Spatiotemporal information during unsupervised learning enhances viewpoint invariant object recognition

Moqian Tian
Department of Psychology, Stanford University,
Stanford, CA, USA

Kalanit Grill-Spector
Department of Psychology, Stanford University,
Stanford, CA, USA
Stanford Neuroscience Institute, Stanford University,
Stanford, CA, USA

Recognizing objects is difficult because it requires both linking views of an object that can be different and distinguishing objects with similar appearance. Interestingly, people can learn to recognize objects across views in an unsupervised way, without feedback, just from the natural viewing statistics. However, there is intense debate regarding what information during unsupervised learning is used to link among object views. Specifically, researchers argue whether temporal proximity, motion, or spatiotemporal continuity among object views during unsupervised learning is beneficial. Here, we untangled the role of each of these factors in unsupervised learning of novel three-dimensional (3-D) objects. We found that after unsupervised training with 24 object views spanning a 180° view space, participants showed significant improvement in their ability to recognize 3-D objects across rotation. Surprisingly, there was no advantage to unsupervised learning with spatiotemporal continuity or motion information than training with temporal proximity. However, we discovered that when participants were trained with just a third of the views spanning the same view space, unsupervised learning via spatiotemporal continuity yielded significantly better recognition performance on novel views than learning via temporal proximity. These results suggest that while it is possible to obtain view-invariant recognition just from observing many views of an object presented in temporal proximity, spatiotemporal information enhances performance by producing representations with broader view tuning than learning via temporal association. Our findings have important implications for theories of object recognition and for the development of computational algorithms that learn from examples.

Introduction

Humans have a remarkable ability to recognize objects across transformations that induce large changes in their appearance. Changes in the viewpoint of objects relative to the observer, especially rotations in depth, present a significant challenge to the visual system as parts of an object visible from one view may only partially overlap those from another view. Even more challenging for the visual system is to obtain both generalization across different retinal images of the same object seen from different viewpoints, and fine-grained distinctions among different objects presented at the same viewpoint, which may have similar appearance. For example, a side view of a Toyota Corolla looks different from its front view, but the same side view of the Toyota Corolla looks similar to a side view of a Honda Civic. A substantial body of literature suggests that the ability to recognize objects across vast changes in viewpoint, or viewpoint invariant recognition, is learned (Bülthoff, Edelman, & Tarr, 1995; Hayward & Tarr, 1997; Logothetis, Pauls, Bülthoff, & Poggio, 1994; Poggio & Edelman, 1990; Tarr, Williams, Hayward, & Gauthier, 1998). Importantly, learning invariant recognition of novel objects can occur without supervision or explicit labels, solely from the natural viewing statistics. Understanding this kind of unsupervised learning is important as learning from the natural statistics is likely the main route that infants and animals learn about objects (DiCarlo & Cox, 2007; Fiser & Aslin, 2002; Hauser, Newport, & Aslin, 2001). However, there is intense debate regarding what information in the natural viewing statistics is used during unsupervised learning of viewpoint invariant recognition (Balas & Sinha, 2008; Harman &

Citation: Tian, M., & Grill-Spector, K. (2015). Spatiotemporal information during unsupervised learning enhances viewpoint invariant object recognition. *Journal of Vision*, *15*(6):7, 1–13, http://www.journalofvision.org/content/15/6/7, doi:10.1167/15.6.7.

doi: 10.1167/15.6.7 Received September 12, 2014; published May 11, 2015 ISSN 1534-7362 © 2015 ARVO

Humphrey, 1999; Liu, 2007; Wallis & Bülthoff, 2001) and how many object views are needed (Biederman, 1987; Bülthoff & Edelman, 1992; Bülthoff et al., 1995).

One perspective suggests that invariant object recognition is achieved by generating an internal representation consisting of a three-dimensional (3-D) model of each object describing its parts and configuration (Biederman, 1987). Exposure to one or a few object views that provide enough information about the parts and configuration of an object should be sufficient to build such a 3-D model. An implication is that training with a single view will generalize to views that are far apart from it (Wang, Obama, Yamashita, Sugihara, & Tanaka, 2005), as long as nonaccidental properties (features that maintain constant geometry across views, such as linearity, parallelism, curvilinearity, and symmetry) are visible (Amir, Biederman, & Hayworth, 2012). This learning may be particularly effective for recognition of objects from different categories that differ in their parts and configuration (basic-level recognition; e.g., car vs. boat), but may not be sufficient for discriminating among objects within a category that share similar parts and configuration (subordinate-level recognition; e.g., Toyota vs. Honda; Amir, Biederman, Herald, Shah, & Mintz, 2014). A second perspective suggests that the representation of objects is view-based (Bülthoff & Edelman, 1992; Grill-Spector et al., 1999; Logothetis & Pauls, 1995), enabling both basic and subordinate recognition (Bülthoff et al., 1995). According to view-based theories, during learning participants need to be exposed to multiple views of a novel object to learn its "view space" (that is, a continous space of all possible views of an object; Bülthoff & Edelman, 1992). However, these theories do not specify what is the expected view-tuning width around each learned view (Bülthoff & Edelman, 1992; Hayward & Tarr, 1997; Wallis, Backus, Langer, Huebner, & Bülthoff, 2009) and if the view tuning depends on learning parameters in addition to the structure of the object. Nevertheless, these theories suggest that spatial and/or temporal continuity among object views during unsupervised training may be key for linking object views into a coherent internal representation (DiCarlo & Cox, 2007; Liu, 2007; Sinha & Poggio, 1996; Wallis et al., 2009; Wallis & Bülthoff, 2001).

Several theories have been suggested to explain how temporal and spatial information can facilitate unsupervised learning of objects. The first hypothesis suggests that presenting participants with views of an object close together in time, namely, in *temporal proximity*, is sufficient for linking them (Liu, 2007) via an associative mechanism (Miyashita, 1988; Sakai & Miyashita, 1991). A second hypothesis suggests that learning the 3-D structure of an object is facilitated by its *motion information* (Balas & Sinha, 2008; Knappmeyer, Thornton, & Bülthoff, 2003; Kourtzi & Shiffrar, 1997, 1999; Roark, Barrett, Spence, Abdi, & O'Toole, 2003; Stone, 1998, 1999; Vuong & Tarr, 2004). Because the way visible features of an object change in time is unique to an object, observers can use motion information occurring from the consecutive presentation of object views to infer how an object may appear across rotation. A third hypothesis suggests that the visual system uses both spatial and temporal continuity, that is *spatiotemporal information*, to link among views of an object (Perry, Rolls, & Stringer, 2006; Vuong & Tarr, 2004; Wallis et al., 2009; Wallis & Bülthoff, 1999). Because under natural viewing conditions the appearance of an object tends to vary slowly over both time and space, both sources of regularity can be used by the visual system to link object views (DiCarlo & Cox, 2007; Isik, Leibo, & Poggio, 2012; Ullman, Harari, & Dorfman, 2012; Wallis et al., 2009). Critically, studies disagree whether spatiotemporal continuity or motion information during unsupervised learning provide additional benefits compared to temporal proximity (Balas & Sinha, 2008; Harman & Humphrey, 1999; Liu, 2007; Wallis et al., 2009; Wallis & Bülthoff, 2001).

The goal of the present study is to untangle these sources of information during unsupervised learning of novel 3-D objects and investigate the conditions under which specific kinds of information are particularly beneficial. Specifically, we asked does spatiotemporal continuity during unsupervised training provide better learning than temporal proximity alone? In Study 1 we asked if training with spatiotemporal continuity among object views produces better recognition performance than training with temporal proximity. In Study 2 we asked if motion among object views during unsupervised learning facilitates learning for either spatiotemporally or temporally continuous sequences of object views. In Studies 3 and 4 we asked if spatiotemporal continuity produces better generalization than temporal proximity and if the benefit depends on the number of trained views. In Study 5 we asked whether differences between spatiotemporal and temporal proximity learning might be due to the differential engagement of subjects during learning under different conditions.

Methods

To examine the effects of different sources of information on unsupervised learning, we generated sequences of novel 3-D object views in which we untangled the effects of temporal proximity, spatiotemporal continuity, and motion information among

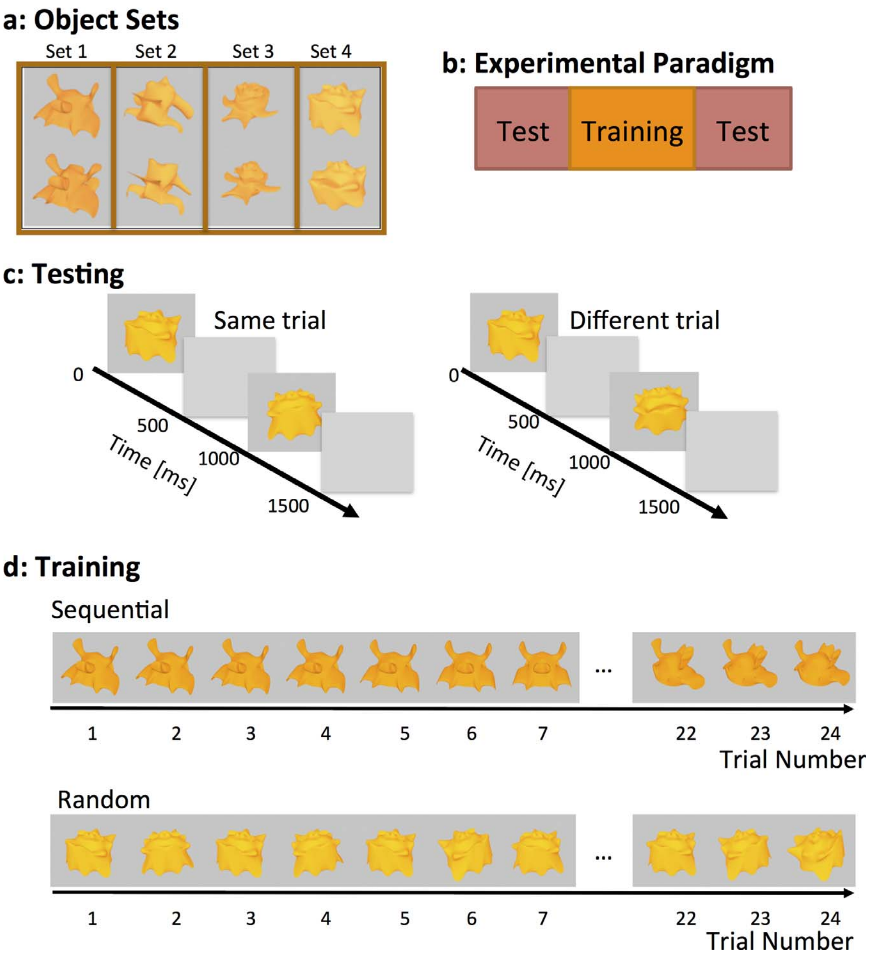

Figure 1. Procedure, paradigm, and stimuli. (a) Examples of the four object sets used in these experiments. (b) General experimental procedure for each object set. (c) Illustration of the discrimination testing paradigm: Subjects indicated whether two consecutive images were the same or different objects. In each trial subjects were presented with different views rotated 90° apart around the vertical axis. (d) Illustration of training paradigm: Subjects watched images of object views either in sequential or random order. The two objects in each set received the same type of training.

views. Note that these different kinds of sources of information are not equivalent.

First, sequences of object views containing spatiotemporal or motion information also contain temporal proximity, but the reverse is not true, as sequences of randomly presented views contain temporal proximity but may not contain spatiotemporal or motion information. Second, spatiotemporal continuity exerts an additional constraint compared to temporal proximity, as it requires a gradual shape change among temporally proximal images. Third, spatiotemporal continuity and motion information are not equivalent because motion can exist without spatiotemporal continuity and spatiotemporal continuity can occur without motion. For example, in random dot displays, apparent motion can occur even when there are no changes in luminance and it is impossible to track individual dots across frames (this is referred to as "second order motion"; Chubb & Sperling, 1989; Lu & Sperling, 1995). Further, it is possible to generate displays with high spatiotemporal correlation, but no perceived motion. For example, displays that change very slowly, below the rate in which motion integration occurs (Simons & Rensink, 2005). In the context of object recognition, Preminger and colleagues (Preminger, Blumenfeld, Sagi, & Tsodvks, 2009) showed that slowly and continuously changing the appearance of a photo of a face over days changed subjects' memory of the face. This is another example of how slow changes that are spatially and temporally proximal affect subjects' face perception, even when images are stationary. Thus, while in many cases motion and spatiotemporal continuity are coupled, they can be separated.

To generate a difficult recognition task that requires learning the entire 3-D structure of an object, not just a representative feature, we generated novel 3-D objects in sets where objects in a set contained the same parts and configuration (Figure 1a). Participants went through a testing phase, training phase, and a second testing phase (Figure 1b). During testing, participants were examined on a subordinate perception task, namely on their ability to determine if two images of an object rotated 90° apart were of the same object or taken from different objects (Figure 1c). We measured participants' improvement post- versus pretraining. During unsupervised learning, participants were presented with sequences of object views spanning a 180° view space of rotation around the vertical axis. We varied the viewing statistics by presenting the same views in different orders. In all experiments, half of the object sets were trained with temporal proximity alone in which object views appeared in random order (Supplemental Movie 1) and the other half of the object sets were trained with spatiotemporal continuity in which object views appeared in a sequential order (Supplemental Movie 2; Figure 1d).

Participants

The experimental protocol was approved by Stanford University's Human Subjects Institutional Review Board. Participants provided written informed consent before participating in the experiment. Participants received monetary compensation or course credit for their participation. Between 14 and 27 participants participated in each of the experiments. These are typical sample sizes used in perceptual learning studies.

Stimuli

Four sets of novel 3-D objects were created using 3DS Max (http://usa.autodesk.com/3ds-max/). Each set contained two objects that had the same parts and configuration, but differed in the shape of features and their aspect ratios (Figure 1a). Objects were yellow-colored to be visible over the gray background and symmetric around a vertical plane passing through their centers. All objects were rendered in the same lighting conditions and were shown on a uniform gray background. Stimuli were calibrated in a pilot study using different participants. We modified the object sets till participants' untrained discrimination performance was similar across sets. A one-way repeated measures analysis of variance (ANOVA) on subjects' performance found largely similar performance accuracy during the object discrimination task across different object sets before and after training in all five experiments; Study 1: before training: $F(3, 57) = 0.61$, $p > 0.05$; after training: $F(3, 57) = 0.87$, $p > 0.05$; Study 2: before training: $F(3, 57) = 0.99$, $p > 0.05$; after training: $F(3, 57) = 0.52$, $p > 0.05$; Study 3: before training: $F(3, 45) = 1.36$, $p > 0.05$; after training: $F(3, 45) = 0.14$, $p > 0.05$; Study 4: before training: $F(3, 45) = 4.62$, $p < 0.05$; after training: $F(3, 45) = 0.87$, $p > 0.05$; Study 5: before training: $F(2, 52) = 1.53$, $p > 0.05$; after training: $F(2, 52) = 0.97$, $p > 0.05$.

Apparatus

All experiments were run on 27-in. iMac computers using the Mac OS Mountain Lion operating system. Stimuli presentation was coded with Matlab (http://Mathworks.com) and Psychophysics Toolbox 3 (http://psychtoolbox.org).

Experimental protocol

All experiments began with an object-discrimination testing block, followed by a training phase, and then a postlearning object-discrimination testing block (Figure 1a). In Studies 1 through 4, each participant participated in four such sequences, one for each set of novel objects (Figure 1b). Half the object sets were learned using sequential sequences and half were learned with random sequences. In Study 5, each participant participated in three sequences, with one of them trained in sequential order, one in random order, and one without object training. All testing of learning effects are within-subjects.

Object discrimination testing

In each trial, two object images were presented sequentially, each for 500 ms, separated by a 500-ms blank screen (Figure 1c). The two images were of objects of a given set shown in two views rotated 90° apart from each other (rotation around the vertical axis; images drawn from the entire range of 0°–180° views). In half of the trials, the two views were of the same object (Figure 1c, left), and in the other half of the trials, the two views were of different objects from the same set (Figure 1c, right). The positions of the two objects were jittered randomly around fixation in a range of 0°–1.3°. Object size varied randomly from 10.4° × 10.5° to 12.4° × 12.4°.
Task: Participants were instructed to indicate via keypress on a keyboard if the images were two views of the same object or two views of different objects. Participants participated in 40 trials per object set in each testing block. In all experiments, subjects' accuracy in the discrimination test was measured using *d*-prime (d', Green & Swets, 1974). We also report the miss rate and false alarm rate separately to test if the improvement in performance comes from decreased miss rate, decreased false alarm rate, or both.

Training

Participants participated in an unsupervised training block in which they were exposed to different views (spanning a 180° rotation) of two objects of a given set either in a sequential or random order (Figure 1d). During the learning phase of each object set, participants received four blocks of training, with two blocks for each object from a given set. Participants were instructed to watch the images. Half the object sets were trained using sequential training and the other half using random training, in alternating order. Both objects of a given set were trained in the same manner (sequential or random). We counterbalanced across participants which object sets were used in the sequential and random conditions as well as the order of the conditions.

Study 1

Twenty participants were trained with 24 views of novel objects rotated in increments of 7.5° and were tested with the same views (Figure 1). During training each view was shown for 667 ms. Learning was done in blocks. In each block, 24 views of an object were shown. During sequential training the views were shown in sequence. Thus, the object appeared to rotate back and forth three times. During random training the same views were shown for the same duration and number of repetitions, but they appeared in random order. After a training block with one object, the subject was trained in an identical manner with the second object of the set. The whole process was repeated twice. Thus, during unsupervised training, participants were shown 24 views × 6 times × 2 sequences = 288 trials per object lasting 400 s. Videos of training sequences can be found in the supplemental materials (sequential training: Supplemental Movie 1; random training: Supplemental Movie 2).

Study 2

Twenty participants participated in Study 2, which used the same paradigm as Study 1 except that we compared sequential and random training with and without motion information. To suppress motion perception we inserted texture masks between consecutive object images in a sequence. Each intact image was presented for 417 ms and each mask was presented for 250 ms, yielding an identical training duration for the motion and no-motion (masked) sequences. Two of the object sets were trained with masks between object images (no motion) and two without masks (with motion). In each motion condition (motion/no motion), views of one object set were shown in sequence and views of another object set were shown in random order. We counterbalanced the assignment of the object sets across participants for the four experimental conditions.

Studies 3 through 5

Studies 3 through 5 were similar to Study 1 with the following changes: (a) pre- and posttesting was conducted on views that were not shown during training and (b) in Studies 4 and 5, only seven views were shown during training, each view was shown for 667 ms (Study 4) and 250 ms (Study 5), and testing was performed on 18 views spanning the 0°–180° that were not shown during training.

Study 3

Sixteen participants participated in Study 3. Testing was performed on object views that were along the in-depth rotation axis as training and were spaced 3.75° away from the trained views.

Study 4

Twenty-six participants participated in Study 4. Training used seven views spanning the same 180° along the vertical axis but spaced 30° apart. Each view was shown eight times in either a sequential or random sequence. In the sequential condition the object appeared to rotate back and forth four times. In the random condition the same views were shown in a random order. Each sequence was repeated twice yielding: 7 views × 8 times × 2 times = 112 training trials per object, which lasted about 80 s. Testing views were 7.5° or 15° away from the nearest trained view.

Study 5

Twenty-seven participants participated in Study 5. Subjects were tested and trained on the same views and procedure used in Study 4. To control for performance during unsupervised learning and to test whether the posttraining improvement is not just because of performing the discrimination test a second time, we implemented the following changes: (a) To control performance during sequential and random training, subjects were asked to monitor changes in an object's contrast during unsupervised training and press a key

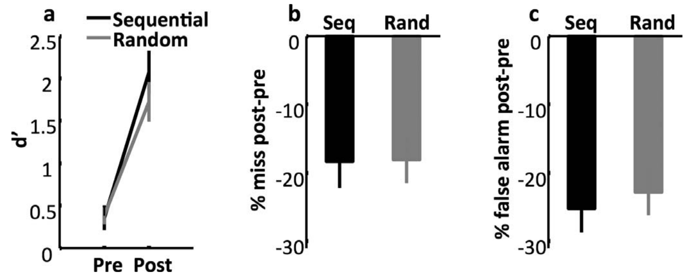

Figure 2. Study 1: Sequential and random learning similarly improve discrimination of in-depth rotated objects. (a) *d′* in discrimination performance before (pre-) and after (post-) training for sequential learning (black) and random learning (gray). (b) Learning-related decrease in the percentage misses. (c) Learning-related decrease in the percentage false alarms. Results are averaged across 20 subjects. Error bars indicate standard error of the mean (*SEM*). *Seq:* sequential. *Rand:* random training.

whenever an image momentarily decreased in contrast. (b) To test if training produces a larger improvement than from performing the discrimination task twice, we added a baseline, no-training condition. In this no-training baseline condition, the learning phase was substituted by a dot task. In the dot task, participants were asked to look at a centrally presented dot and press a key when its color changed. No object images were presented during the dot task. The dot was about 3° in visual angle, and its color changed randomly in 2–4 s intervals. The duration of the dot task was identical to the duration of the sequential and random training. We also changed the presentation duration of each image to be 250 ms to test the effects of presentation time on learning performance.

Results

Study 1: Temporal continuity is sufficient for learning to recognize novel 3-D objects across views

We compared unsupervised learning of 3-D novel objects following presentation of 24 object views spanning a 180° view space presented either in sequential order or random order. The former provides spatiotemporal information and the latter only temporal proximity. We considered three possible outcomes: (a) If spatiotemporal information is necessary for learning invariant object representations, learning will occur for sequential, but not random training sequences; (b) if spatiotemporal information produces more robust representations, the learning effect will be larger for sequential than random training sequences; and (c) if temporal proximity is sufficient, then there will be similar learning for random and sequential training sequences.

Prior to training, participants were poor at discriminating whether two views of objects rotated 90° apart in the image plane were of the same or different objects (accuracy: 55.2% ± 1.5%; *d′*: 0.37 ± 0.09). After training, participants' overall accuracy in discrimination increased (75.7% ± 2.7%; *d′*: 1.90 ± 0.21). A two-way repeated measures ANOVA of *d′*, with factors of training (pre/posttraining) and training sequence (sequential/random) revealed a significant effect of training (Figure 2a; $F(1, 19) = 61.85$, $p < 0.001$). There was no significant difference in the effect of learning across sequential and random training (no main effect of training sequence, $F(1, 19) = 0.49$, $p = 0.41$, and no interaction between training and training sequence, $F(1, 19) = 0.94$, $p = 0.33$). Improvement in both sequential and random training was due to a significantly reduced miss rate (Figure 2b; two-way ANOVA with factors of training × training sequence, main effect of training, $F(1, 19) = 37.00$, $p < 0.001$) and a significantly decreased false alarm rate (Figure 2c; main effect of training, $F(1, 19) = 77.50$, $p < 0.001$). There was no significant difference across training sequences on the reduction of either the miss or false alarm rate (Fs < 1; ps > 0.5).

Results of Study 1 revealed that unsupervised learning of novel 3-D objects was similar for sequential and random training sequences, suggesting that exposure to object views in temporal proximity is sufficient for improving object recognition performance.

Study 2: Motion information during unsupervised training does not enhance learning of view invariant representations of rigid 3-D objects

While Study 1 found no difference in performance after unsupervised training using sequential versus random sequences, both sequences contained motion information. During sequential training the object appeared to rotate smoothly (Supplemental Movie 1), and during random sequences the object's motion appeared irregular (Supplemental Movie 2). Therefore, in Study 2 we examined whether motion information during unsupervised training aids learning. The training was identical to Study 1 except that in half of the sequences, we reduced motion information by presenting masks between consecutive images. We examined three hypotheses: (a) If motion information is necessary for learning, training without motion information will be poorer than training with motion for both sequential and random sequences; (b) if motion aids learning, training with motion sequences will

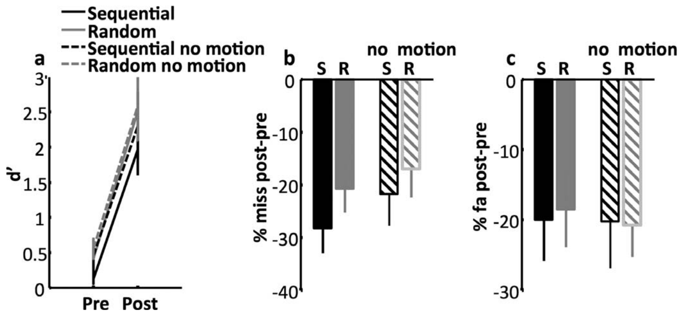

Figure 3. Study 2: Learning with or without spatiotemporal and motion information yields similar improvement in object discrimination across rotations in-depth. (a) *d*′ in discrimination performance pre- and posttraining for sequential learning with motion information (solid black), random learning with motion information (solid gray), sequential learning without motion information (dashed black), and random learning without motion information (dashed gray). (b) Learning-related decrease in the percentage misses. (c) Learning-related decrease in the percentage false alarms. Data are averaged across 20 subjects. Error bars indicate *SEM. S:* sequential. *R:* random training.

improve learning for both sequential and random sequences; and (c) if temporal proximity is sufficient, training with or without motion information will produce the same learning effect.

Study 2 revealed three notable results. First, we found a significant training effect across all four training conditions. The average discrimination accuracy increased from 56.7% ± 1.5% (*d*′: 0.39 ± 0.10) to 77.6% ± 3.7% (*d*′: 2.34 ± 0.36). A three-way repeated measures ANOVA on *d*′ with factors of training (pre/posttraining blocks) × training sequence (sequential/random) × motion (no motion/motion) found a main effect of training, $F(1, 19) = 37.12$, $p < 0.001$ (Figure 3a).

Second, performance did not significantly differ across sequential and random training sequences (Figure 3a; three-way ANOVA, no significant main effect of training sequence, $F(1, 19) = 1.60$, $p = 0.22$, and no significant interaction between training and training sequence, $F(1, 19) = 0.83$, $p = 0.37$). Third, the outcome of training was not different when motion information was available during training compared to when there was no motion information (Figure 3a; three-way ANOVA, no main effect of motion, $F(1, 19) = 0.12$, $p = 0.30$) and no interaction between training sequence and motion, $F(1, 19) = 0.08$, $p = 0.79$).

Learning resulted from both a decrease in miss rate (Figure 3b) and a decrease in the false alarm rate (Figure 3c). All four training conditions yielded a similar decrease in the miss rate (Figure 3b; three-way ANOVA, main effect of training, $F(1, 19) = 34.03$, $p < 0.001$; no significant interaction between training and training sequence, $F(1, 19) = 2.40$, $p > 0.05$) and no interaction between training and motion, $F(1, 19) = 1.13$, $p > 0.05$. We found similar decreases in the false alarm rate across training conditions (Figure 3c; three-way ANOVA, significant main effect of training, $F(1, 19) = 27.31$, $p < 0.001$), and no significant interaction between training and training sequence, $F(1, 19) = 0.01$, $p > 0.05$, or training and motion, $F(1, 19) = 0.06$, $p > 0.05$.

Study 2 replicated results of Study 1 in that sequential and random training yielded similar performance. Further, Study 2 found that motion did not facilitate unsupervised learning of novel 3-D objects because the training effect was similar when training with or without motion for both sequential and random sequences.

Studies 3 and 4: Spatiotemporal information yields better generalization to new object views than temporal information

The surprising finding in previous experiments is the similar improvement in object discrimination following spatiotemporal continuous and random training of object views. One interpretation is that temporal proximity is sufficient for linking among views shown during training, and that spatiotemporal continuity does not provide additional information. However, two additional factors may have contributed to the pattern of the results: (a) Training and testing were conducted on the same views; therefore, temporal proximity during training was sufficient to associate these particular object views, consequently obtaining a learning effect. (b) Training employed a large number of object views, which was perhaps more than sufficient for generating a robust representation of the object view space, irrespective of the training sequence.

To examine whether sequential training produces more robust representations than random training when these factors are controlled, we conducted two additional experiments. As in the prior studies we measured performance before and after unsupervised sequential or random training, where the former contains spatiotemporal continuity and the latter only temporal proximity. However, here we measured performance on untrained views to assess the generalization of the learning effect. Study 3 used the same training with 24 views as Study 1, but tested on new views between the trained views. Study 4 trained with only seven views spaced 30° apart, and tested on new views between the trained views. We reasoned that if sequential training produces a more robust representation than random training, generalization performance will be higher for the former than the latter, in particular when the number of training views is reduced.

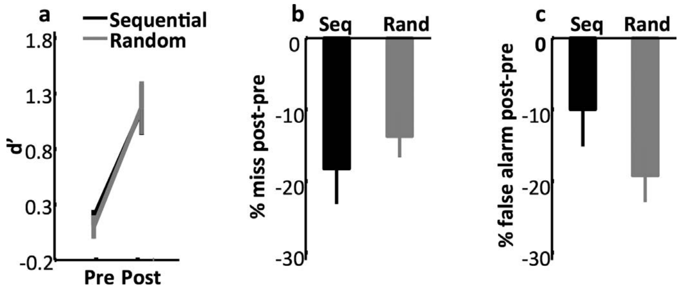

Figure 4. Study 3: Sequential and random learning with many views similarly improve generalization performance in the ability to discriminate between views of in-depth rotated objects. (a) *d*′ in discrimination performance before (pre-) and after (post-) training for the sequential learning (black) and random learning (gray). (b) Learning-related decrease in the percentage misses. (c) Learning-related decrease in the percentage false alarms. Results are averaged across 16 subjects. Error bars indicate *SEM. Seq:* sequential. *Rand:* random training.

Results of Study 3 indicated that when training with 24 views as in the prior experiments, learning generalizes to untrained views. Postlearning, participants significantly improved their accuracy on untrained views. Performance increased from 52.19% ± 0.98% pretraining (*d*′: 0.17 ± 0.05) to 67.50% ± 2.60% posttraining (*d*′: 1.12 ± 0.19). Main effect of training is significant (Figure 4a; $F(1, 15) = 32.16$, $p < 0.001$). The improvement was similar for sequential and random training (no main effect of training sequence, $F(1, 15) = 0.04$, $p = 0.83$, and no interaction between training and training sequence, $F(1, 15); = 0.10$, $p = 0.75$). Again, performance improved both due to a significant decrease in the miss rate (main effect of training, $F(1, 15) = 19.85$, $p < 0.001$) and a significant decrease in the false alarm rate (main effect of training, $F(1, 15) = 10.85$, $p < 0.001$). There was no effect of training sequence on either miss or false alarm rate (all ps > 0.5; Figure 4b, c). Overall, training with 24 views yielded similar generalization to nearby views for both sequential and random training.

Even though we trained with less than a third of the views in Study 4, participants' performance on untrained views significantly increased from 52.7% ± 1.2%, (*d*′: 0.17 ± 0.07) before training to 66.6% ± 1.6% (*d*′: 1.04 ± 0.12) after training. Main effect of training is significant (Figure 5a, $F(1, 25) = 44.60$, $p < 0.001$). Notably, results differed from the prior experiments in that the improvement was significantly greater after sequential than random training, as revealed by a significant two-way interaction between training and training sequence, $F(1, 25) = 5.21$, $p = 0.02$, and a main effect of training sequence, $F(1, 25) = 5.44$, $p = 0.02$.

This larger improvement for sequential rather than random learning was due to a significantly larger decrease in the false alarm rate in the former compared to the latter (Figure 5c; interaction between training and training sequence in false alarm rate, $F(1, 25) = 6.47$, $p = 0.01$). There was also a significant decrease in the miss rate after training (Figure 5b), but it did not differ across training conditions (two-way ANOVA with factors of training and training sequence, main effect of training, $F(1, 25) = 29.36$, $p < 0.001$; no interaction between training and training sequence, $F(1, 25) = 0.89$, $p = 0.35$). This indicates that the greater improvement in discrimination performance after sequential learning stems from fewer errors in which participants judged that views of different objects belonged to the same object. Finally, performance (*d*′, miss rate, and false alarm rate) on untrained views that were 7.5° from the trained views was similar to performance on views that were 15° away, suggesting that generalization in sequential learning extends to at least 15° away from trained views.

Overall, results of Studies 3 and 4 show that when fewer views were shown during training, spatiotemporal continuity enables greater generalization to new untrained views of objects than when training with the same views in temporal, but not spatial, proximity.

Study 5: Differences across sequential and random training are not due to differential engagement during unsupervised learning

Might the difference between sequential and random learning in Study 4 be due to differential allocation of attention or engagement with the stimuli during sequential versus random presentations? For example, one may hypothesize that the random sequences may disorient participants, leading them to look at object

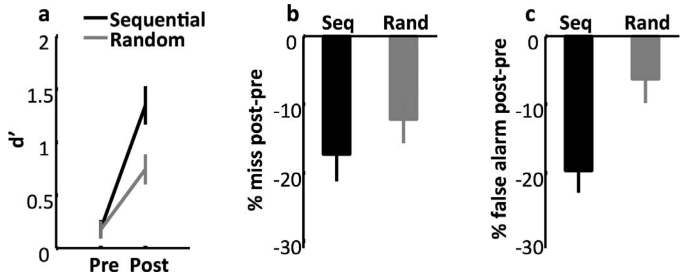

Figure 5. Study 4: When training with a limited number of views, sequential training enables better discrimination among objects shown in untrained views than random training. (a) *d*′ in discrimination performance pre- and posttraining for sequential (black) and random (gray) learning. (b) Learning-related decrease in percentage misses. (c) Learning-related decrease in the percentage false alarms. Results are averaged across 26 subjects. Error bars indicate *SEM. Seq:* sequential. *Rand:* random training.

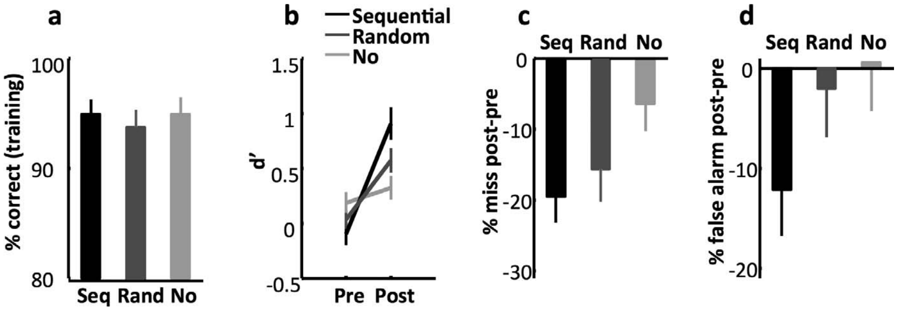

Figure 6. Study 5: When task is controlled during training, sequential training enables better discrimination of untrained views than of random training and no training. (a) Percentage correct performance during the training phase: contrast task (sequential and random training) and dot task (no training). (b) d' in discrimination performance pre- and posttraining for sequential, random, and no training. (c) Learning-related decrease in percentage misses. (d) Learning-related decrease in the percentage false alarms. Results are averaged across 27 subjects. Error bars indicate *SEM*. *Seq:* sequential. *Rand:* random training. *No*: no training.

images less than during sequential presentations. Furthermore, given that learning was lower during random training, perhaps this improvement was not due to training at all, but due to participants' familiarization with the discrimination task. Namely, perhaps participants always perform better the second time they do the discrimination task, irrespective of whether or not they actually watched the stimuli during training.

To address these questions, in Study 5, we examined whether unsupervised training produces a larger improvement in performance than a baseline condition in which the discrimination test was done twice, but with no object training between tests. Here subjects viewed a dot instead of the object during the training phase and reported when its color changed. We also examined whether the difference between sequential and random learning would be maintained when participants' engagement level is controlled during random and sequential training. Here during training subjects were instructed to detect a decrement in the contrast of the object (see Methods).

Measurement of performance during training showed that it was similar across the sequential and random training, and was also not different than performance on the baseline task (Figure 6a; $F(2, 52) = 0.75$, $p > 0.05$). Participants' performance on untrained views significantly increased from 51.1% ± 0.8%, (d': 0.04 ± 0.05) before training to 60.3% ± 1.1% (d': 0.60 ± 0.07) after training. A two-way repeated measures ANOVA on d' with factors of training (pre/posttraining blocks) × training type (sequential/random/none) showed a main effect of learning, $F(1, 26) = 40.04$, $p < 0.001$; a main effect of learning type, $F(2, 52) = 5.18$, $p < 0.01$; and an interaction between learning and learning type, $F(2, 52) = 5.36$, $p < 0.01$ (Figure 6b). Critically, we replicated the results of Study 4. We found that the improvement in discrimination performance was significantly higher after sequential training compared to random training, $F(1, 26) = 4.15$, $p < 0.05$. Improvement was also significantly higher after sequential training compared to the baseline condition, $F(1, 26) = 14.61$, $p < 0.05$, and higher after random training compared to the baseline condition, $F(1, 26) = 4.16$, $p < 0.05$. However, performance differences in the baseline condition post- versus prelearning were not significant, $t(26) < 0$, $p > 0.05$.

Overall, results of Study 5 replicated the results of Study 4 showing that spatiotemporal information provides a greater benefit than temporal information in generalization to new views. Results of Study 5 also rule out the possibility that this benefit is contributed solely by unequal level of engagement during learning in the two paradigms. Finally, we show that learning is necessary to generate an improvement in discrimination performance, since performing the test a second time does not substantially improve performance.

Discussion

Our data show that neither spatiotemporal continuity nor motion information is necessary for unsupervised learning of viewpoint-invariant object representations. Specifically, when many views were provided in temporal proximity during learning, sequential training yielded similar improvement in discrimination performance across views as did training with randomly presented views. However, when fewer views were provided during training and object recognition performance was tested on new views, spatiotemporally continuous training enhanced generalization compared to training with temporal proximity alone. Importantly, results of Study 5 show that the improvement during spatiotemporal compared to temporal proximity training is not due to differential attention or engagement during sequential compared to random training.

These results have two important implications. First, if enough views of an object are shown during training, temporal proximity is sufficient for unsupervised learning of object representations. Second, results of Studies 4 and 5 reveal that a specific benefit of spatiotemporal continuity during learning is enhanced generalization to new object views compared to learning from temporal proximity alone. One prediction from our findings is that that spatiotemporal continuity may aid unsupervised learning compared to temporal proximity in other situations where we have impoverished information on the 3-D structure of the object during learning, such as low contrast, noise, or difficult lighting.

What are the implications of our findings for the types of object representations that are built during learning?

Our results are consistent with theories suggesting a view-based representation of objects (Bülthoff & Edelman, 1992; Bülthoff et al., 1995; Logothetis & Pauls, 1995) as the number of views shown during training affected performance. In contrast, 3-D geon-based representations (Biederman, 1987) would have predicted no gain in sequential versus random presentations and no gain from having more object views during training given that the relevant features were visible in all training conditions.

What happens during learning novel 3-D objects?

We hypothesize that unsupervised learning generated view-tuned units centered on the trained views. The representation of the object view space after training likely depends on the number of learned views, the width of the view tuning around the learned views, and their overlap. We suggest that when many views are learned and their view tunings overlap, the entire view space of the object will be represented after training. In this case temporal proximity may suffice for linking among views, and spatiotemporal continuity will not further improve generalization. However, when training involves fewer and more distant views, the object view space obtained by training might be incomplete as there may be intermediate views between trained views that are not represented. Given the same number of training views during spatiotemporal and temporal learning, our results suggest that spatiotemporal continuity might provide broader view tuning compared to temporal proximity. In turn, this enables better generalization to new views. These predictions can be tested in future computational, neural network simulations (Perry et al., 2006; Wallis & Rolls, 1997; Yamins et al., 2014) or neuroscience experiments (Andresen, Vinberg, & Grill-Spector, 2009; Grill-Spector et al., 1999; Kietzmann, Swisher, König, & Tong, 2012; Logothetis & Pauls, 1995) that will examine the view tuning of object representations after spatiotemporal continuous and temporal proximal learning.

While our findings showed the specific benefit of spatiotemporal learning over learning with temporal proximity alone, our results also generate new questions that can be examined in future research: (a) What is the range of object views in which sequential training provides an advantage compared to random training? Does this number depend on the 3-D structure of the object? (b) What is the minimum number of views for which training with temporal proximity alone can improve object recognition performance? (c) How close in time do views of an object need to be shown for effective temporal proximity learning to occur? Will our findings extend to training in timescales of minutes? Hours? Days?

Temporal proximity during unsupervised learning is sufficient to produce view-invariant object discrimination

Our finding that temporal proximity during unsupervised learning is sufficient to produce view-invariant object discrimination is consistent with and extends Liu's (2007) finding of similar recognition memory performance after unsupervised learning with random or sequential presentation of object views. However, our findings differ from Harman and Humphrey (1999) who reported faster (but not better) recognition memory after random training compared to sequential training. Differences among studies may be due to differences in the level of recognition tested: Our experiments tested subordinate-level recognition among objects that shared the same parts and configuration, but Harman and Humphrey's tested basic-level recognition among objects that varied both in their parts and their configuration. Moreover, Liu's study showed that Harman and Humphrey's results were in part due to lower attention and lower effort due to repetitiveness of the stimuli during sequential compared to the random training, which is not a concern in the present study.

Spatiotemporal continuity versus motion information in learning view-invariant object discrimination

Results of Study 2 revealed similar improvement after unsupervised learning when motion information was present as when it was absent. This aspect of our results is different from findings showing that motion information can contribute to learning object representations (Balas & Sinha, 2008; Vuong & Tarr, 2004, 2006) and that motion can facilitate short-term representation of object views (Kourtzi & Shiffrar, 1997, 1999). Why might this be the case? We suggest that when motion information is a diagnostic feature of an object, it aids learning (Stone, 1998, 1999; Vuong & Tarr, 2004, 2006). In particular, motion information may be critical for learning the statistics of deformable biological entities such as hands (Ullman et al., 2012), bodies (O'Toole et al., 2011), faces (Knappmeyer et al., 2003; Lander & Davies, 2007; Pilz, Thornton, & Bülthoff, 2006), and deformable novel objects (Balas & Sinha, 2008). However, in our experiments participants learned novel rigid 3-D objects for which motion was not a key feature, and therefore it did not appear to be critical for learning.

Another question is whether motion may play a role in learning in scenarios with impoverished information. While motion alone is insufficient to explain the differences in performance after sequential and random learning in Studies 4 and 5, as both sequences contain motion information, it is possible that motion, like spatiotemporal information, may enhance learning when training with a limited number of views. This prediction can be tested in future research.

Conclusions

In sum, our data show that learning invariant object representations may occur from unsupervised observation of many examples of object views presented in temporal proximity. However, we show that there is a specific benefit to learning these views in spatiotemporal continuity: it generates better learning from fewer training examples and obtains better generalization performance to new views. These findings suggest that computational algorithms that learn from examples can obtain effective learning of 3-D objects from either learning from many randomly presented examples of an object, or from a small sample of examples if presented in sequential order. Overall, our results provide important insights for theories of viewpoint invariant object recognition and computational algorithms that learn objects from examples.

Keywords: perceptual learning, natural statistics, invariant recognition, object discrimination

Acknowledgments

KG-S and MT developed the study concept and contributed to the study design. MT collected and analyzed the data. MT and KG-S wrote the manuscript. All authors approved the final version of the manuscript for submission. We thank Nathan Witthoft for fruitful discussions. We thank Rubi Hammer for introducing the 3DS Max software. We thank Heather Bowman and Devangi Vivrekar for their help in the collection of data for Studies 4 and 5. We also thank the editor and the three anonymous reviewers for their valuable suggestions that improved the manuscript. This research was supported by NEI Grant 1 R01 EY019279-01A1 to KGS, Ric Weiland Graduate Fellowship to MT, and the Stanford Center for Mind Brain and Computation trainee support to MT.

Commercial relationships: none.

Corresponding author: Kalanit Grill-Spector.
Email: kalanit@stanford.edu.
Address: Department of Psychology, Jordan Hall, Stanford University, Stanford, CA, USA.

References

Amir, O., Biederman, I., & Hayworth, K. J. (2012). Sensitivity to nonaccidental properties across various shape dimensions. *Vision Research*, *62*, 35–43, doi:10.1016/j.visres.2012.03.020.

Amir, O., Biederman, I., Herald, S. B., Shah, M. P., & Mintz, T. H. (2014). Greater sensitivity to nonaccidental than metric shape properties in preschool children. *Vision Research*, *97*, 83–88, doi:10.1016/j.visres.2014.02.006.

Andresen, D. R., Vinberg, J., & Grill-Spector, K. (2009). The representation of object viewpoint in human visual cortex. *NeuroImage*, *45*(2), 522–536, doi:10.1016/j.neuroimage.2008.11.009.

Balas, B., & Sinha, P. (2008). Observing object motion induces increased generalization and sensitivity. *Perception*, *37*(8), 1160–1174, doi:10.1068/p6000.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–147, doi:10.1037/0033-295X.94.2.115.

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, USA*, *89*(1), 60–64.

Bülthoff, H. H., Edelman, S. Y., & Tarr, M. J. (1995). How are three-dimensional objects represented in the brain? *Cerebral Cortex*, *5*(3), 247–260, doi:10.1093/cercor/5.3.247.

Chubb, C., & Sperling, G. (1989). Two motion perception mechanisms revealed through distance-driven reversal of apparent motion. *Proceedings of* *the National Academy of Sciences, USA*, *86*(8), 2985–2989.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Science*, *11*(8), 333–341, doi:10.1016/j.tics.2007.06.010.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences, USA*, *99*(24), 15822–15826, doi:10.1073/pnas.232472899.

Green, D. M., & Swets, J. A. (1974). *Signal detection theory and psychophysics*. Huntington, NY: R. E. Krieger.

Grill-Spector, K., Kushnir, T., Edelman, S., Avidan, G., Itzchak, Y., & Malach, R. (1999). Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron*, *24*(1), 187–203, doi:10.1016/S0896-6273(00)80832-6.

Harman, K. L., & Humphrey, G. K. (1999). Encoding 'regular' and 'random' sequences of views of novel three-dimensional objects. *Perception*, *28*(5), 601–615, doi:10.1068/p2924.

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition*, *78*(3), B53–B64, doi:10.1016/S0010-0277(00)00132-3.

Hayward, W. G., & Tarr, M. J. (1997). Testing conditions for viewpoint invariance in object recognition. *Journal of Experimental Psychology: Human Perception & Performance*, *23*(5), 1511–1521, doi:10.1037/0096-1523.23.5.1511.

Isik, L., Leibo, J. Z., & Poggio, T. (2012). Learning and disrupting invariance in visual recognition with a temporal association rule. *Frontiers in Computational Neuroscience*, *6*, 37, doi:10.3389/fncom.2012.00037.

Kietzmann, T. C., Swisher, J. D., König, P., & Tong, F. (2012). Prevalence of selectivity for mirror-symmetric views of faces in the ventral and dorsal visual pathways. *Journal of Neuroscience*, *32*(34), 11763–11772, doi:10.1523/JNEUROSCI.0126-12.2012.

Knappmeyer, B., Thornton, I. M., & Bülthoff, H. H. (2003). The use of facial motion and facial form during the processing of identity. *Vision Research*, *43*(18), 1921–1936, doi:10.1016/S0042-6989(03)00236-0.

Kourtzi, Z., & Shiffrar, M. (1997). One-shot view invariance in a moving world. *Psychological Science*, *8*(6), 461–466, doi:10.1111/j.1467-9280.1997.tb00462.x.

Kourtzi, Z., & Shiffrar, M. (1999). The visual representation of rotating, three-dimensional objects. *Acta Psychologica (Amst)*, *102*(2–3), 265–292, doi:10.1016/S0001-6918(98)00056-0.

Lander, K., & Davies, R. (2007). Exploring the role of characteristic motion when learning new faces. *Quarterly Journal of Experimental Psychology*, *60*(4), 519–526, doi:10.1080/17470210601117559.

Liu, T. (2007). Learning sequence of views of three-dimensional objects: the effect of temporal coherence on object memory. *Perception*, *36*(9), 1320–1333, doi:10.1068/p5778.

Logothetis, N. K., & Pauls, J. (1995). Psychophysical and physiological evidence for viewer-centered object representations in the primate. *Cerebral Cortex*, *5*(3), 270–288, doi:10.1093/cercor/5.3.270.

Logothetis, N. K., Pauls, J., Bülthoff, H. H., & Poggio, T. (1994). View-dependent object recognition by monkeys. *Current Biology*, *4*(5), 401–414, doi:10.1016/S0960-9822(00)00089-0.

Lu, Z. L., & Sperling, G. (1995). Attention-generated apparent motion. *Nature*, *377*(6546), 237–239, doi:10.1038/377237a0.

Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature*, *335*(6193), 817–820, doi:10.1038/335817a0.

O'Toole, A. J., Phillips, P. J., Weimer, S., Roark, D. A., Ayyad, J., Barwick, R., & Dunlop, J. (2011). Recognizing people from dynamic and static faces and bodies: Dissecting identity with a fusion approach. *Vision Research*, *51*(1), 74–83, doi:10.1016/j.visres.2010.09.035.

Perry, G., Rolls, E. T., & Stringer, S. M. (2006). Spatial vs temporal continuity in view invariant visual object recognition learning. *Vision Research*, *46*(23), 3994–4006, doi:10.1016/j.visres.2006.07.025.

Pilz, K. S., Thornton, I. M., & Bülthoff, H. H. (2006). A search advantage for faces learned in motion. *Experimental Brain Research*, *171*(4), 436–447, doi:10.1007/s00221-005-0283-8.

Poggio, T., & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, *343*(6255), 263–266, doi:10.1038/343263a0.

Preminger, S., Blumenfeld, B., Sagi, D., & Tsodyks, M. (2009). Mapping dynamic memories of gradually changing objects. *Proceedings of the National Academy of Sciences, USA*, *106*(13), 5371–5376, doi:10.1073/pnas.0802111106.

Roark, D. A., Barrett, S. E., Spence, M. J., Abdi, H., & O'Toole, A. J. (2003). Psychological and neural perspectives on the role of motion in face recognition. *Behavioral & Cognitve Neuroscience Reviews*, *2*(1), 15–46, doi:10.1177/1534582303002001002.

Sakai, K., & Miyashita, Y. (1991). Neural organization for the long-term memory of paired associates. *Nature*, *354*(6349), 152–155, doi:10.1038/354152a0.

Simons, D. J., & Rensink, R. A. (2005). Change blindness: Past, present, and future. *Trends in Cognitive Sciences*, *9*(1), 16–20, doi:10.1016/j.tics.2004.11.006.

Sinha, P., & Poggio, T. (1996). Role of learning in three-dimensional form perception. *Nature*, *384*(6608), 460–463, doi:10.1038/384460a0.

Stone, J. V. (1998). Object recognition using spatiotemporal signatures. *Vision Research*, *38*(7), 947–951, doi:10.1016/S0042-6989(97)00301-5.

Stone, J. V. (1999). Object: View-specificity and motion-specificity. *Vision Research*, *39*(7), 4032–4044, doi:10.1016/S0042-6989(99)00123-6.

Tarr, M. J., Williams, P., Hayward, W. G., & Gauthier, I. (1998). Three-dimensional object recognition is viewpoint dependent. *Nature Neuroscience, 1*(4), 275–277. doi:10.1038/1089

Ullman, S., Harari, D., & Dorfman, N. (2012). From simple innate biases to complex visual concepts. *Proceedings of the National Academy of Sciences, USA*, *109*(44), 18215–18220, doi:10.1073/pnas.1207690109.

Vuong, Q. C., & Tarr, M. J. (2004). Rotation direction affects object recognition. *Vision Research*, *44*(14), 1717–1730, doi:10.1016/j.visres.2004.02.002.

Vuong, Q. C., & Tarr, M. J. (2006). Structural similarity and spatiotemporal noise effects on learning dynamic novel objects. *Perception*, *35*(4), 497–510, doi:10.1068/p5491.

Wallis, G., Backus, B. T., Langer, M., Huebner, G., & Bülthoff, H. (2009). Learning illumination- and orientation-invariant representations of objects through temporal association. *Journal of Vision*, *9*(7):6, 1–8, http://www.journalofvision.org/content/9/7/6, doi:10.1167/9.7.6. [PubMed] [Article]

Wallis, G., & Bülthoff, H. (1999). Learning to recognize objects. *Trends in Cognitive Sciences*, *3*(1), 22–31, doi:10.1016/S1364-6613(98)01261-3.

Wallis, G., & Bülthoff, H. H. (2001). Effects of temporal association on recognition memory. *Proceedings of the National Academy of Sciences, USA*, *98*(8), 4800–4804, doi:10.1073/pnas.071028598.

Wallis, G., & Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology*, *51*(2), 167–194, doi:10.1016/S0301-0082(96)00054-8.

Wang, G., Obama, S., Yamashita, W., Sugihara, T., & Tanaka, K. (2005). Prior experience of rotation is not required for recognizing objects seen from different angles. *Nature Neuroscience*, *8*(12), 1768–1775, doi:10.1038/nn1600.

Yamins, D., Hong, H., Cadieu, C. F., Solomon, E., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences, USA*, *111*(23), 8619–8624, doi:10.1073/pnas.1403112111.