Natural Object Recognition: A Theoretical Framework and Its Implementation

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

Abstract

Most work in visual recognition by computer has focused on recognizing objects by their geometric shape, or by the presence or absence of some prespecified collection of locally measurable attributes (e.g., spectral reflectance, texture, or distinguished markings). On the other hand, most entities in the natural world defy compact description of their shapes, and have no characteristic features with discriminatory power. As a result, image-understanding research has achieved little success toward recognition in natural scenes. We offer a fundamentally new approach to visual recognition that avoids these limitations and has been used to recognize trees, bushes, grass, and trails in ground-level scenes of a natural environment.

1 Introduction

The key scientific question addressed by our research has been the design of a computer vision system that can approach human-level performance in the interpretation of natural scenes such as that shown in Figure 1. We offer a new paradigm for the design of computer vision systems that holds promise for achieving near-human competence, and report the experimental results of a system implementing that theory which demonstrates its recognition abilities in a natural domain of limited geographic extent. The purpose of this paper is to review the key ideas underlying our approach (discussed in detail in previous publications [12, 13]) and to focus on the results of an ongoing experimental evaluation of these ideas as embodied in an implemented system called Condor.

When examining the reasons why the traditional approaches to computer vision fail in the interpretation of ground-level scenes of the natural world, four fundamental problems become apparent:

Universal partitioning — Most scene-understanding systems begin with the segmentation of an image into homogeneous regions using a single partitioning algorithm applied to the entire image. If that partitioning is wrong, then the interpretation must also be wrong, no matter how a system assigns semantic labels to those regions. Unfortunately, universal partitioning algorithms are notoriously poor delineators of natural objects in ground-level scenes.

Shape — Many man-made artifacts can be recognized by matching a 3D geometric model with features extracted from an image [1, 2, 4, 6, 7, 9, 15], but most natural objects cannot be so recognized. Natural objects are assigned names on the basis of their setting, appearance, and context, rather than their possession of any particular shape.

Computational complexity — The object recognition problem is NP-hard [16]. As a result, computation time must increase exponentially as additional classes are added to the recognition vocabulary, unless a strategy to avoid the combinatoric behavior is

Figure 1: A natural outdoor scene of the experimentation site.

*Supported by the Defense Advanced Research Projects Agency under Contracts DACA76-85-C-0004, DACA76-90-C-0021, and 89F737300.

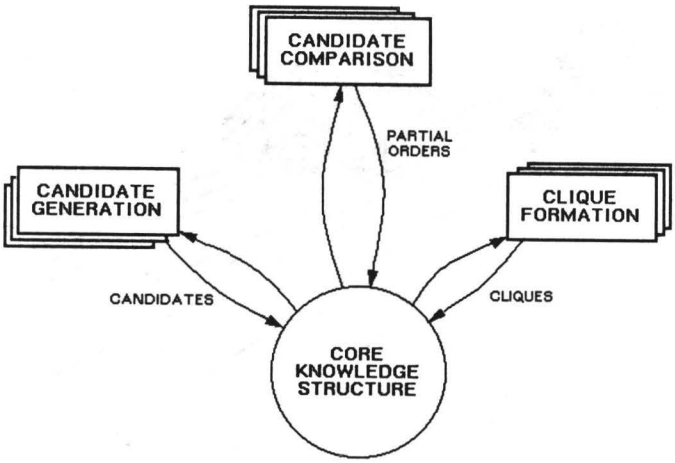

Figure 2: Conceptual architecture of Condor.

incorporated. Such provisions are a necessary component of any recognition system that can be scaled to embrace a real domain.

Contextual knowledge — Despite general agreement that recognition is an intelligent process requiring the application of stored knowledge [3, 5, 14], computer vision researchers typically use artificial intelligence techniques only at the highest levels of reasoning. The design of an approach that allows stored knowledge to control the lower levels of image processing has proved elusive.

Except for the continuing work at the University of Massachusetts [3], the understanding of natural scenes has received surprisingly little attention in the last decade.

2 Approach

A new paradigm for computer vision systems has been developed, which addresses all four of the problems described above. The key provision of this novel approach is a mechanism for the application of stored knowledge at all levels of visual processing. A *context set*, which explicitly specifies the conditions and assumptions necessary for successful invocation, is associated with every procedure employed by the recognition system.

The architecture is organized into three modules as depicted in Figure 2 and described below (a more complete description is also available [13]):

Candidate Generation — Hypotheses concerning the presence in a scene of specific categories of objects are generated by delineating regions in an image using special-purpose operators whose invocation is controlled by context sets, thereby avoiding the need for universal partitioning algorithms. The employment of large numbers of operators ensures that quality hypotheses can be generated in nearly every context and provides redundancy that decreases the reliance on the success of any individual operator.

Candidate Comparison — Hypotheses are accepted only if they are consistent with all other members of a *clique* (consistent subset). Candidate hypotheses for each label are ranked so that the best candidates for each label can be considered before the others. Ranking the candidates ensures that the largest cliques can be found early in the search, thereby limiting the computational complexity of the entire paradigm to a linear growth as the recognition vocabulary is expanded. By constructing only a small number of cliques for each image, the approach loses any guarantee of finding the largest clique, but assures the availability of a credible answer compatible with the computational resources of the system.

Clique Formation — Consistency is enforced by procedures (controlled by context sets) that detect and reject physically impossible combinations of hypotheses. The clique that most completely explains the available data is offered as the interpretation of an image. Thus, individual objects are labeled on the basis of their role in the context of the complete clique, rather than solely on the basis of individual merit.

The invocation of all processing elements throughout the system is governed by context. All processing actions are controlled by context sets, and are invoked only when their context sets are satisfied. Thus, the actual sequence of computations (and the labeling decisions that are made) are influenced by contextual information, which is represented by prior knowledge about the environment and by the computational state of the system.

Definition: A *context set*, CS_k, is a collection of context elements that are sufficient for inferring some relation or carrying out some operation on an image.

Syntactically, a context set is embedded in a *context rule* denoted by

$$L : \{C_1, C_2, \cdots, C_n\} \Longrightarrow A$$

where L is the name of the class associated with the context set, A is an action to be performed, and the C_i comprise a set of conditions that define a context.

Example: The context rule

SKY : {SKY-IS-CLEAR, CAMERA-IS-HORIZONTAL RGB-IS-AVAILABLE } $\Longrightarrow$ BLUE-REGIONS

defines a set of conditions under which it is appropriate to use the operator BLUE-REGIONS to delineate candidate sky hypotheses.

There is a collection of context rules for every class in the recognition vocabulary, and each collection contains rules of three types: candidate generation, candidate comparison, and consistency determination. In theory, Condor performs the actions A that are associated with every satisfied context set.

Figure 3: Result of analyzing Figure 1.

3 The recognition process

For each label in the active recognition vocabulary, all candidate-generation context sets are evaluated. The operators associated with those that are satisfied are executed, producing candidates for each class. The candidate-comparison context sets that are satisfied are then used to evaluate each candidate for a class, and if all such evaluators prefer one candidate over another, a preference ordering is established between them. These preference relations are assembled to form partial orders over the candidates, one partial order for each class. Next, a search for mutually coherent sets of candidates is conducted by incrementally building cliques of consistent candidates, beginning with empty cliques. A candidate is nominated for inclusion into a clique by choosing one of the candidates at the top of one of the partial orders. Consistency-determination context sets that are satisfied are used to test the consistency of a nominee with candidates already in the clique. A consistent nominee is added to the clique; an inconsistent one is removed from further consideration with that clique. Further candidates are added to the clique until none remain. Additional cliques are generated in a similar fashion as computational resources permit. Ultimately, one clique is selected as the best semantic labeling of the image on the basis of the portion of the image that is explained and the reliability of the operators that contributed to the clique.

The interaction among context sets is significant. The addition of a candidate to a clique may provide context that could trigger a previously unsatisfied context set to generate new candidates or establish new preference orderings. For example, once one bush has been recognized, it is a good idea to look specifically for similar bushes in the image. This tactic is implemented by a candidate-generation context set that includes a context element that is satisfied only when a bush is in a clique.

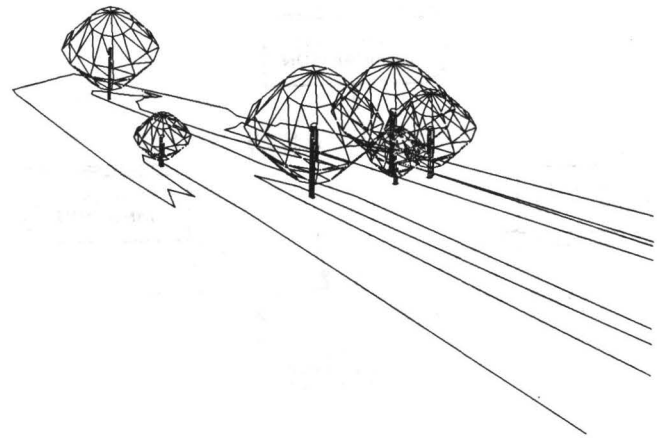

Figure 4: A perspective view of the 3D model produced from the analysis of the image shown in Figure 1.

4 Evaluation scenario

The approach has been implemented in the form of a complete end-to-end vision system, known as Condor. Images that are monochromatic or color, monocular or stereo, provide the input to the system, along with a terrain database containing prior knowledge about the environment. Condor produces a 3D model of the environment labeled with terms from its recognition vocabulary which is stored in the Core Knowledge Structure (CKS) [10, 11] and can be superimposed on the input image (Figure 3) or viewed from another perspective (Figure 4). The model is used to update the terrain database for use by Condor during the analysis of subsequent imagery.

To evaluate the Condor approach, we selected a two-square-mile region of foothills immediately south of the Stanford University campus as our site for experimentation. This area contains a mixture of oak forest and widely scattered oak trees distributed across an expanse of gently rolling, grass-covered hills and is criss-crossed by a network of trails.

We chose 14 classes for the recognition vocabulary on the basis of their prevalence in the experimentation site and their importance for navigation. These terms are:

{sky, ground, geometric-horizon, foliage, bush, tree-trunk, tree-crown, trail, skyline, raised-object, complete-sky, complete-ground, grass, tree}

Procedures have been devised to extract, evaluate, and check the consistency of candidates for each of these classes. Context sets have been constructed to control the invocation of each of those procedures. Currently the knowledge base contains 88 procedures whose invocation is governed by 156 context sets. All the results described in this paper have been generated using this knowledge base.

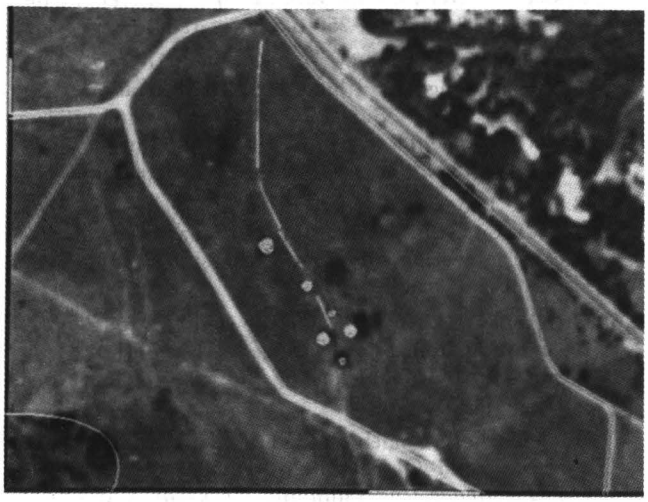

Figure 8: The composite model resulting from the analysis of the image sequence in Figure 7.

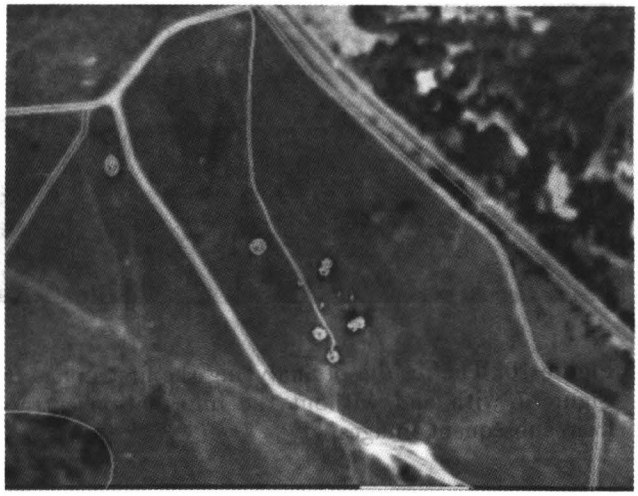

Figure 9: The ground-truth database.

them. No trunk was detectable in the foliage to the left of the image, so Condor labeled it as bush.

Image 6 — The texture in the lower corners of the sixth image was found to more closely resemble foliage than grass, so these regions were erroneously identified as bushes. Beause they are very near the camera, they occupy a significant part of the image, but the 3D model created for them reveals that they are less than 2 feet tall.

Image 7 — Several more trees, grass areas, and part of the trail are recognized in the seventh image.

Image 8 — The primary tree is recognized despite the strong shadows, but the lower portion of the trunk was missed by all the trunk operators. Most of the tree crown operators were unable to provide a decent candidate because of the overhanging branches in the upper-right corner — the only operator that succeeded was the one that predicts the crown based on the size and location of the trunk. The combined effects of the incomplete trunk, the nearness of the tree, and the lack of range data account for poor extraction of the tree crown.

This experiment illustrates how Condor is able to use the results of analyzing one image to assist the analysis of other images. Although some trees and parts of the trail were missed in several images, the 3D model that results is nearly complete. Figure 8 shows an aerial view of the composite model contained in the CKS after processing all eight images. For comparison, Figure 9 portrays a model of the objects actually present on the ground, which was constructed by physically measuring the locations and sizes of the individual objects. Note that all of the trees that were visible in at least one image have been correctly labeled, although some of them were misplaced. Most of the trail has been detected, enough to allow a spatial reasoning process to link the portions into a single continuous trail. Furthermore, everything that was labeled *tree* actually is a tree.

5.3 Experiment 3

Regardless of the architecture, knowledge-based vision systems are difficult to build. If the programmer needed to specify in advance all the information necessary for successful recognition, his task would be hopeless. Therefore, it is essential that a vision system have the ability to improve its competence autonomously, thereby learning through experience how to recognize the objects in its environment.

Assertion 3 *Using context allows Condor to learn how to recognize natural objects.*

To test the validity of this assertion, we return to the first image of the sequence used in Experiment 2 (Figure 7). When originally analyzed, Condor recognized the trail and part of the grass, but not the trees.

Condor was tasked to reanalyze the first image, this time making use of the contents of the entire database constructed during the analysis of the sequence of eight images. The resulting interpretation is depicted in Figure 10.

Two trees that could not be extracted on the first pass are now identified. Condor employed a tree-trunk operator whose context set requires knowledge of the approximate location of a tree in the field of view. The operator projects a deformable 3D model of the trunk onto the image, and optimizes its fit to extract the trunk. This operator successfully identified two of the trees without contradicting any of the original recognition results.

This experiment (along with others not described here) illustrates that the ability to use prior recognition results as context while interpreting subsequent images enables Condor to improve its performance as its exposure to its environment increases.

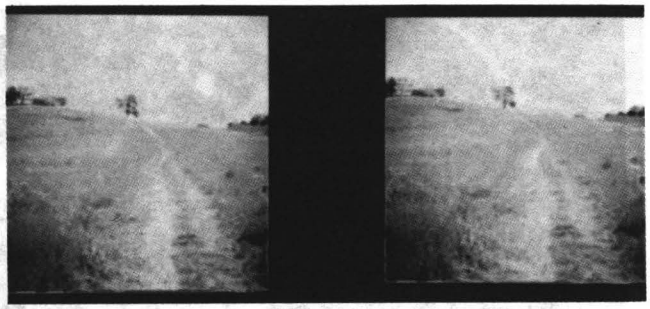

Figure 10: The results of analyzing the first image from Figure 7 with and without the information extracted from subsequent images.

6 Conclusion

In its present embodiment, Condor is still a demonstration system that should be evaluated primarily in terms of its architectural design and innovative mechanisms, rather than its absolute performance. While Condor has demonstrated a recognition ability approaching human-level performance on some natural scenes, it is still performing at a level considerably short of its ultimate potential (even for the Stanford experimentation site). The knowledge acquisition mechanisms, which are a key aspect of the architecture, should allow continued improvement in performance with exposure to additional site imagery.

A new paradigm for image understanding has been proposed, and used to recognize natural features in ground-level scenes of a geographically limited environment. This context-based approach is exciting because it deemphasizes the role of image partitioning and emphasizes the recognition context in a way that has not been attempted before. This new focus could lead to the construction of vision systems that are significantly more capable than those available today.

7 Acknowledgment

Condor includes software provided by many present and former members of the Perception Group at SRI. In addition, Marty Tenenbaum, Jean-Claude Latombe, and Lynn Quam have contributed to the research reported here.

References

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

[2] Brooks, Rodney A., "Model-Based 3-D Interpretations of 2-D Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 5, No. 2, pp. 140–150 (March 1983).

[3] Draper, Bruce A., Robert T. Collins, John Brolio, Allen R. Hanson, and Edward M. Riseman, "The Schema System," *International Journal of Computer Vision*, Vol. 2, No. 3, pp. 209–250 (January 1989).

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

[5] Garvey, Thomas D., "Perceptual Strategies for Purposive Vision," Ph.D. Dissertation, Department of Electrical Engineering, Stanford University, Stanford, CA (December 1975).

[6] Grimson, W.E.L., and T. Lozano-Perez, "Model-Based Recognition from Sparse Range or Tactile Data," *International Journal of Robotics Research*, Vol. 3, No. 3, pp. 3–35 (1984).

[7] Huttenlocher, Daniel P., and Shimon Ulman, "Recognizing Solid Objects by Alignment," *Proceedings: DARPA Image Understanding Workshop*, Cambridge, MA, pp. 1114–1122 (April 1988).

[8] Kass, Michael, Andrew Witkin, and Demitri Terzopoulos, "Snakes: Active Contour Models," *Proceedings, ICCV*, London, England, pp. 259–268 (June 1987).

[9] Ponce, Jean, and David J Kriegman, "On Recognizing and Positioning Curved 3D Objects from Image Contours," *Proceedings: DARPA Image Understanding Workshop*, Palo Alto, CA, pp. 461–470 (May 1989).

[10] Smith, Grahame B., and Thomas M. Strat, "A Knowledge-Based Architecture for Organizing Sensory Data," *International Autonomous Systems Congress Proceedings*, Amsterdam, Netherlands (December 1986).

[11] Strat, Thomas M., and Grahame B. Smith, "The Core Knowledge System," Technical Note 426, Artificial Intelligence Center, SRI International, Menlo Park, CA, (October 1987).

[12] Strat, Thomas M., and Martin A. Fischler, "A Context-Based Recognition System for Natural Scenes and Complex Domains," *Proceedings, DARPA Image Understanding Workshop*, Pittsburgh, PA, pp. 456–472 (September 1990).

[13] Strat, Thomas M., "Natural Object Recognition," Ph.D. Dissertation, Department of Computer Science, Stanford University, Stanford, CA (December 1990).

[14] Tenenbaum, Jay M., "On Locating Objects by Their Distinguishing Features in Multisensory Images," *Computer Graphics and Image Processing*, pp. 308–320 (December 1973).

[15] Thompson, D.W., and J.L. Mundy, "Three-Dimensional Model Matching from an Unconstrained Viewpoint," in *Proc. IEEE Int. Conf. on Robotics and Automation*, pp. 208–220, 1987.

[16] Tsotsos, John K., "A Complexity Level Analysis of Immediate Vision," *International Journal of Computer Vision*, Vol. 1, No. 4, pp. 303–320 (1988).

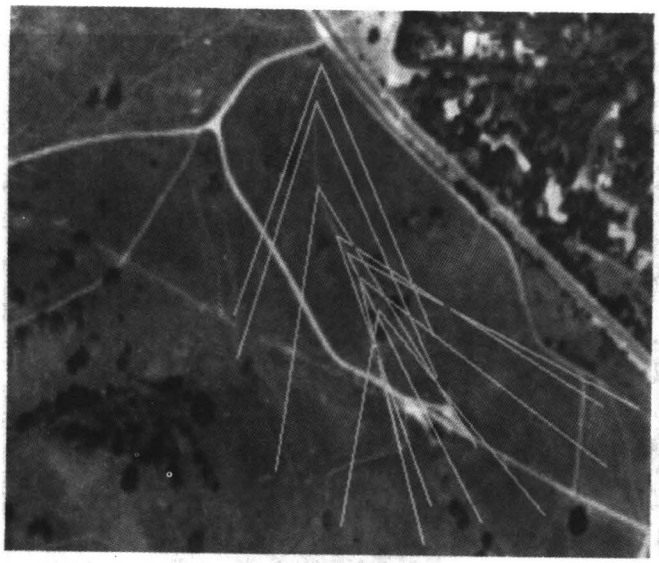

Figure 6: The location and orientation of the camera when each image in Figure 7 was acquired.

analysis are portrayed in Figure 7. Here we highlight a few of the more interesting chains of reasoning and explain the misidentifications that were made:

Image 1 — Condor has correctly labeled the sky, the ground, the trail, and part of the grass, although the trees on the horizon were too indistinct to be reliably identified. These results are transformed into three-dimensional models and positioned in 3-space using depth data acquired from binocular stereo.[1] The resulting models were added to the CKS database to be used as context for the analysis of subsequent images.

Image 2 — The model of the trail from the first image was projected into the second image and used to help identify a portion of the trail. This is accomplished by an operator that superimposes a pair of parallel 3D curves and deforms them to find the model with maximum edge strength while minimizing its curvature (as in [8]). Statistics from the intensity and texture of the grass in the first image were used to help identify the grass in this second image. In this case, the trail-finding operators failed to find the upper half of the trail; as a result, the grass hypotheses in that area were not contradicted.

Image 3 — The tree is finally close enough to allow reliable recognition and a 3D model for it is computed by extracting the envelope of its foliage. The entire visible portion of the trail was correctly identified.

Image 4 — Two additional trees are recognized and stored.

Image 5 — The same trees are recognized by predicting their location and verifying their existence — a much more reliable process than initially extracting

[1] When range data are not available, Condor estimates the depths by projecting each region onto the DEM.

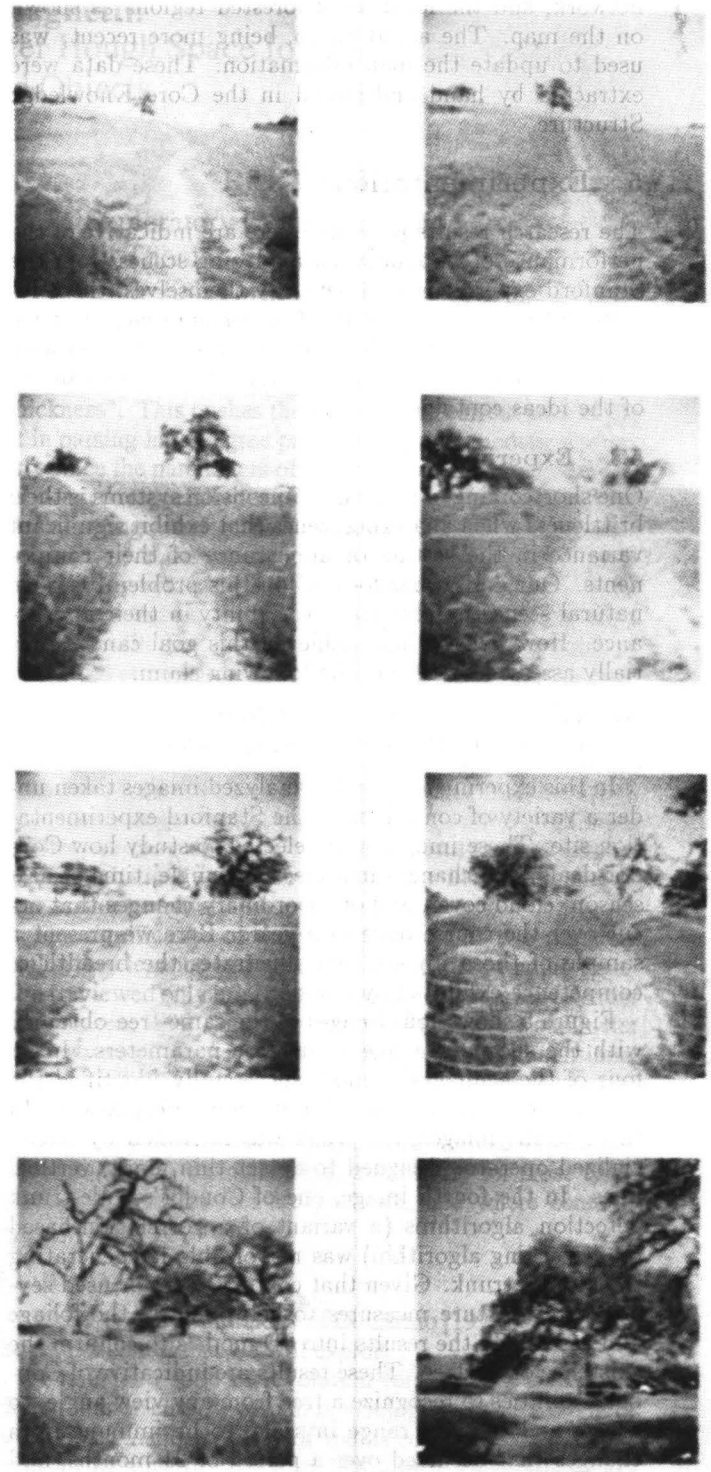

Figure 7: Results of Condor's analysis of the sequence of eight images.

Initial contextual information was extracted from a USGS map and an aerial photograph; this includes a 30-meter-grid digital elevation model (DEM), the road network, and the location of forested regions as shown on the map. The aerial photo, being more recent, was used to update the map information. These data were extracted by hand and stored in the Core Knowledge Structure.

5 Experimentation

The research results presented here are indicative of the performance of Condor when analyzing scenes from the Stanford experimentation site. By themselves, these results do little to endorse the Condor approach, but together with similar results that have been obtained with several dozens of other images, they attest to the validity of the ideas contained therein.

5.1 Experiment 1

One shortcoming of many machine vision systems is their brittleness when analyzing scenes that exhibit significant variance in the setting or appearance of their components. Our design has focused on this problem because natural scenes possess great variability in their appearance. How well we have achieved this goal can be partially assessed by testing the following claim:

Assertion 1 *The Condor architecture is suitable for recognizing natural objects in many contexts.*

In this experiment, Condor analyzed images taken under a variety of conditions at the Stanford experimentation site. These images were selected to study how Condor deals with changes in scale, view angle, time of day, season, cloud cover, and other ordinary changes that occur over the course of several years. Here we present a sample of those images that illustrates the breadth of competence exhibited by Condor.

Figure 5 shows four images of the same tree obtained with the specified image acquisition parameters. In all four of these images, Condor successfully identified the tree without the benefit of any prior information. In three of the images, the trunk was identified by a specialized operator designed to detect thin, dark, vertical lines. In the fourth image, one of Condor's wide-trunk detection algorithms (a variant of a correlation-based road-tracking algorithm) was responsible for generating the correct trunk. Given that context, Condor used several of its texture measures to help identify the foliage and assembled the results into 3D models to confirm the existence of the tree. These results are indicative of Condor's abilities to recognize a tree from any view angle, to accommodate a 7:1 range in scale, to be immune from changes that occurred over a period of 21 months, and to deal with seasonal variation. When Condor has prior knowledge of the existence of this tree, it can be recognized from a distance of at least 590 feet (as demonstrated in Experiment 3), thereby extending its abilities to a 20:1 range in scale.

Experiments applying Condor to other images (not reproduced here) confirm the viability of the approach for recognizing natural objects in a wide variety of settings

range:	194 feet	28 feet
view angle:	160°	124°
date:	12 April 1990	28 July 1988
range:	56 feet	87 feet
view angle:	208°	258°
date:	12 April 1990	12 April 1990

Figure 5: The models of the trees as they were recognized by Condor.

that occur at the experimentation site. The modularity of the context sets makes it possible to expand the breadth of competence still further without degrading previously demonstrated capabilities.

5.2 Experiment 2

To support autonomy in an intelligent, ground-based vehicle, it is necessary to synthesize a reasonably complete description of the entire surroundings, and not just recognize a few isolated objects. This description can be built incrementally because the world does not change very rapidly considering the spatial and temporal scales at which an autonomous ground vehicle would operate. The following assertion summarizes this notion:

Assertion 2 *A geographic database of an extended region can be constructed by combining the recognition results from multiple images, taken over an extended period of time and under multiple viewing conditions.*

To validate this assertion, a sequence of imagery was collected which simulates the movement of a vehicle through a portion of the Stanford experimentation site. The vision system is to construct a labeled, 3D map of the primary features in the vicinity of the simulated vehicle by analyzing each image in turn.

Figure 6 shows the location of the vehicle when each image in the sequence was acquired. Condor was tasked to locate the trees, bushes, trails, and grass in each of these images, beginning with only the information extracted from the USGS map. The results of Condor's