN

- A number of additional computational tricks and techniques have been added to the model over the years, but aren't well documented (*e.g.* center bias)
- IKN is also just referred to as "the Itti model", and remains popular for its intuitiveness and code access

Some Notes for IKN

- A number of additional computational tricks and techniques have been added to the model over the years, but aren't well documented (*e.g.* center bias)
- IKN is also just referred to as "the Itti model", and remains popular for its intuitiveness and code access
- Because saliency is based on peaks of filter activation, it is hard for IKN to detect salient absences

A Motivation from First Principles

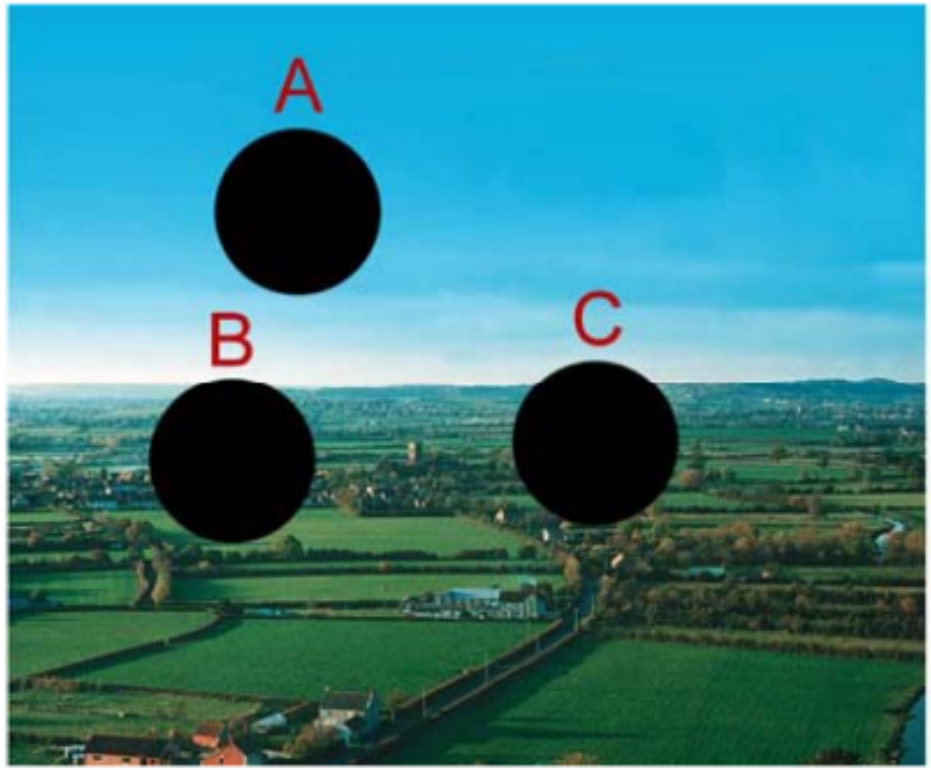


Information theoretic models attempt to frame saliency not based on a set of proscribed perceptual rules, but rather from a mathematical definition of the problem attempting to be solved (*i.e.* determining what parts of a scene carry the most pressing information).

We will discuss in detail one such model: the Attention by Information Maximization (AIM) model.

The Intuition Behind AIM

Try to mentally fill in what you expect to see behind each of the black circles in the image on the right (labeled A-C).



The Intuition Behind AIM

Patch C contains the content which is least predictable, and therefore the most salient.

The Steps of AIM

1. Filter the input image, output a set of feature maps
2. Process the features, output transformed features
3. Combine feature information, output a raw saliency map
4. Perform post-processing, output a final saliency map

Image Filtering

The AIM calculation uses a mathematical simplification which relies on the *independence* of the computed features.

- In the original formulation, independent filters are derived as a set of basis functions computed through Independend Component Analysis (ICA)

Image Filtering

The AIM calculation uses a mathematical simplification which relies on the *independence* of the computed features.

- In the original formulation, independent filters are derived as a set of basis functions computed through Independend Component Analysis (ICA)
- ICA bases are calculated by sampling a large number of random image patches and breaking those patches down into the strongest independent components - a number of pre-computed bases are available in the released code

Image Filtering

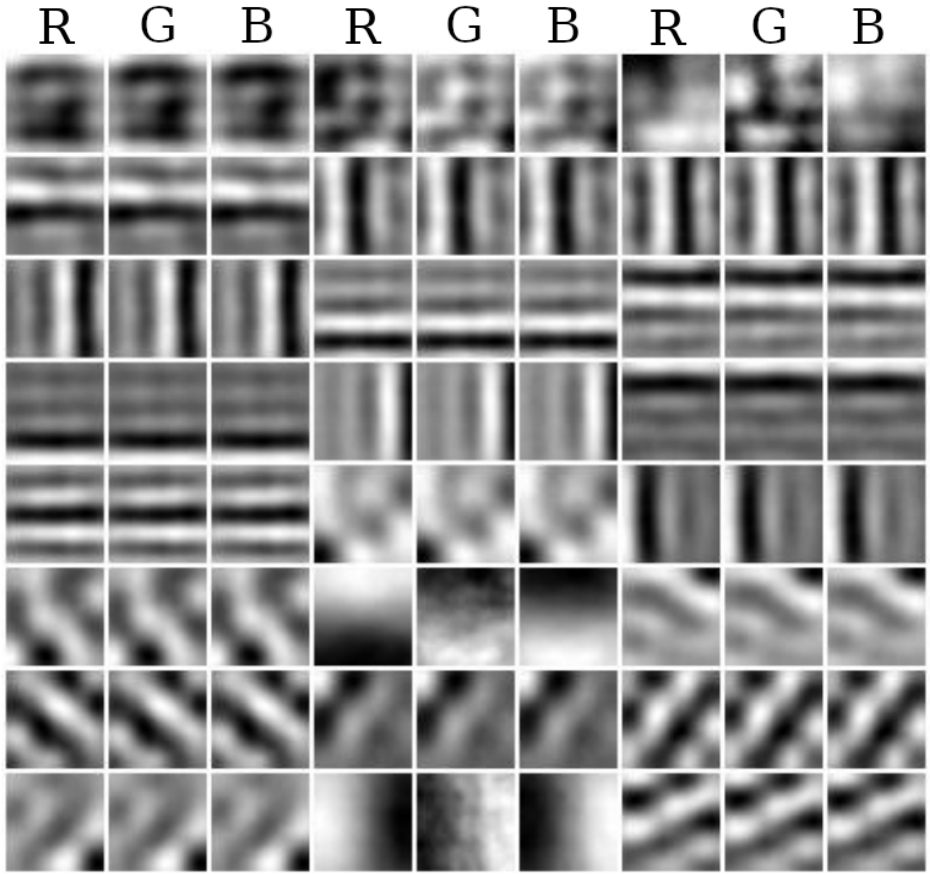
The AIM calculation uses a mathematical simplification which relies on the *independence* of the computed features.

- In the original formulation, independent filters are derived as a set of basis functions computed through Independend Component Analysis (ICA)
- ICA bases are calculated by sampling a large number of random image patches and breaking those patches down into the strongest independent components - a number of pre-computed bases are available in the released code
- Subsequent research (Bruce *et al.*, 2011) demonstrated that it is largely sufficient for the features to be *mostly* independent

AIM Filters Example

An example of the first 24 filters in the `21jade950` pre-trained filter set for AIM.

Each row contains the RGB layers of three filters.



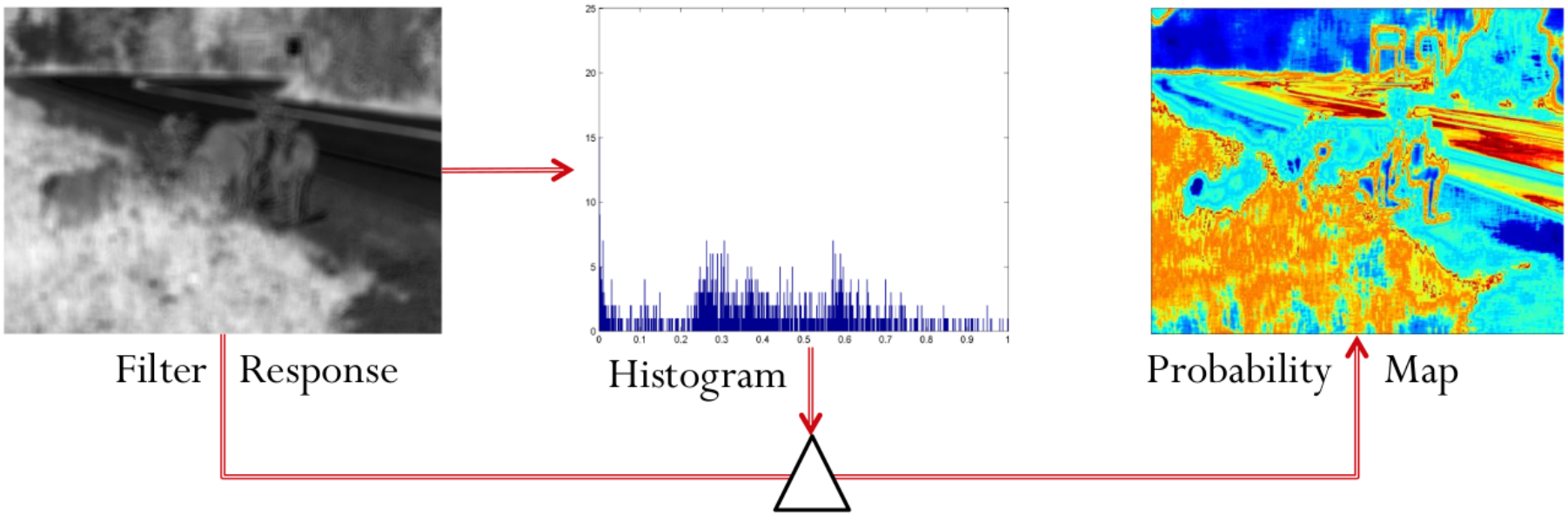
Process the Features

The goal of AIM is to convert feature responses into a measure of the information in an image. The algorithm estimates this by building a probability distribution for each feature.

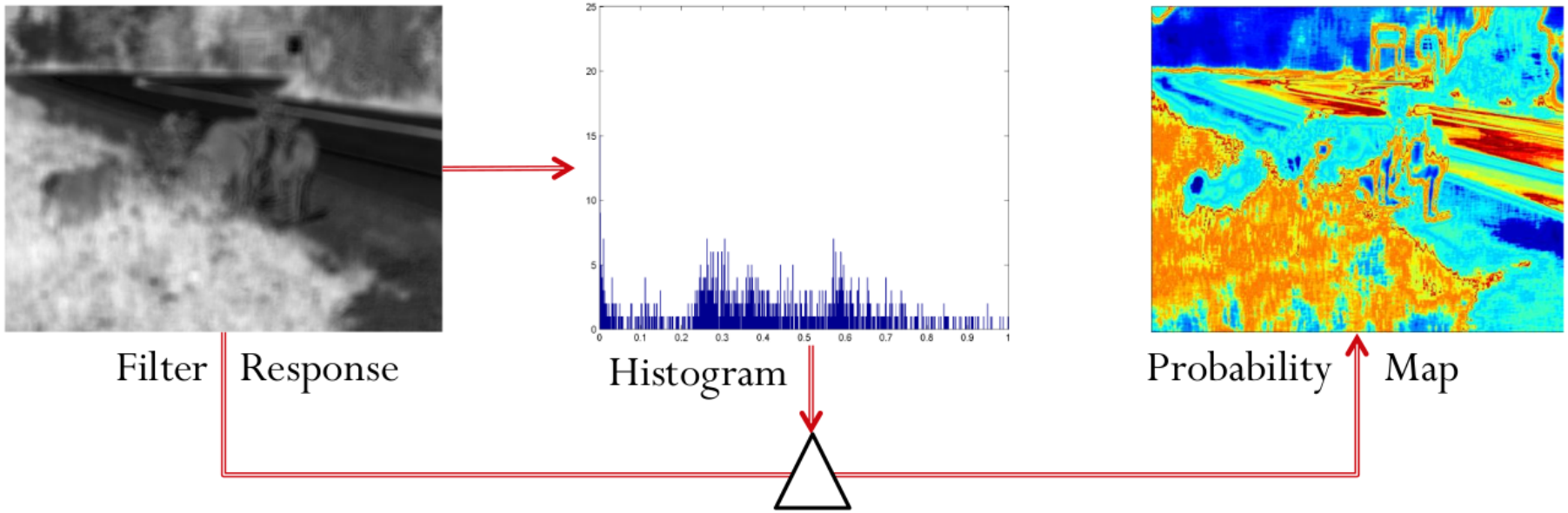
For example, take the input image shown on the right.

Process the Features

The output and processing for each filter is as follows:



Combine Features



Since the filters are assumed to be independent, the probability of pixel (i, j) is the product of the probability over all filters,

$$p_{i,j} = \prod_{f \in \mathbf{F}} p_{i,j,f}$$

where $\mathbf{F}$ is the set of all filters, and $p_{i,j,f}$ is the probability assigned to pixel (i, j) from the contribution of filter f.

Self-Information

Self-information is based on the fact that rare events are more informative, and is calculated as:

$$S = -\log(p)$$

Self-Information

Self-information is based on the fact that rare events are more informative, and is calculated as:

$$S = -\log(p)$$

Thus, we take the salience of pixel (i, j) to be:

$$\mathbf{S}_{i,j} = -\log\left(\prod_{f \in \mathbf{F}} p_{i,j,f}\right)$$

Self-Information

Self-information is based on the fact that rare events are more informative, and is calculated as:

$$S = -\log(p)$$

Thus, we take the salience of pixel (i, j) to be:

$$\mathbf{S}_{i,j} = -\log\left(\prod_{f \in \mathbf{F}} p_{i,j,f}\right) = -\sum_{f \in \mathbf{F}} \log(p_{i,j,f}$$

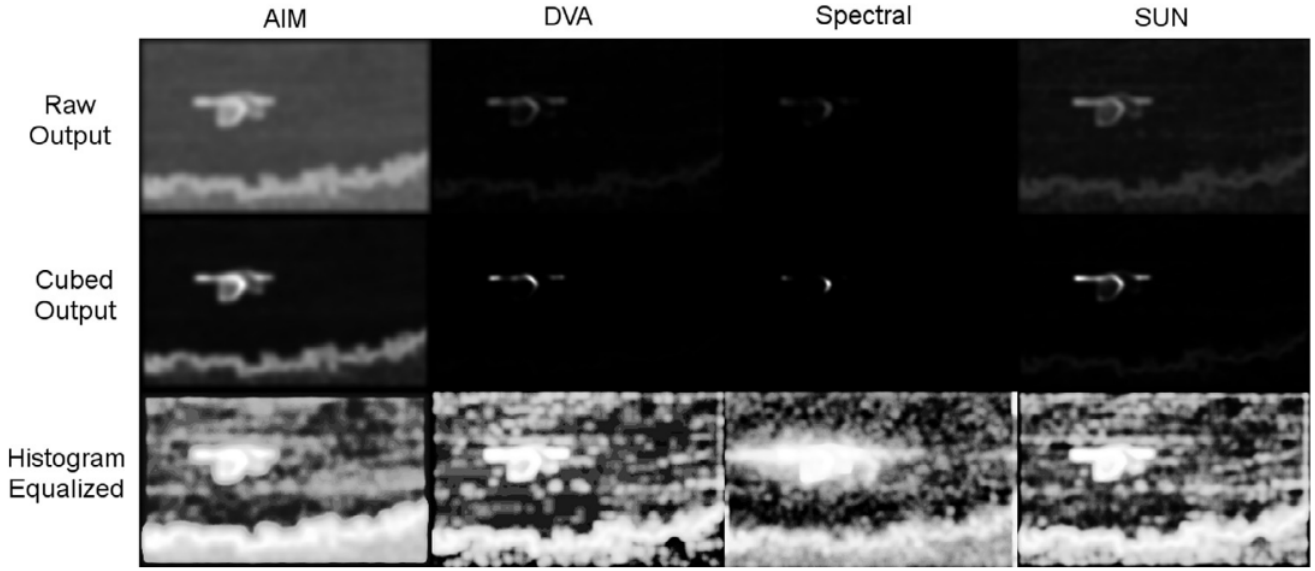
Post-Processing

The AIM model introduced the idea of performing Gaussian smoothing to the saliency map as a post-processing step. This has since become standard practice across many saliency models.

Post-Processing

The AIM model introduced the idea of performing Gaussian smoothing to the saliency map as a post-processing step. This has since become standard practice across many saliency models.

However, numerical manipulation of the output values while still maintaining rank-order can be very useful both for visualization, but also for possible applications such as thresholding and blob detection.



A Focus on Performance

Learned saliency representations interpret saliency as a pattern-matching exercise in which some training signal is taken as indicative of the desired saliency output, and then a mapping from the input to that output is learned.

A Focus on Performance

Learned saliency representations interpret saliency as a pattern-matching exercise in which some training signal is taken as indicative of the desired saliency output, and then a mapping from the input to that output is learned.

A number of methods have been proposed for this, but most recent efforts follow a similar overall pattern, which we will discuss.

Transfer Useful Features

- Rather than attempt to specify or derive specific features for saliency, most learned methods select features from a neural network trained on another visual problem

Transfer Useful Features

- Rather than attempt to specify or derive specific features for saliency, most learned methods select features from a neural network trained on another visual problem
- The most common network chosen is the Visual Geometry Group (VGG, Simonyan & Zisserman, 2014) network, trained for image recognition

Transfer Useful Features

- Rather than attempt to specify or derive specific features for saliency, most learned methods select features from a neural network trained on another visual problem
- The most common network chosen is the Visual Geometry Group (VGG, Simonyan & Zisserman, 2014) network, trained for image recognition
- The top layers of the network (which perform the final recognition steps) are removed, and the internal network features are retained

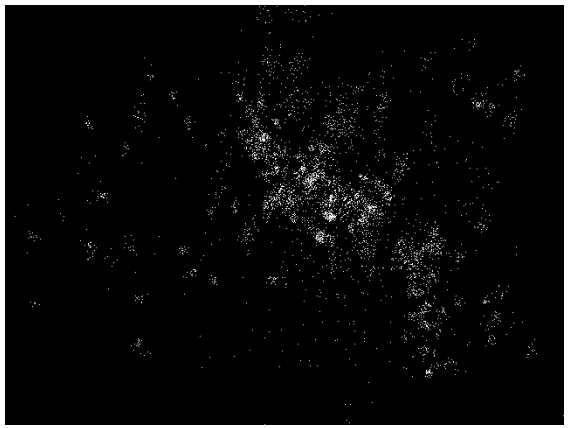
Transfer Useful Features

- Rather than attempt to specify or derive specific features for saliency, most learned methods select features from a neural network trained on another visual problem
- The most common network chosen is the Visual Geometry Group (VGG, Simonyan & Zisserman, 2014) network, trained for image recognition
- The top layers of the network (which perform the final recognition steps) are removed, and the internal network features are retained
- Saliency is then extracted from these remaining features; the way in which this extraction occurs and is trained is typically what separates the different learned models

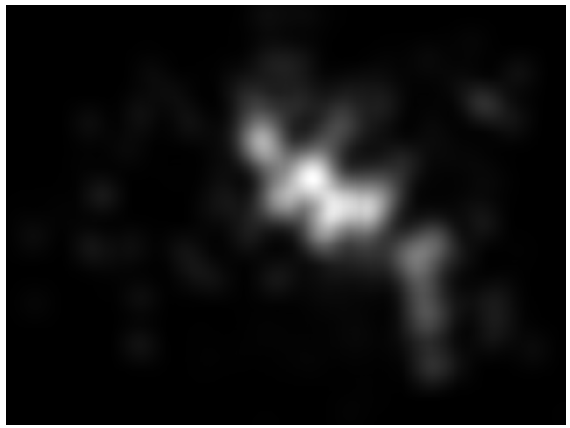
The Training Signal

Almost all learned methods are focused on the task of *fixation prediction* - predicting where in an image a human observer's gaze will most likely fall.

Sample Input



Human Fixations



Human "Saliency Map"

Sample image and human fixations taken from the ImgSal dataset (Li *et al.* 2013).

Saliency Output

Although saliency maps can be useful visualization tools for advertising or graphic design, it is often most effective as part of a larger processing pipeline. Examples include:

- Anisotropic image or video compression

Saliency Output

Although saliency maps can be useful visualization tools for advertising or graphic design, it is often most effective as part of a larger processing pipeline. Examples include:

- Anisotropic image or video compression
- Intelligent thumbnail cropping or image retargeting

Saliency Output

Although saliency maps can be useful visualization tools for advertising or graphic design, it is often most effective as part of a larger processing pipeline. Examples include:

- Anisotropic image or video compression
- Intelligent thumbnail cropping or image retargeting
- Defect detection

Saliency Output

Although saliency maps can be useful visualization tools for advertising or graphic design, it is often most effective as part of a larger processing pipeline. Examples include:

- Anisotropic image or video compression
- Intelligent thumbnail cropping or image retargeting
- Defect detection
- Image quality assessment

Saliency Output

Although saliency maps can be useful visualization tools for advertising or graphic design, it is often most effective as part of a larger processing pipeline. Examples include:

- Anisotropic image or video compression
- Intelligent thumbnail cropping or image retargeting
- Defect detection
- Image quality assessment
- Feature weighting, particularly for embedded systems like robot navigation

The Best Formulation of Saliency is Application Dependent

For specific pipelines, the best formulation for saliency may be different from the best formulation for another application area.

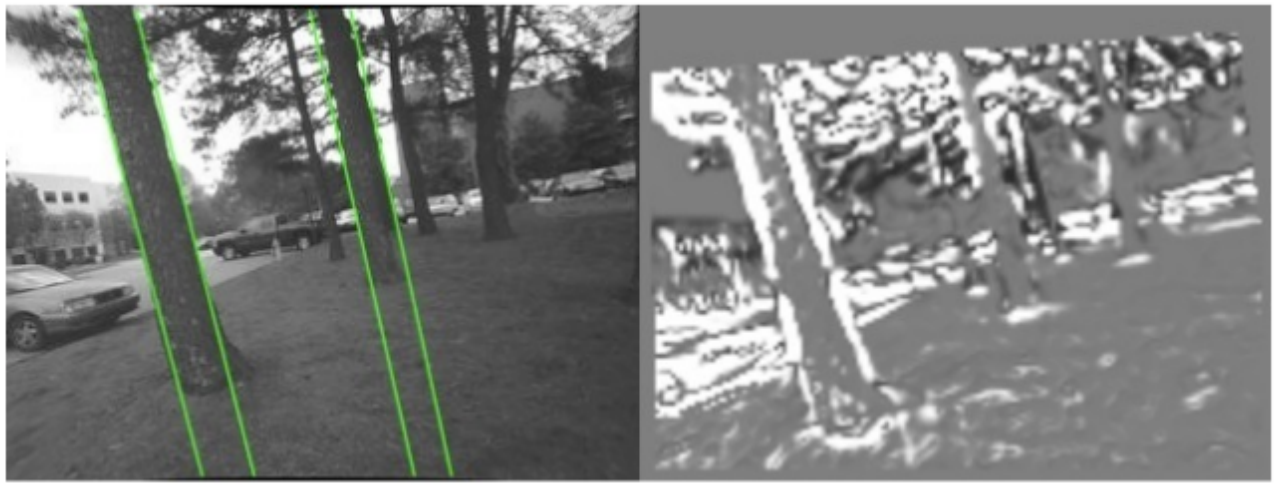


Image source: Roberts *et al.*, Saliency Detection and Model-based Tracking: A Two Part Vision System for Small Robot Navigation in Forested Environments, 2012

The Best Formulation of Saliency is Application Dependent

For specific pipelines, the best formulation for saliency may be different from the best formulation for another application area.

Sometimes it is necessary to develop a novel saliency representation specific for a given application, but often existing methods can be re-purposed. Thought should be given and tests run to determine which method is best suited for a given task.

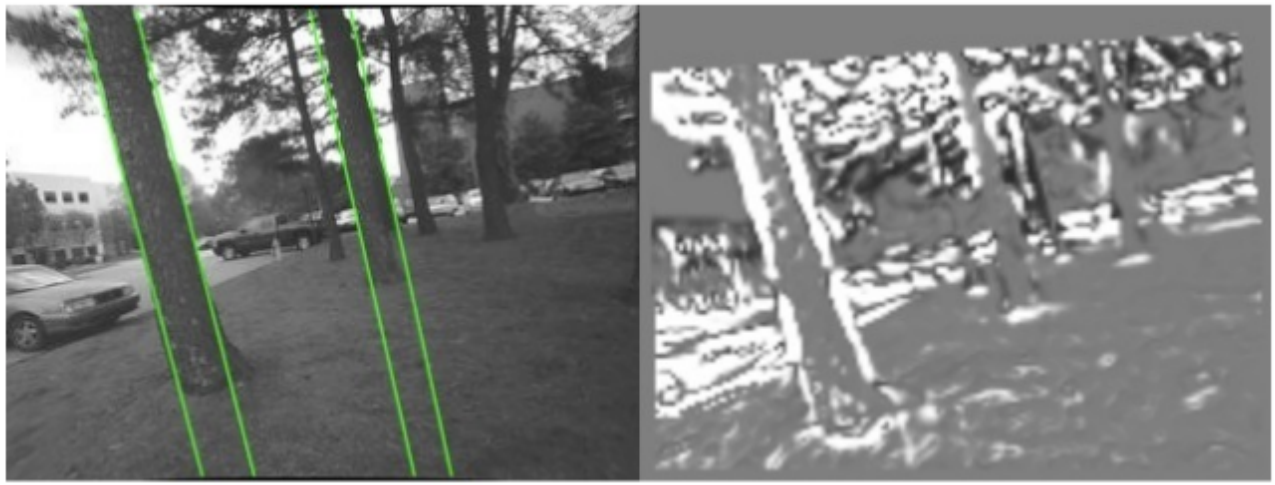


Image source: Roberts *et al.*, Saliency Detection and Model-based Tracking: A Two Part Vision System for Small Robot Navigation in Forested Environments, 2012