To appear in *Empirical Evaluation Techniques in Computer Vision*, K.W. Bowyer and P.J. Phillips, eds, IEEE Computer Society Press, Los Alamitos, CA, 1998.

Overview of Work in Empirical Evaluation of Computer Vision Algorithms

Kevin W. Bowyer[1] and P. Jonathon Phillips[2]

[1]Department of Computer Science and Engineering
University of South Florida
Tampa, Florida 33620-5359
kwb@csee.usf.edu

[2]National Institute of Standards and Technology
Gaithersburg, MD 20899, USA
jonathon@nist.gov

1: Introduction

Computer vision emerged as subfields in computer science and electrical engineering in the 1960s. (The landmark paper of Roberts on object recognition was published in 1965 [29].) Two main motivations for research in computer vision are to develop algorithms to solve vision problems, and to understand and model the human visual system. It turns out that finding satisfactory answers to either motivation is significantly harder than common wisdom initially assumed.

Research in computer vision has continued actively to the current time. Most of the research in the computer vision and pattern recognition community is focused on developing solutions to vision problems. This chapter will address issues from that point of view. There are many researchers who are interested in the human visual system, with collaborations among researchers in computer vision, neuroscience and psycho-physics. Issues of empirical evaluation are important in this area too, but we not directly address this type of research in this chapter.

With three decades of research behind current efforts, and the availability of powerful and inexpensive computers, there is a common belief that computer vision is poised to deliver reliable solutions. Unfortunately, for most applications there are no methods available to test whether computer vision algorithms can live up to their claims. Nor is there any way to measure performance among algorithms, or to reliably determine the state-of-the-art of solutions to a particular problem.

In the absence of accepted methods of empirical evaluation of algorithm performance, advances in computer vision algorithms must naturally come to be judged by other criteria. Such criteria might include considerations such as conceptual elegance, the sophistication of the mathematical methods used and the computational complexity of the algorithm. Unfortunately, conceptual elegance and sophistication of the mathematics are not necessarily correlated in a positive way with performance of an algorithm in application. If the use of more sophisticated mathematics requires more specific assumptions about the application, and these assumptions are not satisfied by the application, performance could even degrade. Conceptually, the situation might be as depicted in Figure 1.

In the computer vision literature, various articles and discussions have argued for methods that would allow comparative assessments of algorithms [10, 12, 16, 20, 25]. The benefits of such methods would include (1) placing computer vision on a solid experimental and scientific ground, (2) assisting in developing engineering solutions to practical problems, (3) allowing accurate assessment of the state of the art, and (4) providing convincing evidence to potential users that computer vision research has indeed found a practical solution to their problems.

Despite these well-founded arguments, the computer vision community for the most part has not yet heeded the call. This has started to change in the last few years. There have been a number of workshops, conferences, and special issues of journals on the topic of empirical evaluation, including papers in this volume from the IEEE Workshop on Empirical Evaluation of Computer Vision Algorithms.

2 Approaches to empirical evaluation

We divide evaluation work into three basic categories. As is the risk with any classification, the categories will not necessarily be clean divisions. Evaluation work could fit into more than one category, or not neatly fit into any category. Despite this risk, we believe that the categories provide insights useful for the developing field of empirical evaluation of computer vision algorithms.

The first category is evaluations that are independently administered. In the prototypical independent evaluation, one group collects a set of images, designs the evaluation protocol, provides images to the testees, and evaluates the test results. This method allows for a high degree of standardization in the evaluation, since all algorithms are tested on the same images and scored by the same method. Thus, independent evaluations usually allow for a direct comparison between competing approaches to a problem. The competing approaches are usually state-of-the-art algorithms, and the individual competitors are often the original developers of the algorithms. Independent evaluation by a non-competitor gives a greater sense of impartiality and objectivity to the results. The major drawback to this form of evaluation is the level of ongoing effort required by the group administering the evaluation. Ideally, the evaluation mechanism needs to evolve and be refined over time.

The second category is evaluations of a set of classification algorithms by one group. The group wanting to do the evaluation will often not be able to get access to original implementations of all of the algorithms of interest, and so will have to implement some of the algorithms based on information in the literature. This introduces the possibility that the version of the algorithm evaluated will not be identical to that used by the original developers of the algorithm. However, implementation and evaluation of a set algorithms by one group can at least establish performance for baseline algorithms. When a new algorithm is first developed, it may not be sufficiently refined to be able to compete against the start-of-the-art algorithms. Comparing against a baseline allows for an initial assessment. Comparing between state of the art and baseline algorithms provides a measure of how much the additional performance costs. The cost could be a function of computational cost or complexity, reliability of the algorithm, or development effort.

An important feature of either of the first two categories is that ground truth is in theory not fuzzy and can be determined accurately. Classification problems often exhibit

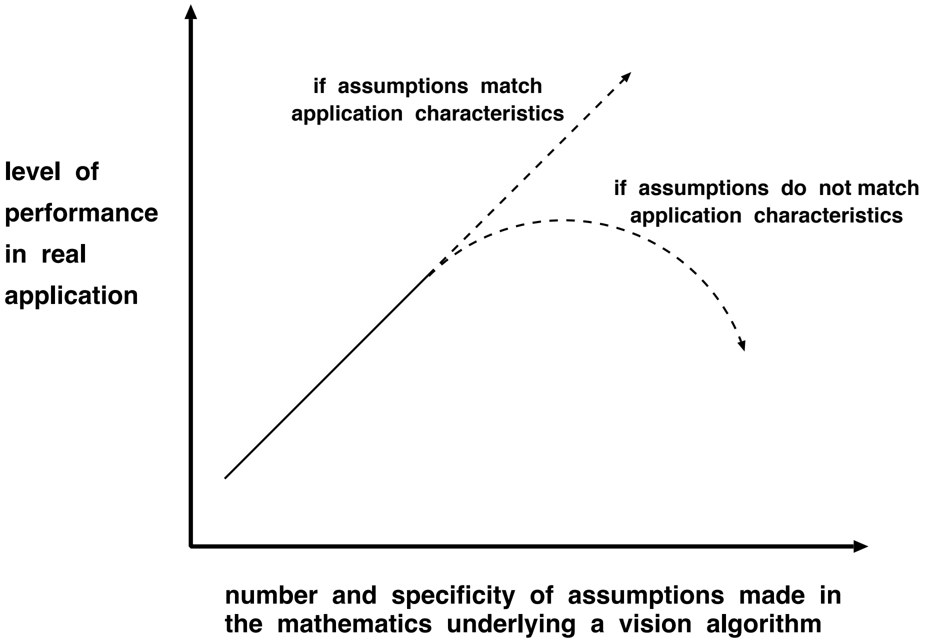
level of performance in real application
if assumptions match application characteristics
if assumptions do not match application characteristics
number and specificity of assumptions made in the mathematics underlying a vision algorithm

Figure 1. Performance as a function of mathematical sophistication.

this property. For example, the identity of a person in a face image is not fuzzy, and the particular character that is written in a certain location is known. As long as provision for recording ground truth is made at data collection time, it should be possible to get reliable and accurate ground truth. However, in practice things are sometimes not so simple. Further discussion of ground truthing techniques is beyond the scope of this overview, but it is a problem touched on in several of the chapters.

The third category is problems where the ground truth is not self evident and a major component of the evaluation process to develop a method of obtaining the ground truth. The classic example here is developing methods of evaluating edge detectors. The question of what "should be" marked as an edge in real images is often problematic. Heath et al [13] managed to avoid a direct answer to the question and still conduct a careful comparative evaluation of five edge detectors. They used human observers to rate the quality of an edge image in the context of object recognition, and then performed a statistical analysis of the resulting quality ratings. Those without experience in conducting human observer experiments may find it difficult to conduct this method of evaluation. Also, it is conceivable that human perception of edge quality may use properties of edges which are different than those needed for machine vision tasks.

3: Tour of Papers in This Volume

The workshop had two invited speakers, Henry Baird and Michael Fitzpatrick. Baird spoke on the impact of standard databases, benchmarks, and generative models on document image analysis research.

The work described in "A blinded evaluation and comparison of image registration methods" by Fitzpatrick and West is an example of an independently administered evaluation procedure. They assessed the performance of multimodal registration algorithms for magnetic resonance imagery (MRI) and position emission tomography (PET), and PET and computer tomograghy (CT). They acquired the images, generated the ground truth, designed the testing protocol, provided the data to the groups participating in the test, and scored the results. (Additional references to their work are West et al. [37, 38, 39].)

The work presented in "A Benchmark for Graphics Recognition Systems" by Chhabra and Phillips is an independent testing procedure for algorithms that automatically analyze engineering drawings. They describe how the synthetic images and the corresponding ground truth are generated, the testing protocol, and the results. Evaluation of document image analysis algorithms is an important topic which has received substantial attention in recent years.

The paper "Performance evaluation of clustering algorithms for scalable image retrieval" by Abdel-Mottaleb, Krishnamachari, and Mankovich provides baseline performance for color-based retrieval of images. They compare performance of two clustering algorithms and three image-to-image metric-based algorithms on a database of 2,000 images.

In face recognition, principal component analysis (PCA) techniques form the basis for a number of face recognition algorithms, and provide a baseline against which other face recognition algorithms are compared. Unfortunately, there are numerous implementations of PCA algorithms. "Analysis of PCA-based face recognition algorithms" by Moon and Phillips presents an empirical evaluation of different implementations of PCA-based algorithms on the standard FERET database of facial images. Also, they look at determining what are statistically significant differences in performance between two implementations of a PCA algorithm. This is done by generating multiple (resampling) test sets from a larger database of facial images.

"Performance assessment by resampling: rigid motion estimators" by Matei, Meer, and Tyler examines resampling techniques for evaluating algorithms. They develop a set of computational tools for applying the bootstrap, jackknife, and empirical influence function to computer vision problems. Their approach is demonstrated on rigid motion estimators.

Algorithm performance can be affected by the quality of the image that goes into an algorithm. Image quality is obviously affected by noise in the image acquisition system. "Sensor errors and the uncertainties in stereo reconstruction" by Kamberova investigates the effects of sensor noise in three-dimensional reconstruction from stereo images. Based on the noise model, the author developed a radiometric correction procedure to reduce the effects of the noise.

"Fingerprint image enhancement: algorithm and performance evaluation" by Hong, Wan, and Jain presents an image enhancement algorithm for fingerprints. The effects of the enhancement are assessed by measuring the performance of a fingerprint matcher on the original and enhanced fingerprints.

In some applications it is not feasible to collect enough real data to design an algorithm. In which case, it is necessary to generate synthetic data and design an algorithm from the synthetic data. "Empirical evaluation of laser radar recognition algorithm using synthetic and real data" by Der and Qinfen discusses the process of designing and evaluating a laser radar autonomous target recognizer. The algorithm is designed from synthetic data generated by a model of a laser radar sensor. They describe the method for validating the model, developing the algorithm from the synthetic data, and validating the algorithm on real data.

The availability of a standard database for a problem makes it possible for researchers to report results on a common database of images and provides a source of images to researchers who do not each individually have the time and money to collect them. "A WWW-accessible database for 3D vision research" by Flynn and Campbell describes a database of three-dimensional models, range images, and code to support research and evaluation of model based recognition and range image analysis algorithms. The database can be browsed and retrieved on the world wide web.

Because numerous algorithms being developed are visual functions that humans can perform, one can gain insight into the visual problem by understanding how humans accomplish the task. "Shape of motion and the perception of human gaits" by Boyd and Little is a case study of the interaction of algorithm development and psycho-physics for the design of an algorithm to understand the human gait.

"Empirical evaluation of automatically extracted road axes" by Wiedemann, Heipke and Mayer looks at the problem of evaluating algorithms that attempt to extract roads in aerial images. Road extraction is a classic problem in aerial image interpretation. Wiedemann et al have manually specified the ground truth for three aerial images, and have an automated method to evaluate an extracted road map in comparison to the manually specified map. Their evaluation includes several quality metrics for the results, such as completeness of the extracted road map, location error, number of gaps per kilometer and mean gap length.

Three chapters in this volume address methods of evaluating edges. "Analytical and empirical performance evaluation of subpixel line and edge detection" by Steger presents a method of analyzing the localization quality of subpixel line and edge detectors. His work includes analytical modeling, simulation experiments and experiments with real data. The nature of his task requires relatively simple images and careful consideration of the details of the experimental setup and image acquisition. This work is most relevant to industrial inspection tasks, as opposed to more general object recognition tasks.

"Objective evaluation of edge detectors using a formally defined framework" by Dougherty and Bowyer presents a pixel-level method of evaluating edge detectors. Their method uses a set of real images for which they have marked ground truth in terms of true edges and no-edge regions. The results of an edge detector are matched to the ground truth to give a (true positive, false positive) count. A given edge detector is adaptively tuned to give its best result in terms of a receiver operating characteristic curve (ROC) for the given set of images. They use four groups of ten images, each representing a different application domain. They present a performance analysis of six edge detectors, including accuracy, ease of tuning parameters and execution time.

"Evaluation of edge detection algorithms using a structure from motion task" by Shin, Goldgof, and Bowyer presents a "higher level" or more "tasked based" evaluation method than that of Dougherty and Bowyer. They have created image sequences for which they have independently measured the 3-D structure and motion. They use the quality of results produced by the line-based structure-from-motion routine of Taylor and Kriegman [32] as the basis of their evaluation. In this context, a better edge detector is one which results in more accurate SFM results. Interestingly, Shin et al reach overall conclusions in their comparison of edge detectors which are similar to the conclusions reached by Dougherty and Bowyer.

4: Other Evaluation Work

This section provides pointers to other recent work in evaluation methods. Some of these works have received substantial attention, and some address areas not covered by the papers in this volume. This section is not meant to be an exhaustive survey. There are undoubtedly many relevant papers which we do not mention. However, we hope that this section will help to give a feel for the current state of the art in empirical evaluation of computer vision algorithms.

Barron et al [1] implemented a number of algorithms for calculating optic flow and studied the performance of these algorithms. However, their work used primarily synthetic images. Use of synthetic images obviously makes the construction of ground truth much simpler, but also introduces the question of how well results obtained on synthetic data extrapolate to real data.

Bolles et al [3] attempted to lead an empirical evaluation of algorithms for shape-from-stereo algorithms. This work was done in the style of distributing data for the evaluation to various groups for them to work on and then report results back. However, only three of five groups were able to complete enough of the evaluation to report results.

Zhang et al [44] report on an empirical evaluation of shape-from-shading algorithms. In this work, all the algorithms were implemented by the group doing the evaluation. This obviously reduces problems arising due to incompatibility of different implementations.

Hoover et al [15] conducted a detailed evaluation of four different range image segmentation algorithms. This comparison used forty images from each of two different types of range sensor. Groups from the University of South Florida, Washington State University, Edinburgh University, and the University of Bern collaborated to compare their respective algorithms. Materials for the comparison were made available on a web site for future researchers to be able to replicate/extend the comparison. The publicly-available materials include both the training and test sets of images, and software tools to automatically compare a machine segmentation to ground truth and score the result.

In the area of face recognition, the FERET evaluation procedures for automatic face recognition algorithms are well known. There have been three FERET tests, the August 1994 and March 1995 tests [23, 24], and another in September 1996 [21, 22, 28]. The FERET evaluation procedures tested the ability of algorithms to recognize faces from still images where the face occupied most of image. The difficulty of the tests increased over time. For the FERET tests, a database of face images was collected. A portion of the images was distributed to researchers for algorithm development and tuning. A second portion was sequestered for testing, which allowed algorithms to be evaluated on images that the researchers had not seen before.

The National Institute of Standards and Technology (NIST) has sponsored two evaluation conferences for optical character recognition (OCR) systems. At the evaluation conference, the results of an independent test are reported and the systems tested are described. For the evaluation, there is a dataset for development and training, and one for testing. A summary, overview, and discussion of the results of these two conferences can be found in Wilson [41] and Wilson et al. [42]. The reports of the evaluation conferences can be found Wilkinson et al. [40] and Geist et al. [11].

A number of researchers have evaluated OCR algorithms. Wilson et al. [2, 43] report performance on a number of classifiers, including nearest-neighbor, multi-layer perceptron, and radial basis function classifiers. The test was performed on NIST datasets. Ho and Baird [14] describe a method for generating synthetic data and report performance of nearest neighbor, decision tree, and distribution map classifiers on the synthetic data. One advantage of synthetic data is that very large datasets can be generated without the cost that would be incurred for real data.

Automatic target recognition (ATR) is a vision problem of interest to the military. An ATR algorithm detects and identifies a class of objects (usually military) in different forms of imagery. Prior to fielding ATR algorithms, the military is evaluating their performance. The US Army has procedures for performing independent of evaluations on infrared ATRs [27] and the US Air Force has procedures for synthetic aperture radar [19, 30]. Li [17] reports results on a number of infrared ATR algorithms that they implemented.

Borra and Sarkar conducted a performance evaluation of edge-based perceptual grouping algorithms [4]. The performance measures used in this work are related to the speed and the accuracy of both constrained-search-based and indexing-based object recognition strategies.

In 1996, a Workshop on Performance Characteristics of Vision Algorithms was held in conjunction with the European Conference on Computer Vision [6]. Two of the organizers of this workshop, Christensen and Förstner, then edited a special issue of the journal *Machine Vision and Applications* containing extended versions of selected papers from the workshop [5]. Their overview to the special issue summarizes a number of philosophical and methodological points related to performance evaluation.

The *MV&A* special issue contains ten papers selected from the workshop. Courtney et al [7] discuss issues of algorithmic modeling. They contrast analytical and numerical approaches, in the context of a feature detection algorithm, a stereo matching algorithm and an object recognition algorithm. Ramesh and Haralick [26] present a method of analyzing the performance of edge finding, linking and gap-filling. They give both analytical predictions and empirical results for simulated data. Wenyin and Dori [36] outline a protocol for evaluating line detection algorithms in document image analysis. They give experimental results from the analysis of synthetic images. Sheinvald and Kiryati [31] analyze the performance of the “subspace-based line detection” (“SLIDE”) algorithm relative to the Hough transform. They conclude that the SLIDE algorithm is less robust than the Hough transform. Vanrell et al [34] use a multi-dimensional scaling approach to analyze a particular texture perception algorithm. They use images from the well-known Brodatz texture album in their work. Eggert et al [9] compare the performance of four well-known algorithms for estimating 3-D rigid body transformations. They reach the conclusion that “... for real-world noise levels, there is no difference in the robustness of the final solutions.” Madsen [18] compares the performance of two pose estimation techniques. He looks at performance both for “unstable viewpoints” and “robust viewpoints.” Venetianer et al [35] investigate the performance of vision-based localization in robotic systems. Their work has the goal of being able to make statements about the performance of autonomous agents. Torr and Zisserman [33] look at an algorithm for estimating the “fundamental matrix” which represents the epipolar geometry between two images. They consider various levels of lossy JPEG compression and the effect that this has on performance. Cozzi et al [8] analyze the performance of phase-based algorithms for estimating stereo disparity. They hope to be able to estimate the expected accuracy of phase-based stereo analyzers in the face of common sources of error.

5 Summary

Empirical evaluation of algorithms is slowly emerging as a serious subfield in computer vision. This is demonstrated by the papers in this volume, recent evaluation workshops, and special issues and sections in computer vision journals. This is laying the foundation to develop accepted practices and methods for evaluating algorithms.

Evaluating algorithms lets researchers know the strengths and weaknesses of a particular approach and identifies aspects of a problem where further research is needed. From a practical point of view, successful evaluations help convince potential users that an algorithm has matured to the point that it can be successfully fielded. This helps to alleviate the situation where a solution is oversold, and the resulting disappointment when the algorithm is fielded.

To help empirical evaluations to become an accepted and expected part of the community we feel that the following steps need to be addressed. For well-established problems, there needs to be standard databases, evaluation protocols, and scoring methods available to researchers. Also, when papers in these areas are submitted for publication, journal editors and referees need to insist that authors evaluate their algorithms. Finally, just as journal reviewers and editors need to insist on appropriate evaluation methods, funding agencies need to give greater attention to these issues. While too great of a rigidity in what constitutes appropriate evaluation may stifle creative new work, the current near-total lack of emphasis of evaluation methods by funding agencies has its own dangers. The FERET program in face recognition algorithms is perhaps the best current example of evaluation methods being enforced by the funding agency.

References

[1] J. L. Barron, D. J. Fleet, and S. S. Beauchemin. Systems and experiment: Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, 1994.

[2] J. L. Blue, G. T. Candela, P. J. Grother, R. Chellappa, and C. L. Wilson. Evaluation of pattern classifiers for fingerprint and OCR applications. *Pattern Recognition*, 27(4):485–501, 1994.

[3] R. C. Bolles, H. H. Baker, and M. J. Hannah. The JISCT stereo evaluation. In *DARPA Image Understanding Workshop*, pages 263–274, 1998.

[4] S. Borra and S. Sarkar. A framework for performance characterization of intermediate level grouping modules. *IEEE Trans. PAMI*, 19(11):1306–1312, 1997.

[5] H. Christensen and W. Foerstner. Special issue on performance evaluation. *Machine Vision and Applications*, 9(5), 1997.

[6] H. Christensen, W. Foerstner, and C.B. Madsen. *Workshop on performance characterization of vision algorithms.* http://www.vision.auc.dk/h̃ic/perf-proc.html, 1996.

[7] P. Courtney, N. Thacker, and A.F. Clark. Algorithmic modelling for performance evaluation. *Machine Vision and Applications*, 9(5):219–228, 1997.

[8] A. Cozzi, B. Crespi, F. Valentinotti, and F. Worgotter. Performance of phase-based algorithms for disparity estimation. *Machine Vision and Applications*, 9(5):334–340, 1997.

[9] D.W. Eggert, A. Lorusso, and R.B. Fisher. Estimating 3-d rigid body transformations. *Machine Vision and Applications*, 9(5):272–290, 1997.

[10] O. Firschein, M. Fischler, and T. Kanade. Creating benchmarking problems in machine vision: scientific challenge problems. In *DARPA Image Understanding Workshop*, pages 177–182, 1993.

[11] J. Geist, R. A. Wilkinson, S. Janet, P. J. Gother, B. Hammond, N. J. Larsen, R. M. Klear, M. J. Matsko, C. J. C. Burges, R. Creecy, J. J. Hull, T. P. Vogl, and C. L. Wilson. The second census optical character recognition systems conference. Technical Report NISTIR 5452, National Institute of Standards and Technology, 1994.

[12] R. Haralick. Computer vision theory: the lack thereof. *Computer Vision, Graphics, and Image Processing*, 36:372–386, 1986.

[13] M. Heath, S. Sarkar, T. Sanocki, and K. Bowyer. A robust visual method for assessing the relative performance of edge detection algorithms. *IEEE Trans. PAMI*, 19(12):1338–1359, 1997.

[14] T. K. Ho and H. S. Baird. Large-scale simulation studies in image pattern recognition. *IEEE Trans. PAMI*, 19(10):1067–1079, 1997.

[15] A. Hoover, G. Jean-Baptiste, X. Jiang, P. Flynn, H. Bunke, D. Goldgof, K. Bowyer, D. Eggert, A. Fitzgibbon, and R. Fisher. An experimental comparison of range image segmentation algorithms. *IEEE Trans. PAMI*, 18(7):673–689, 1996.

[16] R. Jain and T. Binford. Ignorance, myopia, and naivete in computer vision systems. *CVGIP: Image Understanding*, 53(1):112–117, 1991.

[17] B. Li, Q. Zheng, S. Z. Der, and R. Chellappa. FLIR-ATR techniques. *Automatic target Recognition VIII*, Proceedings of SPIE Vol. 3371, (in press 1998).

[18] C.B. Madsen. A comparative study of the robustness of two pose estimation techniques. *Machine Vision and Applications*, 9(5):291–303, 1997.

[19] J. C. Mossing and T. D. Ross. MSTAR evaluation breaks new ground: methodology, results, infrastructure, and data analysis. *Algorithms for synthetic aperture radar imagery V*, Proceedings of SPIE Vol. 3370, (in press 1998).

[20] T. Pavlidis. Why progress in machine vision is so slow. *Pattern Recognition Letters*, 13:221–225, 1992.

[21] P. J. Phillips, H. Moon, P. Rauss, and S. Rizvi. The FERET evaluation methodology for face-recognition algorithms. In *Proceedings Computer Vision and Pattern Recognition 97*, pages 137–143, 1997.

[22] P. J. Phillips, H. Moon, S. Rizvi, and P. Rauss. The FERET evaluation. In H. Wechsler, P. J. Phillips, V. Bruce, F. Fogelman-Soule, and T. S. Huang, editors, *Face Recognition: From theory to applications*. Springer-Verlag, Berlin, (to appear 1998).

[23] P. J. Phillips, P. Rauss, and S. Der. FERET (face recognition technology) recognition algorithm development and test report. Technical Report ARL-TR-995, U.S. Army Research Laboratory, 1996.

[24] P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss. The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing Journal*, in press, 1998.

[25] K. Price. Anything you can do, I can do better (no you can't). *Computer Vision, Graphics, and Image Processing*, 36:387–391, 1986.

[26] V. Ramesh and R.M. Haralick. Random perturbation models for boundary extraction sequence. *Machine Vision and Applications*, 9(5):229–239, 1997.

[27] J. A. Ratches, C. P. Walters, R. G. Buser, and B. D. Guenther. Aided and automatic target recognition based upon sensory inputs from image forming systems. *IEEE Trans. PAMI*, 19(9):1004–1019, 1997.

[28] S. Rizvi, P. J. Phillips, and H. Moon. A verification protocol and statistical performance analysis for face recognition algorithms. In *Proceedings Computer Vision and Pattern Recognition 98*, (to appear) 1998.

[29] L. G. Roberts. Machine perception of three-dimensional solids. In J.T. Tippett et al., editor, *Optical and Electro-Optical Information Processing*, pages 159–197. MIT Press, Cambridge Massachusetts, 1965.

[30] T. D. Ross, S. W. Worrell, V. J. Velten, J. C. Mossing, and M. L. Bryant. Standard SAR ATR evaluation experiments using the MSTAR public release data set. *Algorithms for synthetic aperture radar imagery V*, Proceedings of SPIE Vol. 3370, (in press 1998).

[31] J. Sheinvald and N. Kiryati. On the magic of slide. *Machine Vision and Applications*, 9(5):251–261, 1997.

[32] C. Taylor and D. Kriegman. Structure and motion from line segments in multiple images. *IEEE Trans. PAMI*, 17:1021–1032, 1995.

[33] P.H.S. Torr and A. Zisserman. Performance characterization of fundamental matrix estimation under image degradation. *Machine Vision and Applications*, 9(5):321–333, 1997.

[34] M. Vanrell, Jvitrià, and X. Roca. A multidimensional scaling approach to explore

the behavior of a texture perception algorithm. *Machine Vision and Applications*, 9(5):262–271, 1997.

[35] P.L. Venetianer, E.W. Large, and R. Bajcsy. A methodology for evaluation of task performance in robotic systems. *Machine Vision and Applications*, 9(5):304–320, 1997.

[36] L. Wenyin and D. Dori. A protocol for performance evaluation of line detection algorithms. *Machine Vision and Applications*, 9(5):240–250, 1997.

[37] J. West, J. M. Fitzpatrick, M. Y. Wang, B. M. Dawant, C. R. Maurer, Jr., R. M. Kessler, R. J. Maciunas, C. Barillot, D. Lemoine, A. Collignon, F. Maes, P. Suetens, D. Vandermeulen, P. A. van den Elsen, P. F. Hemler, S. Napel, T. S. Sumanaweera, B. Harkness, D. L. G. Hill, C. Studholme, G. Malandain, X. Pennec, M. E. Noz, G. Q. Maguire, Jr., M. Pollack, C. A. Pelizzari, R. A. Robb, D. Hanson, and R. P. Woods. Comparison and evaluation of retrospective intermodality image registration techniques. *Medical Imaging 1996: Image Processing*, Proc. SPIE 2710:332–347, 1996.

[38] J. West, J. M. Fitzpatrick, M. Y. Wang, B. M. Dawant, C. R. Maurer, Jr., R. M. Kessler, and R. J. Maciunas. Retrospective intermodality registration techniques: Surface-based versus volume-based. In J. Troccaz, E. Grimson, and R. Mösges, editors, *CVRMed-MRCAS '97*, pages 151–160. Springer-Verlag, Berlin, 1997.

[39] J. West, J. M. Fitzpatrick, M. Y. Wang, B. M. Dawant, C. R. Maurer, Jr., R. M. Kessler, R. J. Maciunas, C. Barillot, D. Lemoine, A. Collignon, F. Maes, P. Suetens, D. Vandermeulen, P. A. van den Elsen, S. Napel, T. S. Sumanaweera, B. Harkness, P. F. Hemler, D. L. G. Hill, D. J. Hawkes, C. Studholme, J. B. A. Maintz, M. A. Viergever, G. Malandain, X. Pennec, M. E. Noz, G. Q. Maguire, Jr., M. Pollack, C. A. Pelizzari, R. A. Robb, D. Hanson, and R. P. Woods. Comparison and evaluation of retrospective intermodality image registration techniques. *J. Comput. Assist. Tomogr.*, 21:554–566, 1997.

[40] R. A. Wilkinson, J. Geist, S. Janet, P. J. Gother, C. J. C. Burges, R. Creecy, B. Hammond, J. J. Hull, N. J. Larsen, T. P. Vogl, and C. L. Wilson. The first optical character recognition systems conference. Technical Report NISTIR 4912, National Institute of Standards and Technology, 1992.

[41] C. L. Wilson. Effectiveness of feature and classifier algorithms in character recognition systems. Technical Report NISTIR 4995, National Institute of Standards and Technology, 1992.

[42] C. L. Wilson, J. Geist, M. D. Garris, and R. Chellappa. Design, integration, and evaluation of form-based handprint and ocr systems. Technical Report NISTIR 5932, National Institute of Standards and Technology, 1996.

[43] C. L. Wilson, P. J. Grother, and C. S. Barnes. Binary decision clustering for neural-network-based optical character recognition. *Pattern Recognition*, 29(3):425–437, 1996.

[44] R. Zhang, P. Tsai, J. Cryer, and M. Shah. Analysis of shape from shading techniques. In *Computer Vision and Pattern Recognition*, pages 377–384, 1994.