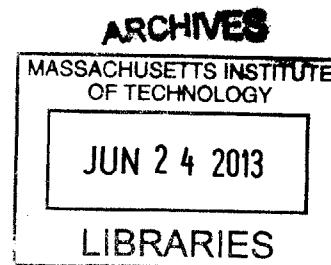


Visual Features for Scene Recognition and Reorientation

by

Krista Anne Ehinger

B.S., California Institute of Technology (2003)
B.Sc., University of Edinburgh (2007)



Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of

Doctorate of Philosophy in Cognitive Science

at the

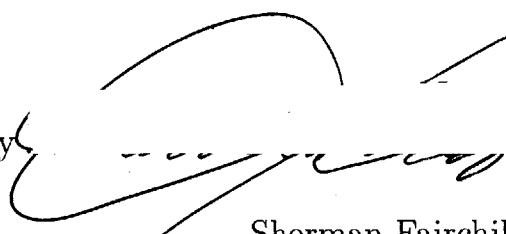
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2013

© 2013 Massachusetts Institute of Technology. All rights reserved.

Author
Department of Brain and Cognitive Sciences
May 21, 2013

Certified by
Ruth Rosenholtz, PhD
Principal Investigator
Thesis Supervisor

Accepted by 
Matthew A. Wilson, PhD
Sherman Fairchild Professor of Neuroscience
Director of Graduate Education for Brain & Cognitive Sciences

Visual Features for Scene Recognition and Reorientation

by

Krista Anne Ehinger

Submitted to the Department of Brain and Cognitive Sciences
on May 21, 2013, in partial fulfillment of the
requirements for the degree of
Doctorate of Philosophy in Cognitive Science

Abstract

In this thesis, I investigate how scenes are represented by the human visual system and how observers use visual information to reorient themselves within a space. Scenes, like objects, are three-dimensional spaces that are experienced through two-dimensional views and must be recognized from many different angles. Just as people show a preference for canonical views of objects, which best show the object's surfaces and shape, people also show a preference for canonical views of scenes, which show as much of the surrounding scene layout as possible. Unlike objects, scenes are spaces which envelope the observer and thus a large portion of scene processing must take place in peripheral vision. People are able to perform many scene perception tasks, such as determining whether a scene contains an animal, quickly and easily in peripheral vision. This is somewhat surprising because many perceptual tasks with simpler stimuli, such as spotting a randomly-rotated T among randomly-rotated Ls, are not easily performed in the periphery and seem to require focal attention. However, a statistical summary model of peripheral vision, which assumes that the visual system sees a crowded, texture-like representation of the world in the periphery, predicts human performance on scene perception tasks, as well as predicting performance on peripheral tasks with letter stimuli. This peripheral visual representation of a scene may actually be critical for an observer to understand the spatial geometry of their environment. People's ability to reorient by the shape of an environment is impaired when they explore the space with central vision alone, but not when they explore the space with only peripheral vision. This result suggests that peripheral vision is well-designed for navigation: the representation in peripheral vision is compressed, but this compression preserves the scene layout information that is needed for understanding the three-dimensional geometry of a space.

Thesis Supervisor: Ruth Rosenholtz, PhD
Title: Principal Investigator

Acknowledgments

First, I would like to thank my advisor, Ruth Rosenholtz. I am immensely grateful for her unflagging support and wise guidance, both in my research and in my career.

I would also like to thank my thesis committee members, Edward Adelson, Pawan Sinha, and Jeremy Wolfe, for all of their advice and feedback on my research, and for always challenging me to look at my results from a new angle.

I would also like to thank my collaborators on work presented in this thesis: Tobias Meilinger, who helped conduct the reorientation experiments, and Aude Oliva, who supervised the work on the canonical views of scenes.

Finally, I would like to thank my parents, Jim and Peg Ehinger, for all of their love and support. And thanks to all of the other friends, family, and colleagues who have been there for me during my years at MIT.

Contents

1	Introduction	13
2	Canonical views of scenes	19
2.1	Methods	21
2.1.1	Participants	21
2.1.2	Materials	21
2.1.3	Design	22
2.1.4	Procedure	22
2.2	Modeling the canonical view	22
2.2.1	Area map	23
2.2.2	Navigational map	23
2.2.3	Stability map	24
2.2.4	Saliency map	25
2.3	Results	25
2.4	Discussion	26
3	A summary statistic model of scene perception	29
3.1	Experiment 1	31
3.1.1	Methods	31
3.1.2	Results	34
3.2	Experiment 2	35
3.2.1	Methods	35
3.2.2	Results	37

3.3	Discussion	38
4	The role of peripheral vision in reorienting in scenes	43
4.1	Methods	45
4.1.1	Participants	45
4.1.2	Design	45
4.1.3	Materials and apparatus	45
4.1.4	Procedure	46
4.2	Results	47
4.2.1	Reorientation performance	48
4.2.2	Reorientation in artificial scenes	49
4.3	Discussion	50
5	Conclusion	53
A	Tables	55
B	Figures	59
	References	75

List of Figures

B-1	A panoramic scene with an outline around the boundary wall (above) and the spatial geometry or “isovist” computed from the scene boundaries (below). The arrow represents the same direction in each image.	60
B-2	ROC curves for each model in predicting the canonical views chosen by human observers.	61
B-3	Example of a stimuli scene (above) and its corresponding mongrel (below).	62
B-4	Comparison of responses to mongrel images and while fixating real scenes, for various scene perception tasks.	63
B-5	Comparison of responses to mongrel images and responses in the go/no-go task.	64
B-6	Mongrel vs. image results from the scene perception tasks, with crowding results from Balas, et al., 2009 and visual search results from Rosenholtz, et al., 2012.	65
B-7	Example of a visual search display (above) and its corresponding mongrel (below). Fixation is assumed to be in the center of the display. .	66
B-8	Two additional mongrels of the same visual search display. Fixation is assumed to be in the center of the display.	67
B-9	Example of a scene search display (above) and its corresponding mongrel (below). The mongrel is synthesized with fixation in the center of the search display.	68

B-10	Examples of the artificial environments used in the reorientation task: a trapezoidal environment with the shape defined by pillars (top), unconnected walls (middle), or connected walls (bottom). Xs repre- sent observers' responses, and the vertical lines represent an 18 degree threshold around the correct response (green X). An overhead diagram of each environment is shown on the right.	69
B-11	Angular error in responses in each training condition, for artificial scenes (left) and real-world scenes (right).	70
B-12	Percent of correct responses in each training condition, for artificial scenes (left) and real-world scenes (right).	71
B-13	Angular error in responses in each training condition, for each of the three artificial scene types.	72
B-14	Percent of correct responses in each training condition, for each of the three artificial scene types.	73

List of Tables

A.1	Mean AUC values for each map in predicting observers' choice of the "best" view of a scene	56
A.2	Correlations between "yes" responses to mongrels and "yes" responses to the same images in the gaze-contingent task	57

Chapter 1

Introduction

Scene perception is extremely fast: people can process the semantics of a scene at least as quickly as they can identify objects and their locations in the scene (Biederman, 1981). People can make scene judgments, such as deciding whether or not a scene contains an animal, after seeing a scene flashed for only 20 ms (Thorpe, Fize, & Marlot, 1996). It takes only a 20-40 ms exposure to a scene to extract many of its properties, such as the depth of the scene or whether it is natural or urban (Greene & Oliva, 2009; Joubert, Rousselet, Fabre-Thorpe, & Fize, 2009). What are the visual features underlying rapid scene perception?

Computer vision approaches to scene perception generally assume that processing a scene in a single glance involves extracting some type of texture representation over the entire image. One dominant algorithm, GIST, analyses a scene by calculating histograms of orientations at various spatial scales and pooling these over large patches of the image (Oliva & Torralba, 2001). This coarse representation of a scene is sufficient to extract many of the scene properties which are visible in a glance, such as the depth and navigability of the scene. It also provides information about the most likely locations of objects in a scene, and can predict where people will look first when searching for those objects (Torralba, Oliva, Castelhano, & Henderson, 2006; Ehinger, Hidalgo-Sotelo, Torralba, & Oliva, 2009). It can also be leveraged to perform more complex tasks, such as determining what city is depicted in an image (Hays & Efros, 2008). Although the GIST representation is clearly much simpler

than the representation of a scene that is extracted by a human observer, it illustrates the power of a scene representation based on a statistical summary of pooled image features. And even the current state-of-the-art approaches to scene recognition rely heavily on global texture analysis, calculating feature statistics over the entire image or over large pooling regions distributed across the image (Xiao, Hays, Ehinger, Oliva, & Torralba, 2010).

This texture-like approach to scene recognition seems to reflect an important property of scenes: they are less structured, and more texture-like, than objects. Both scenes and objects are made up of parts, and in the case of objects, the parts tend to follow a fairly rigid template: for example, a dog has four legs, a head, and a tail. There isn't much flexibility in the number or type of parts allowed, and any new configuration is usually labeled as a different object (eg, a three-headed dog). On the other hand, scenes are usually very flexible in the number and types of parts they allow. For example, a forest scene should contain trees, but they can be any species and in any configuration. A kitchen scene should contain some kitchen appliances, such as a stove, refrigerator, microwave, or dishwasher, but it need not have all of these, and they can be in various locations in the scene. On the other hand, the higher-level statistics of a scene are often central to the scene's identity. The spatial arrangement of trees is what defines the forest scene: randomly-arranged trees are a wild forest, but trees in regular rows are an orchard or garden. Spatial arrangement is also important for the kitchen scene: although it can contain various appliances, they should be lined up against the walls, not piled in the middle of the room. In this way, scenes are similar to textures: the precise locations of the individual elements are not as important as their global statistics (Portilla & Simoncelli, 2000).

Although scenes are more texture-like than objects, they are three-dimensional spaces, and therefore scene and object recognition pose many of the same problems for the visual system. One of the main issues in object recognition is how to recognize the same object from different views, since the same three-dimensional object can look very different from different angles. Similarly, scenes can look very different when explored from different directions. Being able to recognize the same place from

a new perspective is one of the most important requirements for navigating through an environment. What are the visual features that are used to recognize a place, extract information about its spatial extent and layout, and locate the observer in that space?

According to the *geometric module hypothesis*, animals (including people) navigate primarily by the spatial layout of their surroundings. Cheng (1986) observed that rats seeking food in a rectangular enclosure make errors based on the rotational symmetry of the space: they seek food in the correct corner and the diagonally opposite corner. Animals persist in these errors even when there are featural cues that could be used to distinguish between the two locations, such as a different pattern on each corner of the enclosure or a differently-colored wall on one side. Gallistel (1990) used this and other animal studies to argue that orientation in the absence of self-movement cues relies solely on the spatial geometry of the environment. Spatial geometry is computed from visible surfaces and aligned with a stored representation of the space in memory to reorient the animal within the environment. Figure B-1 shows an example of how the shape of an environment can be extracted by measuring the distances to visible surfaces. This representation is also known as an *isovist* (Benedikt, 1979), which is the architectural term for the three-dimensional volume of space visible from a single location.

The spatial geometry of an environment forms the basis of a cognitive map that is used to perform navigational tasks (Tolman, 1948; O'Keefe & Nadel, 1978). The cognitive map is a two-dimensional allocentric representation of space which encodes the geometric layout of the environment and the positions of landmarks, objects, and the observer within that environment. This map-like representation is thought to be stored in the hippocampus, encoded in *place cells*, which fire when an animal is in a particular location in space (*place field*), irrespective of other factors like the animal's current view or heading. These cells were originally observed in rodents (O'Keefe & Dostrovsky, 1971), but have also been identified in humans (Ekstrom et al., 2003). The firing rate of a place cell is a function of the animal's distance from all of the walls of its enclosure (O'Keefe & Burgess, 1996), and a network of place cells with

overlapping place fields can be used to encode locations relative to the boundaries of an enclosure, consistent with the geometric module hypothesis (Burgess, Recce, & O'Keefe, 1994).

Human beings are thought to navigate using two systems: a purely geometric system which is present in young children, and a system that incorporates surface features and landmarks, but which develops more slowly and may be tied to the development of spatial language (Lee & Spelke, 2010; Shusterman, Lee, & Spelke, 2011). Until about the age of 4, children are thought to navigate using the shape of the boundaries of a room, but not landmarks within the room or any features on the boundaries. Young children (18-24 months) make the same errors as rats when searching in rectangular rooms – they search the correct corner and the rotationally-symmetric corner, even when one wall is a different color and could be used to disambiguate between the two locations (Hermer & Spelke, 1994). Children do not seem to be able to reorient themselves to landmarks within the room, such as a rectangular or triangular arrangement of pillars (Gouteux & Spelke (2001), see also Lew (2011) for a survey of similar results in animals). However, children will reorient using a set of four free-standing walls arranged to form a rectangle with open corners, and in this type of room they make the same rotational errors as in a rectangular room (Gouteux & Spelke, 2001). Children will also reorient within rectangles defined by a 2cm-high border on the floor or a pair of 10cm-high parallel ridges on the floor, but not a rectangular mat on the floor (Lee & Spelke, 2011). Children will use landmarks (tall pillars or small boxes) to reorient in a cylindrical room only if they are set right against the wall – in this case, the landmarks apparently become part of the boundary, and their shapes are incorporated into the representation of the shape of the room (Lee & Spelke, 2010). One of the few surface features that young children do use for reorientation is texture density: in a square room with alternating patterns of smaller and larger circles, children make the same rotational errors that they make in rectangular rooms (Huttenlocher & Lourenco, 2007). In this case, children may not be recognizing the pattern at all, but using texture density to determine the shape of the room and incorrectly treating it as a rectangle.

The conclusion drawn from these results is that children and animals can only reorient to the spatial geometry of a space, as defined by its boundary walls. There are clearly some problems with this theory. Although there are easy ways to estimate distances to surfaces in some environments, for example, by using texture gradients or the retinal height at which a surface meets the ground (Gibson, 1979), this process nevertheless poses some shape-from-texture and shape-from-shading problems. In particular, the geometric module hypothesis seems to assume a very accurate shape representation of the surrounding environment, which includes “boundaries” as small as 2 cm in height, and it’s not clear how easily these could be extracted in natural environments with uneven surfaces and irregular textures. It’s also not clear why, after separating geometry from features, feature information should then be discarded (although Gallistel (1990) argues that surface features are unreliable cues for navigation because, in natural settings, they change with the seasons). Finally, the geometric module hypothesis predicts that children should be very accurate at orienting within irregularly-shaped rooms because these have unambiguous geometry, but in fact children do not seem to be able to reorient in these rooms at all (Lew, Gibbons, Murphy, & Bremner, 2010).

An alternate view is that boundary information is used for reorientation not because it is a special type of cue, but because the boundaries of an enclosure tend to provide the best visual information for navigation: far enough away to be fairly stable, but still close enough to provide changing visual information as the observer moves. In support of this are studies showing that animals can use landmarks in some situations (see Lew (2011) for a survey). In general, it seems that landmarks are more likely to be used for reorientation when they are placed closer to the walls of an enclosure, and, as with boundary information, animals use the spatial configuration of landmarks but not surface features like color. This could be taken as evidence for the geometric module hypothesis: perhaps landmarks placed near the boundary become part of the boundary representation and thus can be used for reorientation. However, this could also reflect the fact that landmarks near the boundaries of an enclosure are more stable: because the animal only sees them from a limited set of

viewpoints, these landmarks do not change position relative to each other. In contrast, the landmarks in the middle of an enclosure can present very different views as the animal moves around them – the relative positions of the landmarks in the image plane will change in different views, and the landmarks themselves may look different when seen from different angles. O’Keefe and Nadel (1978) point out that whether a set of landmarks is useful for navigation depends, in part, on their distance from the observer: more distant landmarks will not change as an observer moves, which makes them more reliable, but less useful for estimating the observer’s location.

In this thesis, I investigate how scenes are represented by the human visual system and how observers use visual information to reorient themselves within a space. Because scenes are three-dimensional entities that are experienced through two-dimensional views, they pose many of the same representational problems for the visual system as objects. But unlike objects, scenes envelope the observer, and so much of scene processing must take place in peripheral vision. Hence, an understanding of the representation in peripheral vision is important for understanding how people recognize and reorient within scenes.

Chapter 2

Canonical views of scenes

One of the primary challenges for both scene and object recognition is recognizing a particular scene or object from multiple views. Environmental spaces are three-dimensional, but we perceive them only through two-dimensional views. Being able to recognize a familiar place when we see it from a new angle is critical for navigating through the world. So it is interesting to ask how people represent spaces as views. What aspects of a scene are most important to represent in a single view?

Multi-view recognition has been studied extensively in objects. One important finding is that fact that, although people can recognize familiar objects in any orientation, there seem to be preferred or standard views for recognizing and depicting objects. These preferred views, called "canonical" views, are the views that observers select as best when they are shown various views of an object, and these are the views that people usually produce when they are asked to photograph or form a mental image an object (Palmer, Rosch, & Chase, 1981).

In general, the canonical view of an object is a view which maximizes the amount of visible object surface. The canonical view varies across objects and seems to depend largely on the shape of the object. For most three-dimensional objects (e.g., a shoe or an airplane), observers prefer a three-quarters view which shows three sides of the object (such as the front, top, and side). However, straight-on views may be preferred for flatter objects like forks, clocks, and saws, presumably because the front of the object contains the most surface area and conveys the most information about object

identity (Verfaillie & Boutsen, 1995). In addition, observers avoid views in which an object is partly occluded by its parts, and they avoid accidental views which make parts of the object difficult to see (Blanz, Tarr, & Bulthoff, 1999).

Canonical views of objects may also reflect the ways people interact with objects. People show some preferences for elevated views of smaller objects, but ground-level views of larger objects (Verfaillie & Boutsen, 1995). The ground-level views show less of the object (because they omit the top plane), but seem to be more canonical for large objects such as trucks or trains because these objects are rarely seen from above. However, these sorts of preferences may be due to greater familiarity with certain views, not functional constraints per se. Observers do not consistently select views in which an object is oriented for grasping (e.g., a teapot with the handle towards the viewer), and when subjects do choose these views, they don't match the handle's left/right orientation to their dominant hand (Blanz et al., 1999).

Scenes and places, like objects, are three-dimensional entities that are experienced and recognized from a variety of angles. Therefore, it seems reasonable to expect that certain views of a scene are more informative and would be preferred over others. However, this has not been extensively studied. Studies using artificial scenes (a collection of objects on a surface) have shown that scene learning is view-point dependent, but recognition is fastest not just for learned views, but also for standardized or interpolated versions of the learned views (Diwadkar & McNamara, 1997; Waller, 2006; Waller, Friedman, Hodgson, & Greenauer, 2009). For example, after learning an off-center view of a scene, viewers recognize the centered view of the scene about as quickly as the learned view.

There is also some evidence that there are "best" views of real-world places. Studies of large photo databases have shown that different photographers tend to select the same views when taking photos in the same location, indicating that there is good agreement on the best views of these scenes (Simon, Snavely, & Seitz, 2007). Clustering analyses of the photographs can produce a set of representative views which are highly characteristic and recognizable, but it is not clear that these are the "canonical" views in the sense of Palmer et al. (1981). For example, the most commonly

photographed view in a particular cathedral might be a close-up view of a famous statue in the cathedral, but this view would probably not be considered the “best” view of the cathedral itself, nor would it be the view people produced if they were told to imagine the cathedral.

Determining the canonical view of a scene is more complicated than finding the canonical view of an object, because in addition to rotating the view at a particular location (by turning the head), an observer can walk around within the space, obtaining different views from different locations. The current study looks at only the first part of the problem: what is the canonical view of a scene from a fixed location within that scene? To investigate this question, we use 360-degree panoramic images, which capture all of the views available from a particular location.

2.1 Methods

2.1.1 Participants

195 people participated in the experiment through Amazon’s Mechanical Turk website (www.mturk.com), an online service where workers are paid to complete short computational tasks (“HITs”) for small amounts of money. All of the workers who participated in this task were located in the United States and had a good track record with the Mechanical Turk service (at least 100 HITs completed with an acceptance rate of 95% or better). Workers were paid \$0.01 per trial.

2.1.2 Materials

The stimuli were 1084 panoramic images taken in various indoor and outdoor locations (classrooms, lobbies, chapels, parking lots, gardens, athletic fields, etc.). About half of the images (460) were downloaded from an online collection of panoramic images (www.360cities.net) and the remainder were taken in various locations around the MIT area using a camera fitted with a parabolic mirror. Each image was 3200 by 960 pixels in equirectangular projection, corresponding to 360 degrees horizontally and

about 110 degrees vertically.

2.1.3 Design

Each image was seen by 10 different participants. Participants were allowed to complete as many trials as they wished; on average, each participant performed 32 trials (median 9 trials).

2.1.4 Procedure

On each trial, participants saw one panoramic image in an interactive viewing window (this window was 550 by 400 pixels, corresponding to about 60 degrees by 45 degrees visual angle). Observers could change the view shown in the window by clicking and dragging the image with the mouse, which gave the effect of turning one's head and looking around in the scene. The initial view of the scene was chosen randomly at the start of each trial.

There were two tasks on each trial: first, type a name for the location shown in the panoramic image (for example, "kitchen"); and second, manipulate the viewer window to get the best possible view of the location. Specifically, participants were told to imagine that they were photographers trying to take the best possible snapshot of the scene, a task modeled on the photography task used by Palmer, Rosch, and Chase (1981) in their study of the canonical views of objects.

2.2 Modeling the canonical view

In addition to collecting data on the preferred views of scenes, we wished to determine how these preferred views are influenced by the visual, spatial, and semantic properties of the scene. For example, when choosing which is the "best" view of a scene, people may attempt to maximize the amount of space visible within the view, analogous to choosing a view of an object which shows as much of the object's surface as possible. Or they may consider the functional constraints of the scene, and choose views which

reflect how they would move in the space shown. These navigational views may be preferred because they are functional or because they are familiar: they are the types of views which people experience most often as they move through the environment. In addition, people may prefer stable views, which have minimal perceptual difference from other nearby views. Finally, the choice of a “best” view may be influenced by saliency: a view which contains interesting, unusual, or outlier features may be the best view to distinguish a specific place from other, similar places.

2.2.1 Area map

To characterize the shape of the space around the camera in the panoramic scene, we marked the edges of the ground plane in each image. These edges were defined by the boundaries of the scene (walls, fences, sides of buildings) and ignored small obstructions like furniture, cars, and trees. By measuring the pixel height of this edge in each image, and using the known camera height, we were able to estimate the shape of the visible space around the camera (or “isovist”), as shown in Figure B-1. This allowed us to calculate the distance to the wall in any direction around the camera (“visible depth”), the total volume of space around the camera location, and, for any particular camera view, what percentage of the total space was captured within that view. This percentage, calculated for the full 360 degrees of possible views around the camera, is called the “area map.”

2.2.2 Navigational map

To characterize the navigational affordances of the scene, we marked the walking paths in each image using an online task on Amazon’s Mechanical Turk. Workers participating in this task saw a panoramic image in equirectangular projection and were asked to place arrows on each of the paths, which included sidewalks, hallways, staircases, and navigable spaces between furniture or other obstacles. Since some images did not contain clearly defined walking paths (for example, a large, open field may not contain any marked paths and it is possible to walk in any direction), workers

were given the option to mark a checkbox (“this is a large, open space”) in addition to marking any paths that they did see in the image. Along with instructions, workers were given several examples of correctly- and incorrectly-marked images, followed by a test in which they were required to correctly mark a set of example images.

Three different workers marked the paths in each image; each received \$0.03 per image. None of the workers in this path-marking task had participated in the main experiment. A Gaussian distribution was centered on each of the marker locations in the image and the responses from the three workers were summed to create a “navigational map”. This map gives an estimate of the navigability of all possible views around the camera location: peaks in this map represent directions in which the observer could move from the camera location.

2.2.3 Stability map

To measure stability, we implemented the algorithm of Cyr and Kimia (2001). This algorithm finds stable canonical views in a horizontal circle around an object by clustering nearby views and taking the view with the minimum distance to all other views in its cluster as the "canonical" view. In order to form a valid cluster, views must be similar in some feature space and show a monotonic increase in dissimilarity with rotational distance: views which are physically farther away from the canonical view are more dissimilar than views which are closer. Our implementation used GIST features (Oliva & Torralba, 2001) to represent each view. A low-resolution pixel representation was also tried, but did not improve performance. To build the stability map from the algorithm output, we remove clusters with only one member, and for each remaining cluster, we place a Gaussian distribution over the center of each cluster’s canonical view, weighted by cluster size. Peaks in this map represent the most stable views in the image, views which are similar over a wide angle of rotation in the scene.

2.2.4 Saliency map

To represent saliency, we use the saliency model of Torralba et al. (2006), modified for spherical images. To calculate saliency over a sphere, we sample the sphere using an icosahedral grid. At each sample location, the image pixels are projected onto a tangential plane and convolved with a set of oriented Gabor filters at 4 orientations and 3 spatial scales (filter orientations are defined on the tangential plane). The three color channels (R,G,B) are handled separately. The resulting filter responses are then fit to a multivariate power-exponential distribution to determine their likelihood, as described in Torralba et al. (2006). To obtain a single saliency value for each horizontal direction around the sphere, we take the maximum saliency over the vertical direction (mean saliency was also tested, but gave worse performance).

2.3 Results

Trials were excluded if the worker did not name the location shown in the image (1% of trials) or did not use the viewer to explore the scene and simply submitted the initial view as the best view (3% of trials). In general, agreement on the “best view” of a scene was high: the average circular standard error of the angles selected by observers was 12.7 degrees. Agreement was correlated with the area of the scene ($R = 0.54$), with higher agreement in smaller, indoor spaces, and worse agreement in large, open outdoor scenes.

Model performance was assessed using ROC curves. ROC curves show the detection rate of a model relative to its false alarm rate. In this case, the ROC curves show the proportion of human observers’ “best” views which can be predicted by each map when it is thresholded at a range of threshold values. The area under the ROC curve (AUC) can be used as a measure of a model’s overall performance. A model performing at chance produces an ROC curve that is a diagonal line with an AUC of 0.5. AUC values closer to 1 indicate better model performance.

The ROC curves for each model are shown in Figure B-2, and their corresponding AUC values are given in Table A.1. As an upper bound on our models, we calculated

the agreement between observers: for each image we made a map from all but one of the observers' responses and used that to predict the left-out observer. This process was repeated for all observers and the results were averaged. The AUC for inter-observer agreement calculated in this fashion was 0.80.

Of the predictive models, the area model gives the best prediction of the views selected by observers (AUC = 0.71), while the navigational and saliency models have similar, lower performance (AUC = 0.59 and AUC = 0.61, respectively). The GIST-based implementation of the Cyr and Kimia (2001) stability algorithm performed only slightly above chance (AUC = 0.54). Although it gave a reasonable set of stable views for each image, these views do not seem to predict observers' view preference.

We also measured the performance of models which combined two or more of the above maps. To combine maps, we normalized each map so that it summed to one, and then multiplied them together. The best-performing combined map was a combination of the area and saliency maps, which only slightly outperformed the area map alone (AUC = 0.71). This performance was similar across a range of weights for the two maps. No combination of maps which included the stability map or navigational map outperformed the area map alone, which suggests that, although these maps perform above chance, they don't add any independent predictive power to the area map. The navigational and stability maps perform above chance because they often predict the same views as the area map: views which show a large amount of space tend to be navigable views, and they also tend to be more stable views.

2.4 Discussion

Just as people show clear preferences for certain views of objects, there seem to be agreed-upon "best" views of scenes. This is not surprising, given previous findings in scene research, such as the fact that people tend to use similar viewpoints when photographing famous locations. Overall, it seems that the way people choose a canonical view of a scene may be very similar to the way they select the canonical view of an object. This is because choosing the "best" view of an object or a scene

poses essentially the same problem: how to compress as much three-dimensional visual information as possible into a necessarily limited two-dimensional view.

When selecting canonical views of objects, people seem to try to maximize the amount of visible surface: they select views which show at least two sides of the object, and avoid occlusions and accidental views. Similar constraints seem to apply in scenes. The canonical view from a particular location is dependent on the shape of the space around that location: people show preferences for views that shows as much of the surrounding space as possible. It's not clear whether people choose these large-volume views because they wish to capture the space itself, or because they wish to capture the things that fill that space (objects, textures, etc.). However, the fact that visible area predicts canonical views better than saliency suggests the former: people seem to prefer views which capture the most space over views which capture the most salient objects.

There is also some evidence that the canonical view of an object reflects the way people usually see the object, or the way they interact with the object. However, our results suggest that the canonical view of a scene is not strongly based on functional constraints. Although the canonical view of a scene is often a navigationally-relevant view (a walkway, a corridor), our modeling results suggest that these views may be selected because they show a large amount of the surrounding space, not because they afford navigation. In addition, people are more likely to choose views with high saliency than non-salient views, which may maximize the amount of information shown in the view by including the most unusual or outlier features of the scene.

It may be the case that the canonical view of a scene is not the functional view. There is some evidence that people do not have a specific functional view in mind when they choose canonical views of objects (for example, Blanz et al. (1999) showed that people do not prefer views of objects oriented for grasping). On the other hand, people may consider functional constraints other than navigation when choosing a canonical view of a scene. Navigation is a very general function of scenes; most scenes also afford more specific functions (sitting in a theater, shopping in a store, etc.). If canonical views of scenes do reflect functional constraints, it seems quite likely

that they would reflect these more specific functions rather than a general function like navigation. Further work will be needed to quantify these specific functional constraints and determine how they affect view selection in scenes.

It should also be noted that there are many other factors that could affect choice of view in addition to the two factors modeled here. As noted above, people may prefer views of an environment which show a large number or large variety of the objects within that environment, and this may explain the preference for views which show a large amount of the surrounding space. People may also prefer views which show specific objects, such as ones which are central to the function or identity of a place (such as cars in a parking lot, or the stage in a theater). Aesthetics may also play a role in the selection of a “best” view of a place: people may be biased towards views which have high symmetry or are otherwise aesthetically pleasing. Many of these factors can be quantified and should be included in a full model of view preference in scenes.

Identifying the canonical views of scenes may help in understanding how scenes are represented in memory and perceptual processes. The existence of canonical views of objects has been used to argue for a viewpoint-dependent theory of object recognition, in which objects are stored in memory as a collection of typical or informative views, and recognition involves matching incoming visual information to these stored views (Edelman & Bulthoff, 1992; Cutzu & Edelman, 1994). The existence of canonical views of scenes could suggest a similar view-based representation for memory and perception of scenes.

Chapter 3

A summary statistic model of scene perception

Scene perception has generally been a problem for models of human vision developed to explain performance on more abstract tasks, such as searching within letter arrays. Scene recognition is extremely fast: in less than 100 ms, human observers can name the basic-level category of a scene (Oliva & Schyns, 1997; Rousselet, Joubert, & Fabre-Thorpe, 2005), detect if a scene shows an animal (Thorpe et al., 1996), and recognize scene attributes such as naturalness, openness, and navigability (Greene & Oliva, 2009; Joubert et al., 2009). Some scene tasks, such as determining whether or not a scene contains an animal, can be performed in peripheral vision while attention is engaged by a second, foveal task (Li, VanRullen, Koch, & Perona, 2002; VanRullen, Reddy, & Koch, 2004).

These results are surprising, since many apparently simpler tasks, such as identifying a letter in an array, seem to require focal attention. For example, when searching for a T among Ls, people seem to need to direct focal attention to the individual items in the array in order to find the T; it cannot be seen in the periphery. An influential theory to explain this and other visual search results is the Feature Intergration Theory of Treisman and Gelade (1980). According to this theory, the peripheral visual field represents only isolated feature dimensions, such color or orientation, and it has no information about feature combinations. Focal attention is needed to “bind” the

features from different dimensions to a single location to form a coherent object. According to this theory, peripheral vision can only perform tasks that involve a single feature dimension, such as finding a red square among green squares.

This might lead to the conclusion that scene tasks are easy because they involve only a single feature dimension, but this doesn't appear to be the case. Evidence for this comes from visual search tasks. Animals do not "pop-out" from animal images in a search array (VanRullen et al., 2004), nor do scene properties such as navigability guide visual search (Greene & Wolfe, 2011). And if feature-binding is required to recognize fairly simple objects such as letters, then it should be required to recognize more complex objects such as animals. So why is recognizing a scene in peripheral vision easy, when recognizing a T among Ls in the periphery is not?

Here we investigate whether performance on scene perception tasks can be explained by a summary statistic representation in peripheral vision. Previous work has shown that this representation explains performance on non-scene tasks, including visual search (Rosenholtz, Huang, Raj, Balas, & Ilie, 2012) and crowded letter recognition (Balas, Nakano, & Rosenholtz, 2009). According to this model, the peripheral visual field has a texture-like representation of the world. It measures statistics of low-level features such as luminance and orientation in pooling regions, so it knows about the distribution of features in an image, but does not know precise feature locations. This representation is different from the "unbound" features proposed by Treisman and Gelade (1980). According to the statistical summary model, peripheral vision does have some information about feature conjunctions, such as the correlation between edges in the image, but this information is statistical and pooled over a portion of the image. This representation seems to explain many results in visual search; it's sufficient to discriminate a tilted line in the presence of vertical lines, but not a randomly-rotated T among randomly-rotated Ls (Rosenholtz et al., 2012).

In these experiments, we follow the procedure introduced by Balas et al. (2009). One group of subjects performs a task with scene images. A second group of subjects performs a classification task with "mongrel" images, which have been coerced to match the feature statistics of the stimuli. As in Balas et al. (2009) and (Rosenholtz

et al., 2012), the feature statistics used in these experiments are the rich set of summary statistics proposed by Portilla and Simoncelli (2000) to represent and synthesize textures. The second group of subjects are allowed to view the mongrel images freely, with no time limit, and their performance is used as a measure of the amount of task-relevant information present in the mongrel images.

3.1 Experiment 1

In this experiment, we investigated performance on a range of scene perception tasks. We compared performance when observers fixated centrally in an image to performance on mongrel images synthesized to simulate the same fixation location. We looked at four broad categories of questions: determining whether or not a specific object (such as a car) was present in the scene, identifying the scene category, identifying the spatial layout, and determining where (e.g., in which city) a photo was taken.

3.1.1 Methods

Participants

12 participants were recruited from the Massachusetts Institute of Technology community to participate in the gaze-contingent scene perception task. All were in the 18 to 35 age range and reported normal or corrected-to-normal vision. A second group of 60 participants took part in the mongrel classification on Amazon’s Mechanical Turk service. Demographic data was not collected on these participants. All of the individuals who participated in the Mechanical Turk task were located in the United States and had a good track record with the Mechanical Turk service (at least 100 HITs completed with an acceptance rate of 95% or better). All participants gave informed consent and were paid to take part in the experiment.

Design

In the gaze-contingent task, participants were asked twenty yes/no questions about scenes. Five questions were included from each of the four question groups: presence/absence of an object, scene category or gist, road layout, and geographic location. Each question was presented as a block, in random order. In the mongrel classification task, participants were asked the same questions about mongrel versions of the images used in the gaze-contingent task.

Materials and apparatus

The stimuli for the gaze-contingent experiment were 400 photos of urban environments. The 200 images used as stimuli for the road layout and geographic location questions were collected from Google Streetview, and the 200 images used as stimuli for the object presence and scene category questions were taken from the SUN database (Xiao et al., 2010) or collected from the internet. The images used in the object presence tasks were selected so that the target object appeared in only one location in the image, and object presence was counterbalanced with scene category, so that there were an equal number of target-present and target-absent trials from each scene category. The target scene categories (“downtown,” “parking lot,” “plaza,” “residential neighborhood,” and “shopfront”) were selected from the list of scene categories in the SUN database. Geographic location and road layout was also counterbalanced, so that each road layout class appeared equally often in each city. Images were grayscale and 480 by 640 pixels in size. To ensure that participants could not use foveal information to perform the task, the center of each image was covered with a black circle 1 degree in radius.

During the gaze-contingent task, images were presented at (15 degrees by 20 degrees) on a 34 cm by 60 cm monitor, with participants seated 50 cm away in a dim room. Eye position was tracked with an Eyelink 2000 eyetracking system.

The mongrel classification stimuli were full-field mongrels generated from Gaussian noise images. The synthesis algorithm is as follows: starting at a central fixation

point, the algorithm tiles the image with square, overlapping pooling regions whose size increases with distance from fixation according to Bouma's Law (Bouma, 1970). Within each pooling region, the model computes summary statistics as described in Portilla and Simoncelli (2000). Synthesis is initiated by assuming the foveal region, a circle 1 degree in radius around the fixation point, is reconstructed perfectly. Then, moving outward, each subsequent pooling region is synthesized using the previous partial synthesis result as the seed for the texture synthesis process. The lowest-spatial frequency statistics are synthesized first and then higher spatial frequency information is added in a coarse-to-fine manner. The process iterates a number of times over the whole image. After each iteration, the foveal region and the border between the image and its background are re-imposed on the output. An example of a scene and its corresponding mongrel image is shown in Figure B-3.

Procedure for gaze-contingent task

The twenty scene-perception questions were presented in blocks of 40 images per block. At the start of each block, participants were given the question (for example, "Is this London?") and were shown two example images (for example, a picture of London and a typical distractor scene). Each trial was preceded by a central fixation cross, and the image appeared only after the participant was fixating the cross. Participants were required to maintain fixation on the center of the image; if fixation moved more than 1 degree from the center, the image was replaced with a gray mask. Image presentation time was not limited, but participants were told to respond as soon as they knew the answer to the question, by pressing 1 ("yes") or 2 ("no") on a keyboard. Each block took about 2 to 3 minutes to complete, and participants were allowed breaks after each block.

Gaze position was tracked monocularly (right eye only) at 1000 Hz. Calibration of the eyetracker was performed at the start of the experiment by having the subject fixate 9 targets with a subsequent validation. The same calibration procedure was performed occasionally during the experiment, which could be interrupted at any time for re-calibration.

Procedure for mongrel classification

Participants completed the mongrel classification task on their own computer, using a web interface on the Amazon Mechanical Turk website. Participants were told that the purpose of the study was to determine how well people could recognize images “distorted by digital noise” and were shown examples of images with their corresponding full-field mongrels. In each task, participants were given a single question (for example, “Is this London?”) and were shown mongrel versions of the 40 images which had been used as stimuli for that question in the gaze-contingent experiment. Participants were allowed to study each image for as long as they wished, and then clicked one of two buttons to indicate “yes” or “no”. Participants received feedback after each response. Questions were randomly assigned to participants, and each participant could complete as many as she wished (up to twenty).

3.1.2 Results

One participant in the gaze-contingent task reversed the response keys during one block, so this block of data was dropped from analysis. Average accuracy in each block was calculated for each subject in the gaze-contingent and mongrel classification tasks. Average accuracy on each task is shown in Figure B-4. For the most part, accuracy in the mongrel classification task is very similar to accuracy in the gaze-contingent task: questions that are difficult to answer with mongrel images are also difficult to answer in a single fixation on a scene. The main exception were the tasks which asked observers to detect small, eccentric objects in the scenes: this task seems to be consistently easier in real scenes than would be predicted by mongrel performance. Bonferroni-corrected t-tests were used to compare gaze-contingent and mongrel performance in each task; these differences were significant only for three of the object-detection tasks: detecting a car ($t(22) = 4.21, p < 0.05$), detecting a person ($t(22) = 10.3, p < 0.01$), and detecting a street sign ($t(22) = 18.5, p < 0.01$).

In addition to showing similar overall accuracy, responses to the individual images in the mongrel and gaze-contingent versions of each task were highly correlated for

most of the tasks. The correlations for each task are shown in Table A.2. For the scene category, layout, and location tasks, responses to the individual images in the gaze-contingent experiment were well predicted by responses to the mongrel images. However, correlations were lower for the object-detection tasks.

3.2 Experiment 2

In this experiment, we investigate whether the statistical summary model of peripheral vision can predict performance on two commonly-investigated rapid-perception tasks: a go/no-go animal detection task and a go/no-go vehicle detection task. Previous experiments have shown that both animal/non-animal and vehicle/non-vehicle discrimination can be performed in the periphery without attention (Li et al., 2002).

3.2.1 Methods

Participants

24 participants were recruited from the Massachusetts Institute of Technology community to participate in the go/no-go task. All were in the 18 - 35 age range and reported normal or corrected-to-normal vision. A second group of 24 participants took part in the mongrel classification task on Amazon’s Mechanical Turk service. Demographic data was not collected on these participants. All of the individuals who participated in the Mechanical Turk task were located in the United States and had a good track record with the Mechanical Turk service (at least 100 HITs completed with an acceptance rate of 95% or better). All participants gave informed consent and were paid to take part in the experiment.

Design

In the go/no-go task, participants were asked to identify a target class (“animal” or “vehicle”) in rapidly-presented images. Target class was manipulated between-participants. In the mongrel classification task, participants were asked to sort mon-

grel versions of the go/no-go images into target and non-target (“animal”/“not-animal” or “vehicle”/“not-vehicle”, between-participants).

Materials

The go/no-go stimuli were a randomly selected subset of the images used by Li et al. (2002). The target images for the animal-detection task were 240 scenes containing animals (including mammals, birds, reptiles, fish, and insects). The target images for the vehicle-detection task were 240 scenes containing vehicles (including cars, trains, boats, planes, and hot-air balloons). The distractor set for each task included 120 images from the other target category, plus 120 scenes which contained neither vehicles nor animals (which included images of plants, food, landscapes, and buildings). During the go/no-go task, images were presented in grayscale at 384 by 256 pixels (8.9 degrees by 6.0 degrees) on a 34 cm by 60 cm monitor, with participants seated 75 cm away in a dark room.

The mongrel classification stimuli were full-field mongrels generated from noise using the algorithm described previously with a few variations. Instead of pooling in square image regions, features were pooled and synthesized in elliptical regions oriented along lines radiating out from the fixation center, with the longer (radial) dimension equal to the width/height of the corresponding square pooling region. Because the elliptical pooling regions were narrower than the square regions, the angular overlap of regions was increased, and the radial overlap was decreased. The image/background border was not enforced during synthesis, and therefore a much larger number of iterations were used at each scale to achieve convergence. Mongrels were synthesized to simulate fixation at either the image center or 11 degrees left or right of center, to match the viewing conditions of the go/no-go task.

Procedure for go/no-go task

Participants were instructed to hold down the left mouse button throughout the experiment. At the start of a trial, a central fixation cross appeared for 300 +/- 100 ms, and was followed by an image presented for 20 ms. The image appeared either

at the center of the screen or left or right of the fixation (center of the image at 11 degrees eccentricity); each position occurred equally often. If the image contained a target (animal or vehicle), participants were to respond by releasing the left mouse button as quickly as possible. Participants made no response to non-target images. Participants were given 1000 ms to make their response.

Participants completed 10 blocks of 48 trials, with a break after each block. Each block contained an equal number of target and non-target images, and an equal number of images in each of the three presentation locations (left, center, and right).

Procedure for mongrel classification

Participants completed the task on their own computer, through the Amazon Mechanical Turk website. Participants were told that the purpose of the study was to determine how well people could recognize images “distorted by digital noise” and were shown examples of images with their corresponding mongrels. The experiment consisted of 480 trials which exactly matched one the 24 sessions of the rapid perception experiment. On each trial, participants were shown a mongrel version of an image from the rapid perception task. Mongrel images were always presented in the center of the screen, but had been synthesized to simulate the image’s position in the rapid perception task: left of, right of, or at fixation. Participants were allowed to study each image for as long as they wished, and then responded with a key press to indicate whether or not the mongrel corresponded to the target category for the experimental session (“animal” or “vehicle”). Participants received feedback after each response.

3.2.2 Results

Accuracy in the go/no-go task was averaged across subjects. For both target types, accuracy for centrally-presented images was 94%, and accuracy for peripherally-presented images was 74% for animal targets and 76% for vehicle targets. Performance was considerably lower for the mongrel images: 85% and 85% correct for animal and

vehicle detection, respectively, in mongrel images simulating central presentation, and 60% and 62% correct for mongrels simulating peripheral presentation.

Despite the overall lower performance in the mongrel classification task, responses to individual images in each task were correlated. A comparison of the percentage of “target” responses in each task is shown in Figure B-5. Mongrel images which were more frequently classified as targets (“animal” or “vehicle”) were more likely to be detected in the go/no-go task, in both central and peripheral presentation.

3.3 Discussion

Performance on “mongrel” classification tasks predicts performance on a range of scene perception tasks. This means that the information in the mongrels – a set of feature statistics collected over local pooling regions – is sufficient for many scene perception tasks, such as determining whether a scene is a residential or city street, whether a road turns left or right, or whether or not a scene depicts an animal.

In addition to predicting performance on scene perception tasks, this mongrel image approach has been shown to predict performance on other peripheral visual tasks. Figure B-6 shows the results of previous experiments on crowding (Balas et al., 2009) and visual search (Rosenholtz et al., 2012) overlaid on the results of the scene perception tasks from Experiment 1. These tasks used a slightly different approach, creating “mongrel” images for single patches of the peripheral visual field, instead of simulating the entire image. However, this comparison shows that, across a wide range of tasks, the feature statistics of the mongrel images predict what observers will be able to see in their peripheral vision.

So why are scene tasks, such as recognizing an animal, generally easy in the periphery, while other peripheral tasks such as spotting a letter in a visual search display are generally difficult? The reason is that the summary statistic representation in the periphery preserves the information needed for many scene tasks, but this information is not sufficient for many visual search tasks. For example, consider the search display shown in Figures B-7 and B-8. Figure B-7 shows a visual search display

consisting of randomly-rotated Ls; the task is to find a single randomly-rotated T. Also shown in Figures B-7 and B-8 are three mongrel images with the same statistics as the search display. In one of these mongrels (the lower half of Figure B-7), the T is clearly visible in its true location, but in the other two mongrels (Figure B-8), there is an L in the T's location, and T-like symbols appear in other parts of the display. This means that the statistics of these peripheral patches do not contain enough information to discriminate between a patch of Ls with a T and a patch of Ls with no T, so a search for a T among Ls in this type of display is a moderately difficult task. People cannot spot a T in a peripheral patch – they will need to fixate quite close to this target in order to see it and know it is present. (Note that the Ls nearest the central fixation in each mongrel are reproduced quite faithfully – in the small pooling regions near fixation, the statistical summary representation is sufficient to make out individual letters, but this is not the case in the periphery.)

Although an observer looking at these mongrels might not know exactly where the T had been in the original display, he could nevertheless answer many “scene perception” questions about the original display just by looking at the mongrels. From the mongrels, it's obvious that the original display was a regularly-spaced array of white symbols on a black background, and that these symbols were randomly-rotated Ls or something very similar. This type of scene summary or “gist” is very well conveyed by the texture-like summary statistic representation, and this is true regardless of whether the scene is a real-world scene or an artificial scene such as a letter array.

Although mongrel images predict performance on a wide range of tasks, they do not seem to predict performance on the object-detection tasks of Experiment 1, in which observers were asked detect a small, eccentric object in a larger scene. Many of these tasks seem to be easier when fixating in a real image than would be predicted by mongrel classification performance. One reason for this is that the feature pooling in the mongrels is very likely to destroy the shape of an object – the object features will still be present in the mongrel, but broken apart and rearranged. These statistics may, in fact, contain enough information to say that the object was

present in the original, but because the scrambled object doesn't look much like the original object, people are unwilling to say the object was present. At the same time, random conjunctions of other objects' features in a target-absent image can sometimes produce something that looks very much like the target object. It may be possible to show a better correspondence between mongrel and scene performance by showing observers multiple mongrels per image, which gives a better sense of the statistical nature of the mongrel output and makes it more obvious which features signal true objects and which are accidental.

The fact that mongrel animal images can be distinguished from mongrel non-animals does not mean that search for animal images among non-animal distractors should be an easy pop-out search. And the same is true for other types of scene search tasks – these results do not necessarily mean that a particular scene category or layout would “pop-out” among scene distractors. When multiple images are presented in a search display, features of two neighboring images may be pooled, masking the target or creating illusory targets. The image borders and the spaces between images would also be pooled with the image features, creating a new set of statistics which would complicate target detection. These effects can be seen in Figure B-9, which shows a “mongrel” version of a scene search display.

One concern with the mongrel approach is that the mongrel images could contain too much information. At the extreme, if the “mongrel” images were simply identical to the original images, they should predict performance on all of the easy scene tasks. However, if this were true, we would expect performance on the mongrel images to be much higher than fixating performance on the harder scene tasks, such as the tasks that required people to detect small objects in the scene or to identify the city, and this was not the case in our experiment. Even so, the mongrels may contain more information than needed to do the easier scene tasks, such as identifying scene category or layout: previous work has shown that this information can be extracted from a much coarser representation of the scene (Oliva & Schyns, 1997; Ross & Oliva, 2010). However, the goal of these experiments is not just to find the minimum set of features which can predict performance on scene tasks, but rather to understand

the representation that the visual system has of an image in the periphery. The summary statistic representation described here can predict performance on a range of peripheral visual tasks, including crowding and visual search, while a coarser model might be sufficient for some scene perception but would not predict performance on these other tasks.

The feature statistics which are represented in the mongrel images may or may not be the feature statistics which are represented in the human visual system. There is some evidence that the visual system does, in fact, compute these statistics: Freeman and Simoncelli (2011) have shown that, with an appropriate pooling scheme, mongrel images can be synthesized that are indistinguishable from the original, and they use this approach to argue that this set of statistics forms the representation of images in an early level of the visual system (specifically, in V2). However, other feature sets have been proposed. Crouzet and Serre (2011) have shown that the HMAX model, which was designed to simulate the early stages of the primate visual system (see Serre, Oliva, and Poggio (2007) for details), predicts human performance on a rapid animal/non-animal categorization as well as or better than a model based on the Portilla and Simoncelli (2000) texture statistics. However, these two models have not been compared on a wide range of tasks, so further work will be needed to determine what feature set best matches human perception.

These scene perception results, in combination with previous work, provide strong support for a summary statistic representation in peripheral vision. The peripheral visual system computes a rich set of summary statistics over some feature space, within pooling regions that distributed across an visual field. This representation is sufficient to convey the “gist” of the scene, including the scene category, spatial layout, and some location information, but it is not sufficient to perform tasks that require fine-grained localization, like discriminating letters in a crowded display.

Chapter 4

The role of peripheral vision in reorienting in scenes

When navigating in the real world, our focal attention is often occupied by other tasks such as looking at a person or object in our surroundings, reading a sign, or texting on a cellphone. Even when we don't put effort into interpreting the layout of our surroundings, we have an effortless sense of where we are in space, where we've been, and where we can go from our present location. Because the fovea only covers a small part of the visual field, most of the scene around an observer is being processed by peripheral vision. But how important is peripheral vision for understanding and reorienting in space?

Reorientation is thought to be guided primarily by the shape of the surrounding space: people extract the shape of their surroundings from boundary walls and locate themselves within that shape. It's reasonable to think that peripheral vision would play a role in extracting shape information, since much of the boundary surface would be in the periphery, and scene layout information seems to be very easily visible in the periphery. However, it is not clear that peripheral vision is necessary for extracting shape information: people may be able to learn the layout of their surroundings just as easily by exploring the scene with central vision.

Work in virtual reality and visual prosthetics suggests that navigation and reorientation tasks become more difficult with a narrower field of view. van Rheede,

Kennard, and Hicks (2010) compared performance on a virtual navigation task under two viewing conditions: observers could either view the entire screen or a patch just around their fixation point. The image was downsampled so that the resolution was matched in both cases (and in both cases it was very low, only 30 x 30 samples, to simulate the experience of wearing a visual prosthesis). Wayfinding performance was significantly better with the wider field of view: participants navigated more quickly and chose a more direct path.

In a real-world navigation task, (Rousek & Hallbeck, 2011) asked subjects to find their way around a hospital while wearing goggles that simulated various visual impairments. One of their findings was that people with limited peripheral vision (simulated glaucoma) were significantly worse at navigating the hospital than subjects whose simulated disorders spared peripheral vision. When peripheral vision was impaired, people had more trouble avoiding obstacles in their path, but they also seemed to make more wayfinding errors: they lost their way or mistook their location for another location in the hospital. Similarly, studies which asked people to navigate while wearing blinders have shown that real-world navigation performance is impaired when peripheral vision is blocked, although these tasks tend to look more at obstacle avoidance than at reorientation or wayfinding (Toet, Jansen, & Delleman, 2007, 2008).

Although previous studies have investigated the role of reduced peripheral vision, no previous work has directly compared the role of the central and peripheral visual fields in a reorientation task. In this study, we ask people to reorient themselves within an immersive virtual environment that provides nearly the full field of view that is available in the real world. We investigate how well people can localize themselves in a virtual environment when they have information only from their central visual field or only from their peripheral visual field. In addition, we compare reorientation performance in real-world scenes, where both landmarks and spatial geometry may be used for reorientation, and in artificial environments where people must reorient by spatial geometry alone, and we investigate whether central and peripheral visual information play different roles in these two types of reorientation tasks.

4.1 Methods

4.1.1 Participants

24 participants (age 18 - 40) were recruited from the Max Planck Institute Tübingen community. All reported normal or corrected-to-normal vision. Participants gave informed consent and were paid for their participation.

4.1.2 Design

The experiment consisted of 45 trials, which included 9 artificial scenes interleaved with 36 real-world scenes. Each trial appeared in one of three training conditions: full view, central vision only, or peripheral vision only. Trials cycled between the three conditions in that order.

4.1.3 Materials and apparatus

Artificial scenes were created in Blender, and consisted of two room shapes (kite-shaped or trapezoidal) defined by three different types of structures (connected walls, free-standing walls, or pillars). Examples of each of these room types are shown in Figure B-10. In each artificial scene, spherical panoramic images were rendered at the test location (which was always the center of the room) and training locations (various corners of the rooms). Stimuli for the real-world scene trials were collected from Google Streetview. These photos had been taken from a bicycle-mounted camera about 1.7 meters off the ground. Training and test views were selected from nearby locations (median 17 meters apart; range 10 to 63 meters). The panoramic images used as stimuli in the experiment subtended the entire visual field (360 degrees horizontally and 180 degrees vertically) and were saved in equirectangular projection at a resolution of 8192 by 4096 pixels and 24-bit color.

During the experiment, images were projected onto a wide-area, half-cylindrical screen. Participants were seated at the center of the screen and head position was fixed with a chin rest. In this position, the viewing screen subtended 220 degrees

horizontally and about 63 degrees vertically (23 degrees above fixation and about 40 degrees below). Participants used a game controller to make their responses and to manipulate the view shown on the screen: moving the analog stick on the controller rotated the view, as though the observer were turning around in the scene. A camera mounted below the chin rest was used to monitor eye position, and a tin clicker was used to signal that the participant had broken fixation.

4.1.4 Procedure

Three practice trials preceded the experiment, one practice trial for each training condition. Each trial consisted of a training and test phase. In the training phase, a spherical panoramic image was projected onto the screen, and participants were given 40 seconds to explore the scene by using the analog stick on the game controller to change the view. In all training conditions, subjects were asked to maintain fixation on a cross projected into the center of their field of view. In the full-view training condition, the scene was unobstructed. In the central-only training condition, subjects viewed the scene through a central, circular window of radius 20 degrees while the rest of the image was masked. In the peripheral-only training condition, this was reversed and a circular mask of radius 20 degrees obscured the central portion of the scene.

Participants could terminate the training phase by pressing a button on the game controller; after 40 seconds this phase terminated automatically. The image was replaced by noise for 1 second, and then the participant was shown a second panoramic photo, taken from a location near the training image's location. A cross was projected into the center of the participant's field of view, and the participant was asked to rotate the view to place the cross on the ground at the point where they had apparently been standing during the training phase. During the test phase, participants were permitted to move their eyes, and there was no time limit on responses. In both phases, view rotation was limited to 12 degrees per second to counteract motion sickness.

4.2 Results

Trials were dropped if the participant pressed the response key too quickly (1% of trials) or made a saccade during the trial (9% of trials). Dropping saccades eliminated more than 90% of the peripheral-only trials for two subjects, so these two subjects were dropped entirely from the analysis.

Ground truth correct answers had to be determined manually for the real-world images from Google Streetview. This was done by annotating a set of 25 - 50 matching point pairs in the two images and using RANSAC to find the subset of points with the best alignment and identify the fundamental matrix relating the two views (Hartley & Zisserman, 2004). The fundamental matrix can be decomposed to find the vector from the center of the test panorama to the center of the training panorama (the location of the camera which took the training view image). Assuming that the ground plane is flat, the vector to the point on the ground below the training camera can be determined from a single pair of matched ground points in the two views. We performed this calculation with multiple pairs of ground points in each image and took the median result as the ground truth location of the point on the ground below the training camera.

Reaction times in the test phase were 28 seconds in full-view training condition, 29 seconds in the central-only condition, and 30 seconds in the peripheral-only condition. These differences were not significant ($X^2(2) = 3.08$, $p = 0.21$). However, there was a significant difference in study time across the three conditions ($X^2(2) = 9.75$, $p < 0.01$). This difference is only significant in pairwise comparisons between the central-only and peripheral-only condition: people spent slightly more time studying the scene when they explored with their central vision only (average 36 seconds) than when they had peripheral vision only (average 33 seconds). People were more likely to quit the learning phase before it timed out in the full-view and peripheral-only conditions (48% and 53% of trials) than in the central-only condition (41% of trials), although these differences are not significant ($X^2(2) = 5.63$, $p = 0.06$).

4.2.1 Reorientation performance

One measure of the accuracy on this task is the angular error between the participant's response and the correct location. The participant's response and the correct response are treated as vectors on a unit sphere, and the angular error is the angle between these vectors. This measure of accuracy is shown in Figure B-11. Overall, angular error was much lower when people reoriented in real-world scenes than when they reoriented in artificial scenes, where the only cue for reorientation was the shape of the room. Angular error was 46 degrees, 54 degrees, and 35 degrees for the full view, central-only, and peripheral-only learning conditions in artificial scenes. In the real-world scenes, angular error was 21 degrees, 23 degrees, and 28 degrees for the same conditions.

A 2 (scene type) x 3 (training condition) repeated measures ANOVA showed a significant main effect of scene type ($F(1,42) = 28.0, p < 0.01$), but no main effect of training condition ($F(2,42) = 1.4, p = 0.26$). However, there was a significant interaction between scene type and training condition ($F(2,42) = 7.1, p < 0.01$). Bonferonni-corrected t-tests were used to compare training conditions within each scene type. These comparisons showed a significant difference between the central-only training condition and the peripheral-only training condition in the artificial scenes ($t(21) = 2.9, p < 0.05$). No other comparisons were significant. In the artificial scenes, people were significantly more accurate in reorienting themselves after exploring the scene with peripheral vision only than when they had explored the scene with central vision only.

In addition to looking at the distance between the participant's response and the correct location, we ran another analysis in which we scored each response as correct or incorrect. In this analysis, we only looked at whether the observer had gotten the direction to the training location roughly correct, and we ignored their estimate of the distance to the training location. Responses were marked correct if they were within 18 degrees horizontally of the true location; all other responses were marked incorrect. (Other thresholds were tested, but they gave the same results and are not

reported here.) Figure B-10 shows some example scenes with observers' responses and the threshold for correct responses, while Figure B-12 shows the proportion of correct responses in artificial and real-world scenes. In the artificial scenes, people were able to reorient themselves correctly 48% of the time when they had studied the scene with their full visual field, 28% of the time when they had studied with central vision only, and 51% of the time when they had studied the scene with peripheral vision only. People were more accurate in real-world scenes, reorienting correctly 75%, 67%, and 69% of the time in the same learning conditions.

A 2 (scene type) x 3 (training condition) repeated measures ANOVA showed a significant main effect of scene type ($F(1,42) = 29.6, p < 0.01$), and a significant main effect of training condition ($F(2,42) = 5.4, p < 0.01$). There was also a significant interaction between scene type and training condition ($F(2,42) = 3.3, p < 0.05$). Bonferonni-corrected t-tests were used to compare training conditions within each scene type. As before, these comparisons showed a significant difference between the central-only training condition and the peripheral-only training condition in the artificial scenes ($t(21) = 2.5, p < 0.05$). No other comparisons were significant. In artificial scenes, where shape is the only cue for reorientation, people are better at reorienting themselves after exploring the scene with their peripheral vision only than when they explore with central vision only.

4.2.2 Reorientation in artificial scenes

We also investigated reorientation performance in the three types of artificial spaces, which were defined by pillars, unconnected walls, or connected walls. This analysis was done within images, because it had not been possible to counterbalance the artificial scene types with training condition for each subject. The angular error in each condition across the three types of artificial scenes are show in Figure B-13. A 3 x 3 ANOVA looking at the angular error across room type and training condition showed a significant main effect of training condition ($F(2,12) = 11.9, p < 0.01$) but no main effect of room type ($F(2,12) = 0.67, p = 0.55$). There was a significant interaction between condition and room type ($F(2,12) = 4.7, p < 0.05$).

Bonferroni-corrected pairwise comparisons between the individual room types within each condition were not significant, although it does appear that the interaction may be due to a difference in the peripheral-only condition: observers may have been better at reorienting in the rooms with connected walls than in other scene types.

The proportion correct in each room type by condition is shown in Figure B-14. A 3x3 ANOVA looking at the proportion of correct responses across room type and training condition showed a significant main effect of training condition ($F(2,12) = 16, p < 0.01$) but no main effect of room type ($F(2,12) = 0.86, p = 0.47$). There was a significant interaction between condition and room type ($F(2,12) = 6, p < 0.01$). Again, the Bonferroni-corrected pairwise comparisons between the individual room types within each condition were not significant, but it appears that observers were more accurate in the connected-wall rooms when reorienting with peripheral vision only, while accuracy in the full view and central-only conditions were similar across room types.

4.3 Discussion

People were better at reorienting in artificial scenes that they had explored with peripheral vision only than they were at reorienting in artificial scenes that they had explored with central vision only. When people are asked to locate themselves within an impoverished environment, where the shape of the surrounding walls provides the only cue to location, they seem to be better at extracting that shape information with peripheral vision than with central vision.

However, there was no difference in reorientation performance when people studied real-world scenes with only their peripheral vision, only central vision, or both. This indicates that peripheral visual information is important for reorientation by spatial geometry alone, but it isn't necessary when other, non-geometric cues are available. In the real-world scenes, observers could orient themselves using landmark objects within the space and the surface features of the boundaries (a distinctive set of windows on one wall of a building, for example), in addition to the spatial geometry of the scene.

Central vision alone seems to be sufficient to extract and use these non-geometric cues.

It's not surprising that people can extract some scene layout information from peripheral vision alone. As discussed previously, scene layout is readily extracted from a single fixation on a scene, and it is easy to see in a crowded "mongrel" representation of a scene. In fact, coarse layout information can be extracted from orientation information even when orientations are pooled over quite a large portion of the scene, for example into four quadrants around the image center (Ross & Oliva, 2010). So scene layout information is not likely to be lost to crowding, and it should be possible to determine the spatial geometry of a scene from peripheral vision alone.

More surprising is the fact that it is actually easier to reorient by spatial geometry with only peripheral visual information than it is with only central visual information. However, it is probably easier for the brain to extract the shape of a surrounding space when a larger portion of the space is visible in a single view. In the peripheral-only condition, observers could see a nearly 180-degree view of the space. Even though the central 40 degrees of the scene were obscured, this view would allow people to see two or three of the walls of a room simultaneously, which probably made easier to work out their relative positions. In the central-only condition, observers could only see a small part of the room at one time and had to figure out the relative positions of structures by moving their gaze and trying to relate the two views. There is probably some noise involved in this process, particularly in this virtual reality task where people could not use normal proprioceptive cues from eye and head movement to determine the distance between two views. This would make it much more difficult to work out the shape of the surrounding space, resulting in more errors when people are asked to use that shape information to reorient themselves.

However, in natural scenes, people do not orient by shape alone: they have access to many other cues, including landmark objects within the space and distinctive features on the walls of the space. When these cues are present, people can orient just as easily with central vision alone as with their full visual field or with peripheral vision only. It's possible that people use different strategies in these environments: reorient-

ing by spatial geometry when only peripheral vision is available, and by landmarks when only central vision is available. Or people might adopt a similar strategy in both viewing conditions, reorienting by non-shape cues such as a distinctive pattern on the boundary walls, or a large landmark object. Further experiments would be needed to distinguish between these possibilities.

Finally, it is interesting to note that we found some difference in reorientation accuracy in the three different artificial scene types. Young children are known to perform differently in these different types of spaces: they are better at reorienting in rooms with connected or free-standing walls than in rooms with a shape defined by pillars (Gouteux & Spelke, 2001). Adults typically reorient well in all three room types. Gouteux and Spelke (2001) take this as evidence that young children navigate by room shape as defined by boundary walls, and the ability to navigate by landmark objects such as pillars develops later in life. However, this study shows that when the task is made sufficiently difficult, even adults reorient more easily in a shape defined by connected walls than in a shape defined by unconnected landmark objects. Previous work with adults and children has used real-world orientation tasks which allow the use of many cues to determine the shape of the space, including optic flow and self-movement in the space, while this task required observers to reorient using visual information only from a single position in the space. The fact that we find reorientation differences in adults suggests that it may actually be easier, visually, to determine the shape of a space defined by connected walls than a shape defined by landmarks such as pillar. Children's difficulty in reorienting by landmarks may not due to an innate navigational system which can only use boundary information; rather, they may simply have more trouble integrating the various visual and motion cues that adults use to understand the shape of their environment.

Chapter 5

Conclusion

Scenes and objects pose many of the same problems for the human visual system: scenes, like objects, are three-dimensional entities that we must be able to recognize from many different two-dimensional views. In order to navigate through the world, we must be able to extract and represent the three-dimensional shapes of the spaces around us. Peripheral vision plays an important role in this process, because scenes, unlike objects, surround the viewer and thus are mainly viewed in the periphery.

The visual system can extract a great deal of information about a scene from a crowded, peripheral visual representation. A summary statistic representation of the scene, which captures the distribution of features but not their precise locations, is sufficient to determine the basic-level scene category and the spatial layout of the scene. It also carries enough detail to allow an observer to guess the geographic location of a scene and detect some larger objects in the scene.

The spatial geometry of a scene, as defined by its boundaries, plays a particularly important role in scene representation. Just as the canonical view of an object is the one that shows the object's surfaces as well as possible to give a sense of its shape, the canonical view of a scene seems to be the one that shows as much of the area as possible and gives a sense of the spatial geometry. Representing the spatial layout of the scene in a single canonical view may be particularly important because this layout information is what people use to orient themselves within a space.

The spatial geometry of a scene is easily extracted from peripheral vision because

it is defined over large boundary surfaces, such as walls, buildings, and ground. Extracting layout information from these boundary surfaces is mostly a matter of texture segmentation: determining where two walls meet or finding the edge between the walls and the floor. The texture-like representation in peripheral vision proposed by the summary statistic model is well-suited for this purpose. The fact that spatial geometry is easily processed in peripheral vision may be why it is so important for navigation: this information is available at all times in the periphery, even when central vision is focused on individual objects in a scene. And conversely, landmark objects and surface features may be relatively less useful for navigation because their precise locations and configuration details are more likely to be lost to crowding in peripheral vision.

Instead of being a disadvantage, a crowded, texture-like representation in peripheral vision may actually be a beneficial adaptation for navigation. There is a limit on how much information the visual system can process with reasonable speed: the brain can't afford to process the entire visual field with the same high fidelity as is available in the fovea. One option for dealing with that bottleneck might be to process a smaller field of view with high resolution; the other option is to process a wide-angle view with lower resolution. A wide-angle field of view is extremely useful for representing the three-dimensional shape of an environment: it's easier to understand the relative positions of two structures when both can be seen at once in a single view. So rather than reduce the field of view, the human visual system has opted for a compressed, summary statistic representation in peripheral vision. This representation doesn't preserve the exact locations of features in the periphery, but it captures the scene layout information that is necessary for understanding the three-dimensional geometry of a space.

Appendix A

Tables

Table A.1: Mean AUC values for each map in predicting observers' choice of the "best" view of a scene

0.80	Inter-observer agreement
0.71	Area map
0.61	Saliency map
0.59	Navigational map
0.54	Stability map

Table A.2: Correlations between "yes" responses to mongrels and "yes" responses to the same images in the gaze-contingent task

0.47	bike
0.73	car
0.58	fire hydrant
0.34	person
0.41	sign
0.86	downtown street
0.97	parking lot
0.95	plaza
0.94	residential street
0.96	shopping street
0.93	4-way intersection
0.68	left turn
0.73	right turn
0.86	no turn
0.95	T intersection
0.69	Europe
0.76	London
0.67	Los Angeles
0.75	NYC
0.75	Paris

Appendix B

Figures

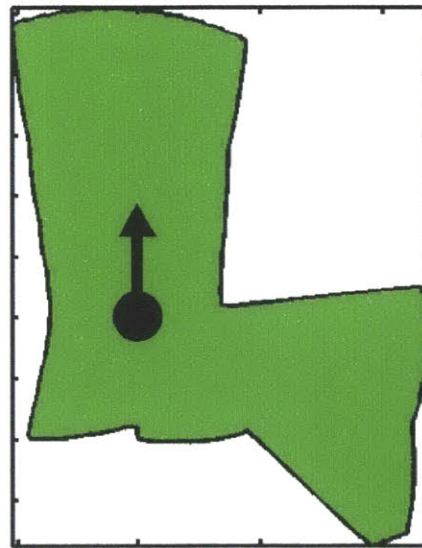


Figure B-1: A panoramic scene with an outline around the boundary wall (above) and the spatial geometry or “isovist” computed from the scene boundaries (below). The arrow represents the same direction in each image.

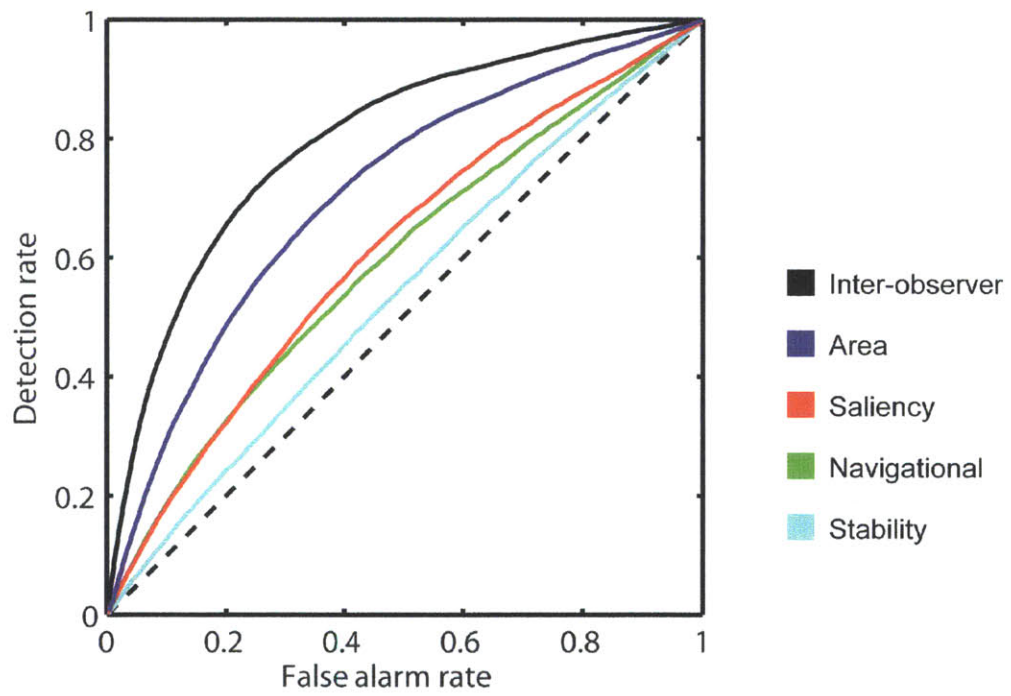


Figure B-2: ROC curves for each model in predicting the canonical views chosen by human observers.

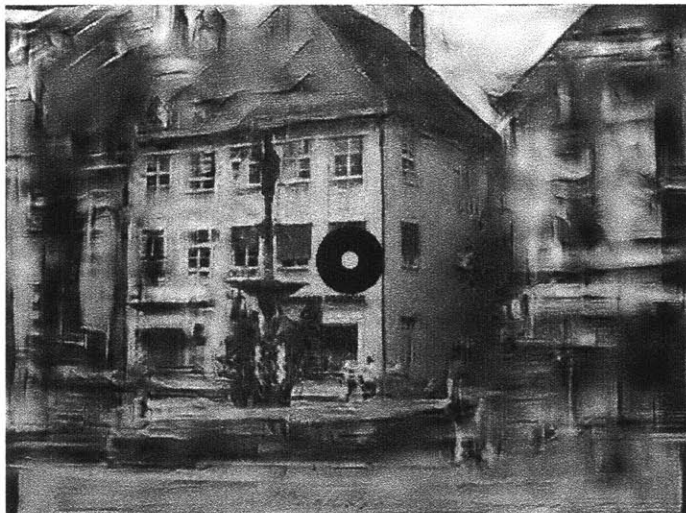
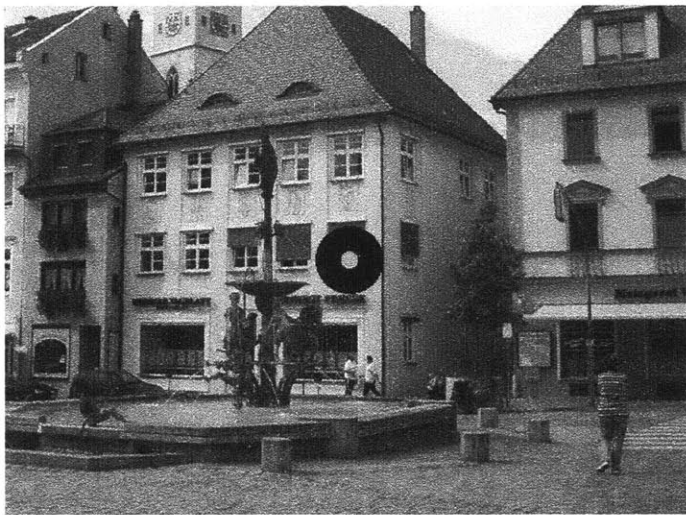


Figure B-3: Example of a stimuli scene (above) and its corresponding mongrel (below).

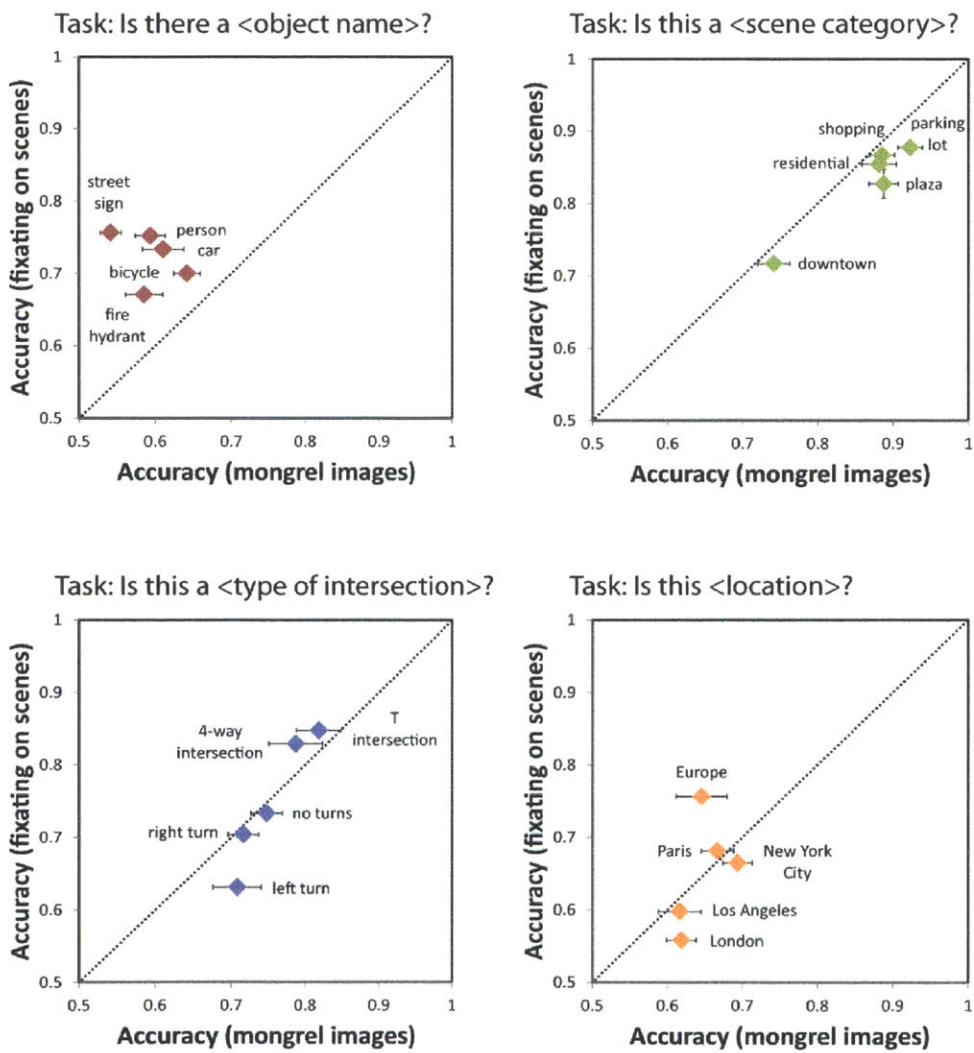


Figure B-4: Comparison of responses to mongrel images and while fixating real scenes, for various scene perception tasks.

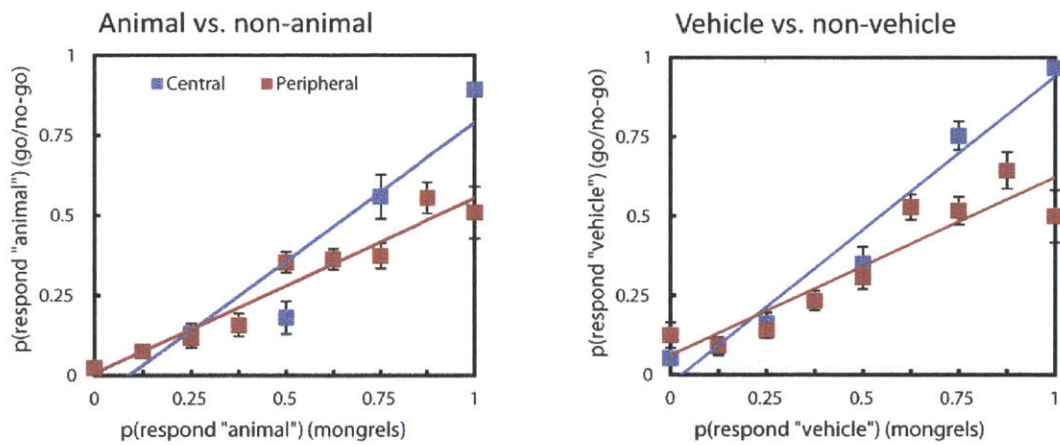


Figure B-5: Comparison of responses to mongrel images and responses in the go/no-go task.

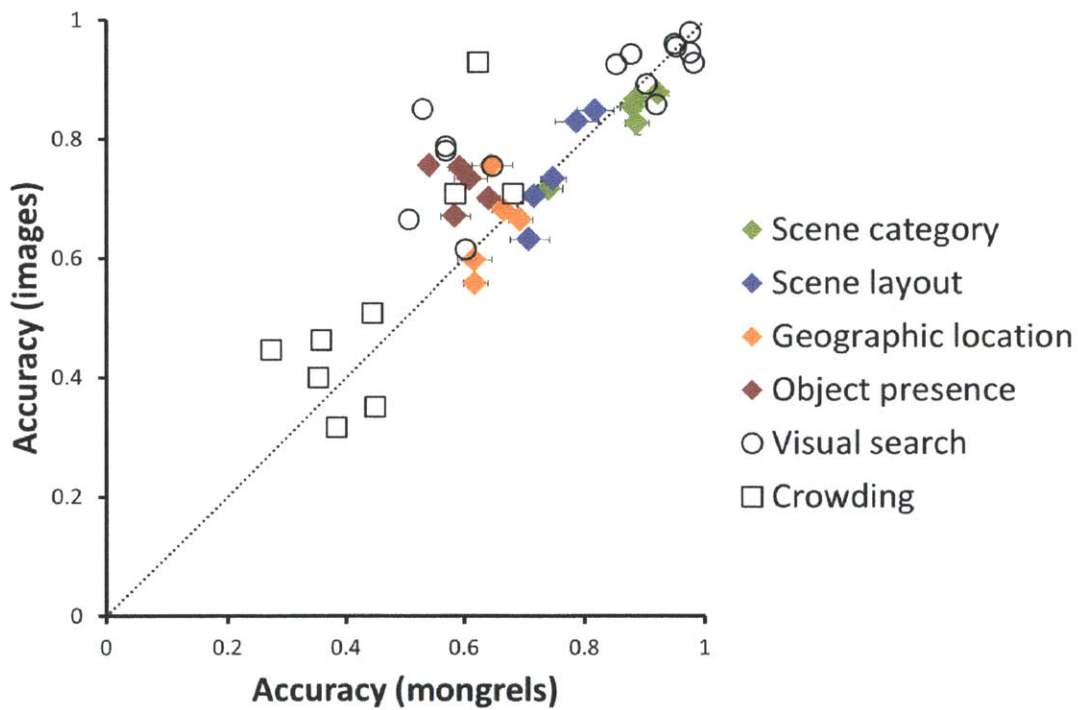


Figure B-6: Mongrel vs. image results from the scene perception tasks, with crowding results from Balas, et al., 2009 and visual search results from Rosenholtz, et al., 2012.

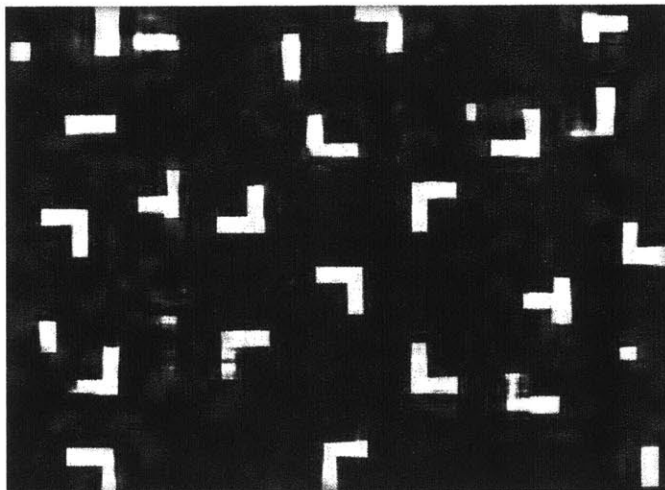
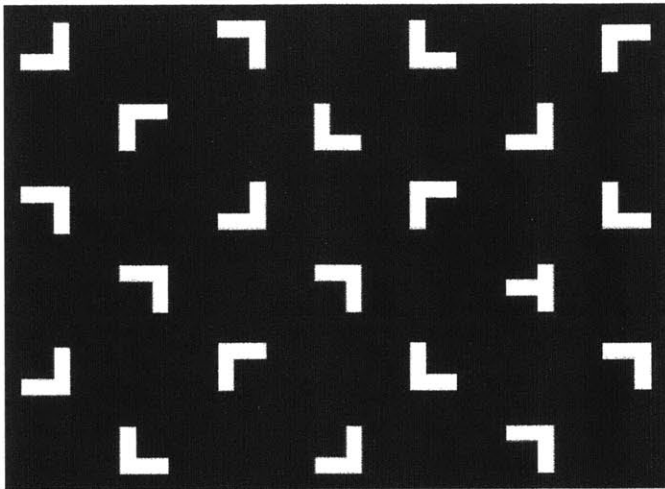


Figure B-7: Example of a visual search display (above) and its corresponding mongrel (below). Fixation is assumed to be in the center of the display.

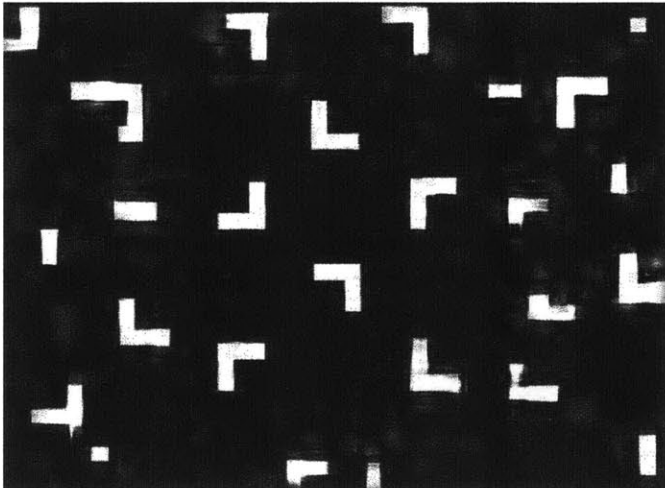
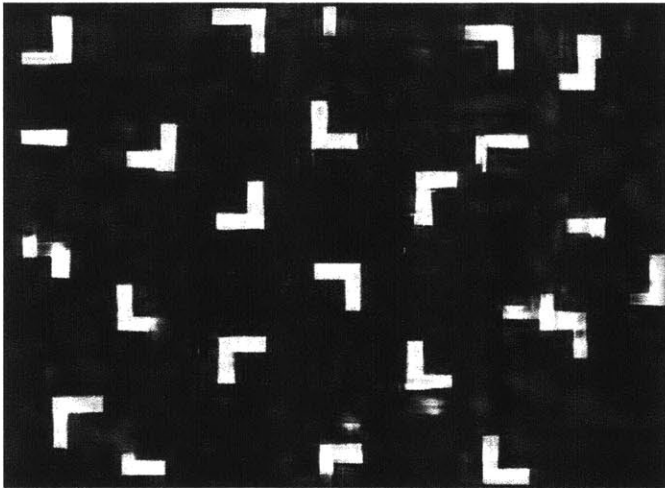


Figure B-8: Two additional mongrels of the same visual search display. Fixation is assumed to be in the center of the display.



Figure B-9: Example of a scene search display (above) and its corresponding mongrel (below). The mongrel is synthesized with fixation in the center of the search display.

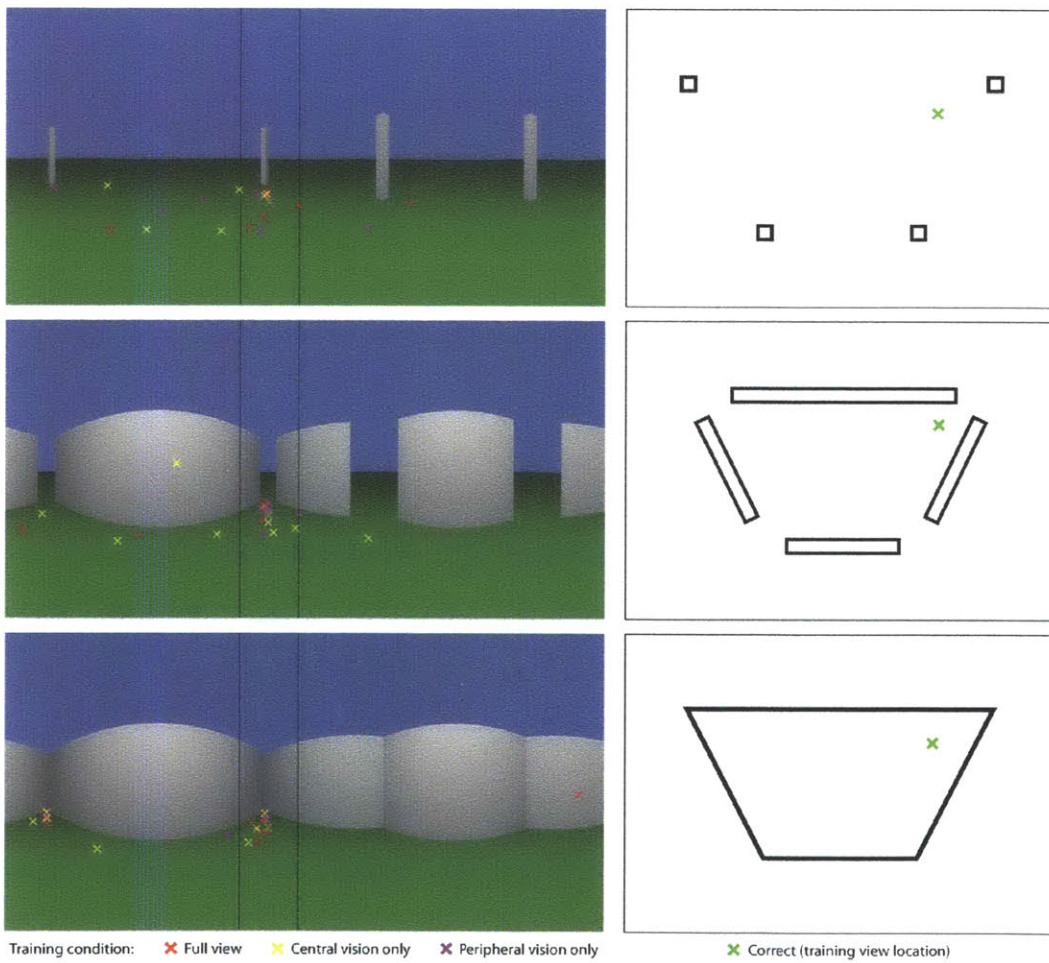


Figure B-10: Examples of the artificial environments used in the reorientation task: a trapezoidal environment with the shape defined by pillars (top), unconnected walls (middle), or connected walls (bottom). Xs represent observers' responses, and the vertical lines represent an 18 degree threshold around the correct response (green X). An overhead diagram of each environment is shown on the right.

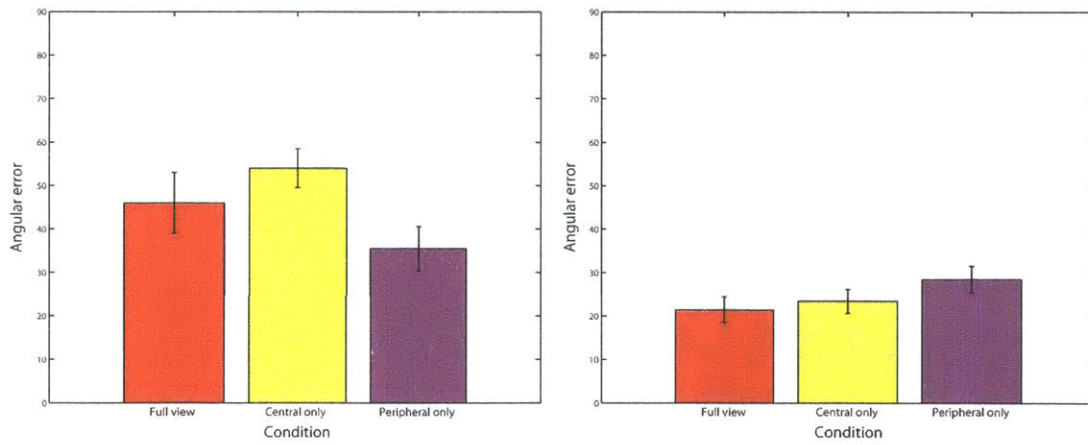


Figure B-11: Angular error in responses in each training condition, for artificial scenes (left) and real-world scenes (right).

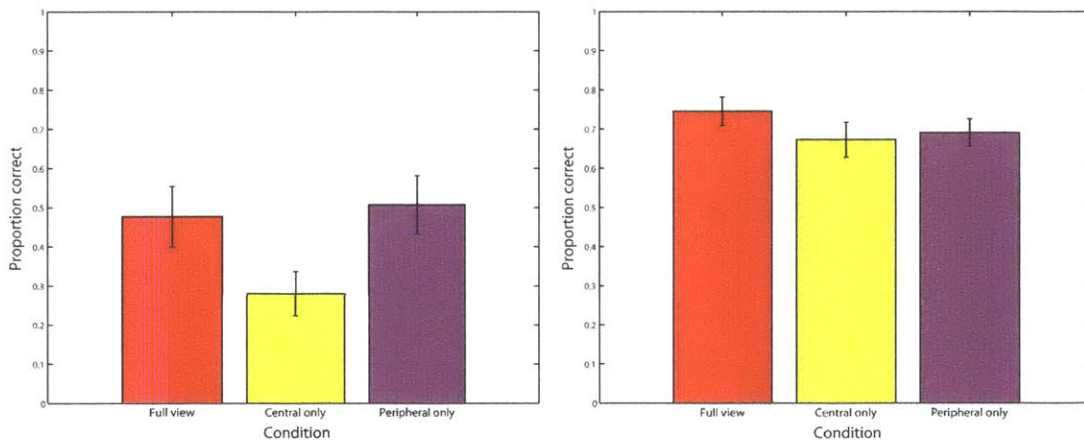


Figure B-12: Percent of correct responses in each training condition, for artificial scenes (left) and real-world scenes (right).

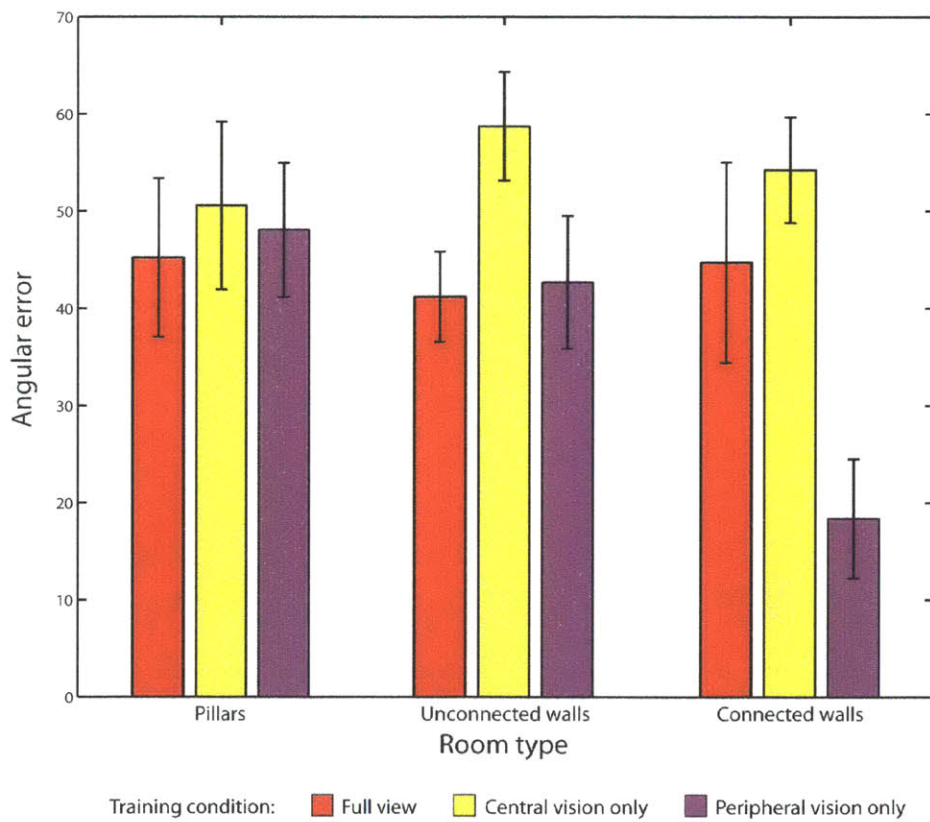


Figure B-13: Angular error in responses in each training condition, for each of the three artificial scene types.

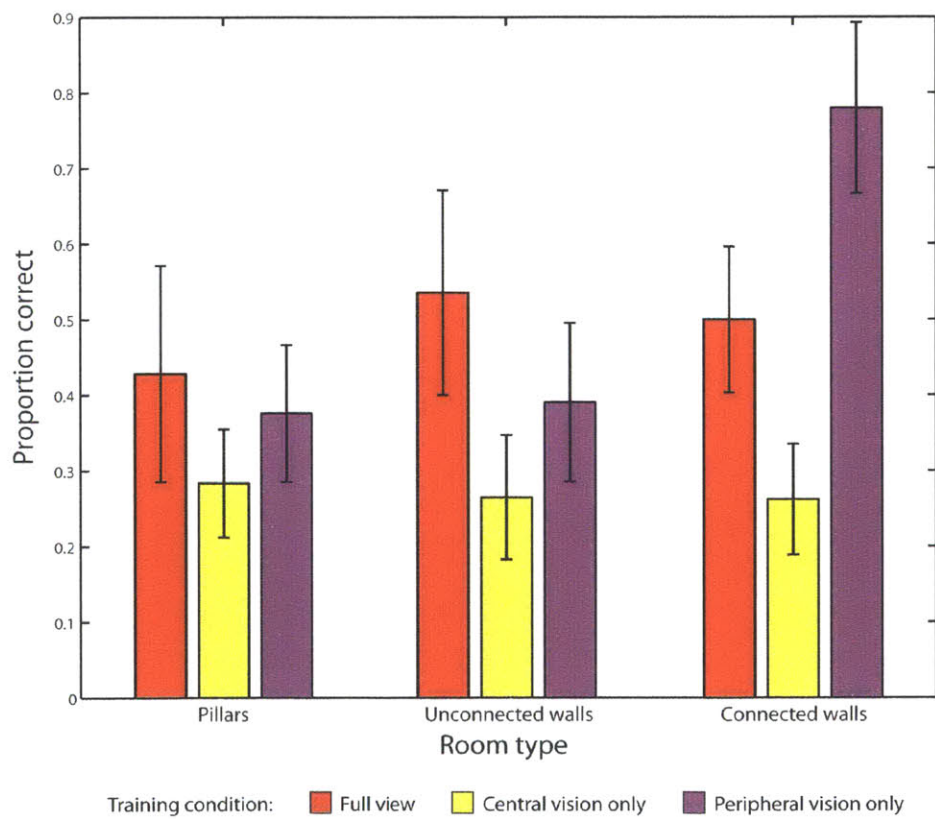


Figure B-14: Percent of correct responses in each training condition, for each of the three artificial scene types.

References

- Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision*, *9*(12), 1-18. doi: 10.1167/9.12.13
- Benedikt, M. L. (1979). To take hold of space: Isovists and isovist fields. *Environment and Planning B*, *6*, 47-65.
- Biederman, I. (1981). On the semantics of a glance at a scene. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual Organization* (p. 213-263). Hillsdale, New Jersey: Lawrence Erlbaum.
- Blanz, V., Tarr, M., & Bulthoff, H. (1999). What object attributes determine canonical views? *Perception*, *28*(5), 575-599.
- Bouma, H. (1970). Interaction effects in parafoveal letter recognition. *Nature*, *226*, 177-178.
- Burgess, N., Recce, M., & O'Keefe, J. (1994). A model of hippocampal function. *Neural Networks*, *7*(6-7), 1065-1081.
- Cheng, K. (1986). A purely geometric module in the rat's spatial representation. *Cognition*, *23*(2), 149-178.
- Crouzet, S. M., & Serre, T. (2011). What are the visual features underlying rapid object recognition? *Frontiers in Psychology*, *2*(326). doi: 10.3389/fpsyg.2011.00326
- Cutzu, F., & Edelman, S. (1994). Canonical views in object representation and recognition. *Vision Research*, *34*(22), 3037-3056. doi: 10.1016/0042-6989(94)90277-1
- Cyr, C., & Kimia, B. (2001). 3D object recognition using shape similarity-based aspect graph. In *2001 Proceedings of the 8th IEEE International Conference on Computer Vision (ICCV)* (Vol. 1, p. 254 -261 vol.1). doi: 10.1109/ICCV.2001.937526
- Diwadkar, V., & McNamara, T. (1997). Viewpoint dependence in scene recognition. *Psychological Science*, *8*(4), 302-307. doi: 10.1111/j.1467-9280.1997.tb00442.x
- Edelman, S., & Bulthoff, H. (1992). Orientation dependence in the recognition of familiar and novel views of 3-dimensional objects. *Vision Research*, *32*(12), 2385-2400.
- Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition*, *17*(6-7), 945-978. doi: 10.1080/13506280902834720
- Ekstrom, A., Kahana, M., Caplan, J., Fields, T., Isham, E., Newman, E., & Fried,

- I. (2003). Cellular networks underlying human spatial navigation. *Nature*, *425*(6954), 184-187.
- Freeman, J., & Simoncelli, E. (2011). Metamers of the ventral stream. *Nature Neuroscience*, *14*(9), 1195 - 1201.
- Gallistel, C. (1990). *The organization of learning*. Cambridge, MA: MIT Press.
- Gibson, J. (1979). *The Ecological Approach to Visual Perception*. Hillsdale, NJ: Erlbaum.
- Gouteux, S., & Spelke, E. (2001). Children's use of geometry and landmarks to reorient in an open space. *Cognition*, *81*(2), 119-148.
- Greene, M. R., & Oliva, A. (2009). The briefest of glances: The time course of natural scene understanding. *Psychological Science*, *20*(4), 464-472. doi: 10.1111/j.1467-9280.2009.02316.x
- Greene, M. R., & Wolfe, J. M. (2011). Global image properties do not guide visual search. *Journal of Vision*, *11*(6), 1-9. doi: 10.1167/11.6.18
- Hartley, R. I., & Zisserman, A. (2004). *Multiple View Geometry in Computer Vision* (Second ed.). Cambridge University Press, ISBN: 0521540518.
- Hays, J., & Efros, A. A. (2008). im2gps: Estimating geographic information from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Hermer, L., & Spelke, E. (1994). A geometric process for spatial reorientation in young children. *Nature*, *370*(6484), 57-59.
- Huttenlocher, J., & Lourenco, S. (2007). Coding location in enclosed spaces: is geometry the principle? *Developmental Science*, *10*(6), 741-746.
- Joubert, O. R., Rousset, G. A., Fabre-Thorpe, M., & Fize, D. (2009). Rapid visual categorization of natural scene contexts with equalized amplitude spectrum and increasing phase noise. *Journal of Vision*, *9*(1), 1-16. doi: 10.1167/9.1.2
- Lee, S., & Spelke, E. (2010). A modular geometric mechanism for reorientation in children. *Cognitive Psychology*, *61*(2), 152-176.
- Lee, S., & Spelke, E. (2011). Young children reorient by computing layout geometry, not by matching images of the environment. *Psychonomic Bulletin & Review*, *18*(1), 192-198.
- Lew, A., Gibbons, B., Murphy, C., & Bremner, J. (2010). Use of geometry for spatial reorientation in children applies only to symmetric spaces. *Developmental Science*, *13*(3), 490-498.
- Li, F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(14), 9596-9601. doi: 10.1073/pnas.092277599
- O'Keefe, J., & Burgess, N. (1996). Geometric determinants of the place fields of hippocampal neurons. *Nature*, *381*(6581), 425-428.
- O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map. preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, *34*(1), 171 - 175.
- O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. New York: Oxford University Press.

- Oliva, A., & Schyns, P. G. (1997). Coarse blobs or fine edges: Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology*, *34*, 72-107.
- Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145-175. doi: 10.1023/A:1011139631724
- Palmer, S., Rosch, E., & Chase, P. (1981). Canonical Perspective and the Perception of Objects. In J. Long & A. Baddeley (Eds.), *Attention and Performance IX* (p. 135-151). Hillsdale, NJ: Erlbaum.
- Portilla, J., & Simoncelli, E. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, *40*(1), 49-71.
- Rosenholtz, R., Huang, J., Raj, A., Balas, B. J., & Ilie, L. (2012). A summary statistic representation in peripheral vision explains visual search. *Journal of Vision*, *12*(4). doi: 10.1167/12.4.14
- Ross, M. G., & Oliva, A. (2010). Estimating perception of scene layout properties from global image features. *Journal of Vision*, *10*(1). doi: 10.1167/10.1.2
- Rousek, J. B., & Hallbeck, M. S. (2011). The use of simulated visual impairment to identify hospital design elements that contribute to wayfinding difficulties. *International Journal of Industrial Ergonomics*, *41*(5), 447-458. doi: 10.1016/j.ergon.2011.05.002
- Rousselet, G., Joubert, O., & Fabre-Thorpe, M. (2005). How long to get to the "gist" of real-world natural scenes? *Visual Cognition*, *12*(6), 852-877. doi: 10.1080/13506280444000553
- Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Science*, *104*(5), 6424-6429.
- Shusterman, A., Lee, S., & Spelke, E. (2011). Cognitive effects of language on human navigation. *Cognition*, *120*(2), 186-201.
- Simon, I., Snavely, N., & Seitz, S. M. (2007). Scene summarization for online image collections. In *2007 IEEE 11TH International Conference on Computer Vision (ICCV)* (p. 274-281).
- Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*(6582), 520-522.
- Toet, A., Jansen, S. E. M., & Delleman, N. J. (2007). Effects of field-of-view restrictions on speed and accuracy of manoeuvring. *Perceptual and Motor Skills*, *105*(3, Part 2), 1245-1256. doi: 10.2466/PMS.105.4.1245-1256
- Toet, A., Jansen, S. E. M., & Delleman, N. J. (2008). Effects of field-of-view restriction on manoeuvring in a 3-D environment. *Ergonomics*, *51*(3), 385-394. doi: 10.1080/00140130701628329
- Tolman, E. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189-208.
- Torralba, A., Oliva, A., Castelano, M. S., & 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(4), 766-786. doi: 10.1037/0033-295X.113.4.766
- Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12, 97-136.
- van Rheede, J. J., Kennard, C., & Hicks, S. L. (2010). Simulating prosthetic vision: Optimizing the information content of a limited visual display. *Journal of Vision*, 10(14). doi: 10.1167/10.14.32
- VanRullen, R., Reddy, L., & Koch, C. (2004). Visual search and dual tasks reveal two distinct attentional resources [Article]. *Journal of Cognitive Neuroscience*, 16(1), 4-14.
- Verfaillie, K., & Boutsen, L. (1995). A corpus of 714 full-color images of depth-rotated objects. *Perception & Psychophysics*, 57(7), 925-961.
- Waller, D. (2006). Egocentric and nonegocentric coding in memory for spatial layout: Evidence from scene recognition. *Memory & Cognition*, 34(3), 491-504.
- Waller, D., Friedman, A., Hodgson, E., & Greenauer, N. (2009). Learning scenes from multiple views: Novel views can be recognized more efficiently than learned views. *Memory & Cognition*, 37, 90-99.
- Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). SUN Database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (p. 3485-3492). doi: 10.1109/CVPR.2010.5539970