Scene Understanding

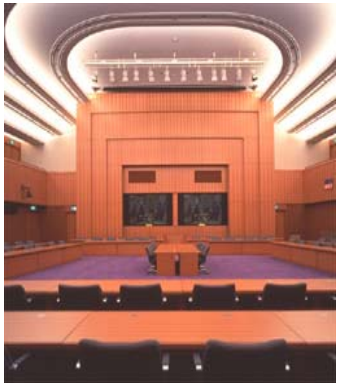
Aude Oliva

Brain & Cognitive Sciences
Massachusetts Institute of Technology
Email: oliva@mit.edu   http://cvcl.mit.edu



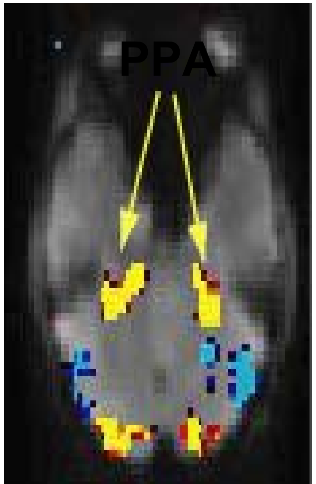












Definition

- A scene is a view of a **real-world environment** that contains **multiples** surfaces and objects, organized in a **meaningful way**.

- Distinction between objects and scenes:

objects are compact and act upon

Scenes are extended in space and act within

The distinction depends on the action of the agent

A tour of Scene Understanding’s litterature

9.912 Scene Understanding Seminar

DATABASE | ARTICLES | MATLAB CODE | HOME 9.912 | CVCL HOME

This material is based upon work supported by the National Science Foundation under CAREER Grant No. 0546262 awarded to Aude Oliva

BOOKS

Computer Vision: a modern approach. Forsyth and Ponce. Prentice Hall. (download slides here)

Vision Science: Photons to Phenomenology. Stephen E. Palmer, MIT Press.

ARTICLES

Section 1: Classics in Scene Understanding and Visual Cognition

Bergen, J.R., & Adelson, E.H. (1988). Early vision and texture perception. Nature, 333, 363-364.

Biederman, I., Glass, A.L., & Stacy, E.W. (1973). Searching for objects in real-world scenes. Journal of experimental psychology,97, 22-27.

Biederman, I., Rabinowitz, J.CV., Glass, A.L., & Stacy, E.W. (1974). On the information extracted from a glance at a scene. Journal of experimental psychology, 103, 597-600.Potter, M.C. (1975). Meaning in visual search. Science, 187, 965-966.

Bruner, J.S, & Potter, M.C. (1964). Interference in visual cognition. Science, 144, 424-425.

Friedman, A. (1979). Framing pictures: the role of knowledge in automatized encodingand memory of gist. Journal of Experimental Psychology: General, 108, 316-355.

Potter, M.C., & Levy, E.I. (1969). Recognition memory for a rapid sequence of pictures. Journal of experimental psychology, 81, 10-15.

Navon, D. (1977). Forest before Trees: The precedence of global features in visual perception. Cognitive Psychology, 353-383.

Shiffrin, W., & Schneider, R.M. (1977). Controlled and Automatic Human Information Processing: I. Detection, Search, and Attention . Psychological Review, 84(1), 1-66.

Sperling, G. (1960). The information available in brief visual presentations. Psychological Monographs: General and Applied, vol 74 (11).

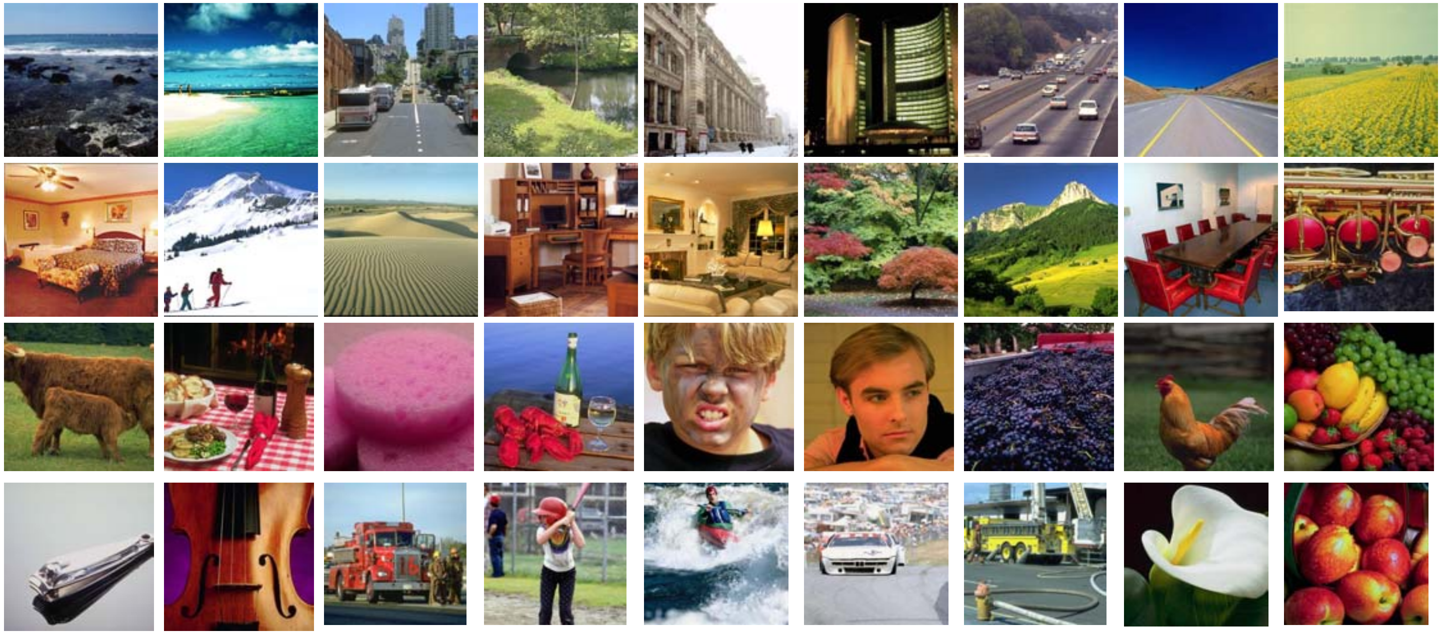
Sperling, G. (1963). A model for visual memory tasks. Human Factors, February (pp 20-31).

http://cvcl.mit.edu/SUNSarticles.htm

I. Rapid Visual Scene Recognition

We move our eyes every 300 msec on average

How do human recognize natural images in a short glance ?



Demonstrations

First, I am going to show you how ***good*** the visual system is

Then, I will show you how ***bad*** the visual system is

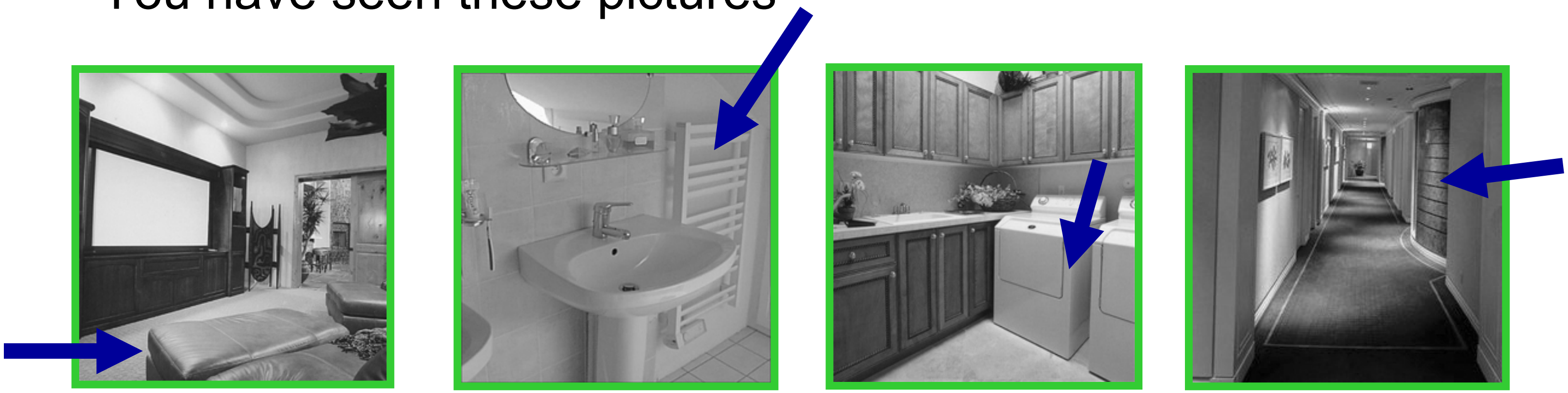
Memory Confusion:
The scenes have the <u>same spatial layout</u>

You have seen these pictures

You were tested with these pictures

Memory Confusion:
The details of some objects are forgotten

You have seen these pictures



You were tested with these pictures

Human fast scene understanding

In a glance, we remember the meaning of an image and its global layout but some objects and details are forgotten

A few facts about human scene understanding

- Immediate recognition of the *meaning* of the scene and the *global structure*

- Quick visual perception lacks of objects and details information. Objects are *inferred, not necessarily seen*

Which One Did You See?

A

B

C

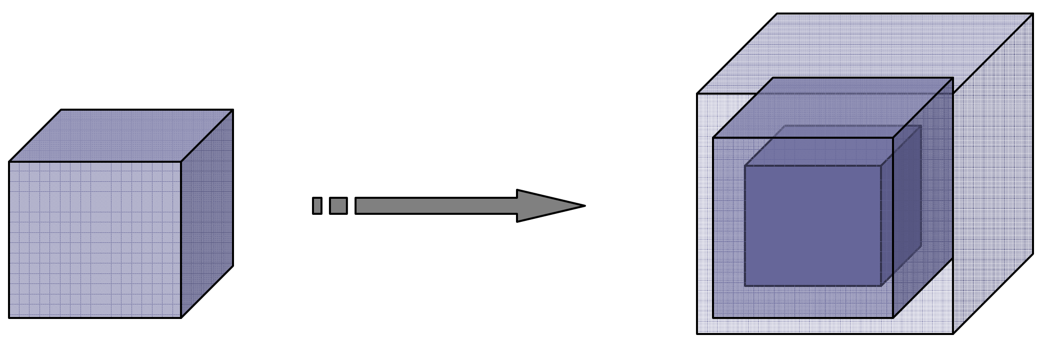
D

Systematic scene memory *distortion*

correct answer

B **C**

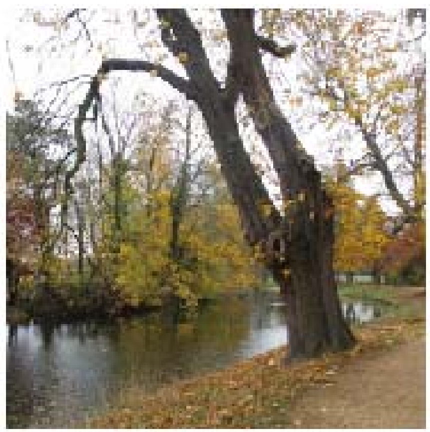
too close ⟷ **too far**

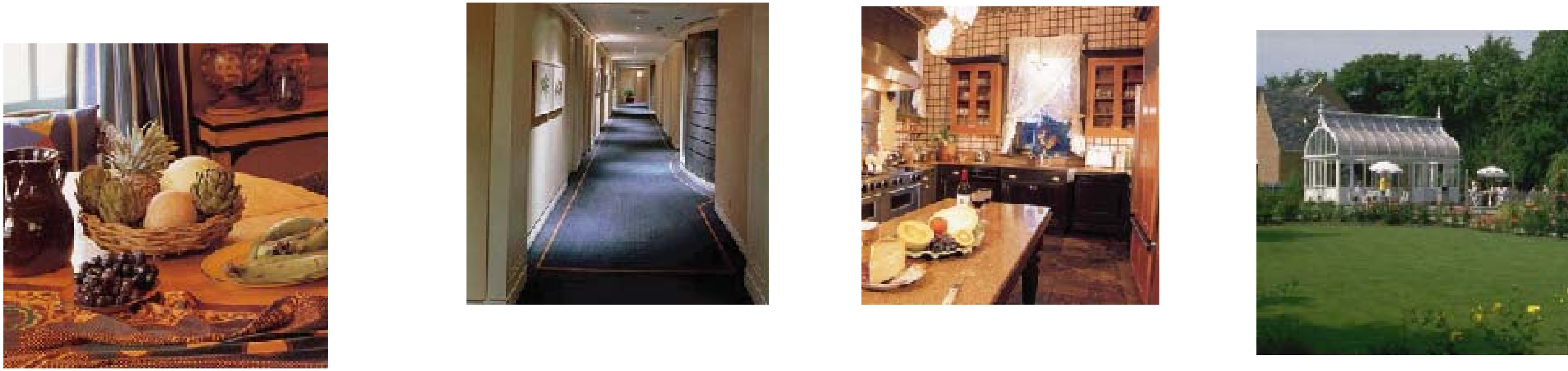


Helene Intraub (Boundary Expansion Effect on pictures of object)







Test images

Scene Representation
Time course of visual information within a glance

- Definition: what is the “gist”
- A few observations : getting the gist of a scene
- How do spatial frequency information unfold?
- What is the role of color ?
- What are the global properties of a scene?

The Gist of the Scene

- Mary Potter (1975, 1976) demonstrated that during a rapid sequential visual presentation (100 msec per image), a novel scene picture is indeed instantly **understood** and observers seem to comprehend a lot of visual information, **but a delay of a few hundreds msec (~ 300 msec) is required for the picture to be consolidated in memory.**

- The "gist" (a summary) refers to the visual information perceived after/during a glance at an image.

- To simplify, the gist is often synonymous with the *basic-level category* of the scene or event (e.g. wedding, bathroom, beach, forest, street)

What is represented in the gist ?

- The "Gist" includes all levels of visual information, from low-level features (e.g. color, luminance, contours), to intermediate (e.g. shapes, parts, textured regions) and high-level information (e.g. semantic category, activation of semantic knowledge, function)

- **Conceptual gist** refers to the semantic information that is inferred while viewing a scene or shortly after the scene has disappeared from view.

- **Perceptual gist** refers to the structural representation of a scene built during perception (~ 200-300 msec).

Oliva, A. (2005). Gist of a scene. In *Neurobiology of Attention*. Eds. L. Itti, G. Rees and J. Tsotsos. Academic Press, Elsevier.

Rapid Scene "Gist" Understanding:
Mechanism of recognition

- Mary Potter (1975, 1976) demonstrated that during a rapid sequential visual presentation (100 msec per image), a novel picture is instantly **understood** and observers seem to comprehend a lot of visual information

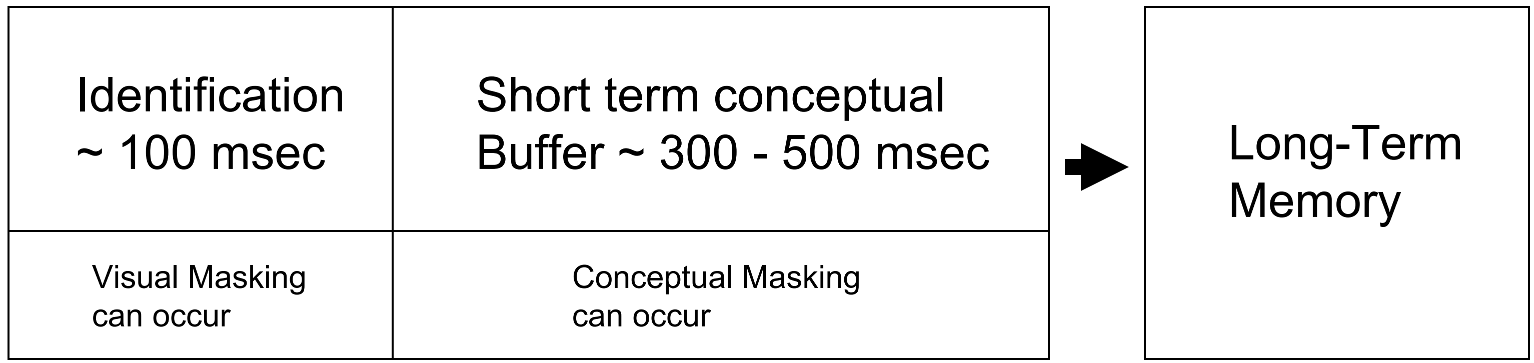
- But a delay of a few hundreds msec (~ 300 msec) is required for the picture to be consolidated in memory.

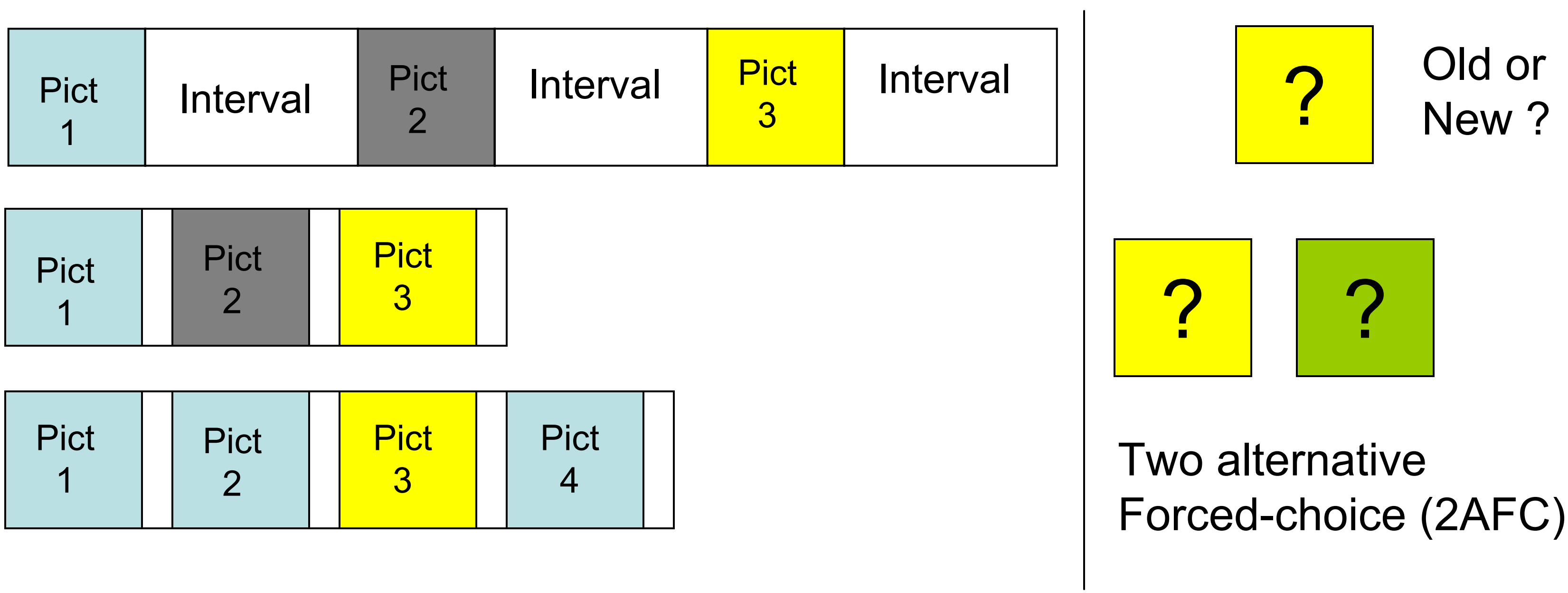
Pict 1	Interval	Pict 2	Interval	Pict 3	Interval

Identification ~ 100 msec	Short term conceptual buffer ~ 300 msec	→	Long-Term Memory
Visual Masking can occur	Conceptual Masking can occur		

Basis of RSVP paradigm

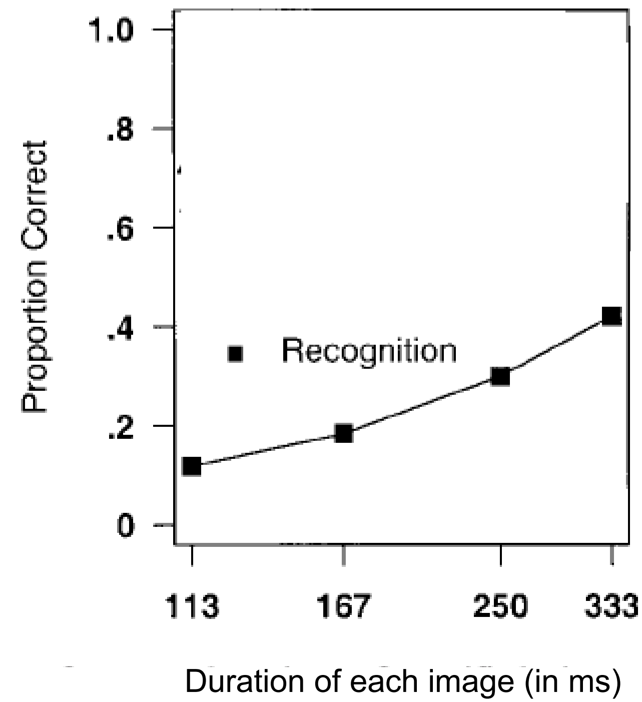
Rapid Sequential Visual Presentation





Molly Potter's work (1976)

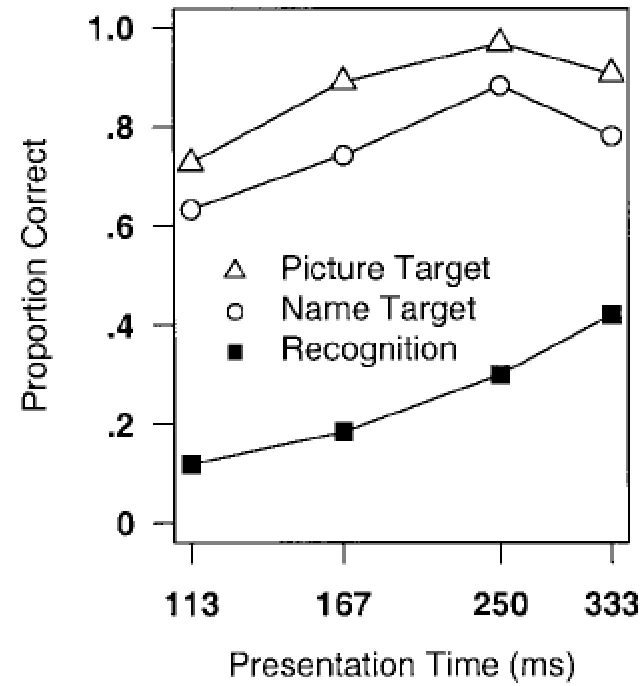
Effect of conceptual masking: the *n+1* picture interferes with the processing of picture *n*.



Is this a fixed "limit" ? Can we beat this limit in temporal processing ?

When cued ahead about which image to search for …

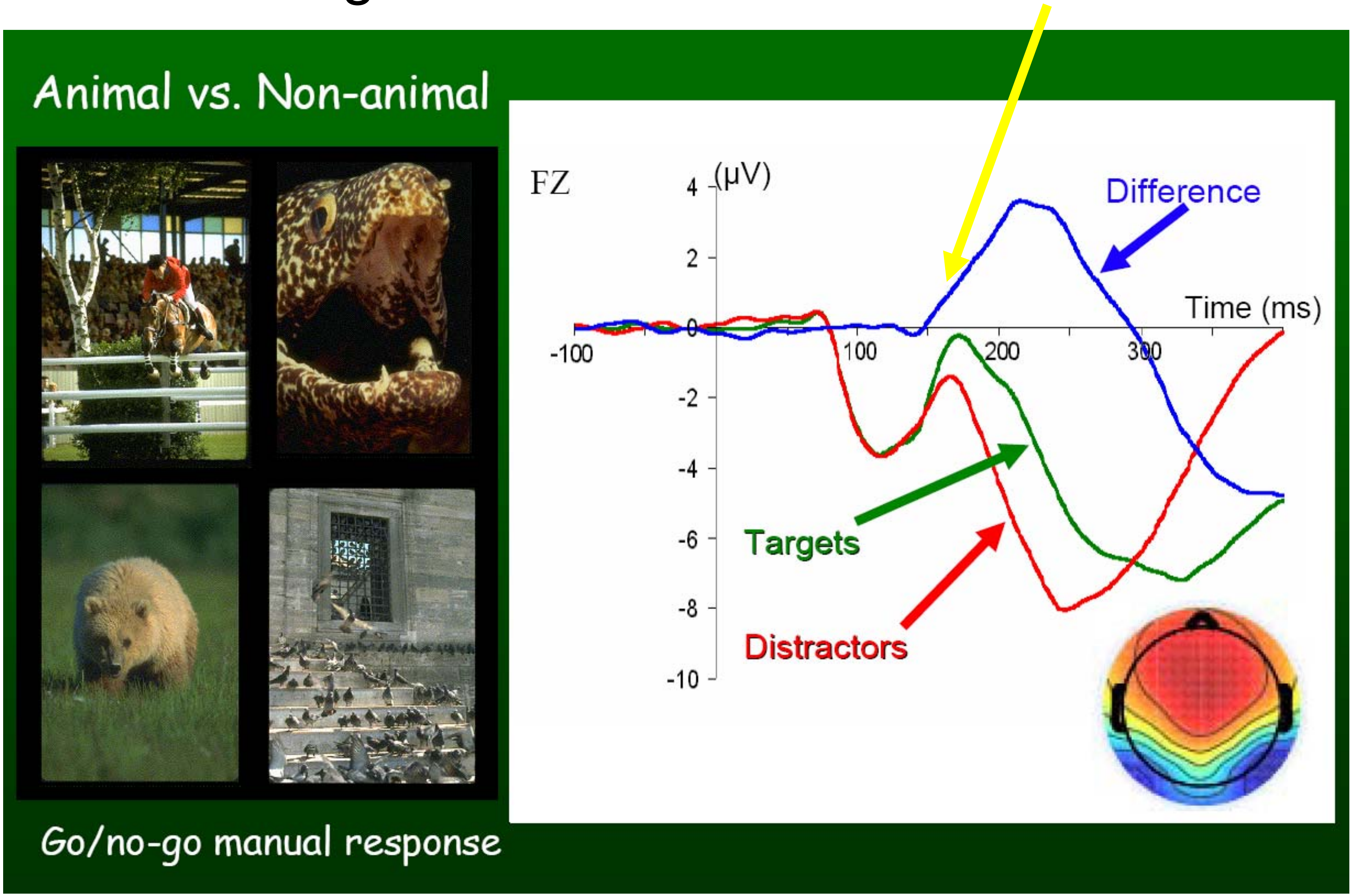
Observers were cued ahead of time about the possible appearance of a picture in the RSVP stream (the cue consisted of a picture, or a short verbal description of the picture, “a picnic at the beach”) and were asked to detect it



A viewer can comprehend a scene in 100-200 msec but cannot retain it without additional time.
At higher temporal rates, pictures are “forgotten”

Thorpe (1998): Detecting an animal among distractors

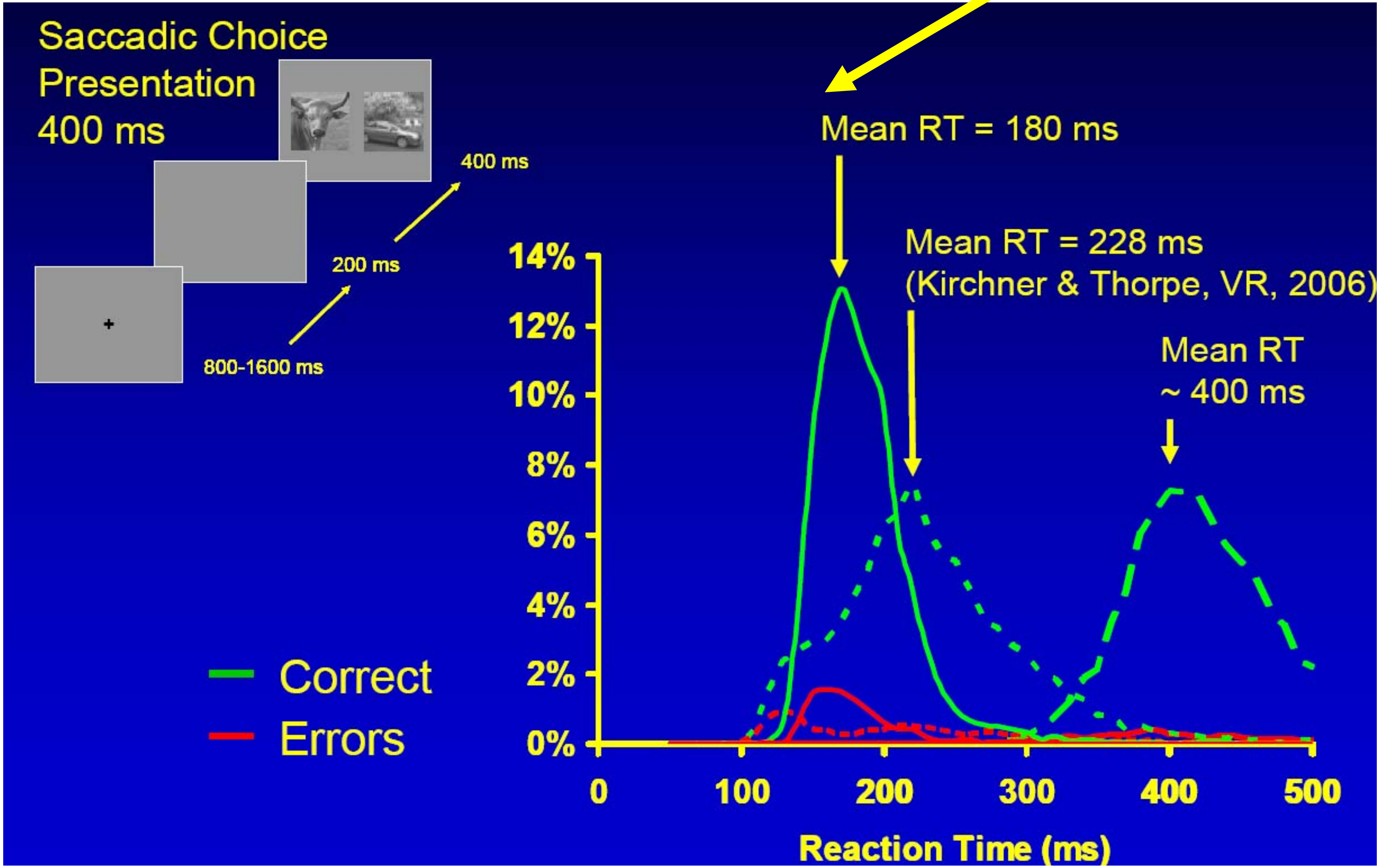
EEG response 150-160 msec after image presentation



http://suns.mit.edu/SUnS07Slides/FabreThorpe_SUnS07.pdf

Kirchner & Thorpe (2006)

Saccadic response 180 msec after image presentation

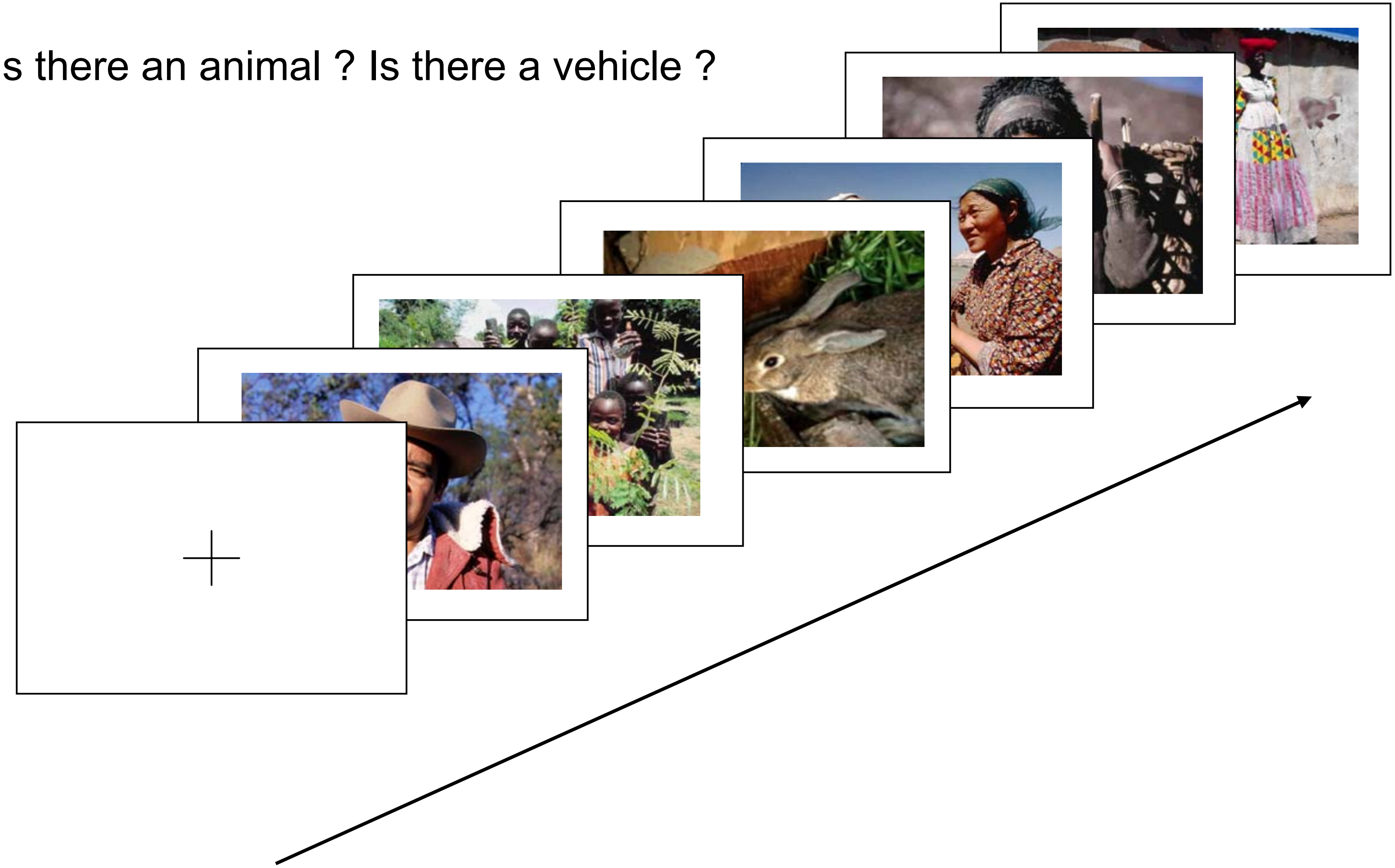


http://suns.mit.edu/SUnS07Slides/Thorpe_SUnS07.pdf

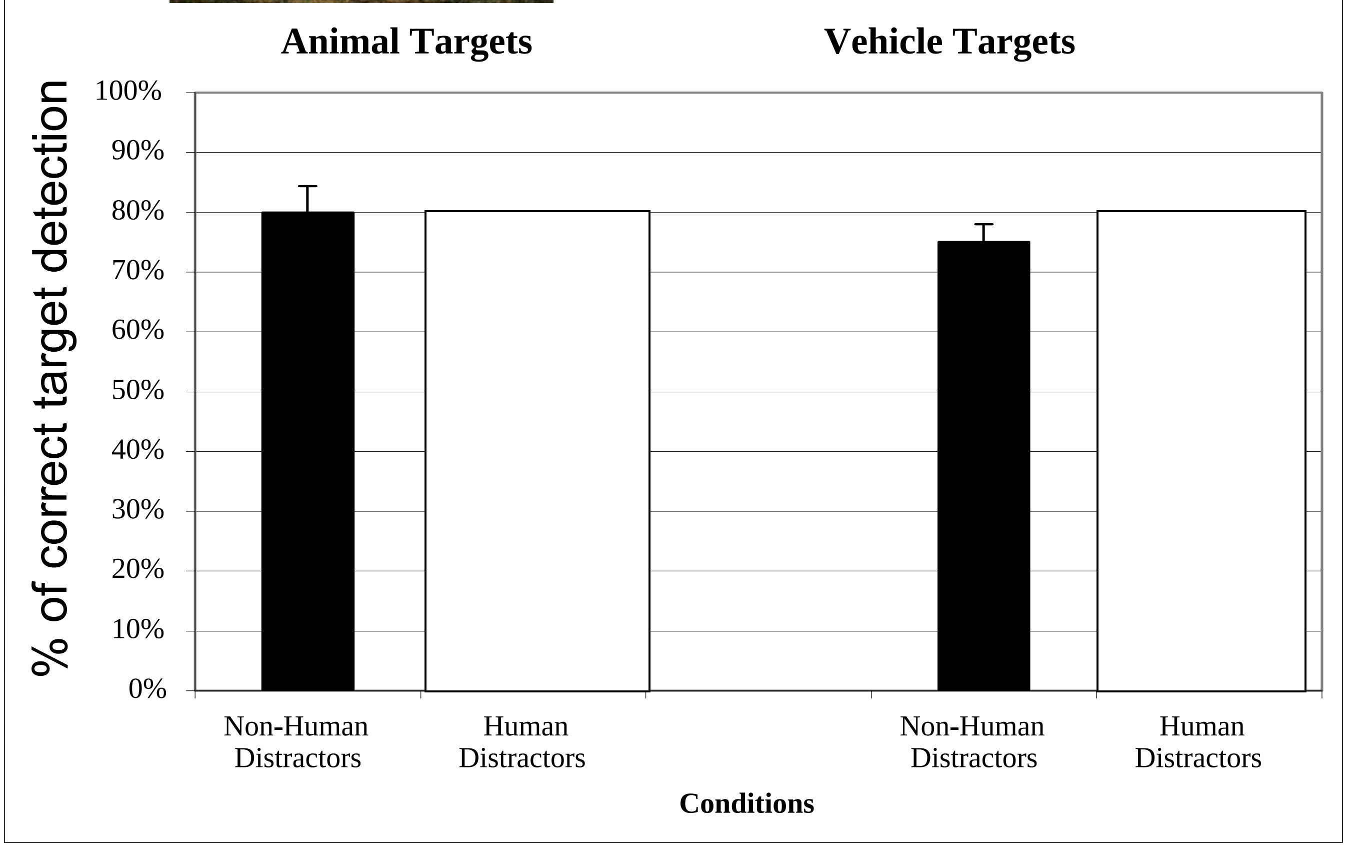
Evans & Treisman (2005): An RSVP task

Hypotheses: Performance should deteriorate when the distractors scenes share some of the *same features* with targets.

Is there an animal ? Is there a vehicle ?

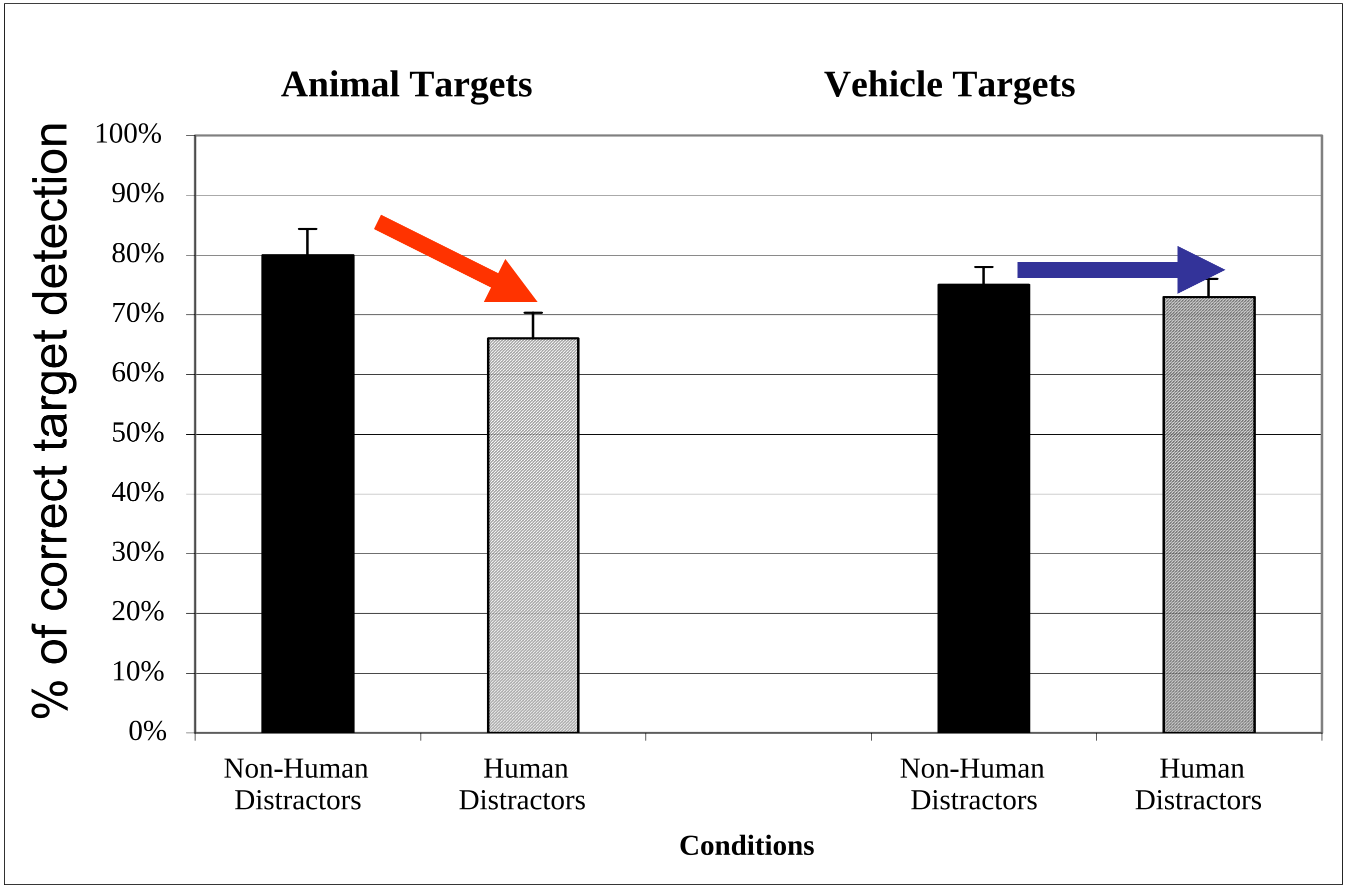


“People” were used as distractors for animal (target) and for vehicle (target)



Features set like parts of head, body, hair are shared between animals and Human: this level of information may help recognition of animals in previous studies

Evans & Treisman: Results

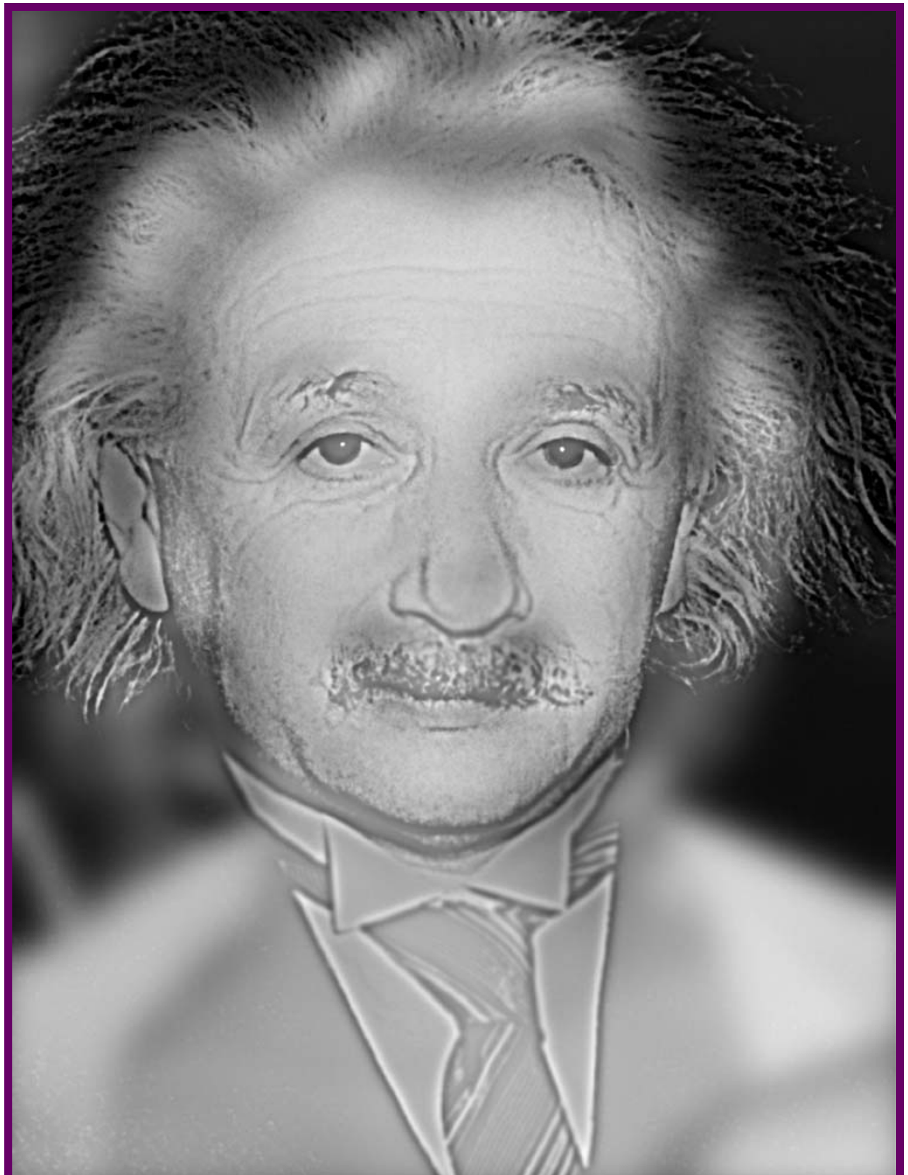


Features set like parts of head, body, hair are shared between animals and Human: this level of “part “information may help recognition of animals in previous studies

Scene Representation
Time course of visual information within a glance

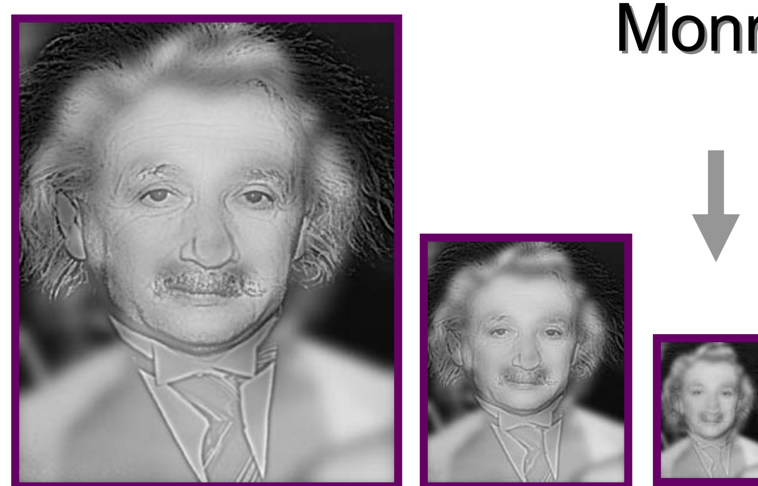
- Definition: what is the "gist"
- A few observations : getting the gist of a scene
- How do spatial frequency information unfold?
- What is the role of color ?
- What are the global properties of a scene?

Hybrid Images :
A method to study human image analysis



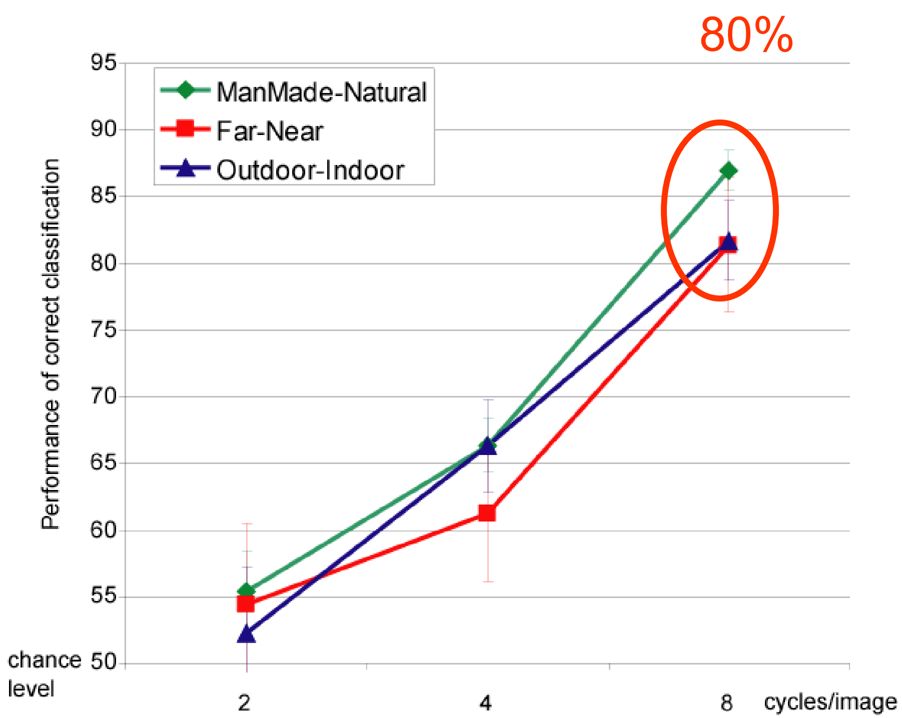
Albert
Einstein

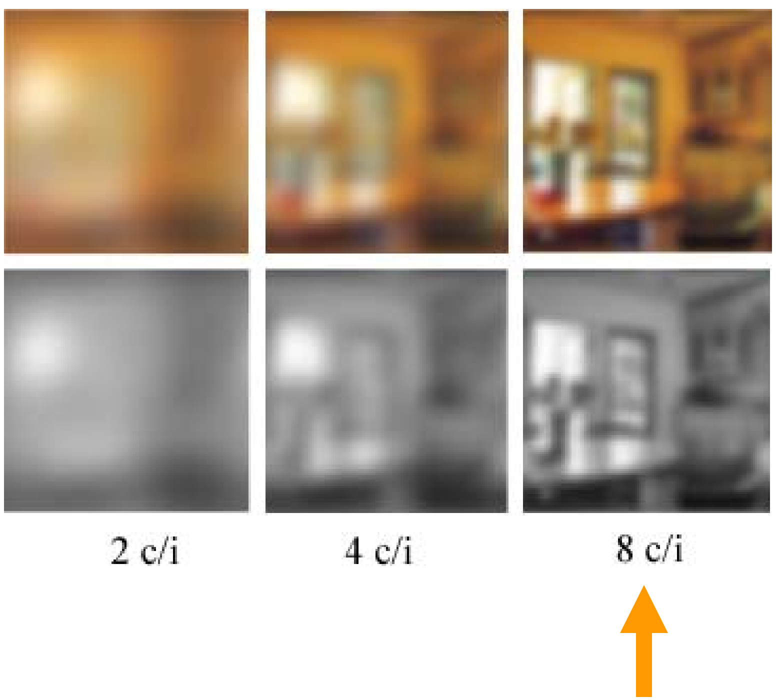
Marilyn
Monroe



Superordinate Classification

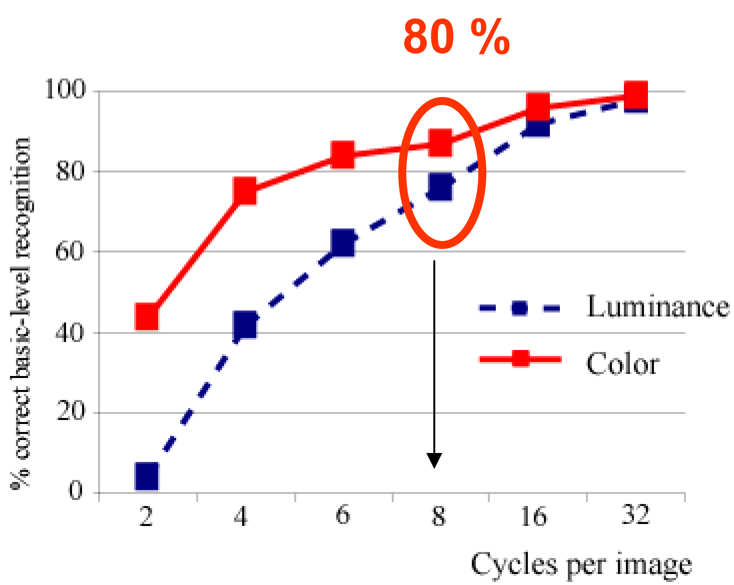
<u>Task</u>: Binary classification in **super-ordinate categories**.
<u>Result</u>: **80 %** of correct classification at a spatial resolution of **8 cycles / image** (image of 16 x 16 pixels size).

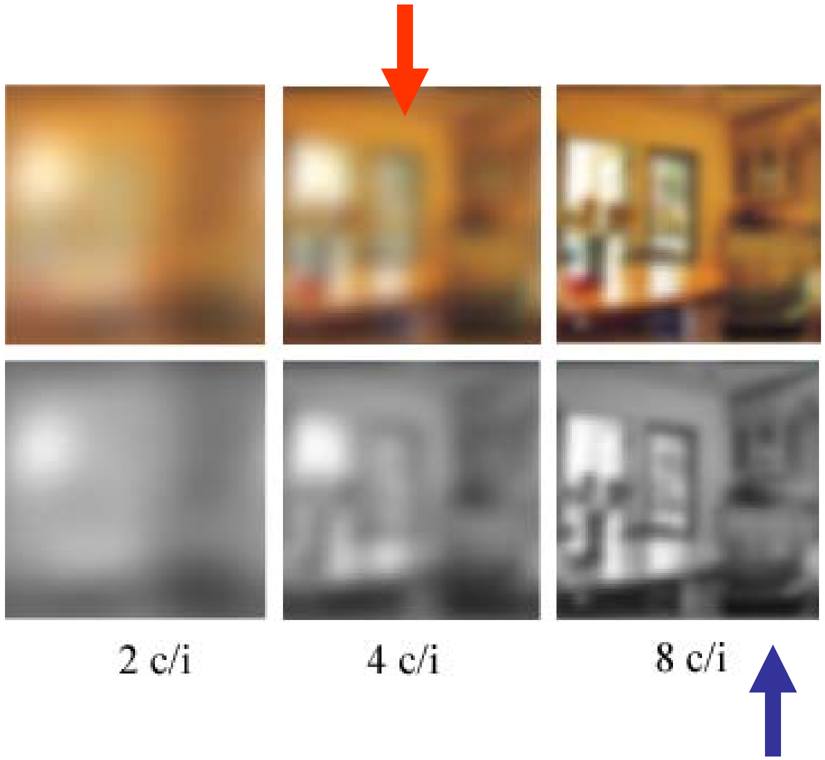




Scene Identification: Basic-Level

Task: Identify the basic-level category of the scene (scenes from 24 different semantic categories).
Result: **80 %** of correct classification at a spatial resolution of **8 cycles / image for grey-level scenes,** and at a resolution of **4 cycles/images for colored scenes**





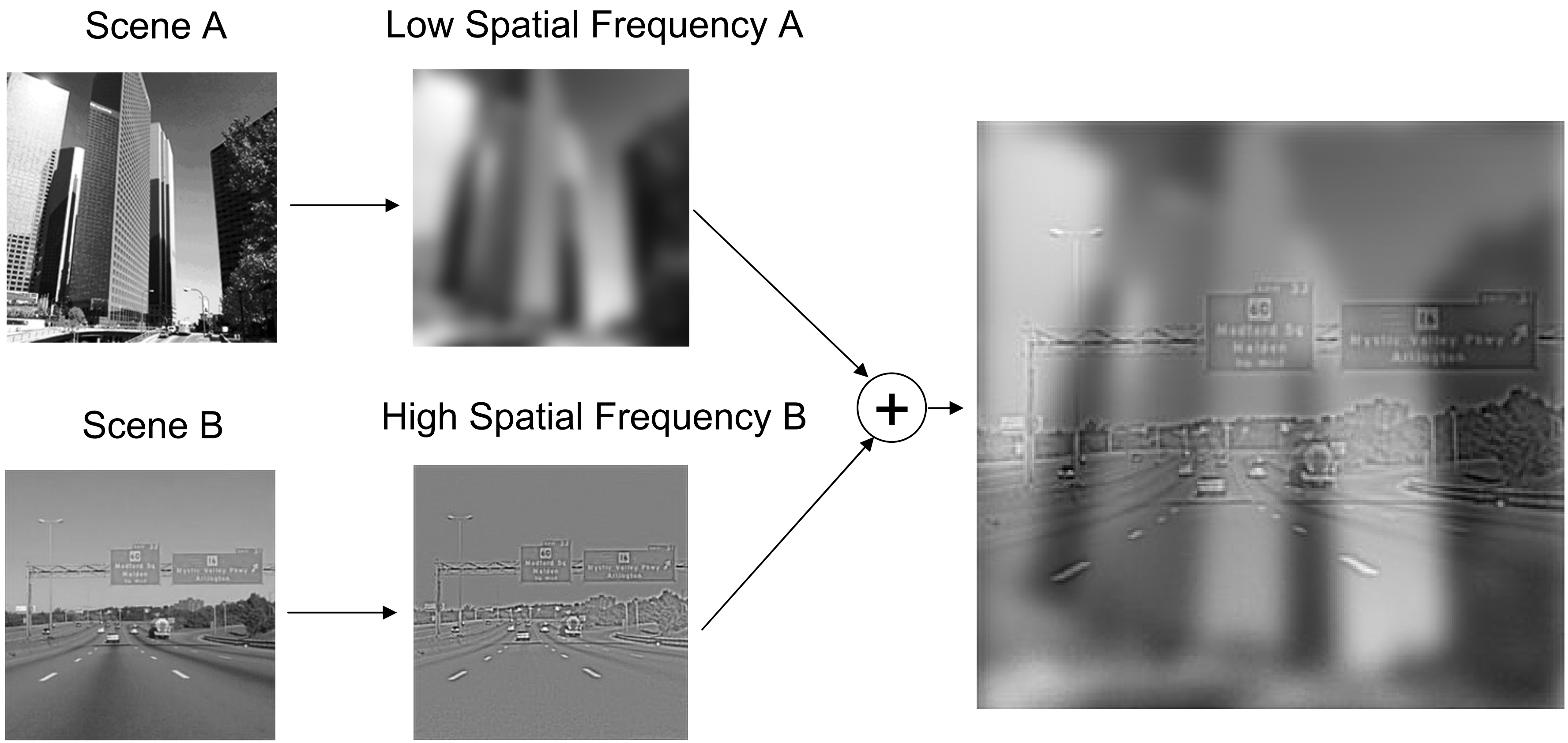
Oliva, A., & Schyns, P.G. (2000). Colored diagnostic blobs mediate scene recognition. *Cognitive Psychology*

Edges or Blobs ?

- Scenes can be identified at a superordinate and a basic-level with only coarse spatial layout (resolution of 4-8 cycles/image)
- At such a coarse spatial resolution, local object identity is not available
- Objects identity can be *inferred* after identifying the scene
- But … natural images are usually characterized by contours and our visual system encodes edges.
- What roles do "blobs" and "edges" play in fast scene recognition?

Torralba & Oliva, 2001

Hybrid Spatial Frequency Images

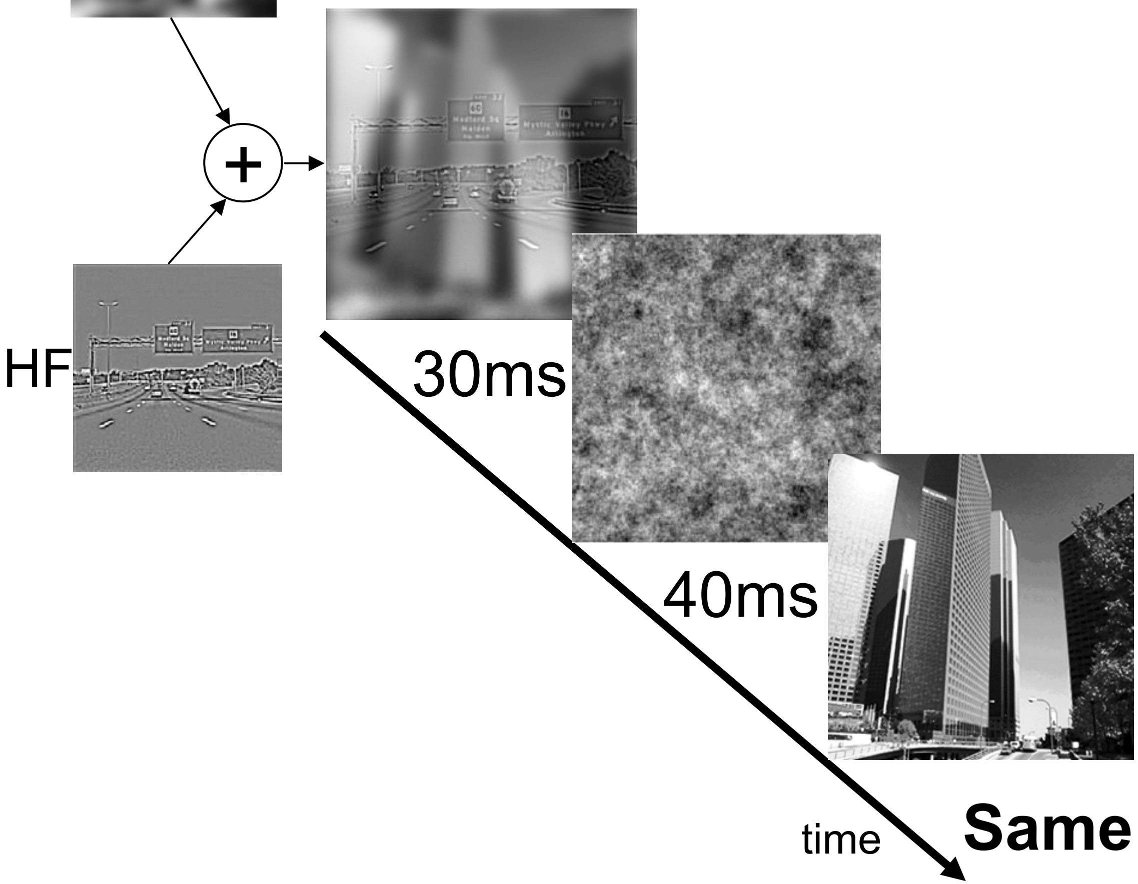


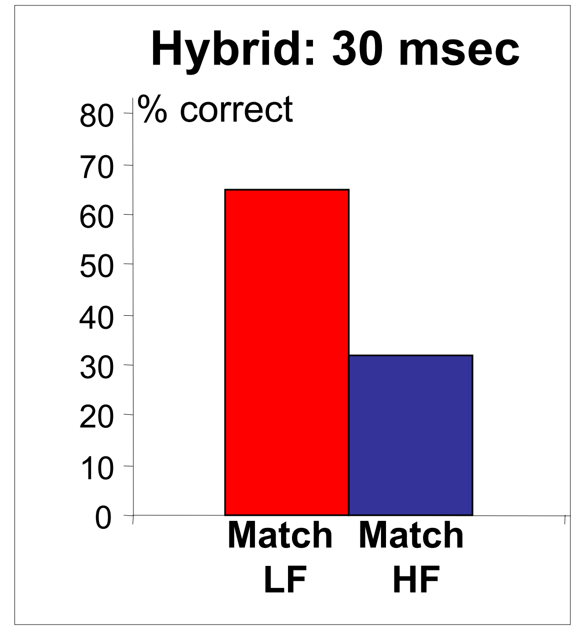
Hybrid images allow to study *concurrently* the roles of "blobs" and "edges" in fast scene recognition. Which information do we process first ?

Schyns & Oliva (1994, 1997), Oliva (1995), Oliva & Schyns (1997)

Exp 1: Detection Task

Subjects were not aware that images were hybrids.





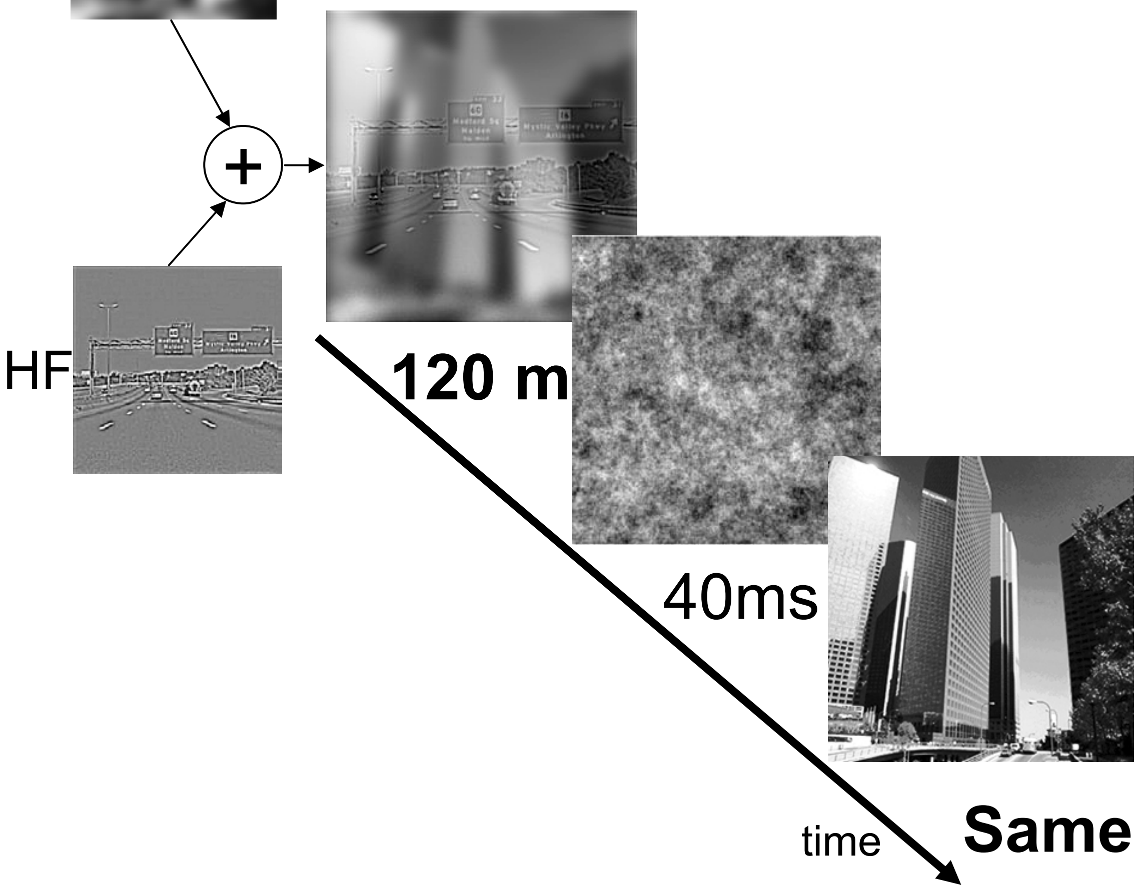
The second image can be:
- New image
- Match to LF
- Match to HF

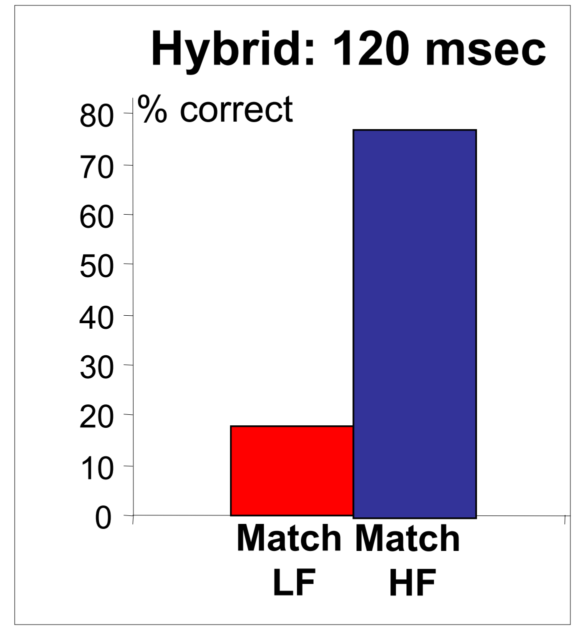
Same or different ?

Schyns & Oliva (1994). From blobs to boundary edges. Psychological Science.

Exp 1: Detection Task

Subjects were not aware that images were hybrids.





The second image can be:
- New image
- Match to LF
- Match to HF

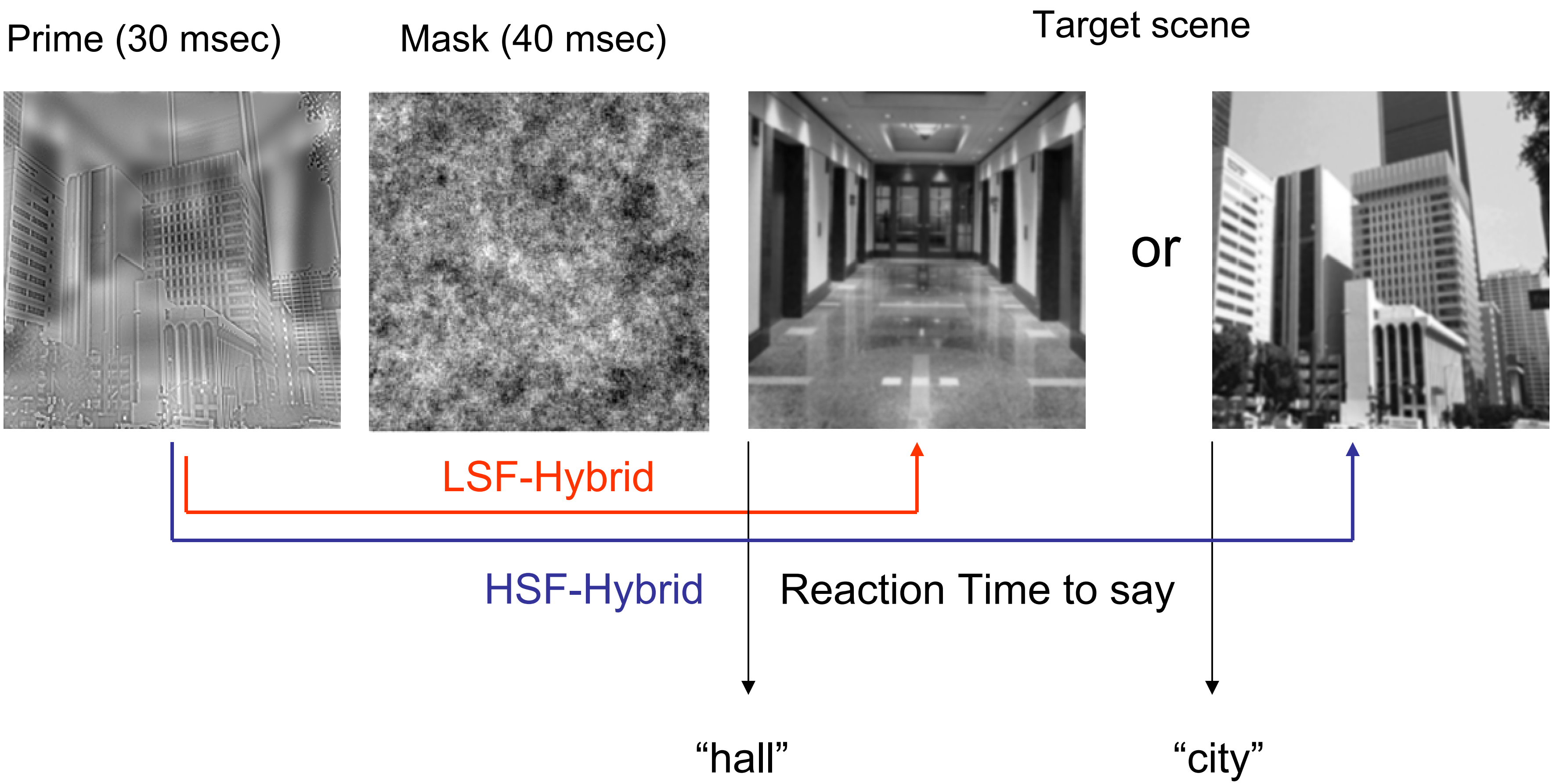
Same or different ?

Schyns & Oliva (1994)

Mandatory or Flexible Coarse to Fine?

- Within a glance, observers are using spatial scales in a *coarse to fine* manner.
- Is coarse-to-fine a mandatory process of visual scene processing or is it due to a task constraint? (i.e. identifying a scene under degraded conditions).
- Are all spatial scales available at the beginning of the visual processing (30 msec of stimulus duration)?
- If so, the brief presentation of one hybrid scene should successfully help the recognition of two scenes.

Exp 2: Naming Task

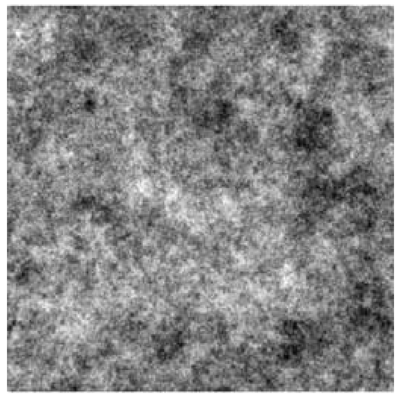


Oliva & Schyns (1997). Blobs or boundary edges. Cognitive Psychology.

Exp 2: Naming Task

Prime (30 msec)

Mask (40 msec)



Target scene

or

Unrelated pair

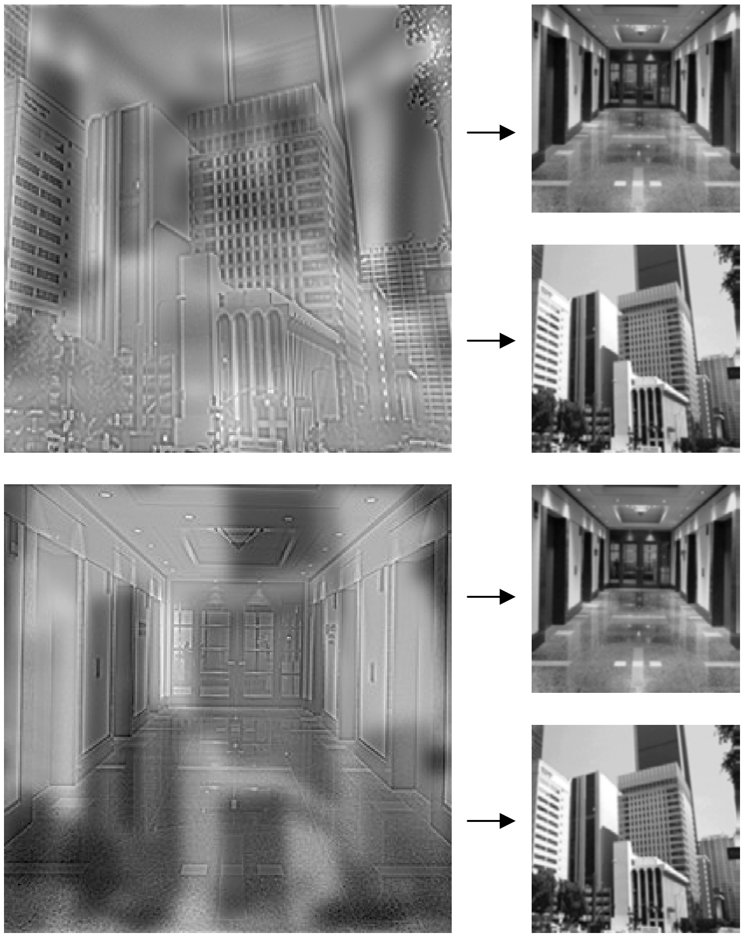
Reaction Time to say

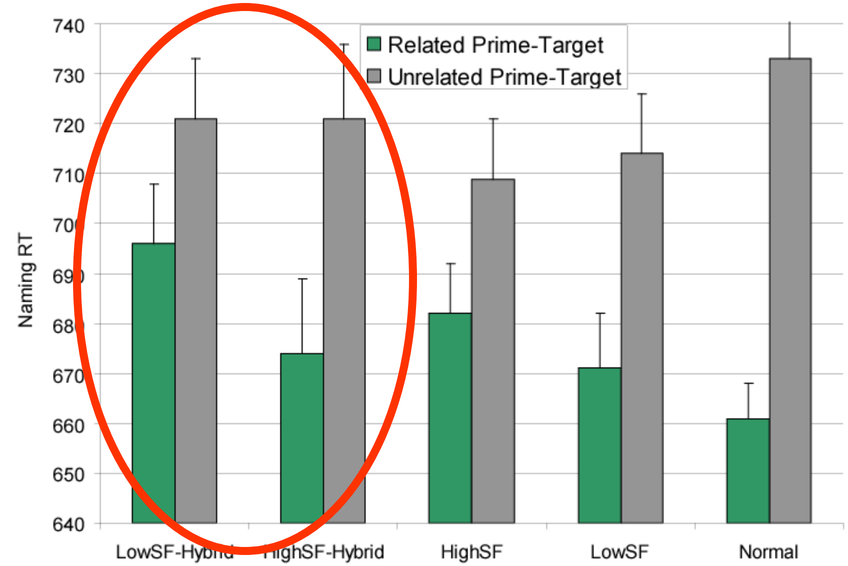
“hall”

“city”

Oliva & Schyns (1997). Blobs or boundary edges. Cognitive Psychology.

Experiment 2: Results





Both Low and High SF seem to be available very early in the visual processing (30 msec of exposure).

Oliva & Schyns (1997). Blobs or boundary edges. Cognitive Psychology.

Spatial Scales Scene Processing

- Spatial resolution around 8 cycles/image are sufficient for recognizing most of scenes at a basic-level category
- Object identification is not a requirement for scene identification
- All spatial scales information available very early (30 msec) in the temporal dynamics of natural image recognition

- What about the role of color in fast scene recognition?

Oliva, A. (2005). Gist of a scene. In *Neurobiology of Attention*. Eds. L. Itti, G. Rees and J. Tsotsos. Academic Press, Elsevier.

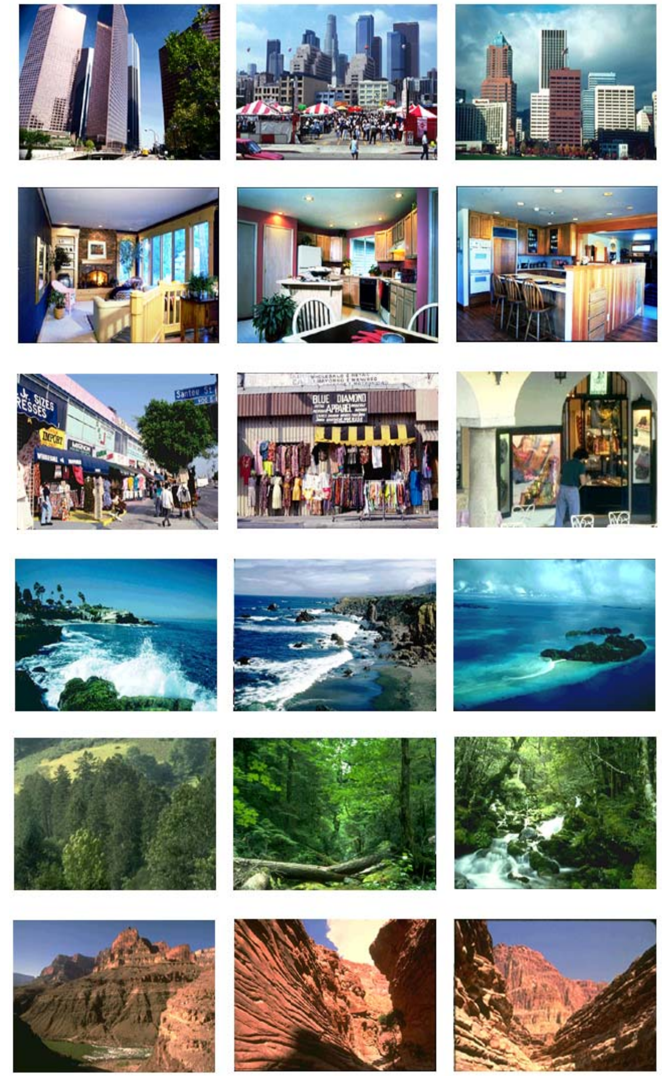
Color *Diagnosticity*

Man-made categories: no specific colour mode

Natural categories: specific and distinctive colour modes

Hypothesis:

- When color is a feature *diagnostic* of the meaning of a scene, altering color information should impair recognition.

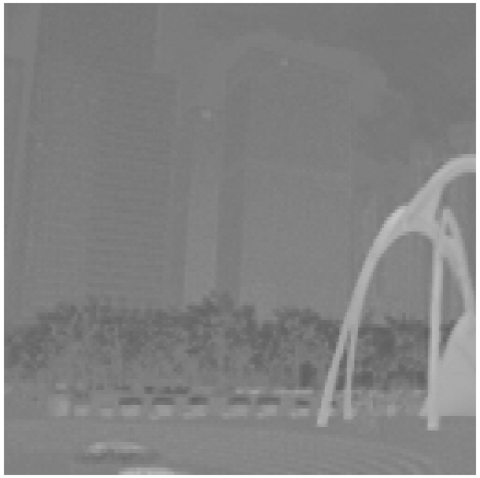


Oliva & Schyns (2000). Colored diagnostic blobs mediate scene recognition. Cognitive Psychology.

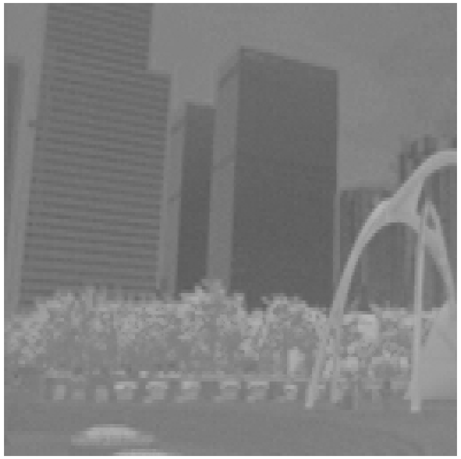
R G B space -> L*a*b*

Lab

Luminance

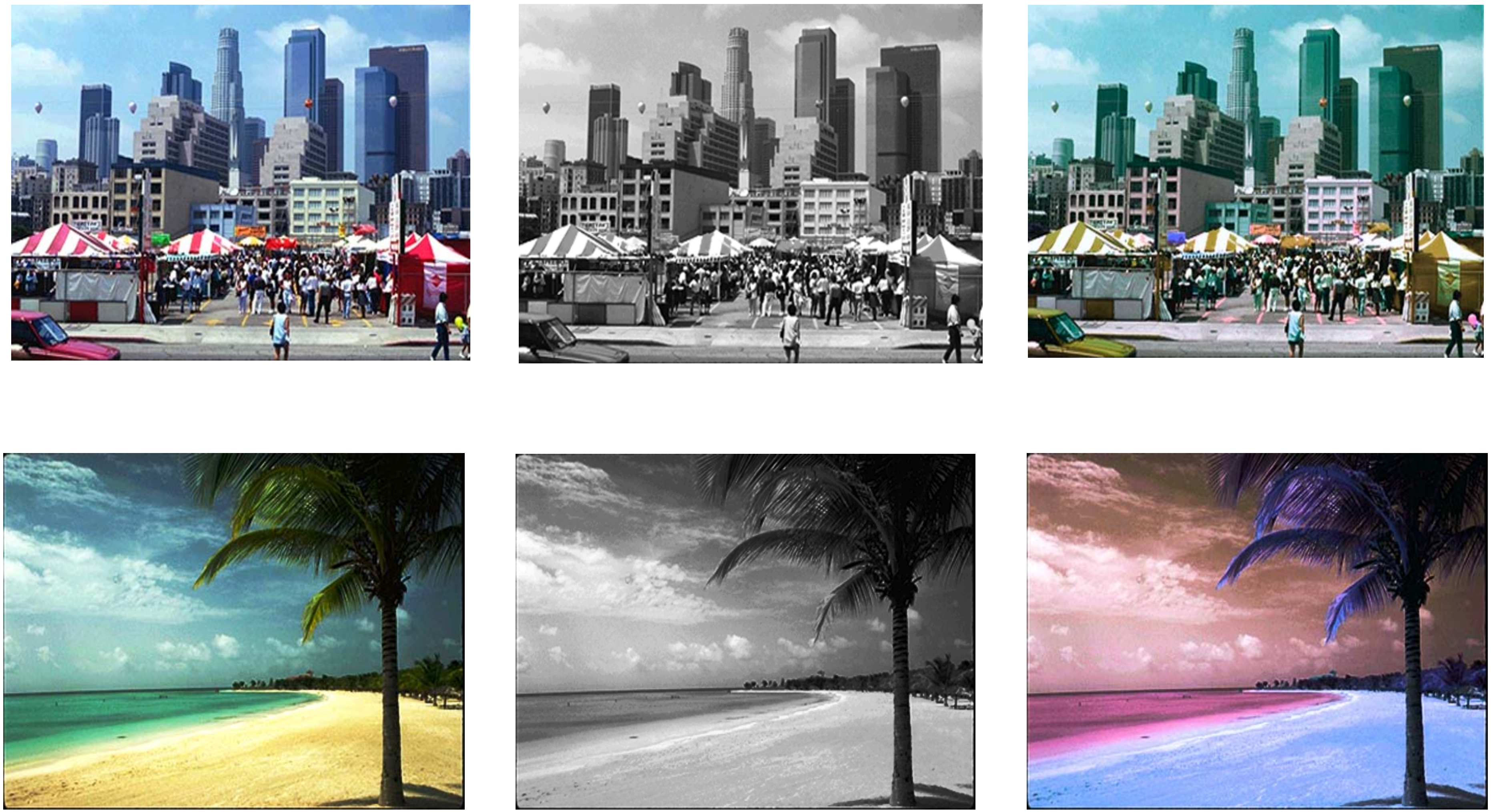


a (red - green)



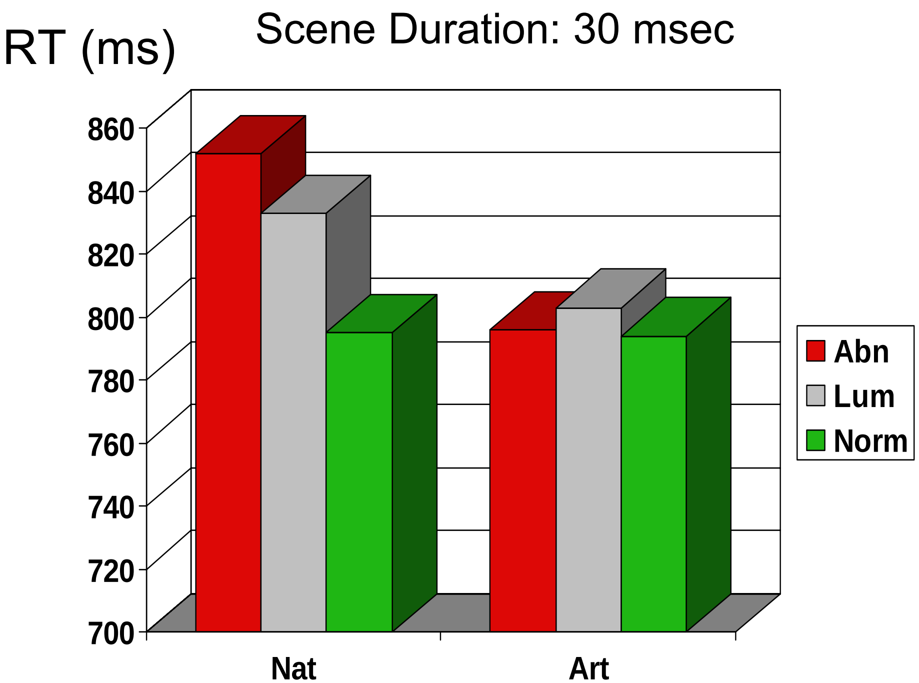
b (yellow - blue)

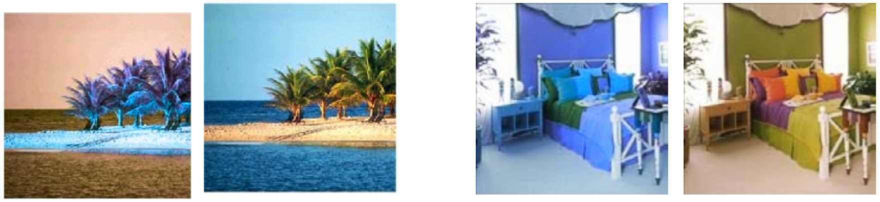
Examples of Stimuli



Normal color | Luminance | Abnormal Color

The role of Diagnostic color

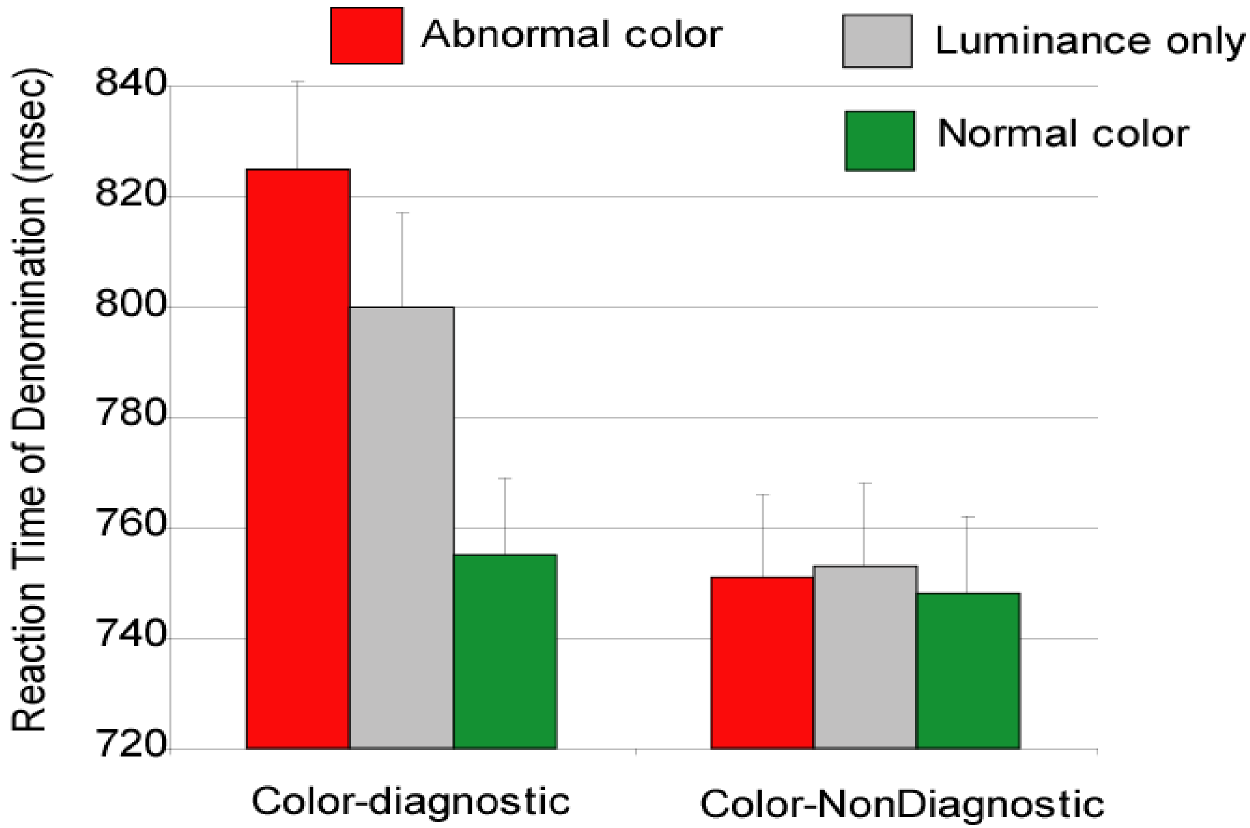


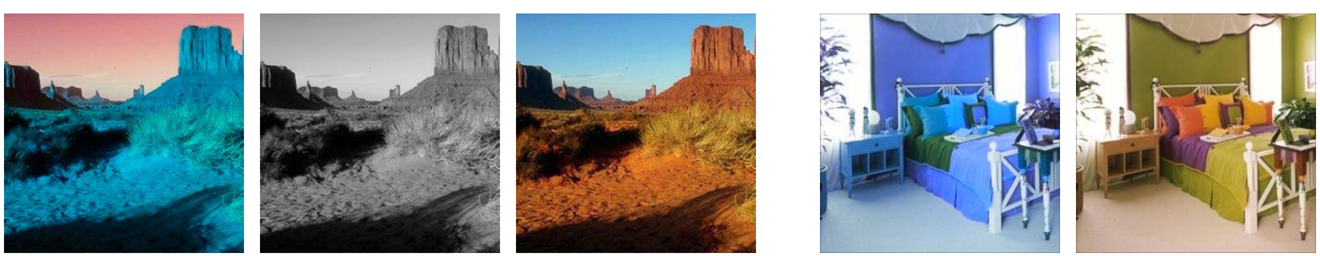


- Color helps scene identification but only when it is a diagnostic feature of the scene category

Oliva & Schyns (2000). Colored diagnostic blobs mediate scene recognition. Cognitive Psychology.

The role of *diagnostic* color

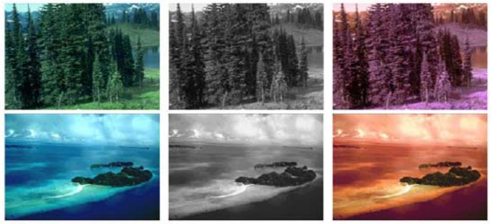


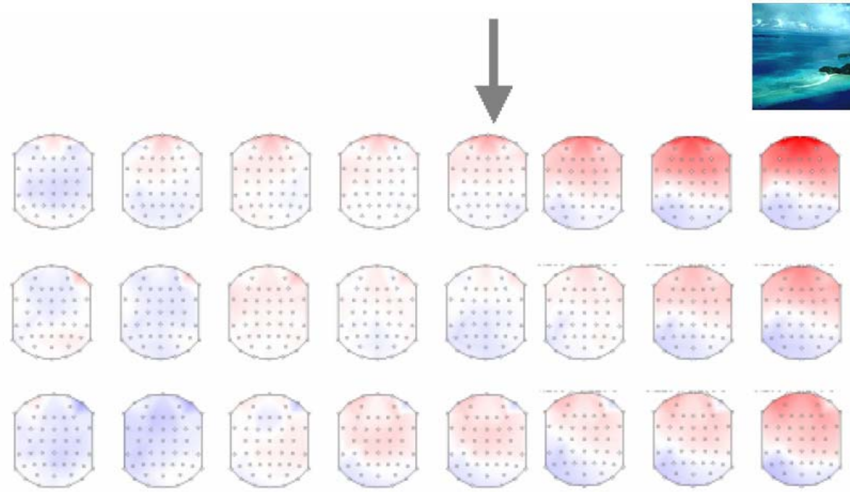


Oliva & Schyns (2000). Colored diagnostic blobs mediate scene recognition. Cognitive Psychology.

The role of Color & Brain Signals

Diagnostic colors contribute to early stages of scene recognition





Significant frontal differential activity for Normal Colored Scenes (vs. gray and abnormal colors) **150 msec** after image onset

Goffaux, V., Jacques, C., Mouraux, A., Oliva, A., Rossion, B., & Schyns. P.G. (2005). *Visual Cognition*.

Scene Representation
Time course of visual information within a glance

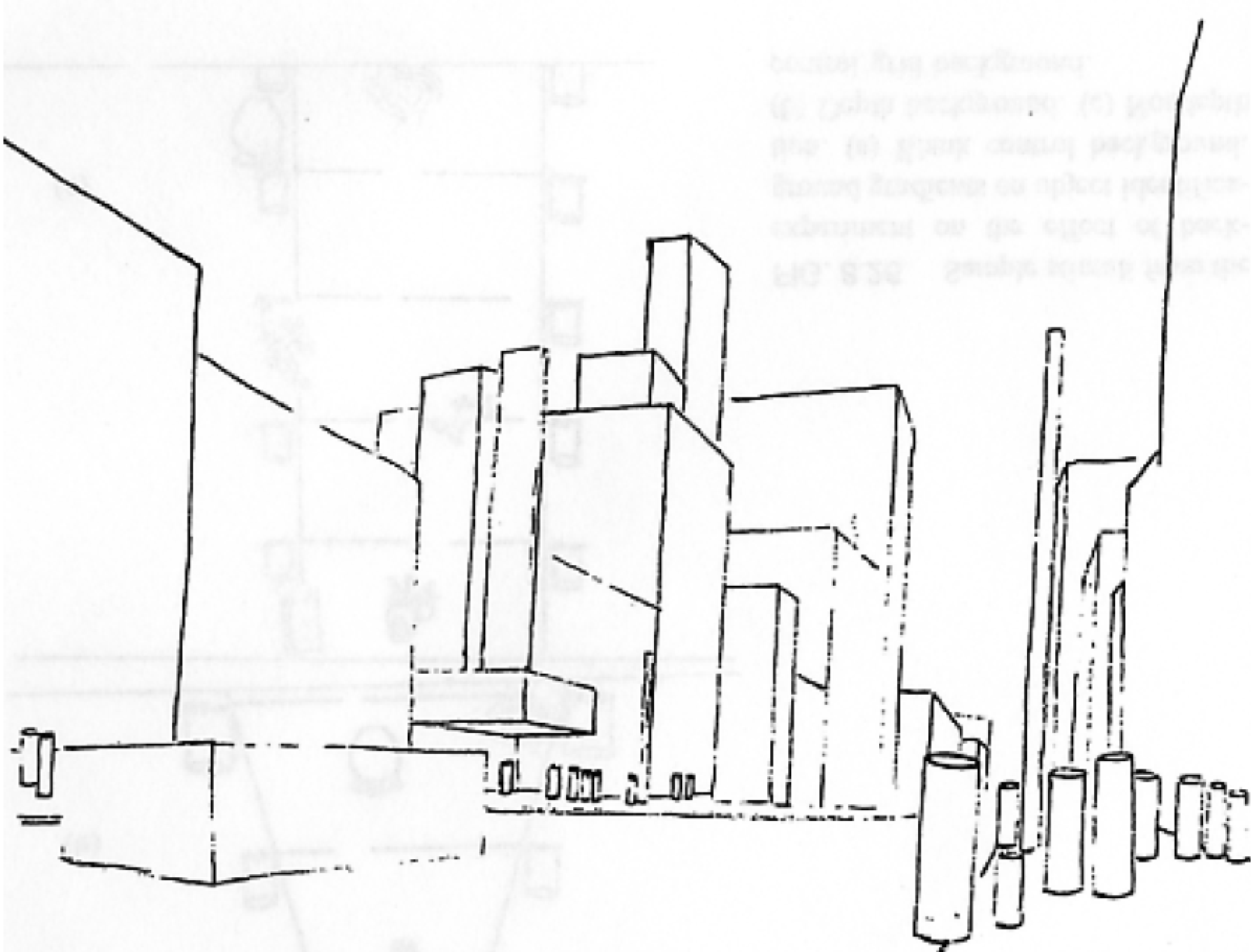
Some simple features are correlated
with scene recognition

What are the other properties of a scene image
that could help "recognition" (gist)?

Reducing the objects
Enhancing the scene

Reducing the objects

Enhancing the scene & global/configural processing



Irving Biederman

Forest Before Trees: The Precedence of Global Features in Visual Perception
Navon (1977)

How do we recognize the forest in the first place?

Navon (1977) says:

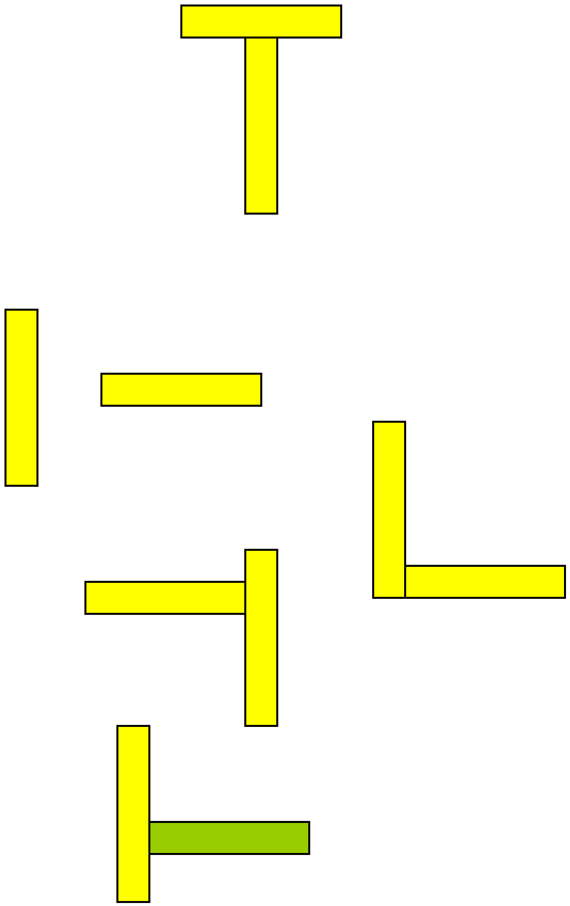
- “No attempt was made here to formulate an operational definition of globality of visual features which enables precise predictions about yhe course of perception of real-world scenes.

- What is suggested in this paper is that whatever the perceptual units are, the spatial relationship among them is more global than the structure within them (and so forth if the hierarchy is deeper).

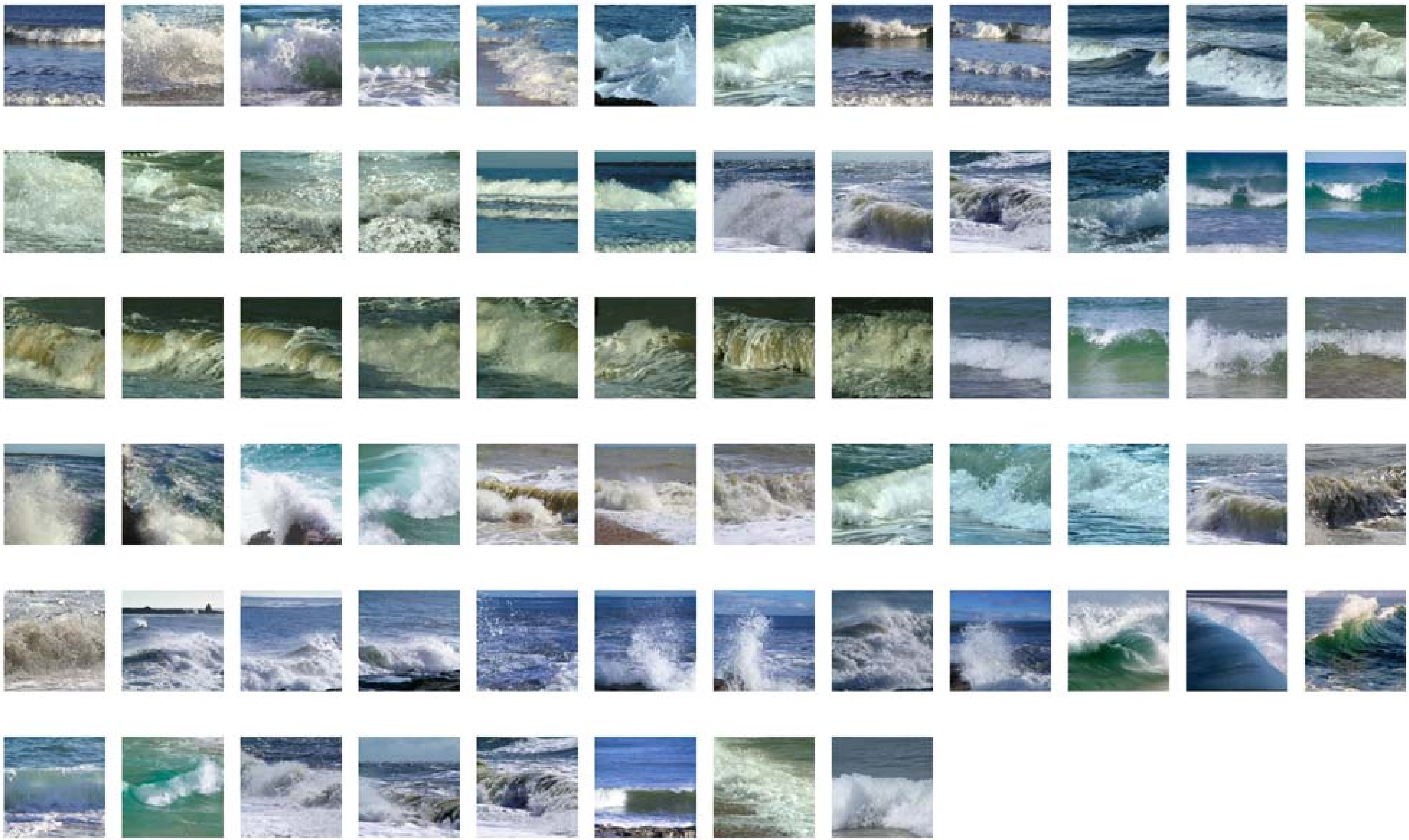
- Thus, I am afraid that clear-cut operational measures for *globality* will have to patiently await the time that we have a better idea of **how a scene is decomposed into perceptual units. “**

What are the perceptual units ☺

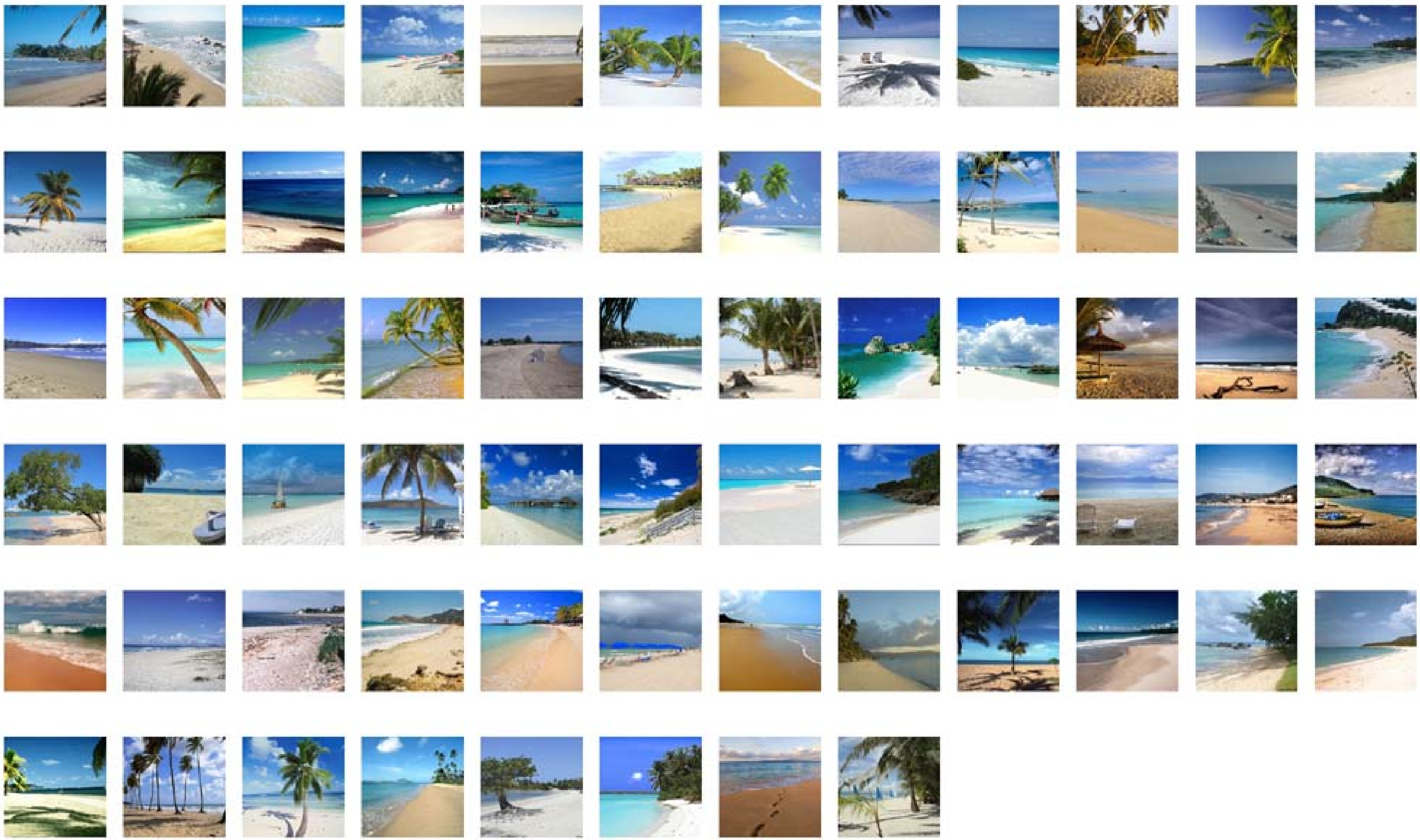
What are the perceptual units ?



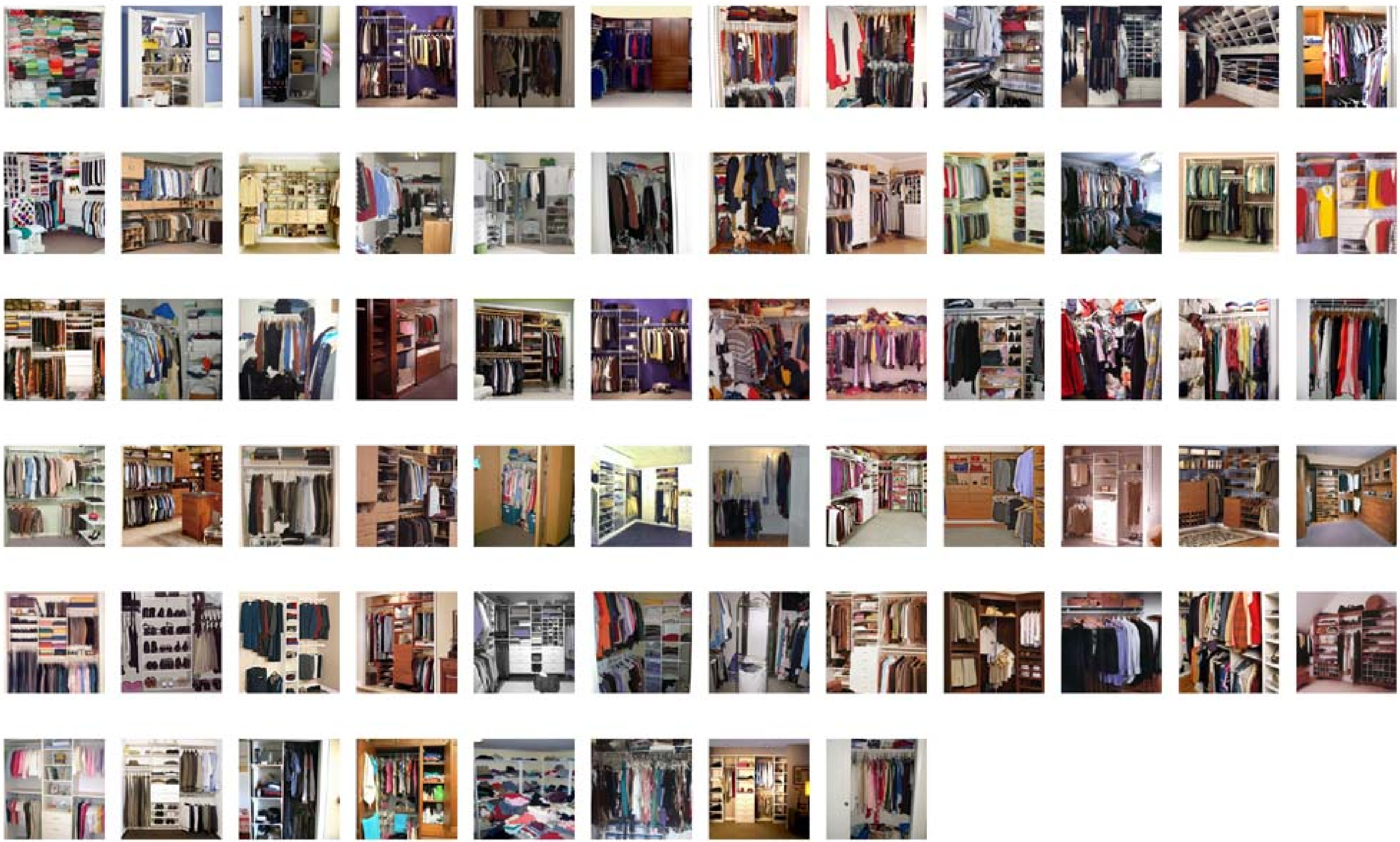
Waves ~ Texture



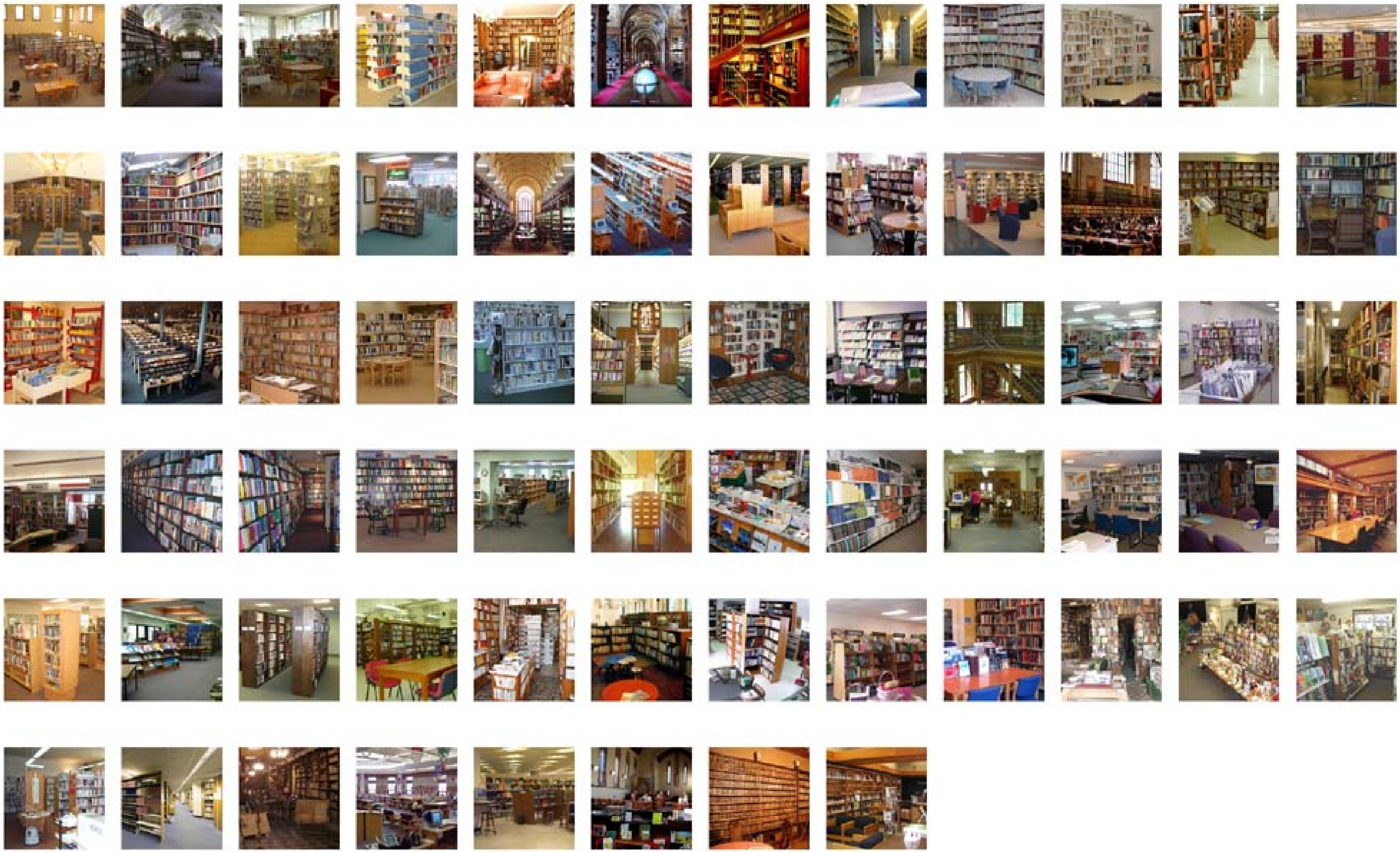
Beach



Closet

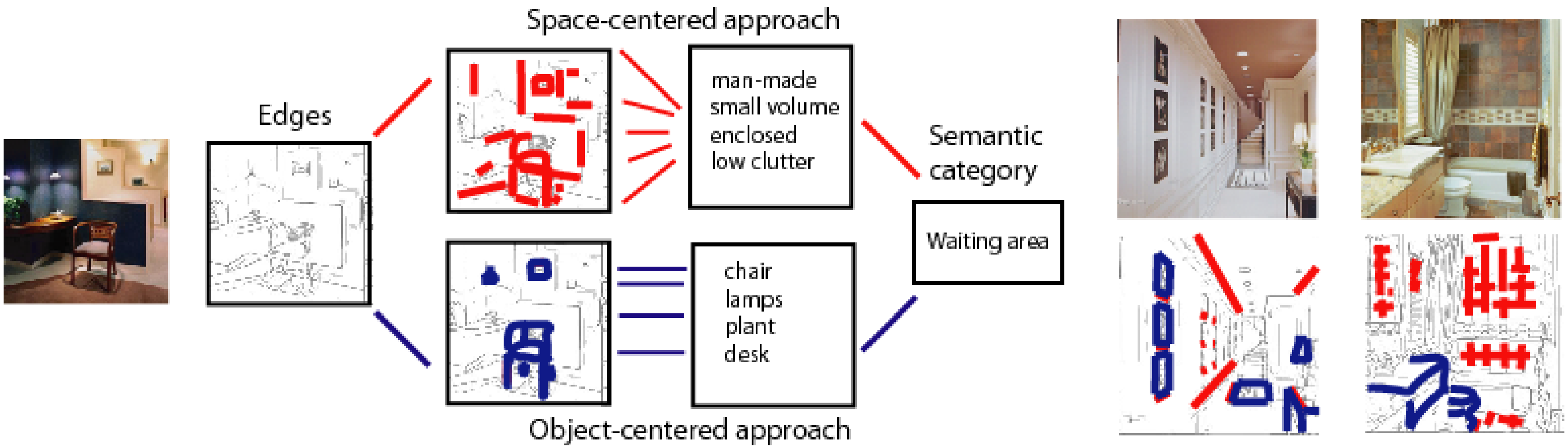


Library



Scene Identification: Basis ?

Scene-Centered Approach



A scene-centered approach proposes another representation of scene information, that is independent of object recognition stages (object-centered approach).

A scene-centered approach does not require the use objects as an intermediate representation. The structure of a scene can be represented by perceptual properties of space and volume (e.g. mean depth, perspective, symmetry, clutter).

Oliva & Torralba (2001). International Journal of Computer Vision. Torralba & Oliva (2002). PAMI.
Oliva & Torralba (2002). 2nd Workshop on Biologically Motivated Computer Vision.

Part-based approach: e.g. *objects*

If you knew the identity of all the objects in a scene, recognition would be perfect

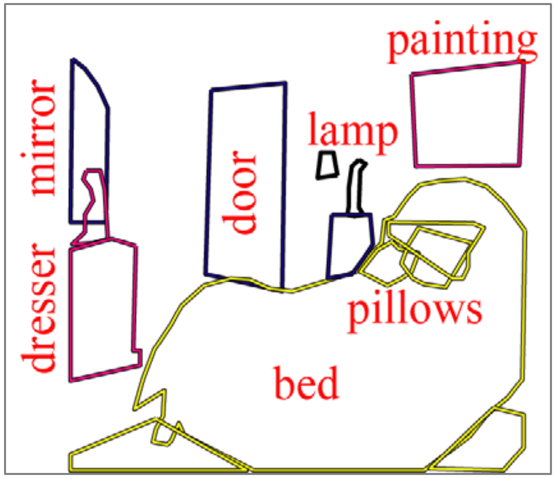
	Bathroom	Bedroom	Conf. room	Dining room	Kitchen	Living room	Office
Bathroom	99				03		
Bedroom		99	02				03
Conference room		02	98	01			07
Dining room			01	98	02	03	03
Kitchen	03			02	99		
Living room				03		99	01
Office		03	07	03		01	97

Labelme: a vector of the list of all objects for each image

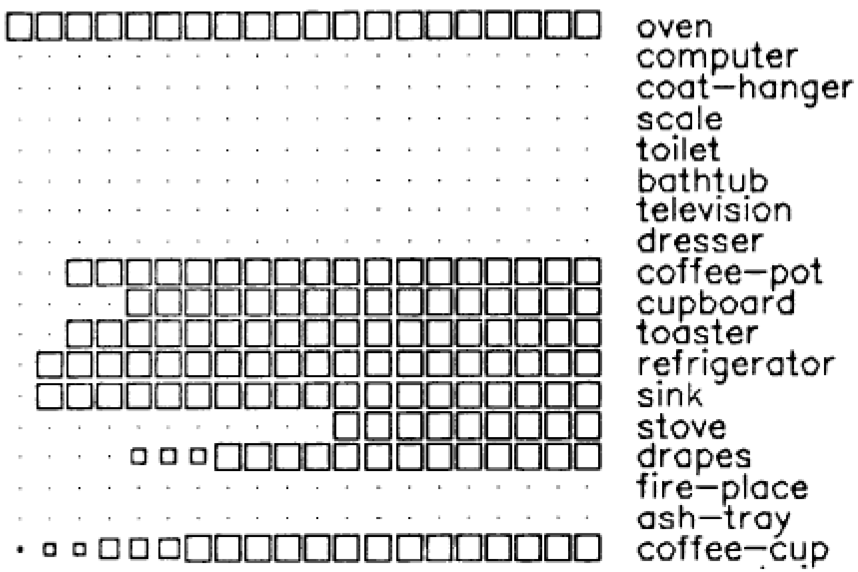
Oliva et al. 2006

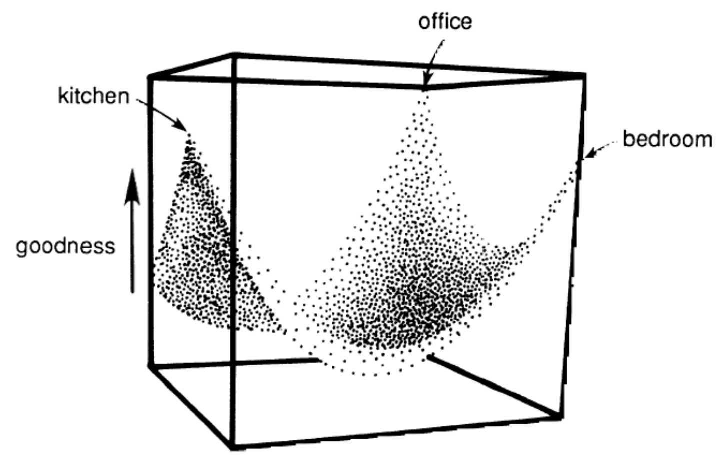
Part-based approach: e.g. *objects*

- Scenes as collections of objects has always been very popular:
  - Schemas (Bartlett; Piaget; Rumelhart)
  - Scripts (Schank)
  - Frames (Minsky)



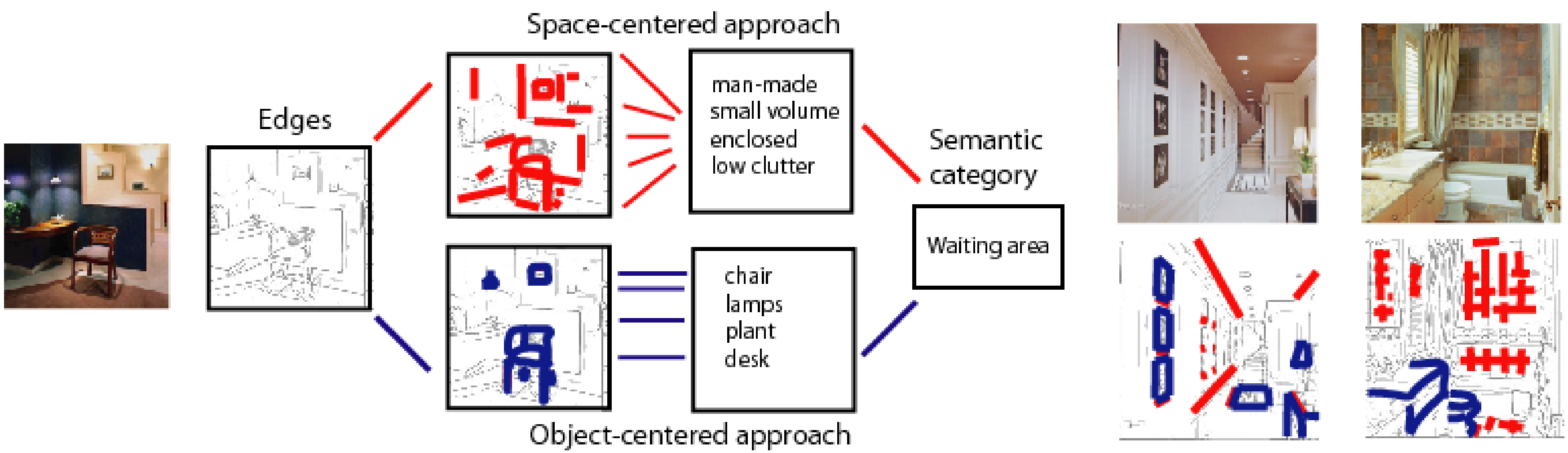
Part-based approach: e.g. *objects*





Rumelhart et al. 1986

Scene-Centered Approach

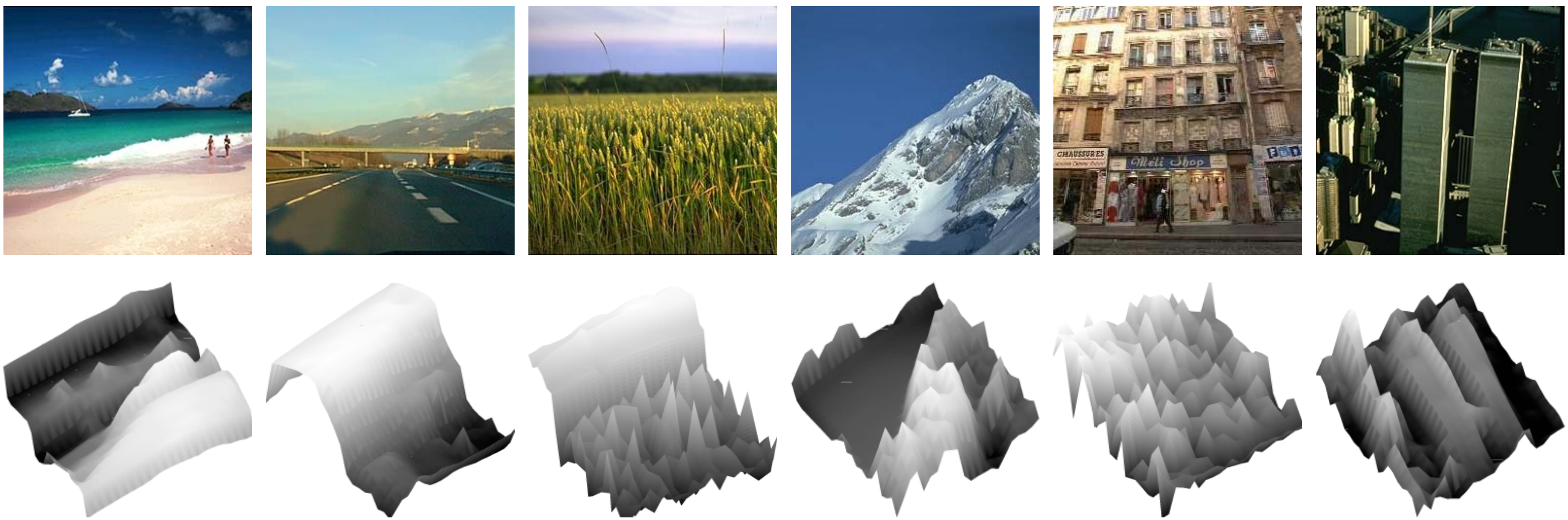


A scene-centered approach proposes another representation of scene information, that is independent of object recognition stages (object-centered approach).

A scene-centered approach does not require the use objects as an intermediate representation. The structure of a scene can be represented by perceptual properties of space and volume (e.g. mean depth, perspective, symmetry, clutter).

Oliva & Torralba (2001). International Journal of Computer Vision. Torralba & Oliva (2002). PAMI.
Oliva & Torralba (2002). 2nd Workshop on Biologically Motivated Computer Vision.

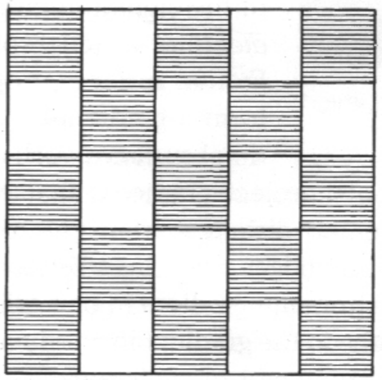
Holistic approach: global surface properties

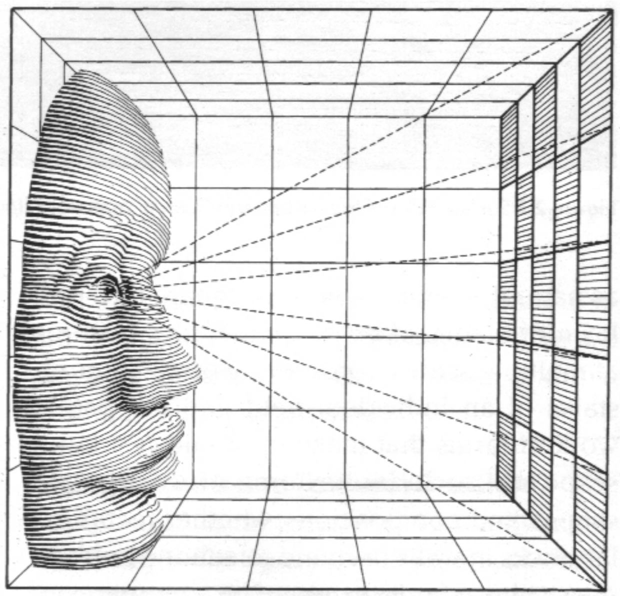


A scene is a single surface that can be represented by global descriptors

Oliva & Torralba (2001)

Textural Signatures of Visual Scenes
"Flat frontal surface"

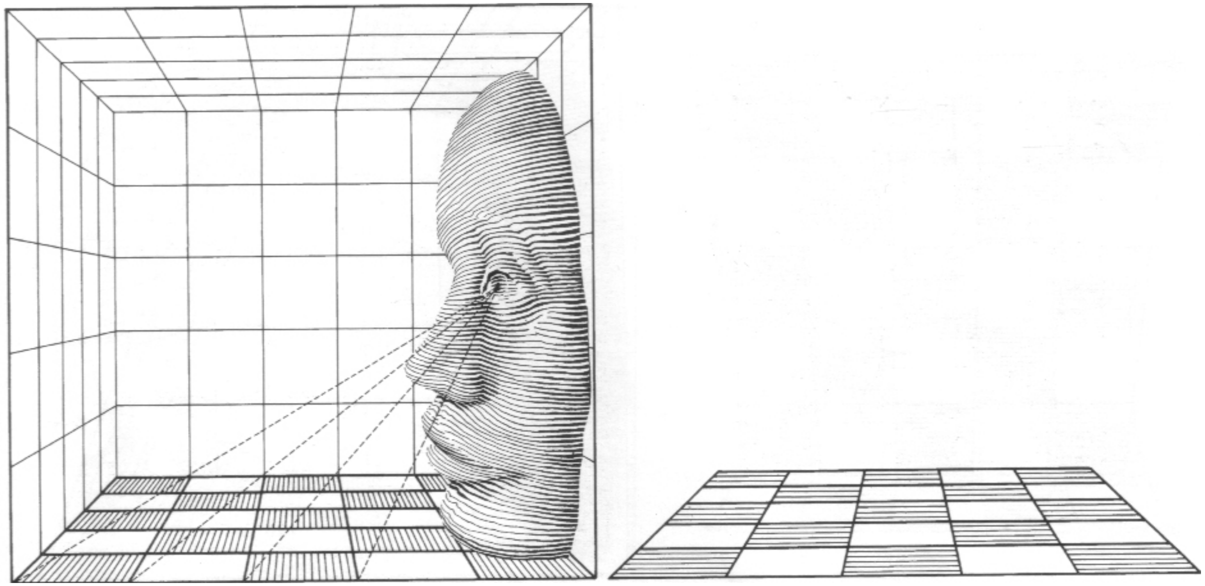




A flat frontal surface projects an array of stimuli on the retina whose gradient (interval between stimuli) is constant

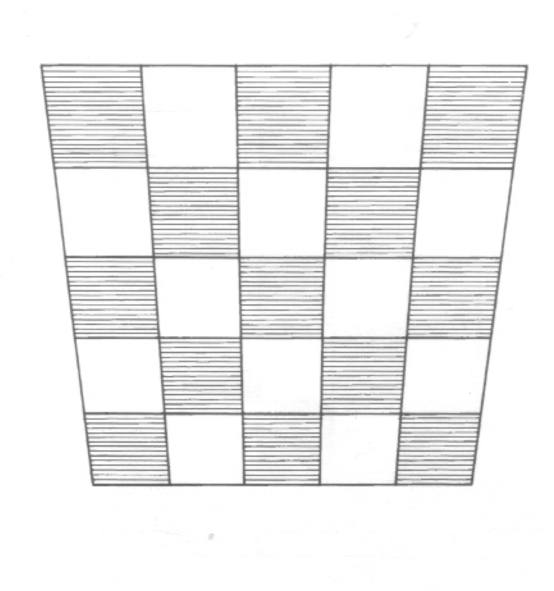
J J Gibson

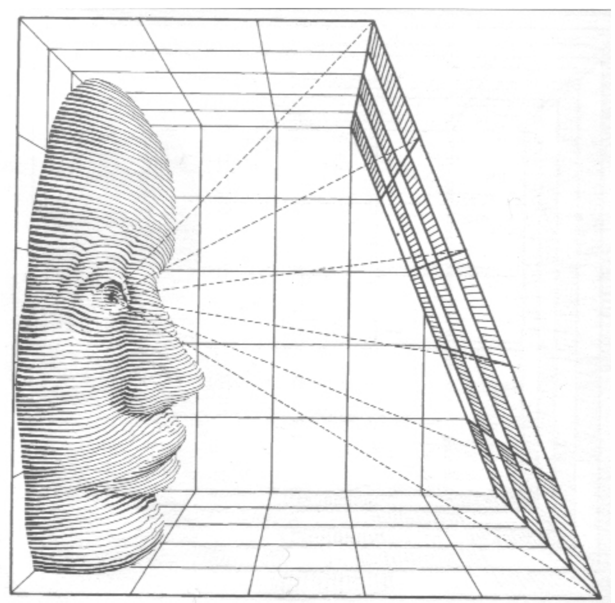
Textural Signatures of Visual Scenes
"Flat longitudinal surface"



A flat longitudinal surface projects an array of stimuli on the retina whose gradient decreases and nears the center of the retina with increasing distance from the observer

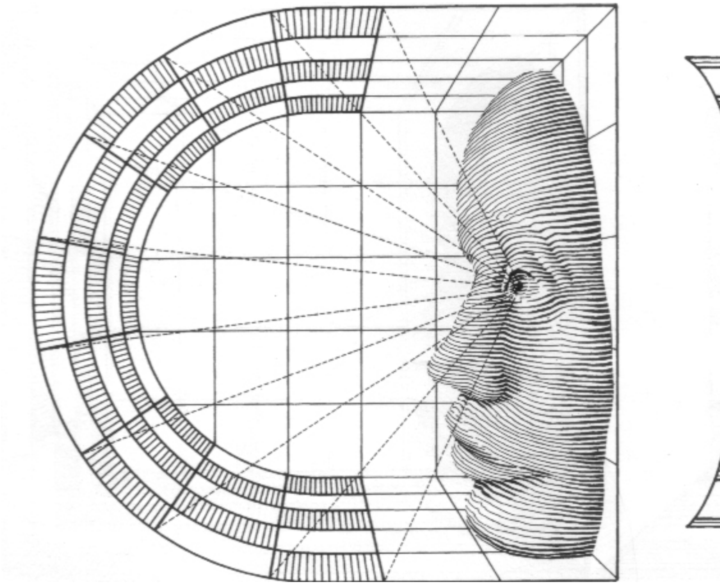
Textural Signatures of Visual Scenes
"Flat slanting surface"

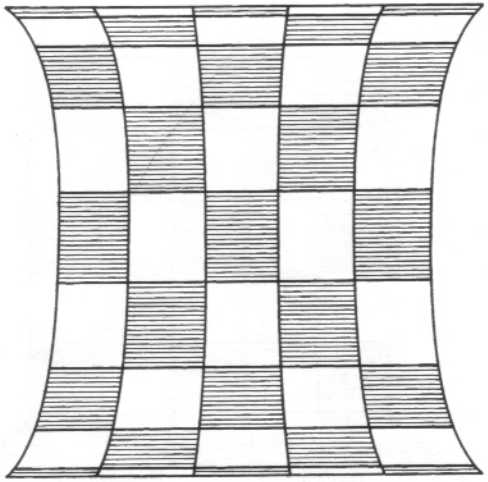




A flat slanting surface projects an array of stimuli on the retina whose gradient decreases and nears the center of the retina either more or less rapidly than that of a longitudinal surface.

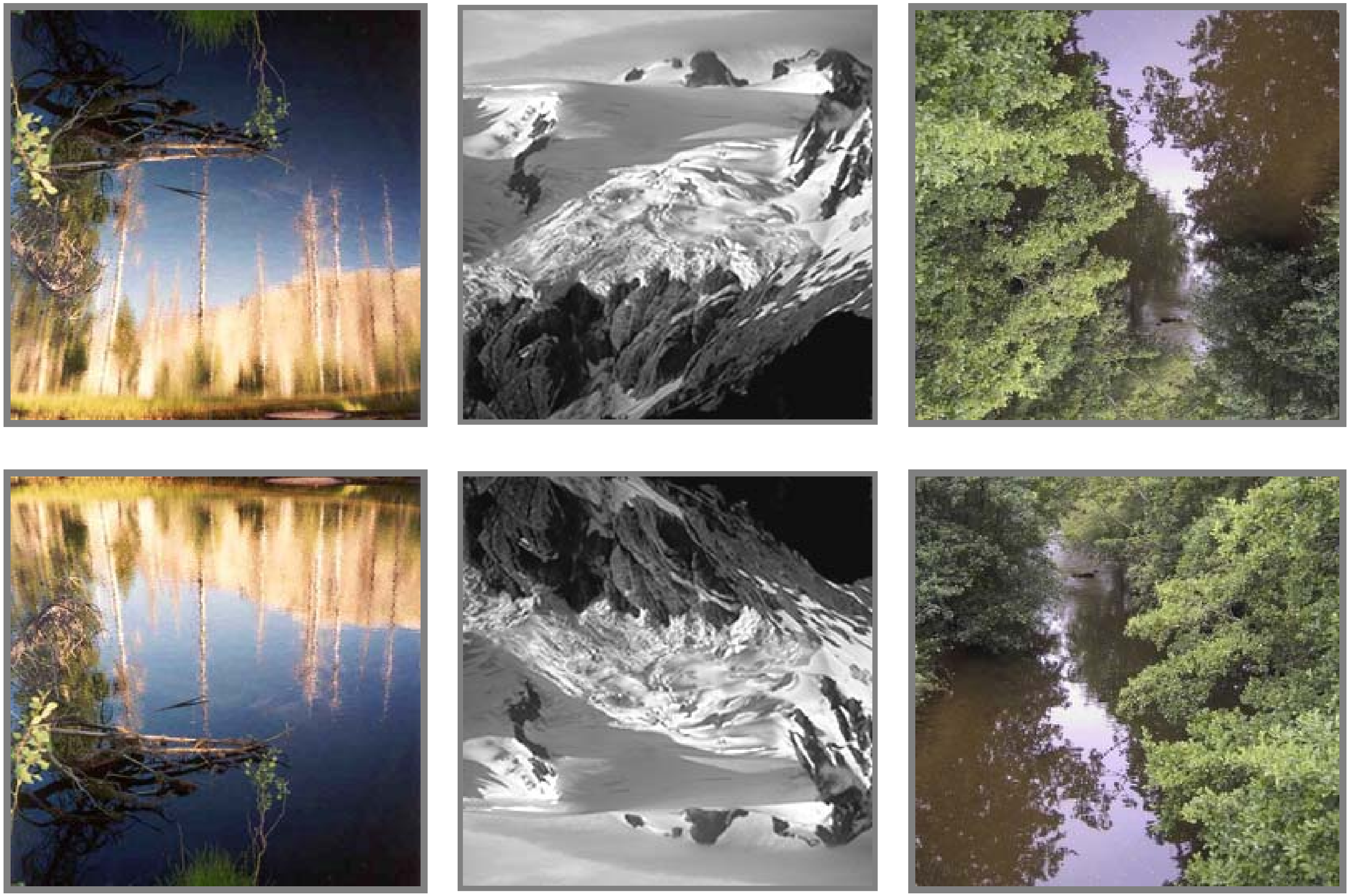
Textural Signatures of Visual Scenes
"A rounded surface"





A rounded surface projects an array of stimuli on the retina whose gradient Changes from small to large to small as the surface curves from a longitudinal to a frontal and back to a longitudinal attitude relative to the observer.

Textured surface layout influences depth perception



Torralba & Oliva (2002, 2003)

Statistical Regularities of Scene Volume

When increasing the size of the space, natural environment structures become larger and smoother.

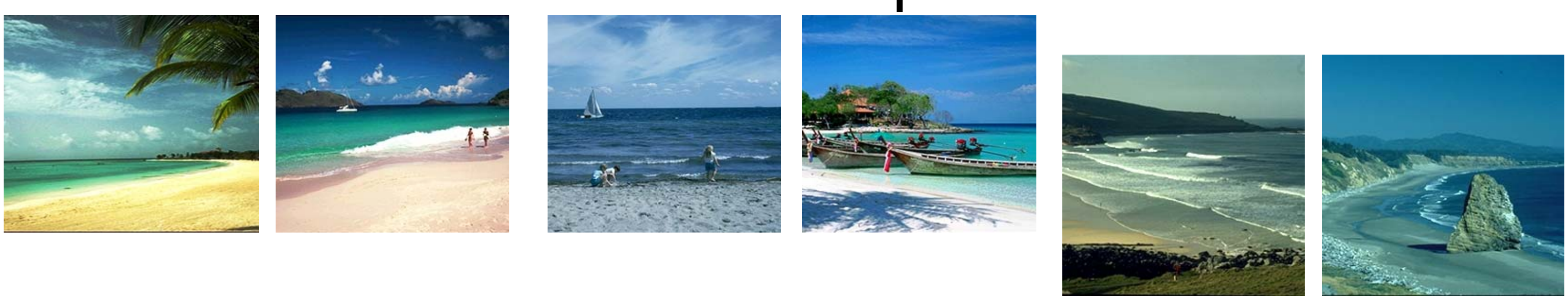
For man-made environments, the clutter of the scene increases with increasing distance: close-up views on objects have large and homogeneous regions. When increasing the size of the space, the scene "surface" breaks down in smaller pieces (objects, walls, windows, etc).

Torralba & Oliva. (2002). Depth estimation from image structure. IEEE Pattern Analysis and Machine Intelligence

Hints of Globality: Spatial Structure

Forests are “enclosed”

Beaches are “open”



"Agnosic" human scene representation: How far can we go with it ?

A lake

Scene-Centered Representation

100% natural space
66% open space
64% perspective
74% deep space
68% cold place

ect-Cente
esentatio

23 sky
35 % ter
18% tre
2 % mount
% gra

Spatial Envelope Theory

As a scene is inherently a 3D entity, initial scene recognition might be based on properties *diagnostic of the space* that the scene subtends and not necessarily the objects the scene contains

"Street"




Degree of clutter, openness, perspective, roughness, etc …

Oliva et al (1999); Oliva & Torralba (2001, 2002, 2006); Torralba & Oliva (2002,2003); Greene & Oliva (2006, in revision)

Spatial Envelope Representation

Global Properties diagnostic of the space the scene subtends provide the basic level of the scene

(1) Boundary of the space

Mean depth

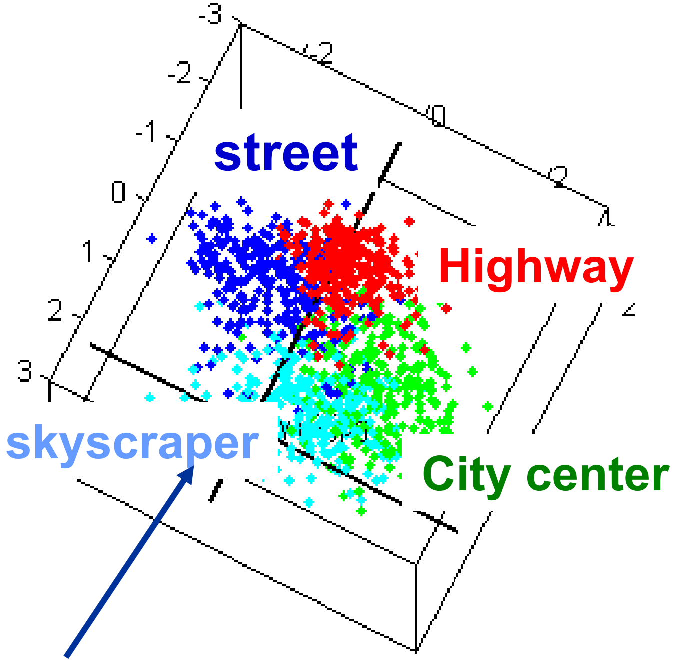
Openness

Perspective

(2) Content of the space

Naturalness

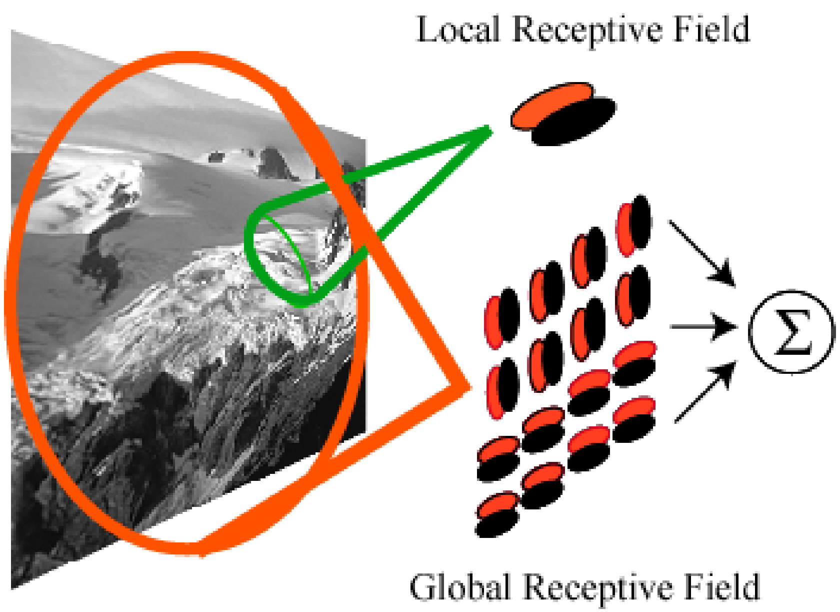
Roughness



Degree of Openness

Given human ranking of how *open to enclosed* a given scene image is, the goal is to find the low level features that are correlated with "openness"

From open scenes to closed scenes



High degree of Openness

Lack of texture

Low spatial frequency horizontal

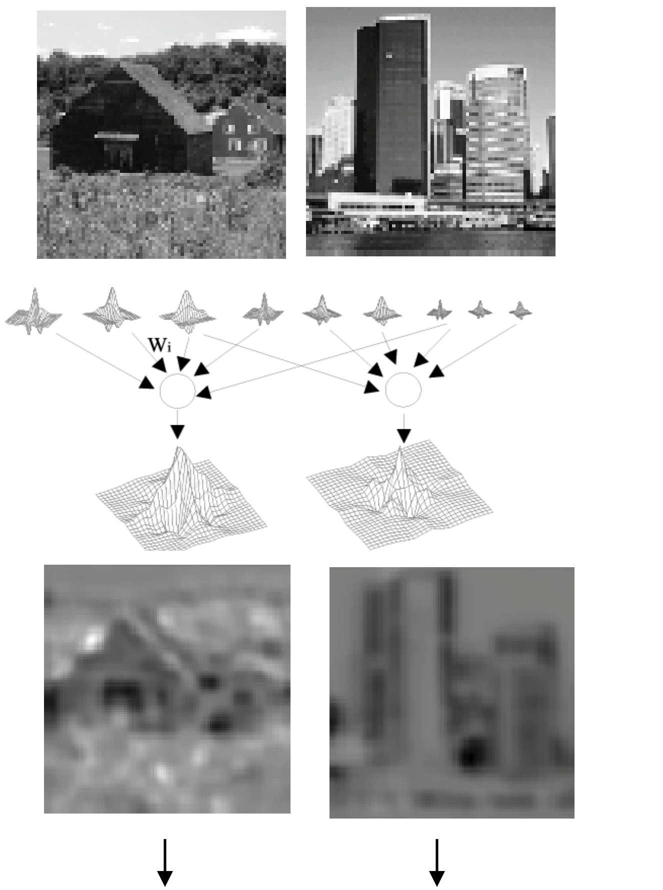
High spatial frequency isotropic texture

Oliva & Torralba (2001, 2006)

Global Scene Property: Openness

Global scene properties can be estimated by a combination of low level features

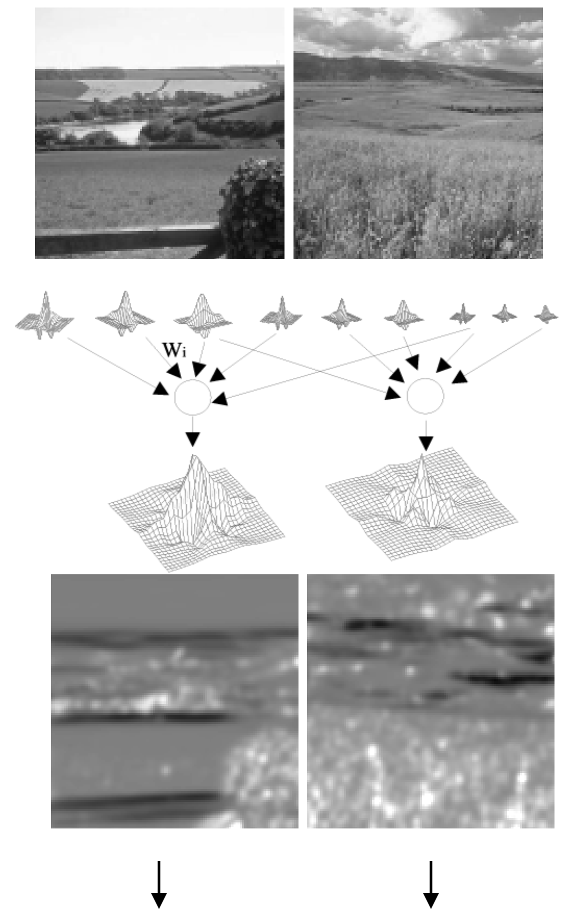
Diagnostic features of Naturalness



Medium level of naturalness

Low level of naturalness (man-made environment)

Diagnostic features of Openness

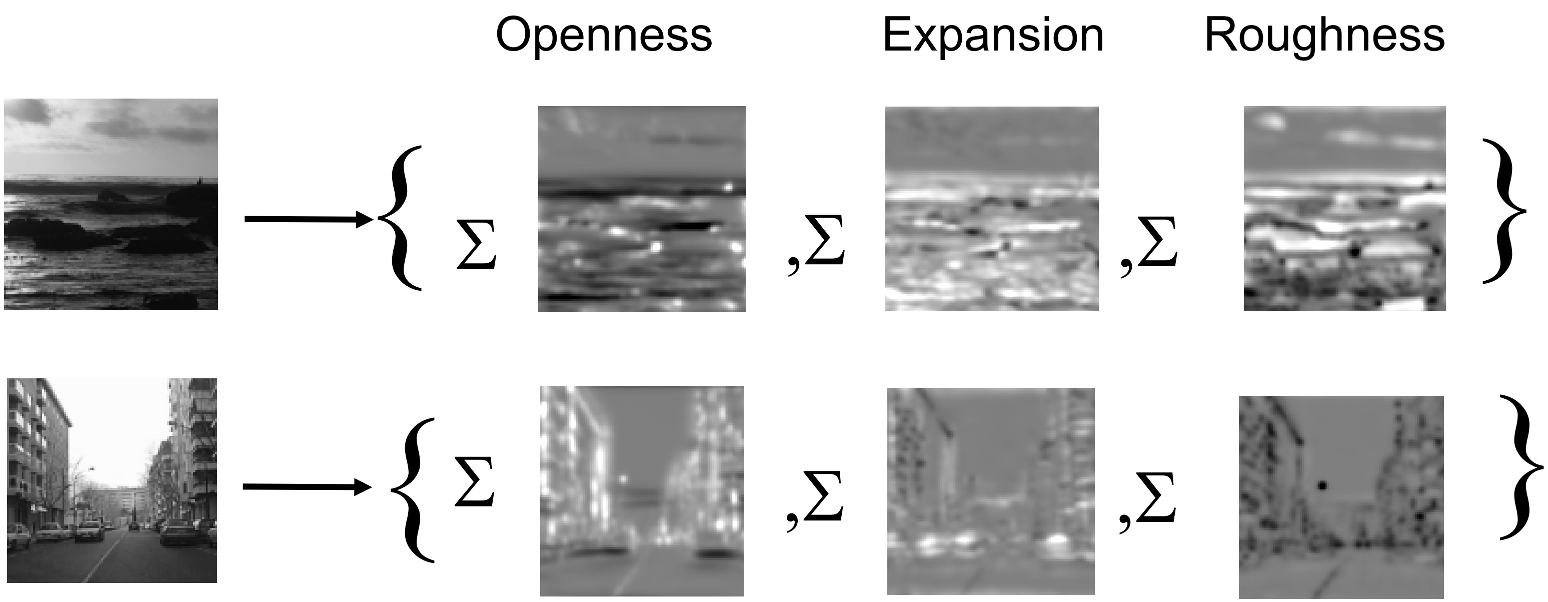


Open scene

Semi-open scene with texture

Spatial Envelope Representation

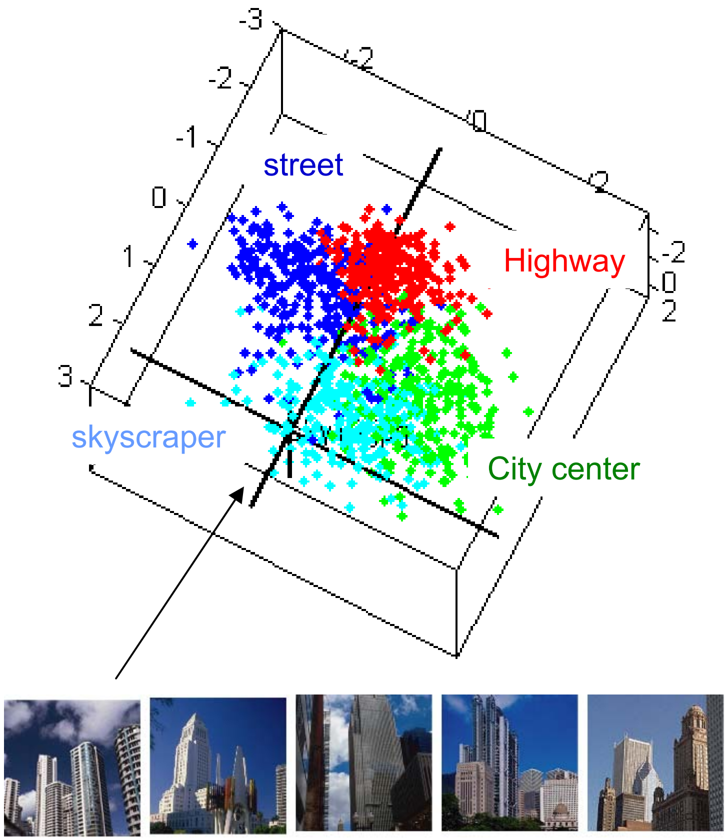
- A scene image is represented by a vector of values for each spatial envelope property.
- For instance:

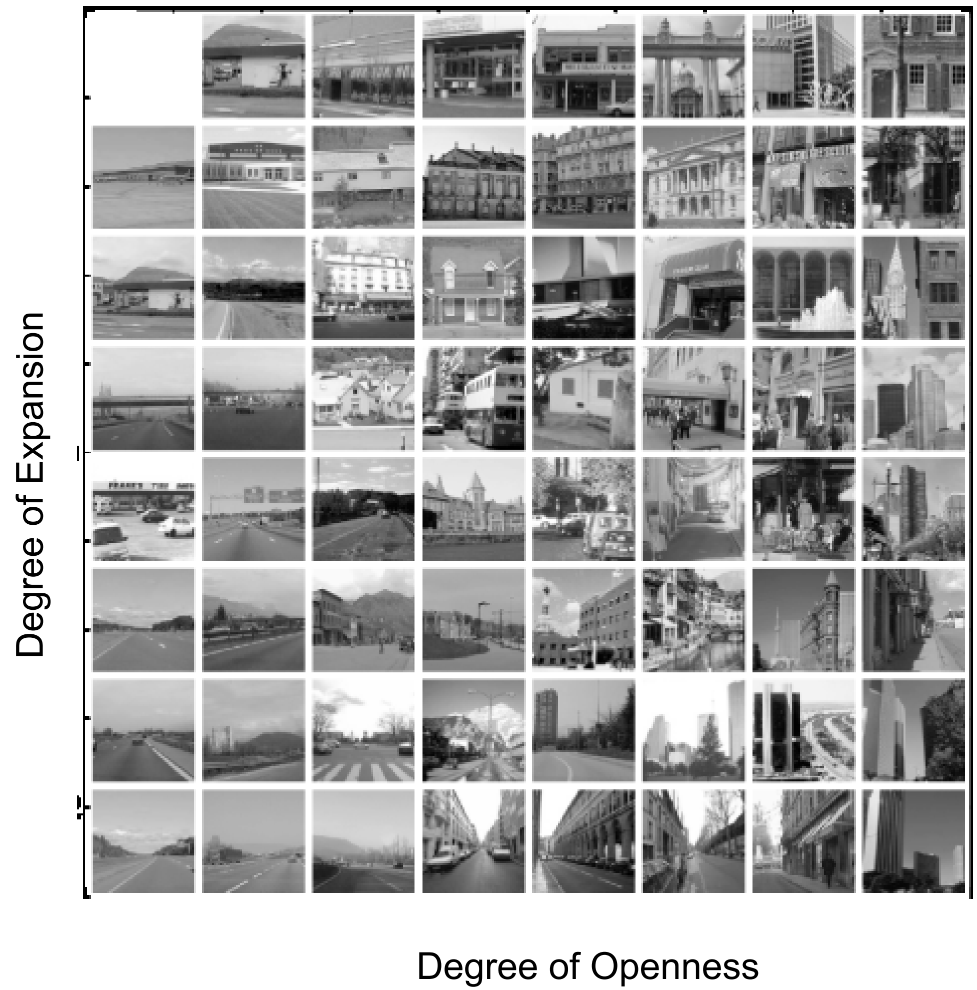


Oliva & Torralba (2001)

Modeling Scene Representation

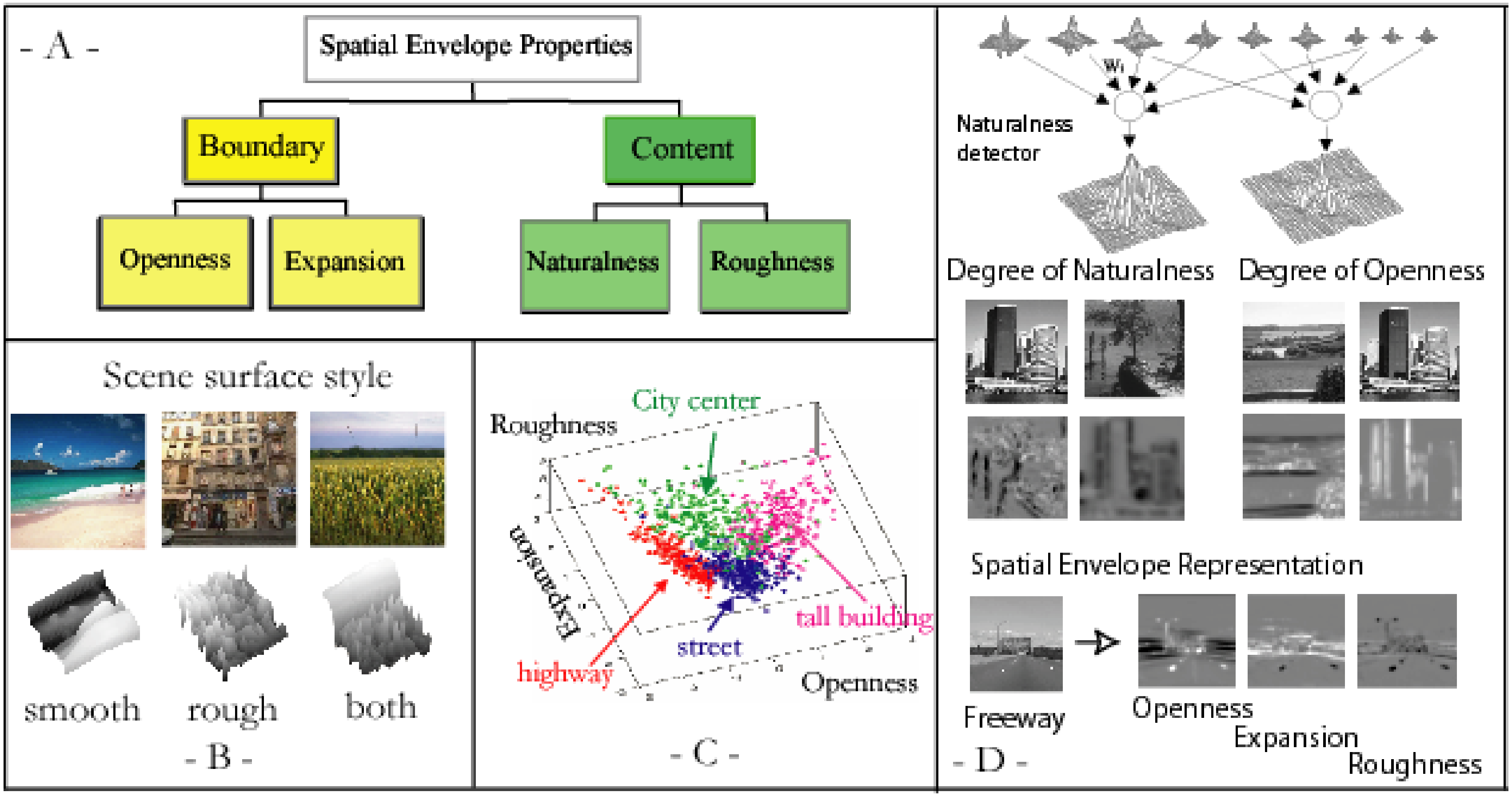
Scenes from the same category share similar global properties





Oliva & Torralba (2001). The spatial envelope model

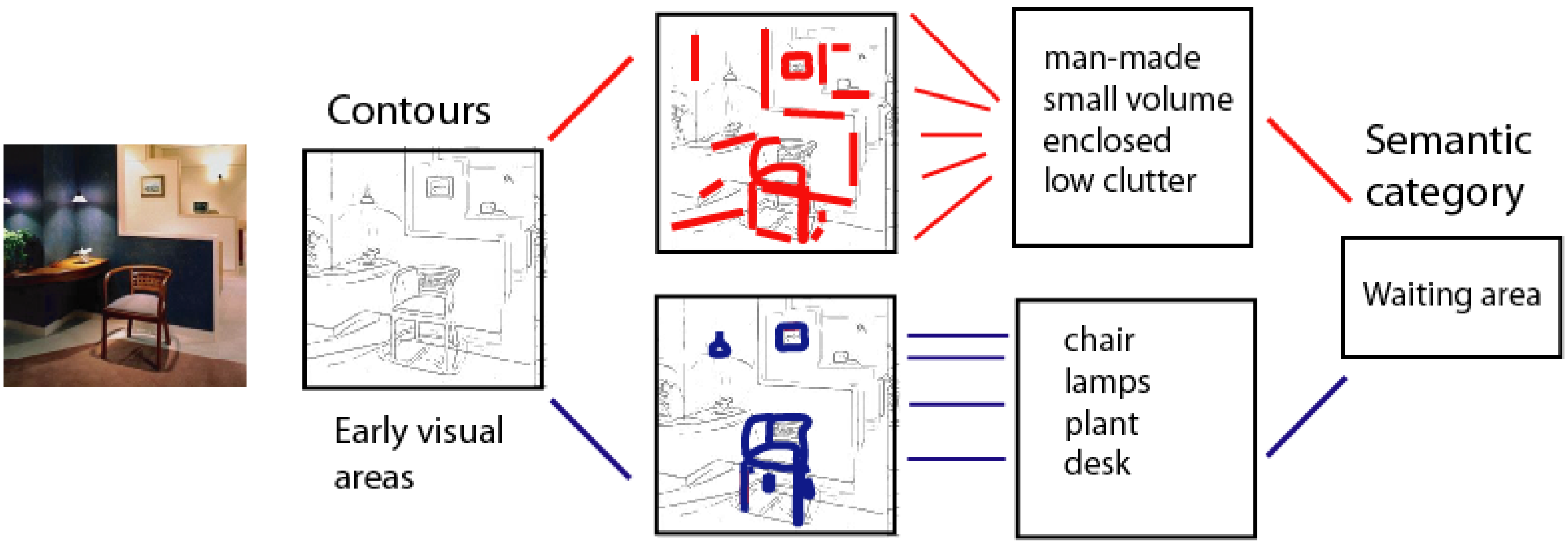
Spatial Envelope Theory of Scene Recognition



Oliva & Torralba (2001). International Journal of Computer Vision.

What about human mechanism of scene recognition ?

Scene-centered representation



Object-centered representation

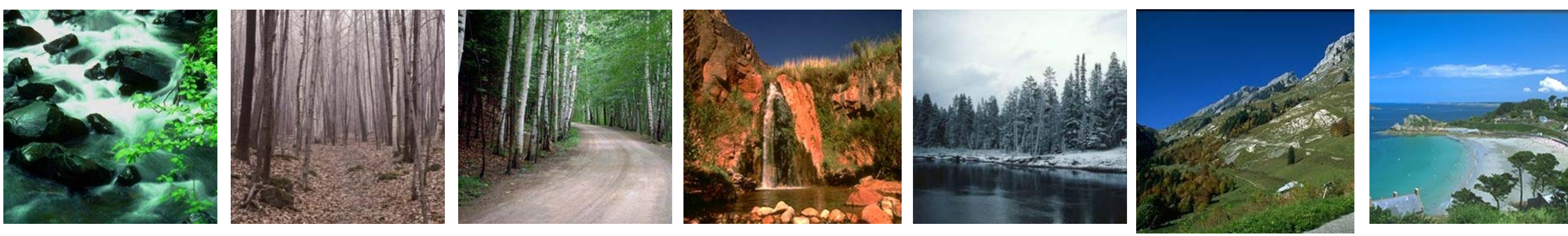
Scene centered representation

0.83 Camouflage
0.39 Movement
0.72 Navigability
0.55 Temperature
0.25 Openness
0.38 Expansion
0.27 Mean depth

Potential for Navigation

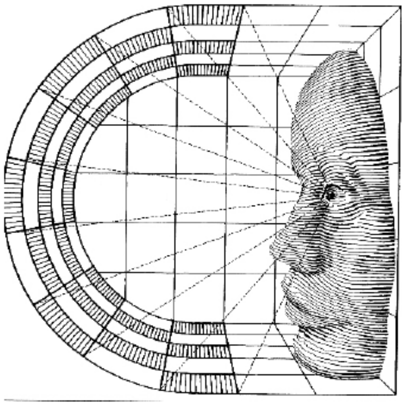
Difficult to walk through ⟷ Easy to walk

Mean depth



Small volume ⟷ large volume

Scene-Centered Representation



- <u>Boundary</u> *Mean depth Openness Expansion*
- <u>Content</u> *Naturalness Roughness Clutter*
- <u>Constancy</u> *Temperature Transience*
- <u>Affordance</u> *Navigability Concealment*

Greene & Oliva (2008). Recognition of Natural Scenes from Global Properties: Seeing the Forest Without Representing the Trees. Cognitive Psychology

Database

Desert

Field

Forest

Lake

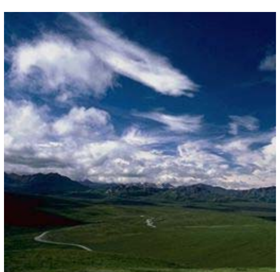
Mountain

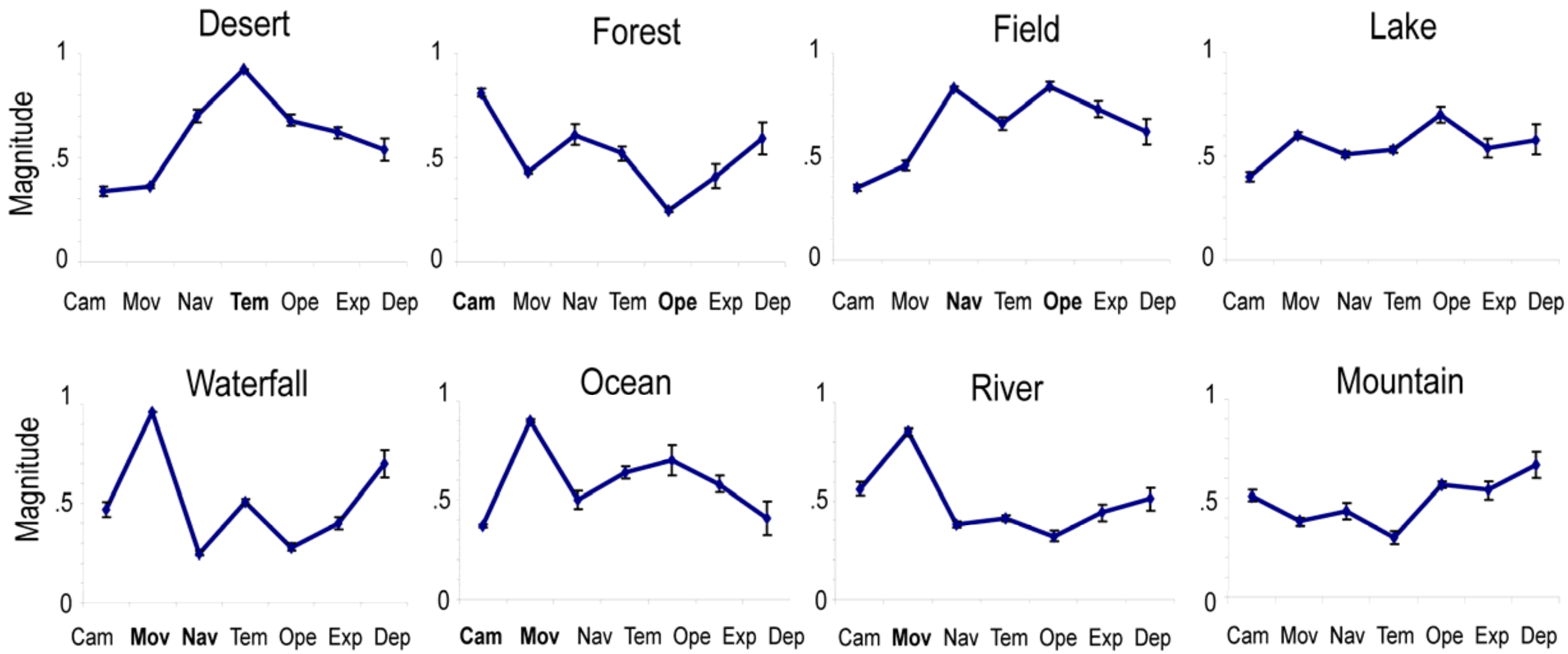
Ocean

River

Waterfall

Global scene properties as similarity metric





Experimental Approach:
Errors Prediction

Two scenes with similar global representation
but different categorical memberships should be confused
with each other (more false alarm)

Closed space
Low navigability

Open space
High navigability

Coast

Forest

Field

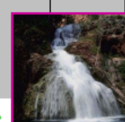
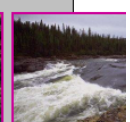
Scene-centered representation predicts human categorical false alarms rate

Scene-Centered Representation

0.76

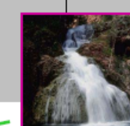
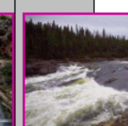
False alarms Scene categories


	Desert	Field	Forest	Lake	Mount	Ocean	River
Desert							
Field	0.7						
Forest	0.1	0.1					
Lake	0.5	0.6	0.2				
Mount	0.3	0.3	0.3	0.6			
Ocean	0.4	0.4	0.1	0.7			
River	0.0	0.0	0.4	0.4	0.4	0.5	
Waterfall	0.0	0.0	0.3	0.3	0.3	0.4	0.7

Matrix of similarity between Scene Categories


	Desert	Field	Forest	Lake	Mount	Ocean	River
Desert							
Field	0.29						
Forest	0.11	0.16					
Lake	0.07	0.15	0.09				
Mount	0.16	0.16	0.10	0.21			
Ocean	0.11	0.16	0.07	0.25			
River	0.09	0.13	0.10	0.19	0.14	0.21	
Waterf	0.06	0.06	0.11	0.09	0.14	0.13	0.29

Matrix of false alarms between Scene Categories

Image analysis (distance of each distractor to the target category) shows the same high correlation.

How *sufficient* is a scene-centered representation?

Method: Compare a naïve Bayes classifier to human performance.

Given a novel image	Scene-centered Signature	Probable Semantic Class
	0.36 Camouflage 0.38 Movement 0.94 Navigability 0.99 Temperature 0.89 Openness 0.68 Expansion 0.83 Mean depth	→ "desert"

A scene-centered classifier predicts correct performances

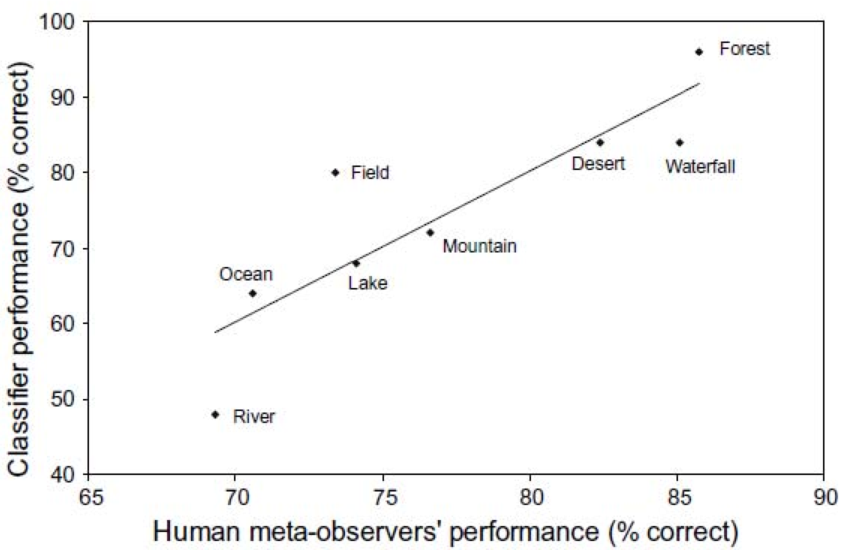


Image		
	H	Mountain, Lake, Ocean
	C	Mountain, Lake, Ocean
	H	Forest, River
	C	Forest, River
	H	Desert, Mountain, Lake
	C	Desert, Lake, Mountain
	H	Mountain, River, Lake,Forest
	C	Mountain, Lake, River, Forest

The classifier selects the same category than human in 62 % of cases for ambiguous, *non-prototypical* images

A scene-centered classifier predicts well the type of human false alarms

Given a misclassification of the classifier, at least one human observer made the same false alarm in 87% of the images (and 66% when considering 5 / 8 observers)

Ocean (error)

river

field (error)

desert

Scene Classification from “Texture”

Oliva & Torralba (2001,2006)

Scene Recognition via *texture*