

Perception, representation and recognition: A holistic view of recognition

CHRIS CHRISTOU¹ and HEINRICH H. BÜLTHOFF²

¹ *Unilever Research, Wirral, UK*

² *Max-Planck Institute for Biological Cybernetics, Tübingen, Germany*

Received 9 June 1999; revised 21 March 2000; accepted 29 March 2000

Abstract—It is clear that humans have mental representations of their spatial environments and that these representations are useful, if not essential, in a wide variety of cognitive tasks such as identification of landmarks and objects, guiding actions and navigation and in directing spatial awareness and attention. Determining the properties of mental representation has long been a contentious issue (see Pinker, 1984). One method of probing the nature of human representation is by studying the extent to which representation can surpass or go beyond the visual (or sensory) experience from which it derives. From a strictly empiricist standpoint what is not sensed cannot be represented; except as a combination of things that have been experienced. But perceptual experience is always limited by our view of the world and the properties of our visual system. It is therefore not surprising when human representation is found to be highly dependent on the initial viewpoint of the observer and on any shortcomings thereof. However, representation is not a static entity; it evolves with experience. The debate as to whether human representation of objects is view-dependent or view-invariant that has dominated research journals recently may simply be a discussion concerning how much information is available in the retinal image during experimental tests and whether this information is sufficient for the task at hand. Here we review an approach to the study of the development of human spatial representation under realistic problem solving scenarios. This is facilitated by the use of realistic virtual environments, exploratory learning and redundancy in visual detail.

1. INTRODUCTION

A mental representation is anything that allows us to think about, visualise and make judgements concerning physical objects or scenes in their absence. There seems little point in referring to mental representation without reference to the function that the representation subserves such as recognition. The forms of recognition that have been studied in experimental psychology include identification, categorisation and discrimination (see Lliter and Bühlhoff, 1998). The basis for all these abilities is the encoding of spatial information derived through perceptual experience (Wallis

and Bülthoff, 1999). Spatial information derived during early vision is encoded and forms spatial memories of a region of space. Recognition is only successful if a mental representation can be matched with the current contents of perceptual experience. One requirement of this process is that it should be flexible. This is desirable because the world around us is always changing and we should never expect to encounter exactly the same viewing condition on different occasions. A flexible mental representation of objects or scenes must therefore be insensitive to variations in illumination, viewing conditions and other factors that may be liable to change on separate occasions. Such flexibility may theoretically be achieved either by a transformation of the representation to match the contents of the perceptual encoding or by a transformation of perceptual encoding to match the representation. Thus, recognising a depth rotated object may be achieved either by transformation of an object in memory or by a transformation on the current contents of perceptual experience. Deciding which of these transformations is more prominent in human recognition (both are theoretically possible) is not easy and both entail problems. However, addressing such issues of flexibility may provide us with a better understanding of how mental representation is achieved in humans. To do this it is informative to consider briefly the philosophical basis of the issue of mental representation.

Empiricist philosophers considered all knowledge to be ultimately derived from the basic sensory qualities of experience. John Locke for instance proposed the existence of both simple ideas (yellow, hot, sweet) and complex ideas (conjunctions of simple ideas). All knowledge therefore is based on experience and complex entities, or objects, are essentially combinations of qualities derived through sensory experience. The intention of John Locke and other empiricist philosophers who followed him was to determine the origin of ideas and therefore mental representation, the basis for knowledge. The influence on this thinking is apparent in modern theories of vision. For instance, the conception of object representation as the result of hierarchical extraction of detail is embodied in David Marr's theory of object encoding. In Marr's view the final (3D) encoding or representation of an object is a product of a so-called 2.5D sketch that only specifies the depth distances and orientation changes directly visible to the observer. Furthermore the 2.5D sketch is formed from a combination of even more basic detail derived from visual cues such as stereopsis, motion parallax, the kinetic depth effect, texture, shading, etc. (see Marr and Nishihara, 1978). Different types of depth and shape information are extracted from each of these cues separately and combined to form the 2.5D sketch. Eventually, the 2.5D sketch is used to form a complete (3D) representation of an object.

In the last few years there have been numerous experiments carried out which delve into the information extraction modules proposed by Marr. The emphasis of this research has been either to produce a working model of each individual module (how it operates) or discovering how these modules interact or combine information to form a unified percept. At the same time, cognitive psychologists and neu-

roscientists have concentrated on higher-level processes that use this information, for instance, in object recognition. There are key findings from both the low-level depth perception studies and also from the higher-level recognition studies that we believe shed light on the entire object recognition process.

For instance, with respect to object recognition there has been much controversy over whether mental representation of objects is view-dependent. That is, does recognition performance strictly depend on the, perhaps limited, views of objects experienced by an observer or is a single view of an object sufficient for generalised recognition of it. Biederman (1987) argued for the latter and proposed a theory of object recognition based on the extraction of view-invariant geometrical features, or geons, which when visible allow both familiar and newly learned objects to be recognised from any view of them. Biederman provides evidence that recognition ability is indeed invariant to scale, position and in some circumstances rotation in depth (e.g. Biederman and Gerhardstein, 1993; but see Tarr and Bülthoff, 1995). Contrary to this, is the finding of several other studies that recognition performance on familiar object classification and novel object recognition varies for different views of objects revealing canonical views that are more easily recognised than other accidental views (e.g. Palmer *et al.*, 1981). Also, it has been reported that for novel (computer generated) objects the time required for successful recognition is a function of how far the object is rotated away from a trained view of it (see Tarr, 1995, for a review).

In terms of low-level feature extraction and depth perception a number of studies have reported systematic distortions of shape particularly along the depth dimension (along the line of sight). For example, Johnston (1991) reported systematic distortions in the perceived shape along the depth axis of stereoscopically portrayed cylindrical surfaces. Even when surface texture was added this distortion persisted (Johnston *et al.*, 1993). The integration of stereo and shading was investigated by Bülthoff and Mallot (1988) who found that reconstructions of shape from subject's settings of a 'depth-probe' yielded an overall flattening or under-estimation of depth. Using a similar probe for surface orientation, Koenderink *et al.* (1992) also made reconstructions of the content of photographs. Although they did not specifically address the issue of veridicality their reconstructions also reveal an under-estimation of curvature. All of these experiments involved simple surfaces. However, in an experiment involving multi-object computer generated scenes, Christou *et al.* (1996) showed that slant settings across these scenes were most noticeably underestimated when surface texture alone defined the constituent surfaces and when no explicit outline contours were visible. Even when realistic shading and surface texture is visible, surface orientation, and in particular surface slant with respect to the observer, is underestimated.

The overall impression from such results is that the depth dimension is squashed or foreshortened and that this is particularly apparent under reductionistic conditions under which both retinal information and observer movement is restricted. The net effect of this on object recognition may be that since the greatest degree of error

in the initial stages or shape perception appears to occur in the reconstruction of the depth dimension then this dimension will produce the most unreliable encoding and therefore the least weighting. Thus, under restricted viewing conditions (i.e. given a single image of a picture) the image plane detail is predicted to be the best encoded and the depth dimension the worst. The addition of depth cues should increase reliability although as mentioned above it appears that even stereo cues are not guaranteed to yield veridical shape and depth perception. This is corroborated by the experiments of Edelman and Bühlhoff (1992) with computer generated wire-like shapes known as paperclips. They found that recognition performance with depth rotated versions of these paperclips varied with the (angular) distance from the original training views. More significantly they showed that, while there was an improvement in performance when these objects were presented in stereo, the dependence of error rate on misorientation relative to the training view was the same in the Mono and Stereo condition. However, this does not amount to a proof of a correlation between depth underestimation and poor recognition and clearly, this is an area for further research. In summary, given a single viewpoint of the observer, there is good reason why 2D detail should form the basis of visual encoding.

The close connection between the properties of perception and the nature of representation considered here can be tested empirically by correlating properties of low-level encoding with recognition ability. These considerations in general lead to constructive guidelines for conducting recognition experiments and in the remainder of this paper we demonstrate these using some of our previous experiments as examples. The guidelines centre on a holistic approach to study representation via performance within realistic recognition tasks. This amounts to a shift of emphasis away from cue reduction and towards the use of measurement schemes that are not sensitive to the presence of diverse (possibly interacting) experimental variables. Such an approach is encompassed in perturbation analysis (Young *et al.*, 1993; Landy *et al.*, 1994; Koenderink *et al.*, 1996) in which correlation is used to assess the effectiveness of the independent variable in the presence of other cues.

Our chosen method of studying recognition within realistic contexts is facilitated by the use of computer simulations for the portrayal of virtual space. Virtual space generated by computer can be likened to the virtual space depicted by a photograph or painting but with the possibility of interaction both with where the camera is and with objects in the scene. Thus we create a scenario in which observers feel they are performing some realistic task and from which we can study cognition under, perhaps, more realistic (everyday) conditions. We have used virtual space interaction to study both scene recognition and object recognition. In the rest of this paper we give a flavour of these experiments.

2. SCENE RECOGNITION

Being 'lost' is usually the result of being in an unfamiliar environment. The environment looks unfamiliar and we do not know how to get to our destination.

What if we were told how to proceed through a building to get to our destination? If the building has a complex structure then it is also possible we may get lost when we try to find the entrance again. This indicates to us that, initially at least, scene recognition depends on the views we have experienced and generalising from these views is hard to do. Can we test this under realistic yet controlled conditions?

In a series of experiments (Christou and Bühlhoff, 1999) we allowed participants to manoeuvre around a simulated building (a virtual attic) in a constrained manner while we monitored what they saw. This familiarisation was devised as a 'random search' in which participants had to find and acknowledge small encoded markers that only appeared when viewed close enough. They were therefore not being 'primed' for a recognition experiment although they were told that they would be given a test about the location of the markers later on. After finding all the markers each participant was shown pictures of the locations of each of the markers together with depth rotated views of these marker locations (see Fig. 1). We could guarantee that these 'novel views' were completely unfamiliar because during familiarisation participants' views of the environment were restricted. An equal number of distractor images taken from a similar 3D 'distractor' environment were also shown to participants. The subjects had to simply respond when they believed the current image was taken from the original environment, which they had traversed during the familiarisation stage.

Figure 2 shows the results expressed in terms of the hit-rate averaged for all 18 participants. The figure shows that familiar views were most easily recognised, the novel direction views produced much more difficulty although performance was greater than chance. Also shown in the figure is the hit-rate for mirror images of the familiar views that we also presented to subjects in the same block. On very

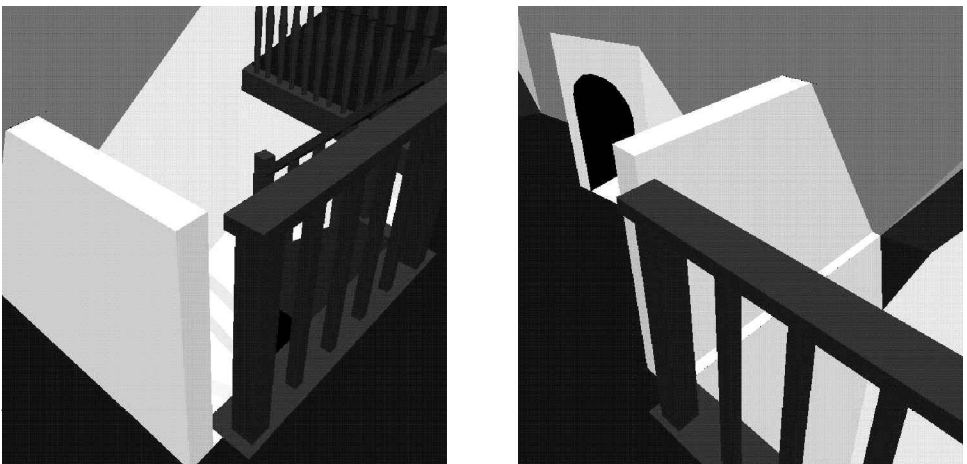


Figure 1. Two images of the virtual environment used for scene recognition. The image on the left shows an example of a 'familiar view'. The image on the right shows the corresponding 'novel view', which was never explicitly experienced by subjects.

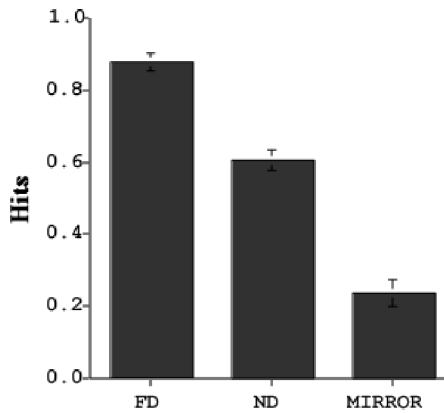


Figure 2. Correct responses (Hits) as a function of view type. FD is the familiar direction views. ND the novel direction views and MIRROR were mirror images of FD.

few occasions did participants identify these mirror views as belonging to the test environment even though in terms of content these images and the familiar views were identical. In general, this experiment shows that after extensive, controlled, and yet realistic learning the restrictions imposed on the content of perceptual experience are still reflected in recognition performance.

In a later experiment a second group of observers were given a similar task although this time the interactive familiarisation stage was replaced with the passive observation of a series of pictures of the environment. These pictures were in fact the views shown to the best performing participant of the previous experiment. The others were generated by us so that the entire environment was viewed. The familiarisation process was devised such that it lasted the same amount of time as in the previous experiment. After testing, this group of subjects suffered a small drop in the familiar view performance but the ability to recognise the novel direction views of familiar (acknowledged) locations was very badly affected. This suggests that the mode of learning is highly important. This diminution in performance however appeared not to be the result of a lack of depth cues during familiarisation. When we repeated this experiment but with stereoscopically presented disparate images of the 3D environment, performance improved a little but did not reach that of the original experiment.

3. RECOGNITION OF NOVEL OBJECTS

In terms of object recognition we wanted to investigate previous findings of view-dependent recognition with novel shapes or objects similar to those reported by Rock and DiVita (1987) and Bühlhoff and Edelman (1992). We conjectured that one of the primary problems for subjects might have been the somewhat artificial conditions of the experiment. For instance Rock and DiVita (1987) presented their (real) shapes on a dark formless background with reduced viewing. The shapes



Figure 3. Shows a ray-traced image of the experimental ‘setting’. The test objects (wire-like figures) were presented to observers on the pedestal in the centre of the room. Observers could adjust their view of the objects in real-time but were tethered to a point on the top of the pedestal.

appeared to be floating in space. We devised an experiment involving such objects but included a realistic familiarisation strategy together with a highly realistic environmental context (see Fig. 3) (Christou *et al.*, 1999). The environmental context may provide additional depth cues by which the metric aspects of the objects’ shape could be scaled. A benefit for depth cues provided by the visual background in object recognition was reported by Humphrey and Jolicoeur (1993) although Biederman (1981) found no such effect. Another benefit of the visual background is to provide an implicit specification of the observers viewing position relative to the environment and therefore relative to the object. Provided that we are familiar with a given 3D environment, the implicit specification of viewer vantagepoint that the environment provides may facilitate object recognition.

A further novelty of this experiment was the element of natural learning, which was encompassed by the task. Participants had learned to discriminate between four ‘paperclips’. During familiarisation, they could use a 6 degree-of-freedom motion input device to rotate themselves (to a limited extent) around the objects. The benefits of this are, firstly, that structural ambiguities caused by accidental occlusion of parts can be eliminated. Also, such movement introduces optic flow from which can be derived the relative depth of parts of a surface. This may help to reduce any underestimation of depth.

We ensured that participants could differentiate between the four objects from a randomly chosen ‘familiar’ viewing direction by training them to reach a criterion

level of identification performance after which they were tested on all directions around the object. Much to our surprise we found a strong dependence of error rates on the magnitude of the angular displacement from the familiar view. This was similar to previous experimental findings with such objects. Within this naturalistic scenario it still appears that participants find it difficult to form a fully three-dimensional characterisation of the objects. This is illustrated by the distinctive inverted U-shaped response profile of the error-rate shown in Fig. 4. Worst performance was for right angle or 90 deg views and general improvement for 180 deg. The significance of this is that for 90 deg views the revelation of new detail is maximal. It seems that even though observers could move around the objects by up to 30 deg in the training phase they still characterised the objects according to 2D (or 2.5D perhaps) detail. When viewed from 90 deg away from the familiar view this detail almost disappeared and they found it very difficult to identify the objects. For 180 deg of course many of the features, which were visible from the 0 deg view become visible again and we conjecture that this is the cause of the improved performance for 180 deg views compared to 90 deg views.

Another fascinating revelation from this experiment was that the room made a significant contribution to the identification of objects. This is shown in Fig. 4 by the separation in the two curves. One curve depicts the response errors for identification tests devoid of the original visual background on which they were initially trained and the other with appropriate background. Having the background in view during the identification stage reduced the error-rate significantly although it did not eliminate view-dependency. We reasoned that there were two possibilities for this advantage as stated above; namely that either the scene provides additional depth cues or that the scene was used to specify observer vantagepoint. We tested these possibilities by repeating the experiments and manipulating the orientation of

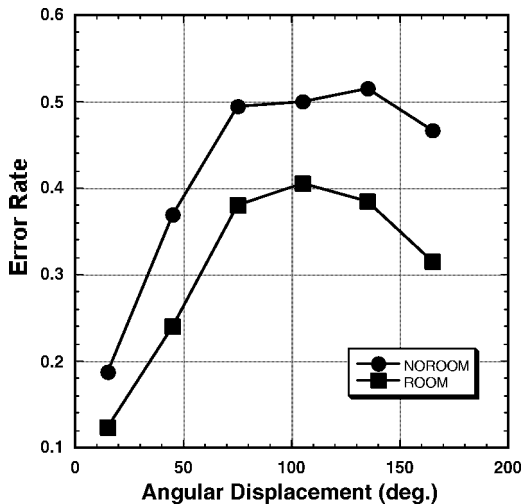


Figure 4. Shows the error rate as a measure of performance for shape identification.

the room with respect to the observer during testing. That is, when the observer was rotated to the test position, the room was also rotated (by a random amount). This meant that observers could not depend on the visual background to tell them from which position they were viewing the objects. We found that this manipulation increased the error rate compared to when the room stayed in a fixed relationship with respect to the test shapes. This is one indication that observers may be able to use the room to provide spatial reference frame information and that this information can be used in shape or object identification.

4. CONCLUSION

We have discussed the form of spatial representation used by humans to facilitate the recognition of objects and scenes. Our starting point was the consideration that the content of representation is determined by the qualities of experience and therefore of perception. If perception is constrained to lie within certain invariant bounds then these bounds may be reflected by the properties of representation and consequently by recognition performance.

This effect however depends on the level of processing that occurs during perception. At one extreme, if no processing occurs then representation is tied to the sensory properties of the retinal image or optical array that brought about the representation. At the other extreme, perception may be an analytical process where, for instance, symbolic descriptions of what we see are actively produced. The ability to generalise or to overcome variations in appearance during identification and recognition will depend on the extent of processing that occurs during perception and the kinds of information extracted. However, this is further complicated by the possibility that during recognition the stored representation (if it is impoverished) is processed further and adjusted to form a match with the current stimulus. This 'processing' of the content of perception may therefore happen during learning or during recollection and recognition; in the former case to enhance encoding, and in the latter to account for changes between encoding and the current content of experience. But, while such processing may enable some form of interpolation and extrapolation from shape encoding (see Bühlhoff and Edelman, 1992), it still remains an issue whether it can overcome foreshortening along the depth dimension. The depth dimension must be reconstructed and the chances of making an error in judging depth and in the extraction of shape from depth cues is higher than for the extraction of 2D shape. This may account for the special emphasis (greater weight) that the visual system places on 2D detail and hence explains why recognition performance is better for views of objects that have been previously viewed.

Our strategy for studying this question has been to make as few assumptions as possible about the kinds of information that is necessary for recognition under realistic contexts and in fact to provide as much visual detail as possible. We have therefore attempted to use realistic illumination, realistic contexts, movement and interactivity to approximate as closely as possible the natural process of spatial

encoding. One way to provide realistic learning under highly controlled conditions is to use computer simulation. Computer generated stimuli are already used for many low-level psychophysical experiments. It is now possible to use much more complicated stimuli but with the same level of control.

Our initial results suggest that the level of processing during perception lies somewhere between the two extremes stated above. That is, we do not believe that retinal images form the basis of representation. Neither do we believe that perception is always analytical. It appears that information can be stored in an uninterpreted fashion, and later be recalled and processed. This is demonstrated in the above-chance performance for novel views in our scene recognition experiments. If only experienced views were stored then such performance would not be possible. However, recognition still appears to be very much a function of what we experience as shown by the obvious view-dependency in our results. This is so even under the naturalistic conditions we have attempted to simulate. These results, coupled with those which utilise real objects and scenes, indicate that mental encoding need not be elaborated to the extent that the true nature of 3D objects are faithfully represented and that the limitations of perception may be reflected in the nature of spatial encoding.

REFERENCES

- Biederman, I. (1981). Do background depth gradients facilitate object identification? *Perception* **10**, 573–579.
- Biederman, I. (1987). Recognition by components: A theory of human image understanding, *Psychol. Rev.* **94**, 115–147.
- Biederman, I. and Gerhardstein, P. C. (1993). Recognizing depth-rotated objects: Evidence and conditions for three-dimensional viewpoint invariance, *J. Exper. Psychol.: Human Perception and Performance* **19**, 1162–1182.
- Bülthoff, H. H. and Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition, *Proc. Natl Acad. Sci. USA* **89**, 60–64.
- Bülthoff, H. H. and Mallot, H. A. (1988). Integration of depth modules: stereo and shading, *J. Optical Soc. Amer. A* **5**, 1749–1758.
- Christou, C. G., Koenderink, J. J. and van Doorn, A. (1996). Surface gradients, contours and the perception of surface attitude in images of complex scenes, *Perception* **25**, 701–713.
- Christou, C. and Bülthoff, H. H. (1999). View dependence in scene recognition after active learning, *Memory and Cognition* **27**, 996–1007.
- Christou, C., Tjan, B. S. and Bülthoff, H. H. (1999). Viewpoint information provided by a familiar environment facilitates object identification, Max-Planck Institute for Biological Cybernetics, Technical Report No. 68.
- Edelman, S. and Bülthoff, H. H. (1992). Orientation dependence in the recognition of familiar and novel views of three-dimensional objects, *Vision Research* **32**, 2385–2400.
- Humphrey, G. K. and Jolicoeur, P. (1993). An examination of the effects of axis foreshortening, monocular depth cues and visual field on object identification, *Quart. J. Exper. Psychol.* **46A** (1), 137–159.
- Johnston, E. B. (1991). Systematic distortions of shape from stereopsis, *Vision Research* **31**, 1351–1360.

- Johnston, E. B., Cumming, B. G. and Parker, A. J. (1993). The integration of depth modules: stereopsis and texture, *Vision Research* **33**, 813–826.
- Koenderink, J. J., van Doorn, A. J. and Kappers, A. M. L. (1992). Surface perception in pictures, *Perception and Psychophysics* **52** (5), 487–496.
- Koenderink, J. J., van Doorn, A. J., Christou, C. and Lappin, J. S. (1996). Perturbation study of shading in pictures, *Perception* **25** (9), 1009–1026.
- Liter, J. C. and Bülthoff, H. H. (1998). An introduction to object recognition, *Zeitschrift für Naturforschung* **53c**, 610–621.
- Landy, M. S., Maloney, L. T., Johnston, E. B. and Young, M. (1994). Measurement and modelling of depth cue combination: in defense of weak of weak fusion, *Vision Research* **35** (3), 389–412.
- Marr, D. and Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes, *Phil. Trans. Roy. Soc. Lond. B* **200**, 269–294.
- Palmer, S. E., Rosch, E. and Chase, P. (1981). Canonical perspective and the perception of objects, in: *Attention and Performance IX*, Long, J. and Baddeley, A. (Eds), pp. 135–151. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Pinker, S. (1984). Visual cognition: An introduction, in: *Visual Cognition*, Pinker, S. (Ed.), pp. 1–62. Elsevier, Amsterdam.
- Rock, I. and DiVita, J. (1987). A case of viewer-centred perception, *Cognitive Psychology* **19**, 280–293.
- Tarr, M. J. (1995). Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects, *Psychonomics Bulletin and Review* **2** (1), 55–82.
- Tarr, M. J. and Bülthoff, H. H. (1995). Is human object recognition better described by geon structural descriptions or by multiple views? Comment on Biederman and Gerhardstein (1993), *J. Exper. Psychol.: Human Perception and Performance* **21** (6), 1494–1505.
- Wallach, H. and O'Connell, D. N. (1953). The kinetic depth effect, *J. Exper. Psychol.* **45**, 205–217.
- Wallis, G. and Bülthoff, H. H. (1999). Learning to recognize objects, *Trends in Cognitive Sciences* **3**, 22–31.
- Wheatstone, C. (1838). On some remarkable, and hitherto unobserved, phenomena of binocular vision, *Phil. Trans. Roy. Soc. Lond.* **33**, 371–394.
- Young, M. J., Landy, M. S. and Maloney, L. T. (1993). A perturbation analysis of depth-perception from combinations of texture and motion cue, *Vision Research* **33** (18), 2685–2696.