Object Recognition Using Spatiotemporal Signatures

James V Stone

Psychology Department, Sheffield University,

Sheffield, S10 2UR, England.

Tel: 0114 222 6522 Fax: 0114 276 6515

Email: j.v.stone@sheffield.ac.uk

Web: http://www.shef.ac.uk/psychology/stone

Key words: Object recognition, motion, learning.

Running Head: Spatiotemporal signatures.

Abstract

The sequence of images generated by motion between observer and object specifies a spatiotemporal signature for that object. Evidence is presented that such spatiotemporal signatures are used in object recognition. Subjects learned novel, three-dimensional, rotating objects from image sequences in a continuous recognition task. During learning, the temporal order of images of a given object was constant. During testing, the order of images in each sequence was reversed, relative to its order during learning. This image sequence reversal produced significant reaction time increases and recognition rate decreases. Results are interpreted in terms of object-specific spatiotemporal signatures.

Introduction

Conventional theories of object recognition seek to explain how objects seen from a single, static view are recognised. However, computational and ethological analyses of vision in terms of spatiotemporal information suggest that motion could be used, not only for recovering three-dimensional (3D) shape information [Ullman, 1979], but also, directly, for the recognition of biological stimuli [Adelson, 1991] *and* rigid objects in motion [Stone, 1993]. More generally, these analyses suggest that visual mechanisms which operate successfully on static views may constitute special cases of a more general mechanism attuned to spatiotemporal stimuli. According to this hypothesis, object recognition from a static view is a special case of a general ability to recognise objects from spatiotemporal sequences.

Observer motion is a common source of retinal image changes. For example, looking at an object whilst moving past it over a small distance generates a temporal sequence of retinal images. These images are related to each other principally by rotational transformations, although components of scaling and shear are also present. Such a sequence is a rich source of visual cues (e.g. motion, shading, texture and stereo) which can be used to estimate an object's 3D shape. The temporal contiguity of inputs suggests that they were derived from similar 3D spatial scenarios. This can act as a powerful cue for learning about atemporal invariants such as surface depth [Stone, 1996] and 'object' identity [Becker, 1996] in artificial neural nets. Additionally, the temporal sequence of retinal images of an object is itself characteristic of the particular object being viewed, and constitutes a spatiotemporal signature of that object [Stone, 1993]. Note that there is a critical difference between information provided by 'shape from' (shading *etc*) mechanisms and information provided by spatiotemporal signatures. Whereas 'shape from' mechanisms use images to provide information about the *atemporal* 3D structure of an object, spatiotemporal signatures *consist* of the temporal evolution of spatial changes over time. The question addressed in this paper is: Do spatiotemporal signatures contribute to object recognition?

In an experiment on biological motion, Mather *et al* [Mather et al., 1992] used Johansson figures [Johansson, 1973] to generate cue-conflict stimuli such that figures had the walking characteristics of one gender, and the structural characteristics (e.g. as defined by width of hips) of the other gender. (A Johansson figure consists of a person with a light placed at each major joint which is viewed in a darkened room). The perception of gender was found to be associated with the motion characteristics of a walking Johansson figure, rather than by atemporal 3D gender-specific structural information.

Further evidence that certain types of objects might learn to become associated with particular types of motion is provided in [Sinha and Poggio, 1996], where subjects interpreted rotation of rigid stick figures of humans as a walking motion. In this case, rotation of a rigid object was interpreted as a deformation, presumably because these types of object (i.e. people) are usually associated with a particular type of deformation (i.e. walking).

Whilst the results reported in [Mather et al., 1992,Sinha and Poggio, 1996] may seem unsurprising in the case of biological motion stimuli [Adelson, 1991], it is not obvious that a similar type of effect might be found for rigid objects in general.

Figure 1 HERE.

Methods

Stimuli: The stimuli consisted of image sequences of rigid, smooth, grey-level objects rotating against a black background (see Figure 1). The obliquely placed light source was constant within and between image sequences. In each 90-image sequence, one object rotated through 360 degrees around an axis which rotated over time. All rotations were around a fixed point, which approximated the centre of mass of the object. All objects underwent the same set of rotational changes, giving the appearance of a tumbling motion. Each image was 300x300 pixels with 128 grey-levels. Image sequences were played at a constant rate of 25 images/second, and were displayed in a darkened room on an Apple Multiple Scan 20 computer screen (set to 1024×768 pixel resolution), using a modified version D Pelli's Videotoolbox software [Pelli, 1997]. Subjects viewed movies at a distance of about 0.75m. The target and distractor objects were different for each subject, and were chosen randomly at the start of the experiment. The starting image of each sequence was chosen at random every time it was played.

Table 1 HERE.

Procedure: The experiment consisted of three learning blocks of about 20 minutes each. In each block, each subject simultaneously learned to recognise four target objects, in a continuous recognition task, with targets being shown for a minimum of 10 trials (see Table 1). At the start of each block, subjects were shown four targets once for two complete rotations (i.e. 180 images). Thereafter, each subject was shown a sequence of image sequences, of which half displayed a target object and half displayed a distractor object. Each distractor was seen once only.

Subjects indicated if each image sequence contained a target by pressing one of two response keys. Subjects were asked to respond as quickly and as accurately as possible at any time after the start of each image sequence. No feedback was given at any time.

Subject performance was evaluated over each trial set within a block. A trial set is defined as the four targets and four previously unseen distractors, shown sequentially in random order. A score for each trial set was calculated as follows. If $T/4$ is the proportion of targets correctly recognised and $F/4$ is the proportion of distractors identified as targets then score = 1 if $T \geq 3$ and $F \leq 1$; else score=0. The learning criterion was reached by obtaining a score of 1 for three out of four consecutive trial sets. After the learning criterion had been reached, each subject continued the task as before for five further sets. Subjects were not informed that the learning criterion had been reached, and the five test sets followed the learning sets without interruption.

Within each block, half of the targets were allocated to the experimental and half to the control condition. In the experimental condition, the order of images in each target sequence was reversed once the learning criterion had been reached. In contrast, the order of images in each target sequence remained unaltered within the control condition. Subjects were informed at the start of the experiment that the order of images in some sequences would be reversed at some points in the experiment.

The order of images was counter-balanced across sequences. Half of the target and distractor image sequences were played in ascending order (e.g. $1 \rightarrow 90$, as denoted by a '$\bullet$' in Table 1) from a randomly chosen starting image, and half in descending order (e.g. $90 \rightarrow 1$, as denoted by a '$\circ$' in Table 1) during learning and testing, in both the experimental and control conditions.

There were 28 subjects, all of whom were undergraduate psychology students.

Graphs: In order to reflect the within-subjects experimental design, all and only the standard errors plotted in Figures 2 and 3 were computed after inter-subject variability had been removed [Loftus and Masson, 1994]. The plotted standard errors were based on transformed observations defined by $y_{st} = x_{st} - x_s + x_G$, where x_{st} is the mean observation (RT, recognition-rate, or false-alarm rate) of subject s on trial t, x_s is the mean of observations for subject s over all trials in three blocks, and x_G is the grand mean of all observations.

Results

Figure 2 and 3 HERE.

A comparison of performance before and after image sequences were reversed showed a significant decrease in recognition rate, and a significant increase reaction time (see Figures 2 and 3). The following analyses refer only to data obtained during 10 trials before, and 5 trials after, the learning criterion had been reached. The mean number of trials required to learn four targets was 12.7.

Table 2 and 3 HERE.

Separate ANOVAs were performed for RT and hit rate data, for the learning phase (trials 1-10) and test phase (trials 11-15). No significant differences between conditions were found during the learning phase (see Table 2). During testing, a significant difference in RT, but not hit rate, was found (see Table 3).

Table 4 and 5 HERE.

These ANOVAs were augmented with paired t-tests (2-tailed) comparing differences in responses across trials 10 and 11 within each condition (see Tables 4 and 5). For both RT and hit rate, significant differences between trials 10 and 11 were found in the experimental, but not in the control condition.

A paired t-test comparing false alarm rates (i.e. the proportion of times that a distractor was classified as a target) on trials 10 (0.131) and 11 (0.161) indicated a non-significant difference ($t=-1.33$, $df=27$, $p=0.194$).

Discussion

Two current theories of object recognition are the feature-based 'geon' theory [Biederman, 1995], and the 2D characteristic view theory [Bulthoff and Edelman, 1992]. The 'geon' theory posits that salient 2D or 3D features are used for recognition, whereas the 2D characteristic view theory posits that an object is recognised by interpolating over a small number of known 2D views of that object. Neither of these theories could account for the current findings because, in the experimental condition, the only difference between a given target during learning and testing was a simple image sequence reversal. Therefore, the set of 2D views, and any 2D 'geons', or 3D 'geons' implied by structure-from-motion were identical for each learned object within both the experimental and control conditions. Other visual cues, such as shading, texture and stereo, that might be used for recognition were also the same for each learned object within each condition.

Given that there were no differences in these purely spatial cues between learning and testing, theories that depend on such cues could not account for the current findings.

The stimuli used in this experiment lie at one extreme of a continuum of spatiotemporal stimuli. This continuum includes biological motion stimuli such as continuously deformable amoebae, articulated (Johansson) figures, and smooth rigid objects. Whereas it seems intuitively obvious that the characteristic motions of deformable and articulated objects might be used for recognition, it is less obvious why the motion of rigid objects might be so used. It is noteworthy that performance was reduced, but still quite good, when a learned target's image sequence was reversed in the test phase. This may suggest that the spatiotemporal signature contributes relatively little to recognition in this experiment, and that other cues were largely responsible for the recognition of targets.

However, such comments are subject to the following caveat. The results presented here imply nothing about whether the visual system uses the direction *and* magnitude of motion recognise rotating objects, because the motion *magnitude* of all targets remained unaltered between learning and testing. Instead, the results imply only that the visual system depends on the direction of motion for recognition. If the visual system does encode both the magnitude and direction of motion then manipulating both of these might induce a larger effect than that reported here.

In any case, the fact remains that spatiotemporal signatures do appear to contribute to object recognition. Given this, the findings presented here can be explained by several mutually non-exclusive hypotheses regarding the nature of spatiotemporal signatures. Either subjects make use of a spatiotemporal signature that consists of the relative motion of points in 3-space (such information is available via structure-from-motion and other cues such as shading), or subjects use signatures that consist of the relative motion of points in the image plane, without reference to their 3D coordinates. Either of these types of spatiotemporal signatures could account for the findings, although evidence presented in [Bulthoff et al., 1997] suggests that recognition of Johansson figures depends on 2D spatiotemporal signatures. Additionally, rather than considering only 2D or 3D abstract motion vectors derived from the image sequence, it is possible that recognition depends directly upon spatiotemporal photometric information (e.g. by the motion of surface texture or specularities).

Conclusion

Primitive animals tend to rely on motion information (e.g. [Tinbergen, 1951,Lettvin et al., 1959]), suggesting that the ability to utilise simple spatiotemporal sequences may be deeply embedded in biological visual systems. If so, it is perhaps to be expected that humans learn the spatiotemporal characteristics of visual stimuli, and use these as a cue in object recognition.

Spatiotemporal continuity is a fundamental property of the physical world. Any visual organism that did not make use of 'raw' motion information, in the form of spatiotemporal signatures, would be discarding a powerful and ubiquitous visual cue for identifying salient events in that world.

Acknowledgements: Thanks to N Hunkin, J Porrill, K Gurney, G Mather, J Frisby, T Valentine, and D Buckley for many useful discussions, and to Z Ghahramani, N Hunkin and two anonymous reviewers for comments on this paper. Thanks to Zoe Driver for assistance with running experiments. Software for generating stimuli was written by J Porrill. Thanks to J Mayhew for suggesting the use of reversed image sequences. This research was supported by a by a Mathematical Biology Wellcome Fellowship (Grant Number 044823).

References

[Adelson, 1991] Adelson, E. (1991). Mechanisms for motion perception. *Optics and Photonics News*, pages 25–31.

[Becker, 1996] Becker, S. (1996). Mutual information maximization: Models of cortical self-organisation. *Network: Computation in Neural Systems*, 7(1):7–31.

[Biederman, 1995] Biederman, I. (1995). Visual object recognition. In Kosslyn, S. and Osherson, D., editors, *Invitation to Cognitive Science*, pages 121–165. MIT Press.

[Bulthoff and Edelman, 1992] Bulthoff, H. and Edelman, S. (1992). Psychophysical support for a 2d view interpolations theory of object recognition. *Proceedings National Academy of Sciences USA*, 89:60–64.

[Bulthoff et al., 1997] Bulthoff, I., Bulthoff, H., and Sinha, P. (1997). View-based representations for dynamic 3d object recognition. Technical Report 47, Max-Planck-Institut fur biologische Kybernetik.

[Johansson, 1973] Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 14:201–211.

[Lettvin et al., 1959] Lettvin, J., Maturana, H., McCulloch, W., and Pitts, W. (1959). What the frog's eye tells the frog's brain. *Proceedings of the IRE*, pages 1940–1951.

[Loftus and Masson, 1994] Loftus, G. and Masson, M. (1994). Using confidence intervals in within subjects designs. *Psychonomic Bulletin and Review*, 1(4):476–490.

[Mather et al., 1992] Mather, G., Radford, K., and West, S. (1992). Low-level processing of biological motion. *Proc. Roy. Soc. London. Series B*, 249:149–155.

[Pelli, 1997] Pelli, D. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, 10:437–442.

[Sinha and Poggio, 1996] Sinha, P. and Poggio, T. (1996). Role of learning in three-dimensional form perception. *Nature*, 384:460.

[Stone, 1993] Stone, J. V. (1993). Computer vision: What is the object? In *Prospects for AI, Proc. Artificial Intelligence and Simulation of Behaviour, Birmingham, England. IOS Press, Amsterdam.*, pages 199–208.

[Stone, 1996] Stone, J. V. (1996). A canonical microfunction for learning perceptual invariances. *Perception*, 25(2):207–220.

[Tinbergen, 1951] Tinbergen, N. (1951). *The Study of Instinct*. Oxford University Pres.

[Ullman, 1979] Ullman, S. (1979). *The interpretation of visual motion*. MIT Press.

Block 1		Learning Trials										Test Trials				
Trial Number		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Exp.	Target 1	○	○	○	○	○	○	○	○	○	○	●	●	●	●	●
Exp.	Target 2	●	●	●	●	●	●	●	●	●	●	○	○	○	○	○
Cntrl	Target 3	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Cntrl	Target 4	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●

Table 1

Experimental procedure for one of three blocks of stimuli, for a subject who requires 10 trials to learn the target objects, followed by 5 test trials. Each trial consists of 4 targets and 4 new distractors (not shown here) presented in random order, and counter-balanced for image sequence order. Targets 1 and 2 are in the experimental condition, and the order of images in these target sequences is reversed in the 5 test trials. Targets 3 and 4 are in the control condition, and the order of images in these sequences remains unaltered across trials. The symbol ($\bullet$, $\circ$) indicates whether an object's 90-image sequence is presented in ascending order ($\bullet$) (e.g. images $1 \rightarrow 90$), or descending order ($\circ$) (e.g. images $90 \rightarrow 1$) from a starting image which was chosen randomly on each trial. Any image could be chosen as the starting image without causing discontinuities in motion because all contiguous images (including frames 1 and 90) showed target views separated by the same angular rotation.

	Reaction Time			Hit Rate		
Effect	F	df	p	F	df	p
Condition	0.380	1	0.543	0.86	1	0.771
Trial	10.83	9	< 0.0001	3.16	9	< 0.005
Cond $\times$ Trial	1.29	1,9	0.245	0.297	1,9	0.975

Table 2

ANOVA results for data between trials 1 and 10.

	Reaction Time			Hit Rate		
Effect	F	df	p	F	df	p
Condition	4.69	1	0.018	2.40	1	0.133
Trial	2.09	4	0.087	0.765	4	0.550
Cond $\times$ Trial	1.62	1,4	0.175	1.68	1,4	0.159

Table 3

ANOVA results for data between trials 11 and 15.

Condition	Experimental	Control
Trial 10	1.77s	1.80
Trial 11	1.95s	1.81
t	2.52	0.145
df	27	27
p	< 0.01	0.443

Table 4

Results of 2-tailed paired t-tests for reaction time data between trials 10 and 11.

Condition	Experimental	Control
Trial 10	0.89	0.86
Trial 11	0.77	0.89
t	4.42	-1.00
df	27	27
p	< 0.001	0.326

Table 5

Results of 2-tailed paired t-tests for hit rate data between trials 10 and 11.

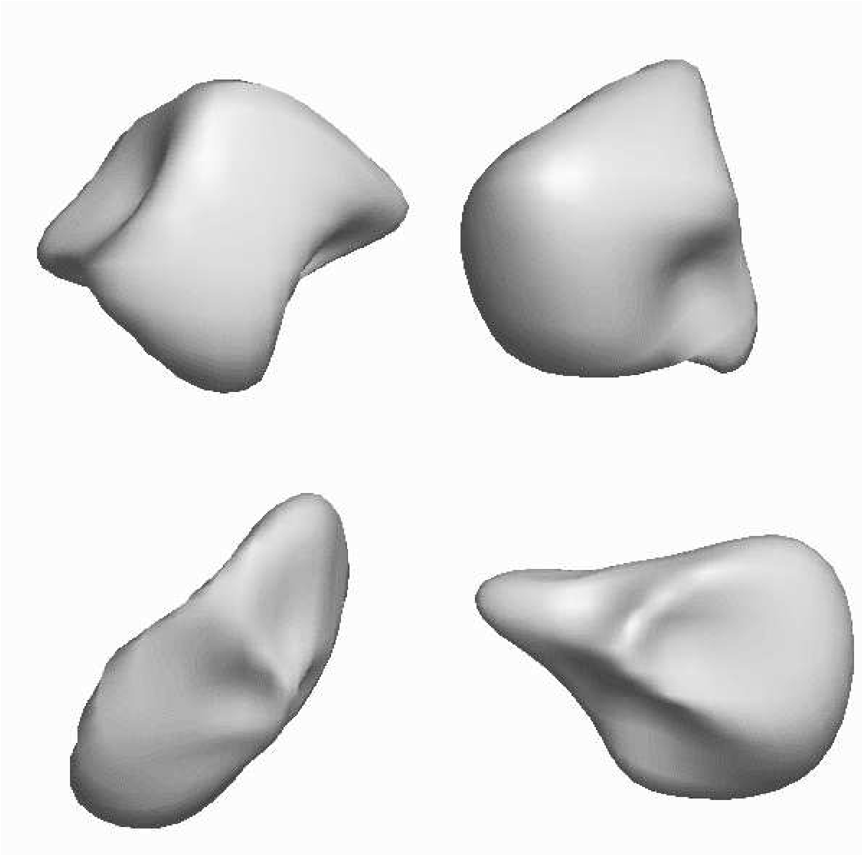

Fig. 1. Images of four different objects.

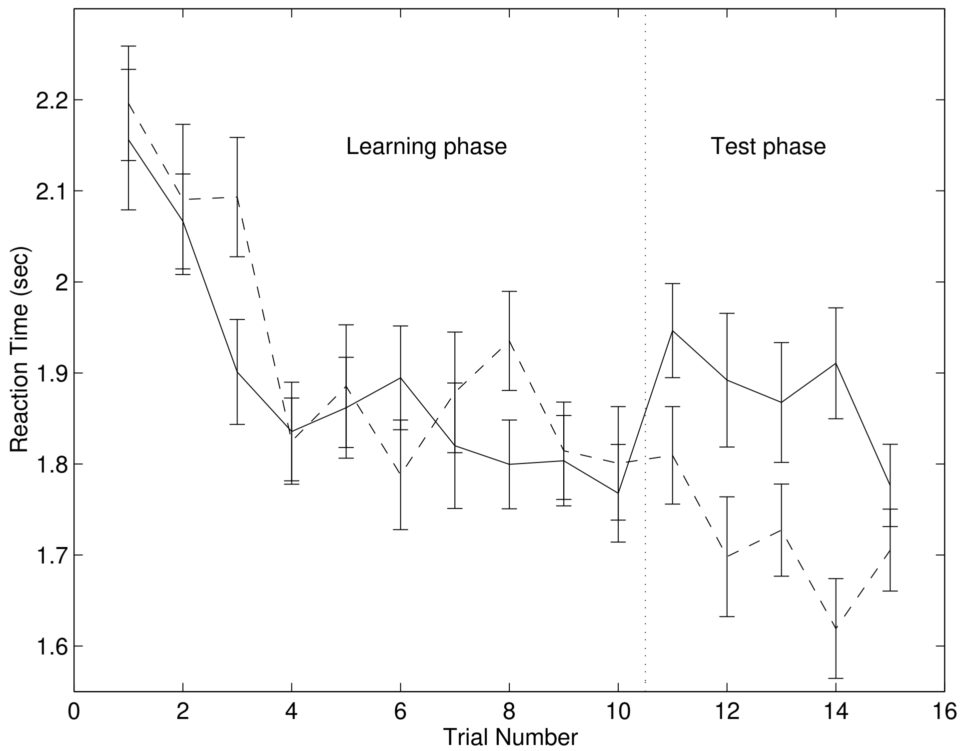

Fig. 2. Mean reaction times during learning (trials 1-10) and testing (trials 11-15), for control (dashed line) and experimental (solid line) conditions. Bars indicate standard errors (see Methods section).

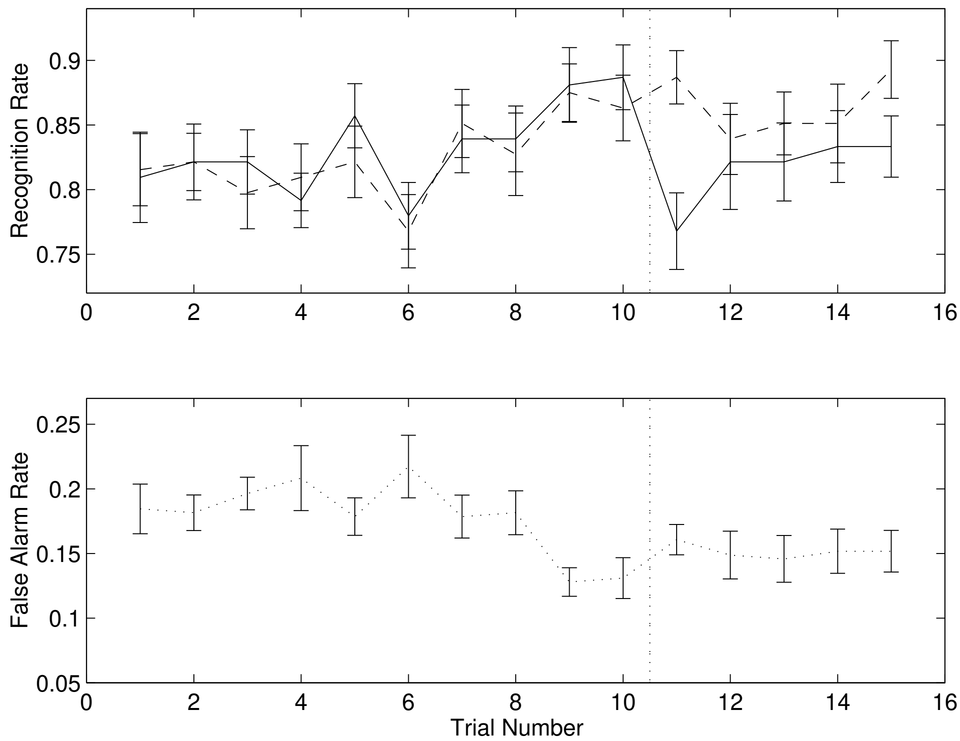

Fig. 3. Upper graph: Mean recognition rates during learning (trials 1-10) and testing (trials 11-15), during control (dashed line) and experimental (solid line) conditions. Lower graph: Mean false alarm rates during learning and testing. Both graphs are drawn to the same scale. Bars indicate standard errors (see Methods section).