Progress in Computer Vision at the University of Massachusetts[1]

Allen R. Hanson, Edward M. Riseman
Computer Vision Laboratory
Computer Science Department
University of Massachusetts
Amherst, MA 0l003
URL: http://vis-www.cs.umass.edu/

Abstract

This report summarizes progress in image understanding research at the University of Massachusetts over the past year. Many of the individual efforts discussed in this paper are further developed in other papers in this proceedings. The summary is organized into several areas:

1. Image Exploitation under RADIUS
2. Terrain Reconstruction
3. Daedalus Battlefield Visualization
4. Learning in Vision
5. Image Databases
6. Content-based Image Indexing
7. Related Research

The research program in computer vision at UMass has as one of its goals the integration of a diverse set of research efforts into systems that are ultimately intended to achieve real-time image interpretation in a variety of vision applications.

1. Image Exploitation under RADIUS

1.1. The Ascender Automated Site Modeling System

Under the ARPA/ORD RADIUS program, UMass is developing a set of image understanding (IU) modules for performing several geometric site modeling tasks including initial site model acquisition, extension, and refinement (Collins, Cheng et al. 1995). The UMass design philosophy emphasizes model-directed processing, rigorous 3D photogrammetric camera models, and fusion of information across multiple images for increased accuracy and reliability.

Three years of design and development has resulted in the UMass ASCENDER system (Automated Site Construction, Extension, Detection and Refinement) for automatically populating a site model with 3D building models extracted from multiple, overlapping images (both nadir and oblique) of the site. To acquire a new site model, an automated building detector is run on one image to hypothesize potential building rooftops. Supporting evidence is located in other images via epipolar line segment matching, and the precise 3D shape and location of each building is determined by multi-image triangulation. Projective mapping of image intensity information onto these polyhedral building models results in a realistic site model that can be rendered using virtual “fly-through” graphics. As new images of the site become available, model extension and refinement procedures are performed to add previously unseen buildings and to improve the geometric accuracy of the existing 3D building models. In this way, the system gradually accumulates evidence over time to make the site model more complete and more accurate.

In mid-April 1995, Version 1.0 of the Ascender system was delivered for testing on classified imagery to Lockheed-Martin; at the same time, an informal transfer was made to the NEL for familiarization and additional testing. Feedback from both of these groups has resulted in several system improvements and an overall “hardening” of the code.

The Ascender system is very effective at finding flat-roofed buildings in “clean” imagery (good resolution and visibility, not too much scene clutter). A thorough evaluation is currently underway to determine system performance with respect to various system variables. To this end, an evaluation test set has been created consisting of seven cropped subimages from the recent

[1] This research has been supported in part by the Defense Advanced Research Projects Agency under contracts DACA76-92-C-0041 (through TEC), F0602-94-C-0042 (through USAF Rome Laboratory), DAAL02-91-K-0047 (through ARL) and DAAG04-93-G-0422 (through Colorado State University); by the National Science Foundation under grant IRI-9208920; and by the Office of Naval Research under grants N00014-95-1-1035 and N00014-89-J-3229.

RADIUS Ft. Hood dataset. Thirty ground truth building models were created by hand using interactive modeling tools provided by the RCDE. Each building is composed of RCDE "cube", "house" and/or "extrusion" objects that were shaped and positioned interactively to project as well as possible (as determined by eye) into the images comprising the test dataset. The Ascender system was then tested on the dataset to determine how well it performed at different grouping sensitivity settings and with different length and contrast settings of the Boldt line extraction algorithm. At the highest sensitivity settings, 71 out of the 73 rooftop polygons were found by combining results from four images in the dataset. A more complete description of the test procedure and an analysis of the results obtained can be found in (Collins, Hanson et al. 1996) in these proceedings.

Although Ascender was initially designed to extract a single generic class of buildings, namely flat-roofed, rectilinear structures (e.g. rectangular box-shapes, L-shapes, U-shapes,), extensions are underway that allow the system to handle other building classes. Some initial results on peaked-roof and multi-level flat-roofed buildings are reported in (Collins, Hanson et al. 1996; Jaynes, Stolle et al. 1996) in these proceedings.

1.2. Context-Sensitive Strategies for Three-Dimensional Grouping

As noted earlier, extensions to the Ascender system must be developed in order to handle other common building types. Examples include multi-level flat roofs (or single-level flat roofs containing significant substructures such as large air conditioner units), peaked-roof buildings, juxtaposition of flat and peaked roofs, curved-roof buildings such as Quonset huts or hangars, as well as buildings with more complex roof structures containing gables, slanted dormers or spires. Furthermore, a general site model reconstruction must include the relationship between buildings and the surrounding terrain. Buildings are often located on slopes and near contextually significant terrain features such as vegetation, river beds and rock outcroppings.

To achieve the desired goal of a more general and flexible building extraction system, a significant research effort is underway at UMass to explore alternative detection and reconstruction strategies that combine a wider set of algorithms for generating and fusing 2D and 3D information. The types of bottom-up grouping strategies being considered involve processing of a wider variety of geometric tokens, as well as techniques for fusing line data with high-resolution Digital Elevation Map (DEM) data. In addition to the 2D line, corner and closed polygons currently extracted by the system, new techniques have been developed to extract 3D lines, corners and surfaces. By verifying geometric consistencies between 2D and 3D tokens associated with building components, larger and more complex 3D structures are organized using context-sensitive, knowledge-based strategies.

Several grouping/fusion strategies for 3D building detection and reconstruction are being developed. For example, one strategy uses grouping of 2.5D lines (horizontal lines with an elevation) to form 3D rooftop polygons. This immediately disambiguates the 2D polygons of a multi-level roof by grouping structures at different heights. Another strategy involves parametric surface fitting constrained to a 2D polygonal rooftop hypothesis which has been projected to the digital elevation map. Three types of rooftop surface fits have been attempted to date: planar, peaked, and curved. Since the DEM data is potentially noisy, due to rooftop clutter and mismatches near occlusion boundaries, robust statistical estimation techniques are used to do surface fitting. The 2.5D lines and surfaces can also be grouped if they satisfy the appropriate geometric constraints for each distinct building type. For further details see (Collins, Hanson et al. 1996; Jaynes, Stolle et al. 1996) in these proceedings.

1.3. Recovery of 3D Edgels via Multi-Image Matching

The problem of feature matching and 3D reconstruction across multiple views is considered in (Collins 1996) . The term "true multi-image matching" is introduced to describe techniques that make full and efficient use of the geometric relationships between multiple images and the scene. Three conditions are placed on a true multi-image method: it should generalize to any number of images, the algorithmic complexity should be linear in the number of images, and every image should be treated on an equal footing, with no one image singled out for preferential treatment as a reference view. Several multi-image matching techniques that have been reported in the literature are found not to pass this set of conditions. A few techniques that do pass the test all explicitly reconstruct scene structure in object space, while determining correspondences in image space. Object space seems to be the conduit through which successful multi-image methods combine information from each image.

A new space-sweep approach to true multi-image matching is presented that simultaneously determines 2D feature correspondences between multiple images together with the positions of observed 3D point, line or edgel features in the scene. The method is based on the premise that areas of space where several viewing rays intersect are the likely locations of observed 3D scene features. The approach can be visualized as sweeping a plane through space, while noting positions on it where many backprojected feature rays intersect. A careful examination of the projective relationships between the images and the plane in space, and between different positions of the sweeping plane, shows that the feature mappings involved can be performed very efficiently. A statistical model is also developed to tell how likely it is that a given number of viewing rays will pass through an area of space by chance, thus determining whether the results of the matching/reconstruction procedure are likely to be correct. Based on a simple random distribution on feature clutter in the image, this statistical model guides the setting of algorithmic parameters in order to optimize various statistical criteria, such as minimizing the predicted number of false positive matches that are found.

2. Terrain Reconstruction

The UMass terrain reconstruction system (Schultz 1995) employs a hierarchical image matching scheme that is specifically designed to recover the three-dimensional structure of textured surfaces from two overlapping images that may have been taken from oblique and widely separated viewpoints and with large base-to-height ratios. While our system can be applied effectively to nadir images, possibly widely spaced, it also takes advantage of quite oblique views that often cause problems with other reconstruction algorithms.

When digital elevation maps (DEMs) are computed from widely separated images, perspective effects may cause a large number of false matches and poor reconstruction accuracy. When the base-to-height ratio[2] (*b/h*) exceeds ~ 0.5, the performance of image matching algorithms begin to deteriorate, and when *b/h* becomes greater than 1, elevation errors caused by perspective effects become large. Thus, different views of the same ground patch at 45 degrees or more from nadir will produce highly foreshortened pixels on opposite sides of the patch, and the effectiveness of standard template correlation can have problems. For operational scenarios involving UGV, RADIUS, and UAV programs, aerial images are often taken from distant reconnaissance sources, and an ability to deal with large base-to-height ratios would be advantageous.

The three key elements of our algorithm are modifications to standard methodology and designed specifically to recover a DEM from images taken from widely varying viewpoints. These techniques include a new correlation match score, and modifications to standard subpixel and hierarchical matching techniques. These variations have been incorporated into a stereo terrain analysis package. We have compared this system to standard techniques by processing sequences of simulated images with values of *b/h* that varied between 0.25 and 2.25, and a single pair of high altitude images with a base-to-height ratio of 0.63. Analysis of the simulated data showed that the reconstruction accuracy of the system remains independent of base-to-height ratio.

The system has also been applied to a series of real aerial image data over the past two years, including Ft. Hood and the Lockheed-Martin Demo B site (Jaynes, Stolle et al. 1996; Riseman, Schultz et al. 1996). No ground truth for these two data sets are available. However, qualitative evaluation of the algorithm shows that the shadows are quite consistent with ground protrusions and the sun angle, roofs are flat where they should be, vehicles stand out from the ground plane, obvious features such as ditches appear to be correct, and in the Fort Hood data the foliage stands clearly above the surrounding ground plane. Until quantitative data is supplied with test data, complete validation and measurement of the accuracy of the algorithm is not possible. The battlefield visualization project described in the next section utilizes the terrain reconstruction system.

3. Daedalus Battlefield Visualization

The Daedalus system is being developed in response to a perceived need to greatly improve battlefield awareness via visualization and information fusion in the ARPA Unmanned Ground Vehicle (UGV) program. Real-time rendering of terrain views provides battlefield commanders a powerful new tool for the planning of operations, as well as an extremely effective interface for briefing the plan to those who will implement it. However, while the advantages of these planning tools are just beginning to be exploited, the staleness of the data they rely on inhibit their use for real-time monitoring and assessment of combat operations. Novel means are required to provide commanders with the ability to instantly comprehend dynamic battlefield situations and to react accordingly.

The goal of the Daedalus program is a revolutionary advance in providing dynamic battlefield awareness at the command center. The problem of battlefield awareness for the future commander is that he will not be co-located with the multiple, distributed sensors of future systems, and typically will not be intimately familiar with the battlefield terrain. The result can be a serious loss of ‘presence’ on the battlefield, leading to a lack of awareness of the true battlefield situation. The Daedalus system fuses information from field reconnaissance vehicles (including ground scout

[2] The base-to-height ratio is defined as the perpendicular distance from the midpoint of the line joining the camera centers and a nominal point on the surface.

vehicles and air vehicles) and collateral intelligence with high resolution terrain reconstructions of the battlefield, which are overlaid with military symbology and cultural features derived from ITD data. Realistic graphical models of both friendly and enemy vehicles are inserted into the terrain model and animated with estimated vehicle velocities. The result is a single, consistent 3D model of the dynamic battlefield. Dynamic control over battlefield viewpoint will provide commanders with the ability to focus on and assess changing conditions and provide unprecedented access to timely and relevant data.

High-performance graphic workstations will allow real-time 3D visualizations of the evolving battlefield situation. The system will allow the commander to roam at will around the battlefield (viewing the situation from any perspective) without being restricted by the original sensor positions. By fusing all of the available information, including both remote sensor and human input, an integrated and consistent view of the battlefield can be represented in a dramatic, easy to interpret, 3D representation, with asynchronous information updates from multiple UGV scouts and other ground sensors. A more complete conceptual overview of the Daedalus system is presented in (Riseman, Schultz et al. 1996) in these proceedings.

4. Learning in Vision

4.1. Need for Learning in Complex IU Applications

Although the field of image understanding has made significant advances over the past 20 years, we have not yet developed the theoretical insight or practical methodology for the effective integration of the many aspects of vision into functioning systems. Advances are taking place in the application of machine learning technology to the integration and control of existing IU algorithms to allow the creation of practical, robust systems for special-purpose applications.

In many ways, the stage has been set for learning in machine vision by the research of the past twenty years. Computer vision researchers have been dividing naturally - without any global consensus or mandate - into 10 or 20 subfields with smaller, better defined problems. This has led to clear focus and development of theories and techniques for the different subdisciplines. There are now several good and improving algorithms for edge and line extraction (straight and curved), stereo analysis, tracking, depth from motion (two-frame and multi-frame), shape recovery, 2D and 3D object recognition, 3D pose determination, and knowledge-based focus of attention, and this list can easily be extended in scope. Indeed, computer vision researchers have made more progress than most outside the field (and many inside) are aware of! This state of affairs is due primarily to our inability to easily produce highly visible results in the form of integrated task-specific systems. It is as though we have the supplies and tools to build a house, but lack the architectural drawings.

Many visual tasks -- even within constrained domains such as ATR, object recognition, change detection, or visual inspection -- require specialized software and hardware solutions. This is a great limitation to the wide and flexible use of computer vision. Trainable vision systems would overcome this limitation and would meet the needs for reliable and autonomous performance.

The need for robust and flexible techniques that can adapt to the user without requiring extensive explicit programming and customization is especially high in some areas of vision such as ATR where the context, models, and algorithms are expected to vary significantly. Vision systems that *learn and adapt* are one of the most important directions in IU research right now. This reflects an overall trend -- to make intelligent systems that do not need to be fully and painfully programmed. It is the only way for us to develop vision systems for the military that are robust and easy to use in many different tasks.

Unfortunately, many IU learning efforts are focused on parameter optimization as a major issue in achieving optimal performance in ATR applications since they involve tuning parameters to the data in the context of particular sensors, CAD target objects, illumination, weather conditions, terrain type, etc. However, as soon as there are changes in data or problem definition (e.g. new targets or environmental context), the system's performance is no longer optimal, and all too often breaks down entirely. Our contribution is a focus on *learning the control strategies* (as discussed in the next section) to combine and sequence algorithms so that systems remain robust in the face of these problems.

New methodologies for learning are being based on recent theoretical and experimental results in “learning from examples” using neural network architectures, statistical classification procedures, reinforcement learning algorithms, and discrete control techniques. One should be able to transform the task of constructing a new vision application to one of training the system with a set of representative input-output examples relevant for the task. Given the libraries of available IU algorithms and representations, the goal is to select sequences of algorithms, their associated parameters, and intermediate representations to optimize specific applications automatically, while minimizing the involvement of the user.

These techniques of automated machine learning provide us a general framework from which the more specific theories of control in computer vision can arise. The potential payoff of developing these capabilities would be to allow the effective compilation of robust and efficient application-specific systems, finally unleashing the full power developed over the past twenty-five years of IU research to solve a myriad of military and industrial problems.

4.2. The Schema Learning System

The UMass Schema Learning System (SLS) uses machine learning techniques to automatically assemble computer vision algorithms into working special-purpose computer vision systems. This project seeks to understand both the scientific and pragmatic issues in *learning efficient and robust control strategies* for specific visual tasks.

The construction of special-purpose computer vision systems is modeled as a control problem, in which the aim is to find the optimal policy for achieving a particular goal, such as recognizing a particular object instance or class of objects in a given domain. System actions are the selection and sequencing of image understanding routines developed here and elsewhere in the IU community. The set of possible legal sequences is syntactically defined by the input and output parameters of each vision routine. The states of the system being controlled are collections of intermediate image understanding representations, such as line segments, image regions and surface patches.

The control policy is built from a library of IU procedures that transform data from one representation to another. Two types of learning algorithms are being studied inside the SLS experimental framework: (1) Symbolic learning systems that exhaustively apply all IU procedures to a (small) set of training images, and then infer minimum-cost policies through a symbolic algorithm similar to Mitchell's Version Spaces algorithm (Mitchell 1978), and (2) reinforcement learning systems, which incrementally refine a control policy using a variant of Sutton's Temporal Difference (TD) algorithm (Barto, Sutton et al. 1983). In both approaches, decision trees and/or neural networks are used to group similar data instances (e.g. regions with similar attributes) into equivalence classes, thereby allowing the problem to be cast as a discrete control problem.

4.3. Learning 2D and 3D Grouping Strategies

Obviously, the construction of effective and robust strategies becomes more difficult as the set of algorithms employed change (or are updated), and as the characteristics of the image domain or semantic content vary. Both are likely, and construction of such strategies is extremely time-consuming, labor intensive and quite likely to be non-optimal. A major experiment in learning in computer vision is under way; see (Draper, Riseman et al. 1996) in these proceedings for details. SLS is being trained to accomplish tasks in building detection and reconstruction from the RADIUS project. The challenge for SLS is to automatically learn a control policy that selects which techniques to use to construct building models and to determine how to combine their (sometimes contradictory) results. If successful, these experiments will demonstrate SLS's ability to customize control strategies for complex, real-world tasks.

The first of these tasks is to recognize two-dimensional roofs of buildings in aerial photographs. The approach being taken in the RADIUS program is one of developing hand-crafted strategies for finding rooftops by grouping line segments, analyzing shadows, and exploiting other 2D and 3D image cues. Our goal for SLS is to automatically learn an equivalent (or better) strategy based on the same type of information *without the human in the loop*.

Ideally, SLS's library should contain the same subroutines used by previous hand-crafted rooftop recognition systems. Unfortunately, SLS's procedures must be stand-alone executable UNIX modules, while many of the subroutines developed by RADIUS researchers are Lisp functions in RCDE (Gerson and Wood Jr. 1994). Therefore, a small set of RADIUS-style vision routines have been reimplemented for this experiment, many although not all from UMass. In particular, algorithms have been employed for extracting straight lines (Boldt, Weiss et al. 1989) corners formed by orthogonal line segments, pairs of parallel line segments, grouping this data into closed polygons, and for re-examining the image top down to look for missing data. Alternatively, the image can be segmented into regions that may (or may not) support the edge/corner/line data, while shadows provide another source of information about rooftops. The 3D heights of line segments can be determined by matching them across multiple, overlapping images; see (Collins, Hanson et al. 1996) in these proceedings as well as (Collins, Cheng et al. 1995; Collins, Hanson et al. 1995). The challenge for SLS is to find a control policy among the huge number of possibilities for applying these algorithms that maximizes system performance. In initial experiments reported in (Draper, Riseman et al. 1996) SLS succeeded at this task in 9 of 10 trials.

As part of the ongoing experiments, a second RADIUS task being assigned to SLS is to reconstruct complex 3D buildings from multiple, overlapping aerial images. Eventually, this task will include the modeling of buildings with multi-level roofs, peaked roofs, curved roofs, and large structures (such as air conditioning and/or water storage units) on the roof; it is likely that each type of building will require a different context-sensitive grouping strategy.

Although all of the algorithms for 2D rooftop recognition are also applicable to this task, what makes 3D building reconstruction possible is the introduction of new visual procedures that extract 3D information from overlapping images. In addition correlation-based stereo matching of overlapping images can produce dense elevation maps (Schultz 1995); by fitting surfaces to the projections of polygons on the elevation data, the position and orientation of roofs and other surfaces can again be determined. Once again, the task of SLS is to find a policy that uses both the 2D and 3D procedures to their greatest advantage.

4.4. Color-Based Focus of Attention Mechanisms for ATR

An object recognition technique based on RGB color information originally developed for finding landmarks in the ARPA UGV program has proven to be remarkably effective at generating target focus-of-attention regions in ATR applications. The system has been extensively tested this year at the UGV Demo C and by an ARPA sanctioned study by LGA, Inc. In Demo C, 100% of the targets were found; in the LGA test on the Ft. Carson RSTA images, about 98% of the targets were detected. In both cases, there were a number false positives (3-4 false positives per image); this suggests that the technique provides very reliable focus-of-attention for ATR that can be used in conjunction with FLIR or other sensors operating outside the visual spectrum. The usefulness of the system is not restricted to military tasks -- it can also be used to recognizing landmarks, traffic signs, and even certain types of bushes and trees. A brief technical description (Buluswar and Draper 1994) follows.

As the lighting conditions of an object vary, the color of many objects forms a continuous region in three-dimensional (red, green, blue or RGB) color space. We use a non-parametric function approximation scheme to estimate this region based on training images. In other words, given samples of the object's color under different lighting conditions, the method 'learns' the boundary of the region in color space. In particular, we use multivariate decision trees, which perform piecewise linear function approximation. After the training phase is complete, every image pixel can be classified as object or non-object according to whether it lies within the confines of the learned color region or not. The result is a binary 'region-of-interest' image that marks all the pixels that lie within the object's color space. This allows the system to use an RGB lookup-table for classification, enabling it to run in almost real-time on standard hardware.

For the machine vision researcher, this is a novel method for object recognition using color information. Our seemingly simple method of pixel classification, without using any of the physics-based models for illumination, reflection or imaging, addresses the problem with far more success than most other techniques. The application of this technique to RSTA is being done under the ARPA RSTA (Reconnaissance, Surveillance and Target Acquisition) program in cooperation with Colorado State University (Beveridge, Draper et al. 1996).

5. Image Data Bases

5.1. Databases and IU

Problems related to image databases are becoming increasingly important in a wide variety of DOD applications. This is partly the result of an ever-increasing number of sensors operating at higher and higher resolutions, leading to an explosion of image information. Database problems also result from increasingly demanding sensory processing tasks: in the very near future, the military expects to interpret data from multiple sensors more quickly and with respect to more tasks than before (e.g. RSTA, photo-interpretation, battlefield monitoring, bomb-damage assessment), and to distribute this derived data to a larger set of recipients, quickly if not in real-time.

For example, the UAV Tier 2+ and Tier 3- programs (Delaney 1995) are expected to deploy reconnaissance aircraft which have the ability to stay aloft for over 24 hours and continuously monitor 40,000 sq. nm per day using visible light and infrared, and eventually synthetic aperture radar sensors. These aircraft are expected to produce data streams of 50 Mbits per second to satellites and 137 Mbits per second to dedicated ground stations, with a total of over 1 terrabyte of data per UAV per day. Large portions of this data must be stored, accessed, and processed to provide timely information about the current state of the dynamic battlefield.

5.2. Issues

Image Understanding applications need to process, store, and access a variety of visual information, including not only images, but also the wide range of representations associated with the intermediate and final results of the many vision algorithms making up an IU system. To support DOD applications like the ones discussed in previous sections, an image database system must actually fulfill three roles:

(1) Interface/Representation. A database is an important component in an image interpretation system since it provides the mechanisms for defining the set of common representations used by the IU algorithms to store and exchange

information. Thus, an image database system must deal not only with the sensor data (images) but also the multitude of data types produced as a byproduct of the interpretation process, some of which must be saved. Furthermore, this representation must be multi-level and extensible if the database is to be applied to new IU domains.

(2) Data exchange. The database system must serve as a data conduit, passing data from sensors through a (possibly distributed) processing system to the end-user. Depending on the application, this communication may or may not have to be real-time.

(3) Database storage and retrieval. As military sensors collect ever more data, the ability to efficiently store this data and relevant intermediate representations derived from interpreting the images, and then retrieve it quickly in response to a user or system query is crucial.

Unfortunately, neither current image databases nor commercial RDBs satisfy these requirements.

5.3. ISR4 - A Persistent Object Store

ISR4 is the fourth generation of a software tool for supporting data management and process integration in complex image understanding systems. It extends ISR3 (Draper, Kutlu et al. 1994), a real-time data store system coded in C++ and serving as an IU software development support tool, by combining it with the capabilities of Mneme (Moss 1990), a persistent object store. Our goal is to achieve visual data persistence in IU applications while overcoming the pitfalls of commercial relational database systems in terms of performance, flexibility and user-defined extensibility by allowing a customizable indexing and query capability

The current ISR3 provides a database programming interface for development of IU software. The interface is equipped with a hierarchy of visual representations, I/O support for visual data in a variety of common file formats, and graphics tools for displaying visual data. It also provides an initial set of indexing techniques for storing and retrieving visual data.

Mneme provides a flexible low-level object store with customizable indexing, clustering, and buffer management policies. These policies can be modified to reduce data transfer delays, page swaps, and total disk space. An example of customizing storage and access for IU applications is the ISR4 *tile-image* format, where an image is clustered on disk, and only required sub-images are brought into memory (on-demand), to significantly reduce the number of page swaps during common image processing operations.

The combined system provides a database programming interface for IU applications that can be customized for highly efficient operation; see (Kutlu, Draper et al. 1996) in these proceedings for more detail. Multi-dimensional indexing techniques are being explored to support efficient storage and retrieval strategies for spatial and temporal image data, and query languages for visual database applications. We are investigating two domains in particular: 1) content-based image retrieval (see next section), in which large numbers of images must be indexed according to color, texture, and/or shape; and 2) the RADIUS photo-interpretation project, in which very large images and their interpretations must be efficiently accessed according to geographic position, time, and content, and where the primary operations are change detection and site model construction.

6. Content-Based Image Indexing

6.1. National Center for Intelligent Information Retrieval (CIIR) at UMass

The National Center for Intelligent Information Retrieval (CIIR), funded by the National Science Foundation, has been established at UMass to conduct basic research and technology transfer in the areas of information retrieval from very large volumes of text. Research in CIIR focuses on the integration of techniques from information retrieval, natural language processing, case-based reasoning, database systems, and image processing/computer vision.

A major development of CIIR is the INQUERY full text retrieval engine (Brolio, Callan et al. 1994). This engine consistently achieves among the highest overall performance in objective evaluations within the annual Department of Commerce/ Department of Defense co-sponsored national Text Retrieval Conference (TREC). INQUERY is in use in major branches of the U. S. government, in private industry, and in academia and has been ported to Japanese, Chinese, Spanish, and Finnish language versions. A joint research project has been established between CIIR and the computer vision lab to enhance INQUERY to include visual information retrieval capabilities.

6.2. Retrieving Similar Images from Databases

Today, text retrieval engines like INQUERY are able to rapidly retrieve relevant textual information contained in documents on-line. However, a considerable amount of information exists in the form of images and it would advantageous to have automatic methods for indexing and retrieving them based on their content. Retrieval may be based on visual attributes such as color, texture, shape or appearance. Retrieval based on color or texture, however, may fail because of the wide variability of colors that even a small class of objects may possess. A characterization based on shape or appearance will not suffer from these problems, although use of shape does require prior segmentation of the database into objects. Since automatic segmentation is an unsolved problem, such segmentation must be done manually and hence is also a laborious process.

Ravela and Manmatha (Ravela, Draper et al. 1996; Ravela, Manmatha et al. 1996; Ravela, Manmatha et al. 1996, to appear) have proposed a strategy for retrieving images using appearance. The appearance of images in the database is encoded by filtering them with Gaussian derivatives at several scales. The user then interactively defines a query template from the query image. This query consists of user-defined regions in the image along with their spatial locations. The user selects regions based on features that characterize the objects or scene in question. For example, for a car viewed from one side, these features could be the two wheels of the car. The query may, therefore, be visualized as a template with certain portions zeroed out. The template is also filtered with Gaussian derivatives. The template, which consists of a vector of Gaussian derivative images, is now correlated with the filter outputs of the database. A search is carried out over scale space to ensure that a variation in scale (size) can be handled. The database images are ranked using the correlation score. The user may then redefine the query template based on the retrieval results to improve retrieval performance. Experiments demonstrate that similar images are usually ranked much higher than dissimilar ones. No prior segmentation of the images is required and the technique works with viewpoint changes up to about 20 degrees as well as illumination changes.

6.3. Image Retrieval Using Color Cues

With the advent of large image databases, efficient query-based retrieval of images has become an important issue. M. Das has been exploring the use of color information as the basis of image retrieval and has applied it in domains that have distinct color signatures; one example of such a domain is product advertising and logos. She has developed a very fast retrieval scheme with a small overhead. This technique uses the peaks of the color histogram in HSV space to describe the color content of a query image and retrieves similar images which match the color peaks in the query image. Peak matching incorporates an inexact matching technique and is robust across images where the query object occurs in different sizes, orientations and backgrounds. An index table is created off-line which can be directly indexed by the peaks during the query stage, resulting in very fast retrieval. A somewhat slower verification phase is added to remove most false matches when accuracy is more important that speed. The verification phase, which is based on localizing the peaks in a region of the original image and matching the characteristics of the region with the query, can also be used for identifying the query object's size and orientation in the image.

This retrieval scheme works best for multi-colored query objects. The scheme has been tested on a database of color product advertising images. A recall score of 96% and a precision score of 86% was attained for the test set which contained product images that varied in size, orientation, and background. Since this scheme is fast and requires small overhead, it is suitable for use in large color image databases which contain multi-colored objects like consumer products, fabrics, company logos, and natural domains like flowers, birds, insects and minerals.

7. Related Research

7.1. Robust Object Tracking Across Aspect Changes for Augmented Reality

A novel augmented reality system for temporal registration of a modeled object using a single camera has been developed as part of a project on developing an interactive repair manual (Ravela, Draper et al. 1996). In this application technicians look through a visor at an annotated object; together, the object's image and the overlaid annotations provide directions for the next repair step. Robust, real-time registration is possible through 360° of out-of-plane object rotation, as demonstrated (Ravela, Draper et al. 1995) in a real-time system (approximately 8Hz update rate on a Pentium laptop) for a complex object (a portable electric generator).

The temporal registration system is initialized with a known set of model-image correspondences, known camera parameters, a pre-compiled aspect table that associates discrete viewpoints with object features visible from those views, and an estimated initial pose obtained from an initialization step. Once initialized, the system follows a simple three-step pose-tracking loop at each new frame: a) pose information is used to index into the aspect table and a list of visible features is extracted, b) these features are projected into the image to hypothesize the new locations of the feature templates, c) these templates are matched in the new frame and a new pose is computed. The tracker localizes feature templates in search windows around hypothesized locations using steerable filters and normalized cross-correlation. This technique is relatively insensitive to changes in lighting conditions and compensates fully for feature translations and rotations in the image-plane; it also compensates for some non-trivial out of plane rotations. Pose computation (Kumar and Hanson 1994) uses camera parameters and the model-image correspondences to robustly solve for a rotation and translation that minimizes the projection error of model points on the object.

7.2. Using Weighted Bipartite Matching to Solve the General 3-D Rigid Motion Problem

When the correspondences between the points are not known and there exist missing data or outliers, the problem of bringing two 3D point sets into alignment is difficult In the case of rigid body motion, this involves recovery of the transformation (rotation and translation) between the point sets. Traditional correspondence methods suffer from combinatorial complexity while correspondenceless methods have a difficult time dealing with outliers.

(Wang, Cheng et al. 1995) have developed a new approach for solving the general 3-D rigid body motion problem. For ideal cases with no missing points, a closed-form solution based on eigenstructure decomposition is presented to recover the correspondences of the two point sets and their motion. For general cases where missing points occur, it is proven that the correspondence problem can be reduced to a weighted bipartite matching problem. Using a heuristic measure of "point pair affinity" derived from the eigenstructure decomposition, an efficient weighted bipartite matching algorithm is used to determine the correspondences. A fast outlier removal algorithm is combined with the weighted bipartite matching algorithm to improve the correspondence recovery. The algorithm is shown to be robust in the face of both noise and outliers in both simulations and experiments on real data.

7.3. Surface Reconstruction from Occluding Contours

For the reconstruction of curved surfaces from occluding contours with known camera motion, one of the boundaries of reconstruction is known as the *frontier curve* (Giblin and Weiss 1995). This trajectory-dependent curve is one that separates regions that can be reconstructed from those that cannot and has many interesting properties. It can be defined as the envelope of the set of contour generators for a continuous sequence of views. Intuitively, this means the envelope is the set of intersections of successive contours. It is also the set of points where the epipolar plane is tangent to the surface. In practice, for discrete camera motion, it is the intersection of the contour generators for two successive views. This property makes it useful when the camera motion is unknown, since these points are visible on the occluding contour in two views. Giblin and Weiss have been continuing their investigation into the properties of the frontier, which may provide robust solutions to the important problem of identifying these points.

7.4. Qualitative and Quantitative Obstacle Detection

During autonomous navigation, it is necessary to know when an obstacle is in the path of the vehicle. Zhang has been examining the problem of obstacle detection from the point of view of the environmental assumptions and the available geometric constraints. In cases where only a qualitative result (yes or no) is necessary concerning the presence of an obstacle in a selected field of view, two algorithms have been developed whose response depends upon the existence of a solution to a linear system. These algorithms are monocular or stereo, do not generate a complete 3D reconstruction, assume a planar road surface, and are fast. A third stereo algorithm continuously estimates the ground plane parameters and reconstructs partial 3D structures by determining the height of each point in the scene with respect to the ground plane. Derivation of the algorithms and experimental results can be found in (Zhang, Weiss et al. 1996, to appear; Zhang, Weiss et al. 1994).

More recently, Zhang has shown that based on four coplanar correspondences of two externally uncalibrated cameras, 3D reconstruction can be achieved in Euclidean space with only one uniform scale factor and up to two real solutions. The scale factor is the physical distance from the camera center to the plane formed by the four points in 3D space. Consequently, if this distance is known a priori, then the 3D structure can be determined completely up to two solutions. To disambiguate the two solutions, a third view is required in general. These results are discussed further in (Zhang and Hanson 1996) in these proceedings.

References

Barto, A. G., R. S. Sutton and C. W. Anderson. (1983). "Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems," IEEE T-SMC, Vol. 13(5).

Beveridge, J. R., B. A. Draper and K. Siejko. (1996). "Progress on Target and Terrain Recognition Research at Colorado State University," Proc. of ARPA Image Understanding Workshop, Palm Springs, CA.

Boldt, M., R. Weiss and E. M. Riseman. (1989). "Token-Based Extraction of Straight Lines," IEEE Trans. SMC, Vol. 19(6), pp. 1581-1594.

Brolio, J., J. Callan and W. B. Croft. (1994). "INQUERY System Overview," Proc. of TIPSTER Text Program (Phase I), San Francisco, CA, pp. 47-67.

Buluswar, S. and B. Draper. (1994). "Non-Parametric Classification of Pixels Under Varying Outdoor Illumination," Proc. of ARPA Image

Understanding Workshop, Monterey, CA, pp. 1619-1625.

Collins, R. (1996). “A Space-Sweep Approach to True Multi-Image Matching,” Proc. of ARPA Image Understanding Workshop, Palm Springs, CA.

Collins, R., Y. Cheng, C. Jaynes, F. Stolle, X. Wang, A. R. Hanson and E. M. Riseman. (1995). “Site Model Acquisition and Extension from Aerial Images,” Proc. of International Conference on Computer Vision, MIT, Cambridge, MA.

Collins, R., A. Hanson, E. Riseman and H. Schultz. (1995). “Automatic Extraction of Buildings and Terrain from Aerial Images,” Proc. of Automatic Extraction of Man-Made Objects from Aerial and Space Images, Ascona, Switzerland, pp. 169-178.

Collins, R. T., A. R. Hanson, E. M. Riseman, C. O. Jaynes, F. Stolle, X. Wang and Y. Q. Cheng. (1996). “UMass Progress in 3D Building Model Acquisition,” Proc. of ARPA Image Understanding Workshop, Palm Springs, CA.

Delaney, W. P. (1995). “The Changing World, The Changing Nature of Conflicts: A Critical Role for Military Radar,” Proc. of IEEE International Radar Conference, pp. 11-15.

Draper, B., G. Kutlu, A. R. Hanson and E. M. Riseman. (1994). “ISR3: Communication and Data Storage for an Unmanned Ground Vehicle,” Proc. of International Conference on Pattern Recognition, Jerusalem, Israel.

Draper, B. A., E. M. Riseman and A. R. Hanson. (1996). “Learning Grouping Strategies for 2D and 3D Object Recognition,” Proc. of ARPA Image Understanding Workshop, Palm Springs, CA.

Gerson, D. J. and S. E. Wood Jr. (1994). “RADIUS Phase II-The RADIUS testbed system,” Proc. of APRA Image Understanding Workshop, pp. 231-237.

Giblin, P. and R. Weiss. (1995). “Epipolar Curves on Surfaces,” Image and Vision Computing, Vol. Volume 13(Number 1), pp. 33-44.

Jaynes, C., F. Stolle, H. Schultz, R. Collins, A. R. Hanson and E. M. Riseman. (1996). “Three Dimensional Grouping and Information Fusion for Site Modeling from Aerial Images,” Proc. of ARPA Image Understanding Workshop, Palm Springs, CA.

Kumar, R. and A. R. Hanson. (1994). “Robust Methods For Estimating Pose and a Sensitivity Analysis,” CVGIP-IU, Vol. 60(3), pp. 313-342.

Kutlu, G., B. A. Draper, J. E. B. Moss, E. M. Riseman and A. R. Hanson. (1996). “Persistent Data Management for Visual Applications,” Proc. of ARPA Image Understanding Workshop, Palm Springs, CA.

Mitchell, T., “Version Spaces: An Approach to Concept Learning”, Ph.D. Thesis, TRMIT78, Dept. of Stanford University, Computer Science, 1978.

Moss, E. (1990). “Design of the Mneme Persistent Object Store,” ACM Transactions on Information Systems, Vol. 8(2), pp. 103-139.

Ravela, S., B. Draper, J. Lim and R. Weiss. (1995). “Adaptive Tracking and Model Registration Across Distinct Aspects,” Proc. of IROS ‘95, Pittsburgh, PA.

Ravela, S., B. Draper, W. J. Lim and R. Weiss. (1996). “Tracking Object Motion Across Aspect Changes for Augmented Reality,” Proc. of ARPA Image Understanding Workshop, Palm Springs, CA.

Ravela, S., R. Manmatha and E. Riseman. (1996). “Scale-Space Matching and Image Retrieval,” Proc. of ARPA Image Understanding Workshop, Palm Springs, CA.

Ravela, S., R. Manmatha and E. Riseman. (1996, to appear). “Retrieval From Image Databases Using Scale-Space Matching,” Proc. of ECCV, Cambridge, England.

Riseman, E., H. Schultz, W. J. Lim, B. Draper and A. Hanson. (1996). “Daedalus Battlefield Visualization System,” Proc. of ARPA Image Understanding Workshop, Palm Springs, CA.

Schultz, H. (1995). “Terrain Reconstruction from Widely Separated Images,” Proc. of SPIE, Orlando, FL, pp. 113-123.

Wang, X., Y. Cheng, R. Collins, A. Hanson, R. Weiss and R. Moll, “Using Weighted Bipartite Matching to Solve the General 3-D Rigid Motion Problem”, TR95-106, Comp. Sci. Dept., Univ. of Massachusetts, Amherst, MA, 1995.

Zhang, Z. and A. Hanson. (1996). “3D Reconstruction Based Homography Mapping,” Proc. of ARPA Image Understanding Workshop, Palm Springs, CA.

Zhang, Z., R. Weiss and A. Hanson. (1996, to appear). “Obstacle Detection on Qualitative and Quantitative 2D Reconstruction,” IEEE T-PAMI.

Zhang, Z., R. Weiss and A. R. Hanson. (1994). “Qualitative Obstacle Detection,” Proc. of IEEE Computer Vision and Pattern Recognition Conference, Seattle, WA, pp. 554-559.