ELSEVIER
Available online at www.sciencedirect.com
SCIENCE DIRECT®
Vision Research xxx (2004) xxx–xxx
Vision Research
www.elsevier.com/locate/visres

When is scene identification just texture recognition?

Laura Walker Renninger *, Jitendra Malik

UC Berkeley Vision Science and Computer Science, The Smith-Kettlewell Eye Research Institute, 2318 Fillmore Street, San Francisco, CA 94115, USA

Received 6 November 2002; received in revised form 25 March 2004

Abstract

Subjects were asked to identify scenes after very brief exposures (<70 ms). Their performance was always above chance and improved with exposure duration, confirming that subjects can get the gist of a scene with one fixation. We propose that a simple texture analysis of the image can provide a useful cue towards rapid scene identification. Our model learns texture features across scene categories and then uses this knowledge to identify new scenes. The texture analysis leads to similar identifications and confusions as subjects with limited processing time. We conclude that early scene identification can be explained with a simple texture recognition model.

Keywords: Scene perception; Texture discrimination; Natural images; Computational vision; Categorization

1. Introduction

1.1. Background

Our visual system can gather an incredible amount of information about an image in a glance. When a rapid sequence of photographs is presented (133–300 ms per image), subjects are surprising accurate at detecting a target image, whether the subject was precued with the target picture or just a verbal description of the objects in the scene (Potter, 1975). Singly presented pictures preceded and followed by a noise mask can be accurately detected in a later recognition task, even when the presentation was less than 120 ms in duration (Potter, 1976). When a natural image is shown for only 20 ms, subjects can detect whether or not an animal is present. Event-related potentials suggest that this decision is reached within 150 ms (Thorpe, Fize, & Marlot, 1996).

From these experiments, it is clear that we are quick to detect objects in the image but can we also detect or identify the place or scene depicted? Fortunately, we have names for scenes, such as "beach", "street" and "forest" (Tversky & Hemenway, 1983). It has been shown that subjects are, in fact, able to identify scene categories from a masked presentation of a photograph shown for only 45–135 ms (Schyns & Oliva, 1994). This identification can be as quick and accurate as the identification of a single object (Biederman, 1998). Rapid scene identification might be useful for creating a context in which objects can be located and identified (see Henderson & Hollingworth, 1999 for a review).

In general, subjects are very good at getting the "gist" of a scene, i.e. the conceptual category *and* layout (the schema) within a single fixation. Although the accurate timing of scene identification has not yet been determined, researchers believe it occurs within 100 ms. What sort of representation or information are we using to identify scenes so quickly? One possibility is that scene processing includes activation of a spatial layout, or schema of the scene. This is supported by phenomenon called boundary extension. Subjects presented with a scene will later remember having seen a greater extent of it than was depicted in the photograph (Intraub & Richardson, 1989). While the first demonstrations used a presentation time of 15 s, later experiments demonstrated that the phenomenon could still occur with 250 ms presentations (Intraub, Gottesman, Willey, & Zuk, 1996). There is also evidence for specialized brain areas that process places: the parahippocampal place area (PPA) is thought to process information about the lay-

* Corresponding author. Tel.: +1-415-345-2097; fax: +1-415-345-8455.
E-mail address: curlee@alum.mit.edu (L.W. Renninger).

0042-6989/$ - see front matter © 2004 Published by Elsevier Ltd.
doi:10.1016/j.visres.2004.04.006

out or geometry of the scene (Epstein & Kanwisher, 1998).

What cues or information in the image allows us to quickly activate the scene schema? Friedman (1979) proposed that the visual system might first recognize a "diagnostic object" that in turn triggers recognition of the scene. For example, a toaster would be diagnostic of a kitchen scene. Others argue that scenes may have distinctive holistic properties. For example, Biederman (1972) found that subjects have more difficulty recognizing and locating objects in a jumbled scene than in a coherent one, even when the objects remain intact. Loftus, Nelson, and Kallman (1983) studied the availability of holistic versus specific feature cues in picture recognition experiments. For brief presentations, subjects performed better when their response depended on the holistic cue. The arguments for a holistic property are consistent with the fact that we do not need to scan an image with our eyes or apply attention to particular objects in order to get the gist of the scene and most research supports this theory (Loftus et al., 1983; Metzger & Antes, 1983; Schyns & Oliva, 1994).

1.2. Texture as a holistic cue

By definition, a holistic cue is one that is processed over the entire visual field and does not require attention to analyze local features. Color is an obvious and strong cue for scene identification (Oliva & Schyns, 2000). Texture can be processed quickly and in parallel over the visual field (Beck, 1972; Bergen & Julesz, 1983), making it a candidate as well. Subjects can rapidly identify scenes without color, so we omit this dimension in our study and focus on the role of texture as a holistic cue.

An image region with one texture seems to "pop-out" or segregate easily from a background region with a perceptually different texture. What are the relevant features within a texture that allow this rapid discrimination? Julesz (1981, 1986) proposed that the first order statistics of "textons" determine the strength of texture discrimination. Just as phonemes are the elements that govern speech perception, textons are the elements that govern our perception of texture. Julesz described them to be locally conspicuous features such as blobs, terminators and line crossings. These features were described for the micropattern stimuli used in early texture discrimination experiments; however, these patterns are a poor representation of the real-world textures our visual system deals with. Filter-based models can represent the relevant local features that compose a texture and are easily applied to more realistic images (Bergen & Adelson, 1988; Fogel & Sagi, 1989; Landy & Bergen, 1991; Malik & Perona, 1990).

1.3. Summary of our approach

We investigate to what extent the texture features in a scene can be used for identification. First, subjects are asked to identify scenes with limited viewing times. Next, we reformulate the idea of textons to be the characteristic output of filters applied to a set of real images. Our model then identifies scenes by matching their texton histograms against learned examples. Finally, we compare our model performance against subject performance and conclude that a simple texture recognition model can mostly account for early human scene identification.

2. Experimental methods

2.1. Subjects

A total of 48 undergraduates were paid to participate in the 1-h experiment. Each participant had normal or corrected-to-normal vision and gave written consent in accordance with the University of California at Berkeley's Committee for the Protection of Human Subjects.

2.2. Stimuli

Images of scenes were taken from the Corel Image Database and various Internet sites. Our image database consists of 1000 images of scenes in 10 basic-level categories: beach, mountain, forest, city, farm, street, bathroom, bedroom, kitchen and livingroom. These scenes can also be placed in three superordinate-level categories: natural/outdoor, man-made/outdoor and man-made/indoor (Fig. 1). We randomly selected 250 of these images as the training set from which the model learned prototypical textures. The remaining 750 images were used as the test set to measure the ability of our subjects and our model to identify scenes.

2.3. Procedure

The experiment was run in a dimly lit room to reduce visual distractions. Subjects fixated a marker that blinked before stimulus onset to reduce spatial and temporal uncertainty. The target was a grayscale image displayed briefly (<70 ms) depending on the test condition. Subjects never saw the same image twice. Following the target, a jumbled scene mask immediately appeared for 20 ms to interrupt perceptual processing and to restrict target availability to the exposure duration. Each rectangular region in the mask was sampled from a different scene category. Next, a uniform gray field was displayed for 500 ms, followed by two word

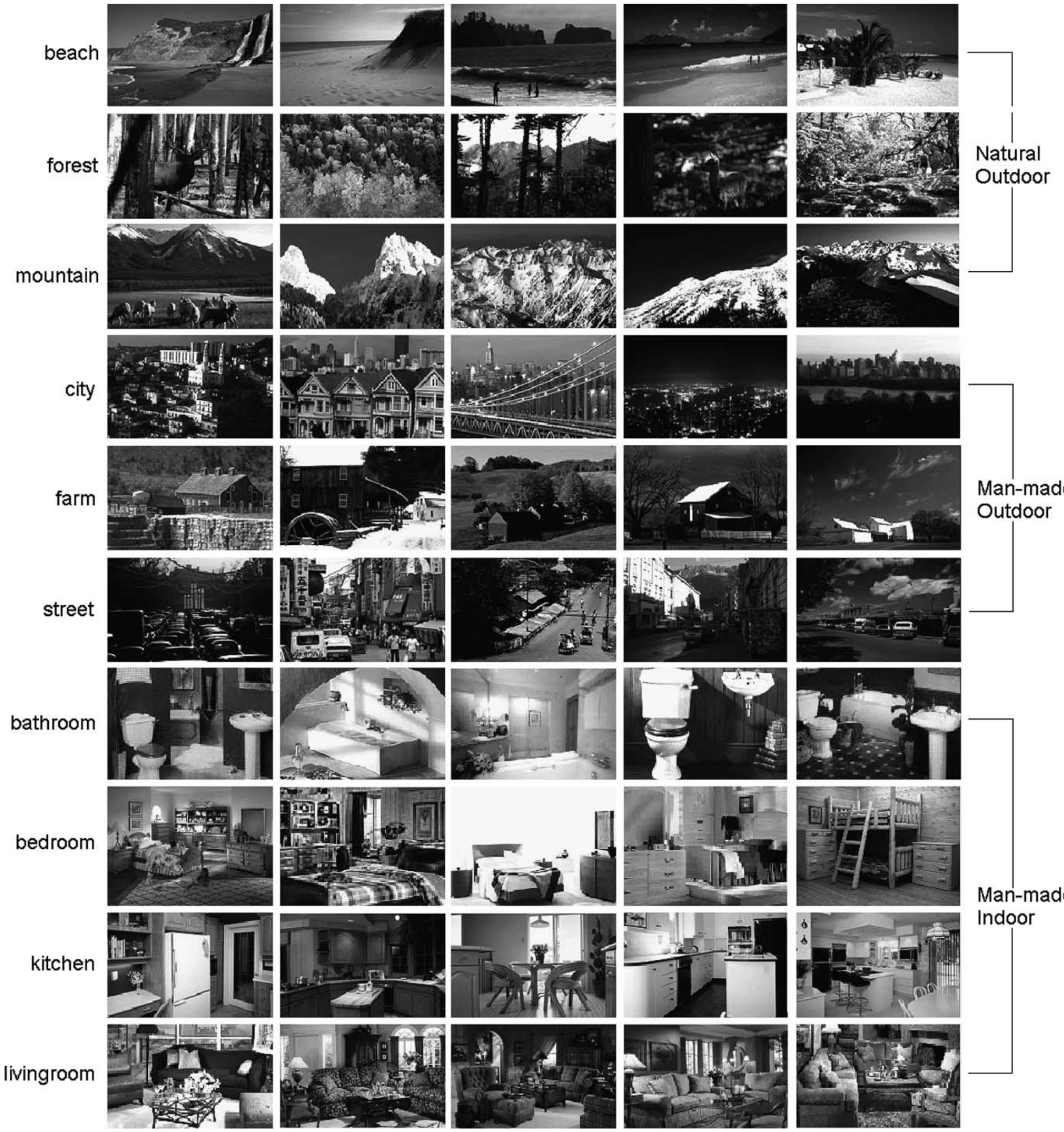

Fig. 1. Pictured here are some example images from the ten scene categories used in this paper. Each row is labeled with its basic-level (left) and superordinate-level (right) category. The dataset is available at http://www.cs.berkeley.edu/projects/vision/shape/.

choices for 2.5 s. One word choice corresponded to the grayscale image presented and the other was chosen randomly from the remaining nine scene labels. Subjects responded in this two-alternative forced choice task by selecting the word on the left or right that best described the target image (Fig. 2).

2.4. Design

A preliminary study in which subjects viewed the scenes for 150 ms was conducted to validate the experimental setup. Performance was near perfect, confirming that the task is reasonable given the labeling of the dataset, choice of mask and viewing distance. With this setup we can study the effects of target exposure dura-

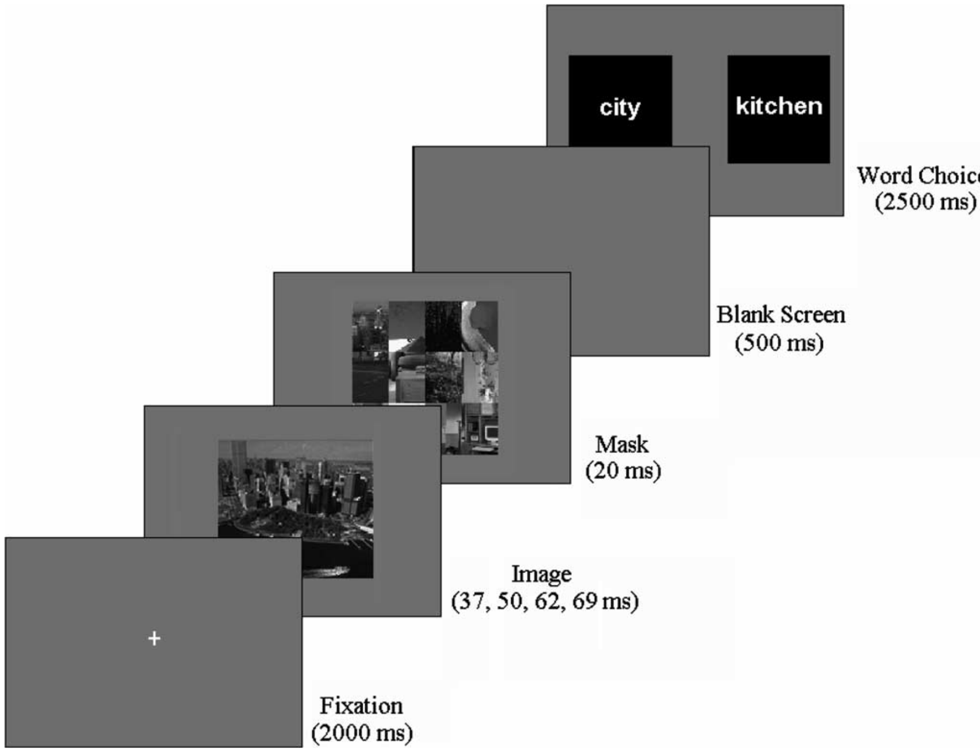

Fig. 2. Subjects were shown grayscale scenes for 37, 50, 62 or 69 ms followed by a jumbled scene mask and two word choices. The 2AFC task was to select the word that best described the target.

tion on scene identification. Four conditions were tested in which the target was displayed for 37, 50, 62 or 69 ms. There were 11, 15, 8 and 14 participants for the respective conditions. On a given trial, the target image was presented followed by its corresponding category label and one of the other nine category labels. To explore all 10 categories, an experimental block consisted of 90 trials. Most subjects completed two experimental blocks during the session.

2.5. Apparatus

Stimuli were presented on a PC running Windows 2000 and the BitmapTools presentation software for Matlab (developed by Payam Saisan, under the supervision of Martin Banks). The display was set at 800×600 pixels and 256 colors with a refresh rate of 160 Hz. Subjects did not use a chinrest, but were instead instructed to sit with their back against the chair to maintain a viewing distance of approximately 2.5 m. Responses were collected on a BTC Wireless Multimedia Keyboard 5113RF. The images were displayed on a mid-gray background and presented foveally. Absolute image dimensions varied, but were scaled to a height of 380 pixels (7.6 in. displayed) to subtend a visual angle of approximately 5.3°.

3. Texture model

Several researchers have constructed algorithms that extract low-level features from images in order to classify them into two categories, for example indoor versus outdoor (Szummer & Picard, 1998), city/suburb versus other (Gorkani & Picard, 1994) and city/suburb versus landscape (Vailaya, Jain, & Zhang, 1998). They achieve reasonable classification performance by weighting particular discriminating features, for example, cities will have more vertical edge energy than flat landscapes (see also Oliva & Torralba, 2001).

The classification schemes mentioned above apply high-level or top-down knowledge in the form of a class-specific template or feature weighting. Because subjects are quick to identify scenes in a glance without prior cues, we avoid learning class-specific features and instead examine the ability of early vision mechanisms to delineate scene categories in a purely bottom-up fashion.

Our model learns what local texture features occur across all scene categories by first filtering the set of 250 training images with V1-like filters, then remembering their prototypical response distributions. The number of occurrences of each feature within a particular image is stored as a histogram, creating a holistic texture descriptor for that image. When identifying a new image, its histogram is matched against stored examples. Another distinction from past work is that a texture analysis deliberately ignores global spatial relationships across the scene.

3.1. Learning universal textons

3.1.1. Training set

The training set contains 250 images (25 examples for each of the 10 scene classes) that were not used in the testing phase. The model learns universal textons from this training set.

3.1.2. "V1" filters

As mentioned earlier, Julesz' formulation of a texton was suited to micropatterns but not to generic images. Filter models can also describe human texture discrimination and are better suited to our purpose. The formulation of these filters follows descriptions of simple cell receptive fields in V1 of the primate visual cortex (DeValois & DeValois, 1988). In particular, these receptive fields can be characterized as Gabor functions, difference of Gaussians and difference of offset Gaussians. For our model, we use first and second derivatives of Gaussians to create quadrature pairs,

$$f_{\text{odd}}(x,y) = G'_{\sigma_1}(y)G_{\sigma_2}(x)$$
$$f_{\text{even}}(x,y) = G''_{\sigma_1}(y)G_{\sigma_2}(x)$$

where $G_\sigma(x)$ represents a Gaussian with standard deviation σ. The ratio σ_2:σ_1 is a measure of the elongation of the filter. The filters are built at three scales for spatial frequency selectivity and rotated for orientation selectivity (Fig. 3). The three filter scales, taking viewing distance of the target stimulus into account, are equal to 3.6, 2.5 and 1.8 c/deg. This range of spatial frequencies is shifted lower than our peak sensitivities under photopic conditions, as might be expected given the brief (high temporal frequency) nature of our stimuli and the lower light levels used during the experiment (DeValois & DeValois, 1988).

3.1.3. Clustering filter response distributions

As a first step in our texture analysis, the image is convolved with the filter bank to produce a vector of filter responses $I * f(x_0, y_0)$, which characterizes the image patch centered at x_0, y_0. Because texture has spatially repeating properties, similar vectors of responses will reoccur as texture features reoccur in the image. To learn what the most prevalent features are, we filter the entire training set of images and cluster the resulting response vectors to find 100 prototypical responses. In particular, we utilized the K-means clustering algorithm available in the Netlab toolbox for Matlab. The prototypical responses found correspond to common texture features in the training images. We call these prototypes "universal textons" to stress that the features are learned across multiple examples of the scene categories, rather than within a single image (Malik, Belongie, Leung, & Shi, 2001; Malik, Belongie, Shi, & Leung, 1999). We can visualize a universal texton by multiplying its filter response vector by the pseudoinverse of the filterbank (Jones & Malik, 1992). Our universal textons are illustrated in Fig. 4(a). They correspond to edges and bars with varying curvature and contrast.

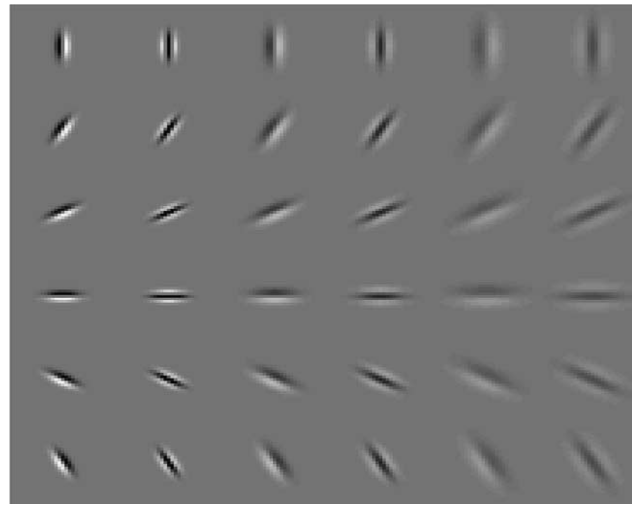

Fig. 3. Our model uses a filterbank of first and second derivatives of a Gaussian to estimate texture features at each pixel in the image. The 36 filters consist of two phases (even and odd), three scales (spaced by half-octaves), and six orientations (equally spaced from 0 to π). Each filter has 3:1 elongation and is L_1 normalized for scale invariance.

3.1.4. Histograms of activity in texton channels

Once we have a vocabulary of universal textons, we can analyze any image into texton channels and examine the resulting histogram. Each pixel in an image is assigned to a texton channel based on the vector of filter responses it induces. The value of the kth histogram bin for an image is then found by counting how many pixels are in texton channel k. The histogram represents texton frequencies in the image:

$$h_i(k) = \sum_{j \in \text{image}} I[T(j) = k]$$

where $I[\cdot]$ is the indicator function and $T(j)$ returns the texton assigned to pixel j (Malik et al., 1999, 2001). In essence, the histogram is a holistic representation of texture in the image that ignores gross spatial relationships (Fig. 4(b)).

3.2. Identifying new scenes

3.2.1. Test stimuli

The 750 images not used for learning universal textons are used here to test the ability of our texture model to identify scenes.

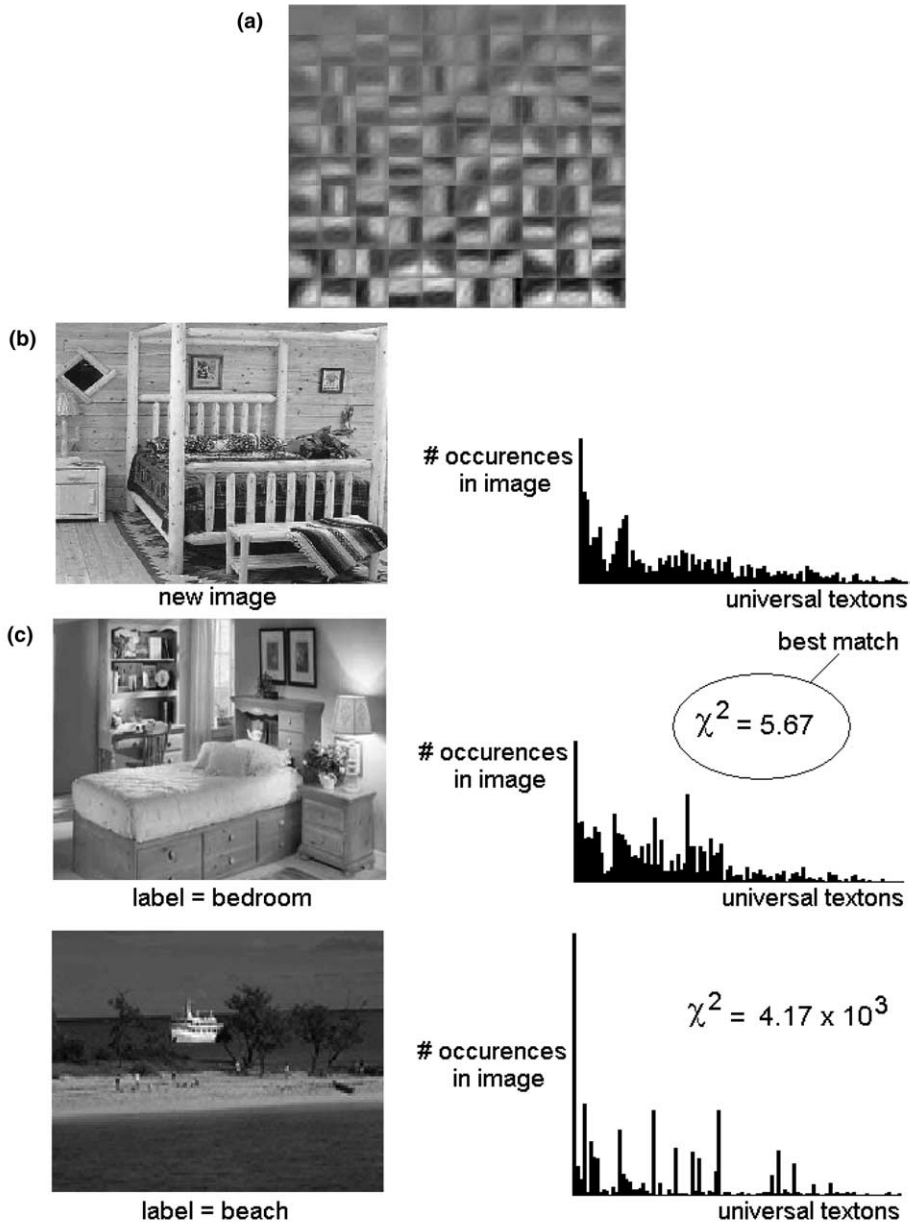

Fig. 4. (a) The 100 texture features found across the training images (sorted by increasing norm). These ‘‘universal textons’’ correspond to edges and bars of varying curvature and contrast. (b) Each pixel in an image is assigned to a texton channel based on its corresponding vector of filter responses. The total activity across texton channels for a given image is represented as a histogram. (c) Test images are identified by matching their texton histograms against stored examples. The χ^2 similarity measure indicates our test image is more similar to a bedroom than a beach scene in this case.

3.2.2. Comparing histograms

For each new image, we can develop a description of its texture by creating a universal texton histogram. To find the closest match, this histogram is compared to stored histograms for the training images using the χ^2 similarity measure

$$\chi^2(h_i, h_j) = \frac{1}{2}\sum_{k=1}^{K}\frac{[h_i(k) - h_j(k)]^2}{h_i(k) + h_j(k)},$$

where h_i and h_j are the two histograms and K is the total number of bins (universal textons). If χ^2 is small, the two images are similar in their texture content (Fig. 4(b) and (c)). The model is tested with the same 2AFC task as our subjects, and the target scene is assigned the label of its closest match.

4. Data analysis

Subjects were not allowed to see the same image more than once to prevent recognition and learning effects on the data, therefore we do not have data for one subject across the time conditions. We are also interested in how the model compares to typical subject performance. For these reasons, we collapse data across subjects within a single time condition. We measure statistics from the consolidated data using bootstrapping techniques (Efron & Tibshirani, 1993). The datasets for each time condition are resampled with replacement at least 1000 times. From each resampling, the statistic of interest is calculated. The central limit effect causes the resulting distribution over the statistic to tend toward normality as the number of samples increases. The mean and standard deviation of this distribution provide the best estimate of the statistic and the standard error of the estimate. The 95% confidence intervals are also taken from this distribution and used to determine statistical significance.

Bootstrapping techniques assume that the observed data is representative of the underlying population. This is a valid assumption given that we collapse data across 48 subjects and trials were fully randomized. When we break the analysis down to examine specific error types, the number of samples available for bootstrapping is drastically reduced. For the error analysis, we discard the 62 ms time condition. This condition had the fewest number of subjects and is somewhat redundant with the 69 ms time condition. It also simplifies our presentation of the confusion analysis.

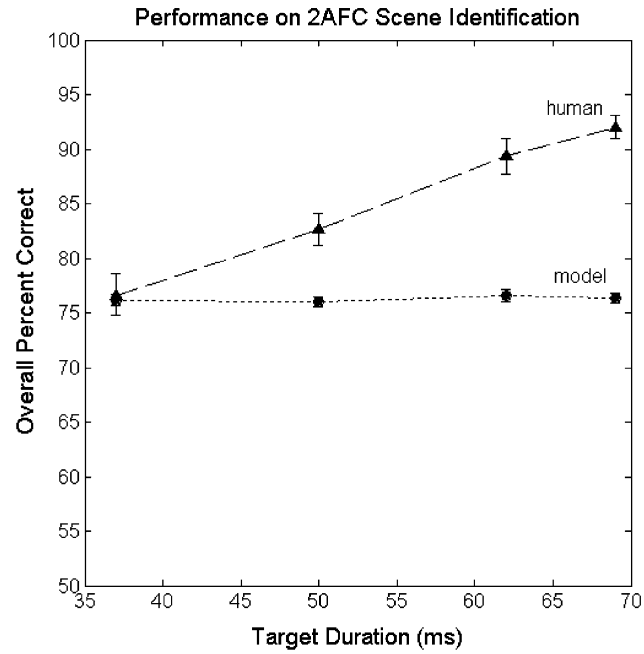

Fig. 5. Subject accuracy in the 2AFC scene discrimination task improves with increased presentation time. The percent correct is plotted with its 95% confidence intervals for 48 subjects (11, 15, 8 and 14 subjects at 37, 50, 62 and 69 ms). Chance performance is 50% correct.

5. Results and discussion

5.1. 2AFC scene identification

Subjects and the model performed well above chance on the 2AFC task for all time conditions. Performance is similar for the model and the subject at 37 ms, but the subjects outperform the model overall at longer durations (Fig. 5). With 69 ms, subjects are performing above 90% correct, confirming that the gist of a scene can be processed within one fixation. The model performance could differ at the four time conditions because it is presented with whatever images the subjects saw for that condition, however, performance stayed nearly constant at 76% correct.

Subjects made comments during the experiments that they saw "*the* kitchen" or "*the* forest" when referring to the stimuli, indicating that they often perceived only one instance of each scene, when in fact, there were many examples of each scene class presented to them during the experiment. This is consistent with previous experiments that suggest we get the gist of a scene quickly, but it takes longer to retain the specific details of those scenes in memory (Loftus et al., 1983; Potter, 1976).

5.2. Correct identification of basic-level categories

The proportion of correct responses for the model is most similar to human responses at 37 ms across the 10 basic-level scene categories (Fig. 6). Identical performance occurs along the diagonal line in this figure. Significant correlation occurs between the model and humans at both 37 and 50 ms. At 37 ms, the model is doing a better job on beach and kitchen scenes, but humans are far superior on mountain scenes. Subjects reported that mountains just seemed to "pop out" at them. In this case, subjects seem to be able to make use of large-scale shape information (the triangle of the mountain against the sky). As time progresses to 50 ms, the performance is still correlated, but humans are doing a better job on categorizing 9 of the 10 basic-level scene categories. With longer exposures, subjects are clearly outperforming the texture model.

5.3. Identification errors

With the briefest exposures, we might expect human errors to be noisy and unpredictable, given the difficulty of the scene identification task. As exposure durations are increased, however, we would expect these errors to become more systematic. Can the pattern of these errors be explained by our texture model?

Both humans and the model can identify a scene as a member of its superordinate category before its basic-level category is identified. When we group error rates at the superordinate-level, we see stronger correlation at 50 ms for both beach and mountain scenes (Fig. 7). Significant positive correlation for basic-level identification does not occur until 69 ms. Correlations at one category

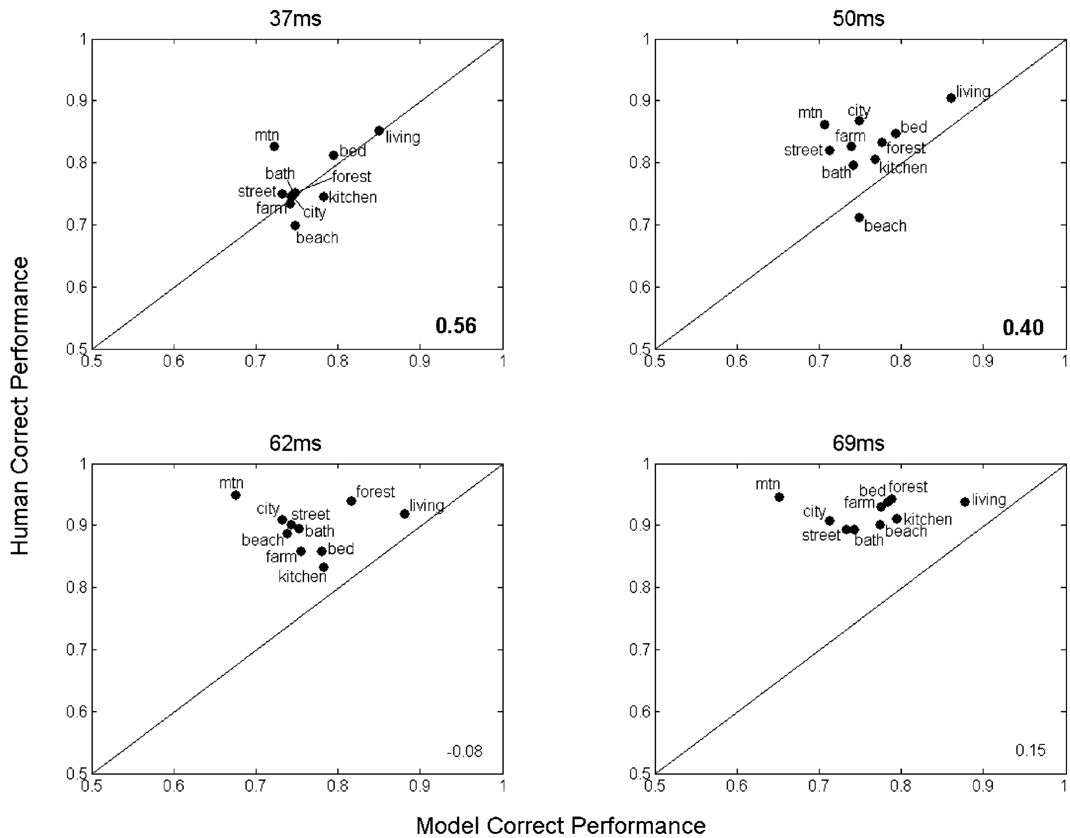

Fig. 6. Comparison of model and human performance in correctly classifying scenes at the basic-level. Identical performance occurs along the diagonal. Correlation coefficients are noted in the lower right corner of each plot. Performance of the model is significantly correlated with human performance at 37 and 50 ms (bold values).

level do not necessarily require correlation at the other, but they are indicative of how the errors cluster together.

Both humans and the model can distinguish between scenes that have distinctive orientation energy profiles. For example, subjects and the model perform similarly on indoor/man-made scenes which have energy at all orientations, and beach and mountain scenes which have energy confined to more specific orientations.

Scenes with visually similar textures are confused by both humans and the model. When error rates are low (69 ms), cities are heavily confused with streets and farms are confused with beaches. Clearly cities and streets have buildings and other man-made structures. If you remove the few man-made structures from a farm scene, they would indeed look much like a beach scene with a distinct horizon line and mostly flat ground.

While the successes of the model are certainly interesting, the failures are also informative. Humans seem to be making an outdoor versus indoor discrimination very early during scene processing. For example, forest and street scenes have a lot of vertical orientation energy and our model gets them confused with indoor as well as outdoor man-made scenes, as would be expected. Our subjects, however, rarely confuse these scenes with indoor man-made scenes, resulting in poor or even significantly negative correlations between humans and the model (Fig. 7). This special ability of our subjects might again be related to the spatial arrangement of regions or textures in the scene. Outdoor scenes will tend to have a horizon line dividing the untextured sky from the textured ground. Clearly, spatial relationships should be captured in a complete model for early scene identification. Several approaches have been described in the object recognition literature (e.g. Belongie, Malik, & Puzicha, 2002; Burl & Perona, 1996) and could be easily adapted to scene identification.

6. Summary

Scene identification is achieved quite rapidly by the human visual system and may be useful in creating context for object localization and identification during real-world tasks. Previous data and this current study demonstrate that subjects can process the gist of a scene within a single fixation. Comparison of our model with human performance demonstrates that texture provides a strong cue for scene identification at both the superordinate and basic category levels during early scene

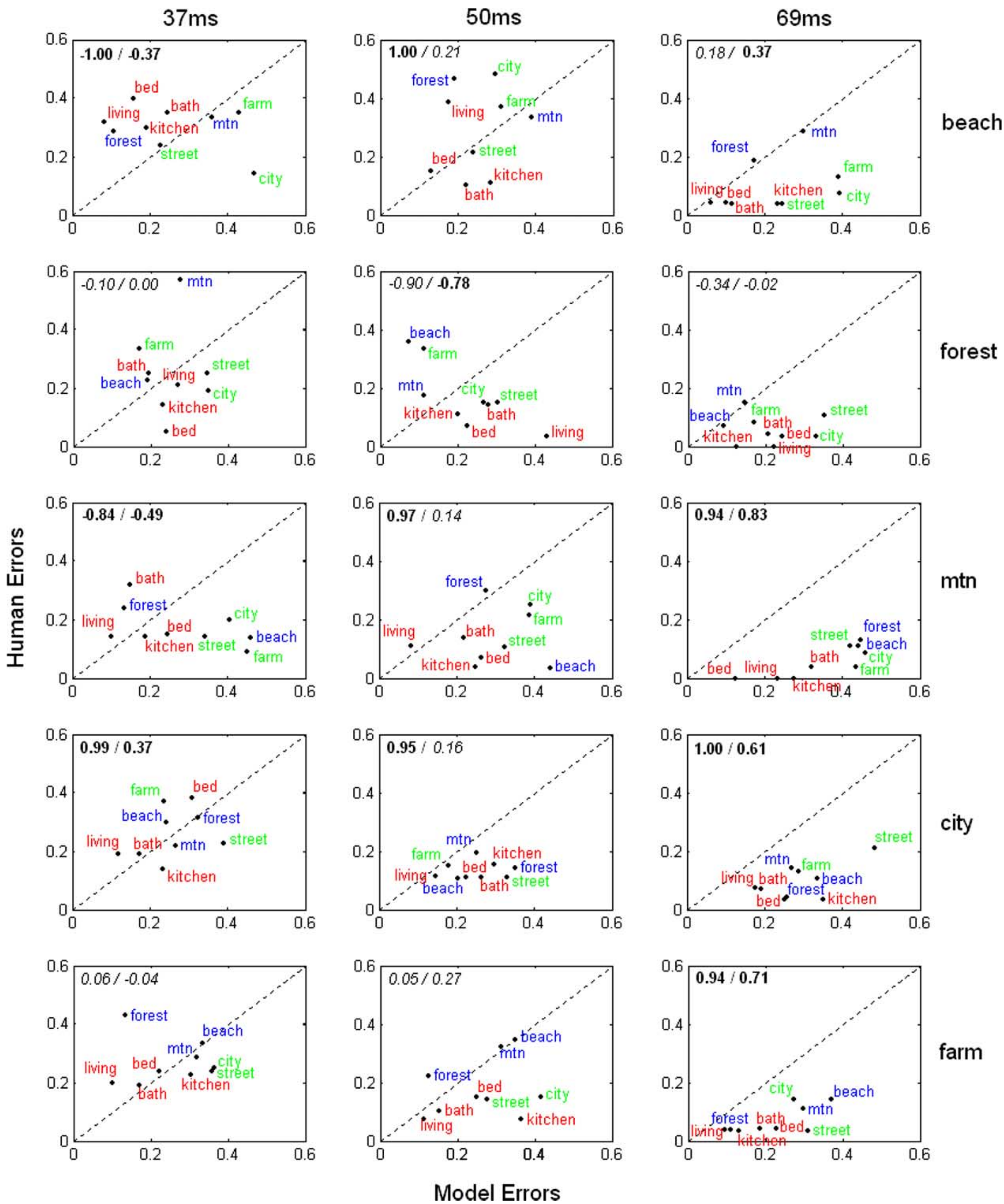

Fig. 7. Comparison of model and human errors when classifying scenes at the basic-level, broken down by scene category. Data from the 62 ms condition has been omitted for simplicity (see Section 4). The superordinate category of each label is indicated by its color. Red = man-made/indoor; Green = man-made/outdoor; Blue = natural/outdoor. Correlation estimates are in the upper left-hand corner for error analysis at the superordinate-level (left) and the basic-level (right). Significant values are in boldface type. Identical error rates fall along the diagonal line. When the subjects are more confused by a scene category, it falls above the line. When our model is more confused by a scene category, it falls below the line.

processing. Failures to describe human performance seem to be due to lack of knowledge of spatial relations. In addition to texture, subjects may have access to coarse segmentation or shape cues in the image. Texture alone was able to account for correct categorization and error patterns on 8 out of 10 scenes categories. From this we conclude that a simple texture recognition model mostly explains early scene identification.

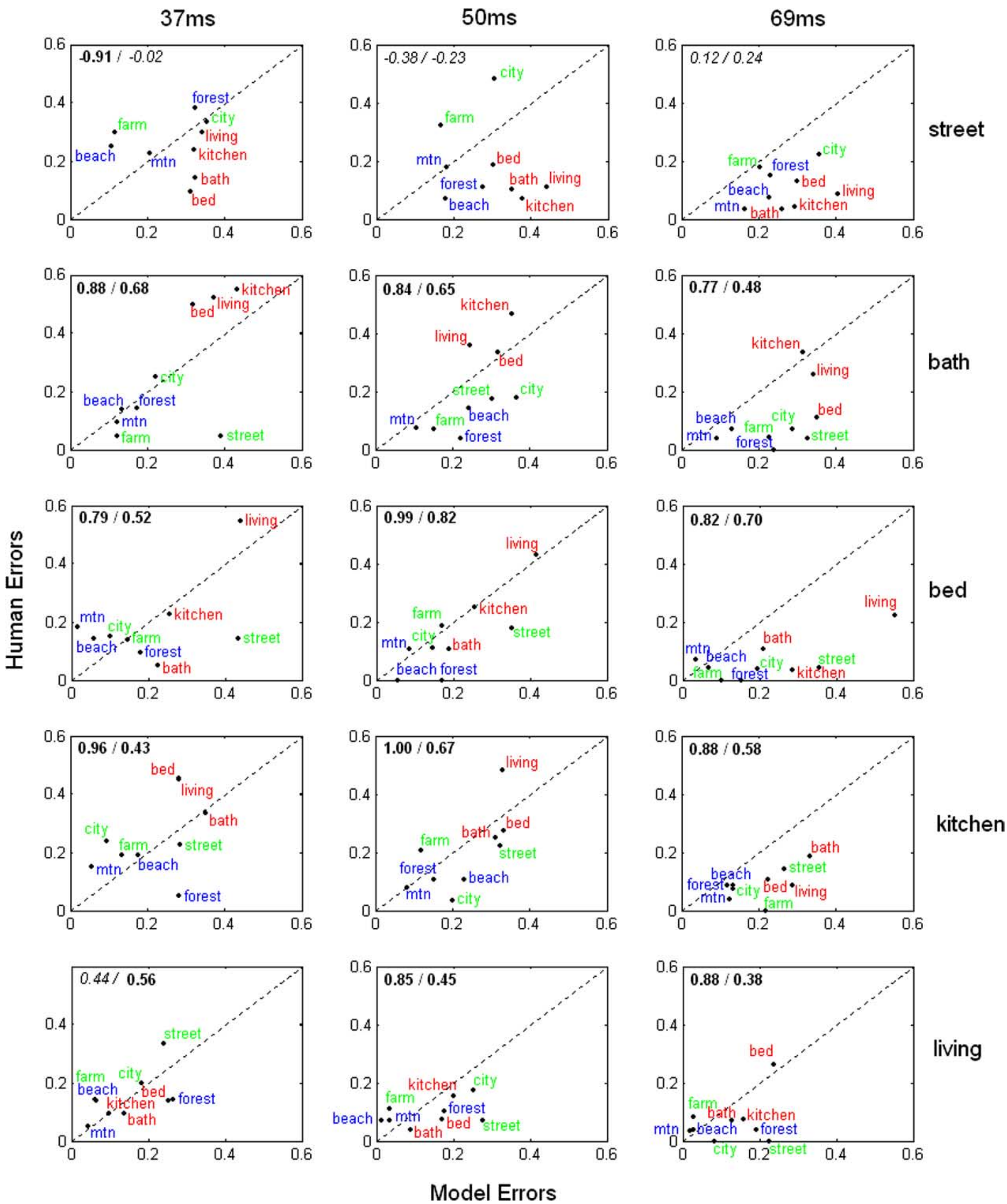

Fig. 7. (*continued*)

Acknowledgements

We would like to thank the UC Berkeley Computer Vision and Vision Science groups, especially Alyosha Efros, Ahna Girschick, Temina Madon, Kim Miller, Laura Sanftner and Neil Renninger for participating in the earliest experiments and for helpful suggestions regarding the manuscript. We would also like to thank the reviewers for their rigorous perusal of this manuscript. This research was supported in part by the Office of Naval Research, grant N00014-01-1-0890.

References

Beck, J. (1972). Similarity grouping and peripheral discriminability under uncertainty. *American Journal of Psychology, 85*, 1–19.

Belongie, S., Malik, J., & Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(4), 509–522.

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

Bergen, J. R., & Julesz, B. (1983). Rapid discrimination of visual patterns. *IEEE Transactions on Systems, Man, and Cybernetics, 13*, 857–863.

Biederman, I. (1972). Perceiving real-world scenes. *Science, 177*, 77–80.

Biederman, I. (1998). Aspects and extension of a theory of human image understanding. In Z. Pylyshyn (Ed.), *Computational processes in human vision: An interdisciplinary perspective*. New Jersey: Ablex Publishing Corporation.

Burl, M. C., & Perona, P. (1996). Recognition of planar object classes. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (pp. 223–230).

DeValois, R. L., & DeValois, K. K. (1988). *Spatial vision*. Oxford: Oxford University Press.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature, 392*(6676), 598–601.

Fogel, I., & Sagi, D. (1989). Gabor filters as texture discriminator. *Biological Cybernetics, 61*, 103–113.

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

Gorkani, M. M., & Picard, R. W. (1994). Texture orientation for sorting photos "at a glance". In *Proceedings of the 12th international conference on pattern recognition* (pp. A459–464).

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology, 50*, 243–271.

Intraub, H., Gottesman, C. V., Willey, E. V., & Zuk, I. J. (1996). Boundary extension for briefly glimpsed photographs: Do common perceptual processes result in unexpected memory distortions? *Journal of Memory and Language, 35*(2), 118–134.

Intraub, H., & Richardson, M. (1989). Wide-angle memories of close-up scenes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(2), 179–187.

Jones, D., & Malik, J. (1992). Computational framework to determining stereo correspondence from a set of linear spatial filters. *Image and Vision Computing, 10*, 699–708.

Julesz, B. (1981). Textons, the elements of texture perception, and their interactions. *Nature, 290*, 91–97.

Julesz, B. (1986). Texton gradients: The texton theory revisited. *Biological Cybernetics, 54*, 245–251.

Landy, M. S., & Bergen, J. R. (1991). Texture segregation and orientation gradient. *Vision Research, 31*, 679–691.

Loftus, G. R., Nelson, W. W., & Kallman, H. J. (1983). Differential acquisition rates for different types of information from pictures. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 35*, 187–198.

Malik, J., Belongie, S., Leung, T., & Shi, J. (2001). Contour and texture analysis for image segmentation. *International Journal of Computer Vision, 43*, 7–27.

Malik, J., Belongie, S., Shi, J., & Leung, T. (1999). Textons, contours and regions: Cue integration in image segmentation. In *Proceedings of the IEEE international conference on computer vision, vol. 2* (pp. 918–925).

Malik, J., & Perona, P. (1990). Preattentive texture discrimination with early vision mechanisms. *Journal of the Optical Society of America A, 7*, 923–932.

Metzger, R. L., & Antes, J. R. (1983). The nature of processing early in picture perception. *Psychological Research, 45*(3), 267–274.

Oliva, A., & Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology, 41*, 176–210.

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

Potter, M. C. (1975). Meaning in visual search. *Science, 187*(4180), 965–966.

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

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science, 5*(4), 195–200.

Szummer, M., & Picard, R. W. (1998). Indoor–outdoor image classification. In *IEEE international workshop on content-based access of image and video databases*.

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

Tversky, B., & Hemenway, K. (1983). Categories of environmental scenes. *Cognitive Psychology, 15*, 121–149.

Vailaya, A., Jain, A. K., & Zhang, H. J. (1998). On image classification: City images vs. landscapes. *Pattern Recognition, 31*, 1921–1936.