Submitted to Vision Research, October 2002

When is scene recognition just texture recognition?

Laura L. Walker and Jitendra Malik

UC Berkeley Vision Science & Computer Science

Abstract

Subjects were asked to discriminate scenes after very brief exposures (37-69 ms). Their performance was always above chance and increased 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 categorize new scenes. The texture analysis leads to similar categorizations and confusions as subjects with limited processing time. We conclude that a simple texture discrimination model mostly explains early scene identification.

Keywords: scene perception; texture discrimination; natural images; computational vision; categorization

1. Introduction

Our visual system can process the gist of a scene within a single fixation. This keen ability has been studied in several ways. When a rapid sequence of photographs is presented, with each image displayed for only 100ms, subjects show a remarkable performance in later recognition tests (Potter, 1976). Our scene processing abilities can also aid us in rapid object detection. When a natural image is shown for only 20ms, subjects can detect whether or not an animal is present. Event-related potentials suggest that this decision is reached within 150ms (Thorpe, Fize & Marlot, 1996).

It is not understood how we are able to process and understand scenes so rapidly. What information is the visual system making use of? 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 thought the objects remained 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 a gist of the scene.

By definition, a holistic cue is one that is processed over the entire visual field and does not require attention to process local features. Color is an obvious and strong cue for scene identification (Oliva & Schyns, 2000). We also know that texture can be

processed pre-attentively 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; Malik & Perona, 1990; Landy & Bergen, 1991).

1.2. Summary of our approach

We investigate whether or not the texture features in a scene aid recognition. First, subjects are asked to discriminate scenes with limited viewing times (37-69ms). Next, we reformulate the idea of textons as the characteristic output of filters applied to a set of real images. Our model then classifies scenes by matching their texton histograms against learned examples. Finally, we compare our model performance against subject

performance and conclude that a simple texture model can account for early human scene recognition.

2. Experimental Methods

2.1. Methods

2.1.1. Scenes

Images of scenes were taken from the Corel Image Database and various Internet sites. Our image database consists of 1,000 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 750 images from the dataset to be used in these experiments. The remaining 250 images were held out and used in training the model.

2.1.2. Participants

A total of 48 undergraduates were paid to participate in the 1-hour 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.

2.1.3. Apparatus and Stimuli

The experiments were run in a dimly lit room to reduce visual distractions. 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 x 600 pixels and 256 colors with a refresh rate of 160 Hz. Subject responses were collected on a BTC Wireless Multimedia Keyboard 5113RF to allow for a 3 meter viewing distance. The images of scenes subtended a visual angle of 12 degrees and were presented on a mid-gray background.

2.1.4. Procedure

Subjects fixated a marker that blinked before stimulus onset to reduce spatial and temporal uncertainty. The stimulus consisted of a grayscale image displayed for 37, 50, 62 or 69ms. Subjects viewed each image for the same length of time, and never saw the same image twice. Following the stimulus, a jumbled scene mask appeared for 20ms to interrupt perceptual processing and to restrict stimulus availability to the exposure duration. Each block in the mask was taken from a different scene category. The mask was followed by two simultaneous word choices for 2.5 seconds. One word choice corresponded to the grayscale image presented and the other was chosen randomly from the remaining scene labels. Subjects responded in this two-alternative forced choice task by selecting the word on the left or right (Fig. 2).

2.2. Results

All subjects performed well above chance in this task. Discrimination performance increased with increasing exposure times, as expected (Fig. 3).

2.3. Constructing Confusion Matrices from 2AFC data

We are interested in how subjects classify each scene when ten scene labels are possible. For our ten scenes, we can represent subject performance in a 10x10 confusion matrix. Across each row of the matrix we enter the proportion of times a particular scene is classified into each of the 10 categories. The diagonal entries will contain the percent correct performance, and the off-diagonal entries represent how often a scene is misclassified or confused for other scene classes.

In a 10AFC task, we would have shown our subjects an image and asked them to which of 10 possible scene categories it belonged. From this data, we could directly construct the confusion matrix, but such a task becomes too difficult for the subject. Instead, we ran the easier 2AFC task. Fortunately, we can expand our observations into a full confusion matrix according to the equations

$$P_A = \frac{1}{\left[1 + \sum_{K \neq A} N_K\right]} \text{ and } P_K = N_K P_A,$$

where P_A is the probability of selecting the correct label when scene A is displayed and P_K is the probability of selecting the incorrect scene label K when scene A is displayed. N_K is the ratio of misses to hits over all trials when scene A is shown with label choices A and K. The development of this conversion and an illustrative example can be found in Appendix A.

To compute confusion matrices for the three superordinate-level categories, we simply sum performance over the basic-level categories. We estimate the variance in our experimental data using the Monte Carlo method by running simulations in which we used only half of our observations (chosen at random) to compute the percent correct

performance. We used 100 of these simulated observations to compute our sample statistics. The confusion matrices for each experimental condition are tabulated in Appendix B for reference. Their main results will be discussed in section 4. Note that after this conversion, chance performance is 10%.

3. Texture model for scene classification

The main goal of our model is to investigate whether or not a texture analysis of the image can explain subject performance in our scene discrimination task. Any reasonable texture discrimination model could be substituted.

Our model first learns what local texture features occur across scene categories by filtering a set of training images with V1-like filters and remembering the prototypical response distributions. The number of occurrences of each feature in an image is stored as a histogram, creating a holistic texture descriptor for that image. When classifying a new image, its histogram is matched against stored examples.

3.1. Learning Universal Textons

3.1.1. Training Set

Roughly 25 examples from each basic-level category were held from the original dataset for learning universal textons.

3.1.2. Filters

As mentioned earlier, Julesz' formulation of a texton was suited to micropatterns but not to general images. Filter models can also describe human texture discrimination

and are better suited to our purpose (Bergen & Adelson, 1988; Fogel & Sagi, 1989; Malik & Perona, 1990; Landy & Bergen, 1991). 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,

$$f_{odd}(x,y) = G^{/}_{\sigma_1}(y) G_{\sigma_2}(x)$$

$$f_{even}(x,y) = G^{//}_{\sigma_1}(y) G_{\sigma_2}(x)$$

where $G_{\sigma}(x)$ represents a Gaussian with standard deviation σ. The ratio $\sigma_2 : \sigma_1$ 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. 4).

3.1.3. Clustering

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 we found correspond to common texture features in the training images. We call these prototypes "universal textons" to stress that these features are learned across a set of images, rather than within a single image (Malik, Belongie, Shi, & Leung, 1999,

2000). 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. 5a. They correspond to edges and bars with varying curvature and contrast.

3.1.4. Histograms of Activity in Texton Channels

Once we know our universal textons, we can analyze an image into texton channels and build its 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 k^{th} 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:

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

where $I[\cdot]$ is the indicator function and $T(j)$ returns the texton assigned to pixel j (Malik et al., 1999, 2000). In essence, the histogram is a holistic representation of texture in the image (Fig 5b).

3.2. Classifying New Scenes

3.2.1. Test stimuli

The 750 images used in the psychophysical experiments are used here to test the performance of our texture-model to identify scenes.

3.2.2. Comparing Histograms

For each new image, we can develop a description of its global 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. 5b,c). Using a simple nearest-neighbor classifier, our new image is assigned the label of its closest match (Hastie, Tibshirani, & Friedman, 2001).

3.2.3. Confusion Matrix

After labeling each test image, we characterized the texture model performance with a confusion matrix (Appendix B). We estimated the variance in our data as before, using the Monte Carlo method. We compute performance for the superordinate-level categories by summing over the performance on basic-level categories. Our classifier assigns the correct label above chance (10%), indicating that the texture in an image gives us some information about scene identity.

4. Analysis of Model

4.1. Discrimination of Superordinate-level Categories

Confusion matrices for categorization at the superordinate-level can be found in Appendix B. We can better visualize this data by representing the numbers with gray

levels (Fig. 6). Correct performance is represented along the diagonal, which will be white for perfect performance. Off-diagonal entries that are not black represent confusion between scene categories. The confusion matrix produced by our model looks most like the matrices for subjects with 37-50ms of exposure to the image. Subjects perform better with increased viewing time, and outperform the model at 62 and 69ms.

4.2. Discrimination of Basic-level Categories

The confusion matrices for basic-level categorization of the scenes can be found in Appendix B. These larger matrices are more difficult to compare with a gray-level representation, so we will first focus on the correct classifications (Fig. 7). The scenes are sorted along the bottom axis according to increasing performance by the texture model.

Our model either outperforms or has equivalent performance as subjects with 37ms of exposure time, with the exception of mountain scenes. When viewing time is increased to 50ms, subjects do better and outperform the model on 4 scenes. By 69ms, subjects consistently outperform the model.

Why was our model weak in explaining subject performance on mountain scenes? In informal interviews after the experiments, 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). Subjects also made comments 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 an experiment. This is

consistent with previous experiments that suggest we get the gist of a scene with one fixation, but it takes longer to retain the specific details of those scenes in memory (Loftus et al., 1983; Potter, 1976).

4.3. Analysis of Confusion

We have discussed how well our model performs compared to subjects in assigning the correct label, but it is also interesting to look at common errors they make. Fig. 8 shows a histogram representation of the confusion matrices for basic-level categorization. The magnitude of the bar represents the frequency that the corresponding label is chosen (key to left of the figure). A good strategy for viewing these plots is to think of the bars as activity in the 10 different "scene channels". In each column, the indicated scene category is presented and subjects' category responses are plotted down the rows as viewing time is increased. The bottom row is the response profile from the texture model. The star marks the correct channel, where a peak in activity will be observed when the scene is correctly labeled most of the time.

Focusing first on the left column and starting with the top plot, subjects are shown a beach image for 37ms. The activity is roughly uniform across all scene channels. Moving down to the next plot, subjects are allowed to view a beach image for 50 ms. Now the activity is a little more specific – it has increased for outdoor scenes and decreased for indoor scenes. Moving down again, subjects are now viewing a beach image for 62ms. At this point, we see a nice peak in the beach channel, but also a good amount of activity in the mountain channel. In the fourth row, subjects are viewing a beach image for 69ms and most of the activity is correctly in the beach channel.

The same sort of scrutiny can be applied to each column in the figure. In general, we find that activity moves first to the correct superordinate-level category and then resolves itself to the basic-level category. In some cases the basic-level category is resolved as early as 37ms (mountains and livingrooms) and the remainder of the categories are resolved by 50ms.

Our model, based on texture in the images, is able to resolve the basic-level categories like subjects with 62ms or less of viewing time. The one exception is mountains, as we saw before. Subjects are able to resolve this category fairly well as early as 37ms, while the model has only shifted towards outdoor scenes. Again, this is likely due to the large shape cue available. The model gets bedrooms scenes highly confused with livingroom scenes, but subjects demonstrate this same confusion until 62ms.

5. Summary

Subjects were asked to discriminate scenes at the basic-level after very brief exposures (37-69 ms). Their performance was always above chance and increased with exposure duration, confirming that subjects can get the gist of a scene with one fixation. We have proposed that a simple texture analysis of the image can provide a useful cue towards this scene identification. Our model learns common texture features, or universal textons, across 10 basic-level scene categories, and then describes each image with a histogram of texton frequencies. By matching texton histograms for new images against those stored in memory, we can classify the new image based on its closest match. When comparing the model performance against subject data, we found that our

texture analysis led to similar classifications and confusions as subjects with limited processing time. We conclude that a simple texture discrimination model mostly explains early scene identification.

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. This research was supported in part by the Office of Naval Research grant number N00014-01-1-0890.

Appendix A

We are interested in the probability of classification over ten categories, but our subject data was gathered for a 2AFC task. Here we demonstrate how a simple conversion can be used to construct a full confusion matrix for the ten scene categories.

In our 2AFC task, scene A is shown with word choices A and K, where K is one of the other 9 scene categories. The proportion of times A is correctly chosen as the scene shown is equal to

$$\frac{Hits}{Trials} = \frac{P_A}{P_A + P_K}$$

and the proportion of times K is incorrectly chosen is equal to

$$\frac{Misses}{Trials} = \frac{P_K}{P_A + P_K}$$

Let N_K be the ratio of misses to hits

$$N_K = \frac{Misses}{Hits}$$

$$\Rightarrow P_K = N_K P_A$$

Given multiple scene choices, the sum of probabilities of choosing any scene must sum to one. In other words, the entries along one row in the final confusion matrix must sum to one.

$$P_A + \sum_{K \neq A} P_K = 1$$

Given this constraint and the ratio of hits to misses for a type of trial, we can convert any 2AFC observance into a row in our confusion matrix.

$$P_A + \sum_{K \neq A} N_K P_A = 1$$

$$\Rightarrow P_A = \frac{1}{\left[1 + \sum_{K \neq A} N_K\right]}$$

To better illustrate the conversion, here is an example. Assume we only have three scene categories, A, B and C, and that we obtain the results in Table 1 with a 2AFC task. Let's compute the row in the confusion matrix for scene A. In other words, given three scene choices, how often is scene A classified as A, B and C?

$$N_B = \frac{20.0}{80.0} = 0.250 \text{ and } N_C = \frac{5.0}{95.0} = 0.053$$

$$P_A = \frac{1}{[1 + (N_B + N_C)]} = \frac{1}{1 + 0.250 + 0.053} = 0.767$$

$$P_B = N_B P_A = (0.250)(0.767) = 0.191$$

$$P_C = N_C P_A = (0.053)(0.767) = 0.041$$

When scene A is shown, it will be classified correctly as A 77% of the time, incorrectly as B 19% of the time and incorrectly as C 4% of the time. The final confusion matrix for this example can be found in Table 2.

Appendix B

Table 3 shows the full confusion matrices for subject and texture model data for basic-level categorization of the scenes. Across each row is the percentage of times an image of a given class is categorized as each of the 10 categories. Correct performance measures occur along the diagonal. The Monte Carlo method was used to compute the standard deviations and mean observations. Simulations were run in which half of the data was used to compute the mean of an observation. 100 observations were used. Table 4 shows the confusion matrices for subject and texture model data for

superordinate-level categorization of the scenes. These matrices are computed by summing across the basic-level categories.

List of References

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

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.

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

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.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer.

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. (1986). Texton gradients: the texton theory revisited. *Biological Cybernetics, 54,* 245-251.

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

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. *Proceedings of the IEEE International Conference on Computer Vision, 2,* 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.

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

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

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

List of Figures

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 category (right). This dataset is available at http://www.cs.berkeley.edu/~lwalk.

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

Fig. 3. Subject accuracy in the 2AFC scene discrimination task improves with increased presentation time. Data is for 48 subjects (11, 15, 8 and 14 subjects at 37, 50, 62 and 69ms). Chance performance is 50% correct. At 69ms, accuracy is around 90%, confirming that the gist of a scene can be processed within one fixation.

Fig. 4. Our model uses this filterbank to estimate texture features at each pixel in the image. The 36 filters consist of 2 phases (even and odd), 3 scales (spaced by half-octaves), and 6 orientations (equally spaced from 0 to π). Each filter has 3:1 elongation and is L_1 normalized for scale invariance.

Fig. 5. (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 categorized 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 in this case.

Fig. 6. Superordinate-level confusion matrices for subjects and the model (Appendix B) are illustrated with gray levels. The order of the scene categories from top to bottom, left to right is: natural/outdoor, man-made/outdoor and man-made/indoor. Correct categorization occurs along the diagonal, which will be white for perfect performance. The amount of misclassification is represented in the off-diagonal blocks. At the superordinate-level, the model performs similar to subjects with 37-50ms of image exposure.

Fig. 7. Percent correct classifications are plotted versus basic-level scene categories, sorted by model performance. To allow direct comparison of the model with subjects, the 2AFC data has been recomputed as 10AFC data – chance performance is 10%. The dotted curves represent standard error measures for the model. (a) The model performs the same or better than subjects with 37ms of exposure, with the exception of mountain scenes. (b) When viewing time is increased to 50ms, the model is still able to account for subject performance on more than half of the scene categories. (c & d) The model cannot account for subject performance with more than 70ms of exposure.

Fig. 8. Plotted here is the response activity in different "scene channels" for subjects and the model. Across the x-axis of each plot are the ten scene categories, and the bars are

colored according to their superordinate category (see the key). In each column, one scene category has been shown. Moving down rows, subjects view that scene for 37, 50, 62 or 69ms. A star represents the activity in the correct channel. The bottom row is response activity based on our texture analysis of the scene. See the text for a discussion.

List of Tables

Figure 1.

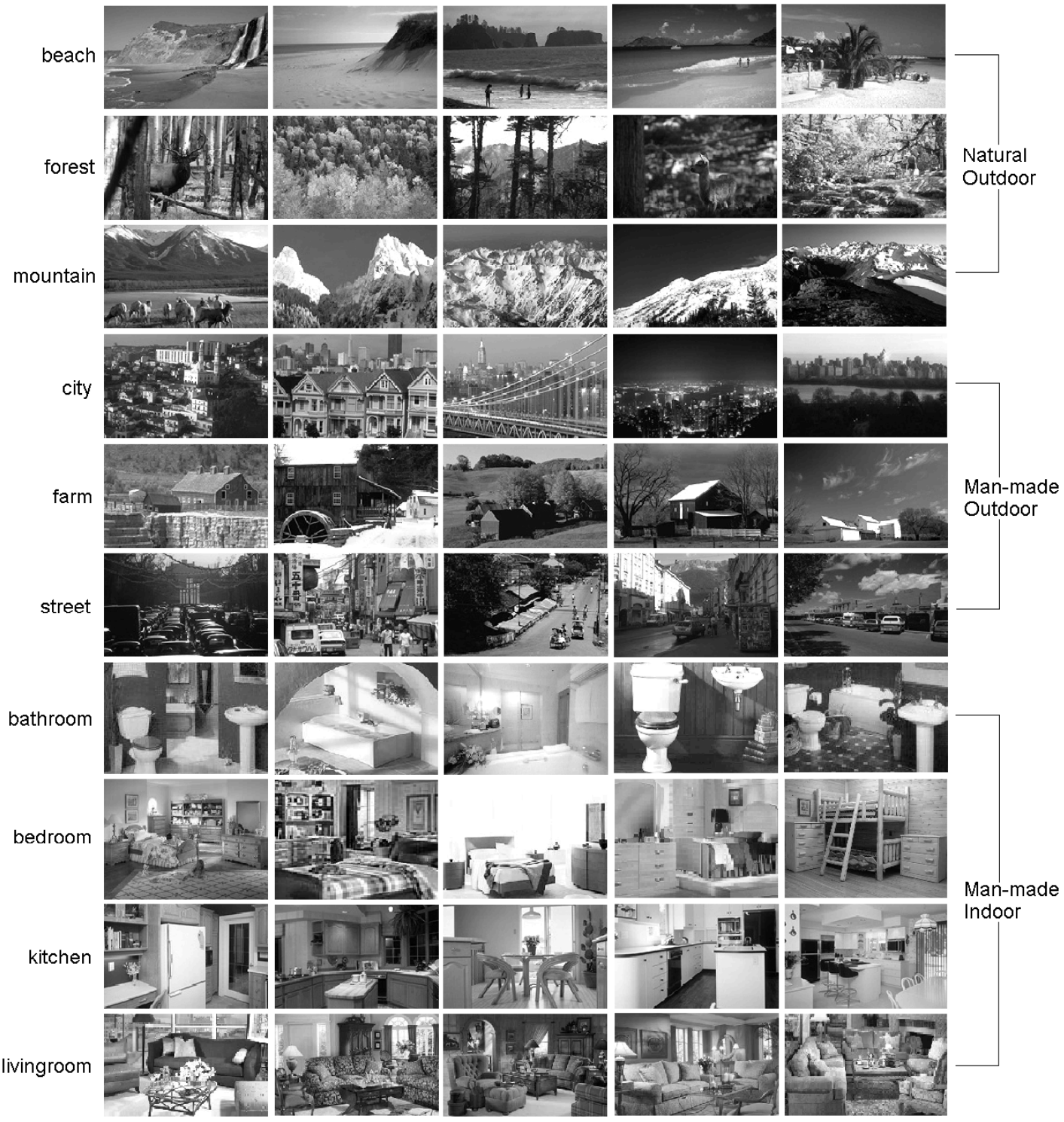
beach
forest
mountain
city
farm
street
bathroom
bedroom
kitchen
livingroom
Natural Outdoor
Man-made Outdoor
Man-made Indoor

Figure 2.

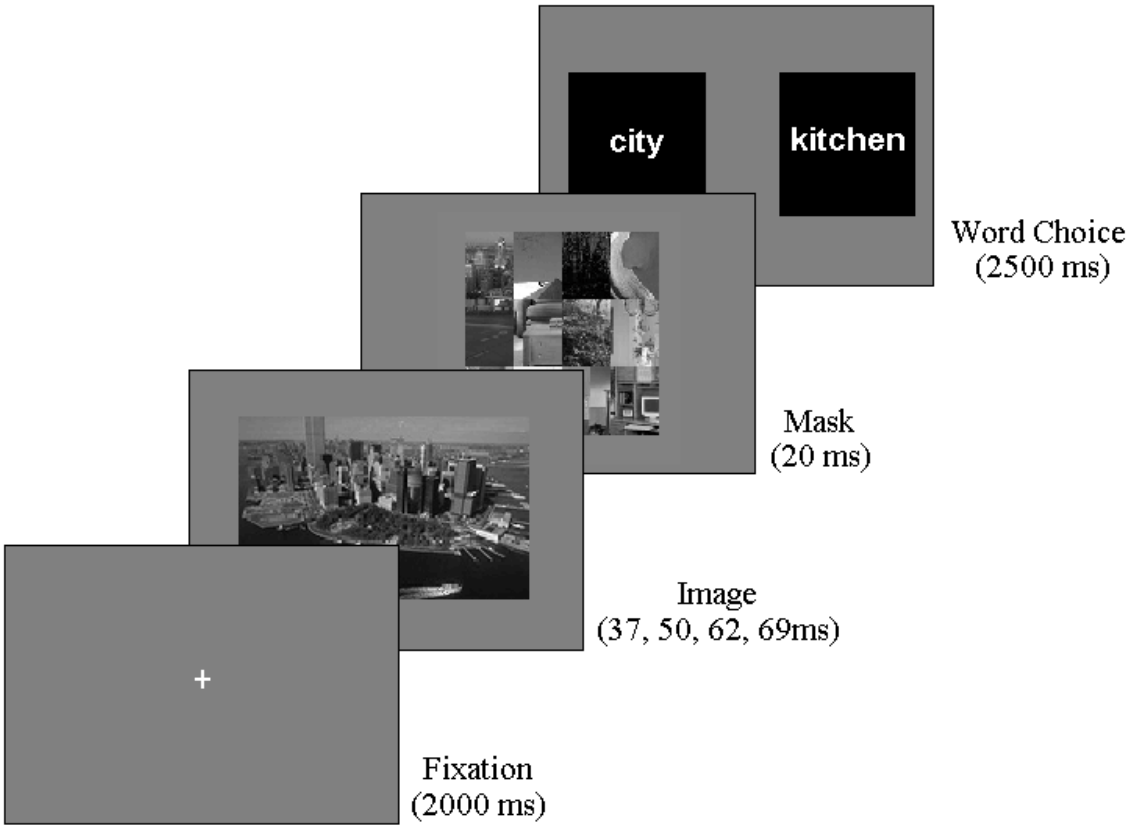
city
kitchen
Word Choice
(2500 ms)
Mask
(20 ms)
Image
(37, 50, 62, 69ms)
+
Fixation
(2000 ms)

Figure 3.

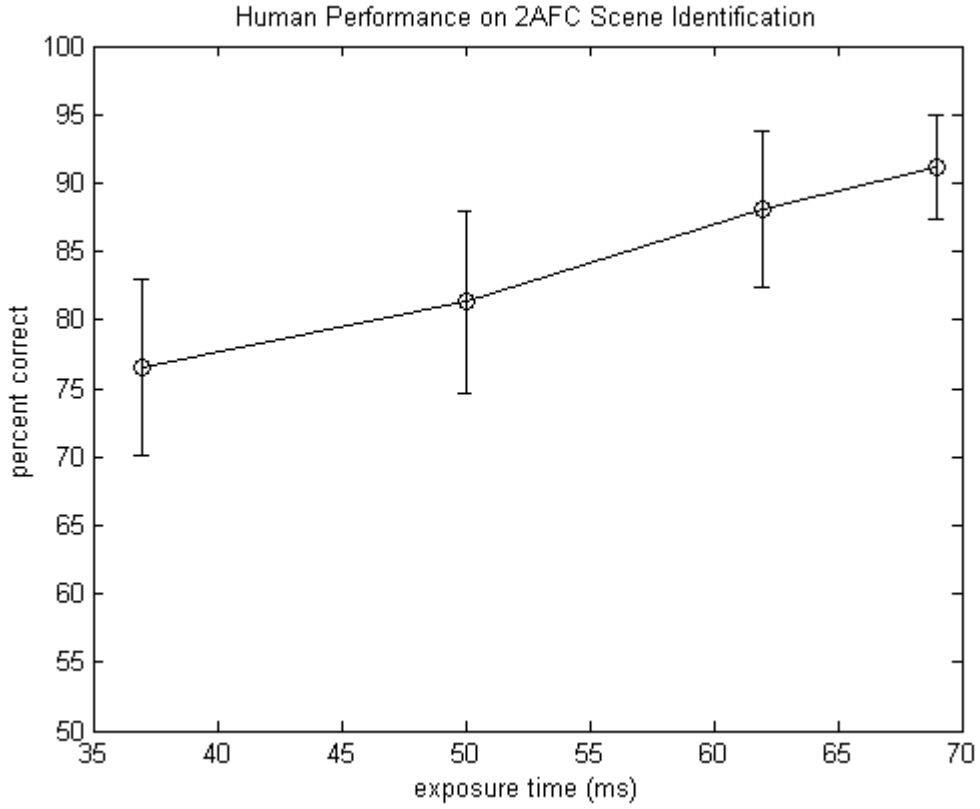
Human Performance on 2AFC Scene Identification
percent correct
exposure time (ms)
100
95
90
85
80
75
70
65
60
55
50
35
40
45
50
55
60
65
70

Figure 4.

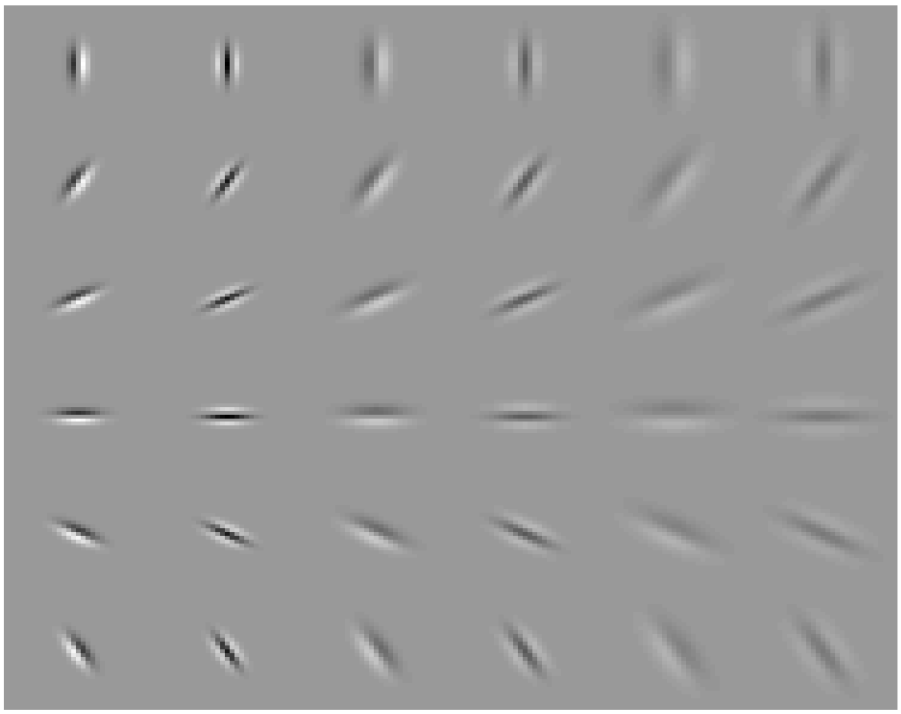

Figure 5.

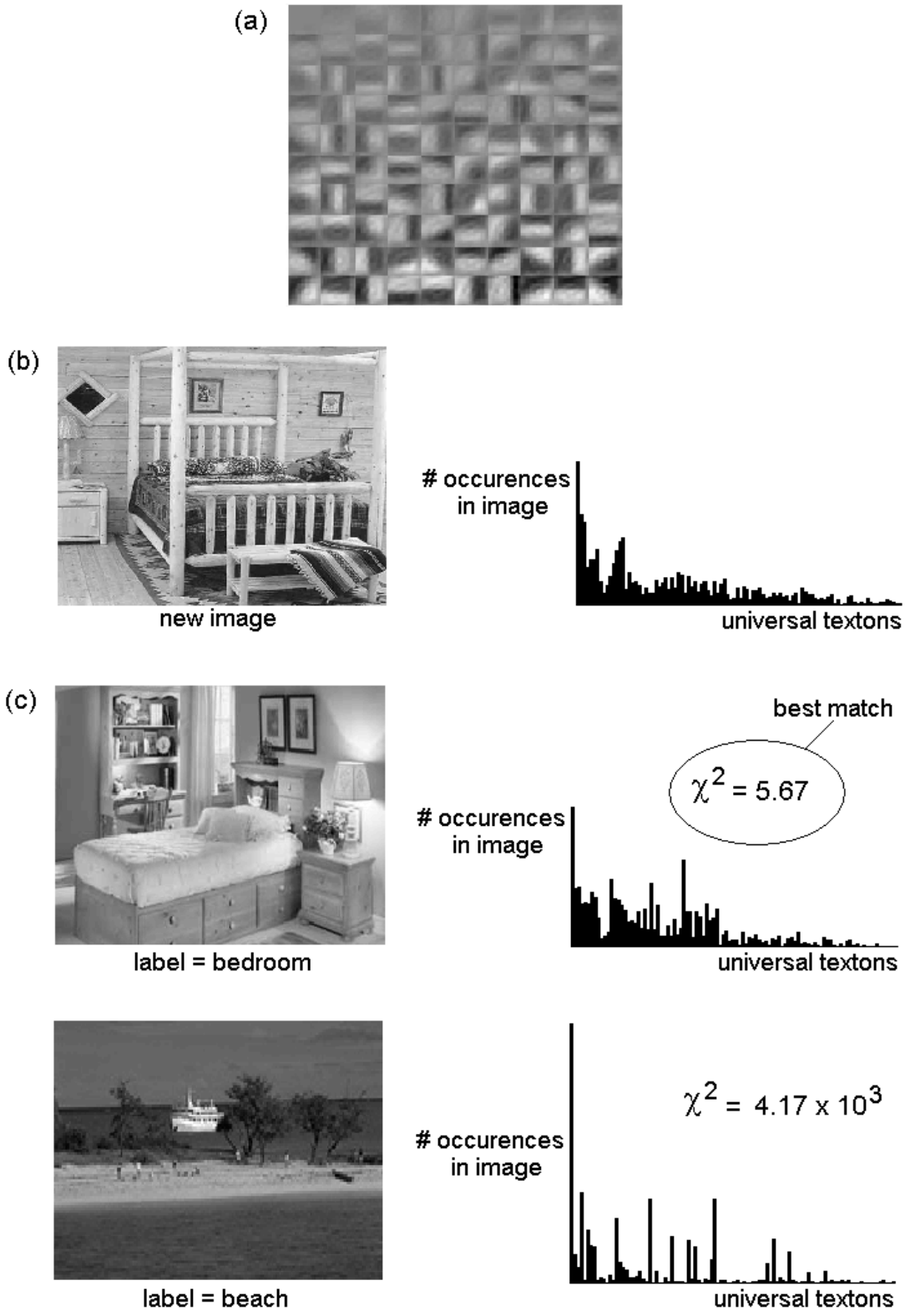
(a)
(b)
new image
occurences in image
universal textons
(c)
best match
$\chi^2 = 5.67$
occurences in image
universal textons
label = bedroom
$\chi^2 = 4.17 \times 10^3$
occurences in image
universal textons
label = beach

Figure 6.

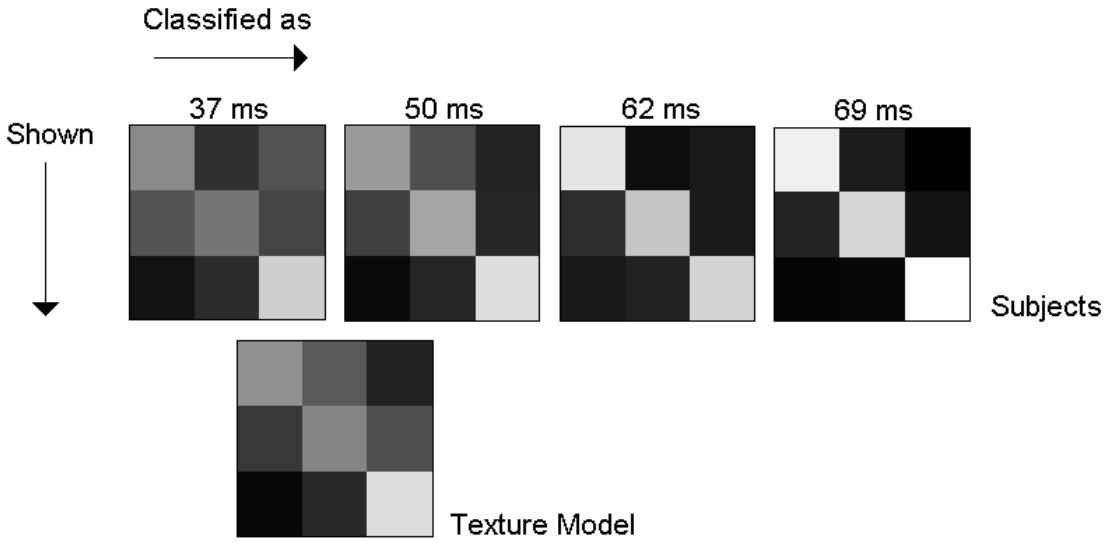
Classified as
37 ms
50 ms
62 ms
69 ms
Shown
Subjects
Texture Model

Figure 7.

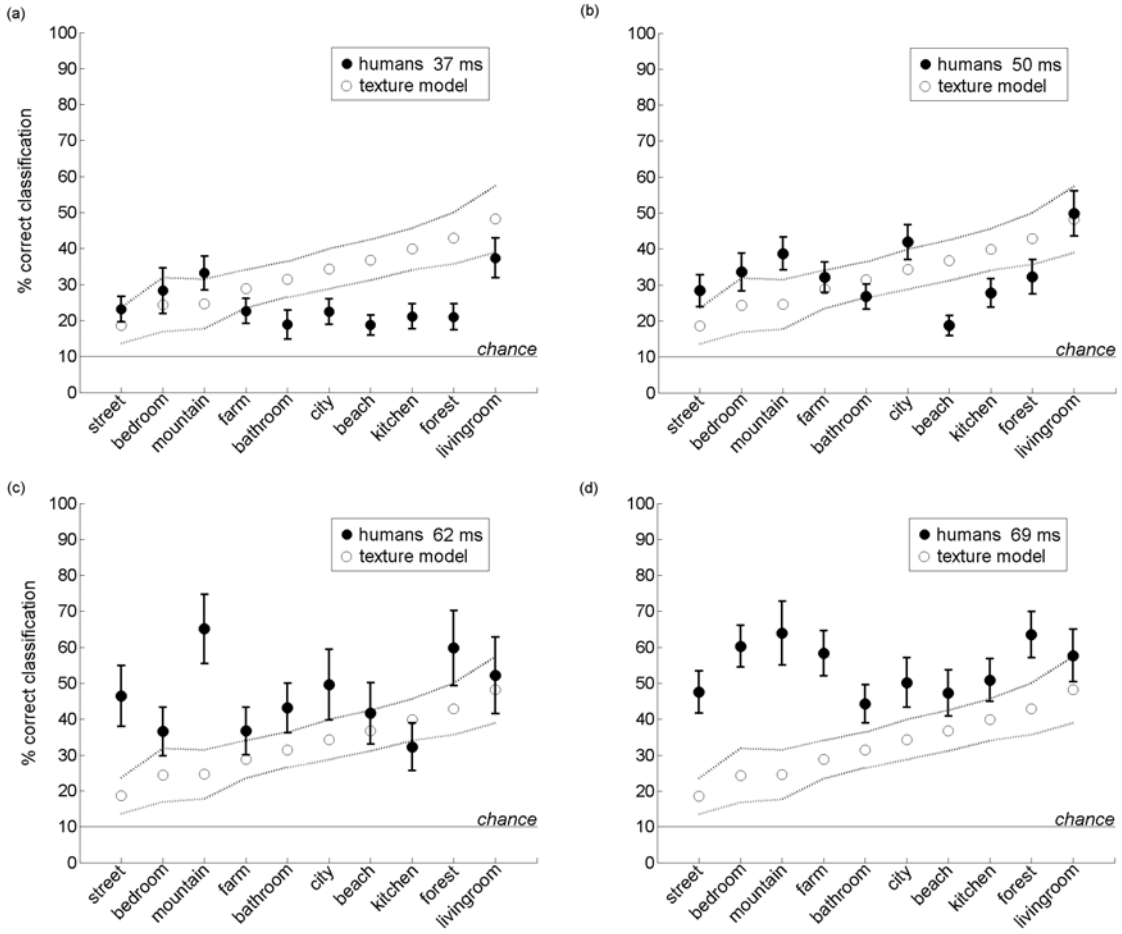
(a)
humans 37 ms
texture model
% correct classification
chance
(b)
humans 50 ms
texture model
chance
(c)
humans 62 ms
texture model
% correct classification
chance
(d)
humans 69 ms
texture model
chance
street
bedroom
mountain
farm
bathroom
city
beach
kitchen
forest
livingroom

Figure 8.

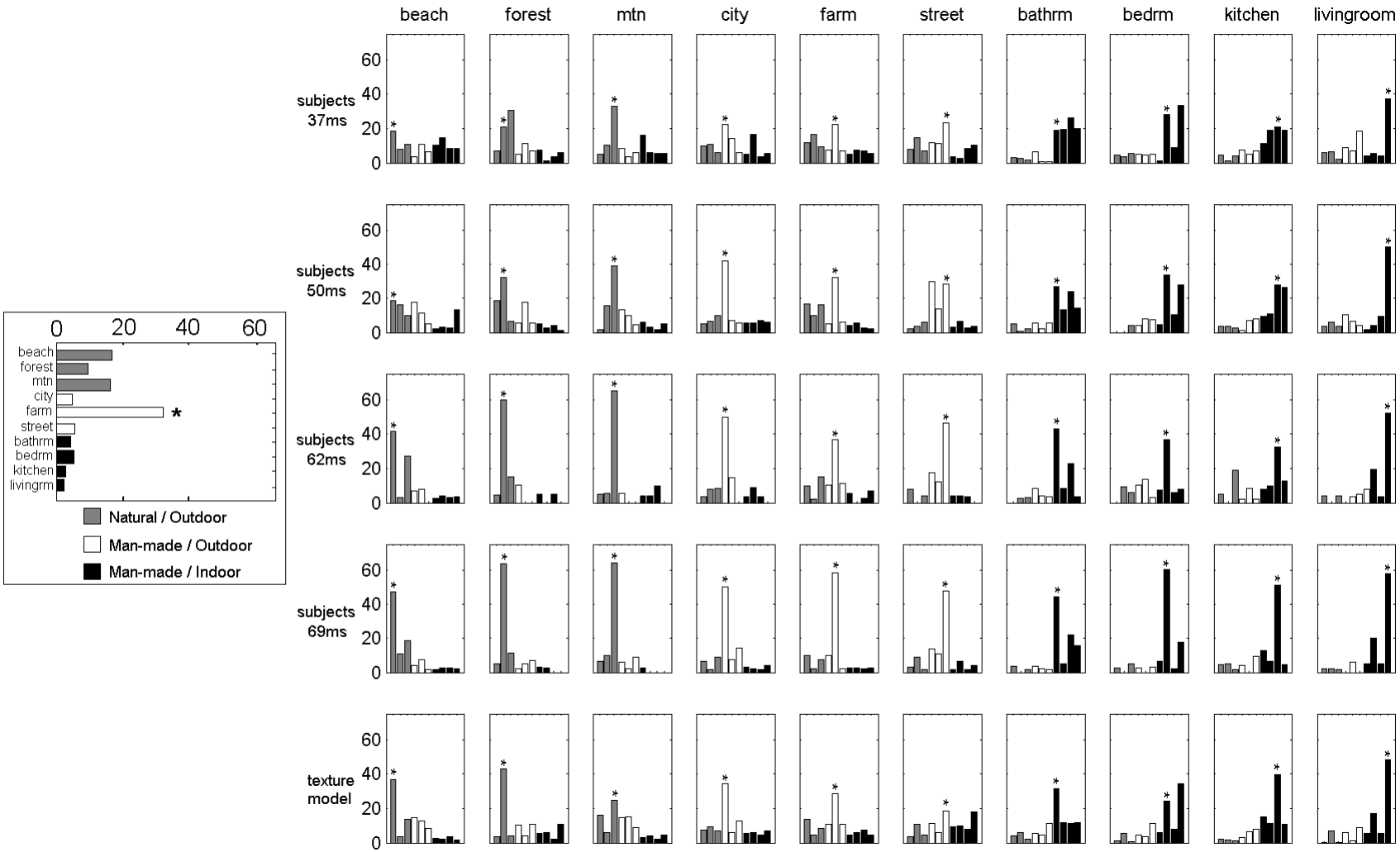
beach
forest
mtn
city
farm
street
bathrm
bedrm
kitchen
livingroom
subjects 37ms
subjects 50ms
subjects 62ms
subjects 69ms
texture model
0
20
40
60
Natural / Outdoor
Man-made / Outdoor
Man-made / Indoor

Table 1.

Scene shown	Other Word Choice	% Correct Measured
A	B	80.0
A	C	95.0
B	A	85.0
B	C	90.0
C	A	75.0
C	B	98.0

Table 2.

Scene shown	Classified as A	Classified as B	Classified as C
A	76.7	19.1	4.1
B	13.7	77.7	8.6
C	24.6	1.5	73.9

Table 3.

Scene Shown	Category Selected (%)									
	Beach	Forest	Mtn	City	Farm	Street	Bath	Bedrm	Kitchn	Livingrm
Subject Classifications at 37ms										
Beach	18.7±2.8	8.2±4.6	10.7±6.5	3.7±2.6	10.7±4.9	6.5±3.4	10.3±5.2	14.5±5.5	8.3±3.9	8.4±3.9
Forest	7.2±4.3	21.0±3.6	30.6±10.6	5.1±3.1	11.1±5.2	7.0±3.8	7.3±4.3	1.1±1.3	3.8±2.3	5.8±3.7
Mtn	4.9±3.6	10.6±5.3	33.2±4.7	8.5±4.9	3.5±2.7	5.8±3.6	16.1±7.1	6.2±4.0	5.7±4.2	5.6±3.3
City	10.0±5.2	10.9±5.4	5.8±2.7	22.5±3.6	14.0±6.4	6.0±3.5	5.1±3.2	16.5±8.3	3.8±2.6	5.4±3.1
Farm	11.8±5.4	16.8±7.2	9.4±4.8	7.4±4.4	22.6±3.5	7.1±3.5	5.1±3.3	7.6±3.7	6.8±3.3	5.4±3.4
Street	7.9±3.7	14.7±6.3	6.7±3.2	11.7±5.1	11.1±4.9	23.1±3.5	3.7±2.6	2.6±1.9	8.2±4.7	10.1±4.7
Bath	3.0±2.0	2.8±1.9	1.8±1.6	6.4±3.5	0.9±1.0	0.9±1.1	18.9±4.0	19.3±8.8	26.0±12.0	20.0±8.8
Bedrm	4.6±2.9	3.5±2.5	5.6±3.1	4.8±3.2	4.4±2.7	5.2±3.1	1.4±1.5	28.3±6.4	8.7±4.5	33.5±10.6
Kitchn	4.5±3.0	1.2±1.2	4.3±2.5	7.6±4.7	5.1±2.9	7.1±4.0	11.2±5.5	18.9±9.2	21.1±3.5	19.0±8.0
Livingrm	6.1±3.5	6.6±3.9	2.2±2.3	9.0±5.2	6.8±4.1	18.3±7.9	4.0±3.0	5.8±3.4	3.9±3.4	37.3±5.5
Subject Classifications at 50ms										
Beach	18.7±2.8	16.1±5.5	9.9±3.9	17.7±6.7	11.3±4.9	5.1±2.2	2.1±1.3	3.2±2.0	2.5±1.5	13.5±5.4
Forest	18.6±6.9	32.2±4.8	6.7±3.5	5.6±2.9	17.7±7.7	5.8±3.5	5.2±2.8	2.7±1.7	4.2±2.7	1.3±1.3
Mtn	1.7±1.5	15.9±6.3	38.7±4.6	13.1±5.0	9.8±4.5	4.6±2.9	6.2±3.4	3.1±2.0	1.7±1.6	4.9±2.9
City	5.0±3.2	6.7±3.8	9.9±5.1	41.9±4.8	7.0±3.7	5.8±3.9	5.4±3.4	5.5±3.1	7.0±3.9	5.8±3.0
Farm	16.7±5.7	9.9±4.8	16.4±6.1	4.8±2.7	32.1±4.2	5.9±3.2	3.9±2.4	5.5±3.6	2.6±2.1	2.2±1.8
Street	2.2±1.7	3.5±2.5	6.1±3.2	29.9±10.0	13.6±4.9	28.4±4.4	3.3±2.2	6.8±3.8	2.5±2.1	3.8±2.5
Bath	4.9±2.2	1.0±1.2	2.2±1.8	5.7±2.9	2.1±1.5	5.6±3.2	26.8±3.5	13.4±5.2	24.1±8.2	14.3±5.5
Bedrm	0.0±0.0	0.0±0.0	4.2±2.8	4.1±2.5	7.9±4.1	7.7±3.9	4.4±2.7	33.6±5.2	10.5±4.0	27.6±9.1
Kitchn	3.4±2.2	3.5±2.1	2.6±1.8	1.2±1.1	7.2±3.2	7.8±3.3	9.5±3.6	11.0±4.5	27.8±4.0	26.2±9.0
Livingrm	3.8±2.6	6.2±3.8	3.7±2.6	10.3±4.3	6.5±3.6	4.3±3.2	1.7±2.2	4.0±2.7	9.6±4.5	49.9±6.3
Subject Classifications at 62ms										
Beach	41.6±8.5	3.1±3.3	27.2±11.5	7.0±5.7	7.7±5.3	0.0±0.0	2.7±3.5	3.9±4.1	3.2±3.8	3.5±3.8
Forest	4.6±5.1	59.8±10.5	15.2±10.3	10.4±8.7	0.0±0.0	0.0±0.0	4.9±5.6	0.0±0.0	5.2±5.5	0.0±0.0
Mtn	5.1±5.5	5.6±5.5	65.1±9.6	5.7±5.3	0.0±0.0	0.0±0.0	4.3±4.4	4.3±5.3	10.0±8.5	0.0±0.0
City	3.4±3.8	8.0±6.8	8.3±6.9	49.6±9.9	14.7±9.0	0.0±0.0	3.7±3.7	8.8±7.4	3.6±3.9	0.0±0.0
Farm	9.7±6.3	2.2±3.1	15.1±9.0	10.2±6.7	36.7±6.6	11.3±7.6	5.3±4.9	0.0±0.0	2.7±3.2	6.8±6.1
Street	8.0±7.0	0.0±0.0	4.1±4.1	17.5±9.7	12.5±7.8	46.4±8.5	4.1±4.5	3.9±4.1	3.5±4.2	0.0±0.0
Bath	0.0±0.0	2.8±3.4	3.2±3.4	8.6±6.9	3.9±3.7	3.6±3.3	43.1±6.9	8.5±5.9	22.8±10.0	3.4±3.8
Bedrm	0.0±0.0	9.2±6.0	6.0±4.6	10.4±7.1	13.7±6.9	3.1±3.3	7.3±5.4	36.6±6.8	6.0±5.1	7.9±10.5
Kitchn	5.0±4.2	0.0±0.0	19.2±10.1	2.3±2.9	8.4±5.9	2.1±2.7	7.8±5.5	10.0±7.1	32.3±6.7	12.9±7.1
Livingrm	4.0±4.8	0.0±0.0	4.0±4.1	0.0±0.0	3.8±3.9	5.0±4.8	7.7±6.7	19.5±10.5	3.8±4.0	52.3±10.6
Subject Classifications at 69ms										
Beach	47.2±6.4	10.8±5.5	18.7±6.8	4.3±2.7	7.7±5.0	1.9±1.9	1.9±1.9	2.5±2.4	2.6±2.3	2.3±2.3
Forest	5.2±3.6	63.5±6.4	11.3±6.0	2.2±2.5	5.1±4.5	6.8±4.4	3.2±2.9	2.7±2.6	0.0±0.0	0.0±0.0
Mtn	6.5±5.3	10.0±5.5	64.0±8.9	5.9±5.2	2.3±2.5	8.7±5.2	2.6±2.7	0.0±0.0	0.0±0.0	0.0±0.0
City	6.5±3.5	1.6±2.1	8.9±4.7	50.2±6.9	7.4±4.4	14.4±7.4	3.3±2.6	2.0±2.0	1.7±1.9	4.0±3.1
Farm	10.0±5.0	2.3±2.5	7.3±4.5	9.9±4.6	58.4±6.3	2.1±2.2	2.5±3.0	2.5±2.8	2.4±2.4	2.6±2.8
Street	3.3±2.7	8.9±4.6	1.7±1.8	13.6±5.2	10.7±4.5	47.5±5.9	1.9±1.9	6.7±4.1	1.8±2.2	3.9±3.2
Bath	3.5±2.3	0.0±0.0	1.8±1.9	3.8±2.4	2.0±2.3	1.6±1.8	44.3±5.3	5.3±3.4	22.2±7.3	15.7±5.7
Bedrm	2.6±2.6	0.0±0.0	4.9±3.6	2.8±2.7	0.0±0.0	3.0±3.0	6.6±3.8	60.3±5.8	2.2±2.4	17.5±6.8
Kitchn	4.6±3.6	5.1±3.7	1.8±1.9	4.3±3.3	0.0±0.0	9.4±4.6	12.6±5.0	6.6±4.5	50.9±6.0	4.6±3.8
Livingrm	2.0±2.2	2.4±2.3	1.8±2.2	0.0±0.0	5.9±4.5	0.0±0.0	5.1±3.6	20.2±8.1	5.0±3.4	57.7±7.3
Texture Model Classifications										
Beach	36.8±5.6	3.6±2.3	13.7±4.3	14.9±3.8	12.6±3.3	8.4±4.0	2.7±1.6	2.3±1.7	3.6±2.0	1.4±1.7
Forest	3.7±2.3	42.9±7.1	4.1±2.7	10.1±6.1	4.3±1.6	10.7±4.3	5.5±3.2	5.9±2.9	2.0±1.6	10.9±3.2
Mtn	16.3±3.9	6.1±3.1	24.6±6.9	14.8±5.5	15.2±4.6	8.8±4.2	3.3±2.5	4.0±2.3	2.4±1.7	4.7±3.0
City	7.6±2.7	9.2±3.4	6.9±3.7	34.3±5.6	5.8±2.1	12.8±4.1	5.4±2.3	6.2±3.2	4.6±3.1	7.1±2.3
Farm	13.8±4.8	4.5±2.3	8.3±5.1	10.9±3.9	28.8±5.3	11.0±4.4	4.5±2.7	6.1±3.3	7.6±2.5	4.5±2.3
Street	3.8±2.2	10.6±3.1	4.7±2.5	11.4±4.9	6.0±2.2	18.6±5.0	9.4±3.8	9.7±3.2	8.1±3.0	17.8±4.7
Bath	4.1±1.7	5.9±3.9	2.1±1.5	5.3±3.0	4.6±3.3	11.5±4.3	31.5±4.9	11.9±4.7	11.3±3.4	11.8±5.3
Bedrm	1.0±1.1	5.7±1.6	0.9±1.1	4.8±2.3	3.6±1.9	11.1±4.3	5.8±2.8	24.4±7.5	8.0±3.0	34.6±9.0
Kitchn	2.4±2.2	1.7±2.1	1.2±1.1	3.0±2.0	6.4±2.5	8.1±3.3	15.3±6.2	11.3±4.8	39.9±5.8	10.7±5.0
Livingrm	0.1±0.4	6.9±3.0	0.4±0.9	5.9±4.5	1.2±1.6	9.1±3.5	5.5±3.2	17.0±5.6	5.5±3.0	48.2±9.2

Table 4.

Scene Shown	**Category Selected (%)**		
	Natural / Outdoor	Man-made / Outdoor	Man-made / Indoor
Subject Classifications at 37ms			
Natural / Outdoor	48.32±4.60	20.64±3.11	31.05±4.39
Man-made / Outdoor	31.33±3.98	41.84±3.62	26.82±3.72
Man-made / Indoor	11.58±2.22	19.12±3.09	69.30±3.15
Subject Classifications at 50ms			
Natural / Outdoor	52.88±3.81	30.27±4.29	16.85±2.93
Man-made / Outdoor	25.44±3.66	56.46±3.79	18.10±3.57
Man-made / Indoor	8.84±1.87	17.54±2.67	73.62±3.03
Subject Classifications at 62ms			
Natural / Outdoor	75.71±5.45	10.28±4.05	14.00±4.57
Man-made / Outdoor	19.58±4.95	66.31±5.95	14.11±4.76
Man-made / Indoor	13.33±3.98	16.20±3.95	70.48±4.87
Subject Classifications at 69ms			
Natural / Outdoor	79.07±4.14	15.00±3.98	5.94±2.09
Man-made / Outdoor	16.85±3.37	71.37±4.00	11.77±2.39
Man-made / Indoor	7.62±1.92	8.19±2.37	84.18±2.84
Texture Model Classifications			
Natural / Outdoor	50.56±4.25	33.26±4.20	16.19±2.52
Man-made / Outdoor	23.14±3.15	46.54±3.38	30.33±3.26
Man-made / Indoor	8.12±1.51	18.69±3.32	73.18±3.10