SRI International

Technical Note 463 • March 1989

RECOGNIZING OBJECTS IN A NATURAL ENVIRONMENT: A CONTEXTUAL VISION SYSTEM (CVS)

Prepared by:

Martin A. Fischler
Program Director

Thomas M. Strat
Senior Computer Scientist

Artificial Intelligence Center
Computing and Engineering Sciences Division

APPROVED FOR PUBLIC RELEASE
DISTRIBUTION UNLIMITED

Supported by the Defense Mapping Agency and the Defense Advanced Research Projects Agency (DARPA) under contracts MDA 903-86-C-0084 and DACA 76-85-C-004. To appear in the Proc. of the DARPA Image Understanding Workshop, 5/89.XS

333 Ravenswood Avenue • Menlo Park, CA 94025-3493 • (415) 326-6200 • FAX: (415) 326-5512 • Telex: 334486

RECOGNIZING OBJECTS IN A NATURAL ENVIRONMENT: A CONTEXTUAL VISION SYSTEM (CVS)

Martin A. Fischler and Thomas M. Strat
Artificial Intelligence Center
SRI International
333 Ravenswood Avenue
Menlo Park, California 94025

Abstract

Existing machine vision techniques are not competent to reliably recognize objects in unconstrained views of natural scenes. In this paper we identify a number of weaknesses in current recognition systems, including an inability to solve the partitioning problem or to effectively use context and other types of knowledge beyond that of immediate object appearance. We propose specific mechanisms for dealing with some of these problems and describe the design of a vision system that incorporates these new mechanisms. The system has been partially implemented and we include some experimental results indicative of its operation and performance.[1]

1 Introduction

Most existing work in robotic vision has focused on issues directly related to the immediate geometry and photometry of the scene and the imaging process, such as recovering the three-dimensional location and orientation of scene surfaces from image data and recognizing objects by their geometric shape or by the presence or absence of some prespecified collection of easily measured attributes (e.g., spectral reflectance, texture, and area). On the other hand, most real-world objects are assigned names based largely on their origin, function, purpose, or context. For example, we recognize an object as a bridge not because it has a specific shape, intensity, or texture, but rather because it links the two sections of a road interrupted by an intervening body of water.

This research effort is concerned with the design of a perception system that not only recovers shape but also achieves a physical and semantic understanding of a scene. Such a development would form a link between the low-level, quantitative geometric information currently produced by vision systems and the stylized, symbolic representations that are the mainstay of formal reasoning and planning programs. Until this competence is available, it will remain impossible to develop robots that react intelligently and flexibly to natural environments.

Completely duplicating the human ability to recognize objects is probably equivalent to duplicating human intelligence. An intermediate goal would be to recognize the objects that can occur in completely natural settings (i.e., no human artifacts). Aside from only having to deal with a restricted class of objects and phenomena (probably fewer than 1000 distinct entities), purpose and intent are reduced to recognizing the requirements of vegetation for light, nourishment, and space, as well as the constraints physical forces impose on both living objects and terrain formations. This particular recognition domain derives special interest and importance from the fact that all living creatures must contend with this world and, at least in part, solve this particular recognition problem. Since human recognition abilities evolved in this context, it is reasonable to assume that this domain is extendable (an attribute missing from most other limited "worlds" chosen in the past for the exploration of machine recognition).

[1]Supported by the Defense Mapping Agency and the Defense Advanced Research Projects Agency under contracts MDA903-86-C-0084 and DACA76-85-C-0004

To appear in the Proc. of the DARPA
Image Understanding Workshop, 5/89.

How difficult is the limited problem of recognition in the natural outdoor world? It is safe to say that few, if any, of the relevant objects can be reliably recognized with existing approaches; even worse, there currently is no credible story about how to proceed. For example, how do we recognize rocks in photographs, what criteria/procedure can be offered to a person who has never seen a rock before (other than offering the advice that after everything else is labeled, the remaining objects are rocks)? What is it that allows us to distinguish a river from a lake, a puddle, or a "run-off" channel, or standing water on a leaf, even after we have locally identified the visible surfaces in a photo that correspond to water?

The investigation we are conducting has a very practical ultimate goal: to enable autonomous vehicles to carry out meaningful work. The research itself has four components:

- To develop an approach that addresses the fundamental limitations of today's systems and that could serve as an architecture for a computational vision system that achieves both a geometric and a semantic understanding of its environment.
- To choose a vocabulary for describing the outdoor world. This must include representations for describing a wide variety of shapes and physical attributes, as well as constructs for expressing semantic properties and relations among objects.
- To develop a system that can recognize instances of this vocabulary in imagery from an outdoor domain. This will involve building a knowledge base and suitable control structures to enable the system to recognize natural objects in that domain.
- To demonstrate the feasibility of the approach by implementing a significant portion of the proposed architecture and performing experiments on representative imagery in the context of an autonomous robot.

Our intent in this paper is to frame the recognition problem for natural scenes, offer a reasonable story about how it can be solved, and describe the progress we have made in building a vision system for this purpose.

2 Framing the Problem

What is it that we want to recognize? What new and prior information do we require (e.g., with respect to sensory input – a single photo, a sequence of photos, stereo or range information, passive or active sensing, etc.)?

We almost invariably recognize things based on partial information. For example, in a photo, we never see the "backs" of the objects we recognize, but rather assume that they are there. If we recognize a tree in a photo, we might actually see the trunk, or some branches and leaves, but typically (e.g., in a forest scene), even if most of the "parts" are visible, we couldn't correctly assign all the leaves to the correct trunk. Thus, even though we know that trees have trunks, branches, leaves, and roots, it is generally their parts that we recognize and can localize; we then deduce the presence of the whole. It is quite possible that some of the reasonably intelligent animals that live in forested terrain do not partition the world the way we do, and may not even have a concept corresponding to the coherent entity we call a tree. Thus, questions we must address include: what is the vocabulary of objects we want to be able to recognize; and, to what extent should we be able to delineate a recognized object?

What is it that we extract from image data that allows us to recognize natural objects? What knowledge must we have in advance? Edges, contours, color, and texture are far from sufficient cues for recognizing a river, and it is not clear what it is that allows us to recognize a rock. In a sense, rivers are shaped

and located by their environment, and it is difficult, if not impossible, to recognize them out of context – i.e., without simultaneously recognizing both the geometry of the surrounding terrain and the co-occurring natural objects. Trees, on the other hand, are shaped by a genetic "blueprint"; they satisfy a set of constraints that are imposed on, rather than derived from, their surroundings. One can probably recognize a line drawing of a tree in isolation, but not of a river. Rocks are intermediate between rivers and trees. They can often be recognized in isolation, but not from a line drawing. They are positioned by their immediate environment, hut not shaped by it; they do not conform to a genetic blueprint, but their appearance must reflect a rather narrow set of formative conditions [10].

As a means of extracting a manageable subproblem for our initial effort, the implemented system is intended to recognize a variety of scene components, but will primarily focus on the task of recognizing trees. Given an image of terrain obtained by an autonomous vehicle, and a crude map of the area, the system attempts to identify the obvious trees. We wish to focus on recognition, and will not be unduly concerned with positional accuracy. Objects that are mislabeled are the primary errors that we wish to avoid – mislabelings are the type of "ugly" mistakes that plague today's vision systems.

3 Ingredients of a Solution

The realization of robust recognition in the natural outdoor world will require that four current limitations of machine vision be overcome: the almost exclusive reliance upon shape, the ill-defined nature of the partitioning problem, the lack of an effective way to use context, and the inability to control the growth and complexity of the recognition search space.

The role of geometry – In most existing approaches to machine recognition, the shape of an object or of its parts has been the central issue. Indeed, many artifacts of human technology can be recognized solely on the basis of shape, and to a large degree this fact accounts for the limited success so far achieved by machine recognition systems [2, 4, 12, 13]. These techniques cannot be extended to the natural world because shape alone is insufficient (even for people) to recognize most objects of interest (e.g., a rock or a river). It is easy to recognize a line drawing of an isolated telephone, but, as previously discussed, doubtful that one could correctly classify a river based upon a line drawing of the river alone. Indeed, most natural objects fail this "line drawing test" – a test that requires identification based solely on shape. The fact that very few natural objects have compact shape descriptions further complicates the use of shape in describing natural scenes. Thus a rather complex and cumbersome description would be required to describe the shape of something as common as a tree or a bush. It is obvious that shape cannot be the sole basis for a general-purpose recognition system.

The role of scene partitioning – A common paradigm in machine vision has been to partition an image into distinct regions that are uniform in intensity, texture, or some other easily computed attribute, and then to assign labels to each such region. For natural scenes, however, it is seldom possible to establish complete boundaries between objects of interest. We have already mentioned the difficulty of associating leaves with the correct trees. Other examples are abundant – where does a trunk end and a branch begin, what are the boundaries of a forest, is a partially exposed root part of the ground or the tree? Despite these difficulties, it remains necessary to perform some form of partitioning to do recognition, otherwise we have nothing to refer to when making a classification. Because of the impossibility of performing scene partitioning reliably (even if such a goal were well-defined), we cannot rely on partitioning in the usual sense. Instead, we need an alternative view that allows object recognition without requiring complete or precise object delineation.

The use of contextual information – It is widely known that an object's setting can strongly influence how that object is recognized, what it is recognized as, and if it is recognizable at all. Psychological

studies have shown that people cannot understand a scene in the absence of sufficient context, yet when such contextual information is present, recognition is unequivocal [9]. Individual objects may project as a multitude of appearances under different imaging conditions, and many different objects may have the same image. From these studies it is clear that computational vision systems will be unable to classify an object competently using only local information. A perceptual system must have the ability to represent and use non-local information, to access a large store of knowledge about the geometric and physical properties of the world, and to use that information in the course of recognition. However, the few computational vision systems that make use of context do so superficially or in severely restricted ways. In our work, we make the use of contextual information a central issue, and explicitly design a system to identify and use context as an integral part of recognition.

A mechanism for control of complexity – The two standard architectures currently employed in scene analysis are (1) top-down or model-driven interpretation, and (2) bottom-up, a breadth-first form of sensor-driven interpretation. Both of these forms of image analysis (and various intermediate versions) lack an essential attribute of intelligent behavior – an explicit mechanism for generating a (potential or actual) problem solution without requiring some form of exhaustive search. In controlled or simple environments, exhaustive search may be computationally feasible, but the complexity of the natural world imposes the requirement for a more efficient solution mechanism. A key aspect of our approach is the provision of an explicit mechanism for generating high-quality assertions about the scene without the need for exhaustive search.

4 The Shape of a Solution

An effective scene-recognition system must provide:

- A "language" for making semantically meaningful assertions about a scene.
- Computationally effective procedures for generating and evaluating hypotheses (assertions) about scene objects and relations.
- Methods that permit validation of a scene description at a global level.

The key ideas underlying our approach are:

1. The selection of the outdoor navigation and mapping tasks as the basis for delimiting the semantic knowledge the system must be able to employ in describing and understanding a scene (some relevant vocabulary items are listed in Table 1).

geometric horizon	sky	ground
skyline	thin vertical raised object	occluding edge
raised object	foliage	tree trunk
tree crown	bush	tree
cloud	ridge line	branch

Table 1: Vocabulary of terms

2. Explicitly encoding the different sets of conditions (contexts) under which the semantic objects we are interested in can be reliably compared and recognized. These recognition conditions, which typically require the prior or co-determination of the presence of other semantic features, are called *context sets*. Some of the criteria that comprise context sets are listed in Table 2. High-level knowledge must be brought to bear if recognition is to be performed competently. As we will see, context sets are the central knowledge source around which our recognition system is constructed.

GLOBAL CONTEXT:	
	foothills on San Francisco peninsula
	daytime
	cloudless sky
LOCATION:	
	local topography
	touching the ground
	coincident with a known tree
APPEARANCE:	
	geometric: shape, size, neighbors, depth, orientation
	photometric: intensity, texture, color
FUNCTIONALITY:	
	supporting another object
	bridging a stream
	counterbalancing a tree limb

Table 2: Examples of criteria that comprise context sets

3. The design of a control structure to limit the computational complexity of the recognition process: (a) control of the quality of hypothesis generation through a biased form of image partitioning, i.e., the use of region delineation based on a context-set-adjusted homogeneity metric tailored to find portions of an image associated with specific scene features; and (b) control of search (in recognition space) by creating partial orderings of all of the hypotheses making assertions about the presence of a particular type of scene feature.

4. Recognition is accomplished by finding mutually consistent groupings (*cliques*) of hypotheses about the different semantic features. Cliques are formed by sequentially adding the highest ranking hypotheses (with respect to the partial orderings) that are consistent with the current global interpretation.

5. Physical and imaging constraints, in the context of a 3-D spatial model, are employed as the primary criteria for the consistency of a global interpretation (rather than probabilities for different arrangements of semantic objects).

4.1 Scene Semantics

There are probably on the order of 100 to 1000 semantic object classes that an organism (such as a rabbit or a robot) needs to be able to identify in order to move safely about in a relatively benign environment (such as the grass- and tree-covered rolling hills in the San Francisco Bay area), but for our initial experiments we will focus on finding trees and will recognize other scene features only to the extent of supporting this primary objective (these additional scene features include sky, ground, thin raised objects, shadows, occlusion edges, ...). As indicated earlier, many of the naively chosen object classes we might deem appropriate for describing

a natural scene require complex reasoning about those parts of the object that are directly observable; thus, in our first experiments, we will actually look for vertically oriented tree trunks rather than actual trees. The determination of a semantic vocabulary that matches the recognition capabilities of a given vision system and satisfies the requirements of a specified task is a non-trivial issue that we are investigating but will not further address in this paper.

4.2 Context Sets

Computer vision presents the following paradox: For natural scenes, in order to recognize an object, its surroundings must first be recognized, but in order to recognize the surroundings the scene objects must often be recognized first. Is it really necessary to recognize everything at once, or can some things be recognized in relative isolation? If so, what are they, and how can they be recognized? It is certainly the case that one does not have to know about everything in the universe to recognize (say) a tree; on the other hand, trying to find the trees in an image by looking through a small "peep-hole" may be impossible. A critical aspect of our approach is to define explicitly sets of conditions (context sets) under which various semantic objects (such as trees) can be recognized – or at least ranked (in terms of their "tree-ness"). Typical conditions employed in context sets are presented in Table 2. Our approach makes use of several forms of context:

- *Global context* – Trees differ greatly from one climatic zone to another. Knowing whether the image is of an arctic or tropical environment should affect perceptual strategies.
- *Location* – An object flying through the air is more likely a bird than a rabbit – and is almost certainly not a tree. These facts can be deduced from context alone, without any reference to an object's shape, color, or texture.
- *Appearance* – Once a tree has been recognized, it is easy to recognize others, because neighboring trees are often similar in appearance.
- *Functionality* – A tree trunk that has fallen across a stream can act as a bridge, but can be identified as a bridge only after recognition of the stream.

There is a collection of context sets for each semantic category in the vocabulary and each context set is associated with exactly one category. The system employs three kinds of context sets:

(1) A *hypothesis generation context set* is used to generate candidate hypotheses for the category it is associated with. A hypothesis is an assertion that, for example, a certain specific region in an image corresponds to a tree trunk. Each context set contains the conditions under which a computational process (*operator*) should he employed and supplies the particular parameter settings that should be used in that context. The operator is only invoked when the conditions of the context set are satisfied. All hypothesis generation context sets whose conditions are satsified are used to generate candidates.

(2) A *hypothesis validation context set* is used to rule out certain candidate hypotheses from further consideration as instances of a class. It specifies a set of conditions such that if the conditions are not satisfied, the hypothesis could not denote a member of the class. For example, an object whose diameter is greater than fifty feet could not be a tree trunk. These context sets eliminate the obviously erroneous hypotheses that might have slipped through the earlier processing.

(3) A *hypothesis ordering context set* is used to rank order a pair of candidate hypotheses. It contains conditions under which one hypothesis should be preferred over another as an instance of the class associated with the context set. If any hypothesis ordering context set asserts a preference between two hypotheses, that preference is recorded. Otherwise, no preference is established. Pairwise comparison of all candidate hypotheses for a class results in a partial ordering of hypotheses for that class.

4.3 Control of Complexity: Hypothesis Generation and Ordering

Perception is appropriately considered a form of intelligent behavior because it nominally involves searching an infinite "description space" to find an appropriate model of some imaged environment that conceivably could have any one of an infinite number of arrangements of its components. Most current vision systems avoid addressing this central problem by restricting the dimensionality of the relevant search space, for example, by dealing exclusively with controlled environments where only a few, completely modeled, objects can occur. That is, we have complete parameterized descriptions of the environment in advance; our task is to find appropriate values for a relatively small number of numeric parameters. Because such systems have no provision for handling the combinatorics of complex natural environments, they are not capable of "scaling up" to real-world scenes.

It is apparent that a central consideration in the design of a "real world" vision system is the development of some mechanism for controlling the quality (and thus the quantity) of the assertions/hypotheses that must be filtered by the system. We deal with this problem by the use of prior knowledge, by context-set-controlled partitioning, and by quality-based ordering of the generated hypothesis for preferential use in constructing global descriptions.

4.3.1 The Use of Prior Knowledge

We envision the system we are constructing as being part of an autonomous robot situated in the world, so a great deal of information will be available to it. Furthermore, the spatial and temporal continuity of the world leads to added opportunities for a robot vision system to improve its performance based on its experience.

The following facts are known by reasonably intelligent animals and are expected to be immediately available to a robot as well. The approach we are taking is designed to make use of these items of information:

- *Height of viewer above the ground:* An animal knows its orientation intrinsically because its eyes are fixed in relation to a given configuration of its body (assuming its feet are on the ground). This value is fixed (or easily computed) for a robot as well.
- *Orientation of the viewer (with respect to gravity):* An animal knows this by virtue of its sense of balance in the inner ear. A robot could know this by employing a simple sensor.
- *Vertical vanishing point:* Knowing the orientation of the viewer with respect to gravity allows one to compute the vertical vanishing point in the image plane.
- *A depth image:* Through stereopsis, an animal can estimate the distance to the things that it sees, although the quality, density of coverage, and accuracy are debatable, and the precision available to it degrades with distance. A robot might employ several means to acquire depth data. Binocular stereo techniques provide depths with various qualities. Laser rangefinders provide dense depth images but have other limitations. Other techniques (motion, texture, shading) and sensors (acoustic, structured light) are also available.
- *The ground (at least one point in an image that can be labeled as ground):* An animal can identify at least one point that is known to be on the ground by looking at its feet (assuming the animal is currently standing on the ground). A robot could perform the same action to locate a ground point.
- *The sky (at least one point in an image that can be labeled as sky):* An animal can find at least one point that is known to be sky by looking straight up (assuming it is not currently underneath a tree, in a cave, etc.) A robot could do the same.

– *Experience:* Unless the terrain is totally new, an animal has knowledge of the relative locations and appearance of some features and can predict the appearance of parts of a scene. While the vocabulary and representation of this information remain research issues, a robot should store and use this type of information as well.

4.3.2 Context-Set-Controlled Partitioning and Hypothesis Ordering

One of the unusual aspects of our approach is its organization. Over the years, many different strategies have been proposed for extracting information from images. It has become apparent that no single general-purpose strategy will suffice to address the many requirements placed on vision systems. Our approach is designed to choose low-level operations based on the semantics and context of each visual task. Rather than applying a fixed operator in all circumstances, we make strong use of semantics to determine the most appropriate low-level processing.

The underlying structure of the design is motivated by the following key observation:

> Within any given image, there is usually a relatively straightforward technique that will find the object or feature of interest.

Of course, that particular technique may fail miserably when applied to some other image. However, given a collection of simple techniques, one can generate a number of candidate features that will nearly always contain the "correct" interpretation. It then remains to choose that feature from among all the candidates. To accomplish this, we make use of a novel scheme:

> First, we bias the feature (hypothesis) generation process to be more responsive to generating good hypotheses relevant to a specific semantic category. Then we compare the candidates pairwise to find instances that are clearly "better" examples of the given class than their opponents. Repeating this comparison process imposes a partial order on the candidate set for each label of interest.

This departs from conventional approaches in two ways. First, comparing two candidates for a given label requires knowledge of the semantics of that label only, whereas the customary approach of comparing two labels for a given region requires knowledge of the relationships between many semantic categories. This orientation provides a basis for believing that sufficient knowledge might be encoded in the system to allow robust comparison. Second, we enforce the condition that the comparisons lead to a determination only if one candidate is clearly a better choice than the other. With this conservative approach, we hope to avoid the "ugly mistake" of misclassifying a candidate based on too little information.

The actual hypotheses generated by the system are motivated and shaped by the context sets in a manner somewhat analogous to a collection of production rules. Separate partitions of an image are created for each relevant semantic object class (e.g., sky, ground, trees ...), and in some cases, for each specific context set. The basic mechanism for creating a partition is an operator that assigns to each pixel in the image a homogeneity score which is a measure of its maximum "feature space distance" from any of its immediate neighbors. The components of the feature space, and their relative weightings, are specified by the currently active context set.

The output of the homogeneity operator is a gray-scale overlay of the original image that can be thresholded (under context set control) to form a binary image in which the "zero points" either are discontinuities or correspond to some feature other than the one associated with the context set; the "one points" are accumulated into coherent regions that are indicative of the presence of the feature of interest (see Figures

1 – 4). The context set may require that other generic operations be performed on the detected regions; e.g., "skeleton" generation as required for tree detection.

In actual practice, the context sets have varying degrees of specificity and precision. The more general and low-resolution context sets will operate first (because their conditions are more easily satisfied), and they typically use the output of the generic image operators with some standard set of parameter values. Thus, it is only necessary to run a few generic operators over the image once in forming an initial scene model. (Table 3 describes the collection of generic operators we currently employ.) Specialized operators and generic operators with customized settings are typically required by the context sets that can only be invoked after the initial model is formed – these context sets make more precise assertions about the scene, or deal with known exceptions to the general rules for finding the relevant object instances. For example, the system might require a special context set to recognize a lightning-struck tree that is known to be present, but no longer looks like a normal tree.

TEXTURE	– Measures the log-variance in a small window
STRIATION	– Measures the orientation and strength of an oriented pattern
THRESHOLD	– Creates a binary mask
SEGMENTATION	– Partitions an image
ASSOCIATION	– Groups contiguous clusters in a binary mask
EDGE DETECTION	– Finds discontinuities
LINEAR DELINEATION	– Finds line-like structure
HOMOGENEITY	– Measures the maximum feature-space difference between pixels in a small neighborhood

Table 3: Operators

The system constructs a global scene model by incrementally assembling mutually consistent collections of hypotheses (generated by the process just described). The hypothesis generation context sets are used to assure that only reasonable assertions are made; the hypothesis ordering context sets are used to provide an ordering for selecting the most likely hypotheses to add to the model. There is a collection of context sets for each class of objects the system is to recognize. If any one of them has a preference during a comparison of two hypotheses, then that preference is taken into account by incorporating it in the partial order. If no context set can establish a preference, or if several context sets disagree (which ideally should never happen and would require a modification of the context sets involved), then no preference is recorded for that comparison. In this way, a partial order is constructed from all hypotheses for each object class.

4.4 Constructing a Global Scene Model

Once partial orders have been constructed for the labels of interest, and for the additional labels that appear as terms in the primary context sets, a search for a consistent interpretation of the scene is conducted. Hypotheses are individually added to cliques, checking for consistency with those already present in the clique. The search begins with those candidates at the tops of the partial orders and progresses until no more hypotheses can be added without causing a contradiction. The clique that explains the largest portion of the image is offered as the best interpretation. The result should represent a reasonable explanation of major portions of the image.

4.5 Model Consistency

Determining the "correctness" of a scene model is analogous to establishing the validity of a scientific tbeory – one can never be assured that the model is valid: the best we can do is to filter out incorrect models by showing they have some internal contradiction or violate some accepted fact. In the approach we have described for constructing a scene model, the problem of assuring global consistency is addressed by the mechanism for clique formation – the addition of a new assertion about the scene should not be able to cause the satisfaction of the conditions of other context sets that cause rejection of already accepted hypotheses.

Assertions about the different semantic categories constrain each other in an absolute way in terms of establishing 3-D geometric and physical relationships that must be consistent with known (a priori) facts about the environment and the object categories (constraints on size; support; orientation; occupancy of solid objects; and free-space imposed by the viewing geometry of the image). For example, if an object was identified as a tree at some unknown distance, and a later assertion established the distance to the tree so as to allow us to check its dimensions against known generic tree size limits, we might find that a contradiction exists between the distance assertion and the tree identity assertion. Thus, if an object is identified as a tree, then a region of the image completely occluding a portion of the trunk certainly cannot be sky. Even if the system has enough knowledge to deduce this fact through activation of appropriate context sets (i.e., (1) an occluding object is closer than the object it occludes, (2) the sky is infinitely far away, (3) if the sky occludes the tree trunk, the tree trunk is infinitely far away, (4) a visible object at infinite distance has infinite size, (5) trees have finite size), explicit knowledge of this contradiction might remain unknown to the system. Since there is no explicit mechanism for accomplishing the above reasoning, the contradiction would only be recognized if we explicitly force the propagation of all distance and size assertions (i.e., make explicit all physical assertions). Rather than attempting to achieve this purpose by random or exhaustive expansion of our knowledge base, we build a (simulated) analogical model of the environment as a way of directing the deductive process in evaluating global consistency. Thus, consistency checking is accomplished by 3-D model construction according to a set of conditions that prevent a non-physically-realizable situation from occuring. The resulting 3-D model is one of the primary outputs of the system.

5 Progress

The adequacy of our approach is largely an empirical question which we address experimentally using real imagery. The implementation (called CVS, for Contextual Vision System) is not complete, but enough has been put in place to provide some preliminary results. The system is built on a number of other components that we routinely use in our work. One of these, the Core Knowledge System (CKS), provides the primary representation and storage mechanism for the actual and hypothesized scene entities derived by the CVS [19, 20].

Our implementation strategy has been first to construct a control structure to carry out all phases of the approach. This has been completed. Second, an instantiation of the knowledge base has been accomplished for a thin slice of the knowledge that must be present in a fully functional system. We have used results obtained from this partial system to guide the design of the remainder of the knowledge base and to provide insight into the merits and limitations of our approach. Constructing the knowledge bases for a perceptual system can be a tedious and difficult task. It is clear that some form of automated knowledge acquisition (learning) is desirable and perhaps necessary. We are exploring ways to add such a facility to CVS.

5.1 Vocabulary

As mentioned earlier, the choice of vocabulary is crucial, and the best terms to include in the vocabulary may not be the most obvious ones. The terms we currently employ for the recognition of trees were listed in

Table 1. They were chosen on the basis of three factors: (1) they can often be recognized in relative isolation – knowledge of the presence of other classes is probably not necessary for finding potential candidates; (2) they appear to be useful for setting the context for the recognition of other objects; and (3) operators can be constructed to reliably locate candidate instances. As the system matures, we will introduce new terms into the vocabulary.

5.2 Control of Complexity

In the customary machine vision paradigm, an image is partitioned into disjoint regions, and a unique label is assigned to each one. If there are r regions and k labels, then r^k possible labelings exist. Various constraints and heuristics are employed to find the best labeling within this search space, which is exponential in the number of potential labels.

Within CVS, each operator generates a small number of candidate regions, which are then ordered by the applicable context sets and added to cliques. Suppose a correct interpretation of an image contains r regions. Let p be the probability that a region that is about to be added to a clique is part of a consistent interpretation. The probability of adding r such regions to a single clique is p^r. On average, one would have to try constructing p^{-r} cliques before a valid one with r regions is obtained. Since each clique requires up to r regions, the complexity of clique construction is $O(rp^{-r})$. Thus, any contextual recognition scheme, such as ours, is potentially exponential in the number of regions in the result. A coarse description is generated rather quickly – more detail can be added but at an exponentially increasing cost. The key to attaining a manageable level of complexity is the ability to offer a candidate region to a clique with a high probability of acceptance. If $p = 1$, then the complexity is $O(r)$. Obviously, the closer that p approaches 1, the longer one can ward off a combinatoric explosion. The quality of the candidate generators certainly affects p – having fewer incorrect candidates yields a higher probability of acceptance. The process of partial order construction was designed to boost p even higher. By ensuring that the best examples of a label are introduced first, one may avoid the need to introduce some of the less likely candidates.

5.3 Hypothesis Generators

A hypothesis generator is implemented as a combination of low-level operations that delineates in an image one or more regions as candidates for a particular vocabulary term. Our strategy has largely been to employ standard image-processing routines that have been developed by the vision community through the years. These are combined in various ways to tailor their output according to the specifications of the context sets. Some of the operators we currently employ were listed in Table 3.

5.4 Hypothesis Ordering

Partial orders among candidate regions are created by pairwise comparison of candidates. Context sets associated with each vocabulary term are employed to perform the pairwise comparison. A context set defines a collection of related criteria that is sufficient to prefer one candidate over another as an instance of its class. Examples of such criteria are listed in Table 2. When some criterion is not satisfied, a context set will offer no preference between two candidates. A preference relation will be added to the partial order if any context set offers a preference. Examples of several partial orderings of candidate regions is shown in Figures 5 – 7.

There is one special "context set" for each semantic category that enforces constraints on the physical properties of the object (e.g. tree trunk width, maximum tree height, maximum branch lengths). If any physical property of an object fails to satisfy a listed constraint, the object cannot be given the corresponding label and is removed from the partial order.

5.5 Global Consistency

A labeled region is not accepted until a mutually consistent set of labeled regions is identified that explains the image features. Cliques of consistent regions are formed incrementally by nominating a candidate for inclusion, evaluating it for consistency with the existing clique, and adding it to the clique. If a nominee is found to be incompatible with a clique, it is removed from its partial order and an alternate nominee is selected. If a nominee is consistent, it is added to the clique, and all context sets are (theoretically) reevaluated in light of the new context represented by the expanded clique.[2]

A 3-D model of the evolving clique is created by inserting object tokens into the Core Knowledge System [19, 20]. The spatial and semantic retrieval mechanisms of the CKS are employed to establish context for further processing by other context sets. The CKS maintains the data associated with each clique as separate opinions so that hypotheses maintained by one clique do not interfere with those of another clique.

Nominees are chosen according to various heuristics that attempt to maximize the chance that a candidate will be accepted into a clique. Currently, we use the simple heuristic of choosing the largest region that is atop any partial order. Later, we intend to allow the context to suggest which candidate from which class should be nominated.

An initial consistency check is performed in the image plane. A nominee is checked for overlap with any region already in the clique. If the overlap exceeds a threshold, the nominee is ruled to be inconsistent. This simple-minded scheme is intended to serve as a place-holder until a module capable of full three-dimensional consistency checking is completed.

5.6 Illustration of Clique Formation

We now present an example showing the preprocessing and result of clique formation. Some aspects of the example have been hand-edited for the sake of perspicuity. Figure 8 portrays an image of some trees on the Stanford campus that has been presented to CVS for interpretation. After applying a number of context sets and their associated hypothesis generators, partial orders for the vocabulary terms sky, foliage, ground, and tree trunk have been created (see Figures 5 – 7).

Suppose that sky candidate number 593 (which was generated by a simple thresholding on intensity) is nominated first. It is added to an empty clique and the corresponding volume is marked as sky in the CKS. Any context sets that might now be satisfied are reevaluated. This reevaluation may cause the generation of new candidates or may change some of the partial orders, but for the remainder of this example, we'll assume that does not happen. Next, foliage candidate 543 is introduced. It overlaps somewhat with the sky region in the clique, but not enough to flag an inconsistency. So 543 is added as foliage; its volume (estimated using range from stereo) is inserted in CKS as foliage, context sets are reevaluated, and processing continues. Next ground region 549 is nominated. Its overlap with the sky region already in the clique is greater than the allowed threshold for ground-sky overlap and is ruled inconsistent. Candidate 549 is removed from the ground partial order and processing continues. Suppose foliage candidate 545 is nominated next. Its overlap with the sky region already in the clique is small enough, so its volume is computed and inserted in the CKS. But, because it is completely contained in the sky volume it has no possibility for support, and is ruled as inconsistent. It and all its inferiors in the foliage partial order are removed from consideration because there are context sets that have already determined that 545 is a better example of foliage than 544 or 594. If 545 is not foliage, than 544 and 594 cannot be foliage either. Further nomination of candidates 554 (sky), 540 and 536 (ground), 546 (foliage) and 537 (tree trunk) yields a clique containing (536 537 540 543 546 593) – its coverage of the image is portrayed in Figure 9(a).

[2] In practice, a "lazy" evaluation scheme is employed to perform the reevaluation efficiently.

Now suppose a new clique is created, but that sky candidate 556 is the first nominated. Foliage candidate 543 is nominated and accepted. When ground candidate 549 is nominated, no overlap with the sky region is found, so 549 is accepted into the clique as ground. Foliage candidate 545 is now found to be supportable by the ground, so it is accepted as well. Further processing results in a clique containing (536 537 540 543 544 545 546 549 556 595). The coverage of this clique is shown in Figure 9(b). Because it explains more of the image than the previous clique, it is accepted as the better interpretation.

6 Concluding Discussion

6.1 Prior Work

Nearly all of the existing work on recognition (by a robotic vision system) has been conducted in a context where precise geometric models of the relevant objects are known beforehand, and the major goal has been to find projections of the various models that best match some part of an image [2, 4, 12, 13]. To relax the requirement for complete and accurate models, Fischler and Elschlager introduced the technique of deformable (spring-loaded) templates [6], which represent objects as a combination of local appearances and desired relations among them (the "springs"). An object represented in this way is located in an image by using a form of dynamic programming to simultaneously minimize local and global evaluation functions. Some geometric recognition systems, such as ACRONYM [3], accept parameterized models to recognize a class of objects, but these too are overly restrictive to be of much use with natural features.

For the natural world, precise geometric models of natural objects are not available, and existing techniques offer little insight on how to proceed. There has been some work directed toward the goal of semantic understanding of natural outdoor scenes [1, 5, 7, 11, 14, 16, 17, 18, 21, 22], but surprisingly, very little new work has been initiated in the last ten years.[3] All of these approaches begin by partitioning the image into regions, which presumably mirrors the "natural" decomposition of the scene into "objects." The regions are then analyzed in one way or another to determine their interrelationships, to merge them into larger regions, and ultimately, to assign each region a label that categorizes it semantically.

6.2 Our Contribution

Some of the key differences between our work and previous efforts include recognition in the absence of explicit shape models; no reliance on accurately partitioned and delineated objects; no requirement for logically consistent absolute constraints; and no use of probabilistic models requiring a priori probability values and independence assumptions.

The critical issue that must be resolved in formulating a viable control strategy is whether it is necessary to recognize everything at once (e.g., via relaxation), or whether some critical scene elements can/must be recognized first in relative isolation. If so, what are these elements and how can they be found? In a similar sense, what volume of space, context, etc. constitutes a smallest interpretable unit? We take a relatively unique position based on the following assertion: almost nothing in nature can be visually identified in isolation. Therefore, every interpretation rule must have an explicitly stated contextual setting to which its use is restricted.

The ability to generate "good" hypotheses and perform reliable comparisons of candidates for a particular label is one of the most important aspects of our system. We have devised a new mechanism, known as *context sets*, to support these functions. The name derives from the need to compile a set of information

[3]The major exception to this statement is the work sponsored by the DARPA Strategic Computing program. However, much of this work, such as that described in this paper, has still not reached a full stage of maturity.

sufficient for deciding whether one candidate should be preferred over another. This information, then, constitutes the context for recognition in that circumstance. The context sets are the mechanism employed to account for such factors as view angle, scale, and geographic location.

The context sets are the main repository of knowledge within our system. In addition to supporting the ordering of hypotheses, and thus the efficient construction of a global model, they play a major role in hypothesis generation and in checking global consistency – the context sets are thus an integrative mechanism linking all the different knowledge levels the system must be concerned with.

A key idea in the way the system is organized is that we do not make direct decisions that choose between two different class labels for a detected region in an image – the context sets and the partial orderings they create deal only with one semantic category. Thus we avoid the combinatorics of having to (explicitly) describe how to distinguish among combinations of different class labels for an unknown object. We also avoid the need to make major modifications in our knowledge base when some new object type is added. Recognition in the CVS is accomplished when a globally-consistent, labeled 3-D model has been constructed.

7 Acknowledgment

We would like to acknowledge the contributions of Helen Wolf and Lynn Quam, who shared their expertise and assisted with the system implementation and with the experiments described here.

References

[1] Barrow, Harry G., and Tenenbaum, Jay M., "MSYS: A System for Reasoning about Scenes," Technical Note 121, Artificial Intelligence Center, SRI International, April 1976.

[2] Bolles, R.C., Horaud, R., and Hannah, M.J., "3DPO: A 3-D Part Orientation System," *Proceedings 8th International Joint Conference on Artificial Intelligence*, Karlsruhe, West Germany, August 1983, pp. 1116–1120.

[3] Brooks, Rodney A., "Model-Based 3-D Interpretations of 2-D Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 5, Number 2, March 1983, pp. 140–150.

[4] Faugeras, O.D., and Hebert, M., "A 3-D Recognition and Positioning Algorithm using Geometrical Matching Between Primitive Surfaces," *Proceedings 8th International Joint Conference on Artificial Intelligence*, Karlsruhe, West Germany, August 1983, pp. 996–1002.

[5] Feldman, Jerome A., and Yakimovsky, Yoram, "Decision Theory and AI: A Semantics-Based Region Analyzer," *Artificial Intelligence*, Vol. 5, No. 4, 1974, pp. 349–371.

[6] Fischler, M. A., and Elschlager, R. A., "The Representation and Matching of Pictorial Structures," *IEEE Transactions on Computers*, Volume C-22, Number 1, January 1973, pp. 67–92.

[7] Fischler, Martin A., Bolles, Robert C., and Smith, Grahame, "Modeling and Using Physical Constraints in Scene Analysis," Technical Note 267, Artificial Intelligence Center, SRI International, September 1982.

[8] Fischler, Martin A., and Wolf, Helen C., "Linear Delineation," Proceedings IEEE Computer Vision and Pattern Recognition, 1983, pp. 351–356.

[9] Fischler, Martin A., and Firschein, O., *Intelligence: The Eye, the Brain, and the Computer*, Addison-Wesley Publishing Co., 1987, pp 220–229.

[10] Fischler, Martin A., and Strat, Thomas M., "Recognizing Trees, Bushes, Rocks and Rivers," *Proceedings of the AAAI Spring Symposium Series: Physical and Biological Approaches to Computational Vision*, Stanford University, March 1988, pp. 62–64.

[11] Garvey, Thomas D., "Perceptual Strategies for Purposive Vision," PhD Thesis, Department of Electrical Engineering, Stanford University, December 1975.

[12] Goad, C., "Special Purpose Automatic Programming for 3D Model-Based Vision," *Proceedings: DARPA Image Understanding Workshop*, June 1983, pp. 94–104.

[13] Grimson, W.E.L., and Lozano-Perez, T., "Model-Based Recognition from Sparse Range or Tactile Data," *International Journal of Robotics Research*, Volume 3, Number 3, 1984, pp. 3–35.

[14] Hanson, A.R., and Riseman, E.M., "VISIONS: A Computer System for Interpreting Scenes," in *Computer Vision Systems*, Academic Press, New York, 1978, pp. 303–333.

[15] Laws, Kenneth, I., "Integrated Split/Merge Image Segmentation," Technical Note 441, Artificial Intelligence Center, SRI International, July 1988.

[16] McKeown, David M., Jr., and Denlinger, Jerry L., "Cooperative Methods for Road Tracking in Aerial Imagery," *Proceedings: DARPA Image Understanding Workshop*, April 1988, pp. 327–341.

[17] Rosenfeld, A., Hummel, R.A., and Zucker, S.W., "Scene Labeling by Relaxation Operations," *IEEE Trans Systems, Man, Cybernetics*, Volume 6, Number 6, June 1976, pp. 420–433.

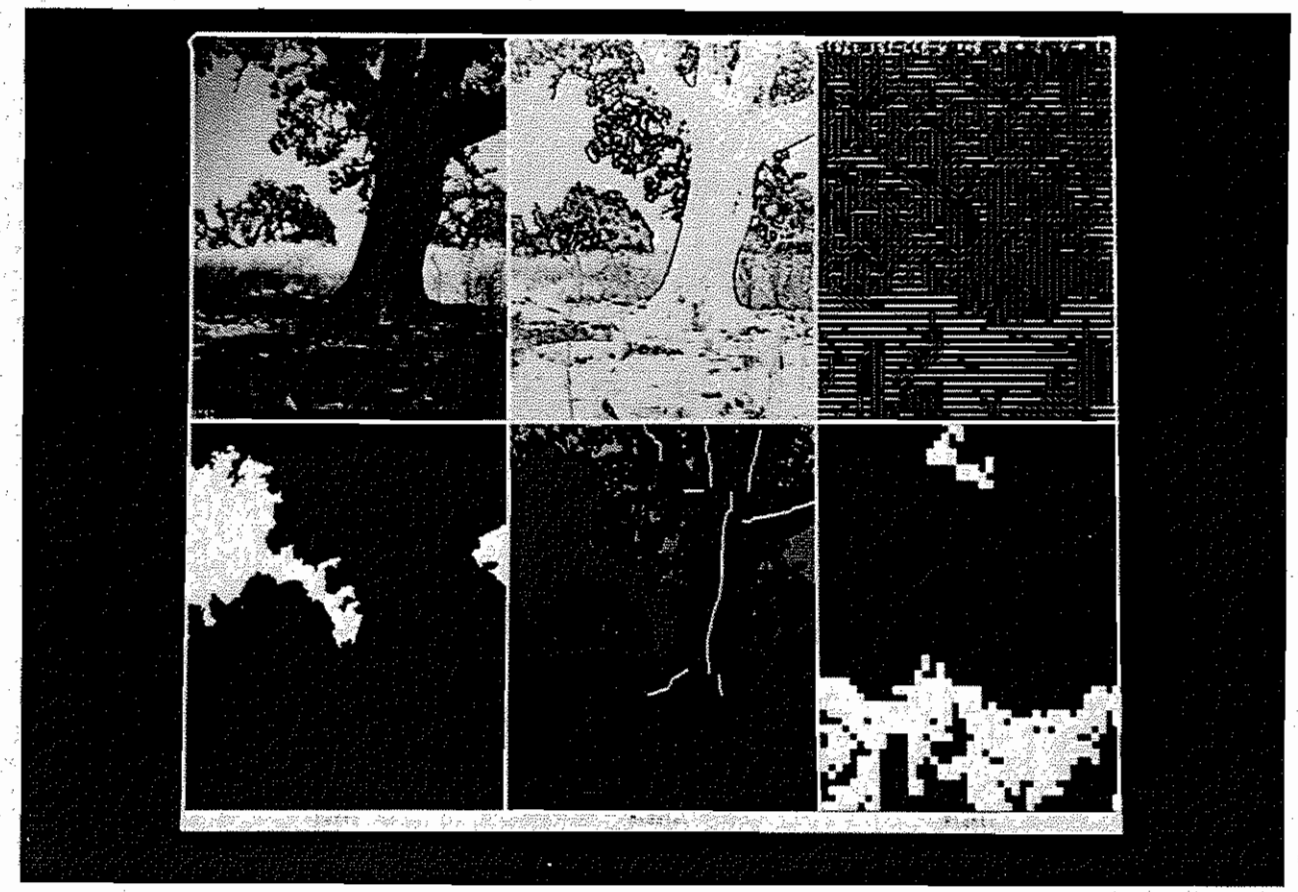
[18] Sloan, Kenneth R., Jr., "World Model Driven Recognition of Natural Scenes," PhD Thesis, Moore School of Electrical Engineering, University of Pennsylvania, Philadelphia, Pennsylvania, June 1977.

[19] Smith, Grahame B., and Strat, Thomas M., "Information Management in a Sensor-Based Autonomous System," *Proceedings: DARPA Image Understanding Workshop*, February 1987, pp. 170–177.

[20] Strat, Thomas M., and Smith, Grahame B., "The Core Knowledge System," Technical Note 426, Artificial Intelligence Center, SRI International, October 1987.

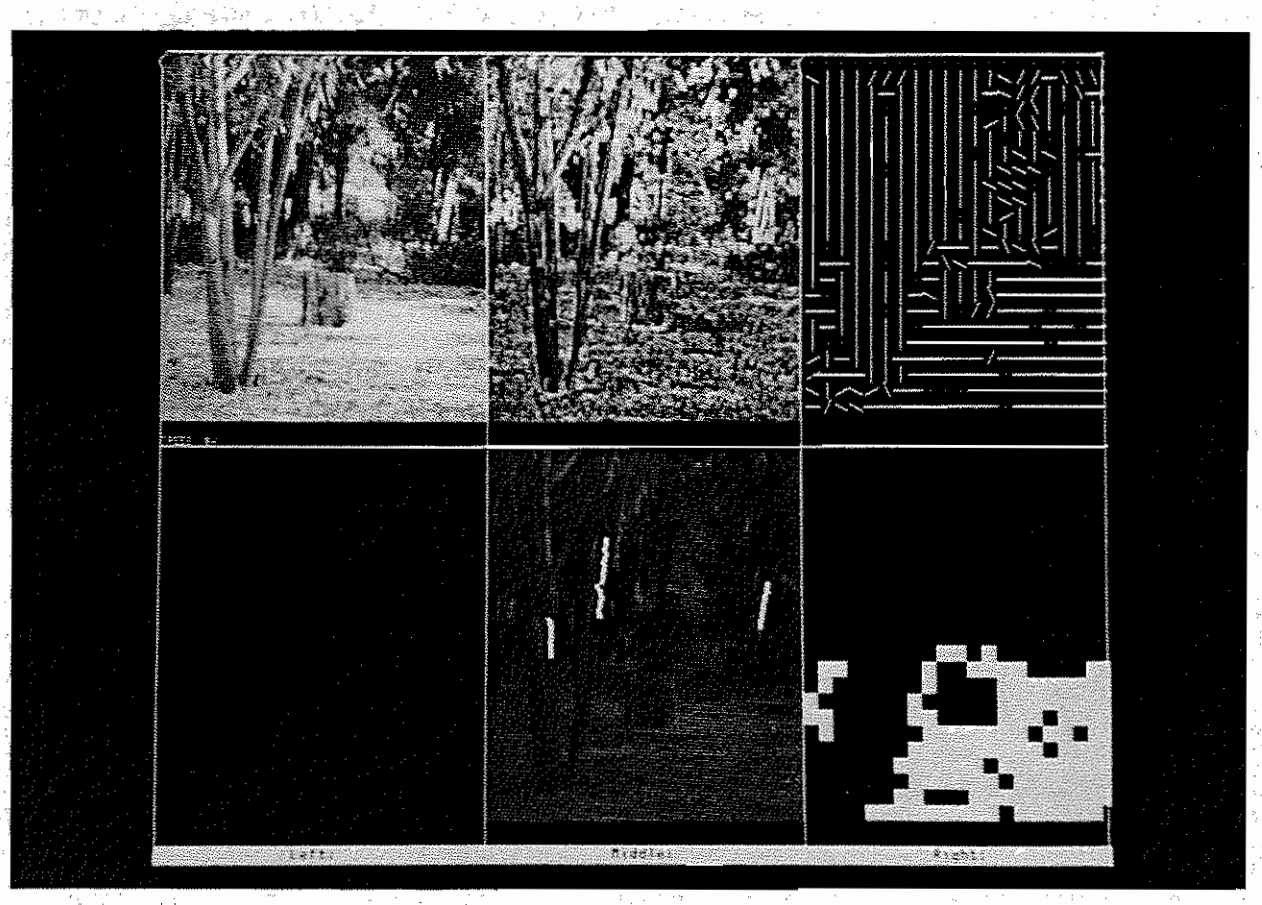
[21] Tenenbaum, J. M. and Weyl, S., "A Region Analysis Subsystem for Interactive Scene Analysis," *Proceedings of the Fourth International Joint Conference on Artificial Intelligence*, September 1975, pp. 682–687.

[22] Yakimovksy, Yoram, and Feldman, Jerome A., "A Semantics-Based Decision Theory Region Analyzer," *Proceeding of the Third Joint Conference on Artificial Intelligence*, August 1973, pp. 580–588.



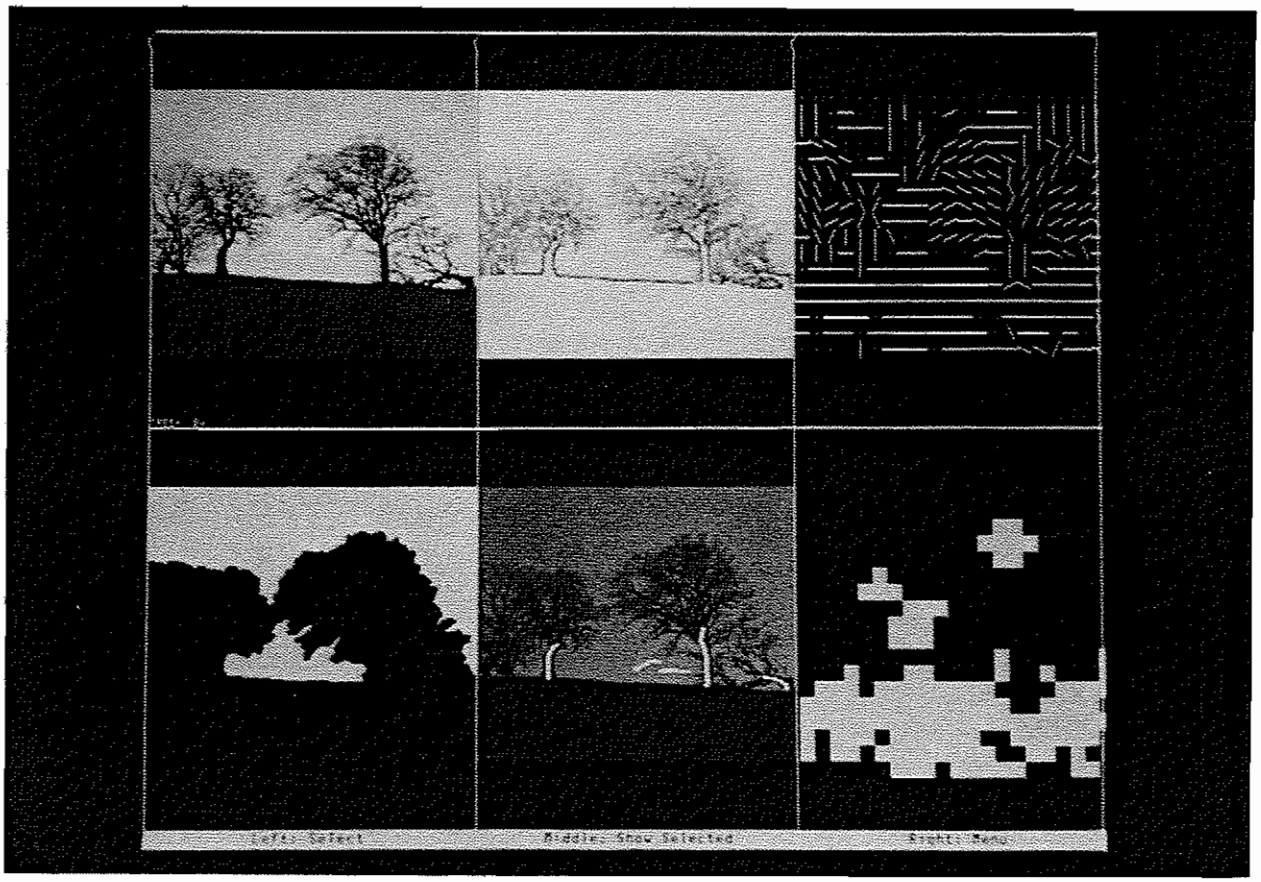
(a) B&W Image of some trees near Stanford	(b) Homogeneity operator – Each pixel value is the maximum difference in intensity between it and all neighboring pixels.	(c) Striations operator – Line segments show the orientation of any texture pattern in a small window.
(d) Sky region hypotheses – The entire scene was partitioned by Laws' segmenter, KNIFE [15]. Each region that is above the geometric horizon, is relatively bright, and is relatively untextured is displayed.	(e) Tree trunk hypotheses – Coherent regions were grown from the output of the homogeneity operator (b) above. Skeletons of the major regions were constructed and filtered to remove short and highly-convoluted skeletons. The tree trunk and its major limbs have been identified (superimposed in white on the original image).	(f) Ground region hypotheses – Regions of horizontal striations were extracted from (c) above. Small regions have been discarded. Horizontal surfaces tend to have horizontal striations when viewed from an oblique angle due to perspective foreshortening.

Figure 1: Output of various operators applied to a natural scene



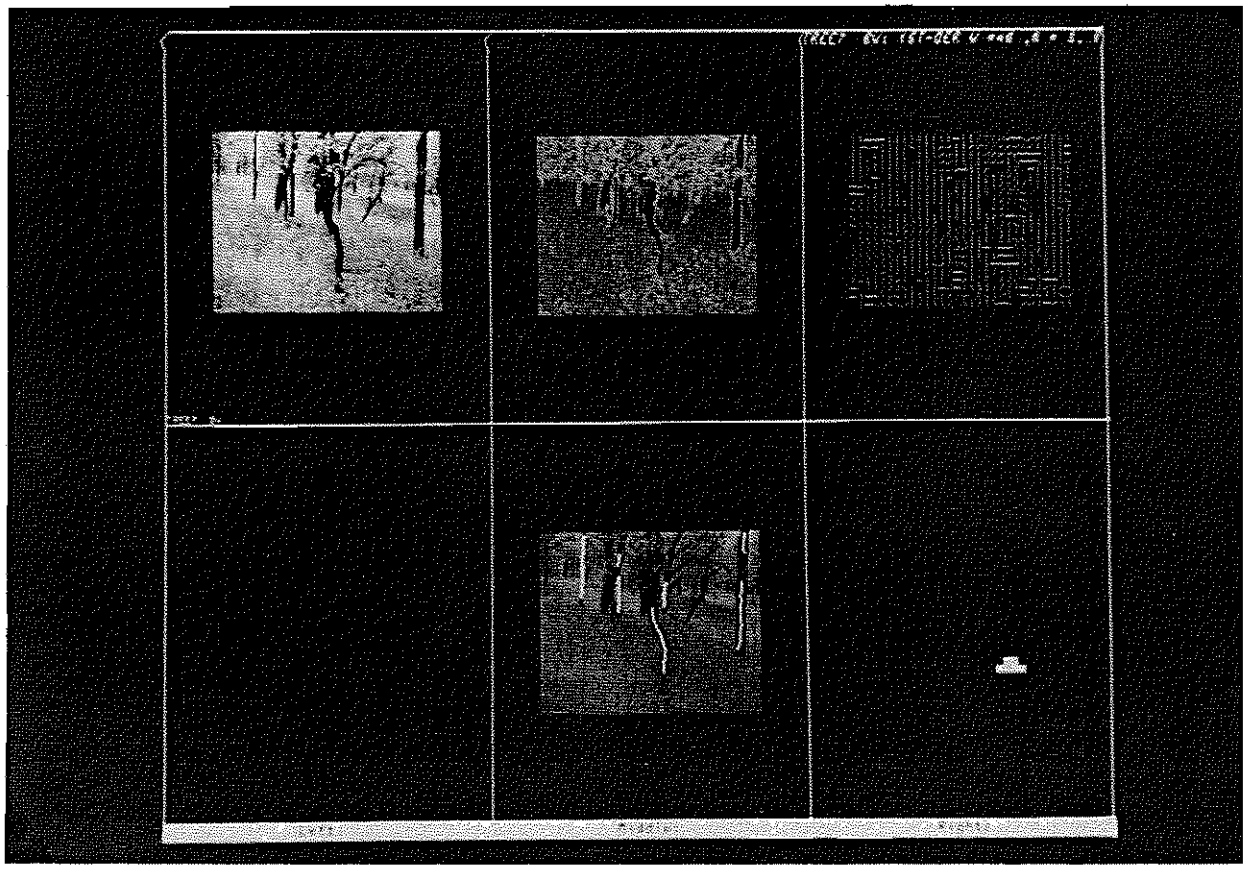
(a) B&W Image of some trees and a stump	(b) Homogeneity operator – Each pixel value is the maximum difference in intensity between it and all neighboring pixels.	(c) Striations operator – Line segments show the orientation of any texture pattern in a small window.
(d) Sky region hypotheses – The entire scene was partitioned by the KNIFE segmenter. There were no relatively bright, untextured regions above the geometric horizon.	(e) Thin object hypotheses – A linear delineation operation [8] was performed on the output of the homogeneity operator (b) above. Short segments and highly-convoluted segments were removed. Portions of several tree trunks have been identified.	(f) Ground region hypotheses – Regions of horizontal striations were extracted from (c) above. Small regions have been discarded. Notice how the stump is not included.

Figure 2: Output of various operators applied to a natural scene



(a) B&W Image of some trees near Stanford	(b) Homogeneity operator – Each pixel value is the maximum difference in intensity between it and all neighboring pixels	(c) Striations operator – Line segments show the orientation of any texture pattern in a small window
(d) Sky region hypotheses – The entire scene was partitioned by the KNIFE segmenter. Each region that is above the geometric horizon, is relatively bright, and is relatively untextured is displayed.	(e) Thin object hypotheses – A linear delineation operation was performed on the output of the homogeneity operator (b) above. Short segments and highly-convoluted segments were removed. The trunks of the trees were successfully delineated, although some spurious candidates remain to be filtered out later.	(f) Ground region hypotheses – Regions of horizontal striations were extracted from (c) above. Small regions have been discarded. Some branches of the trees were picked up, but these candidates should be eliminated during later processing.

Figure 3: Output of various operators applied to a natural scene



(a) B&W Image of some trees	(b) Homogeneity operator – Each pixel value is the maximum difference in intensity between it and all neighboring pixels	(c) Striations operator – Line segments show the orientation of any texture pattern in a small window
(d) Sky region hypotheses – The entire scene was partitioned by the KNIFE segmenter. Each region that is above the geometric horizon, is relatively bright, and is relatively untextured is displayed. No regions survived.	(e) Thin object hypotheses – A linear delineation operation was performed on the output of the homogeneity operator (h) above. Short segments and highly-convoluted segments were removed. Many of the tree trunks have been identified.	(f) Ground region hypotheses – Regions of horizontal striations were extracted from (c) above. Small regions have been discarded. Very little ground was identified because the terrain is covered with tall grass.

Figure 4: Output of various operators applied to a natural scene

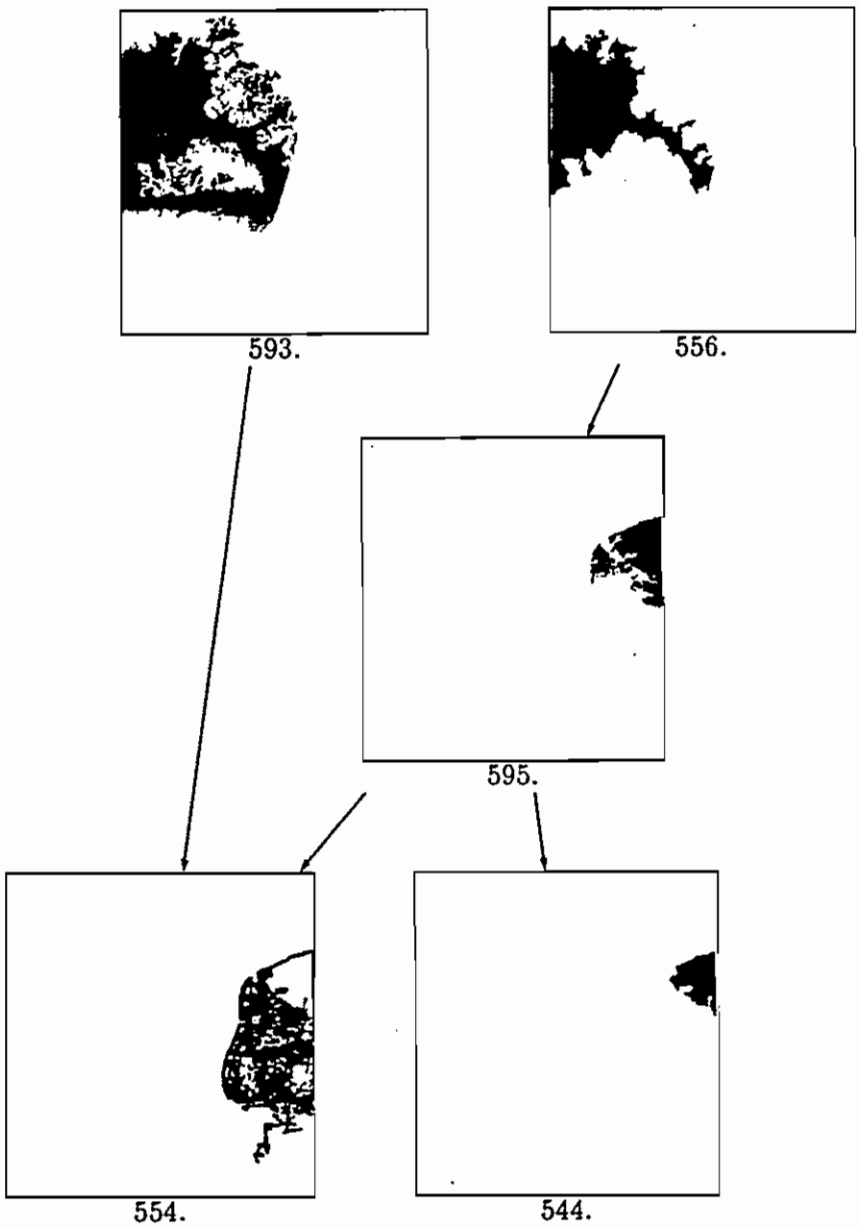


Figure 5: Partial order of sky candidates for the tree image of Figure 8

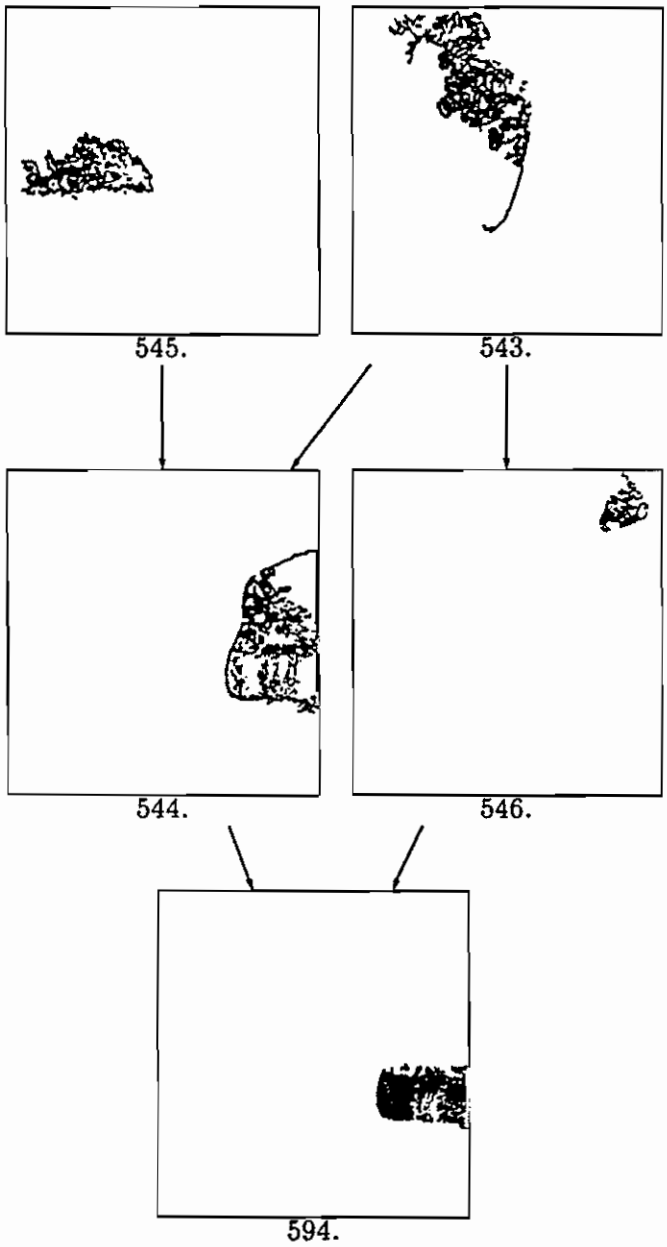
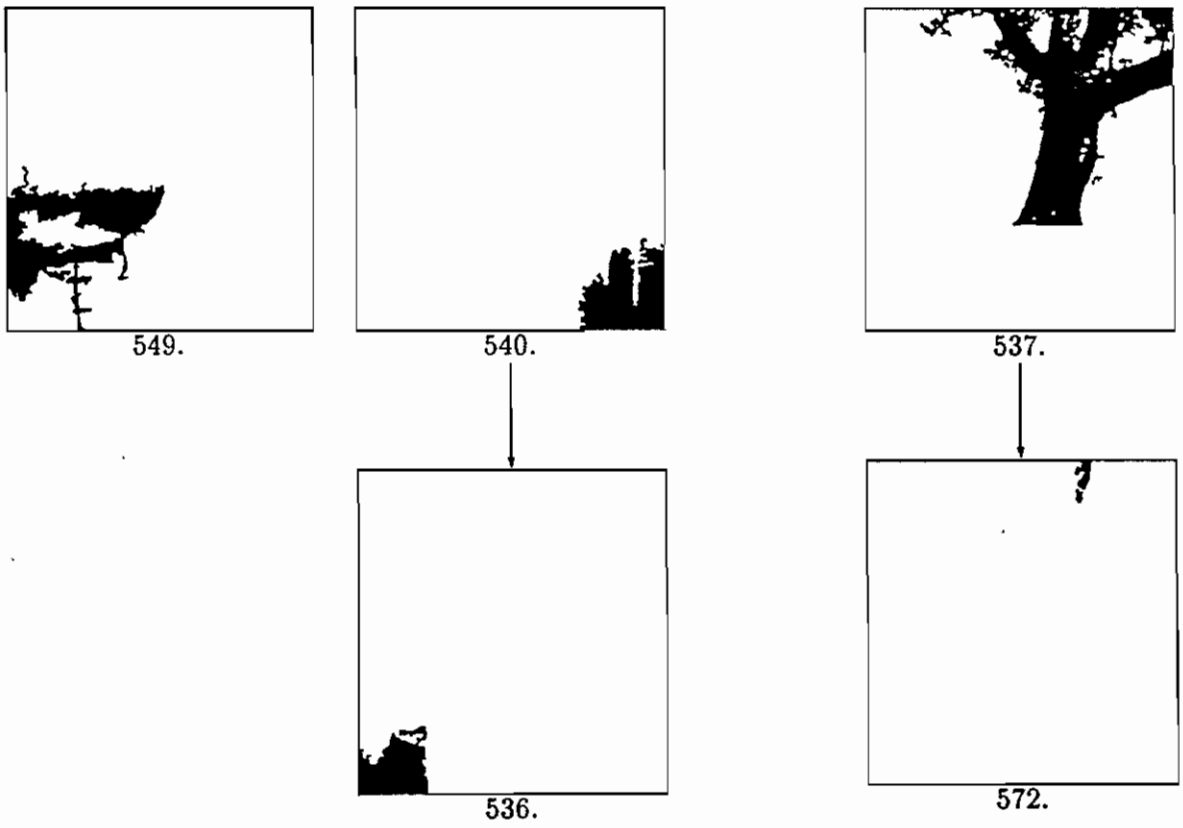


Figure 6: Partial order of foliage candidates for the tree image of Figure 8



(a) Ground candidates

(b) Tree trunk candidates

Figure 7: Partial orders for the tree image of Figure 8

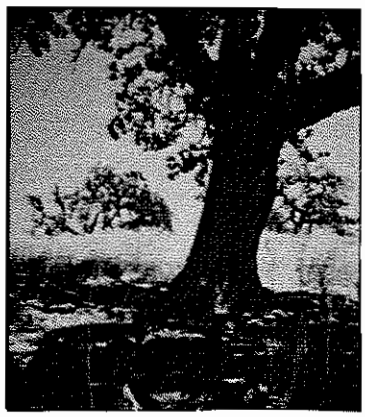


Figure 8: A natural scene of a tree on the Stanford campus

(a)

(b)

Figure 9: Region coverage maps for two cliques formed from the Stanford tree scene