Bayesian object localisation in images

J. Sullivan, A. Blake*, M. Isard and J. MacCormick
Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, UK.
Web: http://www.robots.ox.ac.uk/∼vdg/

In Proc Int. J. Computer Vision, 44(2), 111–135, 2001.

Abstract

A Bayesian approach to intensity-based object localisation is presented that employs a learned probabilistic model of image filter-bank output, applied via Monte Carlo methods, to escape the inefficiency of exhaustive search.

An adequate probabilistic account of image data requires intensities both in the foreground (ie over the object), and in the background, to be modelled. Some previous approaches to object localisation by Monte Carlo methods have used models which, we claim, do not fully address the issue of the statistical independence of image intensities. It is addressed here by applying to each image a bank of filters whose outputs are approximately statistically independent. Distributions of the responses of individual filters, over foreground and background, are learned from training data. These distributions are then used to define a joint distribution for the output of the filter bank, conditioned on object configuration, and this serves as an observation likelihood for use in probabilistic inference about localisation.

The effectiveness of probabilistic object localisation in image clutter, using Bayesian Localisation, is illustrated. Because it is a Monte Carlo method, it produces not simply a single estimate of object configuration, but an entire sample from the posterior distribution for the configuration. This makes sequential inference of configuration possible. Two examples are illustrated here: coarse to fine scale inference, and propagation of configuration estimates over time, in image sequences.

1 Introduction

The paper develops a Bayesian approach to localising objects in images. Approximate probabilistic inference of object location is done using a learned likelihood for the output of a bank of image filters. The new approach is termed *Bayesian Localisation*[1]

Following the framework of "pattern theory" (Grenander, 1981; Mumford, 1996), an image is an intensity function $I(\mathbf{x}),\ \mathbf{x} \in \mathcal{D} \subset \mathcal{R}^2$, taken to contain a template $T(\mathbf{x})$ that has undergone certain distortions. Much

*Current address: Microsoft Research, 1 Guildhall Street, Cambridge, UK

[1]Previously (Sullivan et al., 1999) we have referred to the new approach as "Bayesian Correlation", but have since been persuaded that this is a somewhat misleading term.

of the distortion is accounted for as a warp of the template $T(\mathbf{x})$ into an intermediate image $\tilde{I}$ by an (inverse) warp mapping g_X:

$$T(\mathbf{x}) = \tilde{I}(g_X(\mathbf{x})),\ \mathbf{x} \in S, \tag{1}$$

where S is the domain of T, and g_X is parameterised by $X \in \mathcal{X}$ over some configuration space $\mathcal{X}$ (for instance planar affine warps). The remainder of the distortion in the process of image formation, is taken to have the form of a random process applied pointwise to intensity values in $\tilde{I}$, to produce the final image I:

$$I(\mathbf{x}) = f(\tilde{I}(\mathbf{x}), \mathbf{x}, w(\mathbf{x})), \tag{2}$$

where w is a noise process and f is a function that may be nonlinear. Note that (2) may include a component of sensor noise but in practice, this is emphatically not its principal role. Camera sensor noise is negligible compared with the principal source of variability that needs to be modelled probabilistically: illumination changes, and the residual variability between objects of a given class that is unmodelled otherwise.

Analysis "by synthesis" then consists of the Bayesian construction of a posterior distribution for X. That is, given a prior distribution [2] $p_0(X)$ for the configuration X, and an observation likelihood $L(X) = p(Z|X)$ where $Z \equiv Z(I)$ is some finite-dimensional representation of the image I, the posterior density for X is given by

$$p(X|Z) \propto p_0(X)p(Z|X). \tag{3}$$

In the straightforward case of normal distributions, (3) can be computed in closed form, and this can be effective in the fusion of visual data (Matthies et al., 1989; Szeliski, 1990). In the non-Gaussian cases commonly arising, for example in image clutter or with multiple models, sampling methods are effective (Geman and Geman, 1984; Gelfand and Smith, 1990; Grenander et al., 1991), and that is what we use here.

There have been a number of powerful demonstrations in the pattern theory genre, especially in the field of face analysis (Cootes et al., 1995; Beymer and Poggio, 1995; Vetter and Poggio, 1996) and in biological images (Grenander and Miller, 1994; Storvik, 1994; Ripley, 1992). A great attraction of pattern theoretic algorithms is that they can potentially generate not just a single estimate of object configuration, but an entire probability distribution. This facilitates sequential inference, across spatial scales, across time for image sequence analysis, and even across sensory modalities.

The previous work most closely related to Bayesian localisation is as follows. First Grenander et al. (1991)use randomly generated diffeomorphisms as a mechanism for Bayesian inference of contour shape. Its drawback is that it treats the intensities of individual, neighbouring pixels as independent which leads to unrealistic observation likelihood models. Second, the algorithm of Viola and Wells (1995) for registration by maximisation of mutual information contains the key elements of probabilistic modelling and learning of foreground, but does not take account of background statistics. It computes a single estimate of object pose, rather than sampling the entire distribution of the posterior. Thirdly, Geman and Jedynak (1996) use probabilistic foreground/background learning for road tracking but compute only a single estimate of pose rather than sampling from the posterior; furthermore, the statistical independence of observations, which is a necessary assumption of the method, is not investigated. Attributes of these and other important prior work are summarised in table 1, in terms of elements of Bayesian Localisation as follows.

[2] The problem of how to obtain the prior p_0 is a much debated issue for Bayesian inference in general which is entirely outside the scope of this paper. We simply adopt the common line of developing a methodology in which the role of the prior is at any rate explicit.

IB Intensity Based observations, not just edges.

FL Foreground Learning in terms of probability distributions estimated from one or more training examples.

MS Multiple Scale search is well known to be a sound basis for efficient image-search.

PD Posterior Distributions are generated, rather than just single estimates, facilitating sequential reasoning for image sequence analysis, and potentially across sensory modalities.

BM Background Modelling: in a valid Bayesian analysis, image observations Z must be regarded as fixed, not as a function $Z(X)$ of a hypothesis X. For example, a sum-squared difference measure violates this principle by considering only the portion of an image directly under a given template $T(\mathbf{x})$. In contrast, in a Bayesian approach, evidence about where the object is *not* must be taken into account, and that requires a probabilistic model of the image background.

SI Statistical Independence of observations must be understood if constructed observation likelihoods are to be valid.

	IB	FL	MS	PD	BM	SI	**Comments**
(Burt, 1983)	×		×				multi-scale pyramid
(Witkin et al., 1987; Scharstein and Szeliski, 1998)	×		×				scale-space matching
(Grenander et al., 1991; Ripley, 1992)	×			×	×		random diffeomorphisms
(Viola and Wells, 1993)	×	×					mutual information
(Cootes et al., 1995)	×	×	×				multi-scale active contours
(Black and Yacoob, 1995), (Bascle and Deriche, 1995), (Hager and Toyama, 1996)	×	×					affine flow/warp
(Isard and Blake, 1996)		×		×			random, time-varying active contours
(Olshausen and Field, 1996; Bell and Sejnowski, 1997)	×				×	×	independent components (ICA)
(Geman and Jedynak, 1996)	×	×			×		response learning

Table 1: **Precursors to Bayesian Localisation.**

2 Bayesian framework

2.1 Image observations

Image observations can be based on edges or on intensities (and a combination of the two can be particularly effective (Bascle and Deriche, 1995)). Edges are attractive because of their superior invariance to variations in illumination and other perturbations, but true Bayesian inference (3) with edges is not feasible. This is

because, given a set Z of all edges in an image, there is no known construction for the observation density $p(Z|X)$ that is probabilistically consistent. One feasible approach allows Z to be a function of X, so that $Z(X)$ consists solely of those edges found close to the outline of the object, in configuration X. Then a likelihood $L(X) = p(Z(X)|X)$ can be constructed (Isard and Blake, 1998), but cannot be used for true Bayesian inference as that demands that the observations Z must be fixed, not a function of X. The alternative approach followed here avoids the problem encountered with edges by using a fixed set of intensities covering the entire image. turns out that Bayesian localisation subsumes the need for explicit edge features, because its probabilistic model of intensity naturally captures foreground/background transitions.

2.2 Sum-squared difference and cross-correlation

One approach to interpreting image intensities probabilistically is to make the very special assumption that image distortions are due to additive white noise. Then, a likelihood

$$L(X) = \exp -\Sigma(X) \tag{4}$$

can be defined (Szeliski, 1990) in terms of a sum-squared difference (SSD) function $\Sigma(X)$:

$$\Sigma(X) = \int_{\mathbf{x}\in S} w(\mathbf{x}) \left(T(\mathbf{x}) - I(g_X(\mathbf{x}))\right)^2 , \tag{5}$$

where the weighting $w(\mathbf{x})$ depends on the noise variance. It is worth noting that a likelihood such as (4) is generally multi-modal, having many maxima. Ingenious algorithms (Witkin et al., 1987; Scharstein and Szeliski, 1998) have been needed to find maximum likelihood estimates. Multi-modality is a feature of image likelihood functions generally, whether based on edges or intensities, and is the reason for needing random sampling methods later in this paper.

The likelihood (4) has been used successfully in surface reconstruction (Szeliski, 1990) but is not appropriate for image intensity modelling, for two reasons. The first is that the assumption of additive, white noise is not plausible. It implies statistical independence of adjacent pixels. In practice however, the sources of intensity variation are illumination changes and intrinsic variability between objects of one class. Such variations are spatially correlated (Belhumeur and Kriegman, 1998). If a fine-scale independence assumption is made nonetheless, the resulting likelihood function $L(X)$ can have grossly exaggerated variations (Ripley, 1992), even as great as several hundred orders of magnitude, for minor perturbations of X.

The second reason is that the SSD-based likelihood (5) $L(X)$ depends on the image intensities over a domain $g_X(S)$ that varies with X. This means, effectively, that the observation likelihood is $L(X) = p(Z(X)|X)$, depending on observations $Z(X)$ which are not fixed. This was precisely the problem with edge-based observations which we set out to put right! The problem can be rectified by insisting that observations Z are computed as some fixed function of an image $I(\mathbf{x}),\ \mathbf{x} \in D$, where D is a fixed domain, irrespective of X. The domain D will then be the union of a foreground region $g_X(S) \cap D$, and a background region $D \backslash \{g_X(S)\}$. Any consistently constructed likelihood $p(Z|X)$ must therefore depend both on the foreground and on the statistics of the background. The intuition behind this is that the image contains statistical information both about where the object *is* and where it *is not*. A complete Bayesian theory must take account of both sources of information.

2.3 Filter bank

If assuming independence of adjacent pixels is unreasonable, then some alternative representation of the image I is needed whose elements are either mutually independent or have known statistical dependence. We have opted to seek a set $Z = (z_1, \ldots, z_K)$ of observations, in the form of a bank

$$z_k = \int_{S_k} W_k(\mathbf{x}) I(\mathbf{x}) \mathrm{d}\mathbf{x} \tag{6}$$

of filters W_k, with supports S_k arranged on a regular grid, as in figure 1. The task now is to find a filter bank

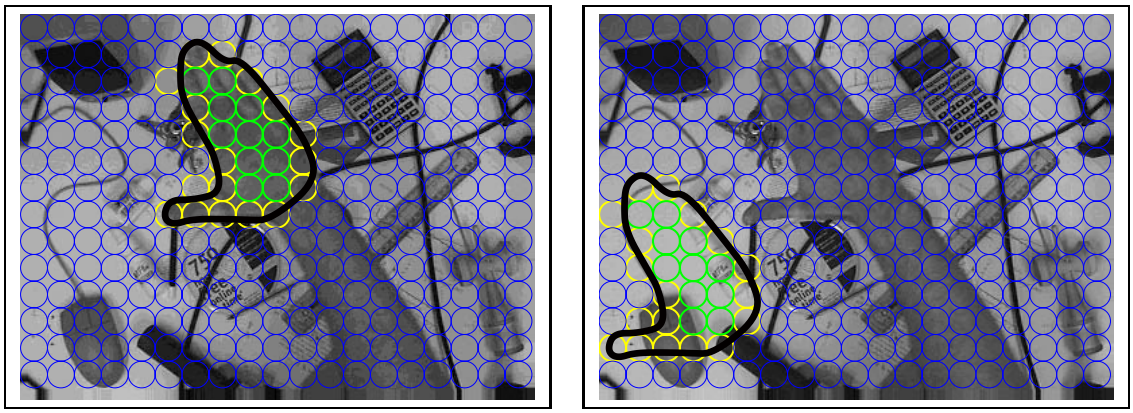


Figure 1: **The world through a filter bank.** A bank $Z = (z_1, \ldots, z_K)$ is illustrated here with circular supports $S_1, \ldots, S_K$ arranged on a regular grid, so that the world is viewed, in effect, through a sieve. Supports are labelled foreground (inside the black hypothesised outline), background or mixed, according to the hypothesised X — left: approximately correct ($X = X_0$); right: X out in the clutter.

$\{W_k\}$ whose outputs (conditioned on object configuration X) are mutually independent, at least approximately, so that a joint conditional density for the bank of outputs — the image observation likelihood function — can be constructed as a product:

$$p(Z|X) = \prod_{k=1}^{K} p(z_k|X). \tag{7}$$

Single filter likelihoods $p(z_k|X)$ are learned directly from training images (Geman and Jedynak, 1996) and details are given later. For simplicity and computational efficiency (Mallat, 1989; Burt, 1983; Shirai and Nishimoto, 1985), we restrict the fixed bank to contain filter functions

$$W_k(\mathbf{x}) = W(\mathbf{x} + \mathbf{u}_k) \tag{8}$$

that are simply copies of a standard filter $W(\mathbf{x})$, translated over some regular grid defined by the displacement vectors $\{\mathbf{u}_k\}$.

2.4 Factored sampling

For the multi-modal distributions that arise with image observation likelihoods, Bayes' formula (3) cannot be computed directly but Monte-Carlo simulation is possible. In *factored sampling* (Grenander et al., 1991), random variates are generated from a distribution that approximates the posterior $p(X|Z)$. A weighted "particle-set" $\{(s^{(1)}, \pi_1), \ldots, (s^{(N)}, \pi_N)\}$, of size N, is generated from the prior density $p_0(X)$ and each particle $s^{(i)}$ is associated with a likelihood weight $\pi_i = f(s_i)$ where $f(X) = p(Z|X)$. Then, an index $i \in \{1, \ldots, N\}$ is sampled with replacement, with a probability proportional to π_i; the associated s_i is effectively drawn from a distribution that converges (weakly) to the posterior, as $N \to \infty$. It will prove useful later to express the sampling scheme graphically, as a "particle diagram"

$$\boxed{p_0} \xrightarrow[N]{} \bigcirc \xrightarrow{\times f} \bigcirc \xrightarrow[N]{\sim} \bigcirc. \tag{9}$$

It is interpreted as follows: the first arrow denotes drawing N particles from a known density p_0, with equal weights $\pi_i = 1/N$. (Particle sets are represented by open circles.) The $\times f$ operation denotes likelihood weighting of a particle set:

$$\pi_i \to f(s^{(i)})\pi_i, \ i = 1, \ldots, N.$$

The final step denotes sampling with replacement, as described above, repeated N times, to form a new set of size N in which each particle is given a unit weight; each particle is therefore drawn approximately from the posterior.

Where the likelihood f is a very narrow function in configuration space, sampling can become inefficient, requiring large N in order to give reasonable estimates of the posterior. In the paper (section 8) it is shown how this can be mitigated by "layered sampling" in which broader likelihood functions are used in an advisory capacity to "focus" the particle set down, in stages. In the vision context, layered sampling is a vehicle for implementing multi-scale processing.

3 Probabilistic modelling of observations

The observation (ie output value) z from an individual filter is generated by integration over a support-set S such as the circular one in figure 2, which is generally composed of both a background component $B(X)$, and a foreground component $F(X)$:

$$z|X = \underbrace{\int_{B(X)} W(\mathbf{x})I(\mathbf{x})\,\mathrm{d}\mathbf{x}}_{\text{MAIN NOISE SOURCE}} + \int_{F(X)} W(\mathbf{x})I(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{10}$$

The main source of variation in $z|X$ is expected to come from the background which is assumed to be a sample from some general class of scenes. In contrast, the foreground relates to a given object, relatively precisely known, though still subject to some variability. This means that there should be a steady reduction in the variance of the distribution of $z|X$ as X changes from values in which the circular support is entirely over foreground, via intermediate locations overlapping both foreground and background, and finally to values in which it is entirely over background. This is supported by experiments in which density functions for z which

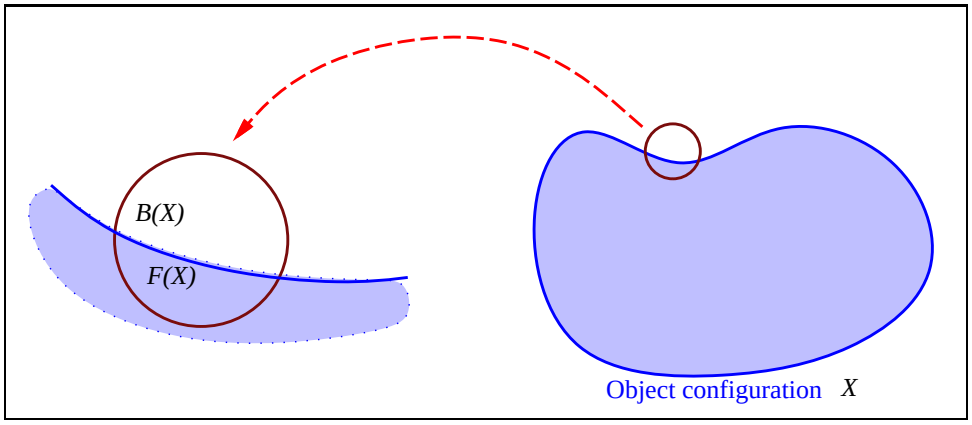


Figure 2: **The support of a mask.** A circular support set S is illustrated here, split into subsets $F(X)$ from the foreground and $B(X)$ from the background.

have been learned from images, both from background regions and also from foreground regions (figure 3). The filter used in the experiment is a Gaussian

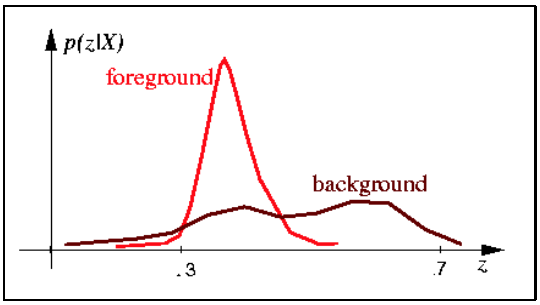


Figure 3: **Learned observation densities for a Gaussian filter.** Densities $p(z)$ are exhibited both for foreground and background, in the case that $W(\mathbf{x})$ is Gaussian, with support radius $r = 20$ pixels. Units of z are intensity, scaled so that intensities in the original image lie in the range $0, 1$.

$$G_\sigma(\mathbf{x}) = \frac{1}{\sigma^2} \exp -\frac{|\mathbf{x}|^2}{2\sigma^2} \tag{11}$$

in a circular support of radius r $(= 3\sigma)$.

The role of $p(z|X)$ in Bayesian localisation is as a likelihood function for X, associated with a particular observation z, as illustrated in figure 4. Note that, although X is generally multidimensional, in the diagram it is depicted as a one-dimensional variable, for the sake of clarity. The entire family of idealised densities can be represented in (z, X)-space as shown in the figure. Then, to construct the likelihood functions, the z-value is considered to be fixed and X allowed to vary. This is illustrated in figure 4 by considering slices of constant z. For example, $z = 2$ in the figure depicts a relative high value which, in the example, is more likely to be associated with a filter-support lying predominantly over the foreground. The resulting likelihood is peaked

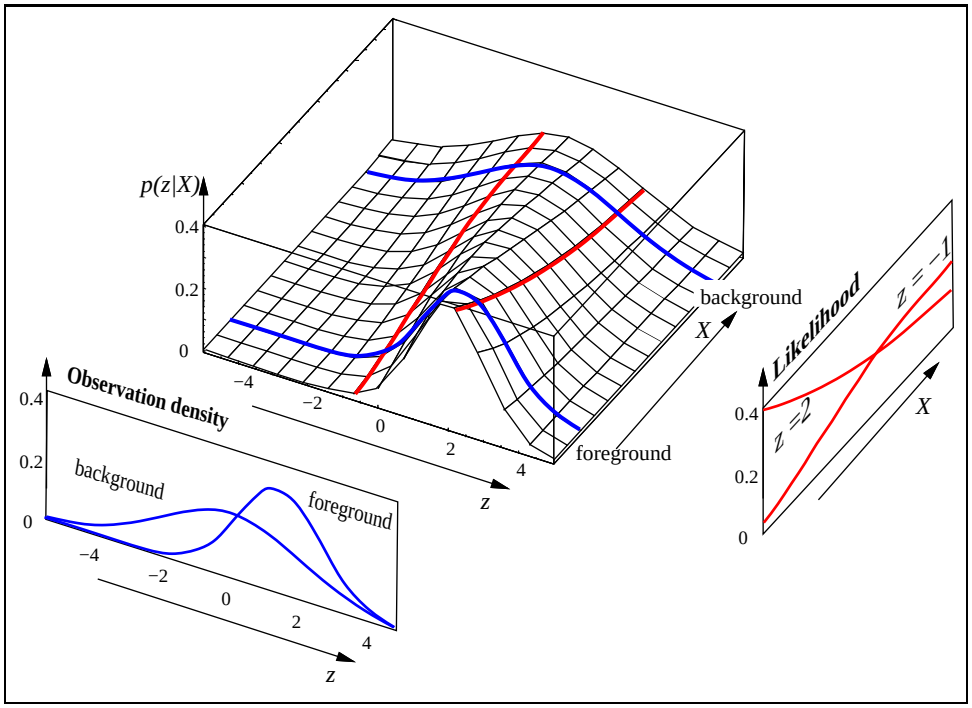


Figure 4: **Observation likelihood.** The density $p(z|X)$ is formally a function of z with X as a parameter, and is illustrated for foreground and background cases. The whole family of such one-dimensional densities, indexed by the continuous variable X, are assembled to synthesise $p(z|X)$, as shown. Now $p(z|X)$ is "sliced" in the orthogonal direction, to generate likelihoods (functions of X for fixed z). In the examples, an observation $z = 2$ biases X towards a foreground value, whereas $z = -1$ biases towards background.

around a value of X corresponding to predominant foreground support. Conversely, for $z = -1$, the support is more likely to be predominantly over the background and the mode of the likelihood shifts towards background values of X. [3]

Likelihood functions from several observations z_k should "fuse" when they are combined (7), to form a joint likelihood that is more acutely tuned (figure 5) than the likelihood for any individual z_k. Note the importance of the z_k from "mixed" supports, lying partly on the background and partly on the foreground. It might be tempting to regard them as contaminated and discard them whereas, in fact, they should be especially informative, responding selectively to the boundary of the object — see figure 1.

4 Filter response-learning

If it were not for mixed supports, learning would be relatively straightforward. Over the background, for instance, it would be sufficient just to evaluate the outputs z (6) of the circular filter repeatedly, at assorted

[3] Note that "slicing" is purely an analytical tool to illustrate the way observation likelihoods exist implicitly within a probabilistic model for filter response. Slicing does not actually form part of any algorithm proposed here.

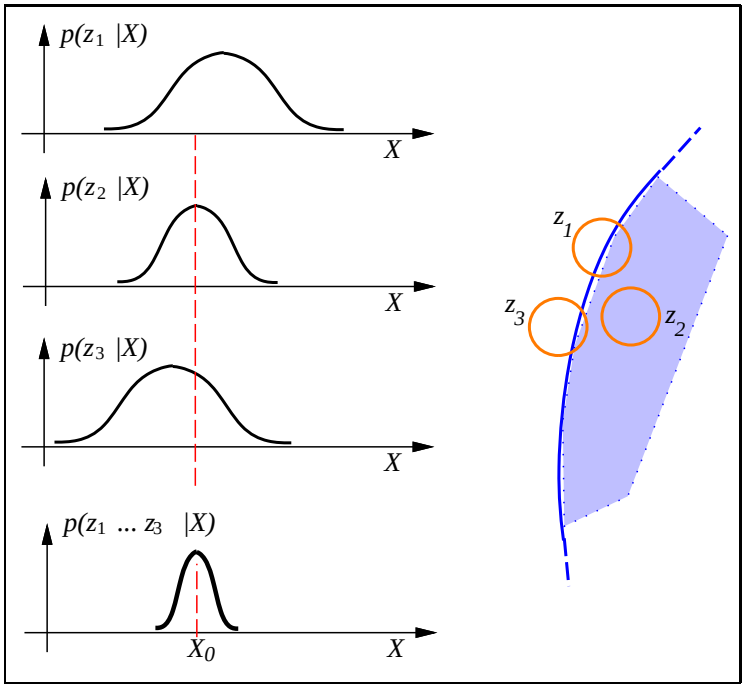


Figure 5: **"Hyperacuity" from pooled observations.** Likelihoods from independent observations combine multiplicatively, to give a joint likelihood narrower than any of the individual constituents.

locations over some training image, and fit a probability distribution $p^{\mathcal{B}}(z)$. However, over a mixed support, only a part of the circle lies over the background. If this part is approximated as a segment of a circle (figure 6), and provided each filter functional $W_k(\mathbf{x})$ is isotropic (or steerable (Perona, 1992)), then the background

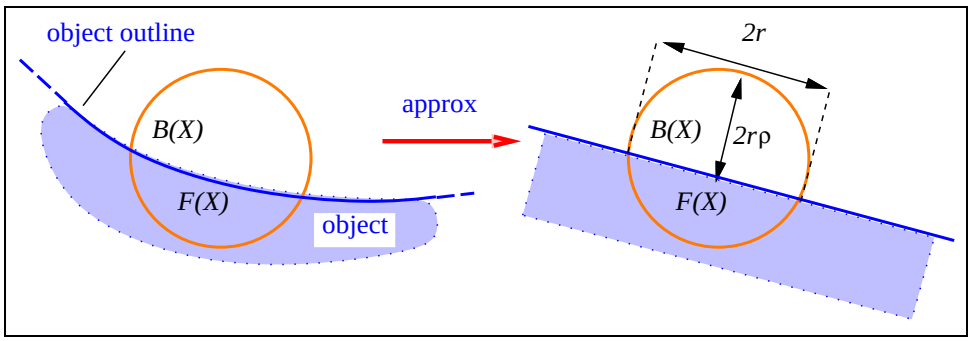


Figure 6: **Approximating foreground/background supports** Assuming that the object's bounding contour is sufficiently smooth (on the scale r of the radius of the filter support) the boundary between foreground and background can be approximated as a straight line. The support therefore divides into segments with offsets $2r\rho$ and $2r(1-\rho)$ for background and foreground respectively.

distribution can be parameterised by a single offset parameter ρ (at a given scale r). This parameter is defined

for $0 \leq \rho \leq 1$, as in the figure so that: when $\rho = 1$ the filter support is entirely over the background; when $\rho = 0$ it is entirely over the foreground; and for $0 < \rho < 1$ it straddles the object boundary.

Training examples for background learning must be constructed over circular segments with offsets throughout the range $0 \leq \rho \leq 1$, to learn background distributions $p_k^B(z|\rho)$. (Clearly, in practice, only a finite number of these can be learned, leaving the continuum of ρ to be filled in by interpolation.) To consider a hypothesised configuration X, the Bayesian localisation algorithm needs to evaluate, for each k, an *offset function* $\rho_k(X)$ and a likelihood $p_k(z|\rho_k(X))$. The likelihood function consists of a sum of background and foreground components, and is therefore constructed as a (numerically approximated) convolution

$$p_k(z|\rho) = p_k^B(z|\rho) * p_k^F(z|\rho) \tag{12}$$

of learned background and foreground density functions.

5 Learning the background likelihood

Statistical independence of image features is an issue that has been studied elsewhere, in the context of neural coding (Field, 1987): if neural codes are efficient in the sense of avoiding redundancy, their components can be expected to be nearly statistically independent. It is also known that independent components of natural scenes tend to have "sparse" or "hyper-kurtotic" distributions — ones with extended tails compared with those of a normal distribution (Bell and Sejnowski, 1997).

5.1 Experiments with response correlation

Experiments on background correlation are done here using statistics collected from each of the four scenes in figure 7. Our experiments are similar to those done by Zhu and Mumford (1997) in which they showed the background distribution is remarkably consistent across scenes, for a ∇G filter. Here we look at the div of that filter output, which should therefore similarly show a consistent distribution, and the small-scale experiments done here support that. A necessary condition for independence is freedom from correlation, so autocorrelation was estimated by random sampling of pairs of supports, separated by a varying displacement. This was done for two choices of filter function $W(\mathbf{x})$: Gaussian $G(\mathbf{x})$ and Laplacian of Gaussian $\nabla^2 G(\mathbf{x})$, and typical results are shown in figure 8. At a displacement such as r $(= 3\sigma)$, corresponding to a typical separation between filters, the $G(\mathbf{x})$ filter shows correlation and hence there cannot be independence. On the other hand $\nabla^2 G(\mathbf{x})$ is uncorrelated at a displacement of r. Further experiments, looking at the entire joint distribution for responses z_k, z_l of two filters with variable spatial separation, support statistical independence, as figure 9 shows.

The independence is obtained at the cost of throwing away information about mean response and the 1st moment, though this is likely to be beneficial in conferring some invariance to illumination variations. These experiments were for complete, circular supports. With part-segments of a circle ($\rho < 1$), statistical independence of $\nabla^2 G(\mathbf{x})$ responses deteriorates. Experiments like the ones in figure 8 show correlation lengths increasing for $\rho < 1$, with $\rho = \frac{1}{4}$ the worst case. This will mean greater statistical dependence between mixed supports, and it is not clear how this could be improved; but note at least that typically it is a minority of filter supports that are mixed.

Figure 7: **Background learning:** training scenes used in experiments.

Fitting the background distribution

A further benefit of the $\nabla^2 G(\mathbf{x})$ filter is that the learned background distributions turn out to be far more constant across scenes(and this is known to be true also for ∇G filters (Zhu and Mumford, 1997)) than for a plain $G(\mathbf{x})$ filter. Background distributions were learned by repeated sampling of z_k (6) for randomly positioned supports, then histogramming and smoothing to estimate $p^{\mathcal{B}}(z)$. The results for complete circular supports ($\rho = 1$), shown in figure 10, show sufficient consistency to indicate that some fixed parametric form should be sufficient to represent the densities. The learned responses turn out not to be normally distributed, but have a hyper-kurtotic distribution, that is one with greater kurtosis than a normal distribution, and this is clearly visible in the extended tails in figure 10. Hyper-kurtotic distributions are known to emerge in independent components of images (Bell and Sejnowski, 1997), and are often found to be well modelled by a single-exponential distribution[4]

$$p^{\mathcal{B}}(z) \propto \exp -|z|/\lambda. \tag{13}$$

The distribution fits the experimental data quite well (figure 11). In that case, a global background likelihood

[4]We refrain from the commonly used term "Laplace" distribution here, to avoid the potential confusion with the Laplacian operator in $\nabla^2 G$.

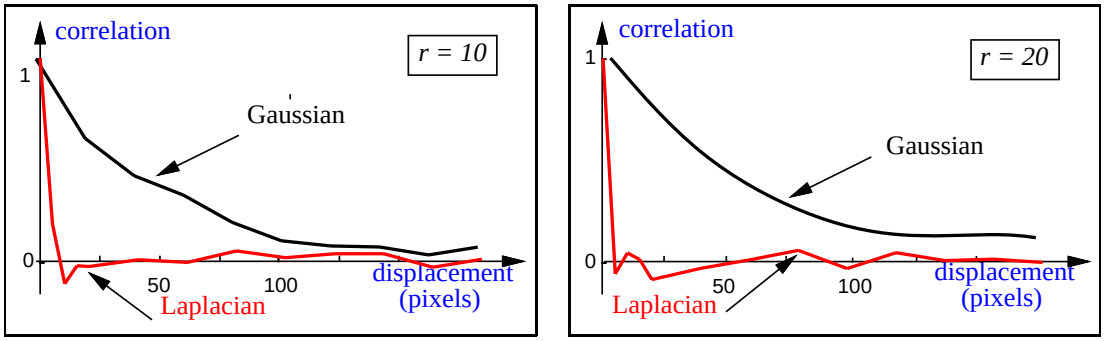
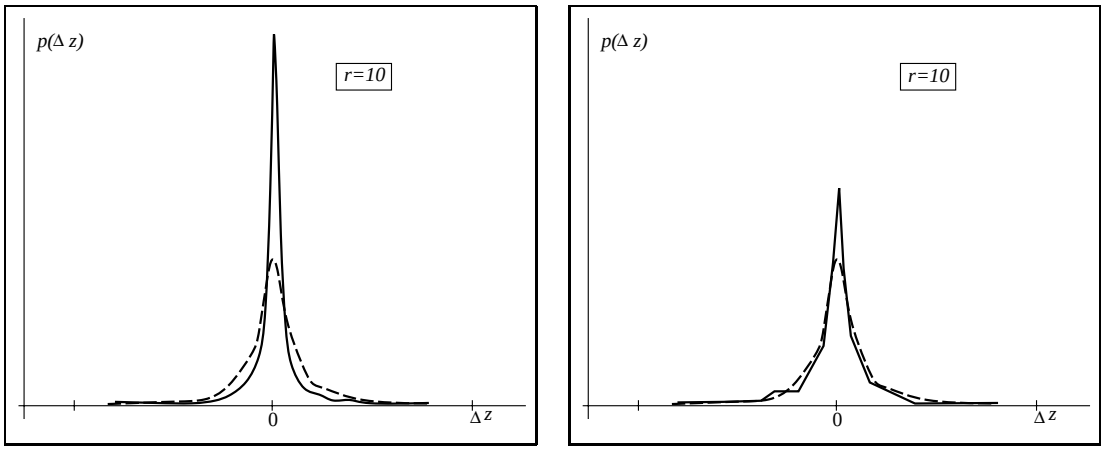


Figure 8: **Autocorrelation of filter output.** Results are for the first (hand) image from figure 7, at two sizes of spatial scale r. The Gaussian filter $G(\mathbf{x})$ shows substantial long-range correlation whereas, for $\nabla^2 G(\mathbf{x})$ correlation falls to zero for non-overlapping supports.



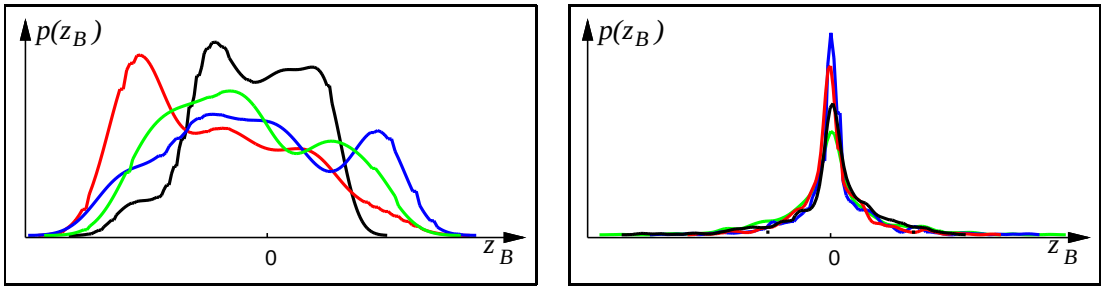
displacement $\delta = 5$ pixels displacement $\delta = 20$ pixels

Figure 9: **Independence of filter output.** Two filters displaced δ apart have outputs z_1, z_2, and the distribution of the difference $\Delta z = z_1 - z_2$ is plotted here. The dashed curve shows a reference distribution for large δ. In the case $\delta = 5$ pixels that correlation is high (see figure 8) z_1, z_2 are clearly not independent — the distribution for Δz does not match the reference distribution. However with $\delta = 20$ pixels, for which z_1, z_2 are uncorrelated, they are shown here also to be approximately independent.

of the form (7), is a product of exponentials of filter responses, just the scene model derived by Zhu et al. via maximum entropy (Zhu et al., 1998, eq. (21)).

For $\rho < 1$ (circle segments), the single-exponential distribution does not fit so well, with $\rho = \frac{1}{4}$ again being the worst case. [This is to be expected, given that $\nabla^2 G$ does not sum to 0 over an arbitrary segment of a circle, except for the semi-circle $\rho = \frac{1}{2}$. This implies that the distribution mean will not be zero, and hence cannot have single-exponential form.]

Since $W = \nabla^2 G$ sums to 0, the means of densities $p^{\mathcal{F}}$ and $p^{\mathcal{B}}$ for foreground and background will also coincide at 0, as in figure 12. Given this loss of the information associated with the means of $p^{\mathcal{B}}$ and $p^{\mathcal{F}}$,



Gaussian $G(\mathbf{x})$ Laplacian of Gaussian $\nabla^2 G(\mathbf{x})$

Figure 10: **Learned background distributions.** Learned densities $p^{\mathcal{B}}(z)$ are shown here for each of the four scenes in figure 7 at scale $r = 20$: they are highly variable for the $G(\mathbf{x})$ filter, but rather consistent for $\nabla^2 G(\mathbf{x})$.

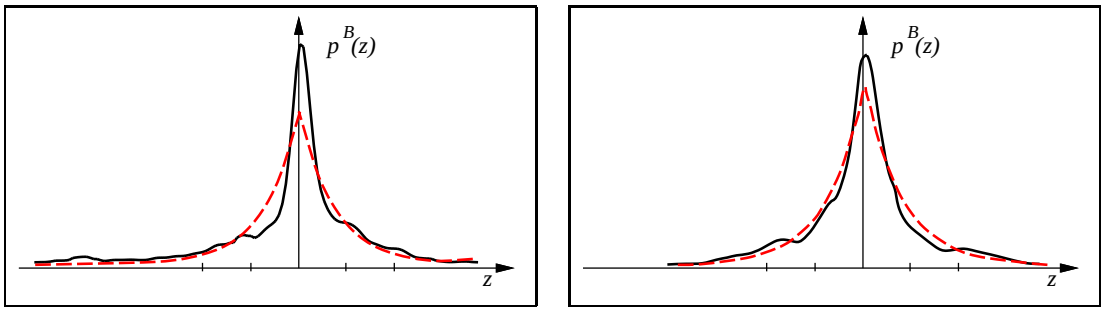


Figure 11: **Exponential model for background distributions.** Learned densities $p^{\mathcal{B}}(z)$ for the first and last of the 4 scenes in figure 7, at scale $r = 20$ with $\rho = 1$, are fitted here (by MLE) to an exponential distribution, which captures the elongation of the tails.

discriminability between foreground and background is reduced, the price paid for improved illumination-invariance. However, the foreground model can be extended in certain ways to improve discriminability again. One way is "foreground subdivision" as in section 6; another uses intensity templates (Sullivan and Blake, 2000).

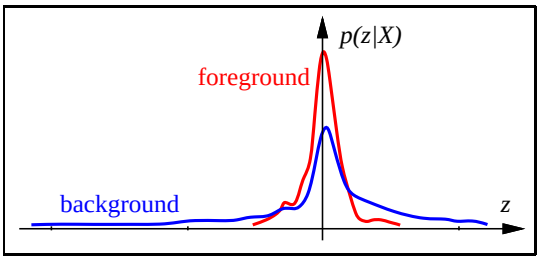


Figure 12: **Foreground and background distributions when** $\int W(\mathbf{x})\mathrm{d}\mathbf{x} = 0$**,** for support radius $r = 20$ pixels. The means of the foreground and background distributions now coincide, cf. figure 3.

5.2 Optimal filter bank grid

At a given spatial scale, the maximum information about an image can be collected by packing filter supports S_k as densely as possible, within the constraint that filter outputs z_k must be uncorrelated. For filters W_k that are isotropic, correlation depends simply on the displacement between pairs of filters. A useful measure is that the correlation function (figure 8) crosses 0 at a displacement of around r $(= 3\sigma)$. The most effective packing of filters, for the given level of correlation, will be the one that maximises the packing density for a given minimum displacement between filter centres. This is well-known to be a hexagonal tesselation, whose packing density is approximately 50% greater than square packing. For the $\nabla^2 G$ filter, the filter support is circular[5] with radius approximately r $(= 3\sigma)$ which is also the displacement for (approximately) zero correlation. Hence supports in the hexagonally tesselated optimal filter bank overlap substantially as in figure 13.

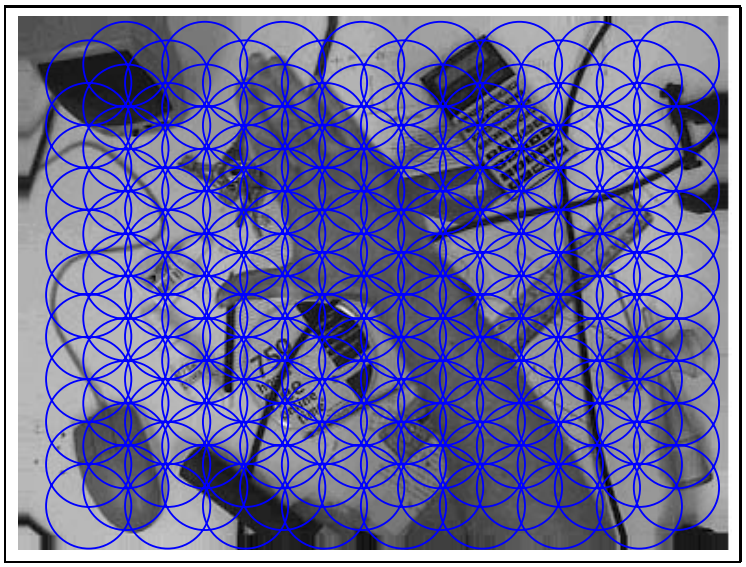


Figure 13: **Optimal tesselation of filter supports.** Maximum density of $\nabla^2 G$ filters, while avoiding correlation between filter pairs, is achieved by a hexagonal tesselation, as shown, with substantial overlap (support radius $r = 40$ pixels illustrated).

6 Learning the foreground likelihood

Learning distributions for foreground responses is similar to the background case. As before, $p^{\mathcal{F}}(z|\rho)$ is learned for some finite set of ρ-values, and interpolated for $\rho \in [0, 1]$. There are some important differences however.

[5] Of course, the filter has theoretically unbounded support, but we take the point at which filter amplitude falls to around 10% of its maximum value.

6.1 Deformations and pooling

Three-dimensional transformations and deformations of the foreground object must be taken into account. Tabulating $p^{\mathcal{F}}$ not only against ρ but also against transformation parameters is computationally infeasible. Variations that cannot be modelled parametrically can nonetheless be pooled into the general variability represented by $p^{\mathcal{F}}(z|\rho)$. This implies that $p^{\mathcal{F}}(z|\rho)$ should be learned not simply from one image, but from a training set of images containing a succession of typical transformations of the object.

6.2 Outline constraint

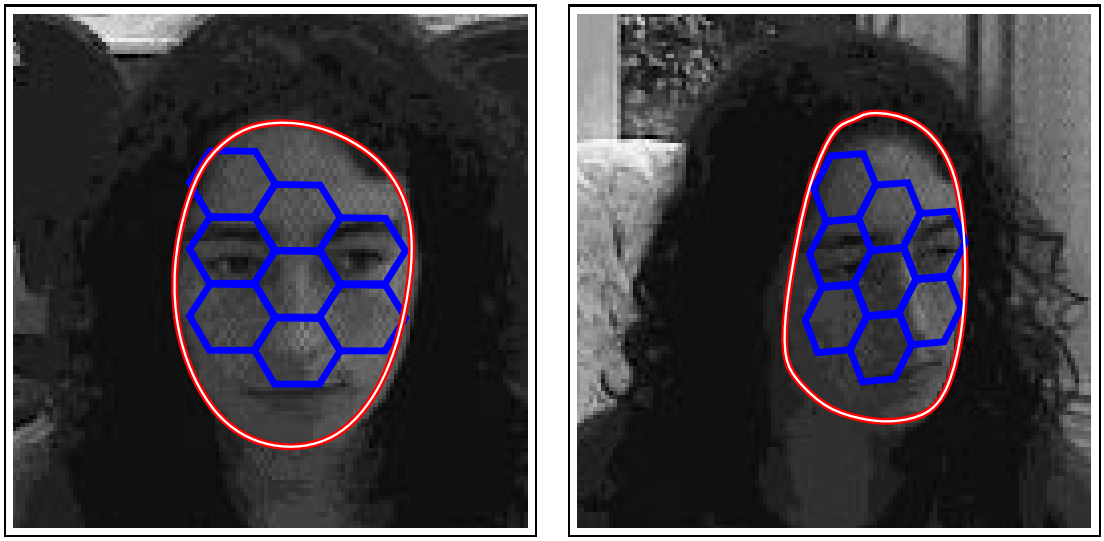
The distribution $p^{\mathcal{B}}(z|\rho)$ was learned from segments dropped down at random, anywhere on the background. Over the foreground, in the case that $\rho = 0$, $p^{\mathcal{F}}(z|\rho)$ is similarly learned from a circular support, dropped now at any location wholly inside the training object. However, whenever $\rho > 0$, the support $F(X)$ must touch the object outline; therefore, for $0 < \rho < 1$, $p^{\mathcal{F}}(z|\rho)$ has to be learned entirely from segments touching the outline.

6.3 Foreground subdivision

For $\rho = 0$, it has so far been proposed that $p^{\mathcal{F}}(z|\rho)$ be learned by pooling responses throughout the object interior. Pooling in this way discards information contained in the gross spatial arrangement of the grey-level pattern. Sometimes this provides adequate selectivity for the observation likelihood, particularly when the object outline is distinctive, such as the outline of a hand as in figure 1. The outline of a face, though, is less distinctive. In the extreme case of a circular face, and using isotropic filters, rotating the face would not produce any change in the pooled response statistics. In that case, the observation likelihood would carry no information about (2D) orientation. One approach to this problem is to include some anisotropic filters in the filter bank, which would certainly address the rotational indeterminacy.

An alternative approach which also enhances selectivity generally, is to subdivide the interior $\mathcal{F}$ of the object as $\mathcal{F} = \mathcal{F}_0 \cup \ldots \cup \mathcal{F}_{N_F}$, as in figure 14, and construct individual distributions $p^{\mathcal{F}_i}(z|\rho = 0)$ for each subregion $\mathcal{F}_i$. A foreground distribution $p^{\mathcal{F}_i}(z|\rho = 0)$ applies to any filter support S_k that lies entirely within $\mathcal{F}$ and whose centre is in $\mathcal{F}_i$. The case $i = 0$ is a "catch-all" region, pooling the responses of any filter whose centre is not in $\mathcal{F}_i$ for any $i > 0$ (the hexagons in figure 14). The choice of the number N_F of sub-regions is of course a trade-off between increasing, with N_F, the specificity of the information that is learned while, at the same time, requiring more data to learn adequate estimates of the $p^{\mathcal{F}_i}$ as the sub-regions $\mathcal{F}_i$ get smaller.

Sub-regions are defined with respect to a standard configuration, say $X = 0$, as in figure 14a. In a novel configuration $X \neq 0$, encountered either in training or evaluation of the likelihood $p(Z|X)$, suitably warped forms of $\mathcal{F}_i$ must be defined (figure 14b)). This could be achieved by defining the configuration space $\mathcal{X}$ as a space of two-dimensional warps g_X, using thin plate splines for example (Bookstein, 1989). A more economical but more approximate approach is adopted here, representing the outline contour as a parametric spline curve (Bartels et al., 1987), and the configuration-space $\mathcal{X}$ is modelled as a sub-space of the spline space. Then the warp of the *interior* of the object is approximated as an affine transform by *projecting* the configuration X onto a space of planar-affine transformations (Blake and Isard, 1998, ch 6). The fact that this affine transformation warps the interior only approximately does not imply that errors are introduced into the Bayesian localisation procedure. Rather, the variability due to approximating the warp is simply pooled during



(a) (b)

Figure 14: **Foreground subregions.** The object interior $\mathcal{F}$ is subdivided (a) as $\mathcal{F} = \mathcal{F}_0 \cup \mathcal{F}_1 \cup \ldots \cup \mathcal{F}_{N_F}$ where sub-regions $\mathcal{F}_1, \ldots, \mathcal{F}_{N_F}$ here are hexagons and $\mathcal{F}_0$ is the remaining part of $\mathcal{F}$. In a novel view (b), sub-regions must be mapped onto the new images, done here by approximating the warp of the interior as a planar-affine map.

learning into the distributions $p^{\mathcal{F}_i}$. The resulting model then loses some *specificity* but is still "correct" in that the variability is fairly represented by probabilistic pooling.

6.4 Statistical independence

Known behaviour for independence of natural scenes, which applied well to background modelling, cannot necessarily be expected to apply for foreground models, given that the foreground is far less variable. Nonetheless, repeating the autocorrelation experiments now for the foreground has produced evidence of good independence for $\nabla^2 G$ filters, as in figure 15.

6.5 Representing the distribution

Whereas filter response z over (highly variable) background texture assumed the characteristic kurtotic form, the foreground is far less variable and does not have extended tails (figure 12). Hence the exponential distribution is unsuitable. A normal distribution might be more appropriate but the safest approach is to continue to represent $p^{\mathcal{F}}$ in a more general fashion, as an interpolated histogram.

6.6 Intensity offset model

Recently, we have developed a more effective form of foreground model which incorporates an *intensity offset*. Briefly it works as follows, but see (Sullivan and Blake, 2000) for details of the approach. Over the foreground

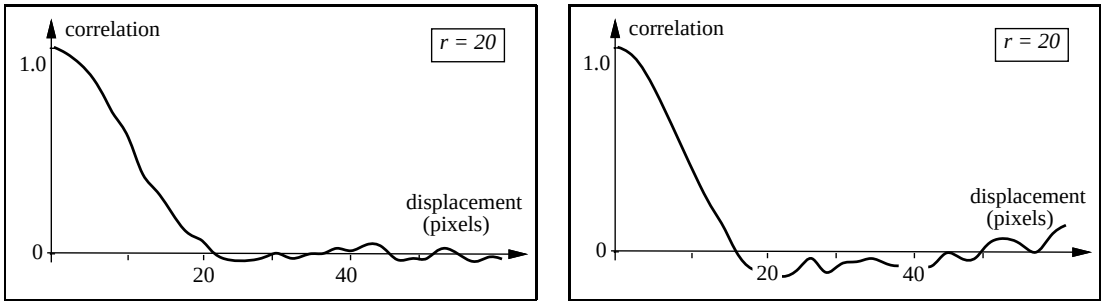


Figure 15: **Foreground autocorrelation** for the $\nabla^2 G$ filter, over two different foreground objects: a hand (left) and a face (right). In both cases, correlation falls to zero at a displacement of around r or 3σ, similarly to correlation of background texture.

$\mathcal{F}(X)$, the intensity $I(\mathbf{x})$ is modelled as having a mean $\bar{I}_X(\mathbf{x})$ generated as a warp

$$\bar{I}_X(\mathbf{x}) = \bar{I}(T_X(\mathbf{x}))$$

of a learned intensity template $\bar{I}(\mathbf{x})$. This then leaves only the difference

$$\Delta I_X(\mathbf{x}) = I(\mathbf{x}) - \bar{I}(T_X(\mathbf{x})), \quad \mathbf{x} \in \mathcal{F}(X),$$

as observed by the filter bank $\{W_k\}$, to be modelled statistically. More of the variation in the intensity pattern $I(\mathbf{x}), \quad \mathbf{x} \in \mathcal{F}(X)$ is accounted for deterministically, leaving a tighter distribution for the random component of the foreground model.

Inclusion of the intensity offset, in this way, fulfils a similar objective to the foreground subdivision of section 6, in using more of the information in the spatial intensity pattern of the object. It turns out (Sullivan and Blake, 2000) to have an additional advantage: that the template model can be extended to take some account of lighting variations deterministically, rather than leaving lighting changes to be modelled entirely statistically.

7 Exercising the learned observation likelihood

Having established, in previous sections, that reasonable densities $p_k(z|r)$ for individual supports can be learned from background and foreground densities, it is now possible to exercise the full joint likelihood function $p(Z|X)$. This is constructed (7) as a product, in which the offset ρ for each support segment is obtained from its offset function $\rho_k(X)$:

$$p(Z|X) = \prod_{k=1}^{K} p_k(z_k|\rho_k(X)). \tag{14}$$

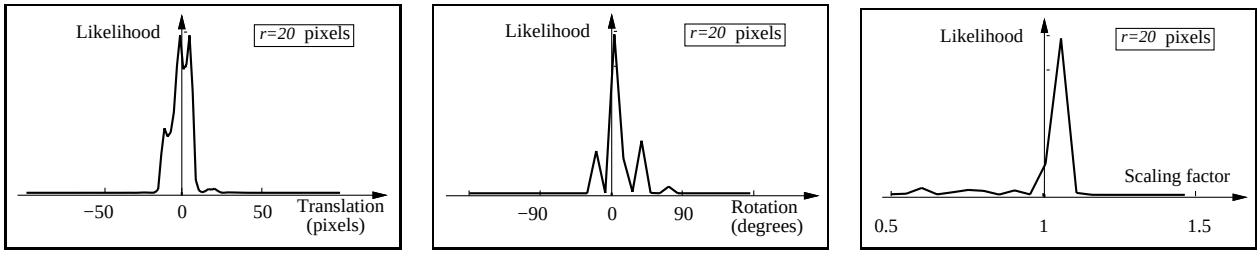
Evaluation of the offset function requires a geometrical calculation of the size of the circle-segment that approximates the intersection of the object (at configuration X) with the kth support. It is interesting to note that, although Bayesian analysis requires that Z should consist of the entire set of filters z_k in figure 1, some economies can legitimately be made. Given a sample $X_1, \ldots, X_N$ of object hypotheses, if some filter support S_k lies always in the background for *all* the X_n, the corresponding term can be factored out of (14). For a

truly parallel, pyramid architecture this may be no real advantage. If image processing is serial a "sampling rehearsal" can tag just those z_k whose likelihoods do not factor out; other z_k need not be computed. The "factoring out" phenomenon also makes another interesting point. The filters that actually contribute to global likelihood variations are those near the boundary of at least some hypothesised configuration X; so despite being intensity-based, it transpires that Bayesian localisation does in fact emphasise edge information.

The learned observation likelihood is exercised here in two ways. First, the likelihood function is explored systematically, with respect to translation, rotation etc., and at various spatial scales. Secondly, the likelihood function is applied to randomly generate samples, to sweep out posterior distributions for pose, again at several scales.

7.1 Systematic variations in observation likelihood

First, for the hand scene of figure 1, $p(Z|X)$ — the joint likelihood composed of a product of likelihoods $p(z_k|X)$ for individual filters, is exercised systematically. This is done as a check that the likelihood does register a peak at the true object position, and has reasonable variations around the peak. In these demonstrations, X is varied over a configuration space of Euclidean similarities; results are displayed in figure 16. The joint likelihood fuses information from individual supports effectively, with a maximal value, as expected,

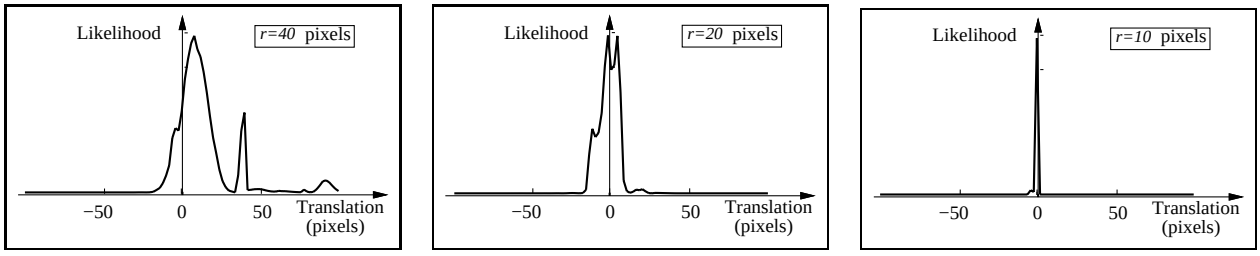


Translation Rotation Scaling

Figure 16: **Exercising the joint likelihood.** The joint observation likelihood $p(Z|X)$ is exercised here as X ranges over coordinate axes in the space of Euclidean similarities. Note that the peak in each case is approximately at the origin ($X = X_0$). (Support radius is $r = 20$ pixels.)

near the true solution X_0. Figure 17 demonstrates the effect of changing the filter scale r. As expected, the likelihood function is more broadly tuned at coarser scales, appearing to have a width of about $2r$, or less due to hyperacuity effects as in figure 5. As a final check, it is interesting to consider the likelihood ratio for two configurations, one correctly positioned over the target, and one way out over background as in figure 1. In such cases, treating pixels as independent typically produces ridiculously large likelihood ratios. Even using Gaussian masks ($r = 20$), which we know are not independent, gives a likelihood ratio in this case of $1 : 10^{55}$ — still very large. However, this falls considerably with $\nabla^2 G$ masks, as expected given the independence of their output over foreground and background, to a more plausible $1 : 10^4$

To summarise, the learned observation likelihood for $\nabla^2 G$ masks has been exercised here, systematically, and found to have reasonable properties. The next task is to use it to compute approximations to the posterior $p(X|Z)$, by means of the factored sampling scheme of section 2.4.



$r = 40$ pixels $r = 20$ pixels $r = 10$ pixels

Figure 17: **Joint likelihood at various scales.** The observation likelihood $p(Z|X)$ shown for translation, at various scales. Again, modes are approximately unbiased, and the width of the likelihood peak increases with r.

7.2 Sampling from the posterior

To locate a hand against a cluttered background, by Bayesian localisation let us assume first that its orientation is known but that the prior $p(X)$ for translation is broad (has high variance). Samples from the posterior, at several scales, are shown in figure 18. For a given scale, the broad prior is focused down to a narrow posterior distribution which, as earlier in figure 17, is narrower at finer scales. It is not clear from figure 18 that coarse scales actually have a useful role — the finest scale, after all, gives the most precise information. However, if the sampling process is "pressed" harder, by expanding the prior without increasing the size N of the particle-set, the fine scale breaks down, as figure 19 shows, while at the two coarser scales, sampling from the posterior continues to operate correctly. That suggests a role for coarser scales in guiding or constraining finer ones, if only a Bayesian sampling mechanism can be found to do it, and that is the subject of section 8.

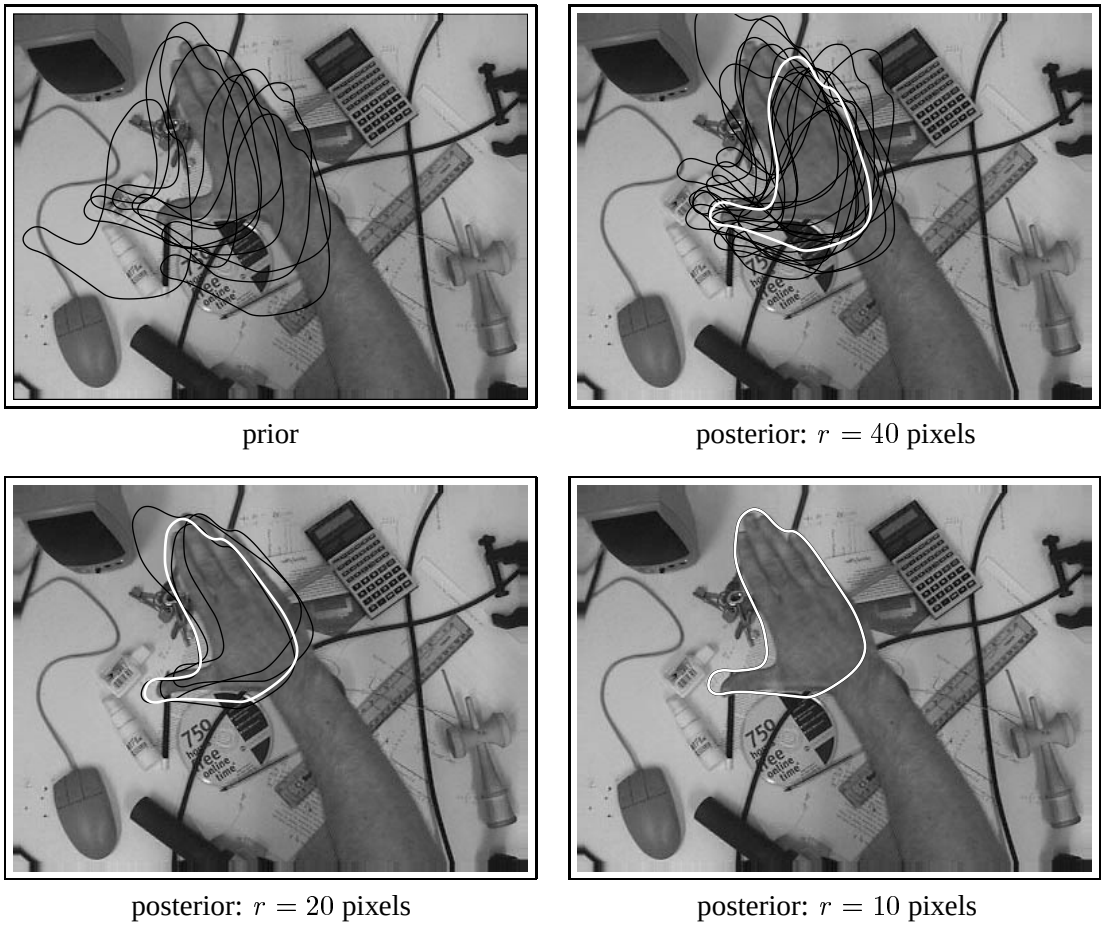
8 Layered sampling

In section 7.2, the problem of "overloading" was demonstrated, that occurs when image observations are made at a fine spatial scale. It results from the observation likelihood $f(X)$ having a support that is narrow compared with the support of the prior $p_0(X)$. A continuation algorithm is used to reduce computational complexity by introducing a sequence of likelihoods f_n whose supports are intermediate between those of $p_0(X)$ and $f(X)$, and which reduce progressively in size. One form of this idea is "annealed importance sampling" (Neal, 2000), in which $f(X)$ is replaced by $f(X)^\beta,\ 0 < \beta < 1$ in order to broaden likelihood function. It is known to reduce the number of particles needed for estimation (to a given accuracy) by importance sampling, from N to $\log N$.

Layered sampling is an alternative form of continuation principle in which the intermediate likelihoods are obtained by making image measurements at a variety of spatial scales. Filter responses at several scales $r = r_1, r_2, \ldots$ are used in coarse-to-fine sequence. so background distributions

$$p^{\mathcal{B}}(z|\rho, r),\ 0 \leq \rho \leq 1,\ r = r_1, r_2, \ldots$$

need to be learned at each scale, and similarly for foreground distributions.



prior | posterior: $r = 40$ pixels

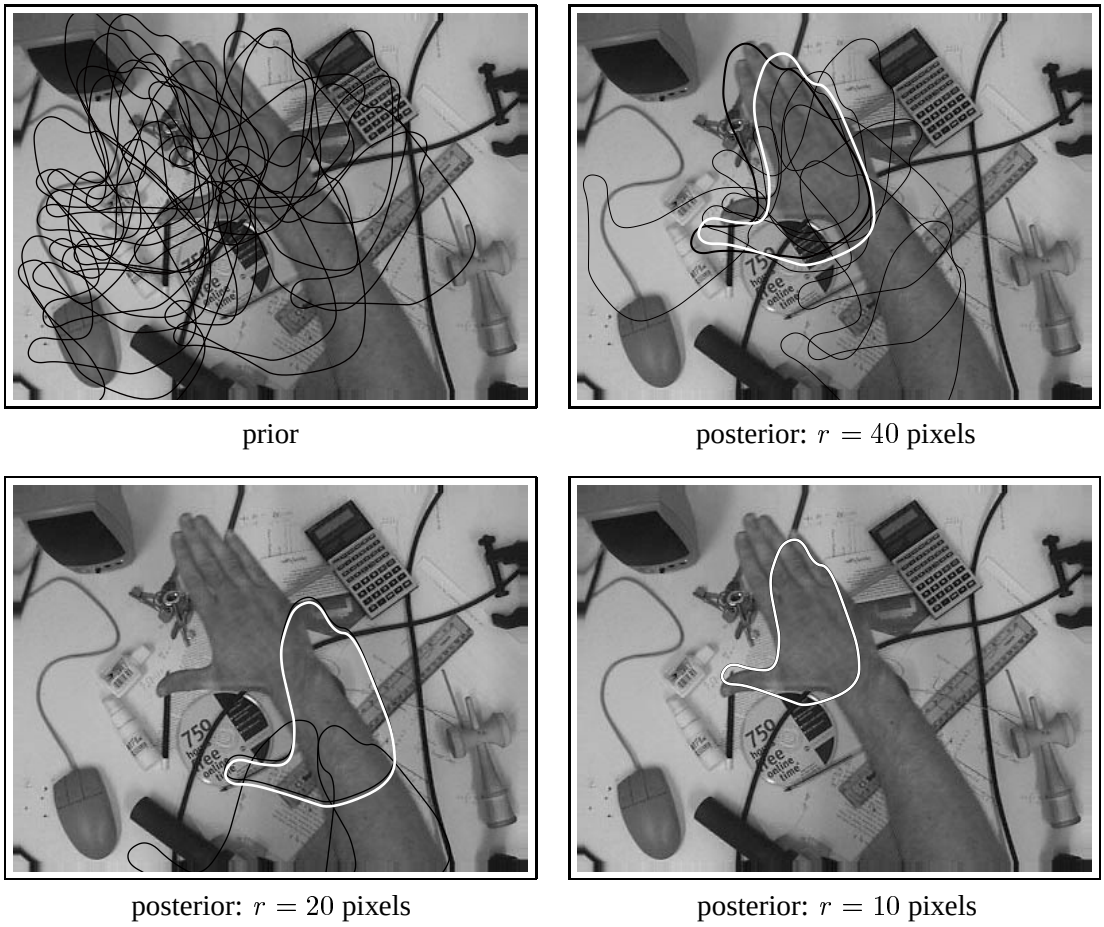
posterior: $r = 20$ pixels | posterior: $r = 10$ pixels

Figure 18: **Random samples from the posterior.** Factored sampling from the posterior density $p(X|Z)$, in which the prior $p(X)$ is a broad distribution of Euclidean similarities (planar rigid motion plus size-scaling). At each scale r, the posterior mean $\mathcal{E}[X|Z_r]$ (white contour) is close to the true configuration X_0 and the variance of the distribution $p(X|Z_r)$ decreases with r, as expected. Particle set size is $N = 80$ per layer. (For clarity, only particles from the posterior accounting for at least 1% of likelihood over sample-set are shown.)

8.1 Importance reweighting

Layered sampling uses what we term "importance reweighting, in which the particles representing some prior distribution $p_0(X)$ are replicated and re-weighted. Particles are replicated to a degree that is proportional to the value of some weighting function $g(X)$, as in figure 20. Following the re-distribution, likelihood weights are adjusted to compensate, so that the particle-set continues to represent the same underlying prior p_0. The re-weighting operation is denoted by a $\sim$ operator with a weighting function. An example of its use follows:

$$\boxed{p_0} \xrightarrow[N]{} \bigcirc \xrightarrow[N]{\sim g} \bigcirc \xrightarrow{\times f} \bigcirc \xrightarrow[N]{\sim} \bigcirc.$$



prior | posterior: $r = 40$ pixels

posterior: $r = 20$ pixels | posterior: $r = 10$ pixels

Figure 19: **A broader prior "overloads" factored sampling** Now the demonstration of figure 18 is repeated, but with a prior 1.5 times as broad, causing sampling at the finest scale to break down (observe the large bias in the mean configurations at scale $r = 10, 20$ pixels. (Again, $N = 80$.)

This is factored sampling (9) with an extra, intermediate, reweighting stage. In terms of particle-sets, the reweighting operation $\sim g$ is defined as follows

$$\{(s^{(i)}, \pi_i),\ i = 1, \ldots, N\} \to \{(s^{(i(j))}, 1/g(s^{(i(j))})),\ j = 1, \ldots, N\}$$

where each $i(j)$ is sampled with replacement from $i = 1, \ldots, N$ with probability proportional to $\pi_i g(s^{(i)})$.

A useful property of the resampling operation $\sim g$ is that it is an *asymptotic identity*: as $N \to \infty$, the difference between the distributions of the two random variables generated by

$$\boxed{p_0} \xrightarrow[N]{} \bigcirc \xrightarrow[1]{\sim} \bigcirc \text{ and by } \boxed{p_0} \xrightarrow[N]{} \bigcirc \xrightarrow[N]{\sim g} \bigcirc \xrightarrow[1]{\sim} \bigcirc$$

converges weakly to 0.

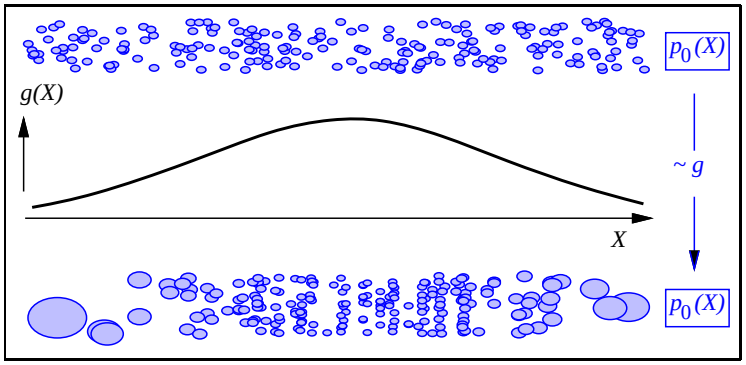


Figure 20: **Importance reweighting.** A uniform prior $p_0(X)$, represented as a particle-set (top), is resampled via an importance function g to give a new, re-weighted particle-set representation of p_0. (The illustration here is for a one-dimensional distribution, though practically X is multidimensional.)

Resampling with the $\sim g$ operation does not, on its own, deal with the problem of a narrow likelihood function. Although it does concentrate sampling to a narrower region of configuration space, the gaps between particles are as great as ever (figure 21). Gaps can be filled, however, by adding a further random variable

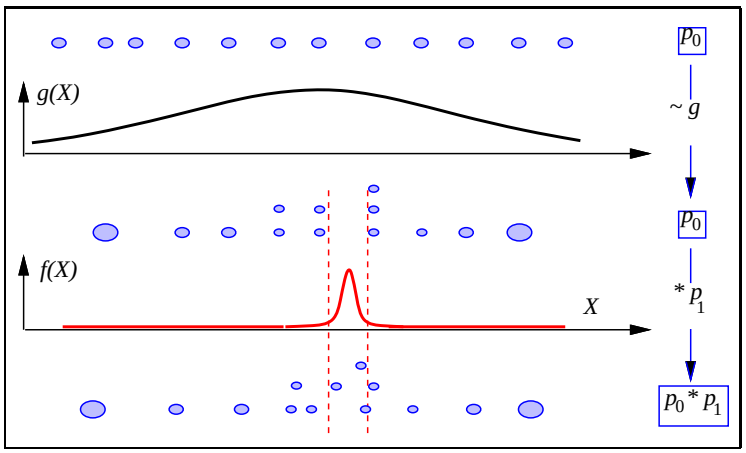
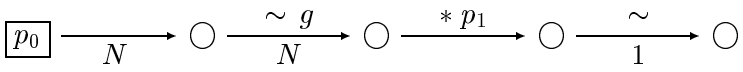


Figure 21: **Resampling followed by convolution.** This simplified example illustrates that importance reweighting on its own cannot repopulate the sparsely sampled support of the likelihood f. Repopulation *can* however be achieved by adding a random increment, corresponding to convolving the prior p_0 with p_1, the density of the random step.

with density p_1, to each particle. This has the effect of diffusing apart identical copies of particles generated in the resampling step. Of course, the combined operation is no longer an asymptotic identity — particles at the output of



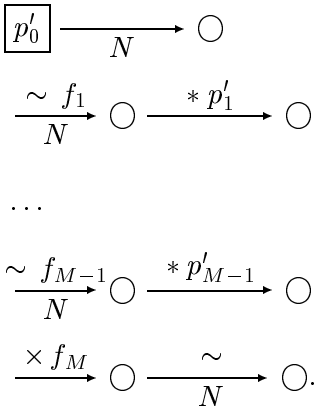
are distributed asymptotically according to the density $p_0 * p_1$.

8.2 The layered sampling algorithm

Layered sampling is applicable when importance resampling functions $f_1, \ldots, f_M$ are available, in which $f_M = f$ is the true likelihood, and each f_{m-1} is a coarse approximation to f_m. In addition, the prior p_0 must be decomposable as a series of convolutions

$$p_0 = p'_0 * p'_1 \ldots * p'_{M-1} \tag{15}$$

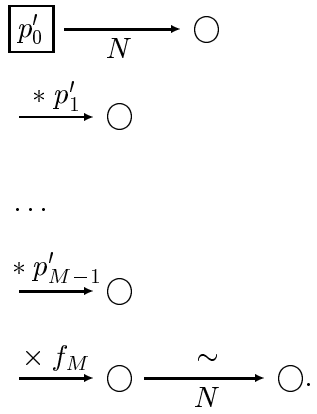
and this corresponds to expressing X *a priori* as a sum of random variables. Functional forms for the densities p'_m need not necessarily be known, provided only that a random sample generator can be constructed for each. For example, in processing motion sequences using the CONDENSATION algorithm (Isard and Blake, 1996), p'_0 could be represented as a set of particles from the previous time $t-1$, and $p_d = p'_1 \ldots * p'_{M-1}$ is some decomposition of a normal distribution $p_d(X(t)|X(t-1))$ for the likely displacement over one time-step, into normally distributed components. With this decomposition of the prior, the sampling process (9) on page 6 can be replaced by a sequence of layers:



(16)

Each layer includes an importance resampling step, with the observation likelihood f_i at the ith scale as the resampling function, until the Mth and final layer, at which the fine-scale f_M acts multiplicatively on likelihood weights, in the usual way.

The asymptotic correctness of layered sampling can be demonstrated by manipulating the sampling diagram. Using the asymptotic identity property of $\sim$, (16) can be rewritten, deleting resampling links, to give



and now the p'_m convolutions can be composed to give

$$\boxed{p'_0 * p'_1 * \ldots * p'_{M-1}} \xrightarrow[N]{} \bigcirc \xrightarrow{\times f_M} \bigcirc \xrightarrow[N]{\sim} \bigcirc.$$

which, from (15), and since $f_M = f$, reduces to the original factored sampling process (9).

8.3 Variance reduction

A remaining problem is how to choose the likelihood functions and the decomposition of p_d in such a way as to minimise the variance of the particle set generated in the final layer. These are complex problems in general, but some progress can be made by setting out the following special case.

1. The prior p'_0 is a rectangular distribution, with a support of volume a_0 in configuration space.
2. Each likelihood function f_m is idealised as a rectangular (uniform) distribution with a support of volume a_m.
3. The support of each f_m is a subset of the support of f_{m-1}.
4. Each p'_m is chosen in such a way that N particles are effectively uniformly distributed over the support of f_m, as depicted in figure 21. This can be done by matching the support of p'_{m-1} to the support of f_m.
5. Variance minimisation is not well-posed for rectangular distributions f_m, since their support is bounded. Instead, we minimise the "failure rate" — the probability that the particle set in some layer is empty.

Under these assumptions it can be shown (see appendix) that the failure rate is minimised by choosing

$$a_{m-1} = \lambda a_m \tag{17}$$

so that successive support volumes are in some fixed ratio λ.

Three further useful results (derivations omitted) can be obtained using analysis of estimator variance for importance sampling (Neal, 2000; Liu and Chen, 1995; Geweke, 1989).

- Using just a single layer (*i.e.* without layered sampling), the number N of particles required to achieve a given failure rate is
$$N \propto a_0/a_M \tag{18}$$
- With layered sampling, the failure rate is minimised by having approximately
$$M = \log_2(a_0/a_M) \tag{19}$$
layers. This means that $\lambda = 1/2$ is the optimal ratio of support volumes.
- With the optimal number M of layers, the total number of particles required falls to
$$NM \propto \log_2(a_0/a_M), \tag{20}$$
a logarithmic speed-up compared with (18).

9 Results

Layered sampling is applied here to the problem of multi-scale localisation. In all cases, a hexagonal tesselation of filters was used with separations of 6σ (sections 9.1, 9.2) or 3σ (sections 9.3, 9.4). [Recall that the support of the filters are truncated at $r = 3\sigma$; filter sizes are specified as r-values in experiments below.] A constant number N of particles was used in each layer; demonstrations with motion in section 9.4 were done with just a single layer, though clearly these also would be expected to benefit from multiple layers.

9.1 Sampling across scales

In the Bayesian localisation application, the f_m from the layered sampling algorithm correspond to observation likelihoods from the coarsest scale $m = 1$ to the finest $m = M$. Operation of the algorithm is illustrated here, in figure 22, for the hand-finding problem that caused the overloading of single-scale sampling earlier, in section 7.2. The normally distributed prior p_0 is split, as a sum of normal variables, into 3 factors

$$p_0 = p_0' * p_1' * p_2',$$

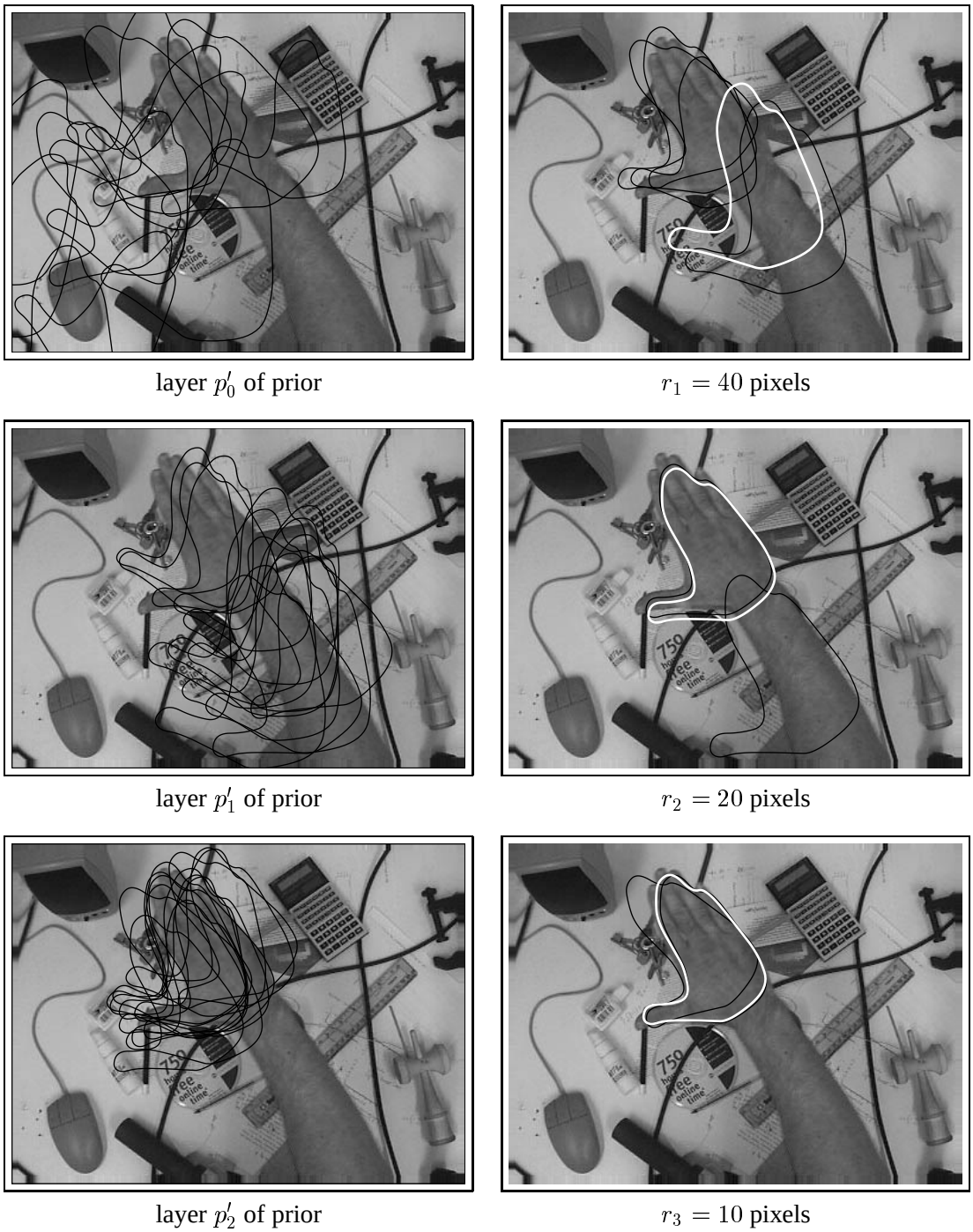
each factor to be used before scales $r_1,\ r_2,\ r_3$ in the coarse-to-fine hierarchy of observations. Scales are chosen to decrease geometrically, as implied by the fixed ratio rule (17) above. (This implication holds on the assumption that observation likelihood functions scale linearly with filter radius r, and demonstrations tend to support this, as in figure 17). The ith scale generates an observation likelihood function f_i, where $f_i(X) = p(Z_i|X)$. Note that the formal likelihood derives from observations only at the finest scale. Observations at other scales are cast by layered sampling in an "advisory" role, their scope limited to importance sampling for the next finer scale. This avoids any need for any formal assumption of statistical independence across scales which may be hard to justify.

9.2 Occlusion

One of the attractions of intensity-based matching is its robustness to disturbances in the image data, and a severe form of disturbance is presented by occlusion. Where occlusion is anticipated, this is addressed in the Bayesian localisation framework simply by treating the occluder as part of the background, and evaluating the appropriate observation-likelihood functions there. More challenging is occlusion that is not anticipated, as in figure 23. The figure illustrates the power of the Bayesian sampling approach to deal with ambiguity. At coarse scale, the part-occluded and blurred representation of shape leaves object-orientation quite ambiguous, though translation is somewhat constrained. Finer scales contain fragments of curve at sufficient resolution to register quite precisely with part of the object outline. Hence the rotational ambiguity is resolved. Even though the posterior at the finest scale has very small variance, nonetheless, the facility to represent ambiguity in the intermediate processes is what has allowed multi-scale information to be propagated effectively.

9.3 Pose variation

Bayesian localisation is capable of handling a configuration space $\mathcal{X}$ that incorporates varying 3D pose, as the demonstration of figure 24 shows. The foreground distributions in this demonstration were learned using

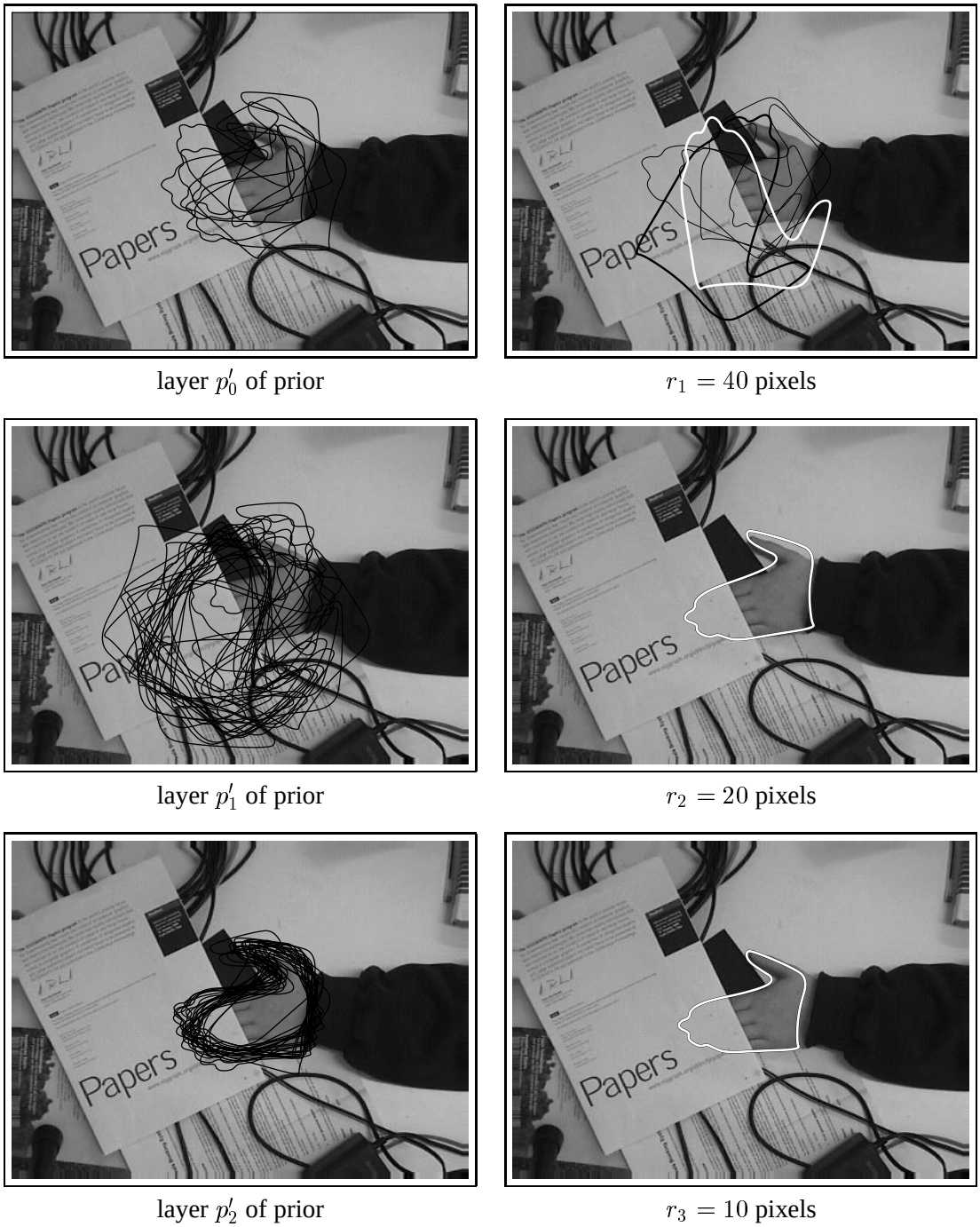


layer p_0' of prior | $r_1 = 40$ pixels

layer p_1' of prior | $r_2 = 20$ pixels

layer p_2' of prior | $r_3 = 10$ pixels

Figure 22: **Layered sampling across spatial scales**: the demonstration of figure 19 is repeated, but now with layered sampling, from coarse to fine scale. Note that the overload evident at finest scale in figure 19, is rectified here, with a similar computational load ($N = 80$ particles per layer).



layer p_0' of prior | $r_1 = 40$ pixels

layer p_1' of prior | $r_2 = 20$ pixels

layer p_2' of prior | $r_3 = 10$ pixels

Figure 23: **Layered sampling with occlusion**: a demonstration like the one in figure 22 but now with the object suffering unpredicted occlusion. Note that, at the coarsest scale, shape information is sufficiently distorted by occlusion, that object orientation is quite ambiguous in the posterior. Finer scales resolve the ambiguity.

Figure 24: **Pose variation**: the prior is approximately uniformly distributed (on the white rectangle) over translations, with normal distributions over pose and zoom. The first and last layers of the posterior from layered sampling with $r = 40, 20, 10$ pixels are shown, for each of three poses of a face. (Means displayed in white; $N = 250$ particles per layer, of which the 15 with highest likelihood are displayed in layer.)

foreground subdivision as discussed in section 6, with subregions of a diameter equal to that of the filter support. In fact, in the coarsest layer, there is space within the face contour for only one subregion, but 7 subregions at $r = 20$ and 33 at $r = 10$. Note the "rogue" face hypothesis appearing on the curtain at the left, which receives a significant weight in layer 1, at the coarsest scale (a blurry hallucination), but does not survive at fine scale.

A further demonstration of face-tracking, free-running at about 1 frame/sec, is given at

`http://www.robots.ox.ac.uk/~vdg/movies/bayes-face.mpg`.

In this case there are two layers with $r = 40, 20$ and $N = 600$ particles per layer, and a foreground intensity model is used, as in (Sullivan and Blake, 2000).

9.4 Motion tracking

Motion tracking demonstrations in this section serve two purposes. First they test the Bayesian localisation algorithm over many separate video frames. Second they underline the importance of Bayesian techniques for sequential inference. The prior for object configuration in each frame is predicted from the posterior for the previous frame, via a learned dynamical model (Blake et al., 1995; Baumberg and Hogg, 1995). The iterated process of prediction and Bayesian localisation forms a particle filter (Gordon et al., 1993; Kitagawa, 1996; Isard and Blake, 1996). A person walking across a room is tracked (figure 25) in the manner of Baumberg and

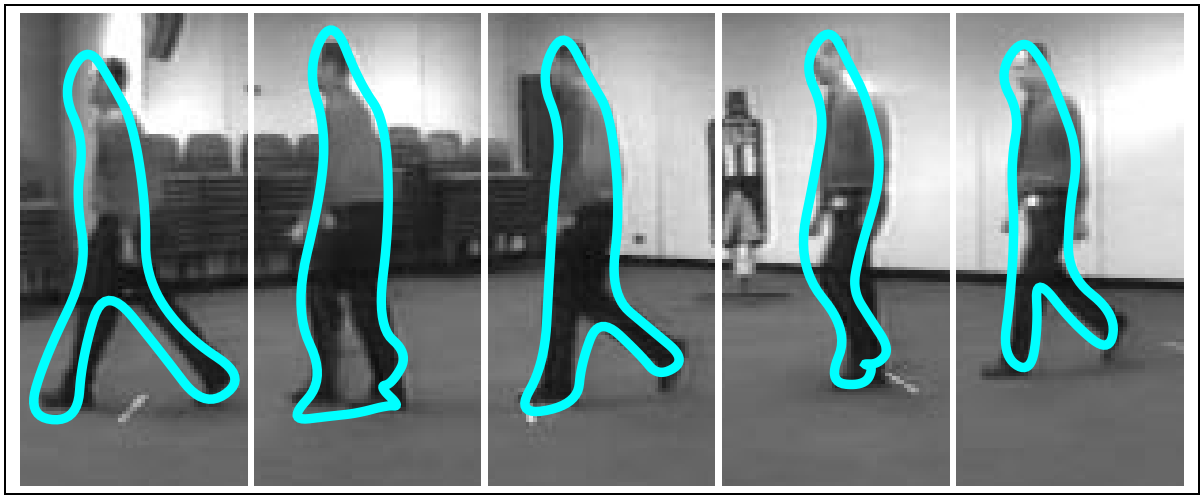


Figure 25: **Deformable motion.** A deformable contour model with 8 free parameters is used to track a walking person. The image sequence contains over 150 image frames. (We used a single layer with $r = 15$ pixels and $N = 1500$ samples.)

Hogg's tracking demonstration (1995), but *without* background subtraction. See also the movie version at

`http://www.robots.ox.ac.uk/~vdg/movies/bayes-walker.mpg`.

Instead, distracting background clutter is dealt with by the learned foreground/background models embodied in the observation likelihood. Consequently, the method not limited to backgrounds that are stationary, or moving in some easily predictable fashion.

A note should be added here on computation time. The task (on-line, excluding learning) here consists principally of image processing to obtain the z_k, and of computation of likelihood (14), of which the offset function $p_n(z_k|\rho_k(X))$ is main burden. The image processing can be done using pyramid filter banks (Burt, 1983) that are available in hardware. The offset function (at scale $r = 40$) can be computed for approximately $N = 500$ particles per time-step, at frame-rate. Bayesian localisation at video frame-rate is therefore quite feasible, in principle.

10 Conclusions

The original elements of Bayesian localisation are: the development of filter-based likelihood functions for matching with particular attention to statistical independence; learning of foreground and background distributions, and distributions for "mixed" receptive fields; probabilistic multi-scale analysis by means of "layered sampling".

The approach has been tested on a variety of foregrounds and backgrounds. It is capable of planar object localisation, even with unpredicted occlusion, and versatile enough to work with 3D pose changes, and with image sequences of moving objects, including nonrigid ones. A number of issues are raised: the choice of partition for the prior in layered sampling; the use of spatio-temporal filters and associated independence arguments; temporal updating of the foreground distribution. These remain for future investigation.

Acknowledgements

We are grateful for the support of the Royal Society of London (AB), EPSRC (AB,JS,MI) and the EU (JM). We have enjoyed and benefited from discussions with D. Mumford, S. Mallat, G. Hinton, B. Buxton, A. Zisserman and P. Torr.

References

Bartels, R., Beatty, J., and Barsky, B. (1987). *An Introduction to Splines for use in Computer Graphics and Geometric Modeling*. Morgan Kaufmann.

Bascle, B. and Deriche, R. (1995). Region tracking through image sequences. In *Proc. 5th Int. Conf. on Computer Vision*, pages 302–307, Boston.

Baumberg, A. and Hogg, D. (1995). Generating spatiotemporal models from examples. In *Proc. British Machine Vision Conf.*, volume 2, pages 413–422.

Belhumeur, P. and Kriegman, D. (1998). What is the set of images of an object under all possible illumination conditions. *Int. J. Computer Vision*, 28(3):245–260.

Bell, A. and Sejnowski, T. (1997). Edges are the independent components of natural scenes. In *Advances in Neural Information Processing Systems*, volume 9, pages 831–837. MIT Press.

Beymer, D. and Poggio, T. (1995). Face recognition from one example view. In *Proc. 5th Int. Conf. on Computer Vision*, pages 500–507.

Black, M. and Yacoob, Y. (1995). Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *Proc. 5th Int. Conf. on Computer Vision*, pages 374–381.

Blake, A. and Isard, M. (1998). *Active contours*. Springer.

Blake, A., Isard, M., and Reynard, D. (1995). Learning to track the visual motion of contours. *J. Artificial Intelligence*, 78:101–134.

Bookstein, F. (1989). Principal warps:thin-plate splines and the decomposition of deformations. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11(6):567–585.

Burt, P. (1983). Fast algorithms for estimating local image properties. *Computer Vision, Graphics and Image Processing*, 21:368–382.

Cootes, T., Taylor, C., Cooper, D., and Graham, J. (1995). Active shape models — their training and application. *Computer Vision and Image Understanding*, 61(1):38–59.

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

Gelfand, A. and Smith, A. (1990). Sampling-based approaches to computing marginal densities. *J. Am. Statistical Assoc.*, 85(410):398–409.

Geman, D. and Jedynak, B. (1996). An active testing model for tracking roads in satellite images. *IEEE Trans. Pattern Analysis and Machine Intell.*, 18(1):1–14.

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

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57:1317–1339.

Gordon, N., Salmond, D., and Smith, A. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc. F*, 140(2):107–113.

Grenander, U. (1976–1981). *Lectures in Pattern Theory I, II and III*. Springer.

Grenander, U., Chow, Y., and Keenan, D. (1991). *HANDS. A Pattern Theoretical Study of Biological Shapes*. Springer-Verlag. New York.

Grenander, U. and Miller, M. (1994). Representations of knowledge in complex systems (with discussion). *J. Roy. Stat. Soc. B.*, 56:549–603.

Hager, G. and Toyama, K. (1996). Xvision: combining image warping and geometric constraints for fast tracking. In *Proc. 4th European Conf. Computer Vision*, pages 507–517.

Isard, M. and Blake, A. (1996). Visual tracking by stochastic propagation of conditional density. In *Proc. 4th European Conf. Computer Vision*, pages 343–356, Cambridge, England.

Isard, M. and Blake, A. (1998). Condensation — conditional density propagation for visual tracking. *Int. J. Computer Vision*, 28(1):5–28.

Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5(1):1–25.

Liu, J. and Chen, R. (1995). Blind deconvolution via sequential imputations. *J. Am. Stat. Soc*, 90(430):567–576.

Mallat, S. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11:674–693.

Matthies, L., Kanade, T., and Szeliski, R. (1989). Kalman filter-based algorithms for estimating depth from image sequences. *Int. J. Computer Vision*, 3:209–236.

Mumford, D. (1996). Pattern theory: a unifying perspective. In Knill, D. and Richard, W., editors, *Perception as Bayesian inference*, pages 25–62. Cambridge University Press.

Neal, R. (2000). Annealed importance sampling. *Statistics and Computing*, in press.

Olshausen, B. and Field, D. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609.

Perona, P. (1992). Steerable-scalable kernels for edge detection and junction analysis. *J. Image and Vision Computing*, 10(10):663–672.

Ripley, B. (1992). Classification and clustering in spatial and image data. In Goebl, H. and Schader, M., editors, *Procs. 15 Jahrestagung von Gesellschaft fur Klassifikation*. Springer-Verlag.

Scharstein, D. and Szeliski, R. (1998). Stereo matching with nonlinear diffusion. *Int. J. Computer Vision*, 28(2):155–174.

Shirai, Y. and Nishimoto, Y. (1985). A stereo method using disparity histograms and multi-resolution channels. In *Proc. 3rd Int. Symp. on Robotics Research*, pages 27–32.

Storvik, G. (1994). A Bayesian approach to dynamic contours through stochastic sampling and simulated annealing. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 16(10):976–986.

Sullivan, J. and Blake, A. (2000). Satistical foreground modelling for object localisation. In *Proc. European Conf. Computer Vision*, volume 2, pages 307–323.

Sullivan, J., Blake, A., Isard, M., and MacCormick, J. (1999). Object localisation by bayesian correlation. In *Proc. 7th Int. Conf. on Computer Vision*, pages 1068–1075.

Szeliski, R. (1990). Bayesian modelling of uncertainty in low-level vision. *Int. J. Computer Vision*, 5(3):271–301.

Vetter, T. and Poggio, T. (1996). Image synthesis from a single example image. In *Proc. 4th European Conf. Computer Vision*, pages 652–659, Cambridge, England.

Viola, P. and Wells, W. (1993). Alignment by maximisation of mutual information. In *Proc. 5th Int. Conf. on Computer Vision*, pages 16–23.

Witkin, A., Terzopoulos, D., and Kass, M. (1987). Signal matching through scale space. *Int. J. Computer Vision*, 1(2):133–144.

Zhu, S. and Mumford, D. (1997). GRADE: Gibbs reaction and diffusion equation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(11):1236–1250.

Zhu, S., Wu, Y., and Mumford, D. (1998). Filters, random fields and maximum entropy (FRAME). *Int. J. Computer Vision*, 27(2):107–126.

A Layered sampling and bounded variance

The result from section 8 about arranging the scales of successive likelihood functions in fixed ratio is derived here. Making the assumptions 1–5 from section 8.3, the density of particles on entering the mth layer in (16) is N/a_{m-1}, assumed uniformly distributed in configuration space. Then the proportion of these particles which lies within the support of f_m has mean

$$\lambda_m = \frac{a_m}{a_{m-1}}$$

and is binomially distributed. The probability $P(F_m)$ of "failure" at the mth layer is therefore

$$P(F_m) = (1 - \lambda_m)^N$$

and the event $F = F_1 \cup \ldots \cup F_M$ of failure at any layer has probability

$$P(F) = 1 - \prod_{i=1}^{M} \left(1 - (1 - \lambda_m)^N\right).$$

Now minimising $P(F)$ under the constraints that $\mu_i \geq 0$ and the constraint (imposed using a Lagrange multiplier) that the product

$$\prod_{i=1}^{M} \lambda_i = \frac{a_M}{a_0}$$

is a constant, gives a unique solution

$$\lambda_1 = \lambda_2 = \ldots = \lambda_M,$$

so that the ratios a_m/a_{m-1} are all equal, as required.