International Journal of Computer Vision 52(2/3), 97–120, 2003

A Distance Measure and a Feature Likelihood Map Concept for Scale-Invariant Model Matching

IVAN LAPTEV AND TONY LINDEBERG
Computational Vision and Active Perception Laboratory (CVAP), Department of Numerical Analysis and Computer Science, KTH (Royal Institute of Technology), SE-100 44 Stockholm, Sweden
laptev@nada.kth.se
tony@nada.kth.se

Received January 15, 2002; Revised September 27, 2002; Accepted November 22, 2002

Abstract. This paper presents two approaches for evaluating multi-scale feature-based object models. Within the first approach, a scale-invariant distance measure is proposed for comparing two image representations in terms of multi-scale features. Based on this measure, the maximisation of the likelihood of parameterised feature models allows for simultaneous model selection and parameter estimation.

The idea of the second approach is to avoid an explicit feature extraction step and to evaluate models using a function defined directly from the image data. For this purpose, we propose the concept of a feature likelihood map, which is a function normalised to the interval [0, 1], and that approximates the likelihood of image features at all points in scale-space.

To illustrate the applicability of both methods, we consider the area of hand gesture analysis and show how the proposed evaluation schemes can be integrated within a particle filtering approach for performing simultaneous tracking and recognition of hand models under variations in the position, orientation, size and posture of the hand. The experiments demonstrate the feasibility of the approach, and that real time performance can be obtained by pyramid implementations of the proposed concepts.

Keywords: matching, scale-space, image features, tracking, recognition, multi-scale representations

1. Introduction

When computing image descriptors for image interpretation, a primary aim is to emphasise and abstract relevant properties in the data while suppressing others. Common approaches for computing image descriptors involve either (i) the computation of sparse sets of image features (feature detection) or (ii) the computation of dense maps of filter responses (direct methods).

In this respect, a main strength of feature based approaches is that they provide an abstracted and compact description of the local image shape. Image features are usually invariant to intensity transformations and can selectively represent characteristic visual properties of image patterns. In particular, using multi-scale feature detection it is possible to estimate the size of image structures and to represent image patterns in a scale-invariant manner. Moreover, when representing real-world objects, an important constraint originates from the fact that different types of image features will usually be visible depending on the scale of observation. Thus, when building object models for recognition, it is natural to consider hierarchical object models that explicitly encode features at different scales as well as hierarchical relations over scales between these.

The purpose of this work is to develop two complementary approaches for comparing such hierarchical object models to image data. Specifically, we will

be concerned with graph-like and qualitative image representations in terms of multi-scale image features (Koenderink, 1984; Crowley and Sanderson, 1987; Lindeberg, 1993; Pizer et al., 1994; Triesch and von der Malsburg, 1996; Shokoufandeh et al., 1999; Bretzner and Lindeberg, 1999a), which are expressed within a context of feature detection with automatic scale selection (Lindeberg, 1998a, 1998b).

For the first, *sparse feature-based* approach, a scale invariant *distance measure* will be proposed for comparing hierarchical object models to multi-scale features extracted from image data, and we will use this measure for evaluating the likelihood of object models.

For the second, *dense filter-based* approach, the notion of *feature likelihood maps* will be developed. The idea is to compute a function on a multi-scale feature space, which is normalised to the interval [0, 1], and which directly approximates the likelihood of image features in such a way that its response is localised in space and scale, with the strongest responses near the centres of blob-like and ridge-like structures.

A main reason behind these constructions is to provide means for verifying feature-based object hypotheses, either with or without explicit feature detection as a pre-processing stage. In particular, we aim at matching schemes that are provably invariant with respect to scale and contrast changes of the image pattern. As will be shown, the developed schemes easily integrate with particle filtering approaches (Isard and Blake, 1996; MacCormick and Isard, 2000) and in this way enable invariant object tracking and object recognition.

The structure of this paper is as follows: Section 2 reviews how multi-scale image features with automatic scale selection can be extracted from grey-level and colour images. Then, Section 3 describes how a distance measure can be formulated for two sets of multi-scale image features, and how the likelihood of object models can be estimated based on this distance measure. Section 4 presents an alternative approach based on feature likelihood maps, by which scale invariant model matching can be performed without explicit feature detection. In Section 5, these two techniques are integrated with a Bayesian approach for estimating model states, applied to simultaneous hand tracking and hand posture recognition. Section 6 gives a brief review of related works, and Section 7 concludes with a summary and discussion.

2. Scale-Space Image Features

2.1. Scale-Space Representation

For any continuous signal $f : \mathbb{R}^2 \mapsto \mathbb{R}$, the linear scale-space $L : \mathbb{R}^2 \times \mathbb{R}_+ \mapsto \mathbb{R}$

$$L(\cdot;\, t) = g(\cdot;\, t) * f(\cdot), \tag{1}$$

is defined as the convolution of f with Gaussian kernels

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

One reason for considering such a representation is that the Gaussian derivatives

$$\begin{aligned} L_{x^m y^n}(\cdot;\, t) &= \partial_{x^m y^n}(g * f) \\ &= (\partial_{x^m y^n} g) * f = g * (\partial_{x^m y^n} f) \end{aligned} \tag{3}$$

constitute a canonical set of filter kernels given natural symmetry requirements on a visual front-end (Witkin, 1983; Koenderink and van Doorn, 1992; Lindeberg, 1994; Florack, 1997). Another reason is that the evolution over scales of a signal and its Gaussian derivatives provides important cues to local image structure.

2.2. Normalised Derivatives

One scale evolution property that we shall make particular use of here is based on the behaviour over scales of γ-normalised Gaussian derivative operators (Lindeberg, 1998b)

$$L_{\xi^m \eta^n} = t^{(m+n)\gamma/2} L_{x^m y^n} \tag{4}$$

where $\xi = x/t^{\gamma/2}$ and $\eta = y/t^{\gamma/2}$ denote γ-normalised coordinates. It can be shown both theoretically and experimentally that such normalised differential entities assume extrema over scales at scales reflecting the size of local image patterns. Hence, γ-normalised operators can be used for, for example, local size estimation and for constructing scale-invariant image descriptors.

2.3. Scale-Space Extrema

Consider any normalised differential entity $\mathcal{D}_{norm} L$ computed from a homogeneous polynomial of normalised derivatives. Then, a scale-space extremum

$(x_0, y_0; t_0)$ is defined as a point where $\mathcal{D}_{norm}L$ assumes a local extremum with respect to space and scale (Lindeberg, 1998b)

$$\nabla(\mathcal{D}_{norm}L) = 0, \quad \partial_t(\mathcal{D}_{norm}L) = 0. \tag{5}$$

Under rescalings of original image pattern $f'(x) = f(sx)$, scale-space extrema satisfy the following scale invariance property. If $(x_0, y_0; t_0)$ is a scale-space extremum of the original pattern f, then $(sx_0, sy_0; s^2t_0)$ is a scale-space extremum of the rescaled pattern f'.

2.4. Grey-Level Feature Detection

In this work, we will make use of scale-space extrema for detecting blob-like and ridge-like image structures in grey-level and colour images. For grey-level images, blobs can be detected from scale-space maxima of the square of the normalised Laplacian operator

$$\mathcal{B}_{\gamma-norm}L = \left(\nabla^2_{norm}L\right)^2 = (t^\gamma(L_{xx} + L_{yy}))^2 \tag{6}$$

where $\gamma = 1$, while ridges can be detected from scale-space maxima of the following normalised ridge strength measure

$$\begin{aligned}\mathcal{R}_{\gamma-norm}L &= (t^\gamma(L_{pp} - L_{qq}))^2 \\ &= t^{2\gamma}\left((L_{xx} - L_{yy})^2 + 4L_{xy}^2\right)\end{aligned} \tag{7}$$

where $\gamma = 3/4$ and L_{pp}, L_{qq} denote the eigenvalues of the Hessian matrix $\mathcal{H}L$.[1]

2.5. Colour Feature Detection

When detecting image features in situations when there is a poor contrast between the object and the background, more robust image features can be obtained if the feature detection step is based on colour images instead of grey-level images. The abovementioned grey-level feature detection method can be extended to colour feature detection in the following way:

The input colour image is transformed from RGB format to an *Luv*-colour space (Billmeyer and Saltzman, 1982), which separates the grey-level information from the chromatic information. A scale-space representation C is computed from each colour channel, giving rise to a set of multi-scale colour channels L_i. Blobs are detected from scale-space maxima of the normalised Laplacian

$$\begin{aligned}\mathcal{B}_{\gamma-norm}C &= \left(\nabla^2_{norm}C\right)^2 = \sum_i (t\nabla^2 L_i)^2 \\ &= \sum_i t^2(\partial_{xx}L_i + \partial_{yy}L_i)^2\end{aligned} \tag{8}$$

while multi-scale ridges are detected as scale-space extrema of the following normalised measure of ridge strength

$$\mathcal{R}_{\gamma-norm}C = \sum_i t^{2\gamma}((\partial_{xx}L_i - \partial_{yy}L_i)^2 + 4(\partial_{xy}L_i)^2). \tag{9}$$

The motivation for the definition of $\mathcal{B}_{\gamma-norm}C$ is that it constitutes a natural extension of the trace of the Hessian for a vector-valued function. The motivation for the definition of $\mathcal{R}_{\gamma-norm}C$ is that it can be derived from a least-squares fitting of ridge responses for the different colour channels (see Appendix A).

2.6. Covariance Matrix Associated with Image Features

Each image feature detected at a point $(x_0, y_0; t_0)$ in scale-space indicates the presence of a corresponding image structure of size t_0 at position (x_0, y_0). To estimate the spatial extent for non-isotropic image structures, we estimate their covariance matrix Σ_0 from a second moment matrix (Bigün et al., 1991; Rao and Schunk, 1991; Lindeberg, 1994; Gårding and Lindeberg, 1996) computed in a neighbourhood of $(x_0, y_0; t_0)$

$$\begin{aligned}\nu = \sum_i \int_{(x,y)\in\mathbb{R}^2} &\begin{pmatrix} (\partial_x L_i)^2 & (\partial_x L_i)(\partial_y L_i) \\ (\partial_x L_i)(\partial_y L_i) & (\partial_y L_i)^2 \end{pmatrix} \\ &\times g(x - x_0, y - y_0; t_{int})\, dx\, dy\end{aligned} \tag{10}$$

at local scale $t_l = t_0$ equal to the detection scale and at integration scale $t_{int} = c^2 t_0$ proportional to the local scale, where $c > 1$. As shown in Appendix B, in the case of a non-isotropic Gaussian blob with covariance matrix Σ_0, the eigenvectors e_1, e_2 of ν coincide with the eigenvectors of Σ_0. Moreover, for a second moment matrix of such a Gaussian blob computed at local scale t_l, the eigenvalues $t_1 > t_2$ of Σ_0 are related to the eigenvalues $\lambda_1 > \lambda_2$, of ν as $t_1 = \sqrt{t_l^2 c^4 + \frac{\lambda_1}{\lambda_2}(t_2^2 + 2t_2 t_l(1 + c^2) + t_l^2(1 + 2c^2))} - (1 + c^2)t_l$. Using this relation and the fact that the smallest eigenvalue t_2 of Σ_0 (corresponding to the width of

Figure 1. Blob and ridge features for a hand: (a) circles and ellipses corresponding to the significant blob and ridge features extracted from an image of a hand; (b) manually selected image features corresponding to the palm, the fingers and the finger tips of a hand; (c) a mixture of Gaussian kernels associated with the blob and ridge features, illustrating how the selected image features capture the essential structure of a hand.

non-isotropic image structure) can be approximated by the estimated scale t_0 of the scale-space maximum, the covariance matrix Σ_0 can be estimated as

$$\Sigma_0 = \begin{pmatrix} | & | \\ e_1 & e_2 \\ | & | \end{pmatrix} \begin{pmatrix} \tilde{t}_1 & 0 \\ 0 & \tilde{t}_2 \end{pmatrix} \begin{pmatrix} -e_1^T - \\ -e_2^T - \end{pmatrix} \qquad (11)$$

where $\tilde{t}_1 = t_0(\sqrt{c^4 + 4\frac{\lambda_1}{\lambda_2}(1+c^2)} - c^2 - 1)$ and $\tilde{t}_2 = t_0$. Alternatively, a more accurate estimate of Σ_0 can be obtained with an iterative method as presented in Lindeberg and Gårding (1997) and Mikolajczyk and Schmid (2002).

To represent non-isotropic image descriptors graphically, we will henceforth use ellipses centred at $\mu_0 = (x_0, y_0)$ and with covariance matrix Σ_0. Figure 1(a) and (b) show an example of such descriptors computed from an image of a hand.

3. Distance Measure for Multi-Scale Image Features

When representing image patterns and object models by multi-scale image features, as described in Section 2, an obvious problem concerns how to formulate a measure by which comparisons can be made between two feature-based representations. For image features defined solely in terms of positions (x_i, y_i) in the image domain, a Euclidean measure could obviously be expected to be sufficient. For scale-space features, which in addition comprise scale measurements, one could attempt to define a scale-space metric by

$$|(x_i, y_i; t_i) - (x_j, y_j; t_j)|^2 = (x_i - x_j)^2 + (y_i - y_j)^2 + c^2(\sigma_i - \sigma_j)^2 \qquad (12)$$

where $\sigma = \sqrt{t}$ and the constant c would then be necessary due to the fact that space and scale are not fully commensurable entities. An obvious problem with such an approach, however, is to determine the degree of penalty with respect to deviations along the spatial dimensions vs. the scale dimension. Yet, other problems concern how to capture the fact that we may be more willing to accept spatial deviations for coarse scale image features than for fine scale image features, and to be more willing to accept spatial variations along the orientation of a non-isotropic oriented structure than in the perpendicular direction.

In this section, we will address this problem by taking the underlying image structures into account that gave rise to the detected image features. The goal is to construct a distance measure that is invariant under translations, rotations as well as isotropic scalings and intensity transformations.

3.1. Distance Between Two Image Features

Consider two image features w_i and w_j, which in this context can be regarded as either blobs or ridges. For these features, we have associated mean vectors, μ_i and μ_j, as well as covariance matrices, Σ_i and Σ_j. As mentioned above, the distance between a pair of such features must take into account the difference in their position, size, orientation and anisotropy. Rather than treating all these attributes separately, we here propose to model these image features by two-dimensional Gaussian functions in the image domain. The motivation for this choice is that such Gaussian functions can be regarded as idealised model patterns representing the original image structures that gave rise to these features:

- either from a maximum-entropy motivation given fixed values of the mean vector and the covariance matrix (Bevensee, 1993),
- or from the facts that
  - the result of applying blob detection according to (6) to an isotropic Gaussian kernel with mean μ_0 and isotropic covariance matrix $\Sigma_0 = t_0 I$ gives rise to a scale-space maximum at position $\hat{x}_0 = x_0$ and scale $\hat{t}_0 = t_0$,
  - the result of applying ridge detection according to (7) to a non-isotropic Gaussian kernel with mean μ_0 and non-isotropic covariance matrix

$\Sigma_0 = \mathrm{diag}(t_1, t_2)$ gives rise to a scale-space maximum at position $\hat{\mu}_0 = \mu_0$ and scale $\hat{t}_0 \approx t_1$, given that $t_2 \gg t_1$. Subsequent computation of the second moment matrix (10) then leads to an estimate of the covariance matrix of the form $\hat{\Sigma}_0 \approx \mathrm{diag}(t_1, t_2)$.

Thus, each such image feature will be represented by a normalised Gaussian kernel

$$\begin{aligned}\bar{g}(x, \mu, \Sigma) &= h(\Sigma) g(x, \mu, \Sigma) \\ &= \frac{h(\Sigma)}{2\pi\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}, \quad (13)\end{aligned}$$

where $h(\Sigma)$ is a normalisation factor that we will determine in order to obtain a scale-invariant distance measure. Figure 1(c) shows an example of such a representation of a hand model in terms of Gaussian distributions.

Then, to define the actual distance between two features w_i and w_j, we propose to compute the square difference between their associated distributions[2] $\bar{g}_i = \bar{g}(x, \mu_i, \Sigma_i)$ and $\bar{g}_j = \bar{g}(x, \mu_j, \Sigma_j)$:

$$\begin{aligned}\phi^2(w_i, w_j) &= \int_{\mathbb{R}^2} (\bar{g}(x, \mu_i, \Sigma_i) - \bar{g}(x, \mu_j, \Sigma_j))^2 \, dx \\ &= \int_{\mathbb{R}^2} \left(\bar{g}_i^2 + \bar{g}_j^2 - 2\bar{g}_i\bar{g}_j\right) dx. \quad (14)\end{aligned}$$

Using the fact that the product of two Gaussian functions is another Gaussian function with covariance $\hat{\Sigma} = (\Sigma_i^{-1} + \Sigma_j^{-1})^{-1}$, mean $\hat{\mu} = \hat{\Sigma}(\mu_i'\Sigma_i^{-1} + \mu_j'\Sigma_j^{-1})$ (where μ' denotes the transpose of μ) and a different amplitude, i.e.

$$\begin{aligned}&g(x, \mu_i, \Sigma_i) g(x, \mu_j, \Sigma_j) \\ &\quad = C \frac{\sqrt{\det\left(\Sigma_i^{-1}\right)\det\left(\Sigma_j^{-1}\right)}}{2\pi\sqrt{\det\left(\Sigma_i^{-1} + \Sigma_j^{-1}\right)}} g(x, \hat{\mu}, \hat{\Sigma}), \quad (15)\end{aligned}$$

where

$$\begin{aligned}C = \exp\Big(&-\frac{1}{2}\big(\mu_i'\Sigma_i^{-1}\mu_i + \mu_j'\Sigma_j^{-1}\mu_j \\ &- \left(\mu_i'\Sigma_i^{-1} + \mu_j'\Sigma_j^{-1}\right)\left(\Sigma_i^{-1} + \Sigma_j^{-1}\right)^{-1} \\ &\times \left(\Sigma_i^{-1}\mu_i + \Sigma_j^{-1}\mu_j\right)\big)\Big), \quad (16)\end{aligned}$$

the integral in (14) can be evaluated in closed form:

$$\begin{aligned}\phi^2(w_i, w_j) = &\frac{h^2(\Sigma_i)}{4\pi\sqrt{\det(\Sigma_i)}} \underbrace{\int_{\mathbb{R}^2} g(x, \mu_i, \Sigma_i/2)\, dx}_{=1} \\ &+ \frac{h^2(\Sigma_j)}{4\pi\sqrt{\det(\Sigma_j)}} \underbrace{\int_{\mathbb{R}^2} g(x, \mu_j, \Sigma_j/2)\, dx}_{=1} \\ &- C \frac{h(\Sigma_i)h(\Sigma_j)\sqrt{\det\left(\Sigma_i^{-1}\right)\det\left(\Sigma_j^{-1}\right)}}{\pi\sqrt{\det\left(\Sigma_i^{-1} + \Sigma_j^{-1}\right)}} \\ &\times \underbrace{\int_{\mathbb{R}^2} g(x, \hat{\mu}, \hat{\Sigma})\, dx}_{=1}. \quad (17)\end{aligned}$$

To be applicable for contrast, translation, rotation and scale invariant matching, ϕ should be invariant to such (simultaneous) transformations of w_i and w_j. Contrast, translation and rotation invariance is straightforward. From (17), (or from a combination of (13) and (14) with scale invariance argument) it can be seen that $\phi(w_i, w_j)$ will be scale-invariant if and only if we choose

$$h(\Sigma) = \sqrt[4]{\det(\Sigma)}. \quad (18)$$

Thus, we obtain

$$\phi^2(w_i, w_j) = \frac{1}{2\pi} - C \frac{\sqrt[4]{\det\left(\Sigma_i^{-1}\right)\det\left(\Sigma_j^{-1}\right)}}{\pi\sqrt{\det\left(\Sigma_i^{-1} + \Sigma_j^{-1}\right)}}. \quad (19)$$

By construction, the distance measure ϕ assumes its minimum value zero only when the two features w_i and w_j are equal, while its value increases when the features deviate in their positions, sizes or shapes. This idea is illustrated in Fig. 2. Obviously, the square difference of two Gaussian functions as illustrated in Fig. 2(c) becomes flat for similar features, while its volume increases for features with deviating means or covariances.

Finally, the fact that ϕ can be regarded as a distance measure follows from the fact that we map image descriptors (μ, Σ) to Gaussian functions, and define ϕ in terms of the norm of functions in L_2. Thus, the triangle inequality holds for ϕ, i.e., for any triplet of multi-scale image features w_i, w_j and w_k we have

$$\phi(w_i, w_j) \leq \phi(w_i, w_k) + \phi(w_k, w_j). \quad (20)$$

In addition, ϕ satisfies $\phi(w_i, w_j) \geq 0$ for any w_i, w_j with equality if and only if $w_i = w_j$.

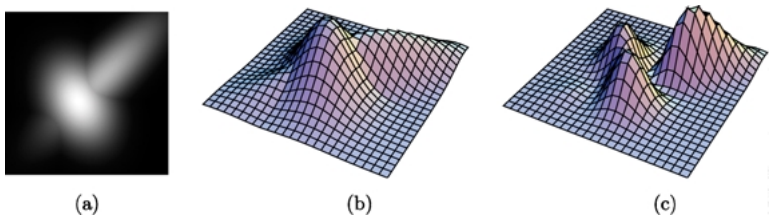

Figure 2. Two overlapping features represented by Gaussian functions in the image domain are shown: (a) as a gray-value image and (b) as a three-dimensional plot. (c) illustrates the square difference of these Gaussian functions, which after integration gives the distance measure between the features.

3.2. Distance Between Model and Data Features

Then, to perform matching between object models and image patterns, let us consider two sets $\mathcal{W}^m$, $\mathcal{W}^d$ with N^m model and N^d data features, respectively. Specifically, let us consider the model and the data as two mixtures of Gaussian distributions in the image domain

$$G^m = \sum_i^{N^m} \bar{g}\big(x, \mu_i^m, \Sigma_i^m\big), \quad G^d = \sum_i^{N^d} \bar{g}\big(x, \mu_i^d, \Sigma_i^d\big), \tag{21}$$

where $\bar{g}(x, \mu_i^m, \Sigma_i^m)$ and $\bar{g}(x, \mu_i^d, \Sigma_i^d)$ are normalised Gaussian functions associated with the model and the data features defined in (13). In a similar way as for the distance between two image features, we define the distance between the composed model and the set of data features by integrating the square difference of their associated Gaussian mixture functions:[3]

$$\Phi^2(\mathcal{W}^m, \mathcal{W}^d) = \int_{\mathbb{R}^2} (G^m - G^d)^2 \, dx. \tag{22}$$

Figure 3 illustrates this idea on two model features and two data features. While the overlapping model and the data features cancel each other, the mismatched features in both the model and the data increase the square difference $(G^m - G^d)^2$ (see Fig. 3 (b)) and consequently increase the distance Φ. By expanding (22) we get

$$\begin{aligned}\Phi^2(\mathcal{W}^m, \mathcal{W}^d) &= \underbrace{\sum_i^{N^m}\sum_j^{N^m} \int_{\mathbb{R}^2} \bar{g}_i^m \bar{g}_j^m \, dx}_{Q_1} + \underbrace{\sum_i^{N^d}\sum_j^{N^d} \int_{\mathbb{R}^2} \bar{g}_i^d \bar{g}_j^d \, dx}_{Q_2} \\ &\quad \underbrace{- 2 \sum_i^{N^m}\sum_j^{N^d} \int_{\mathbb{R}^2} \bar{g}_i^m \bar{g}_j^d \, dx}_{Q_3}\end{aligned} \tag{23}$$

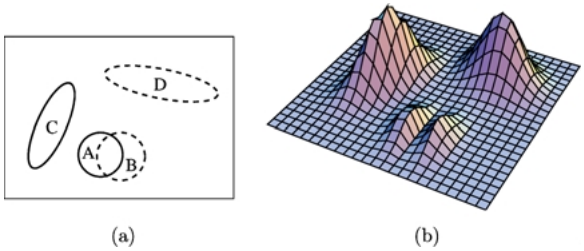

Figure 3. Two model features (solid ellipses) and two data features (dashed ellipses) in (a) are compared by evaluating the square difference of associated Gaussian functions. While the overlapping model (A) and the data (B) features cancel each other, the mismatched features (C and D) increase the square difference.

whose computation requires comparisons of all of the feature pairs. We can note, however, that blobs and ridges in $\mathcal{W}^m$ and $\mathcal{W}^d$ that originate from different image structures should be well separated by their positions scales or both. In practice, overlaps between detected features usually occur due the multiple responses to the same image structure and therefore are not relevant for our analysis and can be discarded. Hence, to save computations, we approximate the terms Q_1 and Q_2 in (23) by

$$Q_1 \approx \sum_i^{N^m} \int_{\mathbb{R}^2} \big(\bar{g}_i^m\big)^2 \, dx, \quad Q_2 \approx \sum_i^{N^d} \int_{\mathbb{R}^2} \big(\bar{g}_i^d\big)^2 \, dx. \tag{24}$$

Additionally, we can note that most of the model and the data features will not overlap either, and most of the products in Q_3 will be close to zero. Thus, we compare each model feature w_i^m only to its closest data feature $w_{k_i}^d$ and approximate Q_3 by

$$Q_3 \approx 2 \sum_i^{N^m} \int_{\mathbb{R}^2} \bar{g}_i^m \bar{g}_{k_i}^d \, dx. \tag{25}$$

Here, the indices $k_{1,\ldots,N^m}$ correspond to data features matched by the model, while the data features with indices $k_{N^m+1,\ldots,N^d}$ (we assume $N^d > N^m$) are considered as outliers. Note, that matching according to (25) does not guarantee a one-to-one correspondence between the model and the data features. Such a constraint, however, could easily be added in the implementation.

Taking advantage of the approximations above, we can finally simplify the expression (23) for Φ into Ψ as

$$\begin{aligned}\Psi^2(\mathcal{W}^m, \mathcal{W}^d) &= \sum_{i=1}^{N^m} \int_{\mathbb{R}^2} \left(\left(\bar{g}_i^m\right)^2 - 2\bar{g}_i^m \bar{g}_{k_i}^d + \left(\bar{g}_{k_i}^d\right)^2\right) dx \\ &\quad + \sum_{i=N^m+1}^{N^d} \int_{\mathbb{R}^2} \left(\bar{g}_{k_i}^d\right)^2 dx \\ &= \underbrace{\sum_{i=1}^{N^m} \phi^2\left(w_i^m, w_{k_i}^d\right)}_{\textit{offset criterion}} + \underbrace{\frac{N^d - N^m}{4\pi}}_{\textit{outlier criterion}}, \quad (26)\end{aligned}$$

where ϕ is the distance measure between features w_i^m and $w_{k_i}^d$ according to (17). Since Ψ is now expressed in terms of distances ϕ, it is clear that Ψ inherits the invariance properties to joint translations, rotations and re-scalings of image features as well as intensity transformations.

An important property of the proposed distance measure is that it equally accounts for mismatches in the model and in the data. Given a set of parameterised models, minimisation of Ψ over different models and their parameters according to (26) provides a trade-off between (i) the distance between the matched model and the data features (an offset criterion) and (ii) the number of mismatched data features (an outlier criterion). In this way, the measure Ψ enables matching and selection of object models that may be subsets of each other. This property will be highly important for the problem of tracking and recognising hand postures as addressed in Section 5.

The measure Ψ is easy to compute in practice. For each model feature f_i^m, we choose the data feature $f_{k_i}^d$ that minimises $\phi(f_i^m, f_{k_i}^d)$. Then, the sum over all $\phi(f_i^m, f_{k_i}^d)$, $i \in [1, N^m]$, together with a constant term for outliers give the desired estimate.

Currently, all the features in the model contribute equally to Ψ. By modelling the uncertainty of the features and multiplying the contribution to Ψ from the individual feature pairs by their estimated certainty values, the influence of uncertain image features can be decreased and thus facilitate matching to highly noisy data.

3.3. *Likelihood Estimate*

When performing tracking and recognition, the matching problem can be expressed as finding the best hypothesis of an object model $M(X_0)$ (represented by a set of model parameters X_0) given a set of image measurements $\mathcal{I}$. Thus, we must search for the minimum of Ψ over all X. For the purpose of tracking and recognition using particle filtering, as will be addressed in Section 5, however, it turns out to be more convenient to maximise a likelihood measure $p(\mathcal{I} \mid X) = p(\mathcal{W}^d \mid \mathcal{W}^m)$ instead of minimising a distance measure. Thus, in analogy with Gibbs distributions (Geman and Geman, 1984), we will define an approximate likelihood function in terms of Ψ as

$$p(\mathcal{W}^d \mid \mathcal{W}^m) = e^{-\Psi^2(\mathcal{W}^m, \mathcal{W}^d)/2\sigma^2}, \quad (27)$$

where σ^2 is a free parameter. Variation of σ enables control over the sharpness of the maxima in the likelihood function. This property is especially useful for optimisation algorithms such as the one that will be used in Section 5. Whereas initial (coarse) estimates of the locations of the maxima can be obtained using high values of σ, lower values of σ enable more precise adjustments of these estimates.

4. Feature Likelihood Maps

In several computer vision applications relating to tracking and recognition, one faces the problem of estimating the likelihood of object models relative to image data. While we in the previous approach first performed feature detection, and then defined a likelihood function in terms of a distance measure between the image features, an alternative approach consists of estimating the likelihoods of object models directly from image data.

In this section, we will address this problem within the context of feature-based object models and in such a way that the approach is compatible with an automatic scale selection mechanism. We will introduce a notion of feature likelihood maps $\mathcal{M} : \mathbb{R}^2 \times \mathbb{R}_+ \mapsto \mathbb{R}$, based on the idea that for a blob of size t_0 located at a point (x_0, y_0) in space, the feature likelihood map $\mathcal{M}$ should satisfy the following basic properties:

– $\mathcal{M}$ should assume its maximum value one at $(x_0, y_0; t_0)$,
– $\mathcal{M}$ should assume high values in a small neighbourhood of $(x_0, y_0; t_0)$, and
– $\mathcal{M}$ should decrease monotonically towards zero elsewhere.

Additionally, $\mathcal{M}$ should not give preference to blobs of any particular size, position or amplitude, and should thus be invariant to scalings and translations in the image as well as to local changes of the image contrast.

With regard to model matching, the underlying idea behind the construction of this likelihood map is that the likelihood of model features should be possible to estimate by mere lookup in $\mathcal{M}$. Moreover, estimation of likelihoods from local neighbourhoods has benefits in efficiency, since computations of $\mathcal{M}$ can be done "on demand". These properties make the approach convenient for object tracking and object recognition using particle filtering, as addressed in Section 5. It should be emphasised, however, that the proposed feature likelihood maps will not constitute true probability estimates. Rather the intention is to provide heuristic estimates that can be derived in closed form without need for an explicit learning stage.

4.1. Likelihood Map for Blobs in the 1-D Case

For simplicity of illustration, let us first construct a likelihood map for a prototype of a one-dimensional blob given by a Gaussian function $f(x) = g(x; x_0, t_0)$ with parameters t_0 and x_0 corresponding to the size and the position of the blob (see Fig. 4(a)). Using the semi-group property of the Gaussian kernel, it follows that the scale-space representation of f is $L(x; t) = g(x; x_0, t + t_0)$. The γ-normalised second-order derivative of this expression is

$$L_{\xi\xi}(\xi; t) = t^{\gamma_2} L_{xx}(x; t) = -\frac{t^{\gamma_2}(t + t_0 + (x - x_0)^2)}{\sqrt{2\pi(t + t_0)^5}} e^{-\frac{(x-x_0)^2}{2(t+t_0)}} \quad (28)$$

If we choose $\gamma_2 = 3/4$, then it can be shown (Lindeberg, 1998a) that $L_{\xi\xi}$ assumes a local extremum over space and scale at the point (x_0, t_0) in scale-space that corresponds to the position x_0 and the size t_0 of the original blob f. Thus, $L^2_{\xi\xi}$ satisfies some of the required properties of the desired likelihood map $\mathcal{M}$. However, $L^2_{\xi\xi}$ is not invariant to the local amplitude of the signal (see Fig. 4(b)).

4.1.1. Quasi-Quadrature. A standard methodology for amplitude estimation in signal processing is in terms of quadrature filter pairs (h_+, h_-), from which the amplitude can be estimated as

$$\mathcal{Q}f = (h_+ * f)^2 + (h_- * f)^2. \quad (29)$$

Strictly, a quadrature filter pair is defined from a Hilbert transform, in such a way that $\mathcal{Q}$ is phase-independent. Within the framework of scale-space derivatives, a quadrature entity can be approximated by a pair of normalised first- and second-order Gaussian derivative operators (Koenderink and van Doorn, 1992; Lindeberg, 1998b):

$$\mathcal{Q}_1 L = A L^2_\xi + L^2_{\xi\xi} = A t^{\gamma_1} L^2_x + t^{2\gamma_2} L^2_{xx} \quad (30)$$

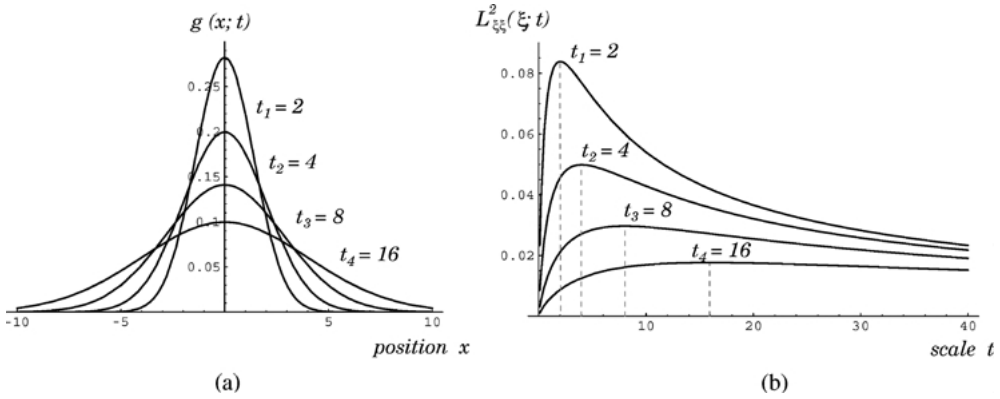

Figure 4. (a): Gaussian kernels of various widths; (b): Evolution over scales of the second order normalised derivative operator $L^2_{\xi\xi}$ in the case when $\gamma_2 = 3/4$.

where A is a constant and $L_\xi = t^{\gamma_1/2} L_x(x;t)$ is the normalised first-order derivative operator. Concerning the choice of the γ-normalisation parameters, we choose $\gamma_2 = 3/4$ for the second-order derivative operator as above, while $\gamma_1 = 1/2$ is chosen to satisfy scale-invariant behaviour of $\mathcal{Q}_1 L$. This scale-invariance follows if we write $\mathcal{Q}_1 L$ as

$$\mathcal{Q}_1 L = t^{-1/2}\left(A t L_x^2 + t^2 L_{xx}^2\right) \tag{31}$$

i.e. as a power function of the scale parameter t multiplied by a non-homogeneous differential expression in terms of γ-normalised derivatives with $\gamma = 1$. Note that $\gamma_1 = 1/2$ has also been established as a natural choice with regard to edge detection (Lindeberg, 1998a).

Concerning the choice of A, we can choose $A \approx 4/e$ such that the response of the $\mathcal{Q}_1 L$ is approximately constant over space in a local neighbourhood of $(x_0; t_0)$ (Lindeberg, 1997).

4.1.2. Incorporating Stability of Image Structures Over Scales. A main idea towards constructing a likelihood map for blob structures is to divide the blob measure $L_{\xi\xi}^2$ by an amplitude estimate $\mathcal{Q}L$

$$\mathcal{M} = \frac{L_{\xi\xi}^2}{\mathcal{Q}L}. \tag{32}$$

If we would take $\mathcal{Q}L = \mathcal{Q}_1 L$, however, the corresponding feature likelihood map would not be suitable for scale selection, since its response at the blob point $x = x_0$ would be constant over scale. To localise image structures over both space and scale, we propose to include the stability of image structures over scales (corresponding to low values of derivatives with respect to scale) as a main component in $\mathcal{M}$. Specifically, we propose to measure the stability of blob structures over scales in terms of the derivative of $L_{\xi\xi}$ with respect to effective scale $\tau = \log t$. Using

$$\partial_\tau = t\,\partial_t \tag{33}$$

and the fact that all Gaussian derivatives satisfy the diffusion equation

$$\partial_t(L_{x^\alpha}) = \frac{1}{2}\,\partial_{xx}(L_{x^\alpha}), \tag{34}$$

it follows that the derivative of the γ-normalised second-order Gaussian derivative with respect to effective scale can be expressed as:

$$\begin{aligned} L_{\xi\xi\tau}(\xi;t) = t\,\partial_t L_{\xi\xi}(\xi;t) &= \gamma_3 t^{\gamma_3} L_{xx} + t^{\gamma_3} L_{xxt} \\ &= \gamma_3 t^{\gamma_3} L_{xx} + \frac{t^{\gamma_3+1}}{2} L_{xxxx}. \end{aligned} \tag{35}$$

By adding this expression to (30), we thus propose to extend $\mathcal{Q}_1 L$ into

$$Q_2 L = A L_\xi^2 + B L_{\xi\xi\tau}^2 + L_{\xi\xi}^2, \tag{36}$$

where we choose $\gamma_3 = 3/4$, since for a Gaussian prototype signal the scale-space maximum of $L_{\xi\xi}^2/\mathcal{Q}_2 L$ is assumed at

$$t_{max} = \frac{2\gamma_3}{3 - 2\gamma_3} t_0. \tag{37}$$

Thus, the maximum response over scales is assumed at a scale reflecting the size of the blob $t_{max} = t_0$ if and only if $\gamma_3 = 3/4$.

Figure 5(a) and (b) illustrate the evolution of the components in this expression, i.e. L_ξ^2, $L_{\xi\xi\tau}^2$ and $L_{\xi\xi}^2$, over space and scale. As can be seen, the responses of L_ξ^2 and $L_{\xi\xi\tau}^2$ complement the response of $L_{\xi\xi}^2$ by assuming high values where $L_{\xi\xi}^2$ is low and vice versa. Thus, one can expect that by an appropriate choice of the weights A and B, $\mathcal{Q}_2 L$ will be approximately constant in a neighbourhood of (x_0, t_0). Such a behaviour is illustrated in Fig. 5(c) and (d).

4.1.3. Invariance Properties. If we consider the ratio $L_{\xi\xi}^2/\mathcal{Q}_2 L$, it is apparent that the amplitude cancels between the numerator and the denominator. Thus, we achieve invariance to local affine intensity transformations. Moreover, since $\mathcal{Q}_2 L \geq L_{\xi\xi}^2 \geq 0$ it follows that the ratio $L_{\xi\xi}^2/\mathcal{Q}_2 L$ will always be in the range [0, 1].

Regarding scale-invariance, the following property holds: If for a signal f the entity $L_{\xi\xi}^2/\mathcal{Q}_2 L$ has a scale-space maximum at a position $(x_0; t_0)$ in scale-space, then under a rescaling of f by a factor s, there will be a scale-space maximum at $(s x_0; s^2 t_0)$ in the ratio $L_{\xi\xi}^2/\mathcal{Q}_2 L$ computed for the rescaled image pattern.

The relative magnitudes of $L_{\xi\xi}^2$ and $\mathcal{Q}_2 L$ are illustrated in Fig. 5(c) and (d). From these graphs it can be seen that the response of the ratio $L_{\xi\xi}^2/\mathcal{Q}_2 L$ will be localised in space and scale around the centre of the Gaussian blob and at a scale reflecting the size of the blob.

To conclude, we have shown that for a Gaussian blob the ratio $L_{\xi\xi}^2/\mathcal{Q}_2 L$ satisfies all the stated requirements

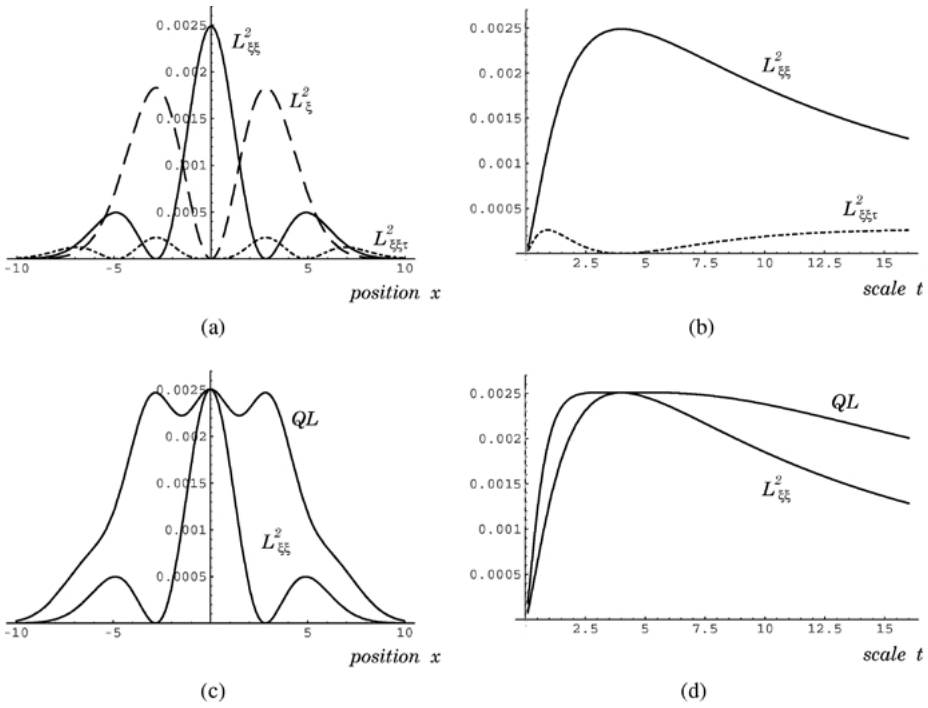

Figure 5. (a) and (b): Evolution of L_ξ^2, $L_{\xi\xi}^2$ and $L_{\xi\xi\tau}^2$ over space and scale when applied to a Gaussian blob centred at $x_0 = 0$ and with variance $t_0 = 4$; (c) and (d): Evolution of $L_{\xi\xi}^2$ and $\mathcal{Q}_2 L$ when using the parameter values $A = 1$ and $B = 2.8$. Note that $\mathcal{Q}_2 L$ is approximately constant over space and scale in the neighbourhood of (x_0, t_0).

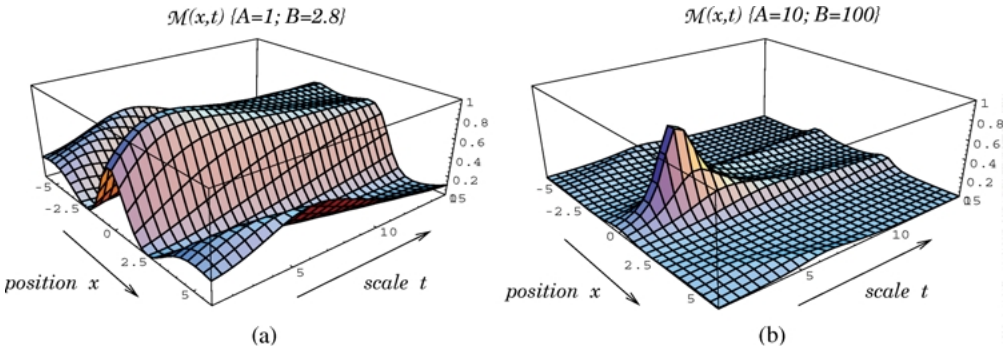

Figure 6. Evolution of the likelihood map $\mathcal{M}$ over space and scale for different values of parameters A and B when applied to a Gaussian blob ($x_0 = 0$, $t_0 = 4$). The more precise localisation of responses over scale in (b) is desirable for e.g. size estimation and matching.

on the feature likelihood map. Thus, we define

$$\mathcal{M} = \frac{L_{\xi\xi}^2}{\mathcal{Q}_2 L} = \frac{L_{\xi\xi}^2}{A L_\xi^2 + B L_{\xi\xi\tau}^2 + L_{\xi\xi}^2}. \qquad (38)$$

4.1.4. Determination of the Free Parameters A and B. Concerning the parameters A and B, which so far are undetermined, it can be verified that $A \approx 1$ and $B \approx 3$ give an approximately constant behaviour of the denominator of $\mathcal{M}$ around $(x_0; t_0)$. This was the original design criterion when the quasi quadrature entity (30) was proposed (Lindeberg, 1998b). Figure 6(a) shows the behaviour of $\mathcal{M}$ in this case. As can be seen, however, the peak around $(x_0; t_0)$ is rather wide in the scale direction, and there are two quite strong side lobes in the spatial direction.

For the purpose of dense scale selection with application to recognition, it is desirable to have a more narrow and localised response with respect to scale and space. For this reason, we manually adjust the parameters A and B to the values $A = 10$ and $B = 100$ and obtain a desired behaviour of $\mathcal{M}$ as illustrated in Fig. 6(b). As an alternative approach, the parameters A and B could also be learned automatically from training data.

4.2. *Likelihood Maps for Blobs and Ridges in the 2-D Case*

The likelihood map defined in Section 4.1 can be easily extended to two dimensions. Consider again a Gaussian kernel $f = g(x, y; x_0, y_0, t_0)$ as a prototype of an image blob of size t_0 and centred at (x_0, y_0). The scale-space representation of this signal is given by $L(x, y; t) = g(x, y; x_0, y_0, t + t_0)$ and its position and scale can be recovered from the normalised Laplacian operator

$$\nabla^2_{norm} L = L_{\xi\xi} + L_{\eta\eta} = t^{\gamma_2} L_{xx}(x, y; t) + t^{\gamma_2} L_{yy}(x, y; t) \tag{39}$$

which assumes a local maximum at (x_0, y_0, t_0) if we take $\gamma_2 = 1$. To construct a quadrature entity $\mathcal{Q}$, we consider the normalised gradient magnitude

$$|\nabla_{norm} L| = \sqrt{L_\xi^2 + L_\eta^2} = t^{\gamma_1/2} \sqrt{L_x^2 + L_y^2}, \tag{40}$$

as the analogue of L_ξ in the one-dimensional case. Moreover, we take

$$\partial_\tau \nabla^2_{norm} L = L_{\xi\xi\tau} + L_{\eta\eta\tau} = \gamma_3 t^{\gamma_3} (L_{xx} + L_{yy}) + \frac{t^{\gamma_3+1}}{2} (L_{xxxx} + L_{yyyy} + 2L_{xxyy}) \tag{41}$$

as the analogue to $L_{\xi\xi\tau}$ in order to complement the response of $\nabla^2_{norm} L$ along the scale direction. Using these two entities, we thus define

$$\mathcal{Q}_3 L = A(\nabla_{norm} L)^2 + B\left(\partial_\tau \nabla^2_{norm} L\right)^2 + \left(\nabla^2_{norm} L\right)^2 \tag{42}$$

from which the *Laplacian likelihood map* is defined as

$$\mathcal{M}_{lap} = \frac{\left(\nabla^2_{norm} L\right)^2}{\mathcal{Q}_3 L} = \frac{(L_{\xi\xi} + L_{\eta\eta})^2}{A\left(L_\xi^2 + L_\eta^2\right) + B(L_{\xi\xi\tau} + L_{\eta\eta\tau})^2 + (L_{\xi\xi} + L_{\eta\eta})^2}. \tag{43}$$

Concerning the choice of γ_1, we can observe that $\gamma_1 = 1$ is a necessary requirement once we have chosen $\gamma_2 = 1$, if we in addition require that the sum of the first- and second-order derivative responses should be written on the form

$$\nabla_{norm} L^2 + \left(\nabla^2_{norm} L\right)^2 = t^{-\Gamma}\left(t\left(L_x^2 + L_y^2\right) + t^2 (L_{xx} + L_{yy})^2\right) \tag{44}$$

for some Γ, to allow for a scale invariant scale selection procedure. Concerning the choice of γ_3, it can be shown that for a Gaussian blob pattern, the local maximum over scale of $\mathcal{M}_{lap}$ is assumed at

$$t_{max} = \frac{\gamma_3}{2 - \gamma_3} t_0. \tag{45}$$

Thus, $t_{max} = t_0$ if and only if $\gamma_3 = 1$.

It is straightforward to show that $\mathcal{M}_{lap}$ is rotationally invariant, and invariant with respect to scale as well as the amplitude of the blob; it assumes values in the range [0, 1] and for a Gaussian blob of size t_0 centred at (x_0, y_0) the maximum value 1 is assumed at $(x_0, y_0; t_0)$. Hence, $\mathcal{M}_{lap}$ has essentially similar properties as the likelihood map (38) in the one-dimensional case. Figure 7(a)–(c) illustrate how, with $A = 10$ and $B = 100$, $\mathcal{M}_{lap}$ assumes a rather sharp maximum at (x_0, y_0, t_0) and rapidly decreases with deviations from this point.

4.2.1. Suppression of Saddle Regions and Noise. Besides blobs and ridges, however, $\mathcal{M}_{lap}$ will also respond to certain saddle points. This occurs when $\nabla_{norm} L = 0$, $\partial_\tau(\nabla^2_{norm} L) = 0$ and when the eigenvalues λ_1 and λ_2 of the Hessian matrix have different signs. To suppress such points, we introduce a saddle suppression factor

$$\mu = \frac{\lambda_1^2 + \lambda_2^2 + 2\lambda_1\lambda_2}{\lambda_1^2 + \lambda_2^2 + 2|\lambda_1\lambda_2|} = \frac{L_{\xi\xi}^2 + L_{\eta\eta}^2 + 2L_{\xi\xi}L_{\eta\eta}}{L_{\xi\xi}^2 + L_{\eta\eta}^2 - 2L_{\xi\eta}^2 + 2\left|L_{\xi\xi}L_{\eta\eta} - L_{\xi\eta}^2\right|}. \tag{46}$$

It can be seen that μ is equal to one when λ_1 and λ_2 have the same sign (i.e., for emphasised blob and ridge structures), while μ decreases towards zero if λ_1 and λ_2 have equal magnitude and opposite sign. Moreover, to suppress the influence of spurious noise structures

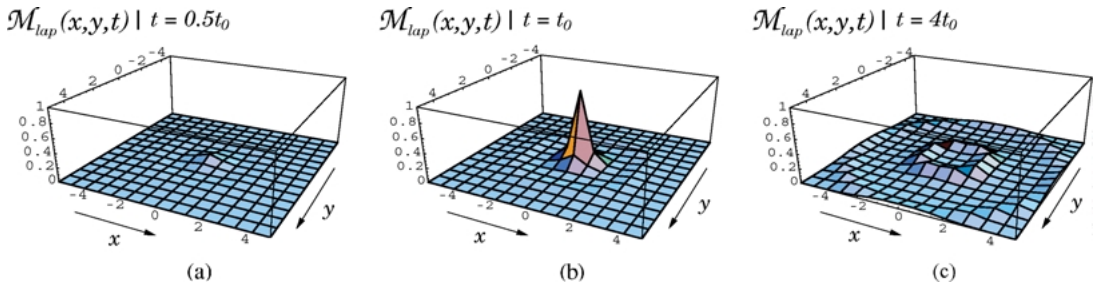

Figure 7. Evolution of the likelihood map $\mathcal{M}_{lap}$ over space and scale for a two-dimensional Gaussian blob defined by ($x_0 = 0$, $y_0 = 0$, $t_0 = 1$). Plots in (a), (b) and (c) illustrate $\mathcal{M}_{lap}$ for scale values $t = 0.5$, 1 and 4.

of amplitude lower than ε, we introduce a small noise-suppression parameter $\varepsilon_{norm} = \varepsilon/t$ in the denominator, where the normalisation with the factor $1/t$ preserves the scale-invariance of $\mathcal{M}$. Thus, we define a saddle- and noise-suppressed feature likelihood map as

$$\tilde{\mathcal{M}}_{lap} = \mu^k \mathcal{M}_{lap} = \frac{\mu^k (L_{\xi\xi} + L_{\eta\eta})^2}{A\left(L_\xi^2 + L_\eta^2\right) + B(L_{\xi\xi\tau} + L_{\eta\eta\tau})^2 + (L_{\xi\xi} + L_{\eta\eta})^2 + \varepsilon_{norm}^2} \tag{47}$$

where we have chosen $k = 4$ and ε is currently chosen manually. Alternatively, ε can be estimated using standard noise level estimation schemes.

4.2.2. Specialised Likelihood Maps for Blobs and Ridges. The proposed Laplacian likelihood map emphasises both blob-like and ridge-like image structures. For the purpose of matching, however, it can be desirable to construct specialised maps with selective responses to different image structures. Hence, to emphasise circular blob-like structures, we introduce a specialised blob likelihood map

$$\mathcal{M}_{blob} = \frac{4\mu^k \left| L_{\xi\xi} L_{\eta\eta} - L_{\xi\eta}^2 \right|}{A\left(L_\xi^2 + L_\eta^2\right) + B(L_{\xi\xi\tau} + L_{\eta\eta\tau})^2 + (L_{\xi\xi} + L_{\eta\eta})^2 + \varepsilon_{norm}^2}, \tag{48}$$

and to emphasise elongated ridge-like structures, we construct a ridge likelihood map

$$\mathcal{M}_{ridge} = \frac{\mu^k \left((L_{\xi\xi} - L_{\eta\eta})^2 + 4L_{\xi\eta}^2 \right)}{A\left(L_\xi^2 + L_\eta^2\right) + B(L_{\xi\xi\tau} + L_{\eta\eta\tau})^2 + (L_{\xi\xi} + L_{\eta\eta})^2 + \varepsilon_{norm}^2}. \tag{49}$$

For more details on the construction of $\mathcal{M}_{blob}$ and $\mathcal{M}_{ridge}$, see Laptev and Lindeberg (2001a).

4.3. Experiments

Figure 8(a) shows the result of computing the Laplacian feature likelihood map for a synthetic image with two Gaussian blobs and one Gaussian ridge. As can be observed, $\tilde{\mathcal{M}}_{lap}$ assumes high values for both blobs and ridges and the responses are well localised in space and scale, with the main peaks corresponding to the positions and the sizes of the original image structures.

One minor artifact arises, however, for the ridge-like structure. Due to the calibration of the scale normalisation scheme for a Gaussian blob, it follows that the selected scale for a Gaussian ridge of width t_0 will be $2t_0$. This problem is avoided when computing specialised likelihood maps for blobs and ridges according to (48) and (49). Figure 8(b) and (c) show the result of computing these feature likelihood maps for the same pattern as the Laplacian feature likelihood map. As expected, the blob likelihood map gives dominant responses to the blobs, while the ridge likelihood map mainly emphasises the ridge. Moreover, by using appropriate γ-values, i.e. $\gamma_1 = 1$, $\gamma_2 = 1$ for $\mathcal{M}_{blob}$ and $\gamma_1 = 1/2$, $\gamma_2 = 3/4$ for $\mathcal{M}_{ridge}$, the variances of the original blobs and the ridge in images are successfully recovered. It can also be noted, that since the end points of the ridge share properties of both blobs and ridges, both the blob and ridge likelihood maps give weak responses around these points.

Figure 9 shows the result of computing $\mathcal{M}_{blob}$ and $\mathcal{M}_{ridge}$ for an image of a hand. Here, it can be seen that $\mathcal{M}_{blob}$ gives dominant responses to blob structures and weaker responses to line terminations, while $\mathcal{M}_{ridge}$ emphasises the elongated ridge-like structures, such as the fingers and the thin patterns in the background.

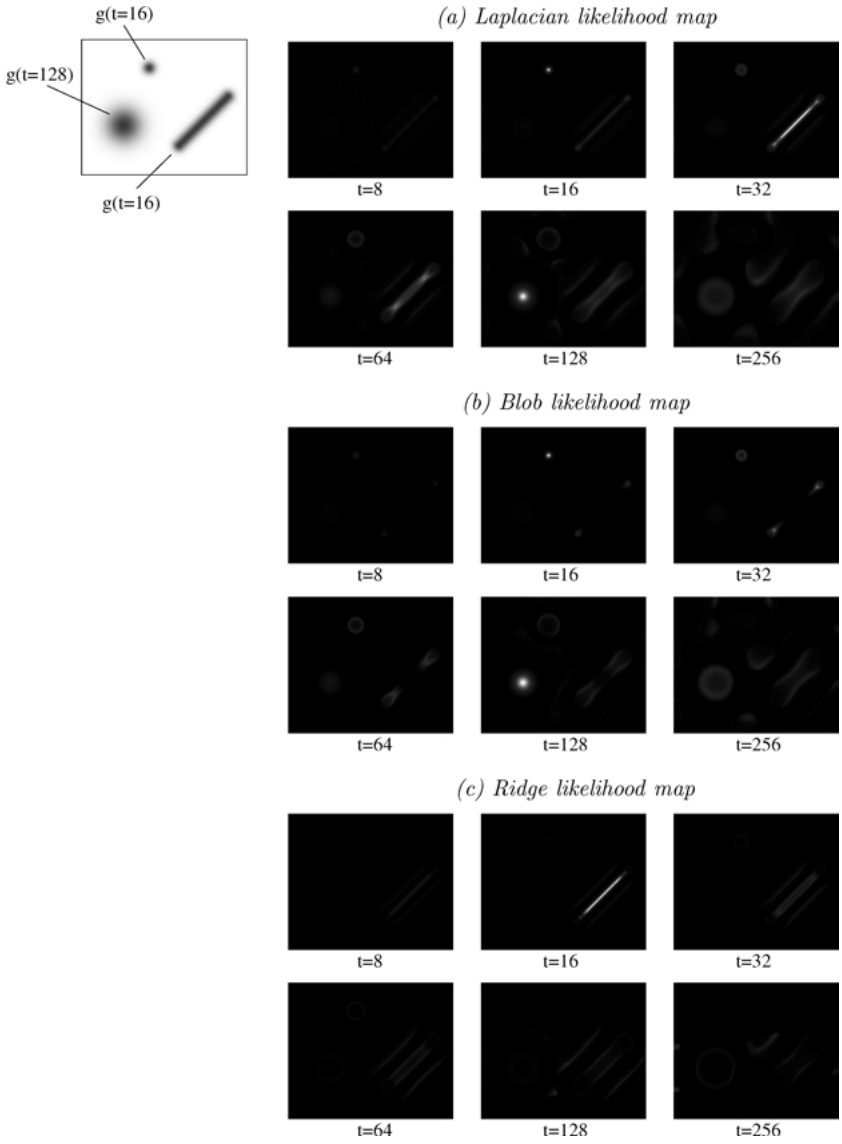

Figure 8. The result of computing feature likelihood maps for a synthetic image containing two blobs and one ridge. As can be seen, the high responses of the likelihood maps are localised at positions and scales of image structures with corresponding parameters. While the Laplacian likelihood map (top) gives high responses to both blobs and ridges, more exclusive responses to blobs and ridges are obtained with the blob likelihood map (middle) and the ridge likelihood map (bottom). Moreover, by separating the Laplacian blob likelihood map into a blob likelihood map and a ridge likelihood map, we can enforce a maximum ridge response at a scale corresponding to the width of the ridge structure.

From these results, it can be clearly seen how we by these feature likelihood maps can separate between the small structures in the background, the fingers and the palm of a hand. Moreover, despite the varying contrast of the image structures, the feature likelihood maps give equally high responses to weak ridges in the background and to the fingers of higher contrast. In many cases, this is a desirable property of a recognition system aimed at classifying local image structures irrespective of illumination variations.

5. Integration with Particle Filtering Applied to Hand Tracking and Posture Recognition

To experimentally investigate the use of the proposed distance measure and feature likelihood concepts for

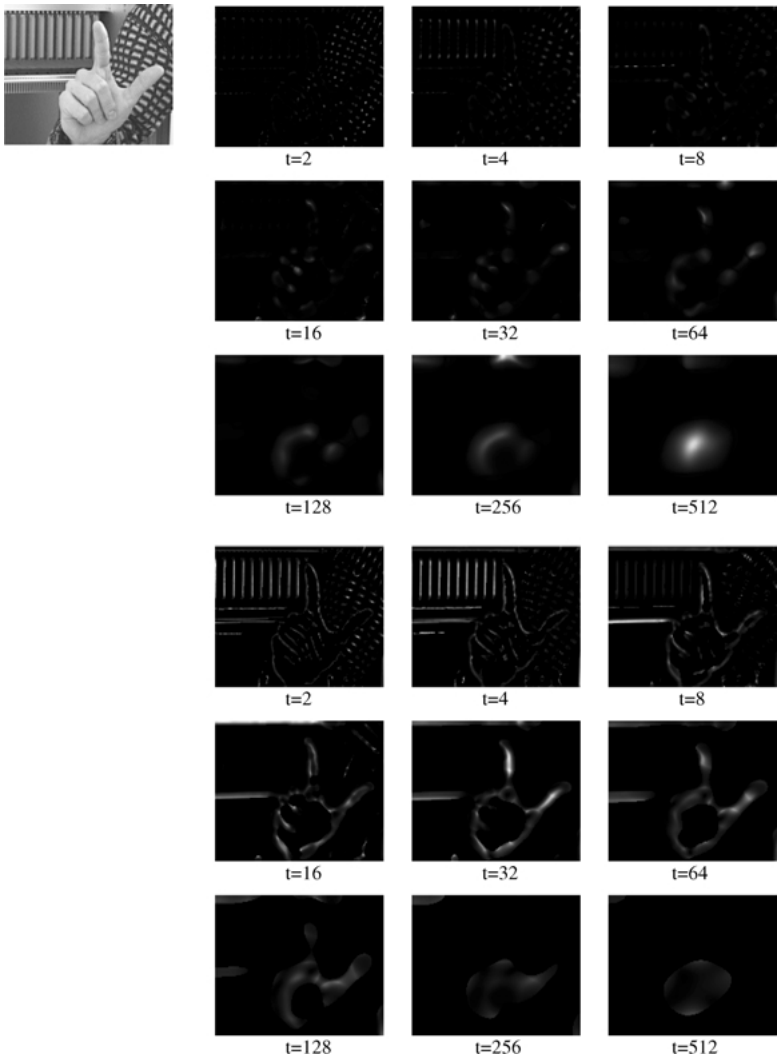

Figure 9. The result of computing the blob likelihood map (upper part) and the ridge likelihood map (lower part) for an image of a hand. (To display only responses corresponding to bright image structures, the likelihood maps have been set to zero for points with $\nabla^2 L > 0$.) Note how the fingers give rise to fine scale responses in the ridge likelihood map, and that the palm of a hand gives rise to a coarser scale response in the blob likelihood map.

model matching, we will in this section apply these notions to the problem of tracking and recognising hands in different postures.

From a statistical viewpoint, the problem of tracking and recognising hands in different postures can be formulated as the problem of estimating a vector X_0 that describes the appearance of the object model in a sequence of images. In this context, it should be noted that if one would use a unimodal approach, such as Kalman filtering, the vector X_0 would be estimated by tracking a single object hypothesis over time. Such a strategy, however, is bound to fail in situations where the hypothesis gets trapped on clutter in the background. For this reason, we shall here apply a multi-modal approach, based on particle filtering to overcome this inherent difficulty, by simultaneously tracking several object hypotheses and estimating X_0 from the probability distribution over the entire parameter space X.

Thus, following this approach, we will develop a particle filter that tracks hands over time and simultaneously recognises their postures. In particular, we shall make use of a compact hand model based on features at multiple scales, and apply the proposed distance measure and feature likelihood maps for efficient

evaluation of these models. It will be demonstrated that the resulting method will be able to track hands over complex backgrounds in real time based on pyramid implementations.

5.1. *Hierarchical and Graph-Like Hand Models*

One idea that we shall explore here is to consider relations in space and over scales between image features as an important cue for recognition. To model such relations, we shall consider graph-like object representations, where the vertices in the graph correspond to features and the edges define relations between different features.

Specifically, we shall make use of quantitative relations between features to define hierarchical, probabilistic models of objects in different states. For a hand, the feature hierarchy will contain three levels of detail; a blob corresponding to a palm at the top level, ridges corresponding to the fingers at the intermediate level and blobs corresponding to the finger-tips at the bottom level (see Fig. 10). Coarse-scale features in such a representation will enable rough localisation of hands in images, while fine-scale features such as finger tips will provide more precise localisation of hands and will increase the discriminability between the different hand poses as well as between the object and the background.

While a more general approach for modelling the internal state of a hand consists of modelling the probability distribution of the parameters over all object features, we will here simplify this task by approximating the relative scales between all the features by constant ratios and by fixing the relative positions between the ridges corresponding to the fingers and the blobs corresponding to the finger-tips. Thus, we model the global position (x, y) of the hand, its overall size s and orientation α. Moreover, we have a state parameter $l = 1, \ldots, 5$ describing the number of open fingers present in the hand posture (see Fig. 10(b)). In this way, a hand model can be parameterised by $X = (x, y, s, \alpha, l)$.

5.2. *Particle Filtering*

Particle filters aim at estimating and propagating the posterior probability distribution $p(X_t, Y_t \mid \tilde{\mathcal{I}}_t)$ over time, where X_t and Y_t are static and dynamic model parameters and $\tilde{\mathcal{I}}_t$ denotes the observations up to time t (Isard and Blake, 1996; MacCormick and Isard, 2000). Using Bayes rule, the posterior at time t can be evaluated according to

$$p(X_t, Y_t \mid \tilde{\mathcal{I}}_t) = kp(\mathcal{I}_t \mid X_t, Y_t)\, p(X_t, Y_t \mid \tilde{\mathcal{I}}_{t-1}), \tag{50}$$

where k is a normalisation constant that does not depend on variables X_t, Y_t. The term $p(\mathcal{I}_t \mid X_t, Y_t)$ denotes the likelihood that a model configuration X_t, Y_t gives rise to the image $\mathcal{I}_t$. Using a first-order Markov assumption, the dependence on observations before time $t-1$ can be removed and the model prior $p(X_t, Y_t \mid \tilde{\mathcal{I}}_{t-1})$ can be evaluated using a posterior from a previous time step and the distribution for model dynamics according to

$$\begin{aligned} &p(X_t, Y_t \mid \tilde{\mathcal{I}}_{t-1}) \\ &\quad = \int p(X_t, Y_t \mid X_{t-1}, Y_{t-1}) \\ &\qquad \times p(X_{t-1}, Y_{t-1} \mid \tilde{\mathcal{I}}_{t-1})\, dX_{t-1}\, dY_{t-1}. \end{aligned} \tag{51}$$

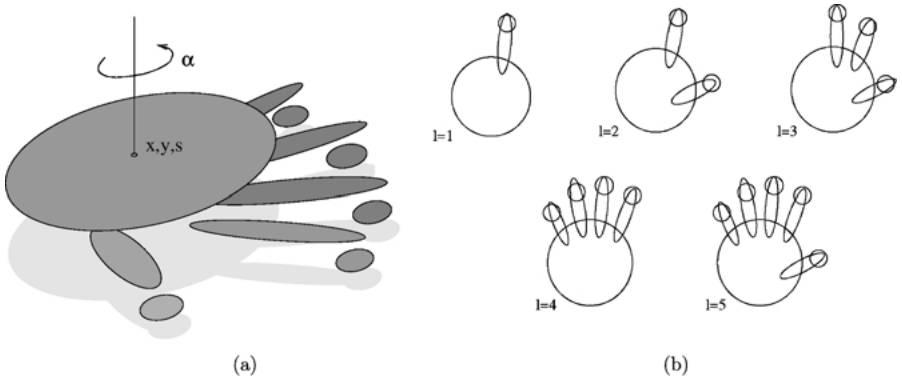

Figure 10. Model of a hand in different states: (a) hierarchical configuration of model features and their relations; (b) model states corresponding to different hand postures.

Since the likelihood function is usually multi-modal and cannot be expressed in closed form, the approach of particle filtering is to approximate the posterior distribution using N particles, weighted according to their likelihoods $p(\mathcal{I}_t \mid X_t, Y_t)$. The posterior for a new time moment is then computed by populating the particles with high weights and predicting them according to their dynamic model $p(X_t, Y_t \mid X_{t-1}, Y_{t-1})$ (Isard and Blake, 1996).

In our case, $X_t = (x, y, s, \alpha, l)$ describes the state of the hand model, while Y_t denotes the time derivatives of the first four variables, i.e., $Y_t = (\dot{x}, \dot{y}, \dot{s}, \dot{\alpha})$. Then, we assume that the likelihood $p(\mathcal{I}_t \mid X_t, Y_t)$ does not explicitly depend on Y_t, and compute $p(\mathcal{I}_t \mid X_t)$ using either the feature-based approach proposed in Section 3 or the feature likelihood maps introduced in Section 4.

5.3. Computation of Model Likelihoods

5.3.1. Feature-Based Method. For the feature based approach, the computation of the likelihood of an object model comprises the following steps: (i) detection of image features according to Section 2, (ii) computation of the approximate distance measure between the model and the data features as described in Section 3.2 and (iii) estimation of the model likelihood according to Section 3.3.

An important property of the distance measure Φ in (22) and its approximation Ψ in (26) is that they take into account both mismatches in the model and in the data. This makes it possible to perform model selection also in cases when some object models are subsets of others. To illustrate this, consider the task of matching a hand model in states with one, two and three open fingers $l = 1, 2, 3$ (see Fig. 10) to an image of a hand as shown in Fig. 1(a). If we match according to an offset criterion only (see Eq. (26)), hypotheses with one and two open fingers ($l = 1, 2$) will have the same fitting error as a hypothesis with three open fingers ($l = 3$). Thus, the offset criterion alone is not sufficient for a correct selection of hand state. To solve the problem, we must require the best hypothesis to also explain as much of the data as possible by minimising the number of mismatched data features (outlier criterion). This will result in a hypothesis that best *fits* and *explains* the data, i.e. the hypothesis with the correct state $l = 3$.

5.3.2. Filter-Based Method. For the feature likelihood based approach, the idea is to use the computed feature likelihood maps for direct lookup of feature likelihoods, which are then combined into a joint likelihood estimate for the entire model. Thus, to evaluate a blob feature $w_b(x_0, y_0, t_0)$ in an object model described by a state vector $X = (x, y, s, \alpha, l)$, we use $\mathcal{M}_{blob}$ defined in (48) and estimate the likelihood according to

$$p_{blob}(I \mid w_b) = \mathcal{M}_{blob}(x_0', y_0', s^2 t_0), \tag{52}$$

where (x_0', y_0') denotes the position of the blob feature after subjecting the original position (x_0, y_0) in the hand model to a translation by (x, y), a rotation by α and a scaling by s, as given by the state vector X

$$\begin{aligned}(x_0', y_0') = (&x + s(x_0 \cos\alpha + y_0 \sin\alpha),\\ &y + s(x_0 \sin\alpha + y_0 \cos\alpha)).\end{aligned} \tag{53}$$

Similarly, for evaluating a ridge feature $w_r(x_0, y_0, t_0, d_0, \beta_0)$, complemented by a state vector $X = (x, y, s, \alpha, l)$, we compute $\mathcal{M}_{ridge}$ defined in (49) for a set of points

$$(x_i, y_i) = (x_0', y_0') + i\,\Delta e_{(\alpha+\beta_0)} \tag{54}$$

and define the likelihood estimate as

$$p_{ridge}(I \mid w_r) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{M}_{ridge}(x_i, y_i, t), \tag{55}$$

where $m \sim d_0$ denotes the number of sampling points along the ridge of length d_0, while (x_0', y_0') again represents the transformed position of the image feature according to (53), and $e_{(\alpha+\beta_0)}$ is a unit vector that describes the orientation of a ridge feature in the hand model (with original orientation β_0), rotated by the overall orientation α of the hand model.

Finally, to equally treat models consisting of different numbers of features n and to enable model selection, we introduce a measure of the maximum admissible feature error $\varepsilon_{max} \in (0, 1)$ and define the matching score of the model as

$$p(I \mid X) \sim (1 - \varepsilon)^{N-n} \prod_{i=1}^{n} p_w(I \mid w_i), \tag{56}$$

where $p_w(I \mid w_i)$ corresponds to likelihoods of individual model features defined in (52) and (55) and N stands for the maximal number of features in all models.

Figure 11. Illustration of the effect of the colour prior. (a) Original image, (b) map of the probability of skin colour at every image point.

5.3.3. Incorporating Information about Skin Colour. When identifying humans in images, skin colour has demonstrated to be a powerful cue for identifying regions corresponding to human skin (Forsyth and Fleck, 1999). To make use of skin colour as a complementary cue to the shape information, we define an estimate $p_{skin}(I \mid w)$ of the probability that the colour at the position of a feature w corresponds to human skin and compute it from colour histograms of human hands accumulated in a training stage (see Fig. 11). Then, to obtain a likelihood estimate of a hand model X with features $w_i, i = 1...n$, we combine estimates for skin colour with the likelihood of hand shape and obtain

$$p_{hand}(I \mid X) = p_{skin}(I \mid X)\, p_{shape}(I \mid X), \quad (57)$$

where

$$p_{skin}(I \mid X) = \prod_{i=1}^{n} p_{skin}(I \mid w_i) \quad (58)$$

and $p_{shape}(I \mid X)$ is estimated either using the feature-based approach as defined in (27), i.e. $p_{shape}(I \mid X) = p(\mathcal{W}^d \mid \mathcal{W}^m)$ or from the filter-based approach as defined in (56).

5.4. *Particle Filtering Dynamics*

Concerning the dynamics $p(X_t, Y_t \mid X_{t-1}, Y_{t-1})$ of the hand model, a constant velocity model is adopted, where deviations from the constant velocity assumption are modelled by additive Brownian motion, from which the distribution $p(X_t, Y_t \mid X_{t-1}, Y_{t-1})$ is computed. To capture changes in hand postures, the state parameter l is allowed to vary randomly for 30% of the particles at each time step.

When the tracking is started, all the particles are first distributed uniformly over the parameter spaces X and Y. After each time step of particle filtering, the best hypothesis of a hand is estimated, by first choosing the most likely hand posture and then computing the mean of $p(X_t, Y_t \mid \tilde{\mathcal{I}}_t)$ for that posture (see Fig. 12). Hand posture number i is chosen if $w_i = \max_j(w_j)$, $j = 1, \ldots, 5$, where w_j is the sum of the weights of all particles with state j. Then, the continuous parameters are estimated by computing a weighted mean of all the particles in state i.

Figure 12. At every time moment, the hand tracker based on particle filtering evaluates and compares a set of object hypothesis. From these hypothesis, which represent the probability distribution of the object, the most likely object state is estimated.

5.5. *Experiments*

To investigate the proposed approaches for model evaluation, we have applied it to multistate hand tracking in a natural office environment. When using the feature-based approach, the particle filtering was performed with $N = 1000$ particles, which were evaluated on the $N^d = 200$ strongest scale-space features extracted from each image of a sequence.

Figures 13(a)–(c) show a few results from such tracking. As can be seen, the combination of particle filtering with the distance measure for hierarchical object models correctly captures changes in the position, scale and orientation of the hand as well as changes in the hand postures.

When evaluating object hypotheses based on feature likelihood maps, we have observed very similar

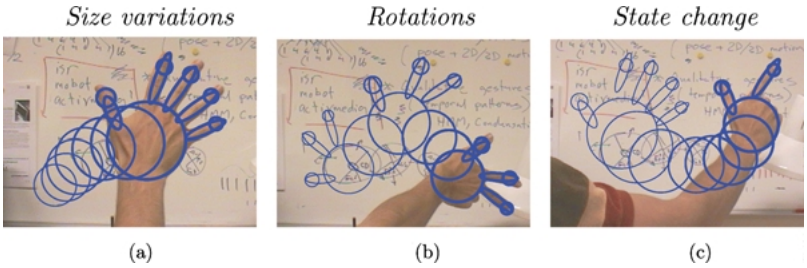

Figure 13. Result of applying the proposed framework for tracking a hand in an office environment. (a): size variations; (b): rotations; (c): a change in hand state $l : 5 \rightarrow 2$.

tracking results as in the case of distance-based evaluation. Moreover, the combination of shape and skin colour information according to (57) enables stable tracking and hand posture recognition to be performed in scenes with cluttered backgrounds as illustrated in Fig. 14.

To evaluate the stability of the hand tracker, we combined it with a simple application where captured hand motions affect a drawing device. The cursor on the screen was controlled by the position of the hand, and depending on the state of the hand, different actions could be performed. A hand posture with two fingers implied a drawing state, while a posture with one finger meant that the cursor moved without drawing. With three fingers present, the shape of the brush could be changed, while a hand posture with five fingers was used for translating, rotating and scaling the drawing. Figure 15 shows a few snapshots from such a drawing session. As can be seen from the results, the performance of the tracker is sufficient for producing a reasonable drawing.

5.6. *Implementation Details*

In practice, real-time tracking was achieved using pyramid implementations of the proposed evaluation schemes. For the feature-based approach, hybrid pyramid representations (Lindeberg et al., 2002) were generated by separable convolution with separable binomial $(1, 4, 6, 4, 1)/16$ kernels combined with spatial subsampling performed in such a way that several levels of scale were represented at the same level of resolution.

For pyramid implementation of the feature likelihood maps, the resolution at scale level t_i was obtained by sub-sampling the original image with a factor $\kappa_i = \sqrt{t_i / t_f}$, where sixteen levels of resolution with $t_i = 2^{i/2}$, $i = 1..16$ were used in total. The derivatives were computed using filter kernels of fixed scale t_f. From the experiments, we found $t_f \approx 2.0$ to be sufficiently large for obtaining a satisfactory quality of $\mathcal{M}$ on one hand, while on the other hand being sufficiently small to enable fast computations.

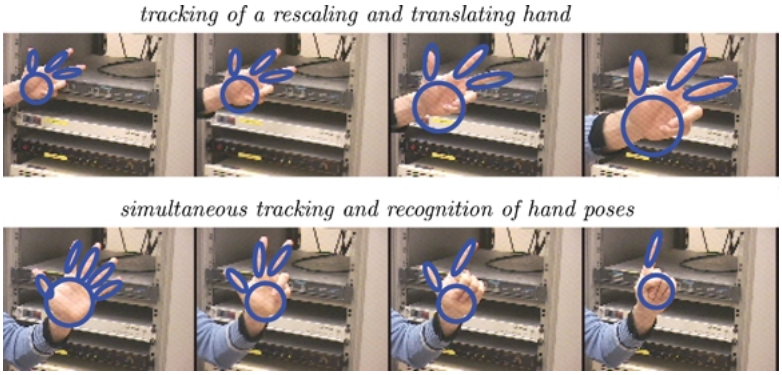

Figure 14. Results of combined hand tracking and pose recognition using particle fitering and evaluation of feature-based hand models on feature likelihood maps.

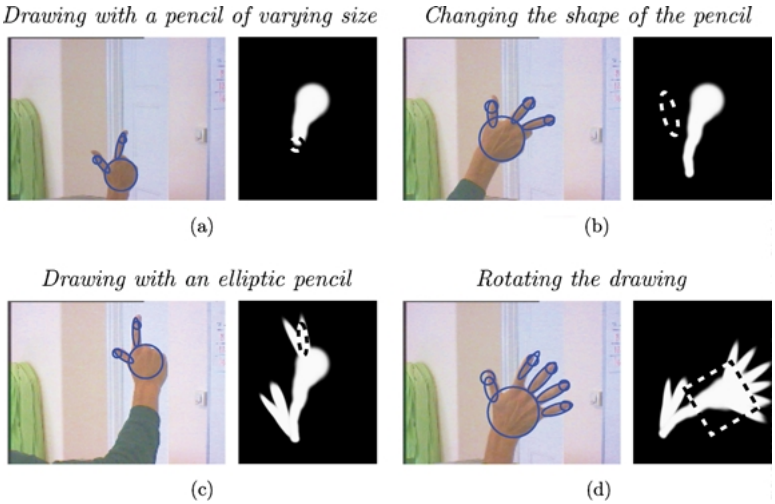

Figure 15. Illustration of the stability of the hand tracker, by letting the estimated hand motions affect a drawing device, where the position, the size and the orientation of a pencil are controlled by the corresponding parameters of the hand in the image (a), (c). In (b) the user is able to change the elliptic shape of a pencil by rotating a hand in a state with three open fingers. In (d) the drawing is scaled and rotated with a hand in a state with five open fingers.

On a modest 550 MHz Pentium III processor, our current implementation (without extensive optimisation) requires about 0.1 s for either the feature extraction step or the computation of the feature likelihood maps on a 100×100 image and about 0.04 s for performing the particle filtering using 1000 hypotheses.

6. Related Works

The subject of this paper relates to multi-scale approaches for image representation, computation of differential invariants, detection of image features as well as tracking and recognition of view-based object models. Because of the scope of these areas, it is not possible to give an extensive review, and only a few closely related works will be mentioned.

Crowley and Sanderson (1987) extracted peaks from a Laplacian pyramid of an image and linked them into a tree structure with respect to their resolution. Lindeberg (1993) constructed a scale-space primal sketch with an explicit encoding of blob-like structures in scale-space as well as the relations between these. Pizer et al. (1994) proposed the use of multi-scale medial-axis representations computed directly from image patterns distributions. Triesch and von der Malsburg (1996) used elastic graphs to represent hands in different postures with local jets of Gabor filters computed at each vertex.

Multi-scale image differential invariants (Koenderink and van Doorn, 1992; Lindeberg, 1994; Florack, 1997) have been computed by several authors, including Schmid and Mohr (1997) who apply such descriptors at interest points for image indexing and retrieval. A histogram approach based on statistics of local image descriptors has been successfully applied by Schiele and Crowley (2000) among the others.

Explicit scale selection for extraction of multi-scale image features has been investigated by Lindeberg (1998a, 1998b). Dense descriptors for estimating characteristic scales at any image point have been considered by Lindeberg (1998b), Almansa and Lindeberg (2000), Chomat et al. (2000) and Hall et al. (2000). Pattern representations by scale-invariant image features providing compact object descriptions have been used for recognition by Lowe (1999) and Mikolajczyk and Schmid (2001).

Shokoufandeh et al. (1999) detected maxima in a multi-scale wavelet transform in a way closely related to the detection of scale-space maxima. The features were then connected into an acyclic graph according to their sizes and positions. Siddiqi et al. (1999) considered representations of binary images by shock graphs and used them for object representation and recognition. The approach by Bretzner and Lindeberg (1999b) is more closely related to the one followed here, by being based on multi-scale blob and ridge features and

by defining explicit qualitative relations between these features across scales.

Other scale-space approaches with related aims have been developed by Lifshitz and Pizer (1990), Griffin et al. (1992), Gauch and Pizer (1993), Burbeck and Pizer (1995), Olsen (1997) and Vincken et al. (1997). Interesting works regarding hierarchical representations in biological vision have been presented by Riesenhuber and Poggio (1999).

With respect to object tracking, Isard and Blake (1996) developed a particle filtering approach for tracking contour-based models. Black and Jepson (1998) used eigenspace models of gray-value patterns for tracking deformable models. Integration of several image cues such as skin colour in particle filtering was proposed in Isard and Blake (1998). The efficiency of particle filters was increased using layered and partitioned sampling (Sullivan et al., 1999; MacCormick and Isard, 2000; Deutscher et al., 2000). With close relations to the notion of feature likelihood maps, Sidenbladh and Black (2001) have performed explicit learning of likelihood functions from training images and applied these to human body tracking based on ridge structures.

With regard to the area of hand gesture analysis (Pavlovic et al., 1997; Cipolla and Pentland, 1998), early work of using hand gestures for television control was presented by Freeman and Weissman (1995) using normalised correlation. Some approaches consider elaborated 3-D hand models (Rehg and Kanade, 1995), while others use colour markers to simplify feature detection (Cipolla et al., 1993) or image differencing using an updated background model (von Hardenberg and Bérard, 2001). Appearance-based models for hand tracking and sign recognition were used by Cui and Weng (1996), while Heap and Hogg (1998) and MacCormick and Isard (2000) used silhouettes of hands. A natural extension consists of integrating such edge based representations with region based representations as proposed in this work. An integration of the proposed methods into a working real-time system for human computer interaction has been presented in Bretzner et al. (2002).

7. Summary

We have introduced and applied two novel methods for evaluating multi-scale feature-based object models. Within the first, feature-based approach, we constructed a scale-invariant distance measure for comparing two sets of sparse multi-scale features. Based on this measure, an approximate likelihood function for object models was defined in such a way that maximisation of this function allows for simultaneous parameter estimation and model selection.

Whereas the computation of a distance measure requires pre-computation of all the features in the input images, the need for pre-computation can be relaxed within the second, dense filter-based, approach. Here, we proposed the concept of feature likelihood maps, which approximate the likelihood of multi-scale blob-like and ridge-like structures in image data. The computation of such maps at any point requires only local image information and can therefore be performed "on demand". Both evaluation methods are invariant with respect to common transformation of image structures including scalings, rotations translations and changes in amplitude.

When combining either of these (complementary) methods with a probabilistic approach for object tracking and object recognition, we have shown that the proposed evaluation schemes can be easily integrated with particle filtering and be used for, for example, simultaneous hand tracking and hand posture recognition. Specifically, by implementing the multi-scale image processing steps based on pyramids, both methods allow for tracking and recognition in real time.

For the chosen application, we manually constructed object models. A natural extension for future work would be to employ learning methods for automatic acquisition of multi-scale feature models and parameters for their optimal matching.

Appendix A. Ridge Strength Measures for Colour Images

This section shows how ridge strength measures previously expressed for scalar grey-level images (Lindeberg, 1998a) can be extended to vector valued colour images (according to Lindeberg and Sjöbergh, personal communication).

Consider the following measures of ridge strength, defined for a scalar image L at a given scale t:

$$\mathcal{R}_1 L = (L_{pp} - L_{qq})^2 \tag{59}$$

$$\mathcal{R}_2 L = \left(L_{pp}^2 - L_{qq}^2\right)^2 = (\nabla^2 L)^2 \mathcal{R}_1 L \tag{60}$$

where L_{pp} and L_{qq} denote the second-order directional derivatives in the main eigendirections $\bar{p}$ and $\bar{q}$ of the Hessian matrix $\mathcal{H}L$. For any orthogonal coordinate system (r, s), the directional derivative operators along the

coordinate directions can be expressed as

$$\partial_{\bar{r}} = \cos\alpha\,\partial_x + \sin\alpha\,\partial_y, \quad \partial_{\bar{s}} = -\sin\alpha\,\partial_x + \cos\alpha\,\partial_y, \tag{61}$$

where α denotes the orientation of the coordinate system relative to a Cartesian frame. After a few algebraic manipulations, it can be shown that following differential entities, inspired by the functional forms of (59) and (60),

$$\mathcal{R}_{1,(r,s)}L = (L_{rr} - L_{ss})^2 \tag{62}$$

$$\mathcal{R}_{2,(r,s)}L = \left(L_{rr}^2 - L_{ss}^2\right)^2 \tag{63}$$

can be expressed as follows when parameterised as function of the orientation α of the (r, s)-system

$$\begin{aligned}\mathcal{R}_{1,(r,s)}L &= \frac{1}{2}\left((L_{xx} - L_{yy})^2 + 4L_{xy}^2\right) \\ &\quad + \frac{1}{2}\left((L_{xx} - L_{yy})^2 - 4L_{xy}^2\right)\cos 4\alpha \\ &\quad + 2L_{xy}(L_{xx} - L_{yy})\sin 4\alpha \\ \mathcal{R}_{2,(r,s)}L &= (L_{xx} + L_{yy})^2\mathcal{R}_{1,(r,s)}L\end{aligned} \tag{64}$$

A special property of the principal (p, q)-coordinate system aligned to the eigendirections of the Hessian matrix $\mathcal{H}L$, is that it maximises the entities $\mathcal{R}_{1,(r,s)}L$ and $\mathcal{R}_{2,(r,s)}L$ over all orientations α of the (r, s)-system. In the following, we shall make use of this property when extending the ridge strength measures (59) and (60) from scalar grey-level images to vector-valued colour images. By summing up these expression over a set of colour channels $C = (L^{(1)}, L^{(2)}, L^{(3)})^T$

$$\mathcal{R}_1 C = \sum_{L\in C} \mathcal{R}_1 L \mathrel{\hat{=}} \sum_{i=1}^{3} \mathcal{R}_1 L^{(i)} \tag{65}$$

$$\mathcal{R}_2 C = \sum_{L\in C} \mathcal{R}_2 L \mathrel{\hat{=}} \sum_{i=1}^{3} \mathcal{R}_2 L^{(i)} \tag{66}$$

and introducing the descriptors (R_1, S_1, T_1) and (R_2, S_2, T_2), respectively, according to

$$R_1 = \sum_{L\in C} (L_{xx} - L_{yy})^2 \tag{67}$$

$$S_1 = \sum_{L\in C} 4L_{xy}^2 \tag{68}$$

$$T_1 = \sum_{L\in C} L_{xy}(L_{xx} - L_{yy}) \tag{69}$$

$$R_2 = \sum_{L\in C} (L_{xx} + L_{yy})^2(L_{xx} - L_{yy})^2 \tag{70}$$

$$S_2 = \sum_{L\in C} 4(L_{xx} + L_{yy})^2 L_{xy}^2 \tag{71}$$

$$T_2 = \sum_{L\in C} (L_{xx} + L_{yy})^2 L_{xy}(L_{xx} - L_{yy}) \tag{72}$$

it follows that $\mathcal{R}_1 C$ and $\mathcal{R}_2 C$ can be written as

$$\mathcal{R}_1 C = \frac{1}{2}(R_1 - S_1) + \frac{R_1 - S_1}{2}\cos 4\alpha + 2T_1 \sin 4\alpha \tag{73}$$

$$\mathcal{R}_2 C = \frac{1}{2}(R_2 - S_2) + \frac{R_2 - S_2}{2}\cos 4\alpha + 2T_2 \sin 4\alpha \tag{74}$$

To determine the direction α that maximises this expression, we differentiate this expression with respect to α and set the derivative to zero, which gives

$$\tan 4\alpha = \frac{4T}{R - S} \tag{75}$$

and

$$\cos 4\alpha = \frac{R - S}{\sqrt{16T^2 + (R - S)^2}} \tag{76}$$

$$\sin 4\alpha = \frac{4T}{\sqrt{16T^2 + (R - S)^2}} \tag{77}$$

Insertion into (73) or (74) finally results in

$$\mathcal{R}C = \frac{1}{2}(R + S + \sqrt{(R - S)^2 + 16T^2}) \tag{78}$$

where the triple (R, S, T) should be replaced by either (R_1, S_1, T_1) or (R_2, S_2, T_2) depending on which of the ridge strength measures (59) or (60) is being used. In practical experiments, it has often been the case that the approximation $\mathcal{R}C \approx R + S$ gives satisfactory performance.

Appendix B. Shape Estimation

In this section, we will show how the shape of an anisotropic Gaussian blob with covariance matrix Σ_0

$$g(x, \Sigma_0) = \frac{1}{2\pi\sqrt{\det(\Sigma_0)}} \exp\left(-\frac{1}{2}x^T \Sigma_0^{-1} x\right)$$

can be estimated from a second moment matrix as defined in (10). For simplicity, let us first assume that

Σ_0 is diagonal, i.e. $\Sigma_0 = \mathrm{diag}(a, b)$. From the semi-group property of the Gaussian kernel, it follows that the isotropic scale-space representation of $g(x, \Sigma_0)$ is $L(x, \Sigma_0, t) = g(x, \Sigma_0 + tI)$. By explicit evaluation of the definition of the second moment matrix (10), it can after a few calculations be shown that

$$\begin{aligned}\nu &= \begin{pmatrix} \nu_{11} & \nu_{12} \\ \nu_{12} & \nu_{22} \end{pmatrix} \\ &= \int_{x \in \mathbb{R}^2} (\nabla L(x; \Sigma_0, t_l))(\nabla L(x; \Sigma_0, t_l))^T g(x; t_i I)\, dx \\ &= \frac{t_i}{4\pi^2 \sqrt{(a + t_l)(b + t_l)(a + t_l + 2t_i)(b + t_l + 2t_i)}} \\ &\quad \times \begin{pmatrix} \frac{1}{(a+t_l)(a+t_l+2t_i)} & 0 \\ 0 & \frac{1}{(b+t_l)(b+t_l+2t_i)} \end{pmatrix} \end{aligned} \tag{79}$$

where we can directly read off the eigenvalues of ν as $\lambda_1 = \nu_{11}$ and $\lambda_2 = \nu_{22}$. From this expression one can observe that unless the local scale t_l is zero and the integration scale t_i is infinite, the ratio between the eigenvalues will be affected by the use of non-infinitesimal local scale and finite integration scale. One way of improving the accuracy in the estimate is to use shape adaptation of the Gaussian kernels as described in Lindeberg and Gårding (1997) and Mikolajczyk and Schmid (2002).

An alternative way of computing an estimate of the shape of the underlying Gaussian blob is by solving for b in terms of a, λ_1 and λ_2. Assuming that the integration scale is proportional to the local scale $t_i = c^2 t_l$, one obtains

$$b = \sqrt{t_l^2 c^4 + \frac{\lambda_1}{\lambda_2}\left(a^2 + 2at_l(1 + c^2) + t_l^2(1 + 2c^2)\right)} - t_l - c^2 t_l. \tag{80}$$

From the facts that (i) the eigenvalues of a second moment matrix are preserved under rotations of the image domain, and (ii) the eigenvectors follow this rotation (Lindeberg, 1994), it follows that this expression is also valid for a non-diagonal covariance matrix parameterised as $\Sigma_1 = R\Sigma_0 R^T$, where R denotes the rotation.

Acknowledgments

This paper is based on two presentations at Scale-Space'01 held in Vancouver, Canada, July 2001 (Laptev and Lindeberg, 2001a, 2001b).

The support from the Swedish Research Council for Engineering Sciences, TFR, and from the Royal Swedish Academy of Sciences as well as the Knut and Alice Wallenberg Foundation is gratefully acknowledged. We also thank Lars Bretzner for many valuable suggestions concerning this work and for his help in setting up the experiments.

Notes

1. One reason for using $\gamma = 1$ for blob detection is that then for a circular Gaussian blob with variance t_0, the selected scale will be $\hat{t} = t_0$. Similarly, when using $\gamma = 3/4$ for ridge detection, the selected scale for a Gaussian ridge with variance t_0 will be $\hat{t} = t_0$.
2. Other measures for comparing distributions, such as the KL-divergence or the $\mathcal{X}^2$-divergence, could also be interesting alternatives. Here, we have chosen to use square difference due to its symmetric property and simplicity.
3. The fact that Φ is a distance measure follows from a similar way of reasoning as for ϕ.

References

Almansa, A. and Lindeberg, T. 2000. Fingerprint enhancement by shape adaptation of scale-space operators with automatic scale-selection. *IEEE Transactions on Image Processing*, 9(12):2027–2042.

Bevensee, R. 1993. *Maximum Entropy Solutions to Scientific Problems*. Prentice Hall: Englewood Cliffs, NJ.

Bigün, J., Granlund, G.H., and Wiklund, J. 1991. Multidimensional orientation estimation with applications to texture analysis and optical flow. *IEEE Trans. Pattern Analysis and Machine Intell.*, 13(8):775–790.

Billmeyer, F. and Saltzman, M. 1982. *Principles of Colour Technology*. John Wiley and Sons: New York.

Black, M. and Jepson, A. 1998. Eigen tracking: Robust matching and tracking of articulated objects using view-based representation. *International Journal of Computer Vision*, 26(1):63–84.

Bretzner, L. and Lindeberg, T. 1999a. Qualitative multi-scale feature hierarchies for object tracking. *Journal of Visual Communication and Image Representation*, 11:115–129.

Bretzner, L. and Lindeberg, T. 1999b. Qualitative multi-scale feature hierarchies for object tracking. In *Proc. 2nd International Conference on Scale-Space Theories in Computer Vision*, O.F.O.M. Nielsen, P. Johansen, and J. Weickert (Eds.), vol. 1682, Springer Verlag, Corfu, Greece, pp. 117–128.

Bretzner, L., Laptev, I., and Lindeberg, T. 2002. Hand-gesture recognition using multi-scale colour features hierarchical features and particle filtering. *Proc. Face and Gesture*, Washington, DC, USA, pp. 63–74.

Burbeck, C.A. and Pizer, S.M. 1995. Object representation by cores: Identifying and representing primitive spatial regions. *Vision Research*, 35(13):1917–1930.

Chomat, O., de Verdiere, V., Hall, D., and Crowley, J. 2000. Local scale selection for Gaussian based description techniques. In *Proc. Sixth European Conference on Computer Vision*, vol. 1842

of Lecture Notes in Computer Science. Springer Verlag: Berlin, pp. 117–133.

Cipolla, R. and Pentland, A. (Eds.) 1998. *Computer Vision for Human-Computer Interaction*. Cambridge University Press: Cambridge, UK.

Cipolla, R., Okamoto, Y., and Kuno, Y. 1993. Robust structure from motion using motion parallax. In *Proc. Fourth International Conference on Computer Vision*, Berlin, Germany, pp. 374–382.

Crowley, J. and Sanderson, A. 1987. Multiple resolution representation and probabilistic matching of 2-D gray-scale shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(1):113–121.

Cui, Y. and Weng, J. 1996. View-based hand segmentation and hand-sequence recognition with complex backgrounds. In *Proc. 13th International Conference on Pattern Recognition*, Vienna, Austria, pp. 617–621.

Deutscher, J., Blake, A., and Reid, I. 2000. Articulated body motion capture by annealed particle filtering. In *Proc. Computer Vision and Pattern Recognition*, Hilton Head, SC, pp. II:126–133.

Florack, L.M.J. 1997. *Image Structure.* Kluwer Academic Publishers: Dordrecht, The Netherlands.

Forsyth, D. and Fleck, M. 1999. Automatic detection of human nudes. *International Journal of Computer Vision*, 32(1):63–77.

Freeman, W.T. and Weissman, C.D. 1995. Television control by hand gestures. In *Proc. Int. Conf. on Face and Gesture Recognition*, Zurich, Switzerland.

Gårding, J. and Lindeberg, T. 1996. Direct computation of shape cues using scale-adapted spatial derivative operators. *Int. J. of Computer Vision*, 17(2):163–191.

Gauch, J.M. and Pizer, S.M. 1993. Multiresolution analysis of ridges and valleys in grey-scale images. *IEEE Trans. Pattern Analysis and Machine Intell.*, 15(6):635–646.

Geman, S. and Geman, D. 1984. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741.

Griffin, L.D., Colchester, A.C.F., and Robinson, G.P. 1992. Scale and segmentation of images using maximum gradient paths. *Image and Vision Computing*, 10(6):389–402.

Hall, D., de Verdiere, V., and Crowley, J. 2000. Object recognition using coloured receptive fields. In *Proc. Sixth European Conference on Computer Vision*, vol. 1842 of Lecture Notes in Computer Science. Springer Verlag: Berlin, pp. 164–177.

Heap, T. and Hogg, D. 1998. Wormholes in shape space: Tracking through discontinuous changes in shape. In *Proc. Sixth International Conference on Computer Vision*, Bombay, India, pp. 344–349.

Isard, M. and Blake, A. 1996. Contour tracking by stochastic propagation of conditional density. In *Proc. Fourth European Conference on Computer Vision*, vol. 1064 of Lecture Notes in Computer Science. Springer Verlag: Berlin, pp. I:343–356.

Isard, M. and Blake, A. 1998. ICondensation: Unifying low-level and high-level tracking in a stochastic framework. In *Proc. Fifth European Conference on Computer Vision*, H. Burkhardt and B. Neumann (Eds.), vol. 1406 of Lecture Notes in Computer Science. Springer Verlag: Berlin, pp. 893–908.

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

Koenderink, J.J. and van Doorn, A.J. 1992. Generic neighborhood operators. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(6):597–605.

Laptev, I. and Lindeberg, T. 2001a. A multi-scale feature likelihood map for direct evaluation of object hypotheses. In *Proc. Scale-Space '01*, M. Kerckhove (Ed.), vol. 2106 of Lecture Notes in Computer Science. Springer-Verlag, Vancouver, Canada, pp. 98–110.

Laptev, I. and Lindeberg, T. 2001b. Tracking of multi-state hand models using particle filtering and a hierarchy of multi-scale image features. In *Proc. Scale-Space '01*, M. Kerckhove (Ed.), vol. 2106 of Lecture Notes in Computer Science. Springer-Verlag, Vancouver, Canada, pp. 63–74.

Lifshitz, L. and Pizer, S. 1990. A multiresolution hierarchical approach to image segmentation based on intensity extrema. *IEEE Trans. Pattern Analysis and Machine Intell.*, 12(6):529–541.

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

Lindeberg, T. 1994. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers: Boston.

Lindeberg, T. 1997. On automatic selection of temporal scales in time-causal scale-space. In *AFPAC'97: Algebraic Frames for the Perception-Action Cycle*, pp. 94–113.

Lindeberg, T. 1998a. Edge detection and ridge detection with automatic scale selection. *Int. J. of Computer Vision*, 30(2):117–154.

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

Lindeberg, T. and Gårding, J. 1997. Shape-adapted smoothing in estimation of 3-D depth cues from affine distortions of local 2-D structure. *Image and Vision Computing*, 15:415–434.

Lindeberg, T., Niemenmaa, J. and Bretzner, L. 2002. Scale selection in hybrid multi-scale representations, in preparation.

Lowe, D. 1999. Object recognition from local scale-invariant features. In *Proc. Seventh International Conference on Computer Vision*, Corfu, Greece, pp. 1150–1157.

MacCormick, J. and Isard, M. 2000. Partitioned sampling, articulated objects, and interface-quality hand tracking. In *Proc. Sixth European Conference on Computer Vision*, vol. 1843 of Lecture Notes in Computer Science. Springer Verlag: Berlin, pp. II:3–19.

Mikolajczyk, K. and Schmid, C. 2001. Indexing based on scale invariant interest points. In *Proc. Eighth International Conference on Computer Vision*, Vancouver, Canada, pp. I:525–531.

Mikolajczyk, K. and Schmid, C. 2002. An affine invariant interest point detector. In *Proc. Seventh European Conference on Computer Vision*, vol. 2350 of Lecture Notes in Computer Science. Springer Verlag: Berlin, pp. I:128–142.

Olsen, O.F. 1997. Multi-scale watershed segmentation. In *Gaussian Scale-Space Theory: Proc. PhD School on Scale-Space Theory*, J. Sporring, M. Nielsen, L. Florack and P. Johansen (Eds.), Kluwer Academic Publishers, Copenhagen, Denmark, pp. 191–200.

Pavlovic, V.I. Sharma, R., and Huang, T.S. 1997. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Trans. Pattern Analysis and Machine Intell.*, 19(7):677–694.

Pizer, S.M., Burbeck, C.A., Coggins, J.M., Fritsch, D.S., and Morse, B.S. 1994. Object shape before boundary shape: Scale-space medial axis. *Journal of Mathematical Imaging and Vision*, 4:303–313.

Rao, A.R. and Schunk, B.G. 1991. Computing oriented texture fields. *CVGIP; Graphical Models and Image Processing*, 53(2):157–185.

Rehg, J. and Kanade, T. 1995. Model-based tracking of self-occluding articulated objects. In *Proc. Fifth International Conference on Computer Vision*, Cambridge, MA, pp. 612–617.

Riesenhuber, M. and Poggio, T. 1999. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2:1019–1025.

Schiele, B. and Crowley, J. 2000. Recognition without correspondence using multidimensional receptive field histograms. *Int. J. of Computer Vision*, 36(1):31–50.

Schmid, C. and Mohr, R. 1997. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–535.

Shokoufandeh, A., Marsic, I., and Dickinson, S. 1999. View-based object recognition using saliency maps. *Image and Vision Computing*, 17(5/6):445–460.

Siddiqi, K., Shokoufandeh, A., Dickinson, S., and Zucker, S. 1999. Shock graphs and shape matching. *International Journal of Computer Vision*, 35(1):13–32.

Sidenbladh, H. and Black, M. 2001. Learning image statistics for bayesian tracking. In *Proc. Eighth International Conference on Computer Vision*, Vancouver, Canada, pp. II:709–716.

Sullivan, J., Blake, A., Isard, M., and MacCormick, J. 1999. Object localization by bayesian correlation. In *Proc. Seventh International Conference on Computer Vision*, Corfu, Greece, pp. 1068–1075.

Triesch, J. and von der Malsburg, C. 1996. Robust classification of hand postures against complex background. In *Proc. Int. Conf. on Face and Gesture Recognition*, Killington, Vermont, pp. 170–175.

Vincken, K., Koster, A., and Viergever, M. 1997. Probabilistic multiscale image segmentation. *IEEE Trans. Pattern Analysis and Machine Intell.*, 19(2):109–120.

von Hardenberg, C. and Bérard, F. 2001. Bare-hand human-computer interaction. *ACM Workshop on Perceptive User Interfaces*, Orlando, FL, USA.

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