CVGIP: IMAGE UNDERSTANDING
Vol. 58, No. 2, September, pp. 137–146, 1993

Velocity-Based Correspondence in Stereokinetic Images*

VALÉRIE CORNILLEAU-PÉRÈS† AND JACQUES DROULEZ

Laboratoire de Physiologie Neurosensorielle, C.N.R.S., France

Received February 22, 1990; accepted October 22, 1992

This paper explores the possibility of using the binocular optic flow as an input for the correspondence process between stereoscopic images. The main advantage of the stereocorrespondence from optic flow (SCOF) is that it does not require the use of any a priori hypothesis concerning the 3D object under analysis. In order to determine its performance relative to noisy data, we applied an algorithm of SCOF on different rigid surfaces undertaking various 3D motions. We found that when SCOF is possible it is rather robust to noise. Moreover, the study of its domain of optimal efficiency shows that SCOF is likely to cooperate well with static stereopsis or structure from motion algorithms, thereby strengthening the processing of dynamic stereo images. As far as human vision is concerned, our psychophysical results indicate that a SCOF process does not seem to be used in the perception of 3D structure. This could be accounted for by the poor contribution of convergence signals to the perception of absolute depth in human vision, which seems incompatible with the precise knowledge of the geometry of the viewing system required by the SCOF.

INTRODUCTION

The cooperation between motion parallax and binocular disparity for the perception of 3D structure has recently become a focus of interest in both computer and biological vision studies. The reason for developing the line of research is twofold. First the models of "structure from motion" and "structure from stereopsis" are mathematically similar and consist generally of two steps:

1. The search for the 2D correspondence between points in different views (successive or simultaneous) taken of the same object,

2. The interpretation of this correspondence field in terms of 3D distances and the reconstruction of a 3D map of the environment.

Second these models or algorithms are now well developed but present generally a high sensitivity to noise when their application is extended to a wide variety of images (reviews of this problem for "structure from motion" and for stereopsis are given in [1, 20], respectively).

Therefore a number of studies have developed algorithms of cooperation between motion and stereopsis as a means for reinforcing the coherence of visual processing and reducing its sensitivity to noise. They generally consider that the correspondence problem is solved for both motion and stereopsis (i.e., the optic flow, and the binocular disparities are known for every point in the two images) and develop algorithms for unifying the search for 3D parameters of the object structure and displacements. This is done in [23] for the case of orthographic projection, where the motion in depth cannot be retrieved from retinal images. If the displacement between the two optical systems is known, Mitiche [18] proposes an algorithm for the computation of the depth map and velocity-in-depth, and a test for the segregation of rigid objects with different motions in space. Alternatively Mitiche [19] shows that the relative displacements between two cameras can be deduced from the optical flow and the stereo correspondence. In [11] a constraint of smoothness on the 3D velocity of images features is applied as a guide for the disparity and temporal matches.

Motion and stereopsis can also cooperate at an earlier stage in the processing of 3D structure, namely in the 2D correspondence step. Waxman and Duncan [26] demonstrate the existence of a set of relationships linking the retinal velocities of corresponding points in the left and right images and use them to find the best disparity distribution. However, the approach developed by these authors is valid only for rigid surfaces of limited curvature (the surface is locally approximated by its tangential plane). Moreover, their scheme was run on artificial images when the 3D motion of the viewing system was known and its robustness under noise was not tested.

In this paper, we show that one component of the retinal velocity can be used directly for stereo matching for any type of surface in motion relative to the viewing system. In order to evaluate theoretically the performance of this method we discuss the case of a rigid object (Section 1). In Section 2 we present the results of computer simu-

* This work was supported by the company Essilor (Convention CIFRE n°85/224).

† Correspondence address: Laboratoire de Physiologie Neurosensorielle, 15 rue de l'Ecole de Médecine, 75270 Paris cedex 06, France.

1049-9660/93 $5.00

lations for *stereo correspondence from optic flow* (SCOF) aimed at estimating the robustness of a simple SCOF algorithm for various surfaces and motions and for small and large fields of view. Then, in Section 3, we report some related experiments performed with human subjects which suggest that the SCOF might not be used by the visual system as a cue for the binocular perception of structure from motion. Finally, in Section 4, we discuss the possible applications of the method for computational vision and we give an interpretation of our psychophysical results.

1. COMPUTING 3D STRUCTURE FROM MOTION DISPARITY: THEORY

1.1. The Dynamic Epipolar Equation

Let us demonstrate that the comparison between the 2D velocity fields obtained on images taken by two eyes/cameras can serve as an input for the processing of stereo correspondences and of the 3D structure of objects. We consider two optical systems of optical centers (or projection points) O_1 and O_2, and of known geometry. Their optical axes are supposed to be coplanar and their focal distances are equal to 1. The projection surfaces (or retinae) are supposed to be hemispheric, of centers O_1 and O_2. This is not a limitation, since a simple metrical transformation can map any retinal surface, and any vector field measured on this surface, onto an hemispheric surface.

The cartesian coordinate system $(OIJK)$ of the 3D space is defined as follows:

—O is the middle of the baseline $[O_1O_2]$,
—I is a unitary vector of the axis (O_1O_2),

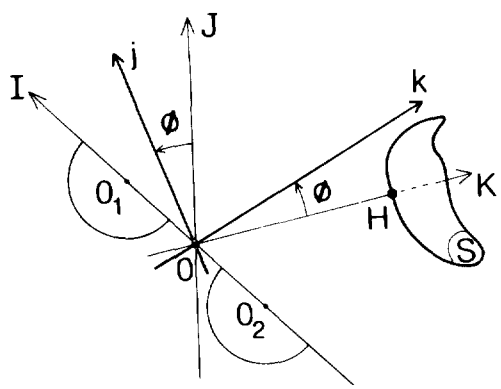

FIG. 1. Definition of the coordinate systems. O_1 and O_2 are the nodal points of the left and right eyes, O is the middle of O_1 and O_2, the axis (OK) is orthogonal to the line (O_1O_2) represented by the axis (OI). The axis J forms an orthogonal coordinate sytem with (OI) and (OK). The coordinate system $(OIjk)$ is obtained from $(OIJK)$ by a rotation of angle ϕ around (OI). The two arcs of circles represent the hemispherical retinae. S is a surface intersecting the axis (OK) in point H.

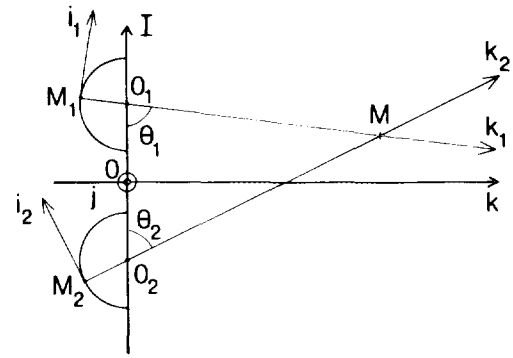

FIG. 2. Definition of the local referentials. O_1, O_2, O, I, j, k, as defined in Fig. 1. M is an object point of images M_1 and M_2 on the left and right retinae, respectively. The local referential $(M_1i_1jk_1)$ and $(M_2i_2jk_2)$ and the angle θ_1 and θ_2 are defined in text.

—the unitary vector K is normal to I and included in the plane of the two optical axes.
—J is a unitary vector normal to I and K.

The cartesian coordinate system $(OIjk)$ is obtained from $(OIJK)$ by a rotation of angle ϕ of J and K around the axis (OI) (see Fig. 1).

A point M_2 of the right retina is a possible correspondent of a point M_1 of the left retina if it lies in the plane $(O_1O_2M_1)$, so that the lines (O_1M_1) and (O_2M_2) intersect in an object point M. This is equivalent to the restriction of the possible correspondents to M_1 to the epipolar line located at the intersection of the right retina and of the plane $(O_1O_2M_1)$. Given a left image point M_1, there is a unique angle ϕ such that the plane (OIk) contains M_1, O_1, and O_2, and we shall now restrain the search for the correspondent point M_2 to M_1 to the points located in this plane which is the plane of Fig. 2.

θ_1 is the angle between the lines (O_1O_2) and (M_1O_1), and the 3D cartesian coordinate system $(M_1i_1jk_1)$ (see Fig. 2) is defined as follows:

—k_1 is the unitary vector of the line (M_1O_1),
—i_1 is the unitary vector orthogonal to k_1 and j.

For a point M_2 on the right retina, θ_2, i_2, and k_2 are defined according to the same rules (Fig. 2).

Let us consider an object point M moving with a 3D velocity U. P_1 and P_2 are the proximities of M relative to the left and right eyes (reciprocals of the distances O_1M and O_2M, respectively). The velocity of the image M_1 of M on the left image is the vector U_1 of coordinates (u_1, v_1) in the basis (i_1, j). The same notations hold for M_2, image of M on the left retina. If $\langle , \rangle$ refers to the inner product, we have

$$\begin{aligned} u_1 &= -P_1 \cdot \langle U, i_1 \rangle \\ v_1 &= -P_1 \cdot \langle U, j \rangle \end{aligned} \qquad (1)$$

$$u_2 = -P_2 \cdot \langle U, i_2 \rangle$$
$$v_2 = -P_2 \cdot \langle U, j \rangle \quad (2)$$

Considering that

$$P_1 \cdot \sin \theta_2 = P_2 \cdot \sin \theta_1, \quad (3)$$

the systems (1) and (2) are compatible if the points M_1 and M_2 verify

$$v_1 \cdot \sin \theta_2 = v_2 \cdot \sin \theta_1. \quad (4)$$

This shows that, given a point M_1 on the left retina, any point M_2 of the right retina verifying coplanarity with O_1, O_2, and M_1, and Eq. (4), is a possible correspondent of M_1. It should be noted that in the case where the two retinae are planes parallel to the line (O_1O_2), the epipolar curves are lines parallel to (O_1O_2) and the dynamic epipolar equation equivalent to (4) states that the velocities of correspondent image points have equal components in the direction orthogonal to the epipolar lines.

More generally, for any pair of viewing systems of projection centers O_1 and O_2, the epipolar equation states that correspondent image points and O_1 and O_2 lie in the same plane. The derivative of this static epipolar equation relative to time is then a dynamic equation involving the image velocities of correspondent points. A possible process of SCOF consists therefore in the pairing of image points that satisfy both the static and dynamic epipolar equations.

In the case of our hemispherical retinae, this process is divided in two steps:

—computation of the functions

$$f_i(\theta_i) = v_i/\sin \theta_i \quad (5)$$

(i being equal to 1 or 2) along the epipolar lines of the left and right images;

—the search for the points M_1 and M_2 of corresponding epipolar lines such that $f_1(\theta_1) = f_2(\theta_2)$.

The stereo correspondence problem from optic flow is thus equivalent to the classical stereo correspondence, except that the values of the luminous intensity are replaced by the values of the function f defined along an epipolar line as

$$f(\theta) = v/\sin \theta$$

(for simplicity, we drop the index i in formula (5)). A false match may occur if the function f takes the same value in different points of an epipolar line. From a purely theoretical point of view, the determination of the 3D structure from the correspondence between optical flows requires that the function f vary with the angle θ along the epipolar lines. In the case of noisy data, if we assume that the noise which perturbates the measure of the velocity field is roughly proportional to the magnitude of this velocity, *the performance of a SCOF algorithm will be an increasing function of the relative variations of* f quantified by the ratio $f'(\theta)/f(\theta)$ ($f'(\theta)$ being the derivative of f relative to θ along an epipolar line).

In theory, the algorithms that are used for the classical stereo correspondence can also be used for the SCOF with the function f as an input. In particular we could use the algorithm of Marr and Poggio [15], which first locates the zero-crossings of the second derivatives of the luminous intensity I, i.e., the points where the variations of I are maximal. However, the two types of input for stereo matching have different properties and are likely to complement one another in many situations for the following reasons:

1. Except along contours corresponding to depth discontinuities, f varies smoothly over the image of any object of the environment, whereas the variations of I are not necessarily related to changes in the depth of object points.

2. On one hand the stereo correspondence using I as an input depends mainly on the variations of I along the direction of the epipolar lines. On the other hand, the aperture problem (see, for instance, [10]) implies that the determination of v, and a fortiori of f, depends mainly on the variations of I along the direction normal to the epipolar.

1.2. The Case of Rigidity

The function f and its variations depend on the 3D structure and motion parameters. Although the SCOF can be applied to any type of visual scene, we restrict our discussion to the case of rigid motions in order to estimate the influence of different structure and motion parameters on the performance of the method. The motion is decomposed into a translation T of coordinates (t_X, t_Y, t_Z) in (IJK), and a rotation W around O, of coordinates (w_X, w_Y, w_Z). With the previous notational conventions, the expression of the function f on the left retina is

$$f(\theta) = \frac{t_Y \cdot \cos \phi - t_Z \cdot \sin \phi + a \cdot (w_Y \cdot \sin \phi + w_Z \cdot \cos \phi)}{z} - (w_Y \cdot \sin \phi + w_Z \cdot \cos \phi) \cdot \operatorname{cotan} \theta - w_X, \quad (6)$$

where a is half the baseline ($a = O_1O_2/2$), z is the k-coordinate of the object point M in the system $(Oljk)$, θ is the angle between (O_2O_1) and (O_1M), and ϕ is the vertical eccentricity of M (Figs. 1 and 2). In particular, if w_Y and

w_Z are null and if the surface is a horizontal cylinder with axis parallel to (OI) (any surface formed with lines parallel to (OI)), the function f is constant along an epipolar line, and the SCOF cannot be used.

More generally, in the case where f is contaminated by a noise proportional to its absolute value, the formula (6) indicates that

1. we can expect good results from a SCOF algorithm when the variations in depth are large in the direction of the baseline (O_1O_2) (i.e., the variations in z with θ are large);

2. the rotation component w_X does not modify $f'(\theta)$ but determines the overall value of $f(\theta)$. Therefore w_X has no influence on the performance of the SCOF for noiseless data but determines the robustness of the algorithm under noisy conditions;

3. the component of translation parallel to (O_1O_2), t_X, is useless for the SCOF;

4. in foveal or perifoveal vision (ϕ small), the only motion components that make the SCOF possible are t_Y and w_Z.

2. COMPUTER SIMULATIONS

We divided our computer simulations into two groups. The first group simulates foveal vision, when the influence of each motion component can be considered as uniform over the image (ϕ small in Eq. (6)). This first group also aimed at simulating the analysis of object structure by a biological system capable of minimizing the retinal slip around the fovea by means of eye movements. Alternatively, in the second group of simulations we consider large planar images, as can be obtained by artificial vision systems. In this case the influence of the different motion components varies in a complex way with the image coordinates, and image stabilization is not necessarily possible.

2.1. Small Field Simulations

2.1.1. *General Description*

In the first group of simulations, we tested the performance of a simple SCOF algorithm in conditions corresponding to the fine analysis of surface structure in human foveal vision. The geometry of the viewing system is defined in Section 1.1. The field of view was small (8° diameter) and the velocity was null in the center of the image (thus reproducing the conditions of our experiments reported in Section 3). Except if otherwise stated, the choice of w_X produces a minimal overall image velocity, and we expect optimal performance of a SCOF algorithm, relative to w_X.

2.1.2. *Surface Motion*

The surfaces intersected the axis (OK) (see Fig. 2) in a point H that was fixed throughout the motion, at a distance of 72 cm from O. As $|\phi|$ remains small (smaller than 4°), the expression of function f (Eq. (6)) indicates that t_Z and w_Y play qualitatively the same role for the SCOF as t_Y and w_Z, respectively, but are of negligible importance relative to them. Therefore we performed simulations for surfaces undertaking the two following motions:

—motion R. Rotation around an axis parallel to (O_1O_2); all motion components are null except t_Y and w_X, with $w_X = t_Y/Z_0$, if Z_0 is the K-coordinate of the point H (Z_0 = 72 cm here);

—motion O. Rotation around the axis (OK); all motion components are null except w_Z.

2.1.3. *Image Resolution and Surfaces*

We used circular images of diameter 40 pixels, obtained on a hemispherical retina, each pixel representing 0.2° visual angle. The baseline (distance O_1O_2) was 6.2 cm. The time unit (tu) was the time separating two images. We used the following surfaces:

—planes, characterized by the couple (β_X, β_Y) of tilt angles relative to the directions I and J, respectively (the plane (0°, 0°) is the plane normal to the axis (OK)): we used tilt angles ranging between 0 and 45°;

—spheres of radius 10 cm with their center located 82 cm from O on (OK);

—"biplanes," consisting of a couple of parallel planes, normal to (OK), and intersecting this axis at a distance of 75 and 72 cm, respectively. The part of the latter plane (i.e., the nearer) corresponding to an I-coordinate smaller than −1.5 cm was taken off, and there was thus a depth-discontinuity along a line parallel to (HJ);

—dihedrons. A vertical dihedron consisted in two half-planes of tilt angles 15° and −15° in the direction I and intersecting along the line (HJ) (as a roof seen from above).

2.1.4. *The Velocity Input*

The exact velocity coordinate v along the axis j was first calculated and then perturbated by a Gaussian noise. Two types of noise were used simultaneously; the first ("proportional" noise) had a standard deviation proportional to $|v|$ and did not perturbate the points of zero v, while the second was a background noise of fixed standard deviation expressed in pixels. The standard deviations of the proportional and background noise were 3% and 1 pix/tu, respectively. In order to keep the influence of background noise roughly constant, we used 3D motion velocities that limited the maximum image velocity to 40 pix/tu (tu being the time unit defined in 2.1.3). It

should be noted that this magnitude was chosen arbitrarily and that all our results are valid for any range of image velocities, provided the background noise is always 1/40th of the maximum velocity norm.

The values of the noise could not be chosen from experimental results since there exists, to our knowledge, no quantified results concerning the computation of optic flow from natural images. However, Horn and Schunck [9] report errors of a few percent on the computed velocity field, and the simulations currently performed in our laboratory [8] indicate that various iterative methods applied on images of smooth surfaces lead to an average error of about 2 to 5% on the computed velocity field (inputs are spatial and temporal derivatives of the luminous intensity and are perturbated by a gaussian background noise of magnitude 5% of the mean value of each input distribution). Therefore the noise values chosen here seem quite appropriate for synthetic images, although they are probably optimistic for the processing of natural images by artificial vision systems. As far as biological vision systems are concerned, the results obtained by Maunsell and van Essen [16] suggest that the response variability of the neurons coding image velocity in area *MT* of the macaque monkey is proportional to the average amplitude of the response and equal to a few percent of this amplitude. Hence our choice of a proportional noise component of 3%.

2.1.5. *The SCOF Algorithm*

The noisy velocity $\mathbf{v}$ thus obtained was first filtered with a 11 × 11 pix Gaussian filter of smoothing factor 3 pix.

A simple algorithm of linear interpolation was then applied between epipolar lines of the left and right images. Given a pixel p_l of the left image, where the value of f is f_l, the algorithm searches for a couple of consecutive pixels p_r and p_r' in the right image such that f_l is between the values of f at p_r and p_r'. This search starts in the pixel p_r, that has the same position as p_l within the limits of the image, and continues by exploring its nearest neighbours alternatively on its left and right along the epipolar line. p_l is termed "high-confidence point" if it satisfies the following conditions:

(a) $|f_l|$ is higher than four times the background noise;

(b) the relative variation of f as measured between the two nearest neighbours of p_l is larger than 1%.

When the latter condition was not verified or when no corresponding value for f_l was found in the right image by the above process, the depth associated with p_l was set to the depth found for the previous pixel on the epipolar line. When condition (a) was not satisfied, the algorithm was nonetheless applied.

For each couple of correspondent points M_1 and M_2 of coordinates θ_1 and θ_2 on the left and right epipolar lines (Fig. 2) the depth z of the object point was calculated from the formula

$$z = 2 \cdot a/(\text{cotan } \theta_1 + \text{cotan } \theta_2).$$

For all the high confidence points (and only for them) we then averaged the relative error (in percentage) of this depth, and we plotted the depth profile in the planes (*OIK*) (horizontal profile).

For all the surfaces used here a disparity error of 1 pix would lead to a depth error of about 4 to 5%. This value is directly related to the spatial resolution but is not a prediction of the final error, since the effects of noise, gaussian filtering, and linear interpolation between pixels combine in a complex way.

2.1.6. *Results*

Motion R. With the noise values fixed as indicated in part 2.1.4, less than 3% of the points were found to be of high confidence in the case of a frontal plane. As the horizontal tilt angle β_X of the plane increased to 1° and 10°, the percentage of image points where depth could be determined increased to 32% and 75% while the mean depth error decreased to 1.5% and 0.4%. The results obtained for the sphere were intermediate (56% of high-confidence points, and 0.7% of mean depth error). More generally, as predicted from the theoretical considerations in Section 1.2, the performance of the algorithm improved as the horizontal tilt β_X of the surface increased. On the opposite, an increase in β_Y (the tilt in the vertical direction) slightly impaired the quality of the results; for instance, when the plane of 10° horizontal tilt was inclined by 10° or 45° in the vertical direction, the percentage of high-confidence points decreased from 75% to 60% and 10%, and the mean depth error increased from 0.4% to 0.6% and 2.9%, respectively. This effect of β_Y can be accounted for as follows: With the same notations as in 1.2, it can be shown that, for motion R, $f(\theta)$ increases with β_Y while $f'(\theta)$ remains constant. As β_Y increases, the relative variations of f thus decrease and the algorithm performs worse in the presence of proportional noise.

The horizontal profiles of some surfaces, as obtained from our SCOF algorithm, are presented in Fig. 3. The part of the profile which corresponds to points of low-confidence is represented by dotted lines. Whenever the algorithm could be applied reasonably well on the image (when the percentage of high-confidence points was, say, higher than 30%) this dashed line was found to fit the theoretical profile of the surface quite well.

These results confirmed that, for motion R, the performance of our SCOF algorithm depends crucially on the

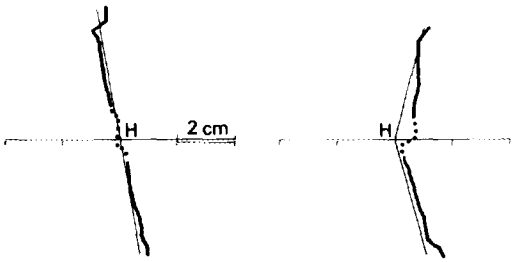

FIG. 3. Results obtained for the small field angle and motion R (rotation around a frontoparallel horizontal axis). The plane of the figure is the plane (OIK) and the graduate axis is (OK). The thin line indicates the exact section of the surface, while the reconstructed section is in thick line. The dashed line shows the low-confidence points (see text 2.1.5). The point O is located 72 cm from H in the left direction of the figure. The scale is identical for all pannels in the horizontal and vertical directions. The surface is a plane (10°, 0°) in (a) (see text section 2.1.3), and a dihedron in (b).

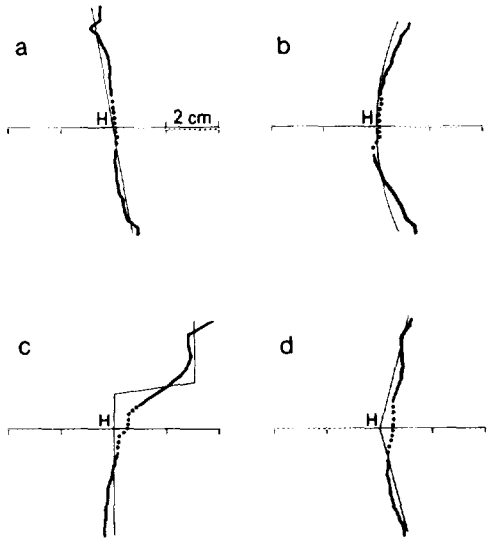

FIG. 4. Same legend as in Fig. 3, except that the depth sections were obtained for motion O (rotation around sagittal axis). The surface is a plane (10°, 0°) in (a), a sphere in (b), a biplane in (c), and a dihedron in (d).

horizontal tilt of the surface. However, this algorithm could be trusted in any situation, in the sense that it yielded a mean error of less than 1% whenever the percentage of high-confidence points exceeded 30%.

Motion O. When the surfaces undertook a rotation around the axis (OK) the algorithm yielded good results for all the surfaces described above: more than 73% of the image points could be matched with high confidence, leading to a mean depth-error of about 0.4%. The corresponding horizontal profiles of some surfaces are shown in Fig. 4. For motion O, the depth computed for points of low-confidence was always close to the theoretical depth. The smoothing of the depth profile observed in Fig. 4c is due to the gaussian filtering of the velocity and to the linear interpolation performed by the algorithm.

Influence of w_X. The strength of the hypothesis of image stabilisation, stated in Section 2.1.1, was evaluated by varying the component w_X for the motions R and O.

For motion R, when the value of w_X was reduced by 12% (the resulting motion was a rotation around an axis paralle to (OI) and located 82 cm rather than 72 cm from O) the velocity ranged between 31 and 40 pix/tu rather than between 0 and 40 pix/tu (recall that the 3D velocity was adjusted so that the maximum image velocity was 40 pix/tu; then, as predicted in Section 1.2, the relative variations of f over the image decreased and caused an impairment of the SCOF process. For instance, the percentage of image points allowing a reconstruction of the plane (10°, 0°) decreased from 74% to less than 2% of the image. When motion O was combined with a component w_X of value equal to 5% of w_Z the resulting motion was a rotation around an axis of angle 5° with (OK). This resulted in a decrease of the relative variations of the function f over the image, and in a slight decrease of performance of the algorithm; for the frontal plane, the proportion of high-confidence points decreased from 73.7% to 50%, while the relative depth error increased from 0.4% to 0.7%.

2.2. Large Field Simulations

2.2.1 *General Description*

In order to determine the performance of our algorithm for images subtending large visual fields such as those used in computer vision systems, we extended our simulations to planar images obtained by viewing systems with parallel optical axes (normal to the image planes). O_1, O_2, a, $(OIJK)$ are defined as in Section 1.1 (see Fig. 5), and the image plane is normal to the axis K. If the

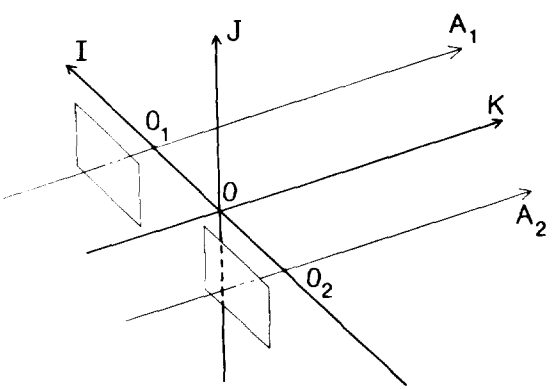

FIG. 5. The coordinate systems for large field simulations. The axes A_1 and A_2 of the two cameras are parallel. $O1$, $O2$, O, I, J, K are defined as in Fig. 1, except that K is chosen parallel to A_1 and A_2. The retinae are planes normal to (OK).

coordinates of an object point M in $(OIJK)$ are (XYZ), the coordinates of its image along axes parallel to (OI) and (OJ) on the left image plane are

$$x_l = -(X - a)/Z$$
$$y_l = -Y/Z.$$

The epipolar lines are then the image lines parallel to (OI), and the dynamic epipolar equation states that the image velocities along the axis (OJ) of corresponding image points are equal. In the case of a rigid object, if (t_X, t_Y, t_Z) are the translation coordinates in (IJK) and (w_X, w_Y, w_Z) are the coordinates of the rotation around O, the function to be matched between pairs of epipolar lines is thus

$$f(x_l) = \frac{(t_Y - y_l \cdot t_Z) + a \cdot (w_Z + y_l \cdot w_Y)}{Z} + (w_Z + y_l \cdot w_Y) \cdot x_l - w_X \cdot (y_l^2 + 1). \quad (7)$$

Here the depth of an object point is calculated from the coordinates x_l and x_2 of corresponding image points as

$$Z = 2 \cdot a/(x_l - x_2).$$

The predictions 1, 2, and 3 of Section 1.2 still hold here. Two other predictions could be stated from formula (7):

4′. Since the coordinate y_l cannot be neglected here (contrary to the vertical eccentricity in the small field study), we can also predict that not only the component w_Z, but also w_Y should allow the determination of depth. Moreover, these two components should play approximately the same role, since they yield identical relative variations of f in the whole image, except along the epipolar line defined by $y_l = 0$.

4″. A calculus of the relative variations of f along the epipolar lines also shows that the performance should decrease in the left and right extremal parts of the image. In the case of motion O, for instance, these relative variations are divided by about 15 from the center to the left or right extremity of the image.

2.2.2. *Simulation Parameters*

The images were 120 × 120 pix, each pixel subtending 0.01 cm, with a focal distance of 1, and covered about 60° viewing angle. For the stimuli used here, a disparity error of 1 pix would result in a depth error ranging between 12 and 18%, which was three to four times higher than for small field images (this factor thus quantifies the loss of resolution). The noise and the algorithm were the same as in the small field simulations, but the maximum image velocity was 120 pix/tu. The surfaces were also the same, except that the sphere was placed in front of a planar background located 92 cm from O, as it did not cover fully the field of view, and that the biplane discontinuity was 10 cm in depth, rather than 3 cm, to overcome the loss of angular resolution.

2.2.3. *Results*

Motions R and O. The results were qualitatively the same as for small field simulations, except that the depth errors were about 8 to 12 times higher. This can be accounted for partly by the loss of resolution and partly by the effect of the horizontal eccentricity described above (prediction 4″). For motion R and surfaces presenting a high-confidence area of more than 70% of the whole image, the mean depth error was generally of about 5%, although it reached 15% for the (10°, 45°) plane. For motion O the surfaces presented more than 80% high-confidence points, with a depth error of about 2%. Several depth profiles are presented on Figs. 6 and 7. As predicted above, the amplitude of the depth oscillations around the exact value increase from the image center to the left and right peripheral parts.

Rotation around a Vertical Axis. When all motion components were null except w_Y, the planes were reconstructed with more than 70% high-confidence points and a mean depth error of less than 3%. As stated above (prediction 4′) these results are similar to those obtained for motion O.

Translation in Depth. When all the motion components are zero except t_Z, formula (7) indicates that the variations of f with x_l depend linearly on the horizontal variations of Z. Indeed our results showed that the algorithm yielded acceptable performance only when the X-tilt angle of the plane reached 45° (high-confidence area, 78%; mean depth error, 8%).

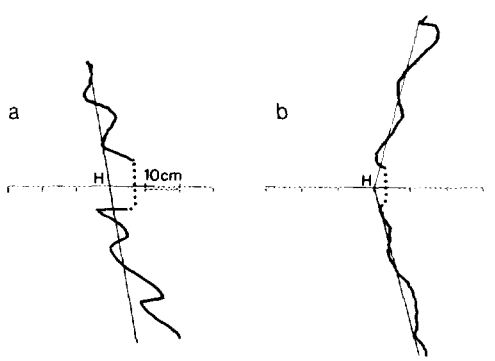

FIG. 6. Results obtained for large field simulations and motion R. Same legend as in Fig. 3. The scale is indicated in (a). The surfaces are the plane (10°, 0°) in (a) and the dihedron in (b).

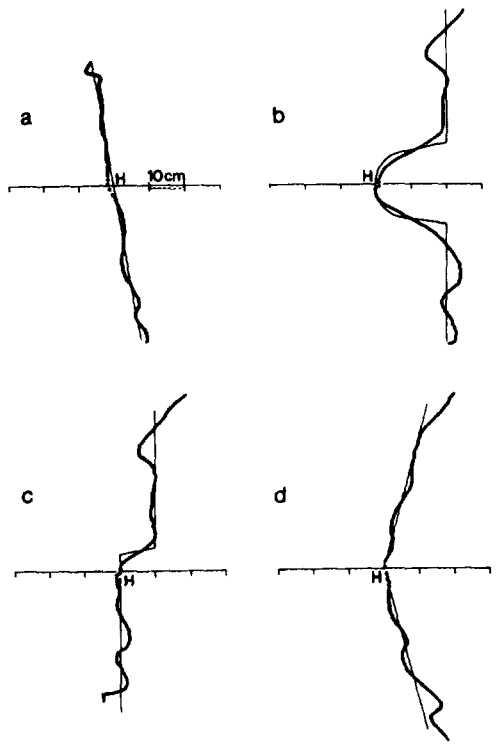

FIG. 7. Results obtained for large field simulations and motion O. Same legend as in Fig. 3. The surfaces are a plane (10°, 0°) in (a), a sphere located in front of a planar background in (b), a biplane in (c), and a dihedron in (d).

Influence of w_X. The performance of the algorithm were less sensitive to a change in the value of w_X than for the small field angle. For motion R, when the value of w_X was reduced by 12% of its value, the high-confidence area and the mean depth error were not modified. For a pure translation along (OJ) ($w_X = 0$), the results became reasonably good only when the X-tilt of the plane reached 30° (high-confidence area, 84%; mean depth error, 8%).

As far as motion O is concerned, the algorithm tolerated the addition of a w_X component equal to half the value of w_Z (for the frontal plane the mean depth error increased from 1.5% to 3%). The rotation around (OJ) tolerated a component w_X of about 10% of the value of w_Y (the mean depth-error then increased from 2.5 to 5%).

Finally it should be stressed that the quality of the results could generally be predicted from the overall value of the variations of the function f with θ (the relative depth-error ranged between 1.5 and 15% when the high-confidence area was larger than 70%).

2.3. Conclusion

The results of our simulations indicate that even when the SCOF is theoretically possible its application to natural images may yield variable results according to the type of 3D object, 3D motion, and the eccentricity of the part of the visual scene. Five main conclusions emerge:

1. the components t_X and w_X alone do not allow SCOF and do not theoretically modify the results of the SCOF for a given 3D motion;

2. the robustness to noise of the SCOF is generally optimal when the component w_X minimizes the global image motion;

3. for any type of surface, a rotation around the sagittal axis (OK) provides the most reliable velocity information for the SCOF which presents then a high robustness under noise;

4. for several type of 3D motions (in particular, the translations along a vertical axis (OJ)) the SCOF allows a good reconstruction of the parts of the environment which present steep variations in depth;

5. when the SCOF can be performed with a high level of confidence (determined by the value of the relative variations along an epipolar line of the image velocity component orthogonal to this epipolar), the reconstruction of the depth map is reasonably good (generally between 2 and 5% relative error for our noisy velocity fields).

3. THE SCOF IN BIOLOGICAL VISION SYSTEMS

Several neurophysiological studies [21, 27, 6, 7, 22, 24, 25] have demonstrated the existence of neurons coding specifically the disparity of retinal motion (i.e., the difference in velocity of a stimulus on the two retinae) in the visual pathway of the cat and monkey (although this is still under discussion for area MT of the monkey, see [17]). In addition, the psychophysical studies by Lee [13] and Beverley and Regan [3] suggest that motion disparity can be used by the visual system for the perception of 3D motion. Since 3D motion and structure parameters are deeply related in the problem of structure from motion, the question arises whether motion disparity could also provide 3D structure information to the visual system.

It has been demonstrated in [4] that stereoscopic depth perception is strengthened by the presence of edges or zero-crossings. However, if such characteristics are lacking, 3D structure can still be perceived from intensity-based stereo. Therefore the fact that the function f, which is the input of the SCOF, presents generally smooth variations does not seem to be an obstacle to its use by the visual system.

Lee [13] demonstrated that a human observer could use motion disparity as a cue for *3D motion*, in the absence of position disparity. We designed a set of experiments to test whether motion disparity could also serve as a cue for *3D structure* in this case [5]. Human subjects

were shown stereo-kinematograms representing a surface defined by a set of dots. Only half of the dots (randomly chosen) was effectively presented to one eye, while the other half was presented to the other eye. Therefore the two kinematograms were stereoscopically motion-correlated, but not position-correlated. Motion parallax was available on each monocular kinematogram, while each pair of stereo images was completely devoid of depth information.

For different types of motion and, in particular, motion R and motion O, our results suggest that the subjects did not use motion disparity in the discrimination between a planar and a spherical surface.

In another set of experiments, we measured the ability of the subjects to detect surface curvature with static stereopsis, motion parallax, or both simultaneously (then motion disparity was also provided). Our preliminary results indicate that the performance obtained with both cues present simultaneously is rather weaker than could be predicted from the results obtained for each cue separately. This also supports the conclusion that, at least with our experimental procedure, motion disparity is not used by the visual system as a cue for the perception of 3D structure.

Therefore, our results, as compared to those obtained by Lee, or Beverley and Regan, suggest that the visual system does not use motion disparity as a cue for the fine analysis of 3D structure. Rather, it seems capable of comparing, between the two retinae, the global velocities of a patch of surface (as in Lee's experiment) relative to a static frame, or to detect that an approaching object is likely to hit the head (as in Regan and Beverley's experiments).

4. DISCUSSION

4.1. Cooperation of the SCOF with Other Processes

First it should be noted that our SCOF algorithm was very simple and could be improved in many ways. In particular a constraint of smoothness of the environment, classically used in robotics (for a review see [2 Chaps. 5, 9]), could be introduced in the search for a correspondent point, by taking into account the disparities already found in the neighborhood of a given pixel (along an epipolar line and between adjacent epipolar lines).

In spite of this simplicity our simulations show that stereo-correspondence can be established by using one component of the velocity field. Like the zero-crossings, or the peaks of intensity, this input has the advantage of being theoretically independent of the absolute level of global luminous intensity and of the absolute value of the contrast at each image point. However, except if the motion of the viewing system relative to the environment is composed predominantly of a rotation around the sagittal axis (OK), the SCOF is generally not sufficient to reconstruct the depth of the environment over a whole image. Rather, it is likely to cooperate efficiently with other depth cues such as static stereopsis and motion parallax. Its similarity and complementarity with static stereopsis have already been shown in Section 1.1. Alternatively, the processes of SCOF and structure from motion parallax are difficult to compare for the following reasons:

—contrary to SCOF, all the algorithms of structure from motion proposed so far require an hypothesis of local or global rigidity of the object under analysis;

—the SCOF yields the absolute depth of image features, whereas the output of structure from motion algorithms is a relative depth map of the environment.

Despite these differences, the cooperation of SCOF and structure from motion is likely to be fruitful because they share the same input (the 2D velocity field) and perform optimally for different 3D movements; for instance, a rotation around the baseline allows the SCOF to be performed in most points of our large field images, whereas structure from motion would be useless in this case. The opposite is true for a translation along the baseline.

Finally a SCOF algorithm could complement the five-step-process used by Waxman and Duncan [26] and replace, when it is possible, the binocular flow correspondence which involves an hypothesis of rigidity of the surface under analysis.

4.2. Interpretation of the Physiological Data

The relative robustness to noise of our correspondence algorithm and its complementarity with other depth cues support the use of SCOF in artificial vision systems. However, the physiological data reported above suggests that, although the differences of velocity between corresponding points of the left and right images is probably coded in the visual pathway, the human visual system seems able to use this depth cue for the perception of the global 3D motion of a surface patch, but not for the fine analysis of 3D structure. We advance two reasons to explain this phenomenon.

First, consider a viewing system such as defined in Section 1.1, but where each eye/camera can move and where the velocity field and retinal positions are coded retinotopically, i.e., relative to a fixed point of the retina termed the fovea. The application of the SCOF as described in Section 1.2 necessitates the knowledge of the geometry of each eye/camera, but also of their positions relative to each other. If the latter are not known, the system can still perform intensity-based stereopsis and obtain a relative, rather than absolute, depth-map. On the opposite, the SCOF does not even yield relative depth, since the function f itself, which is the input of the map-

ping, cannot be calculated (the angle θ in formula (5) must be replaced by the sum of the eccentricity of the image point relative to the fovea, and the angular position of this fovea relative to the axis (OK), which is unknown). Therefore the function to be matched in SCOF along an epipolar line, as expressed in retinotopic coordinates, is not simply translated over the retinae during a change of the convergence angle, as is the case for static stereopsis. This may well be secondary for an artificial vision system, if the positions of the cameras are well characterized, but it could constitute an obstacle to the use of SCOF in human vision, where the angle of convergence is known to provide poor information of absolute depth [14, p. 243].

Second, the use of SCOF requires a rather precise determination of the epipolar lines, which is likely to be lacking in human vision; for instance, Julesz [13] demonstrates that stereopsis resists an expansion of 15% of the left image of a stereogram. A precise definition of the epipolar lines is not necessary if static stereopsis is achieved prior to the use of motion disparity as may be the case in Regan and Beverley's experiment. By contrast, this is absolutely necessary for the SCOF to be performed in our experiment involving images that are not correlated in position.

REFERENCES

1. G. Adiv, Determining three-dimensional motion and structure from optical flow generated by several moving objects. *IEEE Trans. Pattern Anal. Mach. Intell.* **7,** 1985, 384–401.
2. N. Ayache, *Vision stéréoscopique et perception multisensorielle*, InterEditions, Collection Science Informatique, Paris, 1989; English trans., MIT Press, Cambridge, MA.
3. K. I. Beverley and D. Regan, Evidence for the existence of neural mechanisms selectively sensitive to the direction of movement in space, *J. Physiol.* **235,** 1973, 17–29.
4. H. H. Bülthoff and H. P. Mallot, Interaction of different modules in depth perception, in *Proceedings, First ICCV, London, 1987*, pp. 295–305.
5. V. Cornilleau-Pérès and J. Droulez, Stereo-motion cooperation and the use of motion disparity in the visual perception of 3D structure, *Percept. Psychophys.* in press.
6. M. Cynader and D. Regan, Neurons in cat parastriate cortex sensitive to the direction of motion in three-dimensional space, *J. Physiol.* **274,** 1978, 549–569.
7. M. Cynader and D. Regan, Neurons in cat visual cortex tuned to the direction of motion in depth: Effect of positional disparity, *Vision Res.* **22,** 1982, 967–982.
8. J. Droulez and V. Cornilleau-Pérès, The use of the coherence constraints in the 3D processing of moving images, Esprit BRA n°3149 (Mucom), Second Periodic Progress Report, July 1991, A8.6.
9. B. K. P. Horn and B. G. Schunck, Determining optical flow, *Artif. Intell.* **17,** 1981, 185–203.
10. E. C. Hildreth, The computation of the velocity field, *Proc. R. Soc. London B* **221,** 1984, 189–220.
11. M. R. M. Jenkin, *The Stereopsis of Time-Varying Imagery*, Technical Report RBCV-TR-84-3, University of Toronto, 1984.
12. B. Julesz, *Foundations of Cyclopean Perception*, Univ. of Chicago Press, Chicago, 1971.
13. D. N. Lee, Binocular stereopsis without spatial disparity, *Percept. Psychophys.* **9**(2B), 1970, 216–218.
14. Y. Le Grand, *Optique Physiologique. T.III. L'Espace Visuel*, Editions de la revue d'Optique, Paris, 1956.
15. D. Marr and T. Poggio, A computational theory of human stereo vision, *Proc. R. Soc. London B* **204,** 1979, 301–328.
16. J. R. H. Maunsell and D. C. van Essen, Functional properties of neurons in middle temporal visual area of the macaque monkey. I. Selectivity for stimulus direction, speed and orientation, *J. Neurophysiol.* **49,** 1983, 1127–1147.
17. J. R. H. Maunsell and D. C. van Essen, Functional properties of neurons in middle temporal visual area of the macaque monkey. II. Binocular interactions and sensitivity to binocular disparity, *J. Neurophysiol.* **49,** 1983, 1148–1167.
18. A. Mitiche, On combining stereopsis and kineopsis for space perception, in *Proceedings, First Conf. on AI Applications, 1984*, pp. 156–160.
19. A. Mitiche, Three-dimensional space from optical flow correspondence, *Comput. Vision Graphics Image Process.* **42,** 1988, 306–317.
20. H. K. Nishihara and T. Poggio, Stereo vision for robotics, in *The First International Symposium on Robotics Research* (M. Brady and R. Paul, Eds.), pp. 489–505, MIT Press, Cambridge, MA, 1984.
21. J. D. Pettigrew, Binocular neurones which signal change of disparity in area 18 of cat visual cortex, *Nature* London **241,** 1973, 123–124.
22. G. F. Poggio and W. H. Talbot, Mechanisms of static and dynamic stereopsis in foveal cortex of the rhesus monkey, *J. Physiol* **315,** 1981, 469–492.
23. W. Richards, Structure from stereo and motion, MIT AI Memo No. 731, 1983.
24. K. Toyama and T. Kozasa, Responses of Clare–Bishop neurons to three-dimensional movement of a light stimulus, *Vision Res.* **22,** 1982, 571–574.
25. K. Toyama, Y. Komatsu, H. Kasai, K. Fujii, and K. Umetani, Responsiveness of Clare–Bishop neurons to visual cues associated with motion of a visual stimulus in three-dimensional space, *Vision Res.* **25,** 1985, 407–414.
26. A. M. Waxman and J. H. Duncan, *Binocular Image Flows: Steps toward Stereo-Motion Fusion*, University of Maryland, Computer Vision Laboratory Report CAR-TR-119, May 1985.
27. S. M. Zeki, Cells responding to changing image size and disparity in the cortex of the rhesus monkey, *J. Physiol.* **242,** 1974, 827–841.