Translation Tolerance in Vision

Anonymous CogSci submission

Abstract

A fundamental challenge in object recognition is to recognize an image when it is projected across different retinal locations, an ability known as translation tolerance. Although the human visual system can overcome this challenge, the mechanisms responsible remain largely unexplained. The 'trained-tolerance' approach holds that an object must be experienced across different retinal locations to achieve translation tolerance. Previous studies have supported this approach by showing that the visual system struggles to generalize recognition of novel objects to translations as small as 2° of visual angle. The present paper outlines a series of eyetracking studies that show novel objects can be recognized at translations as far as 18° from the trained retinal location, challenging the standard account of translation tolerance in neuroscience and psychology.

Keywords: Translation Tolerance; Translation Invariance; Object Recognition; Vision

Introduction.

We can identify familiar objects despite the variable images they project on our retina, including variation in image size, orientation, illumination, and position on retina. How the visual system succeeds under these conditions is still poorly understood. Here we focus on our ability to identify objects despite variations in retinal location and consider the extent to which the visual system relies on "on-line" vs. "trained" translation tolerance. In the case of on-line tolerance, learning to identify an object at one location immediately affords the capacity to identify that object at multiple retinal locations. At one extreme, the visual system can immediately generalize to all locations (to the limit of visual acuity), what might be called on-line translation invariance; at the other extreme, generalization is limited to a few degrees of visual angle. Trained tolerance, by contrast, refers to the hypothesis that we learn to identify familiar objects across locations by explicitly training the visual system to identify each object across a broad range of retinal locations. On this view, one of the functions of eye-movements is to ensure that objects are projected to multiple locations. These two accounts trade-off on one another: the more restricted on-line translation tolerance is the more trained tolerance is required to support the ability to identify objects across a wide range of retinal locations.

As detailed below, most of the empirical research in psychology and neuroscience suggests that on-line tolerance is restricted to a few degrees of visual angle, and to date, all neural network models of object identification show the same restriction. As a consequence, most theories of vision assume that trained tolerance plays an important role in our ability to identify objects across a range of retinal locations.

Early long-term priming studies by Biederman and colleagues (Biederman & Cooper, 1991; Cooper, Biederman, & Hummel, 1992; Fiser & Biederman, 2001) provided evidence for extensive on-line translation tolerance, and indeed, in some cases, translation invariance. For example, Fiser and Biederman (2001) asked participants to name line-drawings of objects as fast and accurately as possible in a study phase. In a later test block, participants were faster and more accurate to name repeated images compared to a set of control objects (same name, different exemplar) regardless of whether retinal position was the same at study and test or displaced by 10 degrees (°) of viewing angle. A limitation of all these studies, however, is that they assessed priming for familiar objects, and accordingly, participants had seen the same type of objects in a wide variety of retinal locations prior to the experiment. This leaves open the possibility that the findings reflected trained rather than on-line translation tolerance within the visual system, or indeed, trained tolerance outside the visual system with priming effects occurring within semantic or verbal systems (Kravitz, Vinson, & Baker, 2008).

In contrast with the Biederman studies, a number of authors have failed to observe robust translation tolerance for novel objects that participants had not seen prior to the experiment (e.g., Afraz & Cavanagh, 2008; Cox & DiCarlo, 2008; Dill & Fahle, 1997; Dill & Fahle, 1998; Newell, Sheppard, Edelman, & Shapiro, 2005). *Figure 1* outlines a selection of studies that used different experimental paradigms and found highly limited (in one case no) translation tolerance (adapted from Kravitz, Vinson, & Baker, 2008). Based on the outcome of such studies, Chen et al. (2017) state that "the translation-invariance of the human visual system is limited to shifts on the order of a few degrees - almost certainly less than 8°" (p.5). In line with this, Kravitz et al.'s (2008) review of behavioral studies found that "most of the training and matching studies found a significant decrement in discrimination performance with translations varying from 0.5° to 2°" (p. 118).

Neural data are also consistent with the idea that on-line translation tolerance is highly limited. Researchers have identified neurons in inferior-temporal cortex (IT) with a range of receptive fields (ranging from 2.8° to 26°; for review see Kravitz et al., 2008). The larger receptive fields are thought to provide the neural underpinning of translation tolerance, but it is important to note that these receptive fields have only been observed for familiar or newly-trained stimuli that have been seen at multiple

retinal locations (e.g., Gross et al., 1972; Ito, Tamura, Fujita, & Tanaka, 1995; Tovee, Rolls, & Azzopardi, 1994). Accordingly, these large receptive fields could reflect trained or on-line translation tolerance. Consistent with the former hypothesis, Cox and DiCarlo (2008) only observed small receptive fields for their novel stimuli that were trained in one location, as shown in *Figure 1C*. That is the neural data appear to mirror the behavioral data: robust translation tolerance and large receptive fields are found for familiar stimuli, limited generalization and small receptive fields are observed for unfamiliar stimuli. Indeed, Cox and DiCarlo reach a similar conclusion to Kravitz et al. (2008), writing "...the computational machinery of the ventral visual stream is not constructed in a manner that automatically produces position tolerance in IT, even across relatively small changes in retinal position. Instead, the creation and/or maintenance of IT position tolerance might require experience".

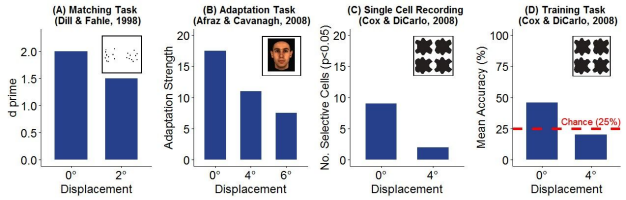

Figure 1. Behavioral studies of translation tolerance.

Similarly, work with artificial neural network models has reported robust translation tolerance for words and objects at trained locations, but highly limited generalization to untrained locations. For example, Dandarund et al. (2013) and DiBono & Zorzi (2013) showed that that their models of visual word identification supported translation invariance, but the models were trained with each word at each location. Elliffe, Rolls and Stringer (2002) showed that a biologically inspired neural network model called VisNet supported on-line translation tolerance to untrained locations for simple stimuli, but each stimuli had to be trained at multiple spatial locations (after training in 7 locations the model generalized to an 8th and 9th location), and the authors tested small translations (translations of 8 pixels in a 128x128 retina). The above behavioral, neural, and computational above findings have led most theorists in psychology and neuroscience to endorse the 'trained' account of translation tolerance.

Despite empirical and computational results, there are still reasons to question the trained tolerance hypothesis. Behavioural studies that failed to observe on-line translation (e.g., *Figure 1*) suffer from a number of limitations. For example, stimuli are typically unlike real objects (e.g., Dill & Fahle, 1997; see *Figure 1a*), and/or are very similar to each other (e.g., Cox & DiCarlo, 2008; see *Figure 1c*). Differentiating between highly similar items may rely on low-level visual representations that are retinotopically constrained (Kravitz et al., 2008). Additionally, stimuli in these experiments were typically trained at a given location for just 100ms (contrary to everyday visual experiences in which stimuli can be encoded for longer intervals). More extended studying time may be required for robust online translation tolerance. Consistent with the first possibility, Dill and Edelman (2001) observed much more extensive translation tolerance for unfamiliar objects that were more object-like and discriminable from one another. Indeed, they reported no significant reduction in performance in five of six experiments when images were displaced by 8 degrees. And consistent with both possibilities, Bowers, Vankov and Ludwig (2016) reported more robust translation tolerance still when participants trained to identify more discriminable stimuli novel stimuli that were studied at one retinal location for longer periods of time. Indeed, in their Experiment 3, participants were ~80 % accurate in naming novel objects following a shift of 13° (when chance was 16.7 %).

In this article we explore on-line translation tolerance in humans given the conflicting empirical evidence regarding on-line translation tolerance and the theoretical implications for theories of vision in psychology and neuroscience. Two gaze-contingent eye-tracking studies are reported and include the following critical design features. First, in all studies, 24 novel objects were used, each of which was a member of a pair of objects built from similar parts but in a different global configuration (see Method). Using a large set of stimuli of this sort should encourage participants to learn the complete objects rather than just the parts. Previous studies have rarely matched items on the basis of their parts (but see Dill & Edelman, 2001), and have typically included far fewer stimuli. Second, the novel three-dimensional objects we included were designed to be more naturalistic compared to the novel stimuli used in previous experiments, such as those depicted in *Figure 1a* and *1c*. This makes it more likely that the visual system will process these new stimuli in a manner more similar to everyday recognition. Third, we included study conditions in which stimuli were presented for unlimited time at study as opposed to the brief display conditions in previous studies that may have artificially reduced on-line translation tolerance. Fourth, we included test conditions in which objects were presented for 100ms durations, reducing the likelihood that participants adopted artificial strategies at test. Note that the Bowers et al. (2016) experiments reporting robust on-line translation invariance included a smaller number of less realistic objects that were displayed for an extended time at test. Accordingly, the current studies provide a much stronger test of the on-line translation tolerance hypothesis.

Experiment 1.

Method

Participants and Equipment. Fifteen participants (Experiment 1a=6, 1b=9) were recruited from the University of Bristol's course credit scheme for Psychology students. Eye-movements were monitored using the Eyelink 1000 plus system (SR Research). Stimuli were presented using Psychopy v1.85.3 (platform: Linux-Ubuntu), and on a 120Hz monitor with a spatial resolution of 1920 x 1080 pixels (screen width = 53cm), at a distance of 70cm.

Stimuli. Twenty-four novel objects were taken from Leek, Roberts, Oliver, Cristino, and Pegna (2016). Each object was part of a pair that had similar local features but a different global configuration (see *Figure 2*). For each participant, one member of each pair was randomly assigned the label 'A' and the other was assigned 'B'.

Figure 2. Twenty-four novel objects. Each column contains a pair of objects that are matched for similar local features, but which differ in global configuration.

Procedure. In the learning phase of the experiment participants were trained to categorize the 24 objects as 'A' or 'B'. Each object was presented one-by-one in the centre of the screen and occupied 5°x5° of visual angle. Participants were required to maintain their gaze on a centrally located fixation-cross for 1000ms for an object to appear. If gaze moved 1.5° beyond the fixation-cross, a mask replaced the object. The learning task was split into two phases: (i) *Familiarization*. The familiarization phase consisted of two presentations of each object. Each object was presented for an unlimited time and was accompanied by a sound-file indicating its category (A or B). (ii) *Training*. Each object was displayed at the same location without the sound file and for unlimited time until the participant pressed a button to indicate the image's category. Audio feedback was then provided. The training phase continued until the participant correctly identified each object consecutively (i.e., 24/24 consecutive correct answers). After completing the first training phase (Block 1), participants completed two additional training phases - Block 2 and Block 3 - which were identical except for their shorter presentation times of 500ms and 100ms respectively.

After the learning phase, participants completed seven test-blocks, each consisting of 24 presentations (one of each object); each test-block differed in terms of horizontal eccentricity from the centre of the object to the central fixation cross (i.e., displacement from training location). Test blocks 1, 2, 3, 4, 5, 6 and 7 presented images at 0° (i.e, trained-position), 3°, 3°, 6°, 6°, 9°, and 9° displacement from the centrally trained position, respectively. Test-blocks with the same displacement (e.g., block 2 and 3 were both 3°) differed in terms of presentation side (left or right). Within each test-block, order of presentation was randomised and no feedback was provided. In Experiment 1a images remained on the screen until participants responded. Experiment 1b was the same as Experiment 1a except that images were presented for 100ms at test in order to reduce possible response strategies. These final presentation durations are similar to previous studies that have failed to find online translation tolerance (see *Figure 1*).

Results.

Table 1. Mean (+/- 95% CI range) Accuracy in Experiment 1. *Columns show displacement of the test presentation from the trained location.*

	Mean (+/- 95% CI range) Accuracy			
	0°	**3°**	**6°**	**9°**
Exp 1a (N = 6)	98% (5%)	98% (3%)	95% (9%)	94% (9%)
Exp 1b (N = 9)	98% (2%)	95% (3%)	91% (4%)	84% (7%)

As shown in *Table 1*, novel objects were recognised with high accuracy at untrained retinal-positions (chance is 50%). Even at the most distal untrained position (9°), objects were recognised with a mean accuracy of 94% when unlimited time was afforded at test (Experiment 1a), and although translation tolerance was reduced when stimuli were presented for 100ms at test (Experiments 1b), accuracy was still 84% when at 9° displacement.

Experiment 2.

Experiment 2 served two purposes. (i) Although we observe near complete translation invariance for newly acquired objects displaced by 9° when objects were presented for an unlimited amount of time, there was a significant reduction when stimuli were presented for 100 ms at test. In an attempt to reduce any effects of retinal specificity, Experiment 2a adopted a learning condition known as 'location-training' (i.e., training at the test location - see below). (ii) Experiment 2b also used location training to examine whether the robust on-line translation reported in our experiments (1a to 2a) could be extended to an even more distal untrained location, 18° from the trained location.

Location training has been used by previous studies to show that participants can overcome retinal-specificity for low-level visual discrimination tasks. Xiao et al. (2008) demonstrated that participants who had been trained to discriminate contrasts at location 1 showed complete transfer of this ability to location 2 only when they had also been trained to discriminate different stimuli on a different dimension (orientation) at location 2 (otherwise, learned contrast discrimination was location specific). Xiao et al. concluded that training at location 2 trained participants to overcome stimulus-nonspecific factors like local noise at the stimulus location, and this enabled complete location transfer. Our Experiment 2 investigated whether location training may also reduce retinal specificity in high-level visual recognition tasks (Experiments 2a and 2b investigated this at displacements of 9° and 18°, respectively).

Participants, Equipment and Stimuli. Experiment 2 used identical equipment, stimuli and recruitment methods as Experiment 1. Ten participants were used in Experiment 2a, and 10 different participants in Experiment 2b.

Experiment 2a: Learning and Test Phase. During learning blocks 1 and 2, objects were trained for unlimited time and 100ms duration respectively, at the centre of the screen (at fixation). In block 3, 'location training' took place: twelve objects were trained at one peripheral location, 9° from the central fixation-cross and the remaining 12 objects were trained at a contralateral peripheral location, 9° to the other side of the fixation-cross (all presentations were 100ms). Participants were trained until they got 12/12 consecutive correct answers at each peripheral location. In the test phase, objects were tested at three test locations: 9° left, 9° right, and centre of fixation, giving three test conditions: "trained-central" (0° displacement from central training location), "trained-peripheral" (0° displacement from peripheral training location) and "novel-peripheral" locations (9° displacement from central training location, on the opposite side to the trained peripheral location). To control for possible order effects, the three test locations were randomly interleaved within each test-block.

Experiment 2b: Learning and Test Phase. Experiment 2b examined whether the robust on-line translation reported in Experiment 2a could be extended to an even more distal untrained location, 18° from the trained location. In order to displace presentations by 18° at test, images were presented at one peripheral location only (and never at central fixation): 12 images were presented 9° to the right, and the remaining 12 were presented 9° to the left of central fixation (images were presented for unlimited time in block 1, and for 100ms duration in blocks 2 and 3). In an attempt to boost performance compared to Experiment 2a, participants were required to get 24/24 consecutive correct answers in each block. At test, objects were tested at two test locations: 9° left, and 9° right of fixation, giving two test conditions: "trained-peripheral" (0° displacement from peripheral trained location) and "novel-peripheral" locations (9° displacement from central fixation, and thus 18° displacement from the opposite peripheral location).

Results.

Table 2. **Mean (+/- 95% CI range) Accuracy scores in Experiment 2a and 2b.** Columns show degrees by which the test presentation was displaced from the nearest training location and the screen position of that test presentation.

	Mean (+/- 95% CI range) Accuracy			
Displacement	**0°**	**0°**	**9°**	**18°**
Screen Position	**Centre (trained)**	**Peripheral (trained)**	**Peripheral (novel)**	**Peripheral (novel)**
Exp 2a (N=10)	93% (5%)	83% (5%)	81% (6%)	not tested
Exp 2b (N=10)	not tested	97% (3%)	not tested	89% (7%)

The results of Experiment 2a and 2b are summarised in *Table 2*. In Experiment 2a, mean accuracy scores at the *novel-peripheral* position (9°) were significantly above chance and were nearly equivalent to those yielded in the *trained-peripheral* position (0° Peripheral). Thus, Experiment 2a provided strong evidence for robust online translation tolerance over 9° displacement even when objects are presented for just 100ms at test.

In Experiment 2b, objects were recognised with a very high degree of accuracy even when the trained location was displaced by 18°. Moreover, 5 of the 10 participants scored above 90% on mean accuracy at 18°. As illustrated in *Figure 3*, the findings from Experiment 2b, show on-line translation tolerance for novel stimuli over much more distal displacements compared to all previous work.

Discussion

The present paper has provided evidence of robust on-line translation tolerance in the human visual system. Participants trained to recognise novel objects at one retinal position could recognise the same objects at untrained distal retinal-locations (up to 18°) with high accuracy.

The findings are contrary to previous studies that demonstrate much more limited generalization over translations as small as 2 and 4° (e.g., Cox & DiCarlo, 2008; Dill & Fahle, 1998) and that have been used to support trained theories of translation tolerance. Indeed,

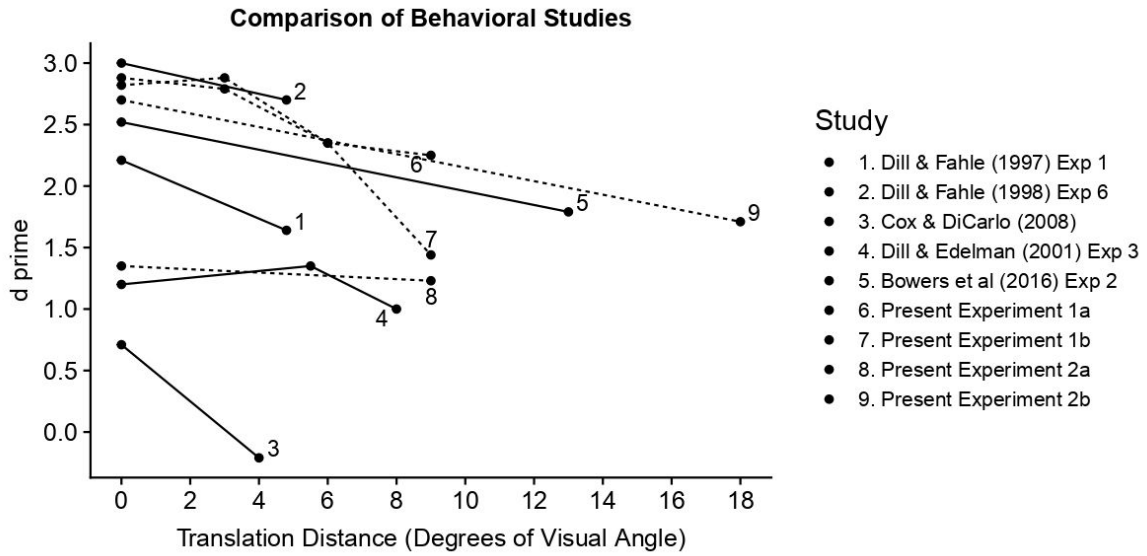

Figure 3. Comparison of d-prime scores for previous and present experiments as a function of translation distance. For each study, the experiment with the best performance at the largest displacement is illustrated. The present experiments (dashed lines) show more robust on-line translation tolerance than the majority of previous experiments. Experiment 2b showed robust on-line translation tolerance over a larger translation distance (18°) than any previous experiment. d-prime scores were calculated using the psyphy package (Knoblauch, 2014) for R (R Development Core Team, 2018).

advocates of the approach have recently claimed "the translation-invariance of the human visual system is limited to shifts on the order of a few degrees - almost certainly less than 8°" (Chen et al., 2017; p. 5). Rather, the present findings indicate that novel object recognition can be generalized on-line to more distal untrained retinal positions than previously demonstrated (see *Figure 3*): objects were recognised over translations as large as 18° with performance near 90%.

Why was robust on-line translation tolerance demonstrated in the present experiments whereas most previous experiments demonstrated highly limited generalization? As described above, previous studies typically used stimuli that are unlike real objects and/or are very similar to each other. Differentiating such stimuli may rely on low-level visual processes that are highly retinotopically constrained. High-level visual processes may also be more retinotopically constrained under these conditions. Indeed, there is some physiological evidence that receptive field (RF) sizes of neurons in the infero-temporal cortex (IT) are a function of stimuli size (DiCarlo & Maunsell, 2003). The present study used more naturalistic stimuli and included a number of variations in the procedure used by most psychophysical studies, including extended sampling times and, in Experiment 2, 'location training'. The more naturalistic conditions may have encouraged recruitment of IT neurons with larger RFs. Other studies that have also observed robust translation tolerance have also used more naturalistic, easily discriminable stimuli (Dill & Edelman, 2001) and extended sampling times (Bowers et al., 2016), but our findings go beyond this work by showing that robust translation tolerance extends to 18° under conditions in which strategic processing is minimized (by flashing items at test for 100 ms and by including a large set of novel objects that differed in the configuration of similar parts).

The findings are also relevant to computational modelling of the visual system. As noted in the Introduction, previous attempts to achieve on-line translation tolerance with artificial neural networks have reported highly limited tolerance. Such demonstrations may have been considered a strength given similarly limited tolerance reported in humans (e.g., Dill & Fahle, 1997; 1998). The present results show the need for these models to capture the robust on-line translation tolerance we have reported in humans. There is reason to believe that at least one class of artificial neural network can achieve this. Deep convolutional neural networks (CNNs) are designed to support translation tolerance by including convolutional layers and global pooling layers. Convolutions involve copying learning that occurs at one retinal location to other locations (the premise that information learned at one location is relevant at others), whilst pooling layers aggregate information from multiple

spatially organized units to a single unit in order to down-size the image. Both of these inbuilt ("innate") mechanisms are widely claimed to support translation tolerance, but there is surprisingly little evidence as to whether these mechanisms can support robust on-line translation tolerance as we have observed. We are in the process of running simulations to assess whether CNNs can support our empirical results.

Overall, the evidence outlined above is a clear demonstration that the human visual system can support recognition of novel objects at untrained distal retinal positions. Since the standard approach within psychology and neuroscience is to deny such robust generalization, there is a need for the field to more widely acknowledge an on-line generalization mechanism that can account for these results.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme.

References

Afraz, S.R., & Cavanagh, P. (2008). Retinotopy of the face aftereffect. *Vision Research, 48,* 42–54.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94*, 115–147.

Biederman, I., & Cooper, E. E. (1991). Evidence for complete translational and reflectional. *Perception, 20*, 585–593.

Biederman, I., Cooper, E.E., Kourtzi, Z., Sinha, P., & Wagemans, J. (2009). Biederman and Cooper's 1991 Paper. *Perception*, *38*, 809–825.

Bowers, J. Vankov, I. & Ludwig, C. (2016). The visual system supports online translation invariance for object identification. *Psychonomic Bulletin Review, 23,* 432-438

Chen, F. X., Roig, G., Isik, L., Boix, X, & Poggio, T. (2017). "Eccentricity dependent deep neural networks: Modeling invariance in human vision," in AAAI Spring Symposium Series, Science of Intelligence, 2017.

Cooper E. E., Biederman I., & Hummel J. E. (1992). Metric invariance in object recognition: A review and further evidence. *Can. Psychol. 46*, 191-214.

Cox, D. D. & DiCarlo, J.J. (2008). Does Learned Shape Selectivity in Inferior Temporal Cortex Automatically Generalize Across Retinal Position? *Journal of Neuroscience*, *28*, 10045-10055.

Dandurand, F., Hannagan, T., & Grainger, J. (2013). Computational models of location-invariant orthographic processing. *Connection Science, 25*, 1–26.

Di Bono, M. G., & Zorzi, M. (2013). Deep generative learning of location-invariant visual word recognition. *Frontiers in Psychology, 4*, 635.

DiCarlo, J. J., & Maunsell, J. H. (2003). Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position. *Journal of Neurophysiology, 89,* 3264–3278.

Dill, M., & Fahle, M. (1997). The role of visual field position in pattern discrimination learning. *Proceedings of the Royal Society B, 264,* 1031–1036.

Dill, M. & Fahle, M. (1998) Limited translation invariance of human visual pattern recognition. *Perception & Psychophysics, 60*, 65–81

Dill, M., & Edelman, S. (2001). Imperfect Invariance to Object Translation in the Discrimination of Complex Shapes. *Perception, 30*, 707–724.

Elliffe M.C.M., Rolls E.T., & Stringer S.M. (2002). Invariant recognition of feature combinations in the visual system. *Biological Cybernetics, 86,* 59– 71.

Fiser, J., & Biederman, I. (2001). Invariance of long-term visual priming to scale, reflection, translation, and hemisphere. *Vision Research, 41*, 221–234.

Gross, C.G. et al. (1972). Visual properties of neurons in inferotemporal cortex of the macaque. *J. Neurophysiol., 35*, 96–111

Ito, M. et al. (1995) Size and position invariance of neuronal responses in monkey inferotemporal cortex. *J. Neurophysiol., 73*, 218–226

Kravitz, D. J., Vinson, L. D., & Baker, C. I. (2008). How position dependent is visual object recognition? *Trends in Cognitive Sciences, 12,* 114–122.

Leek, E. C., Roberts, M. V., Oliver, Z. J., Cristino, F., & Pegna, A. (2016). Early differential sensitivity of evoked-potentials to local and global shape during the perception of three-dimensional objects. *Neuropsychologia, 89,* 495–509.

Nazir T, & O'Regan J.K. (1990). Some results on translation invariance in the human visual system. *Spatial Vision, 5,* 81-100.

Newell, F. N., Sheppard, D. M., Edelman, S., & Shapiro, K. L. (2005). The interaction of shape- and location-based priming in object categorisation: Evidence for a hybrid "what + where" representation stage. *Vision Research, 45*, 2065–2080.

Tovee, M. J., Rolls, E. T., Azzopardi, P., (1994). Translation invariance in the responses to faces of single neurons in the temporal visual cortical areas of the alert monkey. *J. Neurophysiol., 72,* 1049-1060.

Ullman, S. (2007). Object recognition and segmentation by a fragment based hierarchy. *Trends Cogn Sci, 11,* 58–64.

Xiao, L.Q., Zhang, J.-Y., Wang, R., Klein, S. A., Levi, D. M., & Yu, C. (2008). Complete transfer of perceptual learning across retinal locations enabled by double training. *Current Biology, 18*, 1922–1926.