VISUAL COGNITION, 2007, 15 (1), 12–19

Psychology Press
Taylor & Francis Group

Adaptation to statistical properties of visual scenes biases rapid categorization

Daniel Kaping, Tzvetomir Tzvetanov and Stefan Treue

Cognitive Neuroscience Laboratory, German Primate Centre, Goettingen, Germany

The initial categorization of complex visual scenes is a very rapid process. Here we find no differences in performance for upright and inverted images arguing for a neural mechanism that can function without involving high-level image orientation dependent identification processes. Using an adaptation paradigm we are able to demonstrate that artificial images composed to mimic the orientation distribution of either natural or man-made scenes systematically shift the judgement of human observers. This suggests a highly efficient feedforward system that makes use of "low-level" image features yet supports the rapid extraction of essential information for the categorization of complex visual scenes.

The human visual system has a remarkable ability to recognize objects, even in the midst of complex, continuously changing environments. This requires the transformation of a point-by-point retinal image into the neuronal representation of an object that is view-invariant, i.e., largely unaffected by changes in position, orientation, distance, or the presence of other visual objects in the vicinity. The recognition and categorization of scenes and objects is believed to be performed in higher level cortical areas such as the inferotemporal cortex (Logothetis & Sheinberg, 1996; Tanaka, 1996) and the medial temporal lobe (Kreiman, Koch, & Fried, 2000).

Despite its inherent difficulty, detection and categorization of objects and scenes is carried out effortlessly (Li, VanRullen, Koch, & Rerona, 2002), remarkably fast (Grill-Spector & Kanwisher, 2005; Potter, 1976), and is robust to manipulations such as image inversion (Rousselet, Mace, & Fabre-Thorpe, 2003). In a series of experiments Thorpe and colleagues (Rousselet, Fabre-Thorpe, & Thorpe, 2002; Thorpe, Fize, & Marlot, 1996; VanRullen &

Please address all correspondence to Stefan Treue, Cognitive Neuroscience Laboratory, German Primate Centre, Kellnerweg 4, 37077 Goettingen, Germany. E-mail: treue@gwdg.de

This research project has been supported by a Marie Curie Early Stage Research Training Fellowship of the European Community's Sixth Framework Programme under the contract number MEST-CT-2004-007825.

http://www.psypress.com/viscog DOI: 10.1080/13506280600856660

Thorpe, 2001) asked human subjects to decide whether an unmasked picture of a scene presented for only 20 ms contained an animal or not. Measuring event related potentials the authors were able to document different frontal activation between the two picture types only 150 ms after stimulus onset, suggesting that this type of categorization is relying on a feedforward mechanism, rather than on a high-level feature detection system located high up in the visual processing hierarchy (Rousselet et al., 2003).

Such findings point to a system that can rely on low-level image analysis for accurate object detection and scene categorization. Several factors can contribute to such a system: It has been pointed out that the general layout of scenes supports scene recognition after only a short glance (Friedman, 1979). A correct category detection permits an overall scene evaluation along more general, superordinate levels allowing the extraction of categorical properties of the depicted scene independent of detailed object recognition (Biederman, 1981; Oliva & Torralba, 2001).

Additionally, simple hierarchical processing can build upon easily extractable statistical image information (Oliva & Schyns, 1997), such as the spatial frequency composition of an image extracted through image decomposition via Fourier transformation, and the use of the orientation-selective neurons in early visual cortex. This would provide a plausible mechanism for the rapid categorization process.

For such an approach to work, scenes that are to be distinguished should differ in their respective Fourier spectra and these differences need to be large enough to enable reliable scene categorization. Indeed, Torralba and Oliva (2003) showed that the power spectrum of natural environments differ from man-made environments (Figure 1), particularly because of the

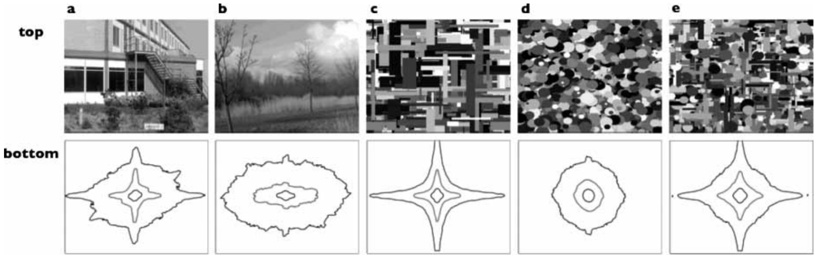

Figure 1. Examples of the images (top row) used in this study with their corresponding power spectrum (bottom row, see also Torralba & Oliva, 2003). The contour plots represent 70% (outer line), 80% (middle line), and 90% (inner line) of the spectrum log amplitude and show that man-made scenes contain more energy along the cardinal axis compared to the natural scenes. Images of (a) man-made (b) and natural scenes. Artificial images used for the adaptation based upon their relating power spectrum to emphasize (c) man-made (d) or natural image statistics. (e) Neutral adapter made up of circles and rectangles, combining man-made and natural power spectrum attributes.

predominance of contours oriented along the cardinal axes in man-made environments. They also point out that the statistics of orientation and scales are a good cue for scene categorization (Oliva & Torralba, 2001), and propose a simple linear model that uses the spectral principal components of these categories to allow semantic categorization between them (Torralba & Oliva, 2003).

While these studies document the presence and sufficient magnitude of statistical differences between images of natural and man-made environments, to date no psychophysical study has demonstrated that humans are able to exploit it for rapid scene categorization. Here we provide such a demonstration by documenting the presence of two aspects of human scene categorization that can be accounted for by a process that computes simple image statistics.

First, we test the effect of image inversions on performance because Fourier analysis is inversion-invariant due to the cardinal axes symmetry of the global frequency spectrum (Torralba & Oliva, 2003; see also Figure 1), i.e., upright and inverted images have identical image statistics and should therefore be equally distinguishable from other images.

Secondly, a scene categorization based on image statistics likely needs to be continuously calibrated, i.e., subjects probably categorize scenes into natural and man-made images by comparing a given scene's spectrum against an internal reference that represents an average of recent inputs. This would resemble similar processes in identity (Leopold, O'Toole, Vetter, & Blanz, 2001) or gender and race (Webster, Kaping, Mizokami, & Duhamel, 2004) categorizations based on images of faces. Such an approach is prone to the effects of adaptation, i.e., extended exposure to images stimulating those processing channels responsible for detecting extreme versions of one of the two categories should shift the subjects' categorization midpoint towards such adapters, if the adapted channels are indeed used in the categorization process.

In our experiments, subjects categorized greyscale environmental images in a two-alternative forced choice (man-made vs. natural) image rating task. We compared categorization performance for upright and inverted images of natural and man-made scenes and determined the effect of adapting with long-duration abstract stimuli that mimicked the prototypical orientation components of either man-made or natural scenes, respectively.

Our results show that performance was unaffected by image inversion and that the subjects' scene categorization was systematically affected by adaptation in line with the prediction sketched out above. Together the findings demonstrate that the human visual system exploits low-level image statistics for performing rapid scene categorization, an approach applicable for many categorization tasks and therefore probably widely employed.

METHODS

Twelve naive subjects (8 female and 4 male, ages 15–29) participated in the study. All subjects had normal or corrected-to-normal vision and gave written informed consent. Subjects sat in a dimly lit room, 57 cm from a computer monitor (85 Hz, 40 pixels/deg resolution) with their head stabilized on a chinrest. They were asked to categorize images briefly presented on a uniform grey background as man-made or natural scenes.

The test stimuli ("scene images") used were 316 grey level still images scaled to 13.3×10.9 deg (530×435 pixels) taken from the van Hateren and van der Schaaf Natural Stimuli Collection (1998). The images were selected from the collection such that about half of them were rated as man-made and half as natural by two of the authors with unlimited viewing time.

In each trial one test stimulus was presented for 12 ms between a spatial frequency adapting sequence and a mask stimulus (Figure 1). The mask (presented for 94 ms) appeared 94 ms after the test stimulus and was used to constrain the perceptual availability as a retinal afterimage. This interstimulus interval was chosen to be as short as possible and as long as necessary to allow acceptable performance.

The adapting stimuli were computer generated images of circles and/or rectangles that were composed such that they either matched the average power spectrum of all scene images (*neutral adapter*, made up of circles and rectangles), the spectrum of those scene images rated as man-made (*man-made adapter*, rectangles only), or that of the natural-scene images (*natural adapter*, circles only). A dynamic adaptation sequence of 10 adapting stimuli (117 ms each) was presented at the beginning of every trial. The adapting image sequence and the test images were separated by a 294 ms uniformly grey blank screen.

The three adapter types were used in separate experimental blocks of 316 test stimuli in a randomized sequence of 50% upright and 50% inverted images. In each block, each image was used upright with four subjects and inverted with another four subjects. Subjects were not told that inverted images were present. Each subject participated in two of the three adapting conditions, thus categorizing each image twice, once upright and once inverted. Results were analysed using standard Z-test for binomial distributions with adjusted p-values for multiple comparisons (Zar, 1999).

RESULTS

For each of the adaptation conditions each of the 316 test images was categorized four times in its upright and four times in its inverted orientation. For each image the number of "natural scene" responses was

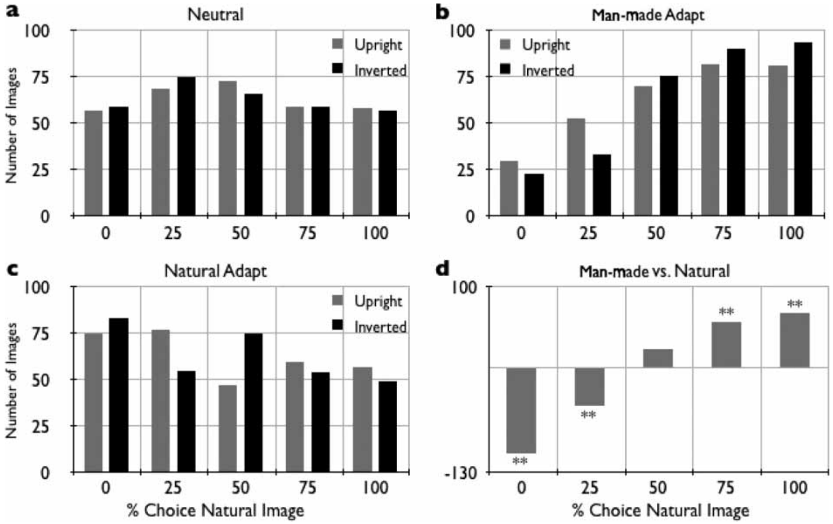

Figure 2. Histograms of number of images rated as man-made scenes (0%) by all four subjects that were shown a particular image, natural scenes (100%) or between, for (a) the neutral condition, (b) man-made-like adapters, and (c) natural statistics adapters. Categorization of upright and inverted images showed no significant difference throughout the three conditions (a–c), allowing to pool responses independent of orientation (d). Comparing man-made versus natural by subtracting the histograms show highly significant differences (d) ($*p_{\text{adjusted}} < .05$, $**p_{\text{adjusted}} < .01$).

counted across the four subjects that rated the image in the same orientation. For each possible count frequency (0, 25, 50, 75, and 100%) the number of images receiving the corresponding rating were counted (Figure 2a–c).

The light bars in Figure 2a show the resulting histogram for upright images in the neutral condition. The homogeneous distribution indicates that the subjects were able to perform the task, that the collection of images were not biased to one or the other category, and that the images varied as to their perceptual unambiguity. Comparing the response distribution against the one for the inverted images (dark bars) reveals no significant difference, indicating that the subjects could rate the inverted images just as well as the upright images.

Similarly, for the man-made and natural adapting conditions no significant differences were found for upright and inverted images. But the response distributions between these two adapting conditions were very different. Figure 2b shows that adaptation to the underlying statistics of man-made environments biased the categorization towards "natural" responses (see Figure 2b and 3b). A significant overall decrease ($Z = 4.21$, $p_{\text{adjusted}} < .01$ inverted, $Z = 3.08$, $p_{\text{adjusted}} < .05$ upright) of images collectively categorized as man-made (following adaptation to man-made image statistics) produced a reshaped response to identify significantly ($Z = 3.37$, $p_{\text{adjusted}} < .01$ inverted) more natural aspects within the test images (Figure 2b). For the natural adaptation paradigm a strong opposite trend was present (see Figure 2c and 3b) and a direct comparison between the response distribution of man-made versus natural adapting stimuli revealed highly significant effects (Figure 2d, $Z > 7.74$, $p_{\text{adjusted}} < .01$, pooling over orientation).

DISCUSSION

Our data show that the human visual system is able to categorize novel environmental scenes rapidly and unaffected by inversion, indicating a neural mechanism not relaying on high-level image orientation dependent identification processes. This interpretation is supported by our finding that adaptation with an abstract image composed to mimic the orientation

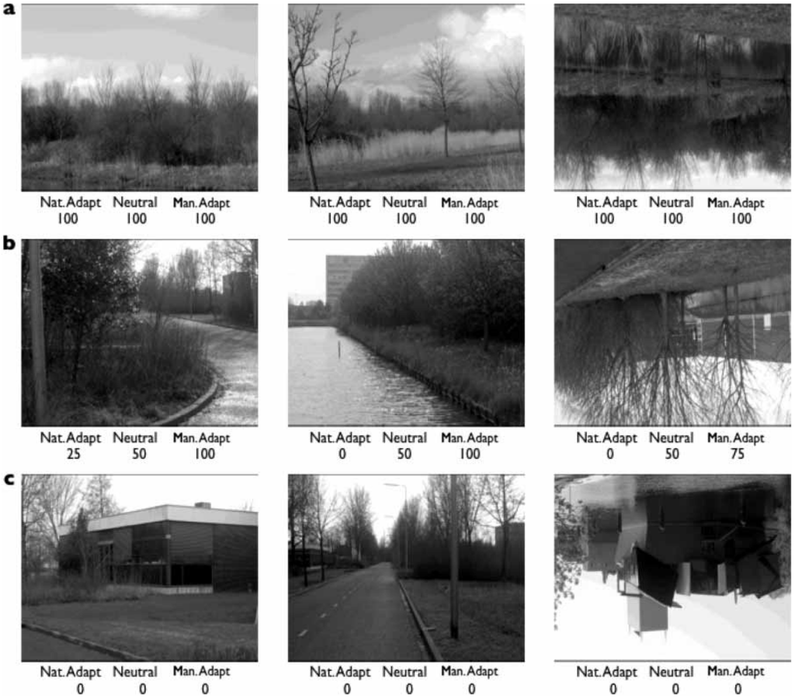
a
Nat.Adapt Neutral Man.Adapt
100 100 100
Nat.Adapt Neutral Man.Adapt
100 100 100
Nat.Adapt Neutral Man.Adapt
100 100 100
b
Nat.Adapt Neutral Man.Adapt
25 50 100
Nat.Adapt Neutral Man.Adapt
0 50 100
Nat.Adapt Neutral Man.Adapt
0 50 75
c
Nat.Adapt Neutral Man.Adapt
0 0 0
Nat.Adapt Neutral Man.Adapt
0 0 0
Nat.Adapt Neutral Man.Adapt
0 0 0

Figure 3. Illustration of natural scenes with their corresponding responses (below the image) in the three conditions. (a) Example of natural scenes unaffected by the adaptation to image statistics; (b) scenes judged to be ambiguous in the neutral condition shifted by the adapting conditions; (c) man-made scenes unaffected by the adaptation. Inverted images are shown in the last column.

content of a man-made scene biased subjects to report a given image as representing a natural scene more often than after exposure to an equally abstract adapting pattern mimicking the orientation composition of a natural scene (Figure 3d). This adaptation effect indicates that the abstract images affected specific processing channels that contribute to rapid scene categorization, documenting that the human visual system is not only highly sensitive to the statistical properties of the visual input but can also exploit patterns in those properties to perform such seemingly complex decisions as whether an image depicts a scene that is natural or man-made.

Two points need to be made when evaluating these findings: First, the rapid feedforward scene categorization process demonstrated by our findings is obviously just a first "best guess" of the visual system. It allows us to recover the "gist" of a scene (Braun, 2003). Scrutinizing the scene, if it remains visible (i.e., without masking), allows the visual system to employ its full range of object recognition systems resulting in a much more reliable categorization (Rosch, 1978) based on a fuller perceptual representation. Nevertheless, our data show a low-level scene analysis system that presumably operates on all inputs and might provide a preattentive screening for basic aspects of the visual signals entering cortex. As such the system could provide important input towards the construction of a saliency map of the visual environment (Treue, 2003).

Second, the approach employed by the visual system in extracting and interpreting the Fourier spectrum of the visual input is just one of many low-level analyses that can be performed by neuronal populations in the early visual system. Such systems could provide rapid estimates of many other categorical assessments of the visual input or even just patches of it.

In summary, our findings reveal a highly efficient system for constructing an internal representation of the visual input that relies on the feedforward extraction of "low-level" image features yet supports sophisticated perceptual judgements previously thought to require "high-level" image processing. This system appears to be particularly useful in case of high processing load, whenever fast judgements are needed and in animals that lack the sophisticated processing abilities of primate extrastriate cortex.

REFERENCES

Biederman, I. (1981). On the semantics of a glance at a scene. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 213–253). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Braun, J. (2003). Natural scenes upset the visual applecart. *Trends in Cognitive Sciences*, *7*, 7–9.

Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General*, *108*, 316–355.

Grill-Spector, K., & Kanwisher, N. (2005). Visual recognition: As soon as you know it is there, you know what it is. *Psychological Science*, *16*, 152–160.

Kreiman, G., Koch, C., & Fried, I. (2000). Category-specific visual responses of single neurons in the human medial temporal lobe. *Nature Neuroscience*, *3*, 946–953.

Leopold, D. A., O'Toole, A. J., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience*, *4*, 89–94.

Li, F. F., Van Rullen, R., Koch, C., & Rerona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences, USA*, *99*, 9596–9601.

Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience*, *19*, 577–621.

Oliva, A., & Schyns, P. G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology*, *34*, 72–107.

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.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, *2*, 509–522.

Rosch, E. (1978). Principles of categorization. In B. E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 28–49). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

Rousselet, G. A., Mace, 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*, 440–455.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, *19*, 109–139.

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

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

Treue, S. (2003). Visual attention: The where, what, how and why of saliency. *Current Opinion in Neurobiology*, *13*, 428–432.

Van Hateren, J. H., & van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society of London Series B*, *265*, 359–366.

VanRullen, R., & Thorpe, S. J. (2001). Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects. *Perception*, *30*, 655–668.

Webster, M. A., Kaping, D., Mizokami, Y., & Duhamel, P. (2004). Adaptation to natural facial categories. *Nature*, *428*, 557–561.

Zar, J. H. (1999). More on dichotomous variables. *Biostatistical analysis* (4th ed., pp. 555–558). Upper Saddle River, NJ: Prentice Hall.

Manuscript received January 2006
Manuscript accepted June 2006
First published online July 2006