

Consideration of Uncertainty in Computer Vision: Necessity and Chance¹

J. Meidow

*Research Institute for Optronics and Pattern Recognition, Gutleuthausstr. 1, 76275 Ettlingen, Germany
e-mail: meidow@fom.fgan.de*

Abstract—Observations and decisions in computer vision are inherently uncertain. The rigorous treatment of uncertainty has therefore received a lot of attention, since it not only improves the results compared to ad hoc methods but also makes the results more explainable. In this paper, the usefulness of stochastic approaches will be demonstrated by example with selected problems. These are given in the context of optimal estimation, self-diagnostics, and performance evaluation and cover all steps of the reasoning chain. The removal or interpretation of unexplainable thresholds and tuning parameters will be discussed for typical tasks in feature extraction, object reconstruction, and object classification.

DOI: 10.1134/S1054661808020053

1. INTRODUCTION

Motivation. The observations used in computer vision are inherently uncertain and often contaminated with outliers. The rigorous and consistent treatment of uncertainty has therefore received a lot of attention, since it not only improves the results compared to ad hoc methods but also makes the results more explainable.

Often, thresholds for decisions can be replaced by quantile values of hypothesis tests, i.e., error probability limits, provided that the stochastic properties of the entities under consideration are known. The statistical interpretation of free tuning parameters gives insights into their meaning and eventually permits them to be set reasonably or even to be eliminated. Thus, expert knowledge for tuning these parameters becomes obsolete, which paves the way toward fully automatic procedures.

Some publicly available computer programs for image processing and computer vision require assumptions about uncertainty as input parameters, e.g., the image noise level, but do not provide any information about the quality of the output. This defeats statements about the reliability of the results. Sometimes this is incomprehensible, because often little additional effort is required to provide this information. Reported quality measures can help the user to decide on the usefulness of the algorithm for the specific application.

Related work. D. Mumford predicted that statistical reasoning and modeling of uncertainty will become the scientific paradigm of the future [1]. Within the artificial intelligence community, P. Cheeseman defended

probability theory as the right tool for dealing with uncertainty: Many forms of reasoning systems are reducible or inferior to the use of probabilistic methods for the same task [2].

Methods for geometric reasoning within the framework of projective geometry can be found in [3] and [4]. To speed up the verification process within the (deterministic) RANSAC algorithm [5], assessment components have been introduced [6, 7]. Studies for the uncertainty of multiview relations are [8, 9]. Statistically based model selection schemes can be found for instance in [10] for structure from motion problems and in [11] for geometric reasoning.

Contribution. The presented considerations refer to all levels of abstraction in vision systems. Applications of these principles can be found in the entire reasoning chain. A complete review covering all aspects of the topic is beyond scope of this paper. In the following, we therefore restrict ourselves to selected examples that demonstrate the usefulness of stochastic approaches in computer vision.

2. MANIFESTATIONS

In the following, we discuss selected examples. We do not discuss representations for uncertainty. We implicitly assume that the uncertainty to be represented can be represented by the first two moments of a probability distribution. Without more specific information, this is justified by the maximum entropy argument, which states that, under certain assumptions, the majority of probability distributions that could have generated a finite data set must have an entropy close to the maximum entropy distribution [12].

¹ The text was submitted by the author in English.

2.1. Removal of Unexplainable Thresholds and Tuning Parameters

In the following, the replacement of thresholds and tuning parameters by better explainable parameters will be presented. For example, significance levels will be introduced for decisions by hypothesis tests. This will be discussed in the context of feature extraction, object reconstruction, and classification.

2.1.1. Feature extraction. Many image processing algorithms expect, for instance, the image noise σ_n as an input parameter. Often this parameter influences the results considerably—at least quantitatively, i.e., based on the number of extracted features. In the following we assume that the image noise has been estimated based on repeated measurements, image models, or gradient information [13, 14].

Edge Extraction. Edge extraction methods usually rely on a threshold for the magnitude $\|\nabla g\|^2 = g_r^2 + g_c^2$ of the local intensity gradients g_r and g_c . Since the expectation value of the gradients is $E(g_r) = E(g_c) = 0$, the extraction process can be interpreted as testing the hypothesis that the magnitudes of the gradients are significantly nonzero. With the standard deviations of the gradients derived by error propagation, the test statistic is

$$\frac{g_r^2}{\sigma_{g_r}^2} + \frac{g_c^2}{\sigma_{g_c}^2} \sim \chi_2^2, \tag{1}$$

which is χ^2 -distributed with two degrees of freedom. Of course, the independence assumptions do not hold exactly. The precision of the position and orientation of the edge element can be given with the known image noise level σ_n [15].

2.1.2. Reconstruction: Statistical interpretation of linear regularization. Another example for the possibility to remove a parameter is given in the context of linear regularization. As an example, we consider the determination of a regularly spaced height profile $u_j = f(x_j)$ from possibly irregular spaced heights $z(x_i)$ observed at the positions x_i . Approximation by the non-parametric graph $z = f(x)$ results in a 1.5 D representation of the profile.

If there are fewer observations than unknown heights, the problem is ill-posed and we have to introduce prior information to obtain a unique solution. With implicit smoothness assumptions about the profile, the functional of the standard Tikhonov regularization reads

$$E_1 = \sum_i e_z^2(x_i) + \lambda^2 \int [u''(x)]^2 dx \tag{2}$$

with the observations

$$z(x_i) + e_z(x_i) = \alpha u_j + (1 - \alpha) u_{j+1} \tag{3}$$

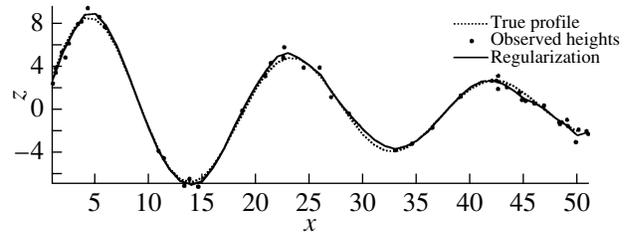


Fig. 1. The true profile, the observed heights with standard deviation $\sigma_z = 0.5$, and the reconstructed profile for an expected minimal curvature radius $r_{\min} = 0.72$.

being a linear interpolation of the unknown heights u_j and the regularization parameter λ^2 . This factor controls the trade-off between the “roughness” of the solution and the infidelity of the data. In a statistical framework, a discrete version of the energy reads [16]

$$E_2 = \frac{1}{N} \sum_i \frac{e_z^2(x_i)}{\sigma_z^2} + \frac{1}{K} \sum_j \frac{\kappa^2(u_j)}{\sigma_\kappa^2}, \tag{4}$$

assuming N uncorrelated observations. The K curvatures are approximated by the finite differences

$$\kappa(u_j) = \frac{u_{j-1} - 2u_j + u_{j+1}}{d^2}, \tag{5}$$

where d is the spacing of the height profile sought.

Multiplying (4) by $N\sigma_z^2$ and comparing it with (2) reveals that the regularization parameter is determined by [15]

$$\lambda^2 = \frac{\sigma_z^2 N}{\sigma_\kappa^2 K}, \tag{6}$$

where the ratio N/K accounts for the different numbers of height observations and curvature “observations.” For the determination of a proper regularization parameter λ^2 or the variances σ_z^2 and σ_κ^2 , there exist various approaches, e.g., cross validation [17], variance components estimation [18], and the method of maximal curvature of the L-curve (cf. Figs. 1 and 2 for an example).

These methods are not appropriate for real-time applications since they are somewhat expensive and time-consuming. However, knowledge of the stochastic properties of the observations or of the expected profile offers simple opportunities to set this parameter suitably:

- If the average variance of the observed heights is known, the variance of the curvatures can be derived from (5) by error propagation, i.e., $\sigma_\kappa^2 = 6\sigma_z^2/d^4$.

- If the variance σ_z^2 is unknown, an expected minimal curvature radius r_{\min} can be specified which sub-

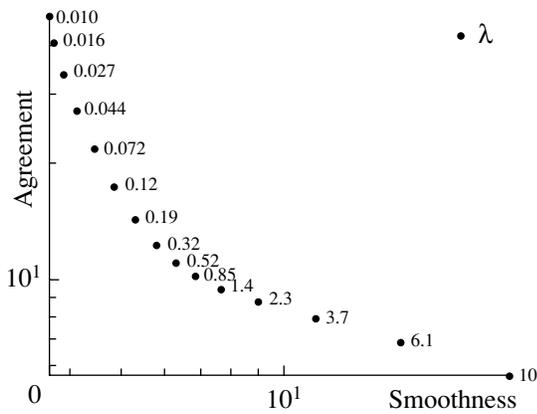


Fig. 2. The sum $\sum e_z^2(x_i)$ as a function of $\sum \kappa^2(u_j)$ for various values of λ . A reasonable solution of $\lambda = 0.52$ can be read off at the bend of the L-curve, which corresponds to a minimal curvature radius of $r_{\min} = 0.72$ (cf. Fig. 1 also).

sumes the expected roughness of the profile with a sound interpretable quantity: Assuming uncorrelated, normally distributed observations $\kappa \sim N(0, \sigma_\kappa^2)$, the probability for the curvature κ lying within the interval $\pm 2\sigma_\kappa$ is $P(-2\sigma_\kappa < \kappa < 2\sigma_\kappa) = 95\%$ and thus $\sigma_\kappa = 1/(2r_{\min})$ holds.

Of course, these relations hold only for values with a problem-specific, reasonable magnitude.

Within a statistical framework, model violations due to non smooth objects causing ridges and breaks can be considered up to some degree also [19]: By applying robust estimation procedures, the influence of outliers can be reduced. This affects both the observed point heights $z(x_i)$ and the fictitious curvature observations $\kappa(u_j)$.

2.1.3. Classification with rejection class. For many applications in uncontrolled environments, the assumption that the model base specifies all objects that might appear is not sustainable. Therefore, for the classification task, a rejection class is usually introduced to relax this closed world assumption [20].

This rejection of classifications can be done either because of large distances to the model instances or because of ambiguities. Both cases can be treated readily by applying a χ^2 -test for the Mahalanobis distances and considering likelihood ratios respectively [21].

2.2. Optimal Estimations and Self-Diagnostics

Within the framework of adjustment theory, we discuss the derivation of weights and covariances (1) to obtain optimal estimations and (2) to apply model validations.

2.2.1. Optimal estimation: Feature correspondences and weights. Knowledge of the variances and

correlations for observations offers the opportunity to obtain optimal parameter estimations. This will be discussed in examples for extracted and tracked interest points.

Optimal estimation. In multiple view geometry, the relations between different images and camera orientations are often formulated by algebraic entities, modeling mappings, or constraints for image features. Within the framework of algebraic projective geometry, these relations are often bilinear in the unknown parameters and the observations using a homogeneous representation for the entities. Thus, in order to minimize algebraic errors, the parameters can easily be derived by solving eigenvalue problems.

These solutions are statistical suboptimal because they do not take the correlations of the observations into account. Often the individual weights (variances) of the observations are not taken into consideration either. Furthermore, the solutions cannot easily be generalized to the estimation of multiple homogeneous entities with multiple constraints.

Therefore, the determination of realistic weights for the observations is desirable for computing optimal estimations.

Weights for extracted and tracked interest points. Common interest point operators search image windows for which the sum of the gradient magnitudes is large [22, 23]. As postulated by the settings of the operator parameters, the error characteristic for the point positions is usually homogeneous and isotropic. In this situation, the treatment of these uncertainties in the position is, in principle, not mandatory [24]. However, some multiview approaches enforce a uniform occurrence and distribution of the interest points in the image in order to achieve a favorable geometric configuration, e.g., by image tiling. It cannot then be expected that this characteristic holds any longer, and the uncertainties of the points can influence the result of the parameter estimation considerably [8]. Furthermore, if we want to use other image features, such as straight line segments for instance, the uncertainty has to be considered to use different image features combined.

In the context of stereo analysis, point correspondences can be in principle established by tracking or by matching [25]. In both cases, the uncertainty of the point positions can be determined without much effort [24]. Consideration of these uncertainties affects the uncertainties of the parameters describing the image orientations as well as positional uncertainties of the reconstructed objects in space.

The left side of Fig. 3 shows enlarged the estimated standard error ellipses for the positions of the interest points extracted with subpixel precision in a first image. The right side shows the tracked image points together with their standard error ellipses of the coordinates. As is to be expected, the overall appearances of the error characteristics are similar in shape and magnitude for tracking and matching.

2.2.2. Self-Diagnostics: Model validation. Within the theory of adjustment, the stochastic properties of the observations are collected in a stochastic model. An initial covariance matrix $\Sigma_{ll}^{(0)}$ of the observations \mathbf{l} is assumed to be known and related to the true covariance matrix Σ_{ll} by

$$\Sigma_{ll} = \sigma_0^2 \Sigma_{ll}^{(0)}, \quad (7)$$

with the possibly unknown scale factor σ_0^2 . As pointed out in the previous paragraph, the covariance matrix Σ_{ll} of the observations \mathbf{l} is often available.

If the initial covariance matrix reflects correctly the uncertainty of the observations, the variance factor is $\sigma_0^2 = 1$. The factor can be estimated from the residuals $\hat{\mathbf{e}}$ via [18]

$$\hat{\sigma}_0^2 = \frac{\hat{\mathbf{e}}^T \Sigma_{ll}^{-1} \hat{\mathbf{e}}}{r}, \quad (8)$$

where r denotes the redundancy of the system.

If the mathematical model actually holds and the observations are normally distributed, the test statistic

$$T = \frac{\hat{\sigma}_0^2}{\sigma_0^2} \sim F_{r, \infty} \quad (9)$$

has the expectation value one and is Fisher distributed with r and ∞ degrees of freedom. If the test is not rejected, data and model fit. Otherwise there is no evidence for conclusions such as incorrect models, presence of outliers, or wrong covariances for the observations. In particular, small errors in the assumptions concerning the precision $\Sigma_{ll}^{(0)}$ of the observations lead to a rejection of the hypothesis. Therefore, initially the usefulness of test statistic (9) is rather limited.

However, outliers can be eliminated by applying robust estimation schemes or outlier detection by hypothesis tests. If we furthermore have reasonable weight matrices Σ_{ll}^{-1} for possibly correlated observations, $\sigma_0^2 = 1$ holds and model violations can be detected.

Thus the hypotheses test (9) enables the validation of functional models if we assume observations to be free of outliers and realistic weight matrices to be within the adjustment procedure.

2.3. Performance Evaluation

The evaluation of results with respect to reference data (ground truth) refers to the values of estimated quantities and as well as to decisions.

2.3.1. Empirical accuracy of estimated parameters. For the evaluation of the empirical accuracy of n results \mathbf{x} with covariance Σ_{xx} , one can check the com-

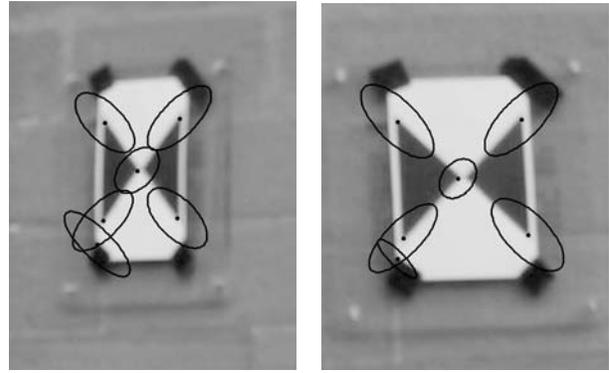


Fig. 3. On the left, the standard error ellipses for the positions of the extracted interest points, enlarged by the factor 50; on the right, the tracked point positions in a subsequent image of the sequence.

plete set in a combined hypothesis test. The Mahalanobis distance

$$d = (\mathbf{x} - \mathbf{x}_0)^T (\Sigma_{xx} + \Sigma_{x_0x_0})^{-1} (\mathbf{x} - \mathbf{x}_0) \quad (10)$$

is χ_n^2 -distributed and includes the covariance matrix $\Sigma_{x_0x_0}$ of the reference data \mathbf{x}_0 . Thus, the often-made assumption of error-free reference data is relaxed. If $d > \chi_{\alpha; n}^2$ holds, it can be concluded that the accuracy potential of the observations is not exploited—provided that the reference data actually are correct and normally distributed.

2.3.2. Competing classifiers. As pointed out by P. Cheeseman, one has to distinguish between probability and the uncertainty of probability [2]. Since probability values depend on the evidence (information) used to derive them, they are random variables. The importance of this insight will be demonstrated in the context of the comparison of competing classifiers based on the evaluation of a confusion matrix.

Error rate estimation. The classification error or simply the error rate $P(e)$ has been stated to be the “ultimate measure” for the performance of a classifier [26]. If τ is the number of misclassified test samples out of n samples, the maximum-likelihood estimator of $P(e)$ is

$$\hat{P}(e) = \frac{\tau}{n}. \quad (11)$$

The number τ is a random variable because it depends on the specific training and test sets used. Therefore, τ and $\hat{P}(e)$ have associated confidence intervals according to the binomial distribution.

The table illustrates the necessity of considering these properties for two competing classifiers 1 and 2. Without further information, classifier 2 would be preferred because of its lower classification error rate. By knowing the number of samples used to determine these rates, it turns out that the confidence interval for

Confidence intervals for the error rates of two classifiers 1 and 2 computed for a significance level of $\alpha = 0.05$

Classifier	Error rate	Confidence interval	Sample size n
1	0.20	[0.145, 0.225]	200
2	0.15	[0.06, 0.260]	50

the classifier 2 is quite large. This might be not acceptable for applications that demand low misclassification rates. Furthermore, in this situation other criteria such as reliability or computational speed might weigh more.

Receiver operating characteristic. Characterizing the performance of a classifier by the error rate (11) is not adequate for all pattern recognition systems. For binary classifiers the Receiver Operating Characteristic (ROC) curve permits the system designer to assess the performance of the recognition system at various thresholds in the decision rule [26]. For each threshold value, the correct detection rate is plotted versus the false acceptance rate. The convex hull of all sample points yields the ROC curve. Figure 4 shows examples of the ROC curves for two different classifiers.

For the evaluation of the overall performance, several measures have been suggested. Among them, the value A for the area under the curve has a proper statistical meaning: The value $A \in [0, 1]$ reflects the probability that a randomly chosen point from class a is correctly rated or ranked with greater suspicion than a randomly chosen point from class b [27].

Which classifier performs significantly better than the other cannot, without further information, be stated

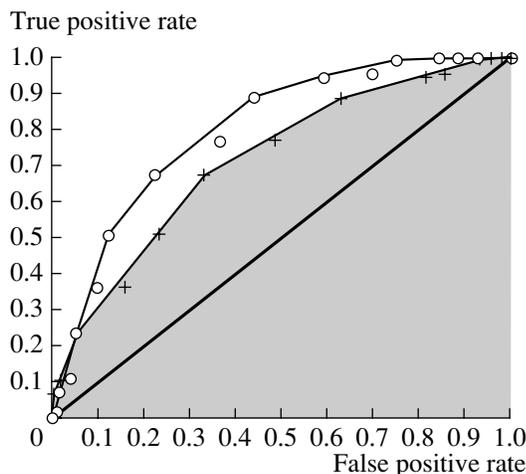


Fig. 4. ROC curves for two classifiers. The gray area for one classifier indicates the area under the curve. The diagonal results for a random classifier.

at all. But with rather general assumptions, the variance of the area can be specified via [27]

$$\sigma_A^2 = \frac{A(1-A) + (n_a - 1)(Q_1 - A^2)}{n_a n_b} + \frac{(n_b - 1)(Q_2 - A^2)}{n_a n_b} \quad (12)$$

with sample sizes n_a and n_b for the two classes, $Q_1 = A/(2 - A)$, and $Q_2 = 2A^2/(1 + A)$. Thus, the uncertainty of A is in an essential way influenced by the number of samples used to determine the corresponding ROC curve.

Whether two classifiers 1 and 2 differ significantly in their performance can be checked with the hypothesis test $A_1 = A_2$ vs. $A_1 \neq A_2$. The denominator of the test statistic

$$T = \frac{(A_2 - A_1)^2}{\sigma_{A_1}^2 + \sigma_{A_2}^2 - 2\sigma_{A_1 A_2}} \quad (13)$$

follows from error propagation for the area difference and takes the correlation of the areas A_1 and A_2 into account [28].

3. CONCLUSIONS

It has been demonstrated that the benefits of using stochastic approaches and interpretations are manifold:

- Unexplainable thresholds and parameters can be removed or replaced by quantities with a sound interpretation. This gives hints about how to generalize the procedures.

- The reduction of the parameter set to a small number of well-defined, meaningful parameters paves the way toward full automatic vision systems and real-time applications.

- The provision of quality measures permits the user to decide on the usefulness of algorithms for specific applications.

- Vision modules that require and provide quality measures can be integrated into the reasoning chain in a consistent manner. For example, the uncertainties of extracted image features can be transmitted to the estimation of parameters describing the image orientation (Section 2.2.1). Also, the (secondary) results of the estimation step can again be used for model validation (Section 2.2.2).

- The determination of realistic weights and correlations for observations offers the opportunity to obtain optimal estimates, to consider features of different kinds, to fuse information from different sensors, and to apply hypothesis tests.

- The use of stochastic models enables self-diagnostics, which is essential for fully automatic systems. Model validation and selection can be applied.

REFERENCES

1. D. Mumford, "The Dawning of the Age of Stochasticity," *Mathematics: Frontiers and Perspectives*, Amer. Math. Soc., Providence, RI 197–218 (2000).
2. P. Cheeseman, "In Defense of the Probability," in *Proceedings of the Ninth International Joint Conference of Artificial Intelligence (IJCAI-85)* (Los Angeles, California, Morgan Kaufmann, 1985), pp. 1002–1009.
3. S. Heuel, *Uncertain Projective Geometry. Statistical Reasoning in Polyhedral Object Reconstruction* (Lecture Notes in Computer Science, Springer, 2004) vol. 3008.
4. W. Förstner, "Uncertainty and Projective Geometry," Ed. by E. Bayro-Corrochano, in *Handbook of Computational Geometry for Pattern Recognition* (Computer Vision, Neurocomputing and Robotics, Springer, 2004).
5. M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for model Fitting with Applications to Image Analysis and Automated Cartography," *Communications of the Association for Computing Machinery* **24** (6), 381–395 (1981).
6. O. Chum and J. Matas, "Matching with PROSAC—Progressive Sample Consensus," in *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR), 2005*, vol. 1, pp. 220–226.
7. E. Michaelsen et al., "Estimating the Essential Matrix: GOODSAC versus RANSAC," *Photogrammetric Computer Vision PCV'06. Symposium of ISPRS Commission III* (vol. XXXVI, part 3 of IAPRS, Bonn, Germany, 2006), pp. 161–166.
8. A. Criminisi, "Accurate Visual Metrology from Single and Multiple Uncalibrated Images," *Distinguished Dissertations* (Springer, London, Berlin, Heidelberg, 2001).
9. S. Utcke, "Grouping based on Projective Geometry Constraints and Uncertainty," in *Proceedings of the Sixth International Conference on Computer Vision* (Narosa Publishing House, 1998), pp. 739–746.
10. P. H. S. Torr et al., "The Problem of Degeneracy in Structure and Motion Recovery from Uncalibrated Image Sequences," *International Journal of Computer Vision* **32** (1), 27–44 (1999).
11. K. Kanatani, "Uncertainty Modeling and Model Selection for Geometric Inference," *IEEE Transactions on Pattern Recognition and Machine Intelligence* **26** (10), 1307–1319 (2004).
12. E. T. Jaynes, "On the Rationale of Maximum Entropy Methods," in *Proceedings of the IEEE, 1982*, vol. 70, pp. 939–952.
13. M. Luxen and W. Förstner, "Characterizing Image Quality: Blind Estimation of the Point Spread Function from a Single Image," in *Proceedings of the PCV'02 Symposium, Graz, 2002*, p. 205.
14. C. Liu et al., "Noise Estimation from a Single Image," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2006*, pp. 901–908.
15. W. Förstner, "10 Pros and Cons Against Performance Characterization of Vision Algorithms," Ed. by H. I. Christensen et al. in *ECCV Workshop on Performance Characteristics of Vision Algorithms* (Cambridge, UK, 1996).
16. U. Weidner, "Parameterfree Information-Preserving Surface Restoration," Ed. by J.-O. Eklundh, in *Computer Vision—ECCV '94* (Vol. 801, Lecture Notes in Computer Science, 1994), vol. II, pp. 218–224.
17. P. Craven and G. Wahba, "Smoothing Noisy Data with Spline Functions," *Numerische Mathematik* **31**, 377–403 (1979).
18. "Manual of Photogrammetry," Ed. by J. C. McGlone, et al., *American Society of Photogrammetry and Remote Sensing*, 5th edition (2004).
19. A. Blake and A. Zisserman, *Visual Reconstruction* (MIT Press, Cambridge, 1987).
20. W. B. Mann and T. O. Binford, "Probabilities for Bayesian Networks in Vision," in *Image Understanding Workshop 1994* (San Francisco, CA, Morgan Kaufmann, 1994), pp. 633–643.
21. R. Muzzolini, et al., "Classifier Design with Incomplete Knowledge," *Pattern Recognition* **31** (4), 345–369 (1998).
22. W. Förstner and E. Gülch, *A Fast Operator for Detection and Precise Location of Distinct Points, Corners and Centres of Circular Features* (ISPRS Intercommission Workshop, Interlaken, 1987), pp. 273–310.
23. C. G. Harris and M. J. Stephens, "A combined corner and edge detector," *Fourth Alvey Vision Conference, 1988*, pp. 147–151.
24. Y. Kanazawa and K. Kanatani, "Do We Really Have to Consider Covariance Matrices for Image Features?" *International Conference on Computer Vision, 2001*, pp. 301–306.
25. B. T. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," in *Proceedings of Image Understanding Workshop, 1981* pp. 212–130.
26. A. K. Jain, et al., "Statistical Pattern Recognition: A Review," *IEEE Transactions on Pattern Recognition and Machine Intelligence* **22** (1), pp. 4–37 (2000).
27. J. A. Hanley and B. J. McNeil, "The Meaning and the Use of the Area under a Receiver Operating Characteristics (ROC) Curve," *Radiology* **143** (1), 29–36 (1982).
28. J. A. Hanley and B. J. McNeil, "A Method of Comparing the Areas under Receiver Operating Characteristic Curves Derived from the Same Cases," *Radiology* **148** (8), pp. 839–843 (1983).



Jochen Meidow studied surveying and mapping at the University of Bonn, Germany, and graduated with a diploma in 1996. As research associate at the Institute for Theoretical Geodesy, University of Bonn, he received his PhD degree (Dr.-Ing.) in 2001 for a thesis about aerial image analysis. Between 2001 and 2004 he was a postdoctoral fellow at the Institute for Photogrammetry, University of Bonn, and since 2004 he is with the Research Institute for Optronics and Pattern Recognition (FGAN-FOM) in Ettlingen, Germany. He is a member of the DAGM (German Pattern Recognition Society). His research interests are adjustment theory, statistics, and spatial reasoning.