A Statistical Approach to Feature Detection and Scale Selection in Images

Dissertation

zur Erlangung des wirtschaftswissenschaftlichen Doktorgrades des Fachbereichs Wirtschaftswissenschaften der Universität Göttingen

vorgelegt von

Peter Majer
aus Frankfurt a.M.

Göttingen, Mai 2000

Erstgutachter: Professor Dr. W. Zucchini (Göttingen)
Zweitgutachter: Professor Dr. T. Lindeberg (Stockholm)

Tag der mündlichen Prüfung: 7.7.2000

Contents

Preface

The first impulse that led to this thesis came up at the Wednesday Workshop of the Institute for Statistics and Econometrics at the University of Göttingen.

At one such workshop Dr. Mauvis Gore from the German Primate Center (Deutsches Primatenzentrum GmbH) showed some ultrasound images of human ovaries which she examined routinely in order to study the development of individual follicles. These are extremely difficult to identify on the images and it requires considerable time and expertise to do so. Her question was whether this task could be done automatically.

The prospect of working on a subject that on one hand sees an increasing number of applications in areas as widely spread as medicine (x-ray, ultrasound, and magnetic resonance), biology (microscopy), geography (areal and satellite images), robotics, automatization, and many others, and on the other hand attempts to understand vision, our most useful sense, appeared very interesting.

Initially I chose Markov random field models to segment the images into ovary and follicles. This was inspired by my background in statistical physics and the popularity of Markov random field models in statistics. Unfortunately I was soon convinced that Markov random field models are not the first choice for the first steps of vision for two reasons: i) They require prior knowledge about the number of "colors" into which an image should be segmented. ii) They are not local, so the results for a given point can depend on the image intensity at positions that are far from the point.

After the scale-space conference in Utrecht in 1997 I decided to change the subject. The concept of "scale" was exactly what I was missing in the Markov random field models. Now the term occurs at 591 places in this thesis.

Chapter 1

Introduction

This thesis addresses the problem of extracting *useful information* from *images* of the *physical world*. The emphasis is on "useful", pertaining to some *task* that one aims to achieve.

Images of the physical world are used for a bewildering variety of tasks. "A pigeon uses vision to help it navigate, fly, and seek out food. Many types of jumping spider use vision to tell the difference between a potential meal and a potential mate. ... The rabbit retina is full of special gadgets, including what is apparently a hawk detector." [Marr, 1982, p. 32] There are many technical applications as well, controlling robot movements or aiding diagnosis and surgery in medicine. All these tasks are most certainly solved in different ways. Any particular solution may turn out to be useful or not in retrospect, when it is applied. So how can one go about *constructing* a *useful* solution?

A very interesting possibility is to look at existing biological visual systems. Understanding biological vision would be interesting in itself and one may hope to learn some tricks for the construction of artificial visual systems. This approach was pioneered in the 1950s and 60s by Barlow [Barlow, 1953], Hubel and Wiesel [Hubel and Wiesel, 1962], [Hubel and Wiesel, 1968], and many others.

The alternative approach of *computer vision* attempts to build a visual system from scratch. This approach focuses on the task to be solved and in principle admits any method to construct a solution as long as the task is solved. At the same time it raises the question whether the task alone provides any guidelines to its solution and if so, what these guidelines are.

A few requirements about the final solution should be dealt with at *all* levels of the construction. These concern some minimal requirements on what type of information must *not* be discarded, formulated as "invariance requirements". If, for example, a rabbit needs to be able to detect a hawk coming from any direction, then all steps of the processing must be able to deal with all possible directions. Discarding information about hawks coming from behind would evidently not be

a useful strategy.

Next the question arises whether to construct a solution to any specific task in one piece or to divide it into several steps some of which may also be useful to other tasks. The general consensus on this is that *some basic steps of processing are useful for very many different tasks*. These basic steps of visual information processing are called *early vision* or *low level vision*, the terminology emphasizing the claim to generality. The following section intends to give the reader a rough idea of what is generally believed to be a set of useful first steps of visual information processing.

1.1 A Feature Detection Recipe

A detailed description of the first steps of image analysis will be given in the following two chapters. To give the reader a rough idea we sketch them as a recipe in three steps:

1. *Smooth the Image*
   An observed image is smoothed. In general several degrees of smoothing should be performed. Figure (1.1) shows a magnetic resonance image of a brain and some smoothed versions of the same image.

   Sometimes the appropriate degree of smoothing is known beforehand due to the setup, e.g. in an industrial application where distance of camera and object are fixed and the observed objects are very similar each time.

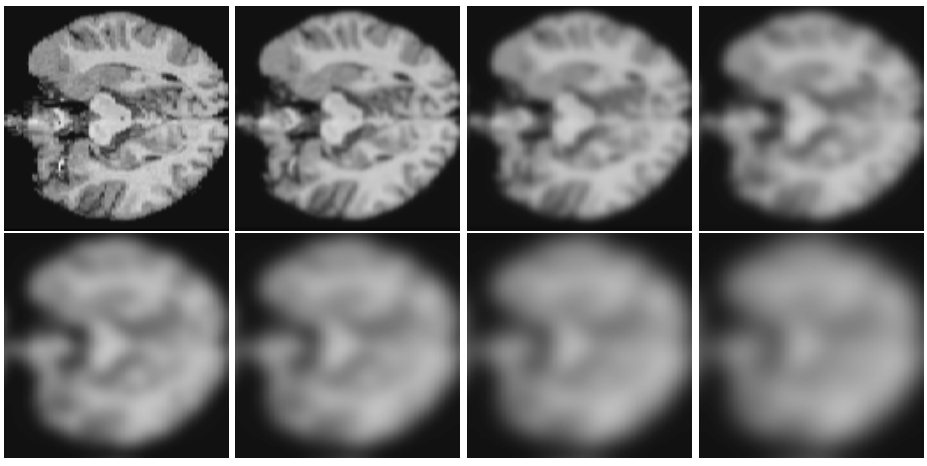


Figure 1.1: Original and smoothed images.

2. *Choose a Feature Detector*
   Local structural properties of the smoothed images are computed. Examples of such properties are the gradient or the principal curvature. They should not depend on any parameters that might require "user-interaction" or an "intelligent guess" of the programmer. Figure (1.2) shows the gradient of the smoothed images. Figure (1.3) shows the principle curvature of the smoothed images.

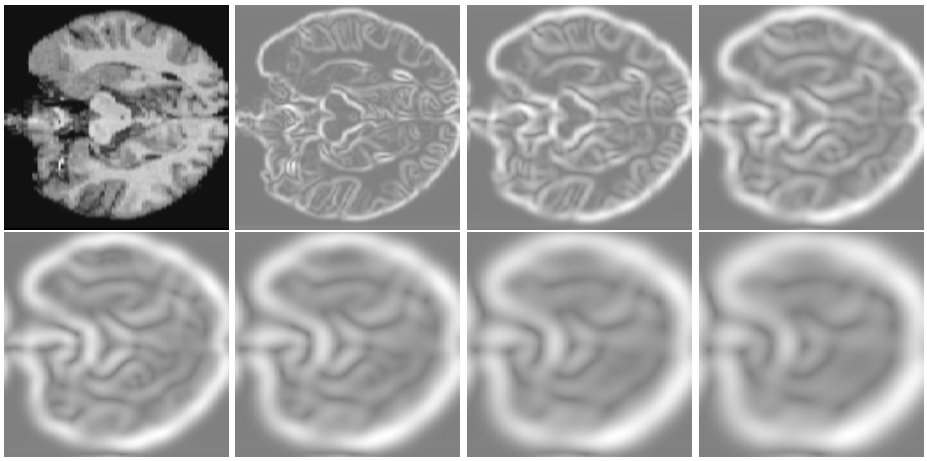


Figure 1.2: Original image and gradient of smoothed images. (For better display the grey values have been adjusted independently in each image.)

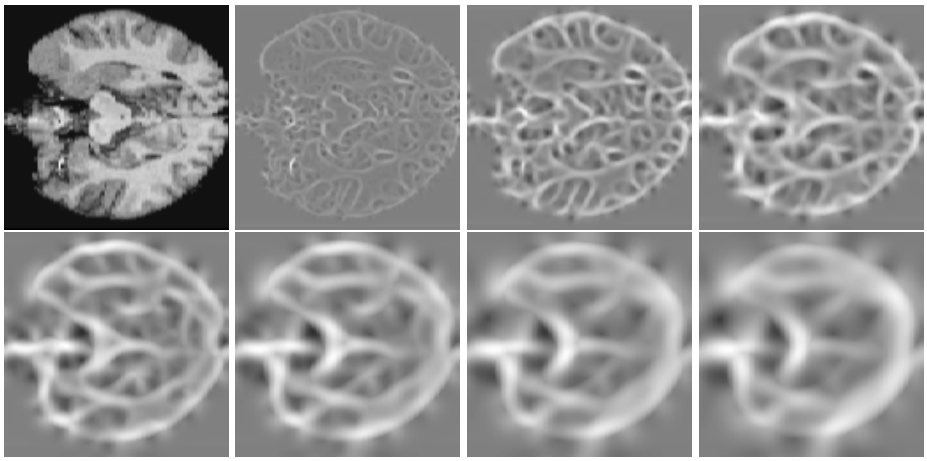


Figure 1.3: Original image and principle curvature of smoothed images.

3. *Compute Local Extrema of a Feature Detectors Response*
   The local extrema of the structural properties are considered "particularly informative" positions. Figure (1.4) shows "edges" of the brain image at different degrees of smoothing. Edges are local maxima of the gradient along the gradient direction. Figure (1.5) shows the "ridges" of the same image. Ridges are a subset of the local minima of the principal curvature along the direction of principal curvature (see Chapter 5).

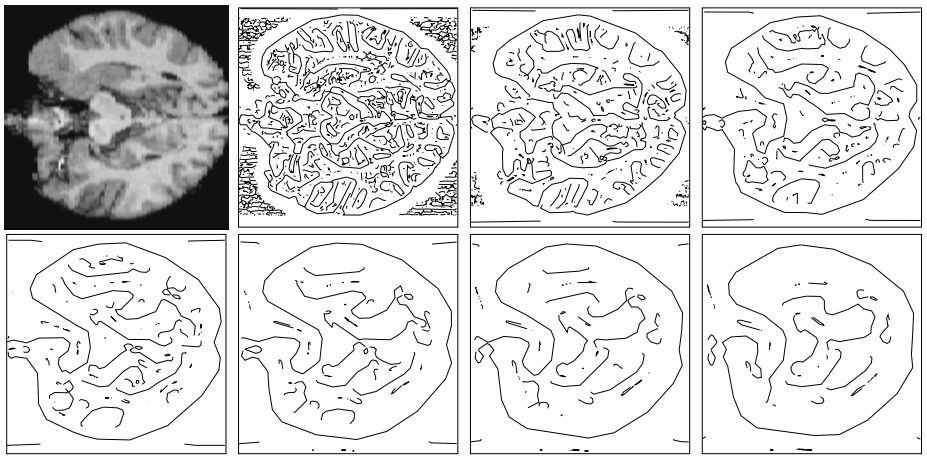


Figure 1.4: Original image and maxima of gradient of smoothed images along gradient direction.

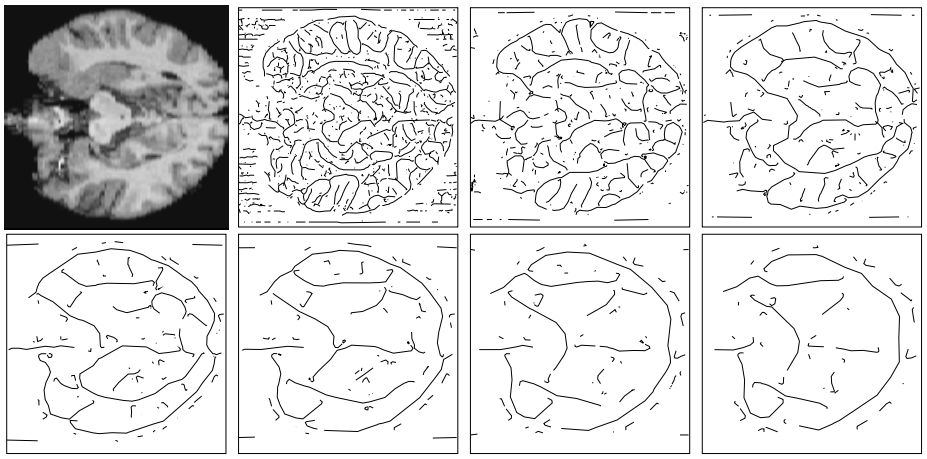


Figure 1.5: Original image and "second derivative ridges" of smoothed images.

Based on the fact that the computed "edges", "ridges" and perhaps other features capture some essential structural information about an image which is closer to a content-based description than the original pixel-based representation one hopes that these features should suffice to solve many tasks of vision.

1.2 Some Questions about Scale

The above recipe involves choosing several degrees of smoothing. Why that? Is there not a single appropriate degree of smoothing? If one follows the "scale-space concept" introduced by Koenderink [Koenderink, 1984] the answer to this is "No, usually not".

The idea of the "scale-space concept" is that the degree of smoothing can reveal the size of "objects" within an image as follows: With increasing degree of smoothing objects vanish from the image, small objects first and larger objects later. The degree of smoothing at which an object vanishes basically measures the size of the object. For this reason the smoothing parameter is also called "scale".

According to the scale-space concept any object within an image has a position and a scale. To find both positions and scales it is evident that an image must be smoothed to all possible degrees (unless the content is known beforehand). As Koenderink wrote in 1984,

> The challenge is to understand the image really on all these levels (scales) *simultaneously*, and not as an unrelated set of derived images at different levels of blurring (smoothing).

Ironically the question of how to determine both positions and scales has eluded scale-space theory for almost a decade. The first systematic integration of "position detection" and "scale selection" was proposed in 1993 by Lindeberg [Lindeberg, 1993b]. The proposal is in many respects similar to the above recipe. The image is smoothed to different degrees, some operators are applied, and then "particularly informative" positions and scales are computed as local extrema of the operator response with respect to position and scale.

Lindeberg's proposal for scale-selection contains a so-called γ-normalization parameter. Different choices of this parameter yield different "particularly informative" scales. It remains a question, what the "right" choice of γ-normalization should be. More generally it is not clear why scales should be selected according to the prescription given by Lindeberg. Both questions are addressed by this thesis.

1.3 Contributions of the Thesis

The study of scale-selection and the approach adopted has affected the author's point of view of scale-space and feature detection in general. For this reason the original contributions of this thesis are not presented as a separate part but integrated into the overall presentation. In the following we give a summary of the original contributions with references to their location in the thesis.

A central idea of the thesis is *stochastic simplification*. This proposes to randomly shuffle the pixels of an image to new positions.

Stochastic simplification is introduced in chapter 2. In section 2.4 we prove that a very natural condition on shuffling produces random images whose expected value is exactly linear scale-space.

Section 2.4 also introduces a *local entropy* defined for any single point in scale-space. We prove that the sum of local entropies over all points of an image increases monotonically with scale. This captures in a mathematically rigorous way the intuitive idea that smoothing (by Gaussian filter kernels) simplifies images both globally and, more importantly, also locally.

In chapter 6 the idea of shuffling is applied to feature detection and scale selection. This chapter proposes to make use of the *local distributions* that shuffling generates at each point in scale-space. These distributions allow us to take a point of view from which feature detection and scale selection appear as special cases of one and the same concept. As a consequence there is a canonical scale-selection operator to any feature detection operator.

Chapters 5 and 7 apply the theoretical concepts to the problem of ridge detection. They contain some original contributions throughout. In particular section 7.4 describes an interesting phenomenon of a second derivative ridge-detector. At fixed scales this operator frequently responds to edges. At variable scales, however, the "correct" choice of scale-selection allows the operator to "escape" from edges along the scale direction.

Section 3 of chapter 8 describes a modification of the well-known marching squares/cubes algorithm that are necessary to compute ridges.

Finally chapter 9 discusses a self-similarity property of normal noise in scale-space. The contribution here is that this property facilitates the computation or estimation of distributions of some "measurements" made on normal noise in scale-space. Such distributions could be useful e.g. to assess the significance or saliency of features.

1.4 Some Remarks Concerning the Formulation

There are several possibilities to formulate and compute the scale-space of an image, either in terms of *integral equations* or *partial differential equations*, in the space domain or in the *frequency domain*, in a continuous or a discrete formulation. As far as the presentation is concerned we have chosen a continuous integral formulation with filter kernels in the spatial domain. The software [1] that the author wrote to implement the theory makes use of a discrete "integral" formulation in the frequency domain. The formulation in terms of partial differential equations is extensively used in the literature on *non-linear scale-spaces* [Weickert, 1998]. A genuinely discrete formulation can be found in [Lindeberg, 1990].

[1] Algorithms for smoothing and computation of derivatives via Fourier transformation in cartesian, gradient and curvature coordinates, as well as algorithms for the computation of zero-crossings in 2 and 3 dimensions, and the computation of ridges without and with scale selection were written based on the free Vista library [Pope and Lowe, 1994] from the University of British Columbia.

Chapter 2

Scale-Space

This chapter introduces the *scale-space representation* of image data that replaces an image by a family of smoothed versions of the same image.

The *scale-space representation* has proved useful to the task of *vision* because with increasing blur details of the original image are lost. This allows a visual system to "concentrate" on the appropriate level of detail and to relate "things" across different levels of detail.

The chapter is organized as follows. The definition of linear scale-space and some examples are given first. Next the purpose of the representation, to serve as a useful starting point for *vision*, is briefly discussed. A number of properties of scale-space that appear particularly useful concerning vision are discussed in section (2.3). Finally the intuitively evident fact that smoothed images are simplified versions of the original image is considered in detail from a stochastic point of view. It is shown that random shuffling of pixels to new positions can create scale-space, and that the *average local entropy* of this process increases monotonically with scale. The latter is a mathematically rigorous formulation of the simplification property of scale-space.

2.1 Linear Scale-Space

Linear scale-space is a *representation* of data that makes explicit some information that is otherwise only implicitly present in the data, namely *scale*. As a representation for vision it was independently proposed by [Iijima, 1959] and [Witkin, 1983].

The *linear scale-space* of $f : \mathbb{R}^N \to \mathbb{R}$ is defined as $L : \mathbb{R}^N \times \mathbb{R}_+ \to \mathbb{R}$ with $L(\cdot, 0) \equiv f$ and for $t > 0$

$$L(\cdot, t) \equiv G(\cdot; t) * f \tag{2.1}$$

where

$$G(\mathbf{x};t) = \frac{e^{-\frac{\mathbf{x}^T\mathbf{x}}{2t}}}{(2\pi t)^{N/2}}$$

is the (rotation symmetric) Gaussian filter kernel of width $\sqrt{t}$ and $*$ denotes the convolution operator [1]. $\sqrt{t}$ is called the *scale*.

Figure (2.1) shows some examples of "slices" from the scale-space of some two-dimensional images. They illustrate how with increasing scale small scale information is lost.

2.2 The Purpose of Scale-Space: Vision

Any representation of data is useful, or not, only together with some information processing task. The scale-space representation is designed for *vision* which Marr characterizes as follows [Marr, 1982, p. 31]:

> Vision is a process that produces from images of the external world a description that is useful to the viewer and not cluttered with irrelevant information.

The sheer amount of data makes the distinction between relevant and irrelevant a primary concern to a visual system. Experimental evidence from the human visual system provides an impressive example [Atick, 1992]: The retina collects data at a rate of more than 10^6 bits/sec [Jacobson, 1951]. Most of these are discarded before arriving at the visual pathway. Studies of the speed of visual perception [Sziklai, 1956] or reading [Kornhuber, 1973] show that the visual pathway in humans transmits around 50 bits/sec.

Considering the task of compressing the information contained in an image scale-space may seem a step in the wrong direction. It obviously requires much *more* storage than the observed data alone which are themselves only the first, $t = 0$, slice of scale-space. It appears, however, that the scale-space representation is better suited for *subsequent detection of relevant information* than the original image representation.

The basic idea is to *describe each "object" at the appropriate scale*. For example it would certainly be inappropriate to describe a tree top on a molecular scale of 10^{-6} meters. Of course a scale of 1000 meters is not a better choice. An efficient description may be achieved on scales around 1 meter. It is self-evident that much information is discarded when replacing a micrometer description of

[1] The convolution of $G(\cdot;t)$ and f is defined as $(G(\cdot;t) * f)(\mathbf{x}_0) \equiv \int d\mathbf{x}\, G(\mathbf{x}-\mathbf{x}_0) f(\mathbf{x})$. See e.g. the chapter on fast Fourier transforms in [Press et al., 1988].

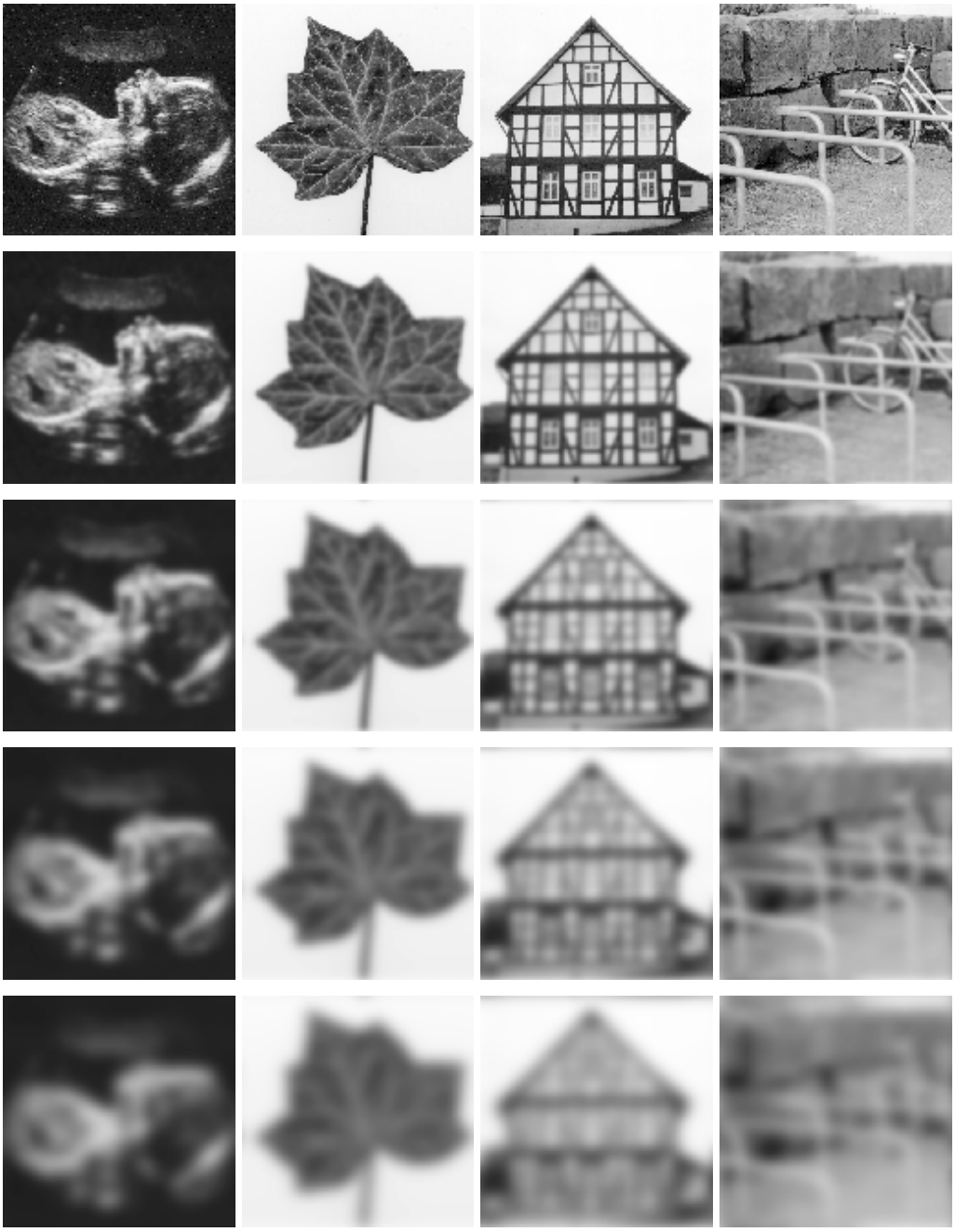


Figure 2.1: Slices from scale-space. All images have 512 by 512 pixels. The displayed scales are $\sqrt{t} = 0$, $\sqrt{t} = 4$, $\sqrt{t} = 8$, $\sqrt{t} = 12$, and $\sqrt{t} = 16$ (where a unit length is the width/height of a pixel).

the tree top by a description on a 1 meter scale. However, *for the purpose of describing the tree top* the gain outweighs the loss.

To find an appropriate description without prior information about the image content it is necessary to study an image at *all* scales as sketched in figure (2.2), the scale-space representation being the natural starting point. Subsequent steps to analyze the image content and find appropriate scales may be sketched as follows: A toolbox of operators, each of which focuses on some different aspect, is used to "look at" the scale-space. The resulting data are then searched for (a small set of) *particularly informative features* across space and scale. These features provide a condensed description of the original image where each feature is associated with its appropriate position and scale.

How to achieve these later steps of the scale-space paradigm will be the subject of subsequent chapters.

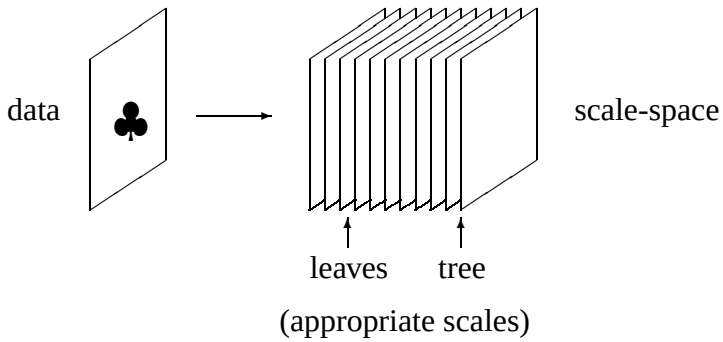


Figure 2.2: The scale-space representation contains appropriate scales for all "objects".

2.3 Useful Properties of Scale-Space

The remainder of this chapter is devoted to some interesting properties of the scale-space representation. These properties give an idea of why the scale-space representation could be useful to vision. They go hand in hand with the question of *what abilities a visual system should possess in order to perceive the physical world around it*. Though we have attempted to present the ideas rather than the technical details, the discussion does become technical at some points. The reader who is more interested in how the first steps of vision might be achieved or implemented should continue with the next chapter.

2.3.1 Simplification

It is apparent from the above examples that with increasing scale detail is lost. From the original data at scale $t = 0$ the slices of scale-space make a transition to constant intensity at infinite scale. Clearly this transition corresponds to a gradual simplification of the image content. Intuitively such a gradual simplification appears a useful property of the scale-space representation because it allows the level of detail to be chosen appropriate to the image content. This has inspired several authors to define simplification in a strict mathematical rather than intuitive sense and regard it as a *necessary* property of a representation of image data for vision. Some of these definitions shall be discussed in the following.

Non-Creation of Local Extrema in One Dimension

Witkin [Witkin, 1983] was first to formulate a *simplification* property of one-dimensional scale-space. He defined this to mean the non-creation of local extrema, i.e. going from small to large scales no new local extrema along space may appear. To exemplify this figure (2.3) shows the scale-space of a one-dimensional image together with the locations of local extrema along space. One can see clearly that local extrema are able to annihilate each other but no new local extrema appear toward larger scales. Babaud et al.[Babaud et al., 1986] showed that linear scale-space is the unique representation with this property.

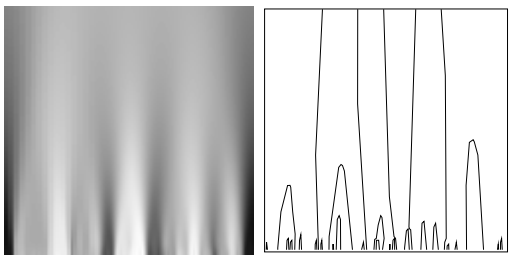


Figure 2.3: One dimensional scale-space and zero-crossings of the first derivative along space.

Non-Enhancement of Local Extrema

In two or more dimensions the simplification property of scale-space must be characterized somewhat differently since here it *is* possible that new local extrema appear with increasing scale.

A view of simplification that applies in any dimension is that all local maxima should decrease with increasing scale and conversely all local minima should increase. This property may be observed in the example images. It is also easily

proved to hold in scale-space since the derivative along scale may be expressed as follows:

$$\partial_t L(\mathbf{x};t) = \frac{1}{2}\sum_i \partial_i \partial_i L(\mathbf{x};t)$$

At a local maximum each of the second derivatives on the right hand side is negative so that the derivative along scale is negative as well, which goes to show that a local maximum of $L(\mathbf{x};t)$ decreases with increasing scale.

Koenderink [Koenderink, 1984] formulated this simplification property, that he called *causality*, as follows. Consider iso-surfaces $L(\mathbf{x};t) = Constant$ in scale-space (e.g. if the image is two-dimensional $L(\mathbf{x};t) = Constant$ describes a two-dimensional surface in a three-dimensional scale-space). At positions on such a surface where $L(\mathbf{x};t)$ is an extremum with respect to $\mathbf{x}$ the surface should point its convex side toward increasing scales $\sqrt{t}$. This, he showed, is equivalent to the above equation and thus to linear scale-space if additionally differentiability, homogeneity and isotropy are demanded [2].

Stochastic Simplification

Still another point of view of simplification is the following. Suppose we randomly shuffle the pixels (intensities) in an observed image $f(\mathbf{x})$ to new positions. This should on average destroy structural information so that the average of the shuffled images is a simplification of the observed image. It remains to define how exactly to shuffle intensities.

To shuffle the pixels around, allow them to jump from position $\mathbf{x}$ at "time" t to position $\mathbf{y}$ at "time" $t+\tau$ with some *transition probability* $p(\mathbf{y},t+\tau|\mathbf{x},t)$ [3]. The position $\mathbf{x}$ of each pixel thus becomes a random variable and one can study how the distribution of pixels evolves with "time" t. To achieve a *gradual* simplification the "time" dependence is important. *For short times the typical length of a jump should be proportional to the time τ between t and $t+\tau$.* In terms of transition probabilities this is expressed as follows (indices running from 1 to N):

$$\begin{aligned}
\int dy_i (y_i - x_i) p(\mathbf{y},t+\tau|\mathbf{x},t) &= A_i\tau + o(\tau) \\
\int dy_i (y_i - x_i)(y_j - x_j) p(\mathbf{y},t+\tau|\mathbf{x},t) &= C_{ij}\tau + o(\tau) \\
\int dy_i (y_{i_1} - x_{i_1})\ldots(y_{i_n} - x_{i_n}) p(\mathbf{y},t+\tau|\mathbf{x},t) &= o(\tau) \qquad \text{for} \qquad n > 2
\end{aligned} \tag{2.2}$$

[2] Koenderink does not explicitly mention differentiability but makes use of it. Differentiability ensures continuity and that is certainly required to prohibit new local extrema from "popping up out of nowhere".

[3] The positions of pixels are shuffled according to a *Markov process*.

These equations for the transition probabilities define a *diffusion process* [Honerkamp, 1990] [Gardiner, 1985]. For short "times" τ the average jump displaces a pixel at $\mathbf{x}$ to $\mathbf{x}+\mathbf{A}\tau$ and the jumps typically deviate from this average by $C_{ij}\tau$. The *drift* A_i and the *diffusion tensor* C_{ij} can in principle be functions of position, time, or even the simplified image.

If one defines a simplification $L(\mathbf{x};t)$ of an image $f(\mathbf{x})$ to be the expected value of images shuffled in the described way then it can be shown that $L(\mathbf{x};t)$ satisfies the partial differential equation

$$\partial_t L(\mathbf{x};t) = \sum_i \partial_i A_i L(\mathbf{x};t) + \frac{1}{2}\sum_{i,j} \partial_i \partial_j C_{ij} L(\mathbf{x};t)$$

with initial condition $L(\mathbf{x};0) = f(\mathbf{x})$. The derivation of this equation will be given in the last section of this chapter. The equation is the generating equation of scale-spaces in general, including the nonlinear scale-spaces where both diffusion coefficient C_{ij} and drift A_i may depend on the local intensity $L(\mathbf{x};t)$ (see e.g. [Perona and Malik, 1990], [Alvarez et al., 1992] or [Weickert, 1998] for an overview).

Imposing isotropy and homogeneity makes $A_i = 0$ and $C_{ii}(\mathbf{x},t) = 1$, $C_{ij} = 0$ for $i \neq j$ so that again linear scale-space can be seen to be the unique solution.

A very interesting consequence of shuffling is that it allows one to define *local* entropies of the random intensity at position $\mathbf{x}$ and time t. The intuitive idea that shuffling simplifies images may then be associated with the fact that the average *local* entropy increases monotonically with time t. A proof hereof is given at the end of the chapter.

2.3.2 Translation and Rotation Invariance

Let us now consider two properties that are not only useful but practically indispensable to a visual system, unless prior information about the image content is available.

If an observer moves relative to a scene the physical content of the scene remains unchanged, of course. For a visual system that aims to "see" the physical scene it is therefore important that the information content of its description remains unchanged as well, apart from the fact that it "sees" the change of position. This is formulated in terms of *translation* and *rotation invariance* as follows: translation (rotation) of an image before computation of scale-space is identical to translation (rotation) after computation of scale-space. Schematically this is shown in figure (2.4).

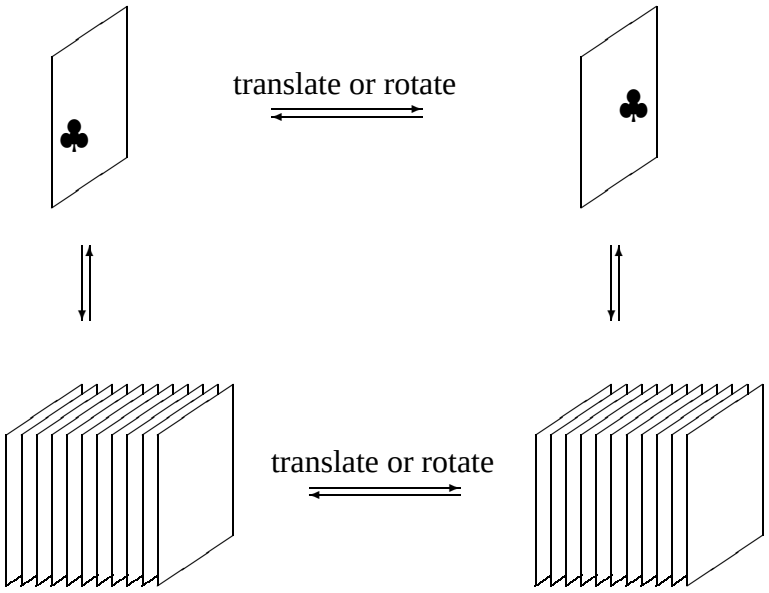


Figure 2.4: Commutative diagram of translation and rotation invariance.

Some restrictions must be made. Invariance with respect to all possible movements between observer and scene is generally not physically achievable due to a limited field of view as well as a limited resolution of the visual system and not least the projection of a three-dimensional scene to a two-dimensional image.

Consider, however, a special situation where these limiting factors do not apply. Let this page of paper be the scene and rotate it or move it left or right changing its distance to your eyes as little as possible. In this situation, too, the visual system should "see" the same information irrespective of the rotation or translation of the scene.

More generally invariance with respect to translations and rotations of the projection of a scene onto the image plain can be achieved (as long as the content is not moved out of the image domain). Technically this if formulated as follows: Call T the coordinate transformation $T(\mathbf{x}) = \mathbf{M}\mathbf{x} + \mathbf{a}$ for some vector $\mathbf{a} \in \mathbb{R}^N$ and some orthonormal $N \times N$ matrix $\mathbf{M}$ and denote by $f \circ T$ the concatenation of T and f, i.e $(f \circ T)(\mathbf{x}) = f(T(\mathbf{x}))$. Then one easily verifies that scale-space satisfies translation and rotation invariance in the following sense:

$$(G(\cdot;t) * (f \circ T))(\mathbf{x}) = ((G(\cdot;t) * f) \circ T)(\mathbf{x})$$

Here we have neglected the image border for convenience.

Concerning the example with the page of paper one remark is in place. Clearly our visual system *cannot* read the text on the page equally well from any orien-

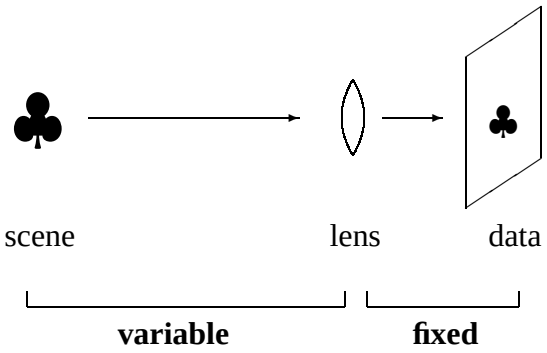
tation. This is a consequence of the fact that we always read text oriented in the same way, the "top" of the page facing up. Whenever such prior knowledge about the environment is available it is possible to increase the efficiency of information processing at the cost of the ability to deal with general situations. The approach pursued by scale-space theory is to attempt to understand and work out the more general methods, imposing translation and rotation invariance. The hope is that many applications can profit from even a small improvement by this approach.

Let us now consider changes of distance between observer and scene. These require special attention because they either enhance or destroy the details of the scene that are visible to the observer.

2.3.3 Observational Blur and Scaling

Any physically observed image is blurred by the measurement device or eye. This *observational blur* makes small scales unobservable and leads to loss of detail as the observer moves away from the scene.

Observational blur is a physically inevitable property of the measurement device or eye [4]. It is a result of the measurement itself, created for example by the lens and the photoreceptors. What is important in the present context is that *the amount of blur is fixed on the scale of the measurement device* as sketched in the following figure. This has been termed the *inner scale* of the measurement device [Florack et al., 1992] [Florack et al., 1994].



The effect of a variable distance between observer and scene is the following. With increasing distance the projections of the scene onto the image plain become smaller. Still all projections receive the same amount of blur on the scale of the measurement device. Conversely this means that *on the scale of the scene distant scenes are blurred more than close scenes*. In effect this is a physical possibility

[4] A measurement device can be optimized to *minimize* observational blur but it cannot be avoided altogether.

to construct a scale-space, which, of course, need not be the linear scale-space. Linear scale-space arises in this way only when the observational blur is Gaussian.

Suppose again the visual system aims to "see" the physical world. If the same physical scene is observed at different distances it would be useful to have some way of identifying the resulting images. One way to facilitate this is to artificially, by computation, subject an observed scene to the scaling and, more importantly, the extra observational blur that *would* result from a physically larger distance. The scale-space representation does just that, as far as observational blur is concerned. As shown in figure (2.5) a simple transformation allows one to match an observation at a large distance and "the same" observation at a shorter distance.

Suppose an image d of a distant scene differs from a closer image f of the same scene as follows: $d = G(\circ;t_o) * f^s$, where $f^s(\mathbf{x}) = f(s\mathbf{x})$ and $s > 1$. Then the scale-space of the distant scene is related to that of the close scene by:

$$(G(;t) * d)(\mathbf{x}) = (G(\cdot;s^2(t+t_o)) * f)(s\mathbf{x})$$

Evidently this equation would not hold if the observational blur was not Gaussian. To set up a similar equation in that case would require a non-Gaussian scale-space. The fact that the equation holds for Gaussian scale-space and Gaussian observational blur is due to the *recursivity principle* or *semi-group property* which states that a Gaussian filter kernel smoothed with a Gaussian filter kernel is again a Gaussian filter kernel.

2.3.4 Differentiability

A technically useful property of the scale-space representation is *differentiability*. $L(\mathbf{x};t) = (G(\cdot;t) * f)(\mathbf{x})$ can be differentiated up to any order by the relation

$$\partial_1^{n_1} \ldots \partial_N^{n_N} L(\mathbf{x};t) = ((\partial_1^{n_1} \ldots \partial_N^{n_N} G(\circ;t)) * f)(\mathbf{x})$$

This property is extensively used in the first steps of processing the scale-space representation as will become apparent in the subsequent chapters.

Particularly notable is that the above relation allows one to differentiate the scale-space of *discretely sampled data* points. While obviously it makes no sense to speak of differentiating discretely sampled data themselves, the equation

$$(G(\circ;t) * (\partial_1^{n_1} \ldots \partial_N^{n_N} f))(\mathbf{x}) = ((\partial_1^{n_1} \ldots \partial_N^{n_N} G(\circ;t)) * f)(\mathbf{x})$$

makes the meaning well-defined [5].

[5] In terms of *regularization theory* [Tikhonov and Arsenin, 1977] differentiation of discretely sampled data is an *ill-posed problem* [Hadamard, 1902] and scale-space is a *regularization* of this problem. For an introduction to regularization theory see [Goutte, 1997].

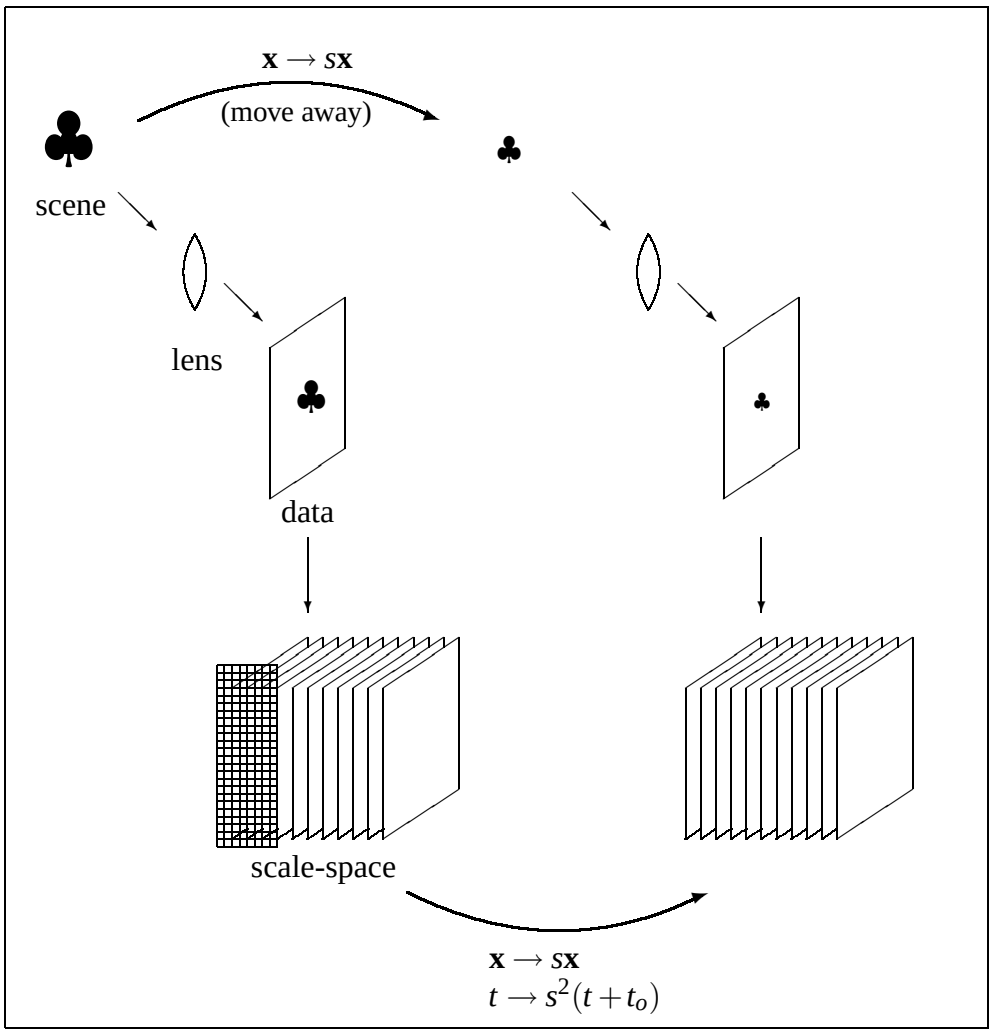


Figure 2.5: Scaling, observational blur, and scale-space. If an image is scaled in size by a factor s, i.e. $\mathbf{x} \rightarrow s\mathbf{x}$ the scale-space is transformed by $\mathbf{x} \rightarrow s\mathbf{x}$, $t \rightarrow s^2 t$. The grid masks that part of the scale-space from the close scene which is unobservable in the distant scene.

2.4 Stochastic Simplification and Scale-Space

This section considers in detail stochastic simplification and scale-space. The idea of stochastic simplification is that random shuffling of pixels should on average[6] destroy structural information of an image. It is thus not surprising that scale-space may be "derived" from shuffling as will be demonstrated.

The second point of this section concerns the intuitive property that scale-space simplifies images both globally and, on average, also locally. This is formulated in terms of local entropies of shuffled images and it is shown that the sum of local entropies increases monotonically with scale.

2.4.1 A Derivation of Scale-Space

An impressive number of approaches have appeared in the vision literature that derive linear scale-space from a number of basic axioms. [Iijima, 1959], [Iijima, 1962a], [Iijima, 1962b], [Iijima, 1963], [Otsu, 1981], [Koenderink, 1984], [Yuille and Poggio, 1986], [Babaud et al., 1986], [Lindeberg, 1990], [Florack et al., 1992], [Alvarez et al., 1993], [Lindeberg, 1994b], [Pauwels et al., 1995], [Nielsen et al., 1997], [Lindeberg, 1997], [Florack, 1997]. The arguments may roughly be divided into two categories. One is based on *simplification* and the other on the *recursivity principle*. A detailed overview is given in [Weickert et al., 1997].

Here it is demonstrated that stochastic simplification produces scale-space under some very natural assumptions about shuffling.

To define shuffling we allow each pixel to jump from its position $\mathbf{x}$ at "time" t to position $\mathbf{y}$ at "time" $t+\tau$ with some *transition probability* $p(\mathbf{y},t+\tau|\mathbf{x},t)$. Each pixel carries around with it the intensity of the observed image at its starting position and all pixels are allowed to jump independently of one another.

The actual condition we impose in order to achieve a *gradual* simplification is that *for short times* τ *between* t *and* $t+\tau$ *the typical length of a jump should be proportional to the time* τ. In terms of transition probabilities this is expressed by equations 2.2 which we repeat here:

$$\begin{aligned} \int dy_i (y_i - x_i) p(\mathbf{y},t+\tau|\mathbf{x},t) &= A_i\tau + o(\tau) \\ \int dy_i (y_i - x_i)(y_j - x_j) p(\mathbf{y},t+\tau|\mathbf{x},t) &= C_{ij}\tau + o(\tau) \\ \int dy_i (y_{i_1} - x_{i_1})\ldots(y_{i_n} - x_{i_n}) p(\mathbf{y},t+\tau|\mathbf{x},t) &= o(\tau) \qquad \text{for} \qquad n > 2 \end{aligned} \tag{2.3}$$

[6] Average with respect to repeated shuffling

The *drift* A_i and the *diffusion tensor* C_{ij} can in principle be functions of position, time, or even the simplified image. In any case, these conditions allow one to derive a partial differential equation for the transition probabilities as follows [Honerkamp, 1990]: Take some function $R(\mathbf{y})$ with vanishing first derivative at the boundary of the image domain. Then

$$\begin{aligned}
&\int d\mathbf{y} R(\mathbf{y}) \partial_t p(\mathbf{y},t|\mathbf{x},t) = \\
&\quad = \lim_{\tau\to 0} \frac{1}{\tau} \int d\mathbf{y} R(\mathbf{y}) \left[p(\mathbf{y},t+\tau|\mathbf{x},t) - p(\mathbf{y},t|\mathbf{x},t) \right] \\
&\quad = \lim_{\tau\to 0} \frac{1}{\tau} \left[\int d\mathbf{y} R(\mathbf{y}) \int d\mathbf{z} p(\mathbf{y},t+\tau|\mathbf{z},t) p(\mathbf{z},t|\mathbf{x},t) - \int d\mathbf{y} R(\mathbf{y}) p(\mathbf{y},t|\mathbf{x},t) \right] \\
&\quad = \lim_{\tau\to 0} \frac{1}{\tau} \left[\int d\mathbf{y} \int d\mathbf{z} \left\{ R(\mathbf{z}) + \sum_i (\mathbf{y}-\mathbf{z})_i \partial_i R(\mathbf{z}) \right.\right. \\
&\qquad\qquad \left.\left. + \sum_{i,j} (\mathbf{y}-\mathbf{z})_i (\mathbf{y}-\mathbf{z})_j \partial_i \partial_j R(\mathbf{z}) + \ldots \right\} p(\mathbf{y},t+\tau|\mathbf{z},t) p(\mathbf{z},t|\mathbf{x},t) \right. \\
&\qquad \left. - \int d\mathbf{y} R(\mathbf{y}) p(\mathbf{y},t|\mathbf{x},t) \right]
\end{aligned}$$

Here we have used the Chapman-Kolmogorov equation and a Taylor expansion of R about $\mathbf{z}$. In the limit $\tau \to 0$ the integrals in $\mathbf{y}$ that involve powers of $(\mathbf{y}-\mathbf{z})$ can be evaluated using the assumptions 2.3. This gives

$$\int d\mathbf{y} R(\mathbf{y}) \partial_t p(\mathbf{y},t|\mathbf{x},t) = \int d\mathbf{z} p(\mathbf{z},t|\mathbf{x},t) \left[\sum_i A_i \partial_i R(\mathbf{z}) + \frac{1}{2} \sum_{i,j} C_{ij} \partial_i \partial_j R(\mathbf{z}) \right]$$

Using the fact that $R(\mathbf{z})$ is an arbitrary function that may be chosen to have vanishing first derivatives at the image boarder one gets by partial integration a partial differential equation for the transition probability:

$$\partial_t p(\mathbf{z},t|\mathbf{x},t) = -\sum_i \partial_i A_i p(\mathbf{z},t|\mathbf{x},t) + \frac{1}{2} \sum_{i,j} \partial_i \partial_j C_{ij} p(\mathbf{z},t|\mathbf{x},t) \qquad (2.4)$$

This is the *Fokker-Planck equation* for the transition probabilities $p(\mathbf{z},t|\mathbf{x},t)$ of a diffusion process. According to [van Kampen, 1981] it was first used by Rayleigh, Einstein, and Smoluchowsky in a form with A_i linear in $\mathbf{z}$ and C_{ij} constant. Subsequently Planck and Kolmogorov derived a more general form.

To return to the shuffled images: A pixel at position $\mathbf{x}$ at time 0 is allowed to jump to new positions at times $t > 0$ according to the above transition probabilites. All the time it carries around with it the intensity $f(\mathbf{x})$ of the observed image at its starting position $\mathbf{x}$. If we let pixels start from all positions $\mathbf{y}$, one from each,

and the pixels jump independently of each other, then the average intensity of the shuffled image at position $\mathbf{x}$ and at time t is

$$L(\mathbf{x};t) = \int d\mathbf{y}\ p(\mathbf{x},t|\mathbf{y},0)\ f(\mathbf{y})$$

This is taken to define the simplified image $L(\mathbf{x};t)$, as already indicated by the notation.

The simplified image satisfies, via the transition probabilities, the partial differential equation

$$\partial_t L(\mathbf{x},t) = -\sum_i \partial_i A_i(\mathbf{x},t) L(\mathbf{x};t) + \frac{1}{2}\sum_{i,j} \partial_i \partial_j C_{ij}(\mathbf{x},t) L(\mathbf{x};t) \tag{2.5}$$

with initial condition

$$L(\mathbf{x};0) = f(\mathbf{x}) \quad .$$

Equation (2.5) is the generating equation of scale-spaces in general, including the nonlinear scale-spaces where both diffusion coefficient C_{ij} and drift A_i may depend on the local intensity $L(\mathbf{x};t)$. An overview of nonlinear scale-space theory can be found in [Weickert, 1998]. Some of the axiomatic formulations of scale-space also consider the nonlinear case: [Alvarez et al., 1993], [Lindeberg, 1997].

Finally, let us require that shuffling should be homogeneous and isotropic in the sense that the transition probabilities $p(\mathbf{y},t+\tau|\mathbf{x},t)$ should only depend on the distance $|\mathbf{y}-\mathbf{x}|$ and the time difference τ. This necessitates zero drift $A_i = 0$ and diagonal and constant diffusion tensor $C_{ii} = 1$, $C_{ij} = 0$ if $i \neq j$ so that we get the generating equation for linear scale-space:

$$\partial_t L(\mathbf{x},t) = \frac{1}{2}\sum_i \partial_i \partial_i L(\mathbf{x};t) \tag{2.6}$$

That completes a derivation of scale-space from the definition of simplification via shuffling.

2.4.2 Stochastic Simplification and Local Entropy

The idea of images being simplified in scale-space suggests a relation to *information theory*. In some sense one would expect the information to decrease with increasing scale and conversely the *entropy* to increase.

Sporring and Weickert [Sporring and Weickert, 1997], [Sporring, 1999], [Sporring and Weickert, 1999] defined a global entropy

$$-\int d\mathbf{x}\, L(\mathbf{x},t)\ \log(L\mathbf{x},t)$$

of a smoothed image $L(\mathbf{x},t)$ and proved that this increases monotonically with t. The examples in figure (2.1) suggest that images are not only simplified globally but also locally. The shuffled images allow us to make this stronger statement in the following sense.

The intensity at position $\mathbf{x}$ and time $t > 0$ in a shuffled image is a random variable I with density

$$p(I;\mathbf{x},t) = \int d\mathbf{y}\; p(\mathbf{x},t|\mathbf{y},0)\; p(I;\mathbf{y},0)$$

where initially the intensity I at position $\mathbf{y}$ is $f(\mathbf{y})$ with certainty:

$$p(I;\mathbf{y},0) = \delta(I - f(\mathbf{y}))$$

(δ denotes the Dirac delta function). At each position $\mathbf{x}$ and scale t the entropy of this random variable is

$$\begin{aligned} S(\mathbf{x},t) &\equiv -\int dI\; p(I;\mathbf{x},t)\; \log\left(p(I;\mathbf{x},t)\right) \\ &= -\int dI\; r\left(p(I;\mathbf{x},t)\right) \end{aligned}$$

where $r(u) \equiv u \log u$. The entropy $S(\mathbf{x},t)$ is *local* in scale-space. At any single position $\mathbf{x}$ there may be times t when the entopy increases with t and other times when it decreases with t. However, *the sum of the local entropies increases monotonically with* t. To see this, consider the *sum of local entropies*

$$\begin{aligned} \bar{S}(t) &= \int d\mathbf{x}\, S(\mathbf{x},t) \\ &= -\int d\mathbf{x} \int dI\; r\left(p(I;\mathbf{x},t)\right) \end{aligned}$$

Its derivative with respect to t is:

$$\begin{aligned} \partial_t \bar{S}(t) &= -\int d\mathbf{x} \int dI\; r'\left(p(I;\mathbf{x},t)\right)\; \partial_t p(I;\mathbf{x},t) \\ &= -\int d\mathbf{x} \int dI\; r'\left(p(I;\mathbf{x},t)\right) \int d\mathbf{y}\; \partial_t p(\mathbf{x},t|\mathbf{y},0)\; p(I;\mathbf{y},0) \\ &= -\frac{1}{2}\int d\mathbf{x} \int dI\; r'\left(p(I;\mathbf{x},t)\right) \int d\mathbf{y}\; \mathrm{div}\nabla p(\mathbf{x},t|\mathbf{y},0)\; p(I;\mathbf{y},0) \\ &= -\frac{1}{2}\int d\mathbf{x} \int dI\; r'\left(p(I;\mathbf{x},t)\right)\; \mathrm{div}\nabla p(I;\mathbf{x},t) \end{aligned}$$

In the second to last equation we have inserted $\partial_t p(\mathbf{x},t|\mathbf{y},0) = \frac{1}{2}\sum_i \partial_i \partial_i p(\mathbf{x},t|\mathbf{y},0) = \mathrm{div}\nabla p(\mathbf{x},t|\mathbf{y},0)$ [7]. The integral with respect to $\mathbf{x}$ may now be evaluated by Gauss' theorem.

$$\partial_t \bar{S}(t) = \frac{1}{2}\int d\mathbf{x} \int dI\; r''\left(p(I;\mathbf{x},t)\right)\; \left(\nabla p(I;\mathbf{x},t)\right)^2$$

[7] The proof also applies to the more general Fokker-Planck equation (2.5).

Since $r''\left(p(I;\mathbf{x},t)\right) \geq 0$ it follows that

$$\partial_t \bar{S}(t) \geq 0$$

Consequently the sum of local entropies increases (or remains constant) with time or scale.

Chapter 3

Feature Detection

In this chapter we consider the problem of distinguishing "relevant" and "irrelevant" information within a single slice of scale-space. Specifically we study *feature detection* which refers to the following procedure: Process a smoothed image with some local operator. Then classify those *positions* as particularly informative where the operator response is locally extremal.

A number of questions arise immediately: Why are local extrema relevant? What type of local operators may or should be used?

The chapter is organized as follows. The very intuitive method of *pattern matching* briefly motivates the use of local extrema and gives a simple interpretation of *feature detection* in terms of a least squares fit. Then we turn to the operators of feature detection. It is argued that the useful properties of scale-space should be shared by the feature detection operators. This makes the *derivative of Gaussian* filter kernels the generic scale-space operators and opens the way for *differential geometry* as a powerful toolbox for the construction of feature detectors.

3.1 Pattern matching

Consider the situation where a *model* $g(\mathbf{x})$ of the feature of interest is given and the position of this feature is sought in an image $f(\mathbf{x})$. For example we might be seeking Hanna's face in the picture on the right.

The conceptually easiest way to find the position of a feature model is *pattern matching*: the model is positioned somewhere over the image and its fit is measured. This is repeated for different positions and the position with optimal fit is identified.

Suppose the fit of the model at position $\mathbf{x}_0$ is measured in terms of the squared difference

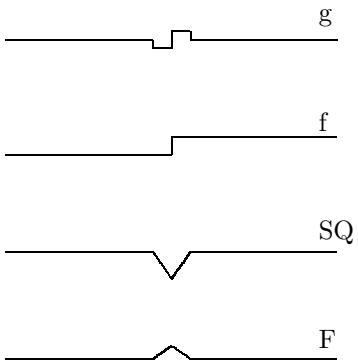
$$SQ(\mathbf{x}_0) = \int d\mathbf{x}\ [g(\mathbf{x}-\mathbf{x}_0) - f(\mathbf{x})]^2$$

(where for simplicity the problem of image boarder is ignored and mathematical tractability is assumed, i.e. g and f square integrable). Then, evidently, the positions of optimal fit may equally be computed from the *operator response* of filtering f with g

$$F(\mathbf{x}_0) = \int d\mathbf{x}\, g(\mathbf{x}-\mathbf{x}_0)\, f(\mathbf{x})$$

The positions $\mathbf{x}_0$ at which F is maximal are exactly those where SQ is minimal since neither $\int d\mathbf{x}\, g(\mathbf{x}-\mathbf{x}_0)^2$ nor $\int d\mathbf{x}\ f(\mathbf{x})^2$ depend on $\mathbf{x}_0$.

As a simple example consider the one-dimensional "edge-model" g and the image f on the right. The squared difference SQ has a single minimum at the position of the step edge in f and the convolution F displays a maximum at the same position.



The interpretation of local extrema of the operator response of a convolution as positions of a least squares fit gives a simple (though restricted[1]) motivation for considering local extrema of operator responses "particularly informative".

3.2 Feature Detection Operators

In 1992 Koenderink and van Doorn wrote [Koenderink and van Doorn, 1992]

> The set of operators in general use comprises an odd lot, with hardly any relations between the various types, nor any clear relations between different versions of the same type (such as edge detectors of

[1] For nonlinear operators the least squares interpretation is not in general possible.

> various orientations), nor especially simple behavior under the action of specific transformation groups (such as translation, rotation, or blurring).

They go on to propose *derivative of Gaussian* filter kernels as the basic feature detection operators because these operators satisfy a *scaling invariance*. We describe the *scaling invariance* property at the end of this section. To begin with the two main approaches to feature detector design are contrasted, followed by a description of derivative of Gaussian feature detectors.

3.2.1 Design Criteria for Feature Detectors

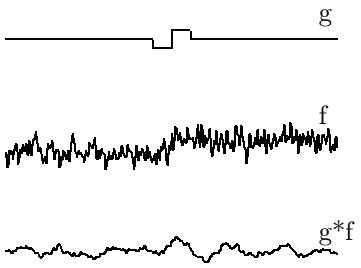
Feature detectors are local operators that are applied to an image in order to subsequently label the local extrema of the response as particularly informative. The design of feature detectors is a fundamental problem of image analysis. The possible gain from a good feature detector is to guide the visual system to *a few* positions in an image which are not only labeled particularly informative during feature detection but which also turn out to *be* particularly useful in the interpretation of the image.

We distinguish two different approaches to the design of feature detectors.

Optimal Design

The historically older approach, most famously pioneered by Canny [Canny, 1986], seeks to optimally balance two opposing qualities of the operators: localization and response to noise.

To demonstrate this consider the "edge detector" of the previous section applied to a noisy step as shown on the right. As can be seen, noise usually produces many extrema in the operator response so that many "false" features are detected. To avoid these, the operator may be constructed to produce a smoother response. In case of our "edge detector" the shape or size of g may be changed, a smoother shape or larger size both leading to a smoother response. Obviously, however, a smooth response is also less sharply peaked at the true positions of features, i.e. the localization error increases.



Invariance Design

The approach of scale-space theory is to construct operators with the same useful properties as possessed by the Gaussian filter kernel. In particular, a translated, rotated, or scaled image should yield the same features as the original image, only translated, rotated, or scaled.

This approach is not as much a design approach to feature detectors as the above. However, as Koenderink and van Doorn write, after fulfilling the invariance requirements "there turns out to be almost no room for "optimization" of operators for various tasks; in most cases that would lead to certain unwanted biases toward certain scales or orientations" [Koenderink and van Doorn, 1992] .

Translation invariance is satisfied by any convolution kernel. Rotation invariance is given either for rotation invariant kernels or when the preferred direction is fixed relative to the image. Scaling invariance is satisfied by derivative of Gaussian filter kernels as will be demonstrated in the following section.

3.2.2 Derivative of Gaussian Feature Detectors

In Gaussian scale-space the only scaling invariant filter kernels are linear combinations of derivative of Gaussian filter kernels (see below). This makes them *the* basic feature detection operators within scale-space theory. All feature detectors of the theory are linear or nonlinear combinations of responses to derivative of Gaussian filters.

Figure (3.1) shows some graphs of one-dimensional derivatives of Gaussians.

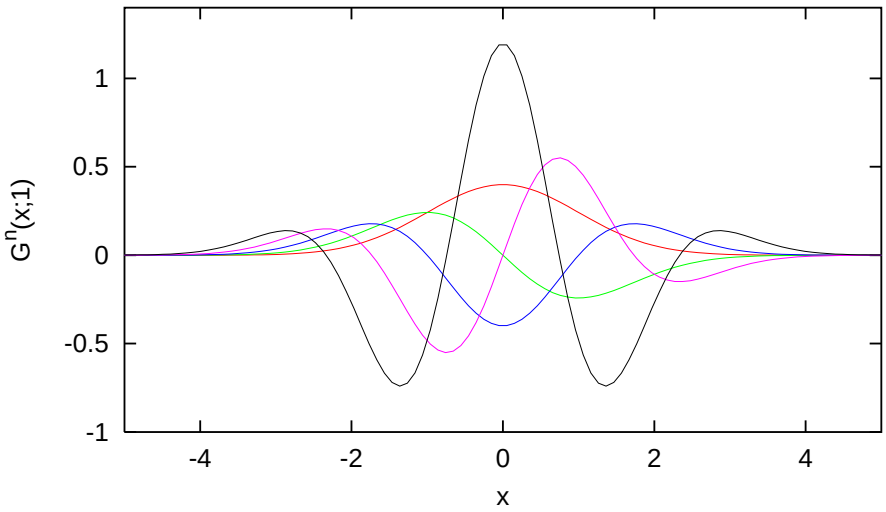


Figure 3.1: One-dimensional derivative of Gaussian filter kernels of orders 0,1,2,3, and 4.

In higher dimensions derivatives may be computed along different directions. The derivatives of Gaussians along Cartesian coordinates are

$$G^{\mathbf{n}}(\mathbf{x};t) = \partial_1^{n_1} \ldots \partial_N^{n_N} \frac{e^{-\frac{\mathbf{x}^T\mathbf{x}}{2t}}}{(2\pi t)^{N/2}}$$

where $\partial_i^{n_i}$ is the n_i-th order derivative along the i-th Cartesian coordinate. To evaluate these functions as shown in figure (3.2) we computed derivatives after Fourier transformation. This is particularly easy, requiring only multiplication of

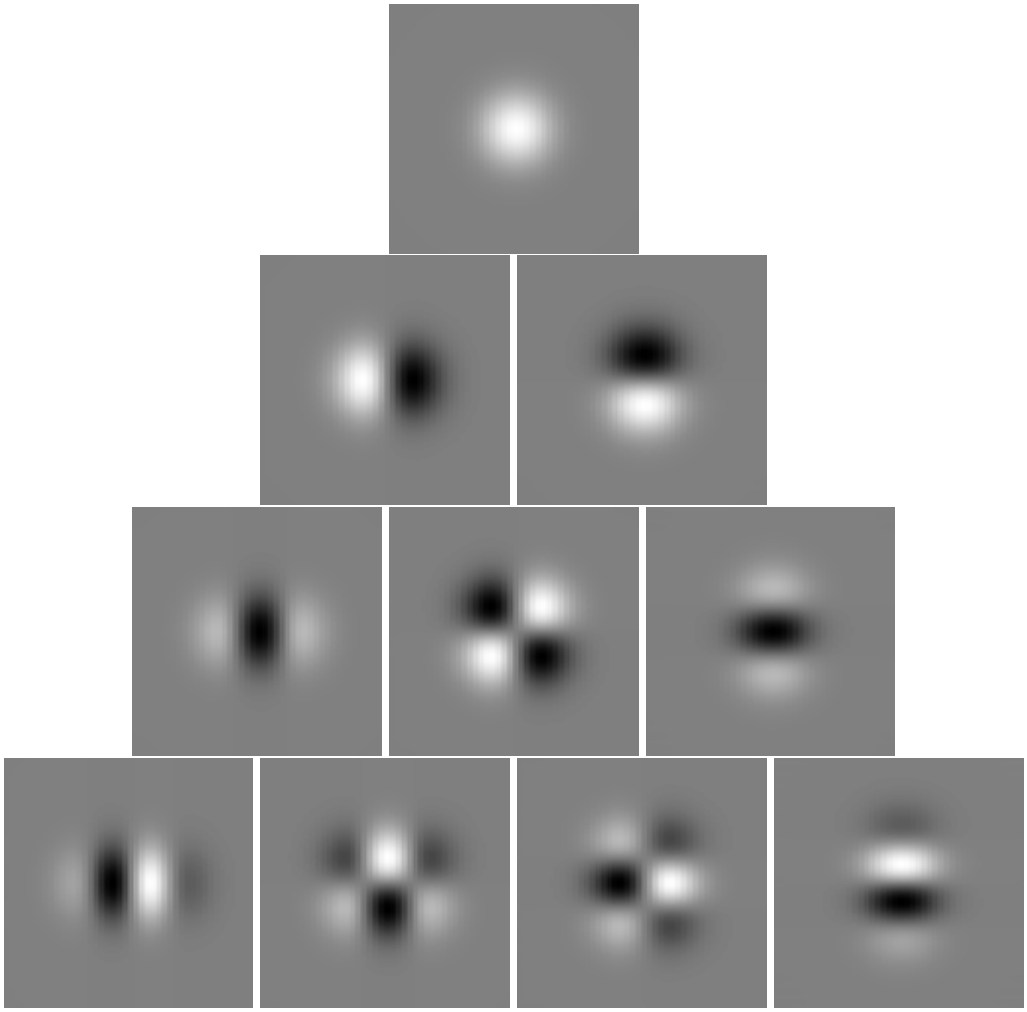


Figure 3.2: Two-dimensional derivative of Gaussian filter kernels of orders 0,1,2, and 3. From top to bottom and left to right: $G^{0,0}$, $G^{1,0}$, $G^{0,1}$, $G^{2,0}$, $G^{1,1}$, $G^{0,2}$, $G^{3,0}$, $G^{2,1}$, $G^{1,2}$, $G^{0,3}$. This figure is reproduced after [Lindeberg, 1994b, p 142].

the (Fourier transformed) Gaussian with a polynomial:

$$G^{\mathbf{n}}(\mathbf{x};t) = \int d\mathbf{k}\, e^{-2\pi i \mathbf{k}^T \mathbf{x}}\, (2\pi i \mathbf{k}_1)^{n_1} \ldots (2\pi i \mathbf{k}_N)^{n_N}\, e^{-2\pi^2 \mathbf{k}^T \mathbf{k} t}$$

The Fourier domain is also used to compute convolutions of images with derivative of Gaussian kernels. This is particularly advantageous at small scales where the *analytical* Fourier transform of the filter kernels, $(2\pi i \mathbf{k}_1)^{n_1} \ldots (2\pi i \mathbf{k}_N)^{n_N}\, e^{-2\pi^2 \mathbf{k}^T \mathbf{k} t}$, may be sampled with higher precision than in the spatial domain, owing to the fact that narrow Gaussians in the spatial domain are wide Gaussians in the Fourier domain.

Derivatives along other coordinates are linear combinations of the Cartesian derivatives. Examples of directional derivatives will be considered in the section on differential geometry below.

Scaling Invariance

Let us now consider the scaling invariance of derivative of Gaussian filter kernels in linear scale-space. The response of a filter kernel $g : \mathbb{R}^N \to \mathbb{R}$ to the scale-space of an image f at scale $\sqrt{t}$ is given by

$$g * G(\cdot;t) * f$$

(where $G(\cdot;t) = G^{\mathbf{0}}(\cdot;t)$ is the rotation symmetric Gaussian). If the image is scaled in size, $f^s(\mathbf{x}) = f(s\mathbf{x})$, the result is $g * G(\cdot;t) * f^s$. Since f and f^s depict the same information apart from the size change it appears reasonable to require that the output too should differ only by a simple transformation of size and scale. Formally this scaling-invariance requires

$$(g * G(\cdot;t(t',s)) * f)(s\mathbf{x}) = (g * G(\cdot;t') * f^s)(\mathbf{x}) \qquad \forall \mathbf{x} \quad \text{and} \quad \forall t' \geq 0$$

for some invertible dependence $t(t',s)$ that satisfies $t(t',s) > t'$ if $s > 1$. This equation should hold for any image f so that it becomes a condition on the filter kernels alone. Choosing $s > 1$ and $t' = 0$ we see that *convolution of g with a Gaussian should be equivalent to rescaling g* (see figure 3.3):

$$(g * G(\cdot;t(s)))(s\mathbf{x}) = g(\mathbf{x}) \tag{3.1}$$

Clearly any derivative of Gaussian $g(\mathbf{x};\tau) = \partial_1^{n_1} \cdots \partial_N^{n_N} G(\mathbf{x};\tau)$ satisfies the scaling-invariance with $t(s) = [s^2 - 1]\tau$:

$$(\partial_1^{n_1} \cdots \partial_N^{n_N} G(\cdot;\tau) * G(\cdot;[s^2-1]\tau))(s\mathbf{x}) = \partial_1^{n_1} \cdots \partial_N^{n_N} G(\mathbf{x};\tau)$$

Since (3.1) is linear in g any linear combination of derivatives of Gaussians is scaling-invariant as well.

It is possible to show that the converse also holds true: Any scaling-invariant operator is a linear combination of derivatives of Gaussians. Koenderink and van Doorn [Koenderink and van Doorn, 1992] prove this result by deriving a complete set of solutions of the diffusion equation that generates linear scale-space.

3.2.3 Interpretations of Derivative of Gaussian Detectors

Several interpretations of the feature detection operators suggest themselves. Particularly interesting is the differential geometric point of view that will be taken up in the next section.

Operators on scale-space

Within the scale-space paradigm it is natural to apply feature detection operators to the slices from the scale-space of an image rather than to the original image. The response to a feature detection operator g is then

$$g * (G(;t) * f) \qquad \forall t \geq 0$$

In this interpretation *one* operator is applied to the data at different scales.

Operators on the image

Alternatively one may see $g * G(;t)$ as a *family of feature detectors* to be applied to the original image f:

$$(g * G(;t)) * f \qquad \forall t \geq 0$$

In this interpretation the invariance requirement (3.1) says that operators of the family should differ only in scale not in shape. As the following figure (3.3) shows, the derivative of Gaussian edge detectors are scaling-invariant while the family of edge-detectors generated from the step model: $g(x) = -1$ for $-1 < x < 0$, $g(x) = 1$ for $0 < x < 1$ and $g(x) = 0$ for $|x| > 1$ is not scaling-invariant.

Differential Operators on Scale-Space

Another point of view is that the convolution of a derivative of Gaussian filter kernel with an image is simply a *derivative of scale-space*:

$$G^{\mathbf{n}}(\cdot;\tau) * G(\cdot;t) * f = \partial_1^{n_1} \cdots \partial_N^{n_N} (G(\cdot;t+\tau) * f)$$

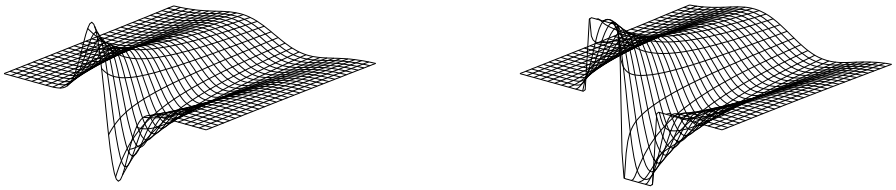


Figure 3.3: Families of one-dimensional edge detectors. Left: scaling invariant first derivative of Gaussian. Right: step edge family. Clearly the step edge family is not scaling invariant.

This suggests that feature detection can be interpreted as a way to search for positions where the *differential geometry* of scale-space is "particularly informative". This point of view has spurred a lot of research applying ideas of differential geometry to vision (see e.g. [ter Haar Romeny, 1994]). It should be noted however that what is "particularly informative" to a visual system need *not necessarily* be special in terms of differential geometry and vice versa. Chapter 6 describes an approach to characterize the "particularly informative" relative to some images that are by definition considered uninformative. This allows a visual system to define what is considered uninformative and informative.

3.3 Differential Geometry of Scale-Space

The response of a derivative of Gaussian feature detector is identical to the corresponding derivative of a slice $L(\cdot;t)$ of scale-space:

$$\left(\partial_1^{n_1}\cdots\partial_N^{n_N}G(\cdot;t)\right)*f=\partial_1^{n_1}\cdots\partial_N^{n_N}\left(G(\cdot;t)*f\right)=\partial_1^{n_1}\cdots\partial_N^{n_N}L(\cdot;t)$$

This interpretation has proved extremely useful to the design of "real" feature detectors that respond to the local *structural* information content of an image. The construction of differential expressions that capture the *intrinsic* (or *structural*, or *geometric*) properties of a "landscape" $L(\cdot;t)$ is the subject of differential geometry which has become an important tool due to the above equation.

This section describes *how to construct structural feature detectors* from the basic derivative of Gaussian feature detectors. Consequently the only "differential geometry" used here are *local coordinates*. (More detailed discussions can be found in [Koenderink, 1990], [Florack, 1993], [Kanatani, 1990] or [ter Haar Romeny, 1994] for an overview.)

First observe that not all partial derivatives of scale-space represent structurally meaningful information about an image. A derivative such as $\partial_1 L$ crucially depends on the orientation of its $x_1, ..., x_N$ coordinate system in relation to the scene. In other words, the information provided by $\partial_1 L$ about the scene changes when the $x_1, ..., x_N$ coordinate system of the observer rotates, as demonstrated in figure (3.4).

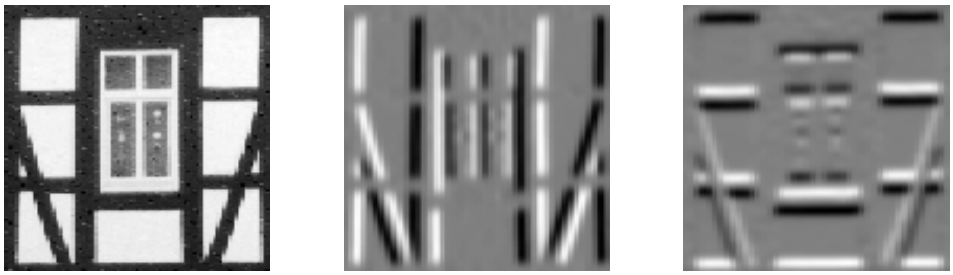


Figure 3.4: Orientation dependence: A first derivative of Gaussian of the window image computed from two different observer orientations. In the central image the derivative was computed from left to right and in the right image from bottom to top.

A structural feature detector must be rotation invariant in the sense that it commutes with rotations: First rotating the image and then applying the feature detector must yield the same as first applying the feature detector and then rotating its response. Structural feature detectors should also be translation and scaling invariant. These properties, however, are already satisfied by any derivative of Gaussian feature detector, so they are automatically inherited by feature detectors constructed from derivatives of Gaussians.

The simplest example of a linear rotation invariant feature detector is the Laplacian:

$$\Delta L(\circ;t) = \sum_{i=1}^{N} \partial_i \partial_i \, L(\circ;t) = \left(\sum_{i=1}^{N} \partial_i \partial_i \, G(\circ;t) \right) * f$$

The image on the right shows the Laplacian of the window image computed at the same scale as above. Rotation invariance of the Laplacian derivative of a slice from scale-space arises from the rotation invariance of the Laplacian of Gaussian filter kernel $\sum_{i=1}^{N} \partial_i \partial_i \, G(\circ;t)$. In terms of the differential geometric point of view the Laplacian is rotation invariant because it is a fully contracted derivative tensor.

A more powerful way to construct rotation invariant operators is to formulate feature detectors in terms of *local coordinates*. This is now considered.

3.3.1 Local Coordinates

Rotation invariance of differential expressions is naturally achieved by introducing *local coordinates* or *gauge coordinates*. At each point of an image these coordinates are fixed by the local image geometry. Thus, when the image rotates the coordinates rotate with it and correspondingly any derivatives computed along local coordinates describe structural or intrinsic properties of the image.

Gradient Coordinates

One choice of local coordinates in two dimensions are the *gradient coordinates*. At each point of a slice of scale-space the vector **v** points along the gradient and the vector **u** orthogonal to it. Figure (3.5) shows the coordinates at two positions of a synthetic slice.

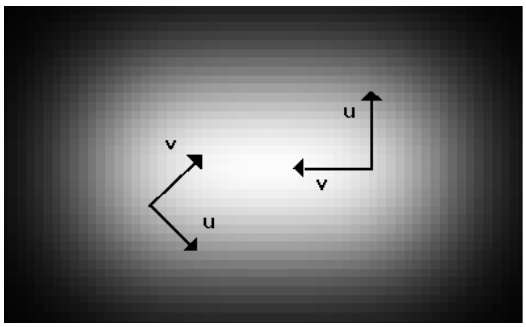


Figure 3.5: Two local gradient coordinate systems. **v** points along the gradient and **u** along the isophote tangent orthogonal to **v**.

The directional derivative $\partial_v L$ along the gradient yields the absolute value of the gradient and the directional derivative along **u** vanishes by definition, i.e. $\partial_u L = 0$.

Figure (3.6) displays $\partial_v L$ and the orientation of **v** for several scales of the window image of figure (3.4) at two different scales. Evidently the coordinate system also depends on the scale $\sqrt{t}$ of the slice $L(\circ;t)$ from scale-space.

Curvature Coordinates

Another choice of local coordinates is a system oriented along the directions of principal *curvature*. In two dimensions these are denoted as **p** and **q**. They may either be defined intrinsically by $\partial_p \partial_q L(\circ;t) = 0$, $\partial_p \partial_p L(\circ;t) \leq \partial_q \partial_q L(\circ;t)$, and

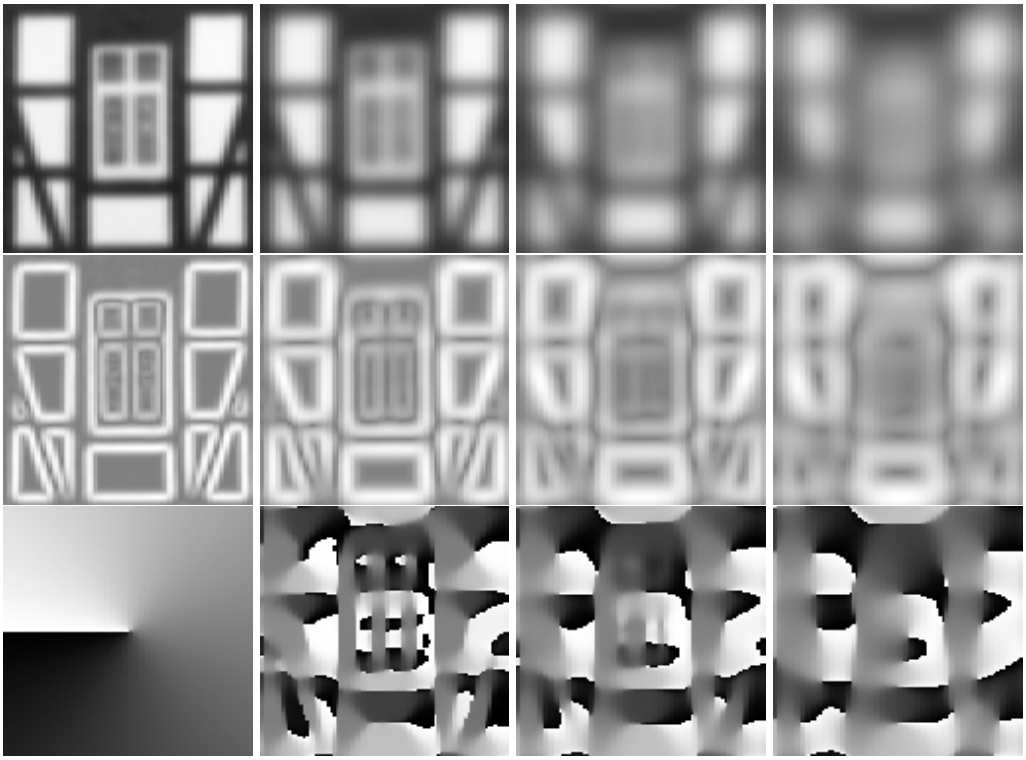


Figure 3.6: Gradient coordinates. The top row shows four slices from the scale-space of the window image of figure (3.4). The image has 128 by 128 pixels and the four scales are $\sqrt{t} = 2$, $\sqrt{t} = 4$, $\sqrt{t} = 6$, and $\sqrt{t} = 8$ (in units of one pixel width/height). The second row shows $\partial_v L(\circ;t)$, the absolute value of the gradient at the same scales. The bottom left image displays the mapping of orientations to gray values that was used to encode the orientations of the gradient at the three scales $\sqrt{t} = 4$, $\sqrt{t} = 6$, and $\sqrt{t} = 8$.

the condition that **p** be orthogonal to **q**, or in terms of extrinsic Cartesian (x,y)-coordinates as the eigenvectors of the Hessian H

$$\begin{array}{l} \mathbf{H}\,\mathbf{p} = L_{pp}\mathbf{p} \\ \mathbf{H}\,\mathbf{q} = L_{qq}\mathbf{q} \end{array} \qquad \mathbf{H} = \begin{bmatrix} L_{xx} & L_{xy} \\ L_{yx} & L_{yy} \end{bmatrix}$$

with eigenvalues $L_{pp} \leq L_{qq}$.

Computation of Directional Derivatives

To compute local directional derivatives they must be expressed in terms of extrinsic coordinates. For the gradient coordinates in two dimensions this is facilitated

by the coordinate transformation from (v,u) to (x,y):

$$T^{v,u} = \frac{1}{\sqrt{L_x^2+L_y^2}} \begin{bmatrix} L_x & L_y \\ L_y & -L_x \end{bmatrix}$$

that allows the first order derivative tensor to be written as

$$\begin{pmatrix} \partial_v \\ \partial_u \end{pmatrix} = (T^{v,u})^{-1} \begin{pmatrix} \partial_x \\ \partial_y \end{pmatrix}$$

i.e. $\partial_v = (L_x\partial_x + L_y\partial_y)/\sqrt{L_x^2+L_y^2}$ and $\partial_u = (L_y\partial_x - L_x\partial_y)/\sqrt{L_x^2+L_y^2}$. Explicit expressions for higher order derivatives in v,u can be constructed by multiplication of these lowest order expressions, e.g. $\partial_v\partial_u = (L_x\partial_x + L_y\partial_y)(L_y\partial_x - L_x\partial_y)/(L_x^2 + L_y^2)$.

3.4 Zero-Crossings

Having discussed the construction of feature detection operators we return to the question of "particularly informative" positions. Within *feature detection* the particularly informative positions are *defined* as local extrema of a feature detector's response.

Why local extrema should be particularly informative *to a visual system* is not immediately apparent. Certainly local extrema satisfy a number of important properties that make them good candidates. Among these properties are: i) local extrema are structural properties of the image alone, i.e. they do not depend on any parameters that might require user interaction. ii) local extrema of rotation (translation, scaling ...) invariant feature detectors share these invariances. These points are often cited to motivate the use of local extrema [ter Haar Romeny, 1994],[Lindeberg, 1994a].

It is the author's opinion, however, that the argument by which particularly informative positions are defined as local extrema *should also be able to deal with the second parameter of scale-space, scale*, in the sense of providing a definition for particularly informative scales.

A method to select particularly informative scales has been proposed by Lindeberg [Lindeberg, 1993b]. It has been very successfully applied among others by Lindeberg [Lindeberg, 1998a], Pizer et al. [Pizer et al., 1998], [Morse et al., 1994], and Lorenz et al.[Lorenz et al., 1997a]. Chapter 6 presents an attempt to motivate *scale-selection* and feature detection from the same principles.

Let us now consider some examples of "particularly informative" positions of different feature detectors.

Figure (3.7) shows the *edges* of the window image computed at several scales. They are defined as *local maxima of the gradient along the gradient direction*. In terms of *zero-crossings* the equivalent definition is:

$$L_{vv} = 0$$
$$L_{vvv} < 0$$

where **v** is the local direction of the gradient at the considered scale.

A comparison of the edges of figure (3.7) with the response of the feature detector L_v in figure (3.6) might evoke some criticism about the particularly informative edges some of which appear rather uninformative. While this criticism is justified it has to be kept in mind that the occurrence of "false" responses is a problem common to all feature detection and thus one that requires attention in a more general setting. Secondly it must be remembered that there is a *qualitative difference* between the response of the feature detector and the edges computed from this. The edges are a set of *positions* while the response of the feature detector is "merely" a mapping that assigns any position a scalar value. Formulated differently, at each position the presence of an edge or not is a binary decision while the response of the feature detector is continuous.

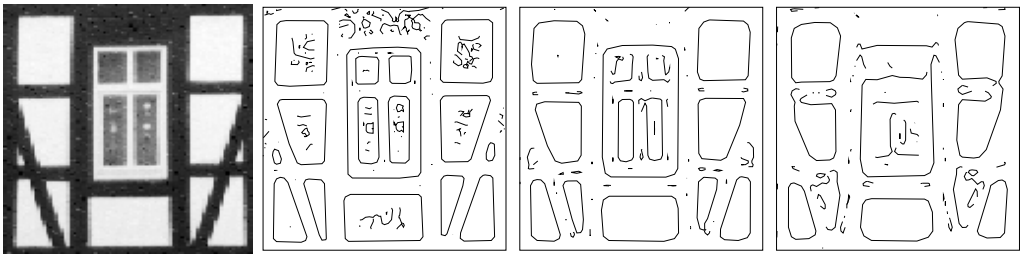


Figure 3.7: Edges of the window image at different scales. The image has 128x128 pixels and the displayed edges were computed at scales $\sqrt{t} = 2$, $\sqrt{t} = 4$, and $\sqrt{t} = 6$. The edges are defined in terms of zero-crossings by $L_{vv}(\circ;t) = 0$, $L_{vvv} < 0$.

Figure (3.8) shows "ridges" of the "Fachwerkhaus" image (which the window is part of) at several scales. The displayed ridges are defined in terms of curvature coordinates as maxima of the image intensity along the direction **p** of minimum second derivative which also satisfy $|L_{pp}| > |L_{qq}|$. A detailed description of ridge detection is presented in the following chapter which also demonstrates the limitations of a fixed scale approach.

Before ending the chapter let us briefly refer to computational aspects. In both examples local extrema were detected in one-dimensional frames within the two-dimensional images. These *generalized maxima* or *critical points* pose some

computational problems that may be approached in several ways. Generally the options are direct maximization or root finding, i.e. the computation of zero-crossings of derivatives [Press et al., 1988]. Further it must be decided whether to sample functions at arbitrary positions or at the discrete positions of the grid on which the original image data are given. Chapter 8 presents algorithms for zero-crossings of discretely sampled functions. A discussion of zero-crossing methods can also be found in [Eberly, 1996] and [Lindeberg, 1998a]. Alternatively [Staal et al., 1999] propose a direct maximization approach.

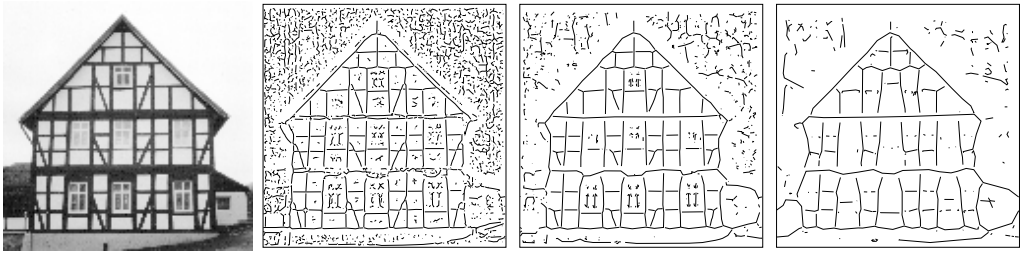


Figure 3.8: Ridges of the "Fachwerkhaus" image at different scales. The image has 512x512 pixels and the displayed ridges were computed at scales $\sqrt{t} = 2$, $\sqrt{t} = 4$, and $\sqrt{t} = 8$.

Chapter 4

Scale Selection

This chapter and chapter 6 deal with the problem of determining "particularly informative" scales in the response of some local operator applied to an image. This chapter describes a method for scale selection that is analogous to the method for feature detection. In combination feature detection and scale selection are performed as follows: Process an image with some local operator. Then classify those (position,scale)-pairs as particularly informative where the operator response has a local extremum with respect to position and scale.

4.1 The Need for Scale Selection

The scale-space concept aims to describe each "object" within an image at its appropriate scale. The basic idea to achieve this links the degree of smoothing within scale-space to the scale of objects as follows: With increasing degree of smoothing objects vanish from the image, small objects first and larger objects later. The degree of smoothing at which an object vanishes basically measures the size of the object. To find the appropriate scales, evidently, one must analyze the image at all scales and then select those that are "particularly informative".

The importance of choosing appropriate scales is best demonstrated by some examples. Figure (4.1) shows fixed scale ridges (particularly informative positions) of a grass image at five different scales. Clearly, at small scales the thick leaves of the grass are not detected while at larger scales the thin leaves and the stem escapes detection. Finally, at very large scales even the large leaves disappear.

The need to select scales may also be motivated from a more technical point of view. Any feature detection operator has some scale or size. For the derivative

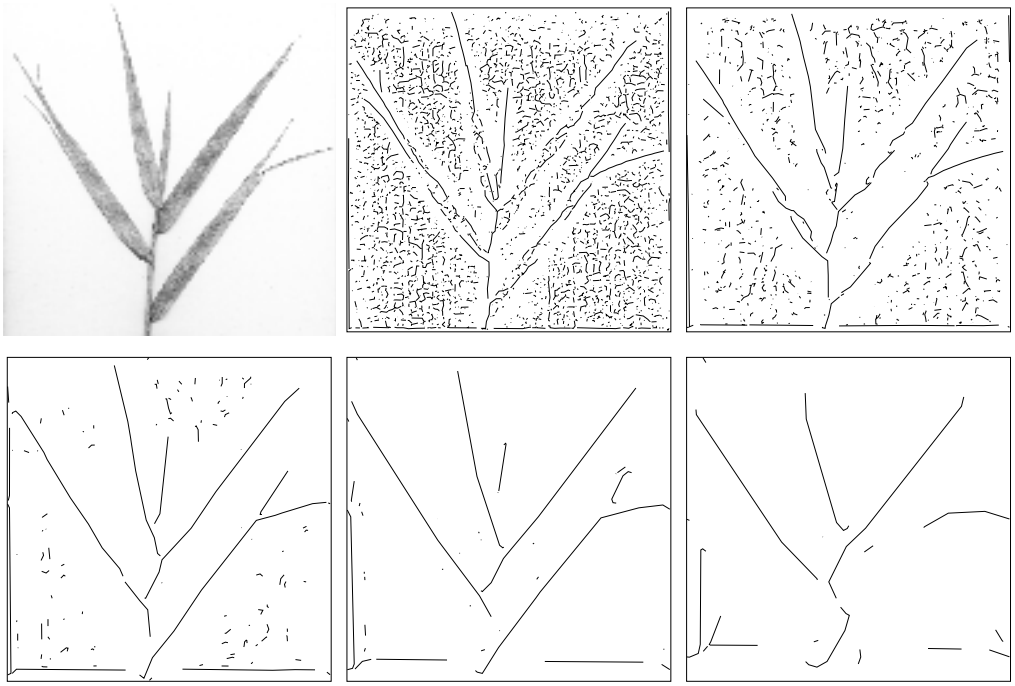


Figure 4.1: Ridges at fixed scales. The images have 512 by 512 pixels and the chosen scale levels are $\sqrt{t} = 2$, $\sqrt{t} = 4$, $\sqrt{t} = 8$, $\sqrt{t} = 16$, and $\sqrt{t} = 32$ (in units of 1 pixel width/height).

of Gaussian operators this is usually made explicit in the notation:

$$G^{\mathbf{n}}(\mathbf{x};t) = \partial_1^{n_1} ... \partial_N^{n_N} \frac{e^{-\frac{\mathbf{x}^T\mathbf{x}}{2t}}}{(2\pi t)^{N/2}}$$

The operators are parameterized by both position and scale. It is then natural to ask not only what the particularly informative positions are but also what the particularly informative scales are.

4.2 Invariance Requirements and Scale Selection

Ironically scale-space theory, which aims to describe each object at its appropriate scale, has not accomplished (in fact not even dealt with) the problem of scale selection for almost a decade. Why not?

The answer can be seen in the fact that scale-space theory has, for a long time, focused on invariance requirements. The theory has profited much from its detailed treatment of invariance requirements. At the same time, however, it appears

to have gone unnoticed [1] that the problem of scale-selection just cannot be solved by an invariance requirement. This has a rather simple reason: Invariance requirements "conserve" information under some specified transformations of the data. Scale selection "destroys" information in the sense that the particularly informative scales, or positions and scales, of some operator response do not contain the same information as the original data. Of course the purpose of scale-selection is not in the first place to destroy information, but rather to distinguish between relevant and irrelevant. This, however, amounts to destroying the irrelevant information and consequently scale-selection cannot be derived from an invariance requirement.

4.3 Normalized Derivatives

The first approach to scale-selection that deals with positions and scales simultaneously was proposed by Lindeberg in 1993 [Lindeberg, 1993b] and, in a context not concerned with derivatives, in [Lindeberg, 1993a].

One observes that the amplitude of the response

$$L_{\mathbf{n}}(\mathbf{x};t) = (G^{\mathbf{n}}(\circ;t) * f)(\mathbf{x})$$

of a derivative of Gaussian operator to an image f tends to decrease with increasing scale because the response is increasingly smoothed. For this reason it makes little sense to define particularly informative scales in terms of local maxima of the operator response with respect to scale, as one might wish to do in analogy to particularly informative positions that are defined as local extrema of the operator response with respect to space.

The amplitude of *normalized derivatives*

$$t^{n/2} L_{\mathbf{n}}(\mathbf{x};t) = t^{n/2} (G^{\mathbf{n}}(\circ;t) * f)(\mathbf{x})$$

where $n = n_1 + \ldots + n_N$, is obviously greater than that of regular derivatives when $t > 1$. This prompted Lindeberg to study local maxima of normalized derivatives with respect to scale and to propose a heuristic principle:

> In the absence of other evidence, a scale level at which some (possibly non-linear) combination of normalized derivatives assumes a local maximum can be treated as reflecting the characteristic length of a corresponding structure in the data.

The idea proved to be very useful and, in its more general form to be considered next, it has been applied to detect blood vessels [Koller, 1995], [Koller et al., 1995], [Lorenz et al., 1997a], and other structures whose size is of interest [Pizer et al., 1998], [Pizer et al., 1994], [Fritsch et al., 1994].

[1] As far as I can judge from the literature.

4.4 γ-Normalized Derivatives

The approach to scale-selection using normalized derivatives can be generalized by considering so-called γ-normalized derivatives [Lindeberg, 1998a]:

$$t^{\gamma n/2} L_{\mathbf{n}}(\mathbf{x};t) = t^{\gamma n/2} (G^{\mathbf{n}}(\circ;t) * f)(\mathbf{x})$$

for some $\gamma > 0$. The possibilities that arise from this generalization are briefly discussed.

To analyze the influence of γ on scale selection it is useful to consider the logarithm of the γ-normalized derivatives.

$$\gamma \frac{n}{2} \log(t) + \log\left(L_{\mathbf{n}}(\mathbf{x};t)\right) \tag{4.1}$$

The local maxima of γ-normalized derivatives occur at the same scales (or positions, if one maximizes with respect to **x**) as the local maxima of their logarithm. Clearly, with increasing γ the influence of the first term increases and the local maxima with respect to scale are pushed toward larger scales.

This γ-dependence suggests that the value of γ may be adjusted such that the "correct" scales are selected in some model situations where a "correct" scale of the image structure is known a priori. To give an example consider the one-dimensional "blob" model $e^{-\frac{x^2}{2w}}/\sqrt{2\pi w} = G(x;w)$ of width w. Suppose one attempts to detect such "blobs" in one dimensional images and one chooses to do so with a second derivative of Gaussian operator. The scale-space of the blob model is $G(\mathbf{x};t+w)$ and the γ-normalized second derivative of Gaussian operator response is $t^{\gamma} G^2(\mathbf{x};t+w) = t^{\gamma}(\frac{x^2}{(t+w)^2} - \frac{1}{t+w}) G(\mathbf{x};t+w)$. The maximum over scales at the center $x = 0$ of the blob occurs at $t = \frac{\gamma}{3/2-\gamma} w$. To achieve a one to one correspondence between the selected scale and the width of the model γ must be set to $\gamma = 3/4$.

A more important consequence of the choice of γ-value is the following. For any specific image structure there is a certain range of γ-values within which the structure is assigned a finite scale. For values of γ greater than some *critical* γ the first term in equation (4.1) dominates so much that the maxima with respect to scale are pushed to infinity. In the blob-example the critical γ-value is $\gamma = 3/2$. For larger values of γ the γ-normalized second derivative has no maximum over scales. In other words, *the choice of* γ *determines which structures can be detected and which not*.

Chapter 7 demonstrates scale-selection for ridge detection using γ-normalized derivatives. It will be seen that a "correct" choice of γ allows a second derivative ridge detector to "distinguish" between ridges and edges.

Chapter 5

Ridge Detection at Fixed Scales

Elongated bright structures on a dark background or dark structures on a bright background are the focus of this chapter. Typical examples picture branches of trees, arteries in medical images, or roads in areal photographs. Inherent to all is a curved *path* along which the structure extends with some characteristic *width*. Here we concentrate on the path *assuming that the width is a priori known*. A later chapter will treat path and width jointly.

The chapter is organized as follows. To begin with, some possible definitions of *ridges* are briefly considered and contrasted. Then we restrict discussion to those *ridges* that can be defined and detected locally in terms of zero-crossings of derivatives. The computation of these zero-crossings presents some technical difficulties which will be discussed in detail. Examples are given throughout.

5.1 Ridge Definitions

An image becomes an "intensity landscape" if one interprets brightness as height. This allows one to speak of local maxima as *peaks*, of local minima as *pits* and of saddle-points (where the gradient vanishes but which are neither maximum nor minimum) as *passes*.

By the same analogy other geomorphological terms may be applied. A *geomorphological ridge* is thus the path of steepest ascent leading from a pass to a peak. A *geomorphological dale* is the path of steepest descent leading from a pass to a pit. This ridge definition has been proposed as a useful feature for image analysis [Koenderink and van Doorn, 1993], [Koenderink and van Doorn, 1994] and has been successfully applied to image segmentation [Griffin et al., 1992].

Notably the geomorphological ridge definition makes no mention of ridge shape. Such ridges may be long and sharply peaked in the direction perpendicular to the steepest ascent. They may also be short and show almost no peak. In fact,

it is not possible to say whether a point lies on a geomorphological ridge or not based only on the *local* shape of the landscape around it.

The global character of geomorphological ridges has created some problems and confusion concerning a precise mathematical definition. The articles [Koenderink and van Doorn, 1993] and [Koenderink and van Doorn, 1994] clarify the matter from a modern differential geometric point of view. In particular they discuss some historical attempts to define geomorphological ridges from local properties alone and conclude that this is "doomed to fail from the very start".

Despite the lacking relation to geomorphological ridges one speaks of "ridges" in image analysis also when these are defined in terms of local properties. An early local ridge definition is due to Haralick [Haralick, 1983]. It is similar to the second of two definitions that will be considered here, the *height ridge* and the *second derivative ridge*. For 2-dimensional images/landscapes the *height ridge* discussed in detail by [Eberly, 1996] is defined as follows: At each point in the landscape choose the direction (axis) along which the landscape has the strongest downward bend. If the landscape falls off to both sides along this axis the point lies on a *height ridge* (equation 5.1).

Another definition looks only at second derivatives of the landscape: As above, at each point the axis of a hypothetical ridge is defined to be orthogonal to the axis of minimum second derivative. The point under consideration is defined to be on a ridge if the second derivative in the direction traversing the ridge has a minimum in that direction (equation 5.3). We call this a *second derivative ridge*.

A different approach is to characterize ridges in terms of edges to the left and right of the ridge [Koller, 1995], [Koller et al., 1995], [Morse et al., 1994]. In the following we consider only the height ridge and the second derivative ridge.

5.2 Height Ridges

The detection of *height ridges* and *second derivative ridges* proceeds in two steps. First at each point of a two-dimensional image a set of orthogonal directions is chosen, one pointing along the hypothetical ridge and the other perpendicular to the ridge. Then, at each point, the image intensity is analyzed along the direction across the hypothetical ridge to see if the point is on a ridge or not.

Directions are determined from second derivatives of the image intensity f at some scale $\sqrt{t}$ from linear *scale-space*, i.e. from second derivatives of $L(\mathbf{x}) = (G(\circ,t) * f)(\mathbf{x})$ where $G(\mathbf{x},t) = e^{-\frac{\mathbf{x}^T\mathbf{x}}{2t}}/(2\pi t)$ is the rotation symmetric Gaussian function, $*$ is the convolution operator (we have omitted t on the left hand side because it is considered constant throughout the chapter). $L(\mathbf{x})$ is a slice from scale-space, i.e. a smoothed image, that may be interpreted as "landscape".

The direction along which a ridge extends is defined as the direction of maximum second derivative and is denoted as **q**. The orthogonal axis of minimum second derivative traverses the ridge and is denoted by **p** (see appendix A for an alternative definition of **p** and **q**). Within these local coordinates the *height ridge* according to Eberly [Eberly, 1996] is defined as:

$$\begin{aligned} L_p &= 0 \\ L_{pp} &< 0 \\ L_{pp} &\leq L_{qq} \end{aligned} \tag{5.1}$$

We follow the slightly different definition of Lindeberg [Lindeberg, 1998a]:

$$\begin{aligned} L_p &= 0 \\ L_{pp} &< 0 \\ |L_{pp}| &\geq |L_{qq}| \end{aligned} \tag{5.2}$$

The two definitions differ when the landscape bends upward along the ridge direction $L_{qq} > 0$ more strongly than it bends down in the orthogonal direction, i.e. when $L_{qq} > -L_{pp}$, $L_{pp} < 0$. Eberly includes points satisfying $L_p = 0$, $L_{pp} < 0$, $L_{qq} > -L_{pp}$ in the height ridge; Lindeberg excludes them. Figure (5.2) shows both types of ridges for a synthetic landscape. Apart from differences at the image border the definition of Eberly produces a longer ridge along the structure and two ridges approaching from the steep side. Let us note again that the detectors respond to local maxima along **p** irrespective of how pronounced these are in the sense that both of the following sequences of three numbers $.97^{-10}, .98^{-10}, .96^{-10}$ and $9, 100, 20$ have a maximum in the middle. This is a property of all feature detectors that requires separate attention.

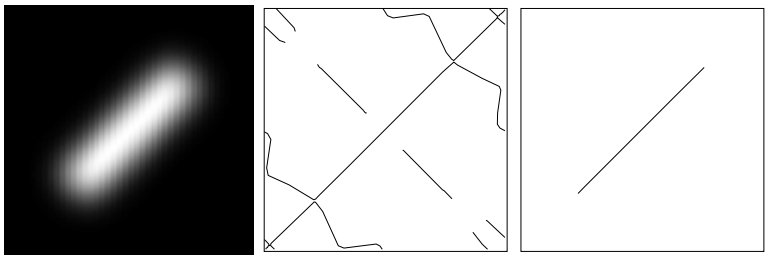


Figure 5.1: Height ridges resulting from definitions (5.1) and (5.2).

Examples of height ridges in real images are given in figure (5.2). Ridges are computed at three different scales for each image showing how different structures appear at different scales. This is particularly pronounced in the ridges of the cotton fabric where at a small scale single threads are detected that merge into orthogonal ridge structures at a larger scale.

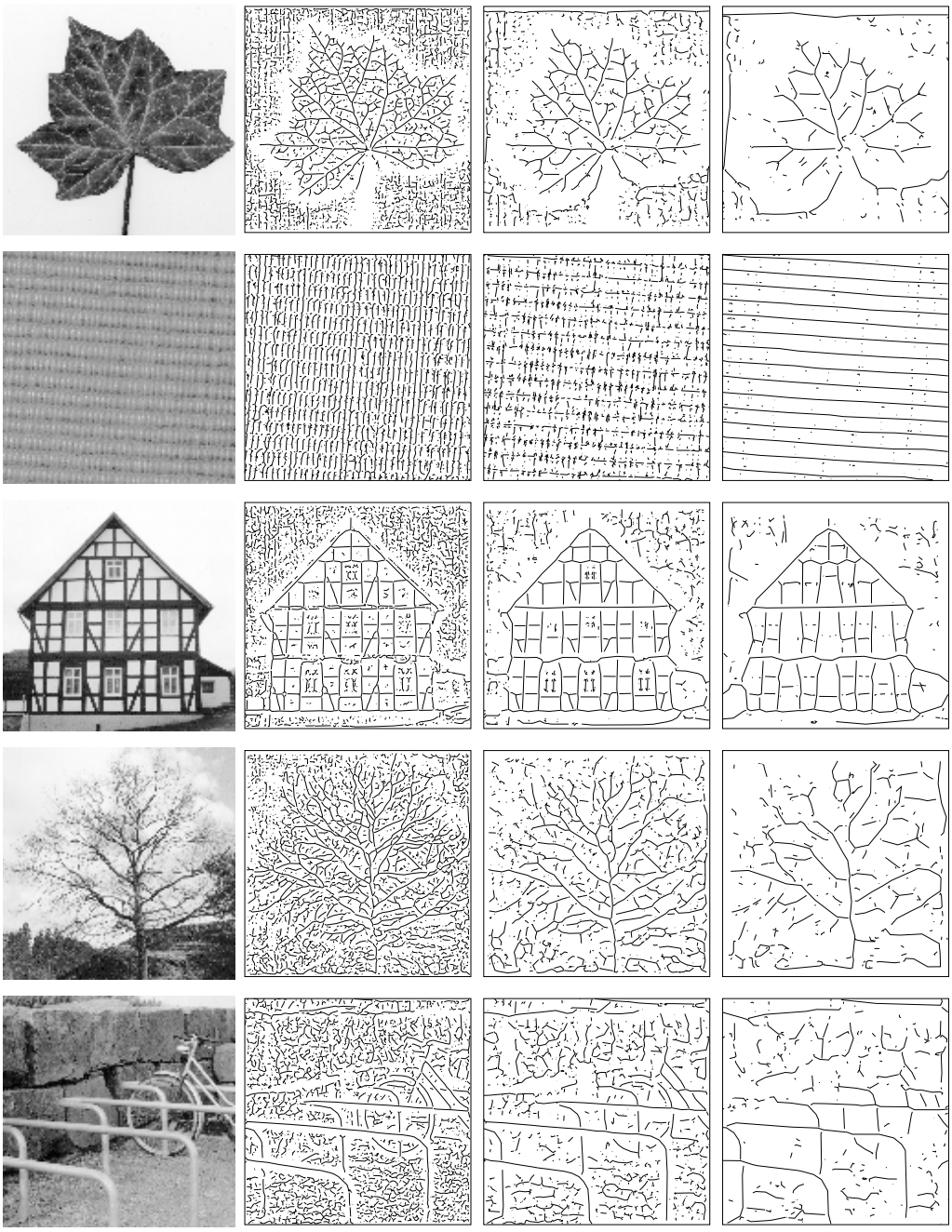


Figure 5.2: Height-ridge examples at different scales. The original images picture from top to bottom a leaf, a piece of cotton fabric, a German “Fachwerkhaus”, a tree, and bicycle stands with a bicycle. All images are 512 by 512 pixels. Ridges are shown for scale levels $\sqrt{t} = 2$, $\sqrt{t} = 4$, and $\sqrt{t} = 8$ (where a unit length is the width/height of a pixel).

5.3 Second Derivative Ridges

Second derivative ridges are characterized in a similar way as height ridges. At each point of a 2-dimensional image an axis is chosen along which some criterion must be fulfilled for the point to be on a ridge. The axis is the same as before, pointing along the direction of minimum second derivative. The criterion applied along the axis is different: a point is on a second derivative ridge if the *second derivative* along the chosen axis is minimal. The defining equations are:

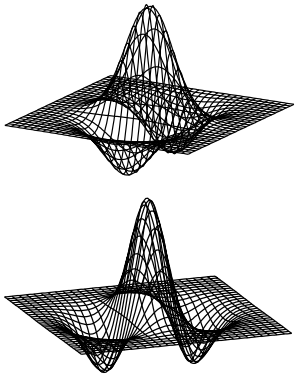
$$\begin{aligned} L_{ppp} &= 0 \\ L_{pppp} &> 0 \\ L_{pp} &< 0 \\ |L_{pp}| &\geq |L_{qq}| \end{aligned} \tag{5.3}$$

Examples of *second derivative ridges* in real images are shown in figure (5.3).

The use of higher derivatives leads to a number of differences between *second derivative ridges* and *height ridges*. Most apparent in the examples is the response to edges seen in second derivative ridges, most notably in the leaf image. This property of second derivatives was even exploited for edge detection by Marr and Hildreth [Marr and Hildreth, 1980]. It should, however, be noted that a response to edges does not occur at all scales, as seen in the "Fachwerkhaus" image. It will be shown in chapter 7 that *it is possible to automatically select just those scales where edges are not detected.*

Another difference between height ridges and second derivative ridges is that the latter may be interpreted as a least squares fit of the "ridge-model"

$$g_{\mathbf{d}}(\mathbf{x}) = -\partial_d \partial_d G(\mathbf{x}, t)$$



where $G(\mathbf{x};t) = \frac{e^{-\frac{\mathbf{x}^T\mathbf{x}}{2t}}}{(2\pi t)}$ is the rotation symmetric Gaussian and $\mathbf{d} \in \mathbb{R}^2$ a vector in $\mathbb{R}^2$ along which the second derivative is taken. Two views of the model are shown on the right. The direction $\mathbf{d}$ and the location of the model along this direction are determined by least squares [1]. For details see appendix B.

[1] To yield solutions of (5.3) those ridge-points found by the least square method where $|L_{pp}| < |L_{qq}|$ must be deleted.

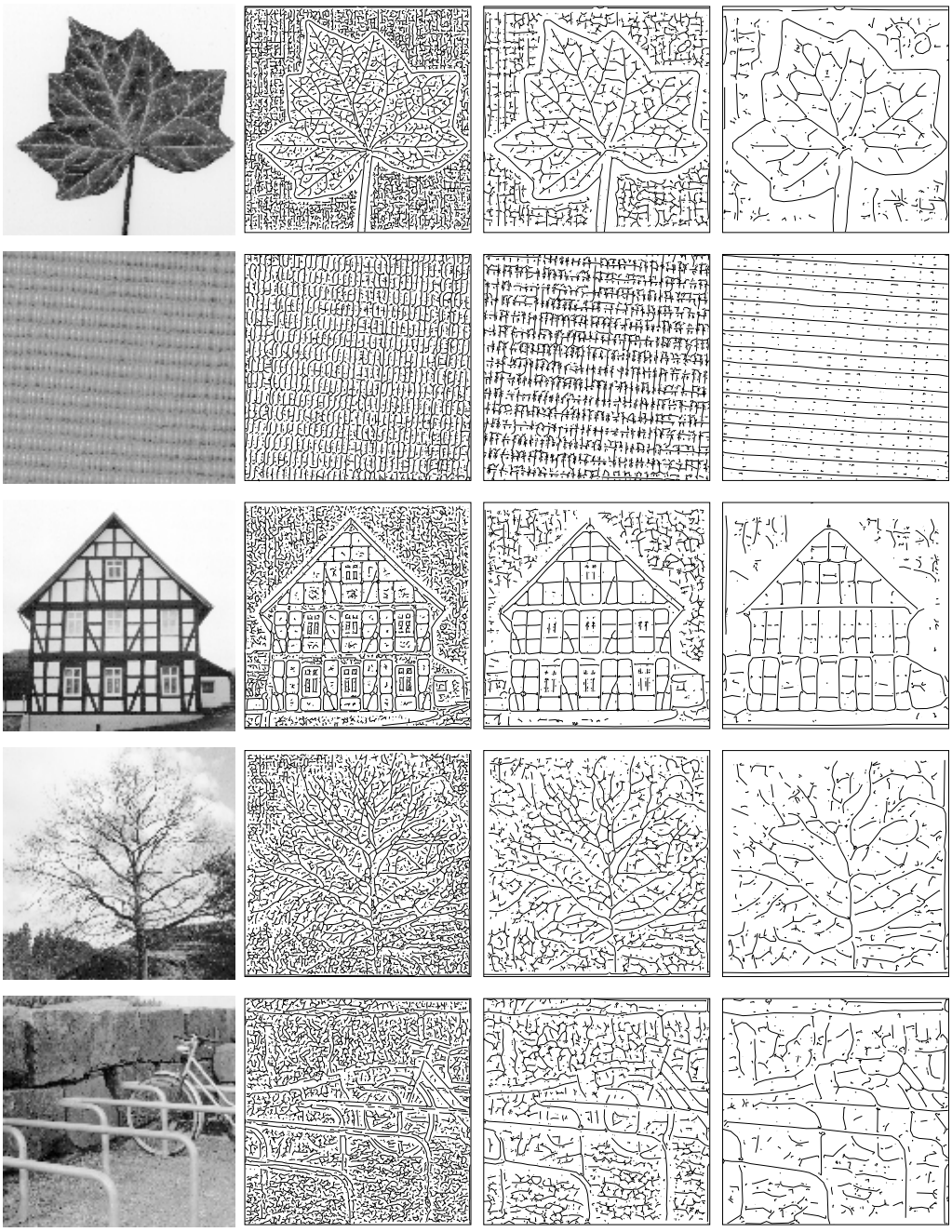


Figure 5.3: Second-derivative-ridge examples at different scales. The original images picture from top to bottom a leaf, a piece of cotton fabric, a German "Fachwerkhaus", a tree, and bicycle stands with a bicycle. All images are 512 by 512 pixels. Ridges are shown for scale levels $\sqrt{t} = 2$, $\sqrt{t} = 4$, and $\sqrt{t} = 8$ (where a unit length is the width/height of a pixel).

5.4 Computation of Ridges

The differential geometric definitions of ridges in terms of directional derivatives may give some intuitive understanding of height ridges and second derivative ridges. However, for the computation of ridges unfortunately they are not very useful. The remainder of the chapter is devoted to the technical problems of ridge computation and will demonstrate one reason why edges and not ridges are the more widely used features in computer vision: edges are much easier to compute.

5.4.1 Direction Discontinuities

The technical problems in ridge detection arise primarily from the discontinuity of the direction of minimum second derivative. The following figure gives an example. At each point the direction **p** of strongest downward bend (minimal second derivative) is displayed by the angle which **p** makes with a fixed direction (as shown on the right). Apparently **p** points radially in the region of the ring. Outside and inside this region the direction of **p** is orthogonal to radial. The direction makes a 90 degree flip between these regions. In addition there are some 180 degree flips along the vertical and horizontal.

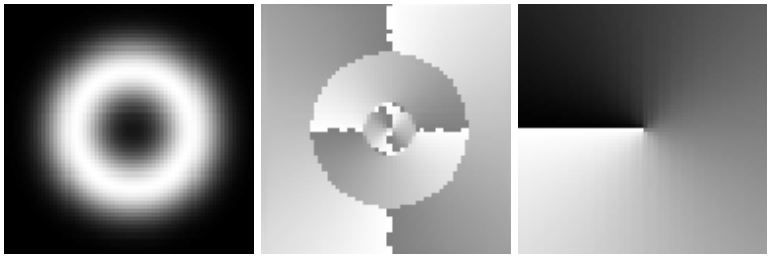


Figure 5.4: Direction discontinuities: The orientation of the vector **p** pointing along the axis of minimal second derivative at each pixel is shown in the center image. The orientation is given relative to a fixed direction as displayed on the right.

The discontinuities in the direction of **p** make the derivatives computed along **p** discontinuous as well. *This interferes with the computation of zero-crossings of e.g.* L_p *between neighboring pixels*: A 180 degree flip of **p** between two pixels changes the sign of L_p, thereby either creating or deleting a zero-crossing. Flips of 90 degrees sometimes interfere with zero-crossings and sometimes not. In any case, zero-crossings computed from derivatives along **p** do not reliably correspond to structural properties of the image .

5.4.2 Continuous Formulation

Alternative formulations of the ridge definitions (5.1), (5.2), and (5.3) based only on continuous quantities may be given provided that the landscape is at least twice continuously differentiable. [Eberly, 1996] discusses three equivalent definitions of a height ridge. [Lindeberg, 1998a] gives another definition. We consider these four ways to formulate the height ridge definition (5.2).

First note that the condition $L_p = 0$ for a point to be on a height ridge requires that **p** is orthogonal to the gradient of L (with the exception of *critical points* where the gradient vanishes). Consequently a necessary condition for a point to be on a height ridge is that the gradient coordinate system **u**, **v** (where **v** is the unit vector along the gradient and **u** is orthogonal to **v**) and the coordinate system **p**, **q** overlap. Such points may be characterized by

$$L_{uv} = 0 \qquad .$$

Figure (5.5) shows the zero-crossings of L_{uv} for a test image.

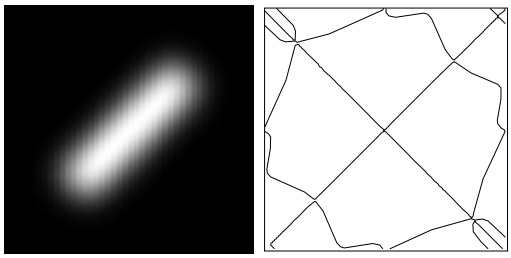


Figure 5.5: Zero-crossings of L_{uv}. At these points **p** and **q** lie along the local gradient coordinates **v** and **u**.

A point on a zero-crossing of L_{uv} is on a height ridge only if it is a maximum along **u** and the second derivative in this direction has a larger absolute value than that along **v**:

$$\begin{aligned} L_{uv} &= 0 \\ L_{uu} &< 0 \\ |L_{uu}| &> |L_{vv}| \end{aligned} \tag{5.4}$$

This definition is used in [Lindeberg, 1998a].

A problem that occurs with height ridges computed from zero-crossings of L_{uv} is that they tend to break up into separate pieces with small gaps. This happens at points where several zero-crossing lines of L_{uv} come very close to one another or even touch. At such points it is not generally clear how to continue a zero-crossing line (see Chapter 8).

To cope with the problems at the intersections of zero-crossings in the above formulation it is essential to get rid of those solutions to $L_{uv} = 0$ where $\mathbf{p}$ is parallel to $\mathbf{v}$ since these points cannot be on a ridge anyway. The points of interest are those where $\mathbf{p}$ is orthogonal to $\mathbf{v}$, i.e. $\mathbf{p}^T\mathbf{v} = 0$. [Eberly, 1996] describes how to reformulate $\mathbf{p}^T\mathbf{v} = 0$ with a continuous left hand side making use of the fact that $\mathbf{p}$ and $\mathbf{q}$ are eigenvectors of the Hessian matrix (see appendix A)

$$\begin{array}{l} \mathbf{H}\,\mathbf{p} = L_{pp}\mathbf{p} \\ \mathbf{H}\,\mathbf{q} = L_{qq}\mathbf{q} \end{array} \qquad \mathbf{H} = \begin{bmatrix} L_{xx} & L_{xy} \\ L_{yx} & L_{yy} \end{bmatrix}$$

with eigenvalues $L_{pp} \leq L_{qq}$. Orthogonality of $\mathbf{p}$ and $\mathbf{v}$ may then be written as either $\mathbf{H}\,\mathbf{u} = L_{pp}\mathbf{u}$ or $\mathbf{u}^T\mathbf{H}\,\mathbf{u} = L_{pp}$. The corresponding two definitions of height ridges are

$$\begin{array}{c} \mathbf{H}\,\mathbf{u} = L_{pp}\mathbf{u} \\ L_{pp} < 0 \\ |L_{pp}| \geq |L_{qq}| \end{array} \tag{5.5}$$

and

$$\begin{array}{c} \mathbf{u}^T\,\mathbf{H}\,\mathbf{u} = L_{pp} \\ L_{pp} < 0 \\ |L_{pp}| \geq |L_{qq}| \end{array} \tag{5.6}$$

It should be noted that L_{pp} and L_{qq} are continuous (if L is twice continuously differentiable) because they are the eigenvalues of the Hessian matrix and all elements of the Hessian matrix are continuous.

Figure (5.6) shows the zero-crossings of $\mathbf{H}\,\mathbf{u} = L_{pp}\mathbf{u}$. Clearly the ridge at the center of the image is seen to continue from the lower left to the upper right. This should be compared with the situation at the center of figure (5.5).

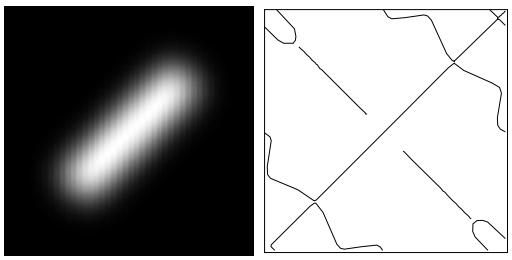


Figure 5.6: Zero-crossings of $\mathbf{H}\,\mathbf{u} - L_{pp}\mathbf{u}$. At these points $\mathbf{p}$ is orthogonal to the gradient $\mathbf{u}$.

Definitions (5.5) and (5.6) require different computational techniques. The first contains a system of two linearly dependent equations in $\mathbf{u}$ whose zero-crossings must be computed. The second requires minimization techniques since $\mathbf{u}^T\,\mathbf{H}\,\mathbf{u} \geq L_{pp}$. In the following section we describe how to solve (5.5).

5.4.3 Stable Solution

The ridge definition (5.5) contains the system of two equations

$$\mathbf{H}\,\mathbf{u} - L_{pp}\mathbf{u} = 0$$

that are linearly dependent in $\mathbf{u}$. (Recall that for 2-dimensional images $\mathbf{H}$ is a 2x2 matrix). One might thus be tempted to solve just one of the equations knowing that the other is then automatically solved as well. Unfortunately this reasoning is incorrect when the coefficients of the chosen equation all vanish so that any $\mathbf{u}$ is a solution to this equation however not necessarily to the other. To demonstrate this, figure (5.7) shows the zero-crossings from both left hand sides of the two equations computed for the test image. Clearly there are some positions where the solution to the first equation is not a solution to the second equation and vice versa. At these points all coefficients of one equation vanish.

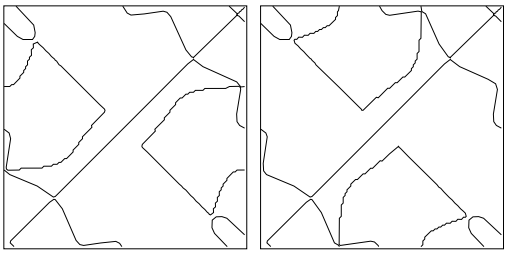


Figure 5.7: Zero-crossings of the both rows of $\mathbf{H}\,\mathbf{u} - L_{pp}\mathbf{u}$.

To find simultaneous zero-crossings of *both* rows of $\mathbf{H}\,\mathbf{u} - L_{pp}\mathbf{u}$ between two neighboring pixels we compute a stability measure for each row and then look for a zero-crossing of the row with greater stability. This is repeated for all pairs of neighboring pixels (see Chapter 8). To get a measure of stability we interpret the coefficients of each row of $\mathbf{H}\,\mathbf{u} - L_{pp}\mathbf{u}$ as a vector: $\mathbf{k}^1(\mathbf{x}) = (L_{xx}(\mathbf{x}) - L_{pp}(\mathbf{x}), L_{xy}(\mathbf{x}))$ and $\mathbf{k}^2(\mathbf{x}) = (L_{xy}(\mathbf{x}), L_{yy}(\mathbf{x}) - L_{pp}(\mathbf{x}))$. The projection of $\mathbf{k}^1(\mathbf{x}_0)$ at a point $\mathbf{x}_0$ onto $\mathbf{k}^1(\mathbf{x}_1)$ at a neighboring point $\mathbf{x}_1$ gives a measure of stability of the first row between the two points.

The resulting zero-crossings need not be closed as can be seen in figure (5.6). This requires modifications to the standard zero-crossing algorithms. Chapter 8 describes both the standard algorithms and their modification to deal with open zero-crossings.

5.4.4 Computation of Second Derivative Ridges

Having described the details of the computation of height ridges we give only a brief summary of the necessary modifications to compute second derivative ridges.

The zero-crossings of L_{ppp} are those points where $\mathbf{p}$ is orthogonal to the gradient of L_{pp}. Denoting the gradient of L_{pp} by $\mathbf{v}^{pp}$ and the orthogonal unit vector by $\mathbf{u}^{pp}$ a continuous formulation of the second derivative ridge definition is

$$\begin{aligned} \mathbf{H}\,\mathbf{u}^{pp} &= L_{pp}\mathbf{u}^{pp} \\ L_{pppp} &> 0 \\ L_{pp} &< 0 \\ |L_{pp}| &\geq |L_{qq}| \end{aligned} \tag{5.7}$$

To find $\mathbf{v}^{pp}$ and $\mathbf{u}^{pp}$ it is convenient to differentiate the analytical expression for the smaller eigenvalue L_{pp} of the Hessian matrix. The two eigenvalues are

$$\begin{aligned} L_{pp} &= \left(L_{xx} + L_{yy} - \sqrt{4L_{xy}^2 + (L_{xx} - L_{yy})^2}\right)/2 \\ L_{qq} &= \left(L_{xx} + L_{yy} + \sqrt{4L_{xy}^2 + (L_{xx} - L_{yy})^2}\right)/2 \end{aligned}$$

and the gradient of L_{pp} is

$$\begin{aligned} v_x^{pp} &= \frac{1}{2}\left(L_{xxx} + L_{yyx} - \frac{4L_{xy}L_{xxy} + (L_{xx} - L_{yy})(L_{xxx} - L_{yyx})}{\sqrt{4L_{xy}^2 + (L_{xx} - L_{yy})^2}}\right) \\ v_y^{pp} &= \frac{1}{2}\left(L_{xxy} + L_{yyy} - \frac{4L_{xy}L_{xyy} + (L_{xx} - L_{yy})(L_{xxy} - L_{yyy})}{\sqrt{4L_{xy}^2 + (L_{xx} - L_{yy})^2}}\right) \end{aligned}$$

The fourth order derivative L_{pppp} is computed from its Cartesian representation

$$L_{pppp} = p_x^4 L_{xxxx} + 4p_x^3 p_y L_{xxxy} + 6p_x^2 p_y^2 L_{xxyy} + 4p_x p_y^3 L_{xyyy} + p_y^4 L_{yyyy}$$

where $\mathbf{p} = (p_x, p_y)$ is a unit vector along the axis of minimum second derivative in the Cartesian frame.

Chapter 6

A Statistical Approach to Feature Detection and Scale Selection

A first step in the analysis of an image by a computer vision system or a biological vision system is to compute the response of some *local translation invariant operators*. These operators are usually constructed in such a way that local extrema of their response are "particularly informative".

As an example consider an "edge detector". Figure (6.1) demonstrates the use of a first derivative of Gaussian filter kernel as an "edge detector". Evidently the local extrema of the response correspond approximately to the edge locations.

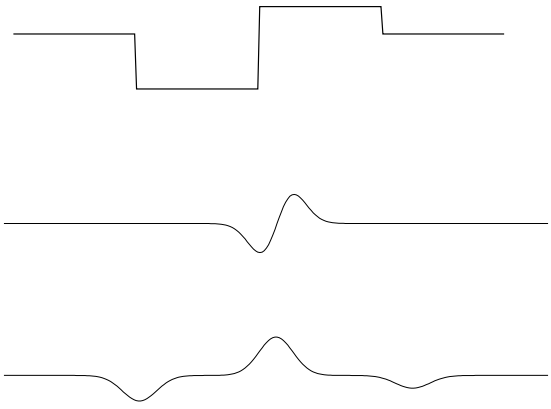


Figure 6.1: Edge detection. From top to bottom: image, filter kernel, response.

Naturally and necessarily any such local translation invariant operators have a *scale* or size, e.g. the above edge detector has a width approximately 1/20 of the

page width. So the question arises not only at which positions but also at which *scales* the operator response is “particularly informative”.

For example the “particularly informative” positions and scales of a “square detector” applied to the image on the right should reveal a *small* square in the upper right and a *large* square in the lower left. Generally, the aim of *feature detection* and *scale selection* together is to find “particularly informative” *(position,scale)-pairs*.

In this chapter we propose a method to generalize feature detection operators whose local extrema in space yield particularly informative scales to scale-selection operators whose local extrema in space and scale yield particularly informative (position,scale)-pairs. From the point of view we take, the generalization is canonical and there is no conceptual difference between feature detection and scale selection.

The chapter is organized as follows. First we describe our idea of how to define the “particularly informative”. Then a simple example is given to show how a square is detected and its size selected. Next we state the approach in more detail. Since this formulation is more general than the usual approach it is shown that the particularly informative positions are in a special case extrema of the operator response. Finally we consider scale selection and construct operators whose extrema with respect to space and scale yield “particularly informative” (position,scale)-pairs.

6.1 “Particularly Informative”

To arrive at a point of view from which particularly informative scales appear conceptually similar to particularly informative positions we need to take a detailed look at why or when it makes sense to define the local extrema of an operator response as particularly informative.

Suppose we are given an image, a local translation invariant operator and the statement that under otherwise identical conditions a large operator response is to be considered more informative than a smaller response. Is this sufficient to call positions where the operator response is locally maximal particularly informative? No, we were not told that the conditions at different positions in the image are to be considered identical. Of course it may be sensible to *assume* this and therefore to consider local extrema of the operator response particularly informative. [1]

[1] As an example of a context in which this assumption is usually not appropriate consider the analysis of image sequences or time series. Here the operator response is generally interpreted relative to responses to previous images.

The point of this is not that we intend to question the usefulness of treating different positions identically. Rather we intend to formulate this assumption concerning *positions* differently and in a way that allows one to formulate similar assumptions about *scales*.

Consider the following alternative *definition of particularly informative positions of an operator* by a three step procedure:

1. Destroy the *structural* information in an image by *shuffling* the pixels to new positions.

2. Measure how much information was destroyed at any position.

3. Label those positions where (locally) most information was destroyed as particularly informative.

Evidently the first two steps require a precise definition. Let us define *shuffling* as follows: First create a bin of intensity values into which the intensity of each pixel of the image is placed exactly once. Then, at any position **x** of the shuffled image randomly draw an intensity value from the bin of intensity values without replacement.

Secondly, let us define at any single position a measure of how much information was destroyed by shuffling: At a position **x** we compute the *operator response to the observed image* as well as the *distribution of operator responses to the shuffled images*. The response to a shuffled image will sometimes be larger and other times smaller than the observed response. The relative frequency with which the random responses are smaller than the observed response is taken as a measure of how informative the observed response is.

This measure might be motivated as follows. The random images contain less structural information *by construction*, so a measure of the structural information must in some way measure a deviation from the random images. The above relative frequency or probability is one possible such measure.

As far as particularly informative positions are concerned the definition via shuffled images is equivalent to the definition in terms of local extrema of the operator response. The shuffling approach has two advantages however: It applies immediately to particularly informative (position,scale)-pairs if one replaces all occurrences of “position” by “(position,scale)-pair”. Secondly the specifics may be modified, shuffling intensities differently or measuring destroyed information differently. A different shuffling method was recently considered in a different context by Koenderink and van Doorn [Koenderink and van Doorn, 1999].

The following section applies this approach to a toy situation.

6.2 A Working Example

Suppose we observed the image of figure (6.2) and intend to detect particularly informative positions of a "square detector". The square detector is defined as follows: it computes at any position $\mathbf{x}$ the number of white pixels [2] in a $l \times l$ square region centered about $\mathbf{x}$. l is a parameter of the square detector that we fix to be 15.

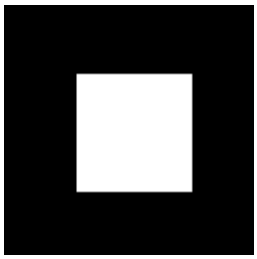


Figure 6.2: Observed image of 32×32 pixels with 225 white and 799 black.

The standard approach to find particularly informative positions is to apply the 15×15 square detector to any position in the image and label those positions particularly informative where the response is locally maximal. From the response shown in figure (6.3) it is evident that there is only one particularly informative position at the center of the image.

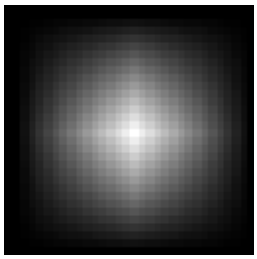


Figure 6.3: Operator response of 15×15 square detector.

The same result may be cast into our statistical interpretation: First all structural information is destroyed by shuffling the pixels to random positions. Two shuffled images are shown in figure (6.4). Then the square detector is applied to the random images and the probabilities of its different responses are observed. This is in principle done at each position of the image. For the shuffling method described above, however, the probabilities at all positions [3] are identical.

[2] The images contain only binary black and white pixels.

[3] With the exception of the image border

Figure 6.4: Shuffled images.

Figure (6.5) displays the probability of observing an operator response of *n* white pixels "under" the square detector at any position within a shuffled image.

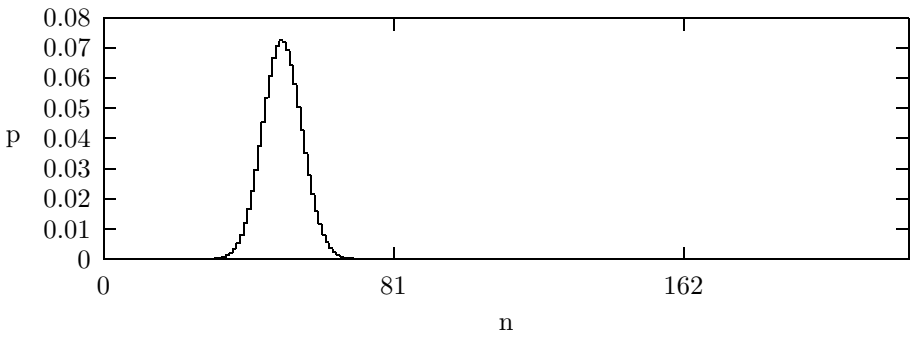


Figure 6.5: Probability of observing n white pixels in 225 pixels randomly drawn with replacement from 799 black and 225 white pixels.

Next, for any observed operator response we compute the probability with which larger responses occur in the shuffled images. Figure (6.6) shows this tail probability for an observation of 70 white pixels in the 15×15 region. Computing tail probabilities for the operator responses at all positions produces an image like figure (6.3) only with a single minimum at the center rather than a maximum. This means that the central position with the smallest tail probability is considered particularly informative according to our statistical approach.

As far as the computation of particularly informative positions are concerned the statistical approach yields the same positions as the maxima of the operator response.

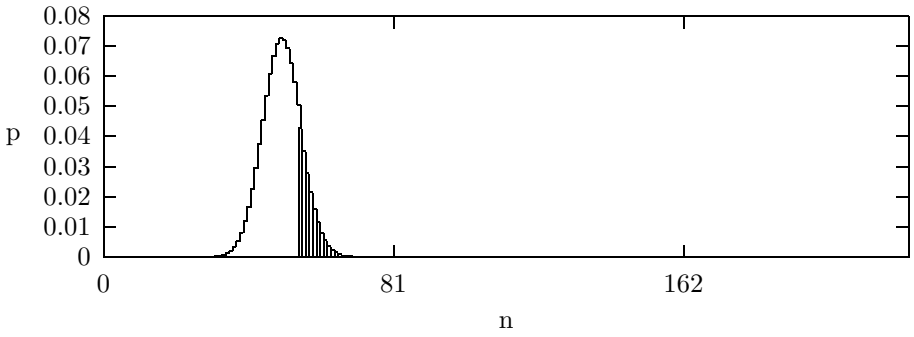


Figure 6.6: Tail probability for an observation of $n = 70$ white pixels in 225 pixels randomly drawn with replacement from 799 black and 225 white pixels. The tail probability is the area of the shaded region.

Let us now look at the determination of particularly informative scales.

Consider two square detectors of different size, 9×9 and 15×15 applied to the image of figure (6.7). For simplicity we apply them only at the centers of the square since we already know that these are the most informative positions.

Figure 6.7: 64×64 image with four 9×9 white squares.

The 9×9 square detector responds to a center of square position with a value of 81 and the 15×15 detector responds with a value of 81 as well. Obviously the responses alone do not permit to call one size more informative than another.

Applying the statistical approach again, we first shuffle the images and then compute the operator response to the shuffled images at the square positions. These are random variables that have the probabilities shown in figures (6.8). This time however *the probabilities for the two different sizes are not the same*. Consequently the tail probabilities of the responses of 81 are different for the 9×9 and the 15×15 square detector. Unfortunately the numerical differences are very small and not immediately apparent from figure (6.8). [4] If we remark

[4]The fact that the numerical differences of tail probabilities are sometimes very small need not

that the probabilities are positive anywhere within the displayed range of operator responses n it becomes apparent that the 9×9 square detectors response of $n = 81$ has a smaller tail probability than the 15×15 detectors 81. By definition we thus call the smaller size more informative.

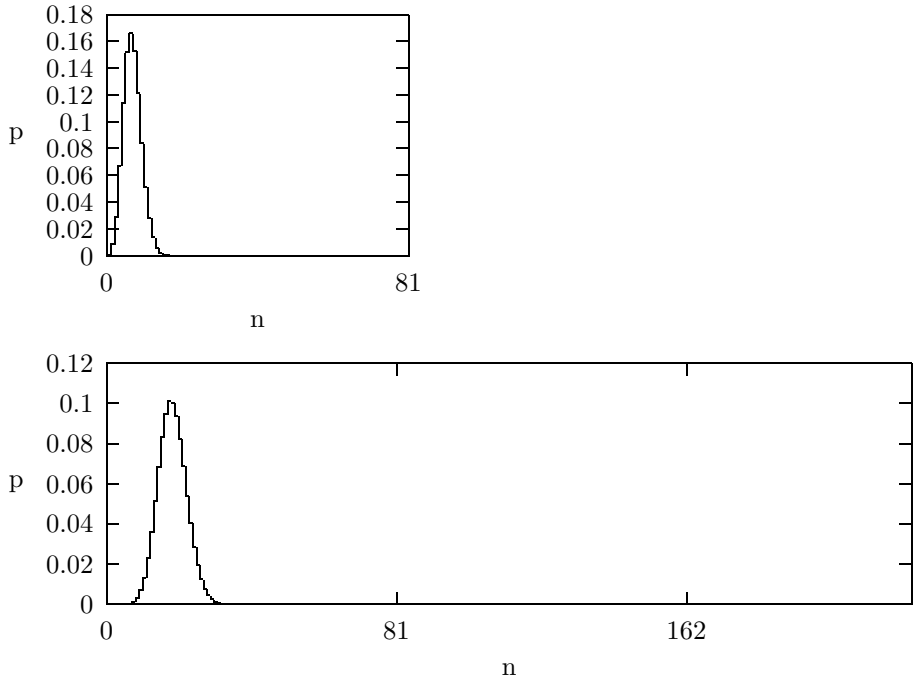


Figure 6.8: Probability of observing n white pixels in 64 and 256 pixels randomly drawn with replacement from 768 black and 256 white pixels.

Finally consider changing the response of the 15×15 square detector for example as follows: instead of computing the number of white pixels below the 15×15 region let it compute the average intensity below this region, i.e. the number of white pixels divided by 15^2. This produces a response of $81/225$ to the center of a square in figure (6.7). It is easily seen however that the result of the statistical approach is unaffected by this. Only the *shape* of a feature detector is relevant to the approach.

These example should demonstrate that our statistical definition of "particularly informative" is at least capable of distinguishing between different positions *and different scales*. Furthermore it is equivalent to the maximum operator response as far as positions are concerned and it appears to produce useful results concerning scales.

concern us because we do not actually propose to compute them as the following sections will show.

6.3 A Statistical Approach

Suppose we are given some feature detection operator. This operator produces a scalar response at different positions of an image and one may ask at which positions the response is particularly informative. Let us assume particularly informative positions are defined as local maxima (with respect to position) of the operator response. If the operator has further parameters such as a scale naturally the question arises which (position,scale)-pairs are particularly informative or in general which parameters are particularly informative.

We now propose a definition of particularly informative parameters that is i) more general than the "maxima of operator response" definition ii) is equivalent as far as particularly informative positions are concerned and iii) can be implemented.

The key issue concerning implementation is that we derive from the given operator new operators whose local extrema yield the particularly informative parameters, position, scale or other. To do so we replace the shuffling of pixel intensities in the approach of section 6.1 by a stochastic model which will be called a *sampling model*. The idea of shuffling was to create images with less structural information so that the structural information of the observed image may be evaluated relative to the shuffled images. The prototypical sampling model that is void of structural information is a *normal white noise* where the intensity at each position is drawn from a normal distribution independently of intensities at other positions.

The mean and standard deviation of the sampling model may be set to the average intensity and standard deviation of the observed image. If however as in scale-space theory one requires invariance of particularly informative parameters with respect to linear intensity transformations, then the choice of mean and (nonzero) standard deviation is irrelevant.

Let us now define particularly informative parameters in a way that is independent of the kind of parameter, position, scale or other. Later we derive from this definition operators whose local extrema yield the particularly informative parameters.

6.3.1 Definition of "Particularly Informative" Parameters

Given some *operator* D_θ that computes for any parameter θ a scalar response from an image f:

$$D_\theta(f) \in \mathbb{R}$$

where the parameter may be position, $\theta = \mathbf{x}$, scale, $\theta = \sqrt{t}$, position and scale $\theta = (\mathbf{x}, \sqrt{t})$ or others. Given also a *sampling model* from which random images may

be sampled that are by definition considered less informative than the observed image. Denote the probability density of the sampling model by p and the random images by ξ.

Applying D_θ to random images ξ creates a distribution of operator responses for each θ. Call the densities of these distributions p_{D_θ}.

For any observed response $D_\theta(f)$ to an observed image f compute the probability by which random images from the sampling model produce larger responses $D_\theta(\xi) > D_\theta(f)$ and call this probability $P(\theta, f)$

Particularly Informative Parameters: *On the basis of a sampling model $p(\xi)$ the "particularly informative parameters" θ of the operator responses $D_\theta(f)$ to an observed image f are the local minima of $P(\theta, f)$ with respect to θ.*

The motivation for this definition is the "information-less" character ascribed to images drawn from the sampling model. The probability $P(\theta, f)$ is one possibility to measure the difference between the observed operator response and the "information-less" responses.

To compute $P(\theta, f)$ the following equations must be evaluated. The probability density p_{D_θ} of operator responses to images sampled from p is

$$p_{D_\theta}(v) = \int d\xi \, p(\xi) \, \delta(D_\theta(\xi) - v)$$

where δ stands for the Kronecker Delta function ($\delta(D_\theta(\xi) - v) = 1$ if $D_\theta(\xi) = v$ and 0 otherwise). Then the probability $P(\theta, f)$ is the *tail-probability* of p_{D_θ}:

$$P(\theta, f) = \int_{D_\theta(f)}^{\infty} dv \, p_{D_\theta}(v) \tag{6.1}$$

shown in figure (6.9)

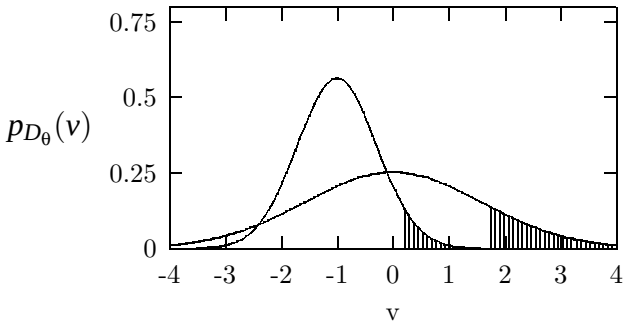


Figure 6.9: Tail probabilities for two observations from different distributions. The probabilities correspond to the shaded areas.

6.3.2 Sampling Models

The purpose of the sampling model is to produce images that are considered less informative than the observed image such that the observed image in return may be evaluated relative to the sampling model. Normal white noise images are certainly a prominent candidate among possible sampling models. Other choices may be useful as well, depending on the situation.

In the context of *image sequences* it is natural to interpret a new image on the basis of the present information about the scene. The sampling model may then describe the prior information about the scene that is known from previous images. The particularly informative parameters are then determined on the basis of the (known and in that sense less informative) prior information. A sampling model that captures the prior information from previous images of a sequence may obviously be quite different from white noise. (For a review of techniques to analyze image sequences see [Barron et al., 1994].)

The sampling model may also describe prior information about the images that a visual system expects to see in its environment. The ensemble of *natural images* for example is not white noise as demonstrated by Field [Field, 1987] and Ruderman and Bialek [Ruderman and Bialek, 1994]. Their model will be considered below.

In the context of a single image without prior information normal white noise appears a useful sampling model. *White* noise guarantees lack of structural information. The choice of distribution could be varied to other distributions than normal. Since however any feature detection operator locally integrates the image intensity operator responses to non-normal sampling models tend toward normal except at very small scales. Thus we prefer to use a normal distribution right away.

6.4 Feature Detection and Homogeneous Sampling Models

In this section we show that the statistical approach produces the same particularly informative *positions* as the maximum operator response of a feature detector if the *sampling model* is *homogeneous*.

Observe that the tail probability defined in equation (6.1) becomes a monotonic transformation of the operator response $D_\theta(f)$ if the integrand is independent of θ. In that case the minima of $P(\theta, f)$ with respect to θ are the maxima of $D_\theta(f)$.

To show that the probability density p_{D_θ} of the operator response is independent of θ when D_θ is a feature detector and the sampling model is homogeneous let us recall or state the definitions of the involved terms.

A feature detector is a local translation invariant operator $D_{\mathbf{x}}$ parameterized by the position $\mathbf{x}$ where it is applied to the image. Translation invariance of the feature detector refers to $D_{\mathbf{x}}(T_{\mathbf{a}} \circ f) = D_{\mathbf{x}-\mathbf{a}}(f)$ where $T_{\mathbf{a}}$ denotes the translation operator that moves an image f by the vector $\mathbf{a}$, i.e. $T_{\mathbf{a}} \circ f(\mathbf{x}) = f(\mathbf{x}+\mathbf{a})$.

A homogeneous sampling model is translation invariant in the sense that any sampled image ξ and the translated image $T_{\mathbf{a}}\xi$ have the same probability density $p(\xi) = p(T_a \circ \xi)$, neglecting difficulties due to image border.

For homogeneous sampling models $p(\xi)$ the particularly informative positions of a translation invariant operator D_x are the local maxima of $D_x(f)$.

Proof: If both the operator and the distribution of images to which it is applied are translation invariant, so is the distribution of responses:

$$\begin{aligned} p_{D_x}(v) &= \int d\xi\, p(\xi)\, \delta(D_x(\xi) - v) \\ &= \int d\xi\, p(T_a \circ \xi)\, \delta(D_x(T_a \circ \xi) - v) \\ &= \int d\xi\, p(\xi)\, \delta(D_{x+a}(\xi) - v) \\ &= p_{D_{x+a}}(v) \end{aligned}$$

Furthermore, if p_{D_x} does not depend on x, then the minima with respect to x of

$$P(D_x(f)) = \int_{D_x(f)}^{\infty} dv\, p_{D_x}(v)$$

are the maxima of $D_x(f)$.•

From our point of view this "justifies" the standard approach to feature detection that makes no mention of a sampling model. Whenever the sampling model is homogeneous all other aspects of the model are *irrelevant* to feature detection. Homogeneity in return appears natural whenever no prior information about preferred positions is available.

6.5 Scale Selection for Derivative of Gaussian Operators

In this section we consider families of operators that are parameterized by *position* x and *scale* $\sqrt{t}$: the derivative of Gaussian filter kernels

$$G^{\mathbf{n}}(\mathbf{x};t) = \partial_1^{n_1} \cdots \partial_N^{n_N} G^0(\mathbf{x};t)$$

of order $n = n_1 + \cdots + n_N$ where $G^0(\mathbf{x};t)$ is the rotation symmetric N-dimensional Gaussian. The operator response to an image f is

$$D^{\mathbf{n}}_{\mathbf{x},t}(f) = (G^{\mathbf{n}}(\circ;t) * f)(x) = \int dy \quad G^{\mathbf{n}}(y-x;t)\, f(y)$$

The previous section showed that on the basis of a homogeneous sampling model one may at any fixed scale compute particularly informative positions in terms of local maxima of the operator response. We now address the problem of also *selecting particularly informative scales* within the statistical approach. Here too it is possible to avoid the computation of probability densities and tail probabilities: For each derivative of Gaussian operator there is a so called γ-normalized derivative of Gaussian operator with the useful property that the maxima of the response of the γ-normalized operator occur at the positions and scales where the tail probabilities of the derivative of Gaussian operator are minimal.

In contrast to feature detection however there is no universal γ-normalized operator for a large class of sampling models. Several sampling models shall be considered: a normal white noise model, a fractal Brownian motion model that describes the statistics of natural images, and a model that has been used in image restoration.

6.5.1 Scale Selection on the Basis of a Normal White Noise Sampling Model

As a sampling model consider images ξ where at each position x the intensity $\xi(x)$ is sampled independently from normal random variables of zero mean and standard deviation σ.

On the basis of a normal white noise sampling model the particularly informative scales ((position,scale)-pairs) of the derivative of Gaussian filter $G^{\mathbf{n}}(\mathbf{x};t)$ are the local maxima with respect to scale (position and scale) of the γ normalized derivative of Gaussian response

$$(\, t^{\gamma} G^{\mathbf{n}}(\circ;t) * f \,)(\mathbf{x})$$

with

$$\gamma = n/2 + N/4 \tag{6.2}$$

Proof: The response $\xi^{\mathbf{n}} = G^{\mathbf{n}}(\cdot;t) * \xi$ of filtering $\xi(x)$ with the derivative of Gaussian kernel $G^{\mathbf{n}}(\cdot;t)$ is a normal colored noise of zero mean and autocovariance

$$\rho^{\mathbf{n}}(\mathbf{x}-\mathbf{x}', t+t') = \sigma^2 (-1)^n \, G^{2\mathbf{n}}(\mathbf{x}-\mathbf{x}'; t+t')$$

that describes the covariance of $\xi^{\mathbf{n}}(\mathbf{x};t)$ and $\xi^{\mathbf{n}}(\mathbf{x}';t')$.

The marginal density $p_{G^{\mathbf{n}}(\mathbf{x};t)}$ at any single position and scale $(\mathbf{x};t)$ is a univariate normal with zero mean and standard deviation

$$t^{-n/2}t^{-N/4}\sigma$$

Evidently multiplication of the operator response by $t^{n/2}t^{N/4}$ makes the marginal density $p_{t^{n/2}t^{N/4}G^{\mathbf{n}}(\mathbf{x};t)}$ at $(\mathbf{x};t)$ of the new response independent of $\mathbf{x}$ *and* t. Then the minima of the tail probability $P(\mathbf{x},t,f)$ of equation (6.1) are the maxima of the operator response $(t^{n/2}t^{N/4}G^{\mathbf{n}}(\circ;t) * f)(\mathbf{x})_{\bullet}$

The γ-normalized derivative of Gaussian operators were introduced by Lindeberg to select particularly informative scales [Lindeberg, 1993b]. The specific choice of γ varies between applications [Koller, 1995], [Lindeberg, 1998b], [Lindeberg, 1998a], [Lorenz et al., 1997a], [Pizer et al., 1998].

6.5.2 Scale Selection with a "Natural" Sampling Model

Studies of natural images [Ruderman and Bialek, 1994] have revealed a power spectrum that is, not surprisingly, different from that of normal white noise. These studies were conducted on ensembles of images of natural scenes. [Ruderman and Bialek, 1994] took images in a New Jersey state park. [Field, 1987] used images of trees, rocks, bushes, and water.

The power spectrum $\tilde{\rho}(\omega)$, the Fourier transform of the auto-covariance, was estimated to be of the form

$$|\tilde{\rho}(\omega)|^2 \propto |\omega|^{-\alpha}$$

corresponding to so-called N-dimensional *fractal Brownian motion* [Pentland, 1984]. The interesting property of this spectrum is that like white noise it has a scaling invariant correlation [Ruderman and Bialek, 1994] [Field, 1987] [Pentland, 1984].

The standard deviation of the response of a derivative of Gaussian operator to normal noise images ξ with a fractal Brownian motion spectrum can be computed by Parseval's theorem [Steenstrup et al., 1999] [Lindeberg, 1994b, section 13.7]

$$\begin{aligned}
\int d\mathbf{x}|G^{\mathbf{n}}(\mathbf{x};t)\xi(\mathbf{x})|^2 &= \int d\omega\, |\tilde{G}^{\mathbf{n}}(\omega;t)\tilde{\xi}(\omega)|^2 \\
&= \int d\omega\, |\omega|^{2n} e^{-t|\omega|^2} |\omega|^{\alpha} \\
&= \int_{r\in[0,\infty[;\phi_1,\ldots,\phi_N\in[0,2\pi]} dr\, d\phi_1 \cdots d\phi_N r^{N-1} e^{-r^2 t} r^{2n-\alpha} \\
&\propto t^{\alpha/2-n-N/2}
\end{aligned}$$

Again multiplication of the operator response with $t^{-\alpha/4+n/2+N/4}$ makes the marginal density at $(\mathbf{x};t)$ of the new response independent of $\mathbf{x}$ *and* t. This allows one to compute particularly informative positions *and* scales as *local maxima of the operator response of*

$$t^{\gamma}G^{\mathbf{n}}(\mathbf{x};t)$$

with $\gamma = -\alpha/4 + n/2 + N/4$.

Lindeberg already noted that in some cases "the normalized derivative model is *neutral* with respect to power spectra of the form $|\omega|^{-2}$" [Lindeberg, 1994b, section 13.7]. It is our proposal that this should not be viewed as a coincidence but that one should choose a sampling model and then adjust γ *in order to* make it neutral with respect to the sampling model.

One "natural" choice of sampling model is the distribution of natural images. Estimated values of α for different natural scenes lie in the range between $\alpha = 1.81 \pm .01$[Ruderman and Bialek, 1994] or between 2 and 3 [Steenstrup et al., 1999]. Interestingly, Steenstrup Pedersen and Nielsen estimate α *via* estimation of γ.

6.5.3 An Image Restoration Model

The white noise model is certainly not a realistic model for "a priori" expected data. One possible improvement was considered in the previous paragraph. Another is to take into account that the measurement device generally blurs the true scene and introduces noise.

For any true scene g consider the data f to be generated by

$$f = k * g + \eta$$

where k is a filter kernel that describes *observational blur* and η is a normal white noise of zero mean and standard deviation ϕ. Assuming now for the true image a normal white noise density $p(g)$ (that is independent of η) of zero mean standard deviation ψ allows one to compute the sampling density $p(\xi)$ of the *a priori* expected data ξ. It is also a normal random field with zero mean and auto-covariance

$$\rho(\mathbf{x}) = \psi^2 k * k(\mathbf{x}) + \phi^2 \delta(\mathbf{x})$$

where $\delta(0) = 1$ and $\delta(\mathbf{x}) = 0 \; \forall \, \mathbf{x} \neq 0$.

The response of a derivative of Gaussian operator $G^{\mathbf{n}}(\cdot;t)$ to samples from this sampling density is again normal of zero mean and auto-covariance

$$\rho(\mathbf{x}) = \psi^2 (G^{\mathbf{n}}(\cdot;t) * k) * (G^{\mathbf{n}}(\cdot;t) * k)(\mathbf{x}) + \phi^2 G^{\mathbf{n}}(\cdot;t) * G^{\mathbf{n}}(\cdot;t)(\mathbf{x})$$

The marginal density of $p_{G^{\mathbf{n}}(\mathbf{x};t)}$ for a single pair $(\mathbf{x};t)$ is a univariate normal of zero mean and variance

$$\sigma(t)^2 = \psi^2 (G^{\mathbf{n}}(\cdot;t) * k) * (G^{\mathbf{n}}(\cdot;t) * k)(0) + \phi^2 G^{\mathbf{n}}(\cdot;t) * G^{\mathbf{n}}(\cdot;t)(0)$$

To compute locally most informative scales one introduces γ-normalized derivatives

$$\frac{1}{\sigma(t)} G^{\mathbf{n}}(\mathbf{x};t)$$

so that the marginal density of the operator response at $(\mathbf{x};t)$ is independent of $\mathbf{x}$ and t and the local maxima of the operator response correspond to locally most informative positions and scales.

Within this model scale selection requires knowledge of k and ψ/ϕ. k and ϕ are properties of the measurement device and thus generally known to the visual system. It is then still necessary to assume a definite value for the variance ψ^2 of the density $p(g)$. [Galatsanos and Katsaggelos, 1992] and [Archer and Titterington, 1995] present some methods for estimating ψ/ϕ from the data.

Only in the limiting case where the noise of the measurement device becomes negligible, $\phi = 0$, the convenient situation arises that informative scales are independent of the variance ψ.

6.5.4 Line-like Structures and Sub-Dimensional Frames

Frequently the aim of feature detection is not to find isolated positions but rather lines. In that case the aim of scale selection can only be to find the particularly informative scale *across the structure* and not along the structure.

To determine line-like structures using local operators generally requires a two step-procedure. First at each position a direction is chosen along which the structure should extend if it went through that position. Then along the perpendicular direction across the hypothetical structure it is checked whether the point under consideration is on the structure or not. Essentially *the detection of particularly informative positions and scales occurs only in the sub-dimensional frame across the structure.* For example to detect one-dimensional ridges in two-dimensional images one detects positions and scales on a ridge in a one-dimensional frame across the hypothetical ridge direction at each point of an image, as described in chapters 5 and 7.

The use of sub-dimensional frames affects the γ-normalization parameter. The dimension N in $\gamma = n/2 + N/4$ should be the dimension of the sub-dimensional frame across the structure, i.e. $N = 1$ for the detection of ridges in two-dimensional images.

6.6 Discussion

Feature detection is so standard that it is "obvious" to determine the particularly informative positions in terms of local maxima of a feature detectors response. However, when it comes to the determination of other particularly informative parameters of an operator this "obvious" approach either does not work or is not obvious.

We have presented a point of view that defines any particularly informative parameters of an operator in conceptually the same way. The idea is to evaluate an operator response *relative to* the operator response to random images that contain by construction/definition less structural information. Loosely speaking this lets those operator responses stand out where there is particularly much structural information of the type that the operator responds to. One way to construct images with less structural information than the observed image is to *shuffle* the pixel intensities to random positions. Another way is to define a *sampling model* from which the random images are drawn. The prototypical sampling model that lacks any structural information is a normal white noise.

From our point of view the "obvious" computation of particularly informative *positions* in terms of local maxima of the operator response is appropriate when the sampling model is homogeneous.

Concerning *scale selection* of derivative of Gaussian operators different sampling models lead to different γ-*normalized derivative operators* whose local maxima with respect to scale correspond to particularly informative scales. From the described point of view this explains why scale selection is not "obvious".

We believe that the *normal white noise* sampling model should present a useful basis for feature detection and scale selection. In the following chapter we study ridge detection with scale selection for a γ-value corresponding to the white noise sampling model.

6.7 Outlook: Nonlinear Scale-Space

The presented approach to feature detection and scale selection may in principle also be applied to feature detection and scale-selection in nonlinear scale-spaces. These scale-spaces are constructed as solutions to the diffusion equation

$$\partial_t L(\mathbf{x};t) = \sum_{i,j=1}^{N} \partial_i \partial_j c^{ij} \, L(\mathbf{x};t)$$

with initial condition $L(\mathbf{x};0) = f(\mathbf{x})$ where f is the observed image [Weickert, 1998]. For constant diffusion coefficients c^{ij} the solution to this partial

differential equation, also known as Fokker-Planck equation [Honerkamp, 1990], is the linear scale-space. When the diffusion coefficient depends on $L(\mathbf{x};t)$ the solution is a nonlinear scale-space.

The derivatives of scale-spaces in general may be interpreted in terms of filter kernels applied to the original image. In linear scale-space these filter kernels are derivatives of Gaussians. In nonlinear scale-spaces they can be of very complicated shapes. In linear scale-space the filter kernels at different positions but identical scale all have the same shape and size, differences in kernel size occur only between scales. In nonlinear scale-spaces the shape and size of kernels may generally vary even between neighboring positions at the same scale. This complicates feature detection and scale selection in nonlinear scale-spaces.

It may be useful to compare the observed response of a derivative operator in nonlinear scale-space to the distribution of responses to random images with less structural information as presented above for the linear scale-space.

Chapter 7

Ridge Detection with Scale Selection

This chapter extends the analysis of elongated bright structures on a dark background or dark structures on a bright background in such a way that *the central line and the width of a structure are jointly determined*. We refer to this as *ridge detection with scale selection* in contrast to ridge detection at fixed scales that was treated in chapter 5.

The need for scale selection is best demonstrated by some examples of ridges computed at fixed scales. Figure (7.1) shows fixed scale ridges of a grass, a cotton fabric, and a synthetic image at three different scales. Clearly, at small scales the thick leaves of the grass are not detected while at large scales the thin leaves (particularly on the left) and the stem escapes detection. The synthetic example and the cotton fabric serve to illustrate the fact that *different scale structures may occur at the same position*. At any point in these images small scale structures run almost vertically and large scale structures run almost horizontally. To capture structures of all scales evidently the data must be analyzed at all scales.

The definition of a *scale-space ridge* discussed in this chapter is closely related to the *multi-scale line filter* proposed by Lorenz et al.[Lorenz et al., 1997b] [Lorenz et al., 1997a], the concepts of *cores* and *medialness* introduced by Pizer and coworkers [Fritsch et al., 1994] [Pizer et al., 1998] as well as the ridge concept discussed by Lindeberg in [Lindeberg, 1998a] and [Lindeberg, 1996] where the term *scale-space ridge* was first introduced. All these approaches build on the idea of γ-normalization introduced by Lindeberg, however, they generally employ smaller values of γ (than proposed here) in order to achieve a one to one association between scale and width.

The chapter is organized as follows. The first section outlines the problems encountered when analyzing data at all scales. The approach taken here is to find particularly informative (position,scale)-pairs in terms of the statistical interpretation described in the preceding chapter. This extends the fixed scale *height ridge* and *second derivative ridge* definitions to variable scales. Some examples

of scale-space ridges are shown for synthetic and real images and differences to ridge detection at fixed scales are discussed. An interesting difference of the second derivative scale-space ridge to its fixed scale cousin is that it produces no "false" responses to edges. This phenomenon is analyzed in detail in section 7.4. Finally the *ridge surface* and *scale-selection surface* of a simple model image are shown to give the reader a geometric impression of the quantities involved in the computation of scale-space ridges.

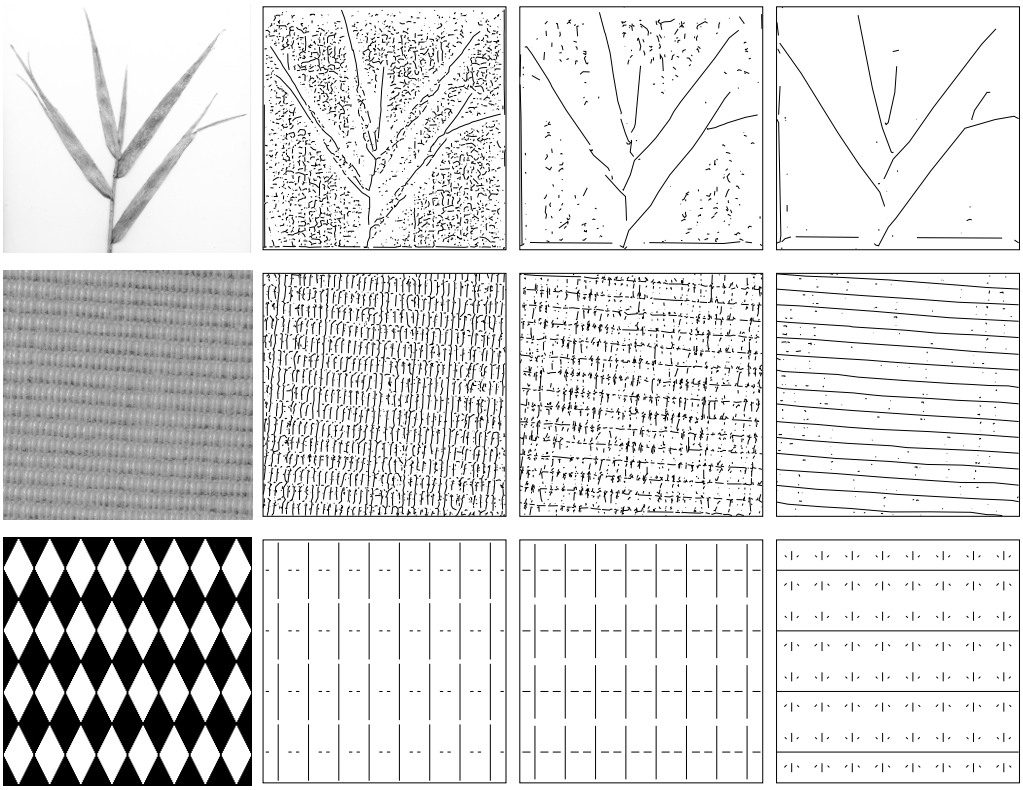


Figure 7.1: Ridges at fixed scales. All three images are 512 by 512 pixels and the chosen scale levels are $\sqrt{t} = 2$, $\sqrt{t} = 6$, $\sqrt{t} = 12$ for the grass image, $\sqrt{t} = 2$, $\sqrt{t} = 4$, $\sqrt{t} = 8$ for the cotton fabric and $\sqrt{t} = 8$, $\sqrt{t} = 16$, $\sqrt{t} = 24$ for the diamonds (in units of 1 pixel width/height).

7.1 The Scale Dimension

If one wishes to take account of the possibility that individual points in an image may be endowed with more than one structure at different unknown scales then obviously the data must be analyzed at all scales. As a direct consequence of this the problem is moved to a higher dimensional space, namely scale-space. This

creates two types of problems. First, the question arises, as to how one can go about detecting particularly informative (position,scale)-pairs in scale-space. We have presented *one* possible approach to this problem in the previous chapter. Based on a statistical interpretation it is possible to take the existing methods for feature detection into the higher dimensional scale-space in a canonical way that treats scale and spatial variables on equal footing.

The second problem concerns the interpretation of results. These too "live" in a higher dimensional space. The ridges of two-dimensional images that will be shown at the end of the chapter are lines in a three-dimensional space. Aside from the technical issues of how to display three dimensional data the more profound problem is that our perception is not trained to interpret these data in any way near as well as it is trained to interpret two-dimensional images. One way to facilitate visual interpretation that will be illustrated here is to project the detected ridges onto the spatial plain. The results are superior to the fixed scale ridges displayed above, at least in certain respects to be discussed.

7.2 Definitions of Scale-Space Ridges

This section describes how ridge detection at fixed scales may be extended to variable scales. We first summarize the computation of ridges at fixed scales as described in chapter 5. This is given a statistical interpretation which canonically generalizes to variable scales, resulting in the definitions of *scale-space ridges* (7.1) and (7.2). These definitions are similar to those suggested and used previously [Lindeberg, 1998a], [Lindeberg, 1998b], [Lorenz et al., 1997b], [Staal et al., 1999].

7.2.1 Ridges at Fixed Scales

At any fixed scale $\sqrt{t}$ of scale-space $L(\mathbf{x},t)$ the *height ridge* and *second derivative ridge* considered in detail in chapter 5 are defined in two steps. First at each image point the direction of a hypothetical ridge is defined to be perpendicular to the direction $\mathbf{p}$ of minimal second derivative L_{pp}. Traversal of the image intensity along $\mathbf{p}$ should thus display the *ridge profile*, if there is a ridge at the point under consideration. The criterion that defines a point on a *height ridge* is that it should be a local maximum along $\mathbf{p}$ and the second derivative L_{pp} along $\mathbf{p}$ should be greater than the second derivative L_{qq} along the orthogonal direction $\mathbf{q}$ [1].

A point lies on a *second derivative ridge at fixed scale* if the second derivative along $\mathbf{p}$ has a local minimum at that point and additionally $L_{pp} < 0$ and $|L_{pp}| >$

[1] The equations that define a *height ridge at fixed scale* in two dimensions may be written as: $L_p = 0, L_{pp} < 0, |L_{pp}| \geq |L_{qq}|$.

$|L_{qq}|$ are satisfied [2].

7.2.2 Fixed Scale Ridges in a Statistical Interpretation

To give a statistical interpretation to ridge detection at fixed scales we view $L(\mathbf{x},t)$ in case of the *height ridge* and $L_{pp}(\mathbf{x},t)$ in case of the *second derivative ridge* as *operator responses* of the Gaussian filter kernel $G(\mathbf{x},t)$ and its second derivative $\partial_p\partial_p G(\mathbf{x},t)$ along $\mathbf{p}$ respectively. Next we do a thought experiment: the information in the observed image is destroyed by randomly shuffling all pixels to new positions. This creates a *distribution of images* which was called a *sampling model* in chapter 6. The operators $G(\mathbf{x},t)$ or $\partial_p\partial_p G(\mathbf{x},t)$ are applied to all shuffled images, which produces a *distribution of operator responses*. Now the observed operator responses are compared to the random operator responses to see, loosely speaking, how much "information" was destroyed by shuffling. This we measure by the *tail probabilities* defined in chapter 6. Finally two observed responses $L(\mathbf{x}_0,t_0)$ and $L(\mathbf{x}_1,t_1)$ are compared by means of their tail probabilities and the one with the smaller tail probability is called "more informative".

If for simplicity the shuffled images are replaced by images drawn from some analytically tractable stochastic model, e.g. a normal white noise, the tail probabilities may be analytically computed (see chapter 6). Based on the normal white noise sampling model (with zero mean and standard deviation 1) the tail probability of $L(\mathbf{x},t)$ is

$$P(L(\mathbf{x},t)) = \int_{L(\mathbf{x},t)}^{\infty} du \frac{e^{-u^2 t^N/2}}{\sqrt{2\pi t^{-N/2}}}$$

where N is the dimension of the image less one, i.e. the dimension orthogonal to the ridge direction. The tail probability of $L_{pp}(\mathbf{x},t)$ is

$$P(L_{pp}(\mathbf{x},t)) = \int_{L_{pp}(\mathbf{x},t)}^{\infty} du \frac{e^{-u^2 t^{2+N/2}}}{\sqrt{2\pi t^{-2-N/2}}}$$

The statistical interpretation allows one to rewrite the definition of a *height ridge at fixed scales* as follows: A point $\mathbf{x}$ is on a *height ridge* if $P(L(\mathbf{x},t))$ has a local minimum along $\mathbf{p}$ and $|L_{pp}| > |L_{qq}|$.

Similarly, a point is on a *second derivative height ridge* if $P(L_{pp}(\mathbf{x},t))$ has a local minimum along $\mathbf{p}$ and $L_{pp} < 0$ as well as $|L_{pp}| > |L_{qq}|$.

[2] The equations that define a *second derivative ridge at fixed scale* in two dimensions are $L_{ppp} = 0$, $L_{pppp} > 0$, $L_{pp} < 0$, and $|L_{pp}| \geq |L_{qq}|$.

7.2.3 Scale-Space Ridges in a Statistical Interpretation

The key advantage of the statistical approach is that this definition canonically generalizes to variable scales defining the *scale-space height ridge* as follows: A (position,scale)-pair $(\mathbf{x},t)$ is on a *scale-space height ridge* if the tail probability $P(L(\mathbf{x},t))$ has a local minimum along $\mathbf{p}$, as well as a local minimum along the scale dimension t, and $|L_{pp}| > |L_{qq}|$.

A *second derivative scale-space ridge* is defined similarly: A (position,scale)-pair $(\mathbf{x},t)$ is on a *second derivative scale-space ridge* if $P(L_{pp}(\mathbf{x},t))$ has a local minimum along $\mathbf{p}$, as well as a local minimum along the scale dimension t, $|L_{pp}| > |L_{qq}|$, and $L_{pp} < 0$

7.2.4 Scale-Space Ridges

Finally, to work with and to compute scale-space ridges we rewrite the definitions of scale-space ridges in a way that does not require the computation of tail probabilities (see Chapter 6 section 6.5.1). A point is on a *scale-space height ridge* (based on a white noise sampling model) if it is a local maximum of $t^\gamma L$ along $\mathbf{p}$ as well as a local maximum of $t^\gamma L$ along t and $|L_{pp}| > |L_{qq}|$ (where for the white noise sampling model $\gamma = .25$). The defining equations are:

$$\begin{aligned} L_p &= 0 \\ L_{pp} &< 0 \\ \gamma t^{\gamma-1} L + t^\gamma L_t &= 0 \\ \gamma(\gamma-1)t^{\gamma-2}L + 2\gamma t^{\gamma-1}L_t + t^\gamma L_{tt} &< 0 \\ |L_{pp}| &\geq |L_{qq}| \end{aligned} \tag{7.1}$$

A point is on a *second derivative scale-space ridge* (based on a white noise sampling model) if it is a local minimum of $t^\gamma L_{pp}$ along $\mathbf{p}$ as well as a local minimum of $t^\gamma L_{pp}$ along t, $|L_{pp}| > |L_{qq}|$, and $L_{pp} < 0$ (where for the white noise sampling model $\gamma = 1.25$). In terms of zero-crossings of derivatives the defining equations are:

$$\begin{aligned} L_{ppp} &= 0 \\ L_{pppp} &> 0 \\ \gamma t^{\gamma-1} L_{pp} + t^\gamma L_{tpp} &= 0 \\ \gamma(\gamma-1)t^{\gamma-2}L_{pp} + 2\gamma t^{\gamma-1}L_{tpp} + t^\gamma L_{ttpp} &> 0 \\ |L_{pp}| &\geq |L_{qq}| \\ L_{pp} &< 0 \end{aligned} \tag{7.2}$$

If one follows our statistical approach these definitions of scale-space ridges are straight forward generalizations of the fixed scale ridge definitions. Even the value of γ is fixed by the statistical model from which the tail probabilities are computed. Based on a normal white noise the scale-space height ridge has $\gamma = .25$ and the second derivative scale-space ridge has $\gamma = 1.25$.

7.2.5 Lack of Invariance to Linear Intensity Transformations

Inspection of the *scale-space height ridge* definition (7.1), in particular the equation $\gamma t^{\gamma-1} L + t^{\gamma} L_t = 0$, reveals that the solution to this equation changes if a constant term is added to the image intensity, i.e. $L(\mathbf{x};t) \rightarrow L(\mathbf{x};t) + const$.

For this reason the scale-space height ridge is not useful to a visual system that aims to "see" the physical world. Why? A constant increase of the image intensity is usually the result of changed lighting conditions rather than a change in the scene itself. If the aim is to "see" the physical scene, the lighting conditions should not affect the features computed from an image.

In the following we consider only the second derivative scale-space ridge.

7.3 Second Derivative Scale-Space Ridges

This section gives some examples of *second derivative scale-space ridges* computed according to definition (7.2).

Figures (7.2), (7.3), and (7.4) show the scale-space ridges of some example images, a grass, a cotton fabric and a synthetic diamond image. The images are scaled down versions of those used for the fixed scale examples above. Figure (7.2) displays *all* scale-space ridges in a projection along the scale-dimension (onto the image plain). Figure (7.3) shows projections along other directions of only the ridges consisting of more than 100 line elements for the grass image, and 50 line elements for the fabric image while all ridges of the diamond image are displayed. In figure (7.3) *boundaries* of ridges were drawn onto the original image. The boundaries are constructed to envelope circles of radius .32 times the selected scale around each position along a ridge.

First thing to observe is that the projection onto the image plain appears useful and is not dissimilar to the fixed scale results. It is not self-evident that this should turn out to be the case. Unlike the fixed scale results these figures display *all* ridges detected across a large range of scales including very small scales. Secondly the fixed scale results shown above are zero-crossings of a first derivative ($L_p = 0$) while the scale-space ridges are zero-crossings of a third derivative ($L_{ppp} = 0$). Generally the number of zero-crossings increases with the order of the derivative.

However, scale selection manages to "escapes" many zero-crossings of the third derivative, as will be discussed in the next section.

The examples demonstrate several differences between ridge detection with scale selection and ridge detection at fixed scales. Those in favor of scale-selection are:

- *Scales can vary along a ridge.* The boundaries of scale-space ridges computed from the grass image are an example hereof.
- *Ridges can cross each other.* This refers to the fact that two lines which are distinct and nonintersecting in three-dimensional scale-space may cross in a two-dimensional projection such as that of figure (7.2). The synthetic diamonds image was constructed to display this property as seen in figure (7.3). Another view of the property is provided by the *ridge surface* shown in a later section.

Last but not least the variable scale approach does not require prior knowledge about the choice of scale.

The examples also reveal some problems of ridge detection with scale selection. Most notably some "ridges" have very steep paths in scale-space, spanning a large range of scales at almost the same position. The interpretation of these structures as ridges appears questionable.

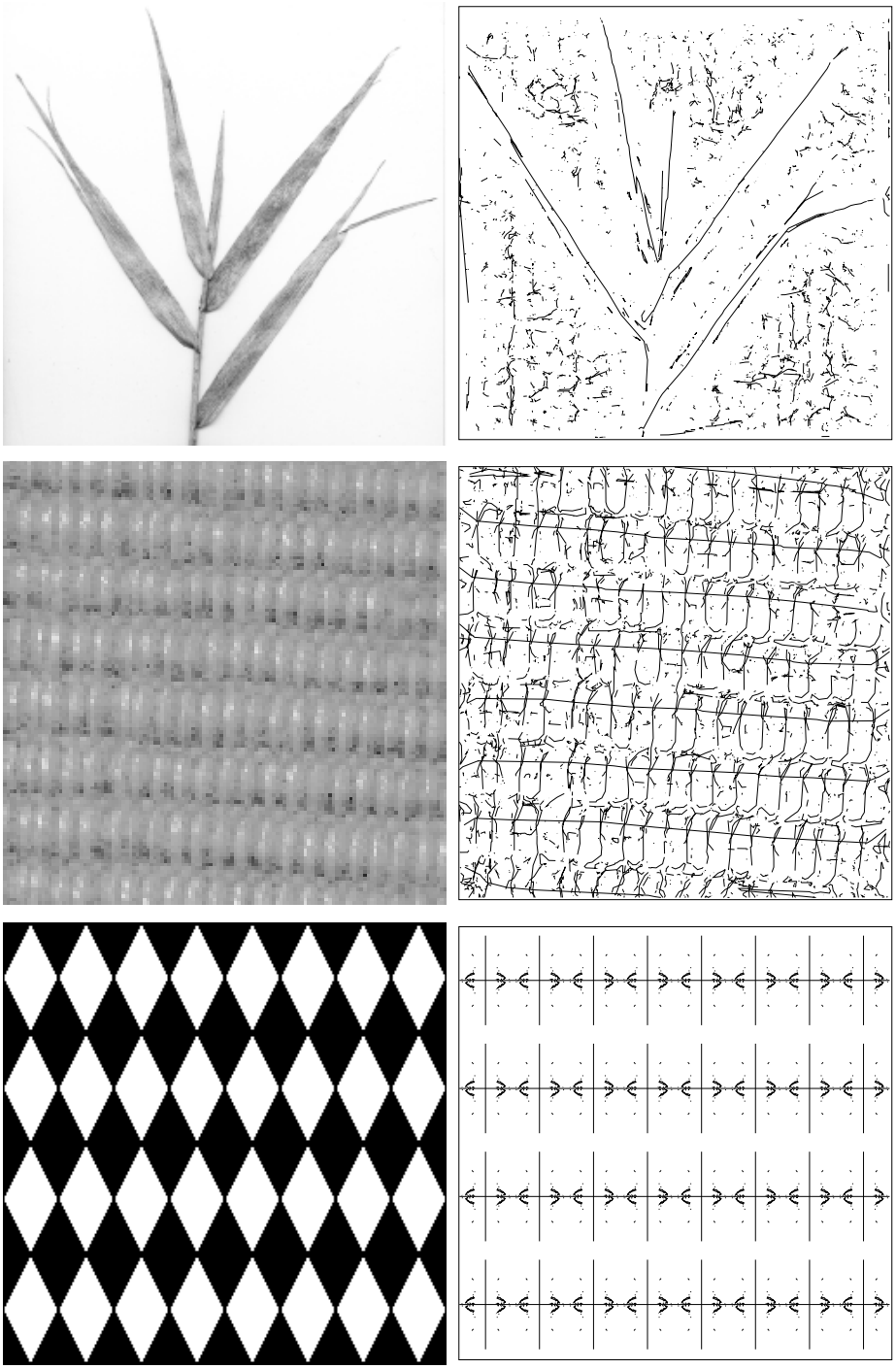


Figure 7.2: Second derivative scale-space ridges projected onto image plain. The original images are 256 by 256 pixels. Ridges were computed for $\gamma = 1.25$ and scales in the following ranges (unit length=1 pixel width): 1.5 to 16 in steps of .5 for the grass image, 1.5 to 13 in steps of .5 for the fabric image, 1 to 28 in steps of 1.0 for the diamond image.

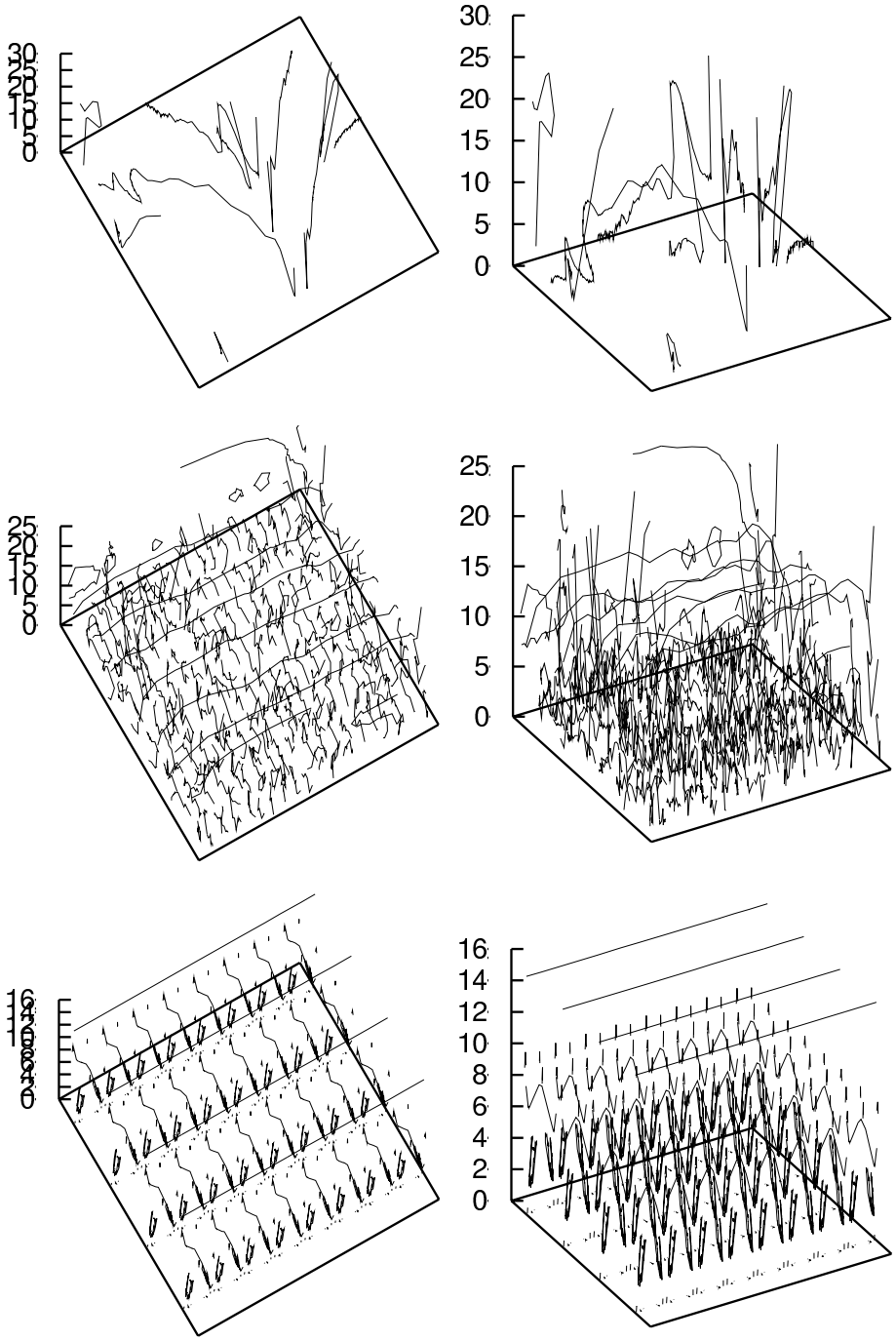


Figure 7.3: Second derivative scale-space ridges projected along different axes. Only ridges consisting of more than 100 line-elements are displayed for the grass image and ridges of more than 50 line-elements for the fabric image. For the diamond image all ridges are displayed.

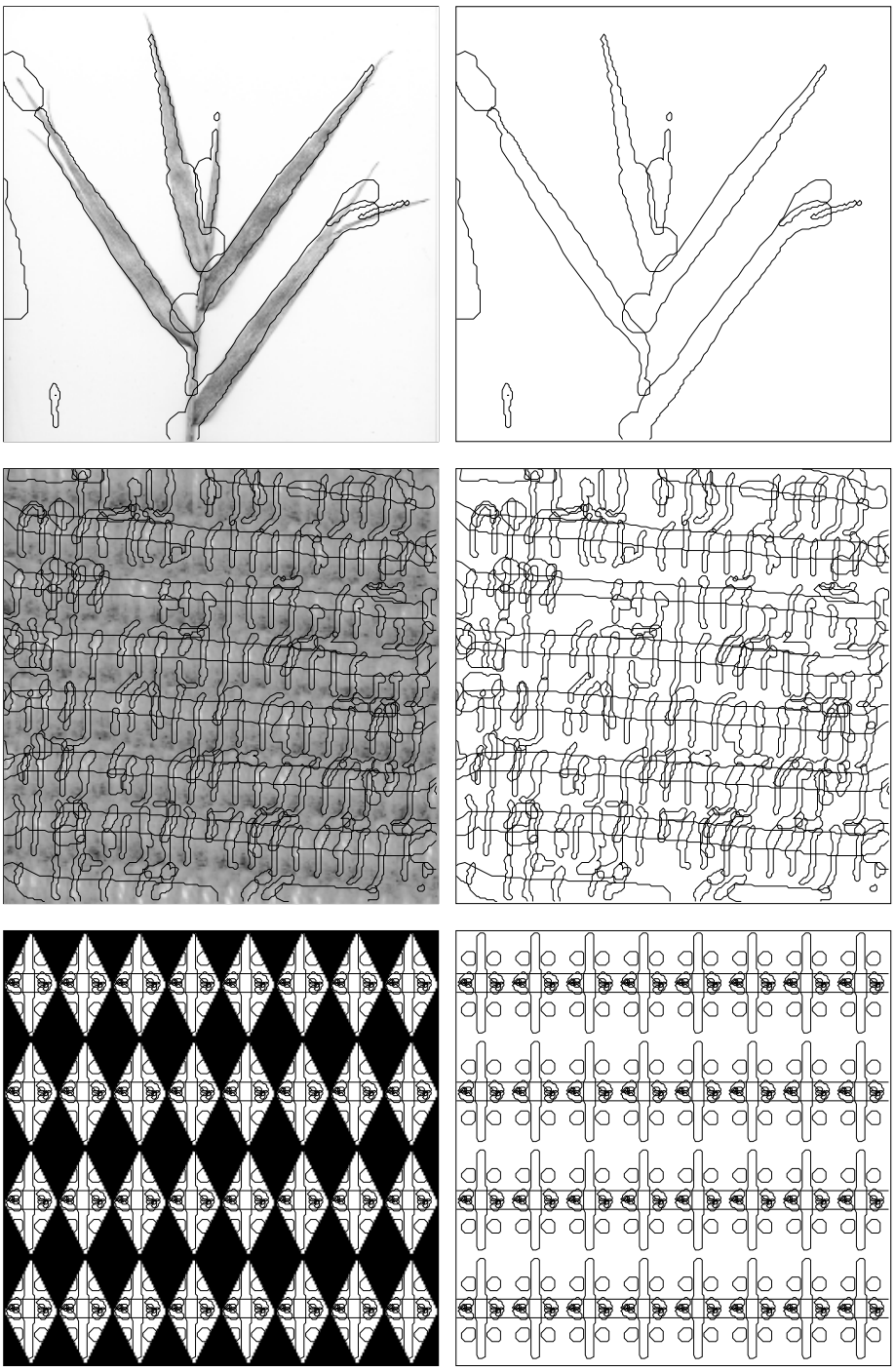


Figure 7.4: Second derivative scale-space ridges depicted by their boundaries (at scale of ridge). Otherwise as in figure (7.3).

7.4 Escape from Edges

The *second derivative scale-space ridge* (7.2) is the generalization of the fixed scale second derivative ridge to variable scales. It might thus be expected to share some properties of the second derivative ridge at fixed scales. Figure (7.5) shows some fixed scale second derivative ridges analogous to the fixed scale height ridges displayed in figure (7.1).

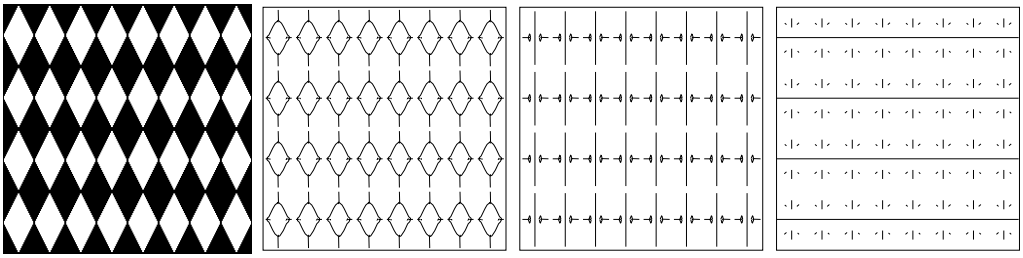


Figure 7.5: Second derivative ridges at fixed scales. The image has 512 by 512 pixels and the chosen scale levels are $\sqrt{t} = 8$, $\sqrt{t} = 16$, $\sqrt{t} = 24$ (in units of 1 pixel width/height). At small scales the diamonds edges are detected.

Concerning the detection of elongated bright structures on a dark background or dark structures on a bright background the *response to edges* seen at the smallest scale in figure (7.5) is most annoying. A number of approaches have been proposed to avoid or suppress these false responses in *scale-space ridges*. Koller [Koller, 1995] suggests a nonlinear operator that combines the response of two edge-detectors on both sides of a hypothetical ridge. Lorenz et al. [Lorenz et al., 1997b] use an edge-indicator to suppress the response to edges. Lindeberg [Lindeberg, 1998a] uses a hybrid approach taking the useful properties from both the scale-space height ridge (7.1) and the second derivative scale-space ridge (7.2).

Interestingly the *second derivative scale-space ridges* resulting from our statistical approach do *not* suffer from false responses to edges. This is a fortunate consequence of the extra scale-dimension and the value $\gamma = 1.25$ of the γ-normalization parameter resulting from the statistical approach based on a white noise sampling model. Intuitively the higher dimensional scale-space opens the possibility to "escape from edges along the scale dimension". A detailed discussion of this phenomenon is given in the following.

7.4.1 One-Dimensional Analysis

The response of the fixed scale second derivative ridge to edges results from the second derivative computed along the direction transversal to the ridge. This sug-

gests to study the behavior only in a one-dimensional frame orthogonal to the ridge direction (along which the second derivative is computed) with the principal advantage that the 2-dimensional scale-space may be completely visualized.

The one-dimensional cut across a two-dimensional ridge displays its profile. Figure (7.6) shows two model-profiles, a Gaussian-ridge profile and a step-ridge profile.

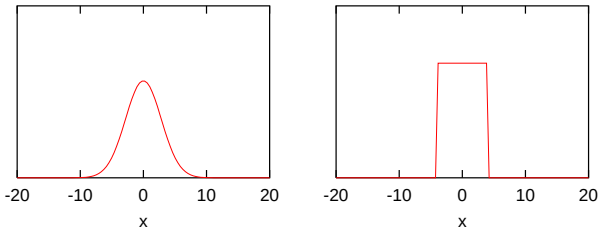


Figure 7.6: Model ridge profiles. Left: a Gaussian ridge profile. Right: a step ridge profile.

In the context of fixed scales a point in the one-dimensional frame lies on a ridge if the second derivative at that point is minimal. Figure (7.7) shows the second derivative of the two model ridges at different scales. Evidently at small scales the second derivative of the step ridge has minima on both sides of the ridge profile. These, instead of the center of the profile, are detected as "ridges".

Let us now examine the response to edges at variable scales. We recall that a (position,scale)-pair lies on a scale-space ridge (profile) if the γ-normalized second derivative $t^\gamma L_{xx}$ has a local minimum in both space and scale, and $L_{xx} < 0$. In terms of zero-crossings of derivatives the defining equations are:

$$\begin{aligned} L_{xxx} &= 0 \\ L_{xxxx} &> 0 \\ \gamma t^{\gamma-1} L_{xx} + t^\gamma L_{txx} &= 0 \\ \gamma(\gamma-1) t^{\gamma-2} L_{xx} + 2\gamma t^{\gamma-1} L_{txx} + t^\gamma L_{ttxx} &> 0 \\ L_{xx} &< 0 \end{aligned} \tag{7.3}$$

Figure (7.8) displays the negative γ-normalized second derivative $-t^\gamma L_{xx}$ for the step-ridge-model at different values of γ. Note that the negative was chosen to improve the display and consequently the *maxima* of the displayed functions correspond to points on a scale-space ridge. Each graph also depicts the zero-crossings of L_{xxx} and $\gamma t^{\gamma-1} L_{xx} + t^\gamma L_{txx}$.

For values of γ in the range $0 \leq \gamma < 1.5$ the graphs of $-t^\gamma L_{xx}$ all have one local maximum somewhere along the axis of mirror symmetry corresponding to

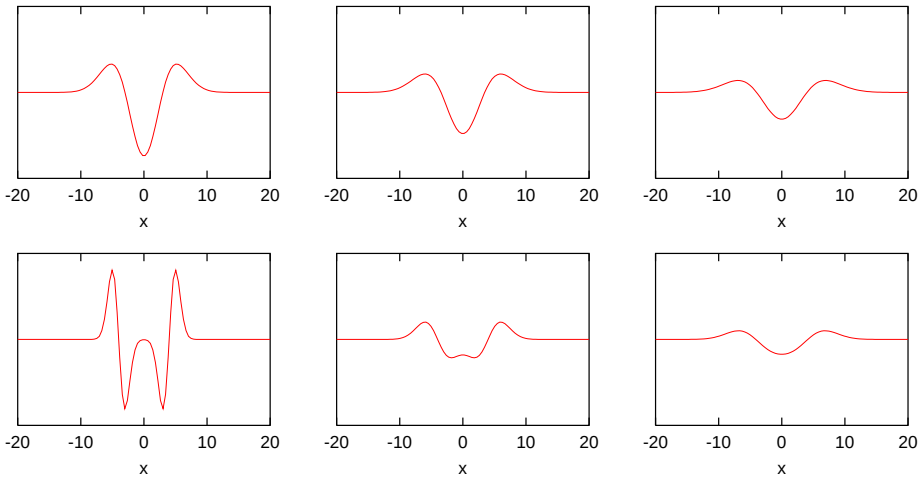


Figure 7.7: Fixed scale ridge detector responses to ridge models. The second derivative of the Gaussian ridge model is displayed in the top row for scales $\sqrt{t} = 1.$, $\sqrt{t} = 4.$, and $\sqrt{t} = 8.$. The bottom row displays the second derivative of the step ridge model for the same scales.

the center of the step-ridge. This maximum appears also as an intersection of the zero-crossings of L_{xxx} and $\gamma t^{\gamma-1} L_{xx} + t^{\gamma} L_{txx}$. We observe that the scale at which the center of the step is detected depends very sensitively on the value of γ. With increasing γ the selected scale increases and diverges as $\gamma \to 1.5$.

Our primary concern here is the behavior of L_{xx} near the edges of the step. As the figures show, maxima at the edges of the step appear only for values of γ less than 1. To see this more clearly, figure (7.9) shows the zero crossings of L_{xxx} and $\gamma t^{\gamma-1} L_{xx} + t^{\gamma} L_{txx}$ for a smaller range of scales. Hence one can conclude that *at values of* $\gamma > 1$ *the second derivative scale-space ridge detector does not detect the edges of the step-ridge.*

Taken together the range of γ where only the center of the step ridge is detected is $1 < \gamma < 1.5$.

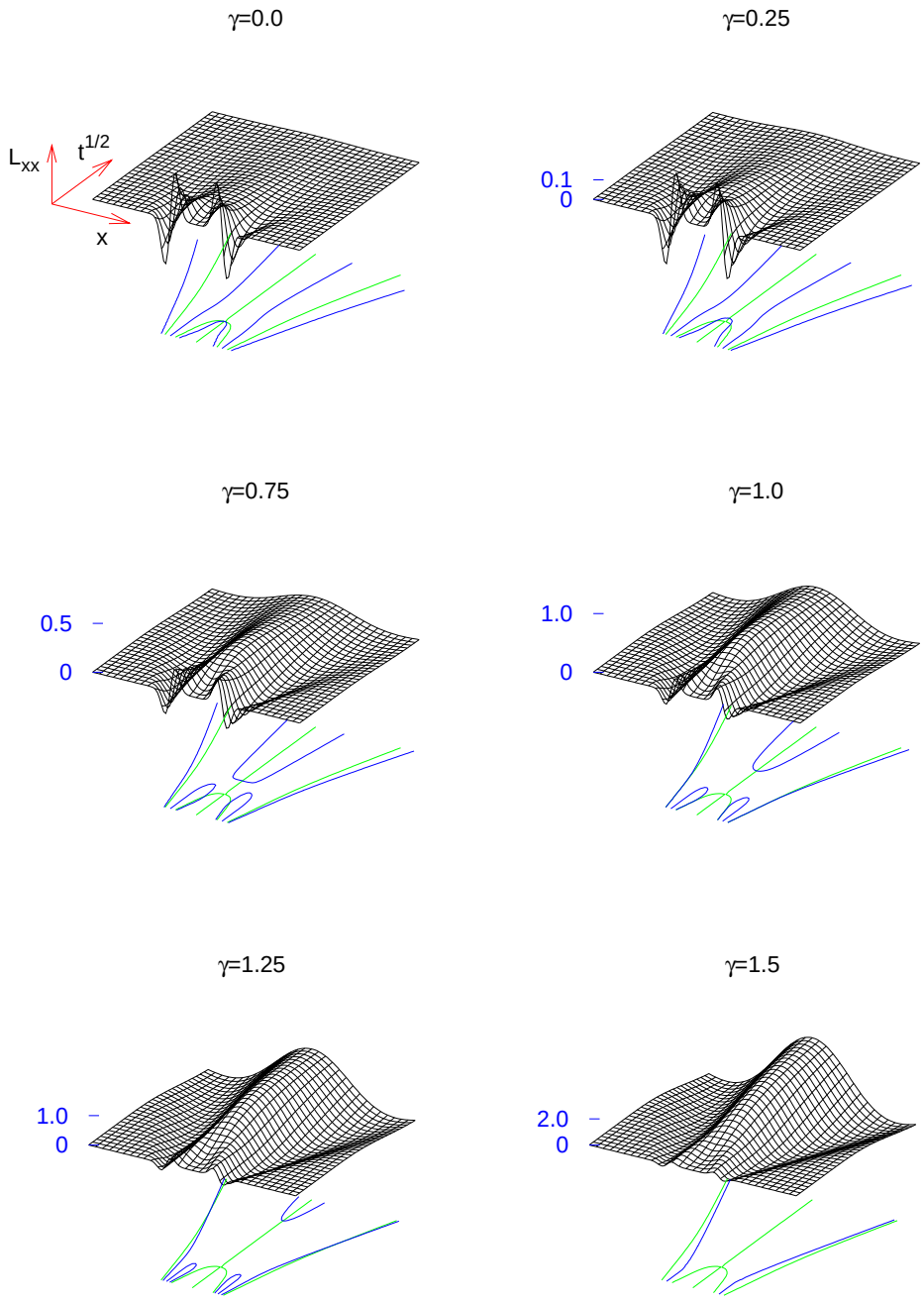


Figure 7.8: Response of second derivative ridge detector $-t^{\gamma}L_{xx}$ to step model. Zero-crossings of the first derivative along space and scale are shown below each surface. With increasing γ the central maximum is "pushed" to larger scales.

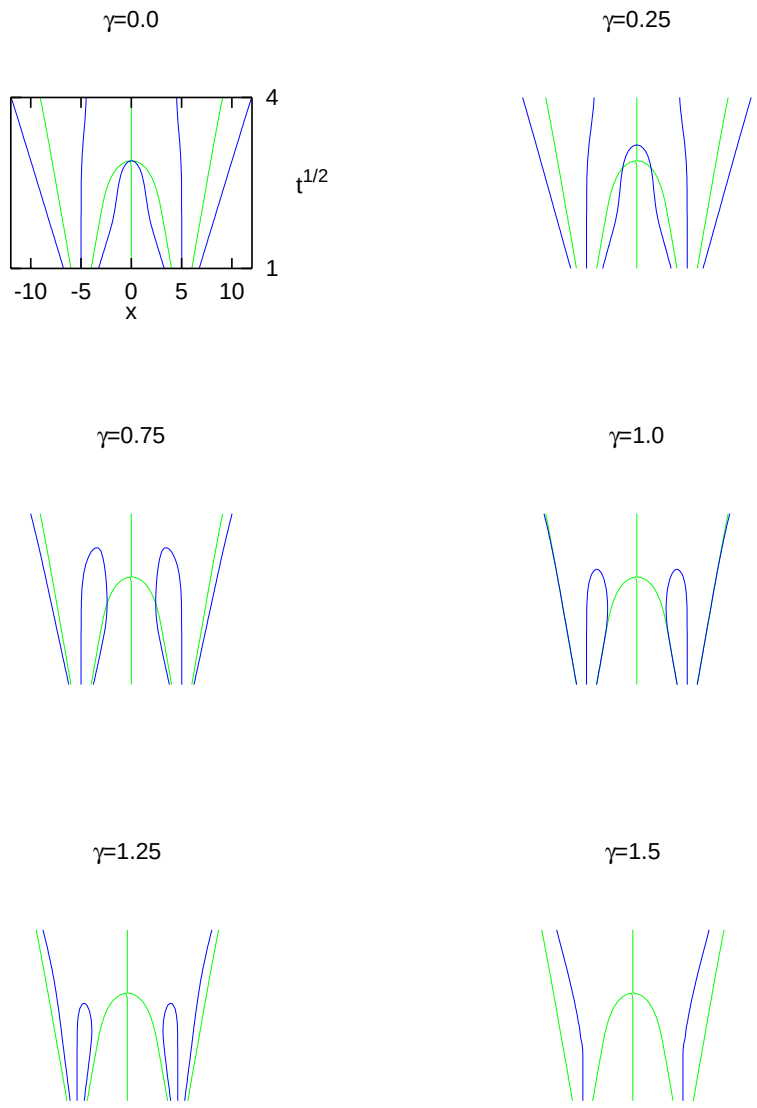


Figure 7.9: Response of second derivative ridge detector $-t^{\gamma}L_{xx}$ to step model. Zero-crossings of the first derivative along space and scale. The minima along space at the edges become unstable in the scale direction at $\gamma \geq 1.0$.

7.5 Ridge Surfaces and Scale-Selection Surfaces

In this section we show some intermediate results of the computation of scale-space ridges. They are interesting in their own right, displaying the geometry of zero-crossings in scale-space. Also they supplement the discussion of computational aspects that was given in chapter 5 for fixed scales. The major technical difference to variable scales is that instead of lines in two-dimensions it is necessary to compute surfaces in three dimensions. (Chapter 8 describes standard algorithms to compute lines and surfaces and their modifications necessary to compute ridges.)

The second derivative scale-space ridge (7.2) lies on the intersection of two surfaces in scale-space, the *ridge surface* [3] and the *scale-selection surface*. We define the *ridge surface* as follows

$$\begin{aligned} L_{ppp} &= 0 \\ L_{pppp} &> 0 \\ |L_{pp}| &\geq |L_{qq}| \\ L_{pp} &< 0 \end{aligned} \tag{7.4}$$

and the *scale-selection surface* as follows

$$\begin{aligned} \gamma t^{\gamma-1} L_{pp} + t^{\gamma} L_{tpp} &= 0 \\ \gamma(\gamma-1) t^{\gamma-2} L_{pp} + 2\gamma t^{\gamma-1} L_{tpp} + t^{\gamma} L_{ttpp} &> 0 \end{aligned} \tag{7.5}$$

Figure (7.10) shows the *ridge surface* of one of the "diamonds" of the synthetic diamond image. One view looks down the scale axis onto the surface. The other has the scale axis oriented pointing from top to bottom of the page and the the long axis of the diamond from left to right.

In the second view the vertical surface that extends farthest left and right contains the scale-space ridge corresponding to the long axis of the diamond. The orthogonal ridge along the short axis lies in the central of the three vertical surfaces that reach highest in scale. The latter surface "bridges" across the former with a small gap along the scale-dimension. In the projection to the image plain this lets the two ridges cross. The crossing of ridges thus goes along with a "bridge" geometry of the ridge surface in this particular case.

[3] Pizer et al. speak of *parameter surface*.

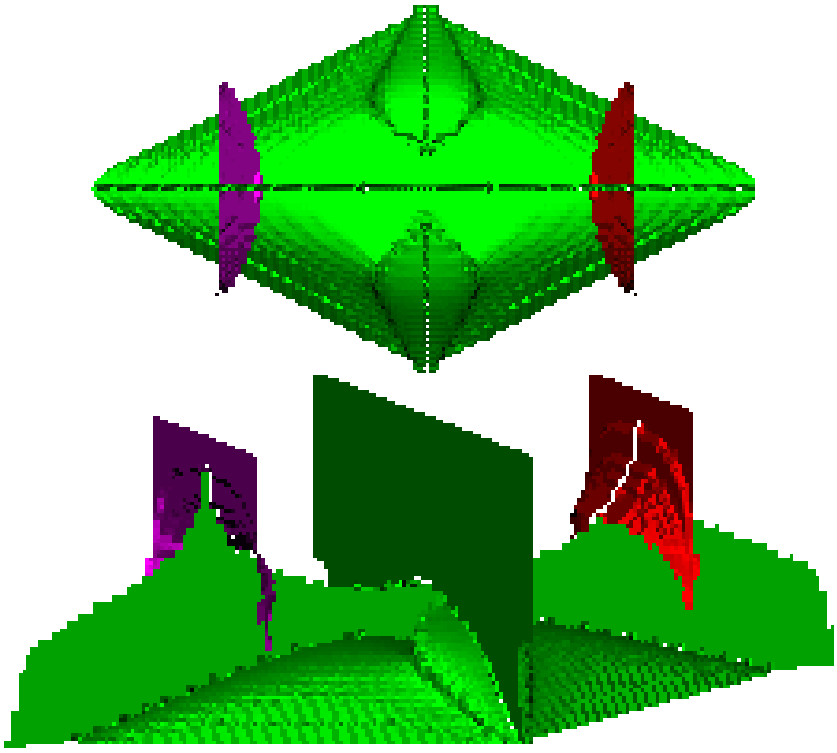


Figure 7.10: Ridge surface of the diamond image from two views.

It should be noted that the ridge surface is not closed and that this results from the discontinuity of the direction **p** (as described in chapter 5) and is not a consequence of the inequalities in the definition above. To demonstrate this figure (7.11) displays the zero-crossing surface $L_{ppp} = 0$ which is also not closed.

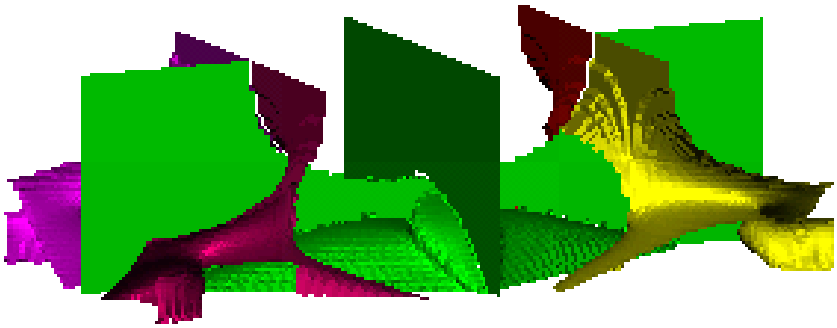


Figure 7.11: Zero-crossing surface of L_{ppp} for the diamond image.

Finally the surface $\gamma t^{\gamma-1} L_{pp} + t^{\gamma} L_{tpp} = 0$ which contains the *scale-selection surface* is shown in figure (7.12) for $\gamma = 1.25$.

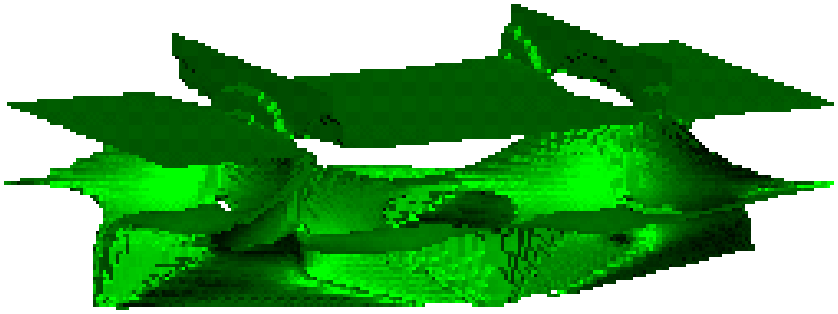


Figure 7.12: Zero-crossing surface of $\gamma t^{\gamma-1} L_{pp} + t^{\gamma} L_{tpp}$ for the diamond image.

The intersection of this surface (less those parts where $t^{\gamma} L_{pp}$ is a maximum rather than a minimum along scale) with the ridge surface yields the *second derivative scale-space ridges* of the diamond image as shown in the previous sections.

Chapter 8

Algorithms for Zero-Crossings

In this chapter we discuss algorithms to compute *zero-crossings* of functions whose values are known only at discrete positions, usually the points of a regular square or cubic grid. Zero-crossings are lines or surfaces that separate the points into regions of positive and negative. The chapter first describes standard algorithms to compute zero-crossings of a single function and then shows how to modify these in order to compute the *simultaneous zero-crossings of two functions* as it is necessary in ridge detection. In contrast to the standard algorithms this allows to compute *open zero-crossings*.

The computation of zero-crossings may be divided into two steps: detection and extraction. From a 2 dimensional image f (or a 2 D function sampled on a regular square grid) one *detects* zero-crossings separating the data into $f > 0$ and $f \leq 0$. Naturally the result of this is a set of line segments each confined to a square of 2x2 neighboring pixels.

To create lines from the segments it is then necessary to establish neighborhood relations between different line segments. We refer to this as the *extraction* of a line.

Similarly, detection of zero-crossing surfaces from 3 dimensional images results in independent surface-patches each confined to a cube of 2x2x2 neighboring pixels. In order to extract surfaces it is then necessary to establish neighborhood relations between individual patches.

A great deal of work concerning the above problems has appeared in the computer graphics literature where they are known (slightly more generally) as *iso-surface* detection and extraction. Their principle application lies in the *visualization* of 3-dimensional data. For this purpose it suffices to detect iso-surfaces and display their patches, so the literature mainly emphasizes iso-surface detection and usually does not introduce the distinction between detection and extraction made here.

The best known algorithm for iso-surface detection, the *marching*

cubes algorithm was independently invented by Wyvill and McPheeters in 1986 [Wyvill et al., 1986] and by Lorensen and Cline in 1987 [Lorensen and Cline, 1987] . The textbook of Alan and Mark Watt [Watt and Watt, 1992] presents a highly efficient implementation that allows subsequent surface extraction.

The computation of two zero-crossing surfaces and their intersection as it occurs in the previous chapter is addressed by [Thirion and Gourdon, 1996]. They begin by computing one of the two surfaces and then detect those lines within this surface that correspond to the intersection of the two surfaces.

A problem that requires special attention in ridge detection is the computation simultaneous of zero-crossings of two functions. In contrast to the zero-crossings of a single function these can be *open*. The necessary modifications to the standard algorithms are described at the end of the chapter. To begin with let us consider zero-crossings of a single function in 2 dimensions.

8.1 Zero-Crossings in two Dimensions

This section treats the computation of zero-crossings from two-dimensional data sampled on a regular square grid as shown in figure (8.1). Discretization results in two types of inaccuracy: i) the exact *location* of a zero-crossing between *any* two neighboring points of the grid can only be estimated and ii) the *topology* of the zero-crossings at *some* places is ambiguous. The following sketch demonstrates both the limited accuracy and the possibility of a topological ambiguity that results from discretization in 2D.

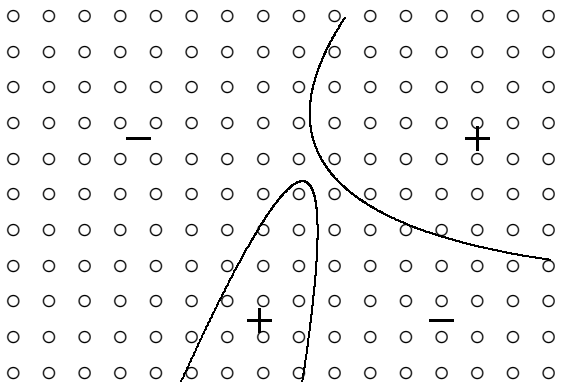


Figure 8.1: Zero-crossings of a 2D-function and grid points of known function values. Within the 4 points of the square grid where the two zero-crossing lines come closest discretization leads to a topological ambiguity.

8.1.1 Detection

Primarily we want to construct *closed lines* (except for possible ends at the image border) that divide the sampled function values into $f > 0$ and $f \leq 0$. This may be achieved with merely 16 different basic elements that describe lines passing through any 2x2 square of neighboring pixels as shown in figure (8.2). Any corner of a square can have a function value either > 0 or ≤ 0 allowing the four corners of a square to have at most $2^4 = 16$ different configurations concerning > 0 or ≤ 0. These can be divided by 16 different configurations of line elements passing through the square.

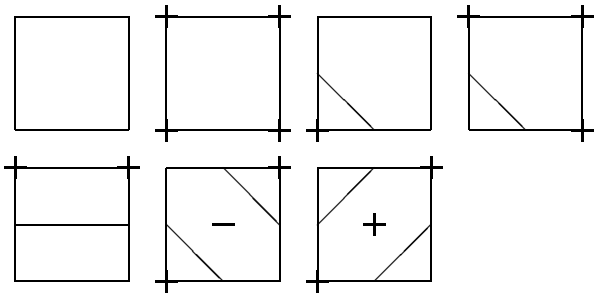


Figure 8.2: Basic line elements. The "+" at the corners denote positive function values. Rotation yields the 16 basic configurations plus two extra configurations resulting from the ambiguity depicted in the last two cases.

A topological ambiguity occurs when all four sides of a square are intersected by a line as depicted in the last two cases of figure (8.2). This information is irreversibly lost by the discretization. A guess at the true situation may be taken by considering the average of the function values at the four corners and choosing the last case when this is positive and the second to last otherwise.

Continuation of lines between squares is automatically achieved if the intersection of a side of a square is determined only from the function values at the two corners of that side. Since these are shared by the two neighboring squares, the same point of intersection is computed in both neighbors. Good results may be achieved by estimating the point of intersection of a side from a linear interpolation of the two function values at the ends. Figure (8.3) gives an example of zero-crossings with linear interpolation contrasted by the same image without linear interpolation (lines intersecting the middle of square-sides).

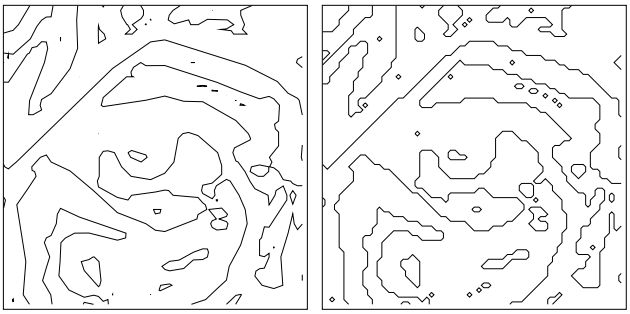


Figure 8.3: Zero-crossings with and without linear interpolation. The original image is 64 pixels wide and high.

8.1.2 Extraction

To establish the neighborhood relations between line elements it is conceptually simplest to "walk" along a line. Recording the encountered intersections with sides of squares in a list captures all the neighborhood information necessary to describe a line. It is even possible to achieve detection and extraction in a single sweep through the data if the order in which squares are processed is determined by the lines (requiring special attention only when a square contains two line elements).

Unfortunately this simple strategy does not generalize to the extraction of iso-surfaces from three-dimensional images. Let us therefore consider another possibility: Initially a detection phase is performed, finding the line elements within any 2x2 square of image pixels. The essential requirement to allow subsequent extraction of lines is that information about neighborhoods is not discarded during the detection phase. This must be facilitated by the data structures.

Optimally, the intersection of any side of a square is computed only once and the point of intersection is stored only once. A line data structure of this type is shown in figure (8.4). The two line elements that share a point of intersection record a reference to the storage position of that point rather than the point itself. They may then be identified as neighbors through this reference. To give an example consider extraction of the line in figure (8.4). The first line element contains points 1 and 4. The only other line element that contains point 4 is the second which is thus the continuing element. The far end of this element is point 5 which is also a point of the third line element, and so on.

The strategy to store each point of a line only once and to identify neighbors by their shared reference to this point may be extended to iso-surfaces of cubic grids in 3D by storing each vertex of a surface just once and having several surface patches refer to a single vertex.

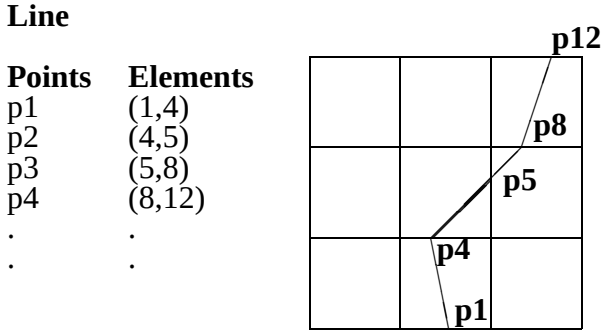


Figure 8.4: A line data structure that stores each point only once captures the neighborhood information.

8.2 Zero-Crossing Surfaces in 3D

This section describes the detection and extraction of zero-crossing surfaces from three-dimensional data sampled on a regular cubic grid. Attention is paid to the treatment of topologically ambiguous situations. In three dimensions a wrong treatment produces holes in the computed surfaces, an obvious topological error. Two different ways to detect surfaces without holes are discussed. The standard approach is to always assume a negative (positive) value for the center of an ambiguous cube-face and to choose a surface that has not only the corners of but also the center on the correct side. Alternatively the value at the center of an ambiguous face may be interpolated from the corners. Finally the related issues of implementation and extraction are treated.

8.2.1 Detection

In 3 dimensions there are $2^8 = 256$ possibilities for the 8 corners of a cube to be inside ($f < 0$) or outside ($f \geq 0$) a surface. Up to simple symmetries these basic cases are displayed in figure (8.5).

The idea of the *marching cubes* algorithm is to determine the surface patches resulting from the 256 different cases just once, leaving open only the exact locations where edges of the cube are intersected, i.e. the exact positions of the surface-vertices. For any cube within an image it then suffices to determine which of the 256 cases the cube belongs to and to fill in the exact vertex positions.

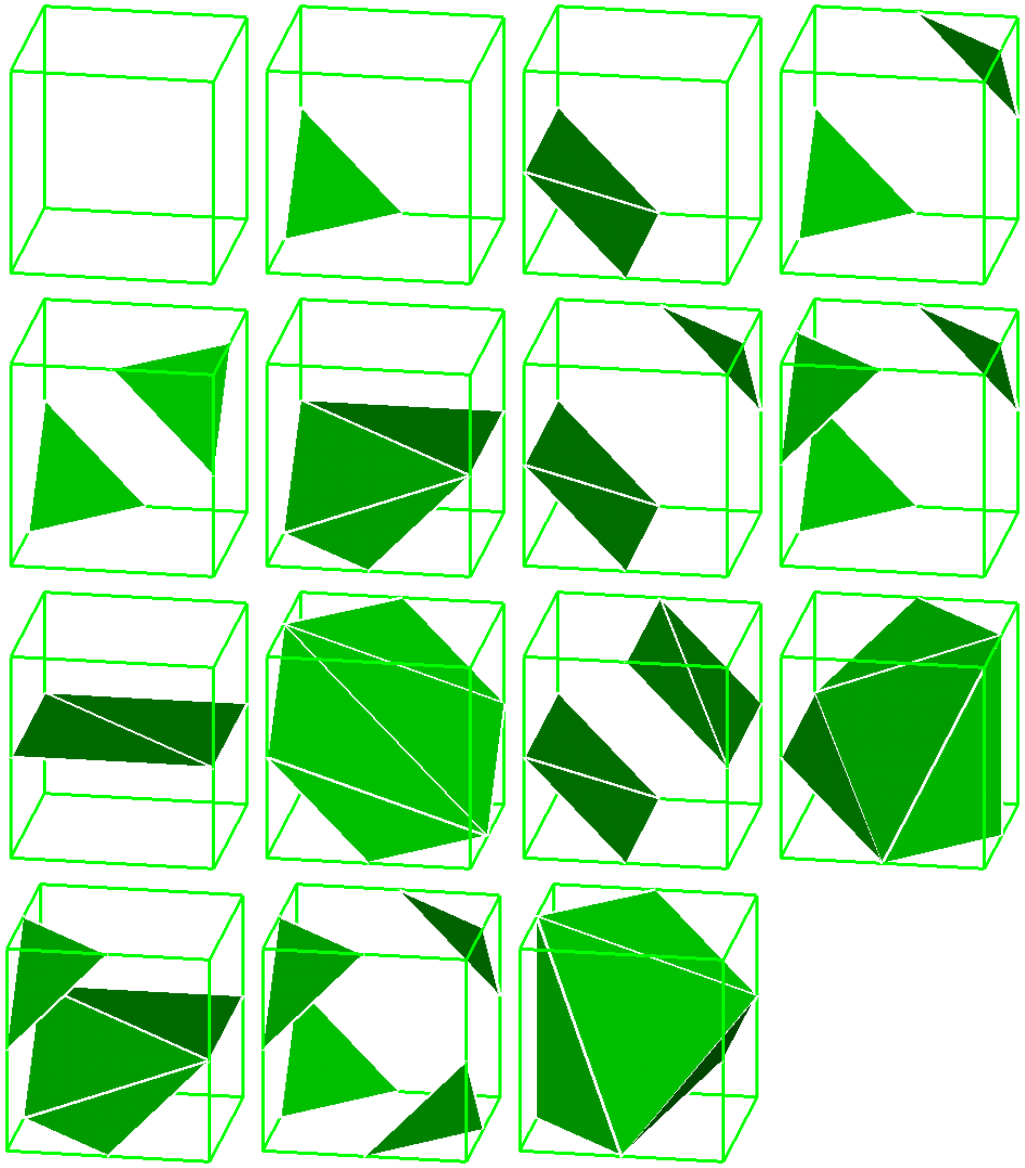
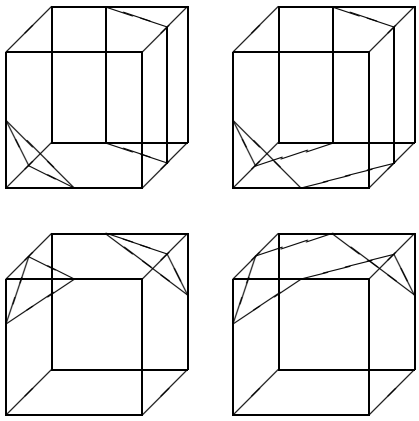


Figure 8.5: Basic surface elements used in the original *marching cubes* algorithm. Eight corners of the cube have 2^8 possibilities to be inside ($f < 0$) or outside ($f \geq 0$) a surface. The figure displays these cases up to simple symmetries (rotation, mirroring, change of inside and outside). Note that cases 4,7,8,11,13, and 14 are ambiguous and the displayed surfaces are just one of several possibilities chosen arbitrarily. The white lines within surfaces indicate possible triangulations.



Unfortunately ambiguous situations occur when all four edges of one cube face are intersected[1], as in cases 4,7,8,11,13, and 14 of figure (8.5). As the figure on the right shows, care must be taken to make a consistent choice across neighboring cubes. Each of the two possibilities displayed for the top and bottom cube must be combined with the correct choice to create a surface without holes.

A natural approach proposed in [Thirion and Gourdon, 1996] is to interpolate the values at the center of each ambiguous face of the cube and to choose the unique surface topology that also has this central point on the "correct" side. Most algorithms, however, just assume all values at the center of a face to be negative (positive). Either way should be "hardwired" into the algorithm in terms of a case-table. We now describe how to create this table.

8.2.2 Generating the Case-Table

To generate a table that maps all possible cubes into unique and consistent surface patches we create all possible cubes once and detect their surface patches once.

The surface-patches within a cube may be systematically constructed in three steps: i) Detect a possible intersection of each of the 12 edges of the cube. ii) Within each face of the cube link the intersections into line elements. iii) Follow the lines around the cube to create surface patches. The first and third step are straight forward. The second step needs to deal with possible ambiguities.

A face with two diagonally opposite corners labeled "+" and the other two labeled "-" may be intersected in either of the two way shown in figure (8.2). To arrive at a unique choice in an ambiguous case one may just assume the center of the face to be "-" and then choose the unique surface that has the central "-" on the correct side.

Thirion et al. [Thirion and Gourdon, 1996] propose to estimate the value at the central point of an ambiguous face by a linear interpolation. Again, of the possible surfaces that one is chosen which has the central value on the correct side. It should be noted that the central value of a face is utilized only when the corners alone are insufficient to select a unique surface topology. The disadvantage of this

[1]The problem of determining surface topologies relative to the 8 corners of the cube from only the information about inside and outside at these 8 corners is ill-posed in the sense of Hadamard.

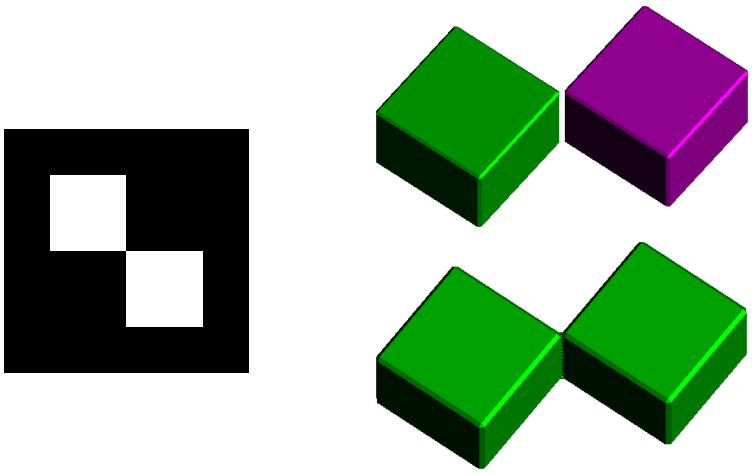


Figure 8.6: Differences between 8 bit or 14 bit case tables: On the left is shown a slice from the test image, the right shows the two possible surfaces. Using an 8bit case table with builtin preference for large negative volumes always splits the positive pixels into two volumes (top). The 14bit case table can also produce a single volume (bottom), depending on interpolated values.

approach is obvious, the number of different cube configurations increases from 2^8 to 2^{14} including the centers of a cubes six faces. [2]

To contrast the two methods consider a test image made up of several slices as shown in figure (8.6) with positive values inside the rectangular regions and negative outside. The algorithm working with 256 cases and the assumption that centers of an ambiguous face are always negative always creates two disconnected surfaces from these data. Using the larger case table and linear interpolation of the centers is able to create one or two surfaces depending on the interpolated value.

8.2.3 Implementation and Extraction

Our implementation, based on the description of [Watt and Watt, 1992] and the case tables above, computes each vertex only once during the detection phase. This minimizes both time and memory and at the same time permits subsequent extraction.

The cubes of the image are processed starting with all cubes of one row, next all rows within a band and finally different bands. Each new cube (except for

[2] Strictly speaking, many of the 2^8 cases are not ambiguous and the larger case table is really only needed for the ambiguous situations. In terms of efficiency, however, it generally pays to choose the larger case table right away.

those on the image border) encounters three edges that have not been dealt with before as displayed in figure (8.7).

The vertices of these three edges, if any, are computed and stored. References to vertices of the other 9 edges are fetched from buffers that are kept specifically for this purpose. Figure (8.8) shows these buffers. Two buffers `xvert[row][col]` and `yvert[row][col]` store all references to horizontal edges lying between consecutive bands. To store references between consecutive rows two buffers `zvert[col]` and `top[col]` are introduced and finally three edges need to be stored between two consecutive cubes in a column.

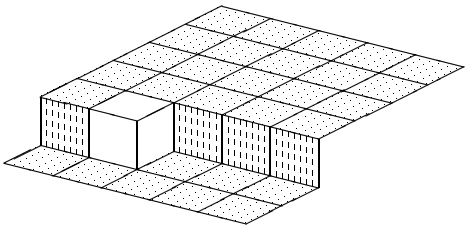


Figure 8.7: Marching of cubes. Only three edges need to be dealt with in a single step, the others have been treated previously.

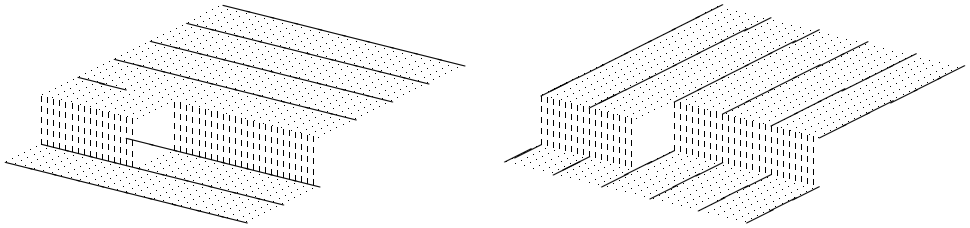


Figure 8.8: Buffers `xvert` and `yvert` used in the implementation.

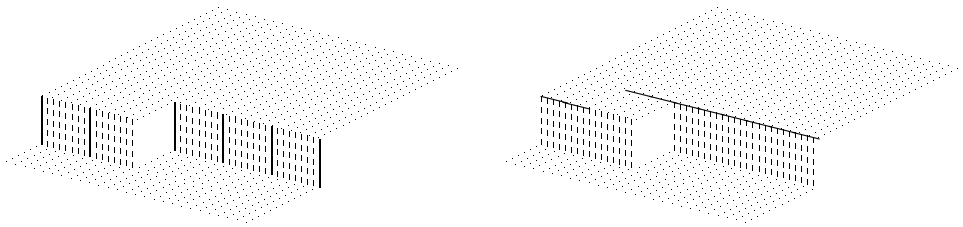


Figure 8.9: Buffers `zvert` and `top` used in the implementation.

The surface patches are broken down into triangles [3] so that detection creates the data structure shown on the right.

Each vertex consists of three coordinates (floating point variables). The order of vertices is the order in which they are encountered by the marching cube. A triangle contains only the integer position of its three vertices.

Surface

Vertices	**Triangles**
v1	(1,2,7)
v2	(2,7,52)
v3	(1,2,23)
v4	(4,6,90)
.	.

To establish neighborhood relations between surface patches we proceed as follows. Initially each vertex is assigned a list of references to all triangles that contain this vertex. To find a neighbor to one side of any triangle it then suffices to look for the unique other triangle that appears in both lists of the two vertices belonging to the side of the first triangle.

Finally, surfaces may be walked using graph algorithms such as breadth first search [Cormen et al., 1990]).

8.3 Open Zero-Crossings

The algorithms described above determine the topology within a square (or cube) from the values of a function at the corners of the square. This ensures consistency between squares, i.e. closed, continuous lines and surfaces. The algorithms may easily be modified to allow more general computation of lines and surfaces that need not be closed.

One such application occurs in ridge detection where the following problem is encountered (as described in more detail in the chapters on ridge detection): Find those points of an image where the gradient vector $\mathbf{v} \in \mathbb{R}^2$ is an eigenvector of the 2x2 Hessian matrix $\mathbf{H}$ to a given eigenvalue λ, i.e. where

$$(\mathbf{H} - \lambda\mathbf{I})\mathbf{v} = 0$$

($\mathbf{I}$ being the identity matrix). $\mathbf{H}$, λ, and $\mathbf{v}$ vary over the image at some points fulfilling the above equations and at others not. The two equations are linearly dependent in $\mathbf{v}$ because λ is an eigenvalue of $\mathbf{H}$, i.e. $\det \mathbf{H} - \lambda\mathbf{I} = 0$. Unfortunately linear dependence does not generally mean that if $\mathbf{v}$ satisfies the first equation it also satisfies the second equation, namely when the first row of $\mathbf{H} - \lambda\mathbf{I}$ is zero: $H_{1,1} - \lambda = 0$ and $H_{1,2} = 0$. If, however, the coefficients of the first equation do

[3] Triangles are created in an arbitrary way. Methods that attempt to fit triangles to the zero-crossings in an optimal manner unfortunately tend to be very time consuming.

not vanish, then a solution of the first equation is also a solution of the second equation. Computationally this requires to compute a zero-crossing between any two pixels from the more *stable* of the two equations. Section 5.4.3 describes the measure of stability that was used for ridge detection. As a consequence of the fact that the zero-crossings are not computed from a single function they need not form closed curves.

The standard zero-crossing algorithms may be modified as follows: Within a square the intersections of all four edges are computed by whatever method is appropriate, for example as zero-crossings of the more stable of two expressions as described in the previous paragraph. Topologically this can result in 16 possible configurations, determined by the intersected edges. Figure (8.10) shows these configurations.

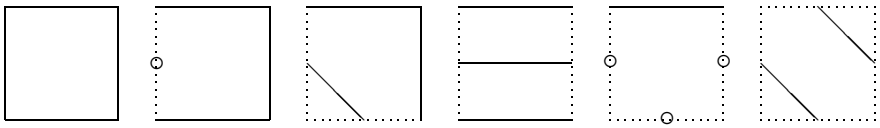


Figure 8.10: Basic line elements. Dotted lines mark intersected edges. Circles symbolize line ends. Rotation yields the 16 basic configurations.

The case table that usually maps configurations of the corners into line topologies is now replaced by a table that maps configurations of edges into line topologies [4]. As far as detection of lines is concerned this is the only necessary modification. Extraction can proceed as before except for the fact that lines must be allowed to terminate within the image.

Figure (8.11) shows zero-crossings computed in ridge detection. Between any two neighboring pixels the more stable row of $\mathbf{H}\,\mathbf{v} - L_{qq}\mathbf{v}$ (where L_{qq} is the larger eigenvalue of $\mathbf{H}$) determines whether an edge is intersected or not. Clearly the resulting zero-crossings can be open.

The generalization to three dimensions is straight forward. A cube has 12 edges that are either intersected or not. This yields 2^{12} cases of edge-configurations that must be mapped to surface-patches. The corresponding case table is created as follows: Within each cube face line elements are created as shown in figure (8.10). It is then attempted to track the line elements around the cube to create closed contours of surface patches as shown in figure (8.5). If this cannot be done the corresponding surface patch is left out.

[4] Ambiguities may be resolved in some way appropriate to the application.

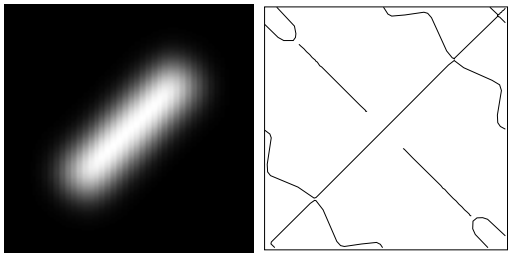


Figure 8.11: Zero-crossings of $\mathbf{H}\,\mathbf{v} - L_{qq}\mathbf{v}$. At these points $\mathbf{p}$ is orthogonal to the gradient $\mathbf{v}$.

8.4 Intersections of Zero-Crossings

To compute the intersection of two zero-crossings we follow the approach of [Thirion and Gourdon, 1996]: extract only one of the zero-crossings from the image data and compute the second zero-crossing within the irregular grid provided by the first. Alternatively one could first compute both zero-crossing surfaces and then find their intersections as is briefly described in the appendix of [Lindeberg, 1998a].

In two dimensions the first zero-crossing yields a set of lines each made up of a list of points. At any two neighboring points the second function is evaluated (by linear interpolation of the four closest neighbors) to see if there is a zero-crossing in between. If so, the location of the zero-crossing is estimated by linear interpolation.

In three dimensions the first zero-crossing yields a set of surfaces each consisting of triangles. Within any triangle we detect zero-crossing lines of the second function, and if a line is detected it is followed to neighboring triangles.

Results computed with these algorithms are shown in figures (7.2), (7.3), and (7.4) of chapter 7 on ridge detection with scale selection.

Chapter 9

Self-Similarity of Noise in Scale-Space

This chapter deals with the statistical properties of normal noise in scale-space. One such property, the standard deviation, was already used in chapter 6 to define scale selection. Here the emphasis is on more general statistical properties such as those of features, e.g. edges or ridges, computed from the images.

The central observation is a scaling invariance of normal noise in scale-space. From this invariance it is easy to derive the scaling behavior of measurements made on normal white noise random fields. Examples of measurements are the number of local extrema per unit volume of scale-space, the length of edges, or the average gradient along edges. The statistical properties of these quantities might be used to assess the significance of different features in the sense that long edges are more significant than short edges because they are less likely to occur. This chapter was presented at a conference [Majer, 1999]. I changed only the layout and corrected an error in equation (9.2).

9.1 Introduction

Properties of normal white noise in scale-space have been studied previously for a number of reasons. Images may be corrupted by noise and scale-space smoothing may improve the signal to noise ratio. Noise has served as a model to study the behavior across scales of properties such as the number of local extrema or the volume of grey-level blobs [Lindeberg, 1994b]. Deviations from the scaling behavior of properties of white noise or ensembles of natural images [Ruderman and Bialek, 1994] can provide useful information to a visual system.

Apart from the covariance of normal white noise in scale-space [Blom et al., 1993] results have been achieved mostly by simulation. The purpose

of this paper is to illustrate that some useful results are available analytically.

9.2 An Invariance of Noise in Scale-Space

It is well known [Koenderink and van Doorn, 1992] that the only *functions* that are form invariant under linear scale-space filtering are the derivative of Gaussian functions

$$G^{\mathbf{n}}(\mathbf{x};t) = \partial_1^{n_1}...\partial_N^{n_N} \frac{e^{-\frac{\mathbf{x}^T\mathbf{x}}{2t}}}{(2\pi t)^{N/2}}$$

Filtering these functions with a Gaussian kernel $G^{\mathbf{0}}$ is equivalent to a rescaling as expressed by the invariance $\mathbf{x} \to s\mathbf{x}$, $t \to s^2 t$, $G^{\mathbf{n}} \to s^{-n-N} G^{\mathbf{n}}$ or $G^{\mathbf{n}}(\mathbf{x};t) = s^{-n-N} G^{\mathbf{n}}(s\mathbf{x};s^2 t)$. N denotes the dimension of space, $x \in R^N$, $\mathbf{n} = (n_1,...,n_N)$ specifies the derivative operator , and $n = \sum_i n_i$ its order. The square-root of the second argument $0 < t \in R$ is the "scale" of $G^{\mathbf{n}}$.

There is also a family of *random fields* that is invariant under scale-space filtering with a kernel $G^{\mathbf{0}}$ in the sense that *a filtering of the random field is equivalent to a rescaling of the joint distribution function*.

Members $\xi^{\mathbf{n}}(\mathbf{x};t)$ of this family are generated (and defined) by filtering a normal white noise $\xi^{\mathbf{0}}(\mathbf{x};0)$ of zero mean and standard deviation σ with a derivative of Gaussian filter kernel $G^{\mathbf{n}}$:

$$\xi^{\mathbf{n}}(\mathbf{x};t) = (G^{\mathbf{n}}(\cdot;t) * \xi^{\mathbf{0}}(\cdot;0))(\mathbf{x})$$

These normal random fields are completely determined by their auto-covariance function

$$\gamma^{\mathbf{n}}(\mathbf{x}-\mathbf{x}', t+t') = \sigma^2(-1)^n \, G^{2\mathbf{n}}(\mathbf{x}-\mathbf{x}';t+t')$$

that describes the covariance of $\xi^{\mathbf{n}}(\mathbf{x};t)$ and $\xi^{\mathbf{n}}(\mathbf{x}';t')$. It follows immediately that the form invariance of $G^{\mathbf{n}}$ is inherited by the random fields:

$$\gamma^{\mathbf{n}}(\mathbf{x}-\mathbf{x}', t+t') = s^{-2n} s^{-N} \gamma^{\mathbf{n}}(s(\mathbf{x}-\mathbf{x}'), s^2(t+t')) \qquad (9.1')$$

The (joint distribution function of the) random field $\xi^{\mathbf{n}}$ is invariant under the rescaling

$$\begin{aligned} \mathbf{x} &\to s\mathbf{x} \\ t &\to s^2 t \\ \sigma &\to s^{-n} s^{-N/2} \sigma \end{aligned} \qquad (9.1)$$

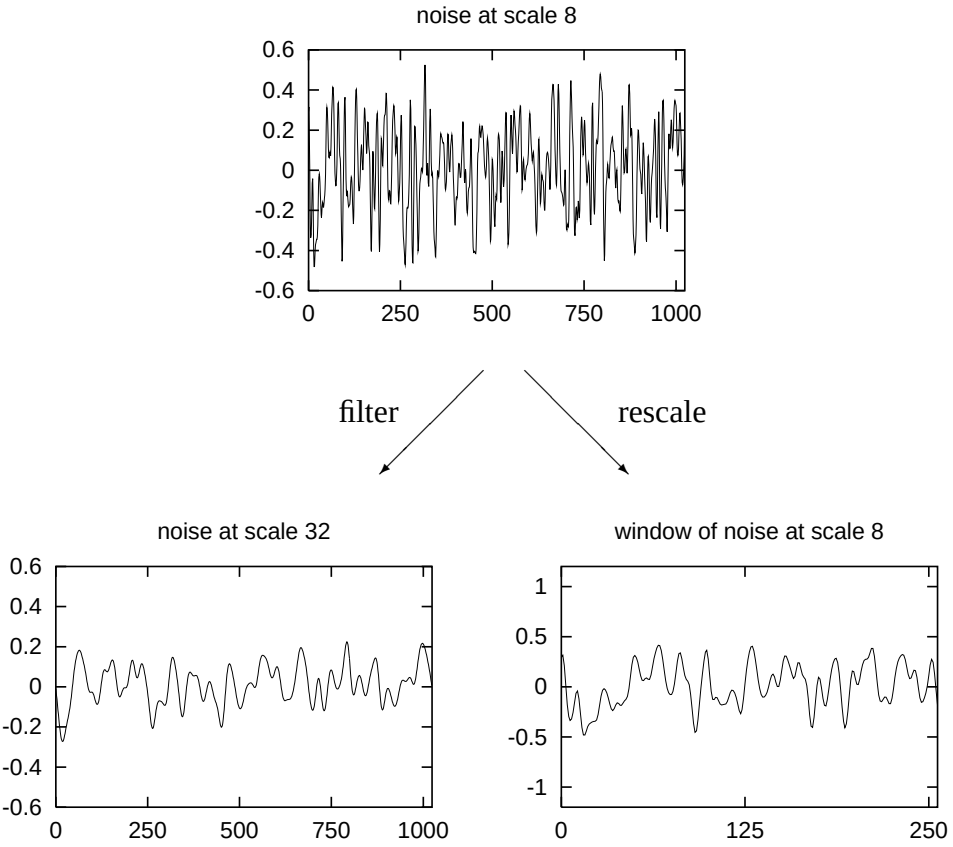


Figure 9.1: Normal noise at scale $\sqrt{t} = 8$, filtered to scale 32, and a rescaled *display* of the noise at scale 8 showing only $0 < x < 256$.

Figure (9.1) displays a one-dimensional realization of $\xi^{\mathbf{0}}(x, 8^2)$, the same realization filtered to $\xi^{\mathbf{0}}(x, 32^2)$, and lastly a rescaled *display* of the first graph.

Obviously the particular *function* that we have realized is not scaling invariant. Filtering the function in the first graph results in the second which is apparently different from the third graph that shows the appropriately rescaled version of the first. However, the similarity of these graphs serves to illustrate the fact that they are generated by identical random mechanisms, i.e. that *the random field is scaling invariant.*

The invariance of normal noise under the scaling transformation (1) allows to derive the *scaling behavior* of any observations made on a random field $\xi^{\mathbf{n}}(\mathbf{x}; t)$.

Some examples follow.

9.3 Density of Local Extrema

The number of local extrema of normal white noise in scale-space has been studied as a model for the scale dependence of the number of features in a signal [Lindeberg, 1994b]. Computation of the expected value of the number of local extrema *at a fixed scale* is extremely difficult [Lindeberg, 1994b]. However, the scale invariance property (1) of normal white noise directly gives a relationship between the distributions of the numbers of local extrema at different scales.

From (1) we find that the distribution of the number $\mathrm{N}^{\mathrm{ext}}$ of local extrema in a volume V of space at t is identical to the distribution of the number of local extrema in a volume $s^N V$ at $s^2 t$. More specifically:

- the probability $P_t(\mathrm{N}^{\mathrm{ext}})$ of observing less than $\mathrm{N}^{\mathrm{ext}}$ local extrema in a unit volume ($\int_\Omega d\mathbf{x} = 1$) of filtered white noise $\xi^{\mathbf{n}}(\mathbf{x};t)$ at scale $\sqrt{t}$ is related to $P_{s^2 t}(\mathrm{N}^{\mathrm{ext}})$ at scale $s\sqrt{t}$ by

$$P_t(\mathrm{N}^{\mathrm{ext}}) = P_{s^2 t}(\mathrm{N}^{\mathrm{ext}} s^{-N}) \tag{9.2}$$

- the expected number $E(\mathrm{N}^{\mathrm{ext}})$ of local extrema over space per unit volume of space behaves as

$$E(\mathrm{N}^{\mathrm{ext}}) \propto t^{-N/2} \tag{9.3}$$

Note that (9.2) and (9.3) hold for any order of the derivative $\mathbf{n}$ in $\xi^{\mathbf{n}}(\mathbf{x};t)$. The scaling behaviour (9.3) of the expected number of local extrema over space has previously been suggested from simulation experiments and a dimensional analysis argument in section 8.7.5 of [Lindeberg, 1994b].

Similar relations hold for the distribution and expectation of the number $\mathrm{N}^{\mathrm{ScSp}}$ of local extrema over scale and space per unit volume of scale and space:

$$E(\mathrm{N}^{\mathrm{ScSp}}) \propto t^{-N/2-1} \tag{9.4}$$

The scale-dependence (9.4) of the number of local extrema over scale and space is verified by simulation experiments. Figure (9.2) shows a plot of $\log \mathrm{N}^{\mathrm{ScSp}}$ against $\log t$ for one-dimensional and two-dimensional white noise.

9.4 Edge Lengths

The distribution of edge lengths l in normal noise $\xi^{\mathbf{n}}(\mathbf{x};t)$ at scale $\sqrt{t}$ is identical to the distribution of scaled edge lengths sl at scale $s\sqrt{t}$. Again this scaling

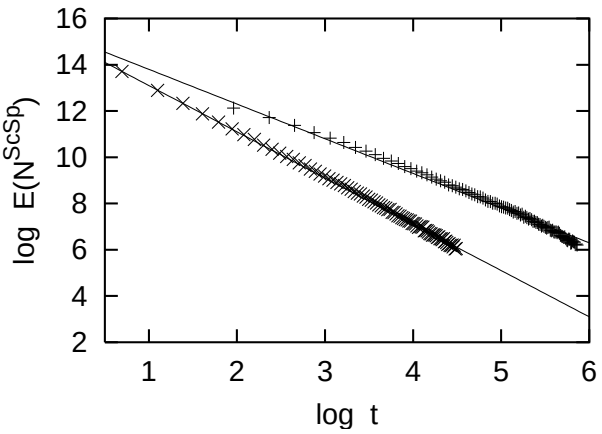


Figure 9.2: Log-Log plot of the number of local extrema against scale for a one-dimensional (top curve) and a two-dimensional normal white noise. The theoretical curves are depicted as lines.

invariance results directly from the invariance of the distribution of normal white noise under the scaling transformation (1) without the need to actually compute the distribution of edge lengths. It should be noted that for this scaling behavior to hold it is essential that edges are computed by an algorithm that commutes with the scaling transformation, e.g. zero crossings of differential invariants.

Let us denote by $P_t(l)$ the relative frequency of edges of lengths less than l occurring in the set of all edges at scale $\sqrt{t}$ in a normal noise image $\xi^{\mathbf{n}}(\mathbf{x};t)$. $P_t(l)$ is identical to the probability $P_{s^2t}(sl)$ of edges of lengths less than sl occurring in a filtered image $\xi^{\mathbf{n}}(\mathbf{x};s^2t)$

$$P_t(l) = P_{s^2t}(sl)$$

so that the expected edge lengths grow linearly in scale $\sqrt{t}$

$$E(l) \propto \sqrt{t} \tag{9.5}$$

as shown on the left of figure (9.3).

9.4.1 Edge Lengths with Boarder Effects

In contrast to dimensionless features the distribution of edge lengths is certainly affected by the image boarder cutting some edges short. We therefore attempt to describe the effect of this on the distribution of edge lengths.

Consider a two-step procedure to arrive at the measured edge lengths. First edges are computed from a hypothetical boarderless image. Then this is cropped

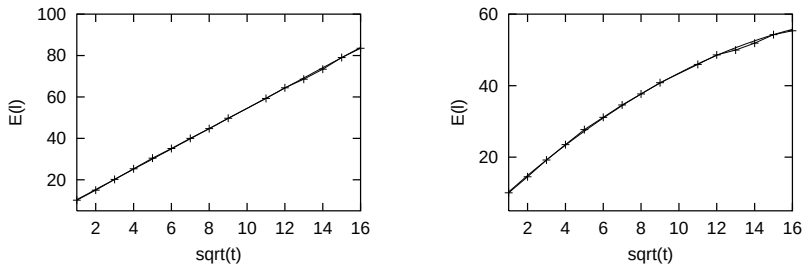


Figure 9.3: Mean length of edges $E(l)$ as a function of scale $\sqrt{t}$. left: without boarder effects. right: with boarder effects. Theoretical relations are shown as lines.

to the observed image size. Thereby some edges are cut into two. One piece of each of these cut edges is kept. With probability one half it will be the long and with equal probability the short piece, so that the expected length after cutting is one half that before. If we denote by $p_t^c(l)$ the probability density of lengths of edges cut by the image boarder, we have

$$p_t^c(l) = 2p_t(2l)$$

where $p_t(l)$ is the density of lengths l, i.e. $P_t(l) = \int_0^l du \quad p_t(u)$. Each edge has a certain probability p^b to be on the boarder of the image. This probability p^b depends on the length l of the edge. If we assume that p^b is linear in l — which should be a good assumption as long as the edge length is smaller than the length of the image — it will scale like

$$p^b(l) = s^{-1}p^b(sl)$$

The density of observed lengths at scale $\sqrt{t}$

$$(1-p^b(l))p_t(l) + p^b(l)p_t^c(l)$$

then scales to

$$(1-s^{-1}p^b(sl))sp_{s^2t}(sl) + 2p^b(sl)p_{s^2t}(2sl)$$

Thus the mean length depends on t as

$$E(l) \propto \sqrt{t} - at$$

with a constant a that depends inversely on the length of the image boarder and on the edge detection and linking algorithm used. Figure (9.3) shows a fit of the scale dependence of edge lengths in 512 by 512 pixel white noise images in scale-space. As edge-detection and linking algorithm we used Canny's non-maximum suppression and hysteresis thresholding [Canny, 1993] (for thresholding see below).

9.5 Blob Volumes

Volumes of so called grey-level blobs have been used to construct a systematic approach for the extraction of important structures in images [Lindeberg, 1994b]. Their significance was assessed from a comparison to the expected blob volume in normal noise.

For the analysis of their scale dependence in normal noise it suffices to know that grey-level blob volumes are integrals of the (smoothed) intensity function over regions of the image domain, and that the regions are defined by geometric properties of the intensity [Lindeberg, 1994b]. Irrespective of whether each region grows or shrinks with increasing scale, the number of regions decreases like $t^{-N/2}$ and thus there average area A increases like

$$E(A) = t^{N/2} \ .$$

More generally, the distribution of areas of the regions of integration shows the invariance $P_t(A) = P_{s^2 t}(s^N A)$.

The values of the intensity function depend on scale as $t^{-N/4}$ so that the integrals over the above areas depend on scale like $t^{N/2-N/4}$

$$E(\text{blob volume}) \propto t^{N/4} \tag{9.6}$$

as reported in simulation studies by Lindeberg [Lindeberg, 1994b].

9.6 Scale-Dependent Thresholds

The described scale dependencies hold only when the measurements commute with the scaling transformation. The introduction, for example, of a threshold in Canny's edge detection and hysteresis algorithm would destroy the scale dependence shown in figure (9.3).

Thresholds may however be modified to depend on scale such that their relative position within the distribution of values they are applied to is independent of scale. Or, conversely, the distribution of values to be thresholded may be rescaled. In the edge detection a threshold on the absolute value of the gradient should be be proportional to t^{-1} (for a two-dimensional image). Alternatively, as in figure (9.3) a fixed threshold was used and 'standardized' gradients

$$t^{1/2} t^{N/4} \, \partial_i \frac{e^{-\frac{\mathbf{x}^T \mathbf{x}}{2t}}}{(2\pi t)^{N/2}}$$

were computed. The use of standardized derivatives is superior to a scale-dependent threshold in that it may be numerically checked by setting the power of the filter kernel equal to 1.

9.7 Summary

Scale dependencies of distributions of properties of white noise in scale-space were derived from a scaling invariance of normal random fields. The method is usually much simpler than a direct computation of the distribution at fixed scales and subsequent derivation of the scale dependence.

Chapter 10

Summary and Outlook

Over the last two decades the scale-space representation of images has become an important element in computer vision. The representation replaces an original image by a family of smoothed versions of the same image such that with increasing smoothness the details of the original image are lost. In principle this allows a visual system to "concentrate" on the appropriate level of detail as follows: With increasing degree of smoothing objects vanish from the image, small objects first and larger objects later. The degree of smoothing at which an object vanishes basically measures the size of an object and provides information about the appropriate level of detail that allows the visual system to "concentrate" on the object under consideration. In this way the degree of smoothing is linked to the size of objects and one speaks of scale-space rather than smoothing-space.

Ironically the question of how to determine particularly informative scales has eluded scale-space theory for almost a decade. The first approach to "scale selection" was proposed in 1993 by Lindeberg [Lindeberg, 1993b]. The proposal is to define particularly informative scales as the local maxima with respect to scale of so-called γ-normalized derivatives.

Different choices of the γ-parameter yield different particularly informative scales. It remains a question, what the "right" choice of γ-normalization should be. More generally it is not clear why scales should be selected according to the proposal of Lindeberg. Both questions are addressed by this thesis.

The central idea that is presented in this thesis in order to deal with scale selection is *stochastic simplification*: the pixels (intensities) of an observed image are randomly shuffled to new positions. On average, if shuffling is defined appropriately, this destroys information and creates simplified versions of the original image.

A Statistical Approach to Feature Detection and Scale Selection

In chapter 6 *stochastic simplification* was used to define particularly informative positions or scales of the response of some operator as follows:

1. Destroy the *structural* information in the observed image by *shuffling* the pixels to new positions.
2. Measure how much information was destroyed at any single position and scale of the operator response.
3. Label those positions where (locally) most information was destroyed as particularly informative.

To define a measure of how much information was destroyed by shuffling at any single (position,scale)-pair one compares the operator response from the observed image at that (position,scale)-pair with the *distribution of operator responses from the shuffled images*. Specifically, in chapter 6 we took the probability to get smaller responses from random images than the observed response as the measure of how much information is lost by shuffling.

This definition of particularly informative positions or scales becomes computationally feasible if shuffling is replaced by sampling random images from some analytically tractable *sampling model*. The prototypical sampling model that produces images without structural information is a normal white noise model.

It was shown that on the basis of any homogeneous sampling model the particularly informative positions occur exactly at those positions where the operator response has a local maximum with respect to position. In other words the statistical definition of particularly informative positions is equivalent to the standard definition of feature detection provided that the sampling model is homogeneous.

The particularly informative scales of derivative of Gaussian operators were seen to correspond to local maxima of γ-normalized derivative of Gaussian operators when the sampling model is normal noise. The γ-normalization parameter is uniquely determined by the sampling model.

The statistical approach to define particularly informative parameters differs from the standard approaches in several ways: i) It provides a single definition for *all* particularly informative parameters of an operator, be they position, scale, or other. ii) It has an intuitive motivation in terms of the above three-step procedure. iii) It defines particularly informative parameters *relative to* a distribution of images that are by construction/definition less informative than the original.

Ridge Detection with Scale Selection

Chapter 7 demonstrated ridge detection with scale selection based on the γ-normalization corresponding to a normal white noise sampling model. It was

observed that this choice of γ allows a second derivative ridge detector to "escape" edges. At fixed scales this detector frequently produces false responses to edges. At variable scales, however, the maxima in response to step edges become unstable at the critical γ-value of $\gamma = 1$. For a step ridge the critical γ-value is $\gamma = 1.5$. This means that at $\gamma = 1.25$, corresponding to a normal white noise sampling model, step edges are not detected while ridges are detected.

Shuffling, Simplification and Scale-Space

In chapter 2, section 4 stochastic simplification was defined as follows: allow a pixel to jump from position $\mathbf{x}$ at "time" t to position $\mathbf{y}$ at "time" $t+\tau$ with some *transition probability* $p(\mathbf{y}, t+\tau|\mathbf{x}, t)$. To achieve *gradual* simplification some simple conditions on the short time behavior were imposed. The expected value of random images generated in this way was shown to be a scale-space.

Section 2.4 introduced a *local entropy* defined for any single point in scale-space. It was proved that the sum of local entropies over all points of an image increases monotonically with scale. This captures in a mathematically rigorous way the intuitive idea that smoothing (by Gaussian filter kernels) simplifies images both globally and, more importantly, also locally.

Self-Similarity of Noise in Scale-Space

Chapter 9 dealt with the statistical properties of normal noise in scale-space. In particular it observed a scaling invariance of normal noise that can be used to assess the scaling behavior of quantities measured in normal noise images. A simple quantity is the standard deviation of the noise. The scaling invariance, however, applies equally to more complicated quantities such as the number of local extrema per unit volume or the length of ridges computed from normal noise images.

Future Work

Several directions for future work suggest themselves.

The *local entropy* introduced in chapter 2 deserves a detailed study. In particular it is clear that the the scale-dependence of the local entropy at any single position is not generally monotonic. Consider a simple example: At the center of a dark spot on a bright background the entropy initially increases until the scale corresponds roughly to the size of the blob. If scale is increased further the density shifts toward the bright pixels and the entropy decreases.

The local entropy is only one property of the local densities created by shuffling. Other properties of local densities could be of interest as well. Recent work

in this direction was presented by [van Ginneken and ter Haar Romeny, 1999].

The local entropy or the local Kullback-Leibler discrepancy should also be the key to an information theoretic treatment of image analysis and feature detection and scale selection in particular. In the authors opinion this is the most important and perhaps the most demanding of the here mentioned directions of work.

Furthermore, scale-selection in non-linear scale-spaces could be approached with the ideas of chapter 6. The key issue that needs to be dealt with is the computation of the power of a filter kernel generated by non-linear scale-space.

It would also be interesting to test whether natural images also satisfy the scaling invariance that holds for normal noise in scale-space (chapter 9). Several studies of the second order statistics of natural images have confirmed the scaling invariance [Ruderman and Bialek, 1994], [Steenstrup et al., 1999]. The question is whether the invariance also holds for "higher order" quantities such as the number of local extrema per unit volume or the length of edges.

Appendix A

Direction of Minium Curvature

Throughout chapter 5 the vector **p** was said to lie along the axis of minimum second derivative. We here show that this is equivalent to **p** being an eigenvector of the Hessian to the smallest eigenvalue.

The second derivative of $L(\mathbf{x})$ along a vector **d** (given in coordinates $\mathbf{x} = (x, y)$) is

$$\mathbf{d}^T \mathbf{H} \mathbf{d} = \mathbf{d}^T \begin{bmatrix} L_{xx} & L_{xy} \\ L_{yx} & L_{yy} \end{bmatrix} \mathbf{d}$$

The direction of minimum second derivative is then obtained by minimizing over all $\mathbf{d} \in \mathbb{R}^2$ restricted to a constant length $\mathbf{d}^T \mathbf{d} = 1$:

$$\min_{\mathbf{d} \in \mathbb{R}^2, \mathbf{d}^T \mathbf{d} = 1} \mathbf{d}^T \mathbf{H} \mathbf{d}$$

Introducing a Lagrange multiplier λ to enforce the constraint this becomes

$$\min_{\mathbf{d} \in \mathbb{R}^2, \lambda \in R} \mathbf{d}^T \mathbf{H} \mathbf{d} - \lambda(\mathbf{d}^T \mathbf{d} - 1) \quad .$$

Denoting the direction **d** along which the second derivative assumes a minimum by **p** it follows that **p** and λ must solve

$$\mathbf{H}\mathbf{p} = \lambda \mathbf{p} \quad ,$$

i.e. λ must be an eigenvalue of the Hessian and **p** an eigenvector. Finally, to have a minimal second derivative **p** must have the smallest eigenvalue.

Appendix B

Least Squares Fit of Second Derivative Ridge

It was mentioned in chapter 5 that *second derivative ridges* may be interpreted as those positions where a *ridge model* optimally fits to the image. The model to allow this interpretation is

$$g_{\mathbf{d}}(\mathbf{x}) = -\partial_d \partial_d G(\mathbf{x}, t)$$

where $G(\mathbf{x}; t) = \frac{e^{-\frac{\mathbf{x}^T \mathbf{x}}{2t}}}{(2\pi t)}$ is the rotation symmetric Gaussian and $\mathbf{d} \in \mathbb{R}^2$ a vector in $\mathbb{R}^2$ along which the second derivative is taken.

At each position $\mathbf{x}_0$ the direction $\mathbf{d}$ is sought along which the model optimally fits to the image $f(\mathbf{x})$ in terms of the square difference

$$\min_{\mathbf{d} \in \mathbb{R}^2, \mathbf{d}^T \mathbf{d} = 1} \int d\mathbf{x} \; (-\partial_d \partial_d G(\mathbf{x} - \mathbf{x}_0, t) - f(\mathbf{x}))^2$$

This is equivalent to the minimum of the convolution of $\partial_d \partial_d G(\mathbf{x} - \mathbf{x}_0, t)$ with $f(\mathbf{x})$:

$$\min_{\mathbf{d} \in \mathbb{R}^2, \mathbf{d}^T \mathbf{d} = 1} \int d\mathbf{x} \; \partial_d \partial_d G(\mathbf{x} - \mathbf{x}_0, t) \; f(\mathbf{x})$$

Pulling the derivative out of the integral and inserting the definition of scale-space $L(\mathbf{x}, t) = (G(\circ, t) * f)(\mathbf{x})$ shows that the least squares direction $\mathbf{d}$ is the direction of smallest second derivative.

$$\min_{\mathbf{d} \in \mathbb{R}^2, \mathbf{d}^T \mathbf{d} = 1} L_{dd}(\mathbf{x}_0, t)$$

This direction was called $\mathbf{p}$ in chapters 5 and 7.

To determine the least squares position in the one-dimensional frame along **p** again one seeks to minimize

$$\int d\mathbf{x}\ (-\partial_d \partial_d G(\mathbf{x}-\mathbf{x}_0,t) - f(\mathbf{x}))^2$$

along **p**. This is equivalent to the maximum of the convolution, and thus to

$$L_{ppp}(\mathbf{x}_0,t) = 0$$
$$L_{pppp}(\mathbf{x}_0,t) > 0$$

Bibliography

[Alvarez et al., 1993] Alvarez, L., Guichard, F., Lions, P., and Morel, J. (1993). Axioms and fundamental equations in image processing. *Arch. Rational Mech. Anal.*, 123:199–257.

[Alvarez et al., 1992] Alvarez, L., Lions, P., and Morel, J. (1992). Image selective smoothing and edge detection using anisotropic diffusion. *SIAM J. Numer. Anal.*, 29:845–866.

[Archer and Titterington, 1995] Archer, G. and Titterington, D. (1995). On some bayesian/regularization methods for image restoration. *IEEE Trans. Image Processing*, 4(7):989–995.

[Atick, 1992] Atick, J. (1992). Could information theory provide an ecological theory of sensory processing? In Bialek, W., editor, *Princeton lectures on biophysics*, pages 223–289.

[Babaud et al., 1986] Babaud, J., Witkin, A., Baudin, M., and Duda, R. (1986). Uniqueness of the gaussian kernel for scale-space filtering. *IEEE PAMI*, 8:26–33.

[Barlow, 1953] Barlow, H. B. (1953). Summation and inhibition in the frog's retina. *J. Physiol. (Lond.)*, 119:69–88.

[Barron et al., 1994] Barron, J., Fleet, D., and Beauchemin, S. (1994). Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77.

[Blom et al., 1993] Blom, J., ter Haar Romeny, B., Bel, A., and Koenderink, J. (1993). Spatial derivatives and the propagation of noise in gaussian scale space. *J. of Vis. Comm. and Im. Repr.*, 4(1):1–13.

[Canny, 1986] Canny, J. (1986). A computational approach to edge detection. *IEEE PAMI*, 8:679–698.

[Canny, 1993] Canny, J. (1993). Finding edges and lines in images. Technical Report AI-TR-720, M.I.T.

[Cormen et al., 1990] Cormen, T., Leiserson, C., and Rivest, R. (1990). *Introduction to Algorithms*. MIT Press.

[Eberly, 1996] Eberly, D. (1996). *Ridges in Image and Data Analysis*. Computational Imaging and Vision. Kluwer Academic Publishers.

[Field, 1987] Field, D. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am.*, 4(12):2379–2394.

[Florack, 1993] Florack, L. (1993). *The syntactical structure of scalar images*. PhD thesis, University of Utrecht, Utrecht, The Netherlands.

[Florack, 1997] Florack, L. (1997). *Image Structure*. Computational Imaging and Vision. Kluwer Academic Publishers.

[Florack et al., 1992] Florack, L., ter Haar Romeny, B., Koenderink, J., and Viergever, M. (1992). Scale and the differential structure of images. *Image Vision Comput.*, 10:367–388.

[Florack et al., 1994] Florack, L., ter Haar Romeny, B., Koenderink, J., and Viergever, M. (1994). Linear scale-space. *Journal of Mathematical Imaging and Vision*.

[Fritsch et al., 1994] Fritsch, D., Pizer, S., Morse, B., Eberly, D., and Liu, A. (1994). The multiscale medial axis and its applications in image regestration. *Pattern Recognition Letters*, 15:445–452.

[Galatsanos and Katsaggelos, 1992] Galatsanos, N. and Katsaggelos, A. (1992). Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation. *IEEE Trans. Image Processing*, 1(3):322–336.

[Gardiner, 1985] Gardiner, C. (1985). *Handbook of Stochastic Methods*. Springer.

[Goutte, 1997] Goutte, C. (1997). *Statistical learning and regularisation for regression*. PhD thesis, Université Paris 6.

[Griffin et al., 1992] Griffin, L., Colchester, A., and Robinson, G. (1992). Scale and segmentation of images using maximum gradient paths. *Image and Vision Computing*, 10(6):389–402.

[Hadamard, 1902] Hadamard, J. (1902). Sur les problemes aux derivees partielles et leur signification physique. *Nul. Univ. Princeton*, 13(49).

[Haralick, 1983] Haralick, R. (1983). Ridges and valleys in digital images. *CVGIP*, 22:28–32.

[Honerkamp, 1990] Honerkamp, J. (1990). *Stochastische Dynamische Systeme*. VCH, Weinheim, Germany.

[Hubel and Wiesel, 1962] Hubel, D. and Wiesel, T. (1962). Integrative action in the cat's lateral geniculate body. *J. Physiol. (Lond.)*, 155:385–398.

[Hubel and Wiesel, 1968] Hubel, D. and Wiesel, T. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol. (Lond.)*, 195:215–243.

[Iijima, 1959] Iijima, T. (1959). Basic theory of pattern observation. In *Papers of Technical Group on Automata and Automatic Control*, Japan. IECE.

[Iijima, 1962a] Iijima, T. (1962a). Basic theory on the normalization of a pattern (in case of a typical one-dimensional pattern). *Bulletin of the Electrotechnical Laboratory*, 26:368–388.

[Iijima, 1962b] Iijima, T. (1962b). Observation theory of two-dimensional visual patterns. In *Papers of Technical Group on Automata and Automatic Control*, Japan. IECE.

[Iijima, 1963] Iijima, T. (1963). Basic theory on the normalization of a two-dimensional visual pattern. *Studies on Information and Control, Pattern Recognition Issue, IECE*, 1:15–22.

[Jacobson, 1951] Jacobson, H. (1951). The informational capacity of the human eye. *Science*, 113:292–293.

[Kanatani, 1990] Kanatani, K. (1990). *Group-theoretical methods in image understanding*, volume 20 of *Series in Information Sciences*. Springer-Verlag.

[Koenderink, 1984] Koenderink, J. (1984). The structure of images. *Biol. Cybern.*, 50:363–370.

[Koenderink, 1990] Koenderink, J. (1990). *Solid Shape*. MIT Press.

[Koenderink and van Doorn, 1992] Koenderink, J. and van Doorn, A. (1992). Generic neighborhood operators. *IEEE PAMI*, 14(6):597–605.

[Koenderink and van Doorn, 1993] Koenderink, J. and van Doorn, A. (1993). Local features of smooth shapes: Ridges and courses. In *SPIE Proc. Geometric Methods in Computer Vision II*, volume 2031, pages 2–13.

[Koenderink and van Doorn, 1994] Koenderink, J. and van Doorn, A. (1994). Two-plus-one-dimensional differential geometry. *Pattern Recognition Letters*, 15(5):439–444.

[Koenderink and van Doorn, 1999] Koenderink, J. and van Doorn, A. (1999). Blur and disorder. In Nielsen, M., Johansen, P., Olsen, O., and Weickert, J., editors, *Scale-Space Theories in Computer Vision, Scale-Space99*, volume 1682 of *Lecture Notes in Computer Science*. Springer.

[Koller, 1995] Koller, T. M. (1995). *From Data to Information: Segmentation, Description and Analysis of the Cerebral Vascularity*. PhD thesis, Swiss Federal Institute of Technology, Zürich.

[Koller et al., 1995] Koller, T. M., Gerig, G., Szekely, G., and Dettwiler, D. (1995). Multiscale detection of curvilinear structures in 2d and 3d medical images. In *Fifth International Conference on Computer Vision ICCV 95*, pages 864–869, Cambridge, MA, USA.

[Kornhuber, 1973] Kornhuber, H. (1973). Neural control of input into long term memory: limbic system and amnestic syndrome in man. In Zippel, H., editor, *Memory and transfer of information*, pages 1–22, New York. Plenum Press.

[Lindeberg, 1990] Lindeberg, T. (1990). Scale-space for discrete signals. *IEEE PAMI*, 12:234–254.

[Lindeberg, 1993a] Lindeberg, T. (1993a). Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus of attention. *International Journal of Computer Vision*, 11(3):283–318.

[Lindeberg, 1993b] Lindeberg, T. (1993b). On scale selection for differential operators. In Heia, K., Hœgdra, A., and Braathen, B., editors, *8th Scandinavian Conference on Image Analysis*, pages 857–866, Tromsœ, Norway.

[Lindeberg, 1994a] Lindeberg, T. (1994a). Scale-space behaviour and invariance properties of differential singularities. In Yang, Y., Toet, A., Heijmanns, H., Foster, D., and Meer, P., editors, *Proc. of the NATO Advanced Research Workshop Shape in Picture*, pages 591–600, Driebergen, Netherlands.

[Lindeberg, 1994b] Lindeberg, T. (1994b). *Scale-Space Theory in Computer Vision*. The Kluwer International Series in Engineering and Computer Science. Kluwer Academic Publishers, Dordrecht, The Netherlands.

[Lindeberg, 1996] Lindeberg, T. (1996). Edge detection and ridge detection with automatic scale selection. In *Proc. Conf. on Computer Vision and Pattern Recognition*, pages 465–470, San Francisco, California.

[Lindeberg, 1997] Lindeberg, T. (1997). On the axiomatic formulations of linear scale-space. In Sporring, J., Nielsen, M., Florack, L., and Johansen, P., editors, *Gaussian scale-space theory*, pages 75–97. Kluwer Academic Publishers.

[Lindeberg, 1998a] Lindeberg, T. (1998a). Edge detection and ridge detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):117–156.

[Lindeberg, 1998b] Lindeberg, T. (1998b). Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):77–116.

[Lorensen and Cline, 1987] Lorensen, W. and Cline, H. (1987). Marching cubes: A high resolution 3d surface construction algorithm. *Computer Graphics*, 21(4):163–169.

[Lorenz et al., 1997a] Lorenz, C., Carlsen, I.-C., Buzug, T., Fassnacht, C., and Weese, J. (1997a). A multi-scale line filter with automatic scale selection based on the hessian matrix for medical image segmentation. In ter Haar Romeny, B., Florack, L., Koenderink, J., and Viergever, M., editors, *Scale-space theory in Computer Vision, ScaleSpace t'97*, volume 1252 of *Lecture Notes in Computer Science*. Springer.

[Lorenz et al., 1997b] Lorenz, C., Carlsen, I.-C., Buzug, T., Fassnacht, C., and Weese, J. (1997b). Multi-scale line segmentation with automatic estimation of width, contrast and tangential direction in 2d and 3d medical images. In *Computer Vision, Virtual Reality and Robotics in Medicine, CVRMed, Grenoble*, volume 1205 of *Lecture Notes in Computer Science*, pages 233–242. Springer.

[Majer, 1999] Majer, P. (1999). Self-similarity of noise in scale-space. In Nielsen, M., Johansen, P., Olsen, O., and Weickert, J., editors, *Scale-Space Theories in Computer Vision, Scale-Space99*, volume 1682 of *Lecture Notes in Computer Science*. Springer.

[Marr, 1982] Marr, D. (1982). *Vision*. W.H. Freeman, N.Y.

[Marr and Hildreth, 1980] Marr, D. and Hildreth, E. (1980). Theory of edge detection. *Proc. Roy. Soc. London B*, 207:187–217.

[Morse et al., 1994] Morse, B., Pizer, S., and Liu, A. (1994). Multi-scale medial analysis of medical images. *Image and Vision Computing*, 12(6):327–338.

[Nielsen et al., 1997] Nielsen, M., Florack, L., and Deriche, R. (1997). Regularization, scale-space, and edge detection filters. *Journal of Mathematical Imaging and Vision*, 7:291–307.

[Otsu, 1981] Otsu, N. (1981). *Mathematical studies on feature extraction in pattern recognition*. PhD thesis, Electrotechnical Laboratory, 1-1-4, Umezono, Sakura-mura, Niihari-gun, Ibaraki, Japan.

[Pauwels et al., 1995] Pauwels, E., Gool, L. V., Fiddelaers, P., and Moons, T. (1995). An extended class of scale-invariant and recursive scale-space filters. *IEEE PAMI*, 17:691–701.

[Pentland, 1984] Pentland, A. (1984). Fractal-based description of natural scenes. *IEEE PAMI*, 6:661–674.

[Perona and Malik, 1990] Perona, P. and Malik, J. (1990). Scale space and edge detection using anistropic diffusion. *IEEE PAMI*, 12(7):629–639.

[Pizer et al., 1994] Pizer, S., Burbeck, C., Coggins, J., Fritsch, D., and Morse, B. (1994). Object shape before boundary shape: Scale-space medial axis. *J. Math. Im. Vis.*, 4:303–313.

[Pizer et al., 1998] Pizer, S., Eberly, D., Morse, B., and Fritsch, D. (1998). Zoom-invariant figural shap: the mathematics of cores. *Computer Vision and Image Understanding*, 69:55–71.

[Pope and Lowe, 1994] Pope, A. and Lowe, D. (1994). Vista: A software environment for computer vision research. In *Proceedings, CVPR '94, IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 768–772, Seattle, WA.

[Press et al., 1988] Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1988). *Numerical Recipes in C*. Cambridge University Press.

[Ruderman and Bialek, 1994] Ruderman, D. and Bialek, W. (1994). Statistics of natural images: Scaling in the woods. *Phys. Rev. Lett.*, 73(6):814–817.

[Sporring, 1999] Sporring, J. (1999). *Measuring and Modelling Image Structure*. PhD thesis, University of Copenhagen.

[Sporring and Weickert, 1997] Sporring, J. and Weickert, J. (1997). On generalized entropies and scale-space. In ter Haar Romeny, B., Florack, L., Koenderink, J., and Viergever, M., editors, *Scale-space theory in Computer Vision, ScaleSpace t'97*, volume 1252 of *Lecture Notes in Computer Science*. Springer.

[Sporring and Weickert, 1999] Sporring, J. and Weickert, J. (1999). Information measures in scale-space. *IEEE PAMI*, 45:1051–1058.

[Staal et al., 1999] Staal, J., Kalitzin, S., ter Haar Romeny, B., and Viergewer, M. (1999). Detection of critical structures in scale space. In Nielsen, M., Johansen, P., Olsen, O., and Weickert, J., editors, *Scale-Space Theories in Computer Vision*.

[Steenstrup et al., 1999] Steenstrup, K., Pedersen, and Nielsen, M. (1999). The hausdorff dimension and scale-space normalisation of natural images. In Nielsen, M., Johansen, P., Olsen, O., and Weickert, J., editors, *Scale-Space Theories in Computer Vision*, volume 1682 of *Lecture Notes in Computer Science*. Springer.

[Sziklai, 1956] Sziklai, G. (1956). Some studies on the speed of visual perception. *I.R.E. Trans. Inf. Theory*, IT-2:125–128.

[ter Haar Romeny, 1994] ter Haar Romeny, B. M., editor (1994). *Geometry-Driven Diffusion in Computer Vision*. Computational Imaging and Vision. Kluwer Academic Publishers.

[Thirion and Gourdon, 1996] Thirion, J.-P. and Gourdon, A. (1996). The 3d marching lines algorithm. *Graphical Models and Image Processing*, 58(6):503–509.

[Tikhonov and Arsenin, 1977] Tikhonov, A. and Arsenin, V. A. (1977). *Solutions of ill-posed problems*. Winston, Washington D.C.

[van Ginneken and ter Haar Romeny, 1999] van Ginneken, B. and ter Haar Romeny, B. (1999). Applications of locally orderless images. In Nielsen, M., Johansen, P., Olsen, O., and Weickert, J., editors, *Scale-Space Theories in Computer Vision*, volume 1682 of *Lecture Notes in Computer Science*. Springer.

[van Kampen, 1981] van Kampen, N. (1981). *Stochastic Processes in Physics and Chemistry*. North-Holland.

[Watt and Watt, 1992] Watt, A. and Watt, M. (1992). *Advanced Animation and Recdering Techniques*. ACM Press.

[Weickert, 1998] Weickert, J. (1998). *Anisotropic Diffusion in Image Processing*. European Consortium for Mathematics in Industry. B.G. Teubner, Stuttgart, Germany.

[Weickert et al., 1997] Weickert, J., Ishikawa, S., and Imiya, A. (1997). Scale-space has been discovered in japan. *Journal of Mathematical Imaging and Vision*, 10(3):237–252.

[Witkin, 1983] Witkin, A. (1983). Scale-space filtering. In *Proc. 8th Int. Joint Conf. Art. Intell.*, Karlsruhe, Germany.

[Wyvill et al., 1986] Wyvill, B., McPheeters, C., and Wyvill, G. (1986). Data structure for soft objects. *The Visual Computer*, 2(4):227–234.

[Yuille and Poggio, 1986] Yuille, A. and Poggio, T. (1986). Scaling theorems for zero-crossings. *IEEE PAMI*, 8:15–25.