

# Performance aspects in a complex cognitive architecture for artificial vision

S. Gaglio

*Dipartimento di Ingegneria Elettrica, Università di Palermo,  
Viale delle Scienze, 90128 Palermo, Italy*

## Abstract

A cognitive architecture for an artificial vision system, aimed for an autonomous intelligent system, is presented with particular attention to performance aspects. The architecture, based on the active vision paradigm, is cognitive in the sense that several cognitive hypotheses have been postulated as guidelines for its design in order to achieve good performance.

## 1 Introduction

An artificial vision system for an autonomous agent must be able to build a rich and concrete internal representation of the external environment. Such an internal representation should allow the system to effectively draw inferences, make decisions and, in general, to perform reasoning processes on its own tasks.

In classical symbolic reasoning systems, the symbols are related to abstract entities according to model theoretic semantics. This turns out to be completely unsatisfactory for an autonomous agent, that needs to find the meaning for its symbols within its internal representation and in its interaction with the external world, overcoming in some way the well known symbol grounding problem.

This paper discusses various aspects of a cognitive architecture for an artificial vision system, which is going to be implemented at the Department of Electrical Engineering of the University of Palermo, Italy. In the architecture an effective internal representation of the environment is built by means of processes defined over a suitable intermediate level that acts as an intermediary between the sensory data and the symbolic level.

From a computational point of view, the architecture is characterized by the coexistence of different heavy processing paradigms: symbolic processing, logic processing, neural processing, vector processing. Performance aspects are therefore addressed in different ways. First of all in the vision process, as pointed out in Wechsler<sup>1</sup>, visual information is subject to bandwidth reduction and coding processes, as it goes from the low image-centered level to higher interpretation levels. At the low level, however, a lot of intrinsic parallelism is present, and it can be exploited through the use of

## 308 High-Performance Computing in Engineering

suitable array processors. At the highest levels in which information is of more symbolic nature, traditional logic inference procedure are known to be wasteful in time and resources. Convenient attentive mechanisms, as present in biological systems, allow to improve performance by avoiding the generations of useless assertions and the need of expensive search procedures. Those mechanisms are implemented in terms of neural networks, that in turn could be fast enough, if programmed on array processors or specialized hardware.

### 2 Coding of visual information and cognitive assumptions

For an efficient coding of visual information, aimed to a drastic bandwidth reduction, three representation levels are hypothesized as the basis of the architectural design: the subsymbolic level, in which the information is strictly related to the sensory data; the linguistic level, in which information is expressed by a symbolic language; and an intermediate, "prelinguistic" conceptual level. At this level the information is characterised in terms of a metric space defined by a certain number of "cognitive" dimensions, independent from any specific language (Gärdenfors<sup>2</sup>). The aim of this level is to generate the very internal representation of the agent's external environment and to provide a precise interpretation for the linguistic level.

A correlated cognitive aspect of the proposed architecture is the role of attentive processes in the link between the linguistic and the conceptual level. In fact, a finite agent with bounded resources cannot carry out a one shot, exhaustive, and uniform analysis of a perceived scene within reasonable time constraints. Furthermore, some aspects of a scene are more relevant than others, and it would be irrational to waste time and computational resources to detect true but useless details. These problems can be faced by hypothesising a sequential attentive mechanism, that scans the internal representation of the scene.

The focus of attention is hypothesised to be driven on the basis of the knowledge, the hypotheses, the purposes and the expectations of the system, in order to detect the relevant aspects in the perceived scene. Hence, it is a task of the higher level components to use the information acquired through the perceptual system to create expectations or to form contexts in which hypotheses may be verified and, if it is necessary, adjusted.

### 3 The design of the architecture

The cognitive assumptions introduced in the previous section are the guidelines for the design and implementation of the proposed architecture for artificial vision. The current implementation is limited to the analysis of static scenes (Chella et al.<sup>3</sup>)

Fig. 1 shows the overall architecture in which the previously described three levels of representations are evidentiatiated.

The block A is the kernel of the subsymbolic level: it receives one or more input pictorial digitised images acquired by a camera and, by means of shape from shading and/or stereo vision algorithms, gives as an output the  $2\frac{1}{2}D$  depth images. The  $2\frac{1}{2}D$  depth images are input to the block B, which builds a scene description in terms of a combination of 3D geometric primitives.

## High-Performance Computing in Engineering 309

The block C implements the associative mapping between the conceptual level and the symbolic level; the aim of this block is to recognize objects and situations. The input to the block C is a vector configuration in the conceptual space, its output is sent to the linguistic level to produce a sentential description of the scene.

The symbolic knowledge base represents the kernel of the linguistic level. The aim of this block is twofold: it describes in a high-level language the perceived scene by interpreting the input coming from the block C, and it generates, by means of its inferences capabilities, the linguistic "expectations" that drive the focus of attention mechanism.

The block D is the block responsible for the linguistic modality of the focus of attention mechanism. It receives as input the instances of concepts from the knowledge base and it suitably drives the focus of attention, in order to seek for the corresponding objects and situations in the acquired scene.

The block E is the block responsible for the associative modality of the focus of attention. Its operation is similar to block D, but it drives the focus of attention looking for the objects in the scene which can be freely associated to the input instances.

The reactive modality of the focus of attention is implemented as an internal mechanism of block D: when the block does not receive any expectations as input, it generates some generic expectations in order to start the operations of the system.

### 4 Geometric primitives

According to the classic approach suggested in Marr and Nishihara<sup>4</sup> the main step towards the recognition of objects present in a scene requires to derive a 3-D model description from a scene image, through the identification, by means of information extracted from the images, of the *primitives* representing the low level "parts" of the description.

The chosen representation is based on models presented in the areas of Computer Graphics, CAD/CAM, known as *CSG* (Constructive Solid Geometry), which describe complicated solids as various "additions" and "subtractions" of simpler solids by means of modified versions of Boolean operators - union, difference and intersection (regularised operators as introduced in Requicha<sup>5</sup>).

A CSG model is also a very efficient and robust coding schema for representing 3-D information: The huge amount of data necessary to tessellate a portion of 3-D space is reduced to a manageable number of parameters describing the geometric primitives and their spatial relationships.

In particular, the approach proposed by Pentland<sup>6</sup> has been followed for the description of both natural and man-made forms, where the representation primitives, i.e. the building blocks of complicated objects, are modelled by using *superquadrics*. Superquadrics are mathematical shapes based on the parametric form of quadric surfaces, in which each trigonometric function is raised to one of two real exponents (Barr<sup>7</sup>). Thus a superquadric is completely defined by its centre, its orientation, its "sides" (length, width and breadth) and the above-mentioned real exponents,  $\epsilon_1$  and  $\epsilon_2$ , called form parameters that affect the superquadric's surface shape.

## 310 High-Performance Computing in Engineering

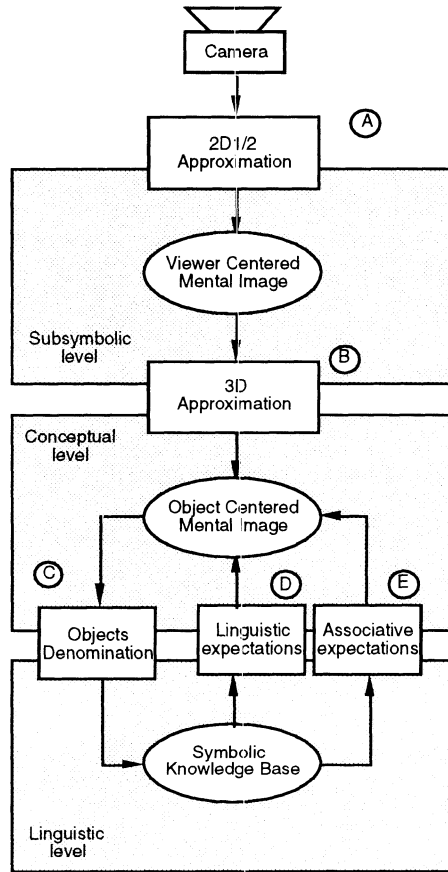


Figure 1 The general architecture of the artificial vision system

For example, superellipsoids in canonical positions are described by the following equations ( $C_\eta = \cos \eta$ ,  $S_\omega = \sin \omega$ ):

- *Position* vector of surface:

$$x(\eta, \omega) = \begin{bmatrix} a_1 C_\eta^\varepsilon C_\omega^\varepsilon \\ a_2 C_\eta^\varepsilon S_\omega^\varepsilon \\ a_3 S_\eta^\varepsilon \end{bmatrix} \quad (1)$$

- *Inside-outside* function:

$$F(x, y, z) = \left( \left( \frac{x}{a_1} \right)^{2/\varepsilon_2} + \left( \frac{y}{a_2} \right)^{2/\varepsilon_2} \right)^{\varepsilon_2/\varepsilon_1} + \left( \frac{z}{a_3} \right)^{2/\varepsilon_1} \quad (2)$$

where  $-\pi/2 \leq \eta \leq \pi/2$ ,  $-\pi \leq \omega < \pi$  and  $a_1$ ,  $a_2$ ,  $a_3$  are respectively the length, the width and the breadth (the 'sides') of the superquadric.

The surfaces so described are parameterized in latitude  $\eta$  and longitude  $\omega$ , and their shape is controlled by the parameters  $\varepsilon_1$  and  $\varepsilon_2$ , in latitude and longitude respectively. Modifying shape factors causes a local surface deformation that can result in squared as well as pinched or rounded shapes (fig. 2). In this way, the simple modification of a few parameters allows for a number of different shapes. Global deformations can also be defined, to take into account stretched, twisted and bent shapes (Ardizzone et al.<sup>8</sup>, Solina and Bajcsy<sup>9</sup>). Thus Boolean combinations and deformations of superquadrics are able to represent with surprising realism scenes and objects of the real world, that can therefore be described, in an extremely compact way, by the list of their parameters.

## 5 From images to shape

Block A of the architecture of fig. 1 is responsible for the extraction of the 3-D information needed to approximate the shape of objects present in the scene through superquadrics. This may be achieved if depth values of the entire image, i.e. a sort of *dense* depth map, are available.

Shape-from-shading and shape-from-stereo are among most popular approaches (Horn<sup>10</sup>) to perform this task, even if their performance is normally acceptable only in presence of limitative hypotheses. The most important of these limitations are on the surface form and characteristics, and on the environmental control of illumination sources. Some of these limitations can be at present accepted in our case. Shapes may be supposed convex and of surface sufficiently smooth and homogenous, and the environment in which the images are acquired is sufficiently controlled.

The recovery of superquadrics from 2-1/2 D information is the task performed by block B of the architecture of fig. 1. A review of reconstruction methods using either superquadrics or other forms of 3-D surface representations may be found in Bolle and Vemuri<sup>11</sup>. Solina and Bajcsy handle the superquadric recovery from 3-D points available from a range imager as a least-squares minimization problem (Solina and Bajcsy<sup>9</sup>).

## 6 Performance aspects

Low level vision, as it emerges from the previous section, is mainly characterised by vector processing and iteration on huge amount of data, with a very high degree of intrinsic parallelism. Classical solutions for high performance computation are based on general purpose array processors, SIMD architectures and specialized hardware.

## 312 High-Performance Computing in Engineering

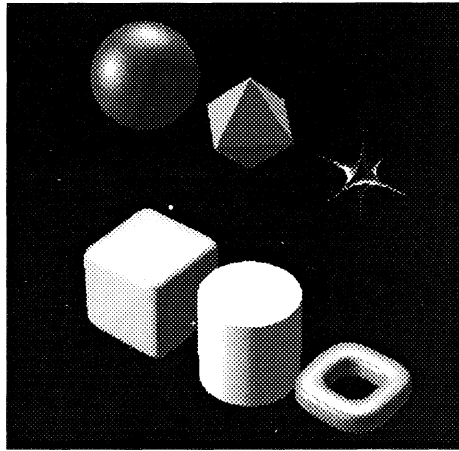


Figure 2: Examples of superquadrics

High level vision emerges from the concept of abstraction, which requires efficient coding of visual information, resulting in a drastic bandwidth reduction. Information becomes of a more symbolic kind. Unfortunately, traditional logic inference procedures introduce a new kind of complexity, which is of combinatorial type and which results in a tremendous waste of time and of other resources.

In the present approach, as discussed in the following sections, efficient computation can be achieved by means of convenient attentive mechanisms, as present in biological systems, that avoid the generation of useless logical assertions and the use of expensive search procedures. Further, these mechanisms are implemented in terms of neural networks, that in turn could be fast enough, if programmed on array processors or specialized hardware.

### 7 The conceptual space

A conceptual space, as introduced by Gärdenfors<sup>2</sup>, is a metric space consisting of a number of quality dimensions. From a formal point of view, a conceptual space is an  $n$ -dimensional space  $CS$  where  $X_i$  is the set of values of the  $i$ -th quality dimension ( $1 \leq i \leq n$ ). Examples of such dimensions can be colour, pitch, mass, spatial coordinates, and so on. The dimensions should be considered "cognitive" in that they correspond to qualities of the represented environment, without references to any linguistic descriptions.

In the case of visual perception, the dimensions of the conceptual space correspond to the parameters of a suitable system of geometric primitives, e.g. simple blocks (as Marr's generalised cylinders), geons (Biederman<sup>12</sup>) and superquadrics (Pentland<sup>6</sup>). In this perspective, the points in  $CS$  correspond to simple geometric building blocks, while complex objects or situations are represented as suitable sets of points. Accordingly, each point is related to measurements, obtained via suitable sensors, of the geometric parameters of "simple basic" objects in the external environment.

## 8 The linguistic level

The role of the linguistic level in the proposed architecture is to provide a rough and concise description of the perceived scene in terms of a high-level logical language, suitable for symbolic knowledge-based reasoning.

In order to describe the symbolic knowledge base we adopt a hybrid representation formalism, in the sense of Nebel<sup>13</sup>. According to this point of view, a hybrid formalism is constituted by two different modules: a terminological component and an assertional component. In our model, the terminological component contains the descriptions of the concepts relevant for the represented domain (e.g., types of objects and of situations to be perceived). The assertional component stores the assertions describing the specific perceived scenes. In this sense, the terminological component can be seen as analogous to a long term memory, while the assertional component to a short term memory. Moreover, terminological formalisms are well suited for our purposes, in that they are centered on conceptual descriptions. This allows a compact description of concepts, whose instances are to be recognised in the perceived scene.

## 9 Performance at the high level: the focus of attention process

A finite agent with bounded resources cannot carry out a one shot, exhaustive, and uniform analysis of a perceived scene within reasonable time constraints. This problem can be faced taking into account the fundamental role of attentive phenomena in vision. Several models of focus of attention mechanisms have been proposed in the active vision literature. The interest in this field has been summarised in Bajcsy and Campos<sup>15</sup> proposing the "active and exploratory" framework for perception.

In the proposed architecture, the conceptual level acts as a "buffer interface" between subsymbolic and linguistic processing. The information coming up from the subsymbolic level has the effect of contemporary activating an (eventually very large) set of points in CS. It is the focus of attention mechanism that imposes a sequential order in the conceptual space according to which the linguistic expressions can be given their interpretation.

In the proposed architecture the focus of attention is determined by three concurrent modalities: the reactive, the associative and the linguistic modality. The "reactive" modality is the simpler one: the grouping paths among points are determined only by the characteristics of the visual stimulus, e.g. the volumetric extension of the forms, or the aggregation density of the perceived objects.

In the associative modality, the grouping paths are determined by an associative, purely Hebbian mechanism determining the attention on the basis of free associations between concepts. Whenever two objects in the same scene are perceived, the weight of the associative connection between the corresponding concepts is increased.

In the linguistic modality, the focus of attention is driven by the symbolic information explicitly represented at the linguistic level.

Both modalities are implemented in terms of attractor neural networks acting as sequence recognizers (Kleinfeld<sup>16</sup>), and characterized by a global energy E, which is the superimposition of three energies :

## 314 High-Performance Computing in Engineering

$$E(t) = E_1(t) + \lambda E_2(t) + \varepsilon E_3(t) \quad (3)$$

In (3)  $E_1$  represents the fast dynamics for period of duration  $t$  and it models the point attractors for the single elements of the sequence;  $E_2$  represents the slow dynamics for period of duration  $t \gg t$  due to time-delayed connections and it models the expectation;  $E_3$  model the global external input to the network.

### Acknowledgements

This work has been partially supported by the Italian Consiglio Nazionale delle Ricerche "Progetto Finalizzato Robotica" and Ministero per l'Universita' e la Ricerca Scientifica e Tecnologica "MURST 40%"

### References

1. Wechsler, H. *Computational Vision*, Academic Press, 1990
2. P. Gärdenfors, P. Three Levels of Inductive Interence, Lund University Cognitive Studies 9, Tech. Rep. LUHFDA/HFKO-5006-SE, 1992.
3. Chella, A, Frixione, M. & Gaglio, S. A Cognitive Architecture for Artificial Vision, CS&AI Lab Tech. Rep, University of Palermo, 1994.
4. Marr, D., Nishihara, K. Representation and recognition of the spatial organization of three-dimensional shapes, *Proc. Roy. Soc. London B*, 1978, **200**, 269-294.
5. Requicha, A.A.G. Representations for rigid solids: theory, methods and systems, *ACM Comp. Surveys*, 1980, **12**, 4, 437-464.
6. Pentland, A.P. Perceptual Organization and the Representation of Natural Form, *Artificial Intelligence*, 1986, **28**, 293-331.
7. Barr, A.H. Superquadrics and Angle-Preserving Transformations, *IEEE Computer Graphics and Applications*, 1981, **1**, 11-23.
8. Ardizzone, E., Gaglio, S., Sorbello, F. Geometric and Conceptual Knowledge Representation within a Generative Model of Visual Perception, *Journal of Intelligent and Robotic Systems*, 1989, **2**, 381-409.
9. Solina, F., Bajcsy, R. Recovery of Parametric Models from Range Images: The Case of Superquadrics with Global Deformations, *IEEE Trans. on PAMI*, 1990, **12**, 131-147.
10. Horn, B.K.P. *Robot Vision*, The MIT Press, 1986.
11. Bolle, R.M., Vemuri, B.C. On three-dimensional surface reconstruction methods, *IEEE Trans. Patt. Anal. Mach. Intell.*, 1991, vol.13, pp. 1-13.
12. Biederman, I. Human Image Understanding: Recent Research and a Theory, *Computer Vision, Graphics and Image Processing*, 1985, **32**, 29-73.
13. Nebel, B. *Reasoning and Revision in Hybrid Representation Systems*, LNAI 422, Springer-Verlag, Berlin, 1990.
14. R. Bajcsy, R. & Campos, M. Active and Exploratory Perception, *Computer Vision, Graphics and Image Processing: Image Understanding*, 1992, 56, 1, 31-40.
15. Kleinfeld, D. Sequential State Generation by Model Neural Networks, *Proc. Nat. Acad. Sci. USA*, 1986, **83**, 9469-9473.