Available online at www.sciencedirect.com
ScienceDirect
Computer Vision and Image Understanding 109 (2008) 305–334
Computer Vision and Image Understanding
www.elsevier.com/locate/cviu

Performance characterization in computer vision: A guide to best practices

Neil A. Thacker [a], Adrian F. Clark [b,*], John L. Barron [c], J. Ross Beveridge [d], Patrick Courtney [e], William R. Crum [f], Visvanathan Ramesh [g], Christine Clark [b]

[a] *Department of Imaging Science and Biomedical Engineering, University of Manchester, Manchester, UK*
[b] *Department of Electronic Systems Engineering, University of Essex, Colchester, UK*
[c] *Computer Science Department, University of Western Ontario, London, Ont., Canada*
[d] *Department of Computer Science, Colorado State University, Fort Collins, CO, USA*
[e] *PerkinElmer, Seer Green, Bucks, UK*
[f] *Centre for Medical Image Computing University College London, London, UK*
[g] *Siemens Corporate Research, Princeton, NJ, USA*

Received 4 May 2006; accepted 30 April 2007
Available online 21 June 2007

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 in literature. Most of these involve, in some way, paying greater attention to the domain of the problem and to performing detailed empirical analysis. The goal of this paper is to review 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.

Keywords: Performance assessment; Performance evaluation; Vision system design

1. Introduction

The aim of this paper is to relate and consolidate common approaches to the task of understanding or characterizing the performance of algorithms in a variety of application tasks. By 'performance characterization' 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; however, it excludes testing hardware, though clearly this is a related task.

This paper consists of two main parts: a review of selected past work in performance characterization 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 both lower-level (feature detection, shape description, *etc*) and high-level visual tasks (*e.g.*, detecting structural changes in medical images) addressed in the computer vision literature. However, the list is by no means exhaustive: the intention is

* Corresponding author. Fax: +44 1206 872900.
E-mail address: alien@essex.ac.uk (A.F. Clark).

1077-3142/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2007.04.006

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 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 [1,2]; again in the early 1990s [3]; continued by Haralick with support from DARPA and elsewhere [4–7] and by Forstner [8]; and was taken up within certain sub-communities, in particular OCR [9–11], document understanding [12,13], graphics recognition [14–20], and photogrammetry [21,22].

The mid to late 1990s saw the organization of the following workshops: ECVnet [23]; ECCV96 [24], leading to a special issue of Machine Vision and Applications [25]; DAGM'97 [26]; CVPR98 [27], a 1998 Dagstuhl workshop [28]; ICVS99 [29]; AVICS99 [30] and ECCV2000 [31], also leading to a book [32]. There were also journal special issues [33–35] and web resources: ECVnet,[1] PEIPA[2] and CMU.[3] These have evolved to a set of well-established workshop series (*e.g.*, PETS,[4] Empirical Evaluation in Computer Vision,[5] and NIST's Performance Metrics for Intelligent Systems[6]), as well as a tutorial series of the PCCV project sponsored by the European Commission.

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 materialize outside a few significant niches, notably semiconductor and pharmaceutical manufacturing. Many potential end-users remained skeptical well into the 1990s, citing a lack of robustness [36]. Recent years have seen the increasing acceptance of vision in a wider range of applications, where highly hand-crafted solutions have been commercialized. 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 behaviors, such as robot navigation and generalized learning exhibiting graceful degradation, have also been frustrated by the brittleness of the vision modules with which they have to work.

As the field of computer vision has developed, so the pool of experienced vision researchers and fresh Ph.Ds 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 that 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 acknowledgment 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 characterization has become established as one of the set of issues to be considered in computer vision research, there is still a lack of consensus on the appropriate questions to be asked. [37] outlines the different levels of analysis for biometric systems, and these also appear 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 characterized by simple parameters (such as gain and 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 (*e.g.*, 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.

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

Ramesh et al. [38] independently summarized a system engineering methodology for building vision systems with two main steps, *component identification* and *application domain characterization*, equivalent to *technology evaluation* and *scenario evaluation* respectively.

3.1. The role of quantitative statistics

We argue that characterization 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 explicit consideration of the statistical character of the input data at all: this makes it difficult to produce any 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 optimization 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. Algorithms viewed as novel in the literature must ultimately be reconcilable with quantitative statistics if they are to be widely accepted as theoretically valid. In our experience it is generally possible to make such an association, and it involves identifying the assumptions necessary to derive such methods from probability theory. 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 transform data or redefine the problem so that they do. For example, if the data have 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 behavior [39]; likewise, a binomial distribution may be converted into an approximation to a Gaussian by the arcsine transform. These are standard techniques already described in the statistical literature [40].

In addition, many algorithms also make assumptions regarding data independence. These assumptions are often overlooked as, unlike issues such as problems with outlier data, independence often does not have catastrophic implications for estimated parameters. Unfortunately, we have insufficient space to discuss this problem further but would like to note that in some cases issues of non-independence can be detected and solved using data pre-processing such as 'whitening'.

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 is by far the most promising theoretical basis for computer vision algorithm design and testing.

A factor that makes machine vision research a particular challenge is that conventional vision and statistical techniques often need to be developed and adapted in order to avoid difficulties with numerical or non-linear aspects of the algorithm. 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 believe 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 is today as a field, we devote a sizable portion of this paper to reviewing the current state of empirical evaluation for selected visual tasks.

3.2. Characterizing variation

A common observation is that algorithms perform differently when run on apparently similar data. Characterizing this variation is both essential and difficult. Underlying this difficulty is the perception that there is one unique mode of variation associated with an algorithm; however, that is rarely the case. To illustrate this, 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 the results from our algorithm might vary if we were to acquire another data set differing only due to measurement noise. For many estimation processes, this variation is already a sufficient result to allow an algorithm to be used in the design of a larger system, by matching the characteristic of the output data to the assumed behavior 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 input, 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 people under, *e.g.*,variations in lighting and pose? The resulting variation characterizes how much an algorithm's behavior will vary from one execution to another on data drawn from a population.

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 1000 images in order to conclude that one algorithm is better than another [41]?

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

3.3. Black-, white- and glass-box testing

Most of the papers in empirical evaluation workshops are aimed at addressing 'black-box' evaluations of vision algorithms. 'Black-box' means that the internal workings of an algorithm are not directly considered, only the output resulting from a particular input. 'Glass-box' is similar but allows examination of the workings of an algorithm. These both contrast with [42–44] which address 'white-box' evaluation, which involves abstracting the essence of the algorithm into a formal mathematical expression or transform, identifying a statistical model of the input data and deriving the statistical model characterizing the output. 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 [44] along with other numerical methods (*e.g.*, bootstrap [45]) perform characterization with explicit analytical statistical models. The tools available include:

Distribution propagation. The input to an algorithm (*e.g.*, 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. Care should be taken while using this tool since the approximation may be good only when the linearization and first-order error approximations are valid. Further details are presented in [46].

Empirical methods. An empirical approach to systems evaluation is described in [47]. 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 [48] for the first description of edge detection performance characterization using bootstrap). Monte Carlo methods can be used for the verification of theoretical results derived in the previous two steps.

Statistical modeling. This involves 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 extensive and encompasses methods from fields such as Bayesian statistics. As this is well-established in computer vision, we shall not discuss it further here.

Machine vision researchers may be sceptical regarding the idea that data distributions can be modelled and analysed in the way we are suggesting. However, this issue should not be considered as being distinct from the algorithm design. If we wish to be able to characterize the algorithms we design then we should ensure that high-level representations are chosen that support the quantitative analysis required. This in turn may lead to the need for research into new approaches for the characterisation of data, extending the range of tools currently available.

3.4. Assessing progress

The purpose of this paper is 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 return here to the definition used to assess the contents of a paper or thesis, that the work must make a contribution to the body of human knowledge. Based upon this, it is our belief that, where possible, research papers should contain results that 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 application areas. In particular, performance figures for any algorithm can vary according to the data set tested. Thus, quantitative results are often less useful for future researchers than some believe. One way to avoid this 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 behavior. We need to ask whether we know enough about the problem to be able to recognize 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 generally thought to be difficult—and, indeed, can work only 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; and for this, we need some agreement within the field as

to which algorithms should be used as benchmarks for comparison (''strawman'' algorithms). The specific details of exactly which algorithms are selected as benchmarks and why are less important than the acceptance that such a mechanism is necessary and a basic agreement of which are acceptable.

Developing a scientific understanding of performance requires us to understand quantitatively how our correct answers would be expected to match those delivered from the algorithm under real world conditions (*e.g.*, 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 ultimately be reconcilable with probability theory, we need to ask whether the body of theory is present which would allow the assumptions underlying algorithms 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 are 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 data sets 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 shall assert 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 encapsulates concepts such as: detection reliability (which can be derived from 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 [37] makes use of more complex metrics such as system reliability expressed as mean time between failures.

4. A review of how vision algorithm performance is analyzed

It is not feasible to review the entire vision literature in the context of performance analysis. Instead, the following subsections perform two services. First, they review essential stages of many image processing pipelines, highlighting research results that, in the authors' opinion, demonstrate good practice in assessing algorithmic performance in a way that encompasses sensing, feature detection, object localization and recognition.

Second, some specific topics are explored in some depth: coding, optical flow, stereo, face recognition, and the determination of structural differences in medical imagery. Though chosen because they reflect the authors' interests and expertise, they demonstrate that individuals and groups of researchers are already active in applying the types of approach expounded above to particular types of vision problem. We have every reason to believe that these areas are representative of the computer vision discipline as a whole.

4.1. Sensor characterization

Although much of computer vision employs straightforward color 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 can describe any sensor. Hence, it is naive to expect that a 'universal' algorithm, effective with data from any sensor, can be devised. A safer and a more scientific approach is to assume that algorithms will not work for any sensor other than that for which it was originally developed and validated, until it is demonstrated that the algorithm's assumptions still hold for data from the new sensor.

Perhaps the earliest step in describing the general characteristics of a sensor is to consider the illumination regime in which it operates. Sensors such as cameras (film, CCD, videcon, X-ray, *etc*) working with illumination from conventional sources operate in an incoherent regime, whereas systems employing stimulated emission sources such as laser stripers, ultrasound systems and radar (including synthetic aperture radar or SAR) result in coherent imaging. Between these two extremes lies partially-coherent imaging, encountered most commonly in high-resolution electron microscopy and systems using low-cost semiconductor lasers. Imagery acquired (partially) coherently should be processed by algorithms that consider both amplitude and phase, which is not the case for incoherent acquisition.

In the context of this paper, arguably the important distinction concerns the statistical distribution of noise present in images. 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 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 is not appropriate and algorithms that assume Gaussian statistics are unlikely to be reliable.

4.1.1. How is testing currently performed?

Sensor specifications are often characterized 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 digitization 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 [21,49] and occasionally by the computer vision community [50–52] but do not appear to form the basis of the majority of algorithms reported in literature.

Furthermore, systems that involve processing as an inherent part of capture will exhibit specific noise characteristics that differ from those described above; *e.g.*, images from MRI scanners exhibit a Rician noise distribution [53]. Likewise, sensors intended for visual presentation of information may employ histogram equalization, introducing discontinuities into the data distribution that may cause failures in downstream algorithms if not accounted for.

4.1.2. 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 analyzed 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 artifacts 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 [22,54] and require careful investigation.

4.1.3. Is there a quantitative methodology for the design of subsystems?

It is absolutely critical to understand the imaging process of the sensor subsystem in order 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.

4.1.4. What should we be measuring to quantify performance?

A parametrized model of sensor stability and repeatability experiments are adequate.

4.2. 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 arcs and circles [17,20,55,56]. 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 [57] famously provided a theoretical derivation based upon a particular choice of 'optimality', though the predictions he made for location performance are not observed in Monte Carlo studies [58]. 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 those by Moravec [59] and Harris and Stephens [60]. Cursory demonstration of function was given by a small number of images and by their inclusion in robotic systems. Scale-independent extensions proposed in [61] were similarly demonstrated on single images and on artificial scenes, illustrating the limitations of the original formulations and providing visual evidence that they 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 [62], moment invariants [63], SIFT [64]).

4.2.1. How is testing currently performed?

For edge and line detection, numerous papers have appeared in literature on boundary segmentation performance evaluation, *e.g.* [42,65,66]. 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, edge location uncertainty and orientation estimate distributions are derived to illustrate that, at low signal-to-noise ratios, the orientation estimate has large uncertainty. Heath et al. [67] visually

compares the outputs from various edge detectors. Cho et al. [48] was the first to use a resampling technique as a tool for studying the performance of edge detection techniques.

Most of the papers described above use simulations or hand-drawn 'ground truth' with which to compare algorithm results. Baker and Nayar [68] is unusual in that it does not require ground truth; rather, the evaluation is made by examining statistics that measure global coherence of edge points detected (*e.g.*, collinearity). Konishi et al. [69] addresses edge detector evaluation using information theoretic principles: it 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 or 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 [70,71]. By 2000 more rigorous evaluations were being reported [72–74] with ROC curves being drawn for a set of viewpoint changes across a database of images.

4.2.2. 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 difficult to design a single data set for every possible feature detector, though an appropriate simulated test data set could be constructed from a theoretical definition of the detector. The RADIUS data set with ground truth [75] has been used in some studies [76].

4.2.3. Are there data sets in common use?

Earlier papers demonstrated effects on the Lenna or Cameraman images. Image data sets are now available on the web [67]. One recent interest operator study [74] utilizes shared code made available by the original developers. A database of 1000 images, taken from a single 3-h video of a planar scene undergoing viewpoint changes, was created, yielding some 300,000 points, and subsequently made available on the web. The study was sufficient to rank consistently the operators but the statistical properties of the data and the relative contribution of various error sources are not discussed further.

4.2.4. Are there experiments that show algorithms are stable and work as expected?

Stability has been explored in terms of some image properties but has been related to a theoretical prediction such as error propagation only in [77]. For well-defined structure, Monte Carlo or resampling techniques could be used if an appropriate definition of the task was available.

4.2.5. Are there any strawman algorithms?

For edge detection, the commonly-cited algorithm is Canny, though implementations of it 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 scenarios such as simple measurement.

Strawman edge detection code, data and scoring tools are available [78].[7] For corner detection, [60] works well and is available[8]; the SUSAN code is also available.[9] As mentioned above, corner detection software is now being shared between research groups [74].

4.2.6. 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 that the image formation process be explicitly taken into account. 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.

4.2.7. 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 [42]. 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 (*e.g.*, vertices).

4.3. Shape- and grey-level-based object localization

The localization and recognition of a known object in an image are tasks that have received a great deal of attention. The two tasks are often linked but we shall try to consider them separately since, from a performance evaluation point

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

of view, different information is sought, with different failure modes and performance metrics. Localization involves estimation of the transformation between the image coordinate frame and the object coordinate frame; this may range from a simple translation vector to full six-degree-of-freedom transformations, depending on the nature of the object and the scene. Our remarks in this section are therefore also relevant to tasks such as image alignment or registration [79].

Techniques proposed range from template matching, Hough transform and 3D wire-frame location, to techniques for deformable objects such as snakes [80], Active Shape Models (ASMs) and Active Appearance Models (AAMs) [81], in which a low-parameter model of an object is aligned over a corresponding image region by the minimization of some cost function. Localization techniques often comprise a representational stage (operating on pixels or derived primitives[10]) and a matching stage based on a cost function.

4.3.1. How is testing currently performed?

Novel localization techniques are generally demonstrated on marked-up data and with localization error expressed in terms of image-plane error. The main issues affecting localization performance include sensor noise (resulting in imprecise image plane feature location), occlusion (missing features) and clutter (spurious non-object features). Lindenbaum [83,84] examined both localization and indexing performance to model the effect of these three factors, and added object model self-similarity to provide estimates of performance bounds. In [85], a black-box comparison of five object localization algorithms 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 transformation, from a commercial imaging library and shared code.

The precision of 3D pose estimates also appears to be strongly dependent on object geometry and viewpoint. In [86] 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 [87]. Haralick studied the six extant formulations for 3-point pose estimation and revealed wide variation in localization accuracy according to the order in which the calculations are performed [88]. Alternatively, for numerically stable techniques, a study of four optimization-based object location techniques on synthetic range data suggest that this choice does not appear to play a significant role [89].

Localization plays a role in mobile robotics [90,91]. Performance of closed-loop pose estimation and tracking of 3D polyhedral objects was reported in [92]. In one co-operative study of co-registration, blinded ground-truthed image sets were distributed [79], though this raised technical and organizational issues as discussed in [93].

4.3.2. Is there a data set for which the correct answers are known?

Image sets with known transformation may be generated either from known mechanical motion [85] or reference marks [79]. Errors in these independent estimates are not given. Image synthesis tools have also been popular for providing synthetic images and ground truth [88,89,92].

4.3.3. Are there data sets in common use?

For some tasks, such as face localization [94] and graphics recognition [13,19], researchers use existing data sets. For other tasks, data sets tend to be created specifically [79,85,86]. The MPEG-7 community has provided a shape data set which has been used in some studies, though others have questioned its quality [95,96].

4.3.4. 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 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 optimization or as a Hough transform; these showed good agreement between predicted and observed object location accuracy (covariance) [97].

4.3.5. 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 detail. There are many variants of ASM and AAM algorithms in use across the community, including a publicly available source at DTU[11] but there does not appear to be any code in common use. There are isolated instances of code sharing [85].

4.3.6. Is there a quantitative methodology for the design of algorithms?

In principle, all such techniques are directly reconcilable with likelihood though the assumptions regarding distributions and resulting accuracy of localization are not tested. Indeed, researchers appear to be working with similarity

[10] See [82] for a recent review of shape representations schemes.

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

measures (*e.g.*, 'mutual information') which have only recently been explicitly reconciled with corresponding quantitative statistical assumptions [98], impeding progress in this area.

4.3.7. What should we be measuring to quantify performance?

Since localization is essentially an estimation task, algorithms should provide quantitative estimates of location and shape parameters complete with error covariances. Major comparative tests such as [79] have limited predictive power without accompanying confidence measures (but see [99]). These should be confirmed on test data sets by comparing predicted error distributions with practical performance. This would produce subsystems which were capable of providing all necessary salient information for use in a larger system.

4.4. Shape-based object indexing: recognition

As with localization, a similar range of model-free and 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* which determine the appropriate representation (*e.g.*, curves or 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 complete *a priori* segmentation of a 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 localization uncertainty [100]. The alternative is to try to select a more stable feature measure. For example [87] 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 recognition task. However, as Maybank [101,102] pointed out, under realistic conditions of measurement noise, the indexing capacity is likely to be limited to some 20 objects, although [103], with an alternate formulation of the feature point PDFs, suggests more optimistic results.

4.4.1. How is testing currently performed?

Testing is normally performed on a set of images and a small set of objects, and measurements are in terms of true detection rates or confusion matrices. Early work on performance issues involved the verification of matches [104] and the sensitivity of the hashing [105]. 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 set up the recognition threshold automatically so that a given false alarm rate can be met by the system. This was extended in [106] to include occlusion. Bounds on indexing performance were established in [83,84] in the presence of uncertainty, occlusion and clutter.

Sarachik [107] 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 false negative and false positive metrics in the form of an ROC curve. From this work, it appeared that uniform clutter models underestimate error rates; but it was shown that detection thresholds can advantageously be set according to estimates of feature density to minimize error rates. In other words, feature-dense regions require more evidence to attain the same confidence level. Knowledge of the object database allows performance to be optimized further.

Shin et al. [108] studied object recognition system performance as a function of the edge operator chosen at the first step of the recognition system, concluding that the Canny edge detector was superior to the others considered. Further evaluation of recognition of articulated and occluded objects using invariance features applied to SAR data was carried out in [109] using a simulator to produce ROC curves for varying orientation angle. Boshra and Bhanu [110] predicted performance bounds using synthetic and real SAR data from the public MSTAR SAR data set [111].

4.4.2. Is there a data set for which the correct answers are known?

Generally, the correct answer here is knowledge of scene contents, though unfortunately pixel-labeled segmentation does not have a unique answer due to the variety of objects and sub-components that may require identification. The definition depends on the intended task (scenario). Nevertheless, labeled data sets have been collected and made available, including the following. One popular source of images is the Corel image data set. Although large ($\sim 10^6$ images) and popular with the information retrieval community, it appears to have the following 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 recognition. An initiative by Microsoft proposed a selection of 10,000 images from the Corel data

set labeled with 100 low-level categories and new metrics [96]. Similarly, the UCID subset has been proposed.[12] The COIL-100 database of labeled everyday objects from Columbia University is an alternative, as is the data set from Washington University,[13] and the SOIL-47 color image data set from Surrey. However, published works on these data sets by groups other than their originators could not be found.[14]

4.4.3. 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 [107,108] thus precluding the possibility of results that will transfer to other domains or for comparison with other algorithms. Whilst large data sets are used (*e.g.*, CMU face detection [112]), it is not clear that they span the space of possible data densely enough to permit good estimation of false detection rates.

4.4.4. 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 have been shown to be useful in this respect [113].

4.4.5. Are there any strawman algorithms?

Fourier descriptors of curvature and moment analysis are common in the literature and available in many libraries. Various ASM and AAM algorithms are in use, though code sharing is limited. Geometric histogram code is also available.[15]

4.4.6. Is there a quantitative methodology for the design of algorithms?

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 widely studied as a tool for recognition [114]. In the case of 2D sample histograms, this can be done using cosine measures, which can be related to Poisson statistics for large histograms [115]. This can be combined with proofs of theoretical completeness of representation (*i.e.* no data are lost) to make statements regarding use of information [100]. Detailed work has also been done to investigate the stability of geometric invariance based indexing schemes [101,102].

4.4.7. What should we be measuring to quantify performance?

Recognition performance for individual models may be described using a true detection and false detection metric.[16] 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 below.). 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 theoretical understanding of the behavior 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.

4.5. Lossy image and video compression

4.5.1. 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 when the ultimate target is the human vision system. There are standardized ways of carrying out and analyzing subjective tests based around the use of forced choices in comparisons of images or video clips.

4.5.2. Is there a data set for which the correct answers are known?

Coding is unusual in that any image or sequence can be used: the original data are always the correct answer.

4.5.3. Are there data sets in common use?

There are several widely-used images (*e.g.*, Lena/Lenna; see [116] for an interesting discussion) and sequences (*e.g.*, salesman) used for basic comparison. Each coding technique is developed in conjunction with an agreed 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, while others

[12] vision.doc.ntu.ac.uk/datasets/UCID/ucid.html.

[13] www.cs.washington.edu/reseach/imagedatabase/groundtruth.

[14] More recently, the ImageCLEF benchmark of some 20,000 images has come to the fore; see eureka.vu.edu.au/~grubinger/IAPR/TC12_Benchmark. html.

[15] www.tina-vision.net.

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

are chosen because they are typical of common classes of imagery or video.

4.5.4. 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.

4.5.5. Are there any strawman algorithms?

Coding is almost unique in that significant effort has been (and continues to be) put into developing international standards: the ITU H.2nn series and the JPEG and MPEG series, for example; see [117] for an overview. The development of data coding techniques involves the development and refinement of algorithms that are distributed amongst the community.

4.5.6. Is there a quantitative methodology for the design of algorithms?

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.

4.5.7. What should we be measuring to quantify performance?

The human visual system is the ultimate target for almost 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.

4.6. Differential optical flow

This section discusses the measurement of 2D and 3D optical flow as measured from first and second 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 data sets (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 [118]. 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 [118]. Other types of 3D data include radial velocity from Doppler storm weather data sets [119], gated MRI data, which can be used to compute 3D volumetric optical flow [120] and multi-view video silhouette sequences [121].

The 2D/3D derivatives are usually computed by repeated application of low-pass and high-pass filters, for example [122]. Parametrization of the data by B-splines and then analytic differentiation of these splines has also been used [123] in a quantitative optical flow calculation. 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 [113,120] or globally via a regularization [120,124].

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 modeled. 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 [125–128], 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 modeled explicitly. Finally, all objects in the scene are rigid; no shape changes are allowed—though this assumption is often relaxed to local rigidity. However, making this assumption assures that optical flow actually captures real motions in a scene rather than expansions, contractions, deformations and shears of various scene objects.

4.6.1. How is testing currently performed?

Optical flow can be evaluated either qualitatively and/or qualitatively. Error can be measured as mean error in magnitude or direction [118], or an angle error measure capturing both the magnitude and direction deviation from ground truth [129].

4.6.2. 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 [130]. 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 [131]. One example of 2D and 3D

synthetic data is sinusoidal images/volumes [120,132] (which are perfectly differentiable) and thus might be considered 'gold standard' data.

4.6.3. Are there data sets in common use?

Barron et al. [132] performed a quantitative analysis for nine optical flow algorithms using the translating/diverging tree sequence [129], the Yosemite fly-through sequence and several translating sinusoidal/square image sequences (and this is a two-frame optical flow method). Otte and Nagel [133] have made a calibrated real image sequence. These data have known ground truth and are publicly available.[17] Although these data sets have been available for a considerable length of time it probably could not be said that these data sets are accepted as any sort of *de facto* standard.

4.6.4. Are there experiments which show that the algorithms are stable and work as expected?

Researchers have derived covariances for optic flow estimates; among them [134] 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 et al. [135] also uses a multi-scale framework and estimates the uncertainties for the optic flow estimate by utilizing the variable bandwidth mean-shift 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 a variable-bandwidth mean-shift filter, are used as the final refined estimate for optic flow.

Other researchers have performed covariance propagation for optic flow [77,136]. Here, we are interested in the propagation of covariance matrices for random input perturbations to the covariances associated with the final computed results. There may be an inherent bias in many optical flow estimators because of the fact that the regularization assumption (*e.g.*, the smoothness assumption in the flow field [124]) is not necessarily correct with all data sets. In other words, the true underlying smoothness constraint is an unknown and the estimation framework naturally has biases. More recent works [137–140] seek to explain perceptual illusions through the estimation framework bias.

Ye and Haralick [141,142] propose a two-stage optical flow algorithm, using least trimmed squares followed by weighted least-square estimators, the first stage of which takes into account poor derivative quality. Nestares et al. [143] used an estimate of optical flow and its covariance at each pixel in a likelihood framework to extract confidence measures of the translational sensor parameters.

4.6.5. Are there any strawman algorithms?

The 'old' optical flow algorithms are still quite effective. There are now newer better, algorithms but algorithms such as [113,124] and, to a lesser extent, [144] and [145] are still good if their underlying assumptions are satisfied. They are readily accessible to researchers and the code is available [132].[18] Many new algorithms have appeared in the literature and all claim to give better optical flow results compared to those in [132]. In particular, [146,147] have the best quantitative results for the Yosemite fly-through sequence. Still, often the results of newer algorithms are only marginally better and the codes not generally available. Some of the classical 2D algorithms also allow simple extensions into 3D.

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. To date, most of the following assumptions have been implicitly made for differential optical flow.

- The data have 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 [148] (which has a reduction factor of $\frac{1}{2}$) may allow differentiation. An implementation of the approach of [148] is described in [149]. Brox et al. and Papenberg et al. [146,147] also used a pyramid with a reduction factor in $[0.8, 0.95]$ to handle larger motions and overcome poor temporal differentiation for two-frame optical flow.
- We assume the input noise in the images is zero-mean IID Gaussian, $N(0, \sigma^2)$. Most, but not all, sensors satisfy this assumption. Algorithms using least squares or total least-squares then give a theoretically optimal solution.
- We assume local translation. If the sensor motion has a rotational component we assume that it can be approximated by one to three small local translations.
- For first-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 used in the subsequent optical flow calculation. Spies and Barron [150] showed that the sum of the normalized squared errors follows a χ^2 distribution.

[17] ftp://ftp.csd.uwo.ca:pub/vision and i21www.ira.uka.de/image_sequences.

[18] The algorithms of Nagel and Uras *et al.* use second-order intensity derivatives which are often difficult to measure accurately. Indeed, Horn's and Schunck's use of first-order intensity derivatives is effectively a second-order method because their smoothness constraint uses derivatives of image velocities, themselves constrained by first-order intensity derivatives.

4.6.6. What should we be measuring to quantify performance?

The information available in a pair of images of the same scene is not sufficient to determine unambiguously point-to-point correspondences with uniform accuracy at all locations [151]. 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 estimate is. An optical flow field with no error estimates cannot be used confidently 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 into a larger vision system.

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

- 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 [152]. If the metric adopted is small in comparison to the expected errors then both the optical flow and the reconstruction method are good; otherwise, one cannot know with certainty if one 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 in the next image. We can compute the intersection of predicted and actual storm ellipsoids in the next image [119] 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 employing feature-based [153–155] or area-based [156,157] measures, often associated with sparse and dense stereo, respectively. Dense depth estimation methods require supplementary interpolation schemes to generate complete depth maps from data with regions of low information content.

4.7.1. How is testing currently performed?

The two approaches have been the subject of direct comparison papers [158–160] as well as coordinated competitions [161,162]. These studies have shown that they generally give good performance for well-behaved scenes with sufficient information content. A typical recent paper [163] uses a standard ground-truthed image pair to validate various improvements to a re-implementation of a basic algorithm using a set 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, the importance of understanding the noise on the input data and the bias in standard techniques have been pointed out [164,165]. This topic was studied further in [166] and developed yet further by considering error propagation for structured lighting systems [167]. In an unusual paper, [168] uses knowledge of errors in the image data, propagating them to generate confidence ranges in the depth estimation and to reject unreliable estimates.

In [169], a set of 16 objective error metrics are proposed to cover aspects of reliability such as badly-matched pixels yielding depth estimates outside acceptable ground truth. Special attention is given to the types of regions involved, so that texture-less, occluded and depth discontinuity regions are treated separately. The gaps in re-projected images for unseen viewpoints were also used for subjective assessment of algorithm performance.

4.7.2. Is there a data set for which the correct answers are known?

The subject of data sets for stereo has led to considerable demand for accurate ground truth. It has also fueled a debate regarding the density and source of such data—whether mechanically derived, manually-annotated, or independently measured (laser rangefinder, by structured lighting, *etc*).

Maimone and Shafer [170] proposed a useful taxonomy of data sets which categorizes them in terms of the extent to which they use real or synthetic data. Synthetic data, with added noise, in the form of indoor corridors are available,[19] though [169] 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.

4.7.3. 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 data sets include: the Stuttgart ISPRS Image Understanding data sets[20] [161], JISCT stereo images[21] [162], the INRIA Syntim stereo databases,[22] various data sets from CMU[23] and the Middlebury dense

[19] www-dbv.cs.uni-bonn.de/stereo_data.
[20] ftp://ftp.ifp.uni-stuttgart.de/pub/wg3.
[21] ftp://ftp.vislist.com/IMAGERY/JISCT.
[22] www-rocq.inria.fr/~tarel/syntim/paires.html.
[23] www.ius.cs.cmu.edu/idb, www.cs.cmu.edu/~cil/cil-ster-html.

stereo data set[24]; the latter two of which include multi-baseline data. Furthermore, Oxford and IN-RIA have made available uncalibrated sequences of Valbonne church[25] [153,154] which can serve as redundant data sets as proposed in [171].

The Middlebury study [169] 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 six test image sets with multiple quasi-planar textured surfaces is established and used to examine the sensitivity of pixel-by-pixel depth estimation reliability to 20 algorithms and parametrizations. 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 a website. The stated intention is to extend the scheme to more complex scenes, including texture-less and more complex surface geometries.

Despite all these efforts, 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.

4.7.4. 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 [172]. This can be tested with simulated data in order to show that the quantitative estimation of depth (for correct matches) has the expected behavior [173].

The collection of additional redundant images of a scene has been proposed [171] to permit self-consistency checks to characterize 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 is available in data sets that provide multi-baseline data (*e.g.*, the Middlebury data sets).

4.7.5. Are there any strawman algorithms?

The creation of the Middlebury website, which provides a set of modules, is a major step forwards. Prior to this, implementations other than the TINA system (which contains feature- and area-based approaches) were not made publicly available.

4.7.6. 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 [156]. Some aspects of the quantitative performance of these algorithms has been addressed using empirically-determined match and mis-match distributions providing a white box analysis of a feature-matching algorithm [174] 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 [175]. The authors of [176] and [177] investigated the stability of the fundamental matrix, a common intermediate representation for relating camera configuration to image plane and disparity. The work of [163] makes some progress in improving performance at discontinuities, but no overall framework for quantitatively validating design decisions has been widely accepted.

4.7.7. 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 quantization error [178] and 3D error [179–188]. Whilst early work emphasized accuracy in the 3D world frame, the $1/Z$ depth-dependence on this measure 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:

- *Reliability*: The ability to identify suitable correspondences, measured in terms of the number of features recovered with correct (usable) geometry.
- *Accuracy*: The geometric precision of correctly-recovered features. Both accuracy and reliability require quantification and are sufficient for feature-based algorithms [189].
- *Smoothness*: 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 an explicit surface model for known object shape or an implicit model hidden within the algorithm (see the discussion of shape-based object indexing above).

Evaluation of feature-based matching algorithms is relatively straightforward. For dense stereo, 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 and to test the performance of algorithms. Algorithms for dense stereo that embed interpolation within the matching process are unlikely to produce results that correspond to the correct surface model, thus requiring the almost impossible task of separate quantification for characteristic surface types. In fact, this issue has much in common with the problems faced by estimation methods for dense optical

[24] www.middlebury.edu/stereo.
[25] www.robots.ox.ac.uk/~vgg/data1.html.

flow, and most of the same arguments restricting quantitative use of dense data still apply [151].

4.8. Face recognition

A practical face recognition algorithm must first detect faces and then recognize them, so performance characterization studies sometimes separate detection and recognition. For example, in the FERET evaluations [190] 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 [191,192]. On the other hand, to characterize the best a particular recognition algorithm can do, absent of 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 does not permit us to cover best practices for performance characterization of face recognition algorithms at the level of detail covered in [37,190,193] and serious readers are strongly encouraged to review these papers. Here, we 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 significantly different in their particulars. These two tasks are identification and verification [37]. Identification is the task of naming a person from a novel image of that person [190]. Verification is the task of deciding if a person is who they claim to be based upon a novel image [193,194]. Several documents provide important background for evaluation of face verification algorithms; one is [193], which covers many aspects of analyzing and comparing different biometric verification systems, including face. In particular, there are subtleties associated with some measures that space does 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 [190,191] are carried out and is consistent with the offline approach to match score-generation advocated in [193]. This abstraction allows analysis to be decoupled from the actual running of the algorithms. Typically, algorithms are run over all imagery included in an evaluation and a single large similarity matrix 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 arises in contexts such as airport passenger screening. There are examples of watch-list evaluations [192] but their evaluation protocols are not as mature as for identification and verification, and we shall not say more about them here.

The following paragraphs review first face detection and localization, and then face identification, as these are quite separate tasks employing different algorithms.

4.8.1. Face detection and 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, secondly, when found how accurately are they localized? Most performance characterization focuses on detection rather than localization, and hence detection rates, false alarm rates [195], and more generally ROC curves are reported [196,197]. Yang et al. [195] is an excellent resource, and there is a related website[26] that provides pointers to face detection data and algorithms as a supplement to the journal article. Compared to analysis of detection, quantitative evaluation of localization behavior

[26] vision.ai.uiuc.edu/mhyang/face-detection-survey.html.

for algorithms is less common and will not be addressed further here.

4.8.1.1. How is testing currently performed?. Face detection is tested in essentially two related ways. One 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-faces chips. The alternative, and somewhat more realistic, way is to pass the detector over large images containing possibly many faces, and score how many faces are detected, how many are missed, and how many detections are false alarms. This later 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.

4.8.1.2. 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.[27] This is divided into test sets A, B, C and the rotated test set. 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.[28] This includes 19×19-pixel image chips of faces and non-faces. The data set comes divided into training and test sets, with 2429 faces and 4548 non-faces in the training set and 472 faces and 23,573 non-face images in the test set.

4.8.1.3. Are there data sets 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.

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

4.8.1.5. 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 of [198] using the AdaBoost [199] learning algorithm. There is an implementation of an AdaBoost frontal face detection algorithms with some refinements [197] available as part of the OpenCV distribution.[29] In [195], detection and false detection rates are compared for nine different face detection algorithms, including [200], which is another possible strawman algorithm. However, while there is public code for the basic algorithm,[30] the code is not packaged to work on face detection specifically. A third possible strawman algorithm is the SvmFu SVM system developed by Rifken[31] and compared with SNoW both in [195] and [201]. The MIT CBCL data are already in the format required by the SvmFu algorithm.

4.8.1.6. 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 that lays down the foundations for systematic application of these approaches. There are arguably as many methodologies as there are distinct learning approaches [198,202,203].

4.8.1.7. 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 as to 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 above mentioned FERET protocol [190] established a common framework for comparing face identification algorithms. Subsequent large scale evaluations include the Face Recognition Vendor Tests [191,192].

4.8.2.1. 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.*

4.8.2.2. Is there a data set for which the correct answers are known?. Good identification experiments require large galleries of human subjects: *e.g.*,96% correct versus 92% on 25

[27] vasc.ri.cmu.edu//idb/html/face/frontal_images/index.html.
[28] fiz:cbcl.mit.edu/cbcl/software-datasets/FaceData2.html.
[29] sourceforge.net/projects/opencvlibrary.
[30] l2r.cs.uiuc.edu/~cogcomp/ and follow link to software.
[31] five-percent-nation.mit.edu/SvmFu/.

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, video imagery is available.

4.8.2.3. 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 useful and competitive data set in terms of the total number of subjects. However, the Notre Dame data set is quickly becoming a useful alternative, with a much higher number of replicate images per subject.

4.8.2.4. 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: same day images with different facial expressions; images taken under changing illumination; and 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, *e.g.* lighting, that have attracted particular attention and for which analytical models exist [207]. 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 [206], data finely-sampled across different illuminations settings are 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 [208,209]. We think such modeling is important and expect to see more of it in the future.

4.8.2.5. Are there any strawman algorithms?. A standard implementation of a Principal Components Analysis (PCA) based algorithm, sometimes called Eigen-faces [210], is the clear choice as a strawman algorithm. A standardized ANSI C implementation is available as part of the CSU Face Identification Evaluation System [211]. 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 well in many cases [211]. 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 principal components analysis and linear discriminant analysis algorithm (PCA+LDA) based upon [212]; a Bayesian intrapersonal/extrapersonal classifier (BIC) based upon [213]; and an

Table 1
Common data sets used for face identification evaluation

Data set	Source	Approx. number of	
		Subjects	Images
FERET [190] www.itl.nist.gov/iad/humanid/feret/	NIST	>1000	>4000
UND Database www.nd.edu/~cvrl/UNDBiometricsDatabase.htm	University of Notre Dame	>500	>30,000
UTD Database [204] www.utdallas.edu/~otoole	University of Texas, Dallas	284	>2500
BANCA [205] www.ee.surrey.ac.uk/banca/	Surrey University	208	>5000
AR Face rvl1.ecn.purdue.edu/~aleix/aleix_face_DB.html	Purdue University	126	>4000
PIE Database [206] www.ri.cmu.edu/projects/project_418.html	Carnegie Mellon University	68	41,368

elastic bunch graph matching (EBGM) algorithm based upon [214].

4.8.2.6. Is there a quantitative methodology for the design of algorithms?. Face identification falls into the theoretical category of labeling 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.[32] However, the most notable sources of image variability are well-studied; and underlying models are used either to guide algorithm design or, actively, as a part of an algorithm. Notable examples include illumination modeling [207] and 3D model-based viewpoint correction [192,215]. Finally, while in no way special to face identification, most algorithm design follows a process of successive refinement as representative training and test data are used to isolate and correct weaknesses.

4.8.2.7. 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 capture explicitly 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 McNemar's test [216]. Monte Carlo resampling may be used to capture other sources of variability, including choice of probe image [217] as well as probe and gallery images [218].

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 [208] and [209] 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. Situations where such comparisons are commonly made are for example where an individual has had several scans of a particular type (modality) over periods from minutes (*e.g.*, for detection of the passage of contrast through a tumour) to months or even years (*e.g.*, for measurement of volumetric change in brain tissue due to dementia) or where two or more individuals have each had a single modality scan.[33] There are particular difficulties involved with analyzing medical images that include: modality-dependent noise characteristics; relatively low resolution the common modalities (*e.g.*, magnetic resonance imaging typically acquires $\geqslant 1\ \mathrm{mm}^3$ voxels); wide variation in scan quality depending on acquisition hardware, scanning protocol, scan-related image artifacts, patient cooperation (or lack of it); normal physiological processes; the need for robustness where image analysis is used to support clinical decision-making; and the small size of structures of interest such as lesions or important anatomic areas (*e.g.*, 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; and 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 tumor 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: identify corresponding features of interest in the images; determine a mapping between the images, usually by some form of image registration; and compute statistics that summarize 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 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 [219,220]. For coverage of segmentation and feature detection see [221,222]. A recent technical survey of algorithms for change detection in images in general may be found in [223].

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

[33] 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.

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,[34] 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 [224]. 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 [99] do not 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 normalization' 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 normalization (which must not introduce anatomical bias but has to preserve real structural differences [225]) 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.*, [99,226]).

A slightly different case is dynamic contrast imaging where a time-series of images is 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 [227]. Analysis of such time-series to identify tumors 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.[35] To analyze 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 [228]. 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) [229], the changes in image intensity caused by shifting tissue boundaries are transformed into a volume change. Other authors use an assumption of voxel intensity linearity between two registered scans in a more automated algorithm [230]. In a complementary approach known as *Structural Image Evaluation using Normalization of Atrophy* (SIENA) [231], 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 [228]. 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 [225,232]. Ashburner and Friston [233] distinguishes 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.

4.9.1. How is testing currently performed?

A variety of testing strategies are employed but there are few standardized 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 [232,234]. Some researchers have investigated the effect of parameter choices on the spatial normalization (registration) component of the algorithms by examining co-localization of manually defined landmarks [235]. 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.*, [236,237].

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

[35] In practice data are aggregated over several heartbeats to obtain higher-quality images. Motion due to breathing can be compensated for in most subjects by breath-holding during the image acquisition.

4.9.2. Is there a data set for which the correct answers are known?

Both real and digital test objects ('phantoms') are often used. Simple phantoms are constructed commercially for quality assurance of imaging systems. Phantoms suitable for testing image analysis techniques must have a realistic appearance when imaged and must be constructed of materials that give acceptable contrast, and in some cases are compatible with more than one imaging modality [238]. 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 [239], or carefully labeling acquired images to form a digital model as in the Zubal phantom[36] [240]). In applications that measure structural change, some way to represent realistic shape and volume change or growth processes must be found. Camara et al. [241] 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. [242] used a biomechanical model of the breast to simulate the effect of plate compression, needle-biopsy, muscle relaxation *etc.*, during dynamic contrast imaging.

4.9.3. Are there data sets in common use?

A variety of data sets is used to test algorithms, although they tend to be application-dependent and not always freely available. The Center for Morphometric Analysis has set up the Internet Brain Repository[37] to make available sets of expertly-labeled 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 [243]. The Montreal Neurological Institute provides the BrainWeb digital brain phantom [244], which 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. More recently a further 20 individual simulated brain volumes have been provided.

4.9.4. 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 that 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 the 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 A_1 and A_2 within a few minutes or hours of each other. The measured changes between A_1 and A_2 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 characterize the stability of algorithms. Unfortunately there are so many sources of variation in medical imaging applications that it 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.

4.9.5. Are there any strawman algorithms?

There are no completely generic strawman algorithms. Although some advances in algorithms are simply 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 to extract directional information from tissue structure) have required the development of novel analysis algorithms that cannot easily be referred back to older methods.

4.9.6. 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, assumptions are made about the intensity ranges associated with particular tissue types. In the case of X-ray CT, the images provide direct quantitative information about the 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 [245]. Algorithms often

[36] noodle.med.yale.edu/zubal/index.htm.
[37] www.cma.mgh.harvard.edu/ibsr/.

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 normalize acquired images using a physics based acquisition model to aid analysis (*e.g.*, the HINT representation in mammography [246]). 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 [99].

4.9.7. 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 cannot 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. Some recent work has been done to integrate previously distinct steps in the image analysis pipeline to put the analysis on a firmer statistical basis *e.g.*, [247].

This problem has much in common with issues relating to the quantitative estimation of optical flow, object localization 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.

4.10. Summary

Here, we attempt to collate the answers to the key questions for each of the visual tasks and draw some conclusions. Table 2 summarizes the response as a clear 'yes' or 'no' where the evidence supports this. In some cases only a qualified 'yes' (indicated by parentheses) can be justified because there are only a few examples or we believe that consensus has not been reached. From the Table we can make several observations. First, there are marked variations in the use of performance characterization techniques by researchers across the various areas of vision considered, and hence presumably in the discipline as a whole. Some tools are used only in some areas, though they could usefully find application in other areas too. The biometric best practice guidelines [193] is probably the best-developed overall guide at present.

There now appears to be fairly widespread use of annotated data sets across most visual tasks, except (curiously) object localization. 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 [248] and more recently for fingerprint biometrics [249], and this may be expected to spread to other tasks. Nevertheless, concerns regarding data set size and coverage remain despite some work on data set design [41,250,251], even though the problems are quite well-understood within fields such as text *corpora* [252–255].

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 [169] and corner detection [74], though most areas are now sharing data. When organized around workshops (the PETS series[38]) or specific tasks (FERET/FRVT [256]), this kind of work appears to be leading to cooperation and accelerated technical progress, 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 [37] 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 data sets. There is convergence towards common evaluation metrics and presentation formats, most notably ROC curves, 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, *e.g.*, in tracking [257–260], and numerical metrics are still not uniformly defined.

Regarding experiments to show that algorithms work 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 generalization protocols 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 limited [83,84], making comparisons between papers very difficult.

[38] visual surveillance.org.

Table 2
Summary of answers to the five closed questions

	Is there a data set for which the correct answers are known?	Are there data sets 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 characterization	Yes	Yes	Yes	N/A	Yes
Video compression	Yes	Yes	Yes	Yes	No
Feature detection	(Yes)	Yes	(Yes)	Yes	(Yes)
Object localization	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)	(Yes)	No
Medical structure	Yes	Yes	Yes	(Yes)	(Yes)

The majority of evaluation experiments still appear to fall into the category of black box analysis—testing on data sets and recording of results—with relatively few attempts to utilize these results to predict performance. Building explicit white-box models of performance [42,44] is a strong idea as it allows the use of explicit modeling to link the properties of the data with performance, and thus gives a means to predict performance on an unseen data set based only on the properties of that data set.

Evaluation protocols have been published and taken up, 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. 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, Ph.D. toolkits) is starting to play a role, despite some quality issues.

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 optimality criteria [57] and Spies' [150,261] studies in optical flow. Likewise, there are white-box-like analyzes of: the behavior of intermediate representations in object recognition (modeling the Hough transform [114] and an invariance scheme [101,102]); types of variability in face recognition (*e.g.*, illumination [207] and 3D viewpoint [192,215]); and stereo (stability of the fundamental matrix [176,177], and match and mis-match distributions [174]).

A firm 'yes' in Table 2 needs valid approaches and appropriate (correct) solutions. An unambiguous, scientific solution of a particular visual task requires a 'yes' to every question. This goes to the core of making scientific progress in this area: the fact that this column has so few unambiguous 'yes' answers suggests that we have only started to lay the foundations that will establish machine vision as a mature scientific discipline. 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. In other words, all the elements needed to achieve real progress in vision as a scientific discipline appear to be in the literature already but there seems to be a lack of motivation to form scientific conclusions.

5. Discussion and future directions

We readily acknowledge that there are many other subject areas of active research in computer vision, though space limitations prohibit their discussion here.[39] However, the subject areas reviewed form a representative sample and show the gamut of characterization techniques that have been applied. We have identified key questions as those which will provide a scientific advance for the area and illustrated them on several topics. The best practice we advocate is anything that can be shown to help answer these questions—in particular, research that conclusively answers questions regarding the most appropriate way to analyse data, and the information that fundamentally appears to be required to extract certain forms of information.

One facet of the vision development process that has not been considered explicitly in the above review is that of software validation, *i.e.* ensuring that the software implementation of an algorithm correctly instantiates its mathematical foundation. We consider how quantitative assessment techniques can aid this task in the following paragraphs. This is followed by a more general discussion of how quantitative techniques can form an inherent part of the algorithm development process.

5.1. Software validation

In implementing a particular algorithm as computer code, the question arises about the correctness of the code. Förstner [262] 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

[39] A more complete bibliography of some 1000 evaluation-related papers can be found at peipa.essex.ac.uk/benchmark/bib/.

the covariance calculation, such as non-linearities [5,46]. Before we can claim that the software is fit for use, we also need to ensure that the assumptions made in any Monte Carlo simulation are consistent with the properties of real data.

Little progress seems to have been made regarding the general problem of changing image domain, *i.e.* being able to predict algorithm behavior 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 intended use of the algorithm. Ultimately, we believe that this cannot be addressed solely by the use of isolated sets of test data. In fact, we contend that what is required is a simulation of the imaging system and its environment. Such simulations must identify the key factors associated with data variability so that the expected variation in computed results can be assessed. This simulation may be in the form of an empirical program, using random numbers to explore the space of possible inputs and output responses, but could also be done analytically for cases where such a mathematical treatment is possible. We should aim to characterize any process to the point that we can predict expected performance of a system on novel datasets, based upon a subset of test data for statistical calibration. Calibration of such a simulation using the test data sets would make it possible to evaluate any other scenario as a function of whichever characteristics were believed to be salient.

To illustrate this point further, Fig. 1 shows three different approaches to validating 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 training and another for validating. The result of the validation process is intended to be representative of the results that would be obtained by validating the system on the entire population. An alternative approach, shown on the left of Fig. 1, is to analyze 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.

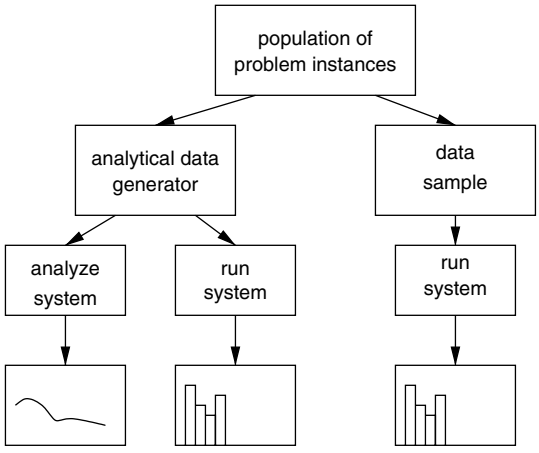

Fig. 1. Different approaches to validating algorithms.

There is a further possible route for validating. 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 synthesized data; the results obtained from this process should be comparable with the results obtained by both methods outlined in the previous paragraph.

5.2. Developments in the science of algorithm development

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

In the future, this might lead to offshoots into the exploration of numerically-stable techniques, for example for stable estimation of covariances from noisy data. In addition, the whole issue of more powerful optimization techniques is worthy of further study but cannot currently be identified as a problem because 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 modeling of the behavior of algorithms. The use of higher-order statistics may be a way to avoid problems with the non-quadratic or non-linear shape of 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 only in specialized circumstances.

Model selection is another topic for further work that is key to the automatic interpretation of arbitrary imagery. 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 Bhattycharia metric and the AIC have been used [263], often illustrated for specific formulations (*e.g.*, in curve-model fitting [264]).

In seeking to apply vision algorithms, it is important to test their validity, not just their novelty, so that the techniques developed by the community may be applied to significant problems. This will have the advantage for researchers that recognition, perhaps 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.

There appear to be common difficulties shared across the subject. However, in some areas, specific toolkits have been developed to address difficult issues and we believe that these could be applied in other application areas too. Unfortunately, what is becoming increasingly apparent is that the knowledge base of many researchers does not 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 shoot-outs. It is particularly worrying that, although much good work has been expended in addressing fundamental issues, there is a lack of general acceptance of what 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. We hope that this review and guide provides a step in this direction.

Acknowledgments

The support of the Information Society Technologies programme of the European Commission is gratefully acknowledged under the PCCV Project ('Performance Characterization of Computer Vision Techniques'), IST-1999-14159. William Crum is grateful for financial and intellectual support from the EP-SRC/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.

References

[1] R.M. Haralick, Computer vision theory: The lack thereof, Computer Vision, Graphics and Image Processing 36 (1986) 272–286.
[2] K.E. Price, Anything you can do, I can do better (no you can't), Computer Vision, Graphics and Image Processing 36 (2/3) (1986) 387–391.
[3] R.C. Jain, T. Binford. Dialogue: ignorance, myopia and naivete in computer vision systems. Computer Vision, Graphics and Image Processing: Image Understanding, 53(1) (1991) 1–128. Contributions from: M.A. Snyder, Y. Aloi-monos, A. Rosenfeld, T.S. Huang, K.W. Bowyer, J.P. Jones.
[4] R.M. Haralick, Dialogue: Performance characterization in computer vision, Computer Vision Graphics and Image Processing: Image Understanding 60 (1994) 245–265, contributions: L. Cinque, C. Guerra, S. Levialdi, J. Weng, T.S. Huang, P. Meer, Y. Shirai; B.A. Draper, J.R. Beveridge.
[5] R.M. Haralick, Performance characterization in computer vision, in: D. Hogg, R. Boyle (Eds.), Proceedings of the British Machine Vision Conference (BMVC92), Springer-Verlag, Berlin, 1992, pp. 1–8.
[6] R.M. Haralick, Overview: computer vision performance characterization, in: Proceedings of the ARPA Image Understanding Workshop, Monterey, USA, 1994, pp. 663–665.
[7] R.M. Haralick, Propagating covariances in computer vision, in: Proceedings of the International Conference on Pattern Recognition (ICPR94), September 1994,Jerusalem, Israel, 1994, pp. 493–498.
[8] W. Förstner, S. Ruwiedel, Robust Computer Vision-Quality of Vision Algorithms, Wichmann, Karlsruhe, Germany, 1992.
[9] T.A. Nartker, S.V. Rice, OCR accuracy: UNLV's third annual test, INFORM 8 (8) (1994) 30–36.
[10] T.A. Nartker, S.V. Rice, J. Kanai, OCR accuracy: UNLV's second annual test, INFORM 8 (1) (1994) 40–45.
[11] S.V. Rice, F.R. Jenkins, 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.
[12] J. Liang, I.T. Phillips, R.M. Haralick, An optimisation methodology for document structure extraction on latin character documents, IEEE Transactions PAMI 23 (7) (2001) 719–734.
[13] J.L. Liang, I.T. Phillips, R.M. Haralick, Performance evaluation of document structure extraction algorithms, Computer Vision and Image Understanding 84 (1) (2001) 144–159.
[14] D. Dori, W. Liu, M. Peleg, How to win a dashed line detection contest, in: R. Kasturi, K. Tombre (Eds.), Graphics Recognition—Methods and Applications, Springer-Verlag, Berlin, 1996, pp. 13–22.
[15] B. Kong, I.T. Phillips, R.M. Haralick, A. Prasad, R. Kasturi, A benchmark: Performance evaluation of dashed line detection algorithms, in: R. Kasturi, K. Tombre (Eds.), Graphics Recognition—Methods and Applications, Springer-Verlag, Berlin, 1996, pp. 270–285.
[16] A.K. Chhabra, I.T. Phillips, A benchmark for graphics recognition systems, in: K.W. Bowyer, P.J. Phillips (Eds.), Empirical Evaluation Techniques in Computer Vision, IEEE Comp Press, CA, USA, 1998.
[17] L. Wenyin, D. Dori, Incremental arc segmentation algorithm and its evaluation, IEEE Transactions PAMI 20 (4) (1998) 424–430.
[18] W. Liu, D. Dori, Performance evaluation of graphics/text separation, in: K. Tombre, A. Chhabra (Eds.), Graphics recognition algorithms and systems, Lecture Notes in Computer Science, vol. 1389, Springer, Berlin, 1998, pp. 359–371.
[19] L. Wenyin, D. Dori, Principles of constructing a performance evaluation protocol for graphics recognition algorithms, in: R. Klette, S. Stiehl, M. Viergever, V. Vincken (Eds.), Performance Evaluation of Computer Vision Algorithms, Kluwer Academic Publishers, Amsterdam, 2000, pp. 81–90.
[20] W. Liu, D. Dori, The arc segmentation contest, in: Proceedings of 4th IAPR Workshop on Graphics Recognition, September 2001, Kingston, Canada, 2001.
[21] American Society of Photogrammetry, Editor. Manual of Photogrammetry, 5th edition, American Society of Photogrammetry, Falls Church, VA, USA, 2004.
[22] T.A. Clarke, J.G. Fryer, The development of camera calibration methods and models, Photogrammetric Record 16 (91) (1998) 51–66.
[23] P. Courtney, Proceedings of the ECVnet Workshop on Performance Evaluation of Vision Algorithms, ECVnet, Paris, France, 1995.
[24] H.I. Christensen, W. Förstner, C.B. Madsen. Proceedings of the ECCV Workshop on Performance Characteristics of Vision Algorithms, ECCV, Cambridge, UK, 1996.
[25] W. Förstner, H.I. Christensen, Special issue on performance evaluation, Machine Vision and Applications 9 (5/6) (1997).
[26] W. Förstner, DAGM workshop on Performance Characteristics and Quality of Computer Vision Algorithms, Technical University of Brunswick, Germany, 1997.

[27] K.W. Bowyer, P.J. Phillips (Eds.), Empirical evaluation techniques in computer vision, IEEE Press, New York, 2000, ISBN 0-8186-8401-1.
[28] R. Klette, F. Wu, S.Z. Zhou, Multigrid convergence based evaluation of surface approximations, in: R. Klette, S. Stiehl, M. Viergever, V. Vincken (Eds.), Performance Evaluation of Computer Vision Algorithms, Kluwer Academic Publishers, Amsterdam, 2000, pp. 241–254.
[29] A. Clark, P. Courtney, Workshop on Performance Characterisation and Benchmarking of Vision Systems, BMVA, Canary Islands, 1999.
[30] J. Blanc-Talon, D. Popescu (Eds.), Imaging and Vision Systems: Theory, Assessment and Applications, NOVA Science Books, Huntington, New York, 2001.
[31] H.I. Christensen, P.J. Phillips, in: Proceedings of the Second Workshop on Empirical Evaluation Methods in Computer Vision, ECCV, Dublin, Ireland, 2000.
[32] H.I. Christensen, P.J. Phillips (Eds.), Empirical Evaluation Methods in Computer Vision, World Scientific Press, Singapore, 2002.
[33] A. Hoover, P.J. Flynn, P.J. Phillips, Special issue on empirical evaluation of computer vision algorithms, Computer Vision and Image Understanding 84 (1) (2001) 1–4.
[34] P.J. Phillips, K.W. Bowyer, Special section on empirical evaluation of computer vision algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (4) (1999) 289–384.
[35] M. Wirth, M. Fraschini, M. Masek, M. Bruynooghe, M. Moonen, Special issue performance evaluation in image processing, EURASIP Journal on Applied Signal Processing (2006), doi:10.1155/ASP/2006/45742.
[36] B.G. Batchelor, J. Charlier, Machine vision is not computer vision, in: B.G. Batchelor, J.W. Miller, S.S. Solomon (Eds.), Machine Vision Systems for Inspection and Metrology VII, vol. 3521, SPIE, 1998, pp. 2–13.
[37] P.J. Phillips, A. Martin, C.L. Wilson, M. Przybocki, An introduction to evaluating biometric systems, IEEE Computer 33 (2) (2000) 56–63.
[38] V. Ramesh, R.M. Haralick, A.S. Bedekar, X. Liu, D.C. Nadadur, K.B. Thornton, X. Zhang, Computer vision performance characterization, in: O. Firshein, T. Strat (Eds.), RADIUS: Image Understanding for Imagery Intelligence, Morgan Kaufmann, San Francisco, USA, 1997, pp. 241–282.
[39] F. Ahearne, N.A. Thacker, P.I. Rockett, The Bhattacharryya metric as an absolute similarity measure for frequency coded data, Kybernetica 34 (4) (1997) 363–368.
[40] G.W. Snedecor, W.G. Cochran, Statistical Methods (8th Edition) Ames (IA), Iowa State University Press, Iowa, 1989.
[41] I. Guyon, J. Makhoul, R. Schwartz, V. Vapnik, What size test set gives good error rate estimates? IEEE Transactions PAMI (1998) 52–64.
[42] V. Ramesh, R.M. Haralick, Random perturbation models and performance characterization in computer vision, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR92), Urbana-Champaign, USA, 1992, pp. 521–527.
[43] V. Ramesh, R.M. Haralick. A methodology for automatic selection of IU algorithm tuning parameters, in: Proceedings of the ARPA Image Understanding Workshop, Monterey, USA, 1994, pp. 675–687.
[44] V. Ramesh, R.M. Haralick, A.S. Bedekar, X. Liu, D.C. Nadadur, K.B. Thornton, X. Zhang. Computer vision performance characterization, in: O. Firschein, T. Strat (Eds.), RADIUS: Image Understanding for Imagery Intelligence, Morgan Kaufmann Publishers, San Francisco, USA, 1997, pp. 241–282.
[45] B. Efron, G. Gong, A leisurely look at the bootstrap, the jackknife, and cross-validation, American Statistician 37 (1983) 36–48.
[46] R.M. Haralick, Propagating covariance in computer vision, International Journal on Pattern Recognition and Artificial Intelligence 10 (1996) 561–572.
[47] P. Courtney, N. Thacker, A. Clark, Algorithmic modelling for performance evaluation, Machine Vision and Applications 9 (5/6) (1997) 219–228.
[48] K. Cho, P. Meer, J. Cabrera, Performance assessment through bootstrap, IEEE Transactions PAMI 19 (11) (1997) 1185–1198.
[49] C.W. Tong, S.K. Rodgers, J.P. Mills, M.K. Kabrinsky. Multisensor data fusion of laser radar and forward looking infared for target segmentation and enhancement, in: R.G. Buser F.B. Warren (Eds.), Infared Sensors and Sensor Fusion,SPIE, 1987.
[50] R. Marík, M. Petrou, J. Kittler. Compensation of the systematic errors of the CCD camera, in: SCIA'97, May 1997, Lapeenranta, Finland, 1997.
[51] G.E. Healey, R. Kondepudy, Radiometric CCD camera calibration and noise estimation, IEEE Transactions on Pattern Analysis and Machine Intelligence 16 (3) (1994) 267–276.
[52] Y. Tsin, V. Ramesh, T. Kanade, Statistical calibration of the CCD process, in: Proceedings of IEEE International Conference on Computer Vision ICCV2001, Vancouver, Canada, 2001.
[53] H. Gudbjartsson, S. Patz, The rician distribution of noisy MRI data, MRM 34 (1995) 910–914.
[54] H.A. Beyer, Accurate calibration of CCD cameras, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR92), Urbana-Champaign, USA, 1992, pp. 96–101.
[55] P.L. Rosin, Techniques for assessing polygonal approximations of curves, IEEE Transactions PAMI 19 (6) (1997) 659–666.
[56] P. Rosin, Assessing the behaviour of polygonal approximation algorithms, Pattern Recognition 36 (2003) 505–518.
[57] J.F. Canny, A computational approach to edge detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 8 (6) (1986) 679–698.
[58] E.R. Davies, Machine Vision: Theory, Algorithms, Practicalities, second ed., Academic press, London, 1997.
[59] H.P. Moravec, Obstacle avoidance and navigation in the real world by a seeing robot rover, Ph.D. thesis, Stanford University, September 1980.
[60] C. Harris, M. Stephens. A combined corner and edge detector, in: Proceedings of the 4th Alvey Vision Conference, 1988, pp. 47–151.
[61] S. Smith, J. Brady, SUSAN—a new approach to low level image processing, International Journal of Computer Vision 23 (1) (1997) 45–78.
[62] W. Freeman, E. Adelson, The design and use of steerable filters, IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (9) (1999) 891–906.
[63] L. van Gool, T. Moons, D. Ungureanu, Affine/photometric invariants for planar intensity patterns, in: Proceedings of the European Conference on Computer Vision, April 1996, Cambridge, England, 1996, pp. 642–651.
[64] D.G. Lowe, Object recognition from local scale-invariant features, in: Proceedings of the International Conference on Computer Vision, Kerkyra, Greece, 1999, pp. 1150–1157.
[65] V. Ramesh, R.M. Haralick, Random perturbation models for boundary extraction sequence, Machine Vision and Applications: Special Issue on Performance Characterization, 1998.
[66] S.J. Wang, T.O. Binford, Local step edge estimation: A new algorithm, statistical model and performance evaluation, in: Proceedings of the ARPA Image Understanding Workshop, 1993, pp. 1063–1070.
[67] M.D. Heath, S. Sarkar, T. Sanoki, K.W. Bowyer, A robust visual method for assessing the relative performance of edge detection algorithms, IEEE Transactions PAMI 19 (12) (1997) 1338–1359.
[68] S. Baker, S.K. Nayar, Global measures of coherence for edge detector evaluation, in: Proceedings of Conference on Computer Vision and Pattern Recognition CVPR 1999, June 1999, Ft. Collins, CO, USA, 1999, pp. 2373–2379.
[69] S. Konishi, A.L. Yuille, J.M. Coughlan, 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, June 1999, Ft. Collins, CO, USA, 1999, pp. 1573–1579.
[70] P. Courtney, T. Skordas, Characterisation de performances des algorithmes de vision—un exemple: le detecteur de coins, in: Proceedings RFIA10, Rennes, France, 1996, pp. 953–962.

[71] P. Courtney, N.A. Thacker, A.F. Clark, Algorithmic modelling for performance evaluation, in: Proceedings of the ECCV Workshop on Performance Characteristics of Vision Algorithms, April 1996, Cambridge, UK, 1996.
[72] V. Gouet, P. Montesinos, R. Deriche, D. Pele, Evaluation de détecteurs de points d'intérêt pour la couleur, in: Reconnaissance des formes et Intelligence Artificielle (RFIA '2000), volume II, Paris, France, 2000, pp. 257–266.
[73] C. Schmid, R. Mohr, C. Bauckhage, Evaluation of interest point detectors, International Journal of Computer Vision 37 (2000) 151–172.
[74] K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors. in: Proceedings of Conference on Computer Vision and Pattern Recognition CVPR 2003, USA, June 2003.
[75] K. Thornton, D.C. Nadadur, V. Ramesh, X. Liu, X. Zhang, A. Bedekar, R. Haralick, Groundtruthing the RADIUS model-board imagery, in: Proceedings of the ARPA Image Understanding Workshop, Monterey, USA, 1994, pp. 319–329.
[76] P.L. Rosin, Edges: saliency measures and automatic thresholding, Machine Vision and Applications 9 (1997) 139–159.
[77] R.M. Haralick, Covariance propagation in computer vision, in: Proceedings of the ECCV Workshop on Performance Characteristics of Vision Algorithms, Cambridge, UK, April 1996.
[78] K.W. Bowyer, C. Kranenburg, S. Dougherty, Edge detector evaluation using empirical ROC curves, in: Proceedings of Conference on Computer Vision and Pattern Recognition CVPR 1999, June 1999, Ft. Collins, CO, USA, 1999, pp. 1354–1359,.
[79] J.M. Fitzpatrick, J.B. West, A blinded evaluation and comparison of image registration methods, in: K.W. Bowyer, P.J. Phillips (Eds.), in: Workshop on Empirical Evaluation Techniques in Computer Vision, June 1998, Santa Barbara, California, 1998.
[80] M. Kass, A. Witkin, D. Terzopolous, Snakes: Active contour models, International Journal of Computer Vision 1 (1987) 321–323.
[81] T.F. Cootes, D. Cooper, C.J. Taylor, J. Graham, Active shape models—their training and application, Computer Vision and Image Understanding 61 (1) (1995) 38–59.
[82] D. Zhang, G. Lu, A review of shape representation and description techniques, Pattern Recognition 37 (2004) 1–19.
[83] M. Lindenbaum, Bounds on shape-recognition performance, IEEE Transactions on Pattern Analysis and Machine Intelligence 17 (7) (1995) 666–680.
[84] M. Lindenbaum, An integrated model for evaluating the amount of data required for reliable recognition, IEEE Transactions PAMI 19 (11) (1997) 1251–1264.
[85] M. Ulrich, C. Steger, Empirical performance evaluation of object recognition methods, in: Empirical Evaluation Methods in Computer Vision, December 2001, Hawaii, 2001.
[86] C.B. Madsen, A comparative study of the robustness of two-pose estimation techniques, Machine Vision and Applications 9 (5/6) (1997) 291–303.
[87] R. Kumar, A.R. Hanson, Robust methods for estimating pose and a sensitivity analysis, Computer Vision, Graphics and Image Processing 60 (3) (1994) 313–342.
[88] R.M. Haralick, C-N Lee, K. Ottenberg, M. Noelle, Review and analysis of solutions of the three point perspective pose estimation problem, International Journal on Computer Vision 13 (3) (1994) 331–356.
[89] D.W. Eggert, A. Lorusso, R.B. Fisher, Estimating 3D rigid body transformations: a comparison of four major algorithms, Machine Vision and Applications 9 (5/6) (1997) 272–290.
[90] P.L. Venetainer, E.W. Large, 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) (1997) 304–320.
[91] K. Kanatani, N. Ohta, Optimal robot self-localization and reliability evaluation, in: H. Burkhardt, B. Neumann (Eds.), European Conference Computer Vision ECCV98, II, Freiburg, Germany, 1998, pp. 796–808.
[92] P. Courtney, 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.
[93] J.C. Gee, Performance evaluation of medical image processing algorithms, in: H.I. Christensen, P.J. Phillips (Eds.), Empirical Evaluation Methods in Computer Vision, World Scientific Press, Singapore, 2002, pp. 143–159, ISBN 981-02-4953-5.
[94] M-H. Yang, D.J. Kriegman, N. Ahuja, Detecting faces in images: a survey, IEEE Transactions PAMI 24 (1) (2002) 34–58.
[95] L.J. Latecki, R. Lakamper, 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, Hilton Head, SC, USA, June 2000, IEEE Computer Society, pp. 424–429.
[96] L. Wenyin, Z. Su, S. Li, Y-F. Sun, H. Zhang, A performance evaluation protocol for content-based image retrieval, in: Empirical Evaluation Methods in Computer Vision, Hawaii, December 2001.
[97] A.P. Ashbrook, N.A. Thacker, P.I. Rockett, C.I. Brown, Robust recognition of scaled shapes using pairwise geometric histograms, in: Proceedings of the British Machine Vision Conference (BMVC95), September 1995, Birmingham, UK, 1995.
[98] P.A. Bromiley, M. Pokric, N.A. Thacker, Computing covariances for mutual information coregistration, in: Proceedings of MIUA, 2004, pp. 77–80.
[99] W.R. Crum, L.D. Griffin, D.L.G. Hill, D.J. Hawkes, Zen and the art of medical image registration: correspondence, homology and quality, NeuroImage 20 (2003) 1425–1437.
[100] N.A. Thacker, P.A. Riocreux, R.B. Yates, Assessing the completeness properties of pairwise geometric histograms, Image and Vision Computing 13 (5) (1995) 423–429.
[101] S.J. Maybank, Probabilistic analysis of the application of the cross ratio to model-based vision: Misclassification, International Journal of Computer Vision 14 (3) (1995) 199–210.
[102] S.J. Maybank, Probabilistic analysis of the application of the cross ratio to model-based vision, International Journal of Computer Vision 15 (1) (1995) 5–33.
[103] D. Huynh, The cross ratio: a revisit to its probability density function, in: M. Mirmehdi, B. Thomas (Eds.), in: Proceedings of the British Machine Vision Conference BMVC 2000, Bristol, UK, September 2000.
[104] D.P. Huttenlocher, W.E.L. Grimson, On the verification of hypothesized matches in model-based recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (12) (1991) 1201–1213.
[105] D.P. Huttenlocher, W.E.L. Grimson, On the sensitivity of geometric hashing, in: Proceedings of the International Conference on Computer Vision (ICCV), 1990, pp. 334–338.
[106] T.D. Alter, W.E.L. Grimson, Verifying model-based alignments in the presence of uncertainty, in: Computer Vision Pattern Recognition CVPR97, June 1997, Puerto Rico, pp. 344–349.
[107] K.B. Sarachik, The effect of gaussian error in object recognition, IEEE Transactions PAMI 19 (4) (1997) 289–301.
[108] M.C. Shin, D.B. Goldgof, K.W. Bowyer, Comparison of edge detectors using an object recognition task, in: Proceedings of Conference on Computer Vision and Pattern Recognition CVPR 1999, Ft. Collins, CO, USA, 1999, p. 1360.
[109] G. Jones, B. Bhanu, Recognition of articulated and occluded objects, IEEE Transactions PAMI 21 (7) (1999) 603–613.
[110] M. Boshra, B. Bhanu, Predicting performance of object recognition, IEEE Transactions PAMI 22 (8) (2000) 956–969.
[111] T.D. Ross, V. Velton, J. Mossing, S. Worrell, M. Bryant, Standard SAR ATR evaluation experiments using the MSTAR public release data set, in: Algorithms for Synthetic Aperture Radar Imagery, vol. 3370, SPIE, April 1998.
[112] H.A. Rowley, S. Baluja, T. Kanade, Rotation Invariant Neural Network-Based Face Detection, Technical Report CMU-CS-97-201, Carnegie-Mellon University, December 1997.

[113] B.D. Lucas, T. Kanade. An iterative image-registration technique with an application to stereo vision, in: Image Understanding Workshop, US Defence Advanced Research Projects Agency, 1981, pp. 121–130 (see also International Joint Conference on Artificial Intelligence '81, pp. 674–679).
[114] J. Princen, J. Illingworth, J. Kittler, Hypothesis testing: A framework for analyzing and optimizing Hough transform performance, IEEE Transactions on Pattern Analysis and Machine Intelligence 16 (1994) 329–341.
[115] N.A. Thacker, F.J. Aherne, P.I. Rockett, The Bhattacharyya metric as an absolute similarity measure for frequency coded data, Kybernetika 32 (4) (1995) 1–7.
[116] J. Hutchinson, Culture, communication and an information age madonna, IEEE Professional Communications Society Newsletter 45 (3) (2001) 1–5.
[117] M. Ghanbari, Standard Codecs: Image Compression to Advanced Video Coding, IEE Press, London, 2003.
[118] H. Spies, B. Jähne, J.L. Barron, Range flow estimation, Computer Vision Image Understanding 85 (3) (2002) 209–231.
[119] X. Tang, J.L. Barron, R.E. Mercer, P. Joe, Tracking weather storms using 3D doppler radial velocity information, in: Thirteenth Scandinavian Conference on Image Analysis, 2003, pp. 1038–1043.
[120] J.L. Barron, Experience with 3D optical flow on gated MRI cardiac datasets, in: First Canadian Conference on Computer and Robot Vision, 2004, pp. 370–377.
[121] C. Theobalt, J. Carranza, M.A. Magnor, H.P. Seidel, Enhancing silhouette-based human motion capture with 3D motion fields, in: Proceedings Pacific Graphics, 2003, pp. 185–193.
[122] E.P. Simoncelli, Design of multi-dimensional derivative filters, in: International Conference on Image Processing, vol. 1, 1994, pp. 790–793.
[123] J.L. Barron, M. Daniel, J.-L. Mari, Using 3d spline differentiation to compute quantitative optical flow, in: Third Canadian Conference on Computer and Robot Vision, CD, see IEEE Xplore, 2006.
[124] B.K.P. Horn, B.G. Schunck, Determining optical flow, Artificial Intelligence 17 (1981) 185–204.
[125] A. Nomura, Spatio-temporal optimization method for determining motion vector fields under non-stationary illuminations, Image and Vision Computing 18 (2000) 939–950.
[126] L. Zhang, T. Sakurai, H. Miike, Detection of motion fields under spatio-temporal non-uniform illuminations, Image and Vision Computing 17 (1999) 309–320.
[127] S. Negahdaripour, Revised definition of optical flow: Integration of radiometric and geometric cues for dynamic scene analysis, Pattern Analysis and Machine Intelligence 20 (9) (1998) 961–979.
[128] H.W. Haußecker, D.J. Fleet, Computing optical flow with physical models of brightness variation, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (6) (2001) 661–673.
[129] D. Fleet, Measurement of Image Velocity, Kluwer Academic Publishers, Norwell, 1992.
[130] J.L. Barron, R. Eagleson, Recursive estimation of time-varying motion and structure parameters, Pattern Recognition 29 (5) (1996) 797–818.
[131] B. Tian, J. Barron, W.K.J. Ngai, H.A. Spies, Comparison of 2 methods for recovering dense accurate depth using known 3D camera motion, in: Vision Interface, 2003, pp. 229–236.
[132] J.L. Barron, D.J. Fleet, S.S. Beauchemin, Systems and experiments: Performance of optical flow techniques, International Journal of Computer Vision 12 (1) (1994) 43–77.
[133] M. Otte, H.-H. Nagel, Optical flow estimation: advances and comparisons, in: European Conference on Computer Vision, 1994, pp. 51–60.
[134] E.P. Simoncelli, Bayesian Multi-scale Differential Optical Flow, vol. 2, Academic Press, New York, 1999, pp. 397–422, Chapter 14.
[135] D. Comaniciu, V. Ramesh, P. Meer, Kernel-based optical tracking, IEEE Transactions PAMI 25 (5) (2003) 564–577.
[136] R. Marík, J. Kittler, 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, April 1996, Cambridge, UK, 1996.
[137] C. Fermüller, Y. Aloimonos, The statistics of optical flow: implications for the processes of correspondence in vision, in: ICPR, vol. 1, 2000, pp. 119–126.
[138] C. Fermüller, R. Pless, Y. Aloimonos, Statistical biases in optical flow, in: Conference on Computer Vision and Pattern Recognition, vol. 1, 1999, pp. 561–566.
[139] C. Fermüller, H. Malm, 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.
[140] C. Fermüller, Y. Aloimonos, 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.
[141] M. Ye, R.M. Haralick, Two-stage robust optical flow estimation, in: Conference on Computer Vision and Pattern Recognition, vol. 2, 2000, pp. 2623–2028.
[142] M. Ye, R.M. Haralick, Optical flow from a least-trimmed squares based adaptive approach, in: ICPR, vol. III, 2000, pp. 1052–1055.
[143] O. Nestares, D.J. Fleet, 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, June 2000, Hilton Head, SC, USA, IEEE Computer Society, 2000, pp. 1523–1530.
[144] H.-H Nagel. Constraints for the estimation of displacement vector fields from image sequences, in: International Joint Conference on Artificial Intelligence, 1983, pp. 945–951.
[145] S. Uras, F. Girosi, A. Verri, V. Torre, A computational approach to motion perception, Biological Cybernetics 60 (1988) 79–97.
[146] T. Brox, A. Bruhn, N. Papenberg, J. Weickert, High accuracy optical flow estimation based on a theory for warping, in: Proceedings of the European Conference on Computer Vision, LNCS Vol. 3024, Springer-Verlag, Berlin, 2004, pp. 25–36.
[147] N. Papenberg, A. Bruhn, T. Brox, S. Didas, J. Weickert, Highly accurate optic flow computation with theoretically justified warping, International Journal of Computer Vision 67 (2) (2006) 141–158.
[148] J.H. Bergen, P. Anandan, K.J. Hamma, R. Hingorani,Hierarchical model-based motion estimation, in: European Conference on Computer Vision, May 1992, pp. 237–252.
[149] J.L. Barron, M. Khurana, Determining optical flow for large motions using parametric models in a hierarchical framework, in: Vision Interface, May 1997, pp. 47–56.
[150] H. Spies, J.L. Barron, Evaluating certainties for image intensity differentiation for optical flow, in: First Canadian Conference on Computer and Robot Vision, 2004, pp. 408–416.
[151] A. Verri, T. Poggio, Against quantitative optical flow, in: J. Michael Fitzpatrick (Ed.), in: Proceedings of the First International Conference on Computer Vision, London, IEEE, 1987, pp. 171–180.
[152] T. Lin, J.L. Barron, in: C. Archibald, P. Kwok (Eds.), Image Reconstruction Error for Optical Flow, Scientific Publishing Co, Singapore, 1995, pp. 269–290.
[153] O. Faugeras, Q. Luong, T. Papadopoulo, The Geometry of Multiple Images, MIT Press, Cambridge, MA, 2001.
[154] R.I. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, Cambridge, MA, 2004.
[155] S.B. Pollard, J.E.W. Mayhew, J.P. Frisby, PMF: a stereo correspondence algorithm using a disparity gradient limit, Perception 14 (1985) 449–470.
[156] H.H. Baker, T.O. Binford, Depth from edge and intensity based stereo, in: Proceedings of the VII International Joint Conference on Artificial Intelligence, August 1981.
[157] S.T. Barnard, W.B. Thompson, Disparity analysis of images, IEEE Transactions on Pattern Analysis and Machine Intelligence 2 (4) (1980) 333–340.

[158] T. Day, J.P. Muller, Digital elevation model production by stereo-matching spot image pairs: a comparison of two algorithms, Image Vision Computing 7 (1989) 95–101.
[159] O. Faugeras, P. Fua, B. Hotz, R. Ma, L. Robert, M. Thonnat, Z. Zhang. Quantitative and qualitative comparisons of some area and feature-based stereo algorithms, in: Proceedings of the Second International Workshop, March 1992, Wichmann, Karlsruhe, Germany, 1992, pp. 1–26.
[160] R. Lane, N.A. Thacker, N.L. Seed, Stretch correlation as a real-time alternative to feature-based stereo matching algorithms, Image and Vision Computing 12 (4) (1994) 203–212.
[161] E. Guelch, Results of test on image matching of ISPRS WG III/4, ISPRS Journal of Photogrammetry and Remote Sensing 46 (1991) 1–18.
[162] R.C. Bolles, H.H. Baker, M.J. Hannah, The JISCT stereo evaluation, in: Proceedings of the ARPA Image Understanding Workshop, Washington D.C., USA, 1993, pp. 263–274.
[163] H. Hirschmueller, Improvements in real-time correlation-based stereo vision, in: IEEE Workshop on Stereo and Multi-Baseline Vision, December 2001, Hawaii, 2001.
[164] J. Weng, T.S. Huang, N. Ahuja, Motion and structure from two perspective views: Algorithms, error analysis and error estimation, IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (5) (1989) 451–476.
[165] K.I. Kanatani, Unbiased estimation and statistical analysis of 3-D rigid motion from two views, IEEE Transactions on Pattern Analysis and Machine Intelligence 15 (1) (1993) 37–50.
[166] M. Brooks, W. Chojnacki, D. Gawley, A. van den Hengel, What value covariance information in estimating vision parameters? in: International Conference on Computer Vision ICCV2001, July 2001, Vancouver, B.C., Canada, 2001.
[167] O. Jokinen, H. Haggren, Statistical analysis of two 3-D registration and modeling strategies, Photogrammetry and Remote Sensing 53 (6) (1998) 320–341.
[168] G. Kamberova, R. Badcsy, Sensor errors and the uncertainties in stereo reconstruction, in: Workshop on Empirical Evaluation Methods in Computer Vision, Santa Barbara, California, June 1998.
[169] D. Scharstein, R. Szeliski, A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, International Journal of Computer Vision, 47 (1–3) (2002) 7–42, April 2002. Microsoft Research Technical Report MSR-TR-2001-81, November 2001.
[170] M.W. Maimone, S.A. Shafer, A taxonomy for stereo computer vision experiments, in: Proceedings of the ECCV Workshop on Performance Characteristics of Vision Algorithms, April 1996, Cambridge, UK, 1996.
[171] Y.G. Leclerc, Q-T Luong, P. Fua, Self consistency: a novel approach to characterising the accuracy and reliability of point correspondence algorithms, in: A. Clark, P. Courtney (Eds.), Workshop on Performance Characterisation and Benchmarking of Vision Systems, January 1999, Canary Islands Spain, 1999.
[172] E.P. Krotkov, Active computer vision by cooperative focus and stereo. Springer series in Perception Engineering, Springer-Verlag, Berlin, 1989.
[173] A.J. Harris, N.A. Thacker, A.J. Lacey, Modelling feature based stereo vision for range sensor simulation, in: Proceedings of the European Simulation Multiconference, June 1998, pp. 417–421.
[174] N.A. Thacker, P. Courtney, Statistical analysis of a stereo matching algorithm, in: Proceedings of the British Machine Vision Conference (BMVC92), 1992, pp. 316–326.
[175] A. Bedekar, R.M. Haralick, A Bayesian method for triangulation, in: ICIP-95: Proceedings, International Conference on Image Processing, vol. II, October 1995, pp. 362–365.
[176] G. Csurka, C. Zeller, Z. Zhang, O. Faugeras, Characterizing the uncertainty of the fundamental matrix, Computer Vision and Image Understanding 68 (1) (1997) 18–35.
[177] P.H.S. Torr, A. Zisserman, Performance characterisation of fundamental matrix estimation under image degradation, Machine Vision and Applications 9 (5/6) (1997) 321–333.
[178] J.J. Rodriguez, J.K. Aggarwal, Stochastic analysis of stereo quantization error, IEEE Transactions on Pattern Analysis and Machine Intelligence 12 (5) (1990) 467–470.
[179] L. Matthies, S. Shafer, Error modelling in stereo navigation, IEEE Journal of Robotics and Automation 3 (3) (1987) 239–248.
[180] S.D. Blostein, T.S. Huang, Error analysis in stereo determination of 3D point positions, IEEE Transactions on Pattern Analysis and Machine Intelligence, 9 (6) (1988) 752–765, November 1987. Correction: on Pattern Analysis and Machine Intelligence 10 (5) September 1988, p. 765.
[181] S.M. Kiang, R.J. Chou, J.K. Aggarwal, Triangulation errors in stereo algorithms, in: CVWS87, 1987, pp. 72–78.
[182] G.G. Mohan, R. Medioni, R. Nevatia, Stereo error detection, correction and evaluation, IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (2) (1989) 113–120.
[183] Y.C. Hsieh, D.M. McKeown Jr., F.P. Perlant, Performance evaluation of scene registration and stereo matching for cartographic feature extraction, IEEE Transactions on Pattern Analysis and Machine Intelligence 14 (1992) 214–238.
[184] D. Ferrari, G. Garibotto, S. Masciangelo, Towards experimental computer vision: performance assessment of a trinocular stereo system, in: Proceedings of the ECCV ESPRIT day workshop, May 1992, Santa Margherita Ligure, Italy, 1992.
[185] S. Thayer, M. Trivedi, Residual uncertainty in 3-dimensional reconstruction using 2-planes calibration and stereo methods, Pattern Recognition 28 (7) (1995) 1073–1082.
[186] Y. Xiong, L. Matthies, Error analysis of a real time stereo system, in: Computer Vision Pattern Recognition CVPR97, June 1997, Puerto Rico, 1997.
[187] M. Petrou, N. Georgis, J. Kittler, Sensitivity analysis of projective geometry 3D reconstruction, in: R. Klette, S. Stiehl, M. Viergever, V. Vincken (Eds.), Performance Evaluation of Computer Vision Algorithms, Kluwer Academic Publishers, Amsterdam, 2000, pp. 255–264.
[188] N. Roma, J. Santos-Victor, J. Tome, A comparative analysis of cross-correlation matching algorithms using a pyramidal resolution approach, in: H.I. Christensen, P.J. Phillips (Eds.), Empirical Evaluation Methods in Computer Vision, World Scientific Press, Singapore, 2002, pp. 117–142, ISBN 981-02-4953-5.
[189] J. Porill, S.B. Pollard, T. Pridmore, J. Bowen, J.E.W. Mayhew, J.P. Frisby, Tina: A 3D vision system for pick and place, in: Proceedings of the Alvey Vision Conference, 1987.
[190] P.J. Phillips, H.J. Moon, S.A. Rizvi, P.J. Rauss, The FERET evaluation methodology for face-recognition algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (10) (2000) 1090–1104.
[191] D.M. Blackburn, M. Bone, P.J. Phillips, Facial recognition vendor test 2000: Executive overview, Technical report, Face Recognition Vendor Test, 2000.
[192] P.J. Phillips, P. Grother, R.J Micheals, D.M. Blackburn, E. Tabassi, J.M. Bone, FRVT 2002: Overview and summary, Technical report, Face Recognition Vendor Test 2002, 2002.
[193] A.J. Mansfield, J.L. Wayman, Best practices in testing and reporting performance of biometric devices, Technical report, National Physical Laboratory, Teddington, Middlesex, UK, August 2002.
[194] S.A. Rizvi, P.J. Phillips, H. Moon, The FERET verification testing protocol for face recognition algorithms, Technical Report 6281, NIST, October 1998.
[195] M.-H. Yang, D. Kriegman, N. Ahuja, Detecting faces in images: a survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (1) (2002) 34–58.
[196] J.P. Egan, Signal Detection Theory and ROC Analysis, Academic Press, New York, 1975.
[197] R. Lienhart, J. Maydt, An extended set of haar-like features for rapid object detection, in: Proceedings of the International Conference in Image Processing, September 2002, vol. 1, 2002, pp. 900–903.
[198] P. Viola, M. Jones, Robust real-time object detection, International Journal of Computer Vision (2004) 137–154.

[199] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, in: Computational Learning Theory: Eurocolt 1995, Springer-Verlag, Berlin, 1995, pp. 23–37.
[200] M.-H. Yang, D. Roth, N. Ahuja, SNoW-based face detector, in: S.A. Solla, T.K. Leen, K.-R. Muller (Eds.), Advances in Neural Information Processing Systems, vol. 12, MIT Press, Cambridge, MA, 2000, pp. 855–861.
[201] M. Alvira, R. Rifkin, An empirical comparison of SNoW and SVMs for face detection, in: MIT AI Memo, 2001.
[202] E. Osuna, R. Freund, F. Girosi, Training support vector machines: an application to face detection, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, 1997, pp. 130–136.
[203] H. Schneiderman, Learning statistical structure for object detection, in: Computer Analysis of Images and Patterns (CAIP), 2003.
[204] A.J. O'Toole, J. Harms, S.L. Snow, D.R. Hurst, M.R. Pappas, J. Ayyad, H. Abdi, A video database of moving faces and people, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (5) (2005) 812–816.
[205] E. Bailly-Bailliére, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruíz, J.-P. Thiran, The BANCA database and evaluation protocol, in: Fourth International Conference on Audio- and Video-based Biometric Person Authentication, 2003, pp. 625–638.
[206] T. Sim, S. Baker, M. Bsat, The CMU pose, illumination and expression database, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (12) (2003) 1615–1618.
[207] P.N. Belhumeur, D.J. Kriegman, What is the set of images of an object under all possible illumination conditions, International Journal of Computer Vision 28 (3) (1998) 245–260.
[208] G. Givens, J.R. Beveridge, B.A. Draper, P.J. Phillips, P. Grother, How features of the human face affect recognition: a statistical comparison of three face recognition algorithms, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, 2004, pp. 381–388.
[209] N. Furl, J. Phillips, A.J. O'Toole, Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis, Cognitive Science 26 (2002) 797–815.
[210] M.A. Turk, A.P. Pentland, Face recognition using eigenfaces, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, June 1991, pp. 586–591.
[211] D. Bolme, J.R.B., M. Teixeira, B.A. Draper, The CSU face identification evaluation system: its purpose, features and structure, in: Proceedings of the Third International Conference on Vision Systems, April 2003, Graz, Austria, 2003, pp. 304–311.
[212] W. Zhao, R. Chellappa, A. Krishnaswamy, Discriminant analysis of principal components for face recognition, in: Face Recognition: From Theory to Applications, 1998, pp. 73–85.
[213] B. Moghaddam, C. Nastar, A. Pentland, A Bayesian similarity measure for direct image matching, Proceedings of the International Conference on Pattern Recognition (1996) B:350–B:358.
[214] K. Okada, J. Steffens, T. Maurer, H. Hong, E. Elagin, H. Neven, C. 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 Soulie, T.S. Huang (Eds.), Face Recognition: From Theory to Applications, Springer-Verlag, Berlin, 1998, pp. 186–205.
[215] V. Blanz, S. Romdhami, 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, 2002, pp. 202–207.
[216] B.A. Draper, W.S. Yambor, J.R. Beveridge, Analyzing PCA-based face recognition algorithms: eigenvector selection and distance measures, in: H. Christensen, J. Phillips (Eds.), Empirical Evaluation Methods in Computer Vision, World Scientific Press, Singapore, 2002.
[217] R.J. Micheals, T. Boult, Efficient evaluation of classification and recognition systems, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, December 2001, vol. I, pp. 50–57.
[218] J.R. Beveridge, K. She, B. Draper, G.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, December 2001, pp. 535–542.
[219] D.L.G. Hill, P.G. Batchelor, M. Holden, D.J. Hawkes, Medical image registration, Physics in Medicine and Biology 46 (3) (2001) R1–R45.
[220] J. Zitova, J. Flusser, Image registration methods: a survey, Image and Vision Computing 21 (11) (2003) 977–1000.
[221] S.D. Olabarriaga, A.W.M. Smeulders, Interaction in the segmentation of medical images: a survey, Medical Image Analysis 5 (2) (2001) 127–142.
[222] J.S. Suri, S. Singh, 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) (2002) 46–76.
[223] R.J. Radke, S. Andra, O. Al-Kofahi, B. Roysam, Image change detection algorithms: a systematic survey, IEEE Transactions on Image Processing 14 (3) (2005) 294–307.
[224] P.A. Bromiley, N.A. Thacker, P. Courtney, Non-parametric image subtraction using grey level scattergrams, Image and Vision Computing 20 (2002) 609–617.
[225] J. Ashburner, R. Hutton, C. Frackowiak, I. Johnsrude, C. Price, K. Fris-ton, Identifying global anatomical differences: deformation-based morphometry, Human Brain Mapping 6 (5–6) (1998) 347–358.
[226] C. Davatzikos, Why voxel-based morphometric analysis should be used with great caution when characterizing group differences, NeuroImage 23 (1) (2004) 17–20.
[227] A. Jackson, Analysis of dynamic contrast enhanced MRI, British Journal of Radiology 77 (2004) S154–S166.
[228] P.A. Freeborough, N.C. Fox, Modeling brain deformations in alzheimer disease by fluid registration of serial 3D MR images, Journal of Computer Assisted Tomography 22 (5) (1998) 838–843.
[229] P.A. Freeborough, N.C. Fox, The boundary shift integral: an accurate and robust measure of cerebral volume changes from registered repeat MRI, IEEE Transactions on Medical Imaging 16 (5) (1997) 623–629.
[230] K. Chen, E.M. Reiman, G.E. Alexander, D. Bandy, R. Renaut, W.R. Crum, N.C. Fox, M.N. Rossor, An automated algorithm for the computation of brain volume change from sequential MRIs using an iterative principal component analysis and its evaluation for the assessment of whole brain atrophy rates in patients with probable Alzheimer's disease, NeuroImage 22 (2004) 134–143.
[231] S.M. Smith, Y.Y. Zhang, M. Jenkinson, J. Chen, P.M. Matthews, A. De Federico, N. Stefano, Accurate, robust, and automated longitudinal and cross-sectional brain change analysis, NeuroImage 17 (1) (2002) 479–489.
[232] C. Gaser, H.P. Volz, S. Kiebel, S. Riehemann, H. Sauer, Detecting structural changes in whole brain based on nonlinear deformations—application to schizophrenia research, NeuroImage 10 (2) (1999) 107–113.
[233] J. Ashburner, K.J. Friston, Voxel-based morphometry—the methods, NeuroImage 11 (2000) 805–821.
[234] C.D. Good, R.I. Scahill, N.C. Fox, J. Ashburner, K.J. Friston, D. Chan, W.R. Crum, M.N. Rossor, R.S.J. Frackowiak, Automatic differentiation of anatomical patterns in the human brain: validation with studies of degenerative dementias, NeuroImage 17 (1) (2002) 29–46.
[235] C.H. Salmond, J. Ashburner, F. Vargha-Khadem, A. Connelly, D.G. Gadian, K.J. Friston, The precision of anatomical normalization in the medial temporal lobe using spatial basis functions, NeuroImage 17 (1) (2002) 507–512.
[236] H. Yabuuchi, T. Fukuya, T. Taijima, Y. Hachitanda, K. Tomita, M. Koga, Salivary gland tumors: diagnostic value of gadolinium-

enhanced dynamic MR imaging with histopathologic correlation, Radiology 226 (2) (2003) 345–354.
[237] A.R. Padhani, J.E. Husband, Dynamic contrast-enhanced MRI studies in oncology with an emphasis on quantification, validation and human studies, Clinical Radiology 56 (8) (2001) 607–620.
[238] N. Boussion, G. Soulez, J.A. de Guise, M. Daronat, Z. Qin, G. Cloutier, Geometrical accuracy and fusion of multimodal vascular images: a phantom study, Medical Physics 31 (6) (2004) 1434–1443.
[239] C. Grova, P. Jannin, A. Biraben, I. Buvat, H. Benali, A.M. Bernard, J.M. Scarabin, B. Gibaud, 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) (2003) 4023–4043.
[240] I.G. Zubal, C.R. Harrell, E.O. Smith, Z. Rattner, G. Gindi, P.B. Hof-fer, Computerized 3-dimensional segmented human anatomy, Medical Physics 21 (2) (1994) 299–302.
[241] O. Camara, M. Schweiger, R.I. Scahill, W.R. Crum, B.I. Sneller, J.A. Schnabel, G.R. Ridgway, D.M. Cash, D.L.G. Hill, N.C. Fox, Phenomenological model of diffuse global and regional atrophy using finite-element methods, IEEE Transactions on Medical Imaging 25 (11) (2006) 1417–1430.
[242] J.A. Schnabel, C. Tanner, A.D. Castellano-Smith, A. Degenhard, M.O. Leach, D.R. Hose, D.L.G. Hill, D.J. Hawkes, Validation of nonrigid image registration using finite-element methods: Application to breast MR images, IEEE Transactions on Medical Imaging 22 (2) (2003) 238–247.
[243] W.R. Crum, D. Rueckert, M. Jenkinson, D. Kennedy, S.M. Smith, A framework for detailed objective comparison of non-rigid registration algorithms in neuroimaging, in: Proceedings of MICCAI 2004, vol. LNCS 3216, 2004, pp. 679–686.
[244] D.L. Collins, A.P. Zijdenbos, V. Kollokian, J.G. Sled, N.J. Kabani, C.J. Holmes, A.C. Evans, Design and construction of a realistic digital brain phantom, IEEE Transactions on Medical Imaging 17 (3) (1998) 463–468.
[245] N.A. Thacker, A. Jackson, Mathematical segmentation of grey matter, white matter and cerebral spinal fluid from MR image pairs, British Journal of Radiology 74 (2001) 234–242.
[246] R. Highnam, J.M. Brady, B. Shepstone, A representation for mammo-graphic image processing, Medical Image Analysis 1 (1) (1996) 1–18.
[247] J.A. Ashburner, K.J. Friston, Unified segmentation, NeuroImage 26 (2005) 839–851.
[248] T.K. Ho, H.S. Baird, Large-scale simulation studies in image pattern recognition, IEEE Transactions PAMI 19 (10) (1997) 1067–1079.
[249] D. Maltoni, D. Maio, A.K. Jain, S. Prabhakar, Handbook of Fingerprint Recognition, Springer, New York, 2003.
[250] I. Guyon, R.M. Haralick, J. Hull, I. Phillips, Data sets for OCR and document image understanding, in: Handbook on Optical Character Recognition and Document Image Analysis, World Scientific Publishing Company, Singapore, 1996.
[251] A.F. Clark, 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.
[252] K. Sparck-Jones, 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.
[253] K. Sparck-Jones, What might be in a summary? in: Krause Knorz, Womser-Hacker (Eds.), Information Retrieval 93: Von der Modellierung zur Anwendung, Konstanz, Universitatsverlag Konstanz, 1993, pp. 9–26.
[254] D. Lewis, K. Sparck-Jones, Natural language processing for information retrieval, Technical Report 307, University of Cambridge, 1993.
[255] J.R. Galliers, K. Sparck-Jones, Evaluating natural language processing systems, Technical Report 291, University of Cambridge, March 1993.
[256] D.M. Blackburn, M. Bone, P.J. Philips, Facial recognition vendor test 2000: evaluation report, Technical report, DARPA, 2000.
[257] T. Ellis, Performance metrics and methods for tracking in surveillance, in: Third IEEE International Workshop on Performance Evaluation of Tracking and Surveillance PETS'2002, June 2002, Copenhagen, Denmark, 2002.
[258] J. Black, T. Ellis, P. Rosin, A novel method for video tracking performance evaluation, in: Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS), 2003, pp. 125–132.
[259] C.J. Needham, R.D. Boyle, Performance evaluation metrics and statistics for positional tracker evaluation, in: M. Petrou, J. Kittler, M. Nixon (Eds.), International Conference on Vision Systems (ICVS03), Graz, Austria, 2003, pp. 278–289.
[260] T. Schloegl, C. Beleznai, M. Winter, H. Bischof, Performance evaluation metrics detection and tracking, in: J. Kittler, M. Petrou, M. Nixon (Eds.), in: Proceedings of the International Conference on Pattern Recognition (ICPR04), vol. IV, Cambridge, England, 2004, pp. 519–522.
[261] H. Spies, Certainties in low-level operators, in: Vision Interface, 2003, pp. 257–262.
[262] W. Förstner, 10 pros and cons against performance characterisation of vision algorithms, in: Proceedings of the ECCV Workshop on Performance Characteristics of Vision Algorithms, April 1996, Cambridge, UK, 1996.
[263] H. Akaike, A new look at statistical model identification, IEEE Transactions on Automatic Control 19 (1974) 716.
[264] N.A. Thacker, D. Prendergast, P.I. Rockett, B-fitting: a statistical estimation technique with automatic parameter selection, in: Proceedings of the 1996 British Machine Vision Conference, Edinburgh, 1996, pp. 283–292.