3D MODELING OF STRUCTURED SCENES THROUGH BINOCULAR STEREO VISION

Angeles López[*], *Filiberto Pla*[**] *and José Ribelles*[**]

[*]Dept. Ingeniería y Ciencia de los Computadores, Universitat Jaume I, E-12080 Castelló, SPAIN
[**]Dept. Lenguajes y Sistemas Informáticos, Universitat Jaume I, E-12080 Castelló, SPAIN
{lopeza, pla, ribelles}@uji.es

ABSTRACT

Structured scenes are usually composed of man-made objects, often with lack of texture, which makes stereo correspondence more difficult. In this paper, we present a new matching algorithm which directly gives a 3D representation of structured scenes, by computing region correspondences and occlusions cooperatively. The new algorithm controls the minimization of a function for each region in several steps, in order to cooperatively guide the process towards a global convergence. We analyze its theoretical computational cost, and we present some experimental results to validate the scope of this approach and the possibilities of a generalization of the method to planar and other surfaces.

1. INTRODUCTION

In this paper, a new method for stereo vision is presented, which aims to produce a 3D representation of a scene from two views. The aim of this approach is to obtain a 3D model of the environment, useful in several applications: path planning of mobile robots and manipulator robots, 3D modeling of scenes and objects for virtual reality applications, product presentations, etc.

We focus on structured scenes, mainly indoors, which are composed of man-made objects. These scenes usually contain planar surfaces, and are often poorly textured. Stereo vision methods are capable of providing a depth map of the scene, which is denser when surfaces are well-textured. In order to obtain a useful 3D representation (i.e. a polygonal mesh model), a range segmentation method must be applied to the range information. In addition, both the stereo vision method and the range data segmentation method have to be capable of detecting and managing occlusions and depth discontinuities.

In order to be able to recognize and match non-textured areas in the images, our method is based on region matching. The matching strategy consists of a cooperative scheme where correspondences, occlusions and 3D reconstruction are obtained in a single process. A representation of the 3D surfaces of the scene is maintained and modified iteratively, in order to minimize the differences between corresponding regions. This strategy is liable to certain constraints on the scene surfaces. In this paper, a simple case is studied in order to verify the algorithm convergence, analyze the computational cost and evaluate the usefulness of the results. The main advantages of our method are: a) all the pixels are assigned a depth, including the pixels which are occluded in the other image, and b) a 3D model is obtained directly, without the need for range data segmentation.

1.1. Background

1.1.1. Structure from stereo

Stereo vision techniques are usually classified [6, 12] into two broad families.

Area-based methods exploit the radiometric resolution of pixels through the use of windows and obtain satisfactory results when scenes are composed of well-textured surfaces. These methods provide a dense disparity map, which is their main advantage, together with the possibility of obtaining sub-pixel accuracy. Their main drawback is that they implicitly assume that the surface is continuous and therefore, they have problems when discontinuities are present on three-dimensional surfaces. Generally, depth data is transformed into structured descriptors through range data segmentation methods [18, 5].

Feature-based methods exploit high level features obtained from the image, which are elements (edge pixels, linear edge segments, corners, curve segments, regions, etc.) with distinctive attributes (position, orientation, curvature, etc.) which are used to find correspondences. The main drawback of these methods is that they provide a sparse disparity map. The higher the feature level, the more robust the correspondences, and the sparser the disparity map. A post-process is needed in order to obtain surface descriptors from the sparse information [18, 7] by means of an interpolation method which should be able to detect the discontinuities. Also hierarchical methods [11] have been developed to reduce the sparseness of the results. Only a few methods take advantage of the matched features to obtain the 3D structure directly [17, 4].

1.1.2. Occlusion detection

Geiger *et al.* [8] showed that occlusions can help in the correspondence computation. It is possible to model occlu-

sions and depth discontinuities explicitly, so that they are part of the problem to solve, and therefore, of the solution. Thus, occlusions and depth discontinuities are not problems to avoid, but a source of information to take into account. Belhumeur [1] also stated that a detailed map of the geometry of the scene (depth, orientation, etc.) should be maintained internally, so that all these elements cooperate in the optimization of the correspondence.

Despite the importance of occlusion detection, only a few methods integrate it into the matching process. Most of these methods model occlusions as elements in the matching space, and use a dynamic programming strategy to obtain the solution [8, 9, 10, 1, 2]. These methods are area-based techniques, which usually fail when the surface is homogeneous. Some smoothing constraint is needed to avoid spurious matches at homogeneous regions and half-occluded regions. The smoothing constraint should be suspended at depth discontinuities and other salient features (i.e. steeply sloping surfaces), which must be preserved to produce accurate reconstructions. Many of the methods used to minimize the complications with homogeneous regions smooth over the salient features in the scene geometry. Belhumeur [1] succeeded in relating salient features, occlusions and stereo correspondence in a very thorough model which provides a global solution. However, the search for the optimal solution is computationally very expensive.

Olsen [16] integrated the detection of disparity discontinuities and occluded areas in a feature-based method, by analyzing the partial derivatives of the reconstructed disparity surface. Reconstruction and correspondence are integrated in a single process, with a coarse to fine strategy. However, it is based on matching the edges of the images, which are usually scarce in homogeneous surfaces.

1.1.3. Use of regions as the matching primitive

The benefits of using regions as the matching primitive are:

- They have a higher semantic level than the other features. Their stability and descriptive capability reduce the number of ambiguities, increment noise tolerance and provide more reliable matches [14, 15].

- They represent homogeneous intensity areas with intensity discontinuities at their boundaries. Because depth discontinuities and occlusions are located at intensity discontinuities in the image, regions represent continuous depth areas, while depth discontinuities should be allowed at their boundaries.

- A 3D model composed of planar surfaces is often enough to represent structured scenes. The existing methods that reconstruct the 3D surface from two regions [17, 4] assume that each region is the projection of a three-dimensional planar surface.

However, higher level features are more difficult to extract from the image. Due to noise, occlusions and limitations of the segmentation techniques, important differences may appear between the regions obtained from both images. This drawback makes the reconstruction of the 3D surface more difficult, if not impossible. For example, Tarel *et al.* [17] used an invariant-based coherence test to reject pairs of corresponding regions which are not well segmented, and Chabbi *et al.* [4] required a trinocular system to obtain triplets of 3D faces which are validated through projective geometry principles.

1.2. Contributions

We propose to use regions as the matching primitive in order to obtain the 3D structure of the scene directly, and to avoid the drawback of using regions by segmenting only one image, which we call the *reference image*.

The corresponding region of each region in the reference image is searched by minimizing the dissimilarity between them. Due to the unicity constraint and the existence of occlusions, the minimization for one region is not independent of the minimization for its adjacent regions. Therefore, this is a n-dimensional optimization problem, where n is the number of regions in the reference image.

This optimization problem can be expressed in terms of variational calculus. However, to obtain an algorithm that guarantees convergence it is necessary to assume some constraint on the shape of the 3D surfaces in the scene. An assumption which is generally made is that each region is the projection of a 3D planar surface [17, 4]. We propose a preliminary approach to this method, which consists of a more restrictive assumption: the 3D surface is planar and parallel to the reference image. This approach allows an easier study of the convergence, the cost, and the results of our algorithm. An initial algorithm was presented in [13], with satisfactory results. However, the accuracy of the results decreased when the minimization of some regions converged much faster than others.

In this paper, we present a new algorithm that controls the minimization of each region in several steps, in order to cooperatively guide the process towards a global convergence. In addition, we show that the order of the theoretical computational cost in the worst case is similar to other area-based methods. Finally, some results with synthetic and real scenes are presented to validate the scope of this approach and the possibilities of a generalization of the method to other types of surfaces.

2. REGION MATCHING AND OCCLUSION DETECTION

A 3D surface is associated to each region R in the reference image, which is constrained to a given shape, in our case a plane characterized by its depth, $Z(R)$. The corresponding region R' of any region R in the reference image consists of

the projection of its associated surface on the other image. Therefore, R' can be expressed as a function of region R and its depth.

The advantage of this approach is that the search for correspondences can be expressed in terms of variational calculus, where the unknowns are the parameters of the associated surfaces. Another advantage is that detection of occlusions is straightforward: when two corresponding regions R'_1, R'_2 intersect, the intersected area is an occluded area of one of the two associated surfaces, due to the unicity constraint. The furthest region is partially occluded by the nearest region.

2.1. Basic strategy

The strategy of this search is based on minimizing the dissimilarity between corresponding regions, which can be computed using any correlation technique, for example, the zero-mean normalized cross-correlation (ZNCC). That is, the minimization of the following function for each region

$$F(R, Z(R)) = -C_{ZNCC}(R, R') \tag{1}$$

where R' is a function of R and $Z(R)$. Therefore, given an estimated depth for each region in the image, the function to be minimized is:

$$E(Z) = \int_{R \subset I_1} F(R, Z(R)) dR \tag{2}$$

According to the Euler equation, for any region R in the reference image, the depth $Z(R)$ that minimizes the function is a solution of the following equation

$$C(R, R')H(R', R') - H(R, R') = 0 \tag{3}$$

where R' is the region corresponding to R regarding the current depth $Z(R)$, $C(R, R')$ is the gray level correlation between corresponding pixels in R and R', and $H(P, Q)$ is a measurement of the correlation between the gray level in P and the partial derivative of the gray level with respect to depth in Q (for more details, see [13]).

Let us call $F_z(R)$ the error computed by the previous equation for each region R.

$$F_z(R) = C(R, R')H(R', R') - H(R, R') \tag{4}$$

If we plot F_z with respect to the region depth, we can observe a zero-crossing at each minimum of the proposed function, with positive values to the right and negative values to the left of the zero-crossing. Therefore, F_z allows us to decide whether depth must be incremented or decremented in order to achieve the associated minimum, which allows the design of an iterative algorithm that increments/decrements the current depth towards the solution. It is important to note that the initial depth should not be far from the solution in order to avoid other local minima different from the global solution. This is the reason why a multi-level (coarse-to-fine) scheme is needed, where the depth map resulting from each level of a pyramidal structure of the images is used to initialize the depth map of the next level (Fig. 1).

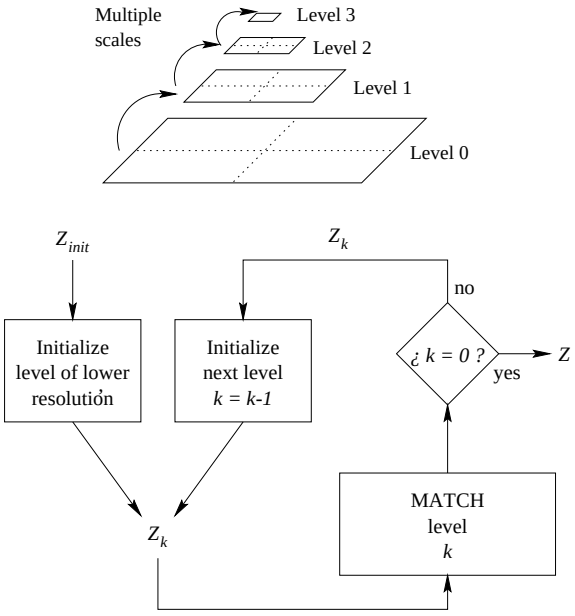

Fig. 1. Multi-level scheme.

2.2. The role of occlusions

The matching algorithm at each level is an n-dimensional optimization algorithm, where n is the number of regions in the reference image. The existence of an occlusion may introduce errors in the correlation measurement of a region. Therefore, occluded areas must be detected and removed from correlation computation. Consequently, the optimization process for each region depends on the optimization process for its adjacent regions, because if the depth of an adjacent region varies, occlusions may appear or disappear, and therefore F_z may vary.

In the proposed strategy, convergence is achieved by performing small increments/decrements at each depth until the minimum for each region is reached. These increments must be small, in order to move slowly towards the solution. To fix the size of the depth increments, Δz, can lead to undesirable disparity increments greater than 1 pixel when objects are close to the camera. We propose to fix discrete disparity increments, Δd, which should be less than 1 pixel.

2.3. Algorithm

The accuracy of the results depends on the selection of Δd. As Δd decreases, the accuracy increases, but the computational cost increases too. A multi-resolution scheme, where different decreasing Δd are used, allows the computation to be accelerated, while refining the converged results. Thus,

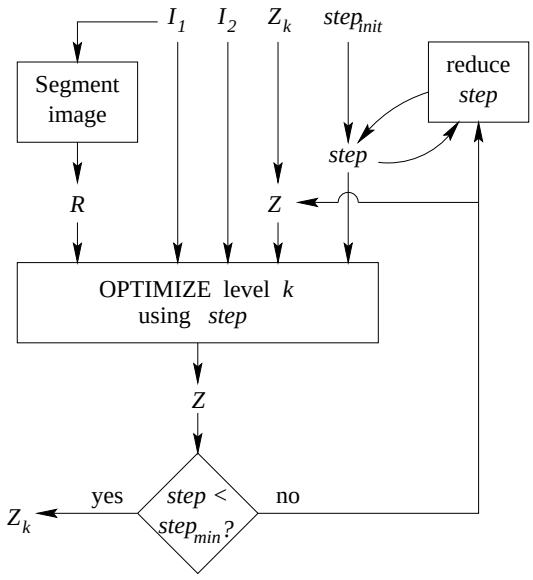

Fig. 2. Matching process at each level.

the search for the minimum for each region in the reference image is performed in several iterations (one per resolution) (Fig. 2). Such a scheme can be applied in different ways:

1. A Δd is associated to each region and convergence is treated independently for each region at each resolution. When a region converges at each resolution, its Δd is reduced and the convergence process is re-started.

   This method does not guarantee the convergence of the algorithm, given that interdependence of results may lead to cycles in its behaviour.

2. A Δd is associated to each region and convergence is treated independently for each region at each resolution. When a region converges at each resolution, its Δd is reduced and the convergence continues in the initial direction [13].

   This method guarantees the convergence, but the accuracy of the results decreases when one region converges much faster than another. This is the case of regions with occlusions that converge rapidly to a local minimum. Although the occlusion can be obtained by the correct convergence of the occluding region, the depth is not corrected due to the convergence strategy.

3. A global Δd is defined and the convergence is treated globally at each resolution. When all the regions converge at each resolution, the global Δd is reduced and the convergence algorithm is re-started.

   This method guarantees the convergence, while the approach to the global minimum is controlled in several steps. In this paper, we show the results obtained with this method and compare with the results obtained with the previous method.

The matching algorithm at each optimization step consists of the iteration of two operations (Fig: 3):

- *Transform* the second image into two maps: T contains the corresponding points for each pixel in the reference image, and S contains a classification of all the pixels in I_1: *active*, *occluded* or *out of bounds*.
- *Increment* (or decrement) depth depending on the previous steps and the comparison of the reference image with T, taking into account only the *active* pixels in S.

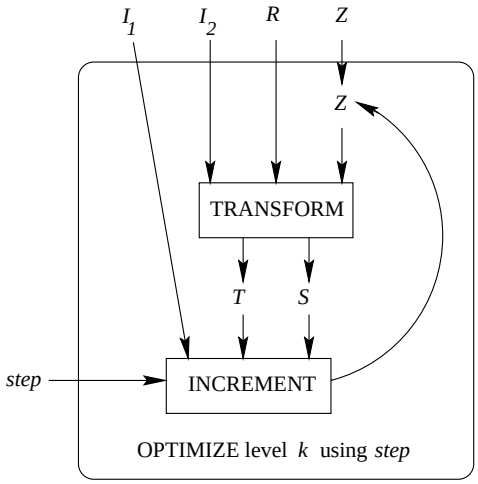

Fig. 3. Optimization at each level and each step.

The accuracy of occlusion detection (and therefore, of depth results) can also be increased by the use of subpixel occlusion detection, which consists of using $u \times v$ cells for the computation of the classification of each pixel.

Finally, the algorithm described in this section can be summarized as follows.

Init Z of top level
for each level k from top to bottom in the pyramid **do**
  { *MATCH level* k }
  for each Δd from coarsest to finest resolution **do**
    { *OPTIMIZE level* k *with step* Δd }
    while not all-regions-converge **do**
      $S, T \leftarrow$ Transform I_1, I_2, R, Z
      $Z \leftarrow$ Increment $I_1, T, S, Z, \Delta d$
    end while
  end for
  if k is not the bottom level **then**
    Init Z of next level using Z of current level
  end if
end for

2.4. Analysis of the computational cost

The computational cost in the worst case, assuming that Δd is reduced by dividing it by 2, is $O(\frac{D}{\Delta d_{min}} NMuv)$, where

$N \times M$ is the size of the images, $u \times v$ is the number of subpixel sections for occlusion detection, D is the size of the disparities range, and Δd_{min} is the minimum Δd required. That is, the product of the number of pixels at the selected subpixel occlusion resolution ($NMuv$, which is the cost of each iteration of the inner loop) by the number of disparities at the selected subpixel disparity resolution ($\frac{D}{\Delta d_{min}}$, which is the total number of iterations of the inner loop). However, it is important to note that we calculated the theoretical cost by assuming that there exists one region that traverses the whole depth interval at every iteration of the inner loop, including the last iteration of the outer loop, which is the most expensive one. This is not true in practice, where the first iteration of the outer loop performs more iterations of the inner loop than subsequent iterations. The first iteration of the outer loop is an approach stage, while the rest of iterations are refining steps which usually require very few iterations.

Generally, the cost of area-based methods is similar to or higher than this, depending on whether they provide local or global optimization. Methods based on correlation windows are $O(DNMnm)$, where $n \times m$ is the size of the correlation window. This cost is similar to the previous one, provided that the Δd_{min} term represents the subpixel accuracy. Methods based on dynamic programming are usually $O(NMD^2)$ when based on pixel differences, or $O(NMD^2nm)$ when based on correlation windows, except for [9], which is $O(3NMDnm)$, but control points must be added in order to be able to detect occlusions [10].

3. EXPERIMENTAL RESULTS

Experiments with synthetic images with ground truth depth show that the accuracy of the results is increased with respect to the previous algorithm [13]. In figure 4, an example of a pyramid with depths ranging from 82 to 117 cm is shown. The mean relative error of the obtained depth map, without considering the background region, is 0.62% with $u = 5$, while it was 2.31% with the previous algorithm and 3.78% when occlusions were not detected.

Darker areas in depth maps are nearer points. White areas in occlusion maps are active points, while points whose corresponding pixel is occluded are drawn in dark, and points whose corresponding pixel is out of the image limits are shown in gray.

In figure 5, another example allows for the comparison of the obtained depth map with previous results. There are some regions that obtain an erroneous depth but, in general, depth obtained with the new algorithm changes more smoothly from left to right than the previous results. For example, the depth of both parking meters and the bush in the background are achieved more accurately. Some erroneous depths are due to errors in the regions obtained from the segmentation method, which merge areas with different depth. The rest of erroneous regions are thin vertical regions which are assigned inexistent occlusions in the results due to the parallelism assumed in the constraint.

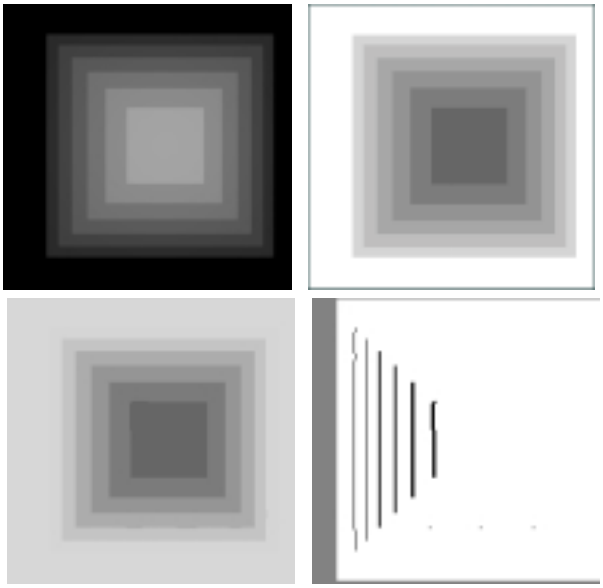

Fig. 4. An example of synthetic images. Top left: reference image; top right: ground truth depth map; bottom left: resulting depth map; and bottom right: occlusion map. Darker areas in depth maps are nearer points.

Experiments with real images of structured scenes show that a complete representation of the scene is obtained (Fig. 6 and 7). As expected, some surfaces, like the floor or the table, can not be modeled with the assumed constraint. In spite of this drawback, the results are satisfactory enough and encourage us to generalize the method to other types of surfaces.

4. CONCLUSIONS

We have presented a new matching algorithm capable of directly providing a 3D model of a structured scene from two views, where depth discontinuities are not smoothed over, and occlusions and correspondences are computed cooperatively. The accuracy of the results has been increased by controlling the convergence of the optimization algorithm. The order of the computational cost is similar to other area-based techniques.

The usefulness of the results is limited by the assumed constraint. An assumption of a scene made of planar surfaces of any orientation would be enough for several applications. Further work is being carried out in order to generalize the method to these surfaces and second order surfaces. Another line of research is the use of left-right consistence to improve the initial gray level segmentation into a more consistent segmentation based not only on gray level but also on range information.

Fig. 5. Example "parking meter" from the JISCT test set [3]. Top: reference image; bottom left and right: depth maps obtained from the previous and new algorithms, respectively.

Acknowledgments

We would like to thank Fundació Caixa-Castelló (grant P1 .1B2000-22) and CICYT of Ministerio de Educación y Ciencia (grants TIC2000-1131 and TAP99-0590-C02-01) for supporting this work. We also would like to thank Generalitat Valenciana for its support.

5. REFERENCES

[1] P. Belhumeur. *A Bayesian Approach to the Stereo Correspondence Problem*. PhD thesis, Electrical Engineering, Yale University, May 1993.

[2] S. Birchfield and C. Tomasi. Depth Discontinuities by Pixel-to-Pixel Stereo. In *Proceedings of the IEEE International Conference on Computer Vision, Bombay, India*, 1998.

[3] R. Bolles, H. Baker, and M. Hannah. The JISCT stereo evaluation. In *Proc. ARPA Image Understanding Workshop*, pages 263–274, Washington, DC, Apr. 18-21 1993. Morgan Kaufmann.

[4] H. Chabbi and M. Berger. Using projective geometry to recover planar surfaces in stereovision. *Pattern Recognition*, 29(4):533–548, 1996.

[5] L.-H. Chen and W.-C. Lin. Visual surface segmentation from stereo. *Image and Vision Computing*, 15:95–106, 1997.

[6] U. Dhond and J. Aggarwal. Structure from stereo - a review. *IEEE Trans. on Systems, Man, and Cybernetics*, 19(6):1489–1510, 1989.

[7] C. Dillon and T. Caelli. Generating complete depth maps in passive vision systems. pages 562–566, 1992.

[8] D. Geiger, B. Ladendorf, and A. Yuile. Occlusions and binocular stereo. *Int. J. of Computer Vision*, pages 221–226, 1995.

[9] S. Intille and A. Bobick. Disparity-space images and large occlusion stereo. In J.-O. Eklundh, editor, *Proc. 3rd European Conf. on Computer Vision*, volume B of *Lecture Notes in Computer Science*, pages 674–677, Stockholm, Sweden, May 1994. Springer Verlag. Extended version in M.I.T Media Lab Computing Group Technical Report No. 220.

[10] S. Intille and A. Bobick. Incorporating intensity edges in the recovery of occlusion regions. In *Int. Conf. on Pattern Recognition*, volume A, pages 674–677, 1994. Also M.I.T Media Lab Computing Group Technical Report No. 246.

[11] G. Jones. Constraint, optimization, and hierarchy: Reviewing stereoscopic correspondence of complex features. *IEEE Trans. on PAMI*, 65(1):57–78, 1997.

[12] R. A. Lane and N. A. Thacker. Stereo vision research: An algorithm survey. Technical Report 94/16, University of Sheffield, Electronic Systems Group, 1994.

[13] A. López and F. Pla. A minimization approach for 3D recovery in region-based stereo vision. In *Proc. 10th Image Processing and its Applications*, pages 47–51, Manchester, UK, 1999.

[14] S. Marapane and M. Trivedi. Region-based stereo analysis for robotics applications. *IEEE Trans. on Systems, Man, and Cybernetics*, 19(6):1447–1464, 1989.

[15] S. Marapane and M. Trivedi. Multi-primitive hierarchical (mph) stereo analysis. *IEEE Trans. on PAMI*, 16(3):227–240, 1994.

[16] S. Olsen. Stereo correspondence by surface reconstruction. *IEEE Trans. on PAMI*, 12(3):309–315, 1990.

[17] J.-P. Tarel and J.-M. Vézien. A generic approach for planar patches stereo reconstruction. In *Proc. 11th Scandinavian Conf. on Image Analysis*, pages 1061–1070, Norway, 1995.

[18] D. Terzopoulos. Reconstruction of visual surfaces: Variational principles and finite element representations. Technical Report A. I. Memo 671, MIT, 1982.

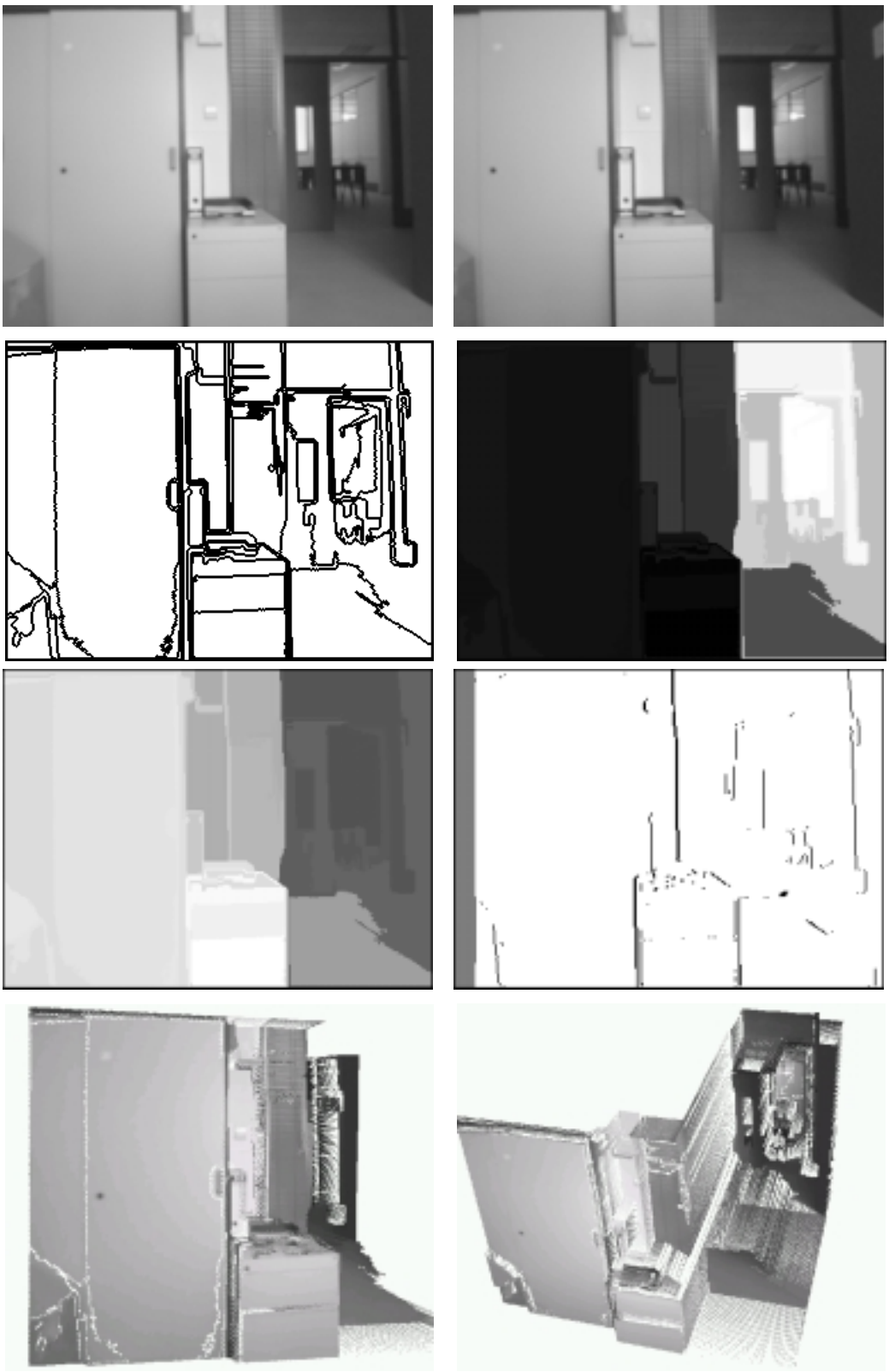

Fig. 6. Real stereo pair of images, *despacho*. From left to right and top to bottom: stereo pair of images, segmentation by region merging, depth map, disparity map, occlusion map and two views of the scene model.

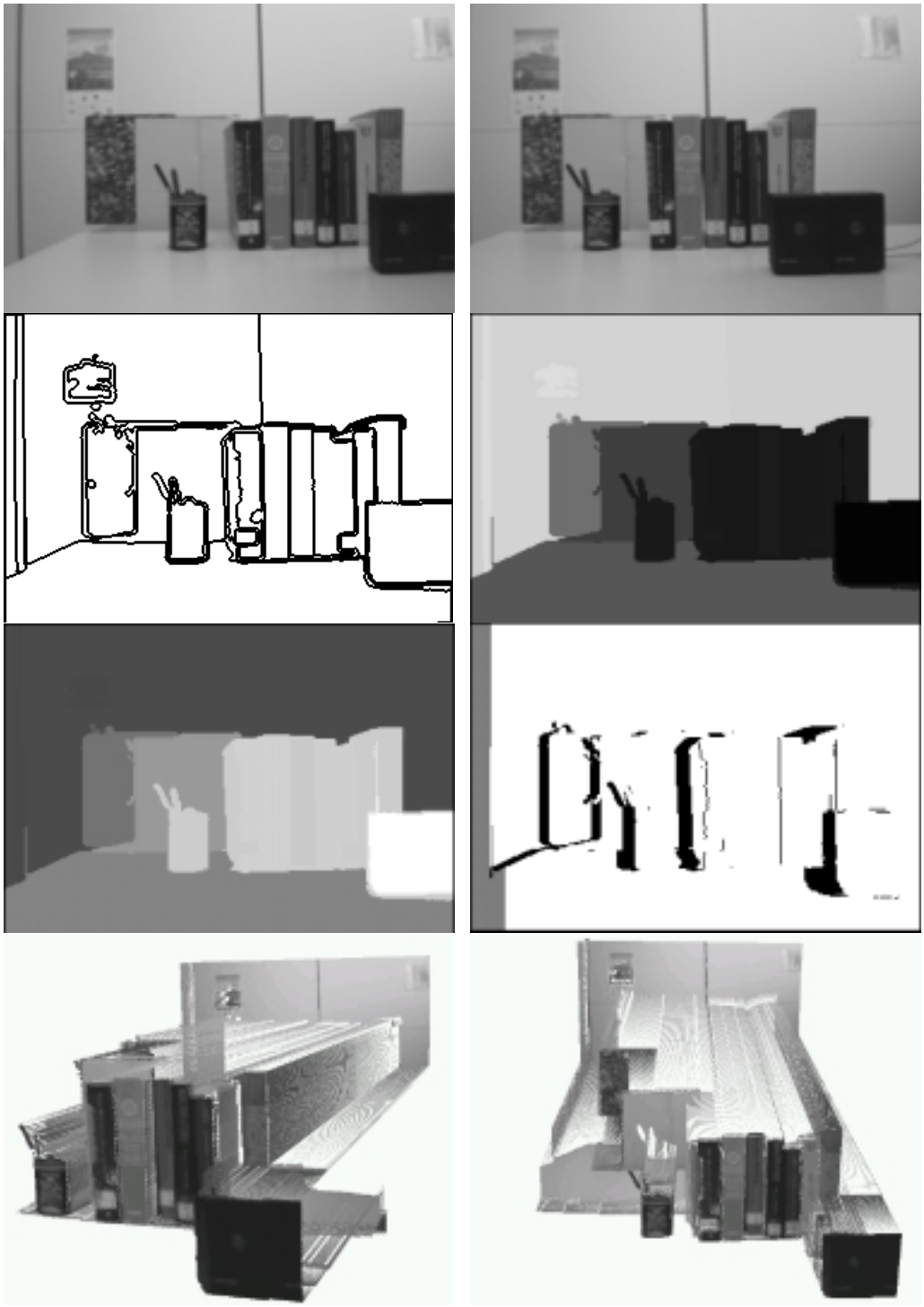

Fig. 7. Real stereo pair of images, *libros*. From left to right and top to bottom: stereo pair of images, segmentation by region merging, depth map, disparity map, occlusion map and two views of the scene model.