Machine Vision and Applications (1997) 9: 215–218

Machine Vision and Applications

Editorial

Performance characteristics of vision algorithms

Henrik I. Christensen[1], **Stockholm,**
Wolfgang Förstner[2], **Bonn**

[1] CVAP, NADA, KTH, S-10044 Stockholm, Sweden
[2] Institute of Photogrammetry, Universität Bonn, D-53115 Bonn, Germany

1 Motivation for this issue

For at least 10 years computer vision has been confronted with papers and discussions on the scientific value of its results and the difficulties in transferring the results to practical systems.

A change of awareness has happened: More than 10 years ago, at the Computer Vision Workshop 1985, two controversial papers with different viewpoints, agreed on the *lack of theoretical research* [3, 7], which should go along with the development of vision procedures: experimental proofs are not enough. Five years ago, the dialogue on 'Ignorance, Myopia, and Naiveté in Computer Vision Systems' initiated by R. Jain and T. Binford [4] and the responses documented the necessity of evaluating theoretical findings, vision procedures algorithms etc. by *using empirical data* in order to increase the number of real world applications of computer vision research.

When observing the increasing number of papers which propose new solutions to classical problems, especially using increasingly more demanding theoretical tools, it seems to become clear that empirical testing of vision algorithms is necessary to allow a clear comparison of the proposed methods by the users of such algorithms. Together with the underlying theories a clear performance characterization of algorithms is necessary.

This special issue is motivated by the belief that the lack of performance characterization of vision algorithms is responsible for the hesitation of industry to use computer vision as one of its tools. Reasons for this situation are manifold: the lack of commonly accepted criteria for evaluation, the lack of a methodology for testing, the lack of translating the experience in testing of other engineering areas to computer vision and possibly also the non-acceptance of empirical or theoretical comparisons of vision algorithms, including their replication, as original research.

2 Common objections against performance characterization

However, when discussing the necessity of empirical testing and performance characterization a number of strong objections are posed repeatedly. Their honesty cannot be debated. It is thus necessary to address these objections in a serious manner but also to show either their shortsightedness or the means to overcome such objections. The main objections are summarized in the following [2]:

1. *Evaluation is task dependent.* Yes, the number of tasks is too large to enable *evaluation* of the algorithms for all such situations. But *characterization* of the performance of algorithms can be parameterized and it allows the user of an algorithm to choose and evaluate it without actually running the algorithm.
2. *Vision is only one module within a complex system.* Yes, this makes evaluation even more difficult as the role of a vision module within a system cannot be predicted by its developer. But if algorithms contain the feature of *self-diagnostics* they enable the calling system to reasonably react on the output of a vision submodule.
3. *Vision is too complex.* Yes, most vision systems consist of many, particularly small algorithms that interact in a data-dependent manner. But modularization is a classic method in systems design; at each level of information aggregation this will allow for compensation for non-optimal decisions and should enable self-diagnosis even for a complete vision system.
4. *The used models are wrong.* Yes, models are approximations to reality. But the decisive question is not the truth of the models – all models are wrong – but their adequacy for solving a certain task. Models should just be acceptable, acceptability being specified by the user of the models and/or algorithms.
5. *Quality measures are not comparable.* Yes, many algorithms, say on edge detection or pose estimation, use their own evaluation criteria, which makes comparisons extremely difficult. But statistics provides transparent measures like variance and probability to characterize performance, which can be linked to reality by hypothesis tests.
6. *No theory is available for many algorithms.* Yes, many algorithms that have been shown to work are not based on a sound theory, or, if they are, the preconditions are not met by many data sets. But one should prefer algorithms if they have a theoretical basis, as their behaviour is then predictable. Requiring algorithms to have pre-

dictable performance and to be linkable to more complex systems stimulates theoretical research, which improves understanding of vision tools.

7. *There are too many tuning parameters.* Yes, the characterization of the the performance of algorithms, as well as their testing, grows exponentially with the number of tuning parameters. But reduction of tuning parameters can be achieved by choosing only tuning parameters with a very well defined meaning or interpreting tuning parameters in a more general framework, e.g. by observing that thresholding always can be interpreted as performing a hypothesis test, where the threshold only depends on the chosen significance level.
8. *Ground truth is too expensive.* Yes, ground truth is expensive. But only algorithms which have been extensively tested will be accepted by users. As experience in engineering disciplines has shown, they are willing to – even significantly – support testing if initial tests and an underlying theory suggest that the acquisition of ground truth is worthwhile. Modularization allows a reduction of cost as only new algorithms need to be empirically tested.
9. *Simulations can not replace experiments with real data.* Yes, only modeled effects can be captured by simulations. But simulations are the only way to compare the theory underlying an algorithm with its implementation. Though, in contrast to theoretical studies, simulations do not allow generalization of the results, they are indispensable for proving the correctness of algorithms and for determining performance values for algorithms under complex conditions.
10. *Performance characterization is not acknowledged.* Yes, testing takes time. Following a rule of thumb, the time relation *theory: implementation: testing* can be approximated by *1: 10: 100*. Together with the call for *new results* this discourages testing and enforces publication of new, unproven theories, including theories working only on one or two examples. But all the above-mentioned arguments demonstrate not only the urgent need for and the usefulness of empirical testing and of characterizing performance but also reveal the necessity to derive and adapt the theoretical basis to enable performance prediction, which definitely is part of research and therefore should be clearly acknowledged by publishers and funding agencies.

3 The papers of this issue

This issue is the first attempt by a journal to focus on the topic of performance characterization and thus is meant to transfer the discussion, already ongoing in workshops (cf. [1, 6]) to a wide audience of both developers and users of vision algorithms.

The goal of this special issue is to present the state of the art in characterizing the performance of vision algorithms at all stages of the development and use, such as the design of algorithms with a pre-specified performance, the testing of algorithms with respect to given specifications and the self-diagnosis of vision algorithms during their use in an automated process.

We encouraged authors to submit papers on the following topics:

- Theory and strategies for performance analysis of vision algorithms
- Linking analysis of vision experiments to the theory underlying the algorithms
- Characterization of the limitations of vision algorithms and/or the class of image data for which a vision algorithm is suited or not suited
- Demonstrations of the usefulness of performance characterization and/or the limitations of statistical testing in computer vision
- Modularization of vision tasks and the characterization of networks of vision algorithms

We received a total of 18 papers, from which 10 were selected by the reviewers. These papers covers most of the aspects of performance characterization which need to be studied and tackled.

3.1 Application areas

Performance characterization starts with a proper *selection of algorithms* to enable a rigorous analysis of vision modules. Obviously low- and mid-level vision processes are to be tackled first as they by far are better understood than high-level processes. The papers reflect this situation. Applications selected by the authors are line and feature detection (Sheinvald and Kiryati, Wenyin and Dori, Courtney et al.), boundary extraction (Ramesh and Haralick), texture perception (Vanrell et al.), stereo matching (Courtney et al., Cozzi et al.), estimation of geometric transformations, such as relative orientation (Torr and Zisserman), pose determination from single images (Madsen, Venetianer et al.) and rigid body transformation (Eggert et al.) but also object detection (Courtney et al.).

3.2 Steps towards performance characterization

The idea of performance characterization is to determine the dependency of the result of an algorithm on the type of the input and the control parameters. These relations may be given in the form of equations, tables or diagrams, enabling the reader or potential user of the algorithm to decide on the usefulness of the vision module in his/her own context of application. In order to accomplish this task one needs to be able (1) to specify the characteristics of the input, (2) to clearly explain the meaning of the control parameters and (3) to precisely define the used measures for characterizing the performance. The papers focus in different manners on these three tasks.

Characterizing the input: either images or geometric structures are used in all studies. The input is characterized by showing the images of the study, or sample images (Torr and Zisserman, Sheinvald and Kiryati, Vanrell et al., Madsen), or by specifying the generation process leading to the simulated data (Sheinvald and Kiryati, Eggert et al., Courtney et al., Cozzi et al., Madsen, Ramesh and Haralick). The

difficult problem of characterizing textured images is treated by Vanrell et al.

One part of specifying the input refers to the noise characteristics. The notion *noise* in the context of performance characterization is broader and refers to all types of deviations of the image/input from the model used in the algorithm. This becomes clear when considering the different types of noise, random perturbations or disturbances, addressed in the articles: besides measurement noise, usually modeled as Gaussian, and explicitly discussed in nearly all papers, we of course find models for outliers (Sheinvald and Kiryati, Ramesh and Haralick, Torr and Zisserman) but also for *background clutter*, which in the most simple case may be other objects of the same type to be detected, as in Sheinvald and Kiryati and Wenyin and Dori. The effect of image compression, which also can be taken as an induced noise process, onto the estimation of the fundamental matrix in stereo is the topic of Torr and Zisserman.

Choosing *control parameters* is a bottleneck in using vision modules and development of complex vision systems. The sensitivity of the result on the control parameters, often thresholds, is addressed in several papers (Venetianer et al., Courtney et al., Cozzi et al., Ramesh and Haralick).

Specifying appropriate *measures for characterizing performance* is a central issue. Good measures form a link between theory and applications. This requires the measures to be embedded into a broad enough theoretical framework that allows consistent reasoning about performance measures. These measures needs to be computable from experiments to allow testing. Finally, the measures need to be simple and intuitive to be acceptable to people on the 'factory floor', who may not have a theoretical background. Probability theory and statistics seem to be an widely accepted framework which is used in nearly all papers. The quality of continuous output values are characterized by their bias (Venetianer et al.), standard deviations (Eggert et al.), covariance matrices (Madsen, Torr and Zisserman) or related test statistics. The quality of classification results is characterized primarily by the mis-detection rate or the probability of false alarm (Sheinvald and Kiryati, Wenyin and Dori, Courtney et al.). More specific quality measures are proposed in order to adapt to the individual application, as the fragmentation quality in line drawing vectorization (Sheinvald and Kiryati) or the mean values for the length of line segments or gaps in boundary detection (Ramesh and Haralick).

Three of the papers explicitly address the problem of defining measures which allow an overall evaluation of a vision module (Wenyin and Dori, Courtney et al.) or the linking of several modules within a sequence of analysis steps (Courtney et al., Ramesh and Haralick).

3.3 Type of studies

When comparing the type of study we find a great variety. Following Maimone and Shafer [5], one may distinguish six steps when evaluating or characterizing performance of vision modules. All these steps are usually required when developing a vision module to be used in practice. All the steps can be found in the papers:

1. *Mathematical analysis*, e.g. analytically deriving performance measures based on a well-defined model of the algorithm. This allows well-understood performance predictions. Examples are the determination of standard deviations, or covariance matrices (Eggert et al., Madsen, Torr and Zisserman), the identification of weak or singular configurations (Madsen) and the prediction of probabilities for misdetection or false alarms (Ramesh and Haralick).
2. *Simulations using data without noise*. This allows verification of the implementation and the identification of artifacts due to finite machine precision (Eggert et al.) or model discretization. It may replace an analytical analysis (Vanrell et al., Cozzi et al.), which may not be feasible for more complex algorithms.
3. *Simulations using data with noise*. This is the classical setup for analyzing and characterizing performance of complex algorithms, as analytical tools usually are not feasible and only in this case are true reference values available. Nearly all papers use this kind of study.
4. *Empirical testing using real data with full control*. This type of analysis is necessary to prove the usefulness of the model underlying the vision module. As full control is required, the effort for this type of study is very high. Therefore this type of study is usually performed when starting to develop a new model or when developing the methodology for evaluating performance, as is the main scope in our context. Venetianer et al., Madsen and Torr and Zisserman show examples how such fully controlled tests may be performed.
5. *Empirical testing using partially controlled real data*. This type of analysis is the standard case for proving the adequacy of model, as parts of the model already have been rigorously checked in previous experiments, so only the additional modifications need to be empirically evaluated. Of course, partially controlled real data are extremely valuable to get initial statements about the performance to enable the formulation of a hypothesis about potential improvements of the algorithms (cf. e. g. the discussion by Torr and Zisserman).
6. *Empirical testing in an uncontrolled environment*. This will be final tests on the practical usefulness (Torr and Zisserman) or tests to show limitations of the procedures by counter examples.

Many authors *re-implemented* previously published algorithms to use them as testbed for demonstrating their methodology, as Vanrell et al. for texture perception, Sheinvald and Kiryati for straight line detection in binary images, or Torr and Zisserman for boundary extraction. Re-implementation only proves that the authors of the published algorithm have provided a useful documentation, which is by far not to be expected on the average, but may lead to the detection of errors in the procedures. Re-implementation in one case was the basis for comparing different algorithms with respect to their performance: Eggert et al. re-implemented four algorithms for rigid body transformation, coming up with the extremely valuable result that there is practically no difference between the algorithms.

Although most aspects of performance characterization are touched on in the papers of this issue, the reader will still

find a great variety of means to achieve this goal, especially in terms of setting up experiments, in terms of measuring performance or in terms of documenting input output relations. It will require a longer discussion within the machine vision community to agree on standards describing performance of vision algorithms. However, the approaches given in the papers are certainly an excellent motivation to critically analyze ones own research and/or criteria for reviewing and increase the awareness of the necessity to invest into research of characterizing performance of vision software.

We wish the reader an interesting time when studying this issue and encourage comments on the approaches.

Acknowledgements. We wish to thank our reviewer panel for a tremendous effort to evaluate the papers within a very limited time, enabling publication of this issue with a minimum of delay from submission to final print (less than 10 months!). The research has been supported by the European Computer Vision Network (ECVnet) sponsored by the European Commission; and by the Centre for Autonomous Systems, sponsored by the Swedish Foundation for Strategic Research.

References

1. Christensen HI, Förstner W, Madsen CB (eds) (1996) Performance characteristics of vision algorithms. ECVNET, http://www.vision.auc.dk/˜hic/perf-proc.html
2. Förstner W (1996) 10 pros and cons against performance characterization of vision algorithms. In: Christensen HI, Förstner W, Masen CB (eds) Workshop "Performance Characteristics of Vision Algorithms"
3. Haralick RM (1985) Computer vision theory: the lack thereof. In: Shapiro A, Kak L (eds) Proceedings of the Third Workshop on Computer Vision: Representation and Control, IEEE CS, pp 113–121
4. Jain TO, Binford RC (1991) Ignorance, myopia, and naiveté in computer vision systems. CVGIP: Image Understanding 53:112–117
5. Maimone WM, Shafer SA (1996) A taxonomy for stereo computer vision experiments. In: Christensen HI, Förstner W, Madsen CB (eds) Performance characteristics of vision algorithms, pp 59–80
6. Meer P, Haralick RM, Förstner W (eds) (1994) Performance versus methodology in computer vision. NSF/ARPA Workshop, IEEE Computer Society, Seattle
7. Price K (1985) I've seen your demo; so what? In: Shapiro A, Kak L (eds) Proceedings of Third Workshop on Computer Vision: Representation and Control. IEEE Computer Society, Seattle, pp 122–124