

# A Cognitive Architecture for Artificial Vision<sup>1</sup>

A. Chella<sup>a,2</sup>, M. Frixione<sup>b,3</sup> and S. Gaglio<sup>a,4</sup>

<sup>a</sup> *Dipartimento di Ingegneria Elettrica, Università di Palermo, Viale delle Scienze, 90128 Palermo, Italy*

<sup>b</sup> *Istituto Internazionale per gli Alti Studi Scientifici, Via G.Pellegrino 19, 84019 Vietri S.M.(Salerno), Italy*

---

## Abstract

A new *cognitive* architecture for artificial vision is proposed. The architecture, aimed at an autonomous intelligent system, is cognitive in the sense that several cognitive hypotheses have been postulated as guidelines for its design. The first one is the existence of a *conceptual* representation level between the *subsymbolic* level, that processes sensory data, and the *linguistic* level, that describes scenes by means of a high-level language. The conceptual level plays the role of the interpretation domain for the symbols at the linguistic levels. A second cognitive hypothesis concerns the active role of a *focus of attention* mechanism in the link between the conceptual and the linguistic level: the exploration process of the perceived scene is driven by linguistic and associative expectations. This link is modeled as a time-delay attractor neural network. Results are reported obtained by an experimental implementation of the architecture.

*Key words:* Perception; Active vision; Robotics; Conceptual spaces; Spatial reasoning; Geometric reasoning; Representation levels; Hybrid processing

---

## 1 Introduction

An artificial vision system for an autonomous agent must be able to build a rich internal representation of the external environment. Such internal represen-

---

<sup>1</sup> Work partially supported by Progetto Finalizzato Robotica and Progetto Coordinato SARI of the CNR (Consiglio Nazionale delle Ricerche), and by MURST 40% of the Ministero per l'Università e la Ricerca Scientifica e Tecnologica.

<sup>2</sup> Corresponding author. E-mail: chella@diepa.unipa.it

<sup>3</sup> E-mail: frix@dist.unige.it

<sup>4</sup> E-mail: gaglio@diepa.unipa.it

tation should allow the system to effectively draw inferences, make decisions, and, in general, perform reasoning processes concerning its own tasks [31,4].

In classical reasoning systems oriented to logic, the meaning of symbols is given by relating them to abstract entities according to model theoretic semantics. This turns out to be incomplete for an autonomous agent, since it needs to find the meaning for its symbols within its internal representation and in its interaction with the external world, thus overcoming the well known symbol grounding problem, as discussed in Harnad [36].

We present a cognitive architecture for an artificial vision system, in which 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. This architecture is not to be considered as a model of human vision: no hypotheses are made concerning its empirical adequacy from a psychological point of view. However, various cognitive results have been used as sources of inspiration.

According to Marr's model [47], visual perception is modeled as a process in which information and knowledge are represented and processed at different levels of abstraction, from the lowest level, directly related to features of proximal stimuli, to the highest one, where knowledge about the perceived objects is of a symbolic nature. Following Marr's seminal work, research in computer vision has exploded, becoming itself a discipline which has provided many working paradigms for object reconstruction and recognition from sensory data (see Besl and Jain [11], Chin and Dyer [24], Bindford [14] for a review).

A general implicit assumption of research in computer vision has been that the vision process ends with the 3D reconstruction of shapes by means of some suitable geometric primitives, as for instance, Marr's generalized cylinders [47]. Models of reasoning about the structure of the reconstructed objects have been proposed only for very special purpose recognition systems (see, e.g., the ACRONYM system [19,20] or the ALVEN system [63]). In the artificial intelligence community, on the other hand, there has been growing interest in spatial reasoning for planning activities of situated agents in a physical environment [1], and for man-machine interaction [44]. But research in this field has failed to result in an effective interaction with the real world environment by means of a working vision system.

The architecture proposed here aims at providing a general vision model for an autonomous agent, that fills a gap between these lines of research by means of a paradigm according to which a reconstructed geometrical scene can be described in symbolic linguistic terms. It also provides a context which can be useful for active vision tasks, as described by Bajcsy [5] and Ballard [7].

This linguistic description should, however, be considered as a first level that is nevertheless sufficient to ground successively higher symbolic reasoning activities.

The three cognitive representation levels proposed by Gärdenfors [33] are the basis of our architectural design: the *subsymbolic* level, in which the information is strictly related to sensory data; the *linguistic* level, in which information is expressed by a symbolic language; and an intermediate, prelinguistic *conceptual* level, where the information is characterized in terms of a metric space defined by a number of *cognitive* dimensions, independent of any specific language. This level aims at generating the essential representation of the agent's external environment and at providing a precise interpretation of the linguistic level.

The interpretation of the conceptual categories at the linguistic level involves some well known problems. For instance, perceptual common sense concepts hardly correspond to clear cut, classic categories which can be described in terms of necessary and sufficient conditions. Membership in perceptive categories is not an all-or-nothing affair: it is usually necessary, for example, to consider a prototype of the category. Moreover, the available information depends strictly on the data acquired through measurement processes. As a consequence, knowledge at the conceptual level is affected by measurement errors. A way of facing these problems is to model the mapping between the conceptual and the linguistic levels in terms of a connectionist device. Neural networks make it possible to avoid an exhaustive description of conceptual categories at the symbolic level: in some sense, prototypes "emerge" from the activity of an associative mechanism during a training phase based on examples. In addition, the measure of similarity between a prototype and a given object is implicit in the behavior of the network and is determined during the learning phase.

A further cognitive aspect is the role of attention processes in the link between the linguistic and the conceptual level. 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. We face these problems by a sequential attention mechanism, which suitably scans the internal representation of the scene. Also the order in which the objects in the scene are analyzed can be relevant (and, obviously, it becomes crucial in the case of the perception of dynamic scenes). Our model drives the *focus of attention* by 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 can be verified and, if necessary, adjusted. The link between the linguistic and the conceptual level is therefore bidirectional: the conceptual level defines the interpretation domain for the symbols at the linguistic level, and the linguistic level generates expectations in order to explore the conceptual level suitably. Three focus of attention modes are the basis of the proposed architecture: a *reactive* mode, in which attention is driven only by the characteristics of scene, a *linguistic* mode in which attention is driven by simple inferences at the linguistic level, and an *associative* mode in which attention is driven by free associations among concepts.

In summary, we extend further and complete the representation levels proposed by Marr, by adding a conceptual and a linguistic level, where understanding takes place. Moreover, the introduced focus of attention provides a systematic and general interaction mechanism among levels, and extends also the active vision paradigm to higher cognitive levels. As a consequence, the limitations of special purpose goal oriented vision systems like ACRONYM and ALVEN, or the more recent systems like TEA-1 [56] and BUSTER [15], are overcome by means of a framework in which general understanding of visual information is modeled in a well-founded manner, and specific goals can be easily expressed.

We are aware of the typical, “hard” and not yet solved problems encountered in real scenes at low level vision, as shadows, poor contrast, occluding objects, segmentation criteria; in Section 9 we will discuss how our architectural design is a contribution towards possible and unexplored solutions.

In the following Sections we present the architecture in a detailed manner, also providing simple experimental results aimed at illustrating the functioning of the various components. It should be noted that also with the adopted reduced experimental setup at the low level, that provides only essential information about scenes, the architecture is able to draw many inferences and to build a rich interpretation context. Specifically, Section 2 delineates the design of the architecture based on the previously exposed principles; Section 3 describes the three levels of representation, while Sections 4 and 5 respectively specify the linguistic level and its interpretation function. Section 6 examines in greater detail the focus of attention mechanism, and Section 7 characterizes the link between the conceptual and the linguistic level in terms of time-delay attractor neural networks. Finally, Section 8 describes the employed experimental setup and the obtained results, and Section 9 presents some concluding remarks and hints on future work.

## 2 The cognitive architecture

The cognitive assumptions introduced in the previous Section provide the guidelines for the design and implementation of the proposed architecture for artificial vision. The current implementation concerns the analysis of static scenes. Fig. 1 shows the overall architecture in which the previously described three levels of representation are pointed out.

Block A is the starting block of the subsymbolic level: it receives one or more input pictorial digitized images acquired by a camera and it gives as output the Marr's  $2\frac{1}{2}$ D description [47] of the input image. This contains information similar to the intrinsic images proposed by Tenenbaum, Fischler, and Barrow [59] and by Barrow and Tenenbaum[9], such as relative depth, local orientation and segmentation maps. Several algorithms and methodologies have been proposed in the computer vision literature to extract this information from the pictorial images (see Bertero, Poggio, and Torre [10], Lee [45], Aloimonos [2] for a review).

The maps extracted by block A are sent as input to block B, which builds, at the conceptual level, a scene description in terms of a combination of 3D geometric primitives. Several types of 3D primitives have been proposed to generate an object-centered description of the scene, like generalized cylinders and cones [51,20], geons [12,26], superquadrics [8,52,57] and deformed superquadrics [61,60,53]. Such primitives can be recovered by many proposed reconstruction methods, which are based mainly on the iterative minimization of suitable non linear error functions (see Bolle and Vemuri [17]).

Block C implements the mapping between the conceptual level and the symbolic level; this block aims at recognizing the objects and the situations. The input to block C is a structure at the conceptual level, its output is sent to the linguistic level to produce a sentential description of the scene. The *symbolic knowledge base* is 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 block C, and it generates, by means of its inference capabilities, the *expectations* that drive the focus of attention mechanism.

Block D is responsible for the linguistic mode 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 the corresponding objects and situations in the acquired scene. Block E is responsible for the associative mode of the focus of attention. Its operation is similar to block D, but it drives the focus of attention by looking for the objects in the scene which can be freely associated with the input instances. The reactive mode of the focus of attention is implemented as an internal mechanism of block D:

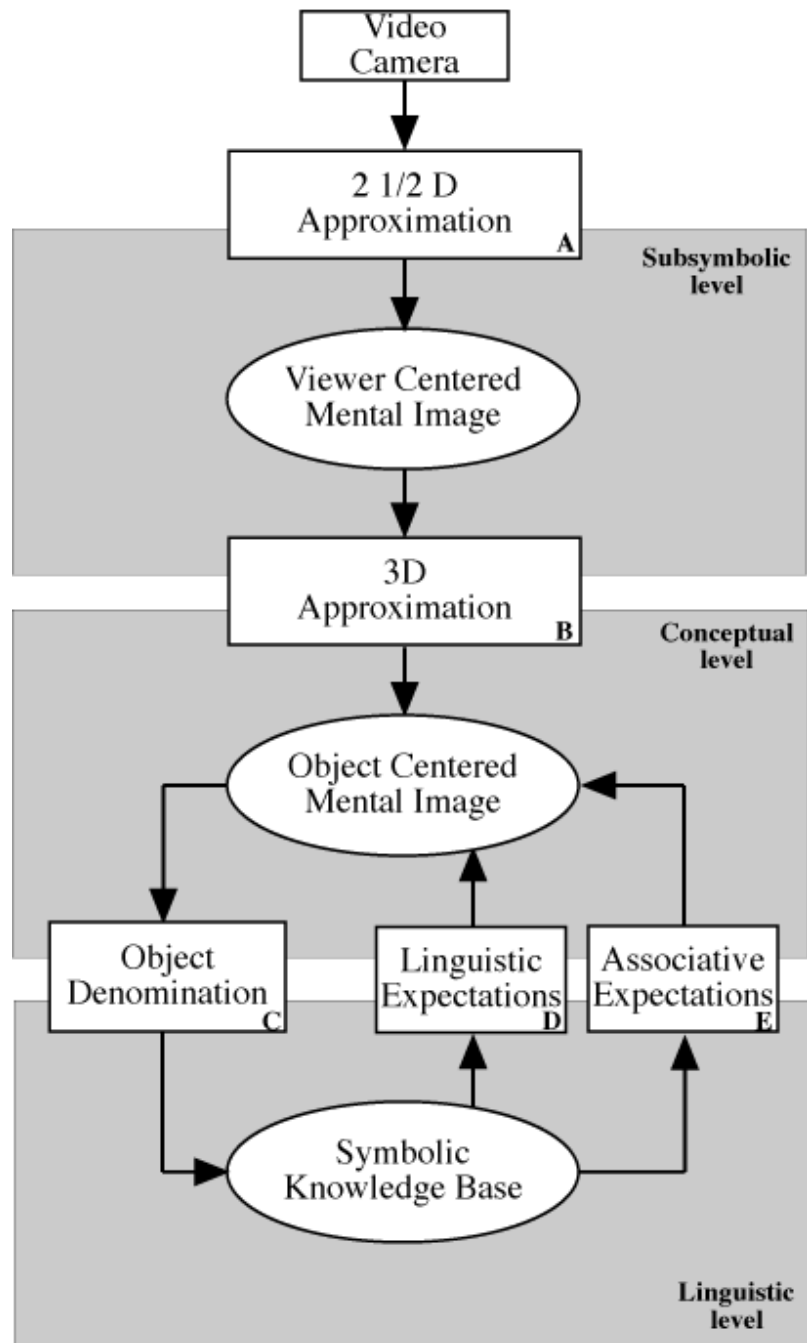


Fig. 1. The proposed architecture in which the three levels of representation are pointed out. Block A receives the input from a camera and gives as output the  $2\frac{1}{2}$ D map images. The maps are sent to block B, which builds a scene description in terms of a combination of 3D geometric primitives. Block C implements the mapping between the conceptual level and the symbolic level. Block D implements the linguistic mode of the focus of attention mechanism, while block E implements the associative mode of the focus of attention.

when the block does not receive any expectations as input, it generates some generic expectations in order to “bootstrap” the operation of the system.

### 3 The levels of representation

According to Marr’s model [47], visual perception is described as an information processing activity at different levels of abstraction. At higher levels visual information is object-centered and is related to the 3D characteristics of the scene. In Marr’s theory a superior symbolic level is limited to a hierarchically organized catalogue of 3D prototypes.

It has been introduced in the mental imagery literature (see Block [16]) the distinction between mental pictures and propositional mental representations: e.g., Kosslyn [43] distinguishes between a short term memory based on mental images, and a propositional long term memory: mental images can be generated and processed starting from the propositional long term information. It has been discussed whether mental images are *viewer centered* or *object centered* representations, that is, whether a mental image depends on the specific observation point or not. Cognitive evidence exists according to which both these kinds of representation coexist and are integrated in human memory, as described by Tarr and Pinker [58] and by Farah, Hammond, Levine, and Calvanio [29].

In the Johnson-Laird theory [39], three levels of representation are hypothesized, that, in some sense, summarize the various points of view sketched above. The “highest” level is a propositional representation, i.e., a symbolic representation similar, for example, to a semantic network. The intermediate representation is a mental model, in some respects analogous to an object centered (or spatial) mental image. The “lowest” level is a visual, viewer centered, mental image.

From a slightly different point of view, Gärdenfors [33] proposes three levels of information representation: a linguistic level, a conceptual level and a sub-symbolic level. At the linguistic level the information is described in terms of a symbolic language, e.g., a first order language; at the subsymbolic level information is characterized directly in terms of the perceptual inputs of the system. Between these two levels, a third level is hypothesized: the *conceptual* level, in which information is described in terms of a *conceptual space*. Our model is inspired by the three representation levels proposed by Gärdenfors. The theory of conceptual spaces provides a robust cognitive background for the definition of the internal representations of the agent’s external environment. Furthermore, this framework may easily be generalized to incorporate well-founded attentional mechanisms, as we will show in Section 6. Further

analogies can be found between the model proposed here and the one proposed by Marr, as well as the models that have emerged from the mental imagery debate.

In Fig.1, the three gray blocks correspond to Gärdenfors' levels of representation. The first level can also be seen as a visual, viewer centered, mental image (or, in Marr's terminology, a  $2\frac{1}{2}$ D sketch). The central level embeds within itself an object centered mental image (in Marr's terminology, a 3D model representation). The upper level consists of a propositional, linguistic, knowledge representation. Such a level can be assimilated to Kosslyn's long term memory and to Marr's hierarchical catalogue of models.

According to Gärdenfors, a conceptual space 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 (for  $1 \leq i \leq n$  with  $n \in \mathbb{N}$ ). Examples of such dimensions would be color, pitch, mass, spatial coordinates, and so on. The dimensions should be considered "cognitive" in that they correspond to qualities of the represented environment, without reference to any linguistic descriptions. In this sense, a conceptual space is prior to any symbolic characterization of cognitive phenomena. Some dimensions in a conceptual space are closely linked to the sensorial input of the system, other dimensions can be related to more abstract concepts.

We call *knoxel*<sup>5</sup> a generic point in a conceptual space (the term is suggested by the analogy with the term *pixel* in digital image processing); knoxels therefore represent epistemological primitives at the considered level of analysis. Formally, a knoxel is a vector  $k = (x_1, x_2, \dots, x_n)$  where  $x_i \in X_i$  corresponds to a parameter associated with a quality dimension of the domain of interest. In our architecture, the dimensions of the conceptual space are the parameters of the 3D geometric primitives which compose the scene. In this perspective, the knoxels correspond to simple geometric building blocks, while complex objects or situations are represented as suitable sets of knoxels. Accordingly, each knoxel is related to measurements, obtained via suitable sensors, of the geometric parameters of simple, basic objects in the external environment. A metric function  $d$  is defined in  $CS$ , which may be considered as a measure of similarity among knoxels in the conceptual space (see Gärdenfors [32]).

In general terms, a precise characterization of the conceptual space poses some problems. This is the case, in particular, when one has to take into account the qualitative difference in the information being represented in each dimension. It is, for instance, a complex task to find a metric that allows for a suitable quantization of the interesting features. Gärdenfors [34] notes that:

---

<sup>5</sup> The term *knoxel* was first introduced by Gaglio, Puliafito, Paolucci, and Perotto [30] with a slightly different meaning.

The main factor preventing a rapid development of a cognitive semantics based on conceptual spaces is the lack of knowledge about the relevant quality dimensions. It is almost only for perceptual dimensions that psychophysical research has succeeded in identifying the underlying topological structures (and, in rare cases, the psychological metric). For example, we only have a very sketchy understanding of how we perceive and conceptualize things according to their shapes. The models developed by Marr and Nishihara [48], Pentland [52], Biederman [13], and Tversky and Hemenway [65] among others, seem to point in the right direction, but there still remains a lot to learn about the “shape space”.

Nevertheless, we claim that our architecture overcomes these problems since we have adopted a very simple (but nonetheless useful) conceptual space in which the dimensions correspond to the parameters of suitable 3D geometric primitives. Their boolean composition, according to schemas of Constructive Solid Geometry (CSG<sup>6</sup>) as described by Requicha [55], permits the representation of a great variety of familiar shapes, particularly those corresponding to human artifacts. We have found convenient to adopt the *superquadrics* as the geometric primitives of the CSG schema. They are widely used both in computer graphics [8] and computer vision [52,57,66] as they offer an acceptable compromise between the compression of information in the scene and the necessary computational costs [57,46]. Furthermore, superquadrics provide good expressive power and representational adequacy [52].

Solina and Bajcsy [57], Gupta and Bajcsy [35], Leonardis, Solina and Macerl [46], among others, have proposed working techniques for recovering superquadrics from real scenes, even when the objects are difficult to segment. Techniques aimed at the recovery of superquadrics, also in the presence of occlusions, have been proposed by Whaite and Ferrie [66] and by Maver and Bajcsy [49].

Superquadrics are geometric shapes derived from the quadrics parametric equation with the trigonometric functions raised to two real exponents. The inside/outside function of the superquadric in implicit form is:

$$F(x, y, z) = \left[ \left( \frac{x}{a_x} \right)^{\frac{2}{\varepsilon_1}} + \left( \frac{y}{a_y} \right)^{\frac{2}{\varepsilon_2}} \right]^{\frac{\varepsilon_2}{\varepsilon_1}} + \left( \frac{z}{a_z} \right)^{\frac{2}{\varepsilon_1}} \quad (1)$$

where the parameters  $a_x, a_y, a_z$  are the lengths of the superquadric axes and the exponents  $\varepsilon_1, \varepsilon_2$ , called *form factors*, are responsible for the shape’s form:

---

<sup>6</sup> According to the CSG schema, the geometric primitives can be considered closed compact sets in Euclidean space, and they can be composed through regularised boolean operators (*R-AND*, *R-OR*, *R-DIFF*) to form general 3D structures.

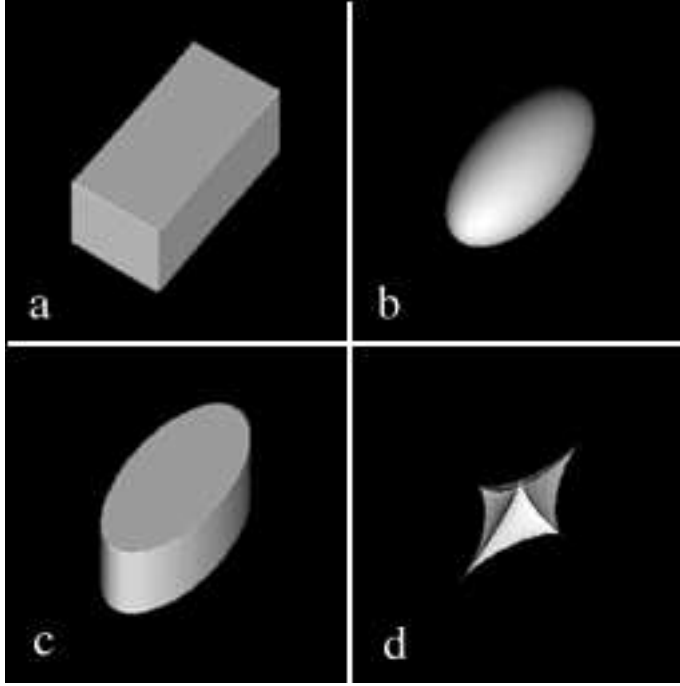


Fig. 2. Aspects assumed by a superquadric by varying its form factors.

$\varepsilon_1$  acts in terms of the longitude, and  $\varepsilon_2$  in terms of the latitude of the object's surface. (1) returns a value equal to 1 when the point  $(x, y, z)$  is a superquadric boundary point, a value less than 1 when it is an inside point, and a value greater than 1 when it is an outside point. Fig. 2 shows the forms assumed by a superquadric by varying only its form factors  $(\varepsilon_1, \varepsilon_2)$ . Form factors less than 1 let the superquadric take on a squared form, as in Fig. 2a where the values  $(0.01, 0.01)$  result in a box-shaped superquadric; values approaching 1 render the shape rounded, as in Fig. 2b, where the form factors  $(1, 1)$  make the superquadric an ellipsoid. When the form factors are  $(0.01, 1)$ , the superquadric assumes a cylindrical shape (see Fig. 2c). Finally, values greater than 1, e.g.,  $(5, 5)$ , tend to generate a cuspidate aspect, as in Fig. 2d.

The previous equation is the parametric equation in canonical form of a superellipsoid: the three center coordinates  $p_x, p_y, p_z$  and the three orientation parameters  $\varphi, \vartheta, \psi$  completely describe a generically displaced superquadric. The expression of the knoxel, describing a generic superquadric is therefore:

$$k = (a_x, a_y, a_z, \varepsilon_1, \varepsilon_2, p_x, p_y, p_z, \varphi, \vartheta, \psi) \quad (2)$$

As an example, let us consider the sample scene in Fig. 3 representing a hammer, a computer mouse and a tennis ball. The knoxels are obtained by approximating each part of the scene by means of the best fitting superquadric (see Fig. 4); details on this operation performed by our experimental setup will be given in Section 8. Each superquadric has been indicated by a tag; the acquired scene is therefore described by the knoxels  $k_1, k_2, k_3, k_4$ .



Fig. 3. A sample scene representing a hammer, a computer mouse and a tennis ball.

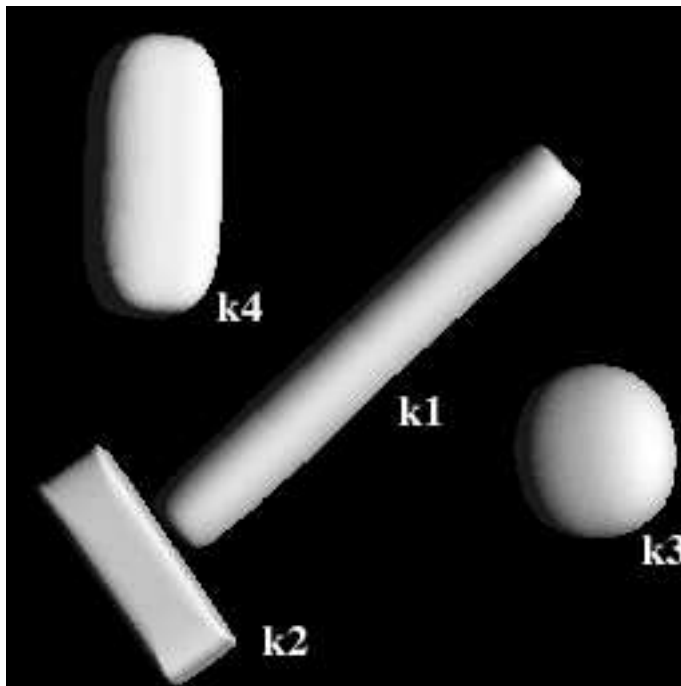


Fig. 4. Results obtained by the superquadric approximation of the scene in Fig. 3.

As previously stated, a knoxel individuates a single superquadric; complex objects and situations are represented by suitable sets of superquadrics according to the CSG schema. It should be noted that the superquadric parameters also code the position and orientation of the superquadric in space; therefore the relative orientation and mutual position of the superquadrics describing a

composite object, e.g., the hammer, are implicitly defined. There is no need of mechanisms such as the *adjunct relations* proposed by Marr [48].

We define a *perception cluster*  $pc = \{k_1, k_2, \dots, k_l\}$  as a finite set of knoxels corresponding to an object or a situation in  $CS$ . Referring to Fig. 4, the perception cluster  $pc_1 = \{k_1, k_2\}$  describes a hammer, while the perception cluster  $pc_2 = \{k_3\}$  describes a tennis ball. The set  $PC$  of all the *perception clusters* in  $CS$  is defined as:

$$PC = \{\{k_1, k_2, \dots, k_l\} | l \in \mathbb{N}, k_i \in CS \text{ for } 1 \leq i \leq l\} \quad (3)$$

The conceptual level is independent of any linguistic characterization. Indeed, the symbols at the linguistic level are interpreted on configurations at the conceptual level. A suitable interpretation function maps linguistic expressions onto conceptual structures of the appropriate type. In Section 5, we describe how this interpretation function may be “computed”.

#### 4 The linguistic level

The role of the linguistic level is to provide a concise description of the perceived scene in terms of a high-level logical language, in itself 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 [50]. Accordingly, 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.

The distinction between terminological and assertional components is useful for maintaining the distinction between the conceptual knowledge, which is largely independent of the specific perceived scene, and the assertions concerning the scene itself. Moreover, terminological formalisms are well suited to our purposes, in that they are centered on conceptual descriptions. This allows for a compact description of concepts, whose instances are to be recognized in the perceived scene. The adopted formalism is completely monotonic (it is well known that in classic terminological system concept description in terms of default attributes is not allowed). Non-monotonic extensions of the conceptual knowledge base would probably demonstrate themselves to be helpful in further developments of the system. Up to now, however, we have chosen to keep the symbolic knowledge base completely monotonic, in order that the prototypical characterization of concepts might emerge entirely from the properties

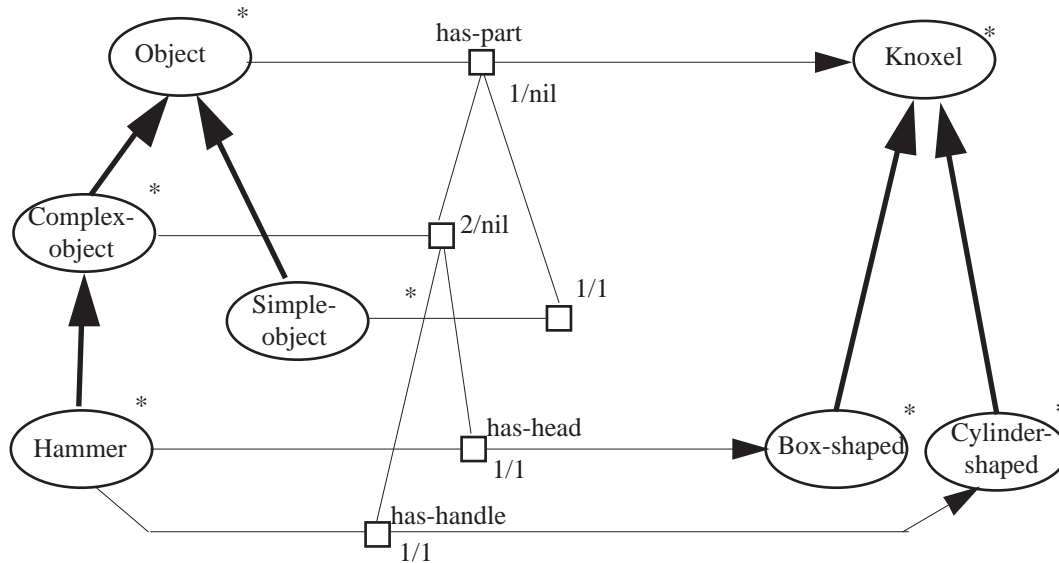


Fig. 5. Graphic description of a fragment of the terminological knowledge base. A generic *Object* is described as composed of at least one knoxel. A *Simple-Object* is described as an object composed of exactly one knoxel; a *Complex-Object* is an object composed of at least two knoxels. *Hammer* is an example of a complex object. The role *has-part* has been differentiated into more distinct roles. The concept *Hammer* has two roles: *has-handle* and *has-head*.

of the conceptual level and from the associative mechanisms linking it to the linguistic level, as proposed by Gärdenfors [32].

As an example, consider in Fig. 5 a fragment of the terminological knowledge base concerning the description of objects. In the figure, the graphic notation developed by Brachman [18] for the *KL-ONE* system has been adopted. A generic *Object* is described as composed of at least one knoxel. A *Simple-Object* is described as an object composed of exactly one knoxel; a *Complex-Object* is an object composed of at least two knoxels. *Hammer* is an example of a complex object. The role *has-part* has been differentiated into more distinct roles. For example, the concept *Hammer* has two roles: a role *has-handle* with exactly one filler, which must be a knoxel with a cylindrical shape, and a role *has-head* with exactly one box-shaped filler.

The assertional component is based on a first order predicate language, in which the concepts of the terminological component correspond to one argument predicates, and the roles (e.g., *has-head* or *has-handle*) correspond to two argument relations. So, for example, in order to assert the existence of an instance *Hammer#1* of the concept *Hammer*, the formula:

*Hammer*(*Hammer#1*)

is asserted. To express that the filler of the role *has-handle* for *Hammer#1* is a specific knoxel *Cylinder-shaped#1*, the formula:

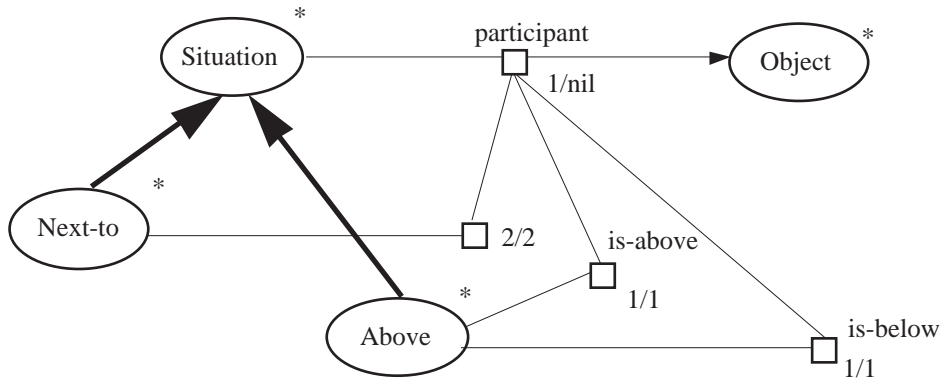


Fig. 6. Graphic description of the *Situation* concept. Every situation has at least one object as participant. *Next-to* and *Above* are described as a particular types of situations, with exactly two participants.

`has-handle(Hammer#1,Cylinder-shaped#1)`

is asserted.

As far as situations are concerned, we choose to represent them as concepts in the terminological formalism. In other words, we assume that situations are reified, i.e., that to every specific situation there corresponds an individual in the domain. This solution is analogous to Davidson's proposal for event representation [25]. Since we have no philosophical worries of ontological parsimony, this choice turns out to be simpler and advantageous in many respects. It is well suited for terminological formalisms, and provides a great flexibility and expressive power. For example, quantification on situations is allowed. Fig. 6 shows the network description of the *Situation* concept, and of two particular types of situation, *Above* and *Next-to*. As shown in Fig. 6, every *Situation* has at least one object as participant. *Next-to* is described as a particular type of situation, with exactly two participants. To assert that an object `O#1` is by the side of a second object `O#2`, an instance `S#1` of *Next-to* is generated, whose participants are `O#1` and `O#2`. In other words, the following assertions are added:

```
Next-to(S#1)
participant(S#1,O#1)
participant(S#1,O#2)
```

The situation *Above* is described by means of two roles, *is-above* and *is-below*, both with exactly one filler. These roles are defined as particular differentiations of the role *participant*.

## 5 Interpreting symbols on the conceptual space

As pointed out in the previous Section, the linguistic level provides a concise symbolic description of the perceived scene. Obviously, in the perception process this stage comes at the end, since it pertains to the most abstract representational level of visual information. What we need at this point is a *denomination function* which maps structures within the conceptual space onto linguistic constructs. A possible solution will be presented in the next Section, and it will be related to the focus of attention mechanisms.

In order to define a denomination function correctly, we cannot avoid proceeding in the opposite direction; i.e., we need to introduce an internal, cognitively oriented semantic interpretation for the symbols at the linguistic level. In particular, we define a suitable *interpretation function* that maps the symbolic structures at the linguistic level onto entities in the conceptual space. This is a general methodological issue in artificial intelligence where it is normally assumed that there is a language that needs a semantics. By contrast, in the perspective of the vision context, the main problem is that there is a perceptual representation that needs a language.

The proposed interpretation function  $\Phi$  associates any individual constant representing an object or a situation at the linguistic level with a perception cluster in  $CS$ , any concept (one-place predicate) with a set of perception clusters, any role (two-place predicate) with a set of pairs of perception clusters, and so on. Therefore, if  $C$  is the set of assertional individual constants and  $\Phi^C$  is the interpretation function  $\Phi$  restricted to  $C$ , then  $\Phi^C$  has the following type:

$$\Phi^C: C \rightarrow PC \quad (4)$$

where  $PC$  represents the set of all perception clusters as defined in (3). As an example referring to the scene in Fig. 4, the interpretation function  $\Phi$  associates, among others, an instance of the concept *Hammer* with the perception cluster  $pc_1 = \{k_1, k_2\}$ , where  $k_1$  and  $k_2$  are the superquadrics representing the hammer head and the hammer handle:

$$\Phi(\text{Hammer\#1}) = \{k_1, k_2\} \quad (5)$$

The compositional aspects of the interpretation of symbolic structures at the linguistic level can be defined according to the usual model theoretic semantics of terminological languages, as described by Nebel [50]. The main difference between the proposed semantics and the usual model theoretic approach is that in our approach individual constants are not interpreted on unstructured

set theoretical entities (the elements of the domain). On the contrary, perception clusters are objects endowed with a rich internal structure. This fact involves relevant consequences. In the traditional model theoretic approach the extension of primitive atomic predicates can be only assumed as given, and is, in a certain sense, completely arbitrary. In our approach, the extension of many primitive predicates can be determined on the basis of the structure of the entities in the semantic model itself.

As a simple example, consider the *has-part* role of the *Object* concept. Given the assertion `has-part(Hammer#1,Cylinder-shaped#1)` in a purely extensional model theoretic semantics, its truth is justified exclusively by the fact that the pair of the extensions of `Hammer#1` and of `Cylinder-shaped#1` belongs to the extension of `has-part`:

$$\langle \Phi(\text{Hammer\#1}), \Phi(\text{Cylinder-shaped\#1}) \rangle \in \Phi(\text{has-part}) \quad (6)$$

In the internal semantics, the truth of the previous assertion can be determined by examining the entities on which `Hammer#1` and `Cylinder-shaped#1` are interpreted in the conceptual space: the assertion is true if the set of knoxels on which `Cylinder-shaped#1` is interpreted is a subset of the set of knoxels on which `Hammer#1` is interpreted:

$$\Phi(\text{Cylinder-shaped\#1}) \subset \Phi(\text{Hammer\#1}) \quad (7)$$

The assumption according to which the individual constants representing objects are interpreted onto perception clusters is a simplification made possible by the fact that we are dealing with static scenes. To characterize objects independently of position and orientation, the perception clusters would be properly parametrized with respect to some of its constituents, i.e., they must be projected onto suitable subspaces of the whole conceptual space. Similarly, in a dynamic context, the internal structure of the semantic entities can be articulated further, in order to justify, at the semantic level, the truth of other kinds of atomic sentences. Consider, for example, object categorization. A given object can be recognized as an instance of a concept *Flexible-object* if, in the set of the perception acts concerning it at different instants, the object itself underwent some kinds of deformation, i.e., if the shape factors or the length of axes varied within certain ranges.

## 6 The focus of attention

As mentioned in the introduction, a finite agent with bounded resources cannot carry out a one shot, exhaustive, and uniform analysis of a perceived

scene within reasonable time constraints. 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. This is a typical problem of traditional symbolic models: Doyle [27] and Cherniak [22] stress the fact that, in order to avoid the proliferation of insignificant true conclusions, the aims and the purposes of an agent must be taken into account in the modeling of inferential activities. In modeling perception, these problems can be faced by taking into account the fundamental role of attentive phenomena in vision, as described in the work of Yarbus [67]. In the psychological literature, the focus of attention has sometimes been described as a spotlight which scans the visual field, individuating relevant aspects (see Posner [54]). This mechanism is analogous to the scanning of a mental image, as described by Kosslyn [43].

Several models of focus of attention mechanisms have been proposed in the artificial vision literature. An interest in some form of active processes during the recognition process is present in Marr's work [47] as well. The early focus of attention models aimed at searching for a particular object in the scene, given a static model of the object. The basic purpose of an attentional mechanism is computational efficiency (see Ballard [7]). The subject has become a key point in the field of the active vision research; the interest in this argument has been summarized by Bajcsy and Campos [6] who propose the "active and exploratory" framework for perception. According to this framework, the perception process of a living or artificial organism is based on four characteristics: it is an active and flexible task, it must have exploratory capabilities, it is a selective process, and it must be able to learn from the environment.

A strategy adopted by the active vision researcher in order to model the focus of attention mechanism aims mainly at choosing an optimal viewing position for the sensors, in order to improve the interpretation of the image and to minimize uncertainty. According to this strategy, Whaite and Ferrie [66] propose a probabilistic measure of the uncertainty of the superquadrics parameters, with respect to a general view position. The observation point is therefore changed in order to minimize this uncertainty. Maver and Bajcsy [49] propose a similar strategy for reasoning about occlusions, that takes into account the knowledge of the sensor geometry. They plan the next positions of the sensor in order to extract information from regions of missing data.

Another well-used strategy to model the focus of attention, (see Burt [21], Tsotsos, Culhane, Wai, Lai, Davis, and Nufflo [64]) is based on the pyramidal approach. Accordingly, the image is represented by a hierarchical data structure; "fine-to-coarse" algorithms generate the image measures, while "coarse-to-fine" search strategies are able to locate objects or situations in the scene. A high level control system drives the gathering mechanism.

Other adopted strategies are based on Bayesian and causal models of the

focus of attention. Rimey and Brown [56] propose TEA-1, a task-oriented system that expends the minimum effort necessary for solving a specific task. The knowledge of the system is structured by Bayesian networks, while the control of action is carried out by a benefit-cost analysis. The system is able to answer to questions about table settings, such as “Is this a fancy or an informal meal?”; the system activates the suitable visual actions controlling the focus of attention movements and the image processing tasks in order to answer the question. Birnbaum, Brand, and Cooper [15] propose the BUSTER system, which is aimed at developing a causal explanation of the scene. The attention is driven by causal semantics in order to find the causal role of elements in the scene and the causal relationships among the elements. BUSTER codes in terms of rules a simple physical knowledge about static scenes made up of structure block stacks incorporating architraves, cantilevers and balanced structures.

It is well known that at the lower, preattentive levels of visual perception there is a global parallel processing of visual information. The data received in input are concurrently processed, in order to produce a global reconstruction of the perceived scene. All data at this level have the same relevance, and no distinction is made between important and irrelevant information. According to Duncan and Humphreys [28], the goal of preattentive processing is a segmentation of the visual field into regions relevant from a purely perceptual point of view. At the attentive level, on the other hand, there is a sequential processing of visual information. From this point of view there is, in general, no “one shot” recognition of an object or of a scene; objects and scenes are, instead, recognized through a sequential exploration of the perceived image.

In our architecture, the conceptual level described in the previous Section acts as a “buffer interface” between subsymbolic and linguistic processing. The information coming up from the subsymbolic level has the effect of temporarily activating an (eventually very large) set of knoxels in the conceptual space. 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 order to describe the focus of attention mechanism, we denote as  $CS^*$  the set of all the possible sequences of elements belonging to  $CS$ , i.e., the set of all the possible sequences of knoxels:

$$CS^* = \{k_1, k_1k_2, k_1k_3, \dots\} \quad (8)$$

We define a *perception act*  $p$  as a generic sequence of knoxels in the conceptual space:

$$p \in CS^* \quad (9)$$

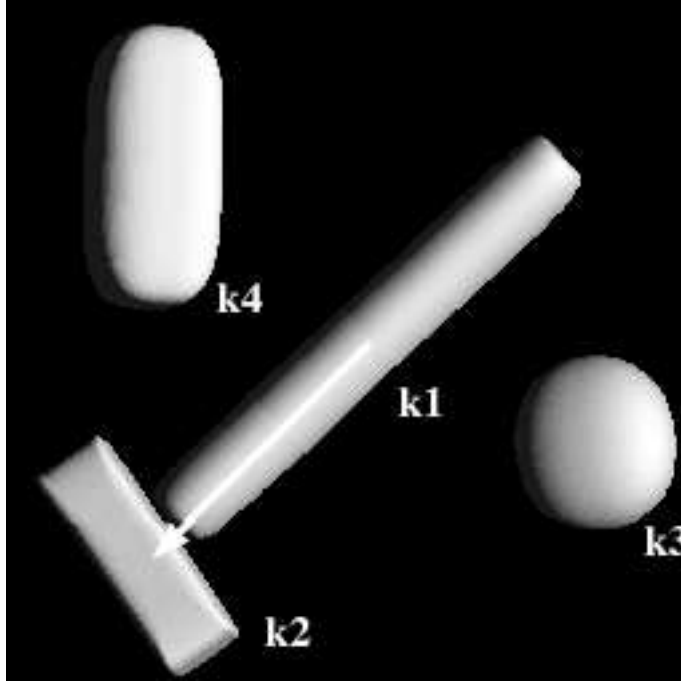


Fig. 7. A perception act related to the scene in Fig. 3; the perception act describes the hammer as a sequence of the hammer handle knoxel and the hammer head knoxel.

Considering the scene in Fig. 3, a possible perception act may be:  $p_1 = k_1k_2$ . This perception act describes a way of perceiving the hammer as a sequence of its handle and its head (see Fig. 7).

With reference to a perception cluster  $pc$ , we say that a perception act  $p$  is associated with the perception cluster  $pc$  if  $p \in pc^*$ , where  $pc^*$  is the set of all sequences of knoxels belonging to  $pc$ . As an example, the previously introduced perception act  $p_1 = k_1k_2$  is associated with the perception cluster  $pc_1 = \{k_1, k_2\}$  describing the hammer. The perception acts associated to a perception cluster therefore correspond to specific ways of perceiving an object or situation described by the perception cluster. It should be noted that the sequence of knoxels that makes up a perception act may not include all the knoxels of the corresponding perception cluster and/or it may include the same knoxels several times. In fact, the perception acts  $pc_2 = k_1k_2k_1$ , or  $pc_1 = k_2k_1$  may also be considered as associated to the perception cluster  $pc$ . They correspond to other ways of perceiving the hammer present in the scene.

We introduce a *denomination function*  $\Theta$  associating perception acts with assertions at the linguistic level:

$$\Theta: CS^* \rightarrow Assertion \quad (10)$$

where *Assertion* is the set of grounded well formed assertional formulas. Given

a perception act  $p$ ,  $\Theta(p)$  is a grounded assertional formula in which a new assertional constant occurs, that denominates  $p$ . This new assertional constant is the “name” that the system associates with the perception act  $p$ . Block C in Fig. 1 implements the denomination function by means of suitable attractor neural networks, as described in the next Section.

According to the previous example, the denomination function maps the perception act  $p_1 = k_1k_2$  related to the hammer, to the instance **Hammer#1** of the concept *Hammer*:

$$\Theta(k_1k_2) = \text{Hammer}(\text{Hammer\#1}) \quad (11)$$

Analogous considerations can be made for concepts describing spatial situations, as *Above* and *Next-to*.

Our proposal for the description of complex concepts in terms of sequences of knoxels is a way of dealing with attentional mechanisms in a well-founded manner, which naturally extends Gärdenfors’ notion of conceptual space. It should be noted that the perception act assumption avoids the needs of augmenting the dimensions of the space in order to describe complex objects or situations made up by several blocks: complex objects can be described by perception acts of arbitrary length.

In order to individuate the grouping paths among knoxels and to generate the most significant perception acts, it is necessary to orient the focus of attention in a suitable manner. In human beings, the focus of attention can be oriented either voluntarily, under the guidance of high level cognitive information and processes, or automatically, in dependence on particular stimuli present in the perceptive field, as described by Posner [54] and Jonides [40]. We assume that the focus of attention is determined by three concurrent modes: the *reactive*, the *linguistic* and the *associative* mode.

The reactive mode is the simplest one: the grouping paths among knoxels 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. As an example related to the previous scene, when the architecture is in reactive mode the focus of attention is directed to the hammer handle and to the hammer head because of their volumetric extension, thus generating the perception act  $p_1 = k_1k_2$  (see Fig. 7). The knoxels related to this perception act are sent to the denomination block to find the corresponding linguistic constants at the linguistic level. The denomination block correctly denominates the input perception act as an instance of the *Hammer* concept. The assertions generated at the linguistic level describing the operation of the architecture in reactive mode are reported in Fig. 8.

---

Knoxel (#k1)  
Knoxel (#k2)

Cylinder-shaped(#k1)  
Box-shaped(#k2)  
Hammer (Hammer#1)  
has-part(Hammer#1,#k1)  
has-part(Hammer#1,#k2)

---

Fig. 8. The assertions generated at the linguistic level related to the perception act represented in Fig. 7.

In the linguistic mode, the focus of attention is driven by the symbolic information explicitly represented at the linguistic level. Consider again the hammer example (Fig. 7). At the linguistic level a hammer is described as composed of a handle of cylindrical shape and a head of boxed shape. Let us suppose that the denomination block has recognized the knoxel corresponding to the *Cylinder-shape*. The description of a hammer at the linguistic level reports that it is made by a cylinder-shaped head and a box-shaped handle. The linguistic level therefore hypothesizes that the cylinder shaped knoxel may be, among other things, a filler for the role *has-handle* of the concept *Hammer*. The linguistic mode of the focus of attention now attempts the identification of the parts of the hammer, in particular its handle and its head, in order to verify the presence of such a hammer in the scene. This corresponds to finding the suitable fillers for the role parts of the object, i.e., a filler for the *has-head* role and a filler for the *has-handle* role. Therefore, whenever a possible hammer handle is recognized, the focus of attention tries to identify a hammer by identifying the possible fillers for its head and its handle. When some of the expectations concerning these Knoxels are satisfied by some corresponding Knoxels in the scene, the perception act made up by the Knoxels that satisfy these conditions is sent to the denomination function in order to recognize the object or the situation, as in the reactive mode. It should be noted, however, that if the cylinder has been recognized, and there is no recognizable hammer head, i.e., the linguistic expectations are not satisfied, the architecture cannot recognize the hammer. Block D in Fig. 1 implements the linguistic expectation function by means of suitable attractor neural networks, as described in the next Section. The assertions generated at the linguistic level related to the described example are reported in Fig. 9.

In the associative mode of the focus of attention, 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. So if hammers and balls have been always present in the same scene, the weight of the association between the

---

Knoxel (#k1)  
 Knoxel (#k2)  
  
 Cylinder-shaped(#k1)  
 Box-shaped(#k2)  
 Hammer (Hammer#1)  
 has-handle(Hammer#1,#k1)  
 has-head(Hammer#1,#k2)

---

Fig. 9. The assertions generated at the linguistic level related to the linguistic expectations for the perception act represented in Fig. 7.

concepts *Hammer* and *Ball* is strong, and they mutually activate each other. As a consequence, whenever a hammer is recognized, the focus of attention tries to identify some balls in the perceived scene, and vice versa.

Let us suppose to have recognized the hammer by the linguistic mode (see the previous example). At the linguistic level, the concept *Hammer* is associated by a Hebbian mechanism to the concepts of *Ball* and *Mouse*, due to a previous learning phase. The linguistic level therefore hypothesizes the presence of these objects in the scene and the associative expectations block generates the corresponding hypotheses. As in the linguistic mode, when some of these expectations are satisfied by some corresponding Knoxels in the scene, the perception act made up by the Knoxels found to be so is sent to the denomination block. Fig. 10 shows the resulting perception act, while Fig. 11 shows the corresponding assertions generated at the linguistic level. Block E (Fig. 1) implements the associative expectations by means of attractor neural networks, just as for the linguistic expectations. The implementation will be described in the next Section.

The task of recognizing the perception acts in the case of spatial situations is similar to the task of recognizing the objects; Fig. 12 shows the assertions generated for the spatial concept *Next-to* (described in Section 5), after the recognition steps of the objects present in the previous scene. In particular the assertions state that the hammer and the ball are side by side: the perception act obtained as the sequence of the Knoxels of the hammer and the Knoxel of the ball ( $k_1k_2k_3$ ) has been recognized from the denomination block as a *Next-to* situation.

It should be clarified that the distinction between the associative mode and the linguistic mode is a soft one: even the linguistic mode in some sense “associates” the perceived object with some expected objects. As it will be explained in the next Section, both modes are implemented by attractor neural networks with suitable associative capabilities trained by a careful learning phase. The main difference between the two modes is that the associative mode captures

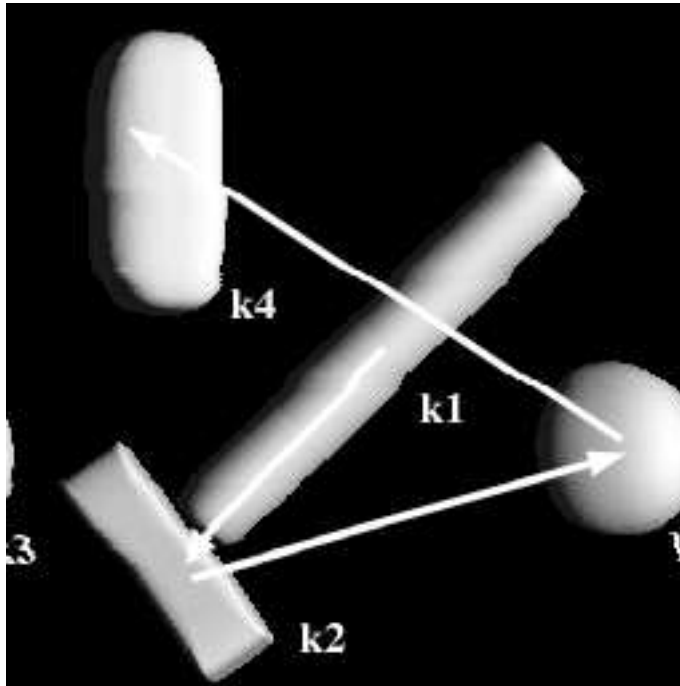


Fig. 10. The resulting perception act related to the previous scene when the focus of attention is driven by the associative expectation to find a ball and a mouse.

---

```

Knoxel (#k1)
Knoxel (#k2)
Knoxel (#k3)
Knoxel (#k4)

Cylinder-shaped(#k1)
Box-shaped(#k2)
Hammer (Hammer#1)
has-handle(Hammer#1,#k1)
has-head(Hammer#1,#k2)

Ball-shaped(#k3)
Ball(Ball#1)
has-part(Ball#1,#k3)

Ellipsoid-shaped(#k4)
Mouse(Mouse#1)
has-part(Mouse#1,#k4)

```

---

Fig. 11. The assertions generated at the linguistic level related to the associative expectations for the perception act represented in Fig. 10.

---

```
Next-to (Next-to#1)
participant(Next-to#1,Hammer#1)
participant(Next-to#1,Ball#1)
```

---

Fig. 12. The assertions generated at the linguistic level related to the spatial situations in the previous scene.

all the free associations not described by the semantic network at the linguistic level, while the linguistic mode associations are driven by the conceptual description at the linguistic level. In the previous scene, for example, the neural networks responsible for the associative mode of the focus of attention have learned to associate a hammer to a ball, but they have not learned to associate the cylinder related to the hammer handle, to the box related to the hammer head, also if the two objects are always present in the same scene. This kind of associations is in fact managed by the linguistic mode.

The main goal of the expectation generation process is to obtain the most exhaustive possible interpretation of the acquired scene by avoiding the generation of true but useless assertions. When the associative and linguistic expectations are not activated, the architecture describes the scene only by means of the simple reactive mode. In this case the architecture has no other choices than to build and denominate all the possible perception acts obtained by combining all the knoxels present in the scene. The reactive mode alone therefore generates a combinatorial exploding number of assertions, the most of which are true but uninformative. It should be noted, in fact, that the denomination of the objects strictly depends on the particular found knoxel sequence: when the input perception act contains the hammer head, the ball, and the hammer handle (Fig. 13), the denomination block does not recognize the hammer, but it recognizes the three knoxels as three distinct objects: a cylinder, a ball and a box. The generated assertions (Fig. 14) are true but they do not describe the scene exhaustively. Furthermore, as the reactive mode has no access to the descriptions of objects of the terminological component, the architecture is not able to fill the roles for the parts of an object: e.g., the reactive mode is able to recognize the hammer (see the assertions in Fig. 8), but it is not able to recognize the cylinder-shaped knoxel as the hammer handle and the box-shaped knoxel as the hammer head.

The denomination and attention mechanisms have been described up to now in isolation. As a matter of fact they operate concurrently by a simple recognition process cycle. The process is bootstrapped by the reactive mode of the focus of attention, which enables the denomination block to recognize objects “evident” in the scene: e.g., referring to Fig. 4, the reactive mode identified the hammer which is recognized by the denomination block. This allows a balancing between the associative and the linguistic modes of the focus of attention to satisfy their own generated expectations. As a default, the ar-

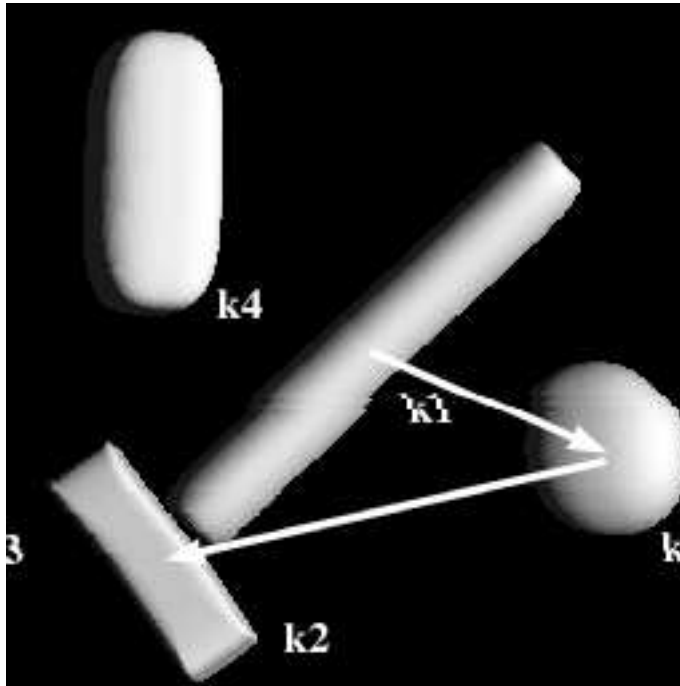


Fig. 13. Another perception act related to the previous scene; the focus of attention is directed to the hammer head, the ball, and the hammer handle.

---

```

Knoxe1 (#k1)
Knoxe1 (#k2)
Knoxe1 (#k3)

Cylinder-shaped(#k1)
Cylinder(Cylinder#1)
has-part(Cylinder#1,#k1)

Ball-shaped(#k3)
Ball(Ball#1)
has-part(Ball#1,#k3)

Box-shaped(#k2)
Box(Box#1)
has-part(Box#1,#k2)

```

---

Fig. 14. The assertions generated at the linguistic level related to the perception act represented in Fig. 13.

chitecture first tries sequentially to recognize the objects anticipated by the linguistic expectations, and then the objects anticipated by the associative expectations. However, it is possible to ignore one or both of them. When no expectations are satisfied, the recognition process restarts through the reac-

tive mode of the focus of attention in search of a new and as yet unrecognized object.

The focus of attention mechanism may be modeled as an *expectation function*  $\Psi$  linking the linguistic to the conceptual level; the function has its domain in the set *Assertion* of assertional grounded well formed formulas, and its range in the set of perception acts. Therefore the function  $\Psi$  is of the following type:

$$\Psi^i: \textit{Assertion} \rightarrow \textit{CS}^* \text{ where } i = 1, 2 \quad (12)$$

where  $\textit{CS}^*$ , as previously discussed, is the set of all perception acts; the index  $i$  indicates the attentive mode: 1 stands for the linguistic mode and 2 stands for the associative mode. The function  $\Psi$  generates expectations on the perception acts to be found in the conceptual space, on the basis of the available assertional information. In other words, the focus of attention looks for specific perception acts belonging to the perception clusters corresponding to the “expected” assertional constant in the perceived scene. We have chosen to model the attentive modes by the same function in order to reinforce the fact that the distinction between the associative mode and the linguistic mode is a soft one, and that they should be considered to be two faces of the same global attentive process.

In the linguistic mode, the assertional constants are generated when the system makes linguistic inferences; as a consequence the focus of attention generates the perception acts made up by knoxel samples for these assertional constants. The system globally computes the expectation function by taking into account the information described in the terminological component. In terms of the previous example, when a cylinder shaped knoxel is found, the focus of attention searches for the fillers of the roles *has-part* of the *Hammer* concept, e.g., a filler for *has-head* and a filler for *has-handle*. Therefore the function  $\Psi^1(\textit{Cylinder-shaped}(\mathbf{k\#1}))$  generates perception acts made up of sample fillers of these roles.

The associative mode is similar to the linguistic mode, except that the focus of attention searches for objects freely associated to the constants introduced at the linguistic level. In the hammer example, when a hammer is found, the function  $\Psi^2(\textit{Hammer}(\mathbf{Hammer\#1}))$  generates perception acts made up by samples of balls and mice.

In the reactive mode, the focus of attention searches for generic objects in the scene. For the purpose of uniformity, this mode is considered a special case of the linguistic mode where the expected object is an instantiation of the most generic class, e.g., an *Object*.

## 7 The connectionist implementation of the link between conceptual and linguistic levels

A perception cluster, as described in Section 3, is a set of knoxels associated to an object or a situation:  $pc = \{k_1, k_2, \dots, k_l\}$ . Each knoxel  $k_i$  may be viewed as a point attractor of a suitable energy function associated to the whole perception cluster. In this way, a set of fixed point attractors models and generates the perception cluster: starting from an initial state representing a knoxel imposed, for instance, from the external input, the system state trajectory is attracted in turn to the nearest stored knoxel of the perception cluster.

The implementation of a perception cluster by means of an *attractor neural network* (see Hopfield [38], Amit [3]), characterized by the corresponding energy function, appears to be a natural choice: each knoxel of the cluster is an activation pattern learned by the network. The implementation of the perception acts associated with a perception cluster is built by means of time delayed connections that learn the corresponding temporal sequences of knoxels, as proposed by Kleinfeld [41] and by Kleinfeld and Sompolinsky [42]. This modification allows the attractor neural network both to recognize and to generate all the perception acts corresponding to a concept. Therefore, to implement the denomination and expectation functions mapping the conceptual onto the linguistic level (the blocks C, D and E of Fig. 1), each concept at the linguistic level is associated to a suitable attractor neural network.

The choice of time-delay attractor neural networks offers several advantages. It is based on the well-studied energetic approach; the learning phase is fast, since it is performed at “one shot”. Furthermore, as it allows for a uniform treatment of both the recognition and the generation of perception acts, the denomination functions and the expectation functions introduced in the previous Section may be implemented by a uniform neural network architecture design.

For the sake of simplicity, we have adopted the *binary unit* version of the attractor neural network; the coding of the knoxels in terms of the binary activation pattern of the network has been computed by the coarse coding algorithm proposed by Hinton, McClelland, and Rumelhart [37].

The general expression of the energy function of an attractor neural network for a perception cluster is:

$$E_1(t) = - \sum_{i=1}^m \sum_{j=1}^m T_{ij} k_i(t) k_j(t) \text{ with } j \neq i \quad (13)$$

where  $m$  is the number of binary units of the network,  $\mathbf{T}$  is the connection matrix storing the attractors representing the knoxels of the perception cluster, and  $k(t)$  is the knoxel representing the current activation pattern of the network. The number  $m$  of units depends on the number  $l$  of knoxels in the perception cluster according to the *low memory load* condition discussed in Amit [3]:

$$l < \alpha_c m \quad (14)$$

where  $\alpha_c \simeq 0.3$ . The connection matrix  $\mathbf{T}$  is given by:

$$T_{ij} = \frac{1}{m} \sum_{\nu=1}^l k_{\nu_i} k_{\nu_j} \text{ with } j \neq i \quad (15)$$

where  $k_{\nu}$  is the  $\nu$ -th knoxel of the perception cluster.

In order to describe a perception act associated to the perception cluster, a sequential operation in the corresponding attractor neural network is implemented by introducing time-delayed connections among units. These connections store the time sequence of knoxels in the perception act; the resulting energy term is:

$$E_2(t) = - \sum_{d=1}^s \sum_{i=1}^m \sum_{j=1}^m D_{ij}^d k_i(t) k_j(t - d\tau) \text{ with } j \neq i \quad (16)$$

where  $\tau$  is the time delay among two subsequent knoxels in the perception act  $p$ ,  $s$  is the amplitude of the time window of interest,  $\mathbf{D}^d$  is the delayed synapses connection matrix related to the time delay  $d\tau$ ,  $k(t)$  and  $k(t - d\tau)$  are respectively the current and the past  $d\tau$ -th knoxel of the perception act.

The connection matrix  $\mathbf{D}^d$  is given by:

$$D_{ij}^d = \frac{1}{m} \sum_{\xi=1}^h k_{(\xi+d)_i} k_{\xi_j} \text{ with } j \neq i \quad (17)$$

where  $k_{\xi}$  and  $k_{(\xi+d)}$  are respectively the  $\xi$ -th and the  $(\xi + d)$ -th knoxel of the current perception act;  $h$  is the length of the considered perception act.

The global external input to the network is modeled by the energy term:

$$E_3(t) = - \sum_{i=1}^m \sum_{j=1}^m F_{ij} k_i(t) I_j(t) \text{ with } j \neq i \quad (18)$$

where  $\mathbf{F}$  is the external input connection matrix,  $\mathbf{I}(t)$  is the actual activation pattern input of the network coming from the conceptual space.

The connection matrix  $\mathbf{F}$  is given by:

$$F_{ij} = \frac{1}{m} \sum_{\nu=1}^h k_{\nu_i} L_{\nu_j} \text{ with } j \neq i \quad (19)$$

where  $L_{\nu}$  is the input corresponding to the knoxel  $k_{\nu}$ .

The global energy function is the sum of (13), (16), (18):

$$E(t) = E_1(t) + \lambda E_2(t) + \varepsilon E_3(t) \quad (20)$$

where  $\lambda$  and  $\varepsilon$  are the weighting parameters of the time delayed synapses and the external input synapses, respectively.

The expectation function  $\Psi^i$ , corresponding to blocks D and E, are implemented by setting the parameters of the energy function  $E(t)$  to  $\lambda > 1$  and  $\varepsilon = 0$ . In fact, the task of these blocks is to generate suitable knoxel sequences representing the expected perception acts for the input assertion. This choice of parameters allows the transitions among knoxels to occur “spontaneously” with no external input. Referring to (20), it can be shown that an attractor is stable for a significant long time period due to the  $E_1(t)$  term, so that the output knoxel is easily observed. As  $\lambda > 1$ , the term  $\lambda E_2(t)$  after some  $d\tau$  is able to destabilize the attractor and to carry the activation pattern of the network toward the following attractor of the sequence representing the next knoxel of the stored perception act. The neural network therefore visits in sequence all the knoxels of the stored perception act related to the input assertion.

The denomination function  $\Theta$ , corresponding to the block C of Fig. 1, is implemented by setting the parameters of the energy function  $E(t)$  to  $\lambda < 1$  and  $\varepsilon > 0$ . The task of this block is the recognition of input knoxel sequences representing the input perception acts. To accomplish this task it is necessary to consider the input term  $E_3(t)$  in order to make the transitions among knoxels happen, as driven from the external input. When  $\lambda < 1$ , the term  $\lambda E_2(t)$  is not able itself to drive the activation pattern transition among the knoxels of the perception act, but when the term  $\varepsilon E_3(t)$  is added, the contribution of both terms will make the transition happen. The neural network therefore recognizes the input perception act as it “resonates” with one of the perception acts previously stored and generates the corresponding assertion.

To examine the operations of the neural networks employed, we adopt the  $\%overlap_k$  measure of performance (see Amit [3]), during network *epochs*, where an *epoch* is an activation cycle of the neural network. This measure

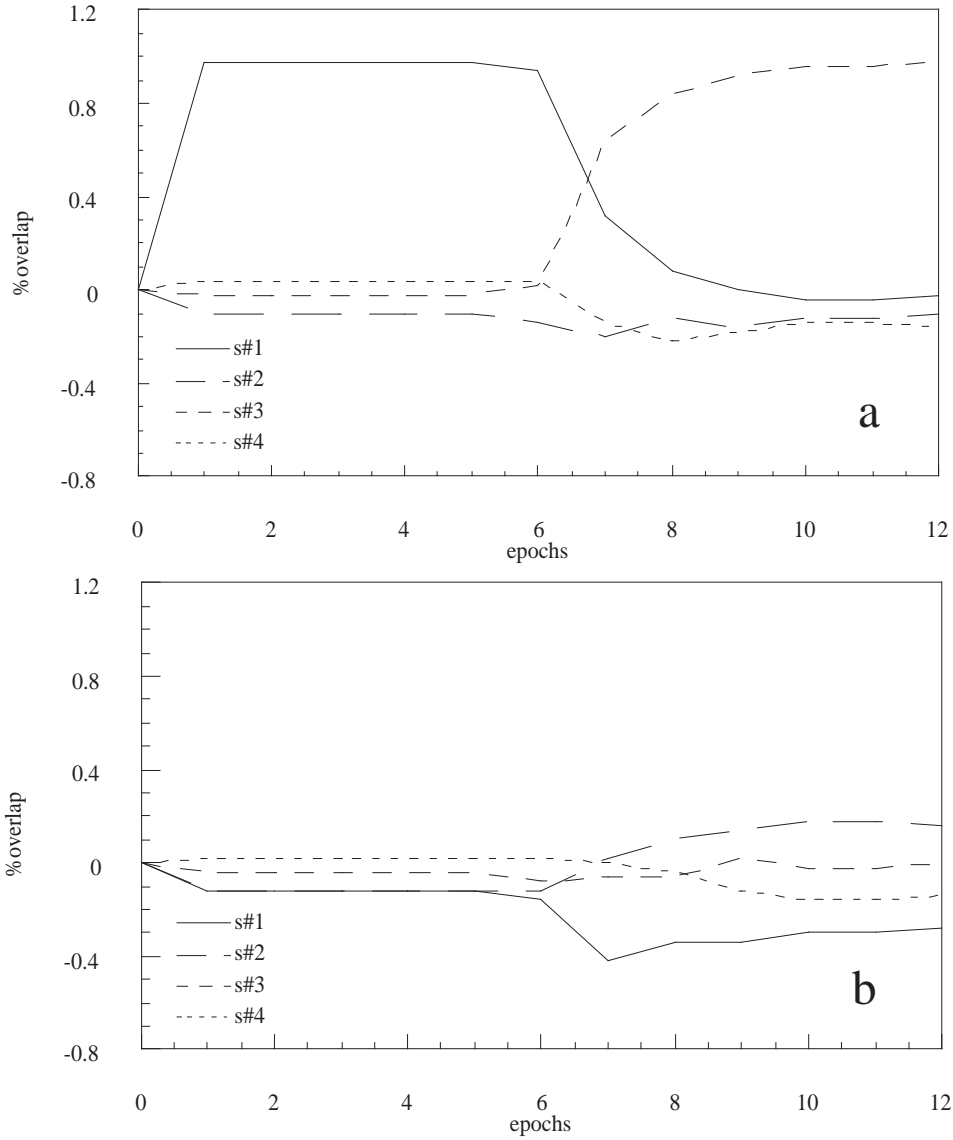


Fig. 15. The diagrams of the  $\%overlap$  Vs  $epochs$  of the neural networks associated to the concepts *Hammer* (a) and *Ball* (b) when the input is the perception act of Fig. 7.

of performance is defined with respect to a previously learned knoxel  $k$  as the time evolution of the overlap, in terms of the normalized dot product, between the current knoxel output of the network  $k(t)$  and the previously learned knoxel  $k$ :

$$\%overlap_k = \frac{k(t) \cdot k}{\|k(t)\| \|k\|} \quad (21)$$

Let us consider the operation of the attractor neural network in the situation depicted in Fig. 7: the focus of attention is directed, by the reactive mode, to the hammer handle and the hammer head, and the knoxels related

to this perception act are sent to the denomination block to generate the corresponding assertion at the linguistic level. Fig. 15 shows the  $\%overlap$  Vs  $epochs$  measures of the neural networks associated with the concepts *Hammer* (Fig. 15a) and *Ball* (Fig. 15b). Each line of the diagrams shows the  $\%overlap$  of the output activation pattern of the networks with respect to a previously learned knoxel, when the input is the perception act  $p_1$  describing the hammer (Fig. 7). It should be noted that the sequence of input knoxels, representing the hammer handle and the hammer head, “resonates” with the previously learned sequence of knoxels **s#1** and **s#3** of the network associated with the *Hammer* concept. On the other hand, the overlap of this sequence of knoxels with the sequences stored in the other network is low. Therefore the denomination block correctly denominates the input perception act as an instance of the *Hammer*, as described in the generated assertions reported in Fig. 8.

Let us consider at this point the operation of the linguistic expectations block (block D of Fig. 1) during the example described in the previous Section: a cylinder has been found and the linguistic level hypothesizes the presence of a hammer in the scene. The linguistic expectations block generates the hypothesized instances of the hammer head and of the hammer handle. Fig. 16 shows the  $\%overlap$  Vs  $epochs$  measure of the neural network generating the possible expected knoxel instances of the *has-handle* filler of the hammer. As in the previous diagrams, each line shows the  $\%overlap$  of the output activation pattern of the network with respect to a previously learned knoxel. It should be noted that the network generates the knoxel hypotheses **s#1**, **s#2** and **s#6** as possible hammer handles. The knoxel **s#4** does not belong to the hypotheses. After this step, the network generates the possible expected knoxel instances of the *has-head* fillers of the hammer. When some of these knoxels are satisfied by some knoxels in the scene, the resulting perception act is sent to the denomination block to recognize an instance of the *Hammer*, thus generating the assertions in Fig. 9.

The operation of the associative expectations block (block E of Fig. 1) in the example considered follows the same guidelines as the linguistic expectations block: at the linguistic level, the *Hammer* is associated, by a Hebbian mechanism, to the *Ball* and the *Mouse*, due to the previous learning phase. The attractor neural network generates the possible expected knoxel instances of the *Ball* and of the *Mouse*; at the end of the operation, the assertions of Fig. 11 are generated.

## 8 Experimental setup

This Section describes the setup adopted to obtain the examples presented throughout the theoretical discussion, along with other more complex exam-

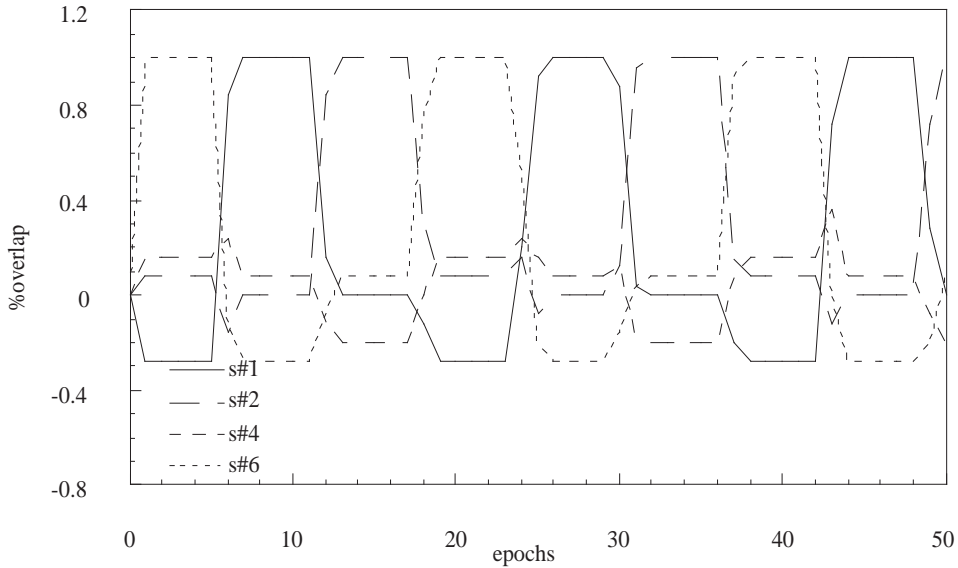


Fig. 16. The diagram of the  $\%overlap$  Vs  $epochs$  measure of the neural network generating the linguistic expectations for the *has-handle* filler of the hammer.

ples of the operation of the architecture. We have chosen an experimental framework that avoids some typical complex problems encountered in 3D vision. Also in this essential framework, our architecture is able to draw interesting inferences and to build an interpretation context. The framework consists of static scenes made up of objects like hammers, tennis balls, computer mice and telephones; all the objects rest on a uniform visually-contrasting planar backdrop. The objects are easy to segment and they are arranged in order to avoid occlusions. Sensory data are 2-D images acquired by a video camera (two-dimensional arrays of pixels) representing an orthogonal view of the observed scene, as in the Fig. 3.

Starting from the acquired pictorial image, the subsymbolic level (see Fig. 1) computes the segmentation map by means of a region growing algorithm (see Zucker [68]): the image is initially partitioned into elementary regions of uniform brightness and the adjacent regions, for which the contrast difference is low, are merged. Fig. 17 shows the segmentation map found after the region growing phase starting from the scene in Fig. 3. The relative depth map is then computed by the Tsai and Shah shape from shading algorithm [62] (see Fig. 18). We do not calculate the local orientation map. Both the depth map and the information about the segmented regions are fed as input to block B of Fig. 1. The first operation of this block is the volumetric representation of the input depth map by a spatial array. The result is a discrete representation of the spatial bulk of the objects present in the scene by *voxels*, i.e., in terms of primitive volume elements (see Fig. 19).

In order to describe the scene in terms of superquadric parameters, and therefore in terms of knoxels, each part of the scene that results from the region

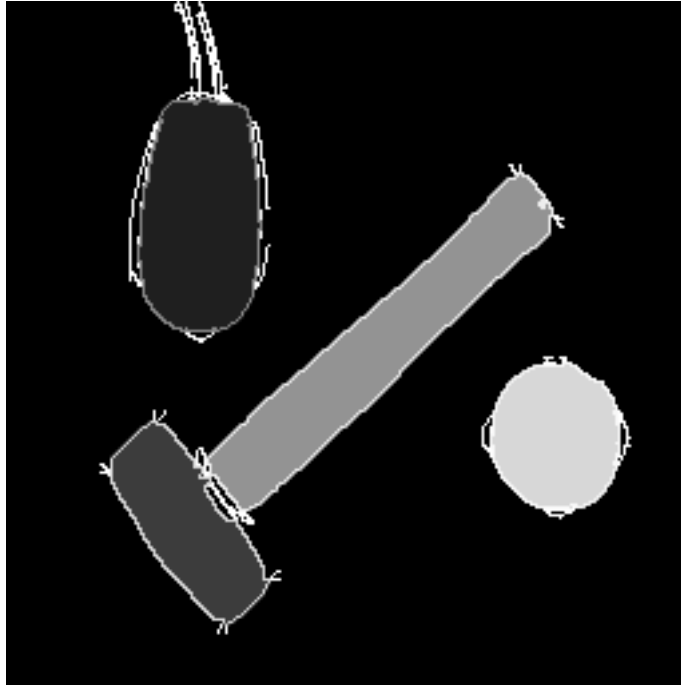


Fig. 17. The result of the segmentation phase. The regions found after the segmentation phase starting from Fig. 3 are set into relief.

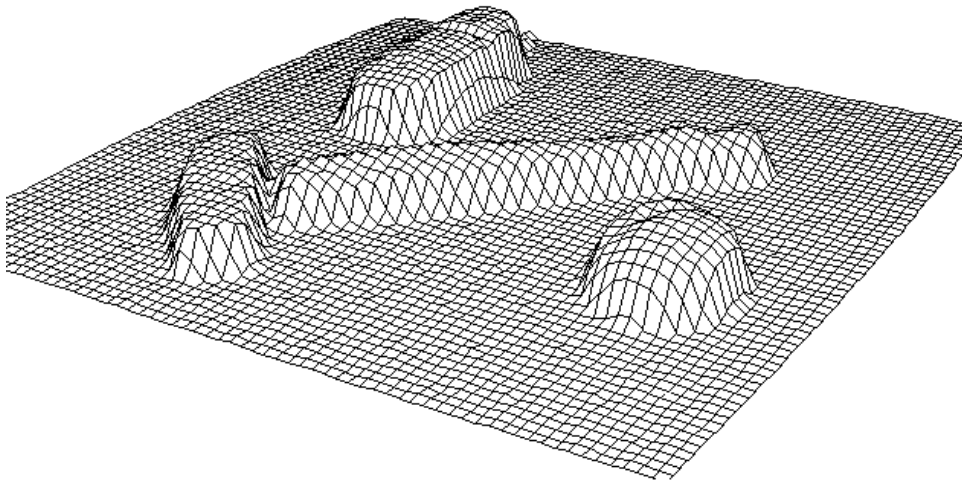


Fig. 18. The depth map of the acquired scene obtained by the shape from shading algorithm.

growing algorithm is approximated by means of the best fitting superquadric. The superquadric approximation operation is carried out by applying a simple two-step algorithm [4]. First, the center  $p_x, p_y, p_z$  and the orientation of the principal axes  $\phi, \theta, \psi$  of the part under consideration are calculated, by determining the point and the unit vectors with respect to which all the products of inertia are zero, by following the algorithm proposed by Chien and Aggarwal [23]. Once the center and the principal axes are known, the compu-

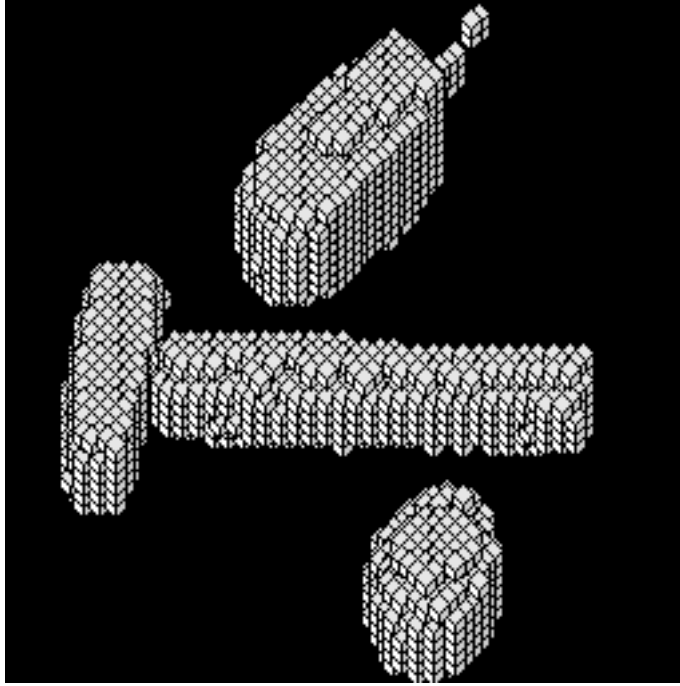


Fig. 19. The voxel representation of the acquired scene.

tation of the lengths  $a_x, a_y, a_z$  of the axes of the superquadric approximating the considered part is trivial. In the second step, the form parameters  $(\varepsilon_1, \varepsilon_2)$  that best correspond to the squareness features of the object are obtained by minimizing the error function proposed by Solina and Bajcsy [57]. Since the center, orientation and axes are known quantities, the error function depends solely on the form parameters and has a minimum value that corresponds to those values defining the superquadric that best fits the given part. The approximation of each part therefore requires an optimization procedure in the two-dimensional space of the form parameters. Fig. 4 shows the results of the recovery of the superquadrics of the acquired scene of Fig. 3: each region of Fig. 17 has been approximated by a superquadric.

Fig. 20 shows a more complex scene made up of a hammer, a cordless telephone, a wood block and a mouse. Fig. 21 shows the superquadric reconstruction of the same scene along with the focus of attention movements during the exploration of the scene. Fig. 22 shows the assertions generated at the linguistic level. By analyzing the focus of attention, it is possible to see that it follows two sequences: a sequence in which the attention is focused on the hammer, the block and the mouse, and another sequence in which the attention is focused on the body and the antenna of the telephone. The assertions generated at the linguistic level and the dynamics of the time-delay neural networks may therefore be analyzed as a concatenation of these two sequences. It should be noted that basing the focus of attention mechanism on the expectations generation allows the creation of “attentional contexts” within which an object is analyzed. In fact, during the analysis of the first sequence, the telephone is ig-

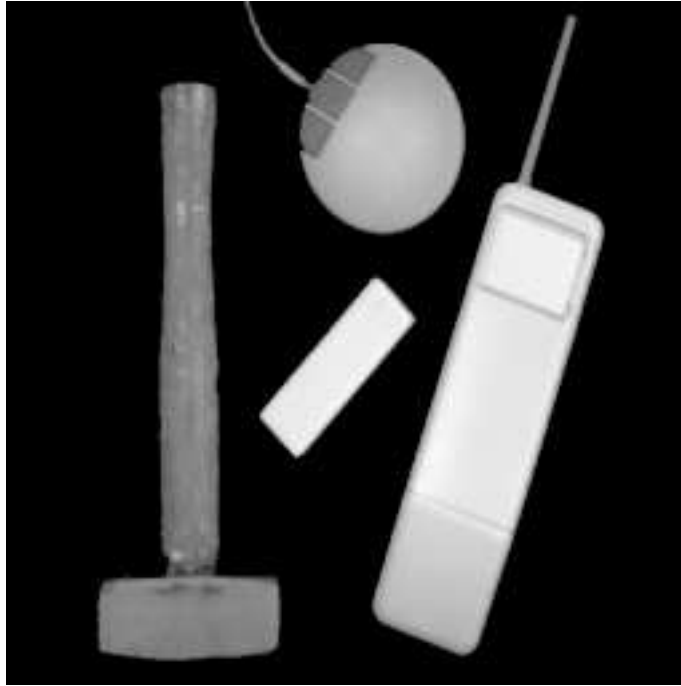


Fig. 20. A complex scene made up of a hammer, a cordless telephone, a wood block and a mouse.

nored, because the object does not belong to the current attentional context. The same thing occurs during the second sequence: the block, the hammer and the mouse are ignored because they do not belong to the same attentional context of the telephone. This allows to avoid the “cognitive overload” problem. The architecture is able to discover the relevant paths, and aggregate the information in order to generate only those linguistic descriptions that are “useful” and “interesting” in the current attentional context.

The scene represented in Fig. 23 demonstrates the same process: the screw and the cylinder make up the first context and the two blocks constitute the second. Fig. 24 shows the superquadric reconstruction of the scene along with the focus of attention movements, and Fig. 25 shows the assertions generated at the linguistic level.

## 9 Discussion and conclusions

The main goal of this work is to link together in a principled way two different research traditions: that of computer vision on one hand, and that of symbolic models of knowledge representation and reasoning, on the other hand. We maintain that this goal can be achieved by taking into account the results obtained in different subfields of cognitive science. The architecture we have described is a first step in this direction. In particular, two main assumptions

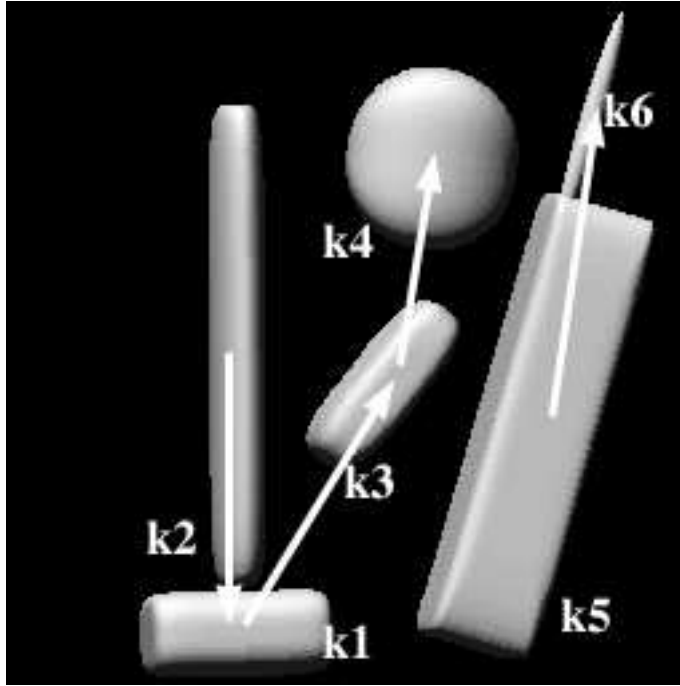


Fig. 21. The superquadric reconstruction of the scene in Fig. 20 along with the focus of attention movements during the exploration of the scene.

are critical for our proposal:

- (1) The existence of a conceptual level, intermediate between the lower vision level and the high level, symbolic representation. The conceptual level has a non linguistic nature (it is independent of any linguistic formulation), and it is modeled in terms of a conceptual space. It is generated starting from the outputs of the vision module, and has the role of providing an interpretation for the symbols of the linguistic level.
- (2) The link between the conceptual level and the linguistic representation is achieved through a focus of attention mechanism, that has the effect of scanning in a sequential way the information processed at the lower levels. This hypothesis stands on the widely shared psychological assumption, according to which lower level vision is based on massive parallel information processing, while high level attentive phenomena are of a sequential nature.

Given these basic assumptions, the specific choices that have been made in working out the architecture make it very general. Obviously it can be adjusted easily enough to accomodate more specific choices.

The architecture extends previous work on scene understanding [19,20,63] by providing a cognitive framework in which to embed 3D reconstruction performed by current artificial vision architectures [11,24]. It provides a well-founded interpretation mechanism that builds a rich linguistic description

---

Knoxe1 (#k1)  
Knoxe1 (#k2)  
Knoxe1 (#k3)  
Knoxe1 (#k4)  
Knoxe1 (#k5)  
Knoxe1 (#k6)

Cylinder-shaped(#k2)  
Box-shaped(#k1)  
Hammer (Hammer#1)  
has-handle(Hammer#1,#k2)  
has-head(Hammer#1,#k1)

Box-shaped(#k3)  
Block(Block#1)  
has-part(Block#1,#k3)

Next-to(Next-to#1)  
participant(Next-to#1,Hammer#1)  
participant(Next-to#1,Block#1)

Ellipsoid-shaped(#k4)  
Mouse(Mouse#1)  
has-part(Mouse#1,#k4)

Above (Above#1)  
is-above(Above#1,Mouse#1)  
is-below(Above#1,Block#1)

Parallelepiped-shaped(#k5)  
Thin-cylinder-shaped(#k6)  
Telephone(Telephone#1)  
has-body(Telephone#1,k#5)  
has-antenna(Telephone#1,#k6)

---

Fig. 22. The assertions generated at the linguistic level related to the perception acts represented in Fig. 21.

of the perceived scene. This linguistic description may be considered as the ground level for complex symbolic spatial reasoning activities, which up to now have been modeled without any reference to actual interaction with the external environment [44]. The proposed focus of attention mechanism complements at the cognitive level the current work on active vision which is mainly modeled in reactive terms [5,6].

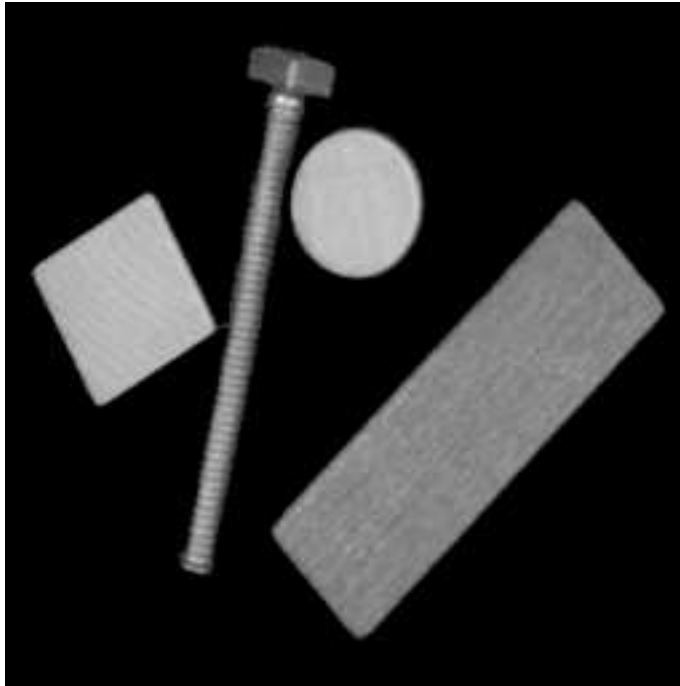


Fig. 23. A complex scene made up of a screw, a cylinder, a square block and a rectangular block.

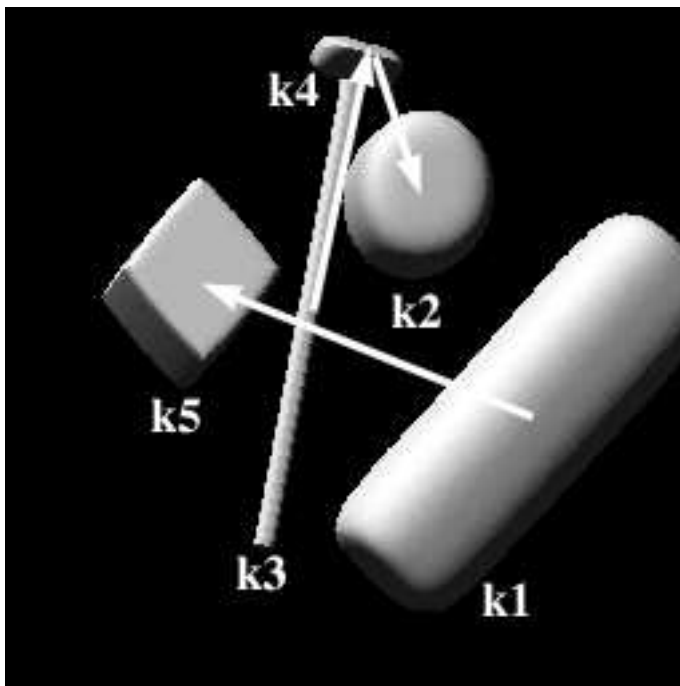


Fig. 24. The superquadric reconstruction of the scene in Fig. 23 along with the focus of attention movements during the exploration of the scene.

---

Knoxe1 (#k1)  
Knoxe1 (#k2)  
Knoxe1 (#k3)  
Knoxe1 (#k4)  
Knoxe1 (#k5)

Thin-cylinder-shaped(#k3)  
Flat-cylinder-shaped(#k4)  
Screw(Screw#1)  
has-peg(Screw#1,#k3)  
has-head(Screw#1,#k4)

Cylinder-shaped(#k2)  
Cylinder(Cylinder#1)  
has-part(Cylinder#1,#k2)

Next-to(Next-to#1)  
participant(Next-to#1,Screw#1)  
participant(Next-to#1,Cylinder#1)

Parallelepiped-shaped(#k1)  
Block(Block#1)  
has-part(Block#1,#k1)

Box-shaped(#k5)  
Block(Block#2)  
has-part(Block#2,#k5)

Next-to(Next-to#2)  
participant(Next-to#2,Block#1)  
participant(Next-to#2,Block#2)

---

Fig. 25. The assertions generated at the linguistic level related to the perception acts represented in Fig. 24.

It is clear that, at the present stage of development, our architecture does not directly address many of the presently unresolved problems of computer vision, although it may well provide a contribution in some of these areas. Typical problems encountered in real vision systems are: non optimal image acquisition conditions, poor contrast, shadows, occlusions between objects, segmentation criteria. Nevertheless, our framework offers interesting hints to face them. For example, the hypothesis generation process at the basis of the focus of attention mechanism can be usefully employed to solve the occlusion problem: the linguistic information and the associative mechanism can provide

interpretation contexts and high level hypotheses that help in interpreting uncomplete structures.

As a matter of fact, occlusions, non optimal image acquisition and segmentation problems can be addressed in a framework in which active vision processes are coupled with our focus of attention mechanism. Symbolic reasoning and attentive processes driven by high level expectations can be essential in orienting low level active processes, in order to acquire new information from the sensors. The fusion of our model in an active vision framework is one of the topics of our future research.

We are presently extending the architecture to the analysis of dynamic scenes. In this case, the subsymbolic level must be able to estimate the motion parameters of the objects in the scene (velocity, acceleration, and so on); the mapping between the conceptual level and the linguistic level must take the dynamic evolution into account; non rigid objects must be recognized in spite of the modifications to their shape. We maintain that the assumptions at the basis of our model can be easily extended to the dynamic scenes. The focus of attention mechanism, and the concept of perception act are by nature dynamic: they introduce a dynamic aspect even into the perception of static scenes. Even more so, they are expected to work in dynamic contexts.

## Acknowledgement

We would like to thank Luigia Carlucci Aiello and Peter Gärdenfors for the interesting discussions about the topics of the paper. Marco Gori, Pino Spinelli and Carmen Usai carefully read and commented on previous versions of this paper. We would also like to thank the anonymous referees for their suggestions which helped improve both the presentation and the contents of the paper.

## References

- [1] P.E. Agre. Computational research on interaction and agency. *Artif. Intell.*, 72:1–52, 1995.
- [2] J. Aloimonos. Visual shape computation. *Proc. IEEE*, 76(8):899–916, 1988.
- [3] D. Amit. *Modeling Brain Function. The World of Attractor Neural Networks*. Cambridge University Press, 1988.
- [4] E. Ardizzone, S. Gaglio, and F. Sorbello. Geometric and conceptual knowledge representation within a generative model of visual perception. *Journal of Intelligent and Robotic Systems*, 2:381–409, 1989.

- [5] R. Bajcsy. Active perception. *Proc. IEEE*, 76(8):996–1005, 1988.
- [6] R. Bajcsy and M. Campos. Active and exploratory perception. *Computer Vision, Graphics and Image Processing: Image Understanding*, 56(1):31–40, 1992.
- [7] D.H. Ballard. Animate vision. *Artif. Intell.*, 48:57–86, 1991.
- [8] A.H. Barr. Superquadrics and angle-preserving transformations. *IEEE Computer Graphics and Applications*, 1:11–23, 1981.
- [9] H.G. Barrow and J.M. Tenenbaum. Computational vision. *Proc. IEEE*, 69(5):572–595, 1981.
- [10] M. Bertero, T. Poggio, and V. Torre. Ill-posed problems in early vision. *Proc. IEEE*, 76(8):889–896, 1988.
- [11] P. Besl and R. Jain. Three-dimensional object recognition. *ACM Comput. Surv.*, 17(1):75–145, 1985.
- [12] I Biederman. Human image understanding: Recent research and a theory. *Computer Vision, Graphics and Image Processing*, 32:29–73, 1985.
- [13] I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94:115–147, 1987.
- [14] T. Bindford. Survey of model-based image analysis systems. *Int. Journal of Robotics Research*, 1(1):18–64, 1982.
- [15] L. Birnbaum, M. Brand, and P. Cooper. Looking for trouble: Using causal semantics to direct focus of attention. In *Proc. ICCV-93*, pages 49–56, Berlin, 1993.
- [16] N. Block. *Imagery*. MIT Press, Cambridge, MA, 1981.
- [17] R.M. Bolle and B.C. Vemuri. On three-dimensional surface reconstruction methods. *IEEE Trans. Pat. Anal. Mach. Intel.*, 13(1):1–13, 1991.
- [18] R.J. Brachman and J.C. Schmoltze. An overview of the KL-ONE knowledge representation system. *Cognitive Science*, 9(2):171–216, 1985.
- [19] R.A. Brooks. Symbolic reasoning among 3D models and 2D images. *Artif. Intell.*, 17:285–348, 1981.
- [20] R.A. Brooks. Model-based 3-D interpretation of 2-D images. *IEEE Trans. Pat. Anal. Mach. Intel.*, 5(2):140–150, 1983.
- [21] P.J. Burt. Smart sensing within a pyramid vision machine. *Proc. of the IEEE*, 76:1006–1015, 1988.
- [22] C. Cherniak. *Minimal Rationality*. MIT Press, Cambridge, MA, 1986.
- [23] C.H. Chien and J.K. Aggarwal. Identification of 3D objects from multiple silhouettes using quadrees/octrees. *Computer Vision, Graphics and Image Processing*, 36:256–273, 1986.

- [24] R. Chin and C. Dyer. Model-based recognition in robot vision. *ACM Comput. Surv.*, 18(1):67–108, 1986.
- [25] D. Davidson. The logical form of action sentences. In N. Rescher, editor, *The Logic of Decision and Action*, pages 81–95. University of Pittsburgh Press, Pittsburgh, PA, 1967.
- [26] S.J. Dickinson, A.P. Pentland, and A. Rosenfeld. From volumes to views: An approach to 3-D object recognition. *Computer Vision, Graphics and Image Processing: Image Understanding*, 55(2):130–154, 1992.
- [27] J. Doyle. Rationality and its roles in reasoning. In *Proc. AAAI-90*, pages 1093–1100, 1990.
- [28] J. Duncan and G. Humphreys. Visual search and stimulus similarity. *Psychological Review*, 96:433–458, 1989.
- [29] M.J.K. Farah, D. Hammond, R. Levine, and R. Calvanio. Visual and spatial mental imagery: Dissociable systems of representation. *Cognitive Psychology*, 20:439–462, 1988.
- [30] S. Gaglio, P.P. Puliafito, M. Paolucci, and P.P. Perotto. Some problems on uncertain knowledge acquisition for rule based systems. *Decision Support Systems*, 4:307–312, 1988.
- [31] S. Gaglio, G. Spinelli, and V. Tagliasco. Visual perception: an outline of a generative theory of information flow organization. *Theoretical Linguistics*, 11(1/2), 1984.
- [32] P. Gärdenfors. A geometric model of concept formation. In S. Ohsuga et al., editor, *Information Modelling and Knowledge Bases III*. IOS Press, Amsterdam, The Netherlands, 1992.
- [33] P. Gärdenfors. Three levels of inductive inference. In D. Prawitz, B. Skyrms, and D. Weststråhl, editors, *Logic, Methodology, and Philosophy of Science IX*. Elsevier Science, Amsterdam, The Netherlands, 1994.
- [34] P. Gärdenfors. Meaning as conceptual structures. Technical Report 40, Lund University Cognitive Studies, Lund, Sweden, 1995.
- [35] A. Gupta and R. Bajcsy. Volumetric segmentation of range images of 3D objects using superquadric models. *Computer Vision, Graphics and Image Processing: Image Understanding*, 58(3):302–326, 1993.
- [36] S. Harnad. The symbol grounding problem. *Physica D*, 42:335–346, 1990.
- [37] G.E. Hinton, J.L. McClelland, and D.E. Rumelhart. Distributed representations. In D.E. Rumelhart and J.L. McClelland, editors, *Parallel Distributed Processing*, volume 1. MIT Press, Cambridge, MA, 1986.
- [38] J.J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proc. Nat. Acad. Sci. USA*, 79:2554–2558, 1982.

- [39] P.N. Johnson-Laird. *Mental Models*. Harvard University Press, Cambridge, MA, 1983.
- [40] J. Jonides. Voluntary versus automatic control over the mind's eye's movement. In J.B. Long and A.D. Baddeley, editors, *Attention and Performance IX*, pages 187–203. Erlbaum, Hillsdale, N.J., 1981.
- [41] D. Kleinfeld. Sequential state generation by model neural networks. *Proc. Nat. Acad. Sci. USA*, 83:9469–9473, 1986.
- [42] D. Kleinfeld and H. Sompolinsky. Associative network models for central pattern generators. In C. Koch and I. Segev, editors, *Methods in Neuronal Modeling*, Bradford Books, pages 195–246. MIT Press, Cambridge, MA, 1989.
- [43] S.M. Kosslyn. *Image and Mind*. Harvard University Press, Cambridge, MA, 1980.
- [44] E. Lang, K.U. Carstensen, and G. Simmons. *Modelling Spatial Knowledge on a Linguistic Basis*, volume 481 of *Lecture Notes in Artificial Intelligence*. Springer Verlag, Berlin, 1991.
- [45] D. Lee. Some computational aspects of low-level computer vision. *Proc. IEEE*, 76(8):890–898, 1988.
- [46] A. Leonardis, F. Solina, and A. Macerl. A direct recovery of superquadric models in range images using Recover-and-Select paradigm. In J.O. Eklundh, editor, *Proc. ECCV-94*, volume 800 of *Lecture Notes in Computer Science*, Berlin, 1994. Springer-Verlag.
- [47] D. Marr. *Vision*. W.H. Freeman and Co., New York, 1982.
- [48] D. Marr and H.K. Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. Lond. B.*, 200:269–294, 1978.
- [49] J. Maver and R. Bajcsy. Occlusions as a guide for planning the next view. *IEEE Trans. Pat. Anal. Mach. Intel.*, 15(5):417–433, 1993.
- [50] B. Nebel. *Reasoning and Revision in Hybrid Representation Systems*, volume 422 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag, Berlin, 1990.
- [51] R. Nevatia and T.O. Bindford. Description and recognition of complex-curved objects. *Artif. Intell.*, 8:77–98, 1977.
- [52] A.P. Pentland. Perceptual organization and the representation of natural form. *Artif. Intell.*, 28:293–331, 1986.
- [53] A.P. Pentland and S. Sclaroff. Closed-form solutions for physically-based modeling and reconstruction. *IEEE Trans. Patt. Anal. Mach. Intell.*, 13(7):715–729, July 1991.
- [54] M.I. Posner. Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32:2–25, 80.

- [55] A.A. Requicha and H.B. Voelcker. Solid modeling: a historical summary and contemporary assessment. *IEEE Comput. Graph. Appl.*, 2(2):9–24, 1982.
- [56] R.D. Rimey and C.M. Brown. Control of selective perception using Bayes nets and decision theory. *International Journal of Computer Vision*, 12(2/3):173–207, 1994.
- [57] F. Solina and R. Bajcsy. Recovery of parametric models from range images: The case for superquadrics with global deformations. *IEEE Trans. Patt. Anal. Mach. Intell.*, 12(2):131–146, 1990.
- [58] M.J. Tarr and S. Pinker. Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, 21(2):233–282, 1989.
- [59] J.M. Tenenbaum, M.A. Fischler, and H.G. Barrow. Scene modeling: A structural basis for image description. *Computer Graphics and Image Processing*, 12:407–425, 1980.
- [60] D. Terzopoulos and D. Metaxas. Dynamic 3D models with local and global deformations: Deformable superquadrics. *IEEE Trans. Patt. Anal. Mach. Intell.*, 13(7):703–714, 1991.
- [61] D. Terzopoulos, A. Watkin, and M. Kass. Constraints on deformable models: Recovering 3D shape and nonrigid motion. *Artif. Intell.*, 36:91–123, 1988.
- [62] P-S. Tsai and M. Shah. Shape from shading using linear approximation. Technical Report CS-TR-92-24, University of Central Florida, Department of Computer Science, Orlando, FL, 1992.
- [63] J.K. Tsotsos. Knowledge organisation and its role in representation and interpretation for time-varying data: the ALVEN system. *Computational Intelligence*, 1:498–514, 1985.
- [64] J.K. Tsotsos, S.M. Culhane, W.Y.K. Wai, Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *Artif. Intell.*, 78:507–545, 1995.
- [65] B. Tversky and K. Hemenway. Objects, parts, and categories. *Journal of Experimental Psychology: General*, 113:169–191, 1984.
- [66] P. Whaite and F. Ferrie. From uncertainty to visual exploration. *IEEE Trans. Patt. Anal. Mach. Intell.*, 13(10):1038–1049, 1991.
- [67] D.L. Yarbus. *Eye Motion and Vision*. Plenum Press, New York, 1967.
- [68] S.W. Zucker. Region growing: Childhood and adolescence. *Computer Graphics and Image Processing*, 5:382–399, 1976.