Telling where one is heading and where things move independently

Niels da Vitoria Lobo and John K. Tsotsos*
Dept. of Computer Science, University of Toronto, Toronto, Canada M5S 1A4.
e-mail: niels@vis.toronto.edu, tsotsos@vis.toronto.edu

Abstract

We summarize our recent novel approach to computing the Focus of Expansion for an observer moving with unrestricted motion in a scene with objects of unrestricted shape. This method also detects points not moving rigidly with the scene. The approach, using collinear image points, is based on an exact method for cancelling effects of the observer's rotation from optic flow. The computational results are being presented elsewhere (da Vitoria Lobo & Tsotsos 1991). Here, we argue that this algorithm is biologically plausible.

Introduction

Many years ago researchers (Helmholtz 1925, Gibson 1957) hypothesized that the 3-D motion and the shape of the environment are perceivable from the projected motion arising out of the relative motion between a monocular observer and the scene. In this paper, we summarize our recent computational solution to this problem and discuss its biological plausibility.

The paradigm we work within assumes that an approximation to instantaneous image velocity (also termed *image flow* or *optic flow*) can be measured, and some progress has been made towards achieving such measurements (see Watson & Ahumada 1985, Heeger 1988, Fleet 1990). The subsequent step, that of computing the 3-D motion parameters and shape information from the instantaneous image velocity, has received ample attention from researchers. However, in addition to the fact that every 3-D algorithm proposed so far is not robust to noise in the input instantaneous velocity, the algorithms typically suffer from other important problems that disqualify them from biological candidacy. Algorithms that permit general rigid motion and unrestricted shapes have to search in spaces that have at least three dimensions (Bruss & Horn 1983), and the non-linear numerical methods used are very sensitive to the initial guess. Others assume some restricted form of motion (eg., no rotation in a certain dimension; see Barron 1988), or restrict the allowable shapes (eg., planarity), in order to get closed-form solutions for the unknowns (Waxman & Wohn 1987), or assume that some parameters are known and solve for the others (Ballard & Kimball 1983, Matthies *et al.* 1989) — all of them too restrictive for biological plausibility.

The basis of our approach is a technique for combining collinear image points which allows rotation to be cancelled in an exact manner. Along any straight line in the image, the rotational contribution to the image velocity component orthogonal to the line varies linearly with length, so that taking approximations to the second derivative of this component of velocity cancels out rotation. Thus despite the unrestricted motion and unrestricted shape involved in the problem, the motion parameters can be unlocked by a search for the correct direction of translation, which is a mere 2-dimensional search and incurs a far lower computational cost than other algorithms that search in higher dimensions. The algorithm, termed the *FOE Algorithm*, uses an operator that simultaneously cancels out rotation exactly and samples the translation contribution to find the direction of translation. Earlier, da Vitoria Lobo & Tsotsos (1990) showed that for three non-collinear image points, the pair-wise relative depths of the three scene points are dependent only on the unknown 3-D direction of translation and the known image velocities and image positions (i.e., that, in principle, knowledge of rotation is irrelevant to the calculation of shape from motion). For other work that cancels rotation see Prazdny (1983), Nelson & Aloimonos (1988) and Heeger & Jepson (1990). Our approach also straightforwardly detects points that do not move in a manner consistent with the assumption of a rigid scene. In this paper we argue that due to its simplicity, low computational cost, inherent parallelism, and robustness to noise, this method is biologically plausible, and we explore its consequences for research in biological systems. We first review our *FOE algorithm* and its extension to detecting independent motion. Then we discuss the importance of the FOE and present the arguments for biological plausibility, along with predictions that result from the model.

Locating the FOE and independent motion

In this section, we summarize work appearing in da Vitoria Lobo & Tsotsos (1991).

Image velocity and scene parameters

When the relative motion between the viewer and a scene point (at depth Z) is described by an instantaneous translation (U, V, W), and an instantaneous rotation (A, B, C), the projected velocity in the image plane at position (x, y) is (u, v), (Longuet-Higgins & Prazdny 1980), where

$$\begin{aligned} u &= \frac{-U+xW}{Z} - Axy + B(x^2+1) - Cy, \\ v &= \frac{-V+yW}{Z} - A(y^2+1) + Bxy + Cx. \end{aligned} \tag{1}$$

The Focus of Expansion

We define the Focus of Expansion (FOE) to be the position of intersection of the imaging surface[1] and the direction of

*John K. Tsotsos is the Canadian Pacific Fellow of the Canadian Institute for Advanced Research. This work was supported by the Natural Sciences and Engineering Research Council and the Information Technology Research Center, a Province of Ontario Center of Excellence. The range data came from the Range Image Database of NRC Canada.

[1] This surface could be planar, hemi-spherical, etc.

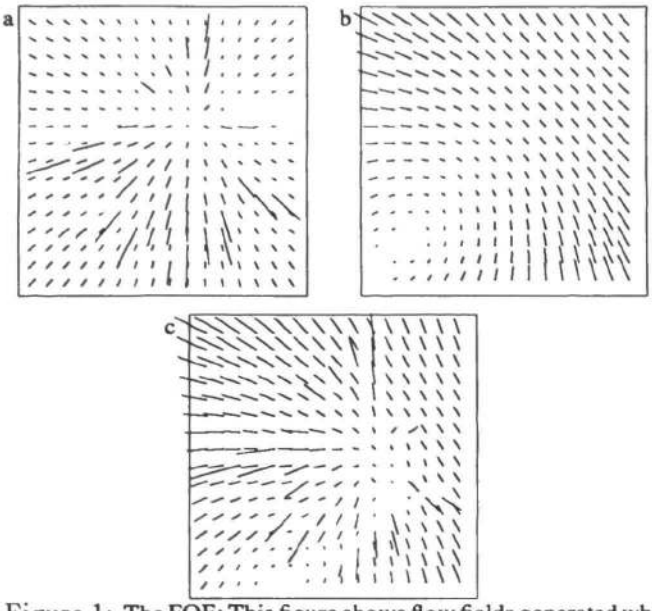

Figure 1: The FOE: This figure shows flow fields generated when observer a) only translates, b) only rotates, and c) moves with both rotation and translation, the most typical situation. In the second case there is no FOE because there is no observer translation, while in the first and third cases the FOE is in the same position for both cases, just above and to the right of the image center. That is, our definition of FOE renders its position in the image completely independent of observer rotation.

instantaneous observer translation. This[2] defines the FOE to be independent of the viewer's instantaneous rotation. Figure 1 illustrates how our definition of FOE makes its position invariant to rotation.

Several authors have researched the computation of the FOE either by restricting the allowable motion to translation or by approximately cancelling rotation(Jain 1982, Reiger & Lawton 1985). In work related to ours, Weinshall (1990) finds the FOE but needs to find elliptical surface patches beforehand. The contribution of the work in da Vitoria Lobo & Tsotsos (1991) is that we find the FOE in an exact manner, even when the motion includes general rotation, with no restriction on surface shape.

FOE From Collinear Measurements

We refer to three image points as a *triplet*. The unit computation of our algorithm is a collinear triplet computation. The computation is a generalization of an approximation to the second derivative of the velocity component that is normal to the line joining the three collinear image points, the "derivative" being taken in the direction of the line. For points subscripted by i, $(i = 1, 2, 3)$, let the image coordinates of the points be (x_i, y_i), their image velocities be (u_i, v_i), and the depths to their counterparts in the scene be Z_i.

[2]This definition differs from that of Regan & Beverley (1982) and Cutting (1986) who define the FOE as the intersection of the image and the direction of motion, thus making it dependent on rotation.

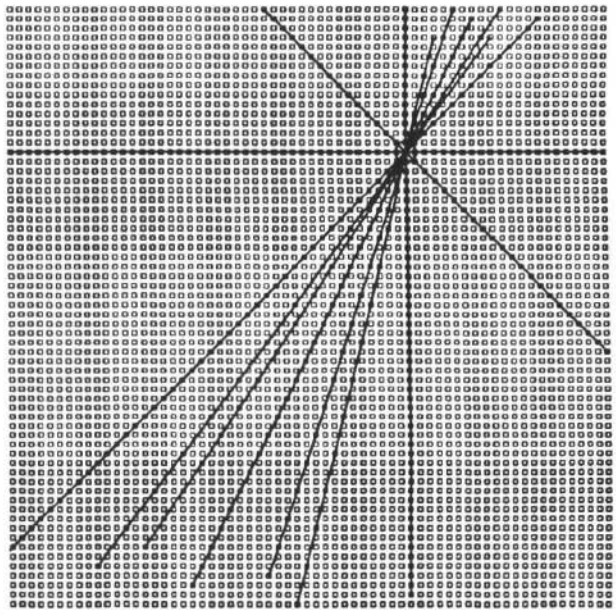

Figure 2: The FOE Operator: This shows how an operator is made up of intersecting lines of points in the image. On a regular 256x256 square optic array, at a single point we can have about 60-70 intersecting lines each of which passes through at least several image points. Here, only 9 lines are shown. The image is a dense grid of points, shown as hollow squares, each with a flow estimate associated with it. Some of the points used by each line have been blackened to identify them. Along each line a *Line Sum* is calculated (see Fig 3). The Line Sums are added to give the response of the operator at position (x,y) which is where the lines intersect.

The computation is the weighted[3] sum[4],

$$Sum \stackrel{\text{def}}{=} \begin{array}{l}(-\sin\theta)(nu_1 - (m+n)u_2 + mu_3) + \\ (\cos\theta)(nv_1 - (m+n)v_2 + mv_3)\end{array} \quad (2)$$

where θ is the angle that the line through the three image points makes with the image x-axis, going from the x-axis to the line in a counterclockwise manner, and where m is the distance between the first and second image points and n the distance between the second and third. By substitution of (u_i, v_i), this generalized *Sum* can be verified to be

$$Sum = \begin{pmatrix} U\sin\theta - V\cos\theta & + \\ Wy_3\cos\theta - Wx_3\sin\theta & \end{pmatrix}\left(\frac{n}{Z_1} - \frac{(n+m)}{Z_2} + \frac{m}{Z_3}\right)$$

i.e., *Sum* = product of translation factor and depth factor.

Sum is zero either if the scene points are collinear, or the translation factor is zero, or if the line passing through the three image points also passes through the FOE. Thus, in general, for a scene containing sufficient depth variation, if we compute many such collinear triplet sums across the complete image, the FOE will be in the position of intersection

[3]When the collinear points are equi-spaced, the weights are (1,-2,1) which is an approximation to the second derivative (Rektorys 1969).

[4]We use the symbol $\stackrel{\text{def}}{=}$ to define a calculation; this is to distinguish it from an expression of equality.

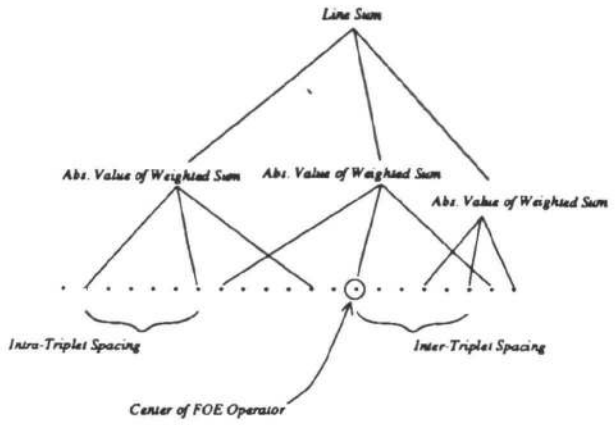

Figure 3: This shows how points on a line were processed. Flow estimates at points are grouped into triplets and summed according to the weighted sum of equation (2). Then the absolute values of these weighted sums are summed to give the *Line Sum* for the line.

of many triplet lines for which, in each case, the triplet sum is zero. Next, we describe an operator and an algorithm to locate the FOE.

The FOE Operator

Consider an operator (called the *FOE operator*) with many lines passing through its center (centered at some (x,y) position in the image) in many directions in the image (see figure 2), such that each line passes through several image points. Along the line, overlapping triplets of image velocity are used and summed (see figure 3).

The absolute values of all triplet sums (as defined by equation 2) along a line are summed to give a *LineSum*, and then the LineSums are summed to get a response at the center. For a rigid scene, there are three reasons why the response could be zero. First, each of the triplets summed by this operator could be a collinear triplet in the scene; however, since each line contains overlapping triplets, for the zero response to be caused by collinear scene points, the whole scene would have to be a single plane, which is quite rare. Second, there could be no translation. Third, the true FOE could be at the position of the center of the operator. Fortunately, if either of the first two cases were to hold, the operator would respond with zero everywhere; whereas, in the third case, assuming the first two do not hold, there will be a unique zero. So, the first two cases can be detected easily and eliminated. Thus, we can sweep this operator across the whole flow field to obtain a response map, detecting where it gives a zero response surrounded by sufficiently non-zero responses.

Testing the FOE computation

Here tests of the FOE algorithm are described. The range image shown in Fig 4 was used to generate the synthetic optic flow field shown in Fig 5. This flow was used as input to our FOE algorithm. For the implementation of the FOE algorithm, an FOE operator was centered at each element in the optic array. Each FOE operator consisted of 24 lines passing through it (see Fig 2). Along each line we centered

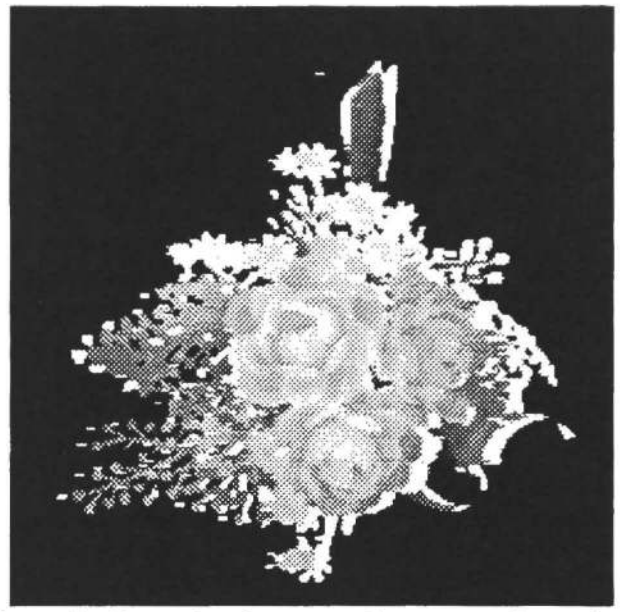

Figure 4: Range data that was used to generate the synthetic flow field used in these experiments. Observer moves with some instantaneous 3-D rotation and translation parameters with respect to these depth points and the flow equations are used to give an image velocity vector at each grid point on the optic array.

a triplet at each point on that line. Within each triplet, the intra-triplet spacing was such that three alternate points on the line were used.

Fig 6 shows the response map for the operator as a function of (x,y). Here, response maps are shown with brightness proportional to the *Log* (response). The darkest point, the global minimum, corresponds to the computed FOE and it is not surprising that it is exactly correct. Even with noise (as high as 8% on average) in the flow, the FOE is found easily. With noise, the pit of the minumum broadens out but is still pronounced (somewhat like Fig 8.)

Functioning in a non-rigid scene

Objects in a moving observer's view will often move independently. Our computation remains competent in these situations. Fig 7 depicts the flow field generated by combining the original rigid flow field used in the previous section with the flow for an independently moving rectangular patch in the upper middle of the image. The frontoparallel patch translates upward and to the right, with an image flow magnitude comparable to those of the flower petals. The FOE algorithm was tested on this combined flow field. The output direction of translation is still correct (filter response map in Fig 8.) Tests indicate that even with larger patches the FOE computation is accurate, demonstrating robustness to patches of non-rigidity in the scene.

Finding points that thwart rigidity

A biological agent needs the ability to detect parts of the

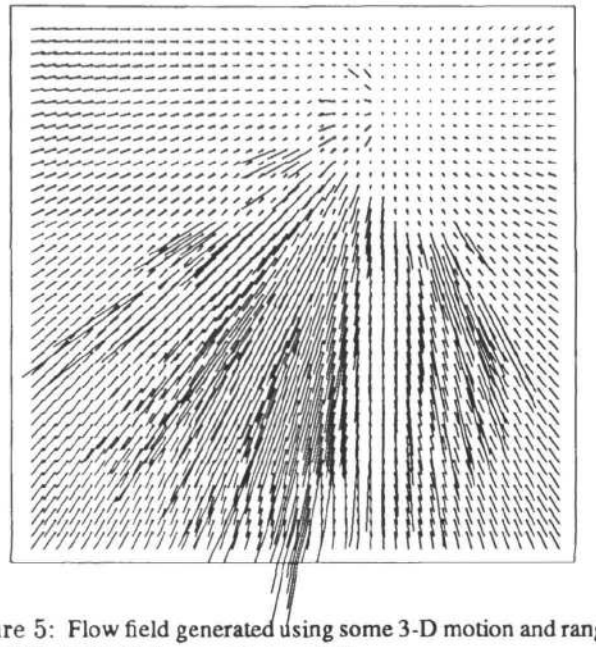

Figure 5: Flow field generated using some 3-D motion and range data of Fig 4. Field shown sub-sampled.

Figure 6: Response map for noise-free flow. In these maps, we show brightness proportional to *Log*(response). For this input, the global minimum is exactly at the true FOE.

scene where the overall-rigidity assumption does not hold. A point may be measured as moving differently from the rest of the scene either because it may be a noisy measurement – in which case any subsequent use of this erroneous flow value such as in the shape reconstruction step, should be approached with caution – or because it could legitimately belong to an independently moving part of the scene. If this is the case, this moving part will probably need additional attention and possibly some special purpose processing, such

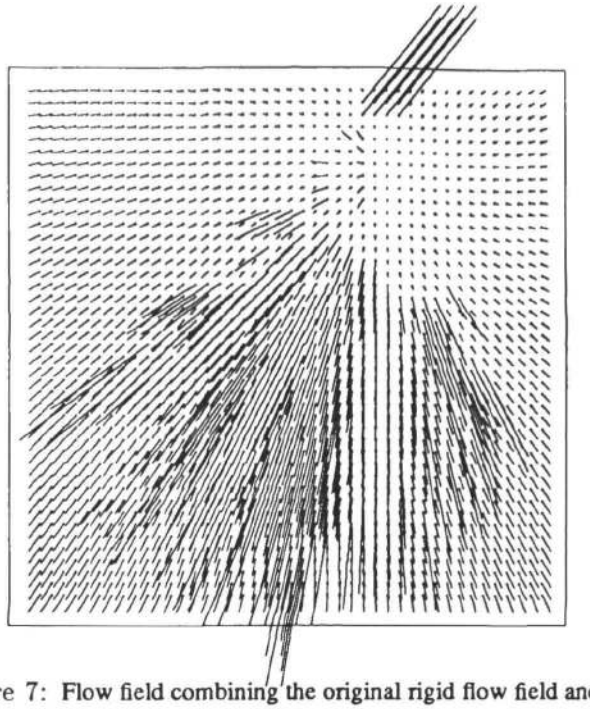

Figure 7: Flow field combining the original rigid flow field and an independently moving patch in upper right part of image. The frontoparallel rectangular patch translates upward, to the right, with an image flow magnitude that is comparable to those of the flower petals.

as that which will segment it from its background.

Our algorithm can easily signal that certain points are moving inconsistently relative to the rest of the scene. We achieve this by using the operator that gave the minimum response indicating that the FOE is at its center. We traverse each of its various lines searching for triplets that don't give near-zero sums. Such triplets correspond to points such that at least one of the three points moves inconsistently with the rest of the scene[5]. Fig 9 shows the region around the patch in the flow field of Fig 7 being marked by the program as sets of inconsistent triplets. Note that because the patch itself is moving in a planar fashion, the triplet sum, when all three points are inside the patch, is zero. So, at an independently moving planar region, only the boundary areas of the region will be detected (this depends on the intra-triplet spacing used). For an independently moving non-planar region, all the triplets, in any way overlapping the region's points, will be detected.

Importance of the FOE

As mentioned earlier, we have already shown in principle that relative depth is a function of the direction of translation and is independent of rotation. In other work, we are proposing that for the noisy cases shape information be recovered in stages, involving qualitative and quantitative recovery based on knowledge of the FOE. The qualitative recovery work has already appeared in Weinshall (1990), and in da Vitoria Lobo & Tsotsos (1990).

[5]We have already stated earlier in this section that if such a collinear triplet were moving with the same rigid parameters as the scene, the triplet *Sum* must be zero.

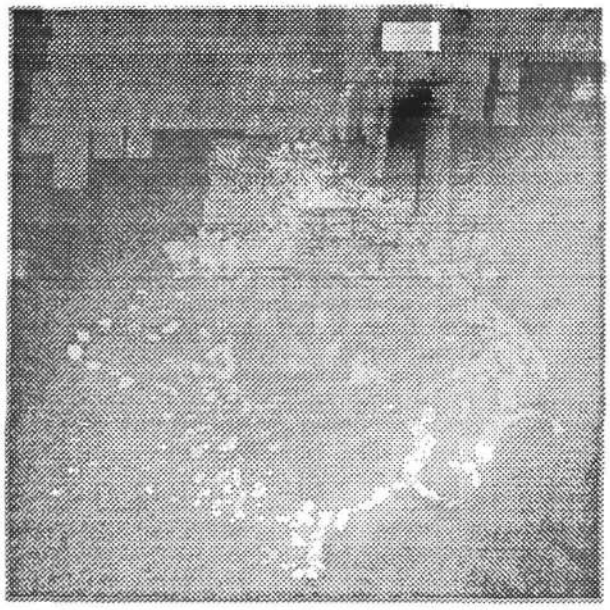

Figure 8: Response map for the flow associated with the nonrigid scene in which a patch moves independently. The global minimum is in the same position, indicating that the FOE computation is robust to non-rigidity in the scene.

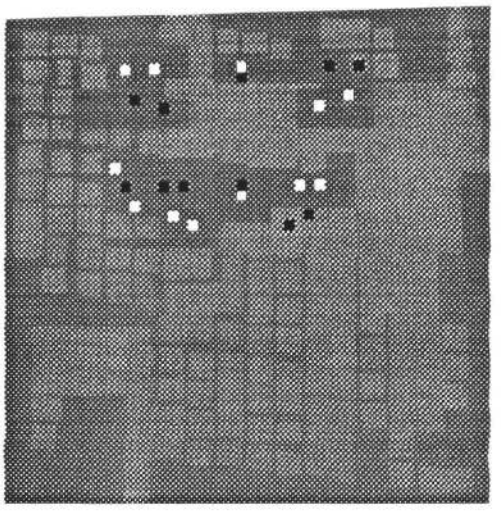

Figure 9: An enlarged portion of the output from the program that detects triplets containing inconsistently moving points, when it was run on the flow of Fig 7. The detected points correspond to areas around the independently moving patch. Because the patch itself is moving in a planar fashion, the triplet sum, when all three points are inside the patch, is zero. So, these internal points are not marked.

In addition to its use in computing shape, the FOE tells where one is heading. This could be useful for navigation tasks. Also, since our computation is a monocular one, a binocular system could compute an FOE for each eye seperately, and this could be used to obtain information about the relative poses of the two eyes.

Relation to biological systems

We suggest that the FOE Algorithm summarized above is biologically plausible because our simple calculations can easily be implemented by the visual cortex. Consider Fig 10. On the left we show one of the motion pathways from Maunsell & Newsome (1987). On the right we show the stages we use to compute our FOE operator response map, suggesting points in the cortical pathway at which our computations could be implemented. Cells in area V1 could compute normal velocities for moving intensity structure, and then connect to area MST either directly or via area MT. (Maunsell & Van Essen 1983, Ungerleider & Desimone 1986).

Tanaka *et al.* (1989) have described cells in macaque dorsal MST that appear to be responsive to patterns composed of points moving outwardly or inwardly along radial lines. These cells have been termed "changing size" or "expansion/contraction" cells. A family of such cells, similar to each other, could be computing in parallel an inverse of our response map, with each cell responding strongly when the FOE is at its center. That is, each cell would be encoding a different direction of translation. If such a cell were computing a sum of our *LineSums*, then a cell that responds to a strictly expanding/contracting radial flow should continue its response even when a rotational field of the kind shown in Fig 1b is added to the expanding/contracting pattern. If this were to be found, this would be very strong evidence for a computation similar to our algorithm.

To compute our FOE operator response at nearby positions, neighboring MST cells would receive input from overlapping triplet *Sums*. Hence it would be reasonable to expect that these *Sums* are not being re-computed each time, but rather that some intermediate cells compute something akin to triplet *Sums*. There are cells in area MT that are known to respond to patterns of activity in which the center differs from its surround. This would be an appropriate substrate in which our triplet *Sums* could be computed. These could appear in the form of elongated cells computing an approximation to the second derivative of the normal velocities, the derivative taken along the long axis. An analogue of such receptive fields has been proposed by Dobbins *et al.* (1987) for curvature detection. To reduce connectivity to the MST cell computing the response for the map, the flow field may be sampled quite sparsely, and we need to study the computational effects of such sparse sampling.

If the response map is being computed in parallel using a family of MST cells, then some subsequent mechanism would need to find the global minimum. This could possibly be accomplished using computations at multiple scales across the map, so that a coarse sampling of the map would indicate the ballpark of the minimum, while further finer grain spatial sampling would give better resolution of the position of the FOE. A framework involving attention would be suitable for achieving this (Tsotsos 1990).

Detecting independently moving points requires top-down activation of the particular MST cell positioned at the FOE. We hypothesize that such feedback exists under attentional control, but its exact nature would need to be discovered.

Finally, there are connections from area MST to areas 7a

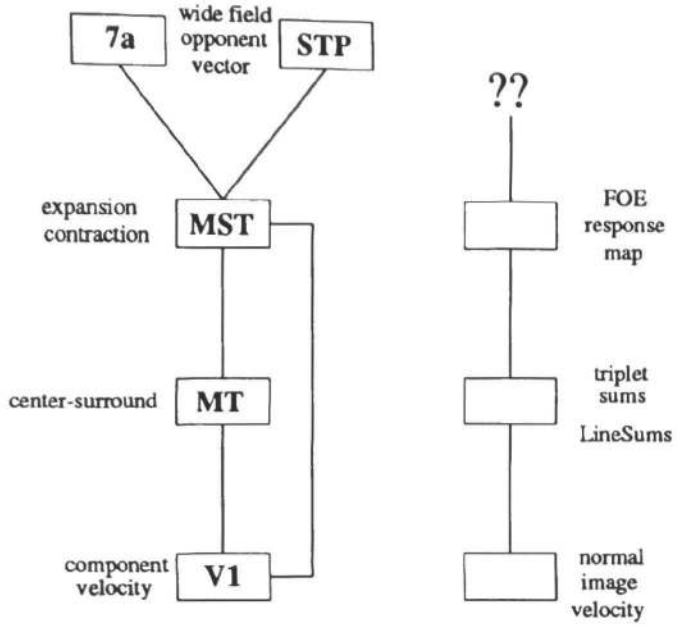

Figure 10: Is the FOE algorithm biologically plausible? On the left, an abstraction of one of the motion pathways described in Maunsell & Newsome (1987). On the right, stages leading up to the computation of our FOE operator response map.

and STP where wide field, "opponent vector" organizations have been found. These cells respond to patterns of radial flow to and from the fixation point (Maunsell & Van Essen 1987). We are studying the possible roles these would play in our framework.

Conclusion

We have summarized a recent novel approach to computing the direction of the translation component of egomotion, and detecting points not moving rigidly with the scene, in the view of an observer moving with unrestricted motion. The detailed computational work appears elsewhere, but here we argued that this approach is biologically plausible and sketched some of its consequences.

References

Ballard D.H. and Kimball O.A. (1983), "Rigid body motion from depth and optical flow," *CVGIP*, **22**, 95-115.

Barron J. (1988), "Determination of egomotion and environmental layout from noisy time-varying image velocity information in monocular image sequences," *Ph.D. Thesis*, Dept. of Computer Science, Univ. of Toronto. Available as RBCV-TR-88-24.

Bruss A.R. and Horn B.K.P. (1983), "Passive navigation," *CVGIP*, **21**, 3-20.

Cutting J. (1986), **Perception with an Eye for Motion**, MIT Press.

da Vitoria Lobo N. and Tsotsos J.K. (1990), "Extracting qualitative shape from image motion: applications to stereo-pairs", *Proc. AAAI Workshop on Qualitative Vision*, pp. 36-40, Boston, July.

da Vitoria Lobo N. and Tsotsos J.K. (1991), "Using collinear points to compute egomotion and detect nonrigidity," *in press, Proc. of Computer Vision and Pattern Recognition*, Hawaii, June 1991.

Dobbins A., Zucker S.W., and Cynader M.S. (1987), "Endstopping in the visual cortex as a substrate for calculating curvature," *Nature*, **329**, 438-441.

Fleet D. (1990), "The Measurement of Image Velocity", *Ph.D. Thesis*, Dept. of Computer Science, Univ. of Toronto.

Gibson J.J. (1957), "Optical motions and transformations as stimuli for visual perception", *Psychological Review*, **64**, No 5, 288-295.

Heeger D. (1988), "Optical flow using spatiotemporal filters", *IJCV*, **1**, 279-302.

Heeger D. and Jepson A. (1990), "Simple method for computing 3-D motion and depth," *Proc. ICCV*, Osaka, Japan, 96-100.

Helmholtz H.(1925), **Treatise on Physiological Optics**, Optical Society of America.

Jain (1982), "An approach for the direct computation of the focus of expansion," *Proc. PRIP-82*, 262-268.

Longuet-Higgins H.C. and Prazdny K. (1980), "The interpretation of a moving retinal image," *Proc. Roy. Soc. Lond.*, **B 208**, 385-397.

Matthies L., Szeliski R., Kanade T. (1989), "Kalman filter-based algorithms for estimating depth from image sequences," *IJCV*, **3**, 181-208.

Maunsell J.H.R., and Newsome W.T. (1987), "Visual processing in monkey extrastriate cortex," *Ann. Rev. Neurosci.*, **10**, 363-401.

Maunsell J.H.R., and Van Essen D.C. (1983), "The connections of the middle temporal visual area (MT) and their relationship to a cortical hierarchy in the macaque monkey," *Jour. of Neurosci.*, **3**, 2563-2586.

Nelson R. and Aloimonos J. (1988), "Using flow field divergence for obstacle avoidance: towards qualitative vision," *Proc. ICCV* Florida, 188-196.

Prazdny K. (1983), "On the information in optical flows", *CVGIP*, **22**(2), 239-259.

Regan D.M. and Beverley (1982), "How do we avoid confounding the direction we are looking and the direction we are moving?" *Science*, **215**, 194-196.

ReigerJ.H. and Lawton D.T. (1985), "Processing differential image motion", *J. Optical Society of America*, **A2**, 354-359.

Rektorys K. (1969), **Survey of Applicable Math**, MIT Press.

Tanaka K., Fukada Y., Saito H. (1989), "Underlying mechanisms of the response specificity of expansion/contraction and rotation cells in the dorsal part of the Medial Superior Temporal area of the macaque monkey," *Journ. of Neurophysiology*, **62**, No. 3, 642-656.

Tsotsos J.K. (1990), "Analyzing Vision at the Complexity Level," *Behavioral and Brain Sciences*,.

Ungerleider L.G., and Desimone R. (1986), "Cortical connections of area MT in the macaque", *J. Comp. Neurol.*, **248**, 190-222.

Watson A. and Ahumada J. (1985) "Model of human visual motion sensing," *J. Optical Society of America*, **A2**, 322-342.

Waxman A.M. and Wohn K. (1987), "Image flow theory: a framework for 3-D inference from time-varying imagery," Chapter Three in **Advances in Computer Vision**, ed. C. Brown, (Erlbaum Publishers).

Weinshall D. (1990), "Direct computation of qualitative 3-D shape and motion invariants," *ICCV*, Osaka, Japan, 230-237.