Environmental Modeling and Recognition for an Autonomous Land Vehicle

D.T. Lawton, T.S. Levitt, C.C. McConnell, and P.C. Nelson
Advanced Decision Systems
Mountain View, CA 94040

1. ABSTRACT

~~We present~~ an architecture for object modeling and recognition for an autonomous land vehicle is presented. Examples of objects of interest include terrain features. fields. roads. horizon features. trees. etc. The architecture is organized around a set of data bases for generic object models and perceptual structures. temporary memory for the instantiation of object and relational hypotheses. and a long term memory for storing stable hypotheses that are affixed to the terrain representation. Multiple inference processes operate over these databases. ~~We~~ describe these particular components: the perceptual structure database. the grouping processes that operate over this. schemas. and the long term terrain database. ~~We conclude with~~ a processing example that matches predictions from the long term terrain model to imagery. extracts significant perceptual structures for consideration as potential landmarks. and extracts a relational structure to update the long term terrain database given.

2. INTRODUCTION

Terrain and object models for autonomous land vehicles (ALVs) are required for a wide range of applications including route and tactical planning. location verification through the recognition of terrain features and objects. and acquiring new information about the environment as it is explored. The following lists important criteria for terrain and object modeling capabilities.

Descriptive Adequacy: The modeling technique should be capable of describing the objects and situations in the environment necessary for the vehicle to function. This includes representing natural as well as man-made objects. It should be a consistent representation that supports modular system development and uniform inference procedures that can operate over different types of objects at different levels of detail. Uniform shape. object subpart and surface attribute affixments are necessary to do this.

Recognition Adequacy: Much of the activity of an ALV is concerned with determining where it is and what is around it. Terrain models should be manipulable for determining the sensor-based appearances of world objects and for controlling recognition processing. This involves the formation of general predictions of sensor derived features from the terrain model. Such predictions will often be uncertain and qualitative due to incomplete prior knowledge of the terrain.

Handling Uncertainty: The existence and exact environmental location of objects will often not be known with complete certainty. Locations will often be determined relative to other known locations and not with respect to a globally consistent terrain map. This is true, for instance, when the sensor displacement parameters are not well determined. It is necessary to represent this uncertainty explicitly in the terrain model so incrementally acquired information can be used for disambiguation.

Learning: A vehicle will learn about the environment as it moves through it. Associating new information with the terrain representation should be straightforward. This is difficult to do, for example, by changing values in a raw elevation array. Types of information to be affixed to the terrain representation include newly discovered objects, details of expected objects, and the processing used in object recognition.

Fusion of Information: The ALV must build a consistent environmental model over time from different sensors. As an object is approached, its image appearance and scale will change considerably, yet it has to be recognized as the same object, with newly acquired information associated with the unique instance of the general object type. In a typical situation, a distant dark terrain patch will be partially recognized based upon distinctive visual characteristics, but may be either a building or a road segment. As it is approached, its image appearance changes considerably, making disambiguation possible. This requires the representation of multiple hypotheses, each formated with respect to the properties of the potential world objects. The structure of the object description should direct the accumulation of information.

A further consideration in developing and evaluating terrain modeling capabilities is that there is not a single ALV. Instead, there are a wide range of autonomous vehicles, indexed by a diverse range of active and passive sensors and assumptions about a priori data. There is a continuum from systems having a complete initial model of the terrain and perfect sensors to those with no a priori model, and highly imperfect sensors. For example, a robot with no a priori data and only an unstabilized optical sensor will probably model the environment in terms of a sequence of views related by landmarks and distinct visual events embedded in a representation that is more topological than metric. An ALV solely dependent on optical imagery will have to deal with the huge variability in the appearance of objects. Experience has shown that even road surfaces have highly variable visual characteristics. Alternatively, a few pieces of highly preselected visual information can serve to verify predictions from a reliable and detailed terrain model and precise position and range sensors.

We call a general object model a schema. A schema can represent perceived, but unrecognized, visual events, as well as recognized objects and their relationships in environmental scenes. The architectural design is focused about the representation, instantiation, and inference over schemas developed by the ALV as it moves through the environment. Schemas are related to similar concepts found in Hanson et.al. - 78 and Ohta - 80. The short term terrain representation consists of schema instantiations that represent accumulated perceptual evidence for objects as attributes and relations that are hypothesized with varying levels of certainty.

Object models are used to organize perceptual processing by integrating descriptive representations with recognition and segmentation control. One aspect of this is the use of different types of attributes and inheritance relations between generic schemas for representation in IS-A and PART-OF hierarchies. A particular object attribute relates three dimensional world properties of an object and sensor dependent view information, either by a set of generic views or viewing procedures. These viewing attributes are also inherited and modified according to different object types. In many systems, objects are treated as lists of attributes that are matched against extracted image features. Here they are treated as specifying an active control process that directs image segmentation by specifying grouping procedures to extract and organize image structures.

Another critical aspect of the architecture is the various types of spatial, localization relations that deal with uncertainty and learning by associating different types of perceptually derived information with terrain models. For example, local (multi-sensor) viewframes affix sets of schemas and un-recognized perceptual structures into local "robot's-eye" views of an ALV's environment. Path-affixments between local viewframes support fusion of information in time without necessarily corresponding to locations in an a priori grid.

This effort has developed an architecture for terrain and object recognition compatible with the wide range of potential sensor configurations and the different qualities of a priori data.

There has been work in artificial intelligence, computer vision, and graphics that satisfy the individual requirements for object modeling capabilities, but little has been done to integrate them. To date, there is no vision system that can interpret general natural scenes, although some can deal with restricted environments Hanson et.al. - 78 while other systems are restricted to artificial objects and environments. Brooks' Brooks - 84 representation based on generalized cylinders meets, or could be extended to deal with, many of these functions. It has well defined shape attribute inheritance between a set of progressively more complex object models, and affixment relations that could be generalized to handle uncertainty. It can also be used to generate constraints on image features from object models. Nonetheless, the system built around this representation has had limited success beyond dealing with essentially orthographic views of geometrically well defined man-made objects. This appears to be partially because the constraints on image structures generated from the abstract instances of object models are too general to generate initial correspondences between models and image structures. Brook's system also used an impoverished set of image descriptions, and the object models could not direct the segmentation process directly during their instantiation. The majority of work in terrain modeling deals with how well a representation can realistically model three dimensional terrain, but not how it is used for recognition. The simplicity of a model that is described by a few parameters is not useful for recognition unless it can direct constrained searches against image data. For example, Pentland's Pentland - 83 use of fractals satisfies aspects of descriptive adequacy for natural terrain, but has been less effective for recognition. Kuipers Kuipers - 82 has produced an interesti[illegible] terrain model for learning and handling uncertainty, but it is non visual. R[illegible]ated to this is Kuan's Kuan - 84 object based terrain representation for planning that is organized in terms of distinct, modifiable objects, but is also not associated with sensor derived processing results.

3. ARCHITECTURE OVERVIEW

The system architecture consists of several databases and inference processes. The inference processes transform the databases, creating additional data structures, and modifying the existing ones. The task interface focuses attention in system processing and monitors progress toward system task goals. This high level architecture is depicted in Figure 1. The boxes with square corners in this figure represent databases, the ellipses represent inference processes, and arrows indicate dataflow.

3.1 SYSTEM DATABASES

At the highest level there are three databases. These are the short term memory (STM), long term memory (LTM), and generic models.

The STM acts as a dynamic scratchpad for the vision system. It has two sub-areas, a perceptual structures database (PSDB) and a hypothesis space. The PSDB includes incoming imagery from sensors, immediate results of extracting image structures such as curves, regions and surfaces, spatial temporal groupings of these structures, and results of inferring 3D information.

The hypotheses space contains statements about objects and terrain in the world. A hypothesis is represented as an instantiated schema. The schema points to the various perceptual structures in the PSDB that provide evidence that the object represented by the schema (such as a terrain patch, road, tree, etc.) exists in the world. Other types of hypotheses include grids, viewframes, and viewpaths. Grids are a special type of terrain representation that contain elevation information and are derived from range data or successive depth maps from motion stereo. Viewpaths, as partially ordered sequences of viewframes, give space time relationships between hypotheses. Viewframes are sets of hypotheses that correspond to what can be seen from a localized position. A hypothesis with no associated perceptual structures is a prediction. As structures and localization are incrementally added to a hypothesis, it progresses on the continuum from predicted to recognized. Hypotheses that have enough evidence associated with them to be considered recognized and stable, are moved to the LTM.

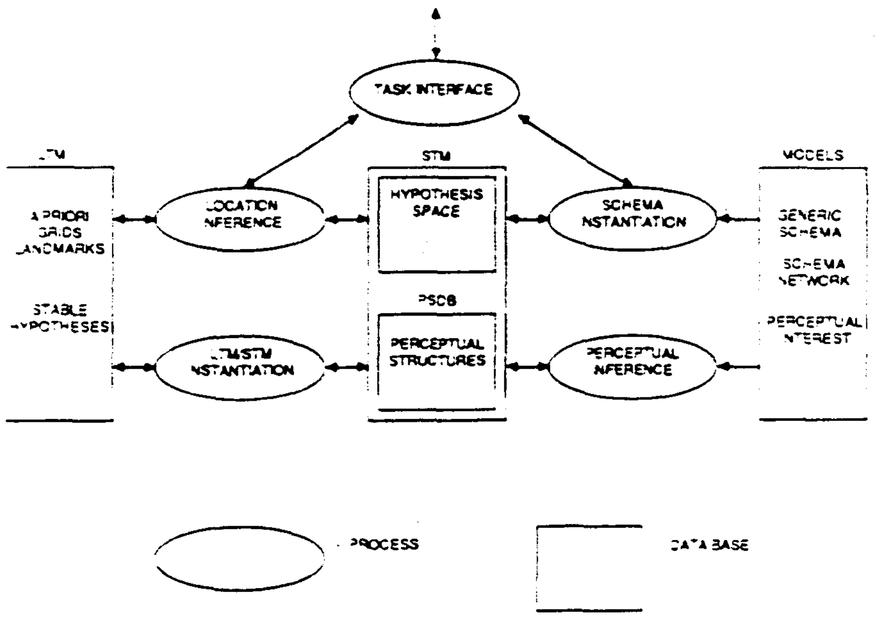


Figure 1: Terrain Modeling and Recognition System Architecture

The LTM stores a priori terrain representations, the long term terrain database, and hypotheses with enough associated evidence to be considered visually stable. A priori data concerning elevation and terrain type information, as well as knowledge of specific landmarks are stored in the LTM. A viewframe, representing a certain location in the world is stored in the LTM if the evidence associated with it could be re-used to recognize the local environment if it was re-encountered. Consistency of one hypothesis with another is not required for storage in the LTM.

The model space stores generic object models, the inheritance relations of the (model) schema network, and a set of image structure grouping processes and rules for evaluating image structure interestingness. Generic models are used dynamically to instantiate and guide search processes to associate evidence to an object instance. Inheritance relations are used by various schema inference procedures to propagate structures, attributes and relations between object instantiations. For instance, the generic two-lane-road schema has an "IS-A" relationship to the generic road schema. It follows, based on the inheritance models, that an instantiation of the two-lane-road schema will inherit the more general characteristics of the generic road schema that in turn inherits the more general characteristics of a terrain patch. Unlike the STM and LTM, the model space is not modified by inference processes.

3.2 INFERENCE PROCESSES

At the highest level, there are five different sorts of inference processes in the vision system. These are perceptual inference, location inference, object instantiation, LTM/STM instantiation, and the task interface.

The PSDB is initialized with the output of standard multi-resolution image processing operations for smoothing, edge extraction, flow field determination, etc. Much subtler inference is required for grouping processes that produce connected curves, textures, surfaces, and temporal matches between image structures. These grouping operations are typically model guided. There are generic models (which may be task dependent) of what constitutes "interestingness" of an image structure.

The hypothesis inference processes produce tasks for the perceptual processes. These may be satisfied by simple queries over the PSDB such as "find all long lines in this region of image", where "long", "line" and "region" are suitably interpreted. Queries can be more complex, requiring, for instance, temporal stability, such as "find all homogeneous green texture regions that are matched (i.e., remain in the field of view) over at least two seconds of imagery", where, again, qualitative descriptors are rigorously defined. Alternatively, the requested perceptual structures may be dynamically extracted. In this case, a history of the processing attempts and results are maintained. If similar requests are made later, such as if we were to view the same environment from a different perspective, these processing histories could be used to recall a processing sequence that produced successful results.

Location processes include a number of different modes of spatial location representation and inference. While exact location information is used when it is available, a key concept is the qualitative representation of relative location. This is fundamental, because the problem of acquiring terrain knowledge from moving sensors involves handling perceptual information that arises from multiple coordinate systems that are transforming in time. The basic approach to location inference is to represent the location of world objects in a qualitative manner that does not require the full knowledge of continuous transformations of sensor coordinates relative to the vehicle the sensors are mounted on, or of transformations of vehicle coordinates relative to the terrain.

The main structures involved in location inference are viewframes, viewpaths, and grids. Viewframes represent both metric location information about world objects derived from range sensors and view-based location information about the directions in which objects are found derived from passive sensor data.

Generic schemas are models of world objects that include information and procedures on how to predict and match the object models in the available sensor data. Besides representing 3D geometric constraints, 2D-3D sensor view appearance including effects of change in resolution and environmental effects such as season, weather, etc., schemas also indicate contextual relationships with other objects, type and spatial constraints, similarity and conflict relations, spatial localization, and appearance in viewframes.

Object schema instantiation may occur by model-driven prediction from a priori knowledge, or directly from another instantiation and a PART-OF relation. The other instantiation process may also occur by matching a distinctive perceptual structure to a schema appearance instance. This sort of "triggering" is more common in situations where there is little a priori information to guide prediction. Object instantiations generate queries to the PSDB grouping process in order to complete matching.

A key idea in object instantiation processing is inference over the model schema network hierarchies. Direct representation and inference over a large enough body of world objects to accomplish outdoor terrain understanding requires very large memory and proportionately lengthy inference procedures over that memory space. Hierarchical representation makes a significant reduction in storage requirements; furthermore, it lends itself naturally to matching schema to world objects at multiple levels of abstraction, thus speeding the inference process. Two basic hierarchies are the IS-A and PART-OF trees.

IS-A hierarchies represent the refinement of object classification. Figure 2 shows part of an IS-A hierarchy for terrain representation. The level of terrestrial-object tells us that we will not see evidence of any schema instance below this node as perceptual structures surrounded by sky. At the level of terrain-patch we pick up the geometric knowledge of adherence to the ground plane, while information stored at the level of a road schema constrains the boundaries of a terrain patch to be locally linear (with other constraints). Types beneath road add critical appearance constraints in color and texture, while the final refinement level in the IS-A hierarchy, the number of lanes, further constrains size parameters inherited from the road schema.

PART-OF hierarchies represent the decomposition of world objects into components, each of which is, itself, another world object. Figure 3 shows a PART-OF hierarchy decomposition for a generic 2-lane-road. PART-OF hierarchies contain relative geometric information that is useful in prediction and search.

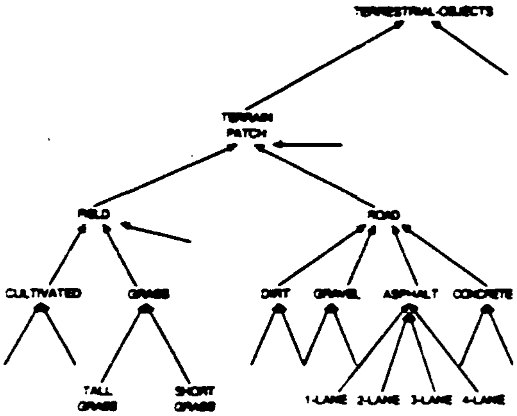


Figure 2: IS-A Hierarchy

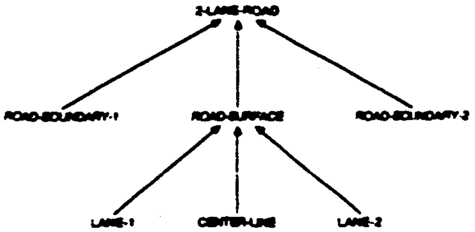


Figure 3: Part of Hierarchy

As object instantiation inference reasons up and down schema network hierarchies, incrementally matching perceptual structures and other data to instances of object appearance in the world, a history mechanism records the inference processing steps, parameters and results. This dynamic data structure is called the schema instantiation structure. One important aspect of this structure is that it can used to extract the inference and processing sequence(s) that worked earlier to see the same object, or ones that are similar. This accounts for the fact that distinctiveness in image appearance is an idiosyncratic process that depends upon many factors which are difficult to model and control, such as current motion, wind, varying outdoor illumination, etc.

4. PERCEPTUAL PROCESSING AND THE PSDB

Perceptual processing is concerned with organizing images into meaningful chunks. The definition of "meaningful" and the development of explicit criteria to evaluate segmentation techniques involves, from a data-driven perspective, that the chunks have characterizing properties, such as regularity, connectedness, and not tending to fragment the image. From a model-driven point of view, segmentation appropriateness corresponds to the extent to which the pieces can be matched to structures and predictions derived from object models. From either perspective, a basic requirement is that image segmentation procedures find significant image structures, independent of world semantics, in order to initialize and cue model matching. This allows for the extraction of world events such as surfaces, boundaries, and interesting patterns independent of understanding perceptions in the context of a particular object. These, in turn, are useful abstractions from image information to match against object models or describe the characteristics of novel objects.

The Perceptual Structure Data Base (PSDB), conceptualized in Figure 4, contains several different types of information. These are classified as images, perceptual objects, and groups. Images are the arrays of numbers obtained from the different sensors and the results of low level image processing (such as contour extraction and region growing routines) that produce such arrays. It is difficult for the symbolic relational representations used for object models, such as schemas, and the processing rules in computer vision systems, to work directly with an array of numbers. Therefore, there are many spatially-tagged, symbolic representations used in image understanding systems that describe extracted image structures such as the primal sketch [Marr - 82], the RSV structure of the VISIONS system [Hanson et.al. - 78], and the patchery data structure of Ohta [Ohta - 80]. We found it useful to build such a representation around a set of basic perceptual objects corresponding to points, curves, regions, surfaces, and volumes.

Groupings are recursively defined to be a related set of such objects. The relation may be exactly determined, as in representing which edges are directly adjacent to a region, or they may require a grouping procedure to determine the set of objects that satisfy the relationship. Groupings can occur over space, e.g., linking texture elements under some shape criteria such as compactness and density, or over time, as in associating instances of perceptual structures in successive images. We stretch the concept a bit, so that groupings also refer to general non-image registered perceptual information, such as histograms.

4.1 INITIALIZATION OF THE PSDB

Whenever new sensor data is obtained, a default set of operations are performed to initialize the PSDB. Edges are extracted at multiple spatial frequencies and decomposed into linear subsegments. The edges are extracted into distinct connected curves, and general attributes such as average intensity, contrast, and variance are associated with them. Similar processing is performed for regions extractions. Histograms are computed with respect to a wide range of object based and image based characteristics in a pyramid like structure. These default operations are used to initialize bottom-up grouping processes and schema instantiations. These, in turn, determine significant structures using heuristic interestingness rules to prioritize the structures for the application of grouping processes or object instantiations.

4.2 IMAGES

Images are the data arrays derived from the optical and laser range sensors and the results of image processing routines for operations including histogram-based segmentation, different edge operators, optic flow field computations, and so forth. Associated with images are several attributes for time of acquisition, relevant sensor parameters, etc. Processing history is maintained in the processing relationship structure that keeps track of the processing history of all objects in the PSDB.

4.3 PERCEPTUAL OBJECTS

Points, curves, regions, surfaces, and volumes are basic types of perceptual structures that are accessible to object instantiations and grouping processes. An example instance of a curve structure is shown in Figure 5. This figure shows many common representational characteristics of perceptual objects. There are default attributes associated with particular objects, such as endpoints, length and positions for a curve. There is also an associated attribute-list mechanism for incorporating more general properties with an object. This list is accessible

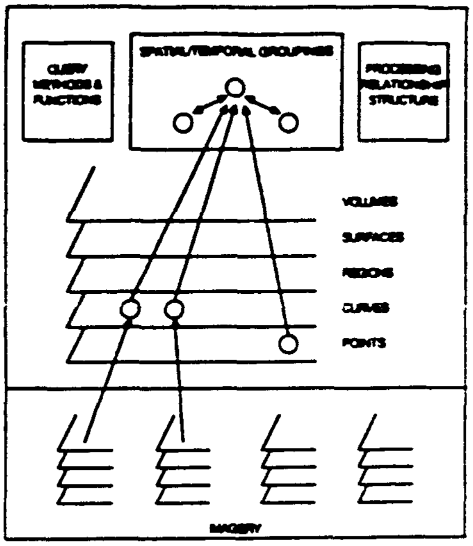
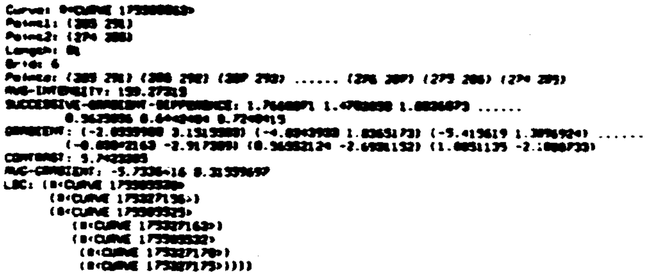


Figure 4: Perceptual Structure Data Base (PSDB)



```
Curve: 6<CURVE 175[illegible]>
Point1: (285 291)
Point2: (274 288)
Length: [illegible]
Grid: 6
Points: (285 291) (286 292) (287 293) ...... (276 289) (275 288) (274 289)
AVG-INTENSITY: 159.27[illegible]
SUCCESSIVE-GRADIENT-DIFFERENCE: 1.76[illegible] 1.47[illegible] 1.[illegible] ......
        0.3625[illegible] 0.64[illegible] 0.72[illegible]15
GRADIENT: (-2.0[illegible] 3.1315[illegible]) (-4.[illegible] 1.[illegible]3173) (-5.415619 1.[illegible]) ......
        (-0.[illegible]2163 -2.917[illegible]) (0.[illegible]2124 -2.69[illegible]152) (1.[illegible]51135 -2.[illegible]733)
CONTRAST: 5.7[illegible]
AVG-GRADIENT: -5.733E-16 0.31599697
LBC: (8<CURVE 175[illegible]>
      (8<CURVE 17532 7156>)
      (8<CURVE 175[illegible]>
        (8<CURVE 17532 7162>)
        (8<CURVE 175[illegible]>
         (8<CURVE 17532 7170>)
         (8<CURVE 17532 7175>))))
```

Figure 5: Curve Example

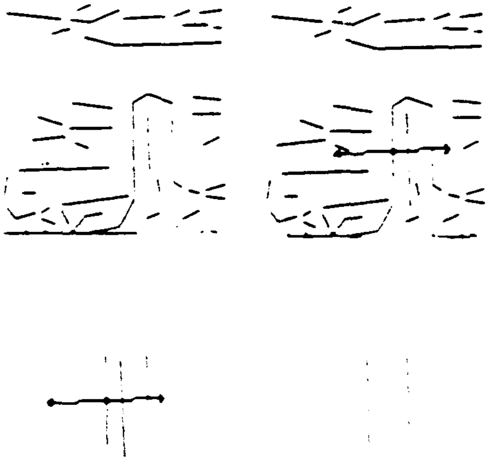


Figure 6: Parallel Grouping

by keywords and a general query mechanism using methods specific to the particular associated attribute. The associated attributes in the example are shown in capital letters. There are many types of attributes that can be consistently associated with a curve using this mechanism.

A useful representation for performing geometric operations and queries over objects is the OBJECT LABEL-GRID (or GRID: in the example curve. The number 6 indicates the index of this structure). This is an image where each pixel contains a vector of pointers back to the set of perceptual objects and groups which occupy that position. This allows geometric operations to be performed directly on the grid. Filtering operations can be applied to the OBJECT LABEL GRID to restrict processing based upon attributes associated with objects. Various types of masks can be associated with objects to reflect a directional or uniform neighborhood to determine object relationships in the OBJECT LABEL GRID.

4.4 GROUPS

A group is a set of related perceptual objects. The relation can be determined directly by a query over an object and those surrounding it, as in finding the set of curves within some distance of a given region. Alternatively, it may require a search process to find the set of objects meeting some, potentially complex, criteria. For example, an ordered set of curves can be grouped together using thresholds on allowable changes in the average contrast and orientation of successive elements. By expressing the grouping process as a search over a state space of potential groups, each group becomes a potential hypothesis in the PSDB. Groups can also reflect temporal relationships; this occurs in matching structures in successive images. A relational grouping procedure is shown in Figure 6 for the determination of nearby parallel lines with opposite contrast directions. This is done for a linear segment by first extracting nearby neighbors using a narrow mask oriented perpendicular from the segment at its mid-point. The intersection of this mask with points in the label grid are determined, and then each candidate is evaluated by checking if it is within allowable thresholds for length, contrast, and orientation. It is then ordered with respect to the smallest magnitude of the difference vector computed from the average gradients. The grouping processes can either produce the best candidate as a potential grouping, or some set of them.

Two different types of grouping processes have been developed: measure-based and interestingness-based. The measure based grouper is a generalization of established edge and region linkers [Martelli - 76]. It uses a measure consisting of:

1) some value to be optimized, such as length, minimal curvature, compactness, or a composite scalar value

2) local constraints on allowable changes in attributes

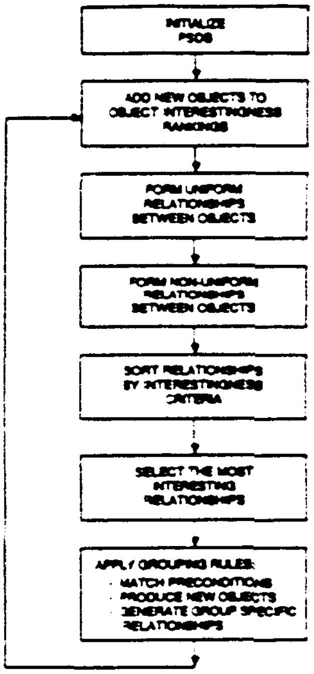
3) global thresholds on attributes

The measure and associated constraints are optimized by a best first search returning several ordered candidate groups. The measure to be used can be associated with a prediction from an object model for substance or shape characteristics. The measure to be optimized can also be determined directly from initially extracted objects by selecting those that are extreme in some attribute or are correlated with the attributes of surrounding objects to derive a measure to be optimized.

The measure based grouper is currently being generalized into one based on interestingness. It involves the basic processing loop shown in Figure 7. Initially, basic perceptual objects including curves, regions, junctions and their associated attributes are extracted using conventional techniques. Extracted objects are represented in label grids to express spatial neighborhood operations over the objects. A uniform neighborhood is established for each object, and directed relations are formed with the adjacent objects in each neighborhood. These relations are represented in a small number of types of match relationships that contain descriptions of the correlation of attributes, subcomponent matching, and composite properties.

Selected attributes of the extracted perceptual objects and the match structures are then sorted into lists with pointers back to the associated objects. These lists are for attributes such as size, average feature values, variance of feature values, compactness, the extent of correlation between the components and attributes of different structures, and the number of groups an object is involved in. These different rankings are then combined using a selection criteria to choose the set of interesting perceptual objects and relationships. The selection criteria sets the required position in different subsets of the sorted attribute lists. An example is to find 100 largest objects in the top 10 of any of the attribute correlation lists. The selection criteria is modifiable during processing and is meant to reflect the influence of model-based predictions.

Interestingness is used to focus the application of grouping rules to a selected set of objects and relations between objects indicated in match structures. The grouping rules then combine perceptual objects to form new perceptual objects, or groups, based upon the type of relation between the objects. Neighborhoods are established with respect to these derived groups to form new relationships. These in turn are sorted in the attribute lists with respect to the previously extracted perceptual objects. In addition to the relations established in uniform neighborhoods, for some groups, non-uniform relations are also established. Processing can continue indefinitely as less and less interesting relations become candidates for the application of grouping rules. Explicit criteria are needed to stop processing; e.g., we can limit processing time, determine when there is a uniform covering of the image with extracted groups, or when structures belong to unique groups.


INITIALIZE PSOS
ADD NEW OBJECTS TO OBJECT INTERESTINGNESS RANKINGS
FORM UNIFORM RELATIONSHIPS BETWEEN OBJECTS
FORM NON-UNIFORM RELATIONSHIPS BETWEEN OBJECTS
SORT RELATIONSHIPS BY INTERESTINGNESS CRITERIA
SELECT THE MOST INTERESTING RELATIONSHIPS
APPLY GROUPING RULES:
- MATCH PRECONDITIONS
- PRODUCE NEW OBJECTS
- GENERATE GROUP SPECIFIC RELATIONSHIPS

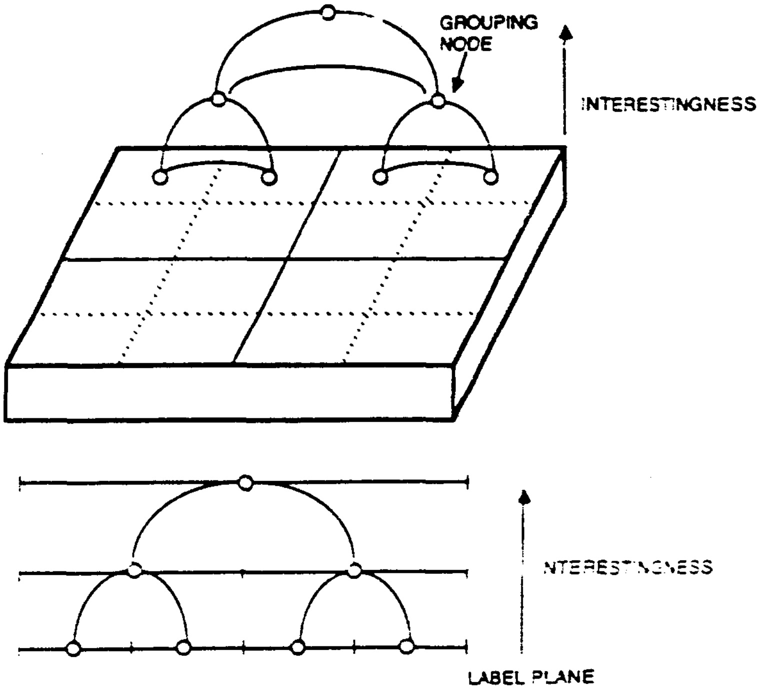


Figure 7: Grouping Processing Flow

Figure 8: Grouping Architecture

These operations are performed by virtual processors called grouping nodes. Grouping nodes are seen as covering regular and adjacent portions of an image area (not necessarily of a single image, because there can be multiple images in a motion sequence). The image area contains some portion of a label plane for accessing the objects based upon their spatial dispositions as well as object-based associated attributes. The grouping nodes are further organized in a hierarchical pyramid shown in Figure 8. Each node is connected to its adjacent neighbors and has a parent and descendants. The transfer of information between nodes at different levels is based upon interestingness. Lower level processes send their most interesting structures up the hierarchy. There are several effects of this. One is that it allows a uniform processing to occur at different levels, so grouping rules can be applied to objects at different levels of interestingness. It also allows relations between nonspatially adjacent structures to be handled in a uniform architecture. It also partitions perceptual structures in a way that corresponds to different levels of control in instantiation of object models.

Organizing segmentation in terms of grouping processes has many advantages for a model based vision system. The grouping processes can be run automatically from extracted significant structures based upon perceptually significant, though non-semantic criteria. Thus, connected curves of slowly changing orientation or compact, homogeneous regions can be extracted purely on perceptual criteria. These image structures correspond to world structure and events, and they are useful for initializing schema instantiations. They correspond to the qualitative image predictions associated with more general schemas. An inference process for compilation from an object model into grouping processes, allows model based vision to have a very active character quite different from single-level attribute matching.

5. SCHEMAS

Schemas represent hypotheses about objects in the world. The process of schema instantiation creates an instance of a schema together with evidence for that schema. Evidence consists of structures in the PSDB, a priori knowledge stored in the LTM, predictions derived from location inference, and relations to already instantiated schema.

Table 1 shows the various slots and relationships in a generic schema. Although this data structure has a frame-like appearance, it is useful to view the schema as a semantic net structure, with slots representing nodes in the net and relationships representing arcs. Schema instantiation inference reasons from a (partially) instantiated node, follows arcs, and infers procedures to execute from the sum of its acquired information in order to obtain more evidence to further instantiate the schema.

The schema network is a generic set of data structures that indicate the a priori relationships between schemas. A key part of this network is the inheritance hierarchies that indicate which descriptions and relationships can be inherited from schema to schema. Inheritance hierarchies allow efficient matching of objects in the world against sensor evidence from progressively coarser to finer levels. As reasoning moves from coarser to finer levels of description in model-based schema instantiations, the schemas inherit descriptive bounds and add new descriptions, and also add constraints to inherited ones. For example, the system

Table 1: Generic Schema Data Structure

- SCHEMA TYPE
- SCHEMA NAME
- SCHEMA INSTANTIATION STRUCTURE
- 3D DESCRIPTION
  - SHAPE
  - SIZE
  - COLOR
  - TEXTURE
  - INDEX TO SENSOR VIEWS
- SENSOR VIEWS
(FOR EACH SENSOR)
(FOR EACH VIEW)
  - PROJECTION RELATIONS
    - PROJECTION FUNCTION
    - 3D BACK CONSTRAINTS
  - DISTINCTIVE IMAGE BASED EVIDENCE
  - PERCEPTUAL STRUCTURE
- COMPONENTS
  - MUST HAVE
  - MAY HAVE
  - 3D SPATIAL RELATIONSHIPS
  - VIEW DEPENDENT RELATIONSHIPS
- PART OFS
- CLASSIFICATIONS
(POINTS UP THE IS-A HIERARCHY ONE LEVEL)
- CONTEXTUAL RELATIONSHIPS
  - ALWAYS OCCURS WITH
  - SOMETIMES OCCURS WITH
  - NEVER OCCURS WITH
  - CONFUSED WITH
  - SIMILAR TO
- LOCATIONAL INFORMATION
  - LOCAL MULTI-SENSOR FRAME ALIGNMENTS
  - GRID ALIGNMENTS
  - 3D SPATIAL RELATIONSHIPS WITH
- RECOGNITION STRATEGIES

may first recognize an object as a terrain patch (because it lies on the ground plane). A road is a type of terrain patch (see Figure 1, that adds linear boundary description, and constrains the visual image appearance of the terrain patch schema in the color and texture descriptors. The two basic types of schema network inheritance hierarchies are IS-A and PART-OF.

Below is a brief explanation of each of the slots and relationships in the generic schema data structure. Schema type refers to the generic name of the schema in the IS-A hierarchy. Schema name is the identification of the schema instance, e.g., if the schema type is "road" then the schema name might be "highway 101". The schema instantiation structure maintains the control history of the schema recognition inference processes for this schema.

The 3D description is an object-centered view of the world object represented by the schema. It includes its 3D geometry and shape description, actual size, and inherent color and texture (as opposed to how its color and texture might appear to a particular sensor). Note that this is the description that matches the schema-object before looking at its structure refined into components. For example, the 3D geometric description of a tree schema does not separate the canopy from the trunk, but gives a single enclosing volume as its representation. The volumetric descriptions of the trunk and canopy appear as the 3D descriptors on their schema further down the PART-OF hierarchy. Thus, inferring down the PART-OF hierarchy corresponds to increasing the resolution of the view of the object represented by the schemas.

The sensor views are descriptions of the stable or frequently occurring appearances of the schema object in imagery. This description is intended to be used for image appearance prediction, evidence accrual for instance recognition, 3D shape inference, and location inference. The reason for storing or runtime generation of explicit (parametrized) image views is that the perceptual evidence matches to these descriptions, not to the three dimensional ones.

The distinctive image appearance slot holds descriptions of perceptual structures that are likely to occur bottom-up in the PSDB. They provide coarse triggers for instantiating the schema object hypothesis without prediction.

The perceptual structure is the dynamically created PSDB query history generated by the schema instantiation as it attempts to fill in evidence matching the various schema slots and relations. The instantiator can re-use successful branches of perceptual structures to improve its recognition speed as it continues to view other instances of the same generic schema type.

Components are pointers to other schema that represent sub-parts of the schema object. They are finer resolution description of the schema, one level down on the PART-OF hierarchy. The MUST-HAVE components are assumed to be parts the represented object must have to exist, although the schema may be instantiated without observing them all. Occasionally occurring components, such as center-lines on roads, can be stored in the MAY-HAVE slot. Spatial relationships between components as they make up the schema object are listed at this level also. Relationships can also be stored on a view dependent basis. These relationships access the sensor-view dependent data in that slot. PART-OF's point upward one level on the PART-OF hierarchy, indicating that this schema is a component of another schema.

Classification points upward and downward one level on the IS-A hierarchy. There may be more than one such pointer, which is to say that the IS-A hierarchy may be partially ordered.

Contextual relationships indicate spatial/temporal consonance or disconsonance between groups of schema types, omitting those which are already indicated in the PART-OF and IS-A hierarchies. Schema that ALWAYS or never-occur with the given one can be used strongly for belief or dis-belief in the schema instance and as focus of attention mechanisms within the instantiation process. SOMETIMES occurs with relationships that are used to store the spatial-temporal aspects of schemas relative appearance in the viewed environment.

CONFUSED-WITH and SIMILAR-TO relationships indicate schema that may be mistaken for the given one, but for different reasons. One schema may be confused with another because they share common evidence pieces, but for which there are sufficient descriptors to disambiguate. Two schema are similar if there is sufficient ambiguity in their appearances, and therefore the available perceptual evidence, that they may be indistinguishable without contextual reasoning. For example, tall grass may be confused with wheat from coarse shape and texture evidence, but can often be disambiguated by color descriptors or finer resolution examination of structure (because of wheat berries, for example). However, roads are similar to runways because they cannot necessarily be distinguished by their intrinsic appearance, no matter how detailed or accurate the descriptors and evidence. Contextual reasoning, e.g., the presence of aircraft on the runway, global curvature of the road, etc. is required.

Locational information points at the various viewframes the schema appears in and inferred 3D relationships with other world objects.

Recognition strategies are prioritization cues for the schema instantiation processes that suggest inference chains likely to pay off to match this schema instance against sensor evidence.

The recognition strategies slot in the schema data structure prioritizes inference approaches relevant to this schema. These approaches include search for components, search for part of schema instance, search on weaker classification, relations with other schema instances, and PSDB matching.

Search for COMPONENTS and search for PART-OF are both inferences along the PART-OF hierarchy in different directions. The instantiator searches the relevant slot to see if there are components to search for or another object of which this schema is a component. If the COMPONENT or PART-OF schemas exist, they can be accessed to continue the inference. Otherwise, each causes an instantiation of the missing schema to be generated as a prediction. Instantiation control can be transferred at this point to the COMPONENT or PART-OF schema. The schema inference process maintains its thread of reasoning relevant to the schema in the schema instantiation structure slot.

8. LONG TERM TERRAIN DATABASE

The long term terrain database is part of LTM. It stores the data necessary for a mobile robot to perform vision-based navigation and guidance, predict visual events, such as landmarks and horizon lines, and to update and refine maps.

The long term terrain database contains a priori map data including government terrain grids, elevation data, and schemas representing instances of stable visual events recorded while traversing paths in the environment. The use of a priori map and grid data to predict percepts and to help guide image segmentation is shown in Section 5. The following presents a summary of a structure for spatial representation and inference that enables a robot to navigate and guide itself through the environment.

We first define the notion of a geographic "place" in terms of data about visible landmarks. A place, as a point on the surface of the ground, is defined by the landmarks and spatial relationships between landmarks that can be observed from a fixed location. More generally, a place can be defined as a region in space, in which a fixed set of landmarks can be observed from anywhere in the region, and relationships between them do not change in some appropriate qualitative sense. Data about places is stored in structures called viewframes, boundaries and orientation regions.

Viewframes provide a definition of place in terms of relative angles and angular error between landmarks, and very coarse estimates of the absolute range of the landmarks from our point of observation. Viewframes allow the system to localize its position in space relative to observable local landmark coordinate systems. In performing a viewframe localization, observed or inferred data about the approximate range to landmarks can be used. Errors in ranging and relative angular separation between landmarks are smoothly accounted for. A priori map data can also be incorporated. A viewframe is pictured in Figure 20.

A viewframe encodes the observable landmark information in a stationary panorama. That is, we assume that the sensor platform is stationary long enough for the sensor to pan up to 360 degrees, to tilt up to 90 degrees (or to use an omni-directional sensor [Cao et.al. - 86]), to recognize landmarks in its field of view, or to buffer imagery and recognize landmarks while in motion.

A sensor-centered spherical coordinate system is established. It fixes an orientation in azimuth and elevation, and takes the direction opposite the current heading as the zero degree axis. Then two landmarks in front of the vehicle, relative to the heading, will have an azimuth separation of less than 180 degrees. If we assume that no two distinguished landmark points have the same elevation coordinates (i.e., no two distinguished points appear one directly above the other) then a well-ordering of the landmarks in the azimuth direction can be generated. We can speak of the landmarks as being "ordered from left to right". The relative solid angle between two distinguished landmark points is now well defined.

Under the above assumptions, the system can pan from left to right, recognizing landmarks, L_i, and storing the solid angles between landmarks in order, denoting the angle between the i-th and j-th landmarks by Ang_{ij}. The basic viewframe data are these two ordered lists, $(L_1, L_2, \ldots)$ and $(\mathrm{Ang}_{12}, \mathrm{Ang}_{23}, \ldots)$. The relative angular displacement between any two landmarks can be computed from this basic list. In [Levitt et.al. - 87] we show how to use this data to essentially parametrize all possible triangulations of our location relative to a set of simultaneously visible landmarks. This localizes the robot's position in space relative to a local landmark coordinate system.

Viewframes contain two basic dimensions of data: the relative angles between landmarks, and the estimated range (intervals) to the landmarks. If we drop the range information, we are left with purely topological data. That is, it is impossible, using only the relative angles between landmarks, and no range, map or other metric data, to determine the relative angles between triples of landmarks, or to construct parametric representations of our location with respect to the landmarks. Nonetheless, there is topological localization information present in the ordinal sequence of landmarks: there is a sense in which we can compute differences between geographic regions, and observe which region we are in.

The basic concept is to note that if we draw a line between two (point) landmarks, and project that line onto the (possibly not flat) surface of the ground, then this line divides the earth into two distinct regions. If we can observe the landmarks, we can observe which side of this line we are on. The "virtual boundary" created by associating two observable landmarks together thus divides space over the region in which both landmarks are visible. We call these landmark-pair-boundaries (LPB's), and denote the LPB constructed from the landmarks L_1 and L_2 by LPB(L_1,L_2).

Roughly speaking, if we observe that landmark L_1 is on our left hand, and landmark L_2 is on our right, and the angle from L_1 to L_2 (left to right) is less than 180 degrees, then we denote this side of, or equivalently, this orientation of, the LPB by $[L_1\ L_2]$. If we stand on the other side of the boundary, LPB(L_1,L_2), "facing" the boundary, then L_2 will be on our left hand and L_1 on our right and the angle between them less than 180 degrees, and we can denote this orientation or side as $[L_2\ L_1]$ (left to right).

More rigorously, define:

$$\text{orientation-of-LPB}(L_1,L_2)$$
$$= \text{sign}(\pi - \Theta_{12}) = \begin{array}{l} -1 \text{ if } \Theta_{12} < \pi \\ 0 \text{ if } \Theta_{12} = \pi \\ -1 \text{ if } \Theta_{12} > \pi \end{array}$$

where Θ_{12} is the relative azimuth angle between L_1 and L_2 measured in an arbitrary sensor-centered coordinate system. Here, an orientation of -1 corresponds to the $[L_1\ L_2]$ side of LPB(L_1,L_2), -1 corresponds to the $[L_2\ L_1]$ side of LPB(L_1,L_2) and 0 corresponds to being on LPB(L_1,L_2). It is a straightforward to show that this definition of LPB orientation does not depend on the choice of sensor-centered coordinate system.

LPB's give rise to a topological division of the ground surface into observable regions of localization, called orientation regions. Crossing boundaries between orientation regions leads to a qualitative sense of path planning based on perceptual information. The three levels of spatial representation given by map or metric data, viewframes and orientation regions are pictured in Figure 9. A

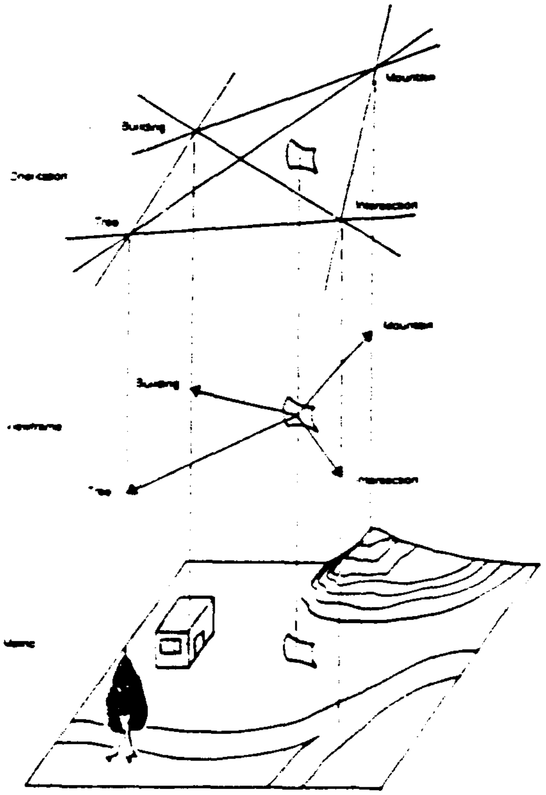


Figure 9: Multiple-Levels-of-Spatial Representation

natural environmental representation based on viewframes recorded while following a path is given by two lists, one list of the ordered sequence of viewframes collected on the path, and another of the set of landmarks observed on the path. We call the viewframe list a viewpath. The landmark list acts as an index into the viewpath, each landmark pointing at the observations of itself in the viewframes. For efficiency, the landmark list can be formed as a database that can be accessed based on spatial and/or visual proximity. Visual proximity can be observed, or computed from an underlying elevation grid and a model of sensor and vision system resolution.

The first occurrence of a landmark points at the instantiated schema or perceptual structure in the vision system database that was used to gather evidence in the landmark recognition process. After that, all recognized re-occurrences of this landmark point back at this initial instance. The same is true for the first occurrences and successful re-recognition of LPB's and viewframes. This mechanism allows multiple visual path representations, built at different times, to be incrementally integrated together as they are acquired by using a common landmark indexing/pointer list.

We use an environmental representation for orientation-region reasoning that is a list of oriented LPB's encountered and crossed in the course of following a path. We call such a list an orientation-path. As with viewpaths, there is an associated landmark list that indexes into the orientation-path.

A dynamically acquirable environmental representation that merges the representations for viewpaths and orientation-paths consists of an ordered list interspersing viewframes. LPB crossings, and appearance and occlusion (or loss of resolution) of landmarks, as well as recording the headings taken in the course of following the path over which the environmental map is being built. Thus, we can integrate the representations required for viewframe and orientation region based reasoning with heading and landmark information to formulate an environmental representation that supports hybrid strategies for navigation and guidance. The representation is formed at runtime and consists of multiple interlocking lists of sequential, time ordered, lists of visual events that include those necessary for the navigation and guidance algorithms presented in Levitt et.al. - 87 .

7. PROCESSING EXAMPLE

The following processing example demonstrates the behavior of some implemented system components. These include the format of predictions from the long term terrain model, the extraction of perceptually significant groupings from the PSDB, how an instantiated schema uses grouping processes and queries over the PSDB, and extracting relevant cues for making viewframe localizations in the long term terrain representation.

Figure 10 shows the elevation contours and road network in the a priori terrain data from the Martin Marietta ALV test site in Denver which was supplied by the U.S. Army Engineer Topographic Laboratories (ETL). The vehicle position on the road is indicated by the arrow in the figure. From this, we are able to roughly determine the correspondence between an image taken from the road

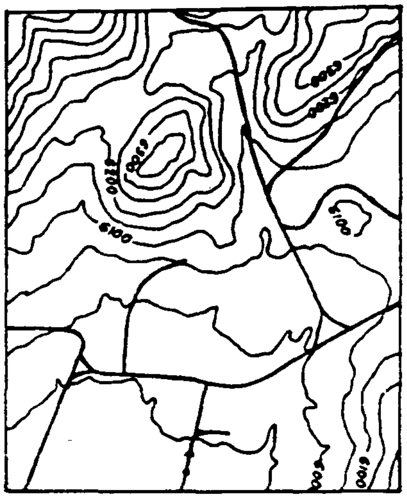


Figure 10: Terrain Data

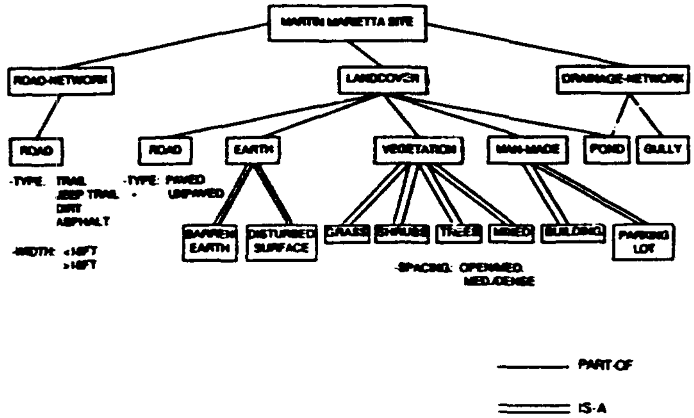
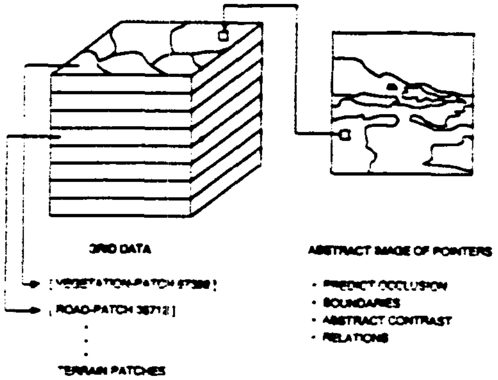


Figure 11: A Prior Terrain Type Classification



ꟻigure 12: Predicted Segmentation From Grid Data

and the terrain data. (the relevant sensor parameters were not available). Figure 11 shows the terrain and feature classification supplied with the a priori data. These correspond to sets of image overlays in register with the elevation data. The road network is stored as a set of curve objects that is decomposed into linear segments with supplied attributes, such as road material and width. Terrain patches are extracted as regions from terrain type information and parametric surface fits to the a priori elevation data.

Figure 12 shows how the grid registered terrain data is instantiated into STM to form a predicted segmentation. The grid data regions from connected analysis correspond to schema instances in the Long Term terrain memory. Established surface display techniques are used to project the elevation with the associated schema instances to form a predicted view. Image positions are then labeled with their associated schema instances. Additionally, there are many schema instances, ordered by depth, at the corresponding image locations. The resulting predicted segmentation is processed as an abstract image where critical perceptual events are determined by size, adjacencies across occlusion boundaries, or types of terrain with high semantic contrast, such as water, fields, or man-made structures. The perceptual structures are merged together based upon distances and semantic type to yield predictions at different resolutions.

Figure 13 shows the predicted terrain patches for the vehicle positioned with respect to the terrain in Figure 10. Figure 14 shows the predicted segmentation after filtering to pull out the horizon line and road/terrain discontinuities for roads near the vehicle. This data is quite coarse (30m sampling), and image areas in the foreground are highly composite containing instances of road and the adjacent grassy fields. Nonetheless, the predicted segmentation yields a qualitative description of predicted image features that is sufficient to initialize and direct grouping processes to find corresponding image features and relationships. The key characteristics of the predicted segmentation are that the vehicle is on a flat plane, and that its field of view consists of road and grassy field terrain patches with some mountains in the distance. Predictions of the dirt road off to the right and the intersection are made from the road-network and the elevation information stored along with it. The predictions are in terms of constraints on region adjacencies across boundaries, and the shape and attributes, such as color contrasts, of the boundaries themselves. The horizon line constraints are that it will tend to have smoothly changing orientation and be adjacent to a large homogeneous region (the sky). In general, the predicted features are described with constrained attributes determined from the visibility components of schemas.

Figures 15 and 16 show some of the contour related structures in the initialized PSDB. Figure 15 shows the edges extracted at one spatial resolution using the Canny edge operator [Canny - 83]. We have found it useful not to apply noise suppression to extracted segments in order to base filtering on structural properties of the contours, including linear deviation and relationships to other image structures. Different linear segment fits for this extracted edge images are shown in Figure 16.

Figure 17 shows the results of grouping processes applied to a set of selected curves in Figure 12 with multiple associated attributes for orientation and color contrasts. The grouping processes were constrained by the predicted segmentation in Figure 14 using constraints on allowable color contrasts, changes in linear segment orientation, and rough image position and extent. Multiple groups are obtained for each predicted image event. Selection of one, or maintaining multiple alternative groups, is explicitly represented in the schema instantiation structure. Here, groups were selected based upon length and uniformity of composite attributes.

Figure 18 shows the results of a road schema instantiation based upon matches to extracted road boundaries in accounting for road surface properties through PART-OF relations. Texture elements adjacent to the road boundary which are consistent with a road surface, such as low contrast, parallel edges corresponding to tread marks, are used to direct queries to instantiate potential road area. Queries are also used to determine the presence of anomalous structures in the road such as anything which is high contrast or oriented perpendicular to the road direction. Such structures require disambiguation through instantiation of another schema (it could be a road marking) cued by the anomaly or elevation estimates derived from motion displacements or range sensing.

Significant image structures near the horizon line are particularly important for landmark extraction. Figure 19 shows extracted interesting perceptual groups near and above the horizon line. Figure 20 shows an extracted viewframe representing the relative visual spatial relationships between some of the objects extracted from this field of view.

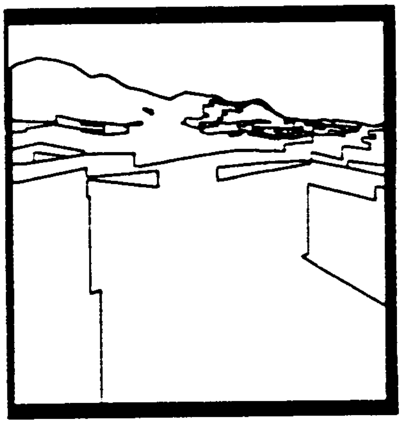


Figure 13: Terrain Patches

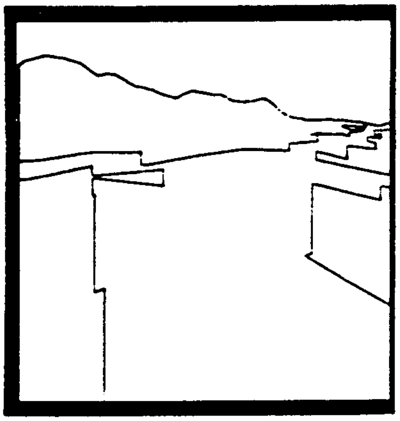


Figure 14: Merged Terrain Patches

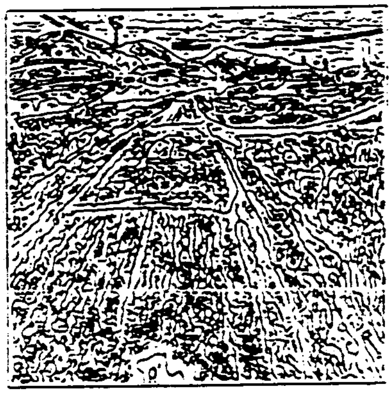


Figure 15: Canny Operator

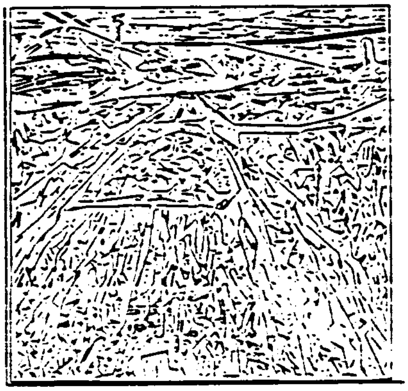


Figure 16: Linear Segment Fits

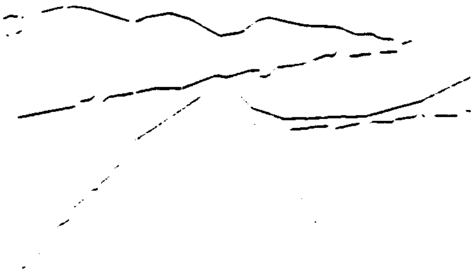


Figure 17: Contour Groupings

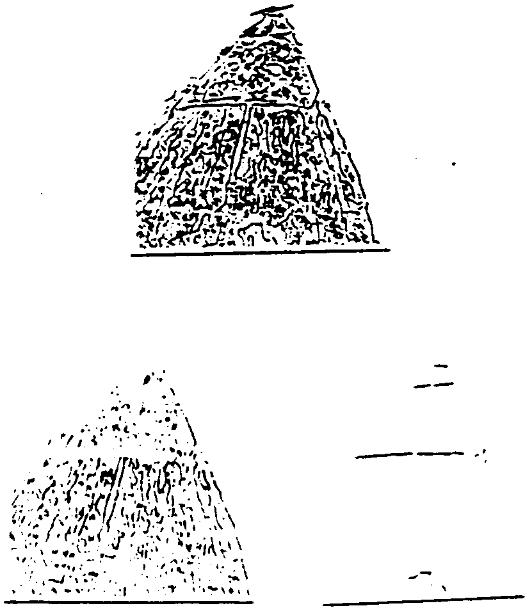


Figure 18: Road Schema Instantiation

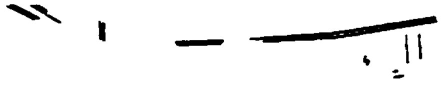
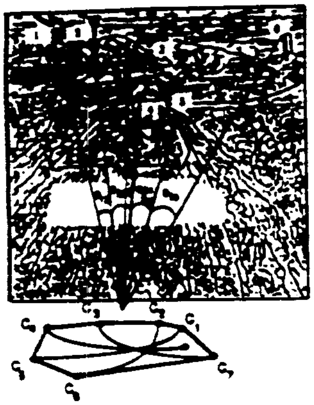


Figure 19: Significant Perceptual Groups



3D-ANGLES

[illegible]

[illegible] FOV)

RANGES

PEAK	[illegible]	[illegible]
POLE	[illegible]	[illegible]
NOTCH	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
INTERSECTION	[illegible]	[illegible]
BUILDING	[illegible]	[illegible]

LOCAL-COORDINATE-SYSTEM: (POLE, NOTCH, INTERSECTION, BUILDING)

PERCEPTUAL-LOCALIZATION-REGION: $(C_1, C_2, C_3, C_4, C_5, C_6, C_7)$

Figure 20: Viewframe Instance

8. SUMMARY

The architecture we have developed, using terrain and road schemas with implemented system components for perceptual processing and manipulating long term terrain data, has been successfully used in tasks for ALV navigation and scene interpretation.

9. ACKNOWLEDGEMENTS

This document was prepared by Advanced Decision Systems (ADS) of Mountain View, California, under contract number DACA76-85-C-0005 for the U.S. Army Engineer Topographic Laboratories (ETL), Fort Belvoir, Virginia, and the Defense Advanced Research Projects Agency (DARPA), Arlington, Virginia. The Contracting Officer's Representative at ETL is Ms. Rosalene M. Holecheck. The DARPA point of contact is Lt. Col. Robert Simpson.

The authors wish to thank Jeanne Gadol and Angela Erickson.

10. REFERENCES

Brooks - 84 - R.A. Brooks, "Model-Based Computer Vision", Computer Science: Artificial Intelligence, No. 14, UMI Research Press, 1984.

Canny - 83 - J. Canny, "A Variational Approach to Edge Detection", In Proceedings of the National Conference on Artificial Intelligence (AAAI-83), pp. 54-58, August 1983.

Cao et.al. - 86 - Z. Cao, S. Oh, and E. Hall, 'Dynamic Omnidirectional Vision for Mobile Robots", Journal of Robotic Systems, Vol. 3, No. 1, 1986, pp. 5-17.

Hanson et.al. - 78 - A.R. Hanson and E.M. Riseman, "VISIONS: A Computer System for Interpreting Scenes", In Computer Vision Systems, Academic Press, 1978.

Kuan - 84 - D.T. Kuan, "Terrain Map Knowledge Representation for Spatial Planning", 1st National Conference on AI Applications, pp. 578-584, 1984.

Kuipers - 82 - B.J. Kuipers, "Getting the envisionment right", In Proceedings of the National Conference on Artificial Intelligence (AAAI-82), Pittsburgh, Pennsylvania, August 1982.

Levitt et.al. - 87 - T. Levitt, D. Lawton, D. Chelberg, and P. Nelson, "Qualitative Navigation", Defense Advanced Research Projects Agency Image Understanding Workshop, Los Angeles, California, February 23-25, 1987.

[Marr - 82] - D. Marr, "Vision", W.H. Freeman, San Francisco, 1982.

[Martelli - 76] - A. Martelli, "An application of heuristic search methods to edge and contour detection", Commun. ACM, Vol. 19, No. 2, February 1976, pp. 73-83.

[Ohta - 80] - Y. Ohta, "A Region-Oriented Image-Analysis System by Computer", Ph.D. Thesis, Kyoto University, Department of Information Science, Kyoto, Japan, 1980.

[Pentland - 83] - A. Pentland, "Fractal-Based Description of Natural Scenes", In Proceedings Image Understanding Workshop, Arlington, Virginia, June, 1983, pp. 184-192.