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 © ARVO

may not require focused attention (Fei-Fei, VanRullen, Koch, & Perona, 2002; Rousselet et al., 2002), although this is not uncontroversial (c.f. Evans & Treisman, 2005) and some performance limitations in the complete absence of attention have emerged (Evans & Treisman, 2005; Rousselet, Thorpe, & Fabre-Thorpe, 2004; VanRullen, Reddy, & Li, 2005). Recently Kirchner and Thorpe (2006) showed that, in a spatial two-alternative forced-choice paradigm, human observers could initiate saccades to the image containing the animal even as short as 120 msec after stimulus onset, while Rieger et al. (2005) measured the magnetoencephalogram (MEG) while studying natural scene perception and concluded that only 90 msec of undistorted processing— processing without a backward mask—enabled their human observers to recognize natural scenes in a match-to-sample task.

Human observers are remarkably fast given the, seemingly, very difficult nature of the task. Thorpe and colleagues argued that the results from ultra-rapid-animal detection experiments pose serious constraints on the class of algorithms that may underlie this capability (Thorpe et al., 1996; Thorpe, Delorme, & VanRullen, 2001; VanRullen & Thorpe, 2002). They argue that the rapid animal detection results suggest a coding scheme in the primate visual cortices where few, initial spikes play a crucial role in a code in which spike timing, rather than rate, is critical (but c.f. Gerstner, 2005). Furthermore, so they argue, their results strongly suggest that information processing sufficient to allow rapid animal detection is purely feed-forward. However, most models of contour integration and object segmentation, thought to *precede* object, and hence animal recognition, rely on re-current interaction provided by feedback—and/or lateral connections (for a recent paper and overview see, e.g., Hansen & Neumann, 2008).

Simple proxies for complex perceptual decisions

Motoyoshi, Nishida, Sharan, and Adelson (2007) presented a compelling and elegant solution to the problem of inferring material properties of objects from images projected onto the retina. While surface appearance is a complex function of illumination, surface geometry and reflectance, Motoyoshi et al. (2007) showed how surface gloss and lightness can be inferred from a simple image statistic: the skewness of the luminance histogram. Thus they argued that there is no need for the visual system to attempt to do inverse optics: skewness of the luminance histogram is a very good proxy to the exceedingly complicated, and typically ill-posed, problem of how to disentangle illumination, surface geometry and reflectance (surface material).

Perhaps images of natural scenes can be classified into animal and no-animal categories based on a simple proxy, too: The relevant information for animal classification may be contained in a simple image statistic that can be extracted quickly and at comparatively limited computational cost without the need for image segmentation and, thus, re-current processing (c.f. Bar, 2004). This is exactly what Oliva and Torralba suggested (Oliva & Torralba, 2001; Torralba & Oliva, 2003). They argued that the relative amount of high spatial-frequency energy in the vertical and horizontal orientation could be used to predict seemingly complex decisions about the absence and/or presence of objects in natural scenes. They showed that the information contained in a small number (N = 16) of spectral principal components (SPC)—principal component analysis (PCA) applied to the normalized power spectra of the images—is sufficient to achieve approximately 80% correct animal detection in natural scenes, and they argue: "*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)." and "*(t)hese 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)." This notion is so attractive because it implies that ultra-rapid-animal detection is based on a simple image statistic which can be calculated *prior* to segmentation: predominantly the relative amount of high spatial-frequency energy in the vertical and horizontal directions—a simple proxy like that of Motoyoshi et al.

Because of its elegance this hypothesis was quickly adopted as an explanation for the rapid detection of animals in natural scenes: "*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)" taken from Henderson (2003), p. 501, or "*… it has been shown that natural images can be classified into animal and non-animal 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*" from Johnson and Olshausen (2003), p. 509.

Unlike Motoyoshi et al. (2007), however, Torralba & Oliva did not present any empirical, psychophysical evidence for their suggested proxy. Even if the spectral differences between the image categories were real, however, this does not necessarily imply that the human visual system is able to exploit them. Wichmann and Henning (1998) showed, e.g., that the amount of high spatial-frequency content *could* be a cue for motion segmentation but it is, alas, not used by the visual system.

The aim of the series of experiments reported in this article was thus to assess psychophysically whether human observers make use of the power spectral differences between animal and no-animal images when rapidly classifying natural scenes.

Experiment 1

In the first experiment we measured our observers' ability to detect animals in natural scenes as a function of presentation time (13 to 167 msec); a noise mask immediately followed all images. In one condition we used the original images, in the other images whose power spectra were equalized (each power spectrum was set to the mean power spectrum over our ensemble of 1,476 images). Were the high-spatial frequency content in the vertical and horizontal orientation critical for rapid animal detection, then subjects should perform worse for the spectrum-equalized images, particularly at the very shortest presentation times, where, according to Torralba & Oliva's spatial-frequency based scene gist hypothesis, differential amounts of high-spatial frequency energy in the vertical and horizontal orientations facilitates the animal/no-animal discrimination prior to a detailed feature analysis.

Methods

All 1,476 images—738 animal- and 738 no-animal or distractor images—were taken from the Corel Stock Photo Library (Corel, 1996) and we used mostly the same images used in earlier experiments on rapid animal detection by Thorpe and colleagues as well as Torralba & Oliva. Figure 1 shows examples of the type of animal images (bottom row) and distractor images used; all images were converted from RGB color to gray using MATLABs (The MathWorks, Inc., 2010) standard rgb2gray routine, which is a simple linear combination of 0.2989*R + 0.5870*G + 0.1140*B. Individual images had rather different mean luminances and RMS-contrast (variance) but the animal and distractor images, as groups, had a different variance, too (approximately 10% higher RMS contrast for distractor images)—despite using nearly 1,500 images, these difference did not average out, suggesting that they are systematic. To avoid this potential cue all images were processed to have the same mean luminance and the average RMS-contrast of the categories was equalized (not each individual image as this would have made many images either almost invisible or look grotesque due to artificially high contrast). Figure 2 shows examples of the processed images.

To generate the set of images with equal power spectrum (and hence equal RMS contrast) we first Fourier

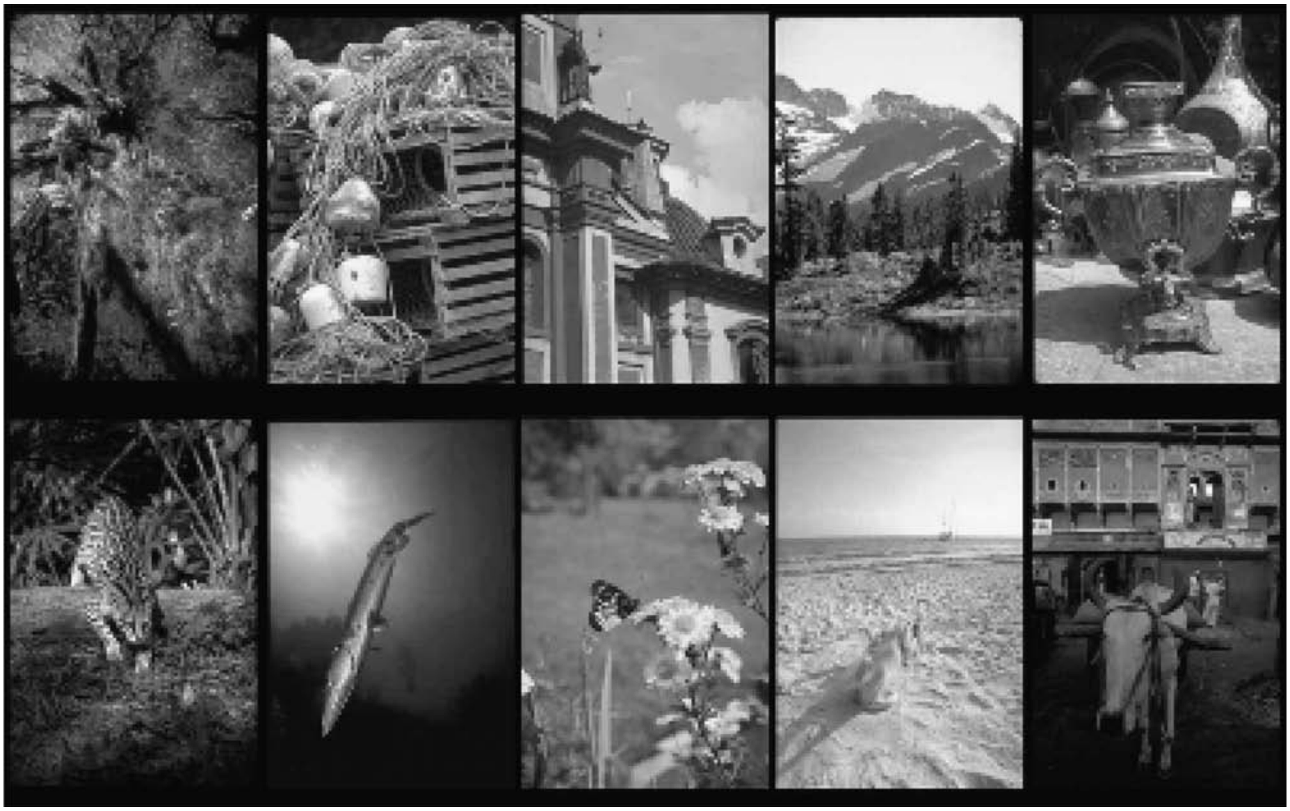

Figure 1. Five animal and five distractor images of the Corel database; the color images of the Corel database were converted to black-and-white.

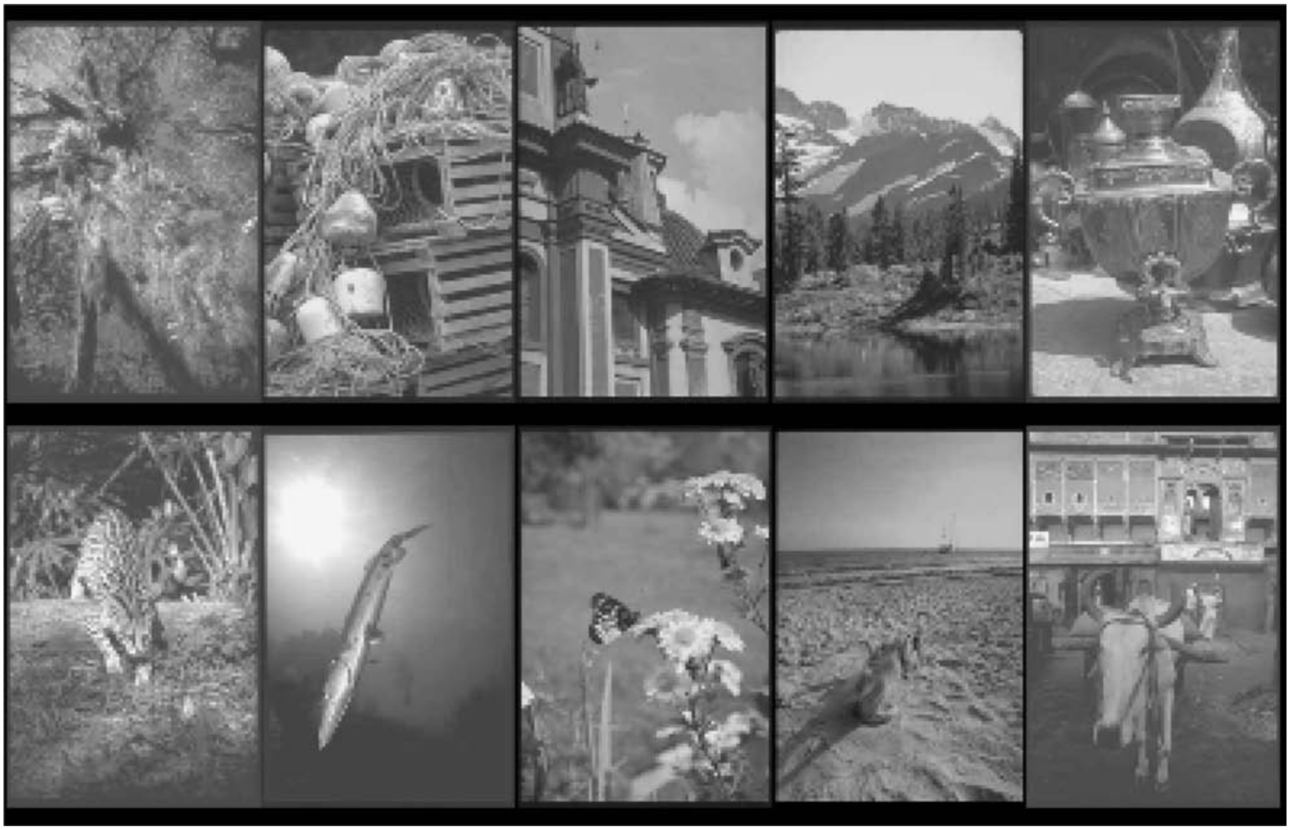

Figure 2. Same as Figure 1 except that the animal and distractor images were processed to have the same mean luminance and, per category, the same RMS contrast (variance; see text for details). In Experiment 1 we used these images to compare performance and refer to them as the original images.

transformed each image and then converted the even (real) and uneven (imaginary) components into amplitude and phase. Then we averaged the amplitude across all 1,476 images, essentially resulting in a 1/f spectrum typical for natural scenes (e.g. Field, 1987). Finally, we combined the original phase spectrum of each image with the mean amplitude spectrum and applied the inverse Fourier transform to obtain the equalized images. As the power spectrum is simply the square of the amplitude spectrum all images thus had the same power spectrum and thus the same RMS contrast. Exemplars from the spectrum-equalized condition are shown in Figure 3. Any algorithm categorizing images into animal and no-animal categories—or any category for that matter—based solely on the power spectrum would thus be at chance with the set of images shown in Figure 3. If humans' ability to rapidly detect animals in images were based on the power spectrum the stimulus presentation time required for correct discrimination should markedly increase.

The experiment was a standard, spatial 2-AFC discrimination task: did the left or the right of the presented images contain an animal? The probability that an animal was presented on the left was 0.5 during the course of the experiment. Stimulus presentation times (SOAs) were 13, 20, 40, 67, 100 and 167 msec immediately followed by a noise mask. The noise masks had the same power spectrum as the spectrum-equalized images but a random phase spectrum. Stimuli were presented using a Cambridge Research Systems VSG 2/5 graphics controller and a carefully luminance calibrated Clinton Monoray monochrome screen at a frame rate of 150 Hz non-interlaced with a spatial resolution of 848 × 636 pixels. Stimuli were viewed from a viewing distance of 150 cm at which individual images subtended 4.62° × 6.15° of visual angle (384 × 512 pixels). The left and right images were separated by a 38 pixel-wide strip of mean gray implying that the total spatial 2-AFC display subtended 9.67° × 6.15° of visual angle. Observers were instructed to fixate the center of the display. Presentation of the stimuli did not change the mean luminance of the display which was set to 70 cd/m^2. During the experiment 114 pairs of images for each SOA were shown, totaling 684 trials per psychometric function, for two conditions (original vs. spectrum-equalized) and five observer (two of the authors (PR, FAW) and three naïve to the purpose of the experiment). Conclusions are thus based on 6,840 2-AFC trials. Experiment 1 and all supplementary experiments to Experiment 1 were conducted at the Max Planck Institute for biological Cybernetics, Department of Empirical Inference in Tübingen while one of us (FAW) was a research scientist there.

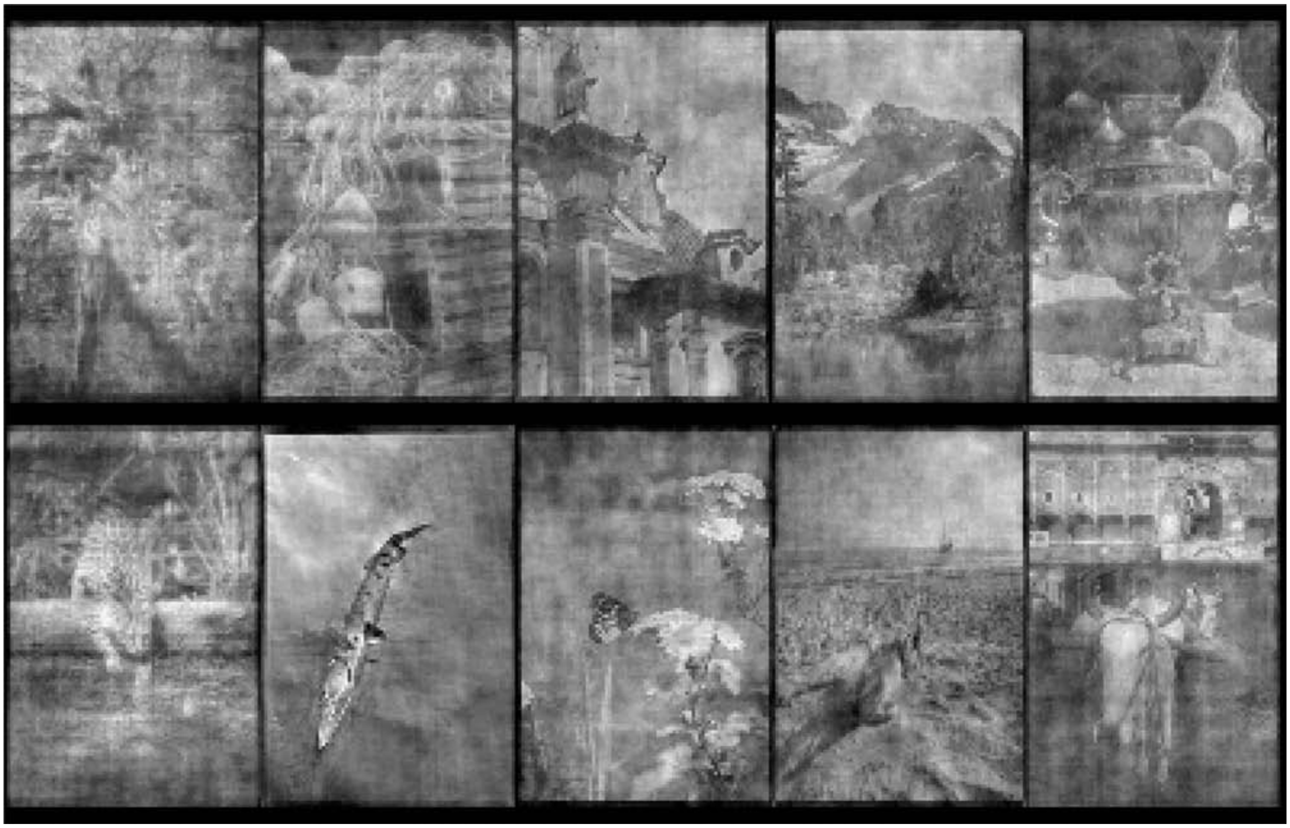

Figure 3. Same as Figure 1 except that all images—animals and distractors—were processed to have exactly the same power spectrum equal to the mean power spectrum of all 1,476 images used (see text for details).

Results

Figure 4 shows proportion correct animal detection as a function of presentation time on semi-logarithmic coordinates for one of the observers (UP). The filled circles show data for the original images, open circles for the spectrum-equalized images. Weibull psychometric functions were fitted to the data using the psignifit toolbox for MATLAB (http://bootstrap-software.org/psignifit) which implements the maximum-likelihood and Monte Carlo resampling methods described by Wichmann and Hill (2001a, 2001b). Inspection of Figure 4 shows, first, that rapid animal detection is indeed very rapid: subject UP only required around 30 msec presentation time (immediately followed by a noise mask) for 75% correct animal detection. Second, there is virtually no discernible difference between the original image and the spectrum-equalized condition, certainly not for the very shortest presentation times. There may be a slight difference in the asymptotic performance level attained by UP, i.e. at the very longest presentation times of 100 and 167 msec.

Figure 5 shows a summary of the results obtained for all five observers. The top panel shows the stimulus presentation times required for 60% correct (downward pointing triangles), the middle panel for 75% correct (circles) and the bottom panel for 90% correct detection

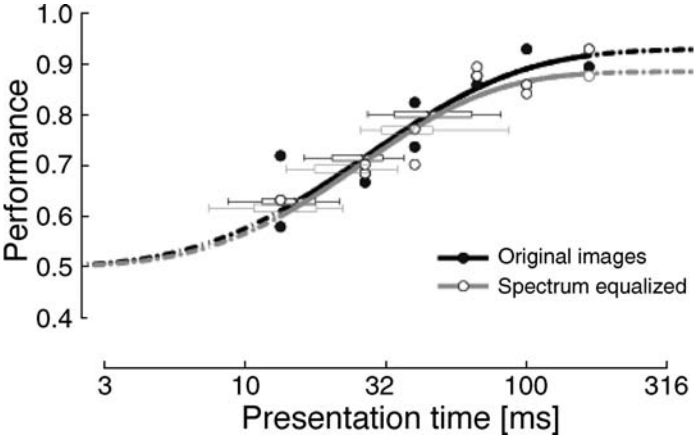

Figure 4. Shows proportion correct animal detection as a function of stimulus presentation time on semi-logarithmic axes; raw data and fitted psychometric functions are shown for one observer, UP. Performance for spectrum equalized images is shown with open symbols and light gray line, for the original images with filled symbols and black line. Boxes and error bars denote the 68% and 95% confidence intervals obtained using a parametric bootstrap procedure (see text for details).

(upward pointing triangles). As in Figure 4 filled symbols denote thresholds for the original images, open symbols thresholds for the spectrum-equalized images. The symbols on the right, with the thick red line around them, indicate the means across observers, as do the thick red lines (dashed for the original images, solid for the spectrum-equalized images). Error bars correspond to the 68% confidence intervals obtained from a parametric bootstrap (Wichmann & Hill, 2001b). Image presentation times are remarkably short, on average around 15 msec for 60% correct, below 25 msec for 75% correct and between 30 and 40 msec for 90% correct animal detection. As for subject UP in Figure 4, the data clearly speak against the Torralba & Oliva hypothesis that the difference in high-spatial frequency content between animal and no-animal images found in the Corel database allows rapid animal detection in human observers. There may be an ever so slight difference for 90% correct (32 msec versus 36 msec) which may have been statistically significant had we conducted many many more trials—it is certainly not statistically significant for the data shown in Figure 5. For the hypothesis under scrutiny this does not matter, however, because, first, the benefit of differential high spatial-frequency content should be largest for the shortest presentation times, not for the longest. Second and more importantly, 36 msec presentation time for 90% correct animal detection is still very rapid indeed, and rapid animal detection in this condition is, by experimental design, definitely not due to the differential high-spatial frequency content in the vertical and horizontal orientations between animal and no-animal images.

Supplementary experiments

In addition, we conducted three supplementary experiments to exclude the possibility that we missed a significant influence of differential high spatial-frequency content on rapid animal detection due to an unfortunate choice of image or experimental design parameters. (By significant we do not only mean statistically significant but in terms of effect size, i.e. much more than a few milliseconds).

Contrast reduction

Figures 4 and 5 suggest that our observers performed remarkably well under the conditions of Experiment 1. A possibility may have been that the differential high-spatial frequency content only boosts performance for the original images if task difficulty is higher. Thus we re-ran Experiment 1 but lowered the RMS contrast of each individual image by 50% (re. the per category RMS contrast equalized images shown in Figure 2, which already have lower contrast than those in the Corel database and shown in Figure 1). Figure 6 shows the same exemplar images as in Figures 1, 2 and 3 but with the same mean luminance, average RMS contrast per category, spectrum-equalized and reduced to 50% contrast. All other experimental parameters and settings as in Experiment 1; the same five observers participated in this supplementary experiment. Figure 7 shows the summary of the results obtained for all five observers. Downward pointing triangles show the stimulus presentation times required for 60% correct, circles those for 75% correct and downward pointing triangles those for 90% correct animal detection. As in Figures 4 and 5 filled symbols denote thresholds for the

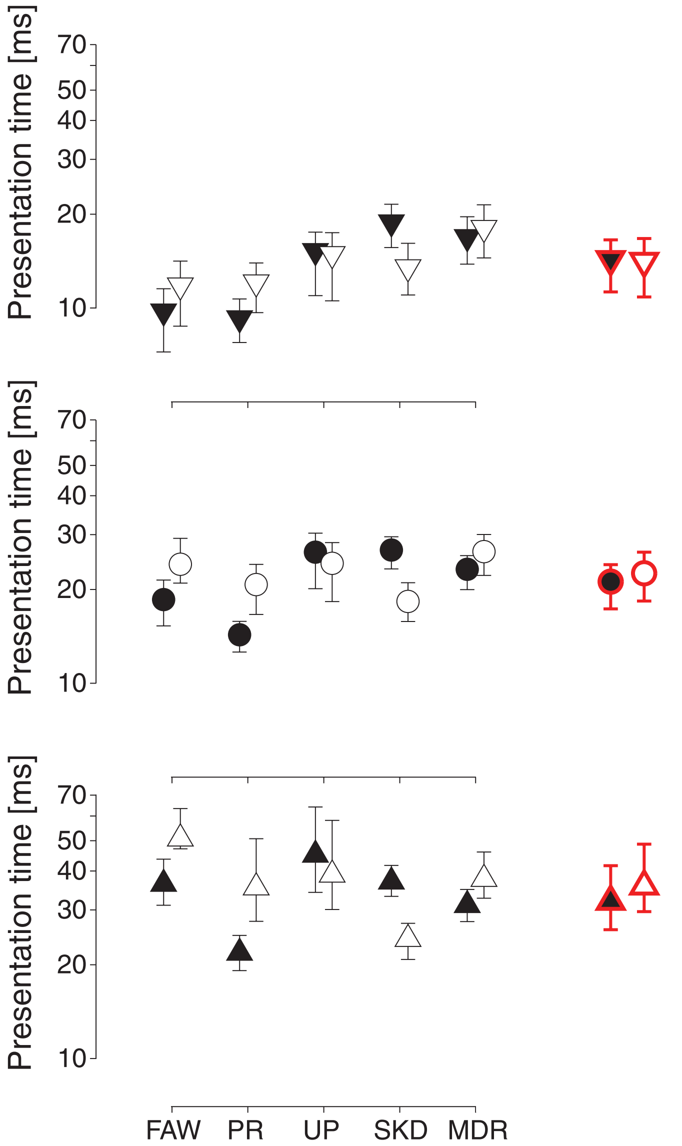

Figure 5. Stimulus presentation time required for 60% (top; downward pointing triangles), 75% (middle; circles) and 90% (bottom; upward pointing triangles) correct animal detection for the five observers who participated in Experiment 1. Open symbols are used for the spectrum equalized images, filled ones for the original images. Error bars correspond to 68% confidence intervals. The symbols on the right, with the thick red edge colors, correspond to the means across observers.

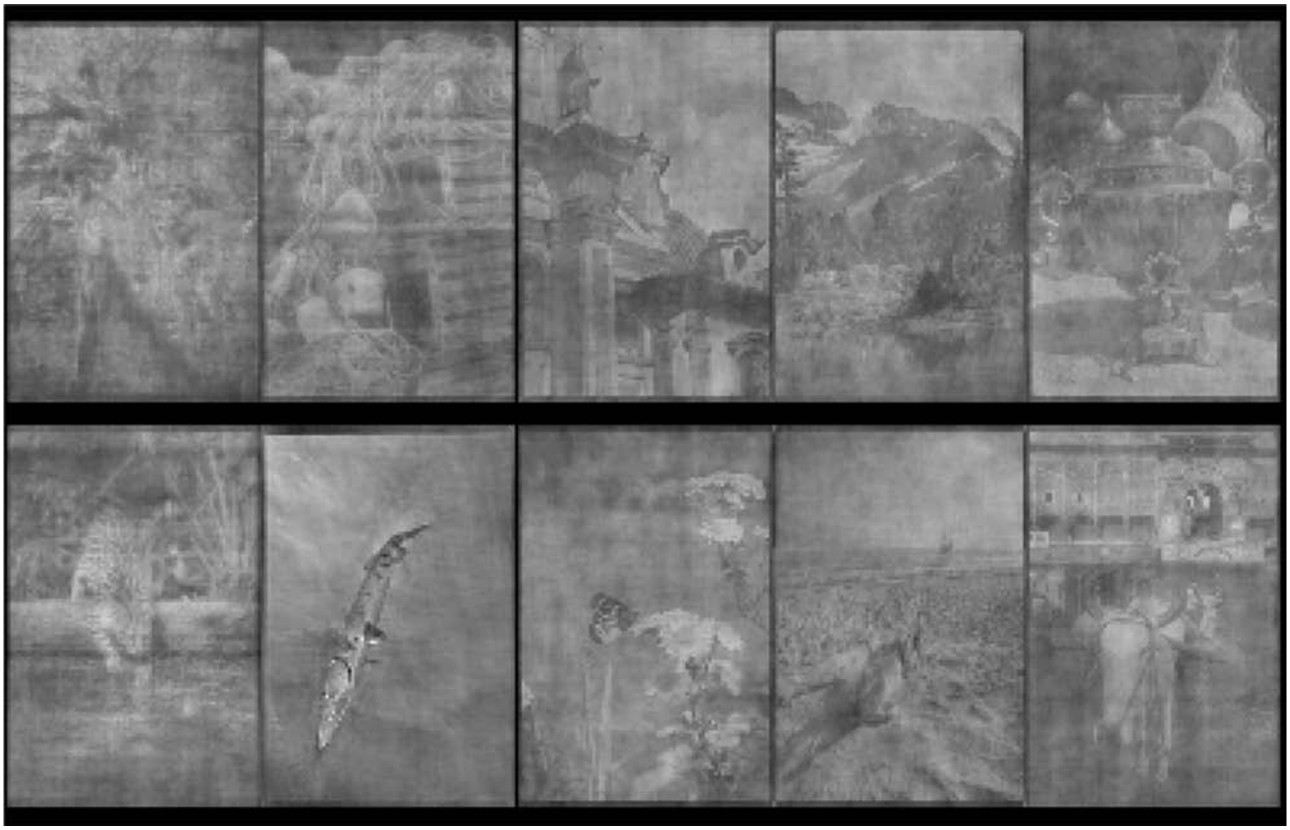

Figure 6. Same as Figure 3 except that the RMS contrast of each image was reduced by 50%.

original images, open symbols thresholds for the spectrum-equalized images. The symbols on the right, with the thick red line around them, indicate the means across observers, as do the thick red lines (dashed for the original images, solid for the spectrum-equalized images). Error bars correspond to the 68% confidence intervals obtained from a parametric bootstrap (Wichmann & Hill, 2001b). Again there are no statistically significant differences between the conditions for at any threshold-level except for observer SKD at 90% correct due to an unusually steep psychometric function in the original image condition. The pattern of results mirrors that for the high contrast condition of Experiment 1 exactly: if there were an effect then perhaps for the highest performance level of 90% (28 msec vs. 34 msec). Not only did the results qualitatively not differ from those of Experiment 1 but in absolute terms the results are virtually identical, too, give or take a few milliseconds (90% correct at 34 msec presentation time immediately followed by a strong visual mask still qualifies to be termed ultra-rapid detection).

Lowering the contrast even further resulted in drastically reduced performance for any SOA, even for 167 msec-informal exploration yielded 50% as the lowest possible contrast at which the task could be performed with at least 90% correct animal detection; at this level the conclusions of Experiment 1 still hold. A similar observation with respect to contrast in natural images was previously

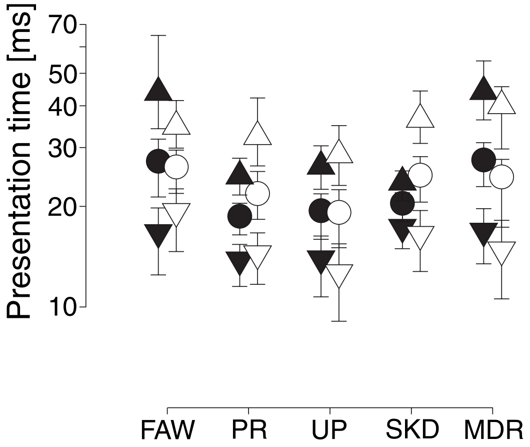

Figure 7. Results for the contrast reduced supplementary experiment. Stimulus presentation time required for 60% (downward pointing triangles), 75% (circles) and 90% (upward pointing triangles) correct animal detection. Open symbols are used for the spectrum equalized condition, filled ones for the condition using the original images. Error bars correspond to 68% confidence intervals. The symbols on the right, with the thick red edge colors, correspond to the means across observers.

reported by Wichmann, Sharpe, and Gegenfurtner (2002) in a visual memory paradigm: once there was sufficient RMS contrast such that stimuli were supra-threshold, memory performance was unaffected by the level of contrast present between "just visible" to original image contrast. Macé, Thorpe, and Fabre-Thorpe (2005) explored the role of contrast in a go/no-go animal/nonanimal categorization task *without noise mask* much more systematically: Macé et al.'s data agree with our informal assessment that around 50% of the original image contrast is needed for the very highest performance (90% correct) and the shortest RTs (see their Figure 3, p. 2011). They were able to show, however, that even at dramatically low contrasts of only 6.25% observers were still performing above chance (57%, see their Table 1, p. 2011).

Experimental paradigm

For a task whose appeal in part stems from its so-called "ecological validity" spatial 2-AFC may be regarded as an unnatural paradigm: if evolutionary pressures had led to extremely efficient animal detection routines in the human brain then a yes-no design may be regarded as more natural. So we re-ran both Experiment 1 as well as the supplementary experiment at reduced contrast as a yes-no experiment (1,368 yes-no trials per image category and observer, i.e. psychometric function; 228 trials per SOA, 50% targets (animals) and 50% distractors). However, results were again virtually identical to those of Experiment 1. Whether spatial 2-AFC or yes-no, whether images at high or low contrast: there is no influence of differential high-spatial frequency content on the stimulus presentation time required for rapid animal detection.

Visual mask

The vast majority of previous animal detection studies did not use a noise masks after showing the images (but see Bacon-Macé, Kirchner, Fabre-Thorpe, & Thorpe, 2007; Bacon-Macé, Macé, Fabre-Thorpe, & Thorpe, 2005). It is perhaps not inconceivable that some backward masking may have interfered with the feedforward high-spatial frequency mechanism postulated by Torralba & Oliva. Thus we repeated Experiment 1 but without a mask following the presentation of each image. Five new observers, all naïve to the purpose of the experiment participated in this supplementary experiment. To reduce the presentation times even further this experiment was conducted using a Sony F-520 display driven at 170 Hz frame rate non-interlaced (800 × 600 pixels spatial resolution); viewing distance was changed to 170 cm in order to keep the subtended angles constant; the mean luminance was slightly lower at 55 cd/m^2. Even at the very shortest possible presentation time of 5.88 msec—a single frame—the average correct animal detection performance was 82 ± 6% in the original condition versus 78 ± 11% in the spectrum-equalized condition: no significant difference. Three of the five observers were slightly better for the original images, but two observers were slightly better for the spectrum-equalized images.

Discussion

Experiment 1 and the three supplementary experiments showed our observers' ability to rapidly detect animals in the Corel database images to be essentially independent of the power spectrum of the images: this result makes it very unlikely that human observers make use of the global power spectrum. Taken together with the results of Wichmann, Braun, and Gegenfurtner (2006), who showed the robustness of animal detection to global phase noise, we are led to conclude that humans use local features, like edges and contours, in rapid animal detection.

Experiment 2

Experiment 2 was conducted to resolve the apparent contradiction that simple, linear pattern classification algorithms can sort images from the Corel database with reasonable accuracy into animal and no-animal images based on the images' Fourier spectrum alone, whereas Experiment 1 and its variants all failed to show any significant influence of the Fourier spectrum on human performance—except perhaps at the asymptotic performance level above 90% correct for presentation times above 100 msec. To investigate this issue we selected a sub-set of images that, according to the spatial-frequency-based classification algorithm, are "special" images: those the classifier classifies easily and correctly (easy images), and those the classifier gets completely wrong, i.e. mis-classifies with confidence (difficult images). We then proceeded to test whether the thus selected images were in any way "special" for our human observers, too.

Methods

Pattern classification and stimulus selection

First we selected a much larger set of 11,000 animal and no-animal images from the Corel database. After gray-scale conversion and square format crops to 480 × 480 pixels, we Fourier transformed each image and binned their amplitude spectra into 8 orientation- and 6 frequency bands, regularly spaced on logarithmic coordinates. Figure 8 shows the average power spectra of animal and no-animal images in the Corel database, before binning. Visual inspection is sufficient to see a systematic difference in the power along the horizontal and vertical directions—exactly replicating the

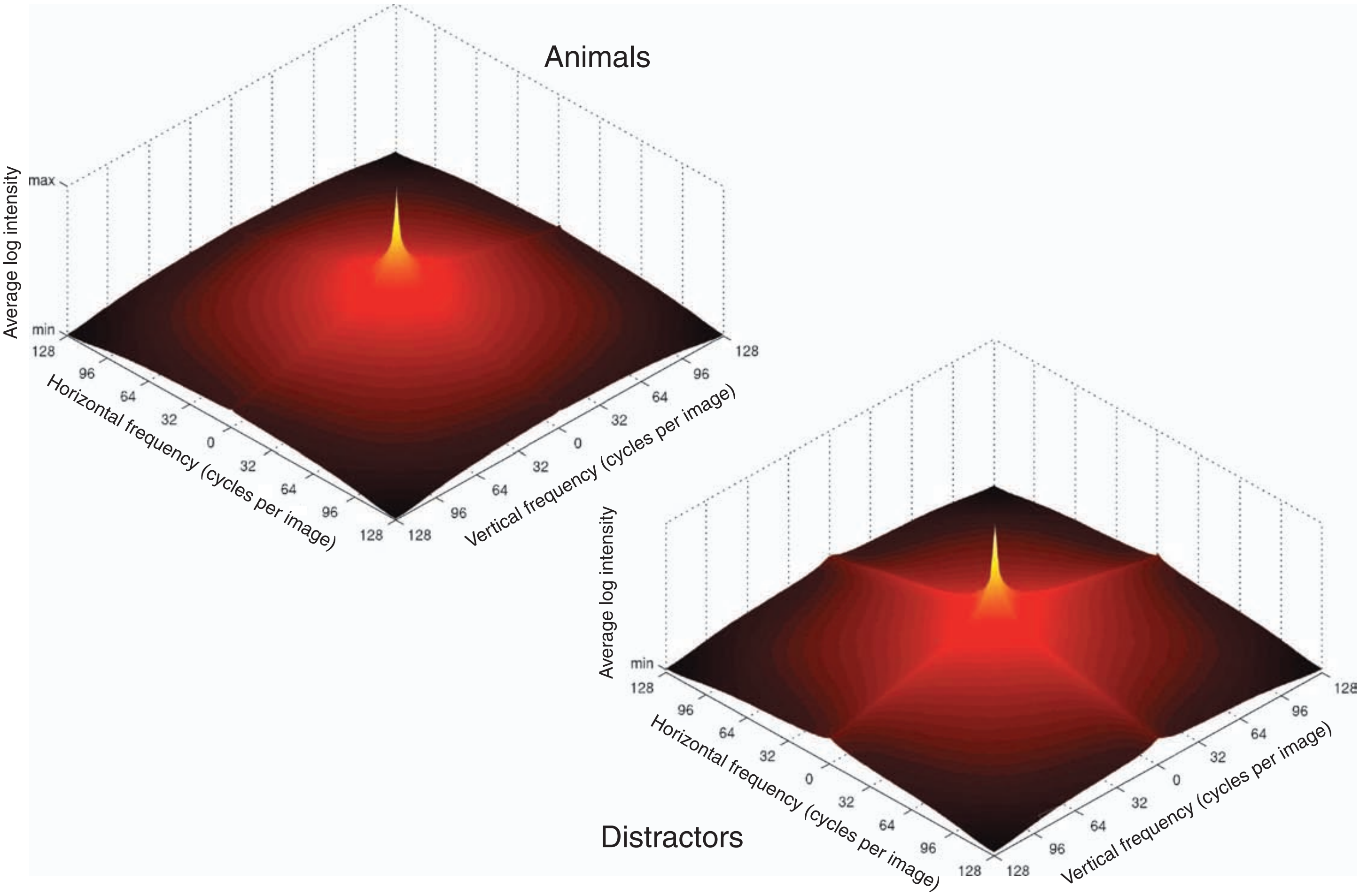

Figure 8. Average animal and distractor power spectra of 11,000 pictures of the Corel Database. The increase in horizontal and vertical high spatial frequencies for the non-animal images is easily visible.

findings of Torralba and Oliva (2003; data on pp. 394–395). We then proceeded to use a linear discriminant analysis to classify the 11,000 images into "animal" and "non-animal" images based on their individual bin values only, achieving slightly better than 74% correct classification. Using principal component analysis (PCA) instead of simple binning as a pre-processing step, we achieved 77% correct classification (similar to Oliva and Torralba, 2001 and Torralba and Oliva, 2003).

After classification we proceeded to sort the images based on the "confidence" (score) which linear discriminant analysis attaches to each classified stimulus. From the 11,000 images we retained only 800: 200 "best animals" (those which the classifier classified correctly as animals with high confidence), 200 "best distractors" (correctly and with confidence classified as distractors), 200 "worst animals" (misclassified with confidence as distractors by our classifer) and 200 "worst distractors" (misclassified with confidence as animals by our classifier). In the following we will refer to the "best animals" and "best distractors" as the *easy* images (for the classifier) and the "worst animals" and "worst distractors" as the *difficult* images. In a final step we again equalized the amplitude spectra of the 800 images as described previously, which, by design, reduced algorithmic performance of our classifier to chance.

Experimental setup

The experiment used a Go/No-Go gap paradigm, similar to that originally used by Thorpe et al. (1996) to elicit very fast responses: Observers were instructed to push and hold a trigger button prior to each trial. When the button was pushed, a fixation dot appeared on the screen and remained uniform at mean luminance for a random period between 500 and 700 msec, followed by a 200 msec gap, in which the fixation dot disappeared. A single target (animal) or distractor (non-animal) image was then shown for 30 ms. We fixed the presentation time at 30 msec without a mask to achieve approximately 90% correct performance in order to be in the ceiling region of Experiment 1, where there may have been a difference between the original and spectrum-equalized images. Thereafter, a small fixation cross re-appeared for 1,000 msec during which the observers were to make their decision by either holding the button pressed steadily

("No-Go"-response, indicating they had believed the image to be a distractor) or releasing the button as quickly as possible ("Go"-response, indicating they had believed the image to contain an animal). After a "Go"-response, the next trial would not start again until the observer pressed the trigger button to signal her readiness. Each of our 800 selected images was shown in random order exactly once to each observer, resulting in a total of 800 trials per condition—i. original images and ii. spectrum-equalized-per subject, taking between 40 and 60 minutes of time depending on the subject's individual pace. Images were presented on a carefully luminance calibrated Iiyama VisonMaster 513 (MA203DT) 21″ CRT screen at a frame rate of 100 Hz non-interlaced with a spatial resolution of 1280 × 960 pixels. Stimuli were viewed from a viewing distance of 45 cm at which the screen subtended 48,2° × 36,9° of visual angle; individual images subtended 19,1° × 19,1° (480 × 480 pixels). A chin rest was used to stabilize head position; ambient lighting was not calibrated to a particular level, but great care was taken to ensure that the same lighting conditions were in effect for all observers. A warm-up period of at least one hour allowed the CRT to stabilize its mean luminance. Twenty-two subjects participated in this experiment; 10 observers did the original images condition and another set of 12 observers did amplitude equalized condition. Overall 9,600 trials were collected for the amplitude equalized condition, and 8.000 for the original-image condition, totaling 17,600 trials. Of those, 58 were discarded because their response time was faster than 200 msec. The remaining trials were considered valid. All observers were students of the Justus-Liebig-University Giessen and were paid for their participation. Observers were between 20 and 31 years of age and had normal or corrected to normal vision.

Results

The mean percent correct and RTs are shown in Table 1. For all ANOVAs reported in this article we corrected the proportions using the variance stabilizing Arcsine-transform (Fleiss, 1981; Hogg & Craig, 1995). Some form of variance stabilization is required when applying ANOVA to binomial data; the Arcsine-transform is reliable unless many entries are either 0 or 1, which they were not for our experiments. In the "original-image" condition, we find the difference between "easy" and "difficult" images to be highly significant both in terms of percent correct (repeated measures one-way ANOVA, $F(1,9) = 30.621$, $p < 0.001$ and response time ($F(1,9) = 8.569$, $p = 0.014$). Overall, observers performed 0.93 ± 0.01 (*SE*) correct for the original images, but 0.95 ± 0.01 (*SE*) for the "easy" images and only 0.90 ± 0.01 (*SE*) for the "difficult" images.

Amplitude spectrum	Image type	Percent Correct	Response Time
All	All	89.3%	450.6 ms
Original images	Easy	95.2%	437.7 ms
	Difficult	90.5%	451.1 ms
	All	92.7%	443.5 ms
Spectrum-equalized	Easy	90.1%	449.3 ms
	Difficult	81.6%	464.4 ms
	All	86.0%	457.7 ms

Table 1. Percent correct and RTs obtained in Experiment 2 for the different image categories (all, original images and spectrum-equalized).

At first sight the significant difference between "easy" and "difficult" images—i.e. a correlation between human observers and the machine classifier—may seem to support the high spatial-frequency hypothesis. However, this is emphatically not the case: When performing the same analysis on the amplitude-normalized image data, we still find a highly significant effect for both percent correct ($F(1,9) = 119.301$, $p < 0.0001$) and response time ($F(1,9) = 27.584$, $p < 0.001$). Overall observers performed 0.86 ± 0.01 (*SE*) for spectrum-equalized images, but 0.90 ± 0.01 (*SE*) for the "easy" images and only 0.81 ± 0.02 (*SE*) for the "difficult" images. Exactly the same pattern holds for the reaction times, where the easy/difficult image difference is 21 msec for original images and 36 msec for equalized images (RT range 471 to 540 msec). This pattern of results suggests that high-spatial frequencies in the vertical and horizontal directions are correlated with whatever (local) cue allows rapid animal detection without being causal.

We also find a very strong effect for the amplitude-normalization in both percent correct ($F(1,9) = 131.8567$, $p < 0.001$) and response time ($F(1,9) = 21.2665$, $p < 0.001$); see above: overall 0.93 ± 0.01 (*SE*) correct for the original images but only 0.86 ± 0.01 (*SE*) for spectrum-equalized images. Thus at high performance levels there is a statistically significant difference between the conditions, as hinted at in Experiment 1 for the longest presentation times. In absolute terms one should not forget that performance for the spectrum-equalized images at only 30 msec presentation time was still 86% correct—far from being "low", or near threshold and still better than our best classifier at just below 80% correct on the original images.[1]

Discussion

Images "easy" or "difficult" to classify as animal or non-animal image based on their spatial-frequency content are "easy" and "difficult" for human observers, too. However, the "easy" images remain "easy" even if the spatial-frequency difference is removed: the same images as before were still classified better and faster by our human observers after spectrum-equalization, strongly suggesting that even in the original condition features other than specifics of the amplitude spectrum made particular images "easy". *High-spatial frequency differences between animal* *and no-animal images are correlated with, but are not causally related to, rapid animal detection for images of the Corel Database*. This finding is clearly at odds with the Torralba & Oliva hypothesis that high-spatial frequency differences play a dominant *causal* role in rapid animal detection.

Supplementary experiment

Experimental paradigm

As for Experiment 1 we wanted to ensure that we did not miss a causal role for high spatial-frequency content differences due to idiosyncrasies of the experimental procedure employed. Thus we repeated the above experiment but this time using standard, spatial 2-AFC instead of a Go/No-Go paradigm. Furthermore, instead of relying on responses being typed into a keyboard as in Experiments 1 and 2 above, here we use saccades as indicator: observers had to perform a saccade to the image containing the animal, i.e. to the left or to the right. Finally, unlike for the main Experiment 2 reported above, where different groups of observers completed the different experimental conditions—original images versus spectrum-equalized images—here the same set of observers completed both experimental conditions.

All eye-movement measurements were performed using SR Research's EyeLink II eyetracking system. Images were presented on the same Iiyama VisonMaster 513 (MA203DT) 21″ CRT screen using the same settings as described above. Ten observers participated in this experiment, completing 800 2-AFC trials each. All observers were students of the Justus-Liebig-University Giessen and were paid for their participation. Observers were between 20 and 31 years of age and had normal or corrected to normal vision. All observers were naïve to the purpose of the experiment. All 10 observers were able to perform eye movements well enough to earn the best calibration accuracy rating from the EyeLink II system. For data analysis we defined three criteria eye-movement traces had to fulfill. First, *goodness of fixation*: Fixation between trial start and stimulus onset was not allowed to diverge from the screen center by more than 70 pixels; this made certain that the observers were not biased in favor of the L or R location. We had to eliminate 129 trials of our 8,000, leaving 7,871 trials for further analysis. Second, *goodness of saccade direction and destination*: Decision saccades were required to end in one of the two smallest possible squares covering the area that was occupied by the target and distractor images. This ensured that the saccades actually were directed into the general area of either of the shown images and eliminated random saccades, e. g. from lack of concentration. Only 7 trials needed to be eliminated to fulfill the requirements, leaving 7,864 trials for further analysis. Third, *goodness of response time:* Observers needed to make their decision no later than 700 msec after stimulus onset to ensure that all responses were reasonably fast. Responses also had to be slower than 80 msec to eliminate random and too early eye movements (reaction times faster than 80 msec can safely be assumed not to be based on stimulus content). 26 trials were removed mainly because they were too fast, leaving a total of 7,838 valid trials for the statistical analysis presented below.

Results and discussion

As perhaps expected from the results of Experiment 2 above, the effect of image difficulty on the hit ratios was significant; additionally, the effect of image difficulty is significant both for the target images (animals, "easy-difficult" in Table 2, $F(1,9) = 82.757$, $p < 0.001$) and the distractor images (non-animals, "difficult-easy" in Table 2, $F(1,9) = 78.161$, $p < 0.001$). The overall performance degradation for of the amplitude equalized images was also significant ($F(1,9) = 87.188$, $p < 0.001$).

As for the main experiment, there is a significant difference between "easy" and "difficult" images, i.e. a correlation between human observers and the machine classifier: if in the 2-AFC both animal and distractor image were taken from the "easy" set, performance was 95.3% correct (*SE* of the means only around ±1–2% for all conditions). Performance dropped dramatically to 76.3% if both images were taken from the "difficult" set. However, for the spectrum-equalized images the drop was, if anything, even more severe, from 89.2% to near chance performance of 60.5% despite that the "easy-easy" and "difficult-difficult" image pairs had *identical* power spectra. Thus the spectral differences in the original images cannot be the cause for the very marked performance difference—they are only correlated with whatever feature(s) cause the behaviorally relevant difference.

Finally it may be instructive to note that for the 2-AFC task the influence of amplitude equalization on RTs is not significant ($F(1,9) = 0.018$, $p = 0.896$), unlike for the Go/No-Go paradigm. Thus it may be that the longer RTs for the amplitude equalized condition reported in Experiment 2

	Type of image pairing	Percent Correct	Response Time
All	All	81.0%	277.5 ms
Original images	All	85.8%	278.4 ms
	Easy-Easy	95.3%	269.9 ms
	Easy-Difficult	81.9%	281.9 ms
	Difficult-Easy	89.7%	269.5 ms
	Difficult-Difficult	76.3%	288.8 ms
Spectrum-equalized	All	76.3%	276.8 ms
	Easy-Easy	89.2%	269.3 ms
	Easy-Difficult	74.9%	279.4 ms
	Difficult-Easy	80.6%	276.9 ms
	Difficult-Difficult	60.5%	282.0 ms

Table 2. Percent correct and RTs obtained in the Supplementary Experiment 2 for the different image categories.

are the result of a response criterion shift—in the Signal Detection Theory sense—for the less "clean" looking spectrum-equalized images. Whether this speculation is true or not, our finding—that classification accuracy shows the same pattern of results for 2-AFC and Go/No-Go, while RTs do not—emphasizes the importance of conducting experiments using different experimental paradigms when evaluating complex visual tasks such as animal detection.

Decidedly non-natural "natural" images

The results of Experiments 1 and 2 and their associated supplementary experiments refute the Torralba & Oliva hypothesis that differential amounts of high-spatial frequency power in the vertical and horizontal orientation helps, let alone underlies, rapid animal detection. What these experiments did show, however, is that there is a non-causal correlation between the amount of high spatial-frequency power in the Corel database images and the ease with which human observers can classify the images into animal and no-animal categories.

One possible explanation is the following: The images in the Corel database—and any other, professional stock photograph database for that matter—were taken by professional photographers. When professional photographers take pictures, particularly of the non-radical or non-arty style found in stock databases, they usually adhere to a small number of rules-of-thumb, listed in every beginners book on photography: Landscapes are photographed with wide-angle lenses at small apertures focused at the hyper-focal distance of the lens; together all of these factors maximize depth-of-field, i.e. everything in the image is sharp (in focus). Portraits, on the other hand, are taken using tele lenses (medium tele for human portraits, long tele lenses for animals, particularly for those of the dangerous type) set to the largest possible aperture with the focus point exactly on the animal (human) to minimize depth-of-field. Minimal depth-of-field is considered desirable for portraits because it pre-segments the (focal) target object (animal, man) from the background.[2] This pre-segmentation is commonly perceived as aesthetically pleasing due to the lack of high spatial frequencies in the background, which is perceived as visually "quiet", and therefore not distracting. If we are correct, then the difference in high spatial-frequency content between animal and no-animal pictures has nothing to do with animals or non-animals per se: it is an artifact of the type of photos customers expect—and pay for—in stock photo databases. Photographers produce this difference because they attempt to pre-segment the portrait photograph for the viewer, and the only way the photographer can do this is by selective depth of field, i.e. high spatial-frequency reduction (low-pass filtering) of the background, typically mainly behind and above the person or animal.

The effect of the very narrow depth-of-field can be clearly seen in the images in Figure 10. All of them belong to the "easy" category of the spatial-frequency based classifier, i.e. our implementation of the Torralba & Oliva algorithm described above classified them correctly and with confidence. Figure 11 shows examples of the misclassified "difficult" animal pictures. All of them were taken with a much smaller aperture and thus more depth-of-field. Figures 11 and 12 show the equivalent selection of distractor images: "Easy" distractors in Figure 11, containing large depth-of-field street or nature pictures, and images at a single depth-plane, i.e. where all information is contained in a single focal plane and is thus completely in-focus. Figure 12, finally, contains the misclassified distractor pictures, and they are strikingly instructive: they are "portraits" of objects, taken at large apertures or still—lives against a uniform background—i.e. devoid of high spatial-frequencies. All the 800 images are available as Supplementary Information.

If we are correct, i.e. that the different orientation content at high spatial-frequencies between animal and no-animal categories are not only non-causal, but actually at least in part a photographic artifact of the Corel Stock database, then we can make the following strong prediction: The algorithm bases its decision about the absence or presence of the animal on the background only—hence *removing the animal from the picture altogether* should affect algorithmic performance very little, whereas it should, of course, lead to chance performance for human observers. If correct, we should thus be able to create a "double-dissociation" between the algorithm and human observers: In Experiments 1 and 2 such an algorithm performs at chance level but human observers are essentially unaffected by the spectral equalization. The animal removal in Experiment 3 should reduce human observers to chance but leave algorithmic performance essentially unaffected.

Experiment 3

In the preceding section we hypothesized that professional photographers cause the spectral differences between animal and no-animal categories and that this pre-*segmentation,* rather than the amount of high spatial-frequency energy, may be helping rapid animal detection. Inspection of Figures 9, 10, 11, and 12 may lend credibility to our claims but cannot, of course, replace a proper psychophysical experiment.

Methods

Stimuli

From the 11,000 Corel database 768 × 512 pixel images we selected a random subset of 800 images, 400 animal and 400 non-animal images; we then manually selected a 480 ×

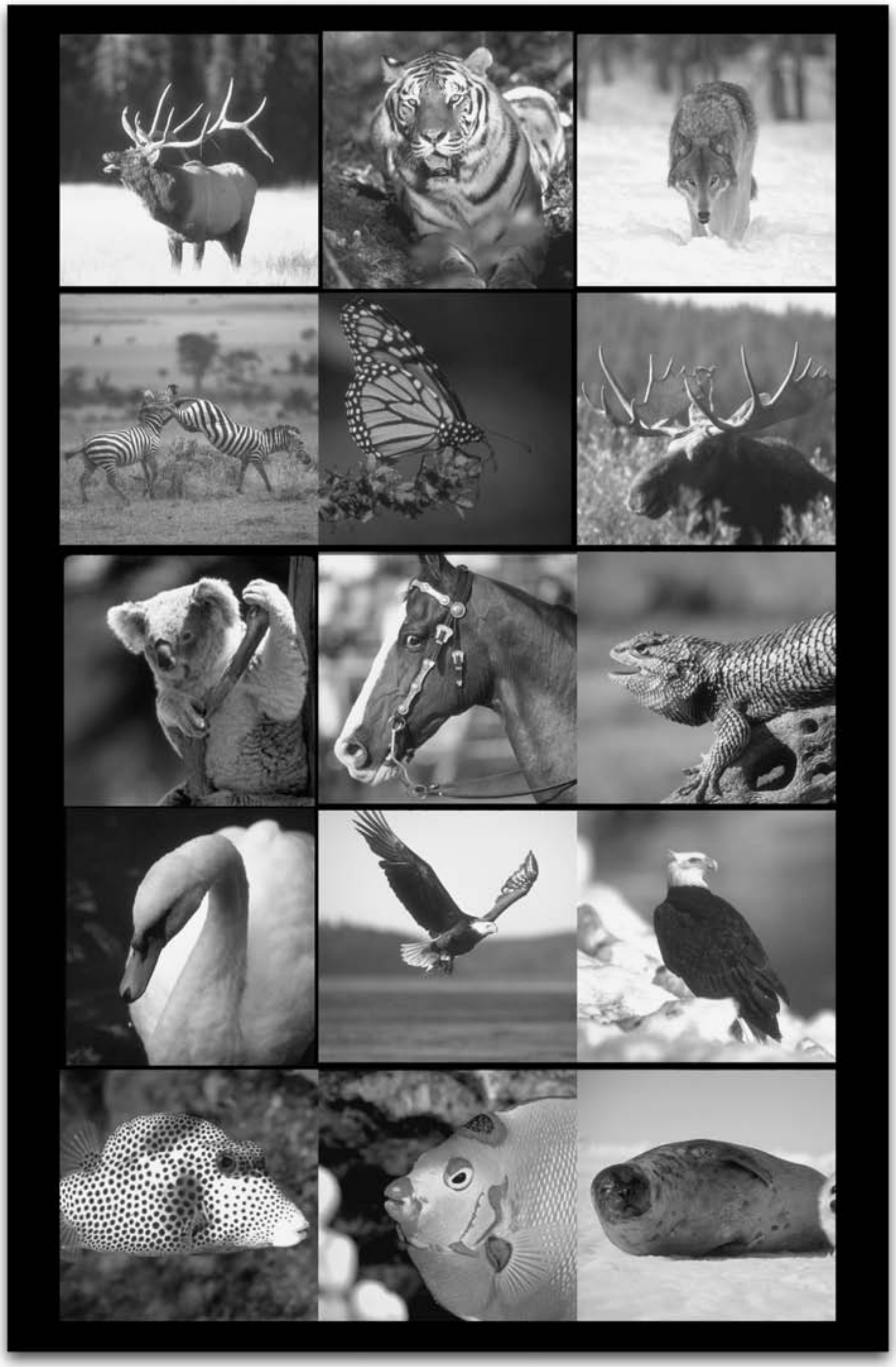

Figure 9. Representative examples (15 of 200) of animal pictures "confidently" classified as animals by a simple algorithm only taking the Fourier spectrum into account—all show animals well pre-segmented with blurred background. See text for details. (All 200 images are available as Supplementary Information.)

480 pixel square region from each image—centered on the animal for the animal images—and then down-sampled all images to 256 × 256 pixels. To cut away the animals, we simply created a circular mask with 204 pixels diameter, always centered on the image. Thus the inner mask had an area of 32,904 pixels or 50.2% of the image area, the outer region consisted of 32,632 pixels or 49.8% of the image area. (A perfect 50–50 distribution of pixels would have required a slightly non-circular mask because within the given pixel grid no perfect circle can account for precisely 50.0% of the image surface.) To reduce the possible interference that a sharp border between masked image and background may exert we smoothed the sharp edge into a Gaussian transition with a total width of 32 pixels

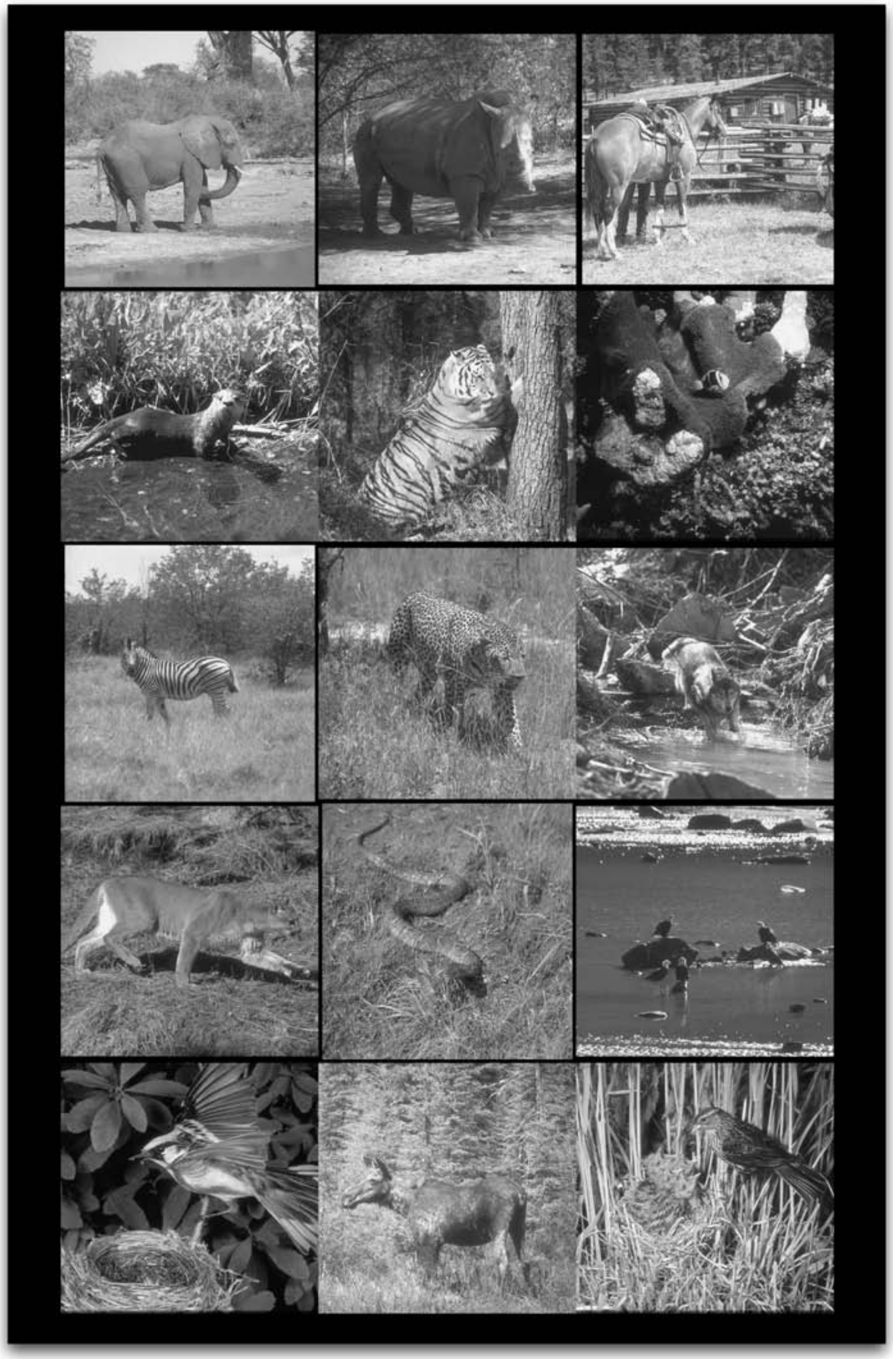

Figure 10. Representative examples (15 of 200) of animal pictures "confidently" misclassified as non-animals by the same algorithm—note the lack of pre-segmentation via blurred backgrounds; see text for details. (All 200 images are available as Supplementary Information.)

between minimum and maximum. The resulting mask (and its inverted form) was then used to cut images into "inner region only" and "outer region only", blending the masked part into neutral gray. As a result of this procedure, most (but not quite all) of the animals were invisible in the "outer region" images. (On some images a few pixels of the animal(s) extended into the outer region and we chose not to try and eliminate them, c.f. the animal examples in Figures 9 and 10. This very likely accounts for a performance slightly better than chance (53.7%, see Table 3 and Figure 17) in the "outer region" condition, see Results section of Experiment 3.) Figure 13 shows examples of two animal and two non-animal images in our three conditions, "Entire Image" in the top row, "Inner Region" in the middle row and "Outer Region" in the bottom row.

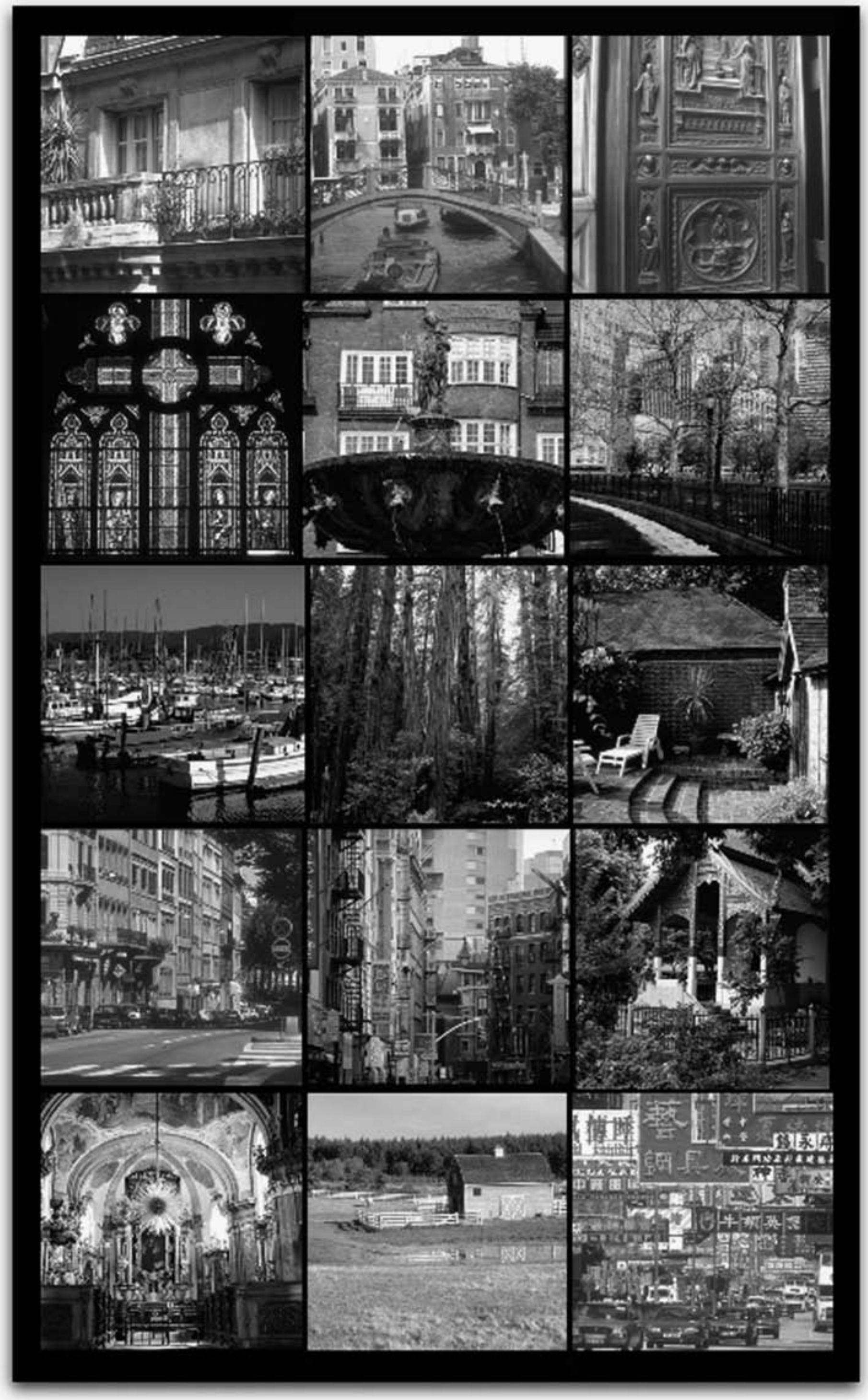

Figure 11. Representative examples (15 of 200) of distractor images very well classified as distractors by a simple algorithm only taking the Fourier spectrum into account. Note how all have either got extended depth-of-field or are in a single image-plane and thus contain spatial detail in all parts of the image. (All 200 images are available as Supplementary Information.)

Look at Figure 13—bottom row, 2nd column: wolf? no wolf? Similarly for all the other images. There *could* be birds in front of the building on the left, there may be a jeep rather than a zebra in the field on the right. In real life, waiting for a short period of time may lead to birds in front of the building on the left, or against the sky in third row images, and the disappearance of the wolf and zebra in the other scenes, i.e. to the exactly same images but now with and without animals, respectively. Thus the scenes *per se* do not contain information allowing reasonable

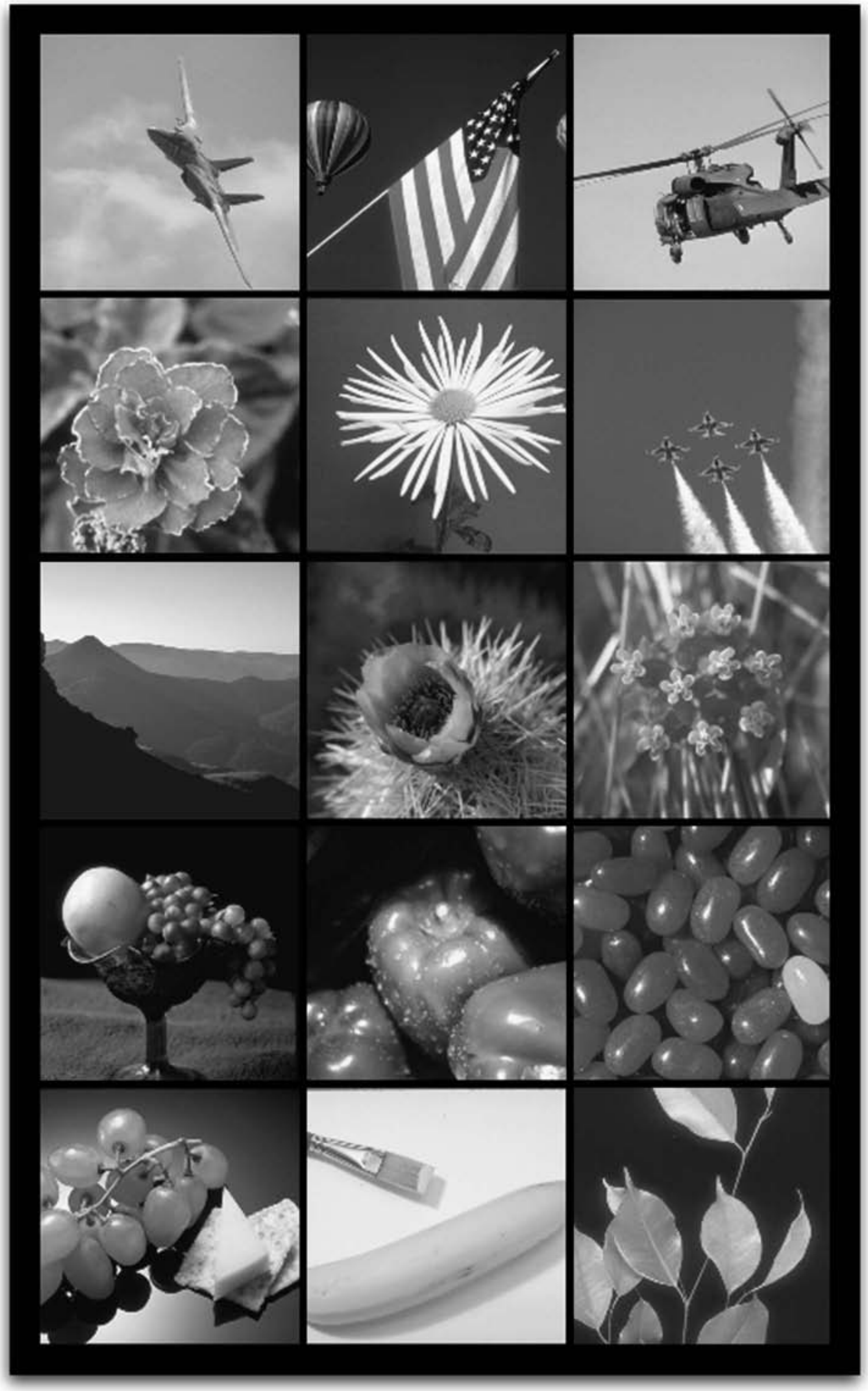

Figure 12. Representative examples (15 of 200) of distractor images misclassified as animals "with confidence" by the same algorithm. Note how these are pre-segmented, "portrait-like" with blurred or blank backgrounds. (All 200 images are available as Supplementary Information.)

guesses about the presence or absence of animals, *at least not in the real world*.

Given that the scenes themselves do not provide any information about the presence or absence of an animal—a non-informative animal—prior if one prefers Bayesian terminology—it would be surprising—and sub-optimal from statistical point of view—if human observers were better than chance deciding whether or not there was an animal in any of the "Outer Region" images. (Unless, of course, one *instructed* and *trained* them to look at the

Image area shown	Percent Correct	Response Time
Entire images	84.4%	294.6 ms
Inner region	78.4%	294.9 ms
Outer region	53.7%	324.8 ms

Table 3. Percent correct and RTs obtained in Experiment 3 for the different image conditions (entire image, inner region only and outer region only).

background and say "animal" if the background is blurry, as for the wolf and zebra images and "non-animal" if in focus, as for the image of the building.)

Algorithmic classification

As pre-processor we chose a representation of the image database in terms of spectral content, but one that also retains information about spatial distribution of spectral content. This can be achieved using a sub-band ("wavelet") transform and we chose Simoncelli's steerable pyramid (second derivative filters) (Simoncelli & Farid, 1996; Simoncelli & Freeman, 1995; Simoncelli, Freeman, Adelson, & Heeger, 1992) because of its nice properties and because elegant MATLAB code is publicly available. To optimally match the requirements of the filtering routines we up-sampled our images to 576 × 576 pixels before filtering; we chose 6 frequency- and 4 orientation bands, i.e. 24 band-pass images as well as the remaining high-pass and one low-pass images. Figure 14 shows the 26 filtered images resulting from the application of the steerable pyramid to an image of the zebra shown in the rightmost column of Figure 13.

Sub-band transforms are over-complete and the steerable pyramid is no exception, although pyramid transforms are already much less over-complete than "standard" sub-band transforms as they allow sub-sampling at each filter stage ($5 \times 576^2 + 4 \times 288^2 + 4 \times 144^2 + 4 \times 72^2 + 4 \times 36^2 + 4 \times 18^2 + 9^2 = 2\ 100\ 897$ or 6.33-times over-complete instead of being 26-times over-complete). To further reduce dimensionality, we down-sampled the resulting individual band-pass images along with the high-pass image to 18 × 18 pixels each, matching the size of the lowest frequency band, resulting in $25 \times 18^2 + 9^2 = 8{,}181$ dimensions per

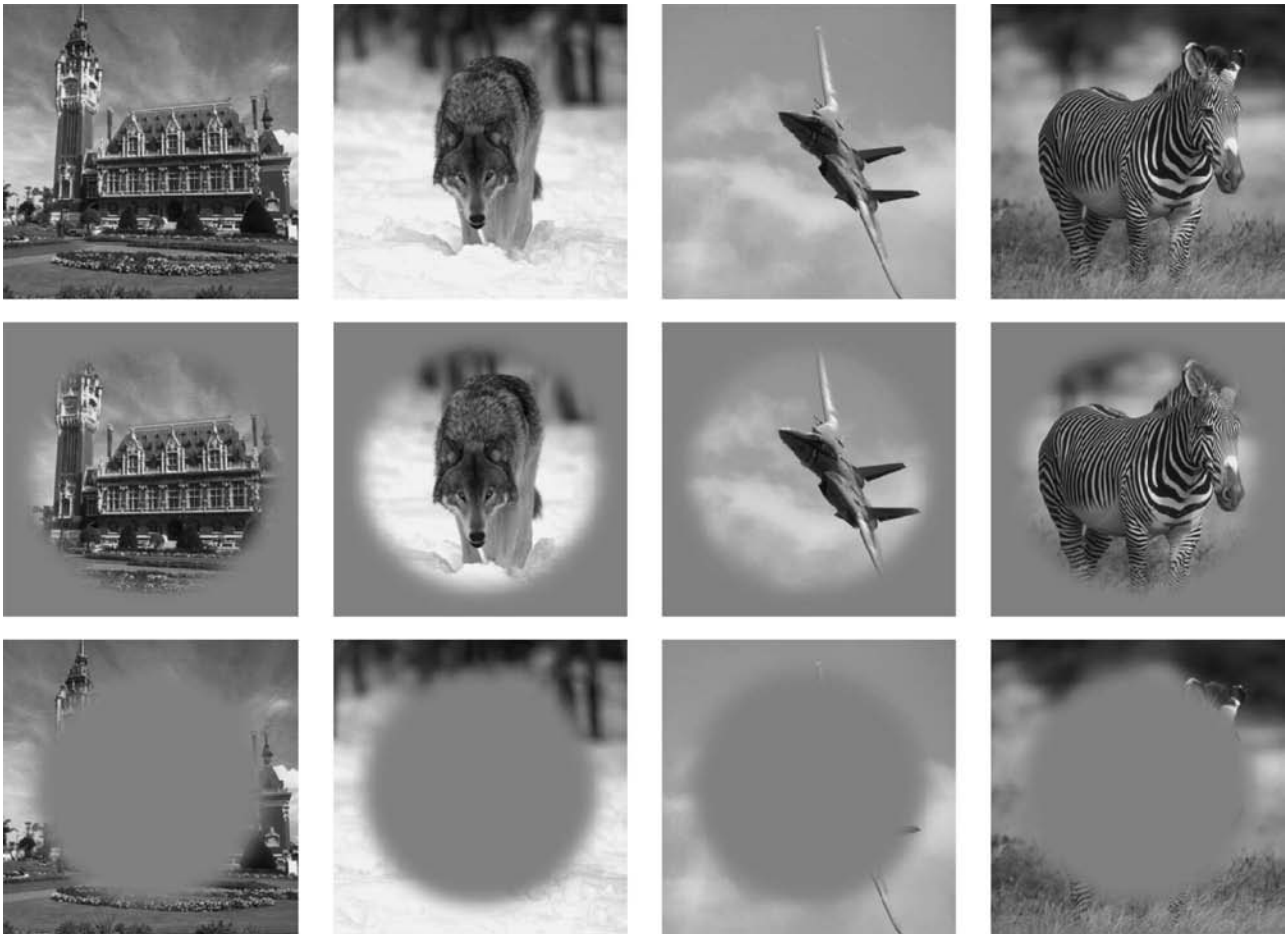

Figure 13. Example stimuli for Experiment 3. The top row images are referred to as "Entire Images", the middle row ones as "Inner Region" and the bottom row ones as "Outer Region."

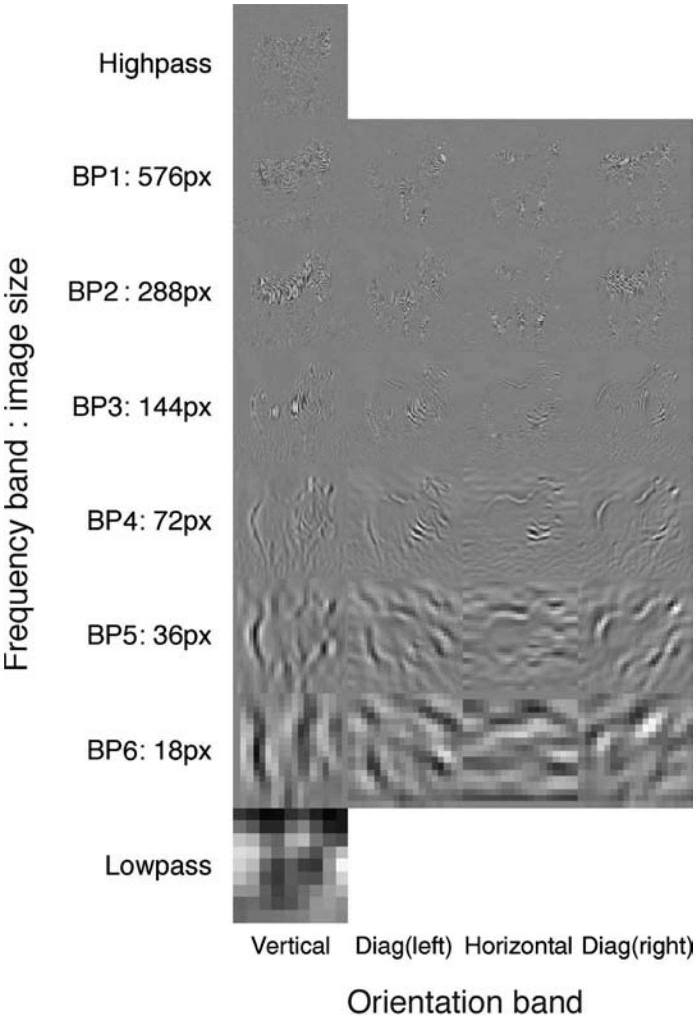

Figure 14. Shows the highpass and lowpass images as well as the 24 filtered subband images resulting from the application of the steerable pyramid to the zebra image of Figure 13, rightmost column, with 6 frequency- and 4 orientation-bands. See text for details.

image. In a final dimensionality reduction step we subsampled each of the twenty-four 18 × 18 pixel sub-band images as well as the high-pass image to a square image with 8 "pixels" only, resulting in a 25 × 8 = 200 dimensional representation per image. (We discarded the low-pass image (9 × 9 pixels) because it was not clear how to assign the 81 pixels to our non-standard 8 pixel representation; furthermore, it did not contain very much useful information for the animal versus non-animal classification, c.f. Figure 16.) Note that the 8 pixels or cells did not form a regular 2 × 4 or 4 × 2 grid; the assignment of the 18 × 18 pixels to the final 8 cells is shown in Figure 15. This non-standard assignment was necessary to allow a circular center of 4 pixels to be compared to the 4-pixel surround for the algorithmic classifier (c.f. Figure 13).

Then we used linear discriminant analysis again to repeat the classification of the 11,000 images into "animal" and "non-animal" images as done for Experiment 2. Using the pre-processing routine described above we were able to achieve a classification accuracy of 78.4% (200 dimensional representation), up from 77% using PCA as dimensionality reduction method. Classifying the "inner region" only (100 dimensional representation), we still achieved 74.3% correct classification. On the "outer region" our classifier still reached a classification accuracy of 73.4% into animal-no-animal categories in the *near-complete absence of animals*. This is in line with our suggestion that the relative amount high-spatial frequencies has nothing to do with animals versus no-animals but everything with an open versus closed aperture while taking the photographs.

To further explore this issue we analyzed the distribution of information relevant for the algorithmic classification in space and spatial frequency. Again we classified all the images from our database, but this time using a single dimension only during each run: our downscaled Simoncelli pyramid had 8,181 dimensions, and we re-classified all images 8,181-times, using a single "pixel" only: from the top left "pixel" in the Highpass image through 8,179 "pixels" to the bottom right one in the Lowpass

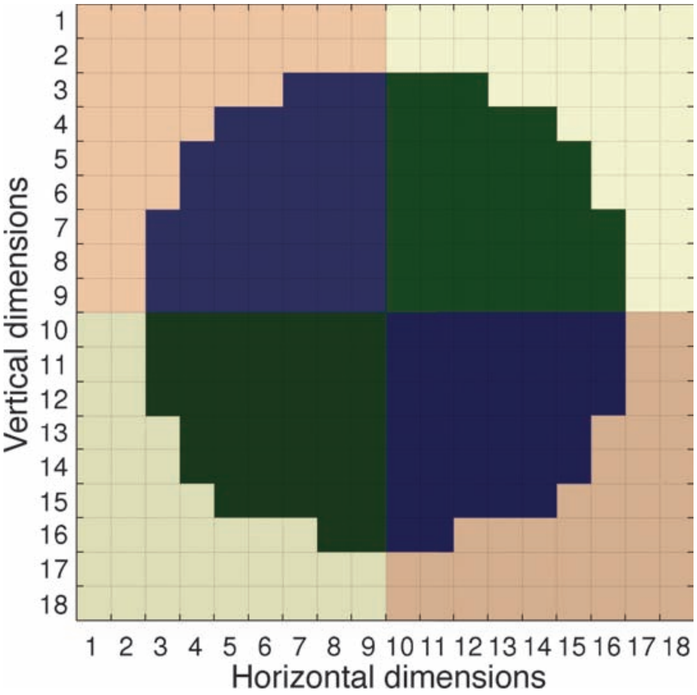

Figure 15. Final pixel assignment of our dimensionality reduction routine. The 18 × 18 pixel sub-band images as well as the high-pass images were simply averaged to the 8 final pixels as shown above (color coded). Thus in our final representation there were 4 "pixels" for the Outer Region (light, de-saturated yellow and orange) and 4 "pixels" for the Inner Region (dark, saturated blue and green). See also Figure 13 and text for details.

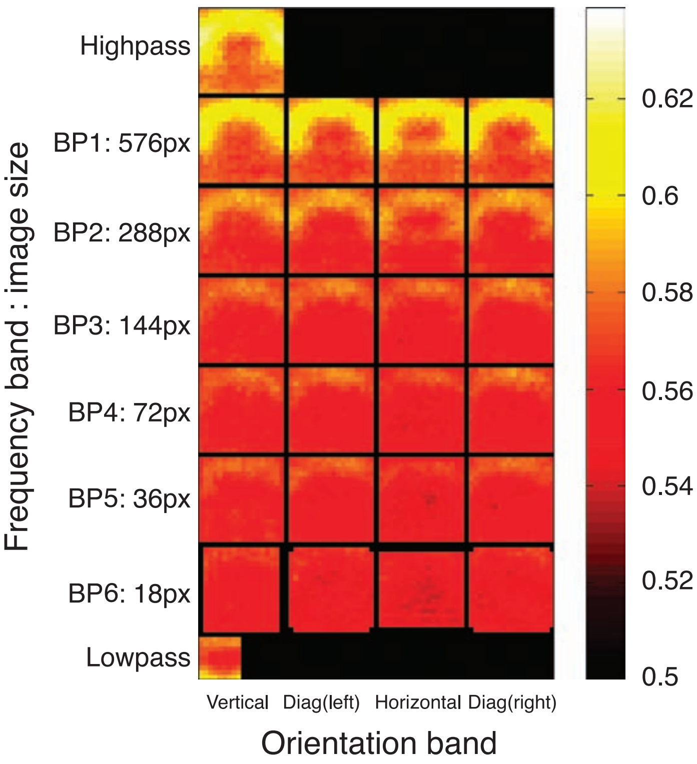

Figure 16. "Information distribution map" of the 8181 "pixel" pyramid representation of the images. Yellow pixels mark the combinations of spatial position, spatial frequency and orientation most useful for classification of Corel database images into animal versus no-animal. See text for details.

image (c.f. Figure 14). This resulted in a classification accuracy per dimension, i.e. per location in space and spatial frequency and orientation. The resulting "information distribution map" is shown in Figure 16, where we color coded the entries according to their importance for the animal-no-animal classification performance: black indicates uninformative pixels, red, pixels of intermediate informativeness, yellow, pixels most informative (up to 63% correct based on a single pixel, i.e. a very small region of the original image within a narrow range of spatial frequency and orientation).

Visual inspection of Figure 16 shows that the information relevant for the classification is located mostly in the highest frequency bands Highpass and BP1: 576px—this in itself is not surprising and simply a confirms the original high-frequency content suggestion using a pyramid rather than the Fourier transform. However, the information is by and large confined to a upper, semicircular region of the image: not where an animal would be located, but in the region surrounding the central object, where in an animal image the blurring effect induced by a typical photographic pre-segmentation is expected to show its maximal effect. This shows that the information most important to the algorithmic classifier is not located *within* the image area covered by the animal, but *around* it, supporting our hypothesis that the Corel image library is affected by an artifact, introduced by the photographers at the time of image capture.

Experimental setup

We employed the 2AFC paradigm exactly as in the supplementary experiment to Experiment 2: Images were presented on a carefully luminance calibrated Iiyama VisonMaster 513 (MA203DT) 21″ CRT screen at a frame rate of 100 Hz non-interlaced with a spatial resolution of 1280 × 960 pixels. Stimuli were viewed from a viewing distance of 45 cm, and a chin rest was used to stabilize head position. The temporal sequence after fixation was 200 msec gap time, 30 msec stimulus presentation time, 1000 msec maximum response time. Saccades were used as response indicators: observers had to perform a saccade to the image containing the animal, i.e. to the left or to the right. All eye-movement measurements were performed using SR Research's EyeLink II eyetracking system. Our 800 selected images resulted in 400 image pairs (2AFC trials). Each subject was shown every image pair in only one of the three conditions ("Entire image", "Inner region", "Outer region"), with the conditions rotated across subjects, ensuring that every image pair was shown in all three conditions, but only once per subject. Of these 400 trials per subject, the first six trials were discarded as training trials. In total we thus had 12 × 394 or 4,728 trials. All observers were students of the Justus-Liebig-University Giessen and were paid for their participation. Observers were between 19 and 31 years of age and had normal or corrected to normal vision. All observers were naïve to the purpose of the experiment. For data analysis we used the same three criteria for eye-movement traces as described above in order to ensure data integrity. First, *goodness of fixation*: We had to eliminate 150 trials of our 4,728, leaving 4,578 for further analysis. Second, *goodness of saccade direction and destination*: 9 trials needed to be eliminated to fulfill the requirements, leaving 4,569 trials for further analysis. Third, *goodness of response time*: 278 trials were removed exclusively because they were too fast, leaving a total of 4,291 valid trials for the statistical analyses presented below.

Results

Classification performance on the subset of entire images averaged 84.4%, with a mean response time of 294.6 ms. When showing only the center region of the images, classification performance averaged 78.4%, a rather modest decline. The mean response time remained unchanged (294.9 ms). As expected, classification accuracy drops to nearly chance performance (53.7%) when showing only the outer regions of the images, with the mean response time increasing to 324.8 ms. A statistical analysis on hit ratio data shows a highly significant effect overall (repeated measures ANOVA, $F(2,22) = 183.713$, $p < 0.001$). Individual, Bonferroni corrected, t-tests on the hit ratios between the three conditions show a significant difference between each of them.

For the mean response times the tiny difference (0.3 ms) between the Entire Image and Inner Region was, unsurprisingly, not significant. The difference for both of them to the Outer Region (mean $\Delta > 30$ ms) is significant (Bonferroni corrected multiple t-tests; overall effect significant; repeated measures ANOVA, $F(2,22) = 17.806$, $p = 0.001$).

Figure 17 summarizes the results of Experiment 3. Algorithmic classification based on the high-spatial frequency content differences is, by and large, unaffected by any of the image manipulations: classification accuracy is around 75% correct for the Inner and Outer Region image, i.e. with or without animals included. There is enough “background” even in the Inner Region images to allow algorithmic classification; remember, we only selected a central circular region for all images and animals are rarely perfectly circular and exactly centered! (See the wolf and zebra images in the middle row of Figure 13 for representative examples and Figures 10 and 11 for further examples.) Human observers, on the other hand, are nearly at chance for the Outer Region images—as we argued they should be given that the scenes *per se* do not provide sufficient context (prior information) allowing an animal/no-animal classification. Thus successful algorithmic classification using relative high spatial-frequency differences in the horizontal and vertical orientations may be based on an artifact in the image capturing process.

General discussion

Unfortunately, not all elegant ideas turn out to correspond to the ways of nature: the simple relative high spatial-frequency content-proxy is not employed by human observers when rapidly classifying natural images because the spectral cue may be, we hypothesize, limited to *photographs* of animals, and not or only weakly present when images of animals in the real world are projected onto the retinas of observers. Hints that human observers do not use the specifics of the amplitude spectrum for scene classification can already be found in previous studies: Thorpe, Gegenfurtner, Fabre-Thorpe, and Bülthoff (2001) studied rapid animal detection in peripheral vision and found that accuracy dropped considerably (in an almost linear fashion): from 93.3% in the fovea, to 60.5% at 70.5 degrees eccentricity. However, this is still clearly above chance for, effectively, extremely low-pass filtered and coarsely sampled images (Geisler & Perry, 1998; Geisler, Perry, & Najemnik, 2006). For images around 60 degrees eccentricity performance was better than 70%

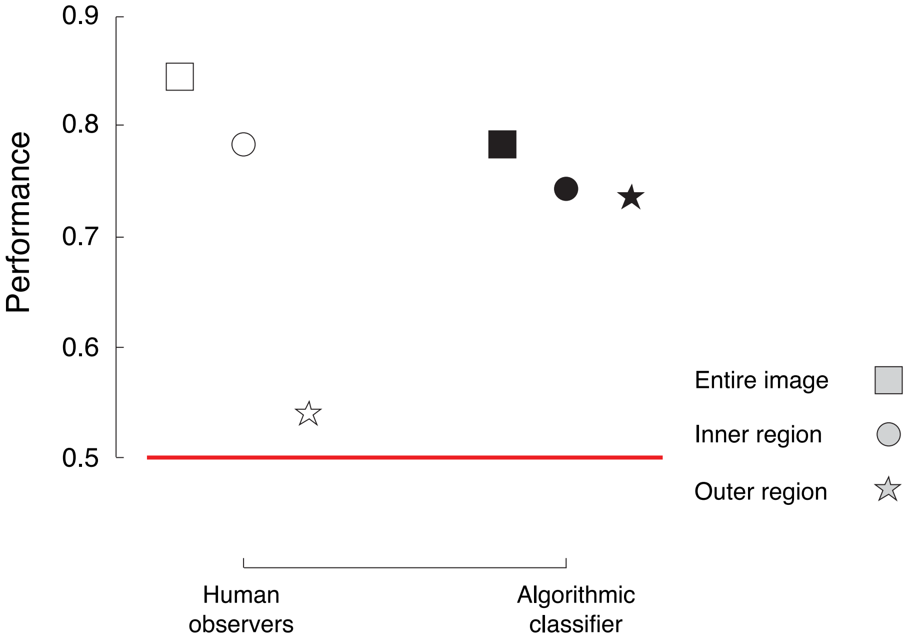

Figure 17. Results of Experiment 3; shown is the average performance of human observers on the left (open symbols; N = 12 observers; 4,291 trials, see text for details) and algorithmic classification based on the relative amount of high spatial frequencies on the right (filled symbols). Different symbols refer to the different experimental conditions: squares to performance on the entire image, circles to performance on the circular inner region only, stars to the performance on the outer region only.

correct—very unlikely due to the relative amount of high spatial-frequencies. Wichmann et al. (2006) explored the influence of phase randomization on animal detection in natural scenes. Classification performance was a monotonically decreasing function of phase randomization, leading to (very nearly) chance performance for fully phase-randomized images (e.g. their Figure 5, p. 1525). Phase randomization does not change the power spectrum, however, and if that—by itself as a proxy—allowed rapid scene categorization, observers in the Wichmann et al. study should have performed significantly better. Recently Loschky and Larsen (2008) specifically explored "animal" detection in fully phase scrambled images and they, too, concluded that differences in the Fourier amplitude spectra by themselves are insufficient to explain rapid animal detection. The results reported in this article go further, however, as we show that the relative amounts of high spatial-frequency energy are not causal even if the images themselves are clearly visible and not only visual noise as in Loschky and Larsen (2008) and Wichmann et al. (2006).

How natural are "natural" images?

The Corel and similar databases begin to be heavily used in the vision community to provide stimuli for experiments with so-called "natural" images (c.f. Rust & Movshon, 2005, versus Felsen & Dan, 2005). We have shown that at least a noticeable subset of these images are likely not entirely "natural": the amount of background blur—most likely through the photographer's choice of aperture and focal length—appears to covary with image content, creating spurious statistical regularities. This should not be surprising as most images in image databases were taken using professional cameras and tripods and, most importantly, by photographers trained to *create* professional photographs. Photographs are not simply a little window onto reality. One of the most well known landscape photographer of all time, Ansel Adams, summarizes this fact aptly in the following quote taken from his book *The Camera*:

> *The term "image management" refers to those controls we employ to alter the image formed by the lens and projected on the film. …. To do so, we must understand the differences between the image seen by the human eye and the one seen by the camera. …. Whether we realize it or not, we observe the world from many points of view, not just one … through continuous movements of the eyes, head, and body. The brain synthesizes this continuous exploration into a unified experience. The novice photographer usually learns about the differences between camera and human vision through a series of disappointments …. Examining a developed photograph … 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)

That there may be problems with commercial image databases was recently suspected by Pinto, Cox, and DiCarlo (2008)—independently of us and for a different image database, the Caltech101 image set. They compared state-of-the-art computer vision algorithms for object detection to a very simple feedforward V1-like model on a "natural image" object recognition task and found the V1-like "null" model to be superior on the Caltech101 images. However, on a seemingly simpler set of isolated, segmented stimuli the V1-like model fails completely, suggesting that it picks up artifacts of the "natural" Caltech101 image database when performing well. Pinto et al. conclude:

> *Taken together, these results demonstrate that tests based on uncontrolled natural images can be seriously misleading, potentially guiding progress in the wrong direction. Instead, we reexamine what it means for images to be natural and argue for a renewed focus on the core problem of object recognition-real-world image variation* (Pinto et al., 2008, p. 0151).

Predator and prey-camouflage, photographic pre-segmentation, and ultra-rapid animal detection

Anecdotally, animal detection in the real-world is sometimes ultra-slow rather than ultra-rapid, at least for the authors and their families. Certainly in most non-urban outdoor settings one *knows* that there must be literally thousands of smallish animals like birds or squirrels around—let alone insects and spiders. Unless an animal *moves*, however, they frequently go unnoticed. (This is even true for many animals in the zoo in spite of the explicit knowledge about the animal and the typically confined space.) The vast majority of animals serve as breakfast, lunch or dinner for other animals and thus camouflage is the norm, not the exception in the animal kingdom. Predators typically do not rapidly detect their prey but either have to be very patient and wait for the prey to move (motion segmentation), or they move themselves for prey-background segmentation through motion-parallax.

The wild-life photographer Art Wolfe published a book entitled *Die Kunst der Tarnung* [The Art of Camouflage], specifically trying not to help pre-segment the animals using the typical tricks of the photographer—Figure 18 shows eight examples from the book (Wolfe, 2005), and in Figure 19 we highlight the highly camouflaged animals

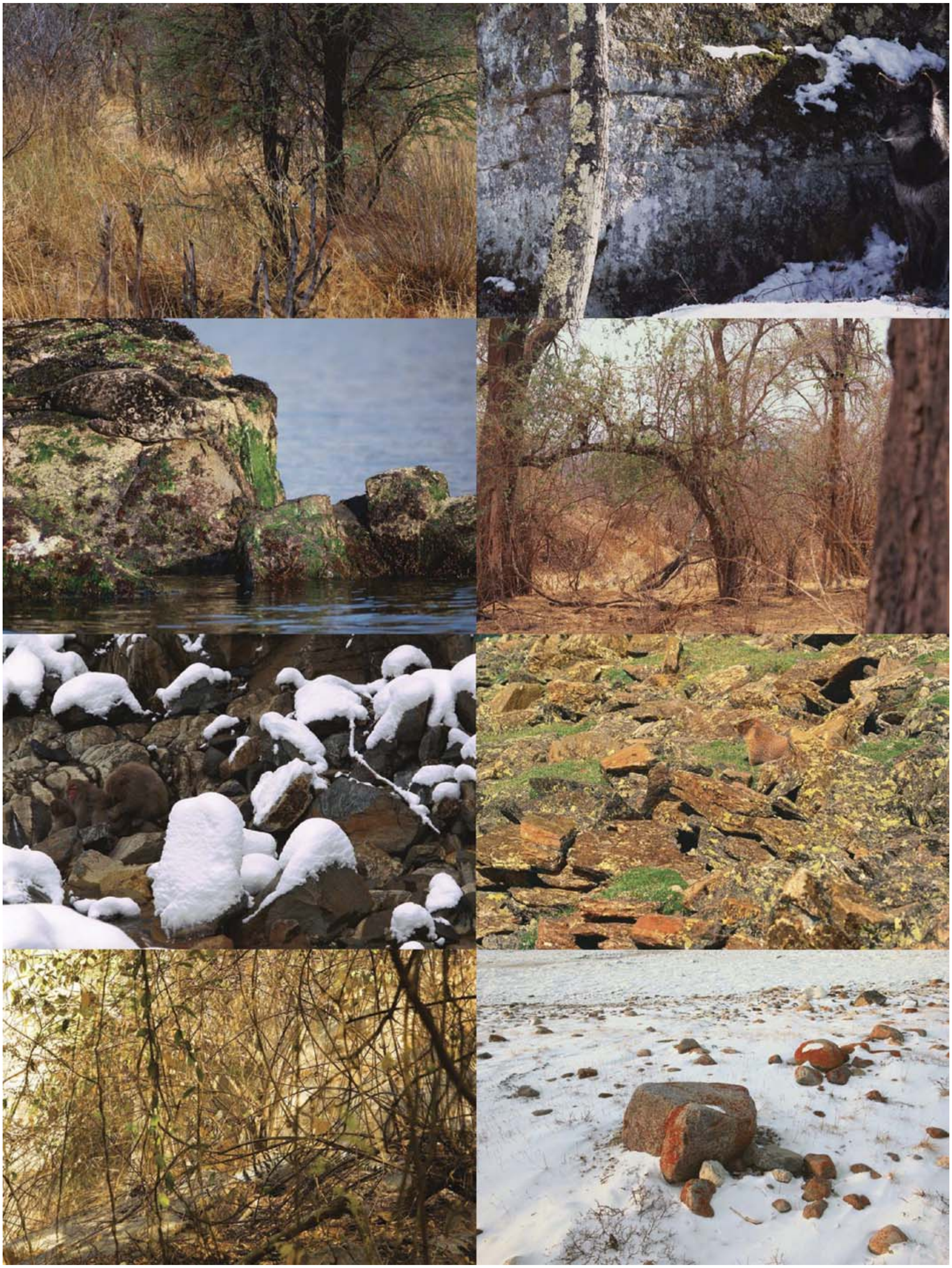

Figure 18. Ultra-slow animal detection in "natural" natural scenes? Top left: Antelope (Impala); top right: Wolf; second row left: Seal; second row right: giraffe; third row left: Macaque (Red face macaque); third row right: Yellow-bellied marmot; bottom row left: Bengal tiger; bottom row right: Polar bear.

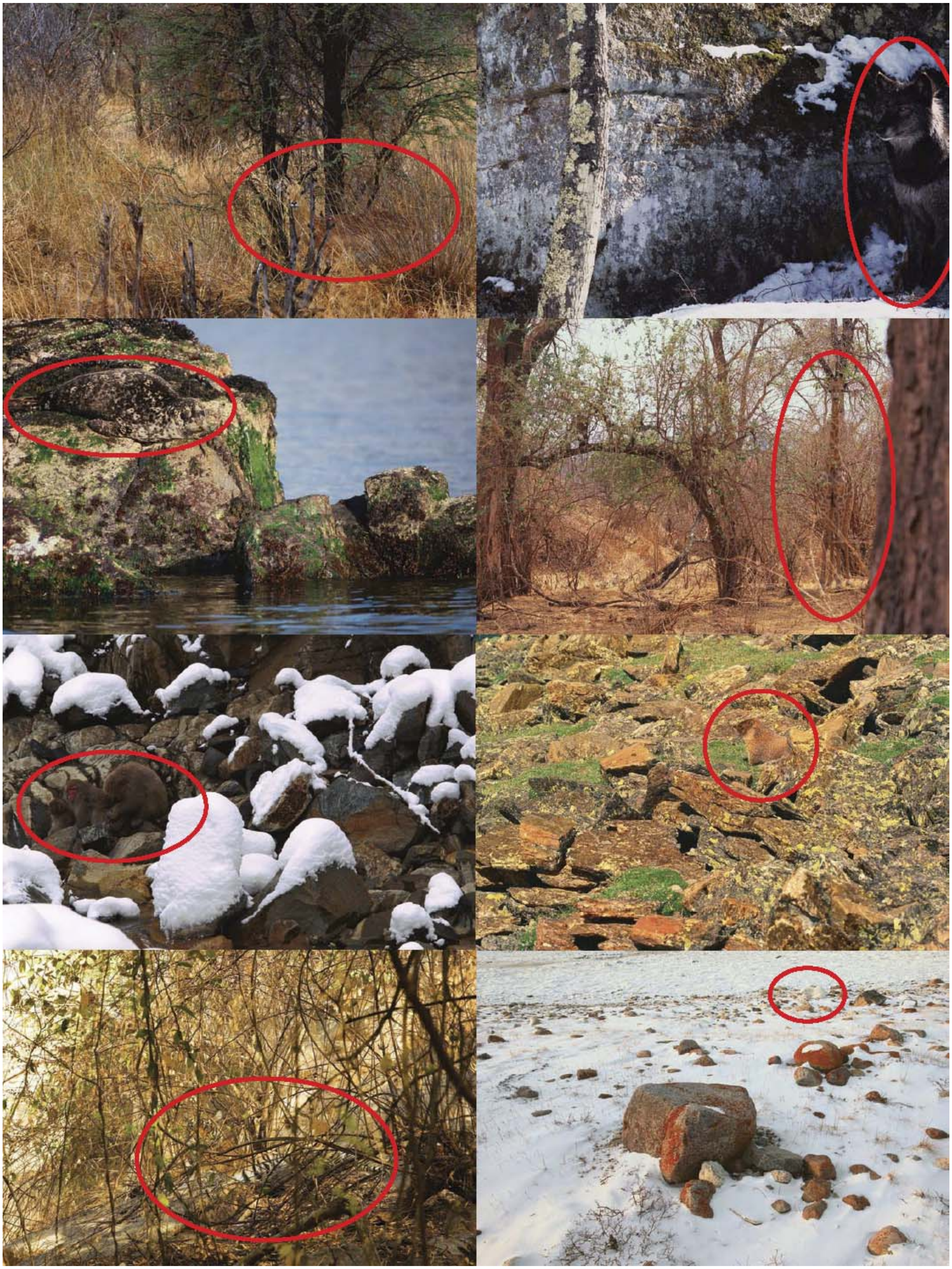

Figure 19. Same as Figure 18 but with animals marked to speed-up ultra-slow detection.

for sake of clarity. Clearly, the images presented in Figures 18 and 19 are extreme examples, and we specifically selected them precisely because animal detection in them is extremely difficult. Data from our own Experiment 2 indicates that even for the difficult-difficult pairings of the Corel database our observers performed well above chance (76.3%), and their detection was very rapid indeed (mean RT of 288 msec). However, observers performed significantly worse than for the easy-easy pairings (95.3% correct; *SEM* < 2% for all conditions; see Table 2). Thus the photographers' image manipulations via aperture and focus point contribute to, but are not the only determinants of rapid animal detection. This is in line with results from Rousselet et al. (2003): in their Experiment 2 observers had to discriminate close-up animal faces versus human faces, and they were able to do so very rapidly, even though both animal as well as human faces were photographed in a similar, large-aperture "portrait-style" with blurry backgrounds. New, Cosmides, and Tooby (2007), finally, reported a number of change-detection experiments where observers had to monitor natural scenes for changes. Either animals or non-animate objects could change in the scenes, but human observers detected the changes in animals more easily than in non-animate objects even when they were, according to New et al., equated for low-level features and intrinsic saliency. Thus New et al.'s results may be taken to support the notion that human observers have developed special hardware to detect animals in natural scenes.

Fletcher-Watson, Findlay, Leekam, and Benson (2008) recently conducted an interesting study where they were interested in rapid person detection. To control for the overall difficulty and scene content, they took photographs of the very same scene twice, once with and once without a person present. This is, we believe, an important first step towards more controlled "natural" image sets.

Summary

1. In a series of psychophysical experiments we have shown that rapid animal detection is independent of the relative amount of high spatial-frequency content differences in the horizontal and vertical orientations between animal and no-animal images shown by images of the Corel database.
2. We conjecture that the spectral differences between the image categories may result at least in part from the photographic process and aesthetic conventions, and many "natural" images may thus be less natural than commonly presumed: The statistics of (professional) photographs are likely different from the statistics of random samples from the natural environment. Because the differences are systematic, they do not disappear by increasing the sample size.
3. We thus point to a potentially important confound in popular image databases perhaps not yet fully appreciated (but c.f. Pinto et al., 2008). This natural image database problem may have implications not only for rapid animal detection research but for all approaches trying to explain scene recognition ("scene gist") on the basis of simple spectral properties (e.g. Bar, 2004; Oliva & Torralba, 2007; c.f. Loschky et al., 2007) or comparatively simple ("null") models of the visual system achieving "good" object recognition performance using Corel database (or similar) images (Serre, Oliva, & Poggio, 2007).
4. Real-world visual object recognition remains a hard problem.

Acknowledgments

This research was supported by the Deutsche Forschungsgemeinschaft Grants Wi 2103/1 and Ge 879/6 as well as the Max Planck Society and the Bernstein Computational Neuroscience Program of the German Federal Ministry of Education and Research. Parts of this work were previously presented at VSS 2005 and 2006 (Drewes, Wichmann, & Gegenfurtner, 2005, 2006; Wichmann, Rosas, & Gegenfurtner, 2005). We would like to thank Hannah Dold, Ingo Fründ and Vinzenz Schönfelder for helpful comments on previous drafts of the manuscript and Alan Allport, Matthias Bethge as well as Frank Jäkel for fruitful discussion. This article benefited considerably from conscientious peer-review; we would like to thank our two anonymous reviewers to help us improve our work, particularly reviewer 1 who brought a number of important studies to our attention—all errors and speculations remaining are, of course, our own.

Commercial relationships: none.

Corresponding author: Felix A. Wichmann.
Email: felix.wichmann@tu-berlin.de.
Address: Modelling of Cognitive Processes, Berlin Institute of Technology & Bernstein Center for Computational Neuroscience Berlin, Berlin, Germany.

Footnotes

[1] Segmentation proper is a computational processes whereby one representation is separated into parts. Photographic means, such as opening the aperture, cannot, of course, segment images but they can very much ease the process of segmentation. This is what we mean by our term pre-segmentation.

[2] Indeed, one may even expect a certain amount of performance degradation from the random-like noise added

by the spectral equalization procedure. The absence of an effect in Experiment 1 for very short presentation times and the persistence of the "easy" versus "difficult" difference in Experiment 2 are evidence against the Torralba & Oliva hypothesis. Conclusions based on a slight drop of asymptotic performance with spectral-equalization, however, cannot be drawn as firmly due to the confound of the random-like noise addition always accompanying the spectral equalization.

References

Adams, A. (1980). *The camera. The new Ansel Adams photography series,* Book 1. Boston: New York Graphic Society.

Bacon-Macé, N., Kirchner, H., Fabre-Thorpe, M., & Thorpe, S. J. (2007). Effects of task requirements on rapid natural scene processing: From common sensory encoding to distinct decisional mechanisms. *Journal of Experimental Psychology: Human Perception and Performance, 33,* 1013–1026. [PubMed]

Bacon-Macé, N., Macé, M. J.-M., Fabre-Thorpe, M., & Thorpe, S. J. (2005). The time course of visual processing: Backward masking and natural scene categorisation. *Vision Research, 45,* 1459–1469. [PubMed]

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience, 5,* 617–629. [PubMed]

Corel (1996). *Corel photo stock library* [CD-ROM]. Ottawa, Ontario, Canada.

Delorme, A., Richard, G., & Fabre-Thorpe, M. (2000). Ultra-rapid categorisation of natural scenes does not rely on colour cues: A study in monkeys and humans. *Vision Research, 40,* 2187–2200. [PubMed]

Drewes, J., Wichmann, F. A., & Gegenfurtner, K. R. (2005). Classification of natural scenes using global image statistics [Abstract]. *Journal of Vision, 5*(8):602, 602a, http://journalofvision.org/5/8/602/, doi:10.1167/5.8.602.

Drewes, J., Wichmann, F. A., & Gegenfurtner, K. R. (2006). Classification of natural scenes: Critical features revisited [Abstract]. *Journal of Vision, 6*(6):561, 561a, http://journalofvision.org/6/6/561/, doi:10.1167/6.6.561.

Evans, K. K., & Treisman, A. (2005). Perception of objects in natural scenes: Is it really attention free? *Journal of Experimental Psychology: Human Perception and Performance, 31,* 1476–1492. [PubMed]

Fabre-Thorpe, M., Richard, G., & Thorpe, S. J. (1998). Rapid categorization of natural images by rhesus monkeys. *Neuroreport, 9,* 303–308. [PubMed]

Fei-Fei, L., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences of the United States of America, 99,* 9596–9601. [PubMed] [Article]

Felsen, G., & Dan, Y. (2005). A natural approach to studying vision. *Nature Neuroscience, 8,* 1643–1646. [PubMed]

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A, Optics and Image Science, 4,* 2379–2394. [PubMed]

Fleiss, J. L. (1981). *Statistical methods for rates & proportions* (2nd ed.). New York: Wiley.

Fletcher-Watson, S., Findlay, J. M., Leekam, S. R., & Benson, V. (2008). Rapid detection of person information in a naturalistic scene. *Perception, 37,* 571–583. [PubMed]

Geisler, W. S., & Perry, J. S. (1998). A real-time foveated multi-resolution system for low-bandwidth video communication. *SPIE Proceedings, 3299,* 294–305.

Geisler, W. S., Perry, J. S., & Najemnik, J. (2006). Visual search: The role of peripheral information measured using gaze-contingent displays. *Journal of Vision, 6*(9):1, 858–873, http://journalofvision.org/6/9/1/, doi:10.1167/6.9.1. [PubMed] [Article]

Gerstner, W. (2005). How can the brain be so fast? In J. L. van Hemmen & T. J. Sejnowski (Eds.), *23 problems in systems neuroscience* (ch. 7, pp. 135–142). Oxford: OUP.

Hansen, T., & Neumann, H. (2008). A recurrent model of contour integration in primary visual cortex. *Journal of Vision, 8*(8):8, 1–25, http://journalofvision.org/8/8/8/, doi:10.1167/8.8.8. [PubMed] [Article]

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences, 7,* 498–504. [PubMed]

Hogg & Craig (1995). *Introduction to mathematical statistics* (5th ed). New Jersey: Prentice Hall.

Johnson, J. S., & Olshausen, B. A. (2003). Timecourse of neural signatures of object recognition. *Journal of Vision, 3*(7):4, 499–512, http://journalofvision.org/3/7/4/, doi:10.1167/3.7.4. [PubMed] [Article]

Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research, 46,* 1762–1776. [PubMed]

Loschky, L. C., & Larsen, A. M. (2008). Localized information is necessary for scene categorization, including the Natural/Man-made distinction. *Journal of Vision, 8*(1):4, 1–9, http://journalofvision.org/8/1/4/, doi:10.1167/8.1.4. [PubMed] [Article]

Loschky, L. C., Sethi, A., Simmons, D. J., Pydimarri, T. N., Ochs, D., & Corbeille, J. L. (2007). The importance of information localization in scene gist recognition. *Journal of Experimental Psychology: Human Perception and Performance, 33,* 1431–1450. [PubMed]

Macé, M. J. M., Thorpe, S. J., & Fabre-Thorpe, M. (2005). Rapid categorization of achromatic natural scenes: How robust at very low contrasts? *European Journal of Neuroscience, 21,* 2007–2018. [PubMed]

Motoyoshi, I., Nishida, S., Sharan, L., & Adelson, E. H. (2007). Image statistics and the perception of surface qualities. *Nature, 447,* 206–209. [PubMed]

New, J., Cosmides, L., & Tooby, J. (2007). Category-specific attention for animals reflects ancestral priorities, not expertise. *Proceedings of the National Academy of Sciences of the United States of America, 104,* 16598–16603. [PubMed] [Article]

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

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences, 11,* 520–527. [PubMed]

Pinto, N., Cox, D. D., & DiCarlo, J. J. (2008). Why is real-world visual object recognition hard? *PLoS Computational Biology, 4,* e27:0151–0156. [PubMed] [Article]

Rieger, J. W., Braun, C., Bülthoff, H. H., & Gegenfurtner, K. R. (2005). The dynamics of visual pattern masking in natural scene processing: A magnetoencephalography study. *Journal of Vision, 5*(3):10, 275–286, http://journalofvision.org/5/3/10/, doi:10.1167/5.3.10. [PubMed] [Article]

Rieger, J. W., Köchy, N., Schalk, F., Grüschow, M., & Heinze, H.-J. (2008). Speed limits: Orientation and semantic context interactions constrain natural scene discrimination dynamics. *Journal of Experimental Psychology: Human Perception and Performance, 34,* 56–76. [PubMed]

Rousselet, G. A., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nature Neuroscience, 5,* 629–630. [PubMed]

Rousselet, G. A., Macé, M. J.-M., & Fabre-Thorpe, M. (2003). Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. *Journal of Vision, 3*(6):5, 440–455, http://journalofvision.org/3/6/5/, doi:10.1167/3.6.5. [PubMed] [Article]

Rousselet, G. A., Thorpe, S. J., & Fabre-Thorpe, M. (2004). Processing of one, two or four natural scenes in humans: The limits of parallelism. *Vision Research, 44,* 877–894. [PubMed]

Rust, N. C., & Movshon, J. A. (2005). In praise of artifice. *Nature Neuroscience, 8,* 1647–1650. [PubMed]

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the United States of America, 104,* 6424–6429. [PubMed] [Article]

Simoncelli, E. P., & Farid, H. (1996). Steerable wedge filters for local orientation analysis. *IEEE Transactions on Image Processing, 5,* 1377–1382. [PubMed]

Simoncelli, E. P., & Freeman, W. T. (1995). The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *Third IEEE International Conference on Image Processing* (vol. 3, pp. 444–447). Washington, DC: IEEE Computer Society.

Simoncelli, E. P., Freeman, W. T., Adelson, E. H., & Heeger, D. J. (1992). Shiftable multiscale transforms. *IEEE Transactions on Information Theory, 38,* 587–607.

The MathWorks, Inc. (2010). 3 Apple Hill Drive, Natick, MA 01760-2098, USA.

Thorpe, S. J., Delorme, A., & VanRullen, R. (2001). Spike-based strategies for rapid processing. *Neural Networks, 14,* 715–725. [PubMed]

Thorpe, S. J., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381,* 520–522. [PubMed]

Thorpe, S. J., Gegenfurtner, K. R., Fabre-Thorpe, M., & Bülthoff, H. H. (2001). Detection of animals in natural images using far peripheral vision. *European Journal of Neuroscience, 14,* 869–876. [PubMed]

Torralba, A. B., & Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems, 14,* 391–412. [PubMed]

VanRullen, R., Reddy, L., & Li, F.-F. (2005). Binding is a local problem for natural objects and scenes. *Vision Research, 45,* 3133–3144. [PubMed]

VanRullen, R., & Thorpe, S. J. (2002). Surfing a spike wave down the ventral stream. *Vision Research, 42,* 2593–2615. [PubMed]

Vogels, R. (1999a). Categorization of complex visual images by rhesus monkeys: Part 1. Behavioural study. *European Journal of Neuroscience, 11,* 1223–1238. [PubMed]

Vogels, R. (1999b). Categorization of complex visual images by rhesus monkeys. Part 2: Single-cell study. *European Journal of Neuroscience, 11,* 1239–1255. [PubMed]

Wichmann, F. A., Braun, D. I., & Gegenfurtner, K. R. (2006). Phase noise and the classification of natural images. *Vision Research, 46,* 1520–1529. [PubMed]

Wichmann, F. A., & Henning, G. B. (1998). No role for motion blur in either motion detection or motion-based image segmentation. *Journal of the Optical Society of America A, Optics, Image Science, and Vision, 15,* 297–306. [PubMed]

Wichmann, F. A., & Hill, N. J. (2001a). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics, 63,* 1293–1313. [PubMed] [Article]

Wichmann, F. A., & Hill, N. J. (2001b). The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Perception & Psychophysics, 63,* 1314–1329. [PubMed] [Article]

Wichmann, F. A., Rosas, P., & Gegenfurtner, K. R. (2005). Rapid animal detection in natural scenes: Critical features are local [Abstract]. *Journal of Vision, 5*(8):376, 376a, http://journalofvision.org/5/8/376/, doi:10.1167/5.8.376.

Wichmann, F. A., Sharpe, L. T., & Gegenfurtner, K. R. (2002). The contributions of color to recognition memory for natural scenes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 509–520. [PubMed]

Wolfe, A. (2005). *Die Kunst der Tarnung* [The Art of Camouflage]. München: Frederking & Thaler.