Fixation could simplify, not complicate, the interpretation of retinal flow

Andrew Glennerster* Miles E. Hansard† Andrew W. Fitzgibbon‡

October 31, 2000

Running title: Fixation could simplify retinal flow

*University Laboratory of Physiology, Parks Road, Oxford, OX1 3PT
†Department of Computer Science, University College London, Gower St, WC1E 6BT
‡Department of Engineering Science, University of Oxford, OX1 3PJ

Abstract

The visual system must generate a reference frame to relate retinal images in spite of head and eye movements. We show how a reference frame for storing the visual direction and depth of points can be composed from the angles and changes in angles between pairs and triples of points. The representation has no unique origin in 3-D space nor a unique set of cardinal directions (basis vectors). We show how this relative representation could be built up over a series of fixations and for different directions of translation of the observer. Maintaining gaze on a point as the observer translates helps in building up this representation.

In our model, retinal flow is divided into changes in eccentricity and changes in meridional angle. The latter, called 'polar angle disparities' for binocular viewing (Weinshall, 1990), can be used to recover the relief structure of the scene in a series of stages up to full Euclidean structure. We show how the direction of heading can be recovered by a similar series of stages.

1 Introduction

Retinal flow must be used to compute both the scene structure and the observer's motion in some co-ordinate frame. It is often assumed that this co-ordinate frame must be 3-dimensional, but it does not have to be. The primary objectives of this paper are (i) to describe a reference frame for visual direction and depth that can be updated without the necessity for explicit 3-D co-ordinate transformations when the observer moves their head or eyes and (ii) to show how the information required to do this can be obtained very simply from retinal flow, provided that the observer maintains fixation as they move.

Most algorithms for interpreting retinal flow assume that it is useful to compute a single 3-D frame in which to describe the rotation and translation of the eye and the layout of points in the scene. If this is the goal, then it is certainly logical to compute the rotational and translational components of retinal flow (Longuet-Higgins and Prazdny, 1980; Regan and Beverley, 1982). A rotational component of flow is generated when an observer moves through a static scene fixating a near object: as the observer translates the eye must counter-rotate to maintain gaze on the object. There is a broad consensus that, somewhere in the visual system, retinal flow must be decomposed into its constituent parts, the rotational and translational flow fields, in order to recover (i) the direction of translation and (ii) the 3-D structure of the scene (Longuet-Higgins and Prazdny, 1980; Regan and Beverley, 1982; Warren et al., 1988; Warren and Hannon, 1990). Longuet-Higgins and Prazdny (1980) were the first to show how this could be done without prior knowledge of the eye's motion. A detailed analysis of several computer vision approaches to this problem is given in Barron et al. (1994). Several biologically motivated models have also been proposed (e.g. Koenderink and van Doorn, 1987; Heeger and Jepson, 1992; Lappe and Rauschecker, 1993, 1994; Beintema and van den Berg, 1998).

Despite the consensus that the visual system performs a decomposition of retinal flow, there is no compelling evidence that it does so (see section 6). Nor is it clear that computing translational flow is necessary or even very useful. The central problem concerns the reference frame in which information might be stored after translational flow is extracted. One suggestion is that information is used to compute the 3-D structure of the scene, first in a head-centred, then a body-centred and finally a world-centred co-ordinate frame (Andersen et al., 1997; Stone and Perrone, 1997; Colby, 1998; van den Berg, 1999; Lappe et al., 1999). This long chain of co-ordinate transformations is avoided in computer vision, where the camera motion and scene structure is computed in a world-based (albeit arbitrary) co-ordinate frame in a single step, without any intervening 'egocentric' reference frames. Not only are the putative biological processes more tortuous, there are also no clear proposed mechanisms for carrying them out. Some models carry out *2-D* transformations, converting retinal signals to a 'head-centred' frame (Zipser and Andersen, 1988) and there is some evidence that transformations of this type are carried out in parietal cortex (Duhamel et al., 1997). Finding evidence of true 3-dimensional transformations, of the type that would be required when an observer translates, is a much greater challenge that has not yet been met.

Briefly, the representation is built up from the *relative visual directions* (RVDs) of points – i.e. the angle subtended at the optic centre between pairs and triples of points (see figure 3). These provide a reference frame for visual direction (section 3), while changes in RVD provide information about the relative depth of objects. RVDs and changes in RVD with respect to the fixation point can be measured very straightforwardly (section

5). According to the hypothesis we present, the act of maintaining gaze as the observer moves is positively beneficial rather than being a complicating factor in the interpretation of retinal flow.

Several components of this representation have been described before. For example, (i) the treatment of retinal flow in terms of polar components (changes in eccentricity, ρ, and meridional angle, θ) has been described in detail for binocular vision (Weinshall, 1990; Liu et al., 1994; Gårding et al., 1995); (ii) fixation has been shown to constrain the estimation of 3-D camera motion (Aloimonos et al., 1987; Bandopadhay and Ballard, 1990; Sandini and Tistarelli, 1990; Daniilidis, 1997) and physiological models of heading estimation (Perrone and Stone, 1994); (iii) a 2-D representation of visual direction plus parallax has been described (Irani and Anandan, 1998); (iv) a reference frame for 2-D location built up from relative positions has been described by Watt (1987) and similar ideas have been suggested to account for saccade-related activity in frontal eye fields (e.g. Goldberg and Bruce, 1990).

The novel aspects of the model we propose are primarily (i) the link between fixation (the maintenance of gaze during observer translation) and the generation of the representation and (ii) the use of relative visual directions (RVDs) and changes in RVDs, which avoids any absolute co-ordinate frame. In addition, we suggest some simple rules for recovering information about the direction of translation (section 5.2) and for storing information gathered during different directions of translation (section 4).

We begin, in section 2, by summarising some of the previous approaches that have capitalized on gaze stabilisation as a way of simplifying the interpretation of retinal (or image) flow.

In section 3, we describe a reference frame for visual direction built up from the relative visual directions (RVDs) of pairs and triples of points. We show, in section 4, how the representation can be extended to include information about the parallax of points as the optic centre of the eye translates (including the case of a moving binocular observer). We describe how the representation is egocentric and yet, at the same time, has some properties of a world-centred (allocentric) frame.

Section 5, relates changes in RVD to the polar components of retinal flow and summarizes previous methods for recovering relief structure (relative depths) using these components (Weinshall, 1990; Liu et al., 1994; Gårding et al., 1995). The methods range from a very simple heuristic for determining whether a point is in front or behind the fixation point up to an algorithm for recovering full, metric 3-D structure. In section 5.2, we describe how a similar hierarchical strategy can recover direction of heading using polar components of flow. Here, hierarchical means that later stages use the solutions of earlier, more approximate stages.

Finally, in sections 6 and 7, we discuss some of the neurophysiological and psychophysical evidence that relates to our proposed representation and set out experimental predictions that could test the theory.

The interpretation of retinal flow is intimately linked with issues of storage and representation. Any successful model must explain how, and in what co-ordinate frame, the visual system combines information from retinal flow generated over several saccades and several translations. One coherent strategy is to continuously update a world-based, 3-D model. The scheme outlined in this paper offers a more biologically plausible alternative.

2 Previous approaches using gaze stabilisation

Some computer vision models have taken advantage of a fixating camera in interpreting retinal flow, but used a different approach from the one we describe. For example, Sandini and Tistarelli (1990) have used non-visual measures of the camera pose to compute the rotational and translational flow. They note that the same ego-motion parameters are useful in computing both flow components, because the two are linked for a fixating camera system. Murray et al. (1997) also use extrinsic signals about the pose of a fixating camera, in their case to control the motion of a robot in relation to a fixated object. In Daniilidis (1997), the problem of egomotion computation from visual signals alone was addressed, and the simplification of the computation was explicitly derived. In their model, a dense flow field is used to compute direction of heading and instantaneous rotation. The approach however, is rooted in a 3D interpretation, and is specifically related to instantaneous egomotion.

A neurophysiologically inspired model for determining heading direction depends on gaze stabilisation (Perrone and Stone, 1994). In the model there is a separate 'template' or neuron for every possible direction of translation (with respect to the fovea) and every fixation distance (see section 6.2). Although this may seem like a large number of possible combinations, it is very much smaller than the total number of templates that would be required if the gaze were not stabilised during translation. Section 6.2 discusses some of the differences between this model and the strategies for estimating heading that we suggest.

The model we describe has features in common with several of these approaches. The principal difference concerns the reference frame for relating information gathered during successive fixations.

3 A reference frame

This section describes a reference frame for visual direction that is built up using only the *relative* visual directions of points (i.e. the angles between pairs and triples of points). These angles do not depend on the rotation state of the eye. The next section shows how changes in RVD, which are produced by translation of the eye, are incorporated in the representation.

Figure 1 shows an idealised eye - a sphere in which the optic and rotation centres coincide - that is 'looking' in different directions. The red and blue arcs show great circles joining the images of points in the visual field. The fovea would move along these arcs during a saccade from one point to another. The arcs correspond to the planes shown in figure 3.

The sphere on the right illustrates a convenient and compact representation for storing visual information if the eye was only free to rotate about its centre and not translate in space. It is essentially a description of the visual direction of points in the optic array (Gibson, 1979) (i.e. the set of light rays arriving at a point in space, in this case the optic centre of the eye). The representation is very similar to the retinal image except that (a) the view is fully panoramic (all visual directions are represented, including behind the head) and, (b) unlike the eye, there is no single co-ordinate system to describe the visual direction of points. Instead, only the relative visual directions (RVDs) of points

are stored. These are the angles between pairs of rays (i.e. the lengths of arcs on the sphere) and between triples of rays (i.e. the angle between two arcs joining at a point on the sphere). These angles can be measured very simply on the retina using a polar co-ordinate frame (see section 5, figure 3). However, the representation as a whole is best described as 'piece-wise polar' or 'piece-wise retinotopic'.

Figure 1: about here

As a simple demonstration of the sufficiency of this method for encoding visual direction, we have placed in a common reference frame a set of images taken with a camera that was free only to rotate about its optic centre. The methods by which this was done are described in the Appendix. The resulting representation of relative visual direction is shown in figure 1b. Points V and A are in fact the same visual feature imaged in both frame 1 and frame 22. Cumulative errors account for the fact that the computed visual directions of V and A do not co-incide exactly.

One purpose of demonstrating this representation using real images is to provide an example of what 'points' might mean when applied to a natural scene. Here we have used MIRAGE centroids (Watt, 1987), which are organised in a hierarchical way so that when the scene is analysed at a finer spatial scale each fine scale 'blob' lies within the boundaries of a coarse scale 'blob'. This means that the relative position of fine scale blobs need only be related to the location of the 'parent' coarse scale blob. (We have not shown these in figure 1.) Without some hierarchical encoding of relative position of this type, it would be problematic to record the relative visual directions of features across the entire optic array and at multiple spatial scales. Surveyors use a multi-scale, hierarchical system of triangulation to map an area. There is some evidence that when viewing a novel scene, human eye movements follow a similar pattern, initially fixating the 'centre of gravity' of a target stimulus configuration and only subsequently the finer detail (e.g. Findlay and Gilchrist, 1997).

The information in a representation like that shown in Figure 1 is sufficient to program a rotation of the eye or camera from one object to another, including to one currently out of view. For each fixation point A to V, we store the relative visual direction of neighbouring points, including the previous and following fixation point. This means that it is possible to calculate, when required, the appropriate angle and axis to rotate the eye from the current fixation point to any other point in the representation. It is not necessary to record either of these in an external co-ordinate frame.

As a result, the representation remains unaffected by rotations of the eye since it records only *relative* visual directions. Instead, a 'pointer' indicating the current fixation direction (and the relative torsion of the eye) changes as the observer makes saccades. This is a common idea in models of a 'stable feature frame' (Bridgeman et al., 1994; Feldman, 1985). There is some experimental support for this type of representation (Henriques et al., 1998), as discussed in section 6.3.

4 Adding depth to the representation of visual direction

A representation of RVD like that shown in figure 1 can be extended to include information about the distance of points as well as their visual direction, and hence form the basis of an 'egocentric' representation.

Figure 2: about here

When the optic centre translates, the RVDs of points change unless the points are infinitely distant. Figure 2a shows how the RVDs in figure 1 change for a single translation of the optic centre (which could include a binocular pair of views). The dotted lines (and dotted white discs) show how some of the visual directions have changed. The white discs in figure 2a indicate the projection of two points that are close to the optic centre, the black discs correspond to points that are ten times more distant. The RVDs of the black discs hardly change, while the visual directions of the near points do change, relative to the distant points, as a result of the translation.

In this example, we have shown the RVD changes across the whole sphere (optic array) as a result of a single translation. In section 5 we will describe a simple way of measuring RVD changes between the fixation point and other points. To recover information about all the RVD changes shown in figure 2a using that method, the observer would have to fixate many different points in succession while making the same translation as, for example, when a static binocular observer fixates different objects in a scene.

The next section considers how information could be stored in the representation when the observer translates in many different directions (figure 2b).

4.1 Properties that persist over many translations

Some properties of images remain invariant when an observer translates through a static scene. Examples are the 'cross ratio' of image lengths that are characteristic of points on a line in space (e.g. Cutting, 1986; Cutting et al., 1992) and affine properties of planar surfaces (e.g. Koenderink and van Doorn, 1987). Several of these apply only to small regions of an image, for example when epipolar lines can be approximated as parallel or the surface can be approximated as a plane. By contrast, the property described in this section applies only for points separated by a large visual angle.

Figure 2b shows how the RVDs of points change when the eye translates in many different directions. In this example, the translations are all of unit magnitude. The colour code (and thickness of the lines) indicates the mean change in the angle between pairs of points subtended at the optic centre over 100 translations in different directions. The change is expressed as a proportion, $\Delta\rho/\rho$, where ρ is the initial angle between the two points. This is a measure of the extent to which the RVD of points changes with translation of the optic centre. The colour code alone is sufficient to distinguish the two near points (whose directions are shown by white discs, as in figure 2a) from the distant ones (black discs). The RVDs of distant points vary very little as the observer translates, as shown by the dark lines joining every pair of black discs (corresponding to distant

points) in figures 2a and b. This is always true for very distant objects like the stars but it also holds in other situations, such as within a room, when the translation is relatively small.

Here is an example of computing one quantity for each pair of points that is useful in distinguishing near from distant points. Figure 2c illustrates how this value, $\Delta\rho/\rho$, which might be loosely be described as the 'elasticity' between two points in the representation, is affected by two factors, (i) the distance to the fixation point (D) and (ii) the difference in distance (measured along each ray) between the fixation point and a second point, P. The initial angle between the two rays, ρ, is 45°. The translations, as in figure 2b, are of unit magnitude. The plot shows that, on average, the difference in depth between the two points has relatively little effect compared to the distance of the fixation point from the observer. This is in marked contrast to the situation that would apply for a small visual angle, e.g. 1°. Then, the values of $\Delta\rho/\rho$ would dip down close to zero when the depth difference between F and P was small. In other words, the bas relief ambiguity would apply - small values of $\Delta\rho/\rho$ could be due either to a large viewing distance or a small depth difference. For large visual angles, on the other hand, when the optic centre translates in many random directions, viewing distance has a much greater effect than depth difference on $\Delta\rho/\rho$, the 'elasticity' of (F, O, P). This means that a lack of 'elasticity' between points identifies them unambiguously as distant. Such points can anchor the reference frame, as explained in the next section.

4.2 Ego- and allo-centric frames united

This section explains how the RVD representation has some of the properties of an allocentric reference frame despite being an ego-centric representation. The link is the set of distant points.

The visual directions of points in figure 2b are all separated by large visual angles. As a result, as discussed above, the lack of 'elasticity' ($\Delta\rho/\rho$) between any pair of points identifies them both as distant. The distant points form a relatively rigid web as the observer translates in different directions (completely rigid if the points are infinitely distant, like the stars). Near points move against the background of distant points. The situation is not symmetrical: the set of near points change their RVD not only with respect to the distant points but also with respect to each other. The two white disks in figure 2b illustrate this well: despite both being at the same distance from the optic centre, the 'elasticity' between them is relatively large. Across the entire sphere, the web of RVDs relating distant points is stable for translations in different directions. No similar web of near points has the same property.

The distant points therefore anchor the representation in a world-based frame. Although the representation remains ego-centric, because it is based on *relative* visual directions the representation of distant points, which remain invariant to rotations *and* translations of the eye, can perform many of the functions usually associated with a world-based or allocentric representation.

It is important to be clear how this apparent sleight of hand is achieved. Normally, ego- and allo-centric representations are described as explicit, 3-D representations with a defined origin and three cardinal directions or basis vectors. It has been proposed that the visual system computes many such representations with origins at, for example, the eye, the cyclopean point (midway between the eyes), the trunk and the hand (e.g. Andersen

et al., 1997; Colby, 1998). By contrast, a representation of RVD does not define the 3-D location of the optic centre. For example, the most distant points (in the limit, stars) provide the least information about the location of the optic centre in space and yet these points provide the world-based 'backbone' or reference frame on which the representation is based. The same principle is used in 'planes-plus-parallax' models that have recently been developed in computer vision (Irani and Anandan, 1998).

Note that the information recorded in figure 2b could be measured over many fixations (recording the polar angle changes only with respect to each fixation point) and over many different directions of translation. So, unlike the example of figure 2a, this information could be recorded by a binocular observer who was free to rotate their head and to fixate on different points, where $\Delta\rho/\rho$ is, in this case, the binocular relative disparity (measured as an inter-ocular difference in eccentricities) for different head and eye positions. The example of using disparity is more straightforward than the monocular case because the magnitude of the translation (the inter-ocular separation) is always constant, whereas motion signals would need to be normalised by an estimate of translation magnitude to be used in the same way. However, even this limited situation poses severe difficulties for any representation that is truly 3-dimensional. The choice of origin and cardinal directions (e.g whether these are head or world-based) critically affects the type of computations that are proposed. Whichever choice is made, relating rapidly changing visual information to a single, 3-D co-ordinate frame is a difficult computational problem.

In the RVD representation we describe here, we have avoided both the problem of choosing a unique 3-D origin and of defining a unique set of cardinal directions (basis vectors). In one sense the fixated object is an origin: its direction corresponds to the origin of the current polar co-ordinate frame for defining direction and, when the relief of points is computed as described in section 5.1, it is at the origin of a 3-D frame. However, it is not a unique origin of the representation as a whole, because it is not maintained across time. This is similar to the use of image-based coordinates in computer vision. For example, Reid and Murray (1996) describe an active vision system which represents its fixation point in relation to four (or more) tracked features in each image. Given the tracked features in three video-frames, and the coordinates of the fixation point in the first two, it is possible to predict the image-coordinates of the fixation-point in the third frame. Thus, they show how some 3-D tasks can be performed without an explicit or stable 3-D frame.

The planes-plus-parallax model (Irani and Anandan, 1998) is similar in only one of these respects. Like the RVD representation we describe, the planes-plus-parallax model avoids defining a 3-D origin and instead records the 2-D parallax of points against a plane of points. When this plane is the plane at infinity, the model is very similar to using the set of distant points as a world-based reference frame in the way we have described. However, the planes-plus-parallax model describes a plane using an absolute co-ordinate frame whose origin is in a fixed visual direction. There is no equivalent absolute frame in the RVD representation. In one sense, the current fixation point defines a primary direction, but there is no 'special' direction for the representation as a whole.

4.3 A primal sketch of the optic array

In summary, the representation of relative visual directions (RVDs) that we have described is something like the 'primal sketch' that Marr and Hildreth (1980) proposed except that

it is of the entire optic array, not just the current retinal image. We have suggested two 'primitives' describing the relationship between points. One describes the relationship between pairs of points. These correspond to the arcs on the spheres in figures 1 and 2. The properties of this primitive are not just the angle separating the pair of points at the optic centre (ρ) but also information about changes in that angle ($\Delta\rho$). In different circumstances (e.g. when carrying out different tasks) the information that is computed and stored about a pair of points could be calibrated to different extents. For example, the property of 'elasticity' described above, ($\Delta\rho/\rho$ averaged over recent translations) is a simple, crude measure. On the other hand, it is possible to compute the fully calibrated metric depth separating the two points (see section 5.1.3) and to store this value as a property of the primitive relating the two points. If this were done all the time for all pairs of points in the representation, it would be formally equivalent to the construction of a 3-D model. We suggest that it is the exception rather than the rule for motion and disparity information to be calibrated to the extent of computing full 3-D structure (see section 7.3), and information is not stored in the representation unless it is computed. As a result, the 'primal sketch' remains sketchy, but has the potential to be made more detailed if required.

The second primitive we propose describes the relationship between triples of points (i.e. the angles between pairs of arcs on the spheres in figures 1 and 2). Again we suggest that information about changes in the angle is stored with greater or lesser degrees of calibration. The use that that visual system might make of changes in these angles is discussed in section 5.

4.4 A reference frame for larger translations

The representation as described so far deals only with small translations. When the observer makes large translations, the relative visual directions of points change significantly and eventually points disappear from view altogether, such as when the observer walks through a doorway. Here we consider two ways in which the reference frame could be extended to be useful for controlling larger translations.

One option is that the visual direction and distance of all objects in the representation is continually updated wherever the observer moves, even for those objects that are currently out of view. This requires the distance of objects to be computed accurately (see section 5.1) in order to update their directions as the observer translates. The representation is then equivalent to a full 3-D model, in which the origin changes as the optic centre translates and the axes change each time the observer makes a saccade. Although this is theoretically possible, it would require as much computation as other types of 3-D representation.

An alternative is that very much less is stored. Simple organisms, such as ants, are known to follow a set of 'image-based rules' when navigating rather than computing a 3-D map of their environment (e.g. Cartwright and Collett, 1983; Judd and Collett, 1998) and similar rules could guide much of the behaviour of more complex animals including humans. In order to navigate, observers must translate in relation to a fixated object, fixate a new target, translate in relation to that, and so on. The rules for each of these movements can be specified in terms of the image changes that are caused by the movement. For example, translation in relation to a fixated object can be controlled by monitoring the output of MSTd neurons of the type described by Perrone and Stone (1994), or using

the rules we describe (section 5.2). Neither of these require the computation of a 3-D frame. Equally, errors in the movement can be detected and corrected using retinotopic signals (e.g. Miall et al., 1993). Indeed, from the perspective of error correction, it is hard to see why visually-guided actions would benefit from any co-ordinate transformation.

The general idea of using the motor system to navigate across a set of sensory states rather than within a 3-D spatial reference frame has been described previously (Gibson, 1950, 1979; Cutting, 1986) and Arbib (1999) has described a 'world graph' model with a similar flavour. Some robot navigation systems use a related approach Müller et al. (2000).

Large scale navigation is made up of a sequence of translations in relation to fixated objects and saccades to new fixation targets. In this paper, we have discussed individual elements of the sequence. Linking the elements together into longer sequences raises new issues which are beyond the scope of this paper. Broadly, however, two things must be stored: (i) the rules for moving in relation to a given fixation point (covered in more detail in section 5.2) and (ii) the rules for choosing new fixation targets. At a given location, the latter amounts to a store of RVDs as described in section 3. This allows rotation to view objects including those that are currently behind the observer. It is also necessary to store the relationship between different locations. These can be specified in terms of the objects that need to be approached (or moved around) to arrive at a new location, rather than using a 3-D frame. To do so requires the storage of sets of RVDs at locations other than the current location. Because RVDs change quite slowly with observer translation, the locations at which it would be necessary to store an entirely new set of RVDs might be quite sparsely distributed in space. These critical locations would be determined both by the layout of an environment and occluding surfaces (as, for example, at a doorway) and also by the demands of the task.

In the remaining part of the paper, we consider in more detail the interpretation of retinal flow when an observer translates and maintains gaze on a point.

5 A polar description of retinal flow

This section describes how RVDs and changes in RVDs can be measured on the retina of a fixating, translating observer. It also reviews how these components of retinal flow can be used to recover relief structure in a series of stages. We show how direction of heading can be recovered in a similar hierarchical manner.

Figure 3a shows the projection of three points, F, P and Q onto the retina of an idealised eye - a sphere in which the optic and rotation centres coincide. F is the fixation point and projects, through the optic centre O, to the fovea, F'. P and Q project to peripheral retinal locations, P' and Q'. These locations can be described in polar co-ordinates (ρ_P, θ_{PQ}):

$$\begin{aligned} \rho_P &= \angle\mathbf{FP} = \angle\mathbf{F'P'} \\ \theta_{PQ} &= \angle(\mathbf{FP}, \mathbf{FQ}) = \angle(\mathbf{F'P'}, \mathbf{F'Q'}) \end{aligned}$$

where $\mathbf{F}, \mathbf{P}$ and $\mathbf{Q}$ are the vectors $(O, F), (O, P)$ and (O, Q). The points F, O, P and Q form two planes meeting along the line (O, F), where O is the optic centre. In general,

points in the world and the points to which they project on the retina define a pencil of planes meeting along the ray (O, F). This pencil of planes has no particular significance when considering general rotations and translations of the eye. However, when the observer translates and maintains gaze on F, the ray (O, F) is special (Weinshall, 1990; Liu et al., 1994; Gårding et al., 1995).

Figure 3: about here

Figure 3b illustrates the consequence of a translation of the optic centre from O_1 to O_2. To simplify the illustration, the translation O_1 to O_2 has been made in the plane (F, O_1, Q). This plane, containing the two positions of the optic centre, O_1 and O_2, and the fixation point, F, we shall call the *base plane* (Weinshall, 1990). In this illustration, there is no cyclotorsion during the translation, i.e. rotation about the line of sight, (O_1, F). As a result, the new projection of Q, Q'_2, lies in the plane (F', O_1, Q'_1). Thus, the projection of Q, and all other points in the base plane, changes only in eccentricity and not in meridional angle as the optic centre translates. In the Appendix, we describe one method for recovering the intersection of the base plane with the retina when there is cyclotorsion during translation.

The projection of P moves from P'_1 to P'_2. This change in retinal location can be described by two polar components. First, P' changes in eccentricity, i.e. the angle $\angle\mathbf{F}'\mathbf{P}'$, which is equal to the angle $\angle\mathbf{FP}$. This component is $\Delta\rho_P$.

Second, P' changes its meridional angle, i.e. the angle θ with respect to some reference plane that contains O, F' and one other retinal point. We refer to the angle between the planes (F, O, P) and (F, O, Q) as θ_{PQ} and the angle between the planes (F, O, P) and the base plane as θ_P. In figure 3b, because Q lies in the base plane, θ_{PQ} and θ_P are the same. As the optic centre translates, these angles change by $\Delta\theta_{PQ}$ and $\Delta\theta_P$. These values are not affected by cyclotorsion (rotation of the eye about (O, F)) since they do not depend on the retinal co-ordinate frame.

On the other hand, the change in meridional angle of P' on the retina, $\Delta\theta_{P'}$, does depend on whether cyclotorsion occurs during translation. Ferman et al. (1987), for example, have measured cycloversion (i.e. conjugate torsion of the eyes) during horizontal oscillations of the head and found it to have a maximum amplitude of about $\pm$ 1 degree. If cyclotorsion during translation is small, the component of retinal motion $\Delta\theta_{P'}$ at P' is a useful approximation to $\Delta\theta_P$. However, when there is significant cyclotorsion (such as when the subject rotates their head around the line of sight as they translate), the rotation of the eye around the line of sight, (O, F) must be determined first (see Appendix) and subtracted from $\Delta\theta_{P'}$ in order to compute $\Delta\theta_P$.

In summary, the two orthogonal components of motion of P' on the retina, $\Delta\rho_{P'}$ and $\Delta\theta_{P'}$, relate to the polar angles ρ_P and θ_P in the following way. $\Delta\rho_{P'}$ (change in eccentricity on the retina) is equal to $\Delta\rho_P$ (i.e. change in the angle $\angle\mathbf{F}'\mathbf{P}'$) provided that the observer maintains fixation during translation. $\Delta\theta_{P'}$ (change in meridional angle on the retina) is equal to $\Delta\theta_P$ provided that the observer maintains fixation *and* that there is no cyclotorsion (rotation about (O, F)) with respect to the base plane during the translation.

The rationale for this particular decomposition of motion at P' is that it can be related in a straightforward manner to the translation of the optic centre relative to F:

translation along the line of sight, (O, F), produces changes in ρ_P but not θ_P; translation perpendicular to the line of sight, produces changes in both ρ_P and θ_P. This is why changes in θ_P are a useful measure when the exact direction of translation is unknown: they provide information about a component of motion in a known direction, i.e. perpendicular to the line of sight.

The following sections review methods of recovering relief structure (Weinshall, 1990; Liu et al., 1994; Gårding et al., 1995) and describe a closely related method for recovering direction of heading based on the polar decomposition of retinal flow.

5.1 Relief

Figure 4: about here

Figure 4 illustrates how changes in θ_P provide information about the depth of P. The scene layout shown in the three examples is very like that shown in figure 3 except that the distance to point P varies in each example. The optic centre translates from O_1 to O_2 to O_3 in the base plane. The contour plot shows $\Delta\theta_P$ for different translations and for different distances to the point P (distance (O_2, P)). The abscissa shows the magnitude of translation in a direction (O_1, O_2, O_3). The units are multiples of the distance (O_2, F). Any component of translation along (O_2, F) produces only radial flow and so has no effect on the magnitude of $\Delta\theta_P$.

5.1.1 Points in the fixation plane

It is straightforward to identify points at the same depth as the fixation point. As has been pointed out (Weinshall, 1990; Liu et al., 1994; Gårding et al., 1995), for small translations there is no change in θ_P for points that lie in the 'gaze-normal plane' i.e. in the plane through F and perpendicular to the line of sight, (O_2, F) (Liu et al., 1994). This is shown by the dotted line on the graph in figure 4. Note that this property ($\Delta\theta_P = 0$) remains true for translations in any 3-D direction.

5.1.2 Points in front and behind fixation

Similarly, it is straightforward to identify points as in front of or behind the fixation point. For a given translation, the sign of $\Delta\theta$ changes for points in front of and behind the gaze-normal plane (figure 4). Since the component of translation along the line of sight has no effect on the sign of θ, this method of picking out points that lie in front of or behind the gaze normal plane applies to a whole range of different directions of translation.

The sign of $\Delta\theta$ also reverses when the direction of translation changes. So, in order to use the sign of $\Delta\theta_P$ to determine the relative depth of P, something must be known about the direction of translation. In fact, the direction of translation has to be known only within a 180 degree range (i.e. directions in the base plane either side of the line of sight (O_2, F)). (This is unnecessary in the binocular case because it is impossible to fixate behind the head. Knowing which is the left and which the right eye's view is sufficient.) What must be known about the translation is (i) the projection of the base plane (O_1, O_2, F) onto the retina (see Appendix) and (ii) the direction of translation

within a range of 180° (either side of the line (O_2, F) or, on the retina, the fovea, F'). On the contour plot shown in figure 4, this division corresponds to the division between positive and negative translations.

5.1.3 Relief structure and metric depth

Gårding et al. (1995) have shown in addition that polar angle disparities ($\Delta\theta_P$) can give the relief structure of a scene when they are scaled by eccentricity (ρ_P). This means that the ratio of depths of points with respect to to the gaze-normal plane is given but not the absolute depths. Further scaling by viewing distance gives full metric structure. In this final stage, the computation is equivalent to recovery of 3-D structure from translational flow (although the co-ordinate frames may differ). However, the benefit of the polar angle method is that there are several intermediate stages, each providing useful information. For many tasks it may be sufficient to stop at an earlier stage in the hierarchy in order to carry out the task. Many examples exist of evidence that the visual system adopts simple strategies when full reconstruction is unnecessary to carry out the task (Cutting, 1986; Cutting et al., 1992; Glennerster et al., 1996; Sun and Frost, 1998). The next section identifies a similar hierarchy of strategies for the recovery of direction of heading.

5.2 Direction of heading

In the previous section, the recovery of depth information using $\Delta\theta_P$ required some knowledge, albeit only in a limited form, about the translation of the optic centre. The logic can be reversed. A limited knowledge of scene structure can be used to help recover information about the direction of translation. Again, this information can be recovered hierarchically: the more specific or complex the algorithm, the greater the precision of the estimate.

The contour plot in figure 4 illustrates the symmetry between scene structure and direction of translation. If the translation of the optic centre is known to be one side of the line of sight (e.g. have a positive value on the x-axis of the plot), then the sign of $\Delta\theta_P$ is sufficient to determine whether P is in front or behind the gaze-normal plane through F. Conversely, if it is known that P is in front of the gaze-normal plane through F, then the sign of $\Delta\theta_P$ is sufficient to determine whether the translation of the optic centre is positive or negative on the x-axis of the plot (where zero is the direction (O_2, F)).

The argument need not be entirely circular. The division of points into those in front of and those behind the fixation point requires one translation about which something is known, after which subsequent, unknown translations can be monitored using changes in θ. Two binocular views can provide the 'known' translation – the inter-ocular separation – so that disparity distinguishes points in front and those behind the fixation plane. Roy and Wurtz (1990) have suggested that an operation very like this is carried out in the dorsal part of the medial superior temporal area (MSTd) of the macaque visual cortex. The neurons they identified were sensitive to both disparity and motion, and their preferred direction of motion was reversed when the stimulus was presented with crossed or uncrossed disparity (in front or behind the gaze-normal plane). As they point out, this pattern of sensitivity is appropriate for detecting translation orthogonal to the line of sight. By considering changes in θ_P, we can extend this idea to detect a *component* of translation orthogonal to the line of sight in the presence of an arbitrary and unknown

component of translation along the line of sight.

This is the simplest strategy in a possible hierarchy of algorithms for recovering information about the direction of translation. A hierarchy of heuristics for recovering direction of heading is described in the Appendix. It is shown how the following information can be derived:

1. Divide a set of translations into two groups, depending on their direction with respect to the line of sight (O, F). For example, if the base plane is horizontal, categorise the directions of heading into those to the left and to the right of the fovea. For an arbitrary point, P, this requires the sign of $\Delta\theta_P$ when the optic centre translates and knowledge of whether P is in front of or behind F.

2. Recover the magnitude of the component of translation perpendicular to the line of sight up to some unknown scale factor, which is constant across translations. This requires, in addition, the magnitude of $\Delta\theta_P$ under each translation.

3. Recover the magnitude of the component of translation along the line of sight up to some unknown scale factor, which is constant across translations. This requires, in addition, the magnitude of $\Delta\rho$ for a point lying in a plane that passes through OF and that is perpendicular to the base plane. For a horizontal translation, this means a point on the vertical meridian.

4. Recover the direction of heading. This requires that the ratio of the two unknown scale factors mentioned above be known. One way to recover this information is by observing the motion of a point that lies in the gaze normal plane over at least two different translations.

All these heuristics require the observer to maintain fixation as they translate.

Cutting (Cutting, 1986; Cutting et al., 1992) describes a strategy for recovering the direction of heading that has some similar features. It uses the 'differential motion parallax' between pairs of points, which is independent of eye rotation. However, the strategy requires a succession of saccades in order to fixate on the direction of heading.

6 Neurophysiological evidence

In this section, we review some of the neurophysiological data that has been used to argue that the visual system (i) decomposes retinal flow into rotational and translational components and (ii) generates a general-purpose, 3-D, head-centred reference frame. We discuss some of the reasons that these may not be necessary conclusions from the data gathered so far. We also provide some examples of data that is better accounted for by a RVD model than one based on explicit three dimensional representation of scene layout.

6.1 Decomposition of retinal flow into rotational and translational components

In neurophysiological studies, the search for an area that might carry out the decomposition of retinal flow into rotational and translational components has focussed in particular

on area MSTd (Saito et al., 1986; Tanaka et al., 1986; Duffy and Wurtz, 1991, 1995; Lagae et al., 1994). Neurons in this area have large receptive fields and respond to complex patterns of motion. There is some evidence that the neurons respond preferentially to patterns of retinal flow that occur during observer translation through a static environment (e.g. Duffy and Wurtz, 1991; Roy and Wurtz, 1990). A wide range of combinations of different flow components have been used to try and classify the responses of cells in this area (e.g. Duffy and Wurtz, 1995).

These experiments have not demonstrated that MSTd divides retinal flow into rotational and translational components. If cells were found to respond predominantly to either the rotational or the translational component of a stimulus despite variations in the other component, there would be good grounds for supposing the visual system treats the two independently, but this has not been shown (e.g. Krekelberg et al., 2000). The fact that such separability is not found does not rule out the possibility that rotational and translational components are extracted at a subsequent stage from the population of responses (e.g. Lappe and Rauschecker, 1993) but, equally, other models are not ruled out either.

6.2 Heading

One of the purposes of extracting translational flow is to recover the direction of heading. We show that heading can be computed from changes in RVDs of points (section 5.2). Here, we compare that approach with neuronal models for recovering heading.

Perrone and Stone (1994) describe a model in which individual detectors pool motions from different parts of the retina. Each detector is 'tuned' to the motions that would be generated by a particular direction of heading (in retinal co-ordinates) and a particular fixation distance (or rotation rate). With these parameters fixed, the possible motions at each retinal location depend only on the depth of the object that projects to that point. This means that, for a particular detector, the input motions at each retinal location form a one dimensional family, all of which contribute equally to the 'template'. Because the model assumes, like ours, that gaze and torsional eye movements are constrained during translation, the number of possible templates is limited.

The details of the model of Perrone and Stone (1994) differ from ours in a number of ways. First, we incorporate disparity. As others have pointed out, disparity provides one way to distinguish points that are nearer than fixation from those that are more distant (e.g. Roy and Wurtz, 1990; van den Berg and Brenner, 1994; Lappe, 1996). This is important because it can disambiguate similar flow patterns that arise from quite different head movements. A good example is an observer fixating a point with a plane of dots behind it and the observer moving leftwards. This produces a very similar pattern of retinal motion to that generated by a plane of dots in front of the fixation point and the observer moving rightwards. Even when there is a component to the observer's translation along the line of sight, the same arguments applies: there are two quite different directions of heading that are hard to distinguish without knowledge of the scene depths. With the addition of disparity information, however (e.g. Roy and Wurtz, 1990; van den Berg and Brenner, 1994; Lappe, 1996), this particular ambiguity disappears.

Many neurons in MSTd respond to binocular disparity, and some have been shown to do so in a way that would be helpful in disambiguating retinal flow patterns in a moving, fixating observer. Wurtz and colleagues (Roy and Wurtz, 1990; Roy et al., 1992) reported

neurons with preferences for opposite directions of motion depending on the disparity of the stimulus. This parallels the strategy we describe of dividing points into those nearer and further than the fixation point before using the $\Delta\theta$ component of their motion (section 5.2) to determine heading. Relatively few models of heading use disparity signals (although see van den Berg and Brenner (1994) and Lappe (1996)) but disparity could be incorporated quite easily into most models, including Perrone and Stone (1994). There is also psychophysical evidence that the addition of disparity information improves heading judgements (e.g. van den Berg and Brenner, 1994).

The second difference between the two approaches is that we describe a series of steps over which the heading estimate is refined whereas Perrone and Stone (1994) propose a single step. In terms of implementation, it is possible to imagine that the steps reflect different sensori-motor strategies rather than different forms of coding in MST. For example, a strategy to correct deviations from a path towards the fixated object might only require a signal giving the sign of the deviation (e.g. 'left' or 'right') or a signal proportional to the angular deviation from the path. These relatively crude signals could be gathered from a larger pool of 'template' detectors than the precise, single template corresponding to a single direction of heading, provided that the system for pooling was appropriate.

A third difference is that most models assume heading direction is converted from a retinotopic frame to a head-centred and finally a world-centred frame (Royden et al., 1994; Stone and Perrone, 1997; van den Berg, 1999; Lappe et al., 1999). In the RVD model, on the other hand, the link between retinotopic and 'world-based' reference frames does not require an intermediate head-centred reference frame (section 4.2).

6.3 A head-centred reference frame

Many different ego-centred representations have been proposed (for reviews see Andersen et al. (1997); Colby (1998)). Here we concentrate on the evidence for a head centred representation because it is often assumed to be the first to be computed from retinotopic signals (e.g. review by Lappe et al., 1999).

Computationally, the recovery of translational flow is assumed be the first step. It is important to realise that translational flow is recovered in a retinal co-ordinate frame, not a head-centred one. It has been proposed (Warren and Hannon, 1990; Lappe and Rauschecker, 1995; Bradley et al., 1996; Beintema and van den Berg, 1998; Stone and Perrone, 1997) that extra-retinal eye position signals are used to convert the information into a head-centred frame, but as yet there are no detailed suggestions about how this might be carried out physiologically.

Recordings from the ventral intra-parietal area (VIP) have been used as evidence of a head-centred representation. For example, neurons in in this area respond to both somatosensory and visual inputs in related regions of space with respect to the head (Colby and Duhamel, 1991; Duhamel et al., 1998). Duhamel et al. (1997) have described neurons in the same area that respond consistently to one region of the visual field independent of the direction of gaze of the animal (see also Galletti et al., 1993). The presumed function of cells in area VIP is to help guide head movements, especially reaching with the mouth (Colby, 1998). However, this is quite different from showing that retinal flow is mapped onto a general, head-centred representation of space. Instead, these neurons in VIP fall into a large class of neurons whose receptive fields appear to reflect the actions that are associated with that sensory input (Colby, 1998), including, for example, neurons

in pre-motor cortex with 'arm-centred' receptive fields (Caminiti et al., 1991; Graziano et al., 1994)). The finding of so-called 'action-oriented' neurons is compatible with many types of representation, including the relative representation we propose. More specific evidence would be required to support the claim that 3-D representations of the entire scene undergo rotations and translations when the head, arm or hand are moved.

There is some evidence that in the lateral intra-parietal area (LIP) the reverse transformation occurs to the one proposed in VIP, that is, from head-centred to retinotopic co-ordinates. This time the transformed signal is an auditory one. Stricanne et al. (1996) found evidence of auditory and visual input converging in a retinotopic frame in LIP, not a head-centred one. Given that auditory information about direction starts off in a head-centred co-ordinate frame, this is a striking finding. Similar mapping of auditory signals into retinotopic co-ordinates occurs in the superior colliculus and frontal eye fields (Jay and Sparks, 1984; Russo and Bruce, 1994).

This type of evidence is compatible with the idea that sensory information is united in a retinotopic frame for the purposes of guiding certain actions. Such actions include the generation of saccades, with which LIP is known to be involved (Shibutani et al., 1984; Barash et al., 1991; Thier and Andersen, 1996) orienting movements (which are often closely related, Freedman and Sparks, 1997) and reaching. Some psychophysical evidence is discussed in the next section that manual pointing is organised in a retinotopic frame (Henriques et al., 1998).

In summary, although the RVD model we propose would be falsified by the demonstration of a general-purpose head-centred representation in the brain (i.e. not one specifically associated with head or mouth movements), no compelling physiological evidence of a representation of this type has yet been presented.

7 Psychophysical evidence

In this section, we first consider some of the psychophysical data that has been presented as evidence of a head-based reference frame. We argue that the results are compatible with other conclusions. We then describe a recent experiment that was designed to distinguish retinocentric from head- or body-centred reference frames. Finally, we discuss psychophysical tests of the RVD model.

7.1 A head-centred frame

Much of the evidence cited in support of a head-centred representation demonstrates that an eye position signal can be used by subjects. For example, it is known that subjects can discriminate the visual direction of a point of light presented in the dark ($\pm$ 3°) (Merton, 1961). This is a sufficient demonstration to show that non-visual cues can provide information about visual direction. However, since the head was fixed in this experiment, it does not show that the representation is head-centred. There is no reason, for example, why a representation of relative direction, very like the visual one we have described, should not be built up from proprioceptive information.

A similar argument applies to experiments on the apparent 'straight-ahead'. For example, Morgan showed that the direction of the apparent visual 'straight-ahead' deviated systematically as a function of the eccentricity of gaze (Morgan, 1978). Prisms placed in front of the eyes also shift the visual scene relative to the apparent straight ahead

(Held and Hein, 1958). These results demonstrate the importance of information about eye position in relating the line of sight to a direction defined in head or body centred co-ordinates, but they are not evidence for a head-centred representation. For example, the RVD representation encodes only the relative visual direction of objects, but it could be related to a motor representation of relative directions. Wearing prisms would affect the registration of these two maps. Given that both maps are relative, neither need to be described as head or body centred.

Experiments on visual stability have been used to argue for the existence of a head centred representation. Helmholtz (1867) argued, from observing that manually moving the eyeball causes apparent movement of the visual scene, that efference copy must normally be used to predict the sensory consequences of the intended eye movement. The results of paralysing the eye muscles, which causes an illusory movement of the scene when the observer intends to make a saccade, lead to a similar conclusion (Perenin et al., 1977).

These experiments show that an expectation of the sensory consequences of a movement is present at the time the movement is made. Indeed, a contemporary view inverts the idea of efference copy (which is thought of as a copy of a motor command) and proposes instead that a prediction of the sensory consequences of a movement is a suitable input to the motor system when generating a motor signal (Miall et al., 1993). In terms of vision, this means a prediction of the 'desired' image. A recent experiment on saccadic adaptation supports this view. Bahcall and Kowler (1999) showed that, provided the sensory consequences of a saccade match the prior expectation, the magnitude of the saccade does not affect the perceived visual direction of the post-saccadic target. A piece-wise retinotopic store such as the RVD representation is better suited to the purpose of predicting the consequences of a saccade than an explicit 3-D representation. Note, as before, that although these experiments on visual stability suggest there is a non-retinotopic representation (i.e one that allows predictions of the sensory consequences of an eye movement), there is nothing to suggest that its co-ordinate frame is head centred.

A different argument in favour of head-centred representation has been raised in the context of heading judgements. For example, Crowell et al. (1998) have shown that heading judgements are affected by the degree of proprioceptive and 'efference copy' information that is available to determine how the fixation point moves with respect to the head or body. However, this does not necessarily imply that the proprioceptive information is being used to construct a model of the scene and the observer's translation in a head-centred frame.

In the experiment of Crowell et al, (and others, e.g. Royden et al., 1994; Stone and Perrone, 1997), the fixation point moves in quite a different way from the rest of the dots in the scene. Often the simulated motion is incompatible with an observer moving through a static scene and fixating on a static object as they move. Under such circumstances, strategies involved in measuring relative motion compared to the fixation point, as described in this paper and by Cutting et al. (1992), and the template model described by Perrone and Stone (1994), would be inappropriate for judging heading.

Instead, there are two separate reference frames in relation to which observers could make their heading judgements: one based on the fixation point, and one based on the simulated 3-D scene. In fact, observers' responses could be described as alternating between the two. When the head and body (or just the head) are passively moved to stay fixed in relation to the fixation point, or when both the observer and the fixation point

remain stationary, heading judgements are biased towards a constant heading in relation to the fixation point (towards it in this case). In conditions where the gaze actively follows the fixation point, and the body remains in a fixed orientation with respect to the simulated 3-D scene, observers are much better able to judge their direction of heading with respect to the simulated scene. One interpretation of the results is that, having set up an abnormal relative motion between the fixation point and the 3-D scene, Crowell et al have shown that proprioceptive and efference copy information are useful for recovering this relative motion and hence helping to solve the task.

A more extensive review of the evidence for a head centred representation of space is given by Henriques et al. (1998). They concluded that there is little evidence in favour at present.

7.2 A retinotopic frame for action

Fortuitous errors or biases in the visuo-motor system make it possible to make deductions about the organisation of the underlying mechanisms. Henriques et al. (1998) used an example of this to test the head centred model. They exploited the fact that observers consistently over-estimate the angular eccentricity of a remembered target when asked to point to it in the dark to distinguish head-centric and retinotopic models of visual space representation. The experiment could not have been done without the consistent bias in pointing, yet neither model would have predicted its existence.

In their experiment, Henriques et al. (1998) found that the pattern of errors in pointing to a remembered target bore a consistent relationship to the direction of gaze *at the moment the subject pointed*, whereas there was no consistent relationship to either the head-centred direction of the target or, critically, to the retinotopic location of the target when it was visible. The conclusion they reached was that pointing commands are coded in a retinotopic frame, rather like the auditory receptive fields described above.

7.3 Using the simplest strategy available

Setting a complex task, by itself, does not provide a test of the RVD representation. For example, we do not suggest that observers *cannot* compute a 3-D model of their environment and their location within it. An architect drawing a plan, side view and elevation of a building is an adequate counter-example. Rather, we propose that a representation of relative visual direction could act as a type of 'primal sketch' like the one proposed by Marr and Hildreth (1980) except that it is of the entire optic array. The purpose of Marr's primal sketch, like the one we propose, was to store relatively 'raw' visual information in a form that could be used by subsequent visual and motor processes.

This presents a difficulty in distinguishing the relative visual direction model and a true 3-D model. It has been suggested that the visual system could use a hierarchy of algorithms to carry out visual tasks, where the complexity of the algorithm depends on the demands of the task (Koenderink and van Doorn, 1991; Tittle et al., 1995; Glennerster et al., 1996). Since this hierarchy could include, at the top level, full Euclidean reconstruction, some subtlety is required in distinguishing the models experimentally. Glennerster et al. (1996) tackled a similar problem in relation to the perception of surface shape where, again, a hierarchy of algorithms could be used depending on the demands of the task. They showed that the systematic distortions in judgements of shape from stereopsis that

had been shown before (Johnston, 1991) disappeared, or were greatly diminished, when the observer's task was changed. Their result can be explained readily if it is assumed that the visual system does not compute the full 3-D structure of objects unless required to do so by the task: when a simpler algorithm could be used (in this case, in order to match the relief of objects at two distances) the visual system uses it. This result fits well with the RVD model we have presented, which stores motion and disparity information in a 'raw' form, available for use in different ways depending on the task.

The hierarchy of strategies we describe for *location* could be investigated using a similar experimental technique. In section 5, for example, we illustrate how points lying on a gaze-normal plane can be picked out by a simple strategy ($\Delta\theta_{PQ}$ is zero for all pairs of points, P and Q, on the plane) as the optic centre translates (or for binocular viewing). When the observer makes a saccade, the same plane is no longer gaze-normal, and a more elaborate algorithm would be required to determine whether a point lies in the plane. Both the scene and the locations of the optic centre(s) remain the same, only the line of sight has changed. Evidence on subjects' performance in these two cases (or similar experiments) could be one way in which to discover whether the visual system uses a hierarchy of strategies for determining the location of objects.

If observers made judgements (or movements) that required information about full Euclidean 3-D structure on every fixation, then the RVD model would lose much of its simplicity. However, the reverse is likely to be true. For example, during natural viewing one saccade is often followed rapidly by another, requiring no computation of 3-D structure. Preliminary studies have been made examining the role of individual fixations during complex natural tasks (e.g. Land and Furneaux, 1997). It would be valuable to extend such studies and to quantify the minimal level of computation required to control the motor behaviour occurring during each period of fixation. The results would help constrain models of the most efficient representation necessary to carry out those tasks.

Conclusion

Whether fixation simplifies or complicates the interpretation of retinal flow depends on what is being computed. If the aim is to compute translational flow and hence scene structure and direction of heading in an explicit 3-D frame, then fixation (and the consequent rotational flow as the observer moves) is indeed a complicating factor. We have argued for a different goal, in which actions are carried out in relation to the fixated object and the location of potential fixation targets is stored by recording their changing RVDs as the observer moves. We have shown how, if this is the goal, the act of maintaining gaze on a point as the observer moves is a positive advantage.

A representation of relative visual directions could act as a type of 'primal sketch' (Marr and Hildreth, 1980) of the optic array — a piece-wise retinotopic store of information lasting at least a few seconds — on which a range of motor and visual processes could draw. We have outlined the ways in which a representation of this sort might allow the visual system to operate successfully in a 3-D world without the need to generate a full 3-D representation of the scene, with all the co-ordinate transformations that implies, every time the observer moves their head or eyes.

Acknowledgement

Supported by the MRC and the Royal Society.

Appendix

This appendix contains details of (i) the image processing described in section 3 and how those images were placed in a common reference frame and (ii) the stage-by-stage recovery of direction of heading described in section 5.2.

Image acquisition and processing

Images were acquired using a video-resolution Pulnix CCD camera, rotating about a fixed point. Before acquisition, the camera intrinsic parameters were recovered using a reference object of known geometry and the calibration method described by Tsai (Tsai, 1986). Note that no information is used about the 3-D location or pose (i.e extrinsic parameters) of the camera.

To obtain image primitives, we processed the images according to the MIRAGE algorithm (Watt, 1987). Convolution with a Laplacian of Gaussian filter at three spatial scales, each separated by one octave, is followed by summation of the positive responses at each scale and summation of the negative responses to form a separate signal. The primitives used in this paper are the 2-D centroids of the zero-bounded regions in the negative response (i.e. the 'dark blobs'). Correspondence between primitives in successive frames was indicated manually.

Representation in one frame

Using the known camera calibration, the 2-D coordinates of primitives are converted to 3-D unit direction vectors ($\hat{\mathbf{n}}$) in the (arbitrary) camera reference frame. The optical centre is at O_1. Let the direction to the fixated feature F be denoted $\hat{\mathbf{n}}_F$, and the directions of to two other primitives P and Q be denoted $\hat{\mathbf{n}}_P$ and $\hat{\mathbf{n}}_Q$. The information recorded comprises: the eccentricities of P and Q, and the dihedral angle θ_{PQ} between the pair of planes (O_1, F, P) and (O_1, F, Q).

$$\begin{aligned} \cos\rho_P &= \cos\angle\mathbf{FP} = \hat{\mathbf{n}}_F \cdot \hat{\mathbf{n}}_P \\ \cos\rho_Q &= \cos\angle\mathbf{FQ} = \hat{\mathbf{n}}_F \cdot \hat{\mathbf{n}}_Q \\ \cos\theta_{PQ} &= \cos\angle(\mathbf{FP}, \mathbf{FQ}) = \frac{\hat{\mathbf{n}}_F \times \hat{\mathbf{n}}_P}{||\hat{\mathbf{n}}_F \times \hat{\mathbf{n}}_P||} \cdot \frac{\hat{\mathbf{n}}_F \times \hat{\mathbf{n}}_Q}{||\hat{\mathbf{n}}_F \times \hat{\mathbf{n}}_Q||} \end{aligned}$$

Figure 1 illustrates these angles where rays (O_1, F), (O_1, P) and (O_1, Q) correspond to the vectors $\mathbf{F}, \mathbf{P}$ and $\mathbf{Q}$ in directions $\hat{\mathbf{n}}_F$, $\hat{\mathbf{n}}_P$ and $\hat{\mathbf{n}}_Q$. From these data, we can recover the original directions—up to an arbitrary rotation of the reference frame—by the following procedure:

1. Choose $\hat{\mathbf{n}}_F = [0, 0, 1]$
2. Choose the YZ plane to contain P ($\hat{\mathbf{n}}_P = [0, \sin\rho_P, \cos\rho_P]$)
3. Set $\hat{\mathbf{n}}_Q = [\sin\theta_{PQ}\sin\rho_Q, \cos\theta_{PQ}\sin\rho_Q, \cos\rho_Q]$

Registration

In order to show that the proposed representation is sufficient to relate a series of eye rotations, we need only consider the registration of a pair of frames. Primitives in the second frame have directions $\hat{\mathbf{n}}'$ in a rotated coordinate system, related to the first by $\hat{\mathbf{n}}' = \mathbf{R}\hat{\mathbf{n}}$, with $\mathbf{R}$ a 3×3 rotation matrix. Recovering the arbitrary frame as above allows the recovery of $\mathbf{R}$.

Hierarchical recovery of direction of heading

It is assumed that (i) fixation is maintained on the point F, (ii) the line on the retina through the fovea (F') corresponding to the base plane (O_1, O_2, F) is known (see below), (iii) all translations are in this plane and (iv) translations are small with respect to the distance (O, F).

Recovering the projection of the base plane

In order to determine the line which is the intersection of the base plane (O_1, O_2, F) with the retina, we first observe that if there is no cyclotorsion, rotation is about the plane normal. Therefore, the image motion of all points in this plane will be restricted to the base line; or equivalently, such points have not tangential rotation, so their $\Delta\theta = 0$.

The effect of cyclotorsion is to add a constant angle to each observed $\Delta\theta$, so that all points which are on the base line will have equal $\Delta\theta$, corresponding to the negative of the amount of cyclotorsion. Therefore, given a reasonably dense image of point motions, lines through the fovea of constant $\Delta\theta$ represent candidates for the base line.

Because $\Delta\theta$ is trivially zero (before adding cyclotorsion) for points at infinity, they would give rise to false estimate of the base line. However, such points will also have zero $\Delta\rho$, and can therefore be easily excluded from the baseline computation.

Heading with respect to the line of sight

Let the direction of translation have two components, $\mathbf{u}$ and $\mathbf{v}$:

$$\mathbf{u} = u\hat{\mathbf{n}}_u \tag{1}$$
$$\mathbf{v} = v\hat{\mathbf{n}}_v \tag{2}$$

where $\hat{\mathbf{n}}_u$ is in the direction OF and $\hat{\mathbf{n}}_v$ is perpendicular to $\hat{\mathbf{n}}_u$ and in the base plane. The sign of v can be defined, arbitrarily, by relating it to the direction of a reference point Q in the base plane, such that:

$$\frac{\hat{\mathbf{n}}_u \times \hat{\mathbf{n}}_v}{||\hat{\mathbf{n}}_u \times \hat{\mathbf{n}}_v||} = \frac{\hat{\mathbf{n}}_u \times \hat{\mathbf{n}}_Q}{||\hat{\mathbf{n}}_u \times \hat{\mathbf{n}}_Q||}. \tag{3}$$

If, for example, the translation $(\mathbf{u} + \mathbf{v})$ is in the horizontal plane, and Q is the the right of the fixation point, then for all translations to the right, v has a positive sign.

For an unknown $\hat{\mathbf{n}}_v$, the sign of v (i.e. $\frac{v}{|v|}$) can be recovered simply from the change in θ of an arbitrary point in the scene, P, (or from many points in the scene), as follows (Weinshall, 1990):

$$\frac{v}{|v|} = \frac{\Delta\theta_P}{|\Delta\theta_P|}\frac{d_P}{|d_P|} \tag{4}$$

where d_P is the depth of P with respect to gaze-normal plane. The sign of d_P (i.e. $\frac{d_P}{|d_P|}$) could be obtained from the binocular disparity of P. θ_P is defined relative to the base plane as:

$$\cos\theta_P = \frac{\hat{\mathbf{n}}_F \times \hat{\mathbf{n}}_P}{||\hat{\mathbf{n}}_F \times \hat{\mathbf{n}}_P||} \cdot \frac{\hat{\mathbf{n}}_F \times \hat{\mathbf{n}}_Q}{||\hat{\mathbf{n}}_F \times \hat{\mathbf{n}}_Q||} \tag{5}$$

where Q lies in the base plane. Defined in this way, near points all have the same sign of $\Delta\theta$ when the optic centre translates.

Ratios of v for different translations

$\Delta\theta_P$ gives not only the sign but also the relative magnitude of v for different translations (Weinshall, 1990). For small translations, we may assume that v is linearly related to $\Delta\theta_P$:

$$v = k_{1P}\Delta\theta_P \tag{6}$$

where k_{1P} is constant across different translations but is specific to the point P.

Ratios of u for different translations

For a point, A, in the plane that is perpendicular to the base plane and which passes through OF:

$$u = k_{2A}\Delta\rho_A \tag{7}$$

where A is a point that lies in this 'perpendicular' plane. For example, if the translation is in the horizontal plane, A is a point lying on the vertical meridian. k_{2A} is constant across different translations but is specific to the point A.

The tangent of the direction of heading is now known up to a scale factor:

$$\tan\rho_e = \frac{v}{u} = k3\frac{\Delta\theta_P}{\Delta\rho_A} \tag{8}$$

where ρ_e is the angle of direction of heading (or epipole) with respect to (OF) and $k3 = k_{1P}/k_{2A}$.

Recovering the direction of heading

There may be a variety of ways to recover $k3$ and hence the direction of heading. For example, if the observer makes a set of head movements in random directions, the expectation is that values of computed ρ_e will have a flat frequency distribution. Incorrect estimates of $k3$ will cause the distribution to be peaked at directions of heading of 0 and 180° or 90 and 270°, where $\hat{\mathbf{n}}_F$ defines 0°. This information could be used to modify and improve the estimate of $k3$. An alternative, more precise method, which we describe here, requires a point in the gaze-normal plane to be identified.

There is a particular direction of heading, $\arctan v_t/u_t$, such that $\Delta\rho_B = 0$ and $\Delta\theta_B = 0$:

$$\frac{v_t}{u_t} = -\cot\rho_B \csc\theta_B \tag{9}$$

which arises when (i) the component of translation in the plane (O, F, B) is tangential to the circle passing through O, F and B and (ii) B is in the gaze-normal plane. The observer does not need to make this translation, but equation 9 allows $k3$ to be computed, as follows.

In the case of translation $(\mathbf{u}_t + \mathbf{v}_t)$, there is no change in ρ_B. $\Delta\rho$ is zero for other translations along the tangent to the circle but, in this particular case, when angle $\angle OFB$ is 90°, the direction of the tangent, $\arctan(v_t \cos\theta_B/u_t)$, can be found simply from ρ_B and θ_B. Points in the gaze-normal plane are readily identifiable (Weinshall, 1990; Liu et al, 1994), having the property that, to a good approximation, $\Delta\theta_B = 0$ for all directions of translations of the optic centre.

In general, $\Delta\rho$ for points has a contribution from each component of translation, $\mathbf{u}$ and $\mathbf{v}$. Thus,

$$\Delta\rho_B = k_{4B}\Delta\rho_A + k_{5B}\Delta\theta_P \tag{10}$$

since $\Delta\rho_A$ is proportional to u and $\Delta\theta_P$ is proportional to v.

Only the ratio of u to v is required to recover the direction of heading. So, from equations 8 and 10:

$$\Delta\rho_B = c[k_{4B}k3u + k_{5B}v] \tag{11}$$

where c is constant across different translations. From equations 9 and 11:

$$k3 = \frac{k_{5B}}{k_{4B}} \cot\rho_B \csc\theta_B. \tag{12}$$

The constants k_{4B} and k_{5B} can be found from observing $\Delta\rho_B$ for two different translations and using equation 10. The solutions are:

$$k_{4B} = -\frac{(\Delta\theta_{P_1}\Delta\rho_{B_2}) - (\Delta\theta_{P_2}\Delta\rho_{B_1})}{(\Delta\rho_{A_1}\Delta\theta_{P_2}) - (\Delta\rho_{A_2}\Delta\theta_{P_1})} \tag{13}$$

$$k_{5B} = -\frac{(\Delta\rho_{A_2}\Delta\rho_{B_1}) - (\Delta\rho_{A_1}\Delta\rho_{B_2})}{(\Delta\rho_{A_1}\Delta\theta_{P_2}) - (\Delta\rho_{A_2}\Delta\theta_{P_1})} \tag{14}$$

Thus, from equations 8, 12 and 13, the tangent of the direction of heading, $\tan \rho_e$, is given by:

$$\tan \rho_e = \frac{v}{u} = \cot \rho_B \csc \theta_B \frac{(\Delta\rho_{A_2}\Delta\rho_{B_1}) - (\Delta\rho_{A_1}\Delta\rho_{B_2})}{(\Delta\theta_{P_1}\Delta\rho_{B_2}) - (\Delta\theta_{P_2}\Delta\rho_{B_1})} \frac{\Delta\theta_P}{\Delta\rho_A}. \tag{15}$$

so direction of heading, ρ_e, can be computed from just $\Delta\rho$ of A and B, $\Delta\theta$ of P and the retinal location (ρ, θ) of B.

References

Aloimonos, Y., Weiss, I., and Bandopadhay, A. (1987). Active vision. *Proceedings of the International Conference on Computer Vision*, pages 35–54, London, UK, June 8–11.

Andersen, R. A., Snyder, L. H., Bradley, D. C., and Xing, J. (1997). Multi-modal representation of space in the posterior parietal cortex and its use in planning movements. *Annual Review of Neuroscience*, 20:303–330.

Arbib, M. (1999). Parietal cortex and hippocampus: from visual affordances to the world graph. In Burgess, N., Jeffery, K. J., and O'Keefe, J., editors, *The hippocampal and parietal foundations of spatial cognition*, pages 416–442. Oxford: OUP.

Bahcall, D. O. and Kowler, E. (1999). Illusory shifts in visual direction accompany adaptation of saccadic eye movements. *Nature*, 400:864–866.

Bandopadhay, A. and Ballard, D. (1990). Egomotion perception using visual tracking. *Computational Intelligence*, 7:39–47.

Barash, S., Bracewell, R. M., Fogassi, L., and Gnadt, J. W. Andersen, R. A. (1991). Saccade-related activity in the lateral intraparietal area 1: temporal properties - comparison with area 7a. *Journal of Neurophysiology*, 66:1095–1108.

Barron, J. L., Fleet, D. J., and Beauchemin, S. S. (1994). Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77.

Beintema, J. A. and van den Berg, A. V. (1998). Heading detection using motion templates and eye velocity gain fields. *Vision Research*, 38:2155–2179.

Bradley, D. C., Maxwell, M., Andersen, R. A., Banks, M. S., and Shenoy, K. V. (1996). Mechanisms of heading perception in primate visual cortex. *Science*, 273:1544–1547.

Bridgeman, B., van der Heijden, A. H. C., and Velichovsky, B. M. (1994). A theory of visual stability across saccadic eye movements. *Behavioural and Brain Sciences*, 17:247–292.

Caminiti, R., Johnson, P. B., Galli, C., Ferraina, S., and Burnod, Y. (1991). Making arm movements within different parts of space - the premotor and motor cortical representation of a coordinate system for reaching to visual targets. *Journal of Neuroscience*, 11:1182–1197.

Cartwright, B. A. and Collett, T. S. (1983). Landmark learning in bees: experiments and models. *Journal of Comparative Physiology*, 151:521–543.

Colby, C. L. (1998). Action-oriented spatial reference frames in cortex. *Neuron*, 20:15–24.

Colby, C. L. and Duhamel, J. R. (1991). Heterogeneity of extrastriate visual areas and multiple parietal areas in the macaque monkey. *Neuropsychologia*, 29:517–537.

Crowell, J. A., Banks, M. S., Shenoy, K. V., and Andersen, R. A. (1998). Visual self-motion perception during head turns. *Nature Neuroscience*, 1:732–737.

Cutting, J. E. (1986). *Perception with an Eye to Motion*. Cambridge, Mass: MIT Press.

Cutting, J. E., Springer, K., Braren, P. A., and Johnson, S. H. (1992). Wayfinding on foot from information in retinal, not optical, flow. *Journal of Experimental Psychology-General*, 121:41–72.

Daniilidis, K. (1997). Fixation simplifies 3D motion estimation. *Computer Vision and Image Understanding*, 68:158–169.

Duffy, C. J. and Wurtz, R. H. (1991). Sensitivity of MST neurons to optic flow stimuli. I. A continuum of response selectivity to large-field stimuli. *Journal of Neurophysiology*, 65:1329–1345.

Duffy, C. J. and Wurtz, R. H. (1995). Response of monkey MST neurons to optic flow stimuli with shifted centers of motion. *Journal of Neuroscience*, 15:5192–5208.

Duhamel, J. R., Bremmer, F., BenHamed, S., and Graf, W. (1997). Spatial invariance of visual receptive fields in parietal cortex neurons. *Nature*, 389:845–848.

Duhamel, J. R., Colby, C. L., and Goldberg, M. E. (1998). Ventral intraparietal area of the macaque: Congruent visual and somatic response properties. *Journal of Neurophysiology*, 79:126–136.

Feldman, J. A. (1985). Four frames suffice: A provisional model of vision and space. *Behavioural and Brain Sciences*, 8:265–289.

Ferman, L., Collewijn, H., Jansen, T. C., and vanden Berg, A. V. (1987). Human gaze stability in the horizontal vertical and torsional direction during voluntary head movements, evaluated with a three- dimensional scleral induction coil technique. *Vision Research*, 27:811–828.

Findlay, J. M. and Gilchrist, I. D. (1997). Spatial scale and saccade programming. *Perception*, 26:1159–1167.

Freedman, E. G. and Sparks, D. L. (1997). Activity in the deeper layer of the superior colliculus of the rhesus monkey: evidence for a gaze displacement command. *Journal of Neurophysiology*, 78:1669–1690.

Galletti, C., Battaglini, P. P., and Fattori, P. (1993). Parietal neurons encoding spatial locations in craniotopic coordinates. *Experimental Brain Research*, 96:221–229.

Gårding, J., Porrill, J., Mayhew, J. E. W., and Frisby, J. P. (1995). Stereopsis, vertical disparity and relief transformations. *Vision Research*, 35:703–722.

Gibson, J. J. (1950). *The perception of the visual world*. Boston: Houghton Mifflin.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Glennerster, A., Rogers, B. J., and Bradshaw, M. F. (1996). Stereoscopic depth constancy depends on the subject's task. *Vision Research*, 36:3441–3456.

Goldberg, M. E. and Bruce, C. J. (1990). Primate frontal eye fields. III. Maintenance of a spatially accurate saccade signal. *Journal of Neurophysiology*, 64:489–508.

Graziano, M. S. A., Yap, G. S., and Gross, C. G. (1994). Coding of visual space by premotor neurons. *Science*, 266:1054–1057.

Heeger, D. J. and Jepson, A. D. (1992). Subspace methods for recovering rigid motion .1. Algorithm and implementation. *International Journal of Computer Vision*, 7:95–117.

Held, R. and Hein, A. (1958). Adaptation of disranged hand-eye co-ordination contingent upon re-afferent stimulation. *Perceptual and Motor Skills*, 8:87–90.

Helmholtz, H. v. (1867). *Handbuch der Physiologischen Optik*. 1st Ed (Voss: Hamburg), 3rd Ed. translated by J P C Southall for the Optical Society of America, 1924.

Henriques, D. Y. P., Klier, E. M., Smith, M. A., Lowy, D., and Crawford, J. D. (1998). Gaze-centered remapping of remembered visual space in an open-loop pointing task. *Journal of Neuroscience*, 18:1583–1594.

Irani, M. and Anandan, P. (1998). Video indexing based on mosaic representation. *Proceedings of the IEEE*, 86:905–921.

Jay, M. F. and Sparks, D. L. (1984). Auditory receptive-fields in primate superior colliculus shift with changes in eye position. *Nature*, 309:345–347.

Johnston, E. B. (1991). Systematic distortions of shape from stereopsis. *Vision Research*, 31:1351–1360.

Judd, S. P. D. and Collett, T. S. (1998). Multiple stored views and landmark guidance in ants. *Nature*, 392:710–714.

Koenderink, J. J. and van Doorn, A. J. (1987). Facts on optic flow. *Biological Cybernetics*, 56:247–254.

Koenderink, J. J. and van Doorn, A. J. (1991). Affine structure from motion. *Journal of the Optical Society of America A-Optics Image Science and Vision*, 8:377–385.

Krekelberg, B., Paolini, M., Bremmer, F., Lappe, M., and Hoffman, K.-P. (2000). Heading encoding in MST during simulated eye movements. *Perception*, 29 (Suppl):119.

Lagae, L., Maes, H., Raiguel, S., Xiao, D. K., and Orban, G. A. (1994). Responses of macaque sts neurons to optic flow components - a comparison of areas MT and MST. *Journal of Neurophysiology*, 71:1597–1626.

Land, M. F. and Furneaux, S. (1997). The knowledge base of the oculomotor system. *Philosophical Transactions of the Royal Society of London Series B- Biological Sciences*, 352:1231–1239.

Lappe, M. (1996). Functional consequences of an integration of motion and stereopsis in area MT of monkey extrastriate visual cortex. *Neural Computation*, 8:1449–1461.

Lappe, M., Bremmer, F., and van den Berg, A. V. (1999). Perception of self-motion from visual flow. *Trends in Cognitive Sciences*, 3:329–336.

Lappe, M. and Rauschecker, J. P. (1993). A neural network for the processing of optic flow from ego-motion in man and higher mammals. *Neural Computation*, 5:374–391.

Lappe, M. and Rauschecker, J. P. (1994). Heading detection from optic flow. *Nature*, 369:712–713.

Lappe, M. and Rauschecker, J. P. (1995). Motion anisotropies and heading detection. *Biological Cybernetics*, 72:261–277.

Liu, L., Stevenson, S. B., and Schor, C. M. (1994). A polar coordinate system for describing binocular disparity. *Vision Research*, 34:1205–1222.

Longuet-Higgins, H. C. and Prazdny, K. (1980). The interpretation of moving retinal images. *Proceedings of the Royal Society, London* B, 208:385–397.

Marr, D. and Hildreth, E. (1980). Theory of edge detection. *Proceedings of the Royal Society of London* B, 207:187–217.

Merton, P. A. (1961). The accuracy of directing the eyes and hand in the dark. *Journal of Physiology*, 156:555–577.

Miall, R. C., Weir, D. J., Wolpert, D. M., and Stein, J. F. (1993). Is the cerebellum a Smith predictor? *Journal of Motor Behaviour*, 25:203–216.

Morgan, C. L. (1978). Constancy of egocentric visual direction. *Perception and Psychophysics*, 23:61–68.

Müller, R., Röfer, T., Lankenau, A., Musto, A., Stein, K., and Eisenkolb, A. (2000). Coarse qualitative descriptions in robot navigation. In Freksa, C., Brauer, W., Habel, C., and Wender, K. F., editors, *Spatial Cognition II. Lecture Notes in Artificial Intelligence 1849*, pages 265–276. Heidelberg: Springer.

Murray, D. W., Reid, I. D., and Davison, A. J. (1997). Steering without representation with the use of active fixation. *Perception*, 26:1519–1528.

Perenin, M. T., Jeannerod, M., and Prablanc, C. (1977). Spatial localisation with paralysed eye muscles. *Ophthalmologica*, 175:206–214.

Perrone, J. A. and Stone, L. S. (1994). A model of self-motion estimation within primate extrastriate visual- cortex. *Vision Research*, 34:2917–2938.

Regan, D. and Beverley, K. I. (1982). How do we avoid confounding the direction we are looking and the direction we are moving. *Science*, 215:194–196.

Reid, I. and Murray, D. (1996). Active tracking of foveated feature clusters using affine structure. *International Journal of Computer Vision*, 18(1):41–60.

Roy, J. P., Komatsu, H., and Wurtz, R. H. (1992). Disparity sensitivity of neurons in monkey extrastriate area MST. *Journal of Neuroscience*, 12:2478–2492.

Roy, J. P. and Wurtz, R. H. (1990). The role of disparity-sensitive cortical neurons in signalling the direction of self-motion. *Nature*, 348:160–162.

Royden, C. S., Crowell, J. A., and Banks, M. S. (1994). Estimating heading during eye-movements. *Vision Research*, 34:3197–3214.

Russo, G. S. and Bruce, C. J. (1994). Frontal eye field activity preceding aurally guided saccades. *Journal of Neurophysiology*, 71:1250–1253.

Saito, H., Yukie, M., Tanaka, K., Hikosaka, K., Fukada, Y., and Iwai, E. (1986). Integration of direction signals of image motion in the superior temporal sulcus of the macaque monkey. *Journal of Neuroscience*, 6:145–157.

Sandini, G. and Tistarelli, M. (1990). Active tracking strategy for monocular depth inference over multiple frames. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:13–27.

Shibutani, H., Sakata, H., and Hyvarinen, J. (1984). Saccade and blinking evoked by microstimulation of the posterior parietal association cortex of the monkey. *Experimental Brain Research*, 55:1–8.

Stone, L. S. and Perrone, J. A. (1997). Human heading estimation during visually simulated curvilinear motion. *Vision Research*, 37:573–590.

Stricanne, B., Andersen, R. A., and Mazzoni, P. (1996). Eye-centered, head-centered, and intermediate coding of remembered sound locations in area LIP. *Journal of Neurophysiology*, 76:2071–2076.

Sun, H. J. and Frost, B. J. (1998). Computation of different optical variables of looming objects in pigeon nucleus rotundus neurons. *Nature Neuroscience*, 1:296–303.

Tanaka, K., Hikosaka, K., Saito, H., Yukie, M., Fukada, Y., and Iwai, E. (1986). Analysis of local and wide-field movements in the superior temporal visual areas of the macaque monkey. *Journal of Neuroscience*, 6:134–144.

Thier, P. and Andersen, R. A. (1996). Electrical microstimulation suggests two different forms of representation of head-centered space in the intraparietal sulcus of rhesus monkeys. *Proceedings of the National Academy of Sciences of the United States of America*, 93:4962–4967.

Tittle, J. S., Todd, J. T., Perotti, V. J., and Norman, J. F. (1995). A hierarchical analysis of alternative representations in the perception of 3-d structure from motion and stereopsis. *Journal of Experimental Psychology: Human Perception and Performance*, 21:663–678.

Tsai, R. Y. (1986). An efficient and accurate camera calibration technique for 3D machine vision. In *Proceedings, Computer Vision and Pattern Recognition*, pages 364–374.

van den Berg, A. V. (1999). Predicting the present direction of heading. *Vision Research*, 39:3608–3620.

van den Berg, A. V. and Brenner, E. (1994). Why 2 eyes are better than one for judgments of heading. *Nature*, 371:700–702.

Warren, W. H. and Hannon, D. J. (1990). Eye-movements and optical-flow. *Journal of the Optical Society of America A*, 7:160–169.

Warren, W. H., Morris, W. M., and Kalish, M. (1988). Perception of translational heading from optical flow. *Journal of Experimental Psychology: Human Perception and Performance*, 14:646–660.

Watt, R. J. (1987). 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.

Weinshall, D. (1990). Qualitative depth from stereo, with applications. *Computer Vision Graphics and Image Processing*, 49:222–241.

Zipser, D. and Andersen, R. A. (1988). A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, 331:679–684.

Figure legends

Figure 1 Combining images related by rotations of the eye or camera.

a) An eye is shown pointing in two directions. The red and blue arcs (portions of a great circle) joining each image feature to the fovea define their relative visual directions (RVDs) with respect to the fovea (and hence the potential saccades required to fixate the corresponding objects). The sphere on the right shows how the retinal images from these two views could be related to form a representation of the RVDs of objects.

b) A set of 22 images (of which 6 are shown here), obtained by rotating a camera about a fixed point, were filtered (2 examples shown), and the directions of 6 features per image were placed in a common reference frame for visual direction (sphere). J is a nominal fixation point in one of the filtered images. The red lines (and corresponding arcs on the sphere) meeting at J define planes through the optic centre, J and five other points. As described in the Appendix, to register these directions with the directions of points visible on the next fixation (K), two correspondences are required (features J and n in this example).

Figure 2: The effect of observer translation on relative visual directions (RVDs).

a) The visual directions of a set of points, as in figure 1a. When the optic centre translates, the visual direction of near points (shown by the white discs) changes with respect to the directions of the distant points (black discs).

b) The colour code here summarises the effect of translating the optic centre in random directions (100 translations of unit magnitude). It shows the mean change in the angle subtended by two points at the optic centre (expressed as a proportion of the initial angle, $|\Delta\rho/\rho_1|$). The width of the arcs varies with the colour. The width is proportional to the log of $|\Delta\rho/\rho_1|$. The near points are 100 and the distant point 1000 times the magnitude of the translations.

c) Change in the angular separation of a pair of points ($|\Delta\rho/\rho_1|$) varies with (i) distance from the observer, D, (ii) the depth difference between points, $(s - D)$ and (iii) their angular separation (here, $\rho = 45°$). Translation magnitude is 1. If ρ was small, e.g. 1°, then the function would dip down towards zero at $s/D = 1$. In the case shown here, near points can be distinguished from a more distant set without knowing the directions of translation or the relative depths of the points $(s - D)$.

Figure 3: Retinal motion and disparity provide a direct measure of changes in relative visual direction (RVD) with respect to the fixated object.

a) Rays from F, P and Q pass through the optic centre at location O_1 and project to the points F', P' and Q' on the spherical retina (centred on the optic centre, O). Taking F' as the fovea, the retinal location P' can be described by its eccentricity, ρ_P (which is also the angle FO_1P) and the polar angle θ_{PQ}, measured with respect to the retinal location Q'. (θ_{PQ} is also the angle between the planes FO_1P and FO_1Q).

b) When the optic centre translates from location O_1 to O_2 while the observer maintains fixation on F, the motion of P' on the retina has two components: a change in eccentricity, $\Delta\rho_P$, and a perpendicular component, $\Delta\theta_P$. In the example shown here, the translation of the optic centre from O_1 to O_2 is in the plane FO_1Q, so the $\Delta\theta$ component of retinal motion at P' signals the change in the angle between the planes (F, O, P) and (F, O, Q) – i.e. in this case, $\Delta\theta_P = \Delta\theta_{PQ}$ (see text). In the more general case, the motions

at both P' and Q' are required to compute the change in the angle θ_{PQ}.

Figure 4: $\Delta\theta$ and scene structure.

The three retinal projections shown here are in the same format as figure 3b, but the distance OP differs in each case. At the top, P is more distant than the gaze-normal plane through F; in the centre, P lies in the gaze-normal plane and at the bottom, P is closer than the gaze-normal plane. In each case, as the optic centre translates from O_1 to O_2 to O_3, the projection of P, P', becomes more eccentric ($\Delta\rho_P$ increases). $\Delta\theta_P$, on the other hand, is negative when P is beyond the gaze-normal plane (top), zero when P lies in the gaze-normal plane (middle) and positive when P is closer than the gaze-normal plane (bottom).

The contour plot shows how θ_P is affected by the distance of P and the translation of O. The x-axis shows the magnitude of the translation in the direction (O_1, O_2, O_3). The y-axis gives the distance (O_2, P). The unit of distance in both cases is the length (O_2, F). The dotted line shows the distance at which P lies in the gaze-normal plane. In this example, the direction of (O_2, P) is (1,1,1) where O_2 is the origin, and the axes are defined by the direction (O_2, F) and the plane (F, O, Q)).

The same overall pattern of θ_P values are observed (i.e. a change in the sign of $\Delta\theta$ depending on the direction of translation and the distance (O_2, P) relative to the gaze-normal plane) independent of the retinal location of P', provided that O_1, O_2, F and P are not co-planar. The pattern is also independent of the direction of translation (O_1, O_2, O_3), provided this remains to one side of (O_2, F).

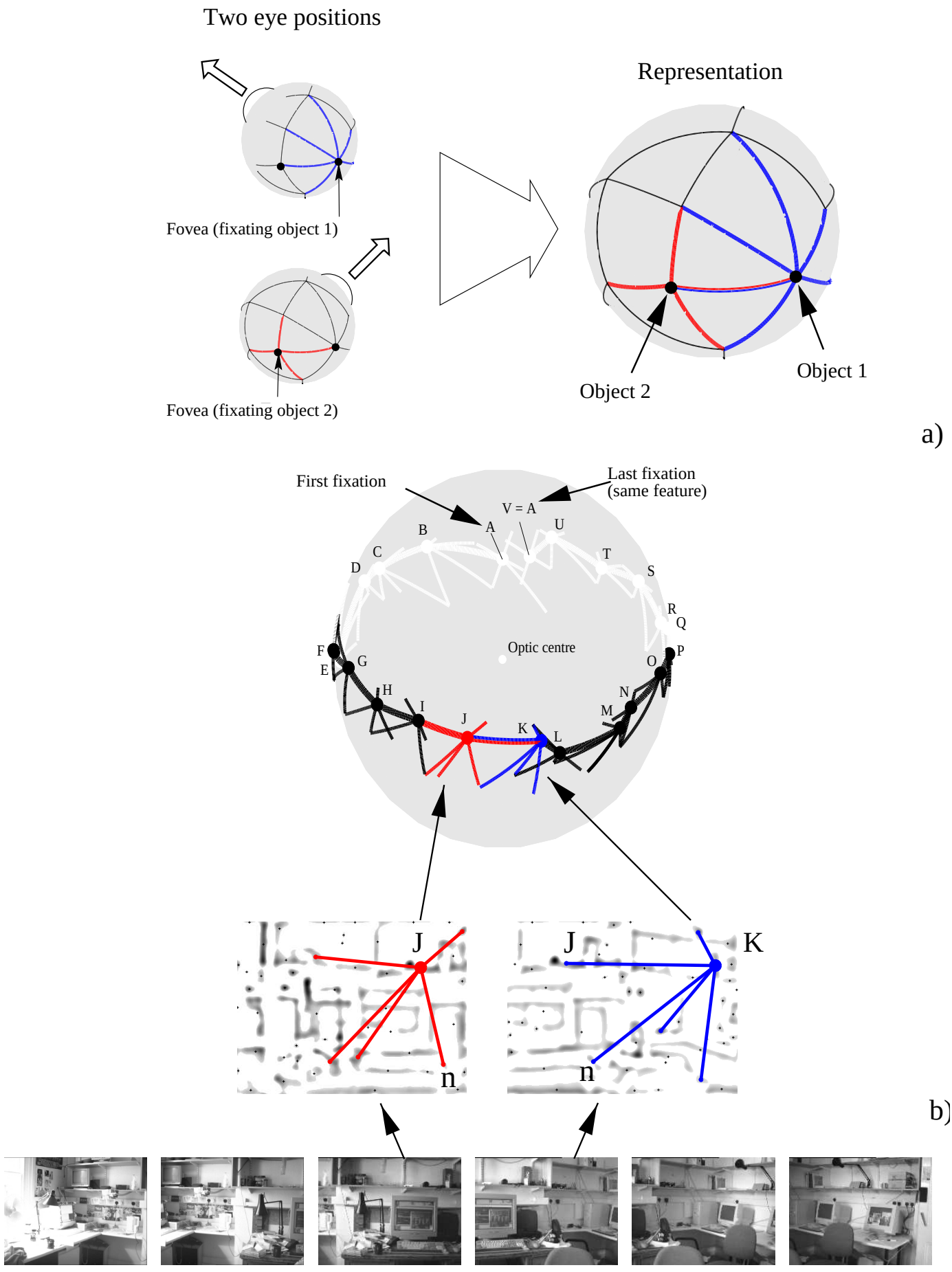
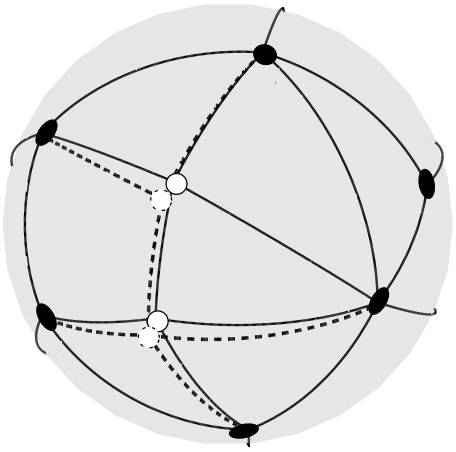
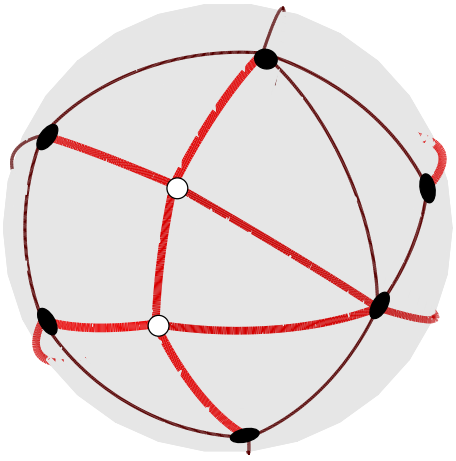

Two eye positions
Representation
Fovea (fixating object 1)
Fovea (fixating object 2)
Object 1
Object 2
a)
First fixation
Last fixation
(same feature)
V = A
A
B
C
D
U
T
S
R
Q
P
O
N
M
L
K
J
I
H
G
F
E
Optic centre
J
n
J
K
n
b)

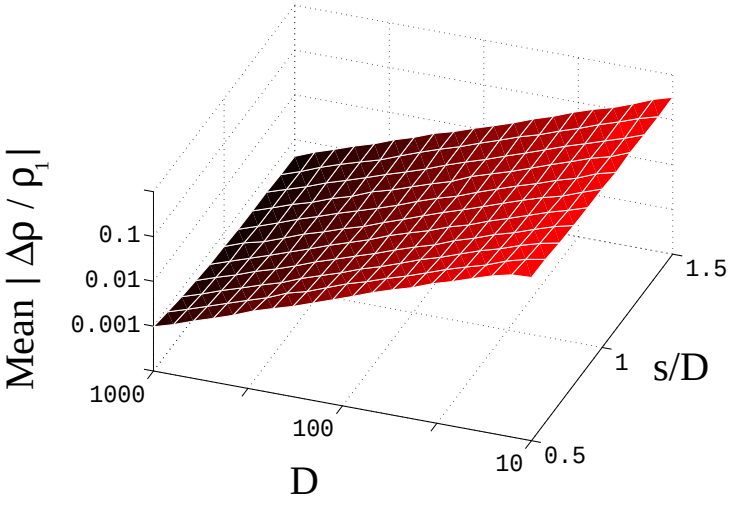
Figure 1

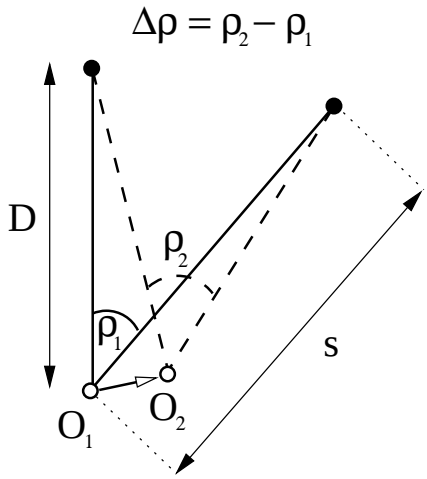


a)


0.01
0.001
0.0001
Mean | Δρ / ρ1 |

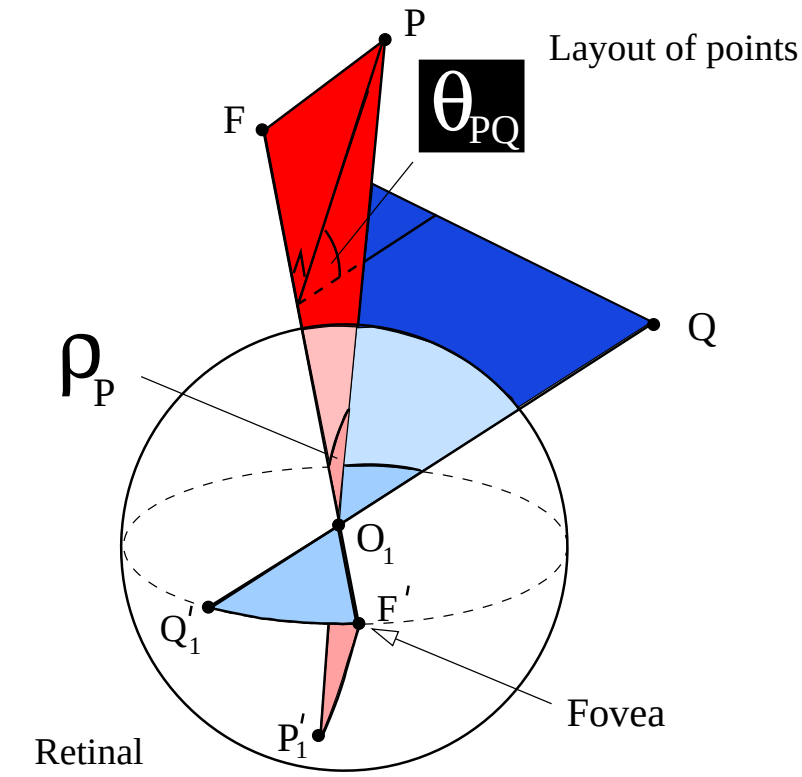
b)


Mean | Δρ / ρ1 |
0.1
0.01
0.001
1000
100
10
D
0.5
1
1.5
s/D

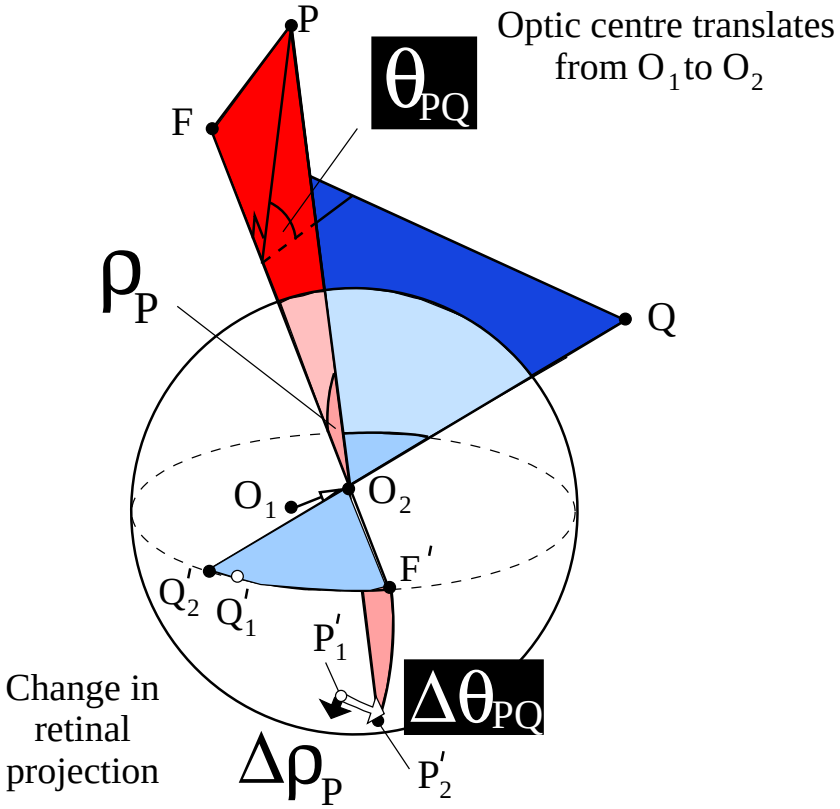

Δρ = ρ2 – ρ1
D
ρ2
ρ1
s
O1
O2

c)

Figure 2


P
Layout of points
F
θPQ
Q
ρP
O1
F′
Q′1
Fovea
P′1
Retinal
projection

a)


P
Optic centre translates
from O1 to O2
θPQ
F
ρP
Q
O1
O2
F′
Q′2
Q′1
P′1
ΔθPQ
Change in
retinal
projection
ΔρP
P′2

b)

Figure 3

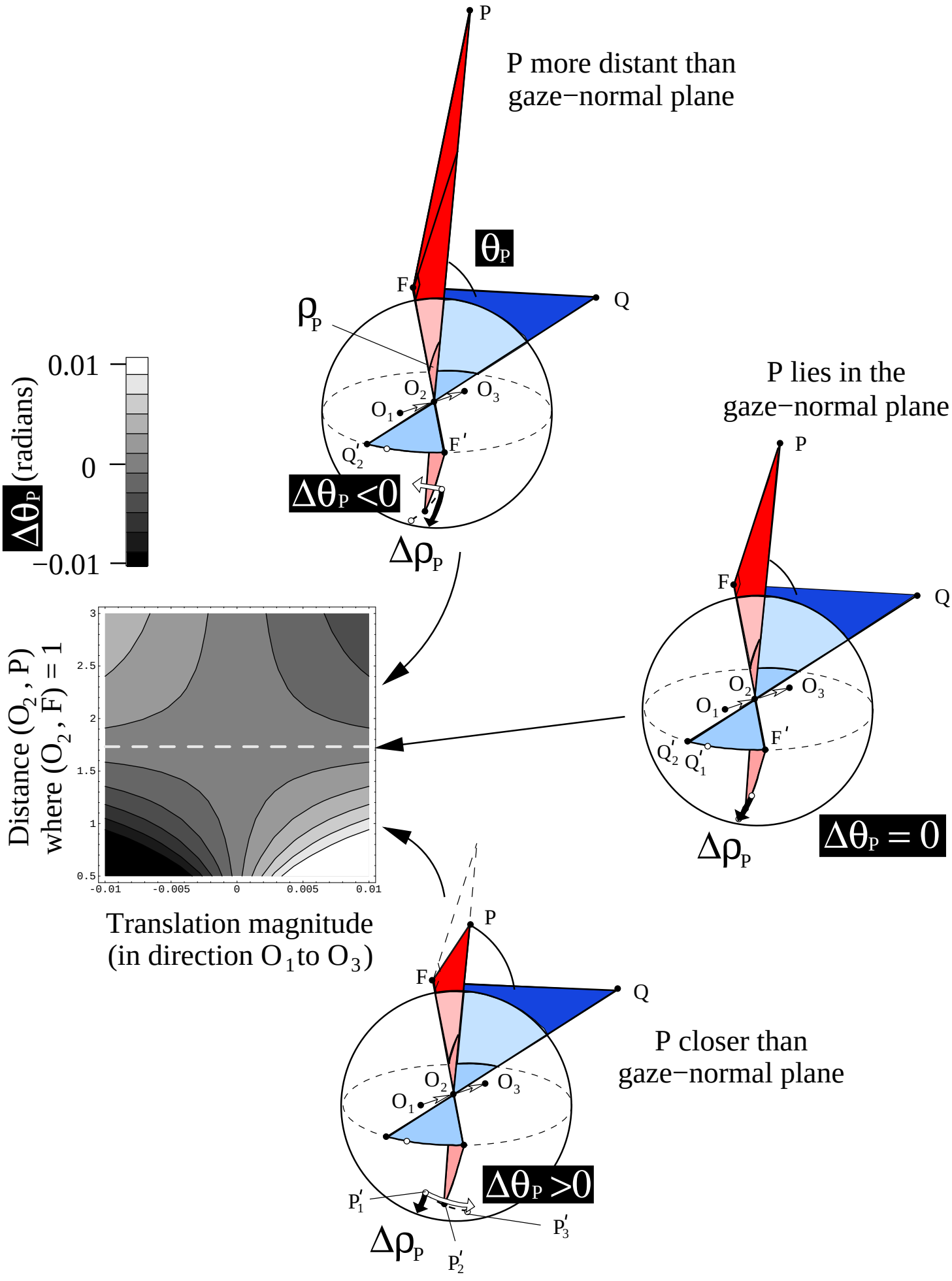

P
P more distant than
gaze−normal plane
θP
F
Q
ρP
O2
O3
O1
F′
Q′2
ΔθP <0
ΔρP
0.01
0
−0.01
ΔθP (radians)
P lies in the
gaze−normal plane
P
F
Q
O2
O3
O1
F′
Q′2
Q′1
ΔρP
ΔθP = 0
Distance (O2 , P)
where (O2 , F) = 1
3
2.5
2
1.5
1
0.5
-0.01
-0.005
0
0.005
0.01
Translation magnitude
(in direction O1 to O3)
P
F
Q
O2
O3
O1
ΔθP >0
P′1
P′3
ΔρP
P′2
P closer than
gaze−normal plane

Figure 4