Learning About Dynamic Objects and Recognizing Static Form

By

Benjamin J. Balas

S.B. Brain and Cognitive Sciences
Massachusetts Institute of Technology, 2002

SUBMITTED TO THE DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN BRAIN AND COGNITIVE SCIENCES
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2007

Signature of Author: _____

Department of Brain and Cognitive Sciences
May 2007

Certified by: _____

Pawan Sinha
Associate Professor of Vision and Computational Neuroscience
Thesis Supervisor

Accepted by: _____

Matt Wilson
Professor of Neurobiology
Chair, Department Graduate Committee

Learning About Dynamic Objects and Recognizing Static Form

by

Benjamin J. Balas

Submitted to the Department of Brain and Cognitive Sciences
on May 7th, 2007 in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy in
Brain and Cognitive Sciences

Abstract

The effects of observed object motion on object perception are examined in two sets of studies. The first section of the thesis provides a thorough examination of various untested aspects of the basic "temporal association" hypothesis, which suggests that object motion provides a principled basis for linking distinct images together if they appear within small time intervals. Using familiar and unfamiliar objects undergoing various forms of non-rigid motion, I ask how well this simple hypothesis predicts behavior in change detection and categorization tasks. The results favor a modified version of the hypothesis which operates over a population of units, such that increases in generalization also produce increases in image sensitivity. The observed effects of long-term knowledge concerning object appearance and expected patterns of motion also force additional modifications of the initial hypothesis to incorporate interactions between learned predictions and recent experience. Specifically, the tendency to alter patterns of generalization following dynamic exposure appears to be contingent on the stability of the direction of movement through appearance space. Consistent with this expanded model, performance in our categorization task appears to depend heavily on whether or not a coherent direction of movement through appearance space can be determined across both categories to be learned.

In the second section of the thesis, I report the results of two parametric analyses of image encoding following dynamic exposure. In each case, I ask how the movement of an object up to the presentation of particular image affects an observers' ability to accurately recall that image. Novel, rigidly rotating objects are used in both cases to characterize the influence of appearance dynamics on short and long-term image encoding. In both cases, I find that local appearance change over time exerts a powerful influence on encoding, suggesting that both immediate percepts and visual memory are modulated by the recent past. The result is a complex picture of dynamic object perception that goes far beyond the basic principle of object motion as a tool for learning invariant recognition.

Thesis Supervisor: Pawan Sinha
Title: Assistant Professor of Vision and Computational Neuroscience

Acknowledgements

I would first of all like to thank Jim DiCarlo, Eric Grimson, and Maggie Shiffrar for agreeing to be on my thesis committee. They bravely took up the treacherous task of reading this thesis and helped me make sense of all this. Their many helpful comments have vastly improved this work beyond its initial chaotic formulation.

Second, I must thank Pawan Sinha for his many years of guidance and his unmatched generosity throughout the course of my graduate career. His creativity, brilliance, and good humor have been an inspiration to me during my time in his laboratory. While he has always provided as much assistance as I needed, I am even more grateful for the freedom he has given me to develop and pursue my own ideas.

There are also many individuals who together fall somewhere in between the words "colleague," "collaborator," and "friend." Some of the following individuals have been all three, and all of them have been invaluable:

Dave Cox has bolstered my optimism and pessimism in whatever direction helped most.

Chris Connor has thoroughly convinced me that we scientists are really getting away with something great by being paid to do research.

Charles Kemp is possibly the best roommate ever and one of the smartest people I know.

Richard Russell taught me to look at all things from the other side, including cognitive science, computation, and doughnuts.

I must also thank many members of the MIT BCS faculty and staff who, while not being my direct advisors, have given their time and their support to me in many ways over the past few years. They include, Ted Adelson, Denise Heintze, Dick Held, Nancy Kanwisher, Aude Oliva, and Ruth Rosenholtz.

Finally, I must thank the three people without whom I cannot imagine having accomplished any of this. My parents, Barry and Janet Balas, have always given me their unwavering love and support, and I find I do not have words to thank them enough for their commitment to my education and my happiness. My wife, Erin Conwell, has been my best friend, my favorite collaborator, and the only person I'll ever really trust onstage. Though it seems meager thanks for all they have done for me, it is to the three of them that I dedicate this thesis.

Biographical Sketch

At the tender age of 5, Ben Balas (having reinvented for himself most of modern mathematics and all of contemporary American Literature) was sent to an ordinary public school to begin his academic career. He progressed from elementary school to upper elementary without incident, proceeding onwards through junior high and high school, spending the majority of those years watching a wide variety of movies on basic cable and simultaneously honing his skills at Street Fighter II: Champion Edition.

In 1998, having a wide variety of interests in the arts and sciences, he chose to attend MIT as a mathematics and physics major so as to limit his studies to only that which could be studied within a clear axiomatic framework.

In 1999, realizing he had no idea what constituted a "clear axiomatic framework," he discontinued his plans for a double degree in mathematics and physics in favor of a course of studies in the department of Brain and Cognitive Sciences. His main interests were in the study of vision, but he also wholeheartedly pursued a minor in film studies until he realized he would have to take several boring, difficult classes to fulfill departmental requirements.

After 4 years of study, it was agreed upon by all that while a degree would certainly be awarded for his work to date, it was "less than clear" that anything of consequence had been accomplished. Graciously ignoring the amount of time spent by the author on improvisational comedy, experimental film, and comic book geekery, it was decided that he would be granted a "Mulligan" this time around, on the condition that he remain at MIT for a PhD in Brain and Cognitive Science. Unaware of his reputation, Dr. Pawan Sinha invited Ben to join his lab, where he was remained up to the present day. The rest, as they say, is history.

Despite an extraordinary amount of time spent playing online arcade games and solving various forms of inconsequential logic puzzles, Ben Balas is the author of the following research articles:

Balas, B. & Tenenbaum, J. (2004). Domain-Specificity in Shape Categorization and Perception. *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, 67-72.

Balas, B. & Connor, C.W. (2004) Look Up and Scream: Analytical Difficulties in Improv Comedy. *Journal of Recreational Mathematics*, 33(1), 32-38.

Sinha, P., Balas, B., Ostrovsky, Y., & Russell, R. (2005). Face recognition by Humans. In "*Face Recognition: Mechanisms and Models*", Editor R. Chellapa, Academic Press.

Gallagher, K., Balas, B., Matheny, J. & Sinha, P. (2005) The Effects of Scene Category and Content on Boundary Extension. *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*.

Balas, B. (2006) Using Computational Models to Study Texture Representations in the Human Visual System. *Vision Research*, 46, 299-309.

Balas, B. & Sinha, P. (2006) Receptive Field Structures for Recognition. *Neural Computation*, 18(3), 497-520.

Balas, B., Cox, D.A. & Conwell, E. (2006) The effect of personal familiarity on the speed of face recognition. ***Proceedings of the 28th Annual Meeting of the Cognitive Science Society.***

Balas, B. & Sinha, P. (in press) Region-based representations for face recognition, ***ACM Transactions on Applied Perception.***

Sinha, P., Balas, B., Ostrovsky, Y. & Russell, R. (in press) Face Recognition by Humans: 19 Results all Computer Vision Researchers Should Know About. ***Proceedings of the IEEE.***

Balas, B. & Sinha, P. (in press) "Filling-in" color in natural scenes. ***Visual Cognition***

Balas, B. & Sinha, P. (in press) Portraits and Perception: Configural Information in Recognizing and Creating Face Images., ***Spatial Vision***

Table of Contents

Introduction

Object recognition takes place in a dynamic world. Discrete objects move around us, constantly changing in pose, illumination conditions, scale, and position while we observe their appearance. The visual world is also "kind" in that object appearance almost always varies smoothly and slowly. This has consequences for both low-level encoding of image structure and high-level representations of object appearance. In terms of image statistics, the substantial redundancy in the spatiotemporal volume corresponding to an arbitrary object's appearance observed over some time interval means that low-level encoding of spatiotemporal image features can be accomplished via "sparse" neural codes. (Dong & Atick, 1995; Olshausen, 2003; van Hateran & Ruderman, 1998). This is important for learning general neural coding strategies of natural scenes. For the purposes of object recognition, the record of what an object looks like over a densely sampled period of time should also provide an extremely rich source of data for constructing a robust appearance model useful for a wide range of recognition tasks. To the extent that high-dimensional image data really corresponds to a low-dimensional manifold (Edelman, 1999; Murase & Nayar, 1995), "walking" slowly through appearance space on a smooth path should make recovery of the true degrees of freedom for object appearance more tractable (Tenenbaum, de Silva, & Langford, 2000).

The idea that the dynamics of the visual world play a pivotal role in perception is not new. Still, despite a great deal of evidence that temporal factors can and do influence high-level visual tasks, we have yet to see a substantive model of object recognition that incorporates object motion. Classical models of object recognition, like recognition-by components or view-based models (Biederman, 1987; Tarr & Bulthoff, 1995), have not yet been modified to accommodate the potential role of dynamic data in learning object representations or carrying out recognition.

Is it necessary to modify these models? Perhaps not, since we're more than capable of recognizing objects in static photographs just as well as if they were in front of us. Also, recent computational models of object concept learning are able to do fairly well at learning and recognizing objects in clutter solely from a large set of labeled static images(Agarwal & Roth, 2002; Fei-Fei, Fergus, & Perona, 2004; Fergus, Perona, & Zisserman, 2003; Weber, Welling, & Perona, 2000). A continually changing visual world provides us with an abundance of static data that we can use to extract useful regularities, but recent work has shown that useful gestalt-like cues for grouping can be learned from video sequences without using the temporal contingencies in that data (Prodohl, Wurtz, & von der Malsburg, 2003). Similarly, structure-from-motion algorithms can be used to extract 3-D form from multiple static views (Ullman, 1979), providing a useful static representation of form. In this case, the temporal structure of the data is important insofar as it aids in obtaining multiple views and establishing correspondence between image features. However, object motion is really only useful here in the service of form recovery. The particular motion observed during learning is of no consequence, so long as it allows for a solution of the underlying 3-D shape.

How relevant is object motion for object learning and recognition? Certainly motion could be useful to get large amounts of static data, or to segment an object from clutter

(Brady & Kersten, 2003), but is there any reason to think that object motion is at all instrumental in the construction of object representations for recognition? In this thesis, I will argue that the answer is yes. I suggest that the observation of coherent object motion facilitates the construction of a coarse population code for object appearance.

My goal in the experiments reported here has been two-fold. First of all, I have set out to investigate the consequences of observed object motion on the discrimination and recognition of static form. To do so, I have designed and carried out a series of experiments inspired by what is known as the "temporal association" hypothesis. This is a relatively recent theory suggesting that temporal proximity is used as an implicit cue by the human visual system to bind dissimilar images of an object together into a common representation. This hypothesis provides an important link between dynamic experience and static recognition. By examining observers' recognition abilities following exposure to dynamic objects, I have determined a set of empirical constraints that rule out strong versions of the temporal association hypothesis and point towards a model of object learning that is more consistent with behavioral and physiological data. My second goal has been to determine what information is extracted from sequences depicting object motion, with a particular emphasis on how privileged views, or what I will call "keyframes," are determined online. That is, after seeing a moving object, what do you actually remember seeing? Here I look for evidence of prototype effects in memory for dynamic objects, and also thoroughly investigate how appearance dynamics modulate the fidelity of image encoding as a function of distance in image-based and model-based representation spaces. The end result is the beginning of a theory concerning how dynamic information is used to define a robust appearance code for recognition tasks at multiple levels.

In the remainder of this introduction, I will outline relevant work concerning object motion and object recognition, and discuss what consequences these results have for models of how dynamic input could be used during the acquisition and application of object representations. Following that, I will explain what contributions I make in the current series of experiments and how my approach differs from previous research.

Object motion and object recognition: A review

There exists a rich body of work describing various interesting effects in which the motion of an object somehow modulates the perception or recognition abilities of an adult observer. What has been generally lacking, however, is an attempt to build a coherent theory of what dynamic information is used for.

A primary difficulty is that much of the data describing effects of object motion on object recognition does not really require appearance dynamics to be incorporated into static models. Instead, there are three broad categories of work: 1) Results that indicate object motion can be used as an independent feature, 2) "Representational momentum" studies that indicate appearance is automatically predicted online, and 3) Results indicating that joint encoding of form and dynamics may be relevant for the recognition of dynamic objects. While intriguing in their own right, none of these three main bodies of research clearly set forth a theory of how object motion contributes to the learning of object form. Below, I discuss each of these main bodies of work in turn and describe the strengths and limitations of the results.

Object motion can be a feature for recognition

The most compelling evidence that a good theory of object recognition should incorporate object motion in some manner is that the motion of an object can be a feature for recognition in its own right. Indeed, human observers are capable of recovering substantial information from stimuli that are almost purely defined by motion. The most striking example of this is the perception of "point-light" walkers, which are images of the human body composed of small, sparse dots usually placed at the joints of the body (Johansson, 1973). While the static position of the dots generally does not provide sufficient information for recognition at any level, once the stimulus is set in motion the percept of a human body is immediate and irresistible. The gender, mood, and even identity of the walker are also readily obtainable from the dynamic stimulus (Cutting, 1987; Kozlowski & Cutting, 1977). When spatial factors diagnostic of category (such as the ratio of shoulder width to hip width as a diagnostic feature for gender) are put in conflict with dynamic features (such as a feminine or masculine gait), the dynamic features typically govern the resulting percept (Thornton, Vuong, & Bulthoff, 2003). Furthermore, recognition judgments with point-light walkers are robust to sophisticated randomization of the spatial position of the constituent dots, a procedure that ensures that neither local motion signals nor spatial form can be used for object recovery (Beintema & Lappe, 2002). Instead, longer-term integration of form over time seems necessary to explain perception under these circumstances. In other cases where spatial form is impoverished, the use of motion information for recognition is apparent. Superimposition of an idiosyncratic motion field on an average face supports identification of the individual who generated the motion, for example (Knappmeyer, Thornton, & Bulthoff, 2003).

These results (and many more) demonstrate that dynamic information is available for recognition and put to use by human observers. Clearly object recognition must be built on some representations of object motion, indicating that a "bag of images" model of either learning or recognition is lacking data that is behaviorally relevant. As an argument for the necessity of a place for motion in object representations, results such as these are unimpeachable.

That said, how do results like these place constraints on the acquisition of object concepts? If object motion is just an independent feature that can be appealed to when static cues are unavailable or non-diagnostic, we are left with no way to understand how learning either set of cues might impact the other. Also, how would motion-based features ever apply to purely static recognition? Naively, they would not. In the absence of dynamic test stimuli, learned dynamic features are simply inapplicable. The flow fields associated with object motion may serve as a useful additional feature for recognition, but aren't particularly useful for determining form. However, the fact that dynamic features could provide observers with the correct label for an object when static form is impoverished suggests a means for associating degraded images with the correct object concept. An "equivalence class" between noisy, blurry, or occluded images of objects and their clearly-viewed counterparts could potentially be built and applied to purely static input using motion cues as a teaching signal. The extent to which this might occur in human object learning has not been directly tested, however.

Ultimately, given the question "What is learned from observation of a dynamic object?" the results of these studies would lead us to answer with something like "Labeled sets of 2-D appearances and flow fields." How either set of features is acquired or encoded is left essentially unconstrained.

Object motion and form may be encoded jointly

There is reason to believe that the human visual system does something more than acquire and maintain separate dynamic and static features for recognition. Specifically, there is some evidence that object motion and form may be encoded jointly.

An interesting result that provided some initial support for this claim was the finding that the dynamics of object production could influence static recognition (Freyd, 1987). In a test of recognition for handwritten characters, observers who learned how various written characters were drawn demonstrated heightened tolerance during a test phase for distortions that were consistent with variability in the underlying motor plan for drawing. Errors that could not have easily resulted from simple deviations in the motor plan they had observed previously were accepted less readily. This result is intriguing, suggesting that some form of dynamic information can influence the recognition (and possibly encoding) of static stimuli, but it does rely on a generative process that is known to the observer. In ordinary object perception, it is rarely the case that generative processes like this exist in a meaningful way, or are known to the observer.

More recent work however, has provided further evidence that object form and object motion are not coded independently. Instead, "spatiotemporal signatures" of object appearance may be formed by combining observed static appearance with observed appearance dynamics. The basis for this hypothesis is the finding that the recognition of rigidly rotating objects is impaired when the direction of rotation at test is different from the direction observed during learning. These results are most evident for objects whose appearance is obscured by fog or sparsely represented by dots (Stone, 1998; Stone, 1999; Vuong & Tarr, 2004), but object-centered and observer-centered motion of an object can affect recognition even for distinct, clearly visible stimuli (Newell, Wallraven, & Huber, 2004).

Finally, there are also several interesting results demonstrating that object form can have profound effects on perceived motion. Sinha and Poggio (Sinha & Poggio, 1996) showed, for example, that learning to associate a particular 3-D form with an ambiguous 2-D image could bias the perceived motion of novel objects. Specifically, after training observers with rigidly rotating "paperclip" objects, it was found that novel objects with a mean-angle projection matching a learned form were perceived as non-rigidly deforming once set in motion. The conclusion was that exposure to a particular image/form relationship during training set up a situation where expectation was violated during viewing of the novel object, leading to the anomalous percept. In a study of apparent motion using the human body, Shiffrar demonstrated that knowledge of biomechanical constraints on human movement substantially biased perceived motion paths. At slow alternation rates between two frames depicting arm movement, longer motion paths

were perceived in cases where the shortest path violated solidity constraints or constraints on joint flexion (Shiffrar & Freyd, 1990). At faster rates, this bias was eliminated in favor of the shortest "impossible" paths. In both sets of experiments, it is clear that the perception of form can have strong influence over perceived motion. Broadly speaking, these results also suggest some sort of joint encoding for dynamic objects.

But what exactly does this kind of work say about the encoding of dynamic objects? For example, what would be a good model of a "spatiotemporal signature?" Arguably, a spatiotemporal signature is akin to the visual system remembering an entire spatiotemporal volume of appearance and putting an object label on it. While not necessarily the most interesting model of dynamic object encoding, it does at least require that form and motion be integrated. Static appearance and flow fields are both obtainable from this volume, of course, making it easy to reconcile this model with the results suggesting that motion can be used for recognition when form is absent or irrelevant. It may also account for some of the data regarding the influence of form on motion, insofar as matching form may imply a matching flowfield. Perhaps when this assumption proves false there is some dissonance that leads to the percepts described above. However, there are reasons to think this is a poor model of object recognition. In particular, it posits that all visual experience with an object is recorded in its entirety. This would require prohibitive amounts of storage, and also leads to indexing problems if individual static images need to be recognized.

I continue by considering one final aspect of dynamic object percept that offers still another set of results demonstrating that static appearance and appearance dynamics interact in interesting ways.

Human observers predict future appearance

Finally, perhaps the most intensively studied aspect of dynamic object perception is the fact that observers appear to formulate predictions concerning the future appearance of objects and scenes. An effect of this kind was first reported by Freyd and Finke in 1983, when they noted systematic impairments for detecting "forward" differences between pairs of images depicting a complex event relative to detecting the same image differences in reversed order (Freyd, 1983). The rated quality of apparent motion in reverse order was also poorer than that of images placed in natural temporal order. The explanation offered for both of these results was that human observers automatically make predictions from images depicting "frozen motion," and that those predictions can systematically interfere with discrimination and motion perception. Beyond the many behavioral studies that followed this initial report, there is also recent neural support for this claim. Kourtzi and Kanwisher reported that passive viewing of "frozen" motion elicited activity from area MT, a cortical area responsible for motion processing (Kourtzi & Kanwisher, 2000).

A related phenomenon that has been investigated by many different groups is the finding that there appears to be "representational momentum" for dynamic objects (Freyd & Finke, 1984). That is, after observing a moving object, observers have an automatic tendency to continue updating object appearance after presentation has ended, as though object perception has inertia and cannot "stop" immediately. The term

"representational momentum" was coined to liken the basic perceptual effects to physical momentum, and has been widely adopted. It is relatively easy to see connections between reports of RM and phenomena like the "flash-lag" effect (Nijhawan, 1994), though it seems these parallels have only been explored by a few researchers thus far (Musseler, Stork, & Kerzel, 2002).

What do the observed effects of appearance prediction and RM on object recognition contribute to a model of dynamic object encoding? Like "spatiotemporal signatures" these effects require form and motion to be jointly encoded. Observed object motion is used to update object appearance, requiring a direct interaction between the two rather than purely independent features. In fact, appearance prediction potentially implies an even tighter connection between object appearance and motion than the spatiotemporal signature model would suggest. This is most evident when we ask what happens under each proposed model when a static image is encountered. If each dynamic object is encoded as a full spatiotemporal volume (or "signature") recognizing a static image of an object is merely an indexing problem. The observer simply has to find a slice of some object volume that matches well and report the label that goes with it. It is not clear in this scenario if the temporal direction of the volume is even relevant. However, under a predictive model, an observer who is given a static image would immediately and automatically generate the next appearance of the object. This suggests that the motion information associated with an image *must* be and is used to generate a new form. This is a much stronger claim than merely proposing that the temporal information is present in a volumetric representation of appearance over space and time. Prediction implies that motion always comes along for the ride.

However, though the influence of prediction over the short and long-term in object perception is intriguing, there are a few caveats that need to be brought up. First of all, it is not even clear that RM exists in they way it was initially described. Obtaining RM appears to be highly contingent on a number of factors that are seemingly unrelated to object perception in an interesting way, for example (Kerzel, 2002). If RM is an artifact of task-related confounds rather than automatic object perception mechanisms, it would be a bad idea to build it too deeply into our model of how dynamic objects are stored and recognized. Second, it is hard to tell in many cases whether the effects of prediction on object and scene perception are cognitive or perceptual. For example, when observers are predicting appearance in a scene based on their understanding of "gravity" or "support," is that something that needs to be built into a model of perception (Nagai, Kazai, & Yagi, 2002; Vinson & Reed, 2002)? This is not to say that cognitive processes cannot influence perception, but it is not my goal in this thesis to examine or model those complex relationships. Though interesting, these interactions do not strike me as core principles of visual learning. Finally, it is far from clear what the relationship is between short-term RM and long-term prediction. Do prediction effects in static images arise from "chronic" RM, for example? How does learned prediction affect momentum in the opposite direction? These are not impossible questions to answer, but they will need to be addressed before prediction can be built into a model of dynamic object encoding successfully.

The need for a coherent theory

I submit that the three previous sets of results do not provide a firm basis for developing a theory of dynamic object encoding that encompasses the recognition of static images. At present, these results constitute a loose confederation of data that successfully demonstrates that observed object motion can affect or supplant the recognition of static and dynamic object form. Unfortunately, a satisfying account of how object motion might contribute to object representation has not been offered. Instead, researchers have generally shown only that there are some reciprocal interactions between motion and form, rather than offering a mechanistic account of how those effects arise. Worse, it is difficult to extend the ideas behind many of these results, as they generally do not offer a framework in which parametric investigation is possible or relevant.

I argue that a heretofore less examined aspect of dynamic object perception, temporal association between images, provides an extremely valuable framework for examining the nature of dynamic object encoding. The basic temporal association makes several easily testable predictions, only some of which have been tested thoroughly. More importantly, the hypothesis also lends itself nicely to parametric investigation and computational modeling. Finally, there are many intriguing questions that follow from a basic temporal association hypothesis that, while not vital to the basic proposal, lead us in very interesting directions that are still highly relevant to dynamic object encoding. I will continue by laying out the basis temporal association hypothesis and discussing what advantages it offers as a framework for studying dynamic objects.

Temporal association and dynamic object perception

The basic principle underlying a temporal association model of object learning is that images that appear close together in time likely have the same physical cause. Assuming this to be the case, one can learn patterns of generalization over distinct static images that are likely to be valid for later recognition. Simply put, images that are close in time should be close in the underlying representation. This is really not much different from other smoothness arguments (Poggio, Torre, & Koch, 1985), except that it singles out time as a privileged dimension. The assumption of temporal smoothness has been found to be very useful for learning low-level tasks like the recovery of stereo information and translation invariance for simple line segments (Foldiak, 1991; Stone & Harper, 1999).

Applied to object recognition, the idea is that one learns to treat different images of the same physical object similarly by observing them close together in time. This proposal provides a unique response to the question "How are dynamic objects encoded?" I suggest that the answer offered by a temporal association model is something like, "Object motion induces a pattern of generalization over static images that primarily encompasses only a set of images observed close together in time during visual experience." This is a bit more long-winded than the answers provided by some of the previous models, but that's mostly because this statement has more interesting content than the others.

Both behaviorally and computationally, there is already compelling evidence that temporal proximity is used by the primate visual system to learn associations between

distinct images. Computationally, implementations of simple "trace rules" that augment artificial neural networks with a decaying memory term for recently viewed items have demonstrated that useful invariant features can be recovered for complex objects with straightforward update rules (Lecun, Bottou, Bengio, & Haffner, 1998; Serre, 2006; Wallis, 1996; Wallis, 1998). Psychophysically, we already know that human adults and infants recover temporal statistics from arbitrary sequences (Fiser & Aslin, 2002; Kirkham, Slemmer, & Johnson, 2002). Beyond learning temporal contingencies for abstract sequences of shapes, it has also been demonstrated that recognition performance for complex objects like faces and "greebles" (Gauthier & Tarr, 1997), can be systematically impaired by manipulating the temporal relationships between distinct images (Cox, Meier, Oertelt, & DiCarlo, 2005; Wallis & Bulthoff, 2001). Finally, there is also some evidence that high levels of the primate visual system may encode temporal contingencies between arbitrary images. Following passive training with random fractal patterns Miyashita et al. found that some cells in inferotemporal cortex displayed responses consistent with a learned association between temporally neighboring images (Miyashita, 1988; Miyashita, 1993; Miyashita & Chang, 1988). Responses in other areas of primate cortex, like the superior temporal sulcus, can be modulated substantially by the immediate history of stimulation (Jellema & Perrett, 2003) suggesting that "actions" might be encoded in these regions.

A representation of object appearance that is based on dynamic experience in this way is nice for several reasons. First, it is based on an assumption that has obvious ecological validity. Second, it is primarily a model of static object recognition that uses time (and thus motion) to learn how to generalize properly. This makes it a nice bridge between classic models of static object recognition and recent data concerning the effects of dynamics on perception. Third, the proposal highlights the importance of visual experience, leading to many testable predictions and open questions that may allow the basic model to refined and extended. Given these advantages, I have chosen to use the basic temporal association hypothesis as a jumping-off point for investigating the perception and recognition of dynamic objects.

Organization of the thesis

This thesis is divided into two main sections.

The first describes a series of studies designed to elucidate the influence of spatiotemporal continuity on object perception. In all of these studies, I build on the basic temporal association hypothesis to arrive at a model for how appearance codes are altered via experience with a dynamic object. Throughout, I emphasize the use of image-level discrimination tasks over procedures which require subjects to retrieve high-level information such as object class or identity. The influence of dynamic input on generalization over distinct images, sensitivity to small appearance changes, appearance prediction, and category learning and recognition are discussed. The main conclusion of this first section is that dynamic object encoding in the human visual system is best explained by a population code for appearance that is modulated in distinct ways by information that can be used for predicting future appearance and information that does not contribute to the formation of new predictions.

In Section 2, I describe two experiments which address the issue of whether or not privileged or "canonical" views of an object (Cutzu & Edelman, 1994; Palmer, Rosch, & Chase, 1981) are recovered during viewing of a dynamic object. I refer to such views as "keyframes" throughout this section, in reference to the animation technique of first drawing only a sparse set of images to represent a full motion sequence, filling it in later with the full set of images needed to achieve smooth animation. I first ask whether or not there is psychophysical evidence that human observers do recover "keyframes" from short exposure to a dynamic object. Then, I examine how spatial and temporal factors affect the fidelity of encoding for particular views of a dynamic object. In terms of the larger model of dynamic object encoding, this section provides insight as to how a population code is initially constructed for a given object. That is, how would the preferred views for units within a population code for a particular object initially be determined? Even assuming that the appearance space viewed needs to be well-covered by units within the population, it is far from clear how the preferred views of those units would be determined.

Each section contains several distinct chapters, which introduce the particular theoretical and experimental issues under consideration therein. While there are also brief summaries between chapters to motivate the progression of ideas throughout, the past literature relevant to each experiment is only discussed in depth in the introduction to each chapter.

References

Agarwal, S., & Roth, D. (2002). Learning a Sparse Representation for Object Detection. *Proceedings of ECCV.*

Beintema, J. A., & Lappe, M. (2002). Perception of biological motion without local image motion. *Proceedings of the National Academy of Sciences, 99*(8), 5661-5663.

Biederman, I. (1987). Recognition-by-Components: A Theory of Human Image Understanding. *Psychological Review, 94*(2), 115-147.

Brady, M. J., & Kersten, D. (2003). Bootstrapped learning of novel objects. *Journal of Vision, 3*, 413-422.

Cox, D. D., Meier, P., Oertelt, N., & DiCarlo, J. J. (2005). 'Breaking' position-invariant object recognition. *Nature Neuroscience, 8*(9), 1145-1147.

Cutting, J. E. (1987). Perception and Information. *Annual Review of Psychology, 38*, 61-90.

Cutzu, F., & Edelman, S. (1994). Canonical Views in Object Representation and Recognition. *Vision Research, 34*(22), 3037-3056.

Dong, D. W., & Atick, J. J. (1995). Statistics of natural time-varying images. *Network, 6*, 345-358.

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

Fei-Fei, L., Fergus, R., & Perona, P. (2004). Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *IEEE CVPR Workshop of Generative Model Based Vision.*

Fergus, R., Perona, P., & Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. *Proceedings of CVPR.*

Fiser, J., & Aslin, R. N. (2002). Statistical Learning of Higher-Order Temporal Structure From Visual Shape Sequences. *Journal of Experimental Psychology: Learning, Memory and Cognition, 28*(3), 458-467.
Foldiak, P. (1991). Learning Invariance from Transformation Sequences. *Neural Computation, 3*, 194-200.
Freyd, J. J. (1983). The mental representation of movement when static stimuli are viewed. *Perception and Psychophysics, 33*(6), 575-581.
Freyd, J. J. (1987). Dynamic Mental Representations. *Psychological Review, 94*(4), 427-438.
Freyd, J. J., & Finke, R. A. (1984). Representational Momentum. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(1), 126-132.
Gauthier, I., & Tarr, M. J. (1997). Becoming a "Greeble" expert: Exploring the face recognition mechanism. *Vision Research, 37*(12), 1673-1682.
Jellema, T., & Perrett, D. I. (2003). Perceptual History Influences Neural Responses to Face and Body Postures. *Journal of Cognitive Neuroscience, 15*(7), 961-971.
Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics, 1*, 201-211.
Kerzel, D. (2002). A matter of design: No representational momentum without predictability. *Visual Cognition, 9*(1-2), 66-80.
Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition, 83*, B35-B42.
Knappmeyer, B., Thornton, I. M., & Bulthoff, H. H. (2003). The use of facial motion and facial form during the processing of identity. *Vision Research, 43*, 1921-36.
Kourtzi, K., & Kanwisher, N. (2000). Activation Human MT/MST by Static Images with Implied Motion. *Journal of Cognitive Neuroscience, 12*(1), 48-55.
Kozlowski, L. T., & Cutting, J. E. (1977). Recognizing the sex of a walker from a dynamic point-light display. *Perception & Psychophysics, 21*, 575-580.
Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278-2324.
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.
Murase, H., & Nayar, S. K. (1995). Visual Learning and Recognition of 3-D Objects from Appearance. *International Journal of Computer Vision, 14*, 5-24.
Musseler, J., Stork, S., & Kerzel, D. (2002). Comparing mislocalizations with moving stimuli: The Frohlich effect, the flash-lag, and representational momentum. *Visual Cognition, 9*(1-2), 120-138.
Nagai, M., Kazai, K., & Yagi, A. (2002). Larger forward memory displacement in the direction of gravity. *Visual Cognition, 9*(1-2), 28-40.
Newell, F. N., Wallraven, C., & Huber, S. (2004). The role of characteristic motion in object categorization. *Journal of Vision, 4*(2), 118-129.
Nijhawan, R. (1994). Motion extrapolation in catching. *Nature, 370*, 256-257.

Olshausen, B. A. (2003). *Learning Sparse, Overcomplete Representations of Time-Varying Natural Images.* Paper presented at the IEEE International Conference on Image Processing, Barcelona, Spain.

Palmer, S., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. In J. Long & A. Baddeley (Eds.), *Attention and Performance IX* (pp. 135-151). Hillsdale, NJ: Lawrence Erlbaum.

Poggio, T., Torre, V., & Koch, C. (1985). Computational vision and regularization theory. *Nature, 317*, 314-319.

Prodohl, C., Wurtz, R. P., & von der Malsburg, C. (2003). Learning the Gestalt Rule of Collinearity from Object Motion. *Neural Computation, 15*, 1865-1896.

Serre, T. (2006). *Learning a Dictionary of Shape-Components in Visual Cortex: Comparison with Neurons, Humans and Machines.* , MIT, Cambridge, MA.

Shiffrar, M., & Freyd, J. (1990). Apparent motion of the human body. *Psychological Science, 1*, 257-264.

Sinha, P., & Poggio, T. (1996). The role of learning in 3-D form perception. *Nature, 384*, 460-463.

Stone, J. V. (1998). Object recognition using spatiotemporal signatures. *Vision Research, 38*, 947-951.

Stone, J. V. (1999). Object recognition: view-specificity and motion-specificity. *Vision Research, 39*, 4032-4044.

Stone, J. V., & Harper, N. (1999). Temporal constraints on visual learning: a computational model. *Perception, 28*, 1089-1104.

Tarr, M., & Bulthoff, H. H. (1995). Is human object recognition better described by geon-structural-descriptions or by multiple views? *Journal of Experimental Psychology: Human Perception and Performance, 21*(6), 1494-1505.

Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science, 290*(5500), 2319-2323.

Thornton, I. M., Vuong, Q. C., & Bulthoff, H. H. (2003). A chimeric point-light walker. *Perception, 32*(3), 377-383.

Ullman, S. (1979). The interpretation of structure from motion. *Proceedings of the Royal Society of London, Series B, 203*, 405-426.

van Hateran, J. H., & Ruderman, D. L. (1998). Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proceedings of the Royal Society of London, Series B, 265*, 2315-2320.

Vinson, N. G., & Reed, C. L. (2002). Sources of object-specific effects in representational momentum. *Visual Cognition, 9*(1-2), 41-65.

Vuong, Q. C., & Tarr, M. J. (2004). Rotation direction affects object recognition. *Vision Research, 44*(14), 1717-1730.

Wallis, G. (1996). Using Spatio-temporal Correlations to Learn Invariant Object Recognition. *Neural Networks, 9*(9), 1513-1519.

Wallis, G. (1998). Spatio-temporal influences at the neural level of object recognition. *Neural Networks, 9*, 265-278.

Wallis, G., & Bulthoff, H. H. (2001). Effects of temporal association on recognition memory. *Proceedings of the National Academy of Sciences, 98*(8), 4800-4804.

Weber, M., Welling, M., & Perona, P. (2000). Unsupervised learning of models for recognition. *Proceedings of ECCV*, 18-32.

Section 1 – Effects of dynamic input on static form generalization, discrimination, prediction, and recognition

In the first section of the thesis, I describe a series of results that place important constraints on the way in which dynamic experience affects the representation of static form. Crucially, in
contrast to previous studies, I generally avoid asking observers for judgments concerning object identity or category since these decisions likely involve complex cognitive mechanisms. Instead, I rely heavily on tasks that require observers to make only image-level judgments of object form.

The advantage of this is two-fold: First, these tasks are very simple for participants to understand, and are thus free of some of the pragmatic difficulties that make previous results in this domain difficult to interpret. Second, since these tasks require only low-level analysis of form (which could in principle be solved using trivial measures such as L2 distance in pixel space) we can be more confident that the effects we obtain have a perceptual basis rather than a cognitive one. The resulting data provides a clear picture of how both category and coordinate judgments within an appearance space are affected by coherent object motion.

Chapter 1 describes a set of experiments in which I examine the basic temporal association
hypothesis using a set of novel, non-rigid stimuli. In contrast to previous work, I examine the
effects of temporal association on both generalization over distinct images and sensitivity to the differences between those same images. By considering how both processes are influenced by the temporal proximity between distinct images, we are able to expand the scope of the basic temporal association hypothesis beyond the singular goal of learning invariance. The results indicate that both generalization and discrimination ability increase following exposure to a coherent dynamic object. I argue that this places an important constraint on how appearance must be encoded, specifically that a coarse population code for appearance is implied given the observed results.

In Chapter 2, I use the perception of human locomotion as a test case for how recent dynamic
experience with an object interacts with prior knowledge concerning the appearance dynamics of a familiar object. First, I characterize how static discriminability between images is affected by the temporal order of stimuli in the absence of dynamic exposure. I find that "forward" discrimination of static images of a walking human is more difficult than discrimination carried out with the same images in reverse temporal order. This basic finding is consistent with previous reports concerning prediction and its effects on image discrimination and apparent motion. In a second experiment, following exposure to a locomoting human figure that is walking either forward or backward, I re-assess sensitivity to "forward" and "backward" image differences. I find that the initial asymmetry favoring backward discrimination is removed by both dynamic stimuli, but that observation of forward locomotion also provides an initial benefit for both discriminations. I suggest that my initial model of a coarse appearance code following

dynamic training must be modified to include a predictive component, the stability of which determines the extent to which fine-grained sensitivity is enhanced following exposure to a dynamic object.

Finally, in Chapter 3, I examine how categorization of static images is affected by the observation of dynamic objects during training. Specifically, I ask whether or not the diagnosticity of object motion (as defined by the path taken through an appearance space) affects static classification when a fixed set of static images are used for defining category membership. Using the same stimulus set employed in Chapter 1, I determine that diagnostic motion during training can actually impair static classification. This result forces us to rule out any account of dynamic object representation positing an ideal observer who retains the full set of observed appearances. Furthermore, in another manipulation I find that the direction of a path through appearance space is not sufficient to induce the impaired classification abilities I find in my first experiment. This supports time-symmetric generalization over appearance, with prediction playing relatively little role in classification.

Introduction to Chapter 1

In this set of studies, I examine the effect of observed object motion on generalization and sensitivity over the set of images contained in the dynamic stimulus. Throughout, behavior is described using image-level judgments, reducing the influence of high-level cognitive mechanisms on response selection. In all of the experiments reported in Chapter 1, observers are exposed to novel, non-rigid, three-dimensional stimuli. This provides the advantage of further limiting the influence of top-down mechanisms since observers are unfamiliar with both the forms and motions they observe during the experiment.

Learning about dynamic objects: increases in generalization and sensitivity

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 on object perception. First, we use an implicit priming criterion to demonstrate 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 discuss the possibility that this seemingly conflicting set of results, one blurring and the other sharpening the perceived distinction between temporally proximate frames, can arise from a model of object representation in which temporal association leads to coarse coding across a population of object units.

Introduction

Object recognition is a computationally difficult task for one primary reason: 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 shape is being viewed (Moghaddam, Jebara, & Pentland, 2000; Moses, Adini, & Ullman, 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, Bottou, Bengio, & Haffner, 1998; Riesenhuber & Poggio, 1999; Ullman, Vidal-Naquet, & Sali, 2002; Weber, Welling, & Perona, 2000), but most of these systems require some form of implicit label, and generalization across large transformations is not very robust. 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 sensitivity 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 & Harper, 1999; Ullman & Bart, 2004; Wallis, 1996; Wallis, 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.

Indeed, a variety of methodologies have provided evidence that this sort of behavior emerges after training with image sequences. For example, neurons in the primate infero-temporal 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; Miyashita, 1993; Miyashita & 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 & Bulthoff, 2001), and the learned sequence of 2D views can impair recovery of 3-D form via the kinetic depth effect (Sinha & Poggio, 1996). Even simple translation invariance can be "broken" by presenting two different objects within a small temporal window (Cox, Meier, Oertelt, & DiCarlo, 2005).

Clearly temporal association can play a pivotal role in object and face recognition (O'toole, Roark, & Abdi, 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. This is usually demonstrated by pairing images that would usually *not* appear close together in time in a natural setting, and demonstrating that subsequent within-pair discrimination is impaired in some way. 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. 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. Learning about an object through temporal association of neighboring images should not impair the ability to discriminate them at the image level. If anything, we argue that such observation should improve one's ability to perform image level discrimination. Having the opportunity to observe a change in appearance over time should alert one 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 sensitivity.

In the current study, we ask whether or not there is behavioral evidence that temporal association can 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 training sequences (approximately 10 minutes) we find that adult observers do in fact display exactly this pattern of behavior. Neighboring images begin 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 basic idea of temporal "smearing" of appearance, we find that observers become more sensitive to appearance changes they have observed in a dynamic sequence. We suggest that the development of a coarse population code for object appearance can explain both of these results. The proposed model makes a clear prediction regarding how neural tuning for objects may change at high levels of the primate visual system as a function of temporal association between images.

Methods

To ensure that observers could not apply previous knowledge concerning how object appearance changes after depth rotation or under varying illumination, we used novel 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 that rely on the construction of a static 3-D object model. Since there is no "ground truth" 3-D form, it is more likely that any effects we observe are the result of temporal linkages between images rather than the augmentation of a 3-D model via structure-from-motion cues.

These objects (which we will refer to as "blobs") are constructed from two spherical harmonics that can be independently rotated through various phase angles (Nederhouser, Mangini, & Biederman, 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 and Supplementary movie files). In both our experiments, training sequences present each observer with objects that change appearance via motion along only one of these directions through appearance space. Following this training period, subjects in our first experiment perform an "instant priming" task (Sekuler & Palmer, 1992) that allows us to determine if image matching can be primed by cue images that are temporal neighbors to the targets. This paradigm has been successfully used by Kourtzi and Shiffrar to probe the representational content of rigid and non-rigid objects undergoing apparent motion (Kourtzi & Nakayama, 2002; Kourtzi & Shiffrar, 1997; Kourtzi & Shiffrar, 1999; Kourtzi & Shiffrar, 2001). To ensure that temporal association (rather than pure spatial similarity combined with repeated exposure) induces generalization, we compare the effects of smooth training sequences to scrambled training sequences across subject groups. In our second experiment, we measure discriminability along both the "horizontal" and "vertical" directions in appearance space both before and after exposure to the training sequences depicting appearance change along only one axis. A simple change detection task is used to determine subjects' sensitivity to image differences in both trained and untrained directions through image space. In both tasks, we determine the

effect of dynamic training on static recognition tasks that do not require the recovery of object labels so as to minimize non-visual cognitive interference.

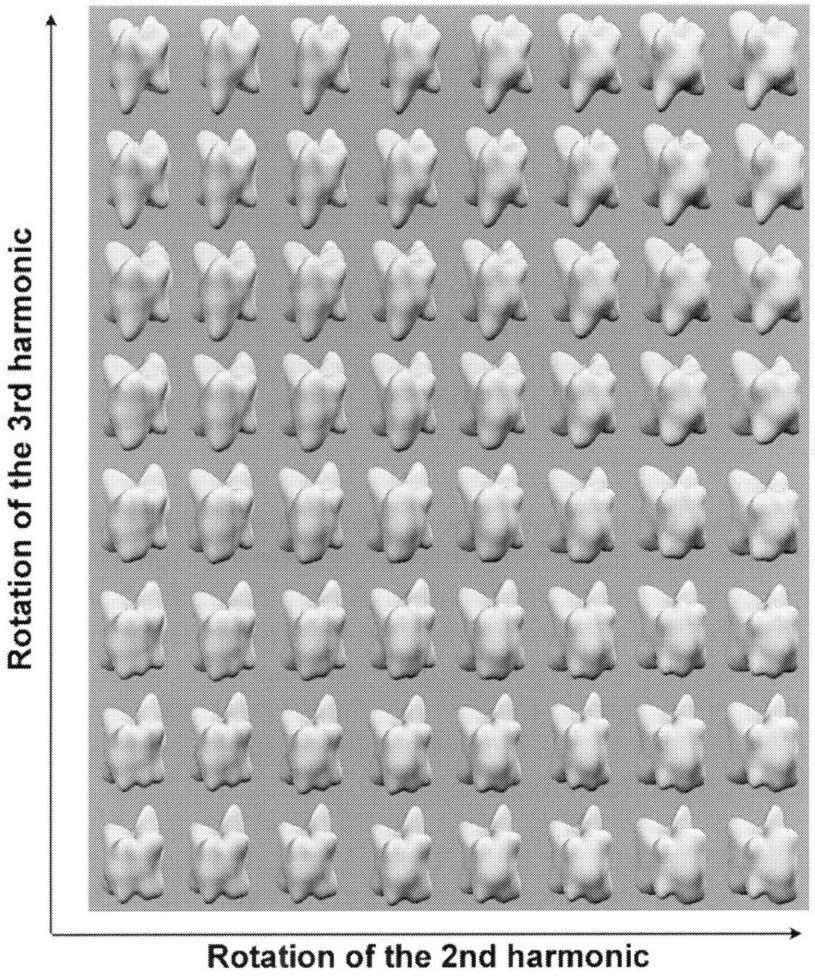


Fig. 1 – An 8x8 space of "blob" stimuli. The horizontal and vertical axes of this space are defined by the phase angle of the 2nd and 3rd harmonic. Movement along each axis induces non-rigid motion that is distinct from that generated by movement along the other axis (see Supplementary movie files). This image represents a schematic view of the full 16x16 blob space used in our experiments. This smaller version has been included for ease of viewing.

Experiment 1 – Priming task

Subjects

24 volunteers from the MIT community participated in this experiment, all between the ages of 18-35. All participants reported normal or corrected-to-normal acuity, and were compensated for their participation. Observed object motion was a between-subjects variable, such that, half of the observers were randomly assigned to the "smooth motion" group, and the remaining half were assigned to the "scrambled motion" group.

Stimuli

We do not include details of the rendering process used to generate the "blob" stimuli given the amplitudes and phase angles of the constituent harmonics, but instead refer

the reader to Nederhouser et al.'s initial report of the stimulus construction process for more details (Nederhouser et al., 2002). An important aspect of the stimulus space for our purposes however, is that the appearance space is scaled relative to a Gabor-jet based image similarity metric such that city-block distance is a meaningful measure of low-level similarity along both axes. The particular similarity metric used has been found to correlate very well with human similarity judgments as well, giving us reason to believe the appearance space is well-scaled for perceptual similarity. For the purposes of the priming experiment, one 16-image "strip" of blob images was used to generate dynamic objects and static test images for use in the "instant priming" task. These 16 images depicted rotation of the 2nd harmonic through an angle of 180 degrees with position of the 3rd harmonic fixed.

Our "smooth" motion sequence was generated by continuously displaying these static frames first in forward, then reverse order at a rate of 12 frames per second. "Scrambled" sequences were created by randomly shuffling frame order for each presentation of a dynamic stimulus to each subject. Additionally, though the exposure time for each frame was matched across smooth and scrambled movies, a 100ms empty frame was also drawn between each image frame in the scrambled movie. The result is that the scrambled movie consists of a flashed presentation of unordered blob frames.

Procedure

Each participant completed three rounds of the priming task, with each round consisting of a training period and a test period. During each round, subjects passively viewed the dynamic stimulus described above for a 3-minute period. No response was required during exposure to the motion sequence.

Following each training session, subjects performed a "go, no-go" image matching task using static images taken from the training sequence. Each trial began with the presentation of one of three cue images for 500ms. Cue images were drawn from the training sequence. After cue presentation, subjects viewed a blank screen for an additional 500ms, and then two target images were simultaneously presented for 3000ms. Subjects were instructed to press any key on the keyboard as fast possible if the two target images were identical, and to withhold their response if they were different. Target images were either identical to the cue image, separated from the cue image by 1, 2, or 3 units in blob appearance space, or images that were never presented during training. In the latter case, a "Control" blob was selected in which the 3rd harmonic was differently oriented from the blobs in the training sequence. Response times to correct "SAME" judgments were recorded. As no response was required for trials on which targets differed from one another, we do not present any data relating to these trials.

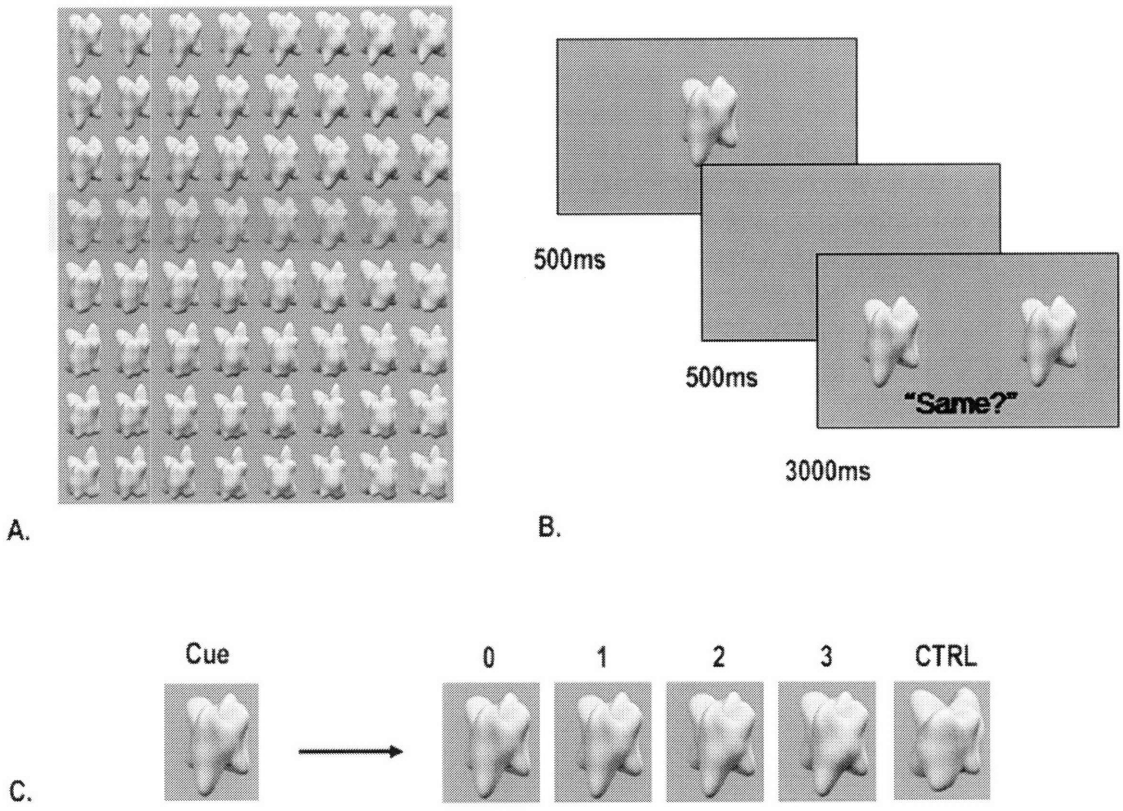


Fig. 2 – Each round in the "Instant Priming" paradigm used in Experiment 1 consists of a training period and a test period. In each training period a dynamic object defined by concatenating images along one "strip" in appearance space is presented to the subject for 3 minutes. (a) Images can be concatenated in their natural order ("smooth motion" group) or in a random order ("scrambled motion" group) Following each round of dynamic exposure, the test period consists of "go, no-go" trials in which a cue image is followed by two images that can either be identical to one another or not. (b) The RT cost for making "SAME" judgments with targets that do not match the cue is calculated for several cue/target distances in blob space (c). An unrelated image is also included as a control for task-specific learning unrelated to image association during training. Again, the depicted 8x8 space is a smaller version of the full blob space used in our tasks.

All stimuli were presented on a 19" Dell Ultrasharp monitor. Training and test stimuli all subtended approximately 2 degrees of visual angle. Dynamic stimuli and cue images were all presented at the center of the monitor, while target images in the test phase were presented at 3 degrees to the left and right of the monitor's center. 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).

Experiment 2 – Change detection

Subjects

An additional 10 volunteers from the MIT community participated in this experiment, again all between 18 and 35 years of age. All participants reported normal or corrected-to-normal acuity, and were compensated for their participation. Subject pools for Experiment 1 and 2 were mutually exclusive.

Stimuli

For this experiment, training and test stimuli were drawn from the space of blob objects described previously. In the appearance space defined by these two transformations, rotation of the 2nd harmonic corresponds to "x-axis" or "horizontal" movement through appearance space, while rotation of the 3rd harmonic corresponds to "y-axis" or "vertical" movement. Multiple image sequences were generated in this space in the manner described for Experiment 1, with one harmonic assuming a particular fixed position and the other free to rotate. Parallel "bands" of images spaced one image apart in our appearance space were always used to create dynamic stimuli (Figure 3).

Procedure

Each subject completed two rounds of the change detection task, one preceding exposure to dynamic stimuli and one following this training period.

During each test period, subjects performed a change detection task using pairs of images drawn from the blob appearance space. On each trial, one blob stimulus was presented for 250 ms, followed by a 200 ms blank period, a 200ms 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 degree interval around the center of the monitor. On each trial, the two blob stimuli presented could be identical, differ in the position of the 2nd harmonic (a "horizontal" difference in our appearance spacee), or differ in the position of the 3rd harmonic (a "vertical" difference). All pairs of "different" stimuli were separated by 2 frames in the appearance space defined previously. Of these, 64 contained images different along a "horizontal" appearance axis, and 64 contained images different along a "vertical" axis (see Figure 3).

Each pair was presented twice, once for each ordering of the stimuli within the pair. An additional 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 8 movies depicting "horizontal" movement through appearance space, while the other half saw 8 movies depicting "vertical" movement. Each individual sequence lasted approximately 30 seconds, and was displayed at a rate of 12 frames per second. Sequences were presented three times each, and the order of sequence presentation was randomized for each subject. The full training period lasted approximately 12 minutes.

In both test sessions, the ability to detect changes in the position of the two harmonics was characterized by calculating d' along each axis in appearance space. The change in sensitivity to image changes observed during the training period was compared to the change in sensitivity for image changes not observed in the dynamic stimulus. Response time was not recorded in this experiment.

Stimulus display parameters in this experiment are identical to those discussed for Experiment 1.

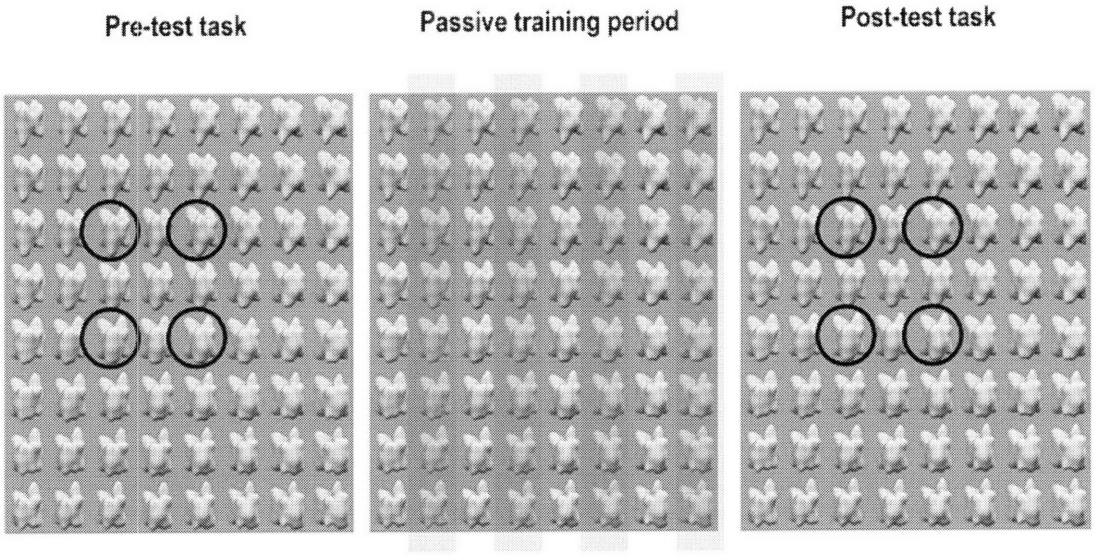


Fig. 3 – In our change detection paradigm, subjects first are tested for sensitivity to static image differences between "horizontal" and "vertical" image pairs in a sequential same/different task. Following this, subjects are exposed to training sequences depicting only one direction of movement through appearance space. Sensitivity to image changes along the trained and untrained axes is reassessed, and changes in sensitivity along each axis are recorded.

Results

Experiment 1 – Temporal association leads to increased generalization

Initially, we expect that cue images that are identical to the targets will be better primes for the same/different judgment than temporal neighbors. As exposure to the smooth motion sequence continues, however, we expect that the temporally associated images will increase in their efficacy to prime "SAME" responses. This same result should not obtain for observers who view the scrambled motion sequence, as there is no consistent temporal relationship between frames.

We compute for each subject the mean response time for "SAME" judgments carried out with each kind of cue/target pair. We then subtract the response time for "SAME" judgments cued by identical images (the *de facto* optimal cue) to yield the RT cost for each condition in which cues do not match targets. If temporal association does indeed induce generalization over neighboring images, we expect that an early positive cost for non-matching cues will give way to a reduced, possibly nil cost after training. This decrease in cost over multiple rounds should not occur following exposure to the scrambled stimulus. We display the results of this analysis for observers in the "smooth" and "scrambled" groups in Figure 4.

We do not analyze observers' accuracy data here as almost all of our participants were at ceiling throughout the task. We note, however, that data from one observer from the "smooth motion" group and one from the "scrambled motion" group were discarded due to extremely low accuracy, perhaps due to a misunderstanding of task requirements.

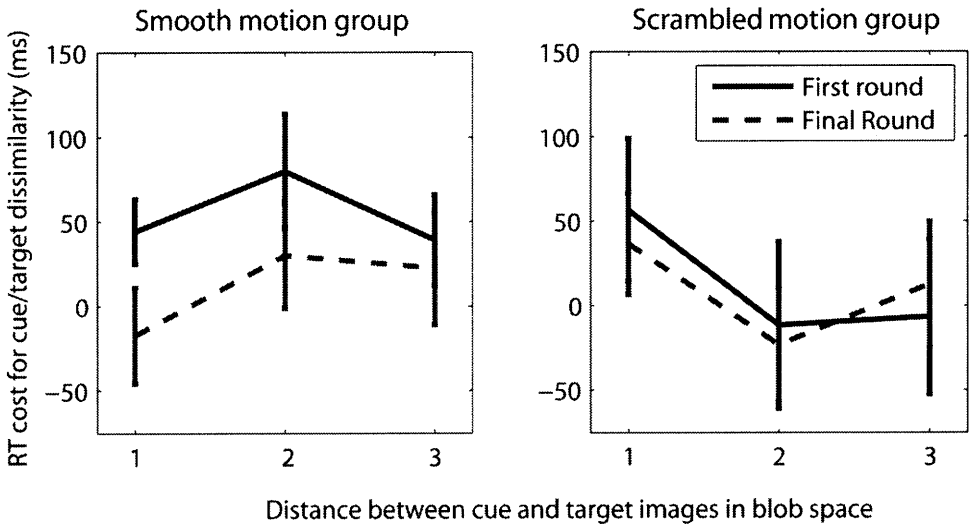


Fig. 4 –After three rounds of training, an initial RT cost for targets that do not match the cue disappears for observers in the "smooth motion" group. Conversely, observers in the "scrambled motion" group show no effects of multiple rounds of dynamic exposure. This difference in conditions is most evident for the smallest cue/target distance. Beyond this point, behavior is less consistent in both groups.

First, we consider the results from observers in the "smooth" motion group. We carried out a two-factor, repeated measures ANOVA over RT differences between identical target/cue pairs and each type of non-identical target/cue pair. Dissimilarity in blob space (1, 2, or 3 units) and training session (first or last) were our two factors. Our analysis reveals a significant effect of training session ($F(1,10) = 5.98$, $p < 0.05$) and a marginally significant effect of cue/target dissimilarity ($F(2,9) = 3.29$, $p = 0.058$). There was no significant interaction between the two factors ($F < 1$).
Second, we conduct the same analysis on the data obtained from observers in the "scrambled" motion group. In this case, we find no effects of training session ($F(1,10)=.009$, $p = 0.928$) or cue/target dissimilarity ($F(2,9)=2.06$, $p = 0.179$). The interaction between the two factors was also not significant. Note that dissimilarity in this case is not related to temporal factors in any way since the order of images in the motion sequence was completely randomized.
We also carry out two pre-planned comparisons on the data points obtained in both conditions from non-identical targets most similar to the cue images (those that are 80ms away in the smooth movie). If mere exposure to the images causes generalization over similar forms, we would expect significant changes in the RT cost over multiple rounds of training in both conditions. However, paired t-tests carried out in each condition yield a significant effect only when smooth motion was viewed during training ($p<0.05$) and not when scrambled motion was viewed ($p = .725$).

Finally, what about images that were never temporally associated with the test images at all? The difference in RT between optimal and unassociated cues does not change as a function of training in either the smooth motion condition (paired t-test, $p = 0.62$) or the scrambled condition ($p = 0.58$), indicating that repeated performance of the matching task is not enough to induce a change in RT differences.

Passive observation of a coherent motion sequence clearly induced a reduction in the priming advantage for the "optimal" prime relative to temporal neighbors of the targets.

The lack of any significant result from the scrambled condition further demonstrates that mere exposure to the static stimuli is not enough for generalization to occur; images must be temporal neighbors.

Experiment 2 – Temporal association causes increased sensitivity

In our second experiment, we determine how the ability to discriminate between subtly different blob images changes as a function of temporal association. Our participants were asked to perform a change detection task with sequentially presented blob images that could be close neighbors along the "horizontal" or "vertical" axes, or could be identical images. In this way, sensitivity along each axis in image space was measured by calculating d' separately for the two possible directions of image change. Following this initial task, observers passively viewed a series of stimuli depicting a blob changing its appearance via either "horizontal" or "vertical" oscillatory motion through appearance space. Afterwards, the change detection task was re-run and sensitivity along each axis in image space was calculated. A difference in d' for each direction of image change (trained or untrained) was then computed for each subject. We ask two questions:

1) Is there any effect of temporal association?
2) If there is an effect of temporal association, is it positive or negative?

We display the results of our analysis in Figure 5.

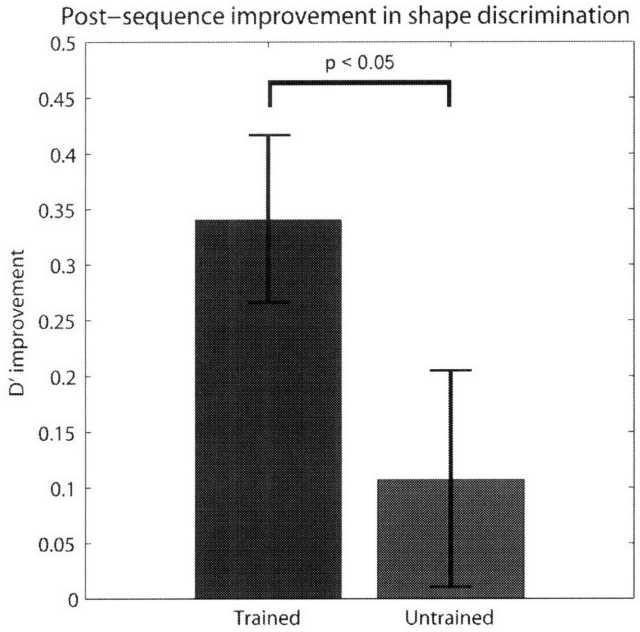


Fig. 5 – After observing object motion brought on by movement along one axis in appearance space, observers are better able to detect static image differences along that axis. Though there is improvement in both directions, there is significantly more improvement in the direction of observed motion.

In both the trained and untrained directions, observers became more sensitive on average to even very subtle image changes. However, the amount of improvement in the trained direction significantly exceeds that of the untrained direction (paired t-test, two-tailed, $p < 0.05$).

Passively viewing blob sequences appears to link distinct images together as evidenced by our "instant priming" task, but here we see that close temporal association improves the ability to detect differences between those same pairs of images in a change detection task. Taken together, results from experiments 1 and 2 suggest that observers display both heightened sensitivity between and increased generalization across frames of a smoothly varying dynamic sequence.

Discussion

We have found that the observation of changing object appearance in dynamic sequences increases both the generalization over temporally close images, and the sensitivity to differences between neighboring images. These results support the notion that object learning is a dual process of constructing invariants and learning to detect subtle variations in object appearance. The visual system learns about objects in such a way that the ultimate goal of a balance between good recognition and good discrimination is met. Our results also suggest that temporal association plays an important role in both aspects of this learning.

The use of an implicit priming task demonstrates that generalization to temporal neighbors is not just evident at the level of object label generation, but also for judgments at the image level. Moreover, by presenting cue-target pairs that were separated by varying temporal distance during training, we gain the ability to assess the strength of image binding over time. Looking at Figure 4, we see for example that the closest temporal neighbors (which are also the most similar) are more effective cues after training than more distant neighbors. Indeed, the slight (but non-significant) negative value of this data point is intriguing in that it suggests that ultimately, temporal neighbors of a stimulus may prove *more* effective primes than an identical image. This is consistent with well-established classical conditioning results. In general, associative learning is strongest when there is a temporal delay between the stimuli to be associated, as opposed to simultaneous presentation. Also, previous work with the flash-lag effect (Nijhawan, 1994) has raised the possibility that an 80-100 ms time window may be a critical interval over which the visual system should learn to make good predictions due to neural transmission limitations. Our work extends this idea to the domain of object recognition by suggesting that part of learning to recognize a new object is learning how appearance will change within a short interval. Finally, our comparison of smooth to scrambled presentation of object appearance 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, Rolls, & Stringer, 2006), our results indicate that temporal continuity is of primary importance for generalization.

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 should subsequently 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?

Several simple explanations can be ruled out based on our design. For example, it is unlikely that observers are using the training sequence to identify local image regions where change is expected. Comparing specific image regions across test stimuli is made difficult both by the smoothness of the blobs and the fact that we randomly jitter the position of both test images on each trial. Also, only attending to regions where change occurs during training would be detrimental to performance on untrained pairs. Given that we observe improvements in both trained and untrained pairs, is not likely that observers adopt this strategy. Similarly, the role of explicit prediction in performing our change detection task is made complicated by the fact that each of the images in our dynamic stimuli predicts *two* images equally well (due to the forward and backward oscillation along appearance axes during training). Since there is no unambiguous prediction to be made from any individual image, it is unlikely to play a major role here. Finally, mere exposure effects can be ruled out. Observers are only tested on images that do not appear in the training sequences.

We suggest that a useful framework for explaining these results, which show simultaneous increases in generalization and sensitivity, involves "coarse coding" of object appearance following dynamic experience. Within certain limits, populations of units with overlapping tuning functions in feature space can provide for superior generalization and better resolution than non-overlapping, highly localized units (Hinton, McClelland, & Rumelhart, 1986). Simulations of coarse or distributed coding have demonstrated that extremely good resolution for "coordinate" judgments can be achieved using redundant representations such as this (Jacobs, 1996; Jacobs & Kosslyn, 1994; Milner, 1974). It is easy to apply this same reasoning 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 initially have tuning curves that do not overlap substantially. Second, we assume that exposure to a dynamic stimulus causes each cell to widen its tuning function so that it responds to a wider range of object appearances. Crucially, widening must not occur symmetrically in feature space. Instead, tuning curves must widen more in the direction commensurate with stimulation. As the curves begin to overlap, a highly redundant code for object appearance emerges. So long as the feature space is not too dense and the tuning functions do not become too large (Hinton, 1986), better generalization ability and better resolution in this space will result from 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, Hietanen, Oram, & Benson, 1992) and novel objects (Logothetis, Pauls, & Poggio, 1995). 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 & Rolls, 1998). Psychophysical data from both humans and monkeys provides further evidence to support population coding for complex objects (Fang & He, 2005; Logothetis, Pauls, Bulthoff, & Poggio, 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 (Jellama & Perrett, 2003), the immediate effects of dynamic stimulation on appearance tuning for arbitrary objects have 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, or exposure to dynamic objects that change appearance very rapidly 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.

We close by pointing out that we have not attempted to incorporate object motion *per se* into our discussion of how temporal association may induce the effects we observe here. In natural settings, temporally neighboring images generally give rise to motion, meaning that some reciprocal encoding of object form and object dynamics may be an important factor in learning object representations. At present, however, we remain agnostic as to the nature of that interaction. Pure temporal association without motion has been demonstrated to give rise to generalization in IT (Miyashita et al., 1993), leading us to believe that it is useful to talk about such linkages between static images in the absence of real motion. Also, given that we only employed static images at test for both tasks, it seems appropriate to not incorporate object motion directly into our proposed model.

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. Based on these data, we have suggested a "coarse coding" model of object learning in which the observation of an object over time induces a distributed representation of object appearance. The model can account for both of our findings with only one proposed change in neural tuning as a function of temporal association. Taken together, these results provide hints about how dynamic input can affect the representation of static form. 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.

References

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. D., Meier, P., Oertelt, N., & DiCarlo, J. J. (2005). 'Breaking' position-invariant object recognition. *Nature Neuroscience, 8*(9), 1145-1147.

Fang, F., & He, S. (2005). Viewer-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 D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1,). Cambridge, MA: MIT Press.

Jacobs, R. A. (1996). Computational studies of the developmental of functionally specialized neural modules. *Trends in Cognitive Sciences, 3*(1), 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*(7), 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*(6), 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*(2), 335-355.

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278-2324.

Logothetis, N., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology, 5*(5), 552-563.

Logothetis, N. K., Pauls, J., Bulthoff, H. H., & Poggio, T. (1994). View-dependent object recognition by monkeys. *Current Biology, 4*(5), 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*(11), 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*, 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*(7), 745a.
Nijhawan, R. (1994). Motion extrapolation in catching. *Nature, 370*, 256-257.
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. *Phil. Trans. R. Soc. Lond. 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.
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*(9), 1513-1519.
Wallis, G. (1998). Spatio-temporal influences at the neural level of object recognition. *Neural Networks, 9*, 265-278.
Wallis, G., & Bulthoff, H. H. (2001). Effects of temporal association on recognition memory. *Proceedings of the National Academy of Sciences, 98*(8), 4800-4804.
Weber, M., Welling, M., & Perona, P. (2000). Unsupervised learning of models for recognition. *Proceedings of ECCV*, 18-32.

Introduction to Chapter 2

Having observed interesting and novel effects of observed object motion on image discrimination in Chapter 1, I continue by examining the effects of observed motion on sensitivity to appearance changes in a familiar object: the human body. Here, rather than avoid the possibility that top-down knowledge may effect perception in this setting, I explicitly examine the interaction between prior expectations of object movement and recent exposure to a particular dynamic stimulus. The human body is ideal for this purpose, since observers are familiar with the form and have temporally asymmetric experience with moving bodies insofar as backwards walking is rarely observed in a natural setting. Using the same change detection paradigm used in Chapter 1, the perceptual consequences of this long-term ecological knowledge are characterized both with and without exposure to a dynamic stimulus in a controlled laboratory setting.

Interactions between prior knowledge and recent experience in the perception of dynamic objects.

Abstract

Temporal association between dissimilar views of an object has been proposed as a tool for learning invariant representations for recognition. We examine heretofore untested aspects of the temporal association hypothesis using a familiar dynamic object, the human body. Specifically, we investigate how recent dynamic experience with an object interacts with long-term memory for expected patterns of movement. In our task, observers performed a change detection task using upright and inverted images of a walking body either with or without previous exposure to a motion stimulus depicting an upright walker. Observers who were exposed to the dynamic stimulus were further divided into two groups dependent on whether the observed motion depicted forward or backward walking. We find that the effect of the motion stimulus on sensitivity is highly dependent on whether the observed motion is consistent with past experience.

Introduction

The way an object moves can be an important aspect of its appearance in many settings. First of all, motion provides information that can be used for recognition independent of static form. For example, any individual image of a point-light walker does not evoke a very vivid percept of a human form, but once it begins to move the percept is irresistible (Johansson, 1973). Despite the sparsity of the point-light stimulus, observers are capable of recovering the gender and identity of walking figures with good accuracy. (Cutting, 1987; Dittrich, 1993; Kozlowski & Cutting, 1977) In other cases where static form is impoverished, motion features can support object or person identification (Knappmeyer, Thornton, & Bulthoff, 2003; Rosenblum et al., 2002). Object motion can thus be considered a useful set of additional measurements available for object recognition alongside static features.

Beyond this basic concept of object motion as a parallel source of information for recognition, there is also evidence that appearance dynamics may be directly incorporated into representations of object form. The recognition of novel written characters can be substantially influenced by observers' understanding of the dynamical process that created them, making form distortions consistent with learned strokes more tolerable than other distortions (Freyd, 1987). The perception of apparent motion in Kanji characters is similarly modulated by observers' previous knowledge of how the character is normally drawn, resulting in different percepts for Asian and Western observers (Tse & Tarr, ????). There is also evidence that dynamic information can interfere with form recognition for complex three-dimensional objects. Observers who learn a set of novel objects undergoing rigid rotation are significantly impaired at recognizing these objects if the direction of rotation is reversed at test (Stone, 1998). The lack of object motion that is consistent with training appears to overshadow the preservation of static form in this case, leading some to posit that dynamic objects are encoded via "Spatiotemporal signatures" of object appearance. This proposal suggests that static figural cues are combined with motion information such that recognition is dependent on consistency of form and movement. Further evidence for such

"signatures" has been obtained in several recent object learning and recognition tasks (Stone, 1999; Vuong & Tarr, 2004).

Finally, we consider one additional proposal regarding the interaction of object motion with object recognition. The temporal association hypothesis suggests that temporal proximity between images allows the human visual system to learn useful invariants for recognition by binding disparate images together into a common representation. Many changes in viewing conditions preserve identity while drastically altering image-level features (Moses, Adini, & Ullman, 1994). This presents a very difficult recognition problem, but temporal proximity provides a principled means of linking very different images to the same physical object. Behaviorally, evidence for such linkages has been demonstrated for both faces and novel objects using paradigms in which temporal association between images leads to increased confusability between objects presented close in time (Cox, Meier, Oertelt, & DiCarlo, 2005; Wallis, 1996; Wallis & Bulthoff, 2001). Several computational models of this process have been implemented as well, lending credence to the idea that learning to recognize a novel object under different viewing conditions may be facilitated by observing that object in motion (Foldiak, 1991; Stone, 1999; Stone & Harper, 1999; Ullman & Bart, 2004; Wallis, 1998). The basic temporal association hypothesis is important in that the theory provides a useful bridge between recognition in a fully dynamic setting and classical work describing static recognition. It suggests that motion contributes to the representation of static form, rather than providing an independent set of dynamic features or serving as half of an integrated representation of particular sequences. Given that many models of static object recognition already exist, we suggest that this proposal provides a nice starting point for examining how models of static object recognition should be altered to incorporate the richness of spatiotemporal data.

In the current study, we aim to extend our current understanding of how temporal association between images affects static recognition by examining the perception of the locomoting human body. Our goal is to determine how sensitivity to image differences changes as a function of object familiarity, motion familiarity, and short-term exposure to a moving object.

The visual perception of the human body is an extremely valuable test case for any theory positing that experience with a dynamic object should influence subsequent performance with static images. First, the human body moves non-rigidly. This severely limits the utility of a static representation of 3-D form, such as a "geon" based encoding scheme (Biederman, 1987; Marr & Nishihara, 1978). Second, observers have a great deal of visual experience with moving bodies. The result is that knowledge of biomechanical constraints and the expected form of the human body can profoundly affect "low-level" processing. For example, Shiffrar and Freyd (Shiffrar & Freyd, 1990) demonstrated that perceived apparent motion of the human body is determined both by the speed of the display and observers' knowledge of possible and impossible motions. Similarly, knowledge of allowable body movements affects the priming of interpolated frames from apparent motion sequences (Kourtzi & Shiffrar, 1999). Top-down influences on the perception of body-like figures were also reported by Sinha and Poggio (Sinha & Poggio, 1996) who demonstrated that a rigidly rotating wire "walker" was generally perceived as non-rigidly deforming, presumably because observers'

expected the human-like form to move in this manner. Similarly, using stereoscopically presented images, Bulthoff et al. (Bulthoff, Bulthoff, & Sinha, 1998) found that expectations of allowable human forms could bias the perceived depth of dots in point-light stimuli. All of these cases demonstrate that human body perception is a good domain for exploring interactions between prior knowledge and perception. Since top-down processes are capable of influencing perceived motion and stereo, it is likely that they can also affect static form perception.

The nature of observers' visual experience with the human body is also particularly useful in that it is asymmetric. Specifically, backward walking is not impossible, but it is far less common than forward walking. This provides us with the opportunity to ask some interesting questions regarding whether or not the visual system learns about object movement in a temporally asymmetric manner, both in the short and long-term. Human observers do use their knowledge of motion asymmetries in the natural world to predict future events, leading to measurable differences in change detection (Freyd, 1983; Freyd & Finke, 1984). Static images of "frozen" motion can even give rise to activation in cortical areas devoted to motion perception (Kourtzi & Kanwisher, 2000), suggesting that predicting future scenes is both psychologically and neurally real. What remains unknown is how predictions based on long-term experience interact with predictions based on very recent exposure.

In the experiments we describe here, we use a simple change detection paradigm to ask several questions regarding the basic temporal association hypothesis and how recent temporal association between images interacts with prior expectations about appearance dynamics. First, we characterize observers' sensitivity to detecting the differences between static images of a normal walking human body that are arranged in natural or reversed temporal order. We find evidence that automatic prediction impairs sensitivity to "forward" changes relative to "backward" changes, and that this impairment is invariant to rotation in depth. Second, we use the same task to evaluate sensitivity to image differences between images of a human performing a strange, but physically possible, gait. Here we find no difference in sensitivity for forward and backward changes, suggesting that it is actual dynamic experience that leads to this performance asymmetry rather than general biomechanical knowledge. Third, we determine how brief exposure to a dynamic stimulus (a walking human) affects subsequent performance at this task. Using distinct groups of participants, we exposed observers to either forward or backward walking to determine how an asymmetry in short-term exposure to a dynamic object interacts with a learned asymmetry in expected image change based on prior knowledge. Our results indicate that recent dynamic experience destroys the forward-backward asymmetry at test, but that the motion observed during training has a profound effect on overall performance levels. Finally, in all of these experiments observers' sensitivity to inverted bodies was measured simultaneously to determine the extent to which previously learned or recently induced temporal factors affected performance with an unfamiliar object. We conclude by discussing the consequences of these results for the basic formulation of the temporal association hypothesis, and suggesting further avenues of investigation.

Experiment 1

In our first task, we ask whether or not prior knowledge concerning expected movement of the human body during walking has consequences for form discrimination. We measure the relative sensitivity to "forward" and "backward" image changes across three different views of the same walking figure in upright and inverted conditions. This allows us to ask first of all whether or not there are consistent effects of appearance prediction on sensitivity, and also whether any such effects are dependent on either the amount of image change or the familiarity of the stimulus.

Method

Subjects

A total of 12 volunteers participated in this task. All participants were between the ages of 18 and 35 and reported normal or corrected-to-normal visual acuity. Also, all participants were naïve to the purpose of the experiments.

Stimuli

All images were created using Poser 6, a 3-D graphics tool for rendering and manipulating models of human bodies (Curious Labs). A male figure was created using the software's default settings and rendered from three viewpoints (side, three-quarters, and frontal views) while walking in place at a normal speed. Figure 1 contains example images of the model at each of the three rendered views. 60 images were rendered at each view depicting the model carrying out a complete walking cycle of two steps, allowing us to continuously loop these images to display ongoing walking. Each image was 278x484 pixels in size (approximately 5 degrees x 10 degrees visual angle) and contained 256 gray levels.

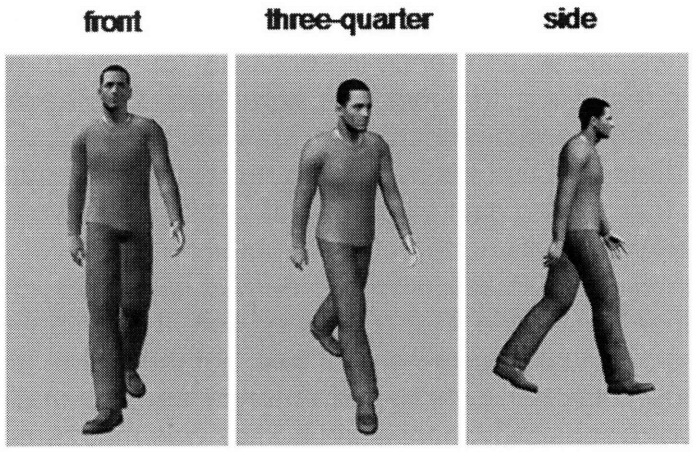


Fig. 1 – Example views of the male walker created for the experiments. The images depict the model in the same pose for each of our three views.

Procedure

All participants performed a change detection task using images from the various walking sequences described above. Volunteers in this experiment carried out this task using all three views of the model. Images were presented in blocks according to rendered view. Block order was counterbalanced across subjects.

On each trial, observers would first see one image of the walker for 500ms, followed by a 500ms delay period, then presentation of a second image for an additional 500ms. (Figure 2) The second image was always translated by +/- 10 pixels horizontally and vertically. Observers were instructed to press "1" if they believed the two images were identical, and "0" if they believed they were different. Errors were indicated with an auditory stimulus.

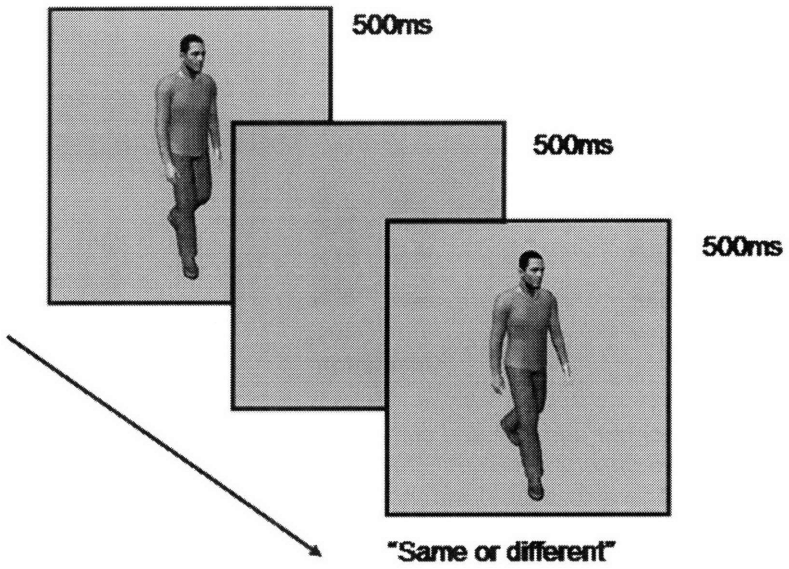


Fig. 2 – A trial from our discrimination task using walker images. On each trial, two images were presented sequentially for 500ms each. The images could be identical, different and in forward temporal order, or different and in reverse temporal order. Furthermore, half of the pairs depicted upright walkers and the other half depicted inverted walkers. Auditory feedback was given when errors were made.

Image pairs contained either two upright or two inverted images and these images could either be identical or different. Pairs of images that were different were further broken down into "forward" pairs and "backward" pairs. "Forward" pairs contained images that were in the correct temporal order for forward walking while "backward" pairs contained images in the opposite order. Each forward pair had a corresponding backward pair in our design containing the same two images presented in opposite order. All "different" image pairs were 2 frames apart in the original rendered sequence (i.e. Frame #5 and Frame #7). At test, the order of pair presentation within each block was randomized separately for each subject. There were 42 image pairs in each of our 6 conditions, yielding a total of 252 trials.

All stimuli were presented on a 19" Sony monitor. Participants were seated approximately 50cm from the monitor, such that all images of the walker subtended

approximately 10x5 degrees of visual angle. Stimulus timing and response collection was monitored by software written using the Matlab Psychophysics Toolbox (Brainard, 1997; Pelli, 1997).

Results

This experiment allows us to characterize the effect of prior knowledge on the discrimination of images placed in typical or atypical temporal order. Our initial hypothesis is three-fold: First, we expect that subjects will find it easier to detect image differences when images are presented in backward order (consistent with involuntary prediction). Second, we expect that this advantage will be greatest when image-level differences are small. Third, this advantage should disappear for inverted images, given their unfamiliarity. We test all these predictions by calculating d' for both forward and backward image pairs at all views, and both upright and inverted stimulus orientations. A graph of these results is displayed in Figure 3.

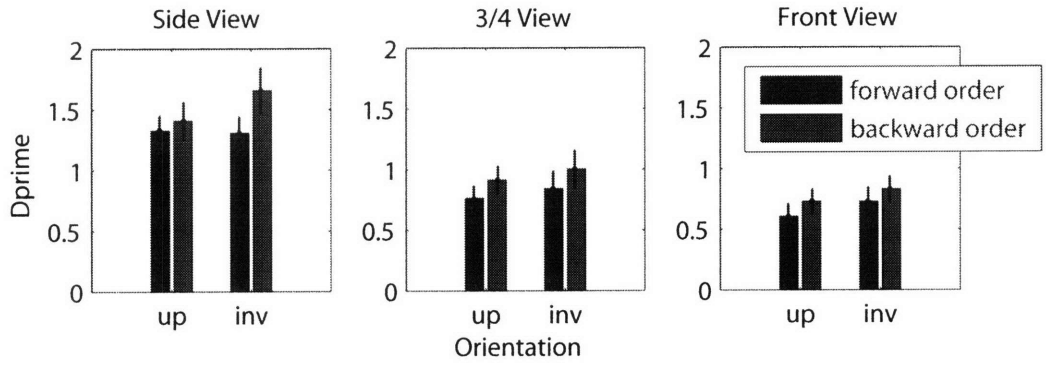


Fig. 3 – Discriminability for upright and inverted stimuli presented at three views in either forward or backward temporal order. We observe main effects of temporal order (favoring backward presentation) and view (favoring larger image differences) with no significant interactions or other main effects. Error bars represent 1 +/- s.e.m. across the group data for each condition.

A 3x2x2 repeated-measures ANOVA with view, orientation, and temporal order as factors yields a main effect of view ($F(2,10)=10.87$, $p=0.003$) and a main effect of order ($F(1,11)=9.25$, $p=0.011$). No other main effects or interactions were significant.

Discussion

These results are in accord with our hypothesis that backward differences should be easier to detect than forward differences. However, there is no interaction between view and temporal order. The consistency of the advantage for detecting backward differences across view suggests that the amount of image difference within a pair does not modulate the effect of temporal order. Instead, change in an object-centered frame of reference seems more relevant. This is somewhat at odds with characterizations of biological motion perception as strictly view-dependent (Verfaille, 1993), insofar as prediction appears to take place in a view-independent way. However, it is not clear that view-dependence in recognition performance should imply view-dependence in predictive perception.

Finally, it is interesting to see that the results from the inverted stimuli are basically identical to those from upright images. Our initial prediction was that order effects should disappear for inverted images since observers have no experience with such images and should therefore be unable to accurately predict future appearance from one frame. Instead, the only difference we see between performance with upright and inverted stimuli is a criterion shift (subjects seeing inverted images are more likely overall to respond "different" across all views), not a change in relative discriminability. This may mean that the same underlying mechanisms are used to process both upright and inverted bodies.

Experiment 2

The results of Experiment 1 suggest that prior knowledge concerning object motion leads to systematic impairment for detecting "forward" image change. However, one possibility we cannot rule out yet is that prediction could be governed by generic knowledge of biomechanical constraints on movement. To address this point, we continue by examining the extent to which these effects generalize to a novel, but physically plausible gait. If observers display temporal ordering effects on sensitivity in this condition, we can conclude that extensive dynamic experience is not necessary for prediction to impair sensitivity. If however, the effects of temporal order disappear, we have more reason to believe that observers' perceptual experience is more relevant in this task than abstract rules concerning limb movement and balance.

Method

Subjects

8 additional volunteers participated in this task. All participants were between the ages of 18 and 35 and reported normal or corrected-to-normal visual acuity. Also, all participants were naïve to the purpose of the experiments.

Stimuli

Additional images were created using the same male figure generated for Experiment 1. In this case, however, the model's gait was edited via built-in tools to be highly unusual, but physically possible. Figure 4 contains example images of the model. As before, 60 images were rendered at each view depicting the model carrying out a complete walking cycle of two steps, allowing for continuous looping. Each image was 278x484 pixels in size (approximately 5 degrees x 10 degrees visual angle) and contained 256 gray levels.

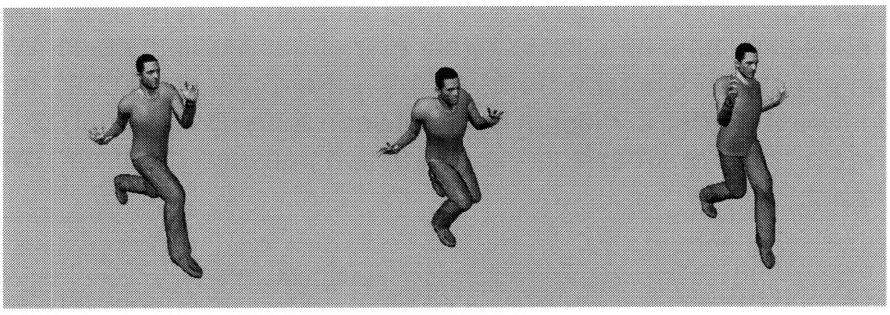


Fig. 4 – Images of the walker model performing the unusual gait used in Experiment 2. This gait is biomechanically possible, but not frequently observed in the environment.

For this task (and for the remaining experiments), we do not measure sensitivity across all three views of the walker employed in Experiment 1. Instead, we only measure sensitivity for the ¾ view of the walker. Such views are generally considered "canonical" object views and provide the most information about object movement and form due to the balance struck between a lack of foreshortening and self-occlusion. While examining all of these various effects across all views could be interesting, we limit ourselves to the ¾ case from here on to provide a more focused analysis and discussion.

Procedure

The change detection task described in Experiment 1 was implemented without alteration in this task. The only change is the stimuli used. Otherwise all display parameters and timing is identical to Experiment 1.

Results

As in Experiment 1, we calculate d' for forward and backward directions of image change in upright and inverted images. In Figure 5 we display the average d' value across subjects in all conditions.

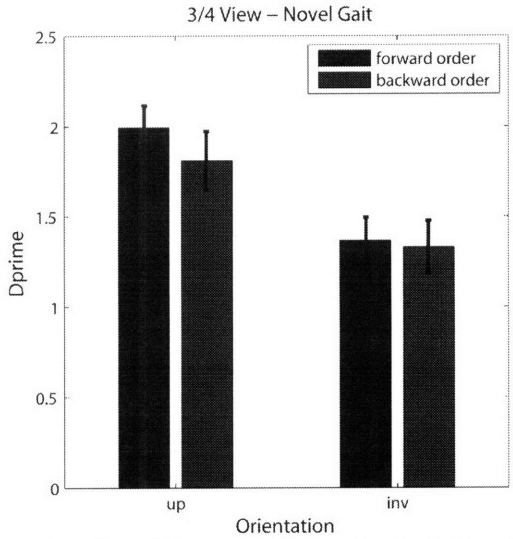


Fig. 5 – Mean dprime scores as a function of image order and orientation. Error bars represent +/- 1 s.e.m.

We carried out a 2x2 within-subjects ANOVA to determine if there were any significant differences in sensitivity as a function of either one of our manipulations. We find only a significant main effect of orientation ($F(1,9)=14.9$, $MSe=3.08$, $p = 0.004$), indicating that inverted images were harder to discriminate between than upright images. There was no significant main effect of order and the interaction between order and orientation was also not significant.

Discussion

This task allows us to determine the extent to which observers' implicit knowledge of biomechanical constraints on movement drives the asymmetry between forward and

backward sensitivity observed in Experiment 1. The lack of any order effect in this task suggests that this knowledge is not the primary determinant of the asymmetry, since biomechanical constraints alone could allow observers to formulate predictions for this setting as well as the previous task. Given that this does not seem to occur, we suggest that the observed effect in Experiment 1 is primarily the result of actual visual experience with the moving human form. This is not to say that biomechanical knowledge plays no role in this task, only that we do not think it is the dominant factor.

One aspect of the data that is very interesting in light of the results of Experiment 1 is the striking inversion effect. Perhaps the constant exposure to upright motion leads to an ability to generalize over picture-plane rotation, but this is counter to traditional interpretations of "Expert" processing. Generally, it is suggested that overexposure to a certain stimulus class leads to an excessive commitment to the representation of the normal stimulus to the detriment of generalization. This is the suggested cause of the famous face inversion effect, for example, and others have used the presence of inversion (not its absence) as evidence of expertise. For the moment, we shall not pursue this issue further since it is tangential to the main thrust of our work. Instead, we offer it as an intriguing touchstone for further inquiry. We continue by returning to our original walking figure to examine the effect of recent dynamic exposure on sensitivity.

Experiment 3

In our last task, we ask how recent exposure to a dynamic object interacts with prior expectations regarding object motion. Following a training period during which observers watch either forward or backward walking, we characterize sensitivity to image change using the change detection task described in Experiments 1 and 2. We ask first of all whether or not there is any effect of the passive training period on the previously observed temporal asymmetry in discrimination. Second, we ask if these effects generalize to the inverted images.

Method

Subjects

A total of 24 volunteers participated in this task. All participants were between the ages of 18 and 35 and reported normal or corrected-to-normal visual acuity. Also, all participants were naïve to the purpose of the experiments.

Stimuli

The images of our walker performing the normal gait as described in Experiment 1 were also used here. As in Experiment 2, we limit ourselves to the ¾ view for simplicity.

Procedure

Volunteers who were placed in either "motion" group began by viewing our model walking either forward (Group 1) or backward (Group 2) for ten minutes. Viewing was broken up into ten 1-minute long blocks, during which the 60 images of the walker (three-quarter view) were continuously looped at a frame rate of 30 frames per second (with a 60hz refresh rate). To ensure that observers attended to the animation during each block, a "cue-dot" detection task was administered during presentation. A small red dot (~0.5 degrees of visual angle in diameter) was drawn at a random location on

the image at randomly selected times for a duration of 32ms. Observers were instructed to press the space bar every time this dot appeared.

Following this exposure period, all observers carried out the change detection task described in Experiment 1 using the images of the three-quarter walker only. There were no other differences in design parameters or procedure.

Results

Having determined in Experiment 1 that prior knowledge regarding human movement affects discrimination for images taken from a coherent walking sequence, we ask how recent exposure to a dynamic stimulus modulates behavior. As we have done in all of our other tasks, we calculate d' for forward and backward discrimination of upright and inverted stimuli. The results are summarized in Figure 6.

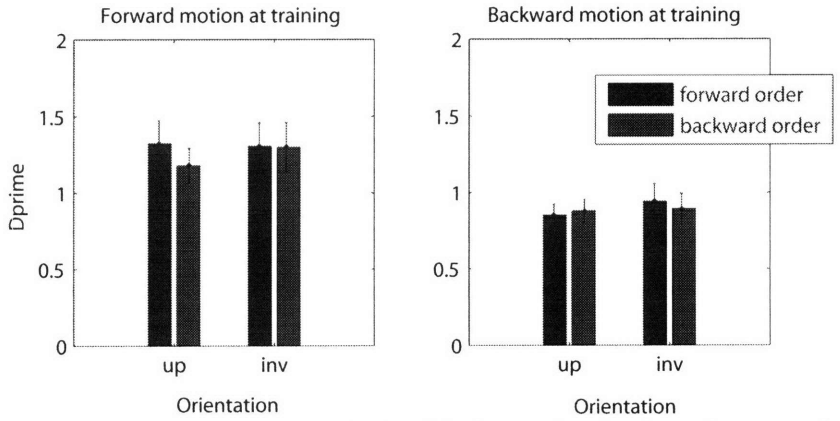


Fig. 6 – Discriminability for upright and inverted stimuli in forward or backward temporal order following exposure to a dynamic stimulus. Observing forward motion results in significantly better performance. Error bars represent 1 +/- s.e.m. across the group data for each condition.

A 2x2x2 mixed-design ANOVA was run, with temporal order and orientation as within-subject factors and forward v. backward training motion as a between-subjects factor. The only significant main effect we observe is an effect of training motion on discriminability ($F(1,22)=7.54$, $p=0.012$). Exposure to forward motion prior to performing the discrimination task yields significantly better performance than exposure to backward motion composed of the same frames played in reverse order. The previously observed effect of temporal order disappears in both cases, and as in Experiment 1, the data from inverted trials is essentially identical to the data from upright trials.

Discussion

Exposure to the walking figure used in our change detection task erases the temporal asymmetry observed in Experiment 1. This occurs regardless of whether or not the figure walks forward or backward during exposure, and the effects of this exposure generalize to the inverted images presented at test. This suggests that dynamic experience with an object has an immediate and broad impact on perception. Furthermore, we find that observing forward motion significantly increases sensitivity relative to observers who viewed backward motion. This also generalizes over upright and inverted images, and indicates that there is a complex interaction between prior expectations of movement and recently observed motion. Finally, as in Experiment 1 we

see no evidence of an inversion effect in our data. Discrimination between images of a normally walking figure appears to be robust to picture-plane rotation, despite what one might expect from previous examinations of image inversion.

We close by attempting to synthesize these various results into a coherent account of how object motion contributes to form processing in both the short and long-term.

General Discussion

In our first experiment, we see evidence that observers automatically predict the future appearance of a human body "frozen" in a walking motion. The result is that discriminating image changes consistent with forward motion is significantly more difficult than discriminating between the same pairs of image presented in reverse temporal order. This is in agreement with a wide range of studies of "representational momentum," with the added feature that by examining performance across different views we can also see that our effect is not contingent on the amount of image change between pairs. Our results are more consistent with a view-invariant prediction mechanism. Finally, the asymmetry in discriminability for forward v. backward sequences extends to inverted images of the model. This last finding was quite unexpected, and indicates a surprising amount of generality in the visual representation of the human body. To summarize, human observers appear to automatically predict future appearance of a body frozen in motion via mechanisms that are invariant to rotations in depth and inversion in the picture plane. To describe this in a basic signal detection framework, forward prediction places images arranged in forward temporal order closer together in the relevant measurement space. The same predictive behavior leads to more separation in the same space between images in backward temporal order. Assuming that intrinsic noise stays constant, the relative difference in separation between the signals in these two cases makes backward sensitivity higher than forward sensitivity. Experiment 2 also lets us rule out the possibility that this process is supported purely by knowledge of biomechanical constraints. Since the order effect vanishes for images of a walker with an unusual gait, we can infer that it is direct exposure to the dynamic stimulus that is the deciding factor. General principles of human movement do not make a substantial contribution.

Our third task provides us with the opportunity to elaborate this basic model by examining the effects of brief exposure to a dynamic stimulus prior to performing discrimination. If dynamic experience serves primarily to facilitate prediction, we should expect that exposure to the forward-moving walker would either have no effect on sensitivity (since the stimulus is consistent with prior knowledge) or exacerbate the existing effect by making forward prediction more robust and forward sensitivity consequently weaker. Under this same assumption, exposure to the backward walker should either nullify the observed asymmetry in sensitivity or reverse its sign.

We also consider the predictions made under a basic "temporal association" model. To review, in this model temporal neighbors are increasingly generalized over after sufficient exposure. To the extent that this occurs, we should expect to see impaired sensitivity following exposure to any moving walker. Since the basic association hypothesis makes no strong claims about the order of image presentation, we submit that in the context of this model, learned generalization should be symmetric with

respect to time. This would lead us to expect a downward baseline shift of equal magnitude in both the "forward" and "backward" conditions. Strictly speaking, this baseline shift in performance should not apply to the inverted images since they were not observed in temporal proximity to one another at any point during training.

What we observe in our third experiment requires us to reject both of these proposals as the sole explanation. Considering the effects of exposure to a backwards walker first, we see evidence that observers effectively "unlearn" the tendency to predict forward appearance. The relative deficit in forward sensitivity has disappeared, with no accompanying decrease in sensitivity to the backward walker. This result is compatible with a simple prediction model, but not the temporal association model. The data from the observers exposed to a "forward" walker proves troublesome for the predictive account, however. Instead of further impairments in forward sensitivity (consistent with a prediction model) or impairments in both conditions (consistent with time-symmetric generalization) we observe instead that the observed asymmetry in sensitivity has disappeared and observers show improvements in both temporal directions.

Neither model alone explains these results satisfactorily, but we will need both to continue. We suggest a two-stage model of how dynamic appearance affects subsequent sensitivity to static images. First, exposure to a dynamic object induces a tendency to predict future appearance. Second, if predictions are solidified, continued dynamic input causes immediate increases in generalization over a population of view-sensitive units. This latter interpretation of the basic temporal association hypothesis (in which increased generalization is linked to increased confusability) is distinct in that it posits the formation of a "coarse code" for object appearance. Distributed representations such as this provide for both increased generalization and increased sensitivity provided that the appearance space is sufficiently sparse (Hinton, McClelland, & Rumelhart, 1986). We have previously demonstrated that exposure to novel dynamic objects situated in a parametrized space gives rise to increased sensitivity for static image differences consistent with observed motion (Balas & Sinha, 2006), suggesting that this proposal is a plausible model for understanding the relationship between dynamic input and static image processing.

Observers who never saw our dynamic stimuli nonetheless entered the laboratory with enough visual experience with walking bodies to predict future appearance, leading to the asymmetric sensitivity profile exhibited in Experiment 1. Exposure to the dynamic walker was then responsible for doing one of two things, dependent on the direction of observed motion. For forward movement, observers' expectations were not violated and thus dynamic input was of no further use in formulating predictions. We suggest that at this point the population of units encoding walker appearance began to generalize more broadly, leading to a highly overlapping and accurate representation of form. Prediction may still be carried out under these conditions, but we propose that the effects of coarsening the appearance representation are larger. Now let us consider the case where backward movement is observed during training. Here, observers' expectations were violated and thus the dynamic input is used to re-formulate predictions for the current stimulus set. The conflict between expectation and observation suppresses the tendency to broaden appearance tuning over the population in any way. Given sufficient exposure to backward movement, our hypothesis predicts that observers should

eventually undergo the same improvement in performance we observe in the forward case. To put it simply, motion is used to increase sensitivity only after we know which way the "arrow of time" points.

To summarize, we propose that dynamic input serves both as data for generating predictions and as the impetus for coarsening the appearance code symmetrically with respect to time. The predictive aspect of dynamic input appears to be the most relevant aspect of dynamic exposure over long time scales. It is clear from our first experiment that learned appearance prediction has long-lasting and robust effects. Moreover, it is impressive that ten minutes of constant dynamic stimulation interacts with prior belief rather than completely overwhelming the results. This is good evidence that learned dynamics are very stable and partially resilient to immediate influences. It is also important to note that both short-term and long-term effects of exposure appear to be capable of operating over a wide range of viewing conditions, as evidenced by the data from inverted normal walkers. The fact that data from these inverted trials is indistinguishable from that taken from upright trials is surprising, and strongly implies that the visual system makes immediate and broad use of dynamic input.

The effects we observe also lead to several interesting questions for further investigation. For example, while it seems reasonable to assume that learned prediction for familiar object appearance is relatively long-lasting, it is unclear what the relevant time-scale is for the decay of the short-term effects we have observed. If subjects in either condition were re-tested a day or a week later, we cannot say as yet what would result. Furthermore, it would be interesting to observe the interplay between asymmetric prediction and symmetric generalization in a novel object. Allowing both factors to come into play during observers' initial experience with a novel object might give us a better sense of the relative priority given to these mechanisms by the visual system by eliminating long-term perceptual history as a confounding factor. Ultimately, understanding the relationship between dynamic and static experience provides an important window into the learning processes that support object recognition in complex and changing environments.

Acknowledgements

BJB was supported in this research by an NDSEG fellowship. PS is supported by a Merck Foundation fellowship, and the Alfred P. Sloan fellowship in Neuroscience.

References

Biederman, I. (1987). Recognition-by-Components: A Theory of Human Image Understanding. *Psychological Review, 94*(2), 115-147.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433-436.

Bulthoff, I., Bulthoff, H. H., & Sinha, P. (1998). Top-down influences on stereoscopic depth perception. *Nature Neuroscience*(1), 254-257.

Cox, D. D., Meier, P., Oertelt, N., & DiCarlo, J. J. (2005). 'Breaking' position-invariant object recognition. *Nature Neuroscience, 8*(9), 1145-1147.

Cutting, J. E. (1987). Perception and Information. *Annual Review of Psychology, 38*, 61-90.

Dittrich, W. H. (1993). Action categories and the perception of biological motion. *Perception, 22*, 15-22.
Foldiak, P. (1991). Learning Invariance from Transformation Sequences. *Neural Computation, 3*, 194-200.
Freyd, J. J. (1983). The mental representation of movement when static stimuli are viewed. *Perception and Psychophysics, 33*(6), 575-581.
Freyd, J. J. (1987). Dynamic Mental Representations. *Psychological Review, 94*(4), 427-438.
Freyd, J. J., & Finke, R. A. (1984). Representational Momentum. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(1), 126-132.
Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed Representations. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1,). Cambridge, MA: MIT Press.
Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics, 1*, 201-211.
Knappmeyer, B., Thornton, I. M., & Bulthoff, H. H. (2003). The use of facial motion and facial form during the processing of identity. *Vision Research, 43*, 1921-36.
Kourtzi, K., & Kanwisher, N. (2000). Activation Human MT/MST by Static Images with Implied Motion. *Journal of Cognitive Neuroscience, 12*(1), 48-55.
Kourtzi, Z., & Shiffrar, M. (1999). Dynamic representations of human body movement. *Perception, 28*, 49-62.
Kozlowski, L. T., & Cutting, J. E. (1977). Recognizing the sex of a walker from a dynamic point-light display. *Perception & Psychophysics, 21*, 575-580.
Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London, Series B., 200*, 269-294.
Moses, Y., Adini, Y., & Ullman, S. (1994). Face recognition: the problem of compensating for illumination changes. *Proceedings of the European Conference on Computer Vision*, 286-296.
Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision, 10*, 437-442.
Rosenblum, L. D., Yakel, D. A., Baseer, N., Panchal, A., Nordase, B. C., & Niehus, R. P. (2002). Visual speech information for face recognition. *Perception and Psychophysics, 64*(2), 220-229.
Shiffrar, M., & Freyd, J. (1990). Apparent motion of the human body. *Psychological Science, 1*, 257-264.
Sinha, P., & Poggio, T. (1996). The role of learning in 3-D form perception. *Nature, 384*, 460-463.
Stone, J. V. (1998). Object recognition using spatiotemporal signatures. *Vision Research, 38*, 947-951.
Stone, J. V. (1999). Object recognition: view-specificity and motion-specificity. *Vision Research, 39*, 4032-4044.
Stone, J. V., & Harper, N. (1999). Temporal constraints on visual learning: a computational model. *Perception, 28*, 1089-1104.
Ullman, S., & Bart, E. (2004). Recognition invariance obtained by extended and invariant features. *Neural Networks, 17*, 833-848.

Verfaille, K. (1993). Orientation-dependent priming effects in the perception of biological motion. *Journal of Experimental Psychology: Learning, Memory and Cognition*(20), 649-670.

Vuong, Q. C., & Tarr, M. J. (2004). Rotation direction affects object recognition. *Vision Research, 44*(14), 1717-1730.

Wallis, G. (1996). Using Spatio-temporal Correlations to Learn Invariant Object Recognition. *Neural Networks, 9*(9), 1513-1519.

Wallis, G. (1998). Spatio-temporal influences at the neural level of object recognition. *Neural Networks, 9*, 265-278.

Wallis, G., & Bulthoff, H. H. (2001). Effects of temporal association on recognition memory. *Proceedings of the National Academy of Sciences, 98*(8), 4800-4804.

Introduction to Chapter 3

In the final chapter of Section 1, I stray somewhat from the use of purely image-level judgments to ask how object motion affects the construction of category representations for novel dynamic objects. In particular, I ask how the diagnosticity of motion during training influences the ability to form a robust representation of object form across stimulus categories. This chapter adds a further layer of complexity to this analysis of object motion and form by asking how static form cues are employed in the service of category representations when dynamic features may or may not provide class information.

Diagnostic Object Motion Weakens Representations of Static Form

Abstract

Past studies have shown that information about how objects move can play an important role in their recognition. Flow-fields associated with an object's intrinsic motion, and also the sequence of views it presents over time can be used to identify the object and also link its disparate appearances. In the current study, we demonstrate that diagnostic object motion is such a perceptually significant cue that it can actually impair classification by de-emphasizing static figural information. Our stimuli comprise exemplars from a synthetic object category. The exemplars can be distinguished from each other on the basis of both static and dynamic cues. When object dynamics perfectly correlate with category membership during training, observers tested at static image classification display significantly longer RTs than observers trained with non-diagnostic object motion. This demonstrates that object motion is a particularly salient aspect of object appearance, capable of suppressing equally useful qualities such as static form, color, or texture.

Introduction

To what extent does object motion play a role in object recognition? This apparently simple question has a complicated answer. In particular, while there is a great deal of evidence suggesting human observers can and do use intrinsic object motion as a cue for identity, it remains unclear how motion and form interact during the acquisition of object concepts. In the current study, we attempt to address this issue by investigating the effects of diagnostic and non-diagnostic motion on the categorization of static object.

Observers do use object motion to categorize stimuli. Though this can be seen in the results of studies using clearly viewed objects (Newell, Wallraven, & Huber, 2004), it is particularly evident when static form is degraded. An extreme version of this is the perception of "point-light walkers" (Johansson, 1973). In the absence of static cues for identity and gender, observers make good use of dynamic input to categorize walkers (Kozlowski & Cutting, 1977). A similar result obtains for face recognition. An "average" face that is made to undergo the idiosyncratic motions of a particular individual can be identified as that individual by naïve observers (Hill & Johnston, 2001; Knappmeyer, Thornton, & Bulthoff, 2003). Finally, there are many studies suggesting that observation of a familiar moving face or body facilitates recognition under degraded viewing conditions (Burton, 1999; Knight & Johnson, 1997; Lander & Bruce, 2000). There remain several open issues, especially the existence of a motion benefit for unfamiliar faces and the possible differences between rigid and non-rigid motion (Christie & Bruce, 1988; Pike, Kemp, Towell, & Phillips, 1997; Schiff, 1986). The overall picture appears to be quite complex, but it seems fair to say that in some circumstances object motion is relied upon for categorization when static form is impoverished.

A second issue regarding the use of motion and form for recognition relates to what happens when motion cues and form cues conflict somehow. By setting motion and

form against one another, we can determine the relative weight allotted to each under clear viewing conditions. Currently, there is some evidence that the motion of an object may take precedence over static form cues. For example, a "chimeric" point-light walker with static cues indicative of one gender (as defined by shoulder-hip ratio) and dynamic cues indicative of the other is categorized according to its movement rather than its form (Thornton, Vuong, & Bulthoff, 2003). Similarly, in face perception there is evidence that infants use dynamic information more than static form as a cue for identity (Spencer, O'Brien, Johnston, & Hill, 2006). Infants will not dishabituate to an old motion pattern superimposed on a new face, indicating that the novelty of the form does not compensate for the familiarity of the motion. Finally, there are several results demonstrating that the direction of rotation for an unfamiliar object becomes an important cue for recognition after relatively little training (Stone, 1998; Vuong & Tarr, 2004). Specifically, reversing the direction of rotation has a strong impact on recognition ability, despite the fact that the same static information is available during training and test periods. Object motion overshadows form in this task, in that the violation of expected object motion has strong consequences even though form is preserved.

These lines of work indicate that observers use object motion for recognition, and even suggest that it is given more importance than static form. In the current study, we extend this idea by examining whether or not observed object motion during training can affect test performance with static images. If object motion provides independent features for recognition, the absence of dynamic features at test should eliminate the effects of dynamic training. However, if dynamic training can affect later static performance, that provides good evidence for an interaction between object motion and the encoding of static form.

Presently, it is unclear whether or not observed object motion can affect static recognition. During rigid rotation, it has been suggested that "structure-from-motion" might allow observers to obtain 3-D information from coherent motion sequences, leading to better recognition. However, recent results indicate that observing object rotation is not a pre-requisite for view-invariant recognition (Wang, Obama, Yamashita, Sugihara, & Tanaka, 2005). Also, though a recognition advantage for temporally coherent vs. incoherent views of a rigidly rotating object has been reported before (Lawson, Humphreys, & Watson, 1994), the exact opposite result has also been reported (Harman & Humphreys, 1999).

If we consider non-rigid motion instead, there is more consistent evidence supporting the possibility that object motion might affect static object perception. For example, dynamic prime images of faces facilitate performance in static image matching (Thornton & Kourtzi, 2002). Also, apparent motion sequences depicting non-rigid objects deforming while rotating effectively prime static matching more than the same sequences displayed without apparent motion (Kourtzi & Shiffrar, 2001). Unfortunately, these studies reveal more about the nature of dynamic encoding than they do about the nature of static encoding following dynamic experience.

Finally, it has also been shown that temporal proximity between images of an object facilitates the binding of those images into a common representation (Cox, Meier, Oertelt, & DiCarlo, 2005; Wallis & Bulthoff, 2001). However, it also seems that

structural similarity can play a similar role even when temporal contingencies are eliminated (Perry, Rolls, & Stringer, 2006).

Given the lack of a clear picture regarding the influence of dynamic training on static recognition performance, we have attempted in the current study to determine whether the observed motion of objects during training can affect the efficiency of static image categorization. This is similar to previous attempts to determine whether motion coherence (usually defined as smooth vs. "random" image ordering) affects performance with static images, but there are several important differences between our work and previous efforts.

First, instead of manipulating motion coherence, we manipulate the diagnosticity of object motion. That is, object motion can either be perfectly indicative of object category (or "diagnostic") or object motion can be highly similar across categories (or "non-diagnostic"). We carry out this manipulation through the use of a class of novel objects called "blobs," created and introduced previously by Nederhouser, Mangini, and Biederman (Nederhouser, Mangini, & Biederman, 2002). The structure of the stimulus appearance space (Figure 1) allows us to define two categories that are always distinguishable by form alone. Within the set of images defining a category, the validity of object motion as a cue for category membership can be determined by how we concatenate still images into dynamic sequences for training. The advantage of using diagnosticity instead of motion coherence is simply that the use of randomized or "strobed" presentation of an otherwise coherent sequence may encourage observers to use very different processing strategies in different conditions. Minimizing this possibility by presenting coherent motion to all participants makes it more likely that we are comparing performance across commensurable tasks.

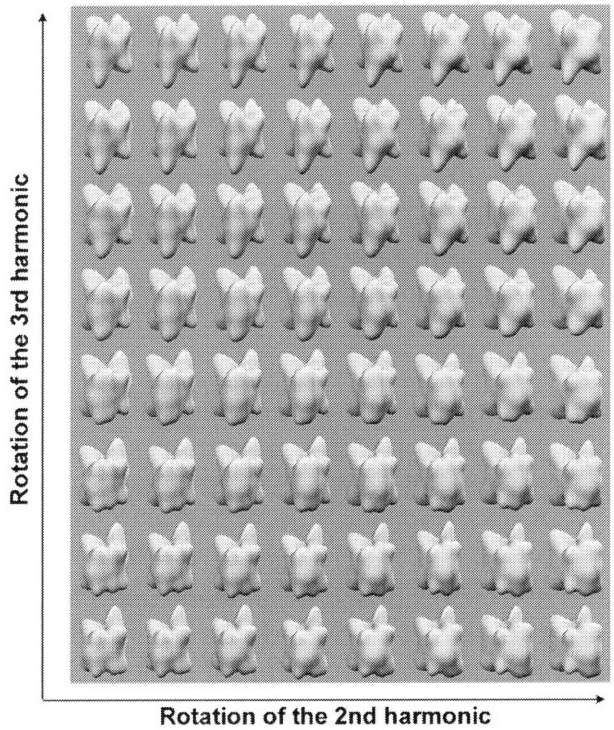


Fig 1. - An 8x8 space of "blob" stimuli. The axes of this space are defined by the phase angle of the 2nd and 3rd harmonic. Movement along each axis induces non-rigid motion that is distinct from that generated by movement along the other axis.

Second, our objects only move non-rigidly. The adult visual system may be so over-exposed to rigid object motion that training effects could be difficult to obtain without introducing novel object deformations. The use of non-rigid motion also confers the additional advantage of making it impossible to explain observers' performance in terms of static volumetric models of object form. Since there is no "ground truth" form, there is no way for observers to build a static object model.
Finally, we suggest that our experiments usefully complement previous work by examining how the validity of a cue, rather than its availability, affect the use of another cue. In some ways this is more natural than placing cues in conflict, or selectively impairing one cue and not another. Under natural viewing conditions, it is probably very common for observers to assess the utility of various cues and weight them accordingly. The question we ask here is if a change in the validity of one cue (object motion) affects the efficacy of a cue with stable validity across groups (object form).

In our first experiment, we manipulate motion diagnosticity by concatenating images into dynamic objects along differently oriented "paths" through blob appearance space. Objects within a category are always built by concatenating images together along paths of the same orientation, but across categories we either allow path orientation to match or not match depending on the experimental condition. We find here that learning to categorize objects with diagnostic motion leads to no difference in accuracy of static image classification, but significantly slower RTs. In our second experiment, we match path orientation across categories and ask whether direction of motion along the path is sufficient to induce the RT difference we observe in Experiment 1. Under these conditions, there is no difference in accuracy or RT, leading us to suggest that it is a symmetric estimate of appearance variability that underlies performance in this task rather than a feature like the motion flow field.

Experiment 1

In this experiment, we define the diagnosticity of object motion in terms of qualitatively different motion patterns obtained by concatenating images along "horizontal" or "vertical" paths through blob appearance space.

Methods

Subjects

Participants were 16 members of the MIT community (8 men and 8 women, with an age range of 18-40 years old), all of whom were naïve to the hypothesis under consideration. All observers reported normal or corrected-to-normal vision.

Stimuli

The "Blob" stimuli created by Nederhouser et al. were used in all the experiments reported here, and we refer the interested reader to their initial report for a more detailed account of blob construction than we present here (Nederhouser et al., 2002). Blobs are defined as a sum of spherical harmonics with varying amplitude and phase and an outer surface interpolated over the resulting object. The space of blobs used in the present study is defined by rotating the phase angles of the 2^{nd} and 3^{rd} harmonic independently, yielding a 16x16 space of images. We display this appearance space in

Figure 1. By starting at one image in the space and rotating the phase angle of only the 2^{nd} harmonic, we end up with what we will call "horizontal" motion through blob space. Rotating only the 3^{rd} harmonic results in what we will call "vertical" motion. It is important to keep in mind that the terms "horizontal" and "vertical" only refer to the arrangement of blobs into the flat space presented in Figure 1. The actual blob motions obtained by concatenating images either "horizontally" or "vertically" are highly complex, global deformations.

Images were assigned to different classes according to their position in blob space. Specifically, "Class I" objects were defined as images depicting a blob with both 2^{nd} and 3^{rd} harmonics oriented between 0 and 90 degrees, while "Class II" objects depicted only blobs with both harmonics oriented between 90 and 180 degrees. The resulting classes are wholly distinguishable by static form alone.

Within the 8x8 appearance space of images defining each class, we constructed dynamic objects by concatenating images together along either the "horizontal" or "vertical" paths, yielding qualitatively distinct non-rigid motions. (Figure 2) All objects in the same class underwent the same object motion, but objects in different classes could either undergo matching motions (non-diagnostic group) or distinct motions (diagnostic group).

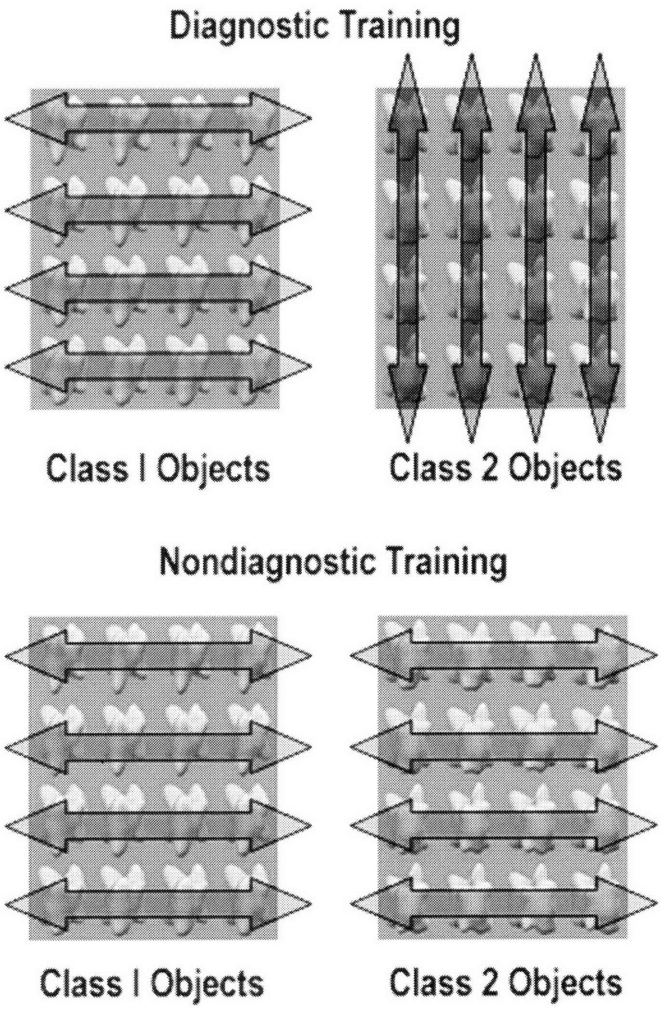


Fig 2. - Object motion diagnosticity as defined in Experiment 1. The top row depicts the construction of dynamic objects for observers in the "Diagnostic" group while the bottom row depicts the same for observers in the "Non-diagnostic" group. The same images are always used to define Class I and Class II, but they are assembled into movies in distinct ways. Note that in the full design, path orientation was balanced across observers such that horizontal and vertical motion occurred in each class the same number of times across both groups.

Procedure

Each image sequence was constructed by oscillating back and forth along one axis in appearance space while maintaining a fixed position on the orthogonal axis. Each movie displayed three complete oscillations (48 frames) and was played at a rate of 12 frames per second. Each object class contained 8 distinct movies, each of which was viewed 12 times during training for a total of 96 dynamic stimuli. Observers classified dynamic stimuli using the "1" and "2" keys on the keyboard and were provided with audio feedback during training. Participants in both groups were told that they were going to have to learn to classify the moving objects into two groups during this training period, and that they would then have to classify still images of the same objects afterwards.

Following training, observers were asked to classify static images as either "Class I" or "Class II" objects according to whatever criterion they had established during training. During this test phase, the 128 frames used to generate the training sequences were each displayed individually 4 times for a total of 512 stimuli. Each stimulus was presented for approximately 750ms. Responses could be collected at any time after initial presentation, and subjects were urged to respond as quickly and accurately as possible. Both accuracy and response time were recorded. All stimulus display parameters and response collection routines were controlled using the MATLAB psychophysics toolbox (Brainard, 1997). Stimuli were displayed on a calibrated 19" Dell Ultrasharp monitor, with a refresh rate of 60Hz. The objects subtended a visual angle of approximately 3 degrees during both training and test and were displayed on a uniform gray background. No feedback was given during this task.

Results and Discussion

All participants rapidly learned to correctly distinguish between dynamic exemplars of Class I and Class II objects. In the 2nd half of the training period, all of our observers attained over 96% correct performance, indicating that in both conditions learning to correctly label dynamic Class I and Class II objects was quite easy. Recognition performance in the test phase of our task was assessed by both accuracy and response time for correct categorization. Both subject groups performed very accurately at the static recognition task (~85% correct and 89% correct for the diagnostic and non-diagnostic groups respectively) with no significant difference between groups. In terms of reaction time however, we observe a strong effect of training condition. Subjects who learned to distinguish Class I objects from Class II objects under non-diagnostic conditions were able to correctly categorize static exemplars from both classes faster than subjects who observed diagnostic motion during training. The Mean RT from the diagnostic group was approximately 1050ms, which proved significantly longer than the 700ms mean RT observed in the non-diagnostic group ($t(14)=2.16$, $p < 0.05$). Figure 3 shows accuracy and RT data from both subject groups.

This result demonstrates that diagnostic object motion can actually repress the formation of a robust representation of static form during learning. Despite explicit instructions that training with dynamic objects would be followed by a test of static recognition abilities, subjects who observed diagnostic motion during training took longer on average to correctly identify still frames from the previously observed image sequences.

Could it be the case that observers in the "Diagnostic" group were simply ignoring object form and attending only to object motion? First of all, we emphasize again that observers were fully aware that their static recognition would be tested following the dynamic training period. Second, given that both subject groups perform accurately at test and do not differ in accuracy, it is difficult to imagine that observers in one group were simply not attending to object form during training. Clearly, both groups were capable of using form to categorize the objects, it is just that members of the "Non-diagnostic" group were able to do this more efficiently.
This result gives us a first piece of evidence that the validity of object motion for categorization can significantly affect the efficiency with which form can be used by naïve observers. Crucially, object form was fully available and fully diagnostic for both

groups, making it all the more surprising that object motion was able to impact static classification in this manner. Furthermore, it is interesting to see that diagnostic motion weakens the efficiency of static form. This result is consistent with a model of categorization in which a limited amount of weight is allocated to features that might be useful for identifying objects. The validity of diagnostic motion may simply draw resources away from representations of the category based on static form, leading to a less useful set of tools for the static test case.

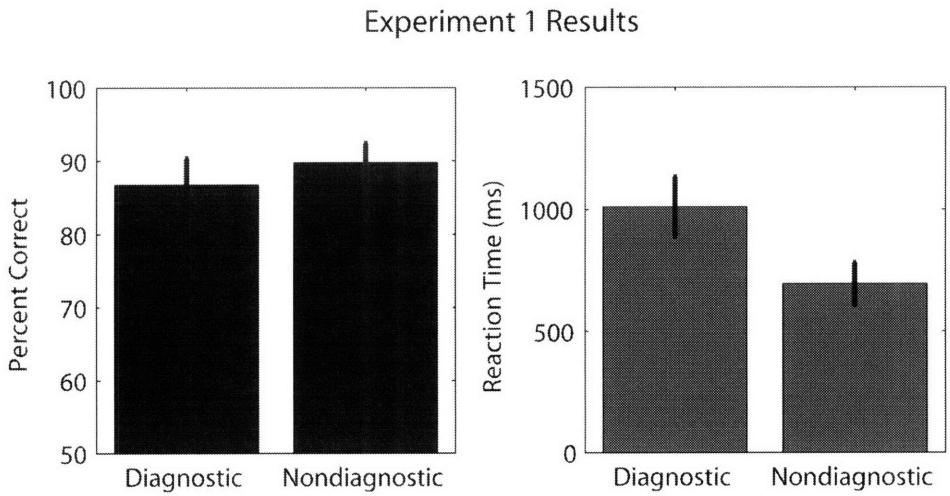


Fig. 3 - Accuracy (left) and response time for correct judgments (right) for observers in Experiment 1. There is no significant difference in accuracy between the two groups, but mean RTs are significantly longer in the Diagnostic group. Error bars represents +/- 1 s.e.m.

We continue by asking a more fine-grained question regarding the nature of diagnosticity for object motion. Specifically, we ask whether or not the direction of blob motion along a "path" of fixed orientation is sufficient to induce the effects we observe here. This experiment provides us with more insight into the nature of the dynamic features that impact static form representations. In particular, it helps us determine the extent to which the sign of motion vectors in a flow field (as determined by an optic flow algorithm, for example) is sufficient to evoke the differences in RT we see following "Diagnostic" and "Nondiagnostic" training.

Experiment 2

Methods

Subjects

16 additional members of the MIT community participated in Experiment 2, all of whom were naïve to the hypothesis under consideration. All observers reported normal or corrected-to-normal vision.

Stimuli

The same space of blob images was used to define object classes and create dynamic stimuli. The partitioning of images into Class I and Class II objects was also preserved so that the form information defining the two categories is matched across conditions and experiments. What differs in this task is that the diagnostic motion no longer results

from differently oriented paths through appearance space, but instead from a difference in the direction of motion through appearance space.

Dynamic objects were created by concatenating images in a consistent direction ("left" or "right" along "horizontal" paths only.) In this case, motion diagnosticity is determined by whether images were concatenated in matching directions along horizontal paths (Non-diagnostic group) or not (Diagnostic group). Figure 4 provides a schematic view of these conditions.

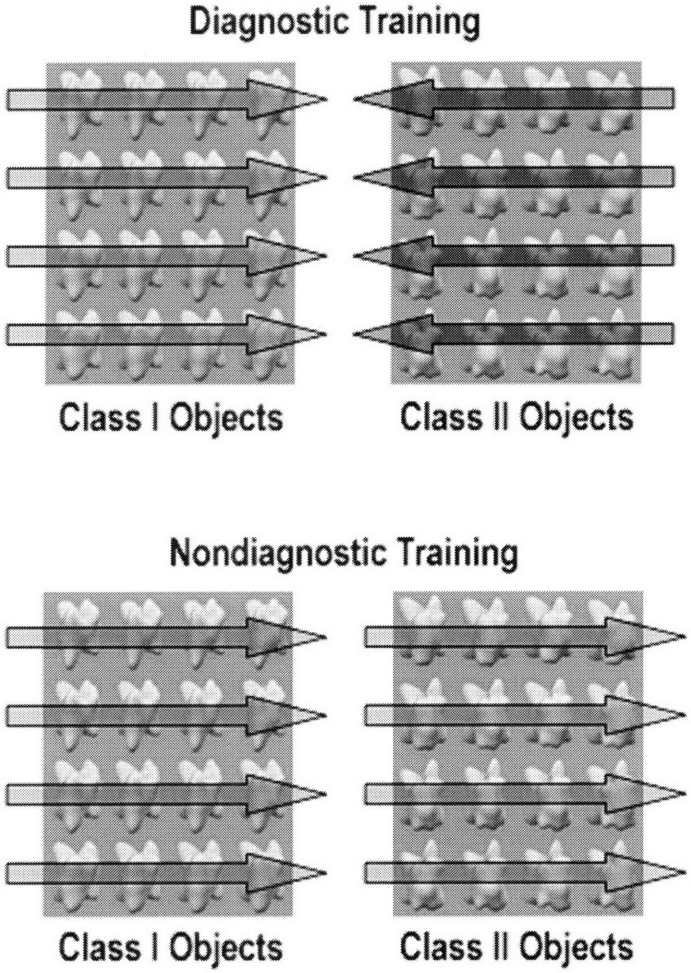


Fig. 4 - Object motion diagnosticity as defined in Experiment 2.

Procedure
The procedure for this task is identical to that described for Experiment 1.

Results

Mean accuracy and response time for accurate classifications in both groups is presented in Figure 5. Contrary to what we found in Experiment 1, there is no difference in performance between groups for RT ($t(14)=0.18$, $p=0.42$, two-tailed test).

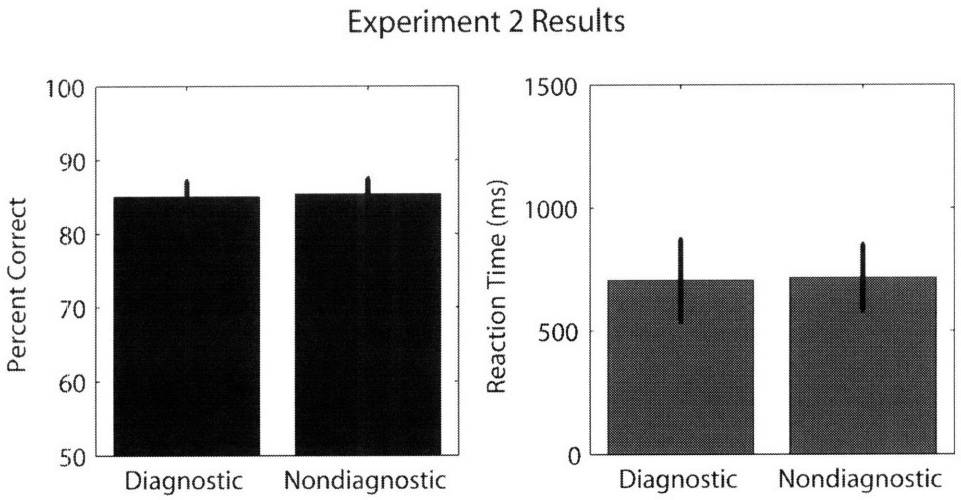


Fig. 5 - Accuracy (left) and response time for correct judgments (right) for observers in Experiment 2. There are no significant differences for accuracy or RT. Error bars represents +/- 1 s.e.m.

General Discussion

Experiment 1 demonstrated that diagnostic object motion could impair static classification performance even when the static images presented during training were identical to those presented to a group who observed non-diagnostic motion. In this case, object motion diagnosticity was defined in terms of qualitatively distinct motions arising from distinctly oriented paths through an appearance space of complex stimuli. In Experiment 2, we find that diagnosticity as defined by the direction of motion along paths of the same orientation in appearance space is not sufficient to induce the RT differences we had observed previously. In this case, object motion across category was qualitatively very similar, only differing in the sign of the flow field arising from object deformation.

Taken together, these two results tell us several useful things about the relationship between observed object motion and representations of object form. First of all, we must reject the notion that observers who see a dynamic object encode all the images in the sequence and maintain a full spatiotemporal volume of object appearance. If this were the case, we should never see differences between groups in either one of our experiments, since the static contents of training were identical across conditions in each task. Second, the particular direction of image change along a path in appearance space has little impact on form encoding. That is to say, the difference between forward motion and its reverse is essentially nil in this context. Qualitatively distinct motions between categories are required to cause a difference in static image processing.

This last observation puts an important constraint on the features of object motion that influence task performance in Experiment 1. As we have already mentioned, a feature like the optic flow field defined by two successive images is not likely to be relevant to this task as it is classically defined. The sign of the flow vectors across the flow field must not be relevant to this task, or else Experiment 2 would have yielded results similar to Experiment 1. Perhaps it is only the pattern of flow vector magnitudes that is

relevant, or some more general measure of variance in image space that is symmetric with respect to time, and thus essentially "non-diagnostic" under the conditions of Experiment 2.

We close by suggesting that a useful way to discuss the effect observed in Experiment 1 may in terms of a model for extracting "common" and "relative" object components for recognition. The decomposition of visual stimuli into components that are shared and the resulting residual components has been a fruitful model for both the perception of motion and surface reflectance (Bergstrom, 1977; Johannson, 1950). To our knowledge, such an analysis has not been carried out in the domain of object perception and recognition. Interpreting our results in this framework, "non-diagnostic" object motion may provide a strong "common" component from which a good representation of form might be extracted as a relative component. "Diagnostic" motion may not allow such a useful decomposition to proceed, leading to a weaker representation of form that does not support efficient classification during our test period. Applying vector analysis to real images may yield many interesting insights regarding dynamic object perception.

Conclusions

We have observed that the observation of diagnostic object motion during training can affect static classification performance at test. Our results suggest that the relevant processes relating object motion to object form are time-symmetric, and that observers do not perfectly encode static form following dynamic training. While further work is needed to elucidate the interaction between motion and form in this context, a vector analysis decomposition of dynamic objects may be an useful model for future study.

Acknowledgments

The authors would like to thank Mike Mangini and Irv Biederman for making the blob stimuli available to us. Dave Cox and Jim Dicarlo also made many helpful suggestions. BJB was supported by a NDSEG Fellowship.

References

Bergstrom, S. S. (1977). Common and relative components of reflected light as information about the illumination, color, and three-dimensional form of objects. *Scandinavian Journal of Psychology, 18*, 180-186.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433-436.

Burton, M. A. (1999). Face recognition in poor quality video. *Psychological Science, 10*, 243-248.

Christie, F., & Bruce, V. (1988). The role of dynamic information in the recognition of unfamiliar faces. *Memory and Cognition, 26*, 780-790.

Cox, D. D., Meier, P., Oertelt, N., & DiCarlo, J. J. (2005). 'Breaking' position-invariant object recognition. *Nature Neuroscience, 8*(9), 1145-1147.

Harman, K. L., & Humphreys, G. W. (1999). Encoding regular and random sequences of views of novel three-dimensional objects. *Perception, 28*, 601-615.

Hill, H., & Johnston, A. (2001). Categorizing sex and identity from the biological motion of faces. *Current Biology, 11*, 880-885.

Johannson, G. (1950). Configurations in the perception of velocity. *Acta Psychologica, 7*, 25-79.
Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics, 1*, 201-211.
Knappmeyer, B., Thornton, I. M., & Bulthoff, H. H. (2003). The use of facial motion and facial form during the processing of identity. *Vision Research, 43*, 1921-36.
Knight, B., & Johnson, A. (1997). The role of movement in face recognition. *Visual Cognition, 4*, 265-273.
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*(2), 335-355.
Kozlowski, L. T., & Cutting, J. E. (1977). Recognizing the sex of a walker from a dynamic point-light display. *Perception & Psychophysics, 21*, 575-580.
Lander, K., & Bruce, V. (2000). Recognizing famous faces: Exploring the benefits of facial motion. *Ecological Psychology, 12*, 259-272.
Lawson, R., Humphreys, G. W., & Watson, D. G. (1994). Object recognition under sequential viewing conditions: Evidence for viewpoint-specific recognition procedures. *Perception, 23*, 595-614.
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*(7), 745a.
Newell, F. N., Wallraven, C., & Huber, S. (2004). The role of characteristic motion in object categorization. *Journal of Vision, 4*(2), 118-129.
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.
Pike, G. E., Kemp, R. I., Towell, N. A., & Phillips, K. C. (1997). Recognizing moving faces: The relative contribution of motion and perspective view information. *Visual Cognition, 4*, 409-437.
Schiff, W. (1986). Recognizing people seen in events in dynamic "mug shots". *American Journal of Psychology, 99*, 219-231.
Spencer, J., O'Brien, J., Johnston, A., & Hill, H. (2006). Infants' discrimination of faces by using biological motion cues. *Perception, 35*(1), 79-89.
Stone, J. V. (1998). Object recognition using spatiotemporal signatures. *Vision Research, 38*, 947-951.
Thornton, I. M., & Kourtzi, Z. (2002). A matching advantage for dynamic faces. *Perception, 31*(113-132).
Thornton, I. M., Vuong, Q. C., & Bulthoff, H. H. (2003). A chimeric point-light walker. *Perception, 32*(3), 377-383.
Vuong, Q. C., & Tarr, M. J. (2004). Rotation direction affects object recognition. *Vision Research, 44*(14), 1717-30.
Wallis, G., & Bulthoff, H. H. (2001). Effects of temporal association on recognition memory. *Proceedings of the National Academy of Sciences, 98*(8), 4800-4804.
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), 1568-1574.

Section 2 – Encoding the appearance of a moving object

In this second section of the thesis I investigate how the motion of an object affects an observer's memory for the appearances observed during exposure. That is, how does the motion of an object affect what we remember of its appearance? This is a distinct question from the studies of discrimination and generalization I have described in Section 1, insofar as I presently address the issue of how strongly individual appearances are encoded rather than how well a given code can be applied to some task. In further contrast to the experiments already described, I work solely with rigid rotation in these studies. The advantage this offers is the ability to completely characterize the state of each stimulus in terms of its true form, position, and orientation in all phases of exposure and testing. This allows us to conduct thorough parametric analyses that were not realizable given the less constrained nature of the stimuli used in Section 1.

In Chapter 4, we carry out just such an analysis via a thorough investigation of the relationship between object motion and the fidelity of immediate recall for object appearance. Using a forced-choice recall task, we characterize the influences of object orientation, sequence smoothness, object speed, and sequence predictability on the encoding of individual appearances. The result is a quantitative model of dynamic object perception that can be applied to arbitrary objects.

Finally, in Chapter 5 I ask whether or not so-called "canonical views" are formed immediately after brief exposure to a moving object. In this study, I use simple "paperclip" objects for which full ground-truth structural data can be obtained readily, allowing for the construction of a plausible *a priori* model of view canonicity dependent on the relationship between 3-D form and projected 2-D appearance. I characterize view canonicity around a densely sampled viewing circle using three distinct behavioral paradigms, and ask in each case how well both static and dynamic implementations of the pre-determined canonicity model can be used to fit the data. I find that privileged views emerge in a systematic fashion in each behavioral setting, but that these measures are highly task-dependent. However, the canonicity model I develop proves quite capable of modeling observers' judgments, and benefits substantially from the inclusion of dynamic information. This latter point suggests that canonicity depends directly on object motion, rather than solely on form. Following this chapter, I present a concluding chapter summarizing the contributions of the thesis.

Introduction to Chapter 4

Chapter 4 describes a comprehensive set of studies designed to characterize the effects of recent perceptual history on the strength of encoding a single static image. As in Chapters 1 and 2, an image-level discrimination task is used to examine how dynamic stimulus characteristics affect static form processing in the absence of high-level cognitive mechanisms. However, in Chapter 4 I am no longer concerned with the ability of an observer to notice the difference between two stimuli. Instead, these tasks are designed to examine the observers' ability to faithfully encode just one stimulus subject to various manipulations of stimulus dynamics.

Object motion and the immediate recall of object appearance

Abstract

We investigate how object motion affects the fidelity of immediate recall for object appearance. Observers viewed complex three-dimensional objects undergoing rigid rotation and were asked to report the last image presented in a sequence. Our results support three main conclusions: First, recall errors in this task reflect uncertainty in the estimate of object appearance rather than uncertainty in an estimate of object position in space. Second, coherent object motion causes observers to maintain a "running average" of object appearance, inducing a bias in recall errors towards images appearing before the target. Finally, absolute error (which disregards the distinction between past and future images) is almost entirely determined by target/distractor similarity and the presentation time of the target. These results are discussed in the context of previous work regarding "representational momentum" and a preliminary model is advanced for predicting recall errors for images in dynamic sequences.

Introduction

The natural visual world is in a constant state of flux. This requires the successful observer to carry out important computations online, interpreting new data and acting on it as soon as it arrives. Dynamic visual input contains multiple cues for a wide range of ecologically relevant tasks and estimating the future state of the world given the recent past is unquestionably valuable. This latter task is especially relevant for object perception insofar as successful interaction with a dynamic object depends critically on both the ability to determine what will happen next and the ability to determine what is happening now.

During natural viewing, we almost always view objects in some sort of dynamic context, yet we know very little about how that context affects our perception. How does the local neighborhood of images in time affect our memory for an individual image in a sequence depicting object motion, for example? Does the predictability of appearance within a sequence make it easier or harder to correctly recall what was seen at a particular point in time? If the immediate past affects perception of the present, how far back in time does that influence reach? These apparently simple questions have yet to be answered completely. While various aspects of object motion and object perception have been investigated, we argue that at present we lack a thorough understanding of how aspects of object dynamics affect the encoding of object appearance. We continue by briefly reviewing some key results relevant to this broad topic, and discuss how our current experiments contribute to our understanding of dynamic object perception.

Representational Momentum

The majority of work relevant to the relationship between object motion and memory for object appearance is dedicated to investigating the nature of so-called "representational momentum." Representational momentum (RM), by analogy to physical momentum, suggests that moving visual stimuli have inertia in appearance space (Anstis & Ramachandran, 1987; Freyd, 1983; Freyd & Finke, 1984). That is, if an observer is watching a moving stimulus (say a dot that is translating to the right) that is suddenly

stopped mid-sequence, the observers' memory for the dot's position will be biased further along in the direction it was moving. This bias can be revealed in a variety of ways. One can ask the observer to report where the dot was last seen, for example. Rather than asking the observer to estimate the position him or herself, one can also present the observer with candidate positions that are to be accepted or rejected. A greater acceptance rate for "advanced" items as opposed to "delayed" items could indicate the presence of RM, as could a faster response time for the rejection of "delayed" items. These measures, as well as others, have been used by many researchers to investigate the characteristics of representational momentum.

Since the initial reports of RM for simple moving objects, there have been a wide variety of studies examining different aspects of the phenomenon (Thornton & Hubbard, 2002). An exhaustive list detailing all the possible factors that can affect RM is beyond the scope of this paper. Instead, we list below a subset of results that are particularly relevant for the current study:

1) RM is characterized by errors of immediate recall biased towards projected visual outcomes. (Freyd & Finke, 1984)
2) The gap between the ending of a sequence and the test phase critically affects RM. A short gap leads to "forward displacement" errors while a longer gap leads to the exact opposite. (Freyd & Johnson, 1987)
3) Greater object speed leads to more pronounced RM, in keeping with the physical analogy. (Freyd & Finke, 1985)
4) RM can be reduced or eliminated by varying the final stimulus in experimental sequences in an unpredictable way. (Kerzel, 2002)

The existence of RM suggests that object motion can directly affect immediate recall for object appearance in a simple way. Object motion is automatically used to generate predictions, such that immediate recall is biased towards an extrapolated future percept. Over several seconds, the influence of the prediction is gradually overwhelmed by the influence of past stimuli. The first part of this proposal is very similar to typical descriptions of the "flash-lag" effect, in which a flashed stimulus "lags" a persisting dynamic stimulus potentially due to continuous predictive updating of the dynamic object's appearance (Nijhawan, 1994). The generality of the proposed mechanism underlying both phenomena is appealing. Both RM and the flash-lag effect have been observed in diverse scenarios , making it tempting to suppose that the main relationship between object motion and object perception is predictive. This basic proposal is satisfying, but still leaves open several issues that we need to address.

First, we point out that both RM and flash-lag experiments tend to confound object position with object appearance, making it difficult to determine exactly what is being predicted during exposure to a dynamic stimulus. For example, the first RM experiments used a rotating rectangle as the dynamic input. Similarly, the first report of the flash-lag effect used a rotating line. This common thread has led to fruitful work suggesting unified models of both phenomena (Musseler, Stork, & Kerzel, 2002), yet in both cases, we cannot say whether observers are updating positions or appearances since those two measures are perfectly correlated. With only a few exceptions, localization is the focus of most RM and flash-lag research, leaving open a very important question for

understanding the perception of dynamic 3-D objects: If object motion impacts perception, is it true world motion that is most relevant or motion through appearance space? For complex objects that self-occlude as they rotate, position change and appearance change are not interchangeable, making this an important issue.

Second, it is also unclear what the relevant time scale for prediction is in either RM or the flash-lag effect. That is, how far back in time does the influence of perceptual history over current perception reach? The flash-lag effect has previously been thought of as a by-product of neural transmission delays, which would suggest that prediction should be governed by a fixed time interval. However, in both phenomena the instantaneous rate of position and/or appearance change that an object experience may also be relevant. This possibility has not been explored in a quantitative manner as of yet.

Overall, while both RM and the flash-lag effect suggest that object motion can affect immediate recall for object appearance, scaling up the proposed mechanisms to incorporate the perception of complex objects will require more work. In particular, teasing apart the contributions of position change and appearance change to the ultimate encoding of a particular image is necessary for understanding natural object perception.

Representation and recognition of complex moving objects

Independent of the RM and flash-lag paradigms, there are multiple experiments designed to probe the representational content of dynamic stimuli. That is, what do we actually extract from experience with a moving object? These studies typically use priming tasks to determine whether or not an apparent motion sequence can prime images that did not actually appear in the stimulus (Kourtzi & Shiffrar, 1997; Kourtzi & Shiffrar, 1999a; Kourtzi & Shiffrar, 1999b; Kourtzi & Shiffrar, 2001). For example, does a two-frame stimulus depicting an object at a 90-degree and 180-degree pose prime an image at the 135-degree position? If so, we can be confident that *interpolation* is a key feature of dynamic object perception. Given that same stimulus, is a 45-degree stimulus also primed? What about a 225-degree stimulus? The answers to these two questions help determine whether or not *extrapolation* occurs in these more complex settings. Overall, the key findings from work with novel and familiar objects undergoing both rigid and non-rigid motion support interpolation far more than extrapolation (Kourtzi & Nakayama, 2002). In these settings, object motion appears to mostly serve as a means for filling in the gaps between stimuli that are actually displayed. This seems counter-intuitive given our discussion of RM, but perhaps the distinction between apparent and "real" motion is responsible for the differing results. Reconciling these findings is an important task for advancing our understanding of dynamic object perception.

Finally, we briefly discuss what is currently known about object motion and its influence on the recognition of static stimuli. While there is a great deal of work concerning various ways object motion may serve as a feature for recognition (Spencer, O'Brien, Johnston, & Hill, 2006; Stone, 1998; Stone, 1999; Vuong & Tarr, 2004; Wallraven & Bulthoff, 2001), we focus here on the issue of whether or not the observation of coherent motion leads to significantly better recognition than exposure to the same set of images in an unordered sequence. The answer to this question is unfortunately very complex at present. An advantage for dynamic stimuli has been reported in several

different tasks (Lander & Bruce, 2000; Lander, Christie, & Bruce, 1999; Lawson, Humphreys, & Watson, 1994; Thornton & Kourtzi, 2002) , but an advantage for the randomized stimuli has also been found (Harman & Humphreys, 1999). Furthermore, while temporal proximity appears to induce automatic image binding over time (Wallis, 1996; Wallis, 1998; Wallis & Bulthoff, 2001), pure structural similarity also appears sufficient to induce the same binding to a lesser degree (Perry, Rolls, & Stringer, 2006).

A parametric study of temporal factors on complex object encoding

In the present study, our goal is to develop a clear understanding of how object dynamics affects immediate recall for appearance. To do this, we will attempt to steer clear of established paradigms like RM and the flash-lag effect. The reason for this is that we want to provide a comprehensive set of results concerning object motion and appearance encoding that is independent of the conflicting methodologies we have mentioned above.

In particular, we are interested in the following questions:

1) Do errors of recall reflect uncertainty in position or appearance?
2) Do local changes in motion affect recall when global motion is preserved?
3) How does complete sequence randomization affect recall?

To answer these questions, we adopt a simple protocol in which observers view a dynamic sequence and then report the last image they saw in a forced-choice task. Throughout, we shall manipulate the sequences our observers see in order to gain insight into the questions we raise above. This strategy allows us to hold test conditions constant during all phases of our experiments, while varying the parameters of our motion sequences as much as we wish. This facilitates comparison across different experiments, making it far easier to discuss a unified mechanism governing behavior in this setting.

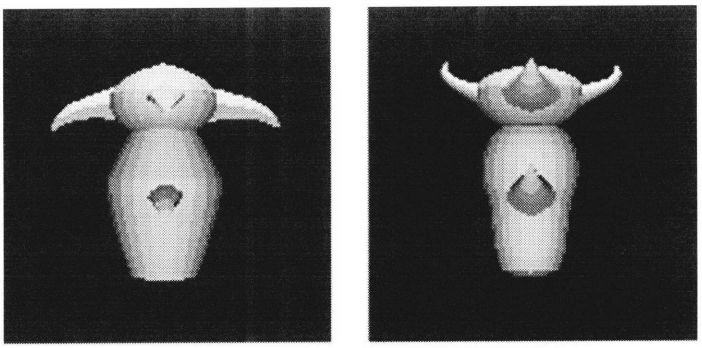
Experiment 1

Our goal in Experiment 1 is to determine whether recall errors for complex dynamic objects reflect observers' uncertainty in positional estimates or appearance estimates. To examine this, we ask observers to view sequences of novel, complex objects undergoing rigid rotation. Our objects are non-uniform, meaning that constant change in angular position does not lead to constant change in appearance. Instead, object appearance varies substantially over the course of the rotation sequence while angular velocity is fixed.

Methods

Subjects – 22 participants (M male, F female, average age ~23 years) were recruited for Experiment 1. All observers were naïve to the purposes of the experiment, and reported normal or corrected-to-normal vision.

Stimuli – Two novel objects were used to generate image sequences. The objects used to construct these short movies were taken from the publicly available "Greeble" stimulus set (Gauthier & Tarr, 1997) and are shown from a frontal view in Figure 1. These objects were selected because their structure is simple, yet provides non-uniform foreshortening and occlusion of object parts during uniform rotation.



Greeble #1 Greeble #2

Fig. 1 – The two "Greebles" used to create simple image sequences of rigid rotation. These items were randomly selected from a larger set of objects.

Each object was rendered while rotating about two axes simultaneously. Rotation about two axes means that at each time step, the object is first rotated about one axis by 12 degrees and then rotated about the second axis by the same amount. Object 1 was rotated a full 360 degrees about its X and Z-axes, starting from an upright, frontally viewed position. Object 2 underwent a full 360 degree rotation about the Y and Z-axes. In both cases, the full animation was rendered using 30 frames. XZ and YZ rotations were used because they produce different degrees of foreshortening and self-occlusion during rotation. Specifically, in one case an extreme "end-on" view of the object is obtained during rotation while in the other sequence one obtains an extreme "side-on" view. "Keyframe" depictions of both rotation sequences are displayed in Figure 2.

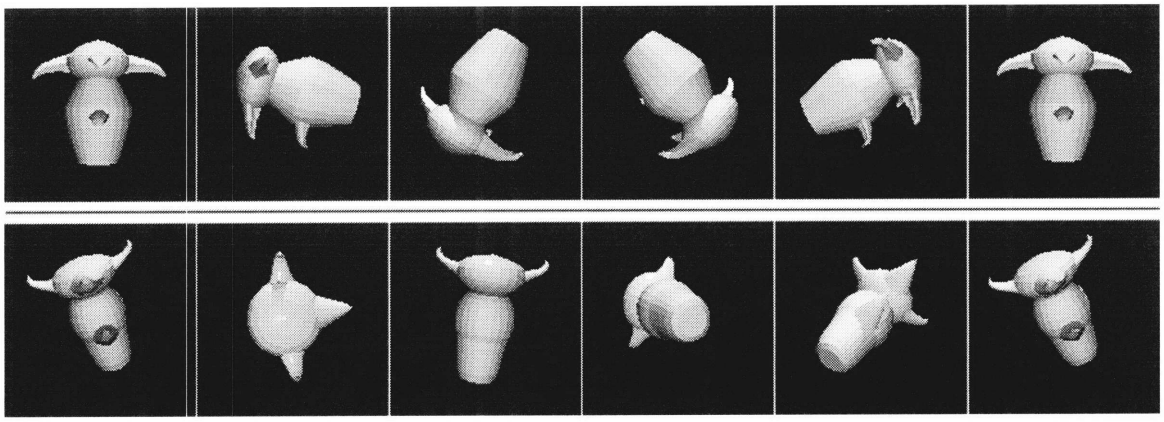


Fig. 2 – "Keyframes" depicting Object1 (top) and Object 2 (bottom) undergoing XZ and YZ rotation, respectively.

Grayscale Greeble images were rendered against a black background under a single point-light source using PovRay v3.5 for Windows. Each image was 123x123 pixels in size.

Procedure

In Experiment 1, observers were asked on each trial to view a short sequence containing multiple images of either Object 1 or Object 2 and remember the last image in the sequence. Each sequence began with the same initial frame and was played at a rate of 12 frames/sec.

Each sequence ended with a 250ms blank period followed by the presentation of a 1/f grayscale noise mask for 100ms. After an additional 2000ms delay period, observers were presented with 9 images of the object arranged in a 3x3 grid. (Figure 3) One of these items was the final image in the preceding sequence, while the other 8 images were distracters. Distracters were always the 8 images closest to the target image in terms of 3-D orientation of the object. The position of the target and distracter images within the grid was randomized on each trial. Observers indicated the position of the target using the numeric keypad and had unlimited time to generate a response. Response time and accuracy were recorded.

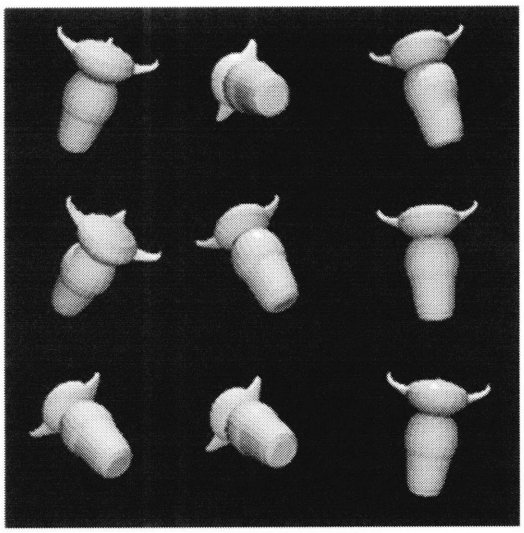


Fig. 3 – An example of the test displays presented to the participants following the observation of an image sequence. One of these images depicts the last image presented in preceding sequence, while the other 8 are its closest neighbors in terms of 3-D orientation.

Our primary manipulation is to vary what frame within the sequence serves as the target frame from trial to trial. Out of the 30 possible images of each object, the 20 images in the middle of the sequence were selected as targets. Each sequence thus began on frame #1, but could end on any frame between #6 and #25, inclusive. Each of these targets was assigned a unique and fixed set of 8 distracters, which were those images "closest" to the target image in terms of 3-D orientation of the object. Though the relative positions of targets and distracters are identical across different targets, target/distracter similarity in appearance space varies substantially.

Each observer carried out 12 trials for each of the 20 target images, for a total of 240 trials per session. Participants viewed the stimuli in a brightly lit room on a 19" Dell

monitor. Observers were seated comfortably approximately 50cm from the display, with no constraints on head or eye movement. All stimulus display and response collection routines were executed with the MATLAB Psychophysics Toolbox for Windows (Brainard, 1997; Pelli, 1997). Observers typically completed the task in about half an hour and were compensated for their time.

Results

For each observer, we calculate the mean absolute error for each target frame. We define error in terms of the increments in angular position that were used to generate the full sequence. This means that on each trial, observers can either pick the correct target (error = zero) or pick one of the distracters in the set (error less than or equal to 4 increments). For the moment, we do not differentiate between the selection of a distracter that appeared prior to the target during the dynamic stimulus and the selection of a distracter that was "in the future."

We define error in this way to make it immediately clear whether recall errors are a function of positional uncertainty or appearance uncertainty. If observers make errors in target selection due to noise in their estimate of 3-D orientation, we should expect a flat error function across all target frames. This is because the positional similarity between targets and distracters does not vary across target frames. If however, this task is carried out via estimates in 2-D appearance, we should expect error to fluctuate significantly across different target frames. In Figure 4, we display plots of mean absolute error across target frames for both Object #1 and Object #2.

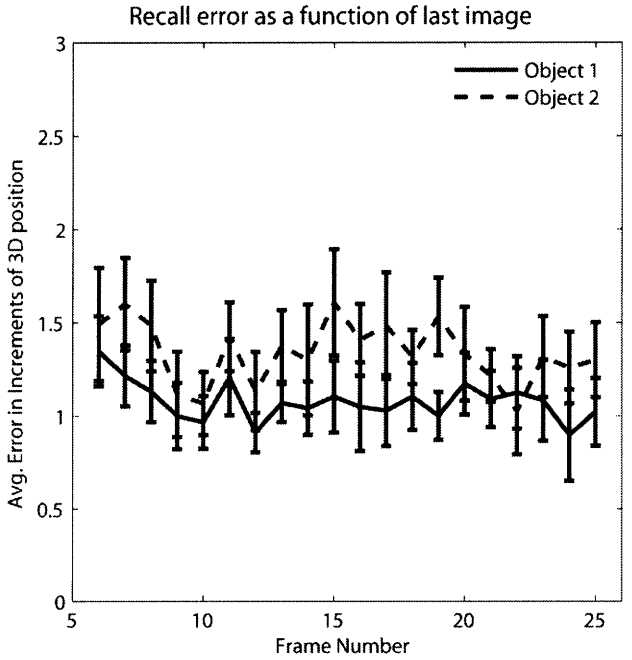


Fig. 4 – Recall error as a function of target image. The x-axis denotes the position of the final frame in the sequence and the y-axis denotes the average error across subjects in increments of 3-D orientation. There is a significant effect of image here, indicating that error is dependent on appearance rather than position.

We used a 2x20 mixed-design ANOVA (with object as a between-subjects factor and target frame as a within-subjects factor) to determine if the differences in error for various target frames were significant. We find a significant effect of target frame ($F(19,20)=1.78$, $MSe=0.176$, $p=0.024$), but no significant effect of object ($F(1,20)=1.37$, $MSe=5.46$, $p = 0.26$). The interaction between object and target frame was also not significant ($F(19,20)=0.721$, $p = 0.80$).

Discussion

Experiment 1 demonstrates that observers do not make errors in immediate recall based on a noisy estimate of position, but instead make smaller or larger errors depending on the similarity target and distracter appearance. This intuitive result is an important first step in developing a coherent theory of dynamic object perception. We have found good evidence that the relevant space for processing dynamic object input is an appearance space rather than a positional space. While this may not be surprising, selecting the right representational space for dynamic objects is the cornerstone of a comprehensive theory.

Now that we have reason to think errors in this task are tied to appearance rather than position, we continue by asking if variations in sequence dynamics can modulate the error function. We shall investigate this possibility by carrying out a simple manipulation on our original object sequences that drastically alters motion on a small time scale, while leaving the large scale movement of the object more or less intact. Errors at each target frame in this manipulated sequence will be directly compared to errors at the same targets when they are part of the original smooth sequence.

Experiment 2

This second experiment serves two important purposes. First of all, it is an important control for the results in Experiment 1. At present, we cannot say whether the motion observed on each trial contributes to the error function at all, or if the conditions at test are responsible for the fluctuations we observe in recall error. Similarly, since each target frame appeared at its own unique timepoint in the original sequence, the length of the sequence up to each target frame is a confounding factor. By manipulating the sequence but preserving the target/distracter sets, we can determine the extent to which object motion modulates error when these conditions are matched. Second, an effect of the manipulated sequence on recall accuracy would suggest that object motion over relatively short time scales is more relevant for immediate encoding than motion over longer intervals.

Methods

Subjects – An additional 16 participants (M male, F female, average age ~25 years) were recruited for Experiment 2. All observers were naïve to the purposes of the experiment, and reported normal or corrected-to-normal vision.

Stimuli – The objects described above in Experiment 1 were also employed for this experiment.

Procedure

As in our first experiment, observers were asked on each trial to view a short sequence containing multiple images of either Object 1 or Object 2 and remember the last image in the sequence. The same 9AFC task described above was implemented again in this experiment, this time with only 10 target images (and associated distracters) for each sequence. The target images selected were the even numbered frames between #6 and #24, inclusive.

In this experiment, observers performed the recall task after viewing both the original smooth tumbling sequences and what we call a "locally scrambled" sequence. In the latter case, the initial ordered sequence of frames is scrambled by flipping the order of each pair of images. For example, if the smooth sequence [ABCDEF] is scrambled in this manner, it becomes [BADCFE]. The resulting sequence depicts an oscillatory "saw-tooth" movement in which the object smoothly rocks backward before jumping forward abruptly. At a long time-scale, object motion is roughly identical across the smooth and locally scrambled sequences. In both cases, the object begins in one position and tumbles around two axes until it returns to that initial pose. At small time-scales, however, the sequences are very different in that the scrambled sequence depicts the object constantly changing direction and angular speed.

Each observer carried out 12 trials for each of the 10 target images in both conditions, for a total of 240 trials per session. Viewing conditions, stimulus display and response collection procedures were all carried over from Experiment 1.

Results

As before, we calculate mean absolute error across target frames in terms of angular increments. Though we have seen in Experiment 1 that position is not the relevant variable in this setting, we will continue to use this error measure for ease of comparison across experiments.

In this task, we are most interested in whether our locally scrambled sequences induce a significant change in recall error when target frame is fixed. We do not have an *a priori* hypothesis as to whether we expect scrambling to make performance better or worse, nor do we characterize errors in terms of the temporal relationship between the true target and the selected distracter. In Figure 5, we display plots of mean error as a function of target frame for both objects in the smooth and locally scrambled conditions.

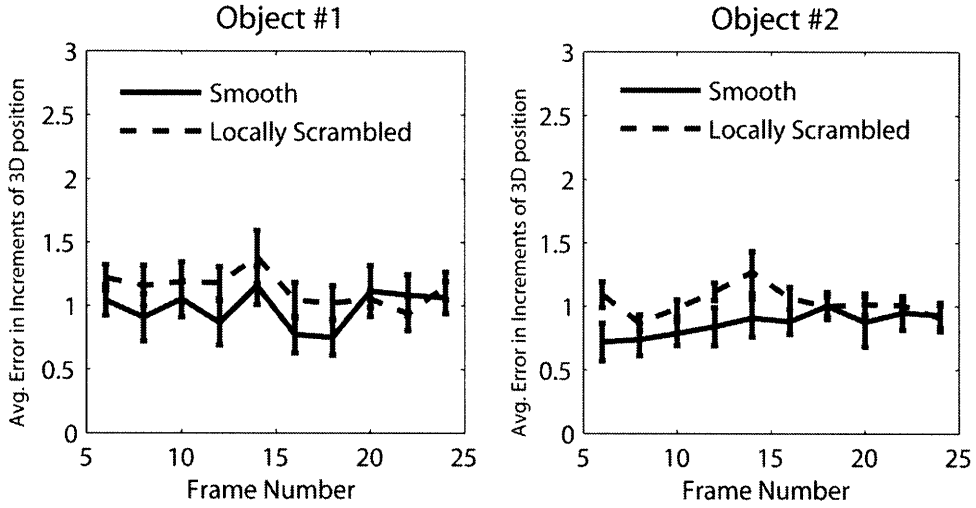


Figure 5 – Recall error as a function of final frame for both objects in the coherent and locally scrambled conditions. We see a main effect of the scrambling manipulation, indicating that appearance change over short time scales does affect recall.

Inspecting Figure 5, it appears that the locally scrambled sequences led to slightly greater amounts of error than the original smooth sequences. To determine if this was indeed the case, we carried out a 2x2x10 mixed-design ANOVA with object as a between-subjects factor, and target frame and sequence condition as within-subjects factors. We find a significant effect of sequence condition ($F(1,14)=6.38$, $MSe=0.326$, $p=0.025$), but no main effect of either target frame ($F(9,14)=1.25$, $MSe=1.50$, $p=0.273$) or object ($F(1,14)=1.65$, $MSe=0.59$, $p=0.22$). No interactions between these factors were significant.

Discussion

Experiment 1 demonstrated that positional estimates were an unlikely basis for recall errors in this task, but did not make a clear case for a contribution of object motion on the immediate memory for object appearance. The results of Experiment 2 however, provide good evidence for a relationship between the dynamics of a sequence and the encoding of its constituent frames. Specifically, we see that locally scrambling frames within a sequence leads to significantly larger errors in our recall task. This occurs despite the fact that the absolute time at which each target frame appears is matched across conditions, as is the set of distracters accompanying each target. The large-scale dynamics of the object are preserved as well, indicating that image change over short time intervals can have a significant effect on the fidelity of object appearance encoding. This result highlights the influence of object motion on the perception of appearance. Clearly the perceptual experience immediately preceding exposure to a stimulus can modulate the accuracy with which that stimulus is remembered.

We conclude by carrying out a final experiment designed to investigate two issues that neither Experiment 1 nor Experiment 2 address. The first of these is the question of how object speed affects recall error. We have seen in Experiment 2 that varying the amount of local image change between frames in our sequence appears to affect error in our task. An important question to ask then is whether sequences with different dynamic properties but identical local image differences gives rise to significantly different error

functions. Parametrically varying object speed achieves this goal, in that local image differences are preserved while the sequence itself is sped up or slowed down relative to some baseline. The second question we ask is the extent to which predictability plays an important role in determining the error at each target frame of a dynamic sequence. In both of our previous experiments the sequences observers viewed were highly predictable, even after local scrambling. This was especially so due to our selection of a common starting point for all sequences an observer viewed during a session. What happens to our error function then, if all predictability is removed? In our final experiment, we compare performance with the original smooth sequences to that obtained with a completely randomized sequence to answer this question. This last manipulation allows us to examine behavior in the absence of true object motion, but with all target/distractor relationships and other display properties preserved.

Experiment 3

Methods

Subjects – An additional 16 participants (M male, F female, average age ~25 years) were recruited for Experiment 3. All observers were naïve to the purposes of the experiment, and reported normal or corrected-to-normal vision.

Stimuli – The objects described in the two previous experiments were used in this experiment also.

Procedure
Our original 9AFC task was implemented once again in this experiment, again with only 10 target images (and associated distracters) for each sequence. The target images selected were the even numbered frames between #6 and #24, inclusive, as in Experiment 2.

To assess the role of speed and predictability on recall error in this final task, we altered our original object sequences in the following ways:

Speed – All sequences were played at three different speeds. In our slow, moderate, and fast conditions, each frame within a sequence was presented for approximately 120ms, 85ms, and 50ms respectively. Varying the speed of coherent sequences allows us to investigate the possible effects of greater image "momentum" on recall, while the same manipulation applied to randomized sequences helps us independently characterize the possibly deleterious effects of limited presentation time.

Spatial Coherence – Images within a sequence were either presented in their natural order (depicting smooth rotation) or in a randomized order independently determined online. This allows us to determine whether or not the observation of smooth, predictable motion has any effect on encoding. The randomized trials also provide us with what we consider a "pure" measure of the effects of target/distracter similarity on recall. Comparing accuracy over target frames across coherent and randomized sequences gives us the ability to see if the amount of encoding error is at all contingent on sequence predictability.

Finally, in this last experiment we no longer use a common starting point for all sequences. We also do not conflate target frame with sequence position as we have done in both of our previous experiments. The reason for this is that we wish to remove all cues to sequence length from both the smooth and randomized sequences. This is vital if we want to claim that the primary difference between the smooth and randomized sequences is the lack of temporal contingencies in the sequence. If we opted to re-use the same design we used in Experiments 1 and 2, observers would be able to use trial length to estimate target appearance for smooth sequence trials, but not for randomized sequence trials. Previously, this strategy was available to observers in all conditions, so we were not concerned about its use. Here, however, we need to take measures to ensure that any differences we see between smooth and randomized sequence display result from differences in predictability rather than the application of different encoding strategies.

On each trial, the number of frames contained in the image sequence was drawn from an exponential distribution with a mean of 20 frames. The exponential distribution was chosen because it has a constant hazard function, making it impossible for observers to guess the remaining length of an individual sequence given its current duration. This makes it unlikely that observers will "ramp up" attention at a certain point since they know a target must be imminent, and nullifies the validity of trial length as a cue for target appearance in both conditions.

Each observer carried out 9 trials each of 10 target frames, 3 speeds, and 2 coherence conditions for a total of 540 trials in a full session. The task was typically completed in an hour and observers were compensated for their participation.

Results

In Figure 6 we display mean recall error across subjects as a function of target frame for both smooth and randomized sequences of both objects at all three speeds. In answer to the questions we posed in this experiment, it looks as though both sequence speed and randomization have surprisingly little effect on performance.

To confirm this intuition, we carried out a 2x2x3x10 mixed-design ANOVA on the data. Speed, sequence type, and target frame were all within-subject factors, while object was a between-subjects factor. In the interests of clarity, we shall only describe the significant effects in the data rather than provide full details of each comparison. We observe main effects of speed ($F(2,14)=10.48$, $MSe=0.149$, $p<0.001$) and target frame ($F(9,14)=3.00$,$MSe=0.491$, $p = 0.003$). There were also marginally significant interactions between target frame and object ($F(9,18)=1.813$, $p=0.072$), as well as between sequence type and object ($F(1,18)=4.20$, $p=0.06$).

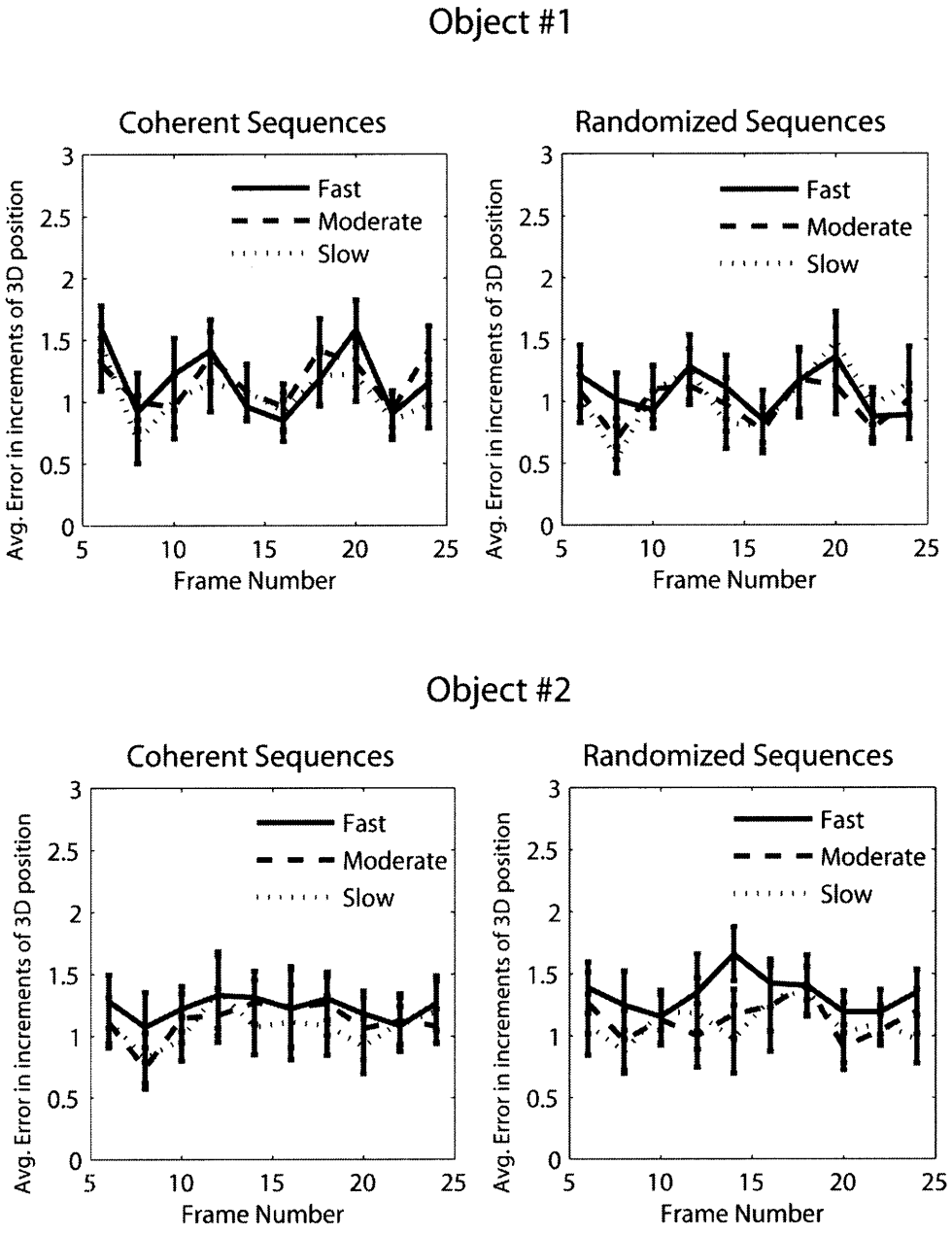


Fig. 6 – Recall error as a function of final frame for both objects at three speeds, in coherent and fully randomized sequences. There is a main effect of frame and speed, but no effect of randomization.

The effect of speed is significant, small in magnitude. This refutes our hypothesis that local image change is the sole determinant of recall error magnitude, suggesting instead that there are deleterious effects of increased speed in both smooth and randomized sequences. A simple model in which encoding noise is modulated directly by presentation time can account for this result. Furthermore, both the main effect of target frame and its interaction with object are compatible with our previous assertions that error is determined by target/distractor appearance similarity rather than position uncertainty.

It is very surprising, however, that there is little effect of randomization on error magnitude. Considering the results from Experiment 2, one might expect that the dramatic change in sequence dynamics brought on by full randomization should greatly disrupt performance. To the contrary, we find no main effect of sequence type at all and only a marginally significant interaction between sequence type and object. This latter result suggests that randomization does affect the error function, but in an object-dependent way. Though this does indicate a potential role of predictability in determining the magnitude of the error function, it also implies that the mechanism governing perception under these conditions is complex.

We conclude our analysis by examining the direction of error in this task, an aspect of our data that we have heretofore ignored. We have only considered absolute error up to this point since we are interested in the uncertainty of observers' appearance estimates rather than their sign, and also to distinguish our work from investigations of RM. However, since we are investigating the role of predictability in this last experiment, a brief examination of the direction of recall errors across conditions seems warranted. Prediction is obviously a directional phenomenon, and so additional differences between our smooth and randomized sequences might be evident if we include this information.

To that end, we define a bias term for each observer in each of the three speed conditions and both sequence types. This bias term reflects the number of errors that could be considered "predictive" based on their sign, normalized by the total number of errors each observer makes. The magnitude of each error does not contribute to the value of this bias term; it is only a proportion of predictive errors to total errors. A value of 0 would indicate that all errors were post-dictive, while a value of 1 would indicated that all errors were predictive. Given that the sign of errors is essentially meaningless in the randomized sequences, we expect that the bias term in all randomized conditions will be close to 0.5, while we may see a different value for the smooth conditions.

The bias term was calculated for all observers, and a 2x2x3 mixed-design ANOVA was run on these values. Speed and sequence type were within-subjects factors and object was a between-subjects factor. We find only a main effect of sequence type ($F(1,14)=15.84$, $MSe=0.0067$, $p=0.001$) with no other significant main effects or interactions. Figure 7 displays the average value of the bias terms across subjects, from which we see that observers' error bias is actually significantly less in the smooth case than the randomized case. This indicates that post-dictive errors are more prevalent in the smooth sequence than in the randomized case, implying that temporal contingency in the sequence induces a sort of retrograde mode of perception.

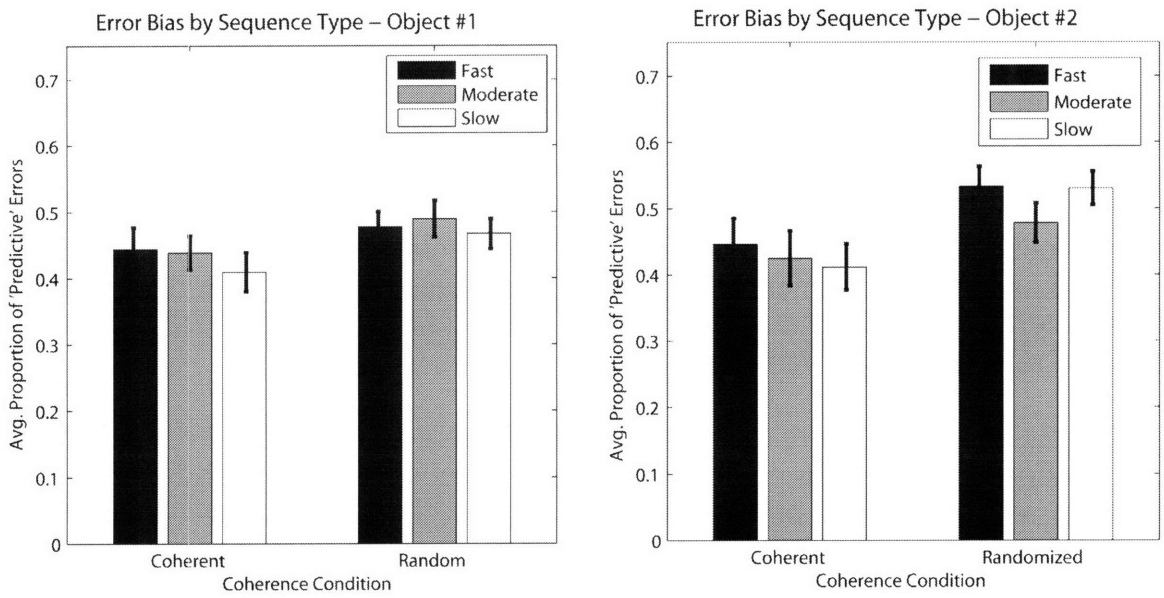


Fig. 7 – Average bias index across observers for all speeds and sequence types for both objects. Smooth sequences give rise to significantly lower bias values than randomized sequences.

Discussion

Speed and sequence randomization both appeared to have small effects on error in this task, but both factors either significantly or nearly significantly changed the error function for appearance recall. Increased speed leads to increased error, consistent with a model where increased presentation time narrows appearance uncertainty. Randomization appears to affect absolute error in a potentially complex way, but by expanding our analysis to incorporate the sign of errors made by observers, we found a clearer relationship between predictability and performance. A stong difference between performance in the smooth and randomized conditions was found in the proportion of post-dictive errors made by each observer. Predictability in the sequence seems to induce a bias for "backwards-averaging" similar to the U-shaped function reported in RM studies.

General Discussion

Overall, these results point to a relatively simple model of dynamic object perception that we can describe in some detail. We conclude by briefly outlining the major features of this model, as informed by the results of our three experiments.

Experiment 1 indicates that observers do not make errors based on a fixed level of uncertainty in object position, but likely do so based on a fixed level of uncertainty in appearance. The result is that varying target/distractor similarity at test causes varying amounts of absolute error across target frames. Our model must therefore operate in an appearance space rather than taking object positions as input. So far, this need not be a model of dynamic object perception since there is no need to account for anything besides the conditions during the static test.

Experiment 2 and Experiment 3 provide additional data that force us to incorporate aspects of object motion into our model, however. The effect of speed we observe in

Experiment 3 indicates that limited presentation time increases uncertainty for both sequence types. Thus, our model must be initialized with a baseline level of appearance uncertainty that can be modulated by presentation time. Though the inclusion of a speed (or more accurately, presentation time) term is a first step towards a model of dynamic perception, this is still essentially a static model since the temporal contingencies between frames do not contribute to the error calculation. This changes when we consider the observed effects of both local scrambling and randomization from our second and third experiments, however. In both cases, we have strong evidence that the order of images in the sequence affects recall error, forcing us to incorporate inter-frame relationships into the model.

We begin by noting that the role of temporal predictability can actually be stated fairly simply. As determined by our comparison between smooth and fully randomized sequences in Experiment 3, predictability leads to a slight retrograde bias in error. When there is no predictability there is no bias. This distinction between predictable sequences and unpredictable sequences can be modeled as a term that shifts the mean percept slightly towards the past by some amount if the input sequence is predictable. Can this shift explain the results obtained in Experiment 2? We suggest that it can. If we allow the magnitude of the shift brought on by sequence predictability to depend on the difference between the target frame and its immediate antecedent, this can explain the significant increase in error related to local scrambling. Specifically, let the perceived last frame in a predictable sequence be an interpolation between the true target and its predecessor. This directly introduces the retrograde bias observed in Experiment 3 and also predicts that large image differences between neighboring frames should lead to larger error rates independent of target/distractor similarity. This is exactly what we observe in Experiment 2, since local scrambling as we have defined it leads to our target frames being more different from their immediate antecedents than they were in the smooth sequence. An interesting prediction that follows from this model is that the effect of local scrambling should be nullified if we simply reversed the sequences, since this restores the image change accompanying target presentation to the same level as in the smooth sequence.

We close by describing our model formally, so that it can be tested against data obtained from arbitrary image sequences.

Given a predictable sequence of images represented in appearance space by [x_1, x_2... x_n] that are played to observers at speed *s*. The probability of selecting image *x* as the target when image x_n is the true target is given by:

$$p(x) = \frac{1}{\sqrt{2\sigma(s)\pi}} e^{\frac{\left\| x - \frac{T}{2}(x_n - x_{n-1}) \right\|^2}{2\sigma^2(s)}}$$

in which T is either 0 or 1 for unpredictable and predictable sequences respectively, and the value of sigma is given by:

$$\sigma(s) = a * s + \sigma_0$$

Determining an appropriate representation of the sequence images to produce the vectors in *x* is, of course, a non-trivial problem. However, once this is done, the only free parameters in this model are the baseline uncertainty in appearance (σ_0) and the coefficient that governs the influence of speed on net uncertainty (a). Both of these parameters can be set from the data obtained from fully randomized trials so that the model can be tested on predictable sequences. The result is an empirically-based quantitative model of dynamic object perception.

Conclusions

Object motion has direct consequences for the fidelity of object appearance encoding. Speed (or presentation time), predictability, and target/distracter appearance similarity all contribute to observers' ability to accurately report the images seen at the end of a dynamic sequence. Object position appears to be an inappropriate variable in modeling behavior in this setting, due to the non-linear relationship between object orientation and object appearance for complex 3-D forms. The model we propose for approximating behavior in this task assumes a level of uncertainty in appearance space that is modulate by presentation time, and has a bias towards past stimuli as determined by an interpolation between the true target and the penultimate image in the sequence. This bias is also modulated by the level of predictability in the sequence. The proposed model is an important first step towards a quantitative theory of dynamic object perception.

Acknowledgments

The author would like to thank the many brave subjects who watched zillions of greebles tumble through space for pay. Thanks also to my good friend, coffee. Two cats also contributed to the manuscript in ways that were subsequently corrected. BJB is a National Defense Science and Engineering Graduate Fellow.

References

Anstis, S., & Ramachandran, V. S. (1987). Visual inertia in apparent motion. *Vision Research, 27*, 755-764.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433-436.

Freyd, J. J. (1983). The mental representation of movement when static stimuli are viewed. *Perception and Psychophysics, 33*(6), 575-581.

Freyd, J. J., & Finke, R. A. (1984). Representational Momentum. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(1), 126-132.

Freyd, J. J., & Finke, R. A. (1985). A velocity effect for representational momentum. *Bulletin of the Psychonomic Society, 23*(6), 443-446.

Freyd, J. J., & Johnson, J. Q. (1987). Probing the time course of representational momentum. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*(2), 259-268.

Gauthier, I., & Tarr, M. J. (1997). Becoming a "Greeble" expert: Exploring the face recognition mechanism. *Vision Research, 37*(12), 1673-1682.

Harman, K. L., & Humphreys, G. W. (1999). Encoding regular and random sequences of views of novel three-dimensional objects. *Perception, 28*, 601-615.

Kerzel, D. (2002). A matter of design: No representational momentum without predictability. *Visual Cognition, 9*(1-2), 66-80.
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*(6), 461-466.
Kourtzi, Z., & Shiffrar, M. (1999a). Dynamic representations of human body movement. *Perception, 28*, 49-62.
Kourtzi, Z., & Shiffrar, M. (1999b). 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*(2), 335-355.
Lander, K., & Bruce, V. (2000). Recognizing famous faces: Exploring the benefits of facial motion. *Ecological Psychology, 12*, 259-272.
Lander, K., Christie, F., & Bruce, V. (1999). The role of movement in the recognition of famous faces. *Memory and Cognition, 27*, 974-985.
Lawson, R., Humphreys, G. W., & Watson, D. G. (1994). Object recognition under sequential viewing conditions: Evidence for viewpoint-specific recognition procedures. *Perception, 23*, 595-614.
Musseler, J., Stork, S., & Kerzel, D. (2002). Comparing mislocalizations with moving stimuli: The Frohlich effect, the flash-lag, and representational momentum. *Visual Cognition, 9*(1-2), 120-138.
Nijhawan, R. (1994). Motion extrapolation in catching. *Nature, 370*, 256-257.
Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision, 10*, 437-442.
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.
Spencer, J., O'Brien, J., Johnston, A., & Hill, H. (2006). Infants' discrimination of faces by using biological motion cues. *Perception, 35*(1), 79-89.
Stone, J. V. (1998). Object recognition using spatiotemporal signatures. *Vision Research, 38*, 947-951.
Stone, J. V. (1999). Object recognition: view-specificity and motion-specificity. *Vision Research, 39*, 4032-4044.
Thornton, I. M., & Hubbard, T. L. (2002). Representational Momentum: New findings, new directions. *Visual Cognition, 9*(1-2), 1-7.
Thornton, I. M., & Kourtzi, Z. (2002). A matching advantage for dynamic faces. *Perception, 31*(113-132).
Vuong, Q. C., & Tarr, M. J. (2004). Rotation direction affects object recognition. *Vision Research, 44*(14), 1717-1730.
Wallis, G. (1996). Using Spatio-temporal Correlations to Learn Invariant Object Recognition. *Neural Networks, 9*(9), 1513-1519.
Wallis, G. (1998). Spatio-temporal influences at the neural level of object recognition. *Neural Networks, 9*, 265-278.
Wallis, G., & Bulthoff, H. H. (2001). Effects of temporal association on recognition memory. *Proceedings of the National Academy of Sciences, 98*(8), 4800-4804.

Wallraven, C., & Bulthoff, H. H. (2001). *Automatic acquisition of exemplar-based representations for recognition from image sequences.* Paper presented at the CVPR 2001.

Introduction to Chapter 5

In Chapter 5, the relationship between object motion and the strength of encoding for individual images is again the subject of study. Here, however, the focus is shifted away from immediate recall for appearance in favor of longer-term memory for the constituent images in a dynamic sequence. Furthermore, I adopt a modeling approach in this last chapter to analyze the contributions of both object form and object motion towards predicting observers' ability to remember the images on display during exposure to a novel moving object.

Recovering canonical views of an object from dynamic input

Abstract

We examine how preferred views of a novel object's appearance might be selected online during exposure to a dynamic stimulus. For our problem domain, we select closed "paperclip" objects. Though they are highly unnatural stimuli, these objects allow us to define a reasonable definition of global canonicity that can serve as a standard for evaluating observer behavior in several tasks. We compare observers' explicit ratings of view canonicity to implicit measures of the same. We conclude that explicit ratings are well predicted by a form-based canonicity model that does not account for dynamic information, but implicit ratings are not. An alternative model based on the observed sequence of images is developed and compared to the form-based proposal using the implicit data.

Introduction

It is by now a well-established fact that observers do not recognize familiar objects in a completely view-invariant manner (Logothetis, Pauls, Bulthoff, & Poggio, 1994; Tarr & Bulthoff, 1995). Instead, it is usually the case that certain views of an object are recognized faster and more accurately than others (Palmer, Rosch, & Chase, 1981). These views are called "canonical views" to reflect that they are the "best" views of an object for recognition.

Previous studies have attempted to determine what makes a particular view of an object canonical, both through psychophysical and computational methods. Empirically, canonical views tend to limit foreshortening of the object along major symmetry axes (Humphrey & Joliceur, 1993; Lawson & Humphrey, 1996) (although some researchers have reported exceptions to this rule of thumb (Perret, Harries, & Looker, 1992)). A ¾ view, often from a slightly elevated position, tends to be selected very frequently by observers who are asked to produce the "best" image of an object (Blanz, Tarr, & Bulthoff, 1999). It has been suggested that this vantage point provides the most information about 3-D form for a wide range of natural objects viewed in typical settings. Another explanation is that such views tend to minimize image redundancies (such as symmetry), making them rich in information as defined within an information theory framework. Computationally, pose robustness has been suggested as a principle governing the selection of canonical views for arbitrary objects (Peters, Zitova, & von der Malsburg, 2002). Under this model the attributes of any individual image are irrelevant. Instead, it is the extent to which small perturbations of the objects position lead to substantial changes in 2D appearance that determines canonicity. A canonical view as defined by pose robustness is a view of the object that will change little when the object moves slightly, making it likely that this view might be selected via unsupervised clustering algorithms, for example.

There are several issues regarding the selection of canonical views left open by all of these studies. First of all, surprisingly little behavioral work has been carried out on observers' judgments of canonicity for views of unfamiliar objects. As a result, there is

not a great deal known about the acquisition of canonical views during object learning. Second, many proposals regarding the selection of privileged object views critically depend on knowledge of the full appearance space of an object, from which a canonical view is selected (Cutzu & Tarr, 1999; Murase & Nayar, 1995). This is especially true of the pose robustness proposal, but still an applicable criticism of most current hypotheses. In general, any model that proposes canonical views are selected by maximizing some image attribute over the space of possible object appearance must contend with the fact that observers only gain experience with novel objects through directed viewing sequences. That is, individual paths through appearance space are observed, but the full set of object appearances is generally unavailable to the observer. Maximizing a proposed image attribute is still a perfectly fine strategy, but it must be conceded that real observers will likely only find local maxima and be forced to make do with them. Third, specific proposals regarding view canonicity that can be expressed computationally are often difficult to test psychophysically. In many cases, this can be attributed to our lack of a good model for translating between perceptual sensitivity to image change and the raw image difference in pixels between two images. The result is that proposals regarding canonical view selection can only be tested in a qualitative way. Finally, the various means of behaviorally assessing view canonicity have not been rigorously compared to the best of our knowledge. In some cases, explicit judgments are requested from the observers. In others, implicit measures like response time and accuracy are used. It is an open question how much these choices affect the process of empirically determining canonical views of any object, familiar or unfamiliar. Moreover, do particular task demands change the results dramatically, or do the same privileged views inevitably surface?

In the current study, we attempt to address all of these issues with a set of experiments examining canonical view selection for novel objects. Rather than use complex, familiar objects, we have opted to use simple "paperclip stimuli" in all of our tasks (Bulthoff & Edelman, 1992; Edelman & Bulthoff, 1992). While these are highly unnatural stimuli, they offer several important advantages. First, there has been a great deal of previous work with these stimuli suggesting that observers recognize them in a view-dependent way, making them a good stimulus set for a study that critically requires some views to be more equal than others (Bricolo, Poggio, & Logothetis, 1997; Edelman, 1999; Logothetis, Pauls, & Poggio, 1995; Logothetis et al., 1994). They are also unfamiliar, providing us with the opportunity to study canonical view selection in its early stages. Additionally, they are simple enough in terms of their 3-D form that we can formalize a straightforward model of view canonicity that is both plausible and easily testable. Furthermore, individual images of these objects are depth-ambiguous under orthographic projects, making it more likely that canonicity will depend on viewing a directed sequence of appearance change. We suggest that this scenario, in which view selection must be carried out online during exposure to a moving object, more closely approximates the natural environment.

In three experiments, we measure view canonicity following exposure to rigidly rotating paperclip. We compare the results obtained from observers' explicit judgments to those obtained by using RT as a measure of canonicity in both 2AFC and 1FC tasks. Throughout, we compare our behavioral data to the predictions made by a form-based

model of canonical view selection, defined specifically for the stimuli under consideration. At each stage, we ask three questions:

1) Are consistent canonical views evident in the behavioral data?
2) If so, can we approximate the behavioral data with a purely form-based model?
3) Does the inclusion of dynamic information improve our model's goodness-of-fit?

We begin by examining observers' explicit selections of "optimal" views following brief exposure to novel paperclip objects.

Experiment 1

In our first experiment, observers were asked to make subjective judgments about image canonicity for novel paperclip objects after viewing short sequences depicting rigid rotation of the objects. These judgments were then compared to a baseline model of view canonicity we describe in more detail below.

Methods

Subjects

24 members of the MIT community volunteered to participate in this experiment. All observers reported normal or corrected-to-normal vision and were naïve to the purposes of the experiment. None of the observers had previous experience with paperclip objects resembling those used in the study.

Stimuli

We created four distinct paperclip objects in MATLAB by randomly selecting 8 points in spherical coordinate space that fell within a sphere of unit radius. These points were then joined together sequentially in random order to form a closed loop (so that line terminators were not a conspicuously salient feature). Each object was then rotated completely about its vertical axis in steps of 12 degrees, yielding a total of 30 images of each object. The objects were arbitrarily labeled A,B,C, and D, and are displayed in Figure 1. Each image was 344x344 pixels in size.

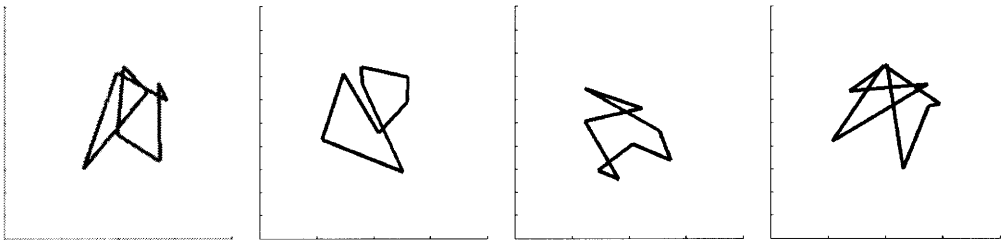


Fig. 1 – From left to right, Objects A,B, C, and D used in each of our three experiments.

Procedure

Each observer provided canonicity ratings for two objects, one after the other, with presentation order balanced across subjects. For each object, the experimental session began with a brief exposure period during which the object under consideration rotated rigidly about its vertical axis. During the exposure period, each object completed 8 full revolutions in which each of the 30 images comprising the full sequence of object views

was displayed. The frame rate of the monitor was 60Hz, and each sequence was played at a rate of 7.5 frames per second.

Following this exposure period, observers were asked to provide canonicity ratings for all 30 images that appeared in the training sequence. Participants were asked to use a 1-7 Likert scale, where "1" corresponded to a very poor view of the object and "7" corresponded to a very good view. Observers were instructed that they should rate the images according to their assessment of how "good" an image each object view was. To be more concrete, we suggested that they consider how likely they would be to show each image to a friend if they had to provide examples of what the training object looked like. During the rating task, each image from the training sequence was shown twice in randomized order. Each image remained on screen until observers provided a rating via the keyboard.

During both the exposure period and the rating task observers were seated approximately 50cm from the monitor in a brightly lit room with no restraints on head position or gaze. Each image subtended approximately 3 degrees of visual angle on screen. All stimulus display and response collection routines were executed with the MATLAB Psychophysics Toolbox for Windows (Brainard, 1997; Pelli, 1997). Observers typically completed the task in about 15 minutes and were compensated for their time.

Results

For each subject, the two ratings provided for each image were averaged, yielding 30 canonicity ratings for each object viewed. In Figure 2, we display the mean canonicity ratings for objects A, B, C, and D.

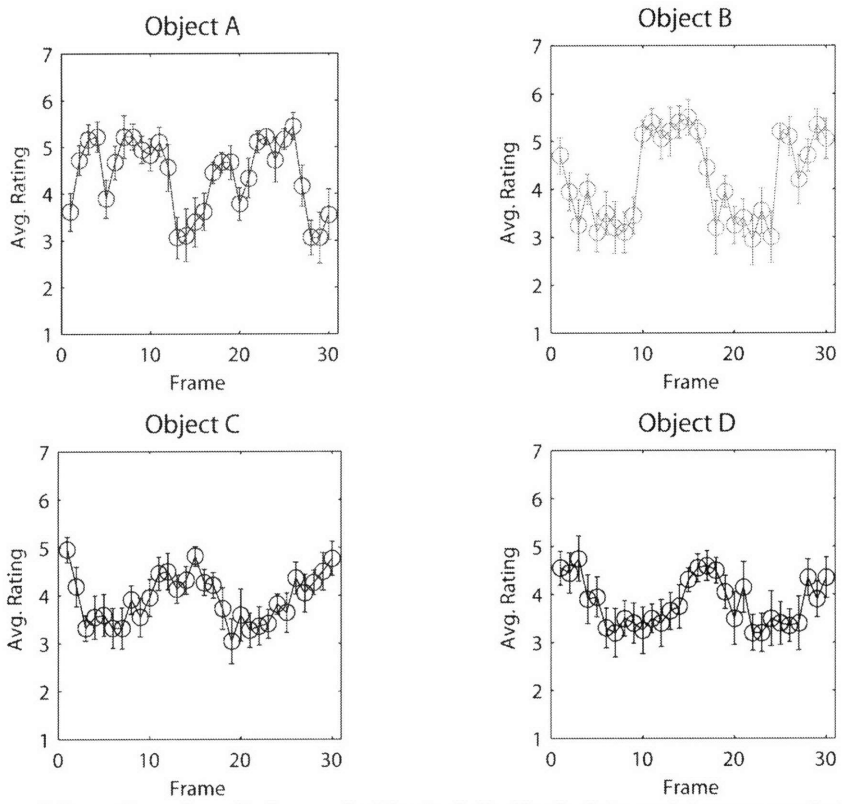


Fig. 2 – Mean canonicity ratings for all views of objects A-D. Each data point represents the average rating across 12 unique observers. Error bars are +/- 1 s.e.m.

It is evident in Figure 2 that observers are quite consistent in their ratings for these objects, leading to strong modes in the data for each stimulus. Specifically, each object appears to have two preferred views that are diametrically opposed on the viewing circle. These images will be mirror images of each other due to our use of orthographic projection in rendering the paperclip objects. It is also important to note that the location of these modes is not the same across objects, as we might expect if canonicity was being driven purely by primacy and recency effects. Instead, it is clear that each set of object ratings has a specific structure.

A baseline model of canonicity

We continue by attempting to model this data using a very simple form-based model of view canonicity for our paperclip stimuli. To motivate this model, consider the limiting case of a paperclip object with only one segment. This segment has a true length, and we propose that a "canonical" view of this minimal object is one in which the projected length is not much different than the actual length. An end-on view of the segment would be the worst view possible, while any view of the segment oriented in the fronto-parallel plane would be best. We can formalize this measure of canonicity for a single segment by dividing the projected length by the actual length to yield a value on the interval [0,1] describing the amount of foreshortening in the view.

To extend this intuitive measure to our 8-segment objects, we calculate this value for each segment in the object and take the mean over all values as our final estimate of canonicity for each view.

Therefore, if we have *n* pairs of 3-D points, (each point in the pair denoted as {x,y,z} and {x',y',z'}) the canonicity of a particular view can be calculated thusly:

$$canonicity = \frac{1}{n}\sum_i \frac{\sqrt{(x_i'-x_i)^2+(y_i'-y_i)^2}}{\sqrt{(x_i'-x_i)^2+(y_i'-y_i)^2+(z_i'-z_i)^2}}$$

This value also ranges between 0 and 1, inclusive. In Figure 3, we display the canonicity values calculated under this model for all views of our four test objects alongside the average observer ratings.

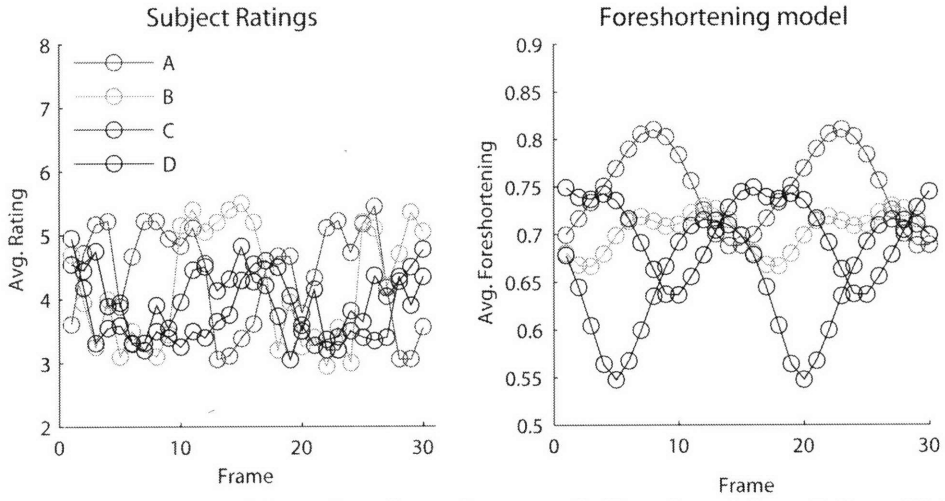


Fig. 3 – Average canonicity ratings from observers (left) and model predictions (right).

To assess the model's goodness-of-fit, we calculate the correlation coefficient between the actual data (using the mean across observers for each view) and the model's predictions. The results of this analysis are displayed in Table 1.

Table 1 – Correlation between baseline model and actual observer ratings.

	Prop. Variance	Significance
Object A	0.56	$p < 0.001$
Object B	0.04	n.s.
Object C	0.59	$p < 0.001$
Object D	0.56	$p < 0.001$

For three of our four objects, the baseline model provides a good fit to the data. Given the unconstrained and subjective nature of the task as well as the simplicity of the model, this is surprising. We note however, that for one of our objects (Object B) this model fails to capture any significant portion of the variance in the observed data. Examining Figure 3, we can see that this object's form-based canonicity function is very shallow compared to the others, suggesting that there is perhaps not enough variability across views to successfully capture the structure in the rating data. We continue by

introducing a slight modification of our original baseline model in which we incorporate the dynamics of our foreshortening measure. This allows us to determine whether or not the motion of our objects plays any significant role in the perceived canonicity of individual views.

To add a dynamic element to our original baseline model, we take the first and second derivatives of our original canonicity function and include these two functions as additional regressors in a multivariate linear regression model. If the inclusion of these dynamical terms improves the fit of our model to the data, we can tentatively conclude that aspects of object motion influence perceived canonicity. In Table 2, we report the results of our regression analysis.

Table 2 – Correlation between full dynamical model and actual observer ratings.

	Prop. Variance	Significance
Object A	0.57	$p < 0.001$
Object B	0.34	$p = 0.012$
Object C	0.76	$p < 0.001$
Object D	0.72	$p < 0.001$

In all four cases, we note that the proportion of variance captured by the model increases, to the extent that Object B is now reasonably well-fit by the model. In Figure 4, we display for each object the original ratings and the best-fit line of the baseline model and the full dynamical model in separate panels.

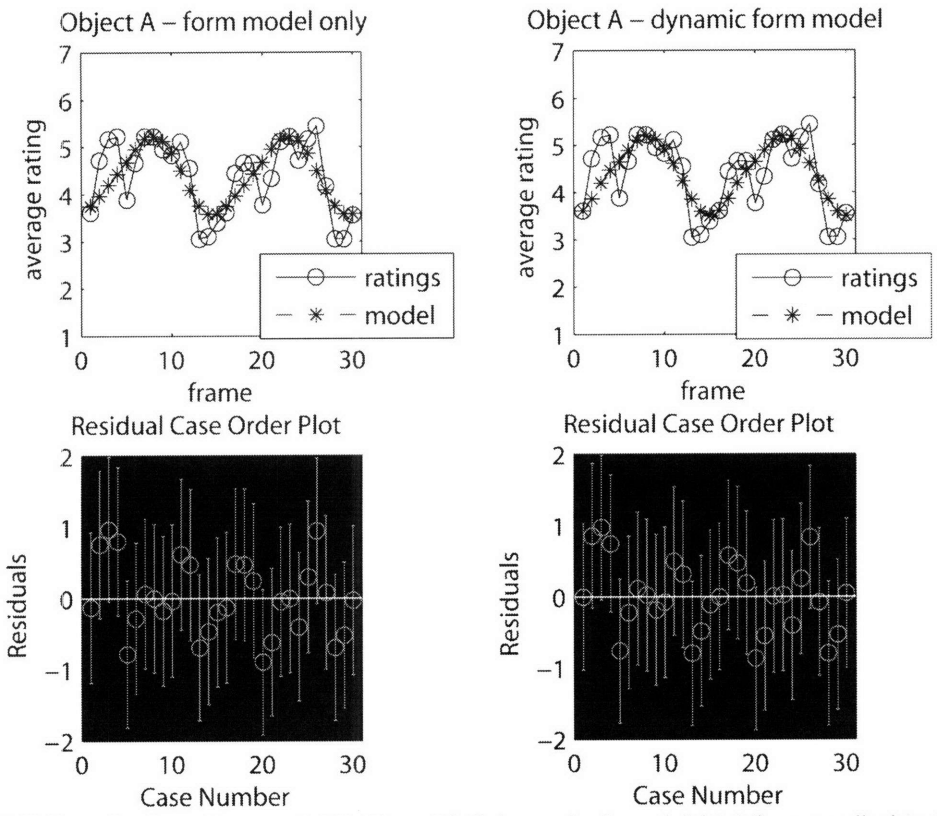


Fig. 4a – Best-fit lines for baseline model (left) and full dynamical model (right) as applied to Object A with the accompanying residual analysis below each plot. In each case, the dynamical model provides a better fit to the data.

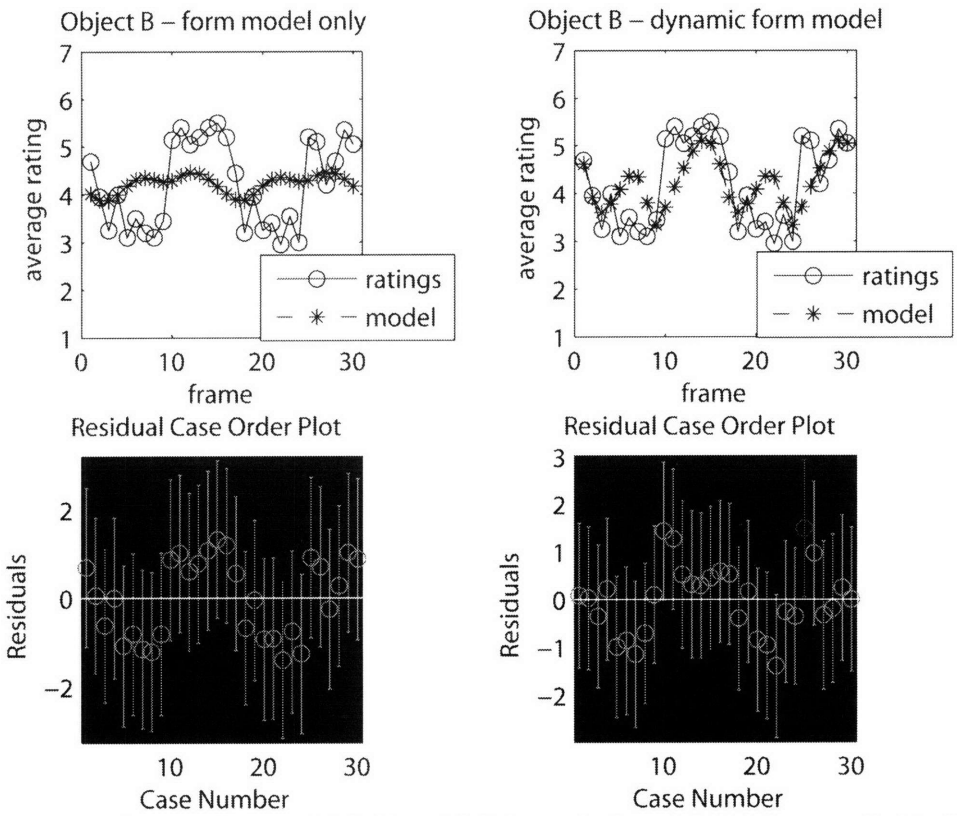


Fig. 4b – Best-fit lines for baseline model (left) and full dynamical model (right) as applied to Object B with the accompanying residual analysis below each plot.

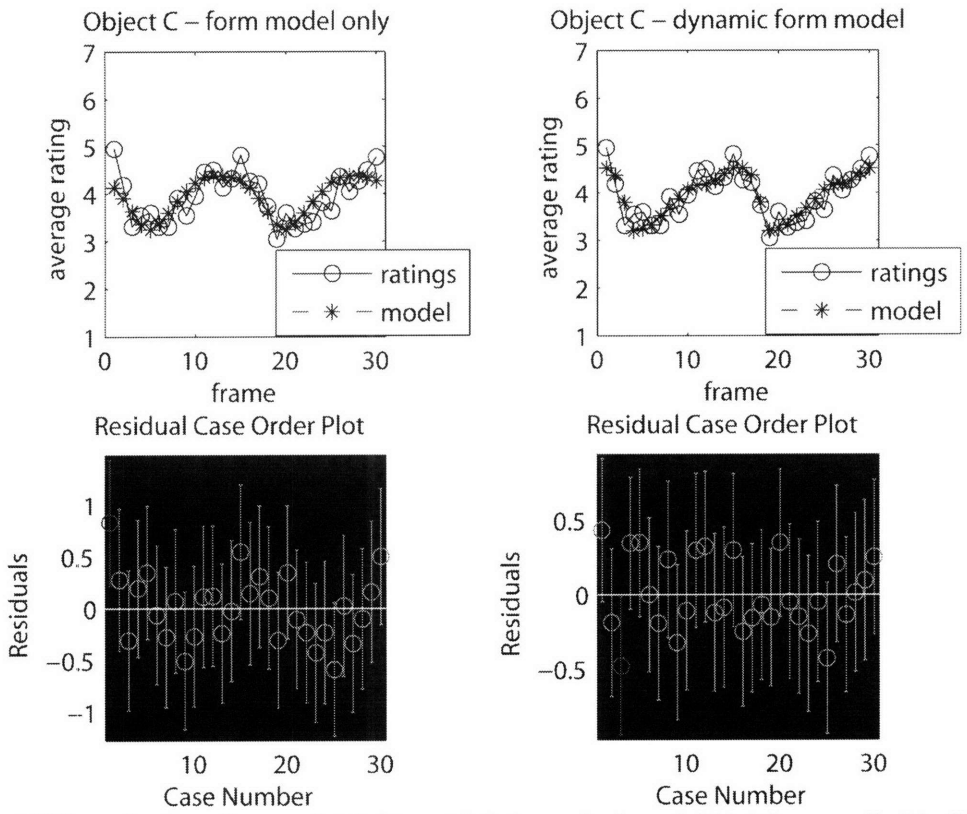


Fig. 4c – Best-fit lines for baseline model (left) and full dynamical model (right) as applied to Object C with the accompanying residual analysis below each plot.

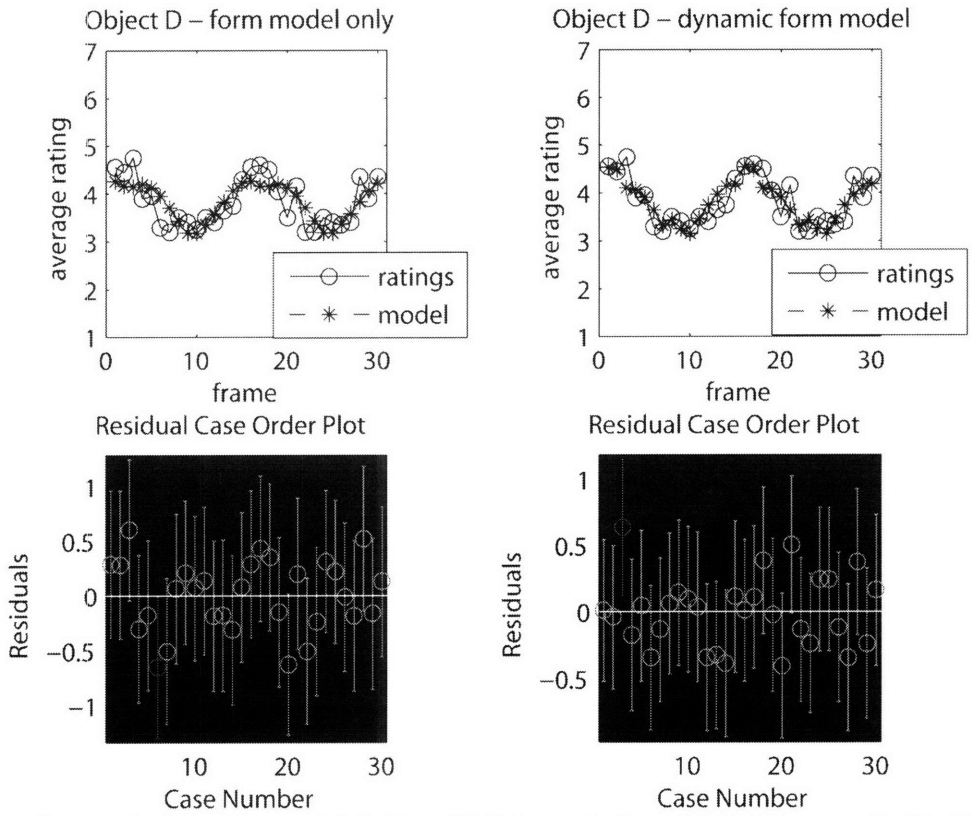


Fig. 4d – Best-fit lines for baseline model (left) and full dynamical model (right) as applied to Object D with the accompanying residual analysis below each plot.

Discussion

Using novel objects and an explicit rating task, we find that observers are very consistent in their selection of canonical views. Moreover, we find that our participants tend to select a single mode and its mirror reflection. The overall structure of the canonicity function also seems to be well-fit by a model of average foreshortening across distinct segments and is improved substantially when dynamic information is included in the model. Dynamic exposure may thus contribute to canonicity in this task in two ways. First, motion may provide cues to each segment's true length via the kinetic depth effect, allowing our observers to estimate the canonicity function we have defined. If this were the only use for motion, stereoscopic exposure to the training objects in a static setting should result in similar canonicity ratings. The added benefits of adding dynamic terms to our model as independent regressors suggests that the change in that function over time also contributes to the canonicity of individual views. Thus, motion affects canonicity beyond providing information for extracting 3-D form.

We continue by examining view canonicity for these objects as determined by an implicit measure rather than by explicit ratings. First, we ask whether or not implicit canonicity judgments resemble explicit ratings. Second, we ask if either of the two models we have developed appear to fit the data reasonably well.

Experiment 2

In this task, we use response time in a 2AFC visual memory task as a measure view canonicity following dynamic exposure to the objects similar to the training period used

in Experiment 1. We also compare observers' ability to correctly recall coherent vs. random paperclip sequences to determine the extent to which memory for coherently moving stimuli is comparable to memory for arbitrary sequences. The baseline model developed in Experiment 1 is applied to the RT data for coherent objects as well, to determine how successful our form-based model is at predicting implicit judgments of canonicity.

Methods

Subjects

The same 24 observers who participated in Experiment 1 also participated in this task. Observers were not asked to perform this task using any objects viewed in Experiment 1.

Stimuli

We use the same paperclip objects described in Experiment 1, along with a set of distracter images created for this task. Each distracter image was created from a specific paperclip image by adding noise to the XY projection of the vertices. Each vertex was moved by +\- 10 pixels in both X and Y, resulting in a new paperclip that is distinct from the original. We refer to the original paperclip stimuli as "object" images and these altered stimuli as "distracter" images.

Procedure

The experiment was divided up into four "coherent" blocks and four "random" blocks, which were interleaved during the full session. Each block began with observers viewing a sequence of paperclip images as described in Experiment 1. In "coherent" blocks, this sequence depicted a paperclip rigidly rotating about its vertical axis. During "random" blocks, the sequence depicted the ordered sequence of distracter images created from a paperclip. The object depicted in a coherent sequence was not the object from which distracters were created for use in the random sequence. Both sequences were played at a rate of 7.5 frames/second for 8 repetitions of the full image set.

Following this exposure period, each block continued with a test phase. On each trial of this task, observers were presented with two images flanking the center of the display, one of which had just been displayed during the exposure period of that block. In coherent blocks (where "object" images are the targets), the other image was the distracter image associated with the target. In random blocks (where "distracter images" are the targets), the other image was the object image associated with the target. Observers' task was to identify the image that had been seen in the previous exposure period as rapidly and accurately as possible. While observers had unlimited time to respond on each trial, it was emphasized that RT was being recorded and so they should try to be as fast as possible without compromising their accuracy. Within a block, each of the 30 target images was presented twice for a total of 60 trials per block and 240 trials per condition in the full session. Target image location for each trial and presentation order were randomized for each subject.

All stimulus display and response collection routines were carried out as described in Experiment 1. Each observer carried out this task for one paperclip object in the coherent condition, and a distinct paperclip object in the random condition.

Results

Accuracy

In Figure 5 we display the mean accuracy across blocks for both coherent and random conditions.

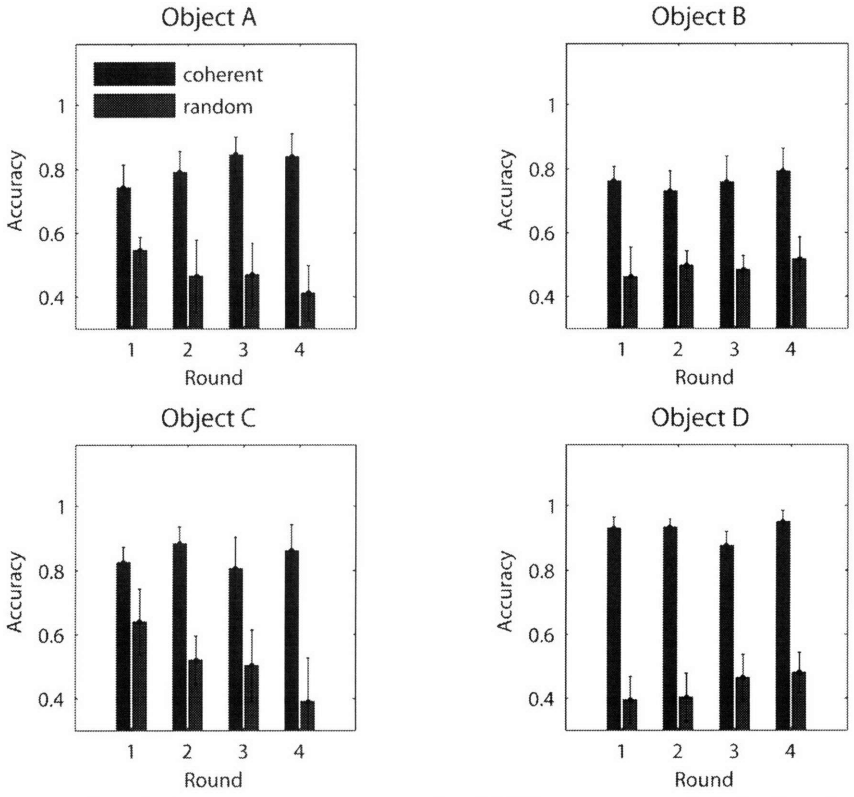


Fig. 5 – Mean accuracy for all observers carrying out the 2AFC memory task in both coherent and random blocks.

A 2x4 between-subjects ANOVA with training sequence type ("coherent" or "random" and block number as factors reveals only a main effect of sequence type ($F(1,176)=175.5$, $Mse=0.033$, $p < 0.001$). No other main effects or interactions were significant. Clearly observers in this task were far more capable at accurately remembering images from the coherent object sequence, performing at chance levels when distracter sequences were presented in random blocks.

Unfortunately, the relatively high rate of performance during coherent blocks means that there are not nearly enough errors on any individual frames for the application of our models to be meaningful. Accuracy simply does not fluctuate enough across frames for us to attempt any parametric explanation of the data. As a result, we do not apply either of our models here. Instead, we continue by considering the reaction time data.

Response Time

For each correct response to a target frame in the coherent condition, we recorded the observers' reaction times, yielding an implicit estimate of canonicity as a function of object view. The intuition is that "better" object views will be recognized faster than poor ones, leading to an RT function over target frames with significant peaks and troughs. To minimize the deleterious effects of inter-subject RT variability, we subtract the mean

RT from all trials (including correct and incorrect responses from both coherent and random blocks) from each observer's mean RT values across all coherent target frames. This allows us to characterize the average RT function across subjects in terms of fluctuations above and below each observer's baseline performance. In Figure 6 we display the mean RT functions recovered for each object.

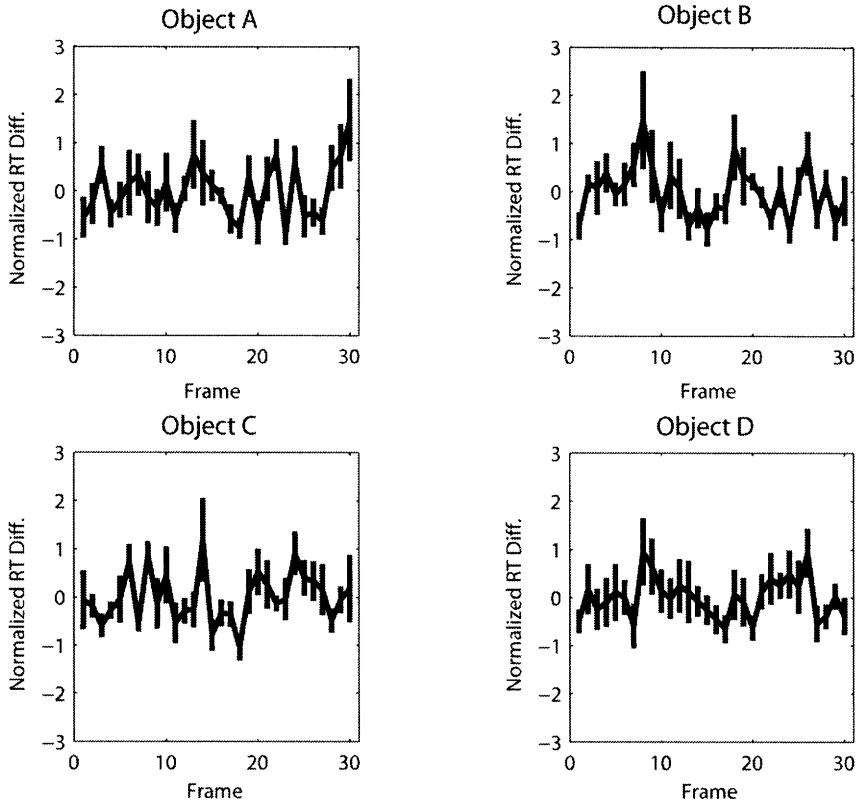


Fig. 6 – Mean RT fluctuations around baseline as a function of target frame for each "coherent" object.

To determine if the variations in RT across target frames and objects were significant, we carried out a 4x30 between-subjects ANOVA with object and target frame as factors. We find a significant effect of target frame ($F(29,600)=1.93$, $MSe=0.89$, $p = 0.0026$) as well as a significant interaction between target frame and object ($F(87,600)=1.42$, $Mse=0.89$, $p = 0.011$). The main effect of rows was nil, by virtue of the normalization procedure which subtracts the mean RT from each row of the data.

The observed effect of target frame indicates that there are indeed systematic differences in RT across target frames during the memory task, but does not rule out general-purpose memory phenomena such as primacy and recency effects. However, the significant interaction of target frame with object supports object-specific (and thus view-specific) contributions to the data. We thus have a dataset concerning view canonicity as defined by an implicit measure that we may proceed to analyze in the context of our proposed models.

Application of the Baseline Model

As in Experiment 1, we compare our observers' data regarding view canonicity to the predictions made by our baseline model. Despite its simplicity, this model proved

capable of capturing a substantial amount of the structure present in observers' explicit ratings of each target view. We therefore apply it to the mean RT fluctuation functions plotted in Figure 5 to determine if the implicit data can be modeled as successfully as the explicit data. We point out that since we are using RT rather than an explicit rating, positive RT fluctuations above baseline indicate poor canonicity for a view and negative fluctuations indicate good canonicity. This means that we will be looking for significant *negative* correlations in this analysis rather than positive values of the correlation coefficient. Table 3 contains the results of our application of the baseline model to the implicit data obtained in this task.

Table 3 – Correlation between baseline model and actual observer RTs.

	Prop. Variance	Significance
Object A	0.029	n.s
Object B	0.0061	n.s.
Object C	0.0001	n.s.
Object D	0.32	$p < 0.001$

For only one object does the baseline model provide a good fit to the RT data. For the remaining objects, average foreshortening explains virtually none of the structure in the data. We are thus forced to conclude that the baseline model is not applicable to this measure of view canonicity. Next, we examine the performance of the full dynamical model

Table 4 – Correlation between full dynamical model and actual observer RTs.

	Prop. Variance	Significance
Object A	0.046	n.s
Object B	0.35	$p = 0.0093$
Object C	0.24	$p = 0.068$
Object D	0.34	$p < 0.001$

We note improvement over the baseline model in each case, but still find that the model is not performing very well. The model's fit is only significant in two cases, marginal in another, and not significant in the fourth. We conclude that neither model is capable of fitting the data obtained from this task, but that the dynamical model is slightly more capable than the baseline model.

Discussion

We are unable to model the significant RT fluctuations we observe over target frame in this task. This leaves us with a few possible explanations for our failure at this stage. First, it may be that the structure we observe in the RT data is simply noise. However, one reason to think this is not the case is that the explicit ratings obtained in Experiment 1 generally correlate well with the implicit data in this task, as we see in Table 5.

Table 5 – Correlation between explicit and implicit canonicity judgments

	Corr. Coeff	Prop. Variance	Significance
Object A	-0.32	0.10	p = 0.089
Object B	-0.39	0.15	p = 0.034
Object C	-0.11	0.0121	n.s.
Object D	-0.41	0.17	p = 0.02

While these correlations are not perfect, they at least suggest that there may be some structure in the RT data relevant to the task at hand.

A second reason we may have failed in our modeling efforts may be that our models are not sophisticated enough. We have admittedly used simple tools and perhaps more complex models could more accurately approximate the implicit data. However, a final explanation that we believe may be very important to consider is the possibility that the presence of distracters encouraged observers to use a flexible strategy for identifying target images. So long as a target and a distracter are present, the observer can use either the "goodness" of the target or the lack of "goodness" in the distracter to form a judgment. Worse, the decision made on each trial need not depend on a fixed weighting of these measurements. If observers are indeed adopting a trial-by-trial strategy alternately relying on the properties of the target and the distracter, or a varying procedure for combining those properties, we are faced with a very difficult modeling challenge.

For now, we suggest that the structure we observe in the RT data in Experiment 2 is real, but may reflect a complex and adaptive strategy for target selection involving both target and distracter images. To proceed, we shall modify our memory test slightly to alleviate the problem of distracter images colluding with target images to influence the structure of the implicit canonicity function.

Experiment 3

In this last task, we use response time in a 1FC visual memory task as a measure of view canonicity following dynamic exposure to the objects used in both Experiments 1 and 2. By presenting targets and distracters in isolation from one another, we hope to minimize observers' ability to rely on cues besides target image properties to perform target selection.

Methods

Subjects

24 new volunteers from the MIT community participated in this task. All observers reported normal or corrected-to-normal vision.

Stimuli

Only the images of the original four paperclip objects were used in this task.

Procedure

The experiment was divided up into two sessions, each with four identical blocks. With a session, only one target object was used. As in Experiment 2, each block began with

observers viewing a paperclip rigidly rotating about its vertical axis. Sequences were played at a rate of 7.5 frames/second for 8 repetitions of the full image set.

Following this exposure period, each block continued with a 1FC test phase. On each trial of this task, observers were presented with one image in the center of the display, which either depicted an image from the exposure period or an image of an entirely different object. Observers' task was to decide as rapidly and accurately as possible whether or not the image had been in the preceding sequence, using a "go,no-go" paradigm. If observers believed the image had been present during exposure, they were to press the space bar as fast as possible. Otherwise they were to do nothing, and the image would disappear after 3 seconds. It was emphasized that RT was being recorded and so they should try to be as fast as possible without compromising their accuracy. Within a block, each of the 30 target and distracter images was presented twice for a total of 120 trials per block and 480 trials in the full session. Presentation order was randomized for each subject.

All stimulus display and response collection routines were carried out as described in Experiments 1 and 2. Each observer carried out this task for two different objects, with distracter objects in each block counterbalanced across subjects.

Results

Accuracy

In Figure 7 we display the mean hit rate across frames for each object. The average false alarm rate for all observers across all objects was approximately 4%, which is low enough that we have not corrected the hit rate for guessing.

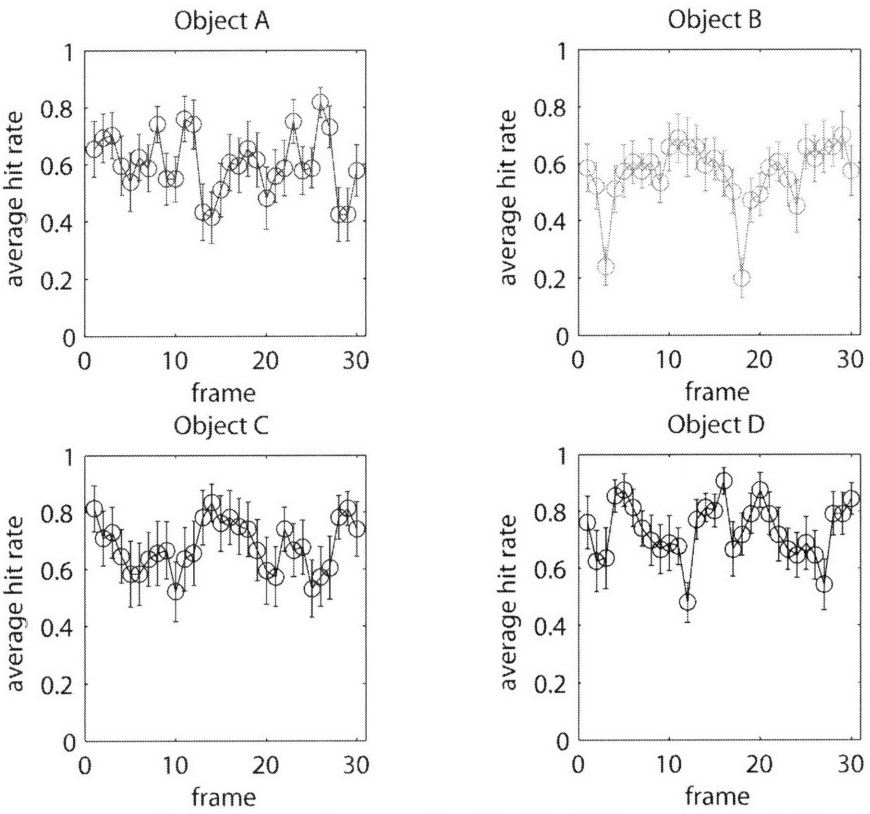


Fig. 7 – Mean hit rate across frames for each paperclip object in 1FC memory task. Error bars represent +/- 1 s.e.m.

To determine if the fluctuations in hit rate across frame were significant, we carried out a 4x30 mixed-design 2-way ANOVA with object as a between-subjects factor and frame as a within-subjects factor. Our analysis reveals a main effect of object ($F(3)=26.2$, $MSe=0.092$, $p < 0.001$) and an interaction between object and target frame ($F(87)=1.64$, $MSe=0.092$, $p < 0.001$). The main effect of target frame was not significant.

We continue by assessing the goodness-of-fit for both the baseline model and the full dynamical model described in Experiment 1. As before, we are particularly interested in whether or not the inclusion of dynamic regressors improves model performance substantially. Table 6 lists R^2 statistics and p-values for both models, while Figure 8 displays the model fits along with residual plots.

Table 6 – Goodness-of-fit for baseline and full dynamical models of 1FC hit rates

	Baseline model	Full dynamical model
Object A	$R^2=0.09$, $p=0.12$	$R^2=0.12$, $p=0.36$
Object B	$R^2=0.54$, $p<0.001$	$R^2=0.63$, $p<0.001$
Object C	$R^2=0.10$, $p=0.081$	$R^2=0.45$, $p<0.001$
Object D	$R^2=0.33$, $p<0.001$	$R^2=0.46$, $p<0.001$

We find that as in the case of the explicit ratings obtained in Experiment 1, the baseline model provides a good fit to the data while the inclusion of the dynamic terms improves the model fit substantially. This suggests both that observers actual memory for the target frames correlates well their subjective evaluation of canonicity, and that both judgments incorporate dynamic information concerning object appearance.

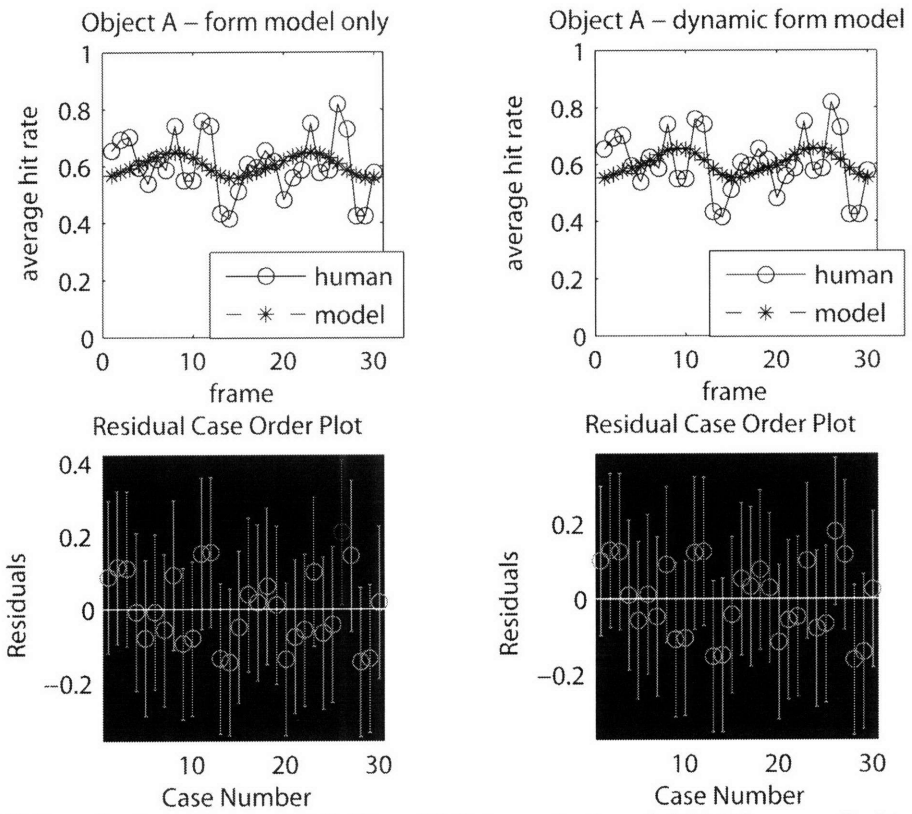


Fig. 8a – Best-fit lines for baseline model (left) and full dynamical model (right) as applied to Object A with the accompanying residual analysis below each plot. In each case, the dynamical model provides a better fit to the data.

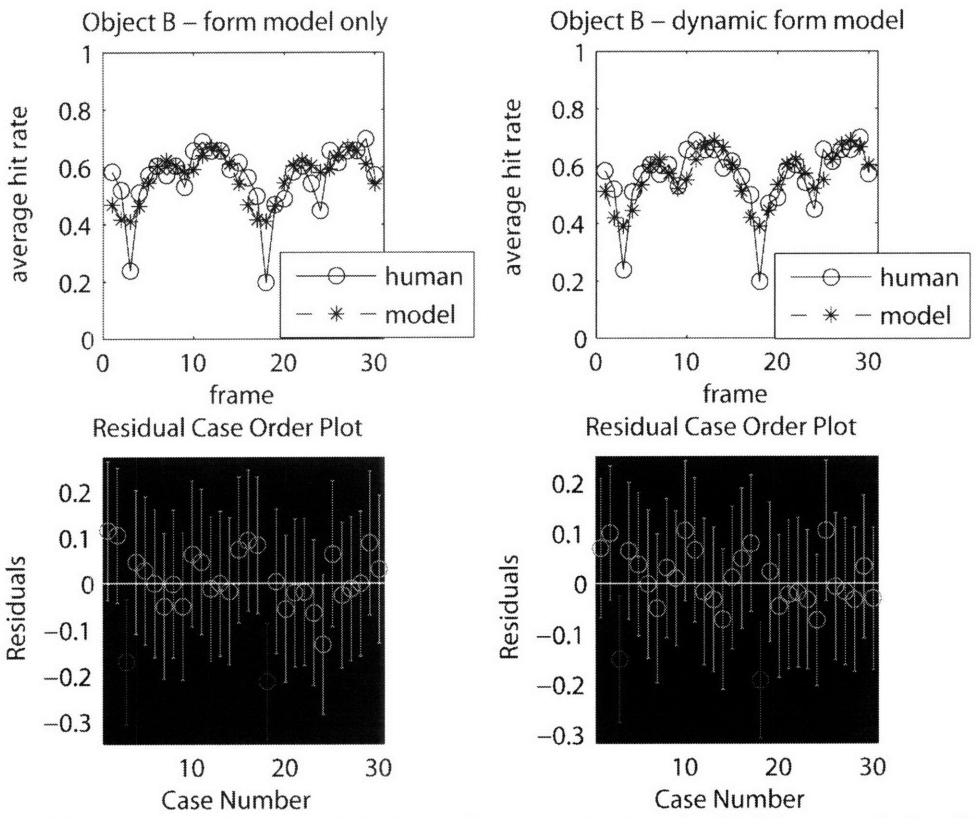


Fig. 8b – Best-fit lines for baseline model (left) and full dynamical model (right) as applied to Object B with the accompanying residual analysis below each plot.

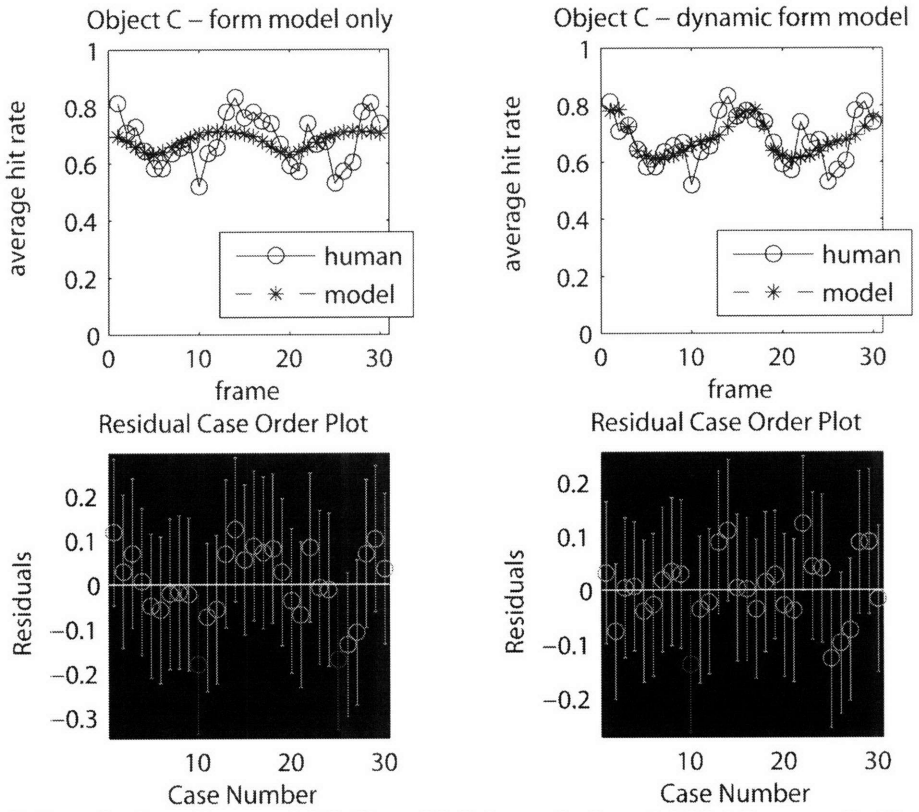


Fig. 8c – Best-fit lines for baseline model (left) and full dynamical model (right) as applied to Object C with the accompanying residual analysis below each plot.

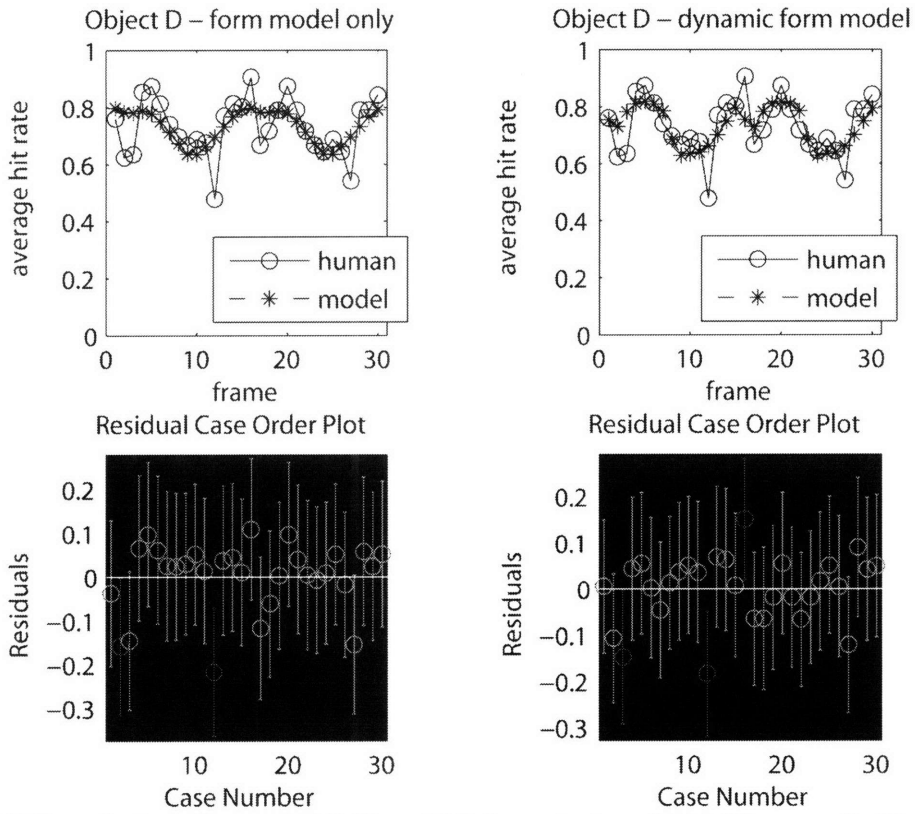


Fig. 8d – Best-fit lines for baseline model (left) and full dynamical model (right) as applied to Object D with the accompanying residual analysis below each plot.

Response time

We conclude by examining the response times to correctly recalled target frames. For each subject, the latency of response to each hit was collected and binned by target frame. The data was then normalized by mean response time to all stimuli so that each data point represents the deviation of RT above or below baseline for each target frame. The average RT across all target frames of all objects is displayed in Figure 9.

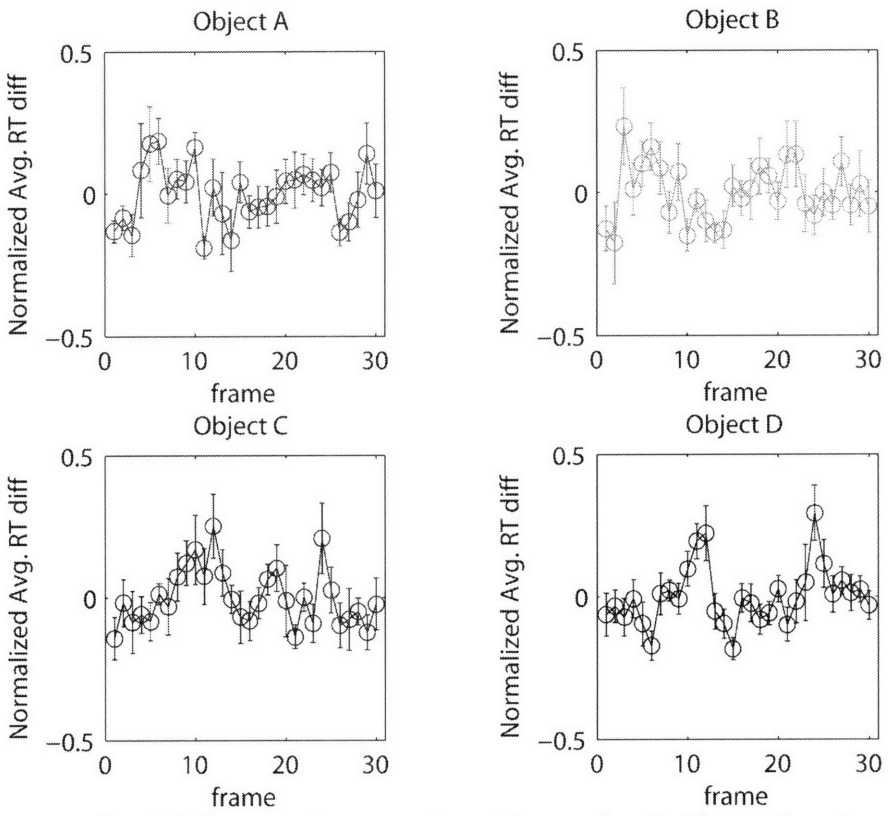


Fig. 9 – Average normalized RT in seconds across target frames for all objects. Error bars represent +/- 1 s.e.m.

As with our hit rate data, we carry out a 2-way 4x30 mixed-design ANOVA to determine if the fluctuations in RT across target frames are significant. We find a main effect of target frame ($F(29)=1.7$, $MSe=0.08$, $p = 0.012$) and an interaction between target frame and object ($F(87)=1.47$, $MSe=0.08$, $p = 0.0043$). The main effect of object was null by virtue of our normalization procedure. We conclude from this that the structure we observe in RT across target frames is statistically real.

We next assess the goodness-of-fit of both the baseline and full dynamical models to the RT data in Figure 8. The results of this analysis are displayed in Table 7.

Table 7 – Goodness-of-fit for baseline and full dynamical models of 1FC RTs

	Baseline model	Full dynamical model
Object A	$R^2=0.18$, $p=0.0175$	$R^2=0.23$, $p=0.074$
Object B	$R^2=0.015$, $p=0.52$	$R^2=0.14$, $p=0.26$
Object C	$R^2=0.02$, $p=0.41$	$R^2=0.20$, $p=0.11$
Object D	$R^2=0.45$, $p<0.001$	$R^2=0.52$, $p<0.001$

Unfortunately, the lack of significant fits to the data makes it difficult to determine whether or not the inclusion of dynamic regressors actually affects model performance at all. While it is tempting to observe the increase in variance captured as we include the dynamic terms, we cannot make a firm conclusion at present.

General Discussion

Across three tasks with the same four novel "paperclip" objects, we have measured canonicity in a variety of ways. Along the way, we have attempted to fit the observed data with a simple model of view canonicity that can either include or exclude terms relevant to the dynamic nature of the input. As a result, we have been able to answer most of the questions that we started with, while leaving a fair amount of interesting work for the future. We conclude by addressing the questions that motivated this study, and discussing possible topics for further investigation.

Are consistent canonical views evident in the behavioral data?
In general, we have observed evidence for "privileged" views in all three of our tasks. Explicit ratings, response times in both the 2FC and 1FC memory tasks, and hit rates in the 1FC task all showed significant fluctuations across target frames that were object-specific. This suggests that canonicity (as defined in each task) was not determined by factors like primacy and recency, but rather a function of object appearance across target frames. As a result, we conclude that "privileged views" can be determined via many different methods.

However, one drawback is that these methods of defining canonicity do not necessarily agree well with one another. Looking at the data across all of our tasks we note many task-dependent differences in the data. In particular, in Experiment 2 we note that the presence of distractors may have added a great deal of noise to our data. Comparing the very sharp transitions in RT fluctuations across target frames in Experiment 2 to the much smoother data in Experiment 3 makes it plain that the canonicity of a view must be considered in the context of the task presented to the observer. However, we did find a fairly high level of agreement between the explicit ratings in Experiment 1 and the hit rates in Experiment 3, which suggests to us that there is actually a fair degree of merit in simply asking subjects for ratings of canonicity rather than adopting implicit criteria.

Can we approximate the behavioral data with a purely form-based model?
Though our baseline model of canonicity was something of a stick-figure caricature of object form, it worked surprisingly well across all of our tasks. There are certainly more factors to consider (such as image-based features and the primacy and recency effects we have so far discounted) but as a starting point, the baseline model holds up fairly well. The primary exceptions to this rule are two-fold: First, we note in Experiment 1 that when the baseline model's canonicity function is very shallow the human data is not similarly circumscribed. Second, response time data seems to be extremely difficult to model successfully. Even if we attempt to model log-transformed RTs or median-filtered data in an effort to avoid outliers (steps taken during exploratory analysis that are not reported here), we have been unable to do much better than we have described already. It may simply be the case that response latency is a poor measure of view canonicity. While this is a rather pessimistic conclusion, at the moment it seems

inescapable, and should at least be remembered as a good rule of thumb for future experiments.

Overall, however, our bare-bones definition of canonicity has worked well for the simplified objects we presented to our observers. Can we scale this model up to work with more complex natural objects? At the moment, we can only speculate, but there is some reason for hope. Tracking image patches from frame to frame would be simple enough, with the initial patches determined by some sort of interest operator. Even without ground truth data concerning the length of the virtual segment joining two patches the change in projected length can still be computed. This simplified model, using only changes in projection rather than ratios of projected length to true length, could be implemented without too much difficulty, but there are several issues that may plague its application. First of all, it is unclear how we should cope with self-occlusion since features can be "lost" during rotation (although see (Wallraven & Bulthoff, 2001) for an interesting computational approach and (Tarr & Kreigman, 2001) for relevant human data). Second, uncertainty in the position of a particular feature due to changes in appearance during rotation or smoothness of object form will add noise to our estimates of foreshortening. The result may be a very messy model that is quite difficult to handle. Integrating the most robust and easily translatable aspects of our baseline model with an appearance-based model of canonicity may be the best way to proceed.

Does dynamic input influence canonicity?
Finally, having examined the evidence for a role of dynamic information in canonicity judgments across three tasks, we must conclude that form dynamics do play a role in defining "privileged views" of a novel object. In every case, the inclusion of dynamic regressors improves the proportion of variance captured by our models, often succeeding where the baseline model has failed. While the role of dynamic input appears to be complex and object-specific, our results suggest that this information is indeed a relevant factor in determining view canonicity.

To make the role of object motion more clear, it may be worthwhile to manipulate sequence dynamics in various ways and examine canonicity judgments under distinct conditions of speed, sequence smoothness and duration. At present, our evidence for the role of dynamic input is somewhat abstract (being based only our modeling) and further psychophysical investigation would help solidify our understanding of how motion contributes to form encoding.

Conclusions

Adult observers do appear to extract prototypical views, or "keyframes", from sequences of novel moving objects after very little exposure. While view canonicity can be determined in many different ways, the results of different tasks are not always in agreement. In particular, response time appears to be an unstable basis for assessing view canonicity. The inclusion of distracters in tasks designed to estimate view canonicity also appears to be problematic, with observers adopting a flexible strategy of using positive or negative evidence to guide behavior. Finally, in our experiments here, a simple form-based model of foreshortening provides a surprisingly good fit to the data. This model is substantially enhanced by the inclusion of dynamic information however,

suggesting that object motion provides information for canonicity judgments beyond a robust estimate of form.

Acknowledgments

This work has been greatly improved by the author watching *Children of Men* just prior to writing the bulk of Experiment 3. There's nothing like a good dystopian action flick to make you glad you're alive to write the end of your thesis.

References

Blanz, V., Tarr, M. J., & Bulthoff, H. H. (1999). What object attributes determine canonical views? *Perception, 28*, 575-600.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433-436.

Bricolo, E., Poggio, T., & Logothetis, N. (1997). *3D object recognition: A model of view-tuned neurons.* Paper presented at the Advances in Neural Information Processing Systems 9.

Bulthoff, H. H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Science, 89*, 60-64.

Cutzu, F., & Tarr, M. (1999). Inferring perceptual saliency fields from viewpoint-dependent recognition data. *Neural Computation, 11*, 1331-1348.

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

Edelman, S., & Bulthoff, H. H. (1992). Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision Research, 32*(12), 2385-400.

Humphrey, G. K., & Joliceur, P. (1993). Visual object identification: Some effects of image foreshortening, monocular depth cues, and visual field on object identification. *Quarterly Journal of Experimental Psychology, 46A*, 137-159.

Lawson, R., & Humphrey, G. K. (1996). View Specificity in Object Processing: Evidence From Picture Matching. *Journal of Experimental Psychology: Human Perception and Performance, 22*(2), 395-416.

Logothetis, N., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology, 5*(5), 552-563.

Logothetis, N. K., Pauls, J., Bulthoff, H. H., & Poggio, T. (1994). View-dependent object recognition by monkeys. *Current Biology, 4*(5), 401-414.

Murase, H., & Nayar, S. K. (1995). Visual Learning and Recognition of 3-D Objects from Appearance. *International Journal of Computer Vision, 14*, 5-24.

Palmer, S., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. In J. Long & A. Baddeley (Eds.), *Attention and Performance IX* (pp. 135-151). Hillsdale, NJ: Lawrence Erlbaum.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision, 10*, 437-442.

Perret, D. I., Harries, M. H., & Looker, S. (1992). Use of preferential inspection to define the viewing sphere and characteristic views of an arbitrary machined tool part. *Perception, 21*, 497-515.

Peters, G., Zitova, B., & von der Malsburg, C. (2002). How to measure the pose robustness of object views. *Image and Vision Computing, 20*, 249-256.

Tarr, M., & Bulthoff, H. H. (1995). Is human object recognition better described by geon-structural-descriptions or by multiple views? *Journal of Experimental Psychology: Human Perception and Performance, 21*(6), 1494-1505.

Tarr, M. J., & Kreigman, D. J. (2001). What defines a view? *Vision Research, 41*, 1981-2004.

Wallraven, C., & Bulthoff, H. H. (2001). *Automatic acquisition of exemplar-based representations for recognition from image sequences.* Paper presented at the CVPR 2001.

Conclusion

At the beginning of this thesis, I suggested that presently we lack enough data concerning human perception of dynamic objects to begin formulating a comprehensive theory. My goal has been to add to our current state of knowledge in a systematic and theory-driven way so that some basic principles of a coherent model can be elucidated. Here, I review the results of the experiments reported here in an attempt to develop a foundation for future efforts. While there is still clearly a great deal of important work left to be done, the data I have presented provides a set of important constraints on dynamic object perception and reveals several key features of human perception and recognition for moving objects. What follows is my attempt to put together the various pieces of this thesis into a coherent whole.

What does object motion do? – Temporal association and its role in appearance coding, prediction, and categorization

At the outset of these experiments, I suggested that the basic temporal association hypothesis represented the most useful framework for further investigation of dynamic object perception. To a substantial degree, the experiments in Section 1 of the thesis were motivated by the desire to further elaborate this proposal by testing the consequences of dynamic object training in a variety of settings. While certain aspects of the original proposal hold up well in light of these tasks, several of my results force us to modify the basic hypothesis substantially.

Appearance coding: generalization and sensitivity following temporal association

In Chapter 1, I reported that observation of a dynamic sequence induces increased generalization and sensitivity for images present during training. This has fairly large consequences for our understanding of how temporal association influences object appearance coding. While later results indicate that there is a lot more than this going on during dynamic object perception, I suggest that this fundamental observation is perhaps the most important contribution of the thesis.

Prediction: An additional role for temporal association

The experiments in Chapter 2 used a familiar object, the walking human body, to demonstrate that what observers know about object motion from long-term experience interacts with recent dynamic input in a complex way. Specifically, it seemed like our conception of a "coarsening" process that increased sensitivity to image change had to be modified to incorporate a prediction mechanism that was allowed to pre-empt changes in appearance coding until accurate predictions had been learned. This proposal was compatible with the results reported in Chapter 1, and leads to several testable predictions for future experiments.

Categorization: Using motion to identify common and relative components of appearance

Finally, Chapter 3 marked a foray into the realm of category learning by demonstrating that the diagnosticity of object motion during learning directly impacted categorization efficiency for static images. Learning object categories under conditions of motion diagnosticity actually impairs performance, so long as diagnostic motion is qualitatively

different across categories. This effect was discussed in terms of the extraction of "common" and "relative" components of form in appearance space, analogous to the famous demonstrations of vector analysis by Gunnar Johannson.

What does object motion do? – Appearance dynamics and recall

While Section 1 provided several important results regarding discrimination and categorization of dynamic objects, Section 2 presented several results relevant to understanding the encoding of individual images during dynamic viewing. Specifically, instead of describing the properties of the established appearance code for a population of images, in Section 2 I examined the variables governing whether or not individual images end up being encoded at all. The dynamics of object appearance to play an important role here, suggesting that at a very deep level our experience of the visual world is determined by movement.

Appearance dynamics and immediate recall

Using thorough parametric studies, I demonstrated in Chapter 4 that our ability to accurately remember what we have just seen is determined in large part by recent perceptual history. In appearance space, our encoding of any individual image is affected by presentation time, temporal contingency, and image differences that are local in time. This last point is particularly important insofar as it highlights what I believe is a core principle of dynamic object perception: spatial factors can be global, but temporal factors are predominantly local. In the closing remarks of this chapter, I proposed a simple probabilistic model for dynamic perception built on the robust finding that regression backwards in time appears to be the dominant feature of immediate recall for moving objects.

Appearance dynamics and keyframes

In chapter 5, I approached the issue of image encoding during dynamic viewing from a different direction. Rather than asking how temporal factors affect immediate recall, the focus of this chapter was on how well observers' preference and memory for individual images could be approximated using static and dynamic models of view canonicity. This analysis has the benefit of providing some data on how "canonical views" are selected during more natural viewing conditions, and also examining if motion contributes to view selection beyond its support of form extraction. I find that the dynamics of form change do make a significant contribution to view encoding, suggesting that local changes in appearance over time do exert an influence on object memory.

What now?

A sketch of dynamic object perception

So what does object motion do? In most models that use temporal continuity as a teaching signal of some kind, motion is reduced to a tool for propagating object labels over time. The emphasis in these models is almost solely on learning how to be invariant to multiple sources of appearance change for complex objects. The data presented in this thesis suggests that the human visual system uses object motion to do more, however. Specifically, I have shown that observers' learn to be sensitive to observed appearance changes while they simultaneously learn to generalize over those same distinct images and that global appearance changes that are local in time appear

to have a profound influence on multiple aspects of form perception and encoding. Given these results, I argue that our understanding of how object concepts are learned from dynamic input must be restructured substantially.

How should we re-conceive the role of object motion in learning representations of object form? This is of course highly speculative, but I think that what we are seeing across all of these sets of studies is the footprint of a learning mechanism that uses object motion to accomplish one primary goal: to identify a primary axis of variation in appearance space.

What this proposal is essentially saying is that object motion helps the visual system do dimensionality reduction. High-dimensional image data usually needs to be described compactly to avoid the "curse of dimensionality" and I suggest that object motion is used to obtain an estimate of something like the first principal component of the incoming image data, yielding a very crude first-order approximation of "intra-stimulus" appearance variability. To put it plainly, if you need to describe an object's appearance with just one number, use its projection on the axis of motion through appearance space.

The results reported in Section 1 look fairly reasonable when viewed through this lens. Chapter 1 tells us that this axis needs to be represented in a way that provides both generalization and sensitivity, possibly via a population of encoding units that stretch and overlap along its length. Chapter 2 tells us that the dominant direction of motion along the axis is probably remembered, and that changes in tuning functions only take place when incoming data doesn't alter the established directionality (or lack thereof). Finally, the categorization results from Chapter 3 suggest that during category learning, the motion of individual objects is used to establish one axis that is applied to all the objects under consideration in *both* categories. The directionality effects we observe in Chapter 2 are wholly extendable to this new case, and are entirely consistent with the reported data.

This is only a tentative conjecture at this point, but it both serves as a fair summary of the first section and provides the foundation for a concrete model of how object representations are built following dynamic exposure. In particular, it suggests that learning the principal component of a distribution is dominated by spatiotemporally smooth paths through that distribution (as revealed by our categorization data). Also, it suggests that dynamic perception is a fundamentally coarse process. Describing a complex, deforming shape with one number is about as quick and dirty as it gets, yet this sketch of a model seems capable of explaining a lot of the behaviors I have described. The simplicity of this proposal also makes it potentially easy to implement, test, and use to generate new and interesting hypotheses regarding human vision.

Can we relate the results in Section 2 to the stick figure sketch I used to summarize Section 1? To be honest, there just isn't enough data to do that in an iron-clad way yet. To be conservative, the real contributions of this thesis are probably best-expressed in a two-fold manner: 1) The effects of dynamic object perception on discrimination and categorization in various settings force a refinement of the basic temporal association

hypothesis to include observed increases in sensitivity, effects of object familiarity and prior expectations of object movement, and effects of motion diagnosticity on categorization, 2) Local appearance changes in time directly effect the strength of encoding for images of a dynamic object. Taken together, these two observations point the way to refinements of existing models and also suggest numerous additional experiments, many of which have been discussed in the preceding chapters.

To look a little beyond this conservative assessment of what we have gained from these experiments, however, there are three aspects of the work I have presented here that I hope can serve as the basis for future work.

First of all, though I can't rigorously defend the full extent of this proposal with the results presented here, I suspect that ultimately most of dynamic object perception will be explainable using a highly redundant model of global appearance like the "principal component" proposal I outlined above. One of the most remarkable and unintuitive aspects of dynamic perception is just how *little* we actually see and remember of a moving object, which suggests to me that our tools for processing such input must be remarkably blunt. For lack of a better word, I have been led to the belief that dynamic object perception is "smooshy," insofar as forms tend to bleed into one another very readily, and the vividness of change over time generally overshadows the ability to appreciate rich spatial structures. This makes it all the more remarkable that sensitivity to small changes can be maintained in the code for a dynamic objects' appearance, a feat which I maintain is best accomplished with a "coarse code" for appearance. This latter idea has already been discussed at length in Section 1, but I want to emphasize here at the conclusion of this manuscript that this idea represents a powerful extension of some old ideas about neural processing that somehow have yet to find their way into high-level vision. To be more specific, what I have been subtly proposing throughout this thesis is a direct analogy between localization in visual space and localization in appearance space. My proposals for "coarse coding" of object appearance are really no different in spirit or in form than earlier explanations of hyperacuity, save for the problem domain. Though this seems like an obvious connection to make, it is remarkable how little attention seems to have been paid to developing this idea further and exploring its consequences. Hopefully, some of the results reported in this thesis may help convince others that the application of population codes to object appearance provides a nice framework for future investigation, potentially leading to a deeper understanding of high-level aftereffects and the timecourse of visual learning.

The second aspect of the work presented here that I hope makes an impact on the study of dynamic object perception is essentially methodological. Throughout these tasks, I have emphasized the use of image-level comparisons rather than category or object-level comparisons. My goal in doing so was to avoid cognitive interference, or to put it plainly, to avoid having subjects only give me the right answers after I taught them what to say. The result is a rich set of data that demonstrates how object motion can influence tasks that are solvable in principle by extremely "dumb" mechanisms. I argue that the use of straightforward change detection tasks and image-level comparisons is crucial to obtaining a clear picture of dynamic object perception. Though the ultimate goal for a theory of object recognition must be to describe how labels are assigned to images, the singular use of naming or categorization tasks can obscure important

aspects of human perception. Worse, it is often very hard to know whether or not one is studying vision or general cognition unless one can show an effect on something like discriminability. In many of the studies I have cited in the preceding chapters, important results hinge on observers reporting numbers on arbitrary scales or producing object labels under circumstances that make the "right" response confusing at best. The result is that we can't build a good model of behavior because we don't know what task the observer was actually performing, nor do we know what would constitute optimal or "ideal" behavior. A deep understanding of dynamic object perception will ultimately require a great deal more data from studies in which we can see the influence of high-level processes in tasks that seemingly require only low-level mechanisms. Though I have experienced first-hand that this is a time-consuming and difficult way to work, I think it is necessary for the development of useful models, be they conceptual or computational.

Finally, though I confess I have not completely adhered to my own advice in this regard (even within the boundaries of this thesis), the more work that can be done with complex three-dimensional objects, the better. At the moment, studies of dynamic perception seem to occupy two extremes: either very simple "objects" are being employed (dots, for example) or fully natural scenes with multiple objects and background clutter are being used as stimuli. What is desperately needed is more work concerning the stimuli "in the middle" like isolated rigid objects with texture and shading, real human figures as opposed to point-light walkers, and objects on backgrounds that we can begin to understand structurally. We are very far from knowing what the right dimensions of appearance space are for object perception, so working within the confines of stimuli that we can obtain *some* amount of ground truth for is still very important. This isn't to say that working at either extreme of the artificial/natural stimulus spectrum is a bad idea, of course, only that I think there's a lot of fruitful work to be done in the middle.

Closing Remarks

With that, I conclude this investigation of dynamic object perception. I have found that the influence of object motion on recognition is a good bit more complicated than earlier results have indicated, but that this complexity may point the way to a more general (and equally elegant) process for object perception. Object motion helps the visual system decide how to pick out discrete images from the continuous flood of sensory data in the natural world, while simultaneously preserving our ability to maintain some representation of an entire sequence. While we are still far from a complete realization of a theory of dynamic object perception, this thesis has highlighted several important constraints on such a theory, and provided a good platform for future work.