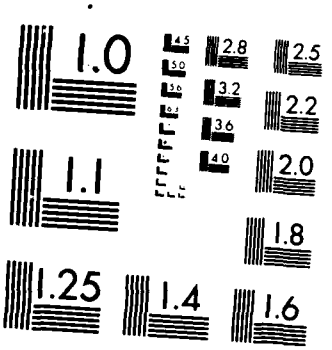
AD-A173 209 DYNAMIC IMAGE INTERPRETATION FOR AUTONOMOUS VEHICLE NAVIGATION(U) MASSACHUSETTS UNIV AMHERST DEPT OF COMPUTER AND INFORMATION SCIENCE E M RISEMAN ET AL. 1/1

UNCLASSIFIED SEP 86 ETL-0437 DACA76-85-C-0008 F/G 9/4 NL



ETL-0437

(2)

AD-A173 209

Dynamic image interpretation for autonomous vehicle navigation

Edward M. Riseman
Allen R. Hanson

University of Massachusetts
Amherst, Massachusetts 01003

September 1986

Prepared for
U.S. ARMY CORP OF ENGINEERS
ENGINEER TOPOGRAPHIC LABORATORIES
FORT BELVOIR, VIRGINIA 22060-5546

E T L

U.S. ARMY ENGINEER TOPOGRAPHIC LABORATORIES
ETL

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No 0704 0188
Exp Date Jun 30 1986

1a REPORT SECURITY CLASSIFICATION: Unclassified	1b RESTRICTIVE MARKINGS: AD-A173209
2a SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION / AVAILABILITY OF REPORT: Distribution of this report is unlimited.
2b DECLASSIFICATION / DOWNGRADING SCHEDULE	
4 PERFORMING ORGANIZATION REPORT NUMBER(S)	5 MONITORING ORGANIZATION REPORT NUMBER(S): ETL-0437
6a NAME OF PERFORMING ORGANIZATION: University of Massachusetts Computer & Information Sci. Dept. / 6b OFFICE SYMBOL (If applicable)	7a NAME OF MONITORING ORGANIZATION: Office of Naval Research
6c ADDRESS (City, State, and ZIP Code): Amherst, Massachusetts 01003	7b ADDRESS (City, State, and ZIP Code): Cambridge, Massachusetts 02138
8a NAME OF FUNDING / SPONSORING ORGANIZATION: U.S. Army Engineer Topographic Laboratories / 8b OFFICE SYMBOL (If applicable)	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: DACA-76-85-C-0008
8c ADDRESS (City, State, and ZIP Code): Fort Belvoir, Virginia 22060-5546	10 SOURCE OF FUNDING NUMBERS: PROGRAM ELEMENT NO. / PROJECT NO. ETL LA/J6/84 / TASK NO. / WORK UNIT ACCESSION NO

11 TITLE (Include Security Classification)
Dynamic Image Interpretation for Autonomous Vehicle Navigation

12 PERSONAL AUTHOR(S)
Riseman, Edward M. and Hanson, Allen R.

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
Annual	FROM 2/26/85 TO 2/25/86	1986 September	19

16 SUPPLEMENTARY NOTATION

17 COSATI CODES			18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	scene interpretation sensor motion spatial reasoning
17	07		
17	08		

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

The University of Massachusetts Autonomous Land Vehicle Project has been concerned with a variety of problems associated with sensor motion analysis and dynamic image interpretation for autonomous navigation.

In particular our research has the following long-range research goals that relate to Task D in the autonomous vehicle navigation program:

1. Determine the motion parameters of a sensor relative to the static environment.
2. Distinguish moving objects from the static environment and determine their motion parameters.
3. Develop algorithms for tracking and predicting the motion and environmental location of the sensor and moving objects.

(over)

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT	21 ABSTRACT SECURITY CLASSIFICATION	
☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS	Unclassified	
22a NAME OF RESPONSIBLE INDIVIDUAL: Rosalene M. Holecheck	22b TELEPHONE (Include Area Code): (202) 355-2767	22c OFFICE SYMBOL: ETL-RI

DD FORM 1473, 84 MAR — 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

UNCLASSIFIED

4. Build a reliable depth map of the environment from combined motion, stereo, and laser range data.

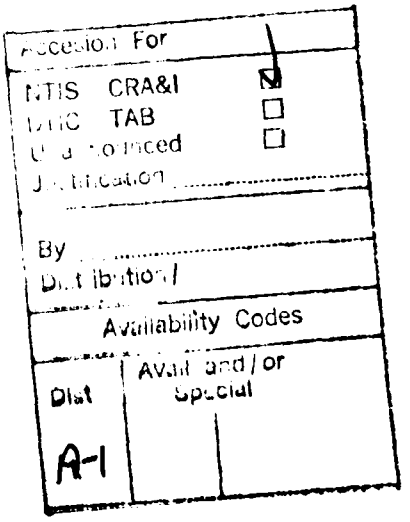
5. Identify major objects (both static and moving) in the environment while the sensor is either stationary or in motion.

6. Interpretation of the environment (i.e., object identification in road scenes) to provide constraints for identifying and tracking moving objects.

7. Provide information to update an environmental model of the moving sensor, including location of the sensor, other moving objects and distinguished stationary objects.

8. Provide control information to an expert navigational and spatial-reasoning system for the purposes of path planning and obstacle avoidance.

9. Integrate all of the above capabilities into a flexible and extensible system for dynamic scene interpretation.



PREFACE

This document was prepared under contract DACA76-85-C-0008 for the U.S. Army Engineer Topographic Laboratories, Fort Belvoir, Virginia, by the University of Massachusetts Computer and Information Science Department, Amherst, Massachusetts. The Contracting Officer's Representative was Ms. Rosalene Holecheck.

Contents

1 Introduction

The University of Massachusetts Autonomous Land Vehicle Project has been concerned with a variety of problems associated with sensor motion analysis and dynamic image interpretation for autonomous navigation.

In particular our research has the following long-range research goals that relate to Task D in the autonomous vehicle navigation program:

1. Determine the motion parameters of a sensor relative to the static environment.
2. Distinguish moving objects from the static environment and determine their motion parameters.
3. Develop algorithms for tracking and predicting the motion and environmental location of the sensor and moving objects.
4. Build a reliable depth map of the environment from combined motion, stereo, and laser range data.
5. Identify major objects (both static and moving) in the environment while the sensor is either stationary or in motion.
6. Interpretation of the environment (i.e., object identification in road scenes) to provide constraints for identifying and tracking moving objects.
7. Provide information to update an environmental model of the moving sensor, including location of the sensor, other moving objects and distinguished stationary objects.
8. Provide control information to an expert navigational and spatial-reasoning system for the purposes of path planning and obstacle avoidance.
9. Integrate all of the above capabilities into a flexible and extensible system for dynamic scene interpretation.

2 Motion Analysis

Our initial research in motion analysis has focussed on two key subproblems and applications that appear to be highly relevant to the ALV Program. The first problem is the ability to recover sensor motion parameters via purely passive sensors. This would be important in the case where a vehicle did not have an active range sensor (due to cost or mission constraints), or an active sensor was

available but malfunctioned. In either case, it is important to have the system be able to use visual data for obstacle avoidance, navigational updating relative to a world map, and moving object detection.

We have restricted our efforts in this area to translational motion (i.e. no rotation of the sensor) both because we believe it is the most important form of restricted motion, and also because we believe we have an algorithm that is computationally tractable and robust. Thus, our first year of the ALV research effort involved the recovery of sensor motion parameters from a pair of images undergoing translational motion.

The second motion problem we have focussed on is the recovery of depth data from a sequence of images produced under known sensor motion. This problem is very important because the vehicle will usually have accurate estimates of its own motion. Consequently, the system can infer the environmental depth of any image point that can be tracked across frames. We have developed an algorithm that will allow depth to be recovered using approximately constant computation over time, while refining coarse depth estimates from the initial pair of frames to more accurate estimates as additional frames continue to arrive.

2.1 Effectiveness In Recovering Translational Motion Parameters

We are continuing the analysis of algorithms for constrained sensor motion [5]. In particular we are evaluating the robustness, accuracy, and efficiency of the algorithm for recovering translational motion parameters and have developed an interim report on this topic [8]. Here the global search for the focus-of-expansion (FOE) requires the computation of the sum of errors (e.g., via correlation) associated with the displacement of a set of feature points in two or more frames. A sparse sampling of the possible location of the FOE provides a global error function whose minimum localizes the FOE, and thus the direction of motion.

The accuracy and robustness of the algorithm is a function of the number of points that are matched for contributions to the error function, which of course must be traded off against the

amount of computation that can be tolerated for real-time motion analysis. Thus far, our experiments on simulated environments imply that there is a wide range of situations for which the motion parameters can be approximately recovered at relatively modest computational expense. Specifically, when the angle between the image plane and the direction of translational motion is less than 45 degrees, then between 4 and 16 points which are widely spaced in the image are sufficient to recover the approximate motion of the sensor. A smaller number of points (4–8 points) is necessary when the camera is oriented approximately in the direction of motion (0-15 degrees) and a larger number of points (8–16) when the camera orientation is at a modest angle (15-45 degrees) with respect to translation. When the angle between camera orientation and translation is large (60-90 degrees) there appears to be a flat error surface around the correct direction of motion, leaving a wide range of ambiguity no matter how many feature points are employed. This result is not surprising in that it states, for example, that when a camera is pointing out the driver's side window, accurate determination of the motion of a vehicle moving down the road is not possible.

We are still investigating more carefully the limits of the accuracy with which sensor motion can be recovered. This will involve smoothing the error surface around the minimum.

Similar analyses for other cases of constrained sensor motion, including pure rotation, and planar motion in a known plane, remain for future work. We believe that they will exhibit similar levels of robustness and computational requirements.

2.2 Refinement and Prediction of Image Dynamics and Environmental Depth Maps Over Multiple Frames

To a large extent research in the interpretation of motion has focussed on the recovery of the motion parameters of a sensor moving through a static environment, and more generally the relative motion between a sensor and a visible object. Under ideal conditions, once these motion parameters are known, a depth map can be recovered from two frames if the displacement (flow) field is exact.

Displacement fields are not perfect, due to changes in lighting conditions, homogeneous image

areas, occlusion, etc. Even with perfect information about sensor motion, displacement vectors from translational motion are a function of the depth of the surface element. Any ambiguity or error in displacements along linear paths emanating radially from the FOE leads to ambiguity in the depth of that surface element. There are several sources of such ambiguity including multiple minima in the matching process for computing displacements, noise affecting the match location, and finally the resolution in the matching process along that radial path. Consequently, we are viewing the matching process as a dynamic refinement of depth over multiple frames.

The work that we have carried out here is a first step in the exploration of several issues involved in the stability, refinement, and prediction of depth maps over multiple frames [2]. We are considering the differences in start-up (when no depth information exists) versus updating an existing (and possibly inaccurate) depth map; in both situations we assume limited computational resources are available, yet increasing accuracy over time is required.

When an image sequence is first acquired, or the visible field changes dramatically (as in the case of coming around a corner), no depth map exists and the situation can be considered as a start-up. Under an assumption of a fixed limit on the computation that can be carried out between any pair of frames, a strategy has been developed to extract a coarse depth approximation from the first pair of frames using a coarse spatial resolution for the matching process. Each subsequent frame that is processed can use the previous estimate of depth to narrow the match area while increasing the match resolution, thereby maintaining constant computation, but finer accuracy in the depth estimates. As this process continues, temporal resolution can also be reduced as necessary. Thus, the approach employed involves a combined hierarchical spatial and temporal resolution as frames continue to arrive.

The refinement strategy that we have just described for the start-up phase of depth map recovery can be generalized for updating, prediction, and error analysis. Under known sensor motion and known environmental depth, the image location and appearance of environmental features can be

accurately predicted and matched from one frame to the next (leaving aside complex issues of image changes due to changes in lighting, highlights, shadows, shape distortion of surface patches, or occlusion). Thus, when one reaches the desired level (or limit) of spatial and temporal resolution, the updating process becomes one of prediction and verification of the environmental model. When predictions are not accurate, then depending upon the representation, the depth of either pixels, points, lines, regions, or surfaces could be refined in a focus-of-attention and refinement process for error reduction. Areas of the image and environment that do not behave as predicted become the focus of processing until their image dynamics over time can be properly predicted. In this manner one has an ongoing mechanism for verification of the current interpretation of the environment.

2.3 Scenarios for ALV Moving Object Data Collection

We have played a lead role in the development of scenarios for data collection associated with moving objects in the ALV scenarios. This information was provided to Carnegie-Mellon University and Martin Marietta to begin to define the needs of the ALV research consortium in the area of motion analysis. There were 5 scenarios that were specified in order of our increasing interest:

1. Moving down a 2 lane road in right lane with vehicle coming towards our vehicle in left lane (i.e. not on collision course).
2. Moving forward with other vehicle moving at 45 degree angle towards our vehicle on a collision course.
3. (Variation of 2) Moving down a 2 lane road towards a perpendicular intersection with the other vehicle moving perpendicularly on a collision course.
4. Moving on a 2 lane road following another moving vehicle for half the sequence on a straight portion of road, and leading into a curved portion for the other half of the sequence.
5. Same as scenario 1 with 2 vehicles moving towards each other around a long curve.

3 Dynamic Interpretation of Images

As part of the ALV project we are further developing our object interpretation techniques for application to the ALV scenario. Work on the VISIONS system for interpretation of images continues

[3,7,9]. A rule-based system for generating initial object hypotheses from image data has been extended to permit information from multiple sources of low level data to be "fused" in a consistent manner. On the basis of the results in a forthcoming thesis by Weymouth [11], we have refined the notion of schemas as a representation of knowledge. We are implementing a new schema system in CommonLisp and translating existing schemas and their associated interpretation strategies into the new format.

3.1 Rule-Based Hypotheses From Complex Aggregations Of Image Events

Recently [10,9] we described a simple type of knowledge source for generating object hypotheses for particular regions in the image. Simple rules are defined in terms of ranges over a scalar feature, and complex rules are defined as combinations of the output of a set of simple rules. The scores of these rules serve as a focus of attention mechanism for other, more complex knowledge-based processes. The rules can also be viewed as sets of partially redundant features each of which defines an area of feature space which represents a "vote" for an object on the basis of this single feature value. The region attributes include color, texture, shape, size, image location, and relative location to other objects. More recently, the approach has been extended to lines, with features including length, orientation, contrast, width, etc. In many cases, it is possible to define rules which provide evidence for and against the semantically relevant concepts representing the domain knowledge. While no single rule is totally reliable, the combined evidence from many such rules should imply the correct interpretation.

Most of the rules previously described are unary, accepting a region as input and returning a confidence for the object label. In addition, simple binary rules, defined over pairs of regions, were used to determine the similarity of the regions and to form aggregations of regions with similar properties. Typically, the rules operate on primitives formed by a single segmentation process (e.g. regions or lines) and result in the merging of the primitives into a more complete description, depending on the confidence returned by the rules. Forming more abstract groups of elements in

this way has advantages when dealing with unreliable segmentation processes: fragmented elements can be grouped to form aggregates which perhaps more closely match object models.

Recently, we have extended this approach to include relational rules, which capture expected relations between the elements of multiple representation (e.g. regions, lines, surfaces) of the image data [1]. Using rules of this form, sets of elements across the multiple representations can be selected and grouped on the basis of relational scalar measures associated with each rule. The result, assuming the confidence value returned by the rule is high enough, is the construction of complex aggregations of elements which satisfy user-specified relations across the multiple representations. One advantage of this approach is that it is modular and extensible; when new representations are added to the system, integration is accomplished by adding the appropriate rules.

In our preliminary work, we are concerned with relational rules defined over regions and lines. Since both are defined in a pixel-based representation, a convenient basis for the rules is intersection of the corresponding sets of pixels. Such relational rules, called intersection rules, are composed of three components:

1. a *relational filtering rule* for selecting lines which intersect a region based on relational measures;
2. a *ranking rule* which ranks the lines which intersect a region based on line attributes; and
3. a *combination function* which calculates the final score of the region-line aggregation based on the scores from the *filtering rule* and the *ranking rule*.

The relational measures are used to measure the type and degree of the relationship between a region and a line. Lines associated with regions are categorized into three types: boundary lines, interior lines, and lines which are neither interior nor boundary. The measures are:

1. interior-line-percentage: the ratio of line area interior to the region to total line area.
2. region-perimeter-percentage: the ratio of region boundary pixels covered by the line area to the region perimeter.
3. line-length-percentage: the ratio of the length of the region boundary covered by the line area to the total length of the line.

The relational filtering rule is then a complex line rule composed of a simple rule for each relational measure; in many cases it simply removes certain combinations of regions and lines from further consideration. The ranking rule ranks each line on the basis of how well it satisfies the associated relational measure. The combination rule is supplied the scores from the relational filtering rule, the line ranking rule, and the relational measures and converts these into a confidence for the hypothesis supported by the rule.

These intersection rules can be used in some very diverse ways. One example is to use a filtering rule on interior-line-percentage to select only those lines which are interior to a region. The ranking rule could then be defined to select short, high-contrast lines. The score of the ranking rule could then be averaged to form a complex texture measure. Alternatively, a density measure could be calculated by counting the occurrences of lines which receive a high score from the ranking rule and then normalizing by the size of the region.

As an additional example, the line-length-percentage measure could be used to select lines which lie mostly on the boundary of the region. The ranking rule could then be defined to favor long lines. The scores from the ranking rule could then be averaged using region-perimeter-percentage as a weighting factor to form a simple shape measure.

A preliminary implementation of the extended rule system has been completed, several simple texture and shape rules have been written, and results have been obtained on urban house scenes and on road scenes. The results [1] are quite promising. For example, we have been able to find roads in several roadscenes by using a rule which implements a simple shape measure. In the future, we intend to write additional rules and apply the system to a larger variety of images, develop new rule types, add additional representations for motion, depth, and surface segmentations, and incorporate the rule-based system into the schema system currently being developed (see next section).

3.2 Schema Networks as a Representation of Knowledge

In the VISIONS system, scene independent knowledge is represented in a hierarchical schema structure organized as a semantic network [3,10,7,4,11]. The hierarchy is structured to capture the decomposition of visual knowledge into successively more primitive entities, eventually expressed in symbolic terms similar to those used to represent the intermediate level description of a specific image obtained from the region, line, and surface segmentations. Each schema defines a highly structured collection of elements in a scene or object; each object in the scene schema, or part in the object schema, can have an associated schema which will further describe it. Each schema node has both a declarative component appropriate to the level of detail, describing the relations between the parts of the schema, and a procedural component describing image recognition methods as a set of hypothesis and verification strategies called *interpretation strategies.*

The schema system provides a hierarchy of memory structures, from vertices (or even pixels) at the bottom level through semantic objects at the top. A further division of knowledge into long term (LTM) and short term memory (STM) across the levels of hierarchy provides a convenient way of differentiating the system's permanent *a priori* knowledge base from the knowledge that it has received or derived from a specific image. The goal of the system is an interpretation, by which is meant a collection of objects at the top level of STM that is consistent with both the image date and the system's *a priori* knowledge of the world as represented in LTM.

A central problem of high-level vision is how to make use of knowledge, not just to categorize the results of lower levels of computation but also to guide those levels through the space of image analysis and feature extraction techniques. Practical systems will need to know about an extremely large number of objects – a prohibitive number for any system that attempts to find each object in each image. Furthermore, there is a computationally explosive number of low and mid-level image operations (segmentation algorithms, texture measures, line finders, rectangle finders, line grouping operators, etc. which collectively are termed 'knowledge sources') which might be

applicable, especially when one realizes that for almost every object there might be a variation of certain operations that would be particularly well suited to recognizing *just that object*. As a result, the combinatorics of what low- and mid-level processes to apply and how to interpret their results is simply too great to expect any near-term increase in the power of computing systems to solve the problem by brute force computation. The high level vision system must control the work being done at the lower levels for computer vision to be computationally feasible in the near future. The goal of this research, then, is to provide a prototype knowledge-driven system called the Schema System, to interpret images and provide control.

The development of the schema system confronts many of the same issues that have come up in other interpretation and control domains, such as speech understanding [6,12]. Among them are questions of the knowledge representation, the communication of information, error recovery and the selection of knowledge sources.

A doctoral dissertation by Terry Weymouth [11] presents our most recent approach to these problems. This dissertation explores the information and control structures needed for knowledge-directed interpretation of natural outdoor scenes. A *schema network* represents object descriptions, relations among objects, and control knowledge. Each node of the network, a *schema*, contains both a declarative structure and references to one or more *interpretation strategies*. The declarative portion of the schema describes the composition of an object including the spatial relations of its parts and their possible appearances in an image. The interpretation strategies are object-specific procedures for creating hypotheses of the existence of the object. In the interpretation strategies, the procedural representation of control information provides a natural form for expressing the dynamic nature of the image interpretation.

A *schema instance* is created when a schema is activated either by a top-down request for a goal or by bottom-up detection of key events in the image. Schema instances continually interact with one another either through a channel set up when a goal is requested or through hypotheses

created in a blackboard data structure. Several schema instances can work simultaneously on relitively independent portions of the interpretation, thus exploiting the potential for parallelism. By selectively grouping line and region primitives into descriptions of parts of a scene, the cooperative activities of the schema instances construct the final interpretation network.

The system was tested on six images from four scenes. Parallel activation of schemas is simulated; overlaping of the timing in the actions of a set of interpretation strategies is illustrated in traces from the simulation. The resulting interpretations contain both the association between object structures and image events and three-dimensional descriptions of some of the objects in the scenes.

4 References

REFERENCES

[1] R. Belknap, E. Riseman, and A. Hanson, The Information Fusion Problem and Rule-Based Hypotheses Applied To Complex Aggregations of Image Events, Proceedings of the DARPA IU Workshop, Miami Beach, FL December 1985.

[2] R. Bharwani, A. Hanson, E. Riseman, Refinement of Environmental Depth Maps over Multiple Frames, Proceedings DARPA IU Worksop, Miami Beach, FL. December 1985.

[3] A. Hanson and E. Riseman, *VISIONS: A Computer System for Interpreting Scenes*, Computer Vision Systems (A. Hanson and E. Riseman, eds.) (1978), 303 - 333, Academic Press.

[4] A. Hanson and E. Riseman, *A Summary of Image Understanding Research at the University of Massachusetts*, COINS Technical Report 83-35 (October 1983), University of Massachusetts at Amherst.

[5] D.T. Lawton, Processing Dynamic Image Sequences From a Moving Sensor, Ph.D. Dissertation (TR 84-05), Computer and Information Science Department, University of Massachusetts, 1984.

[6] V.R. Lesser, R.D. Fennell, L.D. Erman, and D.R. Reddy, Organization of the Hearsay-II Speech Understanding System, IEEE Trans. on ASSP 23, pp. 11-23.

[7] C.C. Parma, A.R. Hanson and E.M. Riseman, *Experiments in Schema-Driven Interpretation of a Natural Scene*, COINS Technical Report 80-10 (April 1980), University of Massachusetts at Amherst.

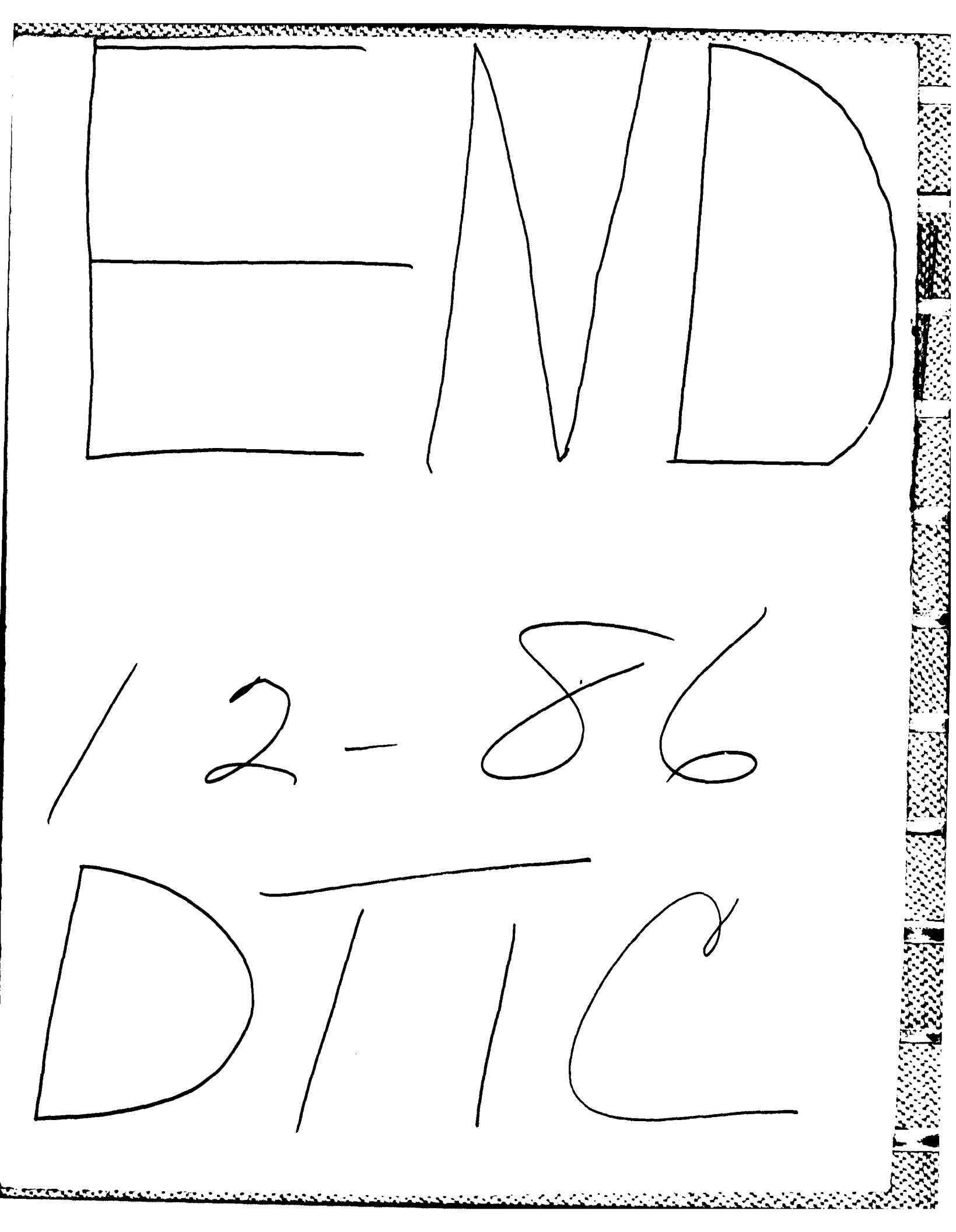
[8] I. Pavlin, A. Hanson, and E. Riseman, Analysis of an Algorithm for Detection of Translational Motion, Proceedings DARPA IU Workshop, Miami Beach, FL, December 1985.

[9] E. Riseman and A. Hanson, A Methodology for the Development of General Knowledge-Based Vision Systems, Proceedings of the IEEE Workshop on Principles of Knowledge-Based Systems, Denver, Colorado, December 1984.

[10] T.E. Weymouth, J.S. Griffith, A.R. Hanson and E.M. Riseman, *Rule Based Strategies for Image Interpretation*, Proceedings of AAAAI-83 (August 1983), 429-432, Washington D.C. A longer version of this paper appears in Proc. of the DARPA Image Understanding Workshop (June 1983), 193-202, Arlington, VA.

[11] T.E. Weymouth, Using Object Descriptions in a Schema Network for Machine Vision, Ph.D. Dissertation, Computer and Information Science Department, University of Massachusetts, Amherst. Also COINS Technical Report, (in progress), University of Massachusetts, Amherst.

[12] W.A. Woods, Theory Formation and Control in a Speech Understanding System with Extrapolation Towards Vision, in Computer Vision Systems (A. Hanson and E. Riseman, Eds.), Academic Press, 1978.


END
12-86
DTIC