Object Recognition: View-Specificity and Motion-Specificity

James V Stone

Psychology Department, Sheffield University, Sheffield, S10 2UR, England.
Tel: 0114 222 6522 Fax: 0114 276 6515
Email: j.v.stone@shef.ac.uk Web: http://www.shef.ac.uk/psychology/stone

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.

Key words: Object recognition, motion, learning, characteristic view, motion-specificity.
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 legitimately 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 et al., 1998). Evidence that neurons in the primate inferotemporal cortex respond selectively to particular types of complex motion was first presented in (Perrett et al., 1990). Neurons were found to respond to human walkers (or to Johansson (Johansson, 1973) figures¹). The neurons

¹ A Johansson figure is a moving human viewed in the dark with lights attached
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 and 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 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 (Bulthoff et al., 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 (Griniasy et al., 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 and 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, 1996; Becker, 1992).

These computational studies are consistent with a range of psychophysical experiments which have been interpreted within a Bayesian framework in (Weiss and Adelson, 1998); the ‘perceptual prior’ adopted implies that motion tends to each major joint.
to be slow and smooth. In a similar vein, it has been shown (Kellman and 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 and 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 and Bruce, 1998). However, recognition of moving (famous, and therefore familiar) faces was found to be significantly better than that of static faces (Lander et al., 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 (Blake et al., 1990; Ullman, 1979), but also as a direct cue for recognition of human walkers (Niyogi and Adelson, 1994), and of characteristic motion of natural objects such as tree canopies (Nelson and 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).
The View-Specificity and Motion-Specificity Hypotheses

The view-specificity hypothesis (Cutzu and Edelman, 1994a; Logothetis and 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 et al., 1992; Sinha and Poggio, 1996), including rotation-reversal experiments (Stone, 1998a; Stone, 1998b), which suggest that objects can be recognised on the basis of characteristic directed view-sequences, or spatiotemporal signatures (Stone, 1993; Niyogi and 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.

Insert Figure 1+2 here

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 Figure 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 ob-
jects ‘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 (Bulthoff and 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 intended to test the hypothesis that a decrease in view-bias after rotation-reversal is associated with a corresponding decrease motion-specificity.

**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.

**Methods**

Two rotation-reversal experiments were run. The first experiment is described in (Stone, 1998a), which used synthetic, grey-level objects (see Figure 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 Figure 2). Results for this experiment are described
below. These experiments will be referred to as GL and DOT, respectively.

**Stimuli:** Experiments DOT and GL have the following features in common. The stimuli consisted of image sequences of rigid, smooth, ‘amoeboid’ objects rotating against a black background (see Figures 1 and 2). Each 90-image sequence was generated using the MatLab programming language. In each 90-image 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 300x300 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 1024x768 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 part of each stimulus. 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.

**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/second. Subjects viewed movies at a distance of about 75cm 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 comparability with data from 24 subjects used in the DOT experiment. Subjects were undergraduate psychology students.
**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 degrees 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/second. Subjects viewed movies at a distance of 57cm, 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.

![Insert Table 1 here](image)

**Procedure**

For both GL and DOT experiments, there were three learning blocks of about 20 minutes 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 4 targets and 4 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.

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 → 90, as denoted by a ‘●’ in Table 1), and half in descending order (e.g. 90 → 1, as denoted by a ‘○’ in Table 1) during learning and testing, in both the experimental and control conditions.
Effects of Rotation-Reversal

Graphs for RT and hit-rate for the DOT experiment are displayed in Figures 3 and 4, respectively. In order to reflect the within-subjects design, the standard error bars plotted in Figures 3 and 4 were computed only after inter-subject variability had been removed (see (Loftus and Masson, 1994) and (Stone, 1998a) for details). The effects of rotation-reversal are larger, but qualitatively similar to, those reported in (Stone, 1998a) for GL objects. 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).

Insert Figures 3+4 here

Insert Tables 2+3+4+5 here

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

\footnote{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)}
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 $\bar{v}$ made by each of 24 subjects. A valid response is a correct response which is made before the presented movie ends. Subject means are presented in Table 6. Results for hit-rate are consistent with those of $\bar{v}$ for the DOT experiment. However, whereas the GL hit-rate results show a significant effect of rotation-reversal, the corresponding $\bar{v}$ 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 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 difference is due principally to the increased reliance on motion induced by use of DOT objects.

**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 rota-
tion direction. However, the view-specificity and motion-specificity hypotheses predict different amounts of view-bias after rotation-reversal, because view-specificity view-bias is unaltered by rotation-reversal, whereas any motion-specificity view-bias is eliminated by rotation-reversal.

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\(^3\) while a particular image of an object’s movie is being displayed (see Figure 5). Briefly, each subject responded to the $k$th movie over a number of trials, and the identity of the image associated with each of $n_k$ correct responses was recorded as an angle between 1 and 360 degrees. The view-bias is then defined as $b_k = n_k |r_k|^2$, where $|r|$ the amplitude of the vector-sum of these angles. The length $|r_k|$ of the resultant vector 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 $n_k |r_k|^2$ 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 1).

Insert Figure 5 here

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 10 responses per subject. These were classified into

\(^3\) Incorrect responses were excluded because, as might be expected, it was found that subjects’ incorrect responses were not associated with particular views of objects.
two control conditions, $C_{\text{pre}}$ and $C_{\text{post}}$, and two experimental conditions $E_{\text{pre}}$ and $E_{\text{post}}$; where the \textit{pre} and \textit{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.

\textit{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 \textit{randomisation} are used to evaluate view-bias here (see Appendix 2).

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 $E_{\text{post}}$ ($p = 0.617$). Thus \textit{view-bias is significantly greater than chance except after rotation-reversal.}

\begin{center}
\textbf{Insert Table 7+8+9 here}
\end{center}

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 $E_{\text{post}}$ 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 $C_{\text{pre}}$ and $C_{\text{post}}$, whereas view-bias approached significance in the experimental condition $E_{\text{pre}}$ ($p = 0.089$) but not in $E_{\text{post}}$ ($p = 0.630$). Similarly, for the DOT experiment, view-bias approached significance in both control conditions $C_{\text{pre}}$ ($p = 0.060$) and $C_{\text{post}}$
(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.

**Does Rotation-Reversal Decrease View-Bias?**

If rotation-reversal decreases view-bias then ($b_{post} < b_{pre}$). For a given change in view-bias associated with rotation-reversal, the significance of the difference $\gamma = (b_{pre} - b_{post})$ in view-bias was computed using a (1-tailed) Wilcoxon matched-pairs signed-ranks test\(^4\). Results are displayed in Table 10.

| Insert Table 10 here |

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 =

\(^4\) This probability should 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 2).
0.057), respectively. As predicted, no significant change in view-bias occurred in the control conditions of experiments GI or DOT with $p = 0.430$ and $p = 0.194$, respectively.

The results are unequivocal for the pooled (GL+DOT) data, showing significant amounts of view-bias within conditions 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.

**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.

The magnitude of the effects of rotation-reversal are 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 (Bulthoff and Edelman, 1992), the largest decrease in recognition (around 35%) was associated with a viewpoint change of 60 degrees using ‘paper-clip’ objects; and it is reported
that similar results were obtained using amoeboid objects like those used here. Similarly, in (Hayward and Tarr, 1997), recognition performance of synthetic objects with well defined visual features decreased by around 6% following a change in viewpoint of 30 degrees. These figures suggest that rotation-reversal produces changes in recognition performance on a par with more radical interventions, such as a change in viewpoint.

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. (Bulthoff and Edelman, 1992)) and demonstrate the existence of view-bias (e.g. (Cutzu and 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 observed. 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 (Bulthoff and Edelman, 1992)); it is intended to suggest that the interpretation of view-bias requires some care if subjects are trained using rotating objects.

From a purely logical viewpoint, what explanations other than motion-specificity might account for the effects of rotation-reversal?
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 the starting image of each movie was chosen at random on each occasion it was presented.

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.

‘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 decide this issue.

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 and 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.

*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)\(^5\). Recognition from optic flow requires that an object can be recognised from the particular optic flow that its motion implies\(^6\). 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 from both optic flow and from temporal sequences.

**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

\(^5\) Thanks to N Hunkin for pointing this out.

\(^6\) 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.
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 as a cue in its own right. Evidence presented here suggests that this is precisely what we do.

**Acknowledgements:** Thanks to N Hunkin, J Frisby, D Buckley, J Porrill, G Mather, J Mayhew, and R Lister for useful discussions, and to two anonymous referees for their detailed comments. This research was supported by a Mathematical Biology Wellcome Fellowship (Grant Number 044823).
References


the combination of local motion signals in human vision. *AI Memo, MIT*, (1624).
Appendix 1: Measuring View-Bias

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

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

The length $|r_k|$ of this resultant vector 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 re-write $r_k$ as $r_k = (C, S)$ then its length is given by $|r_k| = \sqrt{C^2 + S^2}$.

The measure $|r_k|$ tends to over-estimate view-bias for small numbers of valid responses. The value of $|r_k|$ for the $k$th movie is based on $n_k \leq 5$ 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 $b_k = n_k|r_k|^2$ (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 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 be 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 number for a given presen-
tation of a movie is \( i = 15 \) then the movie can be played from image 15 to image \((15+90 \mod 90)\). The conversion from image number to image-angle implies that images displayed in close temporal proximity (e.g. images 85 and 1) are separated by small image-angles.

**Appendix 2: 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 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, \ldots, 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 \( b = \infty \):

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

**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.
<table>
<thead>
<tr>
<th>Block 1</th>
<th>Learning Trials</th>
<th>Test Trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial Number</td>
<td>1 2 3 4 5 6 7 8 9 10</td>
<td>11 12 13 14 15</td>
</tr>
<tr>
<td>Exp Target 1</td>
<td>o o o o o o o o o o</td>
<td>● ● ● ● ● ● ●</td>
</tr>
<tr>
<td>Exp Target 2</td>
<td>● ● ● ● ● ● ● ● ● ●</td>
<td>o o o o o o</td>
</tr>
<tr>
<td>Cntrl Target 3</td>
<td>o o o o o o o o o o</td>
<td>o o o o o o</td>
</tr>
<tr>
<td>Cntrl Target 4</td>
<td>● ● ● ● ● ● ● ● ● ●</td>
<td>● ● ● ● ● ●</td>
</tr>
</tbody>
</table>

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 E. The temporal order of images in each of these target sequences is constant during the 10 learning trials (the $E_{pre}$ condition), and reversed in the 5 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 10 learning trials (condition $C_{pre}$) and the 5 learning trials (condition $C_{post}$). The symbol (●, o) indicates whether an object’s 90-image sequence is presented in ascending order (●) (e.g. images 1 → 90), or descending order (o) (e.g. images 90 → 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.
<table>
<thead>
<tr>
<th>Condition</th>
<th>Experimental</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial 10</td>
<td>1.77s</td>
<td>1.80s</td>
</tr>
<tr>
<td>Trial 11</td>
<td>1.95s</td>
<td>1.81s</td>
</tr>
<tr>
<td>t</td>
<td>2.52</td>
<td>0.145</td>
</tr>
<tr>
<td>df</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td>p</td>
<td>&lt; 0.01</td>
<td>0.443</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Condition</th>
<th>Experimental</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial 10</td>
<td>0.89</td>
<td>0.86</td>
</tr>
<tr>
<td>Trial 11</td>
<td>0.77</td>
<td>0.89</td>
</tr>
<tr>
<td>t</td>
<td>4.42</td>
<td>-1.00</td>
</tr>
<tr>
<td>df</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td>p</td>
<td>&lt; 0.001</td>
<td>0.326</td>
</tr>
</tbody>
</table>

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.
<table>
<thead>
<tr>
<th>Condition</th>
<th>Experimental</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial 10</td>
<td>3.58s</td>
<td>3.82s</td>
</tr>
<tr>
<td>Trial 11</td>
<td>4.46s</td>
<td>3.50s</td>
</tr>
<tr>
<td>t</td>
<td>-4.04</td>
<td>1.78</td>
</tr>
<tr>
<td>df</td>
<td>23</td>
<td>23</td>
</tr>
<tr>
<td>p</td>
<td>&lt; 0.001</td>
<td>0.089</td>
</tr>
</tbody>
</table>

Table 4

DOT Experiment (RT): Two-tailed paired t-tests for the mean reaction times of subjects between trials 10 and 11.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Experimental</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial 10</td>
<td>0.93</td>
<td>0.917</td>
</tr>
<tr>
<td>Trial 11</td>
<td>0.73</td>
<td>0.89</td>
</tr>
<tr>
<td>t</td>
<td>-4.87</td>
<td>1.282</td>
</tr>
<tr>
<td>df</td>
<td>23</td>
<td>23</td>
</tr>
<tr>
<td>p</td>
<td>&lt; 0.0001</td>
<td>0.213</td>
</tr>
</tbody>
</table>

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.
<table>
<thead>
<tr>
<th>Experiment</th>
<th>GL</th>
<th>DOT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condition</td>
<td>Control</td>
<td>Experimental</td>
</tr>
<tr>
<td>$\bar{\pi}<em>{pre} - \bar{\pi}</em>{post}$</td>
<td>-0.132</td>
<td>0.132</td>
</tr>
<tr>
<td>t</td>
<td>-1.860</td>
<td>1.336</td>
</tr>
<tr>
<td>df</td>
<td>23</td>
<td>23</td>
</tr>
<tr>
<td>$p(\bar{\pi}_{pre-post})$</td>
<td>0.076</td>
<td>0.195</td>
</tr>
</tbody>
</table>

Table 6
Results of paired t-tests (2-tailed) for the mean difference $\bar{\pi}_{pre-post}$ in proportion of valid responses $\bar{\pi}$ per subject before and after rotation-reversal, for the control and experimental conditions of the GL and DOT experiments. Note that the significant results for the control condition in the DOT experiment was for an *increase* in $\bar{\pi}$, whereas the corresponding significant results for the experimental condition was for a *decrease* in $\bar{\pi}$.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Control</th>
<th>Experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condition</td>
<td>Pre-reversal</td>
<td>Post-reversal</td>
</tr>
<tr>
<td>View-bias ($b$)</td>
<td>1.121</td>
<td>1.147</td>
</tr>
<tr>
<td>$p(b &gt; 1)$</td>
<td>0.008</td>
<td>0.003</td>
</tr>
</tbody>
</table>

Table 7
Pooled GL and DOT view-bias data: 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. Each value of $p(b)$ is based on 10,000 randomisation samples (see Appendix 2).
<table>
<thead>
<tr>
<th>Condition</th>
<th>Control</th>
<th>Experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condition</td>
<td>Pre-reversal</td>
<td>Post-reversal</td>
</tr>
<tr>
<td>View-bias ($b$)</td>
<td>1.136</td>
<td>1.204</td>
</tr>
<tr>
<td>$p(b &gt; 1)$</td>
<td>0.031</td>
<td>0.003</td>
</tr>
</tbody>
</table>

Table 8
GL Experiment: View-bias $b$ before and after rotation-reversal. See caption of Table 7 for details.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Control</th>
<th>Experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condition</td>
<td>Pre-reversal</td>
<td>Post-reversal</td>
</tr>
<tr>
<td>View-bias ($b$)</td>
<td>1.105</td>
<td>1.089</td>
</tr>
<tr>
<td>$p(b &gt; 1)$</td>
<td>0.060</td>
<td>0.103</td>
</tr>
</tbody>
</table>

Table 9
DOT Experiment: View-bias $b$ before and after rotation-reversal. See caption of Table 7 for details.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>GL</th>
<th>DOT</th>
<th>GL+DOT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condition</td>
<td>Control</td>
<td>Experimental</td>
<td>Control</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>-0.079</td>
<td>0.134</td>
<td>0.018</td>
</tr>
<tr>
<td>$p$</td>
<td>0.430</td>
<td>0.061</td>
<td>0.194</td>
</tr>
</tbody>
</table>

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. Results were obtained using a (1-tailed) Wilcoxon matched-pairs signed-ranks test, with each rank being associated with one movie.
Fig. 1. Example of a grey-level object. Fig. 2. Example of a textured object.
Fig. 3. DOT Experiment: Mean reaction times during learning (trials 1-10) and test (trials 11-15), for control (dashed line) and experimental (solid line) conditions. Bars indicate standard errors (see Methods section).
Fig. 4. DOT Experiment: Mean recognition rates during learning (trials 1-10) and test (trials 11-15), during control (dashed line) and experimental (solid line) conditions. Bars indicate standard errors (see Methods section).
Fig. 5. Diagrammatic representation of small (A) and large (C) view-bias for five consecutive simulated responses to the kth movie, and the corresponding resultant vector \( r \) associated with each set of five responses (B and D, respectively). (A+C) The five image numbers (1-90) of object views which evoke correct responses are multiplied by a factor of four to produce a set of five image-angles. This circular representation of image 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 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 resultant vector shown in D has a length \(|r_k|\) approaching unity, denoting a large view-bias. The effect of different numbers \( n_k \) of correct responses on \(|r_k|\) is eliminated by defining view-bias as \( b_k = n_k |r|^2 \). 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.