Canonical views of scenes depend on the shape of the space

Krista A. Ehinger (kehinger@mit.edu)
Department of Brain & Cognitive Sciences, MIT, 77 Massachusetts Ave.
Cambridge, MA 02139 USA

Aude Oliva (oliva@mit.edu)
Department of Brain & Cognitive Sciences, MIT, 77 Massachusetts Ave.
Cambridge, MA 02139 USA

Abstract

When recognizing or depicting objects, people show a preference for particular "canonical" views. Are there similar preferences for particular views of scenes? We investigated this question using panoramic images, which show a 360-degree view of a location. Observers used an interactive viewer to explore the scene and select the best view. We found that agreement between observers on the "best" view of each scene was generally high. We attempted to predict the selected views using a model based on the shape of the space around the camera location and on the navigational constraints of the scene. The model performance suggests that observers select views which capture as much of the surrounding space as possible, but do not consider navigational constraints when selecting views. These results seem analogous to findings with objects, which suggest that canonical views maximize the visible surfaces of an object, but are not necessarily functional views.

Keywords: canonical view; scene perception; panoramic scenes.

Introduction

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, & Bülthoff, 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, Tarr, & Bülthoff, 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 well studied. Studies using artificial scenes (a collection of objects on a surface) have shown that scene learning is viewpoint 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, et al., 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, suggesting that there is good agreement on the "best" views of these scenes (Simon, Snavely, and 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, Rosch, and Chase (1981). For example, the most commonly photographed view in a particular cathedral could 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, 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 – 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 such as the one shown in Figure 1. These images are taken with a lens attached to a bell-shaped mirror, which captures all of the views available from a particular location.

Method

Materials

The stimuli were 624 panoramic images taken in various indoor and outdoor locations (classroom, lobby, chapel, parking lot, garden, athletic field, etc.). Each image was 3200 by 960 pixels, corresponding to 360° horizontal visual angle and about 110° degrees vertical visual angle.

Participants

195 people participated in the experiment through Amazon's Mechanical Turk, an online service where workers are paid to complete short computational tasks (HITs) for small amounts of money. All of the workers 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.

Design

Each image was seen by 10 different workers. On average, a single worker performed 32 trials (median 9 trials).

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° by 45° visual angle). Observers could change the view shown in the window by clicking and dragging the image with the mouse; this gave the effect of turning 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 (e.g. "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.

Figure 1: An example of a panoramic image used in the experiment. The smaller window shows a portion of the scene as it appeared in the interactive viewer during the experiment (the view shown here is the average "best view" chosen by participants).

Model

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. In addition, people 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.

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 (see Figure 2a). 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 height of this edge in each image, were able to estimate the shape of the visible space around the camera, as shown in Figure 2b. (This field of visible space around a camera location is called the "isovist" in architectural research (Benedikt, 1979).) 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 the "volume map" shown in Figure 2c.

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 an unwrapped panoramic image (as in Figure 1) and were asked 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

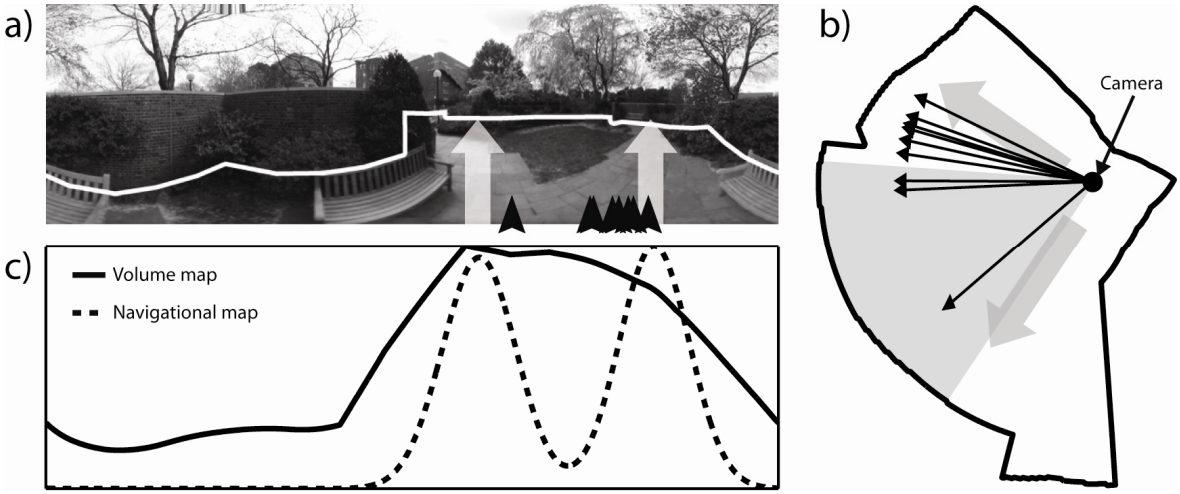

Figure 2: (a) A panoramic image with the ground line outlined in white and arrows marking the navigational paths (black arrowheads represent "best" views of this scene chosen by participants). (b) An overhead view of the same location (the grey region represents the portion of the space captured within a single camera view). (c) The volume and navigational maps for this scene.

marked paths – 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 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 the "navigational map" shown in Figure 2c. This map gives an estimate of the navigability of all possible views around the camera location.

Results

Experiment 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). 251 out of 6240 trials were excluded under these criteria.

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. Significance was measured using Rayleigh's test of nonuniformity, which tests the significance of a mean angle in a circular distribution by comparing it to the mean angle that would be expected from a distribution of random angles. This test returned $p < .01$ for 389 scenes (62% of the image set), and $p < .05$ for 466 scenes (75% of the image set)). This may be a conservative estimate of agreement, since Rayleigh's test does not distinguish between random distributions of views and some types of multimodal distributions (such as views clustered around two angles 180 degrees apart). Examples of scenes with high, moderate, and low agreement are shown in Figure 3.

Agreement (measured as the standard error in the views selected by participants) was correlated with some aspects of the scene layout. Specifically, standard error in views was correlated with the overall volume of space around the camera location, as calculated from the volume map ($r = 0.30$). Similarly, standard error in views was correlated with the percent of subjects who marked the scene as a "large, open space" during the path-marking task ($r = 0.22$). These correlations indicate that agreement on the "best" view was higher in small spaces, and lower in spaces that were large and open. Agreement was also related to the range of distances visible from the camera location. The standard error in views was negatively correlated with the standard deviation of visible depths ($r = -0.40$). In other words, agreement on the "best" view was higher in scenes that showed a variety of closer and farther views than in scenes where all views were about equally distant.

Agreement was significantly higher in indoor than in outdoor scenes ($t(246.8) = 5.81$, $p < .001$). This is likely due to differences in the spatial envelope of these spaces (Oliva & Torralba, 2001): outdoor scenes tend to be much larger and more open than indoor scenes, and indoor scenes are more likely to have complex shapes offering a range of closer and farther views.

There was also a relationship between view agreement and name agreement from the naming portion of the task. The standard error of the angles chosen by observers was negatively correlated with the percent of people giving the dominant name for the scene ($r = -0.44$). This means that when observers agreed on the identity of the scene, they also

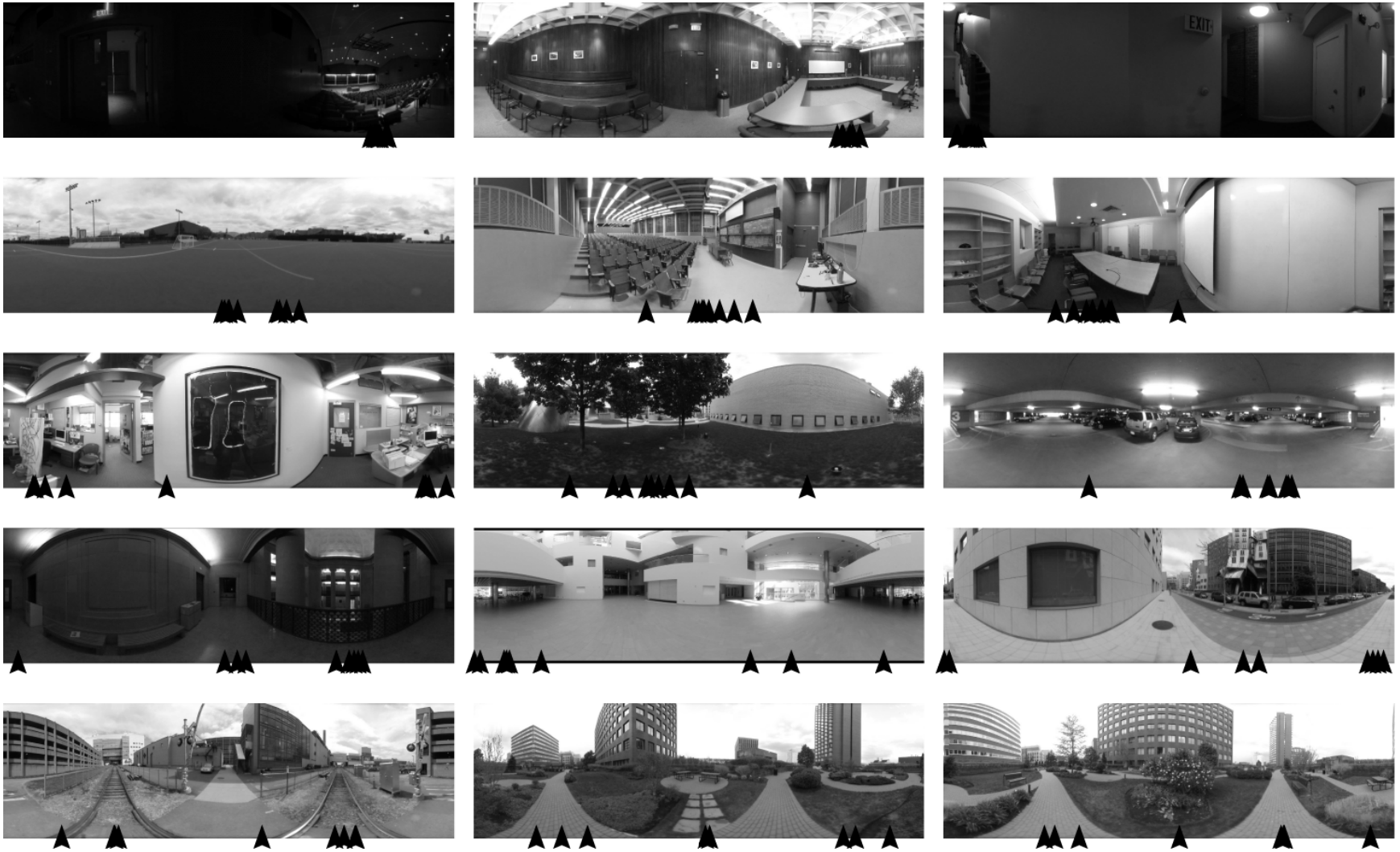

Figure 3: Example scenes; black arrowheads represent views chosen by participants. The top row are the three scenes with highest agreement, followed by scenes at the 75th, 50th, and 25th percentile of agreement. The bottom row shows the three scenes with lowest agreement.

tended to agree on the "best" view of the scene, but when observers disagreed on a scene's identity, they were also likely to choose different "best" views of the scene.

Model performance

One image was dropped from the modeling because it was a very small space with no visible floor, so its volume map was undefined Volume and navigation maps were calculated for the remaining 623 images as described in the previous section. We then tested how well each of these maps could predict the "best" views selected by observers.

Model performance was assessed using ROC curves (Figure 4). 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 threshold at a range of 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 volume model gives the best prediction of the views selected by observers (AUC = 0.75), but the navigational model also performs above chance (AUC = 0.62). The performance of the navigational model does not change when very open scenes (which may not have clear paths) are excluded from the analysis. On 426 "closed" scenes (scenes that were never marked as "open space" during the park-marking task), the navigational model's AUC was 0.61; on the remaining "open" scenes the AUC was 0.62. On the other hand, the volume model does show better performance in closed than in open scenes (AUC = 0.78 and 0.74, respectively). Figure 5 shows examples of high and low performance from the volume and navigational models.

We also tested a combined model, which attempted to predict selected views using both a weighted sum of the volume and navigational maps. However, this model performed worse than the volume map alone, and gave better performances as the weight of the navigational map approached zero. This suggests that the navigational model does not add any indpendent predictive power; it performs above chance because it tends to select the same regions as the volume map (in scenes, a view that shows a large volume usually also affords navigation).

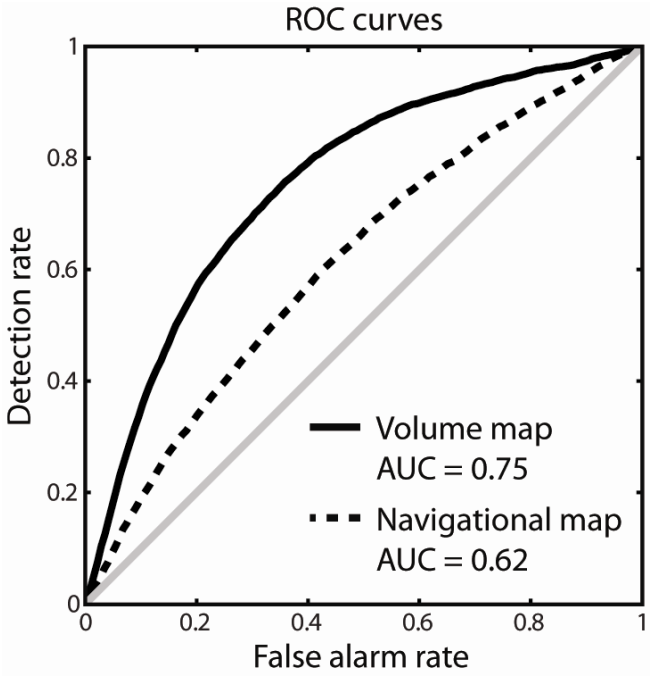

Figure 4: ROC curves for the volume and navigational models. The gray line represents chance performance (chance AUC = 0.5).

Conclusion

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, for example, 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. Choosing the "best" view of an object or a scene poses essentially the same problem: how to compress as much 3D visual information as possible

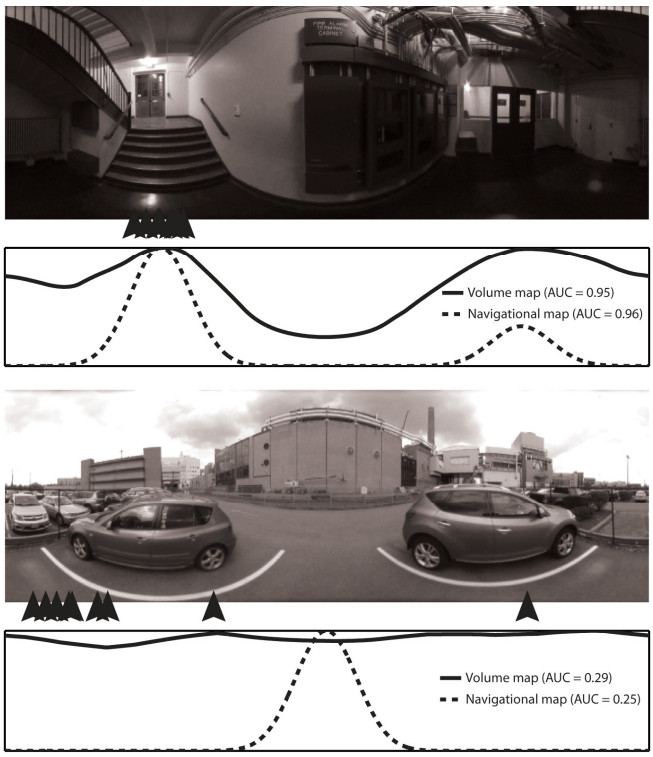

Figure 5: Example of a scene in which both models performed very well (top) and an example of a scene in which both models performed poorly (bottom). Arrowheads mark the "best" views chosen by observers.

into a necessarily limited 2D view.

When selecting canonical views of objects, people seem to be trying 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.). Further work will be required to distinguish between these two possibilities.

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 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 are selected because they show a large amount of the surrounding space, not because they afford navigation.

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, Tarr, and Bülthoff (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 & Bülthoff, 1992; Cutzu & Edelman, 1994). The existence of canonical views of scenes could suggest a similar view-based representation for memory and perception of scenes.

Acknowledgments

The authors would like to thank Ken M. Haggerty for his work preparing stimuli for this project. K.A.E. is funded by an NSF graduate research fellowship. This research was partly funded by an NSF Career award (0546262), NSF grants (0705677 and 1016862) and a NEI grant (EY02484) to A.O.

References

Blanz, V., Tarr, M. J., & Bülthoff, H. H. (1999). What object attributes determine canonical views? *Perception*, *28*, 575-599.

Benedikt, M. L. (1979). To take hold of space: Isovists and isovist fields. *Environment and Planning B, 6*, 47–65.

Cutzu F., & Edelman S. (1994). Canonical views in object representation and recognition. *Vision Research, 34*, 3037-3056.

Diwadkar, V. A., & McNamara, T. P. (1997). Viewpoint dependence in scene recognition. *Psychological Science*, *8*, 302–307.

Edelman S., & Bülthoff, H. H. (1992). Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision Research, 32*, 2385-2400.

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

Palmer, S., Rosch, E., Chase, P., (1981). Canonical perspective and the perception of 40 objects. In *Attention and Performance IX*, Ed. J. Long, A. Baddeley (Hillsdale, NJ: Lawrence Erlbaum), pp. 135-151.

Simon, I., Snavely, N., Seitz, S. M. (2007). Scene Summarization for Online Image Collections. In *Proc. Of the 11th International Conference on Computer Vision*.

Verfaillie, K. & Boutsen, L. (1995). A corpus of 714 full-color images of depth-rotated objects. *Perception & Psychophysics*, *57*, 925-961.

Waller, D. (2006). Egocentric and nonegocentric coding in memory for spatial layout: Evidence from scene recognition. *Memory & Cognition*, *34*, 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.