A cybernetic multiresolution system for disparity estimation in stereo vision*

Patricia Compañ, Rosana Satorre y Ramón Rizo
{patricia, rosana, rizo}@dccia.ua.es
Grupo i3a: Informática Industrial e Inteligencia Artificial
Departamento de Ciencia de la Computación e Inteligencia Artificial
Universidad de Alicante

Abstract

This paper presents a new stereo correspondence algorithm based on an integrated model that incorporates different modules corresponding to every stage. Firstly, original images are scaled in order to reduce its size. A disparity map, which is the basis of the process, is obtained from these reduced images. The disparity map is obtained by building and minimizing an energy function under a multiresolution scheme. The energy function integrates grey level characteristics, non parametric transforms, edges, smoothness and unicity. The disparity obtained at each level of resolution is interpolated and incorporated to the next level. Our model produces a dense disparity map. The algorithm has been tested with several kinds of real images to show its flexibility and a metrics to evaluate the quality of a disparity map is proposed.

Key words: stereo vision, disparity, *simulated annealing*, multiresolution, energy function, interpolation.

Introduction

Because the localization and function of eyes in the human body, our brains receive two very similar images of a scene, taken from two separated points, situated extremely close together on a horizontal line. Two objects, at different distances from the observer, present different relative positions in their retinal images. The brain can measure this displacement (retinal disparity) and use it for estimating depth (Marr, 1979). The disparity of every point in the image forms the so-called *disparity map.* The retinal disparity depends on the distance to the fixation point. To make use of binocular depth cues, an organism must have a binocular visual field: a region of overlapping visibility for the two eyes. Each animal has different extents of binocular visual fields. In general, predators have both eyes at the front of their heads, and consequently have large binocular visual fields. In contrast, prey typically have one eye on either side of their heads, and consequently have small, if any, binocular visual fields.

* Supported by the Generalitat Valenciana, project number GV04B617

Stereo vision is the set of techniques used to extract 3D information from a scene of two or more images taken from different viewpoints. A stereo system must solve several problems:

- Camera calibration, that is, estimating the values of the intrinsic and extrinsic parameters of the camera model.
- Rectification of the epipolar geometry to simplify the search on a scanline.
- Correspondence between tokens of the images. This is the problem of determinating which items in the left image correspond to those in the right image. Our three-dimensional perception of the world is due to our ability to interpret the differences in retinal location between corresponding objects. This problem is considered the most difficult part of the stereo problem (Figure 1).
- Reconstruction of the 3D scene, that is, the calculation of the depth from the disparity.

The correspondence problem

This problem can be seen as a search problem: given an element in the left image, the corresponding element in the right image must be found. From a human point of view, the process of stereovision is so natural that its complexity is not appreciated until attempts are made to automate the process. Uniqueness, smoothness and epipolar geometry are some of the physical constraints that are imposed in the stereo problem to restrict the space of solutions and to obtain a well-posed problem.

1. Epipolar geometry: for a given point in the left image, its possible matchings lie on a line on the right image. As a consequence of this fact, the search space dimension is reduced from two dimensions to one. Obviously, the epipolar restriction is symmetric, that is, for a given point in the right image, its possible matchings in the left image lie on a line too.
2. Uniqueness: if we limit ourselves to opaque objects, each point in the left image has just a single corresponding point in the right image. This is not necessarily true for transparent objects.
3. Smoothness: this restriction is based on the fact that the world is composed principally of smooth surfaces.

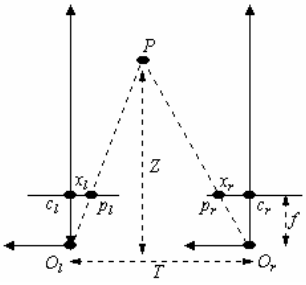


Figure 1: p_l and p_r are corresponding points of 3D point *P*

The correspondence algorithms are classified in two classes: correlation-based and feature-based methods.

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

Feature-based methods firstly extract features and then try to match them. Region-based correspondence is included within the family of feature-based methods. Generally, the higher the semantic level of the primitive, the more robust the matchings. Nevertheless there are important disadvantages: the primitive extraction is more difficult and the disparity map is sparser. In paper (López, 2001) an interesting review of region-based correspondence is shown.

A widely used technique is dynamic programming (Cox, Hingorani, Rao, 1996) (Geiger, Ladendorf, Yuille, 1995). These algorithms are characterized by a global cost function that is minimized.

Some authors have considerated the idea of incorporate multiresolution schemes to detect stereo correspondence with the objective of obtaining and estimating of the depth of a scene. Multi-scale or coarse-to-fine scheme, is a method of efficiently and effectively representing data with the objective of reducing the computational complexity, (Caspary, Zeevi, 2002). At each scale the results of the previous scales are used as the initial estimation.

Stochastic relaxation methods have received a lot of attention in the field of stereovision. These techniques try to obtain an optimal solution to the correspondence problem without falling in a local minimum, (Vogiatzis, Torr, Cipolla, 2003).

Other authors have observed that better results are obtained employing more than two images, (Agraval, Davis, 2002). García, Batlle, and Salvi (Garcia, Batlle, Salvi, 2002) use a trinocular system to estimate both the position and velocity of known objects.

Application to mobile robotics

In the field of design and construction of mobile vehicles able to navigate through unknown environments, it is desirable to incorporate a system to calculate the distance between the vehicle and the objects in the scene. Although environment maps are available, they can change, so the robot must be able to react.

By estimating the disparity from a stereoscopic pair, the system can obtain the distance to the objects and so react, changing its direction of movement. Eq. (1) shows the relation between disparity and object-camera distance.

$$Z = f\frac{T}{d} \quad \textbf{(1)}$$

where *Z* represents the distance from the camera to the object, *f* the camera focal length, *T* the base line and *d* the disparity obtained from the correspondence algorithm.

In later sections, an energy function is described to formulate the correspondence problem, the proposed algorithm to minimize the energy function and to obtain the disparity map is described and the applied multiresolution scheme is indicated. Finally, some experiments and results are shown.

Energy function

The energy function minimized by the *simulated annealing* algorithm is composed by five terms, each weighted by a control parameter (γ_n). Eq. (2) defines this function.

$$U(p) = \sum_{n=1}^{5} \gamma_n U_n(p) \quad \textbf{(2)}$$

The used notation is shown in Table 1.

II	left image
ID	right image
(p_x,p_y)	coordinates of pixel p
N(p)	neighbourhood of p
vL(p), vR(p)	vertical edges of the images, it contains a value 1 if there is an edge between the pixel (p_x,p_y) and the pixel (p_x,p_y-1), and 0 otherwise
hL(p), hR(p)	horizontal edges of the images, it contains a value 1 if there is an edge between the pixel (p_x,p_y) and the pixel (p_x-1,p_y) and 0 otherwise
disp	the disparity map
$\delta(a,b)$	function returning the value 1 if a = b and 0 otherwise
τ	absolute value of the difference of grey levels between two points
$\Xi(p,N(p))$	Census transform of p
H(v1,v2)	Hamming distance between two vectors of bits

Table I: Notation of the energy function

The term U_1 is the grey-level matching at the selected pixel. Instead of comparing one pixel in the left image with just one pixel in the right image, a neighbourhood around pixel *p* is considered.

$$U_1(p) = \sum_{N(p)} \tau(p,q) \,|\, (p \in II) \wedge (q \in ID) \wedge (p_x = q_x) \wedge (q_y = p_y + disp(p)) \quad \textbf{(3)}$$

In previous investigations (Compañ, Satorre, Villagrá, Rizo, 2001) the square difference between intensity values instead of the absolute value of the difference has been used. It has been noticed that the use of the latter improves the results in the presence of atypical values (*outliers*).

The term U_2 is the **transform Census** matching cost (Zabih, Woodfill, 1994). The transform is a non-parametric measure of local intensity. The value of the transform is calculated comparing the intensity value of a pixel with the value of the neighbourhood. The transform must give similar results near corresponding points between the two images.

$$U_2(p) = \sum_{N(p)} H(\Xi(p, N(p)), \Xi(q, N(q))) \mid \\ (p \in II) \wedge (q \in ID) \wedge (p_x = q_x) \wedge (q_y = p_y + disp(p)) \quad \textbf{(4)}$$

The term U_3 is the edge-level matching cost. We use information from both vertical and horizontal edges. It can be supposed that if an edge exists in the left image, then there should also be another edge in the right image corresponding to the first one, but in a position displaced by the amount of pixels determined by the disparity.

$$U_3(p) = \sum_{N(p)} (1 - \delta(vL(p), vR(q))) + (1 - \delta(hL(p), hR(q))) \mid \\ (p \in II) \wedge (q \in ID) \wedge (p_x = q_x) \wedge (q_y = p_y + disp(p)) \quad \textbf{(5)}$$

The term U_4 is the smoothing constraint: it is assumed that disparity varies smoothly between edges. This term switches off smoothing whenever a line field is encountered in the image. Clearly, the disparity value at two locations cannot be similar when there is an edge between them.

$$U_4(p) = (disp(p) - disp(q))^2 * (1 - vL(p)) + (disp(p) - disp(r))^2 * (1 - vL(r)) \mid \\ (p_x = q_x) \wedge (q_y = p_y - 1) \wedge (p_x = r_x) \wedge (r_y = p_y + 1) \quad \textbf{(6)}$$

The final term incorporates the uniqueness constraint. It means that along any row *i,* by calculating the disparity at the j^{th} and the q^{th} column then according to the uniqueness constraint, $j+disp[i,j] \neq q+disp[i,q]$.

$$U_5(p) = \sum_{q=ini}^{fin} \delta(p_y + disp(p), q_y + disp(q)) \mid (p_x = q_x) \quad \textbf{(7)}$$

Simulated annealing

A well-known stochastic relaxation method, called simulated annealing (SA) has been used to obtain the global or, nearly global, optimum solution depending on the annealing schedule. The

algorithm tries to minimize an energy function that incorporates a similitude error measure between corresponding points. There are many versions of SA: Metropolis algorithm, Creutz algorithm, Boltzman machine, Gibbs Sampler, etc. In this paper, the **Metropolis** algorithm (Metropolis, 1953) has been used. Each pixel is visited and the value of the disparity within a given range is updated. Considering the energy function previously defined in Section 2, the algorithm is applied iteratively. The proposed algorithm is shown in Table 2.

Step 1	Assign initial temperature T
Step 2	For each pixel p in the disparity map, 1. Change its value to any of the range of disparity 2. Calculate $\triangle U$ 3. If $\triangle U < 0$, accept it; else, accept if $e^{-\triangle U/T} > \xi$, where ξ is a randomly generated number between [0,1].
Step 3	Lower the temperature by $0<k<1$ such $T^{k+1}=kT^{k}$ and go back to Step 2 a fixed number of iterations.

Table II: S*imulated annealing* algorithm

Multiresolution scheme

In this section, a multiresolution structure is proposed that will speed up the process that initialises the correspondence problem with a solution coming from a lower resolution. The multiresolution scheme is usually represented by a pyramidal structure (Figure 2) whose peak corresponds to the maximum employed level of scale and whose base represents the image at its original scale.

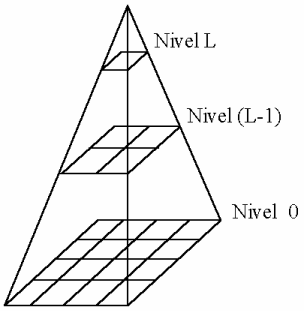
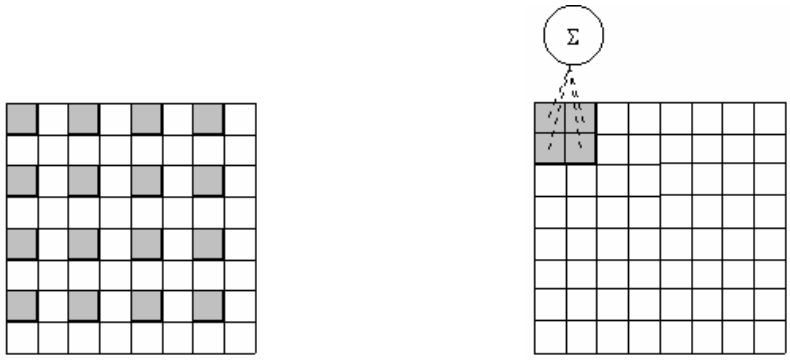


Figure 2: Pyramidal structure

Computations on a coarse grid over a given region are analogous to global computations on a fine grid over the same region (Terzopoulos, 1986). Two types of resolution transformation models are considered:

- Sampling. Some pixels are selected of the original image. Figure 3 (a) shows the process.

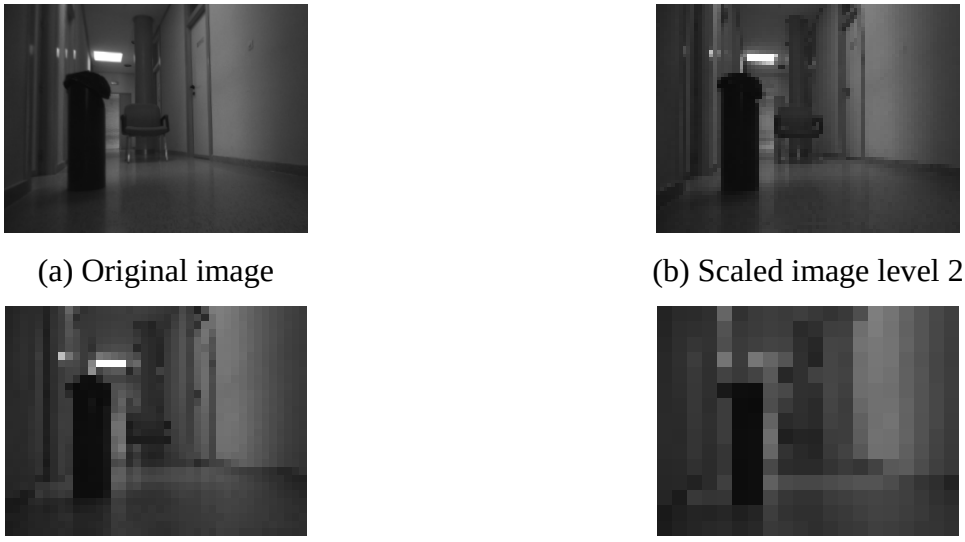
- Block-to-Point. In the original image, blocks are formed and their arithmetic means are calculated. The scaled image is made with these values. The method is illustrated on Figure 3 (b).



(a) Sampling (b) Block-to-Point

Figure 3: Resolution transforms

In this paper sampling methods for scaling are applied: some pixels are selected form the original image. The original images are scaled down by the power of two. The process consists of selecting the first row and deleting the next 2^{n-1} rows, and so on. The process concerning the columns is analogous. Figure 4 shows an example of this scaling method.



(a) Original image (b) Scaled image level 2

(c) Scaled image level 3 (d) Scaled image level 4

Figure 4: Scaled images at different levels

Interpolation

A multiresolution method requires an interpolation technique to allow the use of results from one level of scale in the following level. There are several techniques of interpolation: linear interpolation, Gauss-Seidel, and so on. In this paper linear interpolation has been used.

Starting at the coarsest resolution and obtaining the initial disparity map at that level, the optimal solution is found quickly due to the limited number of elements in the disparity range.

At the following resolution level a new disparity map is generated by interpolating values of pixels obtained from the previous map. Disparity obtained at the previous level is assigned to the corresponding points at the current level, so only some points will have a value. Next, gaps must be filled in using linear interpolation. This map is used as the initial configuration. The algorithm is applied to obtain a disparity map at that resolution. The process is then repeated until level 0 (original image) is reached.

Being $disp_s^{(k)}$ disparity obtained for point *s* at level *(k)*, the process is done as follows:

$$disp_s^{(k-1)} = \begin{cases} 2 \cdot disp_{i/2,j/2}^{(k)} & (i \bmod 2 = 0) \wedge (j \bmod 2 = 0) \\ (disp_{i,j-1}^{(k-1)} + disp_{i,j+1}^{(k-1)}) / 2 & (i \bmod 2 = 0) \wedge (j \bmod 2 = 1) \\ (disp_{i-1,j}^{(k-1)} + disp_{i+1,j}^{(k-1))/2} & (i \bmod 2 = 1) \end{cases} \quad \textbf{(8)}$$

Median filtering is a well-known technique for removing salt-and-pepper noise from images. Each disparity estimation at each level is median filtered. The median filtering step is suitable to correct outlier disparity estimations that deviate from the correct expected estimation (a form of smoothness constraint on the estimations).

Experiments and conclusions

Here some results using real images are presented. Every image depicts exterior and interior scenes taken from real life. By selecting several different types of images, the intention has been to prove the flexibility of the method.

The images were taken with a Digiclops camera interface IEEE 1394 with a resolution of *320x240* pixels. In this section three of the experiments are presented. Images were scaled by a factor of 2^4 to apply the multiresolution scheme, and the search space was limited to a reduced interval.

Quality of the disparity map

Some authors (Scharstein, Szeliski, 2002) use the real disparity map to validate the obtained result. Nevertheless, as our aim is navigation so the most important factor is the distance to the objects. The process to evaluate the quality of the results is as follows. When the images are taken, the distance to the camera from each relevant object is measured. A segmentation algorithm is applied to the disparity map obtained the distance to the objects calculated using the relation between disparity and depth, and so a comparison can be made between the obtained distance and the manual measurement.

The quality measure that we present, Δ_{disp}, gives the error made when the disparity map is calculated. Once the distance from the camera to the relevant objects in the scene is obtained, eq (9) is calculated.

$$\Delta_{disp}(Z,R) = \frac{\sum_{i=1}^{n} | Z_i - R_i |}{n} \quad \textbf{(9)}$$

where Z_i represents the obtained distance for region *i* applying ec (1), R_i the real distance from the object to the camera and *n* the number of objects under consideration in the scene.

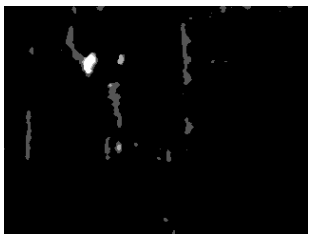
Experiments

Figure 5 shows a first example where images were scaled by a factor of 2^4 before the application of the algorithm. The parameters used were: γ_1=1, γ_2=100, γ_3=100, γ_4=100 and γ_5=150. During the scale process, images of 120 x 160 (scaled by por 2^1), 60 x 80 (scaled by 2^2), 30 x 40 (scaled by 2^3) y15 x 20 (scaled by 2^4) were used. The search space was 25 pixels. Figure 6 (a) shows the disparity map without appliying multiresolution. The result employing a five-level multiresolution pyramid is shown in Figure 6 (b).

(a) Left image

(b) Right image

Figure 5: Outdoor stereo pair



(a) Without multiresolution

(b) With multiresolution

Figure 6: Disparity maps of stereo pair 1

Figure 7 shows the result of applying the segmentation algorithm to the disparity map in Figure 6 (b). Table 3 presents the distances to the objects calculated by Eq. 1 and manually.

Region	Real distance	Calculated distance	Δ_{disp}
1	2.15	2.18	0.05
2	3.20	3.27	

Table III: Distance to the camera of relevant objects of stereo pair 1

(a) Region 1

(b) Region 2

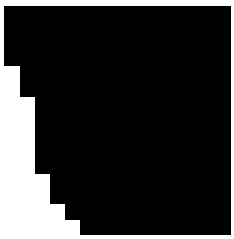
Figure 7: Regions obtained by segmentation algorithm

Figure 8 shows an example of indoor scene. Figure 9 presents the intermediate disparity maps resulting when the algorithm is applied at several resolutions. The values of the parameters are the same as in the previous experiment. The final result is presented in Figure 9 (d). The allowed search space was 40 pixels.

(a) Left image

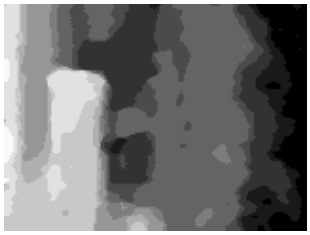
(b) Right image

Figure 8: Indoor stereo pair

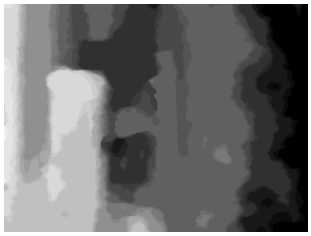


(a) Result level 4

(b) Result level 2



(c) Result level 1



(d) Final result

Figure 9: Intermediate disparity maps of stereo pair 2

Figure 10 presents the result of appliying segmentation algorithm to the disparity map in Figure 9 (d). Table 4 presents the distances to the objects calculated by Eq. 1 and manually.

Region	Real distance	Calculated distance	Δ_{disp}
1	1.50	1.54	0.04
2		2.18	

Table IV: Distance to the camera of relevant objects of stereo pair 2

(a) Region 1 (b) Region 2

Figure 10: Regions obtained by segmentation algorithm

The execution time for all experiments was 8 seconds with the time measured using a Pentium 2.40 Gz processor.

The next experiment is shown in Figure 11. The disparity map is presented in Figure 12 (a). The result of applying the segmentation algorithm is shown in Figure 12 (b). The distance to the object measured manually is 2.50 meters whereas the distance obtained applying Eq. 1 is 2.62 metres, Δ_{disp}=0.12.

(a) Left image (b) Right image

Figure 11: Outdoor stereo pair

(a) Disparity map (b) Region

Figure 12: Disparity map and region obtained

The following conclusions can be highlighted:

- In previous analysis, energy functions applied to the whole image have been used, but it has been observed that using a pixel-by-pixel energy function reduces the computational cost.
- Intensity features, non-parametric transforms, edges, smoothness and uniqueness for building a robust energy function are incorporated.
- The information obtained in previous executions of the algorithm is employing for enhancing the initial configuration for the next iteration. This is why a multiresolution scheme by sampling with linear interpolation is adopted. To eliminate outliers, median filtering is applied.
- A new quality measure for the disparity map is proposed, that allows the quantification of the quality of the disparity map.

At present the method is being adapted to work with colour images. Results will be published shortly. To obtain more precise measures of disparity, we are interested in incorporating subpixel precision to the method.

REFERENCES

Agraval M., Davis L.S. (2002), “Trinocular Stereo Using Shortest Paths and the Ordering Constraint “, IJCV, vol 47, number 1-3, pp 43-50.

Caspary G., Zeevi Y.Y. (2002) “Wavelet-based multiresolution stereo vision”, Pattern Recognition, Proc. 16th International Conferencce on Pattern Recognition (ICPR02), vol3 , pp 680-683.

Compañ P., Satorre R., Villagrá C., Rizo R. (2001), “Visión estereoscópica en un modelo multirresolución”, Actas de la IX Conferencia de la Asociación Española para la Inteligencia Artificial, pp. 1291-1300.

Cox I, Hingorani S., Rao S. (1996), “A maximum likelihood stereo algorithm”, Computer Vision and Image Understanding, Vol 63:3, pp. 542-567.

García R., Batlle J. and Salvi J. (2002), “a new approach to pose detection using a trinocular stereovision system”, RealTimeImg, vol 8, number 2, pp 73-93.

Geiger D., Ladendorf B., Yuille A. (1995), “Occlusions and binocular stereo”, International Journal of Computer Vision, 14, pp 211-226.

López M. (2001). “Visión estereoscópica basada en regiones: estado del arte y perspectivas de futuro”, Actas de IX Conferencia de la Asociación Española para la Inteligencia Artificial.

Marr D., Poggio T. (1979), "A theory for human stereo vision", Proceedings Royal Society London B. pp. 301-328.

Metropolis N. (1953), “Equation of state calculations by fast computing machines”, Journal Chemical, 21, pp. 1087-1091.

Scharstein D., Szeliski R. (2002), “ A taxonomy and evaluation of dense two-frame stereo correspondence algorithms”, IJCV 47 (1/2/3): 7-42.

Terzopoulos D (1986), “Image analysis using multigrid relaxation methods”, IEEE Transactions on Pattern Analysis and Machine Intelligence.

Vogiatzis G., Torr P., Cipolla R. (2003), “Bayesian stochastic mesh optimization for 3D reconstruction”, Proc. 14th British Machine vision Conference, pp 711-718,

Zabih R., Woodfill J. (1994), “Nom-parametric transforms for computer visual correspondence”, Proceedings of the Third European Conference on computer Vision, pp 151-158.