International Journal of Computer Vision, 11:2, 165–186 (1993)

The Role of Fixation in Visual Motion Analysis

CORNELIA FERMÜLLER AND YIANNIS ALOIMONOS

Computer Vision Laboratory, Center for Automation Research, Computer Science Department, and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742-3411; and (Fermüller) *Department for Pattern Recognition and Image Processing, Institute for Automation, Technical University of Vienna, Treitlstraße, A-1040 Vienna, Austria*

Received October 2, 1992; Revised May 21, 1993.

Abstract

How does the ability of humans and primates to fixate at environmental points in the presence of relative motion help their visual systems in solving various tasks? To state the question in a more formal setting, we investigate in this article the following problem: Suppose that we have an active vision system, that is, a camera resting on a platform and being controlled through motors by a computer that has access to the images sensed by the camera in real time. The platform can move freely in the environment. If this machine can fixate on targets being in relative motion with it, can it solve visual tasks in an efficient and robust manner? By restricting our attention to a set of navigational tasks, we find that such an active observer can solve the problems of 3-D motion estimation, egomotion recovery, and estimation of time-to-contact in a very efficient manner, using as input the spatiotemporal derivatives of the image-intensity function (or normal flow). Fixation over time changes the input (motion field) in a controlled way and from this change additional information is derived making the previously mentioned tasks easier to solve.

1 Introduction

Visual navigation problems in the past were mostly studied for the case of a passive observer. In order to demonstrate the computational advantages underlying the perceptual capabilities of an active observer capable of controlling its gaze, we first give a short overview of visual motion techniques for passive observers.

Most studies devoted to visual motion analysis are dominated by the computational approach of Marr (1982). The goal is to recover from a sequence of images the 3-D motion parameters and the structure of the objects in view. The suggested strategies attempt to solve the problem in two steps (Ullman 1979). First, accurate image displacements between consecutive image frames have to be computed, either in the form of correspondence of features or as dense motion fields (optic flow fields). In case the motion between image frames is relatively "large" image features are isolated and tracked through a sequence of frames. This amounts to solving the so-called correspondence problem. Otherwise, in case of "small" motion, the dynamic imagery is considered as a three-dimensional function of two spatial and one temporal variables. And from the spatiotemporal derivatives and some additional information derived by making assumptions about this function, the velocity in the image plane or optical flow, is computed. In a second step the 3-D motion and the structure of the scene is computed from the equations relating the 2-D image velocity to the 3-D parameters. This step is usually called the "structure-from-motion" problem.

The plethora of mathematical models and computational techniques that have been employed in the study of visual motion interpretation can be classified in a variety of ways (according to whether correspondence or optic flow are used as input, on the basis of the type of features used for correspondence or the choice of the image projection model, whether the proposed solutions involve iterative methods or are in closed form, etc.). Here we provide a brief historical account of research in the motion-estimation problem for a passive

observer. Following the two-step approach, the two problems of computing correspondence or optical flow and the estimation of structure from motion have been studied in parallel. In the evolution of the study of the structure-from-motion problem three phases can be distinguished. First, work dealt with the question of the existence of a solution, that is, can we extract any information from a sequence of images about the structure and the 3-D motion of the scene that cannot be found from a single image? Several theoretical results have appeared that deal with questions such as: what can be recovered from a certain number of feature points in a given number of frames (Ullman 1979; Aloimonos & Brown 1989) under either orthographic or perspective projection. Then, the uniqueness aspects of the problem were studied. Nonlinear algorithms for the recovery of structure and motion from point or line correspondences and optic flow appeared increasingly in the literature. Algorithms dealing with correspondence were based on iterative approximation techniques. So, they lacked guaranteed convergence. Later, "linear" algorithms were developed and closed-form solutions introduced (Tsai & Huang 1984; Spetsakis & Aloimonos 1990; Adiv 1985) that allowed proofs of uniqueness. Although research in these lines has been accompanied by many experiments, none of the existing techniques could be used as a basis for an integrated system working robustly in general environments.

The reasons for the lack of applicability to real-world problems are due to the difficulty of estimating retinal correspondence, which is an ill-posed problem; the assumptions that have to be made to derive optical flow; and the sensitivity of 3-D motion estimation to small changes in the input data, that is, image motion. Even optimal algorithms (Spetsakis & Aloimonos 1992)—optimal under the assumption of Gaussian noise—perform quite poorly in the presence of moderate noise. As a result, research on motion analysis has shifted its focus to issues of robustness, with most researchers using redundant information. Algorithms that employ long sequences of image frames have been developed. These techniques, however, still require the correspondence of features and in most studies many assumptions about the continuity of motion and the structure of the scene were employed (with the notable exception of (Spetsakis & Aloimonos 1991)).

Efforts to remove these and many other shortcomings related to the perception of motion and structure, contributed to the development of direct methods for the interpretation of visual motion and to the birth of a new concept, active vision (Aloimonos et al. 1988; Bajcsy 1985; Ballard 1991). Since the computation of optical flow or correspondence is probably an ill-posed problem, it would make sense to circumvent this computational step. The only image-motion representation that is generally well defined is the image-motion component perpendicular to grey-level edges, the so-called normal flow. Algorithms that seek visual motion interpretation on the basis of normal flow belong to the category of direct methods that are gaining increasingly a lot of importance, and this article presents a set of ideas along these lines.

On the other hand, use of normal flow for a passive observer does not in any way alter the local constraints relating image motion to 3-D motion and structure of a moving object which are nonlinear and quite sensitive to small perturbations.[1] The situation is, however, different for the case of an active observer. An observer is active when he has the capability to control the geometric parameters of his sensory apparatus, that is, when he can control the image acquisition process. Aloimonos, Weiss, and Bandopadhay (1988) show that an observer with the ability to control self-motion can solve several recovery problems in a more efficient manner than can a passive observer. One of the reasons for this is that the activities of the observer (verging, fixating, zooming, tracking, focusing, etc.) introduce additional constraints that facilitate the interpretation of the visual signal. This is also the case for problems of visual motion analysis and the activity of fixation.

Here we present the computational advantages of fixation in space-time, usually referred to as gaze control. We show that an active observer with the ability to control its gaze and keep an environmental feature stationary on its image can solve several visual-motion interpretation tasks very efficiently, using well-defined input and spending little computational effort.

2 Overview

The theoretical importance of structure-from-motion can hardly be overemphasized. The reason is that if we are able to compute from a sequence of images the structure of the imaged scene and the relative three-dimensional motion, then subsets of the computed parameters provide sufficient information for the solu-

tion of several problems related to navigation, such as detection of independent motion, kinetic stabilization, obstacle avoidance, target pursuit, hand-eye coordination, automatic docking, etc. Indeed, if a robust solution to the structure-from-motion was available, we would use it in order to solve the above-mentioned problems. Consequently since it has turned out that there exist inherent difficulties in solving the general problem of structure-from-motion, it makes sense to seek direct solutions to the specific problems mentioned above that do not require complete recovery (Aloimonos 1990). In addition, if we are able to supply additional information to the processes solving these specific motion tasks, we may solve problems that were originally considered as ill-posed, ill-conditioned, and nonlinear. Additional information can be obtained by making the observer active and allowing him therefore to manipulate and control certain parameters. This is the approach called for by the paradigm of *active vision* (Bajcsy 1985; Aloimonos et al. 1988; Aloimonos 1990).

In their paper, Aloimonos et al. discuss solutions to some recovery problems for an active observer possessing controlled self-motion, but they consider optical flow as input to their modules. Here, by exploiting the advantages of gaze control, we develop solutions to the 3-D motion-estimation problem that do not rely on optic flow or correspondences but use as input the spatiotemporal derivatives of the image-intensity function.

From the measurements on the image we can only recover the relative motion between the observer and any point in the 3-D scene. The model that has mostly been employed in previous research to relate 2-D image measurements to 3-D motion and structure is the one of rigid motion. Consequently, the case of *egomotion recovery* for an observer moving in a static world has been treated in the same way as the *estimation of an object's 3-D motion* relative to an observer. We argue here that the rigid-motion model is the appropriate one if only the observer is moving, while this holds only for a restricted subset of moving objects—mainly man-made ones. Indeed, all objects in the natural world move nonrigidly. However, considering only a small patch in the image of a moving object, a rigid-motion approximation is legitimate.

Therefore, for the case of egomotion we can use data from all parts of the image plane, whereas for object motion we can only employ local information. Hence, we develop two conceptually different algorithms for explaining the mechanisms underlying the perceptual processes of *egomotion recovery* and *3-D object-motion recovery.*

In particular, we analyze the following two problems:

1. "Given an active observer viewing an object moving in a rigid manner (translation + rotation), recover the direction of the 3-D translation (Focus of Expansion: FOE) and the time to collision by using only the spatiotemporal derivatives of the image-intensity function" (sections 5 and 6). Although this problem is not equivalent to structure-from-motion, because it does not fully recover the 3-D motion, it is of importance in a variety of situations where response to object motion has to be generated and the translation of the moving object is the relevant factor. If an object is rotating around itself and also translating in some direction, we are usually interested in its translation—for example in problems related to tracking, prey catching, interception, obstacle avoidance, etc.
2. Given an active observer moving rigidly in a static environment, recover the direction of its translation and its rotation and determine the relative depth (sections 5 and 7). This is the process of passive navigation, a term used to describe the set of processes by which a system can estimate its motion with respect to the environment.[2]

3 The Input

Due to the aperture problem the only image-motion measurement that can in general be uniquely defined from a sequence of images is the normal flow, the component of the flow perpendicular to the edges (Horn & Schunck 1981). Normal flow can be computed in a variety of ways (Fleet & Jepson 1990; Singh 1990). The difficulty in its estimation is mainly due to the discrete nature of digital images. Computing normal flow in images is as difficult as detecting edges.

Assuming conservation of the image intensity, the normal flow is computed from the spatiotemporal derivatives of the intensity function $I(x, y, t)$ by employing the motion-constraint equation $I_x u + I_y v + I_t = 0$ (Horn & Schunck 1981), where subscripts denote partial differentiation and (u, v) the optic flow. The normal flow is then the projection of the flow on the gradient direction, that is, $(I_x, I_y) \cdot (u, v) = -I_t$. It is by now well established that the optic-flow field and the motion field are not equal in general (they are equal for the

case of a uniformly illuminated Lambertian surface undergoing pure translation) (Verri & Poggio 1989) and it is not clear whether optic flow could be used in quantitative studies of visual-motion problems. However, it can be shown that the normal flow is very close to the normal motion field in places of the image where the local intensity gradient ∇I is high (Singh 1990; Fermüller 1993a). Thus, if we measure normal flow only in regions where the intensity gradients are of high magnitude, we guarantee that the normal-flow measurements can be used for inferring 3-D motion.

Concerning the implementation of the algorithms for finding normal flow in the experiments conducted in this article, the images were first convolved with either a box filter of kernel size 3 to 5 or a Gaussian of the same kernel size and standard deviation in the order of $\sigma = 1.3 - 1.7$. The normal flow was computed by using 3×3 Sobel operators to estimate the spatial derivatives in the x and y directions and by subtracting the 3×3 box-filtered values of consecutive images to estimate the temporal derivatives. Because the direction of normal-flow vectors at corner points cannot be accurately estimated, a preprocessing step was utilized, where through a multiresolution technique (Fermüller & Kropatsch 1992) edge points of high curvature were detected and the normal flow was not estimated there.

4 Previous Research

This work is directed toward the recovery of 3-D motion using the activity of dynamic fixation and the image gradients (normal flow) as input. Thus, it is related to both active vision techniques and direct methods for the perception of visual motion.

The realization that the vision of biological systems is trivially active has initiated a lot of work on active computer vision (Pahlavan 1993; Swain & Ballard 1991; Swain & Stricker 1991). The idea of employing fixation and using the tracking parameters for motion estimation has been used by Aloimonos and Weiss (1988) and Bandopadhay and Ballard (1991), who provide a closed-form solution for the computation of the egomotion parameters for a binocular observer by employing the rotation angle and its first and second derivative (angular velocity and acceleration) along with values of the optic-flow field. Our contribution here lies in the development of a solution for an active monocular observer capable of fixating in space-time that uses normal flow as input, that is, the spatiotemporal derivatives of the image-intensity function. At the same time we explain in computational terms the advantages of dynamic fixation. When an observer fixates at a feature being in relative motion with that observer, visual motion data around the fixation point give rise to new geometric constraints that can be used to decipher the parameters of the three-dimensional motion.

On the other hand, the idea of using the image gradients to directly estimate 3-D motion without going through the intermediate stage of calculating the optic-flow field first appeared in the work of Aloimonos and Brown (1984). They presented a complete solution for the case of pure rotation, whereas a detailed study of translational motion can be found in (Horn & Weldon 1987; Negadharipour 1986; Negadharipour & Horn 1989; and Negadharipour & Ganesan 1992). Finally, a hybrid technique appeared recently (Taalebi-Nezhaad 1990), using optical flow, tracking, and image gradients for addressing 3-D motion estimation in the general case (rotation and translation). However various limiting assumptions about the depth of the scene in view had to be employed in that work. Our contribution here lies in the introduction of several novel geometric normal-flow-field properties due to rigid motion that give rise to simple pattern-matching techniques for recovering egomotion without using any assumptions about the shape of the scene.

Here, we demonstrate that the processes of the estimation of object motion and the estimation of an observer's motion—which were in the past addressed as the same problem—are perceptually different. We develop solutions to these problems for an active observer; in the former case the solution is based on fixation and tracking through the utilization of a small part of the input while in the latter case the solution is based on fixation in time and the analysis of global patterns of image motion. We show how an active observer through a qualitative analysis of normal-flow fields reduces the dimensionality of the motion-perception problem from five dimensions to one thus obtaining a fast and robust solution.

5 The Observer and the Choice of the Coordinate System

Figure 1 depicts a pictorial description of the active observer. Notice that the camera, controllable by a motor, is resting on a platform that can move rigidly. Figure 2 shows a geometric model of the camera. O

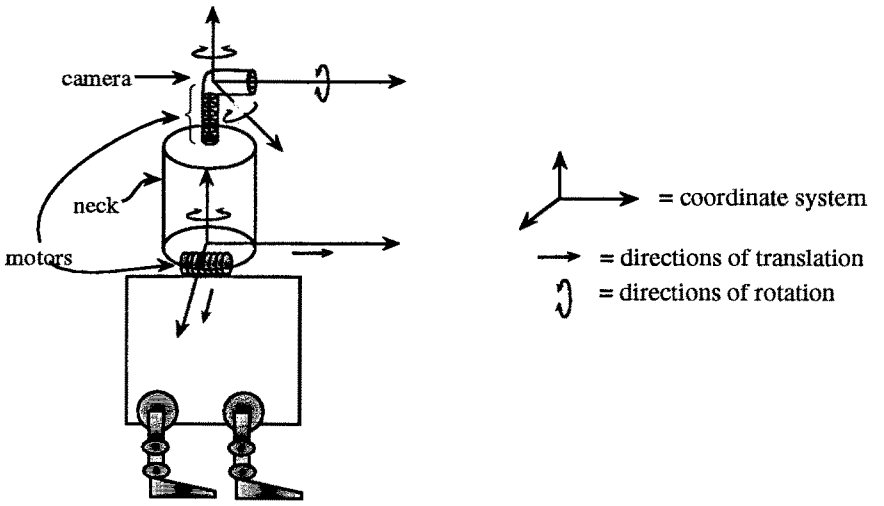


Fig. 1. The active observer.

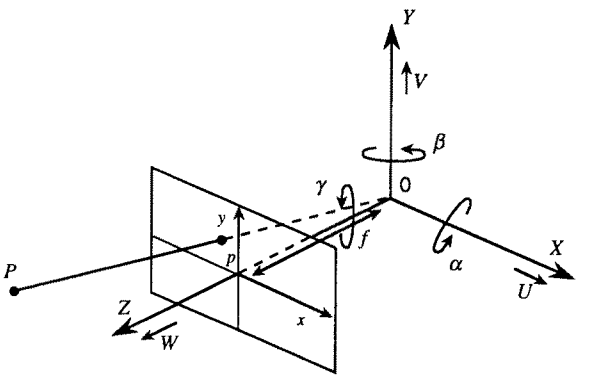


Fig. 2. Imaging geometry and motion representation (camera-centered).

denotes the nodal point of the eye and the image plane is perpendicular to the optical axis OZ at distance f (focal length) from the origin. The image is formed through perspective projection.

Since motion parameters are expressed relative to a coordinate system, prediction of the position of the moving entity (object or observer) at the next time instance is dependent on the choice of the coordinate system. In the case of egomotion, it makes sense to attach the coordinate system onto the observer, simply because the quantities recovered are directly related to the way the observer moves (figure 2). On the other hand, when the observer needs to make inferences regarding another object's motion, the ideal place to put the origin of the coordinate system would be the mass center of the object (the natural system).

Since the mass center is not known, different choices have to be made. Most commonly the camera's nodal point is chosen as the center of the coordinate system (*camera-centered* coordinate system). Rotation is described around the nodal point. In the case of object motion this leads to different values for the motion

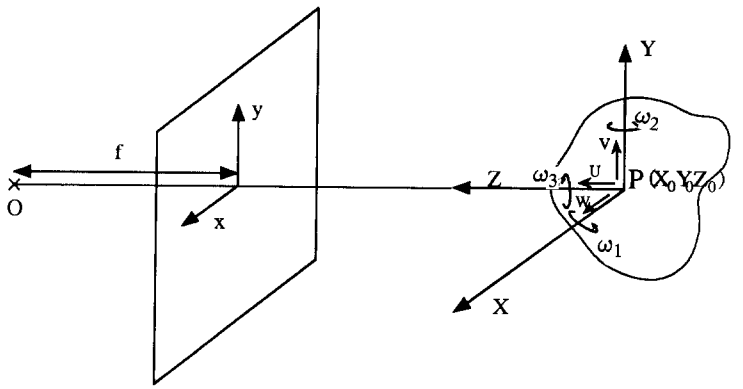


Fig. 3. Object-centered coordinate system.

parameters for each new frame, which is an unwelcome effect in the task of finding translational motion.

We therefore decided to attach the center of rotation to the object's point of intersection with the optical axis (an *object-centered* coordinate system) (see figure 3). The active observer is free in its choice of the center and will therefore decide for a point belonging to a neighborhood of nonuniform brightness with distinguishable features. This way the region under consideration has gradients distributed in various directions and this is essential for the success of the normal-flow-based tracking technique (section 6.1).

This choice is justified by the following argument: When choosing as a fixated point the mass center of the object's image or a point in its near neighborhood, the resulting motion parameters are in many cases close to those of the natural system. In the natural coordinate system with center O_n the velocity v at point P is due to the translational and the rotational component,

$$v = t_n + \omega \times \overrightarrow{O_nP}$$

and in the object-centered coordinate system with center O_o the same velocity is expressed as

$$v = t_o + \omega \times \overrightarrow{O_oP}$$

Therefore the difference in translation between t_n and t_o (see figure 4) is given by

$$\begin{aligned} t_n - t_o &= \omega \times (\overrightarrow{OoP} - \overrightarrow{OnP}) \\ &= \omega \times \overrightarrow{O_oO_n} \end{aligned}$$

This value becomes smaller as $\overrightarrow{O_oO_n}$ decreases.

Throughout this article, in order to stress the different analyses for different coordinate systems, rotation in an object-centered coordinate system is denoted by $(\omega_1, \omega_2, \omega_3)$ and rotation in a camera-centered coordinate system by (α, β, γ).

6 Active 3-D Motion Estimation

It has been argued that fixation and tracking are used in biological vision systems for the sake of simplifying the interpretation of motion. Since our goal is to study computer vision for an active observer, the first question we should ask concerns the nature of the activities employed. What is gained from fixation and tracking? The major advantage of tracking is the accumulation of information over time and therefore the introduction of the parameter of time as additional component to the input information. Another advantage of tracking is that since it is accomplished over a number of steps, the tracking parameters can be sequentially corrected. Thus, there is no need to rely on just one measurement.

The estimation of the translational direction and the time to collision are accomplished in two steps. First, after detecting independent object motion the active observer fixates at a point of the object. This point is considered to be the origin of the "object-centered" coordinate system throughout the analysis. Tracking is then used in order to obtain the projection on the image of the object's translation parallel to the image plane. Using this result and tracking, the observer accumulates depth information over time and acquires information about the translation of the object parallel to the optical axis. Thus the direction of translation (FOE) is estimated and finally the time to collision is obtained by using its relationship to the FOE and spatiotemporal information at the fixated point. The forthcoming nomenclature and analysis follows (Fermüller & Aloimonos 1992).

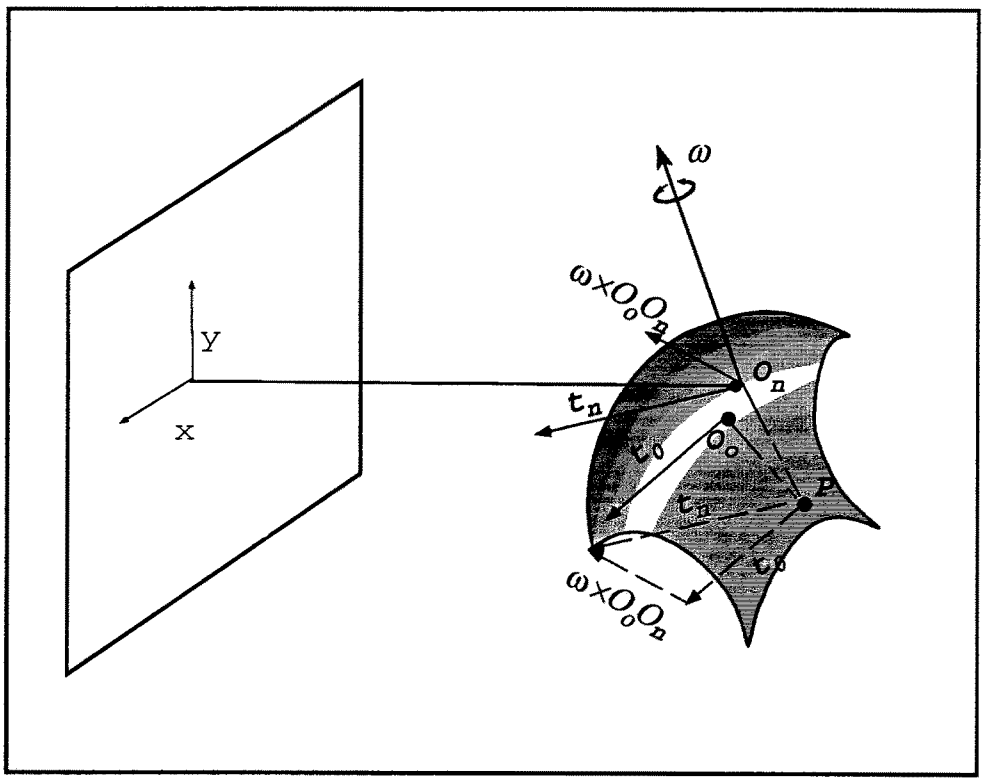


Fig. 4. The difference in translation between t_n in the natural system with center O_n and t_o in the object centered system with center O_o is $\omega \times \overrightarrow{O_o O_n}$.

6.1 Tracking Provides the Translation Parallel to the Image Plane

From a mathematical point of view, fixation in time provides linear relations between the 3-D and the 2-D velocity parameters. An object at distance Z in front of the camera moves in the 3-D environment with translational velocity (U, V, W) and rotational velocity $(\omega_1, \omega_2, \omega_3)$. In an object-centered coordinate system with center $P(X_0, Y_0, Z_0)$ under perspective projection the optical flow (u, v) is related to these parameters through the following equations:

$$\frac{dx}{dt} = u$$

$$= \frac{Uf}{Z} - \frac{Wx}{Z} - \frac{xy\omega_1}{f} + \omega_2\left(\frac{x^2}{f} + \frac{f(Z - Z_0)}{Z}\right) - \omega_3 y$$

$$\frac{dy}{dt} = v$$

$$= \frac{Vf}{Z} - \frac{Wy}{Z} - \omega_1\left(\frac{y^2}{f} + \frac{f(Z - Z_0)}{Z}\right) + \frac{\omega_2 xy}{f} + \omega_3 x$$

In a small area around the center x, y, and $(Z - Z_0)/Z$ are close to zero. The optical flow components due to rotation and due to translation parallel to the optical axis converge to zero, and u becomes Uf/Z and v becomes Vf/Z.

The above equations demonstrate the essence of the fixation-based approach to visual motion estimation. The flow at the center of the image is equal to the parallel translation (U, V) scaled by the depth Z. If we knew the flow at the image center, then subsequent processing would be greatly facilitated. In the sequel we show that fixation in time can be used for the evaluation of this optic-flow value in an iterative manner and

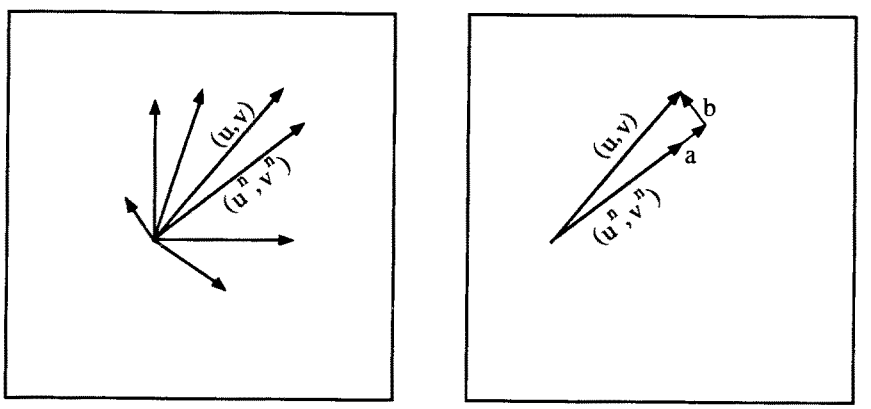
prove the convergence of the method to the exact solution. As this step is accomplished, the direction of parallel translation is obtained. Such a dynamic fixation could obviously be performed in a variety of ways, for instance by using correlation-based methods or other statistical techniques (see for example the articles by Blake et al. and Coombs and Brown in this issue). Here however, as we are mostly interested in the information content of a normal-flow field, we would like to explain how dynamic fixation could be achieved using only normal-flow information.

Consider then the normal flow in a set of directions in a small area around the origin (fixation point). Since normal flow is the projection of the optic flow on the gradient direction, the largest of the normal-flow values in the different directions is therefore the one closest to the optical flow. Let us call this normal-flow vector the *maximum normal flow* and denote it by (u^n, v^n) (see figure 5). We take it as an approximation to the correct optic flow and use it to track the fixated point. The purpose of tracking is to correct for the error in the approximation. In order to keep a point with optical flow (u, v) in the center of the image, the observer has to perform a movement that produces the same value of optical flow in the opposite direction. The way the observer accomplishes this task is by rotating the camera around the nodal point about the x- and y-axis. While the observer is moving it takes the next image and computes again the normal-flow vectors. If the maximum normal flow was equal to the optical flow, a new optical flow (due to object motion and egomotion) of zero will be achieved.

Usually, however, the maximum normal flow and the optical flow will not be equal; there will be a difference in magnitude and/or in direction. An error in magnitude results in a flow vector in the direction of maximum normal flow, and an error in direction creates a flow vector perpendicular to it (see figure 5b). The actual error is usually in both magnitude and direction. Thus the new flow vector is a vector sum of the two components. Again it can be approximated by the largest normal-flow vector measurement. The new measured normal flow is used as a feedback value to correct the optical flow and the tracking parameters; the new normal-flow vector is added to the maximum normal-flow vector computed in the first step. Proceeding by applying the same technique to the successive estimated errors will result in an accurate estimate of the actual flow after a few iterations. The proof of convergence to the exact solution follows:

We use here a simplified model to explain tracking. The change of the local coordinate system during tracking and the fact that the object is coming closer is not considered. Since for the purpose of estimating image motion at the point of fixation the number of tracking steps is small, the error originating from this model is not essential. Concerning a specific application, the algorithm will stop when the computed error is smaller than a given threshold, which will cover model errors.

In each iteration step we are computing an approximation to the difference between the observer's egomotion and the object motion. Considering the possible sources of error we have to show that the approximation error will become zero.



(a) (b)

Fig. 5. (a) Normal flow vectors measured in different directions. (b) The new flow vector (resulting from object motion and tracking) is due to (1) the error in magnitude, and due to (2) the error in direction.

Deviations of the chosen maximum normal flow from the optical-flow value are due to (a) deviations covered through the model, (b) deviations coming from simplifications and discrete computations, and (c) general errors occuring in normal-flow computation.

Regarding (a), the fact that normal-flow measurements are computed in a finite number of directions causes an error in direction of up to half the size of the interval between two normal-flow measurements. If measurements in n directions are performed the maximum error p is bounded by $p < \pi/2n$.

Regarding (b), in the evaluation of flow measurements the parts linear and quadratic in x, y, and $Z - Z_0$ are ignored. Furthermore each measurement in one direction is computed as the average of the normal flow values in a range of directions. These reasons may cause errors in magnitude as well as direction, and a different vector than the closest normal-flow vector may be chosen.

Finally, sensor noise in normal-flow measurements and the numerical computation of the derivatives of the image-intensity function can influence the magnitude and the direction of the estimated value.

Let v be the magnitude of the actual optical flow. The error sources lead to specifying the error in magnitude, q, as a percentage of the actual value. q_i is the magnitude of error in the maximum normal-flow measurement at step i and p_i is the angle between the maximum normal-flow vector and the optical-flow vector, where $q_i < q$ and $p_i < p$. Therefore the difference between the optical flow and the first measurement of maximum normal flow is given by

$$d_1 = \begin{pmatrix} vq_1 \cos p_1 \\ v \sin p_1 \end{pmatrix}$$

where the x-axis is aligned with the maximum normal-flow vector (see figure 6). The square of its magnitude is computed as

$$\| d_1 \|^2 = v^2 q_1^2 \cos^2 p_1 + v^2 \sin^2 p_1$$

The second normal-flow vector, if measured from the direction of the maximum normal-flow vector derived in the second step, is given by

$$d_2 = \begin{pmatrix} \| d_1 \| q_2 \cos p_2 \\ \| d_1 \| \sin p_2 \end{pmatrix}$$

and the square of its magnitude is therefore

$$\begin{aligned} \| d_2 \|^2 = {} & q_1^2 q_2^2 v^2 \cos^2 p_1 \cos^2 p_2 \\ & + q_1^2 v^2 \cos^2 p_1 \sin^2 p_2 \\ & + v^2 \sin^2 p_1 \sin^2 p_2 \\ & + q_2^2 v^2 \sin^2 p_1 \cos^2 p_2 \end{aligned}$$

In general, if we denote by $\{a, b\}$ the fact that either a or b has to be chosen, then $\| d_n \|^2$ can be expressed as

$$\| d_n \|^2 = v^2 \sum_{(\ldots)} \prod_{i=1}^{n} \{q_i^2 \cos p_i^2, \sin p_i^2\}$$

where $(\ldots)$ represents "all permutation."

Since $q_i < 1$ and $\sin p_i < 1$ it follows that $\Pi_i \{q_i^2 \cos p_i^2, \sin p_i^2\}$ and thus the whole term converges to zero. Therefore the convergence of the approximation value to the actual optical-flow value has been shown for the "simplified tracking model."

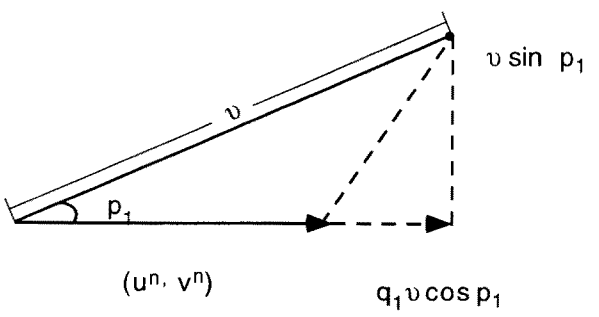


Fig. 6. Difference between optical-flow vector and maximum normal-flow vector.

6.2 Estimating the FOE Using Fixation Through Time

When continuing with fixation over time, as an object comes closer to the observer and the value of Z decreases, the optical-flow value increases. In order to track correctly and adjust to the increasing magnitude of the optical-flow value, the tracking parameters have to be changed too. From the change of the tracking parameters, the change in Z can be derived. If tracking is accomplished by rotation with a certain angular velocity, this just means that the change in depth is derived from the angular acceleration. In the sequel we show the relation between image motion and tracking movement and explain the computation of the tracking parameters, which have to be changed in every step. We explain the exact process of tracking for a geometric setting consisting of a camera that is allowed to rotate around two fixed axes: X and Y. These axes coincide with the local coordinate system of the image plane at the beginning of the tracking process.

We describe rotation by an angle ϕ around an axis, which is given by its directional cosines n_1, n_2, n_3, where $n_1^2 + n_2^2 + n_3^2 = 1$. The transformation of a point P with coordinates (X, Y, Z) before and (X', Y', Z') after motion is described through the linear relation

$$\begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} = R \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$$

where the transformation matrix R is of the following form:

$$\begin{pmatrix} n_1^2 + (1 - n_1^2)\cos\phi & n_1 n_2(1 - \cos\phi) - n_3 \sin\phi & n_1 n_3(1 - \cos\phi) + n_2 \sin\phi \\ n_1 n_2(1 - \cos\phi) + n_3 \sin\phi & n_2^2 + (1 - n_2^2)\cos\phi & n_2 n_3(1 - \cos\phi) - n_1 \sin\phi \\ n_1 n_3(1 - \cos\phi) - n_2 \sin\phi & n_2 n_3(1 - \cos\phi) + n_1 \sin\phi & n_3^2 + (1 - n_3^2)\cos\phi \end{pmatrix}$$

$$= \begin{pmatrix} r_1 & r_2 & r_3 \\ r_4 & r_5 & r_6 \\ r_7 & r_8 & r_9 \end{pmatrix}$$

Since the image coordinates (x, y) are related to the 3-D coordinates through $x = Xf/Z$ and $y = Yf/Z$, we get the following equations:

$$x' = \frac{(r_1 x + r_2 y + r_3 f)\,\mathrm{f}}{(r_7 x + r_8 y + r_9 f)}$$

$$y' = \frac{(r_4 x + r_5 y + r_6 f) f}{(r_7 x + r_8 y + r_9 f)}$$

In order to compensate for the image motion (u, v) of the point P_o, which moves from (0, 0) to (u, v) at one time unit the camera has to be rotated by ϕ, n_1, and n_2, where

$$u = n_2 f \tan\phi$$

$$v = -n_1 f \tan\phi$$

Taking at the center of the image the flow measurements (u, v) at the beginning of the tracking process at time t_1, and assuming that the object doesn't change its distance Z_1 to the camera, we can conclude that during a time interval Δt an image flow $(u\,\Delta t, v\,\Delta t)$ would be measured. The tracking motion necessary for compensation is given by

$$\frac{Uf}{Z_1} = n_2 \tan\phi$$

But at time t_2 the object has moved to distance Z_2 and we measure a rotation

$$\frac{Uf}{Z_2} = n_2' \tan\phi'$$

Figure 7 shows the relationship between the 3-D motion and the tracking parameter. Since $Z_2 - Z_1 = W\,\Delta t$, the change in the reciprocal of the rotation angle is proportional to W/U, because

$$\frac{1}{n_2 \tan\phi} - \frac{1}{n_2' \tan\phi'} = \frac{Z_2 - Z_1}{U\,\Delta t} = \frac{W\,\Delta t}{U\,\Delta t}$$

and the FOE $(U/W, V/W)$ can be computed as

$$\frac{U}{W} = \frac{1}{\dfrac{1}{n_2' \tan\phi'} - \dfrac{1}{n_2 \tan\phi}}$$

$$= \frac{1}{\dfrac{1}{n_2' \tan\phi'} - \dfrac{f}{u\,\Delta t}}$$

and

$$\frac{V}{W} = \frac{1}{\dfrac{1}{-n_1' \tan\phi'} - \dfrac{f}{v\,\Delta t}}$$

It remains to be explained how tracking is actually pursued, since we are facing the problem of a constantly changing local coordinate system. This is described in the appendix, which explains the computation of the tracking parameters.

6.3 Estimating the Time to Collision

If the values of the motion parameters do not change over the tracking time, the value Z/W, the time to collision, expresses the time left until the object will hit

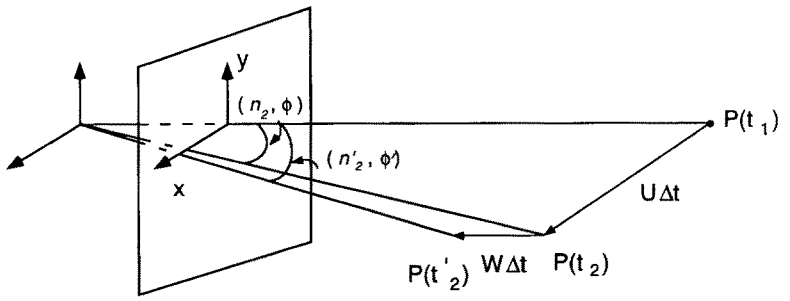


Fig. 7. From the optical flow value, which is due only to translation parallel to the image plane, a translation of P from $P(t_1)$ to $P(t_2)$ is inferred, and therefore the tracking parameters (n_2, ϕ) are expected. But actually the point has moved to $P(t'_2)$ and a rotation described by (n'_2, ϕ') is measured.

the infinitely large image plane. A relationship between FOE and time to collision is contained in the scalar product of the optical-flow vector (u, v) with the vectors in gradient direction (n_x, n_y):

$$\begin{pmatrix} u \\ v \end{pmatrix} \begin{pmatrix} n_x \\ n_y \end{pmatrix} = \| v^n \|$$

For the pixels in the center—for which we ignore the linear and quadratic parts in x, y, and $Z - Z_0/Z$ in the relation between optical flow and 3-D parameters—we get the relationship

$$\frac{Uf}{Z} n_x + \frac{Vf}{Z} n_y = \| v^n \|$$

$$\frac{Uf}{W} n_x + \frac{Vf}{W} n_y = \| v^n \| \frac{Z}{W}$$

Since we know the FOE, we can compute the time to collision from this relationship, by measuring the normal flow value in all directions of the set and by solving an overdetermined system of linear equations through the minimization of the least squares error. At this point it should be mentioned that recent developments on the use of divergence or curl and the theorems of Green or Stokes for deriving exact estimates of the time to collision without going through the intermediate stage of computing the motion parameters (Cipolla & Blake 1992) are only of thematic interest (qualitative vision) and so far of little or no practical value. The reason is of course due to the high order of derivatives employed and the underlying assumptions about continuity and differentiability in the family of Green's theorems.

7 Active Egomotion Recovery

For an active monocular observer undergoing unrestricted rigid motion in the 3-D world we compute the parameters describing this motion. Using a camera-centered coordinate system, the equations relating the velocity (u, v) of an image point to the 3-D velocity and the depth Z of corresponding scene point are (Longuet-Higgins & Prazdny 1984):

$$u = \frac{(-Uf + xW)}{Z} + \alpha \frac{xy}{f} - \beta \left(\frac{x^2}{f} + f \right) + \gamma y$$

$$v = \frac{(-Vf + yW)}{Z} + \alpha \left(\frac{y^2}{f} + f \right) - \beta \frac{xy}{f} - \gamma x$$

where (U, V, W) denotes the translation and (α, β, γ) the rotation vector.

The number of motion parameters that a monocular observer is able to compute under perspective projection is limited to five: the three rotational parameters and the direction of translation. We therefore introduce coordinates for the direction of translation $(x_0, y_0) = (Uf/W, Vf/W)$, and rewrite the right-hand sides of the above equations as sums of translational and rotational components:

$$\begin{aligned} u &= u_{\text{trans}} + u_{\text{rot}} \\ &= (-x_0 + x) \frac{W}{Z} + \alpha \frac{xy}{f} - \beta \left(\frac{x^2}{f} + f \right) + \gamma y \\ v &= v_{\text{trans}} + v_{\text{rot}} \\ &= (-y_0 + y) \frac{W}{Z} + \alpha \left(\frac{y^2}{f} + f \right) - \beta \frac{xy}{f} - \gamma x \end{aligned}$$

Since we can only compute the normal flow, the projection of the optical flow on the gradient direction

(n_x, n_y), only one constraint on the optical flow can be derived at any given poiint. The value u_n of the normal-flow vector along the gradient direction is given by

$$u_n = un_x + un_y$$

$$u_n = \left[(-x_0 + x)\frac{W}{Z} + \alpha\frac{xy}{f} - \beta\left(\frac{x^2}{f} + f\right) + \gamma y\right] n_x + \left[(-y_0 + y)\frac{W}{Z} + \alpha\left(\frac{y^2}{f} + f\right) - \beta\frac{xy}{f} - \gamma x\right] n_y \quad (1)$$

The above equation demonstrates the difficulties of motion computation using normal flow for a passive observer. There is only one constraint at every image point but there are five unknown motion parameters and every new point introduces one more unknown (a scaled depth component $-W/Z$). However, the ability of an active observer to fixate at an environmental point and keep it stationary at the center of the visual field can be exploited to provide additional information and thus simplify the problem. The estimation of an active observer's 3-D motion relative to a static scene is accomplished through four modules.

1. Through the fixation and tracking of a point in the scene, additional information about the location of the FOE is derived. The FOE is constrained to lie on a straight line and this line also supplies partial information about the observer's rotation (section 7.1).
2. Selected normal-flow values form a global pattern in the image plane which is defined by the coordinates of the FOE and one rotational parameter. Using the information provided by the previous module, locating this pattern amounts to one-dimensional search. This procedure provides a set of possible locations for the FOE (section 7.2).
3. In order to further narrow down the possible locations of the FOE and to compute the remaining rotational parameters, a process of "detranslation" is performed. For every candidate FOE provided by the previous module the normal-flow vectors which do not contain that translation are examined to find out whether they are only rotational (section 7.3).
4. Finally, the fourth module (total derotation) eliminates all impossible solutions by checking the validity of the five motion parameters at every image point (section 7.4).

7.1 The Fixation Constraint

Assume that an active observer in rigid motion is tracking, as before, an environmental point whose image (x, y) lies at the center of the visual field ($(x, y) = (0, 0)$). Assume then that during a small time interval $[t_1, t_2]$ the motion of the observer remains constant and that during this time the camera, in order to correctly track, rotates around its X- and Y-axes with rotational velocities $\omega_x(t)$, $\omega_y(t)$ respectively, with $t \in [t_1, t_2]$. The tracking rotation adds to the existing flow field (u, v) a rotational flow field (u_{tr}, v_{tr}), where

$$u = \frac{-Uf + xW}{Z} + \frac{\alpha xy}{f} - \beta\left(\frac{x^2}{f} + f\right) + \gamma y$$

$$v = \frac{-Vf + yW}{Z} + \alpha\left(\frac{y^2}{f} + f\right) - \beta\frac{xy}{f} - \gamma x$$

$$u_{tr} = \omega_x\frac{xy}{f} - \omega_y\left(\frac{x^2}{f} + f\right)$$

$$v_{tr} = \omega_x\left(\frac{y^2}{f} + f\right) - \omega_y\frac{xy}{f}$$

ω_x, ω_y are the tracking velocities at the time of the observation, and Z is the depth of the tracked point.

As before, if tracking rotation is represented by an angle ϕ around a rotation axis $(n_1, n_2, 0)$ with n_1, n_2 directional cosines, then the introduced flow (u_{tr}, v_{tr}) is given by

$$u_{tr} = n_2 f \tan\phi$$

$$v_{tr} = -n_1 f \tan\phi$$

Since the camera is continuously tracking the point at the origin, at any time $t \in [t_1, t_2]$ the introduced tracking motion compensates for the existing flow there, that is,

$$n_{2_t} f \tan\phi_t = \frac{Uf}{Z_t} + \beta f$$

$$n_{1_t} f \tan\phi_t = -\frac{Vf}{Z_t} + \alpha f$$

with the subscript t denoting the time of observation. Writing the above two constraints at times t_1 and t_2 and measuring the involved quantities with regard to the coordinate system at the beginning of the tracking process, we have

$$n_{2_{t_1}} f \tan\phi_{t_1} = \frac{Uf}{Z_{t_1}} + \beta f \quad (2)$$

$$n_{1_{t_1}} f \tan \phi_{t_1} = -\frac{Vf}{Z_{t_1}} + \alpha f \tag{3}$$

$$n_{2_{t_2}} f \tan \phi_{t_2} = \frac{Uf}{Z_{t_2}} + \beta f \tag{4}$$

$$n_{1_{t_2}} f \tan \phi_{t_2} = -\frac{Vf}{Z_{t_2}} + \alpha f \tag{5}$$

Subtracting (4) from (2) and (5) from (3), we obtain

$$f(n_{2_{t_1}} \tan \phi_{t_1} - n_{2_{t_2}} \tan \phi_{t_2}) = Uf\left[\frac{1}{Z_{t_1}} - \frac{1}{Z_{t_2}}\right]$$

$$f(n_{1_{t_1}} \tan \phi_{t_1} - n_{1_{t_2}} \tan \phi_{t_2}) = -Vf\left[\frac{1}{Z_{t_1}} - \frac{1}{Z_{t_2}}\right]$$

or by dividing

$$\frac{V}{U} = \frac{n_{1_{t_2}} \tan \phi_{t_2} - n_{1_{t_1}} \tan \phi_{t_1}}{n_{2_{t_1}} \tan \phi_{t_1} - n_{2_{t_2}} \tan \phi_{t_2}}$$

In the sequel we denote the known quantity $n_{1_{t_2}} \tan \phi_{t_2} - n_{1_{t_1}} \tan \phi_{t_1}/n_{2_{t_1}} \tan \phi_{t_1} - n_{2_{t_2}} \tan \phi_{t_2}$ which is defined by the ratio of the tracking accelerations in the vertical and horizontal directions, by T. If $(x_0, y_0) = (Uf/W, Vf/W)$ is the FOE, the above equation becomes $y_0/x_0 = V/U = T$, which is a linear constraint on the FOE. It restricts the location of the FOE to a straight line passing through the origin of the image coordinate system with slope T (see figure 8).

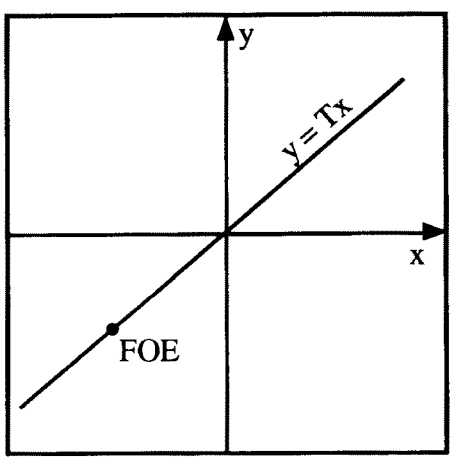


Fig. 8. Fixation constrains the FOE to lie on the line $y = Tx$ and provides the value for the ratio $\beta + \omega_y/\alpha + \omega_x = -T^{-1}$.

7.2 Patterns of Normal Flow

Since the tracking rotation is only around the X- and Y-axes, it would be interesting to examine the structure of the normal-flow-field values not depending on rotation around the Z-axis. In other words, tracking adds a rotational field but does not affect the rotation around the Z-axis.

In the sequel we concentrate on the normal flow vectors not containing rotation around the Z-axis, hereafter called γ-vectors. There are all the normal-flow vectors perpendicular to circles with center at the origin of the image coordinate system. The lines defining the directions of such vectors pass through the origin. Let us also call a γ-vector positive if it points in the direction (x, y) (figure 9); otherwise, its orientation is said to be negative.

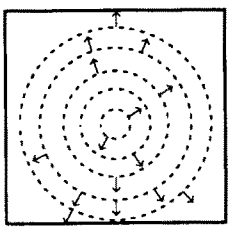


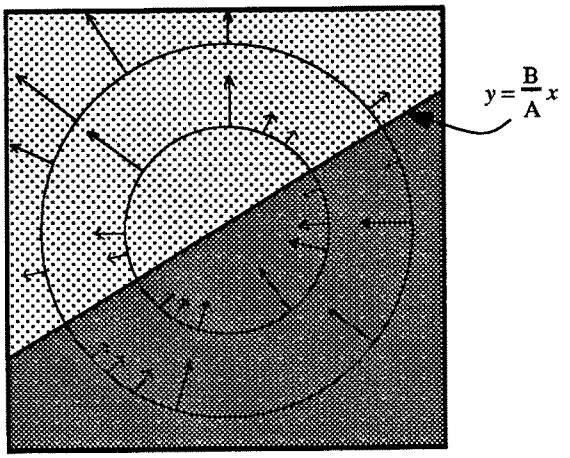
Fig. 9. Positive γ-vectors.

First, we concentrate on the rotational component of the γ-vectors: Along the positive direction, the rotational contribution is

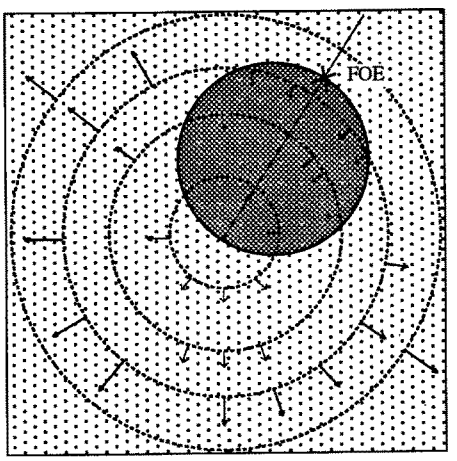
$$u_{\text{rot}}(r, \phi) = -A\left(\frac{r^2}{f} + f\right)\sin \phi + B\left(\frac{r^2}{f} + f\right)\cos \phi$$

where $A = \alpha + \omega_x$, $B = \beta + \omega_y$, r is distance from the image center and the angle ϕ is measured from the x-axis. Thus, the rotational component of the normal flow along a vector pointing away from the image center can be described by a trigonometric function with amplitude max (A, B) and period 2π. Along the line that passes through the image center and makes angle $\phi = \arctan(B/A)$ with the x-axis, the values of the γ-vectors are zero. This line divides the plane into two halves. In one half, the vectors point in the positive direction; and in the other half, they point in the negative direction. In the future we simply refer to them as positive and negative vectors (figure 10a).

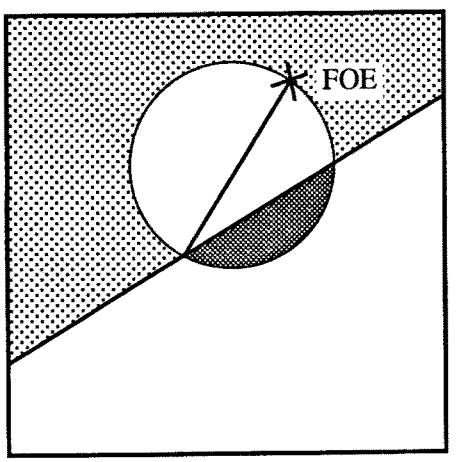
We now turn our attention to the translational component of the γ-vectors: The translational component of the motion field is characterized by the location of the FOE in the image plane. The γ-vectors lie on lines passing through the image center and the optical-flow



(a)



(b)



(c)

Fig. 10. (a) The γ-vectors due to rotation separate the image plane into a half-plane of positive values and a half-plane of negative values. (b) The γ-vectors due to translation are negative if they lie within the circle defined by the FOE and the image center and are positive at all other locations. (c) A general rigid motion defines an area of positive γ-vectors and an area of negative γ-vectors. The rest of the image plane is not considered.

values due to translation lie on lines passing through the FOE. These two lines are at right angles for all points on a circle that have the FOE and the image center as diametrical opposite points. At these points the γ-vectors' translational components vanish. Thus, the geometric locus of all points where there is zero translational normal flow is a circle. The diameter of this circle is the line segment connecting the image center and the FOE. At all points inside this circle the two lines enclose an angle greater than 90° and the normal flow along the γ-vector therefore has a negative value. The normal flow values outside the circle are positive (figure 10b).

In order to investigate the constraints associated with a general motion, the geometrical relations derived from rotation and from translation have to be combined. A circle separating the plane into positive and negative values and a line separating the plane into two half-planes of opposite sign always intersect (in two points, or one point in case the line is tangential to the circle), because both the line and the circle pass through the origin. This splits the plane into areas of only positive or only negative γ-vectors, and into areas in which the rotational and translational flows have opposite signs. In the latter areas, unless we make depth assumptions, no information is derivable (figure 10c).

We thus obtain the following geometrical result for the case of general motion: Points in the image plane at which the gradient direction is perpendicular to circles around the image center can be separated into two classes. For a given FOE, and for a line through the image center which represents the quotient of two of the three rotational parameters, there are two geometrically defined areas in the plane, one containing positive and one containing negative values. We call this structure on the γ-values the γ-pattern.[3] It depends on the three parameters x_0, y_0 and B/A. If we can locate this pattern through some search then in effect we have located the position of the FOE and the value B/A. The γ-pattern depends on three parameters, but the constraints derived from fixation (previous section) reduce the search for the pattern's position to only one dimension.

Indeed, from equations (2) and (3) at the origin we have

$$\frac{n_{1_{t_1}} f \tan \phi_{t_1} - \alpha f}{n_{2_{t_1}} f \tan \phi_{t_1} - \beta f} = -\frac{V}{U}$$

Since at the center $n_{1_{t_1}} f \tan \phi_{t_1}$ is equal to $-\omega_x f$ and $n_{2_{t_1}} f \tan \phi_{t_1}$ is equal to $-\omega_y f$, we obtain

$$\frac{\omega_x + \alpha}{\omega_y + \beta} = \frac{A}{B} = -\frac{V}{U} = -\frac{y_0}{x_0} = -T$$

or $B/A = -T^{-1}$ and $y_0/x_0 = T$.

In other words, tracking provides not only the line $y_0/x_0 = T$ on which the FOE lies, but also defines the line $y = (B/A)x$ which separates positive and negative rotational flow. This reduces the search for the pattern of figure 10c to one dimension. We simply search for a circle with diameter the segment connecting the origin with a point along the line $y/x = T$. This is a robust procedure as it only utilizes the sign of the normal flow. If a wide-angle lens or logarithmic retinae (Tistarelli & Sandini 1992) is employed, most of the directions representing the FOE lie in a bounded area of the image plane. Alternatively, in order to cover all possible cases, the search can be realized in the stereographic space (Sohon 1941) where the space of all orientations is bounded.

Pattern matching, since it does not utilize all values of the normal flow, may provide a set of solutions for the location of the FOE. To further narrow down the space of possible FOE location and to estimate the rotational parameters, the process of detranslation (next section) is performed.

7.3 The Process of Detranslation

By *detranslation* we refer to the process that, given the position of the FOE, selects the normal-flow vectors due to rotation only. Indeed, if the location of the FOE is given, the directions of the translational-motion components are also known. The translational vectors lie on lines passing through the FOE. The normal-flow vectors perpendicular to these lines do not contain translational components; they have only rotational components, (A, B, C). This can be seen from equation (1). If the selected gradient direction at a point (x, y) is $((y_0 - y), (-x_0 + x))$ the scalar product of the translational-motion component and a vector in the gradient direction is zero (figure 11).

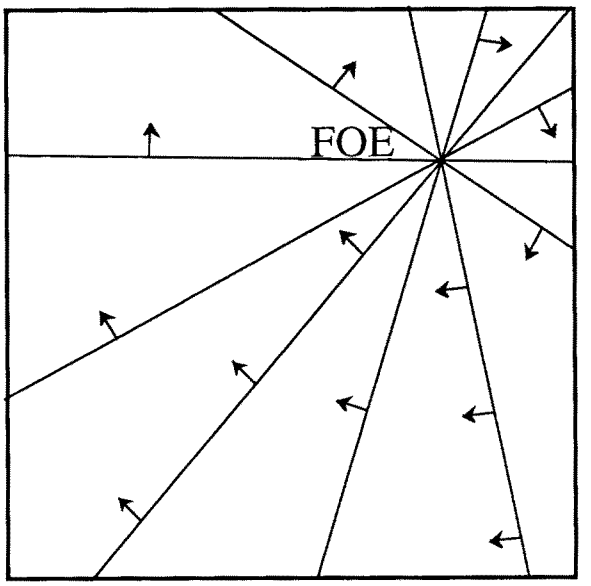


Fig. 11. Normal-flow vectors perpendicular to lines passing through the FOE are due only to rotation.

For each of the possible solutions (x_{0_i}, y_{0_i}), $i = 1, \ldots, n$, for the FOE provided by the pattern matching of the previous section, the normal-flow vectors perpendicular to the lines passing through (x_{0_i}, y_{0_i}) have to be tested to determine if they are only due to rotation (see figure 11). This results in solving an overdetermined system of linear equations, with two unknowns, since the ratio B/A is already known.

Indeed, suppose that we want to test if (x_{0_i}, y_{0_i}) is the correct location of the FOE. Consider all normal-flow vectors $\vec{u}_{n_i} = u_{n_i}(n_{x_i}, n_{y_i})$, $i = 1, \ldots, k$, perpendicular to the lines passing through (x_{0_i}, y_{0_i}). Then,

$$u_{n_i} = \left[A\frac{xy}{f} - B\left(\frac{x^2}{f} + f\right) + Cy\right] n_{x_i} + \left[A\left(\frac{y^2}{f} + f\right) - B\frac{xy}{f} - Cx\right] n_{y_i}$$

and since $A/B = -T$, we have

$$u_{n_i} = \left[-B\left(\frac{Txy}{f} + \frac{x^2}{f} + f\right) + Cy\right] n_{x_i} - \left[BT\left(\frac{y^2}{f} + f + \frac{xy}{f}\right) - Cx\right] n_{y_i}$$

$$i = 1, \ldots, k$$

So, if the above k linear equations in the two unknowns B, C are consistent, then we have found a possible FOE $((x_{0_i}, y_{0_i}))$ and we have computed its corresponding rotation.

7.4 Complete Derotation

Assume that the previous processes do not provide a single solution but a set of solutions $S = \{s_1, s_2, \ldots, s_n\}$ with $s_i = (x_{0_i}, y_{0_i}, \alpha_i, \beta_i, \gamma_i)$ candidate egomotion parameters. In order to eliminate all motion parameters that are not consistent with the given normal-flow field, every normal-flow vector has to be checked.

This check is performed using a "derotation" technique. For every parameter quintuple of S, a possible FOE and a rotation is defined. The three rotational parameters are used to derotate the normal-flow vectors by subtracting the rotational component $(u_{\text{rot}}, v_{\text{rot}})$. At every point the flow vector $(u_{\text{der}}, v_{\text{der}})$ is computed:

$$u_{\text{der}} = u_n n_x - u_{\text{rot}} n_x$$

$$v_{\text{der}} = u_n n_y - v_{\text{rot}} n_y$$

If the parameter quintuple defines the correct solution, the remaining normal flow is purely translational. Thus the corresponding optic-flow field consists of vectors that all point away from one point, the FOE (Horn & Weldon 1987). Since the direction of optical flow for a given FOE is known, the possible directions of the normal flow vectors can be determined. The normal flow-vector at every point is confined to lie in a half plane (see figure 12). The technique checks all points for this property and eliminates solutions that cannot give rise to the given normal-flow field.

7.5 The Algorithm

Assume that a rigidly moving observer is capable of tracking (with tracking velocities ω_x, ω_y) an environmental point whose image is at the origin. Then, the following algorithm outputs the observer's motion.

Step 1. The tracking acceleration provides a line $y = Tx$ on which the FOE lies, as well as the ratio $(\alpha + \omega_x)/(\beta + \omega_y)$ (section 7.1).

Step 2. Using the result of the previous step, a 1-D search along the line $y = Tx$ for the pattern of figure 12c is performed to find solutions for the FOE.

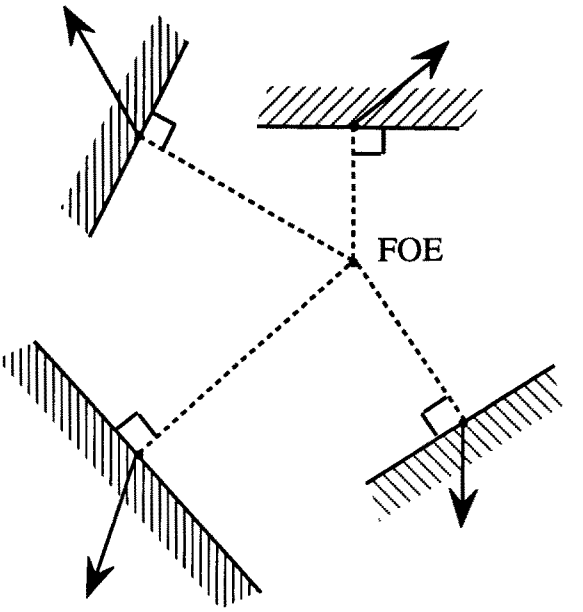


Fig. 12. Normal-flow vectors due to translation are constrained to line in half-planes.

Step 3. The previous step may provide a set $\mathbf{S} = \{(x_{0_1}, y_{0_1}), (x_{0_2}, y_{0_2}), \ldots, (x_{0_n}, y_{0_n})\}$. For each (x_{0_i}, y_{0_i}) we perform the process of detranslation, which may have two consequences. One would be to reject (x_{0_i}, y_{0_i}) as a possible solution and the other would be to accept it with the computed rotation (A_i, B_i, C_i).

Step 4. Step 3 may provide a set S of candidate solutions for the translation and the rotation:

$$S = \{(x_{0_1}, y_{0_1}, A_1, B_1, C_1), \ldots, (x_{0_n}, y_{0_n}, A_n, B_n, C_n)\}.$$

In order to reject impossible solutions complete derotation is performed to check every single normal flow vector for consistency with the motion parameters.

8 Experiments

We have tested the technique of computing object motion on synthetic imagery by using the graphics package Swivel. In this way we were able to simulate object motion as well as camera rotation. In order to analyze the robustness of the method, we evaluated the accuracy of the normal-flow values in the center of the

images. At every point we determined v_{act}, the projection of the known optical flow value on the gradient direction computed there. The error (*err*) in the normal-flow values was defined as standardized difference between v_{act} and the normal-flow value, v_{meas} ($err = (v_{act} - v_{meas})/v_{act}\%$). This way we computed an average error of 76.14% and a standard deviation of 179.64% for the motion sequence at the beginning of the tracking process. This constitutes a large error and is comparable to errors appearing in noisy real imagery.

The object displayed in figure 13 moves in the direction $U/W = 4$ and $V/W = 2$, with an image motion at the center of $u = 0.004$ and $v = 0.002$ focal units, and we tracked it over a sequence of 100 images.

Concerning the implementational details, we computed normal-flow measurements in 10 directions in an area of 9×9 pixels at the center of the image. When testing the first module, with which parallel translation is estimated, we used a threshold of 0.0002 focal units.

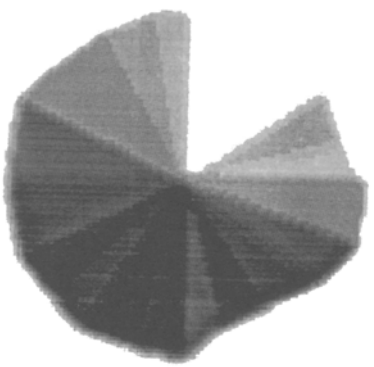


Fig. 13. First image in the sequence used for tracking.

The method converges very quickly, usually after 2 to 3 iterations. We added rotation of growing magnitude to the object motion, and it turned out that the algorithm converges for this set-up even for relatively large rotations. (The object was 25 units away from the camera and moved with translational velocity of $U = 0.1$, $V = 0.05$, $W = 0.025$ units per time unit and the method converges for rotations of up to 0.3° per time unit around the x-, y-, and z-axis.) Some graphical representations are given: Figure 14 shows for the case of no rotation the three normal-flow fields that were computed in the 9×9 pixels large area, before convergence was achieved. In figure 15 two maximum normal-flow-vector sequences are displayed (a: for no rotation, b: for rotation $\omega_x = 0.1°$, $\omega_z = 0.1°$).

Using the estimates of parallel translation from this module and continuing with tracking over 100 steps resulted in FOE values of less than 15% error (e.g., for the case of no rotation we computed an FOE of $U/W = 4.21$ and $V/W = 1.79$). With these experiments we demonstrated that the technique to compute object motion can tolerate a large amount of noise in the input (normal flow).

Especially we showed that tracking can be successfully accomplished using only normal flow under noisy conditions and that tracking acceleration can be employed for robust parameter estimation.

Building upon a successful tracking mechanism in a second series of experiments we tested the last three modules of our egomotion recovery algorithm: pattern matching, detranslation, and derotation. Concerning the implementation of these modules we took the following approach: The elimination of impossible parameters from the space of solutions involves discrimination on the basis of quantitative values. We have implemented

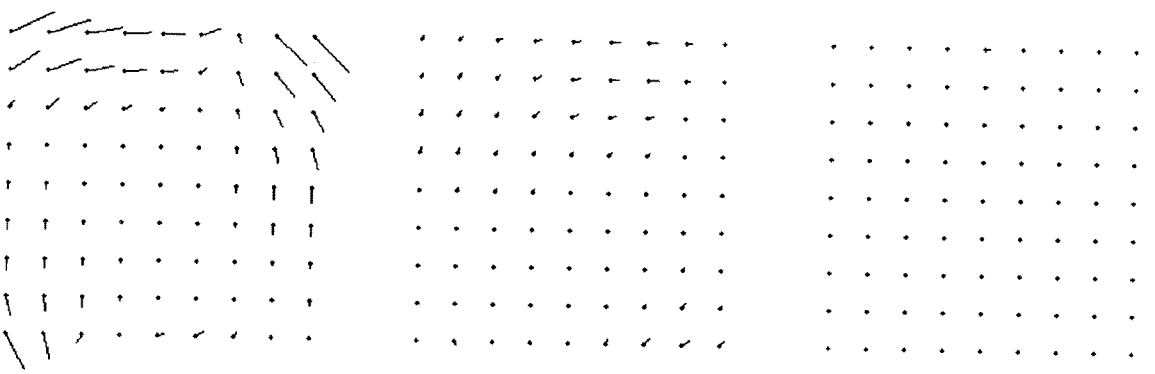


Fig. 14. Normal-flow fields for a tracking sequence.

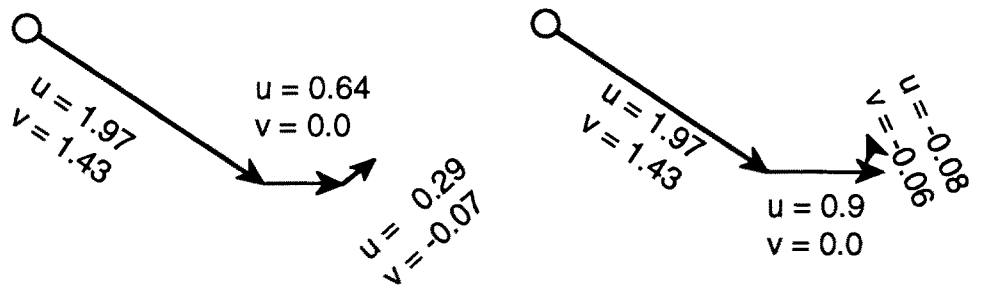


Fig. 15. Maximum normal-flow vectors for (a) no rotation and (b) rotation $\omega_x = 0.1°/\Delta t$, $\omega_z = 0.1°/\Delta t$.

this in the following way: Normal-flow values in certain directions are selected, if they are within a tolerance interval 10°. This relatively large degree of freedom, of course, will introduce some error, but there is a trade-off between accuracy and the amount of data used by the algorithm. In the pattern matching and the derotation modules counting is applied to discriminate between possible and impossible solutions. The quality of the fitting, the "success rate," is measured by the number of values with correct signs normalized by the total number of selected values. The amount of rotation in the derotation module is computed through a simple linear least squares minimization and the discrimination between accepted and rejected motion parameters is based on the value of the residual.

In the pattern-matching and derotation modules no quantitative use of values is made, since only the sign of the normal flow is considered. This particular use of data makes the modules very robust, and the correct solutions are usually found even in the presence of high amounts of noise. To give some quantitative justification of this we define the error in the normal flow at a point as a percentage of the correct vector's length. Since the sign of the vector is not affected as long as the error does not exceed the correct vector in value, our "pattern fitting" and derotation will find the correct solution in all cases of up to 100% error.

Several experiments have been performed on synthetic data. For different 3-D motion parameters normal flow fields were generated; the depth value within an interval and the gradient direction were chosen randomly. Pattern matching was tested by assuming knowledge of the lines $y = Tx$ (where the FOE lies) and $y = (B/A)x$ (which separates positive from negative rotational components) provided by the tracking constraint. The set of possible solutions was then further reduced by detranslation and derotation which were implemented as described above. In all experiments on noiseless data the correct solution was found as the best one. Figure 16 shows the optic-flow field and the

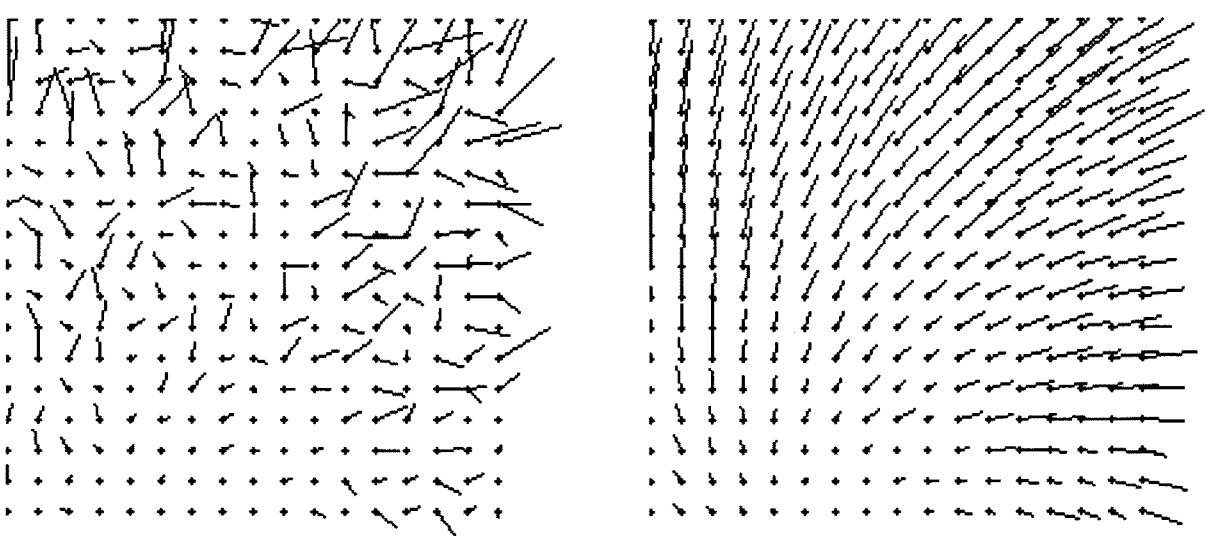


Fig. 16. Flow vectors for synthetic image: (a) Normal-flow field. (b) Optic flow field.

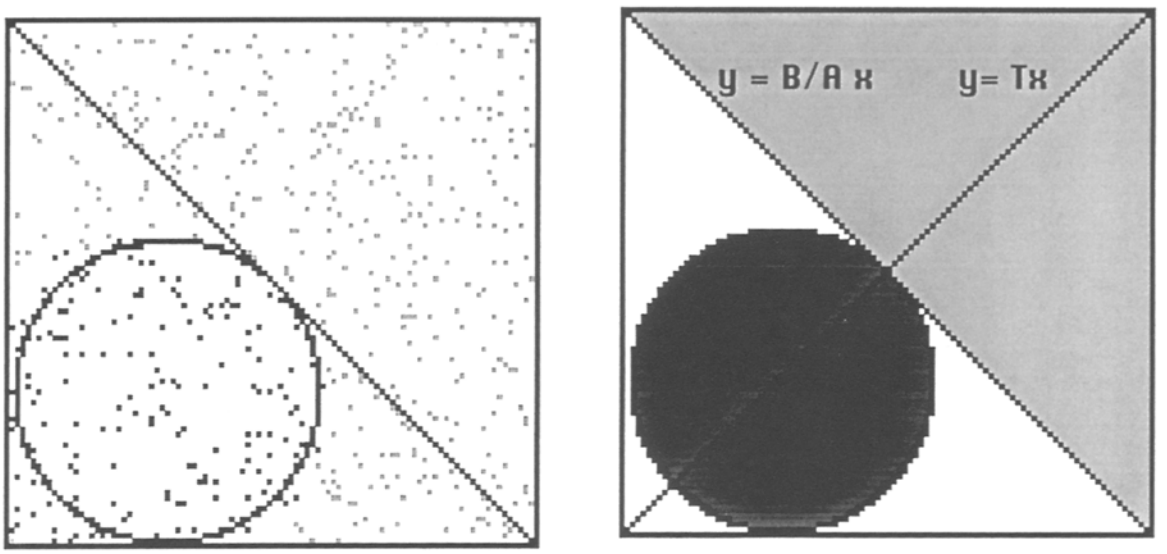


Fig. 17. (a) Positive and negative γ-vectors. (b) Fitting of γ-pattern: The line $y = Tx$ on which the FOE lies and the line $y = (B/A)x$, which separates the rotational components, are found through the fixation constraint.

normal-flow field for one of the generated data sets: The image size was 100×100, the FOE was at $(-40, -40)$ and the ratio of the rotational components was $A : B : \gamma = 1 : -1 : 15$.

In figure 17 the fitting of the γ-pattern to the γ-vectors is displayed.

Points with positive normal flow values are rendered in a light color and points with negative values are dark. Perturbation of the normal-flow vectors' lengths by up to 50% did not prevent the method from finding the correct solution.

As an example of a real scene the NASA–Ames sequence was chosen.[4] The camera undergoes only translational motion, and we added different amounts of rotation: For all points at which translational motion can be found the rotational normal flow is computed, and the new position of each pixel is evaluated. The "rotated" image is then generated by computing the new grey levels through bilinear interpolation. The images were convolved with a Gaussian of kernel size 5×5 and standard deviation $\sigma = 1.4$. The normal flow was computed by using 3×3 Sobel operators to estimate the spatial derivatives in the x- and y-directions and by subtracting the 3×3 box-filtered values of consecutive images to estimate the temporal derivatives. When adding rotational normal flow on the order of a third to three times the amount of translational flow, the exact solution was always found among the best-fitted parameter sets. In figure 18 the computed normal flow vectors and the fitting of γ-patterns for one of the "rotated" images are shown.

Areas of negative normal-flow vectors are marked by horizontal lines and areas of positive values with vertical lines. The ground truth for the FOE is $(-5, -8)$, the focal length is 599 pixels, and the rotation between the two image frames is $\alpha = 0.0006$, $\beta = 0.0006$, and $\gamma = 0.004$. The algorithm computed the solution exactly.

9 Conclusions

It has been argued by psychologists that biological organisms use tracking in the motion-estimation process. Here, we have exploited the advantages of the tracking activity to estimate egomotion and to solve for a monocular observer the problem of computing a moving object's translational direction and its time to collision. We have presented a complete solution to this task by showing how tracking can be pursued when only normal-flow measurements are used and how these parameters are of use in the 3-D motion parameter decoding strategy. The technique for estimating an object's motion consists of three modules. First, tracking is used in combination with fixation to estimate the motion components parallel to the image plane. Second, tracking serves to compute the perpendicular translational components and to estimate the FOE. The output

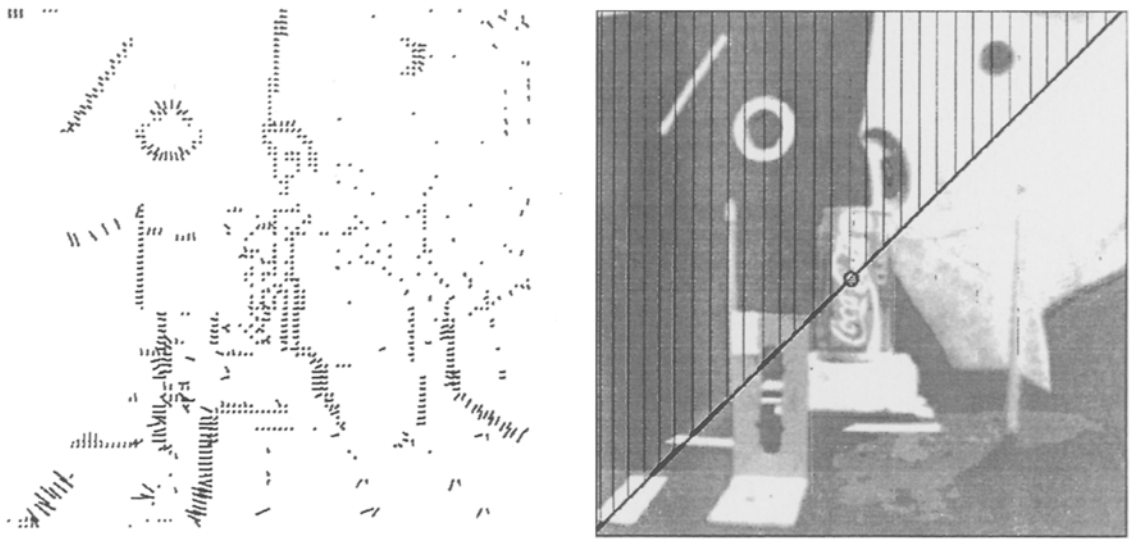


Fig. 18. NASA scene: Normal-flow field and fitting of γ-patterns.

of these modules is then employed to estimate the time to collision. A theoretical analysis of the tracking algorithm in the first module has been performed and the convergence of the method has been proved. Experimental studies have been conducted on synthetic imagery and yielded very good results.

In contrast to the first method where an object-centered coordinate system is used, for egomotion estimation a camera-centered coordinate system is more appropriate. The main difference between the two algorithms described here lies in the fact that object motion is computed from local data while egomotion estimation is based on global data. The technique uses data from all parts of the image plane and exploits geometric relations that are characteristic of a normal-flow field due to rigid motion. The algorithms can be regarded as a search technique in a parameter space where the use of fixation and tracking, along with an appropriate selection of normal-flow values, is used to reduce the dimensionality of the motion-estimation problem from five dimensions to one.

The theoretical analysis and the experiments described in this article demonstrate that the introduced algorithms have the potential of being implemented in real hardware active-vision systems, such as the ones described in by Ballard and Brown (1992) and Pahlavan and Eklundh (1992).

Appendix Computation of Tracking Parameters

Unlike section 6.1 where a "simplified model" was used, here in order to compute the tracking parameters we must take into account the change of the local coordinate system. In addition, we show the necessary parameter transformations between the coordinate systems.

Assume first that the projection of parallel translation at the beginning of the tracking process has been computed as described in section 6.1. From these measurements the rotational parameters ϕ_1, $n_{1,1}$, and $n_{2,1}$ necessary to track for one time interval were derived. However, when continuing with tracking we must consider the fact that through the rotation of the image plane the local coordinate system attached to it changes also. At each tracking step, in the current local coordinate system an optical flow emerges that is due to the change in the Z-distance. The rotation necessary to compensate for this value has to be computed and is added to the old rotation. The summation of rotational vectors is justified, since we are adding a very small vector.

The computation of the rotation vector from normal flow is performed as follows: In the new system the normal-flow vectors are computed in different directions and the maximum value is taken. This vector span from (0, 0) to (u_n, v_n). In order to compensate for this

vector by rotation around the fixed x- and y-axes, the point (0, 0) and the point (u_n, v_n) are transformed back to the old system through the equations (Tsai & Huang 1984; Fermüller & Aloimonos 1992)

$$x_{\text{old}} = \frac{(r_1 x_{\text{new}} + r_2 y_{\text{new}} + r_3 f) f}{(r_7 x_{\text{new}} + r_8 y_{\text{new}} + r_9 f)}$$

$$y_{\text{old}} = \frac{(r_4 x_{\text{new}} + r_5 y_{\text{new}} + r_6 f) f}{(r_7 x_{\text{new}} + r_8 y_{\text{new}} + r_9 f)} \qquad (6)$$

The same formula can be applied to compute from the coordinates the necessary rotation to transform one point into the other.

Acknowledgments

This research was supported in part by ARPA, the National Science Foundation under a Presidential Young Investigator Award to Y. Aloimonos, Alliant Systems Inc., Texas Instruments Inc. and the Österreichisches Bundesministerium für Wissenschaft and Forschung (through a Kurt-Gödel-Stipendium and the Österreichische Bundeskammer der Gewerblichen Wirtschaft.

Notes

1. This is not true in the case of global constraints where data from the whole image can be utilized as, for example, in the estimation of egomotion (see section 7). However, when inference about object motion needs to be made, usually only local constraints can be employed.
2. Passive navigation is a prerequisite for any other navigational ability. A system can be guided only if there is a way for it to acquire information about its motion and to control its parameters. Although it is possible to obtain the necessary information by using expensive inertial guidance systems, it remains a challenge to solve the task by visual means.
3. This pattern constitutes a class from a general class of global patterns of normal flow that can be used in motion analysis (Fermüller 1993).
4. This is a calibrated motion sequence made public for the Workshop on Visual Motion, 1991.

References

Adiv, G. 1985. Determining 3-D motion and structure from optical flow generated by several moving objects, *IEEE Trans. PAMI* 7: 384–401.

Aloimonos Y. 1990. Purposive and qualitative active vision, *Proc. Image Understanding Workshop*, pp. 816–828.

Aloimonos Y., and Brown, C.M. 1984. Direct processing of curvilinear sensor motion from a sequence of perspective images, *Proc. Workshop on Computer Vision: Representation and Control*, pp. 72–77.

Aloimonos Y., and Brown, C.M. 1989. On the kinetic depth effect, *Biological Cybernetics* 60: 445–455.

Aloimonos, Y., Weiss, I., and Bandopadhay, A. 1988. Active Vision, *Intern. J. Comput. Vis.* 2: 333–356.

Bajcsy, R. 1985. Active perception vs. passive perception, *Proc. IEEE Workshop on Computer Vision*, pp. 55–59.

Ballard, D.H. 1991. Animate vision, *Artificial Intelligence* 48: 57–86.

Ballard, D.H., and Brown, C.M. 1992. Principles of animate vision, *Comput. Vis., Graph., Image Process*, special issue on *Purposive, Qualitative, Active Vision*, Y. Aloimonos (ed.), 56: 3–21.

Bandopadhay, A., and Ballard, D.H. 1991. Egomotion perception using visual tracking, *Computational Intelligence* 7: 39–47.

Cipolla R., and Blake, A. 1992. Surface orientation and time to contact from image divergence and deformation. In A. Blake and A. Yuille, eds., *Active Vision*, MIT Press: Cambridge, MA, pp. 39–58.

Fermüller, C. 1993a. Basic Visual Capabilities, Ph.D. thesis, Center for Automation Research, University of Maryland, College Park and Institute for Automation, Technical University of Vienna.

Fermüller, C. 1993b. Motion constraint patterns, *Proc. Image Understanding Workshop*, Washington, D.C.

Fermüller, C. 1993c. Navigational preliminaries, in *Active Computer Vision*, Y. Aloimonos, ed., Erlbaum: Hillsdale, NJ, p. 51–103.

Fermüller, C., and Aloimonos, Y. 1991. Estimating 3-D motion from image gradients, Tech. Rept. CAR-TR-564, Center for Automation Research, University of Maryland, College Park.

Fermüller, C., and Aloimonos, Y. 1992. Tracking facilitates 3-D motion estimation, *Biological Cybernetics*, 67: 259–268.

Pahlavan, K. 1993. Active Robot Vision and Primary Ocular Processes, Ph.D. thesis, Department of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm.

Fermüller, C. and Kropatsch, W. 1992. Multiresolution shape description by corners, *Proc. Conf. Comput. Vis. Patt. Recog.*, pp. 271–276, Urbana–Champaign, IL.

Fleet, D.J., and Jepson, A.D. 1990. Computation of component velocity from local phase information, *Intern. J. Comput. Vis.*, 5: 77–104.

Swain, M.J., and Ballard, D.H. 1991. Color indexing, *Intern. J. Comput. Vis.*, 7: 11–32.

Swain, M.J., and Stricker, M., eds., Promising directions in active vision, Tech. Rept. University of Chicago, CS 91–27.

Horn, B., and Schunck, B. 1981. Determining optical flow, *Artificial Intelligence* 17: 185–203.

Horn, B.K.P., and Weldon, E.J. 1987. Computationally efficient methods of recovering translational motion, *Proc. Ist Intern. Conf. Comput. Vis.*, London, pp. 2–11.

Longuet-Higgins, H.C., and Prazdny, K. 1984. The interpretation of moving retinal images, *Proc. Roy. Soc. London B* 208: 385–397.

Marr, D. 1982. *Vision*, Freeman: San Francisco.

Negadharipour, S. 1986. Ph.D. thesis, MIT Artificial Intelligence Laboratory.

Negadharipour, S., and Horn, B.K.P. 1989. A direct method for locating the focus of expansion, *Comput. Vis., Graph., Image Process* 46(3).

Negadharipour, S., and Ganesan, P. 1992. A simple method for locating the focus of expansion with confidence measures, *Proc. IEEE Conf. Comput. Vis. Patt. Recog.*, Urbana–Champaign.

Pahlavan, K., and Eklundh, J.-O. 1992. A head-eye system—Analysis and design, *Comput. Vis. Graph., Image Process: Image Understanding*, special issue on *Purposive, Qualitative, Active Vision*, Y. Aloimonos (ed.), 56: 41–56.

Sohon, F.W. 1941. *The Stereographic Projection*, Chelsea Press: New York.

Spetsakis, M.E., and Aloimonos, Y. 1992. Optimal visual motion estimation, *IEEE Trans. PAMI* 14: 959–964.

Spetsakis, M.E., and Aloimonos, Y. 1990. Structure from motion using line correspondences, *Intern. J. Comput. Vis.* 4: 171–183.

Spetsakis, M.E., and Aloimonos, Y. 1991. A multiframe approach to visual motion perception, *Intern. J. Comput. Vis.* 6: 245–255.

Singh, A. 1990. Optic flow computation: A unified perspective, Ph.D. thesis, Department of Computer Science, Columbia University, New York.

Taalebi-Nezhaad, M.A. 1990. Direct recovery of motion and shape in the general case by fixation, *Proc. Image Understanding Workshop*, pp. 284–291.

Tistarelli, M. and Sandini, G. 1992. Dynamic aspects in active vision, *Comput. Vis. Graph. Image Process: Image Understanding*, special issue on *Purposive, Qualitative, Active Vision*, Y. Aloimonos (ed.), 56: 108–129.

Tsai, R.Y., and Huang, T.S. 1984. Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces, *IEEE Trans PAMI* 6: 13–27.

Ullman, S. 1979. *The Interpretation of Visual Motion*, MIT Press: Cambridge, MA.

Verri, A. and Poggio, T. 1989. Motion field and optical flow: qualitative properties, *IEEE Trans. PAMI*, 11(5): 490–498.

Waxman, A.M., Kamgar-Parsi, B., and Subbarao, M. 1987. Closed-form solutions to image flow equations for 3-D structure and motion, *Intern. J. Comput. Vis.* 1: 239–258.