In: Vorträge der 42. Photogrammetrischen Woche an der Universität Stuttgart 1989, Schriftenreihe
Institut für Photogrammetrie der Universität Stuttgart, Heft 13, 1989, pp. 205-221

Image Analysis Techniques for Digital Photogrammetry

W. Förstner, Stuttgart University, Institute for Photogrammetry

1 Introduction

Digital Photogrammetry is concerned with photogrammetric techniques and applications based on digital, or digitized images. Though it evolved from *analog* over *analytical* photogrammetry a new quality characterizes the upcoming techniques: whereas the old labels refer to the mechanical or optical and to the computer based realization of the *geometric* relations resp. between object and image space and the interpretation of the image content always was left to the operator, digital photogrammetry inherently covers both the geometric as well as the *semantic* aspects and thus - at least conceptually - aims at full automation of all photogrammetric tasks. Image interpretation, the stepchild of photogrammetric research, will increasingly dominate the geometric analysis and its statistical evaluation. Though a great body of experience concerning semiautomatic map production may help defining goals, showing pitfalls and raising funds and though - especially in the area of classification of multispectral images - quite some techniques already are available, the formalization and theoretical foundation of image interpretation is by far not advanced enough to be of practical help in developing algorithms which could solve specific tasks in Digital Photogrammetry. The admissably great developments in automatic generation of digital terrain models, an important and - besides multispectral calssification - the only applicable technique in digital photogrammetry, only supports this evaluation as also here the necessary steps towards integration of image interpretation into the terrain evaluation have not really been approached up to now.

There seems to be a gap between those techniques concentrating on photometry aiming at image interpretation and those governed by geometry aiming at photo-grammetry, but nearly all algorithms developed for analysing digital images, perhaps except image correlation techniques, contain an interpretation step as first step: detecting targets (fiducials, reseau crosses), extracting distinct points, extracting edges or lines, partitioning images into homogeneous regions, supervised classifications etc. In all these cases an implicit assumption about the usefulness i. e. the semantic content of the extracted image features is made. Even a possible second step of aggregation, e. g. of the edge elements into straight edge segments, in a first instance is an interpretation task. Only the localization steps, in the image or in three dimensional space, deal solely with geometry where classical estimation and evaluation techniques can be applied. But their result is governed by the decisions made before and therefore limited. New concepts which integrate interpretation and geometric analysis and which are as powerfull as those available in parameter estimation therefore need to be developed.

The situation appears to be different in the area of *Computer Vision* where the development of image analysis techniques in the central issue. Here image processing techniques, specifically for restauration and coding, investigations into the human visual system, especially into low level processes and research towards industrial applications of image analysis techniques evolved in parallel with much mutual interference and exchange. Technically image analysis covers many aspects such as feature extraction, motion analysis, shape from shading, stereo, texture, shadow, contours etc. Also quite some theoretical work has been accomplished in specific areas such as edge detection (e. g. Haralick 1984, Canny 1986, Yuille/Poggio 1986), mathematical morphology (e. g. Serra 1982, Haralick et al. 1987, Haralick 1988), texture analysis (e. g. Kashyap 1985, Malik/Perona 1989, Rao/Schunck 1989) or surface reconstruction (e. g. Grimson 1981, Terzopoulos 1986, Blake/Zissermann 1987). Interestingly enough comparably little research has been published in the area of interpreting aerial images (cf. e. g. McKeown et al. 1985, 1988, Brooks 1986, Herman/Kanade 1986, Huertas/Nevatia 1988, McKeown/Denlinger 1988 Hanson/Quam 1988, Mulder et. al. 1988). The reason simply is the complexity of the task, which is orders of magnitude larger than in industrial applications. In spite of intensive development also in this area which can by no means reviewed here, no commonly accepted theory for image interpretaion is available. Knowledge based systems for image analysis due to their broarder goals even require a deeper understanding of image analysis techniques.

There however is a promising approach for image interpretation, which may also play a role in digital photogrammetry. It is based on concepts from information theory and contains a unifying measure for evaluating the result of image interpretation, namely the length of an optimally coded discription of the image, measured in bits (cf.

Georgeff/Wallace 1984; Rissanen 1983, 1987; Leclerc 1988, Fua/Hanson 1987, 1988).

According to Fua and Hanson interpretation consists in a two step procedure:

1. Derive a set of likely hypothesis' of image descriptions using search and/or estimation techniques. Here all available knowledge on the type of objects and on efficient strategies may be explored.
2. Choose the best of the competing hypothesis based on the simplicity of the description. The simplicity or likelihood is measured by the number of bits necessary to describe the specific realization of the model and the deviation of the actual image from the ideal model.

Fua and Hanson stress the importance of *generic models*: These are models with a *structure* of a certain type, specified by rules, and additional *numeric parameters*. An example are cultural objects which appear as homogeneous areas with a rectilinear boundary, thus require the specification of a sequence of polygon sides (involving rectangles) and a simple gray value as description for the interior. These generic models have to be seen in contrast to *specific models* which only require a *fixed set* of numerical parameters to be specified. An additional feature of the generic models is that they contain *geometric and photometric* specifications.

The problem of evaluating competing image descriptions is the incompatibility of photometric data - the original observations in digital image analysis - and the complexity of the model. This paper was motivated by the approach of Fua and Hanson, as it is able to unify photometric and geometric features as well as low and high level structures of models and data within one framework to a large extent. Though its primary concern is the evaluation of hypothesis' it partially provides means for finding good hypothesis, which can be related to robust estimation techniques. Moreover, maximum likelihood and least squares techniques are special cases of the underlying principle derived from information theory. We therefore want to show that also other tasks, such as image matching, image restauration and feature extraction can be derived within the same framework.

The paper on one hand wants to provide the necessary tools from information theory and on the other hand aims at demonstrating the usefulness of the concept for various image analysis tasks relevant for Digital Photogrammetry. The principle of minimum description length encoding using an estimation problem is demonstrated in section 2. The necessary theory is outlined in section 3, being the basis for the collection of image analysis tasks in section 4.

2 Interpreting a Set of Points in a Plane

We want to introduce the principle of minimum description length encording using a simple example similar to the one given by Georgeff and Wallace (1984). Let n_0 points (x_i, y_i) in a plane be given as in Fig. 2-1a. The scope is to explain the data in the most intuitive manner. The figure suggests the larger number $n = 9$ of the $n_0 = 14$ points to approximately sit on a straight line, while the other $\bar{n} = n_0 - n = 5$ points do not belong to this line. Fig. 2-1b shows a different pattern, where we are not sure whether we should assume the 5 points in the middle of the figure to belong to a straight line or whether we rather should treat the figure as consisting of 14 randomly distributed points or even 3 vertical nearly straight lines.

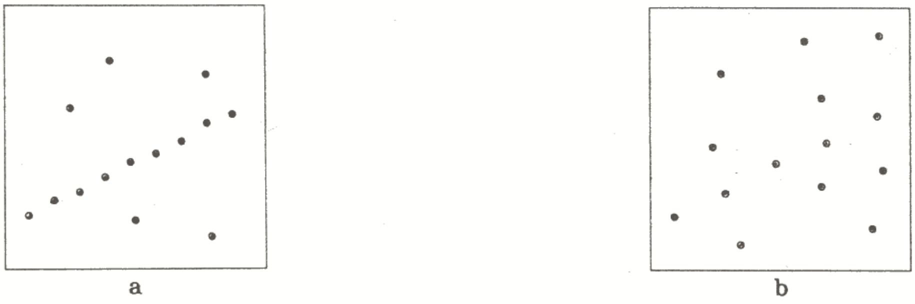


Fig. 2.1 14 points within a square, most likely interpretations:
a) 9 points on a straight line and 5 outliers
b) random set or 5 points on a straight line and 9 outliers?

The situation is representative for a large class of interpretation tasks:

- We have to deal with several competing hypothesis which have a different structure.
- We have to deal with a significant amount of spurious data.
- There may be no explanation of the data within the assumed set of hypothesis.

The problem of explaining the data sets in Fig. 2-1 lies in the fact that the pure fit between a selected number of data points and a set of hypothesized straight lines, say, is not sufficient as a quality measure, as this fit can be made perfect by restricting to just 2 data points or by increasing the number of postulated straight lines. Therefore the evaluation of an explanation has to balance the fit between data and model and the complexity of the model. The principle of description length encoding fullfills these requirements.

We want to derive the description lengths in bits for the case when no model is assumed with the case when the data essentially are assumed to consist of points sitting approximately on a straight line admitting some outliers.

Let the coordinats be given up to a resolution of ε (e. g. 1 pixel) and be within a range R (e. g. 256 pixel). Then $lb(R/\varepsilon)$ [1] bits are necessary to describe one coordinate. The description length for the n_0 points, when assuming no model, therefore is

$$\Phi_0 = \sharp\text{bits (points} \mid \text{no model)} = n_0 \cdot 2lb(R/\varepsilon) \tag{2-1}$$

thus $2 \cdot n_0 \cdot 8 = 16 n_0$ in the case of n_0 points in a $256 \cdot 256$ pixel image or 224 bits on the plot of Fig. 2-1.

If we now assume n points to sit on a straight line and the other $\bar{n} = n_0 - n$ points to be outliers we need

$$\begin{aligned} \Phi_m &= \sharp\text{bits (points} \mid \text{1 straight line)} && \text{(2-2)} \\ &= n_0 + \bar{n} \cdot 2lb(R/\varepsilon) + \left[nlb(R/\varepsilon) + \sum_{i=1}^{n} \left\{ \frac{1}{2ln2} \cdot \left(\frac{v_i}{\sigma}\right)^2 + lb(\sigma/\varepsilon) + \frac{1}{2} lb2\pi \right\} \right] + 2lb(R/\varepsilon) && \text{(2-3)} \end{aligned}$$

where the first term represents the n_0 bits for specifying whether a point is good or bad, the second term is the number of bits to describe the bad points (cf. (2-1)), the third term is the number of bits to describe the good points and the last term is needed to describe the 2 parameters of the straight line. We assumed the good points to randomly sit on the straight line, which leads to the first term in the brackets, and to have gaussian distributed derivations v_i from the line with standard derivation σ. We show in section 3 that $\frac{1}{2ln2} \cdot \left(\frac{x-\mu}{\sigma}\right)^2 + lb\frac{\sigma}{\varepsilon} + \frac{1}{2}lb2\pi$ bits are necessary to describe a Gaussian variable $\underline{x} \sim N(\mu, \sigma^2)$, when μ and σ^2 are given and if it is rounded to multiples of ε.

In the example of Fig. 2-1a, with $n = 9$ and $\hat{n} = 5$ we *on an average* need:

$$\begin{aligned} \bar{\Phi}_m &= n_0 + \bar{n} \cdot 2lb(R/\varepsilon) + n\left(lb(R/\varepsilon) + lb(\sigma/\varepsilon) + \frac{1}{2} lb2\pi e \right) + 2lb(R/\varepsilon) && \text{(2-4)} \\ &= 14 + 2 \cdot 5 \cdot 8 + 9 \cdot (8 + 1 + 2.04) + 2 \cdot 8 \approx 209 \text{ bits} && \text{(2-5)} \end{aligned}$$

to code the point set, when assuming a straight line with outliers. This is less than the 224 bits, thus supporting this explanation. For Fig. 2-1b we however need 229 bits, assuming 5 points sitting on a straight line, which obviously is no explanation for the data.

In this application there exists a close relation to techniques of robust estimation (cf. Huber 1981): Minimizing Φ_m from eq. (2.3) with respect to the parameters of the straight line is identical to minimizing

$$\Phi_m = d + a \sum_{i=1}^{n} \rho(v_i) \tag{2-6}$$

with $a = 1/(2ln2)$, $d = n_0(1 + lb(R/\epsilon) + lb(\sigma/\varepsilon) + \frac{1}{2}lb2\pi) + 2lb(R/\varepsilon)$ and the optimization function

$$\rho(x) = \left\{ \begin{array}{rl} k^2 & \text{if} \quad (x/\sigma)^2 \geq k^2 \\ (x/\sigma)^2 & \text{if} \quad (x/\sigma)^2 < k^2 \end{array} \right. \tag{2-7}$$

[1] lb = logarithm with basis 2

and $k^2 = 2ln(R/(\sqrt{2\pi} \cdot \sigma))$, thus equivalent to minimizing

$$\Phi'_m = \sum_{i=1}^{n} \rho(v_i) \tag{2-8}$$

The function $\rho(x)$ is shown in Fig 2-2. With its flat shoulders, specifically $\rho'(x) = 0$ for $|x| > k$, it reveils its robust properties, in contrast to $\rho(x) = x^2$ with $\rho'(x) = 2x$ being unlimited, as large outliers have no influence onto the estimates.

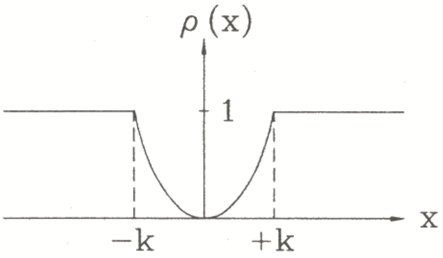


Fig. 2.2 Minimizing function $\rho(x)$ for minimal encoding length

When replacing $\rho(x)$ by $1 - exp(-(x/\sigma)^2)$, thus when blending the shoulders, minimizing Φ' in eq. (2-8) is equivalent to reweighting the residuals with an exponential weightfunction. This has already proposed by Krarup in 1967 (cf. Krarup et al. 1980). The optimization problem formulated there, however, had no link between the number of outliers and the degree of fit, as in eqs. (2-6) and (2-7).

The critical value k essentially depends on R/σ thus on the ratio of the expected range of the outliers to the precision σ of the good data points. In the above mentioned example ($R = 256, \sigma = 2$) we obtain $k = 3.4$, which is close to critical values tradionally chosen on the basis of the significance level of a hypothesis test.

The balance between model complexity and data fit can be used to derive the *minimum number of good data points* which *are necessary to expect an explanation.* In our case of one straight line we obtain from $\bar{\Phi}_m(n) < \Phi_0$

$$n \geq \frac{n_0 + 2lb\frac{R}{\varepsilon}}{lb\frac{R}{\sigma\sqrt{2\pi e}}} \tag{2-9}$$

In our example we obtain $n > 6$, again proving that the 5 points in Fig. 2-1b, which may seem to lie on a straight line, are not sufficient to motivate this explanation. For increasing precision, i. e. for decreasing σ (leaving R, ε and n_0 fixed) we obviously may accept an explanation with less data point supporting it.

The example reveiled several important properties of the description length encoding principle:

- It is able to compare explanations of different structure, here random data with one line plus outliers.
- It is able to cope with spurious data. Any additional explanation of these spurious data using a simple model would further decrease the description length.
- The decision whether data are spurious or not depends on the model not on some signifiance level.
- A decision on the *admissability* of a model or of an explanation is available, rejecting explanations which are too complicated - an extremely usefull and necessary property of the theory.
- The principle of minimum description length encoding for fixed model structure reduces to the principle of maximum likelihood and under the Gaussian assumption to the least squares principle.

The principle of description length encoding obviously goes along with intuition. As it is based on concepts from information theory we want to collect the main result of this theory in the next section.

3 Elements from Information Theory

The theory of information was developed by C. Shannon (Shannon/Weaver 1949) for analysing communication systems. Specifically it deals with measuring the information content of a message and the efficiency of sending the message over a channel which possibly is noisy. The theory is of a statistical nature as it only is concerned with the statistical properties of the message not with its meaning. We only want to lay out the basic notions here, as far as they are necessary in our context (cf. Shannon/Weaver 1949, Berger 1971, Hölzler and Holzwarth 1976).

According to Shannon a discrete information source can be modelled as a Markov-Process, which randomly selects letters out of a prespecified alphabet. The information, which is transmitted per letter, is the larger the less likely the letter is selected and can be interpreted as the degree of surprise when the letter reaches the receiver or as the uncertainity when no knowledge about the letter is available.

In the most simple case the transmitted letters are independent. Let $P(\underline{a} = w_i)$ be the probability that the letter $\underline{a}$ (a random variable) is equal to the value w_i. Then the gain of information when being told w_i, i. e. the *information* of w_i is

$$I(\underline{a} = w_i) = I(w_i) = \log\frac{1}{P(w_i)} = -\log P(w_i) \tag{3-1}$$

If the logarithm is taken to basis 2, the unit of information is the "bit"; if the natural logarithm is taken, the unit of information is the "nat".

In a similar manner one can measure the information which is obtained when being told w_i, but already knows the value of another letter $\underline{b} = w_j$. With the conditional probability $P(w_i \mid w_j)$ we obtain the *conditional information* (cf. Fig. 3.1)

$$I(w_i|w_j) = -\log P(w_i|w_j) = -\log\frac{P(w_i, w_j)}{P(w_j)} \tag{3-2}$$

$$= I(w_i, w_j) - I(w_j) \tag{3-3}$$

In case the events $\underline{a} = w_i$ and $\underline{b} = w_j$ are independent, $P(w_i|w_j) = P(w_i)$, the information we obtain is identical to that without preknowledge. If however, the events are dependent the information obtained when being told w_i is smaller than without preknowledge.

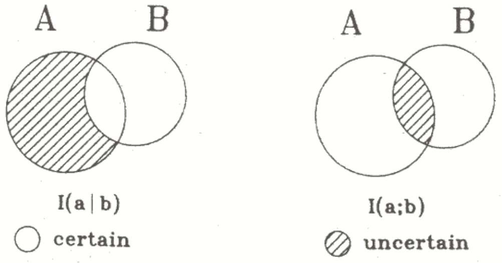


Fig. 3.1 Conditional and mutual information

The difference is the *mutual information* of w_i and w_j

$$I(w_i; w_j) = I(w_i) - I(w_i|w_j) \tag{3-4}$$

$$= -\log\frac{P(w_i, w_j)}{P(w_i) \cdot P(w_j)} \tag{3-5}$$

$$= I(w_i) + I(w_j) - I(w_i, w_j) \tag{3-6}$$

which obviously is symmetric with respect to w_i and w_j (cf. Berger 1971).

The *average information* of a source is called its *entropy* and defined by the expected value of the information:

$$H(\underline{a}) = E(I(\underline{a})) \tag{3-7}$$

$$= -\sum_i P(w_i)\cdot \log P(w_i) \tag{3-8}$$

Analogeously we may obtain the average conditional information or *conditional entropy*

$$H(\underline{a}|\underline{b}) = H(\underline{a},\underline{b}) - H(\underline{b}) \tag{3-9}$$

and the *mutual entropy*

$$H(\underline{a};\underline{b}) = H(\underline{a}) + H(\underline{b}) - H(\underline{a},\underline{b}) = H(\underline{a}) - H(\underline{a}|\underline{b}) \tag{3-10}$$

Now an important theorem of Shannon (1949, theorem 9) states: When coding an information source the number of bits per letter on an average is not less than the entropy of the source. This gives us the possibility to interprete the information of an event as the length of an optimal code describing the event. We have used this interpretation in the example in section 2. It intuitively corresponds to the binary codes used in digital computers to represent numbers. This also is the motivation to use binary logarithms to measure information in units of "bits". We therefore always can interpret the negative binary logarithm $-lbP(\underline{a})$ of the probability $P(\underline{a})$ of an event $\underline{a}$ as the number of bits to describe the event. When applying this concept we need not actually perform the coding, which may be a complicated task. We however only need the number of bits of the optimal code to evaluate, i. e. to compare different hypothesis.

The notion of description length can not directly be applied to real valued random-variables as infinitely many bits would be necessary to code them. The corresponding notion is therefore the *differential information* of a random variable

$$I(\underline{x} = x_i) = I(x_i) = -\log p(x_i), \tag{3-11}$$

where $p(x_i)$ is the value of the density function at x_i. Though $I(x_i)$ may in principle become negative and though it may change when x_i is measured in a different unit, the concept still can be used, as all practical applications refer to differences of information.

This especially holds for rounded values which are used in all application. To see this, we need the differential information and entropy for random variables with equal distribution and Gaussian distribution:

$$\begin{aligned}
\underline{x} \sim Eq[a,b] : I_E(x \mid a,b) &= lb(b-a) \qquad [\text{bit}] && (3\text{-}12)\\
H_E(x \mid a,b) &= lb(b-a) \qquad [\text{bit}] && (3\text{-}13)\\
\underline{x} \sim N(\mu,\sigma^2) : I_N(x \mid \mu,\sigma^2) &= \frac{1}{2ln2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2 + \frac{1}{2}lb2\pi\sigma^2 \qquad [\text{bit}] && (3\text{-}14)\\
H_N(x \mid \mu,\sigma^2) &= \frac{1}{2}lb2\pi e\sigma^2 \qquad [\text{bit}] && (3\text{-}15)
\end{aligned}$$

If we round a gaussian random variable $\underline{x}$ to a resolution of ε, e. g. 10^{-5}, yielding $\underline{x}_r$ then, using eq. (3-3),

$$I_r(x \mid \mu,\sigma^2,\varepsilon) = I_N(x \mid \mu,\sigma^2) - I_E\left(-\frac{\varepsilon}{2},\frac{\varepsilon}{2}\right) = \frac{1}{2ln2}\left(\frac{x-\mu}{\sigma}\right)^2 + \frac{1}{2}lb2\pi\left(\frac{\sigma}{\varepsilon}\right)^2 \qquad [\text{bit}] \tag{3-16}$$

and

$$H_r(x|\mu,\sigma^2,\varepsilon) = \frac{1}{2}lb2\pi e\left(\frac{\sigma}{\varepsilon}\right)^2 \qquad [\text{bit}] \tag{3-17}$$

In case $I_r(\underline{x}_r)$ or $H_r(\underline{x}_r)$ are not positive *no* coding of x_r is necessary, which is valid for $\sigma \leq \varepsilon/\sqrt{2\pi e} = 0.24\varepsilon$. Observe that I_r only represents the bits necessary to code the difference $x-\mu$ to the mean, as we have assumed the mean, the precision and the resolution to be known. As could be expected, minimizing the number of bits when presenting a random number corresponds to only state the necessary digits with respect to its precision.

We now want to investigate the interpretation of observed values $x = (x_1,\ldots,x_n)$ in terms of some explanatory variables y based on a model $E(\underline{x}) = f(y)$. If the covariance C_{xx} of the Gaussian $\underline{x}$ and the y are given, with $v = f(y) - x$ we need

$$I_r(\underline{x} \mid y) = \frac{1}{2ln2}v^T C_{xx}^{-1} v + \frac{1}{2}lb2\pi|C_{xx}| - nlb\varepsilon \qquad [\text{bit}] \tag{3-18}$$

to code the observed values. Minimizing $I(\underline{x}|y)$ with respect to y thus leads to the least squares estimates for y. If we now want to compare different models with respect to their description length we also have to code y. Thus we have to add a term representing the complexity of the model. In the most simple case of no additional information one can show that minimizing $I(\underline{x})$ with respect to all $y \in Y$ out of a prespecified set of models can be obtained from (cf. Rissanen 1987)

$$\begin{aligned}\min_{y\in Y} I(x) &= \min_{y\in Y}[I(x|y)+I(y)] && \text{(3-19)}\\ &\approx \min_{y\in Y}\left[I(x|y)+\frac{k}{2}\log n\right] && \text{(3-20)}\end{aligned}$$

where k is the number of parameters y and n is the number of observed values. The additional term can be motivated by observing that the relative precision $\hat{y}/\sigma_{\hat{y}}$ increases with $\sqrt{n}$ thus the description of $\hat{y}$ increases with $\log\sqrt{n} = \frac{1}{2}\log n$ and k parameters $\hat{y}_i$ are involved. The derivation and the degree of approximation can be found in Rissanen (1983).

We finally want to refer to the mutual information of two Gaussian random variables $\underline{x}$ and $\underline{y}$ (cf. Förstner 1988)

$$H(\underline{x};\underline{y}) = \frac{1}{2}\log\frac{1}{1-\rho_{xy}^2} \qquad \text{(3-21)}$$

which essentially depends on the correlation coefficient, $H(\underline{x};\underline{y})$ being zero if $\rho_{xy} = 0$. If one of both, say y is a vector, the same relation for $H(\underline{x};\underline{y})$ can be used, but now with the total correlation (cf. Giri 1977)

$$\rho_{xy} = \frac{c_{xy}\cdot C_{yy}^{-1}\cdot c_{yx}}{\sigma_x^2} \qquad \text{(3-22)}$$

between $\underline{x}$ and the vector $\underline{y}$, which obviously is a weighted correlation over all correlation ρ_{xy_i} between $\underline{x}$ and $\underline{y}_i$.

We now are prepared to discuss examples of image analysis techniques based on these concepts.

4 Information Extraction by Image Analysis

Information extraction by image analysis, following the concepts discussed so far, can be viewed as finding an optimal description of the photometric data with respect to the models made available. The analysis techniques described in the following make increasing use of the information theoretic concepts, though not all originally were developed with this interpretation in mind.

4.1 Image Matching

Image matching in the simplest form starts from two descriptions, for a right and a left photo, say, and aims at finding a maximal set of correspondencies between elements of the two descriptions or at finding an optimal mapping between the two descriptions. If D_r and D_l are the two descriptions for the right and the left image resp. then we may search for the transformation T of D_l so that

$$\min_T I(D_r \mid T(D_l;p)) \longrightarrow \hat{T} \qquad \text{(4-1)}$$

i. e. after knowing the left image and the transformation the surprise when being told the description D_r is minimum, in the noiseless case being identical to $T(D_l,p)$. A different but mathematically equivalent view of the matching problem aims at maximizing the mutual information between D_r and D_l after applying the transformation

$$\max_T I(D_r;T(D_l,p)) \longrightarrow \hat{T} \qquad \text{(4-2)}$$

Because of $I(\underline{a};\underline{b}) = I(\underline{a}) - I(\underline{a}|\underline{b})$ the two optimization problems are formally identical.

Now classical *correlation* of two signals $x = (x_1,\ldots,x_n)^T$ and $y = (y_1,\ldots,y_m)^T$ searches for the shift where the empirical crosscorrelation coefficient ρ_{xy} after the shift is maximum, which referring to eq. 3-21 maximizes the mutual information, thus corresponds to eq. 4-2. On the other hand all *least squares based techniques* referring to eq. 3-18 minimize the conditional information of the right image, being told the left one and the transformation (parameters).

The advantage of this information theoretic view of image matching is that it includes matching techniques which are based on relational descriptions, which was first proposed and applied by Boyer and Kak (1986, 1988). We use it for model based object location (Vosselman 1989). Relational descriptions consist of image features, such as points, lines and areas, called primitives and of relations between these primitives, e. g. "contains", "is parallel to", "are collinear". Both, primitives and relations, may have attributes, such as "type", "length", "area" or "contrast". Boyer and Kak minimize the distance

$$\min_h D_h = \min_h [D_h(\text{primitives}) + D_h(\text{relations})] \longrightarrow \hat{h} \qquad (4\text{-}3)$$

with respect to all mappings h from the primitives of the left to the right image which do not collide with the corresponding relations. The interprimitive distance D_h is related to the conditional information $I(b_k|a_k)$ of the attributes $\underline{a}_k$ and $\underline{b}_k$ of the left and the right image description which are added over all attributes of corresponding primitives. A similar approach is used for the relations. Thus they refer to the approach eq. 4-1 minimizing the conditional information. They also use information theoretic arguments for the ordering of the primitives according to their uniqueness as discussed in the next subsection.

4.2 Image Feature Extraction

Feature extraction is a basic step in image analysis both for image matching as for image interpretation. Two requirements are essential: local *distinctness* of the features is necessary for geometric precision, whereas global *uniqueness* is useful for decreasing the complexity of search processes during matching or interpretation. Now, distinctness and uniques of features are high if the mutual information with other features is low.

We first want to show that the extraction of distinct points, described by their surrounding intensity function, can be based on this notion of distinctness. Local distinctness of a point can be measured by the average distinctness to all points within a small neighbourhood N_P of P

$$d(P) = \text{average}_{Q \in N_P}[h\{H(P;Q)\}] \qquad (4\text{-}4)$$

where P and Q represent the intensity function around the points in concern and some monotonically increasing function h. As the mutual information $H(P;Q)$ decreases with the correlation coefficient ρ_{PQ}, which itself decreases with an increase of the curvature c_{PQ} of the autocorrelation function, which can be derived from the ratio $(\sigma^2_{f'}/\sigma^2_f)_{PQ}$ of the variance of the gradient to the variance of the intensity function f, the measure $d(P)$ is monotonically depending on the weight

$$w = \sum_{r,c} (f_r^2(r,c) + f_c^2(r,c))/\sigma_f^2 \qquad (4\text{-}5)$$

which is used as indicator for distinct points by the interest operator by Förstner (cf. Paderes et al. 1984, Förstner/Gülch 1986). Here f_r and f_c are the partial derivatives of the intensity function $f(r,c)$ in row and column direction.

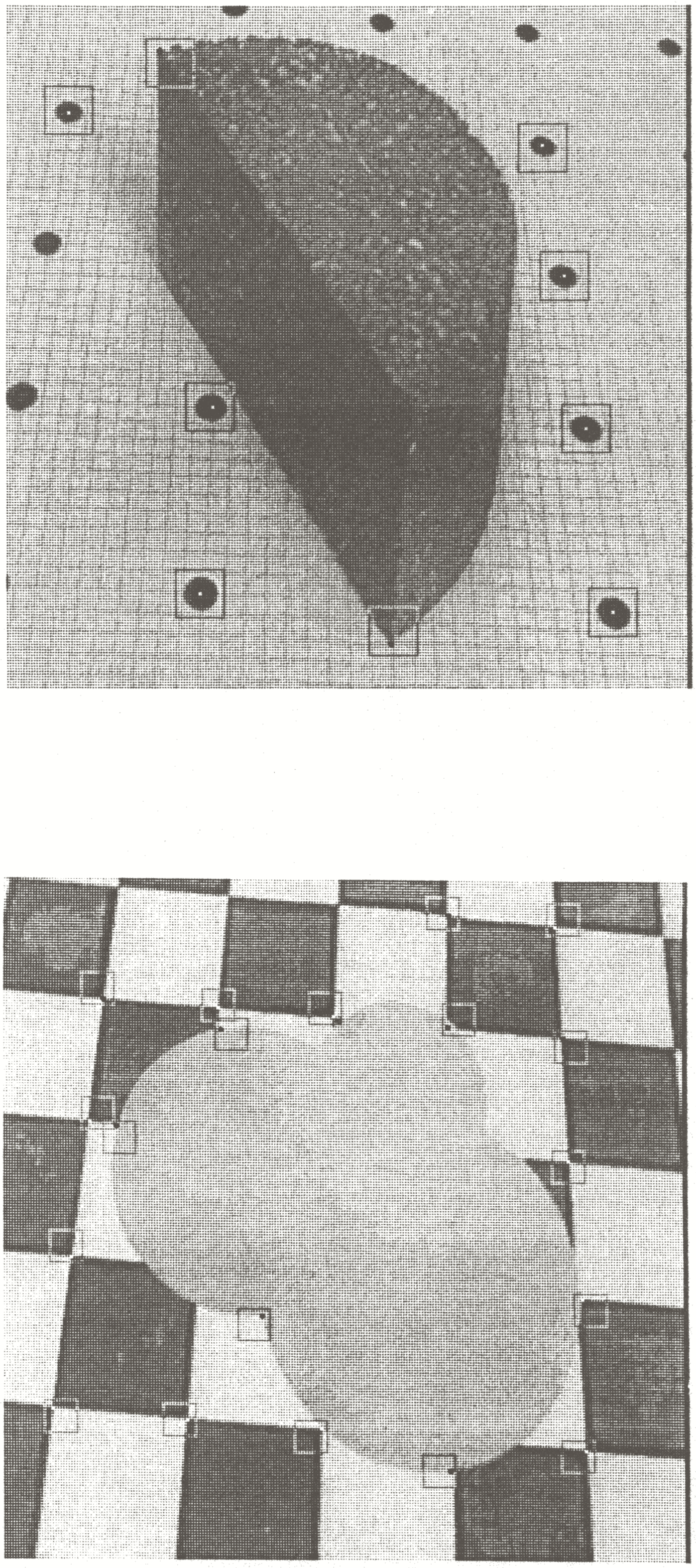
In a similar manner the uniqueness $u(P)$ of a point with respect to a set $Q = (Q_1, \ldots, Q_n)$ can be determined from their mutual information (cf. Förstner 1988) e. g. by

$$u(P) = \frac{1}{H(P;Q)} = \frac{-2}{\log(1-\rho^2_{PQ})} \qquad (4\text{-}6)$$

garanteeing that low mutual information between P and all $Q_i \in Q$ leads to a high uniqueness. Here the total correlation between P and $(Q_1, \ldots, Q_n)$ according to eq. (3-22) is used. Small correlations lead to high uniqueness measures, as to be expected.

Fig. 4-1 shows points automatically selected using eq. 4-5 as criterium for local distinctness. In addition a classification of the window content and an estimation of the optimal position of the point within the window is performed. The selection principle is obviously able to find (nearly] all distinct points.

Fig. 4.1 Automatically selected points using local distinctness according to eq. 4-5



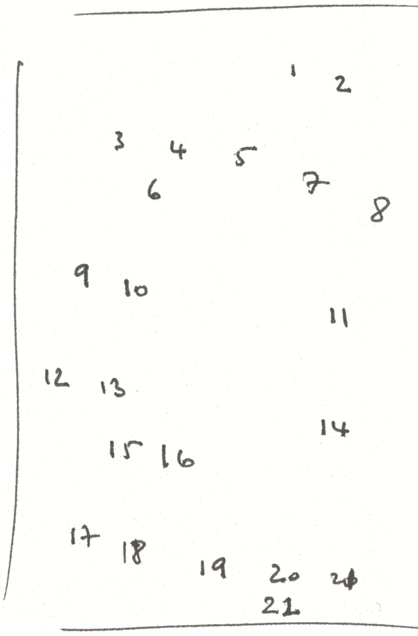


Fig 4-2 shows the uniqueness measures at the corners of checkerboard images, calculated using eq. 4-6. As to be expected the upper left corner and the border of the checker boards reveil the highest uniqueness value whereas the corners in the middle of the field show zero uniqueness values, due to their multiple appearance. When adding noise to the data the distance between the uniqueness measures decreases. Other examples for measuring the uniqueness, specifically of points in feature space, useful for classification, or of strings of symbols, useful for symbolic pattern-matching, can be found in (Förstner 1988).

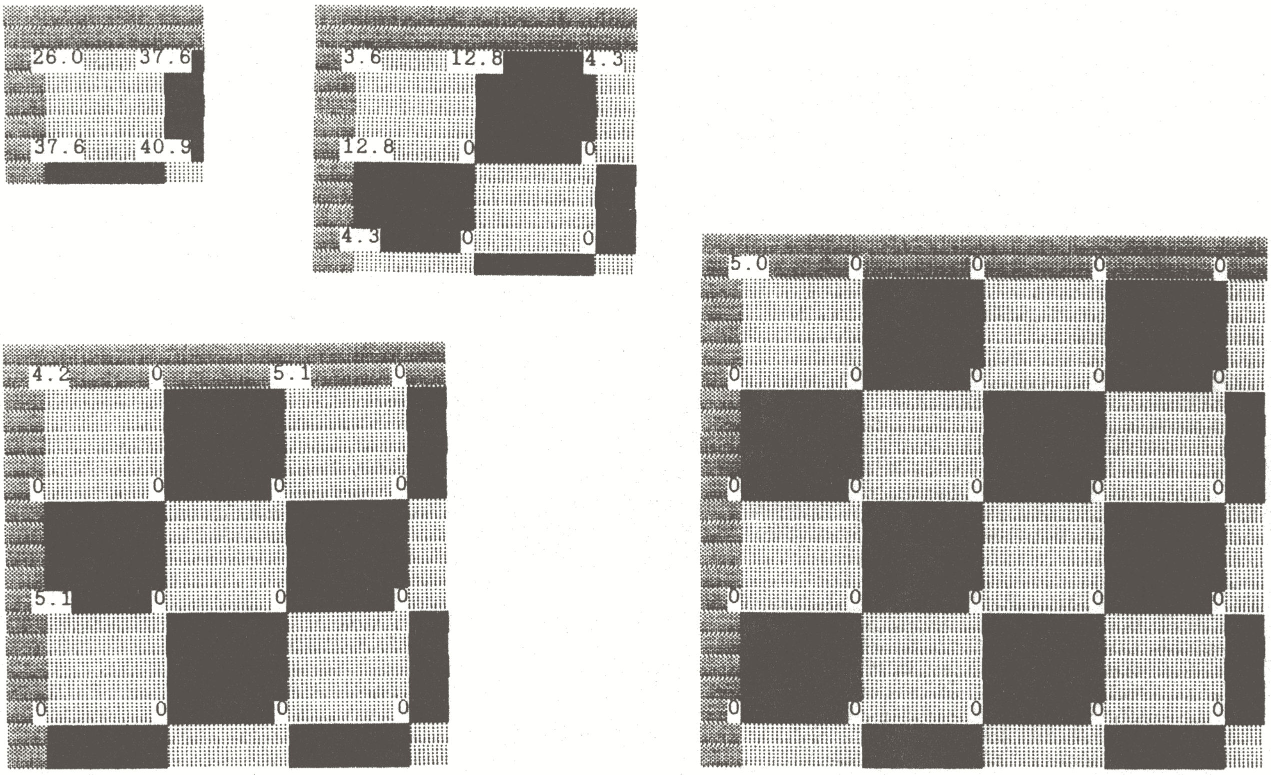


Fig. 4.2 Uniqueness measures of distinct points of parts of a checkerboard according to eq. 4-6 (from Förstner 1988)

4.3 Image Restauration

Image restauration is a prerequisit for image analysis especially if noise and blur prevent proper feature extraction. Classical restauration techniques are based on image models which are oversimplified, main reason for their inadequatness being their inability to preserve edge information, while at the same time supressing noise. Most edge preserving filters are ad hoc and therefore unpredictible in their performance.

We want to present the restauration technique proposed by Leclerc (1988) which is based on the principle of minimum description length encoding and does not show these drawbacks.Its image model states the image to consist of nonoverlapping homogeneous regions. In the most simple case homogenity means constant intensity. The model includes a noise component, e. g. being additive white noise.

We want to demonstrate the principle of the restauration scheme using a one dimensional profile. The profile then consists of intervalls with arbitrary length l_i and height h_i and additional noise v_i (a vector). Thus the observed function can be written as $y = F(l, h, v)$, where l, h and v are vectors. Assuming the components to be independent, thus the probability for a specific y being $P(y) = P(l, h, v) = P(l) \cdot P(h) \cdot P(v)$, we obtain the description length for the observed profile to be

$$D_1(y) = -lbP(y) = -lbP(l) - lbP(h) - lbP(v) \qquad (4\text{-}7)$$

As the coding of each intervall can be assumed to require a constant number $b = b_l + b_h$ of bits (e. g. 8 bits for b_l and b_h each, thus $b = 16$) and as the number of intervalls is one larger than the number of jumps we may write the description length as

$$D_1(y) = a\sum_{i=1}^{n} D(v_i) + b + b\cdot\sum_{i=1}^{n-1}(1-\rho(u_i - u_{i+1})) \tag{4-8}$$

where $D(v_i)$ is the description length for the i-th noise component, u_i is the true and unknown height at position i, $\delta(x)$ is the Kronecker symbol and $a = 1/(2ln2)$. As $D(v_i) = a\left(\frac{v_i}{\sigma}\right)^2 + c$ (cf. eq. 3-16, with $c = \frac{1}{2}lb2\pi\frac{\sigma^2}{\varepsilon^2}$) we have to minimize

$$D_1(y) = b + n\cdot c + a\sum_{i=1}^{n}\left(\frac{v_i}{\sigma}\right)^2 + b\sum_{i=1}^{n-1}(1-\delta(u_i - u_{i+1})) \tag{4-9}$$

with respect to the unknown u_i, which is equivalent to minimizing

$$D_2(y) = (D_1 - b - n\cdot c)/a = \sum_{i=1}^{n}\left(\frac{v_i}{\sigma}\right)^2 + \frac{b}{a}\sum_{i=1}^{n-1}(1-\delta(u_i - u_{i+1})) \tag{4-10}$$

with respect to u_i. Due to the second sum this optimization problem is extremely complex especially when transferred to two dimensions and shows many local minima. Using continuation techniques this type of problem may at least be solved approximately (cf. Blake/Zissermann 1987, Blake 1989). This may be achieved by replacing the $1 - \delta(u_i - u_{i+1})$ term by a less nonlinear function, e. g. having the shape of $\rho(x)$ in Fig. 2-1, thus being $(x/k)^2$ for $|x| < k$ and 1 else. This function for $k \longrightarrow 0$ yields $1 - \delta(x)$, which gives rise to an iteration sequence: namely starting with a large k solving $D_2^k(y) \longrightarrow \min$ and diminuishing k in the next iteration.

Leclerc (1988) has developped this sheme for two dimensional intensity functions including additional features : locally linear or quadratic functions, varying noise variance within regions and taking image blur into account. The result of the restauration scheme is shown ins Fig. 4-3 for a section of a digitized aerial image. Fig. 4-3a shows the original image, Fig 4-3b the restaurated image, Fig. 4-3c the discontinuities and Fig. 4-3d the restaurated image when only taking closed regions into account, which demonstrates the power of the method.

Two remarks are to be made here:

- Though the original model starts from closed regions the minimizing function $D(y)$ does not contain this restriction, but rather only is able to present independent discontinuities between neighbouring pixels, socalled crack edges. This leads to free, even unconnected discontinuities in the reconstruction, which may or may not be used in the analysis (cf. Vosselman 1983).
- The iterative estimation scheme suggests an interpretation as a robust estimation as discussed in section 2. The model $D_2^k(y)$, with k = standard derivation of height differencies, represents an observed Markov-Process of first order with possible outliers in the innovation sequence.

The model is extremely flexible as texture models and multispectral image data may be included and can be used as the basis for extracting generic shapes discussed in the next section.

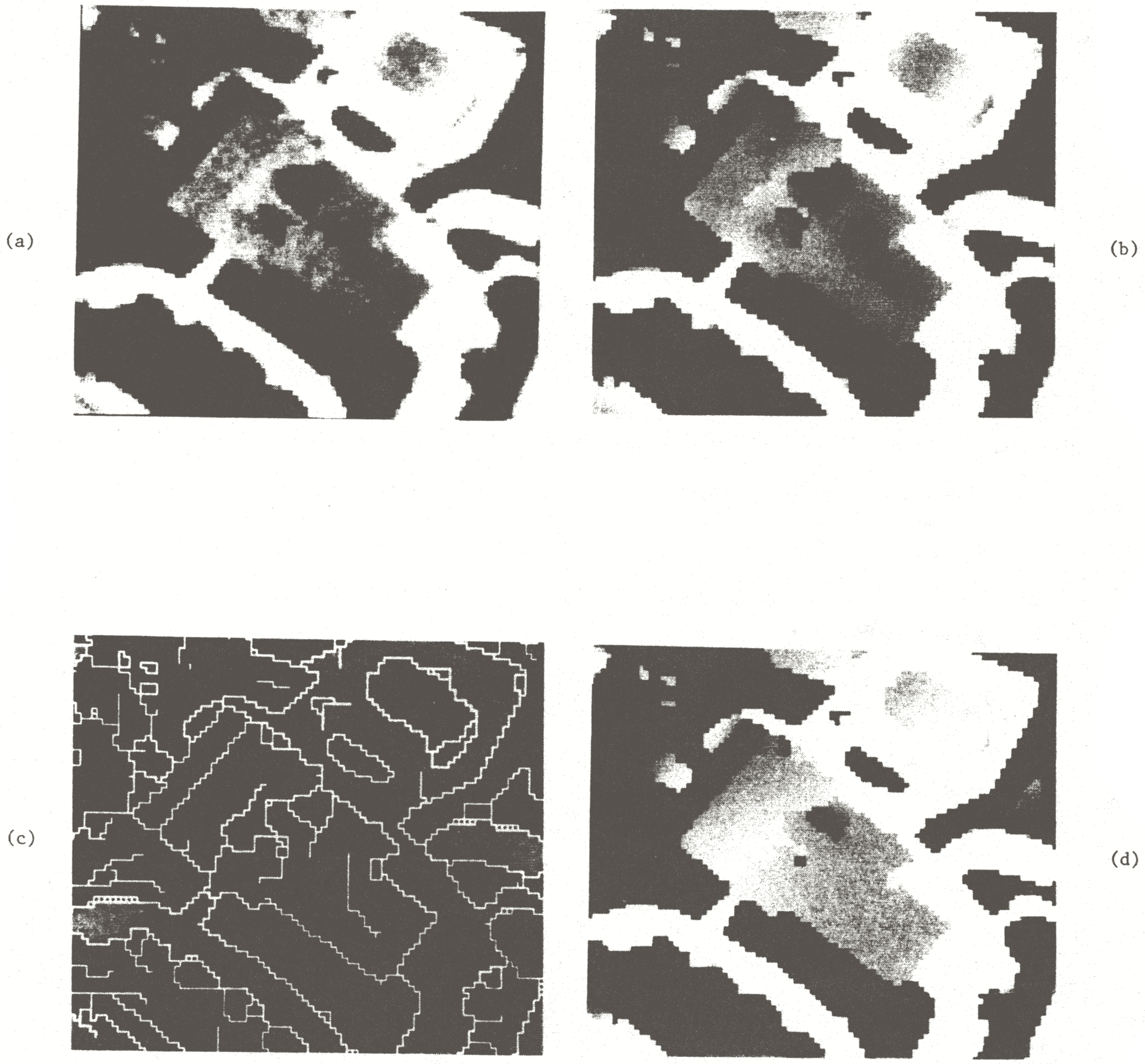
4.4 Extracting Generic Shapes

We finally want to discuss an approach for extracting generic shapes, specifically cultural objects such as buildings, as it has been developed by Fua and Hanson (1987, 1988).

The starting point are restored images as explained in the last subsection. They provide a basis for the generation of hypothesized image descriptions in the form of instances of generic models. The hypothesis generation process is an aggregation procedure collecting neighbouring image features, which show certain relations consistent with the generic model e. g. straight line segments, which are colinear or orthogonal. If a set of such line segments approximately encloses a region of high homogeneity it is closed and forms one candidate hypothesis which has to be evaluated with respect to competing ones. Such competing hypothesis may occur, because different but overlapping

Fig. 4.3 Image partitioning (from Leclerc 1988)

a. original image
b. restaurated image
c. discontinuities (crack edges)
d. restaurated image, only considering regions
and approximating the intensity surfaces by tilted planes



regions are formed and because the closing process may not be unique. We only want to discuss the evaluation process here as it is the core of the interpretation procedure.

The total score S of an interpretation is the logarithm of the probability for the model given the photometric evidence

$$S = -\{\text{description length}\} = lbP(m_0, m_1, \ldots, m_n|e_1, \ldots, e_n) \tag{4-11}$$

where each e_i represents the photometric evidence, i. e. the intensity values, supporting a specific model m_i; m_0 is the background model, not further specified. S, thus P, has to maximized following the principle of minimizing the conditional information and analogeously to eq. (4-1). Fua and Hanson show that under certain independence conditions P in eq. 4-11 can be written as

$$P = p(m_0, m_1, \ldots, m_n) \cdot \prod_{i=1}^{n} p(e_i|m_i)/p(e_i) \tag{4-12}$$

separating the probabilities for data and model. Then eq. 4-11 can be written as

$$S = -I(m|e) = I(e;m) - I(m) = F - G \tag{4-13}$$

It obviously contains two parts:

F: is the mutual information of photometric evidence and geometric model and - using $I(m|e) = I(e) - I(e|m)$ (cf. eq. 3-3) is the number of bits which can be saved in the description of the image when the model is told. This should be large.

G: is the number of bits to specify the geometric model. The more complex the model is, the smaller the score, therefore it has to be subtracted from F.

Though one immediately could have argued with minimizing the conditional information this separation into data and to model description is essential for the development of a practical procedure.

We only want to demonstrate the main features of this approach using a simple example, which actually is a part of the total interpretation scheme of Fua and Hanson.

Let us assume we have to evaluate a homogeneous area. The model free description of the area of A pixels requires $8A$ bits, assuming 8 bits image. When modelling the intensity function by a linear function (3 parameters) one can expect less bits to be necessary to describe the intensity values say kA with $k << 8$. It can be approximated by

$$kA = n(lb\sigma + c) + 8\bar{n} + \left[n lb\frac{n}{A} + \bar{n} lb\frac{\bar{n}}{A}\right] \tag{4-14}$$

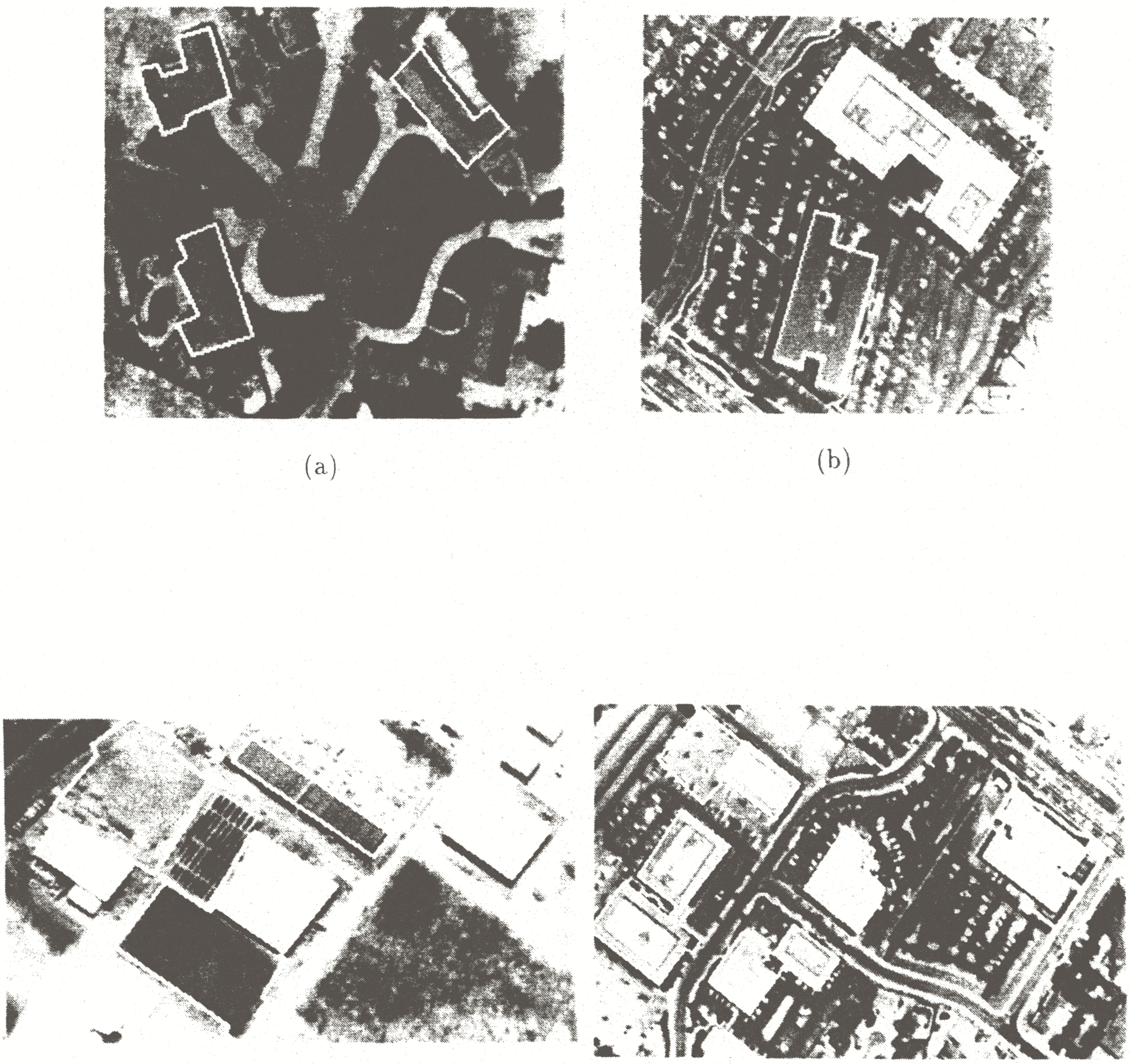
with $c = \frac{1}{2}\log 2\pi e$ (cf. eq. 3-5). The first term is required to describe the data being consistent with the model, the second for the data deviating from the model and the last term to specify whether an intensity value belongs to the model or not. In addition we need to take the description of the model into account, which requires 5 parameters, 3 for the linear function, 1 for the standard derivation of the noise and 1 for the percentage of outliers. Following the argumentation of Rissanen (1983, cf. eq. 3-20) we have to take $\frac{5}{2}lbA$ additional bits into account.

Now, Fua and Hanson introduce a parameter s which to some degree makes the evaluation *independent on the image scale*. They normalize the area A with s^2 leading to the dimensionless area measure A/s^2, which can be interpreted as if images of the same area are oversampled versions of a minimum image and argueing the number of bits should be invariant to the actual image scale. Replacing A by A/s^2 in all equations results in the photometric score

$$F_A = I(e) - I(e|m) = 8\frac{A}{s^2} - (k\frac{a}{s^2} + \frac{5}{2}lb\frac{A}{s^2}) \tag{4-15}$$

with k from eq. 4-14.

Fig. 4.4 Interpretation of four different images with respect to the building model at scale $s = 6$ (from Fua and Hanson 1988)



(a) (b) (c) (d)

The geometric cost simply is the deviation of the extracted geometry from an ideal one. For compact and rectilinear objects one would use

$$G_A = \frac{2L}{s} + \alpha \cdot \bar{\theta} \qquad (4\text{-}16)$$

where L is the length of the boundary in pixels and $\bar{\theta}$ is the average derivation of the sides of the object from multiplers of 90° refering to a principle axis of the object. As G should be small, the first term - again scaled - puts a penalty on objects with rough boundaries, whereas the second term prefers objects with rectilinear form.

As $S = F - G$, better models lead to better scores: let us assume also the walls of a building are contained in the model. Then no addition coding of the geometry is necessary, but F increases as *less* bits are necessary to encode the intensity function in the area of the wall of the building thus more bits are saved when being told the model increases the score. This demonstrates that high level structures of the model can be integrated and their effect onto the description may be evaluated.

In a similar manner Fua and Hanson have developed measures based on the lengthes of edges and on stereoscopic information, and applied it to buildings, roads and vegetation areas. Examples of the fully automatrix extraction of cultural objects are shown in Fig. 4-4 for a fixed scale parameter $s = 6$. Due to the fixed scale only buildings larger than a certain size have been extracted. Also non-building objects being compact, rectilinear and homogeneous areas have been proposed, but which easily could be eliminated interactively.

The image analysis techniques discussed in this section should have convinced the reader that not only theoretical concepts for image interpretation are available but also the implementations are promising for supporting Digital Photogrammetry.

5 Conclusions

The paper wanted to discuss image analysis techniques which may play a role in Digital Photogrammetry. We demonstrated that information theory can be used as a unifying framework for image interpretation, specifically to form a link between the observed intensity values and the in general complex object models which may contain both geometric as well as photometric components. Image matching, feature extraction, image restauration and location of objects; described by generic models can in an intuitive manner be reviewed as information extraction. To the simplicity of the models used so far the results require interactive evaluation, which however consists in comparably simple decisions.

This on one hand is reason enough to integrate automatic image analysis procedures into photogrammetric work stations in order to get experience with the upcoming techniques. This specifically will lead to a better understanding of the possibilities of semiautomatic interpretation techniques and help to more precisely define photogrammetric task. On the other hand the theoretical framework seems to be strong enough to further investigate the meaning and influence of up to now free parameters such as resolution or scale, to work out realistic generic models for objects relevant for automatic mapping and to extend the techniques to include spectral and textural information. Then techniques from artificial intelligence can be based on a solid ground and further increase the capabilities of Digital Photogrammetry.

References

Berger, T. (1971): Rate Distortion Theory. Prentice Hall, N. J., 1971.

Blake, A. (1989): Comparison of the Efficiency of Ddeterministic and Stochastic Algorithms for Visual Reconstruction. IEEE T-PAMI, 1989, pp. 2-12.

Blake, A., Zissermann (1987): Visual Reconstruction. Cambridge, Ma.: MIT Press, 1987.

Boyer, K. L. and Kak, A. C. (1986): Symbolic Stereo from Structural Descriptions. School of EE, Pudue, West Lafayette, TR-EE-86-12.

Boyer, K. L. and Kak, A. C. (1988): Structural Stereopsis for 3-D. IEEE T Pami, Vol. 10, No. 2, 1988, pp. 144-146.

Brooks, R. (1986): Model based 3-D Interpretation of 2-D Images. In Pentland 1986, pp. 292-321.

Canny, J. (1986): A Computational Approach to Edge Detection. IEEE T-PAMI-8, No.6, pp. 678-698.

Förstner, W. (1988): Statistische Verfahren für die Automatische Bildanalyse und ihre Bewertung bei der Objekterkennung und -vermessung. Habilitationsschrift, Stuttgart, 1988.

Förstner, W. and Gülch, E. (1986): A Fast Interest Operator for Detection and Precise Location of Distinct Points, Corners and Centres of Circular Features. Proc. of Intercomm. Conference on Fast Processing of Photogrammetric Data, Interlaken, 1987, pp. 281-305.

Fua, P., Hanson, A. J. (1987): Resegmentation Using Generic Shape: Locating General Cultural Objects. Pattern Recognition Letters 5, pp. 243-252, 1987.

Fua, P., Hanson, A. J. (1988): Generic Feature Extraction Using Probability-Based Objective Functions. Submitted to IJCV

Georgeff, M. P., Wallace, C. S. (1984): A General Selection Criterion for Inductive Inference. Proc. of Advances in Artificial Intelligence, Italy Sept., 1984, T.O'Shea (Ed). North Holland Amsterdam 1984.

Giri, N. C. (1977): Multivariate Statistical Inference. Academic Pr., 1977.

Grimson, W. E. L. (1981): From images to Surfaces. Cambridge, Mass.: MIT Press, 1987.

Hanson, A. J., Quam, L. H. (1988): Overview of the SRI Cartographic Modeling Environment. In Proceedings Image Understanding Workshop, Cambridge, Mass., 1988, pp. 576-782.

Haralick, R. M. (1988): Mathematical Morphology and the Morphological Sampling Theorem. In: Proceedings Image Understanding Workshop, Cambridge, Mass. 1988, pp. 461-487.

Haralick, R. M. (1984): Digital Step Edges from Zero Crossings of Second Directional Derivatives. IEEE T-PAMI-6, No.1, pp. 58-68.

Haralick, R. M., Sternberg, S. R., and Zhuang, X. (1987): Image Analysis Using Mathematical Morphology. In: IEEE T-PAMI-9, No. 4, 1987, pp. 532-550.

Herman, M., Kanade, T. (1986): The 3-D Mosaic Scene Understanding System. In Pentland 1986, pp. 322-358.

Hölzer, E., Holzwarth,E. (1976): Pulstechnik. Berlin, Springer, 1976.

Huber, P. J. (1981): Robust Statistics. Wiley, NY, 1981.

Huertas, A., Nevatia, R. (1988): Detecting Buildings in Aerial Images. CVGIP 41, pp. 131-152.

Kashyap, R. L. (1985): Univariate and Multivariate Random Field Models. In: Digital Image Processing and Analysis. Ed. Chelappa and Sawchuk, Vol. 1, IEEE Computer Society.

Krarup, T. (1967): Internal Report, Kopenhagen. (cf. Krarup et al. 1980)

Krarup, T., Juhl, J., Kubik, K. (1980): Götterdämmerung over Least Squares Adjustment. Int. Arch. of Photogr., Vol. 23, B3, Hamburg, 1980, pp. 369-378.

Leclerc, Y. G. (1988): Image Partitioning for Constructing Stable Descriptions. Proc. of Image Understanding Workshop, Cambridge, MA, 1988.

Malik, J. , Perona P. (1989): A Computational Model of Texture segmentation. In: Proceedings IEEE CVPR Conf., San Diego, Calif. 1989, pp.326-332.

McKeown, D. M., Harvey, W. A., McDermott, J. (1985): Rule-Based Interpretation of Aerial Imagery, IEEE T-PAMI-7, No.5, pp. 570-585.

McKeown, D. M., Denlinger, J. L. (1988): Cooperative Methods for Road Tracking in Aerial Imagery, in Proceedings Image Understanding Workshop, Cambridge, Mass., 1988, pp. 327-341.

McKeown, D. M., Harvey, W. A., Wixson, L. E. (1988): Automatic Aquisition for Aerial Image Interpretation. Submitted to CVGIP, 1988.

Mulder, J. A., Mackworth, A. K., Havens, W. S. (1988): Knowledge Structuring and Constraint Satisfaction: the Mapsee Approach, IEEE T-PAMI-10, No. 6, pp.866-879.

Paderes, F. C., Mikhail, E. M., Förstner, W. (1984): Rectification of Single and Multiple Frames of Satellite Scanner Imagery Using Points and Edges as Control. NASA Symposium on Mathematical Pattern Recognition and Image Analysis, Houston, 1984.

Pentland, A. P. (Ed., 1986): From Pixels to Predicates. Norwood, N. J. Ablex Publ. Co., 1986.

Rao, R., Schunck, B. G. (1989): Computing Orientated Texture Fields. In: Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, Calif. 1989, pp. 61-68.

Rissanen, I. (1983): A Universal Prior for Integers and Estimation by Minimum Description Length, The Annals of Statistics 2, pp. 416-431, 1983.

Rissanen, I. (1987): Minimum Description-Length Principle. In Encyclopedia of Statistical Sciences, 5, pp. 523-527, 1987.

Serra, J. (1982): Image Analysis and Mathematical Morphology. London, Academic Press.

Shannon, C. E., Weaver, W. (1949): The Mathematical Theory of Communication. The University of Illinois Press, Urbana 1949.

Terzepoulos, D. (1986): Regularization of Inverse Visual Problems of Involving Discontinuities, IEEE T-PAMI-8, No. 2, 1986, pp. 129-139.

Vosselman, G. (1989): Relationale Bildzuordnung für die Objektlokalisierung, Seminar "Wissensgestützte Bildanalyse". Stuttgart, Mai 1989.

Vosselman, G. (1989): Symbolic Image Description for Relational Matching. In: Linkwitz, K., Hangleiter, U. (Eds.), High Precision Navigation. Berlin, Springer, 1989, pp. 378-391.

Yuille, A. L., Poggio, T. A. (1986): Scaling Theorems for Zero Crossings. IEEE T-PAMI-8, No. 1, pp. 15-25.

Abstract:
The paper discusses image analysis techniques for interpreting aerial images which can easily be related to methods from information theory, as methods for image matching, image restauration and feature extraction as well for image analysis may be assessed using information theoretic concepts. Specifically the evaluation method, proposed by Fua and Hanson (1987/88) is able to integrate photometric and geometric as well as low and high level structures of object models and image data in a far reaching manner. The basic concepts of information theory, and the relations to least square and robust estimation techniques are discussed using examples from data and image anlysis.

Bildanalysemethoden für die Digitale Photogrammetrie

Zusammenfassung:
Der Beitrag behandelt exemplarisch Methoden der Bildanalyse für die Informationsextraktion aus Luftbildern. Grundlage für die Auswahl der Verfahren ist die Methode die Fua und Hanson (1987, 1988) für die Bewertung von Bildinterpretationen vorschlugen. Sie ist in der Lage photometrische und geometrische, sowie einfache und komplexe Komponenten von Modell und Daten in bisher weitreichendster Weise zu integrieren und stützt sich wesentlich auf die von Shannon entwickelte Informationstheorie. Die Länge der Beschreibung einer Interpretation in bits wird als Maß für die Einfachheit der Erklärung der Daten durch ein Modell verwendet. Es wird gezeigt, daß auch andere grundlegende Bildanalyseverfahren, wie die Bildzuordnung, die Merkmalsextraktion und die Bildrestaurierung sich in das informationstheoretische Konzept integrieren lassen.

Wolfgang Förstner
Institut für Photogrammetrie
Universität Stuttgart
Keplerstr. 11
D-7000 Stuttgart 1