RZ 1170 (#42264) 9/7/82
Mathematics 27 pages

Research Report

LOCALIZATION OF INFORMATION ON FINITE RANDOM FUNCTIONS

Rainer F. Hauser

IBM Zurich Research Laboratory, 8803 Rüschlikon-ZH, Switzerland

Typed by Charlotte Bolliger using the IBM MC82

IBM Research Division
San Jose · Yorktown · Zurich

Copies may be requested from:

IBM Thomas J. Watson Research Center
Distribution Services 38-066
Post Office Box 218
Yorktown Heights, New York 10598

RZ 1170 (#42264) 9/7/82
Mathematics 27 pages

LOCALIZATION OF INFORMATION ON FINITE RANDOM FUNCTIONS

Rainer F. Hauser

IBM Zurich Research Laboratory, 8803 Rüschlikon-ZH, Switzerland

Typed by Charlotte Bolliger using the IBM MC82

ABSTRACT: Primarily for applications in the field of digital image processing, the mathematical concept of information density on finite random functions is defined and some properties are extracted. The idea is to localize the global information content at the individual points in presence of dependences between points. It turns out that the information density is a nonnegative value defined and computable on each point, and that the global information content of the whole random function is the sum of the information densities of all points. Based on these concepts, an edge-detection algorithm is outlined.

August 24, 1982

1. INTRODUCTION

In digital image processing, pictorial information is generally represented as arrays of grey values [1]. The images modeled in this way are discrete, and the grey values representing the brightness or luminance of the point are quantized. For the solution of basic image-processing problems such as scale change, rotation, and requantization, this representation is a good approach. For more sophisticated image-processing functions, it may be convenient to represent images in a stochastic way. Images can be modeled as arrays of random variables called discrete random fields (see Appendix A). This approach is appropriate in image-processing functions such as image compression, noise filtering, halftoning, edge detection, and in most recognition functions.

Roughly, human perception of an image differentiates between two kinds of information. The first is present at the edges or contours. The second is the texture, i.e., large areas devoid of significant detail. Texture representation is possible using only few parameters [2], whereas contours must be treated more carefully. Pattern recognition and image understanding (or their preprocesses) begin with an image modeled as an array of grey values. As a first step, this image must be transformed into a list of contours and texture as the first level of abstract image representation. For one part of this task, a reliable edge-detection or contour-finding algorithm is required. In further steps, recognition will proceed to higher abstractions, but initially we are only interested in this first preprocessing step [3].

As a part of probability theory, an attempt is made to apply information theory to images. This information content of an image is mainly concentrated along the contours, because the local information is high when the prediction, given some other points, is bad and vice versa. The concept of information in communication theory as introduced by Shannon with a sender, a noisy channel

and a receiver is not very useful here, as it implies concepts like causality, etc. which do not have any meaning for images. Instead, we start with a model of prediction, find the probabilities, and take the information as the negative logarithm of the probability. (See Appendix B for more details.) The probabilities are given as joint probabilities, and individual image points are usually dependent on one another. Our intuitive understanding of images makes it desirable to have an expression for the local information density at each point in the image, in addition to the global information content of the whole image and the marginal information content of subimages. In Section 3, such a local information density is defined. The main result of this work is expressed in formula (6).

Based on this definition, an algorithm to detect luminance edges on a grey-level image can be implemented. The information density at a point can be thresholded to detect an edge point. To compare different edge-detection algorithms, one has to consider the requirements for a good edge-dection algorithm. According to [4], three criteria can be stated: 1) The detected parts of edges or contours should be connected (criterion of continuation). 2) The lines representing the edges or contours should be thin (criterion of thinness). 3) The edges or contours should be in the right place (criterion of correct location). As another important criterion, the sensitivity toward noise should be mentioned. For the time being, we are not concerned with testing these criteria, but they should be borne in mind when talking about edge detection and contour finding. The main difficulty of the information-density approach is not the definition of information density, but the model of prediction (i.e., the stochastic model of the image). Whether the algorithm is good with respect to these criteria depends mainly on the stochastic model chosen.

Edge detection is one of the most important steps towards image recognition and image understanding. Typical applications are found in character recognition, where the connections of the contours are important, and automatic

inspection of printed circuit boards, where apart from the connections, the main problem is the small differences in grey values between foreground and background, so that noise plays an important role. The problems and the state of the art in the field of edge detection today are surveyed in [3].

2. RANDOM FUNCTIONS AND MODELING OF IMAGES

Random functions (also called stochastic functions, random processes) are widely used in physics, but they can also be applied to image processing. Some reasons for modeling an image as a random function are given in Appendix A.

2.1. Concepts of probability theory

Given a probability space $(\Omega, \mathcal{A}, P)$, where Ω is a set, $\mathcal{A}$ a σ-algebra over Ω, and P a normed measure on $\mathcal{A}$, a random variable ξ is a measurable function

$$\xi : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B}),$$

where $\mathbb{R}$ is the set of real numbers, and $\mathcal{B}$ the Borel σ-algebra over $\mathbb{R}$. A random variable ξ is defined by its distribution function

$$F_\xi(x) = P(\xi \leq x).$$

An important concept in probability theory is the concept of dependence. Two events A_1 and A_2 in $\mathcal{A}$ are independent if $P(A_1 \wedge A_2) = P(A_1) \cdot (A_2)$. The conditional probability of the event A_1 given the event A_2 is defined as

$$P(A_1 | A_2) = \frac{P(A_1 \wedge A_2)}{P(A_2)},$$

where $P(A_2) > 0$ is assumed. The two events A_1 and A_2 are called conditionally independent given the event A_3 (with $P(A_3) > 0$) if
$P(A_1 \wedge A_2 | A_3) = P(A_1 | A_3) \cdot P(A_2 | A_3)$.

2.2. Random functions

We define a random function ξ as a set $\{\xi_t | t \in T\}$ of random variables ξ_t with the index set T. Instead of this definition, equivalently it can be said that a random function ξ is a function

$$\xi : (\Omega, \mathcal{A}, P) \times T \to (\mathbb{R}, \mathcal{B}), \quad (\omega, t) \to \xi_t(\omega),$$

where the probability space $(\Omega, \mathcal{A}, P)$ must be arranged so that all point-probability spaces $(\Omega_t, \mathcal{A}_t, P_t)$ are contained in $(\Omega, \mathcal{A}, P)$. (This can be forced by taking the product of all Ω_t and $\mathcal{A}_t$.) The dependences of two points ξ_{t_1} and ξ_{t_2} are modeled in the probability distribution P.

We do not assume any structure on the set T. Later, we can give the set T different structures. To treat Markov chains, T must be totally ordered. Discrete images need a relation expressing that two image points are neighbors. But here, we wish to state that in the basic definition a random function is an unstructured set of random variables with a given joint probability distribution P.

2.3. Image models

Mathematically modeled, an image is a function

$$B : R_x \times R_y \to G, \quad (x,y) \to g = B(x,y),$$

and the two sets R_x and R_y are usually intervals of the set $\mathbb{R}$ of real

numbers or of the set $\mathbb{Z}$ of integers. If they are intervals of $\mathbb{R}$, the image is called continuous, and if they are intervals of $\mathbb{Z}$, the image is called discrete [1].

For deterministic images, the set G is also an interval of $\mathbb{R}$ or $\mathbb{Z}$. If G is an interval of $\mathbb{Z}$, the image is called quantized. The numbers in G represent the different grey values possible for image points. The value $B(x,y)$ can be interpreted as the brightness or luminance at (x,y). A discrete deterministic image is nothing but an array of values of G.

A stochastically modeled image is a random function with a two-dimensional set T. This means that G is a set of random variables. The value $B(x,y)$ will be interpreted as the probability distribution of the brightness of the image B in the point (x,y). If all possible random variables in G only take values in a finite subset of $\mathbb{R}$, the image is also called quantized.

The deterministic model is a special case of the stochastic model, since each point can be said to have a probability density function of the form of a δ-function. Each stochastically modeled image is a weighted (by its probability) set of samples, and each sample is a deterministically modeled image.

2.4. The modeling power of stochastic images

A stochastic image is defined as a random function with index set $R_x \times R_y$. This index set has a topological structure, i.e., a neighborhood relation between points. The points close together seem to be highly correlated (see Appendix A). We call these dependences topological dependences. Topological dependences over neighborhoods with a radius of 20 pels and more are normal

in image processing. But there are also dependences between parts of an image far away from each other, which we shall call congruences. (We have an image of a train in mind where all wagons may look the same.) In natural images, the congruences are of no importance. But this is not true for artificial images such as geometrical drawings or text. All parts of a line are congruent to all other parts of the same line. And all letters "A" in a printed text are more-or-less equal, independent of where they occur in the text.

In many image-processing tasks, the stochastic image model may help to develop much more effective algorithms for halftoning, image compression, edge detection and recognition.

2.5. The relation between random function and sample

In image processing, an image is usually given as a sample and not as a random function. The question is how to estimate the random function from one sample (or from a small set of samples). We do not wish to treat all known possible approaches to solve the estimation problem, but we shall add some remarks about this problem from the image-processing point of view.

With a topological model of dependences in mind, one can estimate the dependences. Given some image-processing environment (scanner, printer and their respective resolutions, etc.), one can experimentally determine the topological dependences for a class of images (the class of text images, the class of line-art images, the class of natural images). Especially, one can measure the "strength" of the dependence for a given distance. Here, one is not concerned with the image data of a special image, but with the topological structure of the index set. From the given point-probability distributions P_t, one estimates the joint-probability distribution P. One approach is to take Markov chains generalized to two-dimensional index sets.

One also has to estimate each point-probability distribution P_t. Normal distributions (also called Gaussian distributions) are often used, since among other things, they have the following properties ([5], pp. 22-24): 1) They are uniquely characterized by the expectation and the covariance. 2) The concepts of independence and uncorrelatedness are equivalent. 3) The marginal distributions and the conditional distributions are also normal. 4) Under a nonsingular linear transformation, a normal distribution becomes an easily calculable normal distribution.

We cannot give more hints on how to estimate an image as a random function, but it is a very important task in image processing. Often, image-processing algorithms can be considerably improved by taking better stochastic models, but a lot of research has still to be done in this field.

3. INFORMATION ON RANDOM FUNCTIONS

In this section, we shall define the global information content of a finite random function $\{\xi_t | t \in T\}$, and localize it at the points ξ_t. We assume a random function ξ and one of its samples s to be given. The information content of ξ is here always measured with respect to the given sample s.

3.1. Concepts of information theory

Given the probability space $(\Omega, \mathcal{A}, P)$, the information content $I(A)$ of an event $A \in \mathcal{A}$ is defined by

$$I(A) = -\log P(A).$$

The information content $I(\xi)$ of a random variable ξ with respect to a sample s is defined by

$$I(\xi) = I(\xi = s) = -\log P(\xi = s). \tag{1}$$

The information content of a set of events $\{A_j | j \in J\}$ is defined as the joint information content $I(\bigwedge_{j\in J} A_j)$. It is easy to see from the definitions that $I(A) \geq 0$ and $I(A) = 0$ only if $P(A) = 1$. The information content of an event A depends only on the probability of A and on no other properties of A. The information content $I(\xi)$ of a random variable with respect to the sample s depends only on the σ-algebra and the sample s but not on the value $\xi(\omega)$ in $\mathbb{R}$.

Given a finite random function $\{\xi_t | t \in T\}$ and a sample $\{s_t | t \in T\}$, we use some abbreviations to simplify the notation. We write $P(S)$ and $I(S)$ for $P\left(\bigwedge_{t\in S} (\xi_t = s_t)\right)$ and $I\left(\bigwedge_{t\in S} (\xi_t = s_t)\right)$. In the same way, we write $I(t)$ and $I(t_1,t_2)$ for $I(\xi_t = s_t)$ and $I(\xi_{t_1} = s_{t_1} \wedge \xi_{t_2} = s_{t_2})$, and so on.

3.2. Localization of the information

In the case where all points of $\{\xi_t | t \in T\}$ are independent, the information can easily be localized. The joint probability $P(T)$ is the product of all point probabilities $P(t)$, and the joint information content $I(T)$ is the sum of all point information contents $I(t)$. But dependences connect the different points, and joint information is not the sum of the point information.

We request that the information density $J_T(t)$, i.e., the localized information, on a random function $\{\xi_t | t \in T\}$ satisfies the following two requirements:

1) $J_T(t) \geq 0$ for all $t \in T$,

2) $\sum_{t \in T} J_T(t) = I(t)$.

The information density must reflect the coupling between points by dependences. So we assume that the coupling between one point t and the rest of the points in T can be divided into coupling between the point and no other points, the point and one other point, the point and two other points, and so on. This concept leads to the definition

$$I(S) = \sum_{S' \subset S} D(S') \tag{2}$$

for all $S \subset T$. We interpret the terms $D(S')$ as coupling terms. This definition was inspired by the concept of relative information ([6], pp. 450-451). Since $P(\phi) = 1$, we get $D(\phi) = I(\phi) = 0$, $D(t) = I(t)$ and $D(t_1,t_2) = I(t_1,t_2) - I(t_1) - I(t_2)$. The term $D(t_1,t_2)$ is the negative relative information. These coupling terms $D(S')$ are now assumed to belong in equal parts to all points of S', and the information density $J_T(t)$ at the point t is defined as

$$J_T(t) = \sum_{t \in S \subset T} \frac{1}{|S|} D(S), \tag{3}$$

where the sum goes over all subsets of T containing t. It can easily be seen that $\sum_{t \in S} J_T(t)$ must not be equal to $I(S)$ except for the independent case, because points in T but outside S may have an influence in $J_T(t)$. But for $S = T$, they are equal because no points outside T are involved.

Theorem:

$$I(T) = \sum_{t \in T} J_T(t). \tag{4}$$

Proof:

$$\sum_{t \in T} J_T(t) = \sum_{t \in T} \sum_{\substack{S \subset T \\ t \in S}} \frac{1}{|S|} D(S) = \sum_{\substack{S \subset T \\ S \neq \phi}} \sum_{t \in S} \frac{1}{|S|} D(S) = \sum_{\substack{S \subset T \\ S \neq \phi}} |S| \frac{1}{|S|} D(S) = I(T). \quad \blacksquare$$

It is a question whether the two assumptions expressed in formulas (2) and (3) are good assumptions (see Appendix B).

3.3. Calculation of the information density

The information density $J_T(t)$ is defined via the auxiliary coupling terms $D(S)$. We shall now express the information density of a point with the given information $I(S)$.

The transformation from the $D(S)$ to the $I(S)$ is linear and nonsingular. The inverse transformation is calculated in the following theorem.

Theorem:

$$D(S) = \sum_{S' \subset S} (-1)^{|S|-|S'|} I(S'). \tag{5}$$

Proof:

$I(S) = \sum_{S' \subset S} \sum_{S'' \subset S'} (-1)^{|S'|-|S''|} I(S'')$ must be shown.

$$\sum_{S' \subset S} \sum_{S'' \subset S'} (-1)^{|S'|-|S''|} I(S'') = \sum_{S'' \subset S' \subset S} (-1)^{|S'|-|S''|} I(S'')$$

$$= \sum_{S'' \subset S} \sum_{S'' \subset S' \subset S} (-1)^{|S'|-|S''|} I(S'') = \sum_{S'' \subset S} I(S'') \sum_{S'' \subset S' \subset S} (-1)^{|S'|-|S''|}$$

$$= \sum_{S'' \subset S} I(S'') \sum_{k=|S''|}^{|S|} \binom{|S|-|S''|}{k-|S''|} (-1)^{k-|S''|} = \sum_{S'' \subset S} I(S'') \sum_{k=0}^{|S|-|S''|} \binom{|S|-|S''|}{k} (-1)^{k}$$

$$= \sum_{S'' \subset S} I(S'') \sum_{k=0}^{|S|-|S''|} \binom{|S|-|S''|}{k} (-1)^{k} \cdot (+1)^{|S|-|S''|-k}$$

$$= \sum_{S'' \subset S} I(S'') \, (1 - 1)^{|S|-|S''|} = I(S). \quad \blacksquare$$

With the inverse transformation shown in formula (5), the information density $J_T(t)$ defined in formula (3) can be expressed in the terms $I(S)$. As a preceding step, we prove the following Lemma.

Lemma:

$$\frac{1}{n} \cdot \frac{1}{\binom{n-1}{s}} = \sum_{k=0}^{n-s-1} \frac{(-1)^k}{s+k+1} \binom{n-s-1}{k} \quad \text{for all } 0 \le s < n.$$

Proof:

We set

$$A_n(s) = \frac{1}{n} \frac{1}{\binom{n-1}{s}}; \quad B_n(s) = \sum_{k=0}^{n-s-1} \frac{(-1)^k}{s+k+1} \binom{n-s-1}{k}.$$

We prove $A_n(0) = B_n(0)$:

$$B_n(0) = \sum_{k=0}^{n-1} \frac{(-1)^k}{k+1}\binom{n-1}{k} = \sum_{k=0}^{n-1} \frac{(-1)^k}{k+1}\,\frac{(n-1)!}{k!(n-k-1)!}$$

$$= \sum_{k=0}^{n-1} \frac{(-1)^k (n-1)!}{(k+1)!(n-k-1)!} = \frac{1}{n}\sum_{k=0}^{n-1}\binom{n}{k+1}(-1)^k$$

$$= -\frac{1}{n}\left(\left(\sum_{k=0}^{n}\binom{n}{k}(-1)^k(+1)^{n-k}\right) - 1\right) = -\frac{1}{n}\left((1-1)^n - 1\right)$$

$$= \frac{1}{n} = A_n(0) \text{ for } n \geq 1.$$

Now, we prove $A_{n+1}(s+1) = A_n(s) - A_{n+1}(s)$:

$$A_{n+1}(s+1) - A_n(s) = \frac{(s+1)!(n-s)!}{(n+1)!} \cdot \frac{1}{n-s} - \frac{s!(n-s)!}{n!}\,\frac{1}{n-s}$$

$$= \frac{1}{n-s}\left(\frac{(s+1)!(n-s)! - (n+1)\cdot s!(n-s)!}{(n+1)!}\right)$$

$$= \frac{1}{n-s}\left(\frac{s!(n-s)!}{(n+1)!}(s-n)\right) = -\frac{s!(n-s)!}{(n+1)!} = -\frac{1}{n+1}\,\frac{1}{\binom{n}{s}} = -A_{n+1}(s).$$

Now, we prove $B_{n+1}(s+1) = B_n(s) - B_{n+1}(s)$:

$$B_{n+1}(s) + B_{n+1}(s+1) = \sum_{k=0}^{(n+1)-s-1} \frac{(-1)^k}{s+k+1}\binom{(n+1)-s-1}{k} +$$

$$+ \sum_{k=0}^{(n+1)-(s+1)-1} \frac{(-1)^k}{(s+1)+k+1}\binom{(n+1)-(s+1)-1}{k}$$

$$= \sum_{k=0}^{n-s} \frac{(-1)^k}{s+k+1}\binom{n-s}{k} + \sum_{k=0}^{n-s-1} \frac{(-1)^k}{s+k+2}\binom{n-s-1}{k}$$

$$= \sum_{k=0}^{n-s} \frac{(-1)^k}{s+k+1}\left(\binom{n-s-1}{k-1} + \binom{n-s-1}{k}\right) +$$

$$+ \sum_{k=0}^{n-s-1} \frac{(-1)^k}{s+k+2}\binom{n-s-1}{k} =$$

$$= \sum_{k=0}^{n-s} \frac{(-1)^k}{s+k+1} \binom{n-s-1}{k-1} + \sum_{k=0}^{n-s} \frac{(-1)^k}{s+k+1} \binom{n-s-1}{k} +$$

$$+ \sum_{k=0}^{n-s-1} \frac{(-1)^k}{s+k+2} \binom{n-s-1}{k} = \sum_{k=(-1)}^{n-s-1} \frac{(-1)^{k+1}}{s+k+2} \binom{n-s-1}{k} +$$

$$+ \sum_{k=0}^{n-s} \frac{(-1)^k}{s+k+1} \binom{n-s-1}{k} + \sum_{k=0}^{n-s-1} \frac{(-1)^k}{s+k+2} \binom{n-s-1}{k}$$

$$= \sum_{k=0}^{n-s-1} \frac{(-1)^k}{s+k+1} \binom{n-s-1}{k} = B_n(s).$$

This proves with induction $A_n(s) = B_n(s)$ for all $n > 0$ and all s with $0 \leq s < n$. ■

With this Lemma, we can express the information density $J_T(t)$ with the information terms $I(S)$.

Theorem:

$$J_T(t) = \frac{1}{|T|} \sum_{S \subset T \setminus \{t\}} \frac{1}{\binom{|T|-1}{|S|}} I(t|S). \tag{6}$$

Proof:

$$J_T(t) = \sum_{t \in S \subset T} \frac{1}{|S|} \sum_{S' \subset S} (-1)^{|S|-|S'|} I(S')$$

$$= \sum_{S' \subset T} I(S') \sum_{\substack{S' \subset S \subset T \\ t \in S}} \frac{1}{|S|} (-1)^{|S|-|S'|}$$

$$= \sum_{\substack{S' \subset T \\ t \in S'}} I(S') \sum_{S' \subset S \subset T} \frac{1}{|S|} (-1)^{|S|-|S'|} +$$

$$+ \sum_{S' \subset T \setminus \{t\}} I(S') \sum_{S' \subset S \subset T \setminus \{t\}} \frac{1}{|S|+1} (-1)^{|S|-|S'|+1} =$$

$$= \sum_{t\in S'\subset T} I(S') \sum_{k=|S'|}^{|T|} \frac{1}{k}\binom{|T| - |S'|}{k - |S'|}(-1)^{k-|S'|} +$$

$$+ \sum_{S'\subset T\setminus\{t\}} I(S') \sum_{k=|S'|}^{|T|-1} \frac{1}{k+1}\binom{|T| - |S'| - 1}{k - |S'|}(-1)^{k-|S'|+1}$$

$$= \sum_{S''\subset T\setminus\{t\}} I(S''\cup\{t\}) \sum_{k=0}^{|T|-|S''|-1} \frac{(-1)^k}{|S''| + 1 + k}\binom{|T| - |S''| - 1}{k} +$$

$$+ \sum_{S'\subset T\setminus\{t\}} I(S') \sum_{k=0}^{|T|-|S'|-1} \frac{(-1)^{k+1}}{|S'| + k + 1}\binom{|T| - |S'| - 1}{k}$$

$$= \sum_{S'\subset T\setminus\{t\}} \Big(I(S'\cup\{t\}) - I(S')\Big) \sum_{k=0}^{|T|-|S'|-1} \frac{(-1)^k}{|S'| + k + 1} \cdot$$

$$\cdot \binom{|T| - |S'| - 1}{k} = \frac{1}{|T|} \sum_{S\subset T\setminus\{t\}} \binom{|T| - 1}{|S|}^{-1} \Big(I(S\cup\{t\}) - I(S)\Big).$$

With $I(S\cup\{t\}) - I(S) = -\log P(S\cup\{t\}) + \log P(S) = -\log \frac{P(S\cup\{t\})}{P(S)}$

$= -\log P(t|S) = I(t|S)$. ■

As a side effect, we have proven the following result:

Corollary:

$J_T(t) \geq 0$ for all $t \in T$.

Proof:

In formula (6), we see that it is only necessary to prove $I(t|S) \geq 0$, but this is clear because $0 \leq P(t|S) \leq 1$. ■

With this Corollary and formula (4), we have proven that the information density $J_T(t)$ has the two properties we wanted it to have in Section 3.2, and with formula (6), we are able to express the information density without the help of the auxiliary coupling terms D(S).

In the development of formula (6), no use of a structure on the index set T is assumed. The index set T is nothing but a set.

We now give some remarks on the interpretation of formula (6). The information density is not the conditional information content of point t given the rest of the points, but is a weighted sum of the conditional information content of point t given all subsets of points outside t. One can define the average over all subsets to a given cardinality k,

$$E_k^T(t) = \frac{1}{\binom{|T| - 1}{k}} \sum_{\substack{S \subset T \setminus \{t\} \\ |S| = k}} I(t|S)$$

and the information density $J_T(t)$ is the average over all $|T|$ cardinalities from 0 to $|T| - 1$,

$$J_T(t) = \frac{1}{|T|} \sum_{k=0}^{|T|-1} E_k^T(t),$$

where not each subset, but each cardinality has the same weight. The information density is not I(t) as in the independent case, and not $I(t|T\setminus\{t\})$ as one may have expected, but both values occur in the weighted sum over the $E_k^T(t)$.

4. CONCLUSIONS

We have defined the information density of a finite set of random variables. The definitions are general, but we had especially applications in image processing in mind. As expected, the information density depends on the

stochastic model used to describe an image. The important remaining problem is to find better models to describe the stochastic dependences of an image.

The information density satisfies the two requirements desired, that it is never negative and that the global information of the set of random variables is the sum of the information densities of all points. The definition we gave does not assume any structure as an ordering relation or a distance function.

The intent of this report is mainly to define the concepts and to solve the general mathematical problems. The work shown here can be the basis for further efforts in applying the information density to several image-processing tasks. In particular, an edge-detection algorithm can be developed based on this concept. The contours can be found by a threshold approach or by a grey-tone representation of the information density of the image.

APPENDIX A

REASONS FOR STOCHASTIC IMAGE PROCESSING

There are three main reasons why an image should be modeled as a random function. The first reason is the modeling of noise which is always present in an image. Secondly, an image is not sensitive in some wrong points. And last but not least, the dependences (or correlations) between different points of an image are enormous. (The data-explosion problem demands good compression algorithms which reduce these redundancies.)

A.1. Noise in images

There is only very little knowledge on the deterministic side of noise. With the known and measured parameters, noise can be estimated as a random function. An important impulse to the theory of random functions (random processes) has been the need to develop a model of noise and noise-like physical systems such as Gaussian white noise and Brownian motion. We can conclude that, because of the presence of noise in all images, an image should be modeled as a random function if noise is significant. In deterministically modeled images, there is no control over noise. Image restoration of images distorted by noise is based on the stochastic image model (Wiener filtering).

A.2. Patterns in an image

In an English-language sentence, each "bit" may be important, as one can see from the following two sentences: "I like money" and "I like honey". Images are much less sensitive. Very different bit patterns may show the same image for the human observer. For an image, printed with some hundred pels per

inch and seen with normal observation distance, not the effective local bit pattern is important, but the more global distribution of black points because of the finite aperture of our eyes. This effect is demonstrated in Figs. A1 and A2. With the same observation distance but with bigger print dots, the difference can be seen.

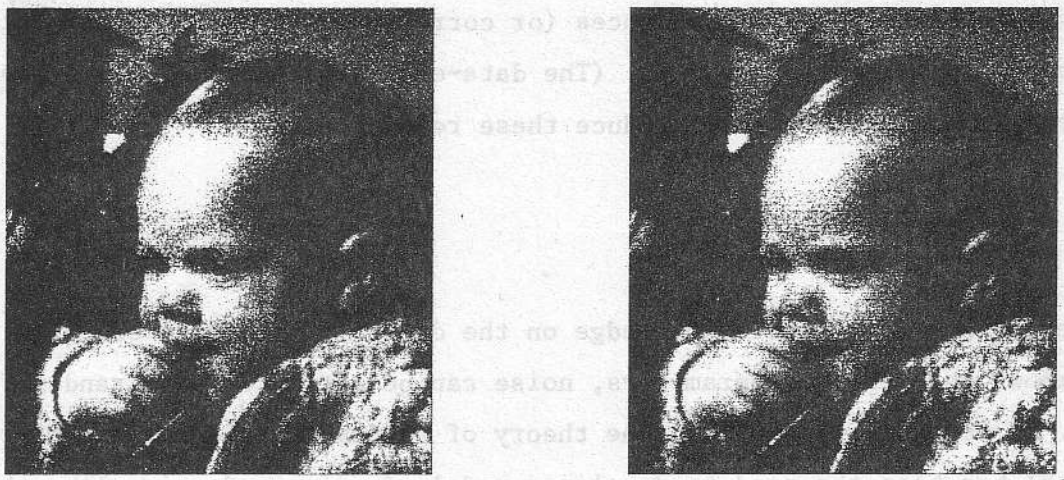
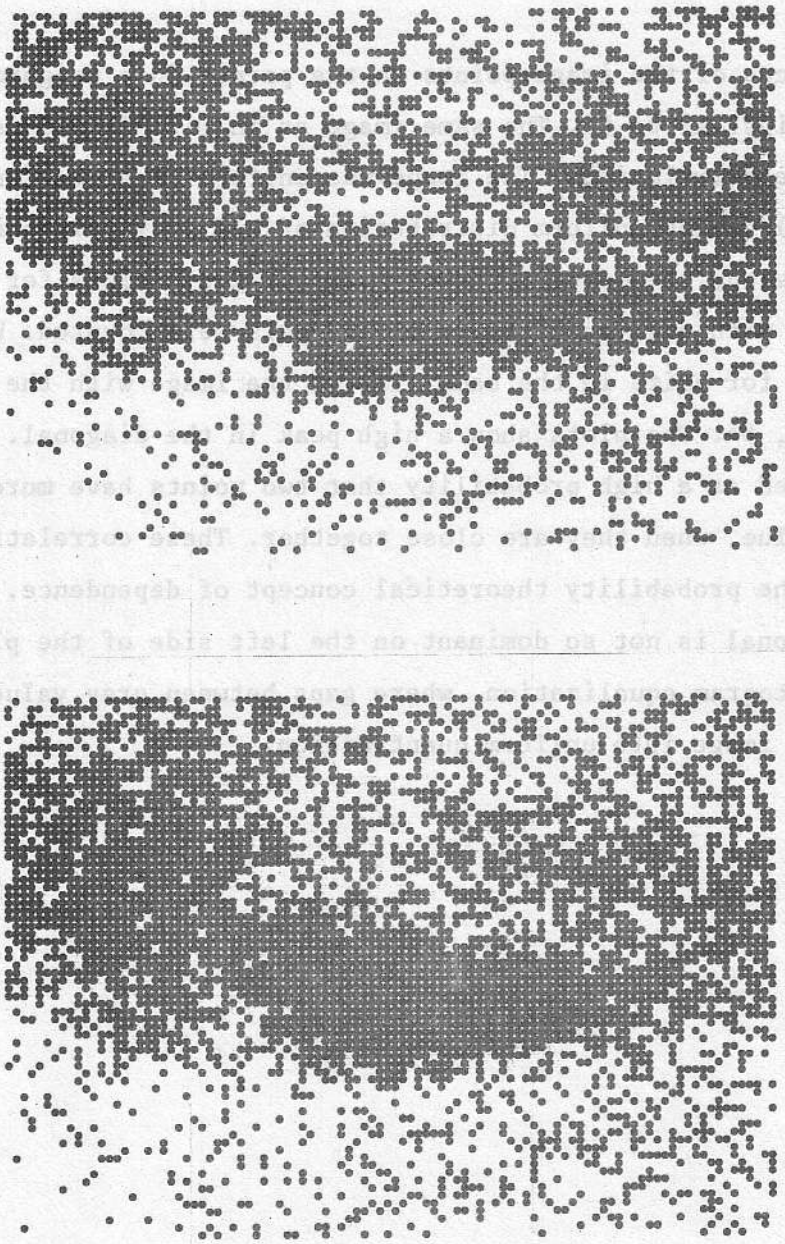


Figure A1. Two different bit patterns showing the same image.

The property of being the same image depends on the receiver of the image. Seen with a microscope, the two images are different. Also for a program working on pel level, the two images are not the same, either.

The effect shown above can be modeled in terms of random functions. Not the luminance of one image point, but the integrated value over some neighborhood is important for the eye of the human observer.



a)

b)

Figure A2. Enlarged part of the images in Fig. A1: part with the eye scaled by a factor 16. a) corresponds to the image on the left in Fig. A1, b) to that on the right.

A.3. Image point correlations

The dependences of the image points on the points in a neighborhood are plotted in Figs. A3 to A6. The same image as in Fig. A1 was used to determine the correlations. The image was uniformly requantized (histogram equalization) to get uniform distribution of all grey values in the image. In the next step, the occurrences of pairs of grey values for all pairs of image points with a given distance of d pels were counted. We plotted the correlations for pairs in the same line of the image with the distances d = 1, 4, 16, 64. The plots show a high peak in the diagonal. This can be interpreted as a high probability that two points have more or less the same grey value, when they are close together. These correlations can be modeled by the probability theoretical concept of dependence. The reason why the diagonal is not so dominant on the left side of the plots comes from the histogram equalization, where gaps between grey values are created to force the image into uniform quantization.

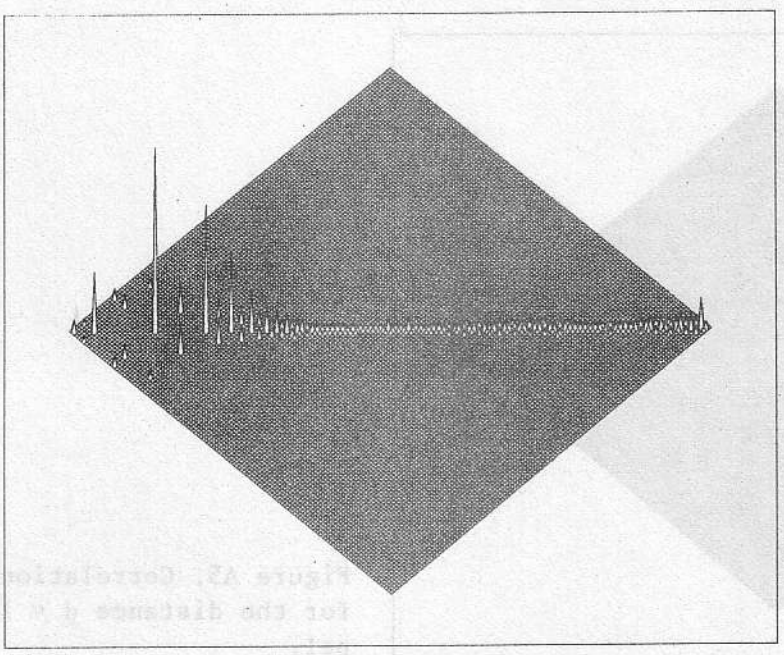


Figure A3. Correlations for the distance d = 1 pel.

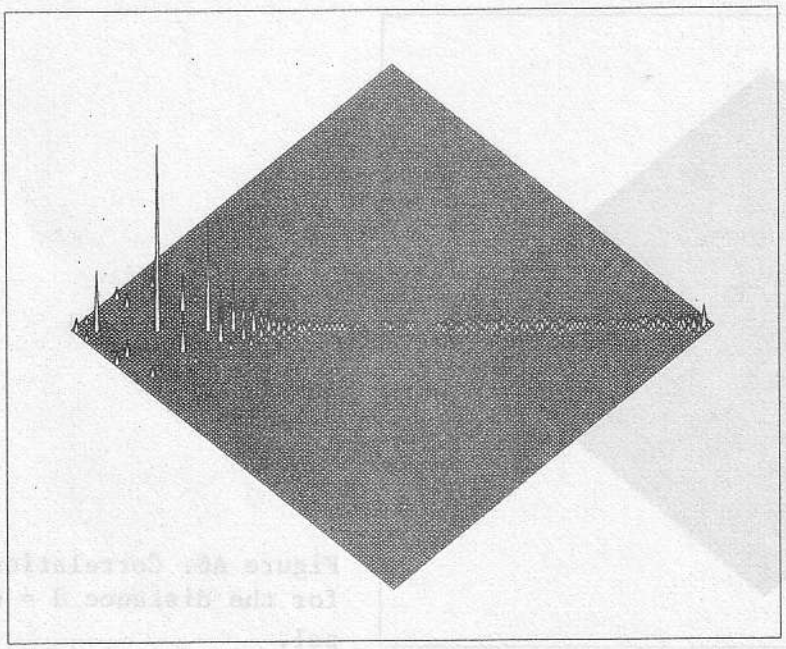


Figure A4. Correlations for the distance d = 4 pel.

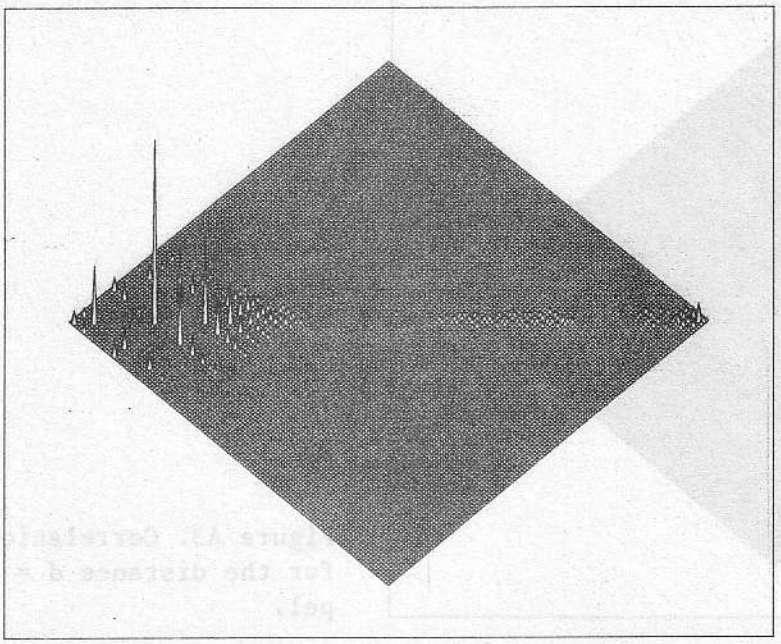


Figure A5. Correlations for the distance d = 16 pel.

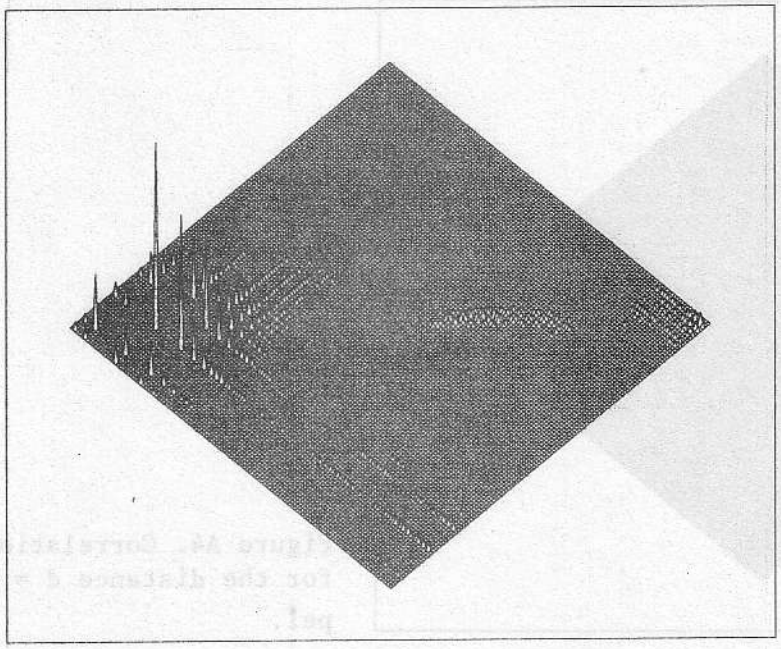


Figure A6. Correlations for the distance d = 64 pel.

APPENDIX B

SOME REMARKS TO THE DEFINITION OF INFORMATION

We have discussed random functions with some of their image-processing applications. In order to speak about information, two systems are necessary which communicate in some way. The information content is a measure for the content of a message. It is usually measured in bits. We are interested here in the information content of an image. The message in this case is a discrete quantized image, the sender is not important, and the receiver is the human observer or a computer program processing this image.

There are two approaches to measure the information content of an image. One can directly count the number of bits necessary to store or generate the given sample of the image. But one can also define the information content of the sample via the probability distribution of the image as a random function. (With optimal compression, both approaches result in the same value.)

Finally, we are interested in how it is possible to localize the information content of a random function in general. The information content seems to be a property of a sample with respect to a random function. We briefly discuss the assumptions in the definition of information density given in Section 3.

B.1. Information as counted bits

An image of the size $N_x \times N_y$ points with N_G grey values can be stored in $N_x \cdot N_y \cdot \log N_G$ bits. To take this value as the information content is not a good solution, because it does not depend on the content of the image. An image stored in this form contains many redundancies that can be removed by compressing the image. The information content of an image is then defined

as the number of bits necessary to store the compressed image. This way leads to Shannon's information theory.

A necessary remark must be made. For each specific image, it is possible to give an algorithm in which this image is compressed to one bit. The whole image is then stored in the decompression algorithm so that the algorithm only has to know whether or not it is the specific image. This is a message possible in one bit. From this point of view, a definition of the information content of an image which uses the concept of all possible compression algorithms has no sense.

As we have described, a computer program can generate an image. This idea can be used to define the information content of an image. The information content of a deterministically modeled image is the length of the shortest program which produces this image without any input [7].

B.2. Information from probability theory

If the message is an English-language sentence, the frequencies of the letters of the alphabet are known. But it is impossible to count the occurrences of all images. One idea is to count the subimages of a certain size and form. With this, we are in the midst of finding the stochastic description for the class of all images. We see that it is a problem to model the stochastic behavior of an image. Now let us imagine that we have a stochastic model for the class of images in which we are interested.

The definition of information in this section depends on the probability distribution of the model given. One can only speak about information with respect to a stochastic model. In the last section, we showed how to

compress a specific image to one bit. This compression algorithm is only reasonable if the image is frequently received. The probability of occurrence of this image must be around 0.5. In image processing, the stochastic model depends on the special problems to be solved therewith. If the class of images contains mainly scanned text pages of a book, the statistics have to reflect this knowledge. Information is, so to speak, only defined with respect to statistics in the same way as the optimal compression [8]. Information is nothing more than the negative logarithm of the probability, and the main task is the definition of useful stochastic models for important classes of images. Here, good solutions are not yet available.

B.3. Local information

Information on a set of random variables is not a local property because of the various dependences. From the mathematical side of information theory, some concepts such as conditional information, information gain, and relative information are known [6]. These concepts are closely related. The relative information is the negative information gain from the independent to the dependent cases of two random variables.

The generalization of the concept of relative information and the property of information to be additive led us to the definition of formula (2). The different coupling terms D(S') are treated as independent parts of the information I(S), and independent information can be added. This remark is not a proof of formula (2), but it is a clue to the interpretation of the terms D(S').

In the definition of relative information, an interpretation can be given ([6], p. 451) that the relative information is the amount of information about one random variable contained in the other. Because of the symmetry

of this definition, the relative information can be distributed to both random variables in equal parts. From here, we stated formula (3). We interpreted the terms D(S') as the symmetrically distributed information belonging to each point in S'.

With the two assumptions stated in formulas (2) and (3), we were able to develop formula (6) for the information density. This formula must now be applied to special classes of finite random functions. In thermodynamics, the stationary case is important and was studied with much effort [9], but this case is less useful in image processing.

Formula (2) is always correct, but it is of special interest when the terms D(S') for higher cardinalities |S'| are only very small values. Then the I(S) terms can be calculated more easily.

REFERENCES

[1] W. K. Pratt, "Digital Image Processing", John Wiley & Sons, New York, 1978.

[2] B. Julesz, E. N. Gilbert, and L. A. Shepp, Inability of Humans to Discriminate between Visual Textures that Agree in Second-Order Statistics — Revisited, Perception, 2, 1973, 391-405.

[3] M. Brady, Computational Approaches to Image Understanding, ACM Computing Surveys, 14, 1, March 1982, 3-71.

[4] L. Kitchen and A. Rosenfeld, Edge Evaluation Using Local Edge Coherence, IEEE Trans. Systems, Man, and Cybernetics, SMC-11, 9, September 1981, 597-605.

[5] K. Fukunaga, "Introduction to Statistical Pattern Recognition", Academic Press, New York, 1972.

[6] A. Rényi, "Wahrscheinlichkeitsrechnung mit einem Anhang über Informationstheorie", VEB Deutscher Verlag der Wissenschaften, Berlin, 1962.

[7] G. J. Chaitin, Algorithmic Information Theory, IBM J. Res. Develop. 21, 1977, 350-359.

[8] G. G. Langdon and J. J. Rissanen, Compression of Black-White Images with Arithmetic Coding, IEEE Trans. Commun., COM-29, 6, June 1981, 858-867.

[9] H. Föllmer, On Entropy and Information Gain in Random Fields, Z. Wahrscheinlichkeitstheorie Verw. Geb., 26, 1973, 207-217.