EXPERIMENTS IN SCHEMA-DRIVEN INTERPRETATION OF A NATURAL SCENE*

Cesare C. Parma
Allen R. Hanson
Edward M. Riseman

COINS Technical Report 80-10

April 1980

This paper will appear in the proceedings of the NATO Advanced Study Institute on Digital Image Processing and Analysis, to be held in Bonas, France, June 23 - July 4, 1980.

*This research was supported by the Office of Naval Research under Grant N00014-75-C-0459, and the National Science Foundation under Grant MCS79-18209.

ABSTRACT

The system under development, VISIONS, is an investigation into general issues in the construction of computer vision systems. The goal is to provide an analysis of color images of outdoor scenes, from segmentation (or partitioning) of an image through the final stages of symbolic interpretation of that image. The output of the system is intended to be a symbolic representation of the three-dimensional world depicted in the two-dimensional image, including the naming of objects, their placement in three-dimensional space, and the ability to predict from this representation the rough appearance of the scene from other points of view. Research in segmentation and interpretation has been separated into the development of two major subsystems with quite different methodologies and considerations.

The focus of this paper is upon the interpretation system. The primary emphasis will be on the development of strategies by which several knowledge sources (KSs) can be integrated using expected knowledge stored in structures called 3D and 2D schemas, each of which may be general or specific to the scene under consideration. A series of increasingly more difficult experiments is outlined as an experimental methodology for developing schema-driven (e.g., top-down) control mechanisms; each succeeding experiment will assume a set of weaker constraints, representing image interpretation tasks where a decreasing amount of knowledge of the situation is available. Experimental results show current capabilities of a number of KSs and the effectiveness of a specific 2D schema in the interpretation of a scene.

IN MEMORIUM

Cesare Parma, 1947-1979

On August 30, 1979 Cesare Parma, a graduate student in the COINS Department, was struck and killed by lightning during a sudden thunderstorm in Amherst, Massachusetts. Many of the results on schema-driven image interpretation in Section VI of this paper were due to the hard work and creativity of Cesare. All the members of the VISIONS group benefited greatly from the blend of his strong intellect and the natural warmth of his personality. We are deeply saddened by this loss, and this paper is dedicated to the memory of this fine individual.

ACKNOWLEDGEMENTS

We wish to acknowledge the COINS graduate students, members of the VISIONS group, who have contributed to this work. This includes Frank Glazer, Ralf Kohler, John Lowrance, Paul Nagin, Terry Weymouth, Tom Williams, and Bryant York. The details of their work have appeared or will appear in a variety of reports and doctoral dissertations.

I. KNOWLEDGE DIRECTED PROCESSING

The system being developed is called VISIONS and is designed to provide an analysis of color images of outdoor scenes, from segmentation through symbolic interpretation. The VISIONS system is decomposed into two major subsystems: a "low-level" system which processes the large numeric arrays of sensory data, and then feeds the "high-level" interpretation processes, which construct a description of the world portrayed in the scene. The output of the system is to be a symbolic model of the three-dimensional world depicted in the two-dimensional image, including the names of objects, their placement in three-dimensional space, and the ability to predict from this model the rough appearance of the scene from other points of view.

The original design of the VISIONS system was heavily influenced by a commitment to knowledge-directed interpretation, and this commitment has been maintained. The emphasis of this paper is on the form of knowledge structures, called schemas, and on the control structures necessary to coordinate a variety of complex processes, which are referred to as knowledge sources, or KSs [LES77]. A knowledge source is a process which specializes in the formation of an hypothesis about an interpretation of the image, based upon a particular type of available visual cue and partially processed sensory data. For example, the perspective KS might infer the physical size of an object depicted by some region in the image, and the object size KS might order, in terms of a confidence measure, the plausible object identities based upon that size. There are serious problems to be faced in the general application of these processes in an integrated fashion. In our system schemas are the means by which we deal with the problems of control of the KSs. A schema is a knowledge structure about a particular visual concept, say a road scene, with procedural components for properly invoking a subset of KSs in a coordinated manner.

The effectiveness of many AI systems appears to be derived from either the constraints available via prior knowledge, or the restrictions of a specific task domain, or a combination of both. The natural language understanding system of Schank [SCH75] is heavily directed in a top-down manner by knowledge structures called scripts; recently, they have proven sufficient for extracting summary descriptions of a large number of actual wire service news stories [SCH79]. The HARPY speech understanding system [LOW76], one of the most effective speech systems to date, embeds a grammar and vocabulary in a network of expected utterances. The system operates top-down by matching paths in the network (which represent possible sentences) against the utterance. One can view this system in terms of a schema for each sentence and the representation of this information in a storage efficient form.

There are various special-purpose vision systems whose effectiveness may be traced directly to the utilization of domain-dependent simplications, for example blood-cell analysis, assembly line parts inspection, etc. It is our belief that the use of schemas ([ARB77] or frames [MIN75], or scripts [SCH77]) provides a bridge between general-purpose and special-purpose systems [BAL78, HAN78c, NEV78]. The development of an individual schema and the verification that it is applicable may be as tractable as the development of a particular strategy in a special-purpose system. Knowledge of the front view of a particular house to some degree should be usable in a manner similar to knowledge of the structure of a complex machine part on a conveyor belt.

It is generally agreed that while research in computer vision is definitely progressing, the problems have been found to be extremely difficult. Our initial efforts have been directed at the construction of a system with sufficient flexibility and generality to explore a variety of issues without requiring substantial systems modifications as the research evolved. As to be expected, the price of such efforts at generality is slower development of the system than we desired, slower than would have been possible with a less flexible special-purpose system. Because of the magnitude of the problem, our research methodology has been to focus on modular components of the system under the constraints of a general system design.

We wish to make it clear that we do not believe that computer vision ought to be primarily a top-down process. Many important mechanisms of human vision appear to be constructive processes which transform sensory data without recourse to semantics [BAR78, HOR77, MAR76, MAR78]. The research effort proposed here, however, attempts to direct the application of some of these processes under the guidance of knowledge-oriented constraints. It will be interesting to see the degree to which this approach can be made general.

II. ADDITIONAL RELATED LITERATURE

There is a very large body of literature that is relevant to the development of effective computer vision systems. In fact it spans the fields of computer science, electrical engineering, cognitive psychology, mathematics, art, etc. with topics that include the physics of light and surfaces, shadows and highlights, image segmentation, color, texture, two- and three-dimensional shape, perspective, occlusion, motion, stereopsis, representation of knowledge, inference, and more. It is not feasible to review this literature here, but a recent book Computer Vision Systems [HAN78a], edited by the authors, documents the state-of-the-art in many of these areas. Here we choose just to mention a few of the many efforts in image understanding systems and leave reference of others for the more detailed sections of the paper.

There have been interesting and somewhat successful attempts to integrate the segmentation and interpretation processes. A decision-theoretic approach to image interpretation by Yakimovsky and Feldman [YAK73, FEL74] produced a region merging process that was integrated with semantic interpretation. Effective results on chest x-rays and road scenes were achieved. Tenenbaum and Barrow [TEN77] demonstrated that a constraint-satisfaction process could be used to block erroneous region merges in their interpretation-guided segmentation system (IGS). This system was generalized into a probabilistic relaxation process for propagating constraints under uncertain interpretation [BAR76].

There are a variety of image interpretation systems where the analysis does not employ three-dimensional representations and processes. In such cases, the output of the system usually is the extraction and labelling of relevant entities in the image, for example the labelling of each 2D region with an object identity. Sakai, Kanade, and Ohta [SAK76] produced a partial labelling of major areas in a building scene (though there were only five possible objects in the data base). Shirai [SHI78] has developed a system which fits smooth curved lines to segmented edges; this system has been used to interpret a desk scene containing a variety of objects. Ballard, Brown, and Feldman [BAL78] are using a flexible knowledge-directed system which has been applied to both aerial images and chest x-rays. Levine [LEV78] has been examining scenes of human figures, cartoons, and landscapes; he has obtained interpretations of several cartoon images. Bajcsy [BAJ76] has used a small semantic network to extract river and bridge regions in aerial images. Uhr [UHR78] has been developing a very general system for both segmentation and scene interpretation using a parallel-array processing system called recognition cones; preliminary results have been produced on a house scene. A group at Hughes Research Labs [DUD77] has developed a system for 2D segmentation and matching of objects with long straight lines (such as buildings). Rubin [RUB77] has extended the basic approach of the HARPY [LOW76] speech understanding system to a scene interpretation system for matching an image of a city skyline with a set of such images from different points of view. Mackworth [MAC78] and Havens [HAV78] have addressed issues of control, based on a cyclic theory of perception, in the context of interpretation of a map relation system.

Interpretation systems using three-dimensional representations can be applied to a wider class of imagery but are correspondingly far more complex. Consequently, much of the work in 3D scene description (interpretation) has primarily been restricted to polyhedral models of objects [ROB65, WAL75], although there has been interesting work on generalized cylinders as a representation for curved surfaces [NEV77, AGI72, MAR77]. Another significant body of research has taken place at levels below object recognition, in particular the extraction of surface information based upon a camera model, illumination model, and surface properties [HOR75, HOR77, WOO77, MAR78, BAR78]. This work promises to provide significant insight into constructive mechanisms in visual perception.

Finally, there is related work in speech understanding that has influenced our research, in particular the Hearsay system

[ERM75, LESS77] whose general structure has been followed in our own research.

III. CONTROL STRATEGIES AND IMAGE INTERPRETATION

III.1. Strategies for Controlling the Interpretation Process

In the past we have raised two important issues of control in our system: the basis upon which KSs are to be invoked and the means by which alternative hypotheses provided by KSs are to be used. Our system was organized to deal with the selection of appropriate KSs and a search space of interpretations by employing a hierarchical modular control strategy [HAN78b, WIL77]. This computational mechanism allows user-defined strategies to be constructed hierarchically out of modular components.

This approach required considerable machinery for dealing with issues of search, and some of these issues drew our attention away from the central issues of vision. The top-down approach that is suggested by schemas bypasses problems of recovering from errors and the inherent combinatorics of a search space of alternatives, at least until we more fully understand the reliability, robustness, and redundancy of our KSs when used in this manner. However, as we will point out, the top-down approach does not imply a complete avoidance of bottom-up issues. Schema instantiation and the application of a general schema to specific images, for example, will require the use of bottom-up processes. In this case, however, the purposes and goals of the bottom-up processes are more specific and well-defined.

III.2. Top-Down Interpretation via Schemas

In the following sections we outline a highly structured approach to the development of general top-down image interpretation. A key problem is to develop effective ways to employ schemas after they are somehow accessed. In some of the experimental stages that we will outline, the goal is to interpret an image using either a specific or a general scene schema from either a known or unknown perspective viewpoint. Thus, the relevant scene schema is assumed to be known, but the specificity of the information varies. Before describing our experimental methodology, let us note the difference between specific vs. general schemas, 3D vs. 2D schemas, and known vs. unknown perspective viewpoints.

specific schema - a schema capturing a particular instance of a given type of scene or object, e.g., a particular house, a familiar section of road, or a specific car such as your own;

general (prototypical) schema - a schema representing a standard or prototypical model of a scene or object, such as a house scene, road scene, or car scene, but not any specific house, road, or car scene;

3D schema - the 3D description of a scene or object in a local coordinate system; this involves the representation of surfaces and volumes, and the relationships between them;

2D schema - the 2D appearance of a 3D schema relative to a viewer-centered coordinate system; this is the way a 3D schema would appear from a particular point of view;

unknown perspective viewpoint - in this case a known schema (general or specific) can only be used as a 3D schema, since the relationship between its local coordinate system and the viewer's coordinate system is unknown.

known perspective viewpoint - if the relationship between the coordinate systems of the schema and viewer is known, then the 3D schema can be used to generate a plan for the scene in terms of a 2D schema.

Under this categorization, a general 3D schema is a structure describing default features of objects and general relationships between sets of objects which are expected to hold across a schema class [MIN75]. A specific 3D schema is a general schema in which features and relationships have been assigned (more) precise values and in which features and relationships unique to the particular environment have been added. In fact top-down interpretation of, let us say, a road scene using a general 3D schema would then involve the construction of a specific 3D schema of that road scene.

A specific 2D schema is a transformation of the corresponding specific 3D schema, given an assumed view angle. The transformation according to view angle is necessary in order to match the specific 3D schema to the image. Similarly, a general 2D schema represents a transformation of the general 3D schema given a view angle; in this case, the general 3D features and relationships are mapped into general 2D features and relationships.

III.3. An Experimental Methodology

In a system as complex as VISIONS, there exists a wide range of plausible strategies for guiding the interpretation process. We propose to explore these strategies by means of a set of carefully defined experiments of increasing difficulty and generality. By controlling the amount and type of information provided, different portions of the system can be exercised and different strategies to use the information can be developed.

We separate the schema-driven operation of our system into distinct tasks:

a) Top-Down Interpretation of Images Via Schemas - this involves the utilization of a relevant schema as a top-down plan for interpretation; it requires coordinated application of the KSs, guided by the schema, to various portions of the image.

and b) Bottom-Up Instantiation of Schemas -- this is the process of selecting a schema that is relevant to the interpretation of the image; in effect, it is the problem of finding cues and paths of inference through long term memory which imply a prototypical context which ought to be used.

These tasks overlap a third task which is one of the most general goals of (computer) vision research:

c) Bottom-Up Interpretation of Images -- the construction of a surface/volume description of the physical world in the image without the use of prior high-level knowledge; it is expected that insights into the mechanisms by which this task might be accomplished will be gained by success in achieving the goals set forth in (a) and (b) above, particularly the use of general schemas in interpreting scenes.

Our research effort is currently focussing on tasks (a) and (b), above. Primary emphasis has been placed on schema-controlled strategies for employing the KSs, but there is continuing effort on the instantiation of the relevant schema. The remainder of this section of the paper will outline experimental stages of system development, and later sections will provide experimental results for the first of these stages.

III.4. Experimental Stages in Schema-Driven Interpretation

Stage 1: The specific scene schema is known;
the viewpoint is known.

In Stage 1 experiments, the system is, in effect, told what it will see. It must merely match its highly constrained expectations to what appears in the particular scene. In these experiments, a specific 2D schema is directly available. The research focus is on the structure of the schema, the control structure for driving the KSs directly from the schema, and on mechanisms for consistently integrating the hypotheses returned by the KSs into the schema. This experiment is an exercise of all the components of the system and its success is fairly well ensured. Since the specific 3D schema is available and the point of view is known, a 2D schema can be generated which closely matches the appearance of the 2D image. The 2D schema provides a powerful plan for directing various KSs in processing the image and interpreting the scene. Some of the results cited later in this paper are a partial exercise of this capability. Those results, we emphasize, should be viewed as exercises in demonstrating the integration of the system.

Stage 2: The general scene schema is known;
the viewpoint is known.

Stage 2 tests the system's ability to interpret a scene using a prototypical schema instead of the specific schema. Thus, the general knowledge of road scenes would be used to interpret an image of some particular road scene. The spatial constraints are more general and any given object in the schema may or may not appear. Since the viewpoint of the general schema is known (e.g., looking down the road), the general 3D schema can be used to generate a general 2D schema which then provides a list of key region, line, and vertex features, as well as rough spatial locations and spatial relationships between features that might appear. Strategies are needed that have flexibility in locking onto any relevant characteristics which are extracted from the 2D image. The processed sensory data must be used by the schema in constructing the description of the particular road scene. While certain relationships are expected, for example converging lines of the sides of the road, their existence and location in the image can only be determined by application of some of the KSs.

<u>Stage</u> 3: The specific scene schema is known;
the viewpoint is unknown.

Stage 3 exercises a different processing capability of the system: the ability to manipulate 3D representations in the selection of the probable view angle. It must rotate and translate a 3D description of a particular scene in order to generate a 2D view which matches the scene. The problem is simplified from the general case because the specific 3D schema is made available. Therefore, <u>if</u> the proper viewpoint can be determined, a very good match is ensured (c.f. results of Stage 1 in Section VI.). Here, important information about the viewpoint may be provided by the orientation of line segments, the 2D shape of regions, and spatial relationships between regions in the image. In addition we can attach information about standard viewpoints to the 3D schema.

<u>Stage 4</u>: The general scene schema is known; the viewpoint is unknown.

Stage 4 is an integration of the techniques developed in the first three. The focus here is on the use of bottom-up information to constrain the general relationships found in the general schema and to obtain the most likely view angle. It is a non-trivial extension of Stages 2 and 3 because even the proper viewpoint still leaves a potentially large degree of variability in the matching and interpretation process. Success here will be dependent upon the quality of the KS's developed during the first three stages and the effectiveness of the control strategies developed in the last two stages.

III.5. <u>Bottom-Up Instantiation of Schemas</u>

<u>Stage 5</u>: The general scene schema is unknown and must be hypothesized and verified.

Even if experiments in Stage 4 are successful and a general 3D schema from an unknown viewpoint can be used for interpreting an image, there is still the serious problem of determining the relevant schema to employ. In a general system for scene analysis, the knowledge base would be expected to contain many schemas. Given the high cost of computation expected to be associated with schema-controlled KS invocation, all possible schemas cannot be applied to see which best fits the situation. Many researchers have worried about problems of search and error recovery in an enormous search space of possibilities. We have decomposed the problem of applying the correct schema from the problem of schema instantiation so that the different issues involved do not get confused.

The accuracy of schema instantiation is dependent upon the degree to which features can be extracted from the sensory data. As bottom-up mechanisms begin to construct a model of the image, features of this model can be matched against the available schemas in long-term memory in order to select a schema that is relevant to the image. The problems here are related to both 2D and 3D schemas. Since the viewpoint is unknown, features of 2D shape which are extracted from the image cannot be matched directly against the schema. Rather, knowledge of possible perspective transformations of the shape features must be used during the matching. This is facilitated by storing with the schema prominent 2D features from important or common points of view; this can be accomplished by means of "standard-view" orientation vectors attached to the schema or to parts of the schema. However, these vectors do not obviate the need for additional mechanisms which can suggest plausible orientations if the given scene does not conform to the standard views.

Inference networks [DUD76] may prove to be effective in integrating the implications of a number of uncertain hypotheses at various lower levels of representation. They allow the effect of multiple hypotheses (in the form of probability updates on nodes) to be simultaneously propagated in the network. After propagating these inferences up to the schema level, schemas with high posterior probabilities can be selected. There are a variety of problems which have not yet been solved, such as the problem of loops (closed paths) in inference paths, the difficulty of estimating joint probability distributions of n nodes, and errors due to inconsistency of binary (or m-ary, m less than n) approximations of the joint probability distributions.

Stage 5 is the least constrained of the experiments thus far, and depends primarily on the ability of the bottom-up constructive mechanisms to transform the scene data in such a way that the appropriate higher level KS's can be applied and a schema instantiated. The development of these constructive mechanisms foreshadows Stage 6, one of the most general and difficult problems in vision.

III.6. Bottom-Up Interpretation of Images

Stage 6: The goal is to construct a (partial) 3D surface/volume description without access to schemas.

It can be argued that research on vision ought to begin with the bottom-up constructive mechanisms and the development of a general theory of vision. Often humans can recognize surface and volume properties and develop a sense of 3D space even when there are virtually no object semantics in the image. There is much to be learned from more constrained approaches which do not involve higher level knowledge [HOR75, HOR77, BAR78, MAR78]. However, they cannot be expected to solve the general vision problem. Given the complexity of our images, we do not expect that the current KSs will be sufficiently reliable, or generally relevant, to be effective over most of the image without guidance by schemas.

Nevertheless, the insights and mechanisms developed in the previous stages should significantly overlap those needed in Stage 6. We expect some of the KSs (e.g., occlusion, 2D shape, spectral attribute matcher) to provide useful information in the general interpretation construction process. Stage 4 experiments require the system to lock onto visual attributes in the image which are consistent with schema expectations. The location, size, and number of objects in a schema (e.g., shrubs in front of a house, the number of windows on a wall of a house, etc.) will vary. Therefore, mechanisms which use the visual characteristics in a manner consistent with bottom-up analysis are required in order to use the general 3D schema.

IV. THE SEGMENTATION ALGORITHMS OF THE LOW-LEVEL SYSTEM

The VISIONS research group has maintained a long-standing research effort in low-level image analysis. Our goal has been to produce a system which can initially provide a segmentation to drive the image interpretation process, and which later can receive semantic feedback to direct low-level processing in the refinement of that segmentation. We cannot discuss the full range of our segmentation efforts; they are documented in a series of reports and papers [NAG79, KOH79, HAN78b, PRA79, PRA80, OVE79, HAN80a]. Here, we limit our discussion to a brief description of two algorithms, an edge relaxation algorithm and a histogram-guided region relaxation algorithm. Both the edge relaxation process and the region formation process are undergoing continuous development.

All algorithms are implemented in a simulation of a parallel hierarchical machine architecture, called a "processing cone", for processing images [HAN74, HAN80b]. The cone is related to similar structures proposed by [UHR74, TAN78, TAN80, ROS79a,b].

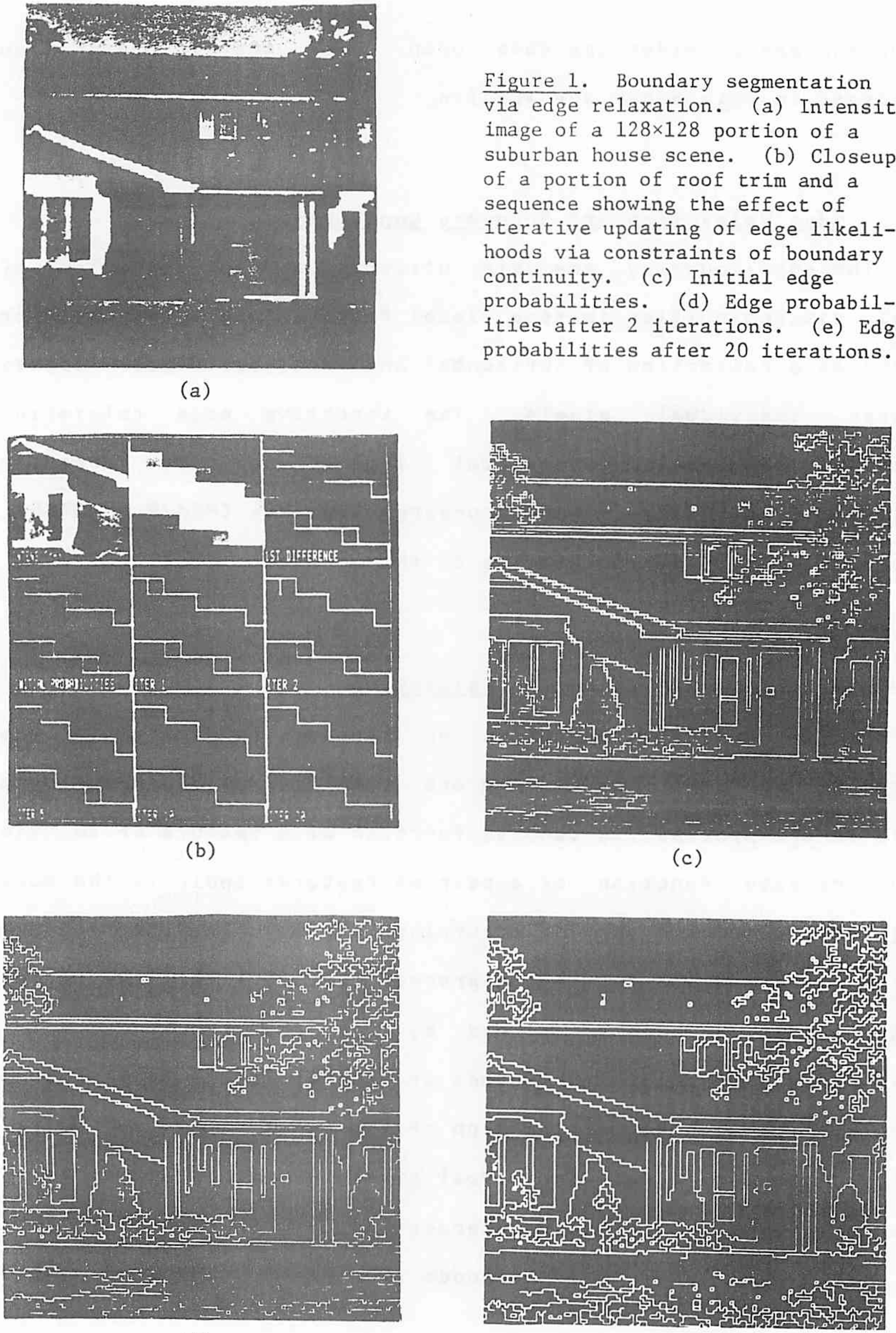
The segmentation processes basically involve two complementary relaxation labelling processes [ROS76, ZUC77, DAV76] for partitioning images into regions and boundaries, either of which can be preceded by a sophisticated smoothing algorithm [OVE79] in a preprocessing pass on the image. The boundary formation process responds to local changes in the data, while the region formation process is sensitive to global similarities in the data. An earlier version of the region algorithm has provided the data upon which the interpretation processes in this paper are applied.

IV.1. Edge Relaxation and Boundary Continuity

The edge/boundary analysis utilizes a representation of local discontinuities in some visual feature (e.g., intensity or color) as a collection of horizontal and vertical edges located between individual pixels. The iterative edge relaxation processes then allow contextual interactions to organize collections of edges into boundary segments [PRA79, HAN78b]. Figure 1 provides sample results of this process.

IV.2. Histogram-Guided Region Relaxation

Region analysis is based on cluster detection in the histogram of some visual feature [HAN78b, NAG79]. Prominent peaks in the probability density function of a feature or in the joint density function of a pair of features indicate the most frequently occurring (or co-occurring) values in the feature space. The region formation process therefore utilizes global histogram cluster labels, defined by the peaks, with pixels. These peaks also allow likelihoods of cluster labels (computed as a function of the spatial location of the peaks relative to the spatial location of each individual pixel in feature space) to be associated with each pixel. Interactions between the label sets of pixels in local neighborhoods are then used to organize



(a)

(b)

(c)

(d)

(e)

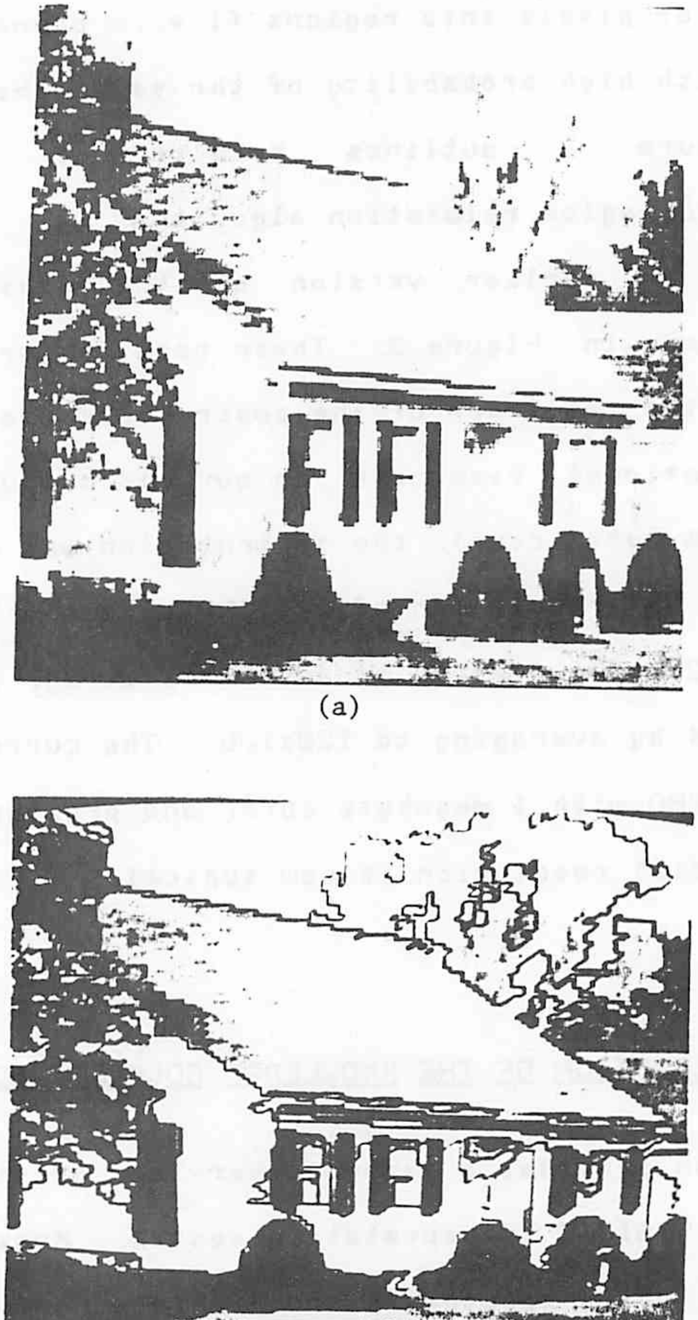
Figure 1. Boundary segmentation via edge relaxation. (a) Intensity image of a 128×128 portion of a suburban house scene. (b) Closeup of a portion of roof trim and a sequence showing the effect of iterative updating of edge likelihoods via constraints of boundary continuity. (c) Initial edge probabilities. (d) Edge probabilities after 2 iterations. (e) Edge probabilities after 20 iterations.

connected sets of pixels into regions (i.e., connected sets of pixels all with high probability of the same label constitute a region). Figure 2 outlines results of applying the histogram-guided region relaxation algorithm.

Results of an earlier version of the region relaxation algorithm appear in Figure 3. These results form the basis of experiments in the remainder of the paper. Because of previous limited computational resources on our old computer facilities (PDP-15 with 96K bytes core), the segmentation was obtained from an image with a resolution of 128x128 pixels. This image was derived from a 256x256 quarter of a 512x512 array, which was then further reduced by averaging to 128x128. The current processing is on a VAX 11/780 with 1 megabyte core, and processing of images with higher spatial resolution is now typical.

V. SUMMARY DESCRIPTION OF THE KNOWLEDGE SOURCES AND INITIAL EXPERIMENTS

This section provides a general overview of the knowledge sources in the VISIONS interpretation system. Knowledge sources are the means by which hypotheses are generated and verified. In some cases, the KSs have been developed only to the point where the results are reasonable. The advantage of this approach is that it allows a minimally complete system to be configured and run. The input/output and functionality of each KS is clearly specified and can be improved as time and resources permit.



(a)

(b)

Figure 2. Region segmentation via relaxation histogram cluster labels. (a) Initial intensity image of a 128×128 portion of a house scene derived by averaging from an image of higher resolution (previous limitations on computational resources dictated this limitation). (b) Resultant segmentation superimposed on intensity image. Note that there is a difference in aspect ratio in this image due to differences in the displays used to generate the picture.

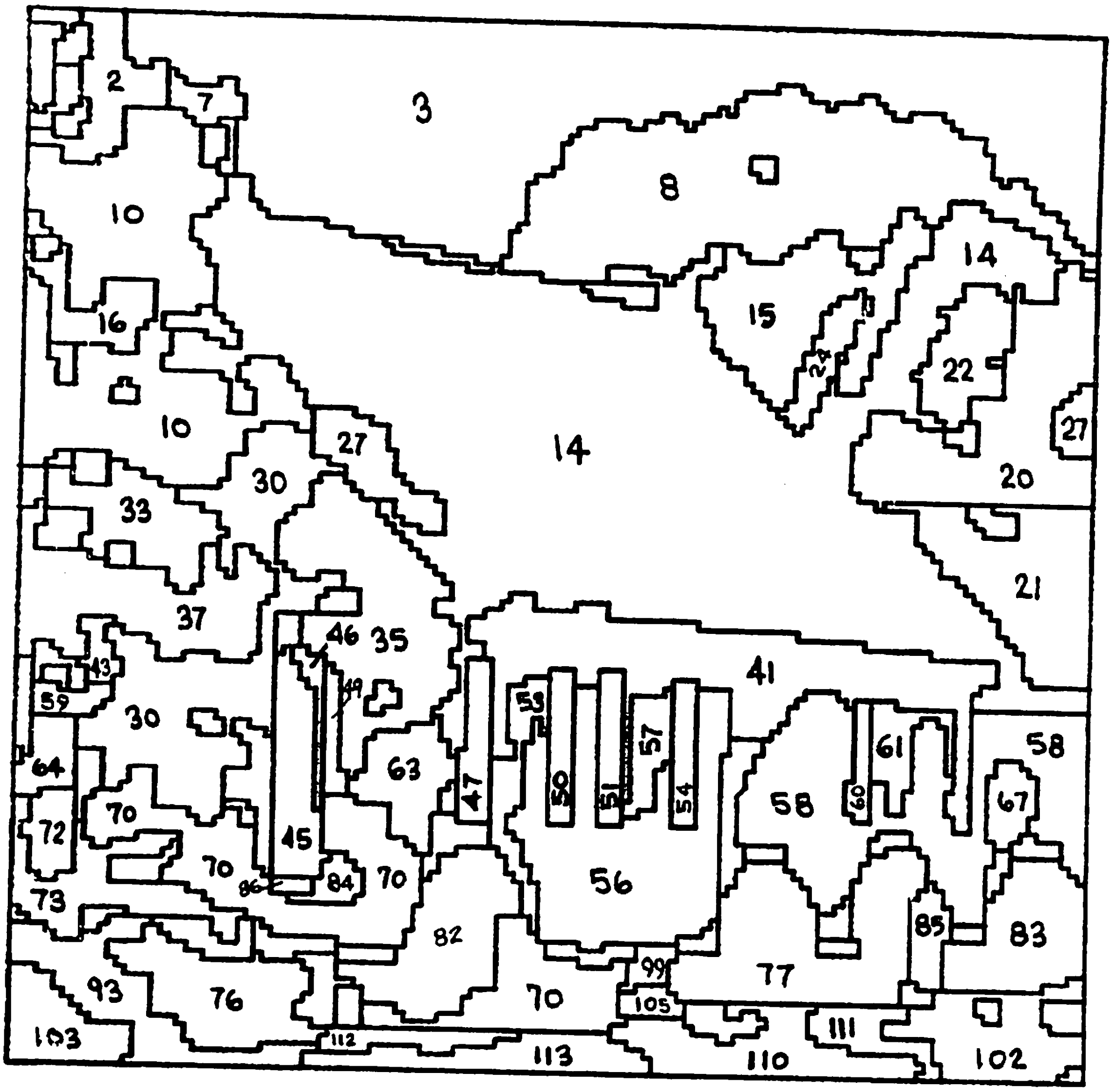


Figure 3. Segmentation data used in experiments. These results of region formation via relaxation on cluster labels were produced by an earlier version of the algorithm which produced the results in Figure 2. The region segmentation has been converted to a region boundary representation and region labels are shown. They form the basis of the experiments described in later sections.* Note that only large regions or regions mentioned in paper are numbered, but all regions have a unique label.

*The integration of the edge and region segmentations is the focus of the current Ph.D. research of Ralf Kohler.

A set of eleven modular KSs and several representations will be briefly reviewed. While we cannot discuss each of these in detail in the limited space of this paper, a short discussion of each KS and, wherever possible, a simple example of local results is provided. However, these local results must be viewed in the context of the evolving design of the whole system [HAN78b,c].

A base-level system has been implemented and is operational to the point where interesting experiments, such as the ones described in the following sections, are being performed. In building this base-level system, an attempt was made to provide sufficient generality of processes and representation -- function and structure -- to allow us to work on different types of scenes, to easily add knowledge in both active and passive form, and to define and execute different types of interpretation strategies.

The reader should note that the results cited in this section were obtained from a version of the system running on the University Computing Center's CYBER-74 time-sharing system. The system is implemented in GRASPER [LOW78], a high level graph processing language built in ALISP [KON75]. The system has been transferred to the COINS Department VAX 11/780 and integration with the VISIONS low-level system is in progress.

Table 1 provides an overview of the set of KSs currently available and briefly discusses the representations employed in various parts of the system. Cross references to more detailed discussions and/or results are included.

TABLE I: SUMMARY OF KNOWLEDGE SOURCES AND REPRESENTATIONS

Name	Brief Statement of Function or Purpose	Cross References
Low-Level Segmentation System*	The goal of the low-level system is the segmentation of an image into visual primitives (regions, boundary segments, and vertices), and the extraction of a range of features to be used by the various knowledge sources (KSs) of the interpretation system.	IV Figures 1, 2, 3
RSV Structure	RSV is a symbolic layered graph structure of regions, line segments, and vertices containing the segmentation results and feature descriptors. This data structure is stored in short-term memory (STM; see below) and represents the processed visual data upon which the interpretation is based.	Figure 8
LTM (Long-Term Memory)	LTM is a hierarchical representation of general (i.e., non-image specific) world knowledge organized into natural levels of abstraction: schemas (stereotypical scenarios), objects, volumes, surfaces, regions, line segments, and vertices.	V.2 Figures 8, 9
STM (Short-Term Memory)	STM is a hierarchical structure of the same form as LTM and used for constructing an interpretation by means of the knowledge sources. An interpretation is then the collection of instantiated nodes in STM. The RSV structure is the bottom three levels -- all other levels are initially empty.	V.2 Figure 8
Inference Net KS	It is a network of a priori probabilities of nodes and conditional probabilities between nodes; it is defined on the arcs and nodes in LTM, and are the means by which implications of local hypotheses may be propagated upward and downward through the layered structure. Any hypothesis generated by a knowledge source can then be used to generate further hypotheses.	III.5 V.6 VI.8 Tables III, V
2D Curve Fitting KS	This KS is designed to produce smooth fits to boundary segments in a segmentation. It utilizes generalized cubic splines, automatic resegmentation of boundaries at points of high curvature, and curve fitting techniques.	V.1 VI.3 Figures 4-7, 21
2D Shape KS	This KS allows symbolic classification of the shape of regions. The confidence that a given image region has a particular primitive 2D shape will be returned. The results allow paths for surface & volume hypotheses via LTM.	V.3 VI.4 Figures 10, 22-24 Table II
Occlusion KS	This KS uses the results produced by 2D shape to analyze junction (vertex) types to produce hypotheses about relative depth relationships between regions. Spline fits to boundary segments produce smoothly varying curves at junctions, which may be analyzed for occlusion cues.	V.4

*Funded by the Office of Naval Research under Contract No. N00014-75-C-0459.

Name	Brief Statement of Function or Purpose	Cross References
Spectral Attribute Matcher KS	It hypothesizes object identities of a region on the basis of a comparison between region attributes (color and texture) and statistics of these features attached to the object nodes in LTM. It is designed primarily for objects for which these attributes are reasonably invariant across images (currently sky, bush, grass, tree, road).	V.5 VI.2 Figures 11, 19, 20
3D Shape KS	It uses a representation for 3D shape with curved surfaces, their organization into objects and object parts, and mechanisms for manipulating the representation. It is called a quilted solid and is defined by collections of Coons' surface patches bounded by cubic splines, in an object centered coordinate system. Quilted solids are joined together by spline blending functions.	V.7 IV.8.3 Figure 12
Perspective KS	The goal is the hypothesis of surface orientation, size, and/or distance in order to produce a partial volume/surface plan of the scene. The current version focusses on relationships between elevation, height, range, and width of surfaces given a camera model and a set of assumptions regarding surface orientation.	V.9 V.10 VI.6 Figures 14, 15, 27 Table IV
Horizon KS	It uses the horizon schema (the most general outdoor schema which relates sky, ground, and horizon) and the camera model to fix the location of the horizon. It is used to filter other hypotheses on the basis of their relationship to the horizon.	V.11 Figures 11, 16
Object Size KS	This module is designed to generate object hypotheses on the basis of the image size of a region. It compares the computed physical (i.e., real world) size of a surface, determined by the perspective module, to the physical size of objects in LTM.	V.12 VI.7 Figures 17, 28
3D Schema	The 3D schema captures stereotypical visual events by organizing subsets of information in LTM into higher order complexes of expected scenarios (e.g., a road scene schema). It may be either specific (a particular known scene) or general. The representation is stored in a local coordinate system and contains control information for top-down interpretation. A projection of a 3D schema produces a 2D schema.	III.2 III.4 V.8 Figure 13
2D Schema	A 2D schema is a projection of a 3D schema from a given point of view. The projection carries along control strategy information and features of the projection (e.g., surface orientation, relative to viewpoint, etc.). It is used to direct top-down interpretation of the image.	III.2 III.4 VI.1-VI.5 Figures 18-28

V. 1. 2D Curve Fitting

The output of the segmentation processes is represented in terms of horizontal and vertical edges for a variety of reasons. They involve concerns about connectedness of edges and the ambiguity that occurs when edges of varying orientation are associated with pixels [RIS77, HAN78b]. It is necessary to transform this rectilinear edge data into a continuous representation. By fitting smooth lines to the data, they more accurately reflect the original visual information. However, various problems occur when the best straight lines are fit to the segments that form the low-level output. The first problem is that the endpoints of a segment do not define the "natural" portion of a boundary over which lines should be fit (refer to Figure 4b). This problem can be avoided by using piecewise linear fits to line segments by decomposing line segments on the basis of points of high curvature, but there are still difficulties. The enlargement of a junction is shown in Figure 4(c) and one can see problems with best-fit straight lines not meeting at a point (Figure 4d), or movement of the location of junctions if pseudo-junctions are formed (Figure 4e). Finally, any type of piecewise straight-line fits cause a discontinuity at the junction in the slope of line segments which are actually portions of a smoothly curving region boundary. This is important in the extraction of surface occlusion cues (Section V.4). These problems are discussed in more detail in [YOR80].

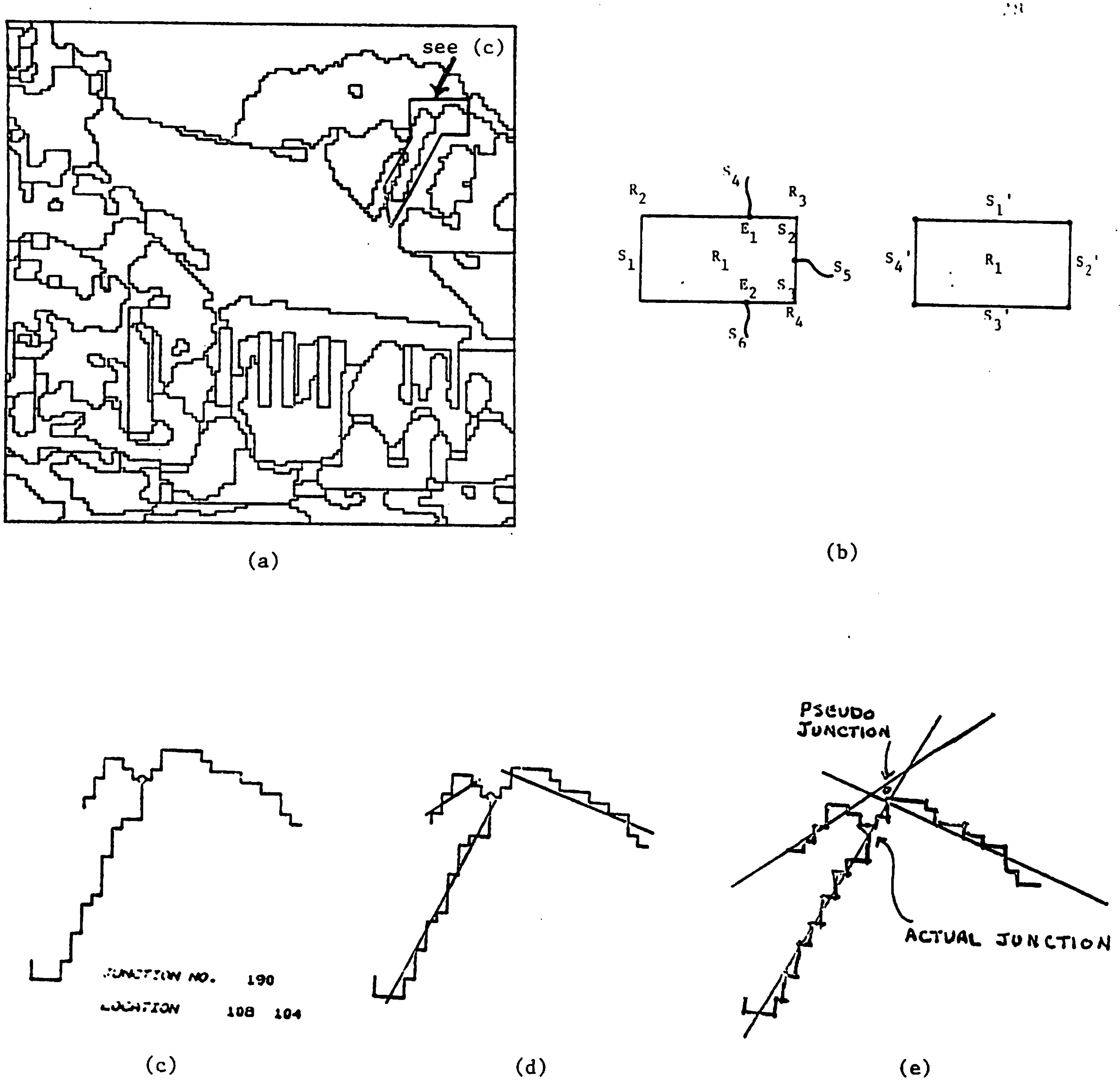


Figure 4. (a) Segmentation of house scene with a typical junction of line segments marked. Line segments are delimited by a line termination or a junction of two or more lines. (b) The segments bounding a region must be restructured by choosing points of high curvature as new junctions in order to obtain correct line fits. (c) Enlargment of junction shown in (a). (d) The best straight line fit to segments emanating from a junction can result in the lines not meeting at the junction. (e) When pseudo junctions are used, actual junction locations are moved and the characteristics of the segments at the junction are lost.

In order to avoid some of these problems, piecewise polynomial functions called splines [AHL67, GOR74] are fit to the set of line segments bounding each region. Splines of degree 1, 2, and 3 are employed: piecewise linear, piecewise quadratic, and piecewise cubic splines [YOR79]. Cubic splines in particular have several nice properties (refer to Figure 5):

a) they are smooth curves -- the function as well as its first two derivatives are continuous in the interval;

b) they are guaranteed to pass through a specified set of points called knots;

c) placement of multiple knots at a single point allows discontinuities to remain;

d) given a set of knots, computation of the spline coefficients is efficiently accomplished via standard algorithms.

The strategy currently in effect is to select points of high curvature as possible knot locations and then use a knot collection procedure to pull nearby knots together. Then splines of all three degrees are fit to the segments. If the piecewise linear straight line has a low RMS error, then the segment between two knots is labelled "straight" and an (R, theta)

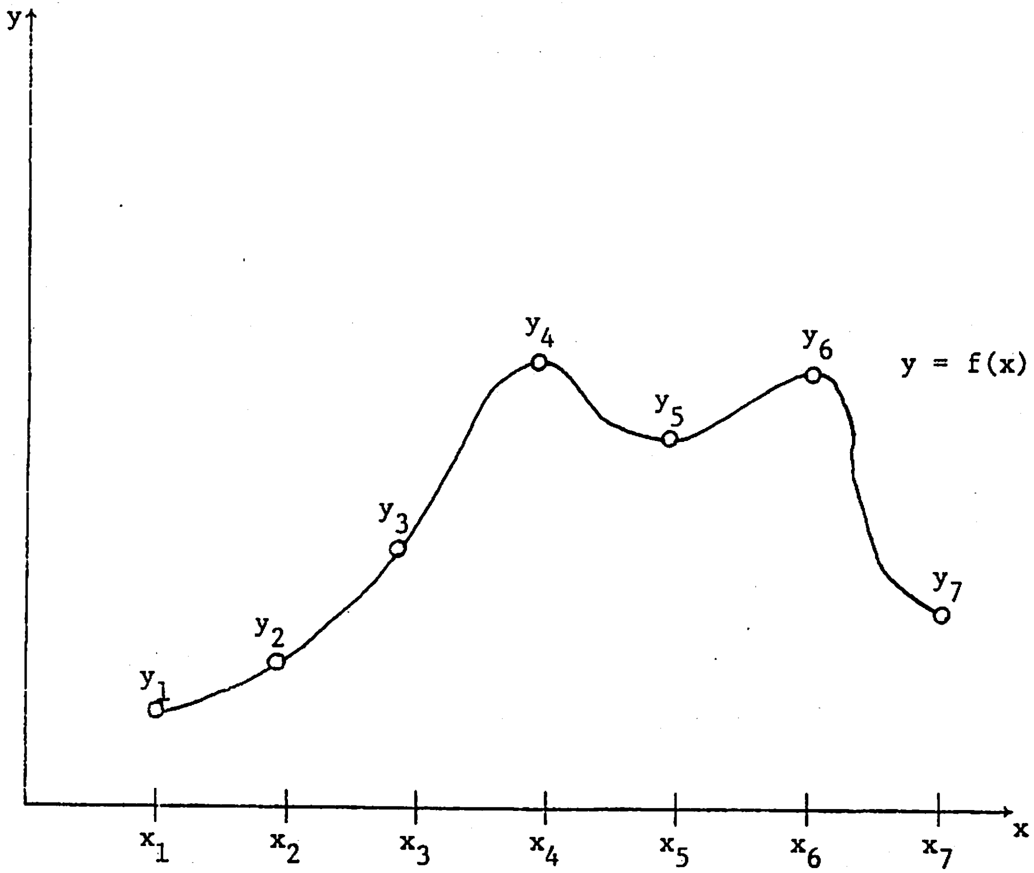


Figure 5. Cubic splines are polynomial functions y = f(x) of degree 3. They define a smooth curve that passes through any specified set of points called knots: (x_i, y_i), i = 1,...,7 in the figure.

parametric representation is used to represent the slope and location (up to co-linearity of the segments). If the straight line fit is not good, then the second degree fit is tested, and if necessary the cubic spline fit is adopted. These points of high curvature are computed on the basis of a modified k-curvature [DAV76], which is the angle that is formed at a given point by straight lines from the given point to the points which are k away in each direction.

The result of knot selection, knot collection, and first and third degree splines for one region is shown in Figure 6. The spline approach has the potential to produce smooth approximations to digital curves and allow a more accurate analysis for junction classification [YOR80], 2D shape analysis, occlusion cues, and surface hypotheses. Although the fits of cubic spline curves shown in Figure 7 are reasonable, there is definitely need for further improvement. The knot selection and collection process was based only upon a local view of curvature; a more global view of curvature may produce more appealing boundary fits.

V. 2. Long-Term Memory (LTM) and Short-Term Memory (STM)

General knowledge about the physical world (or the task domain of interest) is stored in "long-term memory" (LTM). An image will be "understood" in terms of the concepts and relations found in LTM. This knowledge is hierarchically organized into

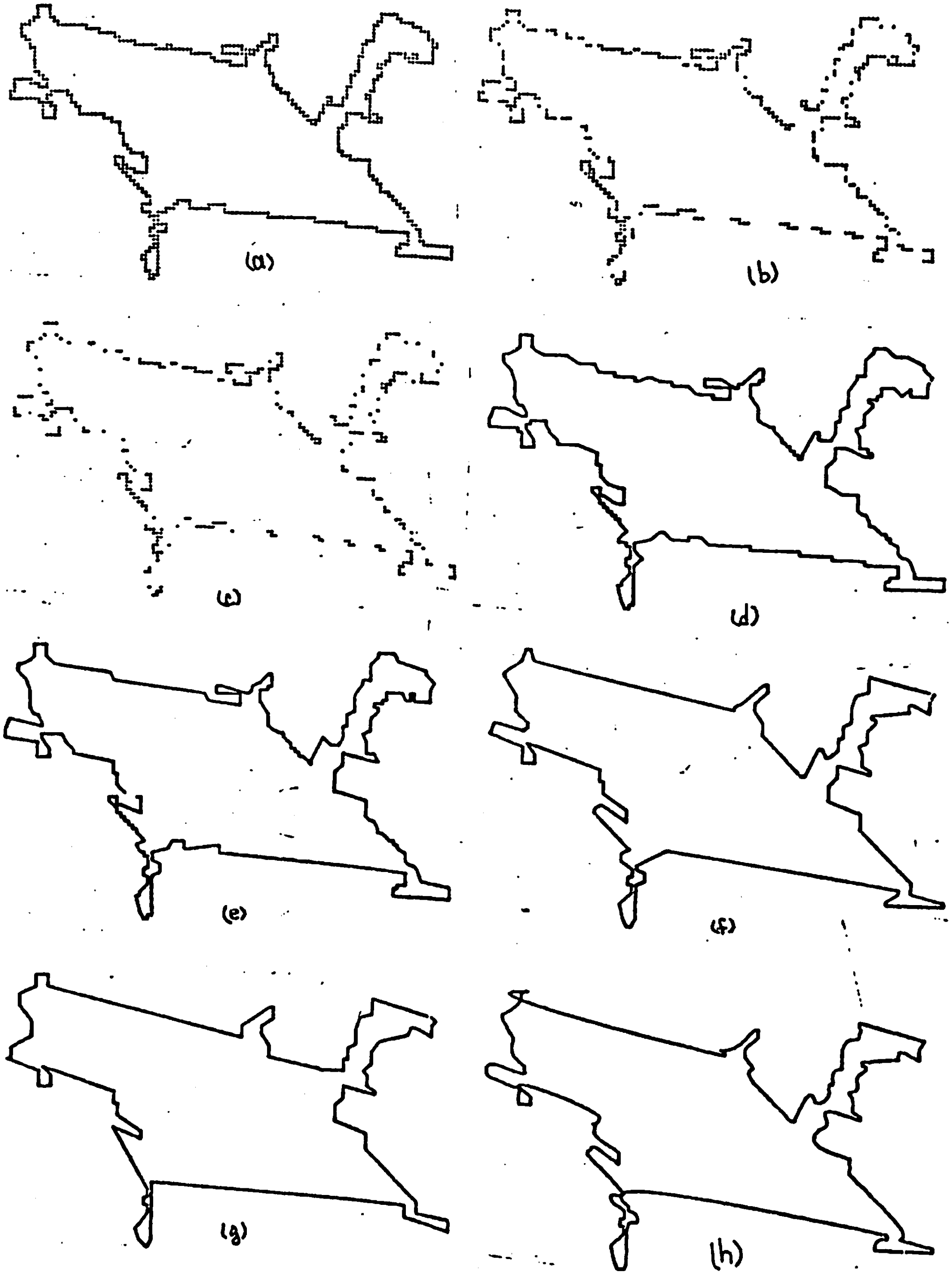


Figure 6. Using splines for 2D curve fitting of a region. (a) Original 646 points along the boundary of region 14. (b) For each point in (a), 3-curvature was computed and all points with absolute value of 3-curvature greater than 0 were retained for the knot collection process. Of the 646 original points, 467 are left. (c) From the 467 points, the knot collection procedure leaves 343: 148 are multiplicity-3 knots and 195 are multiplicity-1 knots. (d) Piecewise linear interpolation of the 3-curvature, 0-thresholded, knot collected boundary of region 14. (e-g) Same as (d), but thresholded at 1, 2, and 3, respectively. (h) Piecewise cubic interpolation to 3-curvature, 2-thresholded, knot collected boundary of region 14.

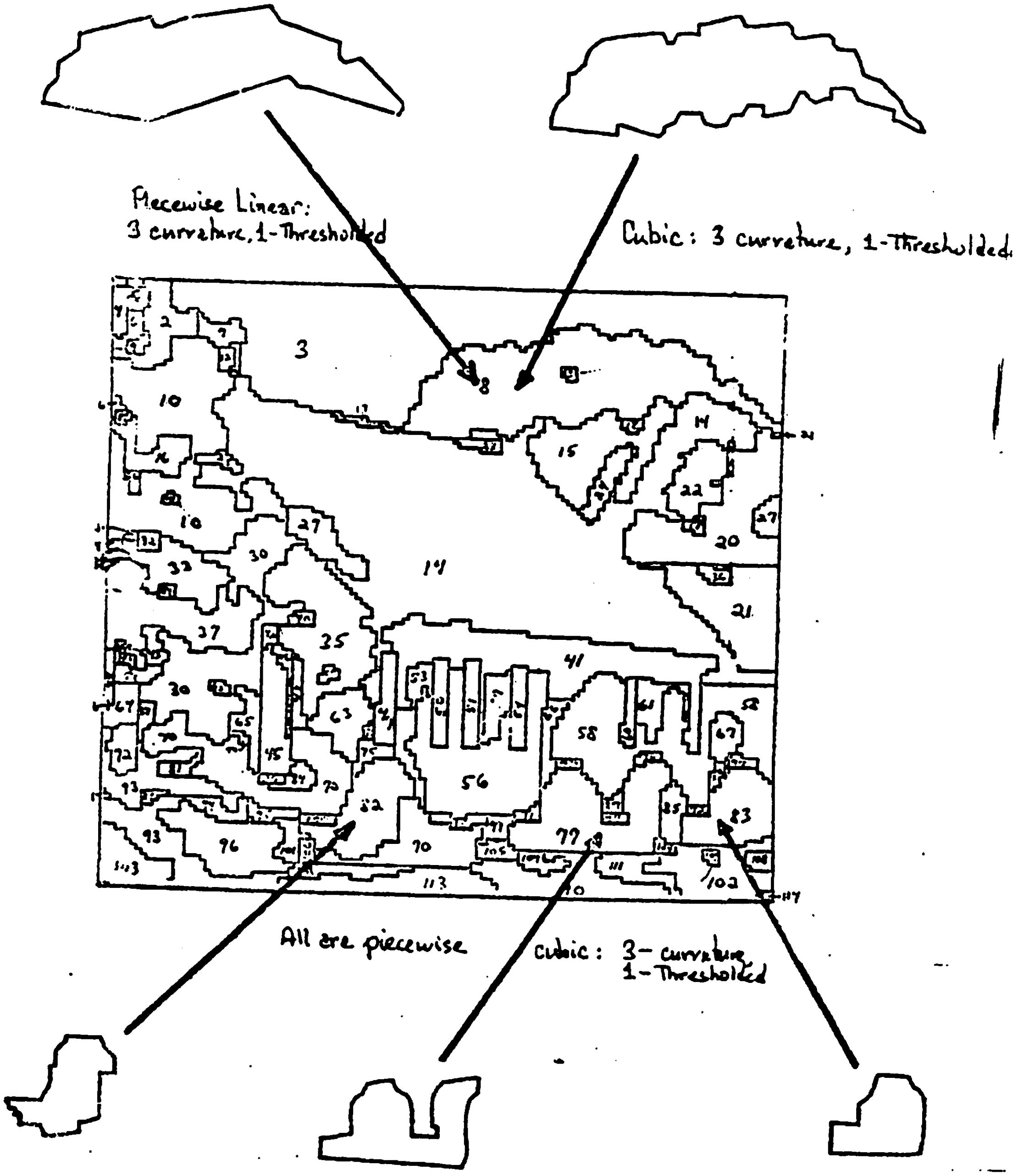


Figure 7. Result of applying 2D curve fitting to selected regions of the image. Knot selection and collection was based on a local view of curvature (three points to either side of the central point); a more global view of boundary curvature may be necessary to produce more appealing fits.

levels which represent a natural abstraction of world knowledge (Figure 8).

Nodes in LTM represent visual primitives with which the system can construct an interpretation, while the arcs represent relations (primarily AND/OR relations) which exist between the primitives. Inter-level arcs represent the paths by which primitives at one level may be related to primitives at levels above and below. These arcs represent paths for hypothesis formation (possible inferences) within LTM; they are used in various ways by other knowledge sources during the interpretation process. Section V.6 discusses how the inference net KS overlays LTM. Figure 9 depicts a representative fragment of the network and describes the size of the network in terms of the number of nodes and arcs.

The interpretation of an image is viewed as a set of instantiations of the nodes in LTM. These instantiations constitute short-term memory (STM) and are shown on the left side of Figure 8. This representation of knowledge, as well as its relationship to the inference net, is the subject of ongoing research by J. Lowrance, a graduate student in our research group. Both STM and LTM are implemented as a layered graph in GRASPER [LOW78], a graph processing language extension to LISP which follows the general approach of [FKR169, PRA71].

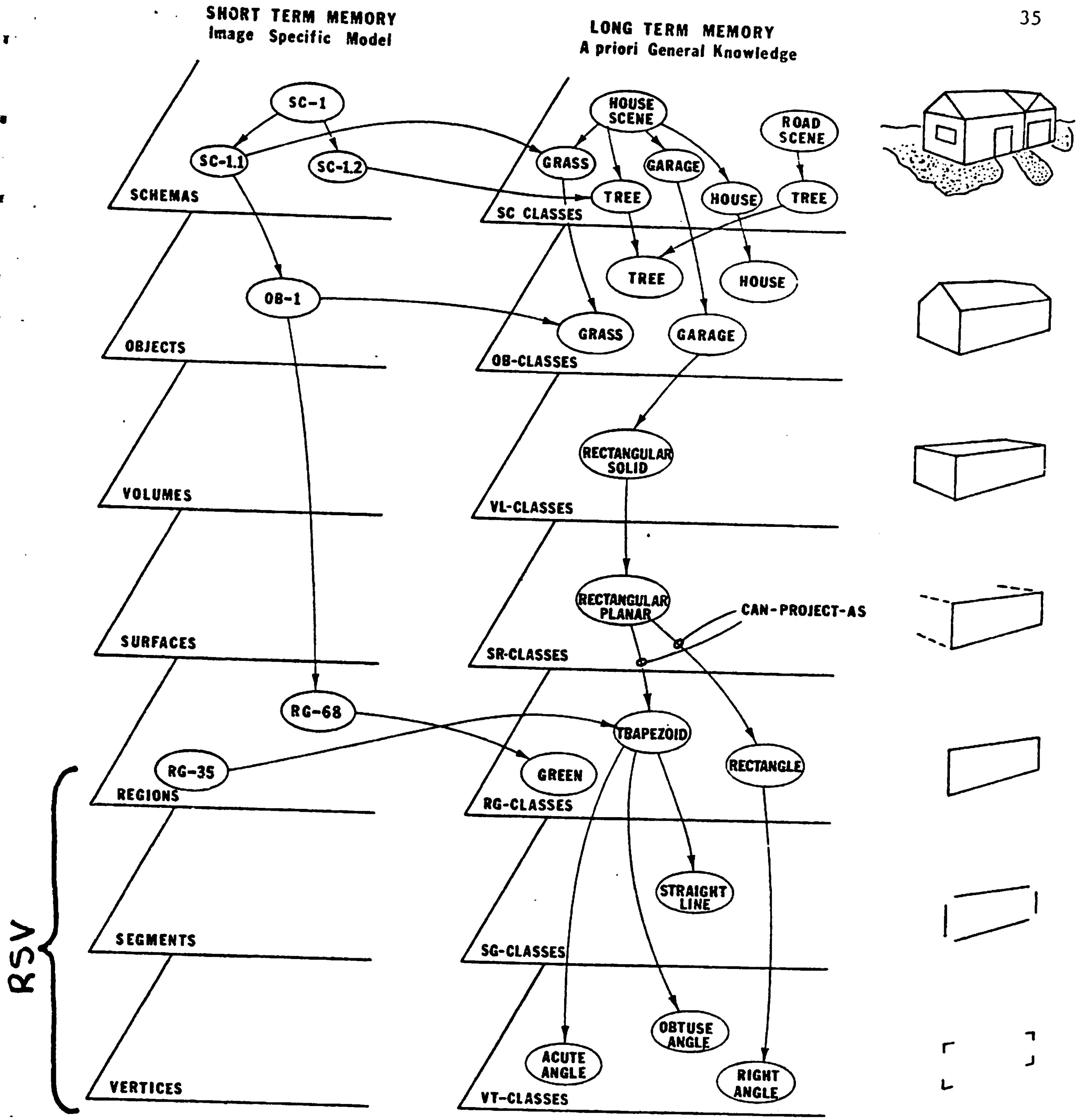
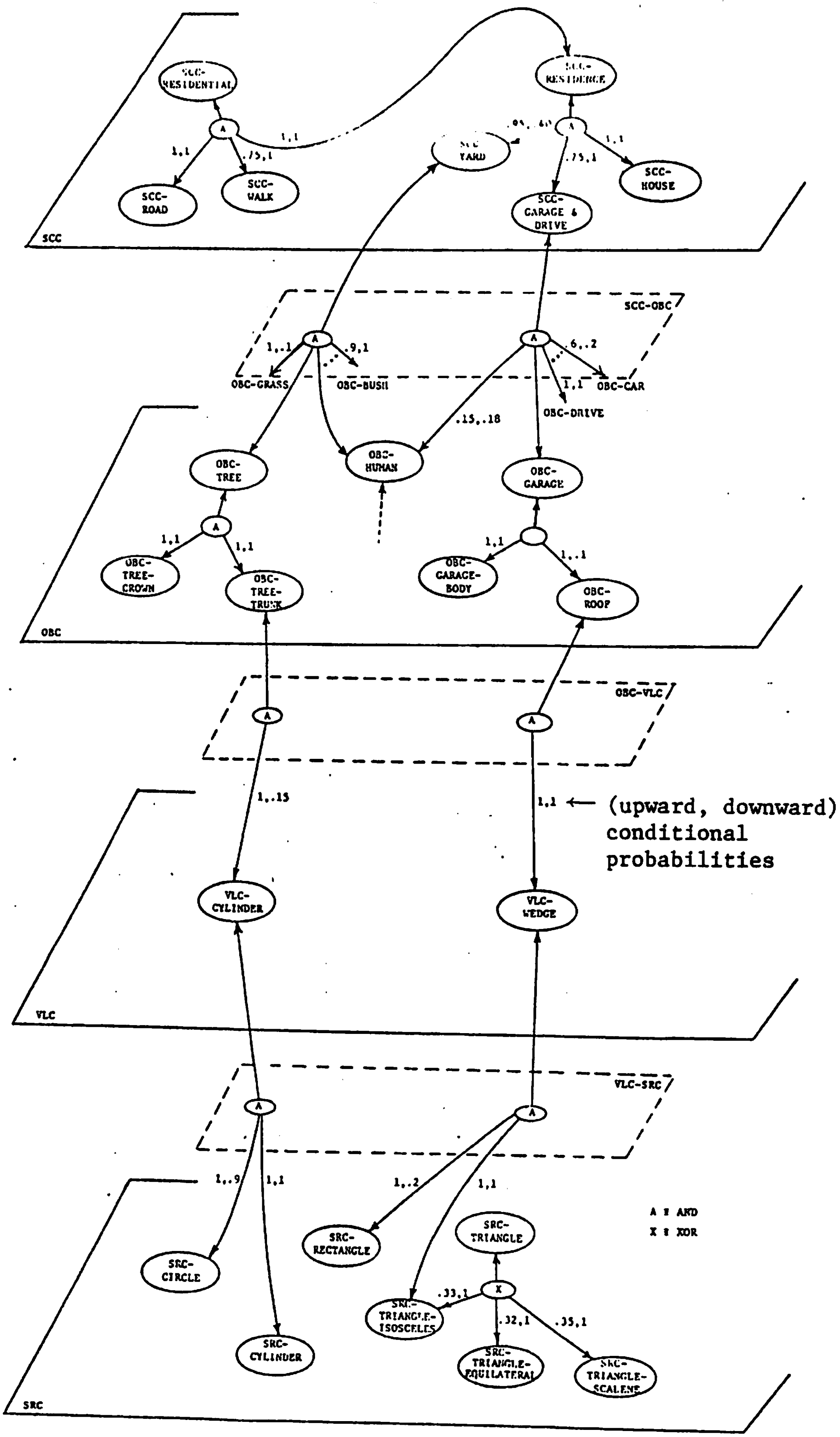


Figure 8. Hierarchical decomposition of long-term memory (LTM) and its relationship to short-term memory (STM). LTM contains the stored knowledge to which the system has access. An interpretation of an image is viewed as a set of instantiations in STM of nodes in LTM.



(upward, downward) conditional probabilities

(a)

Figure 9. (caption on next page)

Level	# Nodes	INTRA-LEVEL RELATIONSHIPS				INTER-LEVEL RELATIONSHIPS				
		# XOR	# XOR-disjuncts	# AND	# AND-conjuncts	# XOR	# XOR-disjuncts	# AND	# AND-conjuncts	Pair of Levels
SCC	10	0	0	3	8					
OBC	33	0	0	9	25	0	0	6	17	SCC-OBC
VLC	10	1	1	0	0	0	0	24	24	OBC-VLC
SRC	27	11	28	0	0	0	0	9	13	VLC-SRC
RGC	21	11	24	0	0	0	0	14	14	SRC-RGC
SGC	8	1	6	0	0	0	0	4	5	RGC-SGC
VTC	17	3	14	0	0	0	0	0	0	SGC-VTC

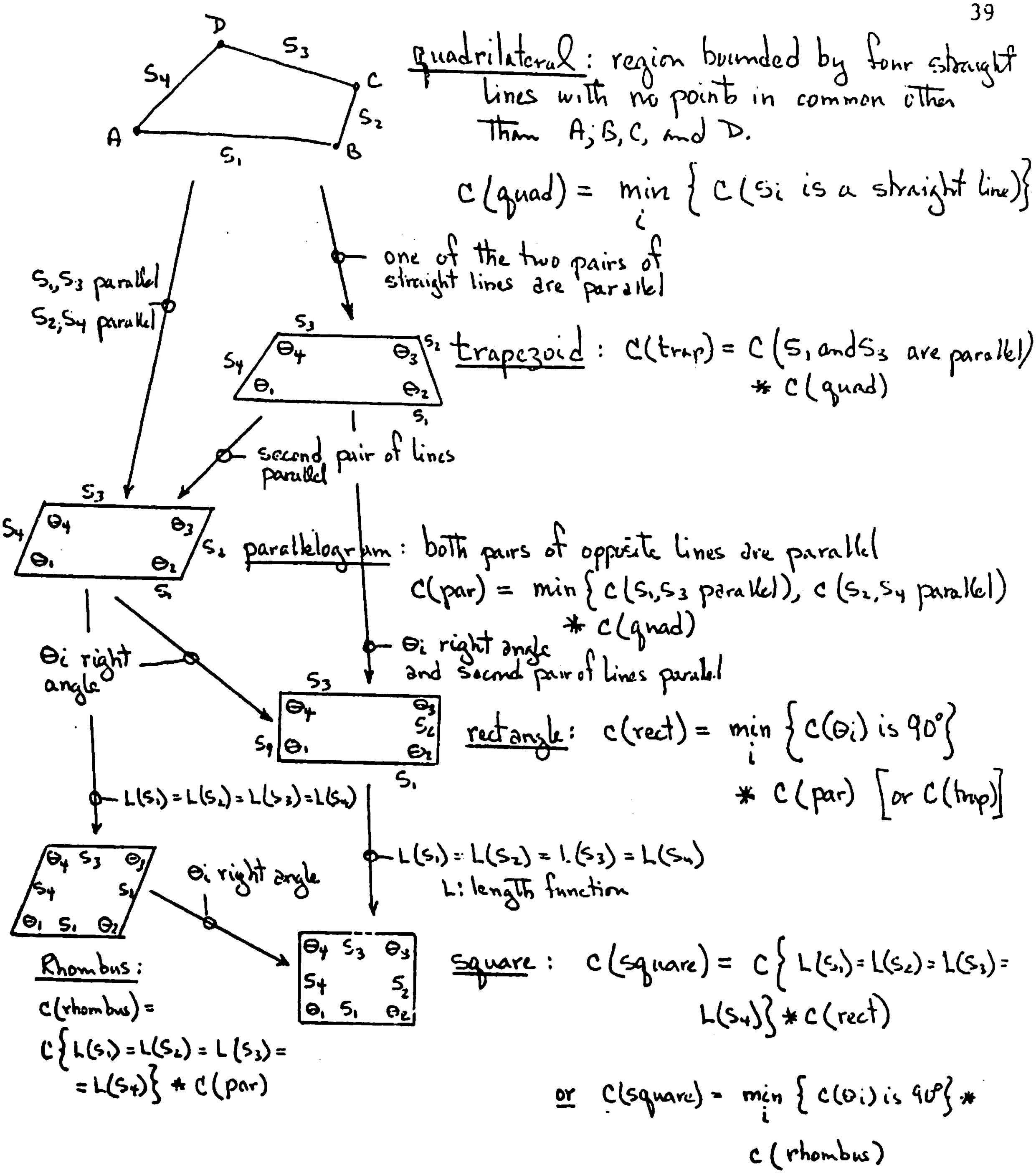
(b)

Figure 9. Detailed Structure of a Fragment of LTM. (a) A typical portion of the knowledge network. Relation nodes (AND and XOR) are circled, while names of primitives are prefixed by the level on which they reside. Between each level in LTM is a plane containing a set of inter-level relations. The pairs (x,y) on solid lines are the upward and downward condition probabilities, respectively, for use by the inference net (see Section V.6). Not shown are the priors for the nodes, attribute lists, control information (strategies) attached to nodes, etc. (b) A summary of the node and relation statistics for the entire network. For both the exclusive OR and AND relations, the first column gives the number of relation nodes (and the number of nodes decomposed according to the relation); the second column gives the number of nodes involved in these relations (i.e., the number of conjuncts and disjuncts).

V.3. 2D Shape

The 2D shape of a region may be an important cue to the identity of an object, or to attributes of a visible surface (such as the 3D orientation of the surface). Many simple relationships between the physical world and its 2D image projection are captured in LTM. For example, the 3D shape of simple volumes (e.g. cylinders and rectangular solids), as well as the 2D shapes of 3D surfaces (e.g., the rectangular surface of a window), are related to standard 2D shapes (e.g. rectangles, trapezoids, circles and ellipses). Therefore, in order to gain access to paths by which 3D hypotheses may be formed, symbolic attributes of shape, where they are relevant, must be associated with regions.

First, we outline the strategy for labelling geometric shapes formed by straight lines. Figure 10 is a portion of LTM which captures an informal definition of several shapes in terms of the straight line segments forming them. The shape classification is hierarchical; that is quadrilaterals are a superclass of both trapezoids and parallelograms, the latter being a superclass of rectangles and rhombi, etc. The definitions of shapes involve increasingly restrictive constraints as the hierarchy is descended. Therefore, if the fit for a quadrilateral is not very good, all shape types which are a subclass of quadrilateral need not be examined. In this way a large amount of computation is avoided.



D
C
A
B
S_1
S_2
S_3
S_4
S_1, S_3 parallel
S_2, S_4 parallel
one of the two pairs of straight lines are parallel
θ_1
θ_2
θ_3
θ_4
second pair of lines parallel
θ_i right angle
θ_i right angle and second pair of lines parallel
$L(S_1) = L(S_2) = L(S_3) = L(S_4)$
$L(S_1) = L(S_2) = L(S_3) = L(S_4)$
L: length function
θ_i right angle

Quadrilateral: region bounded by four straight lines with no point in common other than A, B, C, and D.

$$c(\text{quad}) = \min_i \{c(S_i \text{ is a straight line})\}$$

trapezoid: $c(\text{trap}) = c(S_1 \text{ and } S_3 \text{ are parallel}) * c(\text{quad})$

parallelogram: both pairs of opposite lines are parallel

$$c(\text{par}) = \min\{c(S_1, S_3 \text{ parallel}), c(S_2, S_4 \text{ parallel})\} * c(\text{quad})$$

rectangle: $c(\text{rect}) = \min_i \{c(\theta_i) \text{ is } 90°\} * c(\text{par})$ [or $c(\text{trap})$]

Rhombus:

$c(\text{rhombus}) =$

$c\{L(S_1) = L(S_2) = L(S_3) = = L(S_4)\} * c(\text{par})$

square: $c(\text{square}) = c\{L(S_1) = L(S_2) = L(S_3) = L(S_4)\} * c(\text{rect})$

or $c(\text{square}) = \min_i \{c(\theta_i) \text{ is } 90°\} * c(\text{rhombus})$

Note: $c(S_i \text{ straight line}) = f(\text{Rms error of } S_1 \text{ fit})$

$c(S_1 \,\&\, S_3 \text{ parallel}) = f(\text{slope of } S_1 \text{ and } S_3)$

$c(S_1 \,\&\, S_2 \text{ at right angles}) = f(\text{slope of } S_1 \text{ and } S_2)$

$c(S_1 \text{ and } S_2 \text{ of equal length}) = f(L(S_1), L(S_2))$

Figure 10. Hierarchical definition of a portion of the shape types currently in LTM. c(x) represents the heuristic confidence measure of shape type x. The classification of shapes loosely follows the classification based on affine symmetry [NEW65].

A quadrilateral requires four straight lines and the confidence that a region satisfies this condition can be heuristically specified as the minimum confidence that each of four segments is a straight line. The confidence of a straight line is the RMS error of the best fit to the actual data. Figure 10 outlines the manner in which the computation proceeds and hopefully is self-explanatory. It should be noted that the composition of confidences involves a product of confidences in an attempt to implement a worst-case analysis. One should note, finally, that heuristic functions are needed to specify the confidence of primitive attributes or relationships such as straight line, parallel line, right angle, or equal length; it is expected that any reasonable function will suffice. The result of fitting geometric shapes to segmented regions is shown in Table II.

In addition to primitive shapes formed by straight lines, quadratics are used to detect good fits of ellipses and circles to the regions [AGI72, SHI78]. Originally all types of conics (i.e., the type of curves produced by cutting a right circular cone with a plane, including ellipse , hyperbola , parabola , etc.) were fit, but this has been replaced by spline fits.

Most regions in our outdoor scenes are not classified as any of the simple shapes mentioned and are labelled symbolically as 'blob'. Nonetheless, important information such as the parametric fit of the 2D spline analysis is carried forward for later 3D processing.

Region	Shape	Probability	Aspect Ratio
RG-0047	Rectangle	.937	6.33
	Trapezoid	.96	——
RG-0050	Rectangle	.99	6.33
RG-0051	Rectangle	.99	6.33
RG-0054	Rectangle	.99	6.00
RG-0060	Rectangle	.80	7.5
	Trapezoid	.85	——
RG-0045	Rectangle	.80	6.25
	Trapezoid	.85	——
RG-0049	Rectangle	.85	10.33
	Trapezoid	.90	——
RG-0086	Rectangle	.99	3.00

Table II. Summary of 2D shape fits to selected regions of Figure 3.

V.4. Occlusion

Researchers in image processing have long recognized the importance of picture junctions as loci of surface information. When objects in scenes are limited to planar surfaces forming trihedral vertices, the analysis of picture junctions can be efficiently exploited. The constraint of surface planarity ensures that only straight lines will appear in the image and the trihedral constraint guarantees that there will be a small number of fairly well-understood vertex types [HUF71, CLO71, WAL75, TUR74]. When scenes contain complex curved objects, the problem becomes more difficult.

The cubic spline fits to the image provide useful occlusion information at picture junctions. Placement of knot(s) at the junction ensures that two line segments, meeting at a junction, which are part of a continuous line, will be smoothly fit by the splines. This is a generalization of the "tee" junction in the polyhedral domain, but does not require assumptions about straight lines. A simple strategy for determining the degree of discontinuity (e.g., relative angle) between pairs of line segments approaching the junction yields occlusion hypotheses at the junctions. York [YOR80] is currently examining the improvements obtained by a spline approach vs. piecewise linear fitting on Turner's classification of the 2D junctions formed by the meeting of 3D curved surfaces [TUR74].

V. 5. Object Hypothesis via Spectral Attributes

For a restricted class of objects occurring in outdoor scenes, attributes of color and texture can be expected to remain relatively invariant across a wide range of scenes. The spectral attribute KS matches region attributes to stored attributes of several objects (sky, tree, bush, grass, and road) and returns a measure of the degree of match, ranging from -100 (no match) to +100 (excellent match). The stored attributes were obtained by measuring 60 features across samples of each object extracted from a data base of 25 images. A piecewise linear decision function which reflects the expected variability of each feature of an object is then formed. The matching process extracts an identical set of features from the region (or union of regions) to be identified, and uses the decision function to generate a degree of match for each object. This research is part of the Ph.D dissertation of T. Williams; more detail appears in [HAN78b, WIL80].

The attribute matcher can only be used to hypothesize the presence of certain "target" objects based upon the expected invariance of their color and texture attributes. There are many objects such as cars, shirts, and most other man-made objects which vary in their spectral characteristics. This KS, however, will return a confidence value for any region, regardless of whether the region represents a target object or not. Therefore, we require mechanisms for filtering these hypotheses.

Figure 11 illustrates the results obtained by applying the spectral attribute matcher KS to the 21 largest regions of our example. Of the 21 regions, 14 were target regions (3, 8, 10, 30, 20, 79, 15, 37, 82, 96, 90, 83, 110, and 93) and 7 were non-target regions (14, 58, 41, 56, 35, 70, and 21). Of the 14 targets, 8 were correctly identified on the basis of a maximum confidence decision. If bush and tree are collapsed into a single object (which is not unreasonable given the similarity of spectral attributes), then 11 of the 14 are correctly identified. Of the remaining three target errors, the correct hypotheses had the second highest confidence in two cases (regions 15 and 96); region 8 represents a mixture of sky and small tree limbs and the correct hypothesis is debatable.

Of the 7 non-target regions, 5 of the regions (58, 41, 56, 35, and 70) represent portions of the white house wall and all were hypothesized as sky. In the absence of any additional information, such hypotheses are reasonable and cannot be eliminated. The remaining two regions are both roof (regions 14 and 21) and both were hypothesized as grass, probably due to similarity of values for several crude texture measures. Both of these hypotheses, and three of the previous five, can be filtered if the location of the horizon is known and the ground is assumed to be flat. Regions 14 and 21 cannot be grass and be located above the horizon, while regions 58, 56, and 70 are either below or straddle the horizon and hence cannot be sky. This will be discussed again in the "horizon" KS (Section V.9).

Region	Area		Confidence Measure (maximum confidences circled)					Hypothesized Identity (max. conf.)	Actual Identity (visual)	Correct Hypothesis?	Correct Hypothesis? (Bush/Tree one object)	Comments	Correct Hypotheses Filtered by Horizon KS
	# pixels	% of picture	Bush	Grass	Road	Sky	Tree						
14	3101	18.9	-10	32	-55	-17	-16	Grass	House Roof	No	No	above assumed horizon	no hyp.
3	1939	11.8	-62	-48	41	74	-84	Sky	Sky	Yes	Yes		Yes
8	971	5.9	-20	- 6	31	47	-46	Sky	Tree	?	?	mixture: tree without leaves and sky	?
10	793	5.9	-33	-40	-88	-53	46	Tree	Tree	Yes	Yes		Yes
58	606	3.7	-52	-41	-49	20	-21	Sky	House Wall	No	No	white house wall in sunlight; region straddles assumed horizon	no hyp.
41	560	3.4	- 8	-23	-70	-38	76	Tree	House Wall	No	No	white house wall eaves, & gutter in shadow matches tree onbrightness,texture	No
30	518	3.2	-54	-41	-53	-41	14	Tree	Tree	Yes	Yes		Yes
56	486	3.0	-60	-51	-62	23	-23	Sky	House Wall	No	No	white house wall; region straddles assumed horizon	no hyp.
20	427	2.6	10	37	-88	-50	54	Tree	Tree	Yes	Yes		Yes
35	410	2.5	- 2	13	-61	32	-16	Sky	House Wall	No	No	white house wall with shadow of tree	No
79	373	2.3	45	0	-92	-71	46	Tree	Bush	No	Yes	confidence for bush is almost as large as tree	Yes
70	354	2.2	- 2	7	-61	29	-25	Sky	House Wall	No	No	white house wall; part of region above horizon	?
15	330	2.2	-27	20	-56	-32	2	Grass	Tree	No	No	region above horizon; tree next most	Yes
21	310	1.9	- 8	40	31	-35	-16	Grass	Roof	No	No	region above horizon; likely knocks out road also	no hyp.
37	308	1.9	-33	-30	-92	-53	44	Tree	Tree	Yes	Yes		Yes
82	238	1.5	2	0	-77	-53	- 2	Bush	Bush	Yes	Yes		Yes
96	217	1.3	4	6	-53	-53	-10	Grass	Bush	No	No	bush next most likely	No
90	202	1.2	29	9	-85	-71	0	Bush	Bush	Yes	Yes		Yes
83	198	1.2	39	0	-94	-71	44	Tree	Bush	No	Yes	bush next most likely	Yes
110	196	1.2	-48	32	25	-35	-46	Grass	Grass	Yes	Yes		Yes
93	196	1.2	29	-20	-89	-74	35	Tree	Bush	No	Yes	bush next most likely	Yes

Figure 11. Results obtained by applying the spectral attribute matcher KS to the 21 largest regions (ordered by decreasing size) of Figure 3. The results obtained by filtering the hypotheses by the horizon KS (see Section V.11) are also shown; if there are no positive confidences after filtering, no hypothesis is generated.

The statistics on the remaining 93 regions are approximately the same, although if the size of the region falls below a minimum size, reliable texture measures cannot be extracted and performance falls off. A number of the regions have negative confidence values for all target objects and no hypotheses are generated for these regions.

V.6. The Inference Net in Long Term Memory

The representation of declarative information is a layered, hierarchical graph structure in which nodes represent visual entities and arcs represent the relationships between these entities. By associating probabilities with nodes and conditional probabilities with arcs, an "inference network" [DUD76, KON78] is defined. The arcs and probabilities define weighted paths by which implications of local hypotheses may be propagated upward and downward through the layered network. Any hypothesis generated by any knowledge source which results in a change in the a-priori probability of a node can then be used to generate changes in likelihoods at other nodes via these paths. Moreover, entire partial interpretations may be used to generate hypotheses about likely identities of unexplained portions of the image. The presence of a window and roof, for example, would strongly imply the presence of a house and consequently the house scene schema.

The inference net of the Prospector system [DUD78], as originally formulated, is designed to propagate information in one direction only, from low-level "evidence" nodes towards high level "goal" nodes. The method by which information is propagated is developed from a Bayesian probability formulation of the joint occurrence of the visual entities in the long term memory network. Prospector only employs conditional probability distributions between pairs of nodes (i.e., governed by joint probability distributions of two nodes at a time). In some situations, however, it is desirable (or necessary) to define joint and conditional distributions across n nodes in order to capture higher level dependencies. In any case there are serious theoretical issues inherent in the use of inference nets, such as consistency or loops of inferences which relates to convergence problems in relaxation labelling. These will not be discussed here, but related issues are discussed in [HAN80a, LOW80].

Table III is a summary of the way apriori probabilities of nodes higher in the network change as a result of updating the likelihoods of lower nodes as shown and then propagating upward.

V.7. 3D Shape Representation

There are several important issues involved in the specification of the 3D shape of an object. The more important of these include the choice and representation of the shape primitives, the choice of a coordinate system within which the relationships between primitives can be described, and the ease with which features useful for recognition and/or matching can be extracted [MAR77, AGI72, AGI76, NEV76, NEV77, BAD79]. These

Regions → Volumes		Volumes → Objects		Objects → Schemas	
Instantiations:RGC	prob.	Instantiations: VLC	prob.	Instantiations: OBC	prob
rgc-rectangle R	1.0	vlc-wedge A	.8	obc-bush B	.7
rgc-trapezoid T	1.0	vlc-rect.-solid B	1.0	obc-tree T	.7
				obc-grass G	.7

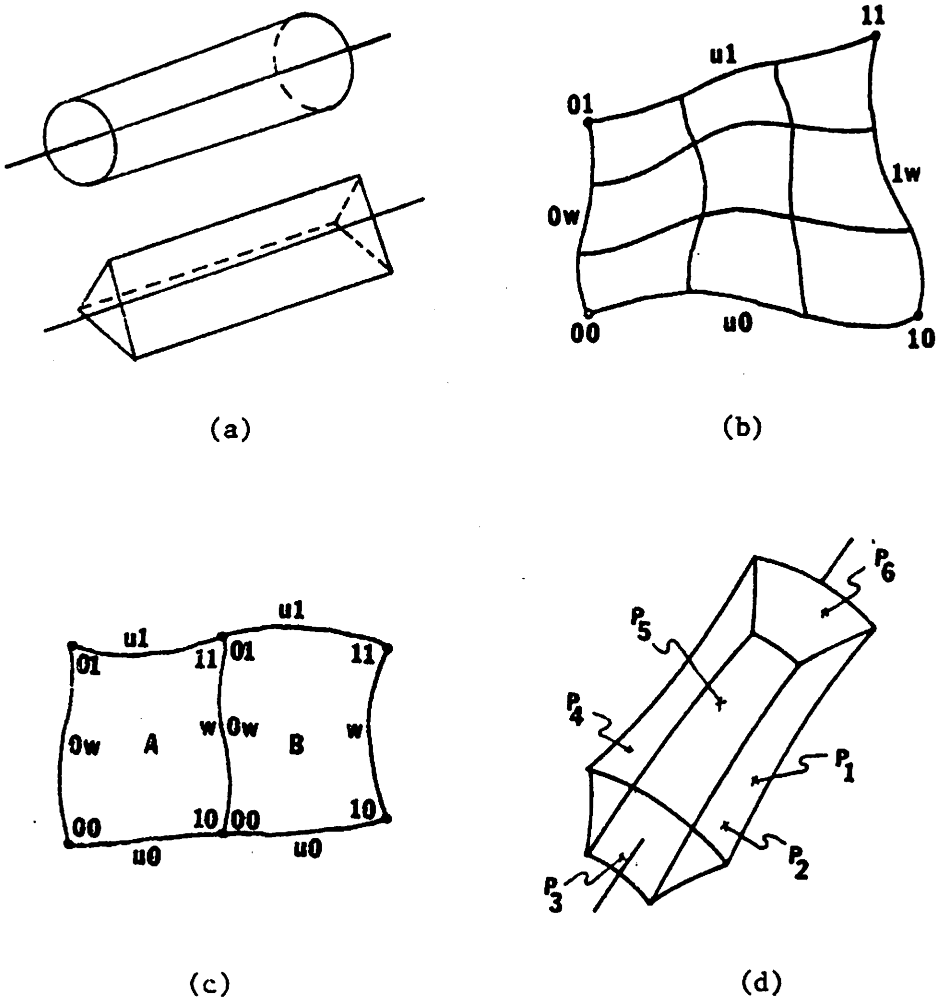
At Volume Level		At Object Level		At Schema Level	
Prior	A-Posteriori	Prior	A-Posteriori	Prior	A-Posteriori
.2	VLC-RECTANGULAR-SOLID (.72508E0 R&T) (.47197E0 T) (.44659E0 R)	.033	OBC-ROOF (.76988E0 A)	.235	SCC-YARD (.98695E0 B G T) (.96297E0 B T) (.89479E0 B G) (.88846E0 G T) (.74512E0 B) (.73249E0 T) (.47243E0 G)
.18	VLC-CYLINDER (.45478E0 R)	.067	OBC-HOUSE (.66089E0 A B) (.47309E0 A) (.13485E0 B)	.234	SCC-RESIDENCE (.50858E0 B G T) (.47189E0 B T) (.40743E0 B G) (.37584E0 G T) (.34480E0 B) (.34206E0 T) (.28574E0 G)
	VLC-CUBE (.37957E0 T)	.017	OBC-GARAGE (.61311E0 A B) (.30874E0 A) (.56834E-1 B)	.100	SCC-RESIDMNTIAL (.42231E0 B G T) (.37918E0 B T) (.30340E0 B G) (.26627E0 G T) (.22978E0 B) (.22656E0 T) (.16035E0 G)
.08	VLC-TRAPEZOIDAL-SOLID (.32124E0 R&T) (.29419E0 T) (.89864E-1 R)	.10	OBC-GARAGE-BODY (.13673E0 B)	.234	SCC-HOUSE (.34487E0 B)
.16	VLC-WEDGE (.20157E0 R&T) (.17997E0 T) (.17972E0 R)	.04	OBC-HOUSE-BODY (.11000E0 B)		
		.032	OBC-BUILDING-SIDE (.10784E0 B)		
		.002	OBC-BUILDING-SHUTTER (.10000E0 B)		
		.005	OBC-BUILDING-WINDOW (.10000E0 B)		
		.026	OBC-BUILDING-WALL (.90000E-1 B)		
		.002	OBC-BUILDING-DOOR (.80000E-1 B)		
		.020	OBC-GARAGE-FRONT-SIDE (.60000E-1 B)		
		.020	OBC-GARAGE-DOOR (.60000E-1 B)		
		.008	OBC-BUILDING-FLOOR (.60000E-1 B)		

Table III. Sample results from the inference net KS. The results shown are inferences upward from one level to the next, assuming the instantiations and associated probabilities as shown. The instantiation(s) represent evidence via some KS for updating the probability of some node(s). The prior and posterior probabilities of nodes higher in the network are shown. The effect of propagating different pieces of evidence from below are labelled with letters after the probability. Actual instantiation of hypotheses on the image of Figure 3 are given in Section VI.8.

issues are being investigated by York in his Ph.D. thesis [YOR80] by applying and further developing techniques from the computer-aided design community [COO67, COO74, GOR74].

The most popular 3D shape representation -- generalized cylinders [NEV77] -- involves formation of a 3D volume developed by sweeping a given planar cross section down an axis (Figure 12a). Thus, an object centered coordinate system is employed and an assembly of subparts is described by relating the local axes to each other [MAR77].

Our efforts are directed at making the relationships between subparts accessible, the relationship of surfaces to volumes more explicit, and the development of a representation for arbitrary curvature of surfaces. The representation (Figure 12b) employs Coons surface patches, whose four sides are delimited by cubic splines [COO74]. The surface patch (Figure 12c) is formed by using an interpolation, or "blending" function, from the pair of opposite sides of the surface patch. The blending function itself is also a cubic spline; it allows a smooth transition between adjacent patches, both those defining a single volume, as well as adjacent volumes, as in a car fender and car body. A "quilted solid" is defined by six surface patches related to a volume-centered coordinate system (Figure 12d). Figure 12(e)-(g) depict surface patches from several different points of view. Many kinds of information can be stored with or derived from a quilted solid (Figure 12h).



(a)

(b)

(c)

(d)

Figure 12. The three-dimensional representation of shape. Much of the representation is based upon cubic splines, Section V.1. (a) The generalized cylinder representation. (b) A Coon's surface patch P(u,w), where u and w are parameterized on the interval [0,1], employs four B-splines P(0,w), P(1,w), P(u,0), P(u,1) to delimit the surface patch boundary; blending functions which are also B-splines interpolate between opposite sides of the surface patch. (c) Two adjacent surface patches A and B can be smoothly joined at a common boundary if the blending functions are constrained properly. (d) Six surface patches can define the shape of a volume around an axis which is used to relate the spatial orientations of such volumes.

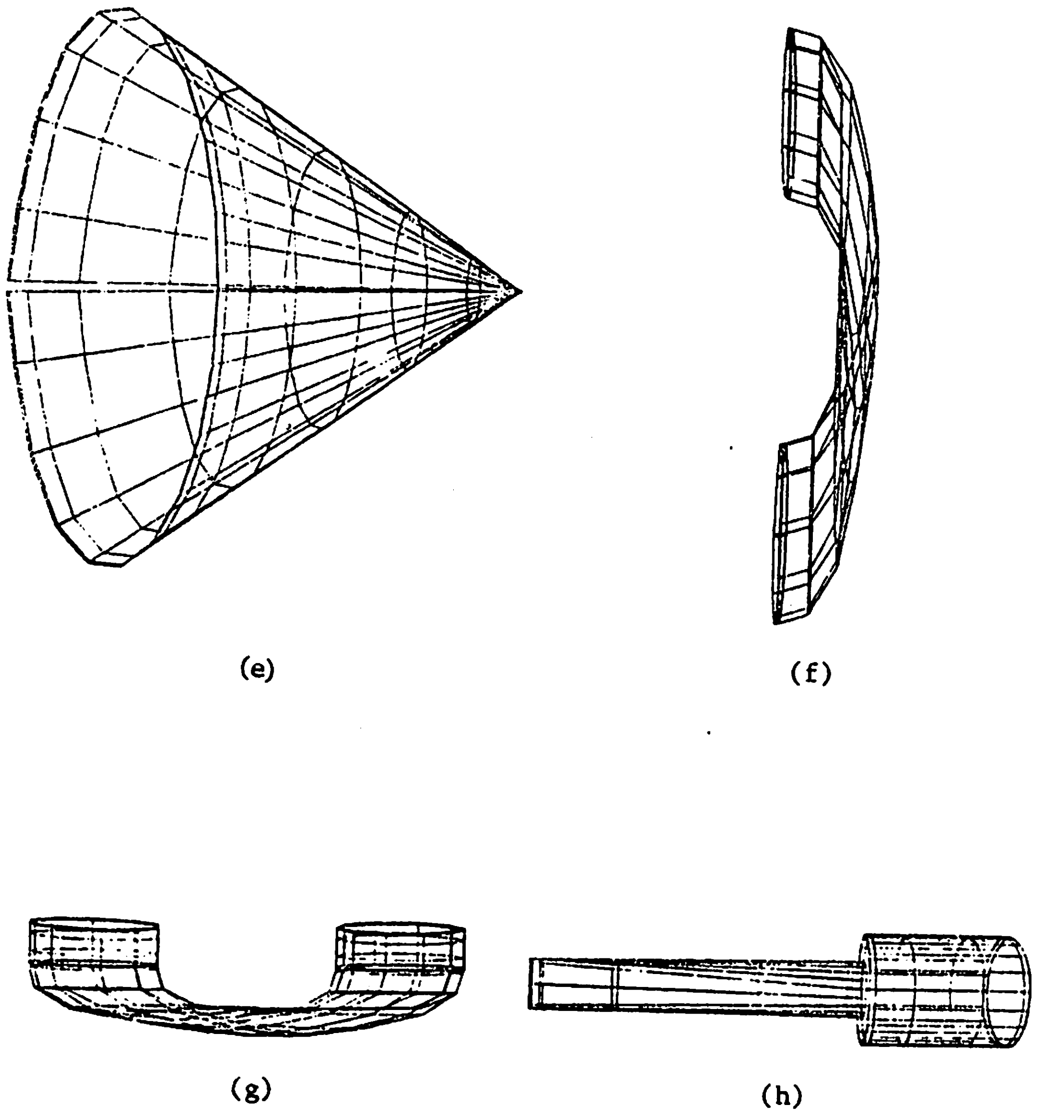


Figure 12. (e) A single surface patch can also be used to fold around upon itself to produce a volume. In this figure, one boundary reduces to a point. (f) Telephone handle using one surface patch. (g) Telephone handle using three surface patches showing smooth join between patches. (h) Screwdriver formed from two patches.

A system for separately defining arbitrary surface patches, combining patches into volumes, combining volumes into objects, building specific 3D schemas, and rotating schemas subject to the assumption of a given point of view is partially developed. However, these components have not yet been integrated into our system.

V. 8. 3D Schemas

There is a great deal of expected structure in our visual environment and it seems evident that such expectations are important in processing visual information. One of the functions of the 3D schema is the organization of subsets of information in LTM into higher order complexes of stereotypical situations in such a way that the spatial relationships between objects, volumes, and surfaces which might occupy or define that space are made explicit. The 3D schema would allow rotation and translation of the prototypical scene so that its appearance from any point of view can be generated. Thus, the processing of a 3D schema allows the generation of potentially relevant 2D schemas.

The results given in Section VI demonstrate top-down interpretation of an image. In order to do this it was assumed that a specific 3D schema was available, that it could be rotated given an assumed point of view, projected onto a 2D image plane, and then hidden lines removed. While those 3D facilities were not available then, and 2D schema information was supplied directly, they are now available.

Our current version of specific 3D and 2D schema have for each schema region a centroid of the expected central location and a radius representing the decreasing likelihood that the schema region appears at that location. Thus, one can think of a spherical or circular probability cloud denoting expected spatial position. This crude representation of location allows selection of regions in the image for matching against schema objects; furthermore, alternative region selections can be ordered by degree of location match. Figure 13 depicts wire frame and surface representations of a model of the house image. The 3D schema we have described attempts to capture approximate relative spatial information of the entities appearing in Figure 13. There are still interesting problems remaining that are associated with the generation of 2D schema from 3D schema. For example, the likelihood that a 2D schema region is visible will be related to the likelihood that another schema region will occlude it. Many issues related to the generation of specific 2D schema from specific 3D schema are under examination.

V.9. Perspective

The perspective knowledge source concentrates on the ways in which the general relationships governing perspective transformations can be used to extract or explain information concerning surface orientation, distance, and size [DUD73, HAR78]. A region (or the union of a group of regions) represents the projection of a 3D surface onto the 2D viewing plane. The

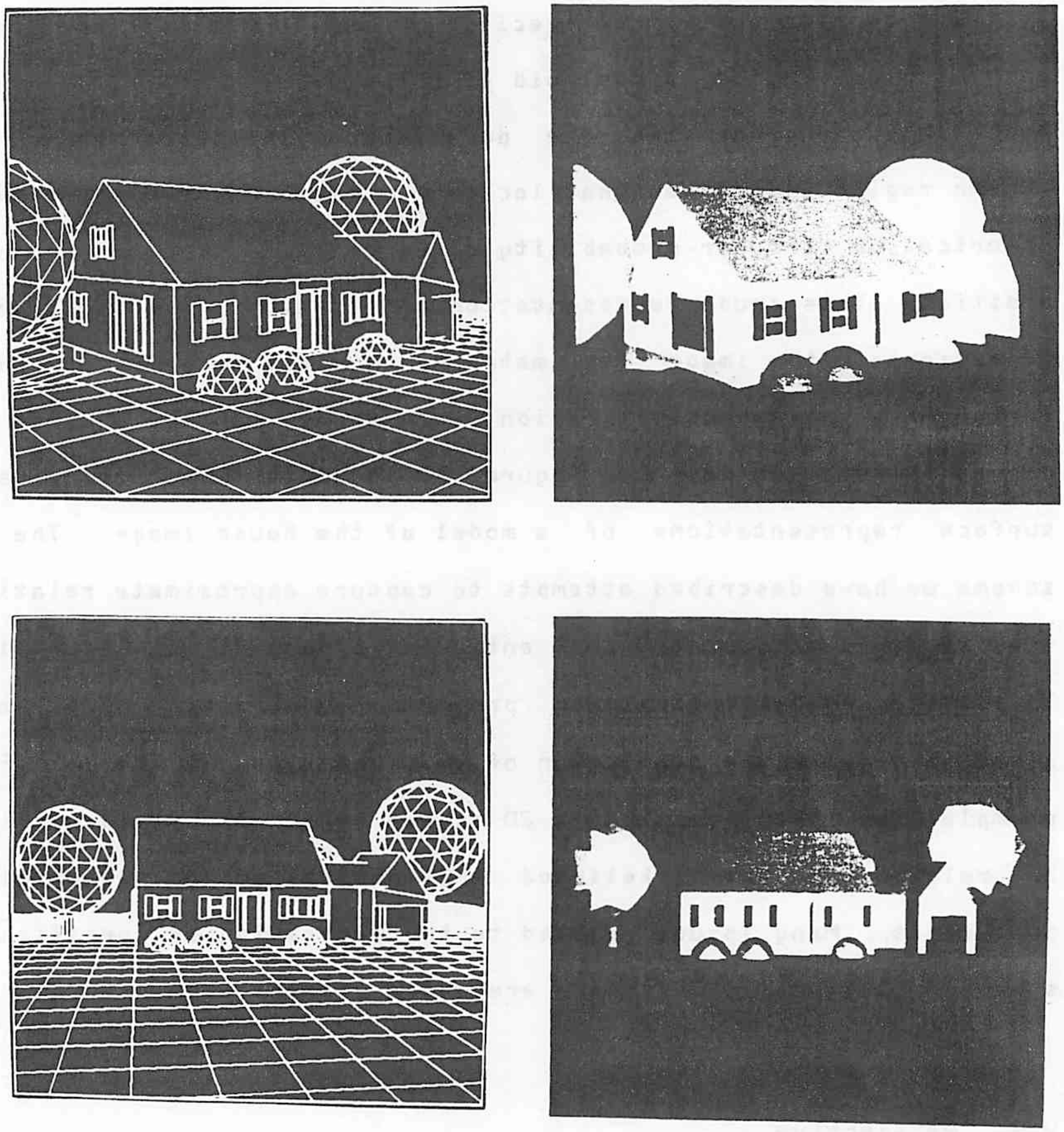


Figure 13. Wire frame and surface representations of a model of the house image seen from two points of view. The current 3D house scene schema is actually an abstract representation of the approximate relative spatial locations of the entities in these images. The components (volumes, surfaces, straight line segments) are actually represented by a position in space and a radius associated with a decreasing likelihood of the component appearing at that location.

problem then is to recover some of the 3D attributes of that surface from the segmented image. Figure 14 is a simple sketch depicting the relationship of the distance and height in the physical world and their associated parameters in the image.

The current version of the perspective KS focusses on the relationship between the following variables:

a) elevation - vertical distance above the ground plane,

b) height - vertical distance from visible bottom edge to visible top edge of surface,

c) range - horizontal distance from viewing location to a distinguished point on the surface,

and d) width - horizontal distance from the visible left edge to the visible right edge of the surface.

The interrelationship of these variables depends on the orientation of the surface in three-space. For simplicity, we assume the orientation is either vertical (i.e., perpendicular to the ground plane, such as a tree) or horizontal (i.e., in the ground plane, such as a road). While these assumptions may appear to be unnecessarily restrictive, they are sufficient to cover many surfaces of interest.

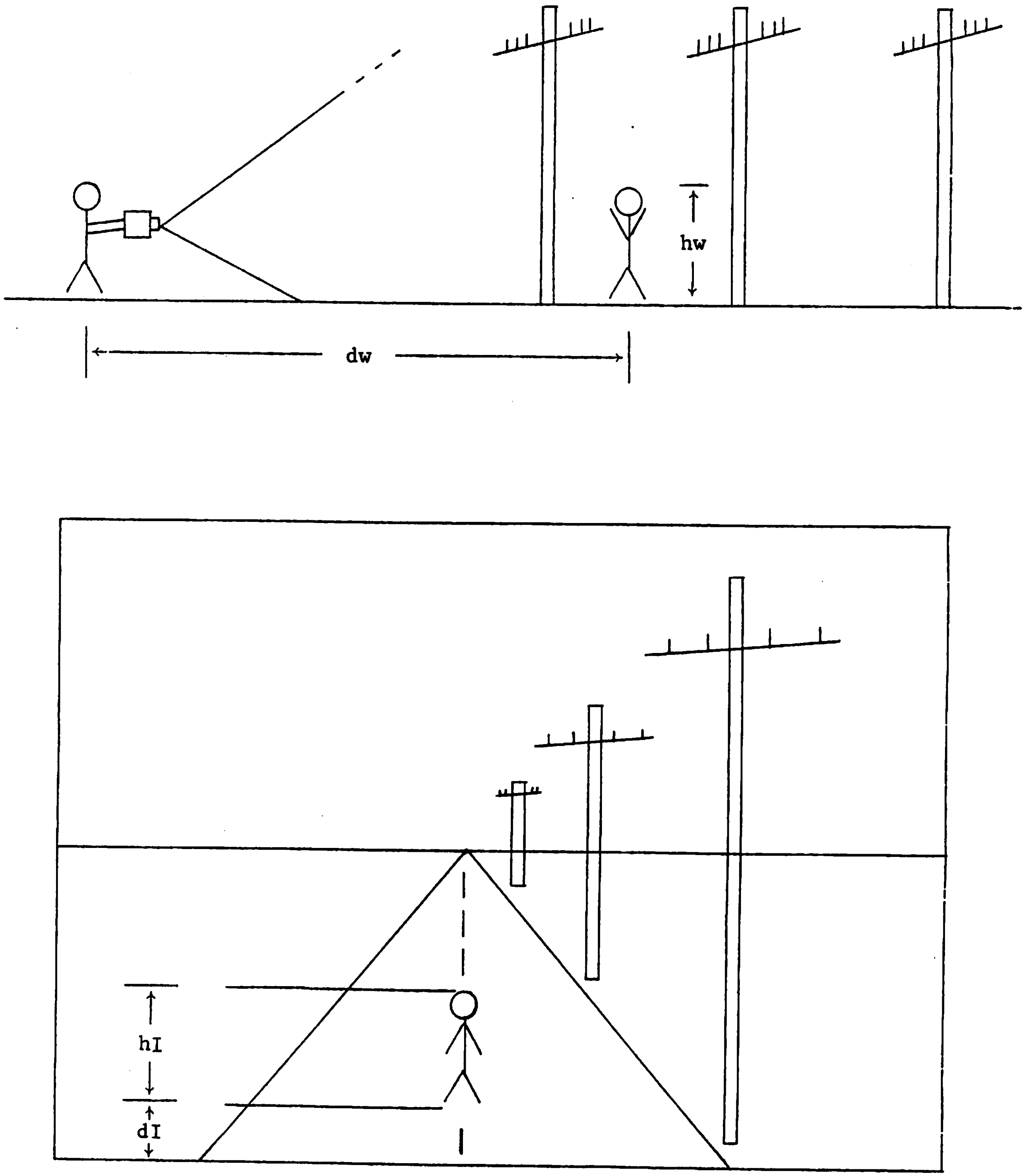


Figure 14. Perspective - ground plane, vanishing points, projective geometry.

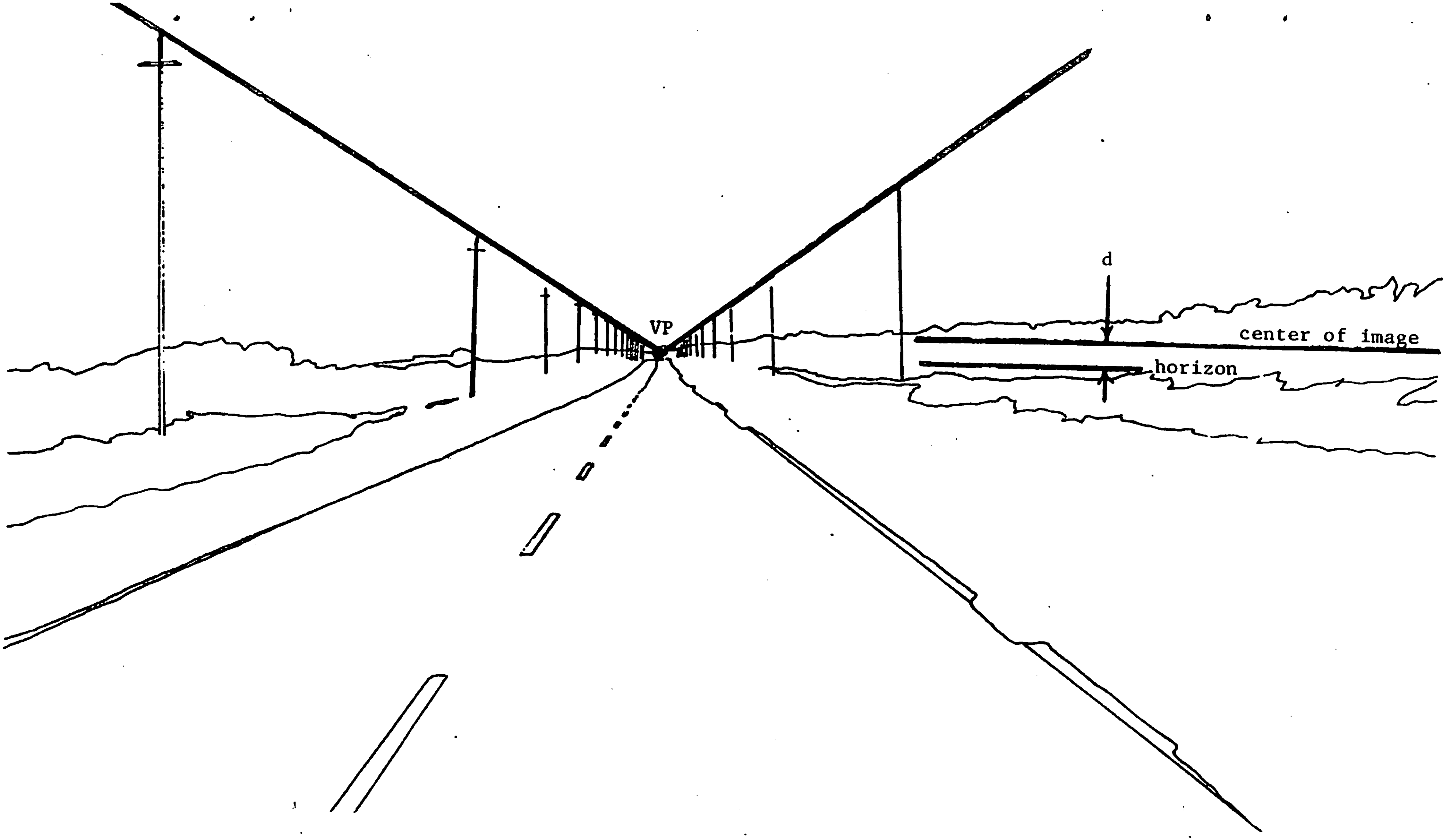
The four variables described above are interrelated. Given the assumption of ground planarity and a camera model (angle of inclination to ground plane, focal length, and height above ground plane), knowledge of any one implies knowledge of the remaining three, although the form of the relationship depends on whether the orientation of the surface is assumed vertical or horizontal. We are continuing to explore ways to use perspective under weaker assumptions in our current research.

In general, there are usually several unknown quantities to be determined and depending on the assumptions made one can solve for different variables. Applying the perspective KS to selected regions of Figure 3, it is easily determined that the range of region 79, for example, is about 37 meters and its height is 1.61 meters; this required assumptions of ground planarity, and that the surface projected as region 79 is perpendicular and attached (i.e., zero elevation) to the ground plane. More extensive results from the perspective KS, and the use of these results for the development of a 3D spatial plan are presented in Section VI.

V.10. Futher Development of the Perspective KS

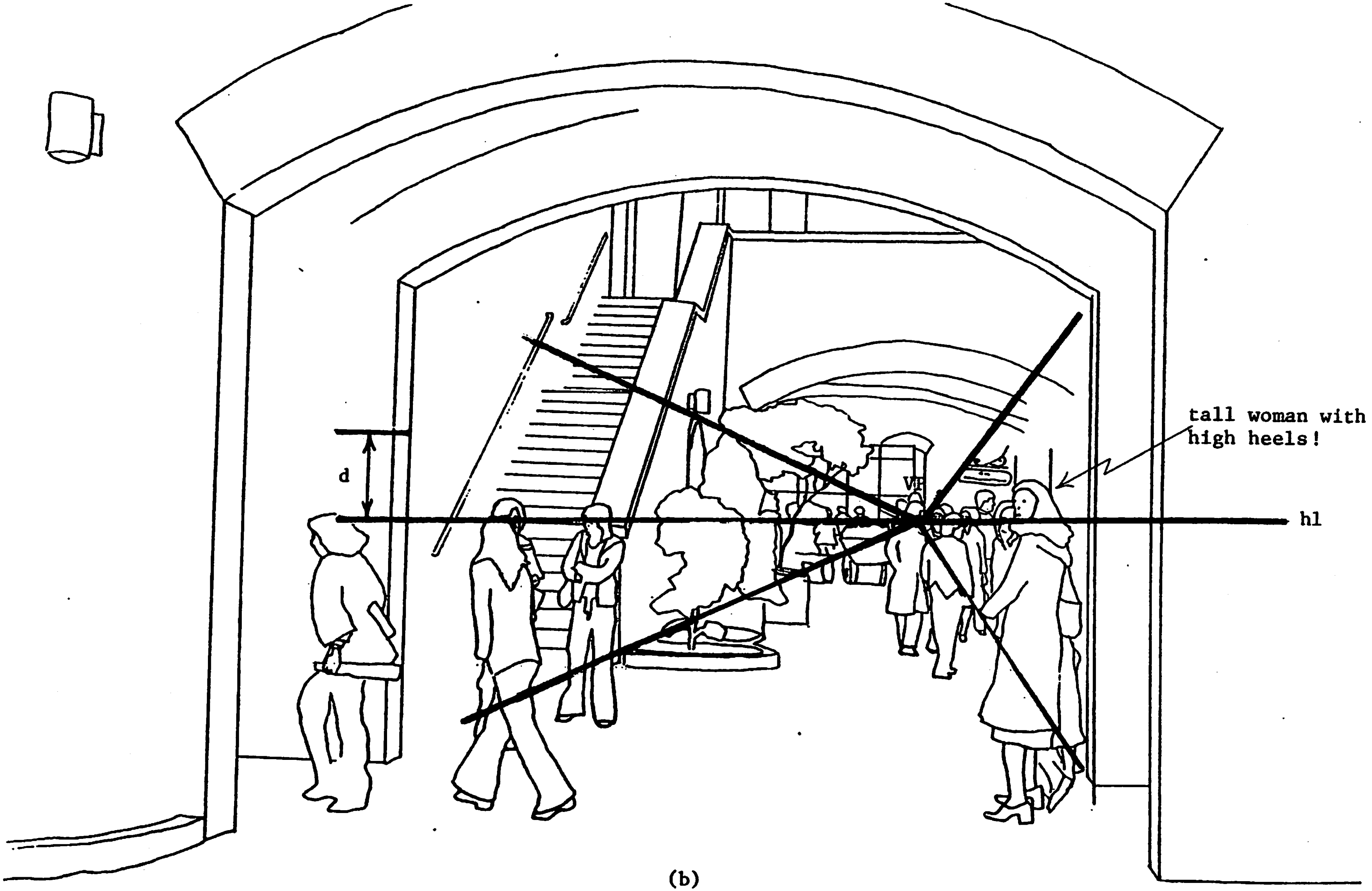
In Figure 15 there are several sources of information in the images that relate to the 2D projection of 3D volumes and surfaces. Figure 15(a) shows a series of identical objects diminishing in size. If it is possible to generate the hypothesis that the objects are of identical size and orientation -- a situation that is not uncommon in various geometrically regular aspects of our man-made world -- then the tops and bottoms of the telephone poles provide lines of convergence to vanishing points on the horizon line. The diminishing size of the telephone poles is a particular example of a feature gradient, known to be important in the perception of space [GIB50].

The use of the perspective equations for size and distance demands knowledge of the tilt angle of the camera relative to the general plane. This information is provided by the position of the horizon in the image when the ground is planar. Figure 15(b) depicts an example where the horizon line can be inferred when in fact it is not visible in the image nor are there convergent lines which could be reliably used. In the physical environment corresponding to Figure 15(b), the plaza provides a flat surface which is defined in the image by the bottom of the feet of the figures. The horizon line lies in this plane. If the relative angle of the camera to an infinitely planar ground surface is 0, then the horizon is in the center of the image, and in general tilt is directly computed from the distance of the horizon to the center of the image. The height and distance of the various figures may be determined directly from the distance of their feet from the bottom of the 2D image. Yet a third plane is roughly described by a least rms error fit of the points corresponding to the eyes of the figures; if the camera is at a similar height to the eyes (not an unreasonable assumption), then this plane is constrained to go through the horizon as well.



(a)

Figure 15. Examples of images in which perspective information provides strong cues to the spatial arrangement of surfaces and objects. (a) If the telephone poles are recognized to be objects of identical sizc that all lie in a plane, then the tops and bottoms of the poles provide lines of convergence to vanishing points on the horizon line. The converging lines of the road provide similar information. The image distance of the horizon from the center of the image provides relative orientation of camera to ground plane. The camera is computed to be tilted .6 degrees upward.



(b)

Figure 15(b). In the plaza scene the horizon line can be inferred when in fact it is not visible. The bottoms of the feet provide points which lie in the ground plane. The line producing minimum RMS error from the heads of the people gives an approximation of the horizon line because the camera is approximately at the same height. The arches if recognized as such could also provide further convergent cues concerning the horizon. The tilt angle of the camera has been computed to be 2.4 degrees upward.

This implies that the eyes of the figures, or more roughly the tops of their heads, must lie at approximately the same height in the image, as is evident in Figure 15b.

There are many other interesting situations which deserve investigation, such as:

a) deriving the orientation for planar surfaces that are at some general orientation, not horizontal or perpendicular to the ground plane;

b) assumptions concerning lines which are near parallel or perpendicular and their implications about the physical world;

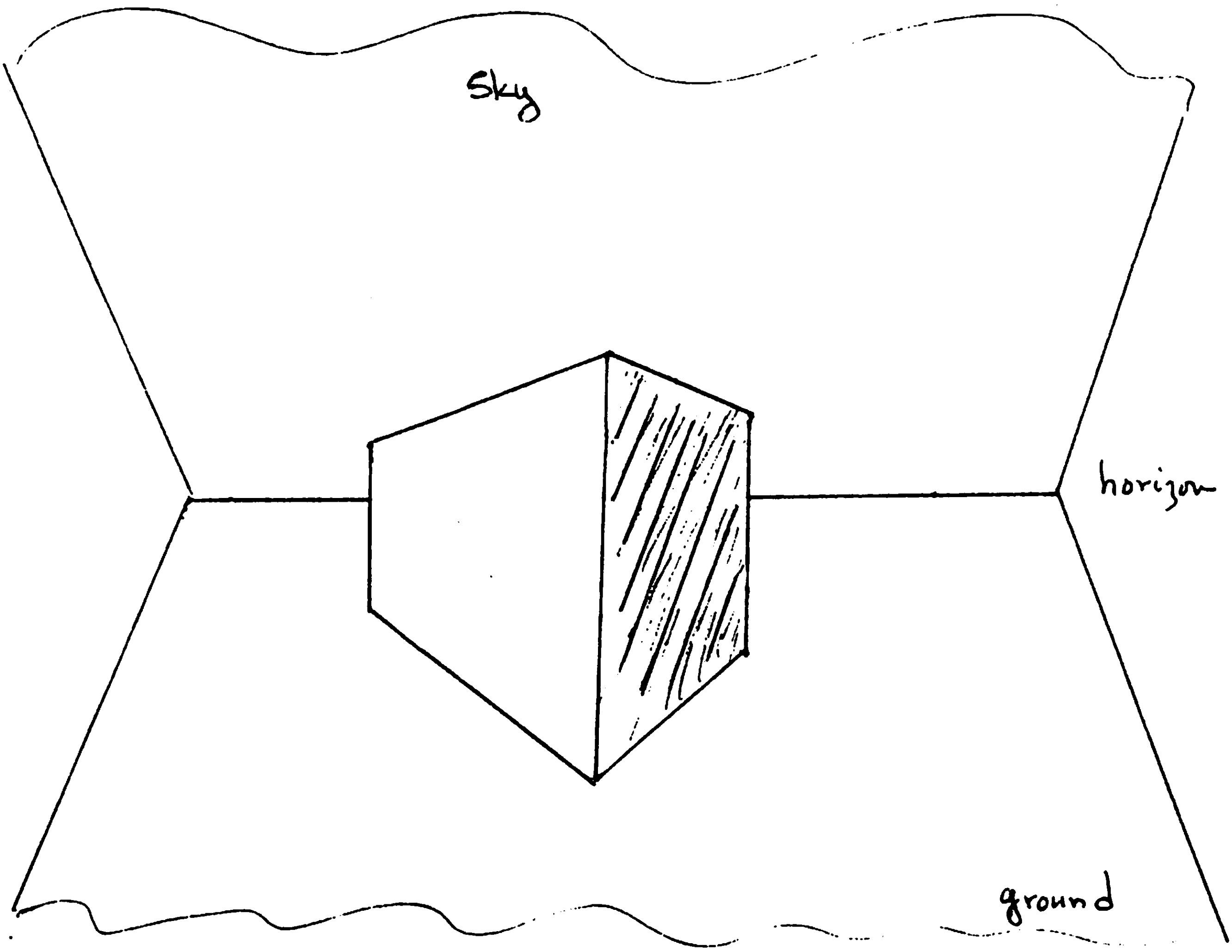
c) deriving distance to objects and camera tilt angle from assumed or known physical sizes of the objects corresponding to regions in the image.

Continuing research on the perspective KS will focus on the information required for the construction of a spatial plan of the 3D scene, the development of a collection of mini-strategies for using this information, the determination of the conditions under which these strategies may be activated, and on methods for extracting this information from the image data.

V. 11. Horizon Schema and Horizon Filter KS

It should be clear that the effects of perspective and distance on the projection of surfaces in the image are determined by the observers position, the camera model (of height, pan, tilt, and focal length), and the orientation of the ground plane. These factors also determine the position of the horizon in the image, if it is visible. The horizon schema is perhaps the simplest and most general of the schemas present in the system. The function of the horizon schema is to define the relationship between sky, ground, and horizon , and to provide the global coordinate system for placing objects and schemas in space (Figure 16).

The horizon schema also provides the basis for a filtering KS applied to the hypotheses generated by other knowledge sources. Since the spectral attribute KS, for example, has no notion of the spatial location of its target objects, some of its hypotheses may be inconsistent with the location of the horizon in the image. By collapsing the more obvious spatial constraints into a knowledge source associated with the horizon schema, many erroneous hypotheses can be eliminated. For example, in Section V.5, Figure 11, region 58 was hypothesized to be sky. While this is a reasonable hypothesis based solely on spectral attributes (white walls tend to "inherit" the color characteristics of the ambient illumination or reflected illuminant characteristics from nearby objects), "sky" regions cannot exist below the horizon and the sky hypothesis can be eliminated. Since no other reasonable hypothesis exists, no hypothesis for this region can be generated by the spectral attribute matcher. For region 15, the hypothesis "grass" is eliminated since the region is above the horizon; the



<u>Figure 16</u>. Illustrative diagram of the Horizon KS and its use as a hypothesis filter. Hypotheses which violate the spatial constraints imposed by the horizon can be eliminated. Regions which extend below the horizon cannot be labelled sky, while regions which extend above the horizon cannot be in the ground plane (e.g., road, grass). Therefore the cross-hatched region cannot be labelled sky, grass, or road.

next most likely hypothesis (tree) cannot be eliminated and becomes the final hypothesis.

The results from the horizon filter KS applied to the output of the spectral attribute matcher KS are shown in Figure 11. Note that the assumption of ground planarity is built into the current version of the horizon schema, and that the real world in many instances presents us with more complex situations.

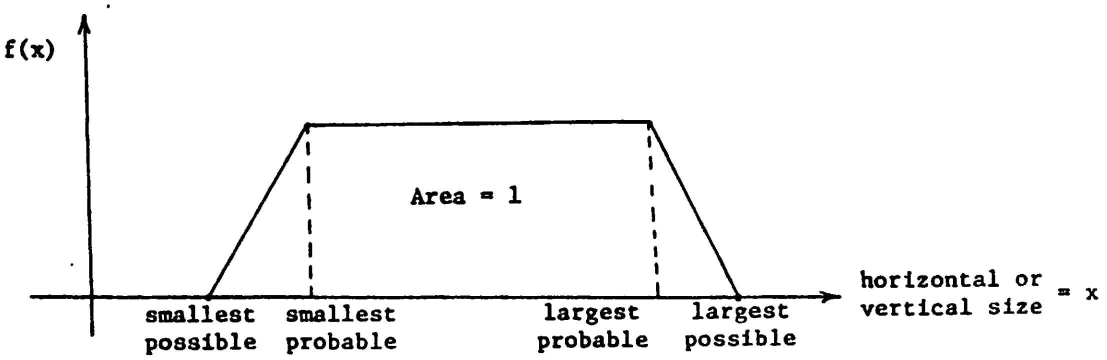
V. 12. Object Size KS

The object size KS is responsible for generating object hypotheses based on the size of a region (or collection of regions) and the results of the perspective KS. For example, once a region is known (or assumed) to represent the projection of a vertical surface, the perspective KS can compute the distance to the surface in the physical world and its physical height and width. The size KS uses this data to return a list of object hypotheses ordered by the confidence that the physical object could be the given size.

The size KS makes use of expected sizes of objects that are stored in LTM. Both major and minor axes and their expected orientation are used where possible. Figure 17(a) shows examples of the ranges of sizes for selected object classes in LTM. A piecewise linear approximation to the size probability density function is formed from these ranges as shown in Figure 17(b). Computation using only the vertical axis (for clarity) of several objects is shown in Figure 17(c); in this figure, the size

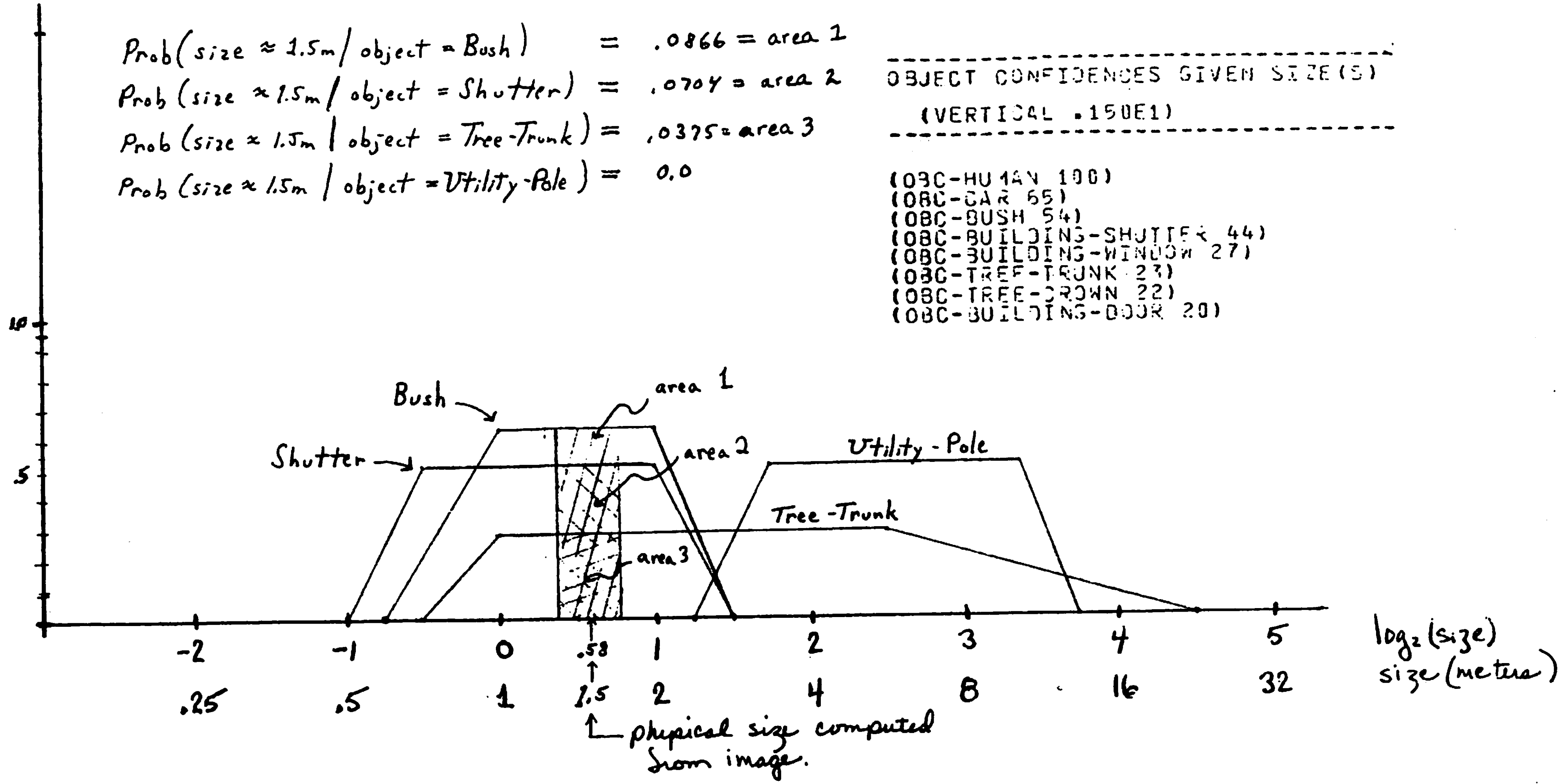
Object	Horizontal Axis				Vertical Axis			
	Smallest Possible	Smallest Probable	Largest Probable	Largest Possible	Smallest Possible	Smallest Probable	Largest Probable	Largest Possible
Building-Door	.6	.84	1.20	1.40	1.40	2.0	2.40	3.40
Building-Shutter	.25	.30	.60	1.0	.5	.71	2.0	2.80
Building-Side	3.40	4.80	11.30	32.0	1.70	2.40	16.0	27.0
Building-Window	.30	.60	1.20	2.40	.25	.35	2.40	3.40
Bush	.60	.84	1.70	4.80	.60	1.0	2.0	2.80
Car	2.0	3.40	4.80	6.70	.60	1.0	1.40	2.0
House	5.70	9.50	16.0	27.0	3.40	4.80	9.50	13.50
Human	.30	.42	.60	.84	1.0	1.40	2.0	2.40
Roof	2.0	4.80	16.0	27.0	2.0	2.40	4.80	6.70
Tree	1,70	3.40	6.70	13.50	2.0	3.40	13.50	32.0
Tree Crown	1.70	3.40	6.70	13.50	.71	1.0	9.50	19.0
Tree Trunk	.25	.25	.60	1.70	.71	1.0	5.70	23.0
Utility Pole	.25	.30	.42	.60	2.40	3.40	9.50	13.50

(a)



(b)

Figure 17. Object Size KS. (a) Typical size ranges for horizontal and vertical axes of some objects in LTM; all sizes are given in meters. (b) Approximation to a probability density function formed from the values in LTM.



Prob(size ≈ 1.5m / object = Bush) = .0866 = area 1

Prob (size ≈ 1.5m / object = Shutter) = .0704 = area 2

Prob (size ≈ 1.5m | object = Tree-Trunk) = .0375 = area 3

Prob (size ≈ 1.5m | object = Utility-Pole) = 0.0

```
------------------------------------
OBJECT CONFIDENCES GIVEN SIZE(S)
  (VERTICAL .150E1)
------------------------------------

(OBC-HUMAN 100)
(OBC-CAR 65)
(OBC-BUSH 54)
(OBC-BUILDING-SHUTTER 44)
(OBC-BUILDING-WINDOW 27)
(OBC-TREE-TRUNK 23)
(OBC-TREE-CROWN 22)
(OBC-BUILDING-DOOR 20)
```

↑ physical size computed from image.

(c)

Figure 17(c). Object confidences given a size are based on the probability of the size falling in a default ± 5% window (exaggerated for clarity), although the actual window can be set by the perspective KS on the basis of an error analysis of the computed size. These probabilities are scaled up, making the highest equal to 100. For expository purposes, only the vertical axis computation is shown; in actual applications, both horizontal and vertical extents are incorporated, resulting in a more constrained set of hypotheses.

coordinate axis is shown in both meters and the logarithm of meters.

The perspective KS returns a computed size and the range of the size; the default range is ±5%. Based on this window of size values, a confidence value is computed for each object in LTM from the ensemble of piecewise approximations. If this window falls outside the size range for the object, the confidence value is defined to be -100. Objects for which the window overlaps the expected range produce positive confidence values. This value is determined for each object by integrating the area under the curve (for that object) within the error window, and then normalizing by the largest value produced for any object (times 100 so that the largest will have a confidence value of 100).

Applying the size KS to RG-50 (a window shutter) of Figure 3 results in the hypotheses: tree trunk with confidence 100, shutter with confidence 35, and all others are negative. More extensive results are provided in Section VI.8.

VI. RESULTS OF INTERPRETATION WITH A SPECIFIC 2D SCHEMA

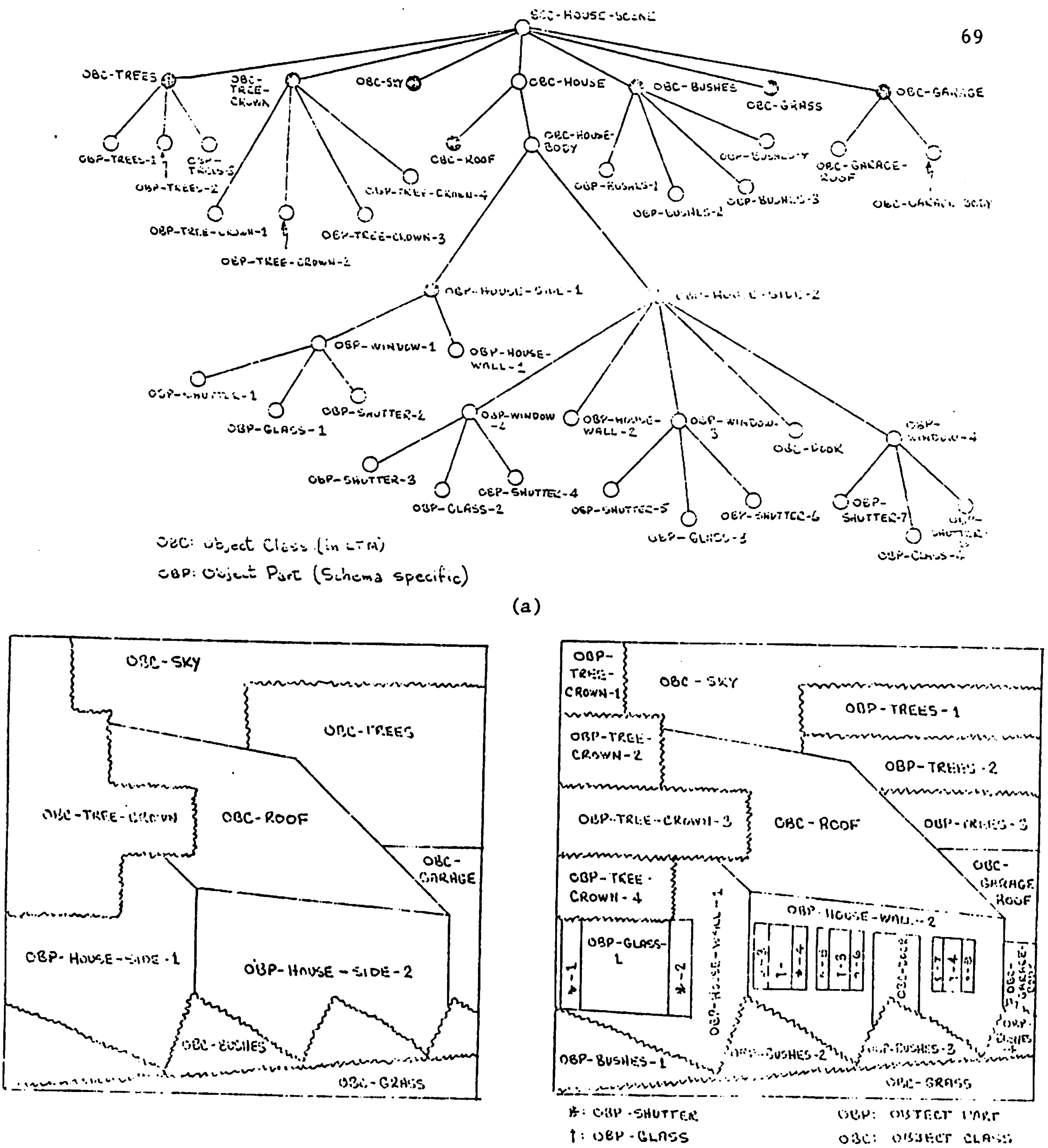
VI.1. Introduction

One of the purposes of 3D schemas is to generate the appearance of prototype scenes from any point of view. For example the 3D schema of a road scene can be rotated and projected to produce the image of a road scene as it would be expected to appear to an observer looking down the road. A particular projection of a 3D schema is referred to as a 2D schema and will be very useful in directing top-down analysis of the image. It is best thought of as a plan (a set of constraints) for interpretation of the image.

A 2D schema of a specific house scene viewed from a front diagonal perspective is implied by the illustration in Figure 18. This 2D schema is not a projection of a general house scene, but rather of a particular house scene. The general house schema would need to specify the expected variability of the general house scene.

The current representation of the 2D schema involves a set of information for each region including location of centroid, area, 2D symbolic shape with an aspect ratio of major to minor axis, color and texture features, location and properties of boundaries, object identity, 3D surface orientation and 3D size. To perform these experiments the 2D schema was generated manually, and current work will make it possible to drive a 3D schema representation and automatically form the projection, estimate likelihood of occluding schema surfaces, and fill out the required attributes from the LTM knowledge base.

It should be clearly understood by now that the current spatial representation of a 2D schema is not a direct copy of the model drawn in Figure 18(b,c), but instead approximates the location of this information.



(a)

(b)

(c)

Figure 18. 2D schema for a specific house scene. (a) Hierarchical structure of schema components. (b) The schema regions represented by the dark nodes in the hierarchy. (c) Schema regions associated with tip nodes in the hierarchy; this is the schema used in all the experiments of Section VI. The squiggly boundaries in the schema are for aesthetic purposes. Currently the position of schema regions is defined by parameters of centroid and area. Schema regions also may have additional parameters of color and symbolic shape, and any subset of these four parameters may be used by a matching function applied to image regions. Straight lines (without squiggles) represent boundaries whose shape and rough position is known, and can also be used to direct matches to nearby straight line segments in the image.

The position of a 2D schema region is defined by two parameters, the position of its centroid and its area. The squiggly boundaries in the 2D schema of Figure 18 are for display purposes. Actually, the positions of the schema regions are not known except to the degree that constraints are implied when there is a distinctive shape, such as rectangle with a particular aspect ratio. On the other hand, there are sometimes boundaries with known characteristics (e.g. long and straight) appearing in expected positions such as those bounding the roof in our house scene schema. Lines whose shape are known are drawn without squiggles where they are roughly expected to appear in the image.

Top-down control of the KSs in the interpretation of an image is relevant in the case where expectations about a given scene are available. The experiments in this section are intended to depict the case where the system is attempting to interpret a known scene via a specific 2D schema, i.e., from a known point of view (Stage 1 from Section III.4). We also assume that the camera model (focal length of lens, tilt angle, pan angle, and height above ground) is known. Results will be presented of the control by 2D schemas of the 2D shape KS, spectral attribute KS, fits of straight line segments, the perspective KS, and the size KS.

The 2D schema KS directs matching of schema regions to image regions and some schema line segments to image line segments. The matching process employs a weighted evaluation function on features of symbolic 2D shape, size, color, and position between regions in the image and in the schema. We will not go over the details of the heuristic match function here, although we note that any non-empty subset of the features can be used for matching. Note that it is necessary, in general, to expand or contract the schema in order to correlate schema size with image size. This is a function of distance and camera parameters and would have to be part of schema processing if it is to be robust in its application.

Matches can also be defined to operate between a schema region and groupings of several image regions, or a schema line segment to a group of image line segments, but then a search is necessary to discover the best groupings. The search for good matches can be directed by a variety of strategies. We will present simple results of a few.

VI.2. Semantically Directed Merging via 2D Schema

The first experiment will demonstrate the matching color and texture attributes in order to improve a fragmented segmentation. It involves the interaction of the spectral attribute KS and the 2D schema in an attempt to merge many adjacent regions whose object identities are the same. The strategy attached to the specific 2D schema for applying KSs to perform semantic merging is outlined in Figure 19. It first involves calls to the spectral attribute matcher to get a list of object types which it can match. The 2D schema contains information on the areas of the image in which these objects (tree, bush, sky, and grass) are

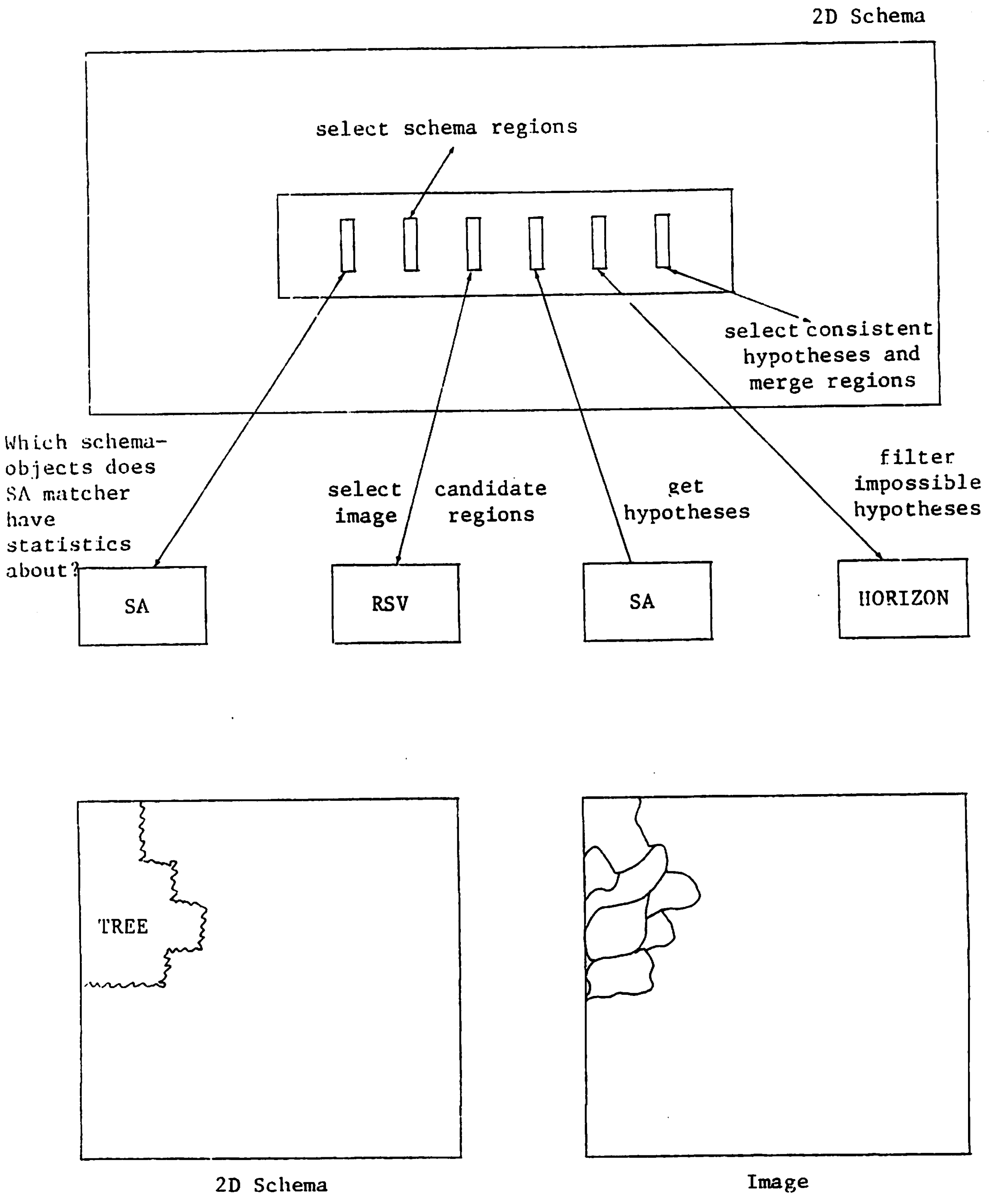

2D Schema
select schema regions
select consistent hypotheses and merge regions
Which schema-objects does SA matcher have statistics about?
select image
candidate regions
get hypotheses
filter impossible hypotheses
SA
RSV
SA
HORIZON
TREE
2D Schema
Image

Figure 19. Semantic merging strategy. The 2D schema determines from the spectral attribute matcher which schema objects it can classify, then selects schema regions which are the expected locations of those objects, then determines all image regions in those vicinities, checks which objects are implied by those attributes, filters object categories which are inconsistent with the horizon model, and then merges regions with identical labels that are consistent with the schema.

expected to appear. Thus, it distinguishes the areas expected to be target objects from the areas of non-target objects. The 2D schema then accesses the region segmentation to select candidate image regions for matching. Each schema region which is expected to contain one of the four types of objects above will be used to direct semantic merging via the attribute match KS. In these areas adjacent regions will be merged if their identities are verified by the attribute KS to be consistent with the schema. Thus, the attribute KS can be viewed as verifying the 2D schema plan.

Figure 20 shows that semantic merging allows most of the fragmentation in the tree to be merged, and separate grass and bush regions to be linked as well. The image is greatly cleaned up and more representative of the semantics of the scene. The results might be further improved by applying the horizon KS to filter the object hypotheses that are inconsistent with the approximate location of the horizon (which has been established via a camera model to be below the center of the image).

VI.3. Straight Line Segment Analysis via 2D Schema

Let us use the long straight lines in our 2D schema to search the image for good candidate matches (refer to Figure 21). The search is constrained by placing a rectangular mask around the selected schema edge (Figure 21b). All lines whose midpoint is inside the mask, and whose slope is within a specified tolerance of the slope of the schema line are selected as

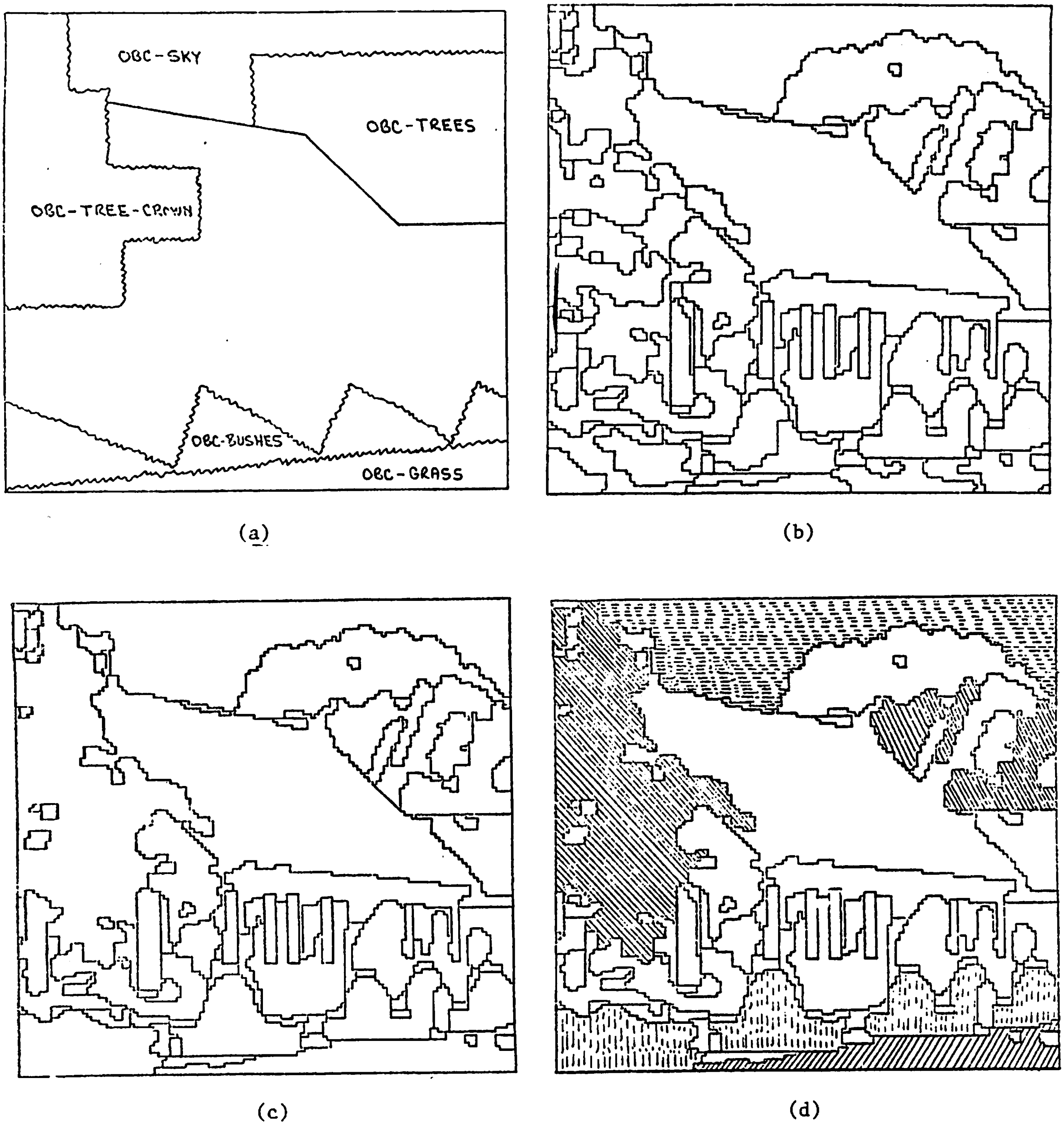


Figure 20. Merged regions via top-down guidance from 2D schema and bottom-up results from spectral attribute matcher. Adjacent regions are merged if the attribute match KS produces identities which are consistent with the 2D schema and horizon KS. (a) 2D schema showing area in image where tree, bush, sky, and grass are expected to appear. (b) Original segmentation. (c) Semantically merged image. (d) Same as (c). Cross-hatched semantically merged image.

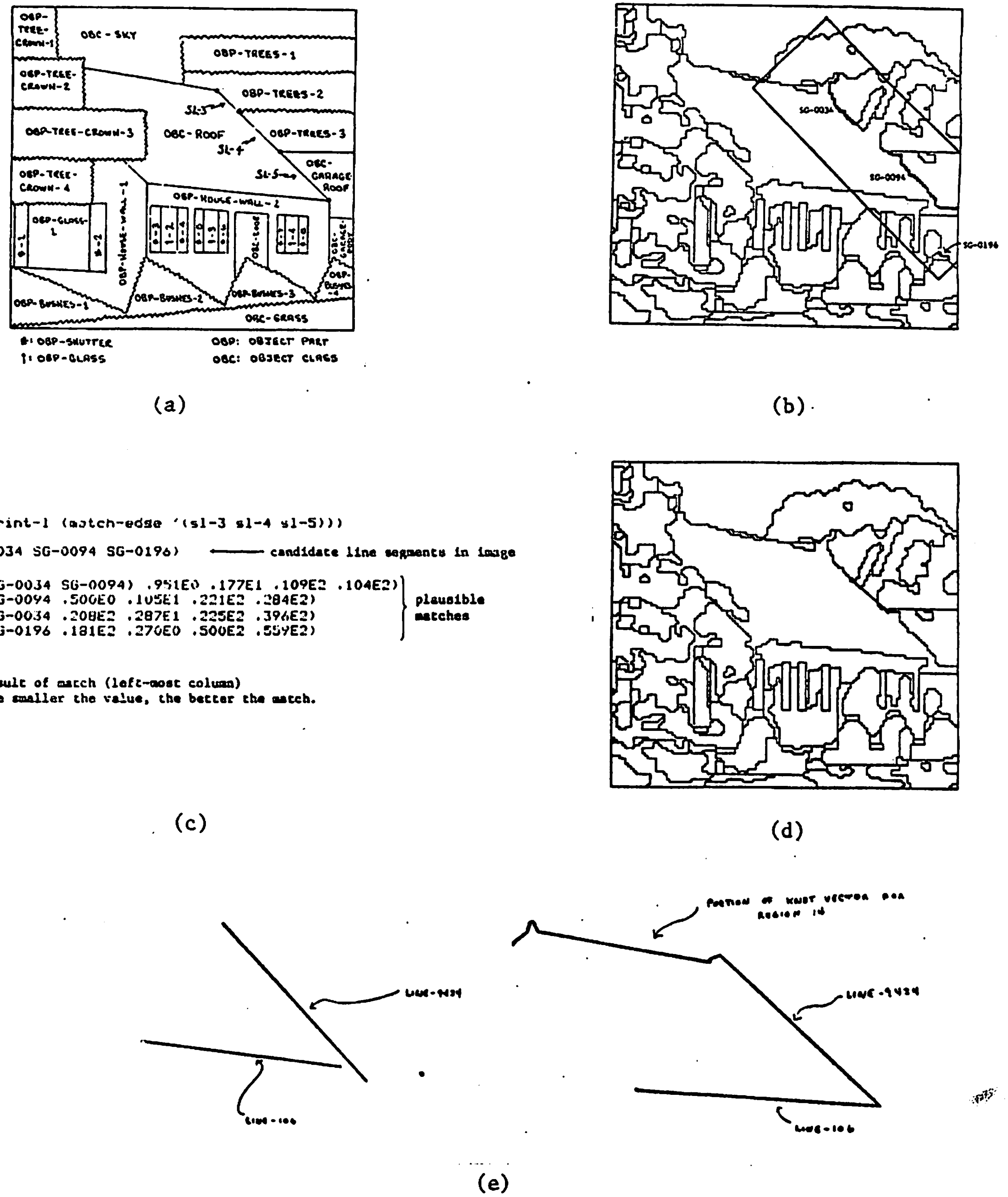
tree	▧	grass	▨
bush	▥	sky	▤

possible matches. The next step is to merge all co-linear segments within the mask into new segments, and then match all the resulting line segments to the schema line. The match is based upon attributes of slope, length, distance between centers, and RMS error, with a best (merged) match for each schema straight line segment. Results for two schema edges -- the right side and lower side of the roof -- are shown in Figure 21(c). The merging of image line segments 34 and 94 clearly produces the best fit to the schema straight line on the right side of the roof and this line segment is completed in Figure 21(d). The lower boundary of the roof also produces a clear match.

If the results of the line segment construction are fed back to the shape KS, region 14 is now identified as a parallelogram with 65% confidence. Note that we expect this to improve further, when the lower straight line of the roof is extended to meet the other straight lines and cut off the region leaks on both sides of the roof. Figure 21(e) shows straight line fits with minimum RMS error. It is estimated that the confidence can be increased to over 90%.

VI. 4. Symbolic Region Shape Matches via 2D Schemas

Certain regions in the schema and the image have symbolic attributes of simple geometric types such as rectangle, trapezoid, ellipse, etc. The shape attributes of schema regions, where they are relevant, are pre-defined (or else will be



(a)

(b)

```
 (pprint-1 (match-edge '(sl-3 sl-4 sl-5)))

(SG-0034 SG-0094 SG-0196)    ←——— candidate line segments in image

(6 (SG-0034 SG-0094) .951E0 .177E1 .109E2 .104E2)
(13 SG-0094 .500E0 .105E1 .221E2 .284E2)             plausible
(19 SG-0034 .208E2 .287E1 .225E2 .396E2)             matches
(29 SG-0196 .181E2 .270E0 .500E2 .559E2)

    result of match (left-most column)
    The smaller the value, the better the match.
```

(c)

(d)

(e)

Figure 21. Results from schema-directed straight line segment analysis. (a) High-level schema used to direct merging of segments. (b) Original segmentation showing mask used to locate candidate line segments. (c) The candidates for matching against schema segments SL-3, SL-4, SL-5 are SG-34, SG-94, SG-134, SG-224, as found in the masked area of (c). The results of matching combinations of these segments are shown in the left-most column. Clearly, segments SG-34 and SG-94 form the closest match. (d) Insertion of roof boundary segment as a result of schema match of SL-3, 4 and 5 to segments SG-34 and 94. (e) If straight line fits are used to improve the roof boundary, the confidence of a parallelogram can be increased sharply.

generated during 3D schema projection). The shape attributes of image regions can be determined by the 2D shape KS.

Let us examine the strategies depicted in Figure 22 for 2D shape matching. The schema requires access to the results of the 2D shape KS and the list of schema regions with distinctive geometrical shape. The shape matching function then can use shape and position to determine a degree of fit. There are three types of matching capabilities of schema and image regions using any subset of the four features:

a) location of centroid,

b) symbolic shape,

c) intensity/color, and

d) area.

The matching function can be applied:

a) directly between a schema region and an image region,

b) between a schema region and a group of adjacent image regions, and

c) between a template (possibly derived from a previous match of image regions) and groups of image regions.

Let us now examine the results summarized in Figure 23.

First, consider an attempted match -- without the use of postion information -- of all schema regions and image regions which have distinctive geometric shapes. This will show that the 2D schema can be robust without an exact spatial plan for the 2D image. The 2D shape KS was run on all regions within the expected house area, and those image regions which have a high

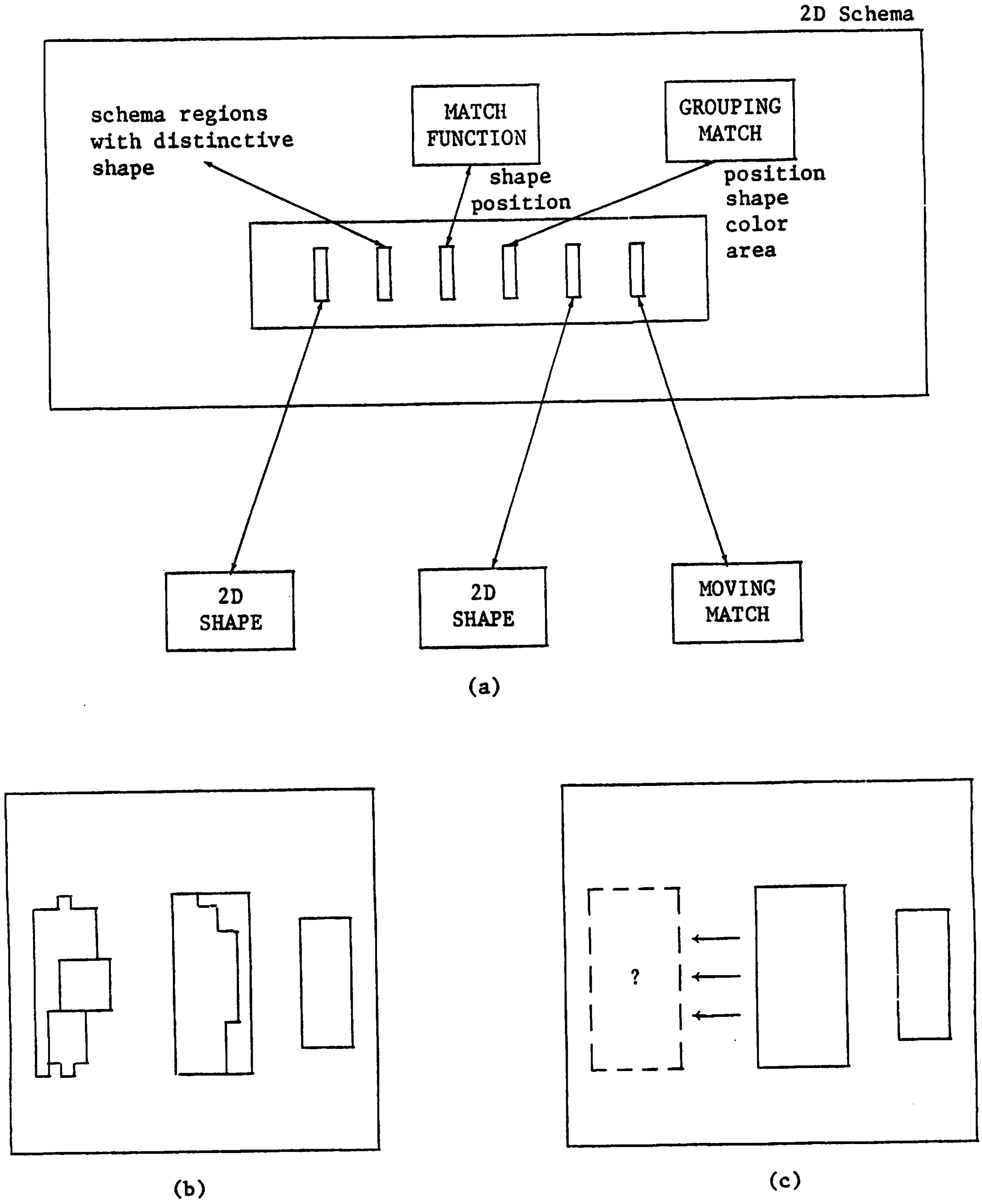
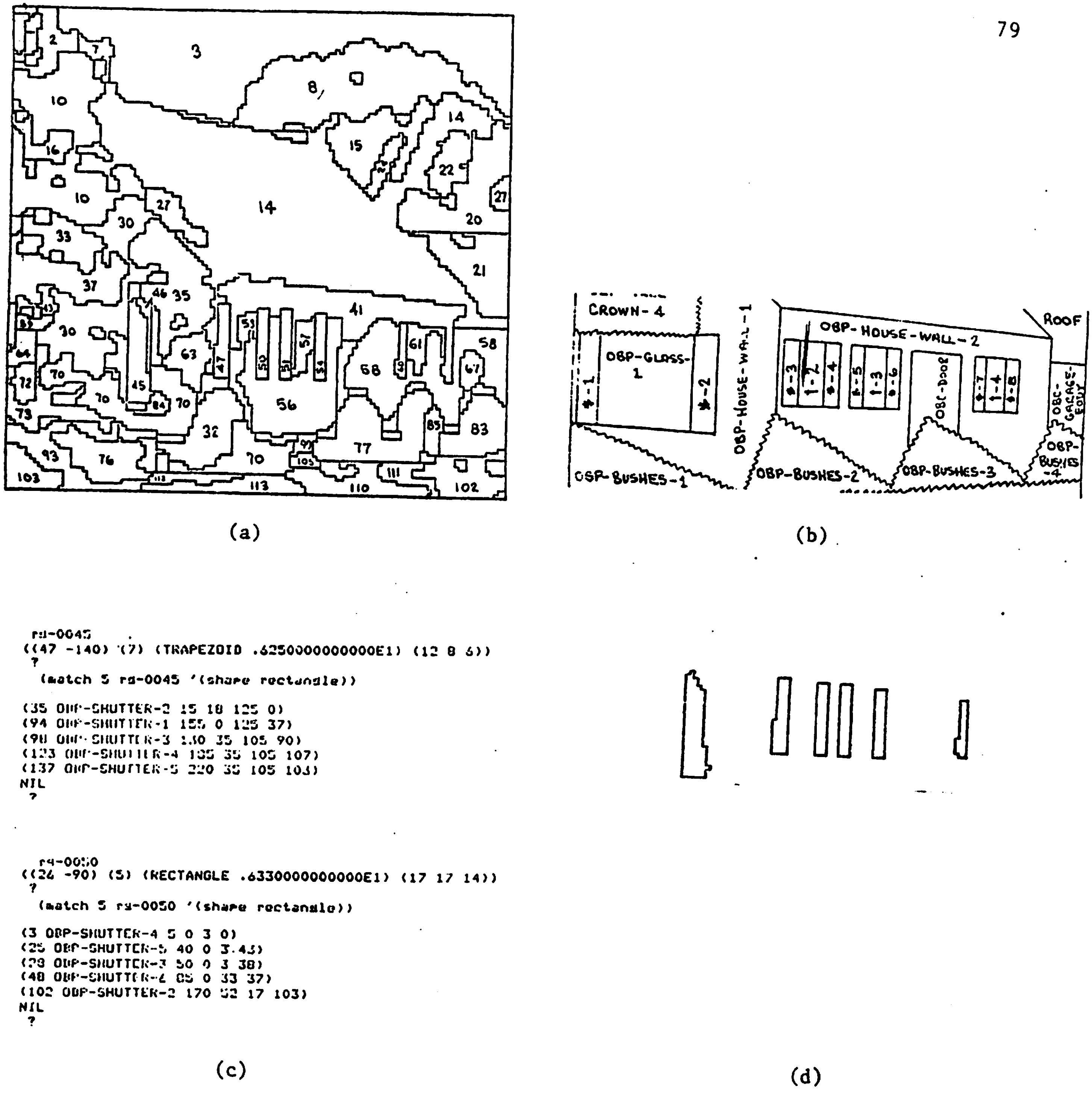


Figure 22. 2D shape matching. The 2D schema calls the 2D shape KS to extract regions with primitive geometrical shapes and then matches them with schema regions by symbolic shape label and position. The match can be applied to individual or groups of image regions. Additionally, features of color and expected area can be used. Image regions which are matched can be used as a template to be moved over the image.



```
  rd-0045
((47 -140) '(7) (TRAPEZOID .6250000000000E1) (12 8 6))
 ?
  (match 5 rd-0045 '(shape rectangle))

(35 OBP-SHUTTER-2 15 18 125 0)
(94 OBP-SHUTTER-1 155 0 125 37)
(98 OBP-SHUTTER-3 130 35 105 90)
(123 OBP-SHUTTER-4 135 35 105 107)
(137 OBP-SHUTTER-5 220 35 105 103)
NIL
 ?


  r4-0050
((24 -90) (5) (RECTANGLE .6330000000000E1) (17 17 14))
 ?
  (match 5 r4-0050 '(shape rectangle))

(3 OBP-SHUTTER-4 5 0 3 0)
(25 OBP-SHUTTER-5 40 0 3.43)
(23 OBP-SHUTTER-3 50 0 3 38)
(40 OBP-SHUTTER-6 85 0 33 37)
(102 OBP-SHUTTER-2 170 52 17 103)
NIL
 ?
```

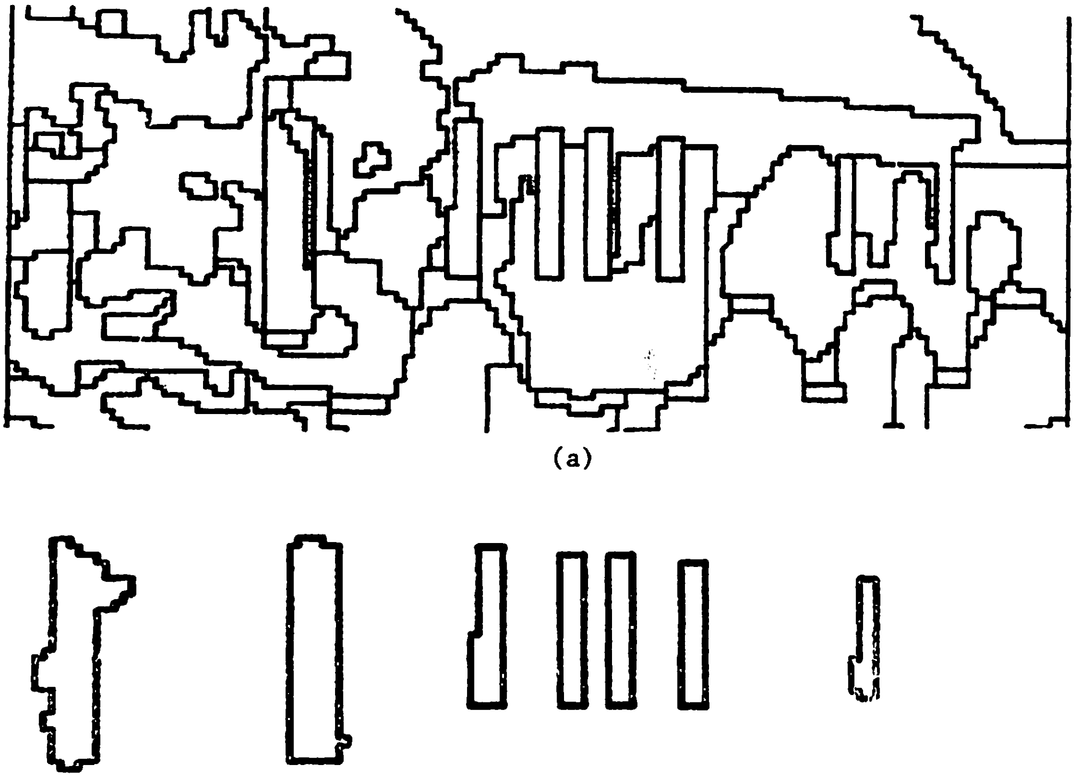
Figure 23. Shape matching via 2D schema.

Results of matching house shutters based upon shape matches and 2D shape KS. (a) Portion of the original segmentation. (b) Portion of the 2D schema. (c) Of the 5 image regions with high confidence of rectangle or trapezoid, two regions, 45 and 50, are matched against schema regions with roughly similar shapes. The match is based upon size, shape, and color and the best five matches are shown. Note that low evaluation is best. The overall match is shown on the left while the match factors of the features in the order given above are shown to the right. (d) The image regions found to match with shutters in the schema. Note that the feature of position (neither schema nor region) was not employed.

confidence of a primitive shape type can be further processed. Each of these regions is used to match against schema regions with similar shape based on attributes of size, shape/aspect ratio, and color. The result of matching regions 45 and 50 with the best five schema regions is tabulated in Figure 23(c). They are found to match reasonably well with the various shutter regions in the schema and poorly with other schema regions. The left shutter has fragmented in the original segmentation and region 45 is closer to a trapezoid than a rectangle. Consequently, it has a poorer match with rectangular shutters than region 50. It should be noted that the evaluation function is scaled into 0 (perfect matches) to 1000 (no match); this evaluation function has not yet been made consistent with the form of other KS outputs.

The second step shows the improvements obtained by the addition of positional information to better form correspondences between schema shutter regions and image regions. There is a good match for five of the six shutters in the front and one of the two on the left (Figure 23c,d). Note that the left-most shutter has not been found and only a part of the next one has been found because of region fragmentation.

Figure 24 demonstrates the grouping capabilities of the 2D schema by focussing on the left two large shutters in the image. The centroid of the schema region is used to select candidate regions for grouping and the match function (based upon all the features) is used to select the subset which matches best. The



(a)

(b)

```
(match-groups left-shutter '(single obj-shutter-2))

(35 (RG-0072 RG-0064 RG-0059 RG-0052 RG-0044) 18 52)
(39 (RG-0072 RG-0064) 35 42)
(60 (RG-0064 RG-0059 RG-0052 RG-0044) 35 85)
(60 (RG-0069 RG-0064 RG-0059 RG-0052 RG-0044) 35 85)
(65 (RG-0049 RG-0040 RG-0046 RG-0045) 18 112)
(78 (RG-0050) 52 103)
(93 (RG-0075 RG-0071 RG-0047) 18 167)
(96 (RG-0075 RG-0047) 35 157)
(131 (RG-0073 RG-0069 RG-0059) 18 243)
(136 (RG-0049 RG-0046) 52 220)
(138 (RG-0053) 35 240)
(142 (RG-0048) 81 203)
(150 (RG-0074 RG-0065) 52 247)
(156 (RG-0075 RG-0071) 68 243)
(160 (RG-0063) 18 302)
(172 (RG-0070 RG-0046 RG-0045) 104 240)
(180 (RG-0095 RG-0086 RG-0070 RG-0065 RG-0045) 120 240)
(187 (RG-0055 RG-0035) 70 303)
(194 (RG-0074 RG-0070 RG-0065) 87 300)
(199 (RG-0062) 92 305)
```

(c)

Figure 24. Schema-directed grouping of image regions with simple and distinct geometric shape. (a) Portion of original segmentation extracted from Figure 3 showing the fragmentation of the left two shutters. (b) Grouped regions found by 2D schema. The rightmost shutter of the left pair was found using the centroid of the schema region to select candidate regions for grouping. (c) Results from match function when a mask formed from the right shutter of the pair is moved to the left and matched against groupings of candidate regions on the basis of color and size. Regions 72, 64, 59, 52, and 44 match best. The merged collection is shown in (b). The confidence that the second region from the left is a rectangle

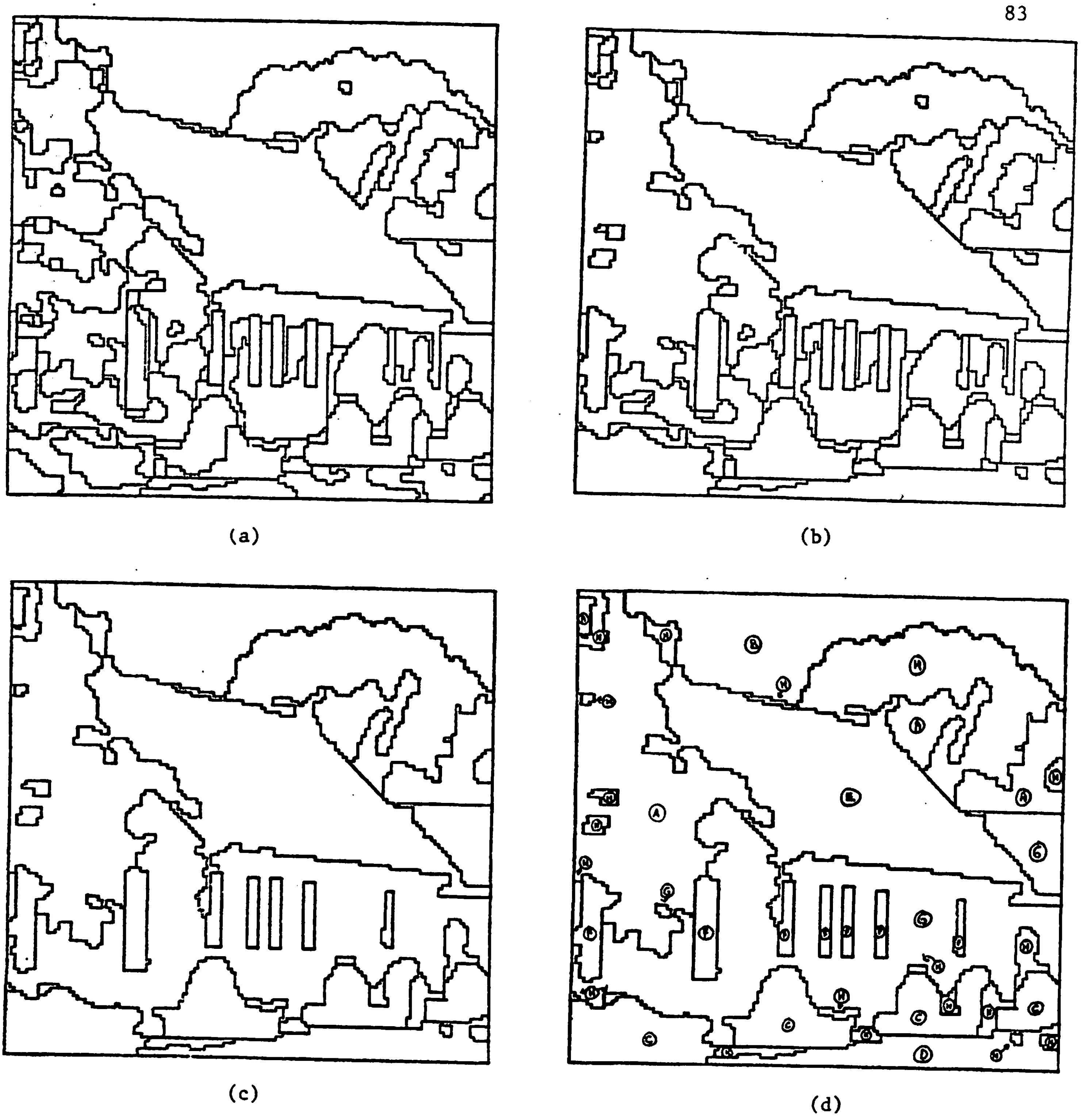
right shutter of the pair is extracted by this technique, but due to the severe fragmentation of the left shutter this technique was not employed. The shutter on the far left was found by moving a template, of the size of the right shutter, towards the left and grouping regions on color and size. The best match is then selected. The 2D shape KS is then applied to determine the rectangular fit on the left two shutters, producing confidences of 24% and 94%, respectively. It is difficult to interpret the 24% value at this point since there has not yet been any tuning of the performance curves of the shape confidence measure; we do not know, as yet, how fast the match values decrease relative to a 'good' match.

VI. 5. Combination of KS Results

The result of integrating the hypotheses of the attribute KS, line segment matches, and the 2D shape KS yields the results in Figure 25. Note that many of the regions in the image are labelled with the proper object identity. Figure 25(c,d) was produced by a clean-up process of merging unlabelled adjacent regions within the house schema region and the remaining background area.

VI. 6. Formation of a Spatial Plan Using Perspective Information

The proper use of the perspective KS requires that a set of assumptions be generated regarding the orientation of surfaces.



(a)

(b)

(c)

(d)

Figure 25. Combination of schema-directed KS results. (a) Original segmentation. (b) Combined results of 2D schema with attribute KS, line segment matching, and region shape KS. (c) All regions without semantic labels are merged under guidance of 2D schema (i.e., unlabelled house regions are kept separate from unlabelled background regions. (d) Same as c, but labels are provided on diagram:

Ⓐ	tree	Ⓔ	roof
Ⓑ	sky	Ⓕ	shutter
Ⓒ	bush	Ⓖ	unlabelled house
Ⓓ	grass	Ⓗ	unlabelled background

In practice they would be determined via the specific 3D schema and other information from long-term memory, but in this case the set of assumptions necessary to drive the perspective KS are obtained directly from the 2D schema. Thus, knowledge that a particular region is bush, and that bushes are usually perpendicular and attached to the ground plane, is available to the 2D schema if it has been generated from a 3D schema. These critical assumptions allow the perspective KS to place that region (bush) in the 3D world model.

Let us consider the strategy for the computation of the distance and size of an unoccluded object which is perpendicular to and touches the ground plane; this strategy will be applied to computing the range and height of the bushes. The strategy by which the 2D schema controls the application of processes is outlined in Figure 26. The spectral attribute matcher KS can be used to validate the regions presumed to be bush and grass. Their common boundary implies that it is unlikely that the bottom of the bush is occluded. Next the perspective KS is called to determine the distance and size of the bush. In this example the range of the bushes is based upon two assumptions: vertical orientation and the elevation of the bottom of the bushes is 0. Then, the identity of regions 102, 110, 112, and 113 as grass implies that there is no occlusion of these bush regions. Hence, the image coordinates of the region can be translated into a range in the physical world. Once the range is computed, then the image size -- region height and width -- allows the physical

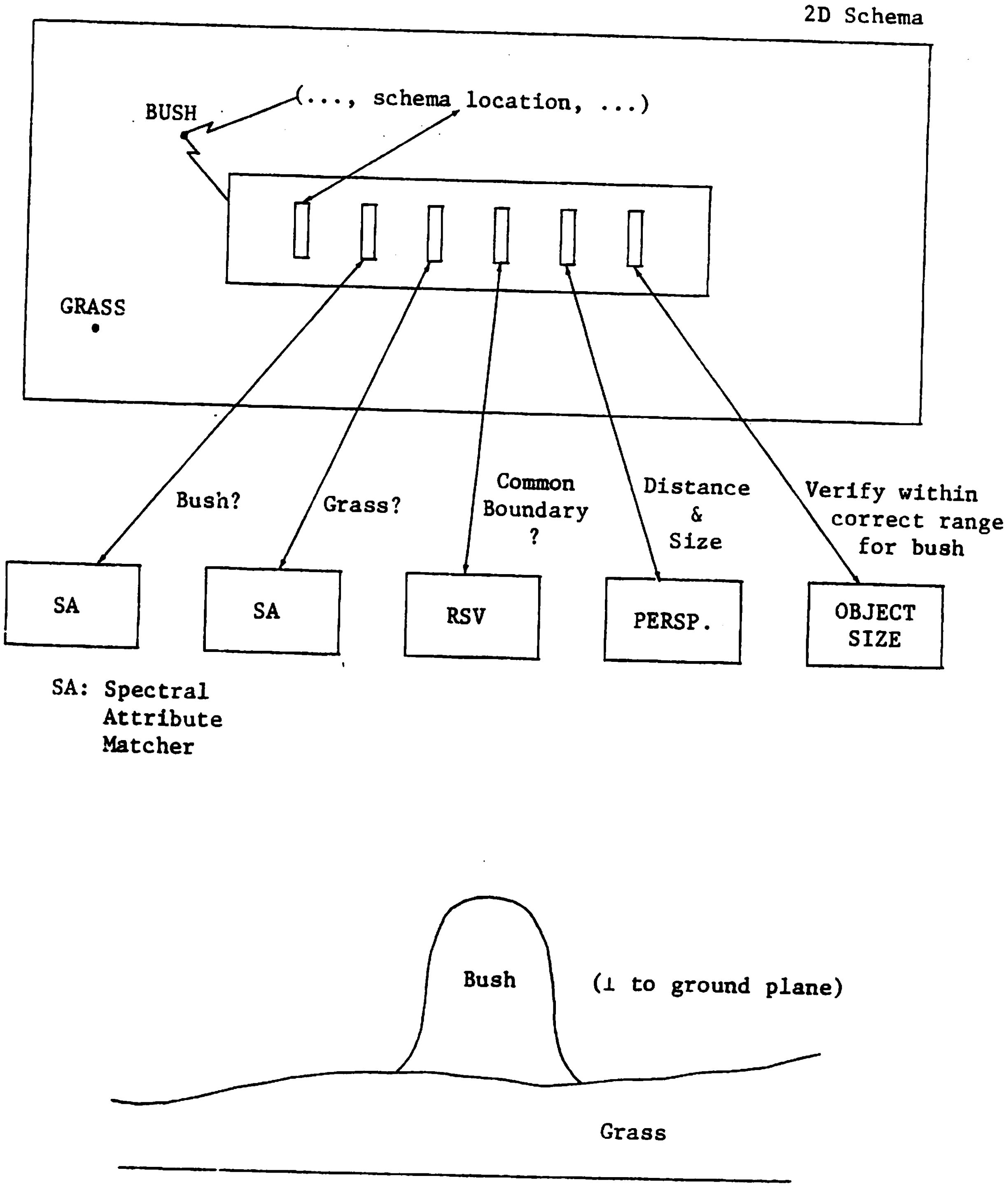


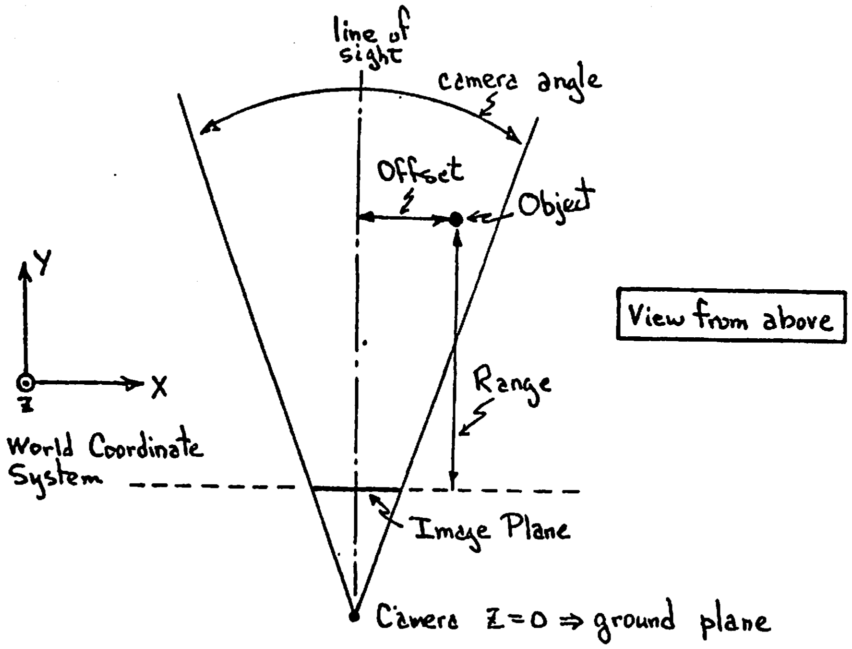
Figure 26. Strategy for computing size and distance of unoccluded object which touches ground. The SA KS is used to verify that the regions expected to be bush and grass. The fact that they have a common boundary implies that the bottom of the bush is not occluded (assuming the ground plane is planar). The perspective KS is used to compute the distance and size, while the object size KS verifies that the computed size is in the expected range of bush sizes.

size to be computed. Note that in order to carry out this analysis, the system employed the 2D schema, the spectral attribute KS, and the perspective KS. The inference drawn from this chain of hypotheses, namely that the region represents a bush, can be partially validated by noting that the computed size falls within the allowed range for bushes stored in long-term memory (see VI.7).

Figure 27(a) describes the camera geometry from a bird's eye view with the image plane shown in front of the focal point for convenience. The range, offset, and elevation of a surface/object in the physical world must be computed in terms of the viewer-centered coordinate system involving the line of sight of the camera. Figure 27(b) lists results of applying the perspective KS, under control of the 2D schema, to selected regions of our test image. All the regions considered (bushes, shutters, house wall) lie roughly (particularly the bushes) in a pair of planes which are vertical to the ground plane and oriented at a diagonal to the right, away from the viewer. The location of objects are graphically portrayed in the bird's eye view of Figure 27(c).

In order to use the results in an effective manner, an error analysis should be taken into consideration. With an assumption of ground planarity and a camera model (focal length = 50 mm, elevation about 2 meters, because the person was standing on higher ground, tilt = 2 degrees upward), then the range of a physical point in the ground plane can be derived directly from the image coordinates of the point (in pixels). However, the computation is not a linear function of this image distance, and both the physical range and its associated error increase exponentially. Table IV lists the absolute and relative error of a one-half pixel error for each row of pixels starting from the top of the image (i.e., row 1 in our 128x128 pixel image). The error in the range is shown superimposed on the location of objects in Figure 27(c). A one-half pixel error in width will produce an error in physical width which is relatively constant over the image unless the camera has a wide-angle lens (e.g., a fish-eye lens). Note that error in range will propagate directly into an error in physical height and width and this must be taken into consideration by the object size KS.

Even such simple perspective results as shown provide the beginnings of a 3D spatial layout. The ranges of the row of bushes in front of the house provide a range of possible orientations for region 56 (the house wall). This partial plan, shown as a bird's eye view, is illustrated in Figure 27(c). The angle of the shutters has been computed to be 24 degrees from the line of sight. The house in Figure 3(a) does not seem to be oriented at such a steep angle, but there is significant foreshortening. This orientation has been determined to be accurate via external physical examination of the environment.



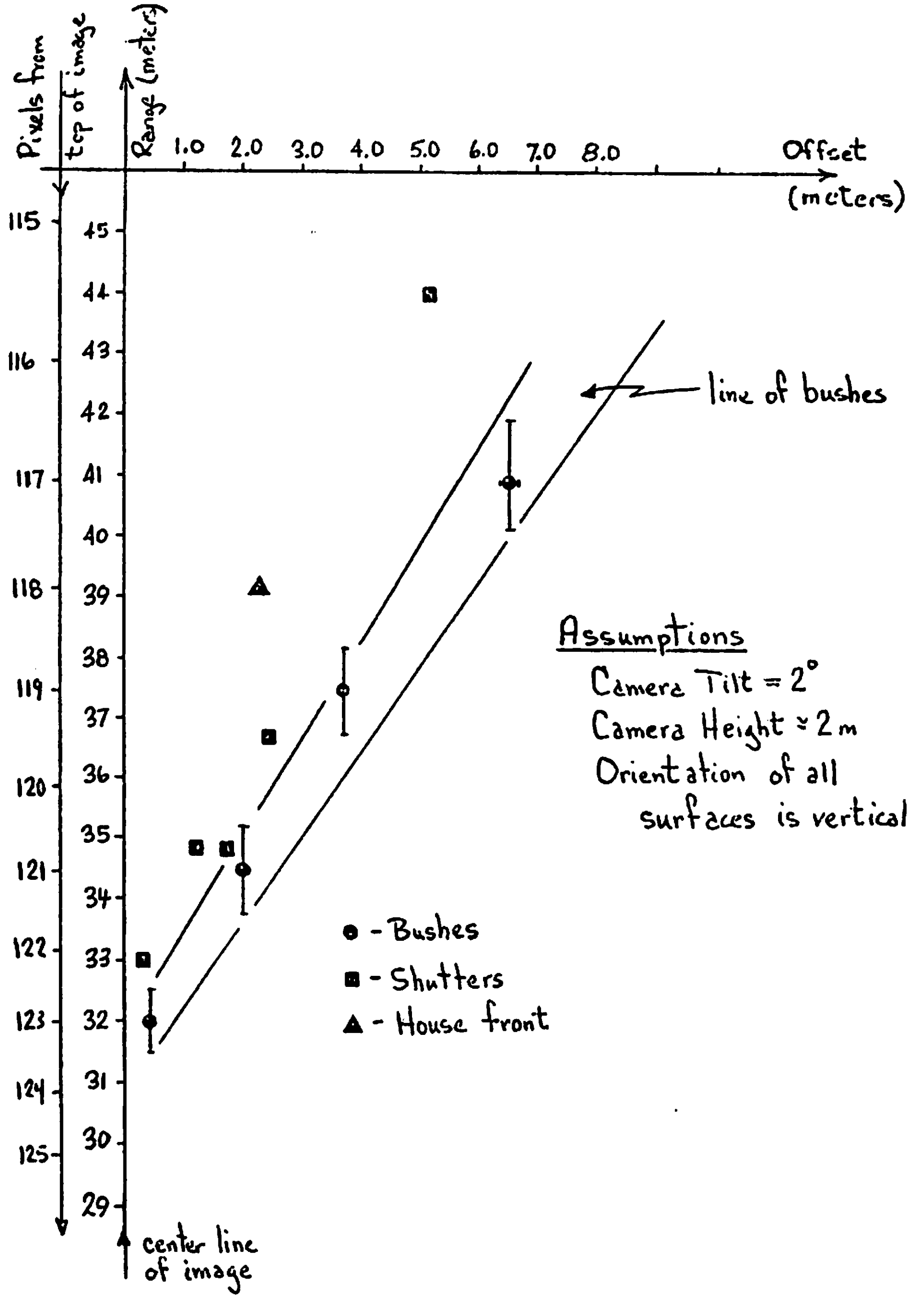
World Coordinates	
Y	Range: distance to object on a line perpendicular to image plane
X	Offset: distance of object to right of line of sight
$\|X_1-X_2\|$	Width of object: |offset of right side — offset of left side|
Z	Elevation: distance of object above ground plane
Z_2-Z_1	Height of object: elevation of top — elevation of bottom

(a)

Region	Region Identity	Range	Offset	Width	Height	Assumptions
90	Bush	32.0		2.32	1.01	vertical, attached to ground
99	Part of House front	39.1				same as 90
82 & 90	Bush	32.0	.472	2.54	1.59	same as 90; 82 vertical with same range as 90
79	Bush	37.5	3.78	2.63	1.61	same as 90
83	Bush	40.9	6.54	1.58	1.48	same as 90
56	House front	39.1		2.39	2.83	99 and 56 lie in one plane; range of 56 is same as 99; 56 is vertical
47	Shutter	33.0	.375	.300	1.50	vertical, height = 1.5 m
50	Shutter	34.8	1.22	.237	1.50	height = 1.5 m
51	Shutter	34.8	1.70	.237	1.50	height = 1.5 m
54	Shutter	36.7	2.54	.250	1.50	height = 1.5 m
60	Shutter	44.0	5.15	.300	1.50	height = 1.5 m

(b)

Figure 27.



(c)

Figure 27. Results of forming a spatial plan using the Perspective KS. (a) Imaging geometry and description of terms used in presenting the perspective results. The Z axis represents the gravitational vertical; for the example image, the line of sight is inclined 2° from the X,Y plane. (b) Computation of physical location and size based upon assumptions shown in the right-hand column. (c) Ground plan of house determined by the perspective KS. The results of (b) in terms of range and offset fix the locations of objects in the X,Y plane. Both range and offset are expressed in meters. The two vertical scales show the correlation between range and rows of pixels in the image. If a pixel in a row is assumed to have elevation 0, then the physical range is obtained by reading the range scale. The error range of Table IV is superimposed as a vertical line through the location of the bushes; the angle of the bushes is computed to be 24° from the line of sight.

Pixel in Row	Computed Range of pixel	Absolute Error (meters)	Relative Error (%)
bottom of → 128	27.1	.416	1.5
image 127	28.0	.442	1.6
126	28.9	.472	1.6
125	29.9	.504	1.7
124	30.9	.540	1.7
123	32.0	.580	1.8
122	33.2	.624	1.9
121	34.5	.674	1.9
120	35.9	.730	2.0
119	37.5	.794	2.1
118	39.1	.865	2.2
117	40.9	.945	2.3
116	42.9	1.042	2.4
115	45.1	1.15	2.5
114	47.5	1.28	2.7
113	50.2	1.43	2.8
112	53.3	1.61	3.0
111	56.7	1.82	3.2
110	60.6	2.01	3.4
109	65.0	2.39	3.7
108	70.2	2.79	4.0
107	76.2	3.30	4.3
106	83.4	3.95	4.7
105	92.2	4.82	5.2
104	102.9	6.01	5.8
103	116.	7.71	6.6
102	134.	10.25	7.6
101	158.	14.3	9.0
100	193.	21.3	11.0
99	247.	35.1	14.2
98	342.	69.0	20.1
97	559.	197.1	35.2
horizon 96	1529.	5294.	346.
→ 95		-6180.	296.

Table IV. Error analysis for perspective KS. It is assumed that a pixel represents a physical point in the ground plane (i.e., at elevation 0). The range of the physical point and its associated error, under an assumption of one-half pixel error in the image, are computed as a function of the row of pixels in which it appears in the image. This table was derived via the camera model for the specific image under consideration: f = 50 mm, tilt = 2°, elevation = 2 m (because the picture was taken from a slight rise in the terrain).

VI.7. Object Hypotheses Based on Size

Once the perspective KS has provided hypotheses about ranges of surfaces and the physical sizes of their projections, the size KS can be used to generate object hypotheses on the basis of the computed sizes. Figure 28 shows the hypotheses and their associated confidences formed by applying the size KS to selected regions from Figure 3. In each case, the default range on size (computed size ±5%) was used, although these values can be set by the result of the perspective KS and the location of the region in the image. Also note that these results could be filtered by spatial location, much as the hypotheses formed from the spectral attribute matcher were. This results in a partial check on the assumptions used by the perspective KS during the computation of the size.

VI.8. Bottom-Up Schema Instantiation

The results discussed in the previous sections (VI.1 to VI.7) were obtained primarily on the basis of top-down guidance from the correct 2D schema. This section describes a simple experiment to instantiate a schema on the basis of bottom-up data.

In this experiment, the inference net was used to propagate data upwards from the object level to the schema level, assuming that bush, tree, four shutters, and grass were instantiated at the object level. From the results cited earlier, it is reasonable to expect that these objects could be obtained from

Region	Height	Width	Actual Identity	Object Size KS Hypotheses
82&90	1.59	2.54	Bush	(OBC-CAR 100) (OBC-BUSH 58) (OBC-TREE-CROWN 25)
54	1.50	.25	Building Shutter	(OBC-TREE-TRUNK 100) (OBC-BUILDING-SHUTTER 35)
41&56	4.30	5.7 to 32	House Body or Building Side	(OBC-HOUSE-BODY 100) (OBC-GARAGE 74) (OBC-ROOF 74) (OBC-HOUSE 71) (OBC-BUILDING-WALL 34) (OBC-BUILDING-SIDE 34) (OBC-TREE 18) (OBC-TREE-CROWN 14)
20	3.5 to 9.7*	3.3	Tree or Tree Crown	(OBC-GARAGE 100) (OBC-TREE 90) (OBC-TREE-CROWN 67) (OBC-ROOF 40) (OBC-GARAGE-BODY 17) (OBC-GARAGE-FRONT-SIDE 12) (OBC-BUILDING-WALL 1) (OBC-BUILDING-SIDE 1)
14	3.75	5.70	Roof	(OBC-GARAGE 100) (OBC-ROOF [illegible]) (OBC-GARAGE-BODY 59) (OBC-TREE 33) (OBC-GARAGE-FRONT-SIDE 32) (OBC-BUILDING-WALL 26) (OBC-BUILDING-SIDE 26) (OBC-TREE-CROWN 24) (OBC-SKY 4) (OBC-HOUSE-BODY 3) (OBC-HOUSE 1)

* Elevation of top of region

Figure 23. Summary of results of object size KS for selected regions of Figure 3. The sizes shown were computed by the perspective KS using the default ± 5% error range (see Figure 17). The actual range can be set by the perspective KS on the basis of the error analysis.

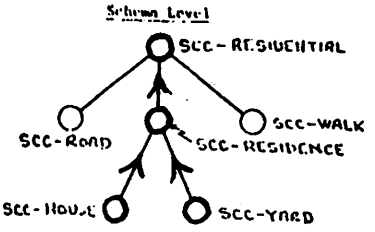
bottom-up analysis of the image: bush, tree, and grass from the spectral attribute matcher, and shutter and roof from the combined horizon filter, 2D shape matcher, perspective, and size.

The results of this experiment are shown in Table V; they overlap somewhat those shown in Table III. The strategy used to obtain the results was very crude. They are based solely on the propagation of the set of object identities to the schema level via the inference net. No attempt was made to validate the results via top-down matches, such as spatial location, or any other information in the schema. Note that instantiating one or two shutters increases the confidence of the house scene schema as expected: additional instantiations will not significantly increase this confidence. Information about the expected number of shutters on a house (or for a more exact example, the number of tires on a car) is stored in arcs in LTM and is thus taken into account by the inference net KS.

VII. CONCLUSIONS

The results cited in this paper represent the current state of development of the VISIONS system. A top-down interpretation of a scene has been successfully performed, although the conditions under which this interpretation was obtained were highly constrained. It demonstrates some degree of integration of the system, from automatic segmentation of the digitized input

Objects --> Schemas



Relevant portion of schema hierarchy

```
Instantiation at Object Level                                  Probability
  Obc-Roof:                 R                                      .7
  Obc-Bush:                 B1                                     .7
                            B2                                     .7
  Obc-Building Shutter:     S1                                     .7
                            S2                                     .7
  Obc-Tree:                 T                                      .7
  Obc-Grass:                G                                      .7
                                                                   .7
```

```
                        At Schema Level

Prior    Schema           Posterior  Objects Instantiated

 .1      SCC-RESIDENTIAL
                          (.98329E0 B1 B2 G R S1 S2 T)
                          (.98067E0 B1 B2 R S1 S2 T)
                          (.97501E0 B1 B2 G R S1 S2)
                          (.97155E0 B1 G R S1 S2 T)
                          (.97155E0 B2 G R S1 S2 T)
                          (.96755E0 B1 B2 R S1 S2)
                          (.96717E0 B1 R S1 S2 T)
                          (.96717E0 B2 R S1 S2 T)
                          (.95769E0 B1 G R S1 S2)
                          (.95769E0 B2 G R S1 S2)
                                 :
                          (.28105E0 G S1)
                          (.26627E0 G T)
                          (.22975E0 B1)
                          (.22656E0 T)
                          (.20894E0 S2)
                          (.16035E0 G)

 .234    SCC-RESIDENCE
                          (.98577E0 B1 B2 G R S1 S2 T)
                          (.98356E0 B1 B2 R S1 S2 T)
                          (.97874E0 B1 B2 G R S1 S2)
                          (.97585E0 B2 G R S1 S2 T)
                          (.97580E0 B1 G R S1 S2 T)
                          (.97240E0 B1 B2 R S1 S2)
                          (.97208E0 B2 R S1 S2 T)
                          (.97208E0 B1 R S1 S2 T)
                          (.96401E0 B2 G R S1 S2)
                          (.96401E0 B1 G R S1 S2)
                                 :
                          (.40743E0 B2 G)
                          (.38841E0 G S1)
                          (.37584E0 G T)
                          (.34480E0 B2)
                          (.34206E0 T)
                          (.32707E0 S2)
                          (.20574E0 G)
```

```
Prior  Schema     Posterior  Objects Instantiated   ||  Prior  Schema      Posterior  Objects Instantiated

.235   SCC-YARD                                     ||  .234   SCC-HOUSE
                  (.99061E0 B1 B2 G T)              ||                     (.97240E0 B1 B2 R S1 S2)
                  (.99527E0 B1 B2 T)                ||                     (.95349E0 B1 R S1 S2)
                  (.98777E0 B1 B2 G)                ||                     (.95349E0 B2 R S1 S2)
                  (.98695E0 B1 G T)                 ||                     (.92264E0 R S1 S2)
                  (.98695E0 B2 G T)                 ||                     (.90097E0 B1 B2 R S2)
                  (.96524E0 B1 B2)                  ||                     (.90097E0 B1 B2 R S1)
                  (.96297E0 B1 T)                   ||                     (.84112E0 B1 R S2)
                  (.96297E0 B2 T)                   ||                     (.75494E0 R S2)
                  (.89477E0 B1 G)                   ||                     (.75494E0 R S1)
                  (.89477E0 B2 G)                   ||                     (.66505E0 B1 R)
                  (.88846E0 G T)                    ||                     (.56765E0 S1 S2)
                  (.74517E0 B1)                     ||                     (.53468E0 R)
                  (.74517E0 B2)                     ||                     (.47491E0 B1 B2)
                  (.73249E0 T)                      ||                     (.45517E0 B2 S1)
                  (.47248E0 G)                      ||                     (.34481E0 B1)
                                                    ||                     (.32711E0 S2)
```

Table V. Experiments in the bottom-up instantiation of schemas. These results are based on the object identities shown, and are expected to be derived predominantly from bottom-up processing. The results are preliminary.

through symbolic output of object identities and generation of a rough plan of the three-dimensional space in the scene.

The primary emphasis of our current efforts is on the development of strategies by which the many knowledge sources can be integrated in order to interpret 2D color images. However, the ability to obtain the correct interpretation is inherently linked to the quality of information provided by these processes: without plausible hypotheses about the image, there isn't any control strategy worthy of investigation! Nevertheless, it is not feasible for us to attempt to perform extensive research in all the areas represented by the KSs. Thus, we must balance our efforts in the development of more complete knowledge sources against the development of interpretation strategies. Currently, we have implemented at least a simple version (and sometimes a complex version) of several KSs.

Each of the KSs developed can be used in different ways to produce several different kinds of hypotheses. The experiments already performed seem to indicate that there may be many mini-strategies for using the KSs in particular ways across the range of images. For example, the perspective KS can determine physical dimensions of surfaces, while the object size KS uses these results to produce a confidence measure for object hypotheses; or the horizon KS can be used to filter implausible object identities from the output of the spectral attribute matcher KS. Interesting strategies can be modelled in terms of the overlap of information related to perspective, occlusion, size, shape, junction analysis, etc. With proper design the set of local processes may be built to answer the questions that are of importance to each other, and this network of processes can be flexibly and incrementally constructed. As the strategies are understood, they can be incrementally embedded in the schemas.

The results presented in this paper were generated via top-down control of the KSs using a specific 2D schema -- in effect a plan -- for a specific house scene. The analysis was highly biased towards success because the schema is tuned to the particular situation: the case of looking at a familiar scene from a familiar point of view. It does, however, show some of the ways that the KSs are able to interact, and can also be viewed as an experiment in verifying that some stored schema is applicable to a given image. The last experiment demonstrates bottom-up interaction of the KSs in an attempt to instantiate the proper schema from a set of schemas.

The facilities now exist for actually developing to a much deeper level some of the ideas we have only been able to suggest as promising. The benefits of some of the interesting developments of our colleagues in the research community over the last few years has led to a deeper appreciation of the problems yet remaining. This is reflected somewhat in a shift of research emphasis, as we propose a highly structured research paradigm for exploring the issues we set forth. A series of increasingly more difficult experiments will provide an experimental methodology for developing schema-driven (e.g., top-down) control mechanisms;

each succeeding experiment will assume a set of weaker constraints, representing image interpretation tasks where a decreasing amount of knowledge of the situation is available. It is worth noting, however, that the basic approach is not substantially different from the initial top-down approach that started the VISIONS project [HAN74, RIS74], although it is considerably richer in detail.

VIII. BIBLIOGRAPHY

[AGI72] G. J. Agin, "Representation and Description of Curved Objects," Stanford AI Memo 73, 1972.

[AGI76] G. J. Agin and T. O. Binford, "Computer Description of Curved Objects," IEEE Transactions on Computers, April 1976, pp. 439-449.

[AHL67] Ahlberg, Nilson, Walsh, Theory of Splines and Their Application, Academic Press, New York, 1967.

[ARB77] M. A. Arbib, "Parallelism, Slides, Schemas, and Frames," in Systems: Approaches, Theories, Applications (W. E. Hartnett, Ed.), D. Reidel Publishing Co., 1977, pp. 27-43.

[BAD79] N. I. Badler and C. Dane, "The Medial Axis of a Coarse Binary Image Using Boundary Smoothing," Proc. of Pattern Recognition and Image Processing Conference, Chicago, Illinois, August 1979, pp. 286-291.

[BAJ76] R. Bajcsy and M. Tavakoli, "Computer Recognition of Roads from Satellite Pictures," IEEE Transactions on Systems, Man, and Cybernetics, Vol. SMC-6, September 1976, pp. 623-637.

[BAL78] D. H. Ballard, C. M. Brown, and J. A. Feldman, "An Approach to Knowledge-Directed Image Analysis," Computer Vision Systems (A. Hanson and E. Riseman, Eds.), Academic Press, pp. 271-281, 1978.

[BAR76] H. Barrow and J.M. Tenenbaum, "MSYS: A System for Reasoning About Scenes," Technical Note 121, Artificial Intelligence Center, Stanford Research Institute, Menlo Park, CA, April 1976.

[BAR78] H.G. Barrow and J.M. Tenenbaum, "Recovering Intrinsic Scene Characteristics from Images," *Computer Vision Systems* (A. Hanson and E. Riseman, Eds.), Academic Press, pp. 3-26, 1978.

[BUL78] B.L. Bullock, "The Necessity for a Theory of Specialized Vision," *Computer Vision Systems* (A. Hanson and E. Riseman, Eds.), Academic Press, pp. 27-35, 1978.

[CLO71] M.B. Clowes, "On Seeing Things," *Artificial Intelligence*, 2, 79-116, 1971.

[COO67] S.A. Coons, "Surfaces for Computer-Aided Design of Space Forms," MIT Project MAC TR-41, June 1967.

[COO74] S.A. Coons, "Surface Patches and B-Spline Curves," in *Computer Aided Geometric Design* (R.E. Barnhill ad R.F. Risenfeld, Eds.), Academic Press, 1974.

[DAV76] L.S. Davis, "Shape Matching Using Relaxation Techniques," Technical Report 480, Computer Science Center, University of Maryland, College Park, MD, September 1976.

[DUD73] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, John Wiley and Sons, 1973.

[DUD76] R.O. Duda, P.E. Hart, and N.J. Nilsson, "Subjective Bayesian Methods for Rule-Based Inference Systems," *Proc. of the National Computer Conference*, 1976.

[DUD78] R. Duda et al., "Development of the Prospector Consultation System for Mineral Exploration," Annual Report, October, 1978.

[DUD77] S.A. Dudani and A.L. Luk, "Locating Straight-Line Edge Segments on Outdoor Scenes," *Proc. of Conf. on Pattern Recognition and Image Processing*, Troy, NY, June, 1977, pp. 367-377.

[DYE80] C.R. Dyer, A. Rosenfeld, and H. Samet, "Region Representation: Boundary Codes from Quadtrees," *Communications of the ACM*, 3, 1980, pp. 171-179.

[ERM75] L. D. Erman and V. R. Lesser, "A Multi-Level Organization for Problem Solving Using Many Diverse Cooperating Sources of Knowledge," Proc. 4th Inter. Joint Conf. on Artificial Intelligence, Tbilisi, USSR, 1975, pp. 483-490.

[FEL74] J. A. Feldman and Y. Yakimovsky, "Decision Theory and Artificial Intelligence: I. A Semantics-Based Region Analyzer," Artificial Intelligence, Vol. 5, 1974, pp. 349-371.

[FRI69] D. P. Friedman, D. C. Dickson, J. J. Fraser, and T. W. Pratt, "GRASPE 1.5 - A Graph Processor and Its Application," Tech. Report, University of Houston, 1969.

[GIB50] J. J. Gibson, The Perception of the Visual World, Greenwood Press, Westport, CT, 1950.

[GOR74] W. J. Gordon and R. F. Riesenfeld, "B-Spline Curves and Surfaces," in Computer Aided Geometric Design (R. E. Barnhill and R. F. Riesenfeld, Eds.), Academic Press, 1974.

[HAN74] A. R. Hanson and E. M. Riseman, "Preprocessing Cones: A Computational Structure for Scene Analysis," COINS TR 74C-7, Univ. of Mass., Amherst, September 1974.

[HAN78a] A. R. Hanson and E. M. Riseman (Eds.), Computer Vision Systems, Academic Press, 1978.

[HAN78b] A. R. Hanson and E. M. Riseman, "Segmentation of Natural Scenes," in Computer Vision Systems (A. R. Hanson and E. M. Riseman, Eds.), Academic Press, pp. 129-163, 1978.

[HAN78c] A. R. Hanson and E. M. Riseman, "VISIONS: A Computer System for Interpreting Scenes," in Computer Vision Systems (A. R. Hanson and E. M. Riseman, Eds.), Academic Press, pp. 303-333, 1978.

[HAN80a] A. R. Hanson, E. M. Riseman and F. C. Glazer, "Edge Relaxation and Boundary Continuity," in Consistent Labeling Problems in Pattern Recognition (R. M. Haralick, Ed.), Plenum Press, 1980.

[HAN80b] A. R. Hanson and E. M. Riseman, "Processing Cones: A Computational Structure for Image Analysis," in Structured Computer Vision (S. Tanimoto and A. Klinger, Eds.), Academic Press, 1980.

[HAR78] R. Haralick, "Using Perspective Transformations in Scene Analysis," Technical Report, Electrical Engineering Department, University of Kansas, Lawrence, May 1978.

[HAV78] W. S. Havens, "A Procedural Model of Recognition for Machine Perception," TR-78-3, Ph.D. Thesis, Department of Computer Science, University of British Columbia, Vancouver, Canada, 1978.

[HOR75] B. K. P. Horn, "Shape from Shading: A Method for Obtaining the Shape of a Smooth Opaque Object from One View," in *The Psychology of Computer Vision* (P. Winston, Ed.), McGraw-Hill, 1975.

[HOR77] B. K. P. Horn, "Understanding Image Intensities," *Artificial Intelligence*, Vol. 8, No. 2, 1977, pp. 201-231.

[HUF71] D. A. Huffman, "Impossible Objects as Nonsense Sentences," in *Machine Intelligence 6* (B. Meltzer and D. Michie, Eds.), Elsevier, 1971, pp. 295-323.

[KOH79] R. Kohler, "Reference Manual for the VISIONS Low-Level Image Processing System," COINS Dept., Univ. of Mass., Amherst, Spring 1979.

[KON75] K. Konolige, "The ALISP Manual," Univ. Computing Center, Univ. of Mass., August 1975.

[KON78] K. Konolige, SRI Technical Report, in preparation.

[LES77] V. R. Lesser and L. D. Erman, "A Retrospective View of the Hearsay-II Architecture," *Proc. Inter. Joint Conf. on Artificial Intelligence*, Cambridge, MA, 1977, pp. 790-800.

[LEV78] M. D. Levine, "A Knowledge-Based Computer Vision System," in *Computer Vision Systems* (A. R. Hanson and E. M. Riseman, Eds.), Academic Press, 1978, pp. 335-352.

[LOW76] B. T. Lowerre, "The HARPY Speech Recognition Systems," Ph.D. Thesis, Department of Computer Science, Carnegie-Mellon University, Pittsburgh, PA, 1976.

[LOW78] J. D. Lowrance, "GRASPER 1.0 Reference Manual," COINS Technical Report 78-20, University of Massachusetts, Amherst, December 1978.

[LOW80] J. D. Lowrance, "Dependency-Graph Models of Evidential Support," Ph.D. Dissertation, COINS Dept., Univ. of Mass., Amherst, expected June 1980.

[MAC78] A. K. Mackworth, "Vision Research Strategy: Black Magic, Metaphors, Mechanisms, Miniworlds, and Maps," in *Computer Vision Systems* (A. R. Hanson and E. M. Riseman, Eds.), Academic Press, 1978, pp. 53-60.

[MAR76] D. Marr, "Early Processing of Visual Information," Phil. Trans. Roy. Soc. B275, 1976, pp. 483-524.

[MAR77] D. Marr and H.K. Nishihara, "Representation and Recognition of the Spatial Organization of Three-Dimensional Shapes," Proc. Roy. Soc. B 200, 1977, pp. 269-294.

[MAR78] D. Marr, "Representing Visual Information," in Computer Vision Systems (A.R. Hanson and E.M. Riseman, Eds.), Academic Press, 1978, pp. 61-80.

[MIN75] M. Minsky, "A Framework for Representing Knowledge," in The Psychology of Computer Vision (P. Winston, Ed.), McGraw-Hill, 1975, pp. 211-277.

[NAG79] P. Nagin, "Studies in Image Segmentation Algorithms Based on Histogram Clustering and Relaxation," COINS Technical Report 79-15 and Ph.D. Dissertation, Univ. of Mass., Amherst, September 1979.

[NEV76] R. Nevatia, Computer Analysis of Scenes of 3-Dimensional Curved Objects, Birkhauser-Verlag, Basel, Switzerland, 1976.

[NEV77] R. Nevatia and T.O. Binford, "Description and Recognition of Curved Objects," Artificial Intelligence, Vol. 8, 1977, pp. 77-98.

[NEV78] R. Nevatia, "Characterization and Requirements of Computer Vision Systems," in Computer Vision Systems (A.R. Hanson and E.M. Riseman, Eds.), Academic Press, 1978, pp. 81-87.

[NEW65] J.R. Newman, The Universal Encyclopedia of Mathematics, The New American Library, July 1965.

[OVE79] K.J. Overton and T.E. Weymouth, "A Noise Reducing Preprocessing Algorithm," Proc. of Pattern Recognition and Image Processing Conference, Chicago, Illinois, August 1979, pp. 498-507.

[PRA71] T. Pratt and D. Friedman, "A Language Extension for Graph Processing and Its Formal Semantics," Communications of the ACM, 4, 1971.

[PRA79] J. Prager, "Analysis of Static and Dynamic Scenes" Ph.D. Dissertation, COINS Dept., Univ. of Mass., Amherst, March 1979.

[PRA80] J. Prager, "Extracting and Labeling Boundary Segments in Natural Scenes," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. PAMI-2, January 1980, pp. 16-27.

[RIS74] E.M. Riseman and A.R. Hanson, "Design of a Semantically Directed Vision Processor," COINS TR 74C-1, Univ. of Mass., Amherst, January 1974.

[RIS77] E.M. Riseman and M.A. Arbib, "Computational Techniques in the Visual Segmentation of Static Scenes," Computer Graphics and Image Processing, 6, 1977, pp. 221-276.

[ROB65] L.G. Roberts, "Machine Perception of Three-Dimensional Solids," Optical and Electro-Optical Information Processing (J.T. Tippet et al., Eds.), MIT Press, 1965.

[ROS76] A. Rosenfeld, R.A. Hummel and S.W. Zucker, "Scene Labelling by Relaxation Operations," IEEE Trans. Systems, Man, and Cybernetics, 6, 1976, pp. 420-433.

[RUB77] S.M. Rubin and R. Reddy, "The Locus Model of Search and Its Use in Image Interpretation," Proc. of Fifth IJCAI, Cambridge, MA, August 1977.

[SAK76] T. Sakai, T. Kanade, and Y. Ohta, "Model-Based Interpretation of Outdoor Scenes," Third Int. Joint Conf. on Pattern Recognition, Coronado, CA, November 1976, pp. 581-585.

[SAM80] H. Samet, "Region Representation: Quadtrees from Boundary Codes," Communications of the ACM, 3, 1980, pp. 163-170.

[SCH75] R.C. Schank and R. Abelson, "Scripts, Plans, and Knowledge," Proc. of Fourth IJCAI, Tbilisi, 1975, pp. 151-158.

[SCH77] R.C. Schank and R.P. Abelson, Goals, Plans, Scripts and Understanding: An Enquiry into Human Knowledge Structures, Erlbaum Press, NJ, 1977.

[SCH79] R.C. Schank, Interdisciplinary Conference, Jackson, Wyoming, January 1979.

[SHI78] Y. Shirai, "Recognition of Real-World Objects Using Edge Cues," in Computer Vision Systems (A.R. Hanson and E.M. Riseman, Eds.), Academic Press, 1978, pp. 353-362.

[TAN78] S.L. Tanimoto, "Regular Hierarchical Image and Processing Structures in Machine Vision," in Computer Vision Systems (A.R. Hanson and E.M. Riseman, Eds.), Academic Press, 1978, pp. 165-174.

[TAN80] S. Tanimoto and A. Klinger (Eds.), Structured Computer Vision, Academic Press, 1980.

[TEN77] J.M. Tenenbaum and H.G. Barrow, "Experiments in Interpretation-Guided Segmentation," Artificial Intelligence, 8, No. 3, 1977, pp. 241-274.

[TUR74] K.J. Turner, "Computer Perception of Curved Objects Using a Television Camera," Ph.D. Thesis, Dept. of Machine Intelligence, School of Artificial Intelligence, University of Edinburgh, 1974.

[UHR72] L. Uhr, "Layered 'Recognition Cone' Networks That Preprocess, Classify, and Describe," IEEE Trans. Computers, 1972, pp. 758-768.

[UHR78] L. Uhr, "'Recognition Cones,' and Some Test Results; The Imminent Arrival of Well-Structured Parallel-Serial Computers; Positions, and Positons on Positions," in Computer Vision Systems (A.R. Hanson and E.M. Riseman, Eds.), Academic Press, 1978, pp. 363-377.

[WIL77] T. Williams and J. Lowrance, "Model-Building in the VISIONS High Level System," COINS Technical Report 77-1, Univ. of Mass., Amherst, January 1977.

[WIL80] T. Williams, Ph.D. Dissertation (in preparation). COINS Dept., Univ. of Mass., expected June 1980.

[WAL75] D. Waltz, "Understanding Line Drawings of Scenes with Shadows," in The Psychology of Computer Vision (P. Winston, Ed.), McGraw-Hill, 1975, pp. 19-91.

[WOO77] R.J. Woodham, "A Cooperative Algorithm for Determining Surface Orientation from a Single View," Proc. 5th International Joint Conference on Artificial Intelligence, MIT, Cambridge, MA, 1977.

[YAK73] Y. Yakimovsky and J.A. Feldman, "A Semantics-Based Decision Theory Region Analyzer," Proc. IJCAI-3, August 1973, pp. 580-588.

[YOR79] B. York, "A Primer on Splines," COINS TR 79-5, Univ. of Mass., Amherst, Mass., March 1979.

[YOR80] B. York, Ph.D. Dissertation (in preparation), COINS Dept., Univ. of Mass., expected June 1980.

[ZUC77] S.W. Zucker, R.A. Hummel, and A. Rosenfeld, "An Application of Relaxation Labelling to Line and Curve Enhancement," IEEE Transactions on Computers, Vol. C-26, April 1977, pp. 394-403.

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER
COINS TR 80-10

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
EXPERIMENTS IN SCHEMA-DRIVEN INTERPRETATION OF A NATURAL SCENE

5. TYPE OF REPORT & PERIOD COVERED
INTERIM

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Cesare C. Parma
Allen R. Hanson
Edward M. Riseman

8. CONTRACT OR GRANT NUMBER(s)
ONR N00014-75-C-0459

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Computer and Information Science Department
University of Massachusetts
Amherst, Massachusetts 01003

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
Office of Naval Research
Arlington, Virginia 22217

12. REPORT DATE
4/80

13. NUMBER OF PAGES
107

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution of this document is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
scene analysis
image processing
knowledge-based systems
interpretation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The system under development, VISIONS, is an investigation into general issues in the construction of computer vision systems. The goal is to provide an analysis of color images of outdoor scenes, from segmentation (or partitioning) of an image through the final stages of symbolic interpretation of that image. The output of the system is intended to be a symbolic representation of the three-dimensional world depicted in the

20.

two-dimensional image, including the naming of objects, their placement in three-dimensional space, and the ability to predict from this representation the rough appearance of the scene from other points of view. Research in segmentation and interpretation has been separated into the development of two major subsystems with quite different methodologies and considerations.

The focus of this paper is upon the interpretation system. The primary emphasis will be on the development of strategies by which several knowledge sources (KSs) can be integrated using expected knowledge stored in structures called 3D and 2D schemas, each of which may be general or specific to the scene under consideration. A series of increasingly more difficult experiments is outlined as an experimental methodology for developing schema-driven (e.g., top-down) control mechanisms; each succeeding experiment will assume a set of weaker constraints, representing image interpretation tasks where a decreasing amount of knowledge of the situation is available. Experimental results show current capabilities of a number of KSs and the effectiveness of a specific 2D schema in the interpretation of a scene.