Selective visual attention: experiments, computational models and performance.

Olivier Le Meur
olemeur@irisa.fr

IRISA - University of Rennes 1

December 15, 2014

VERSION:

- 2010-2011: Document creation, done by O. Le Meur.
- 2011-2012: Correction, adding a slide on face/horizon line, done by O. Le Meur.
- 2013-2014: Minor modif.: adding a video to present Change Blindness, slide on eye tracking device, done by O. Le Meur.

Selective visual attention: from experiments to computational models

Introduction

Human Visual System (1/6)

COMPLEX VISUAL SCENES

(a) Prey vs Predator

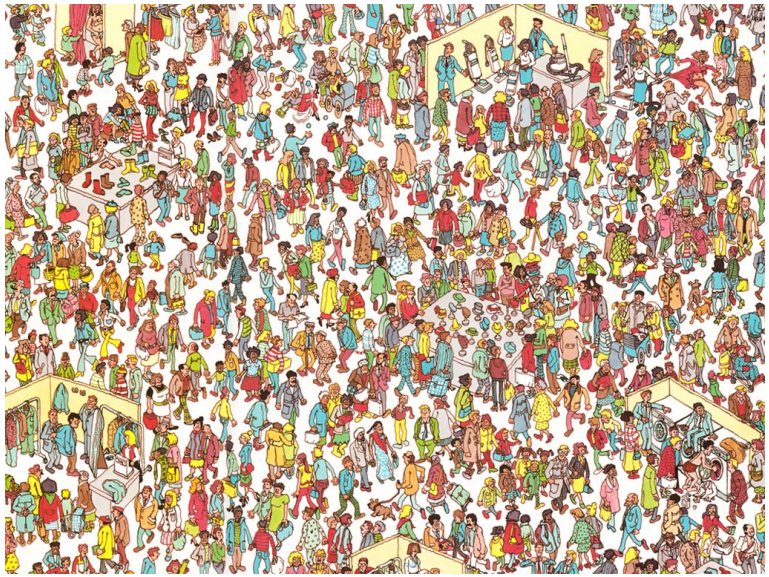
(b) King of the world?

(c) Salvador Dali

(d) René Magritte

Human Visual System (2/6)

Natural visual scenes are cluttered and contain many different objects that cannot all be processed simultaneously.



Where is Waldo, the young boy wearing the red-striped shirt...

Amount of information coming down the optic nerve $10^8 - 10^9$ bits per second

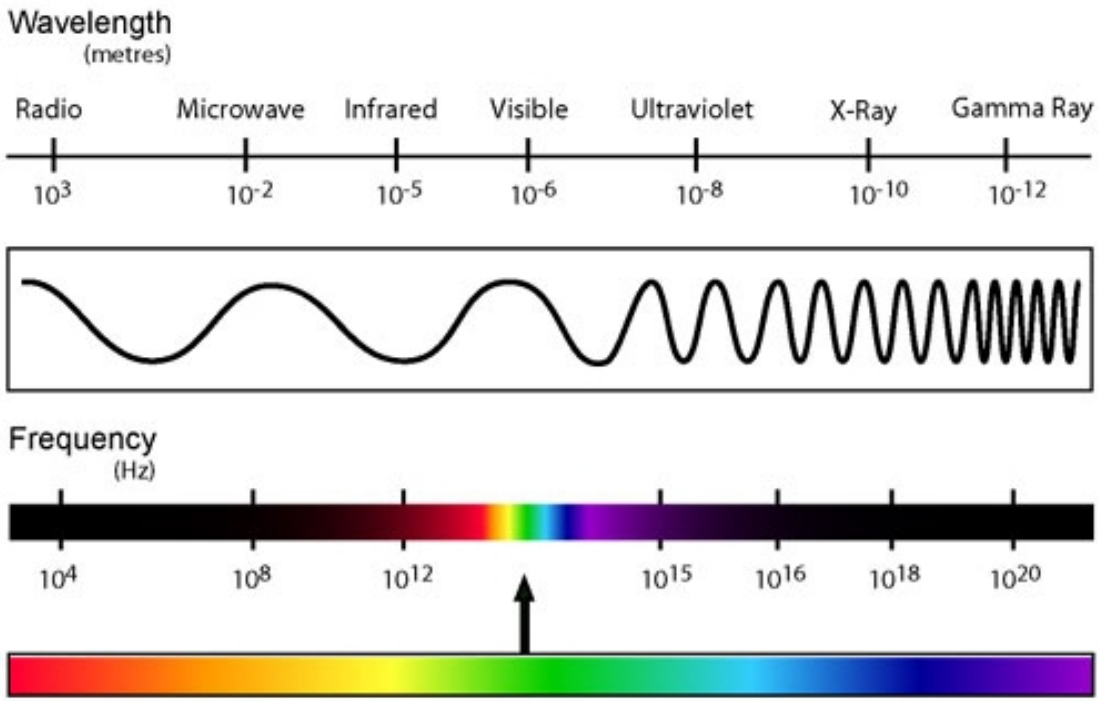
Far exceeds what the brain is capable of processing...

Human Visual System (3/6)

Two kinds of filters are used to reduce the amount of information: passive and active 'filters'.

➟ Passive filters:

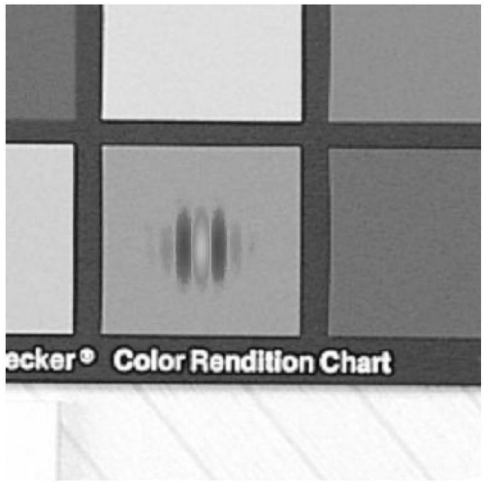
- A typical human eye will respond to wavelengths from about 390 to 750 nm ;



Human Visual System (4/6)

Two kinds of filters are used to reduce the amount of information: passive and active 'filters'.

- Passive filters:
  - The visibility threshold (Contrast Sensitivity Function) and visual masking.



Impact of Gabor patch on our perception (Nadenau, 2000)

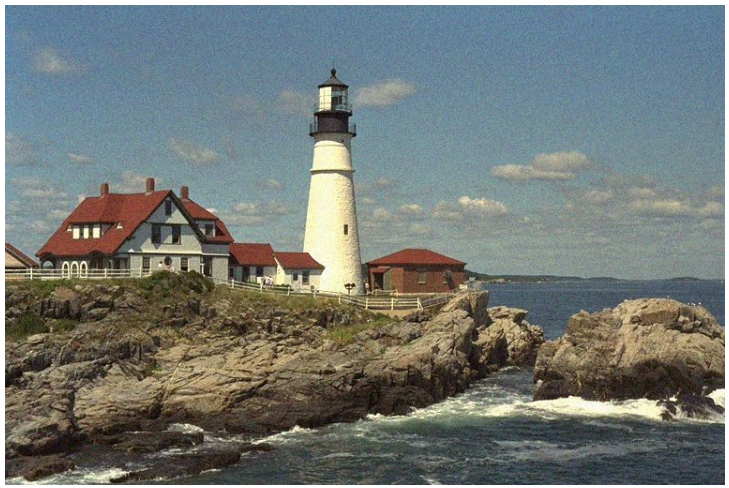
.

Human Visual System (5/6)

Two kinds of filters are used to reduce the amount of information: passive and active 'filters'.

- Passive filters:
  - The visibility threshold (Contrast Sensitivity Function) and visual masking.

(a) Original



(b) Original+uniform noise

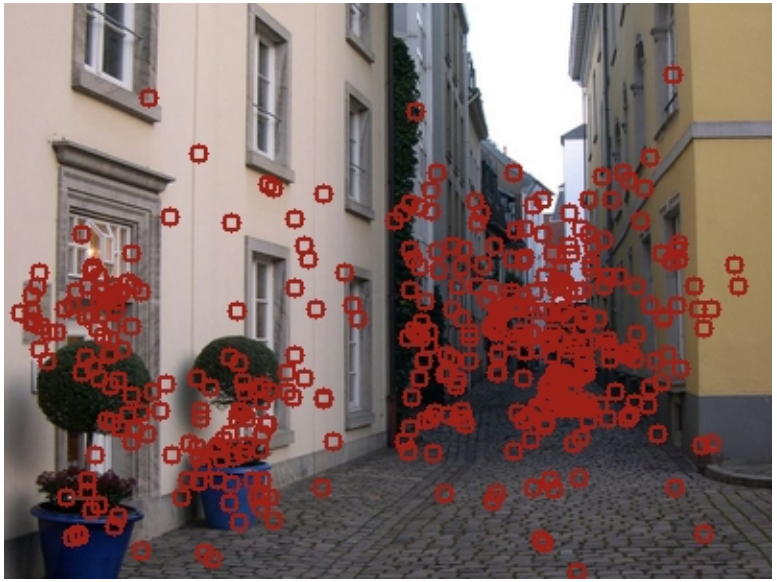
Human Visual System (6/6)

Two kinds of filters are used to reduce the amount of information: passive and active 'filters'.

➠ Active filters:

- To select certain portions of the input to be processed preferentially;
- Shifting the processing focus from one location to another in a serial fashion.

These two steps define what we call the selective visual attention.



Example of fixation map (40 observers).

A powerful demonstration: CHANGE BLINDNESS

WE DO NOT SEE EVERYTHING AROUND US!!!

A number of studies have shown that under certain circumstances, very large changes can be made in a picture without observers noticing them (Rensink et al., 1997).

A powerful demonstration: CHANGE BLINDNESS

What is changing between successive scenes?

A powerful demonstration: CHANGE BLINDNESS

What is changing between successive flashes in the following scenes?

A powerful demonstration: CHANGE BLINDNESS

- Example of change blindness.....

YouTube link: www.youtube.com/watch?v=ubNF9QNEQLA

Fundamental questions

WE DO NOT SEE EVERYTHING AROUND US!!!

Two majors conclusions come from the change blindness experiments:

- Observers never form a complete, detailed representation of their surroundings;
- Attention is required to perceive change.

Fundamental questions

WE DO NOT SEE EVERYTHING AROUND US!!!

Two majors conclusions come from the change blindness experiments:

- Observers never form a complete, detailed representation of their surroundings;
- Attention is required to perceive change.

It raises fundamental questions:

How do we select information from the scene?

Can we control where and what we attend to?

Answers (or part of them) come from different domains such as Psychophysics, Computational sciences, Cognitive Neurosciences, Cognitive Psychology...

Visual attention and eye movements

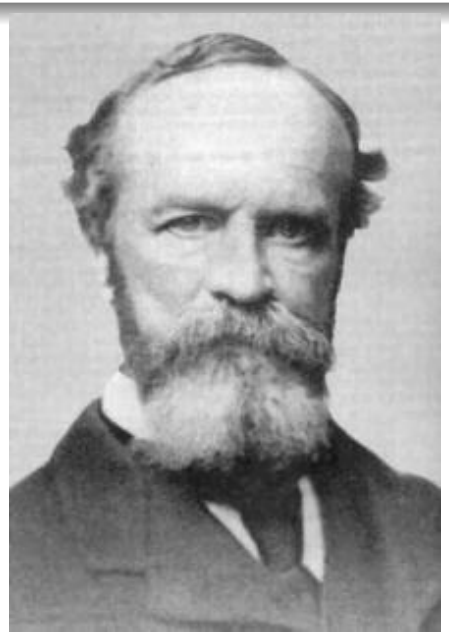
William James, an early definition of attention...

William James (1842-1910), a pioneering American psychologist and philosopher.

Everyone knows what attention is. It is the taking possession by the mind in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought... It implies withdrawal from some things in order to deal effectively with others.

Chapter XI : The Principles of Psychology.

- Attention
- To how many things can we attend at once?
- The varieties of attention
- The effects of attention
- The intimate nature of the attentive process
- Is voluntary attention a resultant or a force?



Visual attention definition

WE DO NOT SEE EVERYTHING AROUND US!!!

Visual attention

Posner proposed the following definition (Posner, 1980). Visual attention is used:

- to select important areas of our visual field (alerting);
- to search a target in cluttered scenes (searching).

There are several kinds of visual attention:

- Overt visual attention: involving eye movements;
- Covert visual attention: without eye movements (Covert fixations are not observable).

Visual attention definition

Bottom-Up vs Top-Down

➟ Bottom-Up: some things draw attention reflexively, in a task-independent way (Involuntary, Very quick, Unconscious);

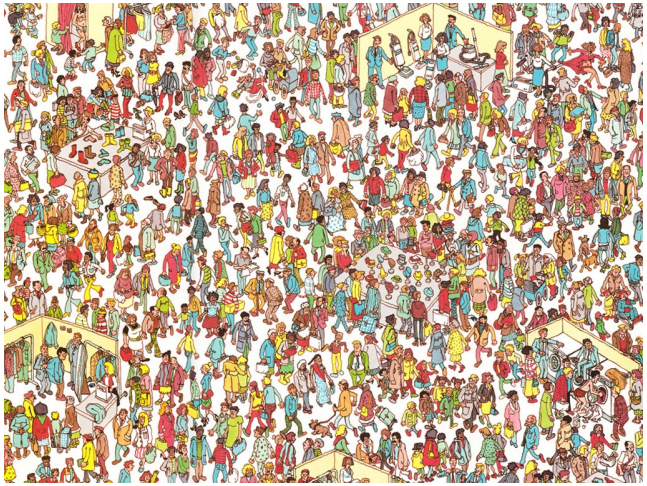
Visual attention definition

Bottom-Up vs Top-Down

- Bottom-Up: some things draw attention reflexively, in a task-independent way (Involuntary, Very quick, Unconscious);



- Top-Down: some things draw volitional attention, in a task-dependent way (Voluntary; Very slow; Conscious).



Eye Movements

There exists different kinds of eye movements:

- Saccade: quick eye movements from one fixation location to another. The length of the saccade is between 4 to 12 degrees of visual angle (spontaneous, reflexive, voluntary);
- Fixation: phase during which eyes is almost stationnary. The typical duration is about 200-300 ms. But it depends on a number of factors (depth of processing (Velichkovsky, 2002); ease or difficulty to perceive something (Mannan et al., 1995));
- Smooth pursuit: voluntary tracking of moving stimulus;
- Vergence: coordinated movement of both eyes, converging for objects moving towards and diverging for objects moving away from the eyes.

Eye Movements

There exists different kinds of eye movements:

- Saccade: quick eye movements from one fixation location to another. The length of the saccade is between 4 to 12 degrees of visual angle (spontaneous, reflexive, voluntary);
- Fixation: phase during which eyes is almost stationnary. The typical duration is about 200-300 ms. But it depends on a number of factors (depth of processing (Velichkovsky, 2002); ease or difficulty to perceive something (Mannan et al., 1995));
- Smooth pursuit: voluntary tracking of moving stimulus;
- Vergence: coordinated movement of both eyes, converging for objects moving towards and diverging for objects moving away from the eyes.

Eye Movements

There exists different kinds of eye movements:

- Saccade: quick eye movements from one fixation location to another. The length of the saccade is between 4 to 12 degrees of visual angle (spontaneous, reflexive, voluntary);
- Fixation: phase during which eyes is almost stationnary. The typical duration is about 200-300 ms. But it depends on a number of factors (depth of processing (Velichkovsky, 2002); ease or difficulty to perceive something (Mannan et al., 1995));
- Smooth pursuit: voluntary tracking of moving stimulus;
- Vergence: coordinated movement of both eyes, converging for objects moving towards and diverging for objects moving away from the eyes.

Eye Movements

There exists different kinds of eye movements:

- Saccade: quick eye movements from one fixation location to another. The length of the saccade is between 4 to 12 degrees of visual angle (spontaneous, reflexive, voluntary);
- Fixation: phase during which eyes is almost stationnary. The typical duration is about 200-300 ms. But it depends on a number of factors (depth of processing (Velichkovsky, 2002); ease or difficulty to perceive something (Mannan et al., 1995));
- Smooth pursuit: voluntary tracking of moving stimulus;
- Vergence: coordinated movement of both eyes, converging for objects moving towards and diverging for objects moving away from the eyes.

Eye Movements

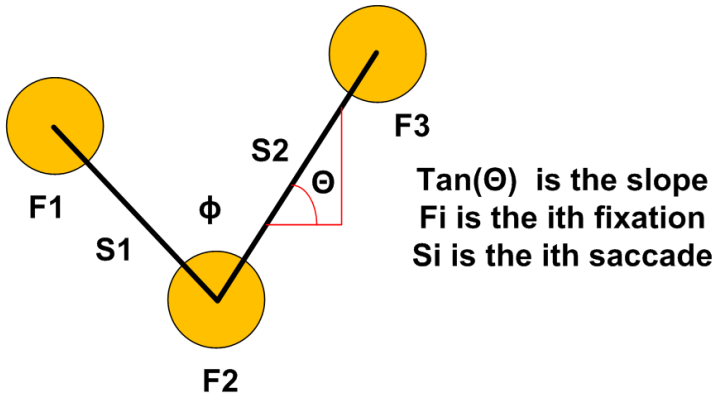
There exists different kinds of eye movements:

- Saccade: quick eye movements from one fixation location to another. The length of the saccade is between 4 to 12 degrees of visual angle (spontaneous, reflexive, voluntary);
- Fixation: phase during which eyes is almost stationnary. The typical duration is about 200-300 ms. But it depends on a number of factors (depth of processing (Velichkovsky, 2002); ease or difficulty to perceive something (Mannan et al., 1995));
- Smooth pursuit: voluntary tracking of moving stimulus;
- Vergence: coordinated movement of both eyes, converging for objects moving towards and diverging for objects moving away from the eyes.

A scanpath is a sequence of eye movements (fixations - smooth pursuit - saccades...).

Eye Movements

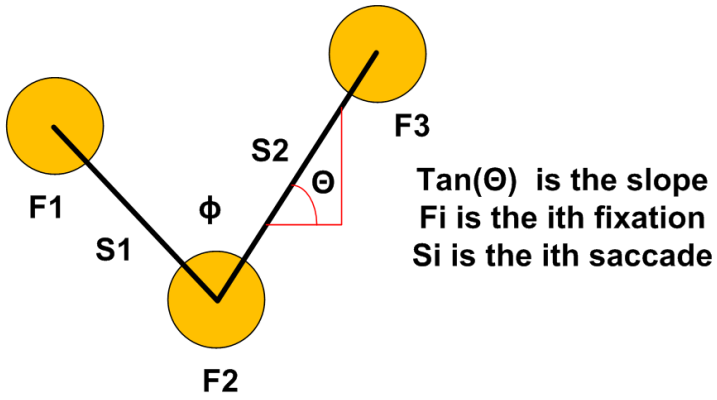
- Slope is the direction of the saccade direction;
- Length of the saccade path estimated by the Euclidean distance (L_2 norm, $\|.\|_2$);
- Orientation is the angle between two succeeding saccades;
- Velocity measures the rate of fixation location change.
  - Mean velocity between the beginning and end of the saccade
  - Peak velocity is the highest velocity reached during the saccade.



- Temporal profile of saccade velocity:

Eye Movements

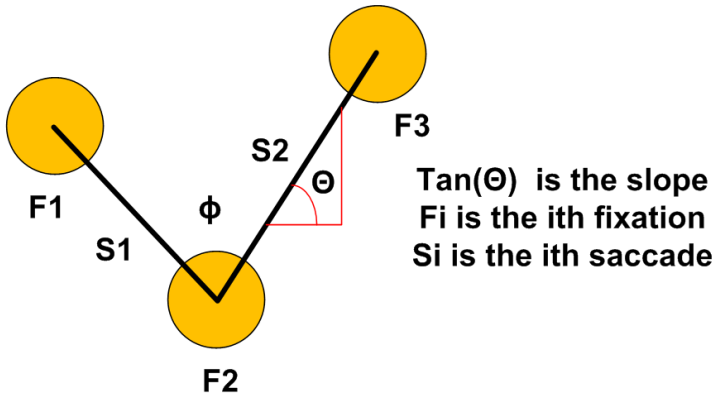
- Slope is the direction of the saccade direction;
- Length of the saccade path estimated by the Euclidean distance (L_2 norm, $\|.\|_2$);
- Orientation is the angle between two succeeding saccades;
- Velocity measures the rate of fixation location change.
  - Mean velocity between the beginning and end of the saccade
  - Peak velocity is the highest velocity reached during the saccade.



- Temporal profile of saccade velocity:

Eye Movements

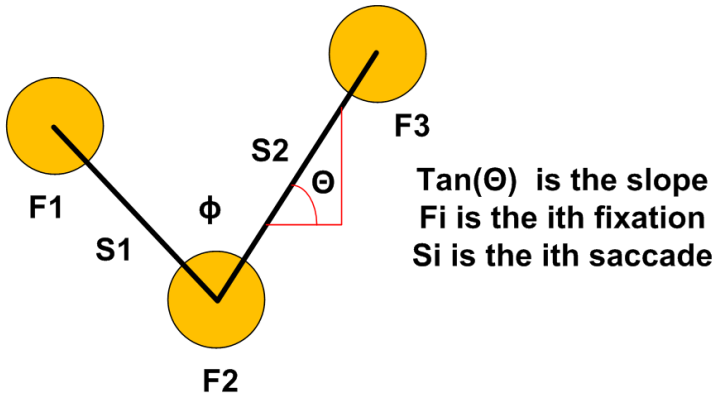
- Slope is the direction of the saccade direction;
- Length of the saccade path estimated by the Euclidean distance (L_2 norm, $\|.\|_2$);
- Orientation is the angle between two succeeding saccades;
- Velocity measures the rate of fixation location change.
  - Mean velocity between the beginning and end of the saccade
  - Peak velocity is the highest velocity reached during the saccade.



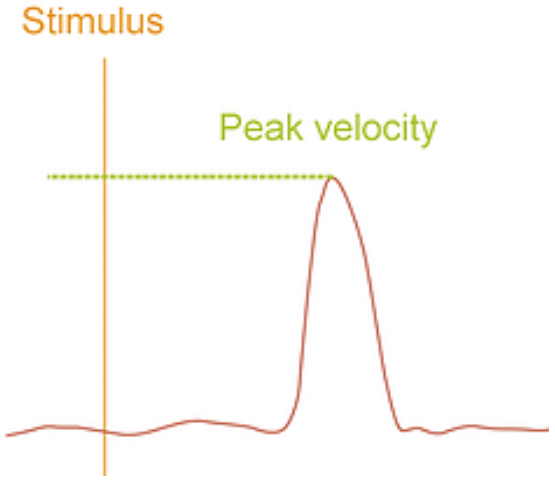
- Temporal profile of saccade velocity:

Eye Movements

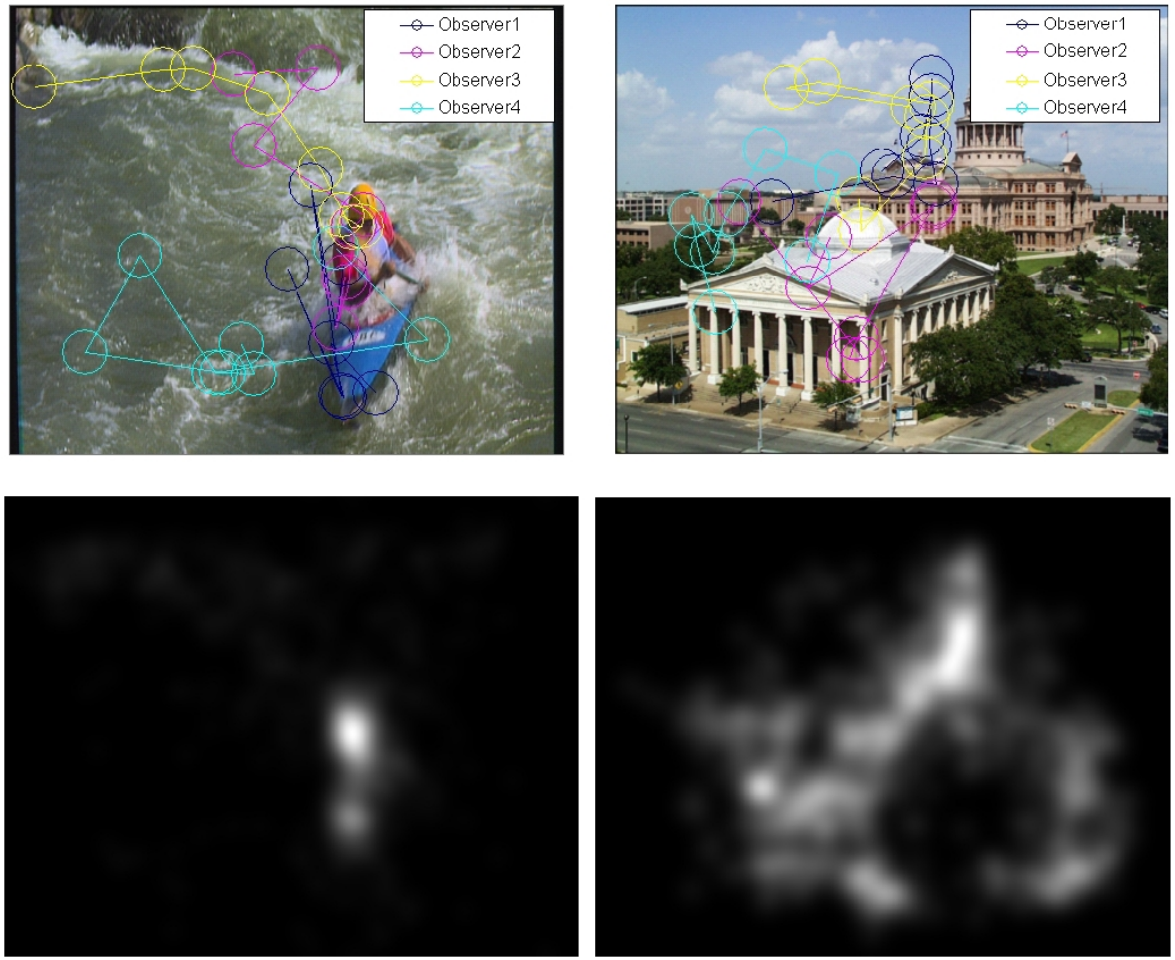
- Slope is the direction of the saccade direction;
- Length of the saccade path estimated by the Euclidean distance (L_2 norm, $\|.\|_2$);
- Orientation is the angle between two succeeding saccades;
- Velocity measures the rate of fixation location change.
  - Mean velocity between the beginning and end of the saccade
  - Peak velocity is the highest velocity reached during the saccade.



- Temporal profile of saccade velocity:

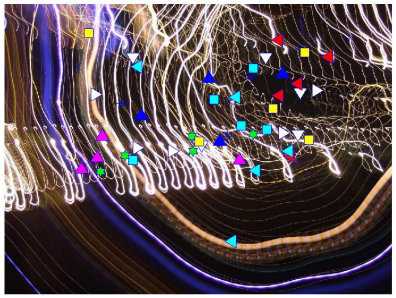


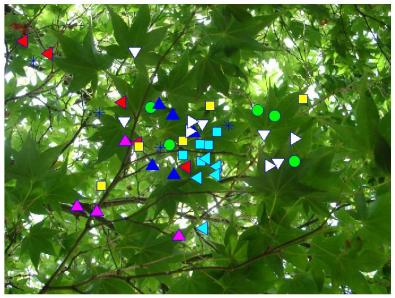
Eye Movements

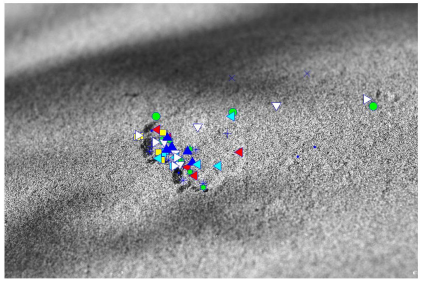
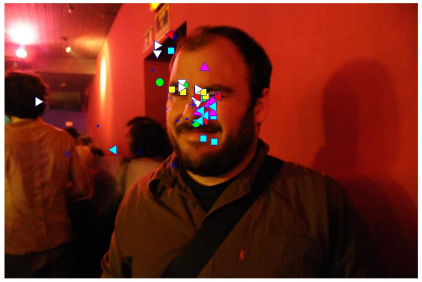


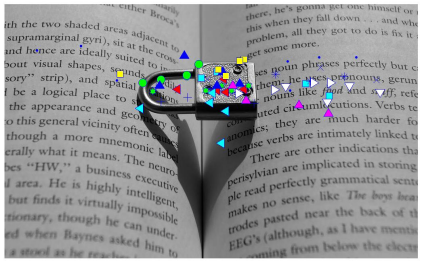
From (Le Meur and Chevet, 2010).

Eye Movements



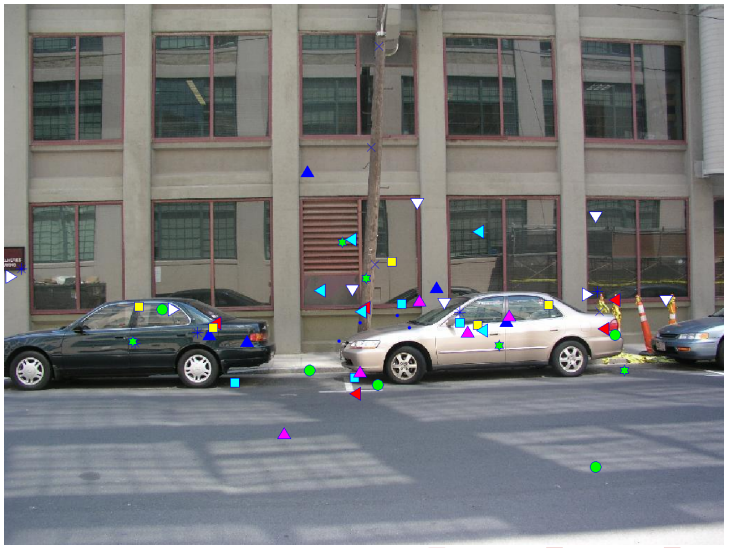




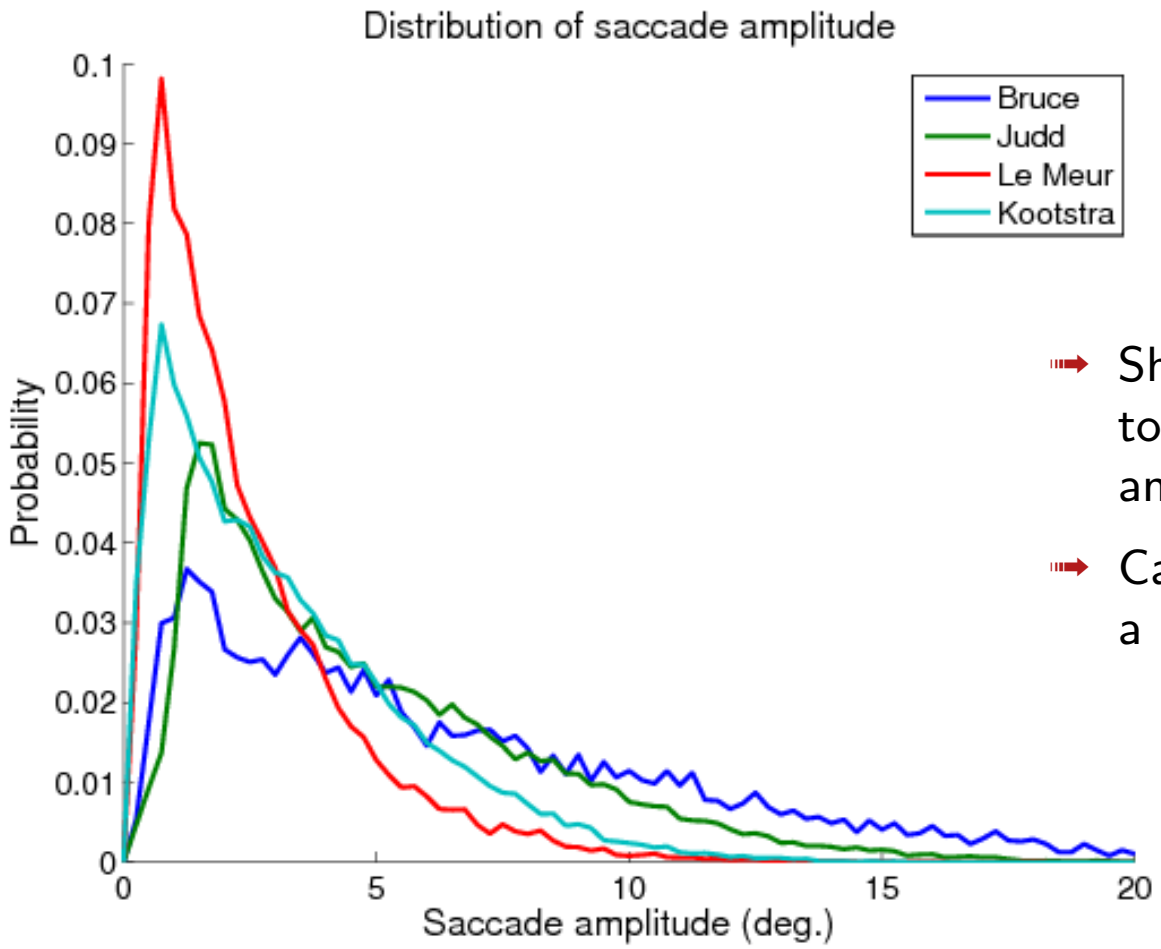

From MIT database.

Eye Movements



Eye Movements

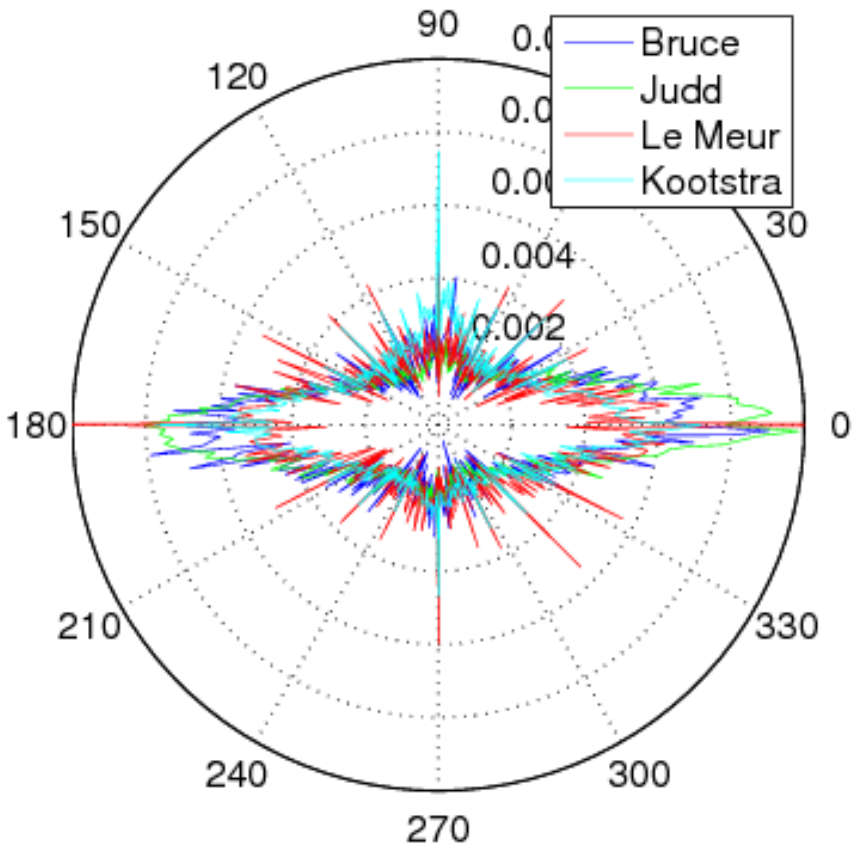
Systematic tendencies on natural scenes



- Short saccades around 1 to 3 degrees of visual angle;
- Can be approximated by a Gamma law.

Eye Movements

Systematic tendencies on natural scenes

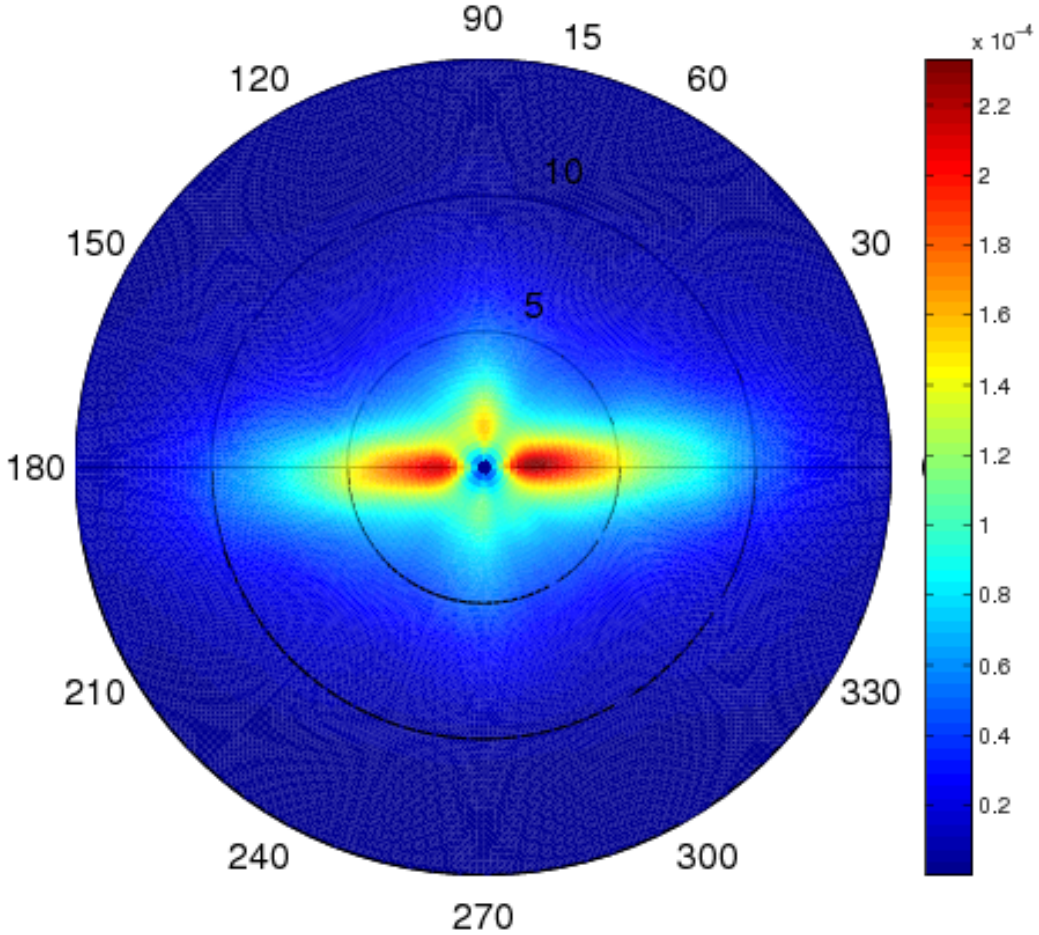


Distribution of saccade orientation

- Anisotropic shape;
- More horizontal saccades than vertical ones;
- Very few diagonal saccades.

Eye Movements

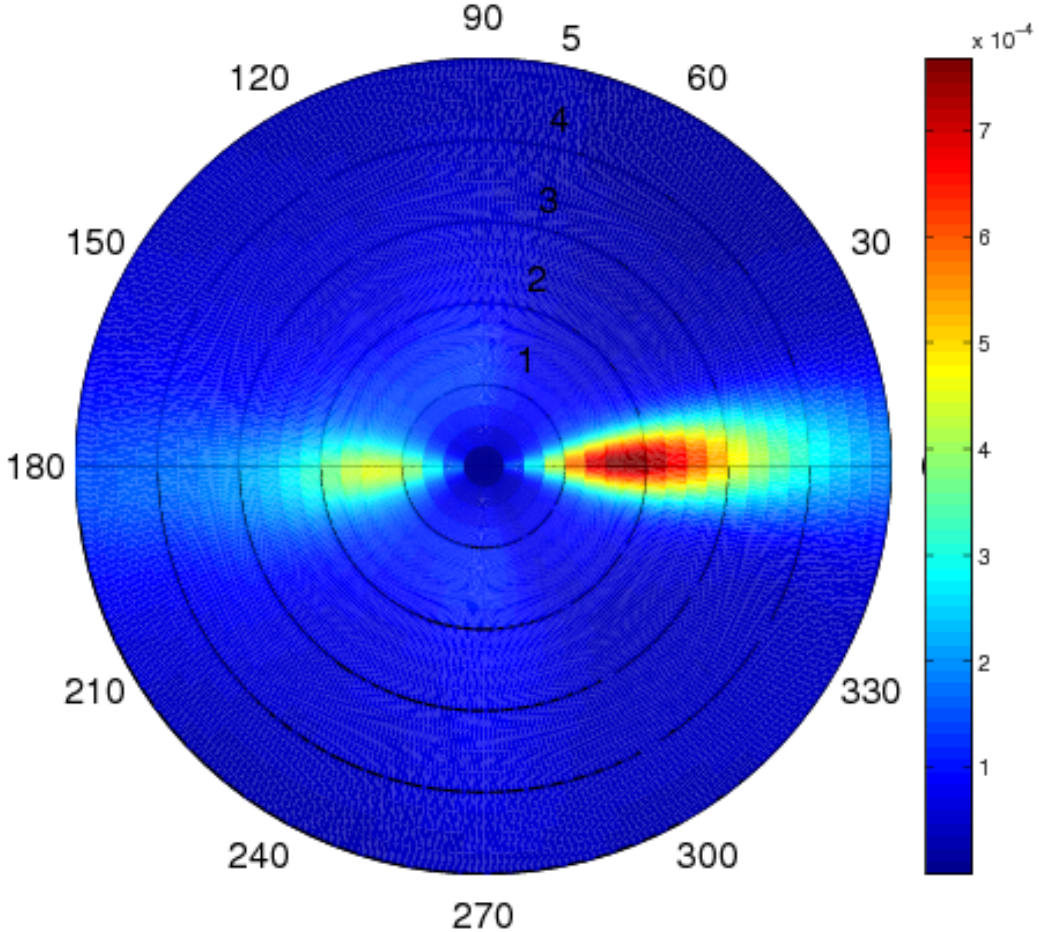
Systematic tendencies on natural scenes



- Joint distribution of saccade amplitudes and orientations;
- Strong horizontal bias and small saccade.

Eye Movements

Systematic tendencies on webpage



- Joint distribution of saccade amplitudes and orientations;
- Strong horizontal bias in the rightward direction;
- F-shaped pattern.

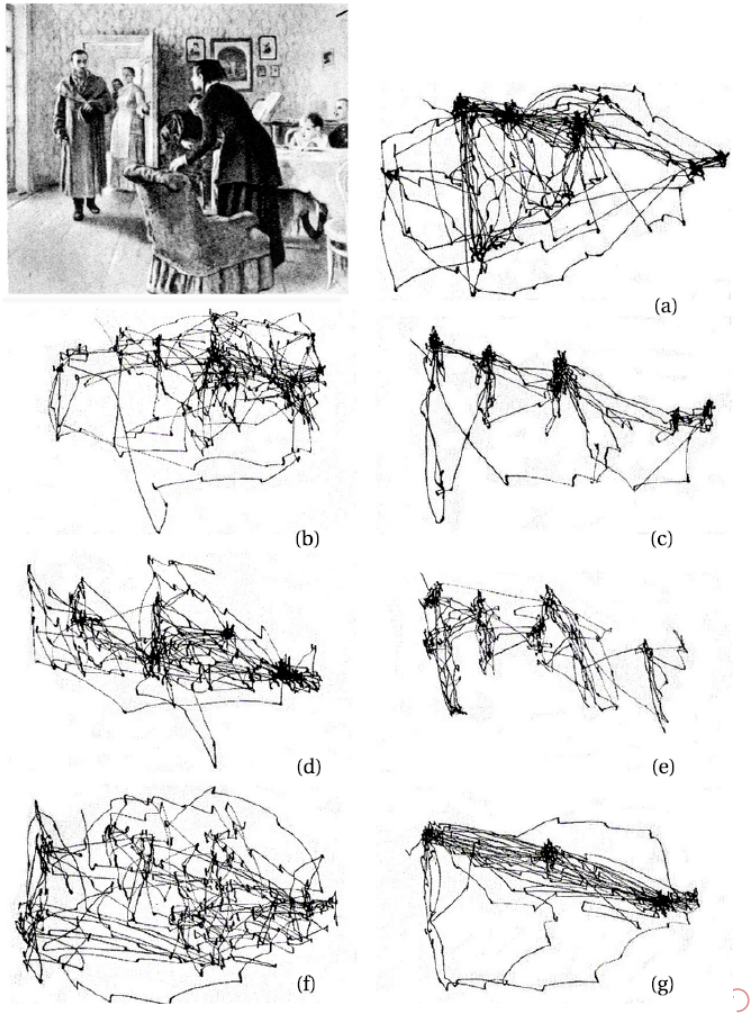
From (Shen and Zhao, 2014)'s eye fixation dataset.

Visual attention and psychophysics

(Yarbus, 1967)

(Yarbus, 1967) demonstrated how eye movements changed depending on the question asked of the subject.

1. No question asked
2. Judge economic status
3. What were they doing before the visitor arrived?
4. What clothes are they wearing?
5. Where are they?
6. How long is it since the visitor has seen the family?
7. Estimate how long the unexpected visitor had been away from the family.



Each recording lasted 3 minutes.

(Yarbus, 1967)

A. Yarbus showed that our attention depends on bottom-up features and on top-down information:

- Bottom-up attention (also called exogenous): some things draw attention reflexively, in a task-independent way...
  - Involuntary;
  - Very quick;
  - Unconscious.

- Top-down attention (also called endogenous): some things draw volitional attention, in a task-dependent way...
  - Voluntary (for instance to perform a task, find the red target);
  - Very slow;
  - Conscious.

(Yarbus, 1967)

A. Yarbus showed that our attention depends on bottom-up features and on top-down information:

- Bottom-up attention (also called exogenous): some things draw attention reflexively, in a task-independent way...
  - Involuntary;
  - Very quick;
  - Unconscious.

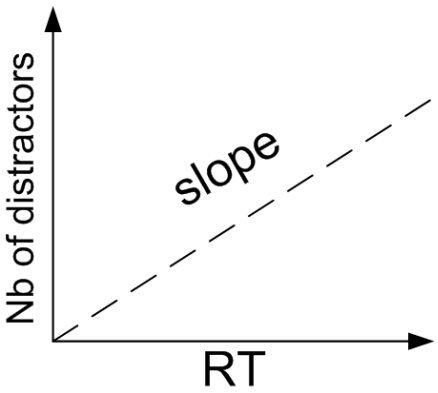
- Top-down attention (also called endogenous): some things draw volitional attention, in a task-dependent way...
  - Voluntary (for instance to perform a task, find the red target);
  - Very slow;
  - Conscious.

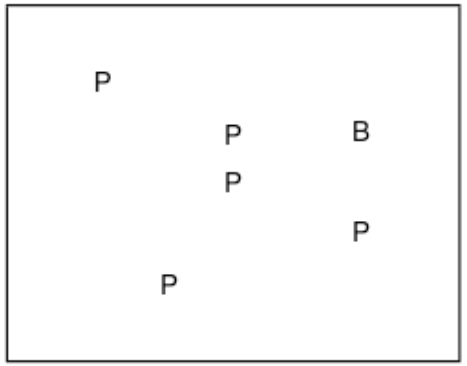
Visual search

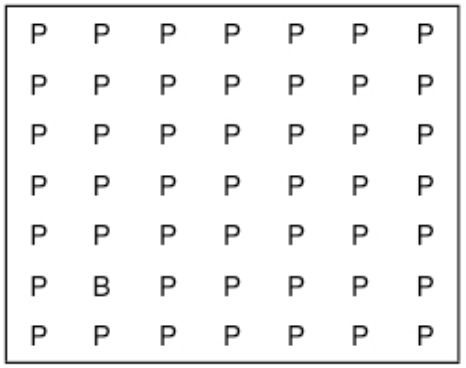
Visual search-based experiments rely on the amount of time taken by an observer to indicate whether a search target is present or not.

Reaction time

An interesting indicator is to have the impact of number of distractors on the reaction time.







The more distractors there are, the longer it takes to find the letter B in the array (From Wikipedia).

Visual search

Different kinds of experiment can be made:

- Feature search:

  The target differs from the distractors by a unique visual featue (color, size, orientation..). They are called pre-attentive features (or early) because the RT is almost constant whatever the number of distractors.

  → There are in the brain specialized visual receptors responding to different visual features...

- Conjunction search:

  The target is not defined by any single unique visual feature, but by a combination of two or more features. The RT increases indicating a serial search.

Visual search

Different kinds of experiment can be made:

- Feature search:

  The target differs from the distractors by a unique visual featue (color, size, orientation..). They are called pre-attentive features (or early) because the RT is almost constant whatever the number of distractors.

  → There are in the brain specialized visual receptors responding to different visual features...

- Conjunction search:

  The target is not defined by any single unique visual feature, but by a combination of two or more features. The RT increases indicating a serial search.

Visual search

Different kinds of experiment can be made:

- Feature search:

  The target differs from the distractors by a unique visual featue (color, size, orientation..). They are called pre-attentive features (or early) because the RT is almost constant whatever the number of distractors.

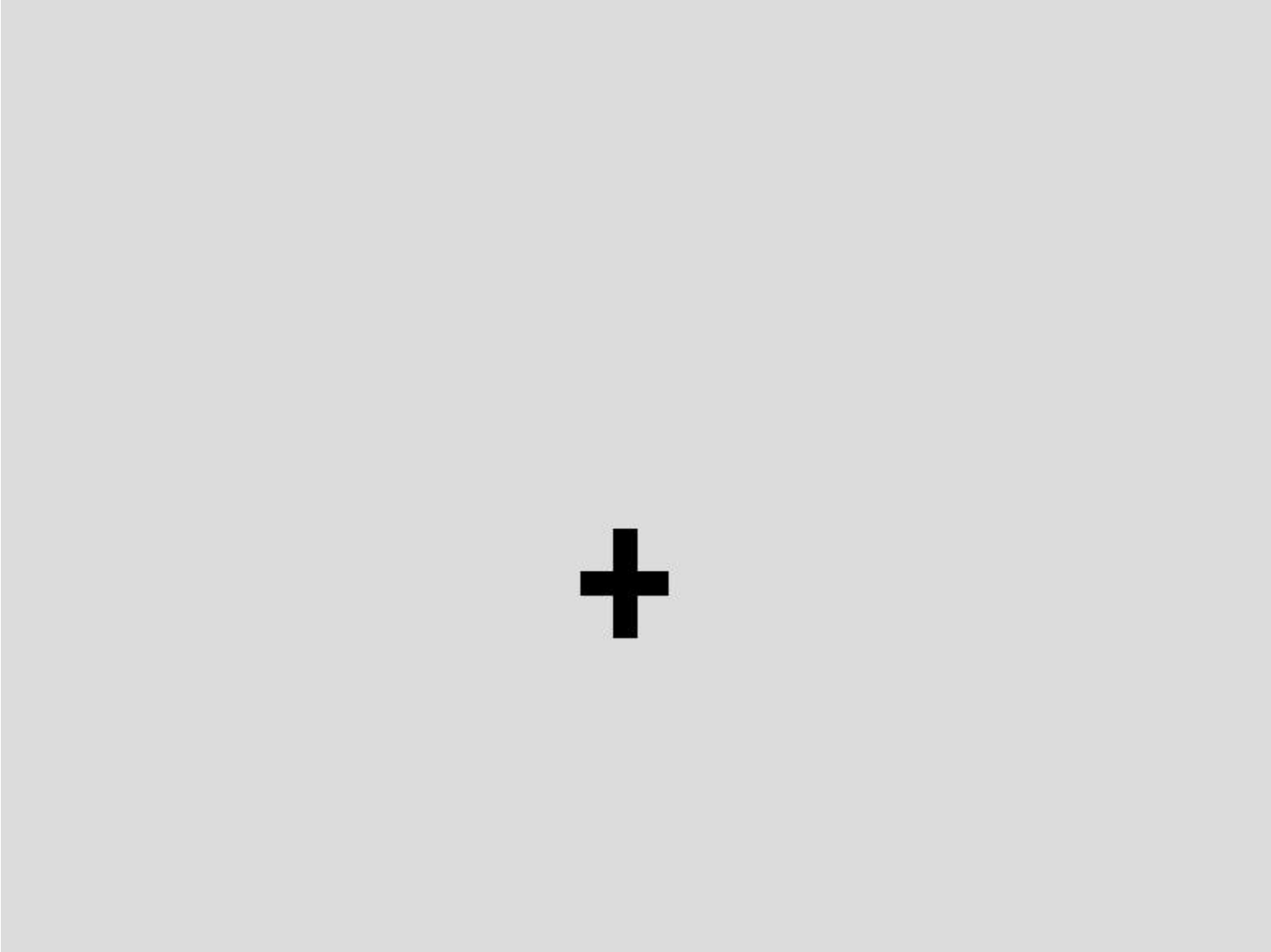
  → There are in the brain specialized visual receptors responding to different visual features...

- Conjunction search:

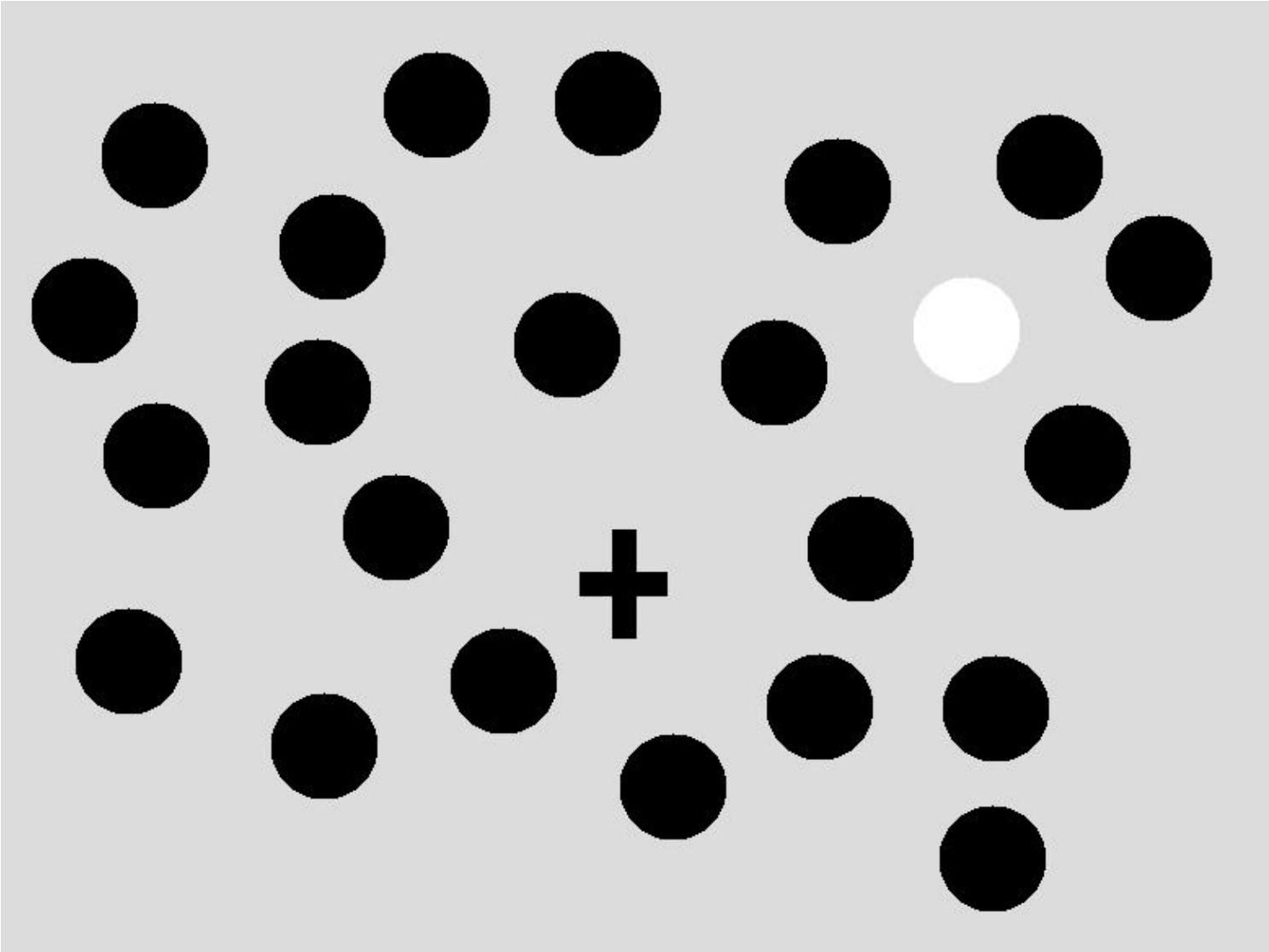
  The target is not defined by any single unique visual feature, but by a combination of two or more features. The RT increases indicating a serial search.

Feature Integration Theory of (Treisman and Gelade, 1980).

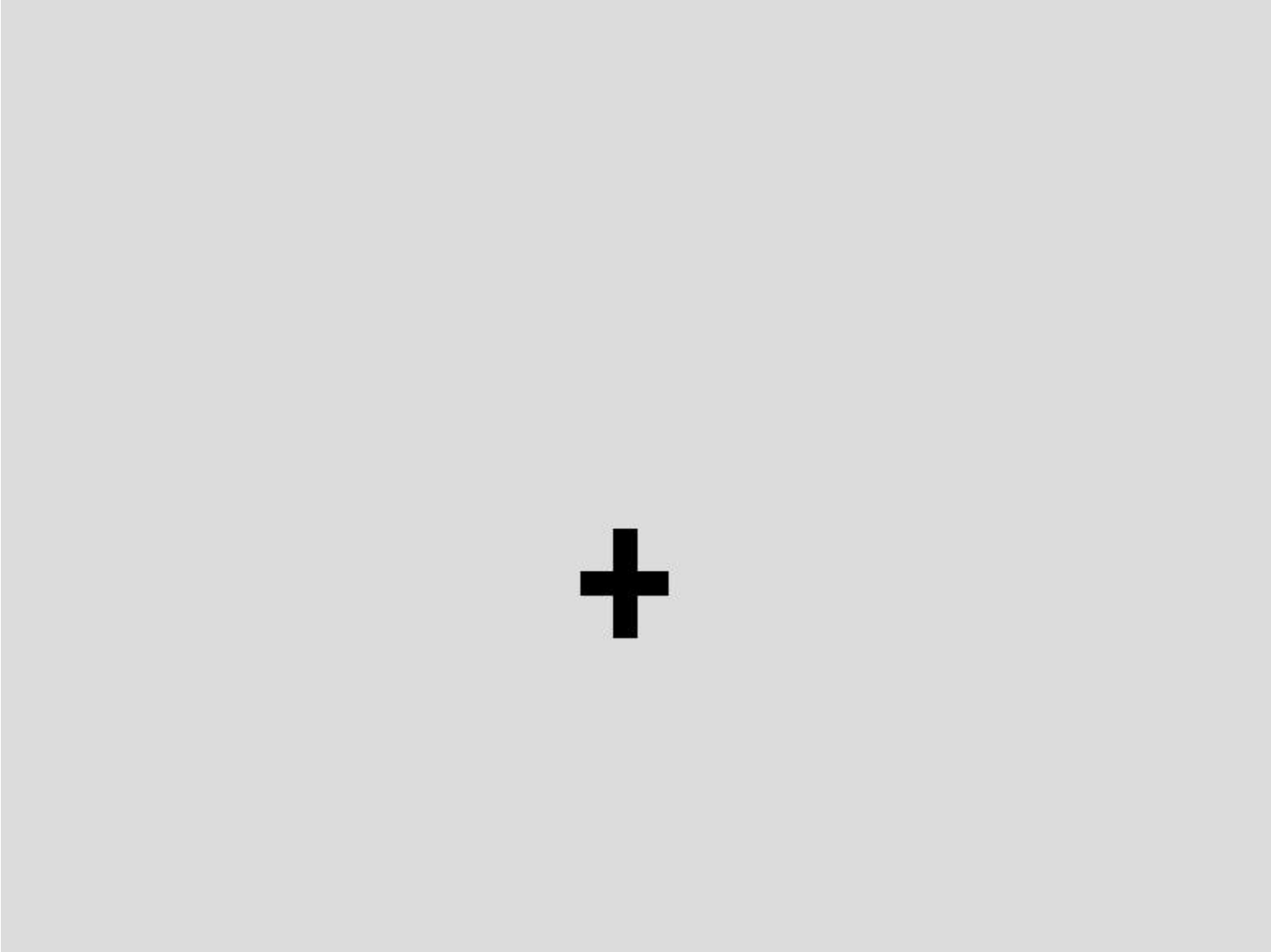
Feature search: Find the white circle.



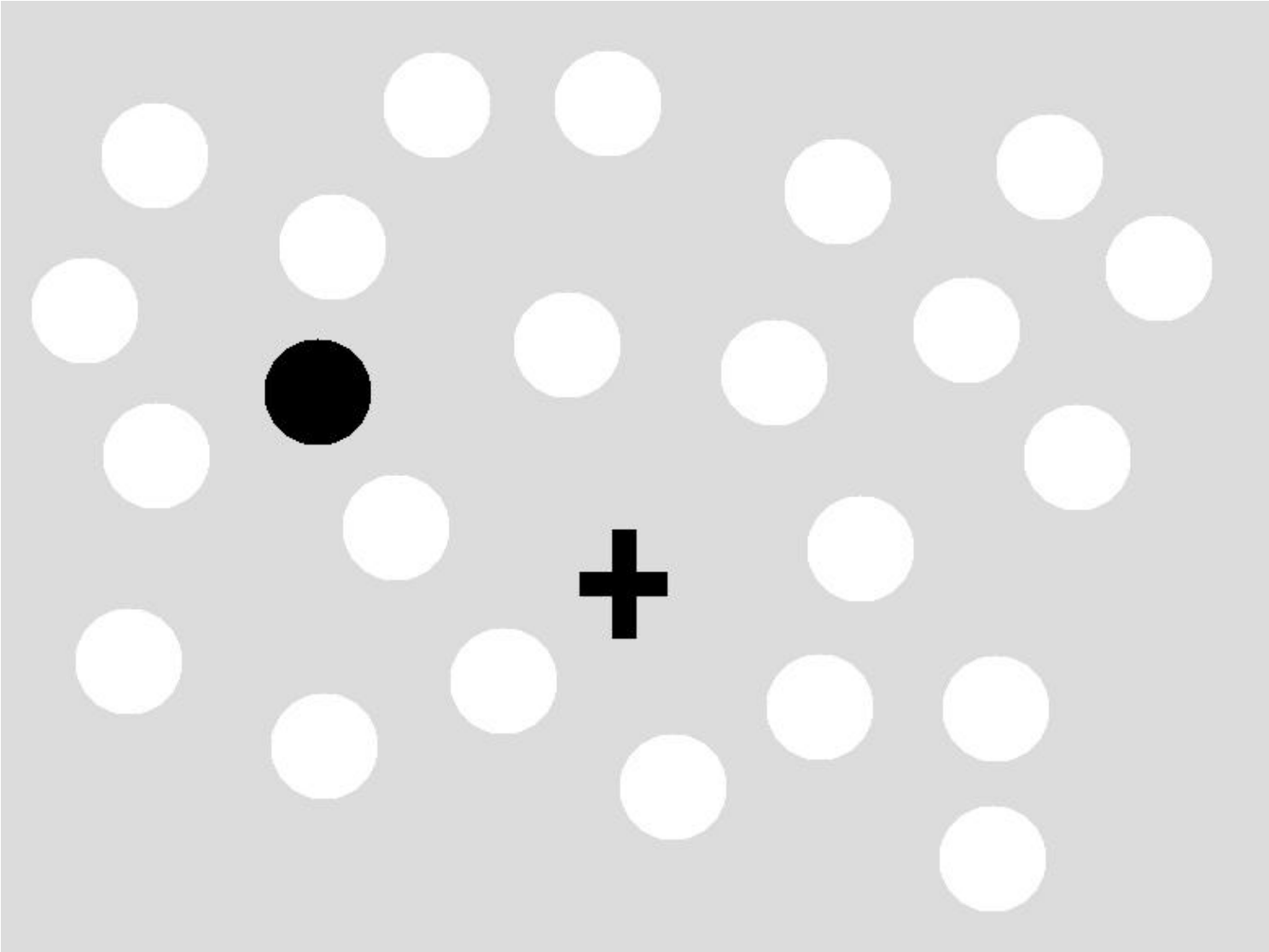
Feature search: Find the white circle.



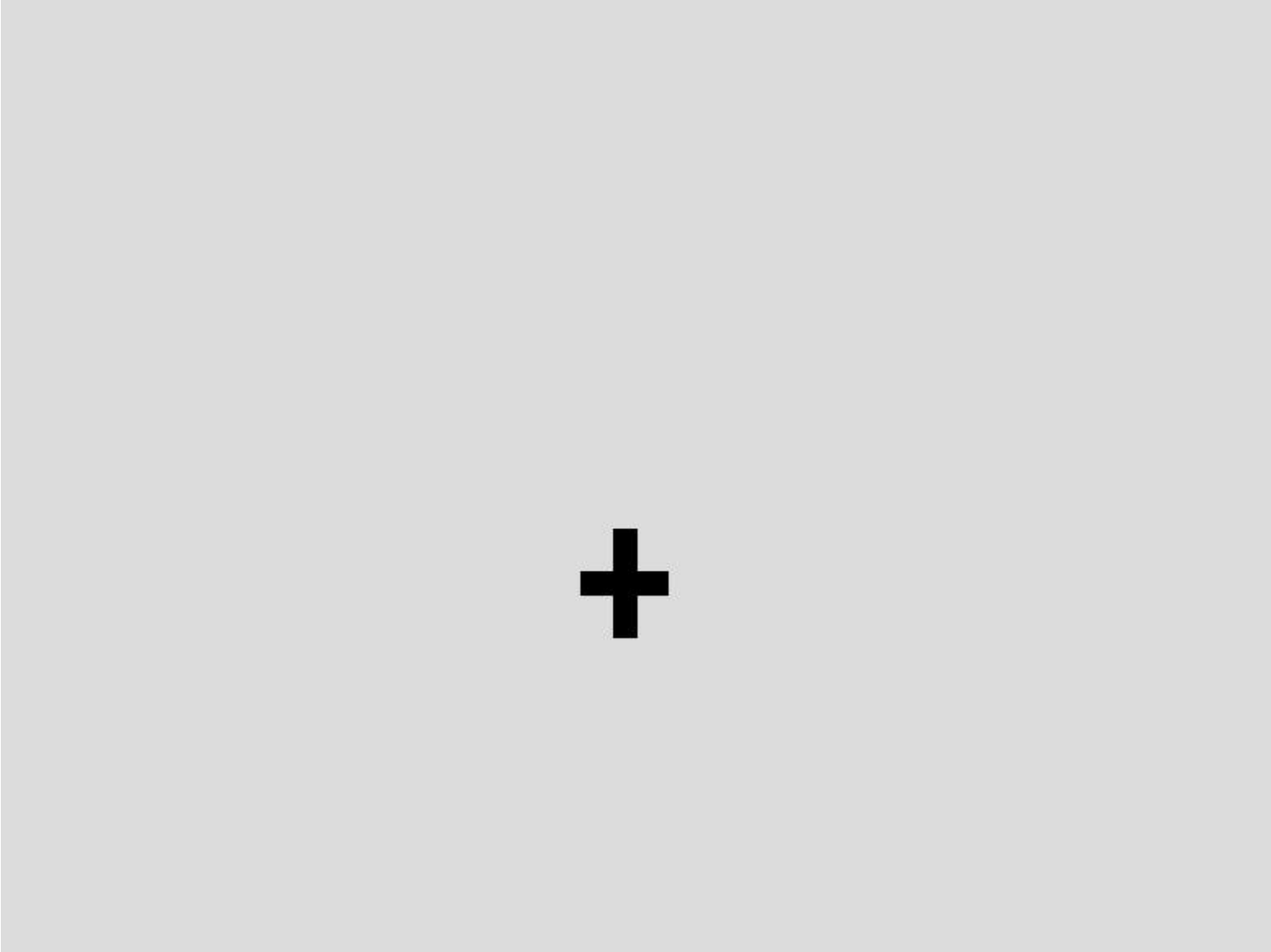
Feature search: Find the black circle.



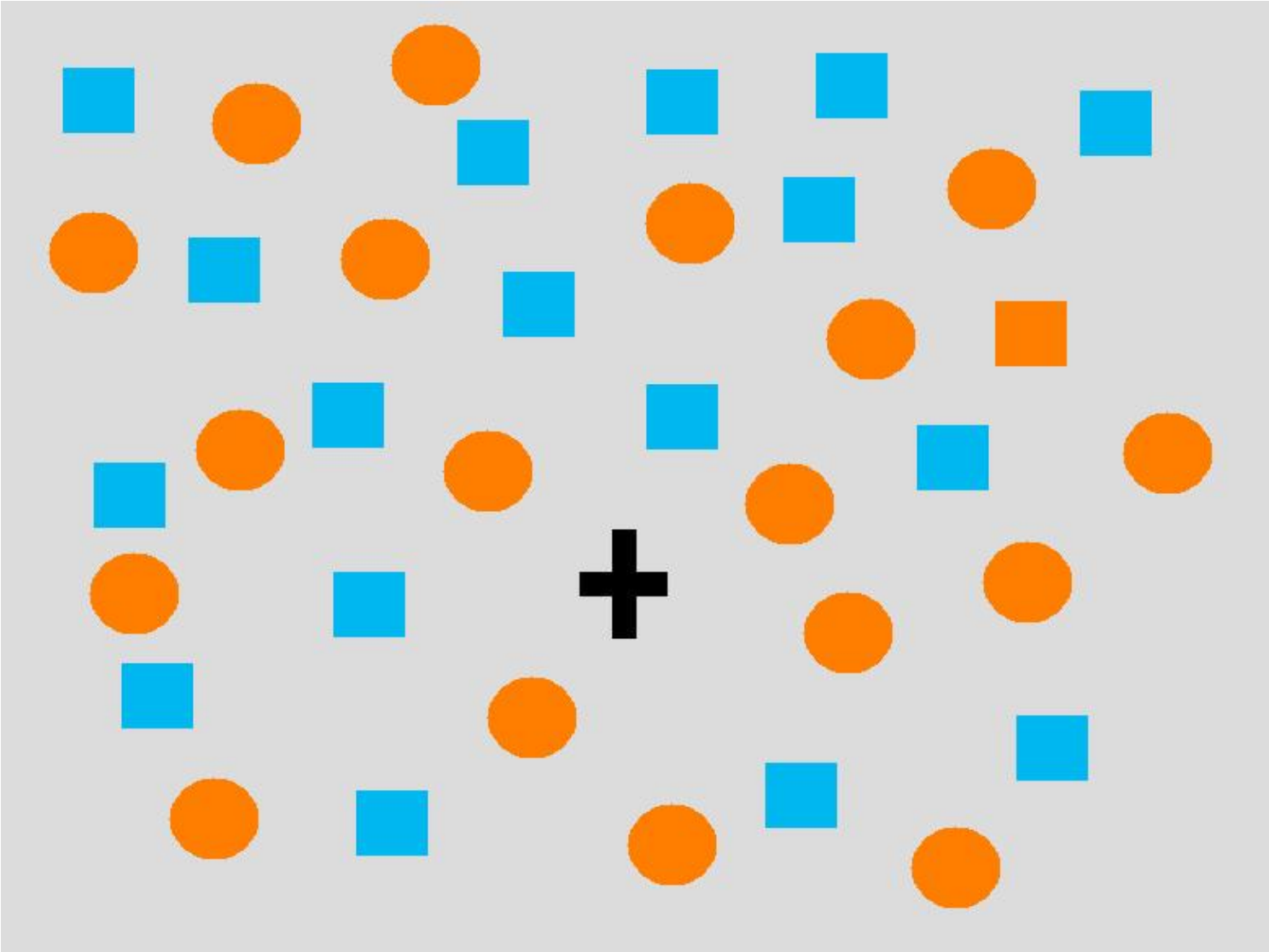
Feature search: Find the black circle.



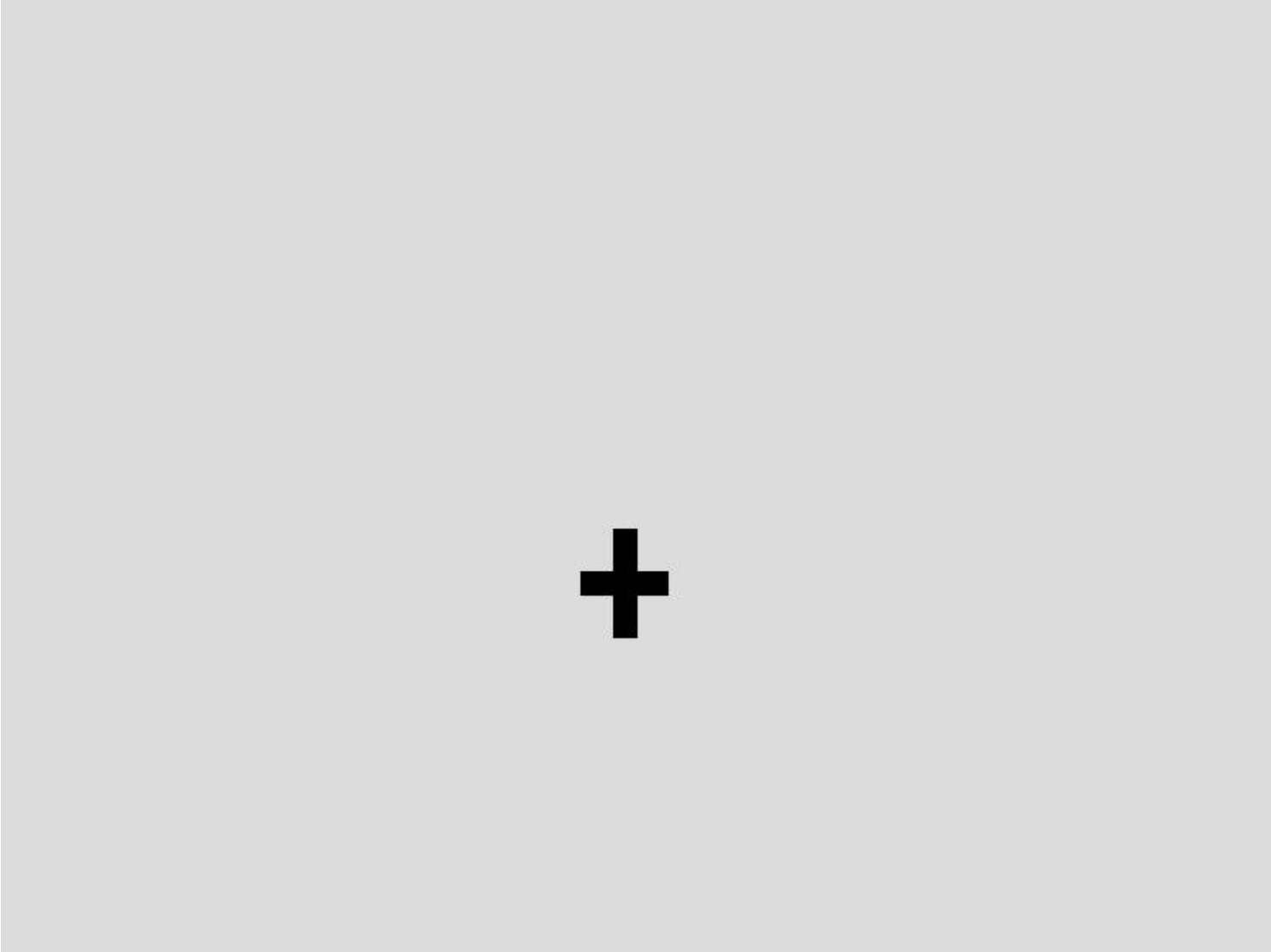
Conjunction search: Find the orange square.



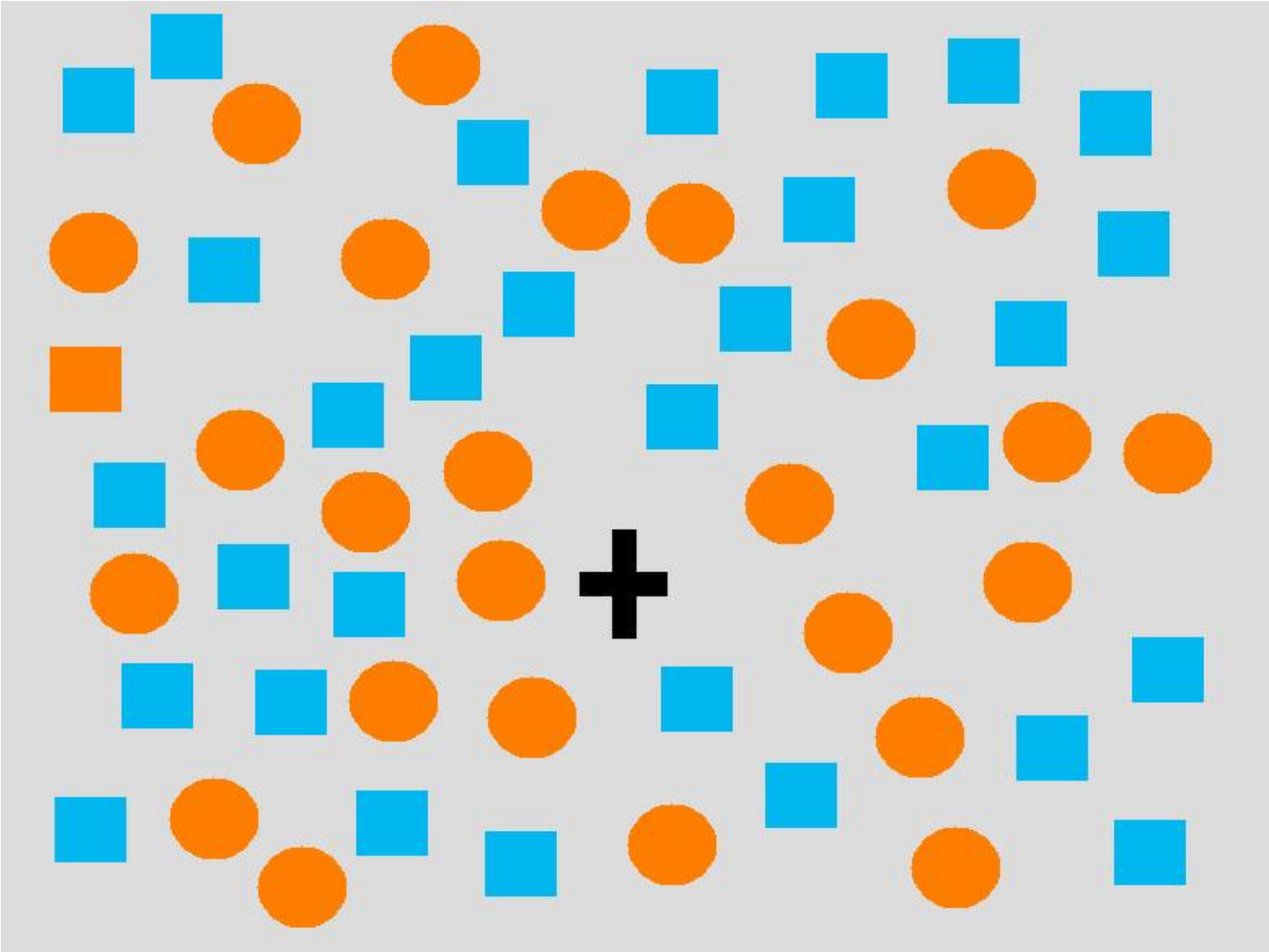
Conjunction search: Find the orange square.



Conjunction search: Find the orange square.



Conjunction search: Find the orange square.

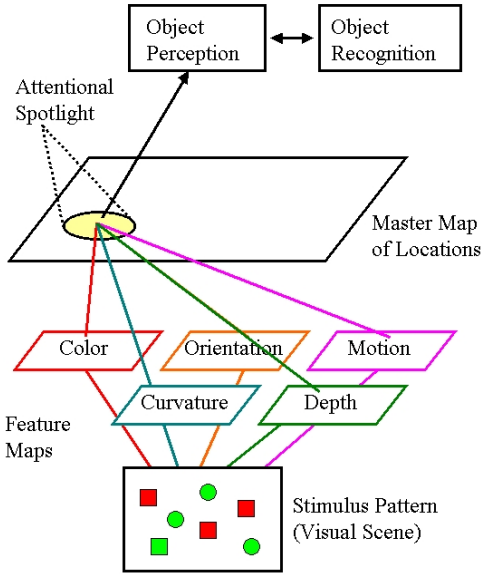


Theories of preattentive and attentive processing

How does preattentive processing occur within the visual system?

➟ Feature Integration Theory (Treisman and Gelade, 1980):

- A list of early or preattentive visual features (orientation, color, size, shape, length, width, curvature, density...);
- A model for the preattentive processing.



Massively parallel

Topographically organized Feature map related to a specific features

Master Map of Locations combination of feature maps

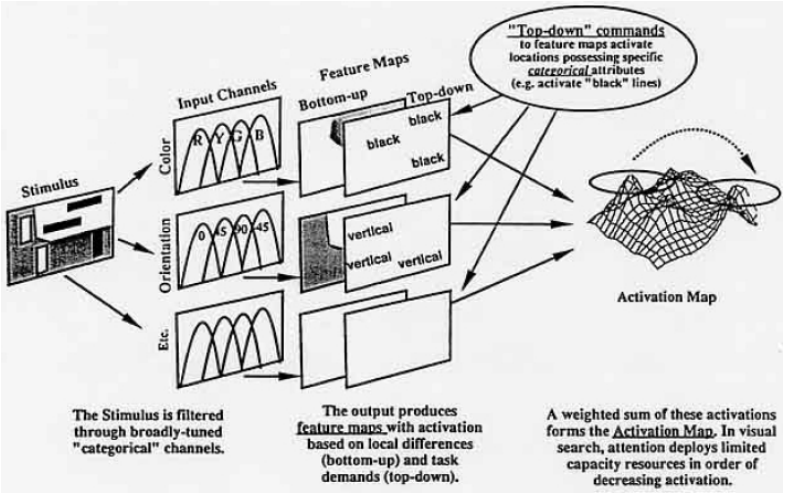
This influential work suffers from several drawbacks:

- Parallel conjunction search (!) (motion and shape);
- Grouping (principles of Gestalt psychology).

Theories of preattentive and attentive processing

- Guided search theory (Wolfe, 1994):
  - A model for the preattentive processing;
  - The activation map is a combination of bottom-up and top-down activation. The weights assigned to these two values are task dependent.

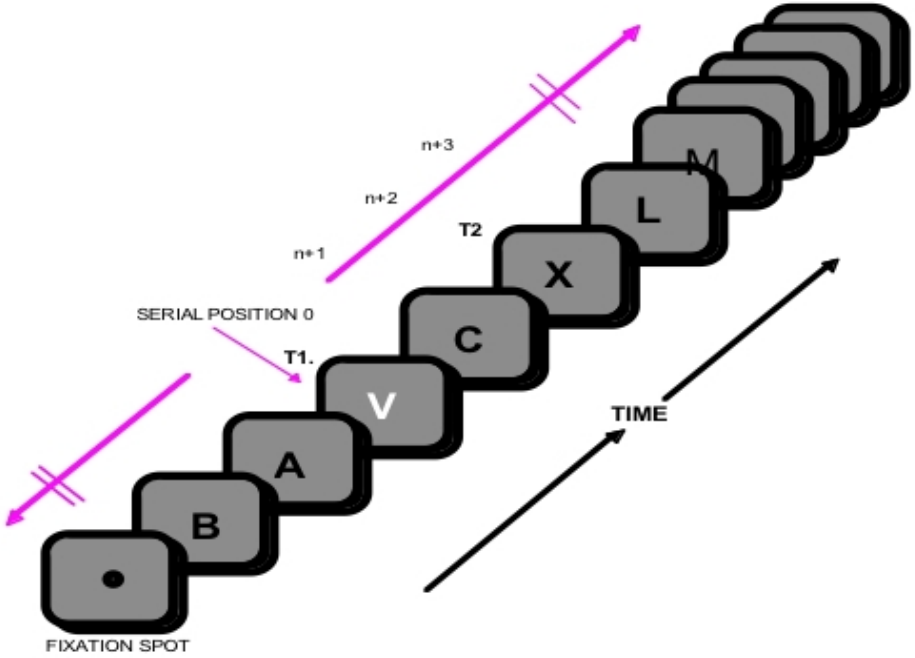
For a feature search → very small weigth on the top-down contribution
For a conjunction search → very small weigth on the bottom-up contribution



The attentional blink

The attentional blink refers to the fact that perception of a second target item is greatly reduced if it is presented within a half second of a first target item (Raymond et al., 1992).

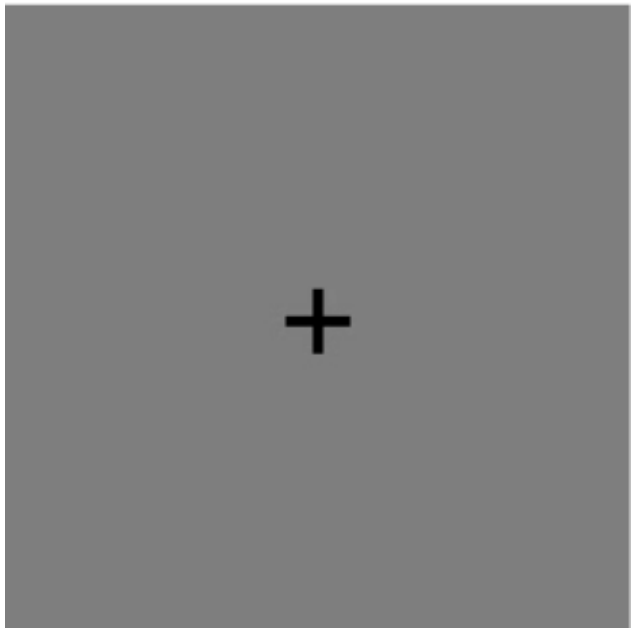
Rapid Serial Visual Presentation (RSVP)

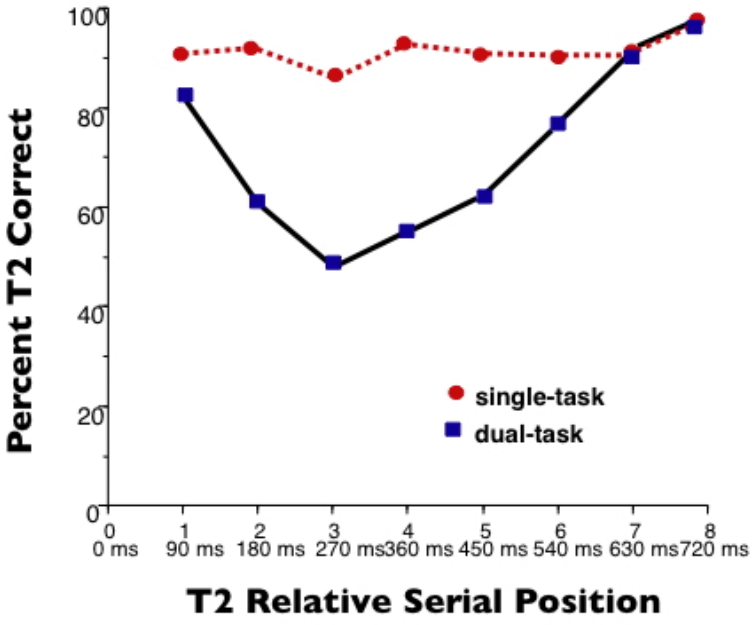


Raymond et al.'s experiment (From wikipedia).

- First target T1 (white letter) in the 10-item per second RSVP stream of black letters.
- Second target T2 (X) occuring after T1 (100 - 800 ms).

The attentional blink





A feedforward pathway

Visual input can be classified rapidly!!!

Ultra-rapid categoriation task (Thorpe et al., 1996) (briefly flashed image belonging or not to the target category):

- Subjects can respond in under 400 ms;
- ERP (Event-Related Potentials) studies have shown that the underlying processing can be done in less than 150 ms.

A feedforward pathway

Based on ultra-rapid categorization task, some conclusions can be drawn:

- It might indicate the existence of a feedforward pathway going through the visual system.
- The foveal vision is not required;

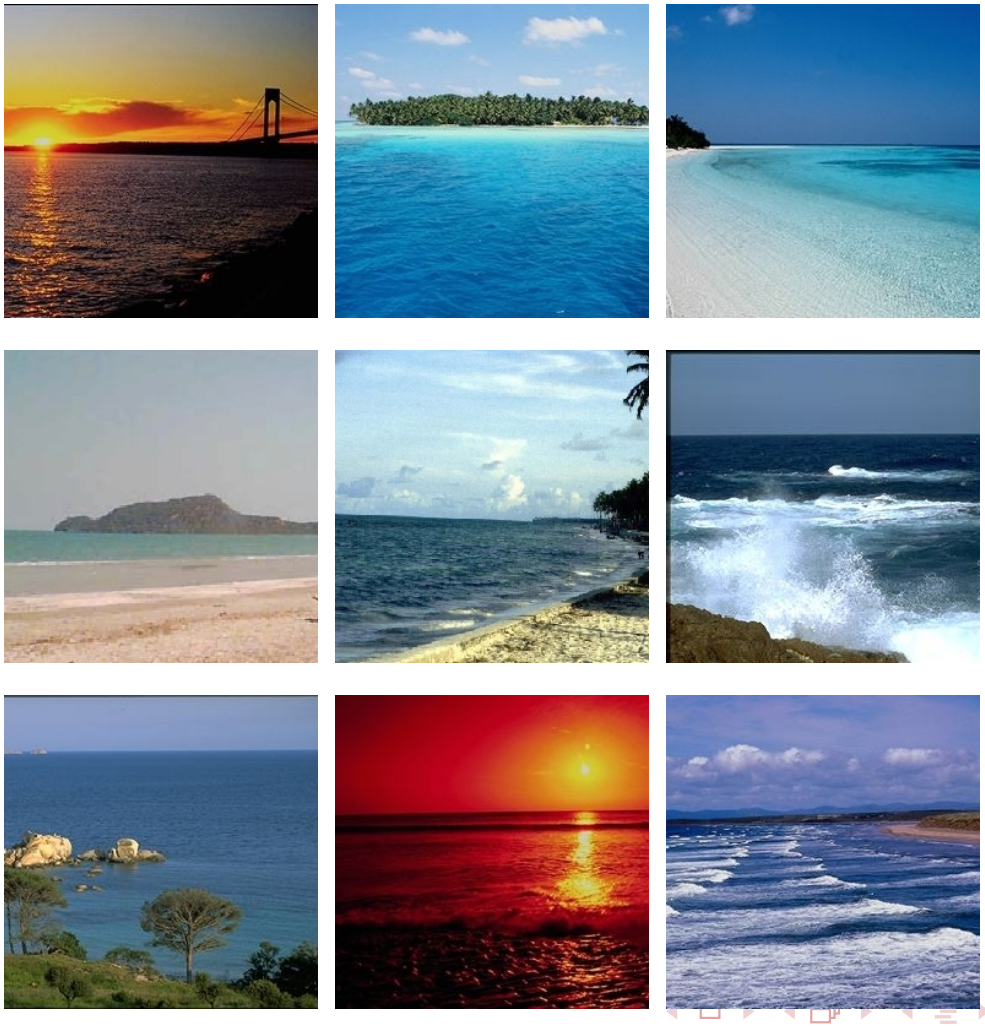
+

- The color information is not necessary (major role of magnocellular pathways);
- Intensive training doesn't modify the reaction times.

See publications on http://www.cerco.ups-tlse.fr/fr_vers/annuaire/simon_thorpe.html

A feedforward pathway

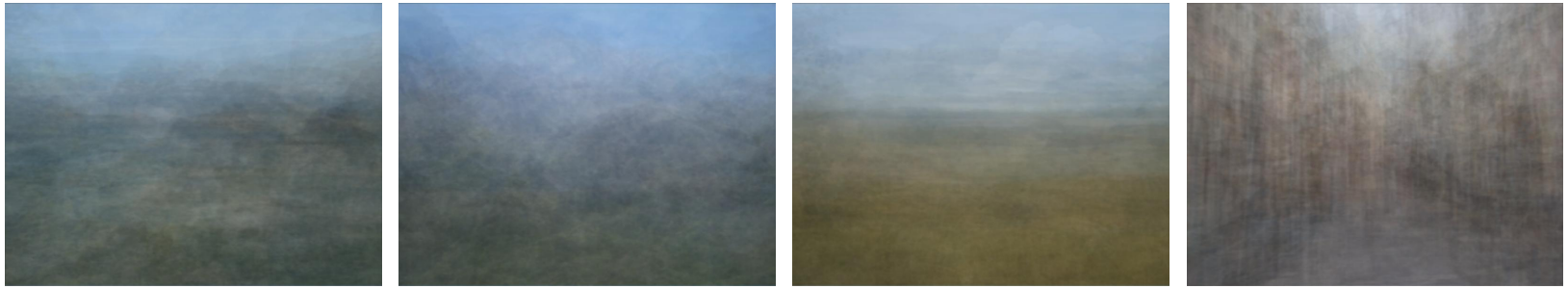
How can a picture be so rapidly recognized despite its variability?



The gist

The gist is an holistic representation of a scene (often compared to a summary):

- Conceptual gist: semantic information inferred from the scene (verbal description related to inferred and perceived information);
- Perceptual gist : structural representation mainly based on bottom-up features.



(a) Coast (b) Mountain (c) OpenCountry (d) Street

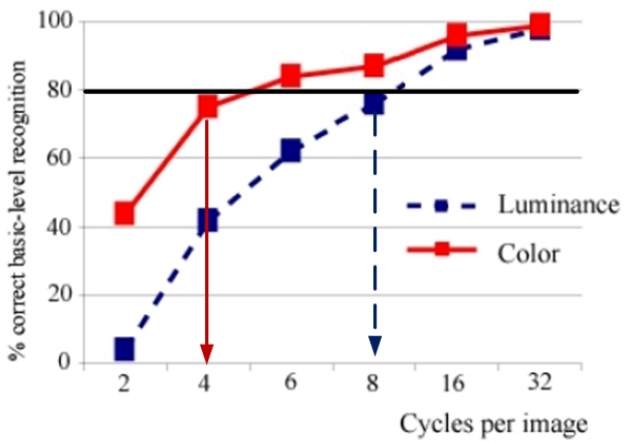
The goal of the gist might be to deliver structural summary in order:

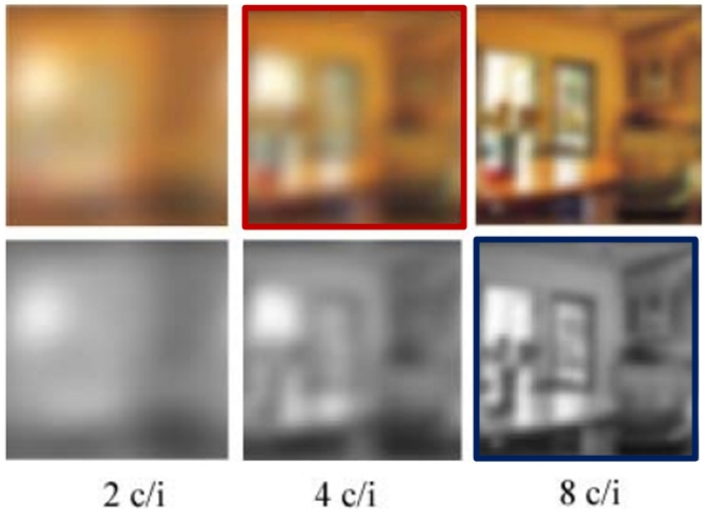
- to identify the scene, its semantic category;
- to get the spatial layout (a coarse version).

A coarse-to-fine process

A COARSE PROCESS

The ultra-rapid categorization might rely on coarse information as suggested by (Oliva and Schyns, 1994, 2000):





- A coarse representation is first used in order to get a rough estimate of the input (Oliva and Schyns, 1994);
- Color influences recognition when it is diagnostic of a scene category (Oliva and Schyns, 2000).

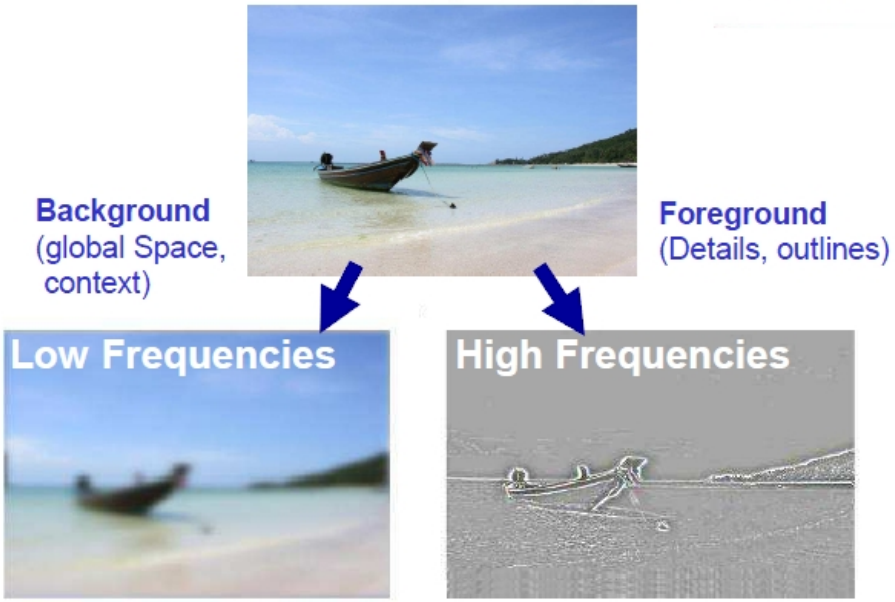
A coarse-to-fine process

BUT ALSO A FINE PROCESS

Edges and objects are also important to recognize a scene.

binding together local contours providing a refinement or refutation of the first decision.

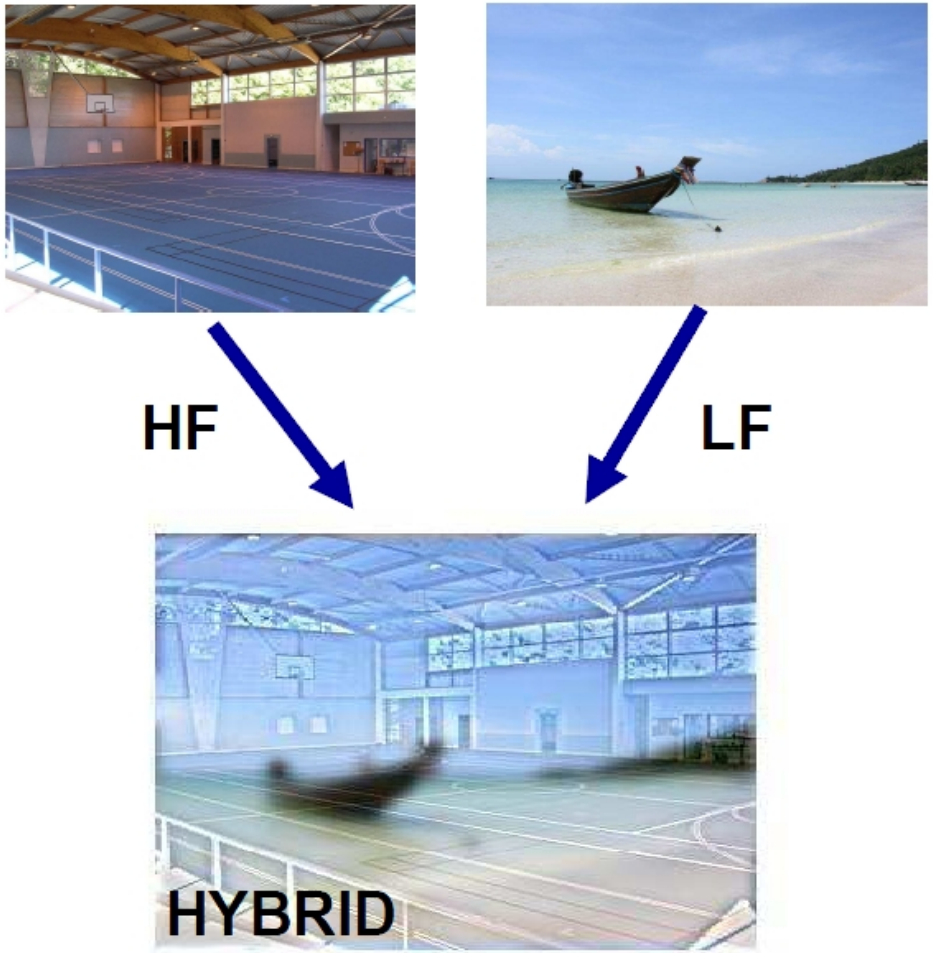
(Oliva and Schyns, 1994) proposed an interesting paradigm to investigate the role of low and high spatial frequencies: Hybrid pictures.



From (Follet et al., 2009).

A coarse-to-fine process

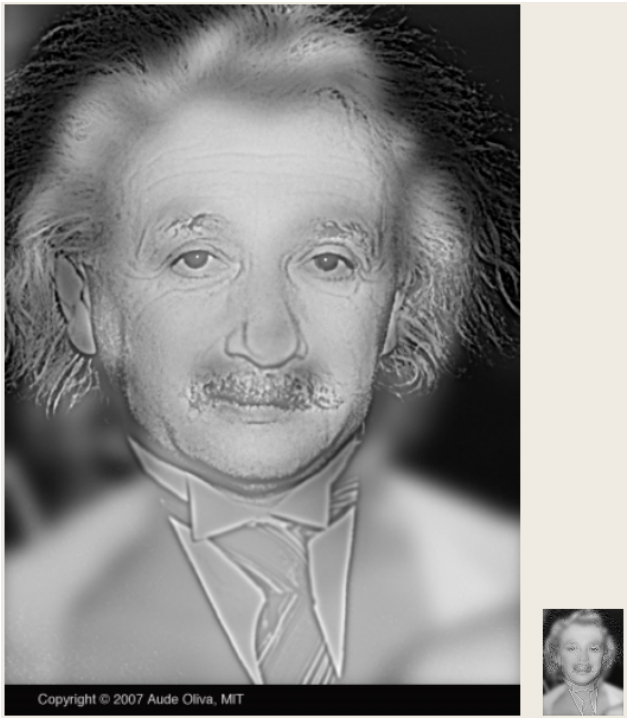
Hybrid pictures



From (Follet et al., 2009).

A coarse-to-fine process

Hybrid pictures



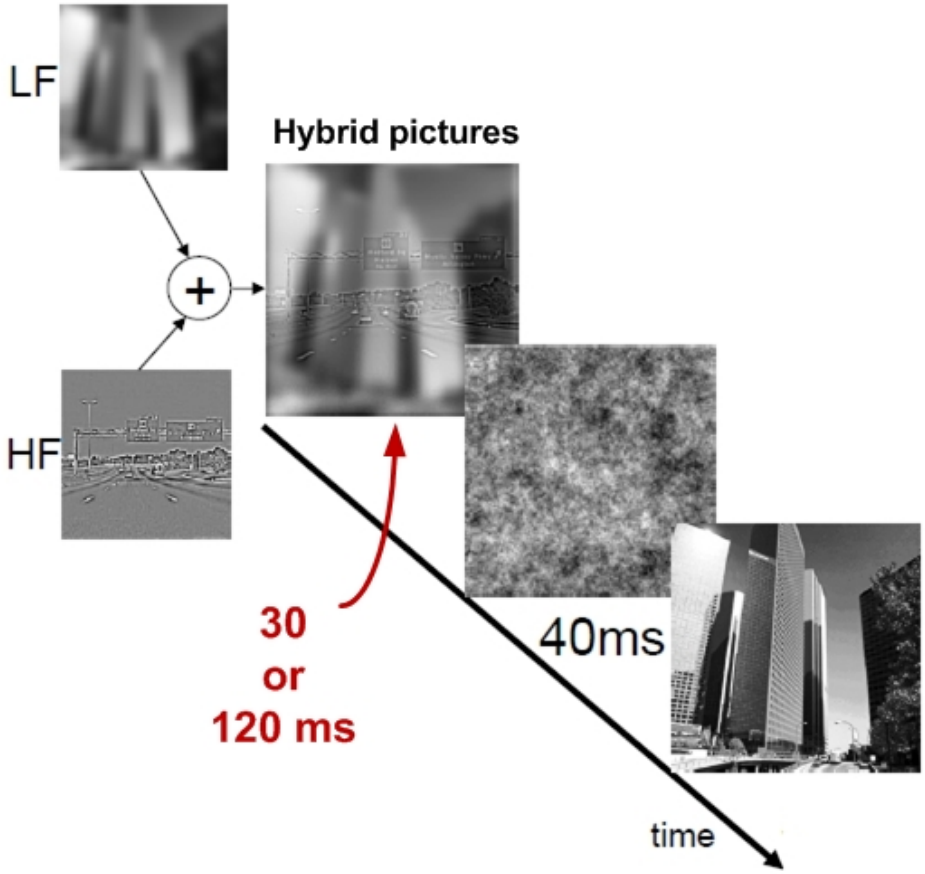
From http://cvcl.mit.edu/hybridimage.htm.

A coarse-to-fine process

In (Oliva and Schyns, 1994), the goal was to make a priming with an hybrid picture. The presentation time was either equal to 30 or 120ms.

Detection task:

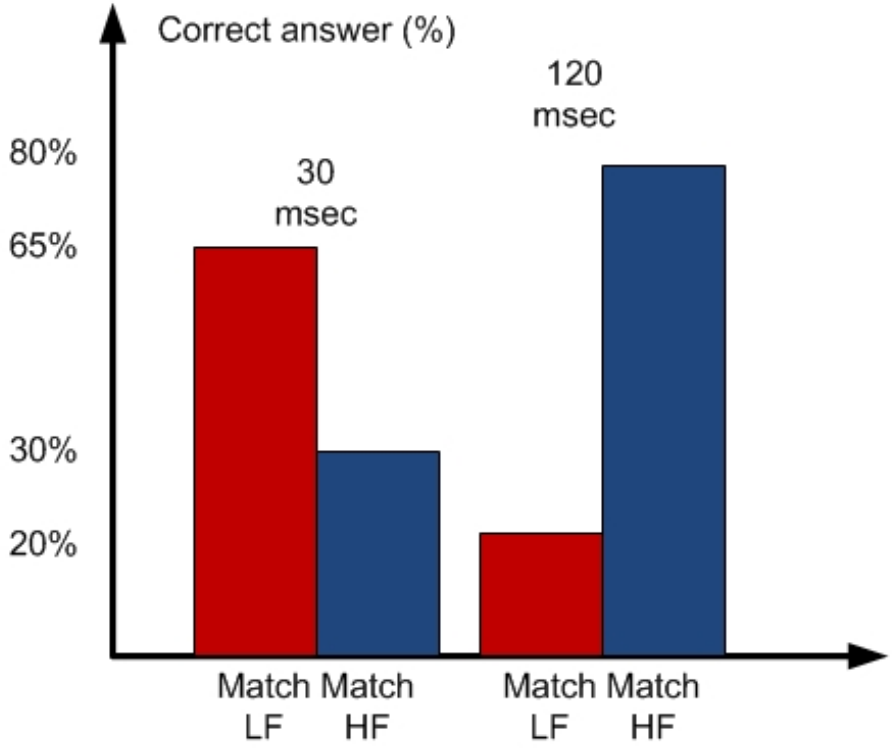
- New picture;
- Low-frequency (for the priming);
- High-frequency (for the priming).



A coarse-to-fine process

Detection task:

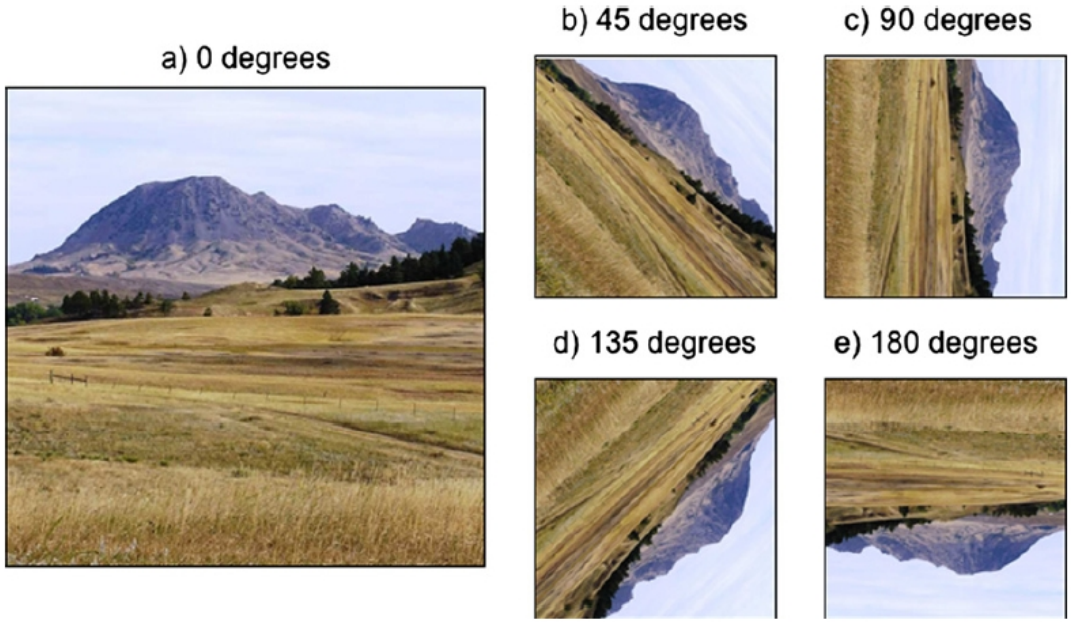
- New picture;
- Low-frequency (for the priming);
- High-frequency (for the priming).



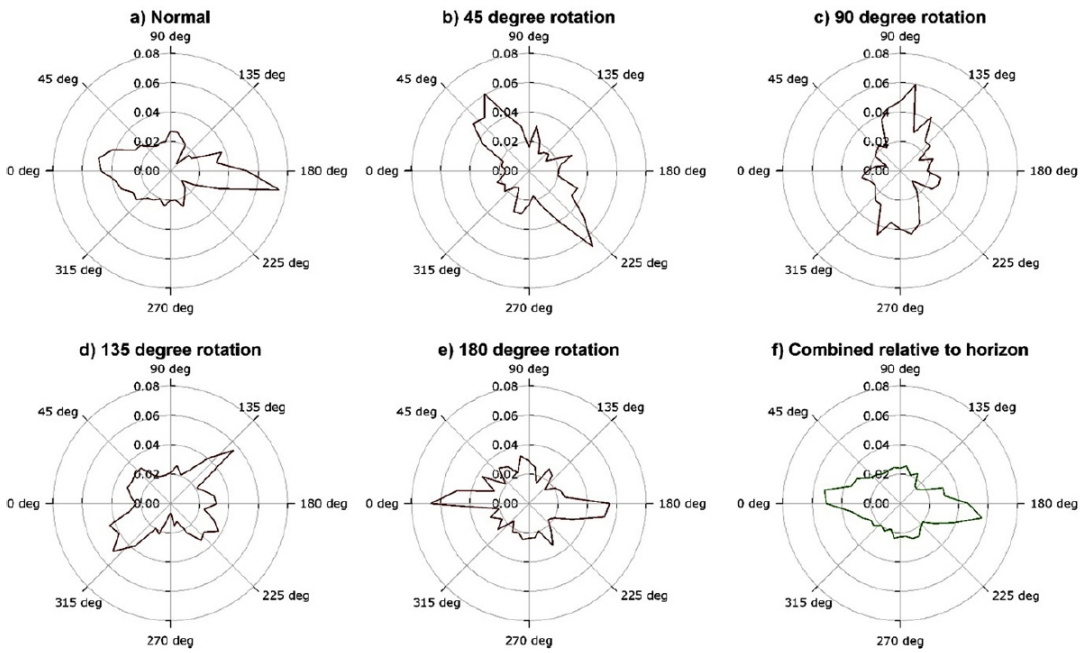
Horizon line

Turning the world around: Patterns in saccade direction vary with picture orientation (Foulsham et al., 2008)

- Participants were instructed to take in as much information as possible, and each picture was followed by a written sentence presented on the screen.



Horizon line



Whichever way the picture was oriented there were more saccades in the axes corresponding to the picture's (horizontal) orientation than elsewhere.

A strong systematic tendency for saccades to occur along the axis of the natural horizon.

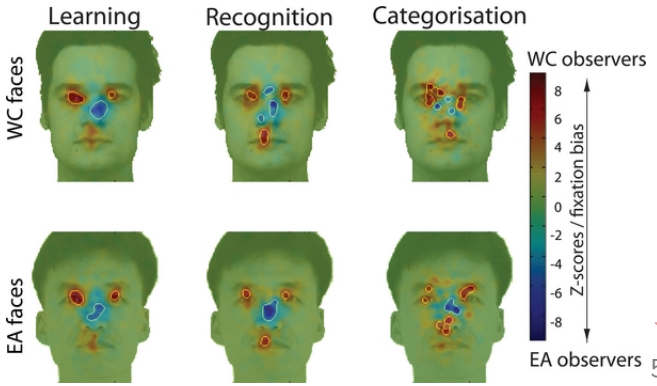
Face detection

Human faces play an important role in the deployment of our visual attention:

- there are neurons in the monkey temporal visual cortex that are sensitive to faces;
- responses are size and position invariant.

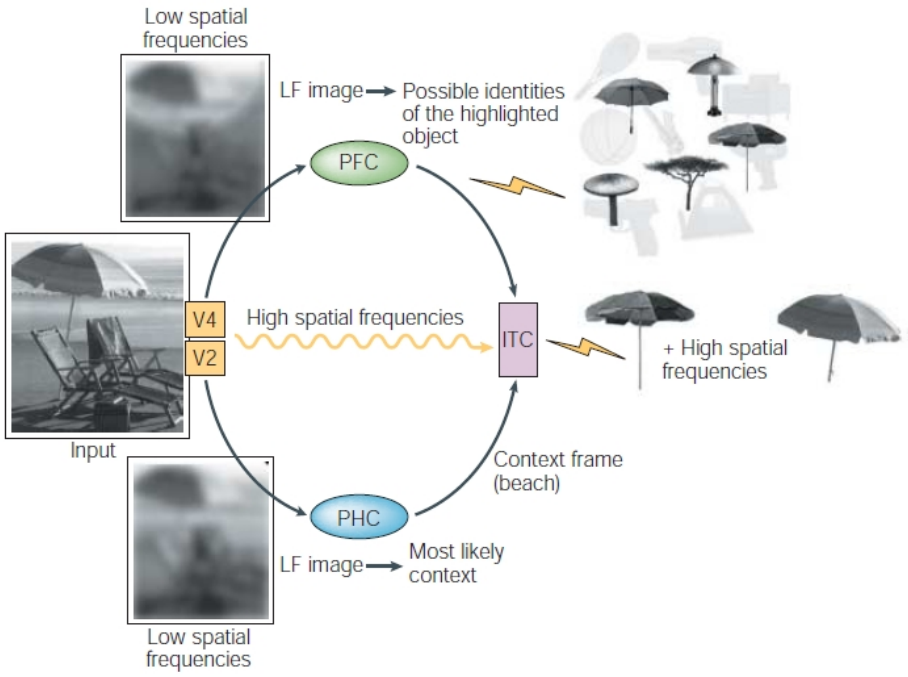
From as early as 1965, studies of eye movements have consistently revealed a systematic triangular sequence of fixations over the eyes and the mouth, suggesting that faces elicit a universal, biologically-determined information extraction pattern.
Based on observations with adults from Western cultures only....
(Blais et al., 2008)

Conclusion in (Blais et al., 2008): The strategy employed to extract visual information from faces differs across cultures. Western Caucasian and East Asian observers differed in how they extracted facial information using eye movements.



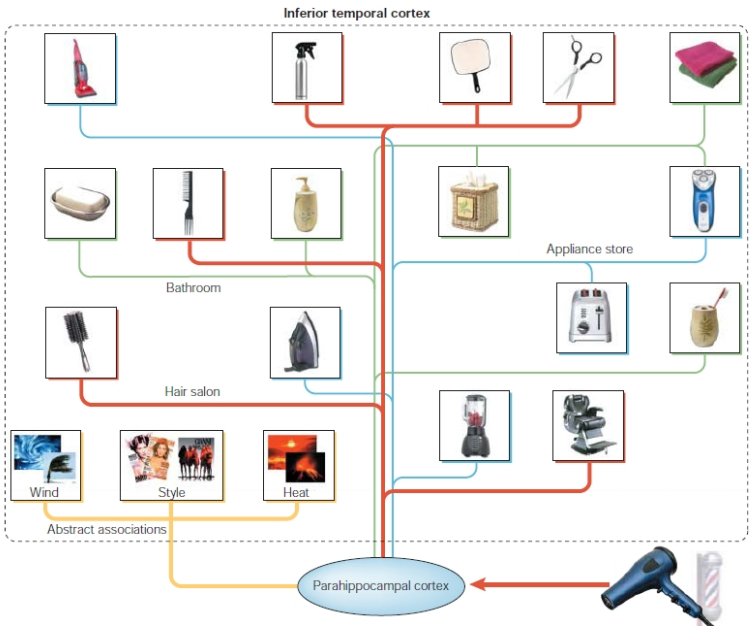
Cognitive models

(Bar, 2004)'s model



1. Global information is conveyed by low spatial frequencies which are used to define a set of probable candidate interpretations of the input: Context Frame;
2. Exact representation and interpretation are subsequently derived from the later arrival of higher spatial frequencies.

(Bar, 2004)'s model



A particular object (for example, a hairdryer) can be associated with several context frames (for example, a hair salon, a bathroom, an appliance store), as well as with abstract concepts (heat, style and wind). The specific context (for example, hair salon), which is established by global scene information and/or by contextual cues that are provided by the recognition of other objects in the scene (for example, a barber pole), dictates which of the experience-based association sets should be activated to facilitate recognition in the particular situation.

(Bar, 2004)'s model

Context frame can be compared to Minsky's frame.

A Minsky's frame is used for knowledge representation and contains several kinds of information:

- Some of this information is about how to use the frame;
- Some is about what one can expect to happen next;
- Some is about what to do if these expectations are not confirmed.

For visual scene analysis, the different frames of a system describe the scene from different viewpoints, and the transformations between one frame and another represent the effects of moving from place to place. Frames also contain a set of probable objects.

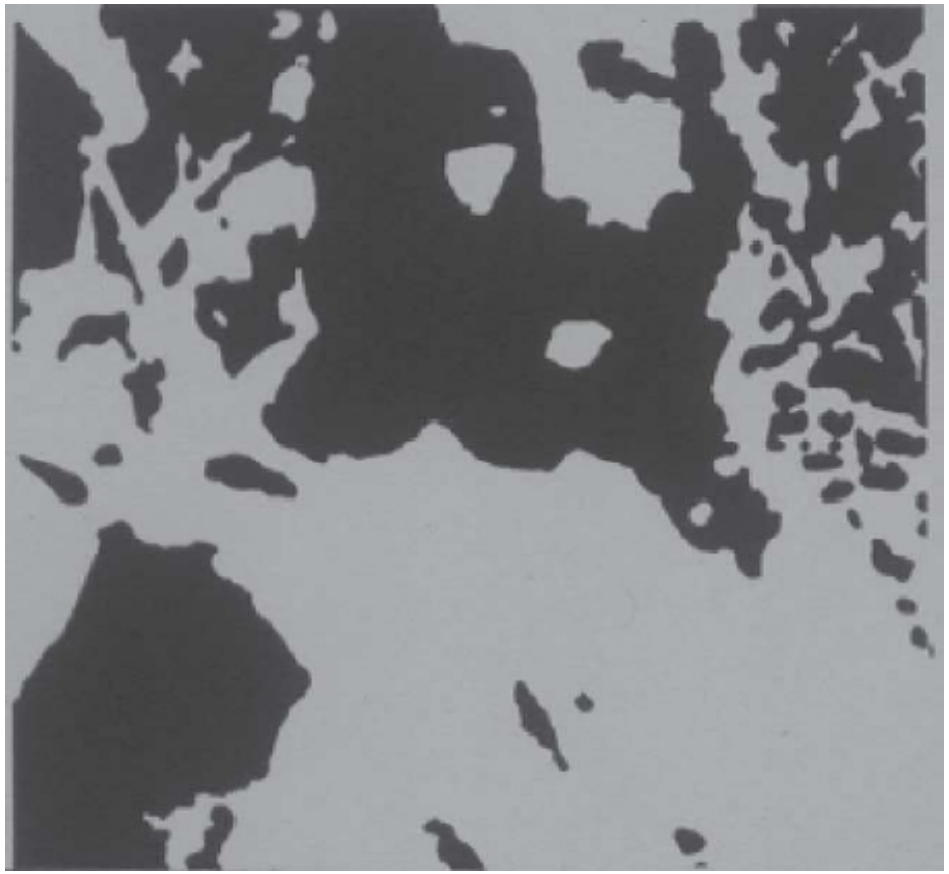
Reverse Hierarchical Theory (Ahissar and Hochstein, 2004)

Reverse Hierarchy Theory dissociates between early explicit perception and implicit low-level vision.

- Implicit low-level vision based on increasingly complex representations in the feedforward pathways;
- Early explicit perception: perception at a glance (high-level information, categorical scene interpretation, identifying *forest before trees*).

Reverse Hierarchical Theory (Ahissar and Hochstein, 2004)

Eureka effect



A set of gray and black regions, without further meaning?

Reverse Hierarchical Theory (Ahissar and Hochstein, 2004)

Eureka effect

Figure 1. A hint for 'understanding' the image in Box 1. A few lines draw attention to those aspects that are essential for perceiving a bearded figure. Now return to Box 1 on the preceding page to see how this 'clue' has affected your perceptual system, perhaps forever!

Computational models

What is saliency?

Salience phenomena put an element forward whether the domain field (visual, linguistic, cognitive...).

The salience can be induced by :

- Top-down mechanisms (expectation, goal, prior and early inferences...);
- Bottom-up mechanisms (visual features, audio...). Bottom-up saliency is a signal-based factor that dictates where attention is to be focussed.

What is saliency?

Salience phenomena put an element forward whether the domain field (visual, linguistic, cognitive...).

The salience can be induced by :

- Top-down mechanisms (expectation, goal, prior and early inferences...);
- Bottom-up mechanisms (visual features, audio...). Bottom-up saliency is a signal-based factor that dictates where attention is to be focussed.

Purposes of saliency/attention are:

- Warning (animals (predator), sudden motion...);
- Exploration (in an efficient way);
- Inspection.

Computational models: two schools of thought

For the computational modelling, two 'schools' can be considered:

- One based on the assumption that there is an unique saliency map (Koch and Ullman, 1985):

Definition (saliency map)

A topographic representation that combines the information from the individual feature maps into one global measure of conspicuity. This map can be modulated by a higher-level feedback.

Computational models: two schools of thought

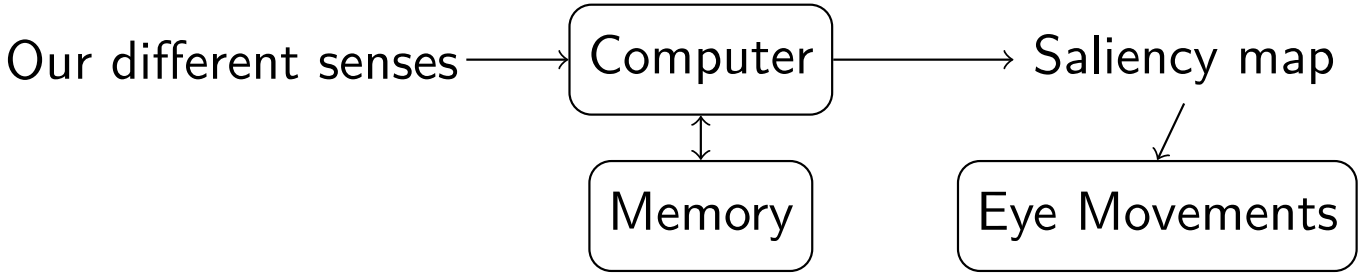
For the computational modelling, two 'schools' can be considered:

➟ One based on the assumption that there is an unique saliency map (Koch and Ullman, 1985):

Definition (saliency map)

A topographic representation that combines the information from the individual feature maps into one global measure of conspicuity. This map can be modulated by a higher-level feedback.

A comfortable view for the computational modelling...



Le cerveau n'est pas un ordinateur, F. Varela, 1998.

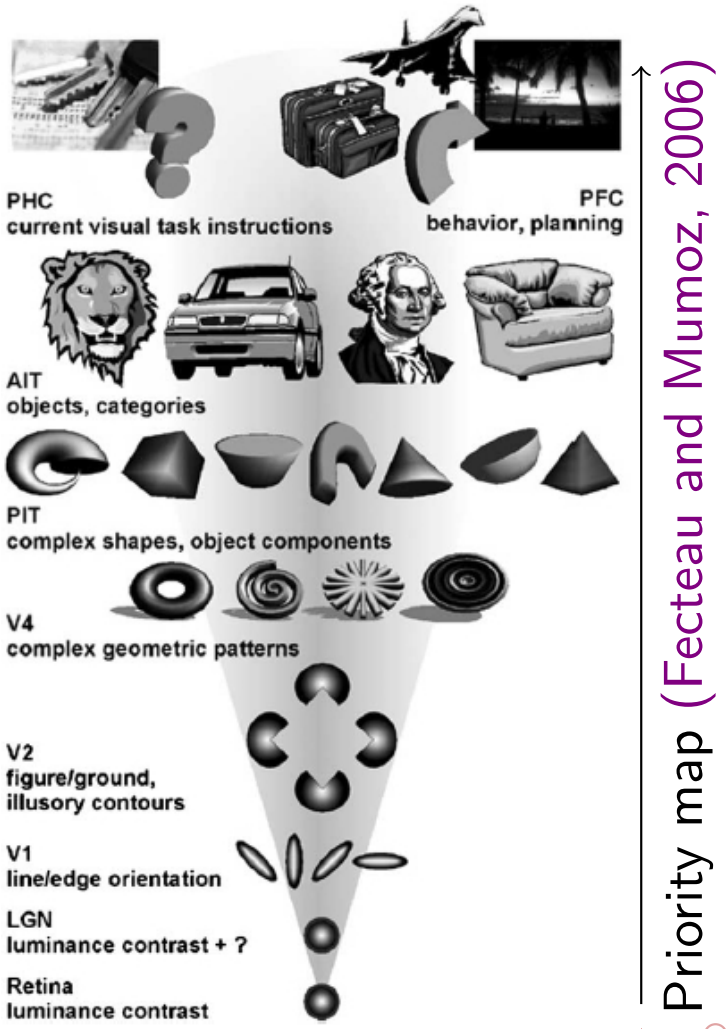
Computational models: two schools of thought

For the computational modelling, two 'schools' can be considered:

- There exist multiple saliency maps (distributed throughout the visual areas) (Tsotsos et al., 1995)).

  Many candidate locations for a saliency map:
  - Primary visual cortex
  - Lateral IntraParietal area (LIP)
  - Medial Temporal cortex

'*At each level, saliency can thus be used as a gain control mechanism to spatially gate relevant information* for the next processing level. From (Van Rullen, 2003).



From (Van Rullen, 2003).

Computational models of Bottom-up visual attention

What is our purpose...

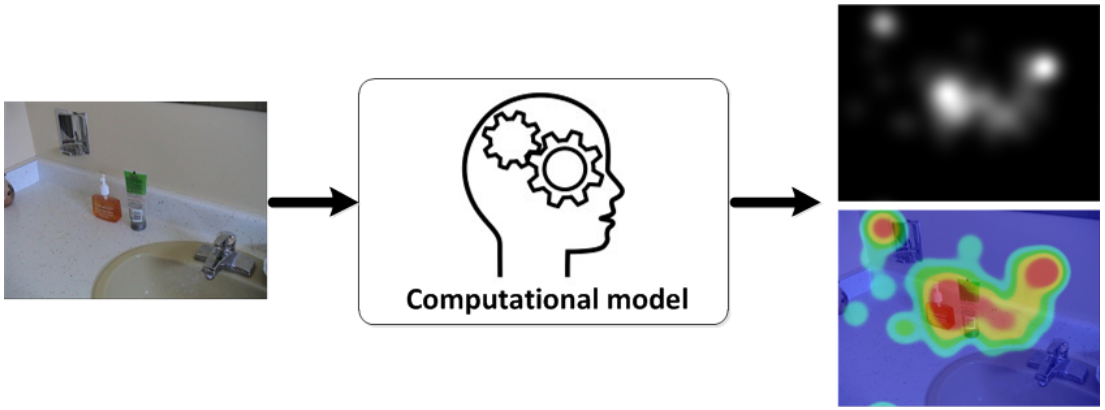
- Bottom-Up: some things draw attention reflexively, in a task-independent way (Involuntary, Very quick, Unconscious).

Computational models of Bottom-up visual attention

What is our purpose...

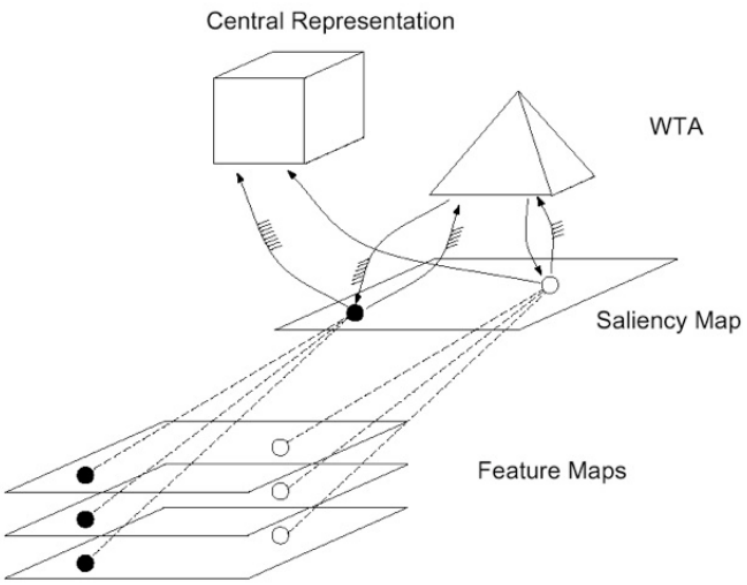
➟ Bottom-Up: some things draw attention reflexively, in a task-independent way (Involuntary, Very quick, Unconscious).

Computational models of Bottom-up overt visual attention



Computational models of Bottom-up visual attention

Most of the computation models of visual attention have been motivated by the seminal work of (Koch and Ullman, 1985).

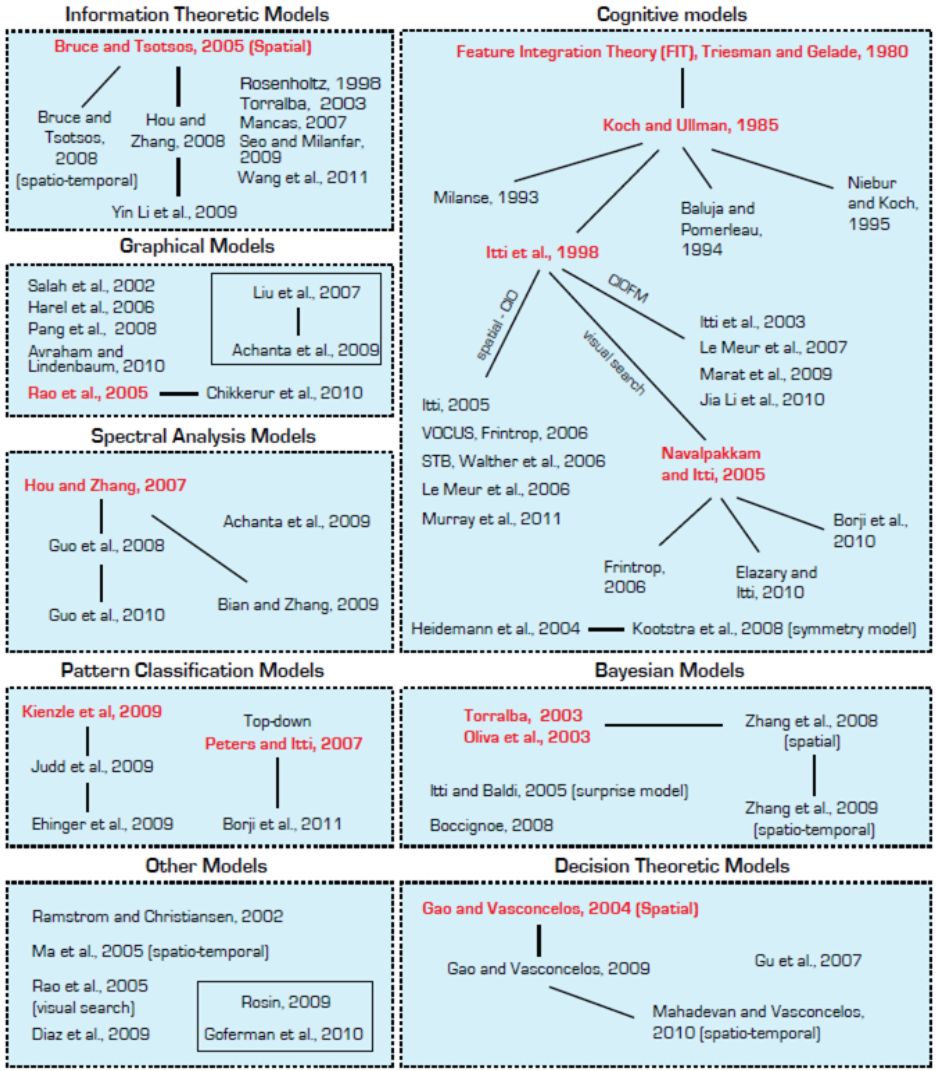


- a plausible computational architecture to predict our gaze;
- a set of feature maps processed in a massively parallel manner;
- a single topographic saliency map.

Computational models of Bottom-up visual attention

Taxonomy of models:

- Information Theoretic models;
- Cognitive models;
- Graphical models;
- Spectral analysis models;
- Pattern classification models;
- Bayesian models.



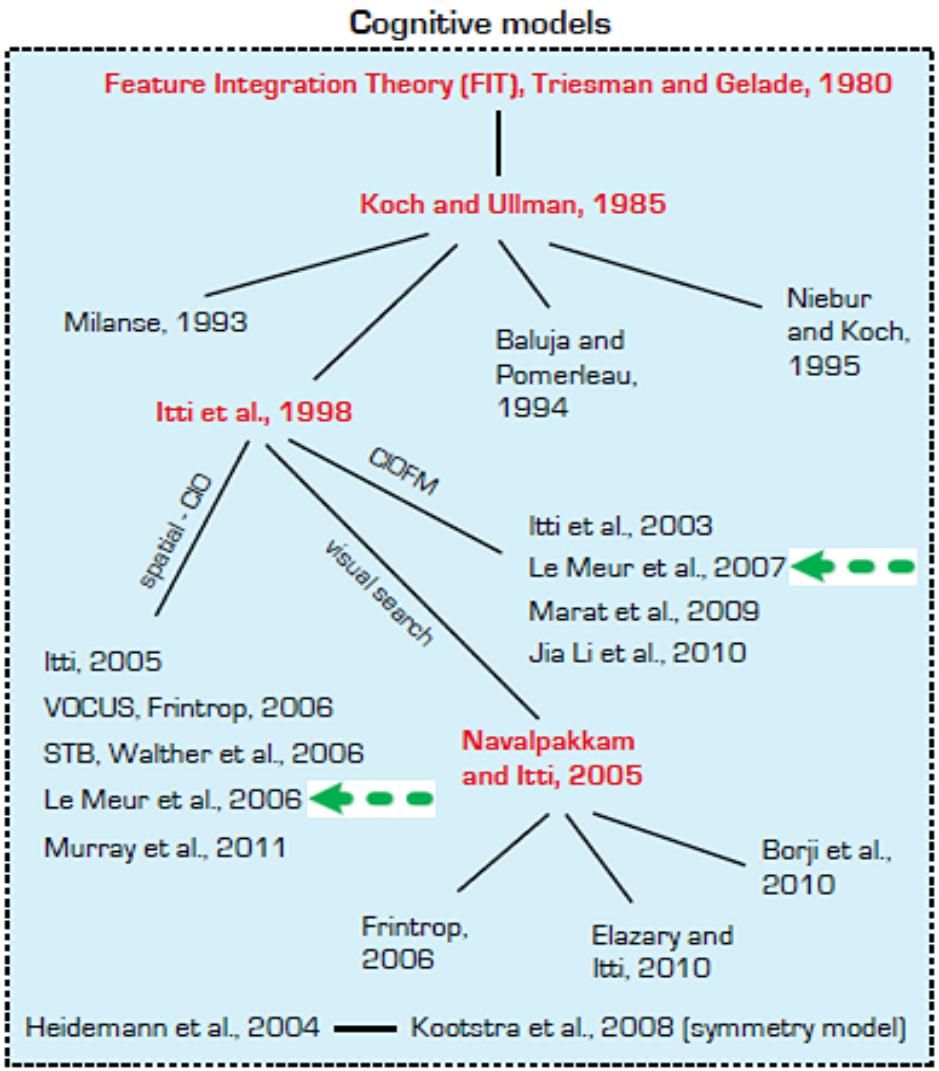
Extracted from (Borji and Itti, 2013).

Computational models of Bottom-up visual attention

Cognitive models:

as faithful as possible to the Human Visual System (HVS)

- inspired by cognitive concepts;
- based on the HVS properties.

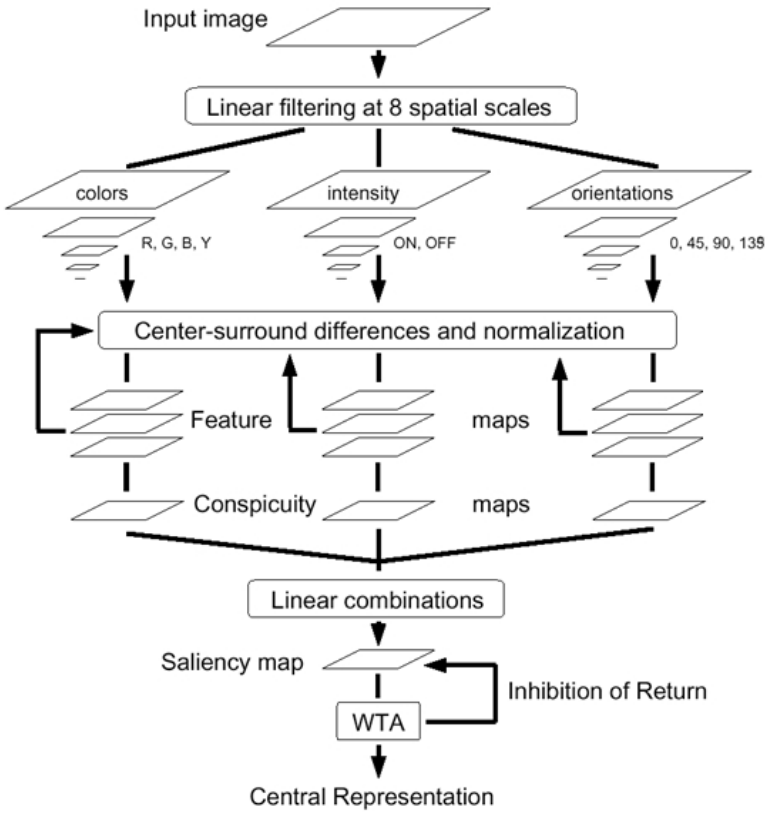


Extracted from (Borji and Itti, 2013).

Cognitive models

Itti's model (Itti et al., 1998), probably the most known...

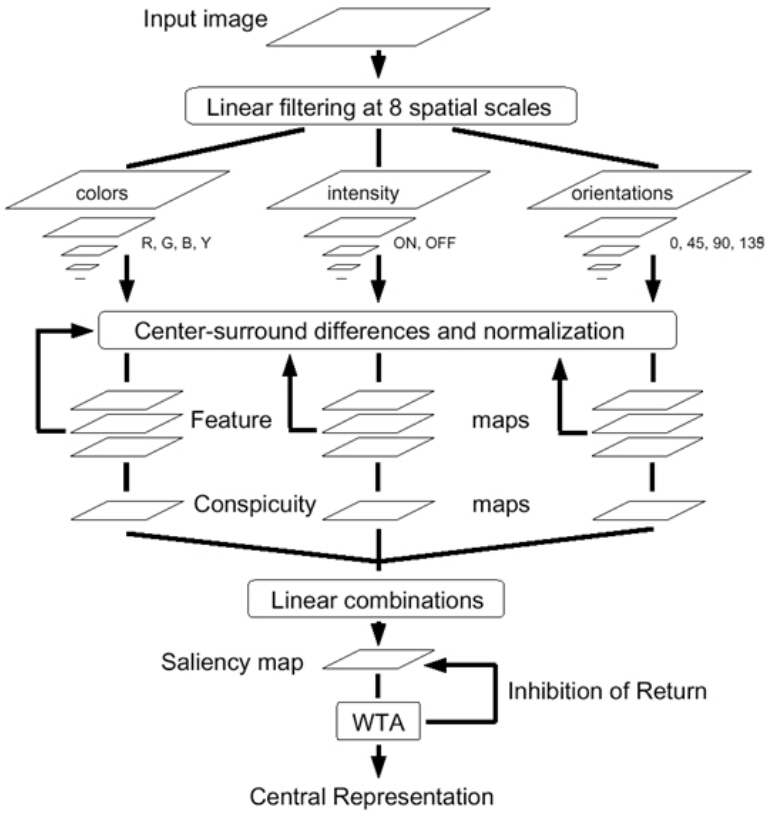
- Based on the Koch and Ullman's scheme;
- Hierarchical decomposition (Gaussian);
- Early visual features extraction in a massively parallel manner;
- Center-surround operations;
- Pooling of the feature maps to form the saliency map.



Cognitive models

Itti's model (Itti et al., 1998), probably the most known...

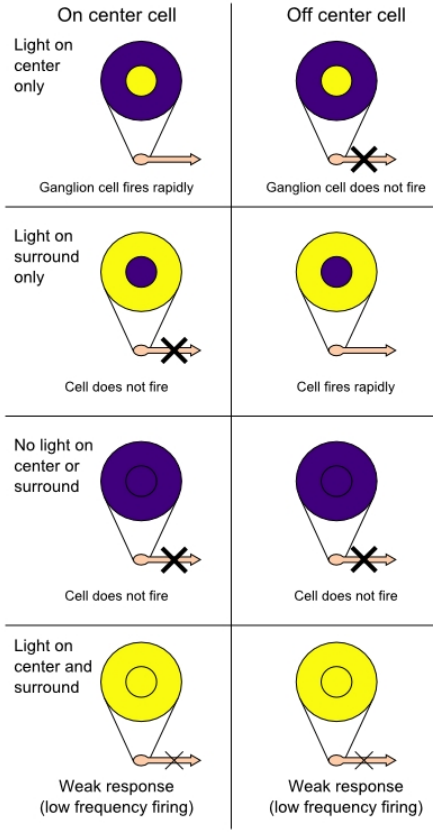
- Hierarchical decomposition (Gaussian):



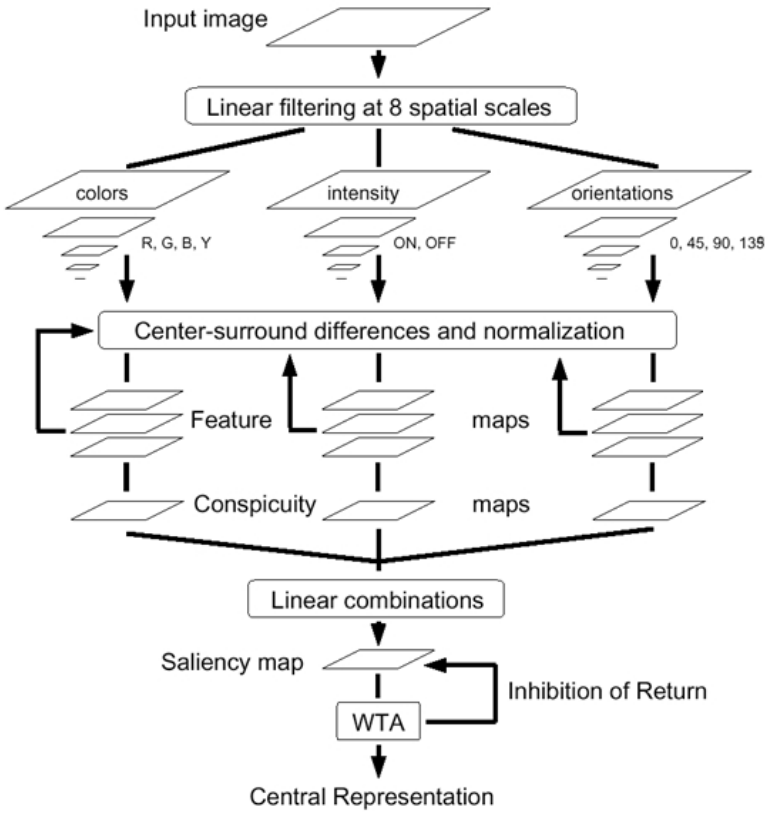
Cognitive models

Itti's model (Itti et al., 1998), probably the most known...

➟ Center-surround operations:



From Wikipedia.

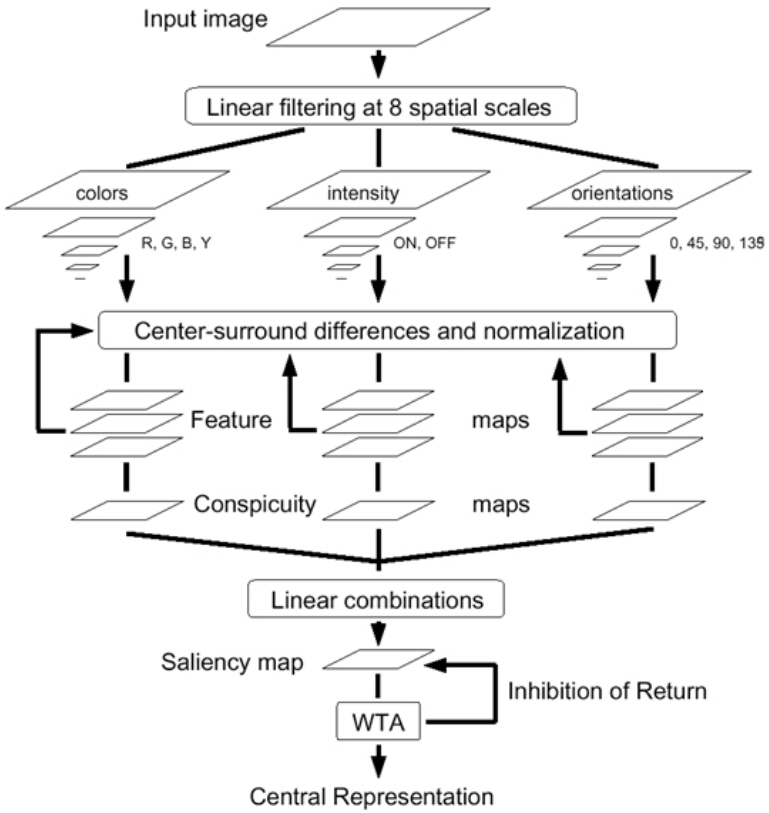


Cognitive models

Itti's model (Itti et al., 1998), probably the most known...

➡ Normalization and combination:

- **Naive summation**: all maps are normalized to the same dynamic and averaged;
- **Learning linear combination**: depending on the target, each feature map is globally multiplied by a weighting factor.
- **Content-based global non-linear amplification**:
  1. Normalize all the feature maps to the same dynamic range;
  2. For each map, find its global maximum M and the average $\overline{m}$ of all the other local maxima;
  3. Globally multiply the map by $(M - \overline{m})^2$.



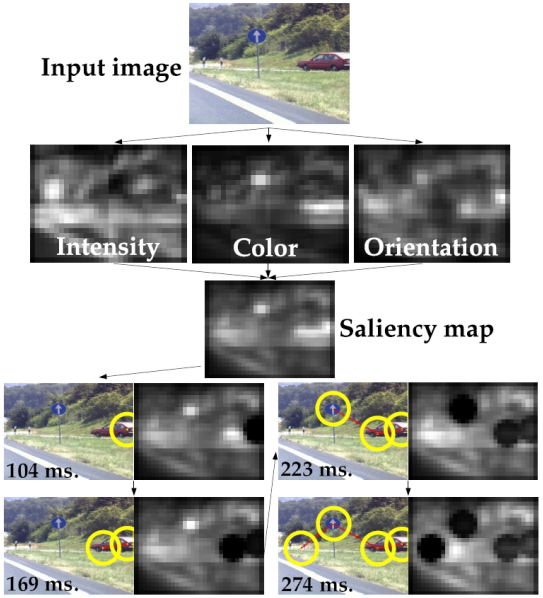
Cognitive models

Itti's model (Itti et al., 1998), probably the most known...

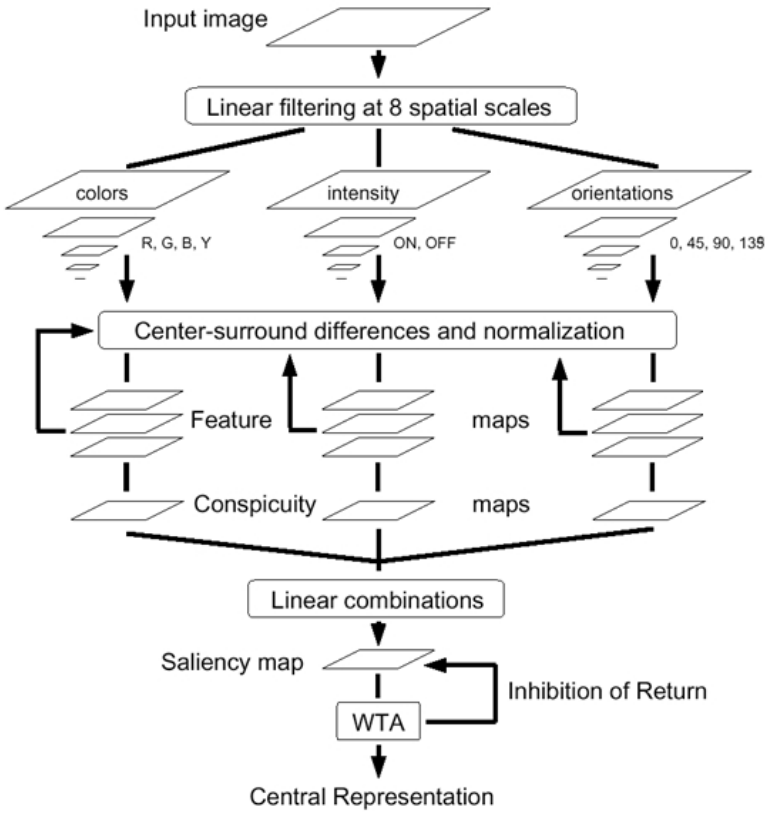
- Winner-Take-All:

  The maximum of the saliency map ⇒ the most salient stimulus ⇒ focus of attention

  Inhibitory feedback from WTA



From Itti's Thesis.



Cognitive models

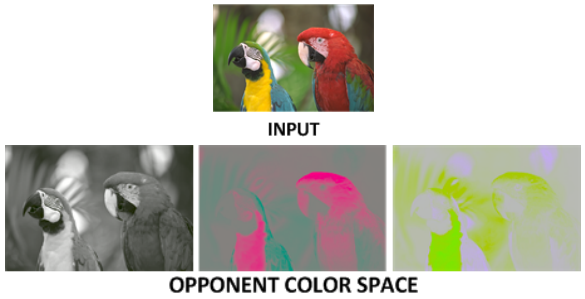
In (Le Meur et al., 2006), we designed a computational model of bottom-up visual attention.

INPUT

1. Input color image;
2. Projection into a perceptual color space;
3. Subband decomposition in the Fourier domain;
4. CSF and Visual Masking;
5. Difference of Gaussians;
6. Pooling.

Cognitive models

In (Le Meur et al., 2006), we designed a computational model of bottom-up visual attention.

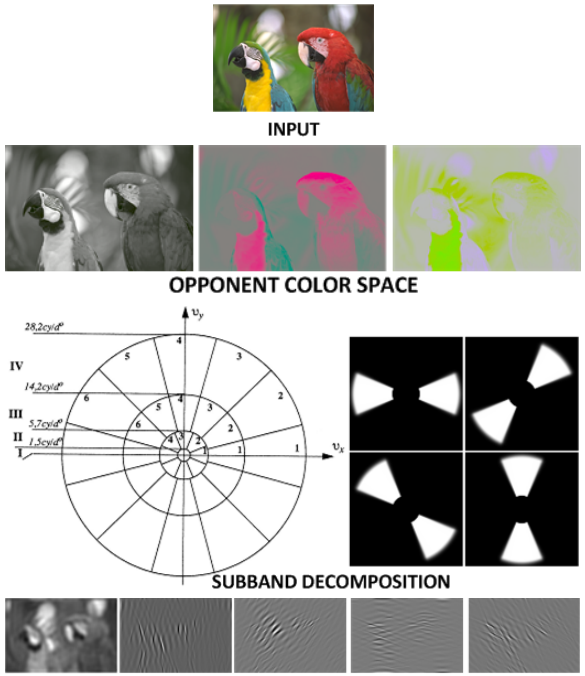


1. Input color image;
2. Projection into a perceptual color space;
3. Subband decomposition in the Fourier domain;
4. CSF and Visual Masking;
5. Difference of Gaussians;
6. Pooling.

Cognitive models

In (Le Meur et al., 2006), we designed a computational model of bottom-up visual attention.

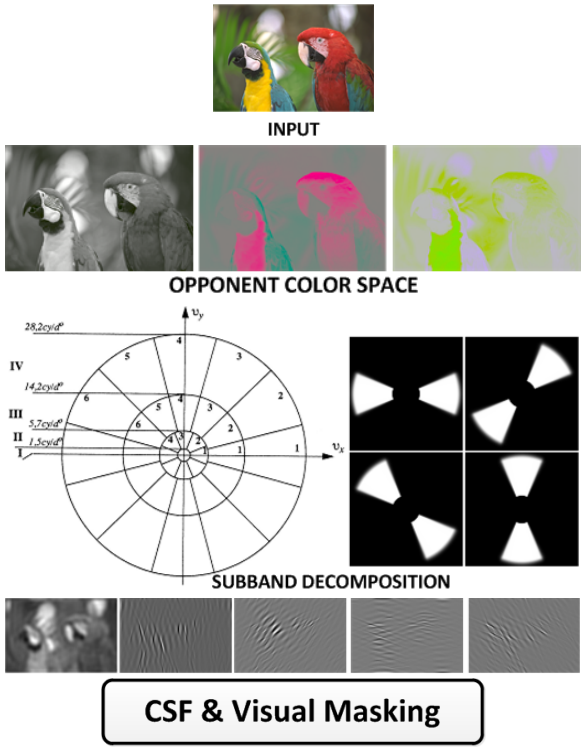
1. Input color image;
2. Projection into a perceptual color space;
3. Subband decomposition in the Fourier domain;
4. CSF and Visual Masking;
5. Difference of Gaussians;
6. Pooling.



Cognitive models

In (Le Meur et al., 2006), we designed a computational model of bottom-up visual attention.

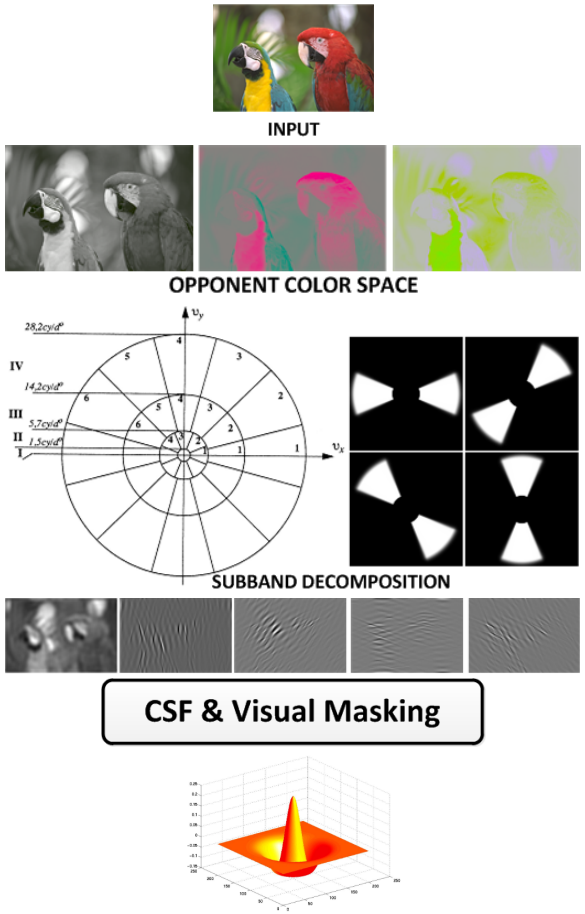
1. Input color image;
2. Projection into a perceptual color space;
3. Subband decomposition in the Fourier domain;
4. CSF and Visual Masking;
5. Difference of Gaussians;
6. Pooling.



Cognitive models

In (Le Meur et al., 2006), we designed a computational model of bottom-up visual attention.

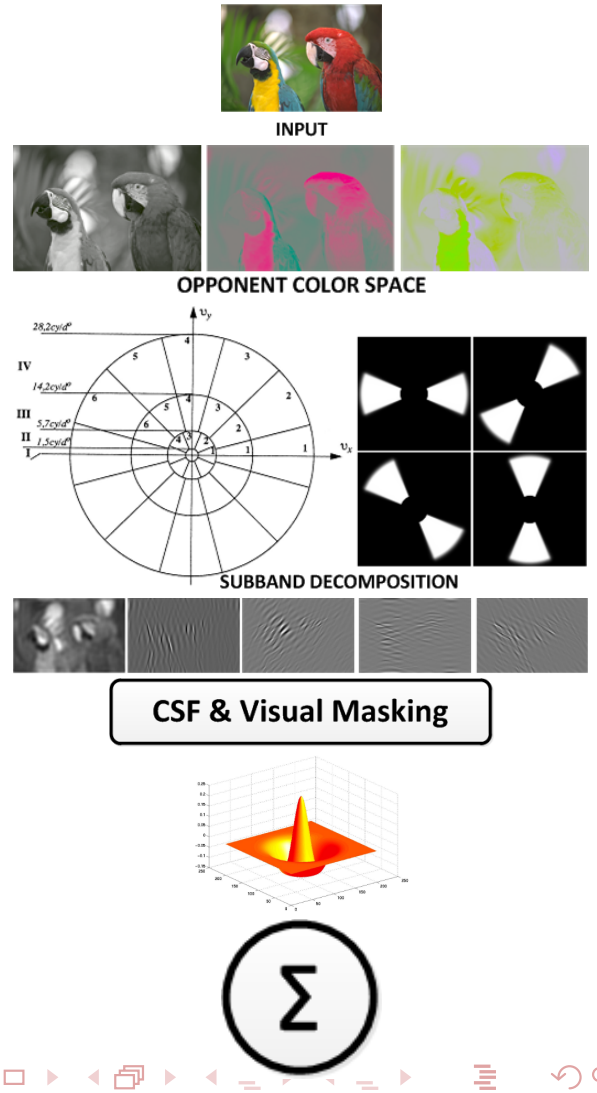
1. Input color image;
2. Projection into a perceptual color space;
3. Subband decomposition in the Fourier domain;
4. CSF and Visual Masking;
5. Difference of Gaussians;
6. Pooling.



Cognitive models

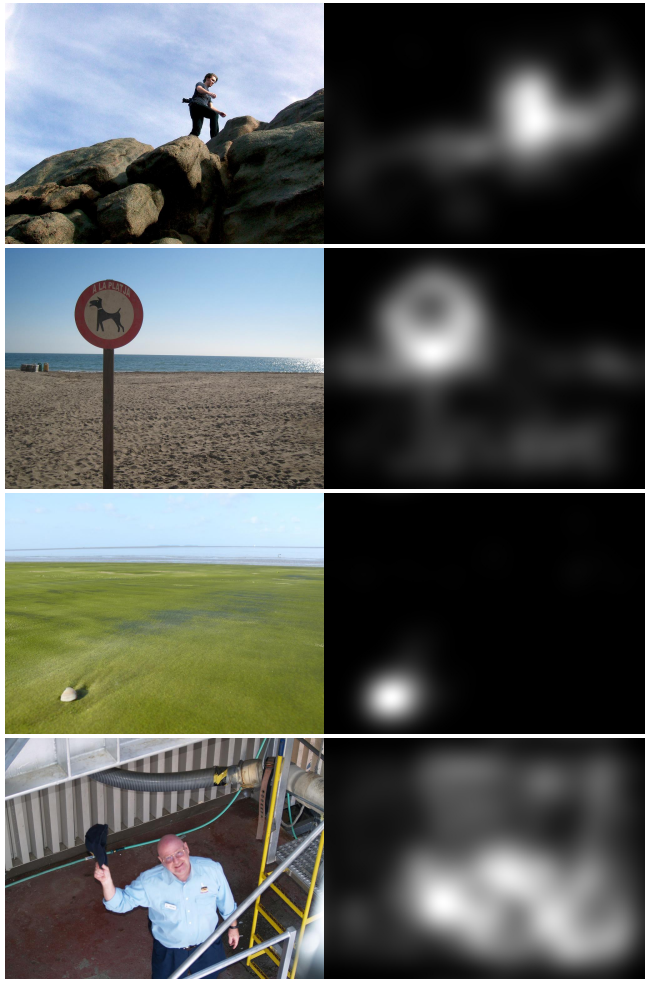
In (Le Meur et al., 2006), we designed a computational model of bottom-up visual attention.

1. Input color image;
2. Projection into a perceptual color space;
3. Subband decomposition in the Fourier domain;
4. CSF and Visual Masking;
5. Difference of Gaussians;
6. Pooling.



Cognitive models

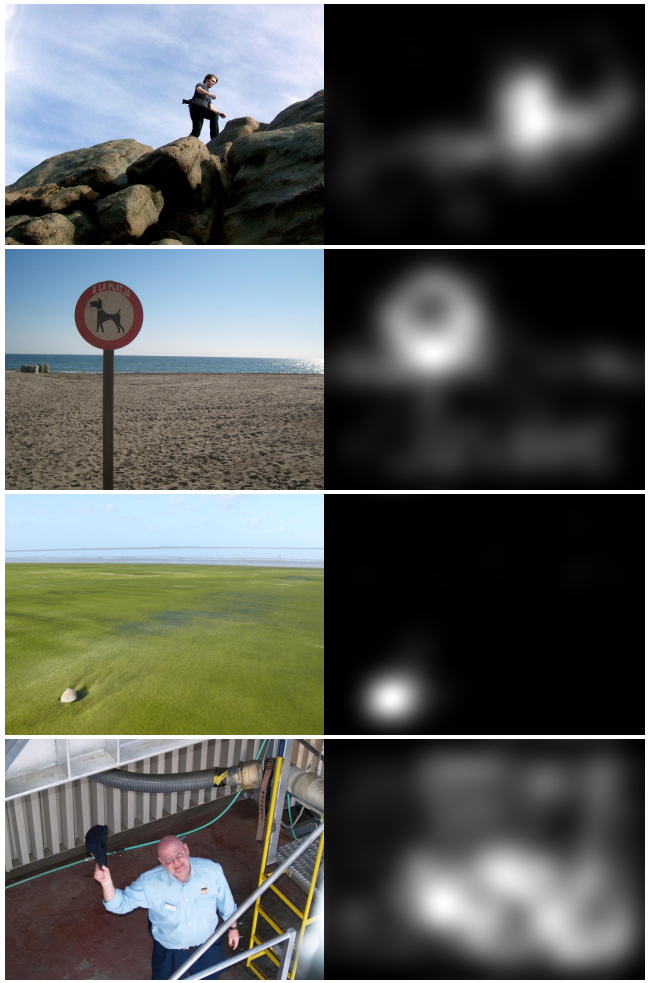
- Good prediction:



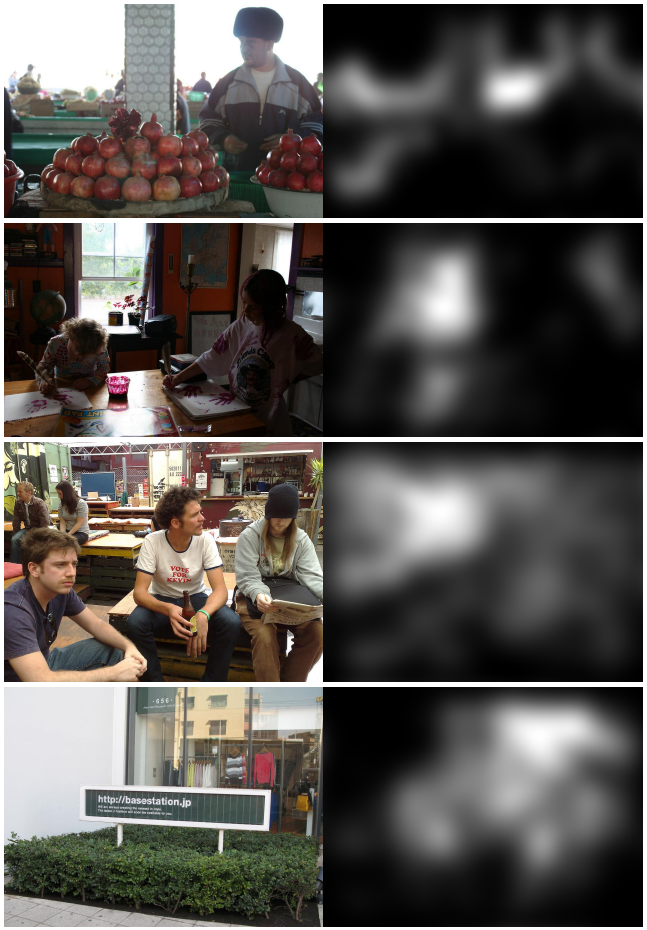
- Failure cases:

Cognitive models

- Good prediction:



- Failure cases:



Probabilistic models

Such models are based on a probabilistic framework taken their origin in the information theory.

Definition (Self-information)

Self-information is a measure of the amount information provided by an event.
For a discrete X r.v defined by $\mathcal{A} = \{x_1, ..., x_N\}$ and by a pdf, the amount of information of the event $X = x_i$ is given by:

$$I(X = x_i) = -log_2 p(X = x_i), \text{ bit/symbol}$$

Properties:

- if $p(X = x_i) < p(X = x_j)$ then $I(X = x_i) > I(X = x_j)$
- $p(X = x_i) \to 0$, $I(X = x_i) \to +\infty$

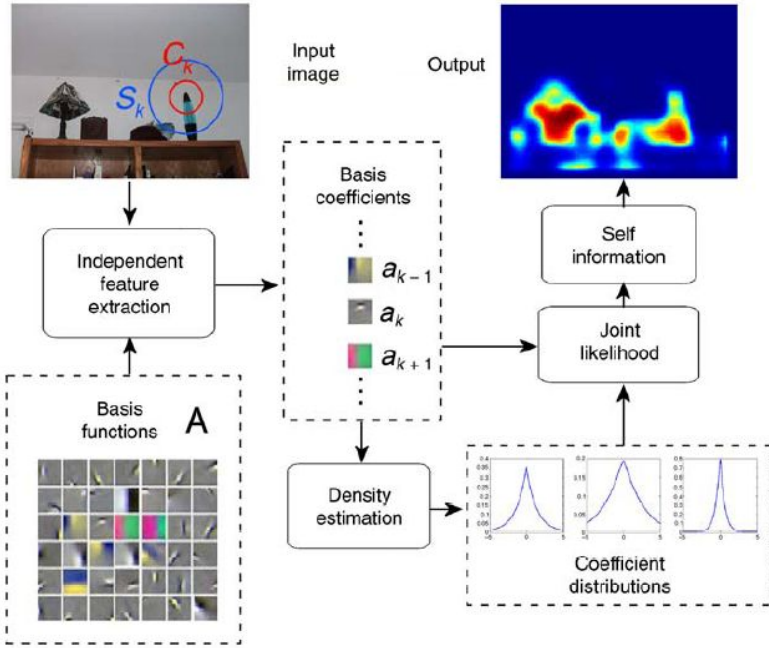
The saliency of visual content could be deduced from the self-information measure.

Self-information $\equiv$ rareness, surprise, contrast...

Probabilistic models

- First model resting on this approach has been proposed in 2003 (Oliva et al., 2003): $S(\mathbf{x}) = \frac{1}{p(\mathbf{v}_l(\mathbf{x}))}$, where $\mathbf{v}_l$ is a feature vector (48 dimensions), the probability density function is computed over the whole image.

- (Bruce and Tsotsos, 2009) modified the previous approach by using the self-information locally on independent coefficients (projection on a given basis).



From (Bruce and Tsotsos, 2009)

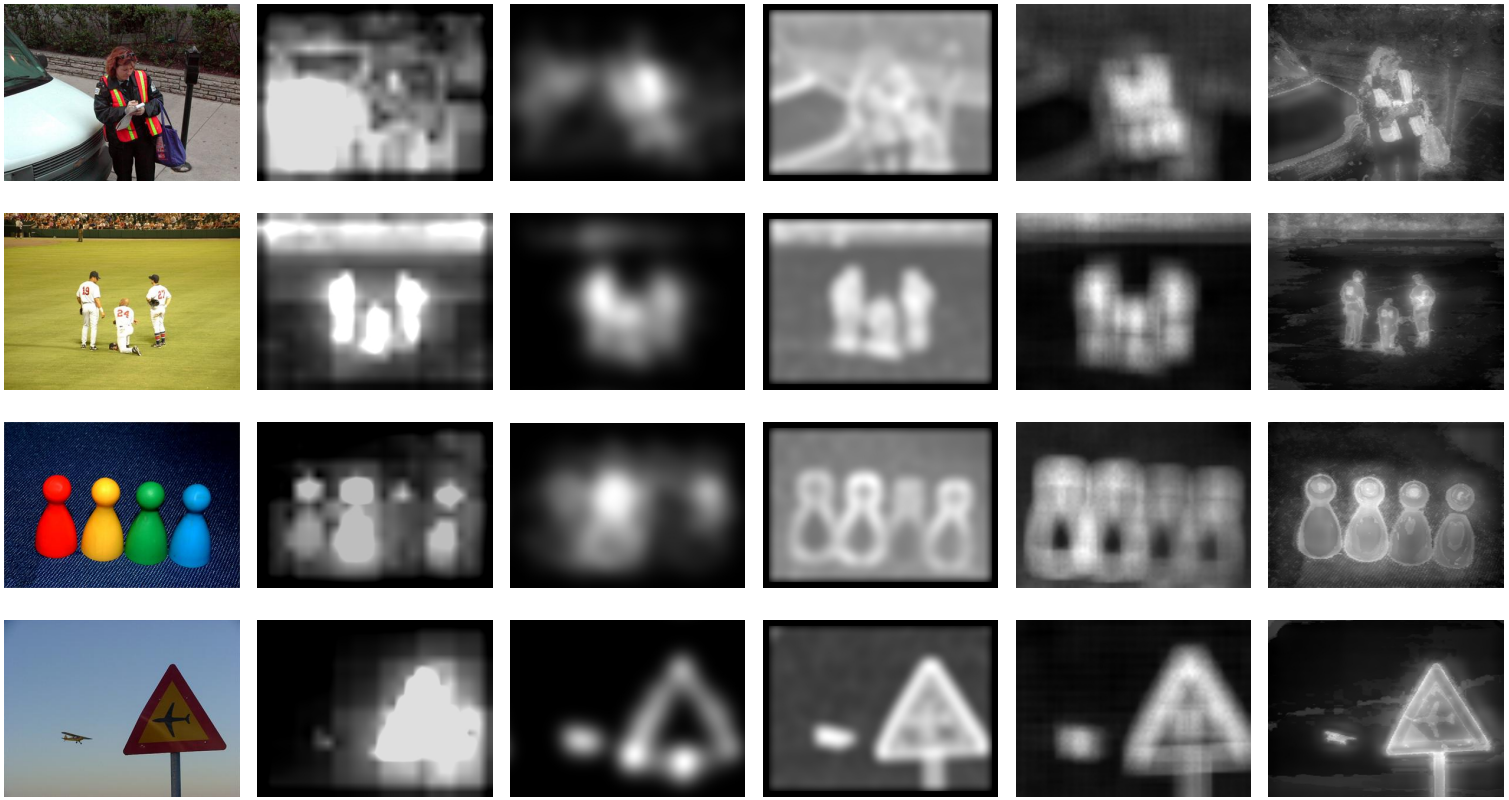
Other models in the same vein: (Riche et al., 2013), (Zhang et al., 2008).

Probabilistic models

The support regions used to calculate the pdf are not the same.

- Local spatial surround (Bruce and Tsotsos, 2009, Gao et al., 2008). Note also that Gao et al. (Gao et al., 2008) uses the mutual information to quantify the saliency;
- Whole image (Bruce and Tsotsos, 2006, Oliva et al., 2003);
- Natural image statistics (Zhang et al., 2008) (self-information).

Some examples



(a) Original (b) Itti (c) Le Meur (d) Bruce (e) Zhang (f) Judd

Saccadic models

Saccadic model to infer the saliency map

Let be an image $I : \Omega \subset \mathbb{R}^n \longmapsto \mathbb{R}^m$, the goal is to predict the visual scanpath (composed of a set of fixations x).

$$x_t^* = \arg\max_{x \in \Omega} p\left(x | x_{t-1}, \cdots, x_{t-T}\right)$$

$$p\left(x | x_{t-1}, \ldots, x_{t-T}\right) \propto p_{BU}(x) p_B(d, \phi) p_M(x, t)$$

- p_{BU} would be the bottom-up saliency map;
- p_B would be the systematic tendencies of oculomotor behavior;
- p_M would be the memory (Inhibition-of-Return);
- Be able to predict the fixation duration.

Performance

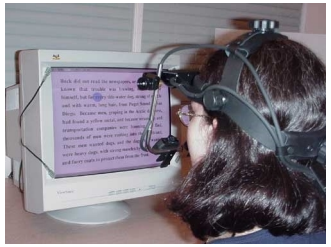
Eye tracking devices

SMI



Cambridge research system

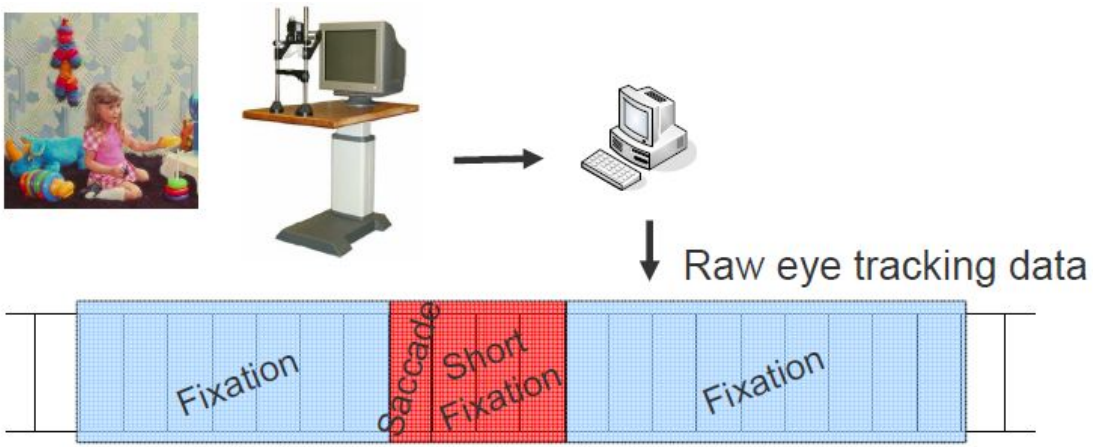


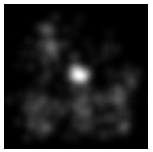


Tobii

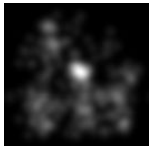


Ground truth

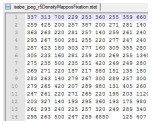




(a) SM,FD



(b) SM,FN



(c) Data

A good review of parsing algorithm in (Salvucci and Goldberg, 2000):

- Spatial:
  - → Velocity-based ;
  - → Dispersion-based;
  - → Area-based.
- Temporal:
  - → Duration sensitive ;
  - → Locally adaptive.

Ground truth

Velocity-based algorithm, the simplest method to identify fixations and saccades:

for All eye tracking data **do**
  Calculate point-to-point velocities for each point in the protocol (distance d between the current point and the next (or previous) one).
  Label each point by using a velocity threshold TH:
  if $d < TH$ **then**
    FIXATION POINT
  else
    SACCADE POINT
  end if
  Collapse consecutive fixation points into fixation groups, removing saccade points.
  Map each fixation group to a fixation at the centroid of its points;
  Return fixations.
end for

Ground truth

Dispersion-threshold identification (belonging tp the dispersion-based algorithms) relies on :

fixation points tend to cluster closely together

while there are still points **do**
  Initialize window over first points to cover the minimum fixation duration.
  Compute dispersion d inside the window
  if $d < TH$ **then**
    Add additional points to the window until $d > TH$.
    Note a fixation at the centroid of the window points.
    Remove window points from the list of points.
  else
    Remove first point from points contained inside the window.
  end if
  Return fixations
end while

Ground truth

The taxonomy and the relevance of different algos:

Table 7: Summary of identification methods.

Method	Accuracy	Speed	Robustness	Impl. Ease	Parameters
Velocity Threshold (I-VT)	√	√√	×	√√	1
Hidden Markov Model (I-HMM)	√√	√	√√	√ / × [a]	8 / 0 [a]
Dispersion Threshold (I-DT)	√√	√	√√	√	2
Minimum Spanning Tree (I-MST)	√	×	√√	×	2
Area-of-Interest (I-AOI)	×	√	√	√	1+ [b]

Key: √√ = very good, √ = good, × = not as good.
[a] I-HMM has 8 parameters in its two-state HMM that can be an be learned through reestimation. Without reestimation, I-HMM has 8 parameters but is simpler to implement; with reestimation, I-HMM effectively has no parameters but is more difficult to implement.
[b] I-AOI has 1 parameter but also requires specification of target areas.

Extracted from (Salvucci and Goldberg, 2000).

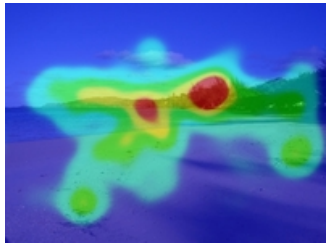
Ground truth

Transformation of the fixation maps into a continuous saliency map:

$$SM(x,y) = FixMap(x,y) * G_\sigma(x,y)$$

where, G_{sigma} is a bi-dimensional Gaussian with a standard deviation σ.

(a) Fixation Map



(b) Continuous saliency Map (heat)

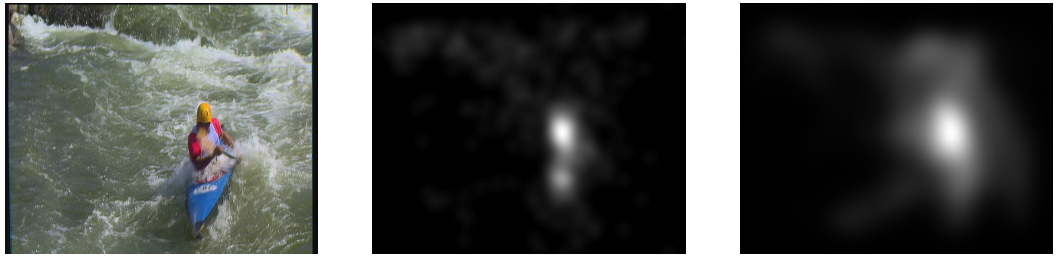
On the web (for natural images):

- Bruce's database http://www-sop.inria.fr/members/Neil.Bruce/
- Le Meur's database (28 color pictures) http://www.irisa.fr/temics/staff/lemeur/
- Rajashekar's database http://live.ece.utexas.edu/research/doves/

Metrics

Different methods are used to assess the degree of similarity between the ground truth and the prediction:

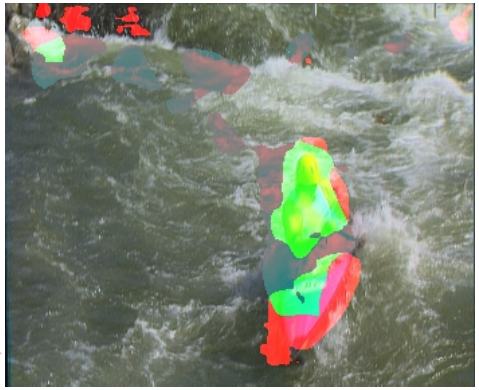
- Saliency-map-based method:



(a) Original (b) Exp. SM (c) Predicted SM

→ ROC (Receiver Operating Characteristic). Each pixel is labeled as fixated or not. Several thresholds are used (AUC (Area Under Curve)).

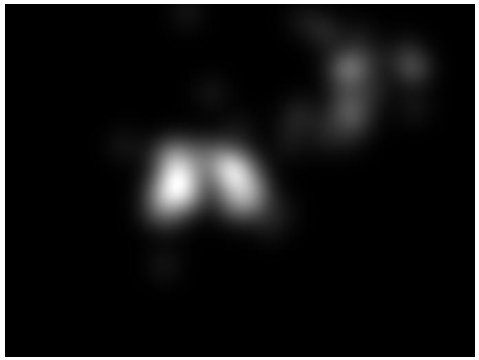
The higher the AUC, the better is the prediction, with 0.50 indicating random performance and 1.00 denoting perfect performance.



Metrics

→ ROC:

(a) Original



(b) Human

(c) Itti's model

(1) Label the pixels of the human map as fixated (255) or not (0):

The threshold is often arbitrary chosen (to cover around 20% of the picture).

Metrics

→ ROC:

(2) Label the pixels of the predicted map as fixated (255) or not (0) by a given threshold T_i:

(3) Count the good and bad predictions between human and predicted maps:

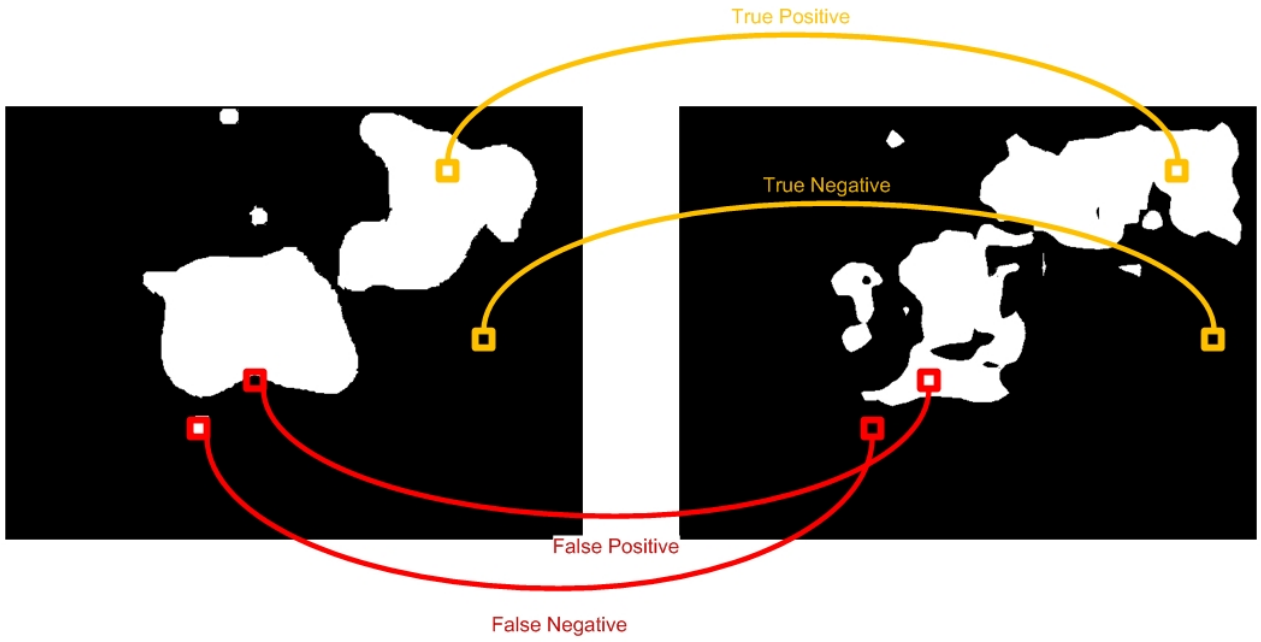
(a) Human Bin.

(b) Predicted Bin.

Metrics

→ ROC:

(3) Count the good and bad predictions between human and predicted maps:

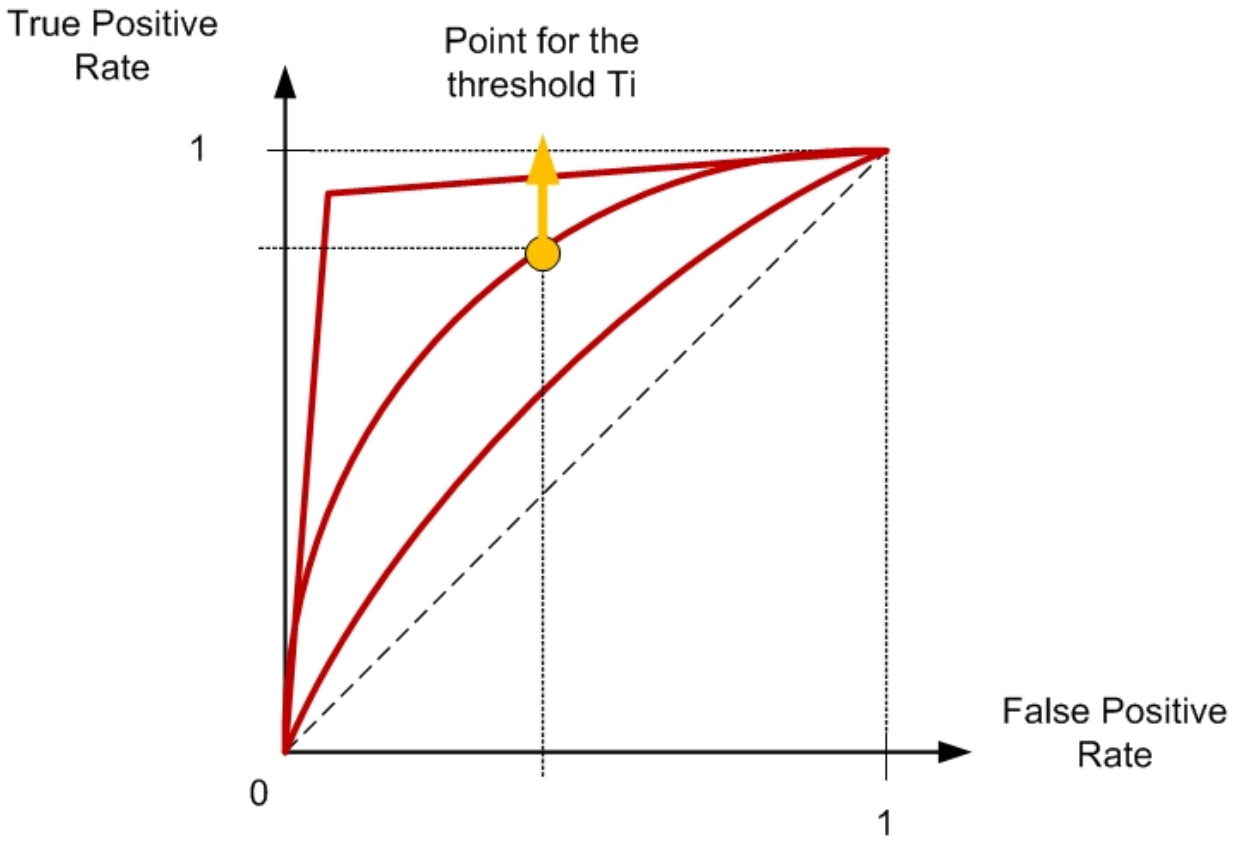


False Positive Rate = True Positive / (True Positive+False Negative)
True Positive Rate = False Positive / (False Positive+True Negative)

Metrics

→ ROC:

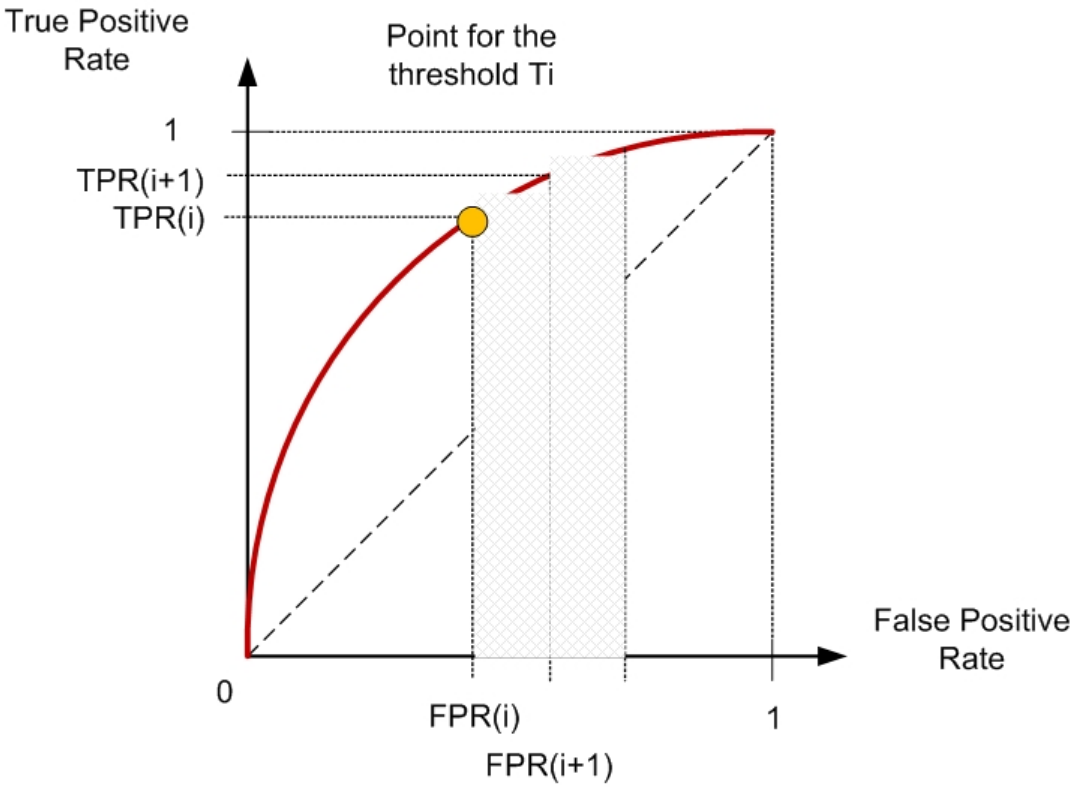
(4) Go back to (2) to use another threshold... Stop the process when all thresholds are tested.



AUC (Area Under Curve)

Metrics

→ ROC:



$$AUC = \sum_{i=0}^{N-2} (FPR(i+1) - FPR(i)) \times \frac{TPR(i+1) + TRP(i)}{2}$$

where, N is the number of threshold used.

Metrics

- KL-Divergence:

$$KL(p|h) = \sum_{i,j} p(i,j) log_2 \frac{p(i,j)}{h(i,j)}$$

where p and h are the pdf of the predicted and human saliency maps.

$$p(i,j) = \frac{SM_p(i,j)}{\sum_{i,j} p(i,j)}$$

$$h(i,j) = \frac{SM_h(i,j)}{\sum_{i,j} h(i,j)}$$

KL is a divergence: $KL = 0$ when p and h are strictly the same, $KL \geq 0$.

- Linear correlation coefficient:

$$CC(p,h) = \frac{cov_{ph}}{\sigma_p \sigma_h}$$

where σ_k is the standard deviation of k and cov_{ph} is the covariance between p and h. CC is between -1 and 1.

Hybrid methods

Four methods involving scanpaths and saliency maps :

- Receiver Operating Analysis;
- Normalized Scanpath Saliency (Parkhurst et al., 2002, Peters et al., 2005);
- The Kullback-Leibler divergence.

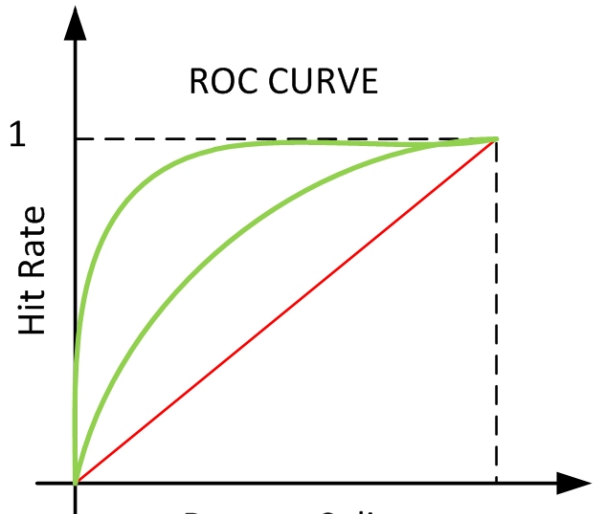
Hybrid methods

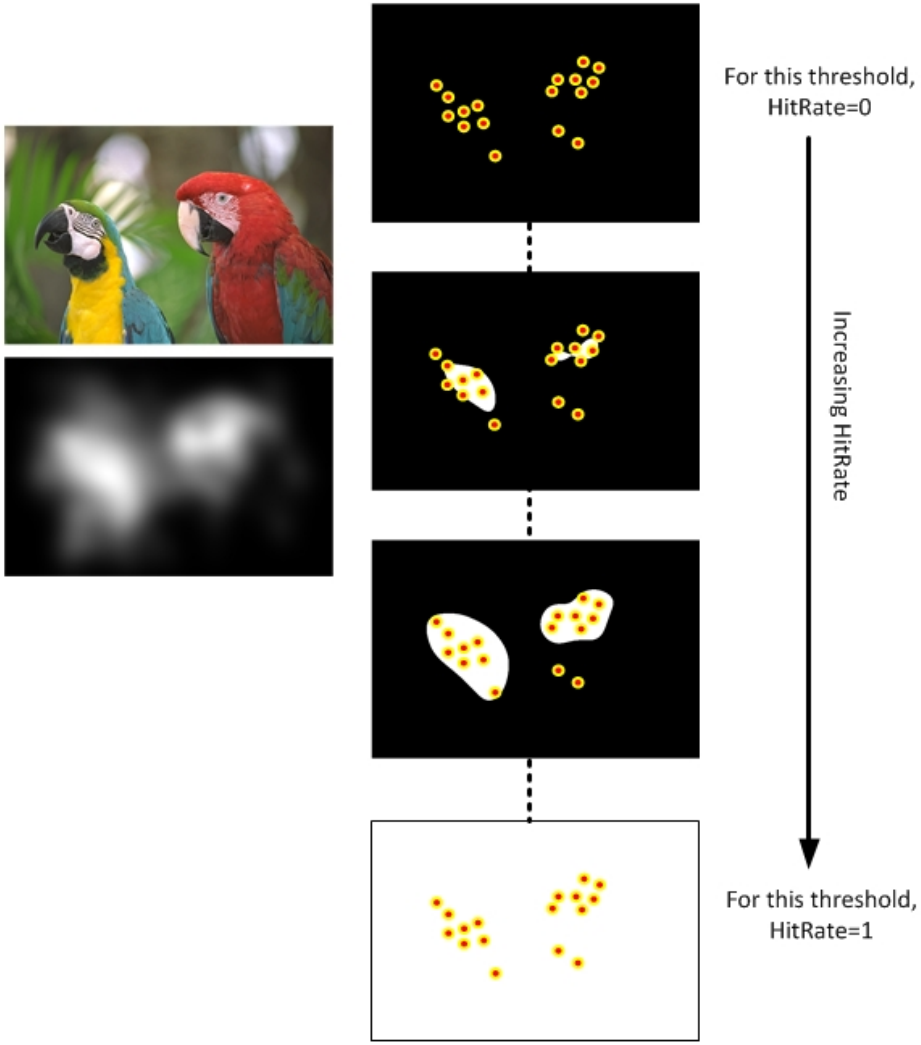
Receiver Operating Analysis

ROC analysis is performed between a continuous saliency map and a set of fixations.

- Hit rate is measured in function of the threshold used to binarize the saliency map (Judd et al., 2009):

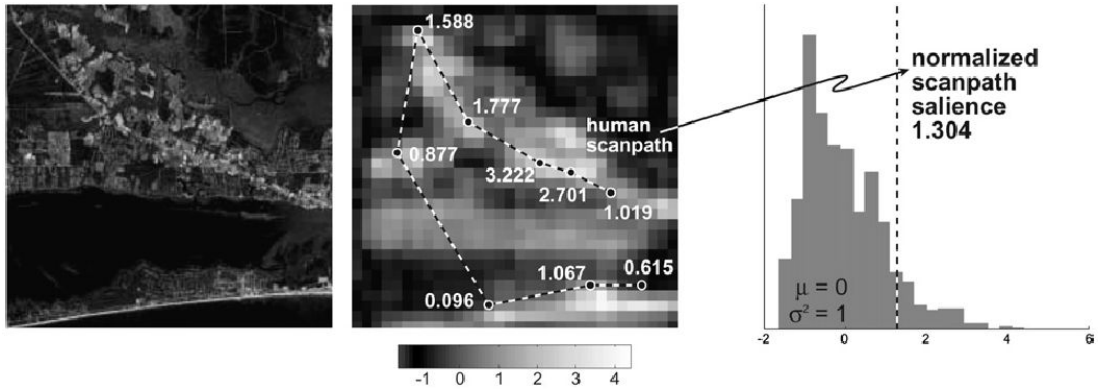
ROC curve goes from 0 to 1!





Hybrid methods

- NSS (Normalized Scanpath salience) gives the degree of correspondence between human fixation locations and predicted saliency maps (Parkhurst et al., 2002),(Peters et al., 2005).
  1. Each saliency map is normalized to have zero mean and one unit standard deviation.
  2. Extraction of the predicted saliency at a given human fixation point.
  3. Average of the previous values.



From (Peters et al., 2005)

$NSS = 0$: random performance;
$NSS >> 0$: correspondence between human fixation locations and the predicted salient points:
$NSS << 0$: anti-correspondence.

Metrics

- Scanpath-based method:
  - → Levenshtein's distance (or string edit) (Levenshtein, 1966). It relies on three character operations: deletion, insertion and substitution. Each is equal to a unit cost.

$$s_1 = abcfeffgdc \text{ and } s_2 = afbffdcdf$$

$s_1 = abcfeffgdc$ $s_2 = afbffdcdf$	start	cost 0
$s_1 = abcfeffgdc$ $s_2 = afeffdcdf$	after substitution of first b by e	cost 1
$s_1 = abcfeffgdc$ $s_2 = abcfeffdcdf$	after insertion of bc after first a	cost 2
$s_1 = abcfeffgdc$ $s_2 = abcfeffdc$	after deletion of last df	cost 2
$s_1 = abcfeffgdc$ $s_2 = abcfeffgdc$	after insertion of g	cost 1

From (Privitera and Stark, 2000).

The Levenshtein's distance between s_1 and s_2 is 6.

Metrics

- Scanpath-based method:
  - → Mannan et al.'s distance (Mannan et al., 1995):

$$I = \left[1 - \frac{D}{D_r}\right] \times 100$$

where D is a measure of dissimilarity given by:

$$D^2 = \frac{n_1 \sum_{j=1}^{n_2} d_{2j}^2 + n_2 \sum_{i=1}^{n_1} d_{1i}^2}{2 n_1 n_2 (a^2 + b^2}$$

where ,

- n_1 and n_2 are the number of fixations in the two traces;
- d_{1i} is the distance between the location of the ith fixation in the first trace and its nearest neighbour in the second trace;
- d_{2j} the distance between j^{th} fixation in the second trace and that of its nearest neighbour in the second one;
- a^2 and b^2 are the squares of the side length of the image;
- D_r is defined as being the distance between two sets of random locations.

Metrics

- Scanpath-based method:
  - → Adaptation of Levenshtein's distance for comparing an artificially generated scanpath to that of human (Privitera and Stark, 2000). The viewing stimulus was partitioned into Region Of Interest by using a segmentation algorithm:

    Two different indices of similarity:
    - S_p : Locus similarity (same areas or not);
    - S_s : Sequence similarity (time dependent).

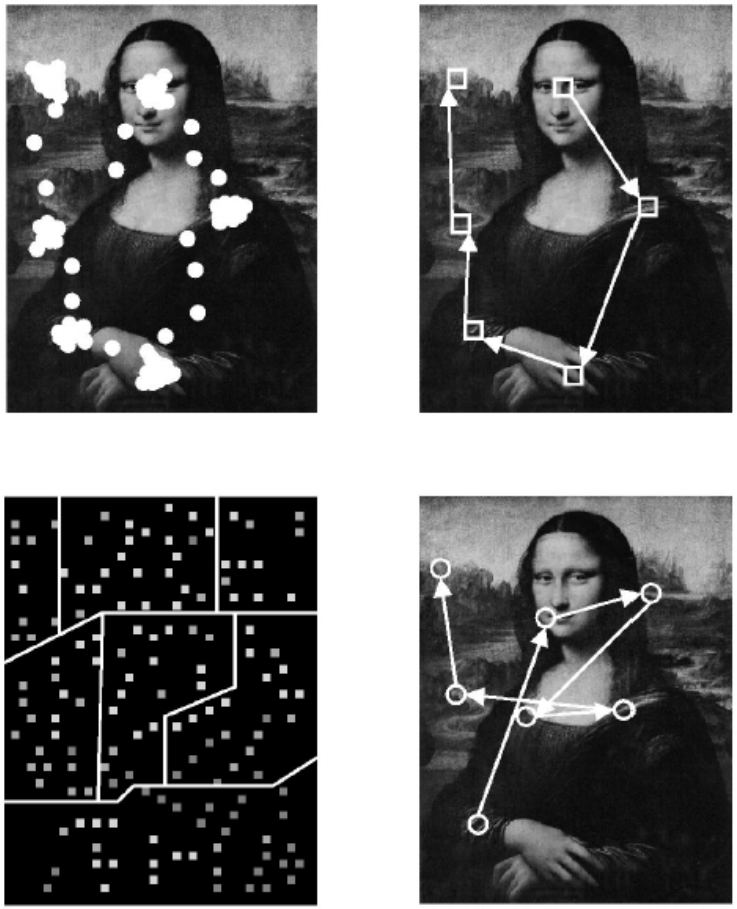


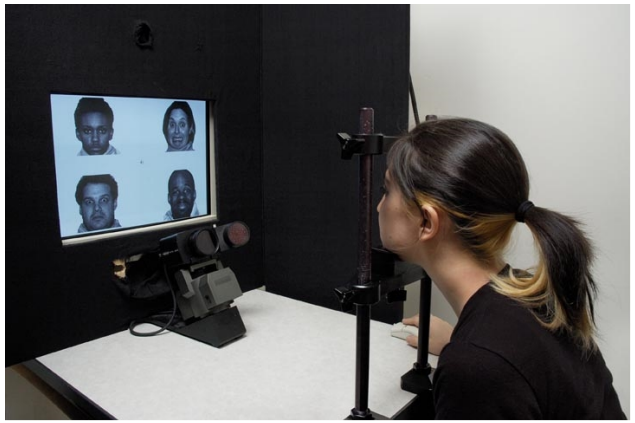
Fig. 1. Computer and human processing—comparing *human* identified Regions-of-Interest, hROIs, (upper right) with *algorithmic* identified Regions-of-Interest, aROIs, (lower right). Note eye movement sampling, (upper left) and local maxima in the processed image (lower left).

From (Privitera and Stark, 2000).

Limitations

Mostly inherent to the building of the ground truth and to the experimental setting...

1. Several parameters can have a significant impact on the results:
   - The apparatus used to record the eye movements (sampling rate, chin rest...). Examples of eye-trackers:



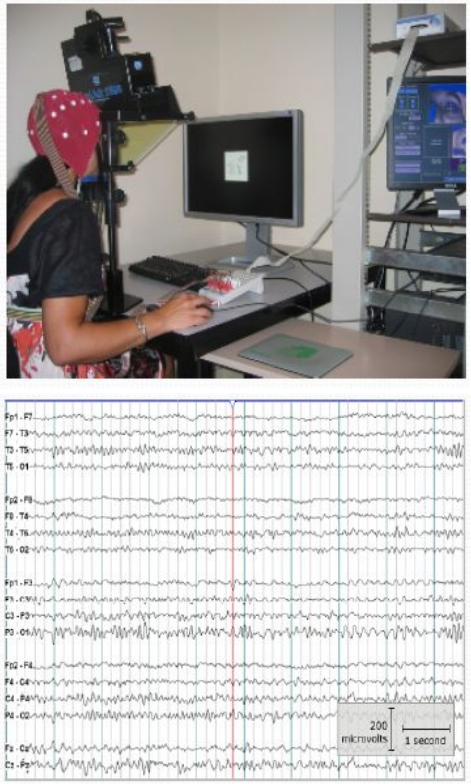
Limitations

1. Several parameters can have a significant impact on the results:
   - A significant central bias (strong artifact):
     - the screen centre might be a privileged location for viewing scenes on computer screen (whatever the task performed, whatever the stimuli) (Tatler, 2007);
     - the central bias might reflect a tendency to re-center the eye in its orbit (Pare and Munoz, 2001).
   - Cognitive constraints (higher-level goals, prior knowledge, expectations...)
   - The task subjects performed. What is the question we should asked?
   - The nature of the stimuli viewed;
2. Has the visual fixation the same meaning?
   - What about the fixation duration? Depth of processing? Is there a correlation between saliency and duration?

Limitations and perspectives

The assessment of computational model mainly rests on the analysis of individual fixations. But...

- Is there a focal-ambient dichotomy? (Follet et al., 2009, Unema et al., 2005).
- Does every fixation convey the same processing?
- Would it be possible to categorize fixations as attentional fixations, semantic fixations...?



Courtesy of T. Baccino.

A promising solution to disentangle different processes is called the EFRP (Eye-Fixation-Related Potentials) (Baccino and Manunta, 2005).

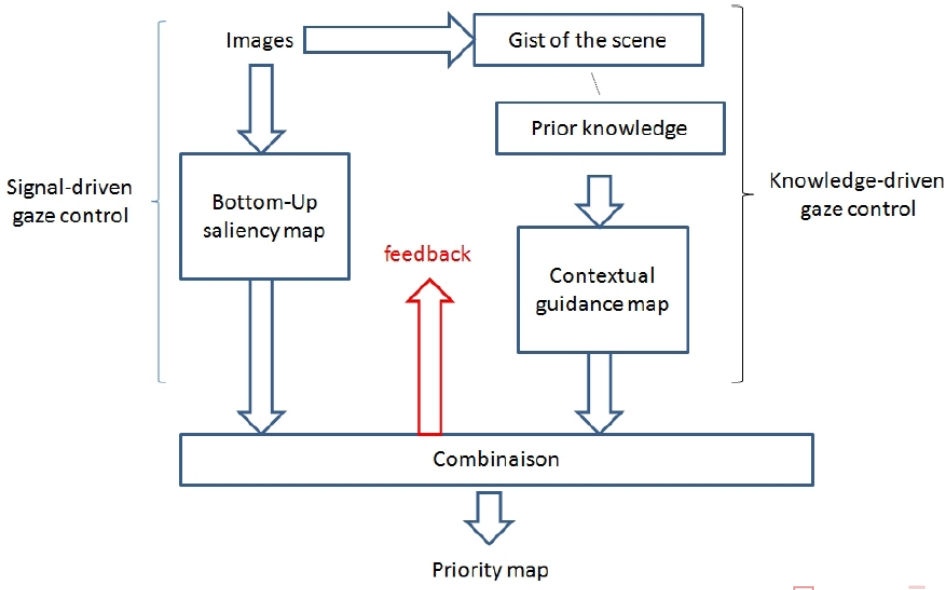
Measuring ERP (Event-Related Potential) and EM conjointly to track the cognitive processes.

Conclusion

The modelling of the human visual attention raises a number of fundamental issues:

- The role of the bottom-up and the top-down, theirs contributions?
- The meaning and the role of visual fixations?
- The role of contextual information...

Future models of visual attention should take into account higher-level information:



Computational models can be used by numerous applications. For instance:

- Video Compression;
- Image quality;
- Retargeting:

From Le Meur et al. (2006).

References

M. Ahissar and S. Hochstein. The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences*, 19(10), 2004.

T. Baccino and Y. Manunta. Eye-fixation-related potentials: insight into parafoveal processing. *Journal Of Psychophysiology*, 19 (3):204–215, 2005.

M. Bar. Visual objects in context. *Nature Reviews:Neuroscience*, 5:617–629, 2004.

C. Blais, R.E. Jack, C. Scheepers, D. Fiset, and R. Caldara. Culture shapes how we look at faces. *PLoS ONE*, 3(8), 2008.

A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35: 185–207, 2013.

N.D.B. Bruce and J.K. Tsotsos. Saliency based on information maximization. In *Advances in Neural Information Processing Systems*, volume 18, pages 155–162, 2006.

N.D.B. Bruce and J.K. Tsotsos. Saliency, attention and visual search: an information theoretic approach. *Journal of Vision*, 9: 1–24, 2009.

J.H. Fecteau and D.P. Mumoz. Salience, relevance and firring: a priority map for target selection. *Trends Cogn Sci*, 10(8): 382–390, 2006.

B. Follet, O. Le Meur, and T. Baccino. Relationship between coarse-to-fine process and ambient-to-focal visual. In *ECEM*, 2009.

T. Foulsham, A. Kingstone, and G. Underwood. Turning the world around: Patterns in saccade direction vary with picture orientation. *Vision Research*, 48:1777–1790, 2008.

D. Gao, V. Mahadevan, and N. Vasconcelos. On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of Vision*, 8:1–18, 2008.

L. Itti, C. Koch, and E. Niebur. A model for saliency-based visual attention for rapid scene analysis. *IEEE Trans. on PAMI*, 20: 1254–1259, 1998.

T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where people look. In *ICCV*, 2009.

C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4: 219–227, 1985.

O. Le Meur and J.C. Chevet. Relevance of a feed-forward model of visual attention for goal-oriented and free-viewing tasks. *IEEE Trans. On Image Processing*, 19(11):2801–2813, 2010.

O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau. A coherent computational approach to model the bottom-up visual attention. *IEEE Trans. On PAMI*, 28(5):802–817, May 2006.

V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Sov. Phys. Dokl.*, 6:707–710, 1966.

S.K. Mannan, K.H. Ruddock, and D.S. Wooding. Automatic control of saccadic eye movements made in visual inspection of briefly presented 2d images. *Spatial Vision*, 9:363–386, 1995.

M. Nadenau. *Integration of human color vision models into high quality image compression*. PhD thesis, EPFL, 2000.

A. Oliva and P.G. Schyns. From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science*, 5:195–200, 1994.

A. Oliva and P.G. Schyns. Colored diagnostic blobs mediate scene recognition. *Cognitive Psychology*, 2000.

A. Oliva, A. Torralba, M.S. Castelhano, and J.M. Henderson. Top-down control of visual attention in object detec- tion. In *IEEE ICIP*, 2003.

M. Pare and D.P. Munoz. Expression of a recentering bias in saccade regulation by superior colliculus neurons. *Experimental Brain Research*, 137:354–368, 2001.

D. Parkhurst, K. Law, and E. Niebur. Modelling the role of salience in the allocation of overt visual attention. *Vision Research*, 42: 107–123, 2002.

R. J. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(18): 2397–2416, 2005.

M. I. Posner. Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32:3–25, 1980.

C. M. Privitera and L. W. Stark. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22:970–982, 2000.

J.E. Raymond, K.L. Shapiro, and K.M. Arnell. Temporary suppression of visual processing in an rsvp task: an attentional blink? *Journal of Experimental Psychology: Human perception & Performance*, 18(3):849–860, 1992.

R.A. Rensink, J. K. O'Regan, and J. Clark. To see or not to see: the need for attention to perceive changes in scenes. *Psychological Science*, 8(5):368–373, 1997.

N. Riche, M. Mancas, M. Duvinage, M. Mibulumukini, B. Gosselin, and T. Dutoit. Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. *Signal Processing: Image Communication*, 28(6):642 – 658, 2013. ISSN 0923-5965. doi: http://dx.doi.org/10.1016/j.image.2013.03.009.

S. Salvucci and J. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Eye Tracking Research and Application (ETRA)*, page 71Ű78, 2000.

Chengyao Shen and Qi Zhao. Webpage saliency. In *ECCV*. IEEE, 2014.

B. W. Tatler. The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7, 2007.

S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381:520–522, 1996.

A. Treisman and G. Gelade. A feature integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

J. K. Tsotsos, S. M. Culhane, Winky Y. Kei W., Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *Artificial intelligence*, 78(1):507–545, 1995.

P.J.A. Unema, S. Pannasch, M. Joos, and B.M. Velichkovsky. Time course of information processing during scene perception: the relationship between saccade amplitude and fixation duration. *Visual Cognition*, 12(3):473–494, 2005.

R. Van Rullen. Saliency and spike timing in the ventral visual pathway. *Journal Of Physiology*, 2003.

B. M. Velichkovsky. Heterarchy of cognition: The depths and the highs of a framework for memory research. *Memory*, 10(5): 405–419, 2002.

J.M. Wolfe. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review 1, 2*, x:202–238, 1994.

A.L. Yarbus. *Eye movements and vision*. Plenum Press: New York, 1967.

L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. Sun: A bayesian framework for salience using natural statistics. *Journal of Vision*, 8(7):1–20, 2008.