REPORT T-5 NOVEMBER, 1972 UILU-ENG 72-2256

CSL COORDINATED SCIENCE LABORATORY

COMPUTER-COGNITION OF NATURAL SCENES OUTLINE OF A PROJECT

F. P. PREPARATA
S. R. RAY

UNIVERSITY OF ILLINOIS - URBANA, ILLINOIS

COMPUTER-COGNITION OF NATURAL SCENES
OUTLINE OF A PROJECT

by

F. P. Preparata and S. R. Ray
COORDINATED SCIENCE LABORATORY
University of Illinois at Urbana-Champaign
Urbana, Illinois 61801

November, 1972

COMPUTER-COGNITION OF NATURAL SCENES

OUTLINE OF A PROJECT

F. P. Preparata and S. R. Ray
Coordinated Science Laboratory
University of Illinois at Urbana-Champaign, 1972

The rationale of the semantic approach to the cognition of natural scenes is critically discussed and analyzed in the context of some recent important experiments in pictorial artificial intelligence. This critique leads to the definition of the desirable specifications of a computer-based image interpretation experiment, whose salient features are: 1) the cognitive approach applies both to the identification and to the feature extraction levels; 2) pre-processing is based on textural properties rather than on intensity alone; 3) the algorithm produces as a description of a scene a tridimensional scheme containing both semantic and geometric inferred attributes. The implementation of such project is to be carried out in the immediate future.

1. INTRODUCTION

Any approach to the interpretation of images by means of an automatic system must make reference to some semantics, which represents the system's knowledge of the classes of images to be interpreted. Thus, semantics is in the domain of the "meanings" of the entities which "appear" in the image.

The early approaches, however, were confined to the recognition of images which referred to the extremely simple semantics of a set of competing hypotheses: for example, the recognition of arabic numerals or of letters of the alphabet. In this case, image interpretation reduces to a standard classification problem, expressed by a zoning of a multidimensional space of parameters (features). The selection of the features to be used in classification is certainly a very important problem. A significant step forward was taken when structure was added to the feature collection, still with reference to a simple semantics. This structure in the feature domain can be legitimately called syntax, i.e., the relational organization of some simple graphical elements (primitives), which can be legitimately called lexicon. Typical in this respect is the recognition of the profiles of chromosomes as the closed concatenations of simple curves (arches, segments, etc.) [1].

It is obvious, however, that the classes of images which lend themselves to such simple description at the semantic level are rather few and not very interesting. In other words, the semantic model of a set of competing hypotheses is totally inadequate for the great majority of images. Therefore, an extremely significant advancement was the introduction of

structure at the semantic level. This is equivalent to recognizing that in most cases images are events to be described rather than samples to be classified. In other words, interesting images represent the "articulation" of simpler constituents, whence the name of articular analysis. The semantic model therefore must reflect the structure of the events to be interpreted. An example of this approach is the semantics of high energy nuclear events (Narasimhan [2]), which is reflected in a syntax operating on a lexicon of traces (trajectories) in bubble-chamber photographs.

Other examples of this approach (referred to for convenience as the "semantics of events" approach) appeared in more recent years. Superficially quite different from Narasimhan's project, it is now possible to associate these other projects from the unifying viewpoint of the semantics of events. In most of these projects, a simple semantics of events was chosen. This was done presumably with the intent to simplify the problem. A typical choice is the semantics of plane-bounded objects [3,4,5], sometimes even with the additional constraint of trihedral vertices [5]. That is, we are dealing with images of heaps of plane-bounded objects, such as cubes, pyramids, and prisms. This choice of semantics--or model--has several effects and implications which should be carefully scrutinized.

1. It is argued intuitively that the choice of a simple semantics reduces the complexity of the recognition task (as we shall see, this is only partially true). A simple semantics, as the one issuing from the plane-bounded objects' constraint, has the interesting consequence of being directly reflected in the image syntax. In other words, the syntactic relationships among the image primitives--such as vertices, edges, and regions--can be used, practically with no additional aid, for the interpretation of the picture.

2. Although a rudimentary semantics induces a direct and formally appealing relationship between semantics and syntax--to the satisfaction of grammar-oriented researchers--at the same time it renounces a wealth of additional constraints, which only a richer semantics can possess and which may be very powerful aids for interpretation. These constraints may not only resolve syntactic ambiguities, but also avoid costly syntactic analysis by making it unnecessary. In other words, in the semantics of plane-bounded objects a "cube" is only an abstract geometrical entity with fixed properties, it is not the shape of an object which is contextually related to its environment (for example, a building in a city). In the semantics of plane-bounded objects, the contextual dependence appears to be intra-object, as opposed to inter-object. And the argument could be made that the inter-object context is a more powerful device for interpretation than the intra-object consistency. On the other hand, a semantics of plane-bounded objects, in spite of its inability to express a meaningful <u>global</u> context, plays a very important role in the analysis of the <u>local</u> inter-object relationships--such as occlusion, support, etc.--since all objects can conveniently be considered as being locally plane.

3. The preceding considerations elicit the view that a simple semantics is a double-edged device. But the most serious criticism to such a simple semantics rests not on pragmatic grounds (i.e., what we can do with the tools at our disposal), but on the philosophical grounds that it is by no means clear that the generalization to the "real-world" semantics is only a quantitative step.

A step in the direction of richer semantics was taken by considering the semantics of scenes which are closer to those of the real-world [6,7].

For convenience we shall refer to it as the "semantics of natural scenes," where "natural" denotes our every-day tridimensional environment. This new viewpoint is discussed in great detail in our previous paper [6], so that we will only recall its highlights. Since we intended to demonstrate the capabilities of contextual interpretation (inter-object), we purposely weakened the image syntax. This was done by using a very coarse resolution in picture acquisition, so that considerable syntactic information would be suppressed. The entire interpretation task reduced to the algorithmic association of picture regions (near-uniform domains) with semantic components, i.e., concepts. The structure of the semantic model (the "map") reflected the inter-object relationships, and geometric proximity of picture regions was used as a heuristic for semantic relatedness. The coarse resolution resulted in the decomposition of the image into a rather small number of regions: although this "clumping" may be objected to, the interpretation of the obtained regions on an essentially contextual basis was pleasantly satisfactory. In the approach of Barrow and Popplestone [7], an object (for example, a cup) is identified as a syntactic construct of simpler geometric constituents. These constituents in turn are identified by interpreting image regions on the basis of their geometric features. Although they adopt a semantics of natural scenes, apparently they make little or no use of inter-object contextual dependencies.

At present, we feel that a satisfactory approach to the computer-cognition of natural scenes must not be limited to the exploitation of purely contextual devices, since this greatly impairs the acquisition of fine image details. Local properties of the image, such as shapes of regions, edges,

etc., must also be used, and the symbolic representation of the interpreted image (i.e., the system's output) must contain at least vestiges of the inferred geometrical structure of the scene. This more mature viewpoint is the informing principle of a project which will be implemented in the immediate future and is outlined in detail in the next section.

2. OUTLINE OF A PROJECT

In the framework of automatic cognition of natural images we shall develop a software system capable of interpreting bidimensional views of tridimensional scenes of the real world. Scenes will be presented to the system as digitized versions of gray-scale photographs. The objective of the system is to produce a stylized representation of the geometric and semantic structure of the scene, which we now describe.

First we shall discuss a sufficiently general format of tridimensional scenes. In each scene two main classes of constituents can be discerned: background and objects. The background can be thought of as a "container" for the objects in the scene. The container has a typical syntactic structure; it consists generally of a horizontal FLOOR and of one or more vertical WALLS. FLOOR and WALL have generic and specific semantics. The generic semantics of FLOOR, for example, is that of being the support of most of the scene objects (since gravity is such a fundamental feature in the real world). The specific semantics concerns types of walls or floor. For example, whether WALL is the sky or the wall of a room, whether FLOOR is that of a room or a field, an ocean, etc. Note that, whereas the generic semantics calls for general-purpose analysis procedures (such as occlusion or support analysis), the specific semantics instead will introduce a global context, suggestive of the objects which are plausible in the

scene. It is the latter device which we feel plays a crucial, although not autonomous, role in scene interpretation; this project is aimed at further substantiating this thesis.

Upon completion of the cognitive interpretation, an adequate representation of the scene will consist of the identification of the container and the production of a partial ordering in the depth dimension (the inferred dimension) of the identified objects. This choice of representation is suggested by the psychological intuition that depth plays more a qualitative than a quantitative role in scene understanding. However, a coarse assessment of depth will help estimate the "real" size of objects, thereby providing an important inferred feature for object interpretation.

In our approach, processing will occur in a sequential top-down fashion. As in our earlier experiments [6], top-down refers to an ordering from general to specific. It is convenient to view the interpretation process as a sequence of levels or stages. The following general criteria govern the execution of the various processing levels:

1. At each level, a small set of heuristics is used (strong heuristics). These heuristics are ranked in order of decreasing strength, to be empirically assessed. The heuristics are tried successively on the scene constituents. Processing of the level terminates either with convincing evidence at some test of the sequence, or with poor evidence by default: at this point, control is transferred to the immediately lower level. This criterion reflects the principle that the processing effort should be commensurate to the information acquired through it. In other words, weak heuristics which provide little confidence and yet may require a substantial

processing effort, should be avoided. In fact, exhaustive testing of all cases conceivable at any level is a mental prejudice originating in the analysis of problems governed by logic, where it has full legitimacy. However there is no reason to assume that this principle should be applied in the analysis of cognitive processes. Rather, it is our conviction that strong heuristics at a lower level may be more illuminating in the total interpretation task than weak heuristics at the current level.

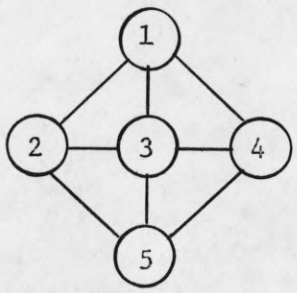
2. The preceding criterion, which is dictated by efficiency, is consistent with the fact that in cognitive processes--as distinct from logical processes--truth-values of statements cover a continuous range. This is equivalent to saying that in cognitive processes statements are tentative, and that their acceptance or rejection is postponed until processing is completed. Needless to say, this calls for the necessity of backtrack provisions at all levels.

3. Preprocessing (feature extraction) and interpretation activities must occur interactively throughout the execution of the cognitive algorithm. In this fashion the cognitive approach can be applied also to the feature extraction phase. In fact, we view this as an important improvement over our original approach in which preprocessing and interpretation were cascaded activities (with no opportunity of feedback). In our current approach, each algorithmic level will consist of carefully matched preprocessing and interpretation: in this manner, depending upon the accumulated context the most rewarding preprocessing will be executed. For example, the complex operation of measuring shape will be performed only on those objects for which, on the basis of their preliminary interpretation, shape is likely to be an important discriminant.

On the basis of these general criteria, we now give a detailed outline of the steps of the cognitive algorithm.

1. Background acquisition. The preprocessing phase consists of the construction of the regions which are likely candidates for forming the background. These regions are clearly those which touch the scene frame. The growth of regions is based, as usual, on intra-region uniformity and inter-region difference. To evaluate uniformity, we feel that the standard approach based on gray level is inadequate, since it easily provides errors of both kinds (misses and false hits). A measure which takes into account the important property of texture appears more adequate. As a compromise between effectiveness and efficiency, we propose to adopt a 3-parameter vector for region formation, obtained as follows. For domains of 8 x 8 picture elements (pixels), we obtain the FFT: from this we derive the 3-component vector $[I,F_x,F_y]$, where I is the average intensity, and F_x and F_y are, respectively, the largest horizontal and vertical frequencies whose intensities exceed a threshold controlled by the average intensity I. Standard classification techniques (training set) will be used to obtain a statistically valid criterion for an evaluation of uniformity based on the given texture vector. Regions will be formed by concentric scanning starting from the frame. The frame is at first partitioned into uniform segments, which are successively extended towards the center until acceptable disuniformities are encountered. The background candidates are selected among the regions previously formed as follows: those regions whose larger dimension is parallel to the peripheral edge they touch, or those regions with the largest contact segments with peripheral edges. After the background candidates have

been found, their adjacency diagram is constructed. With sufficient generality, this diagram is a connected subgraph of the following graph:



where, with reference to a hypothetical "box," vertices 1 and 5 are "ceiling" and "floor," respectively, and vertices 2, 3, and 4 are lateral "walls." Typically, however, a landscape scene consists of the subgraph 3-5. The interpretation of the background is very important because of its great power as a context setter. The broad initial categorization occurs between LANDSCAPE and ROOM (also referred to as outdoors and indoors). We tentatively choose the following strong heuristic: "Choose ROOM with high confidence in the following cases: if the adjacencies 2-3 or 3-4 are near-vertical straight edges, or if the adjacencies 5-i or 1-i (i=2 or 3 or 4) are straight edges consistent with the hypothesis of not being at infinity (we assume that the elevation angle of the observer is known a priori); Choose LANDSCAPE with high confidence in the following cases: The background graph is of type 3-5 and either the adjacency 3-5 occurs at the horizon line or is a nonstraight edge." The weak heuristic is the choice of LANDSCAPE whenever the strong heuristics tests do not yield a satisfactory answer. The tentative decision ROOM requires some verification of the syntactic consistency of the boundary planes, within the framework of the perspective transformation, and the establishment of a coarse frame of reference in the tridimensional model.

The tentative decision LANDSCAPE requires the classification of the types of floor and wall on the basis of measurable attributes.

2. "First Level" Object Acquisition. The next step consists in the interpretation of major (or "first-level") objects in the scene. The first-level objects are selected on the basis of apparent size and adjacency to the picture frame, excluding those regions already interpreted as background. For interpretation, features are derived from the scene with the aid of the depth partial ordering, including:

a) Actual Size (unary)

b) "Supported by" relationship (binary)

c) Aspect Ratio (unary)

as well as the depth ordering itself. (binary)

After correlation of these (strong) features with the semantic map, the need for more specific features will, in general, be selectively indicated by the map itself. When this need is indicated, we derive further attributes (in specific cases) including:

a) Measures of shape

(primarily straightness of edges)

b) Textural Features

(e.g., component frequencies and component frequency ratios)

and select most plausible interpretations on that evidence.

It is to emphasized that the decision to derive attributes and the attributes which are derived is wholly dependent upon each particular situation encountered. This decision is based upon failure of the plausibility measure to indicate sufficient certainty of a particular interpretation and,

if the uncertainty is large, the use of semantic map information to indicate which attributes would be most decisive.

3. "Second Level" Object Acquisition. Acquisition and interpretation of second-level objects proceeds in a manner similar to the treatment of first level objects, except that all picture regions not accepted either as background or first level regions are treated as second level.

All fundamental attributes used in the first level interpretation apply in second level interpretation. In addition, the binary attribute, "enclosed in," will be used as evidence of whole-partness.

There is one important new feature in our current approach that is worth pointing out. We still maintain the hierarchical organization (background--first level--second level) which allows a top-down interpretation from general to specific. Moreover, we process the various levels in a formally identical way, in the sense that the hypotheses formulated at the preceding level establish the context for the current level, and in turn the processing of the current level simultaneously tests the previous hypotheses and establishes new ones. However, in our previous approach we adopted a generality which does not seem to be required by the problem, in that we allowed for an arbitrary number of cascaded levels. It appears--on psychological intuition--that the dynamic range of the levels of details which need to be spanned in analyzing an image rarely exceeds three levels (the three levels discribed above). Some especially detailed object-classes (such as a house in a landscape), however, may require special subprograms for analyzing additional levels: this approach may be viewed as a displacing of the dynamic range of attention.

REFERENCES

[1] R. S. Ledley, "High Speed Automatic Analysis of Biomedical Pictures," Science, Vol. 146, pp. 216-223, October, 1964.

[2] R. Narasimhan, "Labeling Schemata and Syntactic Description of Pictures," Information and Control, Vol. 7, pp. 151-179, June, 1964.

[3] A. Guzmán, "Decomposition of a Visual Scene into Three-Dimensional Bodies," Proc. FJCC, Vol. 33, pp. 291-304, 1968.

[4] P. H. Winston, "Learning Structural Descriptions from Examples," Ph.D. Thesis, M.I.T., September, 1970.

[5] D. A. Huffman, "Impossible Objects as Nonsense Sentences," Machine Intelligence 6, (J. Doran, ed.), University of Edinburgh Press, 1971.

[6] F. P. Preparata and S. R. Ray, "An Approach to Artificial Nonsymbolic Cognition," Information Sciences, Vol. 4, pp. 65-86, Spring, 1972 (also available as Report R-478, Coordinated Science Laboratory, University of Illinois, Urbana, July, 1970).

[7] H. G. Barrow and R. J. Popplestone, "Relational Descriptions in Picture Processing," Machine Intelligence 6, (J. Doran, ed.), University of Edinburgh Press, 1971.