Understanding rapid category detection via multiply degraded images

Chetan Nandakumar

Vision Science Graduate Program, University of California, Berkeley, Berkeley, CA, USA

Jitendra Malik

Department of Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, CA, USA

Rapid category detection, as discovered by S. Thorpe, D. Fize, and C. Marlot (1996), demonstrated that the human visual system can detect object categories in natural images in as little as 150 ms. To gain insight into this phenomenon and to determine its relevance to naturally occurring conditions, we degrade the stimulus set along various image dimensions and investigate the effects on perception. To investigate how well modern-day computer vision algorithms cope with degradations, we conduct an analog of this same experiment with state-of-the-art object recognition algorithms. We discover that rapid category detection in humans is quite robust to naturally occurring degradations and is mediated by a non-linear interaction of visual features. In contrast, modern-day object recognition algorithms are not as robust.

Keywords: rapid category detection, degraded images, object recognition, eye tracking

Citation: Nandakumar, C., & Malik, J. (2009). Understanding rapid category detection via multiply degraded images. *Journal of Vision, 9*(6):19, 1–8, http://journalofvision.org/9/6/19/, doi:10.1167/9.6.19.

Introduction

Thorpe, Fize, and Marlot's (1996) study demonstrated the robustness of the visual system at extremely short time scales. Using an ERP setup, the authors flashed images of natural scenes, where only a portion of the image set contained animals. Subjects were instructed to make a go/no-go response of whether an animal was present in the image. Analyzing the ERP signals between animal-present and animal-absent trials, a significant difference was found 150 ms after stimulus onset. Although there is controversy around this finding (Johnson & Olshausen, 2003; Van Rullen & Thorpe, 2001), it is clear that category detection can occur at very rapid time scales.

The question motivating our study is the following: Can one do rapid category detection when the image is degraded? By the term degradation, we are referring to various ways of reducing information content in the image. Over the years, selected studies have investigated the perceptual effects of degrading images along a singular image dimension. Harmon and Julesz (1973) and Bachmann (1991) demonstrate that with respect to spatial resolution, only 18 × 18 pixels per face are sufficient for robust recognition, and these findings have been extended to the domains of faces, objects, and scenes by Torralba, Fergus, and Freeman (2008) and Torralba and Sinha (2001). Along the dimension of luminance depth, Mooney faces are a classic demonstration of visual processing working in extreme cases of luminance depth degradation (Mooney, 1957). Robustness to degradation falls under the larger umbrella of invariances. Invariance is a percept's tolerance to different transformations such as scaling, lighting, or rotation. There's been extensive study of invariance in psychology and computer science. Studies such as Tarr and Pinker (1989), Rock, Di Vita, and Barbeito (1981), and Yin (1969) investigate rotation invariance in humans. In physiology, Brainard (2004) investigates color constancy and Ito, Tamura, Fujita, and Tanaka (1995) investigates size and position invariance. Many computational models have also been proposed to tackle different invariances such as those of Olshausen, Anderson, and Van Essen (1993), which proposes a model that is scale and shift invariant, and those of Lowe (1999), which proposes the SIFT descriptor, a local descriptor robust to affine transformations and lighting variation.

Why do we think that rapid category detection maybe adversely affected by image degradation? One line of reasoning stems from the SpikeNet computational model proposed by Thorpe (2002). Here the author attempts to build a computational model capable of mirroring the rapid categorization effect originally demonstrated by Thorpe et al. (1996). This model centers on rank order coding—average local contrast determines the timing of neuronal firing. This perspective suggests that degraded images, such as blurred images, would have a delayed response since average local contrast is reduced. This link between reduced contrast and a delayed signal is also echoed in the explanation of the Pulfrich Pendulum phenomenon.[1] A second line of reasoning, which suggests

doi: 10.1167/9.6.19 Received September 12, 2008; published June 29, 2009 ISSN 1534-7362

a delay for degraded stimuli, is that impoverished stimuli rely on top down influences as argued by Cavanagh (1991). This might suggest that feedback mechanisms, and a corresponding increase in processing time would be required for degraded images.

In this study, we have crafted an experiment combining rapid stimulus presentation with degraded stimuli. We use a stimulus set of degraded images where each condition is degraded along a singular image dimension or pair of image dimensions. Through the use of degraded images, we are able to remove image information in a principled manner and by measuring the corresponding effects on perception, we also gain key insights into the mechanisms driving rapid category detection.

With the findings from above, we conducted another experiment to understand how well the latest object recognition algorithms compare to human recognition abilities. It has been argued (Serre, Oliva, & Poggio, 2007) that the current state of the art in computer vision can model the rapid category effect from Thorpe et al. (1996). By running such models with degraded images as input, we further tested the robustness of this claim.

Methods

Experiment 1

Subjects

Twenty-one volunteers with normal or corrected-to-normal vision performed a 2AFC visual discrimination task. The experimental procedures were approved by the UC Berkeley ethical committee.

Experimental setup

The experimental setup is modeled after Kirchner and Thorpe's (2006) 2AFC animal/non-animal detection task. Subjects were seated in a dimly lit room with their heads mounted in a chin rest. Gray scale images were presented to subjects on a CRT monitor placed 60 cm away from the subject. Stimuli were centered laterally 10° from fixation and subtended approximately 6° of visual angle. We use an Arrington eye tracker to track subject's eye movements; the accompanying software package presented the stimuli and appropriately created log files of the subjects' responses. These log files were later analyzed using MATLAB scripts. The stimuli were selected from the COREL collection of natural images. The stimulus set was divided into an equal number of targets and distracters. The target images each contained an animal at an arbitrary pose and location, and the distracter images were natural scenes such as landscapes or forest scenes and did not contain any animals.

Protocol

In each trial, subjects were flashed a fixation cross for 2.5 s, followed by a blank gray screen for 200 ms, and then flashed a pair of images in the right and left hemi-fields for a duration of 30 ms. Each pair contained one target and one distracter image. The target location was equiprobable in the left and right hemi-fields. After stimulus presentation, two fixation crosses were presented for 1 s at ±6°. Subjects were directed to make an eye movement to the cross on the same side as the animal and were measured for performance in different conditions. Each subject was presented with the *Full* or control condition and then a subset of the variable conditions. Please see Figure 1 for a visualization of the experimental protocol.

Stimulus conditions

The first set of conditions contains images degraded to different degrees along one of the following dimensions: spatial resolution, luminance depth, inversion, and reverse contrast. To understand how these different visual dimensions combine to influence perception, the above degradations are paired to create multiply degraded images.

Each dimension of degradation was chosen with a specific intention. Spatial resolution degradation is a naturally occurring degradation with objects at a distance and objects in the periphery, and luminance depth degradation occurs in low-light conditions. The inversion and reverse contrast conditions, although not naturally occurring conditions, offer insight into visual cognition. In the context of face recognition, both transformations have severely impacted recognition performance and thereby offered insight into the underlying representations (Liu, Collin, Burton, & Chaurdhuri, 1999; Yin, 1969). Inversion probes the role of global features as global features are warped in an inverted image while local features (eg. texture) are not affected. The reverse contrast condition probes the importance of edge polarity since each edge switches its polarity under this transformation.

Each condition consists of 50 trials, and between four and seven subjects participated in each condition. Example images from all conditions are shown in Figure 2. The baseline condition, labeled *Full,* contains 8-bit gray scale images of animals and natural scenes.

Along the dimension of spatial resolution, there are three degrees of degradation. The first, labeled *Blur100,* is created by resizing the original image, sized at 512 × 768 pixels, to 66 × 100 pixels. This miniaturized image is then resized back to its original size. The *Blur50* and *Blur25* conditions are created in an analogous manner. This manner of blurring allows us to compute an upper limit of the information contained after degradation. For instance, in the *Blur100* condition, we have 66 × 100 = 6600 dimensions where the value of each dimension is specified by 8 bits.

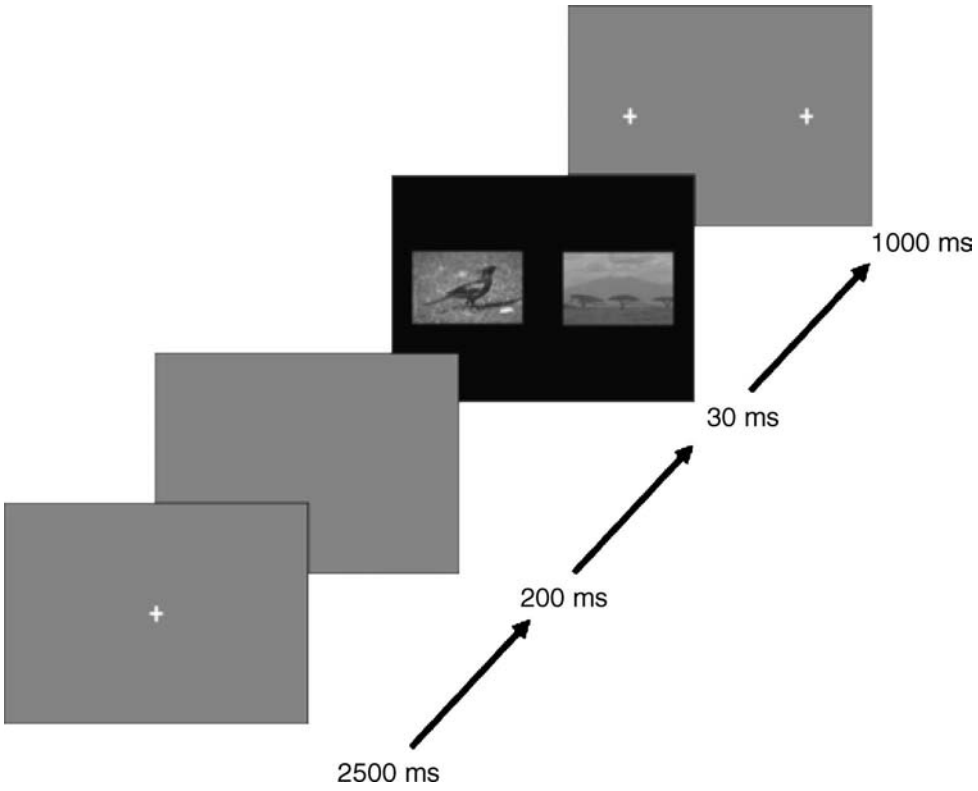

Figure 1. Visualization of the stimulus presentation protocol. On each trial, a fixation cross is displayed, followed by a blank gray screen, then the target–distracter pair, and finally a screen with two fixation crosses. The subject is instructed to saccade to the fixation cross on the same side as the target.

Along the dimension of luminance depth, there are two degrees of degradation. The first, labeled *4Tone,* is created by discretizing the original 8-bit image into a 2-bit image. In other words, the original image, containing 256 levels of gray, is now reduced to only four luminance levels (black, white, and two levels of gray). Likewise, in the *2Tone* condition, each image pixel is either black or white. This first set of conditions also includes the *Inverse* condition where each image is vertically inverted and *Reverse* where the contrast of each image is reversed.

In the second set of conditions, the above degradations are combined in pairs. In the first such condition, spatial resolution degradation is paired with luminance depth degradation: The *Blur100* condition combines with the *4Tone* condition to create the *Blur100 + 4Tone* condition, and *Blur50* combines with *4Tone* to create the *Blur50 + 4Tone* condition. In the next set of conditions, *Inversion* is paired separately with *Blur100* and *4Tone* to create the *Inversion + Blur100* and *Inversion + 4Tone*. Lastly, *Reverse* is paired in a similar manner to create the *Reverse + Blur100* and *Reverse + 4Tone* conditions.

Response recording and detection

Subjects' eye movements are tracked using an Arrington Eye Tracker. This unit consists of a single infrared camera setup with a 30-Hz sampling speed. From the eye tracking signal, the first movement to the left or right side of the screen is extracted and taken to be the binary response for that given trial.

Statistical analysis

A repeated measures ANOVA was conducted to determine significance of a given condition against the baseline condition, *Full*. This test has one independent factor and one dependent factor. The independent factor is the experimental condition, which has 2 levels—the control condition and the experimental condition. This factor served as the repeated measure in the experiment since a given subject participated in both the control and experimental conditions. The dependent factor in this analysis is the accuracy of the subject in the task. The number of subjects participating in each condition varied between 4 and 7.

Experiment 2

To understand how object recognition algorithms cope with degraded images, we test two algorithms: the spatial

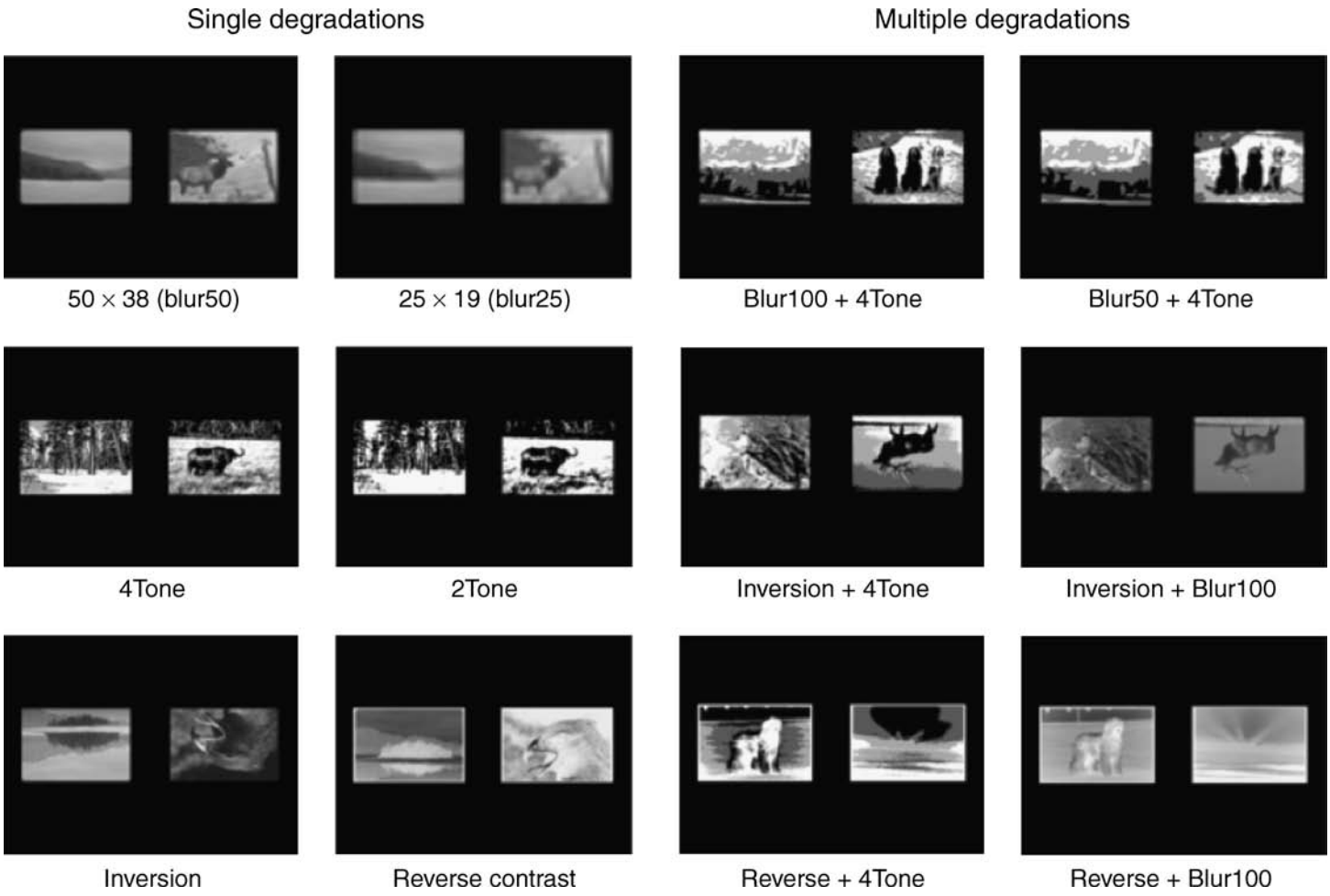

Figure 2. Examples of stimuli from select conditions. The conditions displayed in the left column contain images degraded along a singular condition, and the conditions displayed in the right column contain images degraded along pairs of conditions.

pyramid kernel (Lazebnik, Schmid, & Ponce, 2006) and the GIST descriptor (Oliva & Torralba, 2006). The spatial pyramid kernel serves as a leading approach to object recognition as it has performed well on Caltech 101, one of the key benchmarks in the computer vision community. It works by pooling orientation features across different spatial scales into a feature vector, and this feature vector is then fed through a support vector machine classifier. The classifier uses the training set to find a separating hyperplane between the positive and the negative examples in the feature space. A given test point is classified in reference to this hyperplane—the side it falls on is the class and the distance from the hyperplane is the confidence.

While the spatial pyramid kernel was designed for object recognition, the GIST descriptor was designed to capture the gestalt of an image for scene classification. The GIST descriptor was used to label an arbitrary scene as a beach scene, office scene, etc. The GIST descriptor also uses orientation energy but pools it in a different way. It tiles up the image into blocks and in each block, it sums the energy in different orientation and spatial frequency bands. This feature vector is then fed through a support

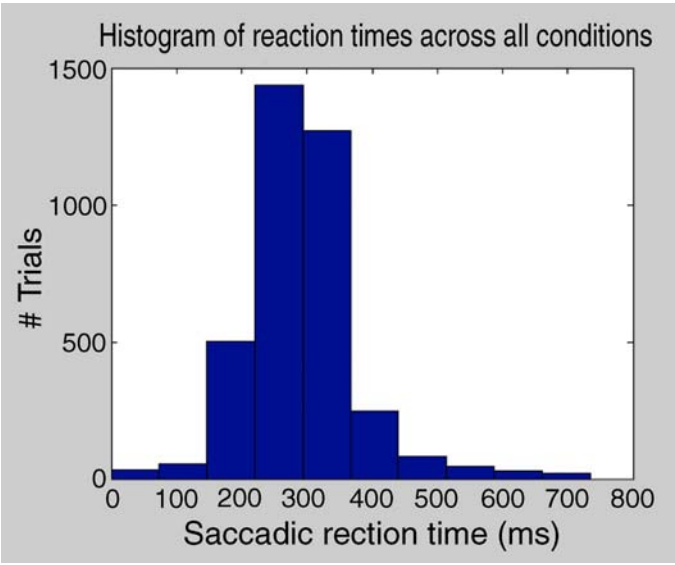

Figure 3. Histogram of saccadic reaction times. Data are pooled across all conditions and all subjects.

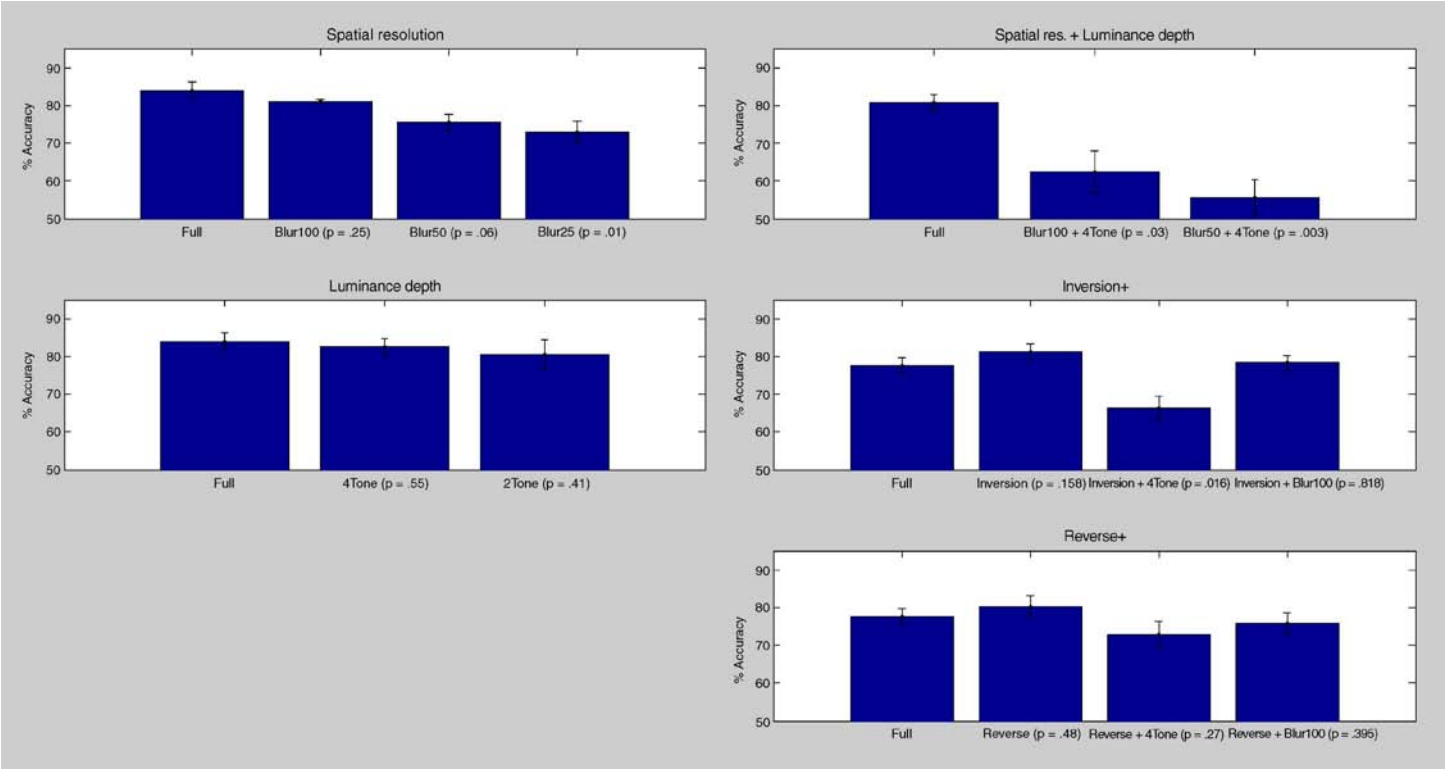

Figure 4. Mean accuracy results along with error bars displayed for each condition. Between 4 and 7 subjects were used in each condition, and a repeated measures ANOVA was used to determine significance. Significance from the baseline condition, labeled *Full,* was found in the following conditions: *Blur25, Blur100 + 4Tone, Blur50 + 4Tone,* and *Inversion + 4Tone.* The *p*-values are printed alongside in the condition labels in the plots.

vector machine classifier. In this experiment, the GIST descriptor was tailored to only use the low spatial frequencies of the image.

We trained each classifier with 200 animal and 200 non-animal images. These images were drawn from the same set used in the human experiments above and each was tagged with its respective category. We tested the performance of the classifiers under different conditions, and to create results comparable to the human results from Experiment 1, we simulated "trials" where each trial contained a randomly paired animal and non-animal image. A correct trial is one where the confidence value assigned to the animal image is greater than that of the non-animal image.

Condition	Human performance (% correct)	Spatial pyramid kernel (% correct)	GIST descriptor (% correct)
Full	84	73	84
Blur100	81	56	72
Blur50	75	59	55
Blur25	73	63	46
4Tone	82	51	75
2Tone	80	57	74
Blur100 + 4Tone	62.4	50	58
Blur50 + 4Tone	55.6	50	52
Inversion	81.1	57	65
Inversion + 4Tone	66.2	51	64
Inversion + Blur100	78.2	57	61
Reverse	80.2	60	76
Reverse + 4Tone	72.8	52	75
Reverse + Blur100	75.7	67	56

Figure 5. Performance of the spatial pyramid kernel in different conditions.

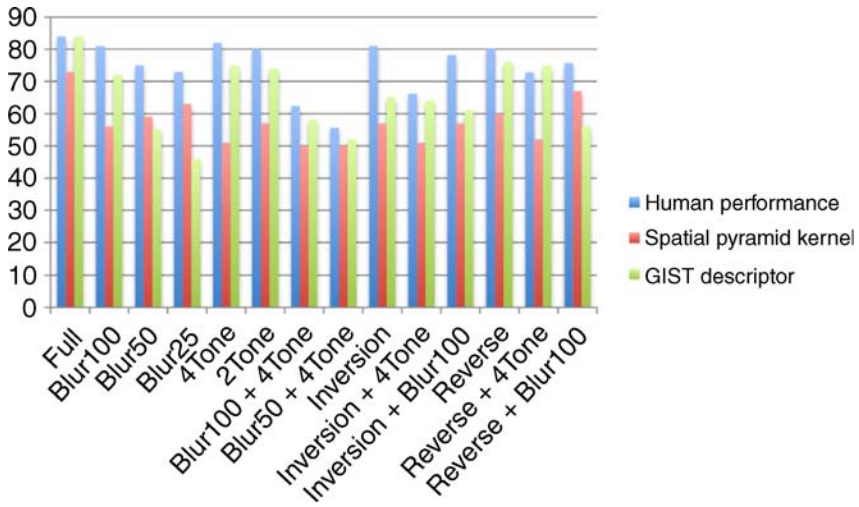

Figure 6. Plot of the data from Figure 5.

As a baseline, we tested the classifier with 200 full resolution (the *Full* condition) images, where 100 were of animal images and 100 were non-animal ones. We then tested the classifiers in all of the degraded conditions.

Results

For Experiment 1, the histogram of reaction times across all conditions is shown in Figure 3. Mean reaction time is at 275 ms. Mean accuracy results along with error bars and p-value are displayed for each condition in Figure 4, where chance performance is at 50% accuracy. Significance ($p < .05$) from the baseline condition, labeled *Full,* was found in the following conditions: *Blur25, Blur100 + 4Tone, Blur50 + 4Tone,* and *Inversion + 4Tone*. Further analysis looking for a link between reaction time and accuracy is displayed in Figure 7 and Figure 8. In Figure 7, we plot speed vs. accuracy across all subjects and conditions, and in Figure 8, we look at the relationship between speed and accuracy in each different condition. We do not find a relationship between speed and accuracy in either analysis.

For Experiment 2, we print the results of the spatial pyramid kernel and GIST descriptor for all conditions in Figure 5 and the graph the results in Figure 6. We observe that neither the GIST descriptor nor spatial pyramid kernel match human performance (Figures 7 and 8).

Conclusions

We can draw several implications from this study for models of visual category recognition. Our experiments show that cues used for rapid visual categorization are robust under significant spatial and luminance depth degradation, and it thereby puts marked constraints on models of rapid visual category recognition. Unlike faces, we also see that this effect is tolerant to both inversion and reversal of edge polarity.

We also note the non-linear interaction among the different visual cues. A striking example is in the *Blur100* and *4Tone* conditions. Individually, neither is significantly different from baseline but by combining the degradations into the *Blur100 + 4Tone* condition, we find a significant drop in human performance. We observe the same phenomenon with the *Inversion + 4Tone* condition (Figure 4). This finding leads to conclusions in direct

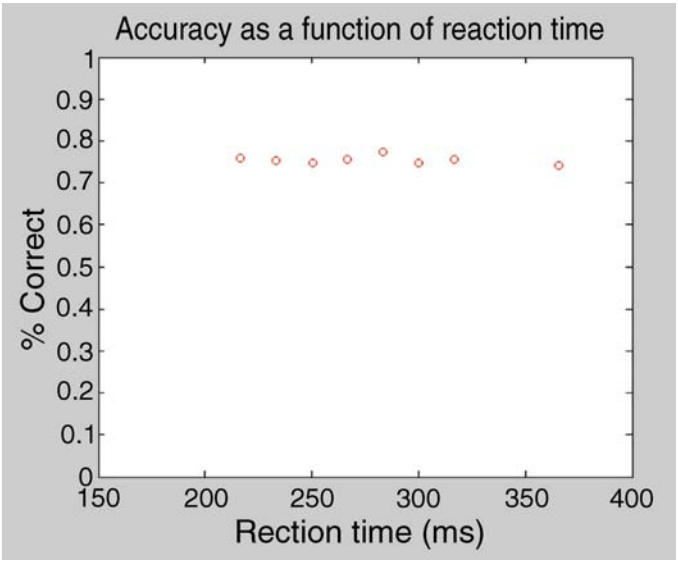

Figure 7. Plot of speed vs. accuracy across subjects and conditions. This plot uses adaptive binning so that each bin contains the same number of points.

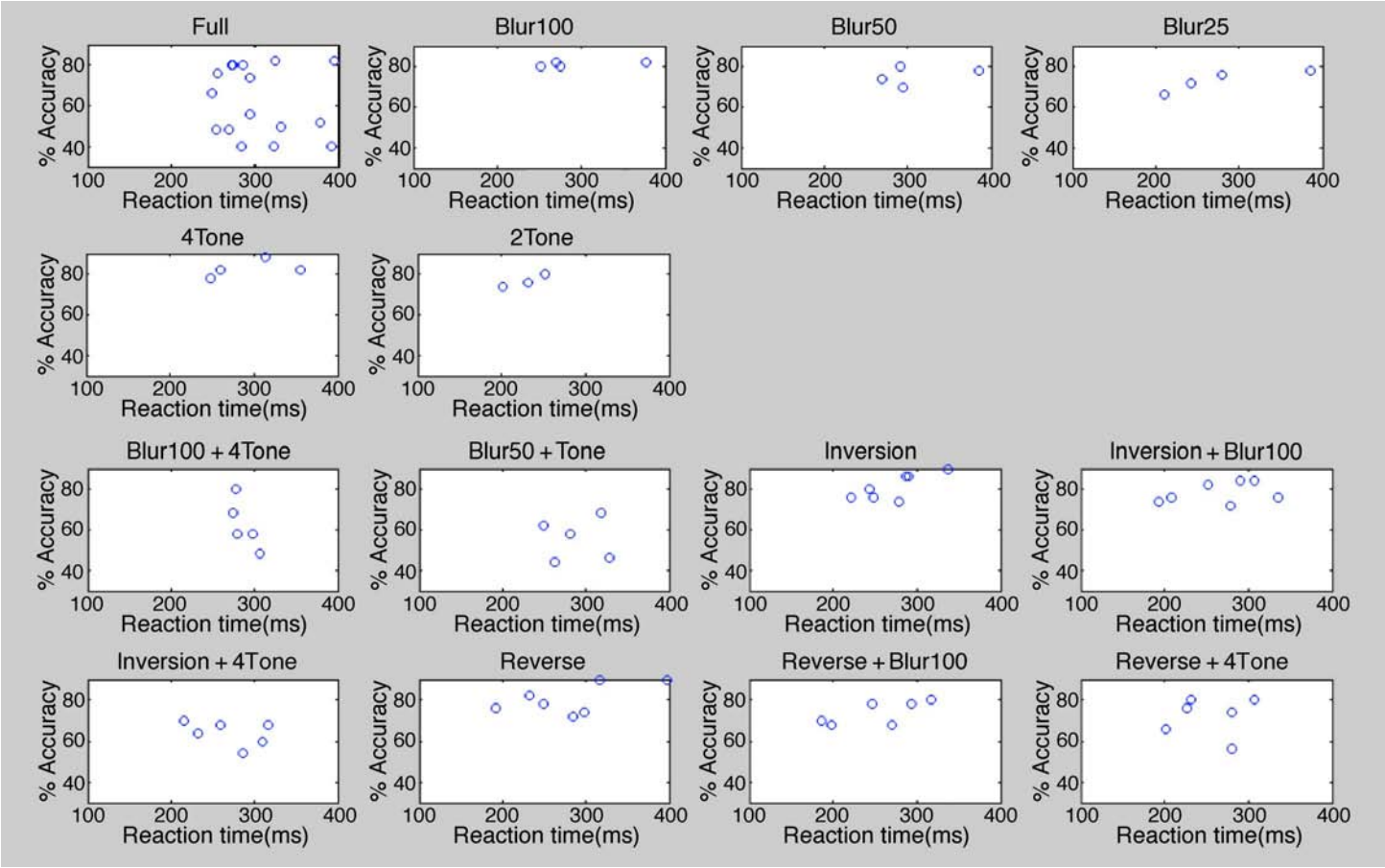

Figure 8. Plot of speed vs. accuracy on a per-condition basis. Each blue dot corresponds to the mean reaction time and accuracy for a given subject in the named condition. Please note that since "Full" was a common condition across all subjects, its plot above is a consolidated plot containing the responses from all subjects.

contrast to claims made by Guyonneau, Kirchner, and Thorpe (2006). In this study, the authors rotated the target and distracter images to different degrees. Since this transformation did not impact performance, the authors concluded that rapid category detection "could not depend on the global distribution of orientations within the image." Since we see significance with the *Inversion + 4Tone* condition, we conclude that both local features and global features are used in rapid category detection.

In addition, this study extends the initial work done by Thorpe et al. (1996) to naturally occurring conditions such as blurring and luminance depth degradation. In such degradations, the image information is not nearly as pristine as in Thorpe's original experiment. This study demonstrates that rapid category recognition is not merely a laboratory effect but is a process that can be at play in real-world settings.

Lastly, many groups are constructing computational models of object recognition, such as Frome, Singer, Sha, and Malik (2007), Lazebnik et al. (2006), and Serre et al. (2007). Since human subjects are quite robust to degradations, we suggest that robustness to degraded input be a critical measure for all such models.

Acknowledgments

This research was supported by a National Science Foundation graduate fellowship and the Arthur J. Chick Professorship. We would also like to thank Jeff Johnson for compiling the stimulus set.

Commercial relationships: none.
Corresponding author: Chetan Nandakumar.
Email: chetan@berkeley.edu.
Address: 750 Sutardja Dai Hall, University of California, Berkeley, CA 94720, USA.

Footnote

[1]In the Pulfrich Pendulum phenomenon, the subject is fitted with a piece of smoked glass over one eye and is presented a pendulum swinging from side to side. Instead of perceiving the lateral motion of the pendulum, the subject perceives the pendulum rotating circularly in depth. The explanation is that the reduced contrast from the smoked glass delays the visual signal from one eye to the brain thereby leading to the subjective perception of depth.

References

Bachmann, T. (1991). Identification of spatially quantized tachistoscopic images of faces: How many pixels does it take to carry identity? *European Journal of Cognitive Psychology, 3,* 85–103.

Brainard, D. H. (2004). Color constancy. In L. Chalupa & J. Werner (Eds.), *The visual neurosciences* (pp. 948–961). Cambridge, MA: MIT Press.

Cavanagh, P. (1991). What's up in top-down processing? *Representations of vision: Trends and tacit assumptions in vision research* (pp. 295–304). Cambridge, UK: Cambridge University Press.

Frome, A., Singer, Y., Sha, F., & Malik, J. (2007). Learning globally-consistent local distance functions for shape-based image retrieval and classification. *IEEE 11th International Conference on Computer Vision* (pp. 1–8).

Guyonneau, R., Kirchner, H., & Thorpe, S. J. (2006). Animals roll around the clock: The rotation invariance of ultrarapid visual processing. *Journal of Vision, 6*(10):1, 1008–1017, http://journalofvision.org/6/10/1/, doi:10.1167/6.10.1. [PubMed] [Article]

Harmon, L. D., & Julesz, B. (1973). Masking in visual recognition: Effects of two-dimensional noise. *Science, 180,* 1194–1197. [PubMed]

Ito, M., Tamura, H., Fujita, I., & Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology, 73,* 218–226. [PubMed]

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]

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (vol. 2, pp. 2169–2178). New York City.

Lowe, D. (1999). Object recognition from local scale-invariant features. *International Conference on Computer Vision* (pp. 1150–1157). Corfu, Greece.

Liu, C. H., Collin, C. A., Burton, A. M., & Chaurdhuri, A. (1999). Lighting direction affects recognition of untextured faces in photographic positive and negative. *Vision Research, 39,* 4003–4009. [PubMed]

Mooney, C. M. (1957). Age in the development of closure ability in children. *Canadian Journal of Psychology, 11,* 219–226. [PubMed]

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research, 155,* 23–36. [PubMed]

Olshausen, B. A., Anderson, C. H., & Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience, 13,* 4700–4719. [PubMed] [Article]

Rock, I., Di Vita, J., & Barbeito, R. (1981). The effect on form perception of change of orientation in the third dimension. *Journal of Experimental Psychology: Human Perception and Performance, 7,* 719–732. [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]

Tarr, M., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology, 21,* 233–282. [PubMed]

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

Thorpe, S. (2002). Ultra-rapid scene categorisation with a wave of spikes. *Proceedings of Biologically Motivated Computer Vision 2nd International Workshop* (pp. 1–15). Tubingen, Germany.

Torralba, A., Fergus, R., & Freeman, W. T. (2008). 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30,* 1958–1970. [PubMed]

Torralba, A., & Sinha, P. (2001). Detecting faces in impoverished images. *MIT AI Tech. Rep.* 2001-028.

Van Rullen, R., & Thorpe, S. (2001). Is it a bird? Is it a plane? Ultra-rapid visual categorization of natural and artifactual objects. *Perception, 30,* 655–668. [PubMed]

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology, 81,* 141–145.