Conceptual Spaces for Computer Vision Representations

A.Chella[1,2], M.Frixione[3] and S.Gaglio[1,2]

[1]Dip. di Ingegneria Automatica e Informatica - Univ. of Palermo, Italy

[2]CERE-CNR, Palermo, Italy

[3]Dip. di Scienze della Comunicazione - Univ. of Salerno, Italy

Abstract

A framework for high-level representations in computer vision architectures is described. The framework is based on the notion of *conceptual space* proposed by Gärdenfors [12]. This approach allows to define a conceptual semantics for the symbolic representations of the vision system. In this way the semantics of the symbols can be grounded on the data coming from the sensors. In addition, the proposed approach generalizes the most popular representation frameworks adopted in computer vision.

1 Introduction

According to Marr [17], computer vision is the process that, starting from bidimensional images, automatically discovers *what* is present in the external world and *where* it is. Computer vision is an information-processing task that receives in input raw and low structured data (the images acquired by a video camera), and gives as its output highly structured data (suitable symbolic descriptions of the scene). Such data are crucial for the effective autonomy of a moving robot [2], for sensing actions in planning [22], for teleautonomy and telepresence [8], and also for advanced man-machine interfaces [11].

The schema of a general architecture for computer vision is shown in Fig. 1. The first block is the *camera*. The data in this block are strictly related to the signals coming from the sensors, and they are generally structured as a matrix of *pixels*.

The information in the camera block is processed by *low level vision* algorithms. The main task of these algorithms is to extract low-level information from the scene, such as contours, edges, textures [14].

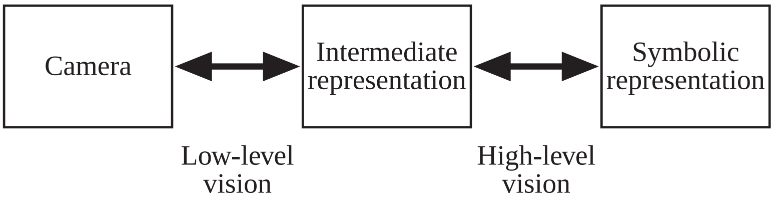

Figure 1: A general architecture for computer vision.

In the *intermediate representation* block the data extracted by the low level processing are represented in terms of the composition of suitable primitives, and grouped on the basis of conceptual categories. In a sense, the data representation in such an intermediate block may be viewed as a compact and compressed representation of the scene acquired by the camera. The minimum requirements of this representation are the independence from illumination conditions and from the specific point of view.

The *symbolic representation* constitutes the output block of this architecture; in it the perceived scene is represented in terms of a high-level formalism, e.g., a first order logical language. In this case the individual constants of the formalism represent specific entities in the scene, one-place predicates represent classes of entities and, in general, n-place predicates represent n-ary relations. Suitable extensions of such type of formalism have been proposed in the literature, which include symbolic descriptions of similarities and analogies among entities [10].

The algorithms linking together the intermediate and the symbolic representations are generally known as *high-level vision* algorithms [27]. Their role is to identify and classify the entities of the intermediate representation in the terms of the symbolic formalism.

In the following, we will concentrate on the intermediate representation block of this schema. In particular, in Sect. 2, we will present the approach proposed in [6, 7] based on the adoption of the notion of *conceptual space* [12]. In Sect 3, we will review some of the most widely adopted representation frameworks for computer vision, and we will compare them with our approach. Finally, in Sect. 4, we will present some conclusions and we will indicate some open research problems.

2 Conceptual Spaces

2.1 Introductory remarks

The theory of conceptual spaces provides a robust cognitive framework for the characterization of the internal representations of the environment of an agent. A *conceptual space* CS is a *metric* space in which entities are characterized by a number of quality dimensions [12]. Examples of such dimensions are color, pitch, volume, spatial coordinates, and so on. Some dimensions are closely related to the sensorial inputs of the system, other may be characterized in more abstract terms. Such dimensions represent qualities of the environment independently form any linguistic formalism or description. In this sense, a conceptual space is prior to any symbolic characterization of cognitive phenomena.

In the domain of artificial vision, we call *knoxel* a generic point in a conceptual space CS (the term knoxel is derived by analogy from *pixel*). A knoxel corresponds to an epistemologically primitive entity at the considered level of analysis. Knoxels are obtained from measurements of the external world performed by the *camera* block, through the subsequent processing of the low-level vision algorithms.

An important aspect of this theory is the possibility of defining a *metric function* in a conceptual space CS. Following Gärdenfors, we maintain that the distance between two knoxels calculated according to such a metric function corresponds to a measure of the *similarity* between the entities represented by the knoxels themselves [25].

A related aspect of this theory is the role of *convex sets* of knoxels in CSs. Gärdenfors proposes the so called *Criterion P*, according to which a *natural category* corresponds to a convex set in some suitable CS. Natural categories are the most informative in taxonomies of real word entities and situations, and are the most differentiated from one another. Natural categories are also the preferred level for reference, they are the first to be learned by children, and categorization at this level is usually faster [23]. According to the *Criterion P*, *betweenness* is significant for natural categories, in that for every pair of knoxels belonging to a convex set (and therefore sharing some features), all the knoxels *between* them also belong to the set itself, and share in their turn the same features.

2.2 Representation of simple objects

2.2.1 Superquadrics

According to our proposal of an intermediate representation level based on conceptual spaces in artificial vision [6, 7], the knoxels are 3D primitive shapes represented according to some Constructive Solid Geometry (CSG) schema (see [20]).

In particular, we adopted *superquadrics* as suitable CSG primitives. Superquadrics are widely used in computer graphics [3] and in computer vision [21, 26, 28]. Various techniques have been proposed for tracking and recovering superquadrics from static and dynamic scenes, even when the objects are difficult to segment or in the presence of occlusions [26, 13, 16, 28, 18].

Superquadrics are geometric shapes derived from the quadric parametric equation with the trigonometric functions raised to two real exponents. The parametric form of a superquadric is:

$$f(\eta, \omega) = \begin{bmatrix} a_x \cos^{\varepsilon_1} \eta \cos^{\varepsilon_2} \omega \\ a_y \cos^{\varepsilon_1} \eta \sin^{\varepsilon_2} \omega \\ a_z \sin^{\varepsilon_1} \eta \end{bmatrix} \tag{1}$$

where $-\pi/2 \leq \eta \leq \pi/2$ and $-\pi \leq \omega < \pi$. The quantities a_x, a_y, a_z are the lengths of the superquadric axes, and the exponents $\varepsilon_1, \varepsilon_2$, are the *form factors*: ε_1 acts in terms of the longitude, and ε_2 in terms of the latitude of the shape. If the form factors are less than 1, then the superquadric assumes a squared shape. For values close to 1 the shape is rounded; greater values tend to generate a cuspidate aspect. Fig. 2 shows the shape assumed by a superquadric by changing the form factors. From top to bottom, ε_1 varies from 0.2 (first row) to 1.5 (last row); from left to right, ε_2 varies from 0.2 (first column) to 1.5 (last column).

Eq. (1) describes a superquadric in canonical form. To describe a superquadric in a generic displacement in 3D space, three center coordinates p_x, p_y, p_z and three Euler angles φ, ϑ, ψ should be added. So, a knoxel $\mathbf{k}$ corresponds to a vector in $\mathbb{R}^{11}$:

$$\mathbf{k} = \begin{bmatrix} a_x & a_y & a_z & \varepsilon_1 & \varepsilon_2 & p_x & p_y & p_z & \varphi & \vartheta & \psi \end{bmatrix}^T . \tag{2}$$

In many cases, it may be convenient to represent a knoxel $\mathbf{k}$ putting in evidence the parameters expressing the shape of the superquadric:

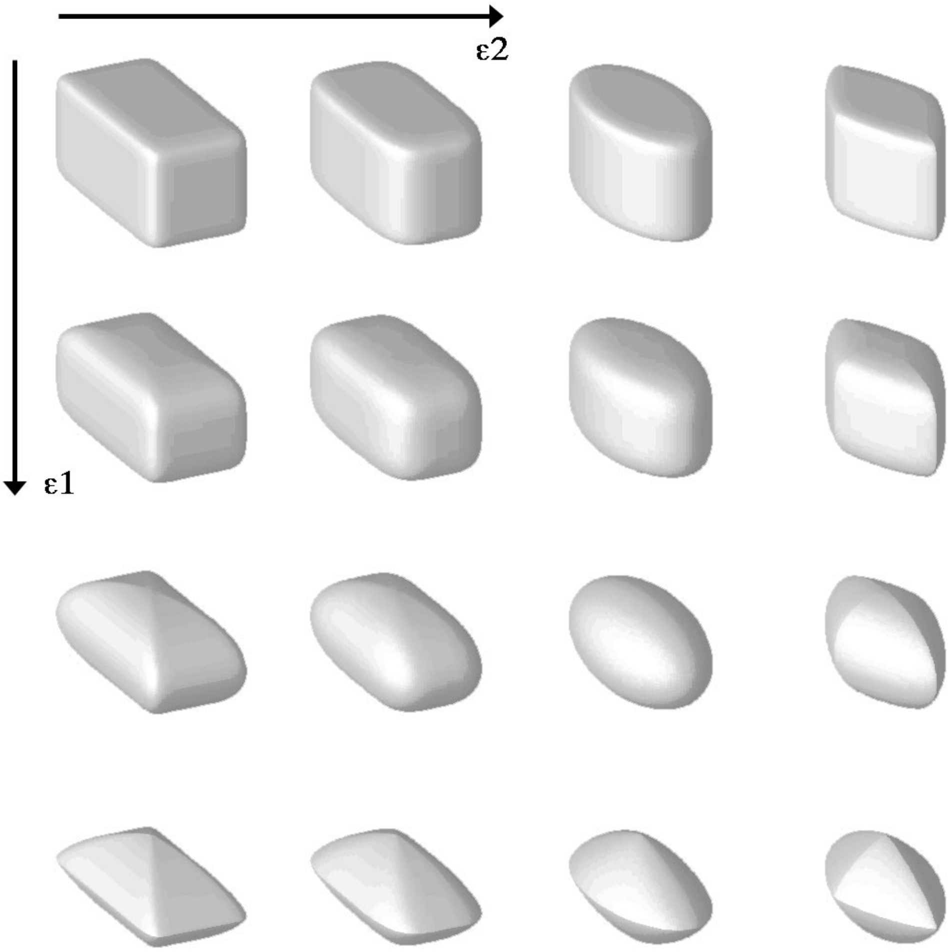

Figure 2: Shapes assumed by a superquadric by varying the form factors ε_1 and ε_2.

$$\mathbf{k}_{shape} = \left[\begin{array}{ccccc} a_x & a_y & a_z & \varepsilon_1 & \varepsilon_2 \end{array} \right]^T , \tag{3}$$

or the parameters corresponding to its spatial displacement:

$$\mathbf{k}_{disp} = \left[\begin{array}{cccccc} p_x & p_y & p_z & \varphi & \vartheta & \psi \end{array} \right]^T . \tag{4}$$

In this way, Eq. (2) now may be written as follows:

$$\mathbf{k} = \left[\begin{array}{c} \mathbf{k}_{shape} \\ \mathbf{k}_{disp} \end{array} \right] . \tag{5}$$

2.2.2 Distance between superquadrics

A *distance* function ds between knoxels based on an Euclidean metric can be defined as follows:

$$ds(\mathbf{k}, \mathbf{k}') = ||\mathbf{k} - \mathbf{k}'|| \tag{6}$$

Taking into account Eq. (5), it is possible to modify ds to obtain a tunable distance measure that differently weights the shape and the displacement of the knoxels:

$$ds(\mathbf{k}, \mathbf{k}', w_s, w_d) = w_s||\mathbf{k}_{shape} - \mathbf{k}'_{shape}|| + w_d||\mathbf{k}_{disp} - \mathbf{k}'_{disp}||. \tag{7}$$

where w_s and w_d are the weights assigned to the distance measures of shapes and displacement, respectively.

Fig. 3 is a pictorial representation of the conceptual space we have adopted: a generic point in CS (a *knoxel*) corresponds to a superquadric along with its displacement in space; concepts as *box* or *cylinder* are represented as sets of knoxels.

We have found [1] that in practical cases the set of knoxels corresponding to "natural" geometric concepts such as *box*, *cylinder*, *sphere*, etc., may be discriminated by a single layer perceptron [19]. This because they correspond to linearly separable sets, i.e., to convex sets in CS, that therefore satisfy the *Criterion P*. So, the choice of superquadrics as knoxels allows to define a conceptual space in which a simple metric function can be defined, and in which simple geometric concepts correspond to convex sets of knoxels. In this way the basic requirements of the conceptual space theory are satisfied.

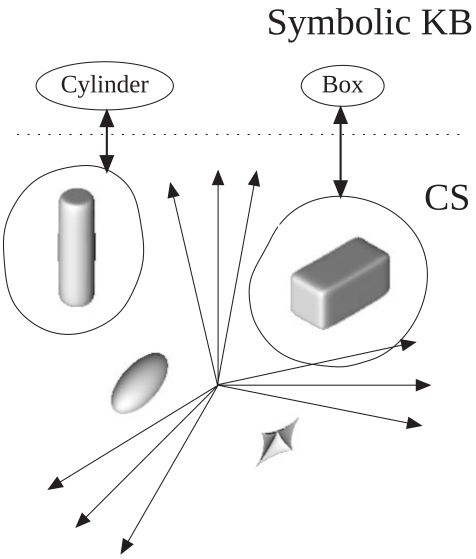

Figure 3: Simple shapes represented in the conceptual space.

2.3 Representation of composite objects

The shape of more complex objects cannot be described In terms of single superquadrics. For example, a chair can be naturally described as the set of its constituents, i.e., its legs, its seat and so on. Analogously, Fig. 4 shows a hammer as composed by two superquadrics, corresponding to its handle and to its head.

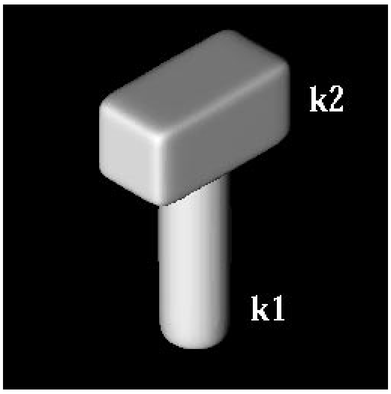

Figure 4: A hammer made up by two superquadrics.

In order to represent composite objects that cannot be described as a single knoxel, we assume that they correspond to sets of knoxels in CS. A generic composite object O is described as the set of knoxels of its n

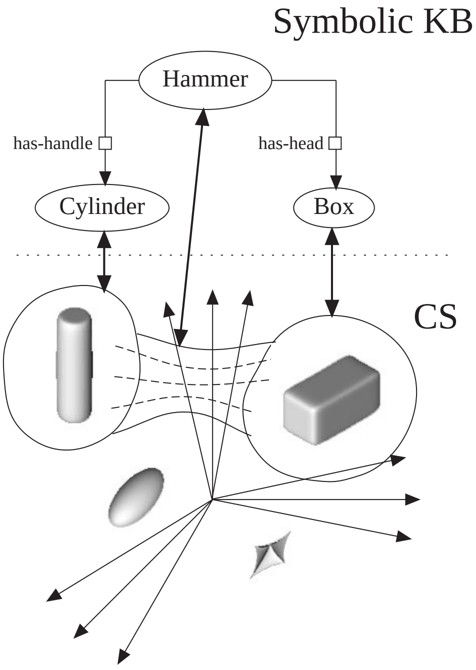

Figure 5: A hammer represented in the conceptual space.

components: $O \equiv \{\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_n\}$. For example, the hammer in Fig. 4 is described as the a of knoxels $\{\mathbf{k}_1, \mathbf{k}_2\}$. In general, every "typical" hammer will corresponds to a suitable pair of knoxels, that correspond respectively to its handle and to its head. Fig. 5 shows a pictorial representation of a hammer in CS: the concept *hammer* is described as a pair of sets corresponding to the hammer's components: the handle and the head.

2.4 Navigating in conceptual space

In order to identify the sequences of knoxels in the CS that correspond to the composite objects described at the level of the symbolic representation, it is possible to imagine a *focus of attention* acting as a light spot that sequentially scans the scene. At the beginning, the focus of attention identifies a zone in CS where a knoxel is expected that matches one of the knoxels of O (say, $\mathbf{k}_1$). If this expectation is satisfied, then the focus of attention searches for a second knoxel of O, say, $\mathbf{k}_2$. This process is iterated until all these expectations are satisfied, and therefore there is enough evidence to assert that an object O is present in the scene.

The movements of the focus of attention may be seen as the movements of a *free-flying* robot navigating in CS. In this sense, the set of knoxels that compose O are target points that may be sequentially reached by the robot; in other terms, the set $\{\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_n\}$ is some sort *motor schema* for the focus of attention in the sense of Arkin [2].

The focus of attention is controlled by two different modalities, namely the *linguistic* and the *associative* modality. According to the linguistic modality, the focus of attention is driven by the symbolic knowledge explicitly stored in the symbolic KB of the system. For example, let us suppose that the system has stored in its KB the description of a hammer as composed by a head and by a handle. When the system recognizes in the scene a knoxel, say $\mathbf{k}_1$, as a possible part of a hammer (e.g., as its handle), it makes the hypothesis that a hammer is present in the scene, and therefore it searches the CS for the lacking parts (in this case, the hammer's head). The black arrow in Fig. 6 shows such a scanning of an hammer by the focus of attention; Fig. 7 shows the corresponding operation in the conceptual space.

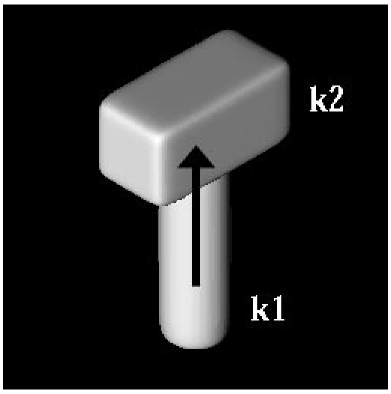

Figure 6: The focus of attention sequentially scanning a hammer.

According to the associative modality, the focus of attention is driven by an associative mechanism based on learned expectations. Let us suppose that the system has seen several scenes where a hammer is present along with a box (as in Fig. 8). As a consequence, the system learns to associate hammers and boxes; when a hammer is present in the scene, it expects to find also a box in the surroundings.

A natural way to implement the focus of attention as it has been described before is to use an *associative memory.* In [6], we describe a mechanism based on a Hopfield neural network with time delayed weights [15] that implements the focus of attention, along with both its *linguistic* and *associative* modalities.

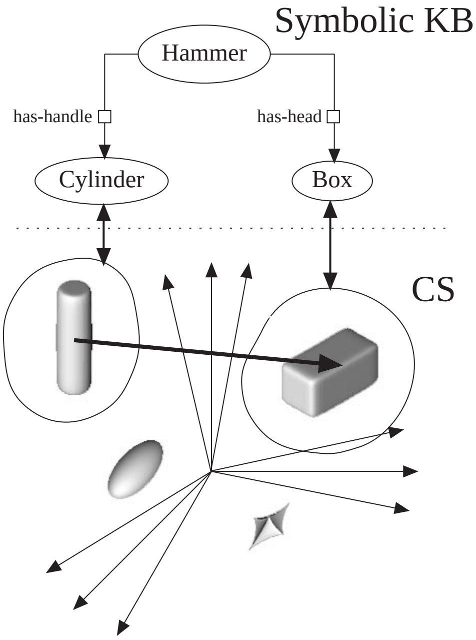

Figure 7: The focus of attention exploring the conceptual space.

3 Comparison with related frameworks

3.1 Feature spaces

Feature spaces are one of the most widely adopted approaches to the problem of the automatic classification of the objects in a scene (see [14] for a tutorial review).

According to this approach, an object is characterized as a vector of feature values corresponding to a point in a feature space in which a metric function is defined. There are many similarities between feature spaces and CSs. In both cases objects are characterized as vectors of components, and a suitable metric function is defined.

However, there are important differences in the basic motivations. The feature space approach is aimed at identifying and classifying objects according to their features [9, 5, 24]. But the adopted features are generally low level and strictly related to the images acquired by the video camera. On the contrary, the conceptual space approach is aimed at the high level interpretation of scenes, and at the cognitive grounding of symbolic representations on perceptual data. The dimensions of a CS are not chosen to be strictly functional for a mere discrimination among objects; rather, they are chosen in order to generate a rich symbolic description of the scene.

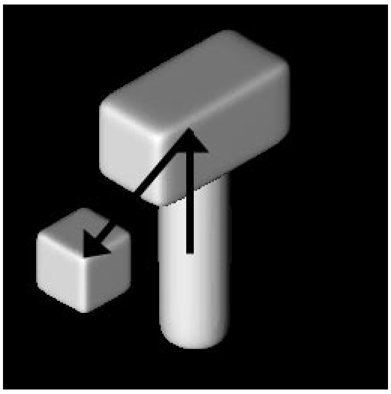

Figure 8: Associative scanning driven by the focus of attention.

Moreover, in the feature space approach objects are treated as a whole, and their internal structure generally is not taken into account. On the contrary, the *CS* approach accounts for the internal structure of the objects in terms of their parts and of the relationships among them. Therefore, from this point of view, the *CS* framework allows far more rich descriptions of scenes and of objects in them.

3.2 Recognition by Components

According to the recognition by component (*RBC*) approach, every object is represented as a set of suitable 3D primitives plus a set of relationships among them; such relationships are usually expressed in the terms of some graph-like formalism [17, 4].

Also the *CS* approach adopts a representation based on 3D primitives. Knoxels are geometrical primitives, and composite objects are represented as the sequence of the knoxels that compose them. In this sense, the *CS* framework can be considered a particular kind of recognition by components, in which the primitive shapes and their relationships are expressed in terms of entities in the conceptual space. By analyzing the top of Fig. 5, it is evident that the description of objects in terms of primitives and of their relationships has an immediate counterpart in *CS*, in the sense that the *CS* representation can be easily mapped on a graph according to the *RBC* approach.

However, if compared to traditional *RBC* representations, knoxel sequences implicitly encode the relationships among knoxels. In this way, an explicit treatment of the relationships as in the *RBC* approach is not mandatory. Furthermore, *CS*s allow for the definition of metric functions for evaluating the similarity between shapes, and this aspect is not immediately available in traditional *RBC* models.

3.3 Chorus of Prototypes

According to the Chorus of Prototypes (*CoP*) approach [10], a generic object O' is represented as the set of the distance measures $d_1, d_2, \dots, d_n$ of O' from some a priori stored objects $O_1, O_2, \dots, O_n$. Such distances are computed by means of suitable neural networks. In this way an object O' is represented exclusively in terms of its relative position with respect to a set of a priori known objects in a suitable space. (see Fig. 9).

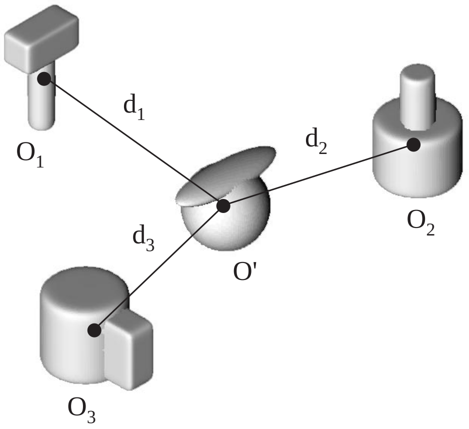

Figure 9: A pictorial representation of the Chorus of Prototypes approach.

Both the *CoP* approach and the *CS* approach are based on metric spaces, where objects are represented in terms of the points of the space. However, in the *CoP* approach the coordinates of the points are not explicitly relevant, in the sense that they do not correspond to features of the objects. On the contrary, in the *CS* approach the coordinates of the knoxels act as effective features, and they have a precise meaning and a precise interpretation at the symbolic level.

In both the approaches the distance between points is significant, and it plays the role of a similarity measure. Thus, both approaches allow for the identification and classification of known objects and for description of unknown objects in terms of their similarity with previously known prototypes.

However, the *CoP* approach does not take into account the parts of the objects: an object, as in the feature space approach, is considered as an unanalyzed whole.

4 Conclusions

Conceptual spaces appear to be a well motivated framework from the point of view of cognitive evidence [12], that generalizes the most common approaches to high level computer vision. In particular:

- It is a feature based representation that is consistent with the use of known algorithms for the identification and classification of objects.

- It is a representation based on parts, allowing for structural description of objects.

- It is a metric representation, allowing for the definition of similarity measures between shapes.

There are many open problems in the research on conceptual spaces as an approach to high level artificial vision. An example is the correct definition of the dimension of conceptual spaces. This point is related to another complex task, namely that of defining suitable metric measures, able to capture relevant relations about shapes.

Nevertheless, we maintain that the conceptual space approach is effective and promising, in that it allows to generalize the main proposal developed in the field of high level computer vision. In addition, conceptual spaces offer a theoretical framework for the development of a conceptual semantics for symbolic representations, that can account for the grounding of symbols on the data coming from the vision system. In this sense, conceptual spaces could give a relevant contribution to a better integration of artificial vision and artificial intelligence techniques in the design of autonomous agents.

Acknowledgements

Author would like to thank Luigia Carlucci Aiello, Christian Balkenius, Peter Gärdenfors, Vito Di Gesù, Marco Gori, Ray Reiter, Pino Spinelli and Achille Varzi for many fruitful discussions concerning the topics of the present work.

This work has been partially supported by MURST "Progetto Cofinanziato CERTAMEN".

References

[1] E. Ardizzone, A. Chella, M. Frixione, and S. Gaglio. Integrating subsymbolic and symbolic processing in artificial vision. *Journal of Intelligent Systems*, 1(4):273–308, 1992.

[2] R.C. Arkin. Integrating behavioral, perceptual, and world knowledge in reactive navigation. *Robotics and Autonom. Systems*, 6:105–122, 1990.

[3] A.H. Barr. Superquadrics and angle-preserving transformations. *IEEE Computer Graphics and Applications*, 1:11–23, 1981.

[4] I Biederman. Human image understanding: Recent research and a theory. *Computer Vision, Graphics and Image Processing*, 32:29–73, 1985.

[5] C.M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford, USA, 1995.

[6] A. Chella, M. Frixione, and S. Gaglio. A cognitive architecture for artificial vision. *Artif. Intell.*, 89:73–111, 1997.

[7] A. Chella, M. Frixione, and S. Gaglio. An architecture for autonomous agents exploiting conceptual representations. *Robotics and Autonomous Systems*, 25(3-4):231–240, 1998.

[8] L. Conway, R.A. Volz, and M.W. Walker. Teleautonomous systems: Projecting and coordinating intelligent actions at a distance. *IEEE Trans. on Robotics and Automation*, 6(2):146–158, 1990.

[9] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis.* Wiley, New York, 1973.

[10] S. Edelman. *Representation and Recognition in Vision.* MIT Press, Bradford Books, Cambridge, MA, 1999.

[11] I.A. Essa. Computers seeing people. *AI Magazine*, 20(2):69–82, 1999.

[12] P. Gärdenfors. *Conceptual Spaces.* MIT Press, Bradford Books, Cambridge, MA, 2000. in press.

[13] A. Gupta and R. Bajcsy. Volumetric segmentation of range images of 3D objects using superquadric models. *Computer Vision, Graphics and Image Processing: Image Understanding*, 58(3):302–326, 1993.

[14] B.K.P. Horn. *Robot Vision.* MIT Press, 1986.

[15] D. Kleinfeld and H. Sompolinsky. Associative network models for central pattern generators. In C. Koch and I. Segev, editors, *Methods in Neuronal Modeling*, Bradford Books, pages 195–246. MIT Press, Cambridge, MA, 1989.

[16] A. Leonardis, A. Jaklic, and F. Solina. Superquadrics for segmentation and modeling range data. *IEEE Trans. Patt. Anal. Mach. Intell.*, 19(11):1289–1295, 1997.

[17] D. Marr. *Vision.* W.H. Freeman and Co., New York, 1982.

[18] J. Maver and R. Bajcsy. Occlusions as a guide for planning the next view. *IEEE Trans. Pat. Anal. Mach. Intel.*, 15(5):417–433, 1993.

[19] M. Minsky and S. Papert. *Perceptrons.* MIT Press, Cambridge, MA, 1969.

[20] M.E. Mortenson. *Geometric Modeling, 2nd ed.* J. Wiley and Sons, New York, 1997.

[21] A.P. Pentland. Perceptual organization and the representation of natural form. *Artif. Intell.*, 28:293–331, 1986.

[22] R. Reiter. Knowledge in action. Logical foundations for describing and implementing dynamical systems. Technical report, Department of Computer Science, University of Toronto, CA, 1999.

[23] E. Rosch. Cognitive representations on semantic cathegories. *Journal of Experimental Psychology: General*, 104:192–233, 1975.

[24] B. Scholkopf, C. Burges, and A. Smola, editors. *Advances in Kernel Methods: Support Vector Learning.* MIT Press, Cambridge, MA, 1999.

[25] R.N. Shepard. Toward a universal law of generalization for psychological science. *Science*, 237:1317–1323, 1987.

[26] F. Solina and R. Bajcsy. Recovery of parametric models from range images: The case for superquadrics with global deformations. *IEEE Trans. Patt. Anal. Mach. Intell.*, 12(2):131–146, 1990.

[27] S. Ullman. *High-level Vision.* MIT Press, Cambridge,MA, 1996.

[28] P. Whaite and F. Ferrie. From uncertainty to visual exploration. *IEEE Trans. Patt. Anal. Mach. Intell.*, 13(10):1038–1049, 1991.