Perception: Psychophysics and Modeling

11 | Scene Perception

Felix Wichmann

Neural Information Processing Group
Eberhard Karls Universität Tübingen

Literature

Thorpe, S. J., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381(6582):520–522.

Torralba, A. and Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems*, 14(3):391–412.

Wichmann, F. A., Drewes, J., Rosas, P., and Gegenfurtner, K. R. (2010). Animal detection in natural scenes: Critical features revisited. *Journal of Vision*, 10(4:6):1–27.

Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175.

Wolfe, J. M., Võ, M. L. H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Sciences*, 15(2), 77–84.

Limits of object recognition

The human visual system is very good at object recognition.

However, we cannot recognise multiple objects simultaneously.

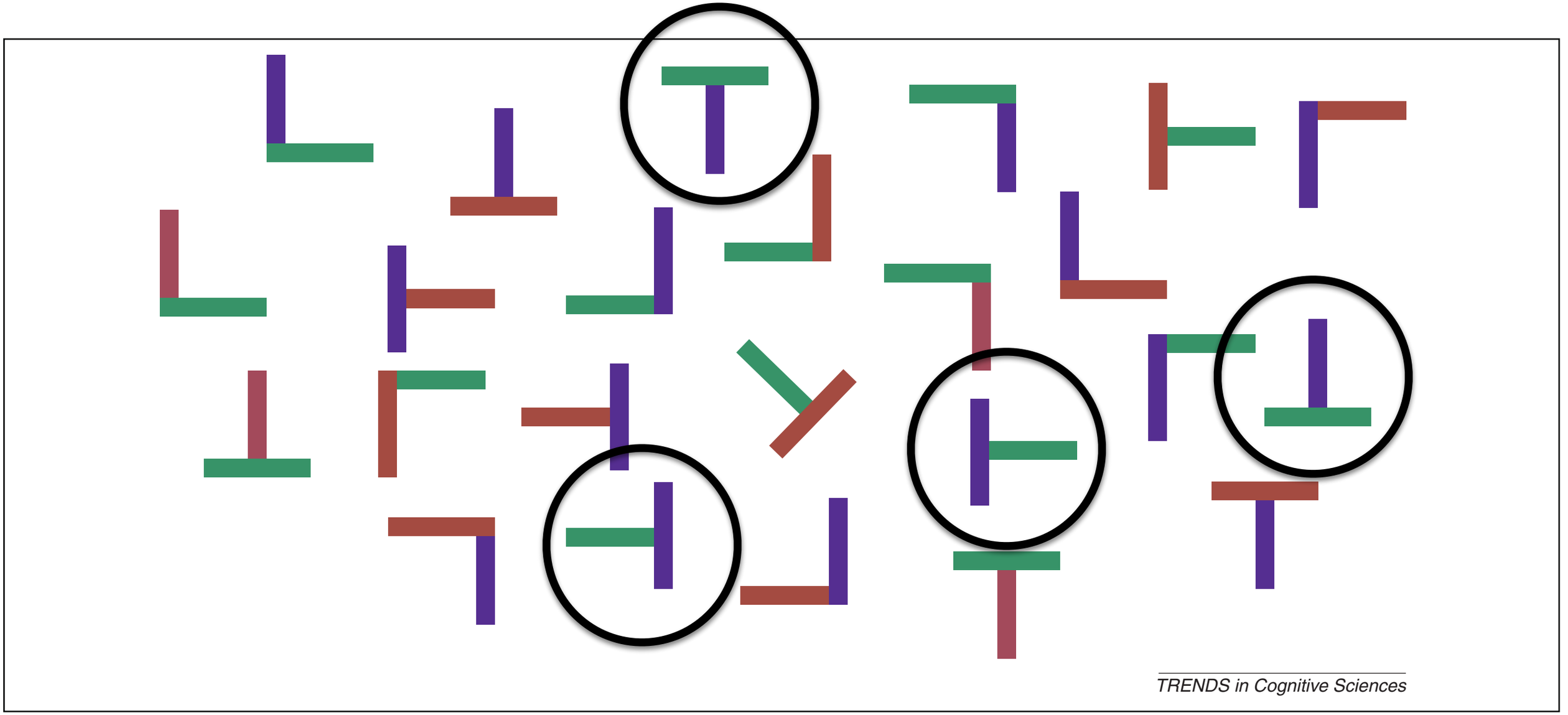


Figure 1. Find the four purple-and-green Ts. Even though it is easy to identify such targets, this task requires search.

From Wolfe, Vo, Evans and Green (2011)

Limits of object recognition

The human visual system is very good at object recognition.

However, we cannot recognise multiple objects simultaneously.

If we suppose that object recognition is a primary goal of the visual system, then additional mechanisms are required for it to be generally useful in complex scenes:

Scene perception (gist)

Saliency (guiding eyes in a scene)

Attention (selection mechanisms more generally)

Perceiving and Understanding Scenes

The gist of a scene can often be apprehended very very quickly, before details are perceived, or objects are recognised (‘‘beach scene’’, ‘‘street scene’’, ‘‘living room’’ , ‘‘open space’’...).

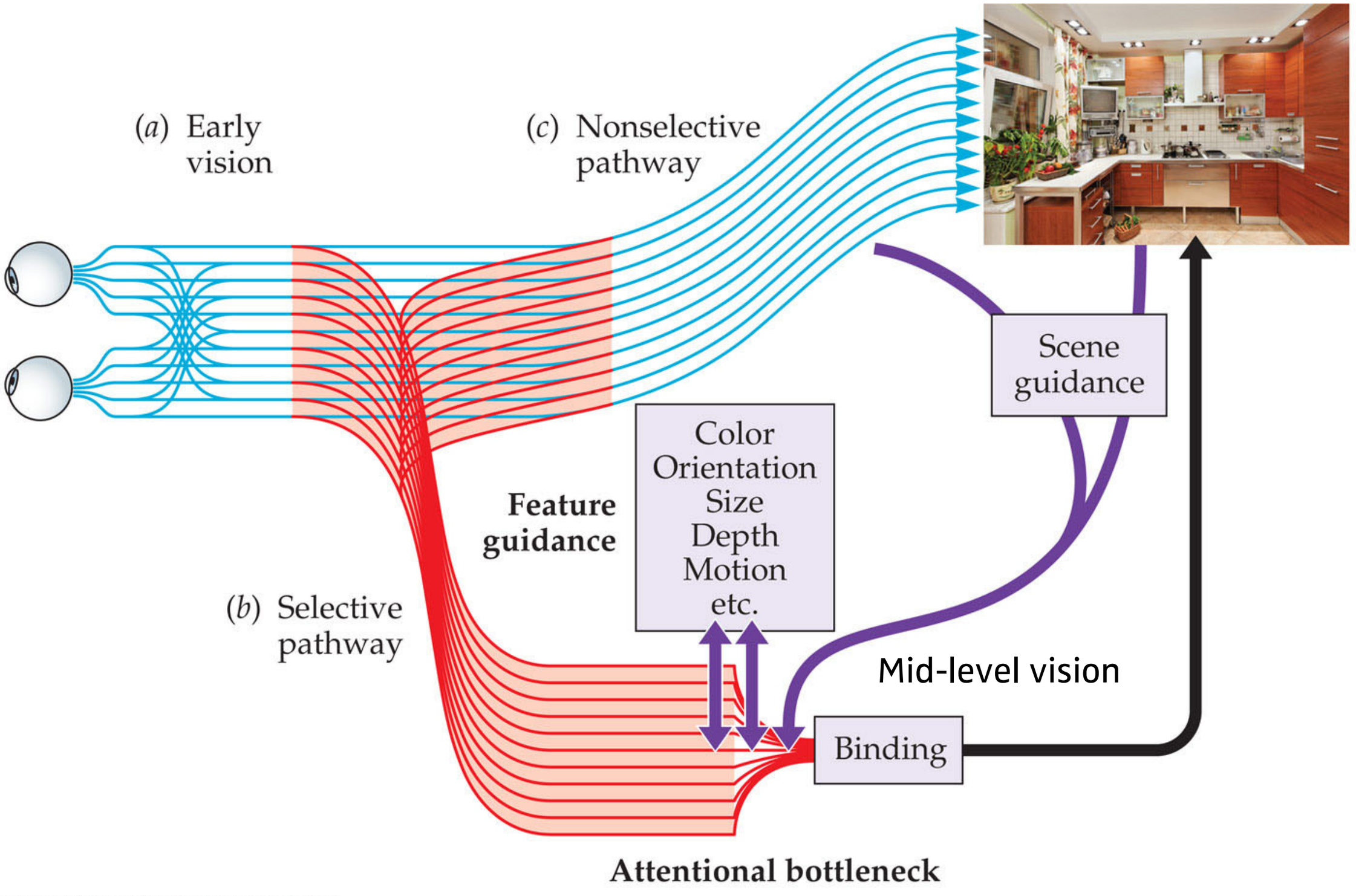
How is this possible?

One suggestion is, that there are two pathways to scene perception:

Selective pathway: Permits the recognition of one or a very few objects at a time; the objects are processed and identified in detail. This pathway passes through the bottleneck of selective attention.

Nonselective pathway: Contributes information about the distribution of features across a scene as well as information about the ‘‘gist’’ of the scene. This pathway does not pass through the bottleneck of attention.

Two pathways from the world to our perception of it



***SENSATION & PERCEPTION 4e*, Figure 7.27**

Two pathways from the world to our perception of it

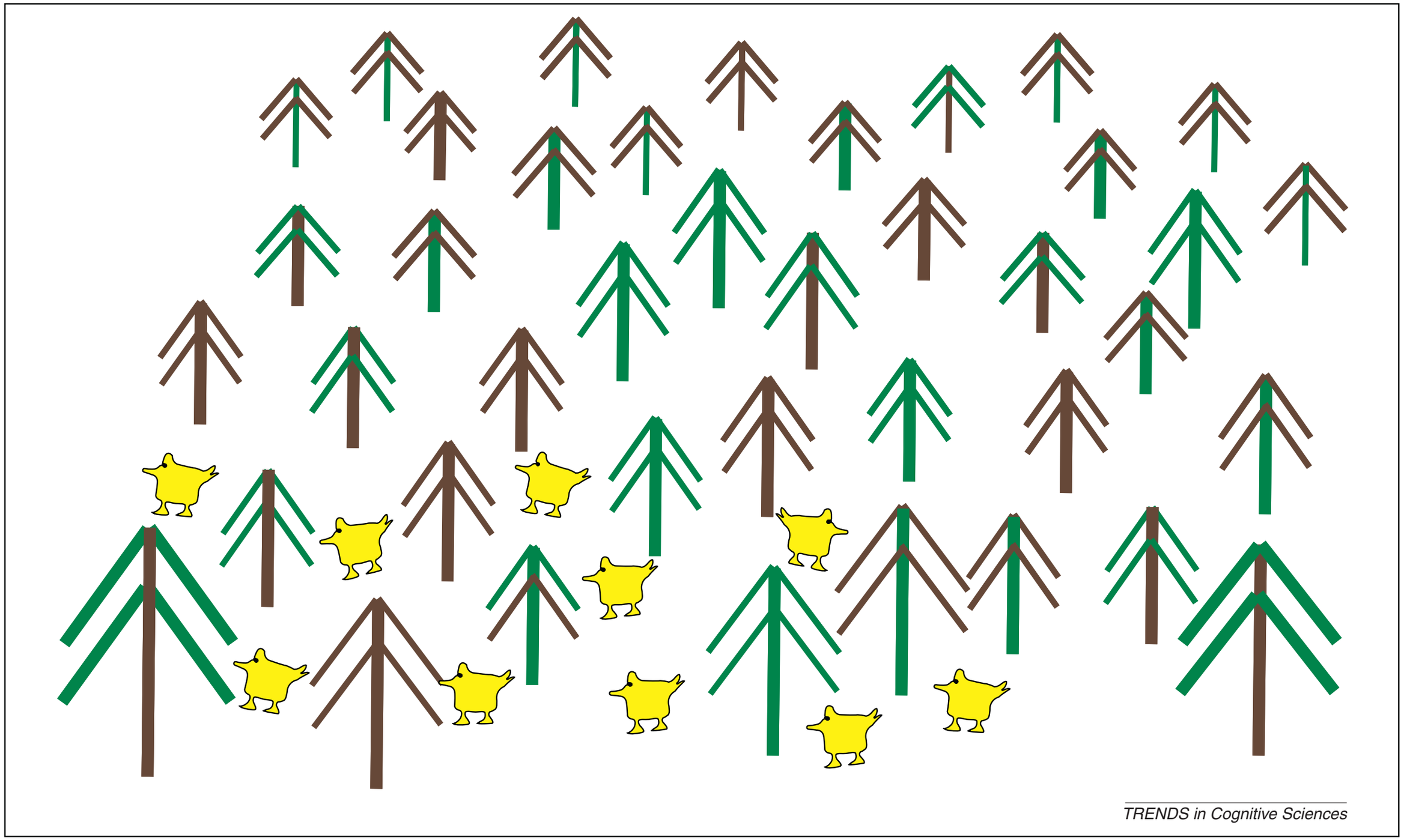


Figure 5. What do you see? How does that change when you are asked to look for an untilted bird or trees with brown trunks and green boughs? It is proposed that a nonselective pathway would 'see' image statistics, such as average color or orientation, in a region. It could get the 'gist' of forest and, perhaps, the presence of animals. However, it would not know which trees had brown trunks or which birds were tilted.

From Wolfe, Vo, Evans and Green (2011)

Perceiving and Understanding Scenes

The nonselective pathway computes scene gist and layout very quickly.

Spatial layout: The description of the structure of a scene (e.g., enclosed, open, rough, smooth) without reference to the identity of specific objects in the scene.

One popular idea of how this might be possible: The nonselective pathway computes ensemble statistics.

Ensemble statistics: The average and distribution of properties, such as orientation or color, over a set of objects or a region in a scene.

Popular idea: Perhaps the ensemble statistics—or *summary statistics*—could help with rapid object recognition, too. Because there are puzzling findings of really very rapid object recognition!

Speed of processing in the human visual system

Simon Thorpe, Denis Fize & Catherine Marlot

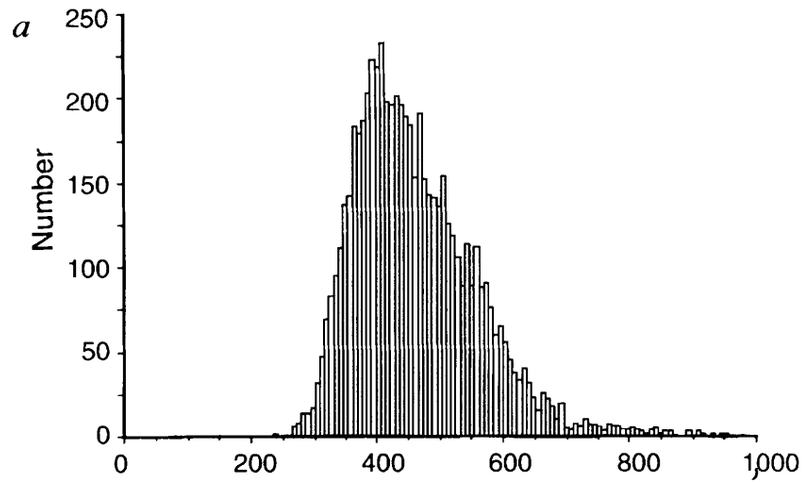
Centre de Recherche Cerveau & Cognition, UMR 5549, 31062 Toulouse, France

How long does it take for the human visual system to process a complex natural image? Subjectively, recognition of familiar objects and scenes appears to be virtually instantaneous, but measuring this processing time experimentally has proved difficult. Behavioural measures such as reaction times can be used[1], but these include not only visual processing but also the time required for response execution. However, event-related potentials (ERPs) can sometimes reveal signs of neural processing well before the motor output[2]. Here we use a go/no-go categorization task in which subjects have to decide whether a previously unseen photograph, flashed on for just 20 ms, contains an animal. ERP analysis revealed a frontal negativity specific to no-go trials that develops roughly 150 ms after stimulus onset. We conclude that the visual processing needed to perform this highly demanding task can be achieved in under 150 ms.

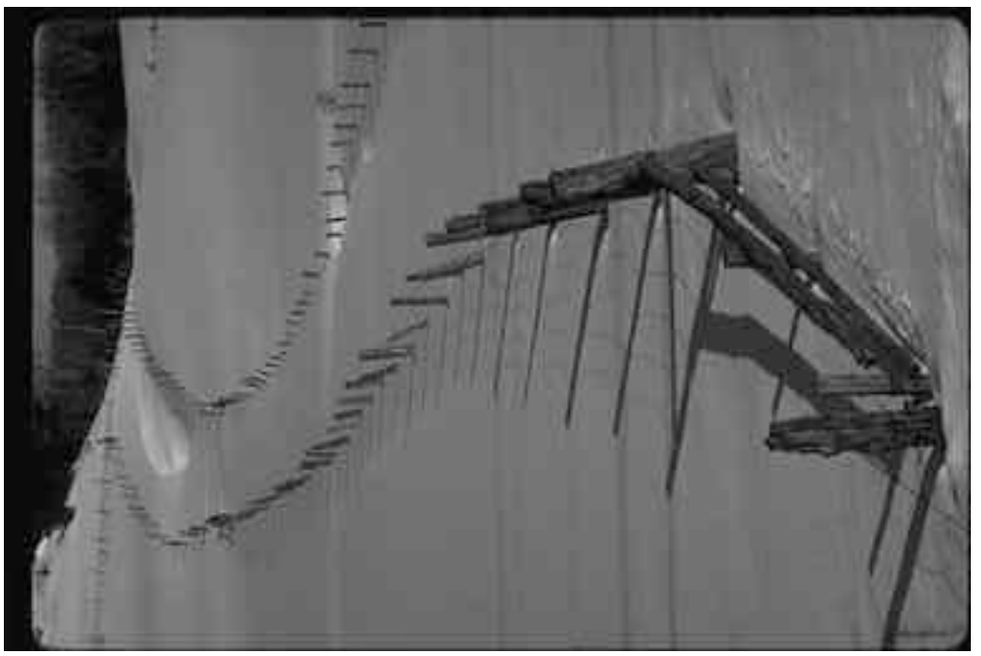
Neurophysiological measurements of the latencies of selective visual responses can be used to provide estimates of visual processing time[3]. For example, it is known that higher-order visual areas such as the primate superior temporal sulcus contain neurons that can respond selectively to faces with latencies of ~ 100 ms[4–6]. In humans, face-selective evoked potentials have been demonstrated using both surface ERP recordings[7,8] and implanted intracerebral electrodes[9–11]. Such potentials typically peak at ~ 200 ms after stimulus onset, but may start as early as 140 ms. It is unclear, however, whether such latencies are typical of visual processing in general. One problem is that face processing may involve highly specialized and optimized neural pathways, and although there have been a few reports of early differential responses to other stimuli, including words[10,12,13] and line drawings[13], no previous ERP studies have attempted to measure processing times for more natural scenes. A second

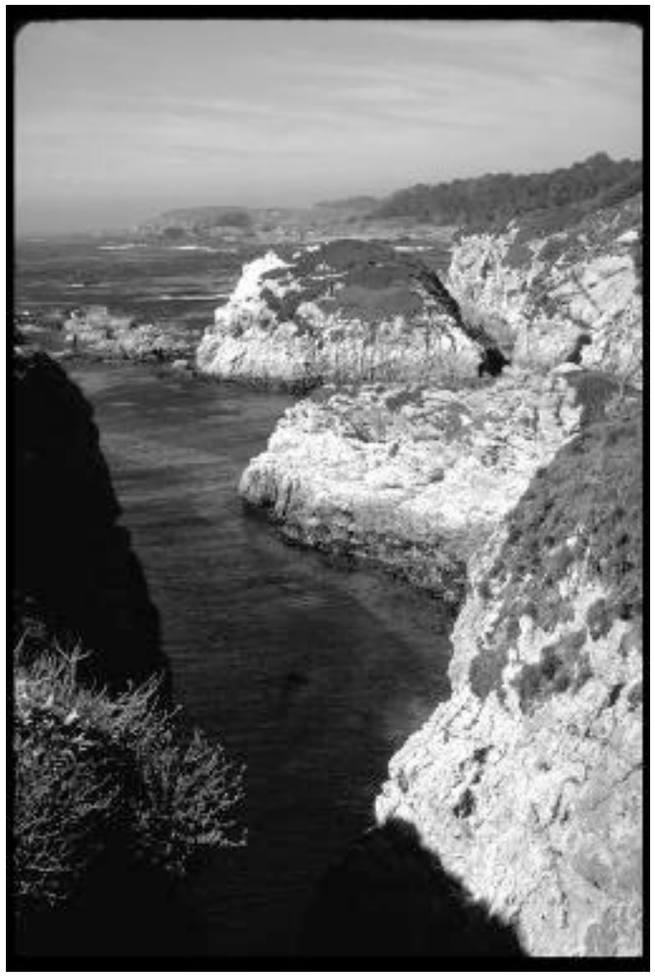
such a task (the subjects had no *a priori* information about the type of animal to look for, its position or size, or even the number of animals present), performance was remarkably good. The average proportion of correct responses was 94%, with one of the fifteen subjects achieving 98% correct responses. The median reaction times on 'go' trials was 445 ms, although this value varied considerably between subjects, from a minimum of 382 ms to as much as 567 ms (Fig. 1). This remarkable level of performance was possible despite the very brief presentations, which effectively rule out the use of eye movements during image processing.

Whereas the behavioural reaction times put an upper limit on the time required for visual processing, the analysis of event-related potentials provided a much stronger constraint. By comparing average brain potentials generated on correct 'go' trials with those generated on correct 'no-go' trials, we were able to demonstrate that the two potentials diverge very sharply at ~ 150 ms after stimulus onset. The effect was particularly clear at frontal recording sites, and was characterized by a nearly linear increase in the voltage difference over the following 50 ms or so, the potential being more negative on no-go trials (Fig. 2). All 15 subjects showed the effect (Fig. 3), and although the onset latency varied somewhat between subjects, the differences were very minor compared with the very large differences in behavioural reaction times. Furthermore, there was no correlation whatsoever between behavioural reaction time and the onset latency for the differential response. This makes it unlikely that the differential




RAIL
ROAD
CROSSING






The Steamship Brand
Molasses Candy

How many animals?

‘‘Problem’’

1. Rapid animal (object) detection is very difficult—‘‘Quite how the human visual system achieves such a phenomenal amount of computation in such a short time is clearly a challenge for current theories of object vision [...].’’ (Thorpe et al., p. 522)

2. ‘‘[...] but given the large number of processing stages involved in the primate visual system, it seems likely that much of this processing must be based on essentially feed-forward mechanisms.’’ (Thorpe et al., p. 522)

Feedforward only? Focal attention required? How is this achieved without segmentation?

Background & Motivation

Results from ultra-rapid-animal detection experiments pose serious temporal constraints on the class of algorithms that may underly this capability (Thorpe et al., 1996 but see, e.g. Gerstner, 2006).

Many models of contour extraction and object segmentation, thought to precede object recognition, rely on feedback-connections, and/or lateral competitive or co-operative interactions, however.

Furthermore, there is evidence that ultra-rapid animal detection is accomplished in the absence of attention (Li, VanRullen, Koch & Perona, 2002).

Ultra-rapid-animal detection may be based on global, simple image statistics which can be calculated prior to segmentation (same as ensemble statistics):

Candidate for a Probabilistic Invariance: High spatial frequencies in the vertical and horizontal orientations allow animal from non-animal images to be discriminated at around 80% correct (Oliva & Torralba, 2001; Torralba & Oliva, 2003).

INSTITUTE OF PHYSICS PUBLISHING NETWORK: COMPUTATION IN NEURAL SYSTEMS

Network: Comput. Neural Syst. **14** (2003) 391–412 PII: S0954-898X(03)53778-2

Statistics of natural image categories

Antonio Torralba[1] **and Aude Oliva**[2]

[1] Artificial Intelligence Laboratory, MIT, Cambridge, MA 02139, USA
[2] Department of Psychology and Cognitive Science Program, Michigan State University, East Lansing, MI 48824, USA

E-mail: torralba@ai.mit.edu and aoliva@msu.edu

Received 16 September 2002, in final form 30 January 2003
Published 12 May 2003
Online at stacks.iop.org/Network/14/391

Abstract

In this paper we study the statistical properties of natural images belonging to different categories and their relevance for scene and object categorization tasks. We discuss how second-order statistics are correlated with image categories, scene scale and objects. We propose how scene categorization could be computed in a feedforward manner in order to provide top-down and contextual information very early in the visual processing chain. Results show how visual categorization based directly on low-level features, without grouping or segmentation stages, can benefit object localization and identification. We show how simple image statistics can be used to predict the presence and absence of objects in the scene before exploring the image.

(Some figures in this article are in colour only in the electronic version)

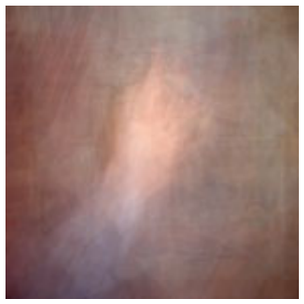
Objects

Face

Pedestrian

Car

Cow



Hand

Chair

Scenes

Mountain

Beach

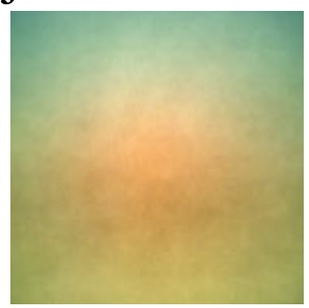
Forest

Highway

Street

Indoor

Objects in scenes



Animal
in natural scene

Tree
in urban scene

Close-up person
in urban scene

Far pedestrian
in urban scene

Car in
urban scene

Lamp in
indoor scene

Figure 1. Averaged pictures of categories of objects, scenes and objects in scenes, computed with 100 exemplars or more per category. Exemplars were chosen to have the same basic level and viewpoint in regard to an observer. The group objects in scenes (third row) represent examples of the averaged peripheral information around an object centred in the image.

from Torralba & Oliva (2003)

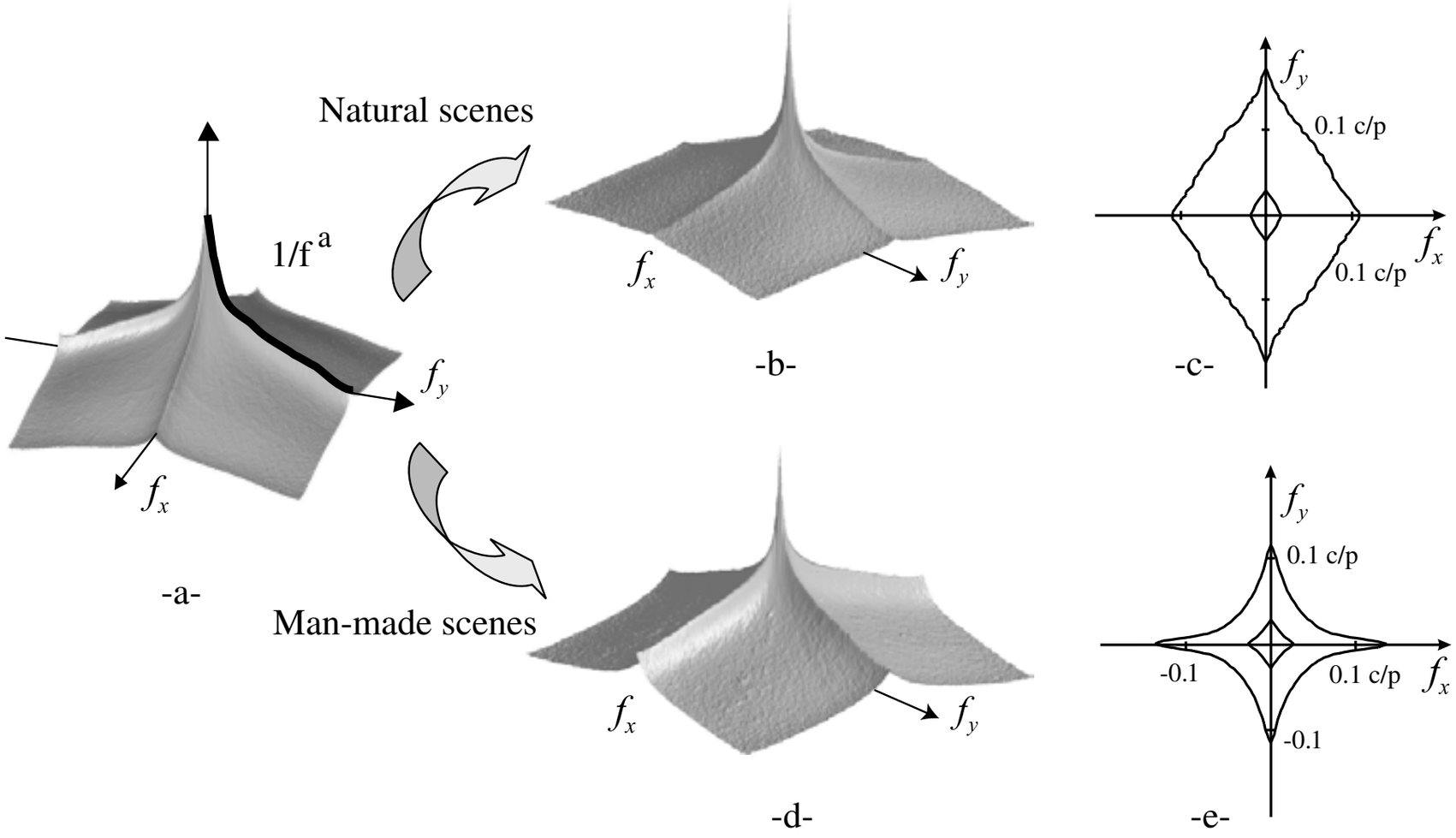


Figure 2. (a) Mean power spectrum averaged from 12 000 images (vertical axis is in logarithmic units). Mean power spectra computed with 6000 pictures of man-made scenes (b) and 6000 pictures of natural scenes (d); (c) and (e) are their respective spectral signatures. The contour plots represent 50 and 80% of the energy of the spectral signature. The contour is selected so that the sum of the components inside the section represents 50% (and 80%) of the total. Units are in cycles per pixel (cf also Baddeley 1996).

from Torralba & Oliva (2003)

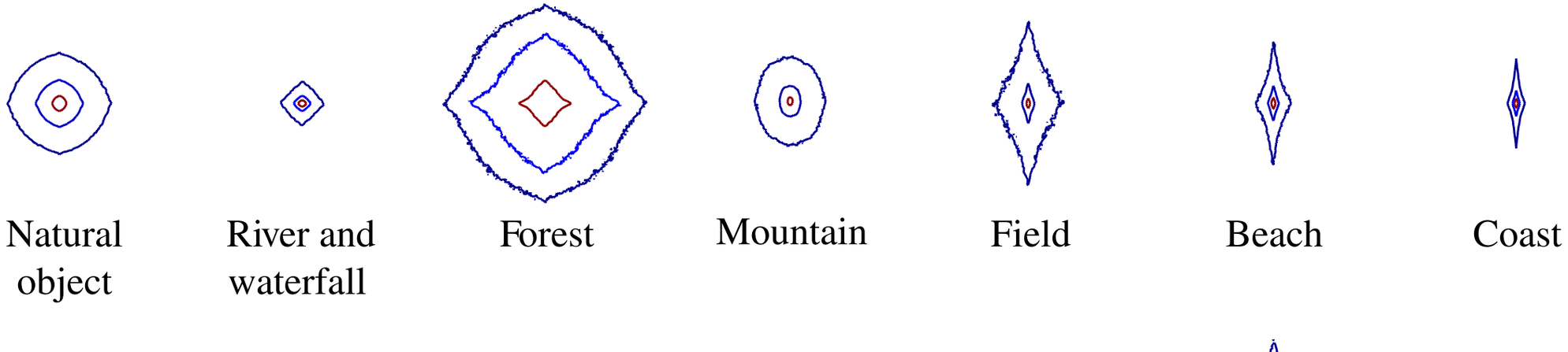


Figure 3. Spectral signatures of 14 different image categories. Each spectral signature is obtained by averaging the power spectra of a few hundred images per category. The contour plots represent 60, 80 and 90% of the energy of the spectral signatures (energy is obtained by adding the square of the Fourier components). The size of the spectral signature is correlated with the slope (α). A large value of α produces a fast decay of the energy at high spatial frequencies, which produces a smaller contour. The overall shape is a function of both $\alpha(\theta)$ and $A(\theta)$.

from Torralba & Oliva (2003)

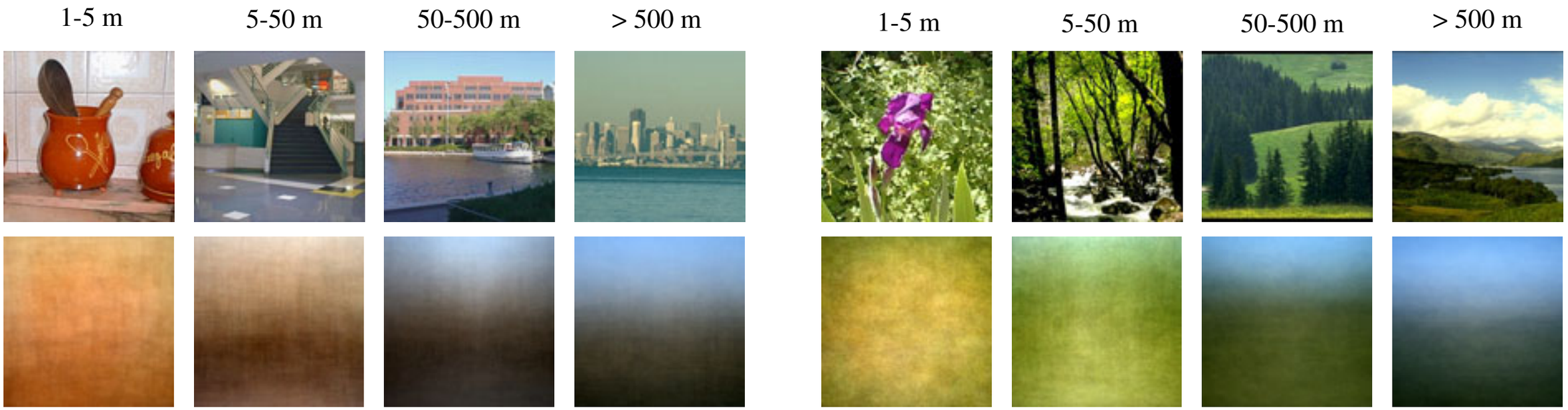


Figure 4. Averaged spatial images and spectral signatures as a function of scene scale. Scene scale refers to the mean distance between the observer and the principal elements that compose the scene. Each image average and spectral signature was calculated with 300–400 images.

from Torralba & Oliva (2003)

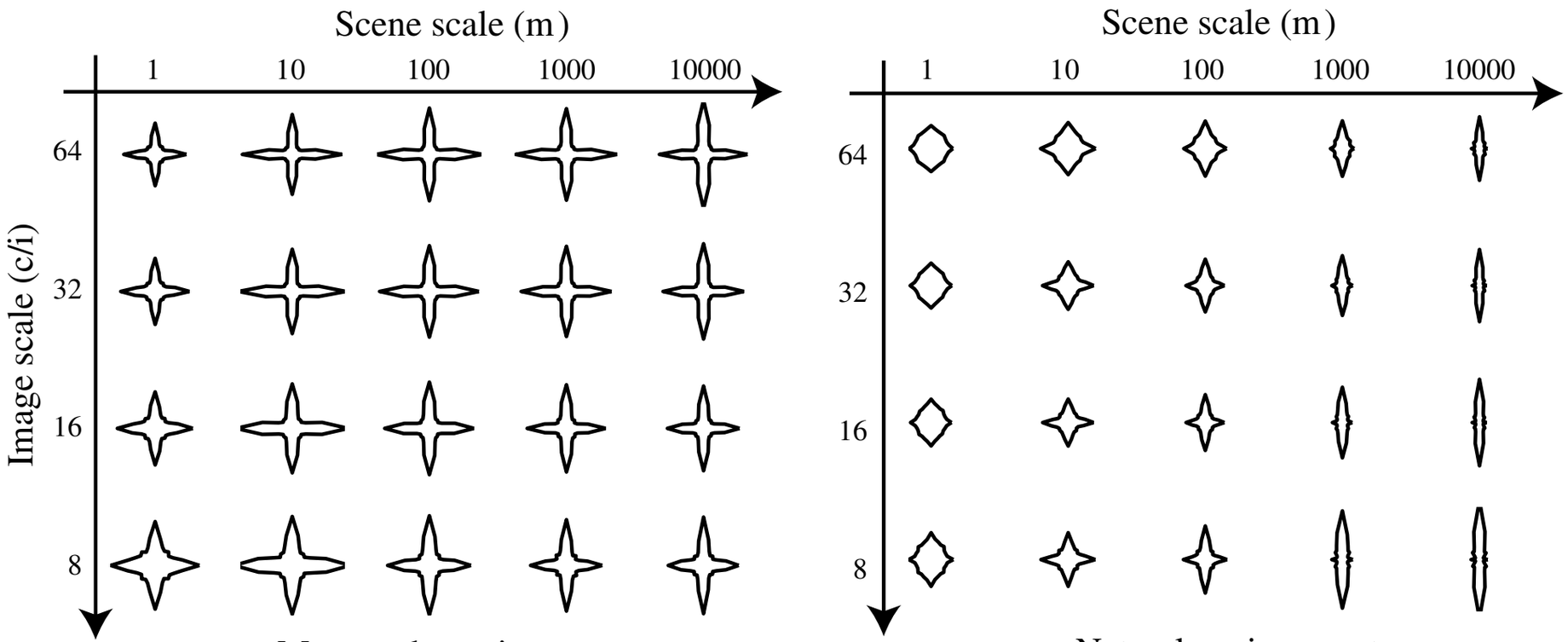


Figure 5. Polar plots of responses of multiscale oriented Gabor filters. The magnitude of each orientation corresponds to the total output energy averaged across the entire image. The energies are normalized across image scale by multiplying by a constant so that noise with $1/f$ amplitude spectrum has the same polar plots at all image scales.

from Torralba & Oliva (2003)

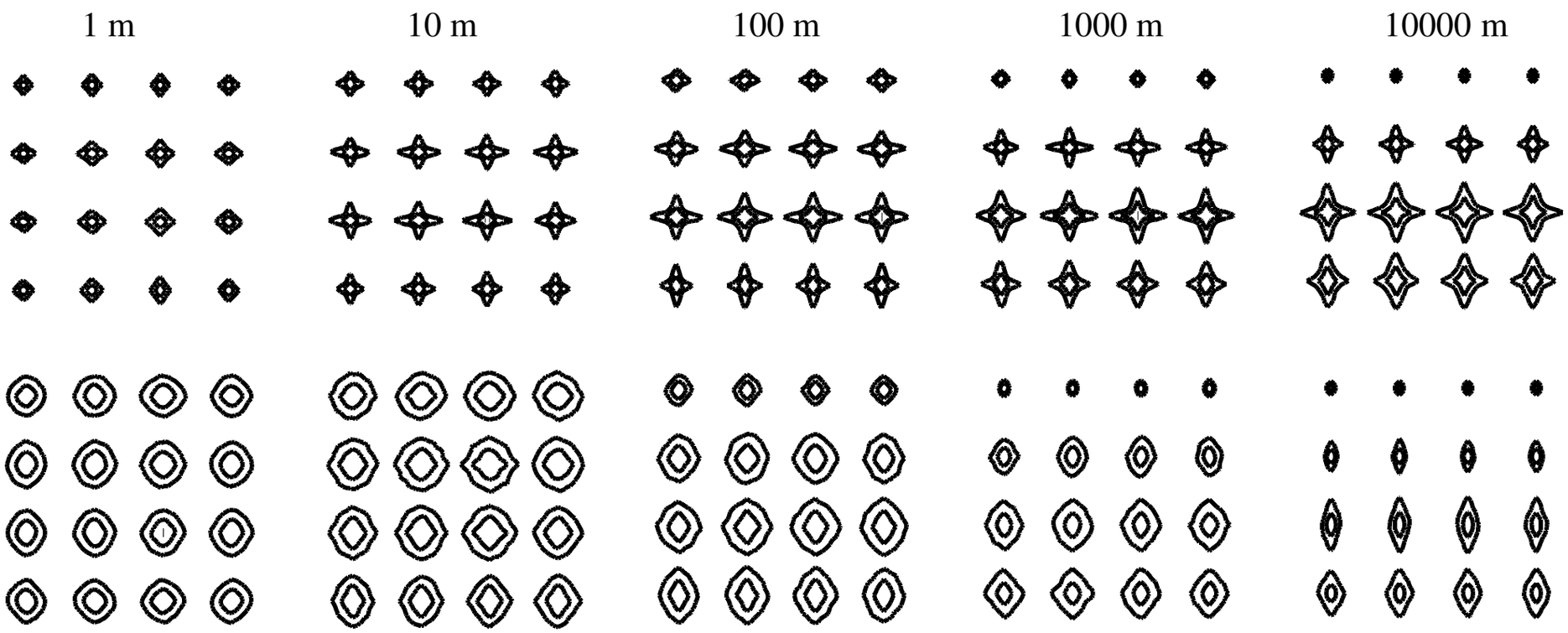


Figure 6. Illustration of the non-stationarity of image statistics in groups of man-made (top) and natural (bottom) environments at different depth scales (from left to right, close-up views to panoramic views). The spectral signatures were obtained by averaging the windowed power spectra at 4×4 locations in the images. As scene scale increases, the image statistics become non-stationary.

Idea: Ensemble or summary statistics not only useful for scene gist, but for object recognition, too!

from Torralba & Oliva (2003)

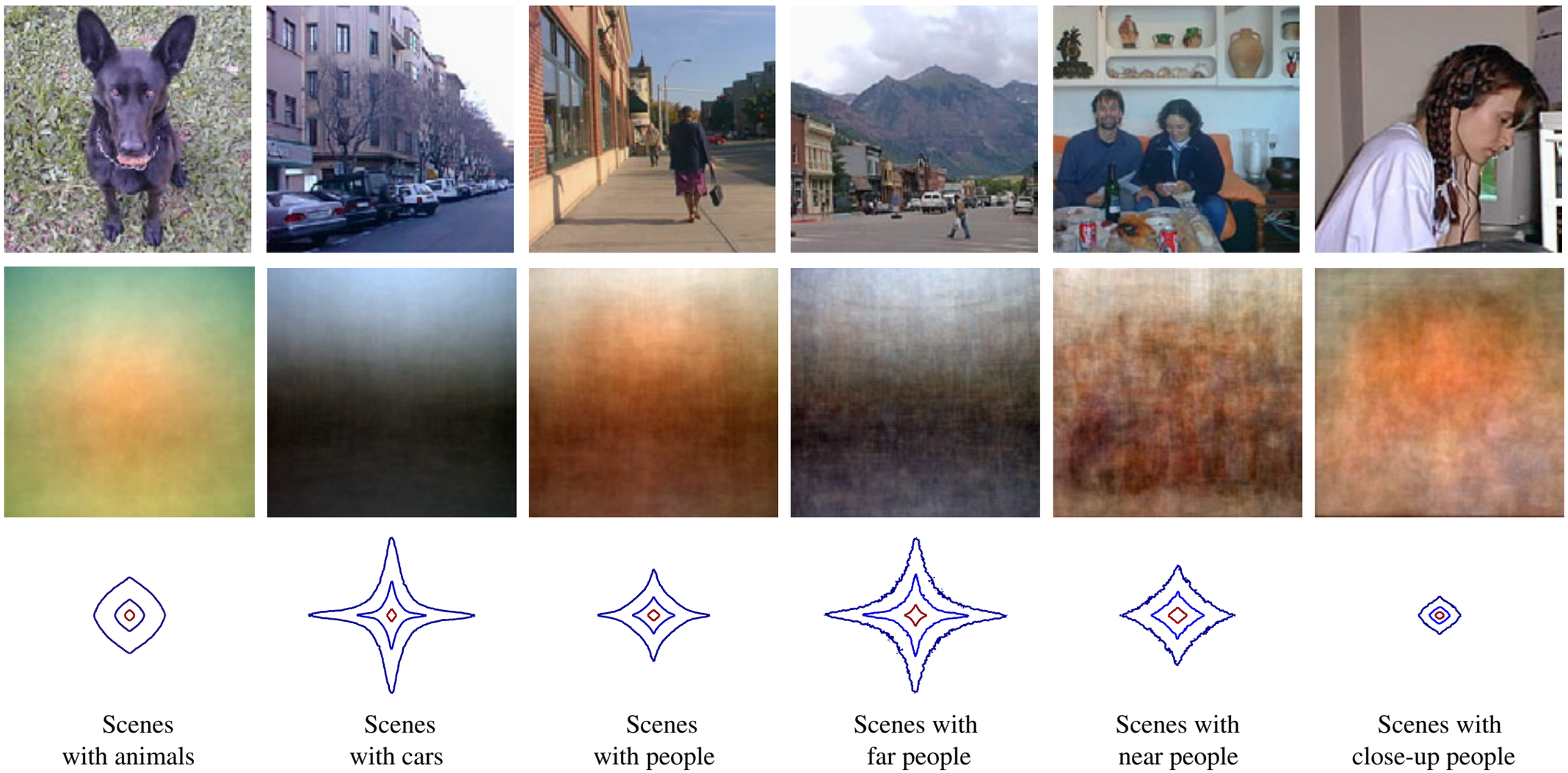


Figure 12. Average intensity and spectral signatures of sets of images constrained to contain specific objects. Image statistics can be predictors of the presence/absence of particular objects in the scene.

from Torralba & Oliva (2003)

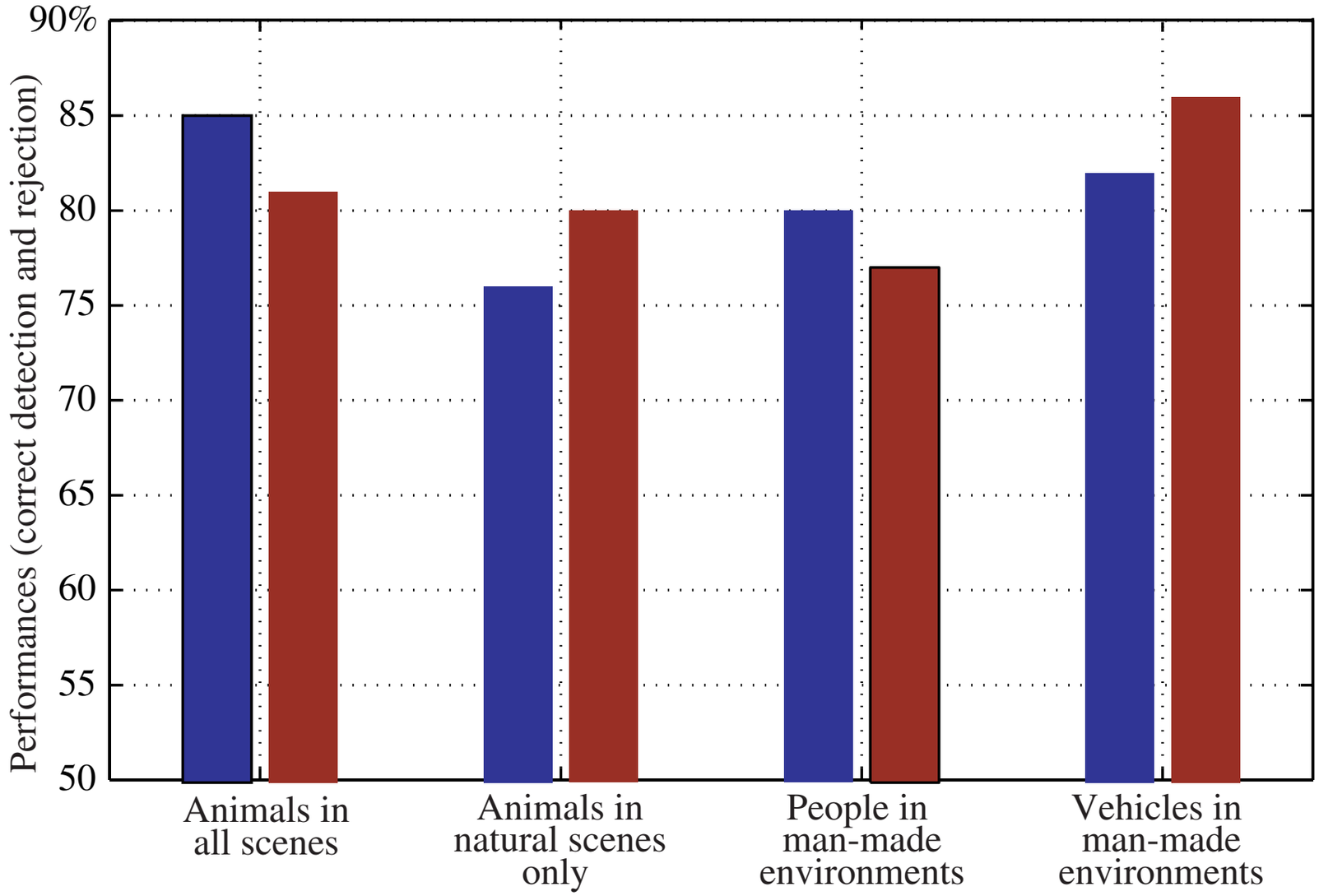
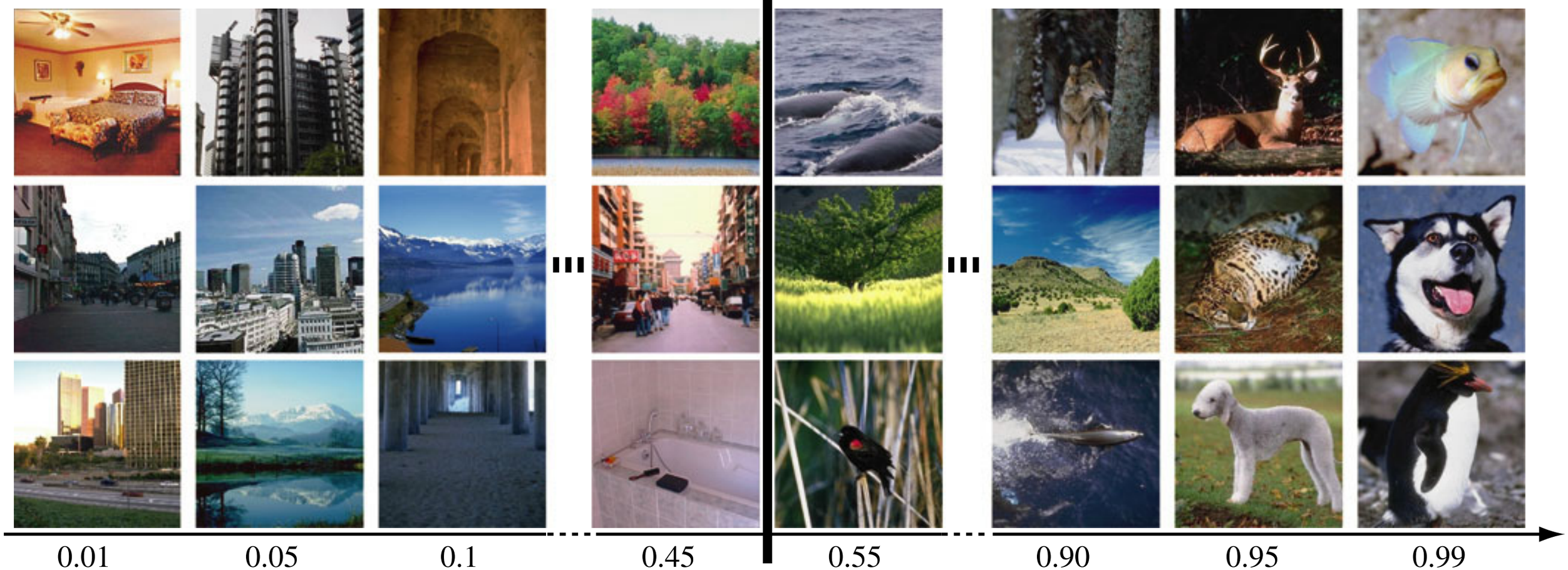


Figure 13. Performance in object prediction. For each object category we show performance for prediction of presence (left bar) of the objects and prediction of absence (right bar).

from Torralba & Oliva (2003)



from Torralba & Oliva (2003)

From Torralba & Oliva Statistics of Natural Image Categories in Network: Computation in Neural Systems, 2003:

... we study the statistical properties of natural images belonging to different categories and their relevance for scene and object categorization tasks. We discuss how second-order statistics are correlated with image categories, ... We show how simple image statistics can be used to predict the presence and absence of objects in the scene before exploring the image (p. 391).

These results corroborate studies by Thorpe ... showing that a cognitive task such as animal versus non-animal categorization could be performed in a feedforward way and without the need of sequential focus of attention or segmentation stages (p. 409).

From J.M. Henderson's Review Human Gaze Control during Real-World Scene Perception in TICS, 2003, p. 501:

Scenes can be identified and their gist apprehended very rapidly, well within the duration of a single fixation ... . This rapid apprehension ... can be based on global image statistics that are predictive of the scene's identity and semantic gist (Oliva & Torralba, 2001; Torralba & Oliva, 2003).

From Johnson & Olshausen *Timecourse of neural signatures of object recognition* in JoV, 2003, p. 509:

... it has been shown that natural images can be classified into animal and nonanimal categories at a success rate of 80% using nothing but measures of global image statistics such as the power spectrum (Torralba & Oliva, 2003), ... Thus, the visual system might be able to build a good template of the features associated with a category and use this template to make preemptive categorizations with reasonable accuracy.

… still …

Even if the spectral differences between the image categories were real, this does not necessarily imply that the human visual system makes use of them.

Animal detection in natural scenes: Critical features revisited

Felix A. Wichmann Modelling of Cognitive Processes, Berlin Institute of Technology & Bernstein Center for Computational Neuroscience Berlin, Berlin, Germany

Jan Drewes Abteilung Allgemeine Psychologie, Universität Giessen, Giessen, Germany

Pedro Rosas Centro de Neurociencias Integradas, Facultad de Medicina, Universidad de Chile, Santiago, Chile

Karl R. Gegenfurtner Abteilung Allgemeine Psychologie, Universität Giessen, Giessen, Germany

S. J. Thorpe, D. Fize, and C. Marlot (1996) showed how rapidly observers can detect animals in images of natural scenes, but it is still unclear which image features support this rapid detection. A. B. Torralba and A. Oliva (2003) suggested that a simple image statistic based on the power spectrum allows the absence or presence of objects in natural scenes to be predicted. We tested whether human observers make use of power spectral differences between image categories when detecting animals in natural scenes. In Experiments 1 and 2 we found performance to be essentially independent of the power spectrum. Computational analysis revealed that the ease of classification correlates with the proposed spectral cue without being caused by it. This result is consistent with the hypothesis that in commercial stock photo databases a majority of animal images are pre-segmented from the background by the photographers and this pre-segmentation causes the power spectral differences between image categories and may, furthermore, help rapid animal detection. Data from a third experiment are consistent with this hypothesis. Together, our results make it exceedingly unlikely that human observers make use of power spectral differences between animal- and no-animal images during rapid animal detection. In addition, our results point to potential confounds in the commercially available "natural image" databases whose statistics may be less natural than commonly presumed.

Keywords: rapid animal detection, natural scenes, power spectrum, amplitude spectrum, scene gist, local features, natural image statistics

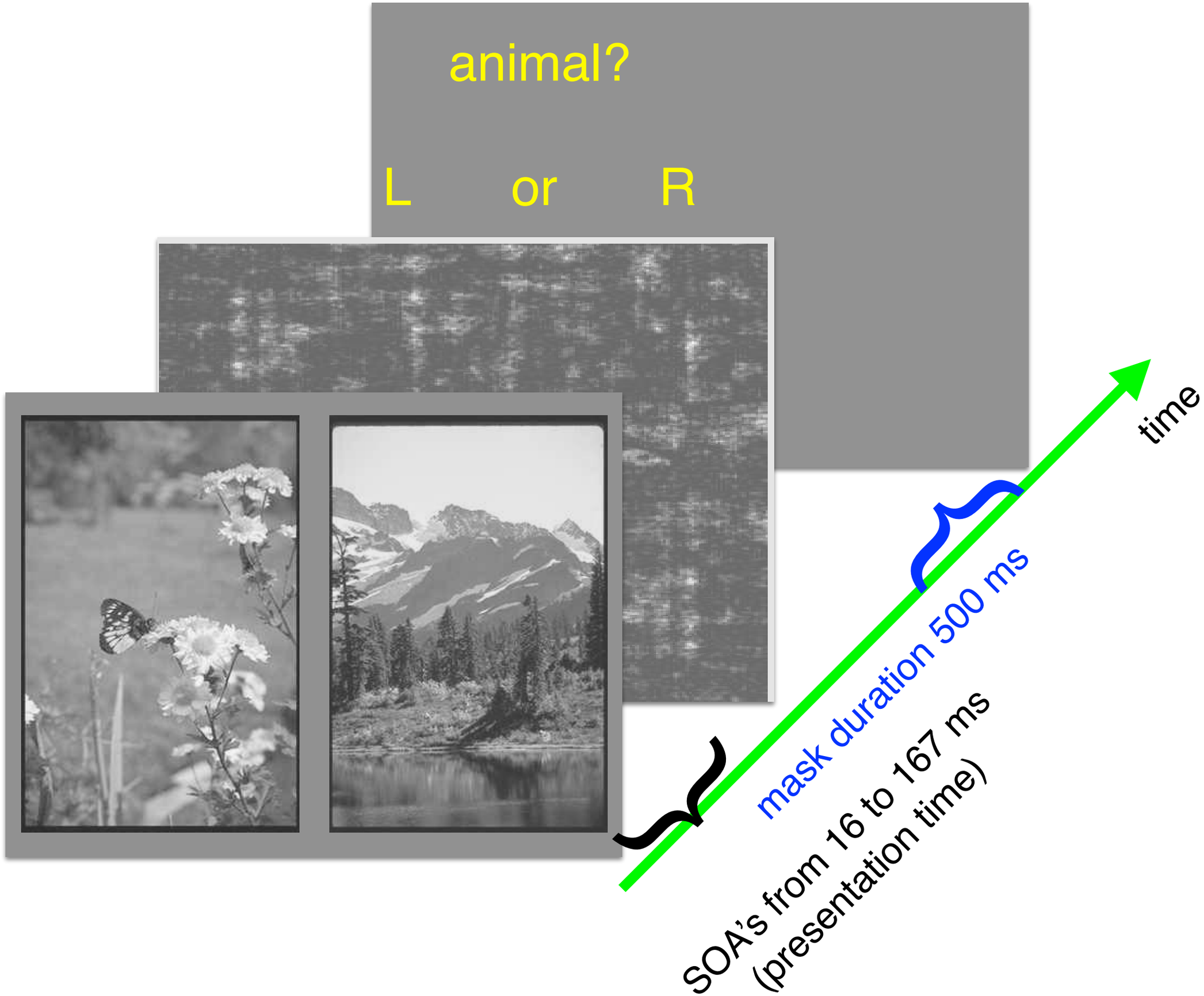
Citation: Wichmann, F. A., Drewes, J., Rosas, P., & Gegenfurtner, K. R. (2010). Animal detection in natural scenes: Critical features revisited. *Journal of Vision, 10*(4):6, 1–27, http://journalofvision.org/10/4/6/, doi:10.1167/10.4.6.

Introduction

The classification of objects in complex, natural scenes is considered a difficult task—certainly from a computational point of view as no computer vision algorithm as yet exists that is able to reliably signal the presence or absence of arbitrary object classes in images of natural scenes. Work by Thorpe, Fize, and Marlot (1996) demonstrated, however, that humans are capable of detecting animals within novel natural scenes with remarkable speed and accuracy: In a go/no-go animal categorization task images were only briefly presented (20 msec) and already 150 msec after stimulus onset the no-go trials showed a distinct frontal negativity in the event related potentials (ERPs). Median reaction times (RTs) showed a speed-accuracy trade-off but for RTs as short as 390 msec observers were already approx. 92% correct (increasing to 97% correct for 570 msec).

This basic result—ultra rapid and accurate animal detection in natural scenes—has been replicated reliably many times: in non-human primates (Fabre-Thorpe, Richard, & Thorpe, 1998; Vogels, 1999a, 1999b), using gray-scale instead of color images (Delorme, Richard, & Fabre-Thorpe, 2000), using different response paradigms and modalities (yes-no or go-no-go versus forced-choice; eye movements versus button presses; e.g. Kirchner & Thorpe, 2006), and while measuring neurophysiological correlates (ERPs; Rousselet, Fabre-Thorpe, & Thorpe, 2002; Thorpe et al., 1996; MEG, Rieger, Braun, Bülthoff, & Gegenfurtner, 2005). Ultra rapid animal detection is even robust to inversion (180 deg rotation) and nearly orientation invariant (Kirchner & Thorpe, 2006; Rieger, Köchy, Schalk, Grüschow, & Heinze, 2008; Rousselet, Macé, & Fabre-Thorpe; 2003; but note that Rieger et al., 2008 found a slight performance decrement for intermediate rotation angles but none for 180 deg inversions). Finally, there are suggestions that rapid animal detection

doi: 10.1167/10.4.6 Received July 30, 2008; published April 15, 2010 ISSN 1534-7362


animal?
L or R
time
mask duration 500 ms
SOA's from 16 to 167 ms
(presentation time)





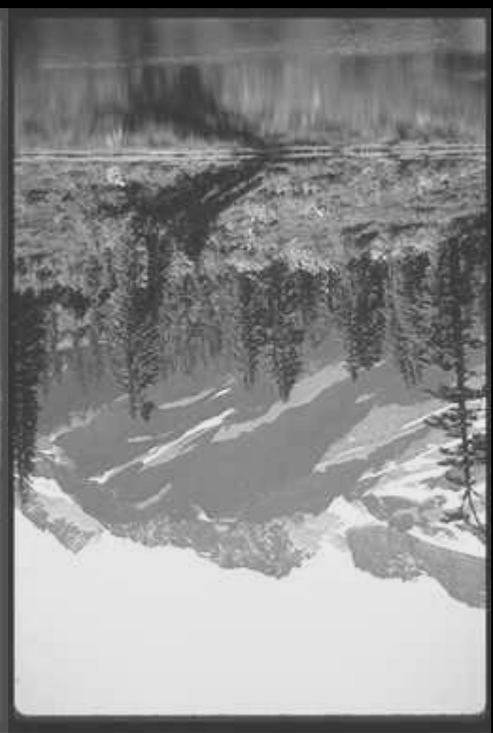


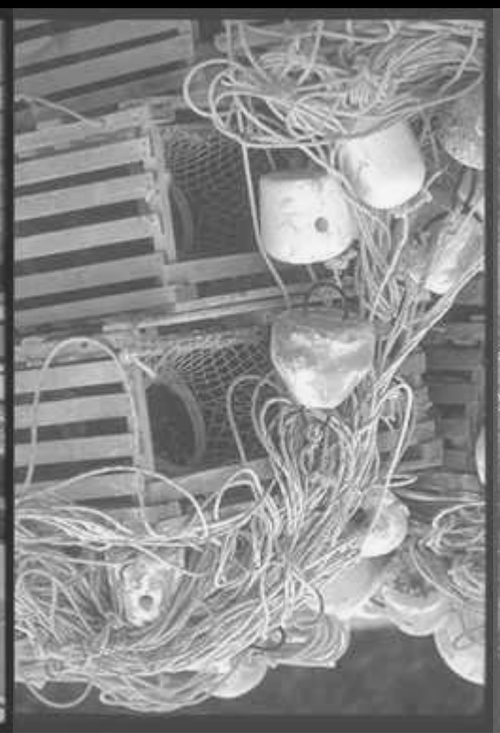


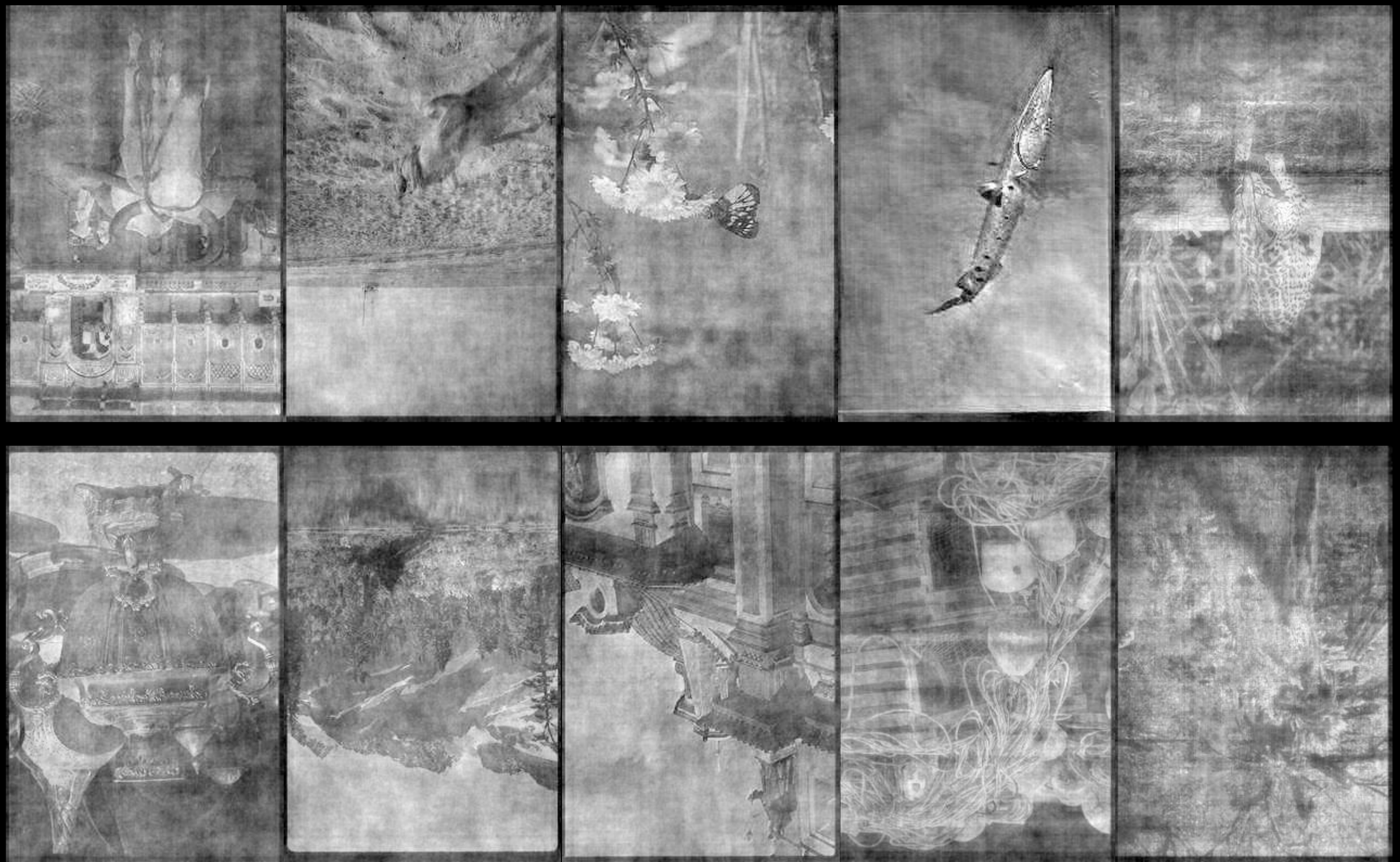




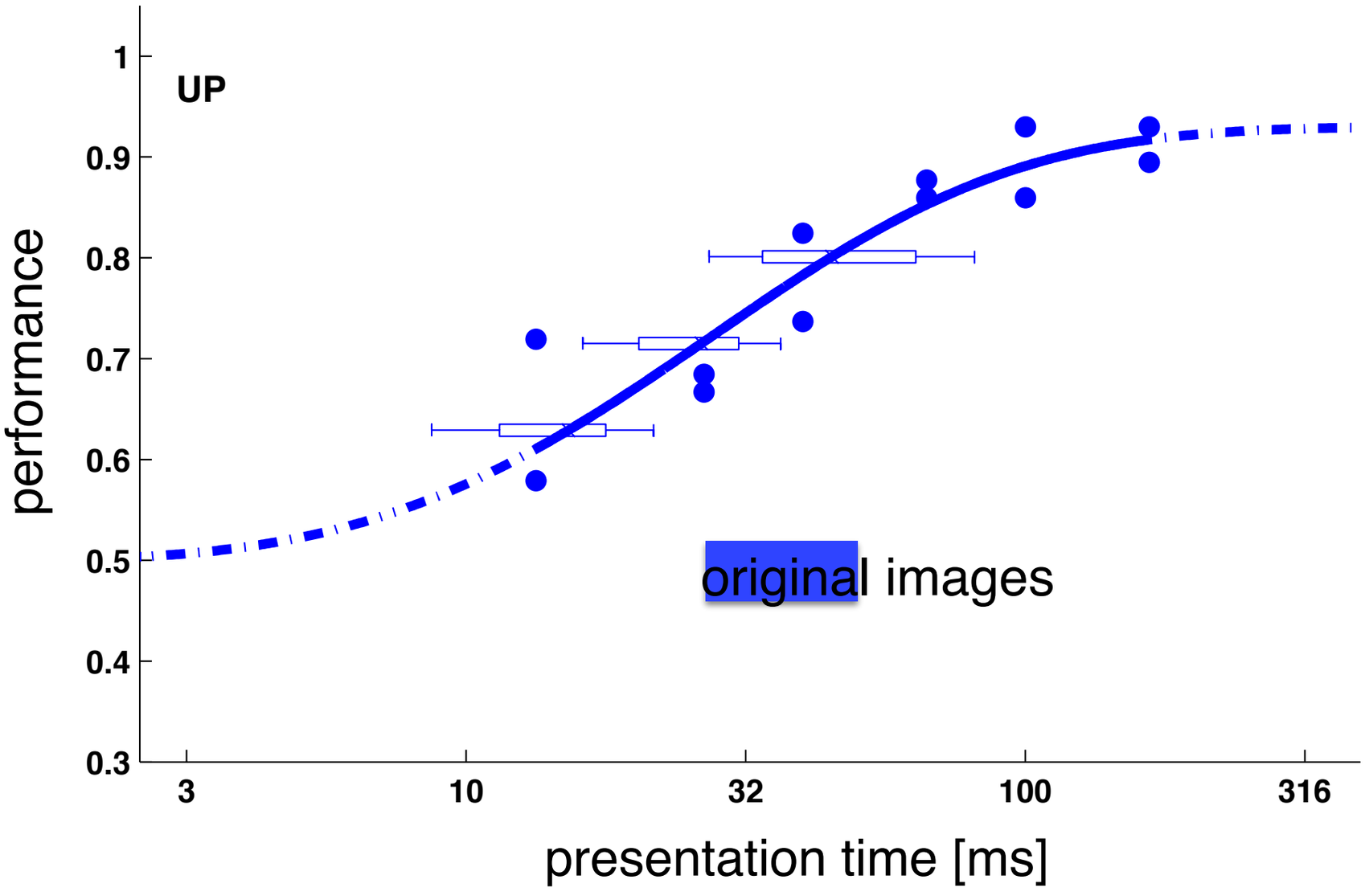


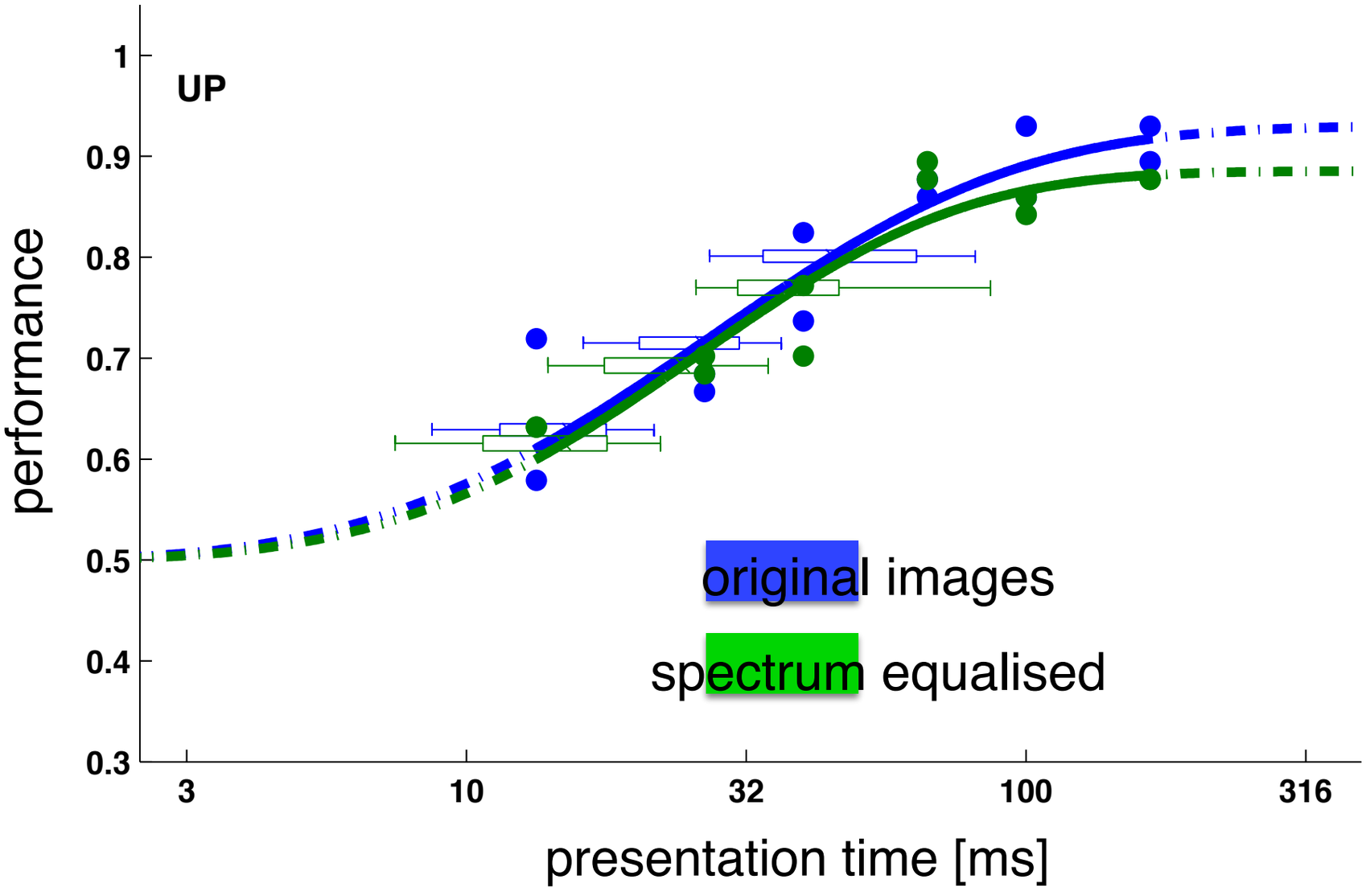


raw data of a single subject



raw data of a single subject



Additional Experimental Variations

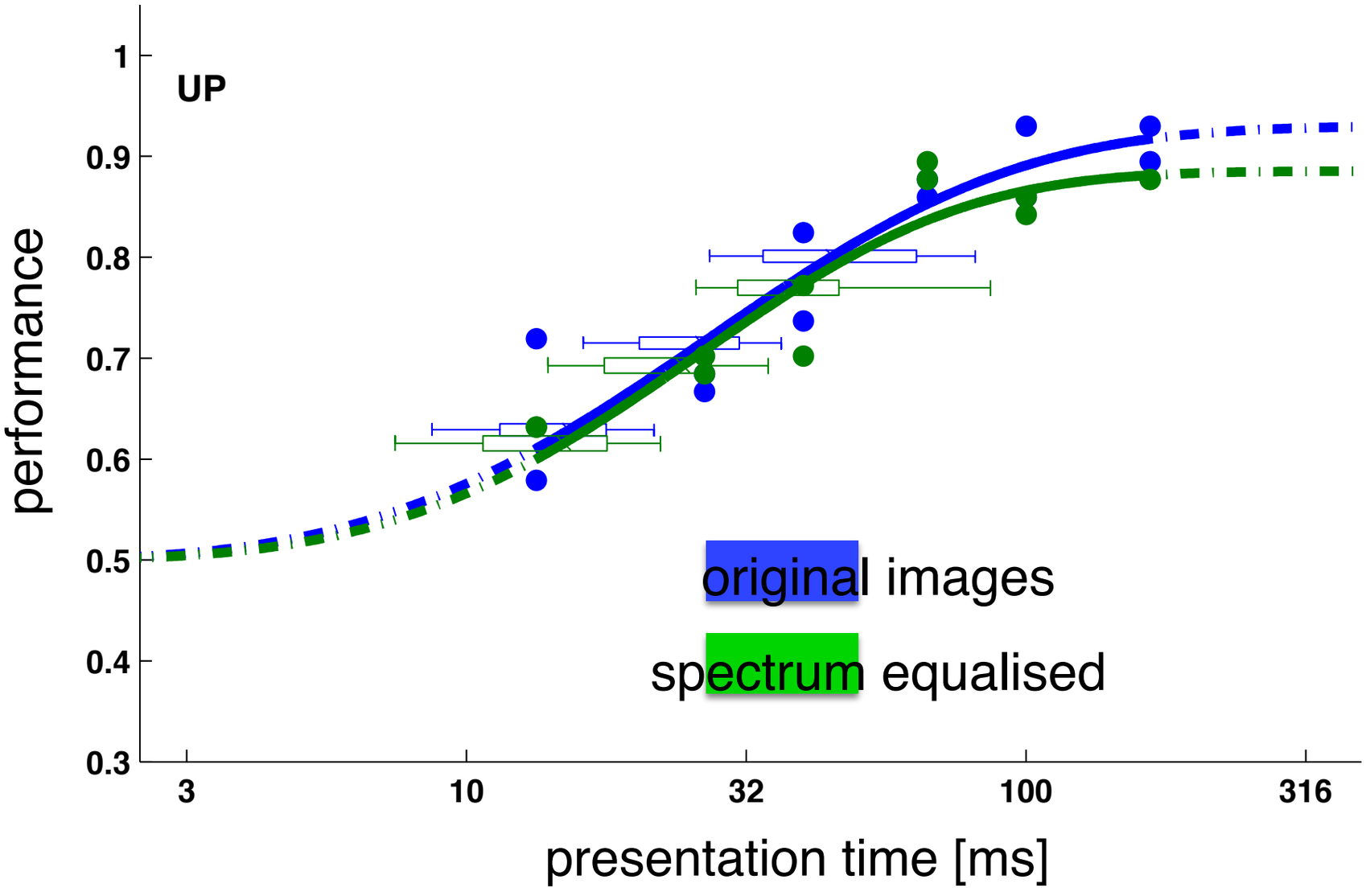
Reaction Time (RT) differences between the conditions?
No appreciable influence.

Task too easy?
Contrast reduction to 50% of original contrast.

Were the distractors critical and not the animals?
Same images but yes-no design.

None of these manipulations made any significant difference.

Asymptotic performance difference?



Second Experiment

Using linear discriminant analysis on the image spectra we classified 11.000 images from the Corel-Database into animals and non-animals, achieving 80% correct classification—*replicating the findings* of Torralba & Oliva (2003).

We retained 800 images, 200 each of the best and worst classified animal and distractor images

Visual Yes-No task: was there an animal?

Again two conditions: i. original images and ii. spectrum equalized.

Presentation time fixed at 30 ms without a mask (approx. 90% correct)

12 naive observers participated in the experiment, resulting in 9.600 trials

animal
or
distractor?

time

30 ms presentation time

Results

- There is an effect on asymptotic performance:
  0.93 ± 0.01 (S.E.) for original images
  0.86 ± 0.01 (S.E.) for spectrum equalised images

- There is an effect of best/worst classified images (original) as predicted by Torralba & Oliva:
  0.95 ± 0.01 (S.E.) for the best classified images
  0.90 ± 0.01 (S.E.) for the worst clasified images

- But there is a at least equally strong effect of best/worst classified images even on the spectrum equalised images:
  0.90 ± 0.01 (S.E.) for the best classified images
  0.81 ± 0.02 (S.E.) for the worst clasified images

- Exactly the same pattern hold for the reaction times (best/worst difference of 21 ms for original, and 36 ms for equalised images, range 471 to 540 ms)

- **High-spatial frequencies appear correlated with, but are not causally related to rapid animal detection of the Corel Database pictures.**

… is any of this really surprising ?

Corel database: professional photographers take professional pictures with professional tripods and cameras. How natural are such ‘‘natural’’ images?

Rapid animal detection works with a single fixation—the quality of the eye, both optics and sensor, is dramatically non-uniform over the visual field!

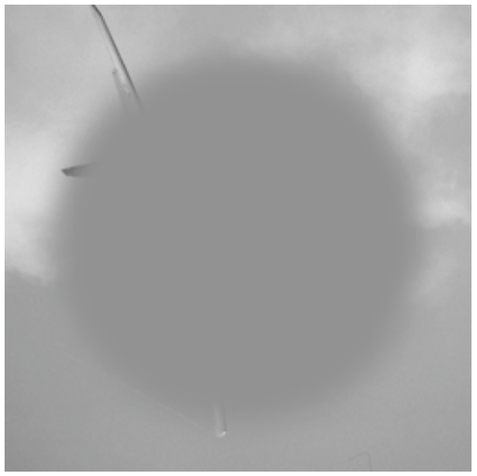
The term ‘‘image management’’ refers to those controls we employ to alter the image formed by the lens and projected on the film. I. To do so, we must understand the differences between the image seen by the human eye and the one seen by the camera. I. Whether we realize it or not, we observe the world from many points of view, not just one I through continuous movements of the eyes, head, and body. The brain synthesizes this continuous exploration into a unified experience. The novice photographer usu- ally learns about the differences between camera and human vision through a series of disappointments I. Examining a developed photograph I the result is not what the photographer believes he saw when he made the exposure, and the effect he recalls is absent or spoiled by intrusions. (Ansel Adams, 1980, ch. 7; pp. 95–96)

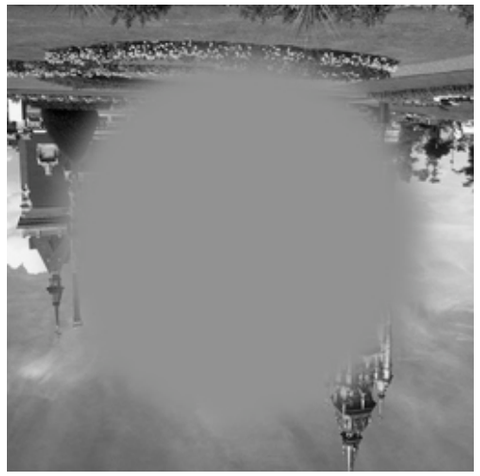
Ansel Adams, "Die Kamera" in *Die neue Ansel Adams Photobibliothek*, Kapitel 7.

Problem 1: Natural Images?

small aperture:
large depth-of-field

large aperture:
shallow depth-of-field

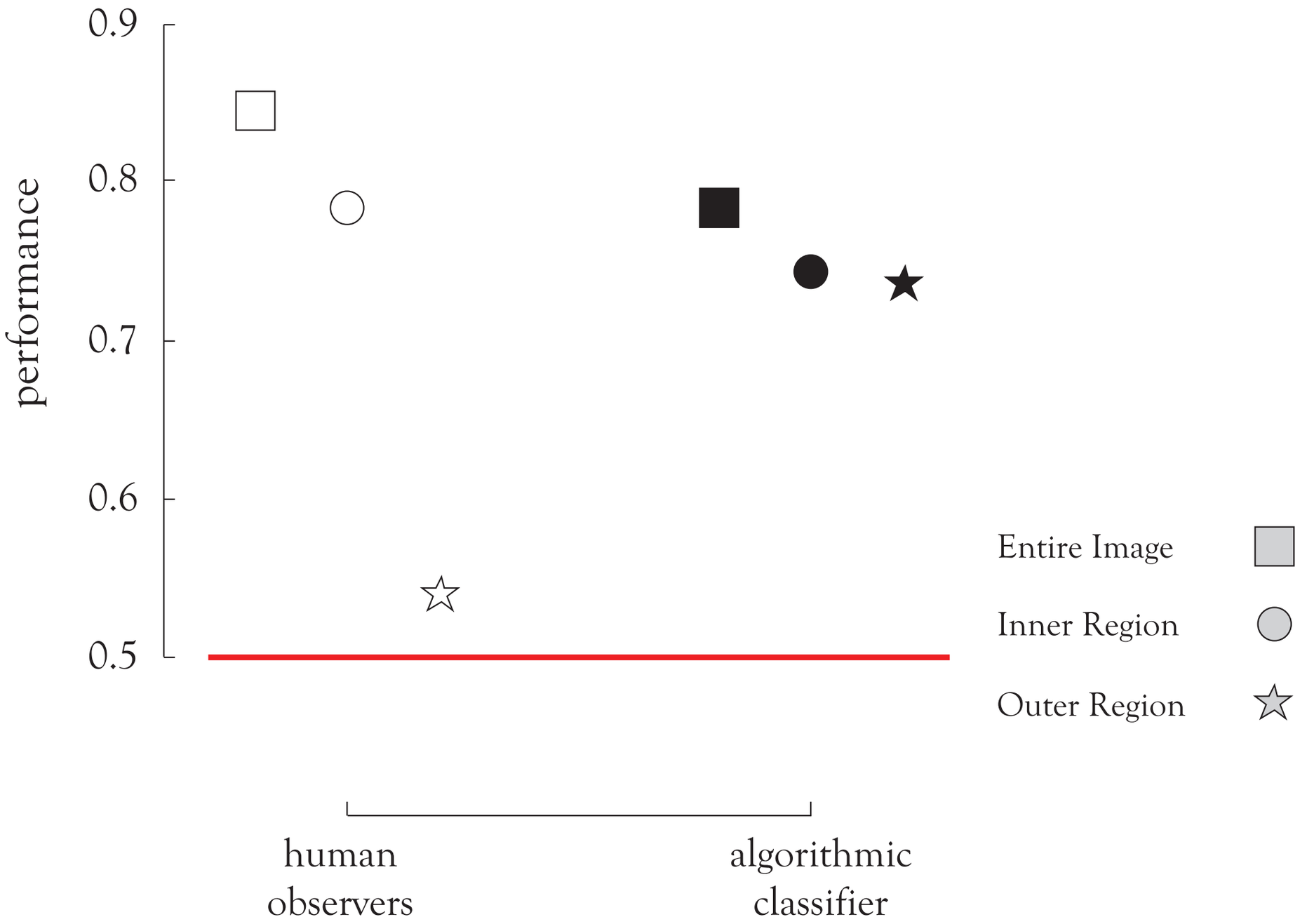









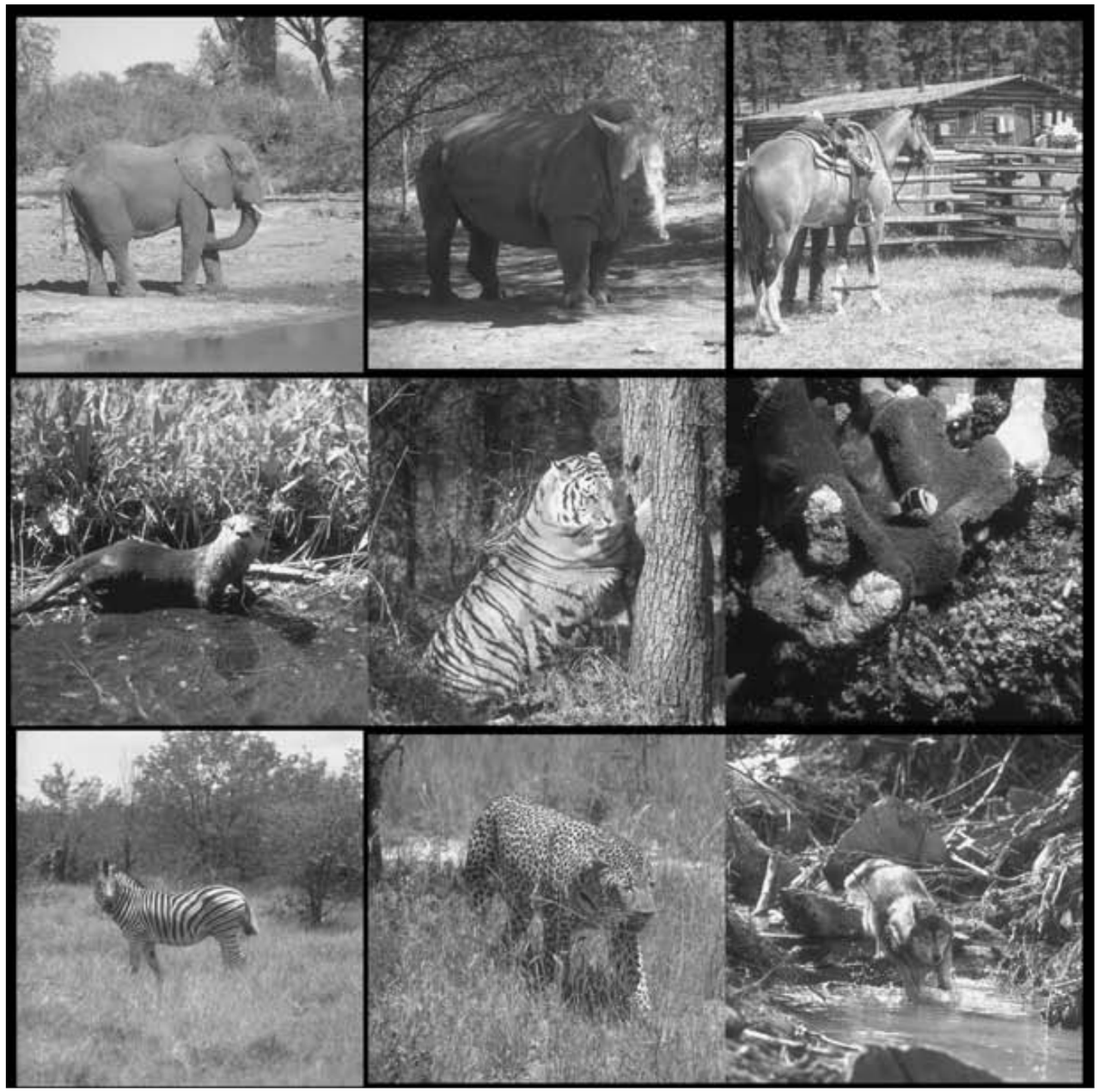



0.9
0.8
0.7
0.6
0.5
performance
Entire Image
Inner Region
Outer Region
human observers
algorithmic classifier

Problem 1: Unnatural ‘‘Natural’’ Images

Effectively, the photographers have segmented the scenes containing animals—the putatively time-consuming contour extraction and object segmentation has been performed.

This is consistent with the finding that the total amount of high spatial-frequencies co-varies with the ease of rapid animal detection, but is not causal.



Problem 2: Spatial Resolution of a Single Fixation

(adapted from Geisler & Perry, SPIE Proceedings, 1998)

Conclusions

Ultra-rapid-animal categorisation is independent of the relative magnitude of high spatial frequencies.

Global image statistics (amplitude and phase spectra, histogram) fail to explain rapid-animal detection: cues are likely local and, alas, more complicated (c.f. Evans & Treisman, *JEP*, 2005; Wichmann, Braun & Gegenfurtner, *Vision Research*, 2006).

For visual tasks that can be accomplished within one or few fixations it may be problematic to try and relate performance to statistics of images with constant spatial resolution.

Many ‘‘natural images’’ may be less natural than commonly presumed.

One would need a database of ‘‘natural’’ natural images, that is, images that technically sound but aesthetically poor. (Camouflage is typically broken using object motion and motion parallax and is pretty effective in still images; c.f. Torralba & Efros, 2011.)

Thus it may be that simply summary or ensemble statistics underly the rapid perception of the gist of a scene, but not the (sometimes) very rapid detection of objects within the scene.