Diagnostics and Performance Evaluation in Computer Vision

Wolfgang Förstner, Institute of Photogrammetry
Bonn University, e-mail: wf@ipb.uni-bonn.de

Contents

Abstract

Increasing the performance of Computer Vision algorithms requires both, robust procedures for handling non-modelled errors and diagnostic tools for achieving autonomy in the evaluation of the achieved results.

The role of diagnostic tools for model evaluation and performance prediction is discussed.

Quality or performance refers to:

1. the precision of the estimated quantities (efficiency) 2. the sensitivity of the estimated quantities with respect to systematic and gross errors. 3. the design of the experiment or the used actual data. 4. the correctness of results and of reports on the correctness of the result.

The performance may be evaluated a. by controlled tests using simulated or real data. This is necessary to prove either the usefulness of the algorithms or the adequateness of the used model. b. by diagnostic tools. This is necessary for achieving autonomy in the chain of automatic procedures within a complete system where generally no reference data are available.

The performance of Computer Vision algorithms can be significantly increased by diagnostic tools, both by detecting singular or weak configurations within high break down estimation, e. g. RANSAC, and by providing a highly reliable selfdiagnosis of the algorithm itself using the internally available redundancy.

Results from extensive empirical tests demonstrate the feasibility of the proposed tools.

1 Motivation

One of the research goals in Computer Vision is to provide tools for *automatic* scene interpretation, which may be used within a system that at least partially relies on information provided by imagery. The quality of the vision tool, as any other tool for supporting automation, is as good as its ability to provide a welldefined output. According to Pregibon [1986] a tool can be termed to be *successful* if

1. *it can handle XX % of all new problems it faces.*
2. *for the (100 – XX) % it cannot handle, it* knows *it.*
3. *You are happy with XX.*

This implicitely requires the tool to contain means for a reliable *selfdiagnosis*.

The design of vision algorithms is hard not only because the lack of a coherent theory (Haralick [1985]) but also because the lack of tools and knowledge to *predict the performance* even for comparably simple tasks as the extraction of image features, the reconstruction of visible surfaces, or the determination of the orientation of the camera from given points. *Error handling* therefore is a central issue in image analysis.

A typical example is the determination of the six parameters for translation and rotation of a camera, which forms a link between observed image features and geometric parameters:

- The *correspondence problem* is far from being solved for general cases. Existing solutions have to deal with large percentages of matching errors. This prevents the direct use of classical estimation procedures and requires to look for robust procedures which, however, make a thorough analysis of the quality of the final result at least difficult, as the underlying theories (!) often only give asymptotic theorems.

- In case *approximate values* are not available or only of poor quality their determination appears to be a far more challenging problem than the refinement via a least squares estimation. The direct solutions, either with minimum or redundant information play a central role, especially in the presence of outliers.

- *Self-Calibration*, where calibration, orientation and generally also scene reconstruction is performed simultaneously, is often required as camera calibration in a laboratory often is not feasible or not sufficient. It increases the difficulty of error analysis by at least one order of magnitude as deficiencies in design, modelling and mensuration have to be handled simultaneously and therefore prevent an algebraic analysis of the system. The difficulty of integrating *all* types of observational values lies in the necessity to formalize the evaluation process in order to adequately handle the different dimensions (pixels, meter, radiants, etc.) of the observations and their influence on the final result.

The paper is motivated by the urgent need to provide tools for analysing the result of vision algorithms both by exploiting the redundancy in the data and by controlled tests. It stresses the need to complement the strong tools from robust statistics by diagnostic tools as both are intimately linked (cf. HUBER [1991]).

The paper is organized as follows: Section 2 discusses the performance evaluation of automatic procedures at a systems level elucidating the places within a vision algorithm where diagnostic tools may come into play. As many Computer Vision tasks can be formalized as estimation, section 3 discusses the different types of errors and the issues in error handling. Section 4 provides the tools for the solution of three classical tasks, 4.2 evaluating the *stability* of estimates in relation to specified requirements, 4.3 evaluation the *sensitivity* of the estimates with respect to model errors, and 4.4 *evaluating* the result of *controlled tests*. The examples in section 5, taken from classical tasks in Computer Vision, demonstrate the feasibility of the concepts, based on extensive tests.

2 Performance Evaluation of Automatic Procedures

Evaluation of the the performance can rely either on internal or external knowledge. We assume the procedures to be evaluated to be describable as a black box with some specified *types* of input and output data. The output data consist of the *result* as such and its *quality*. E. g. a feature detector may give the coordinates of distinct points in the images as result together with their covariance matrix as quality description as output, derived from the raster data of the image as input.

The *internal* evaluation of the result is based on the model, the structure and the values of the given data. It exploits the redundancy in the data with respect to the model as far as possible and is based on an estimate for the accuracy of the given data, hypothesis tests and the analysis of the homogeneity of the data.

The *external* evaluation is based on a comparison of the result of the procedure with some reference. This reference itself may be precisely known, e. g. in the context of simulations, or be derived by some independent measurements, which will have an usually low uncertainty. External evaluation is necessary to prove the validity of the internal model of the procedure. It requires objective measures for evaluating the differences between experiment and reference data. The quality of the difference in general is influenced by both the uncertainty of the result, which may obtained by the internal evaluation and the uncertainty of the reference data.

Thus at this level of evaluation we have four cases depending on whether the result is actually correct and whether the report of the correctness of the result, provided by the internal evaluation is actually correct (cf. Table 1).

The first aim of the development of an algorithm is to increase the probability $p_0 = p_{00} + p_{01}$ so that it actually yields correct results. In the case of no selfdiagnosis, the result is assumed to be correct, thus the percentage of erroneous decisions is $p_1 = 1 - p_0$ leading to costs $C^{(0)} = C_1 p_1$, where C_1 are the expected costs per failure.

Integrating selfdiagnosis, however, changes the scope of the development, e. g. to maximize the probability of correct decisions $p_c = p_{00} + p_{11}$ while staying with p_{00} larger than some specified requirements (cf. XX % in the introduction). The costs $C^{(1)}$ for making wrong decisions now is $C^{(1)} = C_{01}p_{01} + C_{10}p_{10}$ which may be much smaller than $C^{(0)}$ as the costs $C^{(1)}$ (assumed to be $= C_{10}$) for making a decision error of type II (false positives) usually is much larger than the costs for making a decision error of type I and the probability of making erroneous decisions may be made much less than p_1 by diagnostic tools, which are meant to identify incorrect results with a high probability p_{11}.

We therefore will find diagnostic tools to be useful for:

- increasing the performance of an algorithm by increasing the likelihood to provide correct results as such (p_{00}). This refers first to *proper planning* of the design of the experiment, i. e. the boundary conditions for the application of the procedure. This will provide the basis for robust procedures to work reliably, as they require some homogeneity, i. e. lack of leverage points. The

Table 1: shows the possible result of empirical tests together with their probabilities

		report of selfdiagnosis: result is	
		correct	wrong
reality	correct	correct decision p_{00}	wrong decision (I) p_{01}
	wrong	wrong decision (II) p_{10}	correct decision p_{11}

diagnostic tools, however, may also be used to increase the performance of estimators themselves which rely on a stochastic algorithm, as e. g. RANSAC or median least squares estimators, by identifying singular or instable configurations.

- increasing the reliability of the quality measures for self evaluation in order to identify incorrect results (p_{11}).

While this discussion holds for any type of automatic procedure we will concentrate onto tasks which are formalizable as parameter estimation.

3 Parameter Estimation

This section describes the used model especially to introduce the notation, discusses the error types and specifies the issues in error handling in more detail.

3.1 The Model

Let us assume the model to explicitely describe the observation process

$$E(\boldsymbol{l}) = \boldsymbol{g}(\boldsymbol{\beta}) \tag{1}$$

where the expectation of the n observations $\boldsymbol{l} = \{l_i\}$ via $\boldsymbol{g}$ in general nonlinearily depends on the u unknown parameters $\boldsymbol{\beta} = \{\beta_j\}$. The stochastical properties of the observations are captured by the covariance matrix

$$D(\boldsymbol{l}) = \boldsymbol{\Sigma} = \hat{\sigma}_0^2 \boldsymbol{Q}, \tag{2}$$

which may be split into a variance factor $\hat{\sigma}_0^2$ and a weight coefficient matrix $\boldsymbol{Q}$ being the inverse of the weight matrix $\boldsymbol{P}$.

In case this is the only information available the principle of maximum entropy results in the following full model

$$\boldsymbol{l} \sim N(\boldsymbol{g}(\boldsymbol{\beta}), \boldsymbol{\Sigma}) \tag{3}$$

hypothesizing $\boldsymbol{l}$ to be normally distributed. The redundancy of the system is

$$r = n - u \tag{4}$$

In case the redundancy equals 0 or in the unlikely case the observations are consistent, the assumed stochastical properties have no influence onto the estimate. The only task then is to invert (1) to obtain $\hat{\boldsymbol{\beta}} = \boldsymbol{g}^{-1}(\boldsymbol{l}_s)$, where $\boldsymbol{l}_s$ is a subset of $\boldsymbol{l}$ of size u.

3.2 Error Types

In general all components of the model will have an influence onto the result. The key question is how an automatic system handles errors in these assumptions. One may distinguish three types of errors:

1. *Data errors* are errors in the values of $\boldsymbol{l}$. They relate to points, lines or other features in the image or in object space, where measurements are taken. They may really be mensuration errors, e. g. caused by failures in the detection algorithm or matching errors leading to wrong relations between image and object features. Depending on the complexity of the scene and the quality of the used algorithms the percentage of errors may range between a few and over 80 % of the observed values.

2. *Model errors* refer to all three parts of the model: the functional relation $\boldsymbol{g}$, the covariance matrix $\boldsymbol{\Sigma}_{ll}$ and the type of the distribution, here N. Examples for this type of error are manifold:

   - too few, too many or the wrong set of parameters, e. g. when using shallow perspective, projectivity or parallel projection.
   - wrong weighting, e. g. when assuming the same accuracy for all detected points,
   - neglected correlations, e. g. in Kalman-filtering or
   - wrong assumptions about the distribution, e. g. when handling one-sided errors.

   Observe that data errors and model errors formally cannot be distinguished; as a refinement of the model may always specify the type of error in the observations.

3. *Design- or configuration errors* relate to the complete set of functions $\boldsymbol{g} = \{g_i\}$. Such errors cause the estimate $\hat{\boldsymbol{\beta}}$ to be nonunique in some way. Multiplicity of solutions is the best case of nonuniqueness. Depending on the degree of redundancy we may distinguish at least three cases:

   (a) nondeterminable parameters. Critical surfaces belong to this class. An example would be a spatial resection with three points and the projection centre sitting on the critical cylinder.

(b) noncheckable observations or parameters. Here the determination of the parameters may be possible, but errors in the observed values of in the estimated parameters introduced in a Bayesian manner, are not detectable due to a too low redundancy. An example would be a spatial resection with three points in general position.

(c) nonlocatable errors. Here a test may be able to show discrepancies between the data and the model, but no identification of the error source is possible. An example would be a spatial resection with four points in general position.

We concentrate on errors which can be modelled as errors in the expectation of the observations (mean shift model).

3.3 Issues in Error Handling

There are at least two basic questions *automatic* procedures need to be able to answer:

1. How sensitive are the results?

   The results may be uncertain due to the large number of errors mentioned above. Evaluating real cases has to cope with the problem, that several such errors occur simultaneously. *Instabilities* may be hidden within a system of high total redundancy. Then we may discuss

   - determinability of parameters
   - controllability of errors and the effect of nondetectable errors
   - separability of error sources.

   We will formalize this classification in more detail and discuss the first two items explicitely.

2. How small is too small?

   Most algorithms are controlled by *thresholds* or tolerances to be specified by the developper or the user.

   When referring to observations or parameters thresholding may be interpreted as hypothesis testing, which allows to derive the thresholds by specifying a significance level and using error propagation. We will not pursue this topic.

   When evaluating the *design* the formalization becomes less obvious, e. g. when having a small basis in relative orientation (2D – 2D), small angles in spatial resection (3D – 2D) or small distances between the point in absolute orientation (3D – 3D). In all cases the configuration is close to singularity. But then the question arises: how to evaluate *small deviations from a critical surface*? We will show that a generic and formal answer to this question can be given which is based on the *local* geometry of the design.

The final question is "How good is the overall performance?" which requires empirical tests for comparison with real data. It is however not only necessary to look at the differences but to have a rigorous statistical test which checks whether the procedure exploits the information contained in the data. Moreover it should give a clear indication whether the internal performance measures can be used as prediction for the external performance.

The next section will collect the necessary tools needed for internal and external evaluation of estimation procedures.

4 Tools

4.1 Quality Insurance

Treating image understanding tasks as estimation problems allows us to fully exploit the rich arsenal of tools from estimation theory. Regarding the specific problem of data and model errors we specifically need to use the techniques available from robust statistics and regression diagnostics following two different aims (HUBER [1991]):

- The purpose of *robustness* is to have safeguards against deviations from the assumptions.
- The purpose of *diagnostics* is to find and identify deviations from the assumptions.

4.1.1 Robustness

There are two levels of robustness, depending on whether the errors are small or large. Data or model deviations are small in case linear approximations are sufficient. This leads to a rule of thumb that small deviations are deviations less than about 30 % of the values, including all functions of the observations. It e. g. corresponds to a requirement angular errors to be less than approx. 20^o.

1. Robustness with respect to *small deviations.*

   The socalled *influence curve* (HAMPEL et al. [1986]), which measures the effect of errors onto the result, may be used to measure the quality of robust procedures in this case. Maximum-likelihood (ML) type, or *M-estimators* are the corresponding tool to deal with small deviations.

2. Robustness with respect to *large deviations.*

   The *break down point* (ROUSSEEUW/LEROY [1987]) measuring the maximum allowable percentage of errors still guaranteeing the estimateor to yield results with limited bias, may be used to evaluate the quality of procedures in this case. Estimates with a high break down point, up to 50 %, such as least median squares, are the corresponding tool to handle large deviations.

Observe, that the effect of *random* errors onto the result is not covered by the term robustness. These effects usually are measured by the *precision* of the estimates. The reason for this distinction is that random errors are part of the original model thus do not represent deviations from the model, and are taken into account by all basic estimators, like least squares or ML-estimators.

4.1.2 Diagnostics

As already indicated above, there are three levels of diagnostics which all refer to small model errors:

1. *Determinability* of parameters or singularities in the estimation process all measure the instability of the design with respect to random perturbations.

   Standard deviations or in general covariance matrices are the diagnostic tool to detect such a situation. Due to the small size of the random errors a linear substitute model, derived by linearization, may be used to evaluate such instabilities.

   We will discuss this in detail in section 3.2.

2. *Controllability* of observations and detectability of model errors specify the applicability of hypothesis tests.

   The diagnostic tool are minimum bounds of the size of observational or model errors which can be detected by a test with a certain given probability. The *sensitivity* of the result is measured by the effect of nondetectable errors onto the result.

   Both tools may be used for planning as they do not depend on the actual measurements.

   The actual influence of the observations of model parameters measured in a leave-one-out fashion may be decisive for the acceptance of an estimate.

   We will discuss these tools in detail in section 3.3.

3. The *locatability* of observational errors of the separability of model errors specify the ability to correctly classify or identify the error causes.

   This can be described in terms of a confusion matrix, like in statistical pattern recognition, the difference being the here the entries of the confusion matrix depend on the expected size of the errors and on the design or configuration.

   The diagnostic tool therefore are lower bounds for observational errors or model errors which are identifiable or separable with a certain probability.

   In section 4.3 we will formally relate separability to controllability especially with repect to *sets* of observational model errors, but not discuss the notion in detail.

4.2 Instabilities of Estimates or "How Small is too Small?"

Instabilities of parameters occur in case the configuration produces some critical manifold (surface) the solution belongs to. One usually distinguishes:

1. Singularities of the first kind. Here a complete manifold of the parameters is consistent with the observations.

2. Singularities of the second kind. Here small deviations in the observations result in large deviations in the parameters.

An example for a singularity of the first kind is the critical cylinder in spatial resection. It may be formulated as a rule: `IF the projection center` $O \in$ `cylinder`(P_1, P_2, P_3) `THEN` O `is not determinable.`

This rule is the result of an algebraic analysis and due to its theoretical character generally valid in the context of spatial resection and precise.

Such algebraic results, however, have some disadvanteges:

- The statements do not contain any information how to evaluate *deviations from the ideal configuration.*

- The statements do not give any hint to *generalize* to other situations. Other problems, e. g. relative orientation require a separate analysis.

- The statements do not give any *means to evaluate* the orientation even of one image *within a set* of several images to be oriented simultaneously. It may very well be, that in a multi-image setup with a large redundancy the orientation of one of the images can not be determined due to the presence of the above situation.

Such *hidden instabilities* reveil the limitation of purely algebraic approaches which can only be applied to very restricted situations and cannot be generalized.

Thus algebraic techniques cannot be easily transferred into automatic procedures evaluating the stability of an estimate. The solution to this dilemma is based on the observation, that the instabilities are local properties in parameter space and can be fully analysed using the covariance matrix of the parameters. This leads to a shift of the problem. Instead of a deterministic analysis we now are confronted with the problem of evaluating the quality of a covariance matrix. The shift of the problem and it's solution goes back to (Baarda [1973]).

The evaluation method consists of two steps:

1. Specification

   Specifying the user requirements in terms of a so-called *criterion matrix*, say $\boldsymbol{H}$, which gives an upper bound on the desired covariance matrix, corresponding to the desired lowest precision.

2. Comparison

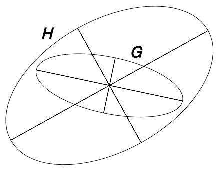
   Checking whether the achieved covariance matrix,

   $$\boldsymbol{G} \doteq \boldsymbol{\Sigma}_{\hat\beta\hat\beta} = (\boldsymbol{X}^T \boldsymbol{\Sigma} \boldsymbol{X})^{-1} \qquad (5)$$

   is better than $\boldsymbol{H}$.

We will discuss this comparison first.

Figure 1: shows the relation $G < H$ between two 2×2 covariance matrices G and H, represented by isolines of constant probability density of the corresponding normal distribution.



4.2.1 Comparing Covariance Matrices

The *comparison* of covariance matrices is interpreted as the requirement the standard deviation of an arbitrary function $\boldsymbol{f}$ to be better when calculated with covariance matrix $\boldsymbol{G}$ than with $\boldsymbol{H}$

$$\boldsymbol{G} \leq \boldsymbol{H} \qquad \cong \qquad \sigma_f^G \leq \sigma_f^H, \text{ with } f = \boldsymbol{e}^T \hat{\boldsymbol{\beta}}, \text{ for all } \boldsymbol{e} \tag{6}$$

Using error propagation, e. g. $\sigma_f^G = \sqrt{\boldsymbol{e}^T \boldsymbol{G} \boldsymbol{e}}$ this leads to (cf. Fig. 1)

$$\boldsymbol{e}^T \boldsymbol{G} \boldsymbol{e} \leq \boldsymbol{e}^T \boldsymbol{H} \boldsymbol{e}, \text{ for all } \boldsymbol{e} \tag{7}$$

or

$$\lambda = \frac{\boldsymbol{e}^T \boldsymbol{G} \boldsymbol{e}}{\boldsymbol{e}^T \boldsymbol{H} \boldsymbol{e}} \leq 1 \tag{8}$$

which requires the determination of the maximum eigenvalue of

$$\boldsymbol{G} \boldsymbol{e} = \lambda \boldsymbol{H} \boldsymbol{e}. \tag{9}$$

The squareroot $\sqrt{\lambda_{max}}$ indicates the maximum ratio of the actual and the required standard deviation.

This evaluation may be simplified using

$$\boldsymbol{K} = \boldsymbol{H}^{-1/2} \boldsymbol{G} \boldsymbol{H}^{-1/2} \tag{10}$$

$$\lambda = \frac{\boldsymbol{e}^T \boldsymbol{K} \boldsymbol{e}}{\boldsymbol{e}^T \boldsymbol{e}} \leq 1 \tag{11}$$

which is equivalent to

$$\lambda_{\max}(\boldsymbol{K}) \leq 1. \tag{12}$$

Equation (10) is favorable in case $\boldsymbol{H}$ easily can be diagonalized (cf. the example below).

The maximum eigenvalue of $\boldsymbol{K}$ in (10) may be replaced by a less tight norm in order to avoid the rigorous determination of the maximum eigenvalue, e. g. the trace:

$$\lambda_{\max}(\boldsymbol{K}) \leq tr \boldsymbol{K} \leq 1. \tag{13}$$

4.2.2 Specification of a Criterion Matrix

The *specification* of a criterion matrix can be based on the covariance matrix $\boldsymbol{\Sigma}_{\hat{\beta}\hat{\beta}}$ derived from an ideal configuration. This has the advantage that the user can easily interprete the result. In case an ideal configuration cannot be given the criterion matrix $\boldsymbol{H} = \boldsymbol{SRS}$ may be set up by specifying the standard deviations σ_i, collected in a matrix $\boldsymbol{S} = Diag(\sigma_i)$ and correlations ρ_{ij}, collected in a matrix $\boldsymbol{R} = \rho_{ij}$, derived from some theoretical considerations, e. g. interpreting the sequence of projection centres in a navigation problem as stochastic process, where the correlations ρ_{ij} depend only on the time or space difference between points P_i and P_j.

The method is able to capture various deficiencies in the design of the configuration of an orientation procedure, without having to discriminate between different types of instabilities (cf. section 5. Such situations also may arise in more complex problems where an algebraic analysis is not possible whereas this method is able to find the instabilities.

When using this method for designing a configuration the eigenvector belonging to the largest eigenvalue gives insight into the most unprecise function of the parameters, which may be used to look for specific stabilization means.

4.3 Model Errors or "How Sensitive is the Result?"

The stability of an estimation, specifically an orientation, evaluated by the covariance matrix only takes random perturbations into account. The result, however, may be wrong due to gross errors, e. g. caused by the matching procedure. As well, an oversimplified model may lead to precise but incorrect results. Both error sources, blunders and systematic errors, only can be detected in the case of redundant observations. This is a necessary but - as we will see - not a sufficient condition. Redundancy allows us to perform tests on the validity of the assumed model *without* reference to additional data used during the estimation. Such tests may lead to the detection or even identification of the error source. Of course, the outcome of these tests may be false. Redundancy, however, increases the stability of the solution and the correctness of the outcome of statistical tests. The theory for performing such a test is described in the literature (cf. BAARDA [1967]/[1968], COOK/WEISBERG [1982], FÖRSTNER [1987]). The structure of that theory, it's use in estimation problems and examples from orientation procedures will be given.

4.3.1 Detectability and Separability of Errors

We first want to discuss the type of evaluation which can be performed depending on the redundancy r of a system.

1. $r = 0$ In the case of no redundancy, one can only evaluate the sensitivity of the result with respect to random errors as shown in the last section. No check of the observation is possible what-

soever. They may remain incorrect without any indication.

2. $r = 1$ In the case of redundancy $r = 1$, a check on the validity of the model is possible. The existence of blunders may be indicated, but they are not locatable, as a "leave-one-out test" always leads to a valid solution.

3. $r = 2$ A redundancy of $r = 2$ is necessary in order to be able to locate simple blunders. A leave-one-out test generally will be able to find the unique consistent set of observations. Double errors are not locatable, however their existence is usually indicated.

4. $r > 2$ For a larger redundancy, $r - 1$ errors are locatable, whereas r errors are only detectable.

The maximum number of detectable errors is $n/2$, i. e. 50 % of the data, as more than $n/2$ observations may mimic a good result. Thus, 50 % is the upper limit for the so-called *breakdown point* of an estimator. The breakdown point of an estimator is the minimum percentage of others which may cause the estimator to give wrong results, i. e. may lead to a bias of any size. The normal mean has the breakdown point 0, the median 50 %, an indication of it's higher robustness. Practical procedures may be better as they may use specific knowledge about the structure of the problem (cf. the straight line detection procedure by (ROTH/LEVINE [1990]).

In case of a small percentage ($< 1\%$) of not too large ($\leq 30\%$) gross errors, the detection and location may be based on the residuals

$$\boldsymbol{vv} = \boldsymbol{g}(\hat{\boldsymbol{\beta}}) - \boldsymbol{y} \qquad D(\boldsymbol{y}) = \sigma_0^2 \boldsymbol{Q} = \sigma_0^2 \boldsymbol{P}^{-1}\,. \quad (14)$$

Using the maximum likelihood estimate

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{P} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{P} (\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\beta}^{(0)}) \quad (15)$$

we can express changes $\Delta \boldsymbol{v}$ of the residuals in terms of changes, thus errors $\Delta \boldsymbol{y}$ of the observations

$$\Delta \boldsymbol{v} = -\boldsymbol{R} \Delta \boldsymbol{y} \quad (16)$$

with the projection matrix

$$\boldsymbol{R} = \boldsymbol{I} - \boldsymbol{U} \quad (17)$$

with the socalled hat-matrix (cf. HUBER [1991])

$$\boldsymbol{U} = \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{P} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{P} \quad (18)$$

(16) is graphically shown in Fig. 2.

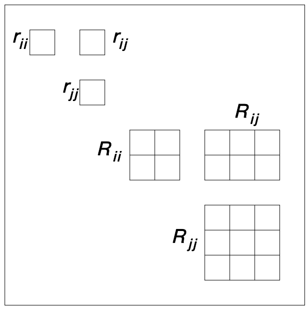
This matrix may be used to analyse the ability of the estimation system to apply selfdiagnosis with respect to errors in the observations, as only effects that can be seen in the residuals are detectable.

We distinguish two levels of evaluation

1. detectability or checkability; and,

2. separability or locatability.

Figure 2: shows the four cases for analysing the projection matrix R with respect to sensitivity (diagonal matrices) and separability (off-diagonal matrices) for single or groups of observations.



Both evaluation measures may refer to single or groups of observations. Thus we have 4 cases.

1. *Detectability* or checkability rely on the diagonal elements or diagonal submatrices of $\boldsymbol{R}$.

   a) *Single* observational errors can only be detected if

   $$r_{ii} > 0\,. \quad (19)$$

   The diagonal elements r_{ii} sum up to the total redundancy r, i.e. $\sum r_{ii} = r$. This indicates how the redundancy is distributed over the observations. The corresponding test statistics for detecting single errors for given σ_0 and uncorrelated observations is

   $$z_i = \frac{-v_i}{\sigma_0} \sqrt{\frac{p_i}{r_i}} \sim N(0\,,\,1) \quad (20)$$

   b) *Groups* of n_i observation can only be detected if the corresponding $n_i \times n_i$ submatrix

   $$\parallel \boldsymbol{R}_{ii} \parallel > 0 \quad (21)$$

   of $\boldsymbol{R}$ is nonsingular. Otherwise a special combination of observational errors may have no influence on to the residuals. The corresponding test statistic is

   $$T_i = \frac{1}{\sigma_0} \sqrt{\frac{\boldsymbol{v}_i^T \boldsymbol{R}_{ii} \boldsymbol{Q}_{ii} \boldsymbol{v}_i}{n_i}} \sim \sqrt{F_{n_i,\infty}} \quad (22)$$

   which reduces to (20). The observations may be correlated within the group, but must be uncorrelated to the others. $\sqrt{\boldsymbol{F}_{n_i m}}$ denotes the distribution of the square root of a random variable being $F_{n_i m}$-distributed.

2. *Separability* or locatability in addition to the diagonal elements of $\boldsymbol{R}$ rely on the off diagonals.

   a) The separability of two *single gross errors* evaluates the likelihood to correctly locate an error, i. e. to make a correct decision when testing both. The decisive measure is the correlation coefficient of the test statistics (20) which is

$$\rho_{ij} = \frac{r_{ij}}{\sqrt{r_{ii} \cdot r_{jj}}}. \qquad (23)$$

   Tables for probabilities of erroneous decisions when locating errors are given by FÖRSTNER [1983].

   Correlation coefficients below 0.9 can be accepted since the probability of making a false decision even for small errors remains below 15 % [1].

   b) The separability of two *groups of observations* $\boldsymbol{l}_i$ and $\boldsymbol{l}_j$ depends on the maximum value

$$\rho_{ij}^2 = \lambda_{max} \boldsymbol{M}_{ij} \qquad (24)$$

   of the $n_i \times n_j$ matrix

$$\boldsymbol{M}_{ij} = \boldsymbol{R}_{ij} \boldsymbol{R}_{jj}^{-1} \boldsymbol{R}_{ji} \boldsymbol{R}_{ii}^{-1} \qquad (25)$$

   which for single observations reduces to (23).

   No statistical interpretation is available due to the complexity of the corresponding distribution.

Example: *Detectability of Errors*

Relative orientation with 6 corresponding points yields a redundancy of $r = 6-5 = 1$. If the images are parallel to the bais and the points are situated symmetrically as shown in (3) then the diagonal elements r_{ii} are 1/12 for points i = 1, 2, 5 and 6 and 1/3 for points 3 and 4.

Obviously errors are hardly detectable if they occur in point pairs 1, 2, 5 or 6. In all cases no location of the false matches is possible as $r = 1$. □

Example: *Separability of Errors*

Spatial resection with 4 points symmetrically to the principle point is known to yield highly correlated orientation parameters. Depending on the viewing angle α, the correlation between the rotation ω (x-axis) and the coordinate y_0 of the projection centre, and between the rotation φ (y-axis) and the coordinate x_0 is (cf. Appendix)

[1]Precisely stated: If the larger of the two test statistics $|z_i|$ and $|z_j|$ in (20) is chosen to indicate the erroneous observation, if the critical value is 3.29, corresponding to a significant level of 99.9 % and a single error can be detected with a probability higher than 80 %, then the probability of making a wrong decision between l_i and l_j is approximately 13 %.

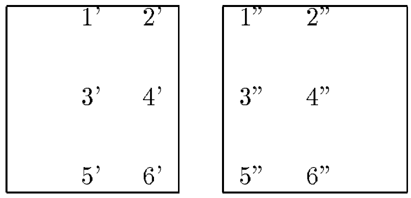


Figure 3: Numbering of 6 points in a stereo pair.

$$\mid \rho \mid = \frac{1}{\sqrt{1 + \sin^4 \frac{\alpha}{2}}} \qquad (26)$$

For a CCD-camera with a focal length of f = 50 mm and sensor size of 5 mm, $\alpha/2 = 1/20$ thus $\mid \rho \mid =$ 0.999997. For an aerial camera RMK 15/23 with a focal length of 15 cm and image size of 23 cm, $\alpha/2 =$ 2/3, thus $\mid \rho \mid = 0.914$.

Thus testing the orientation parameters ω, φ, x_0 and y_0 may easily lead to incorrect decisions for CCD-cameras when testing their significance, whereas errors in these parameters are detectable. □

4.3.2 Sensitivity of the Estimates

In spite of testing for blunders, errors may remain undetected and influence the resulting estimate. The *sensitivity* of the result is often the only information one needs for evaluation. One may determine an upper limit for the influence of a group of observations onto the result.

The influence $\Delta_i f(\hat{\boldsymbol{\beta}})$ onto a function $f(\hat{\boldsymbol{\beta}})$ of the unknown parameters caused by leaving out a group $\boldsymbol{y}_i$ of observation is limited:

$$\Delta_i f(\hat{\boldsymbol{\beta}}) \leq \Delta_i f_{max}(\hat{\boldsymbol{\beta}}) \qquad (27)$$

with (cf. FÖRSTNER [1992])

$$\Delta_i f_{max}(\hat{\boldsymbol{\beta}}) = T_i \cdot \mu_i \cdot \sigma_{f(\beta)} \cdot \sqrt{n_i} \qquad (28)$$

where n_i is the size of the group, $\sigma_{f(\beta)}$ the standard deviation of the function $f(\beta)$ is derivable by error propagation measuring the precision of the result, T_i of the test statistics (22), measuring the quality of the observation group and the geometry factor

$$\mu_i = \lambda_{max}\{(\boldsymbol{\Sigma}_{xx}^{(i)} - \boldsymbol{\Sigma}_{xx})\boldsymbol{\Sigma}_{xx}^{-1}\} \qquad (29)$$

evaluating the mensuration design. The value μ_i explicitely measures the loss in precision, i. e. the normalized increase $\boldsymbol{\Sigma}_{\beta\beta}^{(i)} - \boldsymbol{\Sigma}_{\beta\beta}$ of variance of the result when leaving out the i–th group $\boldsymbol{l}_i$ of observations.

For a single observation it reduces to

$$\mu_i = \frac{1 - r_{ii}}{r_{ii}} \qquad (30)$$

with the diagonal elements r_{ii} of $\boldsymbol{R}$ (cf. (17)).

The value $\Delta_i f_{max}(\boldsymbol{\beta})$ (28) measures the *empirical sensitivity* of the estimate with respect to blunders e. g. matching errors in groups $\boldsymbol{l}_i$; empirical, as it depends on the actual observations via T_i.

If T_i is replaced by a constant δ_0, indicating the minimum detectable (normalized) error, we obtain the *theoretical sensitivity*

$$\Delta_{0i} f(\hat{\boldsymbol{\beta}}) \leq \Delta_{oi} f_{max}(\hat{\boldsymbol{\beta}}) \tag{31}$$

with

$$\Delta_{0i} f_{max}(\hat{\boldsymbol{\beta}}) = \delta_0 \cdot \mu_i \cdot \sigma_{f(\beta)} \cdot \sqrt{n_i}\,. \tag{32}$$

It may be used for planning purposes since it does not depend on actual observations and can therefore be determined in advance. δ_0 is usually chosen to be larger than the critical value k for T_i, e.g. $\delta_0 = 1.5k$ or $\delta_0 = 2k$ and can be linked to the required power of the test (cf. Baarda [1967]/[1968], Förstner [1987]).

Both sensitivity values contain the product of terms representing different causes. This e. g. allows to sacrifice precision, thus increase standard deviation σ_f, by paying more for leaving a larger redundancy, thus lowering the geometric factor μ_i for all observations or vice versa.

Observe that the analysis is based on values which have a very precise geometric meaning. This allows for an easy definition of thresholds, even if one is not aquainted with the theory below. As well, a clear comparison between different configurations is possible even for different types of tasks. Because the evaluation refers to the final parameters, it also may be used when fusing different type of observations. As model knowledge may be formalized in a Bayesian manner, the effect of prior information onto the result of an orientation may also be analysed.

Summarizing the evaluation of the design using the comparison of the covariance matrix of the parameters with a criterion matrix and using the different measures for the sensitivity has several distinct properties:

- it is a general concept
- it works for all types of critical surfaces and solves the problem of critical areas
- it works with all problems of estimation
- it may detect hidden singularities
- it also works in the complex situation where observations of different types are mixed (points, lines, circles, ...) or in the context of sensor fusion where also physical measurements (force, acceleration, ...) are involved
- it is related to a task, thus explicitely depends on user requirements. This enables to argue backwards and optimize the design.
- it provides measures which are easily interpretable.

4.4 Empirical Tests or "How good is the Procedure?"

The final step in developing a procedure is to test its performance using reference information. We therefore assume the procedure to provide estimates $\hat{\boldsymbol{\beta}}$ (15) for the parameters,

$$\hat{\sigma}_0^2 = \frac{\hat{\boldsymbol{v}}^T \boldsymbol{P} \hat{\boldsymbol{v}}}{n - u} \tag{33}$$

for the unknown variance factor σ_0^2 and the covariance matrix $\boldsymbol{\Sigma}_{\hat{\beta}\hat{\beta}}$ (5). We also assume that the result is normally distributed, which, even in the case where robust estimators have been used, is a reasonable assumption due to the central limit theorem.

We now assume to have reference or "true" values $\boldsymbol{\beta}_t$ for the parameters[2]. There may be two situations: either they are derived by some mensuration of (usually superior) precision or they are simulated. This yields:

1. $\boldsymbol{\beta}_t$: $\mathrm{E}(\boldsymbol{\beta}_t) = \tilde{\boldsymbol{\beta}}$, $\mathrm{D}(\boldsymbol{\beta}_t) = \boldsymbol{\Sigma}_{\beta_t\beta_t}$
2. $\boldsymbol{\beta}_t$: $\mathrm{E}(\boldsymbol{\beta}_t) = \tilde{\boldsymbol{\beta}}$, $\mathrm{D}(\boldsymbol{\beta}_t) = \mathbf{o}$.

the second case obviously being a special case of the first one.

The result of the estimation procedure yields either

$$\hat{\boldsymbol{\beta}},\ \boldsymbol{\Sigma}_{\hat{\beta}\hat{\beta}} = \sigma_0^2 \boldsymbol{Q}_{\hat{\beta}\hat{\beta}} \tag{34}$$

or, with the estimated variance factor $\hat{\sigma}_0^2$

$$\hat{\boldsymbol{\beta}},\ \hat{\boldsymbol{\Sigma}}_{\hat{\beta}\hat{\beta}} = \hat{\sigma}_0^2 \boldsymbol{Q}_{\hat{\beta}\hat{\beta}}. \tag{35}$$

The comparison of $\hat{\boldsymbol{\beta}}$ with $\boldsymbol{\beta}_t$ usually is based on the estimated differences

$$\hat{\Delta}\boldsymbol{\beta} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_t \tag{36}$$

by giving the r. m. s. difference

$$\hat{s}^{(0)} = \sqrt{\frac{1}{n}\sum_{i=1}^{u} \hat{\Delta}\beta_i^2} \tag{37}$$

This measure obviously only is meaningful in case the parameters β_i are of the same type, e. g. coordinates in object space during surface reconstruction. Otherwise no averaging can be performed.

By referring to the individual variances we would obtain the (partially) r. m. s. difference

$$\hat{s}^{(1)} = \sqrt{\frac{1}{n}\sum_{i=1}^{u} \left(\frac{\hat{\Delta}\beta_i}{\sigma_{\beta_i}}\right)^2} \tag{38}$$

[2] We assume here that values for *all* unknown parameters are available, otherwise, the expressions have to be modified slightly.

This dimensionless quantity should be close to 1, suggesting it to be a reasonable test statistic for evaluating the estimation. It, however, is an unsufficient statistic, as it neglects the mutual correlations between the estimates, moreover it does not take the uncertainty of the reference data into account.

Therefore we may use the statistic

$$\hat{s}^2 = \frac{1}{u}\hat{\Delta}\boldsymbol{\beta}^T \boldsymbol{\Sigma}^{-1}_{\hat{\Delta}\beta\hat{\Delta}\beta} \hat{\Delta}\boldsymbol{\beta} \qquad (39)$$

with the covariance matrix

$$\boldsymbol{\Sigma}_{\hat{\Delta}\beta\hat{\Delta}\beta} = \boldsymbol{\Sigma}_{\hat{\beta}\hat{\beta}} + \boldsymbol{\Sigma}_{\hat{\beta}_t\hat{\beta}_t} \qquad (40)$$

It can be shown that $\hat{s}^2$ is a sufficient statistic for σ_0^2 *if* the estimate $\hat{\boldsymbol{\beta}}$ is unbiased. Thus, under these conditions, we obtain the following test statistic

$$T_t = \frac{\hat{s}^2}{\hat{\sigma}_0^2} \sim F_{u,r} \qquad (41)$$

which is Fisher distributed, as the parameters $\hat{\boldsymbol{\beta}}$ and the residuals $\boldsymbol{v}$ are independent ($\boldsymbol{\Sigma}_{\hat{\beta}v} = \mathbf{o}$).

The tests statistic T_t obviously may be used to *decide whether the internal estimate* $\hat{\sigma}_0^2$ for σ_0^2 *can be used for prediction of the external accuracy*, captured by $\hat{s}^2$.

The bias $b_\beta = \mathrm{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_t)$ can be estimated empirically in case repeated experiments are available. Then techniques from multivariate statistics may be used for evaluation.

We have to distinguish two cases for applying these statistics.

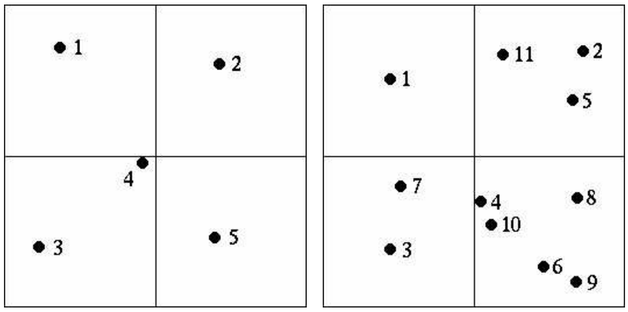
1. The requirements of the *user* usually refer to a certain type of parameter, e. g. object coordinates, which should show a homogeneous precision. This implicitely means, the covariance matrix should be diagonal, with equal entries, representing the required variance: $\boldsymbol{H} = \sigma_\beta^2 \boldsymbol{I}$, which could be checked with the tools from section 4.2.2. With this *assumption* the r. m. s. difference from (37) really gives a statistic for the average precision, in case the uncertainty of the reference data can be ignored. In most practical cases $\boldsymbol{\Sigma}_{\hat{\beta}\hat{\beta}}$ will not be diagonal, which leads to deviations from the ideal structure, resulting in errors in the standard deviations of functions of the estimates up to a factor 3–5 in both directions, thus to relative weight errors in the result of 81 to 625 (!). These worst cases, however, can be avoided in case the number of reference values is large enough ($\geq$ 100) and the variances $\sigma^2_{\hat{\beta}_i}$ do not differ too much.

   Therefore the report of the $\hat{\Delta}\beta_i$ is a reasonable first step towards proving the qualitity of a procedure.

2. The requirements of the *developer*, however, are guided by model building. The model is termed good, if it captures all information provided by the data, here $\boldsymbol{l}$. This can be checked by a hypothesis test using T_t, in case $T_t > F_{u,r,1-\alpha}$ the hypothesis that the differences between experiment and reference data are only caused by random perturbations has to be rejected. This test usually is much sharper than the one performed by a user, as also small effects may influence T_t, whithout necessarily bothering the application.

   No general hints can be given how to change the model, as *all* model errors may lead to a rejection of the hypothesis. Of course, model errors may stay hidden as discussed in the previous section.

Figure 4: shows two sets of image points used for image orientation by spatial resection. Sets of three points may lead to results of different stability as shown in the table for three sets of the left configuration (a).



5 Examples

The following examples are taken from two projects where vision algorithms are used to automate tasks which up to now had to be performed by human operators.

5.1 Analysing the Stability of the Solution

The first three examples are based on the program AMOR (c. f. SCHICKLER [1992]), which contains three different robust procedures, namely clustering, RANSAC and ML-type estimation for automatically determining the exterior orientation of a camera. The procedure is based on the matching of 2D-line segments extracted in the image with straight 3D-line segments describing the form of socalled *control points*, mainly consisting of roofs of buildings. Sets of straight line segments are grouped into "points", which the analysis partly refers to. The aerial images used here usually contain 5-10 such objects/points which are more or less well-distributed over the field of view.

Five image points situated as in Fig. 4 are to be used to estimate the 6 orientation parameters of the image based on given 3D-coordinates with a spatial resection (3D – 3D). Due to gross errors in the data a RANSAC procedure (cf. BOLLES/FISCHLER [1981]) is applied, randomly selecting 3 points and directly solving for the orientation parameters. The quality of this

	configuration	$\sqrt{\lambda_{max}}$
1	1/2/3	0.8
2	2/3/4	88.0
3	1/3/4	13.2

Table 2: shows the stability of sets of three points used for spatial resection (cf. Fig. 2).

selection has to be evaluated automatically in order to immediately exclude unstable configurations. The above mentioned technique for evaluating the stability of a configuration is applied.

The criterion matrix is derived from a least square fit with 4 points symmetrically sitting in the four corners of the image (cf. Appendix). The resulting covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\hat{\beta}\hat{\beta}}$ yields the criterion matrix

$$\boldsymbol{H} = 16 \cdot \boldsymbol{\Sigma} \tag{42}$$

thus requiring the standard deviations of the orientation parameters within the RANSAC-procedure to be better than 4 times the standard deviation of the ideal configuration. $\boldsymbol{\Sigma}$ is sparse allowing to easily, i. e. algebraically determine the matrix $\boldsymbol{H}^{-\frac{1}{2}}$ in (10) (cf. Appendix).

For several triplets of points the ratio $\sqrt{\lambda_{max}}$ is given.

The good triangle (1,2,3) obviously leads to sufficiently precise orientation parameters. The second triplet (2,3,4) consists of three nearly collinear points, which obviously is an undesirable configuration. The third triplet (1,3,4) causes the projection centre to *approximately* lie on a critical cylinder, causing the diagnostic value $\sqrt{\lambda_{max}}$ to be significantly larger than 1, expressing the fact that some function of the orientation parameters in that configuration has a standard deviation being appr. 13 times larger than required. The small triplet (2,5,11) in Fig. 3-1b also leads to a weak determination of the orientation parameters with a value $\sqrt{\lambda_{max}} \approx 4$.

5.2 Analysing the Sensitivity

This example shows the power of the sensitivity analysis for evaluating the success of an automatic procedure. The sensitivity analysis may be used to evaluate the quality of the orientation with respect to

a) matching errors of *individual line segments*; and,

b) matching errors of complete *sets* of line segments, representing one *object*.

The reason for this distinction is that both errors may occur; the first one being very common, the second one (whole sets of line segments) within the clustering procedure performed for each object individually.

a) Matching of 2D image line segments to 3D object line segments.

Table 3: shows the empirical and the theoretical sensitivity of the result of an orientation with straight edge segments.

Edge #	Empirical $\Delta_i f/\sigma_f$	Theoretical $\Delta_{0i} f/\sigma_f$
4	0.07	2.62
5	0.65	1.51
8	0.50	3.44
9	0.80	3.13
10	0.82	2.81
$\vdots$	$\vdots$	$\vdots$
21	0.68	4.42
$\vdots$	$\vdots$	$\vdots$

We have to deal with groups of 4 observations, namely the 4 coordinates representing the start and end point of the line segments. The 4×4 covariance matrix $\boldsymbol{\Sigma}_{l_i l_i}$ of this group also contains the correlations between the coordinates, which may be derived during the edge extraction process. We use a similar approach as Deriche/Faugeras [1990] and Förstner [1992] for representing the uncertainty of the line segments.

A typical result, as given in Table 3, can be summarized in two statements:

1. Empirical sensitivity: The maximum occurs at edge #10. The result may change up to 0.82 its standard deviation if line segment #10 would be left out, which is fully acceptable.
2. Theoretical sensitivity: The maximum occurs at edge #21. The result may change up to 4.42 times its standard deviation if a matching error remains undetected, which is at the border of being acceptable.

Thus, the result appears to be acceptable with respect to the redundancy in the estimations.

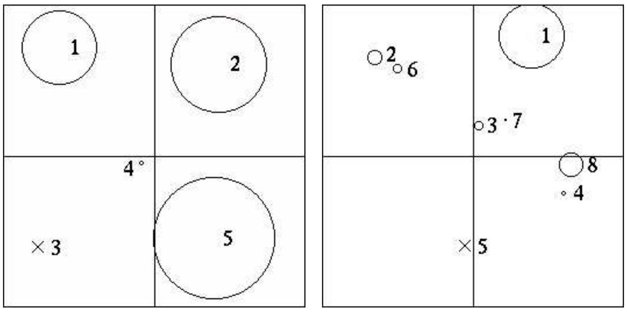
b) Match of a *set* of 2D image line segments to 3D object line segments.

Let us assume the m sets of segments to be matched, have $k_i, i = 1, \cdots, m$ line segments each, and we have to fear a matching error for a complete set. Then the sensitivity analysis has to be based on sets of $4 \times m_i$ coordinates for the m_i line segments.

Figures 5a and 5b show the position of the sets within two aerial images (c = 15 cm) to be oriented.

In Fig. 5a, one of the five sets, namely #3 was not matchable, leaving the spatial resection with 4 objects in the 3 other corners and in the middle of

Figure 5: shows two sets of image points used for image orientation by spatial resection. The readius of the circles indicate the theoretical sensitivity, i. e. the amount the result might change if the point (set of straight line segments) would be wrong without notice. In Fig. a (left) the point #3 has been detected to be wrong, thus only 4 points are left for spatial resection



the image. The circles around these "four points" have a radius proportional to $\delta_{0i} = \Delta_{0i} f_{max}/\sigma_f$ and indicate how sensitive the orientation is with respect to nondetectable errors within the clustering procedure. Because of the geometry factor μ (29) is dominant, the circles indicate how the precision deteriorates if one of the 4 sets is left out:

set 4 the three others 1, 2 and 5 form a well-shaped triangle, and thus guarantee good precision.

set 2 the three others 1, 4 and 5 nearly sit on a straight line leading to a highly unstable solution (near to singularity of first type).

set 1 the three others, 2, 4 and 5, form a well-shaped triangle. However, because the plane going through the sets is nearly parallel to the image plane, the projection centre closely has near to the *critical cylinder*. Leaving out

set 5 also leads to a nearly singular situation.

The situation in Fig. 5b shows a more irregular distribution with 8 sets. Since set 5 was not matched, set 1 is most influential in the orientation, but less than sets 1, 2 and 5 in the case of Fig. 5a.

5.3 Evaluation of the Performance of an Orientation Procedure

Table 4 summarizes the result of 48 image orientations.

The total number of correct and false decisions of the selfdiagnosis is split into the cases where the images contained 6 or more points, i. e. sets of straight line segments and cases with 5 or less points. An orientation was reported as correct if the empirical and the theoretical sensitivity factors $\Delta_i f/\sigma_f$ and $\Delta_{0i} f/\sigma_f$ (cf. eq. (28) and (31)) the standard deviations of the result were acceptable ($\lambda_{max} < 1$ cf. eq. (13)).

46 out of 48 orientations were correct and this was reported by the selfdiagnosis. In one case the orientation was incorrect, which was detected by the analysis. This appeared in an orientation with only 4 points, thus only one redundant point. Therefore altogether in 47 out of 48, i. e. in 98 % of all orientations the system made a correct decision. In about half of the cases (22 out of 46) the RANSAC procedure was able to identify errors which occured during the clustering and correct the result of the clustering, which was repeated with this a priori knowledge.

One orientation failed without being noticed by the system, which corresponds to 2 % false positives. This was an orientation with only 5 points.

The orientation of the 48 images was based on 362 clusterings of model and image line segments. 309, thus 85 %, were correct. As the errors in clustering are either completely wrong and therefore eliminated from the further processing or are wrong by a small amount, it is quite likely that 2 clusterings are incorrect by only a small amount, which may not be detectable by the RANSAC or the robust ML-type estimation, mimicing a good orientation.

The result achieved in this test is a clear reason to require at least 6 points, i. e. sets of straight edges, for a reliable orientation in this application. As can seen in the table, then all 39 orientations not only could correctly be handled by the automatic system, but actually lead to correct orientation parameters.

This example reveals the diagnostic tools to be extremely valuable for a final evaluation of an automatic procedure containing robust estimation procedures as parts.

5.4 Empirical Evaluation of a Surface Reconstruction Procedure

The last example is taken from the empirical investigation of the program MATCH-T reported by KRZYSTEK and WILD [1992]. The program determines the visible surface, represented as a finite element grid as $z = z(x, y)$, using a feature based matching technique (cf. FÖRSTNER [1992]), which uses an M-estimation for eliminating wrong correspondencies. The regularization, based on the curvature of the surface and a bilinear interpolation within the surface patches, leads to a highly overconstraint estimation problem, which is solved directly in a stripwise fashion.

Table 5 summarizes the main characteristics of three projects where empirical tests have been performed. From the many interest points (being in the range of a million) a small percentage has been selected by the robust estimation procedure. Via the interpolated surface they have been checked by independent photogrammetric and tacheometric (field) measurements. The partial inversion of the normal equation matrix yields internal, thus theoretical values $\hat{\sigma}_{\hat{\beta}_i}$ for the precision of the determined height i. e.

Table 4: shows the result of an extensive test of orienting aerial images

		report of selfdiagnosis	
		correct	wrong
reality	correct	46(39/7) (correct decision)	0(0/0) (false positives)
	wrong	1(0/1) (false positives)	1(0,1) (correct decisions)

z-values of the surface.

They are to be compared with the r. m. s. height differences betwen the automatically measured heights amd the heights at the reference positions. In case the results showed significant bias, the differences were bias corrected, which seemed adequate for that test. Thus the *form* of the surface actually is evaluated. The coincidence between theoretical and empirical accuracy is obvious, much more, as the r. m. s. values also contain the uncertainty of the reference measurements.

References

[1967] Baarda W.: Statistical Concepts in Geodesy. Netherlands geodetic Commissison, New Series, Vol.2, No. 4., Delft 1967

[1968] Baarda W.: A Testing Procedure for Use in Geodetic Networks. Netherlands Geodetic Commissison, New Series, Vol.5, No. 1., Delft 1968

[1973] Baarda W.: S-Transformations and Criterion Matrices. Netherlands Geod. Commission New Series, Vol. 5, No. 1, Delft, 1973

[1981] Bolles R. C., Fischler M. A.: Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Comm. of ACM, Vol. 24, No. 6, 1981, 381–395

[1990] Chen C.-H., Mulgaonkar P. G.: Robust Vision-Programs Based on Statistical Feature Measures. Int. Workshop on Robust Computer Vision, Proc., Seattle, October 1990, 39–56

[1982] Cook R. D., Weisberg S.: Residuals and Influence in Regression. Chapmann & Hall, 1982

[1990] Deriche R., Faugeras O.: Tracking Line Segments. Lecture Notes in Computer Science, 427, Springer, 1990, 259–268

[1983] Förstner W.: Reliability and Discernability of Extended Gauss-Markov Models. Deutsche Geod. Komm, A98, München, 1983, 79–103

[1987] Förstner W.: Reliability Analysis and Parameter Estimation in Linear Models with Application to Mensuration Problems in Computer Vision. CVGIP, 40, 1987, 87–104

[1992] Förstner W.: Uncertain Geometric Relationships and their Use for Object Location in Digital Images. Institut für Photogrammetrie, Universität Bonn, Tutorial, 1992

[1992] Förstner W.: A Feature based Corresponcence Algorithm for Image Matching, International Archives for Photogrammetry, Vol. 26-3/3, Rovaniemi, 1986

[1986] Gale W. A. (Ed.): Artificial Intelligence & Statistics. Addison Wesley, 1986

[1986] Hampel F. R., Ronchetty E. M., Rousseeuw P. J., Stahel W. A.: Robust Statistics: The Approach Based on Influence Functions. Wiley, NY, 1986

[1985] Haralick, R. M.: Computer Vision Theory: The Lack Thereof, Proc. of 3rd Workshop on Computer Vision, Bellaire, 1985, pp. 113–121

[1989] Horaud R., Conio B., Leboulleux O., Lacolle B.: An Analytic Solution for the Perspective 4-Point Problem. CVPR 1989, Proc., San Diego, CA, 1989, 500–507

[1981] Huber P. J.: Robust Statistics. Wiley, NY, 1981

[1991] Huber P. J.: Between Robustness and Diagnostics. In: Directions in Robust Statistics and Diagnostics, Part I, W. Stahel and S. Weisberg (eds.), Springer Verlag, 1991, 121–130

[1992] Krzystek P., Wild D.: Experimental Accuracy Analysis of Auomatically Measured Digital terein Models, in: Förstner/Ruwiedel: Robust Computer Vision – Quality of Vision Algorithms, Wichmann verlag, karlsruhe, 1992, pp. 372–390

[1990] Roth G., Levine M. D.: Random Sampling for Primitive Extraction. Int. Workshop on Robust Computer Vision, Proc., Seattle, October 1990, 352–366

[1986] Pregibon A DIY Guide to Statistical Strategy. in: Gale [1986], pp. 389-400

[1987] Rousseeuw P. J., Leroy A. M.: Robust Regression and Outlier Detection. Wiley, NY, 1987

Table 5: shows the result of three extensive tests of the program MATCH-T for surface reconstruction, from KRZYSTEK/WILD 1992

	project 1	project 2	project 3
photo scale	1 : 7 000	1 : 14 000	1 : 12 000
flying height	1 000 m	2 100 m	1 900 m
interest 2D points	0.7 10^6	0.15 10^6	0.9 10^6
measured 3D points	18 712	49 275	300 346
photogrammetric checkpoints	2 123	1 847	4 375
tacheometric checkpoints	135	135	–
theoretical $\hat{\sigma}_{\beta_i}$	0.05 m	0.11 m	0.10 m
r. m. s. $\hat{s}^{(0)}$ photogrammetric	0.06 m	0.11 m	0.11 m
r. m. s. $\hat{s}^{(0)}$ tacheometric	0.075	0.10	–

[1992] Schickler W.: Feature Matching for Outer Orientation of Single Images Using 3-D Wireframe Controlpoints. Int. Arch. of Photogrammetry & Remote Sensing, Vol. 29, ISPRS Comm. III, Washington D. C., 1992

[1989] Sester M., Förstner W.: Object Location Based on Uncertain Models Informatik Fachberichte 219, Springer, 1989, 457–464

[1987] Stockman G.: Object Recognition and Localization via Pose Clustering. CVGIP, 40, 1987, 361–387

A Algebraic expression for the normal equation of the spatial resection with four parts in symmetric position.

Let points $P_i(X_i, Y_i, Z_i), i = 1, \ldots, n$ be given and observed in the image. The linearized observation equations for the image coordinates (x', y') depending on the 6 orientation parameters, namely the rotation angles ω, φ, κ and the position (x_0, y_0, z_0) of the projection centre can be expressed as

$$dx' = -\frac{c}{H} dx_0 z_0 - \frac{x'}{H} dx_0 z_0 + -\frac{x'y'}{c} d\omega \quad (43)$$

$$+c\left(1 + \frac{x'^2}{c^2}\right) d\phi + y' d\kappa \quad (44)$$

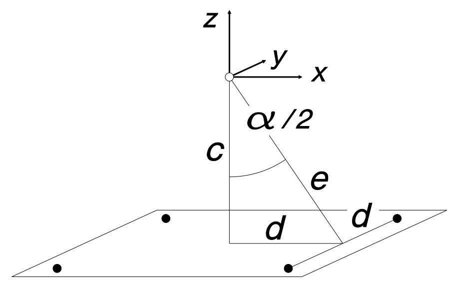
$$dy' = -\frac{c}{H} dy_0 z_0 - \frac{y'}{H} dy_0 z_0 - c\left(1 + \frac{y'^2}{c^2}\right) d\omega \quad (45)$$

$$+\frac{x'y'}{c} d\omega - x' d\kappa \quad (46)$$

valid for each image point $P'(x', y')$. c is the camera constant and $H = z$ is the distance of the object point $p(x, y, z)$.

In case $n = 4$ image points lie in symmetric position $(\pm d, \pm d)$ in the image (cf. Fig. 6) and the z-coordinates of the points $P_i(x_i, y_i, z_i)$ in the coordinate system of the camera are equal to $H = z_i$, we can collect the coefficients of the 8×6 matrix as in the table.

Figure 6: shows the normalized situation for the spatial resection used as a reference for precision.



The algebraic expression for the normal equation system assuming weights 1 is given by

$$N = \begin{pmatrix} 4\frac{c^2}{H^2} & 0 & 0 & 0 & -4\frac{e^2}{H} & 0 \\ 0 & 4\frac{c^2}{H^2} & 0 & 4\frac{e^2}{H} & 0 & 0 \\ 0 & 0 & 6\frac{d^2}{H^2} & 0 & 0 & 0 \\ 0 & 4\frac{e^2}{H} & 0 & 4\frac{e^4}{c^2}(1+\frac{d^4}{e^4}) & 0 & 0 \\ -4\frac{e^2}{H} & 0 & 0 & 0 & 4\frac{e^4}{c^2}(1+\frac{d^4}{e^4}) & 0 \\ 0 & 0 & 0 & 0 & 0 & 8d^2 \end{pmatrix} \quad (48)$$

Discussion

1. The normal equation matrix is sparse. It collapses to two diagonal elements and two 2×2 matrices. This allows algebraic inversion, which may be used for a direct solution of the orientation in real time applications.

2. The correlation between x_0 and $d\phi$ (y-rotation)

i		dx_0	dy_0	dz_0	$d\omega$	$d\phi$	$d\kappa$
1	$x' = +d$	$-\frac{c}{H}$	0	$-\frac{d}{H}$	$-\frac{d^2}{c}$	$\frac{e^2}{c}$	$+d$
	$y' = +d$	0	$-\frac{c}{H}$	$-\frac{d}{H}$	$-\frac{e^2}{c}$	$\frac{d^2}{c}$	$-d$
2	$x' = -d$	$-\frac{c}{H}$	0	$+\frac{d}{H}$	$+\frac{d^2}{c}$	$\frac{e^2}{c}$	$-d$
	$y' = +d$	0	$-\frac{c}{H}$	$-\frac{d}{H}$	$-\frac{e^2}{c}$	$\frac{d^2}{c}$	$-d$
3	$x' = -d$	$-\frac{c}{H}$	0	$+\frac{d}{H}$	$-\frac{d^2}{c}$	$\frac{e^2}{c}$	$-d$
	$y' = -d$	0	$-\frac{c}{H}$	$+\frac{d}{H}$	$-\frac{e^2}{c}$	$\frac{d^2}{c}$	$+d$
4	$x' = +d$	$-\frac{c}{H}$	0	$-\frac{d}{H}$	$+\frac{d^2}{c}$	$\frac{e^2}{c}$	$+d$
	$y' = -d$	0	$-\frac{c}{H}$	$+\frac{d}{H}$	$-\frac{e^2}{c}$	$-\frac{d^2}{c}$	$+d$

(47)

and y_0 and $d\omega$ (x-rotation) are correlated by

$$\rho_{y_0\omega} = -\rho_{x_0\phi} = \frac{N_{24}}{\sqrt{N_{22} \cdot N_{44}}} \tag{49}$$

$$= \frac{4\dfrac{e^2}{H}}{\sqrt{4\dfrac{c^2}{H^2} \cdot 4\dfrac{e^4}{c^2}(1 + \dfrac{d^4}{e^4})}} \tag{50}$$

$$= \frac{1}{\sqrt{1 + \sin^4 \dfrac{\alpha}{2}}} \tag{51}$$

as $d/e = \sin\alpha/2$ (cf. Fig. 6).

3. Taking the square root $\boldsymbol{N}^{1/2}$ of $\boldsymbol{N}$ is trivial for the diagonal elements for dz_0 and $d\kappa$ and requires to take the square root $\boldsymbol{T}^{1/2}$ of two 2×2 matrices $\boldsymbol{T}$ which easily can be determined using the eigenvalue decomposition $\boldsymbol{T} = \boldsymbol{D}\boldsymbol{\Lambda}\boldsymbol{D}^T$ yielding

$$\boldsymbol{T}^p = \boldsymbol{D}\boldsymbol{\Lambda}^p\boldsymbol{D}^T \tag{52}$$

for $p = 1/2$ or, as needed in (10) for $p = -1/2$ ($\boldsymbol{D}^T = \boldsymbol{D}^{-1}$ and $\boldsymbol{\Lambda} = \mathrm{Diag}(\lambda_1, \lambda_2)$).