KNOWLEDGE SOURCES FOR UNDERSTANDING AND DESCRIBING IMAGE SEQUENCES

Bernd Neumann

IfI-HH-M-103/82

November 1982

This paper has been presented at the GWAI-82, 6. Fachtagung über Künstliche Intelligenz, 27.9.-1.10.82 in Bad Honnef, W. Germany. Proceedings as GWAI-82 (W. Wahlster, ed.), Informatik-Fachberichte, Springer-Verlag Berlin/Heidelberg/New York, 1982, 1-21.

Wissensquellen zum Verstehen und Beschreiben von Bildfolgen

Zusammenfassung

Das Verstehen von Bildfolgen wird als ein Vorgang dargestellt, bei dem einzelne Prozesse verschiedenartige Wissensquellen dazu ausnutzen, um aus den Rohbildern Beschreibungen mit zunehmendem Bedeutungsgehalt abzuleiten. Der erste Teil dieser Mitteilung handelt von niedrigen Bilddeutungsprozessen. Es wird eine Übersicht über neuere Arbeiten gegeben, besonders im Hinblick auf die zugrundeliegenden Annahmen über die reale Welt und den Bildentstehungsvorgang. Im zweiten Teil wird Objekterkennung diskutiert. Vorwissen über Objektformen muß durch zusätzliches Wissen ergänzt werden, welches den Erkennungsprozess unterstützt. Abschließend werden Arbeiten zur Bewegungsanalyse behandelt. Hier geht es um Wissensquellen, die zum Erkennen höherer Bewegungskonzepte erforderlich sind.

KNOWLEDGE SOURCES FOR UNDERSTANDING AND DESCRIBING IMAGE SEQUENCES

Bernd Neumann

Fachbereich Informatik
Universität Hamburg
Schlüterstraße 70
D-2000 Hamburg 13

Abstract

The task of understanding image sequences is viewed in terms of processes which exploit various knowledge sources to derive increasingly meaningful descriptions from raw image data. The first part of the article deals with low-level vision. Recent work is surveyed with respect to underlying assumptions about the real world and the image formation process. In the second part object recognition is discussed. Shape knowledge must be augmented by special knowledge which supports the recognition process. Finally, work on motion understanding is reviewed with respect to the knowledge required for recognizing higher-level concepts.

1. Introduction

Image sequence analysis is one of the major tasks of artificial intelligence. It deals with time-varying visual data which is visual information in its most general form. The typical input of the human visual system is time-varying, be it due to observer motion or scene changes. Hence image sequences should be also considered the typical data for computer vision which is the endeavour to do by computer what humans do by eyes and brain. But the historical development of computer vision has led to a single-image paradigm due to the many problems which occur already in this restricted case. Single image analysis is customarily understood to be the task of localizing, describing and identifying objects in the 2D image of a 3D scene. While object recognition is undoubtedly an essential prerequisite of image understanding, it is not the whole problem. This becomes apparent immediately when turning to image sequences.

Image sequences have much the same semantic potential as silent movies. This means that computer analysis of image sequences should ultimately be able to extract results comparable to human understanding of silent movies. This is, of course, an ambitious goal in view of the state of the art in computer vision. But it is the proper view to take when defining the competence of a general vision system.

Silent movie understanding in vision compares to story understanding in natural language processing. In fact, it is likely that the semantics derived from either input should be represented in much the same way. Representational schemes for both, story understanding and image sequence understanding, are still in their infancy, however, so it is premature to decide about commonalities. But the connection between language and images is important for two other reasons.

First, the meaning of a natural language utterance may be clarified by visualization. This amounts to inverting the image understanding process: (imaginary) images are generated from an abstract representation. Waltz has studied this process in some detail using examples like "The dachshund bit on the mailman's ear" (WALTZ 81). He is one of the few researchers active in both vision and natural language.

Another important connection of vision and natural language arises from verbalizing what one sees. Here the results of image (sequence) analysis are taken as input for a process which maps meanings into language. It is not at all clear, what the results of image analysis should be, i.e. where image analysis ends and verbalization begins. Vision researchers tend to consider verbalization the front end of image analysis and natural language researchers vice versa. On the surface this looks like an academic question, but at its core lies the old controversy about the origin of thought: whether thinking is in terms of visual or verbal concepts. This question shall not be persued, however, in this article. It has only been mentioned to

illustrate the uncertainty about higher level vision processes.

This article attempts to give an overview of vision which goes beyond the traditional paradigm of object recognition. It includes and emphasizes contributions which map visual data into higher level concepts, in particular those concepts which support natural language description. The latter part will be quite sketchy and mostly based on motion research.

Computer vision is often scolded for its lack of scientific rigour. Indeed, much of the early literature describes algorithms for visual data manipulation which have a certain adhoc flavour, e.g. algorithms for object recognition in a very restricted environment. But the need for more theory and less empiry has long been responded to. One major step forward resulted from model-based vision concepts. Knowledge about what one is looking for is made explicit using structural representations, while recognition is the conceptually simple process of comparing known descriptions with the unknown (BARROW and POPPLESTONE 71). Other significant achievements resulted from investigating the laws of image formation, i.e. the physics which relate intensity and color of an image point to the corresponding surface element in space (HORN 75). In the same spirit projective geometry was studied to relate the shape of image components to surface shape in space (e.g. GRIMSON 81). All this is mentioned here to motivate a particular point of view which will be taken in this article. Vision will be considered at the knowledge level. This means that processes will be described in terms of knowledge which they exploit or assumptions which they make about the world to attain certain goals. The knowledge level abstracts from particular algorithms and exposes the rationale, see NEWELL 82 for an illuminating discussion of this notion. For example, an algorithm which extracts object boundaries by greyvalue thresholding can be characterized at the knowledge level as a process which assumes that the visual world consists of flat cardboard pieces tossed randomly onto a plane. The deficiencies and limitations are immediately apparent since

we can judge the truth of the underlying assumptions.

The main body of this article is organized according to three phases of a vision system:

(i) low-level vision
(ii) object recognition
(iii) high-level vision

Low-level vision deals with processes which map raw image data into representations below the object level, e.g. pixels into lines or regions into surfaces. From the early beginnings of image processing up to now low-level processes have enjoyed considerable attention, after all most image analysis tasks begin with raw images. In spite of much work the results are still unsatisfactory compared with human vision, and efforts to implement general vision systems usually suffer from poor low-level processing. Recently a series of papers (published jointly in Artificial Intelligence 17, 1981, and also in BRADY 81) contributed significant novel ideas. The major part of chapter 2 gives an account of this work.

The next phase - object recognition - encompasses processes which localize and identify objects using the output of low-level vision and a priori knowledge about object shapes. A priori knowledge is represented in terms of models which capture invariant properties of an object. Chapter 3 reviews the requirements for useful object models from the knowledge point of view.

As has been pointed out before, most vision research ends at object recognition, and anything beyond is much less complete and well-defined. Nevertheless chapter 4 discusses phase 3 of a vision system which is tentatively called high-level vision. Given the output of phase 2 in terms of identified objects and object locations, how can one extract the meaning of a scene? The discussion will mainly focus on motion interpretation. It is shown that in general the meaning of motion cannot be computed by

comparing object trajectories to models. Various other sources of knowledge are required, e.g. domain-dependent standards, models for "events", measures of interestingness. A great deal of affinity to corresponding knowledge structures in natural language reasearch becomes apparent.

2. Low-level vision

It is widely accepted that initial processing of raw images should be general-purpose, i.e. independent of the contents of a particular scene and of a priori knowledge which one might have about it. In fact, low-level vision can be defined as processing images on the grounds of general knowledge about the relationship of images and real-world scenes. In this chapter, the kinds of knowledge which are exploited are first outlined in toto. Then several computational theories are reviewed which tap the knowledge for image processing purposes.

The properties of images are determined by the imaging situation, which can be decomposed into five constituents:

1) Real world. Object shapes and surfaces have certain typical properties irrespective of a particular scene or domain.

2) Illumination. Although images may be taken under a wide variety of lighting conditions, certain restrictions may be expected to hold.

3) View point. Position and orientation of the imaging system usually preclude atypical views.

4) Photometry. Given complete knowledge of 1) to 3), color and brightness of a pixel may be determined from photometric laws.

5) Projective geometry. Similarly, pixel coordinates

corresponding to a real-world point follow from the laws of projective geometry.

The most important property of real-world objects is coherence. Objects are not scattered about in small components but tend to be a connected entity. Of course, there are exceptions or cases of very loose connection, e.g. the branches of a tree in winter, but in general it is valid to assume coherence. From this the coherence principle of low-level vision can be derived:

A1: Assume that neighboring pixels belong to the same object, if there is no evidence to the contrary.

Many segmentation algorithms incorporate this principle, e.g. by merging isolated spots with surrounding regions, although coherence is rarely stated as an underlying assumption.

An equally basic notion is persistence or - more precisely - continuity of change. As a rule, objects do not appear or disappear suddenly or instantaneously undergo drastic changes of shape, color and position. This also applies to illumination and view point properties. In consequence images change only gradually along the time axis if taken at sufficiently small intervals. From this follows an assumption about image sequences:

A2: An object point which is visible in a certain image at a certain location will be visible at approximately the same location in the next image.

This assumption plays an important part in several processes proposed for motion analysis. For example, optical flow (which is the field of pixel displacement vectors between successive images) can be computed from local greyvalue changes given that the displacement vectors are small and certain smoothness assumptions hold (HORN and SCHUNCK 81). If applied to prominent picture points, displacement computation amounts to determine the

correspondence of points in successive images. Proximity, i.e. the assumption of change continuity, has been found to be an essential criterion (BARNARD and THOMPSON 80, DRESCHLER 81), also on the grounds of psychological experiments (ULLMAN 79). It must be noted, however, that A2 is invalid in all but idealized image sequences due to the phenomenon of occlusion. If there is motion, there are always parts of the scene which disappear and others which are uncovered. Nevertheless A2 is employed by a reason which is typical for low-level vision strategies: The number of pixels for which A2 is valid is generally larger than the number of pixels for which it is not valid by one order of magnitude, since the former is a function of object area while the latter depends on contour length. Hence A2 is a good guess. It is not surprising that the above mentioned procedures for optical flow and correspondence computation have problems at object boundaries.

Brightness discontinuities play a key part in low-level vision. They may delineate object boundaries and thus provide a means for segmenting the image into meaningful components. It is worthwhile to consider the underlying physics in order to understand the assumptions on which certain approaches are based. The brightness of a pixel which depicts some surface element depends on

- illumination: the light cast upon the surface element
- reflectivity: physical properties of the surface
- geometry: surface orientation with respect to light sources and observer
- sensor: properties of the imaging device

Hence in principle brightness discontinuities may be due to discontinuities of either of the four factors.

In a thoughtful essay BINFORD 81 elaborates how to exploit brightness discontinuities. First, one should insure a homogeneous sensor response by proper calibration. Binford

conjectures that the microsaccades of human eyes serve this purpose: By comparing the responses of neighboring cells, sensor discontinuities may be evened out. Second, one should suppress unwanted responses due to smooth brightness variations by means of lateral inhibition, i.e. by subtracting from each pixel value the weighted average of its neighborhood. (This operation is known to be used extensively in human perception). The underlying assumption is simple:

A3: Object boundaries do not occur at places of smooth brightness variations.

In view of the interplay of illumination, reflectivity and geometry this is not necessarily true. For example, the effect of an orientation discontinuity at an object boundary may very well be undone by a coinciding illumination discontinuity. A3 is based on a fundamental assumption which reflects the independence of these three factors.

A4: The position of light sources and observer are general, if there is no evidence to the contrary.

A4 has been put forth by several researchers as a guiding principle for image interpretation, see the discussion of STEVENS 81 below as an example. In BINFORD 81 one can even find a more general version:

A5: Perception derives predictions from data using the most general model.

In other words, those interpretations are preferred which impose as few constraints on the unknowns as possible. It would be interesting to tie A4 and A5 to a probabilistic argument, but this has not yet been carried out to the author's knowledge.

Returning to the processing of brightness discontinuities as discussed in BINFORD 81, the next step would be the detection and

localization of elongated step- or peak-like discontinuities - in short: boundary elements - in the lateral inhibition signal. A sense of direction is important for linking boundary elements into boundary lines and for the interpretation of junctions which will be described later. Binford proposes detection of step boundaries by thresholding the gradient of the lateral inhibition signal and localization by finding the zero crossings of its second directional derivative. Conversely, peak boundaries are detected by thresholding the second derivative and localized by zero crossings in the first derivative.

It is interesting to compare this approach with the theory of edge detection developed in MARR and HILDRETH 80. They propose to localize brightness discontinuities by taking the zero crossings of the second derivative of the brightness function in a bandpass filtered version of the raw image data. Filter and derivative operation can be combined into the so-called "mexican hat" operator which is essentially the same as lateral inhibition. Thus Marr locates boundary elements along lines of maximal brightness gradients, while Binford determines position and direction of maximal change of brightness curvature - two levels of differentiation below Marr. Almost all other edge finders which have been proposed for low-level vision are brightness gradient operators of some kind, although only few reflect an underlying theory of edge detection.

For further processing of boundary lines it is crucial to distinguish between illumination, reflectivity and orientation boundaries or combinations thereof. To date, no complete solution of this problem is known, but certain evidence can be exploited which may contribute to a disambiguation. For example, if the brightness ratio across a boundary is approximately constant while individual brightness values vary along the line, then this line is an illumination boundary (BINFORD 81). The reason is assumption A4 from which one can postulate constant surface orientation and reflectivity across an illumination boundary. One must also postulate constant illumination on

either side along the boundary.

Let us assume that boundaries are correctly classified. What can one tell about the surfaces in between the boundaries? This question has enjoyed considerable attention in recent work on low-level vision, and some remarkable progress has been achieved.

Given a single closed contour in terms of the zero crossings of Marr's mexican hat operator and the absence of further zero crossings, GRIMSON 81 investigates the problem of interpolating the 3D surface orientation between the boundaries. Not all surfaces are equally likely since radical surface inflections tend to cause additional zero crossings which are known to be not there. Grimson proposes to choose an interpolation surface which minimizes the probability of such additional zero crossings. A probability distribution can be derived by assuming a uniform distribution for the reflectance normal which captures the effect of unknown illumination, reflectivity and observer properties. This assumption is basically a probabilistic version of A4. Grimson proves that the best surface approximation minimizes the following measure of surface smoothness:

$$\min \iint (s_{xx}^2 + 2s_{xy}^2 + s_{yy}^2)\, dx\, dy$$

(s_{xx}, s_{xy} and s_{yy} are the second derivatives of the surface function in a viewer centered coordinate system). Thus an assumption about "typical" surface shape has been derived:

A6: For the interpolation of surface shape minimize the quadratic variation of the surface gradient.

BARROW and TENENBAUM 81 investigate the same problem and arrive at similar results. They also report about experiments with local operators which carry out the interpolation. For example, a circular boundary was interpolated into a perfect sphere.

The problem of constructing 3D surfaces from boundary lines in an

image is also addressed by STEVENS 81. He deals with surface contours, i.e. reflectivity or illumination boundaries as opposed to orientation boundaries (occluding contours). As the example in Fig. 1 shows, humans are quite capable of inferring an unknown 3D surface from lines on flat paper.

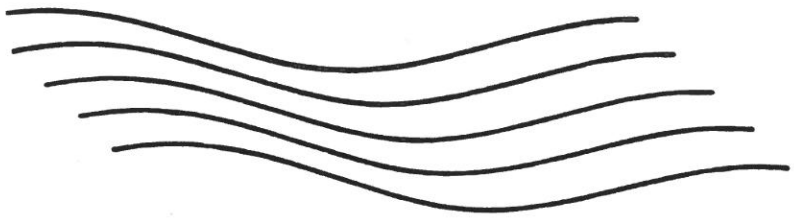

Figure 1: Inferring 3D shape from surface contours

Stevens analyzes the assumptions underlying such a process. As it turns out, A4 is crucial for various inferences. One of the rules which follow from A4 is

A7: Parallel curves in the image are also parallel in space.

If the additional assumption of general placement holds -

A8: Parallel curves remain parallel if slightly displaced on the surface.

- a strong constraint on surface shape ensues:

A9: Parallel image curves lie locally on a cylinder.

Hence low-level vision processes would interpret Fig. 1 in much the same way as humans appear to do it, if the preceding assumptions were adopted.

Surface contours may also occur as texture, i.e. as statistically

distributed surface markings. Texture gradients, i.e. the distortion caused by viewing a textured surface from an oblique angle, are known to provide humans with valuable information about surface orientation even if the undistorted texture shape is unknown. WITKIN 81 proposes a computational approach on the basis of the following assumption:

A10: Textures do not mimick projective effects.

In other words, texture is as irregular or unbiased as evidence allows. If a distortion or directional bias can be explained by projective effects, it is indeed caused by projective effects. Of course, there exist texture patterns which do not conform with A10 and will deceive this process, but human vision seems to be just as fallible. It is interesting to note that A10 can be considered a consequence of A5, underlining the fundamental role of the latter.

In typical images objects occlude each other and object boundaries are only partially visible. From the way boundary pieces are aligned and meet at junctions, constraints can be inferred on their spatial interpretation. LOWE and BINFORD 81 point out several such rules, e.g.

A11: A straight line in the image is also a straight line in space.

A curved line in space requires a special view point to appear straight; hence A11 is valid if the general view point assumption A4 holds. Similarly:

A12: Collinear lines in the image are also collinear in space.

The following assumptions concern junctions:

A13: An image curve which terminates at a continuous curve (forming the stem of a T) cannot be closer to the viewer

than the continuous curve.

A14: If two or more image curves terminate at the same junction (forming an L or Y etc.), they also terminate at the same point in space.

It is important to observe that these rules are not just a bunch of heuristics but all derive from A4. Hence one may very well talk about an emerging theory of low-level vision, although it is difficult to judge whether the rules which have been put forth so far, are complete in any sense.

The remainder of this chapter deals with processes which exploit photometric laws and laws of projective geometry.

Horn's work on shape from shading (HORN 75) has recently been extended to incorporate orientation constraints at object boundaries (IKEUCHI and HORN 81). The surface shape within such boundaries may be obtained from varying brightness values if the reflectance map (which gives the expected brightness for each surface orientation) is known. While there are certain applications which meet this requirement, reflectance properties are usually unknown in low-level vision. It would be interesting to attack the shape-from-shading problem in the same spirit as GRIMSON 81 by searching for the "most likely" surface interpretation compatible with the brightness variations but without knowledge of the reflectance map.

A commonly known process which exploits projective geometry is binocular stereo. It plays an important part in Marr's theory of low-level vision, where depth information is derived from the binocular disparity of the mexican-hat zero crossings (MARR and POGGIO 79). BAKER and BINFORD 81 also propose an edge-based scheme. While the underlying mathematics can be easily derived (see e.g. DUDA and HART 73, p. 398) computer implementations pose problems of accuracy and computational expense (for a survey see NEUMANN 81). It is not yet clear whether binocular stereo may be

expendable in computer vision systems.

Depth information may also be obtained from motion stereo. If an object moves in space, the projected displacements of points on its surface give a clue concerning the spatial configuration and trajectory of these points. For a quantitative analysis one assumes that the object is rigid. The mathematics are not trivial and have only recently been clarified (TSAI and HUANG 81), although they do not exploit anything else than the millenium-old laws of perspective projection. For the purpose of this discussion we only note the additional assumption:

A15: Objects are rigid if there is no evidence to the contrary.

It is not clear, how strong evidence to the contrary should be if A15 is to be abandoned. BARROW and TENENBAUM 81 point out a remarkable phenomenon in human vision, where a rigid wire frame cube appears to be bending and stretching when moved while viewed in Necker reversal. The false interpretation is possibly maintained on the grounds of change continuity (A2) which seems to override A15.

This concludes the overview of low-level vision. Only a small fraction of the pertinent work could be covered due to limitations of space. A useful starting point for further study is BRADY 81. The main results presented in this chapter may be summarized as follows. In low-level vision, object boundaries and surface shape can be computed by exploiting general knowledge about the real world and the imaging process. Various inherent ambiguities concerning a correct spatial interpretation of image features are resolved on the basis of sensible assumptions. A large part of these assumptions can be considered the consequence of the principle of generality (A5) which forms the basis of an emerging theory of low-level vision.

3. Object recognition

Image understanding requires that meaning is assigned to the components of a scene, both individually and as a whole. Object recognition assigns meaning in terms of class membership or identity. In this chapter, the knowledge required for object recognition is characterized by the abstractions which distinguish object models from the object descriptions obtained from low-level vision. It is shown that object models should also contain information tailored to support the recognition process.

The term "recognition" very 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. From the discussion in the preceding chapter it is known that low-level vision provides descriptions for coherent entities in terms of

- visible surface shape
- perceived brightness and color
- position
- time of observation
- illumination

Since an object model must be compared with such a description, it should contain information pertinent to these descriptors. The knowledge captured herewith will be loosely referred to as "shape" knowledge. (Shape is, strictly speaking, only a geometric notion.) There are other kinds of knowledge which may help recognition, e.g. context information from which the position of an object could be inferred. Knowledge of this kind will be discussed in the following chapter. In this chapter we shall only deal with recognition based on visual properties.

There are two conflicting requirements for an object model.

First, it must be an efficient representation for a class of objects. Hence it should abstract from properties which distinguish objects of the same class. For example, objects are typically not distinguished according to illumination or view point. Consequently, models should not contain information which is illumination or view point dependent.

Second, an object model should support recognition. Hence it should provide a description which can be easily compared with an illumination and view point dependent low-level scene description. We shall first discuss object models under the former aspect.

Models must be distinguished according to their use for identification or classification. Identification denotes recognition that object and model are physically the same, whereas classification denotes recognition of class membership. For example, an object can be identified as the dome at Cologne or classified as a church. Classification establishes the traditional ISA-relationship between a class model and a class member, while the identity relation may be called IS.

It is possible to characterize the knowledge contained in the two types of models by the abstractions they perform. Identification usually abstracts from

- instance of time
- position and orientation in space
- view point
- illumination

Classification usually abstracts from all this and also, to some degree, from

- surface properties
- shape

But there are many examples which blur this characterization. Identity may depend very little on appearance (e.g. a person in different ages) while classes may be quite narrowly defined (e.g. a 1 DM coin). In general, however, class membership is less specific than identity.

Several representational schemes have been proposed which have the desired abstraction properties (BINFORD 71, AGIN 72, NEVATIA 74). Because of view point independence object shape is always defined with respect to an object centered coordinate system. One such example is the generalized cylinder. It describes shape in terms of a planar cross section, a space curve spine, and a sweeping rule. It represents the volume swept out by the cross section as it is translated along the spine, held at some constant angle to the spine, and transformed according to the sweeping rule (Fig. 2). Complex objects may be composed of several cylinder primitives by specifying the coordinate transformations between the respective spines.

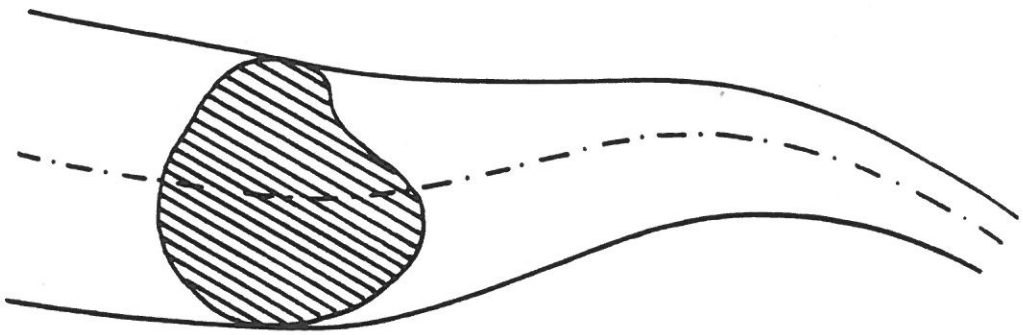

Figure 2: Generalized cylinder

MARR and NISHIHARA 78 point out stick figures as examples for human use of spine-based representations. BROOKS 81 describes the use of generalized cylinders for the vision system ACRONYM. Models for classes of objects with variable shape can be specified by using variable parameters, possibly constrained within certain limits.

Object centered representations are compact and efficient shape descriptions, from which all projections can be readily computed. Thus, in principle, we have answered the question as to the knowledge required for object recognition. Unfortunately, it is not conceivable that object recognition can be carried out efficiently solely on the basis of object centered models. Consider recognition of a simple object, say a spoon. In order to match an unknown object with the spoon model, projections have to be computed for all possible view points. With an angular separation of 10 degrees this amounts to roughly 15000 projections. If applied to all models which might be in question, recognition takes on the form of a gigantic trial-and-error process.

One might propose to use observer-centered models which represent object shape in terms of view point dependent visible surface descriptions. Observer-centered models can be readily matched with scene data. However, each object would be represented by 15000 models according to the number of possible projections - which is again inconceivable.

It does not seem possible to reconcile the requirements of recognition efficiency and storage efficiency by means of trade-offs between object centered and observer centered models. Instead, shape knowledge must be augmented by a separate body of knowledge which supports recognition. In particular, recognition knowledge should provide clues or constraints on the possible view points, given certain image features.

BROOKS 81 gives a detailed account of recognition in ACRONYM based on generalized cylinder models. Recognition is supported by a so-called prediction graph which contains image features predicted from the object model. The prediction graph is different from a tentative projection (which also predicts image features) in one important aspect. It contains features which are invariant or quasi-invariant with respect to a certain range of projections. For example, collinear object features always

project into collinear image features. Similarly, parallel features remain parallel for a subset of projections. Also, the ratio of contour width against length may be quasi-invariant for a certain range of projections. Hence recognition knowledge in ACRONYM specifies invariant image features for ranges of projections.

A complementary view is taken by WALTZ 79. He suggests to specify shape in terms of differences or transformations with respect to well-known prototypes. Applied to different views of an object, this entails complete shape descriptions for a few typical views augmented by recognition knowledge in terms of differences encountered in other views.

IKEUCHI 81 proposes a completely different way of representing shape and recognition knowledge. A surface is described by the extended Gaussian image (EGI), which is the distribution of surface normals normalized with respect to surface area. The surface of a convex polygon can be uniquely reconstructed from its EGI, but in general, different surfaces may have the same EGI, which is a disadvantage of this representation. The main advantage is the use of surface normals which can be immediately related to the visible surface normals supplied by low-level vision. In order to match a visible surface EGI to a model EGI, however, a 3 degree-of-freedom search for the best matching view point would have to be carried out. At this point recognition knowledge comes into play. Ikeuchi proposes to enrich the EGI by two view point dependent features. First, the ratio of area projected onto the image plane against the original surface area can be precomputed for each viewer direction. For example, this value will be large for an ellipsoid viewed from perpendicular to the axis and small viewed along the axis. The same quantity computed for the unknown surface removes one degree of freedom for possible view points. Second, the direction of the axis of inertia of the projected surface area can be precomputed for each viewer direction. The same quantity can be computed for the unknown surface, which removes a second degree of freedom. Thus

EGI matching can be performed in a vastly reduced search space.

Experiments with human vision also seem to indicate the use of special recognition knowledge. Humans can recognize objects with varying ease depending on several factors, including the familiarity of an object, the markedness of typical viewing directions and the amount and direction of rotation with respect to typical views (if any). From experiments reported in ROCK 79 and SHEPARD 79 one can conclude that humans possess both, the ability of visualizing, e.g. performing mental rotation and projection to match model and object, and the ability to use certain shortcuts, which might be called recognition knowledge. One such example is the preference of interpretations which derive from a typical view by a rotation about a vertical axis.

4. High-level vision

Image understanding as discussed up to this point may be summarized shortly as follows:

(i) Extract useful image features
(ii) Interpret image features in terms of 3D surfaces
(iii) Recognize objects by shape

This is the traditional single-image paradigm of computer vision. Some ten years ago a survey article on vision could have ended here, perhaps hinting at context information which might eventually be brought to bear, or pointing out the need of more world knowledge to guide the recognition process.

Today, one can report about work from two sources which have contributed to a changing paradigm of vision: motion analysis and natural language processing. Before starting a discussion of this work it is necessary to clarify the scope of what is called high-level vision in this article.

High-level vision begins where object recognition ends. For simplicity, we shall ignore any interaction of high-level vision with lower-level processes and assume that for each image of a sequence, object recognition has been successfully completed. Hence the input for high-level vision can be assumed to consist of

- object names
- object shapes
- object positions
- view point
- illumination
- instance of time

for each image of a sequence, plus object identities between images. This will be called a geometrical scene description. The output of high-level vision cannot be defined as precisely. It should be an explicit representation of the meaning of an image sequence. In order to gain some understanding of the scope of "meaning", it is helpful to consider several examples.

1) Trees waving in the wind
2) Landscape passing by the window of a moving train
3) Aquarium with fish swimming about
4) Bees performing their dance in front of a bee hive
5) Children playing in the street
6) A goal keeper's parade in a soccer game
7) A street scene showing garbage collection
8) A dachshund biting the mailman
9) Russian tanks crossing the Elbe from east to west
   (DARPA's favourite example)
10) Buster Keaton's silent movie "The General"

The examples range from simple scenes to complex scenes, although an ordering according to "meaningfulness" does not seem possible. Consider example 4: Should a vision program be capable to give a biologist's interpretation of the bee dance? Probably not. How

about example 6: Should a vision program determine whether a goal was scored? After all this is what humans would focus on. Finally, should a vision program "laugh" at Buster Keaton? If it didn't, one could not say that it understands the movie.

It does not seem possible to define the output of an image understanding system other than with respect to typical human image understanding. Furthermore, to achieve human performance a vast amount of knowledge of various kinds is required. It ranges from physical, biological and chemical foundations to social rules and habits, from psychology to history and politics. In this respect computer vision is not different from natural language understanding. More specifically, high-level vision knowledge is to a large part identical with knowledge required for natural language understanding. This explains why natural language research has been mentioned earlier as one of the sources for progress in high-level vision.

High-level vision output also poses a communication problem. How can one verify whether an (abstract) interpretation has captured the right meaning? For methodological reasons, lists, graphics, print-outs of symbol structures, etc. are inadequate, since these data require interpretation by human inspection. Natural language communication is one important way of avoiding this dilemma. (Observable actions are another.) Hence image sequence description (in natural language) must be considered a valuable tool for high-level vision research.

In the remainder of this chapter, work on motion interpretation will be reviewed. Motion concepts are an important ingredient for high-level image sequence understanding. They are also a good subject for research since there are examples ranging from simple to extremely complex motion. This will become apparent when asking the notorious question: What knowledge is required to determine instantiations of certain motion concepts in an image sequence?

There are several basic motion concepts which can be computed from the geometrical scene description using only geometrical templates, i.e. models pertaining to the geometry of motion. In BADLER 75 simple toy scenes are analyzed for "bouncing", "pushing", "hiding", "swinging", etc. Actually, Badler's concepts should not be equated with verb meanings, for example his definition of "bounce" would also apply to a bird landing on the ground and starting again. Yet, his work exposes important properties of such motion. First, complex concepts may be decomposed into simpler concepts by temporal segmentation, e.g. a swing into its back and forth parts. Second, concepts can be organized in a specialization hierarchy, e.g. a horizontal motion may be a roll or a slide.

TSOTSOS 80 presents geometrical motion concepts in a systematical framework and without false semantics. His primitive concepts are area change, location change, length change, and shape change. Higher-level concepts like translate, rotate, contract, etc. are defined in terms of these primitives. Tsotsos' domain of application is left ventricular heart motion. This involves special motion concepts which are only meaningful in this domain. The knowledge required for interpreting "scenes" in terms of these concepts is defined by composition of geometric motion concepts and by specialization using constraints. It does not seem possible, however, to apply this representational scheme to richly structured real world domains as will be seen later.

Many motion concepts correspond to verbs in natural language, thus research on the representation of verb meanings is relevant for high-level vision. One must take care, however, to separate linguistic issues from language-independent concepts. Only the latter are interesting for image understanding. MILLER 72 identified 12 semantic components for English motion verbs. They describe geometrical and physical aspects (change-of-location, change-of-motion-state, change-of-physical-properties, change-of-location-in-some-medium, velocity, direction) as well as intentional and linguistic aspects (causative, permissive,

propellant, instrumental, deictic, reflexive). While the first group of components is similar to the geometrical motion concepts of Tsotsos (which are based on Miller's work), the second group includes features which cannot be computed from a geometrical scene description. Several of these require high-level knowledge about intentional actions. For example, the concept of "avoid" (a car avoiding an obstacle) can only be recognized if knowledge about a typical car-driver's intention to steer clear of obstacles is available. Of course, one may try to recognize avoid-situations solely from geometrical data, but this would be an overinterpretation similar to Badler's bounce. Two of Miller's components describe linguistic features (deictic and reflexive verbs) which do not concern vision.

OKADA 80 pertains to both, motion verbs classification and scene description. Okada uses 20 semantic features, e.g. 'displacement', 'deformation', 'change-in-quality', 'start and stop' to decide which of a set of about 1200 primitive japanese verb concepts applies to a given scene. In his experiments he employs sequences of line-drawings as image data and an extremely simple knowledge base. He does not show how higher-level vision knowledge should be organized to recognize more meaningful verb concepts.

From the preceding one can conlude that the recognition of motion concepts and, by the same token, motion description becomes problematic as soon as non-geometric knowledge is involved. This has also been the experience of project NAOS which deals with natural language description of traffic scenes (NEUMANN 82). While motion concepts such as 'start', 'stop', 'accelerate', turn off' may be recognized by comparing geometrical models with the geometrical scene description (MARBURGER et al. 81), the majority of verbs relevant for traffic scenes requires knowledge which cannot be as easily provided, e.g. 'rasen'. NOVAK 82 points out context knowledge (i.e. the spatial and temporal surroundings), standard properties (i.e. typical velocities) and pragmatic knowledge as three such knowledge categories. It is not possible

to represent this knowledge centered around motion frames as in TSOTSOS' work. Instead representational units similar to MOPs (SCHANK 80), EMOPs (KOLODNER 81) or subscripts (WALTZ 81) may be useful which are designed to bring together all constituents of an event or episode.

This concludes the discussion of high-level vision. It has been restricted to motion understanding since this is the only major body of vision research which goes beyond object recognition. Strong connections to natural language research have been pointed out but have not been followed up. This area diserves much further research before presentation in a survey.

5. Conclusions

The task of understanding and describing image sequences has been discussed from the knowledge point of view. For each of the major phases - low-level vision, object recognition and high-level vision - knowledge required to derive certain descriptions has been characterized. Low-level vision processes exploit general physical knowledge and a small number of fundamental principles. Object recognition is mainly based on a priori knowledge about object shapes and features which support recognition. Higher-level vision bridges the gaps between object recognition and silent movie understanding. Work on motion understanding has been reviewed to point out the open problems rather than solutions of recognizing high-level concepts.

References

Agin 72
Representation and Description of Curved Objects
G.J. Agin
Memo AIM-173, AI Laboratory, Stanford University, 1972

Badler 75
Temporal Scene Analysis: Conceptual Descriptions of Object Movements
N.I. Badler
Report TR 80, Department of Computer Science University of Toronto, Toronto/Canada 1975

Baker and Binford 81
Depth from Edge and Intensity Based Stereo
H.H. Baker and T.O. Binford
IJCAI-81, pp. 631-636

Barnard and Thompson 80
Disparity Analysis of Images
S.T. Barnard and W.B. Thompson
IEEE-PAMI-2 (1980) 333-340

Barrow and Popplestone 71
Relational Descriptions in Picture Processing
H.G. Barrow and R.J. Popplestone
Machine Intelligence 6 (B. Meltzer, D. Michie, eds.)
University Press Edinburgh, 1971, 377-396

Barrow and Tenenbaum 81
Interpreting Line Drawings as Three-Dimensional Surfaces
H.G. Barrow, J.M. Tenenbaum
in: J.M. Brady (ed.), Computer Vision, North-Holland Publishing Co. Amsterdam, 1981, 75-116

Binford 71
Visual Perception by Computer
T.O. Binford
presented to IEEE Conference on Systems and Control (Dec. 1971), Miami/Florida

Binford 81
Inferring Surfaces from Images
T.O. Binford
in: J.M. Brady (ed.), Computer Vision, North-Holland Publishing Co. Amsterdam, 1981, 205-243

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

Brooks 81
Symbolic Reasoning Among 3-D Models and 2-D Images
R.A.Brooks
in: J.M. Brady (ed.), Computer Vision, North-Holland Publishing Co. Amsterdam, 1981, 285-348

Dreschler 81
Ermittlung markanter Punkte auf den Bildern bewegter Objekte und Berechnung einer 3D-Beschreibung auf dieser Grundlage
L. Dreschler
Dissertation, Fachbereich Informatik, Universitaet Hamburg, 1981

Duda and Hart 73
Pattern Classification and Scene Analysis
R.O. Duda and P.E. Hart
Wiley-Interscience, 1973

Grimson 81
From Images to Surfaces
W.E.L. Grimson
The MIT Press, 1981

Horn 75
Obtaining Shape from Shading Information
B.K.P. Horn
in: P.H. Winston (ed.), The Psychology of Computer Vision, McGraw-Hill, 1975, 115-156

Horn and Schunck 81
Determining Optical Flow
B.K.P. Horn and B.G. Schunck
Artificial Intelligence 17 (1981) 185-203

Ikeuchi 81
Recognition of 3-D Objects Using the Extended Gaussian Image
K. Ikeuchi
IJCAI-81 (1981) 595-600

Ikeuchi and Horn 81
Numerical Shape from Shading and Occluding Boundaries
K. Ikeuchi, B.K.P. Horn
in: J.M. Brady (ed.), Computer Vision, North-Holland Publishing Co. Amsterdam, 1981, 141-184

Kolodner 81
Organization and Retrieval in Conceptual Memory for Events or CON54, where are you?
J.L. Kolodner
IJCAI-81 (1981) 227-233

Lowe and Binford 81
The Interpretation of Three-Dimensional Structure from Image Curves
D.G. Lowe and T.O. Binford
IJCAI-81, 613-618

Marr and Hildreth 80
Theory of Edge Detection
D. Marr and E. Hildreth
Proc. R. Soc. London, B207 (1980) 187-217

Marr and Nishihara 78
Representation and Recognition of the Spatial Organization of Three Dimensional Shapes
D. Marr, H.K. Nishihara
Proc. Royal Society B 200 (1978) 269-294

Marr and Poggio 79
A Theory of Human Stereo Vision
D. Marr and T. Poggio
Proc. R. Soc. London, B204 (1979) 301-328

Miller 72
English Verbs of Motion: A Case Study in Semantics and Lexical Memory
G. Miller
in: A.W. Melton and E. Martin (eds.), Coding Processes in Human Memory, V.H. Winston and Sons, Washington/DC 1972, 335-372

Neumann 81
3D-Information aus mehrfachen Ansichten
B. Neumann
in: B. Radig (ed.), Modelle und Strukturen,
Informatik-Fachberichte 49, Springer Verlag
Berlin-Heidelberg-New York 1981, 93-111

Neumann 82
Towards Natural Language Description of Real-World Image Sequences
B. Neumann
GI - 12. Jahrestagung, Informatik Fachberichte, Springer 1982 (to appear)

Nevatia 74
Structured Description of Complex Curved Objects for Recognition and Visual Memory
R. Nevatia
STAN-CS-74-464, Ph.D. Thesis, Computer Science Dept., Stanford University, Stanford/CA

Newell 82
The Knowledge Level
A. Newell
Artificial Intelligence 18 (1982) 87-127

Nishihara 81
Intensity, Visible-Surface and Volumetric Representations
H.K. Nishihara
in: J.M. Brady (ed.), Computer Vision, North-Holland Publishing Co. Amsterdam, 1981, 265-284

Novak 82
On the Selection of Verbs for Natural Language Description of Traffic Scenes
H.-J.Novak
in: W. Wahlster (ed.), GWAI-82, Springer Fachberichte (this volume)

Okada 80
Conceptual Taxonomy of Japanese Verbs and Sentence Production from Picture Pattern Sequences
N. Okada
Information Science and Systems Engineering, Oita University, Oita 870-11/Japan (December 1980)

Rock 79
Form and Orientation
I. Rock
Proc. NSF Workshop on the Representation of Three-Dimensional Objects, R. Bajcsy (ed.), Philadelphia/PA, May 1-2, 1979

Schank 80
Language and Memory
R.C. Schank
Cognitive Science 4, No. 3 (1980)

Shepard 79
Connections between the Representation of Shapes and their Spatial Transformations
R.N. Shepard
Proc. Workshop on the Representation of Three-Dimensional Objects, R.Bajcsy (ed.), University of Pennsylvania, Philadelphia/PA, 1979, pp. N-1 through N-20

Stevens 81
The Visual Interpretation of Surface Contours
K.A. Stevens
in: J.M. Brady (ed.), Computer Vision, North-Holland Publishing Co. Amsterdam, 1981, 265-284

Tsai and Huang 81
Uniqueness and Estimation of Three-Dimensional Motion Parameters of Rigid Objects with Curved Surfaces
R.Y. Tsai and T.S. Huang
Report R-921 (October 1981) Coordinated Science Laboratory University of Illinois at Urbana-Champaign

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

Ullman 79
The Interpretation of Visual Motion
S. Ullman
MIT Press, 1979

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

Waltz 81
Toward a Detailed Model of Processing for Language Describing the Physical World
D.L. Waltz
IJCAI-81 (1981) 1-6

Witkin 81
Recovering Surface Shape and Orientation from Texture
A.P. Witkin
in: J.M. Brady (ed.), Computer Vision, North-Holland Publishing Co. Amsterdam, 1981, 47-74