A Methodology for the Development of General Knowledge-Based Vision Systems

Edward M. Riseman
Allen R. Hanson

COINS Technical Report 86-27

July 1986

This work has been supported by the Air Force Office of Scientific Research under contract AFOSR-85-0005 and the Defense Mapping Agency under contract 800-85-C-0012.

A METHODOLOGY FOR THE DEVELOPMENT OF GENERAL KNOWLEDGE-BASED VISION SYSTEMS

Edward M. Riseman and Allen R. Hanson

Computer and Information Science Department

University of Massachusetts

Amherst, Massachusetts 01003

ABSTRACT

Expert system technology has been successfully applied to many practical problems, but there has been little evidence of transfer to computer vision. In this paper we discuss some of the problems confronting computer vision, and present an approach to the development of general knowledge-based vision systems. Our approach involves building an intermediate symbolic representation of the image data using knowledge-free segmentation processes. From the intermediate level data, a partial interpretation is constructed by associating an object label with selected groups of the intermediate primitives.

The primary mechanism for generation of initial object hypotheses is a rule-based approach applied to the attributes of the lines, regions, and surfaces in the intermediate symbolic representation. Simple rules are defined as ranges over a feature value which are converted to a vote for an object label; complex rules are constructed as a functional combination of the output from the simple rules. The rules are constructed interactively with visual feedback as part of the knowledge engineering process.

Object hypotheses are used to activate portions of the knowledge network which are responsible for verifying or more completely extracting the hypothesized object. Once ac-

tivated, these procedural components direct additional more expensive extraction of object features, as well as the application of further grouping, splitting and labelling processes at the intermediate level. The goal is the construction of intermediate events which are in closer agreement with the stored symbolic object description. We conclude with some principles which could be used to guide knowledge-based vision research.

TABLE OF CONTENTS

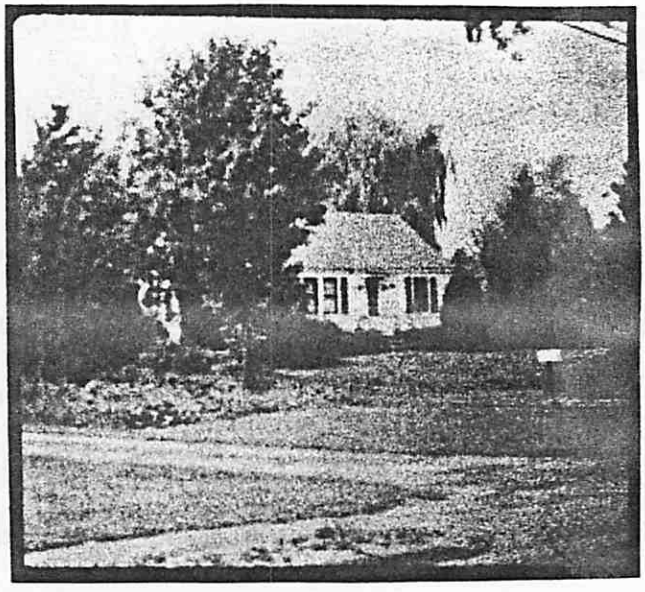
I. INTRODUCTION

Expert system technology, especially techniques for rule-based knowledge engineering, has been successfully applied to many practical problems. Although there are inherent limitations in the complexity and power that can be achieved with traditional expert system approaches [1], particularly when the number of rules are not constrained, there has been little evidence of its application to image interpretation. Typically, vision systems [e.g. 1-3,5,7,11,15,18,21,23,35,28,30,36,37,42] are highly system or application dependent and consequently it has been difficult to transfer them to different task domains. This paper will discuss some of the problems that are specific to computer vision and describe one general methodology for the development of knowledge-based vision systems. The focus of the paper is on the initial iconic to symbolic mapping, which associates portions of the image with hypothesized semantic labels. The proposed approach addresses the start-up problem of interpretation by creating tentative 'islands of reliability' from which context-directed processing can be initiated. Since the initial processes are somewhat independent and are usually associated with separate parts of the images, they can be executed independently and in parallel; multiple processes operating on the same portion of the image can also compete in order to arrive at the best interpretation among a set of alternatives. The relations and expected consistencies between local interpretations form the basis for a cooperative/competitive style of processing among the possible interpretations as the system attempts to extend the islands to uninterpreted parts of the image.

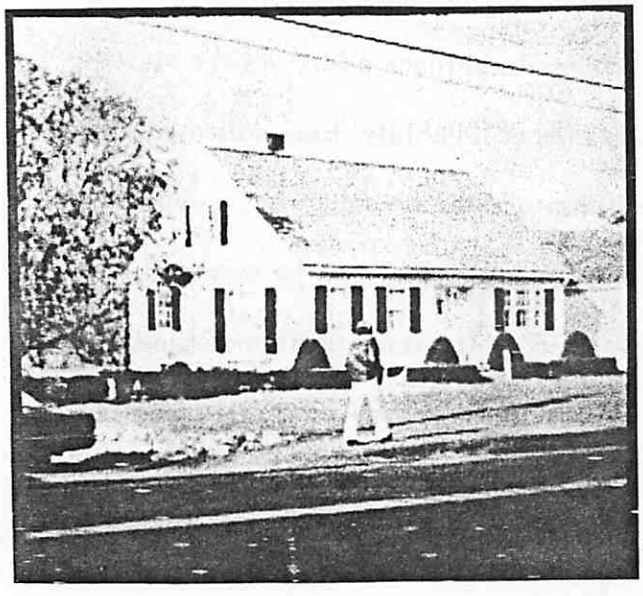
I.1. Complexity of Vision

The complexity of visual tasks can be made explicit by examining almost any complex image. Although this initial discussion is qualitative, we believe the conjectures are

(a)



(b)



(c)

Figure 1. Original images. These images are representative samples from a larger data base. All are digitized to 512 $\times$ 512 spatial resolution, with 8 bits in the red, green, and blue components.

intuitive and reasonable even though it is very difficult to be introspective of one's own visual processing. Humans are rarely aware of any significant degree of ambiguity in local portions of the sensory data, nor are they aware of the degree to which they are employing more global context and stored expectations derived from experience. However, if the visual field is restricted so that only local information is available about an object or object-part, interpretation is often difficult or impossible. Increasing the contextual information so that spatial relations to other objects and object-parts are present makes the perceptual task seem natural and simple. Consider the scenes in Figure 1 and the closeup images in Figure 2. In each case we have selected subimages of objects which show:

a) "primitive" visual elements — these are image events which convey limited information about the decomposition of an object into its parts (which of course is a function at least partly of resolution); note that this implies that the path to recognition of the object via subparts is not available to our perceptual system;

b) absence of context — there is limited information about other objects which might relate to the given object in expected ways; note that this implies that the path to recognition of the object, via the scenes or objects of which it is a part, is not available to our perceptual system.

In Figure 2 as some of the surrounding context of the shoes and the head are supplied, the perceptual ambiguity disappears and the related set of visual elements is easily recognized. In each of the above cases the purely local hypothesis is inherently unreliable and uncertain and there may be little surface information to be derived in a bottom-up manner. It appears that human vision is fundamentally organized to exploit the use of contextual knowledge and expectations in the organization of the visual primitives. However,

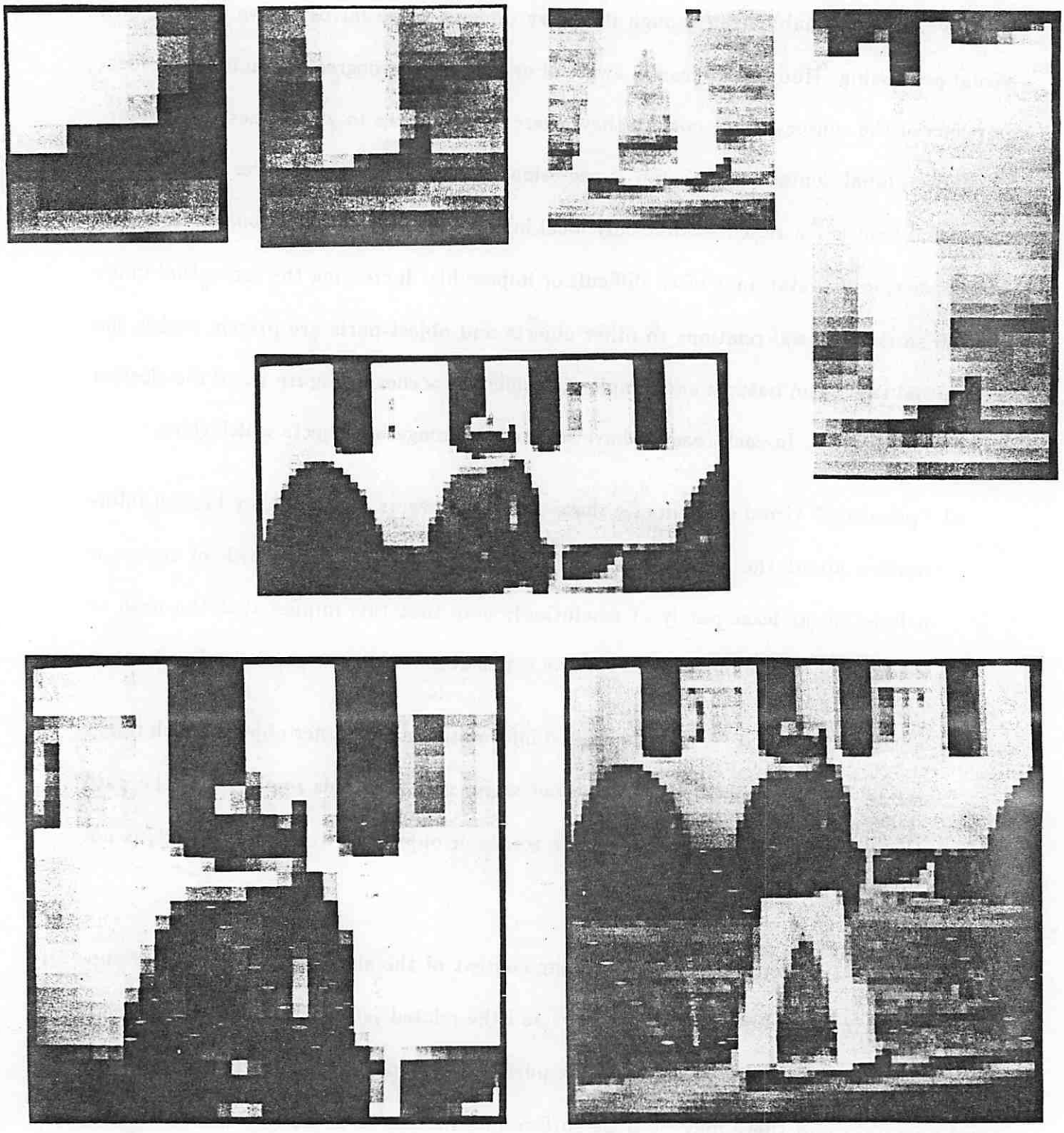


Figure 2. Closeups from original images. In many cases, the identity or function of an object or object part cannot be determined from a small local view. Only when the surrounding context becomes available can the objects be recognized.

it may be impossible to associate object labels with these ambiguous primitives until they are grouped into larger entities and collectively interpreted as a related set of object or scene parts. Thus, the inclusion of knowledge-driven processes at some level in the image interpretation task, where there is still a great degree of ambiguity in the organization of the visual primitives, appears inevitable.

We conjecture that image interpretation initially proceeds by forming an abstract representation of important visual events in the image without knowledge of its contents. The primitive elements forming this representation are then collected, grouped, and refined to bring their collective description into consistency with high-level semantic structures that represent the observer's knowledge about the world.

II. ISSUES FACING KNOWLEDGE-BASED VISION SYSTEMS

The development of knowledge-based vision systems has been hampered by several factors: lack of agreement on what constitutes an adequate representation of image events, lack of low-level processes that can reliably extract relevant image features, lack of satisfactory three-dimensional representations which can capture the inherent variability in the structure of physical objects and scenes, lack of adequate mechanisms for utilizing knowledge during the interpretation process, and finally by the enormous investment in software development that is a necessary precursor to even very simple interpretation experiments. Most of the systems in the current literature only address some of these issues and, perhaps even more discouraging, there do not appear to be ways in which these systems can be easily generalized.

This paper does not attempt to carefully survey the literature in knowledge-based vision systems. However, we do note that most of these systems extract a set of two dimensional image features in some form and then utilize a particular control structure for mapping this information onto concepts in a knowledge base. For example, there have been several rule-based systems which map image features to object identities have been developed [20,25,28], relaxation processes have been exployed for propagating hypotheses under uncertainty [3,11], algebraic constraint manipulation has been used in model matching [7], constraint satisfaction systems have been used to capture relational information [36], and frame-based (or schema-based) approaches for more general (and sometimes more complex) control strategies have been proposed [2,15,37]; see [5] for a survey of some of these approaches. A partial review of image interpretation research can be found in [5,6,14] and descriptions of several individual research efforts are found in [1-3,5,7,11,15,18,21,23,25,28, 30,36,37,42].

Early attempts to interface stored knowledge to image data at the pixel level met with only limited success and little possibility of generalization [36]. For example, blue pixels could immediately be hypothesized to have "sky" labels and appropriate constraints could be propagated, but such an approach to interfacing visual knowledge seems rather futile in the face of increasing numbers of objects and increasing complexity of the task domain. In an image of reasonable resolution there are $512 \times 512 \cong 1/4$ million pixels; hence vision systems must confront the problem of dynamically forming from the large number of individual pixels more useful entities to which propositions will be attached. Transforming the data into a much smaller set of image events is the goal of segmentation processes. However, algorithms for extracting primitives such as 2D regions of homogeneous color and texture, straight lines, simple geometric shapes, and/or local surface patches have proven to be complex and quite unreliable, suggesting that substantial further processing is required before one can expect this intermediate representation to support a globally consistent interpretation.

For a variety of reasons one must expect the data at the level of representation of this first stage of segmentation to be distorted, incomplete, and sometimes meaningless. Segmentation of an image into regions, each of which is composed of a spatially contiguous set of pixels, is a very difficult and ill-formed problem [16]. The sensory data is inherently noisy and ambiguous and this leads to segmentations that are unreliable and vary in uncontrollable ways; for example, regions and lines are often fragmented or merged. In the case of the familiar problem of character recognition, this would be akin to being given joined letters and split letters at a very high frequency. In fact this is one of the major problems in automatic cursive script recognition that makes it a much harder problem than recognition of printed or typed characters. Rather than being concerned only with

the classification of a highly variable set of objects (the cursive characters), the system is also faced with the accompanying problem of organizing the input data into the appropriate segments that form the entities to be classified. Of course general vision is far more complicated than interpretation of handwriting, with a much larger number of more complex objects.

It has been suggested that 2D regions and lines are not appropriate descriptions of the initial image data and that they should be replaced with local estimates of surface orientation, reflectance, depth, and velocity [6,24,36]. In this case the descriptive elements are surface patches which directly capture aspects of the three-dimensional world from which the image was obtained. The implication is that the interpretation task will be far simpler when the surface description is used since it is a representation of the actual physical world that is to be interpreted, and therefore a broader spectrum of domain constraints can be brought to bear upon the information. Although the claim is undoubtedly correct to some degree, reliable extraction of surface range, reflectance, and orientation information from monocular image data has yet to be demonstrated except in highly constrained domains or under very unrealistic constraints on the type of surfaces making up the objects in the scene.

On the other hand, even if a very reliable description of this type could be obtained, the complexity of the natural world will leave us facing many of the same representation, grouping, and interpretation issues. Let us assume for the moment that, in addition to the original spectral information at each pixel, the distance to the corresponding visible surface element at each pixel is also available. Consider the problem of interpreting a complex environment such as a typical crowded city street scene, even if one had such a perfect depth map. How should one partition the information into meaningful entities

such as surfaces, parts of objects, and objects? And then how could this be interfaced to the knowledge base so that control of the interpretation process is feasible? Given that many initial local hypotheses are inherently uncertain and unreliable, how do we achieve globally consistent and reliable integration of the information? This, in fact, is exactly the set of problems that we face the 2D region and line data. We believe that the principles and approaches presented here will be applicable not only to the 2D events extracted from the static monocular color images presented in this paper, but also to the interpretation of 3D depth data recovered from stereo and laser ranging devices, and 2D and 3D motion data derived from a sequence of images.

The problem of forming object hypotheses, or of matching relational models to extracted tokens, is made very difficult by the limitations of low and intermediate level processing. The effect of occlusion leads to the difficult problem of partial pattern matching, where a strong match with part of the pattern is the desired result, as opposed to a weak match of the whole pattern. One must also expect that many region and line samples will not belong to any of the classes because they may be part of shadow regions, portions of occluded objects which cannot be identified, objects that have not been included in the set of object classes in the knowledge base, or object parts which are only identifiable in the context of the object hypothesis. While there has been some success [42] in applying a Bayesian classification viewpoint to these problems, there are many difficulties and we believe standard statistical approaches generally lead to insoluble problems. Classical pattern recognition techniques are not powerful enough by themselves to produce effective classifications in the domains we wish to consider.

Scene interpretation requires processes that construct complex descriptions, where many hypotheses are put forth and a subset that can be verified and which satisfies a

consistent set of relational constraints is accepted. AI systems are often faced with fitting a set of very weak but consistent hypotheses into a more reliable whole. This usually is a complex process that requires great reliance on stored knowledge of the object classes. In such systems, knowledge generally takes the form of object attributes and relations between objects; the relations between parts of objects generally leads to a part-of hierarchical decomposition of the knowledge base.

A related problem involves representing the complexity of the 3D physical world in a form which is useful to the interpretation process [7]. The 3D shape, color, texture, and size of an object class, as well as spatial and functional relations to other objects, often have a great deal of natural variation from object to object and scene to scene. This problem is compounded by the fact that the 2D appearance of these objects in the image are affected by variations in lighting, perspective distortion, point of view, occlusion, highlights, and shadows. These difficulties ensure that the transformation processes for grouping intermediate symbols and matching them to knowledge structures will produce highly unreliable results. The interpretation processes will require general mechanisms for dealing with this uncertainty, detecting errors, and verifying hypotheses.

In summary, there are a variety of issues which must be addressed and resolved before substantial progress in computer vision can be achieved:

a) An effective intermediate symbolic representation must be obtained to serve as the interface between the sensory data and the knowledge base.

b) Knowledge representations must be defined which are capable of capturing the tremendous variability and complexity in the appearance of natural objects and scenes, particularly 3D shape representations.

c) Techniques must be developed for flexibly organizing the intermediate symbols under the guidance of the knowledge base.

d) Mechanisms must be developed for integrating information and data from multiple sources.

e) Inference mechanisms must be available for assessing the indirect implications of the direct evidence.

f) Mechanisms must be developed for coping with the great degree of uncertainty which exists in every stage of data transformation that is part of the interpretation process.

III. OVERVIEW OF THE VISIONS SYSTEM'S APPROACH TO THESE ISSUES

Over the past ten years, the VISIONS group at the University of Massachusetts has been evolving a general system for knowledge-based interpretation of natural scenes, such as house, road, and urban scenes [13,15,16,29,33,34]. The goal of this effort is the construction of a system capable of interpreting natural images of significant complexity by exploiting the redundancies and general constraints expected between and within scene elements.

The general strategy by which the VISIONS system operates is to build an intermediate symbolic representation of the image data using segmentation processes which initially do not make use of any knowledge of specific objects in the domain. From the intermediate level data, a partial interpretation is constructed by associating an object label with selected groups of the intermediate primitives. The object labels are used to activate portions of the knowledge network related to the hypothesized object. Once activated, the procedural components of the knowledge network direct further grouping, splitting and labelling processes at the intermediate level to construct aggregated and refined intermediate events which are in closer agreement with the stored symbolic object description. Figure 3 is an abstraction of the multiple levels of representation and processing in the VISIONS system [15]. Communication between these levels is by no means unidirectional; in most cases, recognition of an object or part of a scene at the high level establishes a strategy for further manipulating the intermediate level primitives within the context provided by the partial interpretation, and for feedback for goal-directed resegmentation. Although the following discussion is based primarily on 2D abstractions of the image data (such as regions and lines), it should be clear that the general ideas extend naturally to 3D abstractions such

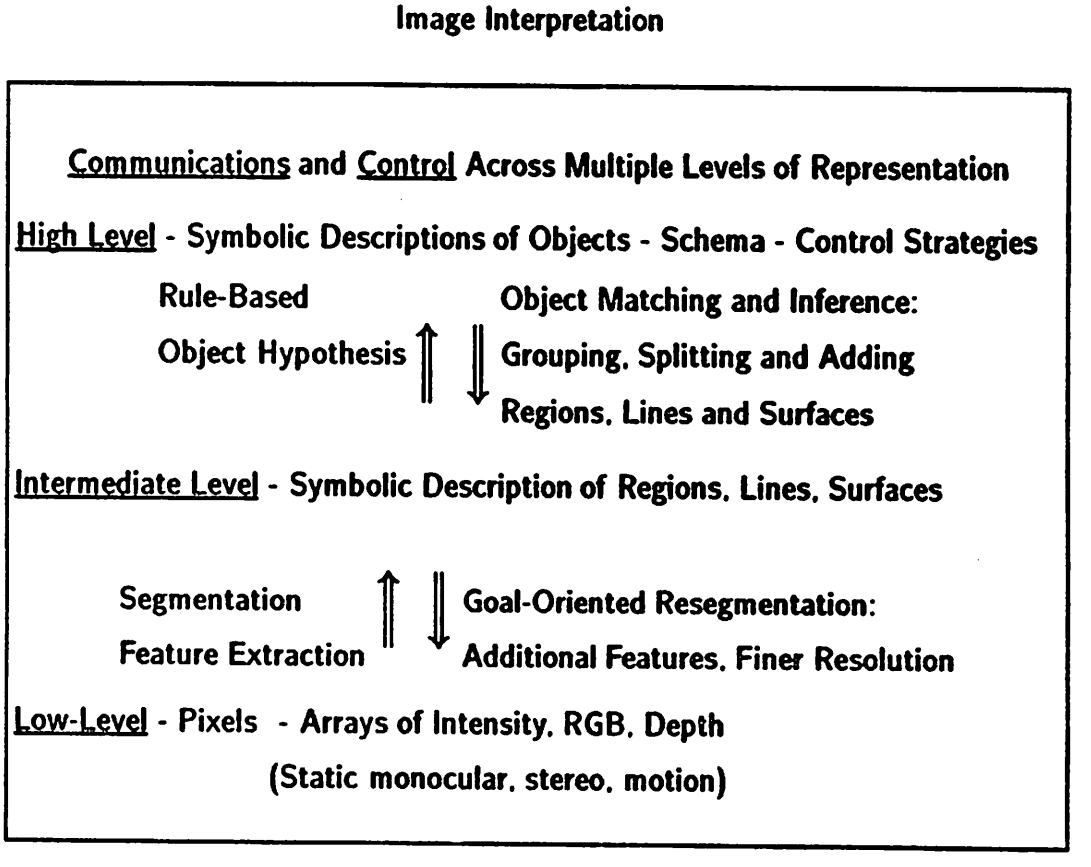


Figure 3. Multiple levels of representation and processing in VISIONS.

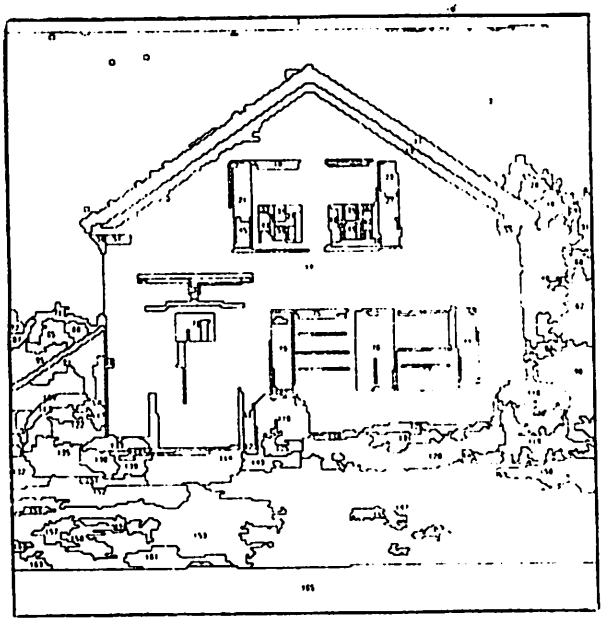
as surfaces as well as to attributes such as motion and depth.

Let us consider some of the stages of processing in a bit more detail.

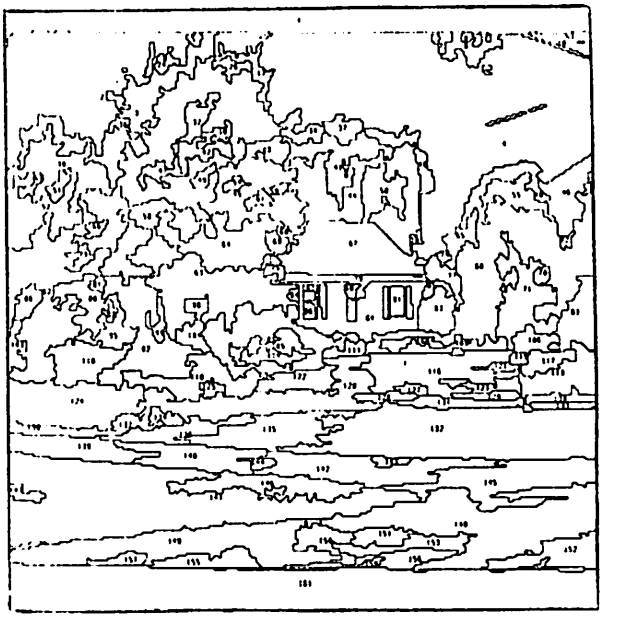
1) Segmentation processes [8,19,26] are applied to the sensory data to form a symbolic representation of regions and lines and their attributes such as color, texture, location, size, shape, orientation, length, etc. Figures 4 and 5 show sample results from two segmentation processes applied to the images in Figure 1. The region and line representations are integrated so that spatially related entities in either can be easily accessed [31]. Two-dimensional motion attributes can also be associated with these entities.

2) Perceptual grouping operations are applied to the low level representations in order to form larger intermediate events, such as straight and curved lines, corners, vertices and T-junctions, parallel lines, various repetitive structures, larger region and surface elements with common properties, and collections of regions, lines, and surfaces satisfying a set of constraints, etc. These operations may be performed bottom-up (data-directed) using general organizational criteria or top-down (knowledge-directed) in a specific context determined by a partially constructed interpretation. Figure 6 shows an example of a bottom-up line grouping process whose goal is to generate long straight lines [38]. An example of top-down grouping is discussed under Item 4 below.

3) Object hypothesis rules are applied to the region and line representation to rank-order candidate object hypotheses [41]; this initial iconic to symbolic mapping provides an effective focus-of-attention mechanism to initiate semantic processing and is described in more detail in Section IV. A simple rule is defined as a range over any



(a)



(b)

(c)

Figure 4. Region segmentations. Regions partition the image into areas which are relatively uniform in some feature (in this case intensity). Mapped into a symbolic structure with a rich set of descriptors, they provide one form of link between the image data and the interpretation system.

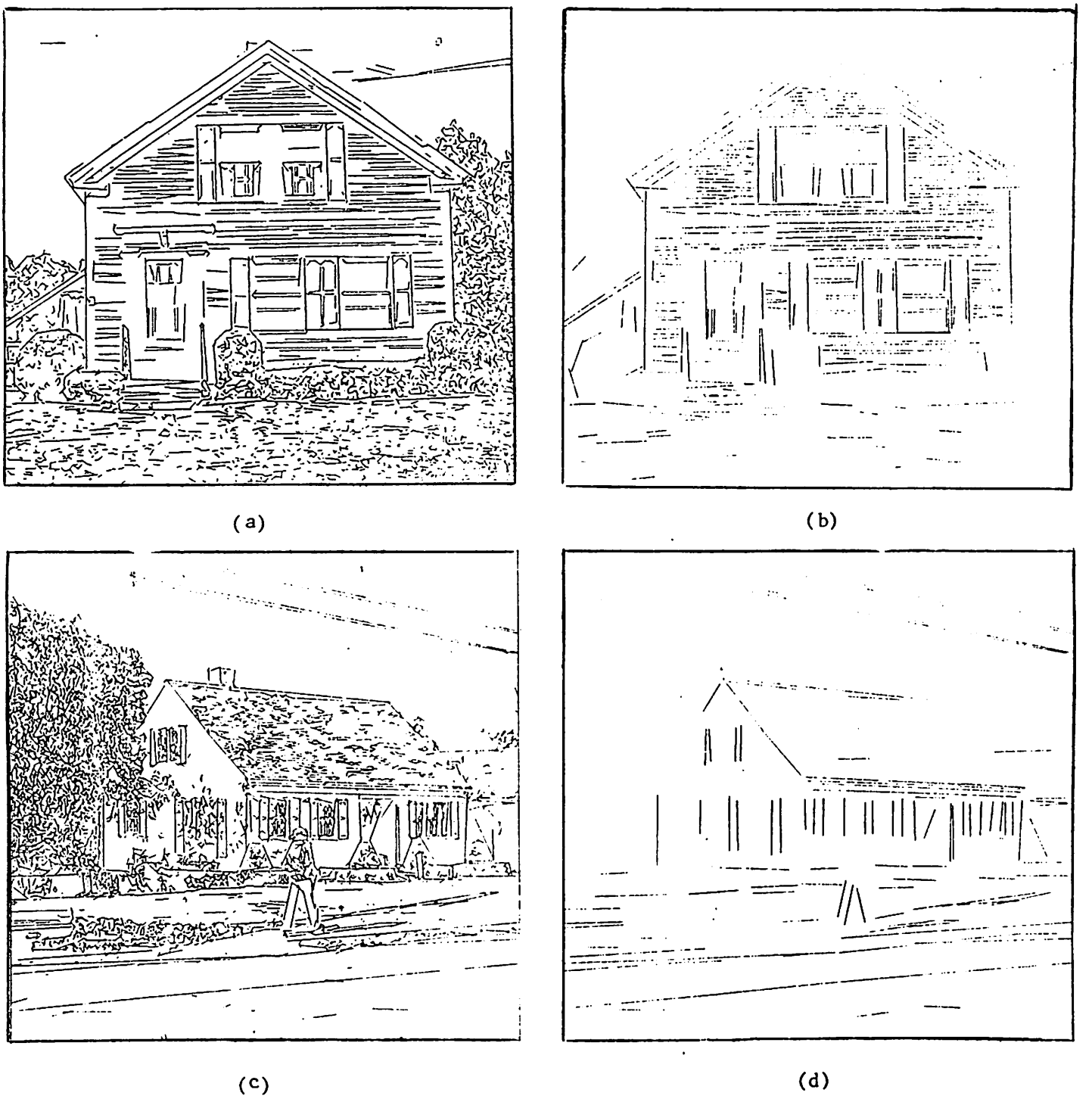


Figure 5. Extraction of straight lines. The straight line algorithm uses a local estimate of gradient orientation to group pixels into regions. A straight line and an associated set of features is extracted from each region. The resulting line image (which contains many lines not shown here) can be filtered in various ways. The two images on the left show all lines whose gradient magnitude exceeds 10 gray levels per pixel; the right images represent a second filtering on the basis of length. The short, high contrast lines in the images in the left images are used as the basis of a texture measure.

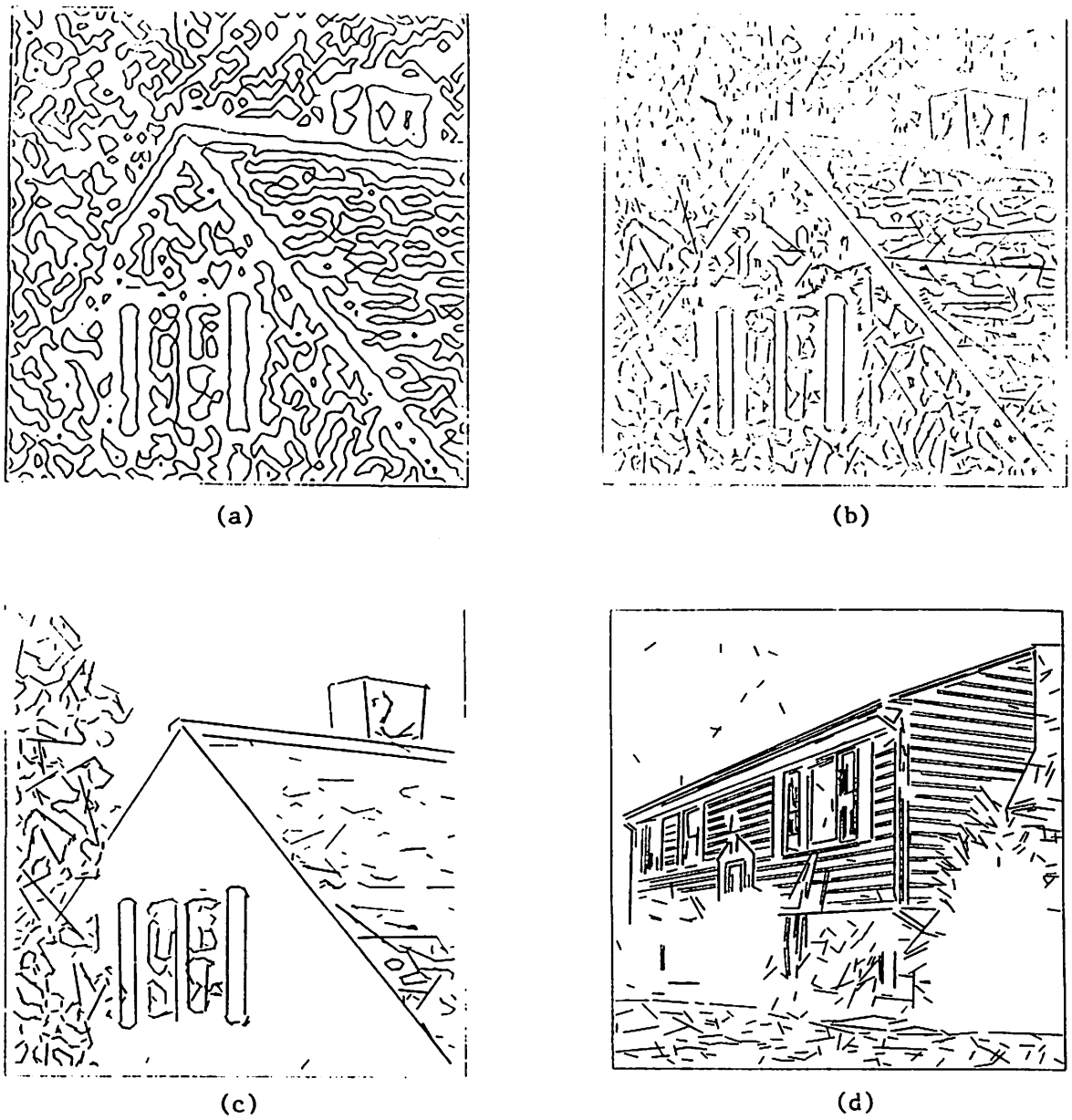


Figure 6. Geometric Line Grouping. The grouping algorithm uses the zero-crossing of the Laplacian to position a straight line of unit length with its center at a zero crossing and with an orientation perpendicular to the gradient measured at the center of the line. An interactive linking and merging processes is then performed to find lines in the next higher level of a hierarchical representation. Almost collinear lines with a similar gradient magnitude that are spatially close are linked and subsequently merged if the straight line approximation to the link candidates is good enough. The resulting representation may be filtered in various ways to retain only those lines satisfying the filtering constraints. (a) zero crossing of the Laplacian (b) unfiltered output of the grouping process (c,d) lines remaining after filtering on length.

scalar feature of the lines or regions. If the attribute of the line/region has a value in the range it will be considered as a "vote" for a particular object label. More complex rules are formed via a logical or arithmetic combining function over several simple rules. The rules can also be viewed as sets of partially redundant features, each of which defines an area of feature space which represents a vote for an object. The region features could include color, texture, size, location in image, simple shape, and motion; line features could include location, length, width, contrast and motion. To the degree that surface patches have been formed, rules can be applied to surface features such as depth, size, location, orientation, reflectance, curvature, and motion. Rules may also be applied to combinations of elements from multiple representations, such as regions and lines, providing a convenient mechanism for fusing multiple representation in a consistent way [4].

4) More complex object-dependent interpretation strategies are represented in a procedural form in knowledge structures called schemas [15,29,40]; these strategies represent control local to a schema node and top-down control over the interpretation process. One interpretation strategy that utilizes the output of the rule set involves the selection of reliable hypotheses as image-specific object exemplars. They are extended to other regions and lines through an object-dependent similarity matching strategy [41]. Thus, as in the HEARSAY paradigm, partial interpretations begin to extend from "islands of reliability" [20]. At this point in the development of the VISIONS system, we are concentrating on the identification and implementation of intermediate grouping strategies for merging and modifying region and line elements to match expected object structures [40]. For this purpose, general knowledge of objects and scenes is organized around the relationships that would be

found in standard 2D views of 3D objects. Verification strategies exploit particular spatial relationships between the hypothesized object and other specific expected object labels or image features. In cases of simple 3D shapes, such as the planar surfaces forming a "house" volume, 3D models and associated processing strategies are employed, and we hope to evolve similar intermediate grouping strategies for complex 3D shape representations in the future.

5) Feedback to the lower-level processes for more detailed segmentation can be requested in cases when interpretation of an area fails, when an expected image feature is not found, or when conflicting interpretations occur. Both the region and line algorithms have parameters for varying the sensitivity and amount of detail in their segmentation output. However, the control of such strategies and the integration of their results is an open problem that is under examination.

6) Due to the inherent ambiguities in both the raw image data and the extracted intermediate representations, a method for handling uncertainty is required if there is to be any possibility of combining this information into a coherent view of the world [14]. Some of the limitations of inferencing using Bayesian probability models maybe overcome using the Dempster-Shafer formalism for evidential reasoning, in which an explicit representation of partial ignorance is provided [35]; we have an ongoing investigation into these issues [22,39] and their potential use. The inferencing model allows "belief" or "confidence" in a proposition to be represented as a range within the [0,1] interval. The lower and upper bounds represent support and plausibility, respectively, of a proposition, while the width of the interval can be interpreted as ignorance.

IV. RULE-BASED OBJECT HYPOTHESIS STRATEGIES

The interpretation task of concern in this paper is that of labelling an initial region segmentation of an image with object (and object part) labels when the image is known to be a member of a restricted class of scenes (e.g., outdoor scenes). An important aspect of this task is the mechanisms for focussing attention on selected areas of the image for which plausible hypotheses of object identities can be generated and for merging regions with a common semantic label. This latter task can occur simultaneously with the labeling process or delayed until a later phase of the interpretation process.

We propose a simple approach to object hypothesis formation, relying on convergent evidence from a variety of measurements and expectations. In the early interpretation phase, when little is known about the scene or its contents, the approach is primarily bottom-up and involves the generation of a few reliable hypotheses about prominent image events. The object hypothesis system thus provides a link between the image data and the knowledge structures. Control can then shift to a more top-down orientation as context and expectations allow the use of further knowledge-dependent processing to validate and extend the initial hypotheses.

Our goal, therefore, is to develop methods for selecting specific image events that are likely candidates for particular object labels, rather than the selection of the best object label for each region and line. For example, given a set of regions in an outdoor scene (and assuming a standard camera position), we might choose to select a few bright blue areas with low texture located near the top of the image as likely "sky" regions. Similarly, in an outdoor scene one could select grass regions by using the expectation that they would be

of medium brightness, have a significant green component, be located somewhere in the lower portion of the image, etc.* For each object, these expectations can be translated into a "rule" which combines the results of many measurements into a confidence level that a region (or small group of regions) represent that object.

IV.1. Knowledge as Rules

Simple rules are defined as the ranges over a scalar-valued feature which will map into a vote for an object label. Typically a feature will be the mean or variance of a property of the pixels or edges composing the regions or lines, respectively. Complex rules involve a combination of simple rules, and they allow fusion of information from a variety of different types of measurements.

We will now develop a simple rule which captures the expectation that grass is green using a feature which is a coarse approximation to a green-magenta opponent color feature by computing the mean of 2G–R–B for all pixels in this region (where R,G,B refers to the red, green, and blue components of the color image, respectively). In order to demonstrate the actual basis and form of knowledge embodied in the rule, in Figure 7 we compare the green-magenta feature distribution of grass pixels to the distribution of the same feature for all pixels. This data was obtained by hand-labelling segmentations from 8 sample images of outdoor house scenes.

The basic idea is to construct a mapping from a measured value of the feature obtained from an image region, say f_I, into a vote for the object. One approach is to define this

---

* Note that a camera model and access to a 3D representation of the environment could dynamically modify the value of these location limits in the image; thus, the use of rules on relative or absolute environmental location in a fully general system would involve modification of expectations about image location as the system orients the camera up or down relative to the ground plane.

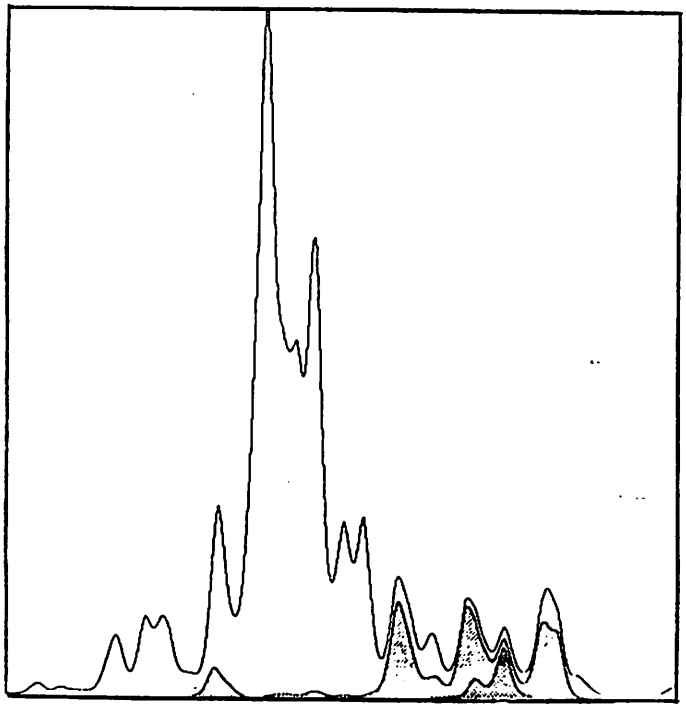


Figure 7. Image histogram of a "green-magenta" feature (2G-R-B). The unshaded histogram represents the global distribution of the feature across all pixels in eight hand-labelled images. The intermediate diagonal shading represents the feature distribution of all grass regions in the eight images. The darkest cross hatched histogram is the feature distribution of grass regions in a single image.

mapping as a function of distance in feature space between the measured value and a stored prototype feature vector which captures the feature properties of the object. Let $d = d(f_P, f) = f - f_P$ be the distance between the measured feature value f and the prototype feature point f_P and let $\theta_1 \leq \theta_2 \leq ... \leq \theta_6$ be thresholds on d (see Figure 8). The response R of the rule is then:

$$R(f) = \begin{cases} 1 & \text{if } -\theta_3 < d \leq \theta_4 \\ \dfrac{d + \theta_2}{\theta_2 - \theta_3} & \text{if } -\theta_2 < d \leq -\theta_3 \\ \dfrac{d - \theta_5}{\theta_4 - \theta_5} & \text{if } \theta_4 < d \leq \theta_5 \\ 0 & \text{if } (-\theta_1 < d \leq -\theta_2) \text{ or } (\theta_5 < d \leq \theta_6) \\ -\infty & \text{otherwise} \end{cases}$$

The thresholds $\theta_i, i = 1, ..., 6$ represent a coarse interpretation of distance measurements in feature space. When the measured and expected values are sufficiently similar, the object label associated with the rule receives a maximum vote of 1. Since small changes in a feature measurement should not dramatically alter the system response, the voting function is linearly ramped to 0 as the distance in feature space increases.

θ_1 and θ_6 allow a "veto" vote if the measured feature value indicates that the object label associated with the prototype point cannot be correct. For example, a certain range of the green-magenta opponent color feature implies a magenta, red, or blue color which should veto the grass label. Thus, certain measurements can exclude object labels; this proves to be a very effective mechanism for filtering the summation of several spurious weak responses. Of course there is the danger of excluding the proper label due to a single feature value, even in the face of strong support from many other features. A

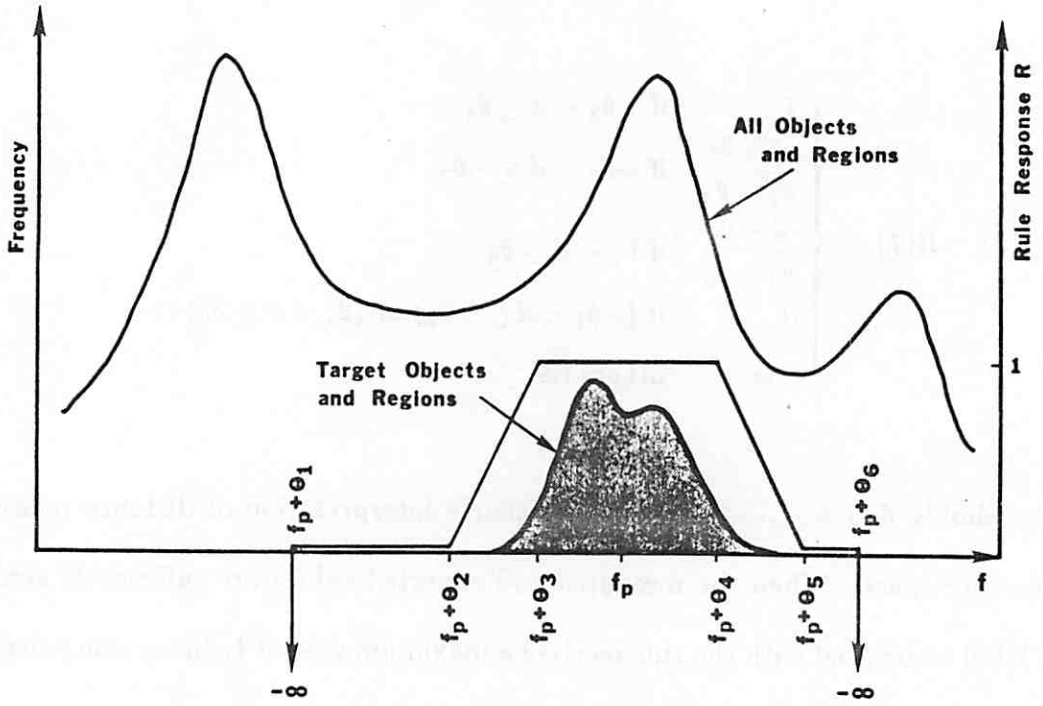


Figure 8. Structure of a simple rule for mapping an image feature measurement f into support for a label hypothesis on the basis of a prototype feature value obtained from the combined histograms of labeled regions across image samples. The object specific mapping is parameterized by seven values, $f_P, \theta_1, ..., \theta_6$ and stored in the knowledge network.

natural extension to the mechanisms presented here would generalize the rule form to be parameterically varied from the fixed form that we have defined. Thus, the ranges could be dynamically varied so that fewer or larger numbers of regions are in the positive voting range of a particular rule. If there are multiple peaks in the histogram, then a simple rule can be defined for each peak independently. Their results can be combined using a simple function, such as the maximum of the responses.

A simple rule is a specification of a constraint on the value of a feature which should be satisfied if the object is present. A complex rule is defined as a (partially redundant) set of simple features that is assembled into a composite rule via a combining function which can take any logical or arithmetic form; this is an extension of the functional form of hypothesis rules in Nagao et al [25]. The premise is that by combining many partially redundant rules, the effect of any single unreliable rule is reduced.

It is useful to impose a hierarchical structure on the set of simple rules. In this paper the rule for each object is organized into a composite rule of 5 components which provide a match of color, texture, location, shape, and size of the object. This allows some flexibility in combining several highly redundant features (e.g., several color features) into a composite rule which is somewhat more independent of the other composite rules (e.g., color vs. location). It should be recognized, however, that this is only one alternative for imposing a hierarchical structure on the set of simple rules; many other combination functions are possible and the choice of the function determines how the features 'cooperate' to provide an initial hypothesis.

Each of the five composite rules is in turn joined into a composite rule. Any rule might have a weight of 0, which means that the rule will have no effect on the weighted

response of the composite rule except that the veto range of the rule can reject a region as a candidate for the object in question. The structure of the composite rule for grass is shown in Figure 9; it consists of a normalized weighted average of the five components C_j:

$$grass\ score = \frac{1}{N}\sum_{j=1}^{5} W_j C_j$$

where the W_j are the weights and $N = \sum_{j=1}^{5} W_j$. Each of the components is in turn a weighted sum of a set of individual rules:

$$C_j = \frac{1}{M}\sum_{k} V_k R(f_k)$$

where $R(f_k)$ is the response from an individual feature rule based on feature f_k and the V_k are the weights $(M = \sum_k V_k)$. Similar rules were developed for sky and foliage.

The weights shown in Figure 9 capture the heuristic importance of each of the contributions to the rule. The weights are integers from 0 to 5, and reflect a belief that only a few levels of relative importance are needed ("*weak*" $\equiv 1$, "*medium*" $\equiv 3$, "*strong*" $\equiv 5$ in importance). The intention is to avoid twiddling of numbers, but to allow obvious relative weightings to be expressed. Since the composite rule response is used only to order the regions on the basis of their similarity to the stored feature templates, rather than classifying them as an instance of a specific class, the expectation is that the rule response is relatively insensitive to small changes in the weights. The rules extend easily and naturally to include other features and attributes, such as depth and motion.

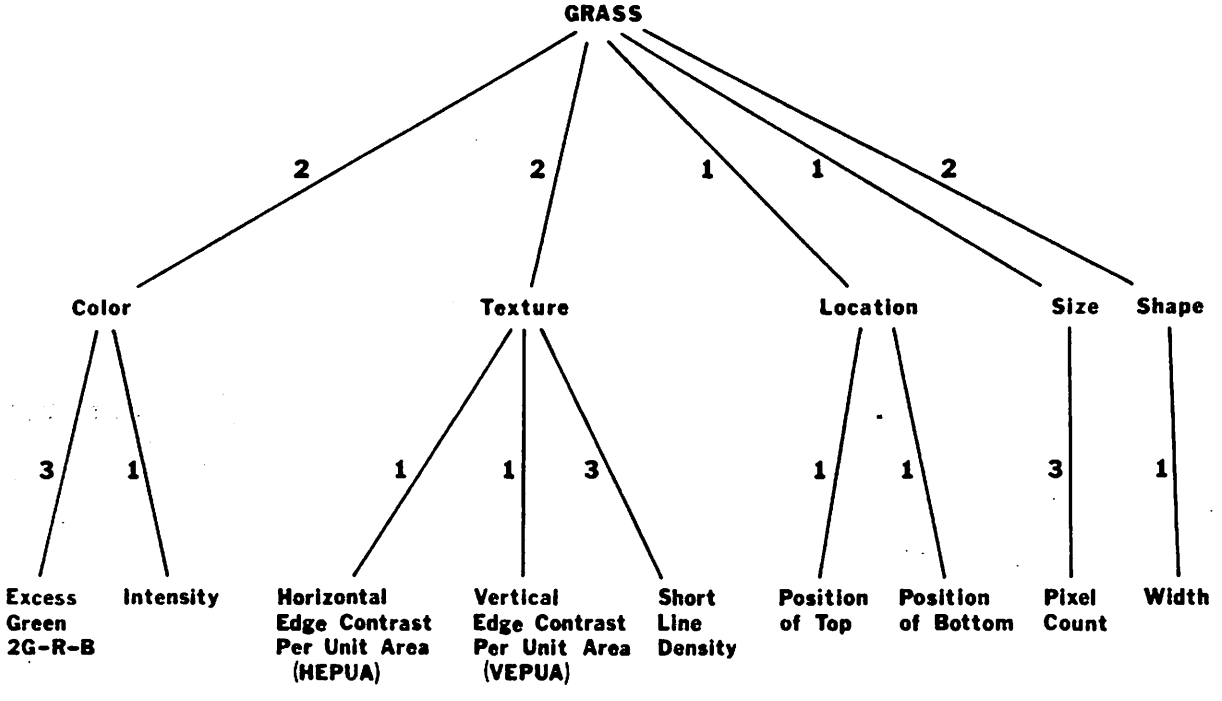


Figure 9. Structure of the grass rule. The rule response is the normalized weighted sum of the resonses of five component rules, each of which is in turn a normalized weighted sum of the responses from simple rules associated with a single feature. Note that a weight could be 0, thereby allowing only the veto range for that feature to be propagated.

IV.2 Relationship to Bayesian Pattern Classification Theory

It is instructive to briefly consider the relationship of the rules to the Bayesian techniques commonly used in pattern recognition, since the goals of both techniques appear to be similar. In the classical character recognition problem the jth character R_j is to be classified as one of a fixed set of classes C_i, i = 1,...,N on the basis of a feature vector $\bar{X}_j$ extracted via measurements on character R_j. One can view the region labelling problem to be equivalent, in that a set of feature measurements can be extracted from all regions, and then each can be classified according to a maximum likelihood decision process.

A training set of characters is usually provided a priori, from which statistical estimates of feature distributions can be extracted; it is necessary that the training set be large enough to capture the expected variations in the domain. The optimal decision process for a given character R_j involves the computation of the a posteriori Bayesian probability for each class given the feature vector $\bar{X}_j$, followed by the selection of the maximum likelihood class as the output decision. Thus, using Bayes rule character R_j is classified as C_i, where

$$\underset{i}{MAX}\ P(C_i \mid \bar{X}_j) = \underset{i}{MAX}\ \frac{P(\bar{X}_j \mid C_i)\ P(C_i)}{P(\bar{X}_j)}$$

Since $P(\bar{X}_j)$ is constant across the N classes, it cannot affect the decision and may be ignored. Thus, the decision rule is now simplified to

$$\underset{i}{MAX}\ P(X_j \mid C_i)\ P(C_i)$$

with $P(\bar{X}_j \mid C_i)$ estimated from the training set and $P(C_i)$ obtained via statistical analysis of the task domain.*

In the pattern recognition/classification case, the set of classes is known, fixed, and usually is not large. Also note that $P(\bar{X}_j)$ could be factored out of the computation of the posterior class likelihoods because the set of feature measurements was the same for each class; i.e., one set of measurements was performed and these results were used to determine all of the $P(\bar{X}_j \mid C_i)$. However, the a priori class probability $P(C_i)$ does vary in the computation of each class likelihood. Finally, the intent of this classification process is the classification of every character sample.

Let us briefly make several points about why the pattern recognition (PR) paradigm of classification is not effective. It assumes a fixed set of known classes that ususally is not large. The samples to be classified are assumed to be directly presented, or to be extractable in a relatively straightforward fashion; in particular there is little difficulty in figure-ground separation of the sample. Finally, the samples are usually assumed to be complete (i.e. no occlusion or missing portions) so that one can avoid the difficult problem of partial matching of portions of the object.

In Section II we discussed the difficulty of image interpretation, and how it relates to these assumptions. Classical pattern recognition techniques are not powerful enough by themselves to produce effective classifications in the domains we wish to consider. The

---

* There are a number of variations to the basic paradigm which we note here, but do not wish to explore in this treatment. Some samples could be rejected if the maximum likelihood is sufficiently low; by avoiding classification of a difficult subset of characters the error rate might be reduced, but more importantly there is the possibility of focussing attention on samples where additional information would be valuable. Another extension involves the dependencies between characters that are a function of the characteristics of the language; contextual analysis via conditional probabilities of letter sequences could be used to improve the estimates of the likelihoods [12,17,32].

general problem of image interpretation involves a possibly large number of classes. The unreliability of low level processes such as line extraction, region segmentation and surface fitting imply that one cannot be sure that any extracted intermediate entity is relevant to any of the classes.

In contrast, we have modified the classification strategy to become an AI "focus-of-attention" process since it is not feasible to initially classify all image events for the reasons discussed in the previous sections. The organization of the input data is not sufficiently well-defined to pose the classical pattern recognition goals in our domain. Thus, rather than the selection of the best object label for each region and line, we are looking for good region and line candidates for a particular object label.

From a Bayesian view, instead of the measurement vector $P(\bar{X}_j)$ being held constant across samples, the a priori class probability $P(C_i)$ is constant across regions to be classified. While there is a common set of features measured on each region, the measurement vector $\bar{X}_j$ may be different for each (i.e. a different set of feature values). This changes the optimal decision rule via a Bayes formulation to

$$\underset{j}{MAX} \frac{P(\bar{X}_j \mid C_i)}{P(\bar{X}_j)}$$

which will decompose into the product of individual feature terms under an assumption of independence.

It should now be clear that the simple piecewise-linear rule form is more than just an approximation to $P(X_j \mid C_i), j = 1, ..., K$. What it also must balance is the relation to $P(X_j)$, which appears in the denominator of the Bayesian ratio. This term is important because it brings in the degree of discrimination of each feature measurement X_j for

class C_i. For example, there would be little value in a feature which exhibited a very tight range (i.e., very low variance) for some object class, if in fact it also exhibited the same distribution for all classes. In actual practice, the vision system designer/knowledge engineer is responsible for the selection of the features used in the object rules. To the degree that a rule developed by this expert covered $P(X_j \mid C_i)$ and excluded $P(X_j \mid C_k), for\ all\ k \neq i$, that rule would be an optimal discriminator. Of course there is still the problem of the usually invalid assumption of feature independence, and therefore our heuristic hierarchical combination of features may be just as reasonable. In fact the use of the veto range for individual features has the same effect as a ratio of zero in the product of probabilities under the assumption of independence.

In summary the PR issues are dealt with in the AI focus-of-attention paradigm by selecting only hypotheses which are more reliable and not attempting to classify all entities. These hypotheses can then be verified via additional processes which make use of contextual knowledge.

IV.3. Results of Rule Application

Figure 10 shows the results of applying selected simple rules for grass to the region segmentation from Figure 4c. For each rule there are two images. The left image of each pair is a composite feature histogram showing the feature distribution across all pixels (the unshaded curve) in a set of images, and the distribution for grass pixels (the cross hatched curve) across the same set of images. The histograms were computed from a set of eight hand-labelled images and smoothed. The right image of each pair shows the strength of the rule response for each region coded as a brightness level: bright regions correspond to high rule output.

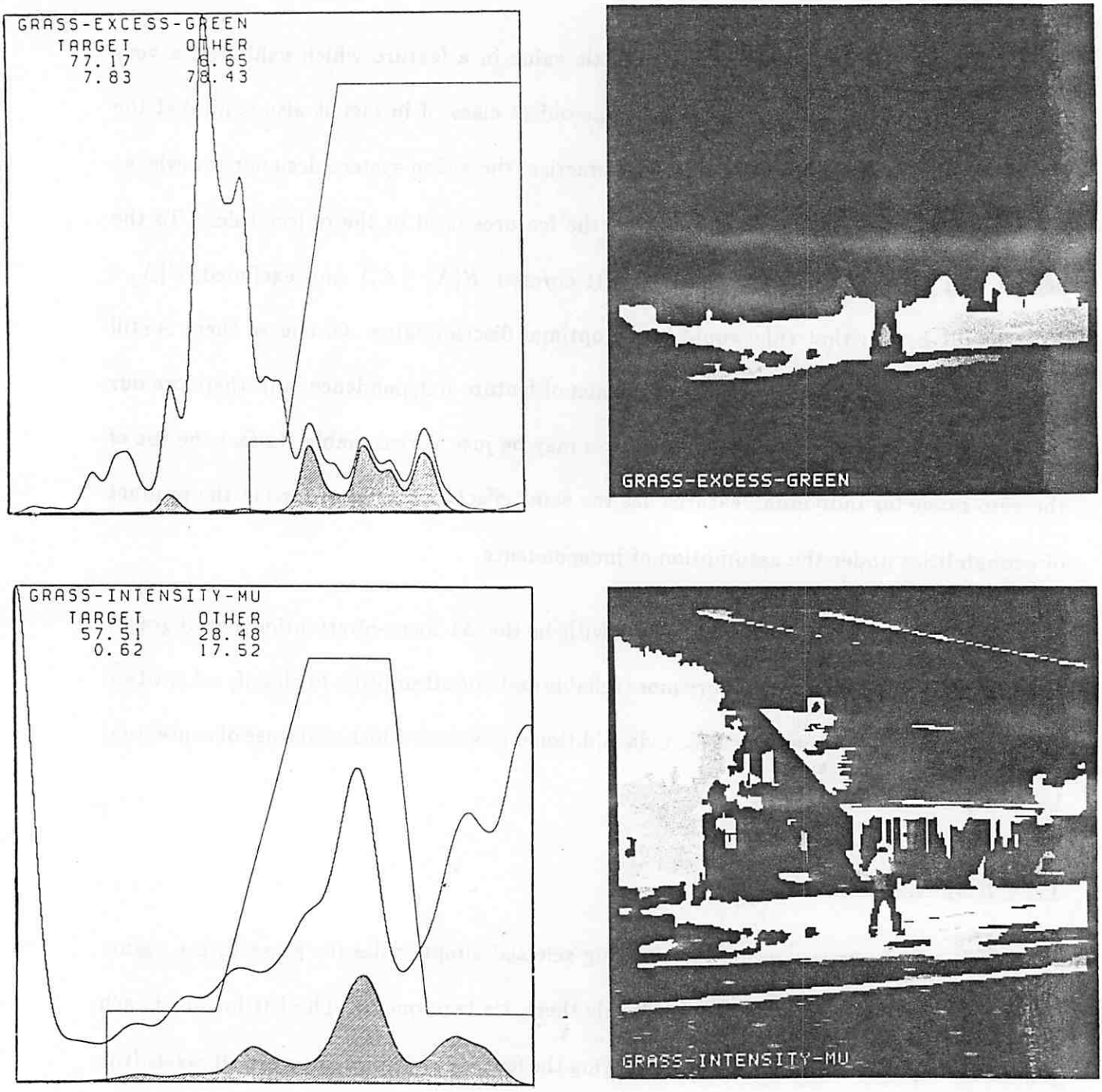
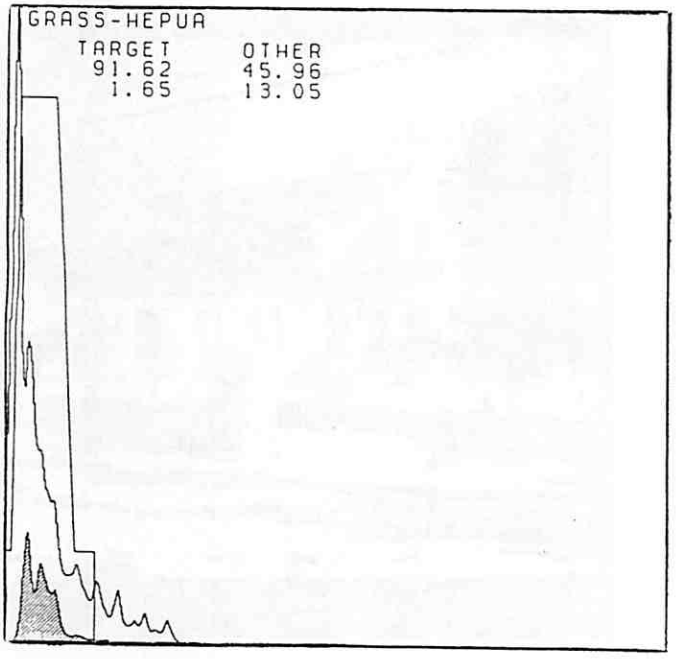
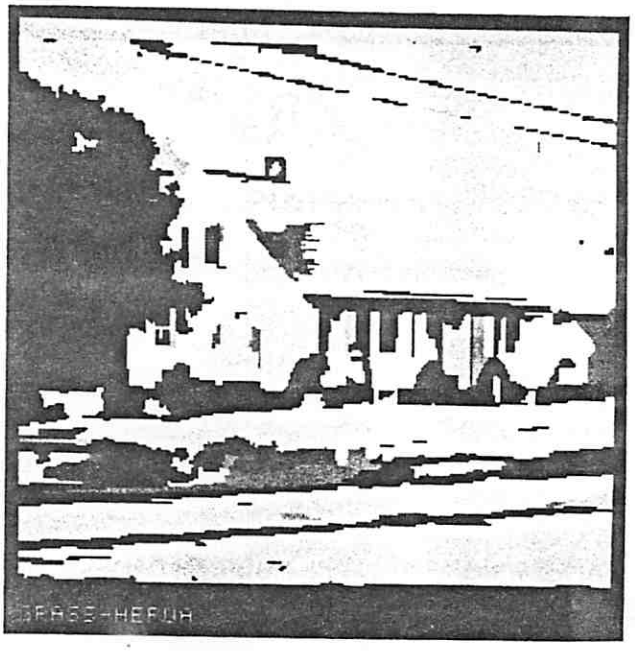
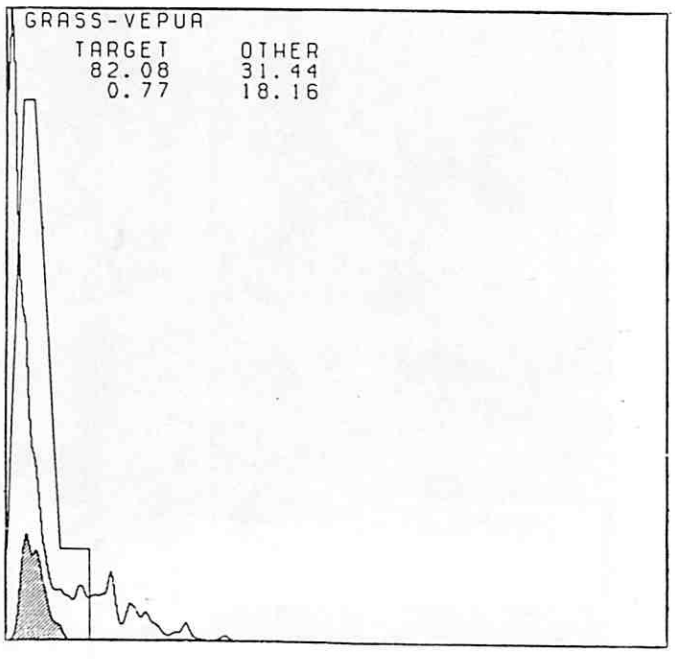


Figure 10. Results from the simple grass rules. In each image pair, the left image is a composite feature histogram showing the feature distribution across all pixels in a set of images, the distribution of grass pixels in the same set, and the rule. The right image shows the brightness encoded strength of the rule response when aplied to all regions in the segmentation of Figure 4c; bright regions correspond to a high rule response. See text for a discussion of the four numbers in the upper left corner of the histograms.







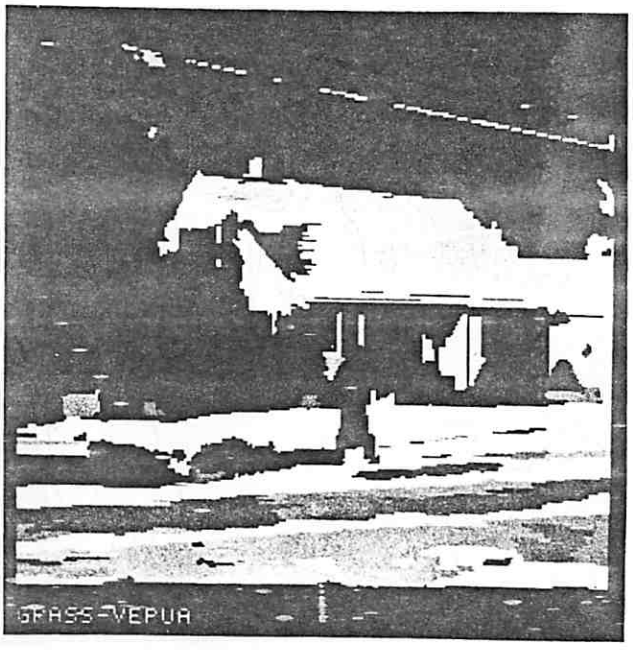
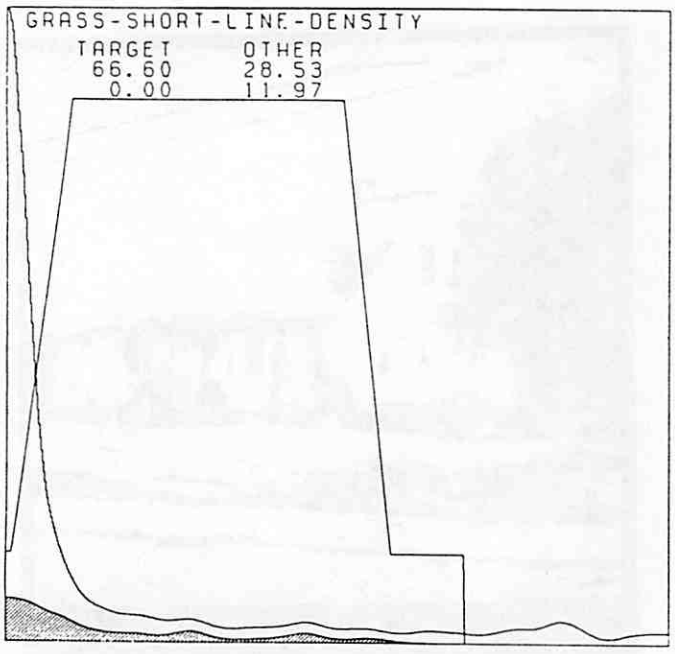


Figure 10, continued



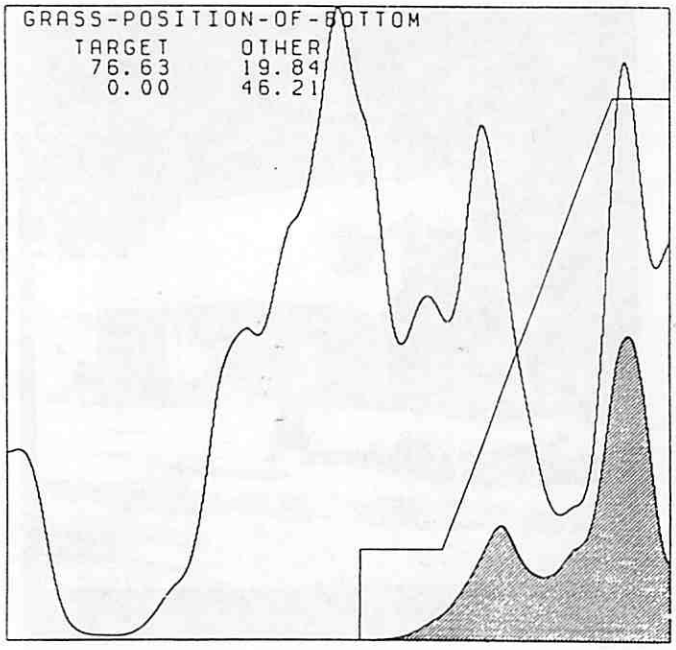
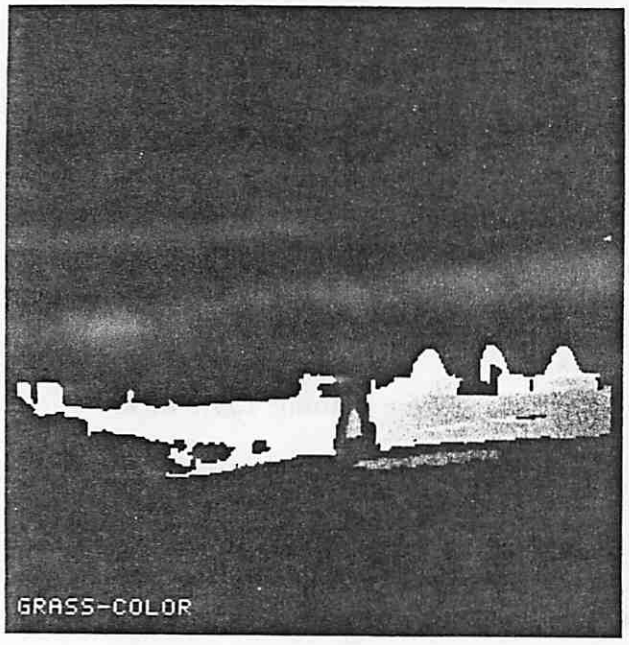


Figure 10, continued

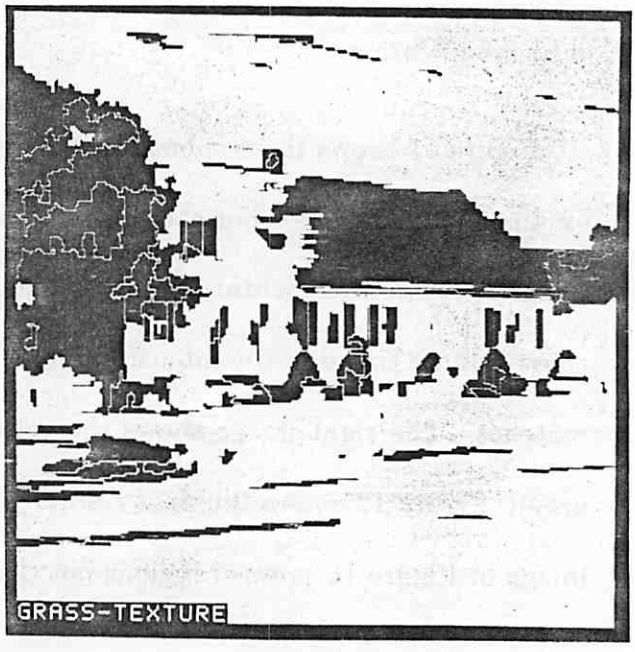
The rule that was developed interactively by the user is superimposed on the histograms in piecewise linear form. In the upper left "Target" refers to the object associated with the rule, in this case grass, while "Other" refers to all objects other than the target object. The first row of numbers shows the weighted average response of grass regions and other regions to the rule function (100 is maximum), while the lower numbers tabulate the percentage of target regions and other regions vetoed. Thus, the ideal rule is one which responds maximally with a value of 100 to the target regions, while vetoing 100% of all other regions. In practice, there is almost always a tradeoff and optimal settings are not at all obvious. In some cases rules for the target object were set to exclude regions associated with other objects, while in other cases the goal was to maximize the response for the target object regions. There is no intent here to put forth these specific rules as a significant contribution or even as a satisfactory set; in fact some of these rules probably need modification.

Figure 11 shows the response for three of the five rule components and the final result for the composite rule. For each rule the region response is shown superimposed over the image in two complementary formats. The left image of each pair shows the strength of the rule response coded in the intensity level of each region; bright regions correspond to good matches. The right image shows the vetoed regions in black (with all others uniformly grey). Figure 12 shows the final results for the foliage, grass, and sky rules in the house image in Figure 1b (vetoed regions not shown).

The effectiveness of the rules can be seen by examining the rank orderings of the regions on the basis of the composite rule responses. For the grass results shown in Figure 11, for example, the two top ranked regions are actually grass. For the grass results shown in Figure 12, the top six regions are grass and 8 of the top 10 regions are grass; the two

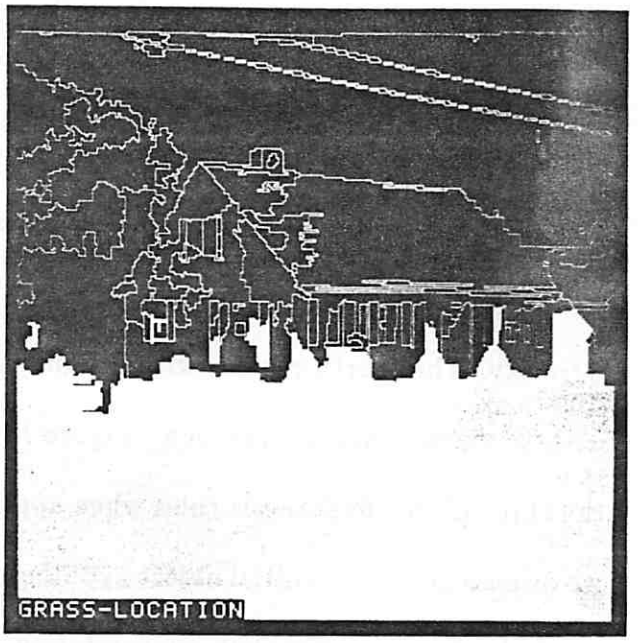


(a)

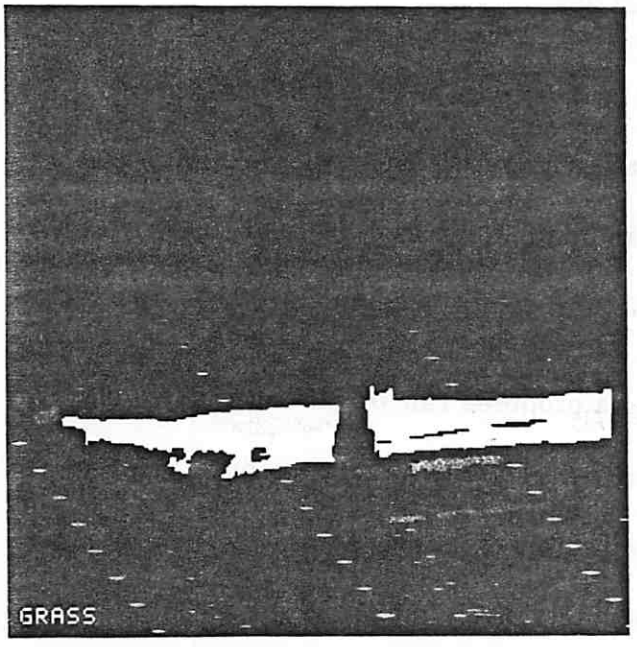


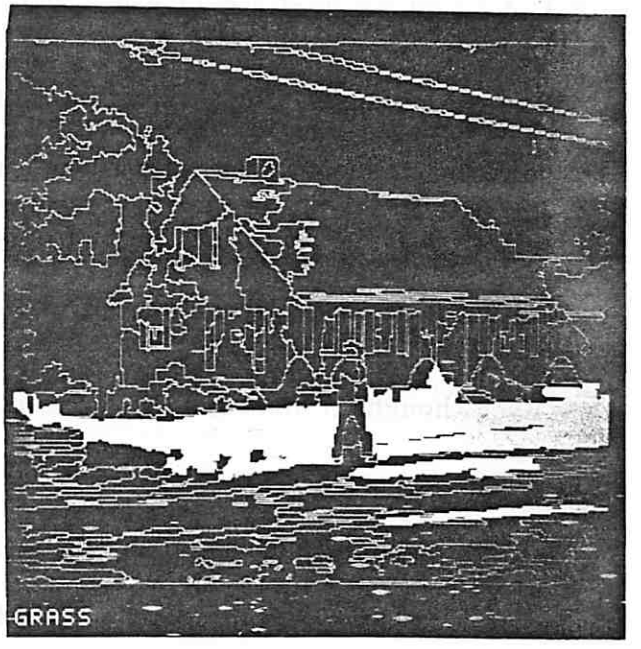
(b)

Figure 11. Rule responses for the grass component rules and the final composite rule. In each pair of images, the left image shows the brightness encoded (bright $\equiv$ high) rule response. The right image shows regions vetoed by the rule in black; all others (non-vetoed regions) are uniformly gray. (a) color component rule; (b) texture component rule; (c) location component rule; (d) final result from grass composite rule.



(c)





(d)

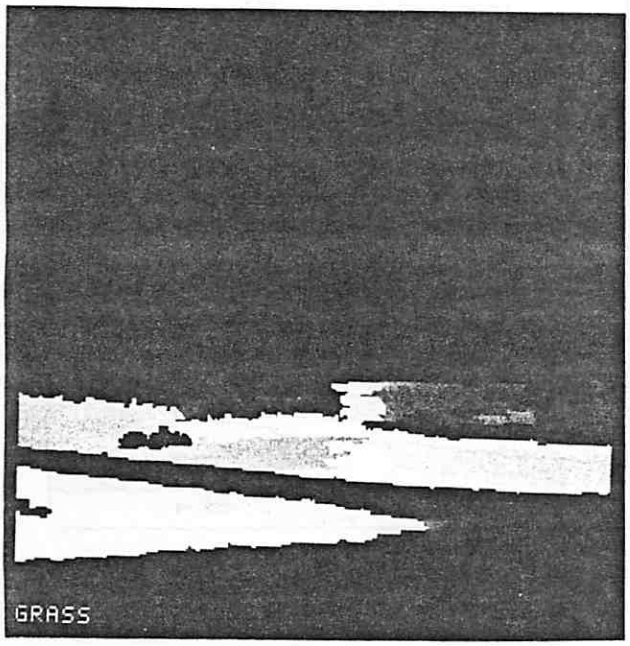
Figure 11, continued

non-grass regions were actually sidewalk and driveway. For the foliage responses shown in Figure 12b, the top 21 regions were some form of foliage (tree, bush, or undergrowth); of the 30 regions not vetoed, there were only 7 non-foliage regions. These 7 regions were actually grass and were among the lowest ranked of the non-vetoed regions (7 of the last 9). For the sky results in Figure 11c, only four regions were not vetoed and the top three were sky. The fourth region, with a significantly lower rule response, was actually foliage with some sky showing through. Figure 13 shows the highest ranked regions for each of the three object hypothesis rules when applied to the three example images. In Section V we discuss how these initial object hypothesis results may be used as the basis of a strategy to produce a more complete interpretation.

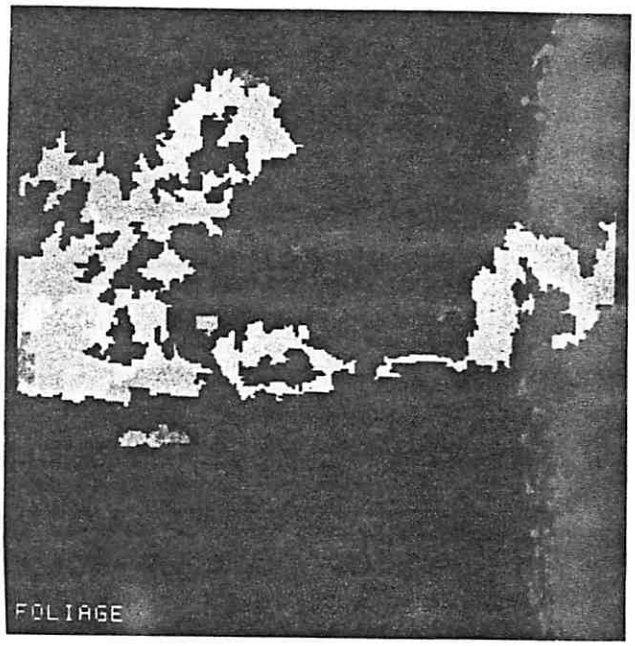
IV.4 A Language Interface for Knowledge Engineering

Knowledge engineering of rules can be greatly facilitated by an interactive environment for rule construction. A user can get an immediate sense of the effectiveness of proposed rules by displaying the rating of each symbolic candidate in intensity or color. Thus, rule development becomes a dynamic process with a natural display medium for user feedback.

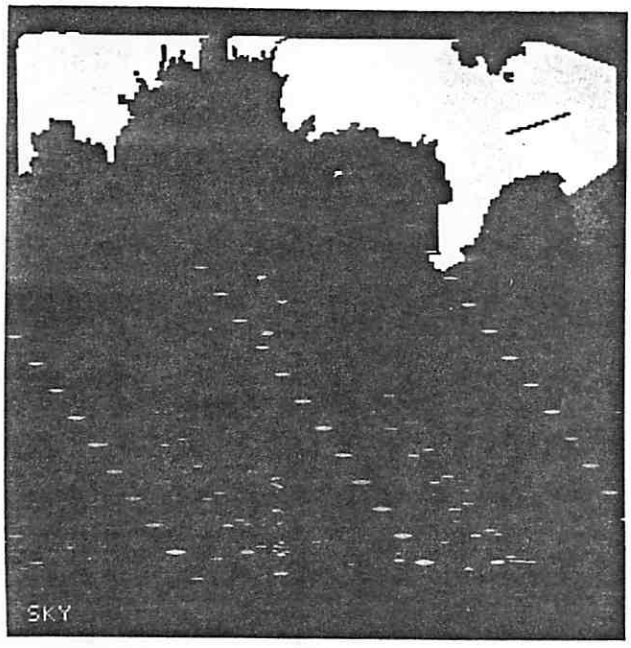
Even though an immediate visual response to a proposed rule or rule set is available, the knowledge engineer must not be forced into a "parameter twiddling" mode. The rules should be robust enough so that a fairly crude specification of the rule parameters generates reasonable results. The specification can then be interactively refined, if necessary. As a first step toward an interactive specification facility, a simple language interface has been constructed so that rules can be specified on any feature in terms of five intervals of the dynamic range of a feature – "very low", "low", "medium", "high", "very high" [33,41]. These labels induce a partition on the range of the feature; for each interval, the user



(a)

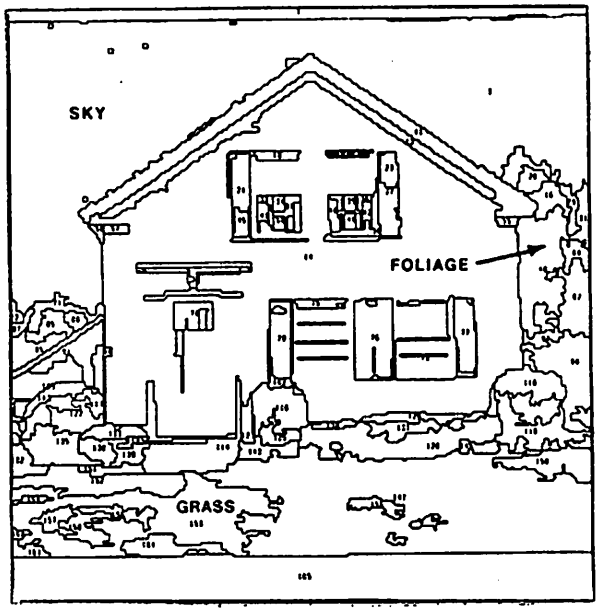


(b)

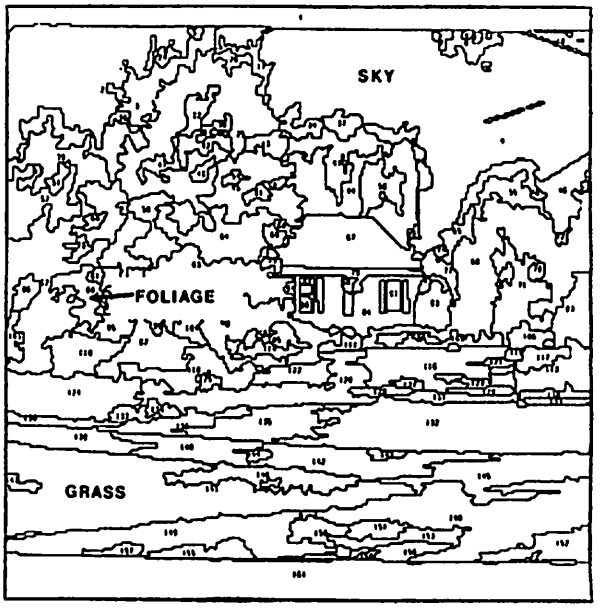


(c)

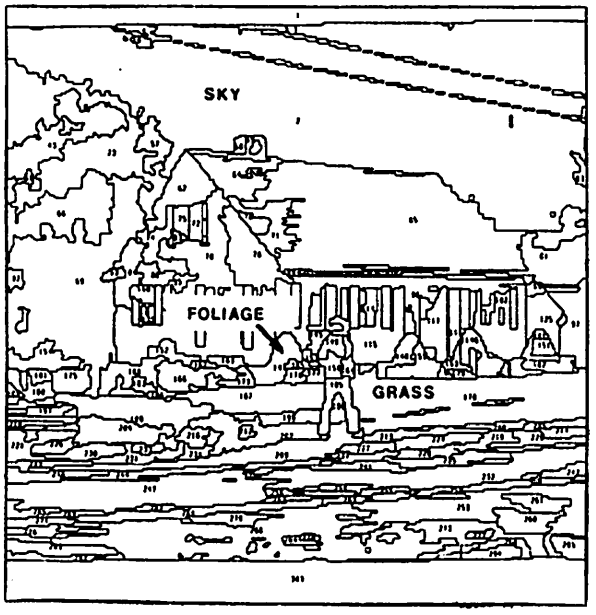
Figure 12. Composite rule responses for grass, foliage, and sky rules, applied to Figure 4b encoded in brightness. (a) grass, (b) foliage, (c) sky.



(a)



(b)



(c)

Figure 13. Highest ranked object hypothesis for the three examples images.

specifies whether the rule response is "ON", "OFF", or "VETO".

The results obtained from the coarsely quantized rules are quite good and are comparable to the results obtained in the previous section using the more carefully defined rules. Typical results from this rule set are shown in Figure 14 using the foliage hypothesis rule on two of the test image segmentations (Figure 4a,c); these results are comparable to those shown in Figure 11. For the segmentation of the house image of Figure 4a, a total of 24 regions survived the vetos. Of these, only two were incorrect (window and grass). For the image in Figure 1b (results not shown), 16 regions were not vetoed and all of them were some form of foliage (tree, bush, tall grass, or undergrowth); only one grass region was included. 28 regions remained after running the rule on the regions making up the house scene in Figure 1c. Of these 28, 20 were bush or tree, 2 were grass, and the remainder were rocks, house window or shadowed areas. All of the incorrect regions were rated fairly low by the rule. In all cases, the most highly rated regions were foliage.

Similar results were obtained for the sky and grass rules. For the image in Figure 1b, for example, 9 regions were not vetoed and of these, 5 were sky, 2 were a mixture of sky and the telephone or power wire (see Figure 4b) and two were a mixture of tree and sky. Applying the grass rule to the segmentation of Figure 1a results in 41 regions, 17 of which are grass and which are among the 18 top rated regions (the other region was bush). The remaining 23 regions, all rated very low, are bushes, trees, windows, house steps, and shutters.

Coarse quantization of the feature range offers the knowledge engineer the opportunity to quickly develop and assess a rule set without detailed examination of feature statistics. It may be possible, in some cases, to completely develop the rule base using semantic

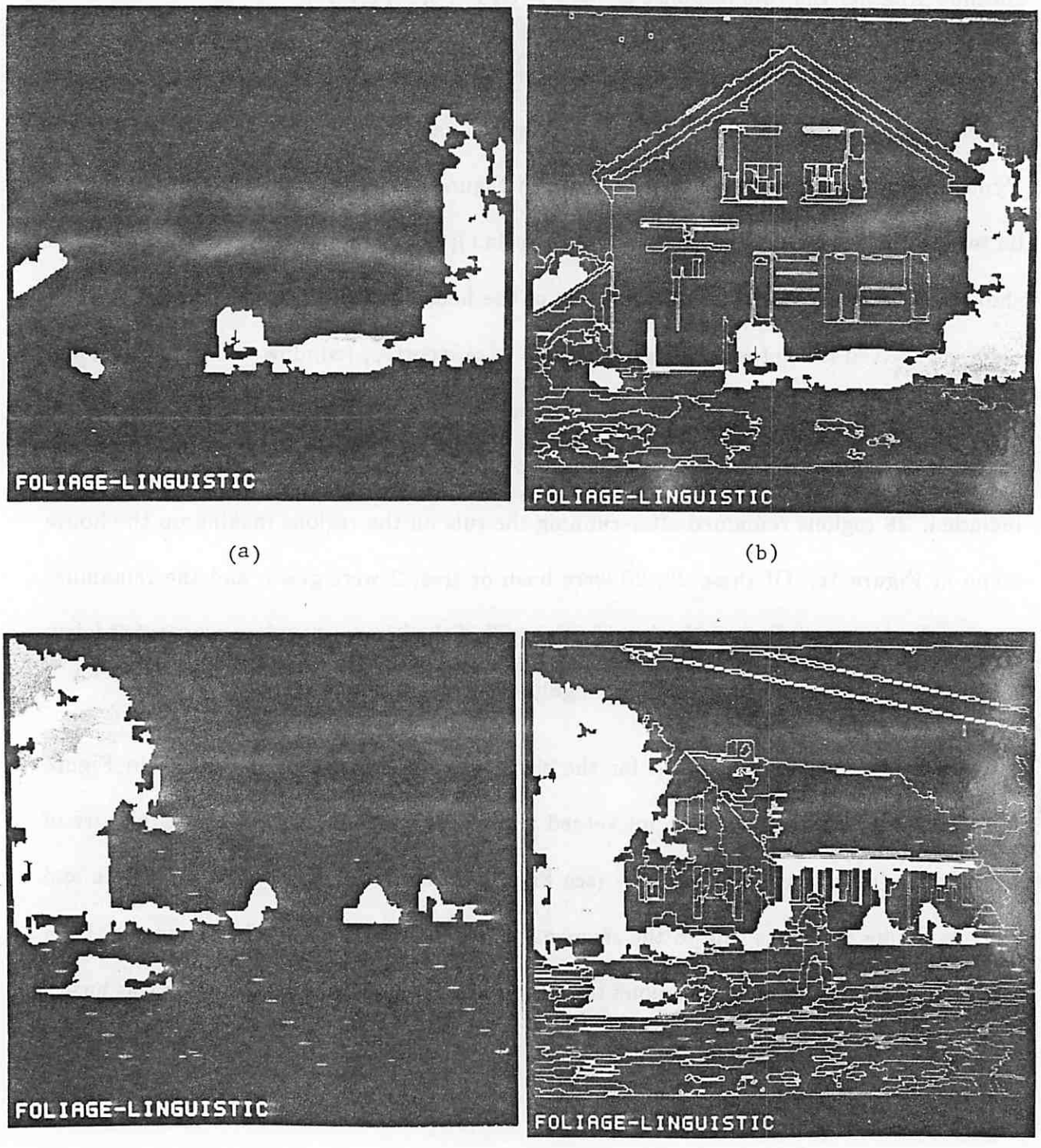


Figure 14. Foliage results from the coarsely quantized rules applied to the segmentations of Figure 4a and c.

terms that are intuitive to the user, including the structure of the composite rules and the relative weights, as well as the setting of the individual rule values.

The rule system, as described in this and preceeding sections, has been applied to a number of outdoor images of several different types (including road scenes). Although quantitive data is not yet available in sufficient quantity to generate believable statistics, qualitatively the results presented here are typical. The rules for grass, foliage and sky appear to be effective in extracting a set of regions which include actual grass, foliage, and sky regions ranked at the top or near the top of the list. Similar results have been obtained for road and sidewalk (concrete and macadam) and, to a somewhat lesser extent, for house roof. The rules appear to be loosely enough defined that normalization of the features has not been necessary. Additional experimental results are being generated for these and other rules; the results will be reported on in the future.

V. SCHEMAS AND THEIR INTERPRETATION STRATEGIES

V.1. Introduction

In the VISIONS system, scene independent knowledge is represented in a hierarchical schema structure organized as a semantic network [13,15,29,41]. The hierarchy is structured to capture the decomposition of visual knowledge into successively more primitive entities, eventually expressed in symbolic terms similar to those used to represent the intermediate level description of a specific image obtained from the region, line, and surface segmentations. Each schema defines a highly structured collection of elements in a scene or object; each object in the scene schema, or part in the object schema, can have an associated schema which will further describe it. For example, a house (in a house scene hierarchy) has roof and house-wall as object-parts, and the house-wall object has windows, shutters, and doors as object-parts. Each schema node (e.g. house, house wall, and roof) has both a declarative component appropriate to the level of detail, describing the relations between the parts of the schema, and a procedural component describing image recognition methods as a set of hypothesis and verification strategies called *interpretation strategies.*

The contextual verification of hypotheses via consistency with stored knowledge leads to a variety of interpretation strategies that are referred to as data-directed (or bottom-up), knowledge-directed (or top-down), or both. In addition these strategies can be domain and object-dependent, or uniform across domains. A rich set of possibilities opens up which unfortunately has not been sufficiently explored by the research community carrying out knowledge-directed vision processing. The first type of interpretation strategy discussed uses the rule system to select "exemplar" regions as object candidates, extending the labels

to similar regions. A second class of strategies uses geometric information to direct the grouping of intermediate events to better match the expected model. A third strategy involves the detection and correction of errors in the interpretation process as shown in an example in the section on final results.

V.2. Exemplar Selection and Extension

The most reliable object hypotheses obtained by applying the object hypothesis rules to the intermediate level data (e.g. regions, lines, and surfaces) can be considered object "exemplars" which capture image specific characteristics of the object. The set of exemplars can be viewed as a largely incomplete kernel interpretation. There are a variety of ways by which the exemplar regions can be used to extend and refine the kernel interpretation [40,41] and we will briefly present one specific implementation for an exemplar extension strategy. The strategy is based on the expectation that the image-specific variation of a feature of an object is less than the inter-image variation of that feature for the same object. In many situations another instance of the object can be expected to have a similar color, size, or shape and this expectation can be used to detect similar objects.

There are a variety of ways by which the exemplar regions can be used and by which the selection of similar regions can be made, depending on the data and knowledge available as well as on the complexity of the object [22,31]. For those objects for which the spectral characteristics can be expected to be reasonably uniform over the image, the similarity of region color and texture can be used to extend an object label to other regions, perhaps using the expected spatial location and relative spatial information in various ways to restrict the set of candidate regions examined. The similarity criteria might also vary as a function of the object, so that regions would be compared to the exemplar in terms of a

particular set of features associated with that object. Thus, a sky exemplar region would be restricted to comparisons with regions above the horizon that look similar (in terms of color and texture) to the largest, bluest region located near the top of the picture. A house wall showing through foliage can be matched to the unoccluded visible portion based upon color similarity and spatial constraints derived from inferences about house wall geometry.

The shape and/or size of a region can be used to detect other instances of multiple objects, as in the case when one shutter or window of a house has been found [10], or when one tire of a car has been found, or when one car on a road has been found. In many situations another instance can be expected to have a similar size and shape. This, together with constraints on the image location, permits reliable hypotheses to be formed even with high degrees of partial occlusion. If one is viewing a house from a viewpoint approximately perpendicular to the front wall, other shutters can be found via the presence of a single shutter since the single shutter provides strong spatial constraints on the location of other shutters. If two shutters are found then perspective distortion can be taken into account when looking for the other shutters, even without a camera model, under an assumption that the tops and bottoms of the set of shutters lie on a straight line on the face of the house. There are many alternatives by which features of an object could be used to determine the full set of regions representing the object. In the current version of the VISIONS systems, these alternatives are represented in the interpretation strategies associated with the object schema.

We have made a basic assumption that exemplar extension will in fact involve a knowledge engineering process that will use different strategies for each object. In some cases, color and texture may be more reliable than shape and size (as in a sky exemplar region), while in other situations shape and size might be very important (as in the shutters).

Spatial constraints may be used very differently in each case. One possibility is to utilize the object-specific set of simple features that were associated with the object hypothesis rules. In fact the same rule system presented earlier can be used to weight the feature differences to form the similarity rating. Similarity extension rules can be implementations of distance metrics or functions, with values of feature differences used to determine the rule response for each. It might be appropriate in one case to employ a piecewise-linear distance function, with high values for small differences and low values for large differences; in another case, a rule might provide a uniformly high response within some threshold. It is also easy to use the veto region for large differences, or for spatial constraints to restrict the spatial area over which region candidates for exemplar extension will be considered.

For the similarity results shown in Figure 15 and 16, the full object hypothesis rules discussed earlier were used to measure the similarity between the exemplar region and each candidate region. The rule response was converted into a distance metric and bright regions correspond to small differences. Each of the rules contained location and size components which enter into the final distance measurement, but in general, there are more intelligent ways of using these features in the interpretation strategy responsible for grouping regions. Our goal here was simply to rank order the regions that were candidates for extension; and again, the specific results shown in the figures are not as important as the overall philosophy.

Figure 15 shows the similarity rating of regions obtained using the features of the grass rule (Figure 9) and comparing them to the exemplar region (see Figure 13). In addition to the final grass similarity response, the similarities obtained from the color component rule and two of its constituent simple rules (green-magenta and intensity) are also shown. Figure 16 shows similar results for foliage.

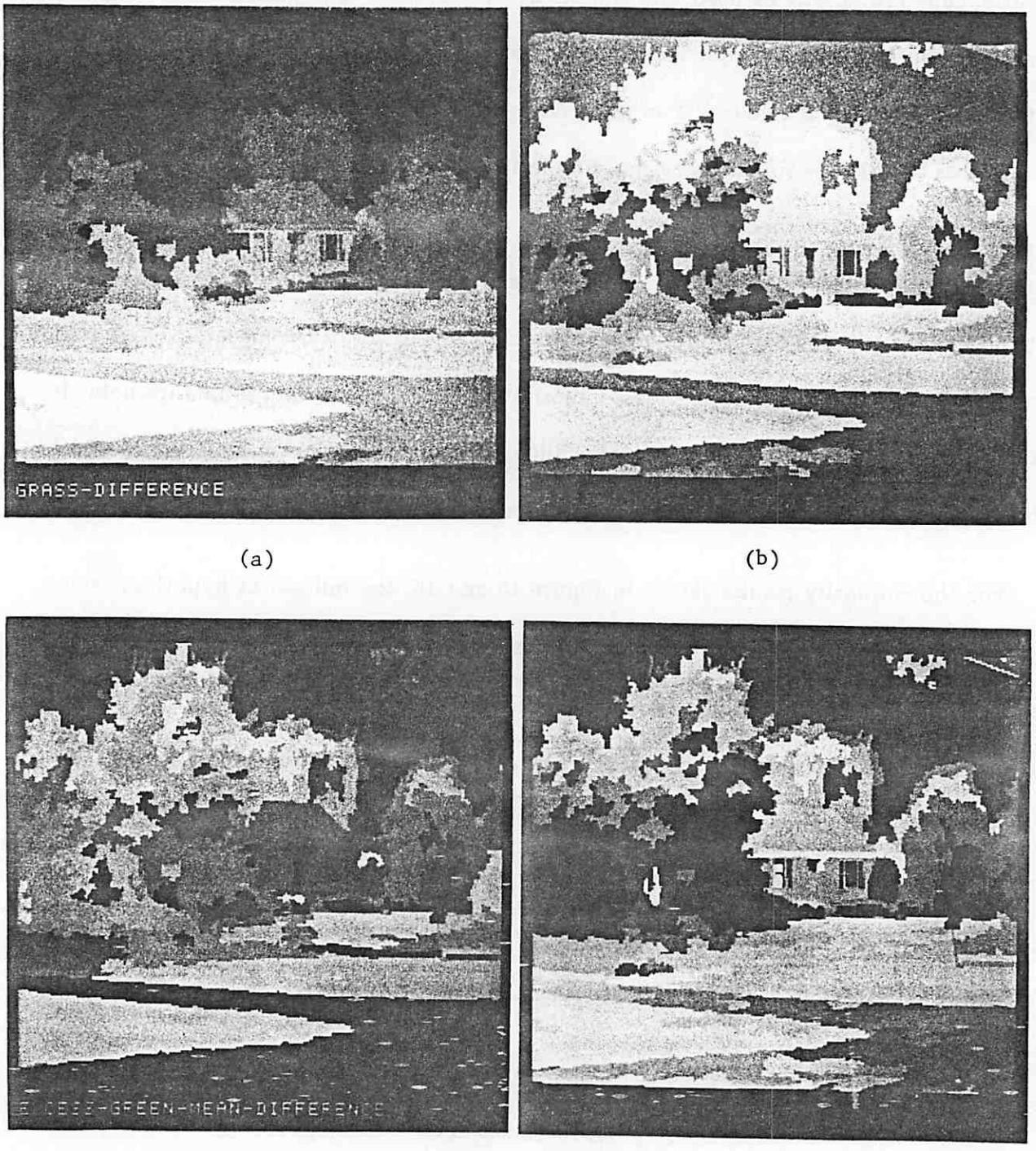


Figure 15. Similarity of the grass exemplar region to all other regions for Figure 4b. (a) similarities from the composite object hypothesis rule; (b) similarities from only the color component of the composite rule; (c,d) similarities from the simple rules associated with excess green and intensity. In all cases, similarity is encoded as brightness.

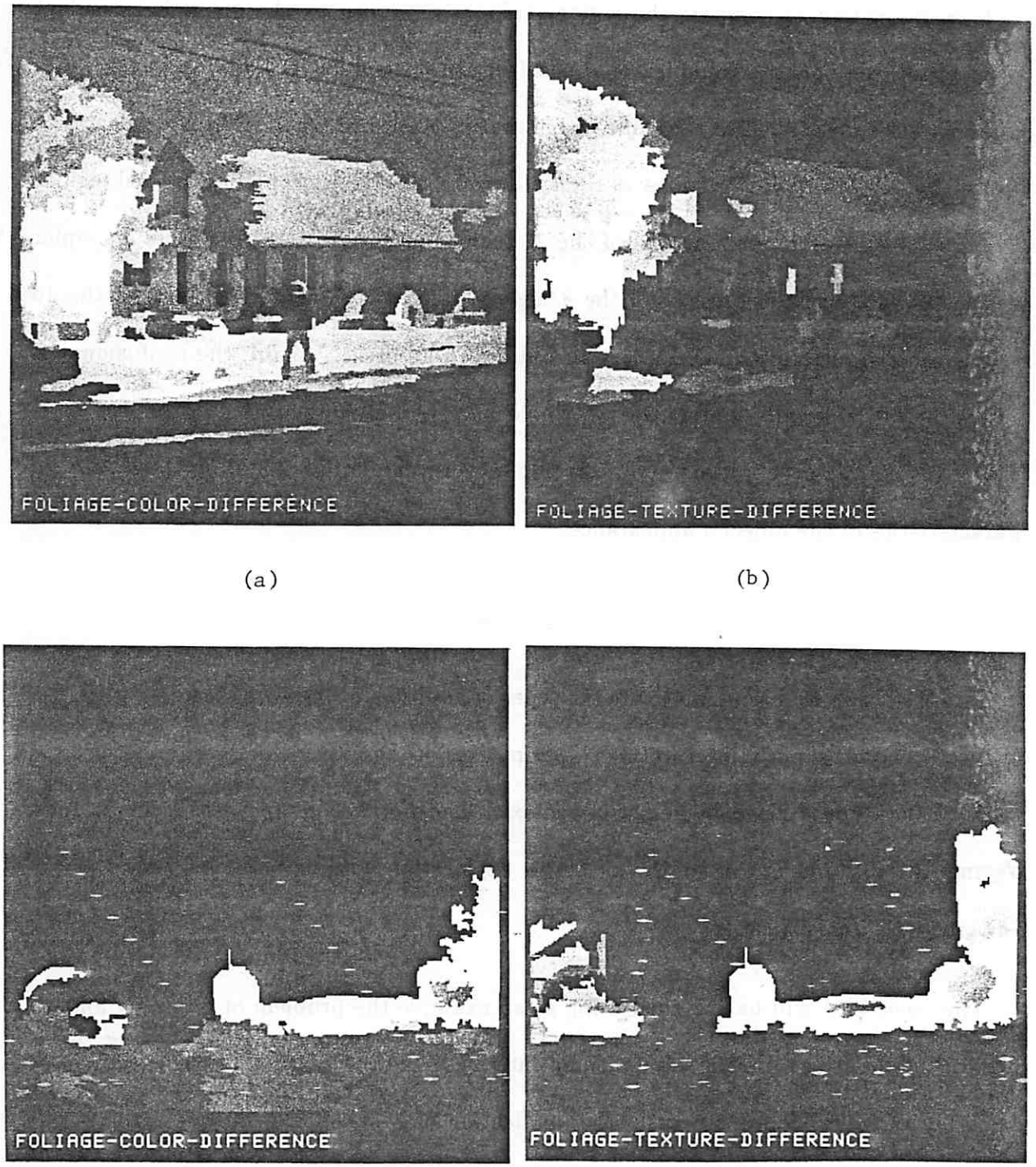


Figure 16. Foliage exemplar matches. (a,b) foliage matches using the largest region in the tree area on the left side of the image (region 69 in Figure 4c); (c,d) foliage matches using the large region in the low bushes in front of the house (region 128 in Figure 4a).

It is interesting to compare the region rankings produced by the image independent initial grass rule and the rankings obtained from the exemplar matching strategy. The original grass rule applied to the same image that in Figure 15 (these results are shown in Figure 12) vetoed all but 27 regions; of these, 15 were grass and the 6 highest ranked regions were grass. By way of comparison, of the 27 regions most similar to the grass exemplar region, 18 were actually grass, and the 8 highest ranked regions were grass. Of the 16 regions most similar to the exemplar, all but two were grass. Again, the confusion was between grass, sidewalk, and driveway. The exemplar matching strategy produces more reliable results than the initial hypothesis rule since it takes into account image-dependent characteristics of the object's appearance.

V.3. Interpretation Strategies for Intermediate Grouping

In this section we briefly motivate the types of additional top-down strategies that will be necessary for properly interpreting the primitives of the intermediate representation in terms of the hypothesized higher level context. The work presented here is taken from Weymouth [40], and is the subject of active exploration within the vision group at the University of Massachusetts.

The basic idea will be sketched using as an example the problem of grouping and interpreting a house roof from a fragmented intermediate representation. Figure 17 shows a number of intermediate stages in the application of a house roof interpretation strategy associated with the house roof schema. Figures 17a and b portray a pair of region and line segmentations that exhibit typical difficulties expected in the output of low-level algorithms; figure 17a also shows the initial roof hypothesis. In this example the region segmentation algorithm was set to extract more detail from the image by producing more

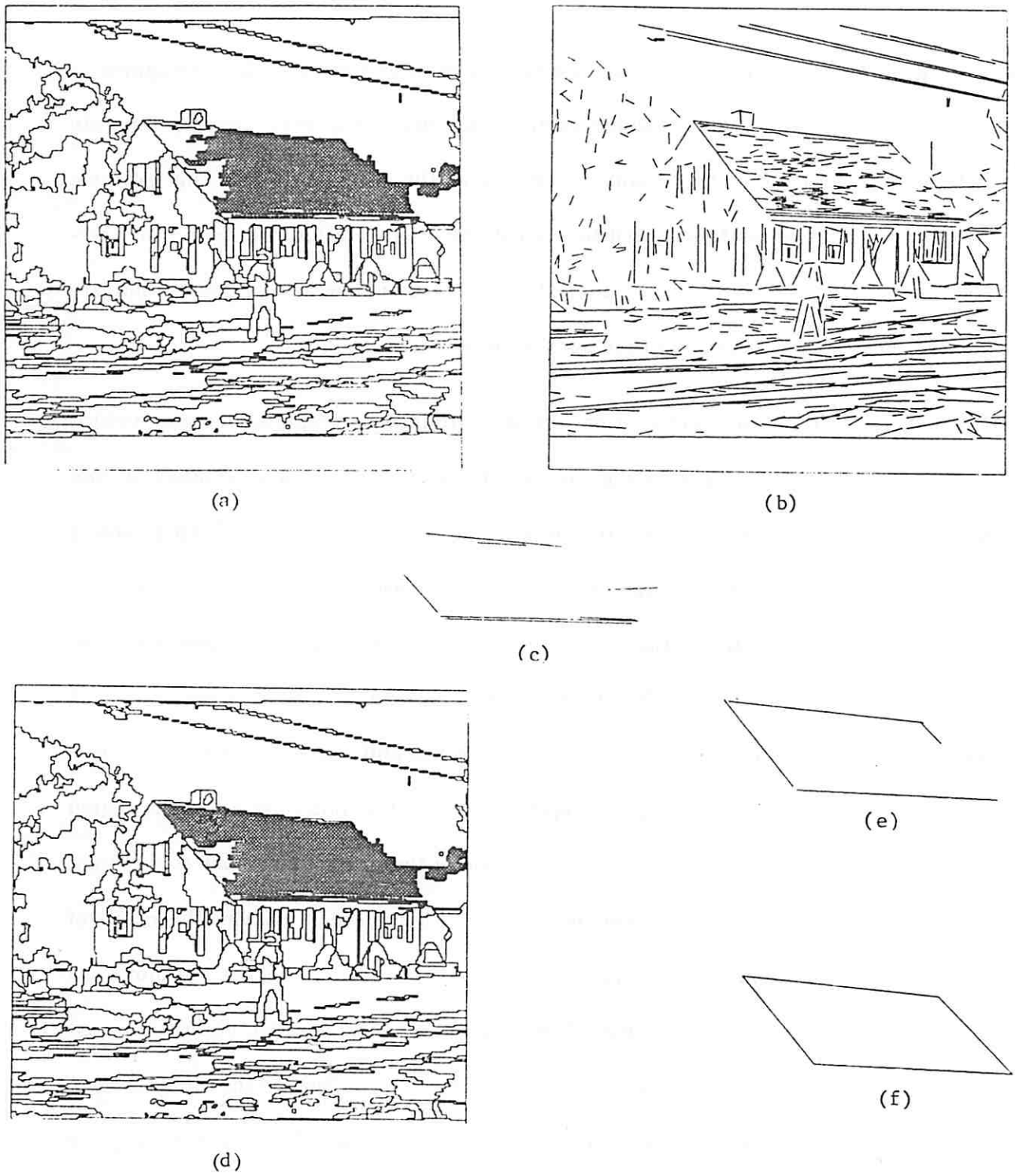


Figure 17. Steps in the schema directed interpretation of a roof. (a,b) region and line representations; the initial roof hypothesis region is crosshatched. (c) long lines bordering the roof hypothesis region boundary. (d) new roof hypothesis after merging regions which are partially bounded by the long lines in the previous image. (e) after joining colinear, nearby segments and filling a straight line to the joined boundaries. (f) the completed hypothesized roof trapezoid.

regions; the result, which is typical of a class of segmentation problems, is the fragmentation of the roof, so that the shadowed left portion was broken into several regions separate from the main roof region. (It should be noted that the segmentation in this example is different from the segmentation used to present the interpretation results in other sections.) When examined carefully, the line extraction results show line segments that are fragmented into shorter pieces, multiple parallel lines, and gaps in lines.

The goal is to use typical segmentation results to produce the trapezoidal region (which is almost a parallelogram) representing the perspective projection of a rectangular roof surface, as well as the orientation in 3D space of that surface. The top-down grouping strategy that we employ here is organized around evidence of the almost parallel lines forming the two sets of sides of the trapezoid. There are alternate strategies for other typical situations where some of this information is missing. Thus, this roof grouping strategy expects some evidence for each of the four sides, and in particular uses the long lines bounding the putative roof region. Figure 17c shows the long lines along the region boundary; "long" is determined as a relative function of the image area of the roof region, here 1/3 of the square root of the roof region area. Figure 17d illustrates the result of merging similar regions which are partially bou.ded by the long lines. Figure 17e shows the lines bounding the extended roof region, after removing shorter, parallel, almost-adjacent lines, joining co-linear nearby segments, and then fitting straight lines to the boundaries to form a partial trapezoid (almost a parallelogram). Finally, Figure 17f shows the complete hypothesized roof trapezoid. The three-dimensional geometry of the roof can then be computed (up to some possibly non-trivial degree of error) based upon either the location of the pair of vanishing points of the two sets of image lines that are parallel in the physical world, or one pair of parallel lines and an assumption of perpendicular angles to a third

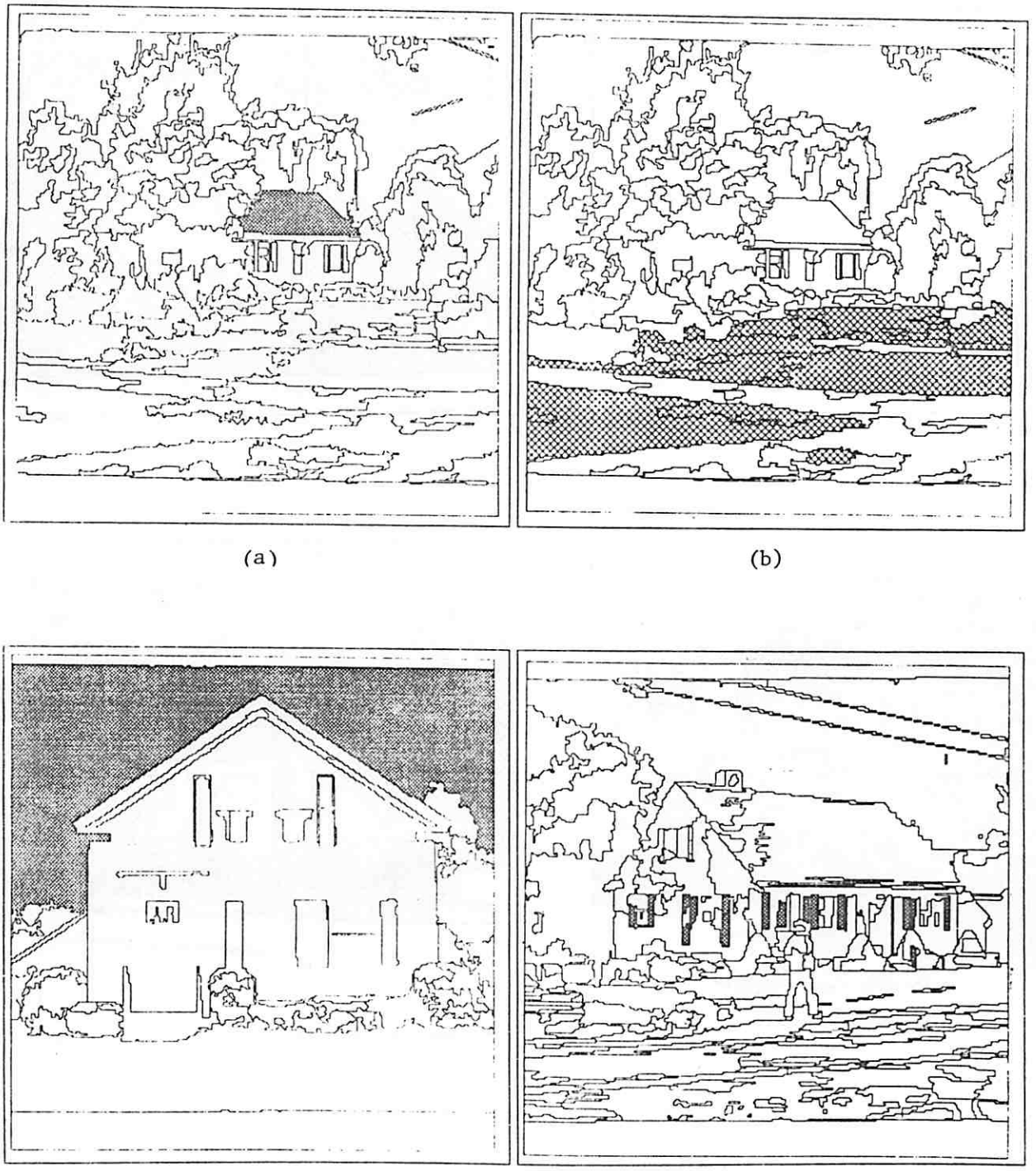
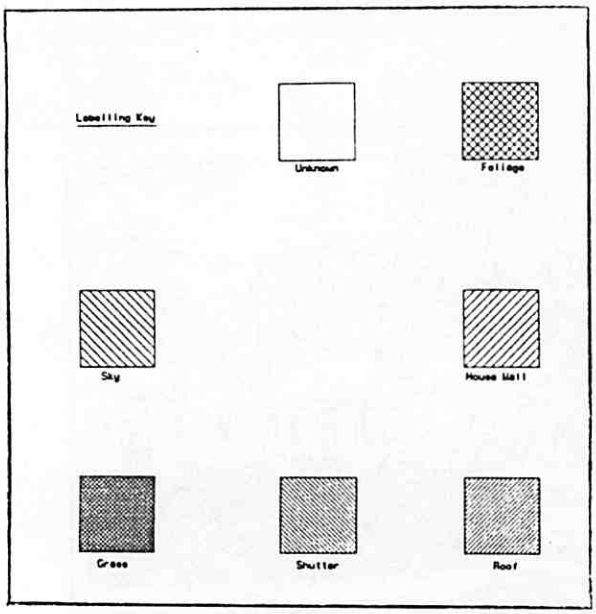
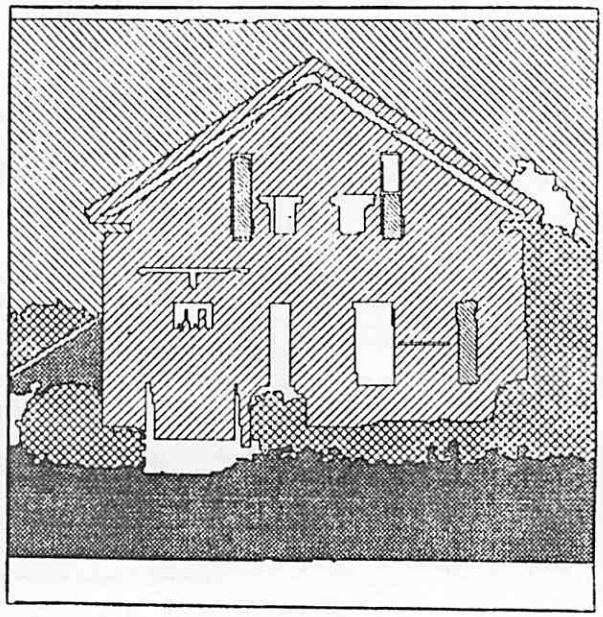


Figure 18. Results from schema interpretation strategies. (a) roof rule-based on roof features and spatial relation to house wall; (b,c) grass and sky-based on exemplar selection and similarity extension; (d) shutter rule-based on shape and spatial relationships to each other and house wall.

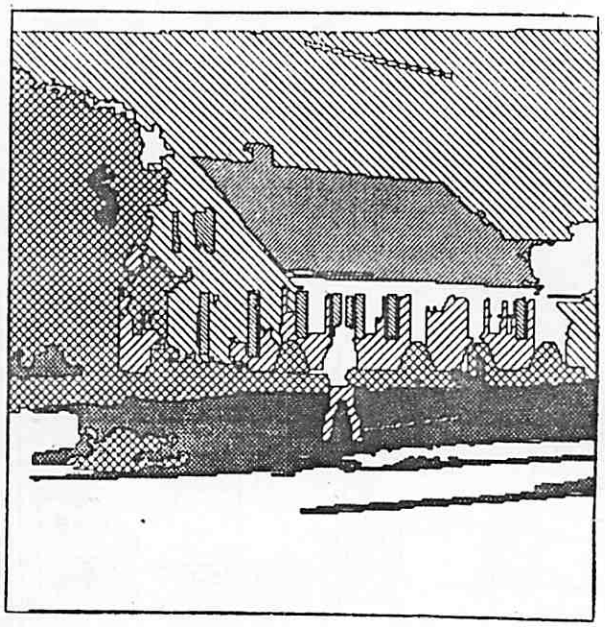


(a)



(b)

(c)



(d)

Figure 19. Final interpretations. These images show the final results obtained by combining the results of the interpretation strategies under the constraints generated from the knowledge base. (a) interpretation key; (b-d) interpretation results. In (d), the missing boundary between sky and wall results in a labelling conflict (the identification shown is sky; a second interpretation has this region labelled house wall).

line [27].

The point of this discussion is that the interpretation process required a flexible strategy for grouping and reorganizing the lines and regions obtained from imperfect segmentation processes. At this point in our understanding we are developing each strategy independently, but we hope to begin to define some standard intermediate grouping primitives that would form the basis of a variety of general top-down strategies.

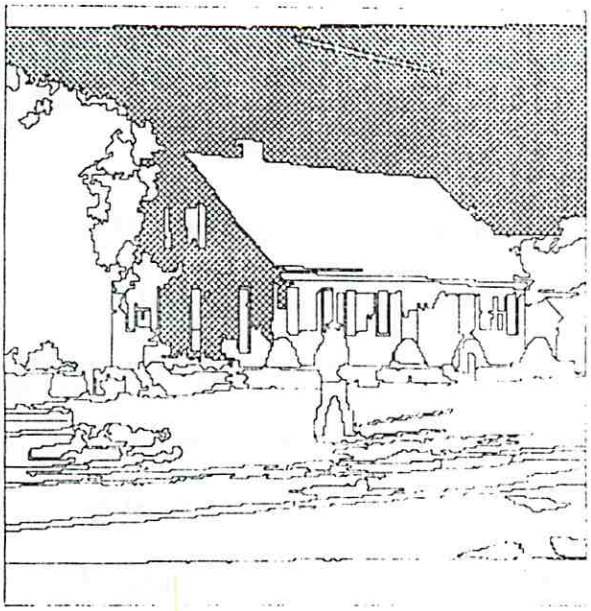
V.4. Results of Rule-Based Image Interpretation

Interpretation experiments are being conducted on a large set of "house scene" images. Thus far, we have been able to extract sky, grass, and foliage (trees and bushes) from many of these images with reasonable effectiveness, and have been successful in identifying shutters (or windows), house wall and roof in some of them. Object hypothesis and exemplar extension rules as described in previous sections were employed. Additional object verification rules requiring consistent spatial relationships with other object labels are being developed. The features and knowledge utilized vary across color and texture attributes, shape, size, location in the image, relative location to identified objects, and similarity in color and texture to identified objects. In the following figures, we show isolated intermediate and final results from the overall system.

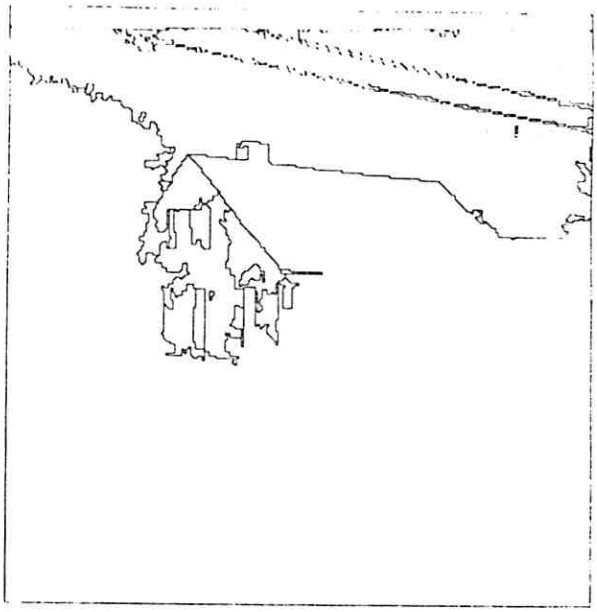
The interpretation results shown were obtained from a version of the VISIONS system that used a different (somewhat coarser grained) set of initial segmentations than those presented earlier, and a set of object hypothesis and exemplar extension strategies that differed in structure (but not in principle) from those presented earlier. Figure 18 shows selected results from the object hypothesis rules after exemplar extension and region merging. Figure 19 illustrates typical interpretations obtained from a house scene schema

interpretation strategy that utilizes a set of object hypothesis rules for exemplar selection, extends the partial model from the most reliable of these hypotheses, and which employs relational information for verifying hypotheses and predicting image location of object parts and related objects. The image areas shown in white in Figure 19 are uninterpreted either because the object did not exist in the knowledge network (and hence no label could be assigned) or because the object varied in some way from the rather constrained set of alternate descriptions of the object stored in the knowledge base.

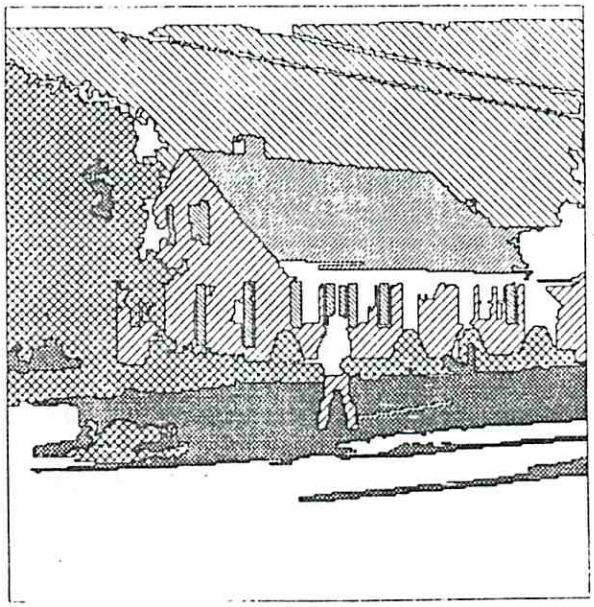
The interpretation shown in Figure 19d illustrates a problem which may be expected to occur quite frequently. The original interpretation was produced using a fairly coarse segmentation, which is desirable from the standpoint of computational efficiency since fewer regions are involved. However, the sky and house wall are merged into one region (as shown more clearly in Figure 20a) since the feature differences between the two areas was below the resolution of the segmentation process. This is the disadvantage of using a coarse segmentation. Parameters of the segmentation processes can be set to produce a finer grained segmentation but many more regions are produced. On the other hand, the smaller regions can be expected to find more of the desired boundaries and sometimes better match the object descriptions in the knowledge base, at the risk of overfragmentation. Because of these conflicting constraints, we believe it is extremely important to closely couple the lower level processes responsible for constructing the intermediate representation and the interpretation processes which operate on the intermediate and higher level representations. The interpretation processes should have focus-of-attention mechanisms for correction of segmentation errors, extraction of finer image detail, and verification of semantic hypotheses. The advantage of top down application of these processes rests in the focussed nature of the processing. Since the processes do not have to be applied ev-



(a)



(b)



(c)

Figure 20. Resegmentation of house/sky region from Figure 18d. (a) the original segmentation showing the region to be resegmented; (b) the regions resulting from the resegmentation of the selected region; (c) final interpretation of the house scene in Figure 1c, after inserting resegmented house/sky regions and reinterpreting the image.

erywhere in the image, they can be more computationally expensive and make heavier use of knowledge specific to the particular problem.

An example of the effectiveness of semantically directed feedback to the segmentation processes is shown in Figure 20. The missing boundary between the house wall and sky led to competing object hypotheses (sky and house wall) based upon local interpretation strategies. The region is hypothesized to be sky by the sky strategy, while application of the house wall strategy (using the roof and shutters as spatial constraints on the location of house wall) leads to a wall hypothesis. There is evidence available that some form of error has occurred in this example: 1) conflicting labels are produced for the same region by local interpretation strategies; 2) the house wall label is associated with regions above the roof (note that while there are houses with a wall above a lower roof, the geometric consistency of the object shape is not satisfied in this example); and 3) the sky extends down close to the approximate horizon line in only a portion of the image (which is possible, but worthy of closer inspection).

In this case resegmentation of the sky-housewall region, with segmentation parameters set to extract finer detail, produces a reasonable segmentation of this region (Figure 20b). It should be pointed out that in this image there is a barely discernable boundary between the sky and house wall. However, once the merged region is resegmented with an intent of overfragmentation, this boundary can be detected. Now, the same interpretation strategy used earlier produces the quite acceptable results shown in Figure 20c. We note that this capability (of detecting labelling conflicts and resegmentation) is not automatic in the current version of the system.

Future work is directed towards refinement of the segmentation algorithms, object hy-

pothesis rules, object verification rules, and interpretation strategies. System development is aimed towards more robust methods of control: automatic schema and strategy selection, interpretation of images under more than one general class of schemata, and automatic focus of attention mechanisms and error-correcting strategies for resolving interpretation errors.

VI. PRINCIPLES TO GUIDE KNOWLEDGE-BASED VISION RESEARCH

In summary, we list some of the principles of our work on knowledge-based vision systems that might provide guidance to other researchers. We do caution the reader, however, that in no way are we asserting that this is the "only" or "correct" or "complete" approach to high-level vision. Rather, the problem domain has proved to be so difficult that there has been little work of any generality. Thus, our statements at this time are distilled from the experience of a partially successful approach to general knowledge-based vision that is continuously evolving as we understand the visual domain more thoroughly.

1) An integrated symbolic representation of 2D image events such as regions and lines, and 3D world events such as surface patches, should be used as the symbolic interface between sensory data and world knowledge. In particular it is the attributes of these elements, potentially including depth (3D) and motion (2D and 3D) information, that provides linkages to stored knowledge and higher-level processing strategies.

2) In the initial stages of bottom-up hypothesis formation, focus of attention mechanisms should be used to selectively group elements of the intermediate representation and construct tentative object hypotheses. The interpretation should be extended from such "islands of reliability". The choice of object classes for initial consideration can be controlled (top-down) via context or expectation.

3) A simple initial interface to knowledge can be obtained via rules defined over a range of the expected values of the attributes of the symbolic events that have been extracted. These rules can be organized around the most likely events or the easiest

events to extract when highly structured situations are expected.

4) Knowledge of the physical world should be organized around scene schema and object schema that can be represented as a structured collection of parts. This allows the contextual relationships to guide the further processing of partial interpretations. In places where 3D shape and spatial relations are complex, the general relationships between image events in typical 2D views can be used to interface to the bottom-up 2D symbolic representation. However, long-range progress is dependent upon more effective 3D shape representations.

5) More complex strategies will be needed for matching salient aspects of the ambiguous, incomplete, and sometimes incorrect intermediate data representation to the object models stored in the knowledge base. They involve a diverse collection of goals, and given our understanding at this time, it may be easier to represent them as procedural knowledge. These strategies include knowledge-directed grouping, deletion, and manipulation of intermediate symbolic entities, as well as goal-oriented feedback to low-level processes.

6) Inference mechanisms for utilizing distributed fine-grained and weak hypotheses will be needed. These inference mechanisms must deal with the issues of high degrees of uncertainty and of pooling a variety of sources of information in order to control processes for extending partial interpretations.

7) Highly interactive user-friendly environments for visually displaying results of knowledge application are very important. The vision domain provides a natural medium for user feedback and interaction.

Acknowledgements

This work has been supported by the Air Force Office of Scientific Research under contract AFOSR-85-0005 and the Defense Mapping Agency under contract 800-85-C-0012. The authors wish to acknowledge the many members of the VISIONS research community, particularly Robert Belknap, Joey Griffith and Terry Weymouth, who have contributed to the technical ideas developed in this paper and the software that produced the results. Our thanks also to Janet Turnbull and Laurie Waskiewicz for their patience and perseverance in producing this manuscript.

REFERENCES

[1] R. Bajcsy and M. Tavakoli, *Computer Recognition of Roads from Satellite Pictures*, IEEE Transactions on Systems, Man, and Cybernetics **SMC-6** (September 1976), 623 - 637.

[2] D. Ballard, C. Brown and J. Feldman, *An Approach to Knowledge-Directed Image Analysis*, Computer Vision Systems (A. Hanson and E. Riseman, eds.) (1978), Academic Press.

[3] H. Barrow and J. Tenenbaum, *MSYS: A System for Reasoning About Scenes*, Technical Note 121 (April 1976), AI Center, Stanford Research Institute.

[4] R. Belknap, E.Riseman and A. Hanson, *The Information Fusion Problem and Rule-Based Hypotheses Applied to Complex Aggregations of Image Events*, Proceedings of IEEE-CVPR Conference (June 1986), 227-234..

[5] T. Binford, *Survey of Model Based Image Analysis Systems*, International Journal of Robotics Research **1** (1982), 18–64.

[6] M. Brady, *Computational Approaches to Image Understanding*, Computing Surveys **14** (March 1982), 3–71.

[7] R. Brooks, *Symbolic Reasoning Among 3-D Models and 2-D Images*, STAN-CS-81-861 and AIM-343 (June 1981), Department of Computer Science, Stanford University.

[8] J.B. Burns, A.R. Hanson and E.M. Riseman, *Extracting Linear Features*, Proc. of the Seventh International Conference on Pattern Recognition (July 30 - August 2, 1984), Montreal, Canada.

[9] R. Davis, *Expert Systems: Where are We? And Where Do We Go From Here?*, AI Magazine 3 (Spring 1982), 3 - 22..

[10] L. Erman, et al., *The Hearsay-II Speech-Understanding System: Integrating Knowledge to Resolve Uncertainty*, Computing Surveys **12(2)** (June 1980), 213–253.

[11] O. Faugeras and K. Price, *Semantic Descriptions of Aerial Images Using Stochastic Labeling*, IEEE PAMI **3** (November 1981), 638-642.

[12] E. Fisher, *The Use of Context in Character Recognition*, Ph.D. Thesis and COINS Technical Report 76-12 (July 1978), University of Massachusetts at Amherst..

[13] A. Hanson and E. Riseman, *A Summary of Image Understanding Research at the University of Massachusetts*, COINS Technical Report 83-35 (October 1983), University of Massachusetts at Amherst.

[14] A. Hanson and E. Riseman, Computer Vision Systems (1978), Academic Press.

[15] A. Hanson and E. Riseman, *VISIONS: A Computer System for Interpreting Scenes*, Computer Vision Systems (A. Hanson and E. Riseman, eds.) (1978), 303 - 333, Academic Press.

[16] A. Hanson and E. Riseman, *Segmentation of Natural Scenes*, Computer Vision Sys-

tems (A. Hanson and E. Riseman, Eds.), Academic Press (1978), 129–163.

[17] A. Hanson, E. Riseman, and E. Fisher, *Context in Word Recognition,* Pattern Recognition, Vol. 8, No. 1 (January 1976), 35–45.

[18] T. Kanade, *Model Representation and Control Structures in Image Understanding,* Proc. IJCAI-5 (August 1977).

[19] R. Kohler, *Integrating Non-Semantic Knowledge into Image Segmentation Processes,* Ph.D. Dissertation and COINS Technical Report 84-04 (1984), University of Massachusetts at Amherst.

[20] V.R. Lesser and L.D. Erman, *A Retrospective View of the Hearsay-II Architecture,* Proc. IJCAI-5 (1977), 790–800, Cambridge, MA.

[21] M. Levine and S. Shaheen, *A Modular Computer Vision System for Picture Segmentation and Interpretation,* IEEE PAMI 3 (September 1981), 540 – 556.

[22] J.D. Lowrance, *Dependency-Graph Models of Evidential Support,* Ph.D. Dissertation and COINS Technical Report 82-26 (September 1982), University of Massachusetts at Amherst.

[23] A. Mackworth, *Vision Research Strategy: Black Magic, Metaphors, Mechanisms, Miniworlds, and Maps,* Computer Vision Systems (A. Hanson and E. Riseman, eds.) (1978), Academic Press.

[24] D. Marr, Vision (1982), W.H. Freeman and Company, San Francisco.

[25] M. Nagao and T. Matsuyama, A Structural Analysis of Complex Aerial Photographs (1980), Plenum Press, New York.

[26] P.A. Nagin, A.R. Hanson and E.M. Riseman, *Studies in Global and Local Histogram-Guided Relaxation Algorithms,* IEEE **PAMI-4** (May 1982), 263–277.

[27] H. Nakatani, R. Weiss, and E. Riseman, *Application of Vanishing Points to 3D Measurements,* Proceedings of the 28th Annual International Technical Symposium on Optics and Electro-Optics (August 1984).

[28] Y. Ohta, *A Region-Oriented Image-Analysis System by Computer,* Ph.D. Thesis (1980), Information Science Department, Kyoto University, Kyoto, Japan.

[29] C.C. Parma, A.R. Hanson and E.M. Riseman, *Experiments in Schema-Driven Interpretation of a Natural Scene,* COINS Technical Report 80-10 (April 1980), University of Massachusetts at Amherst.

[30] K.E. Price and R. Reddy, *Matching Segments of Images,* IEEE PAMI 1 (June 1979), 110-116.

[31] G. Reynolds, Nancy Irwin, Allen Hanson and Edward Riseman,, *Hierarchical Knowledge-Directed Object Extraction Using a Combined Region and Line Representation,* Proc. of the Workshop on Computer Vision: Representation and Control (April 30–May 2, 1984), 238–247, Annapolis, Maryland.

[32] E. Riseman and A. Hanson,, *A Contextual Post Processing System for Error Cor-*

rections Using Binary n-Grams, IEEE Trans. on Computers, C-23,, 480–493..

[33] E. Riseman and A. Hanson, *A Methodology for the Development of General Knowledge-Based Vision Systems*, Proc. of the IEEE Workshop on Principles of Knowledge-Based Systems (December 1984), Denver, Colorado.

[34] E. Riseman and A. Hanson, *The Design of a Semantically Directed Vision Processor*, COINS Technical Report TR 71C-1, University of Massachusetts (February 1975), Revised version COINS Technical Report 75C-1.

[35] G. Shafer, A Mathematical Theory of Evidence (1976), Princeton University Press.

[36] J.M. Tenenbaum and H. Barrow, *Experiments in Interpretation-Guided Segmentation*, Technical Note 123 (1976), AI Center, Stanford Research Institute.

[37] J. Tsotsos, *Knowledge of the Visual Process: Content, Form and Use*, Proceedings of 6th International Conference on Pattern Recognition (October 1982), 654– 669, Munich, Germany.

[38] R. Weiss and M. Boldt, *Geometric Grouping of Straight Lines*, Proceedings of IEEE-CVPR Conference (June 1986), 489–495.

[39] L. Wesley and A. Hanson, *The Use of an Evidential-Based Model for Representing Knowledge and Reasoning about Images in the VISIONS System*, Proc. of the Workshop on Computer Vision (August 1982), 14–25, Rindge, New Hampshire.

[40] T.E. Weymouth, *Using Object Descriptions in a Schema Network for Machine Vision*, Ph.D. Dissertation and COINS Technical Report 86-24,, Computer and Information Science Department, University of Massachusetts at Amherst.

[41] T.E. Weymouth, J.S. Griffith, A.R. Hanson and E.M. Riseman, *Rule Based Strategies for Image Interpretation*, Proc. of AAAI-83 (August 1983), 429–432, Washington, D.C. A longer version of this paper appears in Proc. of the DARPA Image Understanding Workshop (June 1983), 193–202, Arlington, VA.

[42] Y. Yakimovsky and J. Feldman, *A Semantics-Based Decision Theory Region Analyzer*, Proceedings of IJCAI-3 (August 1973), 580–588.