Disparity estimation in stereoscopic vision by *simulated annealing*

Patricia COMPAÑ, Rosana SATORRE, Ramón RIZO

Group i3a: Industrial Computing and Artificial Intelligence
Departament of Computer Science and Artificial Intelligence
University of Alicante
{patricia, rosana, rizo}@dccia.ua.es

Abstract. This paper presents a correspondence algorithm for stereo vision based on an integrated model that includes several units corresponding to different stages: features extraction, minimization of an energy function using simulated annealing, multiresolution and interpolation. Firstly the original images are scaled down to considerably reduce their size. From the reduced images a disparity map is obtained, which is used as the basis to develop the complete process. For this reason an energy function is built and minimized using a multiresolution scheme. The energy function integrates features such as grey level, non parametric transforms, edges, smoothness and uniqueness. The obtained disparity for every resolution is interpolated to work with the following resolution. Our model produces a dense disparity map. The algorithm has been tested with different kinds of real images to show its flexibility.

Introduction

Our brains obtain two similar images of a scene, as if they had been taken from two nearby points on the same horizontal level, this is due to the position and the control of our eyes. Two objects at a different distance from the observer have different relative positions in their retinal images. The brain is able to measure this difference (retinal disparity) and to use it to estimate the depth [9]. The retinal disparity depends on the distance to the fixing instant. To be able to use the binocular capacity to detect depth, an organism must have a binocular visual field, that is, an overlap region of visibility between both eyes. Every animal has a different binocular visual field size. In general, predators have their eyes at the front and so they have large binocular visual fields. On the other hand, their prey typically have their eyes at the side of their heads so they have small binocular visual fields, if any. Stereoscopic vision is a set of techniques that try to recover three-dimensional information from two or more views of a scene. In this process some different stages can be distinguished:

- Calibration of intrinsic and extrinsic parameters involved in the stereoscopic geometry.
- Rectification of the epipolar geometry to simplify the search done when solving the problem of correspondence.
- Correspondence of tokens of the images to obtain a disparity map. Our three-dimensional perception of the world is due to how our brain interprets the difference between the retinal position of correspondent items: the disparity. This problem is considered the main difficulty in stereo vision (Figure 1).

- Reconstruction of the three-dimensional scene, that is, obtaining the depth from the disparity.

1. The correspondence problem

This problem can be seen as a search problem: given an element in the left image, a corresponding element in the right image must be searched.

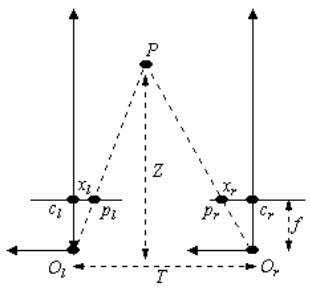


Fig. 1. p_l and p_r correspond to 3D point *P*

From a human point of view, the process of stereoscopic vision is so natural that we cannot appreciate its complexity unless we try to automate it. Uniqueness, smoothness and epipolar geometry are some of the physical constraints that are usually established to make the correspondence search easier.

1. Epipolar geometry: for a given point in the left image, all possible matching points in the right image lie on a line. Therefore, the dimension of the space of search is reduced from two to one dimension. The epipolar constraint is, of course, symmetrical, that is, for a point in the right image, all possible matching points in the left image also lie on a line.
2. Uniqueness: when we are working only with opaque objects, a point in the left image should have only one matching point in the right image, at the most. This is not true in general for transparent objects.
3. Smoothness: this constraint is based on the fact that the world is mainly made of smooth surfaces.

The correspondence algorithms are usually classified into two main groups: those based on correlation and others based on features.

In correlation based methods, the elements to match are fixed-size windows in the image. The similitude criterion is a measure of the correlation between windows in both images. The corresponding element is given by the windows that maximize the similitude criterion within a search region.

The feature based algorithms firstly extract some predefined features and then, they try to match them.

The correspondences based on regions are included in the family of feature based methods. In general, the higher the semantic level of the primitive, the more robust the obtained correspondences are, although some important drawbacks can appear: extracting the primitives can be more difficult and the disparity map is more disperse. In [8] an interesting review of the region based correspondence problem is shown.

A widely used technique is dynamic programming [5], [6] and [2]. These kinds of algorithms are characterized by a global cost function that is minimized.

For a long time several researchers have considered the possibility of including multiresolution models for the detection of correspondence in a stereo pair in order to obtain an estimation of depth in a three-dimensional scene. In [12] an integrated scheme including multiresolution is used. In this example a new approach is formulated and

developed. It integrates several units implied in stereo vision: feature extraction, matching and interpolation. An energy function is built for every unit and every resolution and it is minimized in an integrated manner so that it produces a dense disparity map.

In [7] the authors worked simultaneously with several scales of the image and obtained an error function defined for every scale. They used and compared two methods (GRAPHSEARCH algorithm [11] and gradient fall) to find the correct correspondence between two images.

Other researches have found that better results are obtained using more than two images. Ayache [1] describes a trinocular stereo system in which the initial correspondences between features for cameras 1 and 2 are tested in a verification step that examines a specific point in the third image.

In section 2 an energy function is described to formulate the correspondence problem. The algorithm that is used to minimize the energy function and to obtain the disparity map is explained in section 3. Section 4 presents the applied multiresolution scheme. Finally, in section 5 some experiments are shown.

2. Energy function

The energy function minimized by the simulated annealing algorithm is made up of five terms, each one weighted by a control parameter (γ_n). Eq. (1) defines this function.

$$U(p) = \sum_{n=1}^{5} \gamma_n U_n(p) \qquad \textbf{(1)}$$

Notation is shown in Table 1.

Table 1. Notation of the energy function

II	Left image
ID	Right image
(p_x,p_y)	Coordinates of pixel p
N(p)	Neighbourhood environment of p
vL(p), vR(p)	Vertical edges of the images, it has a value of 1 if there is an edge between the pixels (p_x,p_y) and (p_x,p_y-1), and 0 else.
hL(p), hR(p)	Horizontal edges of the images, it has a value of 1 if there is an edge between the pixels (p_x,p_y) and (p_x-1,p_y), and 0 else
disp	Disparity map
$\delta(a,b)$	Function that returns 1 if a=b and 0 else
τ	Absolute value of the difference of the grey level between two points
$\Xi(p,N(p))$	Census transform of p
H(v1,v2)	Hamming distance between two vectors of bits

The term U_1 is the correspondence cost at intensity level at the selected pixel. Instead of comparing a pixel in the left image with a pixel in the right image, a neighbourhood environment around the pixel p is considered.

$$U_1(p) = \sum_{N(p)} \tau(p,q) \,|\, (p \in II) \wedge (q \in ID) \wedge (p_x = q_x) \wedge (q_y = p_y + disp(p)) \quad (2)$$

In previous works [4] the squared difference between intensity values has been used instead of the absolute value of the difference, but we have verified that the absolute value works better when there are outliers.

The term U_2 is the correspondence cost of the Census transform [14]. It is a non parametric measure of the local special structure. The value of this transform depends on the comparison of the intensity value of a pixel with the intensity values of the pixels in the neighbourhood. The value of the transform for a pixel in the left image must be very similar to that of the corresponding pixel in the right image.

$$U_2(p) = \sum_{N(p)} H(\Xi(p,N(p)),\Xi(q,N(q))) \,|\, (p \in II) \wedge (q \in ID) \wedge (p_x = q_x) \wedge (q_y = p_y + disp(p)) \quad (3)$$

The term U_3 is the correspondence cost at edge level. Information about horizontal and vertical edges is used. It can be assumed that if there is an edge in the left image, there should also be an edge in the right image, corresponding to the first one but with a displacement of the amount of pixels determined by the disparity.

$$U_3(p) = \sum_{N(p)} (1 - \delta(vL(p), vR(q))) + (1 - \delta(hL(p), hR(q))) \,|\, (p \in II) \wedge (q \in ID) \wedge (p_x = q_x) \wedge (q_y = p_y + disp(p)) \quad (4)$$

The term U_4 refers to the smoothness constraint: it is assumed that the disparity varies in a smooth manner between edges. This term does not assume the smoothness when a vertical edge is found in the image. We can expect to have different disparity values in two positions when there is an edge between them.

$$U_4(p) = (disp(p) - disp(q))^2 * (1 - vL(p)) + (disp(p) - disp(r))^2 * (1 - vL(r)) \,|\, (p_x = q_x) \wedge (q_y = p_y - 1) \wedge (p_x = r_x) \wedge (r_y = p_y + 1) \quad (5)$$

The last term includes the uniqueness constraint. If we are working only with opaque objects, every point in the left image should have a unique corresponding point in the right image. This is not true for transparent objects. It means that if the disparity in columns j^{th} and q^{th} is obtained for any row i then, as the uniqueness constrain states, $j+disp(i,j) \neq q+disp(i,q)$.

$$U_5(p) = \sum_{q=ini}^{fin} \delta(p_y + disp(p), q_y + disp(q)) \,|\, (p_x = q_x) \quad (6)$$

3. Simulated Annealing

A stochastic relaxation method called simulated annealing (SA) is used to obtain a global or quasi-global solution depending on a cooling factor. The algorithm tries to minimize an energy function that includes a measure of the similitude error between corresponding points. There are many versions of SA algorithm: Metropolis algorithm, Creutz algorithm, Boltzman machine, Gibbs Sampler, etc. In this study, we have used the **Metropolis algorithm** [10]. Every pixel is visited and its disparity value is modified by another value belonging to a maximum range of disparity.

Considering the previously defined energy function, the algorithm is applied iteratively. The proposed algorithm is shown in Table 2.

Table 2. SA algorithm

Pas 1	Assign initial temperature T
Pas 2	For each pixel p in the disparity map 1. Change its disparity value for another in the given range 2. Calculate $\triangle U$ 3. If $\triangle U < 0$, accept the new value; else, accept if $e^{-\triangle U/T} > \xi$, where ξ is a random value in [0,1].
Pas 3	Cool the system by $0<k<1$ so that $T_{k+1}=kT_k$ and go to step 2 during a fixed amount of iterations

4. Multiresolution

The multiresolution scheme is usually represented as a pyramidal structure (Figure 2) where the peak of the pyramid represents the maximum level of scale and the base is the image in its original scale.

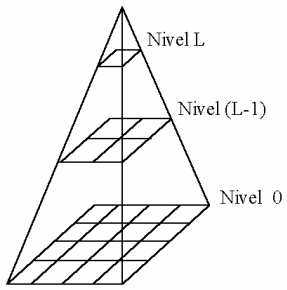


Fig. 2. Pyramidal structure

The multiresolution methods are based on the analogy that can be establish between the operations done in a rough grid of a region and the more global calculations done in a finer grid of the same region [13]. Two ways of transforming a grid into another of a different resolution level can be considered:

- By **sampling**: From the unscaled image some pixels are selected. This procedure is shown in Figure 3.
- By **block averages**: Some blocks are formed from the unscaled image and their average is calculated. Then the scaled image is built with these values.

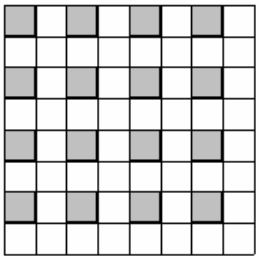


Fig. 3. Resolution transformation by sampling

In our studies we have used the sampling method.

The proposed method to estimate the disparity can be described as follows:

- At the roughest level of resolution (pyramid peak) the optimal solution can be obtained quickly because of the small amount of elements that are in the allowed disparity space.
- At intermediate levels of resolutions, the previous level solution is used to interpolate an initial estimation. The method is then applied to calculate the optimal solution.
- The same process in continued until level 0 of complete resolution (pyramid base) is achieved.

A median filter is applied at every disparity estimation resulting from each scale level. The median filter make the intensity of the pixels smoother in relation to their neighbourhood, eliminating the isolated outliers.

4.1. Interpolation

Multiresolution models require an interpolation technique that allows the results of the method at one level to be used at the next level. Some interpolation techniques have been defined: linear interpolation, Bessel interpolation, Hermite interpolation, and so on. We have used linear interpolation.

5. Experiments and results

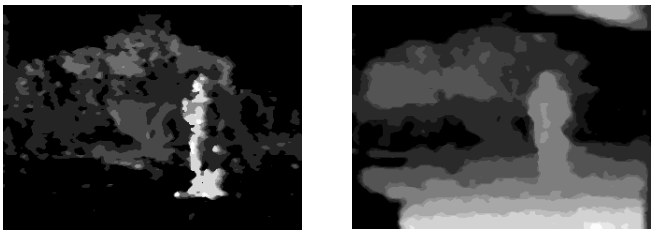
Some experiments conducted with real images are shown. The images belong to several types, both interior and exterior. The main goal of selecting such different kinds of images is to prove the flexibility of our method.

The images have been taken with a Digiclops interface IEEE 1394 camera with a resolution of 240×320 pixels. In the different experiments, the images have been scaled by a factor of 2^4 to apply the multiresolution scheme. Moreover, a range of maximum disparity has been fixed in every stereo pair.

The first example is shown in Figure 4. The parameters that have been used are: γ_1=1, γ_2=150, γ_3=150, γ_4=100 i γ_5=150. During the scale process we have worked with images at a resolution of 120×160 (scale 2^1), 60×80 (scale 2^2), 30×40 (scale 2^3) and 15×20 (scale 2^4). The allowed disparity range is 25. Figure 5 (a) shows the disparity map using the model without multiresolution. The results obtained using multiresolution with 5 levels is presented in Figure 5 (b). In this figure and in the others, the brighter the intensity is, the nearer the object is. The darkest areas represent the most distant objects.

(a) Left image (b) Right image

Fig. 4. Pair of stereo images



(a) Without multiresolution (b) With multiresolution

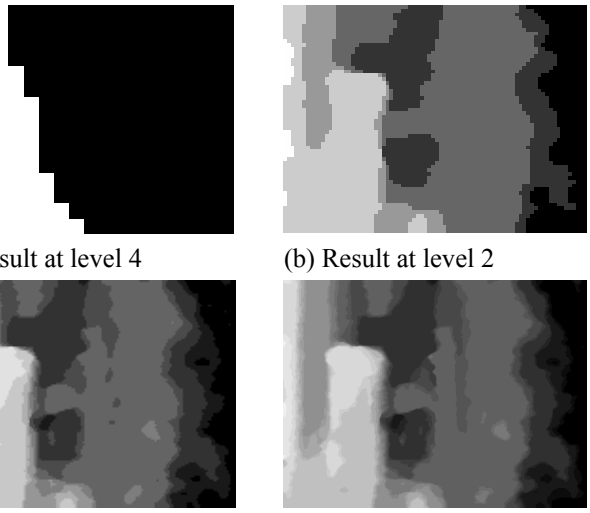
Fig. 5. Disparity maps for the stereo map in Figure 4

The second experiment presents a stereo map taken in an interior scene (Figure 6). Figure 7 shows the intermediate disparity maps resulting from applying the algorithm at several resolutions. The parameters are the same as in the previous experiment. The final result is shown in Figure 6 (d). The allowed range of disparity is 40.

(a) Left image (b) Right image

Fig. 6. Pair of stereo images for experiment 2

The execution time for both stereo pairs is 23 seconds. This time refers to an Athlon XP 1700 processor.



(a) Result at level 4 (b) Result at level 2

(c) Result at level 1 (d) Final result

Fig. 7. Intermediate disparity maps for stereo pair 2

In some previous studies, we have worked with energy functions applied to the complete image, but we have verified that using an energy function defined for pixels decreases the computational cost of the method.

We have included intensity features, edges, non parametric transforms, smoothness constrains and uniqueness restrictions to construct a robust energy function. The fact that humans are able to determine the disparity in a better way when there are edges can constitute evidence that a mechanism based on edges must be included in a stereo algorithm.

We believe that colour features would also allow a better discrimination of pixels, and so we are working to adapt this model to colour images.

6. References

[1] N. Ayache. Artificial Vision for Mobile robots: Stereo Vision and Multisensory Perception. The MIT Press, 1990.

[2] P. Belhumeur and D. Mumford. "A bayesian treatment of the stereo correspondence problem using half-occluded regions". Proc. International Conference on Computer Vision and Pattern Recognition IEEE, 1992

[3] C. Chang and S. Chatterjee. "Multiresolution stereo – A bayesian approach". International Conference on Pattern Recognition, pp. 908-912, 1990.

[4] P. Compañ, R. Satorre, C. Villagrá and R. Rizo "Visión estereoscópica en un modelo multirresolución". Actas de la IX Conferencia de la Asociación Española para la Inteligencia Artificial, pp. 1291-1300, 2001

[5] I. J. Cox, S. L. Hingorani and S. B. Rao. "A maximum likelihood stereo algorithm. Computer Vision and Image Understanding. Vol 63:3, pp. 542-567. 1996.

[6] D. Geiger, B. Ladendorf and A. Yuille. "Occlusions and binocular stereo". International Journal of Computer Vision, 14, pp 211-226, 1995

[7] M. Lew, K. Wong and T. Huang. "Multi-scale stero matching". Int. Conference on Pattern Recognitiion, pp. 600-623, 1992.

[8] M. A. López. "Visión estereoscópica basada en regiones: estado del arte y perspectivas de futuro". Actas de IX Conferencia de la Asociación Española para la Inteligencia Artificial, 2001.

[9] D. Marr and T. Poggio. "A theory for human stereo vision". Proceeding Roy. Soc Lond. B. pp. 301-328. 1979.

[10] N. Metropolis. "Equation of state calculations by fast computing machines". Journal Chem., 21, pp. 1087-1091,1953.

[11] N. Nilsson. Principios de Inteligencia Artificial. Díaz de Santos, 1987.

[12] K. Sunil and U. Desai. "New algorithms for 3D surface description form binocular stereo using integration". Journal of the Franklin Institute, 1994.

[13] D. Terzopoulos. "Image analysis using multigrid relaxation methods". IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986

[14] R. Zabih and J. Woodfill. "Nom-parametric transforms for computer visual correspondence". Proccedings of the Third European Conference on Computer Vision, pp 151-158, 1994.