

Visual Knowledge Representation of Moving Scenes

A. Chella*, M. Frixione[†], S. Gaglio*

(*) - Dipartimento di Ingegneria Automatica e Informatica
University of Palermo, and CERE-CNR,
Viale delle Scienze, I-90128 Palermo, Italy
e.mail: {chella, gaglio}@unipa.it

([†]) - Dipartimento di Scienze della Comunicazione
University of Salerno,
Via Ponte Don Melillo, I-84084 Fisciano (Salerno), Italy
e.mail: frix@dist.unige.it

Abstract

A framework for the representation of visual knowledge, with particular attention to the analysis and the representation of scenes with moving persons and objects is described. Our aim is to integrate in a principled way the models developed within the artificial vision community, and the propositional knowledge representation systems developed within symbolic AI. We claim that this integration requires the introduction of a missing link between these two classes of representations. In our model, such a role is played by the notion of *conceptual space* (CS), a representation in which information is characterised in terms of a metric space. We illustrate our proposal by referring to an experimental set-up, based on a vision system that takes in input real images of moving people.

Key Words: Perception; Artificial vision; Knowledge representation; Conceptual spaces; Action representation; Hybrid models

1. INTRODUCTION

We propose a theoretical framework for the representation of visual knowledge at different levels of abstraction. The phrase “visual knowledge” must be intended in the sense of knowledge extracted from visual data. The aim of our framework is to integrate in a principled way the approaches developed within the artificial vision community, and the propositional systems developed within symbolic knowledge representation (KR) in AI. In this paper we present our proposal with special reference to the problem of the representation of dynamic scenes, i.e., scenes with moving objects and/or people in them.

On the one hand, the computer vision community approached the problem of the representation of dynamic scenes mainly in terms of the construction of 3D models, and of the recovery of suitable motion parameters, possibly in the presence of noise and occlusions (Marr and Vaina 1982, O'Rourke and Badler 1980, Hogg 1984, Rohr 1994, Pentland and Horowitz 1991).

On the other hand, the KR community developed rich and expressive systems for representation of time, of actions and, in general, of dynamic situations (Allen 1984, McDermott 1982, McCarthy and Hayes 1969, Shoham 1988, Israel et al.1991).

Nevertheless, these two traditions evolved separately and concentrated on different kinds of problems. The computer vision researchers implicitly assumed that the problem of visual representation ends with the 3D reconstruction of moving scenes. The KR tradition in classical, symbolic AI usually underestimated the problem of grounding symbolic representations in the data coming from sensors.

It is our opinion that this state of affairs constitutes a strong limitation for the development of artificial autonomous systems (typically, robotic systems). It is certainly true that several aspects of the interaction between perception and behaviour in artificial agents can be faced in a rather reactive fashion, with perception more or less directly coupled to action, without the mediation of complex internal representations. However, we maintain that this is not sufficient to model all the relevant types of behaviour. For many kinds of complex tasks, a more flexible mediation between perception and action is required. In such cases the coupling between perception and action is likely to be “knowledge driven”, and the role of explicit forms of representation is crucial.

The above considerations are corroborated by the data concerning biological organisms, and by the results of cognitive neurosciences. In their book *The Visual Brain in Action*, A.D. Milner and M.A. Goodale (Milner and Goodale 1995, see also Milner 1997, Milner and Goodale 1998), on the basis of clinical and neuroanatomical evidence, make the hypothesis that in humans and in superior mammals the role of the visual system is twofold. On the one hand, vision directly controls a number of motor skills and behaviours of the organism, without any mediation of higher-level cognition (*visuomotor control*). On the other, vision has the purpose of providing the organism with a rich representation of the external environment, which is not directly linked to any specific behavioural output (*perceptual representation*). Rather, such representations are employed in different high-level cognitive capabilities, such as planning, memory, communication, and language. Visuomotor control is phylogenetically older than perceptual representation, and mammals share it with more primitive biological organisms. According to Milner and Goodale, these two aspects of the visual system correspond to different anatomical structures within the brain.

Our system can be intended as a suggestion to provide artificial agents with the latter of the above-mentioned functionalities (with particular emphasis, in this paper, on the perception of dynamic scenes). Of course, we do not claim that *all* the perceptual abilities of an artificial agent must pass through the mediation of explicit, declarative representations. As in the case of biological organisms, it is reasonable to expect that more direct, reactive aspects of visual perception play a central role in many kinds of tasks. However, we maintain that, in a great number of advanced applications, some form of more abstract, knowledge-oriented visual capability, that is less directly linked to behaviour control, could be badly dispensed. Nevertheless, it is our assumption that such a knowledge-oriented visual capability cannot be reduced to a passive, uniform registration of the data coming to the sensors. Rather, it can be profitably modelled as an active process driven by the knowledge and the purposes of the agent.

The existing attempts to integrate visual perception with propositional KR are mostly oriented to natural language interpretation, with particular emphasis on the aspects of man-machine interaction. They face only in a marginal way the general aspects of representation of knowledge. In addition, the existing proposals often concern specific domains of application, such as car traffic (Nagel 1994,

Neumann 1989), biomedical images (Tsotsos 1985), sport scenarios (Herzog and Wazinski 1994, Siskind 1994), simple assembly tasks (Kuniyoshi and Inoue 1993), visual surveillance (Buxton and Gong 1995).

Our aim is to propose a principled integration of the approaches of artificial vision and of symbolic KR. We assume that this integration requires the introduction of a missing link between these two kinds of representation. In our approach, the role of such a link is played by the notion of *conceptual space*. A conceptual space (*CS*) is a representation where information is characterised in terms of a metric space defined by a number of *cognitive* dimensions. These dimensions are independent from any specific representation language. We borrowed the notion of conceptual space from Gärdenfors' proposal (Gärdenfors 2000).

According to Gärdenfors, a *CS* acts as an intermediate representation between subconceptual knowledge (i.e., knowledge that is not yet conceptually categorised), and symbolically organised knowledge. According to this perspective, our architecture is organised in three *computational areas* (this terminology is reminiscent of the cortical areas in the brain).

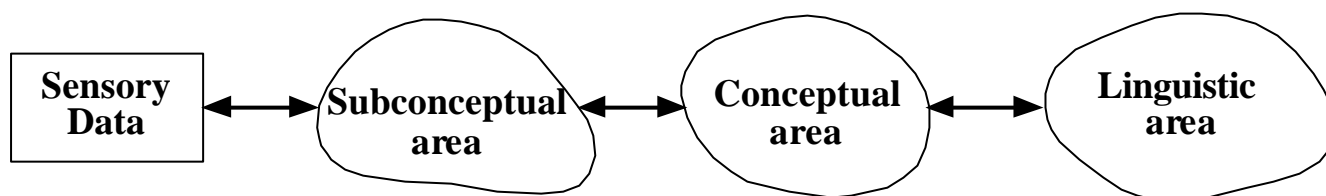


Fig.1. The three areas of representation, and the relations among them

Such areas must not be interpreted as a hierarchy of levels of higher abstraction. Rather, they are concurrent computational components working together on different commitments. Figure 1 schematically shows the relations among them. The *subconceptual* area is concerned with the low level processing of perceptual data coming from the sensors. The term *subconceptual* suggests that here information is not yet organised in terms of conceptual structures and categories. In this perspective, our subconceptual area includes a 3D model of the perceived scenes. Indeed, even if such a kind of representation cannot be considered “low level” from the point of view of artificial vision, in our perspective it still remains below the level of conceptual categorisation.

In the *linguistic area*, representation and processing are based on a propositional KR formalism (a logic-oriented representation language).

In the *conceptual area*, the data coming from the subconceptual area are organised in conceptual categories, which, however, are still independent from any linguistic characterisation.

There is no privileged direction in the communications among the three areas: some computations are strictly bottom up, with data flowing from the subconceptual up to the linguistic area, through the conceptual one; other computations combine top-down with bottom up processing. The mapping between the conceptual and the linguistic areas is based on a *focus of attention* mechanism, implemented by recurrent neural networks.

In a nutshell, the performances of our model can be summarised as follows: it takes in input a set of images corresponding to subsequent phases of the evolution of a certain dynamic scene, and it produces in output a declarative description of the scene, formulated as a set of assertions written in a first order logical language. The generation of such a description is driven both by the “a priori”, symbolic knowledge initially stored within the linguistic area, and by the information learned by the neural network component (see Section 5).

We claim that the proposed framework is not limited to some special purpose domain; rather, it could be extended to most cases in which abstract reasoning on sensory data is needed. In (Chella et al. 1997) this framework has been presented with reference to static scene analysis.

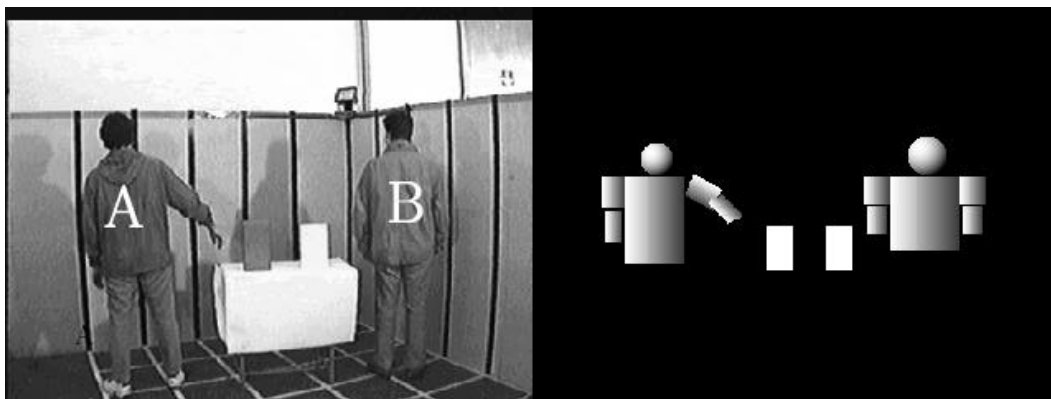


Fig. 2. The experimental set-up: two persons (A and B) stand at the two sides of a table, on which are placed some objects.

In this paper we refer to a scenario in which two persons stand at the two sides of a table with some objects on it (figure 2). We call the person on the left A, and the person on the right B. A and B may perform several actions: they may push blocks, grasp or drop them, lay them down on the table; A may hold out a block to B, or may throw it to him, and so on. Images of the scene are acquired by a colour CCD camera and processed by a 3D computer vision system, which recovers the shape and the motion of objects in the scene in terms of geometric and motion parameters of suitable 3D primitives. As a possible choice for 3D primitives, we adopted *superquadrics* (see Section 2).

This experimental set-up is not intended to be a realistic artificial vision scenario. It has been exclusively developed to illustrate our approach. As a consequence, many aspects that are not strictly relevant for our present purposes have not been taken into account. Thus, for example, occlusions are avoided, and different persons and objects in the scene are characterised by different colours.

The rest of the article is organised in the following sections. Section 2 deals with the conceptual area: here we describe the conceptual space we adopted for the representation of dynamic scenes. Section 3 presents the linguistic area and the symbolic level of representation. Section 4 is concerned with the focus of attention mechanism. Section 5 describes the implementation of the focus of attention by means of recurrent neural networks. Section 6 presents the system at work by describing a more detailed example. Finally, Section 7 is devoted to some conclusions, and to future developments of our research.

2. THE CONCEPTUAL AREA

Representations in the conceptual area are couched in terms of *conceptual spaces* (Gärdenfors 2000). Conceptual spaces provide a principled way for relating high level, linguistic formalisms with low level, unstructured representation of data. A conceptual space *CS* is a metric space whose dimensions are in some way related to the quantities processed in the subconceptual area. Different cognitive tasks can presuppose different conceptual spaces, and different conceptual spaces can be characterised by different dimensions. Examples of possible dimensions of a *CS* could be colour, pitch, mass, spatial coordinates, and so on. In some cases dimensions are strictly related to sensory data; in other cases they are more abstract in nature. In any case, dimensions do not depend on any specific linguistic description. In this sense, conceptual spaces come before any symbolic/propositional characterisation of cognitive phenomena. In particular, in this paper, we adopt a conceptual space devoted to the representation of the motion of geometric shapes.

We use the term *knoxel* to denote a point in a conceptual space. A *knoxel* is an epistemological primitive element at the considered level of analysis. For example, in (Chella *et al.* 1997) we assumed that, in the case of static scenes, a *knoxel* coincides with a 3D primitive shape, characterised according to some constructive solid geometry (CSG) schema. In particular, we adopted superquadrics (Pentland 1986, Solina and Bajcsy 1990) as a suitable CSG schema. Superquadrics offer a flexible and powerful way of representing 3D shapes, and are widely used within computer vision. Various techniques have been proposed for extracting superquadric parameters from static and dynamic scenes, also in scenarios with moving and interacting people. However, we do not assume that this choice is mandatory for our proposal. Our approach could be reformulated by adopting different models of 3D representation.

The entities represented in the linguistic area usually do not correspond to single *knoxels*. We assume that *complex entities* correspond to sets of *knoxels*. For example, in the case of a static scene, a complex shape corresponds to the set of *knoxels* of its simple constituents.

In order to account for the perception of dynamic scenes, we choose to adopt an intrinsically dynamic conceptual space. It has been hypothesised that simple motions are categorised in their wholeness, and not as sequences of static frames. According to this hypothesis, we define a *dynamic conceptual space* in such a way that every *knoxel* corresponds to a simple motion of a 3D primitive. In other words, we assume that simple motions of geometrically primitive shapes are our perceptual primitives for motion perception. In our experimental set-up, we chose superquadrics as geometric primitives. Therefore, a *knoxel* in a dynamic conceptual space is a simple motion of a superquadric.

Of course, the decision of which kind of motion can be considered “simple” is not straightforward, and is strictly related to the problem of motion segmentation. Marr and Vaina (Marr and Vaina 1982) used the term *motion segment* to indicate such simple movements. According to their *SMS* (State-Motion-State) schema, a simple motion is individuated by the interval between two subsequent overall rest states. Such rest states may be instantaneous.

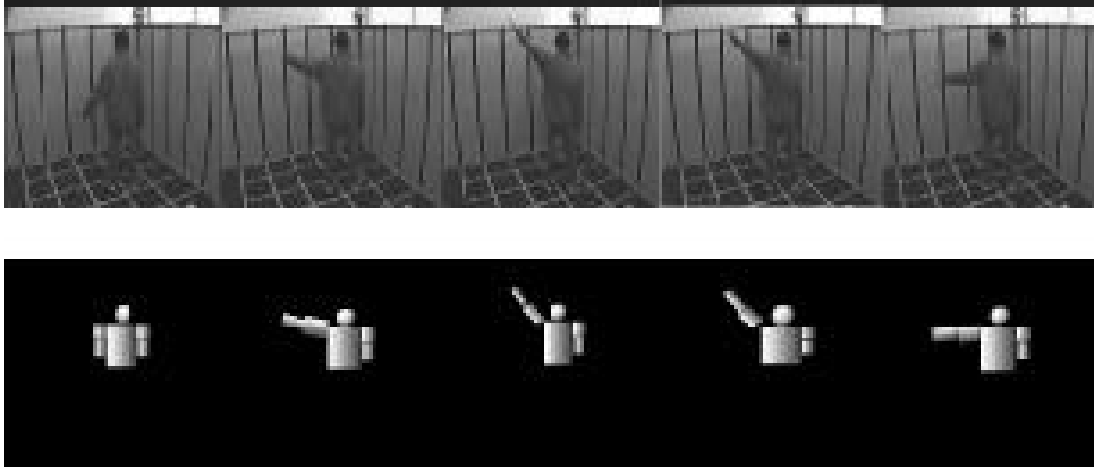


Fig. 3. A dynamic scene in which a person moves an arm.

Consider figure 3. A person moves her left arm up and down. The upward trajectory of the forearm is a simple motion that is represented in CS by a single knoxel, say \mathbf{k}_a . When the arm reaches its vertical position, an instantaneous rest state occurs. The second part of the trajectory of the forearm is another simple motion that corresponds to a second knoxel, say \mathbf{k}'_a . The same holds for the upper arm: the first part of its trajectory corresponds to a certain knoxel \mathbf{k}_b ; the second part (after the rest state) corresponds to a further knoxel \mathbf{k}'_b .

In intrinsically dynamic conceptual spaces, a knoxel \mathbf{k} corresponds to a *generalised* simple motion of a 3D primitive shape. By *generalised* we mean that the motion can be decomposed in a set of components x_i , each of them associated with a degree of freedom of the moving primitive shape. In other words, we have that

$$\mathbf{k} = [x_1, x_2, \dots, x_n,]$$

where n is the number of degrees of freedom of the moving superquadric. In this way, changes in shape and size are also taken into account.

In turn, each motion x_i corresponding to the i -th degree of freedom can be viewed as the result of the superimposition of a set of elementary motions f_j^i :

$$x_i = \sum_j X_j^i f_j^i$$

In this way, it is possible to individuate a set of basis functions f_j^i , in terms of which any simple motion can be expressed. Such functions can be associated to the axes of the dynamic conceptual space as its dimensions. In this way, the dynamic CS results in a functional space. The theory of function approximation offers different possibilities for the choice of basic motions: trigonometric functions, polynomial functions, wavelets, and so on.

For our purposes, we are not interested in the representation of any possible motion, but only in a compact description of perceptually relevant kinds of motion. In the domain we are facing here, the motions corresponding to each degree of freedom of a superquadric can be viewed as the result of the superimposition of the first low frequency harmonics, according to the well-known *Discrete Fourier Transform* (DFT) (Oppenheim and Shafer 1989).

In more detail, the sequence $x_i(n)$ of N samples in time of x_i may be viewed as the result of superimpositions of the first (usually three) harmonics:

$$x_i(n) = \frac{1}{N} \sum_{l=0}^{N-1} X_i(l) e^{j \frac{2\pi}{N} ln}$$

where the $X_i(l)$ are the (complex) coefficients of the DFT:

$$X_i(l) = \frac{1}{N} \sum_{n=0}^{N-1} x_i(n) e^{j \frac{2\pi}{N} ln}$$

The choice of considering only the first low-frequency harmonics is motivated by the fact that, in the present scenario, we generally have smooth motions (people moving their arms and moving objects).

Figure 4 shows an evocative representation of a dynamic conceptual space, along with a generic knoxel \mathbf{k}_a in it. In the figure, each group of axes f^i corresponds to the i -th degree of freedom of a simple shape; each axis f_j^i in a group f^i corresponds to the j -th component pertaining to the i -th degree of freedom.

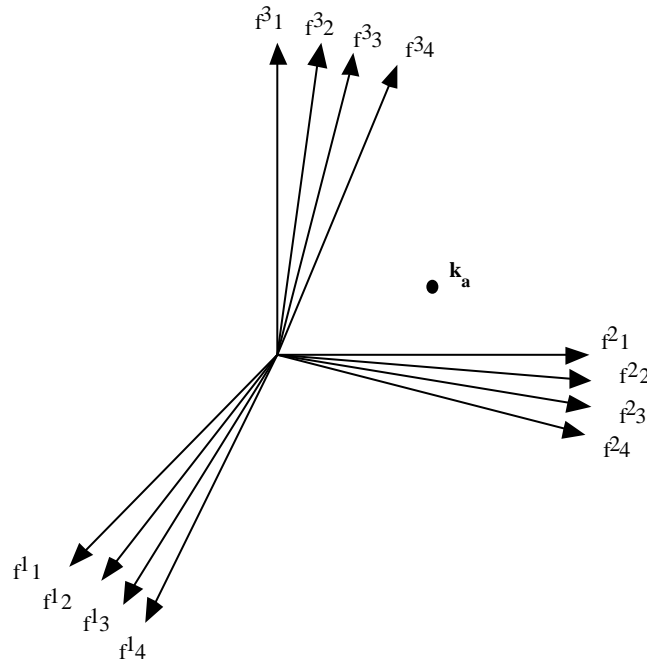


Fig. 4. A dynamic conceptual space.

Up to now we considered only *simple motions*, i.e. motions of *simple shapes* occurring within an interval between two subsequent rest states. We call *composite simple motion* a motion of a composite object (i.e. an object approximated by more than one superquadric). A *composite simple motion* is represented in the *CS* by the set of knoxels corresponding to the motions of its components. For example, the first part of the trajectory of the whole arm shown in figure 3 is represented as a composite motion made up by the knoxels k_a (the motion of the forearm) and k_b (the motion of the upper arm). Note that in composite simple motions the (simple) motions of their components occur simultaneously. That is to say, a composite simple motion corresponds to a single configuration of knoxels in the conceptual space.

In order to consider the composition of several (simple or composite) motions arranged according to some temporal relation (e.g., a sequence), we introduce the notion of *action*. Our use of this term is consistent with Allen's proposal (Allen 1984). An action corresponds to a series of different configurations of knoxels in the conceptual space. We assume that the configurations of knoxels within a single action are separated by instantaneous changes. In the transition between two subsequent configurations, the “scattering” of at least one knoxel occurs. This corresponds to a discontinuity in time, and is associated with an instantaneous event. Marr and Vaina (Marr and Vaina 1982) call these discontinuities *states*, since they correspond to instantaneous rest states of the motion.

3. THE LINGUISTIC AREA

Representation in the linguistic area is based on a high-level, logically oriented formalism. In particular, we adopted a hybrid formalism in the KL-ONE tradition (Brachman and Schmolze 1985, Woods and Schmolze 1992). Such a formalism is hybrid in the sense that is constituted by two different components: a *terminological component* for the description of concepts, and an *assertional component*, that stores information concerning a specific context. In our case study, the concepts in the terminological component describe various types of motion, of action, and so on. The assertional component stores the linguistic information concerning specific perceived situations.

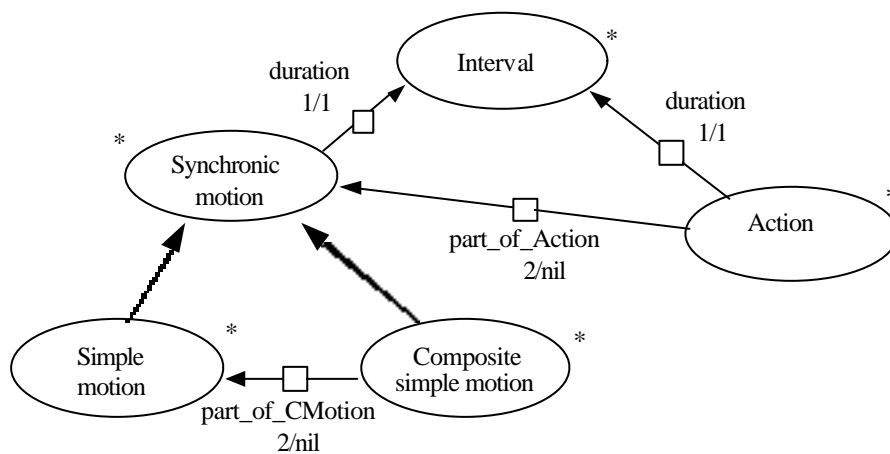


Fig. 5. A fragment of the terminological KB in the linguistic area.

Figure 5 shows a fragment of our terminological knowledge base (we adopted a network notation similar to that of Brachman and Schmolze 1985); it describes the most general concepts representing motion in the KB. Intuitively, a *Synchronic_motion* corresponds to an arrangement of knoxels that simultaneously occurs in the CS: in synchronic motions there is no scattering of knoxels. Every *Synchronic_motion* is associated to a time *Interval* through the role *duration*.

A *Simple_motion* is a synchronic motion that has no parts: it corresponds to a single knoxel in CS. A *Composite_simple_motion* is a *Synchronic_motion* with at least two parts, which are simple motions occurring simultaneously. An *Action* is a motion involving a temporal evolution (a scattering in CS); an *Action* has at least two parts that are instances of *Synchronic_motion* (not occurring at the same time). Also the concept *Action* is related to *Interval* through the role *duration*. All these concepts are considered primitive (they are marked with an asterisk), since the network expresses necessary but not sufficient conditions for their application.

As an example of action description, let us consider the representation of the action of seizing an object (figure 6). The action *Seize* is represented as composed by two parts, both subsumed by *Composite_simple_motion*. The first part, an *Arm_approach*, corresponds to the movement of the arm that approaches the object to be seized. Approaching an object is a particular case of stretching out an arm; therefore, the concept *Arm_approach* is subsumed by *Stretch_out*. The composite motion *Stretch_out* is in turn described as composed by two simple motions: a *Forearm_stretching* and an *Upper_arm_stretching*. When the object has been approached, the *Arm_approach* motion terminates, and a *Grasp* motion begins.

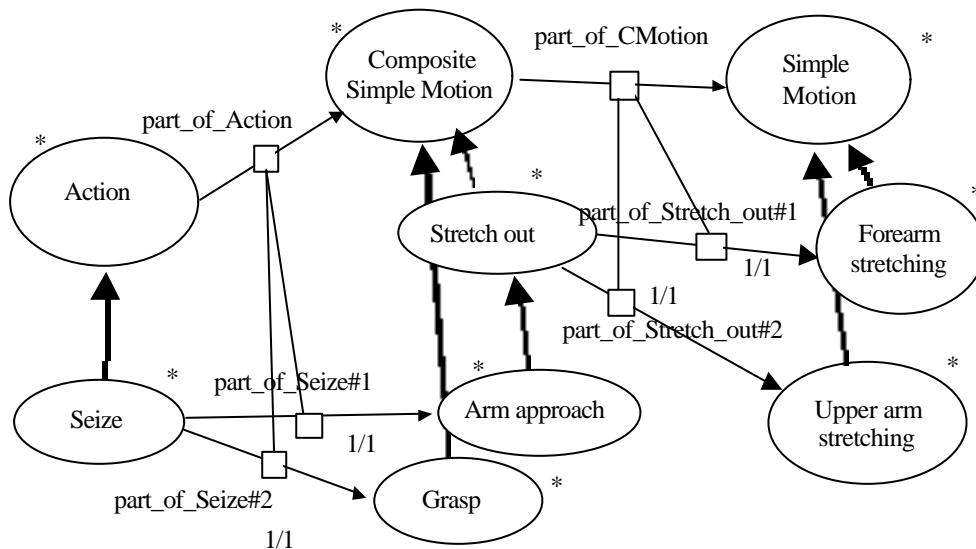


Fig. 6. The description of the *Seize* motion in the terminological KB.

In the above examples, the temporal relations between the various motions and actions are not explicitly represented in the terminological KB. The formalism could be extended with tense operators corresponding, for example, to Allen primitives (Allen 1984; see Artale and Franconi 1998, Wedia and Litman 1992 for similar extensions of terminological languages). However, we do not take into consideration these aspects here. In the following section we will suggest that in our model, for the purposes of action recognition, an explicit declarative representation of temporal relations can be dispensed in many cases. Our position is that it is possible to deal with various aspects of temporal ordering by means of the mechanism of the focus of attention.

The *assertional component* contains facts expressed as assertions in a predicative language, in which the concepts of the terminological components correspond to one-argument predicates, and the roles (e.g., *part_of*, *duration*) correspond to two argument relations. For example, in order to assert the existence of an instance $CtM\#1$ of the concept *Composite_simple_motion*, the formula: `Composite_simple_motion(CtM#1)` is added to the assertional KB.

4. THE FOCUS OF ATTENTION

A finite agent with bounded resources cannot carry out a one-shot, exhaustive, and uniform analysis of the acquired data within reasonable resource constraints. Some of the acquired data (and of the relations among them) are more relevant than others, and it should be a waste of time and of computational resources to detect true but useless details. In order to avoid the proliferation of useless assertions in the linguistic area, in our model the association between symbolic representations and configurations of knoxels in *CS* is driven by a sequential scanning mechanism that acts as some sort of internal *focus of attention*, and is inspired by the attentive processes in human vision.

The focus of attention allows the system to select the relevant aspects of a perceived scene by sequentially scanning the knoxels in the conceptual space. It is crucial in determining which assertions must be added to the linguistic knowledge base: not all true (and possibly useless) assertions are generated, but only those that are judged to be relevant on the basis of the attentive process.

In general, the recognition of a certain kind of simple motion (a certain knoxel in *CS*) will elicit the expectation of other motions simultaneously occurring within the scene (e.g., the motions of other parts of a certain complex shape). In such cases, the mechanism seeks the expected knoxels in the current *CS* configuration. We call this type of process *synchronic attention*, as it refers to a single configuration of knoxels in *CS*.

The recognition of a certain configuration in *CS* could also elicit the expectation of a change in the arrangement of knoxels. In this case, the mechanism seeks the expected knoxels in subsequent configurations of *CS*. We call this process *diachronic attention*, since it involves subsequent configurations of *CS*.

We take into account two main sources of expectations that play some role in attentive processes. On the one hand, expectations could be generated on the basis of propositional information explicitly stored in the symbolic knowledge base. For example, in the semantic network the stretching of an arm is described as composed by two simple motions: a motion of the forearm and a motion of the upper arm. As soon as either one of these two motions is detected in *CS*, the symbolic description elicits the expectation of the other. We call this kind of expectation *linguistic*.

On the other hand, expectations could also be generated on the basis of a purely associative, Hebbian mechanism. Suppose that many examples of scenes have been seen, in which one agent offers a certain object to the other, and the latter takes it. The system could learn to associate these kinds of motion: when one agent stretches his arm holding the object towards the other agent, he is expected to grasp it. We call this kind of expectation *associative*.

Note that symbolic and associative mechanisms have a role in both synchronic and diachronic expectations.

5. A CONNECTIONIST IMPLEMENTATION OF THE MAPPING

In the present implementation of our model, a connectionist device based on recurrent neural networks implements the mapping between the conceptual space and the linguistic area. Concepts that describe primitive motions and actions in the linguistic area are associated to suitable recurrent neural networks acting as “predictive filters” (Psarrou and Buxton 1994) on the sequences of knoxels corresponding to possible instances of the concepts themselves.

Suppose that the set of knoxels $s = \{\mathbf{k}_1, \mathbf{k}_2, \dots, \mathbf{k}_m\}$ corresponds to a prototypical instance of a generic concept C . When a knoxel of s , say \mathbf{k}_1 , has been individuated in the conceptual area and it is given as input to the recurrent network associated with C , the network generates as output another knoxel of s , say \mathbf{k}_2 . In this way, the network predicts the presence of \mathbf{k}_2 in CS . This expectation is considered to be confirmed when the subconceptual area individuates a knoxel \mathbf{k}_2^* close enough (in some suitable sense) to \mathbf{k}_2 . If such expectation is confirmed, then the network receives as input \mathbf{k}_2 and generates the expectation of a new knoxel \mathbf{k}_3 , and so on. The network therefore recognises the configuration of knoxels of the associated concept according to a recognition and expectation loop (Marr and Vaina 1982).

The use of neural networks makes it possible to avoid exhaustive, explicit description of conceptual categories in the linguistic area: in some sense, prototypical motions and actions “emerge” from the activity of this associative mechanism during a training phase based on examples. In addition, the measure of similarity between a prototype and a given motion or action is implicit in the behaviour of the network and is determined during the learning phase.

In particular, we adopted *time-delay attractor neural networks* (Amit 1988, Hopfield 1982). This choice offers an advantage in that such networks are based on the well-studied energetic approach; the learning phase is fast; they allow for a uniform treatment of both the recognition and the generation expectations.

Each primitive motion concept in the linguistic area is associated with a single time-delay neural network. In this way, the conceptual space is implemented by the superimposition of the Hopfield energy landscapes of all the networks associated with the concepts in the linguistic area. By means of the *asymmetric time-delay connections* (Kleinfeld 1986), each network operates as a predictive filter on sequences of knoxels. Given a starting knoxel, each network generates a hypothesis on the subsequent knoxels, and each network competes with the others.

Given a set of knoxels $\{\mathbf{k}_1, \mathbf{k}_2, \dots, \mathbf{k}_l\}$ corresponding to a prototypical instance of a concept C , each \mathbf{k}_i is a *point attractor* of the Hopfield energy landscape of the neural network associated to C . As usual with this kind of network, when a knoxel \mathbf{k}_i^* close enough to \mathbf{k}_i is presented as input, the state of the network evolves so as to settle to the stable point attractor \mathbf{k}_i . The general form of the energy function is:

$$E_1(t) = -\frac{1}{2} \sum_{i=1}^m \sum_{\substack{j=1 \\ j \neq i}}^m T_{ij} s_i(t) s_j(t) \quad (1)$$

where m is the number of units of the network, $\mathbf{T} = \{T_{ij}\}$ is the connections matrix storing the attractors related to the knoxels of C , and $\mathbf{s}(t) = \{s_i\}$ is the current state of the network.

As previously stated, the focus of attention imposes a sequential order on the attractors that correspond to the knoxels of a generic concept C . This is obtained by adding *time-delay* connections among the units of the networks. The resulting “pseudo-energy” term (Kleinfeld 1986) is:

$$E_2(t) = -\sum_{i=1}^m \sum_{\substack{j=1 \\ j \neq i}}^m D_{ij} s_i(t) s_j(t - \tau) \quad (2)$$

where τ is the time delay unit among two subsequent scanned knoxels, $\mathbf{D} = \{D_{ij}\}$ is the delayed synapses connection matrix storing the order of knoxels, $\mathbf{s}(t)$ and $\mathbf{s}(t - \tau)$ are respectively the current and the past τ -th state of the network.

The global energy function of the neural network with the time-delay connections is the sum of equations (1) and (2):

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

This equation captures the dynamic processes in the adopted neural networks. An attractor is stable for some time interval determined by the E_1 term. After several τ times, the term E_2 destabilizes the attractor and carries the state of the network toward the successive attractor of the sequence. In this way, the neural networks generate the expectations of subsequent knoxels of the perception act associated with a concept.

If C is a composite motion, then the sequences of knoxels will refer to the same CS configuration. If C is an action, the sequences of knoxels will refer to subsequent CS configurations. The first case is an example of synchronic attention, the second one is an example of diachronic attention.

The case of simple motions is treated as a special case of composite motions, where the set $s = \{\mathbf{k}\}$ is a singleton. In such cases the network acts as an autoassociator: when \mathbf{k} is given as input to the recurrent network, it generates as output \mathbf{k} itself.

6. SOME EXPERIMENTAL RESULTS

Our experiments are based on a working, even if simplified, artificial vision system. The experimental set up consists of two persons (A and B) standing at the two sides of a table, one in front of the other. Some simple objects are arranged on the top of the table. The camera is fixed, and perpendicular to one of the free sides of the table. Occlusions are avoided, and the persons and objects are characterised by different colours, to simplify the segmentation phase. Images are acquired by a colour CCD camera, and processed in order to recover the moving and the static shapes in terms of superquadrics.

The account of the following pages is oversimplified if compared with the complex problems involved in obtaining a 3D reconstruction of a dynamic scene starting from visual data. For example, we do not take into account the fact that, in any realistic situation, 3D reconstruction is affected by uncertainty. However, this is not the focus of the present article. Here we are interested in the problem of extracting declarative, symbolic knowledge from the outputs of a vision system. The problem of dealing with uncertainty is a central topic of the research on effective vision systems, and we trust in the results achieved in that field. It could turn out to be useful, in further developments of our model, to extend the CS and/or the linguistic representation with some explicit treatment of uncertainty. Similar considerations hold for aspects such as, for example, fuzziness or non-monotonicity. However, we do not deal with these possibilities here.

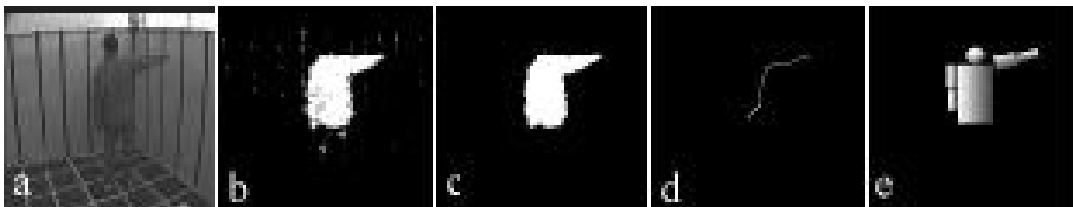


Fig. 7. The phases of the 3D reconstruction performed by the artificial vision system.

Figure 7 shows the phases of 3D reconstruction. In the figure, (a) is the starting image, in which a person wearing a red jacket is shown. In phase (b) the red component of the starting image is extracted. In (c) the contours have been recovered by means of a median filtering operation followed by the application of a region growing algorithm (Zucker 1976). Phase (d) shows the skeletons obtained by applying a suitable skeletonisation algorithm (Zhang and Suen 1984). Finally, in (e) information from skeleton and boundaries is combined in order to approximate the shape by means of occluding cylinders (Marr and Vaina 1982). In order to make possible such an approximation starting from the data shown in (c) and in (d), the system is endowed with a schematic, a priori model of a human body, expressed in terms of superquadrics.

As described in the previous section, the motion parameters of the recovered 3D primitives are segmented according to the *SMS* schema, and the (simple) motions are expressed in terms of low-frequencies harmonics. The knoxels so obtained are sent as input to the recurrent neural networks associated with the concepts, which start the recognition and expectation loop that generates the assertions describing the scene.

Figure 8 shows the evolution of a dynamic scene along with the corresponding 3-D reconstruction. In it the person A seizes an object O on the table, and offers it to B. B stretches out his arm, in order to receive it. Then B drops O on the table. The first row of pictures shows A seizing the object O; the second row shows A offering O to B, who receives it; the last row shows B dropping O on the table. At the end of the sequence, both A and B are in a rest state.

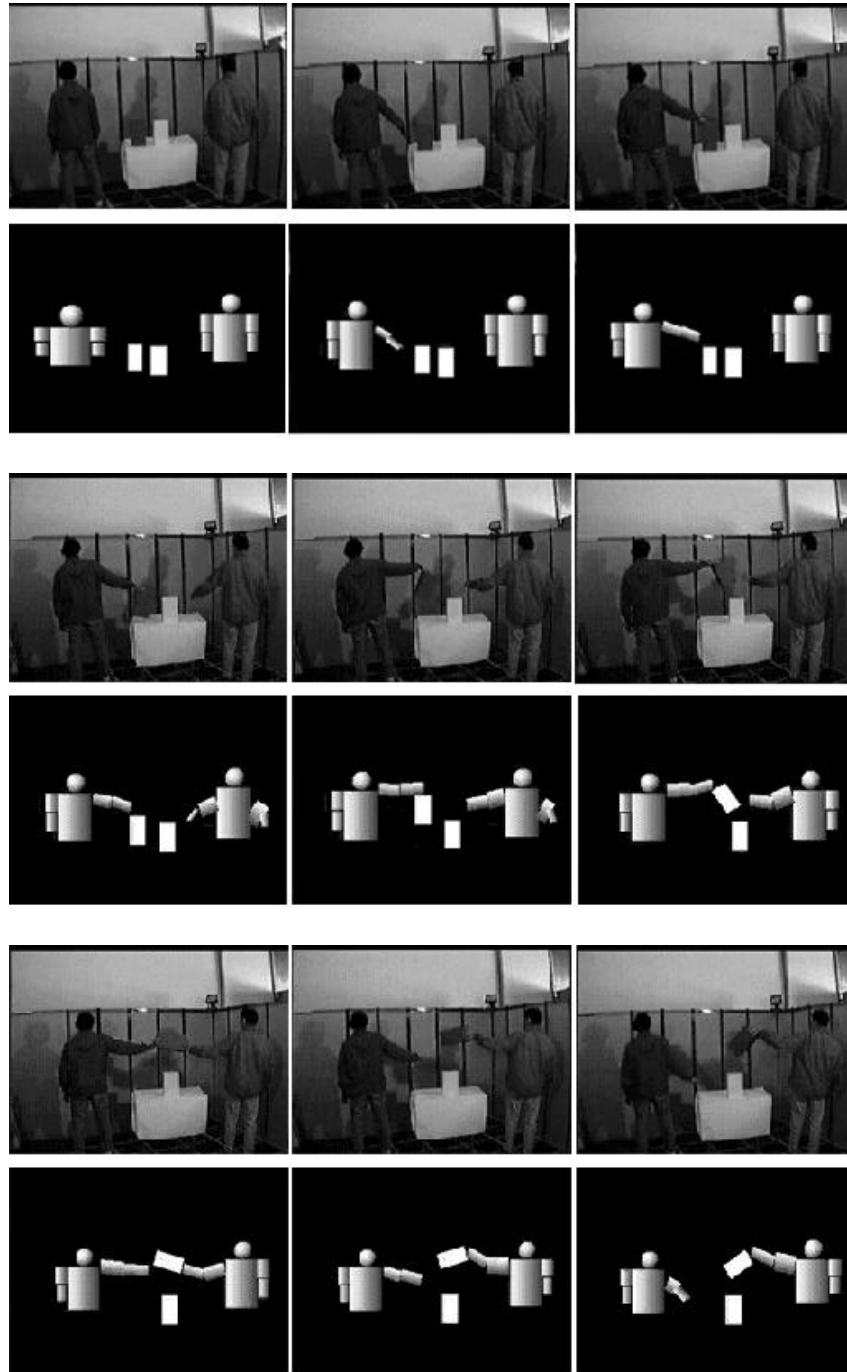


Fig. 8. A complex dynamic scene. A seizes an object, and offers it to B. B receives the object and drops it on the table.

In the following, we report some of the assertions generated in the linguistic area, along with the phases of the process that produces them. The description that follows is schematic, and, for sake of brevity, not every aspect has been reported in detail. This section has the aim of giving a general idea of the overall behaviour of the system.

In reading what follows, it must be remembered that the model does not possess complete, exhaustive characterisations of concepts, neither at the level of the declarative, explicit descriptions of the linguistic area, nor at the level of the knowledge implicitly stored in the connectionist networks. In particular, the latter kind of knowledge heavily depends on the past “experiences” of the system (both regarding the recognition of specific concepts, and the Hebbian mechanism driving the focus of attention and determining the hypotheses that are generated). When the system “recognises” a certain motion as a seizing, or a certain shape as a forearm, this is only *the best hypothesis* among those that are actually available to it. Obviously, the system can make mistakes. The quality of its performances is determined by, among other things, the richness and the exhaustivity of the set of examples on which it was trained, and on the degree of “acquaintance” with its environment and with the kinds of events occurring in it.

In general, the recognition of motion (and action) types is performed by associating certain perceivable motion features (or certain perceivable relationships between motions) to certain concepts. Such features or relationships do not constitute a *definition* of the concepts involved. Up to now, this mechanism has been used to model learned habits. However, it could be extended, in order to capture certain (presumably innate) perceptual mechanisms, such as, for example, those concerning the perception of causality. Consider Michotte’s classical experiments, according to which human perceptual apparatus is pre-disposed to attribute causality or agency to abstract motions of shapes (Michotte 1954). In such cases, built-in mapping procedures would be associated to a pre-set number of concepts concerning a given domain (e.g., the domain of causality relations).

Consider the sequence of figure 8. In the first part, the knoxel \mathbf{k}_1 , corresponding to the stretching out of the forearm of A, is recognised as a simple motion and as an instance of *Forearm_stretching*. Then, on the basis of linguistic expectations, the system expects the motion of the upper arm of A. Therefore, it hypothesises the existence of the corresponding knoxel \mathbf{k}_2 , and searches for it in *CS*. This is an example of synchronic attention, since \mathbf{k}_1 and \mathbf{k}_2 simultaneously occur in *CS*. In more detail, when the knoxel \mathbf{k}_1 (the forearm stretching) is recognised, the assertion

```
Forearm_stretching(F_s#1)
```

is added to the knowledge base (where $F_s\#1$ is a symbolic constant denoting \mathbf{k}_1). Then, \mathbf{k}_1 is fed to the recurrent neural networks associated to those concepts in the terminological KB that have a *Forearm_stretching* as a possible part. Such networks generate hypotheses about which other knoxels could be present in the scene. As soon as the hypothesis generated by the recurrent networks associated with *Stretch_out* is satisfied by \mathbf{k}_2 , this knoxel is recognised to be an instance of *Upper_arm_stretching*. It is given a symbolic name (say, $U_a_s\#1$), and the following assertion is added to the knowledge base:

```
Upper_arm_stretching(U_a_s#1)
```

In addition, the existence of an instance of *Stretch_out* (say, $S_o\#1$) is asserted, and $F_s\#1$ and $U_a_s\#1$ are asserted to be parts of $S_o\#1$:

```
Stretch_out(S_o#1)
part_of_Stretch_out#1(S_o#1, F_s#1)
part_of_Stretch_out#2(S_o#1, U_a_s#1).
```

Then (on the basis of linguistic and associative expectations working together), the *Stretch_out* motion $S_o\#1$ is recognized to be an *Arm_approach*, and to be a part of a *Seize* action. Therefore, expectations for a suitable *Grasp* motion are generated. As such expectations are confirmed, a *Seize* instance $S\#1$ is generated, and the previous motions are recognised to be parts of $S\#1$:

```
Approach(S_o#1)
Grasp(Gr#1)
Seize(S#1)
part_of_Seize#1(S#1, S_o#1)
part_of_Seize#2(S#1, Gr#1).
```

Once this composite motion has been performed, the scene changes, and a new *Stretch_out* motion (say, $S_o\#2$) is detected: now A stretches his arm towards B holding O in his hand. The system hypothesises that A is offering O to B (i.e., that $S_o\#2$ is an instance of *Offer*). As a consequence, a further change in *CS* is expected: B in his turn is expected to stretch out his arm towards A, in order to take O. When this expectation is confirmed, an instance $R\#1$ of the *Receive* motion is generated. In addition, a *Give* action $G\#1$ is instantiated, which has $S_o\#2$ and $R\#1$ as its parts (the description of the concept *Give* in the symbolic KB encompasses two *part_of* roles: the first with restriction *Stretch_out*, the second with restriction *Receive*). This is an example of diachronic expectation, since it involves subsequent configurations of *CS*. The following are some of the generated assertions:

```
Stretch_out(S_o#2)
Offer(S_o#2)
Receive(R#1)
Give(Gv#1)
part_of_Give#1(Gv#1, S_o#2)
part_of_Give#2(Gv#1, R#1)
```

7. CONCLUSIONS

Our approach allows for an effective integration between subconceptual and linguistic computations. This is achieved through the introduction of conceptual spaces as the missing link between these two kinds of representation and processing.

The adoption of attentive mechanisms for the scanning of conceptual spaces allows the adoption of a very concise representation of concepts in the linguistic area for recognition tasks. This is because many details concerning the concepts and the relationships among them are implicitly stored in the associative mechanism performing the mapping, and it is not necessary to represent them explicitly.

For instance, the temporal constraints between the parts of an action can be implicitly coded within the mapping mechanism itself. This approach could turn out to be particularly useful when dealing with non-monotonic aspects in the description of concepts. For example, in recognition problems, all aspects concerning “typicality” may not be represented symbolically, but may be stored in the associative mapping.

As far as further developments are concerned, a central aspect consists in the interaction between visual perception and action planning. In this perspective, a crucial problem is that of the reciprocal relations among vision, reactive forms of planning and more deliberative, knowledge-driven behaviour. According to the distinction developed by (Milner and Goodale 1995), the problem is the interaction between visuomotor control and perceptual representations in determining behaviour. As far as biological organisms are concerned, it is still unclear to what extent the neural representations involved in action perception and in action execution share the same code (see for example Decety and Grézes 1999). In the context of our proposal, the reciprocal roles of subconceptual computations, conceptual area representations, and linguistic knowledge in planning the behaviour of an artificial system must still be studied. In view of this, we are starting a series of experiments in which our artificial vision model is mounted on an autonomous robot.

A particular aspect of the above problem concerns the possibility of performing perceptual explorations. Our present experimental set up rests on the hypothesis that all the information about the scene is recovered by a fixed camera. This is a hard constraint that may be overcome by adopting a moving camera. For example, uncertainty and occlusions may be faced by means of suitable explorations (Maver and Bajcsy 1993). By generalising the behaviours described in the previous sections, the generated expectations may effectively drive the movements of the camera towards the relevant parts of the scene, in order to acquire the lacking information. In this case, the interpretation process may be generated by a direct and effective interaction between the subconceptual and the linguistic areas.

ACKNOWLEDGEMENTS

We would like to thank Roberto Di Martino and Massimiliano Greco for their contribution to the implementation of the experimental set-up. We also wish to thank the anonymous referees that, with their comments, helped us to improve the quality of the paper. This work has been partially supported by the project “Progetto Cofinanziato CERTAMEN” funded by the Italian Ministry for the University and the Scientific and Technological Research (MURST).

REFERENCES

- Allen, J.F. 1984. Towards a general theory of action and time, *Artificial Intelligence*, **23**, 123-154.
- Amit, D. 1988. *Modeling Brain Function. The World of Attractor Neural Networks*, Cambridge, UK, Cambridge University Press.
- Artale, A. and Franconi, E. 1998. A temporal description logic for reasoning about actions and plans, *Journal of Artif. Intel. Research*, **9**, 463-506.
- Brachman, R.J. and Schmoltze, J.C. 1985. An overview of the KL-ONE knowledge representation system, *Cognitive Science*, **9**, 171-216.
- Buxton, H. and Gong, S. 1995. Visual surveillance in a dynamic and uncertain world, *Artificial Intelligence*, **78**, 371-405.

- Chella, A, Frixione, M. and Gaglio, S. 1997. A cognitive architecture for artificial vision, *Artificial Intelligence*, **89**, 73-111.
- Decety, J. and Grèzes, J. 1999. Neural mechanisms subserving the perception of human actions, *Trends in Cognitive Sciences*, **3**(5), 172-178.
- Gärdenfors, P. 2000. *Conceptual Spaces*, Cambridge, MA, MIT Press.
- Herzog, G. and Wazinski, P. 1994. VISual TRANslator: Linking perceptions and natural language descriptions, *Artificial Intelligence Review*, **8**, 175-187.
- Hogg, D.C. 1984. *Interpreting Images of a Known Moving Object*, PhD thesis, University of Sussex at Brighton, UK.
- Hopfield, J.J. 1982. Neural networks and physical systems with emergent collective computational abilities, *Proc. Nat. Acad. Sci. USA*, **79**, 2554-2558.
- Israel, D., Perry, J. and Tutiya, S. 1991. Actions and movements, in: *Proc. 12th IJCAI Sidney, Australia*, 1060-1065.
- Kleinfeld, D. 1986. Sequential state generation by model neural networks, *Proc. Nat. Acad. Sci. USA*, **83**, 9469-9473.
- Kuniyoshi, Y. and Inoue, H. 1993. Qualitative recognition of ongoing human action sequences, in: *Proc. 13th IJCAI Chambery, France*, 1600-1609.
- Marr, D. and Vaina, L. 1982. Representation and recognition of the movements of shapes, *Proc. R. Soc. Lond. B*, **214**, 501-524.
- Maver, J. and Bajcsy, R. 1993. Occlusions as a guide for planning the next view, *IEEE Trans. Pat. Anal. Mach. Intel.* **15**, 417-433.
- McCarthy, J. and Hayes, P.J. 1969. Some philosophical problems from the standpoint of artificial intelligence, in: *Machine Intelligence 4*, edited by Meltzer, B. and Michie, D., Edinburgh, Edinburgh University Press, 463-502.
- McDermott, D. 1982. A temporal logic for reasoning about plans and actions, *Cognitive Science*, **6**, 101-155.
- Michotte, A. 1954. *La perception de la causalité*, Louvain, Publications Universitaires de Louvain, eng. trans. *The Perception of Causality*, London, Methuen & Co. Ltd.
- Milner, A.D. 1997. Vision without knowledge, *Phil. Trans. R. Soc. Lond. B*, **352**, 1249-1256.
- Milner, A.D. and Goodale, M.A. 1995. *The Visual Brain in Action*, Oxford, Oxford University Press.
- Milner, A.D. and Goodale, M.A. 1998. The visual brain in action, *Psyche*, **4**(12).

- Nagel, H.H. 1994. A vision of “vision and language” comprises action: An example from road traffic, *Artificial Intelligence Review*, **8**, 189-214.
- Neumann, B. 1989. Natural language description of time-varying scenes, in: *Semantic Structures*, edited by Waltz, D.L., Hillsdale, NJ, Lawrence Erlbaum.
- Oppenheim, A.V. and Shafer, R.W. 1989. *Discrete-Time Signal Processing*, Englewood Cliffs, NJ, Prentice Hall.
- O'Rourke, J. and Badler, N.I. 1980. Model-based image analysis of human motion using constraint propagation, *IEEE Trans. Pat. Anal. Mach. Intell.*, **2**, 522-536.
- Pentland, A.P. 1986. Perceptual organization and the representation of natural forms, *Artif. Intell.*, **28**, 293-331.
- Pentland, A.P. and Horowitz, B. 1991. Recovery of nonrigid motion and structure, *IEEE Trans. Patt. Anal. Mach. Intell.*, **13**, 730-742.
- Psarrou, A. and Buxton, H. 1994. Motion analysis with recurrent neural nets, in: *Proc. of ICANN 92*, edited by Marinaro, M. and Morasso, P., Berlin, Springer, 54-57.
- Rohr, K. 1994. Towards model-based recognition of human movements in image sequences, *CVGIP:Image Understanding*, **59**, 94-115.
- Shoham, Y. 1988. *Reasoning about change*, Cambridge, MA, MIT Press.
- Siskind, J.M. 1994-5. Grounding language in perception, *Artificial Intelligence Review*, **8**, 371-391.
- Solina, F. and Bajcsy, R. 1990. Recovery of parametric models from range images: The case for superquadrics with global deformations, *IEEE Trans. Patt. Anal. Mach. Intell.*, **12**(2), 131-146.
- Tsotsos, J.K. 1985. Knowledge organisation and its role in representation and interpretation for time-varying data: the ALVEN system, *Computational Intelligence*, **1**, 16-32.
- Weida, R. and Litman, D. 1992. Terminological reasoning with constraint networks and an application to plan recognition, in *Proc. KR-92*.
- Woods, W. and Schmoltze, J.C. 1992. The KL-ONE family, *Computers and Mathematics with Applications*, **23**(5), 133-178, also in *Semantic Networks in Artificial Intelligence*, edited by Lehmann, F. and Rodin, E.Y., 1992, Pergamon Press.
- Zhang, T.Y. and Suen, C.Y. 1984. A fast parallel algorithm for thinning digital patterns, *Comm. ACM*, **27**, 236-239.
- Zucker, S.W. 1976. Region growing: Childhood and adolescence. *Computer Graphics and Image Processing*, **5**, 382-399.