Estimating perception of scene layout properties from global image features

Michael G. Ross

Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology, USA

Aude Oliva

Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology, USA

The relationship between image features and scene structure is central to the study of human visual perception and computer vision, but many of the specifics of real-world layout perception remain unknown. We do not know which image features are relevant to perceiving layout properties, or whether those features provide the same information for every type of image. Furthermore, we do not know the spatial resolutions required for perceiving different properties. This paper describes an experiment and a computational model that provides new insights on these issues. Humans perceive the global spatial layout properties such as *dominant depth*, *openness*, and *perspective*, from a single image. This work describes an algorithm that reliably predicts human layout judgments. This model's predictions are general, not specific to the observers it trained on. Analysis reveals that the optimal spatial resolutions for determining layout vary with the content of the space and the property being estimated. Openness is best estimated at high resolution, depth is best estimated at medium resolution, and perspective is best estimated at low resolution. Given the reliability and simplicity of estimating the global layout of real-world environments, this model could help resolve perceptual ambiguities encountered by more detailed scene reconstruction schemas.

Keywords: space and scene perception, computational modeling, depth, structure of natural images

Citation: Ross, M. G., & Oliva, A. (2010). Estimating perception of scene layout properties from global image features. *Journal of Vision, 10*(1):2, 1–25, http://journalofvision.org/10/1/2/, doi:10.1167/10.1.2.

Introduction

Understanding the visual cues and computations underlying the perception of complex environments is a primary concern of the study of human visual perception. Humans constantly engage in automatic and rapid analysis of scene structure when navigating an environment or searching for objects. Research in human perception has shown that many global properties of a scene are discerned at an early stage of visual processing (e.g. *coarse spatial layout*, Schyns & Oliva, 1994; *naturalness*, Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007; *navigability*, Greene & Oliva, 2009a; *complexity*, Sanocki & Sulman, 2009), and that brief exposure to a specific scene layout facilitates distance perception (Sanocki, 2003; Sanocki & Epstein, 1997). Recent work has also shown that global properties describing the three-dimensional layout of a scene, such as the *dominant depth*, *openness*, or *perspective* of an environment, can be perceived at the very beginning of a glance, and can influence scene categorization (Greene & Oliva, 2009a, 2009b).

Similarly, computational research in scene recognition has succeeded in developing algorithms for semantic scene classification that rely on low-level image features and not on explicit object recognition or segmentation (Fei-Fei & Perona, 2005; Oliva & Torralba, 2001, 2002, 2006; Renninger & Malik, 2004; Torralba & Oliva, 2002, 2003; Vailaya, Jain, & Zhang, 1998; Vogel & Schiele, 2007). Many semantic scene categories, such as *street*, *desert*, or *forest* are partially defined by their spatial layout. However, many of the specifics of layout perception remain unknown, including: which image features are relevant to different layout properties, whether those features are universal or contextual, and what spatial resolution is necessary for the perception of different properties. This paper describes an experiment and a computational model that provide significant new insights on all of these issues.

The determination of scene layout information from a single image has received a great deal of attention over the past 30 years in both computer vision and human psychophysics. Many of the computational methods for recovering layout from non-artificial images focus on relative depth information: shape from shading (Horn & Brooks, 1989); texture gradients (Super & Bovik, 1995); edges and junctions (Barrow & Tenenbaum, 1981), as well as other pictorial cues such as occlusion, relative size, and elevation with respect to the horizon line (for a review, Palmer, 1999). A supplementary source of layout information can be recovered by detecting vanishing points—the apparent intersection of parallel lines produced by perspective projection (Criminisi, Reid, & Zisserman, 2000; Magee & Aggarwal, 1984). More

doi: 10.1167/10.1.2 Received July 10, 2009; published January 8, 2010 ISSN 1534-7362

recently, Divvala, Efros, and Hebert (2008), Hoeim, Efros, and Hebert (2007), Saxena, Sun, and Ng (2009), and Yu, Zhang, and Malik (2008) attempt to infer similar 3D models of scene surfaces by using image segmentation to extract regions that are large enough to provide useful texture, color, perspective or shape cues. The goal of these methods is to characterize the variability of depth across different image locations. The algorithms provide varying degrees of precision, but at the maximum (Horn & Brooks, 1989; Saxena et al., 2009) they create a dense depth map from a single photograph. Unfortunately, no known method is general enough to provide accurate scene information regardless of scene content, camera position, or lighting condition.

Human layout perception is partially due to inter-ocular stereo fusion, but those cues are only effectual at short range and humans easily perceive depth when they are absent—such as when looking at photographs. One hypothesis is that layout perception results from a combination of local cues extracted from parts of a two-dimensional image and then resolved into a consistent overall three-dimensional scene structure. Occlusion and relative position of features with respect to the ground plane permit the segmentation of images into "figure" and "background" regions. Blur information can convey the coarse layout of the scene (Schyns & Oliva, 1994), which, in turn, provides information about the probable location of the surface in depth, and even the size of the observed space (Held & Banks, 2008). Many researchers have observed that human-perceived layout and depth information is not always accurate (Howe & Purves, 2005; Watt, Akeley, Ernst, & Banks, 2005). Surface angles are often underestimated (Girshick, Burge, Erlikhman, & Banks, 2008), and slants of hills are often overestimated (Creem-Regehr, Gooch, Sahm, & Thompson, 2004; Proffitt, Bhalla, Gossweiler, & Midgett, 1995). Distances to objects can be misperceived when a relatively wide expense of the ground surface is not visible (Wu, Ooi, & He, 2004), or when the field of view is too narrow (Fortenbaugh, Hicks, Hao, & Turano, 2007). Although most studies found systematic and consistent biases in how humans estimate object distances and surface orientations, a systematic evaluation of humans' consistency in perceiving scene layout is missing from the literature.

A complementary approach, which we adopt in the current work, is to ignore local variations within an image and instead characterize an environment's global layout properties. For example, instead of detecting that the ground plane at the bottom of an image is near, while the mountain at the top of that image is distant, these methods would reveal that most of the scene elements in this image are distant compared to those in a close-up image of a flower. They could also reveal that one street scene is parallel to the camera, while another is perpendicular, and that the view from a mountaintop is open to the sky, while a forest environment is closed. This type of categorization can be based on the statistical regularities of features and their spatial distribution within environments of similar physical size and shape (Oliva & Torralba, 2001; Torralba & Oliva, 2002, 2003). In fact, three-dimensional spatial and content properties of scenes, like the degree of *openness*, *perspective*, *roughness*, or *naturalness* of an environment, are indicated by corresponding low-level information in a two-dimensional image (e.g., a coast is an "open" environment, characterized by a long horizon line in the middle of the image).

In natural images, surfaces with convex contours are statistically more likely to be nearer to the viewer than those with concave contours (Burge, Fowlkes, & Banks, submitted for publication), and even simple luminance and edge information can be very informative of the overall scene roughness and of the object distances (Yang & Purves, 2003). A variety of image-based cues are correlated with the scale of visual environments and, as scale increases, observer viewpoint is more constrained (Coughlan & Yuille, 1999). Similarly, the building blocks of an environment differ from one scale to another given the functional constraints (e.g. a closet is for small objects, a garage is for large objects), as do the physical processes that shape the space at each scale, particularly for outdoor natural environments. Torralba and Oliva (2002, 2003) showed that visual features of natural images are strongly scale-dependent, which allows a simple model based on the output magnitudes of a bank of localized multiscale oriented filters to determine the absolute depth range of a given scene image. The types of components and materials in a scene (the *content* of the space) can also influence the relationships between image features and scene layouts. For example, Torralba and Oliva (2002) found that the diagnostic features of scale were strikingly different between man-made and natural environments. While man-made scene surfaces become smaller and more heterogeneous with increased viewing distance (i.e. from a door, porch, or building to a city view, the global roughness of the image features increases), natural scene surfaces become larger and more homogeneous (i.e. natural structures become larger and smoother and the horizon line becomes more apparent, breaking the viewed space into distinctive foreground and background regions).

The optimal spatial resolution for determining various spatial layout properties is an open question. As shown by research in visual cognition, the selection of the optimal feature or best spatial resolution in a given image depends on the task to be solved (Gosselin & Schyns, 2001; Oliva & Schyns, 1997; Schyns & Oliva, 1999) and the types of images. For natural scene images, different regions of the phase spectra (McCotter, Gosselin, Sowden, & Schyns, 2005), as well as different spectral magnitude signatures (Torralba & Oliva, 2003) are associated with different basic-level semantic categories (such as forests,

mountains, street), or different image structure properties (Baddeley, 1997). Similarly, Torralba (2009) found that the spatial resolution at which exemplars from a scene category are recognized vary if the scene is an indoor, outdoor man-made, or outdoor natural scene.

Some tasks and image types are well served by features that are spatially invariant and encoded without reference to image location. This is commonly seen in work involving textures (Heeger & Bergen, 1995; Portilla & Simoncelli, 2000) and in the case of natural images with image-spanning homogeneous surfaces (e.g. foliages, forests, and ground views; Field, 1987) or properties related to the fractal dimension of the image (e.g. *roughness* of an image; Heaps & Handel, 1999; Oliva & Torralba, 2001; Pentland, 1984). Natural and man-made environments can be distinguished with high accuracy using spatially invariant features (Torralba & Oliva, 2002; Vailaya et al., 1998). However, the components and surfaces that create the spatial layout of a scene often vary significantly with image location. For example, *open* scenes are characterized by the absence of texture at the top part of the image, whereas *closed* scenes are characterized by the presence of texture across the entire image. Oliva and Torralba (2001) observed that the correlation between human ratings of scenes along the properties of *openness* and *perspective* and a linear discriminant model were higher when the model used a feature representation based on a 4 × 4 spatial grid (2 cycles/image) instead of a spatially invariant encoding.

In this paper, we gain insight into the image features and the spatial resolution that are correlated with the human perception of the global layout of visual environments. Instead of trying to construct detailed three-dimensional scene models or to recover "ground truth" information about the specific locations of objects and surfaces, we focus on recovering global scene layout properties that correspond to dimensions of interest for human perception. Humans can easily group collections of scenes into categories that, although semantically different, share similar dominant depth, openness, perspective, and other properties. The model developed in this paper can successfully translate an image into co-ordinates in the space of perceptual physical scene layout.

The approach presented here differs from previous work in several ways: most notably, it focuses on estimating *perceptual* ground truth rather than *physical* ground truth and it investigates the *spatial resolution* at which image features best represent layout information. By training and testing the model on a large database of human scene layout ratings, we determine the types of image information that best predict the *depth*, *openness*, and *perspective* of a given scene. The results demonstrate that the model replicates a substantial portion of the relationship between image features and human perception. Furthermore, its predictions are general and not specific to the observers it was trained on. Analyzing the model reveals that openness is best estimated at a high spatial resolution, dominant depth is best estimated at a medium spatial resolution, and perspective is best estimated at a low spatial resolution. In many cases, the model's performance is not significantly different from human performance, therefore analyzing its structure and behavior may provide insights for understanding human layout perception or designing new experiments. Not only are these results valuable to furthering our understanding of human scene layout perception, they are relevant to a rich set of computer vision applications, especially in image search and visual navigation.

Human rating experiment

Our goal was to gather human ratings of the dominant depth, openness, and perspective of a large, diverse collection of outdoor images. This database was used to train and test computational models for automatically determining these properties. In order to gather the perceptual ground truth on the three properties, we conducted an experiment in which human observers rated the dominant depth, perspective, and openness of thousands of unique images of natural and human-made outdoor environments (see Figure 1).

While depth is often considered a local property of a given surface, here we refer to depth in a global sense, thereby referencing the size of the space in a scene (e.g. the mean distance between the observer and the boundaries of this space, e.g. Torralba & Oliva, 2002). While dominant depth is not a precisely defined quantity, it has a strong relationship with the physical size of the space, and human judgments are consistent in evaluating this quantity (Torralba & Oliva, 2002).

Openness of a scene refers to the quantity and location of boundary elements of the scene in view. The most open scene is a ground surface stretching to the horizon, with the existence of a horizon line in the absence of any other visual references (e.g. trees, buildings, Gibson, 1986). At the other extreme, a closed scene is composed of surfaces covering the full field of view.

Perspective refers to the degree of expansion of a space. The convergence of parallel lines to a visible vanishing point gives a strong perception of depth gradient to the space represented in an image. However, a flat view on a row of trees, a background mountain, or a building would have no perspective because the scene lacks salient parallel lines or the vanishing point is perpendicular to the camera's direction (a situation also denoted as a vanishing point "at infinity"). Using different sets of images and tasks, previous work has shown that these three properties can be reliably estimated by human observers (Greene & Oliva, 2009b; Oliva & Torralba, 2001; Torralba & Oliva, 2002).

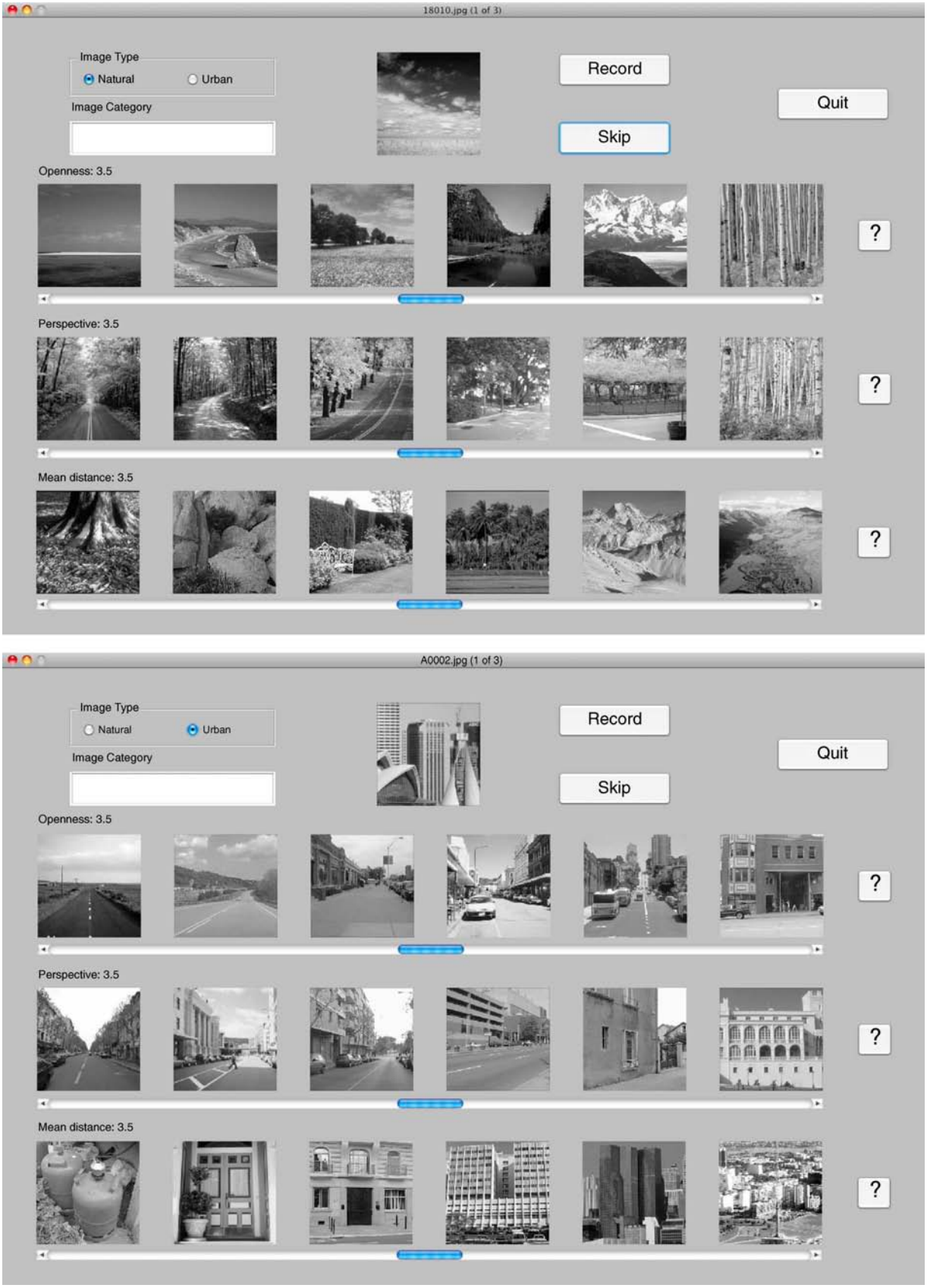


Figure 1. Top: an example display from one trial of the experiment. The sliders are used to indicate the rating of the target image at the top; the "?" buttons are used to indicate an ambiguous layout. Bottom: a similar example display, but with the urban layout prototypes displayed.

Stimuli

A total of 7,138 unique grayscale images of outdoor scenes, representing a variety of outdoor environments, were used. The images were selected from the database used in Oliva and Torralba (2001) and from a larger database containing approximately 40,000 images of outdoor environments collected from a variety of sources, including personal photographs, various Internet search engines, and photography websites. The examples in the Oliva and Torralba (2001) database were originally selected to be prototypical examples of spatial envelope axes. The examples selected from the larger database were chosen at random, but images that did not represent outdoor scenes or were not photographed from a standing position were manually excluded.

Participants

Fourteen participants from the MIT community and others took part in the rating experiment. All participants were between 18 to 40 years old. They all gave informed consent and received monetary compensation of $20/hour. Each observer rated between 300 and 1,300 images. Both authors participated in the experiment, and rated 1,026 and 850 trials respectively. In total, 7,138 unique images were rated by a human observer, and these ratings were used to train and test the models presented in this paper. It would be unreasonable to expect a computer algorithm to predict human ratings better than another human can. Therefore, human rating consistency is a useful performance benchmark. To measure consistency, 838 of the images were rated a second time (in all but 15 cases, the second rating was provided by a different observer than the first rating). The double-rated images were a random subset of the complete set of images.

Procedure

For each image, each of the three properties was rated on a continuous 1–6 scale and observers were provided with prototypical images for the integer values. Figure 1 shows computer displays from two representative trials: participants were shown one target image representing a view of a standing observer on the ground, and asked to rate its degree of perspective, openness and dominant depth on a scale from 1 to 6, as if they were at the scene and observing it from the specific viewpoint represented by the image. We instructed participants to judge the overall dominant depth based on the average depth of the scene across image pixels representing objects or landforms, excluding the depth associated with pixels representing the sky, clouds, or sun ("1" represented near and "6" represented far). We instructed them to judge perspective by estimating the angle between the camera and the perceptually dominant vanishing points in the image, ignoring if the angle was to the left or to the right ("1" represented perpendicular and "6" represented parallel). We instructed them to judge openness by the amount of unobstructed sky in the image and the location of the horizon line ("1" represented open and "6" represented closed).

At the beginning of each trial, observers were shown one randomly selected target image, were asked to provide a one or two-word description of the image (typically, its semantic category), and were required to select if the image was "natural" (primarily consisting of natural objects) or "urban" (primarily consisting of manufactured objects). Each layout rating was represented by a slider bar and a numeric indication of the current rating. At the beginning of each trial, all sliders were set to the middle position, representing rating "3.5" for each property. Above each property slider were a set of six prototype images representing examples of the type of layout associated with the 6 possible integer rating values. There were "natural" and "urban" sets of prototype images and the appropriate set was displayed based on the observer's classification of the target image. Observers were instructed to choose the ratings by sliding the bars to positions that best described the layout properties of the target image. Ratings were recorded as real-valued numbers and observers were free to choose a slider position between two prototypes if they felt that position best represented the image layout. Observers were instructed to choose the rating that best characterized the dominant layout of the scene. For example, a target image might contain objects at a variety of depths, but if the vast majority of pixels are associated with distant objects, they should control the depth rating. The observers were told they could skip rating of a property for a given image if they felt it was ambiguous. The interface provided a "?" button next to each property slider for that purpose.

As described previously, approximately 12% of the images were rated twice and these were used as a performance baseline to compare our model's error rate to human perceptual variability. An appropriate computer model of perceptual spatial layout should achieve a rating prediction error similar to the expected rating variance between two human observers. The inter-observer variance indicates the precision of the perceptual ground truth. A model with lower error is probably over-fitting the data: It is providing information about a specific human rater or a specific type of image, rather than computing generally useful information.

Modeling spatial layout

Oliva and Torralba (2001) have shown that a meaningful scene categorization of outdoor environments can be made in a feature space consisting of the local responses

to frequency-tuned filters of different scales and orientations. In the computer vision literature, this feature space is referred as the GIST descriptor. Previous work by Torralba and Oliva (2002) demonstrated that the GIST features could be used to estimate *physical depth* of natural images using cluster-weighted models (CWMs) (Gershenfeld, 1999). CWMs are a generalization of Gaussian mixture models to linear regression. This work applies CWMs to predicting the *perceptual dominant depth*, *openness*, and *perspective* of real world scene images.

GIST features: A low-dimensional representation of image structure

To compute the GIST features, we used the MATLAB code used in Oliva and Torralba (2001).[1] This algorithm transforms a grayscale image into a collection of feature images by convolving it with a set of Gabor-like filters. The filtered images are each divided into an R by R grid and then summarized by the average complex magnitudes in each grid square. Full details about the GIST features are available in Oliva and Torralba (2001). This representation provides strong information about the structure of the scene, but not about specific objects. For the work presented in this paper, we used GIST features computed from Gabor filters with 8 orientations at 4 scales.

The optimal spatial resolution (the size of R) for estimating each layout property is an open question. Most GIST-based algorithms and analyses have used a resolution of 4 × 4 (or 2 cycles/image). The utility of a finer representation of spatial information in layout perception can be tested by using higher resolutions which represent the average filter responses in more localized image regions. In this work, we varied the grid size (R = 1, 2, 4, 8, 16) to determine the optimal spatial resolution for estimating the perceptual dominant depth, openness, and perspective of outdoor scene images.

Learning algorithm: Cluster-weighted models of scene layout properties

The model introduced in this section is designed to learn the relationship between the image structures (defined by the GIST feature at various resolutions) and the three layout properties. For instance, for dominant depth, the system has to learn that, in the case of a natural environment, long horizontal and oblique contours probably correspond to a large-scale environment, whereas the presence of fine-grained texture all over the image probably indicates a medium-sized environment. We want to determine depth, openness, and perspective estimators which can predict human perceptual judgment of these properties on real-world outdoor images.

Regression is the prediction of one variable's value from the known values of other variables. There are a wide variety of regression models in the scientific literature. In a standard linear regression model, the property of interest is computed with a weighted sum of feature values. For example, a linear regression model of depth might consist of positive weights on high frequency features and negative weights on low frequency features. This model would predict that images containing an abundance of high frequencies are far and images containing an abundance of low frequencies are near. Linear models are valuable because they are simple to fit to data and efficient to compute, but they do not always provide the most accurate results. For example, what if high frequencies are associated with deep city scenes, but shallow forest scenes? This type of context-dependent prediction cannot be described by a single linear regression function, but it could be described by fitting separate regression functions to each scene category. CWMs can be trained to automatically find the regression functions appropriate for each context and achieve more accurate results than simple linear models.

A CWM contains N clusters in a feature space and each cluster is associated with a linear regression function. Regression is performed on a new example by a mixture of all the model's regression functions. The mixture proportions are determined by the CWM's mixture coefficients and the conditional probability of the example's features under the Gaussian distribution representing each cluster c_i. Therefore, the joint probability density of a scene property rating, r_j, and the image feature vector, f_j, is:

$$p(r_j,f_j) \propto \sum_{i=1}^{N} p(c_i)p(f_j|c_i)p(r_j|f_j,c_i), \tag{1}$$

in which

$$p(f_j|c_i) \propto \exp\left(-\frac{(f_j-\mu_i)^T\sum_i^{-1}(f_j-\mu_i)}{2}\right), \tag{2}$$

and

$$p(r_j|f_j,c_i) \propto \exp\left(-\frac{(r_j-\omega_i^T f_j^*)^2}{2\sigma_i^2}\right), \tag{3}$$

where f_j^* indicates the original feature vector f_j with a 1 concatenated to its end. Given an image's features, f_j, the estimated rating that minimizes expected squared error under the model is

$$\hat{r}_j = \frac{\sum_{i=1}^{N}\omega_i^T f_j^* p(f_j|c_i)p(c_i)}{\sum_{i=1}^{N}p(f_j|c_i)p(c_i)}. \tag{4}$$

Given N, the model can be fit to D data samples by choosing random initial parameter values and using the following expectation-maximization (EM) update equations:

$$p(c_i|r_j,f_j) = \frac{p(c_i)p(f_j|c_i)p(r_j|f_j,c_i)}{\sum_{i=1}^{N} p(c_i)p(f_j|c_i)p(r_j|f_j,c_i)}, \quad (5)$$

$$p(c_i)' = \frac{\sum_{j=1}^{D} p(c_i|r_j,f_j)}{\sum_{i=1}^{N}\sum_{j=1}^{D} p(c_i|r_j,f_j)}, \quad (6)$$

$$\mu_i' = \frac{\sum_{j=1}^{D} f_j p(c_i|r_j,f_j)}{\sum_{j=1}^{D} p(c_i|r_j,f_j)} \equiv \langle f \rangle_i, \quad (7)$$

$$\sum_i' = \left\langle (f-\mu_i')(f-\mu_i')^T \right\rangle_i, \quad (8)$$

$$\omega_i' = \left(\langle f^* f^{*T} \rangle_i \right)^{-1} \langle rf \rangle_i, \quad (9)$$

$$\sigma_i'^2 = \left\langle (r - \omega_i^T f^*)^2 \right\rangle_i, \quad (10)$$

where the $\langle\rangle_i$ expectation operation is defined as in the μ_i' update equation. For further details and derivations read Gershenfeld (1999).

For the results reported in this paper, we initialized $p(c_i)$ to a uniform distribution, and the Σ_i and σ_i^2 parameters with the variances of the features and ratings, respectively. The μ_i were initialized to the values of N randomly chosen f_j vectors from the training data, and the σ_i were initialized to all zeros, except for the last (bias) term which was set to predict the rating associated with the initial μ_i. To find the best model for a given N and avoid local minima we re-ran the EM procedure 20 times and chose the result with the maximum log likelihood. The best N was chosen by five-fold cross validation on the training data, searching from 1 to 10 clusters and selecting the N that produced the minimum total mean-squared error on the held-out segments.

Determining the optimal spatial resolutions

We fit CWM models to the human experimental data for three purposes: to determine the optimal spatial resolution for predicting each spatial layout property, to compare the CWM models' performance to human performance, and to determine if the models' predictions would generalize to new observers who were not included in the training data.

For the purpose of determining the optimal spatial resolutions, the 7,138 images were randomly divided into five non-overlapping sets. The duplicate ratings were discarded, so each data point consisted of a unique image and at most one rating for each scene layout category. Five-way cross validation was performed with each set held out as test data in turn.[2]

For each cross-validation split, CWMs were fit to the data in four of the sets and tested on the held-out images in the remaining set. Each layout property was predicted by an independently trained model. Because the division between natural and urban environments is a primary axis of image categorization (Rogowitz, Frese, Smith, Bouman, & Kalin, 1998) and can be automatically determined, the models were trained and tested on three different versions of the data—one that contained all the images, one that only contained the natural images and one that only contained the urban images.[3] Approximately 55% of the training and test data sets consisted of natural images. If human observers did not rate one or more layout properties of a particular image, it was excluded from the test and training sets for only that property.

In training, CWMs were fit using five spatial resolutions of the GIST features: 1×1, 2×2, 4×4, 8×8, and 16×16. For each resolution we used the same 8 orientations and 4 scales of Gabor filters. Principal components analysis (PCA) was used to reduce each type of feature to 24 dimensions. The principal components were computed from 2,000 images that were not used in any training or testing sets. The number of components was chosen based on Torralba and Oliva's (2002) previous experiments with depth estimation using GIST features.[4] Keeping the number of components fixed as the spatial resolution increased helps to avoid the potential need for exponentially more training data at each step due to the increase in feature dimensionality.[5] While our 1×1 GIST representation only has 32 dimensions, 2×2 has 128 dimensions, 4×4 has 512 dimensions, 8×8 has 2,048 dimensions, and 16×16 has 8,192 dimensions. Using a constant

	All	Natural	Urban
Depth	4 × 4	8 × 8	4 × 4
Openness	8 × 8	4 × 4	4 × 4
Perspective	2 × 2	2 × 2	4 × 4

Table 1. The optimal spatial resolutions for computing each layout property with CWMs. Across the five-fold cross validation, each of these models performed significantly better than the CWMs at all lower resolutions (one-sided *t*-test, $p < 0.05$).

number of features helps to ensure that the model generalizes well using the same number of training examples across all resolutions. However, it also means that increasing the spatial resolution available to the model decreases the available frequency information. Therefore, instead of viewing the increasing spatial resolution as providing strictly more information to the CWMs, it is useful to consider it as a control balancing the amount of frequency information versus the amount of spatial information. The 24 dimensions used to represent the 1 × 1 features contain only frequency information about the images, and as the resolution increases some of that information is displaced by spatial information.

The training procedure described in the previous section lead to one CWM for each combination of feature type, layout property, and data set, using only the training data. After training was complete, the models were evaluated on the held-out data to determine which spatial resolution was best suited for predicting each property on each data set.

The cross-validation provides several quantities of interest. First of all, for the remaining experiments that are concerned with measuring the performance of the CWMs for this task, we wish to know which resolution provides the best performance for each estimation problem. To compute the optimal resolutions, we compute the mean squared errors observed for each resolution across all the training/testing splits. We wish to determine the minimum resolution required to produce optimal results for each property. Those resolutions, reported in Table 1, were computed for each property and data set by looking at the mean squared errors recorded for each resolution and using paired one-sided *t*-tests ($p < 0.05$) to determine if they were significantly better than the errors produced by the models trained at all lower resolutions.[6] For example, we would declare a 4 × 4 model optimal if its performance was significantly better than the 1 × 1 and 2 × 2 models' performances and it was not significantly outperformed by the 8 × 8 or 16 × 16 models. For more detail about optimal resolutions, please see Appendix A. It is also interesting to compute the average number of clusters used in each model at the optimal resolutions (Table 2) because it reveals how much context-sensitivity (compared to a linear regression function, which is equivalent to a one-cluster CWM) is necessary to optimally compute each layout property.

Results and discussion

The performance data reveals that the optimal spatial resolution is different for the three layout properties and also depends on whether the image depicts a natural or urban environment.[7] For urban environments, the three layout properties share the same optimal resolution, 4 × 4 (i.e. 2 cycles/image). On the other hand, for natural environments, estimating the dominant depth requires the most spatial resolution (8 × 8), perspective requires the least (2 × 2), and openness falls somewhere in between (4 × 4). Overall, estimating perspective requires the least spatial resolution because perspective is only salient when there are lines stretching across an entire image to indicate the location of vanishing points. It is difficult to gather meaningful data about perspective from a small image patch and it depends more on orientation than location. Openness, on the other hand, is dependent on the location of sky-boundary patches and the location of the horizon line, and depth, especially in complex scenes composed of near objects and distant backgrounds, can depend on a mix of high and low-spatial-resolution cues.

Although the trends in the number of clusters used by the models (Table 2) are less clear, depth and openness estimation appear to generally require more clusters than perspective estimation. This could indicate that perspective regression functions are more generic, or it could be a consequence of the lower precision of perspective perception, which is discussed in the next section.

Comparison to human variance and Mean Model predictions

To measure the quality of the CWM predictions, we compared the accuracy of its predictions to the predictions of three other models: the ratings of another human observer on the same test images, the ratings of the Mean Model (MM), and the ratings of a Linear Model (LM). Figure 2 demonstrates these comparisons using real data on an example image. All error measurements are represented by the squared difference between the model's rating and the rating provided by a human observer. Squared error is the most commonly used error measurement in regression problems.

	All	Natural	Urban
Depth	6.2	3.0	4.6
Openness	6.2	4.6	3.4
Perspective	4.0	1.2	3.4

Table 2. The mean number of clusters used in the optimal resolution CWMs—averages computed across the five-fold cross-validation procedure.

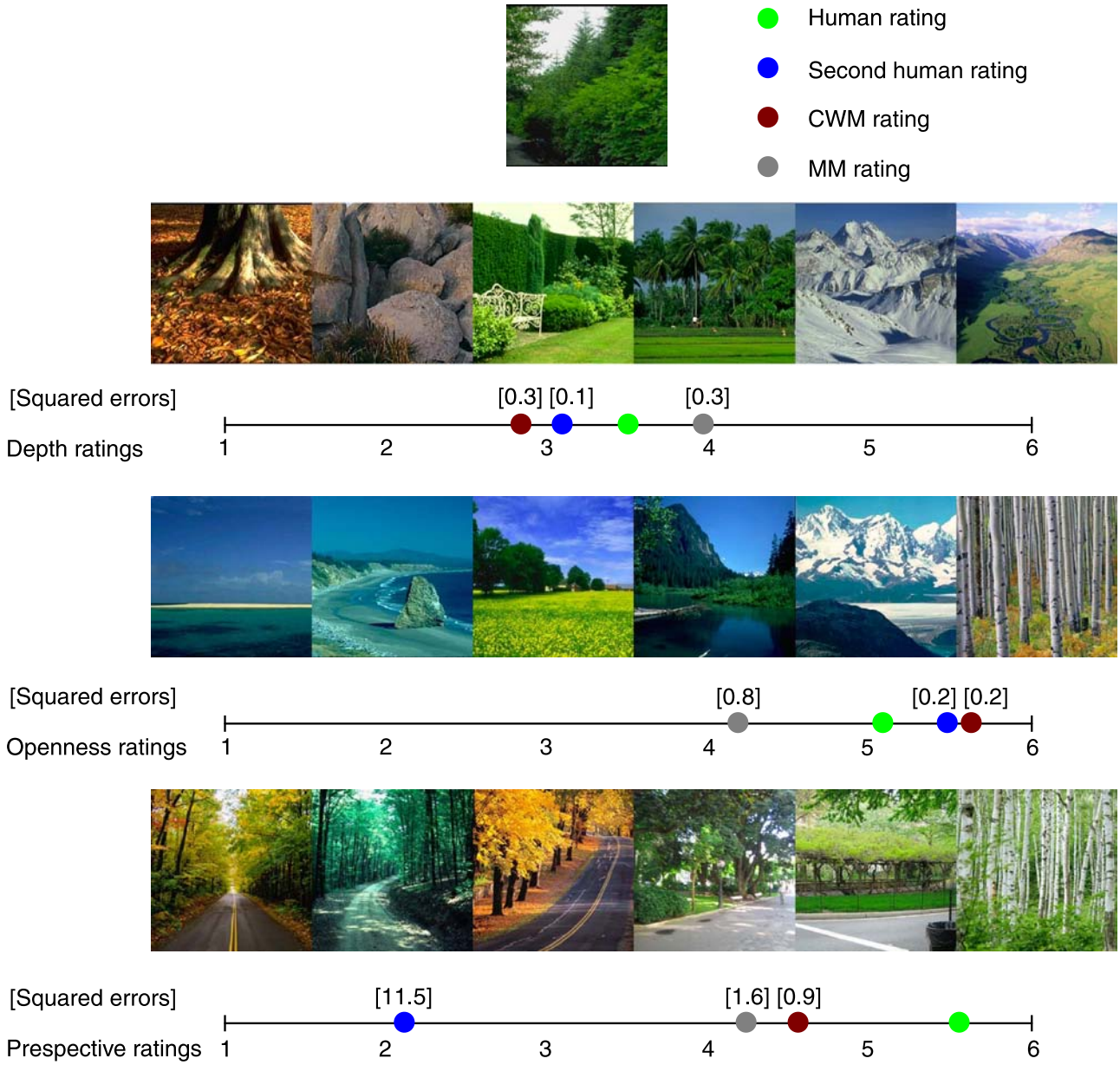


Figure 2. An example of a natural target image and the ratings provided by the original human observer, a second human observer, the CWMs (cluster-weighted models), and the MMs (mean models). Bracketed values are the squared errors between each rating and the original human rating. Color is only used for visualization—all experiments and computer programs used grayscale images.

The choice of these performance comparisons is significant. First, we wish to demonstrate that the CWM model is making use of the information in the image. The MM has the minimum expected squared error for any model that ignores the image, therefore it is an appropriate lower bound for CWM performance—any improvement over its accuracy is due to the CWM capturing useful perceptual information from the images. Comparison to the LM, which is equivalent to a CWM using only one cluster, is significant because it indicates how much of the CWM's performance is due to its context-sensitivity (e.g. if the image is a forest or a city). The CWM automatically implements context sensitivity by clustering images and assigning different regression functions to each cluster. If this provides an important boost to performance, we would expect the CWM squared errors to be significantly smaller than the LM squared errors, which result from using a single linear regression function in all circumstances.

Finally, we need to know a reasonable upper bound for CWM performance. For some properties of some images, multiple human observers will have extremely similar ratings. For example, in Figure 2, both human observers rated the image as nearly completely closed. In these

Depth 4 × 4	Human	CWM	LM	MM
All	**0.46** (0.03)	**0.56** (0.04)	0.61 (0.04)	0.94 (0.05)
Natural	**0.55** (0.04)	**0.64** (0.06)	0.66 (0.05)	1.10 (0.08)
Urban	**0.36** (0.03)	**0.42** (0.03)	**0.44** (0.04)	0.71 (0.05)
Openness 8 × 8	Human	CWM	LM	MM
All	0.71 (0.05)	0.87 (0.06)	1.08 (0.06)	2.84 (0.10)
Natural	0.68 (0.07)	0.93 (0.08)	1.10 (0.09)	2.97 (0.13)
Urban	**0.74** (0.08)	**0.75** (0.08)	0.88 (0.07)	2.65 (0.15)
Perspective 2 × 2	Human	CWM	LM	MM
All	1.80 (0.16)	**1.95** (0.10)	2.00 (0.10)	2.61 (0.11)
Natural	2.15 (0.26)	**2.10** (0.18)	**2.09** (0.18)	2.29 (0.18)
Urban	1.49 (0.18)	**1.57** (0.12)	1.69 (0.13)	2.65 (0.12)

Table 3. Mean squared errors for humans, CWM, LM, and MM (with standard errors of the means). CWM is significantly better than MM in all cases (one-sided Wilcoxon sign-rank test $p < 0.01$). Bold human mean squared errors were not significantly less than their CWM counterparts ($p < 0.01$). Bold LM mean squared errors were not significantly greater than their CWM counterparts ($p < 0.05$).

cases, we would expect a good CWM for openness to achieve a squared error of nearly 0 between its prediction and a human observer's prediction. At the other extreme, perspective is difficult to judge in many natural images because different observers attend to different orientations and vanishing points. In Figure 2, one human rated the image as nearly parallel, while the other attended to other elements and rated it as nearly perpendicular. In this case, even a good CWM might produce a rating substantially different than the human's rating on that scene—it is not a bad model, but simply reflecting that the stimulus is ambiguous. Therefore, when judging CWM performance, it is useful to have a test set where we know how consistently multiple human observers would rate each property of each image. One way to facilitate this is to collect a second set of human ratings on the test images. Then, for each image, we can compare the squared error between the CWM prediction and the original human prediction and the squared error between the original human prediction and the second human prediction. If the two humans are in close agreement, a good CWM model should have low squared error. If the two humans disagree substantially, a good CWM model might have high squared error due to the ambiguity of the stimulus.

In order to compare the human, CWM, LM, and MM performances on the same set of images, the test set was specified to be the set of images with two human ratings. These double-rated images were randomly selected to be a representative sample of the entire image database. On each double-rated image, one rating was treated as the "true" perception and the other used as the human "model" perception. This resulted in a set of 6300 training and 838 test images. In all but 15 cases the duplicate ratings were from two different observers. Using the results from the optimal resolution analysis (Table 1), the CWMs were trained on 4 × 4 GIST features for depth, 8 × 8 GIST features for openness, and 2 × 2 GIST features for perspective, with the number of clusters chosen via cross-validation on the training data as described previously.

Table 3 reports the mean squared errors (and their standard errors) of the human, CWM, LM, and MM perceptual ratings on the test images. To gain an intuition of how the magnitude of squared errors corresponds to differences in rating, look at Figure 2. Note that squared error penalizes large rating differences much more than small ones. Overall humans were most consistent at estimating dominant depth, somewhat less consistent at estimating openness, and least consistent at estimating perspective. The estimation of perspective in natural scene images was particularly difficult for observers, with an inter-observer mean squared error of 2.15, close to the mean squared error of the MM. On the other hand, observers were very accurate in estimating the dominant depth of both natural and urban scenes, as found previously by Torralba and Oliva (2002).

The CWM models' squared errors were significantly smaller than those produced by the MM models in every condition according to the one-sided Wilcoxon sign-rank test ($p < 0.01$, N = 729–837 (all), N = 358–455 (natural), N = 367–382 (urban)). The human ratings had significantly smaller squared errors than the CWM model on all perspective ratings and openness ratings for natural scenes ($p < 0.01$, N = 358–729 (perspective), N = 455 (openness)). On the other hand, human ratings did not significantly outperform CWM for depth on all images and natural images. In addition, humans only marginally outperformed CWM for depth on urban images ($p = 0.05$, N = 367). Human and CWM accuracy at judging the openness of urban scenes were relatively similar (human ratings were more accurate with $p < 0.02$, N = 382). Therefore, it seems likely that the CWM model matches humans' global scene depth perception on this database. Even in the other two categories, at least 90% of the CWM ratings' squared errors were within two standard

deviations of the human squared error distribution, the equivalent percentage for the MM was as low as 50% (for openness prediction on natural images).

The CWM performance trends also match the human performance trends very well. In both cases, depth was the most accurately predicted property, followed by openness, with perspective a distant third. This is unsurprising because in some images, particularly images of natural scenes, there are multiple or ambiguous vanishing points and the determination of a dominant perspective becomes much more subjective, an issue also apparent in the poor human re-rating performance on this property (see Figure 2 for an example). Human observers skipped rating the perspective property more frequently than any other (15% of trials, compared to 5% for depth and 0.5% for openness). The perspective skip rate was much larger for natural scenes (23%) than urban scenes (5%), the largest natural/urban skip rate difference of the three properties (compared to 6%/4% for depth, 0.5%/0.5% for openness). In future work, it would be useful to explore the multi-modal aspects of perspective perception. Running multiple human observers on the ambiguous images and then finding the modes of the $p(r_j|\ f_j)$ function might provide new insights.

Depth and perspective were more accurately predicted by humans and computer models on urban scenes than on natural scenes. Humans were almost equally accurate at rating openness in both contexts, but the CWMs were more accurate for urban images. These trends reflect the simpler structure found in urban environments, especially with respect to perspective, which is strongly defined by rectilinear features such as buildings and streets.

The CWM provided a significant improvement in performance compared to the LM for the depth-all condition and for all openness ratings ($p < 0.01$, N = 799 (depth), N = 382–837 (openness)). It provided a significant improvement ($p < 0.05$, N = 432) for depth-natural, perspective-all, and perspective-urban, but not for depth-urban or perspective-natural. Unsurprisingly, perspective CWM models tend to use the fewest clusters and, therefore, exhibit less non-linearity, than models for depth and openness (see Table 2). Overall, these results indicate that the non-linear models are significantly better than linear models at matching human perception, but a computer vision application concerned with efficiency can use a linear model without sacrificing too much accuracy.

Another useful measure of model performance is the R^2 statistic, which measures the fraction of variance accounted for by a regression model (Myers & Well, 2003). Table 4 contains the R^2 values for all the computer models and the human re-ratings described in Table 3. Notice that although a theoretical maximum of $R^2 = 1$ is possible, the human re-ratings are always substantially below that level due to the aforementioned variance in scene perception between observers. With the exception of natural scene perspective, human re-ratings have the highest R^2 values, followed by CWM, in turn followed by LM. The MM R^2 statistics are always near zero because the expected squared error from the mean rating value is the definition of variance—therefore a MM, by definition, does not account for any rating variance. Table 4 confirms the previously observed pattern that CWMs are generally the best computational models for global scene properties.

Depth 4 × 4	Human	CWM	LM	MM
All	0.51	0.41	0.36	0.00
Natural	0.50	0.42	0.40	0.00
Urban	0.50	0.41	0.38	0.00
Openness 8 × 8	Human	CWM	LM	MM
All	0.75	0.69	0.62	0.00
Natural	0.77	0.68	0.63	0.00
Urban	0.72	0.72	0.67	0.00
Perspective 2 × 2	Human	CWM	LM	MM
All	0.31	0.25	0.23	0.00
Natural	0.06	0.08	0.08	0.00
Urban	0.44	0.41	0.36	0.00

Table 4. R^2 values measuring the fraction of variance accounted for by human, CWM, LM, and MM models—$R^2 = 1$ indicates prefect prediction.

In R^2 terms, humans have the most consistent perception of openness and CWMs also predict openness very well, matching human performance on urban environments. The R^2 values for humans and CWMs are considerably lower for depth. Although depth was the most accurately measured property in terms of mean squared error, the R^2 values account for the fact that depth ratings have less variance across images than openness ratings. The difficulty of predicting natural scene perspective is highlighted even more strongly in the R^2 statistics than in the squared error statistics—it is very clear that no model (not even human re-rating) is accounting for a substantial amount of the variance. This is compatible with the previous hypothesis that natural scene perspective is frequently multi-modal. Variance, the basis of the R^2 statistic, is an inherently unimodal property.

Figures 3, 4, and 5 show the four best and worst results, measured by squared error, for each property on each test data set. Note that although the images are displayed in color for visualization purposes, the experimental observers and all layout-prediction models operated on their grayscale versions. Even with this small sample it is clear that in each category a wide range of images were rated correctly. The correctly rated images in depth and perspective cover some difficult cases and points along those layout spectra (Figures 3 and 5). On this small sample, the correctly rated openness images seem to contain many completely closed images (Figure 4). That end of the openness spectrum is easy to rate consistently, by assigning the maximum rating to any skyless image.

Figure 3. The most and least accurate depth layout predictions made by the CWMs, "h" indicates the human ratings. A rating of "1" indicates the scene is near and "6" indicates that the scene is far. Color is only used for visualization—all experiments and computer programs used grayscale images.

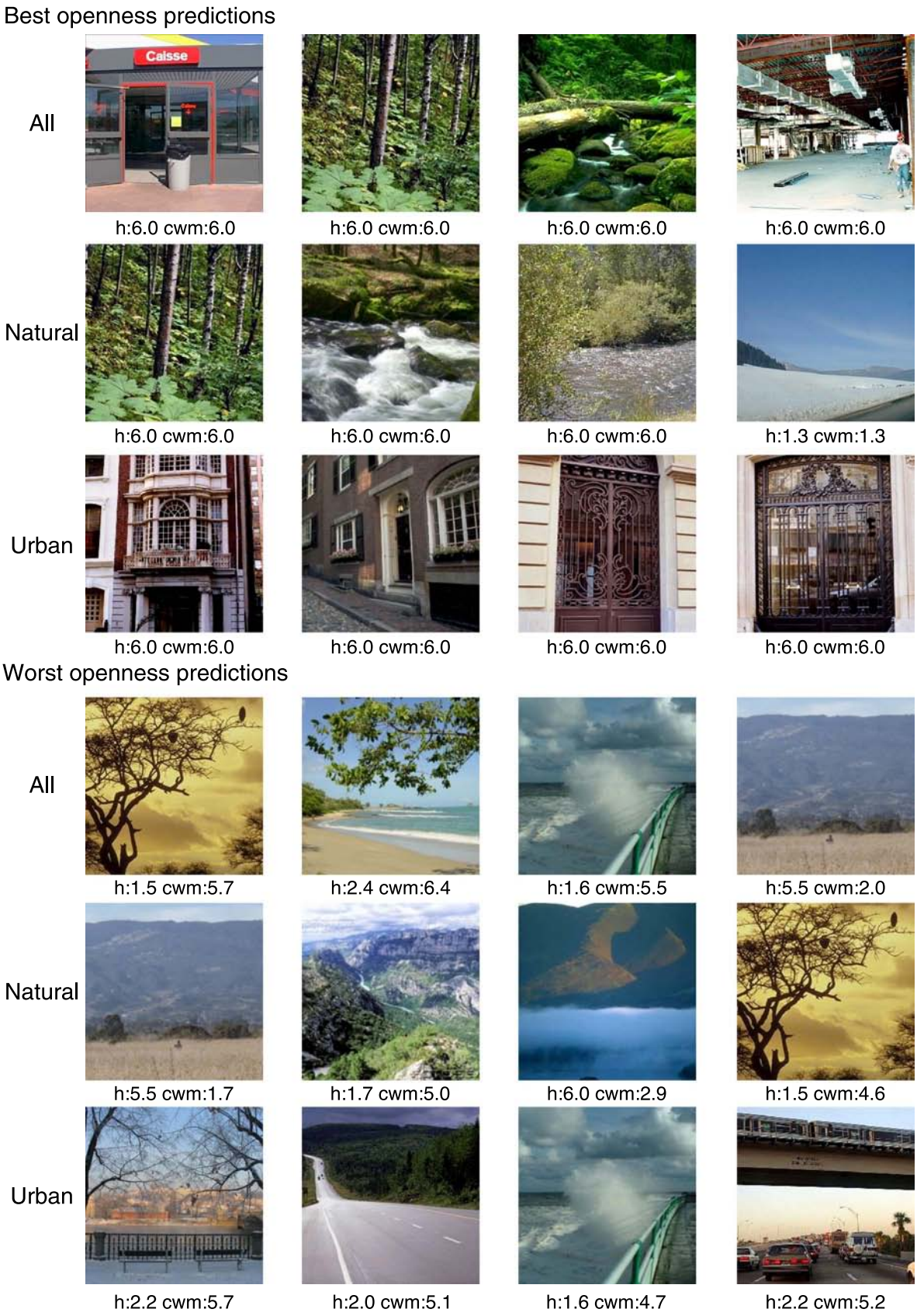


Figure 4. The most and least accurate openness layout predictions made by the CWMs, “h” indicates the human ratings. A rating of “1” indicates the scene is open and “6” indicates that the scene is closed. Color is only used for visualization—all experiments and computer programs used grayscale images.

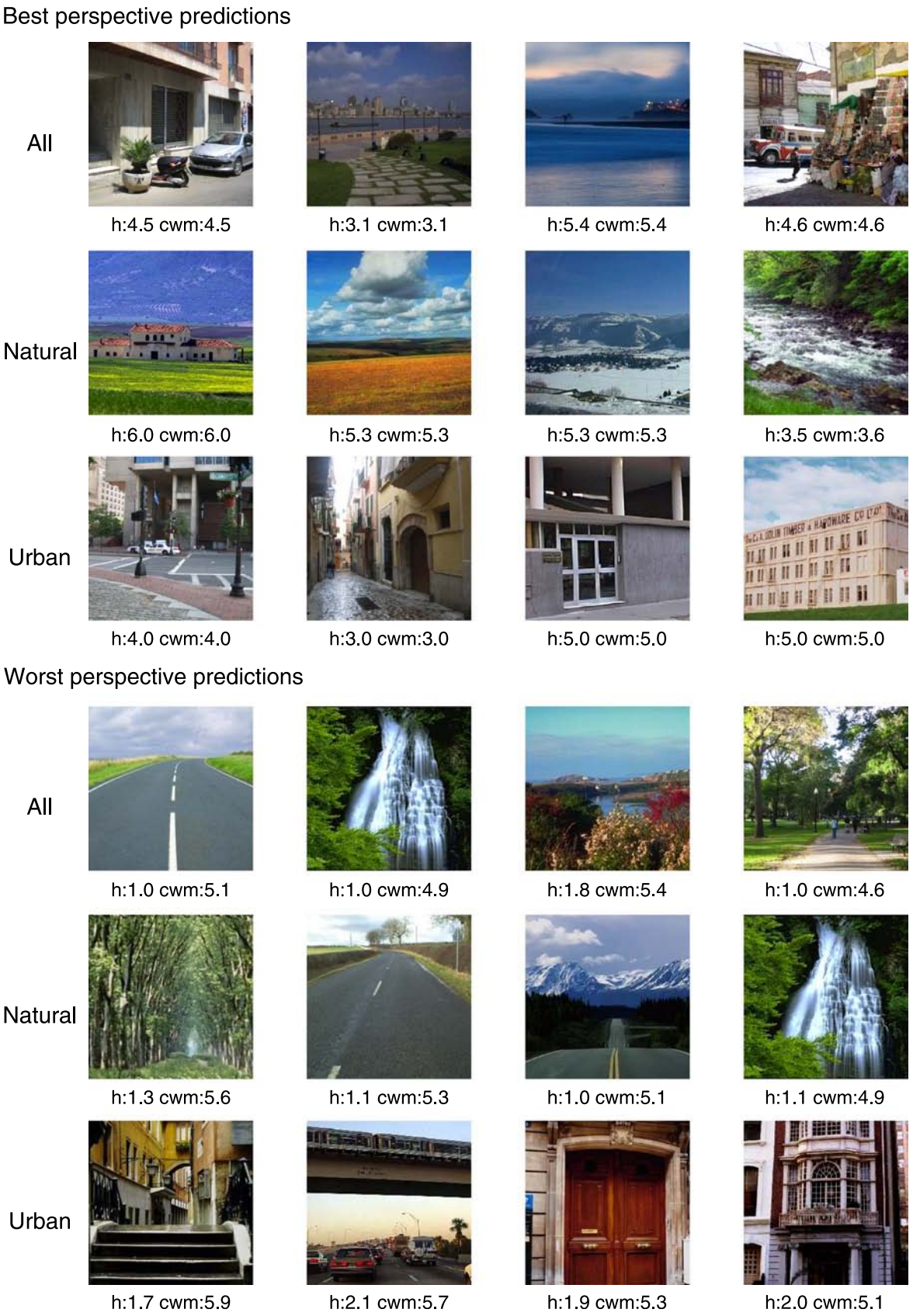


Figure 5. The most and least accurate perspective layout predictions made by the CWMs, “h” indicates the human ratings. A rating of “1” indicates the scene is perpendicular and “6” indicates that the scene is parallel. Color is only used for visualization—all experiments and computer programs used grayscale images.

Several of the worst results are clearly on outlier images. For example, the close-up photo of the tree trunk in the depth examples (Figure 3) is very shallow and photographed at an unusual angle. Similarly, several images are from aerial photographs and one is a view of a balcony corner, two types of images that are not well represented in the database to begin with. Two of the worst-scoring urban perspective images appear to be the result of unusual human rating decisions—the CWM rated two building entrances as approximately parallel while the human rater ranked them as perpendicular. In both cases the second human rater on the images choose ratings very similar to the CWM values. But it would be a mistake to assign all of the blame for poor performance on test examples that were dissimilar from most training examples—a substantial number of incorrectly rated images are very close to the model cluster centers and are not obvious outliers. The most likely path to closing the small gap between the CWMs and humans lies in improving the features or incorporating algorithms that provide more detailed depth maps.

Measurement of cross-observer generalization

In the previous analyses, most observers appear in both the training and test data sets. Therefore, it is possible that the CWMs perform well because they learned to predict the responses of a particular group of observers and their performance would be poor if they were compared to the ratings of a novel observer. To address this concern, we conducted a new analysis. We constructed thirteen testing and training splits. In each split, the testing set contained ratings from only one observer and the training set contained no ratings from that observer. Therefore, the CWM has no opportunity to learn an observer-specific model—if it performs well, that must be the result of learning generic properties of scene perception that generalize to multiple observers.

The results reported in Table 5 compare the average mean squared error over all the possible selections of held-out observer splits to (the "holdouts" condition) to the previously reported mean squared error (the "mixed" condition). For this analysis we only measured performance on all images, not subdividing into natural and urban conditions. Although the average holdout performance was worse than the mixed performance, the mixed performance level was always within one standard deviation of the holdout average. We also compared the performance of CWM and MM across the holdout conditions and found that for every held-out observer and for every property, CWM had significantly better performance than MM (one-sided Wilcoxon sign-rank test, $p < 0.012$, N = 276–1266).

	CWM (mixed)	CWM (holdouts)	MM (holdouts)
Depth (4 × 4)	0.56	0.63 (0.16)	1.00 (0.28)
Openness (8 × 8)	0.87	0.91 (0.20)	2.83 (0.40)
Perspective (2 × 2)	1.95	2.06 (0.40)	2.67 (0.34)

Table 5. Comparison of mean squared error between mixed and single-observer holdout conditions. Average mean squared error (and standard deviations) for CWM (holdouts) and MM (holdouts) reported across all observers. CWM (holdout) outperforms MM (holdout) in all cases ($p < 0.012$).

Based on these results, we conclude that most of the CWMs' performance results from capturing universal perceptual properties and relatively little is the result of learning observer-specific models.

Discussion

Because of their success in replicating human perception, it is interesting to dissect the CWM models to gain better understanding of the information they represent. In order to do this, we need to develop a compact visual representation for the models. Figure 6 depicts the 8 Gabor-filter orientations that we use to compute the GIST features. Note that for each orientation there is a pair of Gabor filters, a sine and cosine pattern of the same frequency and orientation. Furthermore, these patterns indicate that the filters are sensitive to image brightness boundaries that are perpendicular to the filter orientation. For example, if an image has a strong horizontal boundary, such as a horizon, the vertically oriented Gabor filters will have the strongest response to that aspect.

Beneath the filters, Figure 6 shows the responses produced by convolving each filter with an example image (the image was converted to grayscale before filtering). Each response image combines the output from the sine and cosine versions of the relevant Gabor, so they capture brightness variations that match the filter's frequency and orientation regardless of phase. Bright pixels indicate a region of the image containing variations that match that Gabor filter's frequency and orientation. Note that, as mentioned above, the vertically oriented filters are most sensitive to the horizon line in the image. These filtered images can be summarized by computing the average of all the response-pixel values, which indicates the global response of the filter to the image. A convenient graphical representation of the global filter responses is to draw lines representing each filter orientation, setting the length of the line to match the magnitude of the averaged filter responses. At the bottom-left of Figure 6 we can see this representation of the Gabor-filter responses to the sample image. As mentioned previously, we can increase the spatial resolution of the GIST representation by subdividing an image into grid

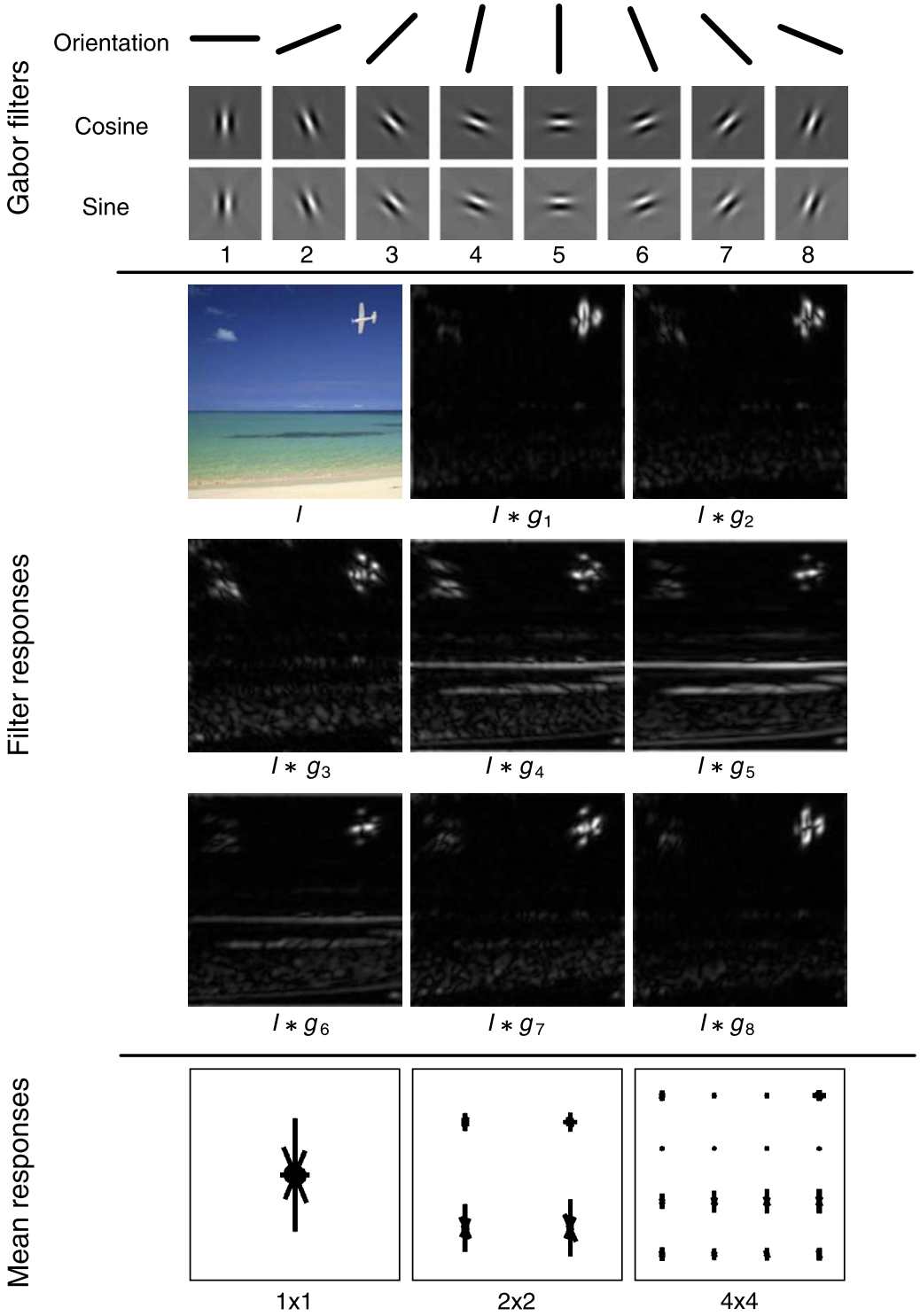


Figure 6. Top: The eight orientations of Gabor filters used on the images, showing the cosine and sine components of each filter. Middle: The responses produced when convolving an example image (I) with each filter. Bottom: Summarizing the responses by indicating the average response to each orientation across the whole image (a 1 × 1 grid), and across localized subregions designated by division of the image into 2 × 2 and 4 × 4 grids. Color is only used for visualization—all experiments and computer programs used grayscale images.

squares and reporting average Gabor-filter responses for each square. The representations produced by subdividing this sample image into 2 × 2 and 4 × 4 grids are also displayed at the bottom of Figure 6.

Figure 7 shows three images that evoke very different Gabor responses. The close-up forest image primarily contains tree-trunks, which primarily excite the horizontally oriented Gabor filters. As mentioned before, the

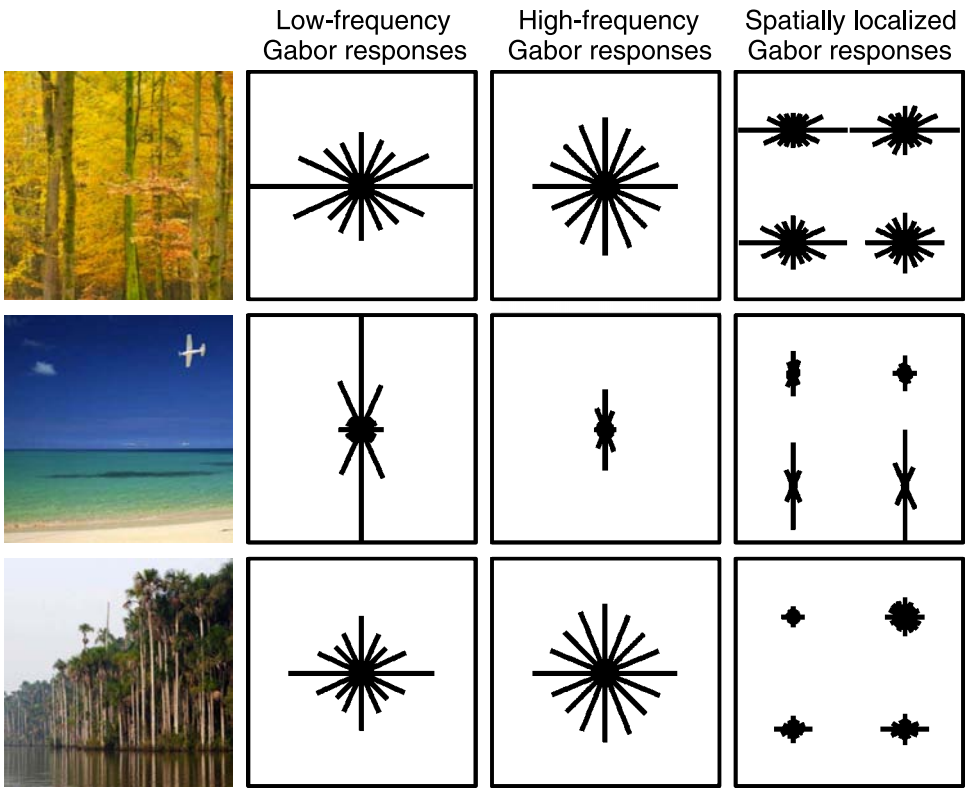


Figure 7. Three images and their responses to Gabor filters. The magnitude of the average responses to different orientations reveals differences in the image structure. Changing the filter frequencies reveals different information, as does computing the spatially localized averages (2 × 2 grid shown). Color is only used for visualization—all experiments and computer programs used grayscale images.

beach image is dominated by its horizon and therefore primarily has vertical Gabor structure. More complex images, such as the third image, which shows palm trees along a shoreline, contain a mixture of orientations and produce a more evenly distributed Gabor output. We can also see that increasing or decreasing the frequency (and, therefore, decreasing or increasing the scale, respectively) of the filters alters the pattern of responses. Some structures, such as the horizon line of the beach scene, are most prominent at a particular scale. As mentioned previously, the GIST representation used in this paper employs Gabor filters at 4 different scales. Finally, note that when we increase the spatial resolution to 2 × 2, new aspects of each image are represented. For example, we can see that the forest close-up scene is very homogeneous, while the palm-tree shoreline image contains very little texture in its upper-left corner.

Figure 8 demonstrates that we can use this same representation to describe the linear regression functions computed on these Gabor-filter responses. The oriented lines represent the same Gabor filters as in Figures 6 and 7, but now the magnitudes represent coefficients that we will apply to the summed responses to those filters. The magnitude of each coefficient is indicated by length and color indicates whether the coefficient is positive (blue) or negative (red). This visual representation can be extended to 2 × 2, 4 × 4, or 8 × 8 Gabor grids just as we extended the Gabor-response representation in Figure 6.

Figures 9, 10, and 11 use these graphical representations to depict the best-fit models (in the all-data condition, with

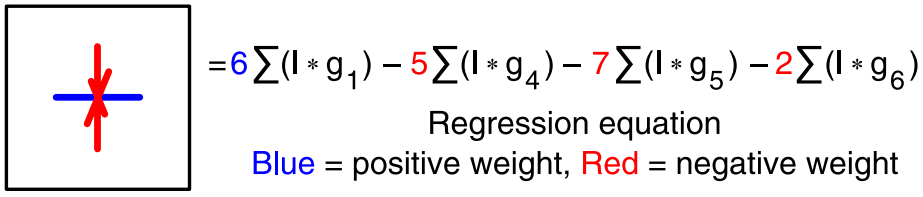
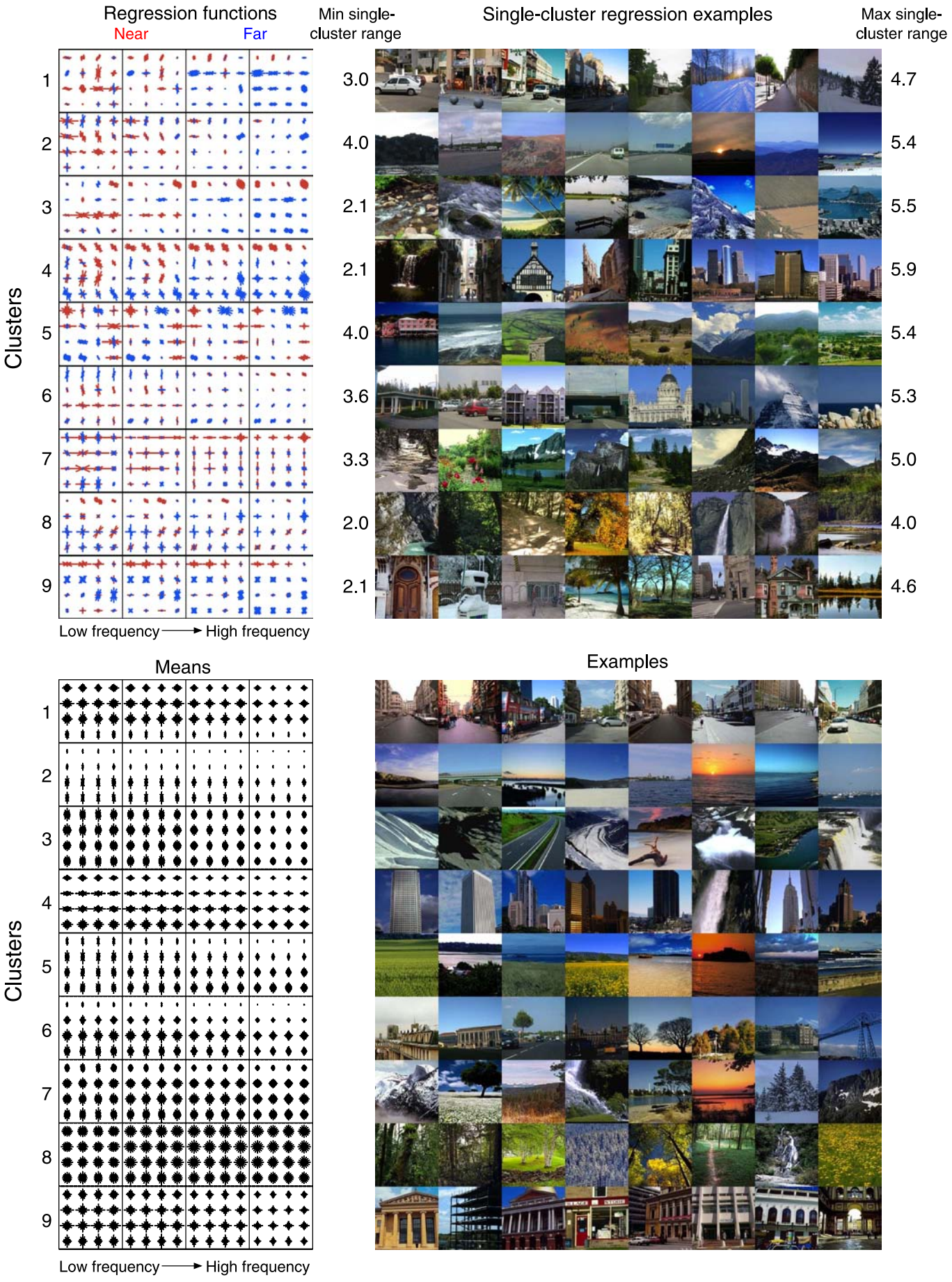


Figure 8. The average-Gabor-response representation can be used to represent regression functions. In this case, line length indicates the magnitude of the coefficient applied to a filter and color indicates if the coefficient is positive or negative. An analogous representation is used for regression on 2 × 2, 4 × 4, or 8 × 8 mean-Gabor-response regression.


Regression functions
Near
Far
Min single-cluster range
Single-cluster regression examples
Max single-cluster range
Clusters
1
2
3
4
5
6
7
8
9
3.0
4.0
2.1
2.1
4.0
3.6
3.3
2.0
2.1
4.7
5.4
5.5
5.9
5.4
5.3
5.0
4.0
4.6
Low frequency → High frequency
Means
Examples
Clusters
Low frequency → High frequency

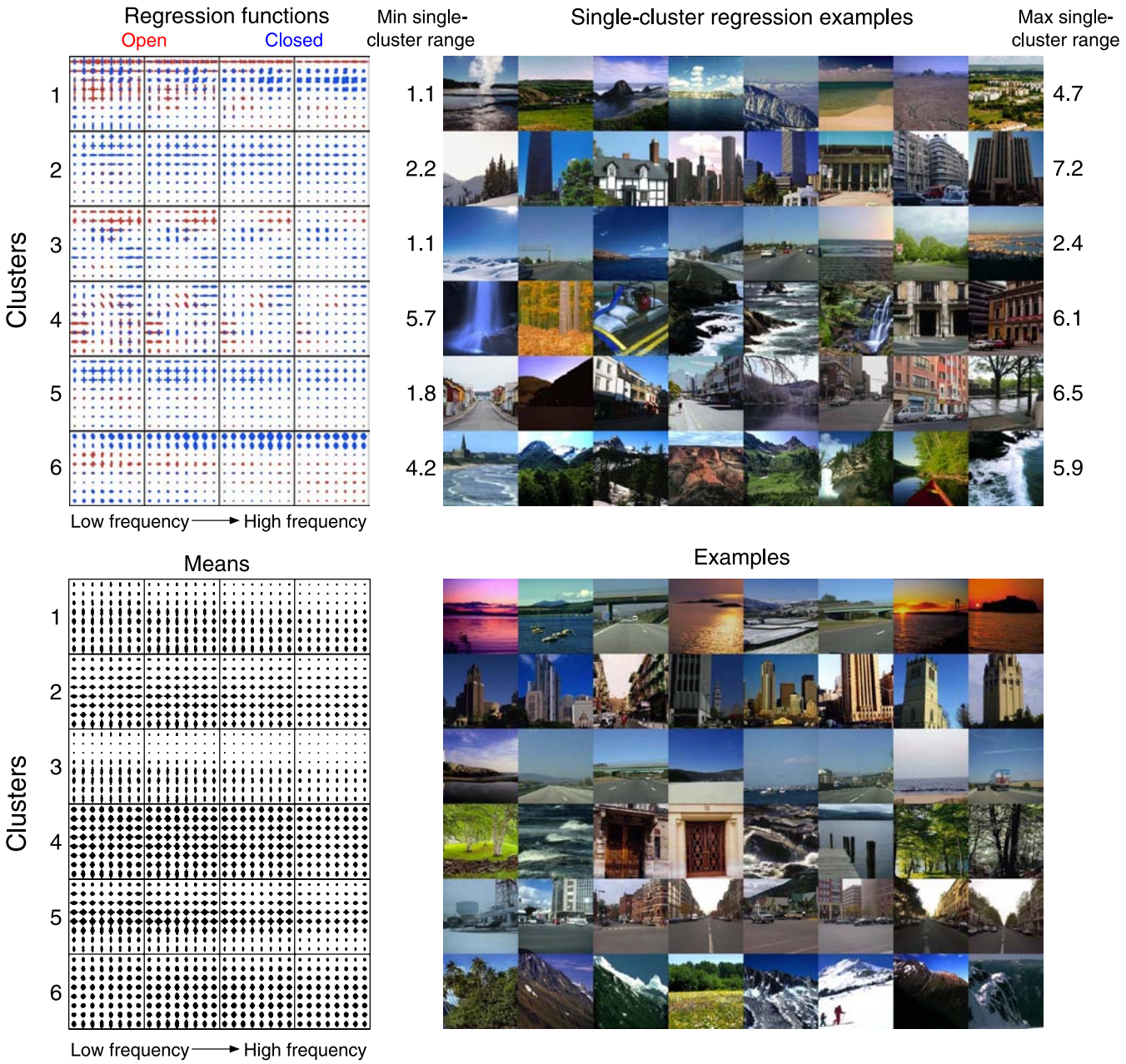


Figure 10. The clusters in the openness (all scenes) CWM model. Top: the regression functions for each cluster and a range of training images whose predicted ratings are mostly (>90%) determined by each cluster. Bottom: the mean feature values for each cluster and the training images nearest to each mean. Functions and means are represented by lines that represent the orientation, frequency, and location of the Gabor filters that comprise the GIST features, magnitudes are represented by length, see text for more detail. Color is only used for visualization—all experiments and computer programs used grayscale images.

Figure 9. The clusters in the depth (all scenes) CWM model. Top: the regression functions for each cluster and a range of training images whose predicted ratings are mostly (>90%) determined by each cluster. Bottom: the mean feature values for each cluster and the training images nearest to each mean. Functions and means are represented by lines that represent the orientation, frequency, and location of the Gabor filters that comprise the GIST features, magnitudes are represented by length, see text for more detail. Color is only used for visualization—all experiments and computer programs used grayscale images.

performance described in Table 3) for depth, openness, and perspective. As previously described, each property model is a collection of linear regression functions (9 for depth, 6 for openness and 4 for perspective) and each regression function is used to predict the property rating for a set of scenes.[8] The number of clusters in each model, as well as the clusters' regression functions and locations in GIST feature space, were learned solely from the training data. In the figures we display two visualizations of each cluster: a visualization of the regression weights

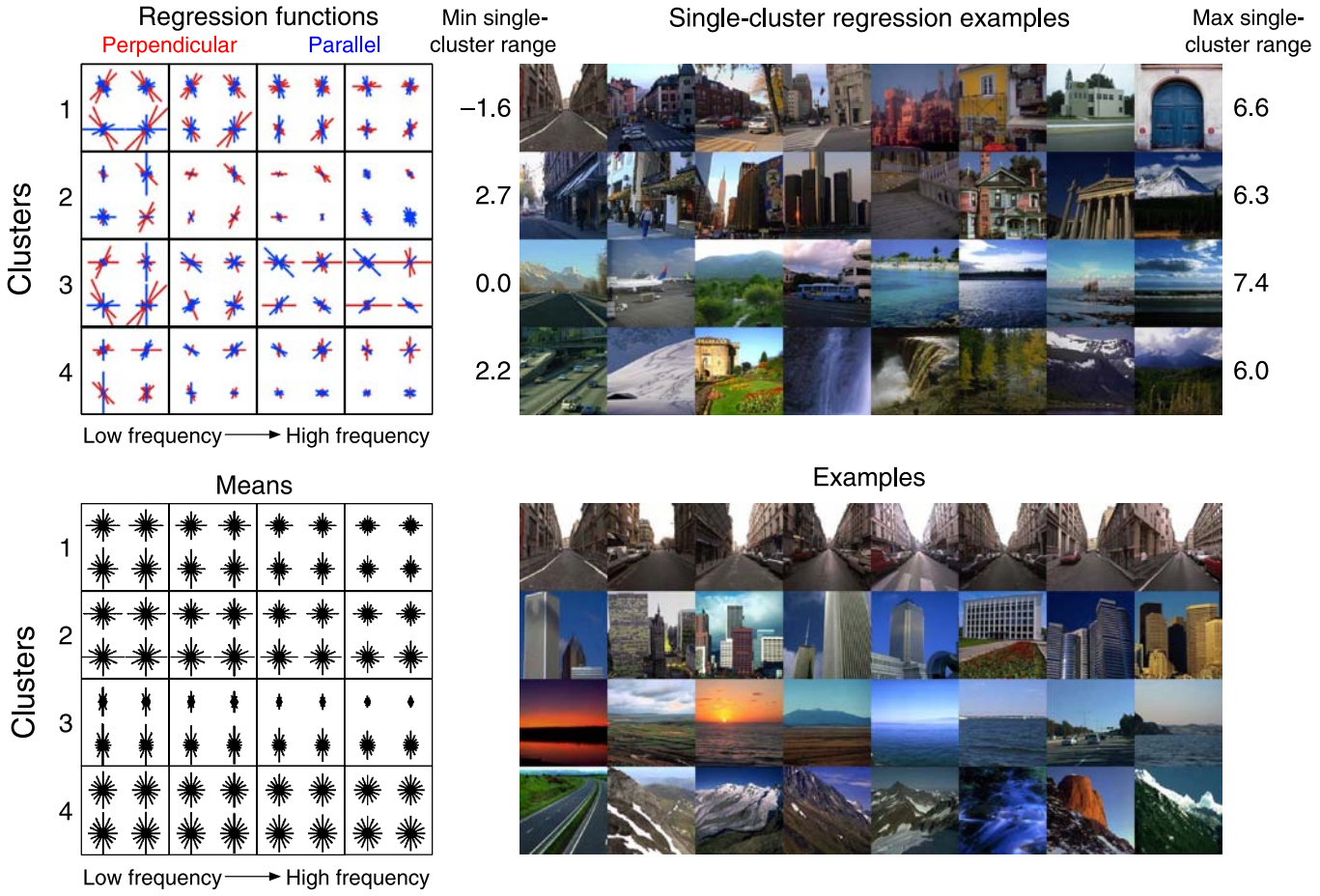


Figure 11. The clusters in the perspective (all scenes) CWM model. Top: the regression functions for each cluster and a range of training images whose predicted ratings are mostly (>90%) determined by each cluster. Bottom: the mean feature values for each cluster and the training images nearest to each mean. Functions and means are represented by lines that represent the orientation, frequency, and location of the Gabor filters that comprise the GIST features, magnitudes are represented by length, see text for more detail. Color is only used for visualization—all experiments and computer programs used grayscale images.

for each cluster, along with some example images and their predicted ratings, and a visualization of the centers (means) of each cluster, along with a sample of images that are very close to those centers. All images shown in the figures are drawn from the training data. The regression images were chosen by first selecting the images whose predicted perceptual ratings were at least 90% determined by that particular cluster's regression function, and then selecting a set that would cover the widest range of ratings in that set. The ratings assigned to the beginning and ending images in each set are displayed. The images are displayed in color for visualization purposes only—all experiments and models utilized grayscale images.

The cluster regression weights and cluster centers are also represented by displaying their values in GIST feature space. As mentioned previously, the GIST representation we used consisted of Gabor filter responses covering 4 scales (low frequency to high frequency) and 8 orientations measured at R × R image locations. The spatial scales are indicated by dividing each representation into four boxes, each one containing the representation or regression weights at one particular scale.

Observing the examples near the mean of each cluster reveal that many clusters are centered on a particular semantic scene class: open landscapes, urban skylines, mountains, streets, etc. This is not surprising given that the GIST features have been previously used to distinguish between these types of categories (Oliva & Torralba, 2001), therefore we would expect that there is a strong tendency for the nearest neighbors of any point in GIST-feature space to contain semantically similar scenes. On the other hand, the regression examples, which are drawn from the entire region of GIST feature space dominated by a particular cluster, show more diversity. For instance there are cases in all three models in which the same regression function is applied to both natural and urban scenes that have some structural similarities. For example,

cluster 9 in the depth model is centered on images of the front of buildings, but when we look at the regression examples for that cluster they also include natural images in which a line of trees provide important depth cues. This generalization can occur because there are structural similarities that transcend the semantic categories.

The influence of context on the regression functions is the most important feature of these models. In each model there are Gabor frequencies and locations that are associated with one end of the rating scale in some clusters and the opposite end in others. For example, in the depth ratings (Figure 9), some spatial frequencies at the top of the image are associated with "near" in cluster 4 and "far" in cluster 5. If we look at the regression examples, cluster 4 mostly describes depth of scenes with buildings, so we would expect frequency content at the top of the image to be associated with a nearby building that is blocking the sky. On the other hand, cluster 5 is associated with landscape scenes and frequency content at the top of these images may be associated with distant cloudbanks or mountains which indicate greater depth. Similar context-dependent effects can be observed in the models for openness and perspective (Figures 10 and 11).

The openness regression functions (Figure 10) also reveal a variety of image cluster-specific representations. For instance, clusters 1 and 3 cover very open scenes (highways, open ocean views, fields) centered on a mixture of natural and urban images. These context-independent clusters are estimated with a mix of high and low-spatial-resolution weights. Structures that are diagnostic of a semantic category (tall front surfaces of buildings for urban scenes, and oblique surfaces of mountainous natural landscapes), are represented by class-specific clusters (clusters 2 and 6, respectively). Finally, the perspective regression functions (Figure 11) demonstrate relatively little spatial structure (given that its optimal spatial resolution is 2×2), but show global feature orientation sensitivity. This makes sense because judging the perspective of the scene is a global estimation that requires detecting the orientation of lines that stretch across the image. Some clusters are especially sensitive to oriented lines in the bottom half of the image, which helps them detect the orientation of streets, paths, and the ground-object boundaries. For instance, in cluster 1, most diagonal filters are associated with perpendicular views (consider the lines of a street extending away from an observer) and horizontal and vertical filters are associated with parallel views.

General discussion

Inference of useful scene layout properties from image-based features has long been a focus of psychological (Gibson, 1986) and computer vision (Marr, 1982) research. Because precise physical models of real-world scenes are inaccessible or computationally intractable, determining the relationship between two-dimensional image information and perceptual properties representing the "shape of the scene" is relevant for both psychological and computer vision research. Even a system designed to produce a detailed scene reconstruction can benefit from a global system that provides holistic scene layout information. Holistic layout knowledge can act as a set of prior probabilities during the detailed reconstruction process, providing knowledge that can resolve local scene perception ambiguities. Here, we show that human observers are very consistent in estimating the dominant depth and openness of natural and urban scene pictures, but their perceptual estimation of perspective is more variable. By using an algorithm that successfully translates images from pixels into a higher-order feature space (spatial-envelope coordinates, Oliva & Torralba, 2001), we observe that human judgments of the three layout properties can be reliably predicted, particularly for estimating the dominant depth or *scale* of a scene. Furthermore, we show that the model's predictions are general and not specific to the observers it was trained on. Importantly, we discovered that the optimal spatial resolutions for determining layout in this higher-order space vary systematically with the content of the space (being natural or manufactured), and the type of layout: openness is best estimated at a high spatial resolution, dominant depth is best estimated at a medium spatial resolution, and perspective is best estimated at a low spatial resolution.

The similarity between the CWMs operating on GIST features and human perceptual layout estimation suggests that the structure of these models may be well suited to encoding structural scene priors. By encoding spatial information that is correlated with the three-dimensional extent of the scene, these models can act as "context" to guide navigation tasks (similar to other works which have used global features context to predict object search and eye movements, Ehinger, Hidalgo-Sotelo, Torralba, & Oliva, 2009; Torralba, Oliva, Castelhano, & Henderson, 2006).

The algorithm can also form the pre-processing stage of a search or clustering algorithm that retrieves or groups images by their three-dimensional physical layout similarity rather than by image features or semantic categories. There are a number of possible uses for this perceptual-layout-based search, including applications in visual robot navigation. Confronted with a new environment, a robot with a large database of past navigational experiences could use a global physical layout estimate to retrieve those examples most likely to apply to the current situation. Given the difficulty of extracting detailed layout, a robust global estimate combined with past experience could serve as a useful prior for algorithms attempting a detailed three-dimensional scene reconstruction. Many tasks, such as picking up an object or traveling towards a building, only require detailed geometric knowledge of part of a scene, while the rest is represented by global

properties. Exploring such mixed-resolution representations could be an interesting topic of future research.

Additionally, a model based on the global physical layout space described in this paper would be able to inform and guide the resolution of local metric cues of depth, perspective, distance, to built more robust spatial estimators. Years of psychophysical studies have shown that human observers may use a variety of local two-dimensional image cues to infer metric information about the scale of a space, its dominant depth or perspective. Recent work in scene recognition has also shown that global properties of layout may be resolved with less exposure time than the semantic category of the scene (Greene & Oliva, 2009b; Joubert et al., 2007), are subject to aftereffects (Greene & Oliva, in press), and may constrain the understanding of the meaning of the scene at the beginning of visual processing (Greene & Oliva, 2009a). A promising avenue would be to merge both global and local cues of scene shape to realize a more complete and accurate representation of three-dimensional natural spaces.

Appendix A

As we discussed in the main body of the paper, increasing the spatial resolution inevitably decreases the precision of frequency information in our PCA representation. Therefore, one might suspect that these results indicate the amount of frequency information necessary to solve these perceptual problems, rather than the optimal spatial resolution. In that case, we would expect performance to decline when the spatial resolution is increased beyond the optimal level because after that there would be too little frequency information present to perform the task. Small declines in performance do occur in most categories once 16 × 16 resolution is reached, but in all cases 16 × 16 performance is better than 1 × 1 performance, which indicates that the main requirement for optimal performance is achieving adequate spatial resolution (Table A1).

	1 × 1	2 × 2	4 × 4	8 × 8	16 × 16
Depth (All)	0.745	0.644	0.607	0.606	0.616
Depth (Natural)	0.828	0.726	0.700	0.675	0.674
Depth (Urban)	0.647	0.503	0.466	0.461	0.476
Openness (All)	1.422	0.914	0.876	0.835	0.843
Openness (Natural)	1.444	0.987	0.932	0.917	0.930
Openness (Urban)	1.288	0.751	0.707	0.716	0.703
Perspective (All)	2.339	1.958	2.016	2.035	2.037
Perspective (Natural)	1.994	1.943	1.941	1.988	1.942
Perspective (Urban)	2.148	1.815	1.769	1.804	1.781

Table A1. The average mean squared error across the five-fold cross validation for each property, data set, and spatial resolution.

Acknowledgments

The authors wish to thank Michelle Greene, who provided a great deal of assistance in creating and administering the experiment. We also thank Timothy Brady and Michelle Greene for providing insightful comments on the draft of this paper. MATLAB code and human ratings' data are available on the authors' website. This work is funded by National Science Foundation grant (0705677) and CAREER Award (0546262) to A.O. Correspondence maybe sent to either author: M.G.R (mgross@broadinstitute.org), A.O (oliva@mit.edu).

Commercial relationships: none.
Corresponding authors: Michael G. Ross; Aude Oliva.
Emails: mgross@broadinstitute.org; oliva@mit.edu.
Address: Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Building 46-4065, Cambridge, MA 02139, USA.

Footnotes

[1]The code is available at http://people.csail.mit.edu/torralba/code/spatialenvelope/.

[2]Note that this cross validation is completely separate from the cross-validation procedure used to choose *N* in the previous section, which only involves the training data.

[3]Torralba and Oliva (2002) and Vailaya et al. (1998) both demonstrated greater than 90% accuracy in automatically classifying images as natural or urban. Our own experiments indicate that for our database, which contains a significant number of images whose natural/urban statuses are ambiguous, the principal components of GIST features at resolutions 1 × 1, 2 × 2, 4 × 4, or 8 × 8 and a linear classifier can produce greater than 83% accuracy on this task.

[4]We used 24 components rather than 25 because an early version of the model used a multi-resolution implementation which made it desirable to use a multiple of 2 and 3 as the number of principal components.

[5]The "curse of dimensionality" is a universal problem in machine learning—see Bishop (2006) or any other textbook.

[6]A one-sided test is appropriate because we are only interested in the case where a higher resolution model outperforms lower resolution models (for information on one-sided tests, see Myers & Well, 2003). For the resolutions we test, a high-resolution GIST grid can

always be transformed into any lower resolution GIST grid by averaging. Therefore, there always exists a high-resolution model that at least equals the performance of the best model at any lower resolution.

[7]Because the PCA representation captures a smaller percentage of the full GIST representation as the resolution increases, it is possible that this resolution analysis is underestimating the potential performance using 8×8 or 16×16 resolutions. However, using the full GIST representations in those cases is impractical because it would exponentially increase the number of training examples required and because the CWM algorithms suffer from numerical precision problems if the feature space is extremely high dimensional.

[8]The number of clusters in each model do not necessarily match the values in Table 2 because the number of clusters is chosen during training and can vary between runs of the training code or due to changes in the training data.

References

Baddeley, R. (1997). The correlational structure of natural images and the calibration of spatial representations. *Cognitive Science, 21,* 351–372.

Barrow, H. G., & Tenenbaum, J. M. (1981). Interpreting line drawings as three-dimensional surfaces. *Artificial Intelligence, 17,* 75–116.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Burge, J., Fowlkes, C. C., & Banks, M. (submitted for publication). Natural scene statistics predict the influence of the figure-ground cue of convexity on human depth perception. *Journal of Neuroscience*.

Coughlan, J. M., & Yuille, A. L. (1999). Manhattan world: Compass direction from a single image by Bayesian inference. In *Proceedings of the IEEE International Conference on Computer Vision,* 941–947.

Creem-Regehr, S. H., Gooch, A. A., Sahm, C. S., & Thompson, W. B. (2004). Perceiving virtual geographical slant: Action influences perception. *Journal of Experimental Psychology: Human Perception and Performance, 30,* 811–821. [PubMed]

Criminisi, A., Reid, I., & Zisserman, A. (2000). Single view metrology. *International Journal of Computer Vision, 40,* 123–148.

Divvala, S. K., Efros, A. A., & Hebert, M. (2008). Can similar scenes help surface layout estimation? *IEEE Workshop on Internet Vision at CVPR* '08.

Ehinger, K., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modeling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition, 17,* 945–978.

Fei-Fei, L., & Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2,* 524–531.

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

Fortenbaugh, F. C., Hicks, J. C., Hao, L., & Turano, K. (2007). Losing sight of the bigger picture: Peripheral field loss compresses representations of space. *Vision Research, 47,* 2506–2520. [PubMed]

Gershenfeld, N. (1999). *The nature of mathematical modeling*. Cambridge: Cambridge University Press.

Gibson, J. (1986). *The ecological approach to visual perception*. Hillsdale: Lawrence Erlbaum Associates.

Girshick, A., Burge, J., Erlikhman, G., & Banks, M. (2008). Prior expectations in slant perception: Has the visual system internalized natural scene geometry? [Abstract]. *Journal of Vision, 8*(6):77, 77a, http://journalofvision.org/8/6/77/, doi:10.1167/8.6.77.

Gosselin, F., & Schyns, P. G. (2001). Bubbles: A technique to reveal the use of information in recognition tasks. *Vision Research, 41,* 2261–2271. [PubMed]

Greene, M. R., & Oliva, A. (2009a). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology, 58,* 137–176. [PubMed] [Article]

Greene, M. R., & Oliva, A. (2009b). The briefest of glances: The time course of natural scene understanding. *Psychological Science, 20,* 464–472. [PubMed] [Article]

Greene, M. R., & Oliva, A. (in press). High-level aftereffects to global scene property. *Journal of Experimental Psychology: Human Perception and Performance*.

Heaps, C., & Handel, C. H. (1999). Similarity and features of natural textures. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 299–320.

Heeger, D., & Bergen, J. (1995). Pyramid-based texture analysis/synthesis. In *Proceedings ACM SIGGRAPH*.

Held, R., & Banks, M. (2008). Perceived size is affected by blur and accommodation [Abstract]. *Journal of Vision, 8*(6):442, 442a, http://journalofvision.org/8/6/442/, doi:10.1167/8.6.442.

Hoeim, D., Efros, A., & Hebert, M. (2007). Recovering surface layout from an image. *International Journal of Computer Vision, 75,* 151–172.

Horn, B. K. P., & Brooks, M. J. (1989). *Shape from shading*. Cambridge: The MIT Press.

Howe, C. Q., & Purves, D. (2005). Natural-scene geometry predicts the perception of angles and line orientation. *Proceedings of the National Academy of Sciences, 102,* 1228–1233. [PubMed] [Article]

Joubert, O. R., Rousselet, G. A., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision Research, 47,* 3286–3297. [PubMed]

Magee, M. J., & Aggarwal, J. K. (1984). Determining vanishing points from perspective images. *Computer Vision, Graphics, and Image Processing, 26,* 256–267.

Marr, D. (1982). *Vision*. New York: W.H. Freeman and Company.

McCotter, M., Gosselin, F., Sowden, P., & Schyns, P. (2005). The use of visual information in natural scenes. *Visual Cognition, 12,* 938–953.

Myers, J. L., & Well, A. D. (2003). *Research design and statistical analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

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

Oliva, A., & Torralba, A. (2002). Scene-centered description from spatial envelope properties. In H. Bulthoff, S. W. Lee, T. Poggio, & C. Wallraven (Eds.), *Lecture notes in computer science series: Proceedings of the second international workshop on biologically motivated computer vision* (pp. 263–272). Tuebingen: Springer-Verlag.

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

Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. Cambridge: The MIT Press.

Pentland, A. P. (1984). Fractal-based description of natural scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6,* 661–674.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision, 40,* 49–71.

Proffitt, D., Bhalla, M., Gossweiler, R., & Midgett, J. (1995). Perceiving geographical slant. *Psychonomic Bulletin & Review, 2,* 409–428.

Renninger, L. W., & Malik, J. (2004). When is scene identification just texture recognition? *Vision Research, 44,* 2301–2311. [PubMed]

Rogowitz, B., Frese, T., Smith, J., Bouman, C., & Kalin, E. (1998). Perceptual image similarity experiments. In *Human Vision and Electronic Imaging III, Proceedings of the SPIE*.

Sanocki, T. (2003). Representation and perception of spatial layout. *Cognitive Psychology, 47,* 43–86.

Sanocki, T., & Epstein, W. (1997). Priming spatial layout of scenes. *Psychological Science, 8,* 374–378.

Sanocki, T., & Sulman, N. (2009). Priming of simple and complex scene layout: Rapid function from the intermediate level. *Journal of Experimental Psychology: Human Perception and Performance, 35,* 735–749. [PubMed]

Saxena, A., Sun, M., & Ng, A. Y. (2009). Make3D: Learning 3D scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31,* 824–840. [PubMed]

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

Schyns, P. G., & Oliva, A. (1999). Dr. Angry and Mr. Smile: When categorization flexibly modifies the perception of faces in rapid visual presentations. *Cognition, 69,* 243–265.

Super, B. J., & Bovik, A. C. (1995). Planar surface orientation from texture spatial frequencies. *Pattern Recognition, 28,* 728–743.

Torralba, A. (2009). How many pixels make an image? *Visual Neuroscience, 26,* 123–131. [PubMed]

Torralba, A., & Oliva, A. (2002). Depth estimation from image structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24,* 1226–1238.

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

Torralba, A., Oliva, A., Castelhano, M., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 113,* 766–786. [PubMed]

Vailaya, A., Jain, A., & Zhang, H. J. (1998). On image classification: City images vs. landscapes. *Pattern Recognition, 31,* 1921–1935.

Vogel, J., & Schiele, B. (2007). Semantic model of natural scenes for content-based image retrieval. *International Journal of Computer Vision, 72,* 2007.

Watt, S. J., Akeley, K., Ernst, M. O., & Banks, M. S. (2005). Focus cues affect perceived depth. *Journal of Vision, 5*(10):7, 834–862, http://journalofvision.org/5/10/7/, doi:10.1167/5.10.7. [PubMed] [Article]

Wu, B., Ooi, T. L., & He, Z. J. (2004). Perceiving distance accurately by a directional process of integrating ground information. *Nature, 428,* 73–77. [PubMed]

Yang, Z., & Purves, D. (2003). Image/source statistics of surfaces in natural scenes. *Network: Computation in Neural Systems, 14,* 371–39. [PubMed]

Yu, S. X., Zhang, H., & Malik, J. (2008). Inferring spatial layout from a single image via depth-ordered grouping. *Computer Vision and Pattern Recognition Workshops,* 2008. CVPRW '08 (pp. 1–7).