ELL 788
Computational Perception & Cognition

July – November 2015

Module 7

Visual Attention

Attention

Dictionary.com:
A concentration of the mind on a single object or thought, especially one preferentially selected from a complex, with a view to limiting or clarifying receptivity by narrowing the range of stimuli.

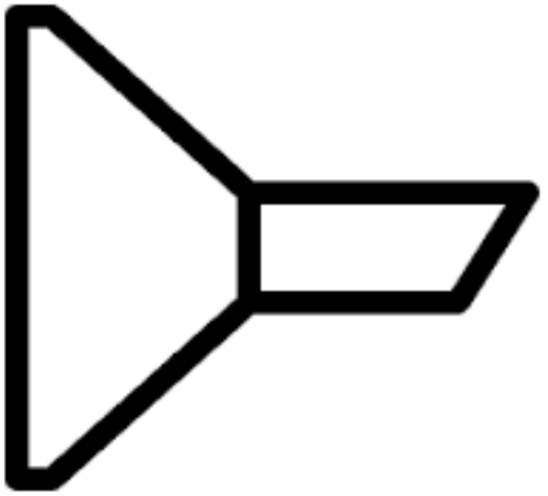
Wikipedia:
Attention is the behavioral and cognitive process of selectively concentrating on a discrete aspect of information, whether deemed subjective or objective, while ignoring other perceivable information.

Attention has also been referred to as the allocation of limited processing resources (to selected stimulii).

Visual attention

- Large volume of visual information on retina
- Processed for scene interpretation and object recognition
  - High-level cognitive and complex processes
- Mechanism to reduce the amount of visual data
  - Selection mechanism and a notion of relevance
- Achieved through retina
  - High-resolution central fovea (1-2 degrees) and a low-resolution periphery
  - Computational mechanisms underlying guidance

Stimulii

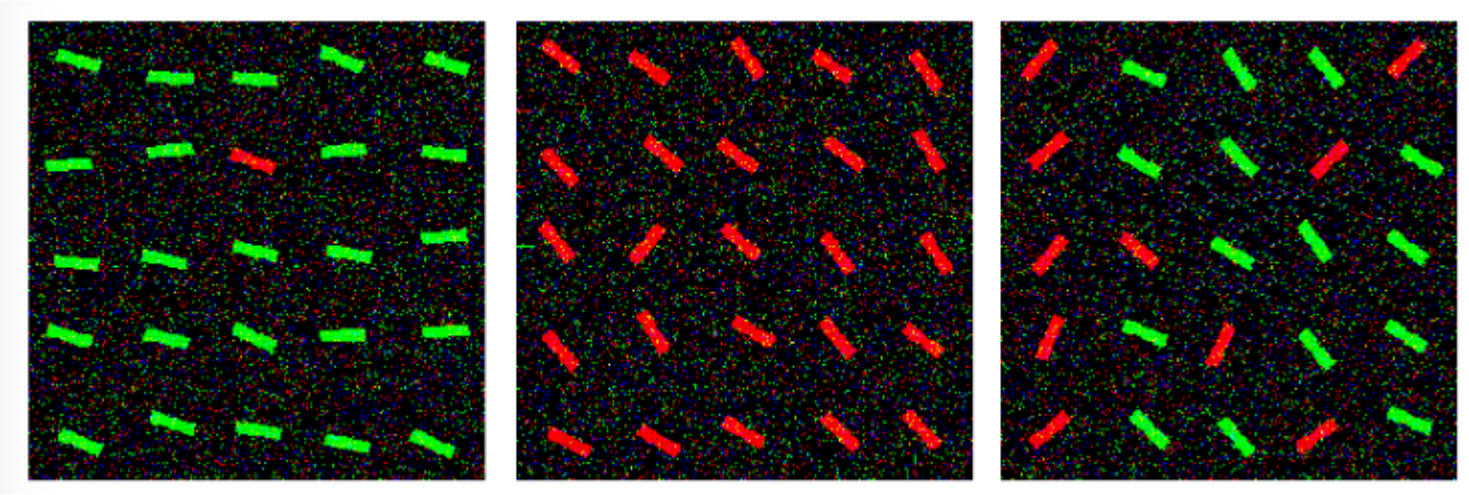


Attention

Awareness

Feature integration theory of visual attention

Treisman 1980



*"... features are registered early, automatically, and in **parallel** across the visual field, while objects are identified separately and only at a later stage, which requires focused attention ..."*

*"... visual scene is initially coded along a number of separable dimensions, such as color, orientation, spatial frequency, brightness, direction of movement. In order to recombine these separate representations and to ensure the correct synthesis of features for each object in a complex display, stimulus locations are processed **serially** with focal attention."*

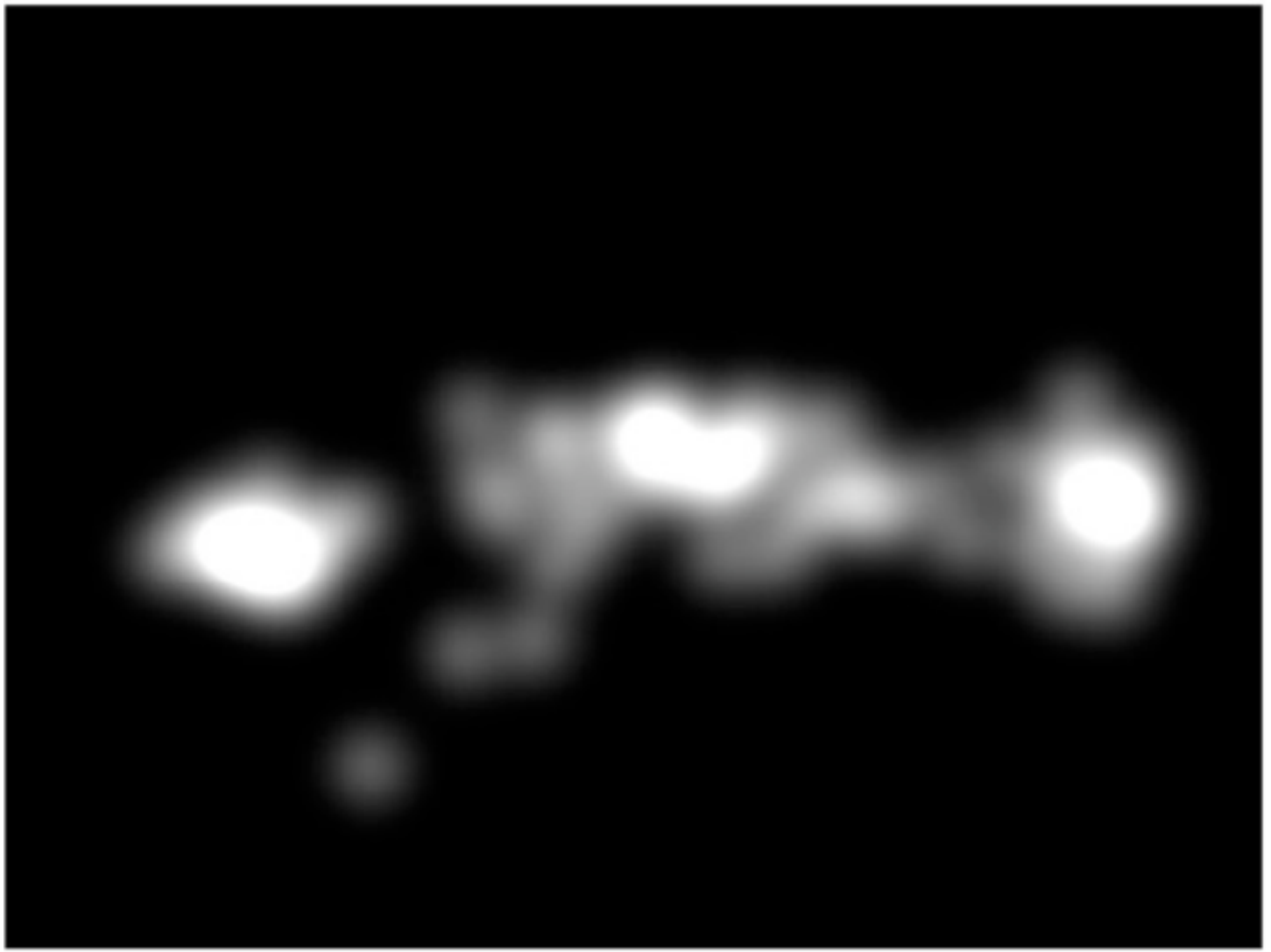

cu motor cu injecție
MARGHITA
360113
AMBULANTA
TRANSPORT
BIHOR-ORADEA
SERVICIUL JUDETEAN DE AMBULANTA


cu motor cu injectie
SERVICIUL JUDETEAN DE AMBULANTA

Psychological experiments with attention models

fixations for one user

first 5 fixations for 15 users

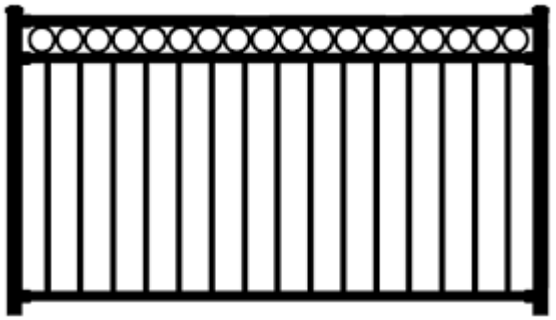


Average fixation locations / continuous saliency map

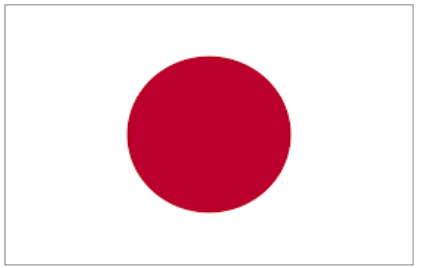
Image features for visual attention

- Intensity
- Color
- Local Orientations (edges)

- Local variations
  - Spatial
  - Temporal



Red ------> Green Blue ------> Yellow

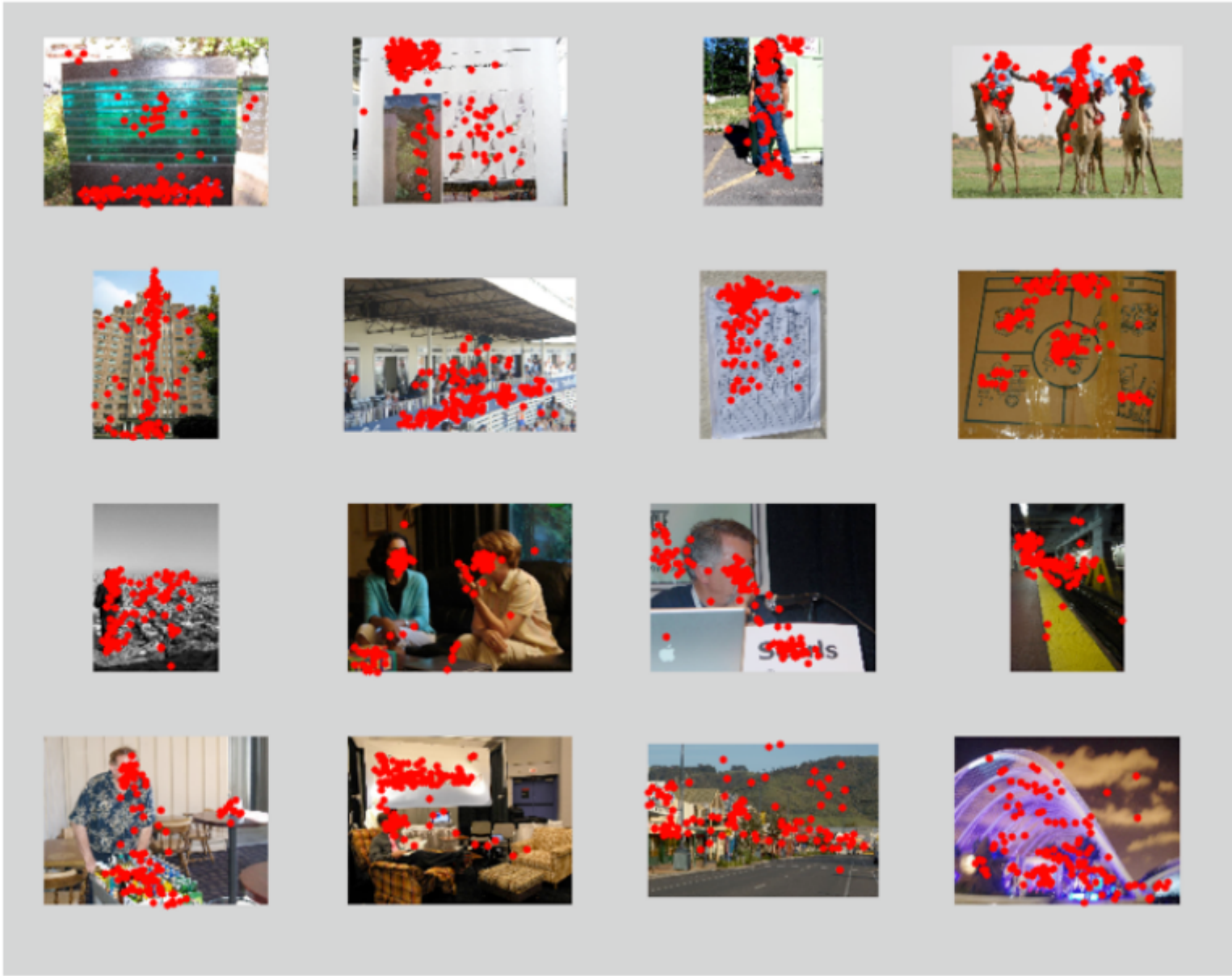


High level features

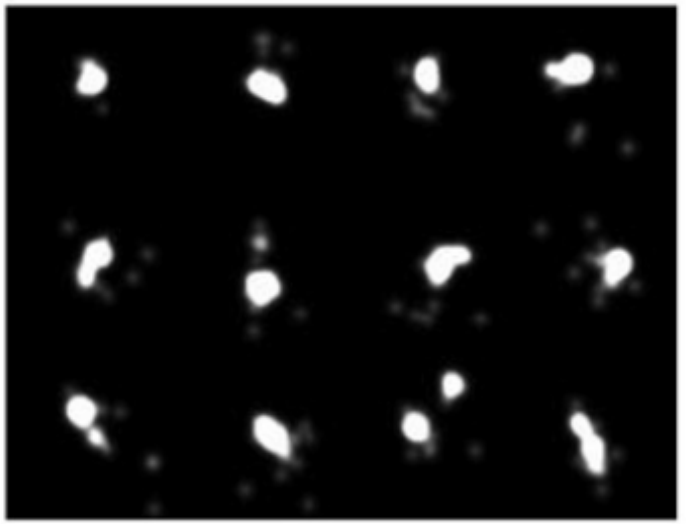
- Human Face, Skin, Hair
- Human, Animal forms
- Cars, other vehicles
- Text

Whatever moves ?

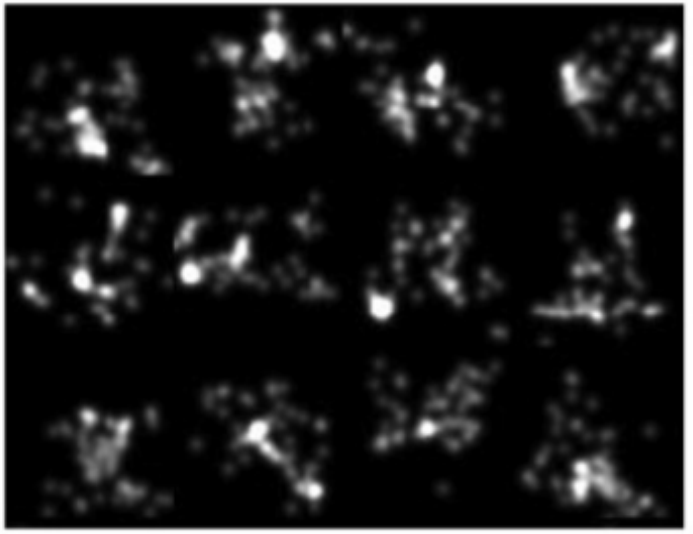
Generally, where people look at



How consistent are humans ?

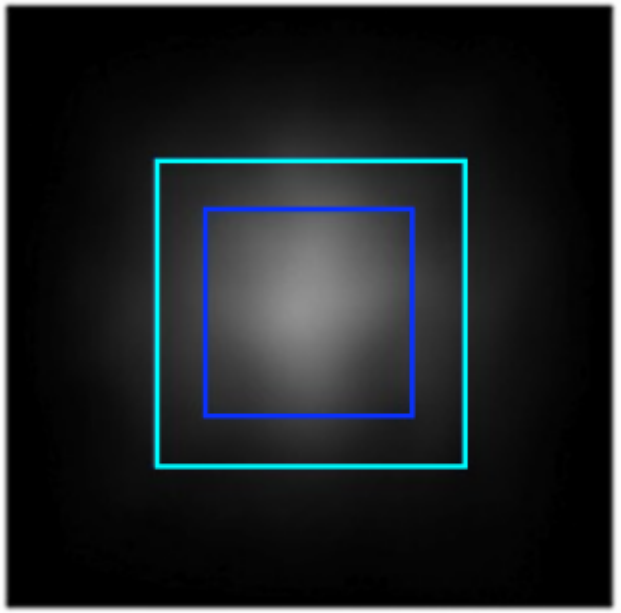


Low entropy saliency maps



High entropy saliency maps

People often look at the centre of the image



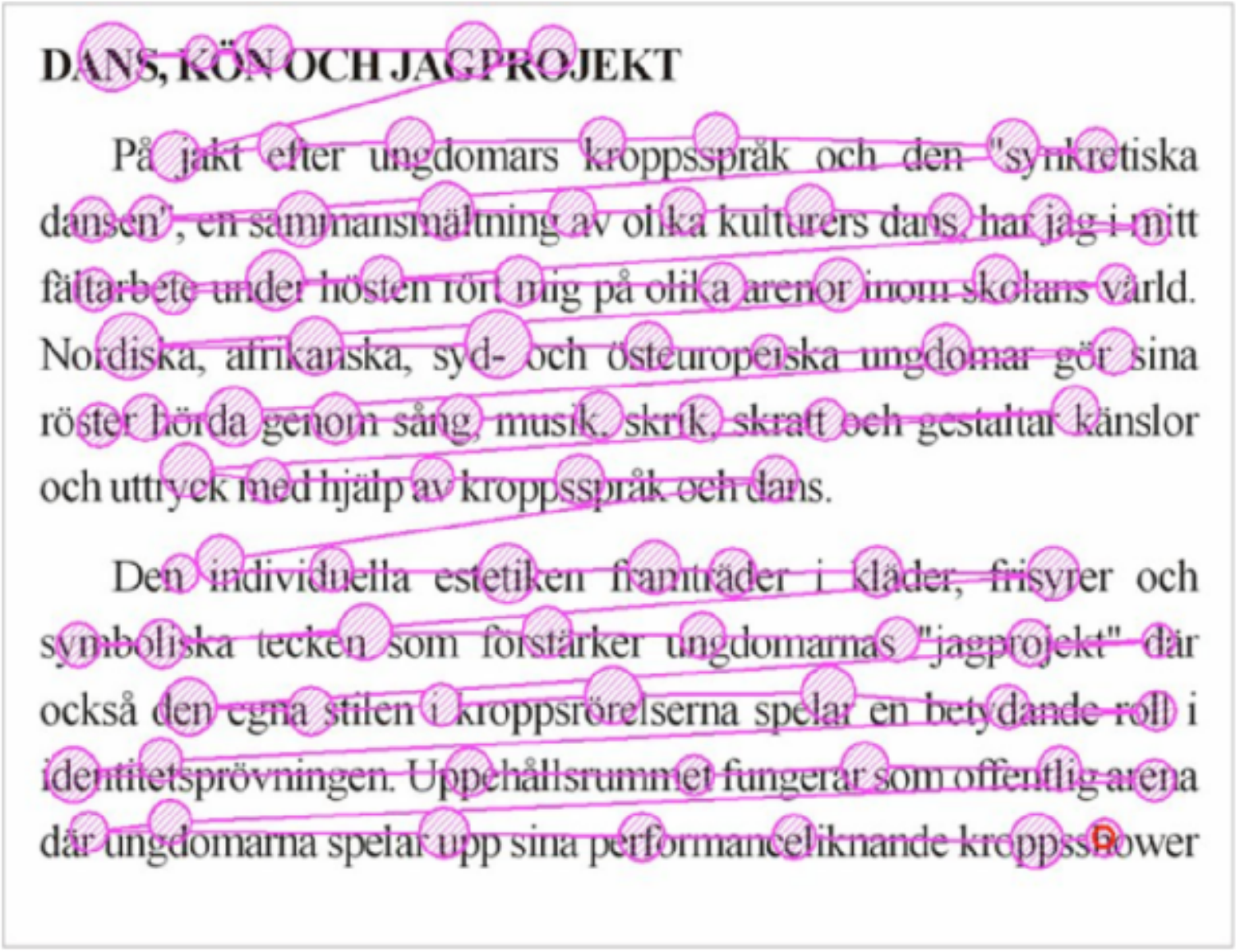
Avg of all fixation maps

40% of fixations within the center 11% of image
70% of fixations within the center 25% of image

Why?

- **Photographic bias**
  people put objects of interest near the center of the frame
- **Viewing bias**
  people EXPECT to see objects of interest near the center
- **Experiment setup**
  viewers sit directly in front of the screen
- **Viewing strategy**
  the center is a good place to START exploring from to best decide where to go next
- **Orbital reserve**
  eyeballs are lazy - they prefer to look straight ahead

Eye movement while reading text



DANS, KÖN OCH JAGPROJEKT

På jakt efter ungdomars kroppsspråk och den "synkretiska dansen", en sammansmältning av olika kulturers dans, har jag i mitt fältarbete under hösten fört mig på olika arenor inom skolans värld. Nordiska, afrikanska, syd- och österuropeiska ungdomar gör sina röster hörda genom sång, musik, skrik, skratt och gestaltar känslor och uttryck med hjälp av kroppsspråk och dans.

Den individuella estetiken framträder i kläder, frisyrer och symboliska tecken som förstärker ungdomarnas "jagprojekt" där också den egna stilen i kroppsrörelserna spelar en betydande roll i identitetsprövningen. Uppehållsrummet fungerar som offentlig arena där ungdomarna spelar upp sina performanceliknande kroppsshower

Why study Visual attention model

- Study human vision system
- Advertizing
- Graphics
- Web usability
- Image and video compression / cropping
- ...

Examples

Doug DeCarlo and Anthony Santella [SIGGRAPH 2002]

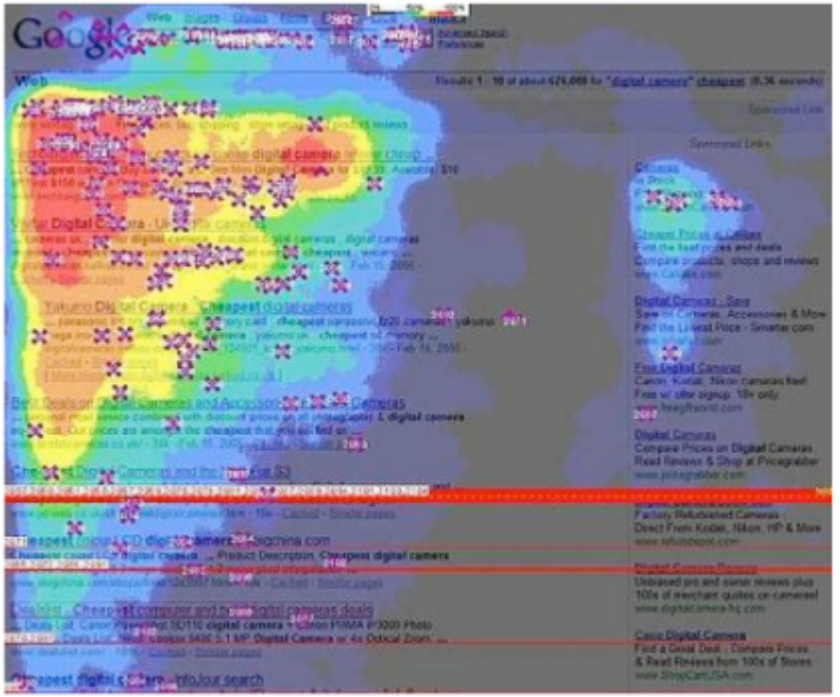
Level of details

(a) original (a) original (b) gaze-based

Santella et al. [SIGCHI 2006]

Autocropping

Examples



Web-page design

Advertising

Visual attention modeling

- Attention is characterized by eye movements
  - **Fixations**
  - **Saccades** (quick, simultaneous movement of both eyes between two phases of fixation in the same direction)
- Attention models try to predict eye movements in a cetain context
  - *Widely different eye movements for the same scene in different task contexts*

Assume K subjects have viewed a set of N images $\boldsymbol{I} = \{ I_i \}_{i=1}^{N}$

Let eye fixations of k-th subject on i-th image be denoted by $L_i^k = \{ p_{ij}^k, t_{ij}^k \}_{j=1}^{n_{ik}}$

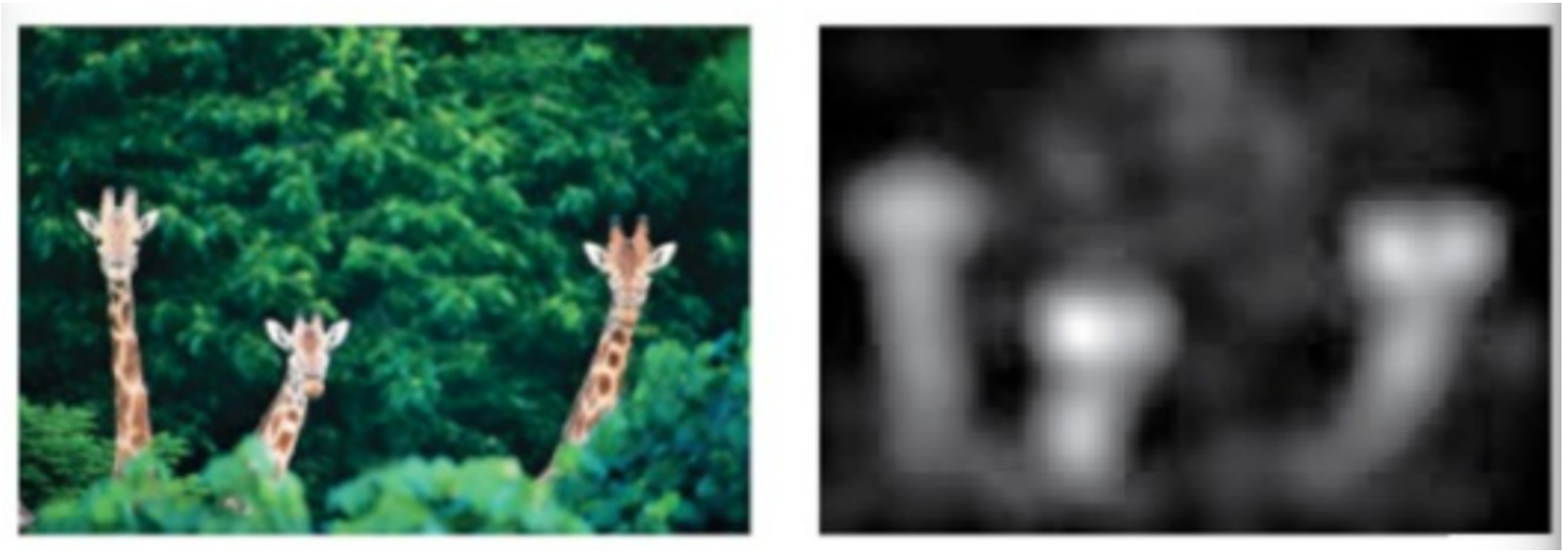
Find a function f that minimizes error on eye-fixation prediction

$$\sum_{k=1..K} \sum_{i=1..N} m\left(f\left(I_i^k\right), L_i^k\right)$$

m is some distance function,
e.g. edit distance

Saliency driven attention

- Based on raw sensory input
  - *Bio-inspired model: Filtering at early stages of human vision system*
- Rapid and Involuntarily (reflexive)
- Distinct regions of interest
    - Color, Orientation, Intensity (Spatial and temporal contrast)



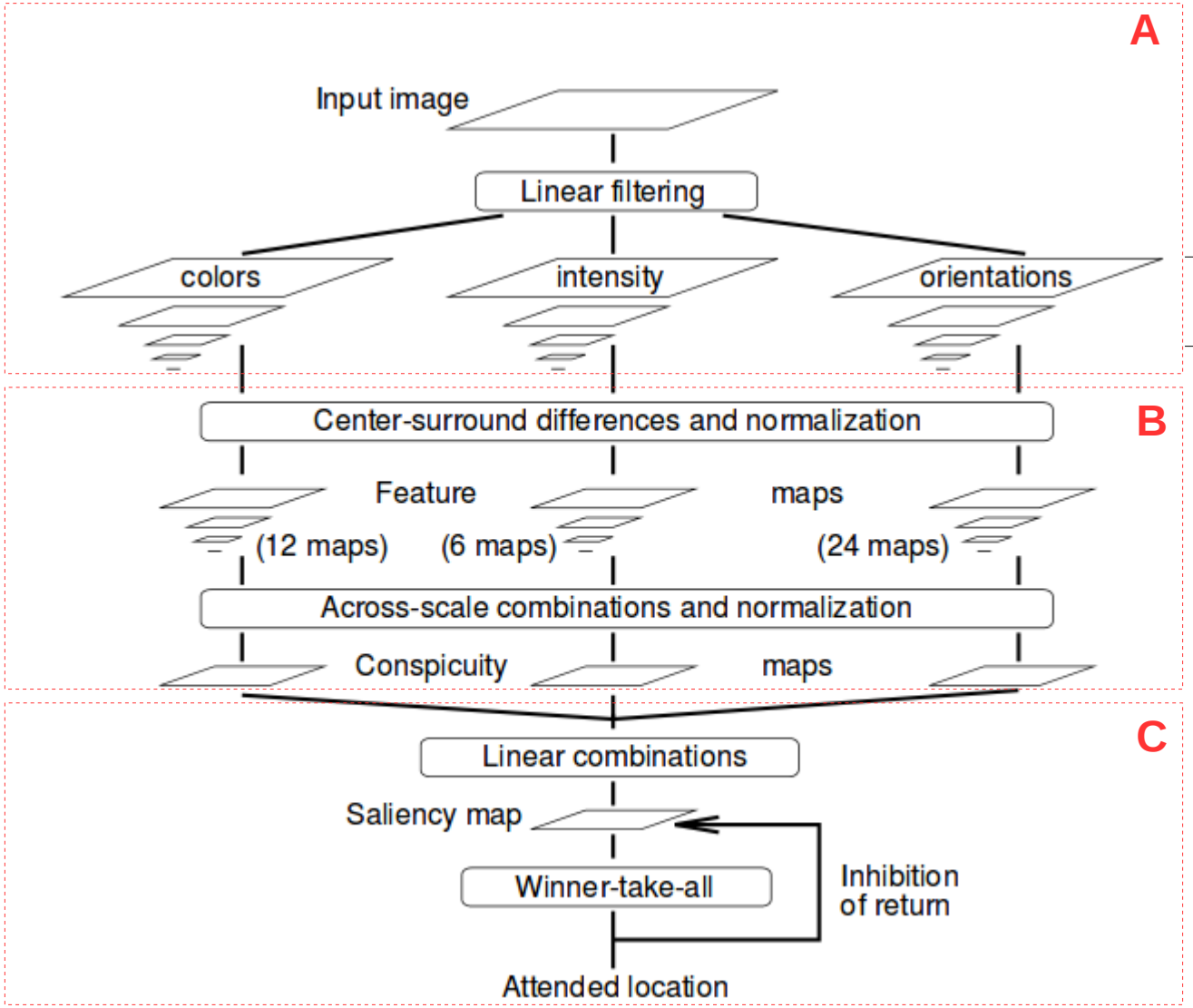
Saliency

- Measure of conspicuity
- Likelihood of a location to draw attention of a human

Property of the image alone ?

A perceptual model for saliency

Itti, et al. 1998

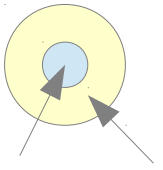


Level 0: 1:1 (640 x 480)

Level 8: 1:256
(Gaussian filter of width 5)

Multi-scale feature comparison

Intensity and color maps are produced at 8 levels of resolution
Local orientation is derived from the intensity maps



At higher resolution *At lower resolution*

Centre: $c \in \{2,3,4\}$
Surround: $s = c + \delta, \quad \delta \in \{3,4\}$

6 maps at different scales

$$M_i(c,s) = |O_i(c) \ominus O_i(s)|$$

Interpolate features at s over c at finesr scale and take pixel-by-pixel difference

Feature maps

Colors

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

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

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

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

Intensity

$$I = (r+g+b)/3$$

$$\mathcal{I}(c,s) = |I(c) \ominus I(s)|$$

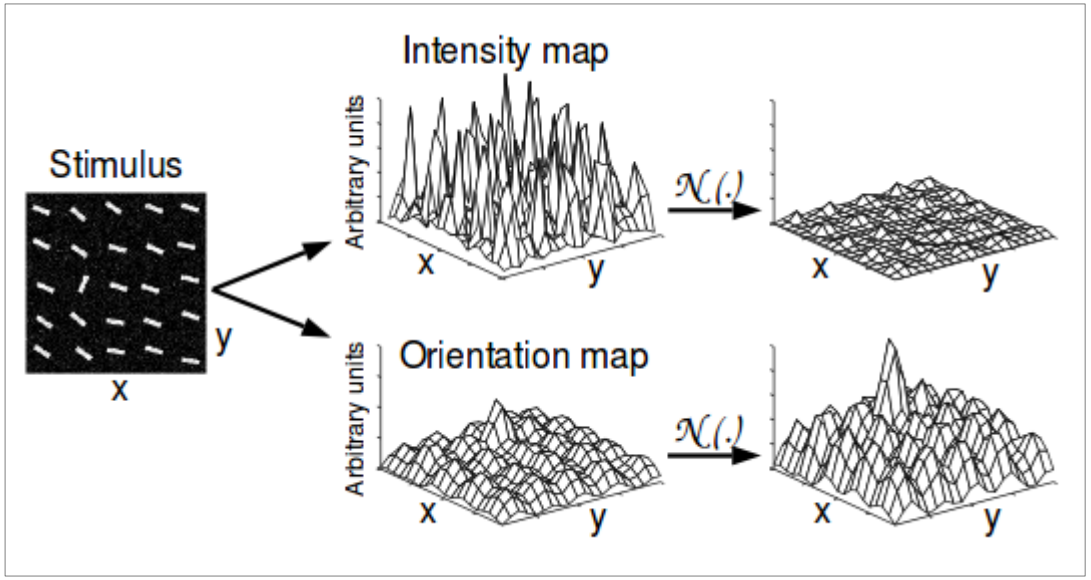
Local Orientation

Gabor filters (wavelets) $O(\theta)$: $\theta = \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$

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

7 x 6 = **42** feature maps are created

Normalization $N(.)$



Features map are not comparable:

- Different dynamic ranges

As there are 42 features, saliency observed in one feature may be masked by others

Normalization

- Homogenize dynamic range $[0,M]$
- Multiply with $(M - \boldsymbol{m})^2$

M: Global maximum, $\boldsymbol{m}$: Average of local maxima

If some of the local maxima are stronger than others, they are more salient

Done for all 42 feature maps separately

Conspicuity maps

Feature maps are combined to three separate conspicuity maps for color, intensity and local orientations.

Involves across-scale addition

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

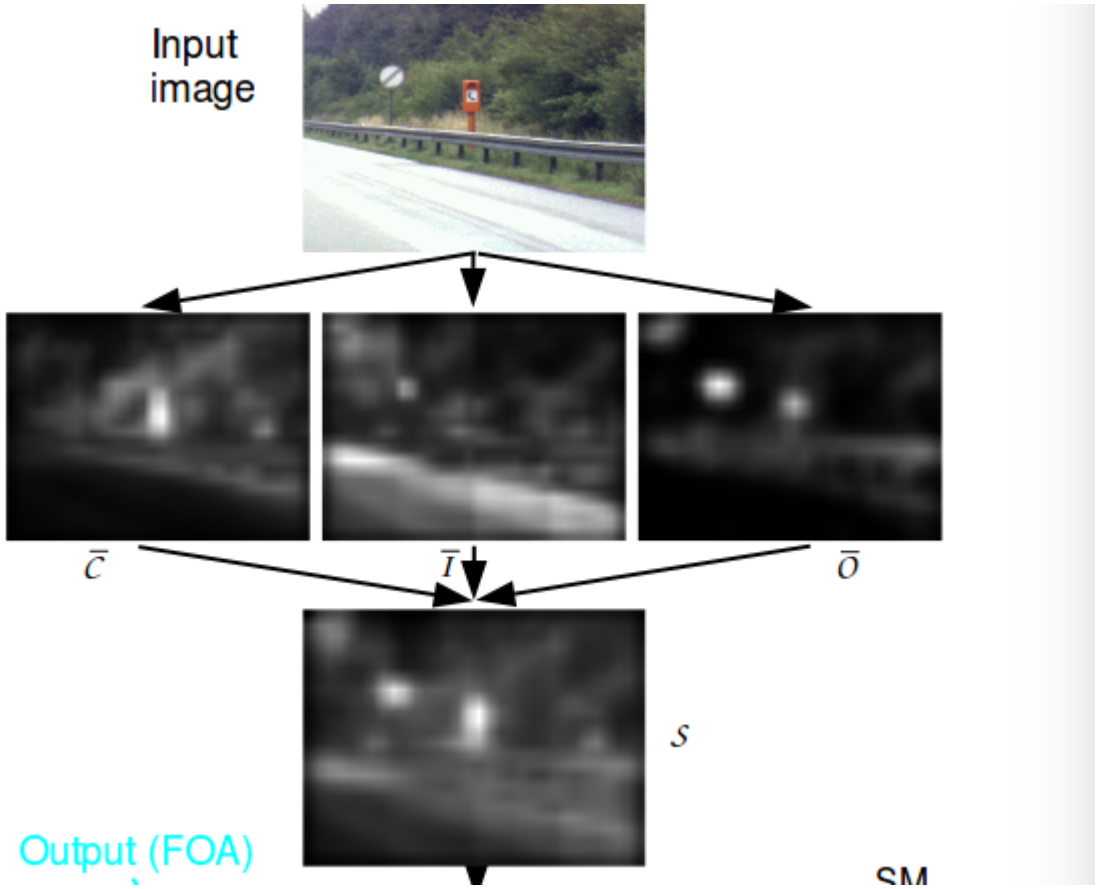
$$\overline{\mathcal{I}} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{I}(c,s))$$

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

Saliency map

Saliency: Linear combination of 3 features (normalized)

$$\mathcal{S} = \frac{1}{3}\left(\mathcal{N}(\overline{\mathcal{I}}) + \mathcal{N}(\overline{\mathcal{C}}) + \mathcal{N}(\overline{\mathcal{O}})\right)$$



Focus of attention (FOA)

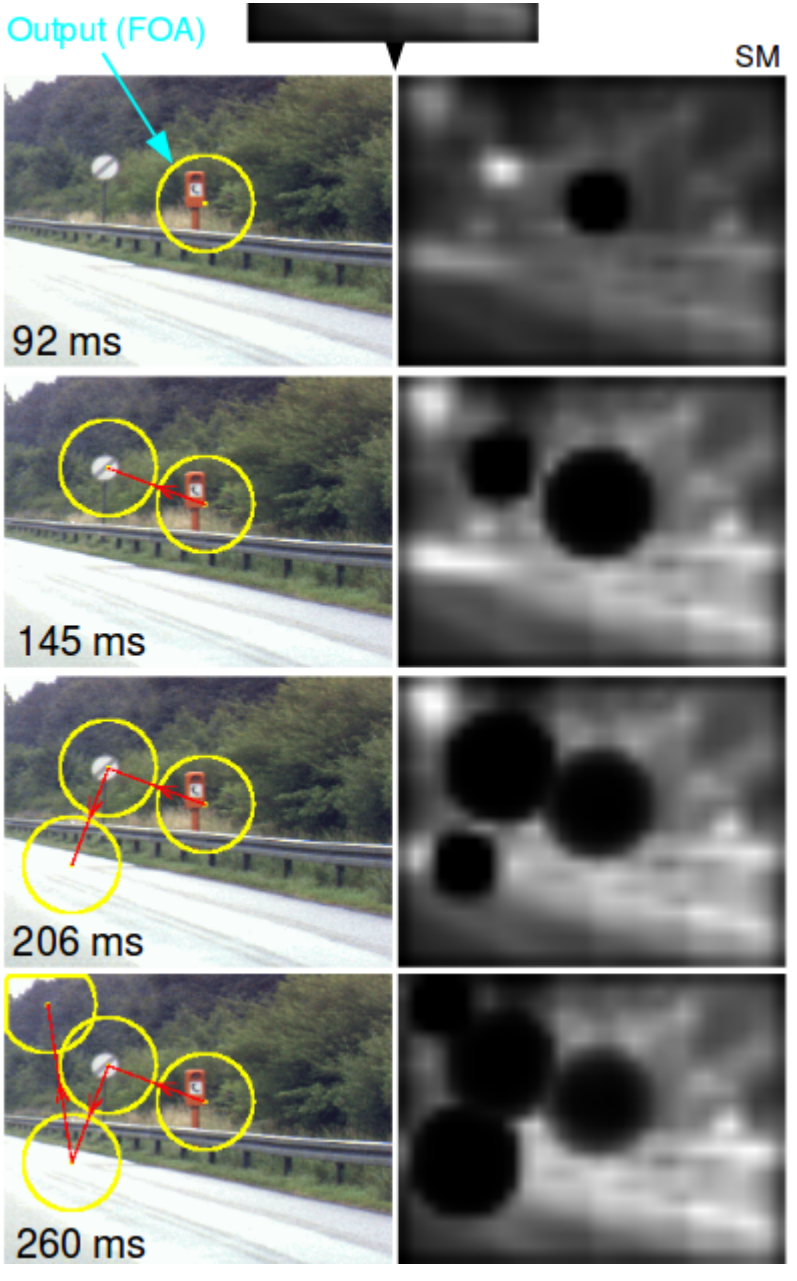
At any given point of time, the maxima in saliency map determines Focus of attention.

Winner-take-it-all strategy:
Neuron attaining highest potential fiers
Determines FOA

Leaks and potential falls to zero

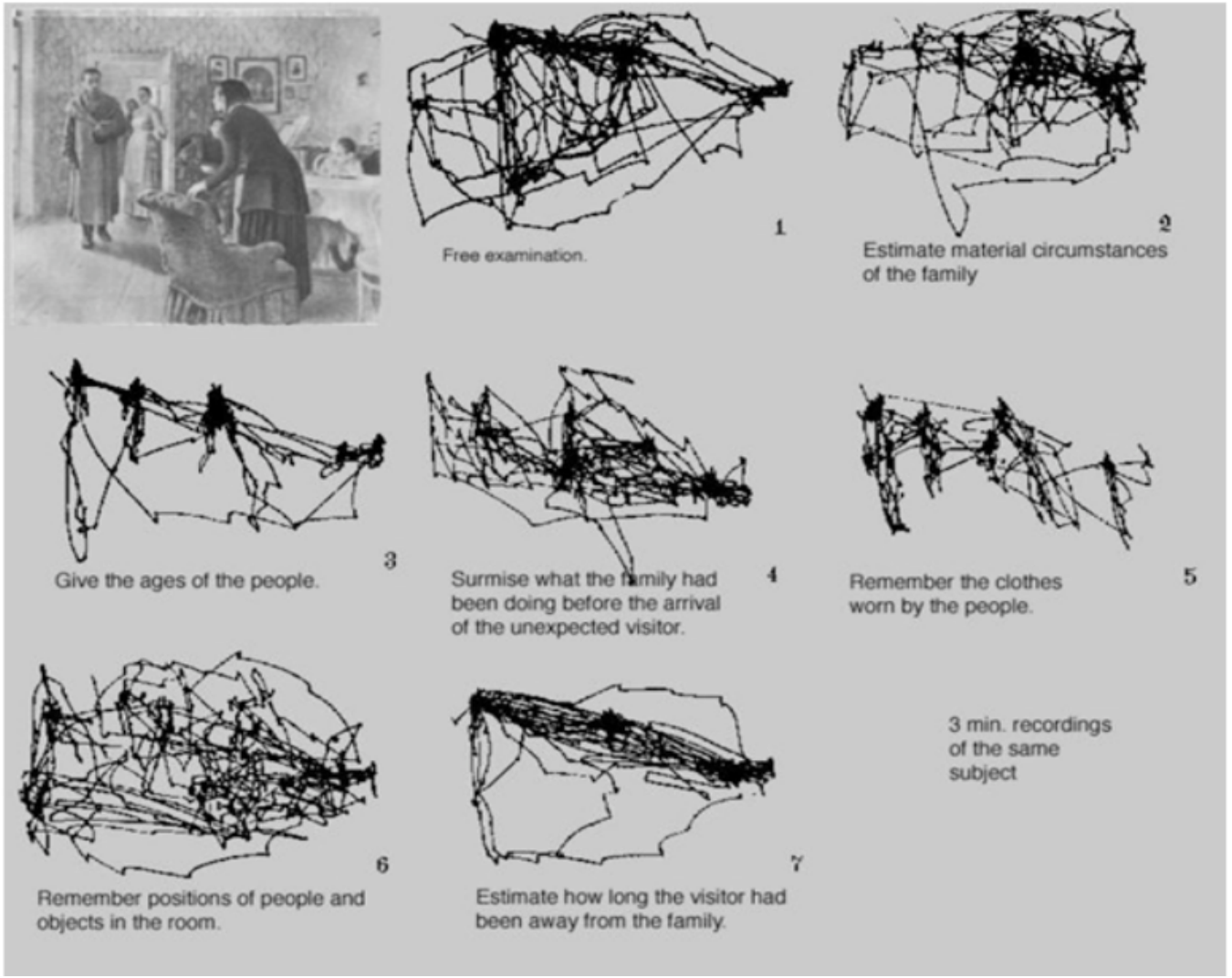
The neuron attaining highest potential now fires
Determines next FOA

...
...



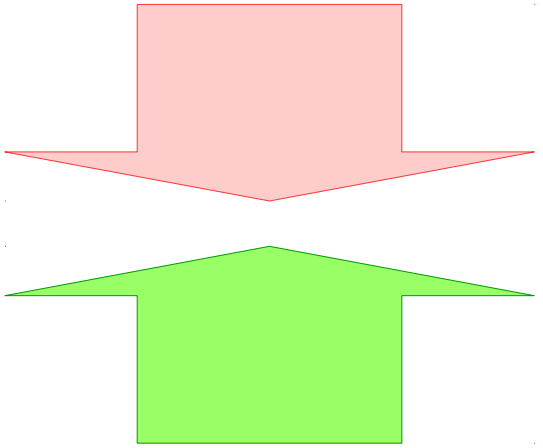
Goal-driven attention

Eye-movements (for the same scene) widely vary depending on the task



Top-down and bottom-up attention

Top-down attention (task-driven)

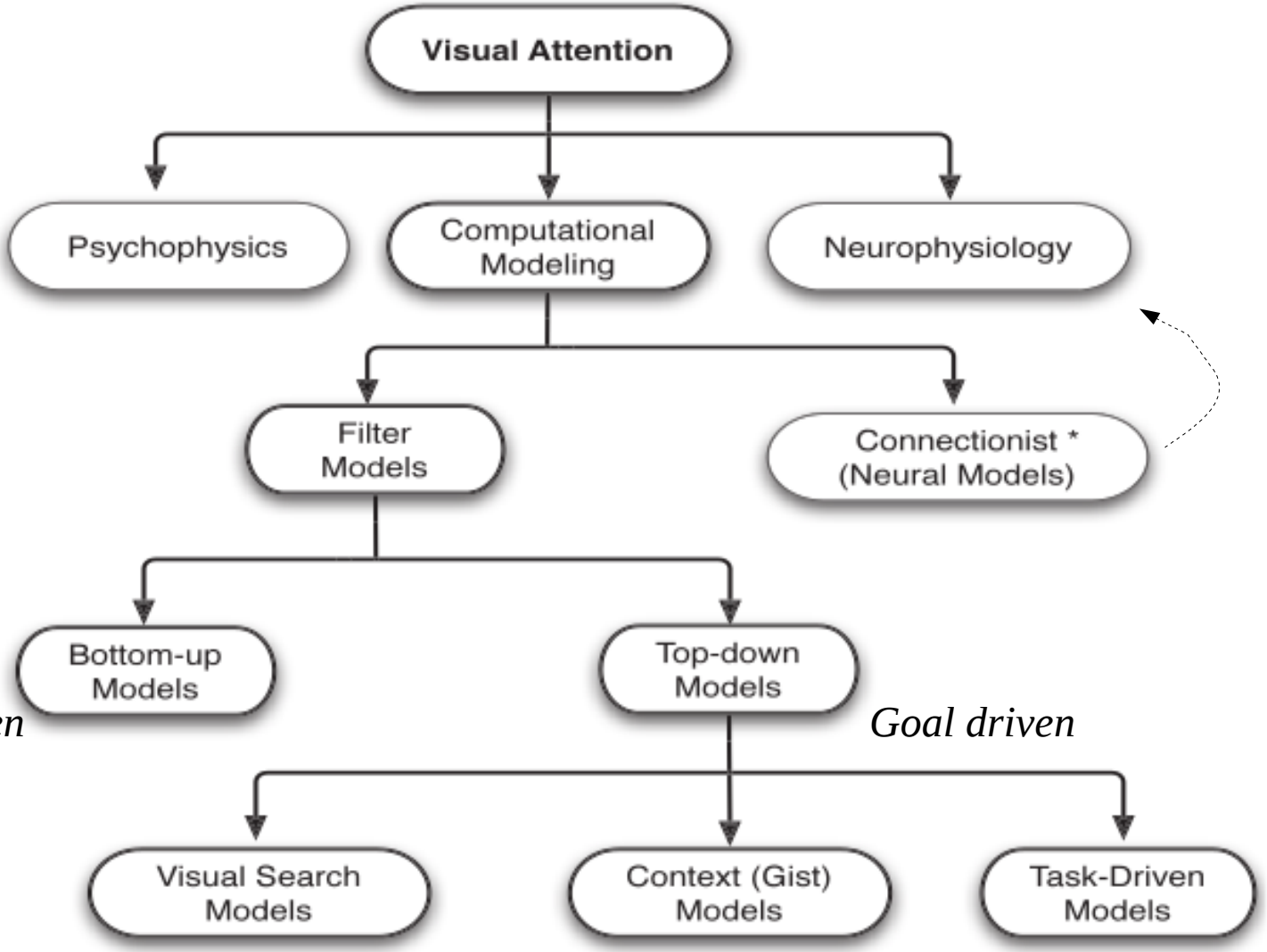


Bottom-up attention (saliency-driven)

Top-down attention

- Goal-driven, voluntary and slow
- Closed loop
- How do we decide where to look ?
  - Visual search model
    - Attention is drawn toward features of a target object
  - Gist model
    - Scene context to constrain locations that we look at
  - Task-driven model
    - More complex: Object features, gist, scene layout, prior knowledge

A taxonomy of visual attention models

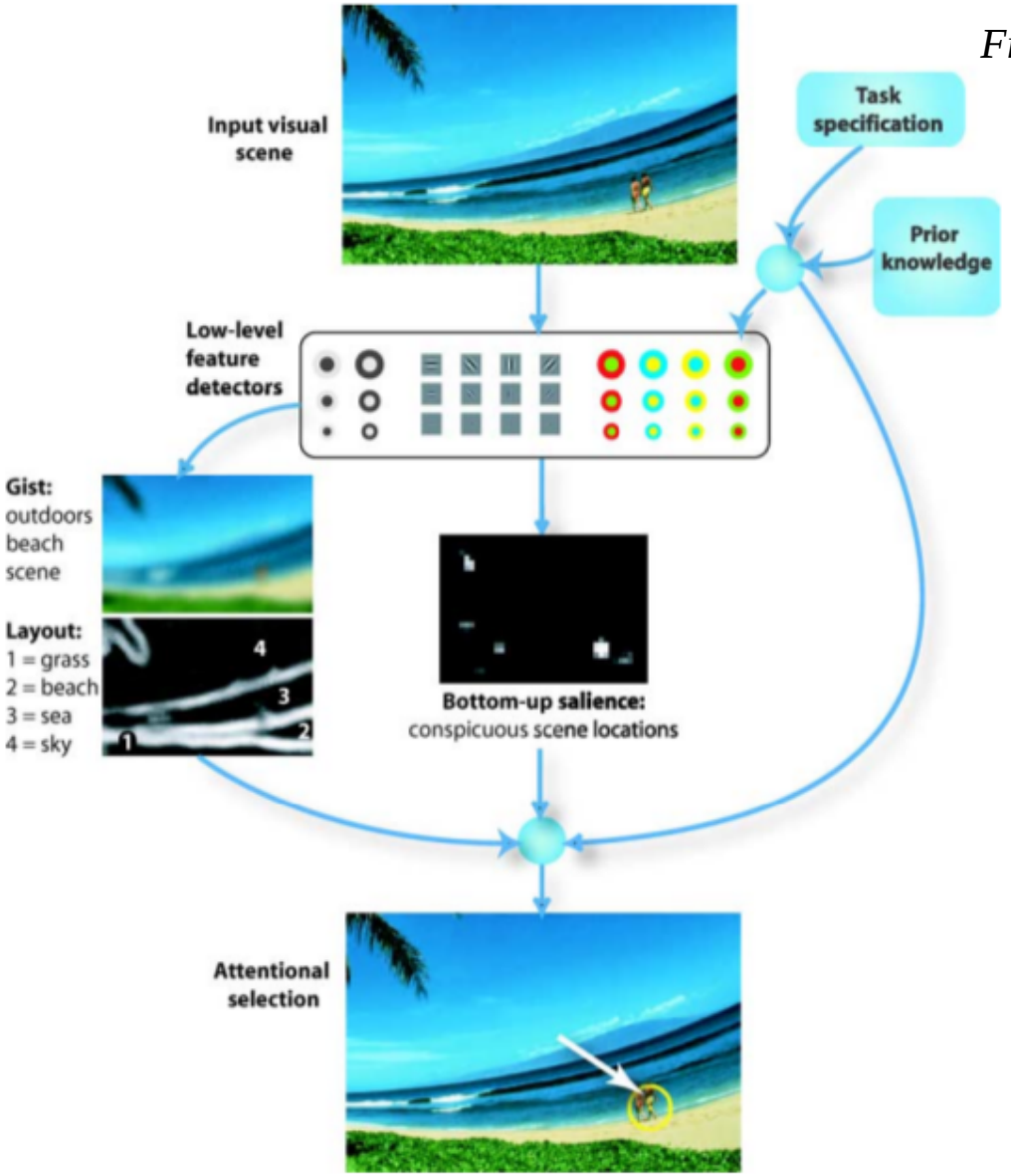


Stimulus driven

Goal driven

Navalpakkam & Itti, 2005

Task-driven attention model



Find humans in the scene

Humans are more likely to be found on the beach / in the water (in a beach scene)

Key components for slecting where to look at

- Task
- Gist and layout
- Knowledge of target

The approach

- Given: Task, Visual scene
- Interpret task with prior knowlege
  - Find task related entities and relations
  - Choose the most relevant entities
- Bias the low-level visual system with known features of the target
  - Modulate the saliency map with task knowledge
  - Combining top-down and bottom-up models
- Recognition of the entity at the focus of attention
  - Traditional and simple feature detectors
  - Hierarchical – generic class to specific instance / view
- Update
  - Remember the entity found in memory – discard irrelevant ones
  - Go to next FOA: find other / more relevant entities – update saliency bias
  - Explore spatial relations: if eyes are found – the mouth must be near

Biasing the saliency

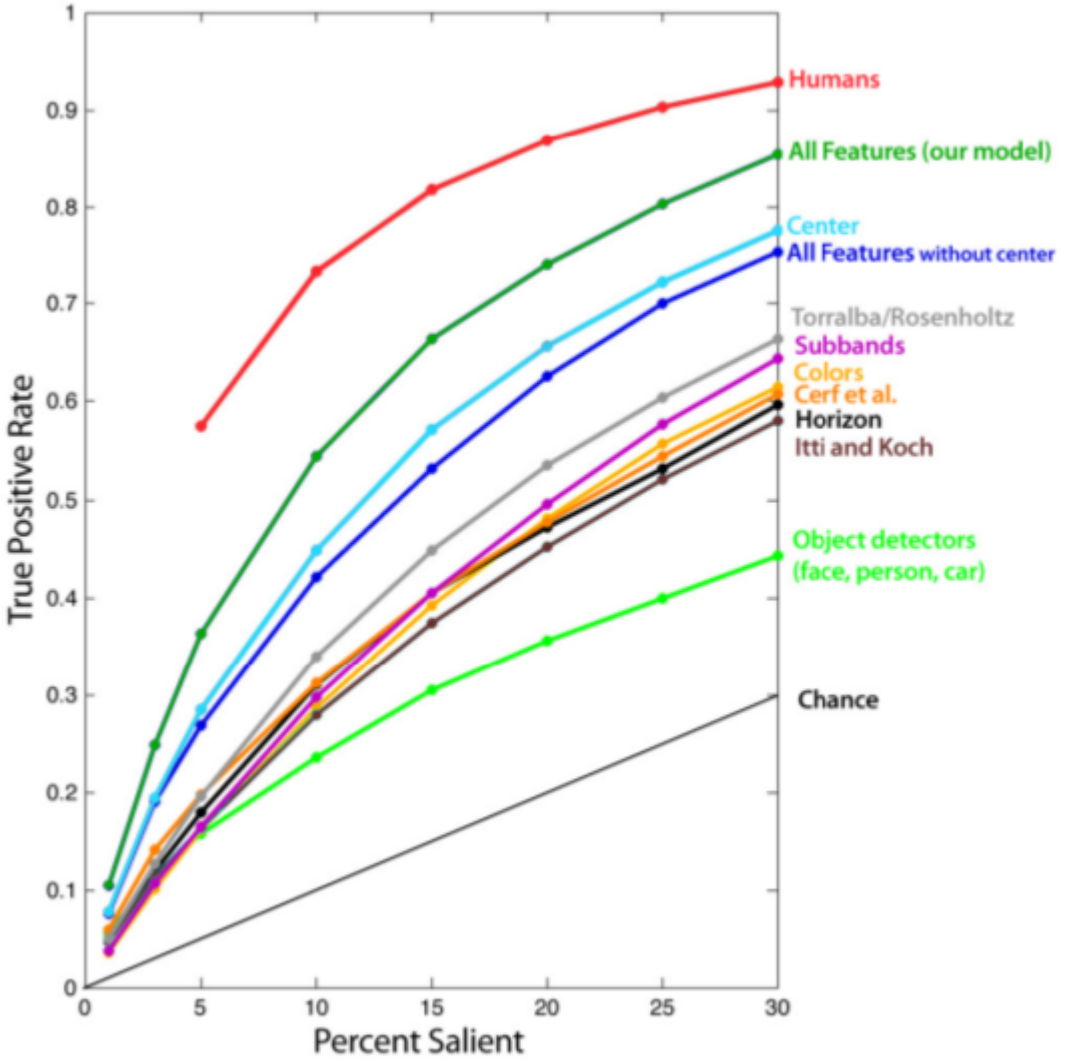
- The features of target object(s) are learnt
  - Same 42 features as in [Itti, et al. 1998]
- Context (local neighborhood) features are also learnt
  - 3 x 3 regions, with the object at the center
- The weights of the features are increased while computing the conspicuity map
  - *see slide 32*

Learning Saliency from humans

Judd & Torralba

- Machine learning based approach
- Training data (saliency as judged by humans)
  - 10 positive samples from top 20% most salient locations
  - 10 negative samples from bottom 70% least salient locations
- Feature set
  - Low level features: Intensity, Orientation and Color contrast
  - Mid-level features: Horizon line
  - High-level features: Face, person
  - Distance from the center
- SVM: Linear kernel

Results:


True Positive Rate
Percent Salient
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
5
10
15
20
25
30
Humans
All Features (our model)
Center
All Features without center
Torralba/Rosenholtz
Subbands
Colors
Cerf et al.
Horizon
Itti and Koch
Object detectors
(face, person, car)
Chance

References

- Borji & Itti (2013). State-of-the-Art in Visual Attention Modeling
- Judd & Torralba. Learning to predict where humans look
- Itti, et al. (1998) Model of saliency based visual attention ...
- Navalpakkam & Itti (2005) Modeling the influence of task on attention.