Progress in Computer Vision at the University of Massachusetts

Edward Riseman
Allen Hanson

COINS TR92-23

March 1992

Progress in Computer Vision at the University of Massachusetts

Edward M. Riseman and Allen R. Hanson
Computer Vision Research Laboratory
Dept. of Computer and Information Science
University of Massachusetts
Amherst, MA 01003

ABSTRACT[1]

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. Mobile Robot Navigation
2. Image Sequence Processing
3. Interpretation of Static Scenes
4. Image Understanding Architecture

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

1. Robot Navigation

The UMass mobile robot navigation project continues to integrate a number of different algorithms with the goal of achieving robust landmark-based navigation. The primary component technologies are described in several sections of this paper. Sections 1.1 and 1.2 discuss the use of landmarks derived from a partial geometric model of the environment to determine the pose of the vehicle. Section 2.2 outlines one mechanism by which an initial partial model of the environment might be automatically acquired from a motion sequence. Section 3.1 discusses techniques for learning recognition strategies within the Schema object recognition system, which is capable of identifying naturally occurring objects. Our goal is to integrate natural objects into the landmark-based navigation system for outdoor navigation and to embed the results of this research into the planning and control framework developed by Fennema [18] which can effectively utilize landmarks at a number of levels,

[1]This research has been supported in part by the Defense Advanced Research Projects Agency under TACOM contract number DAAE07-91-C-R035, HDL contract number DAAL02-91-K-0047, and ETL contract number DACA76-89-C-0016, by the National Science Foundation under grant CDA-8922572 and IRI-9113690, and by RADC under contract number F30602-91-C-0037

including low-level perceptual servoing for producing accurate motor actions, and plan-level perceptual servoing for maintaining adherence to a navigation plan.

1.1. Automated Model Acquisition and Extension

We are continuing our efforts towards the robust determination of pose (location and orientation) of the vehicle in a partially modelled 3D environment via the constraints derived from recognized 3D landmarks [25].

Recently, Kumar has performed experiments on model extension (Kumar and Hanson [25, 26] using basic techniques from pose determination. Points or lines whose 3D positions are known are tracked across frames using the line tracking algorithm of Williams [39], or the point-tracking algorithm of Sawhney [33]. From these points or lines, the relative orientation of pairs of frames are determined. The depths of unmodelled points, which are also tracked over the sequence, are then computed using triangulation. The sensitivity of the acquired depth to errors in the image center has also been investigated. In experiments using two image sequences for which ground truth is available, the 3D positions of the unmodelled points were recovered with an average error in depth of .25% and 1.3%. The error for the second case is larger than for the first in part due to the larger field of view (40^{o} compared to 22^{o}) which increases the sensitivity to errors in the location of the image center. Given that there must be some error in the original 3D positions of landmarks, recovery of new 3D points to this accuracy is a surprising and dramatic result.

1.2. Landmark Recognition

Beveridge [3] continues to develop his model-directed matching algorithms, which are being applied to landmark-based robot navigation. His previous research used a priori knowledge of an approximate robot pose to project landmarks in a 3D model into the image plane. This allows 2D model lines to be matched with 2D data lines extracted from the sensory data by minimizing the error in the spatial alignment under rotation, translation, and scaling transformations in the plane. The resultant correspondence between landmarks and image features for the best 2D-to-2D match is used to recover the actual 3D pose of the robot. This system has been effectively applied to pose recovery in the UMass mobile robot project. However, this system may not be able to recover from the 2D distortion produced when the projection is done from an incorrect sensor pose.

A new 3D-to-2D model matching system has been developed to match 3D landmarks directly to 2D image features. During the iterative matching process, new 3D transformations between the world and the camera are computed, and landmark features are re-projected into the image. This accounts for perspective distortion during the search, and therefore allows recovery of the robot's true position more reliably than the 2D-to-2D matching system. However, the original system usually improves the initial erroneous pose estimate, and does so in roughly one fifth the time required by the second.

2. Image Sequence Processing

2.1. The p-Field: A Computational Model for Binocular Motion Processing

Balasubramanyam [1, 2] is developing an integrated framework for stereo and motion analysis. Given a binocular camera system moving through a static environment, it is possible to obtain a three-dimensional field of vectors, where each vector is parallel to the induced relative 3D motion of an imaged point and scaled in magnitude by the depth of that point. This 3D vector field, referred to as the p-field, is derived from optic flow and disparity, over a pair of stereo frames at successive time instants. The behavior of the p-field is examined for specific cases of restricted motion, as well as for general motion. In particular, the behavior of the p-field under translational vehicle motion promises to be more stable under small vehicle rotations than the behavior of the flow field. We expect that the p-field will allow more robust recovery of the sensor motion parameters, and tracking of 3D points through a sequence of images. Ultimately, this analysis should provide more robust recovery of 3D environmental information than independent stereo and motion analyses whose results are combined after the fact.

2.2. Reconstruction of Shallow Structures

In many man--made environments, obstacles in the path of a mobile robot can be characterized as *shallow*, that is, they have relatively small extent in depth compared to the distance from the camera. Sawhney [32] (these proceedings) presents a framework for segmenting shallow structures from the background over a sequence of images. Shallowness is first quantified in terms of *affine describability*. This is embedded in a tracking system within which hypothesized model structures undergo a cycle of prediction and model--matching. Structures emerge either as shallow or non--shallow based on their *affine trackability*. This work rejects continuity heuristics for purely image motion in favor of temporal continuity defined as the consistency of generic 3D

models, namely shallow structures. This work will be applicable to obstacle avoidance and model acquisition by a mobile robot. In two indoor experiments, object structure represented as frontal planes was recovered to a depth accuracy in the range of 2-5%.

2.3. Multi-Frame Structure from Motion

Recovering structure from motion, even using information from multiple image frames, is difficult, partly because motion error can introduce large, correlated errors in the structure estimate. Thomas and Oliensis [35] (these proceedings) propose a method for recursively recovering structure from motion that can deal with this problem. The algorithm is based on the observation that errors in the motion produce cross-correlations in the structure errors across the 3D points. Conversely, these correlations are the record of the motion error. Thus, to explicitly incorporate motion error in a recursive algorithm, a record of the correlations in the structure errors must be maintained and updated.

Input for the algorithm consists of point correspondences tracked over many image frames. Horn's relative orientation algorithm [23] is used to provide two-frame structure estimates. For this algorithm, a somewhat complex error analysis is used to estimate the expected structure errors, including the cross-correlations. The fusing of the new structure estimate with the old is done using a standard Kalman filter, but with the cross-correlations taken into account. The results on synthetic images show that the structure estimates improve over time as expected; encouraging results on real images are also reported [30, 34, 35].

3. Interpretation of Static Scenes

3.1. Learning 3D Recognition Strategies

In an effort to automate aspects of model acquisition for image interpretation, Draper [13-16] has been examining the role of learning in model-based vision. In particular, he is addressing the automated construction of robust control strategies responsible for creating 'instance-of' relations during interpretation. Given a set of generic parameterized knowledge sources, the goal is to construct the Schema Learning System (SLS) which will learn (from a set of training images) a recognition strategy for a particular object class that minimizes the cost of recognition, subject to a set of accuracy constraints supplied by the user. Recognition strategies are represented by recognition graphs, which are similar in many ways to decision trees. Unlike decision

trees, however, recognition graphs direct hypothesis creation as well as hypothesis verification. Object-specific strategies are learned in a two step process [15] (these proceedings). The first step involves learning which hypotheses should be generated. The second learns how to verify them efficiently. Thus, the task of SLS is to learn control and evidence combination strategies, not new models or knowledge sources. Initial experimental results demonstrate the potential of this approach.

3.2. View Description Networks

Model-directed object recognition becomes much more difficult when the viewpoint of the three-dimensional object is unknown [6-8]. Burns [5] (these proceedings) describes a system designed to effectively match a single 2D image of a potentially cluttered scene to a library containing multiple polyhedral objects and demonstrates its performance on several scenes. This approach to recognition can be characterized by three general ideas. *Description networks* optimize the search for matches to objects from a multiple object library by organizing information about the objects into a single network representation. *View descriptions* contain organized descriptions of the *projections* of the objects from views for which the objects are expected to be seen; these are used during the *match phase*. Finally, the correctness of view description matches are *verified* by estimating the 3D pose of the associated object, evaluating the estimation error and searching for additional assignments between object and images features, given the estimated pose.

An important part of this approach is the design of the *recognition phase* of the system: given a compiled view description network for an object library, the system must direct an effective search for the correct matches in cluttered images. This implies that three key problems are addressed: recognition from an unknown view, the indexing problem (selection of a few high probability candidates based on key features), and model based incremental recognition among the candidate competing hypotheses based on partial matches.

3.3. Model Extension Using Projective Invariants

Collins [9, 11] has developed a new approach to modeling man-made environments based on results from projective geometry. It is well known that the images of coplanar points and lines under rigid-body motion are related by a linear transformation in homogeneous coordinates. Given four known reference points or lines on the plane, the positions of all other points/lines on that plane can be reconstructed, regardless of camera position or intrinsic calibration parameters.

Collins [10] (these proceedings) has extended these results; it is shown that it is possible to obtain partial and in some cases complete 3D reconstructions of those points and lines lying outside the reference plane. The main results are that with a calibrated camera, one reference plane tracked through two images is enough for complete reconstruction of the environment, while for an uncalibrated camera it is sufficient to have two reference planes tracked through two images. The effects of noise in the observations are considered, resulting in a general framework for data fusion in projective space.

3.4. Shape from Shading Revisited

Shape from shading has traditionally been considered an ill--posed problem. However, in recent work, Oliensis [28, 29] has demonstrated that the solutions to shape from shading are often well--determined, with little or no ambiguity. For the case of illumination that is symmetric around the viewing direction (i.e. the light source is behind the camera), it was shown in [27] that there is in general a *unique* solution to shape from shading. This proof is valid for *general* Lambertian objects (without holes), and is the first proof that the problem of shape from shading can be well--posed in general. These arguments were extended to the case of general illumination direction in [29], where it was demonstrated that, in this case also, the solutions to shape from shading are strongly constrained over much of the image.

Recently, Dupuis and Oliensis [17] (these proceedings) has developed a new approach to shape from shading, based on a connection with a calculus of variations/optimal control problem, and has demonstrated its performance on reasonably complex images. The approach leads naturally to an algorithm for shape reconstruction that is simple, fast, provably convergent, and, in many cases, probably convergent to the correct solution. The algorithm is robust against noise and, in contrast with standard variational algorithms, does not require regularization. An explicit representation is given for the surface: its height is expressed as the minimal cost for an optimally controlled trajectory.

4. The Image Understanding Architecture

The IUA project continues as a three-way collaboration between UMass, Amerinex Artificial Intelligence, Inc. (AAI), and the Hughes Research Labs (HRL); this coordinated effort is summarized in [36] in these proceedings. The first IUA prototype hardware has been assembled, tested, and is almost fully functional. The

low-level processor for the second generation IUA and its controller have been designed, and a software simulator has been built for the controller and low-level array. The design for the intermediate level has just been completed.

UMass has developed a SIMD version of the Wormhole Routing technique that takes advantage of the Coterie Network in our low-level processor, in order to provide general permutation routing capability roughly equivalent in performance to that found in the Connection Machine, without the need for special hardware. The significance of this routing capability is that it allows us to build very compact, low-cost, mesh-based parallel processors, of reasonable size (up to about a million processors) that can perform general data-parallel processing.

Our experience with the Coterie Network has resulted in the description of a more general programming paradigm, called multi-associative processing. In turn, the capabilities of the Coterie Network have been explored for directly and indirectly supporting multi-associativity.

Consideration of a set of issues that must be addressed by a parallel symbolic database for intermediate-level processing is in progress. These include the problems of managing data from continuous streams of images, controlling persistence of the data, representations of the data, distribution of the data and maintenance of its consistency, and real-time systems issues.

A C++ class library has been implemented for an image plane data type that supports the development of low-level vision operations that are easily implemented by a parallel processor. This approach to parallel programming has the advantage that it does not involve a non-standard language -- it is merely a new object class written in C++. The only difference between a sequential implementation and a parallel implementation is the run-time library selected for linking.

5. Ongoing and New Work

5.1. Multi-Sensor Dextrous Manipulation

Grupen and Weiss [20, 21] are collaborating on a multi-sensor approach to dextrous manipulation in a robot workcell. Models of objects in the environment are constructed incrementally using an active sensing paradigm in order to support the

ability to form stable grasp configurations with a Utah-MIT hand. The system consists of a camera mounted on one robot arm and the hand mounted on another. The transformation from the camera coordinate system to the hand coordinate system is computed using the pose refinement algorithm developed in [24].

One of the major issues with respect to modeling is fusing information from multiple views and different sensors. The particular application involves the integration of haptic and visual data to produce a triangulation of the surface of an object to be grasped. Haptic sensing here is the determination of the point of contact of the hand with the object based on force measurements. This gives a very rough estimate of the position and normal to the surface. The Giblin-Weiss [19] algorithm provides estimates of position, surface normal and curvature from a sequence of image with known camera motion.

5.2. Figural Completion from Principles of Perceptual Organization

Visual psychology provides strong evidence that generic knowledge of surfaces and occlusion is exploited very early in the perceptual grouping of image contours. Previous work by Williams [38] showed how generic knowledge of this sort could be captured as integer linear constraints and how the problem of segmenting simple scenes into (potentially overlapping) surfaces could be cast as an integer linear programming problem. The first system built along these lines demonstrated the completion of gaps in straight sided figures, such as those caused by occlusion of one (opaque) surface by a second (opaque) surface, and subsequent recovery of the surfaces using only straight line interpolating contours. This limitation severely restricted the range of figures to which the system could be applied (reconstruction of occluded corners, for example, was impossible). In the past year, a new system has been built which uses cubic bezier splines of least energy as the interpolating contours. The system now captures curved illusory contours in figures that have been formulated by Kanizsa and other perceptual psychologists.

5.3. Perceptual Organization of Curves

Token-based grouping has thus far been applied to the problems of recovering straight-line structure [4] and more recently curvilinear structure [12] from the edge data of images. Dolan is currently extending this approach to local parallel implementations of this grouping paradigm. A SIMD model of curvilinear grouping has been designed to be implemented in the CAAPP layer of the IUA [37]. The model is relatively simple and promises many orders of magnitude speedup in extraction of

straight and curved lines. A MIMD version is currently being designed which should alleviate the contention problems in the SIMD design by utilizing both the CAAPP and ICAP layers.

5.4. Qualitative Navigation

One way to solve the computational burden of maintaining accurate geometric maps for navigation is to eliminate such maps altogether. In contrast to model-based approaches to navigation where a map is required that explicitly represents the geometry and location of 3D objects in the world, Pinette [22, 31] is developing a method for qualitative, image-based navigation via homing. This approach maintains only a topological map of the world, representing particular places in the world and the directions between neighboring places. A place is represented explicitly in the map by the image of the world as seen from that location. Spatial reasoning is performed directly on images using only the bearings of landmarks from a current location and a neighboring target location, and does not need to acquire exact shape and range information. The work is developing a theoretical foundation for qualitative reasoning in the incremental homing paradigm, including cases with a lack of precision in recovering the direction of landmarks, and the presence of errors in landmark correspondence.

References

1. Balasubramanyam, P. and M. A. Snyder. (1991). "The P-Field: A Computational Model for Binocular Motion Processing," Proc. of IEEE CVPR, Maui, Hawaii, pp. 115-120.
2. Balasubramanyam, P. and M. A. Snyder. (1992). "A Computational Model for Binocular Motion Processing," Proc. of DARPA IUW, San Diego, CA.
3. Beveridge, J. R. and E. M. Riseman. (1992). "Can Too Much Perspective Spoil the View? A Case Study in 2D Affine and 3D Perspective Model Matching," Proc. of DARPA IUW, San Diego, CA.
4. Boldt, M., R. Weiss and E. Riseman. (1989). "Token-Based Extraction of Straight Lines," IEEE Trans. SMC, Vol. 19, pp. 1581-1594.
5. Burns, J. B. and E. M. Riseman. (1992). "Matching Complex Images to Multiple 3D Objects using View Description Networks," Proc. of DARPA IUW, San Diego, CA.
6. Burns, J. B., R. Weiss and E. M. Riseman. (1990). "View Variation of Point Set and Line Segment Features," Proc. of DARPA IUW, Pittsburgh, PA, pp. 650-659.
7. Burns, J. B., R. S. Weiss and E. M. Riseman. (1992). "Non-existence of General-Case View-Invariants" in Geometric Invariance in Machine Vision (J. Mundy and A. Zisserman, Ed.). To appear.

8. Burns, J. B., R. S. Weiss and E. M. Riseman. (1992). "View-variation of Point-set and Line-segment Features," IEEE PAMI. To appear.

9. Collins, R. and R. Weiss. (1990). "Vanishing Point Calculation as a Statistical Inference on the Unit Sphere," Proc. of IEEE ICCV, Osaka, Japan, pp. 400-403.

10. Collins, R. T. (1992). "Single Plane Model Extension using Projective Transformations," Proc. of DARPA IUW, San Diego, CA.

11. Collins, R. T. and R. Weiss. (1990). "Deriving Line and Surface Orientation by Statistical Methods," Proc. of DARPA IUW, Pittsburgh, PA, pp. 433-438.

12. Dolan, J. and R. Weiss. (1989). "Perceptual Grouping of Curved Lines," Proc. of DARPA IUW, Palo Alto, CA, pp. 1135-1145.

13. Draper, B. (1992). *Learning Object Recognition Strategies*. Ph.D. Thesis, Dept. of Computer Science, University of Massachusetts (Amherst). Forthcoming.

14. Draper, B. and A. R. Hanson. (1991). "An Example of Learning in Knowledge-Based Vision," Proc. of 7th SCIA, Aalborg, Denmark, pp. 189-201.

15. Draper, B., A. R. Hanson and E. M. Riseman. (1992). "The Schema Learning System," Proc. of DARPA IUW, San Diego, CA.

16. Draper, B. and E. M. Riseman. (1990). "Learning 3D Object Recognition Strategies," Proc. of IEEE ICCV, Osaka, Japan, pp. 320-324.

17. Dupuis, P. and J. Oliensis. (1992). "Direct Method for Reconstructing Shape from Shading," Proc. of DARPA IUW, San Diego, CA.

18. Fennema, C. (1991). *Interweaving Reason, Action, and Perception*. Ph.D. Thesis, Dept. of Computer Science, University of Massachusetts. Also COINS Technical Report TR91-56, Department of Computer and Information Science, University of Massachusetts, Amherst, MA, September 1991.

19. Giblin, P. and R. Weiss. (1987). "Reconstruction of Surfaces from Profiles," Proc. of IEEE ICCV, London, pp. 136-144.

20. Grupen, R. and R. Weiss. (1991). "Force Domain Models for Multifingered Grasp Control," Proc. of IEEE Conf. Rob. & Aut., Sacramento, CA, pp. 418-423.

21. Grupen, R. and R. Weiss. (1991). "Issues in System Integration and Control for Interpreting and Manipulating the Environment," Proc. of IEEE Int'l. Sym. on Intelligent Control, pp. 221-226.

22. Hong, J., X. Tan, B. Pinette, R. Weiss and E. M. Riseman. (1990). "Image Based Navigation Using 360 Degree Views," Proc. of DARPA IUW, Pittsburgh, PA, pp. 782-791.

23. Horn, B. K. P. (1990). "Relative Orientation," IJCV, Vol. 4, pp. 59-78.

24. Kumar, R. (1992). *Model Independent Inference of 3D Information from a Sequence of 2D Images*. Ph.D. Thesis, Dept. of Computer Science, University of Massachusetts. Forthcoming.

25. Kumar, R. and A. R. Hanson. (1990). "Sensitivity of the Pose Refinement Problem to Accurate Estimation of Camera Parameters," Proc. of IEEE ICCV, Osaka, Japan, pp. 365-369.

26. Kumar, R. and A. R. Hanson. (1991). "The Application of 3D Pose Determination Techniques to 3D Model Extension," Proc. of IEEE Workshop on Passive Ranging, Princeton, NJ.

27. Oliensis, J. (1989). "Existence and Uniqueness in Shape from Shading," Dept. of Computer and Information Science, University of Massachusetts (Amherst), TR 89-109.

28. Oliensis, J. (1990). "Existence and Uniqueness in Shape from Shading," Proc. of IEEE ICPR, Atlantic City, NJ, pp. 341-345.

29. Oliensis, J. (1990). "Shape from Shading as a Partially Ill-Posed Problem," Dept. of Computer and Information Science, University of Massachusetts (Amherst), TR 90-50.

30. Oliensis, J. and J. I. Thomas. (1991). "Incorporating Motion Error in Multi-Frame Structure from Motion," Proc. of IEEE Workshop on Visual Motion, Princeton, NJ, pp. 8-13.

31. Pinette, B. (1991). "Qualitative Navigation," Proc. of IEEE Symp. IC, Arlington, VA.

32. Sawhney, H. and A. R. Hanson. (1992). "Tracking, Detection, and 3D Representation of Potential Obstacles using Affine Constraints," Proc. of DARPA IUW, San Diego, CA.

33. Sawhney, H., J. Oliensis and A. R. Hanson. (1990). "Image Description and 3D Interpretation from Image Trajectories," Proc. of IEEE ICCV, Osaka, Japan, pp. 494-498.

34. Thomas, J. I. and J. Oliensis. (1991). "Incorporating Motion Error in Multi-Frame Structure from Motion," Proc. of 7th SCIA, Aalborg, Denmark, pp. 950-957.

35. Thomas, J. I. and J. Oliensis. (1992). "Recursive Multi-Frame Structure from Motion Incorporating Motion Error," Proc. of DARPA IUW, San Diego, CA.

36. Weems, C., M. Herbordt, M. Scudder, J. Burrill and R. Lerner. (1992). "Current Status and Research in the Image Understanding Architecture Program," Proc. of DARPA IUW, San Diego, CA.

37. Weems, C., S. Levitan, A. Hanson, E. Riseman, J. Nash and D. Shu. (1989). "The Image Understanding Architecture," IJCV, Vol. 2(3), pp. 251-282.

38. Williams, L. R. (1990). "Perceptual Organization of Occluding Contours," Proc. of DARPA IUW, Pittsburgh, PA, pp. 639-649.

39. Williams, L. R. and A. R. Hanson. (1988). "Translating Optical Flow into Token Matches and Depth from Looming," Proc. of IEEE ICCV, Tarpon Springs, FL, pp. 441-448.