Determination of a depth map from an image sequence

Lionel Marcé, Patrick Bouthemy

INRIA

UNITÉ DE RECHERCHE
INRIA-RENNES

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél.:(1) 39 63 55 11

Rapports de Recherche

N° 765

DETERMINATION OF A DEPTH MAP FROM AN IMAGE SEQUENCE

Lionel MARCE
Patrick BOUTHEMY

DECEMBRE 1987

IRISA

INSTITUT DE RECHERCHE EN INFORMATIQUE
ET SYSTÈMES ALÉATOIRES

Campus Universitaire de Beaulieu
35042 - RENNES CÉDEX
FRANCE
Téléphone : 99 36 20 00
Télex : UNIRISA 950 473 F
Télécopie : 99 38 38 32

Determination of a depth map from an image sequence

Determination d'une carte de profondeur a partir d'une sequence d'images

Lionel MARCE and Patrick BOUTHEMY***

IRISA (*INSA, **INRIA / Centre de RENNES)
Av. du General Leclerc, Campus de Beaulieu
35042 Rennes Cedex
France

Publication Interne n° 382 - Novembre 87 - 14 pages

Abstract

This paper deals wih the issue of building a reliable depth map for robot navigation from a sequence of time−ordered images. More precisely, we investigate the use of successive images, in number equal or more than two, to estimate a depth map at given time. The images are assumed to be acquired by a camera undergoing translation motion. First we show the difficulty of the stated problem considering real images by testing the stability of an interest operator combined with a correlation algorithm. Then we develop a method based on a local modeling of spatio−temporal (or moving) edges. A maximum likelihood scheme, which explicitly incorporates the knowledge of camera motion, allows to determine such features. From the estimated displacement of these edges in the image plane we infer the depth of corresponding object points in space. We give results obtained with a sequence of noisy synthetic images.

Résumé

Nous éxaminons le problème de la construction d'une carte fiable de profondeur à un instant donné, pour la navigation d'un robot, à partir de plusieurs images ordonnées dans le temps en nombre égal ou supérieur à deux. La caméra est animée d'un mouvement de translation uniforme. Après avoir montré la difficulté du problème en testant la stabilité d'un opérateur d'intérêt lié à un algorithme de corrélation, nous proposons une méthode basée sur une modélisation locale d'éléments de contour spatio−temporels dans la séquence d'images et un critère de maximum de vraisemblance, auquel est intégrée la connaissance du mouvement de la caméra. L'estimation des déplacements de ces éléments permet d'induire la profondeur des points correspondants des objets dans l'espace. Nous donnons les résultats de l'application de l'algorithme sur une séquence d'images de synthèse bruitées.

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
(L. A. 227)
UNIVERSITÉ DE RENNES 1 — I. N. S. A. DE RENNES

INSTITUT NATIONAL DE RECHERCHE
EN INFORMATIQUE ET EN AUTOMATIQUE
(LABORATOIRE DE RENNES)

Introduction

Our concern is the following: how to use a single camera for obstacle avoidance purpose in the context of mobile robots or manipulators and assuming that the scene is static. In such applications, camera motion is roughly known, given by sensors, such as odometers for mobile vehicles or optical encoders for manipulators.

From this knowledge of the camera motion, relative depth of possible obstacles can be inferred by using triangulation principle like in stereovision system. Therefore, we have to match features which are projections of the same object element in space into two successive frames. There lies the major difficulty; indeed matching may yield errors sometimes very large.

A bad-accuracy source comes from the image sampling, amplified by the fact that successive frames correspond to very close camera positions. An other problem results from appearing or vanishing areas from one image to the other, because of occlusion of one obstacle by an other.

First we will show the problems associated with the matching process through the use of an algorithm of the kind based on interest points. Then we will describe our method to reduce this difficulty.

Tracking interest points through an image sequence

Choice of interest operator

Some matching methods are based on interest points. Different kinds of such points have been proposed. *Moravec*'s algorithm, [MOR 81], finds large omnidirectional variance points. *Kitchen-Rosenfeld*'s, or *Nagel*'s methods detect corners.

We have tested matching quality considering *Moravec*'s algorithm, well-representative of its class according to *Thorpe*, [THO 84]. Let us briefly recall its principle.

This algorithm measures interest value of a pixel by computing luminance variance along four directions (horizontal, vertical, and the two diagonals) within a small window (for instance 5x5 pixels) around this pixel. Minimum value of these four variances is chosen. The underlying idea in using variances is that uniform areas have a low interest value. Moreover by retaining only the minimum of four directional variances, again a low value is assigned to pixels belonging to an edge. These have a large variance across the edge, but a low one along the edge. Otherwise they could be ambiguity sources in a matching scheme using some correlation algorithm. Only pixels locating at corners or in textured areas will have a large variance along all directions.

A local maximum filter tests every point versus each 24 closest neighbours to prevent from agglomerating interest points in highly textured areas.

Test of *Moravec*'s operator stability

We applied *Moravec*'s operator on a sequence of 125 images acquired at SRI with a Sony camera of 12.5 mm focal distance in an office environment with an uniform movement orthogonal to camera axis [BOL 85]. This data set allowed to test the operator stability, that-is-to-say to know whether it labels the same parts of objects through successive images. (Each acquired image consists of 240 lines and 256 pixels by line. Then these images were interpolated into images of 480 lines and 512 pixels by line.)

If we keep only the five first interest points with the highest value in each image and if we draw these points within the same graph, lines roughly parallel to the camera motion are observed (Fig.1). This means that the operator has a trend to label the same points in successive images.

Combining *Moravec*'s operator with a correlation algorithm.

To go further in the test, then we tried to find again the five interest points deduced from the first image, in all the following images with a pseudo- normalized correlation test.

The reference window is of size 5x5 pixels in the first image and the search window of size 13x13 pixels. If the matched points from the first image to the other following images represented same object parts, we would have five lines parallel to the motion direction in the drawing representing their evolution, but this is not the case (Fig. 2). This means that points are lost from time to time and that we match point projections of different object parts. Results are not better by enlarging the size of the reference area surrounding the interest point in the first image, [MAR 86] .

Nevertheless we could improve interest point tracking by restricting the search area of the correlation algorithm to a horizontal area of 9x1 pixels owing to our camera motion knowledge. In Fig. 3, in order to have a better view of temporal process, we add k pixels to the y-coordinate of the found points to give a perspective impression. If the tracked points represented same object parts, we would get lines with a slant proportionnal to the depth from the object to the camera.

We obtain curves nearly straight and parallel, because detected object points are approximatively located at about the same range from the camera (Fig. 3). Nevertheless the algorithm still looses points from time to time.

If we examine the distortion of the surface around an interest point through successive images, we can see that the interest value, associated with this luminance surface, diminishes quickly. This means that the given point could not be considered any more as an interest point.

Multi-image matching algorithm

Problem statement

Introduction of motion knowledge in the scheme encompassing an interest operator and a correlation algorithm leads to decrease the number of false matches as the above described

experimentation on real images pointed out. Experients reported by *Bharwani*, [BHA 86], on synthetic images corroborate this remark. But resulting improvement is still insufficient.

Hence to obtain better results, we have tried to simultaneously use more than two successive images to realize somehow some kind of smoothing. But we are not concerned with some recursive estimation scheme, as in [ESP 87].

On the other hand, some more sophisticated matching procedure could be chosen, for instance a relaxation method as in [BAR 80], where Moravec interest points were also considered.

However, instead of choosing interest points as primitives, we have taken into account contrast edges, which are much more numerous than interest points. That allows to get a more dense range measurements, but it implies the design of another matching technique. Indeed, we have designed an algorithm derived from the one developped by *Bouthemy*, [BOU 86], which is based upon a local modeling of spatio−temporal edges in an image sequence considered as a 3D space, (x,y,t), two spatial dimensions and one temporal. Our algorithm easily enables to simultaneously take into account in a straightforward manner several successive images, and to incorporate the knowledge of camera motion.

Modeling of a spatio−temporal edge

Due to the relative configuration between camera and objects in space, that−is−to−say moving camera and static scene, all edges in the image sequence can be considered as moving edges. Anyway, an apparent static edge can be assimilated to a moving edge whose displacement is nul. In space (x, y, t) representing the image sequence, a spatio−temporal edge or moving edge then generates a small surface patch.

Let us inspect more precisely the shape of this patch. The spatial part of the moving edge can be represented by a small straight−line segment. As a matter of fact, since the camera motion is supposed to be a translational known one, the displacement of the edge from time t to time $t + 1$ can be completely inferred owing to some local process whereas only the edge−perpendicular displacement component can be derived in the general case, [BOU 86]. Indeed, in the plane Oxy which represents the image plane, successive positions of image points, projections of the same object part, can be determined by using the focus of expansion, concisely denoted as *FOE*. The *FOE* is the intersection of the image plane with the straight−line issued from the projection center and whose direction is given by the translation vector attached to the camera within the subset of successive images considered out of the image sequence. These consecutive projections lie along a line connecting the *FOE* to the initial projection point location at time t. This line will be called an 'expansion ray'.

Besides the edge does not stay parallel to itself from one image to the next, even though its corresponding object part does remain parallel to itself in time. This can be easily explained by the analogous situation in spatial domain. Projections of parallel lines in the 3D space (X, Y, Z) in the image plane give rise to lines intersecting at the vanishing point. Then the surface patch produced by the contour element is complex enough. It can be simplified when the motion between two successive images is small with respect to the distance from the *FOE*, and when this distance is large with regard to the edge length. According to these assumptions, the edge direction can be considered as constant in the considered subpart of the image sequence; the same holds for its length. Then the surface patch locally generated by a moving edge in an image sequence can be modeled as a cylindrical one with an hyperbole as basis.

Relation between moving edge parameters and object point depth

We consider that the camera undergoes some translation motion $\underline{T}$ during the time interval corresponding to the subset of images used for inferring the depth map. Let the projection center be the origin of the coordinate system XYZ. Let the $Z-$axis be the camera axis of view. If d denotes the displacement magnitude of the image point along the expansion ray between time t and time $t+1$, W the translation component of the camera along the $Z-$axis, Z the depth of the object point in space from the camera at time t, D the distance between the corresponding projection point in the image and the FOE at time $t+1$, we have the following relation:

$$d = D \,.\, W / Z \qquad (1)$$

Relation (1) can be easily obtained if appropriate similar triangles are considered based on the straight−line connecting the FOE to the projection center and on the displacement $-\underline{T}$ at a given point in space [WIL 81], and by back−projection on the $Z-$axis.
Hence the process for the estimation of the depth map at a given time t can be summarized as follows. First we have to determine surface patches defined by spatio−temporal edges, whose parameters to be estimated are orientation θ in the image plane Oxy and displacement magnitude d along the expansion ray, from two or several successive images. Afterwards the use of the formula (1) enables to compute the range Z from the corresponding point in space to the camera (Fig.4).

Determination of a spatio−temporal edge

The defined method allows the direct detection of such surface patches and the simultaneous estimation of parameters θ and d. It is based on an hypothesis testing scheme. It implies the definition of a maximum likelihood criterion as explained in [BOU 86]. In [BOU 86] the moving−edge model was merely a planar patch and the estimated motion information was a partial one, i.e. the edge−perpendicular velocity component, since no supplementary assumptions were considered. However the developed formalism can be applied to the more elaborate modeling described as above.
Let us consider the two following competing hypotheses. Given an elementary volume π in the space (x, y, t),

- either, there exists no spatio−temporal edge and the intensity function is modeled as *constant level + noise*, or more precisely defined as $c_0 + n$ for all points of π where c_0 is a constant and n a zero−mean gaussian noise;
- or there exists a surface patch splitting π in two sub−volumes π_1 and π_2; the intensity function is then defined as $c_1 + n$ in π_1 and $c_2 + n$ in π_2, with $c_1 \neq c_2$.

With each hypothesis is associated a likelihood function and the logarithmic ratio of these two functions is considered.

To maximize this ratio comes to maximize the following criterion, [BOU 86]:

$$CRV(p,\theta_j,d_j) = \left| \sum_{m\in M} a_j(m).f(p+m) \right| \qquad (2)$$

where M represents the index set corresponding to all the points of π, p is the current point in (x,y,t) space which the searched edge will be eventually referred to, a_j 's are coefficients depending only on predefined geometries of the surface patch, and $\{f(p+m)\}$ are the observed intensities within π. Assuming predefined geometric configurations means that parameter θ must take value from one set of some known pre-quantized angle measures. The same holds for parameter d_j. Formula yielding coefficients a_j 's are as follows:

$$a_j(m) = \left(\frac{n_2}{2(n_1+n_2)n_1}\right)^{\frac{1}{2}} \qquad if\, m \in \pi_1$$

$$a_j(m) = -\left(\frac{n_1}{2(n_1+n_2)n_2}\right)^{\frac{1}{2}} \qquad if\, m \in \pi_2$$

Computational scheme

The coefficients a_j 's are computed off line. Indeed the implemented algorithm is very close to a convolution procedure with a set of precomputed masks. First let us consider the case of two successive images for the computation of the function *CRV*. They will be called the current image and the next one.

For each point p of the current image and for each mask, corresponding to a given geometry (θ_j,d_j), the expression $CRV(p,\theta_j,d_j)$ is computed. In order to compute the convolution part concerned with the next image, this next image is scanned along an expansion ray from the initial position (x,y) of current point p in current image t, by successive steps out of the set $\{d_j\}$. The computed positions $(x,y)+d_j(r_x,r_y)$, where (r_x,r_y) is the unit vector defining the expansion ray direction, may not coincide with entire sample positions in the next image. Therefore in order to keep the same mask coefficients during the processing, interpolation are rather performed on intensity values if needed. The configuration $\hat{\theta},\hat{d}$ which maximizes the criterion is selected, provided that this maximum exceeds some predetermined threshold λ.

Then a spatio-temporal edge is said to be present at this point p. The range to the camera of the corresponding point in space can be deduced; with this range is associated a confidence coefficient given by $CRV(\hat{\theta},\hat{d})$.

Up to the *FOE* determination, we assume that the knowledge of the camera translation along with some pre-required calibration process provide us with a sufficiently accurate estimation of the *FOE* location. If no such strong statement can be formulated, some refinement procedure could be designed based for example on the technique described in [JAI 83]. An alternative could also be to consider instead of an expansion ray line something like an enlarged ray.

Mask shapes are all the same, namely for each geometrical configuration the union of square submasks in number equal to the number of images considered to compute the expression (2). For a given mask, all submasks are identical corresponding to the intersection of

the generalized cylindrical surface patch with each image plane, which is a straight segment of constant direction as previously assumed. As pointed out in [BOU 87], if only entire pixel positions arise from considered possible displacements, the computational cost can be considerably decreased and an implementation can be proposed the complexity of which is merely equivalent to some usual spatial gradient convolver.

Extension of the method to more than two images.

The algorithm is extended without difficulty to more than two images, as the remark concerning mask set lets it suppose since no inherent distinction is made between the case of two considered images and the case of several considered images. Merely, an elementary volume intersecting more than two image planes is taken into account and the criterion *CRV* is computed within this volume. Indeed to outline this property, expression (2) can be rewritten as follows :

$$CRV = \left| \sum_{s=0}^{\tau-1} \sum_{m \in M_s} a_j(m) f(p+m) \right| \qquad (3)$$

where t designates the current image, τ denotes the number of successive images considered out of the sequence, $M_s = M \cap I_s$, I_s being the sth image plane, (by definition the current image plane corresponds to $s = 0$).

Let us denote d_j^{t+s} the displacement magnitude along the expansion ray between time t and time $t+1+s$. Hence, considering τ (with $\tau > 2$) successive images to compute the depth map at time t only leads to introduce corresponding $\tau - 1$ displacement magnitudes d_j^{t+s} instead of only d_j. However the search space is not widened as the following relation occurs, as illustrated in Fig.5, assuming that the camera displacement is uniform from time t to time $t+\tau-1$:

$$\frac{d_j^{t+s}}{(1+s)D_{t+1+s}} = \frac{d_j^t}{D_{t+1}} \qquad \forall s = 1, \tau - 1$$

Let us precise that supplementary considered images could be as well on both temporal 'sides' of current image t as all subsequent to it. More complicated mask shape than the union of square submasks could also be straightforwardly considered within the same mathematical formalism.

Results

To test the method validity, experiments on synthetic images representing a parallelepiped have been undertaken. Camera motion is parallel to its axis of view and perpendicular to a parallelepiped plane (Fig.6) . Hence, all depicted object points are at the same relative depth with respect to the camera at each given time t. Therefore the standard deviation of the estimated depth has appeared to be an appropriate measure to represent the

algorithm behaviour. More precisely we have considered the relative error expressed as the ratio of the computed standard deviation and the true depth. A sequence of six images has been generated. Let us precise that the standard deviation of the estimated depth is only computed over 200 points corresponding to the largest values of the confidence coefficient.

Use of two images

First the case of only two successive images has been considered. Results are summarized in Tab.1. The measurement noise, due to sampling effect, decreases as the obstacle comes nearer the camera as illustrated in Fig.7. This is characterized by relative error values smaller. If the two images to be considered are current image t and successively image $t+u$ with $u = 2, 3, 4, 5$, we observe a constant decrease of the relative error. The first observation results from the fact that apparent edge motion becomes larger as the camera approaches the obstacle, and so its magnitude relatively gets more precisely determined. Moreover the percentage of misdetermined moving edges whose displacement magnitude is inferior to the search step also becomes lower. The second one is similarly encountered in stereovision system. For a given sampling error in matching, triangulation accuracy increases as the camera positions in space are more distant from each other. However, there is no use to consider larger temporal distances between two processed images, since necessary assumptions for the algorithm are then no more valid and matching difficulty increases.

Use of more than two images

The method has been successively applied to simultaneously 3, 4, 5, and 6 images to compute the depth map referred to the current image. Obtained results are reported in Tab. 2. The relative error decreases with the number of considered images, almost in the same way as when larger temporal distances between two processed images were chosen.

Experiments with noisy images

To strengthen the first results, the same experiments were carried out on images corrupted by gaussian noise of standard deviation successively equal to 10, 20, 30, 40, 50% of the maximum intensity of pixels in the image. The results are given for a standard deviation of 50% in Tab.3. The same trends as previously can be verified, but the improvement is slightly less noticeable when several images are simutaneously considered than when the two processed images are conveniently distant in time. Practically it is nevertheless easier to use simultaneously several images than to determine the best step in time between two images to be processed according to camera motion and object depth, especially when there are several obstacles of different range.

Conclusion

We have investigated the use of a single camera for obstacle avoidance purpose in the case of mobile robots or manipulators whose motion is approximatively known. A matching scheme of interest points through successive images allows to deduce the range of possible obstacles, but combining a correlation algorithm with this interest operator is not robust enough. To solve this issue, a new approach is considered based upon the modeling of a spatio-temporal edge, for which the camera motion knowledge can be integrated. It leads to the computation of a likelihood ratio criterion implemented by convolving the considered images with some appropriate set of masks. Once a moving edge has been detected and its parameters simultaneously estimated, the range of the corresponding object element in space can be directly inferred. This algorithm can be easily applied to more than only two successive images, which enables to improve the method accuracy.

Acknowledgements

The first author would like to thank H.P. Moravec for his welcome in his laboratory at CMU where the first part of this work was done, A. Elfes and L.H. Matthies for helpful discussions.

REFERENCES

[BAR 80] S.T. BARNARD, W.B. THOMPSON, *Disparity analysis of images*, IEEE Trans. on PAMI, Vol.2, No 4, July 1980, pp.323-340

[BHA 86] S. BHARWANI, E. RISEMAN, A. HANSON, *Refinement of environmental depth maps over multiple frames*, IEEE Workshop on Motion: Representation and Analysis, Charleston, South Carolina, May 1986, pp.73-80

[BOL 85] R.C. BOLLES, H.H. BAKER, *Epipolar-plane image analysis: a technique for analyzing motion sequences*, Proc. of the 3rd IEEE Workshop on Computer Vision: Representation and Control, Bellaire, Michigan, Oct. 1985, pp.168-178

[BOU 86] P. BOUTHEMY, *Determining displacement fields along contours from image sequences*, Proc. Conf. Vision Interface ' 86, Vancouver, May 1986, pp.350-355

[BOU 87] P. BOUTHEMY, *A maximum likelihood framework for determining moving edges in image sequences*, Rapport de recherche INRIA Rennes, No 696, Juin 87, 42 pp.

[ESP 86] B. ESPIAU, P. RIVES, *Closed loop recursive estimation of 3d features*

for a mobile vision system, IEEE Conf. on Robotics and Automation, Raleigh, USA, March 1987

[JAI 83] R. JAIN, *Direct computation of the focus of expansion*, IEEE Trans. on PAMI, Vol.5, No 1, Jan. 1983, pp.58–64

[MAR 86] L. MARCE, *Sur l'utilisation des séquences multi–images en robotique*, Publication interne IRISA No 293, Avril 1986

[MOR 81] H.P. MORAVEC, *Obstacle avoidance and navigation in the real world by a seeing rover robot*, Ph.D. Thesis, Stanford University, Sept. 1980

[THO 84] C.E. THORPE, *FIDO : vision and navigation for a robot rover*, Ph.D Thesis, Dep. Computer Science, Carnegie–Mellon Univ., CMU–CS–84–168, Dec. 1984

[WIL 81] T.D. WILLIAMS, *Computer interpretation of a dynamic image from a moving vehicle*, COINS Technical Report 81–22, Univ. of Massachusetts, Computer and Information Science, Amherst, May 1981

Tableau 1.

Images	*Real distance*	*Estimated distance*	*Relative error*
u = 1	22.1	21.7	12%
	20.1	22.9	13%
	18.1	16.3	7%
	16.1	17.8	9%
	14.1	15.1	5%
u = 2	20.1	20.0	7%
	18.1	17.3	5%
	16.1	15.5	4%
	14.1	14.9	4%
u = 3	18.1	16.7	6%
	16.1	15.7	3%
	14.1	13.8	3%
u = 4	16.1	15.3	4%
	14.1	13.9	3%
u = 5	14.1	13.6	3%

Tableau 2.

Images	*Real distance*	*Estimated distance*	*Relative error*
2 successive images	22.1	21.7	12%
	20.1	22.9	13%
	18.1	16.3	7%
	16.1	17.8	9%
	14.1	15.1	5%
3 successive images	20.1	21.8	6%
	18.1	17.8	7%
	16.1	16.9	5%
	14.1	15.1	5%
4 successive images	18.1	16.8	6%
	16.1	16.7	4%
	14.1	14.5	3%
5 successive images	16.1	15.9	3%
	14.1	14.4	3%
6 successive images	14.1	13.9	3%

Tableau 3.

Images	*Real distance*	*Estimated distance*	*Relative error*
u = 1	22.1	21.4	32%
	20.1	20.6	30%
	18.1	16.0	24%
	16.1	18.0	24%
	14.1	14.6	16%
u = 2	20.1	19.5	17%
	18.1	16.9	14%
	16.1	15.6	11%
	14.1	14.5	10%
u = 3	18.1	16.4	11%
	16.1	15.5	9%
	14.1	13.6	7%
u = 4	16.1	15.1	8%
	14.1	13.7	6%
u = 5	14.1	13.4	6%

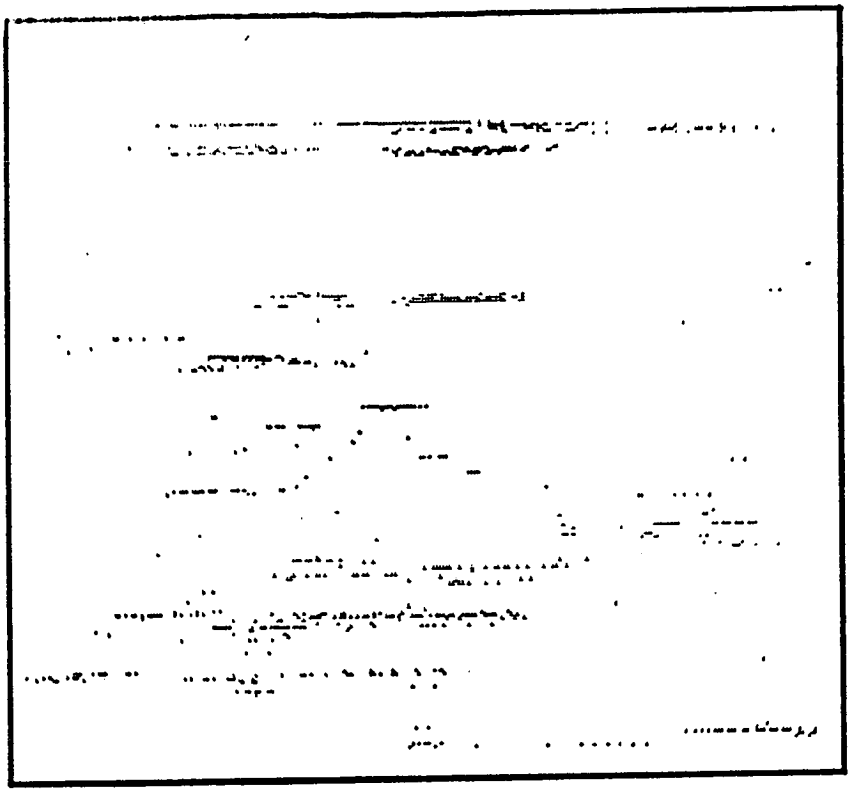

Figure 1. Stability test of Moravec's operator on 125 successive images

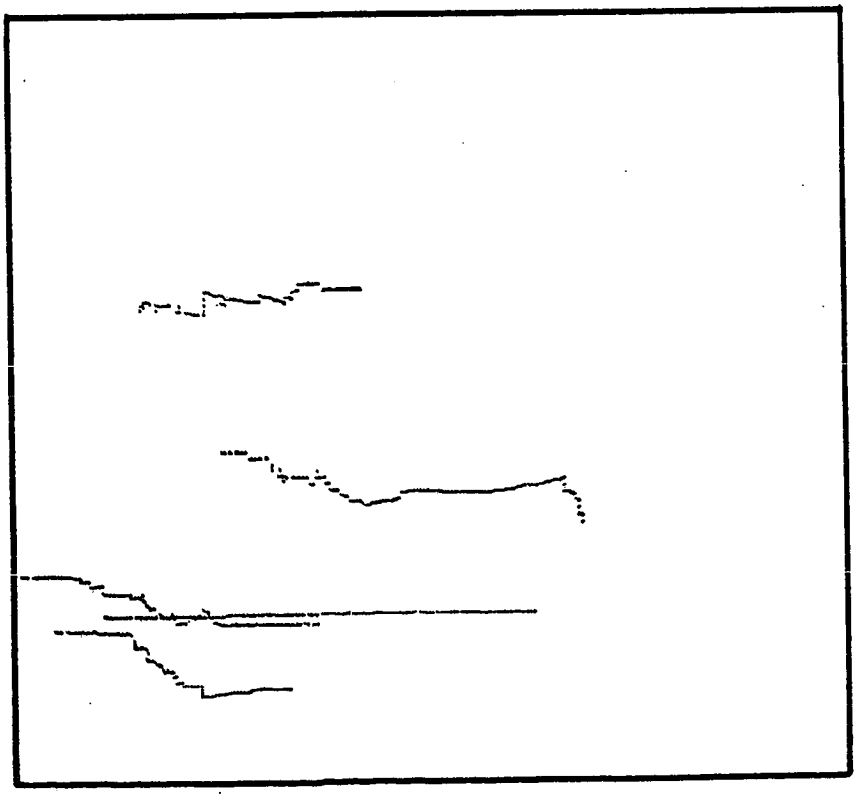

Figure 2. Following of 5 first interest points through the 125 images

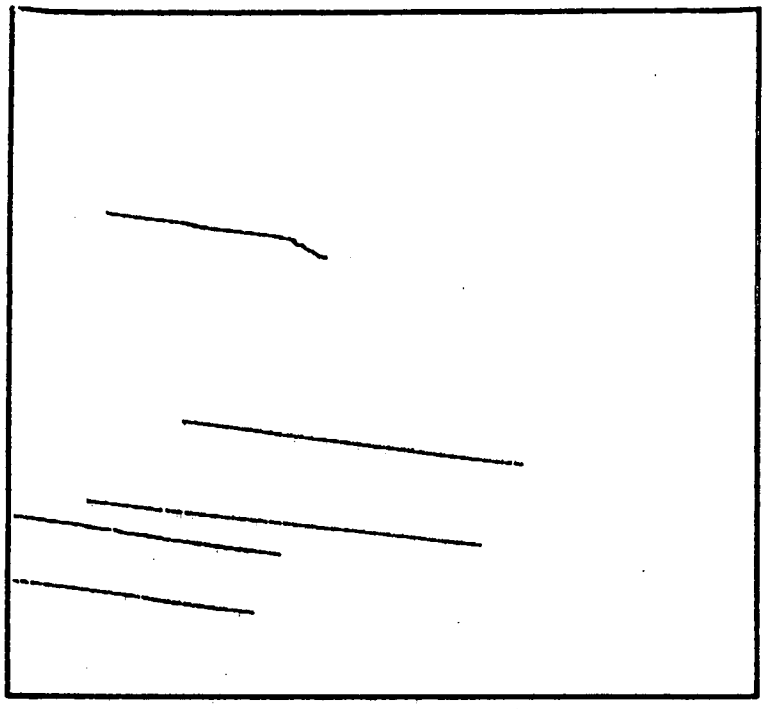

Figure 3. Following interest points with motion knowledge

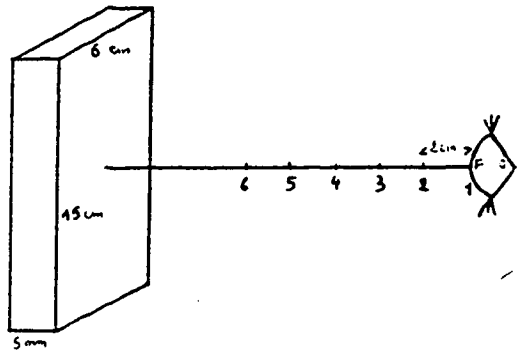

Figure 6. Experimental conditions

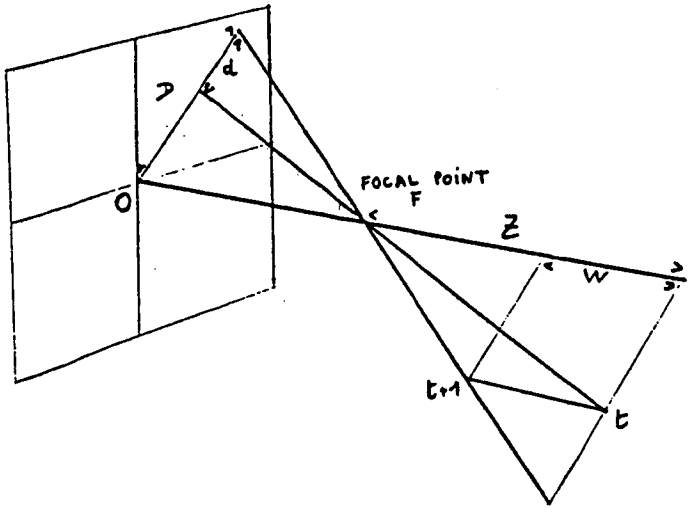

Figure 4. Displacement of one image point

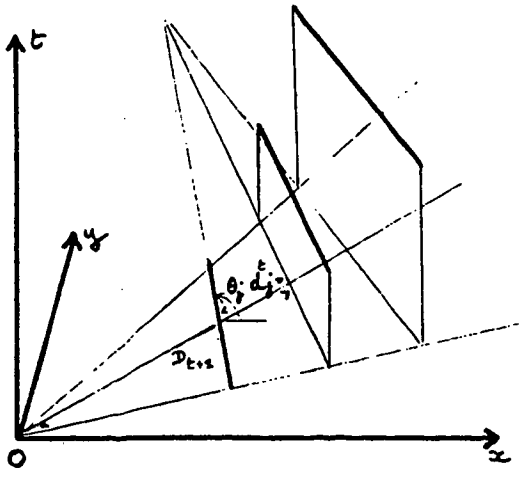

Figure 5. Modelling a spatio-temporal contour element

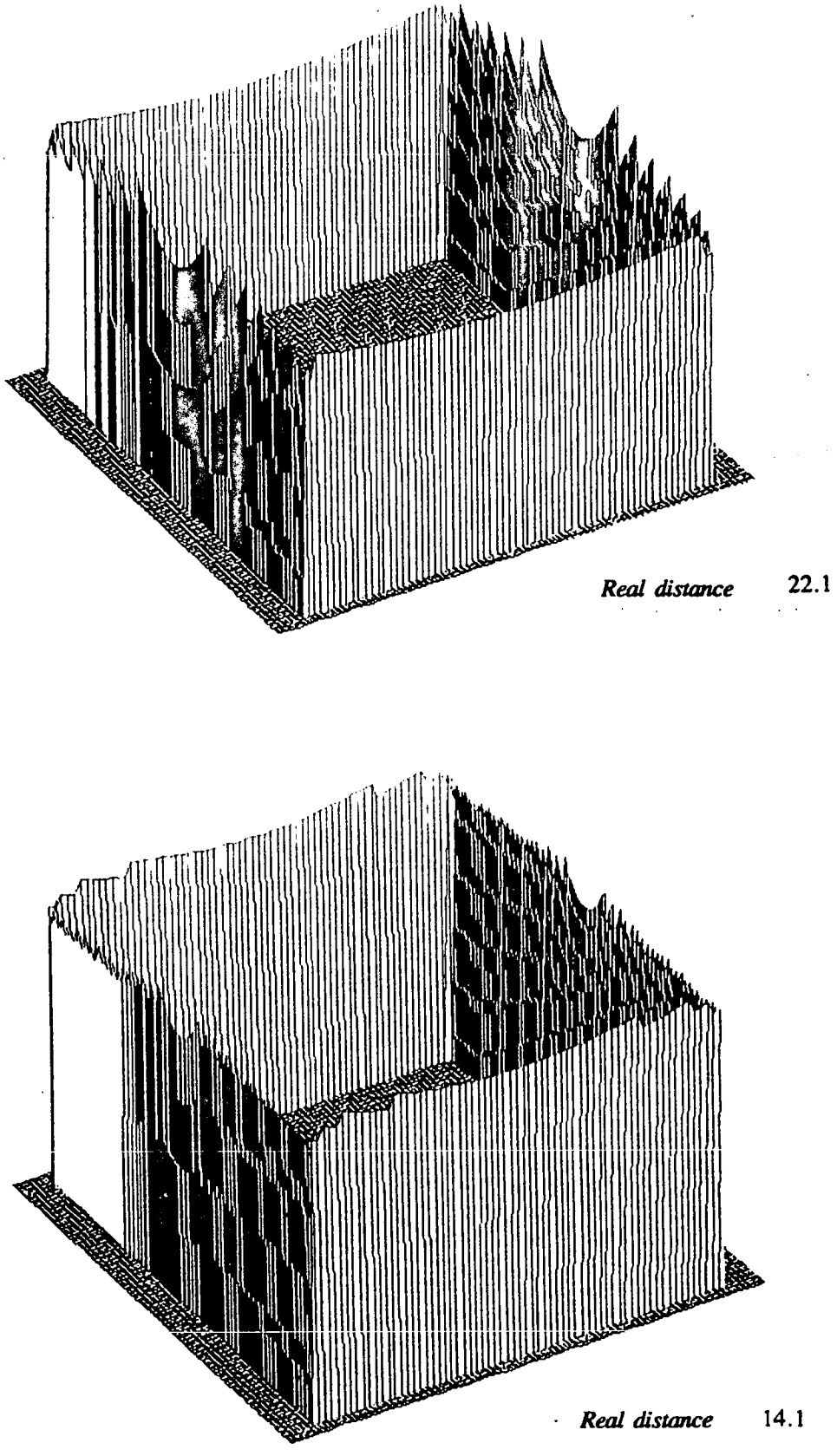

Figure 7. Measurements noise decreases as the obstacle gets nearer

u = 1

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique