Dpto. de Informática e Ingeniería de Sistemas
Universidad de Zaragoza
C/ María de Luna num. 1
E-50018 Zaragoza
Spain

Internal Report: 1995-V01

Motion and Structure from Significant Segments in Man Made Environments

C. Sagüés, J.J. Guerrero

If you want to cite this report, please use the following reference instead:
Motion and Structure from Significant Segments in Man Made Environments, C. Sagüés, J.J. Guerrero C., *2nd IFAC Conference on Intelligent Autonomous Vehicles*, pages 337-342, 1995.

This work was partially supported by project TAP-94-0390 of the Comisión Interministerial de Ciencia y Tecnología (CICYT).

MOTION AND STRUCTURE FROM SIGNIFICANT SEGMENTS IN MAN MADE ENVIRONMENTS

C. Sagüés & J.J. Guerrero *

Abstract

An algorithm to determine the camera motion and the structure in a 3D man made environment is proposed. The algorithm is based on straight edges extracted in the image and some plentiful geometrical relations between them (verticality, horizontality, parallelism, perpendicularity). In this paper the computation of motion and structure using these constraints is tackled. The information of the tips of the straight edges are also used to determine motion. Some partial results using two images are now available. Its application to a mobile robot running in man made environments could be considered.

Vision; visual motion; robot vision; motion estimation; structure constraints.

1 INTRODUCTION

Many mobile robots are able to execute tasks in an indoor environment, where the ground is assumed to be horizontal and robot localization is obtained using known landmarks. However, when there is no assumption like these ones, powerful perception systems to estimate 3D motion of the robot and scene structure are needed. Vision is the sensor more broadly used.

Some methods to recover structure and motion from points and/or lines have been widely studied and revised in the last years [6]. We have treated that problem by using lines with a tip [5]. Working with lines, it is well known that three images at least are needed in order to determine both the camera motion and the 3D estimation of the scene structure. When noise is present, good solutions are difficult to extract but the best solution is obtained with many features and some global nonlinear optimization [11]. Thus, some qualitative information must be used in order to obtain more robust methods in the presence of noise. In [8] the use of polyhedral constraints on surfaces for the computation of 3D structure and motion from 2D visual motion, is described.

In this paper a camera without geometric map of the scene is used. Only some assumptions about the general aspect of the environment are taken into account. The first main assumption is that the segments extracted are mainly vertical and horizontal. In man made environments (indoor or outdoor between buildings, roads, ...) the main static lines have many geometrical constraints (there are many relations of parallelism and perpendicularity between horizontal lines). As in [7] these assumptions are exploited, but in our work the motion is globally computed. The vertical cue is considered in other works [10] trying to provide relevant qualitative information about the structure of the scene to recover more robustly structure and motion.

An algorithm to determine the 3D rigid motion of the camera and the 3D location of segments in the robot environment, using geometrical constraints about the structure, is presented. An initial motion guess is needed, that could be obtained from other sensors or using the expected motion.

2 EXTRACTION AND RECTIFICATION OF SEGMENTS

The first step is the *straight edge extraction* from each image (Fig. 1). The straight edges are obtained as proposed by [1]. The first step in this pro-

*Departamento de Ingeniería Eléctrica e Informática, Centro Politécnico Superior, UNIVERSIDAD DE ZARAGOZA, María de Luna 3, E-50015 ZARAGOZA, SPAIN, Phone 34-76-517274, Fax 34-76-512932, email: csagues@cc.unizar.es

cedure is the extraction of spatial gradients to segmentate the image. Pixels are grouped into regions of similar direction of brightness gradient, with the gradient magnitude larger than a threshold. From here, line-support regions (LSR) containing all information of the straight contours are available. A planar brightness surface is fitted to the LSR by a least-squares approach, predicting the brightness (E) as a function of the image coordinates. In this fitting, a weighting norm proportional to the gradient magnitude is considered. The straight line is obtained as the intersection of this brightness plane and an horizontal plane of mean brightness in the LSR. The parameters of the line in the image are obtained with subpixel accuracy.

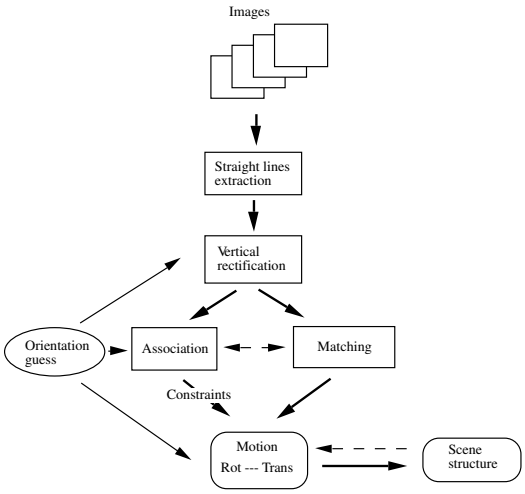

Figure 1: General algorithm

Using this edge detector we obtain, in addition to the geometrical parameters of the edges, some attributes related with their brightness (contrast, average gray level, steepness) and some quality attributes (deviation from straightness). These attributes provide information very useful to identify and to associate them.

Extracted edges are selected in function of its length, its contrast and its deviation from straightness in order to have few but good edges. Thus, the following steps (association and matching) are easier made. Oriented segments in function of the sign of the contrast are finally given, in such a way that the two tips are identified.

Initially a *vertical rectification* of the image segments is carried out. A rotation with two degrees of freedom is obtained, in such a way that the normals of the supposed vertical segments appear horizontal.

The rotation for the rectification $\mathbf{R}_{rc} = Rot(z, \phi_z^r), Rot(x, \psi_x^r)$ is obtained using the vertical segments, to solve the angles ϕ_z^r, ψ_x^r that minimize:

$$\sum_{vertical} [(0,1,0) \cdot \mathbf{R}_{rc}\mathbf{n}^v]^2 \qquad (1)$$

being $\times$ and $\cdot$ the cross and dot product of vectors and $\mathbf{n}^v$ the normal vector of the plane of projection of each supposed vertical line in the camera reference system.

The rectification is made rotating the features in the camera reference system according to $\mathbf{R}_{rc}$ in such a way that every line supposed to be vertical appear in the image as parallel and vertical.

This rectification will be only used to make easy the two following process (matching and association) that are based on some heuristic considerations in the image plane. As can be seen below the motion determination algorithm solve the 3D rotation. It does not take into account the two degrees of freedom of rotation reduced using rectified images because the rotation based only on vertical segments is not well enough conditioned. When a good vertical direction is available the motion algorithm considerably improves the results.

To circumvent the problem of previous determination of vertical image segments, the use of a Hough transform is actually being considered. It searches in directions close to the vertical guessed, assuming the maximum is in this direction (most of the segments will be vertical).

3 MATCHING AND ASSOCIATION OF SEGMENTS

With straight segments extracted in rectified images two processes work in parallel: *matching* and *association* (Fig. 1). The first one, makes the match between target features along the sequence of images. The second one associates segments in each image, in function of hypothesis of verticality, horizontality, parallelism and perpendicularity.

The correspondence problem has been treated by tracking segments in the image, using a Kalman filter. A nearest neighbor tracking approach as in [2],

has been developed. However, besides the classical location values, two image bright attributes of the segment are used in the tracking and matching process [4]. These attributes are the average gray level and the mean contrast. They allow to match segments using not only the geometrical information but also the intensity information, that is in many cases more relevant and selective. Besides that, the matching using these bright attributes can be made nearly in parallel to the geometrical matching adding a little computational overhead.

On the other hand, the bright parameters are crucial to match segments when neither the structure nor the camera motion are known, because geometrical constraints are only valid locally and they can not be imposed in a non heuristic way.

The association process makes a first classification of the segments as vertical and horizontal, using the guessed orientation. The vertical lines vanish near to the vertical vanishing point. The horizontal lines never cross the corresponding line of the horizon. Other segments in the image are not considered in the association step.

After that, hypothesis of relation of parallelism and perpendicularity between horizontal lines is established. These relations can be extracted for close segments belonging to the same surface. Therefore, geometrical information in the image and the brightness coherence are used to determine them.

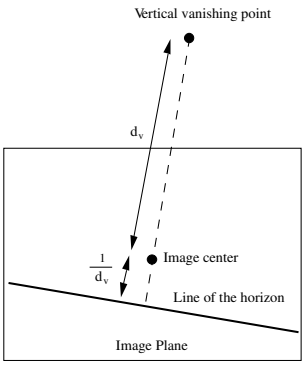

Figure 2: The line of the horizon can be easily obtained as a function of the vertical vanishing point

The final aim of this process is to provide in each image, the vanishing point correspondent to each segment. Normally in man made environment vertical and horizontal lines are dominant. Thus if the vertical direction is determined, the horizontal lines must vanish in its intersection with the line of the horizon. When the vertical vanishing point is available, the line of the horizon is perpendicular to the line which links the image center with the vertical vanishing point. The distance from the origin to this horizon line is the inverse distance from the center to the vanishing point (Fig. 2). Supposing a rectified image the line of the horizon will be horizontal and it will be on the image center.

4 MOTION AND STRUCTURE DETERMINATION

The process goes on with the motion and structure determination. Methods to recover structure and motion based on points usually work better than methods based on lines, but lines are easier to extract and to match than points. Recently some works have been proposed to use segments instead of only its line support in motion and structure determination [12] with good results. The assumption used is that two matched line segments contain the projection of a common portion of the corresponding segment in space.

On the other hand, the well known problem of coupling between translations along an axis with rotations around its perpendicular axis in the image plane can not be solved without depth information (structure knowledge) [11].

In this paper, motion is extracted using the tips of the segments, that are easily extracted and matched. A nonisotropic noise model for the location of the tips that takes into account the noise of the line in the image, is considered (is more probable the real tip moving along the line than across it).

Besides that, the constraints in the orientation of segments provided by the association step are considered in order to have information of relative depth. This is very important because a little error in the orientation obtained brings about large errors in the structure and translation determination.

4.1 Camera model and segment representation

The classical pinhole camera model is assumed. The Z axis is aligned with the focal axis and the focal length is considered to be the unit. A segment in the image is represented using four parameters (Fig. 3). Two of them are used to represent also the infinite line in the image. These are the ϕ and θ angles defining the normal of the projecting plane of the line $\mathbf{n}$, as: $\mathbf{n} = (cos\phi cos\theta, sen\phi cos\theta, -sen\theta)^T$

A third parameter ψ_s defines the location along line of the start tip of the segment, in such a way that the unit vector in the camera reference system that points in the direction of the start tip of the segment is:

$$\mathbf{a}_s = \begin{pmatrix} cos\phi sen\theta cos\psi_s + sen\phi sen\psi_s \\ sen\phi sen\theta cos\psi_s - cos\phi sen\psi_s \\ cos\theta cos\psi_s \end{pmatrix}$$

The parameter ψ_e defines the location of the end point of the segment. The $\mathbf{a}_e$ unit vector of the end tip, in the camera reference system can be given similarly.

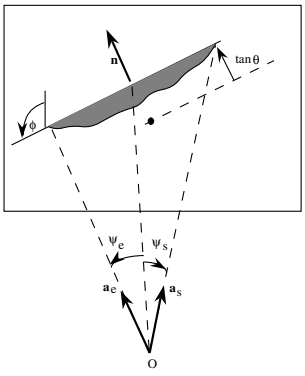

Figure 3: Segment representation for motion computation

We take ϕ $[-\pi, +\pi]$ in such a way that the normal to the plane of projection $\mathbf{n}$ points in the direction of the spatial brightness gradient in image, from dark to light. The angle θ takes values from $-\frac{\pi}{2}$ to $+\frac{\pi}{2}$. Normally using real cameras that have a small field of view, the angle θ will be small for all lines that appear in image. The angles ψ_s and $\psi_e \in [-\frac{\pi}{2}, +\frac{\pi}{2}]$, but using real cameras will also be small. Therefore, we are far from the singularities when using segments that can appear in the image.

4.2 Motion determination

Rigidity is assumed in our motion determination algorithm. By taking the first camera reference system as the basic reference system, the matrix of rotation of the camera from the first to the second camera reference system is named $\mathbf{R}_{12}$. The vector $\mathbf{t}_{12}$ expresses the translation of the camera from the first camera location to the second.

The problem is to estimate motion of the camera given a discrete description of the image deformation from one image to the next. Corresponding segments in two images are used as a description of the deformation in the image due to motion. The image measurements are complemented by a measurement of their uncertainty. The uncertainty of location of the segment is represented by a covariance matrix. It is composed of the line orientation covariance σ^2_ϕ, the line location covariance σ^2_θ and two tip location covariances (along line) $\sigma^2_{\psi_s}$, $\sigma^2_{\psi_e}$ which are supposed to be the biggest.

From two corresponding points in two images the epipolar constraint can be formulated. It constraints the translation vector to be coplanar with the two vectors in the direction of the projection line of the point in the two images. It can be formulated at the ideal case (expressed in the first camera reference system), for each tip (i), as:

$$\mathbf{a}_{1i} \cdot (\mathbf{t}_{12} \times \mathbf{R}_{12}\mathbf{a}_{2i}) = 0 \qquad (2)$$

where the subscript 1 or 2 indicates the first or second image frame.

Besides that, in the association step an hypothesis of direction has been made for several segments in each image. Therefore, a second constraint can be considered which affects to the direction of these segments extracted and to the camera rotation. Thus, for the ideal case, it can be expressed as:

$$\mathbf{d}_{1j} \cdot \mathbf{R}_{12}\mathbf{n}_{2j} = 0\ ;\ \ \mathbf{d}_{2j} \cdot \mathbf{R}^T_{12}\mathbf{n}_{1j} = 0 \qquad (3)$$

being $\mathbf{d}_{1j}$ and $\mathbf{d}_{2j}$ the hypothetic direction in each image corresponding to the j-th line.

It is clear that in the presence of noise there is not a set of motion parameters $\mathbf{R}_{12}$ and $\mathbf{t}_{12}$ that can satisfy these constraints for all segments. So, we try to find a correction for the given observations in such a way that the tips satisfy the epipolar constraint (2), and the infinite lines satisfy the direction guessed (3). This correction is minimized

taking into account the weight of different errors. A constrained least-squares is formulated, that can be solved using Lagrangian multipliers [3]:

$$\begin{aligned} J_d = & \sum_{i,j} \delta\mathbf{a}_{1i}^T \mathbf{\Gamma}_{\delta\mathbf{a}_{1i}}^{-1} \delta\mathbf{a}_{1i} + \delta\mathbf{a}_{2i}^T \mathbf{\Gamma}_{\delta\mathbf{a}_{2i}}^{-1} \delta\mathbf{a}_{2i} \\ + & \lambda_i(\mathbf{a}_{1i} + \delta\mathbf{a}_{1i}) \cdot (\mathbf{t}_{12} \times \mathbf{R}_{12}(\mathbf{a}_{2i} + \delta\mathbf{a}_{2i})) \\ + & \delta\mathbf{n}_{1j}^T \mathbf{\Gamma}_{\delta\mathbf{n}_{1j}}^{-1} \delta\mathbf{n}_{1j} + \lambda_{1j}\mathbf{d}_{2j} \cdot \mathbf{R}_{12}^T(\mathbf{n}_{1j} + \delta\mathbf{n}_{1j}) \\ + & \delta\mathbf{n}_{2j}^T \mathbf{\Gamma}_{\delta\mathbf{n}_{2j}}^{-1} \delta\mathbf{n}_{2j} + \lambda_{2j}\mathbf{d}_{1j} \cdot \mathbf{R}_{12}(\mathbf{n}_{2j} + \delta\mathbf{n}_{2j}) \end{aligned} \quad (4)$$

where $\mathbf{\Gamma}$ correspond with the covariance matrix of the observations uncertainty (see appendix).

Setting the derivatives of this expression equal to zero for the unknowns, and solving the set of equations to eliminate the local variables, an equivalent expression depending only on the five motion parameters (the translation is determined with an scale factor) would be obtained.

There are some nonlinearities and the results are too complicated to do anything useful with them. As in [9] the second order terms of the noise are eliminated and then the derivatives are taken. Besides that, the noise of the line is considered decoupled from the noise of the tip, because normally tips move mainly along line. We prefer to control the approximations to be made, instead of taking correct but unsolvable equations. Therefore the proposed expression to minimize in function of the motion parameters is:

$$\begin{aligned} J_d = & \sum_{i,j} \frac{(\mathbf{a}_{1i} \cdot (\mathbf{t}_{12} \times \mathbf{R}_{12}\mathbf{a}_{2i}))^2}{\mathbf{A}_{1i}^T \mathbf{\Gamma}_{\delta\mathbf{a}_{1i}} \mathbf{A}_{1i} + \mathbf{A}_{2i}^T \mathbf{\Gamma}_{\delta\mathbf{a}_{2i}} \mathbf{A}_{2i}} \\ + & \frac{(\mathbf{d}_{2j} \cdot \mathbf{R}_{12}^T \mathbf{n}_{1j})^2}{(\mathbf{R}_{12}\mathbf{d}_{2j})^T \mathbf{\Gamma}_{\delta\mathbf{n}_{1j}} \mathbf{R}_{12}\mathbf{d}_{2j}} \\ + & \frac{(\mathbf{d}_{1j} \cdot \mathbf{R}_{12}\mathbf{n}_{2j})^2}{(\mathbf{R}_{12}^T\mathbf{d}_{1j})^T \mathbf{\Gamma}_{\delta\mathbf{n}_{2j}} \mathbf{R}_{12}^T\mathbf{d}_{1j}} \end{aligned} \quad (5)$$

where:

$$\mathbf{A}_{1i} = \mathbf{t}_{12} \times \mathbf{R}_{12}\mathbf{a}_{2i} \; ; \quad \mathbf{A}_{2i} = \mathbf{R}_{12}^T(\mathbf{t}_{12} \times \mathbf{a}_{1i})$$

Iterated methods [3] allow to solve the motion with an scale factor for translations, when an initial guess is available. This scale factor can be solved using some knowledge about the structure (like normal height of doors or ceiling) or some knowledge about total translation from odometric sensors.

4.3 Structure determination

From the camera motion, the structure is easily obtained by triangulation, obtaining each 3D line as the intersection of its two projection planes. The direction of each line could be obtained as:

$$\frac{\mathbf{n}_{1j} \times \mathbf{R}_{12}\mathbf{n}_{2j}}{||\mathbf{n}_{1j} \times \mathbf{R}_{12}\mathbf{n}_{2j}||} \quad (6)$$

A tip (i) of the j-th line could be obtained as the intersection of the tip projection line in the first image with the projection plane of the line in the second image:

$$\frac{\mathbf{t}_{12} \cdot \mathbf{R}_{12}\mathbf{n}_{2j}}{\mathbf{a}_{1i} \cdot \mathbf{R}_{12}\mathbf{n}_{2j}} \, \mathbf{a}_{1i} \quad (7)$$

However when the projection plane of a line is nearly parallel to the translation, bad results of structure are obtained because the two projecting planes of the line are parallel. In this case is better to reconstruct the segment using its tips directly. The distance from the origin of the first frame to the 3D tip can be evaluated, and therefore the 3D location of the tip will be:

$$\frac{(\mathbf{t}_{12} \times \mathbf{R}_{12}\mathbf{a}_{2i}) \cdot (\mathbf{a}_{1i} \times \mathbf{R}_{12}\mathbf{a}_{2i})}{||\mathbf{a}_{1i} \times \mathbf{R}_{12}\mathbf{a}_{2i}||^2} \, \mathbf{a}_{1i} \quad (8)$$

5 EXPERIMENTAL RESULTS

Some experiments have been made with two images of our laboratory (see Fig. 4 and Fig. 5). Only 11 segments remain matched after length, gradient and deviation from straightness filtering (see Fig. 6). At the moment the association step is made manually and the tips which are not robust enough (occlusions and overlappings) are not considered for motion computation.

Figure 4: First image of the laboratory to experiment

Figure 5: Second image.

When constraints about line orientation are not used, the algorithm has many problems to disambiguate rotations around the vertical direction from translations along the horizontal axis parallel to the image plane. Using the direction constraints, with lines nearly parallel to the focal axis, the results improve.

A very good guess of the camera rotation is needed to make the algorithm to converge. In Fig. 7 the top view in the first camera reference system of a reconstruction of the scene is shown.

Has been observed that the translation obtained is usually deviated towards the focal axis. The rotation determination is a very critical step, because a little error in rotation makes translation and structure to degenerate. In order to have useful results for a indoor mobile robot a way to determine very accurately the rotations is needed.

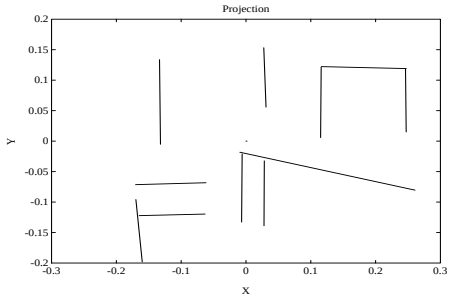

Figure 6: Reconstructed segments projected in the first camera location

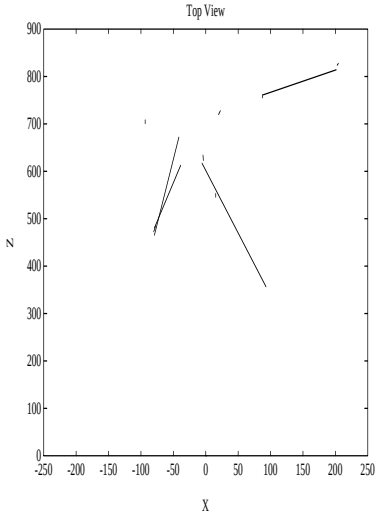

Figure 7: Reconstruction from the motion computed. Top view

6 CONCLUSIONS AND FUTURE WORK

A method to obtain the camera motion and the scene structure using straight edges has been presented. It has been assumed that they are vertical and horizontal and that some geometrical relations between them are satisfied. This assumption is normally achieved in many man made environments.

The motion determination algorithm uses lines and the tips of the segments with a noise model that allow them to move mainly along the line. In order to improve its robustness the geometrical constraints of the lines are used leading to a partial correction of the coupling between rotation and translation in the structure and motion problem. When motion is determined the localization of the features is obtained.

The experiments have shown the difficulties of the motion and structure paradigm to solve the problem if no good rotation information is available, more information of structure is given or more images are taken and fused.

When more than two images are taken an estimation of the 3D location of the lines will be available. If the line has appeared in more than two images the estimation of its 3D location could be used establishing a 2D-3D relation improving the

scene reconstruction and the camera location.

APPENDIX

Small errors are assumed and therefore we can consider the first order approximation to be valid to obtain the covarianze of the errors of the projecting vectors in function of the segment noise. Therefore, the covarianze matrix of the tip and the line vectors can be expressed as:

$$\mathbf{\Gamma}_{\delta \mathbf{a}_i} = \mathbf{J}^{\delta \mathbf{a}_i}_{\psi_i,\theta,\phi} \begin{bmatrix} \sigma^2_{\psi_i} & 0 & 0 \\ 0 & \sigma^2_{\theta} & 0 \\ 0 & 0 & \sigma^2_{\phi} \end{bmatrix} \mathbf{J}^{\delta \mathbf{a}_i}_{\psi_i,\theta,\phi}{}^T$$

$$\mathbf{\Gamma}_{\delta \mathbf{n}_j} = \mathbf{J}^{\delta \mathbf{n}_j}_{\theta,\phi} \begin{bmatrix} \sigma^2_{\theta} & 0 \\ 0 & \sigma^2_{\phi} \end{bmatrix} \mathbf{J}^{\delta \mathbf{n}_j}_{\theta,\phi}{}^T$$

where:

$$\mathbf{J}^{\delta \mathbf{a}_i}_{\psi_i,\theta,\phi} = \begin{bmatrix} -o_x & n_x cos\psi_i & -a_y \\ -o_y & n_y cos\psi_i & a_x \\ -o_z & n_z cos\psi_i & 0 \end{bmatrix}$$

$$\mathbf{J}^{\delta \mathbf{n}_j}_{\theta,\phi} = \begin{bmatrix} -n_x tan\theta & -n_y \\ -n_y tan\theta & -n_x \\ -cos\theta & 0 \end{bmatrix}$$

being:

$$\begin{bmatrix} n_x & o_x & a_x \\ n_y & o_y & a_y \\ n_z & o_z & a_z \end{bmatrix} = Rot(z,\phi)Rot(y,\theta)Rot(x,\psi_i)$$

Acknowledgments

This work was partially supported by project TAP-94-0390 of the Comisión Interministerial de Ciencia y Tecnología (CICYT).

References

[1] J.B. Burns, A.R. Hanson, and E.M. Riseman. Extracting straight lines. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8(4):425–455, 1986.

[2] R. Deriche and O. Faugeras. Tracking line segments. In *First European Conference on Computer Vision*, pages 259–268, Antibes, France, 1990.

[3] P.E. Gill, W. Murray, M.A. Saunders, and M.H. Wright. Constrained nonlinear programming. In G.L. Nemhauser, A.H.G. Rinnoy Kan, and M.J. Todd, editors, *Handbooks in Operations Research and Management Science. OPTIMIZATION*, volume 1, pages 171–210. Nort-Holland, Amsterdam, 1989.

[4] J.J. Guerrero and J.M. Martínez. Determination of corresponding segments by tracking both geometrical and brightness information. In *International Conference on Advanced Robotics*, Barcelona, September 1995. Submitted.

[5] J.J. Guerrero, C. Sagüés, and A. Lecha. Motion and structure from straight edges with tip. In *IEEE International Conference on Systems, Man and Cybernetics*, San Antonio, USA, Oct 1994.

[6] T.S. Huang and A. N. Netravali. Motion and structure from feature correspondences: A review. *Proceedings of the IEEE*, 82(2):252–268, 1994.

[7] X. Lebégue and J.K. Aggarwal. Significant line segments for an indoor mobile robot. *IEEE Transactions on Robotics and Automation*, 9(6):801–815, 1993.

[8] D.W. Murray and B.F. Buxton. *Experiments in the Machine Interpretation of Visual Motion.* The MIT Press, Massachusetts, 1990.

[9] Minas E. Spetsakis. Models of statistical visual motion estimation. *CVGIP: Image Understanding*, 60(3):300–312, 1994.

[10] T. Viéville, E. Clergue, and P. Facao. Computation of ego-motion and structure from visual and inertial sensors using the vertical cue. In *Fourth International Conference on Computer Vision*, pages 591–598, Berlin, May. 1993.

[11] J. Weng, N. Ahuja, and T.S. Huang. Optimal motion and structure estimation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15(9):864–884, 1993.

[12] Z. Zhang. Estimating motion and structure from correspondences of line segments between two perspective images. Rapport

de recherche RR-2340, I.N.R.I.A., Sophia-Antipolis, France, 1994.