HAL
open science

View-based approaches to Spatial Representation in Human Vision

Andrew Glennerster, Miles Hansard, Andrew Fitzgibbon

To cite this version:

Andrew Glennerster, Miles Hansard, Andrew Fitzgibbon. View-based approaches to Spatial Representation in Human Vision. Dagstuhl Seminar on Statistical and Geometrical Approaches to Visual Motion Analysis, Jul 2008, Dagstuhl, Germany. pp.193-208, 10.1007/978-3-642-03061-1_10 . inria-00435556

HAL Id: inria-00435556

https://inria.hal.science/inria-00435556

Submitted on 24 Nov 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

View-Based Approaches to Spatial Representation in Human Vision

Andrew Glennerster[1], Miles E. Hansard[2], and Andrew W. Fitzgibbon[3]

[1] University of Reading, Reading, UK
a.glennerster@reading.ac.uk
http://www.personal.rdg.ac.uk/~sxs05ag/

[2] INRIA Rhône-Alpes, Montbonnot, France
[3] Microsoft Research, Cambridge, UK

Abstract. In an immersive virtual environment, observers fail to notice the expansion of a room around them and consequently make gross errors when comparing the size of objects. This result is difficult to explain if the visual system continuously generates a 3-D model of the scene based on known baseline information from interocular separation or proprioception as the observer walks. An alternative is that observers use view-based methods to guide their actions and to represent the spatial layout of the scene. In this case, they may have an expectation of the images they will receive but be insensitive to the rate at which images arrive as they walk. We describe the way in which the eye movement strategy of animals simplifies motion processing if their goal is to move towards a desired image and discuss dorsal and ventral stream processing of moving images in that context. Although many questions about view-based approaches to scene representation remain unanswered, the solutions are likely to be highly relevant to understanding biological 3-D vision.

1 Is Optic Flow Used for 3-D Reconstruction in Human Vision?

Optic flow, or motion parallax, provides animals with information about their own movement and the 3-D structure of the scene around them. Throughout evolution, motion is likely to have been more important for recovering information about scene structure than binocular stereopsis, which is predominantly used by hunting animals who are required to remain still. The discrimination of 3-D structure using motion parallax signals is known to be highly sensitive, almost as sensitive as for binocular stereopsis [1,2]. Motion parallax has also been shown to provide an estimate of the viewing distance and metric shape of objects, when combined with proprioceptive information about the distance the observer has travelled [3,4]. An important and unsolved challenge, however, is to understand how this information is combined into a consistent representation as the observer moves through a scene.

D. Cremers et al. (Eds.): Visual Motion Analysis, LNCS 5604, pp. 193–208, 2009.
© Springer-Verlag Berlin Heidelberg 2009

In computer vision, the problem is essentially solved. Photogrammetry is the process of using optic flow to recover information about the path of the camera, its internal parameters such as focal length and the 3-D structure of the scene [5,6]. For most static scenes the process is robust, reliable and geometrically accurate. It does not suffer from the systematic spatial distortions that are known to afflict human observers in judgements of metric shape and distance [7,8,9,10,11]. Most approaches to visual navigation in robotics are based on a partial reconstruction of the 3D environment, again using photogrammetric principles [12,13,14].

It is questionable whether human vision is designed to achieve anything like photogrammetric reconstruction. Gibson [15], for example, argued strongly against the idea that vision required explicit internal representation, but did not provide a clear description of an alternative [16]. Similar arguments are still popular [17] but are no more computationally precise. In fact, although there have been some attempts at generating non-metric representations for navigation in robots (see Section 3), there is as yet no well-developed rival to 3-D reconstruction as a model for representing the spatial layout of a scene, either in computer vision or models of human vision.

1.1 An Expanding Room

In our Virtual Reality Laboratory, we have developed a paradigm in which information about the distance of objects from vergence, motion parallax and proprioception conflict with the assumption of scene stability. Observers wear a head mounted display that has a wide-field of view and high resolution. In an immersive virtual environment, they make judgements about the relative size or distance of objects, allowing us to probe the representation of space generated by the visual system and to assess the way in which different cues are combined. Figure 1 illustrates one experiment [18]. As the observer moves from one side of the room to the other, the virtual room expands around him/her. Remarkably, almost all observers fail to notice any change in size of the room even though the room expands by a factor of 4. Even with feedback about the true size of objects in the room, they fail to learn to use veridical stereo and motion parallax cues appropriately [18,19].

Our data suggest a process of scene representation that is very different from photogrammetry. For example, Figure 1b shows data for 5 human observers who were asked to judge the size of a cube seen on the right of the room (when the room is large) compared to a reference cube shown on the left of the room (when the room is small). The reference cube disappears when the observer walks across the room, so they must remember its size. Participants make a binary, forced choice judgement on each trial about the relative size of the reference and comparison cubes. Over many trials, it is possible to deduce the size of the comparison cube that they would match with the reference cube. As Figure 1b shows, this size varies with the viewing distance of the comparison cube: at 6m participants choose a match that is almost four times as large as the reference cube, i.e. almost equal to the expansion of the room, whereas at 1.5m they choose a more veridical match.

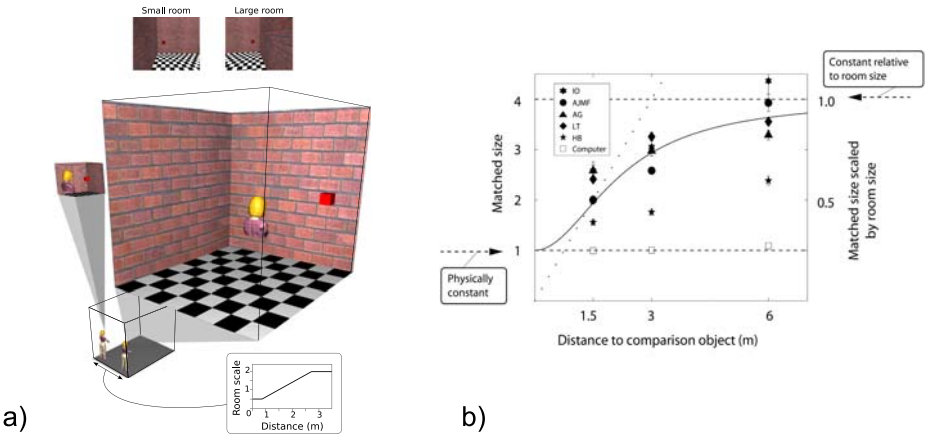

Fig. 1. An expanding virtual room. a) As the observer walks from one side of the room to the other the virtual scene around them expands. The centre of expansion is mid-way between the eyes so the view from that point is unaffected. This stimulus separates purely visual cues to the 3-D structure of the scene from others, such as proprioception. b) Data from an experiment [18] in which participants matched the size of a cube visible when the room was large with one visible only when the room was small. The correct matched size ratio is 1 while a size ratio of 4 is predicted if participants judged the cube size relative to the room or texture elements. Data for 5 participants is shown (solid symbols) and for a simulation using a 3-D reconstruction package (open symbols, see text). The curve shows the predictions of a cue combination model. The dotted line shows the predicted responses if participants matched the retinal size of the cubes. Figure reproduced, with permission, from [18] © Elsevier.

Figure 1b also shows the predicted responses from a 3-D reconstruction package (*Boujou* [20]) supplied with the same images that a participant would see as they carried out the task. In this case, we provided one veridical baseline measure, which sets the scale of the reconstruction, so the height of the simulation data on the y-axis is not informative. The important point, though, is that there is no variation of predicted matching size with viewing distance for the 3-D reconstruction, unlike the pattern of the human data.

In some ways, it is not difficult to explain the human observers' behaviour. The fitted curve in Figure 1b illustrates the predictions of a cue combination model. Matches for a distant comparison object are close to that predicted by a texture-based strategy (judging the cube size relative to the bricks, for example), since here stereo and motion parallax cues to distance are less reliable and hence are given a lower weight [21]. At close distances, stereo and motion parallax are more reliable (they support lower depth discrimination thresholds when compared to texture cues) and hence have a greater effect on matched sizes [22,18,21]. However, combining cues in this way is very specific to the task. It does not imply that there should be related distortions in other tasks, for example those requiring the comparison of two simultaneously visible objects.

The difficulty with a task-specific explanation, such as the cue combination model in Figure 1b, is that it avoids the question of scene representation in the human visual system. It is all very well to say that the world can act as 'an outside memory', and be accessed by looking at the relevant parts of the scene when necessary [23,17]. There must be a representation of some form to allow the observer to turn their head and eyes or walk to the right place to find the information. The representation may not be a 3-D reconstruction, but it must nevertheless be structured. There is as yet no satisfactory hypothesis about what that representation might be but the proposal in this paper is that it may have similarities to 'view-' or 'aspect-graphs', which are introduced in the next section.

1.2 Moving between Views

An alternative to 3-D reconstruction has been described in relation to navigation, in which a robot or animal stores views or 'snapshots' of a scene and records something about the motor output required to move between one view and the next, without integrating this information into a Cartesian map. The snapshots are the nodes in a 'view graph' and the movements are the edges [24,25] (see Section 3). One possible explanation of people's perceptions in the expanding room is that, as they move across a room, observers generally have an expectation of the images that they will receive: the fact that they do not notice that they are in an expanding room implies that they are relatively insensitive to the rate at which those images arrive as they walk.

This is illustrated in Figure 2a, where the grey plane represents a manifold of all the images that could be obtained by a person walking around the room. Each point on the plane represents one image and neighbouring points represent images visible from neighbouring vantage points. The dotted line shows, schematically, the set of images that a monocular observer might receive as they walk from the left to the right of the room. Potentially, that eye could receive exactly the same set of images whether the room was static or expanding: there is no way to tell from the monocular images alone. However, as Figure 2a shows, the rate at which new images arrive for each step the observer takes is different in the expanding room. On the left, where the room is small, new images arrive rapidly, whereas on the right they do so more slowly. Proprioceptive signals from the muscles provide information about the distance walked and these should be sufficient to tell observers about the size of the room. But in the expanding room, observers seem to disregard this information when it conflicts with the assumption that the room is stable, at least in determining their subjective experience of the room.

One attractive aspect of the view graph idea is its close similarity to known biological mechanisms. Moving from image to image or, more generally, from sensory state to sensory state, with an expectation of the sensory input that will be received as a consequence of each movement, is a familiar biological operation associated with well established circuitry, notably in the cerebellum [26,27,28]. The same cannot be said of operations required for general 3-D coordinate

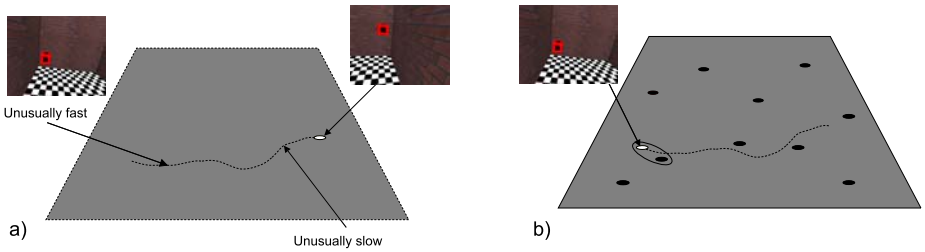

Fig. 2. Possible images in an expanding room. a) The grey plane symbolises a manifold of views of the room shown in figure 1. Each point corresponds to a single view and neighbouring points correspond to views from neighbouring vantage points. The current view is indicated by the white dot. The dashed line indicates the set of views that a hypothetical observer receives as they walk across the room. The manifold of views for the expanding room is identical to that for a static room: the only difference is that when the observer is on the right hand side of the room, where it is large, views change at an unusually slow rate as the observer walks while on the left side of the room, where the room is small, the views change unusually fast. b) The grey plane now symbolises a manifold of potential sensory (or sensory+motivational) states with the current state indicated by the white dot. The state includes visual and proprioceptive input. The black dots indicate stored states. Both the current and the stored states can be described by points in the same high dimensional space.

transformations, particularly translations of the origin. There are no detailed suggestions about how such operations might be achieved in the brain.

On the other hand, Figure 2b illustrates operations that are known to take place in the brain and it looks very similar to Figure 2a. Now the grey plane illustrates a high dimensional space of potential sensory (and motivational) states. Each black dot represents a potential sensory+motivational context that will lead to a particular motor output. The white dot represents the current sensory+motivational context, which is recognised as most similar to one of the stored contexts. The match leads to an action, whose consequence is a new sensory+motivational context, and so the movement progresses.

From this perspective, one can suggest a potential explanation for the fact that observers see the expanding room as static. Even though the proprioceptive signals are different in the two cases, the powerful visual feedback is exactly what is expected in a static room and hence, overall, the sensory context is sufficiently similar for it to be recognised as the expected context rather than an alternative one. In terms of Figure 2b, the path through sensory+motivational space as observers walk across the room is very similar, despite the different proprioceptive inputs, and hence their subjective experience is too.

It is not always true that the proprioceptive input is ignored. When observers move from a small room to a clearly separate large room their size judgements are much better than in the expanding room [18]. In that case, when confronted with a new room of unknown size, there is no strong expectation about the sensory feedback they will receive and so the proprioceptive information becomes

decisive. Even in the expanding room, proprioceptive (and stereo) input contributes to some judgements while not affecting observers' subjective impression of the room size [18,19,21]. Overall, the results in the expanding room suggest a representation that is looser, less explicit and more task-dependent than a 3-D reconstruction.

2 Biological Input and Output of Optic Flow Processing

For photogrammetry, the goal is clear: compute the 3-D structure of the scene and the path of the camera through it. Corresponding features are identified across frames, then the most likely 3-D scene, camera movement and camera intrinsic parameters (such as focal length) are computed given the image motion of the features. The camera motion is described in the same 3-D coordinate frame as the location of the points. The process can be written as:

$$\mathbf{x}_j^i = P^i \mathbf{X}_j \tag{1}$$

where $\mathbf{x}$ gives the image coordinates of points in the input frames, $\mathbf{X}$ describes the 3-D locations of the points in the scene and P is the projection matrix of the camera. P includes the intrinsic parameters of the camera, which remain constant, and the extrinsic parameters (camera pose) which are the only things that change from frame to frame. Equation 1 applies to the jth point in the ith frame.

In computer vision applications, the camera is generally free to translate and rotate in any direction. This is not true in biology. The 6 degrees of freedom of camera motion are essentially reduced to 3. Animals maintain fixation on an object as they move and are obliged to do so for at least 150 ms (this the minimum saccadic latency) [29]. Although the eye can rotate around 3 axes and the optic centre can translate freely in 3 dimensions, the rotation and translation of the eye are tightly coupled so that for each translation of the optic centre there is a compensatory rotation of the eye. Two dimensions of rotation maintain fixation while the third restricts the torsion of the eye with respect to the scene. This means that as a monocular observer moves through a static environment, maintaining fixation on an object, there is only one image that the eye can receive for each location of the optic centre.

A similar pattern of eye movements predominates throughout the animal kingdom and must clearly carry an evolutionary benefit. Land [30] has documented eye movement strategies in many species. Animals fixate while moving, then they make a 'saccade' or rapid rotation of the eyes and fixate a new target as they continue to move. They do this even when their eyes are fixed within the head, as is the case in many insects. In a rare condition in which the eye muscles become paralysed, patients' eyes are fixed with respect to their head [31]. In these cases, the head makes a rapid rotation between periods of fixation, so their visual diet is similar to that of a person with freely moving eyes. And when the eye and head are fixed with respect to the body, for example in a hover fly, the whole body makes saccades.

Why should animals do this? The constraint that eye rotation and translation are tightly linked does reduce the number of free parameters and so can help in the estimation of camera motion and scene reconstruction [32,33,34,35]. However, the rationale we describe here is different. If the goal is to navigate across a manifold of images, as we discussed in relation to Figure 2, then the restrictive pattern of eye movements that animals adopt makes sense. We illustrate this claim in the next sections by considering the role of the dorsal and ventral streams of visual processing during a simple set of movements.

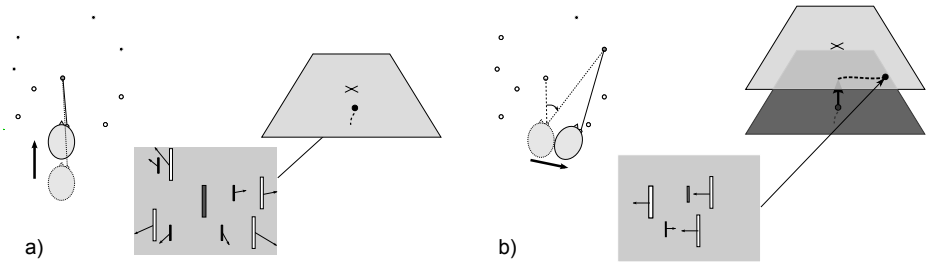

Fig. 3. Moving across a surface of images. a) An observer is shown in plan view moving towards a fixated object. The pattern of flow on the retina is illustrated and, on the right, the movement of the current image across a surface of images, similar to that shown in Figure 2a. b) The observer makes a saccade to a new fixation point which is illustrated by the current image jumping to a new surface of images (upper surface). The observer then moves laterally while maintaining fixation on the same object. Objects with a crossed disparity (open symbols) move in one direction in the image while those with an uncrossed disparity (filled symbols) move in the opposite direction. The cross marks the view that would be obtained if the viewer were at the location of the fixated object.

2.1 Dorsal Stream

Figure 3 shows a simple sequence of head and eye movements to illustrate a possible role of the dorsal stream in controlling head movements. In Figure 3a, the observer moves towards a fixated target. A fast, tight loop from retina to eye muscles keeps the eye fixating as the observer moves. This circuit can even bypass cortical motion processing [36]. The obligation to maintain fixation imposes a tight constraint on the type of image changes that can arise. The plane shown on the right in Figure 3a represents the set or manifold of images that can be obtained when the eye fixates a particular object. Each point on the manifold represents one image. As the observer approaches the fixated object, the current image (black dot) is shown moving towards the view that would be obtained if the observer were at the location of the object (shown by the cross). The retinal flow created by this movement is approximately radial expansion outwards from the fovea. There are many neurons in MSTd (the dorsal part of the medial superior temporal cortex, in the dorsal stream) that are sensitive to flow of this type [37,38].

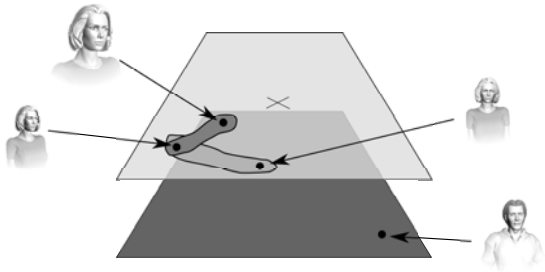

Fig. 4. Ventral stream neuron receptive fields. Neurons in the ventral stream respond selectively to certain objects or complex features while showing considerable invariance to size or viewing direction. The dark grey region illustrates a 'receptive field' of a size-invariant neuron, i.e. the set of images to which it might respond. The lighter grey 'receptive field' shows the set of images to which a view-invariant neuron might respond. As in Figure 3, each surface of images comprises the images that can be obtained by the observer while they fixate on a particular object – in this case a woman and a man.

Figure 3b illustrates in the same way a saccade followed by a lateral movement of the observer. Fixating a new object moves the current image to a new surface of images. The observer then moves laterally, staying the same distance from the fixation point. This produces a quite different pattern of retinal flow in which objects that are closer than the fixation point (shown as open symbols) move one way on the retina while more distant objects move in the opposite direction (closed symbols). The distinction between objects in front of and behind the fixation point can be signalled by disparity (crossed versus uncrossed) and indeed this seems to be a trick the visual system uses to disambiguate the flow. The same area that contains neurons responsive to expansion, MSTd, also contains neurons responsive to one direction of motion when the stimulus disparity is crossed and the *opposite* direction of motion when the stimulus disparity is uncrossed [39]. These neurons are ideally suited to signalling the type of flow that occurs when the observer moves his or her head laterally while fixating an object.

The two components of observer motion shown in Figure 3a and b can be detected independently. The neurons sensitive to forward motion can be used as a signal of progress towards the goal in Figure 3a with the neurons sensitive to lateral motion signalling error; the role of these types of neurons can then be reversed to control the lateral motion shown in Figure 3b. Greater calibration would be required to move on a path between these extremes, but the principle of using these two components remains [40] and the simplicity of the control strategy relies on the restriction that the observer fixates on a point as he or she moves. This is a quite different hypothesis about the role of dorsal stream neurons from the idea that they contribute to a computation of scene structure and observer heading in the same coordinate frame [38].

2.2 Ventral Stream

The ventral stream of visual processing has a complementary role. Neurons in this part of the visual pathway provide, as far as possible, a constant signal despite the observer's head movements, indicating *which* object the observer is looking at (i.e. which surface the current image is on) in contrast to the dorsal stream which signals how the observer is moving in relation to the fixated object, independent of the identity of that object.

As illustrated in Figure 4, it is possible to show on the surface of images the 'receptive fields' of different types of ventral stream neurons. Rather than a receptive field on the retina, here we mean the set of images to which this neuron would respond. The dark grey patch on the left side of Figure 4 shows a hypothetical 'receptive field' for a size-independent cell [41]: it would respond to an object from a range of viewing distances. The overlapping lighter grey patch shows the receptive field of a view-independent cell [42], which responds to the same object seen from a range of different angles.

Combining both of these types of invariance, one could generate a receptive field that covered the entire surface of images. In other words the neuron would respond to the view of a particular object from a wide range of angles and distances. Neurons with this behaviour have been reported in the hippocampus of primates [43]. The hippocampus, which lies at the end of the ventral stream, contains a large auto-association network [44] which has the effect of providing a constant output despite moderate changes in the input. Thus, there are mechanisms throughout the ventral stream that encourage a stable output to be maintained for the period of fixation. In terms of the surface of images, the argument here is that the ventral stream is designed to determine which surface of images contains the current image and the dorsal stream is designed to determine how the current image is moving across it.

So far, we have only considered the example of a short sequence of head and eye movements. In section 3, we discuss view graphs and how these can provide an extended spatial representation of observer location.

3 Representing Observer Location

A view graph (or aspect graph) representation consists of discrete reference views which form the nodes of the graph. In some simple examples, these have been the views at junctions in a maze [45], while in more realistic cases they have been the views at various locations in an open environment [46]. The edges of the graph are the movements that take the observer from one view to the next. A view graph is not a continuous representation of space (in contrast to the coordinate frame of $\mathbf{X}$ in Equation 1) but, when the observer is not at a location from which a reference view was taken, it is still possible to determine which of the reference views is most similar to the current view. This allows space to be divided up into regions, as the example below illustrates.

Figure 5 illustrates this principle for a 2-dimensional world. Suppose the room contained 3 objects as shown and that a 'view' consisted of a list of the angles

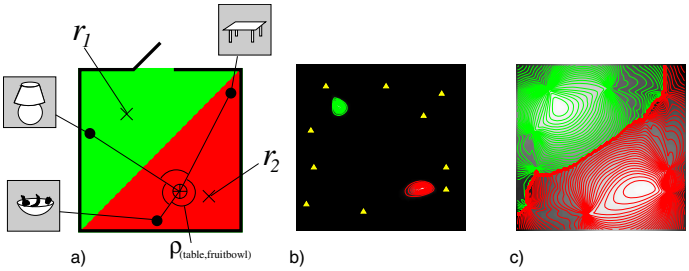

Fig. 5. Reference locations. a) Two reference locations r_1 and r_2 are shown and three objects, with the angle subtended between pairs of objects shown, e.g. $\rho_{(\text{table,fruitbowl})}$. The green and red areas indicate a possible division of the room into locations where the best answer to the question 'Where am I?' would be r_1 and r_2 respectively. b) Plan view of a scene in which the location of objects are marked by yellow triangles. Contour plots show the likelihoods that the observer is at the two candidate locations, colour coded according to which of the two likelihoods is greater. c) Same as b) but showing log likelihoods.

subtended at the optic centre by each pair of objects. Figure 5a shows two reference locations, r_1 and r_2, and the angles subtended between each of the three pairings of objects at a third, arbitrary location. The colours show a hypothetical division of the room into regions where the view is most like that at r_1 (green) and most like the view at r_2 (red). Figure 5b and c illustrate a specific example of this principle. There are now more than three objects, whose locations are shown by yellow triangles, but still two reference locations. The plots show the likelihoods that the view at each location is a noisy version of the view r_1 or r_2, computed as follows.

Scene features at positions (x_p, y_p), $p = 1, \ldots, N$, are imaged from viewpoint (x, y), and represented by the visual angles:

$$\boldsymbol{\rho}_{p,q}(x, y) = \big(\rho_{1,q}(x, y), \ldots, \rho_{p,q}(x, y), \ldots, \rho_{N,q}(x, y)\big), \quad q = 1, \ldots, N, \tag{2}$$

where $\rho_{p,q}$ is the angle between points (x_p, y_p) and (x_q, y_q). In order to avoid the use of a distinguished reference point, we compute all N^2 possible angles from each viewpoint. As well as general views $\boldsymbol{\rho}(x, y)$, we have R different *reference views*,

$$\big\{\boldsymbol{\rho}_{p,q}(x_1, y_1), \ldots, \boldsymbol{\rho}_{p,q}(x_r, y_r), \ldots, \boldsymbol{\rho}_{p,q}(x_R, y_R)\big\}, \quad q = 1, \ldots, N, \tag{3}$$

taken from distinct locations (x_r, y_r). Each reference view $\boldsymbol{\rho}(x_r, y_r)$ is accompanied by a corresponding list of variances, $\boldsymbol{\sigma}^2(x_r, y_r)$:

$$\boldsymbol{\rho}_{p,q}(x_r, y_r) = \big(\rho_{1,q}(x_r, y_r), \ldots, \rho_{p,q}(x_r, y_r), \ldots, \rho_{N,q}(x_r, y_r)\big), \quad q = 1, \ldots, N, \tag{4}$$

$$\boldsymbol{\sigma}^2_{p,q}(x_r, y_r) = \big(\sigma^2_{1,q}(x_r, y_r), \ldots, \sigma^2_{p,q}(x_r, y_r), \ldots, \sigma^2_{N,q}(x_r, y_r)\big), \quad q = 1, \ldots, N. \tag{5}$$

In our simulations, we take $\sigma^2_{p,q} = 1$ in all cases. The fit of the rth reference view to the visual angles obtained at observer position (x, y) is defined as the squared difference between $\boldsymbol{\rho}(x, y)$ and $\boldsymbol{\rho}(x_r, y_r)$, summed over all angles,

$$E_r(x, y) = \sum_{q=1}^{N} \sum_{p=1}^{N} \frac{1}{\sigma^2_{p,q}(x_r, y_r)} \big(\rho_{p,q}(x, y) - \rho_{p,q}(x_r, y_r)\big)^2. \tag{6}$$

We use E_r to compute the likelihood of the current view $\boldsymbol{\rho}(x, y)$, under the hypothesis that the viewpoint coincides with that of model view $\boldsymbol{\rho}(x_r, y_r)$. Specifically, we represent the likelihood as

$$\mathsf{L}\big(\boldsymbol{\rho}(x, y) \,|\, \boldsymbol{\rho}(x_r, y_r)\big) = e^{-E_r(x,y)}, \tag{7}$$

which is proportional to the probability of the rth location-hypothesis:

$$\mathsf{P}\big(x = x_r, y = y_r \,|\, \boldsymbol{\rho}(x, y)\big) \propto e^{-E_r(x,y)}. \tag{8}$$

The normalizing constant is obtained by integrating P over all viewpoints, (x, y). The figures plot the maximum likelihood at each point (x, y), colour-coded by the reference location r for which $\mathsf{L}\big(\boldsymbol{\rho}(x, y) \,|\, \boldsymbol{\rho}(x_r, y_r)\big)$ is maximum.

If the visual system represents places using angles subtended between objects in a similar way, then fixating on an object as the observer moves is a sensible strategy. It allows changes in angles (at least, those with respect to the fixated object) to be monitored easily as the observer moves. Figure 6 illustrates the idea. If the optic centre, O, of the eye/camera is at the location marked by the blue cross and the fixation point is F, then the eccentricity of a point, P, imaged on peripheral retinal is the same as the angle (ρ) between that point and the fixation point subtended at the optic centre. If the observer maintains fixation on the same point, then change in eccentricity of the peripheral object signals change in the angle ρ, Figure 6c. This retinal flow, $\Delta\rho$, is directly relevant to computing changes in likelihood that the current location is r_1 or r_2, as Figures 6b and d illustrate.

The example in Figures 5 and 6 is limited to control of movement between two reference locations. Figure 7 shows a larger set of reference locations chosen by a robot, equipped with an omnidirectional camera, at which it took 'snapshots' or reference views as it explored a real scene [47]. New snapshots were constrained to differ (above a criterion level) from those already stored. In general, views are connected as the nodes in a view graph, where the edges connect 'neighbouring' views. New edges were formed by the robot as it explored: as it left one reference location it compared its current view to all other 'snapshots' it had stored and then used a homing strategy [48,49] to approach the reference location with the most similar view. The edges in the view graph contained no information about the direction or distance between two vertices, only that they were neighbours. This experiment illustrates how the idea of reference views can be extended over a large (potentially limitless) spatial range. Equally, of course, the resolution of the representation can be increased within a particular region of space by including more reference locations there. One can imagine many situations in

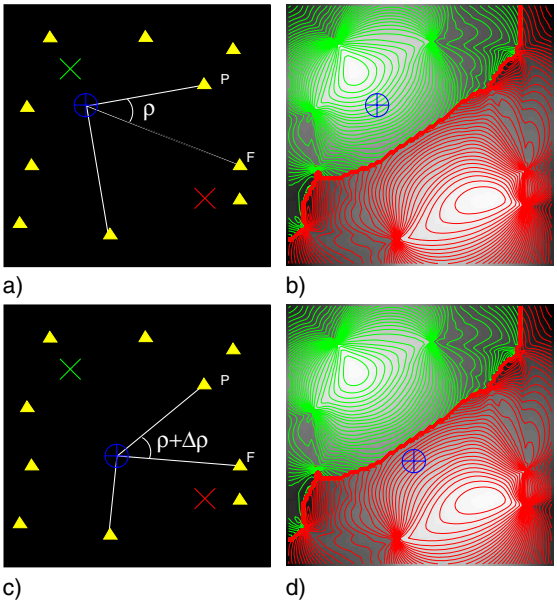

Fig. 6. Retinal flow and fixation. a) The observer's location is marked by the blue target and the fixated object is shown by a yellow triangle, F. Two other objects imaged on peripheral retina are shown, one of which (P) is at an eccentricity of ρ. Movement of the observer causes a change in the angles between pairs of objects subtended at the optic centre (c and d). Because the observer maintains fixation as he or she moves, retinal flow provides a straight-forward record of the changes in these angles with respect to the fixation point. For example, as shown in c), the change in the angle ρ to $\rho + \Delta\rho$ results in flow, $\Delta\rho$, at the point on the retina where the object P is imaged. The change in ρ (and other angles) with respect to the fixation point can be used to determine whether the current view is becoming less like that from reference location r_1 and more like that from r_2 (b and d).

which fine distinctions between different observer locations are important in one part of a scene but less important in others.

View graphs, then, provide a way of extending the ideas about control of observer location to a wider region of space. Section 2 dealt only with observer movement relative to the fixation point, and saccades. Now these can be seen as examples of moving over a small part of an infinitely extendible view graph. However, view graphs do not directly answer the question raised in section 1 about scene representation. For example, in Figure 7, how might the location of the objects (rather than the robot) be represented? The grey object appears in many views, including the snapshots from all the reference locations shown by filled symbols. Is there a sensible coordinate frame in which to unite all the information about an object's location, other than a 3-D world-based frame? Current implementations do not do so and it is not immediately obvious how it should be done.

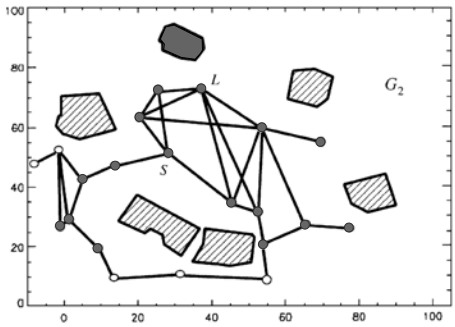

Fig. 7. View graph of a scene. Circles show the locations of reference views (or 'snapshots'), which are the nodes in the view graph (adapted from a figure in Franz et al [47], with permission). The filled circles indicate the reference views that include an image of the grey object. It is not clear how best to define the location of objects in the view graph.

One might argue that in simple animals the distinction does not exist: for an ant to know where an object is located is much the same as knowing how to arrive at it. The same may be true for humans if an object is a long way away (like the Eiffel tower). In that case, the represented location of an object can be a node in the view graph. However, this only avoids the question of uniting estimates of object location across views, which is relevant at closer distances for actions such as reaching or pointing. Humans can point to an object that is not currently visible and can do so in a way that accounts for their rotations *and* translations since they last saw it. They can also direct their hand to an object that they have seen in peripheral vision but never fixated, even when they make a saccade before pointing at it (although they make systematic error in this case [50]). These capacities, however poorly executed, are evidence of a representation of object location that is maintained across head and eye movements. View graphs, or some similar view-based approach, must clearly be able to support similar behaviour if they are to become candidate models of human spatial representation.

4 Conclusion

We have argued that the human perception of space in an expanding virtual room may be understood in terms of a representation like a view graph, or a manifold of images, where the observer has an expectation of the images they will receive as they move across the room but are insensitive to the discrepancy between visual and proprioceptive feedback. It is clear that view graph models are currently inadequate in many ways but we have argued that an explanation along these lines is more likely to explain human perception in the expanding room than one based on 3-D reconstruction. It is clear from the expanding room

and many other perceptual phenomena that the development and successful implementation of non-Cartesian, non-metric representations will be of great relevance to the challenge of understanding human 3-D vision.

Acknowledgements

Funded by the Wellcome Trust. We are grateful to Bruce Cumming, Andrew Parker and Hanspeter Mallot for helpful discussions.

References

1. Rogers, B.J., Graham, M.: Similarities between motion parallax and stereopsis in human depth perception. Vision Research 22, 261–270 (1982)
2. Bradshaw, M.F., Rogers, B.J.: The interaction of binocular disparity and motion parallax in the computation of depth. Vision Research 36, 3457–3768 (1996)
3. Bradshaw, M.F., Parton, A.D., Eagle, R.A.: The interaction of binocular disparity and motion parallax in deptermining perceived depth and perceived size. Perception 27, 1317–1331 (1998)
4. Bradshaw, M.F., Parton, A.D., Glennerster, A.: The task-dependent use of binocular disparity and motion parallax information. Vision Research 40, 3725–3734 (2000)
5. Fitzgibbon, A.W., Zisserman, A.: Automatic camera recovery for closed or open image sequences. In: Burkhardt, H.-J., Neumann, B. (eds.) ECCV 1998. LNCS, vol. 1406, pp. 311–326. Springer, Heidelberg (1998)
6. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge University Press, Cambridge (2000)
7. Foley, J.M.: Binocular distance perception. Psychological Review 87, 411–433 (1980)
8. Gogel, W.C.: A theory of phenomenal geometry and its applications. Perception and Psychophysics 48, 105–123 (1990)
9. Johnston, E.B.: Systematic distortions of shape from stereopsis. Vision Research 31, 1351–1360 (1991)
10. Tittle, J.S., Todd, J.T., Perotti, V.J., Norman, J.F.: A hierarchical analysis of alternative representations in the perception of 3-D structure from motion and stereopsis. J. Exp. Psych.: Human Perception and Performance 21, 663–678 (1995)
11. Glennerster, A., Rogers, B.J., Bradshaw, M.F.: Stereoscopic depth constancy depends on the subject's task. Vision Research 36, 3441–3456 (1996)
12. Basri, R., Rivlin, E., Shimshoni, I.: Visual homing: Surfing on the epipoles. International Journal of Computer Vision 33, 117–137 (1999)
13. Davison, A.J.: Real-time simultaneous localisation and mapping with a single camera. In: Proceedings. Ninth IEEE International Conference on computer vision, pp. 1403–1410 (2003)
14. Newman, P., Ho, K.L.: SLAM-loop closing with visually salient features. In: Proceedings IEEE International Conference on Robotics and Automation, pp. 635–642 (2005)
15. Gibson, J.J.: The ecological approach to visual perception. Houghton Mifflin, Boston (1979)

16. Ullman, S.: Against direct perception. Behavioural and Brain Sciences 3, 373–415 (1980)
17. O'Regan, J.K., Noë, A.: A sensori-motor account of vision and visual consciousness. Behavioural and Brain Sciences 24, 939–1031 (2001)
18. Glennerster, A., Tcheang, L., Gilson, S.J., Fitzgibbon, A.W., Parker, A.J.: Humans ignore motion and stereo cues in favour of a fictional stable world. Current Biology 16, 428–443 (2006)
19. Rauschecker, A.M., Solomon, S.G., Glennerster, A.: Stereo and motion parallax cues in human 3d vision: Can they vanish without trace? Journal of Vision 6, 1471–1485 (2006)
20. 2d3 Ltd. Boujou 2 (2003), http://www.2d3.com
21. Svarverud, E., Gilson, S.J., Glennerster, A.: Absolute and relative cues for location investigated using immersive virtual reality. In: Vision Sciences Society, Naples, Fl (2008)
22. Ernst, M.O., Banks, M.S.: Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415, 429–433 (2002)
23. O'Regan, J.K.: Solving the real mysteries of visual perception: The world as an outside memory. Canadian Journal of Psychology 46, 461–468 (1992)
24. Schölkopf, B., Mallot, H.A.: View–based cognitive mapping and path planning. Adaptive Behavior 3, 311–348 (1995)
25. Koenderink, J.J., van Doorn, A.J.: The internal representation of solid shape with respect to vision. Biological Cybernetics 32, 211–216 (1979)
26. Marr, D.: A theory of cerebellar cortex. J. Physiol (Lond.) 202, 437–470 (1969)
27. Albus, J.: A theory of cerebellar function. Mathematical Biosciences 10, 25–61 (1971)
28. Miall, R.C., Weir, D.J., Wolpert, D.M., Stein, J.F.: Is the cerebellum a Smith predictor? Journal of Motor Behaviour 25, 203–216 (1993)
29. Carpenter, R.H.S.: Movements of the eyes. Pion, London (1988)
30. Land, M.F.: Why animals move their eyes. Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology 185, 1432–1351 (1999)
31. Gilchrist, I.D., Brown, V., Findlay, J.M.: Saccades without eye movements. Nature 390, 130–131 (1997)
32. Aloimonos, Y., Weiss, I., Bandopadhay, A.: Active vision. In: Proceedings of the International Conference on Computer Vision, London, UK, June 8–11, pp. 35–54 (1987)
33. Bandopadhay, A., Ballard, D.: Egomotion perception using visual tracking. Computational Intelligence 7, 39–47 (1990)
34. Sandini, G., Tistarelli, M.: Active tracking strategy for monocular depth inference over multiple frames. IEEE Transactions on Pattern Analysis and Machine Intelligence 12, 13–27 (1990)
35. Daniilidis, K.: Fixation simplifies 3D motion estimation. Computer Vision and Image Understanding 68, 158–169 (1997)
36. Cohen, B., Reisine, H., Yokota, J.-I., Raphan, T.: The nucleus of the optic tract: Its function in gaze stabilization and control of visual-vestibular interaction. Annals of the New York Academy of Sciences 656, 277–296 (1992)
37. Saito, H., Yukie, M., Tanaka, K., Hikosaka, K., Fukada, Y., Iwai, E.: Integration of direction signals of image motion in the superior temporal sulcus of the macaque monkey. J. Neuroscience 6, 145–157 (1986)
38. Perrone, J.A., Stone, L.S.: A model of self-motion estimation within primate extrastriate visual cortex. Vision Research 34, 2917–2938 (1994)

39. Roy, J.P., Wurtz, R.H.: The role of disparity-sensitive cortical neurons in signalling the direction of self-motion. Nature 348, 160–162 (1990)
40. Glennerster, A., Hansard, M.E., Fitzgibbon, A.W.: Fixation could simplify, not complicate, the interpretation of retinal flow. Vision Research 41, 815–834 (2001)
41. Rolls, E.T., Bayliss, G.C.: Size and contrast have only small effects on the responses to faces of neurons in the cortex of the superior temporal sulcus of the monkey. Experimental Brain Research 65, 38–48 (1986)
42. Booth, M.C.A., Rolls, E.T.: View-invariant representations of familiar objects by neurons in the inferior temporal cortex. Cerebral Cortex 8, 510–525 (1998)
43. Georges-Francois, P., Rolls, E.T., Robertson, R.G.: Spatial view cells in the primate hippocampus: allocentric view not head direction or eye position or place. Cerebral Cortex 9, 197–212 (1999)
44. Treves, A., Rolls, E.T.: Computational analysis of the role of the hippocampus in memory. Hippocampus 4, 374–391 (2004)
45. Gillner, S., Mallot, H.A.: Navigation and acquisition of spatial knowledge in a virtual maze. Journal of Cognitive Neuroscience 10, 445–463 (1998)
46. Franz, M.O., Mallot, H.A.: Biomimetic robot navigation. Robotics and Autonomous Systems 30, 133–153 (2000)
47. Franz, M.O., Schölkopf, B., Mallot, H.A., Bülthoff, H.H.: Learning view graphs for robot navigation. Autonomous Robots 5, 111–125 (1998)
48. Cartwright, B.A., Collett, T.S.: Landmark learning in bees: experiments and models. Journal of Comparative Physiology 151, 521–543 (1983)
49. Hong, J., Tan, X., Pinette, B., Weiss, R., Riseman, E.: Image-based homing. IEEE Control Systems Magazine 12(1), 38–45 (1992)
50. Henriques, D.Y.P., Klier, E.M., Smith, M.A., Lowy, D., Crawford, J.D.: Gaze-centered remapping of remembered visual space in an open-loop pointing task. Journal of Neuroscience 18, 1583–1594 (1998)