

A Hybrid Architecture for Shape Reconstruction and Object Recognition

Edoardo Ardizzone, Antonio Chella, Salvatore Gaglio

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

Submitted for publication in "International Journal of Intelligent Systems"

Contacting author:

Antonio Chella
Dipartimento di Ingegneria Elettrica
Universita' di Palermo
Viale delle Scienze
90128 Palermo - Italy
Tel. +39.91.6566273
Fax:+39.91.488452
Email: chella@diepa.unipa.it

A Hybrid Architecture for Shape Reconstruction and Object Recognition

Edoardo Ardizzone, Antonio Chella, Salvatore Gaglio

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

Abstract

The proposed architecture is aimed to recover 3-D shape information from gray-level images of a scene; to build a geometric representation of the scene in terms of geometric primitives; and to reason about the scene. The novelty of the architecture is in fact the integration of different approaches: symbolic reasoning techniques typical of knowledge representation in artificial intelligence, algorithmic capabilities typical of artificial vision schemes, and analogue techniques typical of artificial neural networks. Experimental results obtained by means of an implemented version of the proposed architecture acting on real scene images are reported to illustrate the system capabilities.

1. Introduction

In order to achieve autonomous capabilities of interacting with unknown and unstructured environments, an intelligent system must be equipped with adequate perceptive capabilities and with the ability of building a deep and internal representation of its environment.

Vision is the perceptual modality most largely analysed in the last decades. Since the early works in this field (see e.g. the seminal paper of Marr and Nishihara [Marr & Nishihara 1978]), the role of objects' shape has been recognised as fundamental. So the reconstruction of 3-D shapes from images coming from intensity or range data, has been one of the most investigated fields in computer vision literature.

Shape reconstruction may be described as a low-level process capable to extract meaningful information from sensory data with the weakest possible assumptions about the observed scene. In absence of drastic limitative hypotheses, this may appear to be in contrast with the object recognition activity, which is a higher level process requiring considerable information reduction and abstraction.

Object recognition and classification needs a structured description of reconstructed shapes in such terms that a representation of the perceived scene may be built, which is suitable for reasoning and decision making tasks.

The aim of the present work is the integration of high-level and low-level capabilities within a "hybrid" framework, in which may coexist different approaches: symbolic reasoning techniques typical of knowledge representation in artificial intelligence, algorithmic capabilities typical of artificial vision schemes, and analogue techniques typical of artificial neural networks.

More in details, we hypothesise three representation levels as the basis of our architectural design: the subsymbolic level, in which the information is strictly related to the sensory data; the linguistic level, in which information is expressed by a symbolic language; and an intermediate, "prelinguistic" conceptual level. At this level the information is characterised in terms of a metric space defined by a certain number of "cognitive" dimensions, independent from any specific language [Gärdenfors 1992]. The aim of this level is to generate the very internal representation of the agent's external environment and to provide a precise interpretation for the linguistic level.

The interpretation of linguistic level categories is implemented by a mapping between the conceptual and the linguistic levels in terms of a connectionist device. Neural networks allow to avoid an exhaustive description of conceptual categories at the symbolic level: in a certain sense, prototypes "emerge" from the activity of an associative mechanism

during a training phase based on examples. Moreover, the measure of similarity between a prototype and a given object is implicit in the behaviour of the network and is determined during the learning phases.

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

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

The proposed model must not be considered as a model of human vision: no hypotheses are made about this point, and the model may be referred to an autonomous abstract intelligent system currently under development, in which other components are devoted to the reasoning activities necessary for planning actions, controlling input sensors, coordinating motor activities, and so on.

The paper is structured as follows: Sect. 2 outlines the design of the proposed architecture for artificial vision, Sects. 3 outlines the computational steps from the image to the 3D reconstruction of the scene. Sect. 4 describes the conceptual level of representation, while Sect. 5 and 6 respectively specify the linguistic level and its interpretation function. Sect. 7 particularises in greater details the focus of attention mechanism and Sect. 8 characterises the links between the conceptual and the linguistic level in terms of time-delay attractor neural networks. Finally, Sect. 9 presents the main experimental results obtained with the implemented architecture.

2. The design of the architecture

The cognitive assumptions introduced in the previous section are the guidelines for the design and implementation of an architecture for artificial vision [Chella et al. 1994].

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

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

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

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

Three focus of attention modalities are hypothesised at the basis of the proposed architecture: a reactive modality, in which the attention is driven only by the characteristics of scene, a linguistic modality in which the attention is driven by simple inferences at the linguistic level, and an associative modality in which the attention is driven by free associations among concepts.

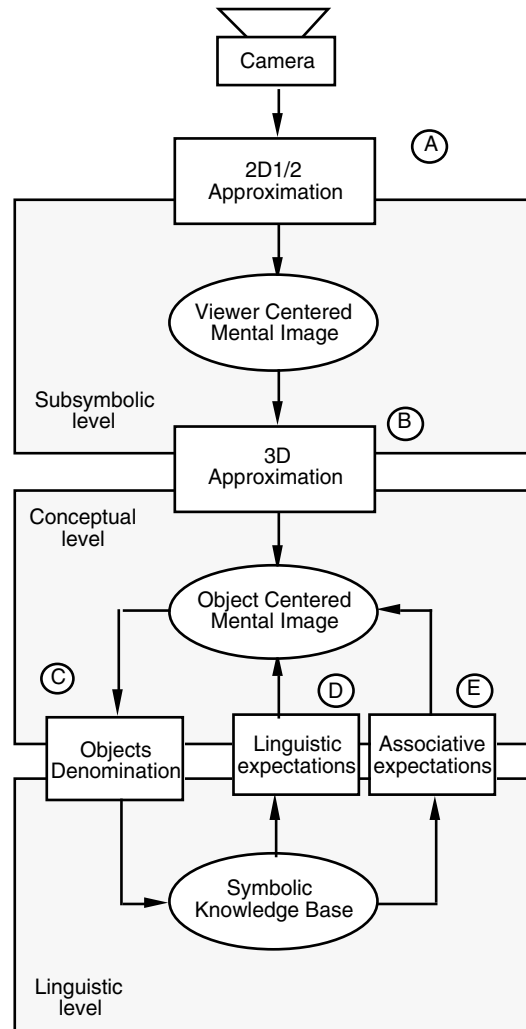


Fig. 1. The proposed architecture. The three levels of representation are evidenced.

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

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

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

3. From the images to the geometric representation

According to the classic approach by Marr and Nishihara, the first step towards the recognition of objects present in a scene requires to derive a 3-D model description from a scene image, through the identification, by information extracted from the image, of the model's coordinate system, of the relative spatial arrangement and sizes of major parts of the model, and of shape components (the *primitives* representing the low level "parts" of the description).

Quite obviously, representations of this kind are strongly influenced by the choice of primitive elements and of composition operators. On the basis of a number of arguments previously reported in the literature, superquadrics have received in the last years a special attention from computer vision scientists. Boolean combinations by regularised operators and deformations of superquadrics are able to represent with surprising realism scenes and objects of the real world.

Considerations of this kind made us to adopt superquadrics as geometric primitives of a 3-D modelling system, falling in the area of CSG modelling systems, in which complex objects can be obtained as recursive, Boolean combinations and deformations of representation primitives.

Even the problem of best fitting superquadrics to single parts on 3-D information (range data or 3-D points extracted from 2-D image data) has been addressed by many authors. A review of reconstruction methods using either superquadrics or other forms of 3-D surface representations may be found e.g. in [Requicha 1980]

Our approach to the primitive recovery from image data directly follows the common notion of "parts" and then takes a straightforward advantage of the CSG nature of the proposed modelling system. Since we are concerned with gray-scale images, early vision methods, like shape-from-x algorithms, can be used to reconstruct the depth information, using one or more 2-D images of the observed scene, in order to build a 3-D shape representation. Shape parts are subsequently approximated by superquadrics. No 3-D shape segmentation is currently attempted: when necessary, 2-D images are directly segmented in order to obtain scene portions to be approximated by single superquadrics. This heavy limitation, that will be removed when effective 3-D segmentation algorithms will be available, may be partially justified by the simplicity of scenes currently considered. In each case, connectivity and other relations among scene parts are taken into account at the conceptual level of the vision system.

In the rest of the paper, experimental results will be shown with reference to the simple real scene, made up by a hammer, a tennis ball and a computer mouse put on a table, shown in Fig. 2. The scene has been acquired as a 256 x 256, 8 bits per pixel image.

We need the extraction from 2-D images of 3-D information adequate to the subsequent fitting with superquadrics. We are not really interested, in this phase of the recovery procedure, to the direct determination of closed form solutions for the surface.

A good, but not necessarily exact estimate of the objects' shape may be sufficient, since the recognition and classification problems are managed at the linguistic level. This may be achieved if depth values of the entire image, i.e. a sort of *dense* depth map, are available.

Robust shape-from-x algorithms are available for this task. Shape-from-shading and shape-from-stereo are among most popular approaches [Horn 1986], even if their performance is normally acceptable only in presence of limitative hypotheses. The most important of these limitations are on the surface form and characteristics, and on the environmental control of illumination sources. Some of these limitations can be at present accepted in our application. Shapes may be supposed convex, as required by the fitting procedure and by the used geometric primitives, and the surface must be sufficiently smooth and homogenous; moreover the environment in which the images are acquired must be sufficiently controlled.

Shape-from-shading algorithms are normally subdivided into two general classes. In *global* methods, the shape is recovered by minimizing some cost function involving constraints such smoothness. The shape is iteratively computed, e.g. using variational techniques [Horn & Brooks, 1989], and it maintains itself globally consistent. In *local* methods, the estimate of the shape is attempted from local variations in image intensity, by

using local constraints about the surface or about the reflectance map. Local methods are more simple and less accurate, while global methods are more complex and more precise.

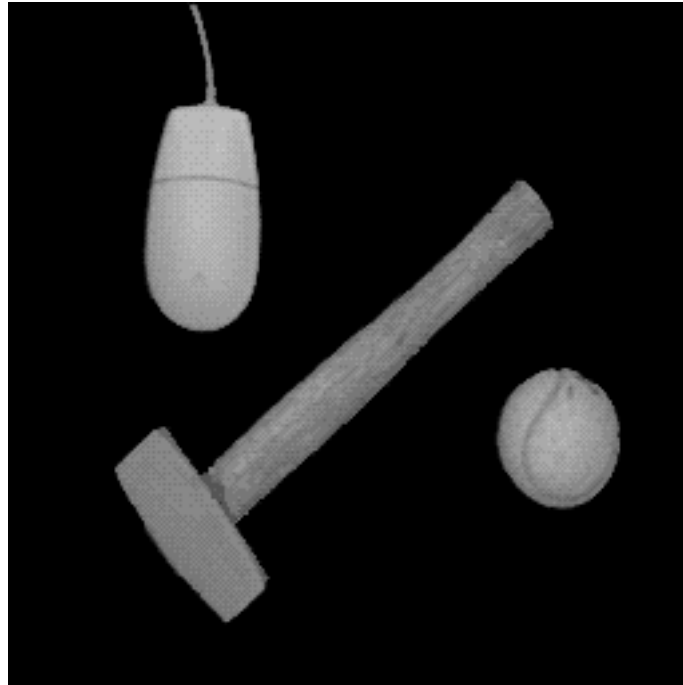


Fig. 2. The scene adopted in the experimental framework.

On the basis of above considerations, we use two methods originally proposed by Pentland, respectively known as *local* shape from shading [Pentland 1989] and *linear* shape from shading [Pentland 1990]. In both cases, the surface is assumed to be locally spherical, and the Lambertian reflectance model is adopted, with constant albedo. In the second method, a linear approximation to the true reflectance function is used. The illuminant direction is computed along with the shape information by the local algorithm, while the linear algorithm requires a separate computation.

Another method, similar to the Pentland's linear algorithm, has been also taken into consideration, i.e. the algorithm proposed by Tsai and Shah [Tsai & Shah 1992]. The major difference lays in the fact that Pentland uses a reflectance map linear in the surface gradient (p,q) , while Tsai and Shah linearize the reflectance in the depth $Z(x,y)$.

These three approaches have been implemented and used on the image shown in Fig. 2. The results are shown in fig. 3, for the Tsai and Shah's method. To give users an adequate impression of the estimated shapes, a spline interpolation, available in our graphic system, has been adopted for the visualisation.

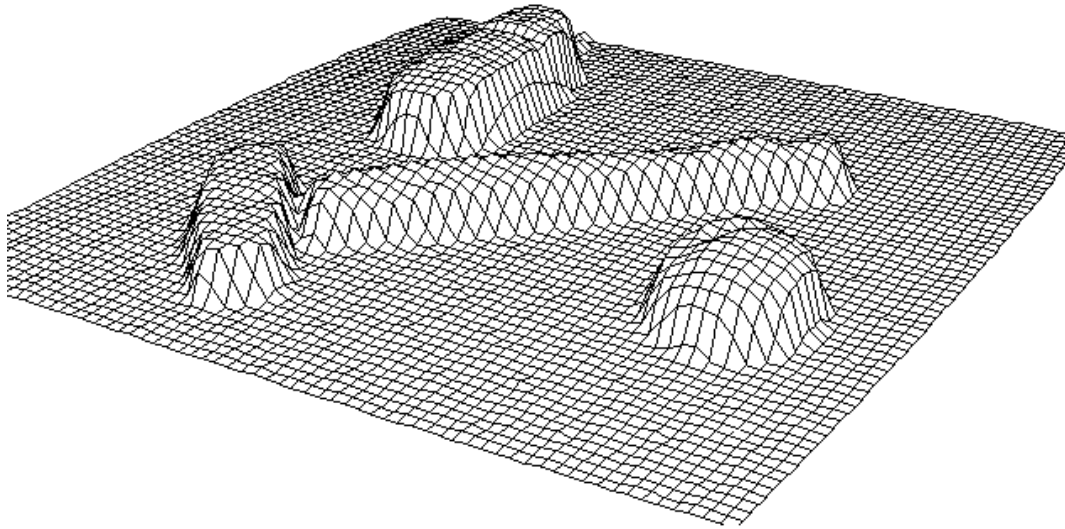


Fig. 3. The shape reconstructed by using the Tsai and Shah's method.

Once the dense depth map of the scene under consideration have been computed, the next step in the visual process is the reconstruction of the geometric model, i.e. the recovery of the superquadrics that best fit the various scene parts. The implemented fitting procedure follows the approach adopted by Solina and Bajcsy [Solina & Bajcsy 1990] and it is based on the minimisation of a suitable error functional. The result of this approach is shown in Sect. 9.

4. The conceptual level of representation

In Marr's theory a superior symbolic level is limited to a hierarchically organised catalogue of 3D prototypes. The mental imagery literature [Glasgow 1993] hypothesises a distinction between mental pictures and propositional mental representations e.g. Kosslyn [Kosslyn 1980] distinguishes between a short term memory based on mental images, and a propositional long term memory. Cognitive evidence exists, according to which both these kinds of representation coexist and are integrated in human memory. [Farah et al. 1988] In Johnson-Laird's theory [Johnson-Laird 1983], three levels of representation are hypothesised, that, in a certain sense, summarise the various points of view above sketched.

From a slightly different point of view, Gärdenfors [Gärdenfors 1992] proposes three levels of information representation: a linguistic level, a conceptual level and a subsymbolic level. At the linguistic level the information is described in terms of a symbolic language, e.g. first order logic; at the subsymbolic level information is characterised directly in terms of the perceptual inputs of the system. Between these two levels, a third level is hypothesised: the conceptual level, in which information is described by means of a conceptual space.

Our model is inspired to Gärdenfors proposal. Many analogies can be found in the Marr model and in many positions emerged from the mental imagery debate.

In fig. 1, the three grey blocks correspond to Gärdenfors' levels of representation. The first level can be seen also as a visual, viewer centered, mental image (or, in Marr's terminology, 2D 1/2 sketch). The central level embeds 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.

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 KS ; the i -th dimension of KS is indicated as X_i . Examples of such dimensions can be colour, pitch, mass, spatial coordinates, and so on. The dimensions should be considered "cognitive" in that they correspond to qualities of the represented environment, without references to any linguistic descriptions. In the case of visual perception, the dimensions of the conceptual space correspond to the parameters of a suitable system of geometric primitives, e.g. simple blocks (as Marr's generalised cylinders), geons [Biederman 1985] and superquadrics [Barr 1981].

We define the *knoxel* as a generic point k in a conceptual space (the term "knoxel" is suggested by the analogy with "pixel"); knoxels therefore represent epistemological primitive elements at the considered level of analysis.

Formally a knoxel is a vector in which each component corresponds to a parameter describing a quality dimension of the domain of interest. In the case of visual perception, these parameters characterise geometric primitives of the kind quoted above. 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 is defined among knoxels, which may be considered as a measure of similarity among knoxels in the conceptual space.

As previously explained, perceived objects and situations correspond to suitable sets of knoxels; we define a perception cluster pc as a finite set of knoxels corresponding to an object or a situation

5. The linguistic level of representation

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

In order to describe the symbolic knowledge base we adopt a hybrid representation formalism, in the sense of Nebel [Nebel 1990]. According to this point of view, a hybrid formalism is constituted by two different modules: a terminological component and an assertional component. In our model, the terminological component contains the descriptions of the concepts relevant for the represented domain (e.g., types of objects and of situations to be perceived). The assertional component stores the assertions describing the specific perceived scenes.

The choice of a hybrid formalism is not constraining. However, the distinction between terminological and assertional components can be useful to keep distinct the conceptual knowledge, which is largely independent from the specific perceived scene, and the assertions concerning the scene itself. In this sense, the terminological component can be seen as analogous to a long term memory, while the assertional component to a short term memory. Moreover, terminological formalisms are well suited for our purposes, in that they are centered on conceptual descriptions. This allows a compact description of concepts, whose instances are to be recognised in the perceived scene.

As an example, consider in fig. 4 a simple fragment of the terminological knowledge base, concerning the description of objects. In the figure, the graphic network notation for the KL-ONE system has been adopted. A brief description of the KL-ONE formalism may be found in Appendix A.

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., *hammer-head* or *hammer-handle*) correspond to two argument relations.

Concerning situations, they are also represented as concepts in the terminological formalism. In other words, we assume that situations are reified, i.e., we assume that to every specific situation correspond an individual in the domain. This gives a great

flexibility and expressive power. Fig. 5 shows the network description of the *Situation* concept, and of two particular types of situation, *On* and *Sided*.

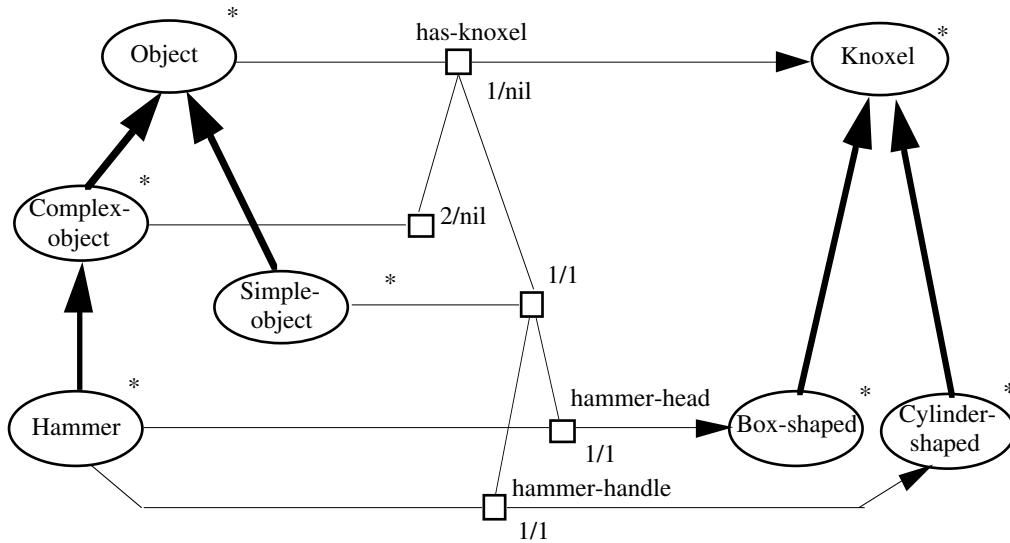


Fig. 4. A fragment of the terminological knowledge base.

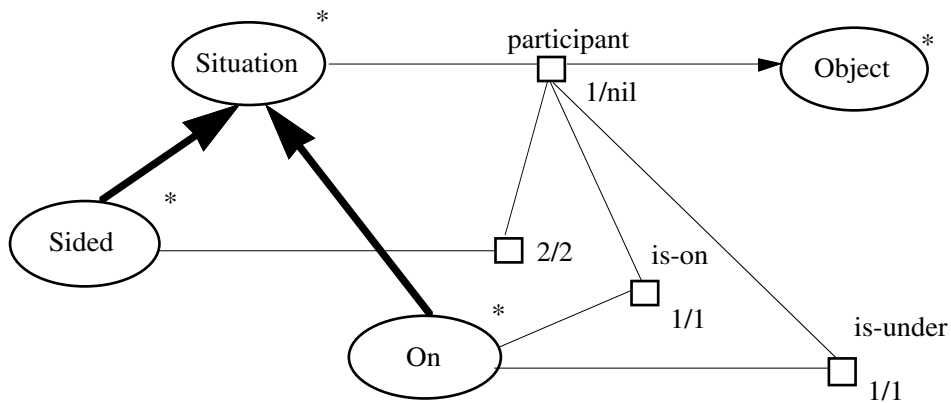


Fig. 5. Network description of the *Situation* concept.

6. The mapping between the conceptual and linguistic levels

The mapping between the conceptual and the linguistic level defines an internal, cognitive oriented, semantic interpretation for the symbols at the linguistic level. The interpretation function associates a perception cluster to any individual constant representing an object or a situation at the linguistic level a perception cluster. Therefore, if C is the set of assertional constants representing objects and situations, the interpretation function φ has the following type:

$$\varphi: C \rightarrow PC \tag{1}$$

where PC represents the set of all perception clusters.

The assertional language can also treat assertional constants, representing more specific features of objects and situations, like, for instance, the main axis of an object or its shape. The set C' of such assertional constants is mapped by the interpretation function to the specific structures of the conceptual space:

$$\varphi: C' \rightarrow X \quad (2)$$

where $X = \bigcup_{i=1}^n X_i$ is the set of all the values of the dimensions of the conceptual space.

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.

The main difference between the proposed internal semantics and the usual model theoretic semantics is that the extensions of predicates are uniquely defined at the conceptual level and, therefore, the truth of atomic assertions in which primitive concepts are involved can be easily determined by specific relations among the conceptual entities.

Consider for example a *participant* role. Given the assertion *participant*(i, j), in a purely extensional model theoretic semantics its truth is justified exclusively by the fact that the pair of the extensions of i and j belongs to the extension of *participant*:

$$\langle \varphi[i], \varphi[j] \rangle \in \varphi[\textit{participant}] \quad (3)$$

In the internal semantics, the truth of *participant*(i, j) can be determined by examining the entities on which i and j are interpreted in the conceptual space: *participant*(i, j) is true if the set of knoxels on which j is interpreted is a subset of the set of knoxels on which i is interpreted:

$$\varphi[j] \subseteq \varphi[i] \quad (4)$$

7. The focus of attention process

As sketched 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.

In modelling perception, these problems can be faced taking into account the fundamental role of attentive phenomena in vision [Yarbus 1967]. In the psychological literature, the focus of attention has been sometimes described as a light spot which scans the visual field, individuating relevant aspects [Posner 1980].

This mechanism is analogous to the scanning of a mental image, as described by Kosslyn²³. Several models of focus of attention mechanisms have been proposed in the active vision literature. The interest in this field has been summarised by Bajcsy [Bajcsy & Campos 1992] which proposes the "active and exploratory" framework for perception.

A well-used strategy to model the focus of attention, followed e.g. by Burt [Burt 1988] is based on the pyramidal approach. Rimey [Rimey & Brown 1994] proposes the TEA-1, a task-oriented system performing the minimum effort necessary to solve a specific task. Birnbaum [Birnbaum 1993] proposes the BUSTER system aimed at developing a causal explanation of the scene.

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 contemporary 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 model the focus of attention mechanism, we define a *perception act* p as a generic sequence of knoxels in the conceptual space: KS^* is the set of all sequences of knoxels (therefore of all perception acts) in the conceptual space KS .

With reference to a perception cluster pc , we say that a perception act p is associated to the perception cluster pc if $p \in pc^*$, where pc^* is the set of all sequences of knoxels belonging to pc . The perception act associated to a perception cluster therefore corresponds to a specific way of perceiving an object or situation described by the perception cluster. It should be noted that the sequence of knoxels making a perception act may not include all the knoxels of the corresponding perception cluster and/or it may include several times the same knoxels.

The description of complex elements in terms of sequences of knoxels seems to be a natural extension of Gärdenfors' notion of conceptual space. It should be noted that the perception act assumption avoids the necessity to augment the dimensions of the space in order to describe complex objects or situations made up by several blocks: they are described by perception acts of several length.

In order to individuate the grouping paths among knoxels and generate the most meaningful perception acts, it is necessary to suitable orient the focus of attention. In the proposed architecture the focus of attention is determined by three concurrent modalities: the reactive, the associative and the linguistic modality. The "reactive" modality is the simpler one: the grouping paths among 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.

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

In the linguistic modality, the focus of attention is driven by the symbolic information explicitly represented at the linguistic level. Consider the hammer example: at the linguistic level (see fig. 4) a hammer is described as composed by a handle and a head. If an object similar to an hammer handle has been recognised, the linguistic level hypothesise that an hammer may be present in the scene. The focus of attention is directed to try to identify its parts, in particular its handle and its head, in order to confirm the presence of such an hammer in the scene. This corresponds to find the suitable fillers for the role parts of the object, i.e. a filler for the *hammer-head* role and a filler for the *hammer-handle* role. Therefore, whenever a hammer handle is recognised, the focus of attention tries to identify a hammer by identifying suitable fillers for its head and its handle.

The focus of attention mechanism may be modelled as an expectation function ψ linking the linguistic to the conceptual level; the function has its domain in the set C of assertional constants representing the expected objects or situation and its range in the set of perception acts belonging to the corresponding perception clusters. 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. The function ψ has the following type:

$$\psi^i: C \rightarrow KS^* \quad i = 1, 2 \quad (5)$$

where KS^* is the set of all perception acts, as stated before; i indicates the attentive modality: 1 stands for the linguistic modality and 2 stands for the associative modality.

8. The neural network implementation of the mapping between the conceptual and the linguistic levels

This section only sketches the neural network implementation of the mapping between conceptual and linguistic level. A more complete exposition of this topic may be found in⁴.

A perception cluster, as previously described, is a set of knoxels associated to a perceived object or situation. Each knoxel may be viewed as a point attractor of a suitable energy function associated to the perception cluster.

A set of fixed point attractors is a good candidate as the model for a perception cluster: starting from an initial state representing a knoxel imposed, for instance, from the external input, the system state trajectory is attracted to the nearest stored knoxel of the perception cluster. Therefore the implementation of perception clusters by means of an attractor neural networks [Hopfield 1982] appears to be natural.

Following this approach the implementation of the perception acts associated to a perception cluster is built by introducing time delayed connections storing the corresponding temporal sequences of knoxels. It should be pointed out that the choice of the time-delayed attractor neural networks is not constraining but 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 an uniform treatment of recognition and generation of perception acts, the denotation functions and the expectation functions introduced in the previous Sect. may be implemented by a uniform neural network architecture design.

In order to describe the dynamics in the conceptual space an adiabatically varying energy landscape E is defined. The energy E is the superimposition of three energies (eq. 6): E_1 represents the fast dynamics for period of duration t and it models the point attractors for the single knoxels belonging to the perception clusters; E_2 represents the slow dynamics for period of duration $t \gg t$ due to time-delayed connections and it models the perceptions acts; E_3 model the global external input to the network.

The global energy function of the time delayed synapses attractor neural network is [Kleinfeld 1986]:

$$E(t) = E_1(t) + \lambda E_2(t) + \varepsilon E_3(t) \quad (6)$$

where E_1 , E_2 , E_3 , are the previously described energy terms; λ and ε are the weighting parameters respectively of the time delayed synapses and the external input synapses.

The expectation functions ψ^i describing blocks D and E of our architecture are implemented by setting of parameters of the energy function E to $\lambda > 1$ and $\varepsilon = 0$. In fact, the task of these blocks is the generation of suitable knoxel sequences representing the expected perception acts.

This choice of parameters allows the transitions occurs "spontaneously" with no external input. Referring to eq. 6, an attractor is stable for a time period significantly long due to the E_1 term. As $\lambda > 1$, the term λE_2 is able to destabilize the attractor and to carry the state of the network toward the successive attractor of the sequence representing the successive knoxel of the stored perception act. The neural network therefore visits in a sequence all the knoxels of the stored perception acts.

The function describing the block C of our architecture of fig. 1 is implemented by setting of parameters of the energy function E to $\lambda < 1$ and $\varepsilon > 0$. The task of the block C is the recognition of input knoxel sequences representing the input perception acts. In order to accomplish this task it is necessary to consider the input term of the energy in order to make the transitions among knoxels happen as driven from the external input.

When $\lambda < 1$ the term λE_2 is not be able to drive itself the state transition among the knoxels of the perception act, but when the term εE_3 is added, the contribution of both terms will make the transition happen. The neural network therefore recognises the input perception act as it "resonates" with one of the perception acts previously stored.

9. Experimental results

In this section some experimental results obtained by the implementation of the described architecture are presented, starting from the real image showed in fig. 2. Fig. 6 shows the 3D approximation of the scene through superquadrics. For the sake of clarity each superquadric has been marked by a tag.

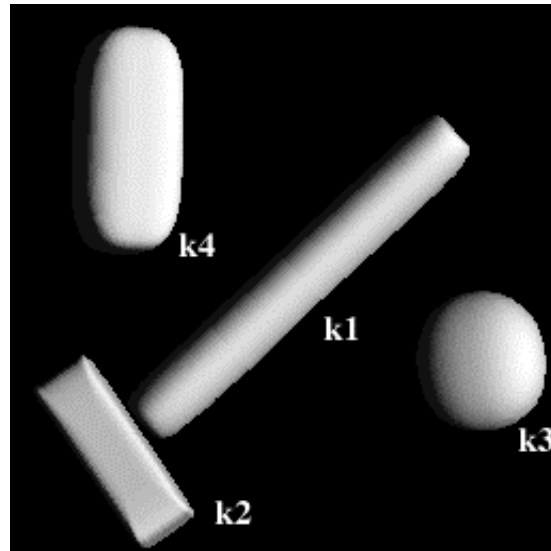


Fig. 6. The approximation of the scene by superquadrics.

When the described architecture is in "reactive" modality the focus of attention searches for generic objects in the scene. In the case of the current scene, the focus of attention is directed to the hammer handle, corresponding to knoxel #k1. The knoxels related to this perception act are sent to the naming block of the architecture in order to find the corresponding linguistic constant at the linguistic level. In this case the knoxel #k1 has been recognised as a *Cylinder_shaped* knoxel.

The generated assertions from the linguistic level describing the operation of the architecture are reported:

```
Knoxel (#k1)
Knoxel (#k2)
Knoxel (#k3)
Knoxel (#k4)
Input_sequence (#k1)
Cylinder_shaped (#k1)
```

The expectation functions suitably drive the focus of attention in order to find the relevant perception acts in the scene. The linguistic expectation function, in particular, generates hypotheses by inferences at the linguistic level. As an example with reference to the previous scene, the linguistic level hypothesises that the cylinder shaped knoxel may be a filler for the role *hammer-handle* of the concept *Hammer*. Therefore the architecture hypothesises the presence of an hammer in the scene; the linguistic expectation block generates perception act hypotheses for the fillers of the role part *hammer-head* and for the filler of the role part *hammer-handle*.

The time-delayed neural networks implementing the linguistic expectation block of the architecture generates the expected perception acts for the hammer-head filler and the hammer-handle filler. When some of these expected knoxels match some corresponding knoxels in the scene, the corresponding perception act is sent to the denomination function

in order to recognise the hammer. The resulting assertions generated at the linguistic level are reported:

```
Linguistic_expectation (Cylinder_shaped, Hammer)
Expected (Hammer, Hammer_head_filler)
Expected (Hammer, Hammer_handle_filler)
Satisfied_by (Hammer_head_filler, #k2)
Satisfied_by (Hammer_handle_filler, #k1)
Hammer (#k1, #k2)
```

The task of the associative expectation function in block E is to suitably drive the focus of attention in order to explore the scene by free associations among concepts. As an example, referring to the previous scene, the concept *Hammer* is associated by a Hebbian mechanism to the concepts of *Ball* and *Mouse*, due to a previous learning phase of the architecture. The linguistic level therefore hypothesises the presence of these objects in the scene and the time-delayed neural network implementing the associative expectation block generate the corresponding perception acts hypotheses.

As in the linguistic modality, when some of these expected knoxels are satisfied by some corresponding knoxels in the scene, the perception act made up by the so found knoxels is sent to the denomination function. The corresponding generated assertion by the linguistic level are:

```
Associative_expectation (Hammer, Ball)
Associative_expectation (Hammer, Mouse)
Expected (Hammer, Ball)
Expected (Hammer, Mouse)
Satisfied_by (Ball, #k3)
Satisfied_by (Mouse, #k4)
Ball (#k3)
Mouse (#k4)
```

The recognition task of perception acts related to spatial relation is similar to the recognition task of objects; the generated assertions related to spatial relations referred to the previous scene are shown. Fig. 7 shows the resulting perception act of the architecture after the previously described operations of the focus of attention.

```
Up (Hammer, Ball)
Sided (Ball, Mouse)
```

Fig. 8a shows a complex scene made up by a hammer, a cordless telephone, a wood block and a mouse. Fig. 8b shows the superquadric reconstruction of the same scene along with the focus of attention movements through the scene exploration. It should be noted that the focus of attention 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 scene is therefore analysed as a concatenation of these two sequences.

The focus of attention mechanism allows in fact the creation of "attentional contexts" in which an object is analysed. During the analysis of the first sequence, the telephone has been ignored, because the object does not belongs to the current attentional context. The same during the second sequence: the block, the hammer and the mouse are ignored because they does not belong to the same attentional context of the telephone.

This allows the system to avoid the "cognitive overload" problem. The system is able to find out the relevant paths, to aggregate the information, in order to generate only the linguistic descriptions "useful" and "interesting" for the system in the current attentional context.

10. Conclusions

Some open questions arise from the operation of the proposed architecture. One of them has been already mentioned and it is related to the lack of capability of an effective 3-D scene segmentation. The segmentation into meaningful parts is currently made up by directly acting on the input images; more complicated input scenes require most effective segmentation tools. To this aim, methods based on the application to 3-D shapes of morphological tools like erosion and dilation seems to be promising.

Another open question is a truly adequate best fitting procedure, as far as the superquadric recovery is concerned. Neither common minimization techniques nor optimization procedures of the kind illustrated in the paper seem to be completely reliable, due to the local minima problems. Possible solutions are based on the direct, one-shot, estimate of superquadric parameters by a search over a rich catalogue of prototypical superquadrics.

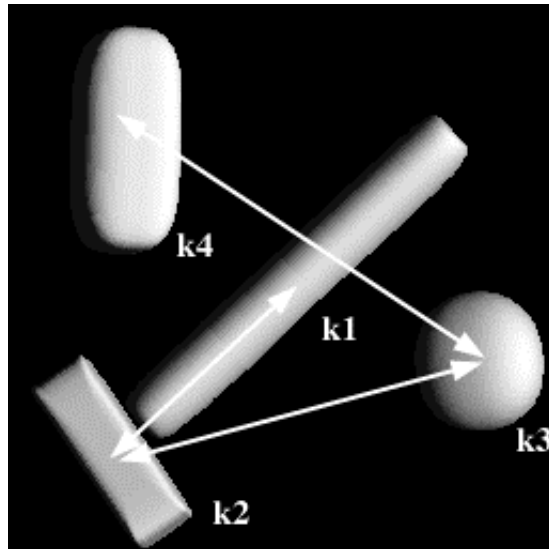


Fig. 7. The resulting perception act related to previous scene when the focus of attentions is driven by the linguistic and associative expectations.

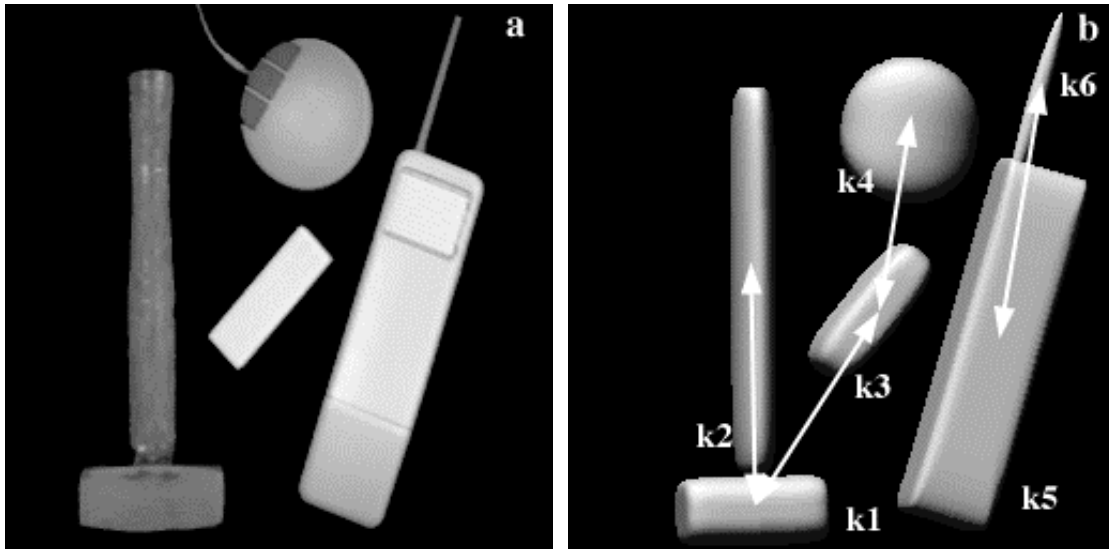


Fig. 8. A complex scene made up by a hammer, a cordless telephone, a wood block and a mouse. a) The acquired scene. b) The superquadric reconstruction along with the focus of attention operation.

Appendix A: A Brief Description of the KL-ONE System

The principal elements of the KL-ONE system [Nebel 1990] descriptions are concepts; the most important type of concept is the generic concept. Potentially many individuals in any possible world can be described by a generic concept; so *Object*, *Hammer*, *Screw*, etc. are all generic concepts each of which are descriptions that could be used to describe individuals in the world. Some of these generic concepts may be formed out of the others: e.g. a *Hammer* is a *Complex-Object*. KL-ONE separates its descriptions into two basic groups: primitive and defined. Primitive concepts are those for which necessary and sufficient conditions are not given in terms of other concepts; they act as incomplete descriptions. Defined concepts are instead derived from other concepts. In the KL-ONE descriptions the concepts are depicted with ellipses; the concept with an asterisk are primitive, while the others are defined (see e.g. fig. 4).

Several structure-forming operations are available for building concepts. They bring together one or more general concepts and a set of restrictions on the concepts. More in details, the components of a Concept are its subsuming concepts and its local internal structure expressed in roles and structural descriptions. Roles describe potential relationships between instances of the concept and those of other associated concepts (e.g. its parts and properties). Structural descriptions express the interrelations among the roles; they are not used in our architecture.

Roles of a concept are taken as a set of restrictions applied to its subsuming concepts. Because a concept is defined in terms of the subsuming concepts, all the parents restrictions must apply to the children. In order to achieve this effect, KL-ONE provides inheritance facilities: e.g. in fig. 4 *Hammer* inherits all the component restrictions of *Complex-Object*.

A role acts as a generalised attribute description, representing potential relationships between individuals of the type denoted by the concept and other individuals. Generic rolesets are the most important type of Roles. They capture the notion that a given functional role of a concept can be played by several different entities for one individual. In KL-ONE representations rolesets are depicted by squares, connected by unnamed links to the concept of which they are components.

The structure of the rolesets is specified with value restriction (V/R) describing necessary type restrictions on rolesets fillers, and with value number restrictions, expressing cardinality information as a pair of numbers, defining a range of cardinality for sets of role-player descriptions (the nil specification means an infinite upper bound).

Roleset restriction allows to describe how the roleset components of a concept can be specified in terms of rolesets belonging to the subsuming concepts. A roleset restriction does not specify a new role but it adds constraints on the filler of a role with respect to a specified concept.

Roleset differentiation allows to specify the sub-roles to fill with subsets of the fillers of the roles they differentiate.

References

- [Ardizzone et al. 1989] E. Ardizzone, S. Gaglio, F. Sorbello, Geometric and Conceptual Knowledge Representation within a Generative Model of Visual Perception, *Journal of Intelligent and Robotic Systems*, Vol. 2, 1989, pp. 381-409.
- [Bajcsy & Campos 1992] R. Bajcsy, M. Campos, Active and Exploratory Perception, *Computer Vision, Graphics and Image Processing: Image Understanding*, Vol. 56, No. 1, 1992, pp. 31-40.
- [Barr 1981] A.H. Barr, Superquadrics and Angle-Preserving Transformations, *IEEE Computer Graphics and Applications*, Vol. 1, 1981, pp. 11-23.
- [Biederman 1985] I. Biederman, Human Image Understanding: Recent Research and a Theory, *Computer Vision, Graphics and Image Processing*, Vol. 32, 1985, pp. 29-73.
- [Birnbaum et al. 1993] L. Birnbaum, M. Brand, P. Cooper, Looking for Trouble: Using Causal Semantics to Direct Focus of Attention, *Proc. ICCV-93*, 1993, pp. 49-56.
- [Bolle & Vemuri 1991] R.M. Bolle, B.C. Vemuri, On three-dimensional surface reconstruction methods, *IEEE Trans. Patt. Anal. Mach. Intell.*, vol.13, 1991, pp. 1-13.
- [Burt 1988] P.J. Burt, Smart Sensing within a Pyramid Vision Machine, *Proceedings of the IEEE*, Vo. 76, No. 8, 1988.
- [Chella et al. 1994] A. Chella, M. Frixione, S. Gaglio: A Cognitive Architecture for Artificial Vision, CS&AI Lab Tech. Rep, University of Palermo, 1994.
- [Dickinson 1992] S.J. Dickinson, A.P. Pentland, A. Rozenfeld, 3-D Shape Recovery Using Distributed Aspect Matching, *IEEE Trans. Patt. Anal. Mach. Intell.*, vol.14, 1992, pp. 174-198.
- [Farah et al. 1988] M.J.K. Farah, D. Hammond, R. Levine, R. Calvanio, Visual and Spatial Mental Imagery: Dissociable Systems of Representation, *Cognitive Psychology*, Vol. 20, 1988, pp. 439-462.
- [Gärdenfors 1992] P. Gärdenfors, Three Levels of Inductive Inference, Lund University Cognitive Studies No. 9, Tech. Rep. LUHFDA/HFKO-5006-SE, 1992.

- [Glasgow 1993] J.I. Glasgow, The Imagery Debate Revisited: A Computational Perspective, *Computational Intelligence*, Vol. 9, No. 4, 1993.
- [Hopfield 1982] J.J. Hopfield, Neural Networks and Physical Systems with Emergent Collective Computational Abilities, *Proceedings of the National Academy of Sciences, USA*, Vol. 79, 1982, pp. 2554-2558.
- [Horn & Brooks 1989] B.K.P. Horn, M.J. Brooks, The variational approach to shape from shading, in B.K.P Horn & M.J. Brooks (Eds.):*Shape from Shading*, The MIT Press, 1989, pp. 173-214 .
- [Horn 1986] B.K.P. Horn, *Robot Vision*, 1986, The MIT Press, Cambridge, Massachusetts.
- [Johnson-Laird 1983] P.N. Johnson-Laird, *Mental Models*. Harvard University Press, 1983, Cambridge, Massachusetts.
- [Kleinfeld 1986] D. Kleinfeld, Sequential State Generation by Model Neural Networks, *Proc. Nat. Acad. Sci. USA* Vol. 83, 1986, pp. 9469-9473.
- [Kosslyn 1980] S.M. Kosslyn: *Image and Mind*. Harvard University Press, Cambridge, MA, 1980.
- [Marr & Nishihara 1978] D. Marr, K. Nishihara, Representation and recognition of the spatial organization of three-dimensional shapes, *Proc. Roy. Soc. London B*, vol. 200, 1978, pp. 269-294.
- [Nebel 1990] B. Nebel, *Reasoning and Revision in Hybrid Representation Systems*, LNAI 422, Springer-Verlag, Berlin, 1990.
- [Pentland 1986] A.P. Pentland, Perceptual Organization and the Representation of Natural Form, *Artificial Intelligence*, Vol. 28, 1986, pp. 293-331.
- [Pentland 1987] A.P. Pentland, Recognition by parts, *Proc. of IEEE 1st Int. Conf. on Comp. Vision*, June 1987, pp. 612-620.
- [Pentland 1989] A.P. Pentland, Local shading analysis, in B.K.P Horn & M.J. Brooks (Eds.):*Shape from Shading*, The MIT Press, 1989, pp. 443-487.
- [Pentland 1990] A.P. Pentland, Linear Shape from Shading, *Int. Journ. of Computer Vision*, Vol. 4, 1990, pp. 153-160.
- [Posner 1980] M.I. Posner, Orienting of attention, *Quarterly Journal of Experimental Psychology*, Vol. 32, 1980, pp. 2-25.
- [Putnam & Subrahmanyam 1986] L.K. Putnam, P. Subrahmanyam, Boolean Operations on n-Dimensional Objects, *IEEE Computer Graphics and Applications*, June 1986, pp. 43-51.
- [Requicha 1980] A.A.G. Requicha, Representations for rigid solids: theory, methods and systems, *ACM Comp. Surveys*, Vol. 12, No. 4, 1980, pp. 437-464.

- [Rimey & Brown 1994] R.D. Rimey, C.M. Brown, Control of Selective Perception Using Bayes Nets and Decision Theory, *International Journal of Computer Vision*, Vol. 12, No. 2/3, 1994, pp. 173-207.
- [Solina & Bajcsy 1990] F. Solina, R. Bajcsy, Recovery of Parametric Models from Range Images: The Case of Superquadrics with Global Deformations, *IEEE Trans. on PAMI*, Vol. 12, 1990, pp. 131-147.
- [Terzopolous & Metaxas 1991] D. Terzopolous, D. Metaxas, Dynamic 3D Models with Local and Global Deformations: Deformable Superquadrics, *IEEE Trans. Patt. Anal. Mach. Intell.*, Vol.13, 1991, pp. 703-714.
- [Tilove 1980] R.B. Tilove, Set Membership Classification: A Unified Approach to Geometric Intersection Problem, *IEEE Trans. Computers*, Vol. 29, 1980, pp. 874-883.
- [Tsai & Shah 1992] P.S. Tsai, M. Shah, A Simple Shape from Shading Algorithm, Techn. Rep. CS-TR-92-24, 1992, Univ. of Central Florida, Orlando, FL.
- [Yarbus 1967] D.L. Yarbus, *Eye Motion and Vision*. Plenum Press, New York, 1967.