A Situation-Bayes View of Object Recognition Based on SymGeons

Fiora Pirri and Massimo Romano
Alcor Group
Dipartimento di Informatica e Sistemistica
Universitá di Roma "La Sapienza"
via Salaria 113, 00198, Roma, Italy {pirri,romano@dis.uniroma1.it

Abstract

We describe in this paper a high level recognition system. The system implements a new approach to model-based object recognition, fully exploiting compositionality of representations : from the analysis of the elementary signs in the image to the analysis and description of an object structure and, finally, to the interpretation of the scene. Perceptual reasoning, likewise the symbolic description of the scene are stated in the Situation Calculus. A description is a specification of an object in terms of its single components which, in turn, are specified using SymGeons, a generalization of parametric Geons. The cognitive process of recognition relies on SymGeons recognition. Here we extend the concepts of aspect graphs and hierarchical aspect graph to obtain a Bayes network integrating composition of aspects together with composition of features.

Introduction

In this paper we present the basic ideas governing the high level recognition system of Mr. ArmHandOne, a tiny mobile manipulator endowed with a pan-tilt binocular head equipped with two color cameras PC100XS from Supercircuits. We have implemented (MIL libraries and C++) all the image processing leading to the Syntactic Analyzer (see item 4 below), the Bayes-Aspect Network for SymGeons recognition (C++ and Eclipse-Prolog), and the high-level descriptions of few simple objects. We have not yet defined the Bayes-network for object descriptions, that would allow us to have a probability distribution on the presence of given objects in the scene, and we are still working on binocular reconstruction of SymGeons. The system has been tested over simple artifacts, although part of it has not yet been implemented. Its architecture is structured as follows:

1. Cognitive level. At this level we represent perception as the reasoning component of the recognition process. Perception is formalized in the Situation Calculus (Pirri & Finzi 1999), and it is essential for solving any dichotomy between what is perceived and what is inferred by an agent using *a priori* knowledge and premises. There are several cognitive problems in connection with the reasoning process concerning perception and knowledge (see e.g. (Reiter 2001a)), requiring a continuous confrontation between an agent inner world representation and sensed/perceived information, like e.g. the anchoring problem[1].
2. Description level. It provides objects and scene descriptions in terms of *categories*. Here a category is a kind gathering the simplest representation of a certain object, using SymGeons as primitive components. E.g. a table can be described as follows: a flat top with one or more vertical supports. The flat top is either a cylindroid or cuboid etc., and it is orthogonally connected with the supports, that could be either cylindroid or cuboid and if there is more than one support each is parallel to the others... So far we have not considered physical or functional specifications in connection with descriptions although they would be very useful.
3. Recognition level. At this level recognition is concerned only with primitive components of the image. Each shape in the image is classified as an aspect/view of a SymGeon, and is given a probability, according to the traits of the shape itself. Classification is done by a Bayes net integrating a hierarchical aspect graph (see (Dickinson & Metaxas 1994)): the two DAGs are fused into an *Aspect-Bayes net*, that is, a Bayes-net in which causal relations between nodes are enriched with compositional relations concerning aspects.
4. Syntactic analyzer. This level is concerned with the construction of a labeled graph of the image. The labeled graph is defined in FOL. The syntactic image analysis delivers a set of segments that we call *primits*. This set forms a graph that we call the *syntactic graph*.
5. Image processing, which is achieved with standard methodologies.

In synthesis our approach is based on the following idea. It is not possible to separate the recognition process from the presence or absence of the knowledge of a specific environment. For the unknown elements of a scene a crucial role is played by the following *abilities*:

[1] Anchoring is the process of creating and maintaining the correspondence between symbols and sensors data that refer to the same physical object (Coradeschi & Saffiotti 2000)

1. *Reference ability*. The ability to conform (map) the current environment to one already known, individuating distinguishing features.

2. *Abstraction ability*. The ability to provide a context free representation of an artifact which is described by its components, drawn from a small set of primitives, and bound together by relations specified in a combinatorial language respecting the principle of compositionality.

3. *Reasoning ability*. The ability to draw hypotheses that can be confirmed: "it could be a hat, a pot, or a lamp shade, and since it is on a lamp it must be a lamp shade".

Generally speaking, the reasoning process taking place in recognition is based on previous (not necessarily accurate) knowledge of the environment from which the current scene is taken. In order to arrange knowledge of objects or scenes into patterns we have to exploit the inner structure of both human artifacts and environments, and this naturally leads to a compositional and constructive representation. This way of thinking is also known as visual perceptual organization which was explored by (Pentland 1986) in his seminal paper on perception. Pentland, in fact, pointed out that perception is possible (likewise intelligent prediction and planning) because of an internal structuring of our environment and because of the human ability to identify the connections between these environmental regularities and primitive elements of cognition. All the model-based approaches to perception have been influenced by this view and the use of knowledge-based models of the world has become a standard paradigm in vision (see e.g. (Edelman 1997; Wu & Levin 1993; Pope & Lowe 1993; Chella, Frixione, & Gaglio 2000; Rock 1997)). Among the model-based approaches, the constructive approach, known as recognition by components (RBC), was pioneered by (Marr & Nishihara 1978), who introduced generalized cylinder, by (Shapiro *et al.* 1984) who use stick, blobs and plates, by (Pentland 1986) who first introduced the use of superquadrics (Barr 1981) in cognitive vision, by (Biederman 1987) who introduced the novel idea of *Geons* (geometric icons), and finally by (Wu & Levin 1993) who modeled *parametric Geons* combining Biederman's ideas with superquadrics.

Following the above line of research on visual perception, we develop our formalization on the recent approach of (Patel 2000) who introduces *SymGeons*, that are an extension of the above mentioned parametric Geons. The advantage of SymGeons over parametric Geons is that, by loosing the symmetry property which we do not need to preserve in our approach, they can be used as coarse descriptions also of asymmetrical objects (e.g. the well known snicker example that was Ullman's point against recognition by components (Ullman 1996)). On the other hand SymGeons have several views, and views are composed of faces which, in turn, are composed of simpler elements as coarse descriptions of primitives geometric signs depicted in the image. This earlier compositional construction is obtained using an Aspect-Bayes network, which plays a role similar to the aspect graph, but here causal relationships are enriched with compositional relationships.

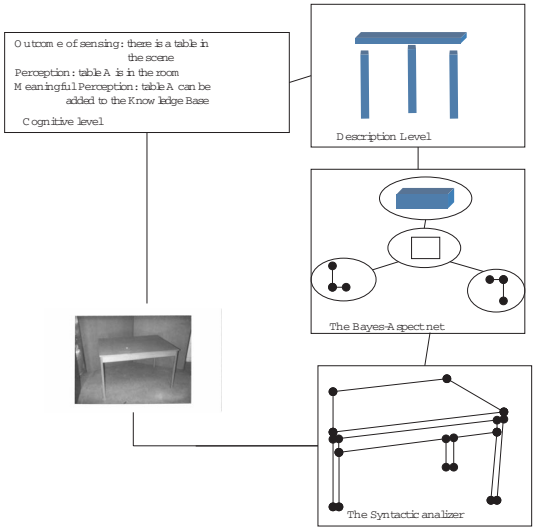


Figure 1: The reasoning process behind perception

Preliminaries

The paper exploits several techniques to model perceptual reasoning. In particular for the scene and object description and for the high level perceptual reasoning, the Situation Calculus is used at all the levels of reasoning for the purpose of formalizing definitions, and also for simple geometric primitives. We consider the Situation Calculus (SC) ((McCarthy & Hayes 1969; Reiter 2001b)) as the core of our logic and language, and we suitably extend it to include new sorts, new symbols for its alphabet and new axioms. We refer the reader to (Pirri & Reiter 1999) for a full presentation of the core logic and language and to (Pirri & Finzi 1999) for more details related to extensions concerning perception. For reasoning on the elementary geometric structure of the image, i.e. at the lines, faces, and aspects levels, both probabilistic and causal reasoning are exploited: hypotheses are drawn about the geometric structure of the image, and these hypotheses are used by perceptual reasoning. However the core of the whole perceptual architecture and the link between the two levels of reasoning, from the image to the scene, are the SymGeons.

The concept of *SymGeon*, introduced by (Patel 2000), is part of a families of related concepts, that have been used in visual recognition, to describe the shape of an object in terms of a relatively few generic components, joined by spatial relationships. SymGeons (which we can consider as a simple generalization of parametric Geons introduced by Kenong Wu and Martine Levine (Wu & Levin 1993)) have their origins, in *qualitative Geons* (Biederman 1987) and in the computer graphic concept of *Superquadrics* (Barr 1981). The Biederman original Geons are 36 volumetric component shapes described in terms of the following qualitative attributes of generalized cylinders: *symmetry, size, edge, axis*: each of these properties can be suitably varied in so determining a unique Geon. Superquadrics were first introduced in computer graphics by Barr in his seminal paper

(Barr 1981). Petel and Holt (Patel 2000) extended the concept of the parametric Geons of Wu and Levine considering the possibility to apply the *tapering* and *bending* transformations at the same time. In such a way they eliminated the intrinsic symmetry of the parametric Geons allowing to model a larger number of asymmetrical objects.

In the rest of the paper we use $\mathcal{G}(\overline{a}, \overline{\epsilon}, \overline{K}, \kappa)$ to denote a generic SymGeon, where $\overline{a} = (a_1, a_2, a_3)$ is the scale vector, $\overline{\epsilon} = (\epsilon_1, \epsilon_2)$ is the squareness vector, $\overline{K} = (K_x, K_y)$ is the tapering vector and κ is the bending parameter. To refer to the coordinates of a SymGeon in the scene we shall use a term pos, so the position of a specific SymGeon $\mathcal{G}(\overline{a}, \overline{\epsilon}, \overline{K}, \kappa)$, with $\gamma = \langle \overline{a}, \overline{\epsilon}, \overline{K}, \kappa \rangle$ will be denoted by $g(\gamma, pos)$ (see (Wu & Levine 1993)).

A classification of SymGeons is given in Figure 2.

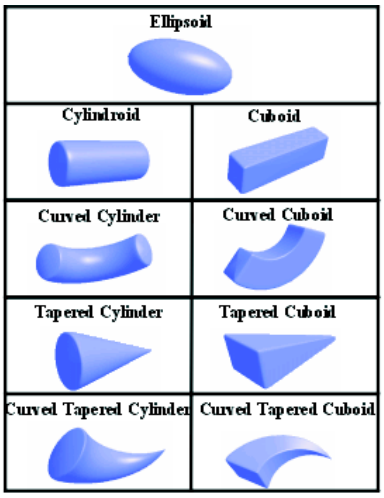


Figure 2: A classification of SymGeons

Finally, we refer the reader to Judea Pearl's Book (Pearl 1988) for an accurate analysis of Bayesian networks.

Cognitive and Description levels

The cognitive part of perception is described and axiomatized in the Situation Calculus. We refer the reader to (Pirri & Finzi 1999) for a full presentation of Perception at the cognitive level. Here we recall some essential features introduced in (Pirri & Finzi 1999). Asserting a "perception" about something that can be perceived, i.e. a perceptible, is denoted by a predicate $percPe(p, v, s)$, in which p is the perceptible of the kind $isT(x)$, $v \in \{0, 1\}$ is the outcome of a sensing action of the kind $sense(isT(x), 1)$, and s is a term of sort situation. For each fluent $F(\vec{x}, s)$ in the language, which is *observable*, i.e. it can be sensed, a perceptible $isF(\vec{x})$ is introduced in the language, together with a successor state axiom of the kind:

$$percPe(isF(\vec{x}), val(s)) \equiv \Psi_{isF}(\vec{x}, s) \quad (1)$$

Here $\Psi_{isF}(\vec{x}, s)$ is a sentence describing what should be true both in terms of other previous percepts and in terms of properties holding in the domain, to make perception hold about the perceptible isF in the situation in which a sensing action has taken place. Obviously there is a frame problem also for $percPe$ and we adopt Reiter's solution (see (Reiter 2001b) for a full and very well detailed description). Each sensing action, in turn, has a precondition axiom of the kind:

$$sPo \quad (sense(isF(\vec{x}), o), s) \equiv \Pi_{isF}(\vec{x}, s)$$

Observe that a successor state axioms like the one for $Percept$ does not interfere with the successor state axioms for fluents, which means that we have two threads of transformations: an inner thread (the agent inner world) traced by the history of actions, through $percPe$ $cept$ plus the perceptible isF, and an outer thread (the agent's external world) traced by the history of actions, through the fluent F. This two threads can be convergent or divergent. If they converge, what is perceived can be added to the database, although the addition is non monotonic, since it is added as an hypothesis. If they diverge, nothing can be added and the agent records a *mistake*. A mistake is not an inconsistency, which can never occur through sensing and percept. This reasoning process is called *meaningful perception*. Inference, according to meaningful perception, is done using regression, so if $\mathcal{D}$ is the theory of action and perception, and

$$\begin{aligned}
&\mathcal{D} \models \\
&percPe(isT(a), 1, s) \wedge percPe(isO(a,), 1, s) \\
&percPe(isC(c), 1, s) \wedge s = [sense(isT(a), 1), .. \\
&\quad sense(isO(a,), 1), .., sense(isC(c), 1), ..]s_0 \wedge \\
&\forall p.p = isT(a) \vee p = isC(c) \\
&\vee p = isO(a,) \rightarrow \neg Mistaken(p, s)
\end{aligned}$$

then meaningful perception would allow the fluents $Table(a, S_0)$, $isC(c, S_0)$, and $isOn(a, , S_0)$ to be added to the initial database; if the history s were mentioning actions other than sensing actions, then meaningful perception would allow the regression of the above fluents to be added to the initial database $\mathcal{D}_{S_0}$.

Now, the problem of perception consists essentially in managing the input data obtained by sensors (e.g. the camera), processing them and suitably adding the results to the theory of action and perception as hypotheses so that the following will hold:

$$\mathcal{D} \cup \mathcal{H} \models \exists percept(p, 1, s) \wedge \exists x.p = isT(x)$$

To understand what $\mathcal{H}$ is and the role of sensing actions, consider the following simple example. There is a table and a chair in a room and an *observation of the scene* is performed, i.e. a shot of the scene is taken (we open our eyes and look into the room); let us cut the instant before we make sense out of what there is in the room. Clearly, at the very moment in which the image is taken no distinction among objects is made. Therefore it is not a single sensing action like $sense(isT(x), v)$ that takes place, but a scene/image acquisition.

From the image acquisition till the inference, leading to an hypothesis that there might be a table in the room, a complex process of revelation takes place. One bringing the shapeless and scattered components identified in the image, to the surface of cognition[2], by giving a structure to

[2]Re-cognition, indeed, means knowing again, to reveal again to cognition.

these components. And there is a step in the structuring that reveals the meaning: "that's a table". In other words the re-cognition process is a thread of revelations (the apperception) giving, attributing, meaning to the elements of the image. This is achieved by conjugating the construction of a tight data structure (a graph of all the SymGeons occurring in the scene together with their topological relations), which is the hypothesis $\mathcal{H}$, together with the meaning given by a description and denoted by a sensing action like $sense(isATable(x), v)$. Therefore the $sense(isATable(x), v)$ action has, indeed, the double meaning of *giving sense* to the elements of the data structure and of bringing to the surface of cognition the existence of an object, a table, in fact.

To understand what we mean let's go through the example of the table. We might simplify the successor state axiom in (1) as follows:

$$\begin{aligned} &typicalPercept(isATable(x), do(a, s)) \equiv \\ &\quad a = sense(isATable(x), v) \wedge InRoom(s) \vee \\ &\quad a \neq sense(isATable(x), v) \wedge Percept(isATable(x), s)). \end{aligned} \tag{2}$$

Now, we constrain the sensing action so that it can be performed just in case a data structure has been released by the re-cognition process. To this end let us distinguish between the object to be localized (the table) and the data structure accounting for the reference primitives of the scene, i.e. the set of all elementary objects appearing in the image, including the legs and the top of the table.

$$\begin{aligned} &possP(sense(isATable(x), v), s) \equiv \\ &\quad (v = 1 \wedge \exists reference. Scene(isATable(x), reference)) \vee \\ &\quad (v = 0 \wedge \forall reference. \neg Scene(isATable(x), reference)). \end{aligned} \tag{3}$$

The above action precondition axiom says that the outcome of sensing is 1 if the description matches over the data structure, the reference, otherwise it will be 0.

We introduce in the following the data structure matching with the term $reference$. First of all observe that we expect to recover, from the image, a set of SymGeons. This process is just hinted in this paper. Once a SymGeon has been recovered, via a Bayes-Aspect Network, see Figure 11, its 3D structure and its localization in the image (depth, height width, etc..) are also returned together with its relationships with the other SymGeons recovered. This process leads to the construction of a *scene graph*, which is totally connected. That is, the graph mentions all the SymGeons recovered in the scene and their relationships: the relationship among each SymGeon and all the others. The set of relations that label the edges of the graph are depicted in Table 1. Although their simplicity, these five relation allow us to describe a large number of complex man made objects, for example in Figure 3 an hammer and a table is represented using connection, orthogonal and parallel relationship between SymGeons. Each relation is suitably defined in terms of a distance and an angle, between two SymGeons. We omit these definitions here.

Therefore the *scene graph* is a triple $\langle S, E, T, \rangle$, where S is the set of the nodes each labelled by a SymGeon recovered in the image, E is the set of $n(n-1)/2$ edges, if $n = |S|$, and T is the set of labels. A label $t \in T$ is defined according

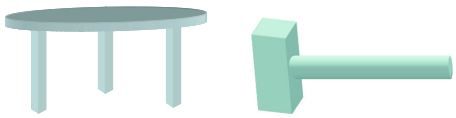


Figure 3: Representation of simple man made objects

to the algebra of connection $\mathcal{AC}$, which is a sestuple

$$\mathcal{AC} = \langle \mathbb{P}, T, \{\{\oplus^i_X\}_{i \leq n}\}_{X \in \mathcal{R}}, \langle \mathcal{R}, \prec \rangle, +, \prec \rangle$$

Here $\mathbb{P} = \{\mathcal{P}, \mathcal{B}, \mathcal{F}, \mathcal{A}, \mathcal{G}\}$ is the set of elementary shapes forming our features taxonomization. $\mathcal{P}$ is the set of $primits$ (Figure 8), $\mathcal{B}$ is the set of $boundits$ (Figure 5), $\mathcal{F}$ is the set of $faces$ (Figure 4), $\mathcal{A}$ is the set of $aspects$, and $\mathcal{G}$ is the set of $SymGeons$ (Figure 2). T is the set of terms of the

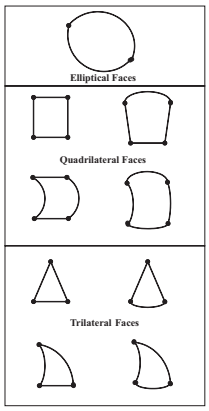


Figure 4: Faces

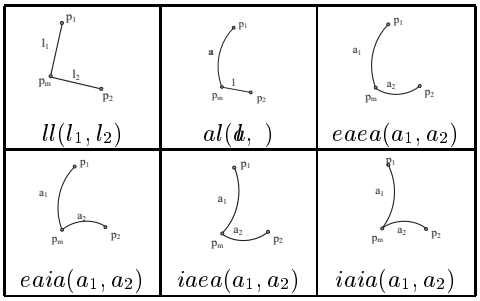


Figure 5: Boundary traits (boundits)

algebra, $\mathcal{R} = \langle \mathcal{R}, \prec, \mathcal{V} \rangle$ is the set of relations reported in Table 1, $\prec$ is a relation over the set of terms T. $+$ is a concatenation operation on T and $\{\{\oplus^i_X\}_{i \leq n}\}_{X \in \mathcal{R}}$ is a family of $n+1$-ary connection operators on T. For each $X \in \mathcal{R}$, there is a set of $\{\oplus^i_X\}_{i \leq n}$, over T, one for each $i \leq n$, and n bounded to the number of SymGeons recovered from the scene. An example of *scene graph* representing a table is depicted in Figure (7).

The set of terms T of the algebra can be further decomposed into a set of *connection terms* $T_\oplus$ and a set of *concatenation terms* T_+. The rules of formation are the following:

1. If $g \in S$, with $S \in \mathbb{P}$ then $g \in T$.

2. If $t, t' \in T$ then $t + t' \in T_+$. $t + t'$ is also denoted by tt'.
3. If $\tau \in T_+$, $\tau = t_1 \cdots t_n$, $g \in T$, and each $t_i \in T$, $1 \leq i \leq n$, then $g \oplus_X^n \tau \in T_\oplus$, for any $X \in \mathcal{R}$. In this case we call g the *head* and τ the *tail* of t.
4. Nothing else is a term of $\mathcal{AC}$.

When in $\oplus_X^n$, $n = 1$ we shall simply write $\oplus_X$, and when X is any relation in $\mathcal{R}$ then we shall omit it. Precedence between the two operators is: $\oplus, +$. E.g. $g \oplus_X^k t_1 + g' \oplus_Y^m t_2 = (g \oplus_X^k t_1) + (g' \oplus_Y^m t_2)$. Substitution of terms for variables is defined as usual.

For a term $t \in T$ we can give a notion of *length* (denoted as $|t|$) defined inductively on its structure. In fact if $t = g$ or $t \in T_\oplus$ then $|t| = 1$, otherwise if $t \in T_+$ we can write $t = t_1 + t_2$ where $t_1, t_2 \in T$ and we have $|t| = |t_1| + |t_2|$ if $t_1 \neq t_2$, or $|t| = |t_i|, i \in \{1, 2\}$ if $t_1 = t_2$.

Example 1 Let $t = g \oplus_P^k t_1 + \cdots + (t_k \oplus_V^2 (g_1 + g_2)) + (g' \oplus_C^2 (t_1' + t_2'))$. Then g is the head in $g \oplus_P^k t_1 + \cdots + (t_k \oplus_V^2 (g_1 + g_2))$ and $t_1 + \cdots + (t_k \oplus_V^2 (g_1 + g_2))$ is the tail. The length of t is $|t| = 2$.

We shall make use of the relation $\in$ to mention a term t occurring in a term t', e.g. $g \in g_1 \oplus_V^2 g_2 + g$. Let $g, g' \in S$, $S \in \mathbb{P}$, let t denote a term according to the above definition of term of $\mathcal{AC}$ and let $\preceq = \prec \vee =$, with $=$ the equality already defined in the language of SC:

Axioms 0.1

0.	$t + t = t$.
1.	$t + t' = t' + t$.
2.	$g \oplus g' = g' \oplus g$.
3.	$g \oplus^0 t = g$.
4.	$g \oplus^{n+1} t_1 \cdots t_{n+1} = g \oplus^n t_1 \cdots t_n + g \oplus t_{n+1}$.
5.	(distributive law 1) $(g \oplus_X^n \tau) + (g \oplus_Y^m \tau') = g \oplus_X^n \oplus_Y^m (\tau)(\tau')$.
6.	(distributive law 2) $(g \oplus_X^m \tau) + (g \oplus_Y^m \tau) = g(\oplus_X \oplus_Y)^m \tau$.
7.	(connection) $t \oplus t' + t' \oplus t'' = t \oplus (t' \oplus t'')$.
8.	$g \oplus_X t = g' \oplus_X t' \equiv g = g' \wedge t = t'$.
9.	$t_1 + t' = t_2 + t'' \equiv t_1 = t_2 \wedge t' = t''$.
10.	$g \not\prec t$.
11.	$\tau \prec t \ \mathrm{op}\ \tau' \equiv \tau \preceq \tau'$, $\mathrm{op} \in \{\oplus, +\}$.
12.	(transitivity of $\prec$) $t \prec t' \wedge t' \prec t'' \rightarrow t \prec t''$.

Other derived properties of $\mathcal{AC}$, whose proofs is omitted here, are:

a. $+$ is reflexive for all $t \in T$;
b. $\oplus$ is reflexive for all $g \in S$, $S \in \mathbb{P}, \forall S$;
c. $g \oplus t_1 + \ldots + g \oplus t_{k+1} = g \oplus^{k+1} (t_1 \cdots t_{k+1})$;
d. $(g \oplus^n \tau) + (g \oplus^m \tau') = g \oplus^k \tau\tau', k = |\tau\tau'|$.

We show, now, how to represent a *scene graph* using a suitable *tree*. First of all we introduce the definition of *principal node*. Given a term t, it has a principal node g, with $g \in Y, Y \in \mathbb{P}$, if $t \in T_\oplus$, i.e. $t = g \oplus_X^{|\tau|} \tau$. We can proof that given a term $t \in T$ of $\mathcal{AC}$ denoting a connected graph, there exist a term t', s.t. $t' = t$ and t' has a principal node.

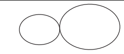
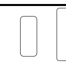
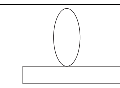
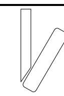
C= Connected: $\oplus_C^n$	
P= Parallel: $\oplus_P^n$	
S= Symmetric: $\oplus_S^n$	
T= Orthogonal: $\oplus_T^n$	
V= Angular: $\oplus_V^n$	

Table 1: Relations between SymGeons in the scene and their functional denotation. All relations are reflexive.

Example 2 Consider the completely connected graph depicted in Figure 6(i), labeled by some relation in $\mathcal{R}$. The term denoting the graph is the following: then τ can be transformed into a term τ' having a principal node, as follows:

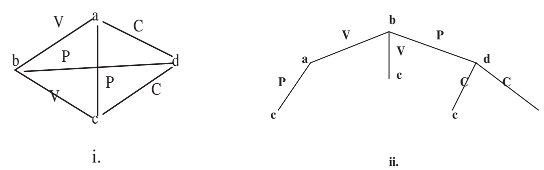


Figure 6:

The above term can be seen as one denoting a tree, e.g. the one depicted in Figure 6 (ii).

This transformation is used in the notion of description. First of all suppose that the graph referencing the scene is the one given in Figure 7. Let $G_1(\vec{a}, e, \vec{K}, k)$ be the cylindroid representing the top of the table, for suitable values of $\vec{a}, \vec{e}, \vec{K}$, and k, and let $G_2(\vec{a'}, \vec{e'}, \vec{K'}, k')$ be the tapered cuboid representing the legs of the table, for suitable values of $\vec{a'}, \vec{e'}, \vec{K'}$, and k'. Let $\gamma_1 = \langle \vec{a}, e, \vec{K}, k \rangle$ and $\gamma_2 = \langle \vec{a'}, \vec{e'}, \vec{K'}, k' \rangle$. Finally let [illegible] denote the position *pos* of each SymGeon (by stereopsis). The term describing the graph is the following:

$$
\begin{array}{l}
(g_1(a, \ ) \oplus_C \oplus_T g_2(b, \ ') + g_1(a, \ ) \oplus_C \oplus_T g_2(c, \ ') + \\
g_1(a, \ ) \oplus_C \oplus_T g_2(d, \ ') + g_2(b, \ ') \oplus_P g_2(c, \ ') + \\
\quad g_2(b, \ ') \oplus_P g_2(d, \ ') + g_2(d, \ ') \oplus_P g_2(c, \ '))
\end{array}
$$

This can be transformed into the following term, with a principal node.

$$
\begin{array}{l}
(g_1(a, \ )(\oplus_C \oplus_T)^3 \\
\quad (g_2(b, \ ') \oplus_P^2 (g_2(c, \ ') \oplus_P^2 g_2(d, \ '))(g_2(d, \ '))) \\
\quad\quad (g_2(c, \ ')) \\
\quad\quad\quad (g_2(d, \ ')))
\end{array} \quad (4)
$$

The term represents the transformation of the graph depicted in Figure 7 into a tree. Now the point is, how do we in-

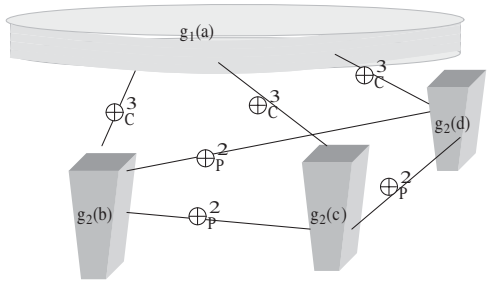


Figure 7: The graph of the scene representing a table, as the relations between SymGeons identified in the image

fer from the above term that it is a term denoting a *table*? This is, in fact, achieved by parsing the term denoting the scene graph, looking for the term denoting a table. Parsing is activated by a description which, in turn, is activated by a sensing action; as we mentioned above, a sensing action is a query to a data structure, namely the graph of the scene.

We shall now use the term $reference$, accounting for the data structure, to define a description. When the inference matches the term $isPo(sense(isIT(x), 1), s)$, a description of the table is asked for. In other words, the description is the *semantics* of a term of the algebra, while its syntax is the structure of the object: e.g. the graph of the table (see 4) is the syntactic structure of a table, the meaning of the table is given through its description (see the next axiom (5). Matching is achieved by a $parse$ function. The $parse$ function is a call to a parsing algorithm that analyzes the term while transforming it by applying the rewriting rules introduced in (0.1).

$$\begin{array}{l} Scene(isIT(x), reference) \equiv \\ \exists t\, \exists top\, \exists supp\, (t = top \oplus_C \oplus_T supp) \wedge \\ (t \in parse(reference)) \wedge \\ \forall l\, l' (l \in parse(supp) \wedge l' \in parse(supp) \wedge \\ \quad l \neq l' \wedge top \oplus_T \oplus_C l \in parse(reference) \wedge \\ \quad top \oplus_T \oplus_C l' \in parse(reference) \wedge \\ \quad l \oplus_P l' \in parse(supp) \wedge \\ \quad \exists p \exists \gamma (top = g(p, \gamma) \wedge \varphi(\gamma, \_) \wedge x = p) \wedge \\ \quad \forall l \exists p' \exists \gamma' (l \in parse(supp) \rightarrow l = g(p', \gamma') \wedge \\ \quad\quad \varphi'(\gamma', p')). \end{array} \tag{5}$$

Here $\varphi(\gamma, \_)$ and $\varphi'(\gamma', p')$ denote the possible SymGeons (e.g. cylindroid, cuboid, etc.) that, respectively, can represent a top and a leg, and in this last case also the possible number of legs. Observe, also, that $x = p$, in the eighth line of the description, is the anchoring of the perceptible $isIT(x)$ to the object, described by the term, through its principal node, which in this case is the top of the table. For each description, in fact, we shall find a major component identifying the term denoting the object we are describing.

Let $t \in T_\oplus$ be the term given by the description $Scene(isF(\vec{x}), reference)$. The $parse$ algorithm verifies if t matches the term $reference$, performing a depth search over $reference$, and performing the following steps:

1. $t = head(\oplus_{X_1} \cdots \oplus_{X_k})^{|tail|} tail, \quad X_1 \cdots X_k \in \mathcal{R}$.
2. If no match for $head$ is found in $reference$ exit with failure. Otherwise:
3. If a match for $head$ is found in $reference$, then:
   (a) If a match for $head(\oplus_{X_1} \cdots \oplus_{X_k})^{|tail|}$ in $reference$ is found, then let $t' = tail$, and go to item (1) with $t = t'$. Otherwise let $t = head(\oplus_{Y_1} \cdots \oplus_{Y_m})^{|tail|} tail$ be the current, where $\oplus_{Y_1} \cdots \oplus_{Y_m}$ is rearranged w.r.t. $\oplus_{X_1} \cdots \oplus_{X_k}$, and $m \leq k$:
   (b) If for some i, $1 \leq i \leq m$, and for some q, a match for $head \oplus^q_{Y_i}$ in $reference$ is found, then:
      i. If there is only one term $reference'$ matching $head \oplus^q_{Y_i}$, and for no other sub-terms $reference''$ of $reference$ there is a match for $head \oplus^h_{Y_j}$, $j \neq i$, then exit with failure. Otherwise:
      ii. rearrange the set of sub-term $reference_i$, $1 \leq i \leq w$, of the term $reference$, matching $head \oplus^q_{Y_i}$, according to the rewriting rules provided, so that $head$ is the principal node of $rearranged(reference)$; let $reference = rearranged(reference)$ and go to (a).
4. Exit with success.

Observe that a match is always reflexive, e.g. $g \oplus g'$ matches with $g' \oplus g$. It is easy to see that the algorithm deals with a rearrangement of a sub-term of the initial term reference until only one term can be rearranged, and therefore it terminates. We have still to prove its completeness: if a description can be matched then it will.

Syntactic Image Analysis

The purpose of the Syntactic Image Analysis is to recognize, in the acquired image, the syntactical categories which are the primitives of the scene structure. The results of such analysis is an *image syntactic graph* which is a graph recording all the elementary traits of the image. Each node of the graph represents a junction point between segments and it is labeled by the information of its 2D positions in the image. Each edge of the syntactic graph represents a segment and is labeled accordingly (straight line or arc of ellipse). At the earlier steps of the analysis we use just classical techniques. In fact, for the execution of some basic operations like filtering and edge detection we use the Matrox Library (MIL), and a convolution based edge detection operator is applied to the image, after filtration. After that a classical edge following algorithm, suitably adapted to filter isolated points or short segments, is used to generate an ordered list of connected points (chain pixels). The resulting edge elements are further approximated with straight lines and ellipse arcs which turn out to be the primitive elements of our taxonomization. We call the set of these primitive elements *primits* i.e. *primitive traits*. A primit, see Figure 8, is a primitive trait defined using the parameters p_1,p_2,α, $C(p_{cu}, p_{cv})$,r_m and r_M, defined in the *image reference frame* $(w, \_)$ as depicted in Figure 9.

The set of primits, independently of any other image features (as color or texture) constitutes the pure syntactic structure of perception. In fact, from this set we obtain the description of the *image structure graph*, $\langle \mathbf{P}, \mathcal{G} \rangle$. Here $\mathbf{P}$ is a

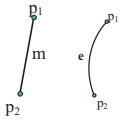


Figure 8: Primitive traits (primits)

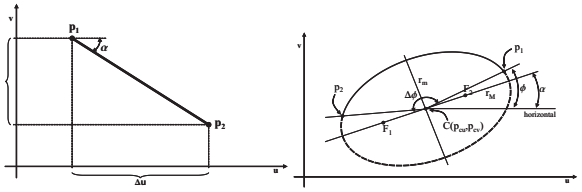


Figure 9: Ellipse and line parameters

set of primits recovered in the image and tCo is a *coterminination* relation between primits defined as follows:

$$tCo\ (g_1, g_2) \equiv_{def} \exists p_1, p_2, p_3, p_4 . primit(g_1, p_1, p_2) \wedge primit(g_2, p_3, p_4) \wedge pwc(g_1, g_2, d) \wedge d < \epsilon$$

Here pwc is a *point wise connection* relation between primits, defined as follows:

$$\begin{aligned} &pwc(g_1, g_2, d) \equiv \\ &\quad primit(g_1, p_1, p_2) \wedge primit(g_2, p_3, p_4) \wedge \\ &\quad d_1 = min_{i,j}\delta(p_i, p_j) \wedge d_2 = \delta(p_h, p_k) \wedge \\ &\quad (p_i = p_1 \vee p_i = p_2) \wedge (p_h = p_1 \vee p_h = p_2) \wedge \\ &\quad (p_j = p_3 \vee p_j = p_4) \wedge (p_k = p_3 \vee p_k = p_4) \wedge \\ &\quad (p_i \neq p_h) \wedge (p_j \neq p_k) \wedge d = \frac{d_1}{d_2} \end{aligned}$$

Here $\delta(\cdot)$ is the Euclidean distance between points:

$$\delta(p_1(u_1, v_1), p_2(u_2, v_2)) = \sqrt{(u_2 - u_1)^2 + (v_2 - v_1)^2}$$

Two primits are coterminant if they have a common end point, i.e. a point whose position is in the range of a given distance.

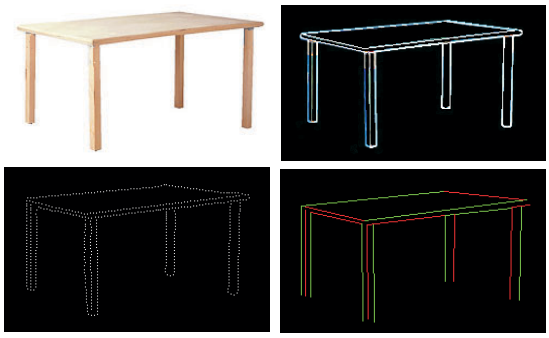


Figure 10: The syntactic analysis of a table

In the sequence of Figure 10 the resulting output of the above steps, applied to the image of a table, is shown.

From HAG to BAN: Hypotheses Formation

In this section we roughly describe the formation of hypotheses drawing the existence of a SymGeon in the image. In the previous Section we have described the primitive components of the image, i.e. the $primits$. By suitably composing primits we form boundits, and by suitably composing boundits we form faces, and finally composing faces we form aspects, where aspects are *views* from different vantage points of a SymGeon. This composition process is formulated in FOL, by explicit definitions. However, due to the image noise and the uncertainty of its structure, the presence of a SymGeon in the scene is gathered using Bayesian inference. E.g. a given closed region of the Syntactic Graph is a cylindroid with probability p.

To this end we construct a Bayes-Aspect Network (BAN), integrating the structure of a HAG (Dickinson & Metaxas 1994) together with causal relations specified by the connection relations $r \in \mathcal{R}$, defined in the algebra $\mathcal{AC}$. The basic BAN topology is obtained by a deterministic construction procedure. Such a procedure, starting from the set of nine SymGeon primitives N_{SP}, extracts the set of aspects N_A, characterizing each primitive $s \in N_{SP}$, the set of faces N_F, describing each aspect $a \in N_A$, and finally, the set of boundits N_B, composing each face $f \in N_F$. N_P is composed of two element n_l and n_a, representing the two kind of primits.

Between each level of the basic BAN, we introduce a connection level. Each node of a connection level is labeled with a relation $r \in R$, defined in $\mathcal{AC}$ (see Figure 11). The conditional probability tables (CPT) linking nodes of a lower level to the nodes of the upper level are defined, in the case of $r \in R$, according to the distance of the features composing the nodes. E.g. the $\oplus_C$ compositional operator CPT, of the portion of the BAN represented in Figure 11, is given below (Table 2) where $\mathcal{N}_{\mu,\sigma}$ is the gaussian distribution with mean

b_1	b_2	$\mathcal{P}(\oplus_C \mid b_1, b_2)$
T	T	$\mathcal{N}_{0,\sigma}(b_1 \oplus_C b_2)$
T	F	0
F	T	0
F	F	0

Table 2: CPT for connection nodes.

μ and variance σ. The CPT for a feature node is suitable defined case by case. Below (Table 3) is shown a feature node's CPT, representing a quadrilateral face with curved boundaries, depicted on the right-hand side feature node in Figure 11. The inference process for recognising a SymGeon, with a given probability, is the following. First of all the root nodes n_l and n_a of the BAN are decorated with the

$\oplus_C$	$\oplus_P$	$\mathcal{P}(\bowtie \mid \oplus_C, \oplus_P)$
T	T	1
T	F	0.5
F	T	0
F	F	0

Table 3: CPT for feature nodes.

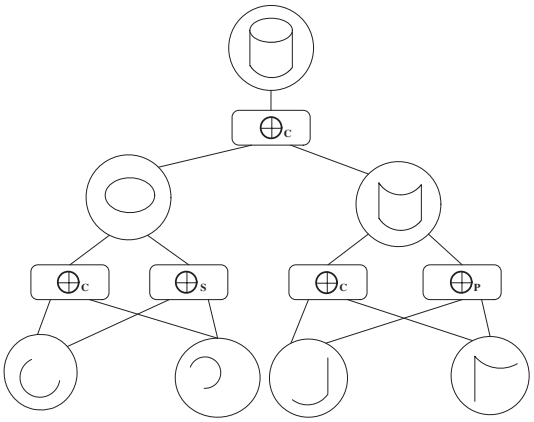


Figure 11: A portion of the Aspect-Bayes.

position of each primits in **P**, with a probability of 1. Once the network is initialized its root nodes constitutes the network *evidence*.

The inference consists in querying the BAN about the probability that, in the image Syntactic Graph, a specific aspect of a SymGeon is mentioned, given the evidence. The query is as follows:

$$\begin{array}{l} \exists p \exists x_1 \cdots x_n aspect(x_1, \cdots, x_n) \wedge \\ \quad bel Pr \quad (aspect(x_1 \cdots x_n) | x_1 \cdots x_n) = p \end{array} \tag{6}$$

It is easy to see that the required inference is double:

1. The first inference requires to construct the terms $x_1 \cdots x_n$ such that each x_i will be a term in $T_{\oplus}$, e.g. $x_i = p_1(\vec{p}) \oplus_C \oplus_S p_2(\vec{p'}) p_3(\vec{p''})$, mentioning only primits, here $\vec{p}, \vec{p'}$ and $\vec{p''}$ denote the set of values characterizing primits. This can be achieved because each aspect is defined in terms of faces and connections, and each face is defined in terms of boundits and connections.
2. The second inference is possible just in case the first returns the set of terms defined by the primits. It is a classical diagnostic inference, requiring to compute the composed probabilities of the paths leading from the specific aspect node to the evidence nodes constituted by the primits composing the specified query.

Once all the primits of the Syntactic Graph have been selected then a list of SymGeons, with their associated probability, is delivered. Obviously the SymGeons with highest probability are chosen, in such a way that they constitute a cover w.r.t. the primits occurring in the Syntactic Graph.

In conclusion we have been trying to define a perceptual architecture that fully exploits compositionality of representation: from the analysis of the elementary signs in the image to the analysis and description of an object structure, and finally to the interpretation of a scene. To achieve this we have been defining a reasoning process that draws suitable hypotheses about each primitive occurring in the observed scene. Hypotheses are evaluated according to their probability and the most probable one is added to the knowledge base for interpreting the scene.

References

Barr, A. 1981. Superquadrics and angle-preserving transformations.

Biederman, I. 1987. Recognition by components - a theory of human image understanding. *Psychological Review* 94(2):115–147.

Chella, A.; Frixione, M.; and Gaglio, S. 2000. Understanding dynamic scenes. *AI* 123(1-2):89–132.

Coradeschi, S., and Saffiotti, A. 2000. Anchoring symbols to sensor data: Preliminary report. In *AAAI/IAAI*, 129–135.

Dickinson, S., and Metaxas, D. 1994. Integrating qualitative and quantitative shape recovery. *IJCV* 13(3):311–330.

Edelman, S. 1997. Computational theories of object recognition. 296–304.

Marr, D., and Nishihara, H. 1978. Representation and recognition of the spatial organization of three-dimensional shapes. In *Proc. R. Soc. Lond. B, vol. 200*, 269–294.

McCarthy, J., and Hayes, P. 1969. Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence* 4:463–502.

Patel, L. N. & Holt, P. O. 2000. Modelling visual complexity using geometric primitives. Orlando: Proceedings, Systemics, Cybernetics and Informatics.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems*. Los Altos, California: Morgan Kaufmann.

Pentland, A. 1986. Perceptual organization and the representation of natural form. *Artificial Intelligence* 28(2):293–331.

Pirri, F., and Finzi, A. 1999. An approach to perception in theory of actions: Part I. *ETAI* 4:19–61.

Pirri, F., and Reiter, R. 1999. Some contributions to the metatheory of the situation calculus. *ACM* 46(3):325–361.

Pope, A., and Lowe, D. 1993. Learning object recognition models from images. In *ICCV93*, 296–301.

Reiter, R. 2001a. On knowledge-based programming with sensing in the situation calculus. *ACM Transactions on Computational Logic (TOCL)* 2(4):433–457. To appear.

Reiter, R. 2001b. *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems*. MIT press.

Rock, I. 1997. *Indirect perception*. MIT-Press.

Shanahan, M. 2002. *A Logical Account of Perception Incorporating Feedback and Explanation*. Proceedings of the Eighth International Conference (KR2002).

Shapiro, L. G.; Moriarty, J. D.; Haralick, R. M.; and Mulgaonkar, P. G. 1984. Matching three-dimensional objects using a relational paradigm. *Pattern Recognition* 17(4):385–405.

Ullman, S. 1996. High-level vision: Object recognition and visual cognition.

Wu, K., and Levin, M. 1993. 3D object representation using parametric geons. Technical Report CIM-93-13, CIM.