The Development of Visual Representations Using Computer Vision Methods

Bernd Neumann
Universität Hamburg

ABSTRACT

This contribution is a comprehensive introduction to Computer Vision. Its primary objective is to give an understanding of the major conceptual building blocks of a general vision system, in particular of the representations of visual information generated in such a system. Vision is viewed as a multilevel knowledge-based process for constructing descriptions of the real world from image data. Raw images are processed to yield edges, regions and other useful image elements. Images elements are grouped into larger aggregates. The crucial step is to interpret image elements as parts of the real world by assigning meaning and three-dimensional shape. Finally, real-world objects are understood as parts of meaningful situations or events. The different stages of this process are discussed in detail.

1. INTRODUCTION

Computer vision is one of the most challenging areas of Artificial Intelligence (AI). It is also an area which promises an extremely rich field of applications. Although the performance of currently available computer vision systems is far from the mark set by the human vision system, computer vision has already been applied to numerous special tasks including, for example, industrial object recognition, quality control, analysis of medical imagery, aerial image classification, object tracking, and others. In addition, a considerable body of knowledge, methods and insights has been assembled concerning vision systems of the future. In fact, many advanced methods have already demonstrated their potential under laboratory conditions.

This contribution is an introduction to Computer Vision organized in correspondence to the overall architecture of a vision system. As it is costumary in the field, vision is viewed as a multi-stage process transforming digital images into increasingly meaningful representations. An overview is given in Section 1.2. The discussions in the following sections are at the "knowledge level": Processes and representations are described conceptually in terms of information content and knowledge rather than procedures and data structures. This is the view point typically taken in Artificial Intelligence. It should be clear, however, that a vision system considered at the implementation level is a highly structured and extremely complex computer program. There may be different forms of implementation, e.g. using parallel or sequential processing. In fact, several new approaches (FELDMAN 88) are stimulated by the advent of massively parallel processing hardware. Implementation aspects will not be elaborated further.

1.1 Purpose of vision

Vision in humans is an activity pursued without conscious intent (for the most part). In order to provide guidance for the construction of artificial vision systems it is necessary to establish the purpose of Computer Vision.

Most researchers of the field take Computer Vision to be "the construction of explicit, meaningful descriptions of physical objects from images" (BALLARD and BROWN 82). This definition is illustrated by the figure below.

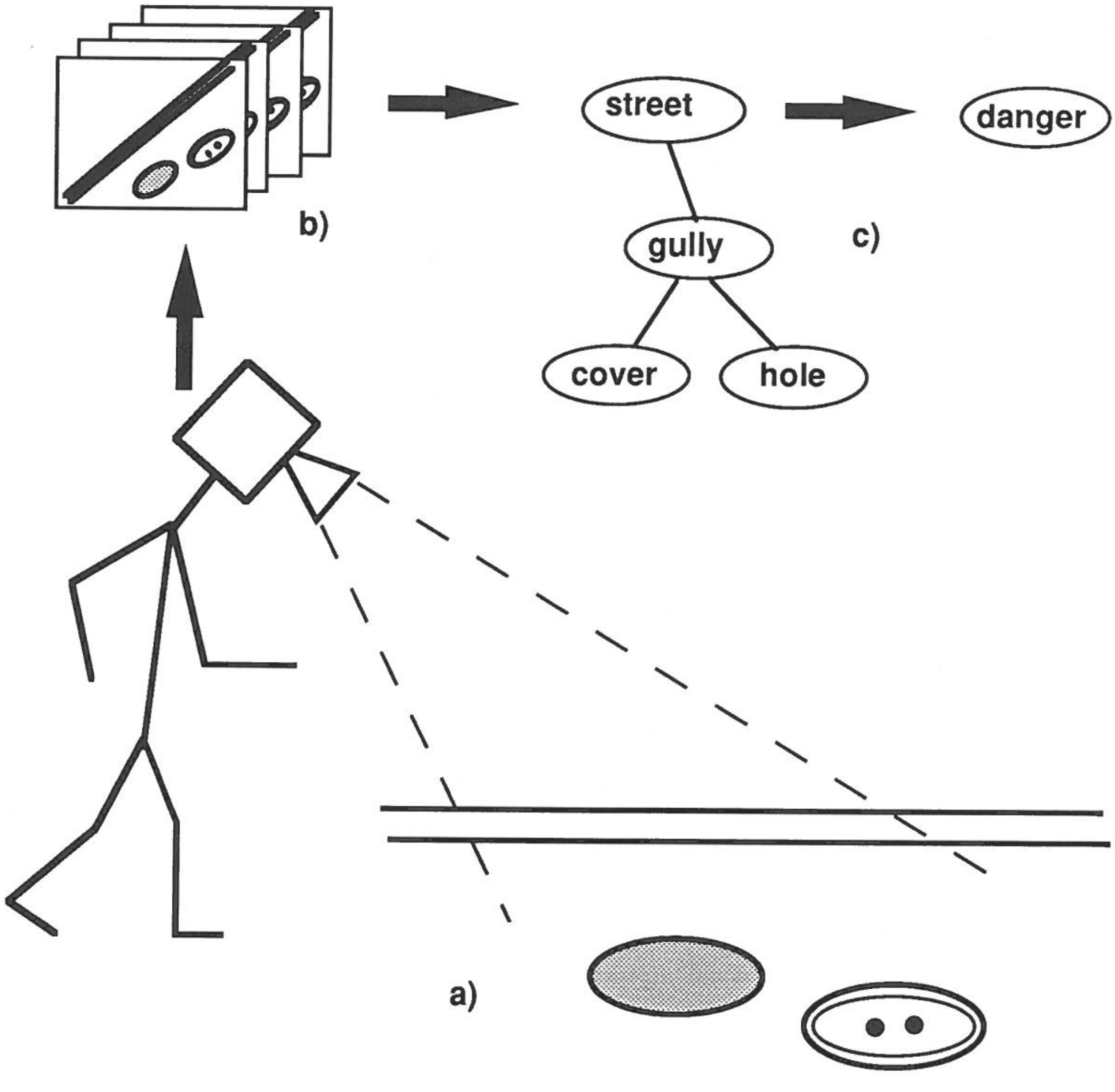

Fig. 1: Constructing a description of physical objects from images

Vision involves:
a) physical objects
b) images of physical objects
c) computer-internal descriptions of physical objects

The physical objects may be any subset of the real world, often called a scene. Scenes are in general 3-dimensional or, if time plays a part, 4-dimensional. Images are 2-dimensional projections of a scene if not defined otherwise. Time-varying scenes are taken to project into image sequences.

The output of Computer Vision is an explicit and meaningful scene description. One may argue about the extent of such a description. Following MARR 81, this description should state "what is where". Hence vision is scene reconstruction ("where") and object recognition ("what"). But vision may also involve higher-level concepts (e.g. "obstacle") as noted by WALTZ 79 and others. NEUMANN 82 suggested that a vision system be able to understand silent movies. Hence vision should encompass "scene interpretation" in the widest sense.

It is not the purpose of computer vision to generate descriptions of images. This is a misconception dating back to the beginnings where work on character recognition prevailed. Image properties of an object are conceptually quite different from real world properties. They describe 2D projections, not 3D objects.

1.2 Framework of a vision system

It is highly unlikely that a single one-step process could be conceived which reconstructs a scene from images. Computer vision is considered a multilevel multiprocess task. Finding useful intermediate representations at various levels of abstraction and procedures to compute these representations is the main business of computer vision research. Actually, the emphasis has been shifting towards representations rather than processes in the recent years. See MARR 78 and BARROW and TENENBAUM 78 for a detailed discussion of representation issues.

In Fig. 2 the framework of a vision system is sketched out in terms of levels of representations which seperate major conceptual building blocks. Traversing the diagram from the bottom to the top, we begin with digitized raw images, possibly organized as an image sequence, at the lowest level. The first processing step is often called segmentation, leading to a compressed representation of each image in terms of image elements, e.g. edges and regions. Image elements must be interpreted as scene elements, i.e. as parts of a real three-dimensional scene. The process which tries to achieve this is called low-level image interpretation. Scene elements are the constituents of object descriptions. The next step is to recognize meaningful objects from all the data gathered so far and using a priori knowledge about object properties. Further processing has the goal to recognize object configurations, events, special situations and other high-level concepts. This task is called high-level image interpretation or scene understanding.

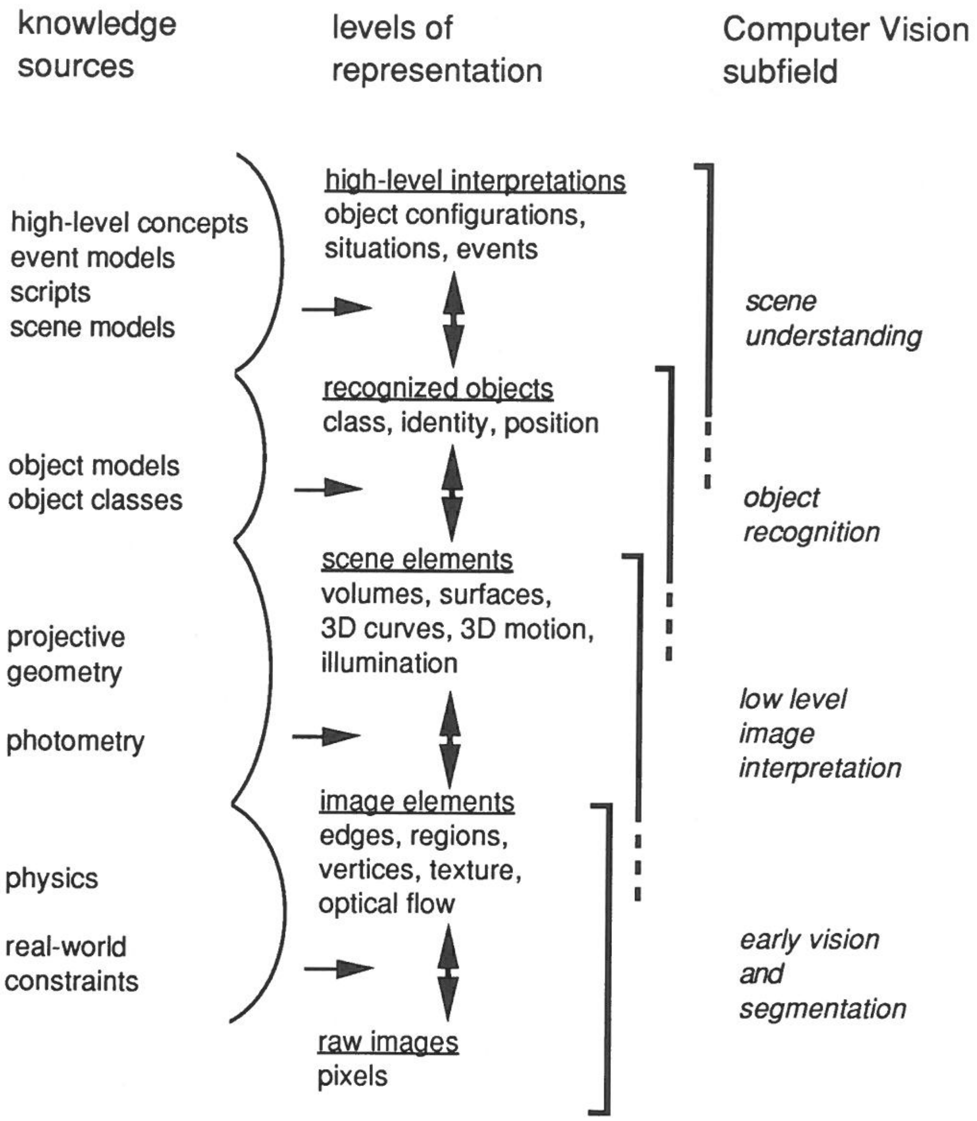

Fig. 2: Conceptual building blocks of a vision system

Subsequent sections will deal with each of these building blocks in more detail. At this point we note the following general properties.

- Vision is a knowledge-based process. At each processing step various types of knowledge must be brought to bear. For example, to recognize a particular object knowledge about the shape of such objects must be exploited. The lower the processing stage, the less specific will this knowledge have to be. Early vision processes reflect real world constraints of a very general nature, e.g. that object boundaries tend to give rise to sharp variations of image intensities.

- Vision involves bottom-up and top-down processing. The bottom-up path may be highly ambiguous by the nature of the problem, hence bottom-up processing often corresponds to hypothesis generation. Vice versa, top-down processing often corresponds to hypothesis verification.

- Scene reconstruction from images is essentially the inverse of generating images from a scene. Hence understanding the image formation process is crucial for low-level image interpretation. Scene reconstruction includes the computation of intrinsic scene characteristics which individually influence image formation, e.g. object surface reflectivity and orientation. At best, the reconstructed scene allows the generation of synthetic images which are identical to the real ones.

Vision systems structured according to this general framework are being developed in several research groups. Up to now, however, they can only perform tasks in very limited domains and not always to satisfaction. A critical survey of such laboratory systems has been written by BINFORD 82.

1.3 Related disciplines

Pattern Recognition

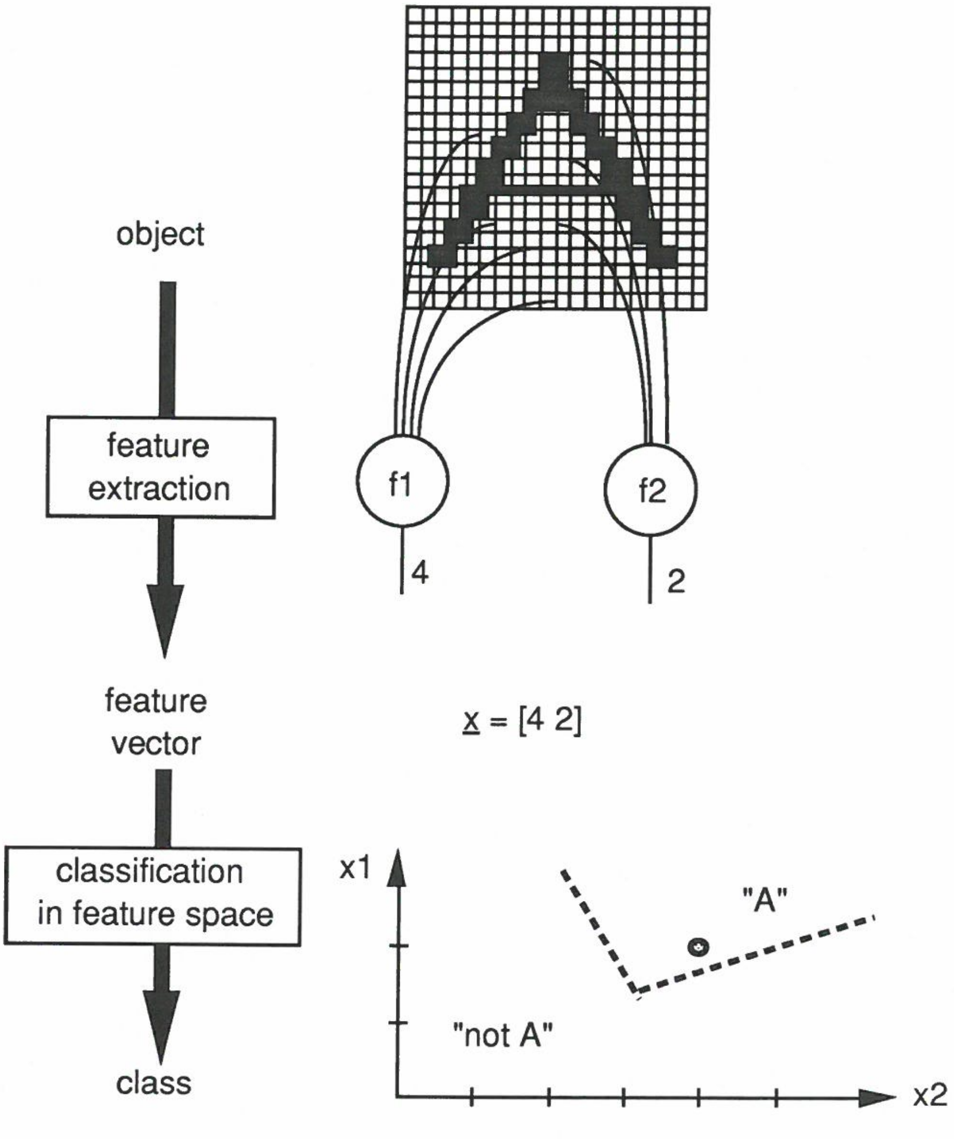

Fig. 3: The pattern recognition paradigm

Pattern recognition is a well developed discipline which offers formal procedures for recognizing different manifestations of objects of the same kind, e.g. handwritten letters. For a comprehensive treatment see DUDA and HART 73, part I, or any other of several excellent books on pattern recognition.

A pattern recognition problem has the structure shown in Fig. 3. Objects are to be represented by feature vectors. Then classification takes place according to the location of the feature vector in feature space. Pattern recognition offers a sophisticated classification theory. It rarely gives hints as to which features should be computed in a concrete problem. Pattern recognition plays a minor role in advanced approaches to computer vision. Most existing commercial vision systems, however, follow the pattern recognition paradigm.

Cognitive science

This is a young discipline devoted to understanding human cognitive processes using computational models. Cognitive science has roots in the established fields of computer science, AI, psychology, neurology, linguistics, and others. Computer vision and AI have a strong tradition of amalgamating ideas and findings from biological vision research. Hence there is no clear-cut distinction between a cognitive and an AI approach to vision. Generally, human vision is important for computer vision because

- humans provide an example of a well-working vision system, and

- the output of computer vision should be in terms of the semantic categories shaped by human thinking.

Biologically oriented computer vision research is carried out at many institutions. For access to the literature see MARR 81, ZUCKER 82, FLEET et al. 84.

2. EARLY VISION AND SEGMENTATION

The task of the first processing stage is to transform a raw image into a representation from which interesting properties of the real world scene can be computed more readily. The approaches which have been taken in computer vision research differ widely according to what properties of the real world scene are deemed interesting and how carefully the processes are designed to meet the objectives. A frequently stated goal is to segment the image into parts which correspond to meaningful entities of the scene, hence the name segmentation for this phase. Meaningful entities are, in general, objects and object parts of which a scene might be composed. For segmentation two complementary approaches can be taken (PAVLIDIS 77):

A Find regions with uniform image properties.

B Find boundaries where image properties are discontinuous.

Region analysis conveniently leads to a segmentation of the image whereas edge analysis requires further processing (e.g. thinning and linking) to obtain closed boundaries which would define image regions. Much work has been invested into segmentation techniques over the past 25 years (see some selected references below) and new proposals are still being published. Yet one can safely say that methods of sufficient generality which are applicable to arbitrary natural scenes, are not available. There seems to be one common cause for most failures: The uniformity computed by the segmentation operator on the image does not correspond to a uniformity of real world objects. For example, regions of uniform greylevel do not generally outline surfaces of uniform reflectance. This raises two issues. First, what are valid assumptions about real world scenes which can be exploited for segmentation? Second, how should the image be processed to exploit these assumptions?

A thorough treatment of these issues is due to Marr who proposed the primal sketch as output of the first processing stage. It consists of edge elements (zero-crossings of the second derivative of filtered versions of the image intensities), blobs, terminations, discontinuities and groupings of such tokens into higher organizational units, e.g. edge segments. Note that the idea of segmentation has been abandoned: The goal is to compute tokens which are generally useful for recovering the 3D-scene geometry. The term "early vision" is used collectively for such processes.

Another major issue of early vision is the problem of resolution. It is widely agreed that multiple resolutions are desirable for reasons of processing efficiency, see TANIMOTO and KLINGER 80 for several contributions along this line. A processing system for images at various resolutions is often called a processing "cone" or "pyramid". Marr argues for obtaining different resolutions by one-octave Gaussian bandpass filters. This appears to be in good agreement with the human visual system.

The last issue which will be pointed out here concerns grouping. At several stages in early vision image elements have to be combined to a larger whole. For example, edge elements are grouped into straight line pieces. The problem is to achieve a globally consistent result by local, possibly parallel decisions. One method is the Hough transform which is applicable whenever the constraints imposed by local evidence can be expressed by possible parameter values of the global result, e.g. by the parameters of possible straight lines through an edge element. Grouping is performed by finding parameter values with a maximal number of votes. Another basic method is relaxation. This is an iterative procedure where local values are modified according to their compatibility with the surrounding neighbourhood. The method is attractive because of its biological plausibility and conceputal elegance, but the behaviour is not always predictable and implementations on a sequential machine may be quite slow.

3. LOW-LEVEL IMAGE INTERPRETATION

The reconstruction of the 3D-scene geometry requires that image elements be interpreted in terms of scene elements, i.e. that scene characteristics are determined which "explain" the corresponding image characteristics. For example, an edge in the image may be interpreted as (or explained by) a shadow boundary in the scene. The basic problem here is that for each pixel there are several intrinsic scene characteristics (BARROW and TENENBAUM 78) which influence the intensity value, hence reconstruction is a one-to-many mapping.

From the physics of image formation the three main characteristics which are encoded into an intensity value are incident illumination, reflectivity of the surface, and surface orientation w.r.t. light sources and sensor. For scene reconstruction it is crucial to recover these constituents as well as other intrinsic properties such as depth values. This is the basic task of low-level image interpretation.

Because of the inherent ambiguity this can only be done, of course, if additional knowledge and assumptions about the scene are brought to bear. For example, one can exploit that surfaces are continuous in space and often have at least piecewise uniform reflectivity. Another important assumption concerns the geometry of light sources, objects and sensors: It is necessary to assume generality, i.e. the absence of deceiving coincidences. This means in particular that small changes of view point do not cause qualitative jumps of image properties. In consequence, collinearity in the image is taken to imply collinearity in space, straight lines in the image are straight lines in space, closed contours in the image are also closed in space, etc.

A typical example for low-level image interpretation is the evaluation of texture gradients, e.g. in the image of a golf ball. Without using prior knowledge about golf balls, based only on the generality assumption, one can reconstruct the three-dimensional surface orientation of the visible part of the golf ball.

Distance and orientation of visible object surfaces in a viewer-centered coordinate system are sometimes called a "2 1/2 D sketch", using a term introduced by MARR 81.

Computer vision research seems to be on the brink of formulating an encompassing set of real world assumptions to be exploited, and interpretation rules derivable from such assumptions. A bulk of contributions towards this end can be found in BRADY 81. NEUMANN 82 contains a discussion of the conceptual dependencies of the assumptions and rules. While such a framework of rules is emerging, much work remains to be done at a technical level. What are efficient procedures which would embody the interpretation rules? How do they interact? How does one deal with exceptions (when one or more of the assumptions are violated)?

4. OBJECT RECOGNITION

Image understanding requires that meaning is assigned to the components of a scene, individually and as a whole. Object recognition assigns meaning in terms of class membership or identity. The term "recognition" aptly describes that something in the scene matches knowledge retained from prior encounters. This knowledge is called a model, while the corresponding part of the scene description is called an instantiation of this model. The main issues of object recognition are

(i) how to represent object models, and

(ii) how to recognize objects given such models.

The following discussion concerns object recognition as conceived for a general vision system. For specialized tasks such as in industrial vision simplified procedures may do the job.

There are two conflicting requirements for an object model. First, it must abstract from properties which distinguish objects of the same class. For example, it should in general not contain illumination and view-point information since these characteristics are usually irrelevant for class membership. Second, an object model must support recognition. Hence it should provide a description easily comparable with a scene description, which is in general illumination and view-point dependent. The burden of providing descriptive features by which scene components can be compared with models may be partly placed on low-level image interpretation processes (Section 3). If they provide discriminating object-specific features, in particular view-point independent 3D shape properties, object models may be comparatively compact. But as low-level vision remains a problem, it is useful to consider object models which also encode view-point dependent information ("mixed models").

Models must be distinguished according to their use for identification (determining physical identity) or classification (recognition of class membership). Some tasks require identification, e.g. tracking a moving object, others require classification, e.g. sorting work pieces on a conveyor belt. Classification establishes the traditional ISA-relationship between a class and a class member, while the identity relation may be called IS. The two types of models differ by the abstractions which they perform. Both models usually abstract from time, position and orientation in space, view point, and illumination. In addition, classification models usually abstract, to some degree, from surface properties and shape.

Recognition involves comparison, or in a general sense, matching. Techniques for matching structural descriptions are well developed formally, but there remain problems of efficiency with large model databases. It is essential that temporal or spatial context and higher-level knowledge be brought to bear to narrow down the possibilities. Object models must be viewed as constituents of larger knowledge frames. See MINSKY 75 for a discussion of such knowledge structures.

Matching may also be performed at a lower level of representation, for example at the iconic level. This requires that a lower-level representation be computed from the hypothesized instantiation of an object model, for example by synthesizing an image. Recognition based on this strategy is sometimes called 'analysis by synthesis'. If geometrical measures are to be used as a goodness criterion for matching, it may be more efficient to use geometrical rather than symbolic representations. Matching will then be performed by comparing surfaces, edges, vertices, etc. in Euclidean 2- or 3-space.

5. SCENE UNDERSTANDING

Scene understanding deals with representations and processes "above" the level of recognized objects. The conceptual units of interest at this stage are, for example, object configurations, situations, motion concepts, events or even episodes. There is no established boundary as to where vision ends and other cognitive processes begin. See WALTZ 79 and NEUMANN 82 for a discussion of the scope of a vision system. It is important to realize, however, that scene understanding is about interpreting (representations of) real world scenes. Images are only indirectly involved.

"Scene understanding" refers to the recognition of higher-level concepts, which is essentially <u>bottom-up</u> processing. However, the <u>top-down</u> path - exploiting higher-level knowledge for scene reconstruction and object recognition - is equally important. In fact, laboratory systems have been able to work on natural scenes because higher-level knowledge guiding the recognition task has been provided by the experimentator.

Research in bottom-up recognition of higher-level concepts is only at the beginning. The remainder of this section will deal with one aspect which has received somewhat more attention: the interpretation of <u>motion</u>. The necessity of recognizing motion concepts in time-varying scenes is evident when one considers the description of a moving object at the object recognition level. It would consist of identity, class membership and shape as well as the spatial positions and orientations for each time slice. Recognition of motion concepts then amounts to finding a more succinct, qualitative description, e.g. 'moving along a straight line'. The question of what concepts should be used and how they should be organized has been treated in depth by TSOTSOS 80. He distinguishes between domain-independent concepts (e.g. rotate, translate, expand, shrink) and domain-dependent concepts expressed in terms of the former ones. Concepts may be defined in accord with natural language expressions - humans tend to name conceptual entities which are important.

The work on traffic scene description in project NAOS (NEUMANN and NOVAK 86) is one of the first examples where a natural-language description is automatically generated from a time-varying scene. Fig. 4 shows the main levels of representation in NAOS. Note that image sequence analysis up to the level of object recognition has been bypassed using interactive techniques.

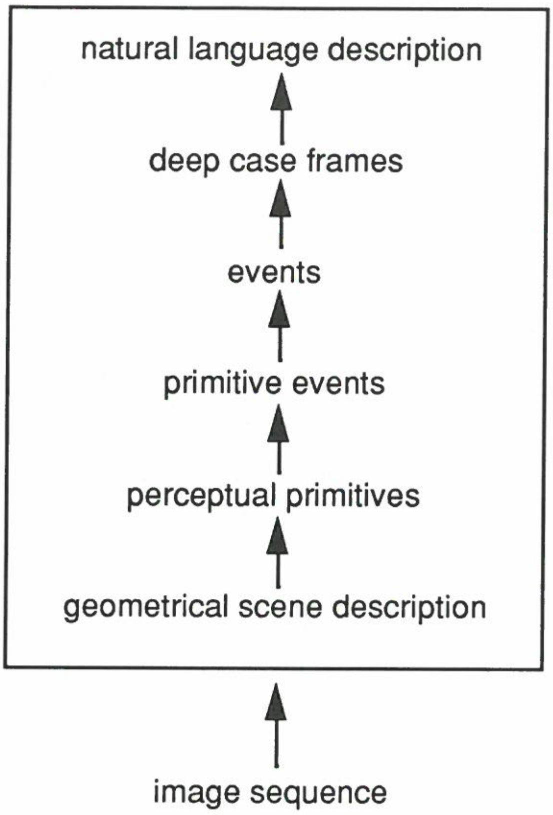

Fig. 4: The main levels of representation in NAOS

With the perspective on future cognitive systems, scene understanding can be considered an interface between vision, natural language, spatial reasoning and other cognitive processes.

The techniques proposed for representation and recognition of motion concepts are in many ways similar to those used in any AI problem where a complex whole is to be recognized by its components: One attempts to instantiate relational or frame-type structures. The need for novel methods arises mainly from the special nature of time and the strong dependency of higher-level concepts on common sense knowledge. In general, scene understanding requires interaction with various knowledge sources and cognitive processes, e.g. common sense knowledge about physics, spatial inference, rules of typical behaviour, etc. Most of these components still require considerable research in their own right.

6. REFERENCES

Ballard and Brown 82
Computer Vision
D.H. Ballard and C.M. Brown
Prentice-Hall, Englewood Cliffs, N.J., 1982

Barrow and Tenenbaum 78
Recovering Intrinsic Scene Characteristics from Images
H.G. Barrow and J.M. Tenenbaum
in: Computer Vision Systems, pp. 3-26 A.R. Hanson and E.M. Riseman (eds.)
Academic Press, New York 1978

Binford 82
Survey of Model-Based Image Analysis Systems
T.O. Binford
Robotics Research 1(1), 1982, 18-64

Brady 81
Computer Vision
J.M. Brady (ed.) North-Holland Publ. Co., 1981, reprinted from Artificial Intelligence 17, 1981

Duda and Hart 73
Pattern Classification and Scene Analysis
R.O. Duda and P.E. Hart
John Wiley Sons, New York, 1973

Feldman 88
Time, Space and Form in Vision
J.A. Feldman
TR-88-011, International Computer Science, 1988

Fleet et al. 84
A Spatio-Temporal Model for Early Visual Processing
D.J. Fleet, A.D. Jepson, P.E. Hallet
RCBV-TR-84-1, University of Toronto, Canada, 1984

Marr 78
Representing Visual Information - a Computational Approach
D. Marr
in: Hanson and Riseman 78a, 61-80

Marr 81
Vision
D. Marr
Freeman, 1981

Minsky 75
A Framework For Representing Knowledge
M. Minsky
in: P.H. Winston (ed.), The Psychology of Computer Vision, McGraw-Hill 1975

Neumann 82
Knowledge Sources for Understanding and Describing Image Sequences
Bernd Neumann
in: W.Wahlster (ed.), GWAI-82, Informatik Fachberichte 58, Springer 1982, 1-21

Neumann and Novak 86
NAOS: Ein System zur natürlichsprachlichen Beschreibung zeitveränderlicher Szenen
B. Neumann and H.-J. Novak
Informatik Forschung und Entwicklung (1), 1986, 83-92, Springer-Verlag

Pavlidis 77
Structural Pattern Recognition
T. Pavlidis
Springer, 1977

Tanimoto and Klinger 80
Structured Computer Vision
S. Tanimoto, A. Klinger
Academic Press, 1980

Tsotsos 80
A Framework for Visual Motion Understanding
J.K.Tsotsos
TR CSRG-114, University of Toronto, 1980

Waltz 79
Relating Images, Concepts, and Words
D.L. Waltz
Proc. NSF Workshop on the Representation of Three-Dimensional Objects
R. Bajcsy (ed.), Philadelphia/PA, May 1-2, 1979

Zucker 82
Early Orientation Selection and Grouping: Type I and Type II Processes
S.W. Zucker
TR 82-6, Dep. Electr. Eng., McGill University, Montreal, Canada, 1982