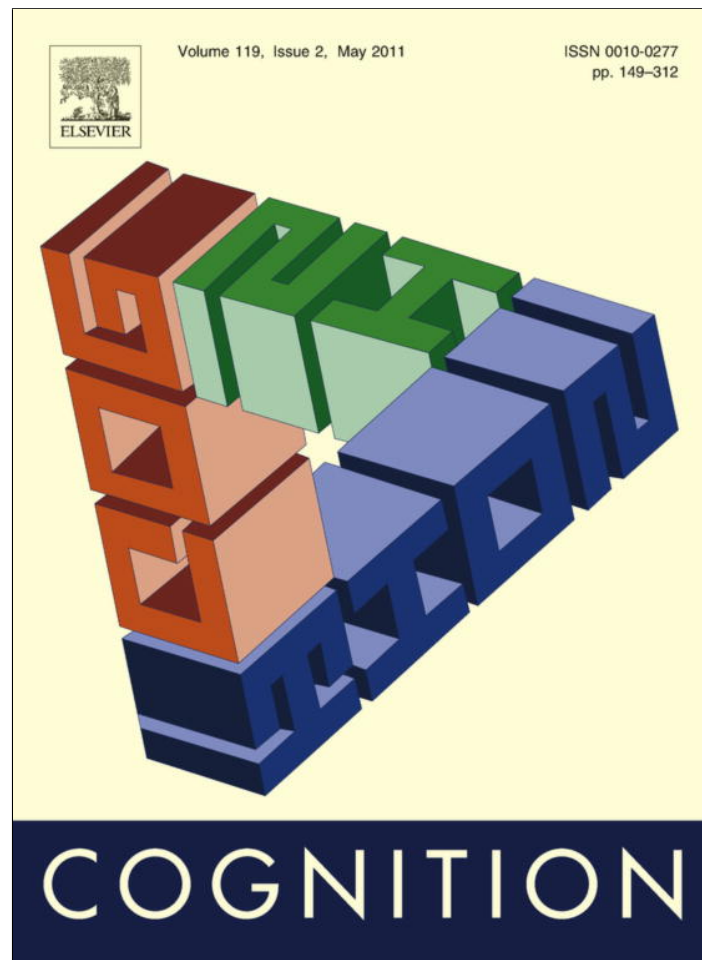


Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

View combination: A generalization mechanism for visual recognition

Alinda Friedman^{a,*}, David Waller^b, Tyler Thrash^b, Nathan Greenauer^b, Eric Hodgson^b^a University of Alberta, Edmonton, Alberta, Canada^b Miami University, Oxford, OH, United States

ARTICLE INFO

Article history:

Received 2 January 2010

Revised 18 January 2011

Accepted 21 January 2011

Available online 21 February 2011

Keywords:

Visual recognition

Scene recognition

View combination

Categorization

Generalization

ABSTRACT

We examined whether view combination mechanisms shown to underlie object and scene recognition can integrate visual information across views that have little or no three-dimensional information at either the object or scene level. In three experiments, people learned four “views” of a two dimensional visual array derived from a three-dimensional scene. In Experiments 1 and 2, the stimuli were arrays of colored rectangles that preserved the relative sizes, distances, and angles among objects in the original scene, as well as the original occlusion relations. Participants recognized a novel central view *more efficiently* than any of the Trained views, which in turn were recognized more efficiently than equidistant novel views. Experiment 2 eliminated presentation frequency as an explanation for this effect. Experiment 3 used colored dots that preserved only identity and relative location information, which resulted in a weaker effect, though still one that was inconsistent with both part-based and normalization accounts of recognition. We argue that, for recognition processes to function so effectively with such minimalist stimuli, view combination must be a very general and fundamental mechanism, potentially enabling both visual recognition and categorization.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Because most objects or situations are unlikely to recur in the exact form and context in which they are first experienced, it is important for human and nonhuman animals alike to have mental processes that can generalize from past to present experiences. Indeed, Shepard, 1987 proposed stimulus generalization as a potential universal law for psychology as it can support both recognition and categorization (Edelman, 1999a). In its simplest conception, stimulus generalization is the probability of responding to a stimulus event that is in some measure similar to a previously experienced one; generalization gradients are metric functions – for example, exponential or Gaussian – of the physical or psychological distance between the

two events based on their physical or psychological similarities.

In the case of vision, most mobile animals require a means to recognize objects, landmarks, and vistas they have previously seen. In past research, we and others have shown that *view combination*, which is a form of generalization that uses information derived from multiple views, plays an important role in object recognition (Bülthoff & Edelman, 1992; Edelman, 1999a; Edelman & Bülthoff, 1992; Edelman, Bülthoff, & Bülthoff, 1999; Friedman, Spetch, & Ferrey, 2005; Poggio & Edelman, 1990; Spetch & Friedman, 2003; Wong & Hayward, 2005). In addition, we have recently shown that view combination processes can also explain recognition performance for real-world scenes (Friedman & Waller, 2008; Waller, Friedman, Hodgson, & Greenauer, 2009; see also Castelano, Pollatsek, & Rayner, 2009). With the present work, we examine whether view combination mechanisms also function for more impoverished stimuli than objects or scenes by investigating whether they operate with visual

* Corresponding author. Address: Department of Psychology, University of Alberta, Edmonton, Alberta, Canada T6G 2E9. Tel.: +1 780 492 2909; fax: +1 780 492 1768.

E-mail address: alinda@ualberta.ca (A. Friedman).

arrays that have no cues to three-dimensional (3D) object structure (Experiments 1 and 2) and minimal or no cues to 3D scene structure (Experiment 3; see Figs. 1 and 2). As we elaborate below, showing that view combination mechanisms operate on such minimalist stimuli supports their fundamental importance to a variety of psychological phenomena in visual recognition and categorization. Additionally, by systematically stripping the visual arrays we used of various features (compared to actual scenes), we were able to arrive at a functional definition of what may constitute a “scene” – an issue that has clear importance for the understanding of human spatial cognition.

The most extensive theoretical development of recognition by combining information from multiple views has been Shimon Edelman’s work with object recognition (1999a, 2002a, 2002b; see also Bühlhoff & Edelman, 1992; Edelman et al., 1999; Ullman, 1998). Edelman (1999a) hypothesized that view combination processes allow people and animals to identify novel views of familiar objects and to categorize altogether novel objects. He proposed that parametric similarities among “prototype” objects are important for recognition and that the physical dimensions of the objects (e.g., attributes such as length, curvature, etc.) form the basis of an internally represented

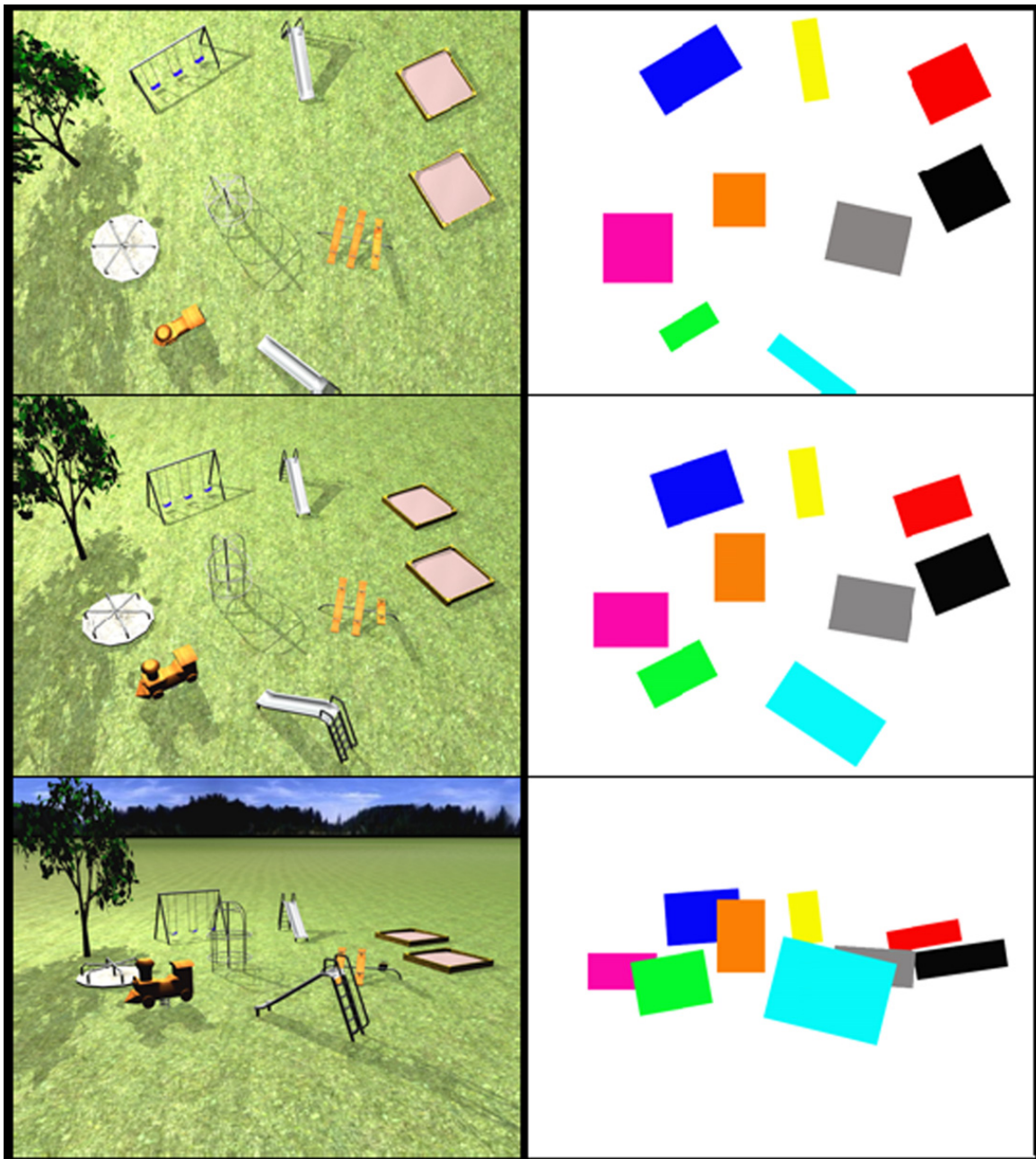


Fig. 1. Example of three target stimuli from Waller et al. (2009) and the stimuli created from them for the present study. The top left panel was the novel Extrapolated view of the playground scene taken at an elevation of 75° , the middle panel was the novel Interpolated view at an elevation of 45° , and the bottom panel was the novel Extrapolated view at a 15° elevation. Each of these views was $\pm 15^\circ$ from the nearest training view and at an arbitrarily assigned azimuth of 0° . The right panels adjacent to each of the playground views correspond to the same conditions in the present study.

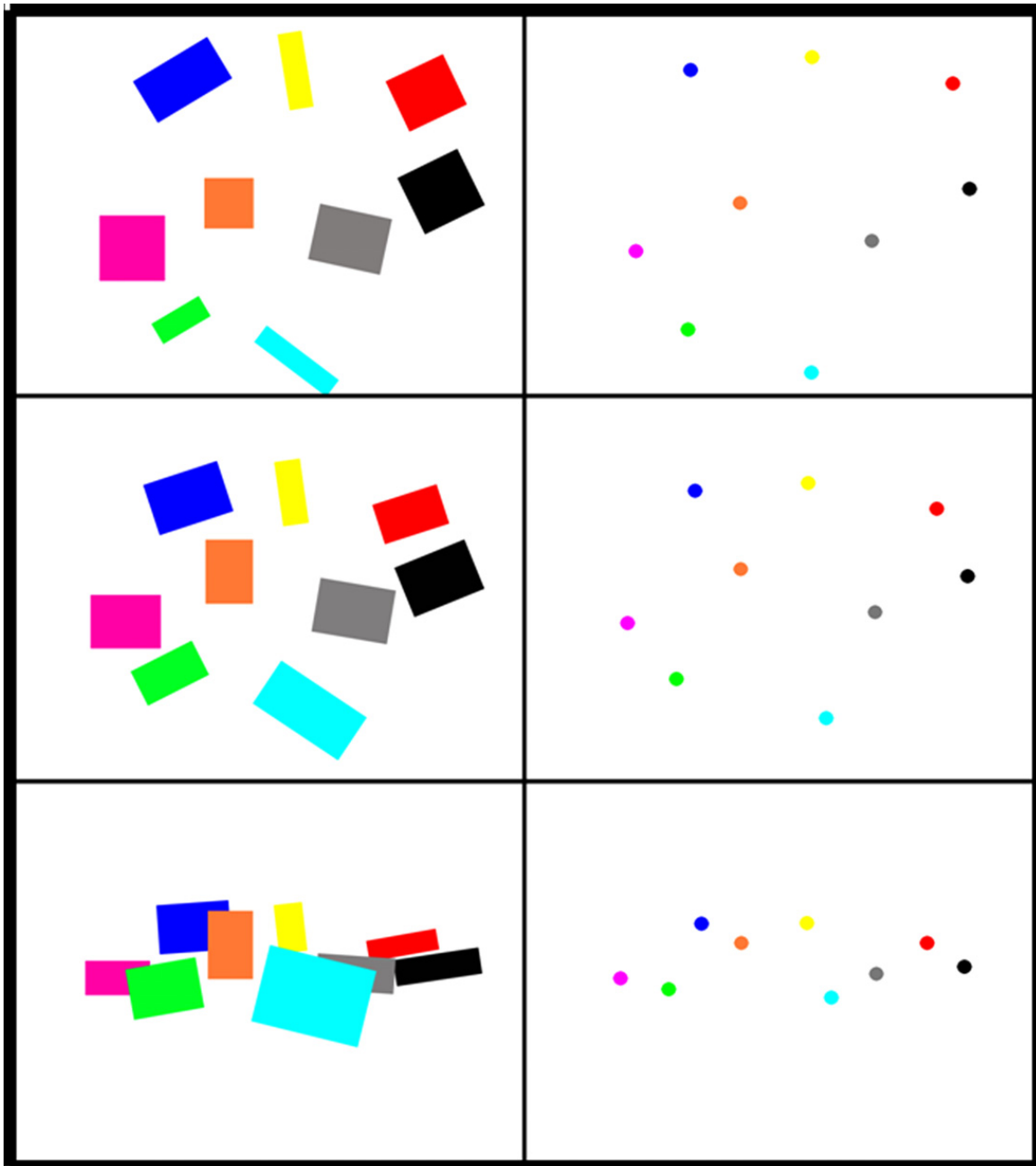


Fig. 2. Example stimulus materials used in Experiment 3 (right panel); each dot is the central point of its corresponding rectangle in the left panel.

multidimensional “shape space.” Similarity is assumed to be a function of the distances between tuning functions (i.e., generalization gradients) representing the prototypes in this space, which in turn are a function of the metric similarities among their multidimensional physical attributes. In this scheme, a novel view of a familiar stimulus, or even a novel stimulus, can be recognized or categorized because the stimulus input causes all of the generalization functions representing stored prototypes that are parametrically similar to the dimensions of the novel input to be simultaneously activated. The activation from these functions is summed to construct a new “view” containing various physical dimensions *in proportion* to the activation they have received. Because of the proportionality of acti-

vation on each dimension, the new view is typically not identical to the input, but it is examined for its metric similarity to the input view. If that similarity is above a certain threshold, recognition (or categorization) occurs.

Because view combination processes sum activation from multiple sources, when a novel input is presented, if the stored generalization functions are sufficiently close in the space, recognition of novel views that span the range of at least two previously experienced views can be as good as (Bülthoff & Edelman, 1992; Edelman, 1999a; Friedman & Waller, 2008; Friedman et al., 2005) or better than (Waller et al., 2009) that for the previously experienced views. Thus, improved (or superior) recognition of novel *Interpolated* stimuli (i.e., those between the span of the Trained

stimuli) relative to equidistant novel *Extrapolated* stimuli (i.e., those beyond span of the Trained stimuli) is the signature evidence in support of view combination accounts of recognition (cf. Friedman & Waller, 2008; Friedman et al., 2005; Waller et al., 2009; Wong & Hayward, 2005). We refer this finding as a *view combination effect*. Because view combination has now been documented with scenes and herein with even simpler arrays, we think that it is appropriate to describe the representation used for recognition as a metric “similarity space”, rather than the “shape space” initially referred to by Edelman, because the representational “manifolds” that Edelman also referred to may be different in kind (for different stimulus types), though the processes that act on them (e.g., generalization; view combination) may not be. Moreover, describing the representations as a similarity space also serves to bring the theoretical construct even further in line with the theoretical implications of Shepard's (1987) proposal of generalization (based on similarity) as a psychological law.

As a model of object recognition, view combination has been contrasted primarily with a part-based representational approach in which recognition is predicted to be viewpoint independent under most circumstances (e.g., Biederman, 1987), as well as a normalization approach in which recognition is usually viewpoint dependent (e.g., Tarr & Pinker, 1989). In a part-based approach (e.g., Biederman, 1987), recognition is based on a relatively small number of volumetric primitives. A two-dimensional (2D) retinal image of an object is segmented into its component primitives through detection of deep regions of concavity as well as non-accidental properties, such as curvature and colinearity. The volumetric primitives and the relations among them are encoded and represented as a 3D *structural description*, which in turn is matched to 3D structural descriptions of objects in memory. Recognition is hypothesized to be viewpoint independent as long as all of an object's volumetric primitives and their relations are identifiable.

It is not clear how a part-based approach could be adapted to make predictions for the range of stimuli that people are able to recognize. For instance, the kinds of simple visual stimuli we examine in the present study (see Figs. 1 and 2) have minimal or no 3D structure, could not involve volumetric primitives, and do not differ with respect to their accidental properties. It is also not clear that this type of approach could, or was meant to, generalize to scene or 2D array recognition. Consequently, we do not discuss part-based recognition further.

In contrast to a part-based approach to recognition, ever since Shepard and Metzler's (1971) demonstration of the mental rotation effect, view-based, *normalization* models of recognition have presumed that representations are 2D, often image-like (e.g., Ullman, 1989), and orientation-specific. In most normalization models, people are assumed to represent multiple orientation-specific views of objects. However, unlike the view combination approach, the multiple representations in a normalization approach capture an object's appearance from different perspectives, rather than explicitly representing the physical dimensions underlying those views as distances in a multidimensional similarity space.

In addition, unlike view combination, normalization models describe recognition as a process of matching a novel view to its nearest specific stored exemplar (even if there are other equidistant stored exemplars), typically by means of a hypothesized transformation mechanism. The “distances” involved in normalizing two views are thus based on the magnitude(s) of transformation(s) required, rather than physical similarity (though the two might be correlated in some cases). That is, for normalization models, similarity is typically based on the physical distance that the novel view must “move through” (e.g., rotate), rather than the physical similarities among the multiple dimensions (e.g., length; amount of curvature) that make up the structure of the stimulus view and previously stored views. Thus, in the normalization approach, the view to be matched is a perspective view of an object or scene and the transformations are geometric. For example, in the *alignment approach* (Ullman, 1989; see also Ullman & Basri, 1991), allowable transformations of the input view include *translation, rotation, scaling, and shear*; after such transformations are made, the views are matched (or not). This approach thus predicts that performance will be monotonically related to the *transformational* distance between the input view and one of the represented views. In the view combination approach, by contrast, the view to be matched to the input may be a perspective view, but it is constructed from activation caused by metric similarities to the input view from many different physical dimensions. If the sum of the activation reaches a certain threshold, the input view is recognized. This approach predicts monotonicity in some circumstances and non-monotonicity in others.

The monotonicity predicted by the normalization account has received much support over a variety of studies and tasks (e.g., handedness discriminations, Shepard & Metzler, 1971; matching or naming familiar objects, Lawson & Bühlhoff, 2008; Palmer, Rosch, & Chase, 1981; naming unfamiliar objects; Tarr, 1995; Tarr & Gauthier, 1998; Tarr & Pinker, 1989). Despite the results that tend to favor normalization, view combination effects are also well-represented in the literature (see Edelman, 1999a, for review). Recent studies have found that recognition of a novel Interpolated view of a scene can be both faster and more accurate than recognition of the Trained view of the scene, even on the very first test trial (Waller et al., 2009). Following the terms used by Palmeri and Nosofsky (2001), we refer to superior recognition of an untrained stimulus to that of a Trained stimulus as a *prototype enhancement effect*. Like the view combination effect, a prototype enhancement effect can be explained by the view combination approach, but is not predicted by normalization models of recognition. Given the extant empirical support for both normalization and view combination accounts of recognition, developing a body of empirical evidence that exposes the limits of human recognition ability will be critical for determining whether (and how) either approach can serve as a general theory of recognition.

With that in mind, in the current experiments, we examined recognition of stimuli that have minimal cues to three-dimensional structure and evaluated whether this performance was better explained by normalization or by

view combination accounts of visual recognition. Our stimuli were arrays of 2D colored rectangles (Experiments 1 and 2) or dots (Experiment 3), each of which was based on a 3D scene stimulus used in the Waller et al. (2009) study. We used abstract arrays because normalization accounts (e.g., Diwadkar & McNamara, 1997) claim to be applicable to them (that is, there should be monotonicity in the data), and view combination accounts, developed for object recognition, may break down with non-semantically related arrays. On the other hand, if 3D structure is absent from stimuli that people can nevertheless recognize as well as or better than familiar stimuli, then view combination must become a more general theoretical construct in order to explain such recognition. For example, Edelman (2002a, 2002b) proposed that a metric similarity space could be used to represent the similarities among pixels on a 2D display, however, this proposal has not heretofore been rigorously examined with respect to 2D stimuli.

As in Waller et al. (2009), in the present study, participants learned to discriminate target from distractor arrays by exposure to four different “viewpoints” and were subsequently tested on views that were either novel interpolations, novel extrapolations, or previously Trained views. Notably, the distractors switched the locations of two rectangles or dots, making it very difficult to do the task without encoding at least identity (color) and object-to-object relations.

According to a normalization approach, recognition performance for both Interpolated and Extrapolated views should be worse than recognition of Trained views. Moreover, it is not entirely clear what transformation(s) could provide a match between the Trained views and any of the novel test views. This is because, although the stimuli were created systematically by changing camera positions by a constant azimuth or elevation in a virtual world, these systematic perspective transformations become difficult to recover when the 3D scenes are rendered as 2D arrays. Current accounts of recognition by normalization would need to be revised considerably in order to explain: (a) a view combination effect with these stimuli, (b) a prototype enhancement effect with these stimuli, and (c) the nature of the transformation(s) that enable recognition of these stimuli. The outcome of the current experiments may thus place serious constraints on a normalization (transformation) account of recognition. On the other hand, view combination models could provide ready and parsimonious explanations of such findings. Demonstrating that view combination accounts for the recognition of simple, minimalistic stimuli will increase its scope and help to establish it as both a fundamental cognitive process and a more general account of recognition than other theoretical models (i.e., normalization and part-based theories).

2. Experiment 1

In Experiment 1, we compared recognition for recently learned stimuli to recognition of novel target stimuli taken from different viewpoints than the training stimuli. These new viewpoints were either between (*Interpolated*) or beyond (*Extrapolated*) the shortest distance between the

Trained views by equal amounts. If view combination mechanisms operate on simple visual arrays, then recognition of Interpolated views may be better than recognition of Extrapolated views (i.e., we will find a view combination effect). Additionally, because the stimuli in Experiment 1 were based on those of Waller et al. (2009), as described below, there may also be better recognition of the untrained Interpolated view relative to the Trained views (i.e., a prototype enhancement effect) because the prototype causes more activation than any single Trained view. Thus, if we find view combination or prototype enhancement effects with these 2D stimuli, it will increase the scope of the view combination approach to that of a very general recognition mechanism.

2.1. Method

2.1.1. Participants

Twenty-six undergraduates (15 men and 11 women) from Miami University participated in the experiment in return for credit in their introductory Psychology course. Eight people (6 men and 2 women) did not reach the 80% criterion in the allotted time, leaving 9 men and 9 women who completed the test trials. The mean age of these participants was 18.9 years ($SD = 1.11$). The percentage of participants who did not reach criterion was approximately the same in Friedman and Waller (2008) and in several similar studies (e.g., Edelman, 1999b; Edelman et al., 1999).

2.1.2. Materials

The stimuli were created from 2D renderings of a digitally modeled 3D playground scene created with 3D Studio Max (see Waller et al., 2009, for a detailed description). The stimuli depicted the original scene from nine different viewing locations; all were above ground level and oriented toward the exact center of the scene. The viewpoint locations are shown schematically in Fig. 3. The *Interpolated* view, which was never seen during training and is depicted by the black circle in the center of the figure, had an angular elevation above the ground plane of 45° and an arbitrarily assigned azimuth of 0° . Four *Training* views, depicted by the gray circles in Fig. 3, were positioned around the Interpolated view, two at the same azimuth and elevations of $\pm 15^\circ$ relative to the Interpolated view (i.e., at 60° and 30° elevations), and two at the same elevation as the Interpolated view and azimuths of $\pm 15^\circ$. Finally, four *Extrapolated* views, also never seen during training and depicted by the white circles in Fig. 3, were similarly positioned around the Interpolated view, differing from it in azimuth or elevation by $\pm 30^\circ$ (e.g., at 75° and 15° elevations) and $\pm 15^\circ$ from the nearest Training view.

For the present experiment, each of the nine central playground objects was masked with a uniquely colored rectangle. The same color always replaced the same object, but the size of the rectangle was changed to fit the visual extent of the object for a given particular perspective view (i.e., a “bounding box”). The remainder of the scene was replaced with a white background. Fig. 1 shows three novel test views of the original playground scene from Waller et al. (2009) and the corresponding stimuli from the pres-

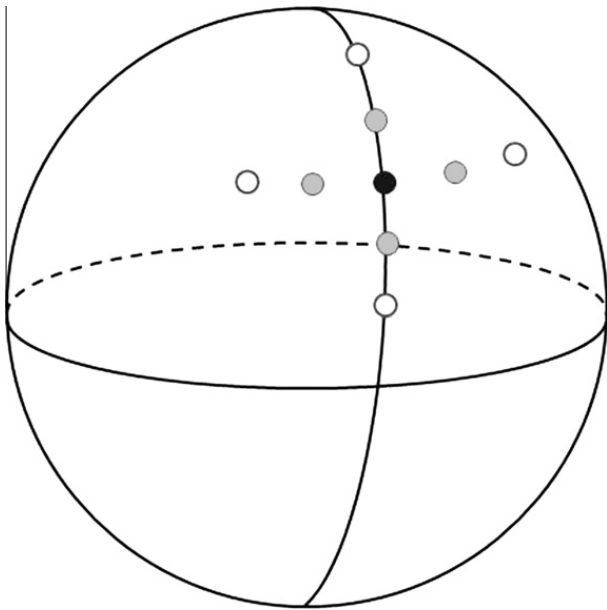


Fig. 3. Schematic diagram of the camera positions from which the different target and distracter views were taken. All viewpoints were equidistant from the center of a scene (not shown in the figure, but represented as the equatorial cross-section of the sphere), and were oriented toward the center of the scene. The camera positions are represented by the small circles. The black circle represents the Interpolated perspective, the four grey circles represent the four Training perspectives, and the four white circles represent the Extrapolated perspectives. This figure appeared as Fig. 1 in Waller et al. (2009), and is used with permission.

ent experiment. As can be seen, the new stimuli have lost much of their 3D perspective information. What is retained are the relative sizes, distances, and angles between objects as a function of the changes in viewpoints as well as the objects' occlusion relationships. However, the lack of semantic content appreciably weakens their interpretation as 3D scenes.

We created either six (for Training and Extrapolated views) or nine (for the Interpolated view) stimuli from each of the nine viewing perspectives. For each view, one stimulus depicted the correct, to-be-learned arrangement of the rectangles; the additional stimuli were distractors that portrayed the scene with the locations of two of the rectangles switched. The Interpolated view had more distractors created for it than the other views because during the test phase for this experiment, the Interpolated view was displayed more frequently than any particular Training or Extrapolated view. We did this to equalize the number of Interpolated, Trained, and Extrapolated views.

Only two rectangles were switched for each distractor and only five of the nine rectangles could potentially be switched. The switches used to create distractors represented a randomly selected subset of possible switches given these five rectangles. Thus, the distractors were designed specifically to disrupt the relative position relations among the array of rectangles, making it virtually impossible to do the task without taking account of object-to-object relations. In addition, and notably, we used different distractors for the training and test trials so that during testing, the distractors, Interpolated, and Extrapo-

lated stimuli were equally novel; only the Trained stimuli were familiar. In sum, for each block of the testing sequence, there were two repetitions of the four familiar Trained stimuli together with two repetitions of their unique novel distractors, two repetitions of four novel Extrapolated stimuli with two repetitions of their unique novel distractors, and eight repetitions of the novel Interpolated stimuli with one repetition of each of eight unique novel distractors.

The experiment was controlled through a computer using EPrime software (Psychological Software Tools, Pittsburgh, PA). Stimuli were presented on a 32.5 cm × 24 cm CRT monitor (85 Hz. refresh rate). Participants responded by pressing buttons on a response box connected to the serial port of the computer.

2.1.3. Procedure and design

The procedure and design were identical to Waller et al. (2009, Experiment 1). It is worth noting that throughout our research, we have used a discrimination learning task to test the view combination approach. We have done so for several reasons. First, as a direct comparison to a normalization approach for identifying abstract stimuli, it is important to know that the generalization functions are actually “created” in long-term memory. Researchers investigating normalization using novel objects (e.g., Tarr, 1995) have thus also used some sort of traditional long-term memory learning task, as have many others who examine scene recognition with unrelated objects (e.g., Mou & McNamara, 2002). Further, though much has been learned from tasks that use rapid stimulus presentation (e.g., ~1 s per view; e.g., Castelano et al., 2009), these tasks are perhaps more appropriate for stimuli that may already be presumed to be in long-term memory because they are familiar. Tasks that use such short display times tend to be more similar to change detection or short-term memory tasks than to recognition or categorization tasks and may thus involve different cognitive strategies.

The participants read instructions that informed them they would be viewing many different arrangements of rectangles and that one particular arrangement was correct. They were instructed to press a green button labeled “Correct” if the arrangement was correct, and a red button labeled “Incorrect” otherwise. Participants were also told that a randomized half of the pictures depicted the correct arrangement and the other half were incorrect. From the participants' point of view, the learning task was to respond whether a given stimulus was “in the correct configuration” or not, using one of two response keys to indicate their answer.

2.1.4. Training trials

During training, participants were required to distinguish the four training stimuli (“views” from the grey circles in Fig. 3) from twelve different distractors (three for each Trained viewpoint). To increase their motivation to work efficiently, feedback was given over headphones; the feedback message said “three points” if participants were correct and answered in less than one second, “two points” if they were correct and answered in one second or more, or “wrong” if they were incorrect. Participants

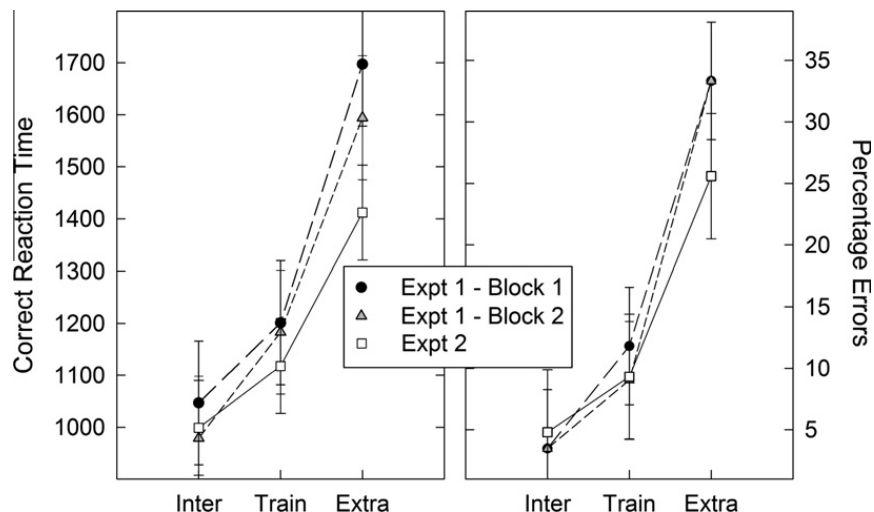


Fig. 4. Correct RTs (left panel) and percentage error rates (right panel) as a function of viewpoint and block for Experiments 1 and 2. Inter = the Interpolated view; Train = the Trained views; Extra = the Extrapolated view. For Experiment 1, error bars are 95% confidence intervals, computed from the block by view interaction sums of squares in the omnibus ANOVAs; for Experiment 2, error bars are 95% confidence intervals computed from the one-way repeated measures ANOVAs on view separately for each group (Loftus & Masson, 1994).

were told that the feedback would stop partway through the experiment but that they would still receive points for correct responses. However, the points had no tangible consequences for the participants.

Each training trial began with a warning beep for 1 s, followed immediately by the stimulus, which was displayed until the participant responded. There was a 1 s delay before the feedback message and then a 250 ms delay before the next trial. Training trials were administered in randomized blocks of 24, with the 12 distractors each presented once and the target presented 12 times (three times from each training view). Participants were required to complete at least two training blocks. If accuracy exceeded 80% in the second or any subsequent training block they proceeded to the testing portion. The number of training blocks ranged between 2 and 13, with a mean of 5.5 blocks and a SD of 3.75.

2.1.5. Test trials

Trials for the testing portion of the experiment used the identical procedure as the training trials, except there was no feedback message. Testing consisted of 96 trials, composed of two blocks of 48 trials. Stimuli representing the Training, Interpolated, and Extrapolated viewpoints were each presented 16 times within each block, and for each of these views, half of the stimuli were targets and half were distractors. Trials depicting Training and Extrapolated viewpoints presented equal numbers of stimuli from each of the four viewpoints. The order of the trials was randomized separately for each block and participant.

2.2. Results and discussion

The reaction times (RTs) to correctly answered trials were averaged across the Interpolated, Trained, and Extrapolated viewpoints separately for each block, and analyzed in a 2 (Block: first or second) \times 3 (View: Interpolated, Trained, Extrapolated) repeated measures analysis of

variance (ANOVA).¹ Only the main effect of view was significant, $F(2, 34) = 16.82$, $p < .0001$, $\eta_p^2 = .49$. The means are shown in the left panel of Fig. 4. Planned contrasts showed that average RTs to the Interpolated viewpoint (1013 ms) were significantly shorter than to the Trained viewpoint (1192 ms); $F(1, 17) = 9.27$, $p < .01$, $\eta_p^2 = .35$, which in turn were significantly shorter than those to the Extrapolated viewpoints (1646 ms); $F(1, 17) = 14.04$, $p < .001$, $\eta_p^2 = .45$.

The percent error data mirrored the RT data in every detail (Fig. 4, right panel). The overall ANOVA yielded only a main effect of viewpoint, $F(2, 34) = 23.19$, $p < .0001$, $\eta_p^2 = .58$. The contrasts showed that the percent error made to the Interpolated view (3.47%) was less than to the Trained views (10.42%); $F(1, 17) = 8.63$, $p < .01$, $\eta_p^2 = .34$, which in turn were less than to the Extrapolated views (33.33%); $F(1, 17) = 23.64$, $p < .001$, $\eta_p^2 = .58$. In a signal detection analysis of recognition sensitivity, the d' for the Interpolated, Trained, and Extrapolated views were 2.34, 0.98, and 0.85, respectively, $F(2, 34) = 41.08$, $p < .0001$, $\eta_p^2 = .71$.

In sum, the present data replicate our previous results with real-world scenes (Waller et al., 2009). By doing so, they extend the view combination model to visual arrays that have little or no formal object or scene structure and no semantic interpretation, and thus broaden the applicability of that approach while constraining models of visual recognition based on normalization.

3. Experiment 2

We designed Experiment 1 so that the number of presentations per “view type” (Trained, Interpolated, and Extrapolated) was equal. Because there was only one Inter-

¹ One participant had no correct responses for the Extrapolated stimulus targets; his mean for that condition was replaced with the group mean. Although this procedure reduces the variability in that condition, in this case it works against the present hypothesis.

polated stimulus and several Trained and Extrapolated stimuli, this meant that more instances of the Interpolated view were presented than any single Trained view during testing. It is thus plausible that the speed and accuracy advantage observed for the Interpolated view derived from this inequality of presentation frequency during test (Zaki & Nosofsky, 2007). In Experiment 2, therefore, we presented each specific novel test view (i.e., all nine views) an equal number of times during test, so that, for example, the combined Trained views were seen 16 times while the Interpolated view was seen only four times. If relative frequency of specific stimuli during testing was responsible for the enhanced performance on Interpolated compared to Trained views, that advantage should disappear in Experiment 2.

In addition to equating for the frequency of views, in Experiment 2, we added an additional first test trial for all participants that was a target trial. That is, the view presented on the first test trial for each group was seen five times during testing, whereas all other views were seen only four times. For about 1/3 of the participants, the first test trial was the Interpolated view, about 1/3 saw one of the four Trained views, and about 1/3 saw one of the four Extrapolated views. This allowed us to determine whether the Interpolated view would be responded to faster and/or more accurately than either the Trained or Extrapolated views on the very first test trial. We also analyzed the data without the 1/3 of the participants whose first trial was the Interpolated view, to ensure that the difference between the three view types was maintained within-subjects in the remaining two groups when, in fact, there were only four (rather than five) Interpolated views presented.

3.1. Method

3.1.1. Participants

Ninety-four undergraduates (63 men and 54 women) from Miami University participated in the experiment in return for credit in their introductory Psychology course. Twenty-three did not complete the training task to the 80% criterion in the allotted time, leaving 50 men and 44 women who provided data for test trials. The mean age of the participants was 19.2 years ($SD = 0.90$). They were randomly assigned to one of three groups defined according to whether their first test stimulus was the novel Interpolated target ($N = 31$), an old Trained target ($N = 32$), or a novel Extrapolated target ($N = 31$).

3.1.2. Materials, procedure, and design

The training stimuli were identical to those used in Experiment 1. The number of training blocks ranged between 2 and 16, with a mean of 5.9 blocks and a SD of 3.59. For the testing stimuli, excluding the very first stimulus, there were two blocks of 36 trials, half of which were targets. Among the targets for each block, there were two presentations of the novel Interpolated view, two presentations of each of the four Trained views, and two presentations of each of the novel Extrapolated views. Thus, across blocks, there were four replications of each individual stimulus, which were thus equated for frequency. In addition, two novel distractors were created for each test-

ing stimulus; in each of these two objects switched places with each other. Thus, as in Experiment 1, the Trained views were the only familiar views during the testing sequence. Finally, the first test trial for each participant was forced to be a target trial; as noted above, 31, 32, and 31 participants were shown either the Interpolated target view, one of the four Trained target views, or one of the four Extrapolated target views. In the latter two cases the particular view seen on the first test trial was approximately balanced across participants. Except for the first test trial, the other test stimuli were presented in a different randomized order for each participant. The procedure was otherwise identical to that used in Experiment 1.

3.2. Results and discussion

We examined correct RTs on the first test trial for each of the three groups for which the first test trial was either an Interpolated, Trained, or Extrapolated target view, using a one-way between-subjects ANOVA. The effect of group did not reach significance, $F(2, 82) = 1.71$, $p = .187$, though the means were in the expected direction: 3262, 3983, and 4321 ms for the Interpolated, Trained, and Extrapolated groups, respectively. The percent of participants who were correct on their first trial when it was an Interpolated, Trained, or Extrapolated target was 93.5%, 90.6%, and 87.1%, which is also in the expected direction.

The correct RTs to the target stimuli presented during test (excluding the first test stimulus, which thus made the number of potential Interpolated, Trained, and Extrapolated targets equal to four, 16, and 16 for each of the experimental groups) were analyzed in an ANOVA in which view (Interpolated, Trained, and Extrapolated) was within-subjects and group (the first test view was either the novel Interpolated view, one of the four novel Extrapolated views, or one of the Trained views) was between-subjects. We did not include Testing Block as a factor because there were only two prototype stimuli per block and seven subjects had no correct trials for one of the other views within a block.

The effect of view was significant, $F(2, 182) = 62.76$, $p < .0001$, $\eta_p^2 = .41$, and is shown in the left panel of Fig. 4. The means for the Interpolated, Trained, and Extrapolated views were 999 ms, 1118 ms, and 1412 ms, respectively. Neither the main effect of group nor its interaction with view were significant, $F_s < 1.00$. The main effect of view was also significant when we excluded the group whose very first trial was an Interpolated view, $F(2, 122) = 43.52$, $p < .0001$, $\eta_p^2 = .42$. We did this to ensure that the advantage for the Interpolated view would remain clear when there were only four such novel test views compared to 16 Trained views and 16 novel Extrapolated views. The means for the Interpolated, Trained, and Extrapolated views in this analysis were 1054 ms, 1180 ms, and 1458 ms, respectively.

Because there was no evidence of a Group \times View interaction in the overall ANOVA, $F(4, 182) < 1.00$, we tested the planned contrasts collapsing over the group factor. The Interpolated view was responded to significantly faster than the average of the Trained views, $F(1, 93) = 15.53$,

$p < .0002$, $\eta_p^2 = .14$. In turn, the Trained views were responded to significantly faster than the Extrapolated views, $F(1, 93) = 65.64$, $p < .0001$, $\eta_p^2 = .41$.

For percent of errors, the view by group ANOVA also yielded only a main effect of view, $F(2, 182) = 60.10$, $p < .0001$, $\eta_p^2 = .40$. The means for the Interpolated, Trained, and Extrapolated views were 4.79%, 9.31%, and 25.60% and are shown in the right panel of Fig. 4. Because the view by group interaction was again not significant, $F(4, 184) = 1.09$, $p = .36$, we again ignored the group factor to test the planned contrasts. Performance on the Interpolated view was significantly more accurate than performance on the Trained views $F(1, 93) = 10.92$, $p < .002$, $\eta_p^2 = .11$, which in turn was more accurate than performance on the Extrapolated views, $F(1, 93) = 60.85$, $p < .0001$, $\eta_p^2 = .41$. Finally, an ANOVA that excluded the group whose first trial was an Interpolated view also had a significant main effect of view, $F(2, 122) = 37.69$, $p < .0001$, $\eta_p^2 = .38$. The mean percent errors for the Interpolated, Trained, and Extrapolated views in this analysis were 2.42%, 7.46%, and 20.57%, respectively.

Analyses of d' also revealed a main effect of view, $F(2, 182) = 71.48$, $p < .0001$, $\eta_p^2 = .44$, and no interaction between group and view, $F < 1.00$. The means for the Interpolated, Trained, and Extrapolated views were 2.55, 1.61, and 1.25, respectively. The planned contrasts across groups revealed that the difference in d' between the Interpolated and Trained views was significant, $F(1, 93) = 67.38$, $p < .0001$, $\eta_p^2 = .42$, and the difference between the Trained and Extrapolated views was also significant, $F(1, 93) = 16.35$, $p < .0002$, $\eta_p^2 = .15$.

The present data, as well as those from Experiment 1, provide two independent demonstrations that a novel Interpolated view of these relatively abstracted 2D stimuli was recognized better than an old Trained view, in terms of both speed and accuracy. In the case of Experiment 2 this finding cannot be the result of differences in frequency of presentation of the different views during the testing sequence.

4. Experiment 3

Experiments 1 and 2 demonstrate that view combination can explain the recognition of stimuli that cannot be explained by other recognition models. Nevertheless, as pointed out in the introduction, view combination was proposed as a model for recognition of 3D objects (e.g., Edelman, 1999a; Edelman & Bulthoff, 1992; Ullman, 1998) – even very unusual ones (e.g., bent straws). Although generalizing view combination to apply to scenes (e.g., Friedman & Waller, 2008; Waller et al., 2009) was an important advance, it may also be somewhat unsurprising, given that there are many respects in which a 3D scene retains a 3D structure that is analogous to an object (e.g., the perceived relations between scene's/objects' parts change systematically in accordance with viewpoint changes).

In Experiments 1 and 2, we used rectangles to cover the original playground objects in each of the nine views. Although this process minimized the scenes' semantic content, the stimuli still retained some cues to 3D structure such as the relative locations, relative sizes, distances,

and angles between objects as a function of the changes in viewpoints, in addition to the objects' occlusion relationships. Arguably, each of these features, either alone or in combination, provided cues to the scene's underlying 3D structure. Consequently, if view combination operates by interpolating between only 3D metric properties we may not have completely ruled out 3D information as an underlying contributor to either the view combination effect or the prototype enhancement effect.

In Experiment 3, we attempted to understand some of the boundary conditions of view combination and the prototype enhancement effects by removing virtually all 3D information from the stimuli. In particular, each rectangle in each viewpoint (Trained, Interpolated, Extrapolated) was reduced to a single colored dot (see Fig. 2, right side). After being converted to dots, there was no occlusion and the stimuli had only identity (color) and location cues; relative size and shape were identical across objects and views. Indeed it seems quite unlikely that one can extract a 3D scene structure from these stimuli.

There are three possible predictions for how people will recognize arrays with such minimal cues to 3D structure. First, it is possible that view combination does not apply to stimuli that lack nearly all 3D structure. If this is the case, then neither a view combination effect nor a prototype enhancement effect should occur (i.e., Interpolated stimuli should be recognized no more efficiently than Extrapolated or Trained stimuli) Second, if view combination applies to recognition processes regardless of the dimensionality of the stimuli, we would expect similar results as those in Experiments 1 and 2 – Interpolated stimuli should be recognized more efficiently than both Extrapolated and Trained stimuli. Finally, if view combination is more effective with 3D than 2D stimuli, even without semantic content in either, and if we have succeeded in removing important aspects of 3D structure, we might observe a view combination effect without the prototype enhancement effect. Either of the latter two findings (view combination with or without an enhanced prototype effect) with the present dot stimuli would constitute the strongest direct evidence to date that view combination is a fundamental aspect of visual recognition, and would be indirect support for the assumption that it works through generalization mechanisms. Further, because the primary difference between Experiment 2 and the present experiment involves the information that was removed from the stimuli, we may also thus begin to be able to say what important features in a stimulus constitute a 3D “scene”.

4.1. Method

4.1.1. Participants, design, and procedure

There were 39 (19 men, 20 women) volunteers from the same pool as in Experiments 1 and 2. Seven people (2 men, 5 women) did not reach the 80% criterion in the time allotted, and one male volunteer had virtually no correct responses so his data were not considered further. The remaining participants required between 2 and 9 training blocks ($Mean = 3.31$, $SD = 1.7$); their mean age was 19.13 ($SD = 1.09$). The design and procedure was identical to that

of Experiment 2 with one exception. Because the analysis of the first test trial in Experiment 2 was inconclusive, we did not divide the subjects into three groups to examine their first RTs.

4.1.2. Stimuli

The stimuli were created by computing the center of each rectangle in each target and distractor stimulus and drawing a colored dot 6 mm in diameter around its center (see Fig. 2). Therefore, none of the stimuli were occluded in any of the viewpoints and all stimulus locations were the same size and shape; that is, relative size and shape did not change with viewpoint.

4.2. Results

The correct RTs for “same” trials were averaged over the Interpolated, Trained, and Extrapolated trials (1744, 1864, and 2325 ms, respectively) and submitted to a one-way ANOVA; the main effect was significant, $F(2, 60) = 11.43$, $p < .0001$, $\eta_p^2 = .276$. The planned contrast between the Interpolated and Trained stimuli was not significant, $F(1, 30) = 1.25$, $p = .27$. However, both the difference between the Interpolated and Extrapolated stimuli, $F(1, 30) = 16.87$, $p < .0001$, $\eta_p^2 = .360$, and between the Trained and Extrapolated stimuli, $F(1, 30) = 11.90$, $p < .002$, $\eta_p^2 = .284$, were significant.

The same pattern was true for the percent error data; the means for the Interpolated, Trained, and Extrapolated trials were 0.80%, 3.0%, and 16.5%, $F(2, 60) = 18.71$, $p < .0001$, $\eta_p^2 = .384$. In addition, the contrast between the Interpolated and Trained stimuli was not significant, $F(1, 30) = 1.60$, $p = .22$, whereas the contrasts between the Interpolated and Extrapolated stimuli, $F(1, 30) = 20.04$, $p < .0001$, $\eta_p^2 = .401$, and between the Trained and Extrapolated stimuli, $F(1, 30) = 23.33$, $p < .0001$, $\eta_p^2 = .437$, were both significant.

The mean d' measures for the Interpolated, Trained, and Extrapolated conditions were 2.60, 2.25, and 1.88, $F(2, 60) = 7.64$, $p < .001$, $\eta_p^2 = .203$. The difference between the Interpolated and Trained stimuli was not significant, $F(1, 30) = 3.14$, $p = .087$, but the difference between the Interpolated and Extrapolated stimuli, $F(1, 30) = 10.84$, $p < .003$, $\eta_p^2 = .265$, and between Trained and Extrapolated stimuli, $F(1, 30) = 8.83$, $p < .006$, $\eta_p^2 = .227$, were both significant.

In summary, for all three measures, there was a viewpoint combination effect, but not a prototype enhancement effect. The Interpolated view was processed more efficiently than the Extrapolated views, but no better (or worse) than the Trained views. This finding suggests that view combination works more efficiently when there are more cues to 3D structure, either at the object or scene level.

5. General discussion

In three experiments, participants were trained to discriminate four particular 2D stimuli comprised of colored rectangles or dots from distractors in which two of the

items had switched positions; all of the stimuli were constructed from a set of 3D playground scenes by masking the objects in the scenes (for rectangles) or using their centers (for dots) and removing background information. After training with the rectangular stimuli, participants went on to recognize a novel stimulus within the span of the training views more quickly and accurately than they recognized the training stimuli themselves; the training stimuli, in turn, were recognized better than novel views outside of the span of the training range but equidistant to the training views. Experiment 2 eliminated frequency of presentation at test as a possible explanation for the observed findings. Importantly, the prototype enhancement effect was eliminated when the stimuli were simplified to mere dots positioned at the center of each of the rectangles for each view. Nevertheless, we still obtained a view combination effect with dot stimuli, which retained only relative location and identity information. We and others have previously documented a view combination effect with static objects (Edelman & Bülthoff, 1992; Friedman et al., 2005) and scenes (Castelano et al., 2009; Friedman & Waller, 2008; Waller et al., 2009). The present study generalizes these previous findings by providing evidence that view combination can explain performance on relatively impoverished 2D stimuli but does so better when those stimuli have some cues (e.g., occlusion) to 3D structure.

The present results add to the theoretical understanding of object and scene recognition by considerably constraining the generality and utility of a normalization account. Simply put, such an account is unable to explain how novel stimuli can be recognized as well as or more efficiently than Trained views. For example, normalization models (e.g., Tarr & Pinker, 1989) could only be applied to the type of arrays used in the present experiments by positing processes that infer the kinds of viewpoint transformations that had been performed on the 3D stimulus configurations. The inverse of those transformations (or some variant) could then be used to perform recognition of the 2D arrays. Yet assuming that this could be done, normalization models would still not predict equal or better performance on the Interpolated view than on the learned views because the transformational process(es) would have to take place.

Similarly, a normalization account cannot easily explain how Extrapolated views at the same transformational distance as the Interpolated view are recognized more poorly than the Interpolated view. However, on the view combination account, in the extrapolated case the more distant learned view(s) will tend to reduce the similarity between the input view and the view constructed from activation of the stored exemplars; the Extrapolated view should be correspondingly more dissimilar to the Trained views' physical dimensions taken together, and thus more difficult to recognize. Thus, both view combination per se and the enhanced prototype effect for either objects or scenes are very easy to account for when they are conceptualized as a function of physical similarities amongst representations in memory on many physical dimensions. Notably also, the enhanced prototype effect has been found in other tasks (e.g., Palmeri & Nosofsky, 2001).

Consequently, although there is empirical evidence for monotonicity in object (e.g., Tarr, 1995) and scene (e.g., Diwadkar & McNamara, 1997) recognition, these results (and others) can be explained more generally (and parsimoniously) by a view combination mechanism of generalization. For example, if the activated representations of the Trained views are relatively far apart, or if the generalization functions are relatively narrow due to overlearning (which decreases the variance of the functions), or the novel input does not span the training range, or there is only one training view, a view combination account can predict performance to decline roughly monotonically with distance between the input and the activated generalization gradient(s) (Bülthoff & Edelman, 1992; Edelman, 1999a; Friedman & Waller, 2008; Friedman et al., 2005). Of course, view combination accounts can also predict non-monotonicity. Both types of predictions can be verified (or made a priori) with metrically derived psychophysical similarity functions (see Shultz, Chuang, & Vuong, 2008) precisely because view combination relies on metric similarities.

A related theoretical advantage of the view combination account of recognition is that it allows similarity to be defined much more broadly than “transformational distance”, because similarity may encompass the many dimensions that underlie structural correspondences, as well as surface features and dynamic cues (e.g., Friedman, Vuong, & Spetch, 2010). Thus, view combination as a mechanism for recognition accounts for a broader class of empirical results, and prescribes the circumstances under which various forms of recognition occur more successfully than normalization. The present results thus provide substantial evidence in favor of a view combination approach to visual recognition in general and against the generality of normalization processes. They also provide strong evidence in favor of view combination as a possible “law of generalization” (Shepard, 1987) for visual recognition of a wide range of stimuli.

The present evidence that view combination mechanisms function in the absence of information about either object or scene structure (Experiment 3) thus places these mechanisms in a more general theoretical context than how they are typically conceptualized and suggest that they are fundamental processes that may support a variety of cognitive functions. Indeed, it is possible that multidimensional similarity spaces, distance metrics, and generalization principles underlying the summation of activation from multiple stored representations could function effectively as a recognition system for many kinds of stimuli, both visual and nonvisual. After all, generalization is a notion that has been construed as the basis of learning, transfer, recognition, and categorization for human and nonhuman animals alike for many years (e.g., Guttman & Kalish, 1956; Hull, 1943; Shepard, 1958, 1987; Spence, 1937; Nosofsky, 1986). Thus, it is possible that the type of view combination documented here – in which an unseen view of a token is better recognized than a familiar one – is the same kind of generalization that underlies performance when people categorize exemplars of different prototypes. For example, in elaborating his original view interpolation/combination

model, Edelman (1999a) allowed that “the same mechanism capable of identifying familiar objects would help make sense of novel ones.” (p. 91). That is, because activation is based on metric similarities novel instances can, in principle, activate stored representations in proportion to levels that correspond to their similarity on the represented dimensions; if there is enough activation overall, the novel instance can be understood as an instance of an extant category.

In the early categorization literature, Posner and Keele (1968, Experiment 3) had people learn four random dot patterns formed systematically from each of three meaningful categories (a triangle, and the letters *M* and *F*). Each exemplar was constructed via a random perturbation at three levels of distortion of each of the dots in the unseen prototype arrays. Participants were tested on the training items as well as on the non-presented category prototypes and other new stimuli at the untrained, more distorted levels. These conditions are analogous to the Training, Interpolated, and Extrapolated conditions in the present experiments, respectively. Posner and Keele's (1968) prototype stimuli were not better recognized than their training stimuli, but they were better recognized than the more extremely perturbed patterns (e.g., the Extrapolated stimuli). They concluded that their findings singled out the prototype pattern as unique, and that “this proposition is stronger than a generalization notion because the schema pattern is, on the whole, as well recognized as the exemplars from which it is abstracted (p. 362)”. In contrast to this claim, the view combination approach predicts that many novel stimuli whose metric properties fall within a given range of training views will be at least as well recognized as the training views, and occasionally better recognized. This is not a different or stronger proposition than generalization; rather it is exactly how generalization should work if sufficient excitatory activation is summed.

The “breakdown” of the prototype enhancement effect in Experiment 3 provides additional clues about the boundary conditions of view combination. In particular, we speculate that view combination works optimally when combining information about the 3D spatial structure of either objects or scenes. According to this view, the rectangle stimuli in Experiments 1 and 2 produced the prototype enhancement effect because they were like scenes “stripped down” to their essential elements and there were no irrelevant features to make different views of the same scene less similar to each other. It is possible that the cues remaining in our rectangle stimuli are some of the features that underlie 3D scene structure and that these features thus help constitute a psychologically-relevant functional definition of a “scene.” In particular, in Experiments 1 and 2 several scene features could have contributed to higher similarity between the training views and the novel Interpolated view, and therefore to better recognition of that view. As a function of viewpoint, these features include: (a) absolute locations of scene elements; (b) object-to-object relations among scene elements; (c) identity, in the form of color; (d) relative locations; (e) occlusion relationships; and (f) relative sizes and shapes among scene elements. Clearly, further

study is needed to validate these features as either necessary or sufficient for scene recognition and to examine other features that we have not mentioned here (e.g., semantic coherence).

In Experiment 3, Interpolated stimuli were more efficiently recognized than Extrapolated stimuli, despite a lack of occlusion relationships and relative sizes and shapes, both of which are clues to 3D structure in even simple 2D arrays. These findings invite the speculation that object identity and location, from which object-to-object relations can be derived, are potentially necessary, and certainly sufficient, for achieving a view combination effect. Additional cues to 3D structure (e.g., occlusion) are perhaps necessary for achieving a prototype enhancement effect. Of course, this does not necessarily imply that occlusion, absolute location, and so on, which are usually view specific, are always necessary for scene recognition. Rather, it is likely that some of these features support scene recognition and are acted on by relatively domain-general recognition mechanisms, such as view combination, to achieve “typical” scene recognition. Ongoing research aims to determine which of these attributes are necessary and which are sufficient in facilitating recognition.

A growing body of literature supports the notion that objects, scenes, visual arrays, and even dot patterns, despite differing radically in their physical characteristics, may all be processed similarly in human recognition and categorization. From an evolutionary point of view, this makes sense: A single all-purpose recognition system is more economical than separate recognition systems for different kinds of visual stimuli. Because object and place recognition, as well as navigation, are so important across the animal kingdom, it is not surprising that mechanisms are in place that integrate information across views of even very simple stimuli. And if stimulus generalization is to take its place as a psychological law, then it should function across many different species, stimuli, and situations.

Acknowledgement

Portions of this research were supported by a grant to the first author from the Natural Sciences and Engineering Research Council of Canada. Work by the fifth author on this project was supported by a gift from Ted Smith to Miami University. We thank Bernd Kohler and Eric Littman for assistance with preparing the stimuli, programming, and conducting the experiments and Quoc Vuong for comments on an earlier version of the manuscript.

References

- Biederman, I. (1987). Recognition by components: A theory of human image understanding. *Psychological Review*, *94*, 115–147.
- Bülthoff, H. H., & Edelman, S. (1992). Psychophysical support for a 2-D view interpolation theory of object recognition. *Proceedings of the National Academy of Science*, *89*, 60–64.
- Castelhano, M. S., Pollatsek, A., & Rayner, K. (2009). Integration of multiple views of scenes. *Attention, Perception, & Psychophysics*, *71*, 490–502.
- Diwadkar, V. A., & McNamara, T. P. (1997). Viewpoint dependence in scene recognition. *Psychological Science*, *8*, 302–307.
- Edelman, S. (1999a). *Representation and recognition in vision*. Cambridge: MIT Press.
- Edelman, S. (1999b). Class similarity and viewpoint invariance in the recognition of 3D objects. *Biological Cybernetics*, *72*, 207–220.
- Edelman, S. (2002a). Multidimensional space: The final frontier. *Nature Neuroscience*, *5*, 1252–1254.
- Edelman, S. (2002b). Constraining the neural representation of the visual world. *Trends in Cognitive Science*, *6*, 125–131.
- Edelman, S., & Bülthoff, H. H. (1992). Orientation dependence in the recognition of familiar and novel views of 3D objects. *Vision Research*, *32*, 2385–2400.
- 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.
- 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.
- Friedman, A., Vuong, Q. C., & Spetch, M. (2010). Facilitation by view combination and coherent motion in dynamic object recognition. *Vision Research*, *50*, 202–210.
- Friedman, A., & Waller, D. (2008). View combination in scene recognition. *Memory & Cognition*, *36*, 467–478.
- Guttman, N., & Kalish, H. I. (1956). Discriminability and stimulus generalization. *Journal of Experimental Psychology*, *51*, 79–88.
- Hull, C. L. P. (1943). *Principles of behavior*. NY: D.Appleton-Century.
- Lawson, R., & Bulthoff, H. H. (2008). Using morphs of familiar objects to examine how shape discriminability influences view sensitivity. *Attention, Perception, & Psychophysics*, *70*, 853–877.
- Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*, 476–490.
- Mou, W., & McNamara, T. P. (2002). Intrinsic frames of reference in spatial memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *28*, 162–172.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.
- Palmer, S., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. In J. Long & A. Baddeley (Eds.), *Attention and performance* (Vol. IX, pp. 135–151).
- Palmeri, T. J., & Nosofsky, R. M. (2001). Central tendencies, extreme points, and prototype enhancement effects in ill-defined perceptual categorization. *Quarterly Journal of Experimental Psychology*, *54A*, 197–235.
- Poggio, T., & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, *343*, 263–266.
- Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*, 353–363.
- Shepard, R. N. (1958). Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology*, *55*, 509–523.
- Shepard, R. N. (1987). Toward a universal law of generalization. *Science*, *237*, 1317–1323.
- Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*, 701–703.
- Spence, K. W. (1937). The differential response in animals to stimuli varying within a single dimension. *Psychological Review*, *44*, 430–444.
- Spetch, M. L., & Friedman, A. (2003). Recognizing rotated views of objects: Interpolation vs. generalization by humans and pigeons. *Psychological Bulletin & Review*, *10*, 135–140.
- Shultz, J., Chuang, L., & Vuong, Q. C. (2008). A dynamic object processing network: Metric shape discrimination of dynamic objects by activation of occipito-temporal, parietal, and frontal cortex. *Cortex*, *18*, 1302–1313.
- Tarr, M. J. (1995). Rotating objects to recognize them: A case study on the role of view point dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin & Review*, *2*, 55–82.
- Tarr, M. J., & Gauthier, I. (1998). Do viewpoint-dependent mechanisms generalize across members of a class? *Cognition*, *67*, 73–110.
- Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, *21*, 233–282.
- Ullman, S. (1989). Aligning pictorial descriptions: An approach to object recognition. *Cognition*, *32*, 193–254.
- Ullman, S. (1998). Three-dimensional object recognition based on the combination of views. *Cognition*, *67*, 21–44.
- Ullman, S., & Basri, R. (1991). Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *13*, 992–1006.

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.

Wong, A. C. N., & Hayward, W. G. (2005). Constraints on view combination: Effects of self-occlusion and differences among

familiar and novel views. *Journal of Experimental Psychology: Human Perception & Performance*, 31, 110–121.

Zaki, S. R., & Nosofsky, R. M. (2007). A high-distortion enhancement effect in the prototype-learning paradigm: Dramatic effects of category learning during test. *Memory & Cognition*, 35, 2088–2096.