Proc. DARPA Image Understanding Workshop
Washington, DC, April 18-21, 1993.

CREATING BENCHMARKING PROBLEMS IN MACHINE VISION: SCIENTIFIC CHALLENGE PROBLEMS

Oscar Firschein*, Martin A. Fischler, and Takeo Kanade*****
DARPA, **SRI International, ***Carnegie Mellon University

Abstract

We discuss the need for a new series of benchmarks in the vision field. to provide a direct quantitative measure of progress understandable to sponsors of research as well as a guide to practitioners in the field. A first set of benchmarks in two categories is proposed (1) static scenes containing manmade objects, and (2) static natural/outdoor scenes. The tests are "end-to-end" and involve determining how well a system can identify instances (an item or condition is present or absent) in selected regions of an image. The scoring would be set up so that the automatic setting of adjustable parameters is rewarded and manual tuning is penalized. To show how far machine vision has yet to go, a Benchmark 2000 problem is also suggested using children's "what is wrong" puzzles in which defective objects in a line drawing of a scene must be found.

1 Introduction

Speech and natural language researchers at DARPA have made extensive use of a benchmarking approach to obtain a measure of the progress in these fields. This exercise has had two distinct positive effects on the fields. First. since the results of the benchmarks provide a direct quantitative measure understandable by people outside these fields, they have been of great use "politically" within DARPA. Second, through the rigorous comparison among various techniques the benchmarking exercise has spurred advances in the field. This paper discusses the strategy for devising and using a new series of benchmarks in the vision field. We believe that the vision field requires such benchmarking efforts to objectively measure its progress. There have been some previous benchmarking attempts in vision at DARPA and elsewhere. but they dealt mainly with measuring the performance of computer architectures running vision algorithms, rather than with the performance of the vision algorithms and systems.

The goal of creating challenging benchmark problems in machine vision is to pose a reasonably comprehensive set of vision problems to which proposed advances can be subjected to experimental evaluation. The problems should be formulated in terms of tasks, for a module or for a whole system, independent of any specific techniques – e.g., evaluation of three-dimensional shape recovery rather than evaluation of a shape-from-X method, or evaluation of natural scene understanding rather than evaluation of an expert system-based interpretation system. There would be challenging problems in various categories. e.g., outdoor scenes, manmade objects, time-varying scenes, and so on. After discussing issues and methodologies in creating vision benchmark problems, this paper presents a few example problems in the domain of static natural scenes and in static scenes containing manmade-objects.

2 Previous Benchmarking in Vision

The DARPA benchmark carried out from 1986-89 [1,2] was an attempt to characterize the performance of machine architectures running IU algorithms. As such, it is not directly applicable to the current IU benchmark effort. However, there were several lessons learned, primarily the time-consuming and somewhat expensive nature of the operation.

A more pertinent benchmark approach is the Unmanned Ground Vehicle set of evaluations for all subsystems, including stereo, LADAR, road-following, and path planning. The stereo evaluation, described in these proceedings [3], is of particular interest in this regard. The overall plan was to pursue a three-pronged approach, including analytic models, qualitative "behavioral" models, and statistical performance models. The analytic models are used to estimate the expected depth precision computable with a specific camera configuration. The qualitative models are used to identify key problems for future research. The statistical model is used to produce quantitative estimates of such key factors as the smallest obstacle detectable at a specified distance. Data gathering and preparation required a large amount of effort. Imagery was collected from five groups: 49 image pairs were selected for analysis and converted to a standard format. An interesting initial result of the evaluation was the identification of the strengths and weaknesses of the various stereo techniques, leading to the possibility of combining them in a system that produces more complete and accurate results than any of the individual techniques.

3 Issues in Vision Benchmark Design

There are a few important issues that arise in vision benchmarking :

- specifying the scope of a problem.
- balancing competitive and collaborative aspects of benchmarking, and
- devising an evolutionary problem selection mechanism for future benchmarks.

3.1 Vision Problem Specification

The critical difference between producing vision benchmarks, and producing those for language and speech, is that the field of machine vision does not yet have widely acceptable specifications for generic problem domains (or representations) on which to base the problem definitions. Further, the number of sample images and supporting data necessary to cover any given problem domain without artificial (unknown) biases seems to be far larger than in language and speech. In language, topics and languages certainly vary a lot; yet a large enough number of news articles, novels, etc. will reasonably cover the problem variations, and text files contain everything the benchmarking algorithms need to work with. In speech. there are variations in frequency, dialects, etc. yet a "numeral digit recognition" problem or a limited vocabulary problem provides some reasonable bound to a problem domain, and a large enough number of speech samples will cover the variations. A high-quality tape of speech (with various types of background noise) is a good universal basic representation for the input data.

In image understanding, the direct analogies do not work as well. Outdoor natural scenes do not seem to have an accessible technical definition, except that people can probably classify a given image as depicting a natural scene or not. A universal representation/media for sensed data describing a scene does not exist. Moreover, the nature of the input devices, the way we acquire images and specify the resolution, the measurable information, etc. can themselves, singly or in combination, constitute major research problems.

3.2 Competitive and Collaborative Aspects

The principal goals of the proposed vision benchmarks are to evaluate scientific progress in specific problem areas, and to make the extent of such progress apparent to the sponsors of the research as well as to the scientists working in this field. Evaluation of a set of alternative solutions to a problem naturally involves comparing the resultant scores and to thus rank the techniques. We can't avoid competition. Making the results of the evaluation difficult to interpret or keeping the identity of the participants secret eliminates the incentive to enter the evaluation and exert the necessary effort to do well.

Nonetheless, it is very important to make sure that we are competing on the right problems, that the competition is fair, and that we don't poison the currently excellent cooperative atmosphere that exists in the DARPA vision research community.

Among its positive benefits, benchmarking will promote collaboration. Many researchers will not be able to afford to develop all the system components themselves in order to enter the evaluation. A module or component that has been proven to have high performance will be transferred from the hands of the developer to other sites whose main research focus is not the module, but rather access to its functionality.

In the specific case of algorithms that are intended to run autonomously, i.e., without manual tuning, it is critical for the purposes of believability that the test data NOT be given to the contestants prior to the benchmark. Further, the problem of automatically finding settings for adjustable parameters (present in almost every vision algorithm) is a key vision problem in which progress should be encouraged – the benchmark could be a positive influence in this regard.

3.3 Evolution of Benchmarks

Since there is enough diversity of opinions about what constitutes the correctness of the output of any componcnt, practical benchmarking tends to be performed on "end-to-end" systems performing a well-understood task. This emphasis on system evaluations can have both positive and ncgativc impacts on the field. Positive effects arc: the promotion of research because there exists accepted criteria of progress; the establishment of some priorities on problems to be solved; and an increased awareness of the availability and usefulness of a broader range of techniques for performance improvement Potential negative effects are: the temptation to use any trick that improves performance on the evaluated task; the premature stifling of new directions in the field, and the reliance on some statistical methods which tends to produce better "average" results. Some of these phenomena. positive and negative, have appeared in the language and speech fields since the introduction of benchmarks.

To achieve the positive effects and avoid the negative ones, the vision benchmarks should allow for evolution and expansion as we improve our understanding of the field. This is especially important in vision because vision tasks are not static – they expand. The benchmark problems cannot be a casually controlled ad-hoc collection of problems, or a set of problems carefully tailored for small cliques, each with a special view of how the problem should be solved. If a group of investigators wants to pursue a promising new approach which cannot be evaluated appropriately within the current set of benchmarks, there must be a mechanism to define a new DARPA benchmark if appropriate criteria arc satisfied.

4 Deriving Benchmark Problems

4.1 Problem Selection

We must first define the problem domains, such as static outdoor scenes, static scenes of manmade objects. outdoor image scene sequences, and image sequences of manmade objects. A panel will be organized to carefully divide the vision field into categories and subcategories because this categorization is one of the most critical issues in the design of the benchmark. A relatively small number of problems (less than four initially) in each category would be carefully selected by community consensus. These problems should be based on some important vision function, NOT some vision architecture, representation, or technique. These should generally be retained in the benchmark until they are "solved" or no longer of scientific importance.

The complete benchmark should cover the vision field by defining between five to ten separate problems for the competition; it may be necessary to have two problems in some categories to separately deal with the main dichotomy of strong vs weak models (eg., man-made environments vs natural outdoor scenes). Each problem category may require separate subcategories for different sensing modalities, viewing conditions, and environmental factors; in the case of sensing modalities. the subcategories could be

- intensity images vs range images
- black-and-white vs color (or multispectral)
- significant perspective distortion vs. essentially orthographic projection

The problems listed in Appendix A are a few abstracted versions of possible benchmark entries for the static outdoor scenes and man-made object scenes. They are offered for discussion in the light of all the above sentiments, but the task of choosing the actual problems still remains. It is hoped that for an initial benchmark a total of no more than four problems will be selected from the set of all submissions. This would allow us to work out the details of the process before an excessive amount of effort is expended.

Finally, it is important to remember that the proposed benchmark will only cover a small subset of the the important problems in machine vision. We have not done away with all the traditional methods of reporting and evaluating progress.

4.2 Evaluation Method

Human examiners would select (but not necessarily reveal to the contestants) a few locations in each image that contain obvious instances (item or condition is present or absent) of, for example, the existence of a road or a material like grass or rock. The scoring at each location is binary – correct or incorrect.

Problems, for example, in recognizing natural objects are believed to be difficult enough so that no currently known technique, or brute force approach, can perform well (i.e., within 50% of human performance on the recognition problems and somewhat higher on the geometry problems depending on the availability of calibration data and the nature of the prior models) without additional constraints on the problem (or the provision of auxiliary information, such as manual parameter adjustment to match the given imagery). An obvious advance would be a performance improvement of, say 5 to 10% over that of the previous best known technique. When performance of a computer vision technique reaches (say) 90-95% of human performance, the corresponding problem is considered to have a reasonable scientific solution and further advance is now also considered in terms of engineering criteria (cost, speed. complexity. etc.).

4.3 Competition Procedure

It is intended that there would be a competition once a year to choose the best performing program in some (or all) problem categories. The programs would have to run on specified machine configurations, must take the input images in a specified format, and must produce answers in a specified time interval. To insure an initial reasonable baseline of performance for the most difficult problems, an operator would be allowed to place a specified number of labeled markers in an overlay of the given test image and/or be given (in advance) a small window from the test image to allow system calibration and parameter adjustment Typical images from each category would be provided in advance, and would not not change in nature or difficulty from year to year.

Entry in the competition implies the entrant is willing to make public the theory (and possibly pseudo-source code) for his algorithms and allow the use of his object code (at least) for scientific purposes.

4.4 specific Proposal

1. A list of 5-10 problem domains will be selected for the benchmark.
2. A panel of experts would be chosen to define the problems and select the sample and test imagery and the contextual information to be made available. Sample imagery would be available prior to the competition. The actual test imagery would be provided to all interested parties after the competition.
3. The nominal approach would be for the panel to select a few locations in each image that contain obvious instances (item or condition absent or present) of the challenge problem subject matter, this information would not be revealed (for the test imagery) until after the competition; scoring at each location is binary, correct or incorrect.
4. The test could be held yearly (e.g., at the IU meeting or at some selected contractor site) on machines provided or

approved by DARPA. Programs must produce answers in a specified time interval. It is intended that the programs will be run without intervention by the contestants. but some provision might be made to allow a contestant to tune his program at a specified penalty to his test score.

5. Theory (and possibly pseudo source *code)* must be provided in report form, and the compiled code actually used in the competition made available (free, but possibly under license) for scientific use.

5 Conclusion

We have taken the initial steps in developing a new set of machine vision benchmarks in the areas of manmade object scenes and natural scenes. The next steps involve more careful delineation of the experimental protocol, selection of the specific problems, and the gathering of imagery and other test data. We welcome comments on this new DARPA benchmarking effort.

References

1. A. Rosenfeld, "A Report on the DARPA IU Architectures Workshop," Image Understanding Workshop, 1987.
2. C. Weems, E. Riseman, and A.R. Hanson, "A Report on the Results of the DARPA Integrated Image Understanding Benchmark Exercise." Image Understanding Workshop, 1989.
3. R.C. Bolles, H.H. Baker, and M.J. Hannah, "The "JISCT" Stereo Evaluation," *in these proceedings.*

APPENDIX

This appendix offers a set of abstracted versions of possible benchmark entries for discussion in the light of the goals and issues. The task of specifying the actual problems still remains.

A.1 Man-Made Object Scenes

The key issues in setting up the problems for man-made object scenes include:

- amount of clutter in a scene
- amount of occlusion of objects
- class of shapes of objects (e.g., polyhedral vs. curved, planar vs. 3d. fixed vs. articulated or deformable)
- class of surfaces (e.g., textured. specular, diffuse)
- lighting conditions
- kind of imagery (2d vs. 3d, grey-scale vs. color)
- class of transformations applied to object model

Even more important is an evaluation method. We can use the ROC (Receiver/Operator Curve) which plots the false negative rate vs false positive rate as the overall indication of the performance of a system. We should evaluate the accuracy of computed pose as well as the number of free parameters in the system. We should also define a series of increasingly harder problems, such as presented below.

PROBLEM 1: Flat parts (with little or no texture on the parts or background) and known camera orientation. But include significant clutter (e.g. as little as 10% of the features in the image are associated with the object) and significant occlusion (perhaps as little as 25% of the object is visible) as well as noise. The goal is to identify and locate as many instances as possible from a small library of known models. This problem is probably fairly well solved by several existing algorithms. Open issues include how to provide/obtain the models and a range of shapes for each of the models. We can allow the objects to scale, rotate, and translate in the image plane.

Example objects include 2D parts (eg. teletype parts) and tools (wrenches, screw drivers, etc).

PROBLEM 2: Solid rigid objects with no articulation, but with significant clutter and occlusion. Texture is allowed on the objects and background. One version of the problem would be 3D shape recovery from a single 2D image; a second version could be 3D shape recovery from a 3D image (i.e. from range imagery).

The objects should include a range of shapes and even several shapes that differ only in a few places. so that saliency *can* be an issue. Examples include models of vehicles or planes, simple office scenes, etc.

PROBLEM 3: Generic objects. This could include parameterized object classes and articulated objects. The idea is to allow for objects that are nonrigid, while still structured. A tank is an obvious example, given the movable turret. Classes of vehicles in general provide the next level of complexity (e.g. categorizing vehicles as a sedan, a station wagon, a van, etc., as well as localizing it).

PROBLEM 4: Recognition by function. A classic example is a "chair", where the recognition system has to identify objects not only by shape, but also by whether they meet certain functional constraints (such as stable support, a flat surface on which to sit, etc.)

A.2 Static Natural Scenes

1. Generic (natural) Object Recognition or Classification: Recognize (point to or delineate) rocks and trees in single images of outdoor scenes. Distinguish between rocks and sand in a desert scene.
2. Specific Object Recognition: Recognize (point to or delineate) the presence of a specific known object in an

image (e.g. a particular person). The object (preferably non-rigid) can be partially occluded or seen from any aspect.

3. Feature Extraction and Delineation:

   -Extract a road network from an aerial image that can be either vertical or oblique and low or high resolution. For example, the image could even be a view of a partially occluded freeway from a window in a nearby building.

   -Given the skeleton of a tree trunk or tree limb (in a forest scene), accurately delineate the corresponding edges and measure the width of the trunk/limb.

4. Geometric Recovery From a Single Image of a Natural Scene:

   -Given a line overlaying an image, determine relative depth (from the camera) along the line.

   - Determine (scene) surface orientation at a given set of points (locations) in an image. The scale of interest will be provided.

5. Geometric Recovery From Multiple Views Of Some Specified Object:

   -Model (i.e., recover the 3D geometry) a building in an urban scene from multiple views.

   -Render the profile of the skyline seen from a specific ground location, given an overhead stereo pair of a mountain or valley.

   -Given a dense sequence of images containing an object of interest (e.g., a tank or a rock) taken from a camera mounted on a moving truck, recover the geometry of the object

6. Track a Specific Moving Object: For example, a specific fish in a fish-bowl containing many fish and other objects that can cause occlusion or temporary disappearance of the fish being tracked. Other possibilities are a person in a crowded store, or a specific tank in a formation moving through a wooded area.

7. Generic Problems in (edge and surface classification) in Image Analysis:

   -Classify specified locations in an image as either edge or not-edge; if edge, then further classify the nature of the edge as either occlusion, orientation, illumination (e.g. shadow), or reflectance edge.

   -Classify the material type of selected surface patches in an image as either wood, metal, glass, water, vegetation, sand/rock, brick, soil, sky, cloud, asphalt, or concrete.

   -Classify, into say three spectral bands, the color spectrum of selected surface patches in a natural outdoor image.

Figure 1: Childrens" Puzzle Picture ["What's Wrong Here?" Kidsbook, Inc. 19901

A 3 Benchmark 2000

To provide a challenge problem for the year 2000, we present the following benchmark problem: Given a line sketch taken from a children's book of "what is wrong" puzzles, see Fig. 1. devise an automatic vision program that can perform as well as a five year old child. The absolute score for any algorithm is the percent of defective objects found in a scene; the relative score can be obtained by comparison with a child's performance.