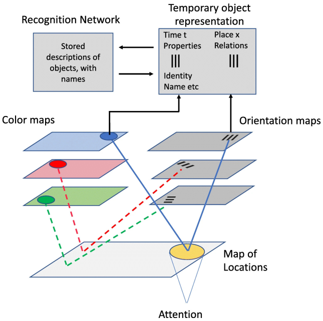
IIT Jodhpur

Biological Vision and Applications
Module 05-02: Cognitive attention models

Hiranmay Ghosh

Cognitive Models

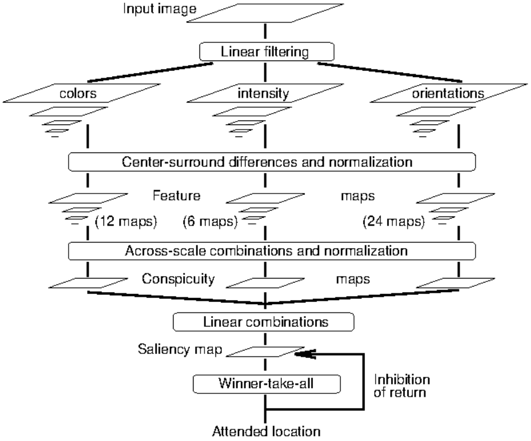
Motivated by Feature Integration Theory



- Based on the observations
  - Early vision distinguishes local contrasts
    - … colors, edges
  - Features are subsequently integrated
    - Treismann's Feature Integration Theory
  - Higher acuity at central vision (5°)
    - … lower at paracentral / macular (8 – 18°)

Itti's model (1998)

Overview

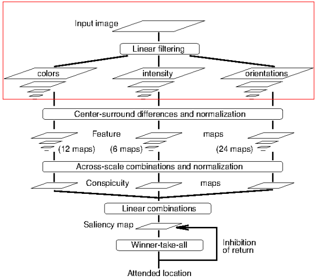


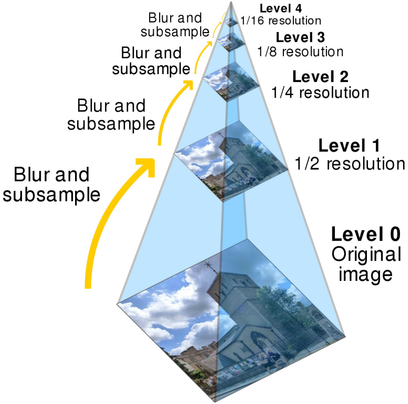
Itti, et al., A saliency based search mechanism ...

Itti’s model: Stage 1

Multi-resolution image analysis

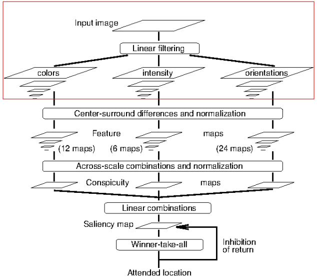
- Multi-resolution analysis of input image
  - Using Gaussian pyramids (9 scales: 0 – 8)





Itti's model: Stage 1

Feature extraction

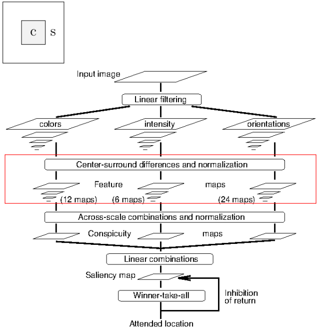


- For images at each resolution level, 3 features are extracted
  - Color (C): R-G and B-Y contrasts
  - Intensity (I): B-W contrast
  - Edge Orientations (O): 0, 45, 90, 135 degrees

- $2 + 1 + 4 = 7$ features extracted for each resolution level

Itti's model: Stage 2

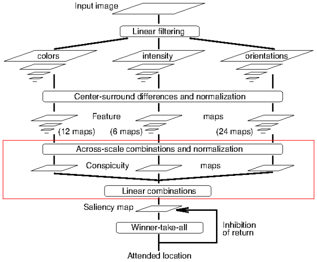
Center-surround operations: Multi-scale feature maps



- Center-surround difference computed for each of 7 features for every location
- Center at hi-res, Surround at lo-res
- Scales used:
  - Center: $c = \{2, 3, 4\}$
  - Surround: $s = c + \delta \; [\delta = \{3, 4\}]$
- Multi-scale differences
  - $\mathcal{F} = | \, F(c) \ominus F(s) \, |$
- 6 scales for each feature
- $7 \times 6 = 54$ "feature maps" (contrasts)
  - Each represents local contrast at a location based on a feature at a certain scale

Itti's model: Stage 3

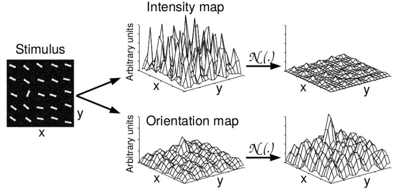
Combining the features: Conspicuity and Saliency Maps



- Feature maps are combined
- Equal weights – normalized $N()$
- Combined in two stages
  - Intra-feature-class, giving three *conspicuity maps*
    - $\bar{I} = \bigoplus_{c,s} N(I(c,s))$
    - $\bar{C} = \sum_{RG,BY} \bigoplus_{c,s} N(C(c,s))$
    - $\bar{O} = \sum_{\theta} \bigoplus_{c,s} N(O(c,s))$
  - Inter-feature-class, giving the final *saliency map*
    - $S = \bar{I} + \bar{C} + \bar{O}$

Itti's model: Stage 3

Normalization



Maxima	**6.00**	**7.00**	**5.00**	**6.00**	**5.00**
Normalized	**0.04**	**0.05**	**0.03**	**0.04**	**0.03**

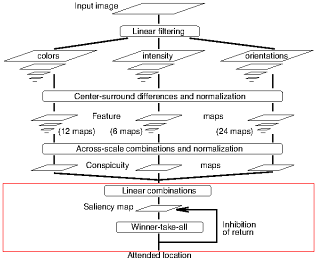
Maxima	**6.00**	**20.00**	**5.00**	**6.00**	**5.00**
Normalized	**0.16**	**0.53**	**0.13**	**0.16**	**0.13**

Local maxima:	6.00	20.00	5.00	6.00	5.00
Choose M	1.00				
Divide by max/M:	0.30	1.00	0.25	0.30	0.25
$\bar{m}$	$(0.30 + 0.25 + 0.30 + 0.25)/4 = 0.275$				
$(M - \bar{m})^2$	0.526				
Normalized values:	0.16	0.53	0.13	0.16	0.13

- Two reasons to normalize
  - Features are at arbitrary scale
  - Normalize to a fixed range $[0, M]$
- Some feature may have many nearly equal peaks, indicating texture
- Steps:
  - Choose M
  - Normalize so that the global max $= M$
  - Compute the average of all other local maxima $\bar{m}$
  - Multiply the map by $(M - \bar{m})^2$

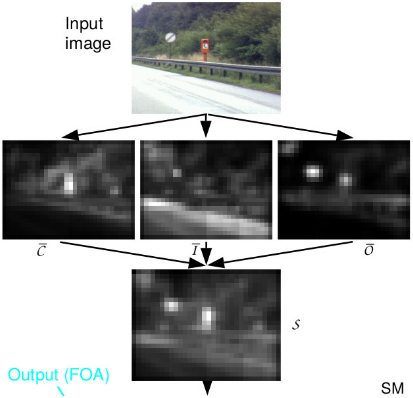
Itti's model: Stage 4

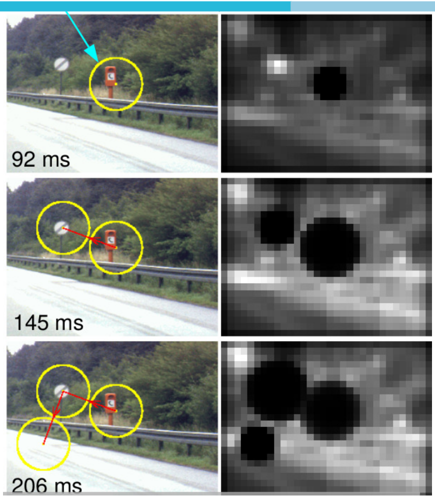
"Winner take it all" and "Return Inhibition" policies



- Winner-take-it-all policy
  - The image location with highest saliency attracts attention
  - All other locations are ignored
- Return Inhibition policy
  - Attention never returns to a location once attended
  - The neurons at the attended place tires out.
  - Attention moves to the location with next highest salience.

Sample Results



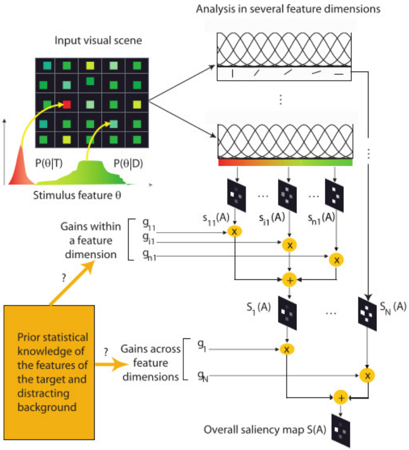


Discussions

- Remains a reference model till date
  - WTA and RI policies are common to all classical models

- Based on cognitive theories of early vision
- Features used: Color, Intensity and Orientations
  - Equal weights to all features
- Models bottom-up attention
- Provides static saliency map
- Eye movement guided by
  - Winner Take All policy
  - Return Inhibition policy

Adaptation to top-down attention


Analysis in several feature dimensions
Input visual scene
P(θ|T)
P(θ|D)
Stimulus feature θ
Gains within a feature dimension
g11
gi1
gn1
s11(A)
si1(A)
sn1(A)
?
Prior statistical knowledge of the features of the target and distracting background
?
Gains across feature dimensions
g1
gN
S1(A)
SN(A)
Overall saliency map S(A)

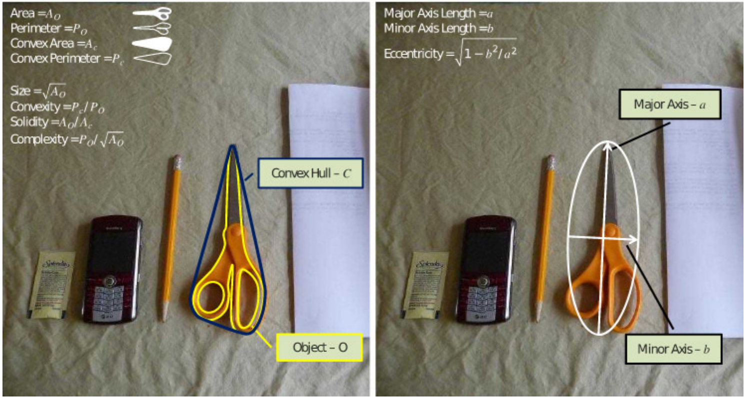
- Visual search task
- Weights assigned to features based on task requirement
- Weights learned from statistical features of target and distractors
- Inflexible

Navalpakkam & Itti. An Integrated Model of Top-Down and Bottom-Up Attention ...

Extension of feature set

Object level attributes

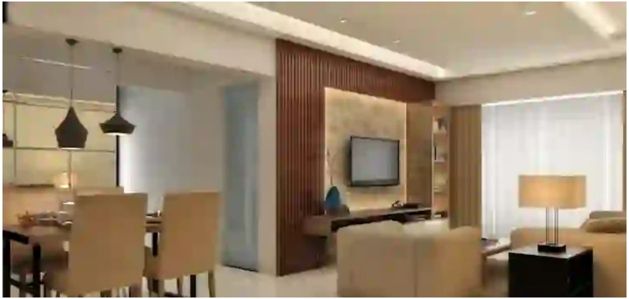
- Recall what is likely to be a foreground object
  - Local motion (for video)
  - convex-ness ...



Extension to feature set (contd.)

What draws human attention? – Rethinking the principles

WELCOME
TO OUR
CAFE!
dreamstime.com



Semantic features

- Semantic features
  - Human face and emotions
  - Text
  - Man-made objects designed to be watched (TV, clock, ...)
  - Objects with sound, smell, taste, touch attributes
  - Objects interacted with (touched or gazed upon by) humans (a computer mouse, ...)
  - ...

Early fusion vs. late fusion

When to fuse the conspicuity maps?

- Early fusion
  - As in Itti's model
  - Fused immediately after normalization
  - Overall saliency map created after fusion
- Late fusion
  - Create saliency map based on one feature
  - Fuse conspicuity maps from the other features for the competing locations
    - One at a time
  - Computationally more efficient
  - Sequence?
    - Color first. No consensus of other features

Khan, et al. Top down color attention ...

Quiz

Quiz 05-02

End of Module 05-02