PsyCh Journal

INSTITUTE OF PSYCHOLOGY
Chinese Academy of Sciences

PsyCh Journal **1** (2012): 82–89
DOI: 10.1002/pchj.12

View combination in recognition of 3-D virtual reality layouts

Hui Zhang,[1,2] Alinda Friedman,[3] Weimin Mou,[3] and David Waller[4]

[1]Institute of Psychology, Chinese Academy of Sciences, Beijing, China, [2]University of California, Davis, CA, USA, [3]University of Alberta, Edmonton, AB, Canada, [4]Miami University, Oxford, OH, USA

Abstract: We investigated whether a normalization model or view combination model fit the performance of scene recognition of 3-D layouts using a virtual-reality paradigm. Participants learned a layout of seven objects from two training views (e.g., 0° and 48°) by discriminating the "correct" layout from distracters. Later, they performed a discrimination task using the training views (e.g., 0° and 48°), an interpolated view (e.g., 24°), an extrapolated view (e.g., 72°), and a far view (e.g., 96°). The results showed that the interpolated view was easier to discriminate than the extrapolated view and even easier than the training views. These results extend the applicability of view combination accounts of recognition to 3-D stimuli with stereoscopic depth information.

Keywords: change detection; scene recognition; view combination

Correspondence: Dr. Hui Zhang, Room 103, 1544 Newton Court, Center for Neuroscience, University of California Davis, Davis, California 95618, USA. Email: hzhang@ucdavis.edu

Received 25 October 2011. Accepted 8 June 2012.

People often have to recognize objects and scenes from views that they have not previously experienced. An ongoing theoretical issue in vision research is to understand what spatial representation(s) people form when they learn particular views of an object or scene and how these are used in the recognition of novel objects or scenes.

One theory of object and scene recognition—a *normalization* approach—stipulates that mental representations of space are viewpoint dependent, such that people represent only experienced views. According to this approach, in order to recognize a novel view after training, the novel percept is transformed with respect to the representation of the closest (single) training view (Christou, Tjan, & Bülthoff, 2003; Diwadkar & McNamara, 1997; Nakatani, Pollatsek, & Johnson, 2002; Tarr, 1995; Tarr & Pinker, 1989). Thus, the normalization approach predicts that scene recognition will be viewpoint dependent, insofar as familiar views should be easier to recognize than novel views, and novel views should be recognized as a monotonically slower or less accurate function of their distance from the learned views. The normalization approach thus predicts that novel views with the same transformational distance to a familiar view should be equally easy to recognize.

The initial evidence for viewpoint-dependent scene recognition came from a study by Diwadkar and McNamara (1997, Experiment 2). Using a discrimination task, Diwadkar and McNamara had people learn a desktop-sized layout from one perspective to a criterion before learning the same layout from three other perspective views simultaneously. Latency appeared to be a linear function of the angular distance between the novel view and the nearest training view. The authors therefore concluded that the spatial relations of the layout were represented in a viewpoint-dependent manner. They also concluded that recognition of the novel views generally involves a process of normalization to the nearest training view.

However, it is still not clear how multiple training views affect scene recognition performance in the normalization model. For example, recently, Friedman and Waller, and their colleagues (Friedman, Spetch, & Ferrey, 2005; Friedman, Vuong, & Spetch, 2010; Friedman & Waller, 2008; Friedman, Waller, Thrash, Greenauer, & Hodgson, 2011; Spetch & Friedman, 2003; Waller, Friedman, Hodgson, & Greenauer, 2009) proposed the view combination model for scene recognition based on the framework developed by Edelman and others (Bülthoff & Edelman, 1992; Edelman,

1999; Edelman & Bülthoff, 1992; Edelman, Bülthoff, & Bülthoff, 1999). They demonstrated that novel views of scenes that were between two training views were recognized more easily than novel views outside of two training views, despite being at the same distance from a learned view. Furthermore, in some cases, the novel views were recognized even more easily than the learned views. Neither result is predicted by a normalization model.

This type of view combination effect was first described as a model of object recognition (Bülthoff & Edelman, 1992; Edelman, 1999; Edelman et al., 1999; Edelman & Bülthoff, 1992). In this approach to object recognition, when a novel view (or even a novel object) is presented, all the learned representations in a parametric shape space that are above a certain threshold of similarity to the input are activated as a function of their similarity, which may be measured in several ways, for example, Gabor similarity (Gabor, 1946), which indicates that the nearer (or more similar) a training view is to the novel view, the greater the contribution of that training view is to overall recognition of the scene. In this approach, the activation of all learned views that are similar to the novel input is summed and used to construct a new view, based on mathematically interpolating between all the parameters that have been activated in the similarity space. If this constructed view is above a threshold of similarity to the novel input, the input is "recognized." For the present purposes, the critical distinction between this model and the normalization model is that the view combination framework allows for multiple familiar views to be activated during the process of recognition, and thus for the recognition of some novel views to be relatively easy.

Friedman and Waller (2008, Experiment 1) had people learn pictures of a playground from two ground-level training views (e.g., 0° and 48°) by discriminating the correct layout from distracters. The targets and distracters were differentiated from one another using either the movement of one object or a switch in position between two objects. The latter manipulation in particular was hypothesized to "force" subjects to learn the locations and spatial relations among the objects; that is, with a "switch" distracter, it was impossible to discriminate the targets from the distracters without learning at least the relative target locations and the targets' identities.

During the test phase, participants discriminated the learned layout from distracters at all five viewpoints (the two training views in addition to three novel views). The novel *interpolated* view was between the span of the two training views (i.e., at 24°), the *extrapolated* view was outside that range by an equal amount of angular distance (i.e., at 72°, which is 24° from the 48° training view), and the far view was the most distant from the training views (i.e., at 96°). Both accuracy and response latency for the interpolated view was better than that for the extrapolated view, and equal to the performance on the training views. This pattern is the behavioral signature of the view combination effect (sometimes also called viewpoint interpolation). More strikingly, in a subsequent study, Waller et al. (2009, see also Friedman et al., 2011) had subjects discriminate a virtual playground from "switch" distracters using four elevated perspective views that surrounded a central view. The central view and the four extrapolated views were never presented during the learning phase. The participants recognized the interpolated central view faster and more accurately than the training views. They referred to this pattern of data as an *enhanced prototype* effect.

Although the stimuli used by Friedman and Waller (2008) and Waller et al. (2009) were depictions of 3-D scenes, they were still presented as 2-D images on a computer screen. Other research (Friedman et al., 2011; McNamara, Diwadkar, Blevins, & Valiquette, 2006) has used even more simplified 2-D images, such as arrays of colored dots. However, there is currently no evidence in the literature that has demonstrated view combination results for actual 3-D layouts. Generalizing view combination effects to the types of layouts and situations commonly experienced by people is an important next step in understanding the scope of a general theory of scene recognition.

Although several cues to 3-D layout have been present in the 2-D stimuli used in previous research, there are additional depth cues (e.g., stereoscopic depth information, motion parallax) that have generally not been available. These cues, which are available from motion and stereo vision, may be critical in 3-D object and scene recognition and may moderate the view combination effects obtained with 2-D computer displays, even when they depicted 3-D scenes. For example, there is some evidence that object recognition differs when the object is presented as flat, 2-D images and when the object is presented with stereoscopic depth information (e.g., Friedman et al., 2005; Pasqualotto & Hayward, 2009). In contrast, Friedman et al. (2011) claim that view combination is a "general recognition mechanism" and, as such, should be evident in a wide range of visual learning situations.

To test the generality of this claim, in the current study we used an immersive virtual-reality paradigm to provide participants with additional cues (e.g., stereoscopic depth) to the 3-D layout that are not present in static 2-D images of scenes. We then addressed which model, the normalization model or the view combination model, fit the data better.

The participants learned a 3-D object array on a virtual desktop from two training views, which were 48° apart from each other, using the discrimination procedure that was used by Friedman and Waller (2008). The distracters were composed of the same objects as the target scene, but two of the objects had switched places with each other (Friedman & Waller, 2008). During each of the five test blocks, the scene was presented from the two training views as well as three novel views (interpolated, extrapolated, and far; Figure 1). We assumed that if the data fit the view combination model, the performance on the interpolated view should be better than that from the extrapolated view and approximately equal to the learned views. A normalization model would assume no performance differences between the interpolated and extrapolated views.

Previous work has shown that, even though learning effects occurred during the testing session (Waller et al., 2009), the interpolated view was responded to more efficiently than the extrapolated view, even on the very first test trial. We speculated that the participants in our study would also display good performance on the interpolated view early in testing, even though they would also learn from all the testing views. Thus, we further hypothesized that the data from the earlier testing trials (e.g., the first test block) would better distinguish between these two models.

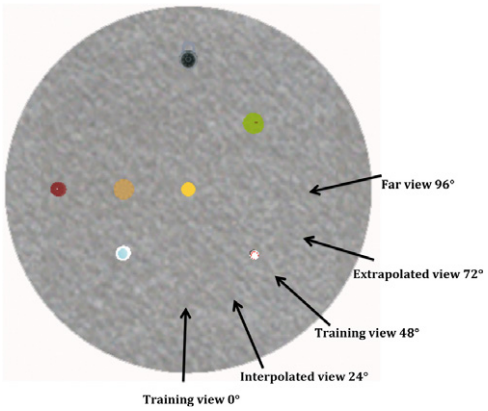

Figure 1. The target layout used in Experiments 1 and 2, viewed from above.

Method

Participants

Sixty-four university students (32 male and 32 female) participated in the study in return for monetary compensation. All were right-handed according to self-report and had normal or corrected-to-normal vision. They gave written informed consent to participate in the study. Of these, 53 (82.6%) participants (26 male and 27 female) met the accuracy criterion of scoring at least 90% correct on each of the two training views during the test trials. The data from the 11 individuals (six male and five female) who did not pass the accuracy criterion were not considered further.

Material and design

The virtual environment with layouts of the objects on a table was displayed in stereo with lightweight (approximately 200 g) glasses-like I-glasses and a PC/SVGA Pro 3-D head-mounted display (HMD; I-O Display Systems, Sacramento, CA). The participants' head motion was tracked with an InterSense IS-900 motion-tracking system (InterSense, Billerica, MA). The HMD supplied 3-D images at a resolution of 800 pixels × 600 pixels and a field of view (FOV) of 26° diagonally for each eye. The virtual objects and the virtual table were rendered with a GeForce 6600GT graphics accelerator. The virtual objects and the virtual table were presented on the origin of the coordinates (superimposed at the center of the table), which was defined by the tracking system and could be recognized by the tracking marker mounted on the HMD. The participants were required to look at the center of the virtual table, so the virtual objects and the virtual table could be seen at the center of the FOV through the HMD.

The apparatus was placed in a 6 m × 6 m laboratory with each wall covered by homogeneous black curtains. As illustrated in Figure 1, the layout consisted of seven common virtual objects (lock, apple, candle, hat, ball, bottle, and battery) with the longest dimension approximately 5 cm. The distance between two nearest objects was 18 cm (e.g., hat and ball).

The objects were placed on a circular virtual table (80 cm in diameter) with a gray matt texture. The table was presented on the floor in the middle of the room. A real chair (seated 42 cm high) was placed 1.9 m away from the center of the virtual table. The participants sat on the chair during

both the learning and test phases. A real bar stool with a mouse on it was placed on the preferred-hand side of the chair for each participant.

Five views of the layout were represented in the HMD. One view was arbitrarily labeled 0°, and the other four views were labeled, with a step size of 24° counterclockwise, as 24°, 48°, 72°, and 96°. The target versions of each view are shown on the left-hand panel of Figure 2.

Half of the participants were trained with the views of 0° and 48° and were tested with those views, as well as with three novel views that were interpolated (24°), extrapolated (72°), and far (96°), relative to the trained stimuli (see Figure 1). The other half of the participants were trained with the 48° and 96° views and tested with those views, as well as with three novel views at 72° (interpolated), 24° (extrapolated), and 0° (far). All of the objects were fully visible from all the viewpoints without overlap of any two objects.

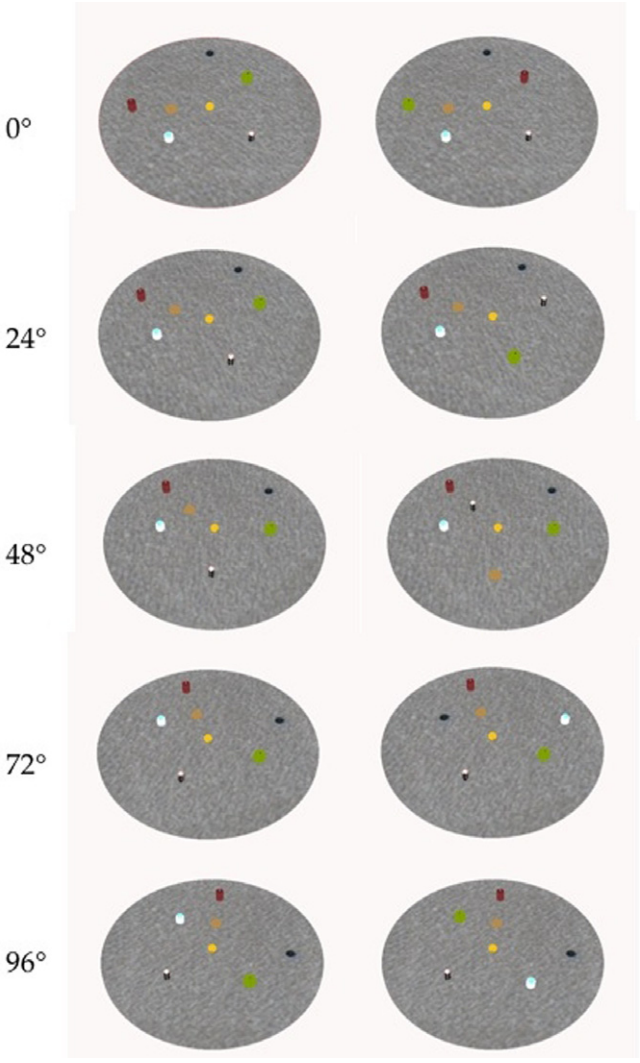

Figure 2. The materials used in Experiments 1 and 2.

Distracters were constructed by randomly switching the positions of two objects in the layout (e.g., apple and candle, see Figure 2). One of the distracters at each view is shown in Figure 2.

A training block consisted of two target trials and two distracters for each training view, for a total of eight trials in the two learning views. The trials in each block were presented in a random sequence. The participants completed at least five blocks of training trials. We calculated their performance online starting with the fifth block. When the participant achieved a 100% accuracy rate on the fifth or later block of training trials, they proceeded to the test trials; otherwise, they continued to do training blocks until the criterion was reached. All the participants reached the criterion during the training session.

A test block consisted of two target trials and two distracters for each of the five views, for a total of 20 trials. The participants received five test blocks. The order of the stimuli was randomized within each block.

Procedure

After the instructions, the participants were blindfolded and guided into the virtual reality room. They were then seated in the chair and put on the HMD.

Before the experimenter initiated the program, a red arrow was presented at the position where the virtual table would be presented. The participants were encouraged to move their head freely to become accustomed to the HMD and then to keep their eyes fixed on the red arrow. The participants held the mouse on the bar stool with their right hand. After the participants indicated that they were ready for the experiment, the first training trial was initiated with a key press by the experimenter. On each training trial, there was a warning sound for 1 s, followed immediately by the stimulus. One of the training arrays was presented on the virtual table and the participants were required to judge whether it was the target array or a distracter. Half of the participants were asked to press the left mouse button when a target array was presented and the right mouse button when a distracter array was presented, while the other half responded the opposite way.

The participants received auditory feedback during the training trials. If the participants made a correct response in less than 2 s on a training trial, the feedback informed them they had received two points. If they were correct but the

response time was longer than 2 s, the feedback informed them that they had received one point. If they were wrong, the feedback said "wrong." All auditory feedback was prerecorded.

The participants were informed that initially they must guess about whether a given arrangement was correct or not and, once they had made a decision, they should respond by pressing the mouse keys as quickly and accurately as possible. They were also informed that there was no feedback in the test phase, but they would still get one or two points for each correct response and no point for wrong responses. This point system was used solely to encourage the participants to engage in the task and had no tangible reward.

Results

In all the tests reported, we adopted a two-tailed alpha level of .05 and an effect size measure of η_p^2.

Response times

ANOVAs with variables of viewing angle (4) and testing block (5) found a significant interaction between viewing angles and testing blocks, $F(12, 624) = 2.58$, $p < .01$, $\eta_p^2 = .05$, as well as a significant main effect of viewing angle, $F(3, 156) = 13.58$, $p < .001$, $\eta_p^2 = .21$, and testing block, $F(4, 208) = 10.81$, $p < .001$, $\eta_p^2 = .17$. The mean response times of all correct responses for different views across all the testing blocks are shown in Figure 3a. Consistent with previous research (e.g., Friedman & Waller, 2008; Waller et al., 2009), because we intended to perform specific comparisons among views from the outset, we used planned comparisons (Rosenthal & Rosnow, 2009) on the theoretically important difference between the interpolated and extrapolated views. The mean response times across all the testing blocks showed that the response time for the interpolated view was significantly shorter than for the extrapolated view, $t(52) = 2.34$, $p = .023$, and was even shorter than that for the training view, $t(52) = 2.46$, $p = .01$.

Additional analyses examined how and whether the view combination effect evolved over the testing blocks. We fit a logarithmic model to the latency data for the extrapolated and interpolated views, respectively. Figure 4 illustrates the simulated data for the response times for the interpolated and extrapolated views across the five testing blocks.

Finally, we examined the differences between the views for only the first test block (Figure 5a, 5b). We again found a significant main effect of response time across training,

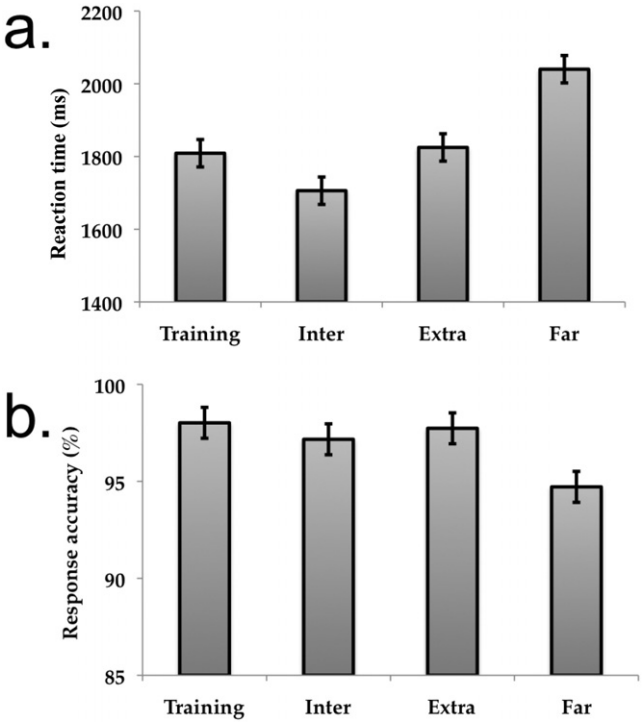

Figure 3. (a) Response latency in discriminating the target layout from the distracters across all the test blocks. (b) Response error in discriminating the target layout from the distracters across all the test blocks. Error bars are the confidence interval corresponding to ±1 standard error of the mean, as estimated from the analysis of variance.

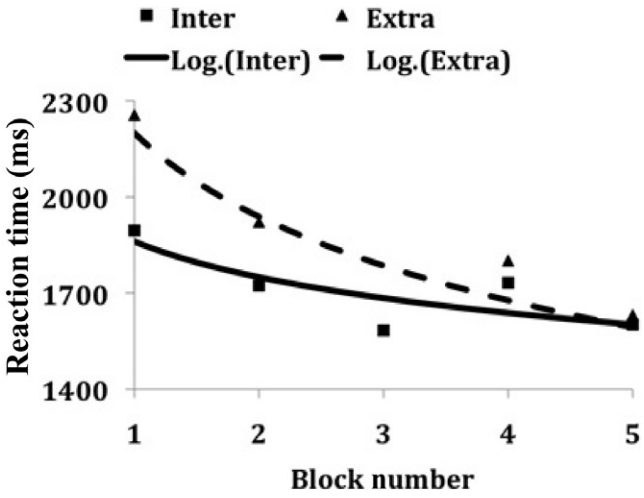

Figure 4. The simulated trend of the performance of the interpolated and extrapolated views across the testing blocks.

interpolated, and extrapolated views, $F(2, 104) = 4.61$, $p = .012$, $\eta_p^2 = .081$. Planned comparisons showed that, even in the first testing block, the response time for the interpolated view was significantly shorter than that for the extrapolated view, $t(52) = 2.3$, $p = .025$. However, no

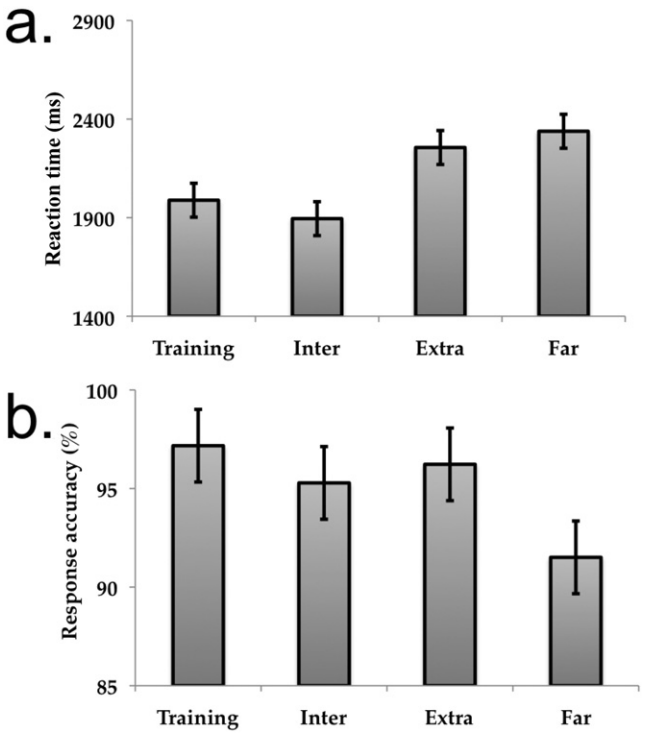

Figure 5. (a) Response latency in discriminating the target layout from the distracters in the first test block. (b) Response error in discriminating the target layout from the distracters in the first test block.

significant difference was found for the response time between the interpolated and training views in the first testing block, $t(52) = 1.1$, $p = .28$.

Response accuracy

The same ANOVA as above was performed for the response accuracy and a main effect of viewing angle was found, $F(3, 156) = 3.56$, $p = .016$, $MSE = .0034$, $\eta_p^2 = .07$. The mean response accuracy across all the testing blocks is shown in Figure 3b. Planned comparisons of the mean response accuracy between the interpolated and extrapolated views across all the testing blocks showed that there was no difference in the accuracy between them, $t(52) = .51$, $p = .62$, nor was there a difference between the interpolated and training views $t(52) = 1.00$, $p = .32$. This result is not surprising, given the training criterion.

For the 11 subjects who failed to reach the 90% accuracy criterion, the mean response latencies for training, interpolated, extrapolated, and far were 1729 ms, 1583 ms, 1770 ms, and 2039 ms, respectively. Their response accuracies for training, interpolated, extrapolated, and far were 80%, 91%, 94%, and 79%, respectively. Here, both the reaction time and the accuracy data mimic those from the subjects who did reach the accuracy criterion.

Discussion and conclusion

In our experiment, by using a virtual reality paradigm, we provided participants with a vivid 3-D layout in which they learned the locations of objects from two training views and then performed a recognition task that required them to discriminate the target layouts from the distracters for the training views as well as for novel views of the layouts. The discrimination performance supported the existence of view combination effects for the first time in a situation in which the participants could receive "natural" cues as to depth. These results extend the findings about scene recognition (Friedman & Waller, 2008; Waller et al., 2009) as well as object recognition (Bülthoff & Edelman, 1992; Edelman, 1999; Edelman et al., 1999; Edelman & Bülthoff, 1992; Friedman et al., 2005; Spetch & Friedman, 2003; Spetch, Friedman, & Reid, 2001) to virtual displays with stereoscopic depth information. Previous studies have shown that people are able to obtain more spatial information from immersive virtual reality than from static visual images of the environment (Waller, Beall, & Loomis, 2004). Accordingly, in comparison with previous studies, our use of an immersive virtual reality paradigm should have enabled the participants to obtain information closer to an actual environmental situation. There are still some constraints on our paradigm, however, because we cannot know the extent to which immersive virtual reality actually simulates a real situation.

Both the view combination and normalization models are able to explain a linear decrease in performance with an increase in offset from the training view, such as the significant linear decline in the response latencies among the training, extrapolated, and far views, $F(1, 52) = 16.26$, $p < .001$, as seen in the current data. Normalization models cannot readily explain, however, the relative ease we observed in the recognition of interpolated views after training with multiple views. However, a view combination model can explain such a result by positing that all prior training views contribute to scene recognition for a novel view, in proportion to their similarity to the training views. For example, in the current study, after a participant trained with the 0° and 48° views, the 48° view could contribute the same to the 24° interpolated view as it could to the 72° extrapolated view. The 0° view, in contrast, would likely contribute more to the 24° view than to the 72° view. Thus, performance for the interpolated view is better than for the extrapolated view. Such findings run counter to Tarr's (1995) proposal that object

recognition of a novel view is based on normalization to the nearest training view, which predicts the same performance for the interpolated and extrapolated views. Thus, the view combination model provides a better explanation of how people perform scene recognition tasks after learning from multiple views.

In the current study, the participants were first trained from two views and then tested from these two training views and three novel test views. Thus, in the current study as well as that of Waller et al. (2009), the participants possibly encoded new information about the layout from the test session. In other words, the initially defined extrapolated view (e.g., 72°), could have become the "interpolated view" between one of the training views (48°) and the far view (96°). We believe that, because of this extra "training" during the test phase, the difference in the performance for the interpolated view and the extrapolated view diminished for the later testing blocks (Figure 4).

In the current experiment, the participants responded to the interpolated view significantly faster than the extrapolated view (Friedman et al., 2011; Friedman & Waller, 2008; Waller et al., 2009). In contrast, there were no differences in the response accuracy among the training, interpolated, and extrapolated conditions. As noted, this may not be surprising, because all the participants had to achieve 100% correct on each of the two training views during the training trials in order to proceed to the test phase. Furthermore, we only analyzed the data from the participants whose response accuracy was higher than 90% on the training views during the test trials. Thus, a ceiling effect for response accuracy probably limited our ability to detect the view combination effects for the accuracy measure.

In sum, in the current experiment we used a virtual reality paradigm with a 3-D table-sized virtual layout of common objects and verified that the view combination model fit our data better than the normalization model. It is now the case that the view combination model has received support for stimulus arrays as simple as dots (Friedman et al., 2011) and as complex as the present 3-D virtual arrays. View combination thus seems a much more general description of visual recognition than does normalization.

References

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences of the United States of America*, *89*, 60–64.

Christou, C. G., Tjan, B. S., & Bülthoff, H. H. (2003). Extrinsic cues aid shape recognition from novel viewpoints. *Journal of Vision*, *3*, 183–198. doi:10.1167/3.3.1

Diwadkar, V. A., & McNamara, T. P. (1997). Viewpoint dependence in scene recognition. *Psychological Science*, *8*, 302–307. doi:10.1111/j.1467-9280.1997.tb00442.x

Edelman, S. (1999). *Representation and recognition in vision*. Cambridge, MA: MIT Press.

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. doi:10.1016/0042-6989(92)90102-O

Edelman, S., Bülthoff, H. H., & Bülthoff, I. (1999). Effects of parametric manipulation of inter-stimulus similarity on 3D object categorization. *Spatial Vision*, *12*, 107–123. doi:10.1163/156856899X00067

Friedman, A., Spetch, M. L., & Ferrey, A. (2005). Recognition by humans and pigeons of novel views of 3-D objects and their photographs. *Journal of Experimental Psychology: General*, *134*, 149–162. doi:10.1037/0096-3445.134.2.149

Friedman, A., Vuong, Q. C., & Spetch, M. L. (2010). Facilitation by view combination and coherent motion in dynamic object recognition. *Vision Research*, *50*, 202–210. doi:10.1016/j.visres.2009.11.010

Friedman, A., & Waller, D. (2008). View combination in scene recognition. *Memory and Cognition*, *36*, 467–478. doi:10.3758/MC.36.3.467

Friedman, A., Waller, D., Thrash, T., Greenauer, N., & Hodgson, E. (2011). View combination: A generalization mechanism for visual recognition. *Cognition*, *119*, 229–241. doi:10.1016/j.cognition.2011.01.012

Gabor, D. (1946). Theory of communication. *Journal of the Institute of Electrical Engineers*, *93*, 429–457.

McNamara, T. P., Diwadkar, V. A., Blevins, W. A., & Valiquette, C. M. (2006). Representations of apparent rotation. *Visual Cognition*, *13*, 273–307. doi:10.1080/13506280544000002

Nakatani, C., Pollatsek, A., & Johnson, S. H. (2002). Viewpoint-dependent recognition of scenes. *The Quarterly Journal of Experimental Psychology*, *55*, 115–139. doi:10.1080/02724980143000190

Pasqualotto, A., & Hayward, W. G. (2009). A stereo disadvantage for recognizing rotated familiar objects. *Psychonomic Bulletin and Review*, *16*, 832–838. doi:10.3758/PBR.16.5.832

Rosenthal, R., & Rosnow, R. L. (2009). *Contrast analysis: Focused comparisons in the analysis of variance*. New York: Cambridge University Press.

Spetch, M. L., & Friedman, A. (2003). Recognizing rotated views of objects: Interpolation versus generalization by humans and pigeons. *Psychonomic Bulletin and Review*, *10*, 135–140. doi:10.3758/BF03196477

Spetch, M. L., Friedman, A., & Reid, S. L. (2001). The effect of distinctive parts on recognition of depth-rotated objects by pigeons (*Columba livia*) and humans. *Journal of Experimental Psychology: General*, *130*, 238–255. doi:10.1037//0096-3445.130.2.238

Tarr, M. J. (1995). Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of

three-dimensional objects. *Psychonomic Bulletin and Review*, *2*, 55–82.

Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, *21*, 233–282. doi:10.1016/0010-0285(89)90009-1

Waller, D., Beall, A. C., & Loomis, J. M. (2004). Using virtual environments to assess directional knowledge. *Journal of Environmental Psychology*, *24*, 105–116. doi:10.1016/S0272-4944(03)00051-3

Waller, D., Friedman, A., Hodgson, E., & Greenauer, H. (2009). Learning scenes from multiple views: Novel views can be recognized more efficiently than learned views. *Memory and Cognition*, *37*, 90–99. doi:10.3758/MC.37.1.90