Real-Time Scale Selection in Hybrid Multi-Scale Representations[*]

Tony Lindeberg and Lars Bretzner

Computational Vision and Active Perception Laboratory (CVAP)
Department of Numerical Analysis and Computer Science
KTH, SE-100 44 Stockholm, Sweden

Abstract. Local scale information extracted from visual data in a bottom-up manner constitutes an important cue for a large number of visual tasks. This article presents a framework for how the computation of such scale descriptors can be performed in real time on a standard computer.

The proposed scale selection framework is expressed within a novel type of multi-scale representation, referred to as hybrid multi-scale representation, which aims at integrating and providing variable trade-offs between the relative advantages of pyramids and scale-space representation, in terms of computational efficiency and computational accuracy. Starting from binomial scale-space kernels of different widths, we describe a family pyramid representations, in which the regular pyramid concept and the regular scale-space representation constitute limiting cases. In particular, the steepness of the pyramid as well as the sampling density in the scale direction can be varied.

It is shown how the definition of γ-normalized derivative operators underlying the automatic scale selection mechanism can be transferred from a regular scale-space to a hybrid pyramid, and two alternative definitions are studied in detail, referred to as variance normalization and l_p-normalization. The computational accuracy of these two schemes is evaluated, and it is shown how the choice of sub-sampling rate provides a trade-off between the computational efficiency and the accuracy of the scale descriptors. Experimental evaluations are presented for both synthetic and real data. In a simplified form, this scale selection mechanism has been running for two years, in a real-time computer vision system.

1 Introduction

Recent works have shown how the notion of automatic scale selection constitutes an essential complement to traditional scale-space representation. While a scale-space representation provides a well-founded framework to represent image structures at different scales, the scale-space representation by itself contains no explicit information about what scales are relevant for further processing.

For addressing the problem of choosing interesting scale levels from image data, a number of different approaches have been developed in the literature

[*] Shortened version in L. Griffin and M. Lillholm (Eds), *Proc. Scale-Space'03*, Springer Lecture Notes in Computer Science, volume 2695, pages 148–163.

(see the review in section 2). If one aims at real-time performance, however, a common problem of most present approaches for automatic scale selection, is computational efficiency. Since scale selection is performed by either minimizing or maximizing feature measures over scales, the algorithms involve explicit search over scales. The purpose of this article is to show how these problems can be remedied for a class of scale selection methods based on normalized derivatives, and how real-time performance can be obtained on a standard PC.

2 Related work

An early approach to scale selection focused on the detection of blob-like image features and scale levels were selected from local maxima over scales of a normalized measure of blob strength (Lindeberg 1993*a*). Later, this idea was generalized to a wide class of differential image features, by selecting scale levels from local maxima over scales of differential invariants expressed in terms of normalized derivatives (Lindeberg 1993*b*, Lindeberg 1994). This principle has been applied to various problems relating to the detection of image features (Lindeberg 1998*b*, Lindeberg 1998*a*, Chomat et al. 2000, Almansa & Lindeberg 2000, Pedersen & Nielsen 2000, Nielsen & Lillholm 2001, Kadir & Brady 2001). In particular, and motivated by the observation that single-scale ridge detection may be highly sensitive to the choice of scale level, special emphasis has been on the detection of ridges for medical image analysis (Pizer et al. 1994, Eberly et al. 1994, Koller et al. 1995, Lorenz et al. 1997, Sato et al. 1998, Staal et al. 1999, Frangi et al. 1999, Majer 2001). Moreover, for the purpose of obtaining zoom invariant image features for further processing, scale selection mechanisms have proven highly useful for interest point detection (Mikolajczyk & Schmid 2002) with applications to object recognition (Lowe 1999, Hall et al. 2000) and tracking (Bretzner & Lindeberg 1998, Laptev & Lindeberg 2001). Other approaches for scale selection have also been presented from the behaviour of entropy measures or error measures over scales (Jägersand 1995, Elder & Zucker 1996, Niessen & Maas 1996, Yacoob & Davis 1997, Lindeberg 1998*c*, Sporring & Weickert 1999, Pedersen & Nielsen 2001, Comaniciu et al. 2001, Hadjidemetriou et al. 2002).

The algorithms for automatic scale selection that will presented in this paper bear close relations to previous work by (Crowley & Parker 1984) for detecting peaks and ridges in a bandpass pyramid, as well as previous works performing scale selection in a regular scale-space representation (Lindeberg 1994, Lindeberg 1998*b*) without spatial subsampling, although reformulated to be expressed in a hybrid pyramid representation (Lindeberg 1995, Grostabussiat 1997, Niemenmaa 2001). Parallel developments of real-time algoritms for automatic scale selection are being made by (Crowley 2002) and (Lowe 2002).

3 Hybrid pyramid representation

Both pyramids (Burt & Adelson 1983, Crowley & Parker 1984, Jähne 1995, Simoncelli & Freeman 1995) and scale-space representations (Witkin 1983, Koenderink

1984, Lindeberg 1994, Florack 1997) have been developed from the idea of representing images at multiple scales in such a way that the resulting representation can be used as input to a large number of visual processes. Computationally, however, these concepts have their relative advantages and disadvantages.

A pyramid representation is highly efficient in the sense that it leads to a rapidly decreasing image size, while a scale-space representation successively becomes more redundant as the scale parameter increases. The highly discretized nature of a pyramid can, however, lead to algorithmic problems at coarse scales, while in scale-space representation the task of operating on the data will be successively simplified at coarser scales.

When processing data at a coarse scale in a scale-space representation, it thus seems natural that a certain amount of subsampling can be performed without affecting the performance too seriously. On the other hand, one could also consider decomposing the smoothing operation in a pyramid into a set of smoothing stages, so as to obtain a denser sampling along the scale direction. In this way, we obtain an *oversampled pyramid*, characterized by the fact that not every smoothing step is followed by a subsampling operation.

The goal of this section is to describe a general class of multi-scale representations, which comprises both regular pyramids, oversampled pyramids and scale-space representation as special cases. Due to space limitations, however, the presentation will sometimes be somewhat condensed. For a more extensive description, see (Lindeberg and Bretzner 2003).

3.1 Reduction operators

Following (Burt & Adelson 1983, Crowley & Parker 1984), let us describe the the transformation between two adjacent scale levels in a pyramid by a reduction operator. For simplicity, let us assume that the pyramid is separable and that the size N of the smoothing filter is odd. Then, the transformation from the representation $L^{(i)}$ at the current scale level i, to the representation $L^{(i+1)}$ at the next coarser level $i+1$ is for some set of filter coefficients $c\colon \mathbb{Z} \to \mathbb{R}$ given by

$$L^{(i+1)} = \textsc{ReduceCycle}(L^{(i)}) \tag{1}$$

$$L^{(i+1)}(x) = \sum_{n=-(N-1)/2}^{(N-1)/2} c(n)\, L^{(i)}(sx-n). \tag{2}$$

Next, let us assume that the smoothing operation can be decomposed into several smoothing steps:

$$\begin{aligned} \textsc{ReduceCycle} := {} & \textsc{SubSample} \\ & \textsc{Smooth}^{+} \end{aligned} \tag{3}$$

where the notation $\textsc{Op}^{+}$ means that several operators of the form $\textsc{Op}$ may occur. $\textsc{ReduceCycle}$ is thus composed of one or more smoothing operations followed by a subsampling. The subsampling operation is here defined by

$$S = \textsc{SubSample}(L;\ s) \tag{4}$$

$$S(x) = L(sx) \qquad (s \in \mathbb{Z}_+). \tag{5}$$

(where we usually choose $s = 2$) and each smoothing step according to

$$S = \textsc{Smooth}(L) \tag{6}$$

$$S(x) = \sum_{n=-N}^{N} c(n)\, L(x-n). \tag{7}$$

For simplicity, let us assume that the coefficients of the smoothing operation originate from a discretization of the diffusion operator repeated K times

$$\textsc{Smooth}(L) = \textsc{DeltaSmooth}(L;\ \Delta t, K) = [\textsc{DeltaSmooth}(L;\ \Delta t, 1)]^K \tag{8}$$

where in one dimension the $\textsc{DeltaSmooth}(L;\ \Delta t, 1)$ operator corresponds to convolution with a binomial diffusion filter of the following form

$$T = \textsc{DeltaSmooth}(L;\ \Delta t, 1) \tag{9}$$

$$T(x) = \frac{\Delta t}{2}\, L(x-1) + (1-\Delta t)\, L(x) + \frac{\Delta t}{2}\, L(x+1) \tag{10}$$

Thus, we can construct kernels such as the binomial three-kernel

$$\textsc{Bin3Kernel} = \textsc{DeltaSmooth}(\cdot;\ \tfrac{1}{2}, 1) = (\frac{1}{4},\quad \frac{1}{2},\quad \frac{1}{4}) \tag{11}$$

and the binomial five-kernel

$$\textsc{Bin5Kernel} = \textsc{DeltaSmooth}(\cdot;\ \tfrac{1}{2}, 2) = (\frac{1}{16},\quad \frac{4}{16},\quad \frac{6}{16},\quad \frac{4}{16},\quad \frac{1}{16}). \tag{12}$$

Moreover, we can define different types of oversampled pyramid representations as illustrated in figure 1. To index the levels in such a hybrid representations, we shall henceforth use the index $i \in [1 \dots I]$ for the subsampling levels and the index $j \in [1 \dots J]$ within each subsampling level.

3.2 Equivalent convolution and derivative approximation kernels

Since the representation at each level is constructed from a set of repeated smoothing and subsampling operations, which are all linear operations, the composed operation can equivalently be modeled as the result of applying one kernel $C^{(i,j)}$, termed *equivalent convolution kernel*, to the original image, followed by a pure subsampling step. If we define a dual operator to the ReduceCycle operator according to

$$\textsc{ExpandCycle} := \begin{array}{l}\textsc{Smooth}^{+}\\ \textsc{Enlarge}\end{array}$$

$$\textsc{Bin3ReduceCycle} := \begin{array}{l}\textsc{SubSample}\\ \textsc{Bin3Kernel}\end{array} \qquad \textsc{Bin5ReduceCycle} := \begin{array}{l}\textsc{SubSample}\\ \textsc{Bin5Kernel}\end{array}$$

$$\textsc{Bin3Reduce6Cycle} := \begin{array}{l}\textsc{SubSample}\\ \textsc{Bin3Kernel}\\ \textsc{Bin3Kernel}\\ \textsc{Bin3Kernel}\\ \textsc{Bin3Kernel}\\ \textsc{Bin3Kernel}\\ \textsc{Bin3Kernel}\end{array} \qquad \textsc{Bin5Reduce3Cycle} := \begin{array}{l}\textsc{SubSample}\\ \textsc{Bin5Kernel}\\ \textsc{Bin5Kernel}\\ \textsc{Bin5Kernel}\end{array}$$

Fig. 1: Examples of regular and oversampled pyramids as generated using the notation for hybrid multi-scale representations defined in (3)–(12). By applying these reduction cycles repeatedly, we obtain pyramids that will be referred to as Bin3Pyramid, Bin5Pyramid, Bin3(6)Pyramid and Bin5(3)Pyramid, respectively.

where the Enlarge operation enlarges any D-dimensional image by a factor s

$$E = \textsc{Enlarge}(L) \tag{13}$$

$$E(x) = \begin{cases} s^D L(x/s) & \text{if all indices in } x \text{ are multiples of } s \\ 0 & \text{if any index in } x \text{ is not a multiple of } s \end{cases} \tag{14}$$

the equivalent convolution kernel corresponding to level (i, j) can be written

$$C^{(i,j)} = \textsc{ExpandAll}(\delta^{(i,j)}) \tag{15}$$

where $\delta^{(i,j)}$ is a discrete delta function at level (i, j) and ExpandAll denotes the ExpandCycle operators corresponding to the set of all the ReduceCycle

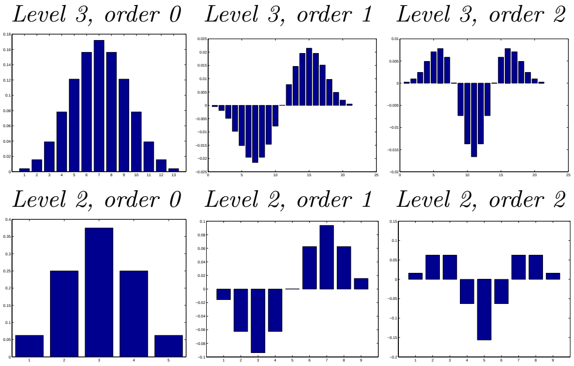


Fig. 2: Examples of equivalent convolution kernels and equivalent derivative approximation kernels for the Bin5Pyramid derived from the Bin5ReduceCycle in figure 1.

operators used for reaching this level. Similarly derivative approximations are computed by taking the grid spacing h at the current into explicit account

$$\partial_{x^r} \approx \mathcal{D}_{x^r} = \frac{1}{h^{|r|}}\, \delta_{x^r}, \tag{16}$$

at any level with resolution h in the pyramid, the corresponding *equivalent derivative approximation kernel* is given by

$$C_{x^r}^{(i,j)} = \textsc{ExpandAll}(\delta_{x^r}^{(i,j)}) \tag{17}$$

where higher dimensional difference approximations $\delta_{x^r} = \delta_{x_1{}^{r_1}} \delta_{x_2{}^{r_2}} .. \delta_{x_D{}^{r_D}}$ are expressed in terms of the one-dimensional rth order difference operator

$$\delta_{x^r} = \begin{cases} (\delta_{xx})^{r/2} & \text{if } r \text{ is even} \\ \delta_x\, \delta_{x^{r-1}} & \text{if } r \text{ is odd} \end{cases} \tag{18}$$

and δ_x and δ_{xx} denote the first-order symmetric difference operators with computational molecules $(-\frac{1}{2},\, 0,\, \frac{1}{2})$ and $(1,\, -2,\, 1)$, respectively (see figure 2).

3.3 Measuring the scale parameter and the subsampling rate

For measuring the scale parameter $t_{(i,j)}$ at any level (i, j) in a hybrid pyramid, we will start from *the covariance matrix of the equivalent convolution kernel*:

$$t_{(i,j)} = (\det V(C^{(i,j)}))^{1/D} = (\det V(\textsc{Expand}(\delta^{(i,j)})))^{1/D} \tag{19}$$

where $V(C)$ represents the spatial covariance matrix of a kernel C and D is the dimension of the signal. At coarser levels of resolution with grid spacing $h \in \mathbb{Z}_+$, the operator $\textsc{DeltaSmooth}(L;\ \mathit{\Delta} t, K)$ in (8) corresponds to scale values at levels k and $k+1$ that are related according to $t^{(i,j+1)} - t^{(i,j)} = K\, h^2\, \mathit{\Delta} t$.

Table 1 shows the scale values for each level computed in this way for some of the pure and oversampled pyramids defined in figure 1.

Bin3Pyramid	Bin5Pyramid	Bin5(3)Pyramid		
0.0	0.0	0.0	1.0	2.0
0.5	1.0	3.0	7.0	11.0
2.5	5.0	15.0	31.0	47.0
10.5	21.0	63.0	127.0	191.0
42.5	85.0	255.0	511.0	767.0
170.5	341.0	1023.0	2047.0	3071.0

Table 1: Scale values for the different levels of two pure and one oversampled pyramid as defined in figure 1.

Then, to describe how the grid spacing h depends on the scale parameter t in a hybrid pyramid let us introduce a *subsampling factor* ρ from the relation

$$h_{max} = \rho\, \sigma = \rho \sqrt{t} \tag{20}$$

where for reasons of computational efficiency we define the actual grid spacing as the maximum power of two that does not exceed this upper bound

$$h(t, \rho) = \begin{cases} \max_{h'=2^{i-1}\,:\, i \in \mathbb{Z}_+ \backslash \{0\}} h' : h' < h_{max}(t, \rho) \text{ if } h_{max} \geq 1 \\ 1 \hfill \text{otherwise} \end{cases} \tag{21}$$

Thus, a subsampling factor of $\rho = 0$ corresponds to preserving the original resolution at all levels of scales, while increasing values of ρ correspond to higher degrees of subsampling at coarser scales.

In this context, self-similarity over scales (implying that $h \leq \rho\,\sqrt{t}$ holds with equality at the lowest pyramid level for any amount of subsampling) is obtained only if we precede the computation of the pyramid by a certain amount of pre-smoothing. If the total amount of smoothing in the composed SMOOTH$^+$ stage between two sub-sampling stages in (3) corresponds to a variance of $h^2 \varDelta t_{cycle}$, where for hybrid pyramids generated according to (8) and (9) we have $h^2 \varDelta t_{cycle} = h^2\, J\, K\, \varDelta t$, then it can be shown that the requirement of self-similarity over scales implies that the pre-smoothing t_{start} (i.e the scale of the first level) and the sub-sampling factor ρ must be given by

$$\rho = \sqrt{\frac{3}{\varDelta t_{cycle}}}, \quad t_{start} = \frac{\varDelta t_{cycle}}{3} \tag{22}$$

Table 2 shows values of ρ and t_{start} computed in this way for a few pyramids. In addition, this table also lists a measure of the average sampling density in the scale direction defined as $d_{mean} = (\tau(t^{(i+1,1)}) - \tau(t^{(i,1)}))/J$ where $\tau(t) = \log_2(t)$.

Pyramid	$t^{(i,j+1)} - t^{(i,j)}$	Levels	ρ	t_{start}	d_{mean}
BIN3PYRAMID	$h^2/2$	1	$\sqrt{6}$	1/6	2
BIN5PYRAMID	h^2	1	$\sqrt{3}$	1/3	2
BIN5(3)PYRAMID	h^2	3	1	1	2/3
BIN5(6)PYRAMID	h^2	6	$1/\sqrt{2}$	2	1/3

Table 2: The subsampling rate ρ and the amount of pre-smoothing t_{start} for a few self-similar pyramids.

4 Scale selection in hybrid multi-scale representation

Our next goal is to express a scale selection mechanism within a hybrid pyramid representation. In previous works, it has been shown that a powerful principle for automatic scale selection consists of selecting interesting scale levels from the scales at which (possibly non-linear) combinations of γ-*normalized derivatives*

$$\partial_{\xi_i} = t^{\gamma/2}\,\partial_{x_i}, \tag{23}$$

assume local maxima over scales (see section 2). Intuitively, this corresponds to selecting scale levels at which the normalized feature response is locally strongest.

General scale invariance property. A basic property of this scale selection method is as follows: If $\mathcal{D}(L)$ is a homogeneous differential expression, and if a local maximum of a signal f is detected at scale t_{locmax}, then under a rescaling of f by a factor s, this local maximum over scale is transferred to the scale level $s^2 t_{locmax}$.

Interpretation in terms of L_p-norms. With respect to the computation of derivatives of the scale-space representation, it can be shown that γ-normalization corresponds to normalizing the corresponding γ-normalized Gaussian derivative operators $g_{\xi^m}(\cdot;\ t) = t^{m\gamma/2} g_{x^m}(\cdot;\ t)$ to constant L_p-norms

$$\|g_{\xi^m}(\cdot;\ t)\|_p = \left(\int_{x \in \mathbb{R}^D} |g_{\xi^m}(\cdot;\ t)|^p dx \right)^{1/p} \tag{24}$$

over scales, where the parameter p in the L_p-norm is related to the parameter γ in the γ-normalized derivative concept according to

$$p = \frac{1}{1 + \frac{m}{D}(1-\gamma)}, \tag{25}$$

where m is the order of differentiation and D denotes the dimension of the signal. Specifically, $\gamma = 1$ corresponds to $p = 1$ and thus to L_1-normalization of all the Gaussian derivative kernels.

4.1 Defining normalized derivatives with spatial subsampling

For transferring this notion of γ-normalized derivatives from a scale-space representation to a hybrid pyramid, our next goal is to define normalization parameters γ_r such that normalized derivative approximations can be written:

$$\mathcal{D}_{x^r,norm} = \gamma_r \, \mathcal{D}_{x^r}. \tag{26}$$

Here, two approaches will be considered and evaluated:

- *variance-based normalization:* multiplying the equivalent derivative approximation kernel (17) at any level in the pyramid by the variance (19) of the equivalent convolution kernel at the corresponding level

$$\gamma_{r,var} = \left(t^{(i,j)} \right)^{|r|/2} = \left(\det(V(C^{(i,j)}))^{1/D} \right)^{|r|/2} \tag{27}$$

- *l_p-normalization:* requiring the l_p-norm of the normalized equivalent derivative approximation kernel to be equal to the L_p-norm of the corresponding Gaussian derivative operator $\partial_{\xi^r} g(x;\ t)$

$$\gamma_{r,l_1} \|C_{x^r}^{(i,j)}\|_p = \|\partial_{\xi^r} g(x;\ t)\|_p \tag{28}$$

Experiments: Scale-space signatures for Gaussian blobs. For a rotationally symmetric Gaussian blob with variance t_0 in two dimensions $f(x, y) = g(x, y;\ t_0)$ it can be shown that the evolution over scales of the γ-normalized Laplacian response at the center of the blob is in the case when $\gamma = 1$ given by

$$(\nabla^2_{norm} L)(0, 0;\ t) = t\,(\partial_{xx} + \partial_{yy}) L(0, 0;\ t) = -\frac{t}{\pi(t_0 + t)^2} \tag{29}$$

and there is a unique maximum over scales in $-(\nabla^2_{norm} L)(0, 0;\ t)$ at $t = t_0$.

Figure 3 shows a few examples of such scale-space signatures computed for Gaussian blobs of different sizes, using a separable Bin3(6)Pyramid with an initial pre-smoothing stage. As can be seen from these graphs, l_p-normalization (stars) gives a closer approximation of the continuous behaviour (the solid curve) than variance-based normalization (crosses). Moreover, for variance-based normalization there are a number of "kinks" in the graph at the scales where subsamplings occur. In these respects, l_p-normalization has clear advantages compared to variance-based normalization.

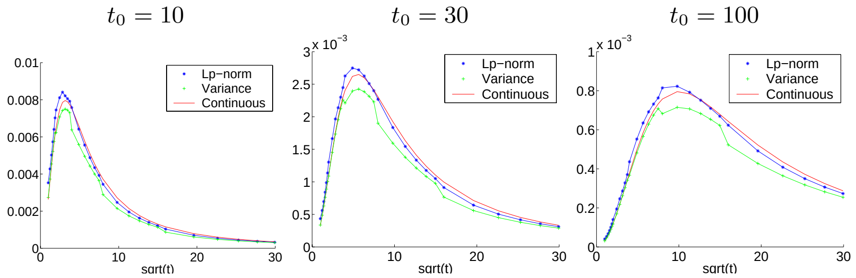


Fig. 3: Scale-space signatures of the (negative) normalized Laplacian response for rotationally symmetric Gaussian blobs with variances $t_0 = 10$, $t_0 = 30$ and $t_0 = 100$, respectively, computed using a separable Bin3(6)Pyramid in two dimensions using l_p-normalization (stars) and variance-based normalization (crosses). For reference, the corresponding continuous behaviour is shown as well (solid curve).

4.2 Detecting scale-space maxima

A method for complementary scale selection and detection of interest points consists of simultaneously maximizing differential entities over both space and scale. If $\mathcal{D}_{space} L$ denotes the differential entity used for spatial selection and if $\mathcal{D}_{scale,norm} L$ is the γ-normalized differential entity used for scale selection, such *interest points with automatic scale selection* can be characterized by

$$\begin{cases} \nabla(\mathcal{D}_{space} L) = 0 \\ \mathcal{H}(\mathcal{D}_{space} L) \text{ negative definite} \\ \partial_t(\mathcal{D}_{scale,norm} L) = 0 \\ \partial_{tt}(\mathcal{D}_{scale,norm} L) \leq 0 \end{cases} \tag{30}$$

where $\mathcal{H}(\mathcal{D}_{space}L)$ denotes the Hessian of $\mathcal{D}_{space}L$. In the special case when $\mathcal{D}_{space}L = \mathcal{D}_{scale,norm}L$ such points are referred to as *scale-space maxima* of $\mathcal{D}_{scale,norm}L$. Our next goal is to investigate how the performance of a blob detector with automatic scale selection depends on the choice of normalization method as well as the subsampling rate ρ in the pyramid.

To quantify the difference between these two normalization approaches, 1000 Gaussian images were generated containing one blob each with random variance between $t_0 = 10$ and $t_0 = 100$ and at a random position within a central 128×128 window in the image. The global maximum over scales of the normalized Laplacian response in the hybrid pyramid representation was detected, and a quadratic interpolation over scales was performed to estimate the scale $\hat{t}$ of the peak in the scale-space signature. The relative error in the estimate was computed

$$\varepsilon_n = \log_2 \left(\frac{\hat{t}_n}{t_{0,n}} \right) \tag{31}$$

and the performance was measured in terms of the following descriptors

$$\varepsilon_{mean} = \frac{1}{N} \sum_{n=1}^{N} \varepsilon_n, \qquad \varepsilon_{spread} = \sqrt{\frac{\sum_{n=1}^{N} \varepsilon_n^2}{N}} \tag{32}$$

where N is the number of blobs. These error measures were then transformed into relative error factors measured in dimension length $\sigma = \sqrt{t}$ according to

$$r_{mean} = \sqrt{2^{\varepsilon_{mean}}}, \qquad r_{spread} = \sqrt{2^{\varepsilon_{spread}}} \tag{33}$$

where the ideal case corresponds to $r_{mean} = 1$ and $r_{spread} = 1$. In addition, the absolute error in the estimated position $(\hat{x}, \hat{y})$ was measured as $\delta = \sqrt{(\hat{x} - x_0)^2 + (\hat{y} - y_0)^2}$ and a relative error measure in relation to the scale level $\sigma_0 = \sqrt{t_0}$ was defined as $\delta_{rel} = \delta / \sigma_0$. This procedure was repeated for different types of separable two-dimensional pyramids as shown in tables 3–4.

As can be seen from the results, there is a substantial variation in the accuracy of the estimate local maximum over scales depending on the type of pyramid — the oversampled Bin3(6)Pyramid and the Bin5(3)Pyramid perform significantly better than the regular Bin3Pyramid and the Bin5Pyramid, and further improvement is obtained if we increase the amount of oversampling by using a Bin3(12)Pyramid or a Bin5(6)Pyramid. In all of these cases, l_p-normalization leads to better performance measures than variance-based normalization. For this reason, we will henceforth prefer l_p-normalization.

Concerning the spatial localization error, we can see how the error decreases as we increase the degree of oversampling in the hybrid pyramid, by decreasing ρ and h_{max}. For the Bin3(6)Pyramid, the Bin5(3)Pyramid, the Bin3(12)Pyramid and the Bin5(6)Pyramid, the average error in all cases corresponds to a fraction of a pixel, and true sub-pixel accuracy is obtained for these data.

Pyramid type	l_p-normalization		variance-based	
	r_{mean}	r_{spread}	r_{mean}	r_{spread}
Bin3Pyramid	0.65	1.61	0.62	1.70
Bin5Pyramid	0.78	1.34	0.77	1.36
Bin3(6)Pyramid	0.93	1.11	0.93	1.15
Bin5(3)Pyramid	0.93	1.12	0.92	1.15
Bin3(12)Pyramid	0.96	1.08	0.95	1.13
Bin5(6)Pyramid	0.94	1.10	0.94	1.13

Table 3: Performance of the scale selection method when performing simultaneous spatial and scale selection based on scale-space maxima of the normalized Laplacian response using different types of hybrid multi-scale representations and either l_p-normalization or variance-based normalization.

Pyramid type	l_p-normalization		variance-based	
	δ	δ_{rel}	δ	δ_{rel}
Bin3Pyramid	1.86	0.32	1.76	0.29
Bin5Pyramid	1.21	0.21	1.21	0.21
Bin3(6)Pyramid	0.18	0.03	0.05	0.01
Bin5(3)Pyramid	0.19	0.03	0.07	0.01
Bin3(12)Pyramid	0.05	0.01	0.03	0.00
Bin5(6)Pyramid	0.05	0.01	0.02	0.00

Table 4: Measures of the spatial localization error when performing simultaneous spatial and scale selection based on scale-space maxima of the normalized Laplacian response using different types of hybrid multi-scale representations and either l_p-normalization or variance-based normalization.

4.3 Post-processing the scale-space maxima from a hybrid pyramid

While the previous results show that scale-space maxima can be detected in a hybrid pyramid using conceptually very clean operations, there is a minor complication with the previous approach. From the quantitative measure r_{mean} shown in table 3, it can be seen that there is a certain bias in the scale selection procedure that leads to an average underestimate of the scale estimate by 4 to 7 % for the sample types of oversampled hybrid pyramid representations that have been evaluated here.

When analysing the image data in more detail, it can be observed that a major reason for this scale bias is due to the detection of local maxima when translational invariance has been violated by the subsampling step. If the position of the original blob is far away from the closest grid point at the scale levels around the scale level t_0 at which it would be detected without spatial subsampling, the magnitude of the normalized Laplacian at the available grid points at the desired scale level $t_k \approx t_0$ may be significantly smaller than they would have been without spatial subsampling. As a result of this, the values of the normalized Laplacian at lower scale levels may be higher (since the grid sampling there is denser), which in turn means that a lower scale level is selected than in the ideal case without spatial subsampling.

To reduce this problem, an additional post-processing stage is applied: If a scale-space maximum is detected at a scale level where the next coarser scale level is at lower resolution, then a computation of image values at (one level of) finer resolution is initiated in a spatial 3×3 neighbourhood around the scale space maximum at this pyramid level. If the magnitude of the normalized differential entity is greater at this scale, then the scale-space maximum is translated to this nearest coarser scale level. Moreover, a tri-quadratic interpolation is performed in a $3 \times 3 \times 3$ neighbourhood in space and scale to estimate the position and the scale of the scale-space maximum with subpixel accuracy.

Table 5 shows the results obtained by adding these two post-processing stages to the previously methodology. As can be seen from a comparison with table 3, for the Bin5(3)Pyramid and the Bin5(6)Pyramid the average bias in the scale estimate is reduced by basically one order of magnitude, from 6–7 % to 0.4–0.6 %. Moreover, the measure r_{spread} of the spread in the scale values is reduced from 10–12 % to 1–3 %.

Pyramid type	l_p-normalization		variance-normalization	
	r_{mean}	r_{spread}	r_{mean}	r_{spread}
Bin5Pyramid	1.196	1.250	1.182	1.239
Bin5(3)Pyramid	1.006	1.032	0.999	1.180
Bin5(6)Pyramid	0.996	1.019	0.999	1.082

Table 5: Performance of the scale selection method when adding extended coarser scale level search and triquadratic interpolation to the previously developed method for performing simultaneous spatial and scale selection based on scale-space maxima of the normalized Laplacian response (see table 3). The numerical values show the mean r_{mean} and the spread r_{spread} of the relative error according to (31) for 1000 Gaussian blobs with random variances between $t_0 = 10$ and $t_0 = 100$.

5 Trade-off: Computational efficiency vs. accuracy

From the experiments on blob detection with automatic scale selection, we have seen how decreasing the value of ρ improves the accuracy of the results. On the other hand, increasing ρ improves the computational efficiency, since fewer grid points are computed. Thus, the hybrid pyramid concept allows us to obtain different trade-offs between computational efficiency vs. accuracy by varying ρ.

To quantify this trade-off, we started out by measuring the computational efficiency in the following way: For a given image size of 384*288 pixels, a threshold on the magnitude of the blob response was determined such that around 500 blobs would be detected between $t_{min} = 4$ and $t_{max} = 2000$ in a Bin5(6)Pyramid. Keeping this threshold fixed, blobs were then detected using the BIN5PYRAMID, BIN5(2)PYRAMID, ... BIN5(5)PYRAMID. A similar experiment was performed using a lower threshold on the blob response, determined in such a way that about 1000 blobs would be obtained in the BIN5(6)PYRAMID. Table 6 shows the computation time for detecting scale-space extrema in this way, with and without using the additional localization stage described in section 4.3. To allow for comparison, a denser estimation of the scale and localization errors for Gaussian blob detection was also performed for the same types of pyramids and using the methodology described in section 4.2 — see table 7.

Pyramid type	ρ	500 blobs		1000 blobs	
		det	det+loc	det	det+loc
BIN5PYRAMID	1.73	16	32	17	45
BIN5(2)PYRAMID	1.22	23	51	25	79
BIN5(3)PYRAMID	1.00	39	66	43	97
BIN5(4)PYRAMID	0.87	55	89	63	127
BIN5(5)PYRAMID	0.77	72	105	81	153
BIN5(6)PYRAMID	0.71	88	121	101	173

Table 6: Computation times (in ms) for blob detection in different hybrid pyramids with and without the additional post-processing stage for scale localization. The timings have been performed on a 2.4 GHz DELL PC with a Pentium 4 processor.

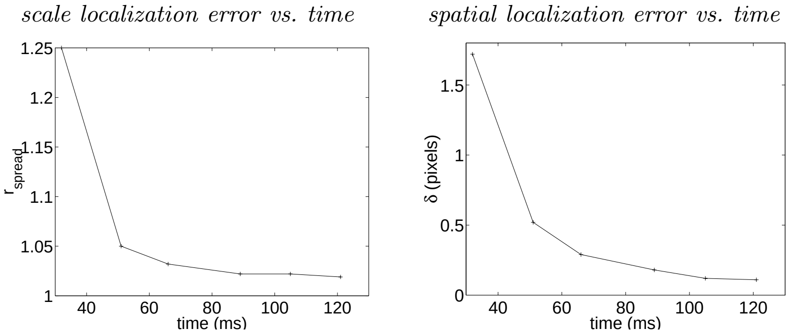
Pyramid type	ρ	δ (pixels)	r_{spread}
BIN5PYRAMID	1.73	1.72	1.250
BIN5(2)PYRAMID	1.22	0.52	1.050
BIN5(3)PYRAMID	1.00	0.29	1.032
BIN5(4)PYRAMID	0.87	0.18	1.022
BIN5(5)PYRAMID	0.77	0.12	1.022
BIN5(6)PYRAMID	0.71	0.11	1.019

Table 7: The spatial and scale localization errors for different subsampling factors ρ using l_p-normalization. The experiments were performed on 1000 Gaussian blobs with random position and random variances between 10 and 100.

If we regard these measures as representative indicators of the computational effort and the computational accuracy in the scale estimates, we thus obtain the trade-off curves in figure 4 for how ρ affects r_{spread} and the computation time.

6 Stability of the scale descriptors

In addition to the abovementioned quantitative experiments on synthetic data with ground truth, it is of particular interest to investigate the stability of the scale descriptors on real-world images. To investigate this, we performed the


scale localization error vs. time

spatial localization error vs. time

Fig. 4: Trade-offs between the localization error (vertical axis) and the computation time (horizontal axis) for hybrid pyramids with different values of ρ: (left) scale localization error, (right) spatial localization error.

following experiment: An image sequence was taken for a set of uniformly spaced distances to an object. In each image, blob detection was performed by detecting scale-space extrema of the normalized Laplacian response in a Bin5(6)-pyramid using l_p-normalization. Five scale-space maxima were selected manually in the first frame, and these features were matched over time as illustrated in figure 5.

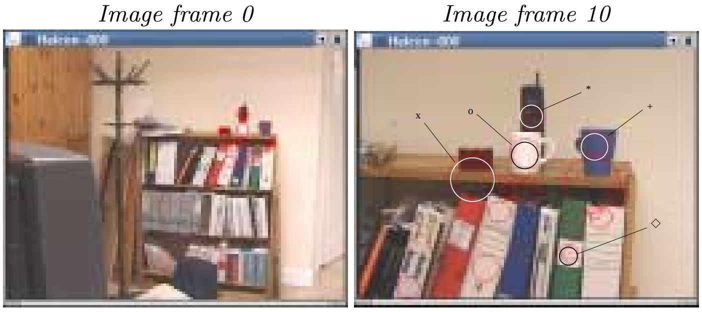


Image frame 0

Image frame 10

Fig. 5: Two out of eleven images in an image sequence used for testing the stability of the scale descriptors over time. In each image, a set of detected image features is indicated, out of which a subset has been matched over time and been used for measuring variations in scale levels over time. In the last image, five scale-space maxima used for scale measurements have been marked by corresponding symbols used in figure 6.

For each one of these five features, a straight line of the form $\frac{1}{\sqrt{t}} = A\tau + B$ was fit to the data (with τ denoting time), and the time to collision was estimated by extrapolating the line to $\tau \to \infty$ (see figure 6). Here, the mean value of the five different estimates of the time to collision was 14.89 time units and the standard deviation 0.30 time units. Considering that these estimates are based on measurements at single points in scale-space, the results show how scale descriptors computed from a hybrid multi-scale representation can be stable enough to be used as a visual cue in its own right.

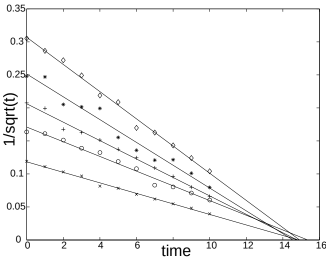


Fig. 6: Graph showing the variation over time of $1/\sqrt{t}$ for five image features matched over time as a camera approaches an object with uniform velocity.

7 Summary and discussion

We have presented a general framework for defining subsampled multi-scale representations in such a way that the theory comprises both traditional pyramid representations and discrete scale-space as limiting cases. Regular pyramids arise as a special case when we have only one scale level between any pair of successive subsampling stages (*i.e.* a reduction cycle with $J = 1$), while a regular discrete scale-space representation is obtained as the limiting case if we let the scale increment Δt in the diffusion smoothing operator tend to zero, while keeping the product of $J\Delta t$ constant and equal to the maximum scale level t_{max} that needs to be accessed. Since this family of multi-scale representations provides a way to express different trade-offs between the relative advantages of pyramids and scale-space representation, we refer to it as hybrid multi-scale representations.

Then, we presented a theory for how scale selection mechanisms based on the maximization over scales of γ-normalized derivatives can be expressed within this family of subsampled multi-scale representations. Two ways of defining normalized derivatives in the presence of spatial subsampling have been studied, and it has been shown that the approach referred to as l_p-normalization performs significantly better than the possibly more straightforward approach of variance-based normalization. Specifically, we have quantified how the steepness of a hybrid representation, parameterized by the subsampling rate ρ, allows us to obtain different trade-offs between computational accuracy as enabled by dense sampling and computational efficiency as promoted by sparse sampling.

We have also shown how the scale descriptors computed from a hybrid multi-scale representation are stable enough to be used as a cue in its own right. Combined with a multi-scale tracking and recognition method described elsewhere (Laptev & Lindeberg 2001), an integrated real-time computer vision based on a simplified hybrid pyramid has been presented in (Bretzner et al. 2002).

Acknowledgments

The support from the Swedish Research Council, the Royal Swedish Academy of Sciences and the Knut and Alice Wallenberg Foundation is gratefully acknowledged.

References

Almansa, A. & 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.

Bretzner, L., Laptev, I. & Lindeberg, T. (2002), Hand-gesture recognition using multi-scale colour features, hierarchical features and particle filtering, *Face and Gesture'02*, 63–74.

Bretzner, L. & Lindeberg, T. (1998), 'Feature tracking with automatic selection of spatial scales', *Computer Vision and Image Understanding* **71**(3), 385–392.

Burt, P. J. & Adelson, E. H. (1983), 'The Laplacian pyramid as a compact image code', *IEEE Trans. Comm.* **9:4**, 532–540.

Chomat, O., de Verdiere, V., Hall, D. & Crowley, J. (2000), Local scale selection for Gaussian based description techniques, *ECCV'00*, Springer LNCS 1842, 117–133.

Comaniciu, D., Ramesh, V. & Meer, P. (2001), The variable bandwidth mean shift and data-driven scale selection, *ICCV'01*, 438–445.

Crowley, J. L. & Parker, A. C. (1984), 'A representation for shape based on peaks and ridges in the Difference of Low-Pass Transform', *IEEE-PAMI* **6**(2), 156–170.

Crowley, J. L. (2002), Personal communication.

Eberly, D., Gardner, R., Morse, B., Pizer, S. & Scharlach, C. (1994), 'Ridges for image analysis', *J. Math. Im. Vis.* **4**(4), 353–373.

Elder, J. H. & Zucker, S. W. (1996), Local scale control for edge detection and blur estimation, *in* 'ECCV'96', 57–69..

Florack, L. M. J. (1997), *Image Structure*, Kluwer, Netherlands.

Frangi, A. F., Niessen, W. J., Hoogeveen, R. M., van Walsum, T. & Viergever, M. A. (1999), Quantitation of vessel morphology from 3D MRI, '*MICCAI*, 358–367.

Grostabussiat, P. (1997), On hybrid multi-scale representations, Licentiate thesis, KTH, Stockholm, Sweden.

Hadjidemetriou, E., Grossberg, M. D. & Nayar, S. K. (2002), Resolution selection using generalized entropies of multiresolution histograms, *ECCV'02*, Springer LNCS 2350, 220–235.

Hall, D., de Verdiere, V. & Crowley, J. (2000), Object recognition using coloured receptive fields, *ECCV'00*, Springer LNCS 1842, 164–177.

Jägersand, M. (1995), Saliency maps and attention selection in scale and spatial coordinates: An information theoretic approach, *ICCV'95*, 195–202.

Jähne, B. (1995), *Digital Image Processing*, Springer-Verlag.

Kadir, T. & Brady, M. (2001), 'Saliency, scale and image description', *IJCV* **45**, 83–105.

Koenderink, J. J. (1984), 'The structure of images', *Biol. Cyb.* **50**, 363–370.

Koller, T. M., Gerig, G., Szèkely, G. & Dettwiler, D. (1995), Multiscale detection of curvilinear structures in 2-D and 3-D image data, *ICCV'95*, 864–869.

Laptev, I. & Lindeberg, T. (2001), Tracking of multi-state hand models using particle filtering and a hierarchy of multi-scale image features, *Scale-Space'01*, Springer LNCS 2106, 63–74.

Lindeberg, T. (1993*a*), 'Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention', *IJCV* **11**(3), 283–318.

Lindeberg, T. (1993*b*), On scale selection for differential operators, *SCIA'93*, 857–866.

Lindeberg, T. (1994), *Scale-Space Theory in Computer Vision*, Kluwer, Netherlands.

Lindeberg, T. (1995), unpublished manuscript on scale selection in hybrid multi-scale representations.

Lindeberg, T. (1998*a*), 'Edge detection and ridge detection with automatic scale selection', *IJCV* **30**(2), 117–154.

Lindeberg, T. (1998*b*), 'Feature detection with automatic scale selection', *IJCV* **30**(2), 77–116.

Lindeberg, T. (1998*c*), 'A scale selection principle for estimating image deformations', *Image and Vision Computing* **16**(14), 961–977.

Lindeberg, T. & Bretzner, L (2003), Real-time scale selection in hybrid multi-scale representations, Technical report, KTH, Stockholm, Sweden.

Lorenz, C., Carlsen, I.-C., Buzug, T. M., Fassnacht, C. & Weese, J. (1997), Multi-scale line segmentation with automatic estimation of width contrast and tangential direction in 2D and 3D medical images, *CVRMed-MRCAS'97*, Springer LNCS 1205, 233–242.

Lowe, D. (1999), Object recognition from local scale-invariant features, *ICCV'99*, 1150–1157.

Lowe, D. (2002), Personal communication.

Majer, P. (2001), The influence of the γ-parameter on feature detection with automatic scale selection, *Scale-Space'01*, Springer LNCS 2106, 245–254.

Mikolajczyk, K. & Schmid, C. (2002), An affine invariant interest point detector, *ECCV'02*, Springer LNCS 2350, 128–142.

Nielsen, M. & Lillholm, M. (2001), What do features tell about images, *Scale-Space'01*, Springer LNCS 2106, 39–50.

Niemenmaa, J. (2001), Feature detection in images with the pyramid representation, MSc thesis, KTH, Stockholm, Sweden.

Niessen, W. & Maas, R. (1996), Optic flow and stereo, *in* J. Sporring et al (eds) *Gaussian Scale-Space Theory*, Kluwer.

Pedersen, K. S. & Nielsen, M. (2000), 'The Hausdorff dimension and scale-space normalisation of natural images', *J. Math. Im. Vis.* **11**(2), 266–277.

Pedersen, K. S. & Nielsen, M. (2001), Computing optic flow by scale-space integration of normal flow, *Scale-Space'01*, Springer LNCS 2106, 14–25.

Pizer, S. M., Burbeck, C. A., Coggins, J. M., Fritsch, D. S. & Morse, B. S. (1994), 'Object shape before boundary shape: Scale-space medial axis', *J. Math. Im. Vis.* **4**, 303–313.

Sato, Y., Nakajima, S., Shiraga, N., Atsumi, H., Yoshida, S., Koller, T., Gerig, G. & Kikinis, R. (1998), '3D multi-scale line filter for segmentation and visualization of curvilinear structures in medical images', *Medical Image Analysis* **2**(2), 143–168.

Simoncelli, E. P. & Freeman, W. T. (1995), 'The steerable pyramid: A flexible architecture for multi-scale derivative computation', *ICIP'95*, 444–447.

Sporring, J. & Weickert, J. A. (1999), 'Information measures in scale-spaces', *IEEE-IT* **45**(3), 1051–1058.

Staal, J., Kalitzin, S., ter Haar Romeny, B. & Viergever, M. (1999), Detection of critical structures in scale-space, *Scale-Space'99*, Springer LNCS 1682, 105–116.

Witkin, A. P. (1983), Scale-space filtering, *8th IJCAI*, pp. 1019–1022.

Yacoob, Y. & Davis, L. S. (1997), Estimating image motion using temporal multi-scale models of flow and acceleration. In: *Motion-Based Recognition*, Kluwer.