Region-based Stereo Vision using a Minimization Approach *

A. López and F. Pla

Departament d'Informatica
Universitat Jaume I
12071 Castellón. SPAIN
e-mail: {lopeza,pla}@inf.uji.es

Abstract

In this paper, a region-based method for stereo vision is presented. The method computes depth directly, without the intermediate calculation of disparities, by the minimization of a correlation-based energy function. Some constraint on the distribution of depth in the regions is assumed, so that the obtained solution is smooth inside the regions while depth discontinuities are allowed in their boundaries. The derived equation indicates whether an increment or decrement in depth is needed in an iterative process in order to achieve the solution. A multiscale approach is used in order to avoid local minima.

Keywords: Region-based Stereo Vision, Energy Minimization, Correlation-based similarity, Depth Map.

1 Introduction

Stereoscopic vision is the set of techniques that try to recover threedimensional information from two or more views of a scene. The following steps are involved in the process: *Calibration* of the intrinsic and extrinsic parameters involved in the stereoscopic geometry; *Rectification* of the epipolar geometry, in order to simplify as much as possible the search involved in the correspondence problem; *Correspondence* between points in both images, which provides a disparity map, preferably a dense map; and *Reconstruction* of the 3D scene, that is, the calculation of depths from disparities.

These steps are sometimes merged or skipped, as in the case of uncalibrated systems that do not need some or all of the calibration parameters [8] or some matching methods that perform the matching without image rectification, like in the case of the minimization and regularization approach [14], which in addition computes depth directly, that is, without the intermediate calculation of disparities.

Depth discontinuities were usually treated during reconstruction, but in the last years there are more works that treat them during correspondence, as in the case of correlation-based techniques using adaptive windows [10] dynamic programming approaches [9] [7] [12] and bayesian approaches [2].

*This work was partially supported by grant TAP95-0710 (CICYT, Ministerio de Educación y Ciencia)

Here we present a method inspired from the Robert and Deriche's approach [14], which computes depth directly and takes into account depth discontinuities. Rectification, correspondence and reconstruction are achieved in an only step. This approach is based on image restoration techniques, which are reviewed in [4] in detail. This method computes depth by minimizing an energy function which consist of the similarity error between corresponding points in both images, based on pixel features.

In our approach, we face the problem of depth discontinuities by using regions obtained from the segmentation of the reference image. Depth discontinuities are usually present at region boundaries, and regions represent areas with smooth depth. There are several methods that use regions as the matching primitive [13, 11, 3]. These works try to match regions in both images, so they try to deal with segmentation errors like regions that are splitted in the other image and so on. Here, we use regions in one image and we try to find its corresponding region without segmenting the other image.

Regions have a higher semantic content compared with other matching primitives such as edge segments or points, so that the resulting correspondences are expected to be more reliable. However, matching regions is not easy due to occlusions and the fact that depth can vary considerably along the region. Several assumptions can be made on the depth distribution in a region: constant depth (scene made of fronto-parallel planes), depth varying linearly (scene made of any type of planes), depth varying arbitrarily (scene made of any type of surfaces).

In robotics applications, the images to be processed often consist of man-made scenes, which are mostly composed by planar surfaces. These surfaces usually contain homogeneous gray levels, which makes difficult to apply a point-based matching technique. However, in a region-based matching, the restriction of planar surfaces permits calculating easily the region corresponding to a given region in the reference image.

Therefore, depth can be calculated by minimizing an energy function based on correlation between regions. For each region in the reference image, its corresponding region in the other image is calculated from the current depth and the calibration parameters. In section 3, we propose a correlation-based energy function to be minimized for region matching. The appropriate multiscale method is developed in order to achieve convergence and selection of the depth increments is discussed. In section 4, some experiments with synthetic image pairs made of constant depth regions, are shown.

2 Previous Work

The minimization and regularization technique faces the stereo problem as the minimization of an energy function. The energy is expressed as a function of the scene depth, so that it can be calculated directly by means of some iterative method.

Let I_1 and I_2 be two different views of the same scene, with I_1 as the *reference image*. Let us define Z as a map of depths corresponding to all the pixels in I_1.

When the stereo system is strongly calibrated, we can calculate the corresponding point $m' = (u', v')$ in image I_2 of a given point $m = (u, v)$ in image I_1 given its depth, $Z(u, v)$,

$$m' = f_{12}(m, Z(m)), \tag{1}$$

where f_{12} depends on the projection matrices of images I_1 and I_2 [6]. This relation between

m and m' depends on the depth of the threedimensional point in the scene,

$$s' \begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} = \begin{bmatrix} J' & 0_3 \end{bmatrix} D' D^{-1} \begin{bmatrix} zJ^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \\ 1 \end{bmatrix}, \tag{2}$$

where J, J', D, D' contain the intrinsic and extrinsic calibration parameters of both images, respectively [14], and $z = Z(u, v)$.

The matching problem can be expressed as an energy minimization problem, where the minimized functional is as follows:

$$E(Z) = \int_m (\varphi(m, Z(m)) + \lambda\phi(|\nabla_m Z|))\, dm, \tag{3}$$

The first term, called the *minimization term*, represents the *error in similarity* of corresponding points from image I_1 to image I_2. The second term, called the *regularization term*, introduces some *constraint in the shape* of the depth function. This term should smooth isotropically the solution when the depth gradient is small (homogeneous regions), and it should smooth anisotropically the solution when the depth gradient is large (presence of depth discontinuities).

The energy function proposed in [14] consists of the squared sum of differences given a set of k features of any given point. According to the Euler equation [5], the solution $z = Z(m)$ is such that verifies

$$\sum_k \left(I_1^k(m) - I_2^k(m')\right) \frac{\partial}{\partial z}\{I_2^k(m')\} + \frac{\lambda}{2}\left(\frac{\phi'(|\nabla_m Z|)}{|\nabla_m Z|} Z_{\xi\xi} + \phi''(|\nabla_m Z|) Z_{\eta\eta}\right) = 0, \tag{4}$$

where $I_i^k(m)$ represents feature k of pixel m in image I_i, $m = (u, v)$ is the pixel in the reference image, $m' = f_{12}(m, z)$ its corresponding point in the other image and $Z_{\eta\eta}$, $Z_{\xi\xi}$ are the second order directional derivatives of $Z(m)$ in the gradient and orthogonal to the gradient directions, respectively.

The solution can be obtained by means of a gradient descent method. Given an initial depth for each pixel in the image, depth is incremented or decremented iteratively in each step in order to achieve the solution. A multiscale approach has to be used in order to achieve convergence. The previous scheme is applied to each level of image pyramids, and the resulting depth map in each level is used to initialize the next level.

3 Energy Minimization for Region-based Correspondence

We propose using regions as the matching primitive, and imposing some constraint in the distribution of the depth of each region, as is the assumption of planar surfaces in the scene. No regularization term is needed given that both smoothness and depth discontinuities in the solution are implicitly considered. Depth smoothness is satisfied by the constraint about depth distribution in regions. Although not all the intensity discontinuities are due to occlusions, almost every occlusion produces some intensity discontinuity. Therefore, the boundaries of regions, which are segmented under the consideration of intensity similarity, are the candidates for depth discontinuities. Thus, the energy function to be minimized becomes,

$$E(Z) = \int_R \varphi(R, Z(R)) dR. \tag{5}$$

where R is any region in the reference image.

The region R' that corresponds to R with a given depth, $Z(R)$, is

$$R' = g_{12}(R, Z(R)) = \{m' = f_{12}(m, Z(m)), \forall m \in R\}, \tag{6}$$

where $Z(m)$ depends on the constraint about the regions depth. That is, function g_{12} depends on the projection matrices of both images, and the constraint about the shape of $Z(R)$. For example, if constant depth is assumed, the contour of R' is exactly the same shape and size than the contour of R and g_{12} consists of calculating the disparities in both axes. If planar surfaces are assumed, the shape of R' is linearly dependent on the shape of R and g_{12} consists of calculating this dependence.

3.1 Notation

Given a pair of regions R and R' of identical size and shape, let us numerate all the pixels within these regions as m_i and m'_i, where $i = 1..N$ and N is the size of the regions, such that m_i and m'_i are the pixels at identical relative positions inside both regions.

Let us denote the *mean intensity* and *standard deviation* of any region R in image I_k as $\mu_k(R)$ and $\sigma_k(R)$, respectively.

Finally, let us define the *zero-mean normalized intensity* in a pixel m_i in image I_k as

$$\mathcal{I}_k(m_i) = \frac{I_k(m_i) - \mu_k(R)}{\sigma(R)}, \tag{7}$$

where R is the region containing m_i, that is, $m_i \in R \subset I_k$.

3.2 Correlation-based Energy Function

Correlation between two regions R and R' of identical size and shape can be computed by the Zero-Mean Normalized Cross-Correlation method (ZNCC), which using the notation in previous section can be expressed as

$$C(R, R') = \frac{1}{N}\sum_{i=1}^{N} \mathcal{I}_1(m_i)\mathcal{I}_2(m'_i). \tag{8}$$

The ZNCC value ranges from -1 to 1, where -1 indicates that intensity values in both regions are completely different and 1 means that they are identical. As we need a measurement of *error in similarity*, we propose the following similarity error function:

$$\varphi(R, Z) = -\frac{1}{N}\sum_{i=1}^{N} \mathcal{I}_1\left(m_i\right)\mathcal{I}_2\left(m'_i\right). \tag{9}$$

According to the Euler equation, and assuming constant depth in the regions, the solution becomes

$$\frac{1}{N}\sum_{i=1}^{N} \mathcal{I}_1\left(m_i\right)\left(\mathcal{I}_2(m'_i)\mathcal{H}(R') - \frac{\partial}{\partial z}\{I_2(m'_i)\}\right) = 0. \tag{10}$$

where $\mathcal{H}(R')$ is the *weighted mean gradient* of region R',

$$\mathcal{H}\left(R'\right) \;=\; \frac{1}{N}\sum_{k=1}^{N} \mathcal{I}_2(m'_k)\frac{\partial}{\partial z}\{I_2(m'_k)\}. \tag{11}$$

3.3 Analysis of the derived Energy Function

In order to analyse the meaning of equation 10, let us develop it as follows,

$$\frac{1}{N}\sum_{i=1}^{N}\mathcal{I}_1\left(m_i\right)\mathcal{I}_2\left(m_i'\right)\mathcal{H}\left(R'\right)-\frac{1}{N}\sum_{i=1}^{N}\mathcal{I}_1\left(m_i\right)\frac{\partial}{\partial z}\{I_2(m_i')\}=0. \tag{12}$$

On the one hand, the first term is proportional to the correlation measurement proposed in 8. On the other hand, the second term is very similar to the definition of $\mathcal{H}$. Let us define a new measurement similar to $\mathcal{H}$, which relates intensities in the reference image with gradients in the other image,

$$\mathcal{H}_{12}(R,R')=\frac{1}{N}\sum_{i=1}^{N}\mathcal{I}_1(m_i)\frac{\partial}{\partial z}\{I_2(m_i')\}. \tag{13}$$

Then, the derivative of F with respect to z can be expressed as

$$C\left(R,R'\right)\mathcal{H}\left(R'\right)-\mathcal{H}_{12}\left(R,R'\right)=0, \tag{14}$$

Both $\mathcal{H}$ and $\mathcal{H}_{12}$ depend on the intensity gradient in the direction of the epipolar line, and they can be considered similar measurements. They mean a comparison between zero-mean normalized intensity and intensity gradient. $\mathcal{H}$ compares intensity and gradient in R', while $\mathcal{H}_{12}$ compares intensity in R with gradient in R'. Correlation between intensities in both neighbourhoods, C, ranges from -1 to 1. If R' is the region corresponding to R in ideal conditions, $C=1$ and $\mathcal{H}=\mathcal{H}_{12}$, so that $C\mathcal{H}-\mathcal{H}_{12}=0$ and depth does not increase neither decrease: the iterative algorithm stops when the appropriate depth is reached. Otherwise, $C\mathcal{H}-\mathcal{H}_{12}\neq 0$ gives us a measurement of how depth should be modified in order to get closer to the solution. That is, the solution is in the zero-crossing of $C\mathcal{H}-\mathcal{H}_{12}$.

3.4 The matching process

The aim is computing a small increment or decrement in depth in each step of the iterative algorithm in order to move depth smoothly towards the solution. In order to avoid achieving some local minima instead of the global solution, a multiscale approach is needed. The pyramidal structure permits the algorithm to converge in an scaled solution at each level, which is used to initialise the next level.

Another important issue is the depth increment/decrement amount, ΔZ. On the one hand, ΔZ that produce R' displacements greater than a few pixels are not desirable in order to avoid jumps that would lead to a different minimum. Therefore, ΔZ should not produce m' increments/decrements greater than, for example, one pixel. As $m'=(u',v')$ can be calculated from $m=(u,v)$ and the current estimated depth, Z^t (equation 2), we propose calculating the new depth, Z^{t+1}, from the desired either $\Delta u'$ or $\Delta v'$.

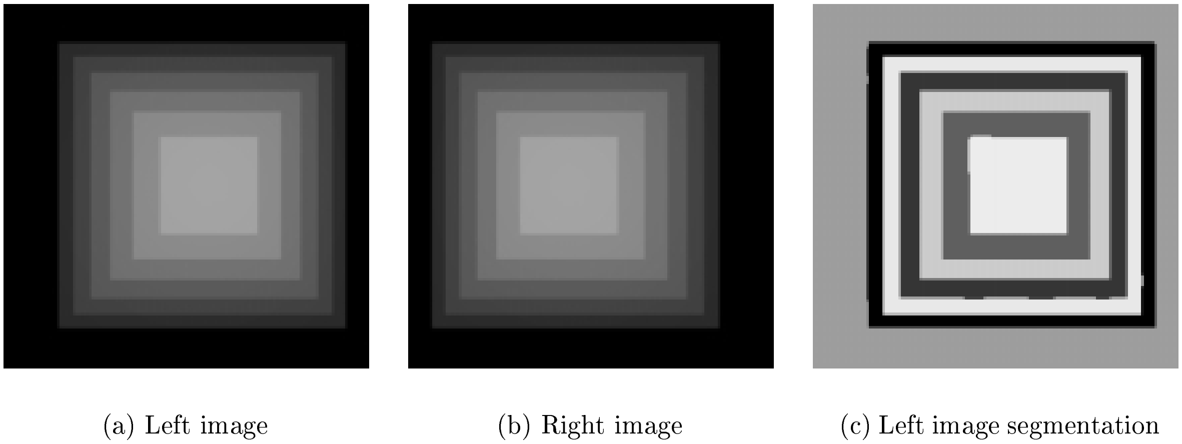
When $C\mathcal{H}-\mathcal{H}_{12}>0$ depth must be decremented and viceversa. If depth has been iteratively decremented in order to make $C\mathcal{H}-\mathcal{H}_{12}$ decrease towards 0 and we obtain a negative value, the last decrement must be reduced in order to get more accuracy in the solution. The idea is using fixed decrements for getting close to the zero-crossing and using smaller decrements until the solution is obtained.

```
ALGORITHM 1: COMPUTEDEPTH
for each region $R$ in $I_1$
        $Z^t =$ initial Z
        $R' = g_{12}(R, Z^t)$
        $E^0 = C(R, R')\mathcal{H}(R') - \mathcal{H}_{12}(R, R')$
        $\Delta u' =$ sign($E^0$)1
        repeat
                $Z^{t+1} = f(\Delta u',$ calibration parameters, $Z_{min}, Z_{max})$
                $R' = g_{12}(R, Z^{t+1})$
                $E^{t+1} = C(R, R')\mathcal{H}(R') - \mathcal{H}_{12}(R, R')$
                if sign($E^{t+1}$) $\neq$sign($E^0$) then
                        $oldZ = Z^t$
                        $Z^t = Z^{t+1}$
                else
                        $\Delta u' = \Delta u'/2$
                endif
        until $E^{t+1} = 0$ or $Z^{t+1} = oldZ$
endfor
```

Algorithm 1: Depth computation at each level in the pyramid.

4 Experimental Results

In the experiments we used different segmentation algorithms: a clustering technique developed in [1] which groups nearby pixels in regions within a certain variance in the gray level and a common region merging segmentation method [15].


(a) Left image

(b) Right image

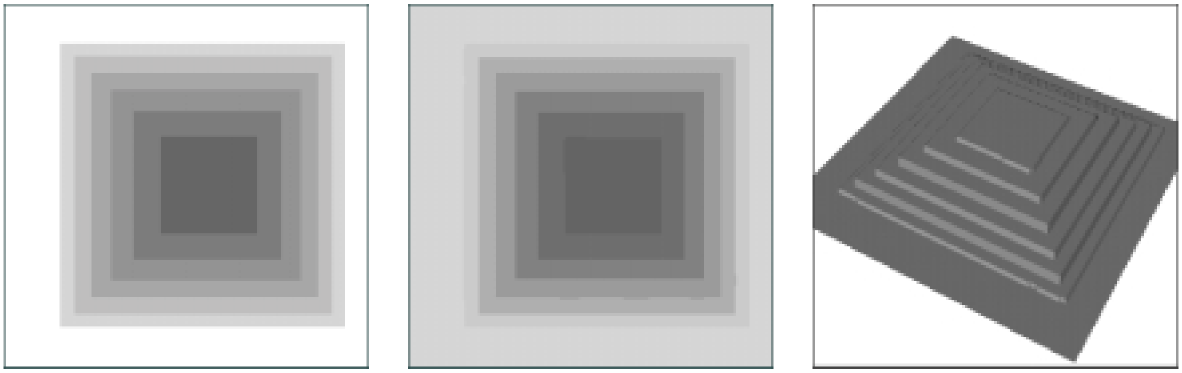
(c) Left image segmentation

Figure 1: Synthetic stereo pair of images.

As depth is considered constant inside the regions, the same results are obtained with any $0 < \Delta u' \leq 1$ for ΔZ computation. Therefore, $\Delta u' = 1$ has been used.

In figure 1, a synthetic stereo image pair is shown. Segmentation of the left image using the clustering technique provides the regions shown in 1(c). The ground truth and the resulting depth maps are shown in figure 2. Clearer areas correspond to further points, while darker areas correspond to nearer points.

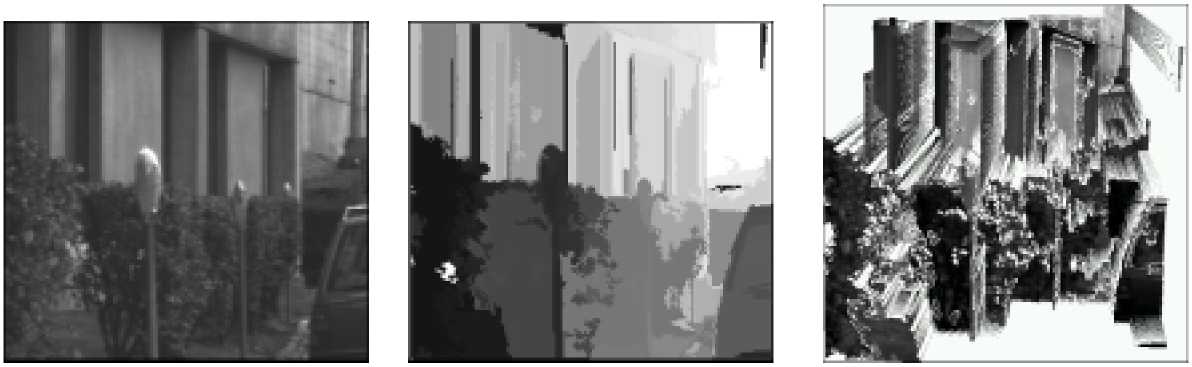
The results show that all the depths have been found, except for the background region of the scene, which has constant grey level, so that it acquires a depth similar to the regions adjacent to it. Note that the obtained depth map 2(b) is quite similar to the



(a) Ground truth depth (b) Depth map (c) 3D reconstruction

Figure 2: Depth maps. Clearer areas are further points.

expected one 2(a). In fact, the mean relative error obtained in all depths calculated is 3.78%, except the background. In this example, depth ranges from 82 to 117 cm.



(a) Left image (b) Depth map (c) 3D reconstruction

Figure 3: A real example.

In figure 3, the results for a real image pair are shown. Note that, although the constraint of constant depth in the regions has been applied, the depth map (figure 3(b)) provides satisfactory results and gives an approximation of the 3D scene (figure 3(c)). Although some areas are not assigned to the expected depth (some little dark regions and one white region) the method obtains satisfactory results even in the case of depth discontinuities, like is the case of the parking meter and the car.

5 Conclusions and Future Work

We have presented the basis of a new stereo vision approach which directly computes, without the intermediate calculation of disparities, depth of the regions resulting from the segmentation of the reference image. The method consist of minimizing an energy function in a multiscale scheme, such that the algorithm converges in an scaled solution at each level, which is used to initialise the next level.

A correlation-based energy function has been proposed in order to include regions similarity in the matching process. The geometric transformation defined by the geometry of the projections and the current estimated depth is taken into account when calculating the corresponding region in the second image at each iteration.

The method has been tested in the particular case of assuming that depth inside the regions is constant. The tests show that the method obtains satisfactory results even in the presence of depth discontinuities. Further work is directed to generalize the method to the assumption of scenes made of planar surfaces as well as other type of surfaces.

References

[1] J. Badenas, M. Bober, and F. Pla. Motion and intensity-based segmentation and its applications to traffic monitoring. In A. del Bimbo, editor, *Proc. of the 9th Int.Conf. on Image Analysis and Processing*, volume 1310 of *Lecture Notes in Computer Science*, pages 502–509, Florence, Italy, May 1997. Springer Verlag.

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

[3] L. Cohen, L. Vinet, P. Sander, and A. Gagalowicz. Hierarchical region based stereo matching. In *Proc. of the IEEE Conf. Computer Vision and Pattern Recognition*, pages 416–421, San Diego, CA, June 1989. IEEE Computer Society Press.

[4] R. Deriche and O. Faugeras. Les EDP en Traitement des Images et Vision par Ordinateur. *Traitement du Signal*, 13(6), 1996. Numéro spécial RFIA'96.

[5] L. Elsgoltz. *Ecuaciones diferenciales y cálculo variacional.* Editorial MIR, Moscu, 1977.

[6] O. Faugeras. *Three-Dimensional Computer Vision: a Geometric Viewpoint.* The MIT Press, 1993.

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

[8] N. Hollinghurst and R. Cipolla. Uncalibrated stereo hand-eye coordination. *Image and Vision Computing*, 12(3):187–192, Apr. 1994.

[9] S. S. Intille and A. F. Bobick. Disparity-space images and large occlusion stereo. In J.-O. Eklundh, editor, *Proc. of the 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] T. Kanade and M. Okutomi. A stereo matching algorithm with an adaptive window: Theory and experiment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(9):920–932, Sept. 1994.

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

[12] Y. Ohta and T. Kanade. Stereo by intra- and inter-scanline search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7:139–154, 1985.

[13] S. Randriamasy and A. Gagalowicz. Region based stereo matching oriented image processing. In *Proc. of the IEEE Conf. Computer Vision and Pattern Recognition*, pages 736–737, Lahaina, Hawai, June 1991. IEEE Computer Society Press.

[14] L. Robert and R. Deriche. Dense depth map reconstruction: A minimization and regularization approach which preserves discontinuities. In B. Buxton, editor, *Proc. of the 4th European Conf. on Computer Vision*, Cambridge, UK, Apr. 1996.

[15] A. Rosenfeld and A. Kak. *Digital Picture Processing*, volume 1. Academic Press, New York, 1982. Second Edition.