OBJECT REPRESENTATION AND SPATIAL KNOWLEDGE: AN INSIGHT INTO THE PROBLEM OF MEN-ROBOTS COMMUNICATION

G. Adorni, A. Boccalatte, and M. Di Manzo

Instituto di elettrotecnica
Universita di Genova, Italy

ABSTRACT

In this paper we are concerned with the problem of a "natural" communication between a human operator and a robot operating in a physical world, like a room.

At first we investigate a formalism for describing objects and spatial relations between objects. This formalism must provide not only a suitable basis for representing spatial knowledge and making spatial inferences, but also a practical interface with the procedures of visual analysis, even if the goal of this paper is the description of scenes by means of natural languages more than the recognition of objects. Then we analyze the translation into this formalism of words and simple sentences that implies perceptions of the world and physical movements.

Thus, for instance, the problem of a proper representation of actions like looking at something or moving an arm or grasping an object is investigated.

Finally we give an insight into the problem of inferences; a classification is given and some particular cases are briefly discussed.

All these problems are discussed with reference to a real world, trying to avoid solutions that could perform only a restricted block of world.

RÉSUMÉ

Cette communication porte sur le problème des communications "naturelles" entre un opérateur humain et un robot fonctionnant dans un cadre physique, à l'intérieur d'une pièce par exemple.

Nous recherchons d'abord sous quelle forme décrire les objets et les relations spatiales entre les objets. Ce formalisme doit non seulement servir de base à la représentation de la connaissance spatiale et aux références spatiales, mais servir aussi d'interface fonctionnelle avec les procédures d'analyse visuelle bien que notre objectif soit beaucoup plus la description d'images au moyen de langages naturels que la reconnaissance des objets. Nous analysons ensuite la transposition, selon ce formalisme, de mots et de simples phrases qui décrivent des perceptions du monde et des mouvements physiques.

Cette méthode permet d'étudier, par exemple, le problème de la représentation adéquate des actions, comme le fait de regarder quelque chose, de bouger un bras ou de saisir un objet.

Enfin, nous abordons le problème des inférences; nous fournissons une classification et présentons une courte étude de quelques cas.

Nous étudions tous ces problèmes en fonction de la réalité de manière à éviter les solutions qui seraient applicables uniquement dans un contexte restreint.

INTRODUCTION

A robot which must operate in the physical world to perform some task must integrate perception with knowledge and moving and manipulating capabilities. In the field of perception only vision has been seriously approached up to now and most of works on computer vision are oriented to the identification of an object in a scene more than to investigating relations among vision, language and understanding. However, as Waltz points out (14), we think that the problem of scene understanding is mainly a problem of scene generation. If a robot is able to build a "mental image" of a scene, starting from its description by words, and to verify its appropriateness, the recognition of objects can be performed by comparing his mental image to the scene model produced by a vision system; the recognition process can be much more goal oriented, because the robot will pay attention only to those features which are significant with respect to its actual task. So, if for instance a man issues the order GRASP THE PENCIL ON THE TABLE, the robot will look only the objects which correspond to its representation of a table and a pencil.

In this paper we face the problem of vision from a knowledge viewpoint. Discussing the features of a model oriented to the representation of objects and spatial relation between objects (1, 3,9,12,13) we will introduce the model for object description through a series of examples which show that all the details of this model properly express some spatial relation. Then we will give few examples on how spatial relations can be formalized, and at least we will conclude with a short discussion of the problem of inference (6,7,10).

THE CHOICE OF PRIMITIVES

Our representation of knowledge is based on the choice of a suitable set of primitive concepts. This choice is quite arbitrary because up to now we have not a proper methodology to evaluate the effectiveness of a conceptual model (11). A qualitative criterion can be based on the capability of supporting a set of inference rules. We could define, for instance, a primitive ABOVE to describe a scene in which *the chandelier is* ABOVE *the table*, and knowing *the table is* ABOVE *the carpet*, we could use an inference rule to deduce that *the chandelier is* ABOVE *the carpet*. However, if we define another primitive, say BEHIND, we should likely use a different rule to infer that *George is* BEHIND *John* from *George is* BEHIND *Paul* and *Paul is* BEHIND *John*, because the meaning of BEHIND is more ambiguous than the meaning of ABOVE, as we will see in the following, and also the orientation of the object must be considered in this case. A large set of inference rules is difficult and expensive to handle and therefore a conceptual model is as better as smaller is the required number of rules. From a computational point of view a very natural way of describing the position of an object is to use a system of coordinate axes. If we are able to transform all the linguistic relations, like *above, under, behind, inside* and so on, into quantitative geometrical relations among the coordinates of some points of the involved objects, a number of inferences can be made by means of few, simple and very general rules which can be directly derived from the analytical geometry. Hence the goal of describing objects and spatial relations by means of a single, non-redundant n-tuple of coordinate axes is very appealing. Unfortunately it seems to be quite far from the psychology of language. In fact, in most cases the position of an object is defined relatively to the position of another one, and the reference object can change within the same sentence as for instance in *the car is parked on the right side of the building which is behind the station.* Moreover, the relation *behind* can be easily described by means of a set of cartesian coordinates, but for the concept of *turn right* a polar system of coordinates is more suitable. If these properties of language are not taken into account, the translation of some relations becomes very cumbersome. A second and even more difficult problem is the description of the structure of objects, both from a static and a dynamic point of view. The knowledge of object structure is often intimately related to our capability of understanding the meaning of a spatial relationship; for instance, the meaning of the sentence *the cat is under the car*, is clear, even if it can depend on the state of the car, moving or parked; on the contrary, the sentence *the cat is under the wall* is not clear, unless the wall is a crashed one or it has a particular shape.

In the following, we will consider some examples of increasing complexity, in order to introduce step by step, all the features of our model of spatial knowledge.

FROM SIMPLE TO COMPLEX RELATIONS

We start with the anaysis of some simple relations, where "simple" stands for "requiring only a simple description of objects". Two simple relations belonging to this subset are, in some cases, *near to* and *far from*. The meaning of the sentence *the house is near to the*

station can be formalized saying that " if X is the distance between a point P of the house and a point Q of the station, then X is less than a target distance L" (1). The distance X can be evaluated referring to any arbitrary system of coordinates, and the amount of knowledge required to give a default value to L is very small, because only a rough evaluation of the typical dimensions of a house and a station is necessary. Hence, objects can be described simply as blocks.

Another sample case is the relation *above*. The concept of "verticality" is much more absolute than all other spatial concepts, because it is related to our sense of pound. Therefore the concept of ABOVE can be always referred to as an absolute vertical axis Z, and a sentence like *the chandelier is above the table* means that "at least a point P ∈ chandelier and a point Q ∈ table exist such that X (P) = X (Q), Y (P) = Y (Q) and Z (P) > Z (Q)" where X (P) represents the value of the coordinate X of P with respect to a cartesian triple. This formulation is independent from the choice of the system of coordinates and the structure of objects, and so we could describe the scene by means of an unique cartesian triple, with the Z axis indicating the vertical direction.

More knowledge about the structure of objects is required by relations like *behind*. If we say that *the man is behind the column*, this clearly means that the column is between the man and an observer who is looking at the man; but if we say that *the man is behind the car*, the position of the man could be related to the car, instead of the observer, because the car has its own front and back. Expressions like the previous one introduce the need of inserting into the description of an object the presence of some privileged part or direction (5). This need can be satisfied by associating to each object a particular cartesian triple, whose X axis indicates the privileged direction of the object, if any, from the back to the front. This kind of relations can be translated referring always to the system of coordinates owned by a particular object; a further translation using an unique absolute system of coordinates can be made only at the inference level.

Let now consider a sentence like *the pen is near the edge of the table*. In this sentence some structural knowledge about the table is necessary: for instance, we must know that the top of the table is typically a part of a plane limit by a curve. Hence, a more detailed description is needed. At this level objects can be described by means of generalized cones which are solid objects generated by a plane section which moves along an arbitrary curve representing the axis of the cone(2,4). The section is always perpendicular to the axis, and its shape is fixed, even if its dimension can vary. The cone is limited by the intersection with two boundary surfaces. The generating section is described by a function $\rho(\theta)$ which gives the distance of the section boundary from the center for each value of the angle θ. Now the table can be described by a cone with a vertical axis having the top delimited by a horizontal plane; so the previous sentence means that "if P is a point of physical contact between the pen and the table, then the distance between P and the axis of the cone on the upper cone surface is less than but near to $\rho(\theta)$,$\rho(\theta)$ being the description of the upper section boundary".

The structure of an object can be further detailed by means of a number of connected cones, but before doing it let us discuss one more kind of relations. Let us consider, for instance,the sentence *the house is three miles after the bridge along the road to Canterbury*. Here the absolute position of the house relatively to a given system of coordinates can be known only if the trajectory of the road is known; if it is not, we can only state the curvilinear coordinate of the house along an unknown curve. Therefore, we must introduce the problem of the identification and description of trajectories. A perfectly known trajectory can be described by a set of parametric equations or something equivalent. However, sometime such a deep detail is not possible or not useful; in thiscase the trajectory can be approximated by stating the origin, destination and eventually a number of intermediate points. This partial description isvery common in human knowledge; for instance, we know that along the railway from Genoa to Rome, there are towns like Pisa and Leghorn, and that, coming from Genoa, we find Pisa before Leghorn, and it takes about two hours to go from Genoa to Pisa and so on, but very few people know exactly the trajectory of the reailway from Genoa to Rome.

OBJECT DESCRIPTION

We can now summarize all the features of the model we used to describe objects and relationships among objects. We said that an object is defined by one or more connected cones (8). Before discussing of connections, let us give a further insight into the problem of cone definition.

A cone is a simple, monolithic object defined by the quadruple <T,C,Ⓗ, S>, where T is a cartesian triple, C is a parametric description

of the cone axes relatively to T, H is the law which gives the orientation, relatively to T, of all the cartesian triples local to each point of the axes, and S is the description of he boundaries of each section (fig. 1).

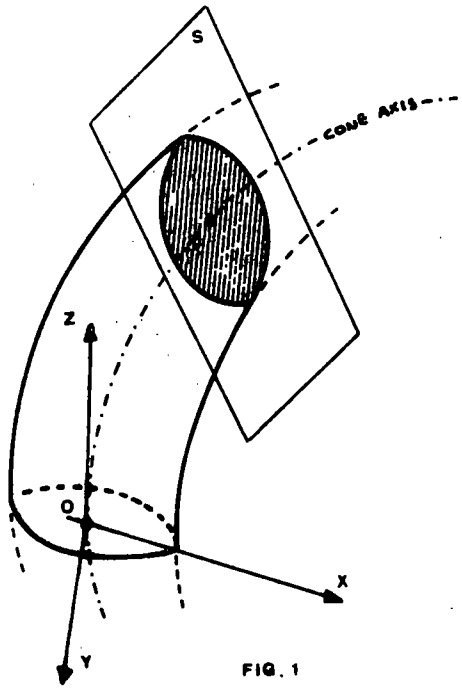

FIG. 1

The purpose of the local triples, one for each point of the cone axis, is to define the alignment of sections; in fact, sections are described by a function S (θ, k), where k is a curvilinear coordinate, and θ is an angle on the plane of section; the point in which θ is equal to zero is fixed relatively to the local triple. This kind of description is quite general; the only limitation is given by the "regularity" of S, which imposes the invariance of the form of the generating section and its convexity. However, this limitation, which avoids very cumbersome description, can be overcome, when it is really needed, using cone connections; for instance, a local anomaly, like a hole, can be described connecting to an empty cone. In many cases we think that only a subset of the features of this definition of a generalised cone are truly necessary; for instance, we expect that a large number of objects can be easily described by means of cones with a straight axis along the Z axis of the triple T. However, to describe a river or a road it is useful to define a cone having an axis which is a curve lying on the horizontal plane.

Cones may be connected by means of rigid connections or points. A rigid connection between a cone A and a cone B is seen as a physical contact between a terminal point Q of the axis of A with an arbitrary point P of the surface of B. The orientation of the connection is defined giving the angles between the axis of A, in the point P, and the local cartesian triple of the section of B which contains the point P (fig. 2).

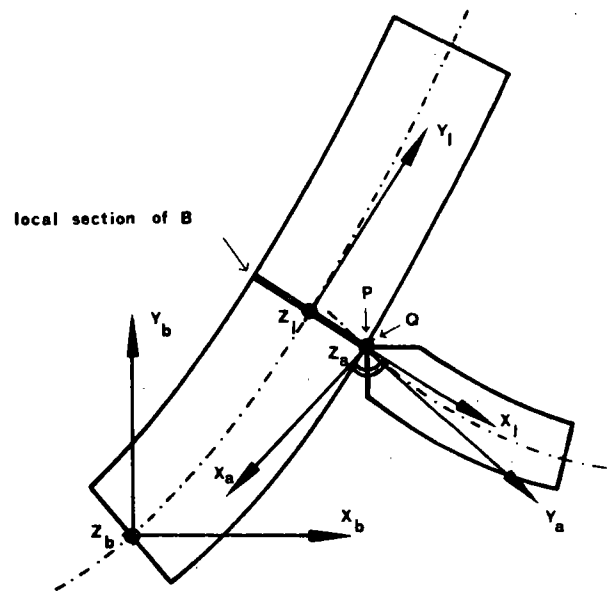

FIG. 2

The definition of a joint is more complicated, because it must avoid the movement to be constrained by physical contacts between the surfaces of objects due only to the approximation of the model. Let consider, for instance, a snake shaped as a number of jointed cones, as in fig. 3.

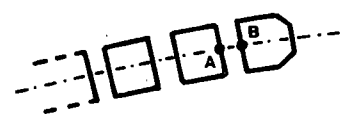

FIG. 3

If we joint the head with the first segment of the body by superimposing points A and B, every rotation of the head relatively to the body leads to an unnatural interpenetration of two pieces of the snake. A possible solution is that of adding a triangular termination to cones which must be jointed, as shown in fig.4.

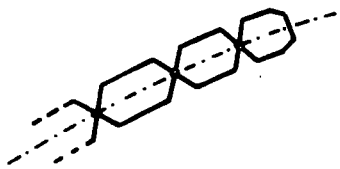

FIG.4

Now rotation is limited only by the sharpness of the two involved terminations. However, this solution is not completely satisfactory because it requires an explicit definition of cone terminations and leads "holes" to the surface. An improved solution consists of using a specific jointing element, that can be seen as a cone generated by a constant plane section flowing along an axis consisting of two jointed segments. A two dimensional case is shown in fig. 5.

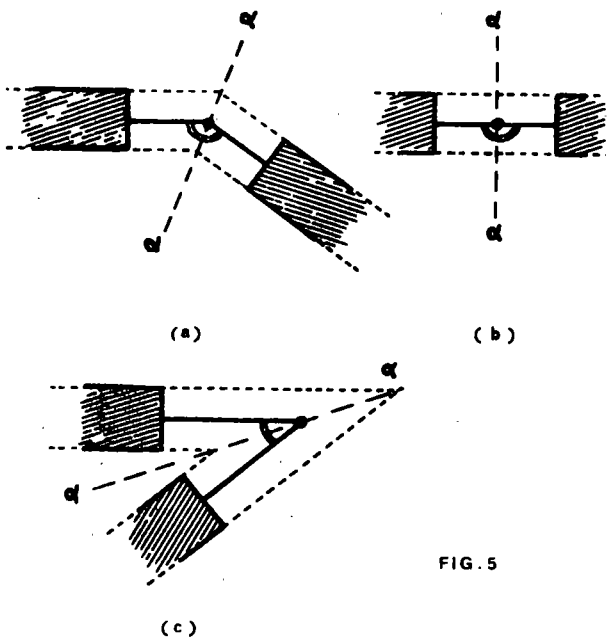

FIG.5

The angle between the two segments in the point of junction is bi-sectioned by a plane α; the surface of the joint is built as a connection of two cones, the left half-joint and the right half-joint, each consisting of a normal cone terminated by a plane α. The shape of the section is arbitrary, so we can have jointing elements like that shown in fig. 6, which is useful to model the human elbow.

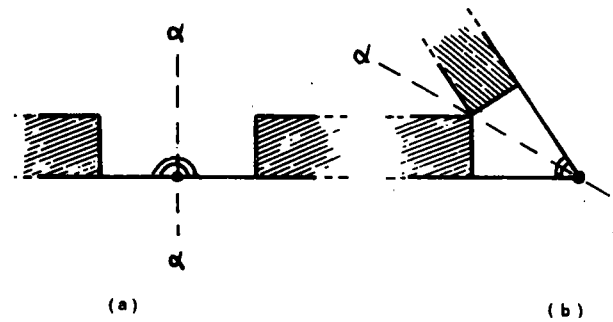

FIG. 6

A jointing element is completely defined by its length and its section. Tortional movements can be allowed, but in this case the section shape must be carefully designed if we want to avoid boundary discontinuities on plane α. Articulation constraints result from the physical contact of the two cones connected to the joint; further, constrints can be explicitly stated. A second kind of joint is a connection like that existing between a desk and a drawer; in this case the point of physical contact can flow on the surface of an object.

MORE ABOUT RELATIONS

Owing to lack of space an exhaustive analysis of the conceptualization of spatial relations it is not possible, we will then limit ourselves to few examples. A more complete description can be found in (1).

At first, let us consider a simple relation like A *is behind* B. If B has not its own front and back, we said that the meaning of the sentence is A *is between B and an observer who is looking at* B. This interpretation can be formalized as follows: "if $\theta(P) = \theta(Q)$, P being a point of A and Q a point of B, relatively to a system of coordinates associated to the observer, then $\rho(P) > \rho(Q)$". The meanings of θ and ρ are visualized in fig. 7.

If B has its own front and back the concept of *behind* can be referred to as the back of B.

In this case (fig. 8), it is easier to use a description in terms of the coordinates X and Y of the cartesian triple associated to B. Then A *is behind* B means: "there exist at least one point P ∈ A and one point Q ∈ B such that Y (P) = Y (Q) and X (P) < X (Q)".

Note that the concept of *behind* is strictly related to the concept of "horizontality". If we associate the cartesian triple of a man to the cone which describes his trunk, we can easily say what means to be behind a standing man, but if the man is lying on his back, we will probably refer to the same relative position as *under*; *the floor is under the man*, in this case, and not *behind* .

The second example is a dynamic one, like *turning right*. The act of TURNING is a physical movement, which can be expressed by a primitive like PTRANS (11). We fill the "directive case" of PTRANS by a trajectory, as it has been previously defined. In this particular case the trajectory is simply described by an origin point S and a destination point D; each point has an associated cartesian triple, T (S) and T (D) respectively. TURNING is described assuming that the planar angle between T (S) and T (D) is about 90°.

The third example is the relation *inside*. If we say that A *is in* B, we must suppose at first that the sections of B have an internal and an external boundary; in many cases this kind of objects can be described by two coaxial cone, a full external cone and an empty internal one (look, for instance at a house, a box, a bottle etc.). Then A *is in* B can be formalized as follow: " if P is a point of A lying on the plane of a section of B, then the distance between P and the axis of B in that plane must be less than the internal boundary of the section of B for the same angle". In fig. 9 the case of *a book in a drawer* is shown.

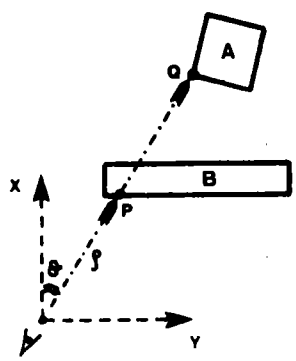

FIG.7

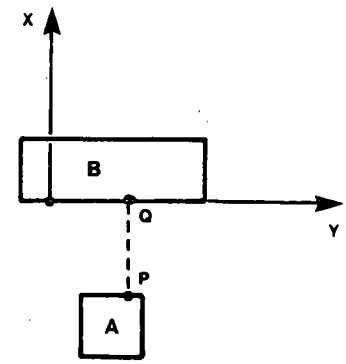

FIG.8

A GLANCE TO THE PROBLEM OF INFERENCE

When we imagine a scene described by words,we add a number of details which are not explicitly stated, but only "reasonable". For instance, to build a scene from the sentence *the man sitting on the chair grasped the pen near the book on the desk*, we must ask two questions like " what is the true meaning of *on*?"; "what are the structures of a man, a chair, a book, a desk and a pen, and their reasonable dimensions?"; "what is a reasonable value of the distance involved by the relationship *near*?"; "what is the position of the man relatively to the desk (he must be able to grasp the pen without leaving the chair)?"; "and the articulation of his body?"; "what are his movements while grasping the pen?". Sometimes these questions can be asked using only spatial knowledge,even if usually we use also a lot of general knowledge; for example, if the man owns the desk, we will imagine him to be *behind* the desk, otherwise he will be likely in front of it. This use of general knowledge is beyond the scope of this paper; in the following we will only briefly discuss some of the inferences wich are more usual in the process of building a scene ,with the aim of introducing problems more than suggesting solutions.

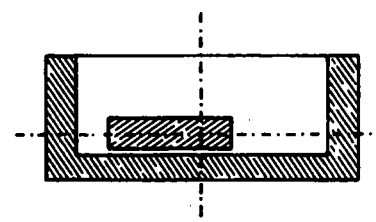

FIG.9

At first, we try a rough classification of inferences, which is probably far from being exhaustive. We can have:

1. Position inferences, to specify the true spatial meaning of ambiguous relations like *on*.
2. Quantification inferences, to give quantitative values to coordinates which are known only through relations like *greater than* or *less than*.
3. Location inferences, to deduce what a relation exists between two objects, starting from known relations between these objects and other ones.
4. Trajectory inferences, to deduce the trajectory of a moving object from knowledge about its origin and/or destination, or vice versa.
5. Structure inferences, to deduce structural characteristics of an object (shape or articulation) from known relations between this object and other objects.

In the following subsection these types of inference are separately analyzed.

Position inference. Some relations, as expressed by the language, are very ambiguous. A typical example is *on*. Usually A *is on* B means that a supporting force is applied by B to A, and therefore there is a typical contact between the two objects. What it really means, from a position viewpoint, depends on the involved objects. If A is much smaller than B, it can be anywhere on the surface of B (look for instance at the sentences *the picture on the wall* and *the fly on the glass*); otherwise *on* usually means *above*, as in the sentence *the book on the glass*. The correct interpretation depends also on other physical characteristics of objects, like their capability of sticking to vertical surfaces; for instance we will not give the same meaning to the sentences *the fly on the window* and *the cat on the window*. Therefore a standard inference for A *on* B can be:

a. *before* (or *behind*), if B has a vertical surface which is large with respect to A, and A can be supported by B; special cases like *the flag on the staff* can be handled considering usual positions of objects;

b. *under*, if B is the lower horizontal surface of an object, as, for instance, the ceiling of a room, an A can be supported by B in such a reverse position (*a fly on the ceiling*);

c. *above*, otherwise.

Quantification inference. The conceptualization of spatial relations usually leads to inequalities between coordinates, but does not give quantitative values. There values can be inferred using knowledge about the typical shape and dimensions of objects. If objects are comparable, their dimensions can give the order of magnitude of distances; for instance, when we say *the building behind the chuch* we expect the distance between the building and the church to be of the order of tens of metres, while in the sentence *the glass behind the bottle*, the expected distance is of the order of tens of centimetres. If an object is much larger than the other one we take the bigger one as reference (*the pen is on the floor behind the table*). Some relations, however, allow to assume the smaller object as a reference; for instance, the distance involved in the sentence *the fly is near the edge of the table* is expected to be of the same order of dimensions of the fly or, alternatively, at least one order of magnitude lower than the dimensions of the table. This usually true whenever a position is described relatively to the boundary of an object. If a dimension of an object is much larger than the other ones, usually this is discarted; when we say *the house is near the river*, we consider the width of the river, and not its length.

In the case of movings, distances can be evaluated also with reference to the total length of the trajectory or time; for instance, the proximity to destination can be evaluated in a very different way according to whether the traveller is flying or walking.

Location inference. In geometrical primitives used to conceptualize spatial relations allow to deduce new relations from a set of known ones simply by mathematical operations like coordinates evaluation, changes of systems of coordinates and so on. However, there is a number of situations which require specific inferences.

1. Reference identification. Sometimes the reference is not explicitly stated; in the sentence *a man on the corner of the square see the house on the right of the church*, we can assume as a reference the man or the church, and obviously the resulting position of the house is not the same in both cases.
2. Direction identification. If we say that *the house is before the church that is behind the Town Hall*, it is not possible to identify the position of the house without making assumptions about the

orientation of the church.

3. Constraints identification. If we say *the coin is under the wardrobe*, we implicitly say also that *the coin is on the floor*, because usually a wardrobe is supported by the floor of the room. Implicit relations of this kind must be identified because they impose constraints, for instance, on the evaluation of distances.

Trajectory inference. Movements are made along trajectories, and sometimes we are interested in knowing fact about these trajectories. Depending on the characteristics of the moving object, trajectory can be variously constrained, and sometimes there is only a very little number of available paths. In this case the chosen trajectory can be identified by means of some knowledge about one or more intermediate crossed points. If constraints are not so strong, the trajectory can be chosen as the shortest path which is consistent with the moving capabilities of the object. For instance, a man moving in a room from the door to the window will probably walk around a table, while a cat will jump on it. Deductions about origin and destination are a particular case of trajectory inference. When moving along heavily constrained paths, we can assume that the origin and destination are known points of the path itself, chosen using informations as for instance the elapsed time or the forseen arrival time or some more general knowledge about the goals of motion. If the movement is loosely constrained, some limited deductions can be made by extrapolation of the actual direction; this kind of inference is useful to forsee the future position of an object in order, for instance, to avoid moving obstacles.

Structure inference. If we say that *the cat is under the wardrobe*, we implicitly assert that there is enough place for a cat between the floor and the bottom of the wardrobe. If the system knows that there are two possible structures for the wardrobe, namely with legs or without, it can infer that in this case the wardrobe must have legs. This kind of inference is used to choose among alternating descriptions of an object; for instance, the sentence *the man is looking for a pen in the drawer*, allows us to infer something about the position of the drawer with respect to the rest of the desk. A very complex strategy of inference is used to deduce facts about articulated objects, like human body. Truly, a good representation of a walking man would require a simulator of human movements which takes into account a lot of physical constraints, as for instance equilibrium problems. In many cases, however, a satisfactory description of the scene can be achieved simply by storing knowledge about few canonical positions, like standing, sitting, lying and so on.

CONCLUSIONS

The problem of scene description in natural language has been only sketched in this paper. Even if more detailed analysis of some particular aspects can be found in the literature, vision is yet a substantially open problem. A lot of work is necessary to answer to a number of basic questions, as, for instance, how to represent abjects with variable shape like a sheet, how to implement and use properly the human capability of finding similarities between shapes, how to use knowledge about the expected goals of an object (of a proper type, of course) to infer its future movements, how to link scene generation with scene analysis and so on. This problem has been neglected for a long time, but now it is receiving more and more attention, and there is an increasing number of research groups currently working on this or on related topics. This interest is justified only by the impact that an integrated vision-manipulation system can have on the applications of robotics, but also by awareness that language is intimately related to the perception of the physical world, and there is a large number of linguistic problems that can not be solved if this perception capability is not achieved.

REFERENCES

1. G.Adorni, M. Di Manzo
Considerazioni su un modello concettuale per la rappresentazione di relazioni spaziali. CNR,ITD-045,I980 (in Italian)

2. G.J. Agin
Representation and description of curved objects. Stanford Artificial Intelligence Project Memo AIM - 173, October, 1972

3. M. Bierwisch
Some semantic universal of German adjectivals. Foundations of language, III, 1-16, 1967

4. T.O. Binford
Visual perception by computer. IEEE Conference on Systems and Control, Miami, December, 1971

5. M.C. Corbalis, I.L. Beale
On telling left from right. Scientific American, 224(3), 94-104, 1971

6. E.J. Crothers
Paragraph structure inference. Ablex Publ. Corp., Norwood, New Jersey, 1979

7. W. Hoeppner
Reprasentation strukturen und inferenzen für zusammensetzte objekte. Projektgruppe simulation von sprochverstchen, Universität Hamburg, Bericht Nr. 15, 1980 (in German)

8. R. Nevatia
Computer analysis of scenes of 3-d curved objects. Birkhauser Verlag, Basel, 1976

9. D. Parisi, C. Castelfranchi
Analisi semantica dei locativi spaziali. III Convegno Internazionale di studi , Roma, Maggio, 1969 (in Italian)

10. C. Rieger
Conceptual memory. Ph.D. Thesis, Computer Science Department, Stanford University, Stanford, Calif., 1975

11. R.C. Schank, B. Nasm-Webber (eds). Theoretical issues in natural language processing ACL Arlington,Va, 1975

12. N.K. Sondheimer
Spatial reference and natural language machine control. Int. J. Man-Machine Studies, 8, 329-336,1976

13. D.L. Waltz
Relating images, concepts and words. Proc. NSF Workshop on the representation of 3-d objects. University of Pennsylvania, Philadelphia, 1979

14. D.L. Waltz
Generating and understanding scene descriptions. University of Illinois, Urbana, working paper, 24, 1980