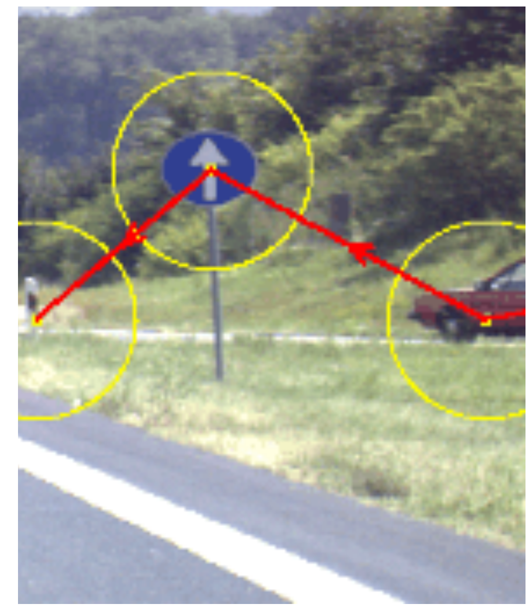
A Model of Saliency-Based Visual Attention for Rapid Scene Analysis

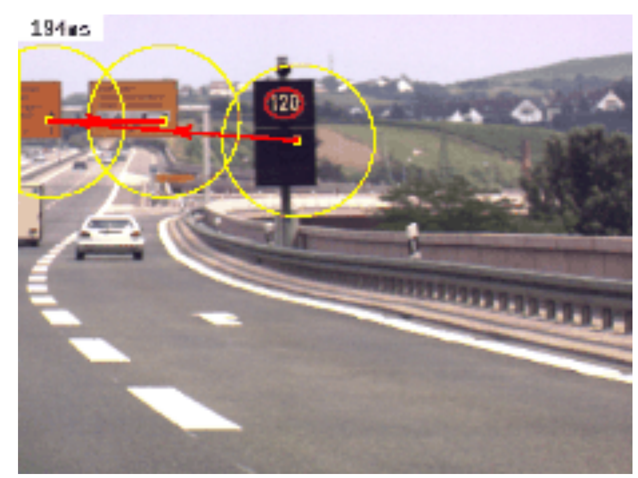
Laurent Itti, Christof Koch, Ernst Niebur

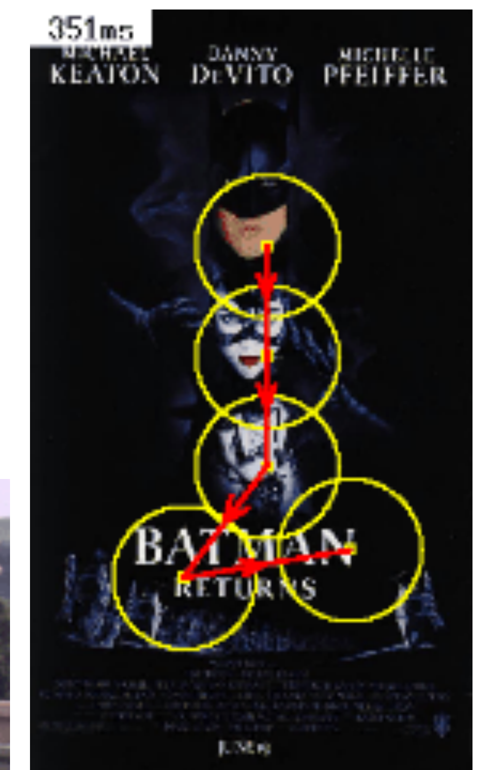
1998

Overview

- Biologically-plausible architecture for visual attention
- Computational model of primate visual attention for static scenes
- Algorithm builds and updates a saliency map to guide focus of attention over time
- Bottom-up approach uses only local information to determine saliency



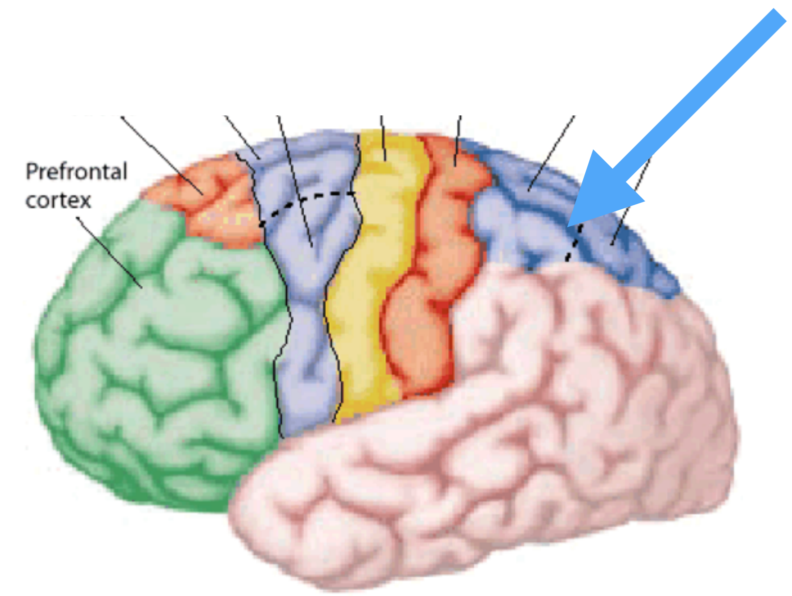


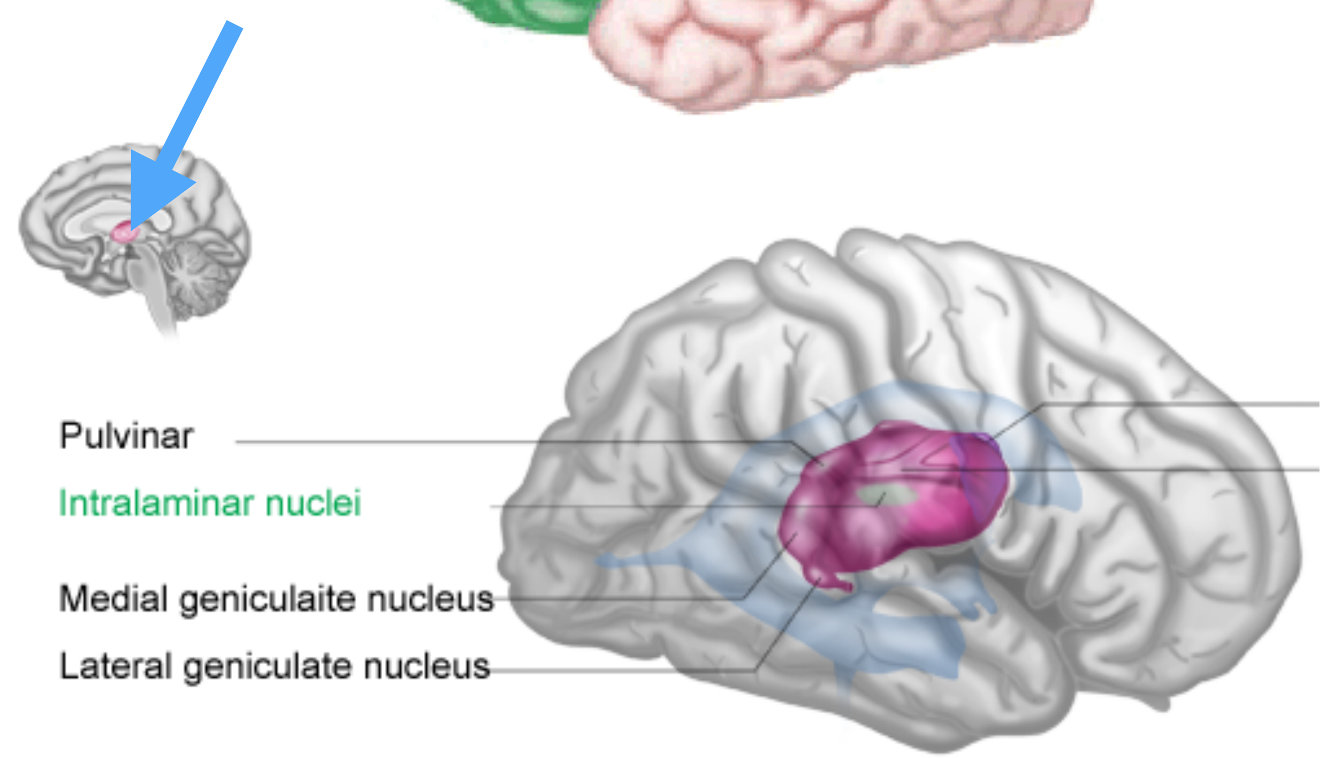


Overview

- In primate species, structures similar to a saliency map are thought to exist in certain regions of the brain
  - Posterior parietal cortex
  - Pulvinar nuclei of the thalamus

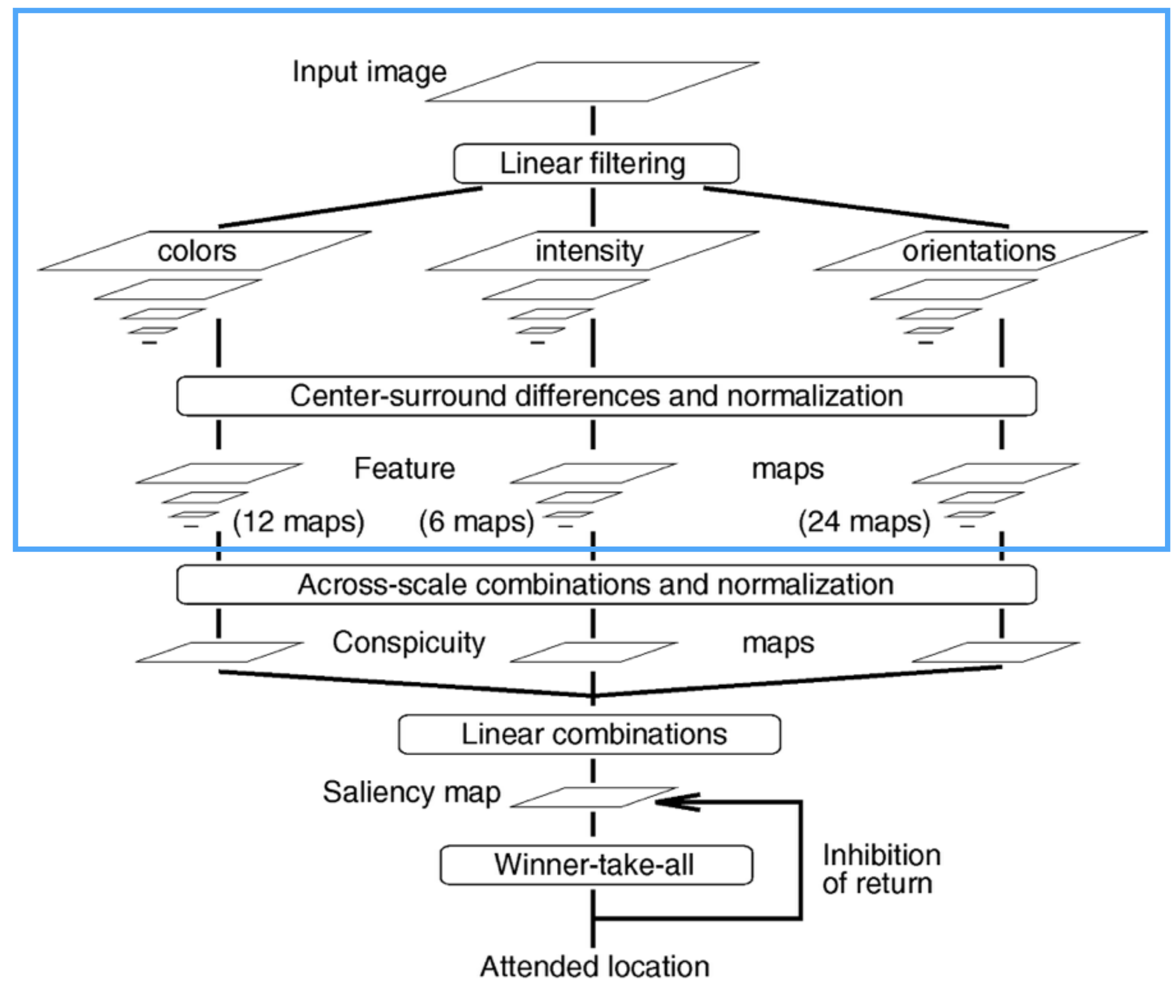
Posterior parietal cortex





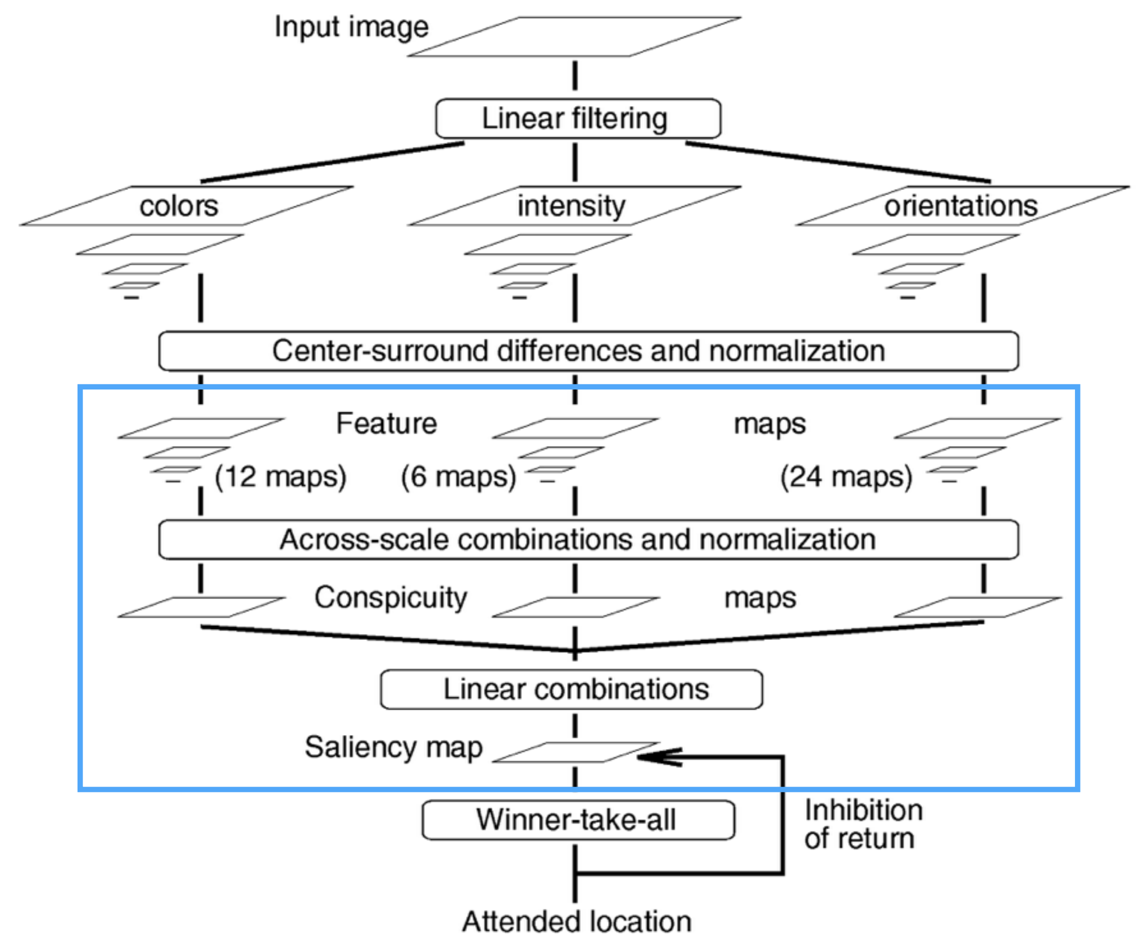
Architecture

- Multi-scale processing using Gaussian pyramids
- Features computed using center-surround difference operations (similar to SIFT)
- Separate feature maps are computed for intensity, orientation, and color at each scale/channel/orientation



Architecture

- Feature maps are combined into three conspicuity maps
- Conspicuity maps are combined by linear combination into a total saliency map

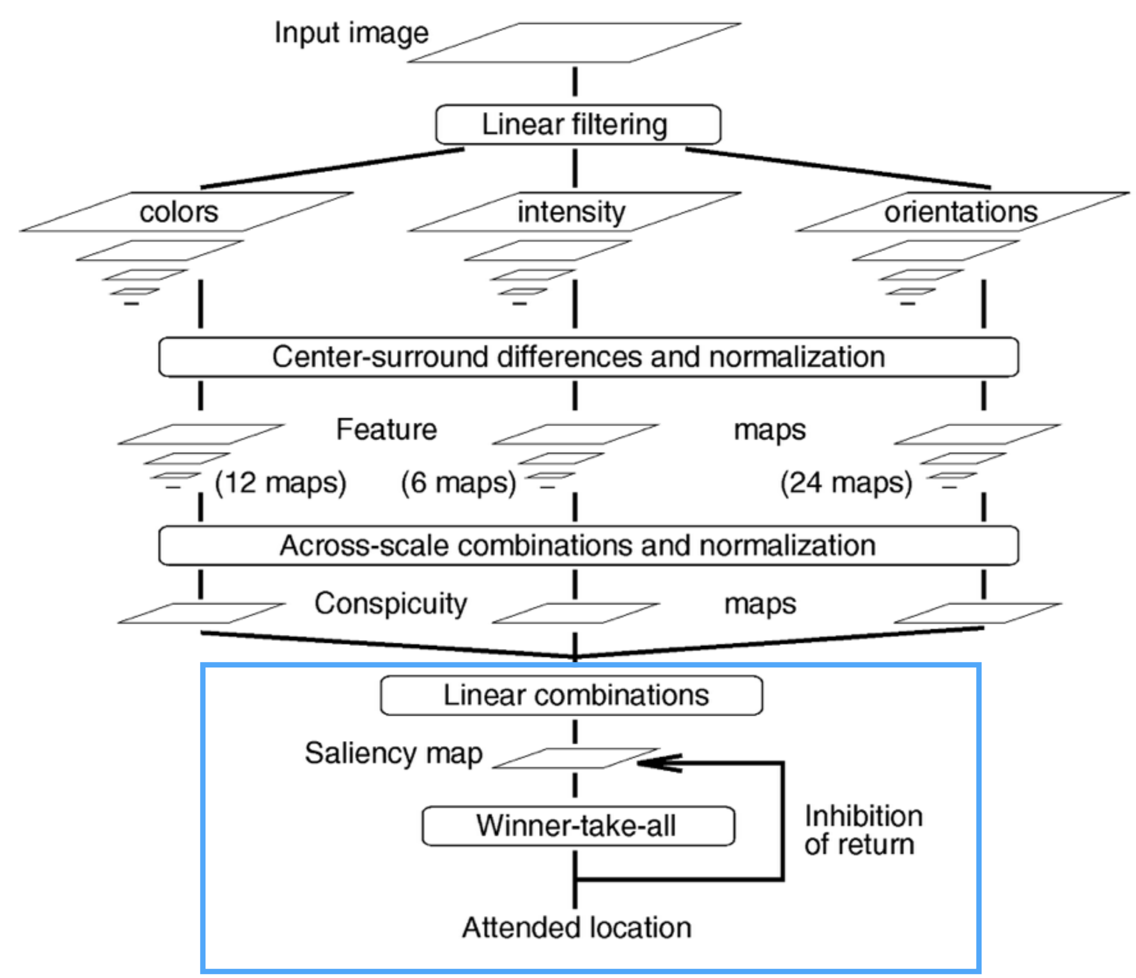


Architecture

- Saliency map is modeled as a leaky 2D “integrate-and-fire” neural network

- A winner is selected by max value and surrounding pixels are suppressed; repeat

- Similar to nonmax suppression



Center-Surround Features

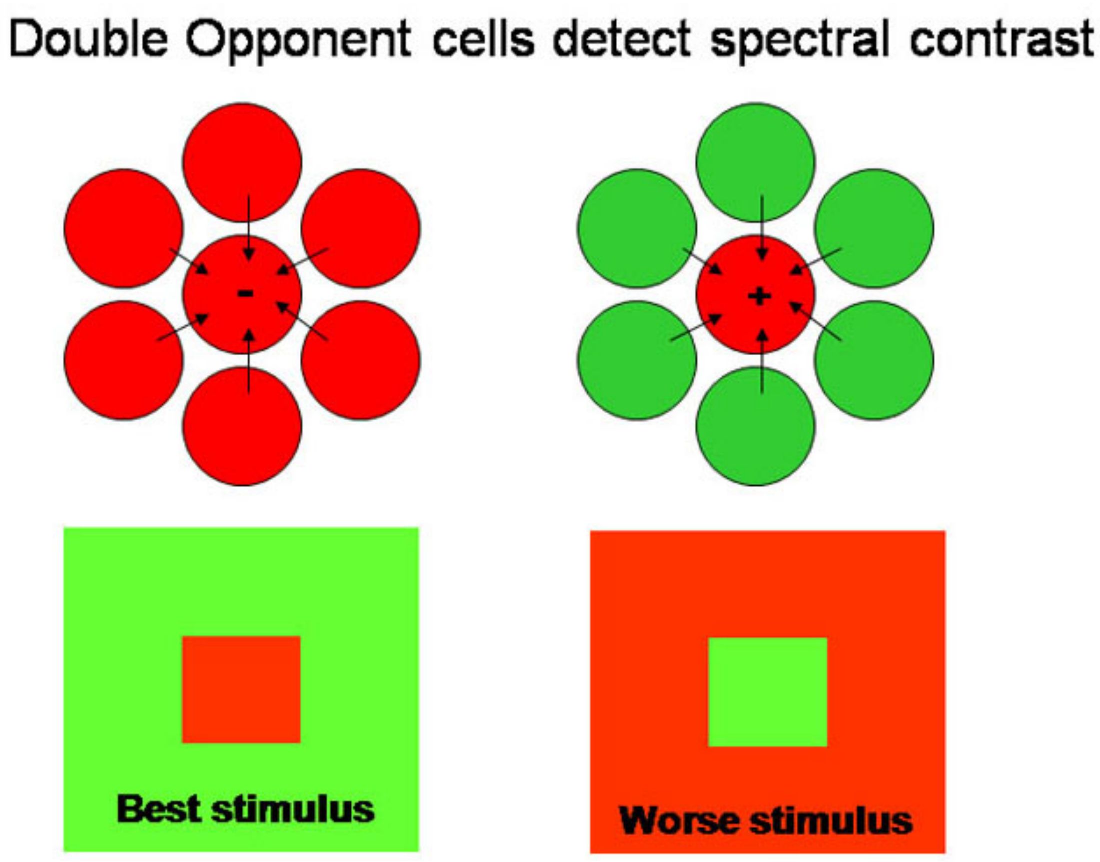
- Comparable to Difference of Gaussian features used in SIFT
- The center is the pixel at scale c and the surround is the corresponding pixel at scale $c + \delta$
- Center-surround difference $c \ominus s$ is computed by interpolating the image at scale s to the finer scale c and subtracting the images
- Using multiple values for c and δ results in multi scale feature extraction
- Unlike SIFT, no scales between octaves are used

Features: Intensity

- Intensity image is obtained as $I = (r + g + b) / 3$
- In mammals, intensity contrast is detected by two types of neurons
  - Sensitive to light centers/dark surrounds, dark centers/light surrounds
- Absolute value of difference covers both types of features
- 6 maps are computed (3 values for $c \times 2$ values for δ)

Features: Color

- Based on color double-opponent system
- Surround pixels inhibit response to same color, increase response to opponent color
- Four color channels: Red, green, blue, and yellow
- Color opponency for R/G, B/Y



Features: Color

$R = r - (g + b) / 2$ $\qquad$ $G = g - (r + b) / 2$

$B = b - (r + g) / 2$ $\qquad$ $Y = (r + g) / 2 - |r - g| / 2 - b$

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

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

Features: Orientation

- Local orientation is extracted using oriented Gabor pyramids

- Orientation feature maps are constructed from the difference in response between scales for the same orientation

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

- 24 feature maps (8 scales (c, s) $\times$ 4 orientations)

The Normalization Operator

- Local maxima are found in each feature map
- For each feature map, the global maximum *M* and the mean of its local maxima *m* is computed
- Each feature map is multiplied by $(M - m)^2$
- This emphasizes maps with clear feature responses and attenuates those which are mostly noise

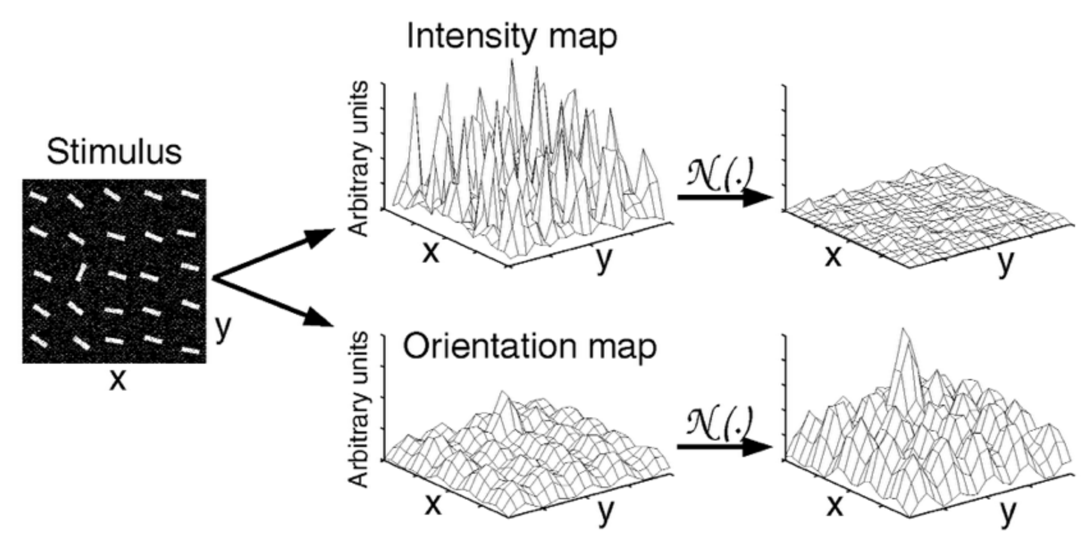


Fig. 2. The normalization operator $\mathcal{N}(.)$.

Conspicuity Maps

- Three conspicuity maps are constructed by adding the feature maps together
  - Intensity, Color, Orientation

$$\overline{I} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c=4} \mathcal{N}\left(I(c,s)\right)$$

$$\overline{C} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \left[\mathcal{N}\left(\mathcal{RG}(c,s)\right) + \mathcal{N}\left(\mathcal{BY}(c,s)\right)\right]$$

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

Saliency Map

- Finally, the three conspicuity maps are averaged into a saliency map
- The saliency map is implemented as a 2D layer of neurons
- The magnitude of the computed saliency determines the synaptic input to each neuron

Saliency Map

- A winner-take-all network is used to select the neuron with the max charge, which then “fires”
- When a neuron fires, the focus of attention is shifted to that neuron’s location
- Charge is drained from nearby neurons
- This process is repeated until time is elapsed

Results

- The authors showed that this model is superior to previous spatial frequency content (SFC) based models in the presence of noise
- Resulting FOA trajectories were not directly compared against human visual trajectories, but agreed with other models when identifying regions of high saliency

Discussion

- As in the primate visual system, the saliency map can be used as a filter between low-level and high-level systems in computer vision
- The model presented here is fairly complex (on the order of SIFT feature extraction)
- May be useful as a precursor to more advanced CV algorithms