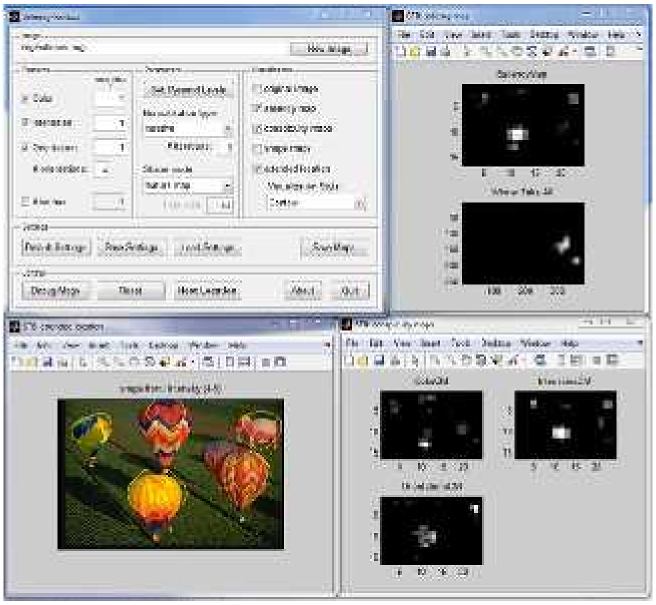
Meeting 2

Feb, 2017

Matlab Salience

- GUI – Salience toolbox
  - http://www.saliencytoolbox.net/
- Source code
  - http://www.vision.caltech.edu/~harel/share/gbvs.php



Further reading: Schiller, 1998

- Great flow chart of saccadic/visual system
- Lets use this as a guide to follow I&K

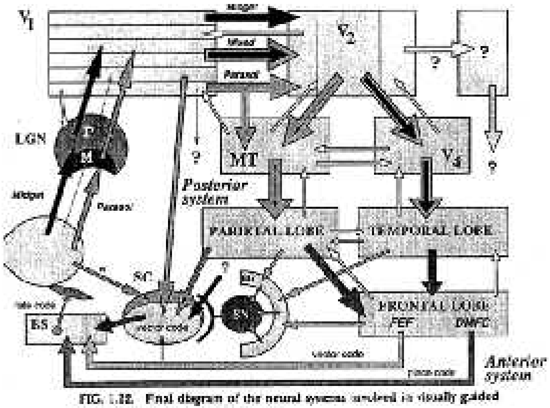


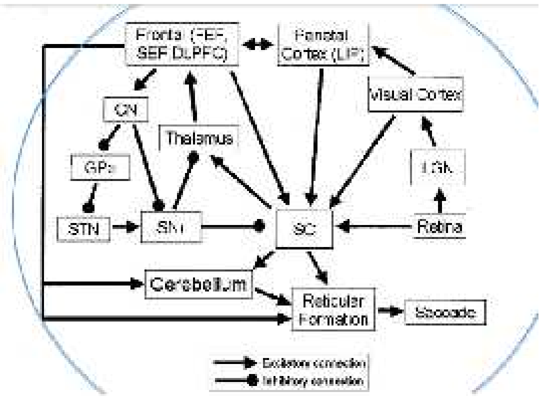
FIG. 1.22. Final diagram of the neural systems involved in visually guided saccadic eye-movement generation.

Itti and Koch

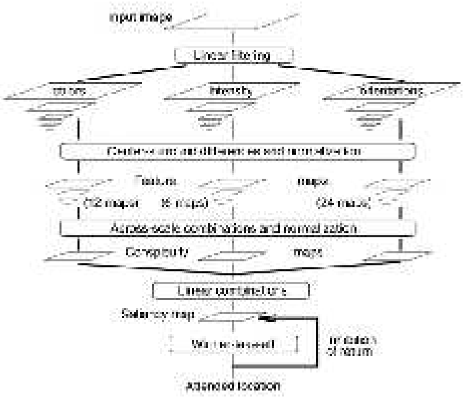
- One of the most influential models of the last 20 years
  - Vision in particular, but well known in all subfields
- Why?
  - Grounded in theory
  - Neurally plausible
  - Testable
  - Controversial? MANY attempts to improve, but not to challenge
- How much theory does a model need?
  - Itti&Koch marked up example

Munoz, 2002

- More programmatic schematic



- Start with graphic
- Pixel coordinate == retinal coordinate
- Like most models, there is a lot of data pre-processing
- But I&K made an effort to make that pre-processing part of the model itself

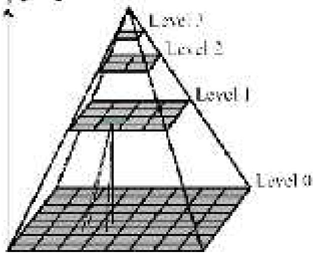


Dyadic Gaussian pyramid

- 1. low pass filter (blur)
- 2. Downsample (keep only every nth pixel)

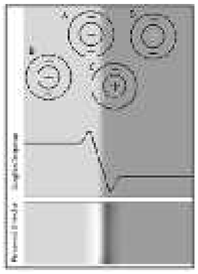
- This is done with nine spatial scales from 1:1 (no downsampling) to 2^8 (1:256)
- Massively parallel

- Implementation?
  - Matlab matrices
  - JAMA for Java
  - Imread() to load image
  - Imfilter() image processing toolbox

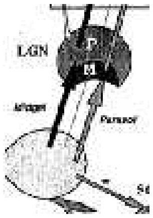


Feature extraction

- Modelling after edge detection from neurons with centre surround receptive fields
- Sensitivity to centre, while inhibiting surround are perfect for detecting features and edges
- I&K implemented as difference between pixel values at fine vs course scales
- Increased smoothing and downsampling reduces those edges,
  - Smoothing a flat surface multiple times, the change will be small
  - Smoothing a sharp image multiple times, the change will be greater
- Each pixel on scale n contains a local average that corresponds to an entire pixel **neighbourhood** on scale n + x of the pyramid.

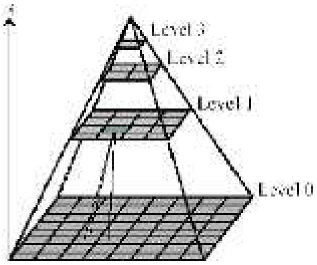


Eg: groups of retinal ganglion cells



Extraction (2)

- Centre(C) defined as pixel at scale 2,3 or 4
- Surround defined as pixel at scale C + 3 or 4
  - Its downsampled, remember? So that pixel includes averages of many surrounding pixels from that lower scale
- 'Multiscale' feature extraction improved single scale used previously

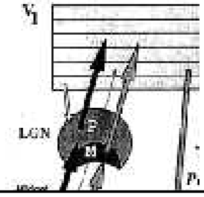


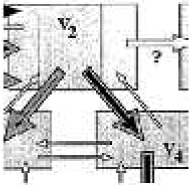
Intensity map

- Image pixels are often defined as 0..255 (8 bit) for each colour
- Intensity for image in a colour image is I = (R+G+B)/3
- 6 maps
  - Centre chosen from scale 2,3,4
  - Surround chosen from centre + 2 or 3
  - No justification given for these choices, and may have been trial and error
- Maps combine light centre with dark surround and dark centre with light surround
  - Performed separately then combined (rectified) though details on how are not provided

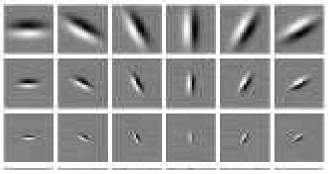
Colour map

- Red blue and green channels are normalized by the intensity channel
  - Otherwise colour maps would hue + intensity, and we already have separate intensity maps
- Colour values less than 1/10th of max were dropped
  - Hue differences at low intensity are difficult to perceive
  - Again, a good example of *theory driving the model*
- *Colour double opponent system*
  - Centre activated by one colour and surround inhibited by a second
  - Implemented Green/Red, Red/Green, Blue/Yellow and Yellow/Blue
- 12 maps in total
  - Possibly again, trial and error
- Those who study colour vision can claim this is a gross simplification of colour, but again... cat

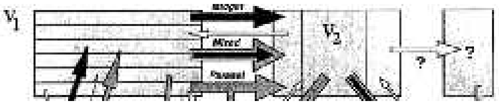




Orientation

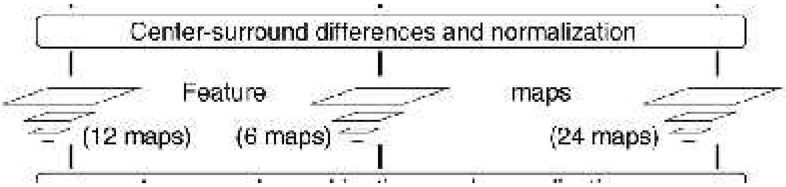


- Gabor filters mimic orientation selective neurons in primary visual cortex
  - (Maybe up to V4 which interprets colour and form)
  - 0, 45, 90 & 135 degrees



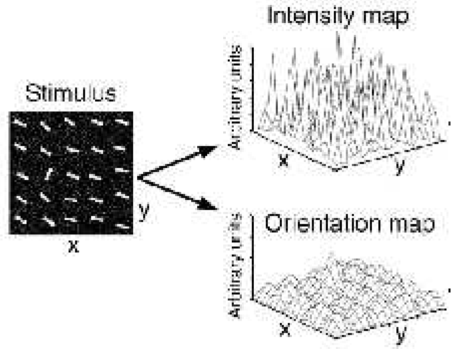
Layers

- 6 intensity maps
  - 3 centres x 2 surrounds
- 12 normalized colour maps
  - 3 centres x 2 surrounds x 2 colour dyads
- 24 orientation maps
  - 3 centres x 2 surrounds x 4 orientations

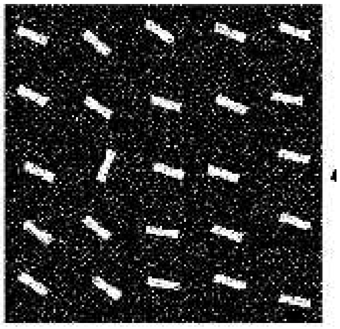


Problem

- ‘salience’ is relative



Can the most salient thing in this image be defined by intensity alone?

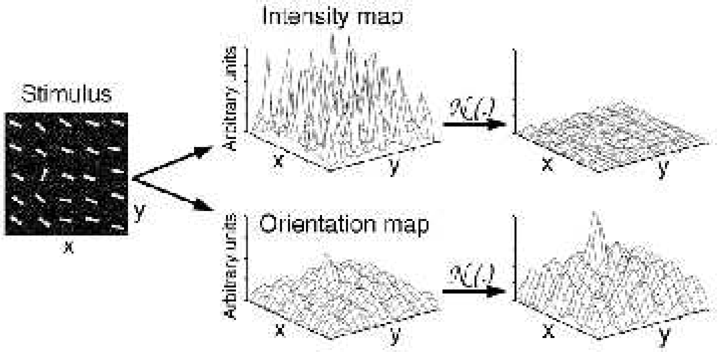


Problem

- Feature maps may not have immediately comparable modalities or ranges
- When combining all 42 maps, strong features in one or two maps can be drowned out by weak features that appear in all
- **This is an experimentally testable prediction for the model!**
- *Normality operator* in model promotes local features when combined
- Is this neurally plausible?
  - Not everything has to be
  - But could be related to cortical lateral inhibition
  - Large areas of activation inhibit activation in nearby areas
  - And multiple large area would be mutually inhibitory

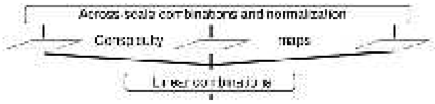
Normality

- normalizing the values in the map to a fixed range [0..M], in order to eliminate modality-dependent amplitude differences;
- finding the location of the map's global maximum **M** and computing the average *m* of all its other local maxima; and
- globally multiplying the map by (**M**-*m*)^2
  - How ***different*** is a spike from the other spikes
- Promotes salient spikes that are larger than typical for that map



Conspicuity maps

- Keep within I, C and O channels to begin with as intermediary conspicuity maps
  - Saliency competition is strong within features as opposed to between features
- For each feature
  - Downsample all maps to scale 4
  - Then add all points to get total map
  - Orientation combines within orientation first, then between orientation
- Final step, normalize each of the three intermediate maps again, then combine
  - Equal weight for each feature



Across-scale combinations and normalization

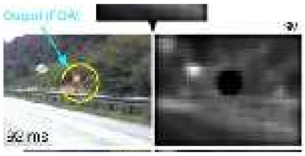
Conspicuity maps

Linear combinations

Final map and selection

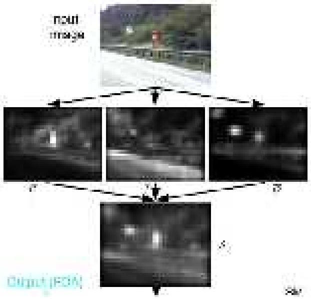
- Area of highest peak now suggests the salient feature where attention should focus
  - But they wanted to model attention selection as well
- This final map is a 2D layer of leaky *integrate-and-fire neurons* at scale four
  - Scale 4 means attention doesn't focus at 'single pixel'
  - Capacity builds until a threshold is reached, then reset
  - Dynamic neural net since input/output change over time
- 'Winner take all' network
  - Only one possible output from a noisy system with many potential options
  - Saccades, attention, forced choice RT, decision making,...
  - In this network, connections suppress all but the most active neurons

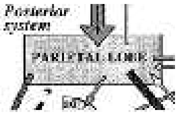
This will be important for us to access

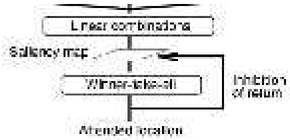


This example is easy for the model. Notice the box will cause spike in colour, orientation and intensity

- The neurons in the neural Salience Map receive excitatory inputs from combined salience scores
- Neurons are all independent.
- The potential of SM neurons at more salient locations increase faster
  - (these neurons are used as pure integrators and do not fire).
- Each SM neuron excites its corresponding Winner take All neuron.
  - This works similar to a diffusion model







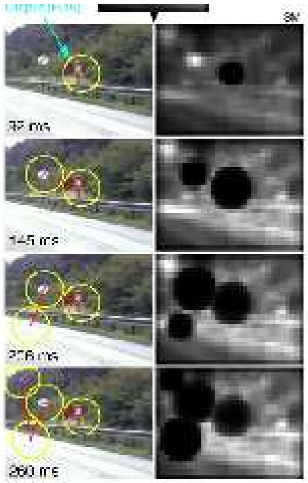
- All WTA neurons also evolve independently of each other, until one (the “winner”) first reaches threshold and fires. This triggers three simultaneous mechanisms (Fig. 3):

1) The FOA is shifted to the location of the winner neuron;

2) the global inhibition of the WTA is triggered and completely inhibits (resets) all WTA neurons;

3) local inhibition is transiently activated in the SM, in an area with the size and new location of the FOA; this not only yields dynamical shifts of the FOA, by allowing the next most salient location to subsequently become the winner, but it also prevents the FOA from immediately returning to a previously-attended location.

(Modelled after the way IOR works in search)
(Klein and MacInnes, 1999)



Final saliency map network

- Recurrent network
  - At any given time, the maximum of the saliency map reflects the focal point for attention
  - Slight bias is added for next FOA to be spatially near the current FOA (average saccadic shift in search is <4 dva)
- FOA is fixed at 1/6th of image width of height.
- Times chosen to simulate shifts of attention every 30-70 ms (though this is incorrect)
- Times chosen to inhibit locations 500-900 ms (this is correct)
- No top-down attention is modelled
  - Always state limitations of the model

tests

- Compared to Spatial Frequency Content Model (SFC)
  - Eye tracking study suggested spatial frequency content higher at fixated over non-fixated locations
  - Fast Fourier transform (FFT) basis of model
- Tested against simple images, but also many with artificial noise
  - I&K handle noise better than SFC
  - Human vision is also great at handling noise in images
- Replicated 'pop-out' parallel search from Treisman and Gelade (1980)
  - When the popout was defined by colour, orientation or luminance, the target was always the first chosen
  - Also showed linear search times for conjunction search
- Model attended textures like letters, arrows, stripes and circles
  - Even though it wasn't directly programmed to do so

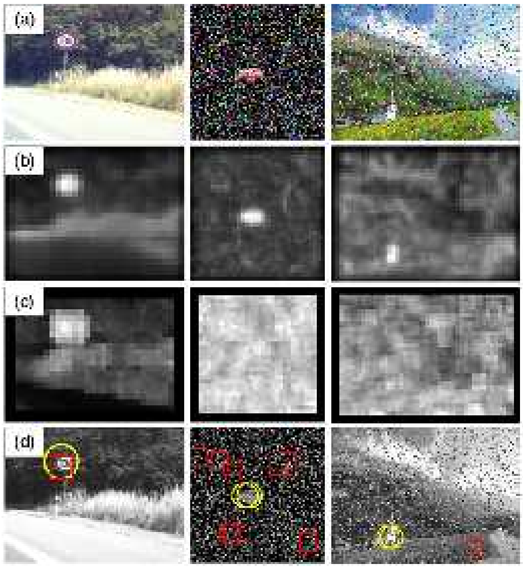
Other models

Extreme cases

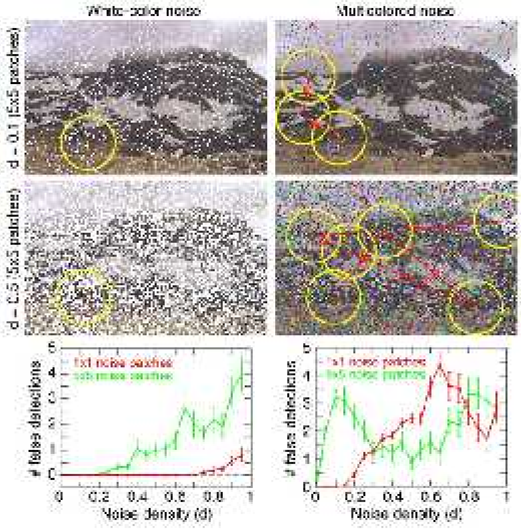
Human data

Unexpected results

- Fig. 4.
- (a) Examples of colour images.
- (b) The corresponding saliency map inputs.
- (c) Spatial frequency content (SFC) maps.
- (d) Locations at which input to the saliency map was higher than 98 percent of its maximum (yellow circles) and image patches for which the SFC was higher than 98 percent of its maximum (red squares). The saliency maps are very robust to noise, while SFC is not

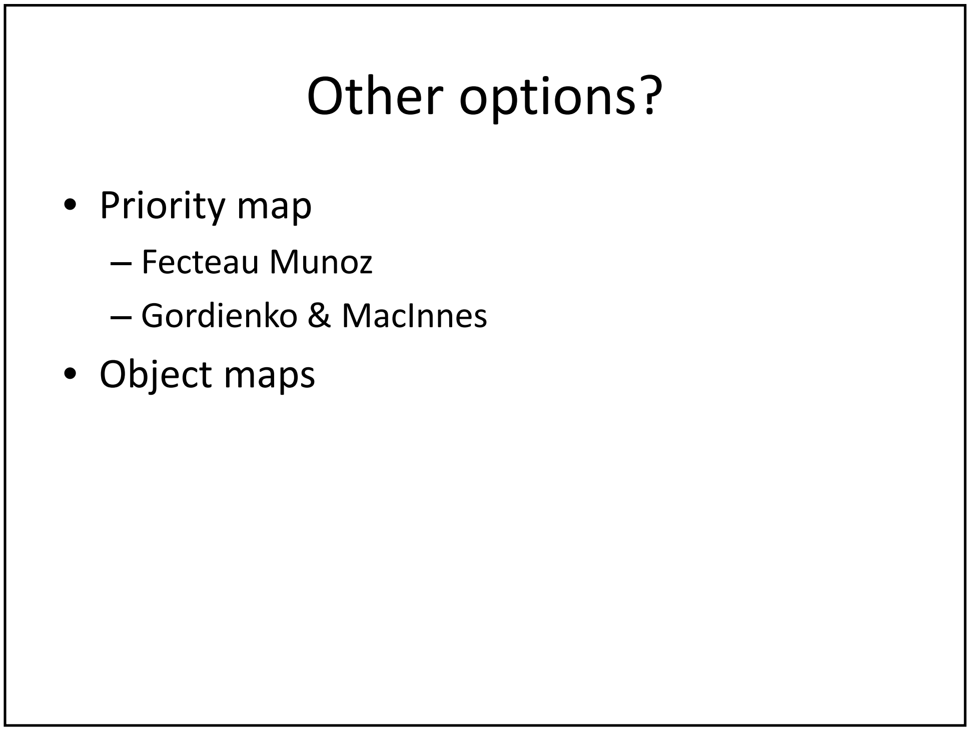


- ‘Visual search’ task with white and coloured noise
- Noise with similar features to search target interfered strongly
  - Coloured noise searching for object defined by colour
  - This is also a good testable prediction for a human experiment
- ‘Although this result does not necessarily indicate similarity between human eye fixations and the model’s attentional trajectories, it indicates that the model, like humans, is attracted to “informative” image locations, according to the common assumption that regions with richer spectral content are more informative.’



Summary from article

- Model similar to aspects of primate visual cortex neuropsysiology
- Massively parallel and feedforward
  - Faster than previous iterative algorithms
- Covers early feature extraction and attention selection
  - Feature integration theory
- Normalization allows for realistic conjunctions of features
- Limitations
  - Only three feature types
  - Feedforward does not include feedback mechanisms
  - No top down attention
  - Fails for non-implemented feature types (corners)



Other options?

- Priority map
  - Fecteau Munoz
  - Gordienko & MacInnes
- Object maps