Tina Memo No. 2005-009
A short revised version can be found in CVIU, 109, 305-334, 2008.

Performance Characterisation in Computer Vision: A Guide to Best Practices.

N. A. Thacker, A. F. Clark, J. Barron, R. Beveridge, C. Clark, P. Courtney, W.R. Crum, V. Ramesh

Last updated
16 / 5 / 2005

Imaging Science and Biomedical Engineering Division,
Medical School, University of Manchester,
Stopford Building, Oxford Road,
Manchester, M13 9PT.

Performance Characterisation in Computer Vision: A Guide to Best Practices

N. A. Thacker, A. F. Clark, J. Barron, R. Beveridge,
C. Clark, P. Courtney, W.R. Crum, V. Ramesh

Preface

In September of 2003 a one week meeting of the authors [1] *of this document was held near Colchester, funded by EU IST project 1999-14159, to discuss best practices in the design and evaluation of computer vision systems. The main aim of the meeting was the generation of this document, an attempt by the authors to relate common problems and scientific issues across a broad range of topics in the field. It is our intention to publish this work in an international journal, in the meantime this working version is to be distributed from our web pages. The intended audience is primarily PhD students, though we expect that established researchers may find some of the contents interesting.*

Abstract

It is frequently remarked that designers of computer vision algorithms and systems cannot reliably predict how algorithms will respond to new problems. A variety of reasons have been given for this situation, and a variety of remedies prescribed [99, 213, 33]. Most of these involve, in some way, paying greater attention to the domain of the problem and to performing detailed empirical analysis. This document is not intended to be a a standard review article. Rather, the goal of this paper is to review only what we see as current best practices in these areas and also suggest refinements that may benefit the field of computer vision. A distinction is made between the historical emphasis on algorithmic novelty and the increasing importance of validation on particular data sets and problems.

[1] With the exception of Bill Crum who was invited to contribute later.

Contents

List of Figures

List of Tables

1 Introduction

The aim of this paper is to relate and consolidate common approaches to the task of understanding the performance characteristics of algorithms in a variety of application tasks. By performance characterisation we refer specifically to obtaining a sufficiently quantitative understanding of performance that the output data from an algorithm can be interpreted correctly. This can be understood as an entirely statistical task. This touches on other aspects of building complete systems, such as validating software though not testing hardware, though clearly this is a related task.

This paper consists of two main parts: a review of selected past work in performance characterisation as it has appeared across a range of subject areas in computer vision; and the presentation of a conceptual framework that ties these efforts together and describes the new insights gained by doing so, as well as pointing to current scientific challenges and possible new directions.

We begin by describing a framework that is based on a set of key questions aimed at revealing the current state of development of methodologies and best practices, as well as the results obtained. We review previous published work in the context of this framework, covering a number of problem domains addressed in the computer vision literature. The areas covered include both lower level (feature detection, shape description, *etc.*) and high-level visual tasks (eg detecting structural changes in medical images) but is by no means exhaustive. The intention is to demonstrate the commonality between the research questions that define scientific progress in the subject, while giving practical examples of the range of potential issues that these questions raise. We then discuss the consequences of applying the framework to the current work.

2 Background and Motivation

Discussion over the need for and role of rigorous performance evaluation of vision algorithms was raised as a specific question within the academic community from 1986 [97, 204]; again in the early 1990s [127]; continued by Haralick within DARPA and elsewhere [83, 98, 100, 101] and by Förstner [79] and was taken up within certain communities, in particular OCR [176, 177, 215], document understanding [149, 150] and graphics recognition [62, 137, 41, 266, 154, 267, 260], and photogrammetry [186, 47].

The mid to late 1990s saw the organisation of a number of workshops (ECVnet [51]; ECCV96 [114] leading to a special issue of Machine Vision and Applications [78]; DAGM'97 [77]; CVPR98 [32], a 1998 Dagstuhl workshop [136]; ICVS99 [45]; AVICS99 [25] and ECCV2000 [43], also leading to a book [44]), journal special issues ([199, 196]), as well as web resources ECVnet[2], PEIPA[3] and CMU[4]) which have evolved today to a set of well-established workshop series (eg PETS[5], Empirical Evaluation in Computer Vision[6], and NIST's PerMIS Performance Metrics for Intelligent Systems[7]), as well as a tutorial series of the PCCV project sponsored by the European Commission[8].

In the industrial arena, the initial high expectations of machine vision were not fully met and the anticipated growth of a vision industry did not materialise outside a few significant niches (notably semiconductor manufacturing and pharmaceutical manufacturing). Many potential end-users remained sceptical well into the 1990s, citing a lack of robustness [15]. Recent years have

[2] http://www-prima.inrialpes.fr/ECVNet/benchmarking.html
[3] http://peipa.essex.ac.uk/
[4] http://www-2.cs.cmu.edu/afs/cs/project/cil/ftp/html/vision.html
[5] http://visualsurveillance.org/
[6] http://www.cs.colostate.edu/eemcv2005/
[7] http://www.isd.mel.nist.gov/PerMIS_2004/
[8] http://www.tina-vision.net/

seen the increasing acceptance of vision in a wide range of specific applications where highly hand-crafted solutions have been commercialised. Our observation is that these successes have generally been developed on a one-off basis by experts in their own fields and not as a result of systematic application of results published in the computer vision literature. Researchers interested in system building and higher-order behaviours, such as robot navigation and generalised learning exhibiting graceful degradation, have also been frustrated by the brittleness of the vision modules they have to work with.

As the field of computer vision has developed, the pool of experienced vision researchers and fresh PhDs has increased. New application areas have been sought out and explored in collaboration with technical experts in those fields. The computer vision community is now interacting with a much wider scientific and funding community than ever before. Although the comparative testing of algorithms has been slow to establish itself, these wider groups are asking questions about the validity of specific tools for their given data analysis problems.

The availability of the World Wide Web as a means of exchanging and sharing common data sets and code has greatly facilitated the increase in performance comparisons in the presentation of new work, as witnessed by the increasing reference to servers where data and/or code may be downloaded. The rise of the Web and multimedia tools have also played a role in presenting new vision applications. Increased computing power has enabled the running of experiments and simulations which would not have been possible before.

Finally, the few published evaluation methodologies have started to be taken up in certain areas. There is a gradual acknowledgement that sharing code and data is important for replication as the basis of the scientific method, and that the community needs to build on the work of others for the field to advance, just as it has in other scientific disciplines.

3 A Framework for Understanding the Performance of Vision Algorithms and Systems

Whilst the subject of algorithm performance characterisation has become established as one of the set of issues to be considered in computer vision research (as evidenced by recent conference calls), there is still a lack of consensus on the appropriate questions to be asked. Phillips [197] outlined the different levels of analysis for a biometric system, and this also appears to be applicable to vision systems in general:

Technology evaluation concerns the characteristics of the technology as conditions are changed (sensor noise, contrast, model database size, *etc.*) for a range of parameter settings. Pre-recorded data sets are used, so that the tests are repeatable and can be used for comparison purposes. It is at this level that algorithm developers can determine the characteristics of their algorithms using generic metrics such as ROC (Receiver Operating Characteristic) curves. A good analogy is with measuring the properties of a transistor: it is characterised by a number of simple parameters (gain, voltage drop) which describe the relationship between inputs and outputs (voltage, current) over a range of conditions (voltage, temperature), and which is independent of the actual use in a large system (switch, amplifier, detector, filter, etc).

Scenario evaluation concerns how the system behaves for a particular application domain for a specific functionality eg recognition, verification with its set of variables (type of lighting, number of users) and how the parameters should be set to obtain the best performance.

Ramesh [213] independently summarised a system engineering methodology for building vision systems with two main steps; *component identification* and *application domain characterisation*,

which are equivalent to the *technology evaluation* and *scenario evaluation* just described.

3.1 The Role of Quantitative Statistics

Characterisation of an algorithm requires a more formal statistical approach than is conventionally attempted in computer vision. In particular, many algorithms seem to have been designed without explicitly considering the statistical character of the input data at all and this makes it very difficult to proceed to the next step of quantitative prediction of performance. It is useful to consider algorithms as estimation processes wherever possible. In general, it is possible to take any algorithm that has a defined optimisation measure and relate the computational form to likelihood. Similarly, one can take any algorithm that applies a threshold and relate it to hypothesis tests or Bayes theory. In the process of doing this, one discovers the assumptions necessary in order to achieve the association. Although the computational details of many algorithms may look very different to standard methods in use in statistics, all seemingly novel data analysis methods are ultimately reconcilable with existing ideas in quantitative statistics. In fact, this can be seen as inevitable once one accepts that probability theory is the only self-consistent form of data analysis.

Once the algorithm has been interpreted in a quantitative statistical form, it is then possible to apply the conventional methods of stability analysis and prediction in order to test these assumptions.

This in turn leads to the proposition that if the assumptions made do not match the true data distributions, it may be possible to pre-process data or re-define the problem so that they do. For example, if the data has a Poisson distribution but a Gaussian is assumed in the algorithm, then the data can be preprocessed using a square-root transform to regain Gaussian behaviour[66]. Likewise a binomial distribution may be converted into an approximation to a Gaussian by the arcsine transform, Indeed these are standard techniques already described in the statistical literature [236].

In addition, the quantitative nature of these approaches has the advantage that it is possible to make definite statements regarding the expected performance of a system using the theory upon which the method is based, and against which actual performance can be compared. This approach should perhaps be regarded as the only possible theoretical basis for computer vision algorithm design and testing. We summarise many of the methods in common use below.

A factor that makes machine vision research a particular challenge is that conventional vision/statistical techniques often need to be further developed and adapted, in order to avoid difficulties with numerical or non-linear aspects of the algorithm. Clearly simpler algorithms are the most straightforward to analyse. It then remains for these results to be validated using data for the forms of variation which most influence the performance of the method.

We suggest that the questions an algorithm developer should ask are ones that relate directly to the identification of sources of variation and the use of a quantitative methodology. Historically, and we feel erroneously, these issues have not been seen as being of primary importance in the computer vision literature. In order to give an assessment of where computer vision as a field is today, we devote a sizeable portion of this paper to reviewing the current state of empirical evaluation for a number of visual tasks.

3.2 Characterising Variation

A common observation is that algorithms perform differently when run on apparently similar data. Characterising this variation is both essential and difficult. Underlying this difficulty is the the perception that there is one unique variation associated with an algorithm. To underscore

the idea that there are different ways of formalising the concept of variation let us consider three specific examples.

The first type of variation we might wish to consider is the *repeatability of the estimation process* — that is, how much might the results from our algorithm vary if we were to acquire another, identical data set differing only due to measurement noise. For many estimation processes, this variation is already a sufficient result to allow this process to be used in the design of a larger system, by matching the characteristics of the output data to the assumed behaviour needed for the design of subsequent modules. It also tests the stability of the algorithm by allowing us to establish whether small changes in the data, at the level expected due to natural variation, result in correspondingly small changes in the output.

A second type of variation is due to *intrinsic variability in a problem description*. Using face recognition as an example, what might we expect the recognition rate to be given different images of the same 100 people under for example variations in lighting and pose? The variance due to intrinsic variability characterises how much an algorithm's behaviour will vary from one execution to another on data drawn from a population/domain (rather than sensor noise).

Another type of variation occurs in test reliability due to *sample size*. It essentially answers the question: "We have observed only n samples from a population; how sure are we about the population as a whole?" Imagine we observe that a face recognition algorithm has a recognition rate of 80% on 100 test images. What range of recognition rates would we expect on a different set of 100 images *from the same domain/population*? More importantly, do we require a sample of $1,000$ images, in order to conclude that one algorithm is better than another [95].

In fact, for any algorithm, you can choose to fix any given set of variables on the input data and vary those remaining to generate a variance estimate. Each will inform us about different aspects of algorithm performance and give corresponding insights which might lead to modification of the algorithm. Equally however, this degree of freedom can also lead to confusion as to what is the most appropriate way of testing systems and the conclusions we can expect to draw.

3.3 Black Box and White Box Testing

Most of the papers in the empirical evaluation workshops are aimed at addressing black box evaluation methodologies for vision algorithms. By this we mean that the internal workings of an algorithm are not directly considered, only the output which follows from an input. This is in contrast to Ramesh and Haralick ([208], [211]) and Ramesh et al [214] that address white box evaluation.

The white box approach to performance characterization requires abstracting the essence of the algorithm into a formal mathematical expression or transform, identifying a statistical model of the input data and deriving the output statistical model. The complexity of the process depends on the nature of the algorithm, the input data types and the output data types.

To facilitate the propagation of variation through white box models, tools used in [214] along with other numerical methods (e.g. bootstrap ([63])) perform the characterization with analytical statistical models. The tools/steps available include:

- Distribution propagation: The input to an algorithm (i.e. an estimator) is characterized by one or more random variables specifying the ideal model, its parameters, and by a noise model with a given probability density function (pdf). The output distribution is derived as a function of the tuning constants, the input model parameters, and the noise model parameters.
- Covariance propagation: The algorithm output is thought of as a non-linear function of the input data and noise model parameters. Linearization is used to propagate the uncertainty

in the input to the output uncertainty. Care should be taken while using this tool since the approximation may be only good when the linearization and first order error approximations are valid. Further details are presented in [103].

- Empirical methods: Courtney et al [55] describe an empirical approach to systems evaluation. Statistical resampling techniques (e.g. bootstrap) are used to characterize the behavior of the estimator. For example, the bias and uncertainty can be calculated numerically (see Cho and Meer [42] for the first description of edge detection performance characterization using bootstrap). Monte-Carlo methods are used for the verification of theoretical results derived in the previous two steps.
- Statistical modeling: Modeling at the level of sensor errors (Gaussian or other perturbations in input), prior models for 3D geometry, spatial distribution of objects, and modeling of physical properties (e.g. constraints on the types of light sources in the scene), etc. The related literature is rather vast and encompasses methods from fields such as Bayesian statistics, Spatial statistics, and Computer Vision and we shall not discuss this further here.

3.4 Assessing Progress

The purpose of this paper is to try to identify what might be considered best research practices in the field. In order to do this we must establish a criterion for what constitutes scientific progress.

We go back here to the definition used to assess the contents of a paper or a thesis, that the work must make a contribution to knowledge. Based upon this, it is our belief that where possible, research papers should contain results which are as independent as possible of the details of the particular analysis method selected. As explained previously, this raises an immediate challenge in a field with so many particular application areas. In particular performance figures for any algorithm can vary according to the dataset tested. Thus quantitative results are often less useful for future researchers than some might think they should be. One way to avoid this problem is to develop an understanding of how well an algorithm is expected to work, in order to confirm that a given implementation really conforms to expected behaviour. We need to ask whether we know enough about the problem to be able to recognise the right answer when we get it. Sometimes preconceptions of what we expect to see as results can take the place of a more formal definition of what is actually required or intended. This is a particular issue when our preconceptions of what we would like are actually impossible given the information contained in the data available.

Another way to finesse these problems is to define and work with common data sets, but this is accepted to be very difficult and can only work up to a point. In cases where we cannot define a single test data set, then we need an alternative method which will put the results from an algorithm into the context of other approaches, ie: a comparison, and for this we need some agreement within the field for which algorithms should be used for comparative purposes. The specific details regarding which algorithms are selected for these purposes, and why, might be less important than an acceptance that such a mechanism is necessary and a basic agreement of which would be acceptable.

Developing a scientific understanding of performance requires us to quantitatively understand how our correct answers would be expected to match those delivered from the algorithm under real world circumstances (such as noise). This level of understanding begins to make possible the use of the algorithms in larger systems. Such exploitation of methods represents another form of direct contribution to scientific knowledge. Assuming that all algorithms should be ultimately reconcilable with probability theory, we need to ask whether the body of theory is present which would allow the assumptions, required to derive a particular analysis approach, to be identified. This gives us the theoretical underpinnings of the subject, so that we know not only how to approach solving a problem but why that approach is fundamentally correct for a specific definition of the

task. On the basis of all of these levels of understanding, we can finally make recommendations regarding the methods available for the design and testing of specific algorithms. Making sure that we match the definition of what was required to how output data is intended to be used.

In this paper we attempt to take the issue of technology evaluation a step further by defining the following KEY QUESTIONS to highlight the current state of development of methodologies and best practices.

- *How is testing currently performed?*
- *Is there a data set for which the correct answers are known?*
- *Are there datasets in common use?*
- *Are there experiments which show algorithms are stable and work as expected?*
- *Are there any strawman algorithms?*
- *What code and data are available?*
- *Is there a quantitative methodology for the design of algorithms?*
- *What should we be measuring to quantify performance?* What metrics are used?

The idea here is to encapsulate what may be seen as genuine scientific progress, culminating with the identification of a theory of image data analysis. One aspect of our approach is to try to define what might be appropriate metrics by appealing to the statistical definition of the task. In doing so we will be asserting that all algorithm constructs should come from a probabilistic statement for the solution of the problem, such as: the probability of a particular interpretation of the image data given a set of assumptions regarding expected data distributions. In particular, a good test metric describes concepts such as detection reliability (which can be derived from the concept of a statistical hypothesis test), estimates of parameter covariance (derived from statistical definitions of likelihood) and classification error (which have origins in Bayes error rates). In all cases, the technology evaluation stage requires simple metrics related to the function (estimation, detection, classification), whereas scenario evaluation [197] makes use of more complex metrics relating to the task (such as system reliability rate expressed as mean time between failure).

4 Review of Performance Analysis Work in Specific Classes of Algorithms within Computer Vision

4.1 Sensor Characterisation

Although much of computer vision employs straightforward colour or monochrome imagery, the range of sensors that can be employed for particular applications is vast — so much so that it is impractical to provide a single model that will describe any sensor. Once again, this situation can lead to confusion, particularly if there is an expectation that algorithms should work on any image regardless of the assumptions needed to relate the method to a principled statistical method. A safer and a more scientific approach would lead us to always assume that methods will not work on any other sensor than that for which it was originally developed and validated, until the work was done to identify the underlying assumptions and show that these still hold for the new data set.

Perhaps the earliest step in describing the general characteristics of a sensor is to consider the illumination regime in which it operates. Illumination may be considered in three basic forms: incoherent, coherent and partially-coherent. Sensors such as cameras (film, CCD, videcon, X-ray,

etc.) working with reflected or re-emitted illumination from conventional sources operate in an incoherent regime. Illumination by stimulated emission sources such as laser stripers, ultrasound systems and radar (including synthetic aperture radar) result in coherent imaging.

The important distinction, at least to a first approximation, concerns the underlying statistical distribution of random noise. For incoherent illumination, where the arrival of any photon is an independent event, Poisson statistics apply. As the Poisson distribution can be approximated well by a Gaussian for large mean (large numbers of photons per pixel), this gives rise to the model of spatially-independent Gaussian noise commonly used throughout image processing and analysis. With coherent illumination, however, the arrival of photons is not independent and, consequently, Poisson statistics do not apply; hence the Gaussian noise distribution used in incoherent imaging is not appropriate.

Partially-coherent imaging occurs for example in devices such as high-resolution electron microscopes and those based on low cost semiconductor laser. The choice of appropriate noise model depends to a great extent on the degree of coherence present.

How is testing currently performed? Sensor specifications are often characterised in terms of photometric quality by their signal-to-noise ratio. This measure, however, is not usually directly applicable to image processing and analysis as it is a measurement of how well the sensor works rather than the information content it produces. It may confound issues of intrinsic detector noise, electronic noise and digitisation precision. A sensor will also suffer from geometric error, including lens distortion, sensor deformation, etc.. Such issues have been extensively studied by the photogrammetric community for visible and IR detectors [186, 252] and occasionally the computer vision community [165, 110, 254], but do not appear to form the basis of the majority of algorithm development work reported in the literature.

Furthermore, systems that involve processing as an inherent part of capture will exhibit specific noise characteristics that differ from those described above; for example, images from MRI scanners exhibit a Rician noise distribution [91]. Likewise, sensors intended for visual presentation of information may employ histogram equalisation which introduces discontinuities into the data distribution and may cause failures in the downstream algorithms if not accounted for.

Are there experiments that show subsystems are stable and work as expected? Different sensor types have vastly different noise characteristics. The various noise distributions that apply in different illumination regimes can be measured from imagery.

The experiments which demonstrate that the sensors have the expected characteristics are those of repeated acquisition of a static scene. The expected noise characteristics can be analysed by looking at differences between images of identical targets.

Many types of imagery exhibit *systematic* errors: for example, synthetic aperture radar imagery may suffer from "blocking" and other artefacts due to the way in which the raw data returns are processed. Non-stationary noise sources and longer term thermal effects in CCD cameras were described in [20, 47] and require careful experimentation.

Is there a quantitative methodology for the design of subsystems? It is absolutely critical to understand the imaging process of the sensor subsystem to develop processing algorithms that stand any chance of working robustly. As an example, conventional edge detectors, which implicitly assume additive noise, will not work well on images where the noise is high and multiplicative, such as radar and low light cameras.

What should we be measuring to quantify performance? A parameterised model of sensor stability and repeatability experiments are adequate.

4.2 Lossy Image and Video Compression

How is testing currently performed? The most common quantitative methodology is mean-square error or similar between the original image and the decompressed image after compression; but this is acknowledged as being significantly inferior to subjective testing, as the ultimate target is the human vision system. There are standardised ways of carrying out and analysing subjective tests based around the use of forced choices in comparisons of images.

Is there a data set for which the correct answers are known? Coding is perhaps unusual in that any image or sequence can be used: the original data are always the correct answer.

Are there data sets in common use? There are several widely-used images (*e.g.,* Lenna)(see [121] for an interesting discussion) and sequences (*e.g.,* salesman) used for basic comparison. Each coding technique is developed in conjunction with a standard set of images or image sequences. Practically all researchers working on the coding technique use the data set. Some images or sequences are chosen because they have specific properties that are likely to be problematic; and others are chosen because they are typical of particular types of common imagery or video.

Are there experiments that show algorithms are stable and work as expected? There *are* such experiments but it is unlikely that they prove stability in any statistical sense.

Are there any strawman algorithms? Yes. Coding is almost unique in image processing in that significant effort has been (and continues to be) put into developing international standards: the ITU H.2*nn* series and the JPEG and MPEG series, for example. The development of data coding techniques involve the development and refinement of algorithms which are distributed amongst the community.

Is there a quantitative methodology for the design of algorithms? No. Although there is sound theory in the form of "information metrics" for data compression in general, these measures are not expected to be relevant to lossy compression of images. The algorithms used make assumptions about the nature of images, such as an expectation of exponential residuals in block-matching for motion compensation in video coding. Although this appears to work well in practice, this has not been formally tested.

What should we be measuring to quantify performance? The human visual system is the ultimate target for all image and video coding, so that is what *should* be measured. The various quantitative measures in common use must be considered as approximations to that. Ideally, in order to eliminate subjective variability, we would like to have access to a realistic computational model of the human vision system, or at least the early parts of the visual pathway. This is clearly a significant challenge.

4.3 Feature Detection

This section covers primarily edge and corner features, though the conclusions of this analysis could be applied to the extraction of larger scale image structures such as arc and circles [222,

220] [266, 260].

A great many papers over the years have evaluated novel feature detectors by producing feature extraction (or enhancement) images. Unfortunately, this has often been based on unique data sets and offers no quantitative performance estimate. Canny famously provided a theoretical derivation based upon a particular choice of "optimality" [39], though the predictions he makes for (location) performance are not observed in Monte-Carlo studies [60]. This is a key point, as if we are to use definitions of optimality to design algorithms these measures are not meaningful unless they have quantitative agreement with actual performance.

Early proposals for interest operators included variants of auto and cross-correlation such as operators by Morovec [174] and Harris [106]. A cursory demonstration of function was given by a small number of images, and by their inclusion in robotic systems (DROID [38]). Scale-independent extensions proposed [234, 233] were similarly demonstrated on single images as well as artificial scenes, demonstrating the limitations of the original formulations and providing visual evidence that these had been overcome. More recently, the emergence of a new family of operators with affine and/or scale invariance has started to appear (steerable filters [82], moment invariants [143], SIFT [156]).

How is testing currently performed? **Edge and line detection** Numerous papers have appeared in the literature on boundary segmentation performance evaluation. Some of the early papers include Ramesh and Haralick [208], [210], [212], Wang and Binford [263].

The first set of papers evaluate edge parameter estimation errors in terms of the probability of false alarm and mis-detection as a function of the gradient threshold. In addition, the edge location uncertainty and the orientation estimate distribution is derived to illustrate that at low signal to noise ratios the orientation estimate has large uncertainty. Heath [111, 112] visually compares the outputs from various edge detectors. [42] was the first to use a resampling techniques (*e.g.,* bootstrap) as a tool for studying the performance of edge detection techniques.

Most of the papers described above use simulations and/ or hand-drawn ground truth to compare algorithm results with ground truth results. Baker and Nayar [9] is unusual in that it does not require ground truth. The evaluation is made by examining statistics that measure global coherence of detected edge points (e.g. collinearity etc.). Konishi[138] addresses edge detector evaluation by using information theoretic principles. More specifically, they use estimates of the Chernoff bound to compare various multi-scale edge detection filters.

Interest operators behaving as local image descriptors (sometimes called 'corners') have been sought to overcome the aperture problem inherent with line/edge descriptors for a range of applications (matching for motion and stereo estimation, object recognition). Anecdotal evidence of poor reliability was explored by studies on synthetic images which revealed non linear dependence on image contrast, noise and confounding edge features [53, 54]. By 2000 more rigorous evaluations were being reported [89, 228, 170] with detection rate against false positive rate (ROC curves) being calculated for a set of viewpoint changes across a database of images.

Is there a data set for which the correct answers are known? There is no mutually agreed data set with reliable ground truth. It would be very difficult to design a single data set for every possible feature detector, though an appropriate simulated test data set could be constructed from an appropriate theoretical definition of the detector. The RADIUS data set with ground truth [250] has served in some studies [221].

Are there data sets in common use? Earlier papers demonstrated effects on the Lena/Lenna or the Cameraman image. Image data sets are now made available on the web [112]. The latest

interest operator study [170] utilises shared code made available by the original developers. A database of 1000 images taken from a single 3 hour video of a planar scene undergoing viewpoint change was created, yielding some 300,000 points and subsequently made available on the web. The study was sufficient to consistently rank the operators but the statistical properties of the data and the relative contribution of various error sources is not discussed further.

Are there experiments that show algorithms are stable and work as expected? Stability has been explored in terms of some image properties but has not been related to any theoretical prediction, such as error propagation (except Haralick [102]). For well defined structure, Monte-Carlo or resampling techniques could be used, if an appropriate definition of the task was available.

Are there any strawman algorithms? For edge detection the commonly quoted algorithm is Canny, though implementations vary. Furthermore, the original Canny work included a scale-space analysis, though under most circumstances only the edges at the highest resolution scale are of quantitative use for measurement. In addition, use of the scale-space analysis also presupposes that the data will be used for the same tasks as in Canny's original work, which differs from tasks such as simple measurement.

Strawman edge detection code, data and scoring tools are available [31] [9] [10]. For corner detection we have found that the Harris and Stephens algorithm works well [107] and is available[11]. The SUSAN code is available[12]. As mentioned above, corner detection software is now being shared between research groups [170].

Is there a quantitative methodology for the design of algorithms? All feature detection algorithms, based upon a single process to enhance the feature and then thresholding, should be interpreted as a hypothesis test. Generally, the hypothesis under test is that there is no feature present and the results of the feature enhancement stage can be accounted for entirely by noise. This requires explicit account of the image formation process to be made. Here statistical limits are often replaced by equivalent empirical thresholds. Alternative methods, based upon the Bayesian approach which requires the identification of both background and signal distributions, are not expected to be strictly quantitative without considerable care.

What should we be measuring to quantify performance? Following from the previous statistical interpretation, we should evaluate the probability of false alarm and mis-detection as a function of threshold as in [209]. Algorithms can also be compared using ROC or FROC (fractional receiver operating characteristic) curves. Ideally, any specification for a feature detector should come complete with quantitative measures of estimation performance, such as orientation and location accuracy as a function of image noise. Some uses of the data also require strict geometrical interpretations of the located features (such as vertices).

4.4 Shape- and Grey-Level-Based Object Localisation

The localisation and recognition of a known object in an image are tasks that have received a large amount of attention. The two tasks are often linked but we shall try to consider them separately, since from a performance evaluation point of view, different information is sought, with different failure modes and performance metrics. Localisation involves estimation of the transformation

[9] marathon.csee.usf.edu/edge/edge_detection.html
[10] figment.csee.usf.edu/edge/roc
[11] www.tina-vision.net
[12] www.fmrib.ox.ac.uk/~steve/susan

between the image coordinate frame and the object coordinate frame. This may range from a simple translation vector to full six degree of freedom Rt transformations, according to the nature of the object and the scene. Our remarks is this section are therefore also relevant to tasks such as image alignment or registration [74].

A great many techniques have been proposed, from template matching, Hough transform, 3D wire-frame location, to techniques for deformable objects such as Snake models [134], Active Shape Models (ASM) and Active Appearance Models (AAM) [50], where a low-parameter model of an object is aligned over a corresponding image region via the minimisation of some cost function. Localisation techniques often comprise a representation stage (operating on pixels or derived primitives)[13], and a matching stage based on a cost function.

How is testing currently performed? Novel localisation techniques are generally demonstrated on marked-up data and with localisation error expressed in terms of image-plane error. The main issues affecting localisation performance include sensor noise (resulting in imprecise image plane feature location), occlusion (missing features) and clutter (spurious non-object features). Lindenbaum [152, 153] examined both localisation and indexing performance to model the effect of these three factors, and added object model self-similarity to provide estimates of performance bounds.

In [256] a black box comparison of five object localisation techniques was carried out on an extended data set of an integrated circuit, to determine accuracy in translation and orientation, and robustness to varying degrees of occlusion, clutter, and changes in illumination. The techniques used were based around grey level and edge intensity information, including template matching and Hough transform, from a commercial imaging library and shared code.

The precision of 3D pose estimates also appears to be be strongly dependent on object geometry and viewpoint. In [159, 160] a large (more than 100:1) variation in stability was found across the range of viewpoint and geometry of 3D polygonal objects using an error propagation approach. In addition, camera parameters have an influence as studied in [140, 141, 142].

Haralick studied a number of formulations for 3-point pose estimation and revealed wide variation in localisation accuracy according to the order in which the calculations are performed [104]. A study of four alternative optimisation techniques on synthetic data suggests that this choice does not appear to play a significant role [155, 64].

Localisation plays a role in mobile robotics [258, 132]. Performance of closed loop pose estimation and tracking of 3D polyhedral objects was reported in [52].

In one co-operative study of co-registration, blinded ground-truthed image sets were distributed [74] thought this raised a number of technical and organisational issues [86].

Is there a data set for which the correct answers are known? Image sets with known transformations may be generated either from known mechanical motion [256] or reference marks [74]. Errors in these independent estimates are not given. Image synthesis tools have also been popular for providing synthetic images and ground truth [104, 155, 64, 52].

Are there data sets in common use? For some tasks, such as face localisation [271], and graphics recognition [267, 150] researchers use existing data sets. For other tasks, data sets tend to be specifically created [160, 74, 256]. The MPEG-7 community have provided a shape dataset which has been used in some studies, though others have questioned the quality [145, 268].

[13] See [275] for a recent review of shape representations schemes

Are there experiments that show algorithms are stable and work as expected? The choice of representation scheme and cost function encode key assumptions about the data and task which may be tested. The establishment of upper and lower performance bounds permits empirical validation of this performance and thus the validation of assumptions about the properties of the data.

There appears to be no published quantitative methodology applied to AAM and ASM work, though in principle error propagation could be applied and confirmed using repeatability experiments. Error propagation has been applied to wire-frame approaches for location of 3D rigid objects, for both stereo geometry and projected image models, formulated either as an optimisation or as a Hough transform shared good agreement between predicted and observed object location accuracy (covariance) [5].

Are there any strawman algorithms? Template-based and Fourier techniques are well supported in the many public domain and commercial libraries, albeit with substantial variation in implementation details. There are many variants of ASM and AAM algorithms in use across the community, including a publicly available source at DTU[14], but there does not appear to be any code in common use. There are isolated instances of code sharing such as reported in [256].

Is there a quantitative methodology for the design of algorithms? Yes. In principle all such techniques are directly reconcilable with likelihood though the assumptions regarding distributions and resulting accuracy of localisation are not tested. Indeed researchers appear to working with similarity measures (for example "mutual information") which have only recently been explicitly reconciled with corresponding quantitative statistical assumptions [34], thus otherwise impeding progress in this area.

What should we be measuring to quantify performance? Since localisation is essentially an estimation task, algorithms should provide quantitative estimates of location and shape parameters complete with error covariances. Major comparative tests such as [74] have limited predictive power without accompanying confidence measures (but see [56]). These should be confirmed on test datasets by comparing predicted error distributions with practical performance. This would produce systems which were capable of providing all necessary salient information for use in a larger system. Use of the shape parameters in object indexing (recognition) is covered in the next section.

4.5 Shape-Based Object Indexing (*i.e.* Recognition)

As with localisation, a similar range of model-free or model-based techniques have been proposed for indexing. These often comprise a representation stage (parametric and non-parametric, operating on pixels or derived primitives), and a matching (correspondence) stage based on a cost function. The choice of representation scheme and cost function encode key assumptions about the data and task. These may be treated separately (white box) or together (black box). The problem of identifying the best correspondence is something that needs to be done by applying the appropriate similarity measure to the chosen representation.

The concept of *scope* is important in this context. This is the theoretical class of problem — degree of allowable clutter, occlusion, etc.). This determines the appropriate representation, e.g. curves vs. histograms. As a consequence some algorithms have obvious limitations of scope that make them impractical for many classes of scenes. In particular, Fourier descriptors require

[14] www.imm.dtu.dk/~aam

complete *a priori* segmentation of the curve, while moment descriptors require scene segmentation — both in the absence of knowledge of the shapes that are expected. The method of geometric histograms has scope to work on cluttered scenes of disjoint edge features while also taking account of expected feature localisation uncertainty[247]. The alternative is to try to select more stable feature measures. For example [142] suggested that edge segment orientation was more stable than length or mid-point. Similarly, the match criterion may be designed according to object number and similarity, and type of allowable variation or articulation.

Some workers have proposed invariance measures such as the cross-ratio as a potential representation for the recognition task. However, as Maybank [168, 167] pointed out, under realistic conditions of measurement noise, the indexing capacity is likely to be limited to some 20 objects, although a more recent paper with an alternate formulation of the feature point PDFs suggests more optimistic results [122].

How is testing currently performed? Testing is normally performed on a number of images and a small set of objects, in terms of true detection rates, or confusion matrices. Early work on performance issues involved the verification of matches [148, 262] and the sensitivity of the hashing [261]. They studied the false alarm characteristics of the recognition technique when a spatially random clutter model is assumed with a given density and a bounded error model is assumed for the object feature points that are detected. The analysis provides a mechanism to automatically set up the recognition threshold so that a given false alarm rate can be met by the system. This was extended in [2] to include occlusion. Bounds on indexing performance were established in [152, 153] in the presence of uncertainty, occlusion and clutter.

Sarachik [225] studied the effect of a Gaussian noise model comprising occlusion, clutter and sensor error on planar object recognition, deriving pdfs for correct and incorrect hypotheses to present true positive and false positive metrics in the form of an ROC curve. From this work it appeared that uniform clutter models under-estimate error rates but it was shown that that detection thresholds can advantageously be set according to estimates of feature density to minimise error rates, that is to say, feature-dense regions require more evidence to attain the same confidence level. Knowledge of the object database allows performance to be optimised further.

Shin [229] studied object recognition system performance as a function of the edge operator chosen at the first step of the recognition system. Their conclusion was that the Canny edge detector was superior to the others compared.

Further evaluation of recognition of articulated and occluded objects using invariance features applied to SAR data was carried out in [129] using a simulator to produce ROC curves for varying orientation angle. Boshra [29] predicted performance bounds using synthetic and real SAR data from the public MSTAR SAR data set [244].

Is there a data set for which the correct answers are known? The correct answer here is generally knowledge of scene contents, though unfortunately pixel-labelled segmentation does not have a unique answer due to the variety of objects and sub-components which may require identification. The definition depends on the intended task (scenario), however a number of labelled data sets have been collected and made available [ref].

One popular source is images is the Corel image dataset. Although very large (1M images) and popular with the information retrieval community, there appear to be a number of drawbacks. The images are well framed and of good contrast, unrealistically so compared to what would be expected from a roaming camera. In addition, the labels are too abstract in nature for use in a recognition task.

An initiative by Microsoft proposed a selection of 10,000 images from the Corel dataset labelled

with 100 low-level categories and new metrics [268]. Similarly the UCID subset has been proposed[15].

A number of alternative labelled datasets of everyday objects have been made available: COIL-100 from Columbia University, a data set from Washington University[16], and the SOIL-47 Surrey Object Image colour dataset [**?**]. However, few if any outside groups appear to have published work based on these data sets.

Are there data sets in common use? Although some large data sets are available, researchers tend to select arbitrary subsets, or generate their own data [225, 229] thus precluding the possibility of results that will transfer to other domains or for comparison with other algorithms.

Whilst large data sets are used eg CMU face detection [223] it is not clear that they span the space of possible data densely enough to permit good estimation of false detection rates.

Are there experiments that show algorithms are stable and work as expected? Stability of algorithms can be assessed by repeated acquisition of images, and by acquiring data under multiple conditions, such as orientation, lighting and occlusion. Image synthesis tools are useful in this respect.

Are there any strawman algorithms? Fourier descriptors of curvature and moment analysis are common in the literature and available is many libraries. Various ASM and AAM algorithms are in use and there is very limited code sharing. Geometric histogram code is also available[17].

Is there a quantitative methodology for the design of algorithms? Yes. For Fourier descriptions the process of curve-fitting is a well-understood technique as it is common to many scientific fields. This can be used as the basis of a likelihood-based shape similarity measure. Likewise, the Hough transform has been well studied as a tool for recognition [205].

In the case of 2D sample histograms this can be done using Cosine measures which can be related to Poisson statistics for large histograms [245]. This can be combined with proofs of theoretical completeness of representation (*i.e.,* no data are lost) to make statements regarding use of information [247]. Detailed work has also been done to investigate the stability of geometric invariance based indexing schemes [168, 167].

What should we be measuring to quantify performance? Recognition performance for individual models may be described using a true detection and false detection metric[18]. Whilst the first is common, the latter is much rarer (see above). For a library of models, the confusion matrix is more appropriate, together with a rejection rate if a non-library object is allowed. Within a particular application, object-dependent data distributions, the prior probabilities of individual objects, and differing mis-recognition are likely to make the performance an issue of *scenario evaluation*. We consider the case of faces and other biometrics, as well as articulated and motion tracking in separate sections. However, we can say this here; it is impossible to define one single data set which will be applicable to all applications. In addition, performance figures for individual applications would appear to be of limited scientific value, in the sense that the results do not transfer to other data sets. It would therefore seem crucial that these areas develop a

[15] http://vision.doc.ntu.ac.uk/datasets/UCID/ucid.html

[16] www.cs.washington.edu/research/imagedatabase/groundtruth

[17] www.tina-vision.net

[18] Terminology from the information retrieval area is sometime used: 'recall' (ratio of number of similar shapes retrieved : total number of similar shapes in database, equivalent to 1-false rate) and 'precision' (ratio of number of similar shapes retrieved : total number retrieved, equivalent to true rate).

theoretical understanding of the behaviour of algorithms which is sufficient to allow approaches to be compared both at the level of the assumptions made and the effects these assumptions have on performance for characteristic data types.

4.6 Differential Optical Flow

This section dicuses the the measurement of 2D and 3D optical flow as measured from 1^{st} and 2^{nd} intensity derivatives. In 2D, optical flow is an approximation to the local 2D image motion (the 2D velocity of pixels in units of pixels/frame). 3D optical flow, an approximation to 3D voxel motion, can be measured either volumetrically or on a surface. 3D volumetric flow is what one would normally consider as 3D optical flow: we compute the motion of each voxel in the datasets (in units of voxels per volume). 3D surface optical flow is computed with respect to a moving surface via derivatives of the surface's depth values, Z_X, Z_Y and Z_t, as computed from depth values, Z, measured, for example, by a range sensor in millimeters/second. Surface optical flow is often referred to as 3D range flow [241]. Some range sensors also provide an intensity image and in some situations the fusion of depth and image data allows a 3D motion calculation where range flow or optical flow alone cannot [241]. Other types of 3D data include radial velocity from Doppler storm weather datasets [243], gated MRI data which we can use to compute 3D volumetric optical flow [11] and multi-view video silhouette sequences [249].

The 2D/3D derivatives are usually computed by repeated application of lowpass and highpass filters, for example the filters proposed by Simoncelli [231]. Thus the computation of differential optical flow is, essentially, a two-step procedure:

1. measure the spatio-temporal intensity derivatives (which is equivalent to measuring the velocities normal to the local intensity structures) and
2. integrate normal velocities into full velocities, for example, either locally via a least squares calculation [157, 11] or globally via a regularization [119, 11].

Such algorithms are generally designed to work on a specific form of image. There can be no occlusion (one object moving in front of/or behind another object), again unless this is modelled for. The images should be "textured" in some way so that derivatives can be computed. For example, no optical flow can be computed for a rotating textureless sphere. The lighting must be uniform or changing in a known way [185, 276, 181, 109], so that intensity derivatives must be due to scene motion only and not to illumination or other changes. Similarly we assume there are no specularities in the scene (otherwise the light source(s) and sensor(s) positions would have to be explicitly modelled).

Finally, all objects in the scene are rigid, no shape changes allowed. This assumption is often relaxed to local rigidity. This assumption assures that optical flow actually captures real motions in a scene rather than expansions, contractions, deformations and/or shears of various scene objects.

How is Testing Currently Performed? Optical flow can be evaluated either qualitatively and/or quantitatively. Quantitatively, error can be measured as average error in magnitude or direction [241] or an angle error measure capturing both the magnitude and direction deviation from ground truth [75]. Qualitative flow evaluation is only useful for general judgements and as a proceed/don't proceed measure. It is usually performed in the absence of ground truth and does not produce results which can be compared in any meaningful way to other work. Synthetic or real data must be accompanied by ground truth to perform quantitative error analysis.

Are there Image Sequences for which the Correct Answers are Known? We can compute optical flow for a real image sequence made using an optical bench with known camera motion and scene structure and use it in a motion and structure calculation [12]. Similarly, one can measure optical flow for synthetic image sequences where true dense depth is known and perform a quantitative error analysis on the computed depth maps [251]. One example of 2D and 3D synthetic data is sinusoidal images/volumes (which are perfectly differentiable) and thus might be consider **gold standard** data. 2D/3D sinusoidal image sequences were used in [13, 11]. Tagged MRI data may supply accurate 3D motion information for 3D MRI analysis but is not yet commonly available.

Are there datasets in Common Use? Barron et al. [13] performed a quantitative analysis for 9 optical flow algorithms using the translating/ diverging tree sequence made by David Fleet [75], the Yosemite fly-through sequence made by Lynn Quam at SRI and a number of translating sinusoidal/square image sequences. Otte and Nagel [191] have made a calibrated real image sequence. These data have known ground truth and are publicly available[19]. Though these datasets have been available for a considerable length of time it could probably not be said that these datasets are accepted as any sort of de facto standard.

Are there experiments which show that the algorithms are stable and work as expected? A number of researchers have derived covariances for optic flow estimates (see for example, [232, 49, 58]). Simoncelli used a Bayesian framework assuming input Gaussian errors for the image values and multi-scale Gaussian priors for optic flow estimates and computed the optic flow estimates along with uncertainties. Comaniciu and his colleagues [58] also use a multi-scale framework and estimates the uncertainties for the optic flow estimate by utilizing the variable bandwidth meanshift estimation framework. The main significance of their work is that a non-parametric density representation for the local optic flow distribution allows for multiple motion regions in a local patch. The mode estimate of the density function, and the covariance around that mode obtained via variable-bandwidth meanshift filter is used as the final refined estimate for optic flow. Other researchers have performed covariance propagation for optic flow [102, 164]. Here we are interested in the propagation of covariance matrices for random input perturbations to the covariances associated with final computed results, in this case, optical flow. There may be an inherent bias in many optical flow estimators because of the fact that the regularization assumption (e.g. the Horn and Schunck smoothness assumption in the flow field [119]) is not necessarily correct with all datasets. In other words, the true underlying smoothness constraint is an unknown and the estimation framework naturally has biases. More recent work was done by Fermüller and Aloimonos and colleagues [69, 72, 71, 70] and seeks to explain perceptual illusions through the estimation framework bias.

Ye and Haralick [272, 273] propose a 2 stage optical flow algorithm, using 'least trimmed squares' followed by weighted least square estimators. The 1^{st} stage takes into account poor derivative quality. Nestares et al. [182] use an estimate of optical flow and its covariance at each pixel [183] in a likelihood function framework to extract confidence measures of the translational sensor parameters.

Are there any Strawman Algorithms? Some of the "old" optical flow algorithms are still pretty good. Of course, there are now better algorithms, but algorithms such as Lucas and Kanade [157] and Horn and Schunck [119], and, to a lesser extent, Nagel [175] and Uras et al. [257] are pretty good, readily accessible to researchers and the code is available [13][20]. Lots of new algorithms

[19] anonymous ftp to ftp.csd.uwo.ca, cd to pub/vision and http://i21www.ira.uka.de/image_sequences.

[20] The algorithms of Nagel and Uras et al. use 2^{nd} order intensity derivatives which are often difficult to measure accurately. Indeed, Horn and Schunck's use of 1^{st} order intensity derivative is effectively a 2^{nd} order method because their smoothness constraint uses derivatives of image velocities, themselves constrained by 1^{st} order intensity derivatives.

have appeared in the literature (one only has to scan the main computer vision journals and conferences since 1994) and all claim to give better optical flow results compared to those in [13]. Still, often the results are only marginally better and the codes are not generally available. Some of these classical 2D algorithms also allow simple extensions into 3D. For example, 3D optical flow on gated MRI data has been computed using simple extensions of 2D Lucas and Kanade and 2D Horn and Schunck [11].

Is there a Quantitative Methodology for the Design of Algorithms? Every algorithm has some assumptions from which the method could be derived using quantitative statistics (generally likelihood). The task therefore falls into the category of a constrained estimation problem. Until now most of the following assumptions have been implicitly made for differential optical flow.

1. The data has to be appropriately sampled to avoid aliasing, i.e. the Nyquist sampling conditions are satisfied (no aliasing). In other words, the data must allow the calculation of good derivatives! For large motions, hierarchical structures, such as a Gaussian pyramid [18] may allow differentiation. An implementation of Bergen et al.'s hierarchical approach is described in [14].
2. We assume the input noise in the images is mean zero i.i.d. Gaussian error, $N(0, \sigma^2)$. Most but not all sensors satisfy this assumption. Algorithms using least squares/total least squares then give an optical solution.
3. We assume local translation. If the sensor motion has a rotational component we assume that it can be approximated by a number of small local translations.
4. For 1^{st} order derivatives the data should fit a straight line: the deviation from a straight line (a residual) could be used as a measure of the "goodness" of a derivative and these goodness values could be use in the subsequent optical flow calculation. Spies [239] showed that the sum of the normalized squared errors follows a χ^2 distribution [240].

What Should We be Measuring to Quantify Performance? The information available in a pair of images of the same scene are not sufficient to unambiguously determine point to point correspondences with uniform accuracy at all locations [259]. Yet the common interpretation of the definition of such tasks is for the delivery of dense data. Optical flow algorithms should provide not only an estimate of flow but also a measure of how good the flow is. An optical flow field with no error estimates cannot be confidently used to provide input to other applications. Covariance matrices are one good candidate for such a confidence measure and Haralick's propagation framework is a good way to integrate such information in a larger vision system.

There are also a number of approaches which allow us to address the fundamental problem of constructing a gold standard for quantitative testing.

- We can use reconstruction error: with the computed flow and the current image, generate the image at the next time and compare that constructed image to the next image [151]. If the metric adopted is small (in comparison to the expected errors) then both the optical flow and the reconstruction method are good, otherwise you probably don't know with certainity which/or both is not working.
- Given good flow, it should correctly predict a future event. For example, the velocity of a Doppler storm represented as a 3D ellipsoid should be able to predict the storm(s) in the next image. We can compute the intersection of predicted and actual storm ellipsoids in the next image [243] as a measure of the 3D velocity accuracy. Quantitative comparison requires the error covariances.

4.7 Stereo Vision

Stereo reconstruction is a well established topic and has given rise to a range of algorithms, generally based around features [67, 108, 201] or area-based [8, 10] measures, which are often associated with sparse and dense stereo, respectively. Dense depth estimation methods require supplementary interpolation schemes to generate complete depth maps from data with low information content regions.

How is testing currently performed? The two approaches have have been the subject of direct comparison papers [61, 68, 144] as well as coordinated competitions [92, 93, 206] which have shown that they give generally good performance for well-behaved scenes with sufficient information content.

In a typical recent paper [117] uses a standard ground-truthed image pair to validate various improvements to a reimplementation of a basic algorithm using a small number of metrics defined specifically for the study. Although performance improvements are demonstrated, thereby giving some indication of the value of the idea, such work provides little opportunity to understand the underlying statistics of the input data or results, nor to understand the validity of the assumptions made.

Whilst much progress has been made in projective geometry in determining the minimal formulations, Kanatani pointed out the importance of understanding the noise on the input data and showed the bias in standard techniques [265, 133]. This was further studied in [36, 37] and further developed error propagation for structured lighting systems [128]. In an unusual paper [131] uses knowledge of errors in the image data, which is propagated to generate confidence ranges in the depth estimation and reject unreliable estimates.

In Scharstein [226, 227] a number of objective error metrics are proposed to cover aspects of reliability as bad matching pixels defined as depth estimates outside acceptable ground truth; rms depth error in disparity units (depth precision). Special attention is given to the types of regions involved, so that textureless, occluded and depth discontinuity regions are treated separately. As an exercise in visualisation - the gaps in reprojected images for unseen viewpoints were also used for a subjective assessment of algorithm performance.

Is there a data set for which the correct answers are known? The subject of data sets for stereo has led to considerable debate about the need for accurate ground truth, as well as the density, source as well as accuracy of the ground truth - whether mechanically derived, manually-annotated, or independently measured (laser rangefinder, by structured lighting, etc).

Schafer proposed a useful taxonomy of data sets which categorises them in terms of the extent to which they use real or synthetic data [161]. Synthetic data, with added noise, in the form of indoor corridors[21] though [226] points out that the lack of realistic textures adversely affects the performance of area-based algorithms.

Some scenes would appear to be simpler than others, for example, the classic Pentagon, Tsukuba/Office-head and SRI/Trees images, although seemingly complex outdoor images, are mostly composed of in-plane surfaces which do not challenge most of the techniques that are based around this assumption.

Are there data sets in common use? Several ground-truthed data sets have been collected, made available and their use reported, with the Pentagon, Tsukuba/Office-head and SRI/Trees images being particularly popular. Available datasets include: the Stuttgart ISPRS Image Understanding

[21] www-dbv.cs.uni-bonn.de/stereo_data

datasets[22] [92, 93], JISCT Stereo Images[23] [206], the INRIA Syntim stereo databases[24] and various datasets from CMU[25] and the Middlebury dense stereo dataset[26] both of which include multi-baseline data. Furthermore Oxford/INRIA have made available uncalibrated sequences of Valbonne church[27] [67, 108] which can serve as redundant datasets as proposed in [146].

The Middlebury study [226, 227] describes an analysis of dense stereo algorithms as four steps in which the first is equivalent to the matching stages common with sparse stereo algorithms.

A collection of 6 test image sets with multiple quasi-planar textured surfaces are established and used to examine the sensitivity of pixel-by-pixel depth estimation reliability to a number of algorithms and parameterisations. The other algorithm steps, notably sparse-to-dense interpolation are also exercised. Test code, test data, ground truth and scoring code are offered to other groups via the Middlebury website. The stated intention is to extend the scheme to more complex scenes including textureless and more complex surface geometries.

Consensus still appears to be missing on the appropriate test data, and no work appears to have been carried out on relating the characteristics of a data set which would allow the performance on unseen data to be estimated.

Are there experiments that show algorithms are stable and work as expected? It is possible to use error propagation on edge-based estimation of depth (following correspondence) [139]. This can be tested with simulated data in order to show that the quantitative estimation of depth (for correct matches) has the expected behaviour [105].

The collection of additional redundant images of a scene has been proposed [146] to permit self-consistency checks to characterise accuracy and reliability of the correspondence stage. In this manner, reconstruction of the scene from two pairs permits comparison of extracted 3D structure in the common image frame. Some aspects of this redundancy are available in datasets which provide multibaseline data, e.g. the Middlebury datasets.

Are there any strawman algorithms? Although a great many stereo vision algorithms have been described in the literature, implementations other than the TINA system ref TINA (which contains feature and area based approaches) were not made publicly available until the creation of the Middlebury site which provides a set of modules, so the situation is much improved [4].

Is there a quantitative methodology for the design of algorithms? Correspondence matching is essentially a statistical selection problem which has been described in terms of probability theory [8]. Some aspects of the quantitative performance of these algorithms have been addressed using empirically-determined match and mis-match distributions providing a white box analysis of a feature matching algorithm [246] describing the relationship between the statistical assumptions, statistics of the data and parameter settings. The correspondence problem has also been considered as a MAP estimation of a geometric triangulation [16].

Torr and others [57, 253] investigated the stability of the fundamental matrix, a common intermediate representation for relating camera configuration to image plane and disparity.

The work of [117] makes some progress in improving performance at discontinuities, but no overall framework for quantitatively validating design decisions has been widely accepted.

[22] ftp://ftp.ifp.uni-stuttgart.de/pub/wg3
[23] ftp://ftp.vislist.com/IMAGERY/JISCT
[24] www-rocq.inria.fr/~tarel/syntim/paires.html
[25] www.ius.cs.cmu.edu/idb, www.cs.cmu.edu/~cil/cil-ster-html
[26] www.middlebury.edu/stereo
[27] www.robots.ox.ac.uk/~vgg/data1.html

What should we be measuring to quantify performance? Several metrics have been proposed to quantify the ability to generate accurate and reliable depth estimates.

Several authors have also examined the performance of individual algorithms in terms of quantisation error [218, 217] and 3D error [166, 27, 135, 173, 120, 73, 248, 269, 195, 219].

Whilst early work emphasised accuracy in the 3D world frame, the $1/Z$ depth-dependence on this measure had resulted in a move to disparity or image plane error estimation as a more representative measure.

The performance of stereo vision algorithms actually has three aspects:

1. Reliability: the ability to identify suitable correspondences, measured in terms of the number of features recovered with correct (usable) geometry.
2. Accuracy: The geometric precision of correctly-recovered features. Both of these properties require quantification and are sufficient for feature-based algorithms [203].
3. For the determination of dense depth data the accuracy of interpolation of stereo across smooth (featureless) surfaces becomes an issue. This can be achieved either on the basis of either an explicit surface model on the basis of known object shape or an implicit model hidden within the algorithm (see shape-based object indexing above).

One could logically argue that if knowledge of the correct model were available and used explicitly, then a likelihood-based determination of surface shape would give an optimal interpolation of surface position. Conventional techniques could then be used to provide error estimates in dense regions. Algorithms that embed interpolation within the matching process are unlikely to produce results that correspond to the correct surface model, thus requiring separate quantification for characteristic surface types. In fact this issue has much in common with the problems faced by estimation methods for dense optical flow, and most of the same arguments restricting quantitative use of derived data still apply [259].

4.8 Face Recognition

A practical face recognition algorithm must first detect faces and then recognize them. Performance characterization studies sometimes separate detection and recognition. For example, in the FERET evaluations [198] the distinction is drawn between fully automatic algorithms that detect, localize and recognize faces, versus partially automatic algorithms, that assume a face has already been detected and localized. Whether evaluation should decouple detection from recognition depends on what one wants to understand. To understand the operational characteristics of a complete algorithm, one should study fully automatic algorithms, as for example is done in the Face Recognition Vendor Tests [24, 200]. On the other hand, to characterize the best a particular recognition algorithm can do, absent errors in detection and localization, one should study just the recognition component behavior. Much of the academic literature on face recognition has adopted this latter approach.

Space will not permit us to cover best practices for performance characterization of face recognition algorithms at the level of detail covered in [198, 192, 1] and serious readers are strongly encouraged to review these papers. Here we will briefly summarize the current state-of-the-art as well as give a few indications of important directions for the future.

In characterizing face recognition algorithms, it is essential to distinguish between three distinct problems or tasks. The first task is detection and localization, as already suggested. The other two both fall under the broad term recognition, but are very different in their particulars. These two tasks are identification and verification [192]. Identification is the task of naming a person

from a novel image of that person [198]. Verification is the task of deciding if a person is who they claim to be based upon a novel image [216, 1]. Several documents provide important background for evaluation of face verification algorithms. One is the best practices document by Mansfield and Wayman [1]. It covers many aspects of analyzing and comparing different biometric verification systems, including those for faces. In particular, there are subtleties associated with some measures that space will not allow us to explore here.

For face recognition systems, as with most other biometric systems, one distinguishes between enrollment and operation. During enrollment, stored examples of known faces are added to the system, for face recognition these are typically stored in a gallery. During operation, when recognition is carried out, one or several novel images of a subject are compared to images in the gallery. Novel images are often called probe images, and performance in general depends upon how well a system can match probe images to the gallery.

The similarity matrix is a nearly universal abstraction and is used to characterize both face identification and verification algorithms. The presumption is that all verification and identification algorithms generate similarity scores between pairs of images. This abstraction lies at the heart of how modern large-scale evaluations such as FERET [198] or the Face Recognition Vendor Tests [24, 200] are carried out and is consistent with the offline approach to match score generation advocated in the "Best Practices in Testing and Reporting Performance of Biometric Devices" [1] . This abstraction allows analysis to be decoupled from the actual running of the algorithms. Typically algorithms are run over the union of all imagery that will be included in an evaluation and a single large similarity matrix is recorded. Analysis can then be carried out using just the information stored in the similarity matrices.

For identification, the gallery is sorted by decreasing similarity relative to the probe image, and the rank of the first gallery image of the same subject as the probe image is defined as the recognition rank. A recognition rank value of one indicates the corresponding gallery image of the same subject is more similar to the probe image than is any other subject's image in the gallery. Recognition rank two indicates one other subject had a probe image that was a better match to the probe image than the correct subject's image, and so on for higher recognition rank values. Recognition rate at rank k is defined over a set of probe images, and is the number of probe images recognized correctly at rank k over the total number of probe images. Some algorithms also normalize similarity scores for identification, and as with verification, so long as the normalization technique is known and only dependent upon the other scores, it can be accounted for when analyzing performance.

A combination of verification and identification arises when the task is to decide whether a person is among a specified set of people, and if so, who they are. This is sometimes called the watch-list task and it arises in contexts such as airport passenger screening. There are examples of watch-list evaluations [200], but the evaluation protocols for watch-list tasks are not as mature as for identification and verification, and we shall not say more about it here.

4.8.1 Face Detection & Localization

Face detection algorithms determine where faces are present in an image, and consequently there are essentially two issues from a performance characterization standpoint. First, are faces found and when found how accurately are they localized. Most performance characterization focuses on detection rather than localization, and hence detection rates, false alarm rates [171], and more generally ROC curves are reported [124, 207]. The face detection survey paper by Yang, Kriegman and Ahuja [171] is an excellent resource and Yang also maintains a website [28] that provides pointers to face detection data and algorithms as a supplement to the journal article.

[28] http://vision.ai.uiuc.edu/mhyang/face-detection-survey.html

Compared to analysis of detection, quantitative evaluation of localization behavior for algorithms is less common, and will not be addressed further here.

How is testing currently performed? Face detection is tested in essentially two related ways. One particularly easy way to test is to use only image chips that are either faces or not faces, and then simply score an algorithm according to whether it correctly distinguishes the face chips from the non-face chips. The alternative, and somewhat more realistic way, is to pass the detector over large images containing possibly many faces, and scoring how many faces are detected, how many are missed, and how many detections are false alarms. This latter approach, while more realistic, does introduce minor judgments relative to when a detection overlaps a true face sufficiently to be categorized as a true detection, as well as how to handle multiple detections of the same face.

Is there a data set for which the correct answers are known? There are two related data sets in common usage. One is the CMU Frontal Face Data [29]. This dataset is divided into Test Sets A, B, C and the Rotated Test Set. Test Sets A, C and the Rotated Test Set were collected at CMU by Henry A. Rowley, Shumeet Baluja, and Takeo Kanade. Test Set B was collected by Kah-Kay Sung and Tomaso Poggio at the AI/CBCL Lab at MIT. Ground truth is available for images in each of the four Test Sets. Eye, nose and corner of the mouth coordinates are provided for 169 faces in 40 images for Test Set A, 157 faces in 25 images for Test Set B, 185 faces in 55 images in Test Set C and 223 faces in 50 images in the Rotated Test Set. The other data set is the MIT CBCL Face Data Set [30]. This data set includes 19 by 19 pixel image chips, some of which are faces and some are not. The data set comes divided into a training and a test set, with 2,429 faces and 4,548 non-faces in the training set and 472 faces and 23,573 non-face images in the test set.

Are there datasets in common use? The two data sets just mentioned are the most common. Face detection can also be tested on more standard face data sets such as FERET and others listed below. However, a lack of non-face images makes serious evaluation of face detectors using only such data sets problematic.

Are there experiments that show algorithms are stable and work as expected? Most all face detection algorithms are trained, and so standard protocols for separating training data from test date are employed to test generalization. Beyond these standard tests of generalization, there are no other commonly used tests for predictability or stability.

Are there any strawman algorithms? There is no single universally recognized strawman algorithm against which to test new algorithms. However, several algorithms are emerging as possible standards of comparison. One is the face detector developed by Viola and Jones [194] using the AdaBoost [274] learning algorithm. There is an implementation of an AdaBoost frontal face detection algorithm with some refinements by Lienhart [207] available as part of the OpenCV distribution [31]. In the survey of Face Detection paper by Yang et al. [171] detection rates and false detection rates are compared for 9 different face detection algorithms, including Yang and Ahuja's SNOW face detector [158]. The SNOW algorithm is another possible strawman algorithm. However, while there is public code for the basic SNoW algorithm [32], the code is not packaged to work on face detection specifically. A third possible strawman algorithm is the SvmFu SVM

[29] http://vasc.ri.cmu.edu//idb/html/face/frontal_images/index.html
[30] MIT CBCL data - http://cbcl.mit.edu/cbcl/software-datasets/FaceData2.html
[31] http://sourceforge.net/projects/opencvlibrary
[32] http://l2r.cs.uiuc.edu/ cogcomp/ and follow link to software.

Data Set Summaries.			
		Approx. number of	
Data Set	Source	Subjects	Images
FERET[198]	NIST	¿ 1,000	¿ 4,000
UND Database	University of Notre Dame	¿ 500	¿ 30,000
UTD Database [190]	University of Texas, Dallas	284	¿ 2,500
BANCA [7]	Surrey University	208	¿ 5,000
AR Face	Purdue University	126	¿ 4,000
PIE Database [230]	Carnegie Mellon University	68	41,368
Links and/or contacts for Data Sets.			
FERET	www.itl.nist.gov/iad/humanid/feret/		
UND Database	http://www.nd.edu/ cvrl/UNDBiometricsDatabase.htm		
UTD Database	Contact Alice O'toole – http://www.utdallas.edu/∼otoole		
BANCA	http://www.ee.surrey.ac.uk/banca/		
AR Face	rvl1.ecn.purdue.edu/∼aleix/aleix_face_DB.html		
PIE Database	http://www.ri.cmu.edu/projects/project_418.html		

Table 1: Common data sets used for face identification evaluation.

system developed by Rifken at MIT [33] and compared with SNoW both in [171] and [3]. The MIT CBCL data includes data already in the format required by the SvmFu algorithm.

Is there a quantitative methodology for the design of algorithms? Theoretically, the measurement of biological parameters should be treated as an estimation task, with subsequent decision stages treated as hypothesis tests. There appears to be no global acceptance of this fact or any body of work which lays down the foundations for systematic application of these approaches. There are as many methodologies as there are distinct learning approaches [189, 194, 113].

What should we be measuring to quantify performance? Face detection is a classic detection problem, and standard ROC analysis is sufficient in most cases. There is an open question when an observed difference is statistically meaningful, and we are not aware of work that has addressed this specifically in the context of face detection.

4.8.2 Face Identification

The FERET protocol [198] already mentioned established a common framework for comparing face identification algorithms. Subsequent large scale evaluations include the Face Recognition Vendor Tests [24, 200].

How is testing currently performed? Typically a cumulative match characteristic (CMC) curve is presented for one or more algorithms tested on a specific set of probe and gallery images. The horizontal axis on such a curve is the recognition rank. As described above, the recognition rank for a specific probe image is the rank of the matching image of the same subject in the gallery when the gallery is sorted by decreasing similarity. The vertical axis is the recognition rate. Thus, a CMC curve reports how many probe images are correctly recognized at rank 1, rank 2, etc. Different combinations of probe and gallery sets may be used to capture such things as images taken on different days, under different lighting, etc.

[33] http://five-percent-nation.mit.edu/SvmFu/

Is there a data set for which the correct answers are known? Good identification experiments require large galleries of human subjects: 96% correct versus 92% on 25 subjects is a difference of one subject and not meaningful. Opinions on the exact number of subjects required vary, but it is probably a good rule of thumb that data sets of fewer than 100 subjects are inadequate. Thus, while there are many public face data sets, there are comparatively few with a sufficiently large number of distinct human subjects. Table 1 summarizes some of the most common publicly used data sets. All but the PIE database contain over 100 subjects, and PIE is included because it has an unusually high number of images per subject and these images were acquired under carefully controlled and varied conditions. In some cases, for example U.T. Dallas, and the BANCA datasets, video imagery is available.

Are there data sets in common use? The FERET data set is probably the most commonly used, and while it is poor in terms of the number of replicate images per subject, it is still a very useful and competitive data set in terms of the total number of subjects. However, the Notre Dame data is quickly becoming an extremely useful alternative with a much higher number of replicate images per subject.

Are there experiments that show algorithms are stable and work as expected? Common practice when addressing these types of questions is to conduct experiments that compare results obtained by an algorithm across a modest set of sample problems in which some external variable is sampled. For example, the original FERET tests compared performance under three qualitatively stated conditions: 1) same day images with different facial expressions, 2) images taken under changing illumination, and 3) extended elapsed time between acquisition of images. This general approach of dividing test data into a small number of cases and qualitatively comparing performance across cases is the most common approach to characterizing stability.

This partitioning of test data and coarse sampling is a limited methodology in so far as it lacks any underlying model of the data and provides only a coarse indication of future performance on novel data sets. There are some sources of variability, for example lighting, that have attracted particular attention and for which analytical models exist [17]. Hence, lighting variability is arguably a more mature and easier form of variation to study. Also, as illustrated by the Yale data set and the more recent and comprehensive PIE data set [230], data finely sampled across different illuminations settings is available.

Stability relative to changes in human subjects is of course measured whenever algorithms are trained on one set of subjects and tested on another. However, explanatory models relating attributes of human subjects to recognition difficulty are rare [87, 184]. We think such modeling is very important and expect to see more of it in the future.

Are there any strawman algorithms? A standard implementation of a Principle Components Analysis (PCA) based algorithm, sometimes call Eigenfaces [255], is the clear choice as a strawman algorithm. The algorithm is readily implemented in an environment such as MatLab. There is also a standardized ANSI C implementation that is available as part of the CSU Face Identification Evaluation System[28].

Some care must be taken, however, if doing one's own implementation, since apparently small details can alter performance greatly. Choice of distance measure in particular is critical. Using Euclidean distance will, for example, yield relatively weak results in many cases, while a measure that uses a cosine measure between images in a variance corrected, whitened, subspace performs very well in many cases [28]. For these reasons, It is best to use not only a standard algorithm, PCA, but a standardized implementation and competitive distance measure as well.

The CSU Face Identification Evaluation System also includes re-implementations of three of the original algorithms included in the FERET evaluation:

- A Combined Principle Components Analysis and Linear Discriminant Analysis algorithm (PCA+LDA) based upon the work of Zhao and Chellapa[277] at the University of Maryland.
- A Bayesian Intrapersonal/Extrapersonal Classifier (BIC) based upon the work of Moghaddam and Pentland[172] at MIT.
- An Elastic Bunch Graph Matching (EBGM) algorithm based upon the work of [187] et. al. at the University of Southern California.

Is there a quantitative methodology for the design of algorithms? Face identification falls into the theoretical category of labelling via probability density estimation. There is no single accepted quantitative methodology, just as there are no adequate and complete first principles models of how face imagery behaves. In other words, there are no comprehensive models that capture all the pertinent sources of variability. [34]. However, some of the most notable sources of images' variability are well studied and in some cases underlying models are used to either to guide algorithm design or actively as a part of an algorithm. Notable examples include illumination modeling [17] and 3D model-based viewpoint correction [26, 200]. Finally, while in no way special to face identification, most algorithm design follows a process of successive refinement as representative training and test data is used to isolate and correct weaknesses.

What should we be measuring to quantify performance? Current practice is to measure recognition rate at a given rank. As rank is varied, this leads to the Cumulative Match Characteristic (CMC) curve defined above. At a minimum, an author today wishing to argue for the superiority of new algorithm must provide comparative results demonstrating improved performance relative to some baseline using CMC analysis.

The most compelling problem with current analytical techniques is their sensitivity to unmodeled changes in scenario or problem. To put matters simply, too often a marked difference in performance on one data set does not consistently generalize. A first step toward greater confidence in the persistence of an observed difference is to explicitly capture some aspect of the uncertainty in the measured difference. A simple way of doing so that captures only uncertainty due to sample size, i.e. the number of observed differences, is McNamar's test [264]. Monte Carlo resampling may be used to capture other sources of variability, including choice of probe image [169] as well as probe and gallery images [19].

Finally, statistical models that make explicit connections between identifiable attributes of the face identification task and expected algorithm performance will, we think, prove to be very important. Thus, as already suggested, works such as [87] and [184] are early examples of what we trust will become a more common and fruitful approach to identifying and quantifying the factors affecting performance.

4.9 Measuring Structural Differences in Medical Images

A common task in medical image analysis is to compare two scans and quantify the differences between them in a clinically useful way. Situations where such comparisons are commonly made include (i) where an individual has had several scans of a particular type (modality) over periods from minutes through to days or even years (ii) where two or more individuals have had a single

[34] Nor does it seem likely that complete models will arise any time soon given the open nature of human subjects and appearance.

modality scan each [35]. . There are particular difficulties involved with analysing medical images that include (i) modality-dependent noise-characteristics (ii) relatively low resolution in some of the most common modalities (e.g. MRI (Magnetic Resonance Imaging) typically acquires $>= 1mm^3$ voxels) (iii) wide variation in scan quality depending on acquisition hardware, scanning protocol, scan-related image artefacts, patient cooperation (or lack of it) (iv) normal physiological processes (v) the need for robustness where image analysis is used to support clinical decision making and (vi) the small size of structures of interest such as lesions or important anatomic areas (e.g. the subthalamic nuclei) relative to the acquired voxel dimensions.

The differences of interest (tasks involved) fall into four broad categories:

- Appearance (or disappearance) of features.
- Differences in volume of corresponding features.
- Differences of shape of corresponding features.
- Differences of texture of corresponding features.

It is common for more than one of these processes to occur concurrently. For instance the size, shape and textural appearance of a tumour can change over time either due to natural growth processes or as a result of chemical, radiological or surgical intervention. Clearly this may make the identification of corresponding structures difficult. Measurement of differences is a three step process (i) identify corresponding features of interest in the images (ii) determine a mapping between the images, usually by some form of image registration (iii) compute statistics which summarise the differences between the features of interest. The order of the steps can vary (e.g. feature identification can be used to constrain image registration or image registration derived on the basis of voxel similarity can be used to identify corresponding features).

As discussed above, methods for detecting structural change usually rely on feature-detection (or segmentation) and image registration pre-processing steps to achieve correspondence of anatomy. Here we focus on the methods for quantifying the differences themselves and comment briefly where the detail of the pre-processing steps is important or where particular caveats apply. For more in depth coverage of image registration see the papers by Hill et al [116] and Zitova and Flusser [278]. For coverage of segmentation and feature detection see Olabarriaga and Smeulders [188] and Suri et al [130]. The appearance of new features can in principle be detected automatically using digital subtraction images and knowledge of the noise characteristics of the images. Simple statistical tests are used to identify voxels in the difference images with intensities sufficiently bright or dark that they are unlikely to be part of the noise spectrum. Clearly this approach can only be used when the intensity characteristics of tissues are the same in a set of images [36] so this is essentially a within-modality technique where, if necessary, the images are pre- registered so that corresponding features are in alignment. This approach has been most successful in brain applications such as lesion detection in pairs of pre- and post-contrast images in multiple sclerosis but there remain traps for the unwary [35]. The brain can often be considered a rigid-body for the purposes of image registration and as this analysis is performed within-subject, problems of anatomical correspondence [56] don't arise. A somewhat related brain application is so-called Voxel Based Morphometry (VBM) where groups of subjects are compared to discover population anatomical differences. One standard technique requires "spatial normalisation" of all subjects to an anatomical standard frame of reference followed by voxel-wise statistical analysis to identify regions where differences between the groups cannot be explained by intra-group variation. There are many methodological details to do with both the normalisation (which must not

[35] A complementary problem which will not be discussed here is where an individual has had several different scans (e.g. x-ray Computed Tomography, Magnetic Resonance Imaging, Ultrasound, radio-isotope imaging etc) which are acquired on different machines but which must be viewed (and fused) in a single frame of reference.

[36] Or an intensity mapping between the images can be estimated.

introduce anatomical bias but has to preserve real structural differences) [6] and the statistical framework (which must properly account for structural correlations within the images) that have lead to some debate about the validity of the framework (e.g. Crum et al. [**?**]; Davatzikos [59]). A slightly different case is dynamic contrast imaging where a time-series of images are acquired during injection of a contrast agent that is designed to identify structures of interest (such as blood-flow in vessels or leakage from lesions) by changing their intensity characteristics over time compared with uninteresting structures [126]. Analysis of such time-series to identify tumours or inflamed tissue requires a parametric model of time- varying contrast enhancement to be fitted at each voxel. Patient motion during image acquisition is an obvious confound which can in principle be addressed by image registration. However the time-varying nature of the contrast of different tissues can confound registration schemes with large degrees of freedom which operate by matching image intensity alone. In cardiac imaging, the dynamic change in volume and shape of the chambers of the heart can be measured by registering MR images acquired as snapshots during the cardiac cycle [37]. To analyse function in more detail, models of mechanical deformation can be constructed from the registered series of images to identify regional abnormalities in the myocardium (heart muscle). Over much longer timescales, there is growing interest in monitoring the progressive reduction in brain volume associated with dementia. Here, MR images of individuals' brains acquired months or years apart are registered either assuming the brain is a rigid structure, or by allowing the registration to deform one image to match the other [81]. In the former approach, analysis of brain volume change can proceed by directly integrating significant intensity differences over the registered brain region. In this so-called Brain Boundary Shift Integral (BBSI) [80] the changes in image intensity caused by shifting tissue boundaries are transformed into a volume change. In a complementary approach known as Structural Image Evaluation, using Normalisation, of Atrophy (SIENA), (Smith et al, [235]) brain edge-maps are constructed explicitly on each of the registered brain images and the distances between corresponding edge-points are integrated to obtain the volume change. Where non-rigid registration is used to match the images, differences in shape and volume are encoded in the deformation field. Local volume change can be estimated by computing the Jacobian determinant of the transformation [81]. In comparisons of groups of subjects, statistical analysis can be performed directly on the parameters of the deformation fields to detect significant group differences in the location and shape of structures - so-called deformation based morphometry [6, 85]. Ashburner et al. [123] distinguish this from tensor-based morphometry where statistical analysis is performed on quantities derived from the deformation fields, the simplest example being the determinant of the Jacobian of the transformation which is an estimate of local volume change.

How is testing currently performed? A variety of testing strategies are employed. There are few standardised benchmarks. For lesion detection concordance with expert manual observers can be used as a gold standard as can images featuring simulated lesions. Automated estimates of volume change can be tested by comparison with manual measurement or by image pairs displaying proscribed (simulated) atrophy. Techniques such as VBM rely on a carefully thought out statistical framework and correlation with other clinical indicators rather than explicit testing against a gold standard. Some attempts to compare this class of algorithm with manual methods have been inconclusive [85, 88]. Some researchers have investigated the effect of parameter choices on the spatial normalisation (registration) component of the algorithms by examining co-localisation of manually defined landmarks [224]. In some clinically applied situations a direct comparison of the results of dynamic contrast imaging against analysis of excised tissue can be made (e.g. [270, 193]).

[37] In practice data is aggregated over several heartbeats to obtain higher quality images. Motion due to breathing can be compensated for in most subjects by breath-holding for between 10s and 40s during the image acquisition.

Is there a data set for which the correct answers are known? Both real and digital test objects ("phantoms") are often used. Phantoms are constructed commercially for quality assurance of imaging systems and are often relatively simple. Phantoms suitable for testing image analysis must have a realistic appearance under imaging and must be constructed of materials that give acceptable contrast and in some cases are compatible with more than one imaging modality[30]. Phantoms constructed to closely resemble human anatomy are referred to as "anthropomorphic". Often so-called "digital" phantoms are employed where a simulation of the imaging of a known anatomical model can create realistic test objects (e.g. generating nuclear medicine brain images from an MR image, [90], or carefully labelling acquired images to form a digital model as in the Zubal phantom [279] [38]). In applications that measure structural change, some way to represent realistic shape and volume change or growth processes must be found. Castellano Smith et al. [40] used a model of diffuse brain atrophy applied to the BrainWeb digital phantom to produce images of a brain subject to varying amounts of atrophy. Schnabel et al. [125] used a biomechanical model of the breast to simulate the effect of plate compression, needle-biopsy, muscle relaxation etc. during dynamic contrast imaging.

Are there datasets in common use? There are a variety of datasets used to test algorithms although they tend to be very application dependent and not always freely available. The Centre for Morphometric Analysis has set up the Internet Brain Repository [39] to make available sets of expertly labelled MR brains to use as a gold standard for evaluation of novel segmentation schemes. They have also been used for evaluating non-rigid registration algorithms (Crum et al., 2004). The Montreal Neurological Institute provide the BrainWeb digital brain phantom [48] that comprises a simulated MR brain volume (either normal or with MS lesions) and is widely used as both a test object and an anatomical standard.

Are there experiments that show that the algorithms are stable and work as expected? Consistency over multiple scans is used to check the robustness of algorithms. For algorithms which quantify change over time, a common check is to ensure that for three scans, A, B and C, of a subject, adding the measured change between scans A and B, and B and C, gives change measured between scans A and C. A related consistency check is to ensure that the change measured between scans A and C is minus the change measured between scans C and A. To check for scanner related robustness, "same-day" scans are used during testing where a subject has two scans A1 and A2 within a few minutes or hours of each other. The measured changes between A1 and A2 and a later scan B are then compared. The assumption is that the changes of interest in the patient are occurring over a longer time-scale then normal physiological changes or factors that contribute to noise in the scanner. Simulated images with varying levels of noise and/or artifact are also used to characterise the stability of algorithms. Unfortunately there are so many sources of variation in medical imaging applications that is extremely difficult to test all eventualities, but the use of simulation data makes it at least possible to test for bias and errors on 'normal' data.

Are there any strawman algorithms? There are no completely generic straw-man algorithms. Although some advances in algorithms are purely improvements on previous methods and so can be tested against them, others reflect advances in image acquisition and cannot be easily related to more established techniques. As an example, measurements of volumetric change have been performed using manual tracing or semi-automatic region-growing techniques for well over a decade and can often be used as a yardstick for newer algorithms. Newer imaging techniques such as MR tagging (used to track dynamic tissue motion) and Diffusion Tensor Imaging (used

[38] http://noodle.med.yale.edu/zubal/index.htm
[39] http://www.cma.mgh.harvard.edu/ibsr/

to extract directional information from tissue structure) have required the development of novel analysis algorithms that cannot easily be referred back to older methods.

Is there a quantitative methodology for the design of algorithms? Most algorithms display at least some awareness of the noise properties of the images. Usually they assume Gaussian stationary noise that can be estimated from examining zero signal regions of the image. This is often an approximation (e.g. structural MR images are nearly always magnitude images reconstructed from two-channel data with noise on each channel; the resulting noise distribution in the reconstructed data is Rician) and in Ultrasound images much of the speckle originates from coherent scattering in the images which can in principle provide more information about the underlying anatomy. Much effort has been expended into incorporating knowledge of image artifacts into analysis algorithms, a common example being MR tissue segmentation algorithms which simultaneously recover a low frequency intensity bias artifact. In general algorithms make assumptions about the intensity ranges associated with particular tissue types. In the case of X-ray CT, the images provide direct quantitative information about x-ray attenuation coefficients that can be used to discriminate between tissues. MR is more complicated as the returned signal is a relative value that depends on several physical parameters and can vary greatly depending on how the scanner is configured. Algorithms usually specify some generic MR image type (or combination of types) and assume, for instance, that in intensity terms CSF $<$ grey matter $<$ white matter in the case of T1-weighted volumetric imaging [178]. Algorithms often try to account for the so-called partial volume effect where the intensity at a single voxel results from a combination of more than one tissue present in that volume. There have also been efforts to normalise acquired images using a physics based acquisition model to aid analysis (e.g. the HINT representation in mammography [115]). As illustrated by algorithms attempting non-rigid registration, the one quantitative aspect lacking from much of this work is an attempt to provide error estimates for estimated parameters, though this failure has been noted [56].

What should we be measuring to quantify performance? The two standard approaches are to compare results against the same measurements made by an expert human observer or to use the results to classify subjects in a way that can be tested against classification based on independent clinical observations. Both of these approaches, while a necessary part of initial testing, only estimate expected error rather than gauging success on previously unseen data. In principle a statistical analysis can assign a confidence to the detection of a novel structure against a background of Gaussian noise of known mean and variance. In fact one of the most popular forms of Voxel-Based Morphometry works, essentially by looking for inter-group differences that can't be explained by intra- group variation. However one of the main unresolved problems at the time of writing is that the detection task is not isolated but is at the end of a pipeline of multiple stages each of which introduces uncertainty to the analysis. Many of these sources of error can either be measured (e.g. imaging noise) or guarded against (e.g. careful scanner calibration to reduce non-linear spatial distortion in MR). In most applications where structural change over time is being measured an image registration step is required to bring scans acquired at different times into spatial correspondence. The propagation of errors in rigid landmark-based registration is well understood but in the voxel-based and non-rigid cases, despite being increasingly used for clinically applied applications, significant additional manual effort is required to estimate errors in new data. Therefore the priority is to develop new methods for measuring registration error, as this will greatly simplify analysis of structural change. This problem has much in common with issues relating to the quantitative estimation of optical flow, object localisation and dense stereo, as described above. The solution, to make quantitative predictions of measurement error, would appear to be the same and require the same technical approaches.

	Is there a data set for which the correct answers are known?	Are there datasets in common use?	Are there experiments which show that the algorithm works as expected?	Are there any strawman algorithms?	Is there a quantitative methodology for the design of algorithms?
Sensor characterisation	Yes	Yes	Yes	N/A	Yes
Video Compression	Yes	Yes	Yes	Yes	No
Feature Detection	(Yes)	Yes	(Yes)	Yes	(Yes)
Object Localisation	Yes	No	(Yes)	No	(Yes)
Object Indexing	Yes	(Yes)	(Yes)	No	(Yes)
Optical Flow	Yes	(Yes)	Yes	Yes	(Yes)
Stereo Vision	Yes	(Yes)	(Yes)	Yes	(Yes)
Face Recognition	Yes	(Yes)	(Yes)	(one)	No
Medical Structure	Yes	Yes	Yes	(Yes)	(Yes)

Table 2: Summary of answers to the five closed questions

4.10 Summary

In this section we attempt to collate the answers to the five closed key questions for each of the visual tasks and draw some conclusions.

Table 2 summarises the response as a clear 'yes' or 'no' where the evidence supports this. In some cases only a qualified '(yes)' can be justified, because there are only a few examples or we believe that the situation is not generally accepted.

From the table we can observe the following:

- There is clearly still variable use of performance characterisation techniques by researchers across in the various areas of vision. Some tools are in use in some areas, whilst others could be applied across other areas. The biometric best practice guidelines [163] is probably the best developed overall guide at present.
- There now appears to be fairly widespread use of annotated datasets across most visual tasks, except curiously object localisation.

  There is still some dissatisfaction expressed about limited data set size and coverage. While some work on data set design has been published [94, 46, 95] and the problems are quite well understood within other fields such as text corpora [237, 238, 147, 84], these lessons do not appear to have been applied systematically within computer vision. There is considerable sharing of such data sets via the internet or digital media, thus reducing the cost of data collection and annotation.

  The use of data generation/synthesis was demonstrated within OCR [118] and more recently for fingerprint biometrics [162], and this may be expected to spread to other tasks.
- There is now a substantial appreciation of the importance of the replication of results using common data and protocols. This is most strongly demonstrated by work in stereo [226] and corner detection [170] though most areas are now sharing data.
- When organised around workshops (the PETS series[40]) or specific tasks, (FERET/FRVT [22,

[40] http://visualsurveillance.org/

23]) this kind work appears to be leading to co-operation and accelerated technical progress, just as previously experienced within the speech community as a result of NIST workshops and comparisons.

- There appears to be broad acceptance of the value of testing at the technology level [197] using generic metrics independent of the context of any specific application task (although it is rarely described as such) as evidenced by the wide availability of datasets.

  There is convergence towards common evaluation metrics and presentation formats, most notably ROC curves, which are now standard practice, after a long period when few detection algorithms discussed false detection rates. There is still some variation in terminology (which is perhaps to be expected) and some ongoing debate [65, 21, 180, 242] and numerical metrics are still not uniformly defined.

- Regarding experiments which show that algorithm works as expected, consistency measures across multiple images of the same scene are now commonly used to demonstrate stability for almost all classes of algorithm. Training generalisation protocol are followed in object and face recognition and Monte Carlo studies can reveal bias. However estimation pdfs are rarely examined, and the use of covariances to determine significance of differences is still very limited [152, 153] making comparisons between papers very difficult.

- The majority of evaluation experiments would still appear to fall into the category of black box analysis - testing on data sets and recording of results - with relatively little ability to utilise these results to predict performance. The building of explicit white box models of performance [208, 214] is a very strong idea as it allows the building of explicit modelling linking the properties of the data with performance, and thus gives a means to predict performance on an unseen dataset, based only on the properties of that data set.

- The availability of shared or strawman code in most areas has made comparison of multiple existing algorithms easier and more consistent. Open source code (such as OpenCV, PhD toolkits) is starting to play a role, despite some quality issues.

- A number of evaluation protocols have been published and taken up, most particularly within biometrics and stereo vision. Downloadable packages comprising annotated data, strawman code and unambiguous scoring code have started to become available, but are not yet widely used.

- Regarding a quantitative methodology for the design of algorithms, there is limited work which addresses this aspect. One can point to Canny's definition of an optimality criteria[39] and Spies' [239, 240] studies in optical flow. Likewise white box-like analyses of the behaviour of intermediate representations in object recognition (modelling the Hough transform [205] and of an invariance scheme [168, 167]); types of variability in face recognition (e.g. illumination [17] and 3D viewpoint [26, 200]) and stereo (stability of the fundamental matrix [57, 253]; match and mis-match distributions [246]). A firm 'yes' would be required to provide valid approaches and appropriate (correct) solutions. This goes to the core of what would be regarded as making scientific progress in this area. However the fact that this column has so few unambiguous 'yes' answers suggests that as a subject, we have only just started to lay the foundations which would establish machine vision as a mature scientific discipline.

An unambiguous (scientific) solution of a particular visual task would require a 'yes' to every question. It is important to note that the table would have looked very different just a few years ago, as the majority of the 'yes' answers are due to recent publications. So in general, all of the elements to achieve this would appear to already be in the literature, but the subject seems to lack the motivation to form scientific conclusions.

In this paper we have tried to highlight the main performance issues by showing the commonality of themes across a variety of topics. This has given us the opportunity to identify, what we feel to be, the insights which may take the subject toward firm conclusions.

We readily acknowledge that there are many other subject areas of active research in computer vision which cannot be explored further in this paper due to space limitations [41].

5 Discussion and Future Directions

5.1 Software Validation

In implementing a particular algorithm as computer code, the question arises about the correctness of the code. In [76] Förstner proposed that vision algorithms should include 'traffic light' outputs as a self-diagnosis tool to indicate the level of confidence in their own operation: green for good, red for poor or no output, and amber for intermediate.

Looking at the covariance and comparing with the Monte Carlo performance can demonstrate that all aspects of the theory have been correctly implemented. However, disagreement does not necessarily indicate that the algorithm is wrong - since there may be numerical problems with the covariance calculation such as non-linearities [98, 101, 103]. Before we can claim that the software is fit for use, we also need to ensure that the data matches the assumptions made in the Monte Carlo simulation.

Little progress seems to have been made regarding the general problem of changing *image domain*. This is, being able to predict algorithm behaviours on unseen images, with different characteristics. While it is expected that use of covariance propagation will at least provide self-consistent estimates of accuracy, issues involving correspondence require scenario evaluation. This problem is intimately linked with the ultimate intended use of the algorithm. Ultimately we believe that this cannot be addressed solely by the acquisition of isolated sets of test data sets in order to test with them. In fact we believe that what is required is a simulation of the imaging system and its environment.

Calibration of such a simulation using the test data sets would make it possible to evaluate any other performance scenario as a function of whichever characteristics were believed to be salient.

Figure 1 shows three different approaches to testing an algorithm. The simplest approach, the one most commonly followed, is shown on the right: given a population of imagery, one samples a set of images for any training and another set for testing. The result of the testing process is intended to be representative of the results that would be obtained by testing the system on the entire population.

An alternative approach, shown on the left of Figure 1, is to analyse the types of variation encountered in the population. This can be used to develop a *data generator*, both as an analytical model and as a piece of software. The analytical model can be used to derive a prediction of how well the algorithm should perform. If the analytical model is correct, it should predict performance that is — within statistical sampling limits — equivalent to the results obtained by the purely test-based approach.

There is a further possible route for testing. The analytical model can form the basis of a piece of software, and the latter can then be used to generate synthetic imagery that exhibits the same variations as the population as a whole. (Of course, the software may be hybrid, incorporating both analytic and sampling components.) The algorithm can then be applied to the synthesised data; the results obtained from this process should be comparable with the results obtained by

[41] A more complete bibliography of some 1,000 evaluation-related papers can be found at http://peipa.essex.ac.uk/benchmark/bib/index.html

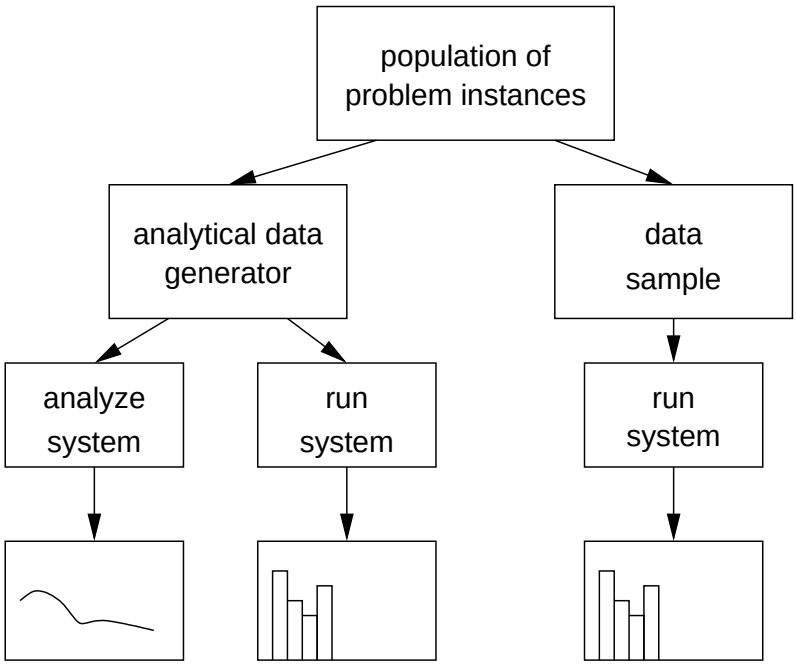

Figure 1: Different approaches to testing algorithms

the methods outlined in *both* the previous paragraphs.

5.2 Developments in The Science of Algorithm Development

As we have shown, there are number of tools that are available today that have been used in a number of different areas to provide quantitative evaluations of algorithms. In doing so, the basic tools have had to be adapted to the specific circumstances which revolve around numerical issues, in common with the algorithms themselves. This often requires a substantial amount of creative thought to overcome. The skills required to solve problems of numerical stability are ones with which computer vision researchers are already familiar but which still require substantial amounts of creative thought.

In the future, this might lead to off-shoots into research exploring numerically stable techniques, for example for stable estimation of covariances from noisy data. In addition, the whole issue of more powerful optimisation techniques is worthy of further study but cannot currently be identified as a problem as we do not know what an algorithm could be expected to have delivered.

In some cases, further development in quantitative statistical methods is required for better understanding and modelling the behaviour of algorithms.

The use of higher order statistics may be a way to avoid problems with the non-quadratic or non-linear shape of some functions where reformulation is not possible. However, reliable estimation of higher order statistics places greater demands on the quality of the data. In the future we might expect that these techniques would be required, but only in specialised circumstances.

Model selection is another topic for further work that is key to the automatic interpretation of arbitrary images. Likelihood techniques require the use of an appropriate model and the determination of the appropriate number of free parameters. Unfortunately, likelihood methods do not return values that can be directly used to select methods. The neural information criteria, the Bhattacharyya metric and the AIC have been used [96], as often illustrated for specific formulations (e.g. in curve model fitting [179]).

5.3 New Application Areas

In seeking to apply vision algorithms, it is important to test their validity not just novelty, so that the techniques developed by the community may be applied to problems outside. This will have the advantage for the researchers that recognition, in the form of additional collaborative publications may be gained from the joint validation and application of the techniques, and that the techniques justify being added to the armory of validated tools for new application problems.

6 Conclusions

In general there appear to be common difficulties which are shared across the subject, in many areas specific toolkits have been developed to address difficult issues and we believe that these could be applied in other application areas. Unfortunately, what is becoming increasingly apparent is that the knowledge base of many researchers does not generally provide them with access to these techniques, it is therefore important that workers in the field make efforts to increase the general breadth of skills, particularly in areas such as probability theory and quantitative statistics. This would support a better theory of algorithm design and methodology. We need to be able to get to the position that algorithms are correct by design, rather than relying entirely upon empiricism and shootouts. It is particularly worrying that although a lot of good work does appear to have been addressing fundamental issues there is a lack of general acceptance for what would appear to be scientifically justifiable opinions. We therefore need better ways of arriving at consensus within the community for such results, so that these approaches are more widely accepted as valid within the community.

Acknowledgements

The support of the Information Society Technologies programme of the European Commission is gratefully acknowledged under the PCCV project (Performance Characterisation of Computer Vision Techniques) IST-1999-14159. William Crum is grateful for financial and intellectual support from the EPSRC/MRC Medical Images and Signals IRC (GR/N14248/01). John Barron gratefully acknowledges support from a NSERC (National Science and Engineering Research Council of Canada) Discovery grant.

A Honest Probabilities

In the process of finding the best interpretation of a set of data, we would anticipate the need to compute the probability of a particular hypothesis, either as a direct interpretation of scene contents or as the likely outcome of an action. Statistical measures of performance can be obtained by testing on a representative set of data. The correct use of statistics in algorithm design should be fully quantitative. Probabilities should correspond to a genuine prediction of data frequency, this property has been described as **honest**. From the point of view of algorithmic evaluation, if an algorithm does not make quantitative predictions then it is by definition untestable in any meaningful manner.

The term honest simply means computed probability values should directly reflect the relative frequency of occurrence of the hypothesis in sample data. Examples are; classification probabilities P(C—data) should be wrong 1-P(C—data) of the time, Hypothesis tests, (that a particular set of data could have been generated by a particular model), should have a uniform probability

distribution by construction, and error estimates from Likelihood should match the true distribution of estimated quantities. Direct confirmation of such characteristics provide powerful tools for the quantitative validation of systems and provide mechanisms for online self test.

Some approaches to pattern recognition, such as k-nearest neighbours, are almost guaranteed to be honest by construction. In addition the concept of honesty provides a very powerful way of assessing the validity of probabilistic approaches. In the case of combining regional information to refine a label category, a topic often referred to as probabilistic relaxation labelling, Poole [202] was able to argue that any system which drives iterated probability estimates to conclusive 0 or 1 labels, and yet still had a finite performance error, could not be honest. It was shown that this problem originated with the assumption that the update function was equal at all update iterations, and how this could be improved by estimating these functions from sample data.

B Systems Analysis Methodology Review: Based on Ramesh&Haralick94

In this section we outline the systems analysis methodology described in [214]. The methodology essentially addresses the problem of analyzing and setting up tuning constants for a vision system with a chosen architecture. The methodology does not address computational issues nor the choice of the architecture itself.

B.0.1 Systems Analysis Given Chosen Algorithm Sequence

Let A denote an algorithm. At the abstract level, the algorithm takes in as input, a set of observations, call them input units U_{In}, and produces a set of output units U_{Out}. Associated with the algorithm is a vector of tuning parameters $\mathbf{T}$. The algorithm can be thought of as a mapping $A : (U_{In}, \mathbf{T}) \rightarrow U_{Out}$. Under ideal circumstances, if the input data is ideal (perfect), the algorithm will produce the ideal output. In this situation, doing performance characterization is meaningless. In reality the input data is perturbed, perhaps due to sensor noise or perhaps due to the fact that the implicit model assumed in the algorithm is violated. Hence the output data is also perturbed. Under this case the inputs to (and the outputs from) an algorithm are observations of random variables. Therefore, we view the algorithm as a mapping: $A : (\hat{U}_{In}, \mathbf{T}) \rightarrow \hat{U}_{Out}$, where the $\hat{}$ symbol is used to indicate that the data values are observations of random variables. This brings us to the verbal definition of performance characterization with respect to an algorithm:

"Performance characterization or Component Identification for an algorithm has to do with establishing the correspondence between the random variations and imperfections on the output data and the random variations and imperfections on the input data."

More specifically, the essential steps for performance characterization of an algorithm include:

1. the specification of a model (with parameter $\mathbf{D}$) for the ideal input data.
2. the specification of a model for the ideal output data.
3. the specification of an appropriate perturbation model (with parameter $\mathbf{P_{In}}$) for the input data.
4. the derivation of the appropriate perturbation model (with parameter $\mathbf{P_{Out}}$) for the output data (for the given input perturbation model and algorithm).
5. the specification and the evaluation of an appropriate criterion function (denoted by Q_{Out}) relative to the final calculation that the algorithm makes to characterize the performance of the algorithm.

The main challenge is in the derivation of appropriate perturbation models for the output data and relating the parameters of the output perturbation model to the input perturbation, the algorithm tuning constants, and the ideal input data model parameters. This is due to the fact that the specification of the perturbation model must be natural and suitable for ease of characterization of the performance of the subsequent higher level process. Once an output perturbation model is specified, estimation schemes for obtaining the model parameters have to be devised. In addition, the model has to be validated, as theoretical derivations may often involve approximations.

The ideal input data is often specified by a model with parameters specified by a vector $\mathbf{D}$ and the algorithm is often an estimator of these parameters. First, we note that the ideal input data is nothing but a sample from a population of ideal inputs. The characteristics of this population, i.e. the exact nature of the probability distributions for $\mathbf{D}$, are dependent on the problem domain. The process of generation of a given ideal input can be visualized as the random sampling of a value of $\mathbf{D}$ according to a given probability distribution $F_{\mathbf{D}}$.

Let $\mathbf{P_{In}}$ denote the vector of parameters for the input perturbation model. Let $Q_{Out}(\mathbf{T}, \mathbf{P_{In}}, \mathbf{D})$ denote the criterion function that is to be optimized[42]. Then the problem is to select $\mathbf{T}$ so as to optimize the performance measure Q, over the entire population, that is given by:

$$Q(\mathbf{T}, \mathbf{P_{In}}) \int Q_{Out}(\mathbf{T}, \mathbf{P_{In}}, \mathbf{D}) dF_{\mathbf{D}} \tag{1}$$

In the situation where the perturbation model parameters, $\mathbf{P_{In}}$, are not fixed, but have a specific prior distribution then one can evaluate the overall performance measure by integrating out $\mathbf{P_{In}}$. That is:

$$Q(\mathbf{T}) \int Q(\mathbf{T}, \mathbf{P_{In}}) dF_{\mathbf{P_{In}}} \tag{2}$$

Having discussed the meaning of performance characterization with respect to a single algorithm, we now turn to the situation where simple algorithms are cascaded to form complex systems.

Let Φ denote the collection of all algorithms. Let $A^{(i)} \in \Phi$, then $A^{(i)} : U_{In}^{(i)} \rightarrow U_{Out}^{(i)}$ is the mapping of the input data $U_{In}^{(i)}$ to the output $U_{Out}^{(i)}$. Note that the unit for $U_{In}^{(i)}$ may not be the same as the unit for $U_{Out}^{(i)}$ and perturbations in the input unit type causes perturbations in the output unit type. A performance measure, $Q^{(i)}$, is associated with A_i. Associated with each algorithm is the set of input parameters $\mathbf{T^{(i)}}$. The performance measure is a function of the parameters $\mathbf{T^{(i)}}$.

An algorithm sequence, S, is an ordered tuple:

$$S : (A^{(1)}, A^{(2)}, \ldots, A^{(n)})$$

where n is the number of algorithms utilized in the sequence. Associated with an algorithm sequence is a parameter vector sequence

$$\mathbf{T} : (\mathbf{T}^{(1)}, \mathbf{T}^{(2)}, \ldots, \mathbf{T}^{(n)})$$

and a ideal input data model parameter sequence:

$$\mathbf{D} : (\mathbf{D}^{(1)}, \mathbf{D}^{(2)}, \ldots, \mathbf{D}^{(n)})$$

. The performance at one step of the sequence is dependent on the tuning parameters, and the perturbation model parameters at all previous stages. So:

$$Q_i f_i(\mathbf{T}^{(i)}, \mathbf{T}^{(i-1)}, \ldots, \mathbf{T}^{(1)}, \mathbf{P_{In}}^{(i-1)}, \ldots, \mathbf{P_{In}}^{(1)}).$$

[42] Note that the input data $\hat{U}_{In}$ is not one of the parameters in the criterion function. This is correct if all the input data do not violate any of the assumptions about the distribution(s) of $\mathbf{D}$ and $\mathbf{P_{In}}$.

So the overall performance of the sequence is given by:

$$Q_n(\mathbf{T}, \mathbf{P_{In}}) \qquad f_n(\mathbf{T}^{(n)}, \mathbf{T}^{(n-1)}, \ldots, \mathbf{T}^{(1)}, \mathbf{P_{In}}^{(n-1)}, \ldots, \mathbf{P_{In}}^{(1)}).$$

The free parameter selection problem can now be stated as follows: *Given an algorithm sequence S along with the parameter sequence T and performance measure Q_n, select the parameter vector $\mathbf{T}$ that maximizes Q_n* . Note that Q_n is actually the integral:

$$Q_n(\mathbf{T}, \mathbf{P_{In}}) \int \ldots \int f_n(\mathbf{T}^{(n-1)}, \ldots, \mathbf{T}^{(1)}, \mathbf{P_{In}}^{(n-1)}, \ldots, \mathbf{P_{In}}^{(1)}, \mathbf{D}^{(n-1)}, \ldots, \mathbf{D}^{(1)}) dF_{\mathbf{D}^{(n-1)}} \ldots dF_{\mathbf{D}^{(1)}}.$$

Note that at each stage a different set of prior distributions $F_{\mathbf{D}^{(i)}}$ comes into play. Also, the perturbation model parameters $\mathbf{P_{In}}^{(i)}$ is a function $g_i(T^{(i-1)}, \mathbf{P_{In}}^{(i-1)}, \mathbf{D}^{(i-1)}, A^{(i-1)})$. In other words, the perturbation model parameters at the output of stage i are a function of the tuning parameters at stage $i-1$, the input perturbation model parameters in the stage $i-1$, the ideal input data model parameters, and the algorithm employed in the stage $i-1$. It is important to note that the functions g_i depend on the algorithm used. No assumption is made about the form of the function g_i.

The derivation of the optimal parameters $\mathbf{T}$ that maximize $Q_n(\mathbf{T}, \mathbf{P_{In}})$ is rather tedious and involved. Therefore in practice the thresholds $\mathbf{T}$ are selected in each individual stage relative to the final task. For example, in [210], thresholds for a sequence of operations involving boundary extraction and linking were chosen relative to the global classification task of extracting building features to satisfy a given misdetection rate for building feature detection and a given false alarm rate for clutter boundary pixels. In the examples that follow, we will adopt a similar strategy. That is, we will set up pruning thresholds in a video surveillance application by defining probability of missing valid hypotheses and probability of false hypotheses as criterions. The reader must note that these criterion functions are essentially functions of the ideal parameters $\mathbf{D}$'s and one has to integrate over the prior distribution of the $\mathbf{D}$'s.

References

[1] A. J. Mansfield and J. L. Wayman. Best Practices in Testing and Reporting of Biometric Devices: Version 2.01. Technical Report NPL Report CMSC 14/02, Centre for Mathematics and Scientific Computing, National Physical Laboratory, UK, August 2002.

[2] T. D. Alter and W. E. L Grimson. Verifying model-based alignments in the presence of uncertainty. In *Computer Vision Pattern Recognition CVPR97*, pages 344–349, Puerto Rico, June 1997.

[3] M. Alvira and R. Rifkin. An empirical comparison of snow and svms for face detection. In *MIT AI Memo*, 2001.

[4] X. Armangue, J. Pags, J. Salvi, and J. Batlle. Comparative survey on estimating the fundamental matrix. In *Proc. 9th Symp. Nacional de Reconocimiento de Formas y Analisi de Imagenes*, pages 227–232, June 2001.

[5] A. P. Ashbrook, N. A. Thacker, P. I. Rockett, and C. I. Brown. Robust recognition of scaled shapes using pairwise geometric histograms. In *Proceedings of the British Machine Vision Conference BMVC95*, Birmingham, UK, September 1995.

[6] Frackowiak R-Johnsrude I Price C Ashburner J, Hutton C and Friston K. Identifying global anatomical differences: Deformation-based morphometry. *Human Brain Mapping*, 6(5-6):358–347, 1998.

[7] Enrique Bailly-Bailliére, Samy Bengio, Frédéric Bimbot, Miroslav Hamouz, Josef Kittler, Johnny Mariéthoz, Jiri Matas, Kieron Messer, Vlad Popovici, Fabienne Porée, Belén Ruíz, and Jean-Philippe Thiran. The banca database and evaluation protocol. In *4th International Conference on Audio- and Video-based Biometric Person Authentication*, pages 625–638, 2003.

[8] H. H. Baker and T. O. Binford. Depth from edge and intensity based stereo. In *Proceedings of the VII International Joint Conference on Artificial Intelligence*, August 1981.

[9] S. Baker and S. K. Nayar. Global measures of coherence for edge detector evaluation. In *Proceedings of Conference on Computer Vision and Pattern Recognition CVPR 1999*, pages 2373–2379, Ft. Collins, CO, USA, June 1999.

[10] S. T. Barnard and W. B. Thompson. Disparity analysis of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(4):333–340, 1980.

[11] J. L. Barron. Experience with 3D optical flow on gated MRI cardiac datasets. In 1^{st} *Canadian Conference on Computer and Robot Vision*, pages 370–377, 2004.

[12] J. L. Barron and R. Eagleson. Recursive estimation of time-varying motion and structure parameters. *Pattern Recognition*, 29(5):797–818, May 1996.

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

[14] J. L. Barron and M. Khurana. Determining optical flow for large motions using parametric models in a hierarchical framework. In *Vision Interface*, pages 47–56, May 1997.

[15] B.G. Batchelor and J. Charlier. Machine vision is not computer vision. pages 2–13, 1998.

[16] A. Bedekar and R. M. Haralick. A bayesian method for triangulation. In *ICIP-95: Proceedings, International Conference on Image Processing*, volume II, pages 362–365, October 1995.

[17] P.N. Belhumeur and D.J. Kriegman. What is the set of images of an object under all possible illumination conditions. *IJCV*, 28(3):245–260, July 1998.

[18] J. H. Bergen, P. Anandan, K. J. Hamma, and R. Hingorani. Hierarchical model-based motion estimation. In *European Conference on Computer Vision*, pages 237–252, May 1992.

[19] J. Ross Beveridge, Kai She, Bruce Draper, and Geof H. Givens. A nonparametric statistical comparison of principal component and linear discriminant subspaces for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 535 – 542, December 2001.

[20] H. A. Beyer. Accurate calibration of ccd cameras. In *Proceedings of Conference on Computer Vision and Pattern Recognition CVPR92*, pages 96–101, Urbana-Champaign, USA, 1992.

[21] J. Black, T. Ellis, and P. Rosin. A novel method for video tracking performance evaluation. In *Joint IEEE Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, pages 125–132, 2003.

[22] D. M. Blackburn, M. Bone, and P. J. Philips. Facial recognition vendor test 2000: evaluation report. Technical report, DARPA, 2000.

[23] D. M. Blackburn, M. Bone, and P. J. Phillips. Facial recognition vendor test 2000. Technical report, DARPA, December 2000.

[24] Duane M. Blackburn, Mike Bone, and P. Jonathon Phillips. Facial Recognition Vendor Test 2000: Executive Overview. Technical report, Face Recognition Vendor Test (www.frvt.org), 2000.

[25] J. Blanc-Talon and D. Popescu, editors. *Imaging and Vision Systems: Theory, Assessment and Applications*. NOVA Science Books, 2001.

[26] V. Blanz, S. Romdhami, and T. Vetter. Face identification across different poses and illuminations with a 3d morphable model. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, pages 202 – 207, 2002.

[27] S. D. Blostein and T. S. Huang. Error analysis in stereo determination of 3D point positions. *IEEE Transactions PAMI*, 9(6):752–765, November 1987. Correction: PAMI(10), No. 5, September 1988, p. 765.

[28] David Bolme, J. R. Beveridge, Marcio Teixeira, and Bruce A. Draper. The csu face identification evaluation system: Its purpose, features and structure. In *Proceedings of the Third International Conference on Vision Systems*, pages 304–311, Graz, Austria, April 2003.

[29] M. Boshra and B. Bhanu. Predicting performance of object recognition. *IEEE transactions PAMI*, 22(8):956–969, September 2000.

[30] De Guise JA Daronat M Qin Z Cloutier G Boussion N, Soulez G. Geometrical accuracy and fusion of multimodal vascular images: A phantom study. *Medical Physics*, 31(6):1434–1443, 2004.

[31] K. W. Bowyer, C. Kranenburg, and S. Dougherty. Edge detector evaluation using empirical ROC curves. In *Proceedings of Conference on Computer Vision and Pattern Recognition CVPR 1999*, pages 1354–1359, Ft. Collins, CO, USA, June 1999.

[32] K. W. Bowyer and P. J. Phillips, editors. *Empirical evaluation techniques in computer vision*. IEEE Press, 1998. ISBN 0-8186-8401-1.

[33] K. W. Bowyer and P. J. Phillips, editors. *Empirical evaluation techniques in computer vision*. IEEE Press, 2000. ISBN 0-8186-8401-1.

[34] P. A. Bromiley, M. Pokric, and N. A. Thacker. Computing covariances for mutual information coregistration. In *Proc. MIUA*, pages 77–80, 2004.

[35] Thacker N.A. Bromiley P.A. and Courtney P. Non-parametric image subtraction using grey level scattergrams. *Image and Vision Computing*, 20:609–617, 2002.

[36] M. Brooks, W. Chojnacki, D. Gawley, and A. van den Hengel. What value covariance information in estimating vision parameters? In *International Conference on Computer Vision ICCV2001*, Vancouver, B.C., Canada, July 2001.

[37] M. J. Brooks, W. Chojnacki, A. van den Hengel, and D. Gawley. Is covariance information useful in estimating vision parameters ? In *Videometrics and Optical Methods for 3D Shape Measurement, Proceedings of SPIE*, volume 4309, pages 195–203, San Jose, USA, January 2001.

[38] M. Pike E. Sparks C. Charnley, G. Harris and M. Stephens. The droid 3d vision system - algorithms for geometric integration. Technical report, Plessey Research Roke Manor, December 1988.

[39] J. F. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.

[40] Hill DLG Thacker NA Castellano Smith AD, Crum WR and Bromiley PA. Biomechanical simulation of atrophy in mr images. In J. Michael Fitzpatrick, editor, *Proceedings of SPIE 5032 Medical Imaging 2003*, volume Image Processing, pages 481–490, Milan, May 2003. Sonka.

[41] A. K Chhabra and I. T. Phillips. A benchmark for graphics recognition systems. In K.W. Bowyer and P.J. Phillips, editors, *Empirical Evaluation Techniques in Computer Vision*, IEEE Comp Press, CA, USA, 1998.

[42] K. Cho, P. Meer, and J. Cabrera. Performance assessment through bootstrap. *IEEE transactions PAMI*, 19(11):1185–1198, November 1997.

[43] H.I. Christensen and P.J. Phillips. Procs. 2nd workshop on empirical evaluation methods in computer vision. Dublin, Ireland, 2000.

[44] H.I. Christensen and P.J. Phillips, editors. *Empirical Evaluation Methods in Computer Vision*. World Scientific Press, 2002. ISBN 981-02-4953-5.

[45] A. Clark and P. Courtney. Workshop on performance characterisation and benchmarking of vision systems. Canary Islands Spain, 1999.

[46] A. F. Clark and P. Courtney. Databases for performance characterisation. In *DAGM workshop on Performance Characteristics and Quality of Computer Vision Algorithms*, Technical University of Brunswick, Germany, September 1997.

[47] T.A. Clarke and Fryer. The development of camera calibration methods and models. *Photogrammetric Record*, 16(91):51–66, 1998.

[48] Kollokian V Sled JG Kabani NJ Holmes CJ Collins DL, Zijdenbos AP and Evans AC. Design and construction of a realistic digital brain phantom. *IEEE Transactions on Medical Imaging*, 17(3):463–468, 1998.

[49] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean-shift. In *Proceedings of Conference on Computer Vision and Pattern Recognition CVPR 2000*, pages II: 142–149, Hilton Head, SC, USA, June 2000. IEEE Computer Society.

[50] T. F. Cootes, D. Cooper, C. J. Taylor, and J. Graham. Active shape models — their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, January 1995.

[51] P. Courtney. Proceedings of ecvnet workshop on performance evaluation of vision algorithms. Paris, France, 1995.

[52] P. Courtney and J. T. Lapreste. Performance evaluation of a 3D tracking system for space applications. In *DAGM workshop on Performance Characteristics and Quality of Computer Vision Algorithms*, Technical University of Brunswick, Germany, September 1997.

[53] P. Courtney and T. Skordas. Caracterisation de performances des algorithmes de vision - un exemple: le detecteur de coins. In *Proceedings RFIA10*, pages 953–962, Rennes, France, 1996.

[54] P. Courtney, N. Thacker, and A. Clark. Algorithmic modelling for performance evaluation. In *Proceedings of the ECCV Workshop on Performance Characteristics of Vision Algorithms*, Cambridge, UK, April 1996. http://www.vision.auc.dk/ hic/perf-proc.html.

[55] P. Courtney, N. Thacker, and A. Clark. Algorithmic modelling for performance evaluation. *Machine Vision and Applications*, 9(5/6):219–228, April 1997.

[56] Jenkinson M. Kennedy D. Crum W.R., Rueckert D. and Smith S.M. Zen and the art of medical image registration : Correspondence, homology and quality. *NeuroImage*, 20:1425–1437, 2003.

[57] G. Csurka, C. Zeller, Z. Zhang, and O. Faugeras. Characterizing the uncertainty of the fundamental matrix. *Computer Vision and Image Understanding*, 68(1):18–35, October 1997.

[58] V. Ramesh D. Comaniciu and P. Meer. Kernel-based optical tracking. *IEEE Transactions PAMI*, 25(5):564–577, May 2003.

[59] C. Davatzikos. Why voxel-based morphometric analysis should be used with great caution when characterizing group differences. *NeuroImage*, 23(1):17–20, 2004.

[60] E. R. Davies. *Machine Vision: Theory, Algorithms, practicalities*. London Academic press, 2nd edition, 1997.

[61] T. Day and J. P. Muller. Digital elevation model production by stereo-matching spot image pairs: A comparison of two algorithms. *Image Vision Computing*, 7:95–101, 1989.

[62] D. Dori, W. Liu, and M. Peleg. How to win a dashed line detection contest. In R. Kasturi and K. Tombre, editors, *Graphics Recognition - Methods and Applications*, pages 13–22. Springer Verlag, 1996.

[63] B. Efron and G. Gong. A Leisurely Look at the Bootstrap, the Jackknife, and Cross-validation. *American Statistician*, 37:36–48, 1983.

[64] D. W. Eggert, A. Lorusso, and R. B. Fisher. Estimating 3D rigid body transformations: a comparison of four major algorithms. *Machine Vision and Applications*, 9(5/6):272–290, April 1997.

[65] T. Ellis. Performance metrics and methods for tracking in surveillance. In *3rd IEEE International Workshop on Performance Evaluation of Tracking and Surveillance PETS'2002*, Copenhagen, Denmark, June 2002.

[66] N.A. Thacker F.Ahearne and P.I.Rockett. The bhattacharryya metric as an absolute similarity measure for frequency coded data. *Kybernetica*, 34(4):363–368, 1997.

[67] Luong Q Faugeras, O. and T. Papadopoulo. *The Geometry of Multiple Images*. MIT Press, 2001.

[68] O. Faugeras, P. Fua, B. Hotz, R. Ma, L. Robert, M. Thonnat, and Z. Zhang. Quantitative and qualitative comparisons of some area and feature-based stereo algorithms. In *Proceedings of the Second International Workshop*, pages 1–26. Wichmann, Karlsruhe, Germany, March 1992.

[69] C. Fermüller and Y. Aloimonos. The statistics of optical flow: Implications for the processes of correspondence in vision. In *ICPR*, volume 1, pages 119–126, 2000.

[70] C. Fermüller, Y. Aloimonos, and H. Malm. Bias in visual motion processes: A theory predicting illusions. In *Statistical Methods in Video Processing*. (in conjunction with European Conference on Computer Vision), 2002.

[71] C. Fermüller, H. Malm, and Y. Aloimonos. Uncertainty in visual processes predicts geometrical optical illusions. Technical Report CR-TR-4251, Computer Vision and Mathematics at Lund Institute of Technology, Sweden, May 2001.

[72] C. Fermüller, R. Pless, and Y. Aloimonos. Statistical biases in optical flow. In *Conference on Computer Vision and Pattern Recognition*, volume 1, pages 561–566, 1999.

[73] D. Ferrari, G. Garibotto, and S. Masciangelo. Towards experimental computer vision: Performance assessment of a trinocular stereo system. In *Procs. ECCV ESPRIT day workshop*, Santa Margherita Ligure, Italy, May 1992.

[74] J. M. Fitzpatrick and J. B. West. A blinded evaluation and comparison of image registration methods. In K.W. Bowyer and P.J. Phillips, editors, *Workshop on Empirical Evaluation Techniques in Computer Vision*, Santa Barbara, California, June 1998.

[75] D. Fleet. *Measurement of Image Velocity*. Kluwer Academic Publishers, Norwell, 1992.

[76] W. Foerstner. 10 pros and cons against performance characterisation of vision algorithms. In *Proceedings of the ECCV Workshop on Performance Characteristics of Vision Algorithms*, Cambridge, UK, April 1996. http://www.vision.auc.dk/ hic/perf-proc.html.

[77] W. Förstner. DAGM workshop on performance characteristics and quality of computer vision algorithms. Technical University of Brunswick, Germany, 1997.

[78] W. Förstner and H.I. Christensen. Special issue on performance evaluation. *Machine Vision and Applications*, 9(5/6), 1997.

[79] W. Forstner and S. Ruwiedel. *Robust Computer Vision-Quality of Vision Algorithms*. Wichmann, Karlsruhe, Germany, March 1992.

[80] P.A. Freeborough and N.C. Fox. he boundary shift integral: An accurate and robust measure of cerebral volume changes from registered repeat mri. *Transactions on Medical Imaging*, 16(5):623–629, 1997.

[81] P.A. Freeborough and N.C. Fox. Modeling brain deformations in alzheimer disease by fluid registration of serial 3d mr images. *Journal of Computer Assisted Tomography*, 22(5):838–843, 1998.

[82] W. Freeman and E. Adelson. The design and use of steerable filters. *IEEE transactions PAMI*, 13(9):891–906, 1999.

[83] from R. M.Haralick. Dialogue: Performance characterization in computer vision. *Computer Vision Graphics and Image Processing: Image Understanding*, 60:245–265, 1994. contributions L.Cinque, C. Guerra, S. Levialdi, J. Weng, T. S.Huang, P. Meer, Y. Shirai; B. A.Draper, J. R.Beveridge.

[84] J. R. Galliers and K. Sparck Jones. Evaluating natural language processing systems. Technical Report Technical report 291, University of Cambridge, March 1993.

[85] Kiebel S Riehemann S Sauer H Gaser C, Volz HP. Detecting structural changes in whole brain based on nonlinear deformations - application to schizophrenia research. *NeuroImage*, 10(2):107–113, August 1999.

[86] J. C. Gee. Performance evaluation of medical image processing algorithms. In H.I. Christensen and P.J. Phillips, editors, *Empirical Evaluation Methods in Computer Vision*, pages 143–159. World Scientific Press, September 2002. ISBN 981-02-4953-5.

[87] Geof Givens, J. Ross Beveridge, Bruce A. Draper, P. Jonathon Phillips, and Patrick Grother. How Features of the Human Face Affect Recognition: a Statistical Comparison of Three Face Recognition Algorithms. In *Proceedings: IEEE Computer Vision and Pattern Recognition 2004*, pages 381–388, 2004.

[88] Fox NC Ashburner J Friston KJ Chan D Crum WR Rossor MN Good CD, Scahill RI and Frackowiak RSJ. Automatic differentiation of anatomical patterns in the human brain: Validation with studies of degenerative dementias. *NeuroImage*, 17(1):29–46, 2002.

[89] V. Gouet, P. Montesinos, R. Deriche, and D. Pele. Evaluation de détecteurs de points d'intérêt pour la couleur. In *Reconnaissance des formes et Intelligence Artificielle (RFIA'2000)*, volume II, pages 257–266, Paris, France, 2000.

[90] Biraben A Buvat I Benali H Bernard AM Scarabin JM Gibaud B Grova C, Jannin P. annin p, biraben a, buvat i, benali h, bernard am, scarabin jm, gibaud b, a methodology for generating normal and pathological brain perfusion spect images for evaluation of mri/spect fusion methods: application in epilepsy. *Physics in Medicine and Biology*, 48(24):4023–4043, December 2003.

[91] Hakon Gudbjartsson and Samuel Patz. The rician distribution of noisy MRI data. *MRM*, 34:910–914, 1995.

[92] E. Guelch. Results of test on image matching of ISPRS WG III/4. *International Archives of Photogrammetry and Remote Sensing*, 27(3):254–271, 1988.

[93] E. Guelch. Results of test on image matching of ISPRS WG III/4. *ISPRS Journal of Photogrammetry and Remote Sensing*, 46:1–18, 1991.

[94] I. Guyon, R. M. Haralick, J. Hull, and I. Phillips. Data sets for OCR and document image understanding. In *Handbook on Optical Character Recognition and Document Image Analysis (in press)*. World Scientific Publishing Company, 1996.

[95] I. Guyon, J. Makhoul, R. Schwartz, and V. Vapnik. What size test set gives good error rate estimates? *IEEE Transactions PAMI*, pages 52–64, 1998.

[96] H.Akaike. A new look at statistical model identification. *IEEE Trans. on Automatic Control*, 19:716, 1974.

[97] R. M. Haralick. Computer vision theory: The lack thereof. *Computer Vision Graphics and Image Processing*, 36:272–286, 1986.

[98] R. M. Haralick. Performance characterization in computer vision. In Boyle R. Hogg D., editor, *Proceedings of the British Machine Vision Conference BMVC92*, pages 1–8. Springer-Verlag, 1992.

[99] R. M. Haralick. Overview: Computer vision performance characterization. In *Proc. DARPA Image Understanding Workshop*, volume 1, pages 667–674, November 1994.

[100] R. M. Haralick. Overview: Computer vision performance characterization. In *Proceedings of the ARPA Image Understanding Workshop*, pages 663–665, Monterey, USA, 1994.

[101] R. M. Haralick. Propagating covariances in computer vision. In *Proceedings of International Conference on Pattern Recognition ICPR94*, pages 493–498, Jerusalem, Israel, September 1994.

[102] R. M. Haralick. Covariance propagation in computer vision. In *Proceedings of the ECCV Workshop on Performance Characteristics of Vision Algorithms*, Cambridge, UK, April 1996. http://www.vision.auc.dk/ hic/perf-proc.html.

[103] R. M. Haralick. Propagating covariance in computer vision. *Intl. Journal. Pattern Recogn. Art. Intelligence*, 10:561–572, 1996.

[104] R. M. Haralick, C-N Lee, K. Ottenberg, and M. Noelle. Review and analysis of solutions of the three point perspective pose estimation problem. *International Journal on Computer Vision*, 13(3):331–356, 1994.

[105] A. J. Harris, N. A. Thacker, and A. J. Lacey. Modelling feature based stereo vision for range sensor simulation. In *Proceedings of the European Simulation Multiconference*, pages 417–421, June 1998.

[106] C. Harris and M.Stephens. A combined corner and edge detector. In *Proc. 4th Alvey Vision Conference*, pages 147–151, 1988.

[107] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of the 4th Alvey Vision Conference*, pages 47–151, 1988.

[108] R.I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004. ISBN: 0521540518.

[109] H. W. Haußecker and D. J. Fleet. Computing optical flow with physical models of brightness variation. *Trans. Pattern Analysis and Machine Intelligence*, 23(6):661–673, June 2001.

[110] G. E. Healey and R. Kondepudy. Radiometric CCD camera calibration and noise estimation. *IEEE Transactions PAMI*, 16(3):267–276, March 1994.

[111] M. Heath, S. Sarkar, T. Sanocki, and K. Bowyer. Comparison of edge detectors: A methodology and initial study. In *Computer Vision Pattern Recognition CVPR96*, pages 143–148, 1996. pairwise human rating.

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

[113] Henry Schneiderman. Learning Statistical Structure for Object Detection. In *Computer Analysis of Images and Patterns (CAIP)*, 2003.

[114] W. Förstner H.I. Christensen and C.B. Madsen. Proceedings of the eccv workshop on performance characteristics of vision algorithms. Cambridge, UK, 1996. http://www.vision.auc.dk/ hic/perf-proc.html.

[115] R. Highnam, M. Brady, and B. Shepstone. A representation for mammographic image processing. *Medical Image Analysis*, 1(1):1–18, 1996.

[116] Holden M Hill DLG, Batchelor PG and Hawkes DJ. Medical image registration. *Physics in Medicine and Biology*, 46(3):R1–R45, March 2001.

[117] H. Hirschmueller. Improvements in real-time correlation-based stereo vision. In *IEEE Workshop on Stereo and Multi-Baseline Vision*, Hawaii, December 2001.

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

[119] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–204, 1981.

[120] Y. C. Hsieh, D. M. McKeown Jr., and F. P. Perlant. Performance evaluation of scene registration and stereo matching for cartographic feature extraction. *IEEE Transactions PAMI*, 14:214–238, February 1992.

[121] J. Hutchinson. Culture, communication and an information age madonna. *IEEE Professional Communictions Society Newsletter*, 45(3):1–5, May/June 2001.

[122] D. Huynh. The cross ratio: A revisit to its probability density function. In M. Mirmehdi and B. Thomas, editors, *Proceedings of the British Machine Vision Conference BMVC 2000*, Bristol, UK, September 2000. URL http://www.bmva.ac.uk/bmvc/2000/papers/p27.pdf.

[123] Ashburner J and Friston KJ. Voxel-based morphometry - the methods. *NeuroImage*, 11:805–821, 2000.

[124] J. P. Egan. *Signal Detection Theory and ROC Analysis*. Academic Press, 1975.

[125] A.D. Castellano-Smith A. Degenhard M.O. LeachO D.R. Hose D.L.G. Hill D.J. Hawkes J.A. Schnabel, C. Tanner. Validation of nonrigid image registration using finite- element methods: Application to breast mr images. *IEEE Transactions on Medical Imaging*, 22(2):238–247, February 2003.

[126] A. Jackson. Analysis of dynamic contrast enhanced mri. *British Journal of Radiology*, 77:S154–S166, 2004.

[127] R. C. Jain and T. Binford. Dialogue: Ignorance, myopia and naivete in computer vision systems. *Computer Vision Graphics and Image Processing: Image Understanding*, 53, January 1991. contributions from: M. A. Snyder, Y. Aloimonos, A. Rosenfeld, T. S. Huang, K. W. Bowyer and J. P. Jones.

[128] O. Jokinen and H. Haggren. Statistical analysis of two 3-D registration and modeling strategies. *Photogrammetry and Remote Sensing*, 53(6):320–341, December 1998.

[129] G. Jones and B. Bhanu. Recognition of articulated and occluded objects. *IEEE transactions PAMI*, 21(7):603–613, July 1999.

[130] S. Singh J.S. Suri and L. Reden. Computer vision and pattern recognition techniques for 2-d and 3-d mr cerebral cortical segmentation (part i): A state-of-the-art review. *Pattern Analysis and Applications*, 5(1):46–76, 2002.

[131] G. Kamberova and R. Badcsy. Sensor errors and the uncertainties in stereo reconstruction. In *Workshop on Empirical Evaluation Methods in Computer Vision*, Santa Barbara, California, June 1998.

[132] K. Kanatani and N. Ohta. Optimal robot self-localization and reliability evaluation. In H. Burkhardt and B. Neumann, editors, *European Conference Computer Vision ECCV98*, pages II: 796–808, Freiburg, Germany, 1998.

[133] K. I. Kanatani. Unbiased estimation and statistical analysis of 3-D rigid motion from two views. *IEEE Transactions PAMI*, 15(1):37–50, January 1993.

[134] M. Kass, A. Witkin, and D. Terzopolous. Snakes: Active contour models. *International Journal of Computer Vision*, 1:321–323, 1987.

[135] S. M. Kiang, R. J. Chou, and J. K. Aggarwal. Triangulation errors in stereo algorithms. In *CVWS87*, pages 72–78, 1987.

[136] R. Klette, F. Wu, and S. Z. Zhou. Multigrid convergence based evaluation of surface approximations. In R. Klette, S. Stiehl, M. Viergever, and V. Vincken, editors, *Performance Evaluation of Computer Vision Algorithms*, pages 241–254, Amsterdam, 2000. Kluwer Academic Publishers.

[137] B. Kong, I. T. Phillips, R. M. Haralick, A. Prasad, and R. Kasturi. A benchmark: Performance evaluation of dashed line detection algorithms. In R. Kasturi and K. Tombre, editors, *Graphics Recognition - Methods and Applications*, pages 270–285. Springer Verlag, 1996.

[138] S. Konishi, A. L. Yuille, J. M. Coughlan, and S. C. Zhu. Fundamental bounds on edge detection: An information theoretic evaluation of different edge cues. In *Proceedings of Conference on Computer Vision and Pattern Recognition CVPR 1999*, pages 1573–1579, Ft. Collins, CO, USA, June 1999.

[139] E. P. Krotkov. *Active computer vision by cooperative focus and stereo*. Springer series in Perception Engineering. Springer-Verlag, 1989.

[140] R. Kumar and A. R. Hanson. Pose refinement: Application to model extension and sensitivity to camera parameters. In *ARPA Image Understanding Workshop*, pages 660–669, 1990.

[141] R. Kumar and A. R. Hanson. Sensitivity of the pose refinement problem to accurate estimation of camera parameters. In *International Conference Computer Vision ICCV90*, pages 365–369, 1990.

[142] R. Kumar and A. R. Hanson. Robust methods for estimating pose and a sensitivity analysis. *Computer Vision Graphics and Image Processing*, 60(3):313–342, November 1994.

[143] T. Moons L. Van Gool and D. Ungureanu. Affine / photometric invariants for planar intensity patterns. In *European Conference Computer Vision ECCV96*, pages 642–651, Cambridge, England, April 1996.

[144] R. Lane, N. A. Thacker, and N. L. Seed. Stretch correlation as a real-time alternative to feature-based stereo matching algorithms. *Image and Vision Computing*, 12(4):203–212, 1994.

[145] L. J. Latecki, R. Lakamper, and U. Eckhardt. Shape descriptors for non-rigid shapes with a single closed contour. In *Proceedings of Conference on Computer Vision and Pattern Recognition CVPR 2000*, pages 424–429, Hilton Head, SC, USA, June 2000. IEEE Computer Society.

[146] Y. G. Leclerc, Q-T Luong, and P. Fua. Self consistency: a novel approach to characterising the accuracy and reliability of point correspondence algorithms. In A. Clark and P. Courtney, editors, *Workshop on Performance Characterisation and Benchmarking of Vision Systems*, Canary Islands Spain, January 1999.

[147] D. Lewis and K. Sparck Jones. Natural language processing for information retrieval. Technical Report Technical report 307, University of Cambridge, 1993.

[148] W. E. L.Grimson and D. P. Huttenlocher. On the verification of hypothesized matches in model-based recognition. *Lecture notes in Computer Science*, 427, 1990.

[149] J. Liang, I. T. Phillips, and R. M. Haralick. An optimisation methodology for document structure extraction on latin character documents. *IEEE transactions PAMI*, 23(7):719–734, July 2001.

[150] J. L. Liang, I. T. Phillips, and R. M. Haralick. Performance evaluation of document structure extraction algorithms. *Computer Vision and Image Understanding*, 84(1):144–159, October 2001.

[151] T. Lin and J. L. Barron. *Image Reconstruction Error for Optical Flow*, pages 269–290. Scientific Publishing Co., Singapore, (C. Archibald and P. Kwok, (eds.)), 1995.

[152] M. Lindenbaum. Bounds on shape-recognition performance. *IEEE Transactions PAMI*, 17(7):666–680, July 1995.

[153] M. Lindenbaum. An integrated model for evaluating the amount of data required for reliable recognition. *IEEE transactions PAMI*, 19(11):1251–1264, November 1997.

[154] W. Liu and D. Dori. Performance evaluation of graphicstext separation. In K. Tombre and A. Chhabra, editors, *Graphics Recognition Algorithms and Systems (Lecture Notes in Computer Science, vol. 1389)*, pages 359–371. Springer, 1998.

[155] A. Lorusso, D. W. Eggert, and R. B. Fisher. A comparison of four algorithms for estimating 3-D rigid transmformation. In *British Machine Vision Conference BMVC95*, Birmingham, UK, September 1995.

[156] D. G. Lowe. Object recognition from local scale-invariant features. In *Int. Conf. Comp. Vision ICCV*, pages 1150–1157, Kerkyra, Greece, 1999.

[157] B. D. Lucas and T. Kanade. An iterative image-registration technique with an application to stereo vision. In *Image Understanding Workshop*, pages 121–130. US Defence Advanced Research Projects Agency, 1981. (see also International Joint Conference on Artificial Intelligence81, pp674-679).

[158] M.-H. Yang, D. Roth, and N. Ahuja. SNoW-Based Face Detector. In S.A. Solla, T. K. Leen, and K.-R. Muller, editor, *Advances in Neural Information Processing Systems 12*, pages 855–861. MIT Press, 2000.

[159] C. B. Madsen. A comparative study of the robustness of two-pose estimation techniques. In *Proceedings of the ECCV Workshop on Performance Characteristics of Vision Algorithms*, Cambridge, UK, April 1996. http://www.vision.auc.dk/ hic/perf-proc.html.

[160] C. B. Madsen. A comparative study of the robustness of two-pose estimation techniques. *Machine Vision and Applications*, 9(5/6):291–303, April 1997.

[161] M. W. Maimone and S. A. Shafer. A taxonomy for stereo computer vision experiments. In *Proceedings of the ECCV Workshop on Performance Characteristics of Vision Algorithms*, Cambridge, UK, April 1996. http://www.vision.auc.dk/ hic/perf-proc.html.

[162] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar. *Handbook of Fingerprint Recognition*. Springer, New York, 2003.

[163] A. J. Mansfield and J. L. Wayman. Best practices in testing and reporting performance of biometric devices. Technical report, National Physical Laboratory, Teddington, Middlesex, UK, August 2002.

[164] R. Marik, J. Kittler, and M. Petrou. Error sensitivity assessment of vision algorithms based on direct error propagation. In *Proceedings of the ECCV Workshop on Performance Characteristics of Vision Algorithms*, Cambridge, UK, April 1996. http://www.vision.auc.dk/ hic/perf-proc.html.

[165] R. Marik, M. Petrou, and J. Kittler. Compensation of the systematic errors of the CCD camera. In *SCIA'97*, Lapeenranta, Finland, May 1997.

[166] L. Matthies and S. Shafer. Error modelling in stereo navigation. *IEEE J. Robotics and Automation*, 3(3):239–248, June 1987.

[167] S. J. Maybank. Probabilistic analysis of the application of the cross ratio to model-based vision. *International Journal Computer Vision*, 15(1):5–33, September 1995.

[168] S. J. Maybank. Probabilistic analysis of the application of the cross ratio to model-based vision: Misclassification. *International Journal Computer Vision*, 14(3):199–210, April 1995.

[169] Ross J. Micheals and Terry Boult. Efficient evaluation of classification and recognition systems. In *IEEE Computer Vision and Pattern Recognition 2001*, pages I:50–57, December 2001.

[170] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *Proceedings of Conference on Computer Vision and Pattern Recognition CVPR 2003*, USA, June 2003.

[171] Ming-Hsuan Yang, David Kriegman, and Narendra Ahuja. Detecting Faces in Images: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 24(1):34–58, 2002.

[172] B. Moghaddam, C. Nastar, and A. Pentland. A bayesian similarity measure for direct image matching. *ICPR*, B:350–358, 1996.

[173] G. G. Mohan, R.and Medioni and R. Nevatia. Stereo error detection, correction and evaluation. *IEEE Transactions PAMI*, 11(2):113–120, February 1989.

[174] H.P. Moravec. Obstacle avoidance and navigation in the real world by a seeing robot rover. Technical report, Stanford Univ., September 1980.

[175] H.-H Nagel. Constraints for the estimation of displacement vector fields from image sequences. In *International Joint Conference on Artificial Intelligence*, pages 945–951, 1983.

[176] Thomas A. Nartker and Stephen V. Rice. OCR accuracy: UNLV's third annual test. *INFORM*, 8(8):30–36, September 1994.

[177] Thomas A. Nartker, Stephen V. Rice, and Junichi Kanai. OCR accuracy: UNLV's second annual test. *INFORM*, 8(1):40–45, January 1994.

[178] A.Jackson N.A.Thacker. Mathematical segmentation of grey matter, white matter and cerebral spinal fluid from mr image pairs. *British Journal of Radiology*, 74:234–242, 2001.

[179] D.Prendergast N.A.Thacker and P.I.Rockett. 'b-fitting: A statistical estimation technique with automatic parameter selection. In *Proc, BMVC 1996*, pages 283–292, Edinburgh, 1996.

[180] C.J. Needham and R.D. Boyle. Performance evaluation metrics and statistics for positional tracker evaluation. In M. Petrou J. Kittler and M. Nixon, editors, *International Conference on Vision Systems ICVS03*, pages 278–289, Graz, Austria, 2003.

[181] S. Negahdaripour. Revised definition of optical flow: Integration of radiometric and geometric cues for dynamic scene analysis. *Pattern Analysis and Machine Intelligence*, 20(9):961–979, 1998.

[182] O. Nestares, D. J. Fleet, and D. J. Heeger. Likelihood functions and confidence bounds for total-least-squares problems. In *Proceedings of Conference on Computer Vision and Pattern Recognition CVPR 2000*, pages 1523–1530, Hilton Head, SC, USA, June 2000. IEEE Computer Society.

[183] O. Nestares and R. Navarro. Probabilistic multichannel optical flow analysis based on a multipurpose visual representation of image sequences. In *IS&T/SPIE 11th Intl. Symposium on Electronic Imaging*, pages 429–440, 1999.

[184] Nicholas Furl, P. Jonathon Phillips and Alice J. O'Toole. Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis. *Cognitive Science*, 26:797 – 815, 2002.

[185] A. Nomura. Spatio-temporal optimization method for determining motion vector fields under non-stationary illuminations. *Image and Vision Computing*, 18:939–950, 2000.

[186] American Society of Photogrammetry, editor. *Manual of Photogrammetry (5th ed.)*. American Society of Photogrammetry, Falls Church, Va. USA, 2004. ISBN 1-57083-071-1.

[187] Kazunori Okada, Johannes Steffens, Thomas Maurer, Hai Hong, Egor Elagin, Hartmut Neven, and Christoph von der Malsburg. The Bochum/USC Face Recognition System And How it Fared in the FERET Phase III test. In H. Wechsler, P. J. Phillips, V. Bruce, F. Fogeman Soulié, and T. S. Huang, editors, *Face Recognition: From Theory to Applications*, pages 186–205. Springer-Verlag, 1998.

[188] S.D. Olabarriaga and A.W.M. Smeulders. Interaction in the segmentation of medical images: A survey. *MEDICAL IMAGE ANALYSIS*, 5(2):127–142, June 2001.

[189] E. Osuna, R. Freund, and F. Girosi. Training support vector machines: An application to face detection. In *CVPR97*, pages 130–136, 1997.

[190] A. J. O'Toole, J. Harms, S. L. Snow, D. R. Hurst, M. R. Pappas, J. Ayyad, and H. Abdi. A video database of moving faces and people. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page (to appear), 2005.

[191] M. Otte and H.-H. Nagel. Optical flow estimation: Advances and comparisons. In *European Conference on Computer Vision*, pages 51–60, 1994.

[192] P. J. Phillips, A. Martin, C.I. Wilson and M. Przybocki. An Introduction To Evaluating Biometric Systems Technology. *Computer*, 33(2):56–63, 2000.

[193] A.R. Padhani and J.E. Husband. Dynamic contrast-enhanced mri studies in oncology with an emphasis on quantification, validation and human studies. *Clinical Radiology*, 56(8):607–620, August 2001.

[194] Paul Viola and Micheal Jones. Robust Real-time Object Detection. *International Journal of Computer Vision*, pages 137–154, May 2004.

[195] M. Petrou, N. Georgis, and J. Kittler. Sensitivity analysis of projective geometry 3d reconstruction. In R. Klette, S. Stiehl, M. Viergever, and V. Vincken, editors, *Performance Evaluation of Computer Vision Algorithms*, pages 255–264, Amsterdam, 2000. Kluwer Academic Publishers.

[196] P. J. Phillips and K. W. Bowyer. Special section on empirical evaluation of computer vision algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(4), 1999.

[197] P. J. Phillips, A. Martin, C. L. Wilson, and M. Przybocki. An introduction to evaluating biometric systems. *IEEE Computer*, 33(2):56–63, 2000.

[198] P.J. Phillips, H.J. Moon, S.A. Rizvi, and P.J. Rauss. The FERET Evaluation Methodology for Face-Recognition Algorithms. *T-PAMI*, 22(10):1090–1104, October 2000.

[199] Hoover P.J Flynn, A and P.J. Phillips. Special issue on empirical evaluation of computer vision algorithms. *Computer Vision and Image Understanding*, 84(1), 2001.

[200] P.J. Phillips, P. Grother, R.J Micheals, D.M. Blackburn, E Tabassi, and J.M. Bone. FRVT 2002: Overview and Summary. Technical report, Face Recognition Vendor Test 2002 (www.frvt.org), 2002.

[201] S. B. Pollard, J. E. W. Mayhew, and J. P. Frisby. PMF: A stereo correspondence algorithm using a disparity gradient limit. *Perception*, 14:449–470, 1985.

[202] I. Poole. Optimal probabilistic relaxation labeling. In *Proc. BMVC 1990*. BMVA, 1990.

[203] J. Porill, S. B. Pollard, T. Pridmore, J. Bowen, J.E.W. Mayhew, and J. P. Frisby. Tina: A 3D vision system for pick and place. In *Proceedings of the Alvey Vision Conference*, 1987.

[204] K. E. Price. Anything you can do, I can do better (no you can't). *Computer Vision Graphics and Image Processing*, 36(2/3):387–391, 1986.

[205] J. Princen, J. Illingworth, and J. Kittler. Hypothesis testing: A framework for analyzing and optimizing hough transform performance. *IEEE Transactions PAMI*, 16:329–341, 1994.

[206] H. H. Baker R. C. Bolles and M. J. Hannah. The JISCT stereo evaluation. In *Proceedings of the ARPA Image Understanding Workshop*, pages 263–274, Washington D.C., USA, 1993.

[207] Rainer Lienhart and Jochen Maydt. An Extended Set of Haar-like Features for Rapid Object Detection. In *Proceedings of IEEE International Conference in Image Processing*, volume 1, pages 900 – 903, September 2002.

[208] V. Ramesh and R. M. Haralick. Random perturbation models and performance characterization in computer vision. In *Proceedings of Conference on Computer Vision and Pattern Recognition CVPR92*, pages 521–527, Urbana-Champaign, USA, 1992.

[209] V. Ramesh and R. M. Haralick. Random perturbation models and performance characterization in computer vision. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 521–527, Urbana-Champaign, USA, 1992.

[210] V. Ramesh and R. M. Haralick. Automatic tuning parameters selection for feature extraction sequence. In *Proceedings of Computer Vision and Pattern Recognition*, pages 672–677, Seattle, Washinton, USA, June 1994.

[211] V. Ramesh and R. M. Haralick. A methodology for automatic selection of IU algorithm tuning parameters. In *Proceedings of the ARPA Image Understanding Workshop*, pages 675–687, Monterey, USA, 1994.

[212] V. Ramesh and R. M. Haralick. Random perturbation models for boundary extraction sequence. *Machine Vision and Applications: Special Issue on Performance Characterization*, 1998.

[213] V. Ramesh, R. M. Haralick, A. S. Bedekar, X. Liu, D. C. Nadadur, K. B. Thornton, and X. Zhang. Computer vision performance characterization. In O. Firshein and T. Strat, editors, *RADIUS: Image Understanding for Imagery Intgelligence*, pages 241–282. Morgan Kaufmann, San Francisco, USA, 1997.

[214] V. Ramesh, R. M. Haralick, A. S. Bedekar, X. Liu, D. C. Nadadur, K. B. Thornton, and X. Zhang. Computer vision performance characterization. In O. Firschein and T. Strat, editors, *RADIUS: Image Understanding for Imagery Intelligence*, pages 241–282. Morgan Kaufmann Publishers, San Francisco, USA, 1997.

[215] S. V. Rice, F. R. Jenkins, and T. A. Nartker. The fifth annual test of OCR accuracy. Technical Report ISRI TR-96-01, Information Science Research Institute, University of Nevada, Las Vegas, April 1996.

[216] Syed A. Rizvi, P. Jonathon Phillips, and Hyeonjoon Moon. The feret verification testing protocol for face recognition algorithms. Technical Report 6281, NIST, October 1998.

[217] J. J. Rodriguez and J. K. Aggarwal. Stochastic analysis of stereo quantization error, quantization error in stereo imaging. In *Computer Vision Pattern Recognition CVPR88*, pages 153–158, 1988.

[218] J. J. Rodriguez and J. K. Aggarwal. Stochastic analysis of stereo quantization error. *IEEE Transactions PAMI*, 12(5):467–470, May 1990.

[219] N. Roma, J. Santos-Victor, and J. Tome. A comparative analysis of cross-correlation matching algorithms using a pyramidal resolution approach. In H.I. Christensen and P.J. Phillips, editors, *Empirical Evaluation Methods in Computer Vision*, pages 117–142. World Scientific Press, September 2002. ISBN 981-02-4953-5.

[220] P. Rosin. Assessing the behaviour of polygonal approximation algorithms. *Pattern Recognition*, 36:505–518, 2003.

[221] P. L. Rosin. Edges: saliency measures and automatic thresholding. *Machine Vision and Applications*, 9:139–159, 1997.

[222] P. L. Rosin. Techniques for assessing polygonal approximations of curves. *IEEE transactions PAMI*, 19(6):659–666, June 1997.

[223] H. A. Rowley. CMU face detection. Technical report, Carnegie Mellon University, 1998.

[224] Vargha-Khadem F Connelly A Gadian DG Salmond CH, Ashburner J and Friston KJ. The precision of anatomical normalization in the medial temporal lobe using spatial basis functions. *NeuroImage*, 17(1):507–512, 2002.

[225] K. B. Sarachik. The effect of gaussian error in object recognition. *IEEE Transactions PAMI*, 19(4):289–301, April 1997.

[226] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondance algorithms. *Intl. Journal of Computer Vision*, 47(1-3):7–42, April 2002. Microsoft Research Technical Report MSR-TR-2001-81, November 2001.

[227] D. Scharstein, R. Szeliski, and R. Zabih. A taxonomy and evaluation of dense two-frame stereo methods. In *IEEE Workshop on Stereo and Multi-Baseline Vision*, Hawaii, December 2001.

[228] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *Intl. J. Computer Vision*, 37:151–172, 2000.

[229] M. C. Shin, D. B. Goldgof, and K. W. Bowyer. Comparison of edge detectors using an object recognition task. In *Proceedings of Conference on Computer Vision and Pattern Recognition CVPR 1999*, pages 1360–, Ft. Collins, CO, USA, June 1999.

[230] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination and expression databas. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1615 – 1618, December 2003.

[231] E. P. Simoncelli. Design of multi-dimensional derivative filters. In *International Conference on Image Processing*, volume 1, pages 790–793, 1994.

[232] E. P. Simoncelli. *Bayesian Multi-scale Differential Optical Flow*, volume 2, chapter 14, pages 397–422. Academic Press, 1999.

[233] S. Smith and J. Brady. SUSAN – a new approach to low level image processing. *International Journal of Computer Vision*, 23(1):45–78, 1997.

[234] S. M. Smith. A new class of corner finder. In *Proc. 3rd British Machine Vision Conference*, pages 139–148, 1992.

[235] Jenkinson M Chen J Matthews PM Federico A De Stefano N Smith SM, Zhang YY. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. *NeuroImage*, 17(1):479–489, 2002.

[236] G. W. Snedecor and W. G. Cochran. *Statistical Methods (8th Edition) Ames (IA)*. Iowa State University Press, 1989.

[237] K. Sparck-Jones and C. van Rijsbergen. Report on the need for and provision of an ideal information retrieval test collection. Technical report, Computer Laboratory, University of Cambridge, 1975.

[238] Karen Sparck Jones. What might be in a summary? In Krause Knorz and Womser-Hacker, editors, *Information Retrieval 93: Von der Modellierung zur Anwendung*, pages 9–26, Konstanz: Universitatsverlag Konstanz, 1993.

[239] H. Spies. Certainties in low-level operators. In *Vision Interface*, pages 257–262, 2003.

[240] H. Spies and J. L. Barron. Evaluating certainties for image intensity differentiation for optical flow. In 1^{st} *Canadian Conference on Computer and Robot Vision*, pages 408–416, 2004.

[241] H. Spies, B. Jähne, and J. L. Barron. Range flow estimation. *Computer Vision Image Understanding*, 85(3):209–231, March 2002.

[242] M. Winter T. Schloegl, C. Beleznai and H. Bischof.

[243] X. Tang, J. L. Barron, R. E. Mercer, and P. Joe. Tracking weather storms using 3D doppler radial velocity information. In *13th Scandinavian Conference on Image Analysis*, pages 1038–1043, 2003.

[244] J. Mossing S. Worrell T.D. Ross, V. Velton and M. Bryant. Standard sar atr evaluation experiments using the mstar public release data set. In *SPIE, Algorithms for synthetic Aperture Radar Imagery V*, volume 3370, 1998.

[245] N. A. Thacker, F. J. Aherne, and P. I. Rockett. The Bhattacharyya metric as an absolute similarity measure for frequency coded data. *Kybernetika*, 32(4):1–7, 1995.

[246] N. A. Thacker and P. Courtney. Statistical analysis of a stereo matching algorithm. In *Proceedings of the British Machine Vision Conference*, pages 316–326, 1992.

[247] N. A. Thacker, P. A. Riocreux, and R. B. Yates. Assessing the completeness properties of pairwise geometric histograms. *Image and Vision Computing*, 13(5):423–429, 1995.

[248] S. Thayer and M. Trivedi. Residual uncertainty in 3-dimensional reconstruction using 2-planes calibration and stereo methods. *Pattern Recognition*, 28(7):1073–1082, July 1995.

[249] C. Theobalt, J. Carranza, M. A. Magnor, and H. P. Seidel. Enhancing silhouette-based human motion capture with 3D motion fields. In *Proc. Pacific Graphics*, pages 185–193, 2003.

[250] K. Thornton, D. C. Nadadur, V. Ramesh, X. Liu, X. Zhang, A. Bedekar, and R. Haralick. Groundtruthing the RADIUS model-board imagery. In *Proceedings of the ARPA Image Understanding Workshop*, pages 319–329, Monterey, USA, 1994.

[251] B. Tian, J. Barron, W. K. J. Ngai, and Spies H. A comparison of 2 methods for recovering dense accurate depth using known 3D camera motion. In *Vision Interface*, pages 229–236, 2003.

[252] C.W. Tong, S.K. Rodgers, J.P. Mills, and M.K. Kabrinsky. Multisensor data fusion of laser radar and forward looking infared for target segmentation and enhancement. In R.G. Buser and F.B. Warren, editors, *Infared Sensors and Sensor Fusion*. SPIE, 1987.

[253] P. H. S. Torr and A. Zisserman. Performance characterisation of fundamental matrix estimation under image degradation. *Machine Vision and Applications*, 9(5/6):321–333, April 1997.

[254] Y. Tsin, V. Ramesh, and T. Kanade. Statistical calibration of the ccd process. In *Proceedings of IEEE International Conference on Computer Vision ICCV2001*, Vancouver, Canada, 2001.

[255] M. A. Turk and A. P. Pentland. Face Recognition Using Eigenfaces. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 586 – 591, June 1991.

[256] M. Ulrich and C. Steger. Empirical performance evaluation of object recognition methods. In *Empirical Evaluation Methods in Computer Vision*, Hawaii, December 2001.

[257] S. Uras, F. Girosi A. Verri, and V. Torre. A computational approach to motion perception. *Biological Cybernetics*, 60:79–97, 1988.

[258] P. L. Venetainer, E. W. Large, and R. Bajcsy. A methodology for evaluation of task performance in robotic systems: a case study in vision-based localisation. *Machine Vision and Applications*, 9(5/6):304–320, April 1997.

[259] A. Verri and T.Poggio. Against quantitative optical flow. In J. Michael Fitzpatrick, editor, *Proc. 1st International Conference on Computer Vision*, pages 171–180, Londodn, 1987. IEEE Computer Soc. Press.

[260] Liu W. and Dori D. The arc segmentation contest. In *Proc. of 4th IAPR Workshop on Graphics Recognition*, Kingston, Canada, September 2001.

[261] D. P. Huttenlocher W. E. L.Grimson. On the sensitivity of geometric hashing. In *IEEE International Conference on Computer Vision ICCV*, pages 334–338, 1990.

[262] D. P. Huttenlocher W. E. L.Grimson. On the verification of hypothesized matches in model-based recognition. *IEEE transactions PAMI*, 13(12):1201–1213, December 1991.

[263] S. J. Wang and T. O. Binford. Local step edge estimation: A new algorithm, statistical model and performance evaluation. In *Proceedings of the ARPA Image Understanding Workshop*, pages 1063–1070, 1993.

[264] Bruce A. Draper Wendy S. Yambor and J. Ross Beveridge. Analyzing pca-based face recognition algorithms: Eigenvector selection and distance measures. In H. Christensen and J. Phillips, editors, *Empirical Evaluation Methods in Computer Vision*. World Scientific Press, Singapore, 2002.

[265] J. Weng, T. S. Huang, and N. Ahuja. Motion and structure from two perspective views: Algorithms, error analysis and error estimation. *IEEE Transactions PAMI*, 11(5):451–476, May 1989.

[266] L. Wenyin and D. Dori. Incremental arc segmenation algorithm and its evaluation. *IEEE transactions PAMI*, 20(4):424–430, April 1998.

[267] L. Wenyin and D. Dori. Principles of constructing a performance evaluation protocol for graphics recognition algorithms. In R. Klette, S. Stiehl, M. Viergever, and V. Vincken, editors, *Performance Evaluation of Computer Vision Algorithms*, pages 81–90, Amsterdam, 2000. Kluwer Academic Publishers.

[268] L. Wenyin, Z. Su, S. Li, Y-F. Sun, and H. Zhang. A performance evaluation protocol for content-based image retrieval. In *Empirical Evaluation Methods in Computer Vision*, Hawaii, December 2001.

[269] Y. Xiong and L. Matthies. Error analysis of a real time stereo system. In *Computer Vision Pattern Recognition CVPR97*, Puerto Rico, June 1997.

[270] Taijima T Hachitanda Y Tomita K Koga M Yabuuchi H, Fukuya T. Salivary gland tumors: Diagnostic value of gadolinium-enhanced dynamic mr imaging with histopathologic correlation. *Radiology*, 226(2):345–354, 2003.

[271] M-H. Yang, D. J Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE transactions PAMI*, 24(1):34–58, January 2002.

[272] M. Ye and R. M. Haralick. Optical flow from a least-trimmed squares based adaptive approach. In *ICPR*, pages Vol III: 1052–1055, 2000.

[273] M. Ye and R. M. Haralick. Two-stage robust optical flow estimation. In *Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2623–2028, 2000.

[274] Yoav Freund and Robert E. Schapire. A Decision-theoretic Generalization of On-line Learning and an Application to Boosting. In *Computational Learning Theory: Eurocolt 1995*, pages 23 – 37. Springer-Verlag, 1995.

[275] D. Zhang and G. Lu. A review of shape representation and description techniques. *Pattern Recognition*, 37:1–19, 2004.

[276] L. Zhang, T. Sakurai, and H. Miike. Detection of motion fields under spatio-temporal non-uniform illuminations. *Image and Vision Computing*, 17:309–320, 1999.

[277] W. Zhao, R. Chellappa, and A. Krishnaswamy. Discriminant analysis of principal components for face recognition. In *In Wechsler, Philips, Bruce, Fogelman-Soulie, and Huang, editors, Face Recognition: From Theory to Applications*, pages 73–85, 1998.

[278] J. Zitova and J. Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21(11):977–1000, 2003.

[279] Smith EO Rattner Z Gindi G Zubal IG, Harrell CR and Hoffer PBl. Computerized 3-dimensional segmented human anatomy. *Medical Physics*, 21(2):299–302, February 1994.