Chapter 1
Representing 3D shape and location

Andrew Glennerster

1.1 A primal sketch that survives eye rotation

Unlike much of the book, which is concerned with 2D shape, this chapter discusses the problem of representing 3D shape. However, I will argue that there is a strong link between these. 3D shape may be better understood in terms of the 2D image changes that occur when an observer moves than 3D ego-centred or world-centred coordinates frames. The same applies to representations of 3D location. 3D shape and 3D location are properties that remain the same as an observer moves through a static world, despite rapidly changing images. Two different conceptions for visual stability emerge. One relies on generating a representation that is like the world and is stable in the face of observer movements. The other relies only on an ability to predict the sensory consequences of a movement. The implications for representation of 3D shape (and location) are quite different under these two frameworks.

Most of the literature on visual stability focuses on a situation that is relatively straightforward from a computational perspective, namely that of a camera (or the eye) rotating around its optic centre[3,4,5,6,7]. In this case, all the light rays we wish to consider arrive at a single optic centre from all possible directions (a panoramic view, what Gibson called the 'optic array' at a single point). In computer vision, the process of 'mosaicing' a set of such images is now standard[8,9]. In principle, it requires only that the rays corresponding to each pixel in each image to be registered in a common 2D coordinate frame, or sphere, of visual directions from the optic centre. Nevertheless, this is a sensible starting point for considering visual stability in general. If points in the scene are all very distant (take, as an extreme

School of Psychology and Clinical Language Sciences,
University of Reading,
Reading RG6 6AL, UK
e-mail: a.glennerster@reading.ac.uk

example, the stars at night), the optic array remains unchanged wherever you move. If these points are stable in the representation, we have a sound foundation for explaining visual stability in general.

We are now in a position to consider translation of the optic centre, either for a moving observer or the case of binocular vision. Translation of the optic centre causes a change in the optic array. Two aspects of this change can be examined separately: first, the image change generated by a small patch in the scene and, second, the changes in the relative visual direction of objects that are separated by wide visual angles. The first is relevant for the representation of 3D surface shape; the second is relevant for encoding object location.

1.2 Translation of the optic centre

1.2.1 Representing surface slant and depth relief

When viewing a small surface patch, the rays reaching the eye can be considered to be parallel (orthographic projection). This means that the ways the image of the surface deforms when the optic centre translates are relatively simple. For example, the component of eye translation along the line of sight causes expansion (or contraction) while the orthogonal component causes 1D shear or stretch. The *axis* of the shear/stretch depends on the tilt of the surface, corresponding to the intersection of the plane perpendicular to the line of sight with the plane of the surface. The *direction* of the

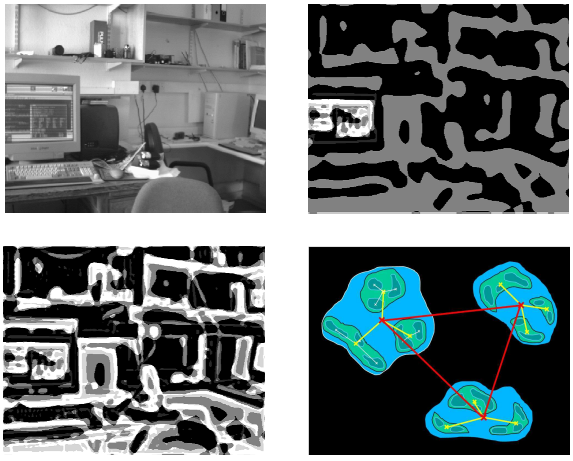

Fig. 1.1 Hierarchical encoding of position. An image (top left) is bandpass-filtered to show regions that are darker than the local mean luminance, including finer scale features in one part of the image, such as the fovea (top right image) or across the whole scene, *e.g.* after many saccades (bottom left). Because the combination of filter outputs follows the MIRAGE algorithm [1], there is a natural hierarchical encoding of position as shown schematically in the bottom right image (see also Figure 1.3).

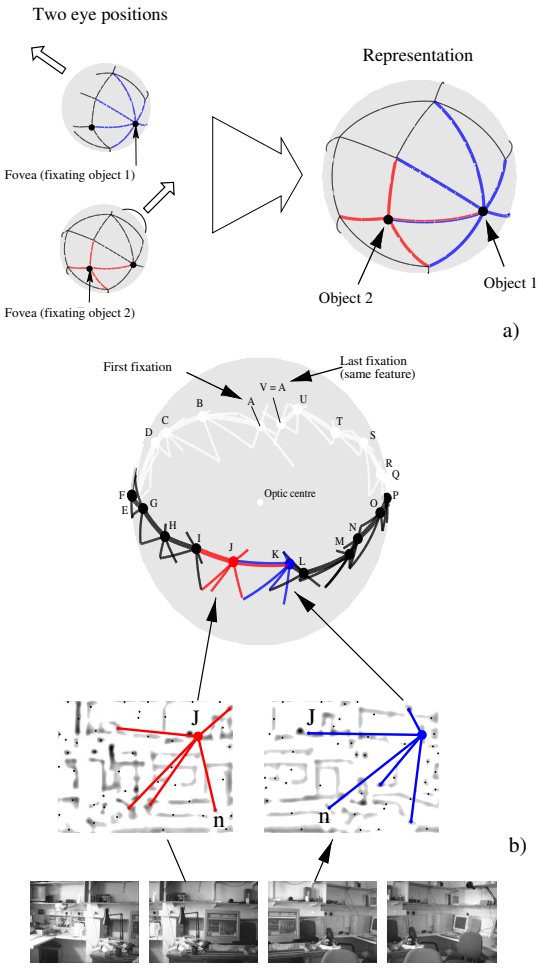

Fig. 1.2 A representation of visual direction. a) An eye that rotates about its optic centre (which is an approximation to the truth in most cases) provides information about the relative visual direction of objects. Fixating different objects provides different sets of relative visual directions (e.g. blue and red arcs) which can be combined across the entire sphere to provide a single, stable representation of relative visual directions. b) An illustration of forming this type of representation from images taken using a camera that rotates about its optic centre, including the same image and primitives as used in Figure 1.1. Features J and n appear in two of the images allowing them to be registered with the correct orientation (adapted from Glennerster *et al.*[2]).

shear/stretch depends on the direction of the observer translation. The *magnitude* of the shear/stretch is influenced by the slant of the surface away from fronto-parallel. Figure 1.3 shows one 'patch' or blob that has been stretched as a result of observer translation. It also shows how a hierarchical encoding of spatial location could help to implement a method of recording image changes. Koenderink and van Doorn[10] have proposed that surface structure could be represented using an image-based coordinate frame that would not require the generation of a 3D object-based representation. Because the three basis vectors of the frame are image based, the coordinates of all points on a rigid object remain unaffected by changes in viewpoint, rather like the coordinates of points on a deformable rubber sheet. A similar approach can be

applied to the deformation of the blob shown in Figure 1.3. The centroids of the blobs at each scale are recorded in relation to the centroid of the blob at a larger scale. If the coordinate frame for measuring these relative positions is inherited from the scale above, i.e. the distance metric is not measured in minutes of arc at the eye but relative to the width and height of the blob at the next coarsest scale, this would lead to a representation of location with similar properties to those advocated by Koenderink and van Doorn[10]. Shear, stretch or expansion of an image region caused by moving laterally or closer to a planar surface patch (as shown in Figure 1.3) would yield no change in the relative position of the finer scale features if positions are measured in this locally-defined, hierarchical coordinate frame. Similarly, any depth relief of points relative to the surface plane would give rise to a change in hierarchical position when the viewpoint changes but this would be independent of the slant of the surface and signal only the relief relative to the surface[11,12].

One difference between this hierarchical scale-based scheme and that of Koenderink and van Doorn[10] concerns the basis vectors used. In Koenderink and van Doorn's scheme, provided that the points defining the three basis vectors are not co-planar, the coordinate of every point on a rigid object is recorded using the same basis vectors. But in the hierarchical system illustrated in Figure 1.3, the coordinate frame is local and scale-based. This means that the representation amounts to something like a set of planar patches at each scale, each patch having a location, depth, tilt and slant defined relative to the 'parent' patch at the scale above. With this proviso, the scale-based hierarchy is very similar to the object-based representation Koenderink and van Doorn proposed and has the advantage of avoiding an explicit 3D coordinate frame.

A series of psychophysical studies support the hypothesis that the visual system may use a surface-based coding system of this sort. Mainly, these studies have investigated the processing of binocular disparity but there is also some evidence from structure from motion experiments[13]. Mitchison and McKee[14] showed that binocular correspondences in an ambiguous stereogram were determined not by a nearest-neighbour rule using retinal coordinates to define proximity, as had always been supposed, but by proximity to an invisible 'interpolation' surface drawn between the edges of the patch. This is equivalent to the prediction of the hierarchical 'rubber sheet' representation outlined above, in which the metric for measuring the location of dots in the left and right eyes is determined by the shear/stretch of the patch in that eye. Like correspondence, perceived depth relief is also determined by the disparity of a point relative to a local surface even when observers are remarkably insensitive to the slant of the surface[15,16,17]. Finally, sensitivity to depth perturbations are determined not by the disparity of a point relative to neighbouring points but instead by its disparity relative to an invisible interpolation plane[18,12,19], as a 'rubber sheet' model would predict.

As an aside, it is worth noting that the hierarchical encoding of blob location proposed here (following Watt and Morgan[20,1]) brings some theoretical

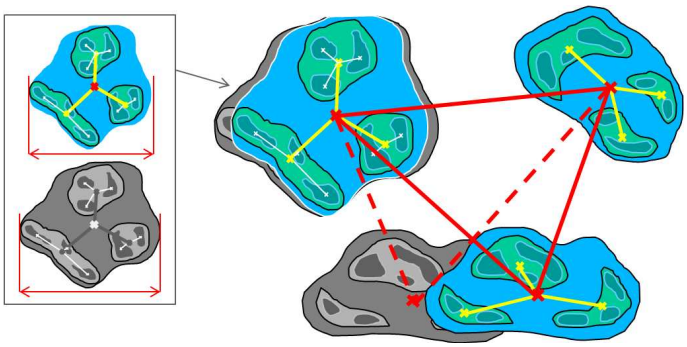

Fig. 1.3 Consequences of translating the optic centre. The 'blobs' shown in Figure 1.1 are repeated here with, in grey, the changes that would be caused by a movement of the observer or a change from the left to right eye's view. The lower blob has shifted to the left without any change in width, size or the configuration of the finer scale blobs within it. This is compatible with the surface being fronto-parallel and at a different depth from the other blobs. The centroid location of the top left blob has not changed so it is at the same depth as the top right blob. However, the width of the blob has changed, compatible with these features being on a slanted surface. The inset shows that in this case all the relative visual directions of the features (yellow and white lines) have changed together, as if drawn on a rubber sheet. These features all lie in the same slanted plane.

disadvantages but there is experimental evidence to suggest that the visual system may be prepared to pay this cost. In the coarse-to-fine stereo correspondence algorithm proposed by Marr and Poggio[21], the 'coarse scale' version of an image is always sparse, with large spacing between features (in their case, 'zero-crossings'). This means that there will always be relatively wide gaps between true and false matches along any given epipolar line and hence a nearest-neighbour rule will yield correct correspondences over a wide range of disparities. In Watt and Morgan's MIRAGE scheme, however, the 'coarse scale' representation is generated by summing the 'on' responses of filters at all spatial scales and, separately, the 'off'-responses. While this has the merit that the fine scale features *always* lie within the boundary of coarse-scale blobs, the disadvantage is that in certain situations the 'coarse scale' representation can be much more densely packed with features than the pure low frequency channel output envisaged by Marr and Poggio. Figure 1.4 shows such a situation: a dense random dot pattern with, on the right, a MIRAGE 'coarse scale' output and a schematic version to illustrate how the 'valleys' between the low frequency blobs have been 'filled in'. A random dot pattern has much greater power at high frequencies than natural images and perceptually it appears far more crowded than most images. Glennerster (1998)[22] measured the ability of the visual system to find matches when random dot patterns were shifted (either in motion or by adding disparity) and showed that MIRAGE primitives predicted well the magnitude of shift that the visual system could tolerate before the perception of motion or stereo depth

broke down. This price (a small D_{max} for high density patterns) appears to be an acceptable sacrifice for the visual system. The positive benefit is that fine scale features always have a simple, hierarchical 'address' to define their location.

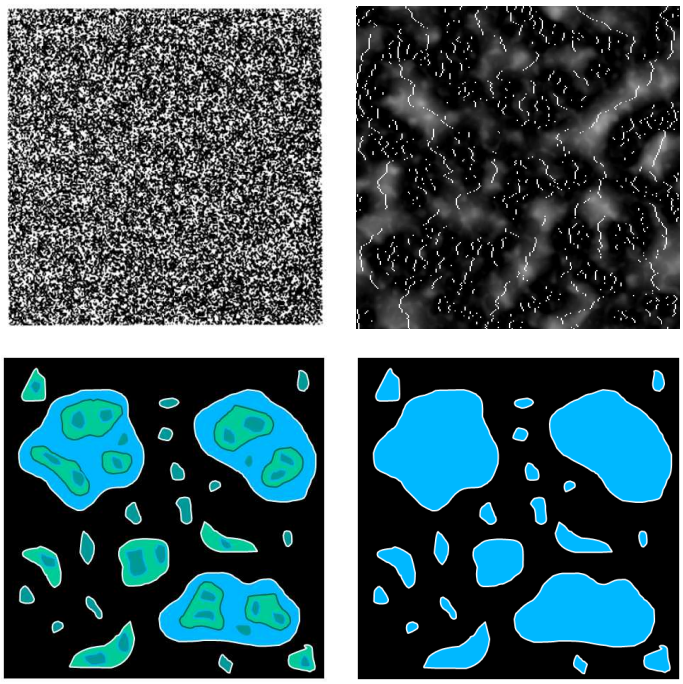

Fig. 1.4 A penalty for hierarchical encoding. If fine scale features are always to lie within the boundaries of coarse scale features, as they do in the MIRAGE algorithm[20,1] and illustrated in Figures 1.1 and 1.3, then the 'coarse scale' representation must inevitably be more crowded than a low-pass version of the image. This is particularly evident in white noise images such as the random dot pattern shown here. In a D_{max} task (see text), observers behave as if their representation of this type of image is quite crowded with features, as shown on the top right (reproduced, with permission, from[22]). The white dots mark the centroids of each blob measured along horizontal raster lines. The 'coarse scale' representation is crowded, as shown schematically in the bottom right panel, because blobs originating from different low, medium and high spatial frequency filters all contribute to the representation (see bottom left panel) and 'fill in the sea' between low spatial frequency 'islands'.

1.2.2 Representing location

Having considered the effect of observer translation on a small patch of the visual field, we now turn to the consequences for widely separated features. There are strong similarities between these two scales but also important differences. In particular, disparity and motion of a small patch provide useful information about surface shape while changes in relative position of widely separated features, such those shown in Figure 1.1a, provide information about object location.

Unlike the image changes in a small region of the visual field, the changes in relative visual direction of widely separated features do not suffer from the 'bas relief ambiguity'. This refers to the fact that a small disparity or motion can be due either to the depth relief being small or to the patch being far away. By contrast, for two widely separated features, if the angle separating them does not change when the observer moves (or there is no change between the left and right eye's view) then, in general, the points are distant: the bas relief ambiguity has disappeared (discussed in detail by Glennerster *et al.*[2]). The tendency for the relative visual direction of two features to change as the observer moves gives useful information about whether those features are part of near or distant objects. The most distant points in a scene form a set whose relative visual directions (the angles separating each pair and triple of points) are the most stable when the observer translates. Against the background of these distant objects, nearer objects 'slide around' as the observer moves[8]. One could turn this around and propose, in Gibsonian fashion, that an observer moves themselves from one place to another by 'grabbing' an object (visually, by fixating it) and 'pushing it' one way or another against the background (by walking, say) until it is in the desired place relative to the background.

The advantage of this representation is that the 3D origin of the coordinate frame is never defined. This makes sense. If you are star-gazing and see only stars, their relative visual directions do not change as you move and hence they provide no information about where you are on earth. The location of the 3D origin is impossible to define. Distant mountains allow your location to be defined more precisely, nearby trees even more so. The closer the objects in view, the more it becomes possible to pinpoint the location of the origin. Only with near objects in view would it make sense to distinguish between the origin of a coordinate system being at the eye, head, body or hand. If, however, the goal is not to build a 3D coordinate frame at all but instead to build an image-based representation, then the stars, the mountains, trees and very near objects provide a hierarchical method of locating the current image in that representation. These ideas are discussed in detail by Glennerster, Hansard and Fitzgibbon[2,23].

In summary, both 3D shape and 3D location can be considered as properties derived from the changes in relative visual direction of features produced by observer translation. The way that each of these are encoded in the visual

system should leave traces when we test psychophysical performance, as we have discussed. Two further examples are described in the final section (1.4).

1.3 Implementation of a universal primal sketch

There is no pretence that the suggestions raised in this chapter are anything like a recipe for implementation, but they do provide some useful pointers. The case of a camera rotating around its centre is an exception. In that case, a solution was described by Watt 25 years ago[1,24], with the location (visual direction) of features defined hierarchically across scale space for the entire optic array. But once the optic centre of the camera or observer translates, practical issues emerge that are considerably more tricky.

One example is the matching process that must link data structures describing the same surface seen from different view points. For example, if a surface is viewed from two distances, the spatial frequency of the filters responding to features on the surface will be higher for the farther viewing distance but if scales, like positions, are defined relative to one another, then the data structure recording fine scale features and a coarse scale outline of the object might be relatively unchanged by this alteration in viewing distance. Relative measures are likely to be a prominent aspect of the primal sketch. Of course, in the real world, with real images, complex changes occur with changes in viewpoint due to cast shadows, occlusions and specularities. The suggestions made in this chapter provide no quick fix for these problems.

It is also worth questioning the extent to which a view-based representation could underlie *all* visual tasks, not just the ones described here. One particularly problematic class of tasks involves imagining you were at a different location and making responses as if you were there. In a familiar environment, the observer may have visited that location in the past, in which case it is possible that an observer could 'run the tape' instead of actually walking to the new location and solve the task that way. But people are able to imagine being on the other side of a room that they have never seen before and to make judgements as if from that location. In our lab, we are currently exploring ways to model behaviours of this type using view-based methods, without relying on the assumption that the brain generates a Cartesian representation of the scene. In general, it is not yet clear what the limits will be to the set of tasks that could be carried out using a primal sketch or view-based framework.

1.4 Apparent paradoxes in the representation of 3D shape and location

The primal sketch outlined in this chapter is a source of 'raw' visual information that could be used for many different tasks. We discuss here two experiments that show how participants' performance appears paradoxical if we assume the visual system uses a 3D representation but readily explained if we suppose that the visual system extracts 'raw' visual information once the task is defined[25]. In one case, the task is a judgement of object shape and in the other it is a judgement of object location.

Figure 1.5 illustrates the shape task. We know that under rich-cue conditions, people show good size constancy and good depth constancy when they compare the size or depths of similar objects across different distances[26,27] but exhibit large biases when asked to make a judgement of the metric shape of a surface such as comparing the depth to the half-height of a horizontal cylinder[28,29,27]. In the case shown in Figure 1.5, the visual system must apparently estimate four values, namely the depths and half-heights of two semi-cylinders presented at two distances: d_1, h_1 and d_2, h_2. If these values were all available to the visual system, independent of the task the participant was set, then it would not be possible for participants to judge $d_1 \approx d_2$, $h_1 \approx h_2$ and yet, under the same viewing conditions, $d_1 > h_1$, $d_2 < h_2$ (*i.e.* d_1 judged as reliably larger than h_1 but d_2 judged to be reliably smaller that h_2). Yet, this is what observers see. If they built a single consistent representation of the scene and accessed the values d_1, h_1, d_2 and h_2 from this representation for all tasks, then the data would present a paradox. However, comparisons of height (h_1 versus h_2) can be done with other short-cuts, such as comparing the retinal size of test objects to other objects in the scene and the same is true of the comparisons of depths. By contrast, comparing d_1 to h_1 or d_2 to h_2 requires an estimate of absolute (not relative) viewing distance which means that these estimates are open to a source of bias that does not affect the other judgements[27]. The important point is that these data provide compelling evidence that the visual system uses information in a more 'raw' form than the metric values d_1, h_1, d_2 and h_2 when carrying out these judgements of 3D shape.

For 3D location, a good example of an apparent paradox is the case illustrated in Figure 1.6 from Svarverud *et al.*[30]. Several experiments using immersive virtual reality have shown that moving observers fail to see a room changing in size around them, by as much as a factor of four in all directions, provided that looming cues are eliminated[31,32,33]. This is compatible with earlier evidence on observers' poor sensitivity to change in disparity in the absence of looming cues[34] and raises interesting questions about the type of representation that observers must be building of the scene. Svarverud *et al.*[30] measured subject's biases when they judged the relative depth of objects either with or without an expansion of the room between the presentation of

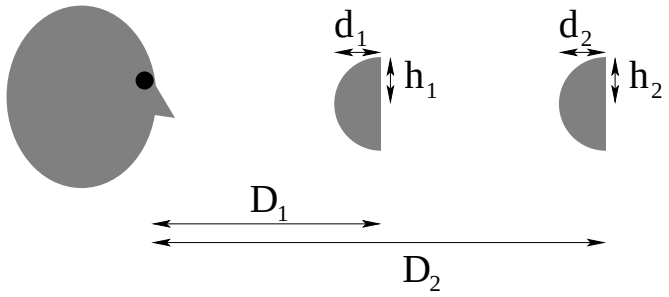

Fig. 1.5 Paradoxical representations of shape. Observers are good at size constancy ($h_1 = h_2$) and depth constancy ($d_1 = d_2$) but, under essentially identical viewing conditions, they make systematic errors when judging the shape of objects ($d_1 > h_1$ while at the same time $d_2 < h_2$). The solution to the apparent paradox is to assume that in each case, once the task is defined, the visual system acquires the relevant information and computes the solution. One task depends on an estimate of viewing distance (*e.g.* D_1) while the other requires only an estimate of the ratio of viewing distances to the two objects (D_1/D_2)[27].

the two objects. Observers did not notice any difference between these two types of trial. As Figure 1.6 illustrates, although their perception of the room was stable throughout, their pairwise depth matches cannot be explained by a single, consistent 3D representation. There is, therefore, no one-to-one mapping between a participant's internal representation of the room and a single static 3D room. It does not matter that the stimulus is an unusual one. The point is that the observer's perception is one of an ordinary, stable room so the conclusions we draw from probing the representation underlying that perception should apply to other ordinary, stable scenes.

These examples raise questions about what the minimum requirements are for a useful representation of the scene. It is no use claiming, as Gibson often appeared to[35], that an internal representation is unnecessary. More recent accounts emphasise the importance of information stored 'out in the world' rather than in the head[25], but these still require a coherent set of rules that will allow the information 'out there' to be accessed. The stored information must remain useful even if the object or visual information is not within the current field of view. This chapter outlines a possible primal sketch of for blob location that is an example of a representation of 'raw' visual information. Something like this might, with further elaboration, fulfil the criteria for a store that could be used to access information 'out there'. Such a representation must store sufficient information to allow the observer to turn their gaze to any object they remember and, if necessary, walk in the right direction until the object comes into view. It must also contain information about the slant of surfaces and the depth relief of points compared to local surfaces. These requirements fall short of the attributes of a full 3D reconstruction, but psychophysical evidence suggests the same is true of human vision.

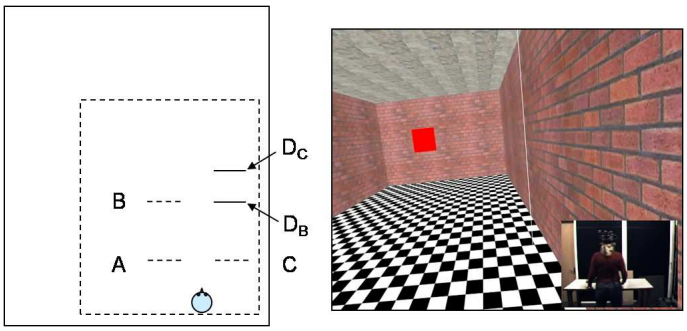

Fig. 1.6 Paradoxical representation of location. In virtual reality, observers judged the relative depth of two squares presented in separate intervals. Sometimes the room expanded between intervals (A to B and C to D), although the participants never noticed a change in room size[30]. On the other trials, the room stayed still (small room: A to C or large room B to D). It is impossible to determine a single location of D relative to A that is compatible with all the pairwise settings observers make. However, similar to Figure 1.5, there is no paradox if the visual system acquires the relevant information for any given comparison once the task is defined.

1.5 Acknowledgements

Supported by the Wellcome Trust.

References

1. Watt, R. J. Scanning from coarse to fine spatial scales in the human visual system after the onset of a stimulus. *Journal of the Optical Society of America A*, 4:2006–2021, 1987.
2. Glennerster, A., Hansard, M. E., and Fitzgibbon, A. W. Fixation could simplify, not complicate, the interpretation of retinal flow. *Vision Research*, 41:815–834, 2001.
3. Duhamel, J. R., Colby, C. L., and Goldberg, M. E. The updating of the representation of visual space in parietal cortex by intended eye-movements. *Science*, 255:90–92, 1992.
4. Zipser, D. and Andersen, R. A. A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, 331:679–684, 1988.
5. Bridgeman, B., van der Heijden, A. H. C., and Velichovsky, B. M. A theory of visual stability across saccadic eye movements. *Behavioural and Brain Sciences*, 17:247–292, 1994.
6. Melcher, D. Predictive remapping of visual features precedes saccadic eye movements. *Nature Neuroscience*, 10(7):903–907, 2007.

7. Burr, D. and Morrone, M. Spatiotopic coding and remapping in humans. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1564):504–515, 2011.
8. Irani, M. and Anandan, P. Video indexing based on mosaic representation. *Proceedings of the IEEE*, 86:905–921, 1998.
9. Brown, M. and Lowe, D. Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision*, 74(1):59–73, 2007.
10. Koenderink, J. J. and van Doorn, A. J. Affine structure from motion. *Journal of the Optical Society of America A-Optics Image Science and Vision*, 8:377–385, 1991.
11. Mitchison, G. Planarity and segmentation in stereoscopic matching. *Perception*, 17(6):753–782, 1988.
12. Glennerster, A. and McKee, S. P. Sensitivity to depth relief on slanted surfaces. *Journal of Vision*, 4:378–387, 2004.
13. Hogervorst, M. and Eagle, R. Biases in three–dimensional structure–from–motion arise from noise in the early visual system. *Proceedings of the Royal Society, London, B*, 265(1406):1587–1593, 1998.
14. Mitchison, G. J. and McKee, S. P. The resolution of ambiguous stereoscopic matches by interpolation. *Vision Research*, 27:285–294, 1987.
15. Mitchison, G. J. and McKee, S. P. Mechanisms underlying the anisotropy of stereoscopic tilt perception. *Vision Research*, 30:1781–1791, 1990.
16. Cagenello, R. and Rogers, B. J. Anisotropies in the perception of stereoscopic surfaces - the role of orientation disparity. *Vision Research*, 33:2189–2201, 1993.
17. Bradshaw, M. F. and Rogers, B. J. Sensitivity to horizontal and vertical corrugations defined by binocular disparity. *Vision Research*, 39:3049–3056, 1999.
18. Glennerster, A., McKee, S. P., and Birch, M. D. Evidence of surface-based processing of binocular disparity. *Current Biology*, 12:825–828, 2002.
19. Petrov, Y. and Glennerster, A. The role of a local reference in stereoscopic detection of depth relief. *Vision Research*, 44:367–376, 2004.
20. Watt, R. and Morgan, M. Mechanisms responsible for the assessment of visual location: theory and evidence. *Vision Research*, 23:97–109, 1983.
21. Marr, D. and Poggio, T. A computational theory of human stereo vision. *Proceedings of the Royal Society, London, B*, 204:301–328, 1979.
22. Glennerster, A. $D_{\max}$ for stereopsis and motion in random dot displays. *Vision Research*, 38:925–935, 1998.
23. Glennerster, A., Hansard, M. E., and Fitzgibbon, A. W. View-based approaches to spatial representation in human vision. *Lecture Notes in Computer Science*, 5064:193–208, 2009.
24. Watt, R. J. *Visual processing: computational, psychophysical and cognitive research*. Lawrence Erlbaum Associates, Hove, 1988.

25. O'Regan, J. K. and Noë, A. A sensori-motor account of vision and visual consciousness. *Behavioural and Brain Sciences*, 24:939–1031, 2001.
26. Burbeck, C. A. Position and spatial frequency in large scale localisation judgements. *Vision Research*, 27:417–427, 1987.
27. Glennerster, A., Rogers, B. J., and Bradshaw, M. F. Stereoscopic depth constancy depends on the subject's task. *Vision Research*, 36:3441–3456, 1996.
28. Johnston, E. B. Systematic distortions of shape from stereopsis. *Vision Research*, 31:1351–1360, 1991.
29. Tittle, J. S., Todd, J. T., Perotti, V. J., and Norman, J. F. A hierarchical analysis of alternative representations in the perception of 3-D structure from motion and stereopsis. *J. Exp. Psych. : Human Perception and Performance*, 21:663–678, 1995.
30. Svarverud, E., Gilson, S., and Glennerster, A. A demonstration of 'broken' visual space. *PLoS One*, 7, 2012.
31. Glennerster, A., Tcheang, L., Gilson, S. J., Fitzgibbon, A. W., and Parker, A. J. Humans ignore motion and stereo cues in favour of a fictional stable world. *Current Biology*, 16:428–43, 2006.
32. Rauschecker, A. M., Solomon, S. G., and Glennerster, A. Stereo and motion parallax cues in human 3d vision: Can they vanish without trace? *Journal of Vision*, 6:1471–1485, 2006.
33. Svarverud, E., Gilson, S. J., and Glennerster, A. Cue combination for 3d location judgements. *Journal of Vision*, 10:1–13, 2010.
34. Erkelens, C. J. and Collewijn, H. Motion perception during dichoptic viewing of moving random-dot stereograms. *Vision Research*, 25:583–588, 1985.
35. Gibson, J. J. *The perception of the visual world*. Boston: Houghton Mifflin, 1950.