PERGAMON

Vision Research 39 (1999) 4032–4044

Vision Research

www.elsevier.com/locate/visres

Object recognition: view-specificity and motion-specificity

James V. Stone *

Psychology Department, Sheffield University, Sheffield S10 2UR, UK

Received 30 October 1998; received in revised form 4 February 1999

Abstract

This paper describes an experiment to distinguish between two theories of human visual object recognition. According to the *view-specificity hypothesis*, object recognition is based on particular learned views, whereas the *motion-specificity hypothesis* states that object recognition depends on particular directed view-sequences. Both hypotheses imply a degree of *view-bias* (i.e. recognition of a given object is associated with a small number of views). Whereas the view-specificity hypothesis attributes this view-bias to a preference for particular views, the motion-specificity hypothesis attributes view-bias to a preference for particular directed view-sequences. Results presented here suggest that recognition of 3D rotating objects involves significant view-bias. This view-bias appears to be associated with an underlying bias for particular directed view-sequences, and not for particular views.

Keywords: Object recognition; Motion; Learning; Characteristic view; Motion-specificity

1. Introduction

The physical world does not present a cacophony of random events to the nervous system. There is structure, but it is hidden amongst the many layers of the sensory array. More importantly, there exist powerful and generic short-cuts for recovering this structure that do not require near-infinite spans of time for their biological evolution, nor near-infinite numbers of neurons for their learning.

The laws of physics impose severe constraints on the behaviour of an object in motion. It would, therefore, be surprising if perceptual systems had not evolved to take advantage of these physical constraints as a means of imposing corresponding computational constraints on learning. For example, the temporal proximity of images of a given object provides a *temporal binding* of its perceptually salient properties, such as orientation in 3-space. It is this temporal binding which permits us to infer that temporally proximal images are derived from similar physical scenarios (e.g. similar 3D pose). Conversely, it also permits us to infer that images separated by long time intervals are likely to be derived from different physical scenarios. The compelling logic of this type of argument suggests that a perceptual system would be at a selective disadvantage if it did not utilise temporal aspects of stimuli.

It is now well established that a proportion of neurons in the primary visual cortex are selective for particular types of motion within their small receptive fields, and that motion is detected over large retinal areas by single neurons in MT (Bradley, Grace & Anderson, 1998). Evidence that neurons in the primate inferotemporal cortex respond selectively to particular types of complex motion was first presented in (Perrett, Harries, Mistlin & Chitty, 1990). Neurons were found to respond to human walkers (or to Johansson (1973) figures[1].). The neurons were highly selective; a large neuronal response to a Johansson figure was observed, but if the figure was walking forwards whilst translating backwards then the neuronal response was much reduced. In a related study, (Mather & Murdoch, 1994) Mather created male-female cue-conflict Johansson figures by combining the 3D structure of a male figure with the dynamics of a female figure, and vice versa. Results suggested that observers' gender judgements

* Tel.: +44-114-2226522; fax: +44-114-2766515.
E-mail address: j.v.stone@sheffieldac.uk (J.V. Stone)

[1] A Johansson figure is a moving human viewed in the dark with lights attached to each major joint.

0042-6989/99/$ - see front matter
PII: S0042-6989(99)00123-6

were based on the different dynamics of male and female Johansson figures, and not on their 3D structure. This finding contradicts conventional theories which assume that motion contributes to recognition only via structure-from motion mechanisms. As to whether these types of stimuli are represented in terms of 2D or 3D information, evidence presented in (Bülthoff, Bülthoff & Sinha, 1997) suggests that they are represented as 2D traces.

Neurophysiological evidence suggests that the temporal order of static stimuli is a factor in neuronal firing rates. After prolonged exposure to a constant (randomly chosen) sequence of fractal patterns, the temporal proximity of static pictures was found to affect the activity of neurons in the anterior inferior temporal lobe of monkeys (Miyashita, 1988). These results have been modelled in (Wallis, 1998) using a Hebbian learning rule derived from (Griniasty, Tsodyks & Amit, 1993) which includes a temporal trace of the activity of model neurons. Using a similar learning rule, a model of temporal binding in recognition of faces which rotated in depth was presented in (Bartlett & Sejnowski, 1998). On a more general level, temporal continuity has been proposed as a generic heuristic for learning of perceptual invariances (Stone, 1996a). The utility of this heuristic has been demonstrated in a number of artificial neural network models (Stone, 1996b; Becker, 1992, 1996).

These computational studies are consistent with a range of psychophysical experiments which have been interpreted within a Bayesian framework in (Weiss & Adelson, 1998); the 'perceptual prior' (probability density function) adopted implies that motion tends to be slow and smooth. In a similar vein, it has been shown (Kellman & Short, 1987) that infants dis-habituate to static views of 3D objects only if they had been previously presented in continuous motion. In contrast, previous exposure to a set of static views failed to induce dis-habituation to static views. Again, these psychophysical results are consistent with computational studies which make use of temporal associations to learn to recognise objects (Edelman & Weinshall, 1991). However, Edelman and Weinshall argue for a view-specific mechanism which uses temporal proximity only as a means of providing information about which particular views belong to the same object. This contrasts with Kellman's interpretation, which states that dis-habituation could not be based on matching to particular learned views, and also with evidence presented in (Stone, 1998a) that motion per se is involved in object recognition.

The role of temporal change in terms of shifts in facial expression has also been investigated as a cue for face recognition. No advantage for moving over static faces was found for recognition of unfamiliar faces (Christie & Bruce, 1998). However, recognition of moving (famous, and therefore familiar) faces was found to be significantly better than that of static faces (Lander, Christie & Bruce, 1998), and the authors argue that the difference in performance between moving and static faces cannot be explained away by mechanisms such as shape-from-motion.

Computational vision research has demonstrated utility for motion, where it has not only been used as a means of recovering shape (Ullman, 1979; Blake, Cipolla & Zisserman, 1990), but also as a direct cue for recognition of human walkers (Niyogi & Adelson, 1994), and of characteristic motion of natural objects such as tree canopies (Nelson & Polana, 1992). Collectively, these studies suggest that motion is critical for learning to recognise objects and faces, and perhaps that motion per se might be used as a cue for recognition in human vision (Stone, 1993)

2. The view-specificity and motion-specificity hypotheses

The v*iew-specificity* hypothesis (Cutzu & Edelman, 1994a; Logothetis & Pauls, 1995) states that an object is recognised by utilising learned characteristic views, and that recognition from other views is mediated by a process of interpolation over these characteristic views. The present research was motivated by several studies (Tinbergen, 1951; Mather, Radford & West, 1992; Sinha & Poggio, 1996), including rotation-reversal experiments (Stone, 1998a,b), which are consistent with a *motion-specificity* hypothesis. This hypothesis posits that objects can be recognised on the basis of characteristic directed view-sequences, or *spatiotemporal signatures* (Stone, 1993; Niyogi & Adelson, 1994; Stone, 1995).

Both the view-specificity and motion-specificity hypotheses predict that particular views tend to be favoured over others during recognition. If a given object is repeatedly recognised on the basis of a short sub-sequence of learned images then the precise timing of recognition tends to occur toward the end of that sub-sequence, creating an *apparent* bias for views toward the end of the sub-sequence. However, whereas view-specificity attributes this *view-bias* to a preference for certain *views*, motion-specificity attributes the observed view-bias to a preference for certain directed *view-sequences*.

In Stone (1998a), it was demonstrated that, after subjects learn to recognise a novel 3D object which has a constant rotational direction, subsequent recognition performance is compromised if the direction of rotation is reversed. Specifically, subjects learned to recognise novel, 3D, rotating objects (see Fig. 1) from short movies, each of which was played in a constant temporal direction during learning. After learning, the temporal order of images in certain movies was reversed. This

rotation-reversal (i.e. playing movies of learned objects 'backwards') produced a significant reduction in recognition performance. The effects of rotation-reversal cannot be interpreted in terms of atemporal 2D nor 3D shape information derived via shape-from-shading/motion/texture cues, 2D characteristic views (Bülthoff & Edelman, 1992), nor 'geons' (Biederman, 1995), because all of these are invariant with respect to the temporal order of images. Results were, therefore, interpreted in terms of object-specific spatiotemporal signatures.

The experiments presented here are intended to investigate whether view-bias is associated with view-specificity and/or motion-specificity. Specifically, these experiments are intended to test the hypothesis that a decrease in view-bias after rotation-reversal is associated with a corresponding decrease motion specificity.

2.1. Predictions

If recognition of an object is mediated only in terms of a small number of views then the amount of view-bias should not be altered by rotation-reversal, because rotation-reversal does not affect which views are seen by a subject. However, if recognition is mediated by *directed view-sequences* then rotation reversal should reduce view-bias.

3. Methods

Two rotation-reversal experiments were run. The first experiment is described in (Stone, 1998a), which used synthetic, grey-level objects (see Fig. 1). The other experiment was similar in design: the principal difference being that objects were defined only in terms of an homogeneous texture of white dots on a black background (see Fig. 2). Results for this experiment are described below. These experiments will be referred to as GL and DOT, respectively.

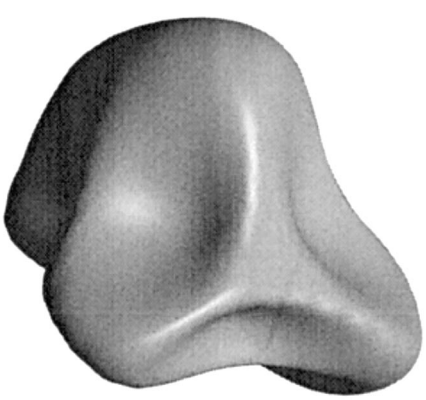

Fig. 1. Example of a grey-level object

Fig. 2. Example of a textured object

3.1. Stimuli

Experiments DOT and GL have the following features in common. The stimuli consisted of image sequences (movies) of rigid, smooth, 'amoeboid' objects rotating against a black background (see Figs. 1 and 2). Each 90-image sequence was generated using the MatLab programming language. In each sequence, one object rotated around an axis which itself rotated over time, giving the appearance of a tumbling motion. All rotations were around a fixed point, which approximated the centre of mass of the object. Each image was 300×300 pixels. Image sequences were played at a constant number of frames per second, and were displayed in a darkened room on an Apple Multiple Scan 20 computer screen (set to 1024×768 pixel resolution), using a Videotoolbox software (Pelli, 1997). Note that the starting image of each sequence was chosen at random every time it was played to prevent subjects from attending only to the first image in each movie sequence. Each movie displayed an object which began and ended with the same 3D pose, and could therefore be played for 90 frames from any starting image.

3.2. GL stimuli

Aspects of the stimuli that were unique to the GL experiment are as follows. The stimuli consisted of image sequences of grey-level objects. The obliquely placed light source was constant within and between image sequences. In each sequence, one object rotated through 360 degrees around an axis which rotated over time. All objects underwent the same set of rotational changes. Each image had 128 grey-levels. Image sequences were played at a constant rate of 25 images/s (3.6 s per movie). Subjects viewed movies at a distance of about 75 cm without a chin rest. The target and distractor objects were different for each subject, and were chosen randomly at the start of the experiment. Data from a randomly selected 24 of the 27 subjects reported in (Stone, 1998a) subjects were used, for com-

Table 1
Experimental procedure for one of three blocks of stimuli, for a subject who requires ten trials to learn the target objects, followed by five test trials[a]

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	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●

[a] Each trial consists of four targets and four 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 *E*. The temporal order of images in each of these target sequences is constant during the ten learning trials (the E_{pre} condition), and reversed in the five test trials (condition E_{post}). Targets 3 and 4 are in the control condition C, and the order of images in these sequences remains constant throughout the ten learning trials (condition C_{pre}) and the five learning trials (condition C_{post}). The symbol (●, ○) indicates whether an object's 90-image sequence is presented in ascending order (●) (e.g. images $1 \rightarrow 90$), or descending order (○) (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 a small angular rotation.

parability with data from 24 subjects used in the DOT experiment. Subjects were undergraduate psychology students.

3.3. DOT Stimuli

Aspects of the stimuli that were unique to the DOT experiment are as follows. The stimuli consisted of image sequences of uniformly textured objects. The texture consisted of white dots on a black object, displayed against a black background, so that the object outline was invisible. In each sequence, one object rotated through 600° around an axis which itself rotated over time. Each object underwent a unique set of rotational changes. Image sequences were played at a constant rate of 15 images/s (6 s/movie). Subjects viewed movies at a distance of 57 cm, with a chin rest. The target objects were the same for each subject, and were chosen randomly from an initial set of stimuli. The allocation of the 12 learned objects (and their learned direction of motion) to the experimental and control conditions was counter-balanced within the male and female groups. The subjects were 12 male and 12 female undergraduate psychology students. Note that the angular speed of object rotation was 10° per second in the DOT and GL movies.

3.4. Procedure

For both GL and DOT experiments, there were three learning blocks of about 20 min each (see Table 1). In each block, each subject learned to recognise four target objects, in a continuous recognition task, with targets being shown for between 10 and 20 trials. 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 series of test image sequences, of which half displayed a target object and half displayed a distractor object.

Target objects were presented as part of a *trial set*, which comprised four targets and four previously unseen and randomly chosen distractor objects. Elements of each trial set were displayed sequentially and in random order. 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. The starting image of each sequence was chosen at random every time it was presented, and each sequence was presented in full irrespective of when a response was made. No feedback was given at any time.

Subject performance was evaluated over each trial set within a block. 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 test trial 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.

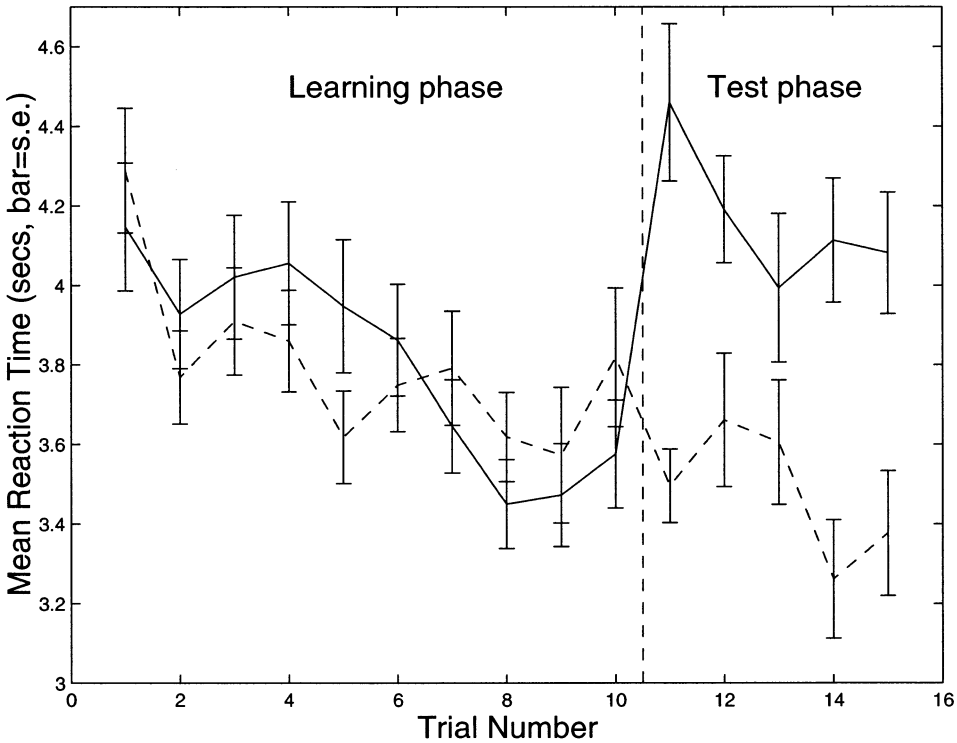

Fig. 3. DOT experiment: mean reaction times during learning (trials 1–10) and test (trials 11–15) for control (dashed lines) and experimental (solid line) conditions. Bars indicate standard errors (see Methods section). Each movie presentation lasted for 6 s.

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

4. Effects of rotation-reversal

Graphs for RT and hit-rate for the DOT experiment are displayed in Figs. 3 and 4, respectively. In order to reflect the within-subjects design, the standard error bars plotted in Figs. 3 and 4 were computed only after inter-subject variability had been removed (see (Loftus & Masson, 1994) or (Stone, 1998a) for details). The effects of rotation-reversal are larger, but qualitatively similar to, those reported in (Stone, 1998a) where GL objects were used. Indeed, the reason for using DOT objects was to increase the effects of rotation-reversal by precluding subjects from using atemporal cues such as shading. Whereas the percentage change in hit-rate was -13% in the GL experiment, it was -22% in the DOT experiment. Corresponding percentages for the RT were $+10\%$ (GL) and $+25\%$ (DOT).

Results are summarised in Tables 2–5. Rotation-reversal significantly increased RT and significantly reduced hit rates between the final learning trial (defined as trial 10) and the first test trial (defined as trial 11), for both the GL and DOT experiments[2].

It is noteworthy that in both experiments, false alarm rates were essentially unaffected by rotation-reversal, suggesting that RT and hit-rate results are not due to general cognitive effects of reversal. The false alarm rates for trials 10 and 11 were 0.131 and 0.161 for the GL experiment, and 0.229 and 0.233 for the DOT experiment; the results of paired two-tailed t-tests on these false alarm rates were $P = 0.194$ (GL) and $P = 0.846$ (DOT).

For comparison with view-bias data to be presented, the data were re-analysed in terms of the mean proportion of *valid* responses $\bar{\mu}$ made by each of 24 subjects. A valid response is defined as a correct response which is made before the presented movie ends[3]. Subject means are presented in Table 6. Results for hit-rate are consistent with those of $\bar{\mu}$ for the DOT experiment. However, whereas the GL hit-rate results show a significant effect

[2] For notational convenience, the final learning trial is defined as 'trial 10', and the first test trial is defined as 'trial 11' (in practice, subjects typically required 12 learning trials

[3] Responses made 200–300 ms after a given movie ends could have been counted as valid because these were likely to have been associated with the final displayed images of that movie. However, with mean RTs of around 3.5 s and standard errors of 0.2 s, the number of such uncounted valid responses is small, and the impact of ignoring these is therefore minimal.

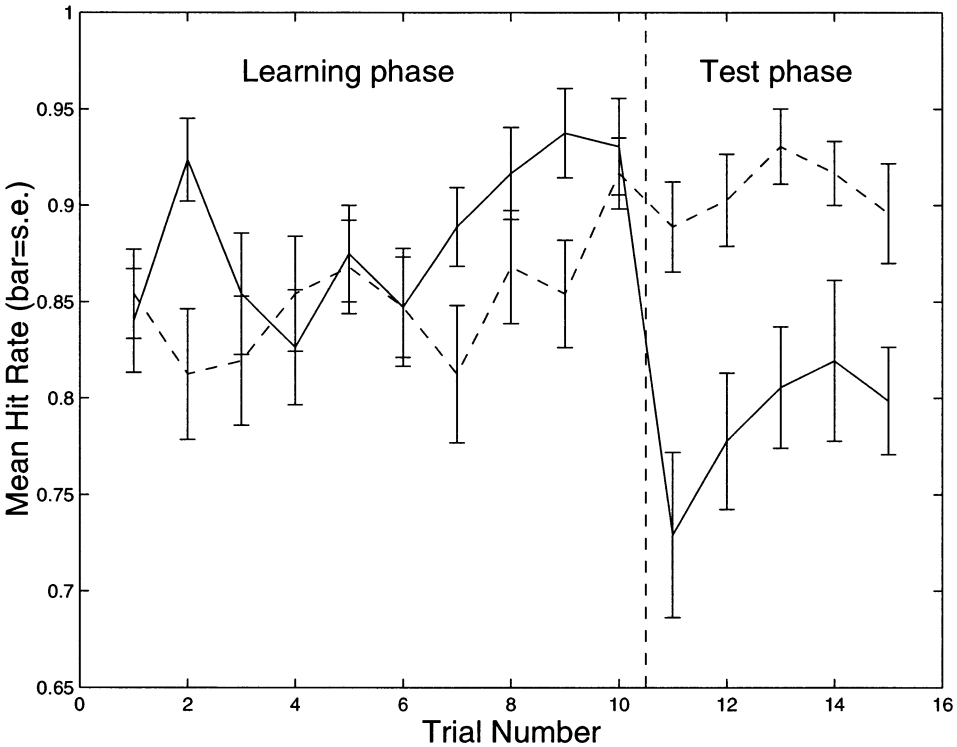

Fig. 4. DOT experiment: mean recognition rates during learning (trials 11–15), during control (dashed line) and experimental (solid line) conditions. Bars indicate standard errors (see Methods section).

of rotation-reversal, the corresponding $\bar{\mu}$ results do not. This suggests that a relatively high proportion of correct responses in the experimental condition of the GL experiment were classified as invalid because they were made after the presented movie had ended, and not because they were incorrect responses.

Overall, the effects of rotation-reversal were larger for DOT than GL objects. These two experiments differ in several respects (as described above), and it is possible that these differences contributed to the difference in results. However, it is conjectured that the differences in results are due principally to the increased reliance on motion induced by use of DOT objects.

5. Effects of rotation-reversal: view-specificity and motion-specificity

An obvious measure of view-specificity is the degree of similarity between views of an object that evoke a correct recognition response from a given subject. Unfortunately, this is also an obvious choice for measuring the degree of motion-specificity. Thus, both the view-specificity and motion-specificity hypotheses predict significant view-bias for objects learned with a single rotation direction. However, the view-specificity and motion-specificity hypotheses predict *different* amounts of view-bias after rotation-reversal, because view-bias resulting from view-specificity should be unaltered by rotation-reversal, whereas view-bias resulting from motion-specificity should be eliminated by rotation-reversal.

5.1. Measuring view-bias

The measure of view-bias b used here is defined in terms of the propensity of subjects to make a correct response[4] while a particular image of an object's movie is being displayed (see Fig. 5). Briefly, each subject responded to the kth movie over a number of trials, and the identity of the image associated with each of n_k valid responses was recorded as an angle between 1 and

Table 2
GL experiment (RT): results of two-tailed paired t-tests for the mean reaction times of subjects between trials 10 and 11

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

[4] Incorrect responses were excluded because, as might be expected, it was found that subjects incorrect responses were not associated with particular views of objects.

Table 3
GL experiment (hit rate): results of two-tailed paired t-tests for the mean hit rates of subjects 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
DOT experiment (hit rate): results of two-tailed paired t-tests for the mean hit rates of subjects between trials 10 and 11

Condition	Experimental	Control
Trial 10	0.93	0.917
Trial 11	0.73	0.89
t	−4.87	1.282
df	23	23
P	<0.0001	0.213

360°. The view-bias is then defined as $b_k = n_k|r_k|^2$, where $|r_k|$ is the length of the mean resultant vector of these angles. The value of $|r_k|$ is close to zero if randomly chosen images are associated with valid responses, and approaches unity if similar views are associated with correct responses. Defining b_k as $n_k|r_k|^2$ (Fisher, 1995) ensures that that the expected value of b_k is unity for any number of valid responses, with large values of b_k implying a large view bias (see Appendix A).

View-bias for each movie was measured over the final five learning trials (the *pre* trials), and also over the five test trials (the *post* trials). Each of the S subjects learned to recognise six control and six experimental objects, making a total of $R = 12S$ sets of ten responses per subject. These were classified into two control conditions, C_{pre} and C_{post} and two experimental conditions E_{pre} and E_{post}, where the pre and post subscripts refer to trials before and after rotation-reversal, respectively. Recall that control objects did not undergo rotation-reversal after learning whereas experimental objects did. View-bias for each condition was defined as the mean view-bias over all movies in that condition.

5.2. Do subjects respond to particular views?

View-bias is based on variables with a circular distribution, so that its significance cannot be assessed using statistics based on Gaussian distributions (Fisher, 1995). Consequently, statistical tests based on *randomisation* are used (see Appendix B).

Evidence that subjects respond to particular views is given by pooled view-biases from the GL and DOT results. as shown in Table 7. Significant levels of view-bias ($P < 0.01$) were found in all conditions except the post-reversal experimental condition E_{post} ($P = 0.617$). Thus *view-bias is significantly greater than chance except after rotation-reversal.*

A similar pattern emerges if data from the two experiments are analysed separately (see Tables 8 and 9). Again, a non-significant view-bias was obtained for experimental condition E_{post} in both GL and DOT experiments, with ($P > 0.45$) in both cases. For the GL experiment, significant view-biases ($P < 0.05$) were found in control conditions C_{pre} and C_{post}, whereas view-bias approached significance in the experimental condition E_{pre} ($P = 0.089$) but not in E_{post} ($P = 0.630$). Similarly, for the DOT experiment, view-bias approached significance in both control conditions C_{pre} ($P = 0.060$) and C_{post} ($P = 0.103$), whereas significant view-bias was found in experimental condition E_{pre} ($P = 0.003$) but not in E_{post} ($P = 0.476$).

These results suggest that the view-bias observed before rotation-reversal was due to motion-specificity, and not to view-specificity, because any view-specificity acquired during learning should survive rotation-reversal. In contrast, and by *definition*, any motion-specificity acquired during learning could not survive rotation-reversal.

Table 4
DOT experiment (RT): two-tailed paired t-tests for the mean reaction times of subjects between trials 10 and 11.

Condition	Experimental	Control
Trial 10	3.58s	3.82s
Trial 11	4.46s	3.50s
t	−4.04	1.78
df	23	23
P	<0.001	0.089

Table 6
Results of paired t-tests (2-tailed) for the mean difference $\bar{\mu}_{\text{pre-post}}$ in proportion of valid responses $\bar{\mu}$ per subject before and after rotation-reversal, for the control and experimental conditions of the GL and DOT experiments[a]

Experiment	GL		DOT	
Condition	Control	Experimental	Control	Experimental
$\bar{\mu}_{\text{pre}} - \bar{\mu}_{\text{post}}$	−0.026	0.026	−0.046	0.144
T	−1.860	1.336	−2.264	5.986
df	23	23	23	23
$p(\bar{\mu}_{\text{pre-post}})$	0.076	0.195	0.033	<0.001

[a] Note that the significant results for the control condition in the DOT experiment was for an *increase* in $\bar{\mu}$, whereas the corresponding significant results for the experimental condition was for a decrease in $\bar{\mu}$.

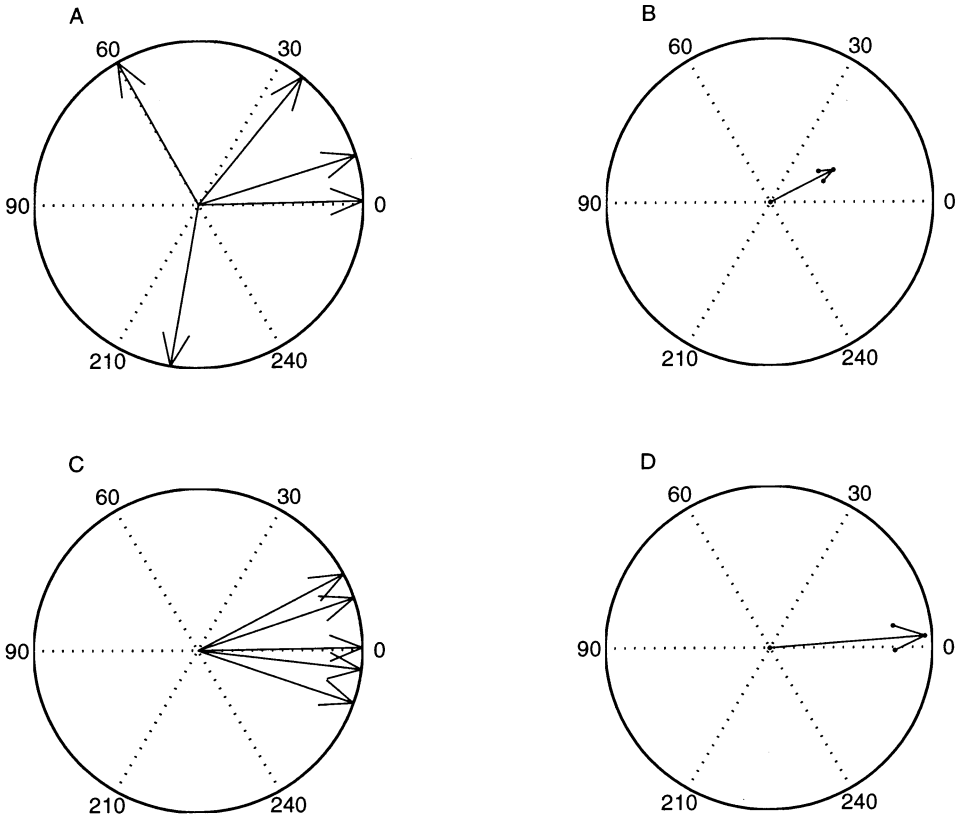

Fig. 5. Diagrammatic representation of small (A) and large (C) view-bias for five consecutive simulated responses to the kth movie, and the corresponding mean resultant vector r_k associated with each set of responses (B and D, respectively). (A, C) The five image identity numbers (1–90) of object views which evoke correct responses are multiplied by four to produce a set of five image-angles. This circular representation of image identity numbers is justified because images 1 and 90, which correspond to image angles 4 and 360, show similar views of an object. As displayed in A and C, each image-angle can be plotted as a unit vector with a direction defined by one image-angle. (A) For $n_k = 5$ randomly chosen response image angles, the mean resultant vector shown in B has a short length $|r_k|$, denoting a small view bias. (C) For responses that cluster around a single image-angle, the mean resultant vector shown in D has a length $|r_k|$ approaching unity, denoting a large view bias. In practice, view bias is defined as $b_k = n_k|r_k|^2$ (see Appendix A). View-bias b within a single condition is defined as the mean value of b_k over a total of 144 sets of five responses.

5.3. Does rotation-reversal decrease view-bias?

If rotation-reversal decreases view-bias then ($b_{post} < b_{pre}$). The significance of the difference $\gamma = (b_{pre} - b_{post})$ in view-bias before b_{pre} and after b_{post} rotation-reversal was computed using a (1-tailed) Wilcoxon matched-pairs signed-ranks test[5]. Results are displayed in Table 10.

For the pooled (GL + DOT) view-bias data, a significant decrease ($P = 0.024$) in view-bias was found between conditions E_{pre} and E_{post}, but not between the control conditions C_{pre} and C_{post} ($P = 0.389$).

For the individual GL and DOT experiments, decreases in view-bias for the experimental conditions approached significance, with ($P = 0.061$) and ($P = 0.057$). respectively. As predicted, no significant change in view-bias occurred in the control conditions of experiments GL or DOT with $P = 0.430$ and 0.194, respectively.

Table 7
Pooled GL and DOT view-bias data: a mean view-bias b before and after rotation reversal, and the probability $p(b)$ of obtaining the observed values of b (or greater) by chance[a]

Condition	Control		Experimental	
	Pre-reversal	Post-reversal	Pre-reversal	Post-reversal
View-bias (b)	1.121	1.147	1.145	0.986
$P(b>1)$	0.008	0.003	0.003	0.617

[a] Each value of $p(b)$ is based on 10 000 randomisation samples (see Appendix B).

[5] This probability could not be computed using a randomisation test because the numbers of valid items for a given movie can differ in the pre- and post-reversal conditions. Therefore, the underlying distributions of pre- and post-reversal view-biases are different. In the absence of an appropriate normalising term (e.g. standard deviation) such a difference precludes the use of randomisation procedures (see Appendix B).

Table 8
GL experiment: view-bias b before and after rotation-reversal[a]

Condition	Control		Experimental	
	Pre-reversal	Post-reversal	Pre-reversal	Post-reversal
View-bias (b)	1.136	1.204	1.093	0.972
$P(b>1)$	0.031	0.003	0.089	0.630

[a] See caption of Table 7 for details.

The results are unequivocal for the pooled (GL + DOT) data, showing significant amounts of view-bias except after rotation-reversal, and a significant reduction in view-bias only after rotation-reversal. This suggests that the less consistent results for the individual GL and DOT data are due to a lack of statistical power. The overall pattern of results is consistent with the hypothesis that motion-specificity, and not view-specificity, was responsible for the observed view-bias and for the reduction in view-bias following rotation-reversal.

6. Discussion

The findings presented here are consistent with the hypothesis that recognition of objects in motion depends on learned directed view-sequences. The resultant view-bias is largely eliminated by rotation-reversal, suggesting that view-bias is due to motion-specificity, and not to view-specificity in the experiments reported here.

6.1. Effect Size of rotation-reversal

The magnitude of the effects of rotation-reversal are comparable with the effects of other visual transformations used in object recognition experiments. Rotation-reversal reduces recognition rates by between 13% and 22% in the experiments reported here. However, object recognition is notoriously robust with respect to experimental manipulation, and reductions in performance of the magnitude observed here can be induced only by quite radical changes, such as a change to a novel viewpoint. For example, in (Bülthoff & Edelman, 1992), the largest decrease in recognition (around 35%) was associated with a viewpoint change of 60° using 'paper-clip' objects; and it is reported that similar results were obtained using amoeboid objects like those used here. Similarly, in (Hayward & Tarr, 1997), recognition performance of synthetic objects with well defined visual features decreased by around 6% following a change in viewpoint of 30°. These figures suggest that rotation-reversal produces changes in recognition performance on a par with more radical interventions, such as a change in viewpoint.

Table 9
DOT experiment: view-bias b before and after rotation-reversal[a]

Condition	Control		Experimental	
	Pre-reversal	Post-reversal	Pre-reversal	Post-reversal
View-bias (b)	1.105	1.089	1.198	1.00
$P(b>1)$	0.060	0.103	0.003	0.476

[a] See caption of Table 7 for details.

6.2. Implications of motion-specificity

Evidence that recognition is mediated by particular static views has previously been obtained from experiments in which subjects were trained with rotating objects and then tested with static views (such experiments use objects which rotate back and forth through a small angle). These experiments show that recognition is sensitive to viewpoint (e.g. Bülthoff & Edelman, 1992) and demonstrate the existence of view-bias (e.g. Cutzu & Edelman, 1994b). However, results presented here suggest that this type of experimental design produces motion-specificity, and that the observed view-bias can be attributed to a preference for particular learned view-sequences. It follows that, if subjects are tested with static views which happen to be part of a learned view-sequence then a preference for particular views will be inferred. Thus, the view-bias observed in such experiments may depend on an underlying preference for particular view-sequences (learned view-sequences should be independent of a 'forwards' or 'backwards' directions in these experiments because presented objects rotate back and forth). This is not intended to suggest that view-specificity cannot mediate recognition (indeed, it has been demonstrated that static views of novel objects are sufficient for learning and recognition (Bülthoff & Edelman, 1992)); it is intended to suggest that the interpretation of view-bias requires some care if subjects are trained using rotating objects.

6.3. Experimental critique

From a purely logical viewpoint, what explanations other than motion-specificity might account for the effects of rotation-reversal?

6.3.1. Subjects learned 'early' views of objects

It may be thought that the observed view-bias resulted from preferential learning of views at the beginning of different presentations of a given movie. However, as stated above, this cannot occur because

Table 10
Difference in view-bias $\gamma = (b_{pre} - b_{post})$ before and after rotation-reversal, and associated 1-tailed significance p for control and experimental conditions of the GL, DOT and (GL+DOT) pooled data[a]

Experiment	GL		DOT		GL+DOT	
Condition	Control	Experimental	Control	Experimental	Control	Experimental
γ	−0.079	0.134	0.018	0.164	−0.031	0.149
P	0.430	0.061	0.194	0.57	0.389	0.024

[a] Results were obtained using a (1-tailed) Wilcoxon matched-pairs signed-ranks test, with each rank being associated with one movie presentation.

the starting image of each movie was chosen at random on each occasion it was presented.

6.3.2. Cognitive effects

General cognitive effects can be discounted because, not only is there no reduction in the recognition of control objects which are inter-leaved with experimental objects, but the false alarm rate for distractor objects is essentially unaltered by rotation-reversal.

6.3.3. 'Forward' and 'backward' movies imply different 3D shapes

It is possible that the 3D perception of an object depends on the direction of rotation. From a purely information-theoretic standpoint, the 3D shape information implied by a movie is independent of whether its images are displayed in ascending or descending order. Nevertheless, the human visual system may interpret a movie of a rotating object as one shape, and a 'backward' version of that movie as a slightly different shape. The experiments presented here cannot be used to address this issue.

6.3.4. The stimuli are 'unnatural'

The objects were not everyday objects, but they were not unnatural in any obvious sense. In order to argue that the observed effects were peculiar to the type of object used, one would have to demonstrate that the effects of rotation-reversal only applied to certain classes of objects. On the basis of recent results on recognition of faces (Christie & Bruce, 1998; Lander et al., 1998), such a class is more likely to be defined in terms of familiarity than in terms of physical properties, such as shape.

6.4. Motion-specificity or sequence-specificity?

Throughout this paper, the effects of rotation-reversal have been attributed to motion-specificity. However, motion could contribute in at least two ways to object recognition: recognition from optic flow, or from temporal order (*sequence-specificity*)[6]. Recognition from optic flow requires that an object can be recognised from the particular optic flow that its motion implies[7]. In contrast, recognition from temporal order requires that an object can be recognised from the particular view-sequence its motion generates, without appealing to mechanisms for computing the object's motion per se. The current paper is not designed to investigate this critical distinction, and the effects reported here could depend on both optic flow and sequence-specificity. However, it is noteworthy that rotation-reversal should have a profound effect on recognition based on both optic flow and temporal sequence.

7. Conclusion

Rotation-reversal does not completely prevent recognition, but it does largely eliminate view-bias. A proportion of recognition performance must therefore depend on motion-independent mechanisms. The point of this paper is not to show that motion-specificity is the predominant cue for recognition, but that the decrease in recognition performance following rotation-reversal can be attributed to motion-specificity.

It might be argued that there is a contradiction between motion as a cue to recognition via structure-from-motion mechanisms, and motion as a direct cue to recognition. In principle, motion can be used as a cue for either or both of these. It follows that an efficient perceiver would make direct use of temporal cues for recognition, as well as using motion as a means of estimating the underlying atemporal structure of the physical world.

As efficient perceivers, we should be able to make use of temporal change, not only as a cue for gauging the atemporal structure of the physical world, but also as a cue in its own right. Evidence presented here suggests that this is precisely what we do.

[6] Thanks to N Hunkin for pointing this out.

[7] An object's motion defines two related optic flow patterns, one on the retina, and one in 3-space; and either of these may contribute to the recognition process.

Acknowledgements

Thanks to N. Hunkin, R. Lister, J. Frisby, D. Buckley, J. Porrill, G. Mather, and J. Mayhew for useful discussions, and to B. Tjan and one anonymous referee for their detailed comments. This research was supported by a Mathematical Biology Wellcome Fellowship (grant no. 044823).

Appendix A. Measuring view-bias

Given five successive presentations of the kth movie to a subject, each of ($n_k \leq 5$) correct responses is associated with a particular image identity number of between 1 and 90. Each image identity number can be converted to an image angle $\alpha_t = 4a_t$, where $t = \{1,\ldots, n_k\}$. Each image-angle can then be plotted as a unit vector, and the mean resultant n_k of the set of n_k unit vectors is given by:

$$r_k = \left(\frac{1}{n_k}\sum_{t=1}^{n_k}\cos(\alpha_t), \frac{1}{n_k}\sum_{t=1}^{n_k}\sin(\alpha_t)\right).$$

The length $|r_k|$ of r_k is close to zero if randomly chosen images are associated with valid responses, and approaches unity if similar views are associated with correct responses. If we rewrite r_k as $r_k = (C_k, S_k)$ then its length is given by

$$|r_k| = \sqrt{C_k^2 + S_k^2}.$$

The measure $|r_k|$ tends to over-estimate view-bias for small numbers of valid responses. The value of $|r_k|$ for the kth movie is based on $n_k < 5$ responses, and tends to increase as n_k decreases. This can be seen from two limiting cases. First, if $n_k = 1$ then $|r_k| = 1$. Second. if $n_k = \infty$ randomly chosen angles, a scattergram of these angles describes an isotropic distribution of points, so that $|r_k| \approx 0$. The undesirable effect of n_k on $|r_k|$ can be accommodated by the transformation $b_k = n_k|F_k|^2$ (Fisher, 1995). If responses are made randomly then the expected value of b_k is unity. with larger values indicating increased view-bias. The view-bias b for a given condition is defined as the mean value of b_k over all K movies in that condition,

$$b = (1/K)\sum_{k=1}^{K} b_k.$$

The conversion of image identity numbers to image-angles can be justified because the first and final images in each movie display similar views of an object. Movies were, therefore played for 90 images from a randomly chosen starting image without producing a motion discontinuity between images 1 and 90. For example. if the randomly chosen starting image identity number for a given presentation of a movie is $i = 15$ then the movie can be played from image 15 to image (15 + 90 modulo 90). The conversion from image identity number to image-angle implies that images displayed in close temporal proximity (e.g. images 85 and 1) are separated by small image-angles.

A.1. Reaction time does not affect measured view-bias

It might be supposed that view-bias measured before rotation-reversal cannot be compared to view-bias measured after rotation-reversal, due to the effects of subject reaction time. However, consider a subject who consistently responds some (reaction) time Δt s after a particular image I_j *is* shown, such that N intervening images are shown before a key is pressed. The image identity number associated with the response is then $j + N$[8]. Now, if the movie is played backwards, and the subject continues to respond Δt s after image I_j *is* shown, then the image identity number associated with the subject's response is now $j - N$ (because image identity numbers increase as the movie is played forwards' and decrease as it is played backwards). Even though the subject responds to the same image before and after rotation-reversal, the recorded image identity number before rotation-reversal is $j + N$, and is $j - N$ after rotation-reversal. However, the measured view-bias before rotation-reversal is the same as that measured after rotation-reversal, because view-bias measures the extent to which any single view is associated with subject responses, without regard to which particular view is associated with responses. Thus, if a subject consistently presses a response key while the $(j + N)$th image is displayed (e.g. before rotation-reversal) then the measured view-bias is the same as if that subject consistently presses a response key while the $(N - j)$th image is displayed (e.g. after rotation-reversal). Therefore, view-bias can be compared before and after rotation-reversal for a given move.

Appendix B. Randomisation for estimating the significance of view-bias

Randomisation is a numerical technique which permits levels of significance to be associated with data values even if the data is derived from a population with an unknown statistical distribution. Given a data array D with K columns (movies) and $n = 5$ rows (trials per movie), the view-bias b_k associated with each of the K presented movies can be computed, from which the mean value b_D of b_k for that condition can be obtained. What is the probability $P(b)$ that a value of b equal to (or greater than) b_D could have been generated by a subject whose responses coincide with all movie images

[8] These image identity numbers are computed modulo 90, but this is omitted for the sake of clarity here.

with equal probability (i.e. a subject who responds at random)?

An answer can be obtained by repeatedly filling D with data generated from a uniform circular distribution, and computing b for each array of random data. Like the observed data, the random data values were chosen from the set {4, 8, 12, 16,…, 356, 360}. A histogram $h(b)$ of values of b is similar to what would be observed if subjects responded at random. After $h(b)$ has been normalised so that it has unit area, it is an approximation to the probability density function (pdf) $f(b)$. The probability of obtaining an observed value b_D (or greater) by chance is defined by the integral of the function $f(b)$ between $b = b_D$ and $b = \infty$:

$$p(b_D) = \int_{b=b_D}^{\infty} f(b)\, db. \quad (2)$$

Given that $h(b)$ *is* an approximation to f *(b)*, $p(b_D)$ can be estimated by finding the area under $h(b)$ defined by all values greater than (or equal to) b_D.

B.1. Randomisation with invalid data items

Randomisation is complicated by the presence of invalid data items in the $K \times n$ data array D for a given condition, where K is the number of movies learned and n *is* the number of trials per movie. For each column, only the $n_k \leq n$ valid responses are used in the computation of view-bias. A valid response is a correct response made before a movie ends. Invalid data were accommodated by leaving the locations of invalid items in D fixed, and by changing values only of valid items in the randomisation process (invalid data is ignored when computing means, etc.).

For a given movie, the number of valid responses determines the underlying distribution from which randomly generated view-bias values are computed, and from which the observed view-bias is derived. The significance of view-bias could therefore be computed for each movie, and implies that the significance of the mean view-bias across all movies can be computed (as here) using randomisation. In contrast, the significance of the difference between pre-reversal and post-reversal view-biases of a given movie cannot be computed using randomisation because corresponding pre-reversal and post-reversal view-biases have different underlying distributions (due to unequal numbers of valid responses in pre- and post-reversal conditions). By implication, this precludes computing the difference in *mean* view-bias between pre-reversal and post-reversal conditions using randomisation. Instead, a Wilcoxon matched-pairs signed ranks test is used with each rank is associated with one movie.

References

Bartlett, M. S., & Sejnowski, T. J. (1998). Learning viewpoint-invariant face representations from visual experience in an attractor network. *Network*: *Computation in Neural Systems*, *9*(3), 399–418.

Becker, S. (1992). Learning to categorize objects using temporal coherence. In J. E. Moody, S. H. & Lippmann, R, *Neural Information Processing Systems* 4, (pp. 361–368). San Mateo, CA: Morgan Kaufmann.

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

Biederman, I. (1995). Visual object recognition. In Kosslyn, S. & Osherson, D., *Invitation to Cognitive Science* (pp. 121–165). MIT Press.

Blake, A., Cipolla, R., & Zisserman. A. (1990). Toward qualitative vision: Motion parallax. In *BMVC90*, pp. 115–120.

Bradley, D., Grace, C., & Anderson, R. (1998). Encoding of three-dimensional structure-from-motion by primate area MT neurons. *Nature*, *392*, 714–717.

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

Bülthoff, I., Bülthoff, H., & Sinha, P. (1997). View-based representations for dynamic 3D object recognition. *Technical Report 47*, Max-Planck-Institut fur biologische Kybernetik.

Christie, F., & Bruce, V. (1998). The role of dynamic information in the recognition of unfamiliar faces. *Memory and Cognition*, *26*(4), 780–790.

Cutzu, F., & Edelman, S. (1994a). Viewpoint-dependence of response time in object recognition. *Vision Research*, *34*, 3037–3056.

Cutzu, F., & Edelman, S. (1994b). Viewpoint-dependence of response time in object recognition. *Vision Research*, *34*, 3037–3056.

Edelman, S., & Weinshall, D. (1991). A self-organising multiple-view representation of 3D objects. *Biological Cybernetics*, *64*, 209–219.

Fisher, N. I. (1995). *Statistical analysis of circular data*. Cambridge: Cambridge University Press.

Griniasty, M., Tsodyks, M. V., & Amit, D. J. (1993). Conversion of temporal correlations between stimuli to spatial correlations between attractors. *Neural Computation*, *5*, 1–17.

Hayward, G., & Tarr, M. (1997). Testing conditions for viewpoint invariance in object recognition. *Journal of Experimental Psychology*, *23*(5), 1511–1521.

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

Kellman, P., & Short, K. (1987). Development of three-dimensional form perception. *J. Experimental Psychology*: *Human Perception and Performance*, *13*(4), 545–557.

Lander K., Christie, F., & Bruce, R, (1998). The role of movement in the recognition of famous faces. (Submitted).

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

Logothetis, N. K., & Pauls, J. (1995). Psychophysical and physiological evidence for viewer-centred object representations in the primate. *Cerebral Cortex*, *3*, 270–288.

Mather, G., & Murdoch, I. (1994). Gender discrimination in biological motion displays based on dynamic cues. *Proceedings of the Royal Society*, *London*, *Series B*., *258*, 273–279.

Mather, G., Radford, K., & West, S. (1992). Low-level processing of biological motion. *Proceedings of the Royal Society*, *London*, *Series B*, *249*, 149–155.

Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature*, *335*, 817–820.

Nelson, R., & Polana, R. (1992). Qualitative recognition of motion using temporal texture. *CVGIP*: *Image Understanding*, *56*(1), 78–89.

Niyogi, S., & Adelson, E. (1994). Analyzing gait with spatiotemporal surfaces. *Workshop on Non-Rigid Motion and Articulated Objects*, Austin, USA, (wwm-bcs. mitt edu/people/adelson/papers.html).

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

Perrett, D., Harries, M., Mistlin, A., & Chitty, A. (1990). Three stages in the classification of body movements by visual neurons. In Barlow, H., Blakemore, C., & Weston-Smith, N. J., *Images and Understanding*, (pp. 94–97). Cambridge: Cambridge University Press.

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

Stone, J. V. (1991) Computer vision: What is the object? In *Prospects for AI*, *Proc. Artificial Intelligence and Simulation of Behaviour*, Birmingham, England. (pp. 993–208) IOS Press, Amsterdam.

Stone, J. V. (1995). The sinking is green stone: using spatio-temporal cues in vision. *Image Processing*, *7*(4), 20–25.

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

Stone, J. V. (1996b). Learning perceptually salient visual parameters through spatiotemporal smoothness constraints. *Neural Computation*, *8*(7), 1463–1492.

Stone, J. V. (1998a). Object recognition using spatiotemporal signatures. *Vision Research*, *38*(7), 947–951.

Stone, J. V. (1998b). Recognition of rigid johansson objects mediated by spatiotemporal signatures. ECVP Abstract. *Perception*, *27*, 119.

Tinbergen, N. (1951). *The Study of Instinct*. Oxford: Oxford University Press.

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

Wallis, G. (1998). Spatio-temporal influences at the neural net level of object recognition. *Network*: *Computation in Neural Systems*, *9*, 265–278.

Weiss, Y., & Adelson, E. (1998). Slow and smooth: a Bayesian theory for the combination of local motion signals in human vision. *AI Memo JIIT 1624*, 23.