Diagnostic Object Motion Weakens Representations of Static Form

Benjamin Balas (bjbalas@mit.edu)
Pawan Sinha (psinha@mit.edu)
Department of Brain and Cognitive Sciences, 43 Vassar St.
Cambridge, MA 02140 USA

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.

Keywords: Object recognition; object motion; categorization

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.

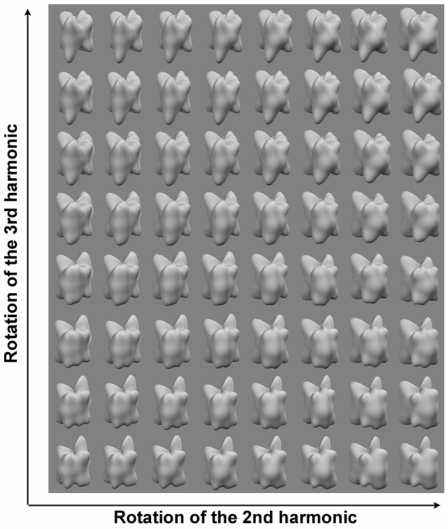

Figure 1:An 8x8 space of "blob" stimuli. The axes of this space are defined by the phase angle of the 2nd and 3rd harmonic. Movement along each axis induces non-rigid motion that is distinct from that generated by movement along the other axis.

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 2nd and 3rd 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 2nd harmonic, we end up with what we will call "horizontal" motion through blob space. Rotating only the 3rd 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 2nd and 3rd 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).

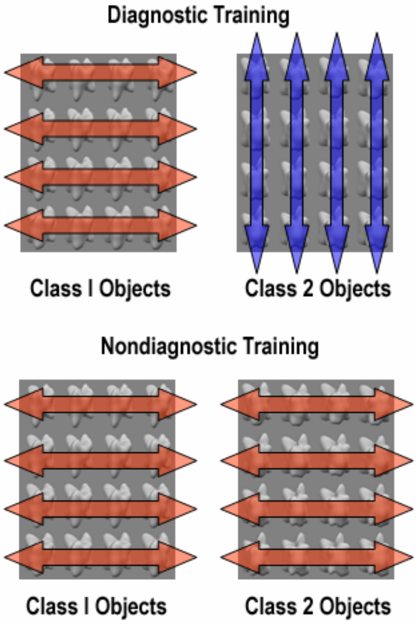

Figure 2: Object motion diagnosticity as defined in Experiment 1. The top row depicts the construction of dynamic objects for observers in the "Diagnostic" group while the bottom row depicts the same for observers in the "Non-diagnostic" group. The same images are always used to define Class I and Class 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 2^{nd} 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.

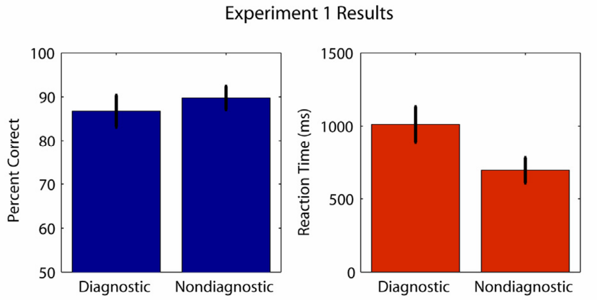

Figure 3: Accuracy (left) and response time for correct judgments (right) for observers in Experiment 1. There is no significant difference in accuracy between the two groups, but mean RTs are significantly longer in the Diagnostic group. Error bars represents +/- 1 s.e.m.

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.

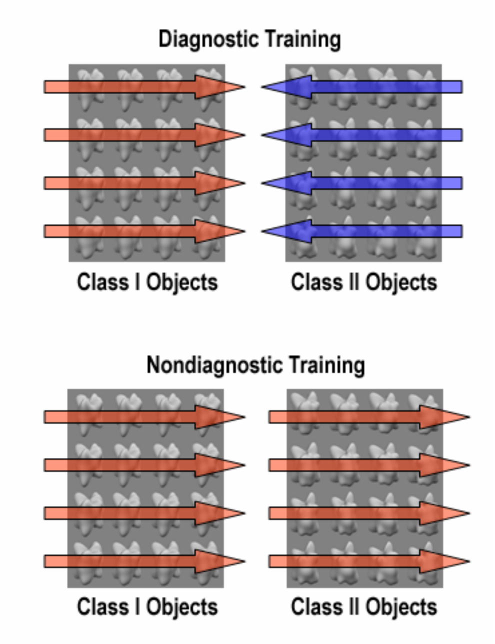

Figure 4: Object motion diagnosticity as defined in Experiment 2.

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).

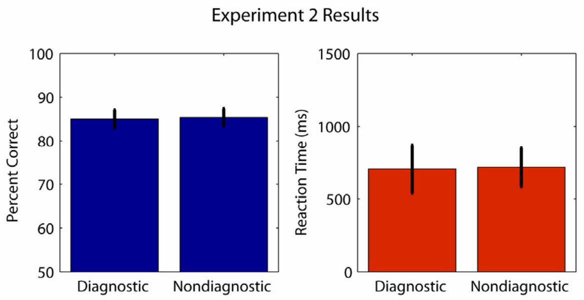

Figure 5: Accuracy (left) and response time for correct judgments (right) for observers in Experiment 2. There are no significant differences for accuracy or RT. Error bars represents +/- 1 s.e.m.

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.