Robot Vision for Autonomous Object Learning and Tracking

Alberto Sanfeliu

Institut de Robòtica i Informàtica Industrial
Universitat Politècnica de Catalunya (UPC)
sanfeliu@iri.upc.es
http://www-iri.upc.es/groups/lrobots

Abstract. In this paper we present a summary of some of the research that we are developing in the Institute of Robotics of the CSIC-UPC, in the field of Learning and Robot Vision for autonomous mobile robots. We describe the problems that we have found and some solutions that have been applied in two issues: tracking objects and learning and recognition of 3D objects in robotic environments. We will explain some of the results accomplished.

1 Introduction

Computer vision in autonomous mobile robotics is a very well known topic that is being treated by many research groups [1]. However, the use of perception techniques to automatically learn and recognize the environment and the objects located on it is probably not so well known. One part of our research has concentrated in the development of techniques to capture and process the information that surrounds a robot, taking into account that this information can be captured by diverse perception sensors (colour video cameras, stereo vision, laser telemeter, ultrasonic sensors, etc.) and the sensors related to robot movement (odometers).

We have focused our research in the development of "robust" techniques that must be as much as possible, "invariant" to illumination, colour, surface reflectance, sensor uncertainty, dead reckoning and dynamic environments. However, this wish is not always possible. We also orient our research to develop techniques to learn the perceptive world, in order to create a data base that can be used later on, by robots.

In this paper we describe our research in two topics in the field: adaptive learning and tracking of moving objects; and learning and recognition of 3D objects of the environment. Although these topics lead to different techniques and methodologies, they share the same perception information formats, colour images and depth maps.

However, we also use other kind of perception information formats which are captured by means of stereo vision, laser telemeter, ultrasonic and odometer sensors. The diverse information captured by these sensors is combined to obtain redundancy in order to improve the robustness of the techniques.

A. Sanfeliu and J. Ruiz-Shulcloper (Eds.): CIARP 2003, LNCS 2905, pp. 17–28, 2003.

Before explaining the methods, we will start describing the typical “problems” that we find in the perception of a dynamic environment where the robot or the objects are moving.

2 Common Problems on the Acquisition of Perception Information Based on Colour Images and Depth Maps

Colour represents a visual feature commonly used for object detection and tracking systems, especially in the field of human-computer interaction. When the environment is relatively simple, with controlled lighting conditions and an uncluttered background, colour can be considered a robust cue. The problem appears when we are dealing with scenes with varying illumination conditions and varying camera position and confusing background.

The colour of an object surface can be modified by several circumstances, which limits the applicability of the use of colour images in robot vision. The following issues modify the colour perception of an object surface:

- the type of the illumination source, the illumination orientation, the number and distribution of the sources of illumination,
- the surface reflectance and the surface orientation,
- the texture of the surface,
- and the shadows produced by other objects or by the own concavities of the object.

Some of these problems can be diminished in static scenes, by controlling the illumination (for example, for indoor robot environments, the type and position of the illumination), the object surfaces (for example, by choosing objects with Labertian surfaces) or the type of objects (for example, by using convex objects).

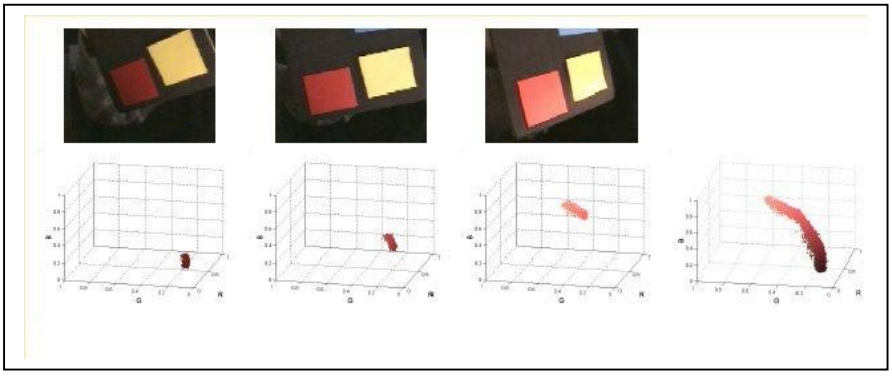

Fig. 1. Typical reflectance problems of a colour (red) planar surface: (a) a sequence of a red planar surface; (b) RGB map of the colour distribution of the sequence of the planar surface.

However, although we can impose some of these constraints in indoor environments, still many of the aforementioned drawbacks persist, due to the relative position and orientation of the robot sensors, the illumination devices and the scene objects. This relative position not only involves the passive sensor (colour camera), but also the illumination sources, for example, the robot can interfere with the illumination of an object surface by means of its shadow cast or a new "virtual" illumination source appears due to the reflection of another surface. A typical example of the last case is the reflectance of the "ground".

Other typical problems are due to the camera sensor, for example, the optical aberration and geometrical deformation, the separation of the channel colour bands, the colour sensibility, the sensor sensibility to the illumination, the problems associated with the shutter speed or the resolution of the camera.

Fig. 2. Some problems with the reflectance of the ground

With respect to the capture of depth information, we have also other drawbacks. In the case of a laser telemeter, the sensor drawbacks are due to the features of laser source, the resolution or the speed of depth acquisition and processing, or the problems related to partial surface occlusion. If the depth sensors (for example, stereo vision or laser telemeter) are in the mobile robot, then other problems come around. For example, the relative position and orientation of the sensors with respect to the scene, because of the "skew" of the elevation, pitch or roll of cameras and laser telemeter with respect to ground.

Additionally to the abovementioned problems, we always find that in robot perception, the uncertainty is an important issue that must be taken into account when discerning from sensory data the objects and the limits of the robot environment. This perception uncertainty must be incorporated in the models for robot navigation, object tracking, object recognition and landmark identification.

3 Developing "Robust Techniques" for Object and Face Tracking

Object and face tracking are two typical problems in robot vision which have been studied using different types of sensors and techniques [2, 3]. One of the most used sensors is the colour video camera, since it provides enough information to follow an object and avoid uncertainties. However, in a real unconstrained environment, the varying illumination conditions, camera position and background create important problems to the robot tracker. Different approaches have been presented to solve these problems in robot tracking, but still this is an open problem from the point of view of robustness.

The important challenge in colour tracking is the ability to accommodate to the variations of the illumination and the environment, that is, the tracker must modify its parameters depending on the circumstances. However, the use of independent adaptive techniques, many times, is not enough to cope with the problem, since the adaptation only takes into account one of the potential variations, for example the colour reflectance, however the variations are usually multivariate. For this reason, we have studied solutions that combine different techniques to take into account the multivariable effect.

One of our first approaches combines information of colour changes and depth for face tracking in real time [4]. The purpose is to follow a face or an object that has colour and depth continuity avoiding the loss of them due to the presence of similar colour in the background. The technique fuses colour adaptation and stereo vision, in such a way, that the tracked objects only is analysed in a surface with similar depth information. The technique uses an ellipse to model the face of a person similar to the work of Birchfield [5] and adaptive colour models, for example [21]. The face is adapted by means of intensity gradients and colour histograms, and the stereo vision information dynamically adapts the size of the tracked elliptical face. The system uses the Kalman filter to predict the new position of the cameras and robot, and it runs at 30 Hz, that is, in real time.

A second approach, [6] tries to solve two important problems in object tracking: the change of the colour and the confusing background. As it was mentioned before, the colour of an object surface changes with the orientation of the surface (in principle only the intensity, but due to the illumination conditions and surface reflectance, the colour can also change). Moreover, if the background is confusing, then the tracking of an object surface becomes very difficult. In order to solve these two problems, we propose a solution based on fusing colour adaptation with shape adaptation. We have developed a method that, by using the CONDENSATION technique [7], combines the use of colour histograms adaptation with snake shape adaptation [8]. The algorithm formulates multiple hypotheses about the estimate of the colour distribution in the RGB space, and validates them taking into account the contour shape of the object. This combination produces a very robust technique whose results can be seen in Fig. 4. The technique is described in detail in [6].

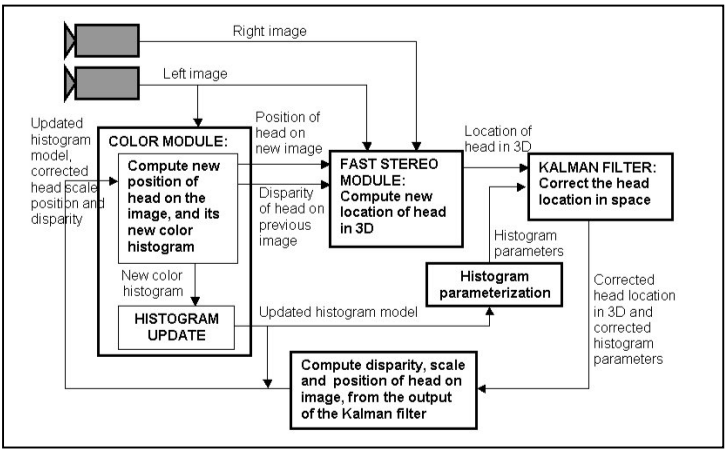

Fig. 3. The tracking vision system

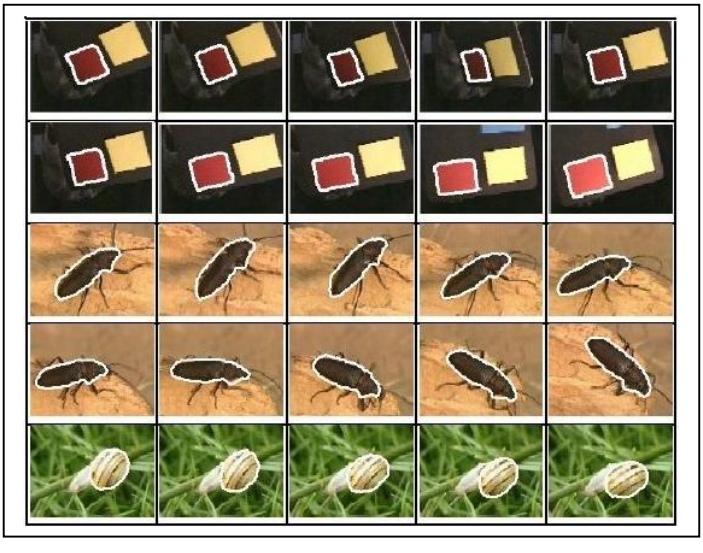

Fig. 4. Four experiments: (1) tracking of circles that change the colour; (2) tracking an object surface with different orientations and illumination; (3) tracking an insect in real environment; (4) tracking a snail in real environment

4 Learning and Identifying of Objects in Mobile Robotic Environments

The process of learning and identifying new 3D objects in robot environments has been treated using different methodologies, for example [9][20], however these techniques only work for very constrained environments. Unfortunately, many of the

proposed methods fail in real unstructured environments, due to problems of illumination, shadows, object and camera position, or confusing background.

In order to overcome some of these problems we are designing our methods taken into account the following criteria:

- the perception features must be, as much as possible, robust and relative invariant to changes of the environment,
- the representation models must be flexible and must include the statistical variations of the structure and of the perception features that are intrinsic in the learning process,
- the recognition, or matching, process must be robust against local variations and have to take into account the problems derived of partial occlusion.
- in the recognition process, the matching must be guided to reduce the number of potential model candidates.

The first criteria is one of the most difficult to solve, since the perception features depend too much of uncontrolled environment conditions. For these reason we have selected as basic perception features, the surface colour and surface shape. The first one can be obtained from colour images and the second one from depth sensors (for example, stereo vision and laser telemeter). The invariance of surface colour is a difficult task, but we are diminishing its influence by using colour constancy methods and statistical information of the feature variations. However, colour constancy algorithms are not yet given us the results that we expect, although our new developments are promising [22]. In the other hand, the surface shape obtained from the depth sensors is a robust feature.

One of the preliminary works to obtain robust results was the fusion of colour segmentation and depth, to improve the segmentation results. The method [23] processes independently colour segmentation and depth map, and then combines both outcomes. The idea of the method is to balance the over-segmentation and under-segmentation, by joining or splitting the singular areas. Fig. 5 shows the results of this method in a colour scene.

Fig. 5. Fusion of colour segmentation and depth map to segment a colour scene

In the rest of this section, we will describe the solutions adopted for representation models, the recognition and the learning processes. The basic representation models that we are using are structural representations, chain of symbols and graphs. In the first case we use cocircuits (of the matroid theory) and in the last case, we use random graphs which combine structural information with statistical information of the attributes of the nodes and arcs. In this way, we have a representation model that can be learned directly from the colour images taken into account the potential variations of the perception features.

Our research group has developed several methods to learn and recognise 3D objects described by multiple views in a scene. These methods have been oriented in two directions: a first one, whose goal is to reduce the number of candidates in object recognition by an indexing technique in 3D object hypothesis generation from single views; and a second one, whose goal is to identify the input object with respect to the model candidates by looking for the minimum measure distance between the object and the model candidates. The first direction allows the reduction of the number of potential model candidates to a few ones, which can be done very fast. The second direction allows to identify the best candidate.

4.1 Indexing Views of 3D Objects

In the first group of techniques, the idea is to represent a 3D object view by means of topological properties of the regions of the segmented image and then to create a table with each of the topological representations of the object. Then the identification process is based on indexing the input representation of one scene view to the table of the topological representations of the 3D object views.

A topological representation is created using the oriented matroid theory by means of encoding incidence relations and relative position of the elements of the segmented image, and by giving local and global topological information about their spatial distribution. The result is a set of cocircuits [10] of sign combinations that relate segmented regions with respect to the convex hull of two selected regions of the scene. The details of this process are explained in [11, 12]. The set of cocircuits obtained is projective invariant, which is an important feature for the representation of the model objects. Fig. 6 shows the segmentation and process indexing of one object and Table 1 shows the resulting indexes of the object.

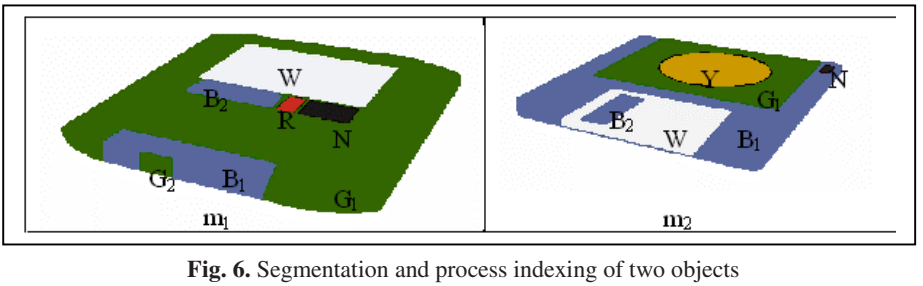

Fig. 6. Segmentation and process indexing of two objects

The result of the process indexing looks as follows:

Table 1. Index result of the process indexing of the images of Fig. 6. The first column is the baseline area from where the segmented regions are related. 0 means the region is inside the baseline area; - the region is one the left side; + the region is on the right side; and * means the region does not exist in the segmented image.

	W	**R**	**Y**	**G_1**	**G_2**	**B_1**	**B_2**	**N**	**Object**
WR	0	0	*	0	0	0	-	+	m_1
WY	0	*	0	0	*	0	0	-	m_2
WG_1	0	*	*	0	*	*	*	*	m_1
WG_1	0	*	0	0	*	0	0	0	m_2
WG_2	0	0	*	0	0	+	0	0	m_1
WB_1	0	0	*	0	0	0	0	0	m_1
WB_1	0	0	*	+	+	+	0	+	m_2
WB_2	0	0	*	+	+	+	0	+	m_1
WN	0	0	*	-	-	-	-	0	m_1
WN	0	*	+	+	*	0	0	0	m_2
RG_1	*	0	*	0	*	*	*	*	m_1
…	…	…	…	…	…	…	…	…	
B_2N	+	0	*	-	-	-	0	0	m_1
B_2N	-	*	+	+	*	+	0	0	m_2

4.2 Learning and Recognising 3D Objects Represented by Multiple Views

In the second group, the idea is to represent 3D object views by means of graphs and then to obtain the model as the synthesis from the graphs that represent the views of a 3D object. Once the model has been learned, the recognition process is based on applying a distance measure among the input graph (the graph that encodes the 3D view of a scene object) and the object models. The input graph is assigned to the model graph with the minimum distance measure value. Fig. 7 shows the process of learning (synthesis of the object graph views) and recognition.

Object views are often represented by graphs, and one of the most robust representations is based on attributed graphs. When a synthesis of these attributed graphs is required to learn a complete object through its views, then a good model representation are the Random Graphs. The generalization of these graphs is denominated General Random Graphs (GRG) which has theoretically, great representation power, but they need a lot of space to keep up with the associated data. We have defined several simplifications to the GRG to reduce the space and also to diminish the time matching complexity to compare among graphs. Wong and You [13] proposed the First-Order Random Graphs (FORGS) with strong simplifications of the GRG, specifically they introduce three assumptions about the probabilistic independence between vertices and arcs which restrict too much the applicability of these graphs to object recognition. Later, our group introduced a new class of graphs called Function-Described Graphs (FDG) [14][15] to overcome some of the problems of the FORG. The FDG also considers some independence assumptions, but some useful 2° order functions are included to constrain the generalisation of the structure.

Specifically an FDG includes the antagonism, occurrence and existence relations which apply to pairs of vertices and arcs. Finally, we have expanded this representation, [17][18] by means of Second-Order Random Graphs (SORG), which keep more structural and semantic information than FORGs and FDGs. These last types of representation have led to the development of synthesis techniques for model object generation (by means of 3D object views) and graph matching techniques for graph identification.

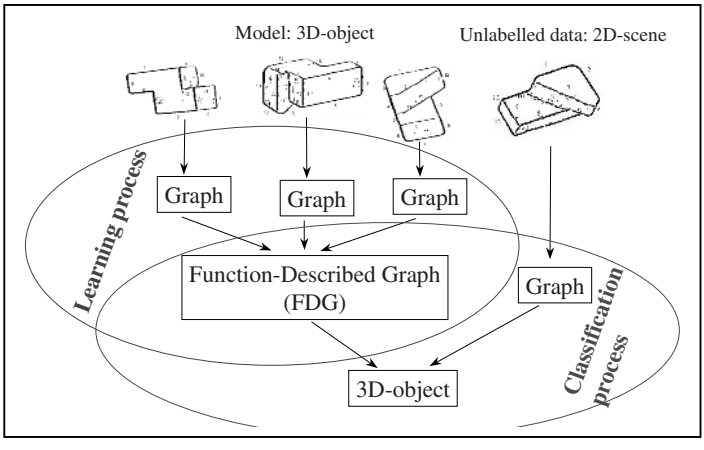

Fig. 7. Learning and classification processes in the classifiers that use only one structural representation per model

We show in this article, one example of unsupervised learning and recognition of 3D objects represented by multiple views. The set of objects was extracted from the database COIL-100 from Columbia University. We did the study with 100 isolated objects, where each one is represented by 72 views (one view each 5 degrees). The test set was composed by 36 views per object (taken at the angles 0, 10, 20 and so on), whereas the reference set was composed by the 36 remaining views (taken at the angles 5, 15, 25 and so on).

The learning process was as follows: (1) perform colour segmentation in each individual object view image; (2) create an adjacency graph for each one of the segmented regions of each object view; (3) transform the adjacency graph in an attributed graph (AG) using the hue feature as the attribute for each node graph; (4) synthesize a group of 35 object views in a FORG, FDG and SORG using the algorithms described in [16][19] (we use groupings of varying number of graphs to represent an object in order to evaluate the results, concretely we used 3, 4, 6 and 9 random graphs for each 3D object). The recognition process follows a similar procedure, but instead of synthesizing the graphs a measure distance between them was applied to evaluate to which 3D object the input graph belonged.

Fig. 8 shows 20 objects at angle 100 and their segmented images with the adjacency graphs. FORGs, FDGs and SORGs were synthesised automatically using the AGs in

the reference set that represent the same object. The method of incremental synthesis, in which the FDGs are updated while new AGs are sequentially presented, was applied. We made 6 different experiments in which the number of random graphs, FORGs, FDGs and SORGs, that represents each 3D-object varied. If the 3D-object was represented by only one random graph, the 36 AGs from the reference set that represent the 3D-object were used to synthesise the random graph. If it was represented by 2 random graphs, the 18 first and consecutive AGs from the reference set were used to synthesise one of the random graphs and the other 18 AGs were used to synthesise the other random graph. A similar method was used for the other experiments with 3, 4, 6 and 9 random graph per 3D-object. Note that if 4 random graphs are used, then each random graph represents 90 degrees of the 3D object.

The best result appears when the SORG and FDG representations were used, although the best is the SORG representation. Fig. 9 shows the ratio of recognition success of the 100 objects using different object representation and distance measures. This figure also shows the result of describing individually each object view by means of an AG and then comparing each input AG against the rest of the prototype AG.

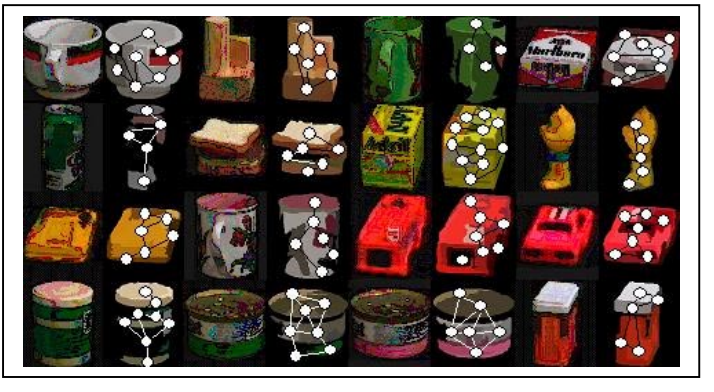

Fig. 8. Some objects at angle 100 and the segmented images with the AGs

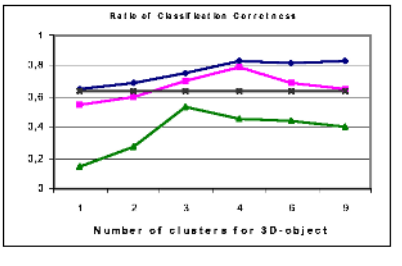

Fig. 9. Ratio of recognition correctness of the 100 objects using SORG, FDG, FORG and AG-AG SORG: ; FDG: ; FORG: ; AG-AG:

5 Conclusions

Robot vision methods require close attention to two important issues. First the real time issue: the methods must have adaptable mechanisms to overcome the variance in the sensing of the basic perception features and they must be robust. Another desirable feature in robot vision is that the objects, map, motion and control models must be learned on line. Not in only in one path, but in successive robot motions. In this article we have presented some of the methods, in tracking and object learning that we are developing following these ideas. We have also applied the same ideas for map building.

References

[1] G.N. DeSouza and A.C. Kak: Vision for mobile robot navigation, a survey. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 24, No. 2, Feb. 2002.

[2] B. Thomas, Moeslund and E. Granum: A Survey of Computer Vision-Based Human Motion Capture. Computer Vision and Image Understanding: CVIU, vol.81(3), pp.231–268, 2001.

[3] D.A. Forsyth and J. Ponce: Computer Vision: A Modern Approach- Tracking (Chapt. 19). Prentice Hall. 1st edition ,2002.

[4] F. Moreno, A. Tarrida, Juan Andrade-Cetto and A. Sanfeliu: 3D real-time head tracking fusing color histograms and stereovision. Proceedings of the 16th International Conference on Pattern Recognition, Quebec, August 12–15.

[5] S. Birchfield: Elliptical head tracking using intensity gradient and colour histograms. Intl. Conference in Computer Vision and Pattern Recognition, pp. 232–237, 1998.

[6] F. Moreno, J. Andrade-Cetto and A. Sanfeliu: Fusion of color and shape for object tracking under varying illumination. Pattern Recognition and Image Analysis, Lecture Notes in Computer Science 2652, Springer-Verlag, pp. 580–588, 2003.

[7] M. Isard and A. Blake: Condensation-conditional density propagation for visual tracking. Int. J. Computer Vision, 28(1), pp.5–28, 1998.

[8] M. Kass, A. Witkin and D. Terzopoulos: Snakes: active contour models. Int. J. Computer Vision. 1(4), pp. 228–233, 1987.

[9] C. Chen and A. Kak: Robot vision system for recognizing objects in low-order polynomial time. IEEE Trans. On System, Man and Cybernetics, 18(6), pp. 1535–1536, Nov. 1989.

[10] A. Björner, M.L. Vergnas, B. Sturmfels, N. White, G.M.: Oriented matroids. Volume 43 of Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1993.

[11] E. Staffetti, A. Grau, F. Serratosa, A. Sanfeliu: Oriented matroids for 3-D object views representation and indexing. Pattern Recognition and Image Analysis, First Iberian Conference; IbPRIA 2003. Lecture Notes in Computer Science, LNCS 2652, Springer Verlag, Mallorca, 4–6 June. 2003

[12] E. Staffetti, A. Grau, F. Serratosa and A. Sanfeliu: Shape representation and indexing based on region connection calculus and oriented matroid theory. In Discrete Geometry for Computer Imagery, Lecture Notes in Computer Science, Springer-Verlag, Napoles, 2003.

[13] A.K.C. Wong and M. You: Entropy and distance of random graphs with application to structural pattern recognition. IEEE Trans. On Pattern Analysis and Machine Intelligence, Vol. 7, pp. 599–609, 1985.

[14] A. Sanfeliu, R. Alquezar, J. Andrade, J. Climent, F. Serratosa and J. Verges: Graph-based representations and techniques for image processing and image analysis. Pattern Recognition, Vol. 35, pp. 639–650, 2002.

[15] F. Serratosa, R. Alquezar and A. Sanfeliu: Function-described graphs for modelling objects represented by sets of attributed graphs. Pattern Recognition, Vol. 36, pp. 781–798, 2003.

[16] F. Serratosa, R. Alquezar and A. Sanfeliu: Synthesis of function-described graphs and clustering of attributed graphs. Int. J. Pattern Recognition, 16(6), pp. 621–655, 2002.

[17] F. Serratosa, R. Alquezar and A. Sanfeliu: Estimating the joint probability distribution of random graphs vértices and arc by means of 2º order random graphs. SSPR'2002, Lecture Notes in Computer Science, LNCS 2396, pp. 252–262, 2002.

[18] A. Sanfeliu and F. Serratosa: Learning and recognising 3D models represented by múltiple views by means of methods based on random graphs. IEEE Int. Conference on Image Processing, Barcelona 2003 (ICIP 2003), Sept. 2003.

[19] A. Sanfeliu, F. Serratosa and R. Alquezar: Second-order graphs for modelling sets of attributed graphs and their application to object learning and recognition. Int. J. Pattern Recognition (in press).

[20] H. Murase y S.K. Nayar. Visual learning and recognition of 3D objects from appearance. International Journal of Comp. Vision, 14(1): 5–24, 1995.

[21] J. Vergés-Lalhi, A. Tarrida and A. Sanfeliu: New approaches for colour histogram adaptation in face tracking tasks. Proceedings of the 16th International Conference on Pattern Recognition, Quebec, August 12–15, 2002.

[22] J. Vergés-LLahí, A. Sanfeliu: Colour constancy algorithm based on object function minimization. Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA), Puerto Andrax, Mallorca, 4–6 June. Lecture Notes in Computer Science. 2003.

[23] J. Andrade and A. Sanfeliu: Integration of perceptual grouping and depth. Proceedings of the 15th International Conference on Pattern Recognition, Barcelona, September 3–7, IEEE Computer Society, Vol. 1, pp. 295–298, 2000.