Perception, 2008, volume 37, pages 1160 – 1174

doi:10.1068/p6000

Observing object motion induces increased generalization and sensitivity

Benjamin Balas, Pawan Sinha

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 43 Vassar Street, Rm 46-4089, Cambridge, MA 02140, USA; e-mail: Benjamin.Balas@childrens.harvard.edu

Received 23 July 2007, in revised form 8 January 2008; published online 4 August 2008

Abstract. Learning to recognize a new object requires binding together dissimilar images of that object into a common representation. Temporal proximity is a useful computational cue for learning invariant representations. We report experiments that demonstrate two distinct psychophysical effects of temporal association via observed object motion on object perception. First, we use an implicit priming criterion to demonstrate that observation of a dynamic object induces generalization over close temporal neighbors. Second, in contrast to predictions from previous work, we find that shape discrimination between images actually improves following the same training procedure. We suggest that these apparently conflicting sets of results, one demonstrating blurring and the other demonstrating sharpening of the perceived distinction between temporally proximate frames, are consistent with a highly redundant code for object appearance.

1 Introduction

One of the key difficulties faced by all models of object recognition is the fact that any complex 3-D object can give rise to a highly varied set of 2-D images. If a vision system (computational or biological) is to accurately recognize an object in a variety of settings, it must be capable of generalizing over image-changing transformations that preserve object identity while remaining sensitive to image differences that indicate a different object is being viewed (Moghaddam et al 2000; Moses et al 1994). There have been multiple attempts to achieve invariant recognition in computer vision systems by detecting local features and pooling across object parts in a hierarchical manner (Fukushima 1980; LeCun et al 1998; Riesenhuber and Poggio 1999; Ullman et al 2002; Weber et al 2000), but most of these systems require some form of implicit label to function adequately. Ultimately, if an observer were forced to learn about novel objects solely from a set of unlabelled 2-D views, it is unclear how the correct pattern of generalization and selectivity could develop.

Luckily, the world does not force us to learn about objects in this manner. Instead, we are able to observe persisting objects in a dynamic world. Furthermore, the world is 'kind' in that object appearance tends to change smoothly and slowly over time. This scenario offers a great advantage to the observer attempting to learn to recognize complex objects. In a dynamic world, the ways in which an object's 2-D appearance can change within some interval will become apparent, with temporal proximity providing a link between images that may be substantially different from one another. Recent years have seen the development of computational vision systems that use temporal proximity within image sequences as a means for learning specific object invariants (Foldiak 1991; Stone and Harper 1999; Ullman and Bart 2004; Wallis 1996, 1998), demonstrating that this is indeed a useful strategy for building robust object representations. Given the simplicity and computational power of using temporal association as a cue for object learning, understanding the role of dynamic information in visual recognition is a fundamental challenge. In the current study we attempt to gain insight into how dynamic input influences the subsequent representation and recognition of static images. Understanding how dynamic input affects static recognition is a vital

step towards linking classical work on static object recognition to ongoing efforts to characterize recognition in real-world dynamic settings.

How can one tell whether or not biological recognition systems make use of temporal proximity to bind together distinct object views? If such linkages are indeed created following exposure to a dynamic stimulus, temporal neighbors that are bound together should give rise to the same neural or behavioral response. One can think of this as a temporal 'smearing' of appearance, whereby images that appear close in time become less distinguishable as object labels are propagated forward. A variety of methodologies have demonstrated that this sort of behavior emerges after training with image sequences. For example, neurons in the primate inferotemporal cortex begin to respond similarly to highly distinct fractal patterns if those patterns are consistently presented as temporal neighbors during prolonged viewing of a training sequence (Miyashita 1988, 1993; Miyashita and Chang 1988). Human observers also demonstrate intriguing behavioral effects of temporal association across a range of tasks. Increased confusability between individual faces can result from temporal association of those faces in smooth motion sequences (Wallis and Bülthoff 2001), and the learned sequence of 2-D views can impair recovery of 3-D form via the kinetic depth effect (Sinha and Poggio 1996). Even simple translation invariance can be 'broken' by presenting two different objects within a small temporal window (Cox et al 2005). Clearly, temporal association can play a pivotal role in object and face recognition (O'Toole et al 2002). We note, however, that most accounts of the effects of temporal association on recognition have stressed its role in linking images together for invariant recognition. We suggest that such studies consider only one aspect of a learning process that has two important parts.

The ability to generalize object identity across appearance changes is undeniably important, but so, too, is the ability to detect these changes. The goal of an object-recognition system should be to decouple changes in appearance from object identity, rather than to achieve the singular goal of invariance (Ullman 1996). An observer who is perfectly invariant to object transformations by virtue of an inability to discern appearance changes is likely to be at a profound disadvantage. For example, a head-on view of a car requires a very different response than a side view, even though both are to be classified as depicting the same object. Ideally, learning about an object through temporal association of neighboring images would not impair the ability to discriminate between them at the image level. Having the opportunity to observe a change in appearance over time could potentially alert observers to specific regions of the image that are likely to change, or, as we will suggest later, provide for a neural representation of global appearance that supports both generalization and selectivity.

In the current study we asked whether temporal association could lead to both increased generalization over neighboring images and increased sensitivity to the differences between those same images. Using relatively brief amounts of exposure to novel moving objects we found that adult observers did in fact display exactly this pattern of behavior. Neighboring images began to be treated as the same stimulus (as determined by an implicit priming criterion), yet in another task these same images become more discriminable after training. Contrary to the hypothesis of temporal 'smearing' of appearance, we found that observers became more sensitive to appearance changes they observed in a dynamic sequence. We suggest that our results can be interpreted if one assumes an underlying population code for appearance (a proposal that is supported by previous behavioral and physiological studies) that undergoes particular changes in neural tuning in response to observed object motion.

2 Experiment 1: The effect of observed object motion on generalization

In this experiment we used a priming task to examine the effect of passively observed object motion on the representation of static form. Our goal was to extend previous work regarding temporal association and object perception by using novel objects, novel non-rigid motion, parametric variation of object appearance, and the use of an implicit measure of generalization following training. To ensure that observed motion (rather than spatial similarity between static frames or repeated stimulus exposure) was specifically relevant to increased generalization, we compared the effects of training sequences depicting smooth object motion to sequences containing interstitial 'blanks' that disrupted the motion percept.

2.1 *Subjects*

Nineteen volunteers from the MIT community participated in this experiment, all between the ages of 18 and 35 years. All participants reported normal or corrected-to-normal acuity, and were compensated for their participation. Ten observers were randomly assigned to the 'smooth-motion' group, and the remaining nine observers were assigned to the 'no-motion' group.

2.2 *Stimuli*

To ensure that observers could not apply previous knowledge concerning how object appearance might change, we used a novel class of objects that underwent non-rigid deformation during the training sequences presented to our participants. The use of non-rigid motion has the additional benefit of ruling out representational strategies based on a static 3-D object model. Since non-rigid objects have no 'ground truth' 3-D form, observers must learn invariant recognition by linking together distinct appearances rather than by applying general-purpose mechanisms based on rigid-body geometry.

The objects we used (which we will refer to as 'blobs') are constructed from two spherical harmonics that can be independently rotated through various phase angles (Nederhouser et al 2002). By separately rotating each harmonic through the complete range of distinct angles in equal increments, one can construct a toroidal space of images in which 'horizontal' and 'vertical' paths through the space give rise to complex and distinct non-rigid motions (see figure 1). An important aspect of this stimulus set is that the appearance space has been scaled relative to a Gabor-jet-based image similarity metric such that city-block distance is a meaningful measure of low-level similarity along the appearance axes. In experiment 1, one 16-image 'horizontal' strip of blob images was used for training and test stimuli. We used blob images to create training sequences depicting objects in motion and also as static test stimuli.

We created two types of training stimuli by concatenating images into movie sequences (figure 2b). The 'smooth-motion' training sequence was generated by displaying the 16 selected images in forward, then reverse, order at a rate of 12 frames s^{-1}. The resulting movie depicted a blob smoothly deforming in an oscillatory fashion. 'No-motion' sequences were created by inserting a 100 ms empty frame between each frame in the smooth-motion movie. The resulting sequence depicted the same sequence of blob appearances as the smooth-motion sequence, but without a strong motion percept (though it is possible that some apparent motion may have been evident to some subjects).

2.3 *Procedure*

We used an instant priming task (Sekuler and Palmer 1992) to characterize the extent of generalization over distinct object appearances following exposure to a moving object. In instant priming, observers display an RT advantage for judging simultaneously presented targets to be identical if they have been precued with an image that matches the two targets (figure 2a). This paradigm has been successfully used by

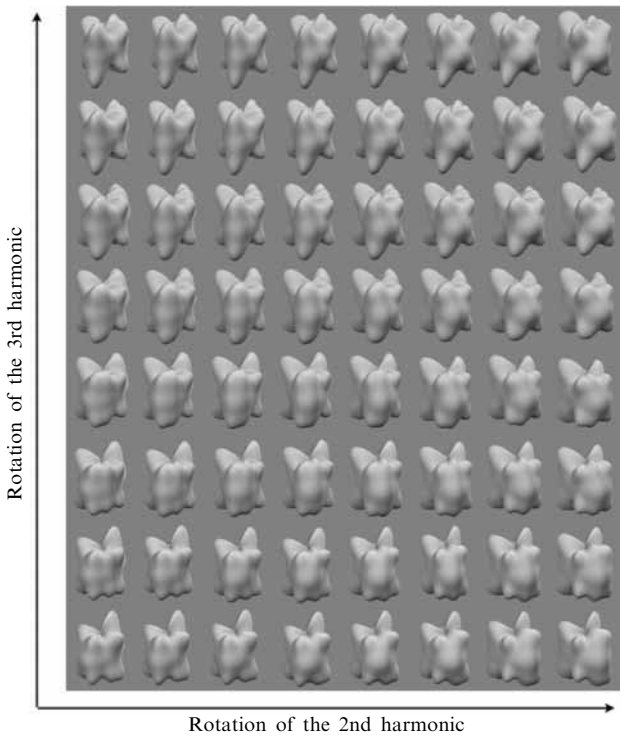

Figure 1. An 8×8 space of 'blob' stimuli. The horizontal and vertical axes of this space are defined by the phase angle of the 2nd and 3rd harmonic. Movement along each axis induces non-rigid motion that is distinct from that generated by movement along the other axis. This image is a schematic view of the full 16×16 blob space used in our experiments. This smaller version has been included for ease of viewing.

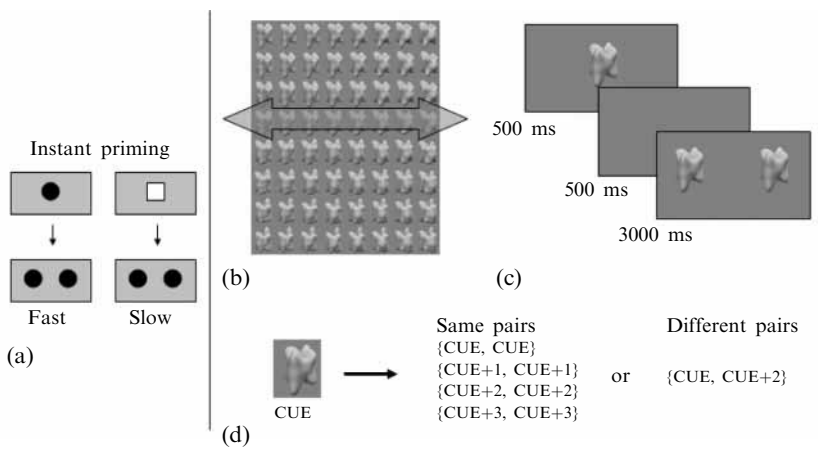

Figure 2. (a) In instant priming, responding that simultaneously presented targets are identical is facilitated by pre-cuing with an image that also matches the targets. There is an RT cost for cue/target dissimilarity. (b) In our training periods, we presented observers with either a smooth-motion stimulus or a no-motion sequence consisting of the same images. Our hypothesis was that continuing exposure to the smoothly deforming object, but not the no-motion sequence, would increase generalization over the strip of object appearances selected for presentation. (c) Following each training round, the test period consisted of go – no-go instant priming trials in which a cue image was followed by two images that either matched each other or differed. RT was measured for "same" responses. (d) Cue/target dissimilarity was parametrically varied during the test period. We calculated the RT cost for making "same" judgments with targets that do not match the cue for several cue/target distances in blob space. Note that, as before, the depicted 8×8 space in (b) is a smaller version of the full blob space used in our tasks.

Kourtzi and Shiffrar to probe the representational content of rigid and non-rigid objects undergoing apparent motion (Kourtzi and Nakayama 2002; Kourtzi and Shiffrar 1997, 1999, 2001). For our purposes, instant priming offers an attractive means for characterizing the relationship between arbitrary cue images and targets via an implicit behavioral effect based on an image-level judgment.

Participants completed three rounds of our task, each round consisting of a training period and an instant priming test period. During each training period, subjects passively viewed the training sequence assigned to them (either the smooth-motion movie or no-motion movie) for 3 min (figure 2b). No response was required during exposure to the motion sequence. Following each training period, subjects performed a go–no-go image-matching task using the static images that composed the training sequences. During this task, each trial began with the presentation of a cue image for 500 ms. Then, after a 500 ms blank interval, two target images were simultaneously presented for 3000 ms (figure 2c). These target images were either identical to one another or different. Subjects were instructed to press the spacebar on the computer keyboard as fast as possible if the two target images were identical and to withhold their response if they were different. RTs to correct "same" judgments were recorded. Since no response was required when targets differed from one another, data from these trials were not analyzed.

Critically, four types of matching target pairs were characterized by their similarity to the cue image presented on each trial. Matching pairs were separated from the cue image by 0, 1, 2, or 3 units in appearance space (figure 2d). The speed advantage conferred by instant priming depends on cue/target similarity, so parametric manipulation of the distance between the cue and target images should lead to systematically increased RTs for responding "same" to targets of increasing dissimilarity to the cue image. Target pairs that were different from one another always contained the cue image and a blob that was 2 units away from the cue in appearance space. Note that all distances between blob images refer to distances along the horizontal strip of images used in this task.

Observers completed 96 "same" trials per round (24 trials $\times$ 4 types of "same" target pairs), and 96 "different" trials. The presentation order of each type of target pair was fully randomized within each test period. Observers typically completed the task in approximately 1 h.

All stimuli were presented on a 19 inch Dell LCD monitor with a 60 Hz refresh rate. Training and test stimuli subtended approximately 2 deg of visual angle. Training sequences and cue images were all presented at the center of the monitor, while target images presented during the test period were displayed 3 deg to the left and right of the center of the monitor. Subject head and eye movements were not restricted or monitored. All stimulus display parameters and response collection routines were controlled by the Matlab Psychophysics toolbox (Brainard 1997; Pelli 1997).

2.4 *Results*

Our hypothesis was that before observers were given much exposure to the training sequences, "same" responses would be more effectively primed by cues that matched the target pairs than by cue images that were dissimilar to the targets. Specifically, we expected that "same" responses would be produced fastest when the cue and target images were the same, and slowest when the cue was very dissimilar to the targets. The difference in RT between the ideal case (cue image matches targets) and the other conditions was expected to provide a measure of the 'cost' of dissimilarity between the cue and the targets.

We further hypothesized that our two training sequences would have differential effects on the RT cost following continued exposure. We expected that, as observers

received continued exposure to the smooth-motion training sequence, temporal association between nearby frames in the sequence would lead to increased generalization over neighboring images. That is, we expected that neighboring images would be bound together into a common representation as training continued, leading to a decrease in RT cost during subsequent test periods. Given the lack (or at least weakening) of a strong-motion percept in the no-motion training sequence, we predicted that continued exposure to this stimulus would not reduce the RT cost substantially.

To test these predictions, we computed the mean RT for correct "same" judgments in each cue/target condition for each subject. We then subtracted out the mean RT for "same" judgments primed by identical cue images (the de facto optimal cue) to yield the RT cost for each condition in which cues did not match targets. If generalization over neighboring images occurred, an initial positive cost for non-matching cues should have given way to a reduced, possibly nil cost after training was completed. Furthermore, if our initial hypotheses were correct, this decrease in RT cost should not have occurred after exposure to the no-motion training sequence. We display the results of this analysis for observers in the smooth-motion and no-motion groups in figure 3.

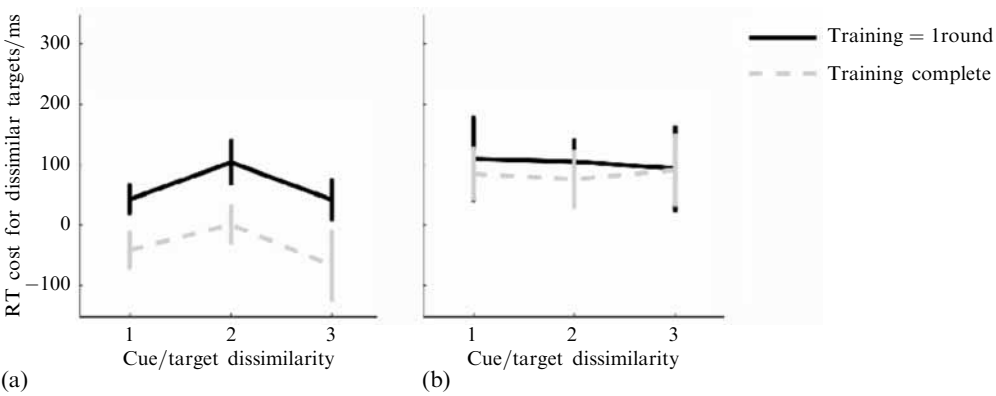

Figure 3. (a) After three rounds of training, an initial RT cost for responding "same" to targets that do not match the cue disappears for observers in the smooth-motion group. (b) Conversely, observers in the no-motion group show no effects of multiple rounds of dynamic exposure. Error bars represent ±1 SEM across the group data in each condition.

First, we examined the results from observers in the smooth-motion group. We carried out a 3×2 repeated-measures ANOVA with cue/target dissimilarity and testing round (first or last) as within-subjects factors. Our analysis revealed a significant effect of testing round ($F_{1,9} = 8.26$, MSE $= 0.146$, $p = 0.014$, $\eta_p^2 = 0.48$) but no significant effect of cue/target dissimilarity ($F_{2,8} = 2.12$, MSE $= 0.024$, $p = 0.15$, $\eta_p^2 = 0.19$). There was no significant interaction between the two factors ($F_{2,8} < 1$). The data from this group of observers thus support our initial hypothesis that continued exposure to object motion can reduce the RT cost of cue/target dissimilarity in the instant priming task.

We continued by conducting the same analysis on the data obtained from observers in the no-motion group. We found no effects of training round ($F_{1,8} < 1$) or cue/target dissimilarity ($F_{2,7} < 1$). The interaction between the two factors was also not significant ($F_{2,7} < 1$). These results support our additional prediction that the absence of perceived motion during passive training periods does not reduce the cost of cue/target dissimilarity.

To examine the extent of generalization over cue/target dissimilarity, we conducted a posteriori comparisons of the RT cost at each cue/target distance in the smooth-motion condition. A two-tailed paired-differences t-test was significant at the smallest

level of cue/target dissimilarity ($t_9 = 2.84$, $p = 0.019$), marginally significant at the next level ($t_9 = 2.25$, $p = 0.05$), and not significant at the greatest dissimilarity level ($t_9 = 1.8$, $p = 0.11$). From this analysis, we conclude that while observed motion led to increased generalization over object appearance, this is most evident over the nearest neighbors in appearance and in time.

2.5 *Discussion*

Passive observation of a smooth-motion sequence induced a reduction in the RT cost for cue/target dissimilarity in our priming task, from which we infer that generalization over the presented object appearances occurred following observed object motion. This result extends previous work by demonstrating measurable effects of temporal association via observed object motion on a novel object class that moves in an unfamiliar and non-rigid way. Furthermore, our results from the smooth-motion group reveal that the extent of generalization over object appearance is strongest for the nearest neighbors in appearance and/or time. Finally, the use of an implicit criterion for characterizing the effects of temporal association between images provides important evidence that the effects of temporal proximity on appearance encoding are not driven by high-level cognitive factors or by response uncertainty. The effect of temporal association on object perception is observable in the context of a purely image-level matching task.

Critically, the data from the no-motion group rule out several trivial explanations of the reduced RT cost observed in the smooth-motion group. The lack of a significant effect of testing round in this group makes it unlikely that the flattened RT costs in the smooth-motion group resulted from task repetition, for example, or repeated exposure to the same static images. Also, since image order was matched in both types of training sequence, we can infer that temporal contingency between frames is not sufficient to drive associations between neighboring images; it is likely that 'real' object motion must be perceived between frames for generalization to occur. This latter point is difficult to state unequivocally, since it is extremely difficult to selectively remove motion from a dynamic stimulus while preserving all other spatial and temporal factors of the input (for example, our insertion of blank frames also lengthens the sequence). At present, we thus cannot firmly conclude that observed object motion per se must be experienced to induce the effects we have observed. For our purposes, this is acceptable insofar as we are presently more interested in exploring the consequences of observed object motion rather than determining the full set of critical properties a stimulus must possess to induce effects like those we have observed in experiment 1.

In experiment 2, we continue by examining the effects of observed object motion on sensitivity for differences in static form. Given that passively observed object motion can lead to measurable changes in generalization within the context of an image-level matching task, how does exposure to the same stimulus modulate sensitivity to image differences? The answer to this question provides us with a more complete description of how object perception is affected by ongoing observation of objects in motion.

3 Experiment 2: The effect of observed object motion on sensitivity

In this task, we investigated how exposure to object motion affects sensitivity to differences in object appearance. We used a change-detection task to measure image-level sensitivity before and after exposure to training sequences with the same parameters as the smooth-motion and no-motion sequences used in experiment 1. This experiment thus allowed us to determine whether exposure to a training stimulus that leads to increased generalization over appearance (as observed in experiment 1) simultaneously induces an increase or decrease in sensitivity to object appearance.

3.1 *Subjects*

Twenty-one volunteers from the MIT community participated in this experiment (eleven in the smooth-motion condition and ten in the no-motion condition), all between 19 and 35 years of age. All participants reported normal or corrected-to-normal acuity, and were compensated for their participation. Subject pools for experiments 1 and 2 were mutually exclusive. The data of an additional participant who completed the task was discarded owing to a failure to follow task instructions.

3.2 *Stimuli*

For this experiment, training and test stimuli were drawn from the space of blob objects described previously. Multiple training stimuli were generated by selecting horizontal and vertical strips of images and concatenating them into extended sequences. As in experiment 1, each individual sequence depicted a blob deforming in an oscillatory fashion. The full sets of 'horizontal' and 'vertical' training sequences were composed of parallel strips of images spaced one image apart in the appearance space (figure 4). Organizing the composition of the training sequences in this manner allowed us to use mutually exclusive image sets for training and test by selectively drawing test images from the gaps left in between the strips of images used to create training sequences.

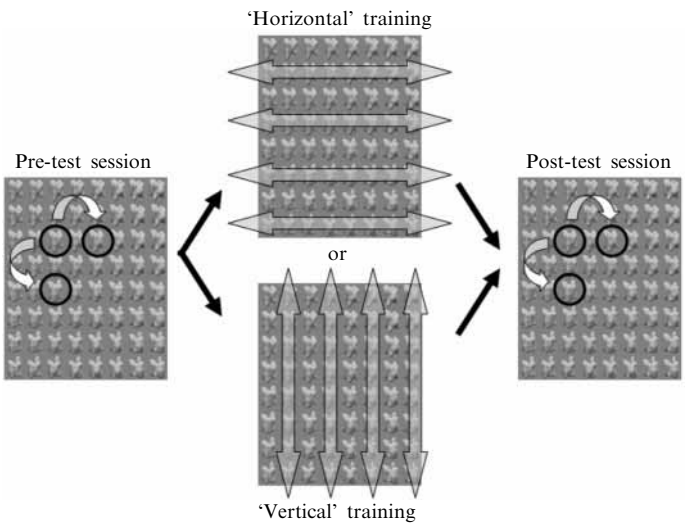

Figure 4. In our change-detection paradigm, we first measured subjects' sensitivity to static image differences between 'horizontal' and 'vertical' image pairs in a same/different task. Following this, subjects were exposed to training sequences depicting blobs deforming along one axis (either horizontal or vertical) through appearance space. Sensitivity to image changes along the trained and untrained axes was reassessed after the training period, and changes in sensitivity along each axis were recorded. As in the previous figures, an 8×8 schematic view of blob space is displayed here for ease of viewing. The full stimulus set used consisted of 16×16 images.

3.3 *Procedure*

Participants in this task completed two rounds of a change-detection task, both before and after passive exposure to either 'horizontal' or 'vertical' training sequences.

On each trial of the change-detection task, observers sequentially viewed two blob images and had to decide if they were identical or different. First, one blob image was presented for 250 ms, followed by a 200 ms blank period, a 200 ms presentation of a $1/f$ fractal noise mask, and the presentation of a second blob stimulus for an additional 250 ms. The position of each of the two blobs was randomly jittered within a ± 1 deg interval around the center of the monitor. On each trial, the two blob stimuli presented could be identical, differ in their position along the 'horizontal' axis,

or differ in their position along the 'vertical' axis. Stimulus pairs that were different were always separated by 2 distance units in the underlying appearance space. Observers were presented with 256 'different' pairs, half of which differed in the 'horizontal' direction and half of which differed in the 'vertical' direction. Also, 256 "same" trials were included for a grand total of 512 trials per test session. The order of pairs presented during the experiment was randomized for each subject. During the pre-training session, auditory feedback was given to subjects to indicate incorrect responses. During the post-training session, no feedback was given.

During the training period, participants passively viewed 8 unique image sequences. Half of our observers were shown only the 8 training sequences in the 'horizontal' set of movies and the remaining half were shown only the 8 training sequences in the 'vertical' set of movies. Each individual sequence lasted approximately 30 s, and was displayed at a rate of 12 frames s^{-1}. Each individual sequence was presented 3 times, and the order of sequence presentation was randomized for each subject. The full training period lasted approximately 12 min.

Smooth-motion sequences were not manipulated further for display during the training period. No-motion sequences were modified such that a blank frame lasting 100 ms was inserted between all image frames in the original sequence. The resulting stimulus (as in experiment 1) depicted the same sequence of static images as each corresponding smooth-motion sequence, but with a substantially weakened motion percept.

3.4 *Results*

Our goal in this experiment was to determine how exposure to object motion affects sensitivity to static form. In experiment 1 we observed that exposure to smooth object motion increased generalization over static images presented close together in time, leading to a reduced RT cost for dissimilar cue images in our priming task. In experiment 2, we asked how exposure to similar smooth and 'jerky' sequences modulated observers' ability to discriminate between images arranged along the trained and untrained appearance axes.

In both pre-test and post-test sessions, the ability to detect image differences along the horizontal and vertical axes of appearance space was characterized by calculating d' for 'horizontal' and 'vertical' pairs. We collapsed d' values across observers by relabeling trials as having contained 'trained' and 'untrained' form differences, according to the set of sequences observed during the training period. This was done to obviate the need to consider potential differences in baseline sensitivity along the two appearance axes. Within a test session, d' values in the trained and untrained direction were calculated with a shared false alarm obtained from all "same" trials presented in that session. This means that differences in d' observed between the trained and untrained direction within a session are really only differences in hit rate. However, we report d' values here rather than hit rates, since the false-alarm rate could differ between the pre-test and post-test sessions and we wished to meaningfully compare performance across these two sessions.

In figure 5 we display the mean d' values across all smooth-motion observers in all conditions.

We carried out a 2×2 repeated-measures ANOVA on the d' values from smooth-motion observers, with type of stimulus pair (trained or untrained form differences) and testing session (pre-exposure and post-exposure) as within-subjects factors. This analysis revealed a main effect of testing session ($F_{1,10} = 11.68$, MSE $= 0.55$, $p = 0.007$, $\eta_p^2 = 0.54$), but no significant effect of stimulus type ($F_{1,10} = 1.14$, MSE $= 1.14$, $p = 0.31$, $\eta_p^2 = 0.10$). The interaction between these two factors was also not significant ($F_{1,10} = 1.41$, MSE $= 0.057$, $p = 0.26$, $\eta_p^2 = 0.12$).

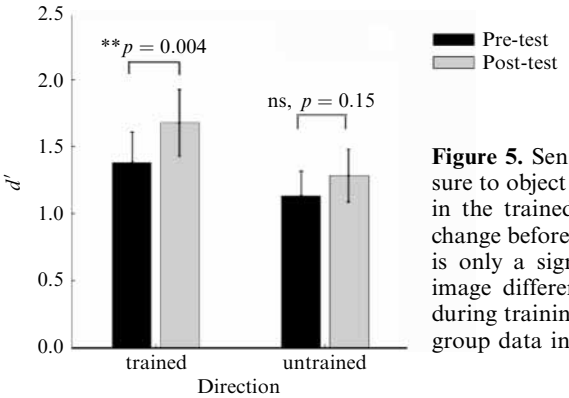

Figure 5. Sensitivity to form differences following exposure to object motion: average d' scores across observers in the trained and untrained directions of appearance change before and after a passive exposure period. There is only a significant effect of the exposure period for image differences along the appearance axis observed during training. Error bars represent ± 1 SEM across the group data in each condition.

We also carried out two planned comparisons on these d' values to examine the effects of the training sequences more closely. We compared pre-test performance to post-test performance in the trained and untrained directions. This analysis revealed a significant effect of the exposure period on d' values recorded in the trained direction (two-tailed paired-differences t-test, $t_{10} = 3.68$, $p = 0.004$) but no effect of training on performance in the untrained direction (two-tailed paired-differences t-test, $t_{10} = 1.56$, $p = 0.15$). This analysis suggests that the previously observed main effect of testing session is driven primarily by improved sensitivity for differences along the appearance axis observed during the training period.

Given only these data, it is difficult to conclude whether or not increased sensitivity in the trained direction depends critically on observing object motion, or depends solely on observing a collection of static images. To help disambiguate between these two possibilities, we next turn to the data from the no-motion group. We analyze the data from the pre-test and post-test change-detection sessions in the same manner as described above for the smooth-motion group. The mean d' values for pre-test and post-test performance in the trained and untrained direction are reported in table 1.

Table 1. Sensitivity (d') to appearance changes in the trained and untrained directions following exposure to the 'no-motion' sequences. As reported in the text, there are no significant differences between pre-exposure and post-exposure means. The data reported in each cell are the mean d' values across subjects, with the standard deviation in parentheses.

Direction	Pre-exposure test	Post-exposure test
Trained	1.14 (0.51)	1.37 (0.69)
Untrained	1.15 (0.52)	1.28 (0.43)

A 2×2 repeated-measures ANOVA reveals that there are no main effects of stimulus type ($F_{1,9} < 1$) or testing session ($F_{1,9} = 2.8$, $p = 0.13$). The interaction between these factors was also not significant ($F_{1,9} = 1.8$, $p = 0.23$). Finally, as in our analysis of the smooth-motion data, we also carried out two pre-planned comparisons of the d' values in the trained and untrained directions before and after exposure to the blob sequences. This also revealed no significant effects of testing session in either the trained ($t_9 = 1.72$, $p = 0.12$, two-tailed paired-differences t-test) or untrained direction ($t_9 = 1.36$, $p = 0.21$). The lack of any increases in sensitivity following observation of the no-motion sequences supports the conclusion that observing object motion (rather than simply many static images) is critical in this task.

3.5 *Discussion*

Given exposure to the same type of smooth-motion sequences that induced increased generalization in experiment 1, observers in experiment 2 showed improved sensitivity for the changes observed during training. This was not observed in the group that was exposed to no-motion sequences, indicating that exposure to the static stimuli is not sufficient to induce this effect. This result (an increase in sensitivity following observed motion) was somewhat unexpected, since increased confusion between images presented close together in time is often used as an index of associative mechanisms at work during exposure to a moving object. The results of previous research thus might have led one to predict decreased sensitivity in the trained direction of appearance space, or no change at all due to the benefits of practice being cancelled out by reduced sensitivity brought on by association between neighboring images.

Several simple explanations can be ruled out that are based on our design. For example, it is unlikely that observers used the training sequence to identify local image regions where change was expected. Comparing specific object-centered image regions across test stimuli was made difficult both by the smoothness of the blobs and the fact that we randomly jittered the position of both test images on each trial. Also, given the global, non-rigid, and non-uniform nature of the deformations depicted in the training sequences, singling out any one region of the image on randomly selected trials would have been an ineffective strategy. The role of learned prediction in performing this task was also minimized. Each image in the training sequences predicted two images equally well (owing to the forward and backward oscillation along appearance axes during training), so that there was no unambiguous prediction to be made from any individual image. Finally, mere exposure effects can be ruled out. Observers were only tested on images that did not appear in the training sequences, and our use of the no-motion condition as a control further rules out this explanation.

As we discussed following experiment 1, it is difficult to state unequivocally that real object motion is necessary to induce these effects. However, we can safely conclude that the effects observed in experiments 1 and 2 are consequences of object motion, even if some more fundamental aspect of the stimulus actually causes these changes in behavior. Our goal throughout this study has been to characterize the impact of observed object motion on subsequent static form processing, and the results of these two experiments provide a rich and complex picture of that relationship. We conclude by presenting a unified interpretation of the data from experiments 1 and 2, arguing that the combination of increased generalization and sensitivity may reveal an important aspect of the underlying perceptual code for object appearance.

4 General discussion

We have found that the observation of object motion was followed by increased generalization over temporally close images and also increased sensitivity to the differences between those images. These results support the notion that object learning is a dual process of acquiring invariance and learning to detect subtle variations in object appearance. The visual system learns about objects such that the ultimate goal of a balance between good recognition and good discrimination is met. Our results also suggest that temporal association (in our case, via observed object motion) plays an important role in both aspects of this learning.

In experiment 1 we demonstrated that generalization to temporal neighbors is not just evident at the level of object naming, but also affects performance in an image-level task. Moreover, we assessed the strength of image binding over increasing dissimilarity between images, finding that image binding is strongest over images that are the closest neighbors in time and appearance. Finally, our comparison of smooth motion to discontinuous presentation of distinct object appearances makes a strong case for temporal

continuity as a stronger cue for object learning than mere exposure to static images. Though observers can use spatial continuity to bolster view-invariant performance (Perry et al 2006), our results indicate that temporal continuity is of primary importance for generalization. The results of experiment 2 are also important in that the use of an image-level judgment makes a strong case for a perceptual, rather than cognitive, effect of observed object motion on form processing. Furthermore, the result that sensitivity improves along the axis of observed appearance change is novel to the best of our knowledge, and leads us to an important insight regarding the nature of object representation.

It appears difficult at first to account for both our priming results and our change-detection results with one mechanism. If the sole function of temporal association is to bind images together into a common representation, we might expect that increased generalization would lead to impaired sensitivity. Learning to treat two images as though they were the same might be thought to make them hard to discriminate, but this is not what we observe in our change-detection task. If observers are becoming better at generalizing over observed appearance changes, how are they also becoming more sensitive to the same changes?

One possibility is that the implicit measure used in experiment 1 and the explicit measure of discriminability used in experiment 2 tap into different levels of object representation. This distinction between the results obtained with implicit and explicit measures has been discussed before (Biederman and Gerhardstein 1993; Lawson 2004; Seamon and Delgado 1999) and our data are generally consistent with the proposal that invariant recognition tends to be more evident with the use of implicit, rather than explicit, measures. To the extent that this is an accurate characterization of the relationship between implicit and explicit measures of object perception, our results represent a generalization of this model beyond 3-D view sensitivity to more general (non-rigid) appearance change.

On a neural level, we must also recognize the possibility that our results reveal the operation of two distinct neural populations: one in which motion is used to increase invariance, and one in which motion is used to improve discrimination. This proposal explains the data, but requires separate mechanisms for learning invariance and sensitivity. During execution of a particular task, this proposal also seemingly requires the observer to selectively recruit distinct neural populations in a task-dependent way. Our data do not allow us to rule this possibility out, but we continue by discussing an interesting alternative.

We suggest that our results can be well explained by positing an underlying population code for object appearance that uses observed motion to develop a representation of form that is not highly local. Encoding object appearance with a highly local code would tend to lead to a direct trade-off between generalization and selectivity. If we imagine a single unit with a typical bell-shaped tuning curve over appearances, it is clear that broadening this curve to generalize over more appearances simultaneously decreases the ability to discriminate between them (figure 6a). By contrast, encoding appearance to using a population of units can potentially support both superior generalization and better sensitivity (Hinton et al 1986). Simulations of coarse or distributed coding have demonstrated that extremely good resolution for 'coordinate' judgments can be achieved with redundant representations in which tuning curves overlap substantially (Jacobs 1996; Jacobs and Kosslyn 1994; Milner 1974). It is easy to apply this idea to object-appearance encoding, and thus explain both of our results with one mechanism. First, we assume some initial population of object-selective cells that differ in their preferred stimulus and that initially have tuning curves that do not overlap substantially. Second, we assume that exposure to a moving object causes each cell to widen its tuning functions so that it responds to a wider range of object appearances. Crucially, widening must not occur symmetrically in appearance space.

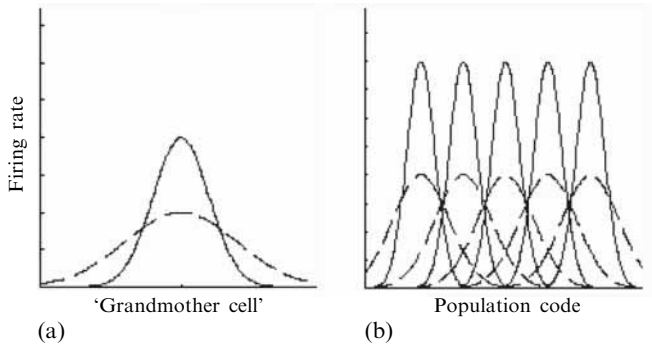

Figure 6. Schematic view of how tuning curves might change following exposure to object motion under two distinct processing frameworks. In each case, we posit that object motion induces broadening of the underlying tuning function (dashed lines), in keeping with previous behavioral work concerning temporal association in primates and humans. (a) A 'grandmother cell' forces a trade-off between generalization and sensitivity. (b) Within certain limits, a population code can support increases in both, making it possible to interpret our results in this framework.

Instead, tuning curves must widen more in the direction commensurate with stimulation (along the observed axis of appearance change during our training periods, for example), leading to some degree of overlap (figure 6b). So long as the feature space is not too dense and the tuning functions do not become too large (Hinton et al 1986), better generalization ability and better resolution in this space can result from the overlapping tuning curves present in this appearance code.

Our theoretical proposal of a population code with units tuned for particular appearances or views is consistent with several physiological results. View-tuned neurons have been found in primate inferotemporal cortex for familiar (Perrett et al 1992) and novel (Logothetis et al 1995) objects. There are reports of highly view-invariant responses for familiar objects as well, but even in these studies the majority of cells show selectivity for particular appearances (Booth and Rolls 1998). Psychophysical data from both humans and monkeys provide further evidence to support population coding for complex objects (Fang and He 2005; Logothetis et al 1994). To our knowledge, the effects of dynamic exposure on the tuning of view-selective cells has not been directly examined. Though there is evidence that preceding action can affect the response of cells specific for body posture in the macaque temporal lobe (Jellema and Perrett 2003), the immediate effects of dynamic stimulation on appearance tuning for arbitrary objects has not been examined.

This proposal points towards some intriguing avenues for future research. For example, coarse coding in a feature space ceases to provide gains in resolution once the tuning curves grow too large. This suggests that there should be a point where further generalization can occur, but sensitivity does not increase. Providing observers with extensive exposure to dynamic objects may reveal this limiting behavior. It would also be interesting to examine how a predictive relationship between image pairs interacts with the effects of dynamic exposure we have reported here. This could be studied in the context of objects like the human body, that are familiar to the observer in form and characteristic motion. Alternatively, one could continue to use novel objects, building predictive relationships into the dynamic stimulus.

5 Conclusions

We have demonstrated in two psychophysical tasks that temporal association between images results in both increased generalization over distinct images and increased sensitivity to the differences between those stimuli. We have suggested that these data can be explained within the context of a population code for object appearance in

which observed motion leads to increased overlap of the tuning functions along the direction of appearance change in the relevant feature space. This proposal can account for all of our main results with only one proposed change in neural tuning following observed object motion. Taken together, these results provide insight into how dynamic input can affect the representation of static form, while simultaneously revealing an important aspect of the underlying code for object appearance. This bridge between dynamic and static object appearance is an important first step towards understanding how the visual system rapidly and simultaneously learns representations to support multiple visual tasks in the fully dynamic world.

Acknowledgments. BJB is supported by a National Defense Science and Engineering Graduate Fellowship. This work has benefited greatly from the observations and advice of David Cox, Dick Held, and Yuri Ostrovsky. Two anonymous reviewers also provided many helpful suggestions.

References

Biederman I, Gerhardstein P C, 1993 "Recognising depth-rotated objects: Evidence for 3-D viewpoint invariance" *Journal of Experimental Psychology: Human Perception and Performance* **19** 1162 – 1182

Booth M C A, Rolls E T, 1998 "View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex" *Cerebral Cortex* **8** 510 – 523

Brainard D H, 1997 "The psychophysics toolbox" *Spatial Vision* **10** 433 – 436

Cox D, Meier P, Oertelt N, DiCarlo J J, 2005 "'Breaking' position-invariant object recognition" *Nature Neuroscience* **8** 1145 – 1147

Fang F, He S, 2005 "View-centered object representation in the human visual system revealed by viewpoint aftereffects" *Neuron* **45** 793 – 800

Foldiak P, 1991 "Learning invariance from transformation sequences" *Neural Computation* **3** 194 – 200

Fukushima K, 1980 "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position" *Biological Cybernetics* **36**(4) 193 – 202

Hinton G E, McClelland J L, Rumelhart D E, 1986 "Distributed representations", in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* Eds D E Rumelhart, J L McClelland (Cambridge, MA: MIT Press) pp 77 – 109

Jacobs R A, 1996 "Computational studies of the development of functionally specialized neural modules" *Trends in Cognitive Sciences* **3** 31 – 38

Jacobs R A, Kosslyn S M, 1994 "Encoding shape and spatial relations: the role of receptive field size in coordinating complementary representations" *Cognitive Science* **18** 361 – 386

Jellema T, Perrett D I, 2003 "Perceptual history influences neural responses to face and body postures" *Journal of Cognitive Neuroscience* **15** 961 – 971

Kourtzi Z, Nakayama K, 2002 "Distinct mechanisms for the representation of moving and static objects" *Visual Cognition* **9** 248 – 264

Kourtzi Z, Shiffrar M, 1997 "One-shot view invariance in a moving world" *Psychological Science* **8** 461 – 466

Kourtzi Z, Shiffrar M, 1999 "The visual representation of three-dimensional rotating objects" *Acta Psychologica* **102** 265 – 292

Kourtzi Z, Shiffrar M, 2001 "Visual representation of malleable and rigid objects that deform as they rotate" *Journal of Experimental Psychology: Human Perception and Performance* **27** 335 – 355

Lawson R, 2004 "Depth rotation and mirror-image reflection reduce affective preference as well as recognition memory for pictures of novel objects" *Memory & Cognition* **32** 1170 – 1181

LeCun Y, Bottou L, Bengio Y, Haffner P, 1998 "Gradient-based learning applied to document recognition" *Proceedings of the IEEE* **86** 2278 – 2324

Logothetis N, Pauls J, Poggio T, 1995 "Shape representation in the inferior temporal cortex of monkeys" *Current Biology* **5** 552 – 563

Logothetis N K, Pauls J, Bülthoff H H, Poggio T, 1994 "View-dependent object recognition by monkeys" *Current Biology* **4** 401 – 414

Milner P M, 1974 "A model for visual shape recognition" *Psychological Review* **81** 521 – 535

Miyashita Y, 1988 "Neuronal correlate of visual associative long-term memory in the primate temporal cortex" *Nature* **335** 68 – 70

Miyashita Y, 1993 "Inferior temporal cortex: where visual perception meets memory" *Annual Reviews of Neuroscience* **16** 245 – 263

Miyashita Y, Chang H S, 1988 "Neuronal correlate of pictorial short-term memory in the primate temporal cortex" *Nature* **331** 307 – 311
Moghaddam B, Jebara T, Pentland A, 2000 "Bayesian face recognition" *Pattern Recognition* **333** 1771 – 1782
Moses Y, Adini Y, Ullman S, 1994 "Face recognition: the problem of compensating for illumination changes" *Proceedings of the European Conference on Computer Vision, Seacaucus, NJ* (Springer, New York) pp 286 – 296
Nederhouser M, Mangini M C, Biederman I, 2002 "The matching of smooth, blobby objects—but not faces—is invariant to differences in contrast polarity for both naive and expert subjects" *Journal of Vision* **2** 745a (abstract)
O'Toole A J, Roark D A, Abdi H, 2002 "Recognizing moving faces: A psychological and neural synthesis" *Trends in Cognitive Sciences* **6** 261 – 266
Pelli D G, 1997 "The VideoToolbox software for visual psychophysics: transforming numbers into movies" *Spatial Vision* **10** 437 – 442
Perrett D I, Hietanen J K, Oram M W, Benson P J, 1992 "Organization and functions of cells responsive to faces in the temporal cortex" *Philosophical Transactions of the Royal Society of London, Series B* **335** 23 – 30
Perry G, Rolls E T, Stringer S M, 2006 "Spatial vs temporal continuity in view invariant visual object recognition learning" *Vision Research* **46** 3994 – 4006
Riesenhuber M, Poggio T, 1999 "Hierarchical models of object recognition in cortex" *Nature Neuroscience* **2** 1019 – 1025
Seamon J J, Delgado M R, 1999 "Recognition memory and affective preference for depth-rotated solid objects: part-based structural descriptions may underlie the mere exposure effect" *Visual Cognition* **6** 145 – 164
Sekuler A B, Palmer S E, 1992 "Perception of partly occluded objects: A microgenetic analysis" *Journal of Experimental Psychology: General* **121** 95 – 111
Sinha P, Poggio T, 1996 "The role of learning in 3-D form perception" *Nature* **384** 460 – 463
Stone J V, Harper N, 1999 "Temporal constraints on visual learning: a computational model" *Perception* **28** 1089 – 1104
Ullman S, 1996 *High-Level Vision* (Cambridge, MA: MIT Press)
Ullman S, Bart E, 2004 "Recognition invariance obtained by extended and invariant features" *Neural Networks* **17** 833 – 848
Ullman S, Vidal-Naquet M, Sali E, 2002 "Visual features of intermediate complexity and their use in classification" *Nature Neuroscience* **5** 682 – 687
Wallis G, 1996 "Using spatio-temporal correlations to learn invariant object recognition" *Neural Networks* **9** 1513 – 1519
Wallis G, 1998 "Spatio-temporal influences at the neural level of object recognition" *Neural Networks* **9** 265 – 278
Wallis G, Bülthoff H H, 2001 "Effects of temporal association on recognition memory" *Proceedings of the National Academy of Sciences of the USA* **98** 4800 – 4804
Weber M, Welling M, Perona P, 2000 "Unsupervised learning of models for recognition" *Proceedings of the European Conference on Computer Vision, Part 1* (Berlin: Springer) pp 18 – 32

p
© 2008 a Pion publication

ISSN 0301-0066 (print) ISSN 1468-4233 (electronic)

PERCEPTION
VOLUME 37 2008

www.perceptionweb.com

Conditions of use. This article may be downloaded from the Perception website for personal research by members of subscribing organisations. Authors are entitled to distribute their own article (in printed form or by e-mail) to up to 50 people. This PDF may not be placed on any website (or other online distribution system) without permission of the publisher.