*COLING 82, J. Horecký (ed.)*
*North-Holland Publishing Company*
*© Academia, 1982*

# COGNITIVE MODELS FOR COMPUTER VISION *

G.Adorni A.Boccalatte M.Di Manzo

INSTITUTE of ELECTROTECHNICS
UNIVERSITY of GENOA
GENOA, ITALY

This paper is focused on the relations existing between language and vision. Its goal is to discuss how linguistic informations about objects,shapes,positions and spatial relations with other objects can be integrated into a cognitive model tailored to spatial inferencing operations.

## INTRODUCTION

A common approach to the problem of scenes interpretation is to generate hypothesis about the position and size of objects and try to use these expectations to guide the search for picture areas which exhibit the expected features |4,8,15|. But where this expectation came from? If a robot operates in a known environment,expectations can be self-generated on the basis of built-in knowledge and previously experienced situations. Another very common source of informations can be some kind of external input, often based on natural language communication. A piece of conversation as "look for the pencil","where?", "on the table" conveys a lot of informations about the presence of a reference object (table) and the characteristics of a surface (top of table) which must be located in order to restrict the search for the target object (pencil). To take advantage of these linguistic information sources we must be able to extract from a qualitative expression like "on the table" all those quantitative constraints which are relevant from a geometric modelling point of view |2|. These problems could seem much more related to the generation of visual analog representations than to the understanding of a scene; but what does it mean exactly to "understand" a scene? When we analyze a scene, we use a lot of not geometric knowledge; we are not surprised to find smoken cigarettes into an ashtray, and a glance is enaugh to classify them, but we could have some troubles to recognize that it contains a company of goldfishes, and this surely not only because of geometric constraints! Therefore the processing of visual knowledge must be based on cognitive models that are able to handle different kinds and sources of informations, and in this sense we feel that there is not a clear cut between scene analysis and scene generation |14,16|.

In the following we will deal mainly with the representation of objects and the formalization of spatial relationships, trying to point out how linguistic informations can be related to visual ones.

* Work supported by Italian National Research Council under grant 80.01142.07

## OBJECT DESCRIPTION AND SPACE MODELLING

The knowledge of the structure of an object 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 may 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 crashed or it has a very particular shape. Every object modelling technique must deal at least with the following issues |6,7| :

1. Object must be described at several levels of detail.To understand the sentence "put the chair near the table" only a rough definition of chair and table dimensions can be sufficient,while to build a model of "a man sitting on a chair" a more sophisticated knowledge abaut the structure of a chair and a man is requested.
2. The articulation of movable object parts must be properly described. The sentences "open the door" and "open the drawer" have different geometric meanings because the movements of doors and drawers usually obey different rules.
3. Characteristic features of objects must be pointed out.Often these features are free surfaces,as the top of a table,in canonical positions.The recognition of a feature allows the generation of hypothesis about the presence of an object.
4. Typical relations between objects must be described.When we look for a pencil we do not start analyzing a wall or a window,but we look at first for a table or some other piece of furniture in which or on which it is reasonable to find a pencil.

Our conceptual definition language allow the definition of lines,surface and solid objects.Solid objects are described by means of GENERALIZED CONES |9,10|,at several levels of detail.Cones can be interconnected by means of fixed or movable points,with arbitrary constraints on rotations and shifting.Specific jointing elements are defined to properly describe the surface of an articulated object;so we can correctly answer to the question: "is the fly on the snake?" indipendently of how the snake is actually coiled.More details can be found in |1|.

From a computational point of view,the use of a system of coordinated axes represent a very natural way to describe the position of an object. If we are able to transform linguistic relations into quantitative geometrical ones,the well knows methodologies of analytical geometry can be used as a simple,general purpose set of inferencing rules.Hence the goal of describing objects and spatial relations by means a simple,non redundant n-tuple of coordinated axes is very appealing.Unfortunately it seems quite far from the psychology of language |2|.
Therefore we associate a redundant FRAME OF REFERENCE (FOR) to every object,consisting of :

- an axis,Z,having direction of the "major" axis of the cone.Two points are specified on it,$Z_{min}$ and $Z_{max}$,corresponding to the extremities of this major axis;
- a point O,on the Z axis,which is the origin of the frame;
- an axis,X,orthogonal to Z,that specifies a further privileged direction of the object;this axis is definible only for some objects(eg.

a man) in which a front and a back can be distinguished.Objects for which the X axis is definable are to be called CLASS 1 objects; those for which the X axis is not definable (eg. a pole) are to be called CLASS 2 objects;

- an axis,Y,orthogonal to X and Z.The Y axis is obviously not definable for class 2 objects;
- a radial coordinate $\rho$ whose origin is at O;
- the coordinates $\rho$ and $\theta$ specified on the X-Y plane;
- a curvilinear coordinate **t** originating at point O.

The use of cones simplifies the FOR;it allows a homogeneous representation of an object shape and of its spatial relations with the external world;it proves particularly useful in situations like "the ball is inside the box".

## SPATIAL RELATIONS BETWEEN OBJECTS

Let's now analyze some spatial relations between objects,in order to discuss how they can be translated in terms of geometrical primitives. Spatial relations involving the Z axis generally use a "major" axis perpendicular to the earth surface;this is the only absolute reference used in language perhaps because the concept of "high" and "low" is directly related to the line of action of the force of gravity.Therefore the sentence "the object A is above the object B" can be conceptualized as :

// $\exists$ P-point $\in$ CONE(A),Q-point $\in$ CONE(B) : X(P)=X(Q),Y(P)=Y(Q), Z(P) $\geqslant$ Z(Q) | FOR does not require further specification //

Note that we can state conditions only for pairs of points whose horizontal projections are the same.In fact, even the"pure" meaning of "above" is much more constraining |4,13|,this relationship is used in a number of "impure" meanings,in which we cannot say that the horizontal projection of A is included in the horizontal projection of B(Fig. 1a), or Z(P) $\geqslant$ Z(Q) for any pair or points P $\in$ CONE(A) and Q $\in$ CONE(B) (Fig.1b).

The preposition "on" is often synonymous of "above",but in some cases it can mean "below", as in "on the ceiling", or involve horizontal relations as in "the lamp is on the wall".Usually "A on B" requires B to support A against the action of gravity,by means of some kind of physical contact.Hence,the conceptualization of "a man on a chair" is the same as "a man above a chair",plus an assertion about physical contact and supporting action :

// $\exists$ P-point $\in$ CONE(MAN),Q-point $\in$ CONE(CHAIR) : X(P)=X(Q), Y(P)=Y(Q),Z(P) $\geqslant$ Z(Q) | CONE(CHAIR) applies a force to the CONE(MAN) | FOR does not require further specification //

Horizontal relations are much more ambigous.Sometimes FOR is explicity stated,as in "looking at the church,the post office is on your right"; otherwise a default assumption is to use FOR associated with the speaker or the listener.

If we consider the sentence "the object A is behind the object B",two interpretations are possible :

a) FOR is the n-tuple associated with the object B;

b) FOR is external to both objects A and B.

Case **a** can be assumed only if B is a class 1 object;case **b** is always assumed when B is a class 2 object,but it is not usual even when B is a class 1 object.In the case **a** the previous sentence is conceptualized as follows :

$$// \exists \text{ P-point} \in \text{CONE(A)}, \text{Q-point} \in \text{CONE(B)} : Y(P)=Y(Q), X(P) < X(Q) \mid$$
$$\text{FOR associated with CONE(B) (ie. FOR} \subset \text{CONE(B)} \quad //$$

This definition and the next one allow to handle situations as those shown in Fig. 2a-b ; **the** situation of Fig.2c does not represent a proper use of "behind";if such a preposition is used,more inferencing capabilities are needed. In the case **b** the previous sentence means that B is (partially) hiding A to an observer,who can be assumed to be one of the actors in the story;hence the conceptual representation is :

$$// \exists \text{ P-point} \in \text{CONE(A)}, \text{Q-point} \in \text{CONE(B)} : Y(P)=Y(Q), X(P) > X(Q) \mid$$
$$X(P), X(Q) > \emptyset, Y(P), Y(Q) \cong \emptyset \mid \text{FOR} \subset \text{K-point} \notin (\text{CONE(A) or CONE(B)}) //$$

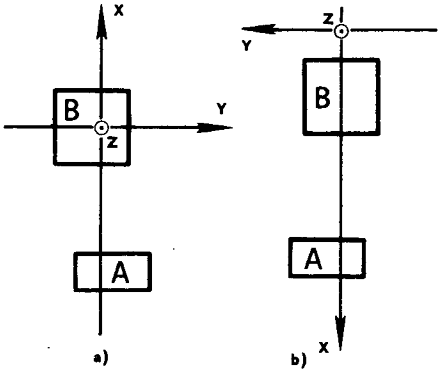

FIG. 2

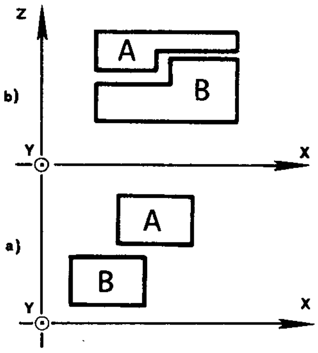

FIG.1

Let's now to consider relations as "on the edge of","on the surface of", "in the middle of" and so on.For every point P on the surface of the cone which describes the object A,its possible to find the corresponding cross-section,that is characterized by a value $\bar{Z}$ of the coordinate along the cone axis.The boundary of this section is described by a radial coordinate $\rho(\vartheta,\bar{Z})$.Therefore the sentence "the pen is in the middle of the table" can be conceptualized as follows,assuming as reference the cross section of the table cone which corresponds to the table top :

$$// \exists \text{ P-point} \in \text{CONE(PEN)}, \text{Q-point} \in \text{CONE(TABLE)} : \rho_P(\vartheta, Z) \cong \emptyset,$$
$$Z(P)=Z_{max}(Q) \mid \text{CONE(TABLE) applies a force to the CONE(PEN)} \mid$$
$$\text{FOR} \subset \text{CONE(TABLE)} \quad //$$

Let's conclude looking at sentences as "the house is before the bridge", "two miles after the lights" and so on.In these cases spatial relations are referred to a path,usually not straight.This type of relations can be conceptualized using a curvilinear coordinate **t** associated with a trajectory s originating in the center of FOR.If the analytical description of such a trajectory is unknown,the robot will be able to make inferences only about the relative positions of objects along the path;so,for instance,from the sentence "the house is two miles after the bridge along the road to Florence" it is possible to deduce that a man wolking towards Florence will meet at first the bridge and then the house,after an evaluable time.If more informations are available (eg. the path is a road and the map of the town is known),the position relative to other FOR can be evaluated from the actual value of t,in order to infer that "two miles after the bridge" means exactly "on the right of the station".The formal description of "the object A is after the object B", is :

$$// \exists \text{ P-point} \in \text{CONE(A)},\text{Q-point} \in \text{CONE(B)} : \text{P} \in \text{s-trajectory}, \\ \text{Q} \in \text{s-trajectory}, \ t(\text{P}) > t(\text{Q}) \mid \text{s-trajectory starts from CONE(B)}//$$

Finally,we should discuss how to quantify all the inequalities which result from the previously analyzed conceptualizations.Such a quantification can be considered as a special case of spatial inference, which unfortunately we cannot introduce here because of lack of space. An attempt to classify inferences can be found in |1|.

## CONCLUSIONS

The problem of robotic vision has been only sketched in this paper. Even if more detailed analysis of some particular objects can be found in the literature |3,7,10,13|,vision is yet a substantially open problem. A number of basic questions as,for example,the representation of objects with variable shapes,or the use of knowledge about the expected goals of an actor to infer its future movement,and the proper linking of cognition with image-processing procedures,are still waiting for a suitable answer.However,these topics are receiving more and more attention,both because of impact that an advanced,integrated vision-manipulation system could have on the applications of robotics,and because artificial intelligence people are aware that there is a large number of linguistic problem that cannot be solved if this perception capability is not achieved.

The work described in this paper is part of a larger project,whose goal is the development of a cognitive background,based on conceptual dependency and related concepts |11,12|,for an integrated vision-manipulation system.

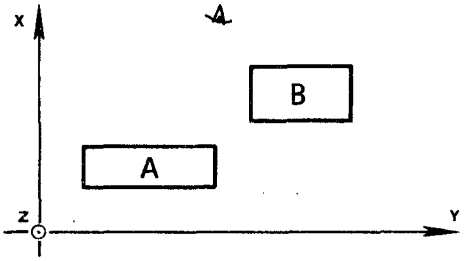

FIG. 2 – c

## REFERENCES

|1| Adorni,G.,Boccalatte,A.,Di Manzo,M., Object representation and spatial knowledge: an insight into the problem of men-robot communication, 7th.Conf. Canadian Man-Computer Communication Society, Waterloo,Canada (june 1981).

|2| Adorni,G.,Di Manzo,M., Some considerations about a conceptual model oriented to the representation of spatial relationship(in Italian), National Research Council ITD-045,Genoa,Italy (march 1980).

|3| Agin,G.J., Vision systems for inspectation and for manipulation control, J.Automatic Control Conf.,S.Francisco,CA,USA (june 1977).

|4| Hanson,A.,Riseman,E.(eds.), Computer vision systems (Academic Press,New York,1978).

|5| Boggess,L.C., Computational Interpretation of english spatial preposition, Tech-Rep. T-75,Coordinated Science Laboratory,Univ. of Illinois (february 1979).

|6| Kuipers,B.J., Modelling spatial knowledge, 5th. Int.J.Conf. on A. I.,Cambridge,MA,USA (august 1977).

|7| Lehnert,W.G., Representing physical objects in memory,Res-Rep.131, Dept. of Comp.Sc.,Yale Univ. (may 1978).

|8| Mackworth,A.K.,Havens,W.S., Structuring domain knowledge for visual perception, 7th. Int.J. Conf. on A.I.,Vancouver,Canada (august 1981).

|9| Marr,D.,Nishihara,H.K., Representation and recognition of the spatial organization of 3-D shapes,Proc.Royal.Soc.Lond.B. (1978).

|10| Nevatia,R., Computer analysis of scene of 3-D curved objects, (Birkhauser Verlag,Basel,1976).

|11| Schank,R.C.(ed.), Conceptual information processing (North-Holland,Amsterdam,1975).

|12| Schank,R.C.,Abelson,R.P., Scripts Plans Goals and Understanding (Lawrence Erlbaum,Hillsdale,1977).

|13| Waltz,D.L., Relating images concepts and words, NFS Workshop on the representation of 3-D objects,Univ.of Pennsylvania,Phyladelphia (1979).

|14| Waltz,D.L.,Boggess,L.C., Visual analog representation for natural language understanding, 6th.Int.J.Conf. on A.I.,Tokyo,Japan (august 1979).

|15| Weymounth,T.E., Experiments in knowledge-driven interpretation of natural scene, 7th.Int.J.Conf. on A.I.,Vancouver,Canada (august 1981).

|16| Winston,P.H.(ed.), The psychology of computer vision (Mc Graw Hill,New York,1975).