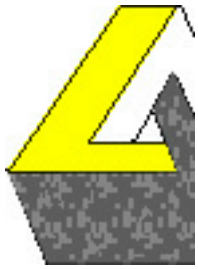

9th Workshop "Theoretical Foundations of Computer Vision"

Evaluation and Validation of Computer Visi Algorithms

Schloß Dagstuhl/Wadern (near Saarbrücken/Germany)

March 16 - 20, 1998

Chairmen

Robert Haralick (Seattle), Reinhard Klette (Auckland),
H. Siegfried Stiehl (Hamburg), Max Viergever (Utrecht)

This workshop addressed a subject which has been under active discussion in computer vision for several ye The evaluation and validation of algorithms is of basic importance for the configuration of computer vision applications.

> "...the authors should clearly state their assumptions and justify their claims based on results that can be obtained under those assumptions. Vague justifications, such as subjective evaluations of images ... should not be allowed." (R.C. Jain, Th.O. Binford: Ignorance, myopia, and naivetÈ in computer vision systems. Computer Vision, Graphics, and Image Processing: Image Understanding 53 (1991), pp. 112 - 117.)

In the ideal case certain "data sheets" should allow to qualify algorithmic solutions in a specific context, e.g. defined by image data, goal of image analysis, or software environment ("edge detection is not equal to edge detection"). There is a lack of methodological fundamentals in the field of performance analysis.

Workshop Program (pdf file, 25k)
Draft of Workshop Report (pdf file, 189k)

Participants presented ideas towards general methodology of evaluation and validation of algorithms or/and t application of these techniques to computer vision algorithms.

The **proceedings** of the workshop will be published by Kluwer.

Related Web Sites

ECV NET Benchmarking and Performance Evaluation
IEEE Computer Society Workshop 1998
DAGM-Working Group Quality Characteristics of Pattern Recognition Algorithms

---

Previous Workshops "Theoretical Foundations of Computer Vision":

8th (1996): Advances in Computer Vision (ed. by F. Solina, W. Kropatsch, R. Klette and R. Bajcsy), Advances in Computing Science, Springer Wien New York, 1997 (ISSN 1433-0113).
7th (1994): Theoretical Foundations of Computer Vision (ed. by W. Kropatsch, R. Klette and F. Solina), Computing Supplement 11, Springer Wien New York, 1996 (ISSN 0344-8029).
6th (1992): Theoretical Foundations of Computer Vision (ed. by R. Klette and W. Kropatsch), Mathemati Research, Akademie Verlag Berlin, 1992 (ISSN 0138-3019).
5th (1990): Theoretische Grundlagen der Digitalen Bildverarbeitung, Mägdesprung/Harz (no proceedings published).
4th (1988): Theoretische Grundlagen der Digitalen Bildverarbeitung (ed. by R. Klette and W.-D. Klix), H 104/88, Technische Universität Dresden, Sektion Mathematik, 1988.
3rd (1986): Digitale Bildverarbeitung (Theoretische Grundlagen) (ed. by R. Klette and W.-D. Klix), Heft Technische Universität Dresden, Sektion Mathematik, 1986.
2nd (1984): Methoden und Algorithmen der Digitalen Bildverarbeitung (ed. by R. Klette and W.-D. Klix) 77/84, Technische Universität Dresden, Sektion Mathematik, 1984.
1st (1982): Digitale Bildverarbeitung (ed. by R. Klette and W.-D. Klix), Heft 60/82, Technische Universit Dresden, Sektion Mathematik, 1982.

CITR: last update: 29 April 1998

Theoretical Foundations of Computer Vision: Evaluation and Validation of Computer Vision Algorithms and Methods

March 16 – March 20, 1998

The third TFCV meeting in Dagstuhl addressed a subject which has been under intensive (and partly controversial) discussion in the computer vision community for several years. The evaluation and validation of computer vision algorithms and methods is of key importance, particularly for the configuration of reliable and robust computer vision systems and as well as for the dissemination of reconfigurable systems in novel application domains. Although in certain areas of computer vision a plethora of literature is already available on this subject, the research community still faces a lack of a well-grounded, mutually accepted, and (possibly) standardized methodology.

The range of fundamental problems encompasses, e.g., the value of synthetic images in experimental computer vision, the selection of a representative set of real images related to specific domains and tasks, the definition of ground truth given different tasks and applications, the definition of experimental test-beds, the analysis of algorithms with respect to general features such as computation time, convergence, stability, or domains of admissible input data, the definition and analysis of performance measures for classes of algorithms, the role of statistics-based performance measures, the generation of data sheets with performance measures of algorithms supporting the system engineer in his configuration problem, etc. The workshop thus attempted to bring together experienced colleagues from the international computer vision community both to discuss the state-of-the-art and to formulate recommendations for future activities.

36 talks, grouped in several topical sessions, were given by 32 speakers from 14 countries. Out of a total of 41 participants, 11 were from Germany; 5 from The Netherlands and the U.S.A. each; 3 from Denmark and United Kingdom each; 2 from Canada, Czech Republic, Japan, New Zealand, and Slovenia each; and 1 participant from Hungary, Israel, Slovakia, and Taiwan each. In addition to the oral presentations, four working groups - partly working during wood-walking tours - were established to discuss issues of common interest in more detail (see appendix for summaries). The participation of a number of younger scientists from Denmark, Germany, The Netherlands, and United Kingdom was rendered possible through financial support from the TMR (Training and Mobility of Researchers) programmme of the European Union which is gratefully acknowledged. Moreover we are pleased to state that the presentations at our meeting were of such a high quality that a refereed proceedings book is planned to be published soon by Kluwer Academic Publishers in the Computational Imaging and Vision series.

We are also grateful to the administration of the Dagstuhl enterprise for creating such an inspiring and free-of-duty environment as well as for providing excellent facilities

which significantly contributed to the success of our meeting.

Eventually the workshop has stimulated different future activities, ranging from the establishment of an algorithmic web site for the international computer vision community, to the recommendation of organizing a similar meeting at Schloß Dagstuhl in Y2K on the subject of theory, methodology, and practice of performance measures.

Robert M. Haralick Reinhard Klette H. Siegfried Stiehl Max A. Viergever

List of Speakers

Monday

P. Courtney, A.W.M. Smeulders, W. Förstner, R. Marik, R.M. Haralick, N. Otha, A. Imiya, J.M. Buhmann

Tuesday

W. Niessen, M.A. Viergever, M.H. Loew, P. Courtney, D. Skocaj, I.T. Phillips, R.M. Haralick, D. Dori, V. Hlavac, P. Meer, P. Zamperoni

Wednesday

M. Petrou, D. Dori, I.T. Phillips, H.S. Stiehl, A. Leonardis

Thursday

B. Jähne, J.L. Barron, S.S. Beauchemin, L.M.J. Florack, M. Nielsen, C.-F. Chen, V. Ramesh, D. Chetverikov, G.L. Gimel'farb

Friday

D. Richter, R. Klette, F. Sloboda

PS: Please note that in the subsequent collection of abstracts the first name of the speakers is printed in full length.

Session I: General Issues

A Perspective from a Vision Company

PATRICK COURTNEY
Visual Automation Ltd., Manchester

High-reliability continuously operating systems of real social and economic value are routinely produced in disciplines based upon civil, aeronautical, and electrical engineering. This relies, in part, on a sound understanding of the system failure modes and the establishment of a set of compensating actions. Such knowledge is largely abscent in the engineering of vision systems. The problem is illustrated by case studies of projects for vision-based surveillance, non-contact measurement, and tracking systems. Under such circumstances, vision solutions are often rejected by the marketplace. Where careful evaluation work has been carried out, simple but useful design guidelines can be developed and can be used to produce workable systems. Still more work of this kind is required from the community of algorithm developers - both academic and industrial.

Design and Performance Analysis in Computer Vision

ARNOLD W.M. SMEULDERS, C. DE BOER
University of Amsterdam

In this communication we conceive of computer vision as a process of design. Designing a function in computer vision can be divided in 5 steps: 1. the design of the computational method, 2. the algorithmic design, 3. implementation of the code, 4. embedding the code in the system, and 5. putting it into use.

In an analysis of the life cycle of each of these steps, we define the computational method as the rationale behind the algorithm, usually in the form of a sequence of mathematical operations which transform the target images, captured in the set Ω_τ, into the result α. α can be anything from an image captured in Ω_α, to a number, a histogram, or a verbal description, where Ω is the space of all images. A computational method is the level of description of a function what is commonly reported in scientific journals. It will come to the end of its life cycle when it has no suitable implementation, when the method has unknown Ω_τ, when another method has a wider Ω_τ, or when another method is easier to understand. Performance analysis may help to extend the life of a method.

According to Cormen's standard book, an algorithm is a well defined computational process taking the input and transforming it into the output by a decomposition of calculations mapped onto an architecture of data structures. It is remarkable that the

metier of algorithmic design has gained little ground in computer vision, due to the fact that few methods are stable. An algorithm will come to its end when the method the architecture changes, when another algorithm has better efficiency, or when another algorithm is simpler to understand.

Of these 5 levels most of the daily effort goes into coding the implementation, whereas discussions in literature are few. This is a remarkable fact demonstrating underestimation of the importance of coding for reuse. The code of a vision function frequently comes to an end when the language, the interface, the libraries stop being used, but also when the previous programmer has left the place, or when it is unclear what and when the code works. In fact, in most practices, without performance analysis code is lost at first event.

The embedding in the surrounding system is frequently encountered as the gap between theory and practice. Embedded performance analysis is the proof for real. Therefore, for long term survival of functions, modular embedding of building blocks is essential using well defined interfaces, maintaining well defined internal states, and taking care that all supplier blocks are encapsulated in the function. For interactive or human inspection applications the implementation does not end with the embedding. In such cases systems may be left unused when all preceding steps function well but are incomprehensible or inoperable for the user. Performance evaluation at the level of use dominates everything else here. In fact, performance here equals online satisfaction measurement.

We discuss a technique called Ω sampling for which support tools have been build: an image synthesizer and a statistical evaluation platform called BESSI. An unknown detector algorithm was evaluated in a week after assignment. It is concluded that performance analysis is needed at all 5 levels of design. It is also concluded that performance analysis is feasible in practice.

Methodology and Experiences with Performance Characteristics in Computer Vision

WOLFGANG FÖRSTNER
University of Bonn

Progress in Computer Vision (CV) research and transfer of results of CV to practical applications requires performance characteristics of the developped methods. Experiences in Photogrammetry with experimental studies, especially within the "European Organisation for Experimental Photogrammetric Studies" (OEEPE) founded in 1953, demonstrate the high effort, the feasibility, and the positive practical response. Methodology relied on a dense interaction of model based theoretical predictions and experimental testing of the models together with experimental protocols which are homogeneous for a wide class of tasks and transparent to users. Transfer of this experience to CV problems requires the setup of standard (sub-)tasks, of standard quality measures, of

reference data sets, and a joint effort to setup experiments. The complexity of CV tasks forces careful performance analysis of basic methods of image processing, feature extraction, calibration and orientation, shape-from-X, and object recognition. Evaluating a segmentation procedure served as an example for the feasibility of such a path.

Quality in Computer Vision

RADEK MARIK
ProTys s.r.o., Prague

Quality of manufacturing processes is often assessed using theory of quality. Theory of quality could be used as a framework for performance analysis in computer vision. In this paper we present selected notions and demonstrate how performance analysis of computer vision algorithms could benefit from clearly defined results and concepts of theory of quality. We start with the definition of quality measure in terms of the total loss to society due to functional variation and harmful side effects. In addition, the roles of true product performance and its substitute are identified. Furthermore, we discuss principles of robust design that could help with setting of control parameters attached to vision algorithms. The second half of the paper is dedicated to computer vision oriented examples demonstrating old and new Japanese tools.

Session II: Statistics

Covariance Propagation

ROBERT M. HARALICK
University of Washington, Seattle

Computer vision algorithms are composed of different sub-algorithms often applied in sequence. Determination of the performance of a total computer vision algorithm is possible if the performance of each of the sub-algorithm constituents is given. The problem, however, is that for most published algorithms, there is no performance characterization which has been established in the research literature. This is an awful state of affairs for the engineers whose job it is to design and build image analysis or machine vision systems.

In this paper we give the solution to the simplest case: how to propagate approximately additive random perturbations through any kind of vision algorithm step in which the appropriate random perturbation model for the estimated quantity produced by the vision step is also an additive random perturbation. We assume that the vision algorithm step can be modeled as a calculation (linear or non-linear) that produces an estimate that minimizes an implicit scaler function of the input quantity and the calculated estimate. The only assumption is that the scaler function has finite second partial derivatives and that the random perturbations are small enough so that the relationship between the scaler function evaluated at the ideal but unknown input and output quantities and the observed input quantity and perturbed output quantity can be approximated sufficiently well by a first order Taylor series expansion.

Finally we discuss the issues of verifying vision programs by comparing their experimentally observed statistical behavior with that derived from covariance propagation.

Computer Vision Problems Viewed as Statistical Inference

NAOYA OHTA, K. KANATANI
Gunma University, Kiryu

Theories are very important for validation of computer vision algorithms. That is because theories predict the behavior and performance of the algorithms without experiments, once the input data are specified. We are taking a statistical approach to establish a general theory for computer vision.

A statistical approach has several advantages. One is that it covers a large range of computer vision problems. Computer vision problems are viewed as a physical measurement plus a pattern recognition problem. Statistics is good at both fields. Another is

that statistics represents data in probability terms such as estimated values and covariance matrices, which have clear meanings. Therefore, modules which construct a whole system have well specified interfaces.

We show one computer vision system as an example of our statistical approach. The system computes object shape and detects moving objects using an image sequence. It consists of three modules: optical flow detection module, structure-from-motion module, and moving object detection module. The first two are categorized in the physical measurement modules and the last is in pattern recognition modules. The optical flow detection module computes optical flow and its covariance matrices from two adjacent frames in an image sequence. The structure-from-motion module uses the optical flow and its covariance matrices and estimates the motion of the camera and depth to the environment. Variances for the depth and a covariance matrix for the motion parameter are also computed. The moving object detection module uses the depth, the motion parameter, and the associated variance information and makes a decision where moving objects are. This system is only one example, but it suggests that our statistical approach is applicable to a wide range of computer vision problems.

Randomized Method for Computer Vision

ATSUSHI IMIYA
Chiba University, Chiba

Some model equations of computer vision problems are given in the forms that the inner product of a datum vector and the parameter vector is zero, and that the determinant of a matrix whose row vectors are data themselves is zero. Sometimes a model equation is given as a system of equations described using both of these forms. For instance, planar lines and planar conics are expressed as the inner product forms, and motion equation is expressed as a system of equations of the inner product forms. Furthermore, the shape reconstruction problem is expressed in the forms of the determinant is zero for stereo pairs and the determinants are zero as a system of equations for multiple views. Ambiguities of rotation parameters of motion for symmetric shapes decide the folding numbers of planar symmetric shapes. These are problems with the linear constraints, the quadric constraints, and the nonlinear constraints.

In this talk, we first show that these expressions of problems permit us to estimate parameters and geometric properties using the random sampling and voting procedure, which estimates parameters by collecting evidence from randomly sampled data. The method derives high-performance parallel algorithms for solving model-fitting problems in computer vision as in the case of the Hough transform for line and conic detection problems. Furthermore, this method permits us to solve non-point-correspondence problems of motion analysis and shape reconstruction for which we do not assume any predetermined point correspondences between or among frames, described in these forms. Second, we derive algorithms being numerically stable against both digitization errors and noise for motion classification, motion analysis, shape reconstruction, and

symmetry detection problems. Third, we evaluate the performance of these algorithms using synthetic data. Finally, we analyze the complexities of these algorithms and we prove trade-off relations between the computational time and the amount of memory for the computation of some algorithms that we propose.

Robust Computer Vision and Statistical Learning Theory

JOACHIM M. BUHMANN
University of Bonn

The maximum entropy method as a design principle for computer vision algorithms has been successfully applied to a large variety of low–level and intermediate–level vision problems like image restoration, motion estimation, shape-from-X problems, image segmentation, and other computer vision tasks modeled with Markov Random Fields. This talk specifies a vision problem as an optimization task of a suitable cost function. Stochastic optimization schemes like simulated annealing are employed to solve the problem at a preselected approximation level. The variables which have to be optimized in the vision task have to be considered as random variables. Large uncertainty/noise in the stochastic search effectively smoothes the cost function. The solution of a vision problem is parametrized by the expectation values of these random variables which are distributed according to the Gibbs distribution.

The maximum entropy approach can be justified by statistical learning theory which provides a formal framework for statistical inference. The difference between the empirical risk which is calculated on the basis of image information and the expected risk is proposed as a measure for robustness since it controls the generalization properties of computer vision solutions. This concept is explained for the case of K-means clustering of vectorial data. Results of the maximum entropy approach are presented for unsupervised texture segmentation on the basis of a multiscale Gabor pyramid with pairwise clustering as the cost function for compact and homogeneous texture segments.

Session III: Segmentation and Feature Detection

Accuracy and Evaluation of Medical Image Segmentation

WIRO J. NIESSEN, K.L. VINCKEN, C.J. BOUMA, M.A. VIERGEVER
University Hospital Utrecht

There is a strong need for algorithms performing objective, reproducible segmentations of medical image data. However, the introduction of these techniques into clinical practice has been hampered by the lack of thorough evaluation of performance. A full validation study is costly, since it should ideally comprise both a large number of datasets and a large number of trained medical experts to inspect these. In case of evaluating segmentation results one should compare the weighted average of these observers with the result of the algorithm. An algorithm performs well enough if it is closer to the average than the variance of the medical experts. Two important steps in this procedure are determining a gold standard (the weighted average) and an error metric to define how much an individual result differs from the gold standard. Based on a clinical evaluation of different segmentation techniques for intravascular ultrasound images, we show that there is a strong task dependence in both these steps.

A promising tool to partially replace costly evaluation procedures is the use of simulations which model the main image degrading effects in acquisition, and the anatomical variability of individuals. We tested an algorithm for segmenting grey matter, white matter, and cerebrospinal fluid from 3D MRI data on simulated MR data which were generated from a digital phantom. Since ground truth is available, we could compare results to other algorithms. Moreover, dependence on image degrading effects can be tested. This showed that in case of MR imaging, modeling of partial volume voxels (i.e. voxels containing multiple tissue types) is one of the essential steps to improve accuracy in segmentation results. We believe that further sophistication of the simulation of image formation will improve the design of future algorithms.

Model-Based Evaluation of Image Segmentation Methods

A.S.E. KOSTER, K.L. VINCKEN, O.C. ZANDER, C.N. DE GRAAF,
MAX A. VIERGEVER
University Hospital Utrecht

A method is proposed for evaluating image segmentation methods, based on the costs needed to manually edit a segmented image until an objectively defined quality limit has been reached. In order to measure the quality of a segmentation, the true or desired object distribution should be known. Within the editing scenario, the number and the

type of actions to transform the segmented image into the eventual distribution can be emulated. By labeling these actions with costs, a segmentation can be evaluated quantitatively. This allows a comparison of different segmentations, and hence of different segmentation methods. The evaluation approach is illustrated using a straightforward editing scenario that features splitting and merging of segments. The images used in the evaluation are artificial as well as medical, 2D as well as 3D.

Performance Characterization of Landmark Operators

K. ROHR, H. SIEGFRIED STIEHL
University of Hamburg

We describe our studies on the performance characterization of operators for the detection and localization of point landmarks. We consider 2D as well as 3D landmark operators. Our general approach to the evaluation of image analysis methods consists of three principal steps: 1) modelling the signal structure of landmarks, 2) analysis of the degrees-of-freedom, and 3) theoretical and experimental performance analysis.

In the case of 2D landmark operators, we consider the localization of corners of polyhedral objects. We have performed an analytic study of ten well-known differential corner detectors. This study is based on an analytic model of an L-corner. The dependency of the accuracy on all model parameters given the full range of the parameter values has been analyzed. A different approach to evaluate corner detectors is based on projective invariants.

In the case of 3D landmark operators, we consider the localization of anatomical landmarks in tomographic images of the human brain. We have carried out several studies to assess the performance of different classes of operators. This includes an investigation of the localization accuracy in dependence of image blur and noise, as well as the application of statistical measures to assess the detection performance. The studies are based on synthetic data (e.g., tetrahedra and ellipsoids), where ground-truth is available, as well as on tomographic images of the human brain, i.e., MR and CT images. (Acknowledgement: Support of Philips Research Laboratories Hamburg, project IMAGINE (IMage- and Atlas-Guided Interventions in NEurosurgery), is gratefully acknowledged.)

Issues in Multimodality Medical Image Registration

MURRAY H. LOEW
George Washington University, Washington D.C.

By the mid-1990s a number of computer hardware, image processing, display, and networking technologies had advanced to the point where it became feasible to overlay images from different medical imaging modalities and to place corresponding anatomical

locations in registration. At present multimodality medical imaging is finding increasing use in diagnosis and in treatment planning, as physicians combine information about form and function to sharpen their understanding of location and pathology in clinical situations.

Registration of images is a prerequisite to the effective use of several sources of image data because anatomic landmarks that are important to physicians may be absent in some of the modalities. For example, precise image registration is critical in computer-assisted surgery where there is an unknown physical relationship between (1) pre-operative images (e.g., MRI; CT; PET/SPECT) used when developing the surgical plan, and (2) the patient when on the operating room table. Similarly, accurate registration is needed in mammography follow-up examinations to identify changes in the size and character of lesions previously identified. Often, the images are of different sizes, resolutions (in space, time, and/or intensity), and orientations. Non-rigid transformations occur frequently, further complicating the problem. In recent years a variety of registration algorithms have been developed and used by researchers for their own purposes; in some cases those algorithms were used by others (perhaps after modification), possibly for quite different purposes.

Developers and users of registration methods have measured the accuracy and precision of their methods in many ways. The multiplicity of anatomic sites, imaging parameters, sample sizes, computation strategies, and intended uses of the registered images have made it nearly impossible to compare the methods quantitatively, or even to characterize them. Thus it is difficult to provide a principled basis for the selection of a registration method for use in a given application.

Two one-day workshops were held in 1997 that brought together a group of 38 experts from nine countries to address several aspects of the issues surrounding multimodality medical image registration. The first workshop produced an agenda that defined the significant problems and assigned them to subgroups for study. The second reviewed the results, limited the scope of the problem, and proposed an action plan that focused initially on the development of standard data sets.

This paper discusses the nature of the problem (why it's interesting; why it's difficult) and the consensus plan that is emerging.

Session IV: Databases and Testbeds

On Databases for Performance Characterisation

A. CLARK & PATRICK COURTNEY
University of Essex & Visual Automation Ltd., Manchester

We describe a survey of image databases intended for comparing the performances of computer vision algorithms. The organization of the survey is described briefly and preliminary findings outlined; this gives an indication as to where the community is putting most effort into "benchmarking" activities. We then examine three particular areas in which databases have been particularly useful in characterizing and comparing performance: optical flow and face identification. Some shortcomings in some of these databases are pointed out. Finally, we consider skills that the vision community needs to develop: the design of databases and experiments.

Sharing Computer Vision Algorithms Over the World Wide Web

DANIJEL SKOCAJ, A. LEONARDIS, A. JAKLIC, F. SOLINA
University of Ljubljana

Researchers in the field of computer vision work on different computers using various operating systems. Therefore, it is very difficult to make the software available to others (e.g., for evaluation and comparison tests) in a convenient way. Fortunately, the Internet provides the means for communication across different platforms. We explore different possibilities of using the Internet for making computer vision algorithms publicly available. We describe how to build an interactive client/server application which uses the World Wide Web for communication. The client program which performs the interaction to the user is a Java applet, while the server program, containing computer vision algorithms, works on the server side as a CGI program. As an example, a stand-alone program for range image segmentation was transformed into a Java-client/CGI-server application and is now available as a service on the World Wide Web for use, testing, and evaluation.

A Methodology for Using UW Databases for OCR and Image Understanding Systems

IHSIN T. PHILLIPS
Seattle University

The ultimate goal for systems that do OCR or any aspect of document image understanding is to perform nearly perfectly over a broad range of document conditions and types. To develop such a system, suitable databases are needed for the training and testing, and the performance evaluation of the system during its development. The three carefully ground-truthed and attributed document databases (UW-I, UW-II, and UW-III) which have been issued recently from the University of Washington are suitable for this purpose.

This paper discusses methodologies for automatically selecting document pages and zones from the UW databases, having the desired page/zone attributes. The selected pages can then be randomly partitioned into subsets for training and testing purposes.

This paper also discusses three degradation methodologies that allow the developers of OCR and document recognition systems to create unlimited "real-life" degraded images – with geometric distortions, coffee stains, and water marks. Since the degraded images are created from the images in the UW databases, the nearly perfect (and expansive) original ground truth files in the UW databases can be reused. The process of creating the additional document images, the associated ground truth and attribute files require only a fraction of the original cost and time.

Session V: Performance Evaluation

Object-Process Methodology as a Framework for Performance Evaluation

DOV DORI
Technion, Haifa

Object-Process Methodology (OPM) is a systems development approach that models the system's structure and behavior within a single model. By using one model the method avoids the model multiplicity problem, of which current system development methods suffer. Complexity is managed by seamless recursive and selective scaling of objects and/or processes to any desired level of detail. Through OPCAT (Object-Process CASE Tool) OPM has been employed in a variety of real-life industrial system applications. OPM is generic as it is based on general system theory principles. It can therefore serve as an effective vehicle for specifying computer vision systems, which are object- and process-intensive, along with the performance evaluation that should be associated with each system and each algorithm within it.

A Case of Performance Evaluation in 3D Image-Based Vision

VACLAV HLAVAC, T. WERNER, T. PAJDLA
Czech Technical University, Prague

The image-based 3D scene representation allows to display the scene from any viewpoint from a set of stored 2D reference images. This approach is an alternative to rendering a full 3D model. The image interpolation was believed to be the proper rendering tool. We developed an optimization method for automatically selecting the set of reference views. Its performance was evaluated using synthetic images (ground truth). The results were difficult to explain near points on the surface where self-occlusion appeared.

We encountered that even for objects as simple as a convex polyhedron the image interpolation is not enough. The novel finding is that the configuration of visible surfaces from two views uniquely determines one of the three cases: (a) image interpolation, (b) image extrapolation, and (c) need for a full 3D model.

We analyzed the newly proposed extrapolation case and proposed a practically usable technology. Having two reference images in correspondence, it (1) does the projective reconstruction of a triangulated surface from a pair of images with correspondences, (2) allows to manually set desired viewpoint, and (3) effectively fills the interiors of

transferred triangles by correctly re-mapping the texture stored with one of the reference images. The oriented projective geometry solves the front/behind ambiguity. If the homography of a transfer is replaced by an affine transformation then the visibility can be solved by z-buffering the distance of a surface point from the epipole.

The Use of Resampling Methods for Performance Evaluation

PETER MEER, J. CABRERA, K. CHO, B. MATEI, D. TYLER
Rutgers University, Piscataway

A new performance evaluation paradigm for computer vision systems is proposed. In real situation, the complexity of the input data and/or of the computational procedure can make traditional error propagation methods unfeasible. The new approach exploits a resampling technique recently introduced in statistics, the bootstrap. Distributions for the output variables are obtained by perturbing the input with a noise process derived from it. Using these distributions rigorous performance measures can be derived solely from the given input. Covariance matrices and confidence intervals can be computed for the estimated parameters and individually for the corrected data points. The potential of the new paradigm is shown by employing it to compare different methods used in the low-level task of edge detection and in the middle-level task of rigid motion estimation.

Dissimilarity Measures Between Gray-Scale Images as a Tool for Performance Assessment

PIERO ZAMPERONI
Technical University of Braunschweig

The measure of dissimilarity D(A,B) between arbitrary gray-scale images A and B presented in this contribution is useful for a quantitative performance evaluation of image restoration, edge detection, thresholding, and segmentation methods. The performances of such methods are ranked on the basis of the lowest dissimilarity between the processed image and a ground truth, e.g. the original image or a-priori known edge maps. Further applications of the D(A,B) measure are: (1) quantitative evaluation of the cumulative dissimilarity caused by small amounts of shift, rotation, affine deformations, illumination changes, and different types of noise, and (2) selection, from an image database, of the most similar images with respect to a given comparison image.

After an overview of the state-of-the-art in gray-scale image comparison approaches, this contribution examines a set of properties that D should have, in order to cope with the tasks mentioned above. Then it proposes a multi-stage dissimilarity measure, in

which each stage, i.e. point-to-point, point-to-image, local image-to-image, and global image-to-image, can be realized by means of different distance measures, thus originating a manifold of variants of D. Some properties of these variants, related to the requirements posed to D and to the nature of the compared images, are examined and made the object of experimental verification.

Numerous experimental results, obtained with real-world images and with widespread low-level image processing operators to be assessed, illustrate the performances of the proposed measure in relation with the following tasks or desired properties: shift and rotation measurements, edge-preserving smoothing operators, edge detection operators, image segmentation by labeling, binarization with automatic threshold selection, robustness with respect to spike noise and to varying scene illumination, selection of the "most similar images" from a database, and face recognition

Session VI: Recognition and 3D Reconstruction

Sensitivity Analysis of Projective Geometry-Based 3D Reconstruction

MARIA PETROU
University of Surrey, Guildford

Projective geometry is a method that allows the reconstruction of the imaged scene, provided that the coordinate positions of at least 6 reference points lying on two planes are known, both on the image plane and in the world coordinate system. The method relies on the solution of a succession of algebraic equations. These equations are non-linear and as such they introduce inhomogeneous errors in the solution space, when there is some error in the measurements performed on the image plane and/or in the location of the reference points. The work presented is the complete sensitivity analysis of the method. The error amplification factor of each equation is defined as the coefficient that multiplies the variance of the distribution of the errors in the position of the reference points, to produce the variance of the distribution of errors in the output quantity (it is assumed that the error in both coordinates of the reference points are independent and distributed with the same variance). The error amplification factor is used to identify which parts of the scene can be reliably reconstructed, and how the reference points can be chosen so that the reconstruction is most accurate. The motivation of this work was the need to infer the 3D shape of a block of granite with the help of 4 cameras around it, so that the optimal way to cut it into slabs is determined in order to minimize waste.

Performance Evaluation of Graphics Recognition

DOV DORI, L. WENYIN
Technion, Haifa

Accurate and efficient vectorization of line drawings is essential for any high-level processing. However, there exists no standard for the evaluation of vectorization algorithms. We propose a protocol for evaluating the extraction of straight and circular line segments to help compare, select, improve, and design new algorithms. The protocol considers a variety of factors at the pixel level and at the vector level, including pixel overlap, endpoints proximity, width, shape, style, and fragmentation. Application to arc detection is demonstrated.

A Performance Evaluation for Graphics Recognition Systems

IHSIN T. PHILLIPS & A. K. CHHABRA
Seattle University & Bell Atlantic Network Systems, White Plains

This paper presents a benchmark for evaluating graphics recogniton systems. The benchmark is designed for evaluating the performance of graphics recognition systems on images that contain straight lines (solid or dashed), circles (solid or dashed), partial arcs of circles (solid or dashed) as well as the locations of text blocks within the images. This benchmark gives a scientific comparison of vectorization software and uses practical performance evaluation methods that can be applied to complete vectorization systems. Three systems were evaluated under this benchmark and their performance results are presented in this paper. We hope that this benchmark will help assess the state of the art in graphics recognition and highlight the strengths and weaknesses of current vectorization technology and evaluation methods.

Performance Evaluation of Eigenimage Recognition

ALES LEONARDIS, H. BISCHOF
University of Ljubljana

The basic limitations of the current appearance-based matching methods using eigenimages are non-robust estimation of coefficients and inability to cope with problems related to outliers, occlusions, and segmentation. We propose a robust approach which successfully solves these problems. The major novelty of our approach lies in the way how the coefficients of the eigenimages are determined. Instead of computing the coefficients by a projection of the data onto the eigenimages, we extract them by a hypothesize-and-test paradigm using subsets of image points. Competing hypotheses are then subject to a selection procedure based on the Minimum Description Length principle. The approach enables us not only to reject outliers and to deal with occlusions but also to simultaneously use multiple classes of eigenimages.

We performed an extensive comparison of the standard eigenspace method with the robust method (we designed two versions of it) on a standard database of 1440 images (20 objects under 72 orientations) under various noise conditions. A performance evaluation of the eigenimage recognition has determined the range of applicability for the methods under different noise conditions.

Session VII: Optic Flow and Low-Level Motion

Performance Characteristics of Low-level Motion Estimators in Spatiotemporal Images

BERND JÄHNE, H. HAUSSECKER
University of Heidelberg

Motion leads to oriented structures in spatiotemporal images. A unified concept is introduced to estimate motion by nonlinear spatiotemporal filtering. This concept includes the standard differential, tensor, and phase based techniques. It is taken as a basis for a detailed performance analysis without any specific assumptions and approximations about the spatial structure of the images. The formulation in continuous space also enables a separation into principal flaws of an approach and errors introduced by inadequate discretization. It is discussed how noise, accelerated motion, inhomogeneous motion, and motion discontinuities bias motion estimation. Another serious source for errors is the approximation of derivative operators by discrete difference operators. A new nonlinear optimization technique is introduced that minimizes the error in the direction of the gradient. For high levels of accuracy, it is also necessary to correct for the photoresponse nonuniformity (PRNU) of the image sensor. The PRNU leads to a static pattern in image sequences that biases the motion estimate towards zero velocity in regions with low contrast.

Using Optical Flow to Measure 2D and 3D Corn Seedling Growth

JOHN L. BARRON & A. LIPTAY
The University of Western Ontario, London & Greenhouse and Processing Crops Centre, Harrow

We use optical flow to measure 2D and 3D growth of corn seedlings. Our method is ultra-sensitive and operates in a non-intrusive, non-contact manner and can measure motions whose magnitude is as minuscule as 5 microns. Our 2D analysis, started in 1994, uses the standard motion constraint equation and a least squares integration method to compute optical flow, which is used as a measure of plant growth. A key assumption of our 2D analysis is that growth motion occurs locally in a 3D plane and the accuracy of the measurements depends on this assumption being satisfied. We describe two 2D growths experiments: the first shows high correlation between root temperature and plant growth and the second shows the growth/swaying of the tip of a corn seedling and of an emerging leaf over a 3 days period.

Both the 2D and 3D methods assume orthographic projection, which holds locally in the image plane. Our 3D optical flow calculation uses a single least square integration calculation to compute a single 3D velocity for each image. Each image in the sequence consists of two views of the same seedling; one view of the corn seedling is front-on while the second view is an orthogonal view (at 90 degrees) of the seedling, made by projecting the plant's orthogonal image onto a mirror oriented at 45° with respect to the camera.

We compute 3D optical flow at the corn seedling's tip by using a simple extension of the 2D motion constraint equation. Results show that 3D motion exhibits swaying circular motion. The accuracy was verified subjectively by superimposing growth vectors on the plant images and viewing the image sequence as a movie, and objectively by observing that the difference between the sum of the vertical components of the velocity in the front and side images between images 10 and 400 was within 1% of the difference of the center of masses of the points used in the velocity calculation of the front and side views of the same images.

Computing Multiple Image Motions

STEPHEN S. BEAUCHEMIN, R. BAJCSY
The University of Pennsylvania, Philadelphia

The computation of image motion for the purposes of determining egomotion is a challenging task as image motion includes discontinuities and multiple values mostly due to scene geometry, surface translucency, and various photometric effects such as surface reflectance. We present algorithms for computing multiple image motions arising from occlusion and translucency which are capable of extracting the information content of occlusion boundaries and distinguish between those and additive translucency phenomena. Sets of experimental results obtained on synthetic images are presented. These algorithms are based on recent theoretical results on occlusion and translucency in Fourier space.

Motion Analysis

LUC M.J. FLORACK
Utrecht University

Vicious circularities have always pervaded the field of image analysis. For example, "features" such as "edges" only exist by virtue of a fiducial "edge detector" used to extract them, whereas these detectors in turn are being constructed with the aim to detect those features that one is intuitively inclined to call "edges".

The paradox arises from abuse of terminology. The polysemous term "edge" is used in two essentially different meanings here: once as an operationally defined concept (output of an edge detector), and once as a heuristic feature pertaining to our intuition (a matter of interpretation) begging the question of appropriate detector design.

Clarity is served by a manifest segregation of "structure" (which, once conventionally defined, becomes a matter of public evidence) and "interpretation" (selection of one among many possible hypotheses all consistent with the evidence). A convenient way to do this is to embed "structure" into a framework of duality and to model "interpretation" as an external constraint (gauge condition) imposed on the class of possible interpretations (gauge invariance).

The proposed framework can be applied successfully to motion extraction. Duality tells us exactly what the effect is of using such-and-such filters, and the gauge condition represents a priori knowledge that suffices to solve the "aperture problem". The aperture problem as such pertains to an intrinsic local invariance (i.e. gauge invariance), which cannot be resolved on the basis of image evidence by itself.

Evaluation by Characterization of Generic Output: Optic Flow

MADS NIELSEN, O. F. OLSEN
University of Copenhagen

The statistical evaluation of an algorithm comprises the same complexity as the development of the statistically optimal algorithm. We therefore suggest a thorough analysis prior to testing in terms of statistic sampling of the input space. The following is an example of an analysis showing "What is the *generic* output of an optic flow computation?". The optic flow field is defined as preserving the intensity along flow-lines. Due to singularities in the image at fixed time, poles are created in the optic flow field. We describe the generic types of flow singularities and their generic interaction over time. Furthermore we analyze the generic structures and events in the optic flow field when analysed in a multi-scale framework. In a general analytic flow field, normally the topology is characterised by the points where the flow vanish again subdivided into repellers, attractors, whirls, and combinations hereof. We point out the resemblance, but also the important differences in the structure of a general analytic flow field, and the structure of the optic flow field expressed through its normal flow. Finally, we show examples of detection of these singularities and events detected from non-linear combinations of linear filter outputs.

Session VIII: Motion, Tracking, and Stereo

Performance Evaluation of Using Motion Estimation in Video Coding

CHI-FA CHEN
I-Shou University, Kaohsiung

This paper investigates the performance of using motion estimation in a video coder. The characteristics of motions in image sequences are firstly studied. It has been found that the distribution of motion vector in both horizontal and vertical directions can be described by the uniform function. It was also found from the simulation results that the grey level distribution of difference images with motion estimation becomes more compact than that without motion estimation. Motion estimation errors on translationally and rotationally moving areas are then discussed. They can be expressed respectively by a random variable with uniform distribution. The performance of motion estimation in video coding is evaluated analytically based on the model of a hybrid coder that was proposed by Chen and Pang. According to the derived results, there is nearly a 7-dB difference when the temporal correlation coefficient changes from 0.95 to 0.99 with the optimal prediction coefficient. The simulation results show a 3-dB to 7.2-dB improvement using block-match motion estimation with unity prediction coefficient. The efficiency of the motion estimator increases dramatically when the estimation inaccuracy is small, i.e., temporal correlation coefficient approaches unity. The derived results can be readily applied to the situations that use methods other than block matching algorithms.

Performance Characterization in the Context of Video Analysis

VISVANATHAN RAMESH
Siemens Corporate Research, Princeton

Image understanding systems are complex and are composed of several algorithms applied in a sequence. In past work, we outlined a systems engineering methodology for characterizing/designing computer vision systems. The methodology has two components: "performance characterization or component identification" and "application domain characterization". Performance characterization involves the determination of the deterministic or stochastic behavior of components in the system. This is done by

specifying an ideal data model, a perturbation model in the input, and determining the output data and perturbation models. Application domain characterization (relative to an algorithm component) involves the derivation of probability distributions describing the input population. This is done by using annotations. The expected performance of the system can then be obtained by simply integrating the performance criterion over the input population distribution.

While we illustrated the two steps for a fixed algorithm sequence, it is not clear how (for various application domains) one can develop a system to automatically configure the algorithm sequence and its parameters. We argue that contextual knowledge plays a dominant role in this translation of a task specification to an algorithm sequence and parameters. We illustrate this in the context of video analysis. More specifically, we present the design of a tennis video interpretation system and we show that domain knowledge regarding scene geometry, video editing, and rules of the game can be used in algorithm design. An annotation tool is used to learn statistics from the video data. These statistics are used to design appropriate decision rules at the various stages of the system.

Experimental Comparative Evaluation of Feature Point Tracking Algorithms

DMITRY CHETVERIKOV, J. VERESTOY
Hungarian Academy of Sciences, Budapest

In this study, dynamic scenes with multiple, independently moving objects are considered. The objects are represented by feature points whose motion is tracked in long image sequences. The feature points may temporarily disappear, enter, and leave the view field. This situation is typical for surveillance and scene monitoring applications.

Most of the existing approaches to feature point tracking have limited capabilities in handling incomplete trajectories, especially when the number of points and their speeds are large, and trajectory ambiguities are frequent. Recently, we have proposed a new tracking algorithm which efficiently resolves these ambiguities.

Tests reported in the previous studies are not statistically relevant and usually consider only sparse point sets, e.g., 4 - 10 points per frame. We have recently initiated a systematic, comparative performance evaluation study of feature point tracking algorithms. As a part of this evaluation, experimental results are presented for our algorithm and the algorithms by Rangarajan and Shah and by Salari and Sethi.

A motion model called Point Set Motion Generator is used to generate the synthetic motion sequences for testing the tracking algorithms. In the current version of the model, the randomly generated points move independently. The advantages of such motion are the variable frequencies of different trajectory ambiguities and the possibility of their statistical assessment. Later, simulations of assorted coherent motions will be added in order to test the capability to cope with correlated trajectories.

We use two different merits of tracking performance. The first one applies the strict trajectory-based criterion which only accepts perfect trajectories. The second merit of performance is link-based. This merit accounts for partially tracked trajectories. Based on these criteria, comparative experimental results are shown and discussed.

Session IX: 3D Objects and Scenes

Pros and Cons of Using Ground Control Points to Validate Stereo and Multiview Terrain Reconstruction

GEORGY GIMEL'FARB
The University of Auckland

The problem of validating binocular, trinocular, or multiple view terrain reconstruction by ground control points (GCP) is addressed. Computational low-level stereo reconstructs a terrain by searching for a 3D surface that gives the closest match between corresponding points (or patches) in given stereo images of the terrain. It is an ill-posed mathematical problem because, generally, different terrains can produce the same stereo images.

Some regularising heuristics may reduce this lack of uniqueness but, in any case, the choice of the GCPs has to be based on estimating confidences of these points. The confidence depends on robustness of matching the corresponding image points. In particular, uniform or repetitive patterns of a terrain colouring result in low-confident GCPs. The reconstructed digital models of a terrain should be supplemented with a supporting confidence map to facilitate the validation process.

Pros of using the GCPs in this process are dictated by the fact that they are "a must" for calibrating (orienting) the images and for measuring actual deviations between the true and the reconstructed terrains. Their cons result from the fact that only the highly confident GCPs can be checked for the validation purposes. The deviations in low-confident areas can hardly be considered as poor performance of a stereo algorithm. But, the computational confidence measures depend on the image features used for matching. Therefore, the GCPs, sufficiently confident for human vision, may possess low confidences in respect to a particular computational stereo algorithm, and vice versa. Also, in some cases, visually chosen GCPs may not correspond to the true terrain due to specific colouring that cheats human stereo perception. Some experiments in reconstructing terrains by symmetric dynamic programming approach and in mapping the confidences of the obtained terrains are presented and discussed. (Acknowledgement: This research is supported in part by the University of Auckland Research Committee Grant T01/XXXXX/62090/F3414076.)

A Systematic Approach to Error Sources for the Evaluation and Validation of a Binocular Vision System for Robot Control

DETLEF RICHTER
Fachhochschule Wiesbaden

The automatic control of production lines may be done in appropriate cases by robots controlled by digital image processing. CCD video cameras are used as optical sensors. When a two-dimensional image of a three-dimensional scene is generated the depth information is lost, however applications with a high degree of automation commonly require an exact recognition of three-dimensional objects and their positions. Therefore, binocular video cameras are used. For the application of image processing algorithms and subsequent robot control, a comprehensive calibrating procedure is required for the cameras and for the robot basis. An evaluation procedure of the corresponding software estimates only the reliability of the whole system, ignoring single effects of errors. In the paper presented, model dependent error sources are recorded systematically to find appropriate procedures for independent verification. It is presumed that no logical errors or errors due to computer internal process communications are present. This approach may be applied to similar complex model dependent software systems.

Multigrid Convergence of Surface Approximations

REINHARD KLETTE
The University of Auckland

The talk suggests a study of multigrid convergence as a general approach for model-based evaluations of computer vision algorithms. This criterion is common in numerical mathematics. Multigrid convergence is discussed in this talk for four different tasks in 3D surface approximation or 3D surface recovery.

Jordan surfaces and Jordan sets provide the proper model for discussing surface approximations. In this talk, an object in 3D Euclidean space is a simply-connected compact set bounded by a (measurable) Jordan surface. Objects in 3D Euclidean space are studied based on given digitizations. The applied multigrid digitization model assumes cube inclusion or cube intersection (side length of cube = grid constant) where the grid constant (2 to the power -r) converges towards zero.

The existence of convergent algorithms is discussed for the problems of approximations of polyhedral surfaces based on sampled data (the least-square error approximation is a convergent technique for planes), volume and surface area measurement for Jordan sets or 3D objects (convergent volume measurement is known for about 100 years, a provable convergent surface area measurement is a recent result in the theory of geometric approximations), and surface recovery by solving a (special) linear differential equation system (see also below for the talk by Kozera and Klette). The existence of

convergent algorithms for surface reconstruction based on gradient information (integration of conservative vector fields) or for the analysis of surface curvature are stated as open problems. For these problems it is suggested to discuss the existence of convergent techniques for specified subclasses of 3D objects.

Evaluation of Finite-Difference Based Algorithms in Linear Shape From Shading

R. KOZERA & REINHARD KLETTE
The University of Western Australia, Perth & The University of Auckland

A number of algorithms based on the finite difference method in linear shape from shading are presented and analyzed. The evaluation of different numerical schemes is achieved by analyzing and comparing stability, convergence, and domains of influence of each scheme in question. The whole stability and convergence analysis is supplemented by the well-posedness proof of the corresponding initial boundary value problem.

The discussion of the linear reflectance map exposes (as both authors hope) a relatively difficulty appearing during the analysis of a seemingly simple linear case. Finite difference techniques as discussed for this case can also be applied for the non-linear PDEs.

In a general sense this contribution shows that the evaluation of different shape reconstruction algorithms may be directed on specific algorithmic features. This approach may complement the statistical performance analysis of algorithms in computer vision.

On Approximation of Jordan Surfaces: A Topological Approach

FRIDRICH SLOBODA
Slovak Academy of Sciences, Bratislava

Approximation of measurable Jordan surfaces in R^3 homeomorphic with the unit sphere on the basis of the notion of a minimal Jordan surface is described. Given two simple polyhedra $P_{S_1}, P_{S_2}, P_{S_1} \subset P^o_{S_2}$, $\partial P_{S_1} = S_1, \partial P_{S_2} = S_2$, the problem is to find a minimal closed Jordan surface S in $G := P_{S_2} \setminus P^o_{S_1}$, $\partial G = S_1 \cup S_2$ surrounding P_{S_1}. It is shown that this problem is related to the approximation problem of measurable Jordan surfaces homeomorphic with the unit sphere.

Two types of methods are introduced: the method of shrinking (compression) and the method of expanding (exhaustion). Both methods are performed on a orthogonal grid by gridding technique. The methods represent a topological approach to the approximation of Jordan surfaces homeomorphic with the unit sphere in R^3 and have a direct impact on integral calculus, computer aided geometric design, computer graphics, and image processing.

Appendix: Reports of Working Groups I - IV:

WG I: Performance Analysis, Web Sites, and Teaching Computer Vision

Chair: ROBERT M. HARALICK
University of Washington, Seattle

Our group initiated an activity to bring together in a coordinated set of web pages material that would be useful in helping researchers and students to learn about characterizing and evaluating the performance of vision algorithms as well as material in helping students to learn about computer vision. The web page material will be both tutorial in content by way of reports, papers, and viewgraphs as well as by way of software implementing algorithms and software characterizing and evaluating the algorithms implemented. The main site will be http://george.ee.washington.edu.

The software will have some web page front end so that it can be demonstrated over the internet.

Each algorithm web page will have places for the following:

1. A self-contained viewgraph tutorial of the algorithm, suitable for teaching
2. Reports or papers describing the algorithm and properties of the algorithm such as time complexity, convergence, numerical stability etc.
3. References to the literature
4. Code for the algorithm including a web interface to enable running of the algorithm by a web browser
5. Documentation for the running of the algorithm
6. Data sets and their ground truth, possibly including some negative examples on which the algorithm does not perform satisfactorily.
7. Software for generating synthetic data sets with ground truth
8. Software for evaluating the results of the algorithm against the ground truth
9. A self-contained viewgraph tutorial of how the characterization of the algorithm performance or the evaluation of the algorithm is done
10. A demo involving a run on real and/or synthetic data

The web pages will be started without necessarily having proper material for all the above places. The community of interested researchers will help fill in any empty place.

WG II: Task-Based Evaluation of Computer Vision Algorithms

Chair: MAX A. VIERGEVER
University Hospital Utrecht

The discussions in the working group on task-based evaluation of computer vision algorithms reflected the dichotomy of the workshop as a whole: analysis of performance characteristics of computer vision algorithms in a well-defined context versus evaluation of methods to carry out an image analysis task in a complex environment.

Most of the research on performance characterization is carried out in the framework of a highly constrained application as, e.g., document analysis or photogrammetry. In such areas, the model underlying the computer vision algorithms generally reflects physical reality well. In complex real world problems as, e.g., in seismic imaging or medical imaging, hard model constraints can be imposed only to a limited extent. In medical imaging, for instance, the biological variability between patients as well as the occurrence of unknown pathologies contest the validity of case-independent model assumptions. This creates a situation in which the model error may be orders of magnitude larger than pure algorithmic inaccuracies.

The first round of the working group discussions was taken up by recognizing and analyzing the above differences. The very establishment of the dichotomy between algorithmic performance characterization and application dependent system evaluation has been the main result of this working group.

The second discussion round focussed on computer vision in medical imaging, the primary field of interest of most of the working group participants. In medical imaging, the crucial issue for acceptance of a computer vision algorithm is whether it improves the diagnosis or the treatment, or at least the diagnostic/therapeutic procedure (similar results, but faster and/or more objective). Evaluation of algorithms needs to be carried out in this perspective.

The major point of discussion was how to reconcile ground truth and realism. Computer-generated data are a perfect gold standard, but provide little realism. They may serve to establish a consistency check of computer vision algorithms, but cannot predict how well the application will benefit. On the other hand, real world images (patient data) lack the objectivity needed for a gold standard. For example, image segmentation may be evaluated on patient data that have been segmented manually by experts, but the evaluation will then have to deal with issues as intra- and inter-operator dependency and inter-patient variations. This will require very large and thereby expensive studies, which will - if at all - be performed only on entire software procedures, certainly not on individual algorithmic components.

While the discussions did not result in practical suggestions for evaluation of algorithms in ill-defined environments, the working group considered the recognition and analysis of the complexity of this type of evaluation quite valuable. In consequence, the establishment of the working group has well served its purpose.

The working group consisted of the following participants: L.M.J. Florack, S. Frantz, M.H. Loew, P. Meer, M. Nielsen, W. Niessen, I.T. Phillips, H.S. Stiehl, M.A. Viergever, and P. Zamperoni.

WG III: Specification and Propagation of Uncertainty

Chair: WOLFGANG FÖRSTNER
University of Bonn

The scope of the working group was to discuss the two questions:

1. How to specify the required performance of an algorithm?

2. How to perform propagation of uncertainty within an algorithm?

We proceeded in a top down way to attack the problem. We hypothesize each algorithm can be specified by its input (x, Q_x), its output (y, Q_y) and the theory A of the transition between input and output. The input and output values x and y are uncertain, their quality is described by Q_x and Q_y. This allows to link different atomic algorithms to form an aggregated algorithm. Specification then is to pose constraints on the range of y and Q_y, which are task specific.

The discussion gave the following results:[1]

1. Proof of the hypothesis: This was done by investigating a complicated enough task (3d-object recognition) which contains enough representative subtasks.

2. Discussion of the possible structures of input and output: The discussion showed that these may be simple or composed structures with subparts being: integers, reals, labels sets, lists, vectors, matrices, tensors graphs, trees, relational structures etc. The *problem* therefore is *to establish a theory of quality measures for the different types of data.*

3. Discussion of the representation and propagation of uncertainty: It appears that statistics provides the appropriate tools, including covariance matrices, confusion matrices, Markoff-Chains, Markoff-Random-Fields, Bayesian networks, graphical models. However, *the task of modelling the uncertainty, of estimating the free parameters, of propagating the uncertainty and of testing the validity of the chosen models is by far not worked out for the basic tasks in image analysis.* The numerous examples from literature however show, that the tools are available and suggest the statistical framework to be applicable.

The working group consisted of the following participants: S. Bailey, C. de Boer, P. Faber, W. Förstner, R.M. Haralick, H. Haussecker, N. Ohta, and J. Sporring.

[1] (cf. also http://www.ipb.uni-bonn.de/ipb/lit/abstracts98/foerstner98.dagstuhl-wg.html)

WG IV: Modular Design

Chair: ARNOLD W.M. SMEULDERS
University of Amsterdam

We consider the methods of performance evaluation as a useful technique to enhance the usability and the life time of techniques of computer vision, it our case by splitting them up in basic building blocks: the modules. We are of the opinion that proper evaluation is against what the module claims it does. So composed modules have to be tested against their intended purpose and with the data corresponding to that purpose. A description of its purpose useful for a general public ("this module finds edges") may not nearly be precise enough for evaluation. A central thesis to investigate is the following. When the inputs and outputs of two modules each are of the same type, the purpose of the performance evaluation is equal, the performance evaluation shows great similarity.

As we have restricted ourselves to low-level image processing, the first focus of attention is to evaluate known intensity patterns. We can restrict ourselves to synthetic images where the ground truth is known. This opens the way for automation of the performance evaluation process, and sampling the space of all images at will. The second focus of attention must be on images with a real world meaning. They have a semantic interpretation. Such will change the character of performance evaluation as here synthetic images with a computational ground truth are not available. It may be even so that the ground truth is only available as a best-consensus opinion. For one thing, this limits the number of images to consider as well as it changes the measure of quality.

The purpose of the working group is to establish patterns in performance analysis methods for basic building blocks. The study will be continued on after the workshop. The case of segmentation reads as an operation of "image" $\rightarrow$ "map". The ground truth in this case is also of the type "map". All essentially different data configurations are listed as point (blobs, corners), line (edges, curves) or regions. Each of these possibilities requires a specific detector with a specific metric of performance. The metric has as input: metric(map, map). For a line segmentation algorithm a line–line metric is needed. In the metric, there should be included: under- and overdetection (has it been detected at all) and locality (is the result on the right place). The alternative is that there is a separate metric for detection as well as for locality. A secondary performance evaluation is the decay of the distance measure when the noise is increased until the point where the algorithm looses track. This is a measure of robustness against noise and other perturbations of the computational model. Other robustnesses, e.g against different noise models, may also be considered.

In the composition of modules, an intriguing question is performance evaluation for composed modules as a product of the performance evaluation of each of the atoms. If such a scheme exists it would be of great help as modules can be evaluated each in their own right and do not need to be evaluated again when put into a bigger system. Such orthogonality would serve performance evaluation considerably as evaluation can be done once. There would be no need to repeat the evaluation any time the module is

embedded in a bigger system. However, there is doubt whether this orthogonality can exist.

The working group consisted of the following participants: M. Petrou, R. Klette, J. Puzicha, and D. Richter.