To appear in *Int. J. of Computer Vision.*

Direct computation of shape cues using scale-adapted spatial derivative operators

Jonas Gårding and Tony Lindeberg

Computational Vision and Active Perception Laboratory (CVAP)*
Department of Numerical Analysis and Computing Science
KTH (Royal Institute of Technology), S-100 44 Stockholm, Sweden
Email: jonasg@bion.kth.se, tony@bion.kth.se

February 1993; updated December 1994

Abstract

This paper addresses the problem of computing cues to the three-dimensional structure of surfaces in the world directly from the local structure of the brightness pattern of either a single monocular image or a binocular image pair.

It is shown that starting from Gaussian derivatives of order up to two at a range of scales in scale-space, local estimates of (i) surface orientation from monocular texture foreshortening, (ii) surface orientation from monocular texture gradients, and (iii) surface orientation from the binocular disparity gradient can be computed without iteration or search, and by using essentially the same basic mechanism.

The methodology is based on a multi-scale descriptor of image structure called the windowed second moment matrix, which is computed with adaptive selection of both scale levels and spatial positions. Notably, this descriptor comprises two scale parameters; a local scale parameter describing the amount of smoothing used in derivative computations, and an integration scale parameter determining over how large a region in space the statistics of regional descriptors is accumulated.

Experimental results for both synthetic and natural images are presented, and the relation with models of biological vision is briefly discussed.

*We would like to thank Jan-Olof Eklundh for continuous support and encouragement, as well as Narendra Ahuja at University of Illinois, and John P. Frisby at University of Sheffield for kindly providing several of the images used in the paper. This work was partially performed under the Esprit-BRA project InSight and the Esprit-NSF collaboration Diffusion. The support from the Swedish National Board for Industrial and Technical Development, NUTEK, and the Swedish Research Council for Engineering Sciences, TFR, is gratefully acknowledged. The first author has carried out part of this work while visiting the AIVRU group at University of Sheffield, and he is grateful for their hospitality as well as for the financial support of the Foundation Blanceflor Boncompagni-Ludovisi, née Bildt, and the Swedish Institute.

Contents

1 Introduction

Virtually all methods for inferring properties of the three-dimensional world from one or more images require an initial stage of retinotopic processing in which the raw image brightness pattern is transformed into some more useful representation. In practical computer vision applications this representation is often tailored for the specific task at hand, but a number of attempts have been made at defining general principles for the structure of a more general-purpose set of low-level operators capable of computing useful representations without any specific prior knowledge of the image structures to be processed.

One such approach, based primarily on theoretical considerations, is the *scale-space representation*, introduced by Witkin (1983) and Koenderink (1984). Perhaps the most important conclusion of this theory is that if the low-level operators are unbiased in the sense that they do not single out particular locations, orientations, or sizes, then the only permissible linear operations are convolutions with Gaussian kernels and their derivatives at various scales (Koenderink and van Doorn, 1992; Florack et al., 1992; Lindeberg, 1994a).

An alternative approach is to try to emulate the structure and characteristics of the early stages of primate vision, either for the purpose of gaining a better understanding of it, or simply because the performance of biological vision systems is superior to that of existing computer vision systems. This approach has generated many interesting and useful results, despite the fact that the current understanding of biological vision is far from complete. For example, general considerations regarding the information processing requirements of the visual system led Marr (1976) to propose the computation of a *primal sketch* in which low-level features of the brightness pattern, such as bars and blobs, are explicitly represented. Other models, e.g. (Turner, 1986; Bergen and Adelson, 1988; Malik and Perona, 1990), have been based on neurobiological studies of the structure of the receptive fields in the mammalian retina and the primary visual cortex. These models have been quite successful at predicting human pre-attentive texture discrimination, and have largely replaced the earlier texton theory by Julesz (1981). Interestingly, the theoretical scale-space approach and the more empirical receptive field approach are to a certain extent in agreement; simple receptive fields in the mammalian retina and primary visual cortex are well described by Gaussian derivatives (Young, 1985, 1987; Jones and Palmer, 1987a, 1987b) but also by similar models such as Gabor functions.

Retinotopic processing models are often based on considerations of relatively low-level visual tasks, such as feature detection and two-dimensional texture discrimination. One might therefore be led to think that visual tasks concerning three-dimensional interpretations of the environment require a qualitatively different type of information processing, which would have little in common with such basic operations as can be performed by a single cell or processing unit. In this paper, however, we show that at least some visual tasks of this type can be implemented as bottom-up retinotopic processing sequences, without the need for iterations, search, or a priori knowledge.

More specifically, we consider the task of estimating the shape and orientation of three-dimensional surfaces in the scene from (i) perspective distortion of surface texture observed in a monocular image, and (ii) the gradient of disparity observed in a binocular image pair. We show that this can be achieved using in principle only the following types of visual front-end operations (Lindeberg, 1993b): (large support) diffusion smoothing, (small support) derivative computations from smoothed brightness data, and (pointwise) non-linear combinations of these derivatives.

The framework is based on the computation of a local (regional) descriptor of the structure of the brightness pattern, referred to as the *windowed second moment matrix*, which describes the local variance of blurred first-order directional Gaussian derivatives. We emphasize and analyze the need for two different scale parameters; a *local scale* parameter describing the amount of smoothing used for suppressing irrelevant fine scale structures when computing pointwise non-linear descriptors of the image brightness pattern, and a second *integration scale* parameter describing the size of the spatial window used for accumulating statistics of the pointwise descriptors.

Thus, the multi-scale nature of image structures is explicitly taken care of, and is built into the representation. We do not attempt to make the representation "complete" in the sense of allowing reconstruction of the original image from the descriptors. On the contrary, we emphasize adaptive *selection* of both scale levels and spatial positions, for the purpose of providing an explicit representation of precisely the information needed by the later stage processes. Moreover, the representation is normalized in such a way that selection of interesting scale levels and spatial positions is achieved simply through detection of local extrema with respect to scale and position of the computed non-linear entities.

The presentation is organized as follows. Section 2 provides a formal definition and description of the basic multi-scale image texture descriptor we propose. Section 3 describes scale problems arising in this context. The notions of local scale and integration scale are formalized, and it is shown how relevant scale values for these two scale parameters can be automatically selected. Section 4 demonstrates how the basic principles for scale selection can be applied to spatial selection, resulting in what can be viewed as a multi-scale blob detection method. These components are then combined in Section 5, which reviews the shape-from-texture problem and demonstrates how estimates of surface shape and orientation can be computed directly from the multi-scale texture descriptor. Section 6 treats the problem of estimating shape from gradients of binocular disparity, and demonstrates that the proposed approach can be successfully applied to this problem as well. Finally, in Section 7 some general conclusions are made, and their implications are discussed.

2 A local texture descriptor

The task of computing meaningful texture descriptors is often referred to in the literature as extraction of texture elements or "texels". Considering the great variability of natural textures, it is not surprising that there is no generally accepted definition of precisely what a texel is. A first and rather obvious requirement on a texel definition is that it must be computable for a large class of natural images, but this still leaves many degrees of freedom.

Here, we shall take a functional approach to texel extraction. Rather than postulating any particular structure of the texture, we consider the requirements of the higher-level processes that need to use the local texture description. The basic principle of shape-from-texture estimation is to use the observed perspective distortion of the texture pattern to estimate the parameters of the distorting transformation, which in turn allow properties of surface and/or viewing geometry to be inferred. The principle of shape-from-disparity-gradient estimation is analogous, the difference being that it uses the distortion from the right to the left image, rather than the distortion from a surface to its image. Hence, for both these processes, the texture description must reflect perspective distortion of the texture in a predictable way, so that the parameters of the distorting transformation can

be recovered from the texture description.

A great simplification of the problem comes from the observation that for many purposes it is sufficient to recover the *linear* part of the perspective distortion. The analysis behind this observation is given in Sections 5 and 6; for the moment we take it as a given fact.

2.1 The windowed second moment matrix

We propose that a texture descriptor expressed in the form of a two-dimensional *second moment matrix* is well suited for the purpose of estimating local linear distortion. Such a second moment matrix can be thought of e.g. as a covariance matrix of a two-dimensional random variable, or, with a mechanical analogy, as the moment of inertia of a mass distribution in the plane. It can be graphically represented by an ellipse, and as will be shown, a linear transformation applied to the spatial coordinates affects the ellipse precisely as it would affect a physical ellipse painted on the surface.

Various forms of second moment descriptors have previously been successfully applied to a number of visual tasks. For estimation of shape from texture, Brown and Shvaytser (1990) used the second moment of the image brightness autocorrelation function to estimate foreshortening, Gårding (1991, 1992) used the second moment of the local Fourier spectrum to estimate foreshortening and texture gradients. Super and Bovik (1992) used the same moment to estimate relative foreshortening. Second moments of the directional statistics of image contours have been used by Kanatani (1984), Blake and Marinos (1990a) and Gårding (1993) for estimation of foreshortening. Moreover, second moment descriptors of brightness gradients have been used by Bigün et al. (1991) and Rao and Schunk (1991) for analysis of oriented or flow-like texture patterns, as well as by Förstner and Gülch (1987) as an "interest" operator in the context of junction detection and stereo matching.

Here, we shall use a particular type of second moment matrix similar to some of those described in the above cited articles. It is defined as follows (Lindeberg and Gårding, 1993): Let $L : \mathbb{R}^2 \to \mathbb{R}$ be the image brightness, and let $\nabla L = (L_x, L_y)^T$ be its gradient. We now define the second moment descriptor[1] $\mu_L : \mathbb{R}^2 \to SPSD(2)$ of L by

$$\mu_L(q) = \begin{pmatrix} \mu_{11} & \mu_{12} \\ \mu_{21} & \mu_{22} \end{pmatrix} = E_q \begin{pmatrix} L_x^2 & L_x L_y \\ L_x L_y & L_y^2 \end{pmatrix} = E_q((\nabla L)(\nabla L)^T), \tag{1}$$

where E_q denotes an averaging operator centered at $q = (x, y)^T \in \mathbb{R}^2$. $\mu_L(q)$ has a number of convenient properties. Clearly, it is invariant to translations, and it can easily be shown that the trace and determinant of μ_L are also invariant to rotations. Moreover, uniform rescaling in the spatial domain and affine brightness transformations only affect μ_L by a uniform scaling factor.

We define the averaging operator E_q as the local weighted mean using a symmetric and normalized window function $w : \mathbb{R}^2 \to \mathbb{R}$. Hence, the components μ_{ij} of $\mu_L(x, y)$ can be expressed as

$$\mu_{ij}(x, y) = \iint_{(x', y') \in \mathbb{R}^2} w(x - x', y - y')\, L_{x_i}(x', y')\, L_{x_j}(x', y')\, dx'\, dy', \tag{2}$$

The invariance properties are preserved provided that w is rotationally symmetric (see below) and has a nice scaling behaviour. A natural choice of window function is the Gaussian;

[1] The notation $SPSD(2)$ stands for the cone of symmetric positive semidefinite 2×2 matrices.

in fact, as described in Section 3.1 this is the *only* translationally invariant choice that leads to scale-space behaviour of μ_L.

2.1.1 Spatial frequency interpretation

μ_L can also be understood in terms of the spatial frequency distribution of $L(x, y)$. Rename temporarily the coordinates $(x, y)^T$ to $(x_1, x_2)^T$, and let $\Phi_L : \mathbb{R}^2 \to \mathbb{R}$ be the power spectrum of L, i.e.,

$$\Phi_L(\omega_1, \omega_2) = \hat{L}(\omega_1, \omega_2)\, \hat{L}^*(\omega_1, \omega_2), \tag{3}$$

where $\hat{L} : \mathbb{R}^2 \to \mathbb{C}$ denotes the Fourier transform of L

$$\hat{L}(\omega_1, \omega_2) = \int_{(x_1, x_2) \in \mathbb{R}^2} L(x_1, x_2)\, e^{-(\omega_1 x_1 + \omega_2 x_2)} dx_1\, dx_2 \tag{4}$$

and $\hat{L}^*$ its complex conjugate. Using Plancherel's relation it follows that

$$\iint_{(x_1, x_2) \in \mathbb{R}^2} L_{x_i} L_{x_j}\, dx_1\, dx_2 = \frac{1}{(2\pi)^2} \iint_{(\omega_1, \omega_2) \in \mathbb{R}^2} \omega_i \omega_j\, \Phi_L(\omega_1, \omega_2)\, d\omega_1\, d\omega_2. \tag{5}$$

Hence, if $L \in \mathbb{L}_2(\mathbb{R}^2)$, the inner products of the first derivatives are proportional to the components of the second moment of the power spectrum.

2.1.2 Visualization by ellipses

Since the second moment matrix is positive semidefinite, it follows that the equation

$$(\xi - q)^T \mu_L(q)\, (\xi - q) = 1 \quad (\xi, q \in \mathbb{R}^2) \tag{6}$$

defines an ellipse (possibly degenerated to a line) centered at q. The semi-axes of this ellipse are the square roots of the inverse of the eigenvalues (λ_1, λ_2) of $\mu_L(q)$, while the orientations of the axes give the directions of the corresponding eigenvectors (see Figure 1). It is easily verified that the distance from the center to the perimeter of the ellipse in some direction is equal to the inverse of the average squared magnitude of the directional derivative of $L(x, y)$ in that direction.

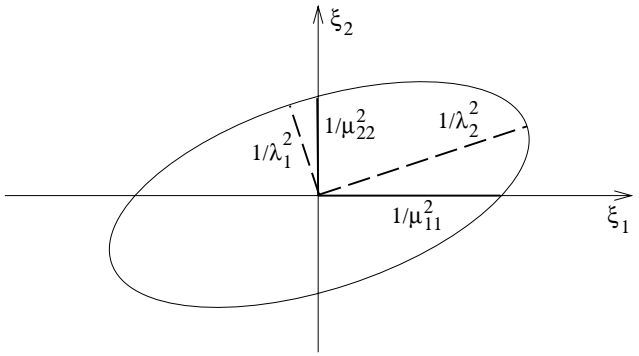

Figure 1: The ellipse representation of the second moment matrix μ_L. For simplicity, the ellipse is shown centered at the origin of the coordinate system.

2.2 Transformation properties

As mentioned in the beginning of this section, the (linear) transformation properties of the local texture descriptor are crucial to the higher-level processes (shape-from-texture and shape-from-disparity-gradients) that are going to operate on the description. Because these processes attempt to recover the parameters of the transformation from the properties of the texture descriptors, the descriptors must be affected in a predictable way by a linear transformation $B : \mathbb{R}^2 \to \mathbb{R}^2$, representing e.g. the linearized perspective mapping from the surface to the image in the shape-from-texture case, or the linearized projective mapping from the left to the right image in the shape-from-disparity-gradient case.

For the windowed second moment matrix the relation is straightforward. Given a brightness pattern L, let $R : \mathbb{R}^2 \to \mathbb{R}$ represent the brightness pattern subjected to an invertible linear transformation of the spatial coordinates $\eta = B\xi$, i.e.,

$$L(\xi) = R(B\xi) \tag{7}$$

where $\xi, \eta \in \mathbb{R}^2$. Moreover, let $\mu_R(p) \in SPSD(2)$ be the local second moment of R at the point $p = Bq$ computed with respect to the "backprojected" normalized window function $w' : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$w'(\eta - p) = (\det B)^{-1} w(B^{-1}(\eta - p)) = (\det B)^{-1} w(\xi - q). \tag{8}$$

It is then straightforward to show that (see Appendix A.1)

$$\mu_L(q) = B^T \mu_R(p) B. \tag{9}$$

In the rest of this section, the arguments p and q to μ_L and μ_R will be dropped to simplify the notation.

It is easily verified that (9) also describes the effect of the coordinate transformation B to the ellipse (6) representing $\mu_L(q)$. Hence, it is justifiable to think of μ_L as analogous to an ellipse that is "painted" on the surface. This analogy often provides sufficient intuition to directly predict the behaviour of μ_L in various situations.

If μ_L and μ_R are known, then the linear transformation B is clearly constrained by (9). However, it is not determined uniquely, since μ_L and μ_R are symmetric and hence only contain three independent components, whereas B may contain four unknown parameters. It can be shown (Gårding, 1991) that the general solution to $\mu_L = B^T \mu_R B$ is

$$B = \mu_R^{-1/2} W^T \mu_L^{1/2} \tag{10}$$

where W is an arbitrary orthogonal matrix, and the notation $\mu^{1/2}$ indicates some (e.g., the unique positive semidefinite symmetric) solution to the equation $X^2 = \mu$.

Fortunately, in the applications to shape estimation from texture and disparity gradients considered in this paper, the ambiguity represented by the rotation matrix W is eliminated by geometric constraints which reduce the degrees of freedom of the linear transformation B (see Sections 5.3 and 6.2).

2.3 The structure of the second moment descriptor

In this section we shall take a closer look at the structure of $\mu_L(q)$, and define a number of derived entities that will turn out to be useful later on.

For any two-dimensional second moment matrix μ, the following entities can be defined from its components μ_{ij}:

$$P = \mu_{11} + \mu_{22}, \qquad C = \mu_{11} - \mu_{22}, \qquad S = 2\,\mu_{12}. \tag{11}$$

Applied to μ_L (with the argument q dropped), these definitions can be rewritten:

$$P = E_q(L_x^2 + L_y^2), \qquad C = E_q(L_x^2 - L_y^2), \qquad S = 2\,E_q(L_x L_y). \tag{12}$$

The first descriptor $P : \mathbb{R}^2 \to \mathbb{R}$ is a natural measure of the strength of operator response; it is the average of the square of the gradient magnitude in a region around q. The two other entities $C, S : \mathbb{R}^2 \to \mathbb{R}$ contain directional information, and it is natural to treat them as a vector (C, S), the magnitude of which is

$$Q = \sqrt{C^2 + S^2}. \tag{13}$$

We also define the normalized entities

$$\tilde{C} = C/P, \quad \tilde{S} = S/P, \quad \tilde{Q} = Q/P. \tag{14}$$

It can easily be shown that $\tilde{Q} \in [0, 1]$; it holds that $\tilde{Q} = 0$ if and only if $E_q(L_x^2) = E_q(L_y^2)$ and $E_q(L_x L_y) = 0$, while $\tilde{Q} = 1$ if and only if $(E_q(L_x L_y))^2 = E_q(L_x^2) E_q(L_y^2)$. $\tilde{Q}$ is a natural measure of the *anisotropy* of $\mu_L(q)$; in terms of the ellipse representation, $\tilde{Q} = 0$ corresponds to a circle, and $\tilde{Q} = 1$ to a line. For example, a rotationally symmetric brightness pattern has $\tilde{Q} = 0$, while a translationally symmetric pattern[2] has $\tilde{Q} = 1$. Rotational symmetry is, however, not necessary in order to obtain $\tilde{Q} = 0$. For example, any pattern with $N \geq 2$ uniformly distributed dominant (unsigned) directions also satisfies $\tilde{Q} = 0$. A second moment matrix with $\tilde{Q} = 0$ will be referred to as *weakly isotropic*.

Q and P are invariant under rotations of the coordinate system provided that the window function w is rotationally symmetric, and they allow the differential invariants of μ_L to be succinctly expressed as follows:

$$\begin{aligned} \operatorname{trace} \mu_L &= P, \\ \det \mu_L &= \frac{1}{4}(P^2 - Q^2) = \frac{1}{4} P^2 (1 - \tilde{Q}^2), \\ \lambda_{1,2} &= \frac{1}{2}(P \pm Q) = \frac{1}{2} P (1 \pm \tilde{Q}), \end{aligned} \tag{15}$$

where $\lambda_1 \geq \lambda_2$ are the eigenvalues of μ_L.

The normalized components $(\tilde{C}, \tilde{S})^T$ can also be understood as representing the local statistics of unsigned gradient directions. A standard technique (Mardia, 1972) for computing statistics of unsigned directions in $\mathbb{R}^2$ is to map a direction angle α to the point $(\cos 2\alpha, \sin 2\alpha)^T$ on the unit circle. This mapping has the desired property that α and $\alpha + \pi$ are mapped to the same point. Using this representation, map each gradient vector $(L_x, L_y)^T = \rho(\cos\alpha, \sin\alpha)^T$ to the point $(\cos 2\alpha, \sin 2\alpha)^T$, and give it a "mass" proportional to the squared gradient magnitude ρ^2 multiplied by the window function. It is then easily shown that the center of mass of this distribution is given by $(\tilde{C}, \tilde{S})^T$. Hence, the average unsigned gradient direction is $\arg(\tilde{C}, \tilde{S})/2$, which is also the direction of the eigenvector corresponding to the largest eigenvalue of μ_L; see (Lindeberg and Gårding, 1993) for more details.

[2] A (two-dimensional) translationally symmetric brightness pattern $f : \mathbb{R}^2 \to \mathbb{R}$ can be written $f(x, y) = h(ax + by)$ for some one-dimensional function $h : \mathbb{R} \to \mathbb{R}$ and some (scalar) constants a and b.

3 Representing and selecting scale

An intrinsic property of objects in the world and details in images is that they only exist as meaningful entities over certain ranges of scale. This issue is of crucial importance when using perspective distortion to derive shape cues; size variations of image structures can occur both because a surface texture contains structures at different scales, and because of perspective effects in the image formation process. Analysing image structures at wrong scales often leads to meaningless results. Concerning the computation of the windowed second moment matrix (or, indeed any other non-trivial texture descriptor which involves integration of statistics of pointwise properties over finite-sized local image neighborhoods) there are two fundamental scale problems, which manifest themselves as follows.

First, the image statistics must be collected from a region large enough to be representative of the texture. Yet, the region must not be so large that the local linear approximation of the perspective mapping becomes invalid. For example, for an ideal texture consisting of isolated blobs, a lower limit for the extent of the integration region is determined by the size of the individual blobs, while an upper limit may be given by the curvature of the surface or interference with other nearby surface patches. This scale controlling the *window function* is referred to as *integration scale* (denoted s).

Second, the image statistics must be based on descriptors computed at proper scales, so that noise and "irrelevant" image structures can be suppressed. The descriptor considered in this paper is based on first order spatial derivatives of the image brightness, and it is obvious that useful results hardly can be expected if the derivatives are computed directly from unsmoothed noisy data, although this problem disappears in ideal noise-free data if the sampling problems are handled properly. This scale determining the amount of *initial smoothing* in the (traditional first-stage) multi-scale representation of the image is referred to as *local scale*[3] (denoted t).

3.1 The multi-scale windowed second moment matrix

A general framework for handling image structures at different scales is provided by scale-space theory (Witkin, 1983; Koenderink, 1984; Babaud et al., 1986; Yuille and Poggio, 1986; Lindeberg, 1990, 1993b, 1994a; Koenderink and van Doorn, 1990, 1992; Florack et al., 1992). In summary, this theory basically states that the natural way to process a given two-dimensional continuous signal $f : \mathbb{R}^2 \to \mathbb{R}$ is by embedding it into the scale-space representation $L : \mathbb{R}^2 \times \mathbb{R}_+ \to \mathbb{R}$ defined as the solution to the diffusion equation

$$\partial_t L = \frac{1}{2} \nabla^2 L = \frac{1}{2} (\partial_{xx} + \partial_{yy}) L \tag{16}$$

with initial condition $L(\cdot;\, 0) = f(\cdot)$. Equivalently, this representation can be obtained by convolution with the Gaussian kernel $L(\cdot;\, t) = g(\cdot;\, t) * f(\cdot)$, where

$$g(x, y;\, t) = \frac{1}{2\pi t}\, e^{-(x^2+y^2)/(2t)}. \tag{17}$$

Based on this framework, a formal definition of the *multi-scale windowed second moment matrix* can be stated as

$$\mu_L(\cdot;\, t, s) = w(\cdot;\, s) * ((\nabla L)(\cdot;\, t)\, (\nabla L)(\cdot;\, t)^T), \tag{18}$$

[3] This terminology refers to local operations (derivatives). Concerning the use of two scale parameters for texture analysis, see also (Casadei et al., 1992).

where s is the integration scale parameter associated with the window function w, and t is the local scale parameter in the scale-space representation of the original image.

In Section 2.1, it was indicated that the Gaussian is a natural choice of window function in (2). This choice could, in principle, be motivated by the fact that this kernel is rotationally symmetric with a nice scaling behaviour, which means that the invariance properties described in Section 2.1 are preserved. More importantly, however, it holds that *if and only if* the window function is a Gaussian, then the components of μ_L, μ_{ij}, constitute *scale-space representations* of the components of $(\nabla L)(\nabla L)^T$, $L_{x_i} L_{x_j}$, respectively (Lindeberg, 1994a). This is a direct consequence of the uniqueness of the Gaussian kernel for scale-space representation given natural front-end postulates (e.g. the causality condition introduced by Koenderink (1984), or the scale invariance used by Florack et al. (1992)).

3.2 Scale selection: Review

The second moment matrix depends upon two scale parameters. In general, appropriate values for these parameters can be expected to vary substantially between different images, and even between different locations in a single image, depending on the type of surface texture, the distance to the surface and the noise in the image formation process. It is thus highly desirable (or even necessary) to include some automatic and adaptive mechanism for selecting appropriate scale levels.

A general method for scale selection has been proposed by Lindeberg (1993c, 1994b). It is based on the idea of studying the evolution over scales of differential invariants expressed in terms of *normalized scale-space derivatives* defined by

$$\partial_\xi = \sqrt{t}\, \partial_x \tag{19}$$

where $\xi = x/\sqrt{t}$ are normalized coordinates. More precisely, the method for scale selection states that scale levels for further processing should be selected from the scales where normalized differential entities assume maxima over scales, based on the following 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 a characteristic dimension of a corresponding structure contained in the data.

This principle is similar although not equivalent to the method for scale selection described in (Lindeberg, 1993a), where scales were selected from from maxima over scales of a normalized measure of the strength of a blob response. This principle can be justified theoretically for a general class of differential invariants as well as a number of specific local brightness models; see (Lindeberg, 1993c, 1994a) and Section 3.3, but its practical usefulness must be verified empirically. Here, we shall apply it for selecting scale levels for computing second moment descriptors.

Figure 2 illustrates the variation over scale of three differential entities related to the second moment descriptor. The graphs show from left to right the variation over scales of (i) the normalized square of the *gradient magnitude* $||\nabla_{norm} L||_2^2$, (ii) the local average of the gradient magnitude using a Gaussian window function with the integration scale proportional to the local scale (this is the *trace* of $\mu_L(q)$), and (iii) the *determinant* of $\mu_L(q)$. These graphs are called the *scale-space signatures* of the entities considered.

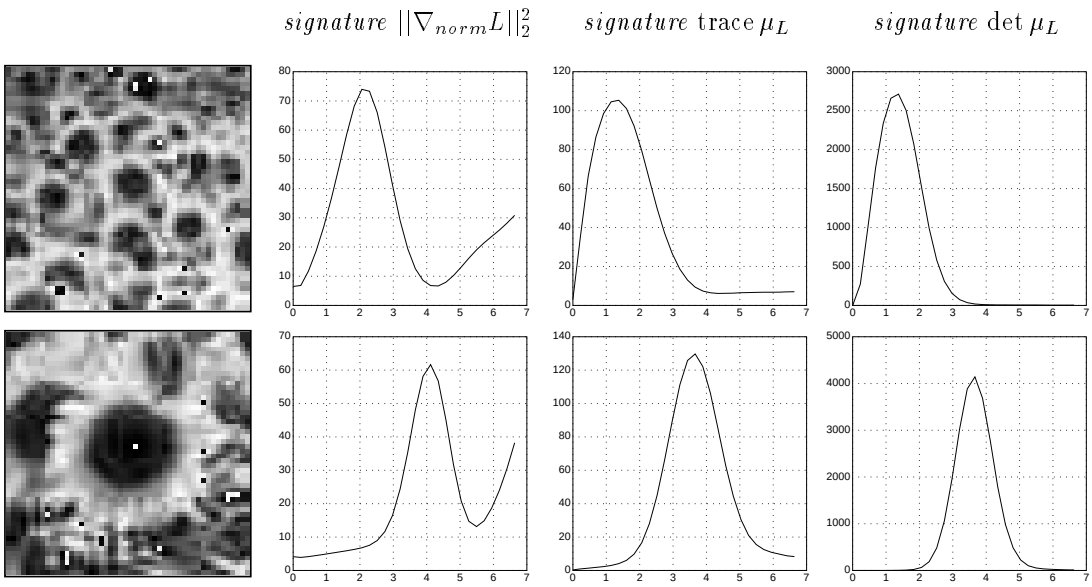

Figure 2: Scale-space signatures of the pointwise and integrated normalized gradient magnitude ($||\nabla_{norm} L||_2^2$ and trace μ_L respectively), as well as the determinant of the second moment matrix (det μ_L) for two details of a sunflower image; (left) grey-level image, (middle left) signature of $||\nabla_{norm} L||_2^2$, (middle right) signature of trace μ_L, and (right) signature of det μ_L. Observe that the maxima in the top row are assumed at finer scales than the maxima in the bottom row. (All entities are computed at the central point. The scaling of the horizontal axis is basically logarithmic, while the scaling of the vertical axis is linear.)

As can be seen, the maxima over scales in the top row of Figure 2 are obtained at finer scales than in the bottom row. Moreover, the ratio between the scale values for which the graphs attain their maxima is roughly equal to the ratio of the sizes of the sunflowers in the centers of the two images respectively, in agreement with the heuristic principle.

This principle for scale selection is not restricted to texture analysis; see (Lindeberg, 1993c, 1994b) for a more general treatment concerning other feature detection tasks, such as junction detection, blob detection, edge detection and ridge detection.

3.3 Properties of the scale selection method

This section lists some more specific properties of the heuristic principle for scale selection. A more extensive treatment can be found in the references cited above.

For two *parallel* (two-dimensional) sine waves

$$f_{par}(x, y) = \sin \omega_1 x + \sin \omega_2 x \tag{20}$$

(where $\omega_1 \leq \omega_2$) it is easy to show that for both $||\nabla_{norm} L||_2^2$ and trace μ_L there is a unique scale maximum when ω_2/ω_1 is close to one, while there are two scale maxima for sufficiently large ω_2/ω_1 ($\omega_{bifurc} \approx 2.4$). A similar result holds for two *orthogonal* sine waves,

$$f_{orth}(x, y) = \sin \omega_1 x + \sin \omega_2 y. \tag{21}$$

If the latter signal is interpreted as the orthographic projection of an isotropic pattern with

foreshortening $\epsilon = \omega_1/\omega_2$, then the interpretation is that the response changes from one to two peaks at slant $\sigma_{bifurc} = \arccos(1/\omega_{bifurc}) \approx 65°$.

The determinant of the windowed second moment matrix, $\det \mu_L$, behaves somewhat differently; it is identically zero for f_{par}, while there is *always* a unique peak in f_{orth}.

More generally, for an *isotropic* pattern (with $\tilde{Q} = 0$, or equivalently, $\lambda_1 = \lambda_2$) the scale maxima of $\operatorname{trace} \mu_L$ and $\det \mu_L$ coincide. This is easily proved from $\operatorname{trace} \mu_L = \lambda_1 + \lambda_2 = 2\lambda_1$ and $\det \mu_L = \lambda_1 \lambda_2 = \lambda_1^2$, which gives $\partial_t \det \mu_L = 0 \Leftrightarrow \partial_t \operatorname{trace} \mu_L = 0$.

For a *unidirectional* pattern (with $\tilde{Q} = 1$, or equivalently, $\lambda_2 = 0$) $\det \mu_L$ is identically zero, while $\operatorname{trace} \mu_L$ is non-zero. Hence, $\det \mu_L$ only responds when there are significant variations along *both* the coordinate directions, typically for blob-like signals.

The behaviour of the normalized derivatives can be understood also in the context of signals having a dense Fourier spectrum. For a signal f with a (fractal) power spectrum $\Phi_f = \hat{f}\hat{f}^* = |\omega|^{-2\alpha}$ it follows from Plancherel's relation that

$$P_{norm}(\cdot;\, t) = t\,(E(L_x^2(\cdot;\, t)) + E(L_y^2(\cdot;\, t))) \sim t^{\alpha - 1}. \tag{22}$$

This expression is independent of scale if and only if $\alpha = 1$. In other words, in the two-dimensional case the normalized derivative model is *neutral* with respect to power spectra of the form $|\omega|^{-2}$, which commonly occur in natural imagery (Field, 1987).

3.4 Scale selection for computing μ_L

Computation of the windowed second moment matrix μ_L requires selection of suitable values for both the local scale parameter t and the integration scale parameter s. In its most general form, the adaptive scheme we propose for setting these scales can be summarized as follows. Given any point in the image;

1. vary the two scale parameters, the local scale t and the integration scale s, according to some scheme;

2. accumulate the scale-space signature for some (normalized) differential entity;

3. detect some special property of the signature, e.g., the global maximum, or all local extrema, etc;

4. set the integration scale(s) used for computing μ_L proportional to the scale(s) where the above property is assumed;

5. compute μ_L at the fixed integration scale while varying the local scale between a minimum scale, e.g. $t = 0$, and the integration scale, and then select the most appropriate local scale(s) according to some criterion.

Our specific implementation of this general scheme is described below.

Scale variation. A completely general implementation of Step 1 would involve a full two-parameter scale variation. Here, a simpler but quite useful approach will be used; the integration scale is set to a constant times the local scale, $s = \gamma_1^2\, t$ (typically $\gamma_1 = \sqrt{2}$). In light of the scale selection heuristic, this scale invariant choice means that the size of the integration region is proportional to the characteristic length of the local smoothing kernel. For example, in the case of periodic patterns, this implies that the size of the integration

region at each local scale is proportional to the wavelength for which the normalized first derivative at that scale would give a maximum response.

Selecting integration scales. Concerning Steps 2–3, we propose to set the integration scales from the scales, denoted $s_{\det \mu_L}$, where the normalized strength of μ_L, represented by $\det \mu_L$, assumes a local or global maximum. This choice is motivated by the observation that for both simple periodic and blob-like patterns, the signature of $\det \mu_L$ has a single peak reflecting the characteristic size (area) of the two-dimensional pattern, while for the pointwise and integrated gradient magnitude the response changes from one to two peaks with increasing (linear) distortion.

Once $s_{\det \mu_L}$ has been determined, it is advantageous to compute μ_L at a slightly larger integration scale $s = \gamma_2^2 \, s_{\det \mu_L} = \gamma_1^2 \gamma_2^2 \, t_{\det \mu_L}$ (typically $\gamma_2 = 2$), in order to obtain a more stable descriptor. More formally, using $\gamma_2 > 1$ can be motivated by the analysis in (Lindeberg and Gårding, 1993; Lindeberg, 1994a) which shows that the estimates of the directional information in μ_L are more sensitive to small window sizes than are the magnitude estimates. The factor $\gamma = \gamma_1 \gamma_2$ is referred to as *relative integration scale*.

Selecting local scales. The local scale at which μ_L is computed in Step 5 should be chosen to suppress noise and irrelevant fine-scale structure without introducing excessive shape distortions due to smoothing. In simple situations it may be acceptable to set it to a fixed value reflecting the overall noise level in the image. A more general and adaptive principle is to set the local scale(s) at each point to the scales, denoted t_Q, where the *normalized anisotropy*, $\tilde{Q}$, assumes a local maximum. This is motivated by the fact that in the absence of noise and interfering finer scale structures, the main effect of the first stage scale-space smoothing is to *decrease* the anisotropy. For example, the aspect ratio of a non-uniform Gaussian blob $f(x, y) = g(x; \, l_1^2) \, g(y; \, l_2^2)$ varies as $(l_2^2 + t)/(l_1^2 + t)$, and clearly approaches one as t is increased. On the other hand, suppressing isotropic noise and interfering finer scale structures *increases* the anisotropy. Selecting the maximum point gives a natural trade-off between these two effects.

Experiments. Figure 3 illustrates these effects for a synthetic image with different amounts of additive white Gaussian noise. Note that the scale-space signature of $\det \mu_L$ has a unique maximum when the noise level, ν, is small, and two maxima when ν is increased. Table 1 gives numerical values obtained by using the proposed method for scale selection. Notice the stability of $s_{\det \mu_L}$ with respect to noise. The selected local scale t_Q increases with the noise level ν, while $\tilde{Q}$ decreases at $t = 0$.[4]

In Section 5.3 it is shown that under a certain assumption about the surface texture (weak isotropy), the estimate of surface orientation is directly related to the normalized anisotropy $\tilde{Q}$, and to the eigenvector of μ_L corresponding to the maximum eigenvalue. Table 1 illustrates the accuracy in estimates of $\tilde{Q}$ and surface orientation computed in this

[4]In these curves there is also a minimum in the signature of $\tilde{Q}$ at coarse scales. The reason why this occurs is that the higher-frequency sine component is suppressed much faster than the lower-frequency sine component. At a certain scale, the contributions to μ_L from these two components are equal (corresponding to $\tilde{Q} = 0$). Then, when the higher-frequency component is suppressed further, the local image structure asymptotically approaches a translationally symmetric pattern; see also (Lindeberg and Gårding, 1993) for a theoretical analysis.

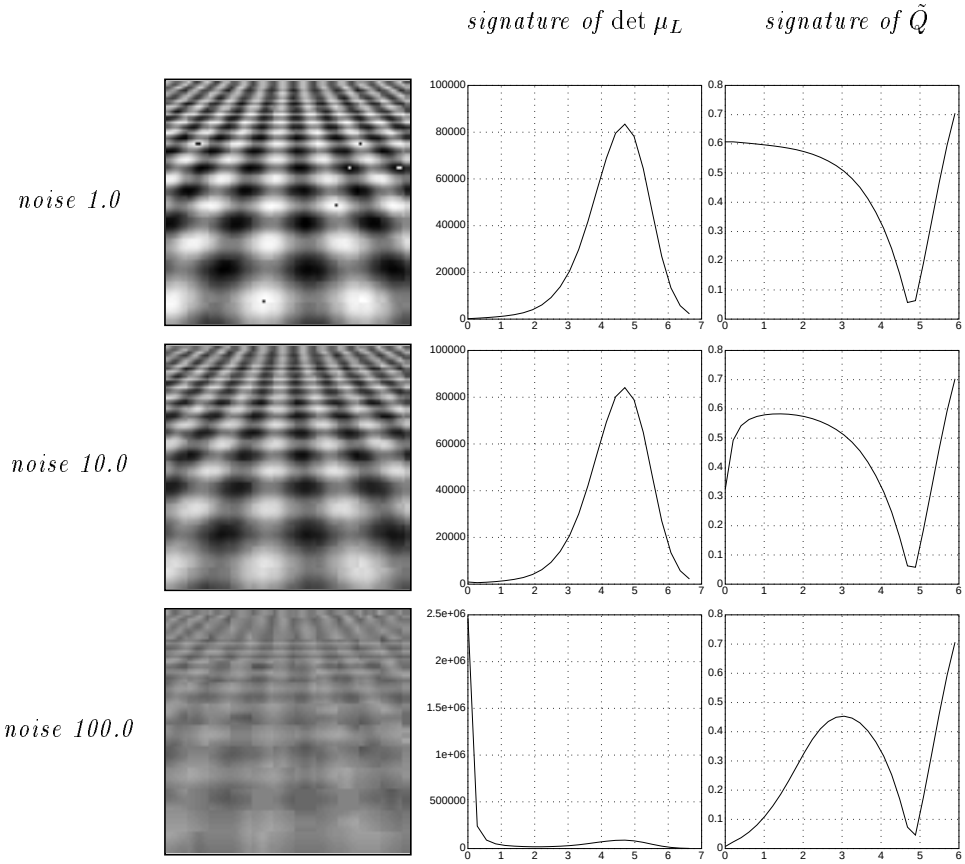

Figure 3: Scale-space signatures of $\det \mu_L$ and $\tilde{Q}$ (accumulated at the central point) for a synthetic texture with added (white Gaussian) noise of standard deviation $\nu = 1.0$ (top row), 10.0 (middle row), and 100.0 (bottom row). The range of grey-levels is [0..255]. The columns show; (left) grey-level image with noise, (middle) signature of $\det \mu_L$, and (right) signature of $\tilde{Q}$.

noise level	$s_{\det \mu_L}$	t_Q	$\tilde{Q}(t_Q)$	$\tilde{Q}(t=0)$	$\Delta\phi_n(t_Q)$	$\Delta\phi_n(t=0)$
1.0	34.9	0.0	0.602	(0.602)	0.2°	(0.2°)
10.0	34.4	2.0	0.579	(0.329)	1.1°	(15.3°)
31.6	34.1	4.2	0.510	(0.033)	4.7°	(45.3°)
100.0	31.4	8.5	0.456	(0.006)	7.8°	(53.7°)

Table 1: Numerical values of some characteristic entities obtained at the central point of the image in Figure 3 using different amounts of additive Gaussian noise and automatic scale selection. Note the stability of the selected integration scale (proportional to $s_{\det \mu_L}$) with respect to variations in the noise level ν, and that the selected local scale t_Q increases with ν. Observe also the increasing difference between the estimates of the normalized anisotropy $\tilde{Q}$ computed at the selected local scale, and at zero local scale (true value 0.600). The last two columns show the error in surface orientation $\Delta\phi_n$ computed by monocular shape-from-texture under a specific assumption about the surface texture (weak isotropy).

way. The error in surface orientation is measured by the angle $\Delta\phi_n$ between the estimated and true surface normal.

Figure 4 illustrates these results graphically, by ellipses representing the second moment matrices, with the size rescaled to be proportional to $s_{\det \mu_L}$. As a comparison, Figure 5 displays a typical result of using non-adaptive (globally constant) scale selection. Here, useful shape descriptors are only obtained in a small part; the window size is too small in the lower part, while the first stage smoothing leads to severe shape distortions in the upper part.

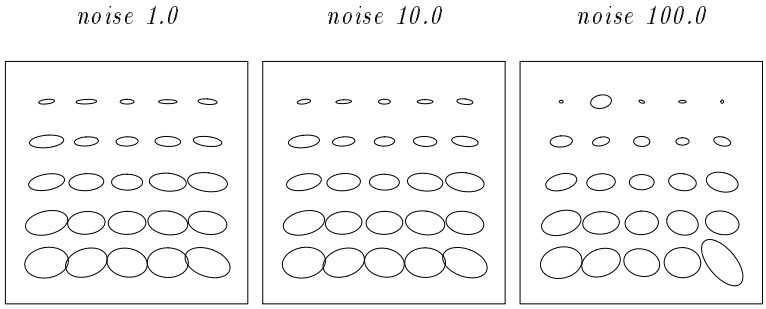

Figure 4: Ellipses representing μ_L computed at different spatial points using *automatic scale selection* of the local scale and the integration scale — note the stability with respect to variations of the noise level.

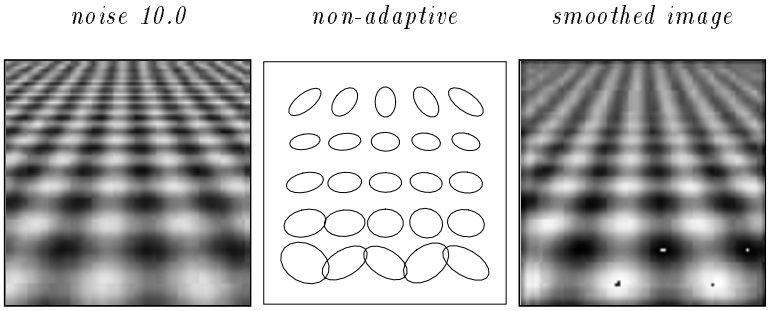

Figure 5: Typical example of the result of using *non-adaptive* selection of the (here constant) local and integration scales — geometrically useful shape descriptors are obtained only in a small part of the image.

3.5 The ellipse representation revisited

The ellipse given by (6) graphically represents the local statistics of the first-order directional derivatives computed at the local scale t and the integration scale s. In particular, the area $A = 1/\sqrt{\det \mu_L}$ of the ellipse reflects the average magnitude of these derivatives. A scaling of the image brightness by some factor k scales A by $1/k^2$, whereas the shape of the ellipse remains unchanged. Hence, ellipses computed in a dim region on average tend to be larger than those computed in areas of higher contrast. For this reason, the absolute magnitude of μ_L is not used for shape estimation.

Size information about characteristic image structures is instead available from the scale selection procedure, and for the purpose of graphically visualization we normalize μ_L by scaling its components to make the area of the ellipse proportional[5] to the scale at which the maximum of $\det \mu_L$ is assumed.

4 Spatial selection and blob detection

The previous sections treated the problem of selecting appropriate scales for local smoothing and regional integration at a given image point. In this section, we shall consider the complementary problem of selecting *where* in the image to apply the multi-scale analysis. This problem is referred to as *spatial selection*.

Spatial selection could in principle be avoided by computing a texture descriptor at every image point, but this is typically not an acceptable solution; it can lead to unnecessarily poor estimates since many image points often contain little or no useful image structure.[6] In particular, many natural textures seem to consist of fairly similar texture elements randomly scattered on the surface. This is quite unlike the idealized case of a perfectly periodic texture, in which all image points provide more or less the same information provided that integration is performed over one period of the pattern.

Here, we shall use the scale selection method for guiding spatial selection process as well. The resulting simultaneous selection of scale and spatial position can be interpreted as a form of *multi-scale blob detector*, where each detected blob is represented by its position, its detection scale, and a second moment matrix. This multi-scale blob detector has obvious limitations compared to more general approaches, e.g. (Blostein and Ahuja, 1989a; Lindeberg, 1993a), since it only represents the shape of each blob by a second moment matrix. However, it is well suited as a pre-processing step for the shape estimation processes described in Sections 5 and 6, since it produces precisely the information needed for estimating local linear distortion and size changes.

4.1 Spatial selection: Basic principle

In Section 3.4, scales were selected at a given image point from local maxima over scale of some (possibly non-linear) combination of normalized spatial derivatives. This principle can be applied to spatial selection as well, by selecting points $(x, y)^T$ and scales t that are simultaneously maxima with respect to scale *and* position. Such points are called *normalized scale-space maxima* of the differential entity considered.

The most straightforward implementation of this general principle is to use the same normalized entity for spatial selection as was used in the selection of integration scale, i.e., $\det \mu_L$ (see Section 3.4). This method has the advantage that spatial selection and scale selection are performed simultaneously. Alternatively, the spatial selection can be performed independently of the scale selection. In particular, it may be desirable to use an operator based on second order derivatives (even operators), since such an operator

[5] The scale factor is selected such that for a circular binary blob the ellipse area is equal to the area of the blob. This only affects how ellipses are displayed; in the computations of various shape cues from μ_L the scale factor always cancels out.

[6] When implementing the algorithm on a serial computer there are obviously efficiency considerations as well.

typically gives rise to spatial maxima at the centers of high contrast blobs that stand out from the surrounding.

Previous methods for blob detection have often been based on the Laplacian of the Gaussian, $\nabla^2 g$; see e.g (Marr, 1982; Blostein and Ahuja, 1989b, 1989a; Voorhees and Poggio, 1987) It is common for methods utilizing $\nabla^2 g$ or similar operators to be combined with some thresholding operation in order to suppress false alarms, and also to contain a more or less complex spatial post-processing step, in which blobs may, e.g., be split or merged according to some geometric criterion. In contrast, the scheme we propose contains neither thresholding nor spatial post-processing.

For the purpose of spatial selection, we have investigated the use of three different non-linear combinations of normalized derivatives, all of them well-defined in the sense that they do not depend on the choice of coordinate system:

- The determinant of the second moment matrix, $\det \mu_L$, i.e., the same property as was favoured for scale selection previously.
- The squared[7] Laplacian $(L_{\xi\xi}+L_{\eta\eta})^2$, i.e., the squared trace of the normalized Hessian, $\mathrm{trace}^2\, \mathcal{H}_{norm} L$.
- The determinant of the normalized Hessian matrix, $\det \mathcal{H}_{norm} L = L_{\xi\xi} L_{\eta\eta} - L_{\xi\eta}^2$.

An analysis concerning the scales at which these entities assume local maxima over scales for a periodic and a blob-like pattern respectively is given in (Lindeberg and Gårding, 1993; Lindeberg, 1994b); some results are summarized in Table 2. Note that the scales at which the maxima are assumed are related by constant factors.

Model signal	$t_{\mathrm{trace}\, \mu_L}$	$t_{\det \mu_L}$	$t_{\mathrm{trace}\, \mathcal{H}_{norm} L}$	$t_{\det \mathcal{H}_{norm} L}$
Periodic: $\sin \omega_1 x + \sin \omega_2 y$	$1/\omega_0^2$	$2/(\omega_1^2 + \omega_2^2)$	$2/\omega_0^2$	$4/(\omega_1^2 + \omega_2^2)$
Blob: $g(x;\, t_1)\, g(y;\, t_2)$	$t_0/\sqrt{1+2\gamma_1^2}$	$\sqrt{t_1 t_2}/\sqrt{1+2\gamma_1^2}$	t_0	$\sqrt{t_1 t_2}$

Table 2: Closed-form expressions for the scale levels where the local maxima over scales are attained for a periodic model signal and a blob-like model signal. For the entities based on the trace of μ_L and $\mathcal{H}_{norm} L$ respectively, only the results from the isotropic cases ($\omega_1 = \omega_2 = \omega_0$, and $t_1 = t_2 = t_0$) are shown. For the periodic signal the trace based entities have two extrema when the foreshortening is sufficiently large, while the maximum is unique for determinant based entities. For the blob signal, the maximum is unique in all four cases.

In practice, each of these entities is computed at an integration scale $s = \gamma_1^2 t$ proportional to the local scale t. In the first case, the integration is applied to μ_L before the determinant is taken, since $\det \mu_L$ is identically zero when considered pointwise. In contrast, the pointwise representations of the other two operators are not singular, so in these cases the integration step could in principle be omitted (i.e., $\gamma_1 = 0$). Nevertheless, such smoothing will be used here as a simple way of suppressing less significant responses, and hence reducing the computational load.

[7] The squaring is performed only in order to obtain uniform treatment of bright and dark blobs. The same effect could, of course, also be achieved by considering both normalized scale-space maxima and normalized scale-space minima of the ordinary Laplacian operator (although the effect of the second stage smoothing then would become somewhat different).

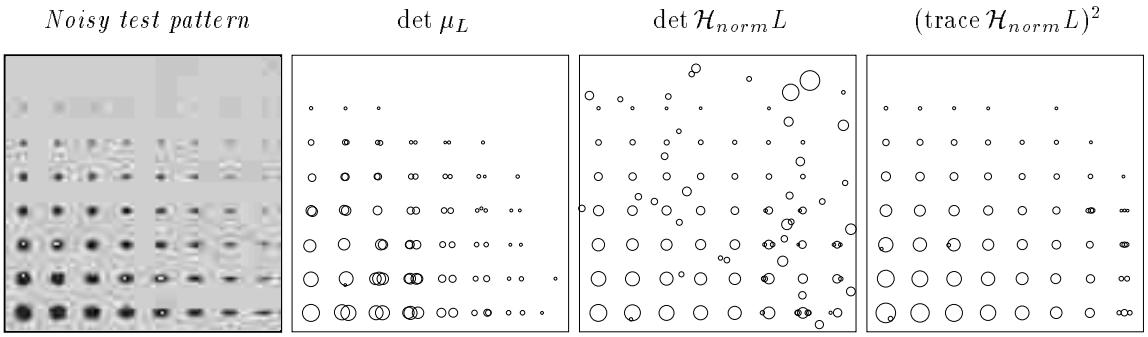

Figure 6: (a) Synthetic image with dark elliptical blobs with varying sizes and aspect ratios on a brighter background, and additive Gaussian noise with a standard deviation equal to 20% of the brightness difference between the blobs and the background. (b)–(d) Normalized scale-space maxima detected using $\gamma_1 = \sqrt{2}$. From left to right, the operator used was $\det \mu_L$, $\det \mathcal{H}_{norm} L$, and $(\mathrm{trace}\, \mathcal{H}_{norm} L)^2$. The size of each circle indicates the scale at which the maximum was assumed.

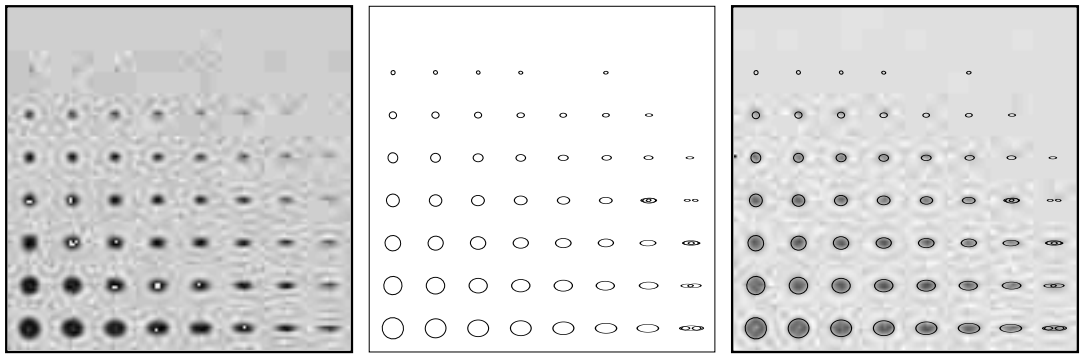

Figure 7: Multi-scale blob detection using normalized scale-space extrema of the square of the Laplacian of the Gaussian. (Left) Original image. (Middle) Detected ellipses. (Right) Ellipses representing the second moment matrix superimposed onto a bright copy of the original grey-level image.

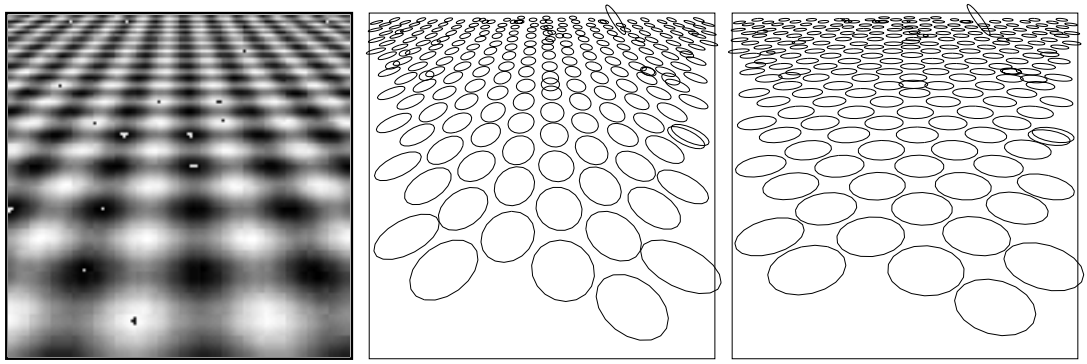

Figure 8: Multi-scale blob detection using normalized scale-space extrema of the square of the Laplacian of the Gaussian. (Left) Original image. (Middle) Detected ellipses before adaption of local scale. (Right) Detected ellipses after adaption of local scale.

4.2 Experimental results

The properties of the spatial selection process will now be illustrated using two synthetic test images. Additional experiments, using natural images, are given in Section 5.

The result of the first experiment is shown in Figure 6. The image to the left contains dark elliptical blobs with varying sizes and aspect ratios on a brighter background, and additive Gaussian noise with a standard deviation equal to 20% of the brightness difference between the blobs and the background. The blob positions detected by each of the three operators in this image are shown to the right.[8] Since no shape information is computed at this stage, the detected blobs are displayed as circles, with the area of each circle proportional to the detection scale.

The performance of all three operators is somewhat similar, but it is clear that they differ in the number of spurious maxima they generate, as well as in their tendency to generate multiple spatial maxima for elongated blobs. Clearly, $\det \mu_L$ generates the largest number of maxima, and $(\mathrm{trace}\, \mathcal{H}_{norm} L)^2$, i.e., the squared normalized Laplacian, generates the fewest. Subsequent experiments on spatial selection will therefore be based on the latter operator, but it should not be ruled out that the other two operators can be advantageous in some situations.

Figure 7 shows the final blobs found by the method, using the squared Laplacian for spatial selection and $\det \mu_L$ for computation of blob size and shape as explained previously. In the scale selection step, the integration scale parameter was coupled to the local scale parameter by $s = \gamma_1^2 t$ with $\gamma_1 = \sqrt{2}$. Then, when computing the second moment matrices, the integration scale was set to $s = \gamma_2^2 s_{\det \mu_L}$ with $\gamma_2 = 2$, where $s_{\det \mu_L}$ denotes the integration scale for which the maximum in $\det \mu_L$ was assumed. Here, only the global maxima with respect to scale have been retained.

The last example of this section demonstrates the importance of adapting the local scale in the computation of μ_L. Figure 8 shows the blobs detected in an image at the local scale that maximizes $\det \mu_L$, as well as the final blobs obtained by adapting the local scale to maximize anisotropy.

5 Shape from texture

This section shows how the proposed multi-scale texture descriptor can be used for estimating the shape or orientation of three-dimensional surfaces in the scene from perspective distortion of surface texture observed in a monocular image.

5.1 Background

The image of a slanted textured surface contains several more or less independent cues that can be used to estimate the shape and orientation of the surface. Pioneering work on this subject was done by Gibson (1950) who studied so-called texture gradients, i.e., systematic variations in the image texture due to perspective distortions. One example is the familiar "perspective effect" which makes the image of a near surface patch smaller than that of a far patch. Several algorithms for estimation of surface orientation from texture gradients have later been proposed, e.g. (Aloimonos, 1988; Blostein and Ahuja,

[8] The scale interval used was $t \in [1, 256]$, with three samples per octave distributed in uniform logarithmic steps, and the image size was 512×512. The integration scale was $s = \gamma_1^2 t$ with $\gamma_1 = \sqrt{2}$.

1989b; Kanatani and Chou, 1989; Blake and Marinos, 1990b). Witkin (1981) pointed out that the foreshortening effect, i.e., the systematic compression of a slanted pattern in the direction of slant, can also be a cue to surface orientation. For example, the image of a slanted circle is an ellipse, and the degree and orientation of the elongation of the ellipse indicates the magnitude and direction of slant. Whereas texture gradients are primarily due to perspective effects, the foreshortening effect can also be observed in orthographic projection of a planar pattern. Various extensions of Witkin's method have later been described, e.g. (Davis et al., 1983; Kanatani, 1984; Blake and Marinos, 1990a; Gårding, 1993). Related methods include (Pentland, 1986; Brown and Shvaytser, 1990).

5.2 Review of image geometry

In order to understand how a local texture description can be interpreted in terms of three-dimensional surface shape, it is necessary to take a closer look at the surface and viewing geometry.

Consider a smooth surface S viewed in perspective projection. The local perspective distortion of the projected surface pattern results from two factors; firstly, the distance and orientation of the surface with respect to the line of sight, and secondly, the angle between the line of sight and the image surface. The latter factor is often referred to as the "position effect". Since it only depends on the internal camera geometry, it can be eliminated by reprojection of the image from the focal point. Hence, for analytical clarity we represent the image by a unit viewsphere Σ, and let it be understood that in practical computations with a planar image the coordinates on Σ are obtained by a local coordinate transformation.

Fortunately, it can be shown that in order to estimate local surface orientation from texture, it suffices to consider the first-order (linear) terms of the perspective projection at each image point. To give a more precise formulation of this statement, it is necessary to introduce a few definitions. Following (Gårding, 1992) and using standard notation from differential geometry (see e.g. (O'Neill, 1966)), consider a perspective mapping of a smooth surface S onto a unit viewsphere Σ (see Figure 9). At any point p on Σ let $(\bar{p}, \bar{t}, \bar{b})$ be a local orthonormal coordinate system defined such that the $\bar{p}$ direction is parallel to the view direction, $\bar{t}$ is parallel to the direction of the gradient of the distance from the focal point, and $\bar{b} = \bar{p} \times \bar{t}$.

Denote by $F : \Sigma \to S$ the perspective backprojection from Σ to S, and by F_{*p} the derivative of this mapping at any point p on Σ. The mapping F_{*p}, which constitutes a linear approximation of F at p, maps point in the tangent plane of Σ at p, denoted $T_p(\Sigma)$, to points in the tangent plane of S at $F(p)$, denoted $T_{F(p)}(S)$. In $T_{F(p)}(S)$, let $\bar{T}$ and $\bar{B}$ be the normalized images of $\bar{t}$ and $\bar{b}$ respectively. In the bases $(\bar{t}, \bar{b})$ and $(\bar{T}, \bar{B})$ the expression for $F_{*p} : T_p(\Sigma) \to T_{F(p)}(S)$ is

$$F_{*p} = \begin{pmatrix} r/\cos\sigma & 0 \\ 0 & r \end{pmatrix} = \begin{pmatrix} 1/m & 0 \\ 0 & 1/M \end{pmatrix}, \tag{23}$$

where $r = ||F(p)||$ is the distance along the visual ray from the center of projection to the surface (measured in units of the focal length) and σ is the slant of the surface. Two *characteristic* (dimensionless) *ratios* (m, M) have been introduced here to simplify later expressions and because of their geometric significance. These entities are the inverse eigenvalues of F_{*p}, and they basically describe how a unit circle in $T_{F(p)}(S)$ is transformed

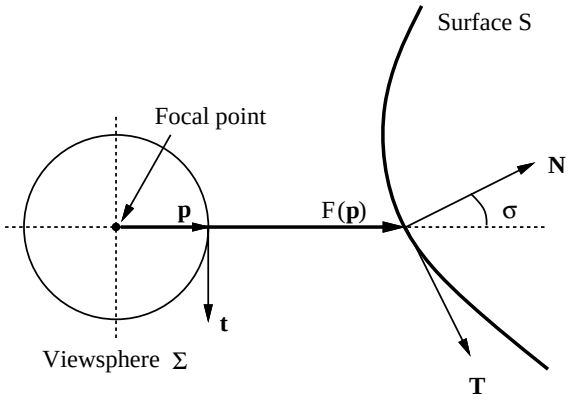

Figure 9: Local surface geometry and imaging model. The tangent planes to the viewsphere Σ at p and to the surface S at $F(p)$ are seen edge-on but are indicated by the tangent vectors $\bar{t}$ and $\bar{T}$. The tangent vectors $\bar{b}$ and $\bar{B}$ are not shown but are perpendicular to the plane of the drawing, into the drawing. (Adapted from (Gårding, 1992).)

when mapped to $T_p(\Sigma)$ by F_{*p}^{-1}; it becomes an ellipse with m as minor axis (parallel to the t direction) and M as major axis (parallel to the b direction).

From F_{*p} several useful relations between local perspective distortion and surface shape can be derived. Firstly, surface orientation is directly related to (m, M) and the corresponding eigenvectors $(\bar{t}, \bar{b})$. The *tilt* direction, defined as the direction of the gradient of the distance from Σ to the surface, is parallel to the eigenvector $\bar{t}$ corresponding to the smaller inverse eigenvalue m. *Foreshortening* is defined as the ratio m/M, and is directly related to surface slant σ by the relation $\cos\sigma = m/M$. Together, tilt $\bar{t}$ and slant σ determine the surface orientation (up to the sign of tilt; both $\bar{t}$ and $-\bar{t}$ are eigenvectors corresponding to the eigenvalue $1/m$). Secondly, "texture gradients" can be computed from the spatial rate of change of various measures derived from the eigenvalues/eigenvectors of F_{*p}. For example, the local area ratio between the image and the surface is $1/\det F_{*p} = mM$, and the normalized *area gradient* which contains information about surface shape and orientation is thus $\nabla(mM)/(mM)$. In Section 5.3 we will return to these relations, and show how they can be exploited in practice.

Normally, the brightness pattern is provided in a planar image Π, rather than in the viewsphere Σ. This is of little consequence, however, because the mapping $G : \Pi \to \Sigma$ from a point q on the planar image to the corresponding point p on the viewsphere can be pre-computed as long as the internal camera geometry is known.[9] Hence, a representation of the brightness structure on Σ can always be obtained by applying G (or its derivative G_{*q}) to the corresponding representation in the planar image Π.

A more detailed discussion of the shape cues that can be derived from the components of F_{*p} and its derivatives can be found in (Gårding, 1992).

5.3 Deriving shape cues from the second moment descriptor

In order to use a texture description derived from a monocular image to infer properties of the surface geometry, it is necessary to introduce some assumptions about the surface

[9]The mapping G is often referred to as the *gaze transformation*. In practice it is usually very close to the identity mapping.

texture. These assumptions can have many different forms. In this section two useful examples are considered; firstly, *weak isotropy*, which allows estimation of "shape from foreshortening", and secondly, *constant size*, which allows estimation of "shape from the area gradient". In both cases the general idea is to compute properties of the local surface geometry by combining estimates of various properties of the image brightness descriptor μ_L with assumptions about the corresponding properties of the surface reflectance descriptor μ_S.

We first need to establish how the relation between μ_L and μ_S depends on the local geometry. The analysis is simplified by introducing an intermediary descriptor $\mu_\Sigma(p)$, which is defined in the tangent plane $T_p(\Sigma)$ to the unit viewsphere Σ at the point $p = G(q)$. $\mu_\Sigma(p)$ describes the structure of the intensities transformed from the image to $T_p(\Sigma)$ by the linearized mapping G_{*q}, and weighted by the transformed window function $w'(p) = w(G_{*q}^{-1}p)$. By (9) we have

$$\mu_L(q) = G_{*q}^T \, \mu_\Sigma(p) \, G_{*q}, \tag{24}$$

where $G_{*q} : T_q(\Pi) \to T_p(\Sigma)$ is the derivative map between (the tangent plane to) the planar image Π at q and the tangent plane to the viewsphere at p. Hence, the practical procedure is first to estimate $\mu_L(q)$ in the image plane and then to compute $\mu_\Sigma(p)$ by inverting (24).

Analogously, $\mu_S(F(p))$ describes the structure of the intensities transformed from $T_p(\Sigma)$ by the linearized mapping F_{*p}, and weighted by the window function transformed accordingly. Assuming that the image brightness is directly proportional to the surface reflectance, it holds that $\mu_S(F(p))$ describes the structure of the linearized and windowed surface reflectance at the point $F(p)$.

In practice, the second moment matrices $\mu_{\Sigma(p)}$ and $\mu_S(F(p))$ cannot be directly expressed in terms of the $(\bar{t}, \bar{b})$ and $(\bar{T}, \bar{B})$ bases respectively, since the orientations of these bases are not known a priori. Introduce rotation angles θ and φ describing these orientations relative to some reference systems. (For $\mu_\Sigma(p)$ we define this reference as the gaze-transformed image coordinate frame, whereas the precise definition concerning $\mu_S(F(p))$ is left open.) Then, (9) gives that the second moment matrices in the reference systems are related by

$$R_\theta^T \; \mu_\Sigma(p) \; R_\theta = F_{*p}^T \; R_\varphi^T \; \mu_S(F(p)) \; R_\varphi \; F_{*p}, \tag{25}$$

where

$$R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

gives the tilt direction relative to the gaze-transformed image coordinate frame and R_φ represents a corresponding rotation relative to some coordinate system in the tangent plane to the surface.

To simplify the notation, the arguments to μ_L, μ_Σ and μ_S will be dropped in the remainder of this section.

5.3.1 Shape from foreshortening

If μ_S is known and if μ_Σ can be computed from the image data, then (25) provides three equations for the four unknowns (m, M, θ, φ). From this viewpoint, the problem is, in general, underdetermined.

To compute surface orientation, however, it is only necessary to know the angle θ (which gives the tilt direction) and the ratio m/M (which gives the slant angle). Moreover, if μ_S

is *weakly isotropic*, i.e., if

$$\mu_S = cI \tag{26}$$

for some (unknown) constant $c > 0$, then the orientation ambiguity concerning φ disappears, since the representation of μ_S is in this case invariant under rotations. Such a distribution (for which $\tilde{Q}_S = 0$) has the property that there is no single preferred direction in the surface texture, i.e., that the surface texture is not systematically elongated. Under this condition and assuming that F_{*p} is non-degenerate, (25) can be rewritten as

$$\mu_\Sigma = c\, R_\theta\, F_{*p}^T\, R_\varphi^T\, R_\varphi\, F_{*p}\, R_\theta^T = c\, R_\theta\, F_{*p}^T\, F_{*p}\, R_\theta^T. \tag{27}$$

It follows that the eigenvectors of μ_Σ are $(\bar{t}, \bar{b})$ expressed in the gaze-transformed image coordinate frame, and that the eigenvalues of F_{*p} are proportional to the square roots of the eigenvalues (λ_1, λ_2) of μ_Σ;

$$m \sim 1/\sqrt{\lambda_1} \sim 1/\sqrt{1+\tilde{Q}}, \qquad M \sim 1/\sqrt{\lambda_2} \sim 1/\sqrt{1-\tilde{Q}}. \tag{28}$$

From the analysis in Sec. 5.2 we then obtain that the tilt direction, $\bar{t}$, is (plus/minus) the eigenvector, $\bar{e}_1$, corresponding to the maximum eigenvalue, λ_1, and the slant is given by

$$\cos\sigma = \frac{m}{M} = \sqrt{\frac{1-\tilde{Q}}{1+\tilde{Q}}}. \tag{29}$$

Hence, if the assumption of weak isotropy can be justified, an easily computed estimate of local surface orientation is directly available. Unfortunately, many natural textures violate this assumption, and it is therefore often necessary to exploit alternative assumptions.

5.3.2 Shape from the area gradient

Assuming that the local "size" of the surface texture does not vary systematically, it is obvious that the gradient of size of the projected texture is an important cue to surface shape and orientation.

Consider a point p in $T_p(\Sigma)$, and let F_{*p} be the local linear part of the perspective backprojection. The area ratio is then equal to $\det F_{*p}^{-1} = mM$, i.e.,

$$A_\Sigma = mMA_S, \tag{30}$$

where A_Σ is the area of a small surface patch on the viewsphere Σ, and A_S is the area of the corresponding patch in the surface S. Hence, assuming that $A_S = c$ where c is some unknown constant, we can define the *normalized area gradient*

$$\frac{\nabla A_\Sigma}{A_\Sigma} = \frac{\nabla(mM)}{mM}. \tag{31}$$

Note that the unknown scale constant c has been eliminated. This means that no assumptions about the absolute scale of the surface texture are necessary. Moreover, no assumptions are made about the elongation of the surface texture (given by the ratio of the eigenvalues of μ_S).

The area A_Σ can be computed from the corresponding area A_L in the planar image Π using $A_\Sigma = (\det G_*)A_L$, where G_* is the gaze transformation discussed in Section 5.2.

In our current implementation A_L is estimated from the scale at which $\det \mu_L$ assumes its maximum, as described in Section 3.5. A more detailed description of how to estimate the normalized area gradient from A_L is given in Appendix A.2.

It has not yet been mentioned how the normalized area gradient should be interpreted in terms of surface shape and geometry. It turns out that its information content is considerably more complex than that of foreshortening; in (Gårding, 1992) it is shown that

$$\frac{\nabla(mM)}{mM} = -\tan\sigma \begin{pmatrix} 3 + r\kappa_t/\cos\sigma \\ r\tau \end{pmatrix}, \tag{32}$$

with respect to the $(\bar{t}, \bar{b})$ basis. Here, r is the distance from the viewer, σ is the slant of the surface, κ_t is the normal curvature of the surface in the tilt direction, and τ is the geodesic torsion, or "twist", of the surface in the tilt direction.

Hence, the normalized area gradient can either be used to recover information about the surface curvature (scaled by distance) if the surface orientation is known, or to recover the surface orientation if the curvature is known or (assumed to be) small. In the latter case there is no ambiguity in the sign of the tilt direction, unlike the case of foreshortening.

5.4 Estimating surface shape and orientation: Basic scheme

Our method for computing monocular shape-from-texture cues from image data can be summarized as follows:

1. Compute local texture descriptors μ_L as described in Section 3.4. This can either be done at selected spatial positions corresponding to normalized scale-space extrema as described in Section 4, or at a (uniform) grid of points generated by some default principle.

2. Determine a set of points where estimates of surface orientation are to be computed. This set of points can be the same as that used for computing the texture descriptors, or it can be a smaller set of points, e.g. a uniform grid. Associate with each point a (Gaussian) window that specifies the weighting of the texture descriptors in the neighborhood of the point. The scale of this window function will be referred to as the *texel grouping scale*.

3. Estimate surface orientation:

   (a) Apply the assumption of *weak isotropy* as described in Section 5.3 to compute foreshortening. This leads to a direct estimate of surface orientation up to the sign of tilt.

   (b) Apply the assumption of *constant area* as described in Section 5.3 to compute the normalized area gradient. This permits a unique estimate of surface orientation under the additional assumption that the local curvature of the surface can be neglected in (32).

   (c) Optionally, apply other assumptions about the surface texture, e.g. compute the foreshortening gradient, and use these assumptions to estimate surface shape and/or orientation.

5.5 The texel grouping scale

In order to compute an estimate of surface orientation at a specified point, the local texture descriptors in the neighbourhood of the point must somehow be combined. As was described in previous sections, the second moment descriptor computed by spatial and scale selection can be informally thought of as a single "texture element". In the case of a perfectly regular surface texture, the shape of this texture element can be relied upon to provide information about local perspective distortion. Most natural textures, however, exhibit a considerable degree of randomness in their structure, and it is therefore necessary to consider more than one texture element in order to detect the systematic geometric distortions due to the perspective effects. Attempts have been made at modeling such randomness statistically (Witkin, 1981; Kanatani and Chou, 1989; Blake and Marinos, 1990a, 1990b), but here such specific models are replaced by the basic principle of reducing variance by integration. For this reason, the concept of *texel grouping scale* has been introduced in the scheme above; it refers to the scale used for combining texture descriptors computed at different spatial points into entities to be used for computing geometric shape descriptors.

If the texture descriptors are combined by weighted averaging into a descriptor of the same type, as in the case of methods based on foreshortening, then the texel grouping scale is closely related (or even equivalent) to the relative integration scale. More precisely, from the semi-group property of Gaussian smoothing, $g(\cdot;\, s_2) = g(\cdot;\, s_2 - s_1) * g(\cdot;\, s_1)$, it follows that, if the local smoothing scale t is held constant, then the second moment matrix at any coarse integration scale, s_2, can be computed from the second moment matrices at any finer integration scale, s_1,

$$\mu_L(\cdot;\, t, s_2) = g(\cdot;\, s_2 - s_1) * \mu_L(\cdot;\, t, s_1). \tag{33}$$

Hence, if the local scale parameter in the scale-space representation is constant (e.g. equal to the scale level in the input image), then in the basic version of the method of estimating surface orientation from foreshortening and weak isotropy, the texel grouping scale is equivalent to the relative integration scale.

However, the cascade smoothing property (33) is not applicable when the texture descriptors are combined into a descriptor of a different type. For example, estimation of shape from texture gradients is based on the average rate of change of some property of the local texture descriptors, so in this case it is clearly not meaningful to compute an average texture descriptor for the whole region. Rather, the appropriate texture property (e.g. area) is estimated from each windowed second moment descriptor separately, and the corresponding texture gradient is then estimated using Gaussian weights given by the texel grouping scale. (The procedure for the case of the area gradient is described in Appendix A.2.)

So far no method for automatic selection of the texel grouping scale has been implemented. In the experiments presented below, the estimates are computed on sparse regular grids, and the size of the Gaussian grouping window is proportional to the grid cells. An alternative approach is, of course, to let the texel grouping scale be proportional to the selected integration scale.

5.6 Experimental results

The examples shown in this section have been computed using the same parameters as in the previous sections, i.e., using $\gamma_1 = \sqrt{2}, \gamma_2 = 2$. The surface orientation will be represented

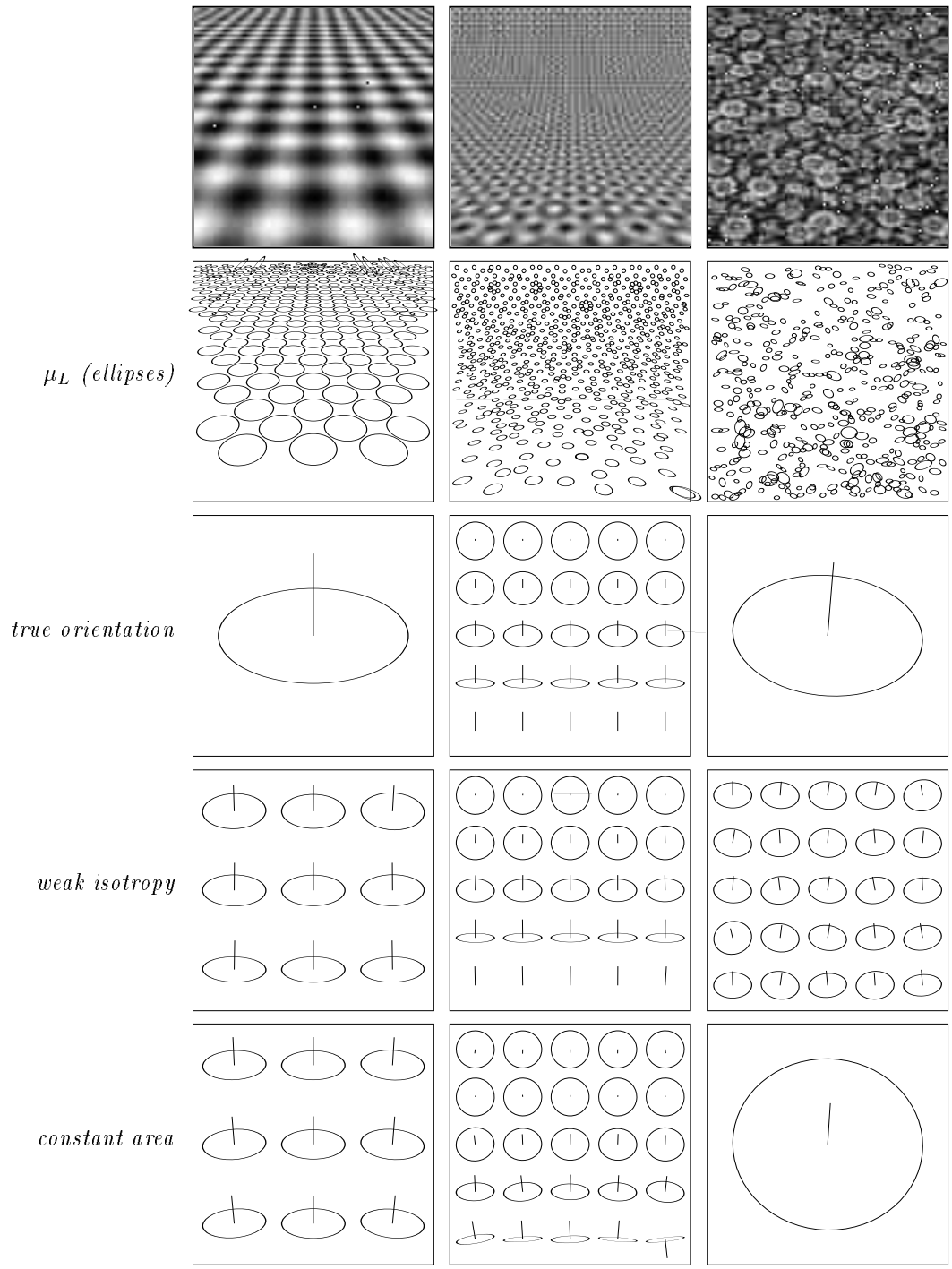

Figure 10: Estimating local surface orientation in a synthetic image of a planar surface with 5% noise (left), a synthetic image of a cylindrical surface with 25% noise (middle), and a real image of a planar surface with known orientation (right). The rows show from top to bottom; (a) the grey-level image, (b) elliptical blobs detected by the adaptive multi-scale method, (c) reference surface orientation, (d) surface orientation estimated from foreshortening, (e) surface orientation estimated from the area gradient.

numerically by (σ, θ), where σ is the slant, i.e., the angle between the surface normal and the optical axis, and θ is the angle between the tilt direction $\bar{t}$ and the horizontal axis of the image coordinate frame.

Figure 10 shows results[10] from two noisy synthetic images and one real image, all with known camera geometry and surface orientation. From top to bottom, the rows show the grey-level image, the detected blobs, the true surface orientation, the surface orientation estimated from foreshortening (only the first of the two estimates is shown), and the surface orientation estimated from the area gradient.

The synthetic image in the left column (also shown in Figure 7) shows a planar surface pattern consisting of the sum of two sine waves and 5% additive Gaussian noise. The true orientation of the surface is $(\sigma = 60°, \theta = 90°)$, and at the center the estimates from foreshortening and the area gradient are $(\hat{\sigma} = 61.1°, \hat{\theta} = 90.0)$ and $(\hat{\sigma} = 62.3°, \hat{\theta} = 90.0)$, respectively.

The middle column shows the same cylindrical surface image that was used in the first row in Figure 4. Here, 25% white Gaussian noise has been added; a noise level high enough to ensure that direct computations on unsmoothed data are bound to fail (compare with Table 1). It is quite obvious that the adaptive multi-scale blob detection technique is able to handle this noise level without much difficulty. At the center the true orientation is $(\sigma = 55°, \theta = 90°)$, the estimate from foreshortening is $(\hat{\sigma} = 54.1°, \hat{\theta} = 90.3°)$, and the estimate from the area gradient is $(\hat{\sigma} = 33.5°, \hat{\theta} = 89.1°)$. The fact that the slant of this surface is underestimated by the area gradient is entirely in keeping with the theory; the estimate is based on the assumption that $\kappa_t = 0$ in (32), but here $\kappa_t < 0$ since the surface is concave rather than flat. In fact, by using (32) the scaled curvature $r\kappa_t$ can be estimated from the difference between the slant estimates from weak isotropy and the area gradient. At the central point the estimate obtained this way is $\widehat{r\kappa_t} = -0.92$, which should be compared to the true value $r\kappa_t = -0.87$.

The right column of Figure 10 shows the results obtained with a real image of a planar surface with known surface orientation. The true surface orientation is $(\sigma = 50.8°, \theta = 85.3°)$, and at the center the estimate from foreshortening is $(\hat{\sigma} = 48.7°, \hat{\theta} = 82.7°)$. Due to the narrow field of view the measurable effects of the area gradient in this image are very small. This fact is reflected by a fairly inaccurate estimate $(\hat{\sigma} = 25.7°, \hat{\theta} = 86.2°)$ obtained from the area gradient in the whole image. On a 3×3 grid the estimates break down completely.

Figure 11 shows the results obtained with five images from (Blostein and Ahuja, 1989b). The camera geometry is unknown, and it is therefore impossible to compute absolute estimates of the surface orientation. To estimate surface orientation from foreshortening, the second moment matrix $\mu_L(p)$ must first be transformed to $T_p(\Sigma)$, but the parameters of this transformation depend on the camera geometry and are hence unknown. Foreshortening is therefore visualized directly by ellipses representing the weighted second moment matrices in the image on which the estimate would be based. To estimate surface orientation from the area gradient, the focal length must be known. However, the position of the *horizon* of the plane, i.e., the line where projected area is estimated to vanish, can still be determined.

[10] In the examples in this section the surface orientation is indicated graphically by a dish with an attached needle parallel to the surface normal. In contrast to the previous illustrations of the second moment matrices, the dishes are from now on viewed in *parallel* projection along the visual ray through the image center. With this convention, the shape of each projected dish specifies the surface orientation regardless of the internal camera geometry and the position of the dish in the image.

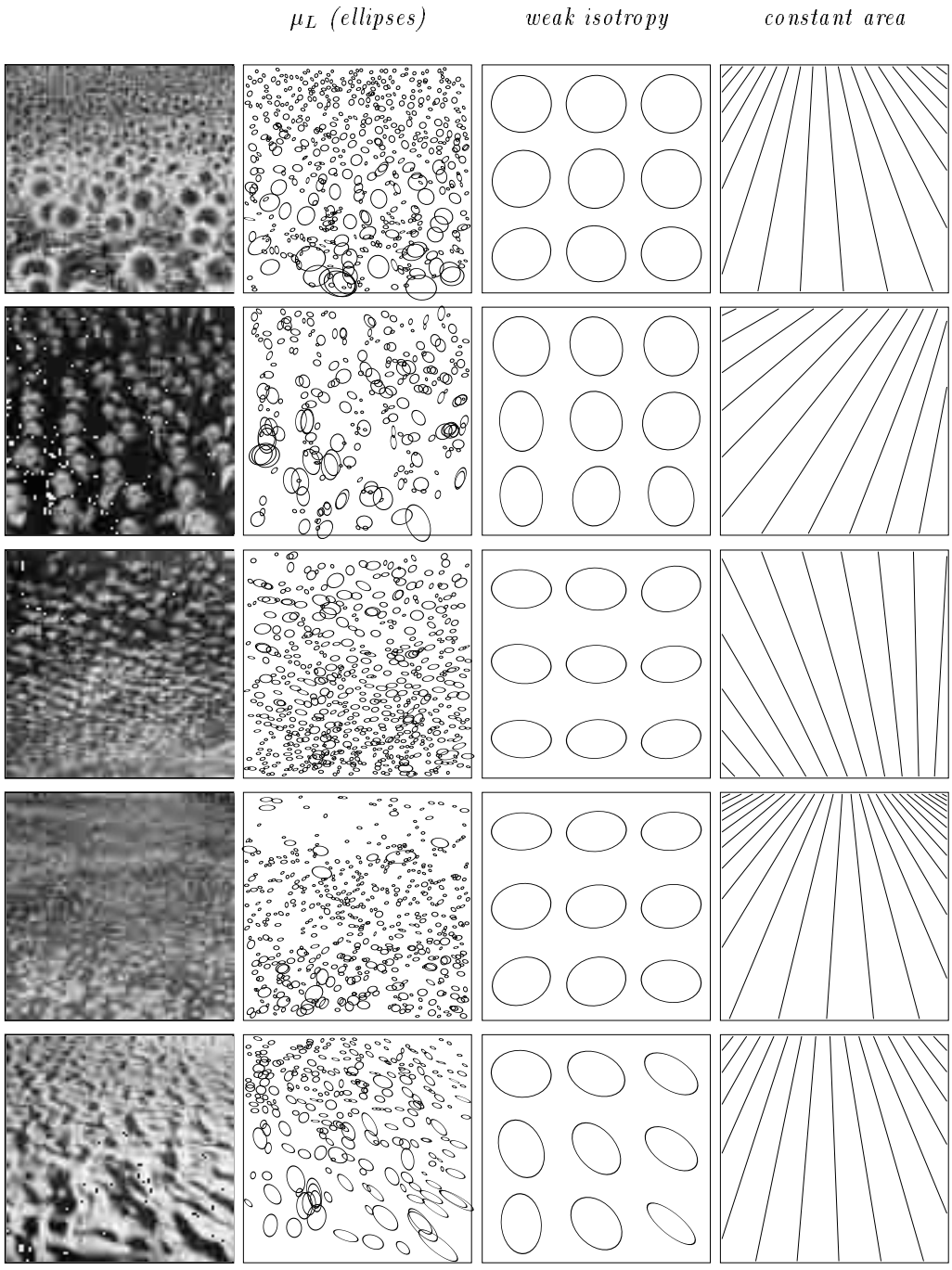

Figure 11: Estimation of foreshortening and the area gradient in real images from (Blostein and Ahuja, 1989b). (a) Real grey-level image. (b) Elliptical blobs detected by the adaptive multi-scale method. (c) Estimated foreshortening, here represented by weighted averages of the second moment descriptors associated with each blob. (d) Estimated area gradient, visualized by lines aligned with the tilt direction converging to a point on the horizon.

The estimated horizon typically lies outside the image, but in the rightmost column of Figure 11 it is indirectly represented by a set of projected lines parallel to the tilt direction in the surface.

It is interesting to note that the foreshortening in these examples often reflects the orientation of the individual texture elements (e.g., the sunflowers), whereas the area gradient corresponds to the orientation of the underlying surface.

6 Shape from disparity gradients

In this section we shall apply a similar methodology, based on multi-scale second moment descriptors, to shape estimation from binocular (stereo) vision. A more detailed account of the geometric aspects of this problem can be found in (Gårding and Lindeberg, 1994).

Traditionally, binocular stereopsis has often been associated with recovery of three-dimensional *depth*. Here, however, we shall be concerned with estimation of *surface orientation*, i.e., the rate of change of depth. Many computational models of stereopsis are based on sparse but salient features such as edges or corners (see e.g. (Pollard et al., 1985)). This approach is often quite successful, but has the drawback that it only produces sparse depth estimates. If higher-order properties are needed, such as local surface orientation or curvature, they could in principle be estimated by first applying an additional stage that interpolates the surface between the data points to obtain a dense depth map and then differentiating this representation.

An alternative approach, which we shall pursue here, is to derive higher-order surface properties directly from the properties of corresponding image patches, without using depth as an intermediate representation. This can be achieved either by first computing a dense disparity map and then estimating derivatives of the disparity field, or by directly using differences in local image properties, e.g. the local statistics of the orientation or curvature of contours.

In both cases, the estimation of surface orientation can be formulated in terms of modelling the local transformation from the right eye's view of a small surface patch to the left eye's view of the same patch by an *affine* transformation, rather than a simple displacement. Analogously, surface curvature can be estimated from the second-order properties of the local left-to-right transformation. The local affine transformation gives rise to *orientation disparity* as well as *spatial frequency disparity*, and several computational models based more or less directly on these cues have been described in the literature (Blakemore, 1970; Koenderink and van Doorn, 1976; Tyler and Sutter, 1979; Rogers and Cagenello, 1989; Wildes, 1981; Jones and Malik, 1992). The methodology presented here builds on, and extends, several of these models.

6.1 Viewing geometry and binocular disparity

A representation of the binocular viewing geometry is shown in Figure 12. We represent visual space with respect to a virtual cyclopean eye, constructed such that the cyclopean visual axis (the Z axis) bisects the left and right visual axes. The X and Z axes as well as the centers of the eyes lie in a common plane, called the fixation plane.

We define left and right coordinate systems (X_L, Y_L, Z_L) and (X_R, Y_R, Z_R) such that the origin of each system is at the center of projection, the Z_L, Z_R and Z axes intersect at the fixation point p with cyclopean coordinates $(0, 0, R)$, and the X_L, X_R and X axes

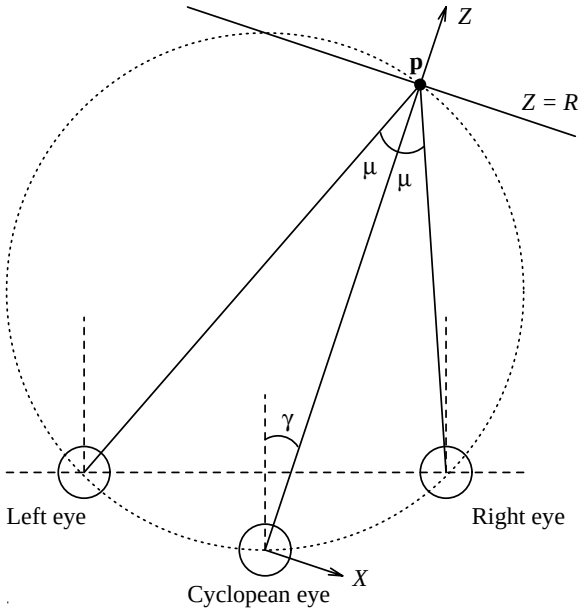

Figure 12: Representation of the binocular viewing geometry. The plane of the drawing is the fixation plane. The primary direction (indicated by dashed lines) is defined as the direction in the fixation plane that is perpendicular to the interocular baseline. The dotted circle through the fixation point and the eyes indicates a part of the point horopter, i.e., the locus of points that yield zero horizontal and vertical disparity.

are contained in the fixation plane Normalized cyclopean image coordinates are defined by $x = X/Z$, $y = Y/Z$; left and right image coordinates are defined analogously. These coordinates are related to the pixel coordinates through the intrinsic camera parameters, which are assumed to be known.

This representation of the viewing geometry does not require p to be the actual fixation point of the viewing system, nor indeed that the eyes fixate any point at all, since a rotation of either eye around the optical center does not affect the information content of the image. Conceptually, we represent the eyes or cameras by the unit viewspheres Σ, Σ_L and Σ_R independently of the physical shape (e.g., spherical or planar) of the physical imaging surface. Left and right image coordinates are then defined in the tangent planes to the viewspheres at the images of any given point p in space. These image coordinates are related to the image coordinates defined with respect to some other fixation point q by a projective transformation which is independent of the structure of the scene. However, to simplify the presentation we shall continue to refer to p as the fixation point.

6.1.1 Vergence and version

Let φ_L and φ_R be the angles between the primary (straight-ahead) direction and the left and right visual axes respectively. The *vergence* angle μ and the *version* (or *gaze*) angle γ are then defined by

$$\mu = \frac{1}{2}(\varphi_L - \varphi_R), \qquad \gamma = \frac{1}{2}(\varphi_L + \varphi_R). \tag{34}$$

As a consequence of this definition, the angle between the cyclopean visual axis and the primary direction is equal to γ (see Figure 12).

6.1.2 Binocular disparity

The retinal disparity of a point in the scene is defined as the difference in retinal position of the left and right projections of the point. Consequently, the retinal disparity of the fixation point is zero by definition. We define horizontal and vertical retinal disparity (h, v) by

$$h = x_R - x_L, \quad v = y_R - y_L, \tag{35}$$

where (x_L, y_L) and (x_R, y_R) are the normalized left and right image coordinates corresponding to the same point in the scene.

If the fixation point p lies on a smooth surface $Z(X, Y)$, a differentiable mapping M is induced from points in the left image to points in the right image in some neighbourhood of the images of p. A Taylor expansion to first order in (x_R, y_R) can then be expressed as

$$\begin{pmatrix} x_R \\ y_R \end{pmatrix} = \begin{pmatrix} 1 + h_x & h_y \\ v_x & 1 + v_y \end{pmatrix} \begin{pmatrix} x_L \\ y_L \end{pmatrix}. \tag{36}$$

In the following we shall denote the matrix in (36) by M_* and refer to it as the *derivative map*. The components $(h_x, h_y; v_x, v_y)$ constitute the *disparity gradient*.

6.2 The disparity gradient

The disparity gradient depends on the viewing geometry and the local surface orientation. Let M_* be the derivative map from the left image to the right image. The disparity gradient is $M_* - I$, where I is the unit matrix, and at the fixation point it holds that

$$M_* = \begin{pmatrix} 1 + h_x & h_y \\ v_x & 1 + v_y \end{pmatrix} = \frac{\cos(\gamma - \mu)}{\cos(\gamma + \mu)} \begin{pmatrix} \dfrac{\cos\mu + Z_X \sin\mu}{\cos\mu - Z_X \sin\mu} & \dfrac{2 Z_Y \cos\mu \sin\mu}{\cos\mu - Z_X \sin\mu} \\ 0 & 1 \end{pmatrix}, \tag{37}$$

where $(Z_X, Z_Y) = (\frac{\partial Z}{\partial X}, \frac{\partial Z}{\partial Y})$ is a gradient based parametrization of surface orientation relative to the cyclopean coordinate system. These parameters are related to the slant-tilt representation used in the previous section by

$$Z_X = \tan\sigma \cos\theta, \quad Z_Y = \tan\sigma \sin\theta. \tag{38}$$

A derivation of (37) can be found in (Gårding and Lindeberg, 1994). The size of the region where M_* provides a reasonably accurate approximation of the disparity field depends on the shape of the surface; for planar surfaces it is in fact valid over quite large visual angles.

6.2.1 The information content of the disparity gradient

What do the non-vanishing components (h_x, h_y, v_y) of the disparity gradient at the fixation point tell us about the local scene structure and the viewing geometry? First, note that the disparity gradient (37) depends on four parameters; two for the viewing geometry (μ, γ) and two for the surface orientation (Z_X, Z_Y). It is thus impossible to recover both the viewing geometry and the local surface orientation from a single measurement of the disparity gradient. If the viewing geometry is known, however, then surface orientation can be estimated and vice versa.

Denote the components of M_* by m_{ij}. Then, a few algebraic manipulations on (37) give

$$Z_X = \frac{(m_{11} - m_{22}) \cos \mu}{(m_{11} + m_{22}) \sin \mu}, \qquad Z_Y = \frac{m_{12}}{(m_{11} + m_{22}) \sin \mu}. \tag{39}$$

These expressions are homogeneous in the components of M_*. Therefore, to estimate the surface orientation it suffices to estimate M_* up to an arbitrary scale factor. In particular, there is no need to know the angle γ of asymmetric gaze, since this parameter only affects M_* by a uniform scaling factor (see (37)).

6.2.2 Estimating the disparity gradient

Using the transformation property (9) of the second moment matrix (with $B = M_*$), it is reasonably straightforward to derive explicit expressions for the disparity gradient in terms of second moment matrices in the left and right images respectively. The system of quadratic equations

$$\begin{pmatrix} \mu_{L_{11}} & \mu_{L_{12}} \\ \mu_{L_{12}} & \mu_{L_{22}} \end{pmatrix} = \begin{pmatrix} m_{11}^2 \mu_{R_{11}} & m_{11}(m_{12}\mu_{R_{11}} + \mu_{R_{12}}) \\ m_{11}(m_{12}\mu_{R_{11}} + \mu_{R_{12}}) & m_{12}^2 \mu_{R_{11}} + 2 m_{12} \mu_{R_{12}} + \mu_{R_{22}} \end{pmatrix} \tag{40}$$

gives rise to two real solutions[11]

$$\begin{aligned} m_{11} &= \alpha\,(1 + \tilde{C}_L)\,\tilde{F}_R, \\ m_{12} &= \alpha\,(\tilde{S}_L\,\tilde{F}_R \mp \tilde{S}_R\,\tilde{F}_L), \\ m_{22} &= \pm\alpha\,(1 + \tilde{C}_R)\,\tilde{F}_L, \end{aligned} \tag{41}$$

where

$$\tilde{F}_L = \sqrt{1 - \tilde{C}_L^2 - \tilde{S}_L^2}, \tag{42}$$

$$\tilde{F}_R = \sqrt{1 - \tilde{C}_R^2 - \tilde{S}_R^2}, \tag{43}$$

$$\alpha = \frac{1}{\tilde{F}_R} \frac{1}{\sqrt{1 + \tilde{C}_L}\sqrt{1 + \tilde{C}_R}} \sqrt{\frac{P_L}{P_R}}, \tag{44}$$

and the '$\pm$' and '$\mp$' signs are coupled. Notably, α occurs as a common factor in all m_{ij} and cancels in (39). Hence, only the *directional* structure of μ_L and μ_R (i.e. $\tilde{C}_L$, $\tilde{S}_L$, $\tilde{C}_R$ and $\tilde{S}_R$) influences the surface orientation estimates, while any difference in magnitude (represented by P_L and P_R) is ignored.

By adding the natural requirement that $\det M_* > 0$, i.e., that the left-to-right ordering is the same in both images, a unique solution is obtained (with $m_{22} > 0$). This constraint is closely related to the *disparity gradient limit* used e.g. by Pollard et al. (1985).

6.3 Experimental results

6.3.1 Procedure

In the experiments described below, estimation of surface orientation from disparity gradients was performed as follows. The windowed second moment descriptors were computed as

[11] The indeterminacy with respect to rotations referred to at the end of Section 2.2 disappears in this case, since $B = M_*$ has only three degrees of freedom ($b_{21} = m_{21} = 0$).

described in previous sections (using $\gamma_1 = \sqrt{2}, \gamma_2 = 2$ as usual) for the left and right images separately. To reduce clutter, a subset of the descriptors for the left image was extracted by applying a threshold to the magnitude of the scale-space maximum of the Laplacian (which was computed in the spatial selection stage). A very simple matching algorithm, based on the epipolar constraint and similarity of detection scale, was then applied to find a corresponding descriptor from the right image for each remaining descriptor from the left image. For each of the resulting left-right pairs of windowed second moment descriptors, the normalized horizontal disparity gradient was estimated using (41), and surface orientation was computed using (39).

It is worth pointing out that the positional disparities which are obtained as a result of the matching process can be used to obtain a depth estimate for each descriptor pair. Hence, the left and right sets of second moment descriptors in fact contain *two* binocular cues, in addition to the texture cues which were discussed in the previous section. However, in the examples below the positional disparity is not used.

6.3.2 Results

Three examples of results obtained with the method are shown in Figure 13.

The first two image pairs were created by perspective texture mapping onto a planar surface with orientation $(\sigma = 60°, \theta = 50°)$. The visual angle across the diagonal of each image is 32°. The first texture is a sinusoidal pattern with 5% additive Gaussian noise. In order to reduce cluttering of the graphical display, a subset of the matched descriptor pairs was selected manually. The second texture is a natural gray-level image of a pebble pattern. The gaze angle is $\gamma = -5°$ for both image pairs, and the vergence half-angle is $\mu = 10°$ for the first pair and $\mu = 6.9°$ for the second pair.

The third image pair was acquired with two CCD cameras and depicts a nursery wallpaper. The camera geometry was $(\mu = 5.6°, \gamma = -4.0°)$.

At the fixation point of the first image pair, the estimated surface orientation was $(\hat{\sigma} = 58.5°, \hat{\theta} = 52.2°)$. The error in the estimate, expressed as the angle between the estimated and true surface normals, is 2.4°. Similar results were obtained at the remaining sixteen points; the maximum error is 3.3°.

The results obtained with the second and third image pairs were slightly more variable but still fully acceptable, as can be seen from the graphical representation. The two or three large errors in each image pair are due to incorrect matches, which could probably be eliminated by a more sophisticated matching algorithm.

7 Summary and discussion

We have shown that a representation of local image structure computed by multi-scale bottom-up retinotopic processing can be directly used for deriving non-trivial cues to the local structure of three-dimensional surfaces in the scene, without iterations, search, or high-level knowledge.

In the first part of the paper, we treated the problem of computing such a representation, and introduced the windowed second moment matrix to represent the local statistics of first order Gaussian normalized derivatives of image brightness. We showed that linear transformations of the spatial coordinates affect this descriptor in a simple way, which

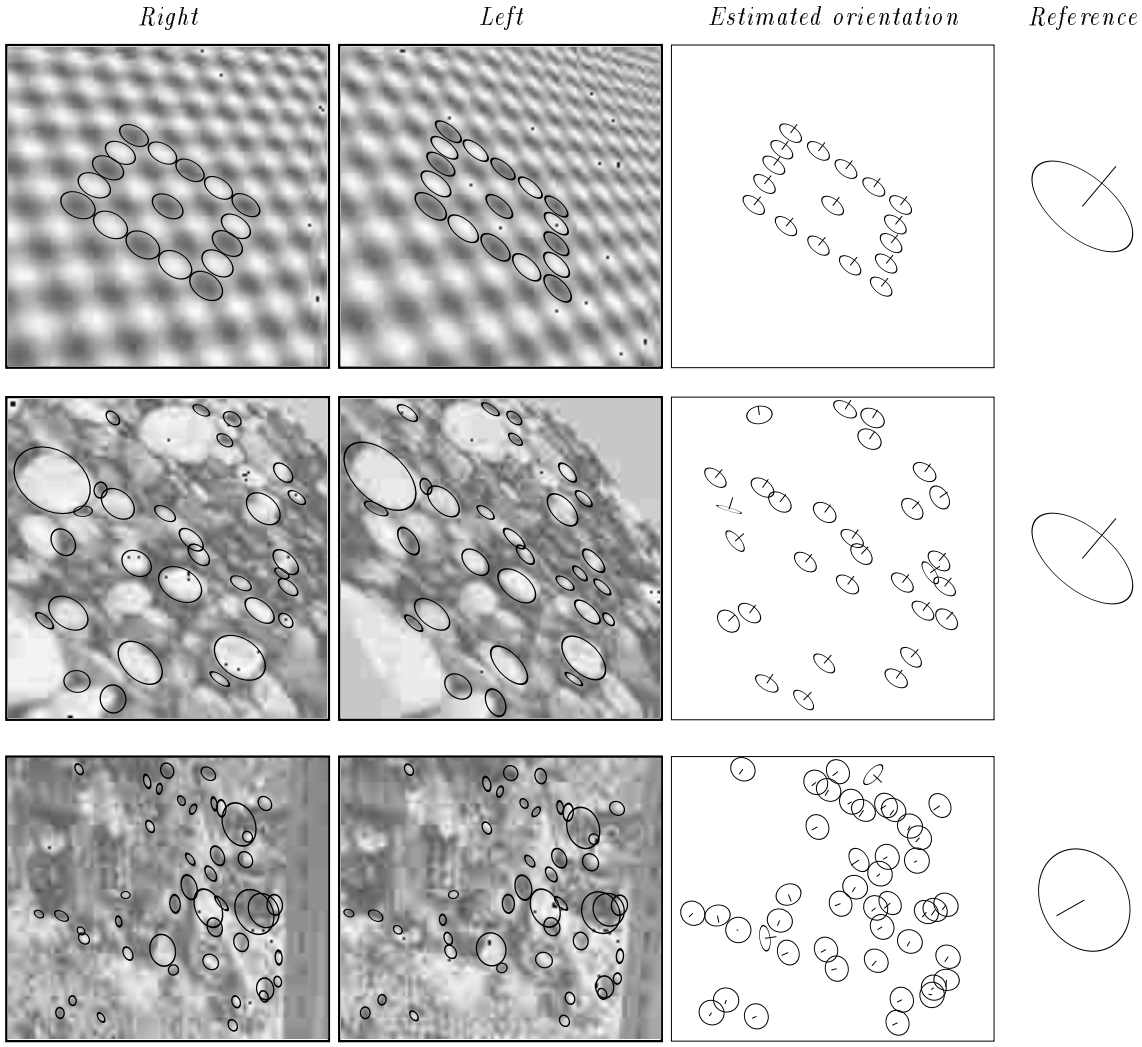

Figure 13: Local surface orientation estimated from the gradient of horizontal disparity in three stereo pairs. The images in the two first rows are generated by texture mapping. The images in the bottom row are taken with a pair of CCD cameras. The columns show from left to right; (a-b) Bright copies of the right and left images with computed texture descriptors superimposed. (c) estimated surface orientation, (d) reference surface orientation.

allows the parameters of the transformation to be estimated from the properties of the descriptor.

The computation of this descriptor involves two scale parameters; first, the smoothing scale at which derivatives of the image brightness are computed, and second, the scale of the window used to integrate statistics of nonlinear descriptors of the differential image structure. We proposed a systematic two-stage method for adaptively choosing these scale parameters. The characteristic dimensions of salient image structures at any given point are first estimated by detecting local maxima with respect to scale of certain differential invariants derived from the windowed second moment matrix. The integration scale is then set proportional to the estimated characteristic dimensions, while the smoothing scale is adapted to obtain a trade-off between suppression of noise and irrelevant fine-scale structures on the one hand, and distortion of the shape of local image structures due to smoothing on the other.

The principle used to determine characteristic dimension was also applied to guide the selection of where in the image to compute the texture descriptors. Whereas the second moment descriptor which describes local image "shape" is based on first derivatives, the entities used for spatial selection were based on second derivatives in order to favour centers of blob-like structures.

In the second part of the paper we treated the problem of using the multi-scale second moment descriptor to derive cues to local three-dimensional surface shape and orientation. We first discussed estimation of shape from texture in a monocular image, based on two independent cues referred to as foreshortening and the area gradient, respectively. It was shown that these two cues can be reliably computed both in noisy synthetic images and natural images.

We then showed that the same methodology can be used to recover local surface orientation by estimating the gradient of horizontal disparity in a binocular image pair. This method has the advantage that it does not depend on any specific assumptions about the surface texture. Experimental results were shown for both synthetic and natural images.

7.1 Relations to biological vision

As mentioned in the introduction, we have not attempted to model biological vision. However, the general principles on which the methodology is based appear to be compatible with current understanding of the structure of the first stages of the primate visual pathway.

For example, it is worth noting that the computation of the windowed second moment descriptor follows the pattern "linear filtering – nonlinearity – spatial averaging". Processing sequences of this type have in recent years been proposed as models for human texture discrimination, e.g. (Caelli, 1985; Bergen and Adelson, 1988; Malik and Perona, 1990). The initial linear filtering stage in our model is based on directional Gaussian derivatives, which have been used to model the receptive fields of simple cells in the mammalian visual cortex (Young, 1985). Moreover, selection of scale levels and spatial positions by detection of local maxima could easily be implemented by lateral inhibition between cells.

The spatial detection process we discussed was based on rotationally symmetric operators such as the Laplacian, which limits the ability to detect very elongated blob-like structures based on the response of a single operator. However, in this context it is interesting to note that in a psychophysical study of the visibility of elliptical Gaussian blobs, Bijl and Koenderink (1993) found that their results can be predicted by a model based on

Pythagorean summation of the responses of rotationally symmetric receptive fields.

7.2 Further research

Some issues not directly addressed by the present work are discussed below.

Grouping We have tacitly assumed that integration of local properties is always a meaningful operation, but in general situations it may be necessary to restrict the integration to some coherent subset of the descriptors in the window. This can have any of a number of reasons, e.g. that the image contains more than one surface, that a surface contains more than one type of texture, or that an image region contains textures resulting from more than one physical process.

Furthermore, we have in most cases used only the most dominant scale at each spatial position; a more general approach would be to detect all local maxima, and then apply spatial grouping based on similarity of characteristic dimension. For example, a noisy image of a slanted pattern might give rise to maxima at small scales due to the noise, in addition to the maxima at coarser scales corresponding to the surface texture. Separate estimation of the area gradient for the fine-scale maxima would then correctly indicate a fronto-parallel surface corresponding to the noise in the image plane.

Cue combination This paper has treated local estimation of surface shape and orientation, using three independent processes. Clearly, some mechanism is needed for unifying these independent estimates into hypotheses about coherent surfaces.

Brightness discontinuities The linear transformation property (9) is strictly valid only if the brightness pattern is differentiable. Non-differentiable structures such as sharp discontinuities may therefore invalidate (9) to a greater or lesser extent. For example, compression of an ideal step edge in the direction perpendicular to the edge obviously does not affect the magnitude of derivatives estimated by finite differences at all, unlike the case of a smooth edge for which the compression would affect the slope of the edge. We plan to investigate this problem in more detail.

Non-uniform smoothing The reason for adapting the local scale in the computation of the second moment descriptor was to obtain a reasonable trade-off between on the one hand suppression of noise and irrelevant fine-scale structures, and on the other hand distortion of the shape of the brightness pattern due to the isotropic Gaussian smoothing.

However, if the shape estimation methods are based on an affine scale-space representation (Lindeberg, 1994a) instead of the linear scale-space based on rotationally symmetric smoothing, then the shape of the smoothing kernel can be adapted to the local image structure and the shape distortion effects be reduced (Lindeberg and Gårding, 1994). This observation is related to the suggestion by Stone (1990) to adapt the local operators used in shape-from-texture estimation to be isotropic when backprojected to the surface, rather than in the image.

A Appendix

A.1 Transformation property of the second moment matrix

The transformation property (9) of the windowed second moment matrix can be verified as follows. Assume that $L, R : \mathbb{R}^2 \to \mathbb{R}$ are two intensity patterns related by $L(\xi) = R(B\xi)$, where $\xi \in \mathbb{R}^2$, and B is a non-singular linear transformation. Without loss of generality assume $\det B > 0$. Then,

$$\nabla L(\xi) = B^T \nabla R(B\xi), \tag{45}$$

which when substituted into the definition of the windowed second moment matrix yields

$$\mu_L(q) = \iint_{\xi \in \mathbb{R}^2} w(q-\xi)(\nabla L(\xi))(\nabla L(\xi))^T \, d\xi = \iint_{\xi \in \mathbb{R}^2} w(q-\xi) B^T (\nabla R(B\xi))(\nabla R(B\xi))^T B \, d\xi. \tag{46}$$

Substituting $\eta = B\xi$ (with $p = Bq$) we obtain

$$\mu_L(q) = B^T \left\{ \iint_{\eta \in \mathbb{R}^2} w(B^{-1}(p-\eta))(\nabla R(\eta))(\nabla R(\eta))^T (\det B)^{-1} d\eta \right\} B. \tag{47}$$

The integral within brackets is the second moment of R at p computed with respect to the backprojected window function $w'(\eta - p) = (\det B)^{-1} w(B^{-1}(\eta - p))$. This window function is normalized as long as the original window function is, because

$$\iint_{\eta \in \mathbb{R}^2} w(B^{-1}(\eta - p))(\det B)^{-1} \, d\eta = \{\text{let } \eta = B\xi \text{ with } p = Bq\} = \iint_{\xi \in \mathbb{R}^2} w(\xi - q) \, d\xi, \tag{48}$$

which verifies (9). Note, however, that the window function w' will not, in general, be rotationally symmetric.

A.2 Estimating simple distortion gradients

In this appendix a practical procedure for estimation of surface orientation from the area gradient in the case of a locally planar surface will be described. A more detailed description is given in (Lindeberg and Gårding, 1993). The same procedure can with only minor modifications be applied to estimation of surface orientation from any simple distortion gradient.

Equation (32) relates the normalized gradient of projected texel area in the viewsphere Σ to surface orientation and curvature. If the curvature is assumed to be small, an estimate of the surface tilt is given by the negative direction of the area gradient, and an estimate of surface slant is given by $\tan^{-1} ||(\nabla A_\Sigma(p))/(3A_\Sigma(p))||$.

In principle, the area gradient can be estimated by applying a central difference operator to the product mM obtained from the pointwise estimate of F_*. However, for a planar surface the product $A_\Sigma = mM$ is not a linear function of the image coordinates, and so a central difference estimate of the first derivate would be biased by the higher derivatives of A_Σ. A more consistent approach is to transform A_Σ to a form that is linear in the image before the central difference operator is applied, thereby eliminating the bias. This procedure can be simplified even further by transforming the image texel area A_L, rather than the viewsphere texel area A_Σ, to linear form, thus bypassing the need to apply the gaze transformation G_*.

For a planar surface with slant σ and tilt θ, it can be shown (Lindeberg and Gårding, 1993) that

$$(A_L(x,y))^{1/3} = k(f\cos\sigma - (x\cos\theta + y\sin\theta)\sin\sigma), \tag{49}$$

where k is an unknown constant.[12]

Hence, a practical procedure for estimating the local surface orientation from estimates of, for example, the area $A_L(x,y)$ in some region of the image can be described as follows. First, compute (samples of) $h(x,y) = (A_L(x,y))^{1/3}$. Then, estimate the parameters $(h_x, h_y, h(0,0))$, either by central differences or, more robustly, by a weighted least-squares fit of $h(x,y) = h_x x + h_y y + h(0,0)$. Finally, compute the estimated local surface orientation using (49) which can be rewritten

$$\hat{\sigma} = \cos^{-1}\left(\frac{h(0,0)}{\sqrt{f^2h_x^2 + f^2h_y^2 + h^2(0,0)}}\right), \tag{50}$$

$$\hat{\theta} = \arg(h_x, h_y). \tag{51}$$

Note that this procedure only requires $A_L(x,y)$ to be computed up to an arbitrary scale factor.

References

Aloimonos, Y. (1988). Shape from texture. *Biological Cybernetics*, *58*, 345–360.

Babaud, J., Witkin, A. P., Baudin, M., and Duda, R. O. (1986). Uniqueness of the Gaussian kernel for scale-space filtering. *IEEE Trans. Pattern Anal. and Machine Intell.*, *8*, no. 1, 26–33.

Bergen, J. and Adelson, E. (1988). Early vision and texture perception. *Nature*, *333*, 363–364.

Bigün, J., Granlund, G. H., and Wiklund, J. (1991). Multidimensional orientation estimation with applications to texture analysis and optical flow. *IEEE Trans. Pattern Anal. and Machine Intell.*, *13*, no. 8, 775–790.

Bijl, P. and Koenderink, J. J. (1993). Visibility of elliptical Gaussian blobs. *Vision Research*, *33*, no. 2, 243–255.

Blake, A. and Marinos, C. (1990a). Shape from texture: estimation, isotropy and moments. *J. of Artificial Intelligence*, *45*, 323–380.

Blake, A. and Marinos, C. (1990b). Shape from texture: the homogeneity hypothesis. In *Proc. 3rd Int. Conf. on Computer Vision*, pp. 350–353, Osaka, Japan. IEEE Computer Society Press.

Blakemore, C. (1970). A new kind of stereoscopic vision. *Vision Research*, *10*, 1181–1200.

Blostein, D. and Ahuja, N. (1989a). A multiscale region detector. *Computer Vision, Graphics, and Image Processing*, *45*, 22–41.

Blostein, D. and Ahuja, N. (1989b). Shape from texture: integrating texture element extraction and surface estimation. *IEEE Trans. Pattern Anal. and Machine Intell.*, *11*, no. 12, 1233–1251.

Brown, L. G. and Shvaytser, H. (1990). Surface orientation from projective foreshortening of isotropic texture autocorrelation. *IEEE Trans. Pattern Anal. and Machine Intell.*, *12*, no. 6, 584–588.

Caelli, T. (1985). Three processing characteristics of visual texture segmentation. *Spatial Vision*, *1*, 19–30.

Casadei, S., Mitter, S., and Perona, P. (1992). Boundary detection in piecewise homogeneous images. In Sandini, G., editor, *Proc. 2nd European Conf. on Computer Vision*, volume 588 of *Lecture Notes in Computer Science*, pp. 174–183. Springer-Verlag.

Davis, L., Janos, L., and Dunn, S. (1983). Efficient recovery of shape from texture. *IEEE Trans. Pattern Anal. and Machine Intell.*, *5*, no. 5, 485–492.

[12] Similar expressions have been derived e.g. by Blostein and Ahuja (1989b) and Kanatani and Chou (1989).

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *J. of the Optical Society of America*, *4*, 2379–2394.

Florack, L. M. J., ter Haar Romeny, B. M., Koenderink, J. J., and Viergever, M. A. (1992). Scale and the differential structure of images. *Image and Vision Computing*, *10*, no. 6, 376–388.

Förstner, M. A. and Gülch, E. (1987). A fast operator for detection and precise location of distinct points, corners and centers of circular features. In *Proc. Intercommission Workshop of the Int. Soc. for Photogrammetry and Remote Sensing*, Interlaken, Switzerland.

Gårding, J. (1991). *Shape from surface markings.* PhD thesis, Dept. of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm.

Gårding, J. (1992). Shape from texture for smooth curved surfaces in perspective projection. *J. of Mathematical Imaging and Vision*, *2*, 329–352.

Gårding, J. (1993). Shape from texture and contour by weak isotropy. *J. of Artificial Intelligence*, *64*, no. 2, 243–297.

Gårding, J. and Lindeberg, T. (1994). Direct estimation of local surface shape in a fixating binocular vision system. Technical Report ISRN KTH/NA/P--94/08--SE, Dept. of Numerical Analysis and Computing Science, Royal Institute of Technology. Shortened version in Eklundh, J.-O., editor, *Proc. 3rd European Conf. on Computer Vision*, Stockholm, Sweden, volume 800 of *Lecture Notes in Computer Science*, pp. 365–376. Springer-Verlag.

Gibson, J. (1950). *The Perception of the Visual World.* Houghton Mifflin, Boston.

Jones, D. G. and Malik, J. (1992). Determining three-dimensional shape from orientation and spatial frequency disparities. In *Proc. 2nd European Conf. on Computer Vision*, pp. 661–669, Santa Margherita Ligure, Italy.

Jones, J. and Palmer, L. (1987a). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J. of Neurophysiology*, *58*, 1233–1258.

Jones, J. and Palmer, L. (1987b). The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *J. of Neurophysiology*, *58*, 1187–1211.

Julesz, B. (1981). Textons, the elements of perception and their interactions. *Nature*, *290*, 91–97.

Kanatani, K. (1984). Detection of surface orientation and motion from texture by a stereological technique. *J. of Artificial Intelligence*, *23*, 213–237.

Kanatani, K. and Chou, T. C. (1989). Shape from texture: general principle. *J. of Artificial Intelligence*, *38*, 1–48.

Koenderink, J. J. (1984). The structure of images. *Biological Cybernetics*, *50*, 363–370.

Koenderink, J. J. and van Doorn, A. J. (1976). Geometry of binocular vision and a model for stereopsis. *Biological Cybernetics*, *21*, 29–35.

Koenderink, J. J. and van Doorn, A. J. (1990). Receptive field families. *Biological Cybernetics*, *63*, 291–298.

Koenderink, J. J. and van Doorn, A. J. (1992). Generic neighborhood operators. *IEEE Trans. Pattern Anal. and Machine Intell.*, *14*, no. 6, 597–605.

Lindeberg, T. (1990). Scale-space for discrete signals. *IEEE Trans. Pattern Analysis and Machine Intell.*, *12*, no. 3, 234–254.

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*, no. 3, 283–318.

Lindeberg, T. (1993b). Discrete derivative approximations with scale-space properties: A basis for low-level feature extraction. *J. of Mathematical Imaging and Vision*, *3*, no. 4, 349–376.

Lindeberg, T. (1993c). On scale selection for differential operators. In K. A. Høgdra, B. Braathen, K. H., editor, *Proc. 8th Scandinavian Conference on Image Analysis*, pp. 857–866, Tromsø, Norway. Norwegian Society for Image Processing and Pattern Recognition.

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

Lindeberg, T. (1994b) Scale selection for differential operators. Technical Report ISRN KTH/NA/P--94/03--SE, Dept. of Numerical Analysis and Computing Science, Royal Institute of Technology. (Submitted).

Lindeberg, T. and Gårding, J. (1993). Shape from texture from a multi-scale perspective. Technical Report ISRN KTH/NA/P--93/03--SE, Dept. of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm. Shortened version in Nagel, H.-H., editor, *Proc. 4th International Conference on Computer Vision*, pp. 683–691, Berlin, Germany. IEEE Computer Society Press.

Lindeberg, T. and Gårding, J. (1994). Shape-adapted smoothing in estimation of 3-D depth cues from affine distortions of local 2-D brightness structure. In Eklundh, J.-O., editor, *Proc. 3rd European Conf. on Computer Vision*, Stockholm, Sweden, volume 800 of *Lecture Notes in Computer Science*, pp. 389–400. Springer-Verlag.

Malik, J. and Perona, P. (1990). Preattentive texture discrimination with early vision mechanisms. *J. of the Optical Society of America*, *7*, 923–932.

Mardia, K. V. (1972). *Statistics of Directional Data.* Academic Press, London.

Marr, D. (1982). *Vision.* W.H. Freeman, New York.

Marr, D. C. (1976). Early processing of visual information. *Phil. Trans. Royal Soc (B)*, *27S*, 483–524.

O'Neill, B. (1966). *Elementary Differential Geometry.* Academic Press, Orlando, Florida.

Pentland, A. P. (1986). Shading into texture. *J. of Artificial Intelligence*, *29*, 147–170.

Pollard, S. B., Mayhew, J. E. W., and Frisby, J. P. (1985). PMF: A stereo correspondence algorithm using a disparity gradient limit. *Perception*, *14*, 449–470.

Rao, A. R. and Schunk, B. G. (1991). Computing oriented texture fields. *CVGIP: Graphical Models and Image Processing*, *53*, no. 2, 157–185.

Rogers, B. and Cagenello, R. (1989). Orientation and curvature disparities in the perception of three-dimensional surfaces. *Investigative Opthalmology and Visual Science*, *30*, 262.

Stone, J. V. (1990). Shape from texture: textural invariance and the problem of scale in perspective images of surfaces. In *Proc. British Machine Vision Conference*, pp. 181–186, Oxford, England.

Super, B. J. and Bovik, A. C. (1992). Shape-from-texture by wavelet-based measurement of local spectral moments. In *Proc. IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition*, pp. 296–301, Champaign, Illinois.

Turner, M. (1986). Texture discrimination by Gabor functions. *Biological Cybernetics*, *55*, 71–82.

Tyler, C. and Sutter, E. (1979). Depth from spatial frequency difference: An old kind of stereopsis? *Vision Research*, *19*, 859–865.

Voorhees, H. and Poggio, T. (1987). Detecting textons and texture boundaries in natural images. In *Proc. 1st Int. Conf. on Computer Vision*, London, England.

Wildes, R. P. (1981). Direct recovery of three-dimensional scene geometry from binocular stereo disparity. *IEEE Trans. Pattern Anal. and Machine Intell.*, *13*, no. 8, 761–774.

Witkin, A. P. (1981). Recovering surface shape and orientation from texture. *J. of Artificial Intelligence*, *17*, 17–45.

Witkin, A. P. (1983). Scale-space filtering. In *Proc. 8th Int. Joint Conf. Art. Intell.*, pp. 1019–1022, Karlsruhe, West Germany.

Young, R. A. (1985). The Gaussian derivative theory of spatial vision: Analysis of cortical cell receptive field line-weighting profiles. Technical Report GMR-4920, Computer Science Department, General Motors Research Lab., Warren, Michigan.

Young, R. A. (1987). The Gaussian derivative model for spatial vision: I. Retinal mechanisms. *Spatial Vision*, *2*, 273–293.

Yuille, A. L. and Poggio, T. A. (1986). Scaling theorems for zero-crossings. *IEEE Trans. Pattern Anal. and Machine Intell.*, *8*, 15–25.