Object Segmentation Scheme for a Mobile Robot

Aldavert Miró, David
Ramisa Ayats, Arnau
López de Màntaras, Ramon
Toledo Morales, Ricardo
Centre de Visió per Computador, Universitat Autònoma de Barcelona
Institut d'Investigació en Intel·ligència Artificial, CSIC

Abstract. Having a good representation of the environment is crucial in mobile robotics. Mapping methods are insufficient to model objects within the environment. Segmentation is a fundamental step to represent objects. In this paper we present a schema based on MPEG-4 segmentation techniques to segment objects of the scene using the depth and intensity.

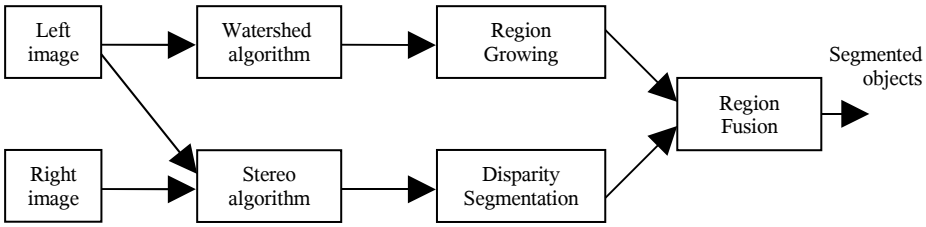
Keywords: Robot Vision, Object Segmentation, Stereovision.

Introduction

Having a good representation of the environment is basic in mobile robotics. Approximations with acceptable results were made using artificial landmarks but these methods are useless in unknown environments. Therefore extraction of natural landmarks arises as a challenging problem. The earliest approximations to solve this problem used information of depth range finders to build maps of unknown environments. These constructions had two implicit problems: map cannot be build because robot position is unknown due to accumulative errors in odometry, and position cannot be corrected with sensors data due to the unavailability of a map of the environment, which is unknown. The techniques that try to solve these two problems jointly are known as SLAM (Simultaneous Localization And Mapping). In the literature there are many techniques that attempt to solve the SLAM problem. Most of them can be categorized as algorithms based on Kalman filtering, Expectation Maximization algorithms and hybrid approximations [1, 2]. Although a map can be constructed using SLAM techniques, this information is not enough for the robot to have a higher semantic knowledge of the environment. For example, this higher level knowledge may be necessary to facilitate user interaction as well as for a more accurate comprehension and interaction of the robot with its environment. Therefore more information than that given by a depth range finder is necessary. Images obtained by a stereo pair of cameras appear as an acceptable solution.

Our main objective is to develop a robot stereo vision system which is able to obtain its own landmarks in a completely unsupervised manner and our main problem is that the acquisition of natural landmarks must be done without any knowledge of the objects of the scene. Image segmentation can be considered the first step towards object detection. Some classical approaches to image segmentation are: morphological watershed [3], split and merge [4] or region growing [5]. However, real objects are composed of many

different colors and patterns. Consequently segmentation techniques will over-segments an object into multiple regions. Therefore we need a property that is uniform in the entire object such as depth or motion. However in depth or motion methods the object boundaries cannot be precisely estimated. For this reason, hybrid approaches have been proposed which combine color and depth/motion properties [6, 7, 8, 9].



Natural landmark extraction schema

Our proposed schema for segmenting the objects in the scene is inspired on Visual Objects segmentation techniques from MPEG-4. But in MPEG-4 the time restrictions are not as strong as in our case; therefore the algorithms at each step of the segmentation process are modified attempting to increase the speed, so that it can work in real time.

Disparity map estimation

In order to properly segment meaningful objects, our method uses depth to group the over-segmented intensity regions. To obtain the depth information of the scene, a disparity map is calculated with correlation-based stereovision. Although in the literature we can find other stereo algorithms with more accurate results, correlation-based stereo has a low computational cost, a regular structure and a fixed execution time, non depending of the scene contents, which makes it the most used method in real-time applications like mobile robot vision [13,15]. Moreover, recent versions of correlation-based stereo algorithms take advantage of the graphics accelerator hardware, producing more than 30 disparity maps per second without any supplementary cost for the computer CPU [10].

On the other hand, correlation-based methods are affected by some problems: Blurring of object borders, lack of texture, repetitive texture and occlusions. The first problem is caused by the violation of the constant depth assumption that occurs when the correlation window overlaps a depth discontinuity. In these zones the disparity value is usually assigned to the higher textured object. Usually this means that the nearer objects grow over the lesser textured background. In zones with lack of texture or where texture is very repetitive the correct disparity cannot be determined because the correlation function may have multiple local minima. Finally, occlusions are points of the scene that are seen only by one camera. In these areas the real disparity is impossible to determine. Many solutions to these problems are present in the literature, but most of them are too computationally expensive for our scheme.

An important choice for an area-based stereo algorithm is the correlation operator. The most used operators in real time stereo are Sum of Absolute Difference (SAD) and Sum of Squared Differences (SSD) [13]. These operators have a low computational cost, but they are very sensible to image noise. To reduce sensibility its normalized version is used. Although this increases the computational cost, the results obtained are much better.

Some constraints can be applied in order to simplify the disparity estimation. The continuity or smoothness constraint states that the surroundings of an image point have the same disparity as the interest point. This is true for all points in an image except for those near depth discontinuities. This constraint makes possible the use of windows to determine the correct disparity, but when a window violates this assumption incorrect matches are found. A well-known strategy is the epipolar constraint, which assures that the corresponding point relies at the same epipolar line. This restriction supposes a great reduction of computational cost, and makes possible the real-time stereo. However, the epipolar constraint is only true if the images are rectified [12]. But in mobile robotics, due to vibrations caused by the robot movement, the cameras relative position can change and the rectification procedure may fail. With low vibration rates one possible solution is extending the search of the maximum correlation to the lines surrounding the theoretical epipolar line. This can partially solve vibration problems but with the cost of highly increasing the computation time.

As we stated above the disparity estimation may have some problems. To compensate these problems a post-process step where the invalid matches are eliminated from the disparity map is necessary. Amongst the numerous techniques present in the literature we found three suitable for our purposes: First, the right-left consistency check, which is the most common technique for false matches rejection. This method consists in comparing the disparity values obtained in the direct disparity map (left image as reference) with its symmetric counterpart (right image as reference), and considering valid only those for which the match in the inverse map falls at the initial point in the direct map. Another technique consists in analyzing the correlation function obtained at each point to reject those that have more than one suitable minimum or a constant correlation function [11,14]. Finally, the Moravec operator detects the points without enough texture to properly find the correct match.

Another common problem for stereo algorithms in indoor environments is the specular reflection of the objects in the ground. That is, the stereo algorithm treats the reflected objects the same way as original objects, and assigns the same disparity to them. One way to solve this problem without finding explicitly symmetries, which is too time consuming for our algorithm, is to have an estimation of the ground plane disparity to check the matches found by our stereo algorithm. If one disparity map point has a value smaller than the theoretical ground plane disparity, then it must be erroneous because is under the ground. In the literature, the ground plane estimation is a common

technique for obstacle detection in mobile robotics and in assisted driving. This solution only works in indoor environments where ground can be approximated by a plane.

The reference ground plane can be thought as a plane-to-plane projective transformation between the two images [16]. Given the projection of at least four ground points in both images, we can estimate the parameters of the projective transformation. Then we can determine for each point of the left image the corresponding point of the theoretical ground plane in the right image. Finally the displacement between calculated points is the theoretical ground disparity.

Object Segmentation

The disparity map contains the information about the depth structure of the scene. Analyzing the distribution of the disparity we can segment the different objects of the scene, but the shape of the objects is deformed by the distortions mentioned in the previous section. To correctly extract the shape of the objects a segmentation is done in the intensity image and this segmentation results are fused with the results of the disparity field segmentation.

Intensity segmentation

For an initial estimation of intensity segmentation we use the morphological watershed that gives us regions of uniform intensity value of the image. This initial segmentation is used as a seed of a region growing algorithm.

The region-growing algorithm [6] fuses the regions given by the watershed using two criteria: proximity and homogeneity. The proximity criterion allows only neighboring regions to be fused. The homogeneity criteria is determined by a function that measures the similarity between regions. Each region is modeled as N pixels sampled from a normally distributed region with mean μ and standard deviation σ. Then two neighboring regions must satisfy two conditions to be fused. First, the mean difference between two regions must be lower than a maximum difference.

$$|\mu_1 - \mu_2| \leq \mu_{\max} \tag{1}$$

If the first condition is satisfied, the number of standard deviation (see equation 2) between the mean of the two regions is calculated. This is achieved by calculating the new mean of the fused region. Then the difference between the first region mean and the new mean is divided by the first region's standard deviation. This is compared with the

maximum number of standard deviation σ_{max} which is the maximum difference allowed between the first mean and new mean.

$$\frac{\mu_{new} - \mu_1}{\sigma_1} \leq \sigma_{\max} \tag{2}$$

Where μ_{new} is given by:

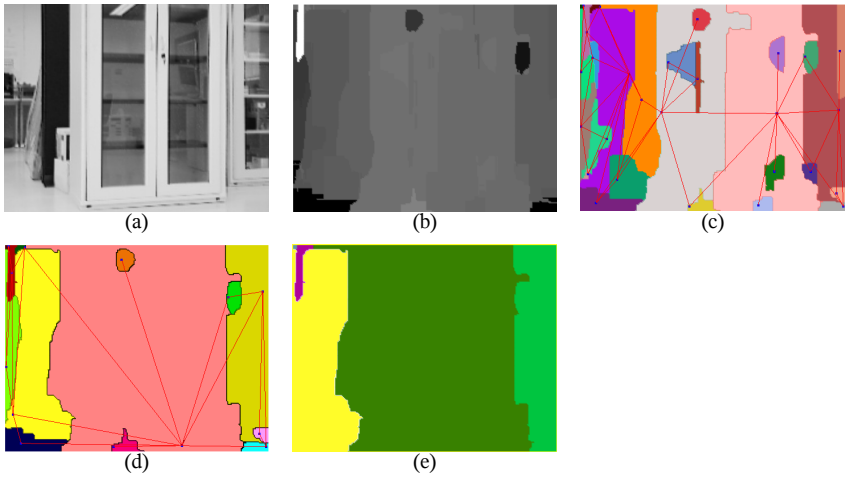
$$\mu_{new} = \frac{N_1\mu_1 + N_2\mu_2}{N_1 + N_2} \tag{3}$$

Where, N_1 and N_2 are the pixels contained in the two compared regions. The first comparison allows merging only regions with similar mean. The second comparison allows the absorption of small regions by large neighbors, but it ensures that regions of significant size are preserved. If the two comparisons are satisfied the two regions are fused, and the mean and the standard deviation of the new region is computed. The algorithm is performed iteratively until no new fusion is done. Finally, regions that are smaller than a given threshold N are fused with the neighbor that has the lower mean difference.

Disparity segmentation

In some approximations the disparity field segmentation is done by applying a classical method of segmentation similar to the region growing algorithm [6]. However this approach only works if the different objects have an homogeneous depth and it is significantly different from its neighbors. We propose a new region growing method for disparity map segmentation were we take advantage of the structural information of the scene. The outline of the method is as follows:

a) The watershed algorithm is used to segment the disparity map in regions of uniform disparity.
b) Then a graph is constructed where nodes are the disparity regions and arcs the neighborhood relations between disparity regions. The value assigned to each arc is the absolute difference between the mean of its nodes.
c) The region growing algorithm is applied, but instead of scanning the regions to fuse from left to right, they are scanned from higher to lower disparity.
d) Beginning with the higher disparity region, for each node the arc of lowest value (minimum mean difference between neighbor regions) is chosen, and the fusion criterions of region growing are applied. Then the second lowest arc is chosen, and so on.
e) If two regions are fused, the resulting new region has all the neighbors of the two regions and all the arc values are recalculated.
f) The process finishes when every node has been visited.
g) Finally the result is post-processed and all regions smaller than a certain number of pixels are fused with the most similar neighbor region.



Disparity map segmentation: (a) Image, (b) estimated disparity map, (c) disparity map regions and graph, (d) graph segmented disparity map, (e) final disparity map segmentation.

Region fusion

When the disparity field and intensity image are segmented, the results are combined to obtain more accurate objects. Being I the intensity region set and D the disparity region set, the region O of the objects in the scene is formed by the regions of I corresponding to the region of D whose area of intersection is maximized. Formally:

$$O_j = \arg\max\left\{A\left(I_i \cap D_j\right)\right\}, i = 1,2,\ldots N^i \tag{4}$$

where $A(\cdot)$ is the area and N^i the number of intensity regions. Then all the regions of I, corresponding to the same region of O, are merged in a new region. The set of new regions form the mask from which objects can be extracted.

Results and Conclusions

In this section we evaluate the performance of the proposed method. From four image sequences of two hundred frames each, acquired in our laboratory, we extracted four representative frames to show the performance of the method. The difference between manually segmented regions (we use as reference) and automatic segmented regions is analyzed. To compute the percentage error of two corresponding regions, the area of symmetric difference is divided by area of the intersection of the regions:

$$E = \frac{A(M) + A(R) - 2A(M \cap R)}{A(M \cap R)} \tag{5}$$

where M is the manually segmented region, R is the region obtained by our method and $A(\cdot)$ is the area of the region. In the following table the obtained results are shown. Rows correspond to images and columns correspond to the results of the indicated object in the segmented images.

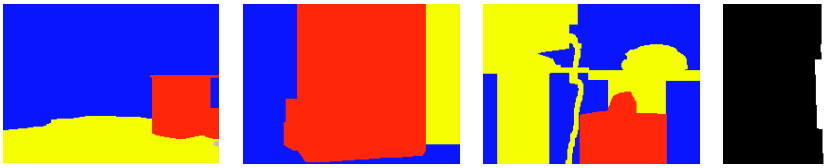
	Red object	Blue object	Yellow object
Image a	24,97 %	5,04 %	11,64 %
Image b	7,36 %	13,34 %	7,92 %
Image c	18,78%	74,74 %	41,83 %
Image d	4,11%	4,94%	----------

(a) (b) (c) (d)

4 selected sample images

(a) (b) (c) (d)

Resulting segmented images (in grayscale: blue = darker and yellow = lighter)



(a) (b) (c) (d)

Manually segmented images (in grayscale: blue = darker and yellow = lighter)

The results obtained in image “a”, and especially in image “c”, show the weakness of the method. In the calculated disparity map, the boundaries of the nearest objects have grown and two close objects may fuse in one disparity region. This can be seen in image “c” where parts of the blue object (background) are hidden by the overgrown yellow object. Other problems, that are inherent in uncontrolled environments, are due to: low texture ratios, zones badly illuminated and low contrast between objects.

Future work

We believe that our results can be refined taking advantage of the motion of the robot to obtain more accurate segmentation results.

Our method is fast enough to segment an image every 4 seconds using a 320x240 stereo pair images. However, more than 90% of the time is spent in the stereovision step. To speed it up, graphic card accelerators can be used to compute more than 30 disparity maps per second, as mentioned in the second section, making our algorithm suitable for real-time.

We pretend to evaluate the presented schema in outdoor environments to check its performance and find ways to improve it.

A data fusion strategy with the other sensors of the robot may allow us to define a reliability measure of the current segmentation results to automatically adjust the parameters of the algorithm. This multi-sensor approach, together with other segmentation methods that we plan to add, increases the architecture complexity enough to justify a multi-agent system implementation.

References

[1] S. Thrun, "*Robotic mapping: A survey*", in Exploring Artificial Intelligence in the New Millenium, G. Lakemeyer and B. Nebel, Eds. Morgan Kaufmann, 2002, to appear.

[2] F. Lu and E. Milios. "*Globally Consistent Range Scan Alignment for Environment Mapping",* In Autonomous Robots, vol. 4, pages 333--349, 1997.

[3] L. Vincent and P. Soille, *"Watersheds in digital spaces: an efficient algorithm based on immersion simulations"*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 13, pp. 583--598, June 1991.

[4] T. Pavlidis, *"Algorithms for Graphics and Image Processing",* Computer Science Press, Rockville, MD.

[5] J.C. Tilton, *"Image Segmentation by Iterative Parallel Region Growing and Spliting"*, 12th Canadian Symposium on Remote Sensing. Vol. 4, pp 2420-2423, July 1989.

[6] P. An, C. Lu and Z. Zhang, *"Object Segmentation Using Stereo Images",* International Conference on Communications, Circuits and Systems, vol. 1, pp 534-538, June 2004.

[7] E. Izquierdo, *"Disparity/Segmentation Analysis: Matching with an Adaptive Window and Depth-Driven Segmentation"*, IEEE Transactions on Circuits and Systems for Video Technology, vol 9, No. 4, June 1999.

[8] R. Venkatesh Babu, K. R. Ramakrishnan and S. H. Srinivasan*, "Video Object Segmentation: A Compressed Domain Approach"*, IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, No. 4, April 2004.

[9] C. Y. Vicent and T. Tjahjadi, *"Obstacle detection by direct estimation of multiple motion and scene structure from a moving stereo rig*",* Proceedings IEEE International Conference SMC, Washington, pp. 2326-2331, Oct. 2003.

[10] R. Yang and M. Pollefeys, *"Multi-Resolution Real-Time Stereo on Commodity Graphics Hardware"*, Proceeding on Computer Vision and Pattern Recognition, vol. 1, pp. 211-217. June 2003.

[11] H. Hirschmüller, *"Improvements in Real-Time Correlation-Based Stereo Vision",* Proceeding of the IEEE Workshop on Stereo and Multi-Baseline Vision, pp. 141-148. December 2001.

[12] A. Fusiello, E. Trucco and A. Verri, *"A Compact Algorithm for Rectification of Stereo Pairs"*. Springer Machine Vision and Applications, vol. 12, pp. 16-22. July 2000.

[13] C. Zhang. *"A survey on stereo vision for mobile robots"*. Technical report, Dept. of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh.

[14] D. Scharstein and R. Szeliski, *"A taxonomy and evaluation of dense two-frame stereo correspondence algorithms",* International Journal of Computer Vision 22, pp. 7-47, April-June 2002.

[15] O. Faugeras et alt. Real time correlation-based stereo: algorithm, implementations and application, INRIA Reseach Report 2013.

[16] J.L Mundy and A. Zisserman (editors). "*Geometrical Invariance in Computer Vision*" MIT Press, 1992.