FROM IMAGE MEASUREMENTS TO OBJECT HYPOTHESES

Allen R. Hanson
Edward M. Riseman

COINS Technical Report 87-129

December 1987

This work has been supported by the Air Force Office of Scientific Research under contract AFOSR-85-0005, the Defense Mapping Agency under contract 800-85-C-0012, and the National Science Foundation under contracts DCR-8500332 and DCR-8318776.

From Image Measurements to Object Hypotheses

Allen R. Hanson and Edward M. Riseman

Computer and Information Science Department
University of Massachusetts
Amherst, Massachusetts 01003

This work has been supported by the Air Force Office of Scientific Research under contract AFOSR-85-0005, the Defense Mapping Agency under contract 800-85-C-0012, and the National Science Foundation under contracts DCR-8500332 and DCR-8318776.

Contents

ABSTRACT

A basic step in the construction of a symbolic interpretation of an image is the initial iconic to symbolic mapping, which associates portions of the image with hypothesized object identities. 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 abstracted image events that are represented as symbolic tokens. This bottom-up step is necessary in order to activate potentially relevant knowledge structures, and to provide spatial constraints on the application of such knowledge.

This paper describes a simple mechanism for generating object hypotheses that relies on covergent evidence from a variety of attributes of region tokens produced by a segmentation process. We introduce the idea of a constraint function, which is an extended real-valued function defined over a single region attribute. A simple constraint maps a single attribute value of a region token into a graded response which can be viewed as a 'vote' for the associated object. Simple constraint functions are hierarchically organized into compound constraints that are applied to a set of token attributes in order to generate initial object hypotheses. Our use of these constraint-based techniques as a focus-of-attention mechanism is compared to classical pattern classification methods and goals.

The object hypothesis system may also be viewed as creating tentative 'islands of reliability' from which knowledge-driven processing may be initiated. The most highly rated object hypotheses can be considered object exemplars which capture image-specific characteristics of the object. Constraint functions can be defined on the difference in the values of a pair of token attributes in the same manner as the values of the token attributes themselves. Thus, a similarity constraint function can be applied to compare all tokens to a given token in an exemplar extension strategy.

1 Introduction

The interpretation of complex natural images, such as those typically encountered in the outdoors domain, has proven to be a formidable task. The variability in the structure of common objects and their appearance in an image, the complications arising from both object and sensor motion, and the image effects of the relationship between the observer, objects, and light sources produce an environment in which purely bottom-up approaches to interpretation cannot be expected to be generally effective. It has been our position for some time that world knowledge, expectations, and image scene content must be used to guide the construction of complete scene-describing symbolic interpretations[16,17]. 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, several rule-based systems which map features to object identities have been developed [23,28,30], relaxation processes have been employed for propagating hypotheses under uncertainty [3,14], algebraic constraint manipulation has been used in model matching [10], constraint satisfaction systems have been used to capture relational information [24,40], and frame-based (or schema-based) approaches for more general (and sometimes more complex) control strategies have been proposed [2,18,41]. A partial review of image interpretation research can be found in [2,8,9,17,42] and descriptions of several other individual research efforts beyond those cited above are found in [1,10,21,17,24,25,32,33,34,38,41,43,44,47].

Early attempts to interface stored knowledge to image data at the pixel level met with only limited success and little possibility of generalization [40]. 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 512×512 ≅ 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.

A basic step in the construction of a symbolic interpretation is the initial iconic to symbolic mapping, which associates portions of the image with hypothesized object identities. This bottom-up step is necessary in order to activate potentially relevant knowledge and to provide spatial constraints on its application. The approach presented in this paper addresses the start-up problem of interpretation by creating tentative 'islands of reliability' from which knowledge-driven processing can be initiated. Since these 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.

1.1 Complexity of Vision

The complexity of visual tasks can be made explicit by examining almost any unconstrained natural image. Although this initial discussion is qualitative, we believe the conjectures are 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 global context and stored expectations derived from experience are employed. 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. Here, we have selected some subimages of objects which show to varying degrees:

- *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;
- 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 is hard to imagine a bottom-up system deriving a surface model of the shoes based upon the intensity information shown in Figure 2. 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, 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.

2 Constraint-Based Object Hypothesis Strategies

The interpretation task of concern in this paper is that of labelling an initial region segmentation of an image [7,29] (such as those shown in Figure 3) 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 (see Figure 4).

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.[1] For each object, these expectations can be translated into

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

a complex constraint function on the attributes of a region token (see next section); the constraint function combines the results of many measurements into a confidence level that a region token (or small group of region tokens) represents that object. In previous papers [5,35,36], each constraint function on a single attribute was referred to as a *rule* which mapped attributes (or features) of a token to a vote for an object label. However, in order to avoid confusion with the use of the term *rule* in productions systems in the AI expert system literature, we have changed the terminology. Perhaps the term fuzzy constraint would be more appropriate, but this, too, carries potential confusion with existing terminology and approaches that we wish to avoid [48].

2.1 The Choice of Initial Token Primitives

It has been suggested that 2D region and line tokens are not appropriate object descriptions of the initial image data and that they should be replaced with local estimates of surface orientation, reflectance, depth, and velocity [4,9,26]. 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 a 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. In fact it has proven very difficult even when stereo, motion, and range data are available.

On the other hand, even if a very reliable surface description 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 such a perfect depth map was available. How should one partition the numerical depth array into meaningful entities such as surface patches, 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 with the two-dimensional region and line data. We believe that the principles and approaches presented here will be applicable not only to the two-dimensional tokens extracted from the static monocular color images presented in this paper, but also to the interpretation of three-dimensional depth data recovered from stereo and laser ranging devices, and both two- and three-dimensional motion data derived from a sequence of images.

2.2 Knowledge as Constraint Functions

The general idea is to represent expectations such as *grass is green* in the form of constraints on the attributes of region tokens, so that if the constraint is satisfied, that attribute *votes* for the associated object label. To do this, we introduce the idea of a simple constraint function as an extended real-valued function on a token attribute, which maps from attribute (or feature) space into a real-valued response:

$$C(R_{ij}) \in \mathbf{R} \cup \{VETO\}$$

where C is the constraint function, R_{ij} is the value of the i^{th} attribute/feature of the j^{th} region token, $\mathbf{R}$ is the set of real numbers, and VETO is a special response indicating that, on the basis of R_{ij}, the region token cannot possibly represent an instance of the associated object (see below). Simple constraint functions, as described above, are thus defined as the ranges over a scalar-valued feature which will map into a real-valued vote for an object label. Typically a feature will be the mean or variance of a property of the pixels composing the region. Belknap et al [5] demonstrate how this general idea can be extended to include multiple types of tokens (e.g. region and line tokens) by defining relational measures between tokens of the different representations. Compound constraint functions involve a combination of simple constraint functions, and they allow fusion of information from a variety of different types of measurements. Hereafter, constraint functions often will be referred to simply as constraints.

We will now develop a simple constraint that captures the expectation that grass is green using a feature that is a coarse approximation to a green-magenta opponent color feature, obtained by computing the mean of 2G–R–B for all pixels in this region.[2] In order to demonstrate the actual basis and form of knowledge embodied in the constraint, Figure 5 compares the green-magenta feature histogram of grass pixels to the histogram of the

[2] Note that R,G,B refers to the red, green, and blue components of the color image, respectively.

same feature for all pixels. This data was obtained by hand-labelling region segmentations of 8 sample images of outdoor house scenes with their object identities.

One approach to defining C for scalar-valued features is to construct this function using distance in feature space between the measured value and a stored prototype feature value. Let $d = d(f_P, R_{ij}) = |R_{ij} - f_P|$ be the distance between the measured feature value R_{ij} and the prototype feature point f_P and let $\theta_1 \leq \theta_2 \leq \ldots \leq \theta_6$ be real-valued thresholds on d. The value C of the simple constraint is then:

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

As shown in Figure 6, this defines a piecewise linear mapping in which the thresholds $\theta_i, i = 1, \ldots, 6$, 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 simple constraint receives a maximum vote of 1. Under the principle of *graceful degradation*, small changes in a feature measurement should not dramatically alter the system response, and therefore beyond some distance the voting function is linearly ramped to 0 as the distance in feature space increases. Note that the function C would have to be modified if the feature was not scalar-valued, as in a circular range for an orientation attribute, or if multiple modes were desired.

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 natural extension to the mechanisms presented here would generalize the constraint 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 constraint. If there are multiple peaks in the histogram, then a simple constraint can be defined for each peak independently; their results could be combined in a straightforward manner using a function, such as the maximum of the responses.

A simple constraint defines a range over the value of a feature within which a measured feature must fall if the region is to be considered an instance of the object label. A compound constraint is defined as a (partially redundant) set of simple constraints on token attributes that is assembled into a composite response via a combination function applied to the responses of the simple constraints. The combination function can take many logical or arithmetic forms; this is an extension of the functional form of hypothesis rule in Nagao et al [28]. The premise is that by combining many partially redundant constraints, the effect of any single unreliable constraint is reduced.

It is useful to impose a hierarchical structure on the set of constraints associated with an object. In the experiments described below, a compound constraint for each object is organized into a hierarchical structure containing five compound constraints which provide an expectation for the color, texture, location, shape, and size of region tokens relative to the object. Each of these compound constraints is composed of a set of simple constraints on various attributes related to the overall constraint. This allows some flexibility in combining several highly redundant features (e.g., several color features) into a compound constraint which is somewhat more independent of the other constraints (e.g., color vs. location), although this is still not theoretically sound as we shall discuss later. It should be recognized, however, that this is only one alternative for imposing a hierarchical structure on the set of constraints; 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 compound constraints for color, texture, location, shape, and size is in turn joined into a compound constraint. The structure of the heuristic compound constraint for grass is shown in Figure 7; it consists of a normalized weighted average of the five compound constraints $\mathbf{C}_h$:

$$\textit{grass score} = \frac{1}{N}\sum_{h=1}^{5} W_h \mathbf{C}_h$$

where the W_h are the weights and $N = \sum_{h=1}^{5} W_h$. Each of the compound constraints is in turn a weighted sum of a set of simple constraints:

$$\mathbf{C}_h = \frac{1}{M}\sum_i V_i C(R_{ih}))$$

where $C(R_{ij})$ is the response from an individual feature constraint based on feature i measured on region R_j, the V_i are the weights, and $M = \sum_i V_i$. Note that a VETO propagates through the summation as a VETO, so that a lower level constraint can still reject a region as a candidate for the object in question. The value associated with any compound constraint might have a weight of zero, which means that the constraint will have no effect on the weighted response of a higher level constraint.

The weights shown in Figure 7 capture the heuristic importance of each of the contributions to the response of the higher-level compound constraint. The weights are integers from 0 to 5, and reflect our belief that only a few levels of relative importance are needed ($weak \equiv 1, medium \equiv 3, strong \equiv 5$ in importance)[27]. Since many simple constraints are expected to be used in a compound constraint, the constraint response should be relatively insensitive to small changes in the weights. The intention is to allow obvious relative weightings to be expressed, but to avoid twiddling of numbers. As we shall discuss in Section 3 of this paper, the constraint response will be used only to order the regions for further processing on the basis of their similarity to the stored feature templates, rather than classifying them as an instance of a specific class. Thus, in this focus-of-attention approach, verification processes are expected to detect errors in the next stages of the image interpretation process.

2.3 Results of Constraint Application

Figure 8 shows the results of applying two simple constraints for grass to the region segmentation shown in Figure 3c. For each constraint there are two images. The left image of each pair is a composite feature histogram showing the feature distribution across all pixels in a set of images, and the distribution for grass pixels across the same set of images. The histograms were computed from a set of eight hand-labelled images and then smoothed. The right image of each pair shows the value of the constraint for each region coded as a brightness level: bright regions correspond to high values and dark regions correspond to low values.

The constraint that was developed interactively by the user is superimposed on the histograms in piecewise linear form. In the upper left portion of the constraint figures, *Target* refers to the object associated with the constraint, 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 constraint function (scaled to a maximum value of 100), while the lower numbers tabulate the percentage of target regions and other regions vetoed. Thus, the ideal constraint 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 between increasing the response of target regions and increasing the percentage of other regions that are vetoed, and optimal settings are not at all obvious. In some cases constraints for the target object were interactively set to exclude regions associated with other objects, while in other cases the goal was to maximize the response for the target object regions. Since the constraints are not independent, they were specified interactively in an empirical knowledge-engineering fashion. There is no intent here to put forth these specific constraints as a significant contribution or even as a satisfactory set; in fact some of these constraints probably need modification.

Figure 9 shows the response for three of the five compound constraints and the final result for the overall compound constraint. For each constraint the region response is shown superimposed over the image in two complementary formats. The left image of each pair shows the strength of the constraint response coded in the intensity level of each region; bright regions correspond to good matches. Since low response and vetoed region both appear dark, a second format is presented. The right image shows the vetoed regions in black with all others uniformly grey. Figure 10 shows the final results for the foliage, grass, and sky hypotheses for the house image in Figure 1b (vetoed regions are not shown).

The effectiveness of the constraints can be seen by examining the rank orderings of the regions on the basis of the values of the compound constraint responses. For the grass results shown in Figure 9d, for example, the two top ranked regions are actually grass. For the grass results shown in Figure 10a, the top six regions are grass and 8 of the top 10 regions are grass; the two non-grass regions were actually sidewalk and driveway. For the foliage responses shown in Figure 10b, the top 21 regions were some form of foliage (tree, bush, or undergrowth); of the 30 regions not vetoed, only 7 were non-foliage regions (all of them were grass and were among the lowest ranked of the non-vetoed regions (7 of the last 9). For the sky results in Figure 10c, only four regions were not vetoed and the top three were sky. The fourth region, with a significantly lower response, was actually foliage with some sky showing through. Figure 11 shows the highest ranked regions for each of three object hypothesis constraints (sky,grass, and foliage) when applied to the three example images. In Section 2.5 we discuss how these initial object hypothesis results may be used as the basis of a strategy to produce a more complete interpretation.

2.4 A Language Interface for Knowledge Engineering

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

Even though an immediate visual response to a proposed constraint is available, the knowledge engineer must not be forced into a "parameter twiddling" mode. The constraints should be robust enough so that a fairly crude specification of the constraint 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 constraints can be specified on any feature in terms a uniform scale of five intervals of the dynamic range of a feature - VERY-LOW LOW, MEDIUM, HIGH, VERY-HIGH [37,46]. These labels induce a partition on the range of the feature; for each interval, the user specifies whether the rule response is *ON*, *OFF*, or *VETO*.

Coarse quantization of the feature range offers the knowledge engineer the opportunity to quickly develop and assess a set of constraints without detailed examination of feature statistics. It may be possible, in some cases, to completely develop the knowledge base using semantic terms that are intuitive to the user, including the structure of the compound constraints and the relative weights, as well as the setting of the individual constraint parameters.

The results obtained from the coarsely quantized constraints are quite good and often are comparable to the results obtained in the previous section using the more carefully defined constraints. Typical results are shown in Figure 12 using the foliage hypothesis constraint on two of the test image segmentations (Figure 3a,c). These results are comparable to those shown in Figure 10, and demonstrate that precisely specified constraints are not always necessary to produce usable hypotheses for knowledge-based interpretation. For the segmentation of the house image of Figure 3a, a total of 24 regions survived the vetoes; 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. In Figure 1c, 28 regions remained after applying the foliage constraint to the image; of these, 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 constraint, and in all cases the most highly rated regions were foliage.

Similar results were obtained for the sky and grass constraints. 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 3b) and two were a mixture of tree and sky. Applying the grass constraint to the segmentation of Figure 1a results in 41 regions, 17 of which are grass and 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.

The constraint system, as described in this and preceding 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 meaningful statistics, qualitatively the results presented here are typical. The constraints 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 constraints appear to be defined sufficiently loosely that normalization of the features has not been necessary.

2.5 Extending Initial Object Hypotheses

The most highly rated object hypotheses, obtained by applying the object hypothesis system 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 [45,46], and we will briefly present one specific implementation.

The exemplar extension strategy presented next uses the similarity of region features between the exemplar and other regions. The strategy is based on the expectation that the image-specific variation of a feature of an object (i.e. intra-image variation) is less than the inter-image variation of that feature for the same object; Figure 5 is a typical example of this point. In many situations another instance of the object in the same scene can be expected to have a similar color, size, or shape and this expectation can be used to detect similar objects, while larger variations almost always occur across images due to changes in lighting, season, geographic locale, object class, etc.

There are many alternative subsets of features of an object which could be used to determine the full set of regions representing the 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 (for example, when extending one window shutter to other shutters). In the current version of the VISIONS systems, it is the responsibility of the knowledge engineer to encode within the interpretation strategies associated with the object schema, information about which of these alternative strategies is most appropriate, when they should be applied, and to what set of primitives they should be applied [13,36,45]. 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, and in that sense the general VISIONS system can be viewed as a collection of cooperating special-purpose vision systems[12].

A simple feature differencing mechanism can be used to extend hypotheses using the same set of features and associated weights that were employed in the original object hypothesis constraints. Assume that region R_k (the exemplar) has been hypothesized to be an instance of a given object T, and that attributes $i = 1, ...N$ will be used to compute similarity with other regions to extend that object hypothesis. Then, a relational measure of similarity S_j, applied to region R_j, is defined as a weighted distance function:

$$S_j = \sum_{i=1}^{N} W_i(|R_{ik} - R_{ij}|)$$

where W_i is the relative weight of attribute i. The city-block distance was used above for computational simplicity, but alternative distance metrics could be used for the relational similarity measure, including Euclidean distance or other forms of feature-normalized distances.

Constraint functions can be defined over the *difference* in the values of a pair of token attributes in the same manner as for the token attributes themselves (see Section 2.2). Thus, a similarity constraint can be applied to compare all tokens to a given token in an exemplar extension strategy. The piecewise-linear functional form allows significant flexibility. It might be appropriate in one case to employ an inverse linear function with high values for small differences and low values for large differences; in another case, the similarity constraint might provide a uniformly high response within some threshold. It is also easy to use a veto for large differences, as in the case of spatial constraints based upon a difference in location attributes to restrict the spatial distance from a given region token over which other region candidates will be considered for exemplar extension.

Results applying the similarity relational measure are depicted in Figure 13 and 14 and would form the basis of exemplar extension strategies as discussed above. The attributes in the full object hypothesis constraint (such as the one shown in Figure 7) were used to measure the similarity between the exemplar region and each candidate region, using the direct feature differences for the simple constraints and combining them into a weighted distance for the compound (multidimensional) feature distances; in the figures, bright regions correspond to small differences. The location and size differences were used in computing the compound weighted difference, 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 are candidates for extension; again, the specific results shown in the figures are not as important as the overall philosophy.

Figure 13 shows the similarity rating of regions obtained using the same set of features as those in the grass constraint (Figure 7) and comparing them to the exemplar region (see Figure 11). In addition to the final grass similarity response, the similarities obtained from the color compound constraint and two of its constituent simple constraints (green-magenta and intensity) are also shown. Figure 14 shows similar results for foliage.

It is interesting to compare the region rankings produced by the image-independent initial grass constraint that was applied to all regions and the rankings obtained from the exemplar similarity strategy, which tunes the region rankings to the initial best choice in the same image. The exemplar matching strategy produces more reliable results than the initial hypothesis constraint since it takes into account image-dependent characteristics of the object's appearance. The original grass constraint shown in Figure 10, when applied directly to all regions, 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. Thus, these limited results demonstrate the possibility that exemplar extension may be effective in practice as we have argued, although it is clear that more experimental results are necessary before strong conclusions are reached.

3 Relationship of the Constraint System to Bayesian Classification Techniques

Constraint-based object hypotheses and the resulting object hypothesis strategies just presented exhibit a number of characteristics that make them appealing to vision practitioners. The knowledge about the properties of objects is simple and there is an obvious relationship to semantic information that is intuitive to human experts who must build the knowledge base. Nevertheless, it will be instructive to consider the relationship to the standard statistical approaches used in Bayesian pattern classification. In this section, we will contrast at a general level the goals and problems of *classification* in a Pattern Recognition (PR) sense with *focus-of-attention* in an Artificial Intelligence (AI) sense.

3.1 Theoretical View of Classification

In the classical character recognition problem, the jth character in a sequence of characters, say 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 $\overline{X}_j$ extracted via measurements on character R_j, where $\overline{X}_j = (x_{1j}, \cdots, x_{Mj})$. One can view the region labelling problem to be equivalent, in that a set of feature measurements can be computed for all regions, and then each can be classified with respect to their object class 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; of course it is necessary that the training set be large enough to capture the expected variations in the domain. Under a reasonable set of assumptions, 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 $\overline{X}_j$, followed by the selection of the maximum likelihood class as the output decision. Thus, using Bayes rule sample R_j is classified as character C_i, where

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

Since $P(\overline{X}_j)$ is constant across the N classes, it cannot affect the decision and may be ignored. Thus, the decision rule typically is simplified to choosing class i such that:

$$\underset{i}{MAX}\ P(\overline{X}_j \mid C_i)\ P(C_i)$$

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

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 [15,20].

To summarize, in the pattern recognition/classification case, the set of classes is known, fixed, and usually is not large. Also note again that $P(\overline{X}_j)$ could be factored out of the computation of the posterior class likelihoods only 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 $P(\overline{X}_j \mid C_i)$. However, the a priori class probability $P(C_i)$ does vary in the computation of each class likelihood. Finally, the intent of the process is the classification of every character sample.

3.2 AI Focus-of-Attention Vs. Pattern Recognition Classification Approaches

Let us briefly make several points about why the pattern recognition (PR) classification paradigm is not effective in the current problem. It assumes a fixed set of known classes that usually 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 character.

The region labelling problem encountered in image interpretation is far more difficult than the typical pattern recognition problem. For a variety of reasons one must expect the data at this level of representation and at this stage of processing 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 [7,17]. As we have pointed out earlier, the sensory data is inherently noisy and ambiguous and this leads to segmentations that are unreliable and vary in uncontrollable ways; regions and lines are fragmented and merged, and thus, the decision processes often will be forced to rely on region and line samples, of which only some are meaningful. In the character recognition problem this would be akin to being given joined and split letters at a very high frequency. In fact this is one of the major problems in cursive script recognition that makes it much harder than recognition of printed or typed characters. In addition to the classification process, the system is repeatedly faced with the difficulty of organizing the input data into appropriate segments.

In addition to the above, 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. This would require a very different organization of the classification process. One must also expect many region samples which do not belong to any of the classes because they may be shadow regions, portions of occluded objects which cannot be identified, objects that have not been included in the set of target classes, or object parts which are only identifiable in the context of the object hypothesis. While there has been some success [47] in applying a Bayesian classification viewpoint to these problems, difficulties abound and we believe this approach generally leads to insoluble problems. We conclude that classical pattern recognition approaches are not powerful enough by themselves to produce effective classification in the domains we wish to consider.

Scene interpretation involves processes that construct complex descriptions, where many hypotheses are put forth; only those subsets that can be verified and which satisfy a variety of constraints are 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 a great reliance on stored knowledge of the object classes. This knowledge takes the form of object attributes and relations between objects, particularly relations between parts of objects which leads to a part-of object and scene hierarchy in the knowledge base. The techniques developed to utilize this type of information have been rather different than the classification strategies typically employed in pattern recognition systems. The conclusion that we have drawn is rather simple — initially there should not be an attempt to classify all image events. The organization of the input data is not sufficiently well-defined to attempt direct classification in the vision domain. However, the Bayesian classification strategy can be modified to become a *focus-of-attention* process, within which the unreliable nature of the features and the imprecision of the numerical values are explicitly considered. The hypotheses must be used in a somewhat qualitative manner, with need for further processing and verification[12,36,45].

With that viewpoint in mind let us consider the problem of labelling the regions of a given segmentation with relevant object identities. In particular, we re-examine Bayes rule in terms of AI strategies for focussing attention upon particular image events that are likely candidates for particular object labels. Thus, rather than the selection of the best object label for each region, we are looking for good region candidates for a particular object label. In other words, given a set of regions in an outdoor scene we wish to rank order the likelihood that each represents an object, such as *sky*. As we have seen in the preceding sections, the goal of this ordering is the selection of *exemplar* candidates, followed by the subsequent extension of the exemplar labels to other regions, and then the application of additional structured knowledge to infer a more complete interpretation.

From a Bayesian viewpoint, instead of the measurement vector $P(\overline{X}_j)$ being held constant across samples, the a priori class probability $P(C_i)$ for some fixed i is constant across regions to be classified. While there may be a large number of object classes, each class is processed independently. This changes the optimal decision rule via a Bayes formulation to selecting, for a given object class i, the region j such that:

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

An assumption of independence of the features in the measurement vector $\overline{X}_j$ allows us to decompose the joint probability into a product of feature terms:

$$\underset{j}{MAX} \prod_{k=1}^{M} \frac{P(x_{kj}|C_i)}{P(x_{kj})}$$

In summary, for a given class C_i all regions are analyzed in order to rank order and select the "best" candidates for further consideration. While there may be a common set of features measured on each region, it should be noted that the measurement vector $\overline{X}_j$ that is the basis of the decision could be different for each object, since it should be composed of those feature measurements most useful for hypothesizing the given object.

3.3 Difficulties with a Statistical Approach to Focus-of-Attention

By changing the classification paradigm to a hypothesis-and-test paradigm involving a focus-of-attention, we hope to finesse some of the inherent difficulties in the organization and quality of the data that was described earlier. However, this still leaves the problem of developing a representative and sufficiently large training set from which the statistics can be derived for $P(\bar{X} \mid C_i)$ and $P(\bar{X})$ for classes $i = 1, ..., n$.

Consider the size of a suitable training set: 50-100 samples of each class might exhibit much of the expected variation in handprinted characters (and far fewer samples would be needed for clean typewritten characters of a single font). However, an enormous number of training images containing objects of interest would be necessary before the variations in lighting, distance, viewpoint, as well as natural variations in object shape, size, and surface reflectance would be adequately represented. How many different views of tree and grass examples are needed if 50 to 100 training samples of a handprinted letter *a* are necessary to achieve statistical reliability?

The second major difficulty which makes the previous problem even worse is that the measurement vector $\bar{X} = (x_1, x_2, ..., x_M)$ will usually include several feature variables, and it is their joint probability distribution which must be estimated. The size of the training set needed will grow exponentially with the number of features. To avoid this problem, even in simple pattern classification problem domains, there is a typical assumption of independence of features:

$$P(\bar{X} \mid C_i) = P(x_1 \mid C_i) \dots P(x_M \mid C_i);$$

in other cases the feature set is assumed to be only pairwise dependent. In fact, when the independence assumption is violated, the Bayesian decision rule based upon feature independence is really a heuristic without theoretical foundation. Nevertheless, many problems have been solved with just such an approach. The point that is being made in this paper is that the world cannot be approached from a purely statistical paradigm, and that knowledge-based manipulation of a range of constraints by other mechanisms will be necessary.

3.4 In Defense of the Hypothesis by Constraint Paradigm

In the light of the discussions of the last few sections, one might argue that many of the difficulties with statistical approaches are inherent in the constraint-based approach as well. In this section we briefly revisit the methodology of generating hypotheses by constraints on the ranges of token attributes, and relate it to some useful aspects of the Bayesian ratio and a pragmatic interactive knowledge-engineering methodology.

It should now be clear that functional form of the the simple piecewise-linear graded constraint (without the veto range) is more than just an approximation to $P(x_k \mid C_i)$. Ignoring the problems of feature independence for a moment and considering a single scalar-valued feature x_k, the positive voting range for object class i is intended to cover a significant portion of the mass in $P(x_k \mid C_i)$. What it fails to include is the information contained in $P(x_{kj}) = \sum_i P(x_{kj} \mid C_i)$ which appears in the denominator of the Bayesian maximum-likelihood ratio (MLR) of

$$\underset{j}{MAX}\,\frac{P(x_{kj} \mid C_i)}{P(x_{kj})}$$

for rank ordering region candidates. The denominator is important because it brings in the degree of discrimination of each feature measurement x_k for class C_i. For example, there would be little value in a feature x_k 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.

When the compound constraints are developed for a particular object or object part, it is the responsibility of the vision system designer/knowledge engineer to select the most appropriate features and constraint parameters, such as the features and weights introduced in the grass and sky hypothesis constraints. To the degree that a constraint covered $P(x_k \mid C_i)$ and excluded the rest of $P(x_k)$, that constraint would be effective. Since this information is captured in $\frac{P(x_k|C_i)}{P(x_k)}$, when these statistics are available this information would be very valuable. Even if they cannot be expected to be very reliable, it might be better to present the ratio for use by the knowledge engineer during the construction of the piecewise-linear function.

The next question in the reader's mind might be: Why bother with the piecewise-linear approximation at all? If the ratio represents the optimal information associated with a scalar feature, then why not use it directly, rather than mapping it through an approximating function? This argument is quite sound and with an adequate training set. the MLR can be used directly (this underlies the automatic system developed in [34]); it clearly will be valuable in evaluating the importance of a given feature in discriminating a class C_i from other object classes. However, a major problem is that it may be very difficult to get a reliable approximation to $P(x_k)$ for reasons presented in earlier sections. The background objects in the world vary significantly and it is hard to imagine that such extensive statistics can be collected or are necessary.

On the other hand, there is still merit in the piecewise linear approximation. The central range represents a semantically intuitive and concise summary of the information in $P(x_k \mid C_i)$. We have argued fairly strongly that the values returned by the constraint when applied to the intermediate data must be used qualitatively with further verification of the best hypotheses. Therefore, fine numerical precision of the MLR should not be necessary. Furthermore, it will make the interface to human vision designers far more difficult. Since the knowledge base is expected to grow, evolve and be maintained, it is quite desirable to simplify the user interface. It is much easier for a person to consider that *a sky region will exhibit high intensity, low texture, ...*, rather than to evaluate the information in a set of complex continuous functions.

Finally, we have the very serious problem of the independence assumption of the feature set. We allow the designer of the knowledge base to choose any of a library of combining functions, including weighted average, multiplicative, logical or to write his own. However, here, we agree with the critics of Bayesian approaches. If the set of features are not independent, then joint probability functions must be estimated and the complexity of the problem explodes. We believe that in most cases features will be partially (and sometimes almost totally) redundant. The end result is that the actual use of this information becomes heuristic even when it is applied within equations derived from a Bayesian formulation. Our basic strategy is to utilize a certain degree of redundancy because of the expected lack of reliability in any single feature. There is no computationally feasible theoretical way around this problem. We just combine the information as if it were independent. However, the performance of the resulting constraints can be empirically evaluated, and the knowledge engineering process becomes an interactive paradigm of evaluation and refinement. In this regard, we share the same experimental methodology that has been at the heart of the development of most existing expert systems.

4 Conclusions

Simple constraints on the range of region token attributes, hierarchically combined into more complex constraints, are a simple and straightforward way of generating initial object hypotheses in an interpretation system. The technique is useful for those objects which may be described in terms of attributes such as color, texture, simple shape, size, location etc. The same technique may be used to extend initial hypotheses into other areas of the image by comparing attributes of the initial hypotheses with the attributes of unlabeled region tokens. The approach has been shown to be effective in extracting initial hypotheses for objects such as grass, sky, and foliage in unconstrained outdoor scenes; it has also been effective in extracting initial hypotheses for roads, roadsides, and house roofs [13].

The technique is related to the traditional classification methods found in pattern recognition. However, the intent is not to classify each region token as one of a fixed number of objects. Rather, the initial set of region hypotheses are intended to be used as a focus-of-attention set for higher level interpretation strategies; within this set the tokens can be rank-ordered on the basis of the values of the constraint function.

In a companion effort, Lehrer, Reynolds, and Griffith [22] have explored the use of the

Shafer-Dempster approach [39] to evidence combination as the mechanism for combining information from multiple attributes in object hypotheses. Their system automatically generates a knowledge-base using statistical information obtained from a set of training objects that is not infeasibly large. Attributes of region tokens are viewed as providing evidence for and against the object labels in the knowledge. Evidence from multiple attributes is combined using a computationally efficient version of Dempster's rule [11]. Results from the system applied to road scenes have been very promising, and are of similar quality to the results shown here. The technique is, of course, dependent upon a hand-labelled set of training images that have reasonable segmentations of the objects under consideration.

In another related effort, Riseman, Hanson, and Belknap [5,6] have extended the techniques presented here into a system which relates information across multiple types of tokens (e.g. regions and lines). In this extension, relational measures are defined between symbolic tokens, so that sets of tokens across representations can be selected and grouped on the basis of relational constraints applied to the measures. This work is a generalization of constraints defined on the similarity relational measure between region tokens (i.e. a relational measure between tokens of the same type) as described in Section 2.5. Control strategies which affect the ordering of constraint application can be used to reduce the computation required for producing token aggregations.

Acknowledgments

The authors wish to acknowledge the many members of the VISIONS research community, particularly Robert Belknap, Joey Griffith, Terry Weymouth, and Tom Williams, all of whom have contributed over an extended period of time to the ideas developed in this paper and the software that produced the results. Our thanks also to Janet Turnbull and Laurie Waskiewicz for aiding us 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*, September 1976, Vol. SMC-6, pp. 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. M. Tenenbaum, "MSYS: A System for Reasoning About Scenes", *Technical Note 121*, AI Center, Stanford Research Institute, April 1976.

[4] H. Barrow and J.M. Tenenbaum, "Computational Vision", *IEEE Proc. 69*, 1981, pp. 572-595.

[5] 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, pp. 227-234.

[6] R. Belknap, E. Riseman, and A. Hanson, "The Information Fusion Problem: Forming Token Aggregations Across Multiple Representations", COINS Technical Report 87-48, University of Massachusetts at Amherst, in preparation.

[7] J.R. Beveridge, R. Kohler, J. Griffith, A. Hanson, and E. Riseman, "Segmentation Using Localized Histograms and Region Merging", COINS Technical, University of Massachusetts at Amherst, in preparation.

[8] T. Binford, "Survey of Model Based Image Analysis Systems", *International Journal of Robotics Research*, Vol. 1, 1982, pp. 18 - 64.

[9] M. Brady, "Computational Approaches to Image Understanding", *Computing Surveys*, Vol. 14, March 1982, pp. 3–71.

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

[11] A.P. Dempster, "A Generalization of Bayesian Inference", *Journal of the Royal Statistical Society*, Series B, Vol. 30, 1968, pp. 205-247.

[12] B. Draper, R. Collins, J. Brolio, A. Hanson, and E. Riseman, "The Schema System", COINS Technical Report, in preparation.

[13] B. Draper, R. Collins, J. Brolio, J. Griffith, A. Hanson, and E. Riseman, "Tools and Experiments in the Knowledge-Based Interpretation of Road Scenes", *Proc. of the DARPA IU Workshop*, Los Angeles, CA, January 1987. Also COINS Technical Report 87-05, University of Massachusetts at Amherst, January 1987.

[14] O. Faugeras and K. Price, "Semantic Descriptions of Aerial Images Using Stochastic Labeling", *IEEE PAMI*, Vol. 3, November 1981, pp. 638–642.

[15] E. Fisher, "The Use of Context in Character Recognition", Ph.D. Thesis and COINS Technical Report 76-12, Computer and Information Science Department, University of Massachusetts at Amherst, July 1976.

[16] A. Hanson and E. Riseman, "The VISIONS Image Understanding System - 1986", in *Advances in Computer Vision*, (Chris Brown, Ed.), to be published by Erlbaum Associates, 1987. Also COINS Technical Report 86-62, University of Massachusetts at Amherst, December 1986.

[17] A. Hanson and E. Riseman, *Computer Vision Systems*, Academic Press, 1978.

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

[19] A. Hanson and E. Riseman, "Segmentation of Natural Scenes", *Computer Vision Systems*, (A. Hanson and E. Riseman, Eds.), Academic Press, 1978, pp. 129–163.

[20] A. Hanson, E. Riseman, and E. Fisher, "Context in Word Recognition", *Pattern Recognition*, Vol. 8, No. 1, January, 1976, pp. 35-45.

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

[22] N. Lehrer, G. Reynolds, and J. Griffith, "A Method for Initial Hypothesis Formation in Image Understanding", COINS Technical Report 87-04, University of Massachusetts at Amherst, January 1987. Also appeared in the *Proc. of the DARPA Image Understanding Workshop*, Los Angeles, CA, February 1987, pp. 521-537.

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

[24] M. Levine and S. Shaheen, "A Modular Computer Vision System for Picture Segmentation and Interpretation", *IEEE PAMI*, Vol. 3, September 1981. pp. 540 – 556.

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

[26] D. Marr, *Vision* W.H. Freeman and Company, San Francisco, 1982.

[27] M. Minsky, "A Framework for Representing Knowledge" in the *Psychology of Computer Vision*, (P.H. Winston Ed.), McGraw-Hill, 1975.

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

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

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

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

[32] K.E. Price and R. Reddy, "Matching Segments of Images", *IEEE PAMI*, Vol. 1, June 1979, pp. 110-116 .

[33] 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*, Annapolis, Maryland, April 30–Mary 2, 1984, pp. 238–247.

[34] G. Reynolds, N. Lehrer and J. Griffith, "A Method for Initial Hypothesis Formation in Image Understanding", *Proc. of the DARPA Image Understanding Workshop*, Los Angeles, CA, February 1987. Also COINS Technical Report 87-04, Computer & Information Science Department, University of Massachusetts at Amherst, January 1987.

[35] E. Riseman, A. Hanson and R. Belknap, "The Information Fusion Problem: Forming Token Aggregations Across Multiple Representations", COINS Technical Report 87-48, Computer & Information Science Department, University of Massachusetts at Amherst, December 1987.

[36] E. Riseman and A. Hanson, "A Methodology for the Development of General Knowledge-Based Vision Systems", in *Vision, Brain, and Cooperative Computation* (M. Arbib and A. Hanson, Eds.), MIT Press, Cambridge, MA, 1987, pp. 285-328. Also COINS Technical Report 86-27, University of Massachusetts at Amherst, July 1986.

[37] 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*, Denver, CO, December 1984.

[38] E. Riseman and A. Hanson, "The Design of a Semantically Directed Vision Processor", COINS Technical Report TR 71C-1, University of Massachusetts at Amherst, January 1974. Revised version COINS Technical Report 75C-1, February 1975.

[39] G. Shafer, *A Mathematical Theory of Evidence*, Princeton University Press, 1976.

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

[41] J. Tsotsos, "Knowledge of the Visual Process: Content, Form and Use", *Proceedings of 6th International Conference on Pattern Recognition*, Munich, Germany, October 1982, pp. 654- 669.

[42] J. Tsotsos, "Image Understanding" in *Encyclopedia of Artificial Intelligence*, (S. Shapiro, Ed.), John Wiley, New York, 1987, pp. 389-409.

[43] J. Tsotsos, "Representational Axes and Temporal Cooperative Processes", in *Vision, Brain, and Cooperative Computation*, (M. Arbib and A. Hanson, Eds.), MIT Press, Cambridge, MA, 1987, pp. 361-417.

[44] 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*, Rindge, New Hampshire, August 1982, pp. 14–25.

[45] T.E. Weymouth, "Using Object Descriptions in a Schema Network for Machine Vision", Ph.D. Dissertation and COINS Technical Report 86-24, University of Massachusetts at Amherst, 1986.

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

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

[48] L. Zadeh, "Approximate Reasoning Based on Fuzzy Logic", *Proc. 6th IJCAI*, 1979, pp. 1004-1010.

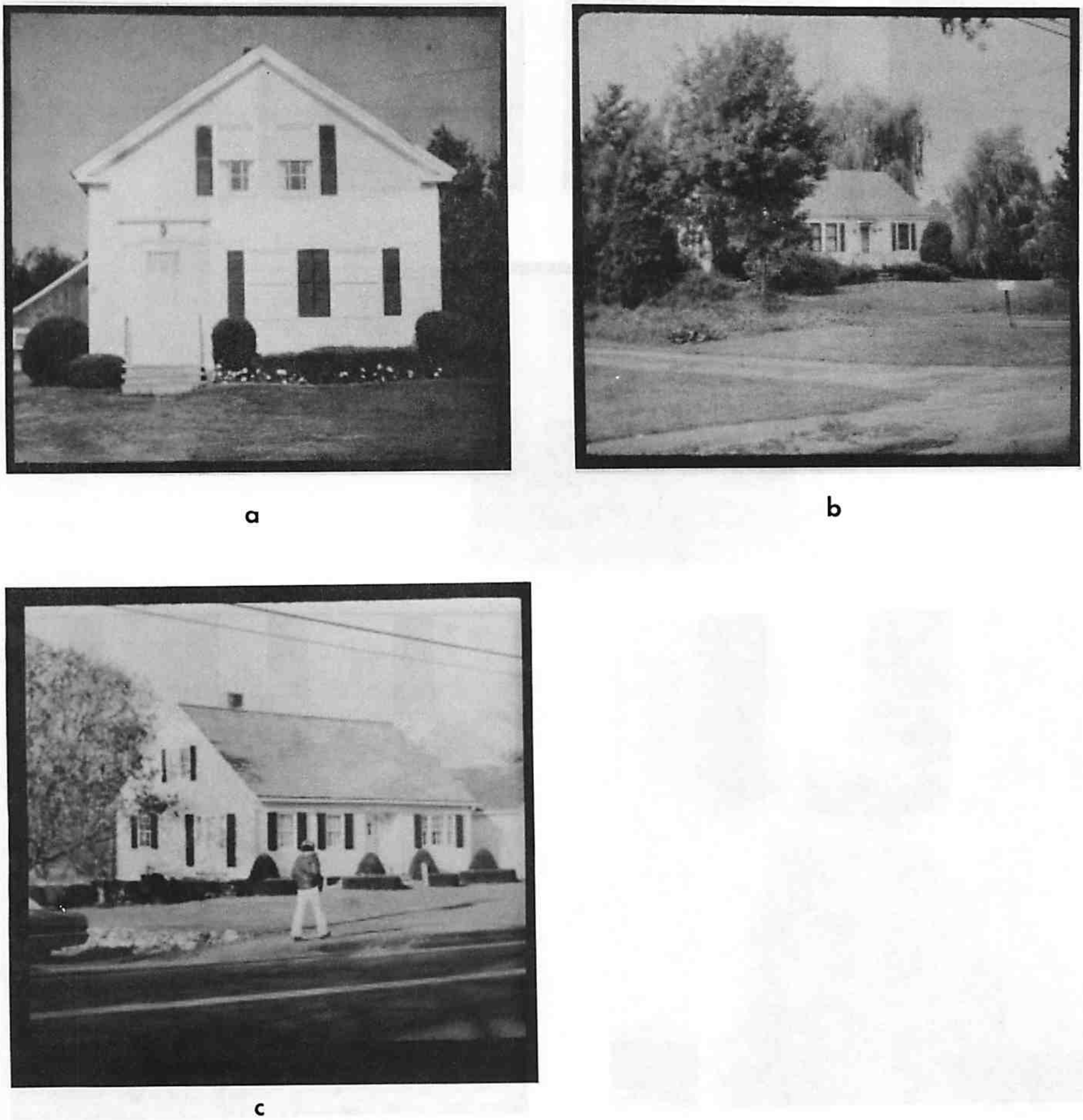


Figure 1. Original images. These images are representative samples from a larger data base. All are digitized to 512 x 512 spatial resolution, with 8 bits of gray scale resolution in the red, green, and blue components.

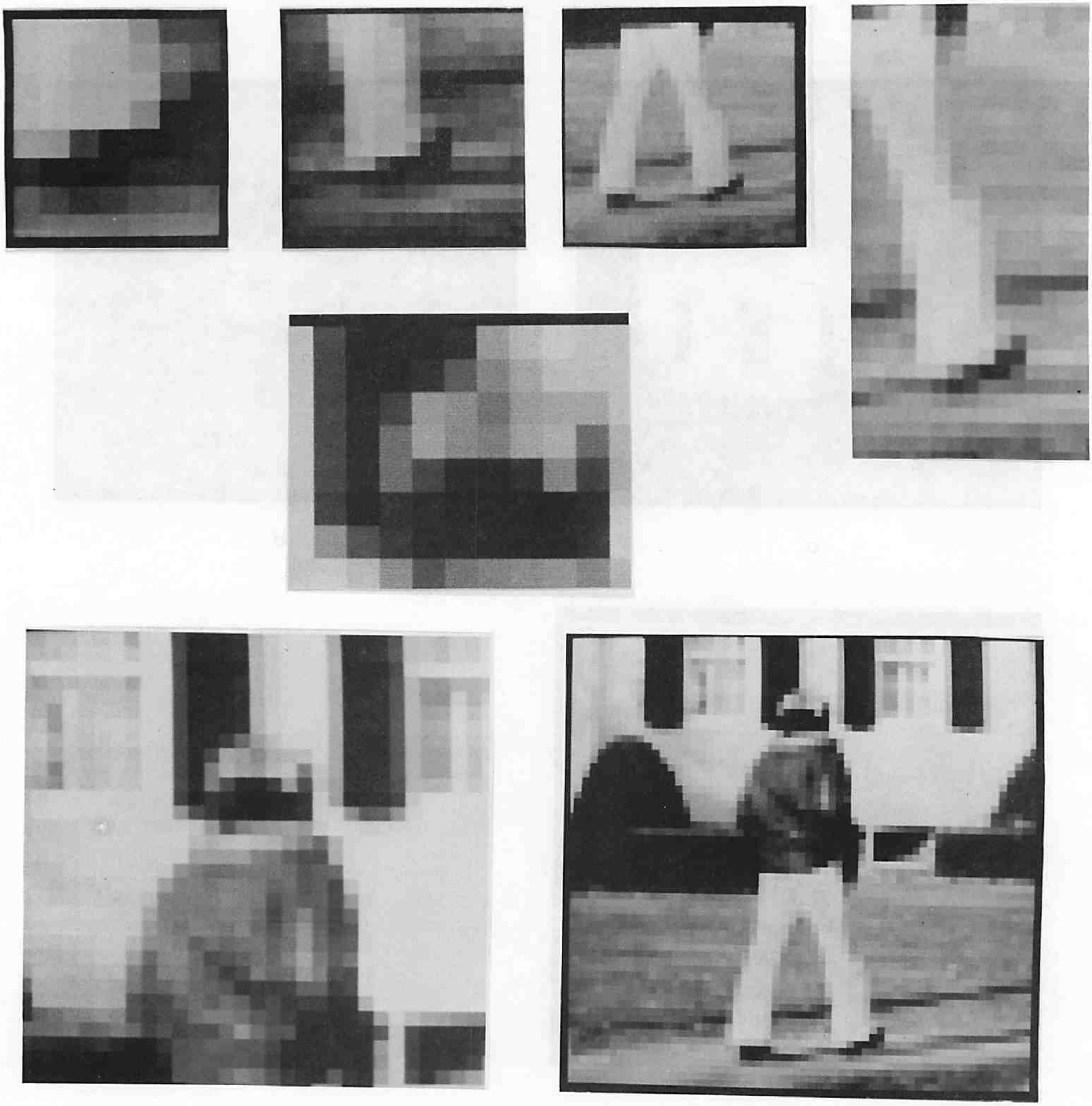


Figure 2. The Relevance of Context. Closeups from one of the original images with differing amounts of contextual information. In most cases, the identity or function of an object or object part cannot be determined from a small local view. Only when the subpart hierarchy or the surrounding context becomes available can the objects be recognized.

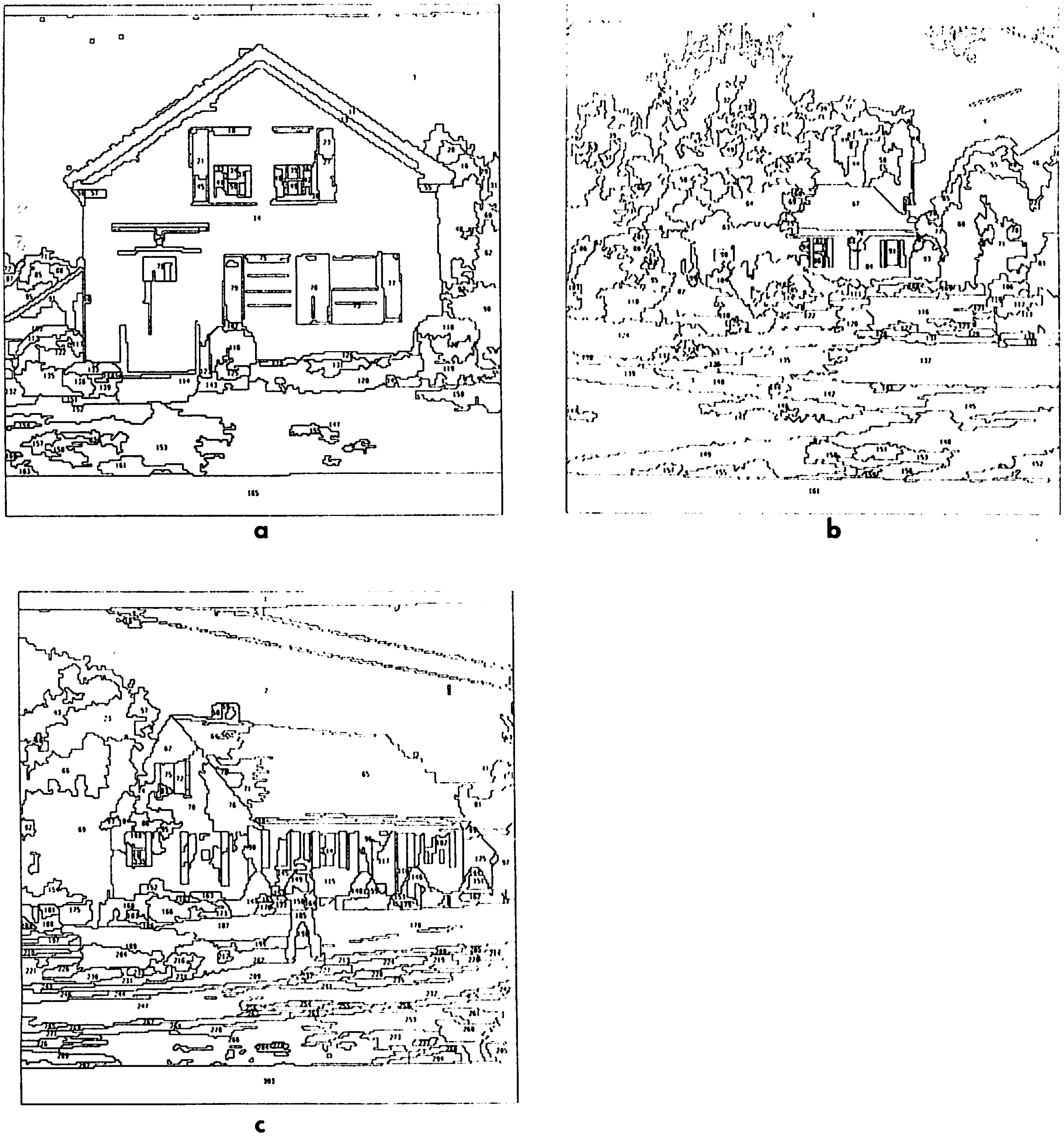


Figure 3. Region Segmentations. Regions partition the image into areas which are relatively uniform in some feature (in this case intensity). When region tokens are mapped into a symbolic structure with a rich set of descriptors, they provide the basis for an early linkage between the image data and a knowledge-base.

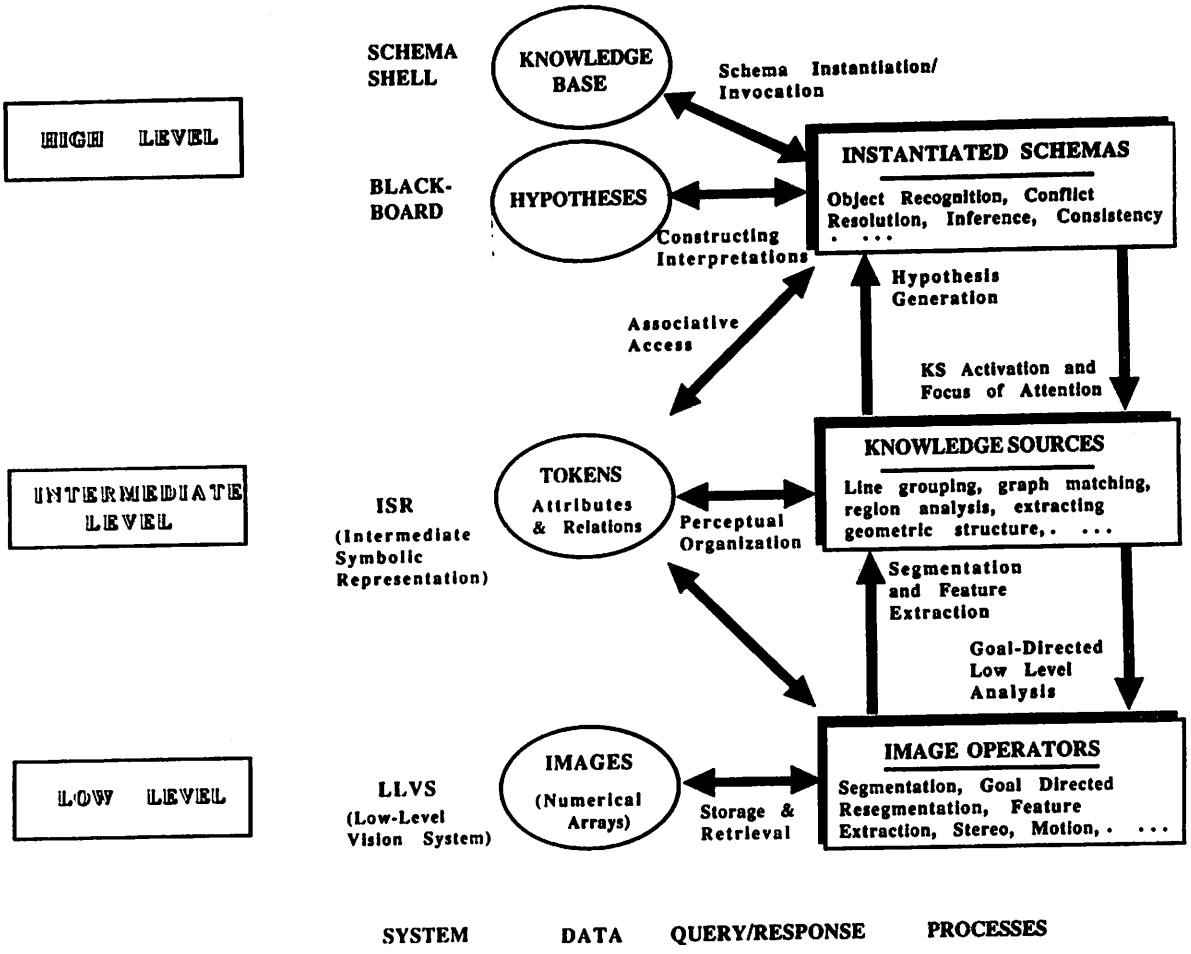

SCHEMA SHELL
KNOWLEDGE BASE
Schema Instantiation/ Invocation
HIGH LEVEL
INSTANTIATED SCHEMAS
Object Recognition, Conflict Resolution, Inference, Consistency • •••
BLACK-BOARD
HYPOTHESES
Constructing Interpretations
Hypothesis Generation
Associative Access
KS Activation and Focus of Attention
INTERMEDIATE LEVEL
TOKENS
Attributes & Relations
ISR
(Intermediate Symbolic Representation)
Perceptual Organization
KNOWLEDGE SOURCES
Line grouping, graph matching, region analysis, extracting geometric structure,. •••
Segmentation and Feature Extraction
Goal-Directed Low Level Analysis
LOW LEVEL
LLVS
(Low-Level Vision System)
IMAGES
(Numerical Arrays)
Storage & Retrieval
IMAGE OPERATORS
Segmentation, Goal Directed Resegmentation, Feature Extraction, Stereo, Motion,• •••
SYSTEM
DATA
QUERY/RESPONSE
PROCESSES

Figure 4. Multiple levels of representation and processing in the VISIONS system [14].

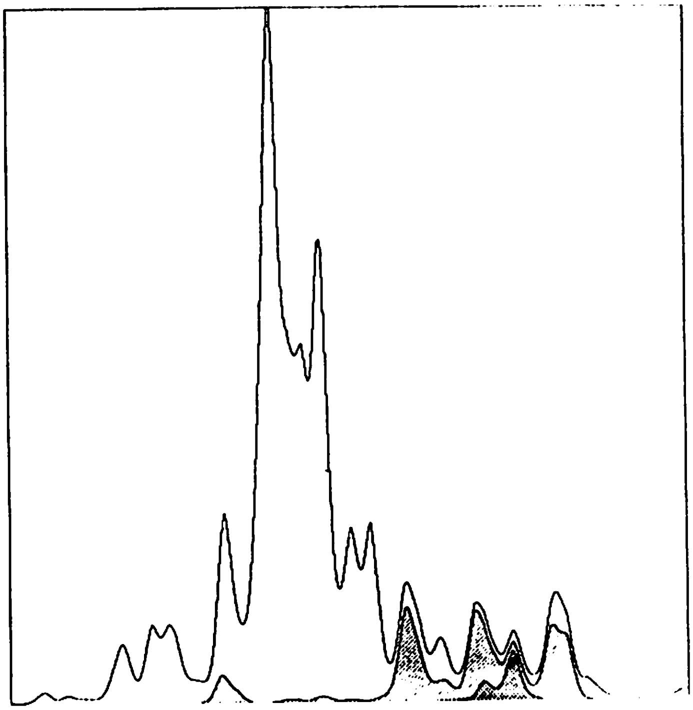
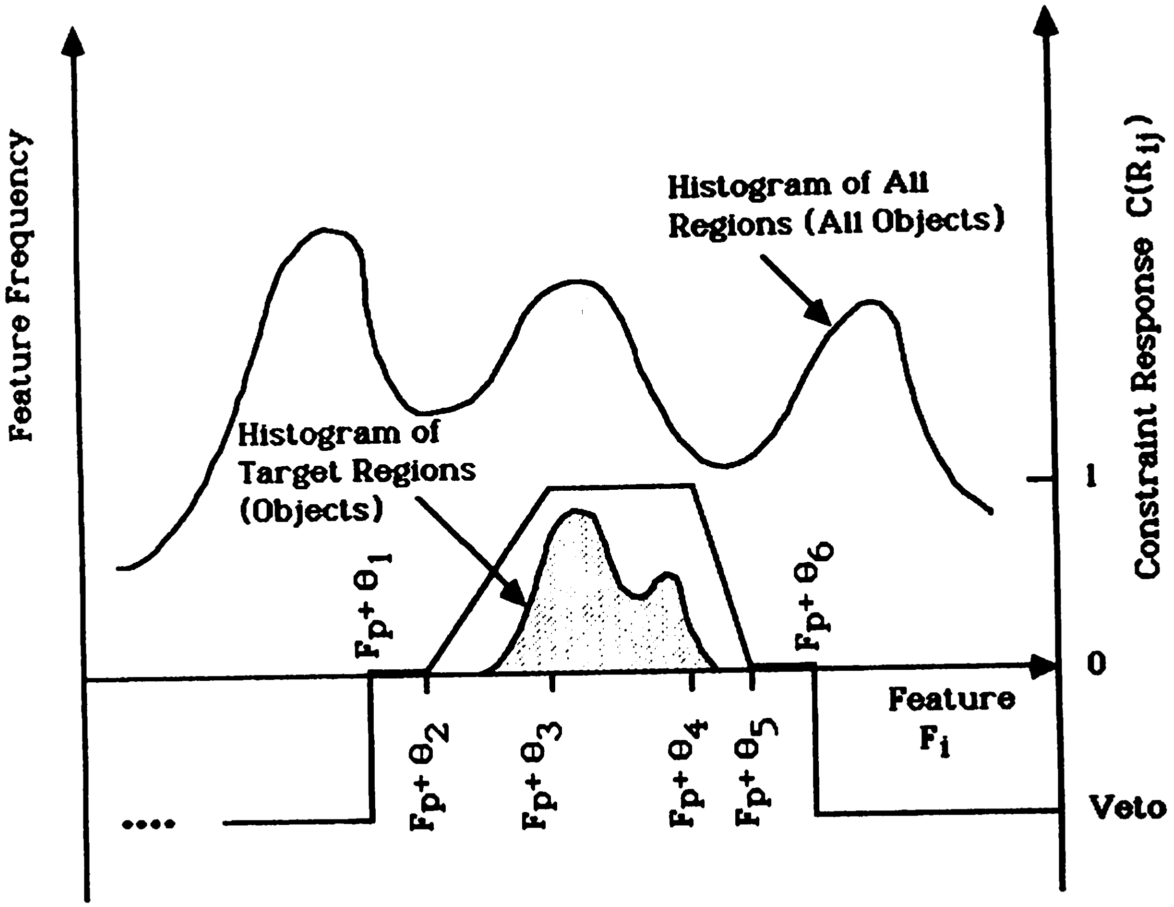


Figure 5. Image histogram of the "green-magenta" opponent color feature (2G-R-B). The global histogram of the feature across all pixels in eight hand-labelled images is shown unshaded. The intermediate diagonal shading represents the feature histogram of all grass regions in the eight images. The darkest cross-hatched shading represents the feature histogram of grass regions in a single image.



Feature Frequency
Histogram of All Regions (All Objects)
Histogram of Target Regions (Objects)
Constraint Response C(Rij)
1
0
Veto
Feature Fi
Fp+Θ1
Fp+Θ2
Fp+Θ3
Fp+Θ4
Fp+Θ5
Fp+Θ6
....

Figure 6. Simple Graded Constraint. The structure of a simple piecewise-linear constraint is shown for mapping a token attribute value R_{ij} into support for an object label hypothesis. Here, it is shown constructed on the basis of a prototype feature value f_p obtained from the combined histograms of labelled regions across image samples. The object-specific mapping is parameterized by seven values, $f_P, \theta_1, \ldots, \theta_6$ and stored in the knowledge network. Alternatively, instead of viewing the constraint in terms of a prototype point f_p in feature space, the construction of this constraint can be viewed as defining the ranges over which the feature values imply the possibility of the object with different degrees of confidence.

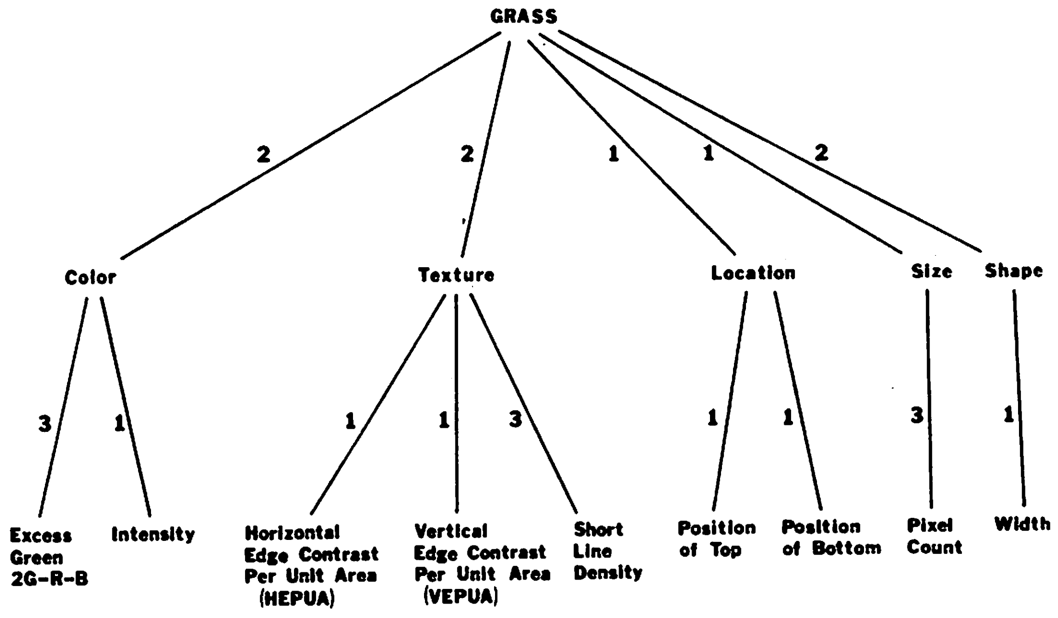


Figure 7. Structure of the compound constraint function for grass. The constraint value is the normalized weighted sum of the responses of five component constraints, each of which is in turn a normalized weighted sum of the responses from simple constraints associated with a single feature. Note that a weight could be 0, thereby allowing only the veto range for that feature to be propagated.

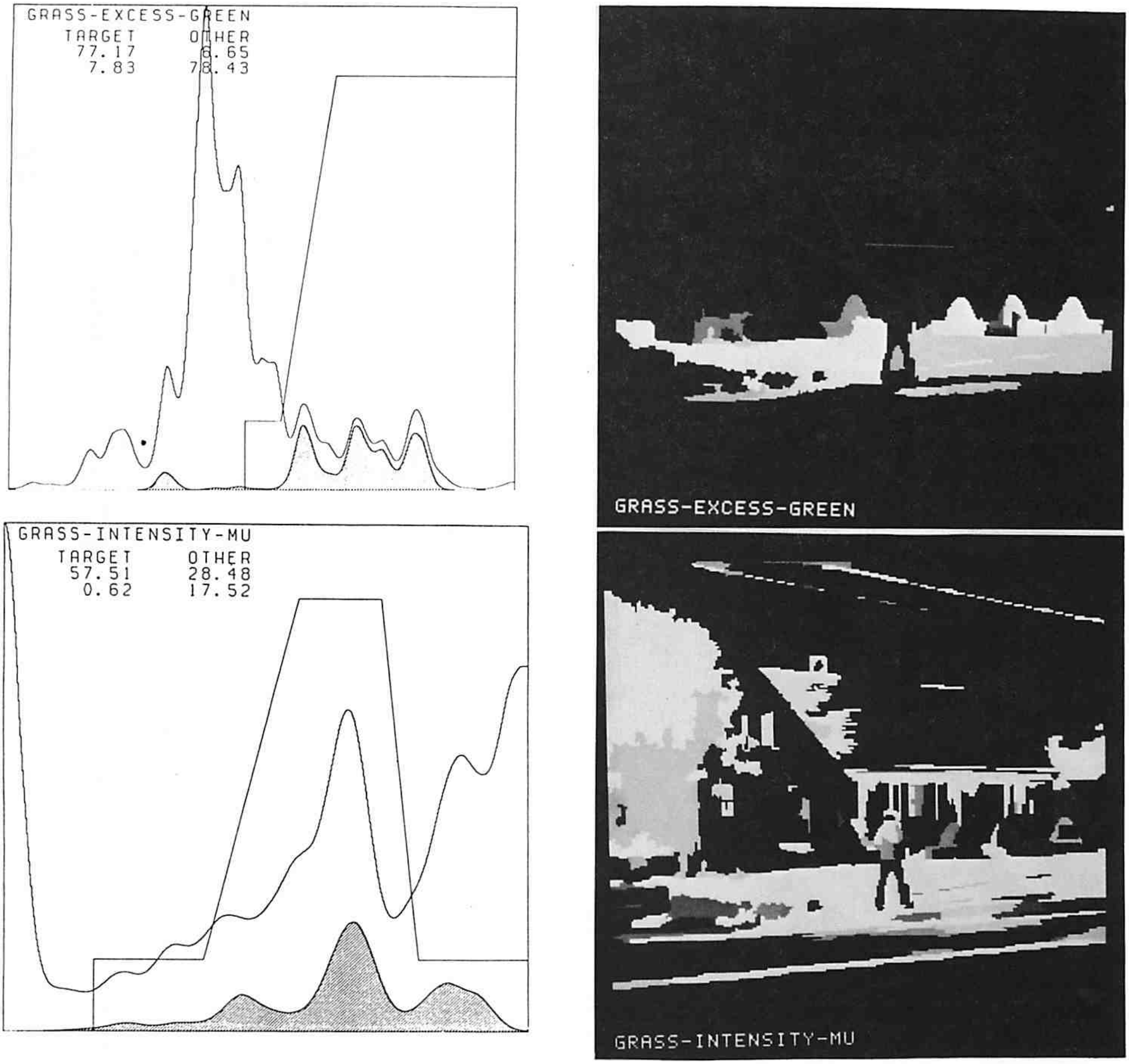


Figure 8. Results from selected simple constraint functions. In each image pair, the left image is a composite feature histogram showing the feature distribution across all pixels in a set of images (the unshaded curve), the distribution of grass pixels in the same set (the shaded curve), and the constraint function. The piecewise-linear constraint function, that is shown superimposed, can be interpreted by noting that the ramps vary over a [0,1] response range, and that the vertical discontinuity in the response delimits the VETO range. The right image shows the brightness-encoded strength of the constraint response when applied to all regions in the segmentation of Figure 3c; bright regions correspond to a high constraint response. Note that very dark regions denote both low confidence regions or vetoed regions. See text for a discussion of the four numbers in the upper left corner of the histograms.

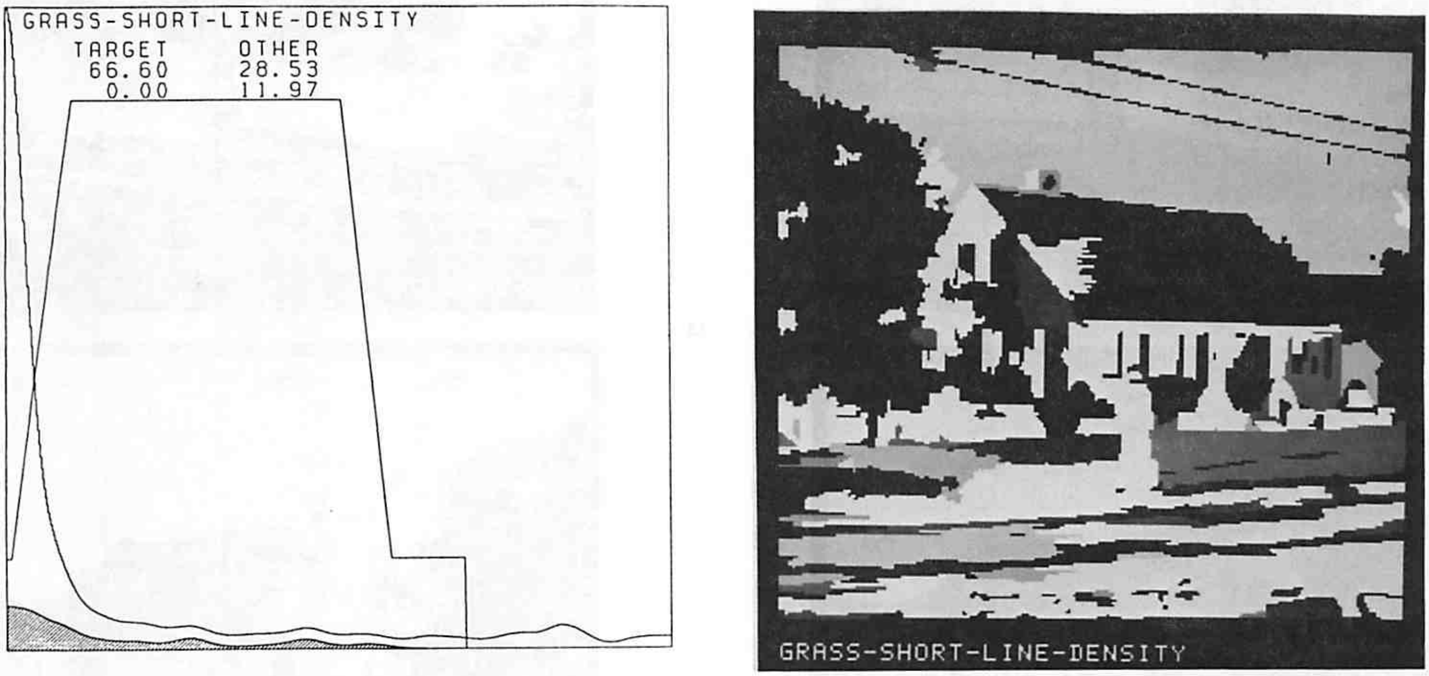


Figure 8, continued

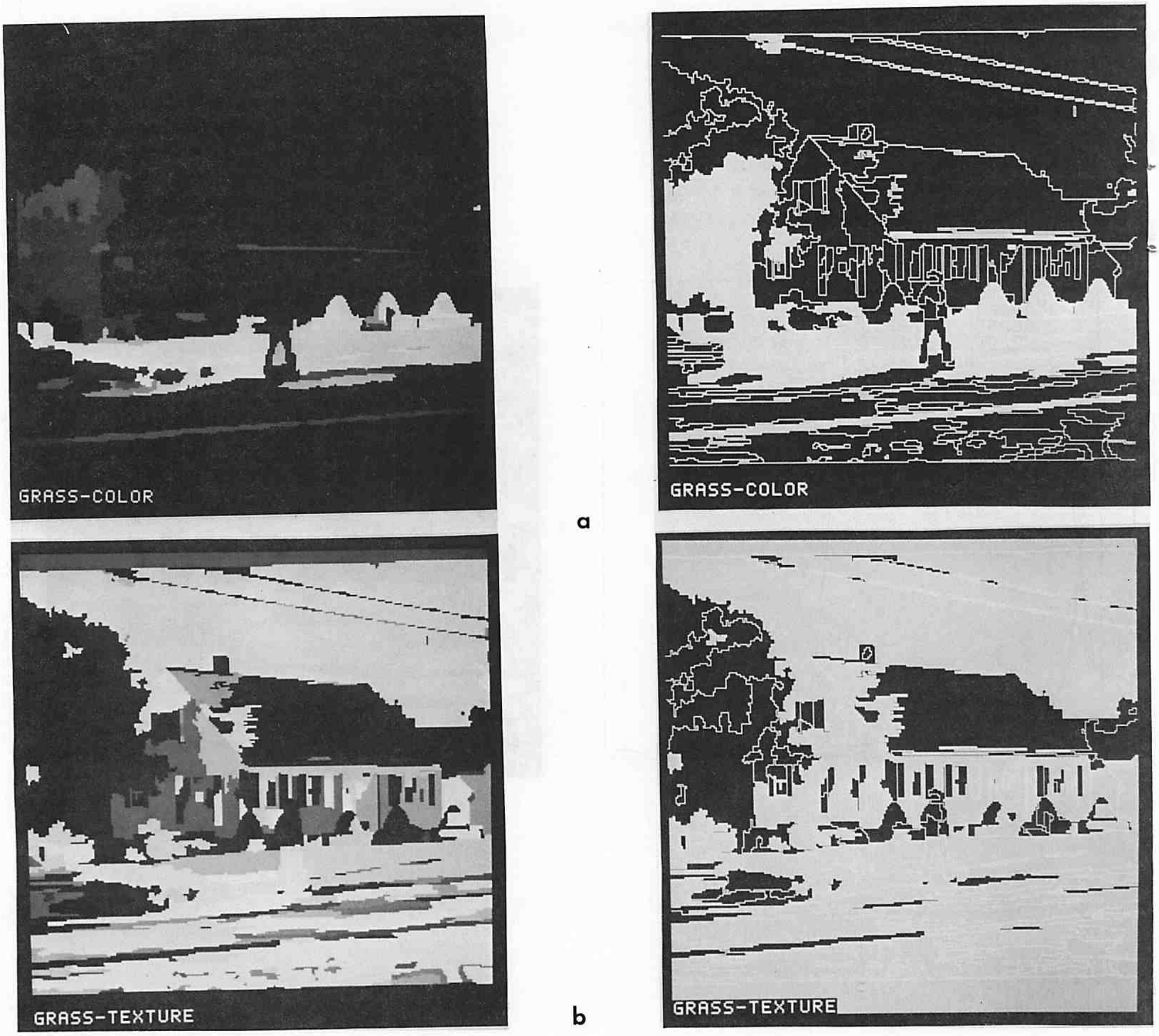


Figure 9. Example Results of Compound Constraints for Grass. The responses for two of the individual grass compound constraint and the overall compound response are shown. In each pair of images, the left image shows the brightness-encoded (bright ≡ high) constraint response. The right image shows regions vetoed by the constraint in black; all other non-vetoed regions in the right image are uniformly gray. (a) color compound constraint; (b) texture compound constraint; (c) final result from overall grass compound constraint.

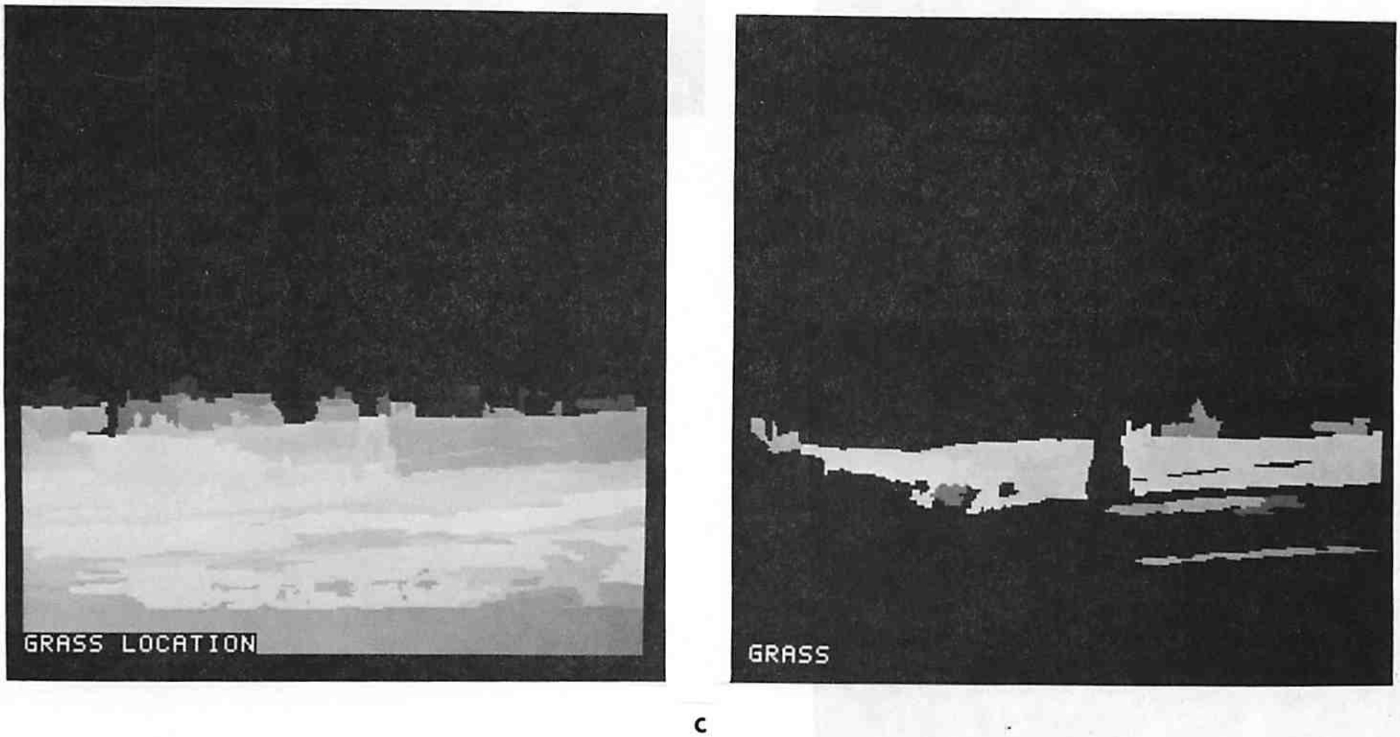


Figure 9, continued

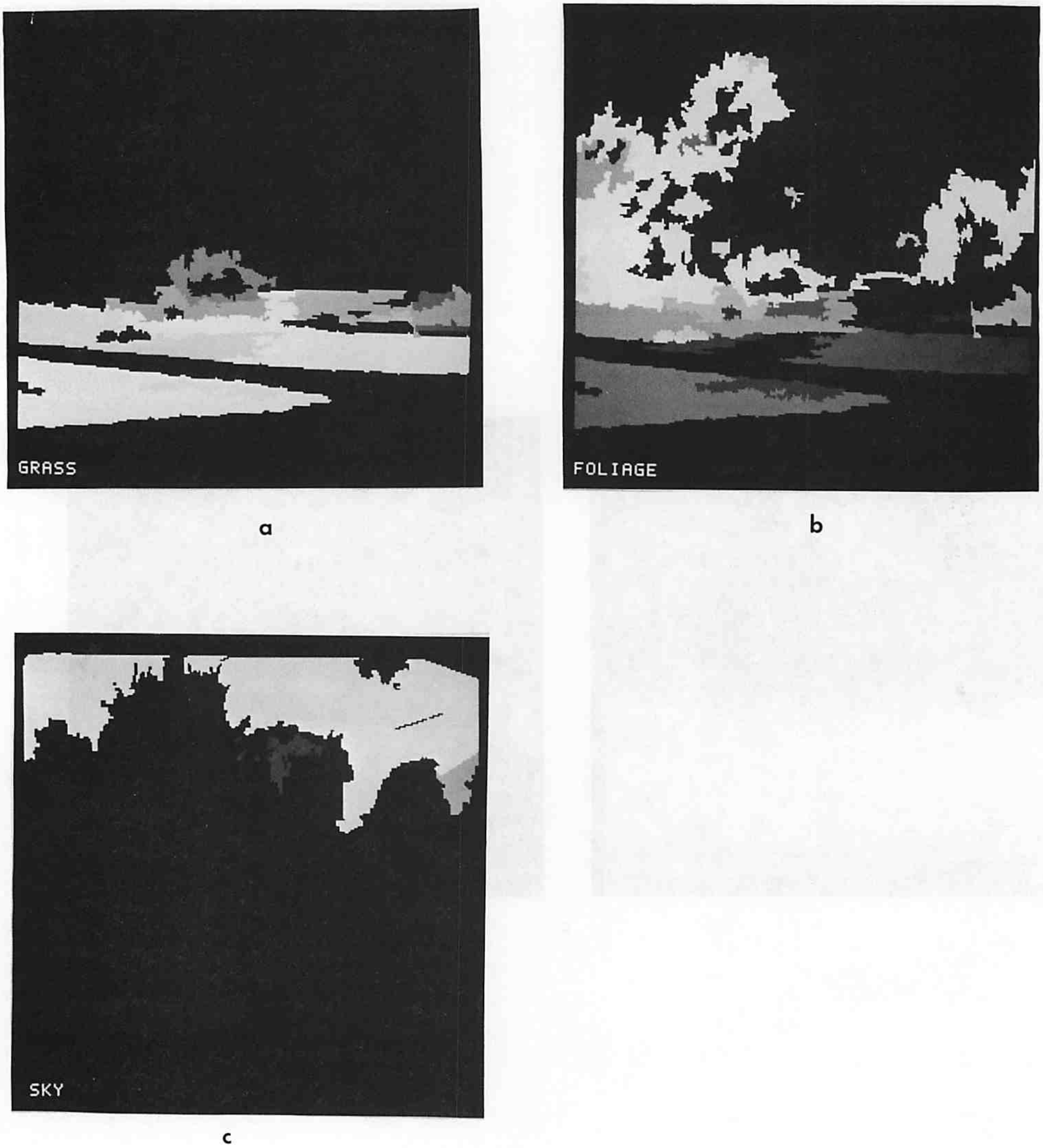


Figure 10. Example Results of Compound Constraints for Grass, Foliage, and Sky. Initial hypotheses from the compound constraints for grass, foliage, and sky rules, applied to Figure 3b, are encoded in brightness. (a) grass, (b) foliage, (c) sky.

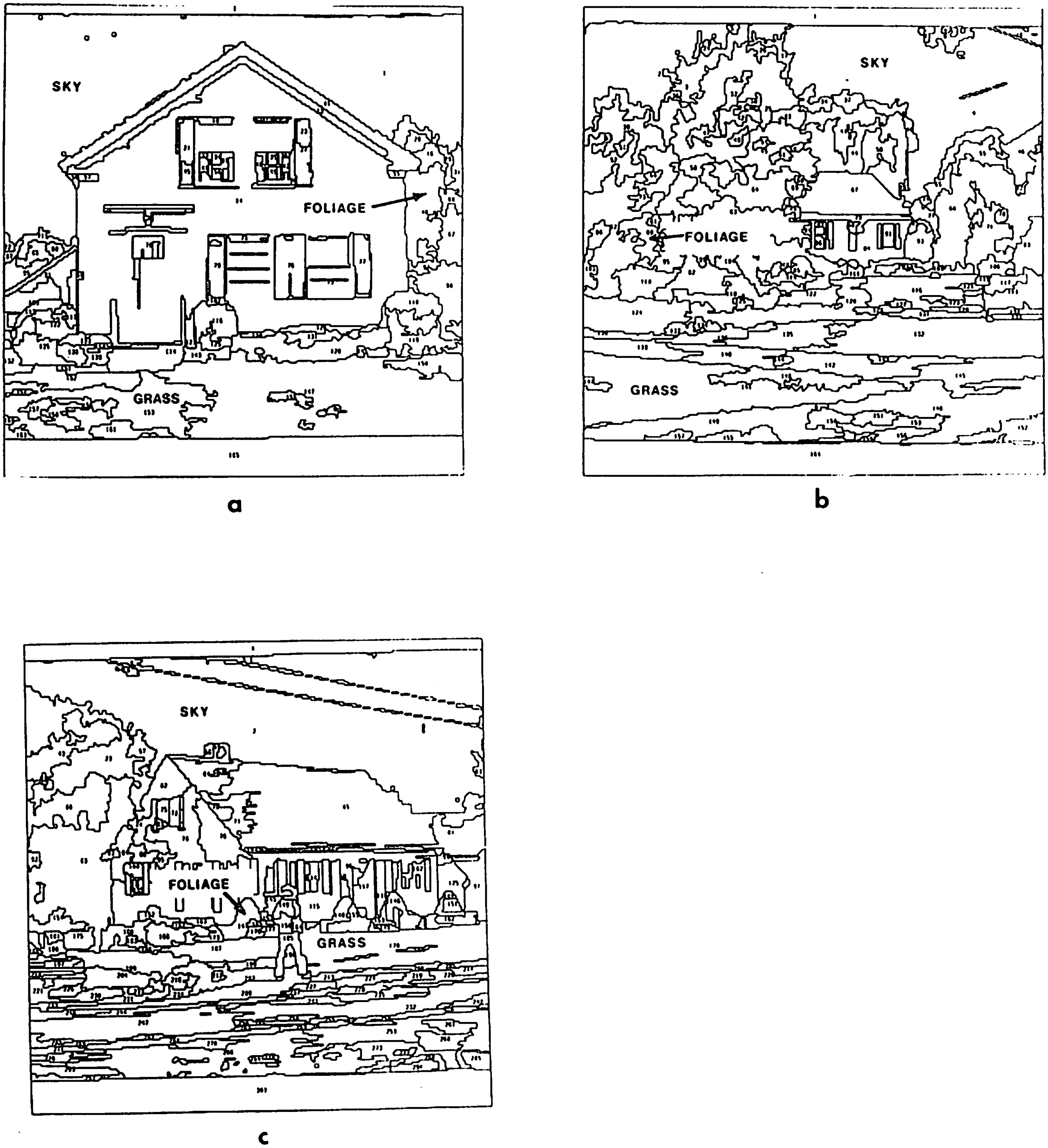


Figure 11. Highest ranked object hypotheses for sky, grass, and foliage for each of the three example images.

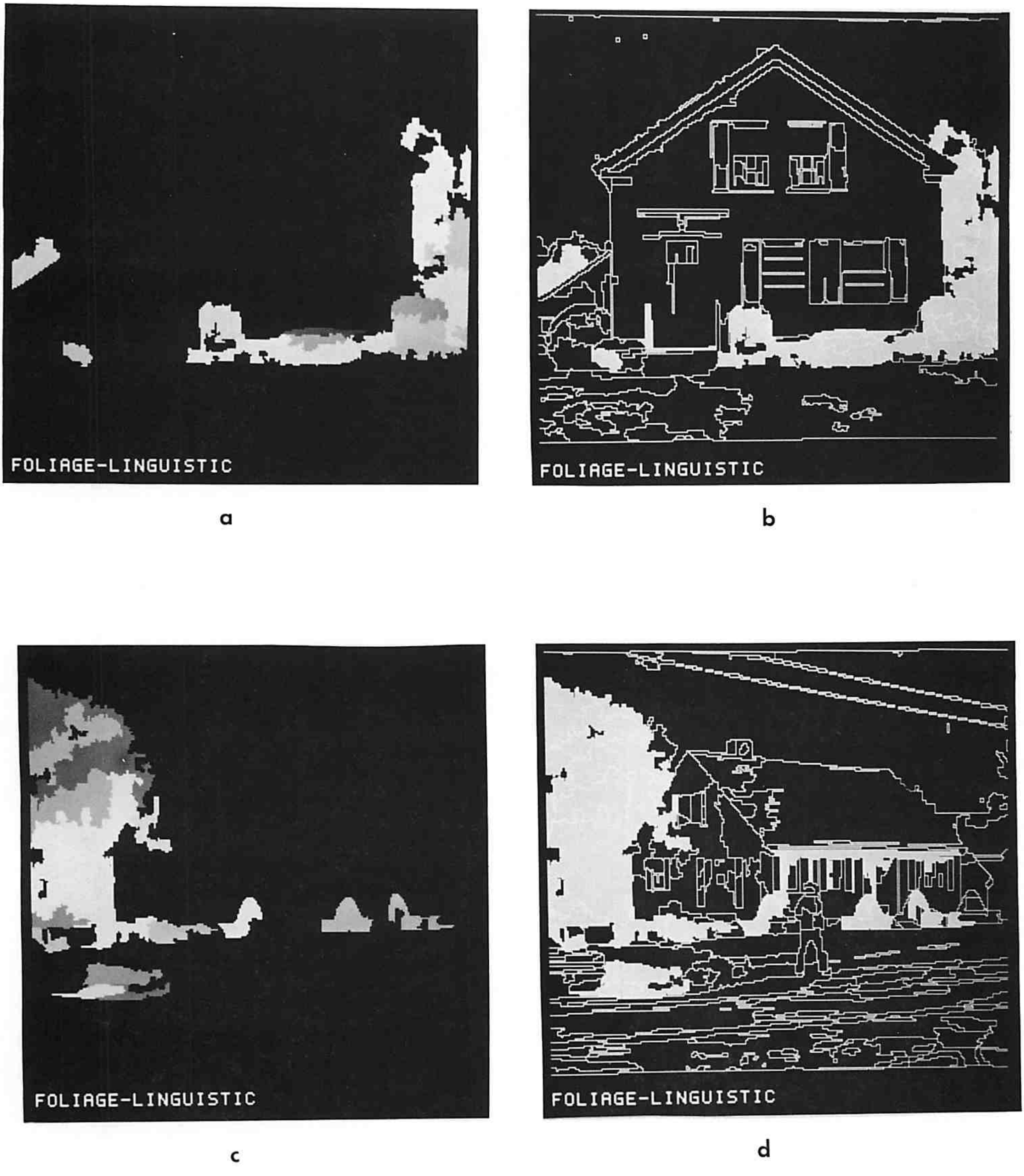


Figure 12. Example Results of Coarsely Specified Constraints via a Language Interface. Foliage results from the coarsely quantized constraints applied to the segmentations of Figure 3a and c.

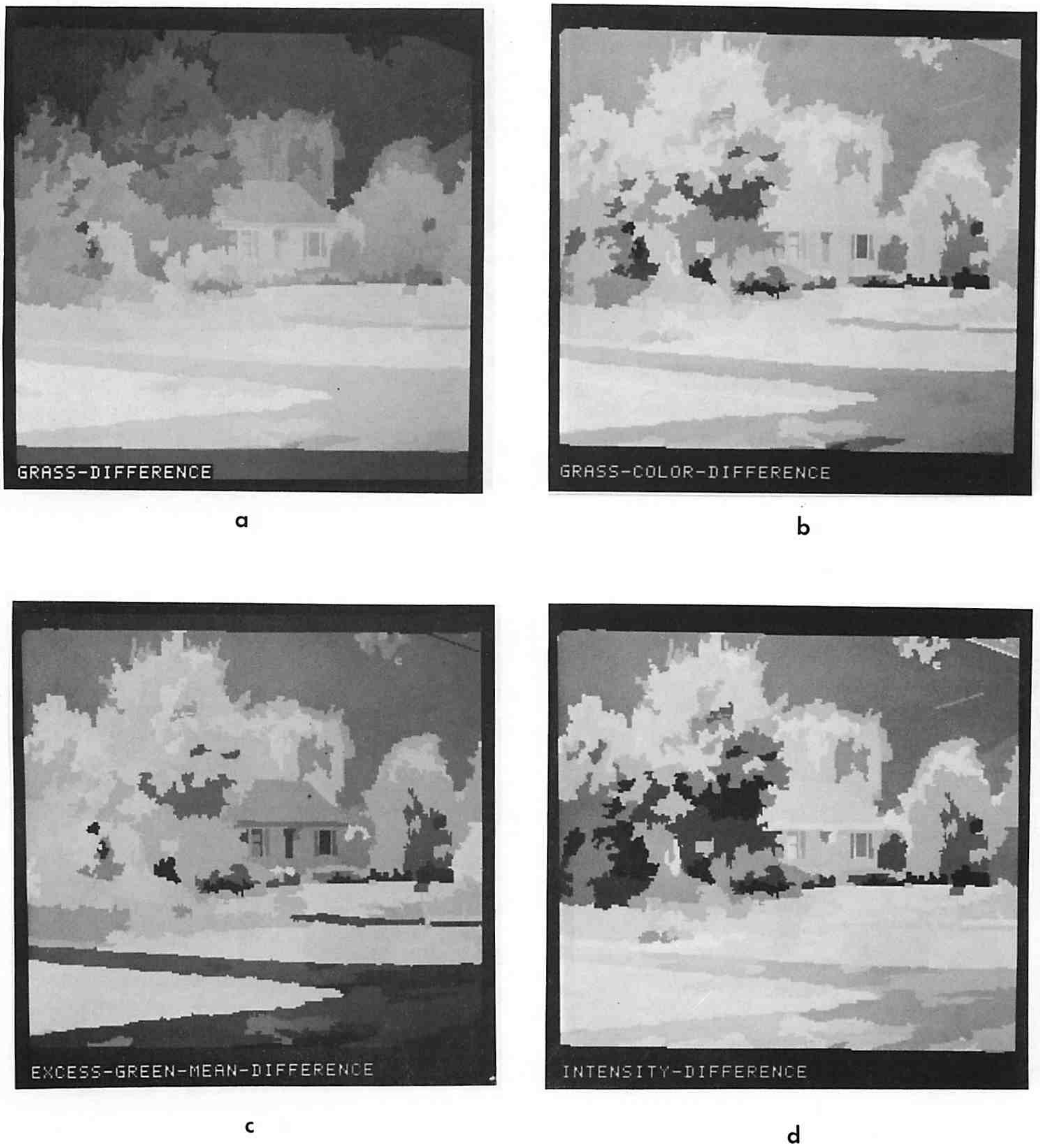


Figure 13. Similarity of Grass Exemplar Tokens for Exemplar Extension. In the scene of Figure 3b, the similarity of the grass exemplar regions token to all other regions tokens is displayed. (a) similarities from the compound object hypothesis constraint; (b) similarities from only the color component of the compound constraint; (c,d) similarities from the simple constraint associated with excess green and intensity, respectively. In all cases, similarity is encoded as brightness (i.e. small difference is bright). Note that subjectively the full compound similarity constraint appears to be more accurate, in particular in removing the foliage regions.

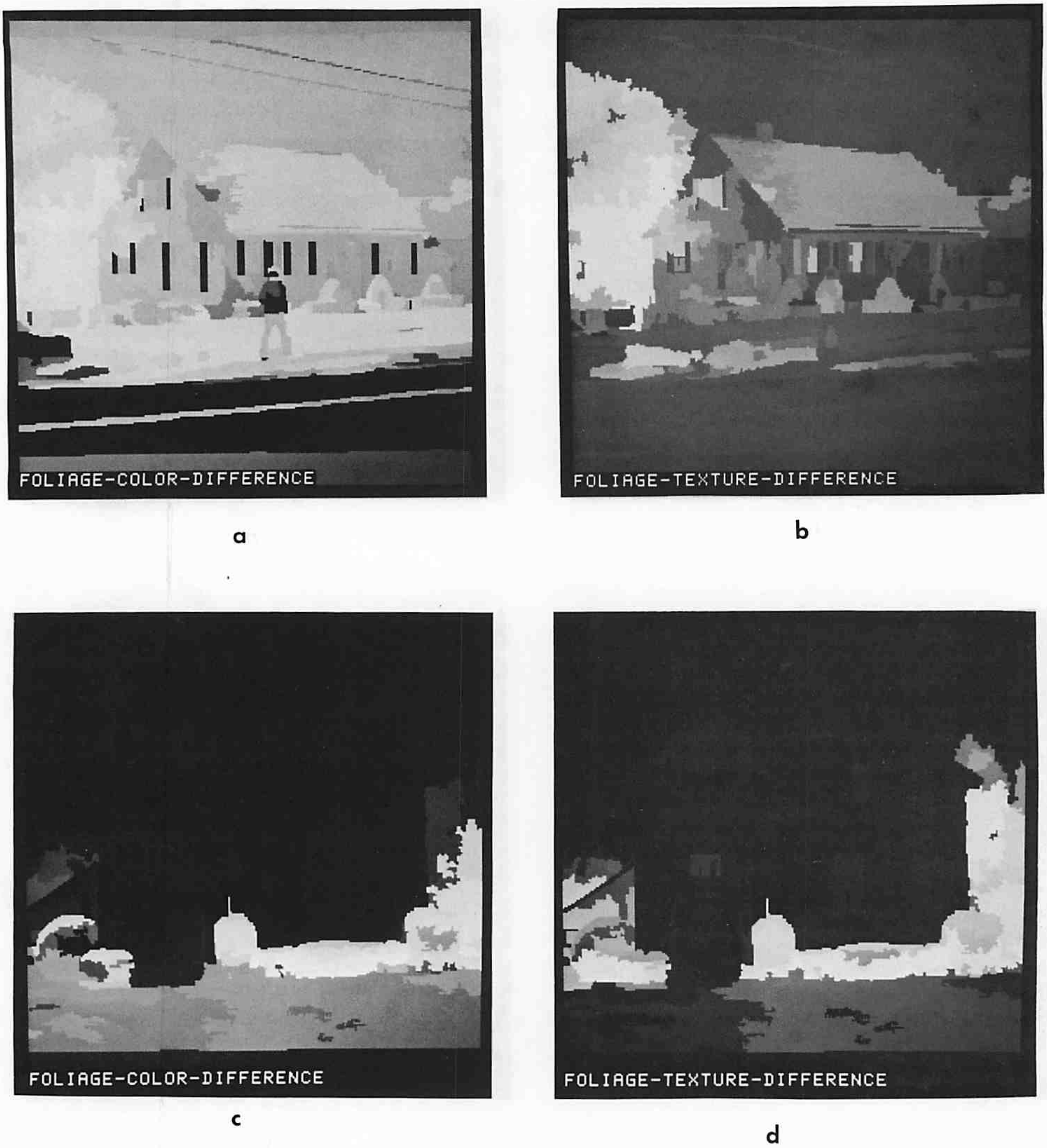


Figure 14. Similarity to Foliage Exemplar Tokens. (a,b) Color and texture similarity between image regions and the largest region in the tree area on the left side of the image (region 69 in Figure 3c) as the image-specific exemplar; (c,d) color and texture foliage matches using the large region in the low bushes in front of the house (region 128 in Figure 3a) as the image-specific exemplar. Note that brightness encodes similarity.