Object Localization by Bayesian Correlation

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/

Abstract

Maximisation of cross-correlation is a commonly used principle for intensity-based object localization that gives a single estimate of location. However, to facilitate sequential inference (eg over time or scale) and to allow the representation of ambiguity, it is desirable to represent an entire probability distribution for object location. Although the cross-correlation itself (or some function of it) has sometimes been treated as a probability distribution, this is not generally justifiable.

Bayesian correlation achieves a consistent probabilistic treatment by combining several developments. The first is the interpretation of correlation matching functions in probabilistic terms, as observation likelihoods. Second, probability distributions of filter-bank responses are learned from training examples. Inescapably, response-learning also demands statistical modelling of background intensities, and there are links here with image coding and Independent Component Analysis. Lastly, multi-scale processing is achieved, in a Bayesian context, by means of a new algorithm, **layered sampling**, *for which asymptotic properties are derived.*

1 Introduction

Object localization in an image $I(\mathbf{x})$ can be viewed as the problem of recovering the warp $g_X(\mathbf{x})$ that transported a certain template $T(\mathbf{x})$ into the image. Here, the warp g_X is parameterised by $X \in \mathcal{X}$, where $\mathcal{X}$ is a configuration space for the warped template, for example, planar-affine space or some space of non-rigid deformations. Following the warp, it is assumed that random imperfections are introduced as a result of sensor-noise and unmodelled variations.

"Analysis by synthesis" [14] then consists of the Bayesian construction of a posterior distribution for X. Given a prior distribution $p_0(X)$ for the configuration X, and an observation likelihood $p(Z|X)$ where $Z \equiv Z(I)$ is some finite-dimensional representation of the image I, then the posterior density for X is given by

$$p(X|Z) \propto p_0(X)p(Z|X). \tag{1}$$

In more straightforward, Gaussian cases, (1) can be computed in closed form. In the non-Gaussian cases commonly arising, for example in image clutter or with multiple models, sampling methods are needed [8], and random sampling underlies the development of Bayesian correlation here.

Relation to previous work Key elements of the work presented here are:

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 searching of images.

PD *Posterior Distributions* for object location, rather than just a single estimate, supports sequential reasoning for multi-scale and image-sequence analysis, and potentially across sensory modalities.

BM *Background Modelling:* in a valid Bayesian analysis, image observations Z must not be a function $Z(X)$ of the hypothesis X. For example, sum-squared difference violates this principle by considering only the portion of an image directly under the template $T(\mathbf{x})$. A Bayesian approach must use evidence about where the object is *not*. That requires a probabilistic model of the image background.

SI *Statistical Independence* of observations must be ensured if constructed observation likelihoods are to be valid. For instance, assuming independence across adjacent pixels is unjustified, and leads to exaggerated variations in the likelihood $p(Z|X)$ for even minor perturbations of X.

There are three outstanding precursors to Bayesian correlation; one concerns random diffeomorphisms [8]; the second is an algorithm [17] for registration by maximisation of mutual information; third is localisation by foreground/background learning [7]. Attributes of these and other important prior work are summarised in table 1, in terms of elements of Bayesian correlation as listed above.

2 Probabilistic inference of shape

A natural choice for the set Z of image observations is a filter-bank consisting of inner-product elements z_k, $k = 1, \ldots, K$ applied to the image I. Each filter-element has the form

$$z_k = \int_{S_k} W_k(\mathbf{x})I(\mathbf{x})\mathrm{d}\mathbf{x}, \tag{2}$$

* for a version of this paper with colour figures and a movie of figure 15

	IB	FL	MS	PD	BM	SI	**Comments**
Burt [5]	×		×				multi-scale pyramid
Witkin *et al.* [18], Scharstein & Szeliski [16]	×		×				scale-space matching
Grenander *et al.* [8]	×			×	×		random diffeomorphisms
Viola & Wells [17]	×	×					mutual information
Cootes *et al.* [6]	×	×	×				multi-scale active contours
Black & Yacoob [3], Bascle & Deriche [1], Hager & Toyama [9]	×	×					affine flow/warp
Isard & Blake [11]		×		×			random active contours
Olshausen & Field [15], Bell & Sejnowski [2]	×				×	×	independent components (ICA)
Geman & Jedynak [7]	×	×			×		response learning

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

computing an inner product of the image and the element function W_k, over a finite support S_k. Element functions may consist of copies of a single response-function $W(\mathbf{x})$, translated to the nodes of a regular grid, so that the world is effectively being viewed through a sieve, as in figure 1. In the familiar case that the space of warps $\mathcal{X}$ consists of

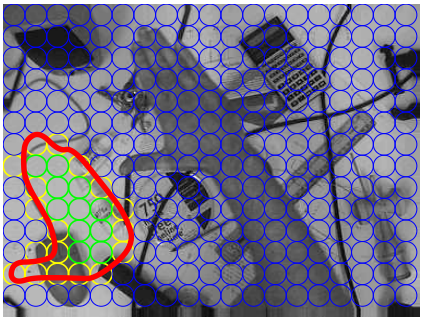


Figure 1: **The world through a filter bank** $Z = (z_1, \ldots, z_K)$, with circular supports $S_1, \ldots, S_K$ on a regular grid. Given some hypothesised X (thick hand-shaped outline), supports are labelled foreground (green), background (blue) or mixed (yellow). (Note: example shows an X that is far from the true hand configuration.)

two-dimensional translations, the bank can be thought of as a discrete sampling of the cross-correlation of W with I. In that case, the response-function $W(\mathbf{x})$ could well be a translated copy of the object template $T(\mathbf{x})$, which would have the effect of tuning $z(X)$ to respond to object position. For the higher-dimensional warp-spaces $\mathcal{X}$ (e.g. planar affine) that we want to deal with here, systematic sampling of $z(X)$, $X \in \mathcal{X}$ is no longer feasible. Generalising the filter bank Z therefore has to take a different tack. The two-dimensional grid layout can remain, but the response-function W becomes something more general. The translated copies W_k generate a set of linear functionals to encode (partially) an image I, with the necessary statistical independence, but no longer tuned to any particular object. Of course, an important generalisation is that there may be more than one type of response-function (eg for various scales), each of which is replicated over the grid to form the components z_k of Z. The entire filter-bank scheme has the attraction that fixed, computationally efficient architectures can be used to compute Z, for instance wavelets [13], pyramids [5] or biological "hypercolumn" hardware [10].

Learning We have argued that $p(Z|X)$ contains both foreground and background components. Tackling image-texture modelling head-on would be complex. An oblique approach is to learn filter-likelihoods $p(z_k|X)$ directly from training images, as [7] but, crucially, also tackling the inescapable issue of *mixed* supports (figure 1). This side-steps any need for a complete model of foreground or background, modelling them *only as they appear in the sieve* of figure 1. Then, provided also that the W_k can be chosen to give the necessary statistical independence, the full observation likelihood can be constructed as a product:

$$p(Z|X) = \prod_{k=1}^{K} p(z_k|X). \tag{3}$$

Factored sampling For non-Gaussian problems, (1) can be simulated by generating random variates from a distribution that approximates the posterior $p(X|Z)$. In *factored sampling* [8], 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"

$$\boxed{p_0} \xrightarrow[N]{} \bigcirc \xrightarrow{\times f} \bigcirc \xrightarrow[N]{\sim} \bigcirc. \tag{4}$$

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: $(s_i, \pi_i) \to (s_i, f(s_i)\pi_i)$, $i = 1, \ldots, N$. 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 equal weight, and which is drawn approximately from the posterior.

3 Probabilistic modelling of observations

The observation (ie output value) z from an individual filter is generated by integration over a support-set S (figure 2) and generally has both a background component $B(X)$ and a foreground component $F(X)$:

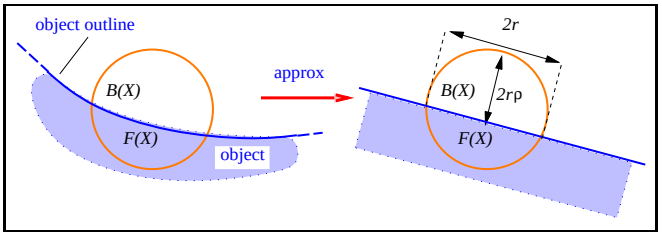


Figure 2: **The support of a filter** is split into subsets $F(X)$ — foreground and $B(X)$ — background. The boundary between subsets is approximated as a line, so $B(X)$ and $F(X)$ are segments of a circle with offsets $2r\rho$ and $2r(1-\rho)$ respectively.

$$z|X = \underbrace{\int_{B(X)} W(\mathbf{x}) I_B(\mathbf{x})\,\mathrm{d}\mathbf{x}}_{\text{MAIN NOISE SOURCE}} + \int_{F(X)} W(\mathbf{x}) I_F(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{5}$$

Then, considering the output z_k of the kth filter at run-time, under a hypothesised configuration X, the Bayesian correlation algorithm needs to compare the measured z_k against the likelihood $p_k(z_k|X)$. Likelihood $p_k(z|.)$ represents a sum of background and foreground components, and is therefore constructed as a (numerically approximated) convolution $p_k(z|X) = p_k^B(z|X) * p_k^F(z|X)$ of learned background and foreground density functions.

The main source of variability in $z|X$ is expected to come from the background which is a sample from some class of scenes, assumed large and only generally known. In contrast, the foreground relates to a given object, relatively precisely known, though still subject to some ambient- and class-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 over foreground, via mixed foreground/background, to pure background. This is supported by experiments shown later. Distributions $p(z|X)$ are learned for fixed values of X and effectively assembled and sliced to give observation likelihoods (figure 3). For example, $z = 2$ in the figure depicts a relatively high value which, in the example, is more likely to be associated with a filter-support lying mainly over the foreground. The resulting likelihood is peaked around a value of X corresponding to predominant foreground support. Conversely, for $z = -1$, the mode of the likelihood shifts towards background values of X.

4 Learning the background likelihood

It proves efficient to approximate the curve dividing $F(X)$ and $B(X)$ as a straight line (figure 2). In that case, if each filter functional $W_k(\mathbf{x})$ is isotropic the background distribution $p^{\mathcal{B}}$ can be parameterised by a single offset parameter ρ, so that the background density $p_k^B(z_k|X)$ for the kth filter takes a simpler form as $p^{\mathcal{B}}(z_k|\rho_k(X))$. Then, at run-time, the Bayesian correlation algorithm will repeatedly evaluate the *offset function* $\rho_k(X)$ in order to evaluate likelihood.

Now training examples must be constructed over circular segments with offsets throughout the range $0 \le \rho \le 1$, to learn the distributions $p^{\mathcal{B}}(z|\rho)$. (In practice, ρ-values are sampled and interpolated.) The $p^{\mathcal{B}}(z_k|\rho_k(X))$ should be independent so that a joint likelihood can be constructed, aggregating all observations z_k (3). Independence is an issue also in "neural coding" [15]: efficient codes that avoid redundancy need statistically independent components. Independent components of natural scenes are known to have "kurtotic" or "sparse" distributions — ones with extended tails compared with those of a normal distribution [2]. A necessary condition for independence is freedom from correlation, so autocorrelation was estimated, for four different scenes (desk-top, rooms, tree), by random sampling of pairs of supports, with varying separation. This was done for two filter functions $W(\mathbf{x})$: Gaussian $G(\mathbf{x})$ (positive mean) and Laplacian of Gaussian $\nabla^2 G(\mathbf{x})$ (zero mean), where $G(\mathbf{x}) = \frac{1}{\sigma^2}\exp -\frac{|\mathbf{x}|^2}{2\sigma^2}$, in a circular support of radius $r = 3\sigma$, as in figure 4. As expected, the $G(\mathbf{x})$ filter is correlated at a reasonable displacement such as $2r$ and hence cannot be independent. The $\nabla^2 G(\mathbf{x})$ filter is uncorrelated at $2r$ and further experiments, looking at the entire joint distributions for responses z_k, z_l of two filters with variable separation, confirm statistical independence. The independence is at the cost of

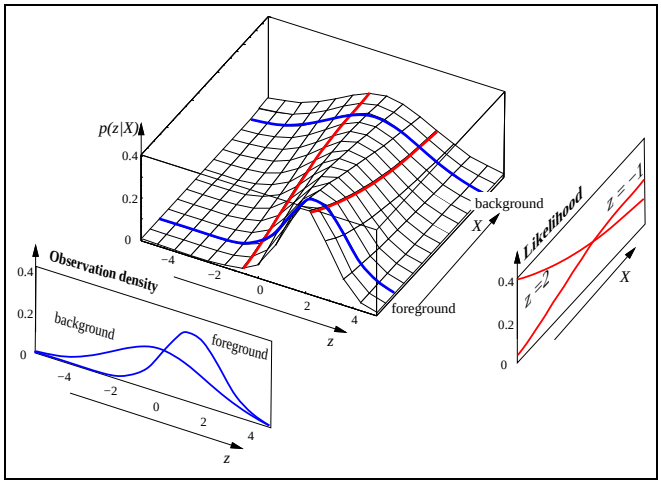


Figure 3: **Observation likelihood.** The density $p(z|X)$ is formally a function of z with X as a parameter, and is illustrated for foreground and background cases. Now $p(z|X)$ is "sliced" in the orthogonal direction, to generate likelihoods — functions of X for fixed z

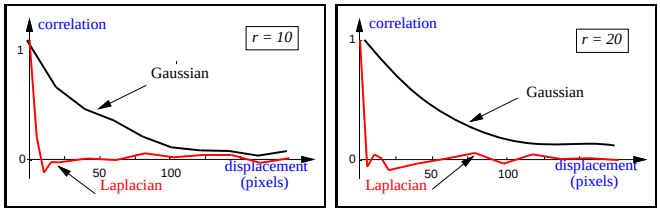


Figure 4: **Autocorrelation of filter output.** Results for the desk-top scene, at two spatial scales r. The Gaussian filter $G(\mathbf{x})$ shows substantial long-range correlation whereas, for $\nabla^2 G(\mathbf{x})$ correlation falls to zero for non-overlapping supports.

discarding information about mean response, but this can be beneficial in conferring some invariance to illumination variations. Experiments so far have been for complete, circular supports. With part-segments of a circle ($\rho < 1$), statistical independence of $\nabla^2 G(\mathbf{x})$ responses deteriorates. This is established by experiments like the ones in figure 4, but now with $\rho < 1$, that show correlation lengths increasing for $\rho < 1$, with $\rho = \frac{1}{4}$ the worst case. This means greater statistical dependence between mixed supports, and it is not clear how this could be improved, but note at least that it is typically a minority of filter supports that are mixed.

It is known that, for ∇G filters, the learned background distributions turn out to be strikingly constant across scenes [19]. Our own experimentation confirms that this holds also for $\nabla^2 G(\mathbf{x})$ filters and that the distribution is quite well modelled as a single-exponential distribution $p^{\mathcal{B}}(z) \propto \exp -|z|/\lambda$, like those emerging in independent components of images [2] and from maximum entropy arguments [20]. The model fits experimental data quite well for $\rho = 1$, though not so well for mixed supports $\rho < 1$, and could be used directly to represent background density, rather than carrying entire histograms.

5 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. There are some important differences however.

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 frame, but from a set of frames containing a succession of typical transformations of the object. These frames may either be separately captured images, or be generated by applying random deformations to one image.

Outline constraint: $p^{\mathcal{B}}(z|\rho)$ for $0 < \rho \leq 1$ was learned from segments dropped down at random, anywhere on the background. Over the foreground, and for 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 foreground support $F(X)$ must touch the object outline; therefore $p^{\mathcal{F}}(z|\rho)$ is learned entirely from segments abutting the outline.

Foreground subdivision: Learning $p^{\mathcal{F}}(z|\rho)$ by pooling responses throughout the object interior is effective with distinctive object outlines (eg hand), but pooling does discard information concerning gross spatial layout. Gross layout can be captured by sub-dividing the object (figure 5) and pooling separately over each subregion. This is especially beneficial with more nearly

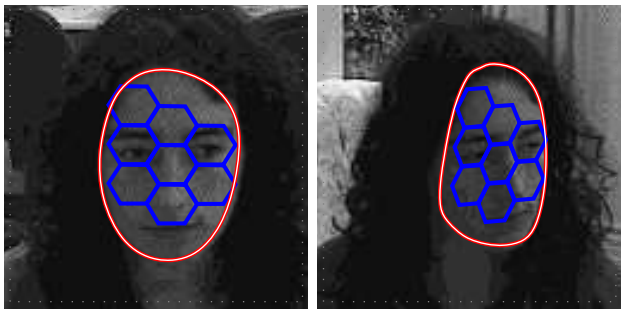


Figure 5: **Foreground subregions.** Object interior $\mathcal{F}$ is subdivided (left): $\mathcal{F} = \mathcal{F}_0 \cup \mathcal{F}_1 \cup \ldots \cup \mathcal{F}_{N_F}$, where sub-regions $\mathcal{F}_1, \ldots, \mathcal{F}_{N_F}$ here are hexagons, and $\mathcal{F}_0$ is the remainder of $\mathcal{F}$. Then sub-regions must be warped (right) onto any novel view.

circular objects (eg faces) for which, if isotropic filters are used, the observation likelihood is insensitive to 2D rotation of the object.
Sub-regions are defined for a standard configuration (figure 5a)); for a general configuration, warped forms of $\mathcal{F}_i$ are needed (figure 5b)). An affine approximation to the interior warp is obtained by projection in function-space [4, ch 6] and approximation error is dealt with simply by pooling it into the learned distributions $p^{\mathcal{F}_i}$.

Statistical independence: known behaviour for independence of natural scenes, which applied well to background modelling, could not necessarily be expected to apply for foreground models, given that the foreground is far less variable. Nonetheless, repeating the autocorrelation experiments has produced evidence of good independence for $\nabla^2 G$ filters over the foreground too.

Distribution model: whereas filter response z over (highly variable) background texture assumed the characteristic kurtotic form, the foreground is far less variable and therefore does not have extended tails (figure 6).

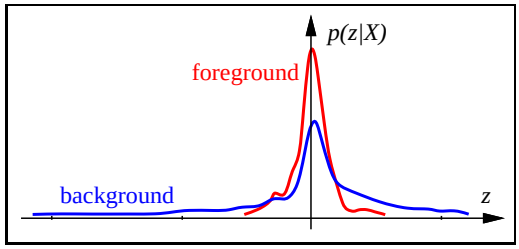


Figure 6: **Foreground and background distributions for $\nabla^2 G$ filter**, with support radius $r = 20$ pixels. As expected, the background distribution is more "kurtotic".

Results: learned observation likelihood First, for the hand scene of figure 1, $p(Z|X)$ — the joint likelihood composed of a product (3) of likelihoods $p(z_k|X)$ for individual filters — is exercised systematically, over a configuration space of Euclidean similarities (figure 7). The joint likelihood fuses information from individual supports effectively,

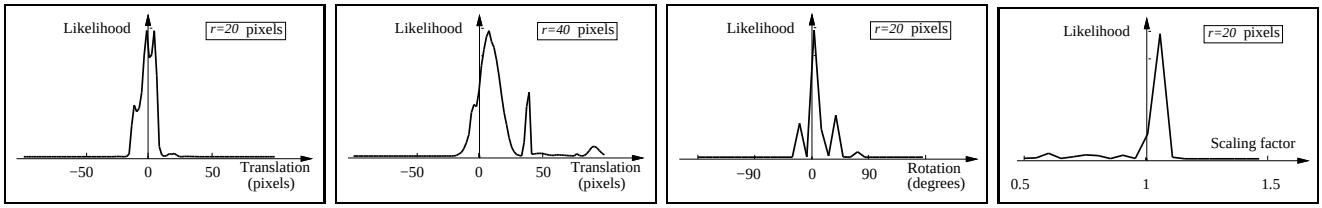


Figure 7: **Exercising the joint likelihood** $p(Z|X)$, as X ranges over coordinate translation (at 2 different scales), scaling and rotation.

with a maximal value, as expected, near the true solution (graph origin). The effect of changing the filter scale r is also demonstrated; as expected, the likelihood function is more broadly tuned, at a coarser scale, appearing to have a width of about $2r$ or (due to hyperacuity effects) rather less.

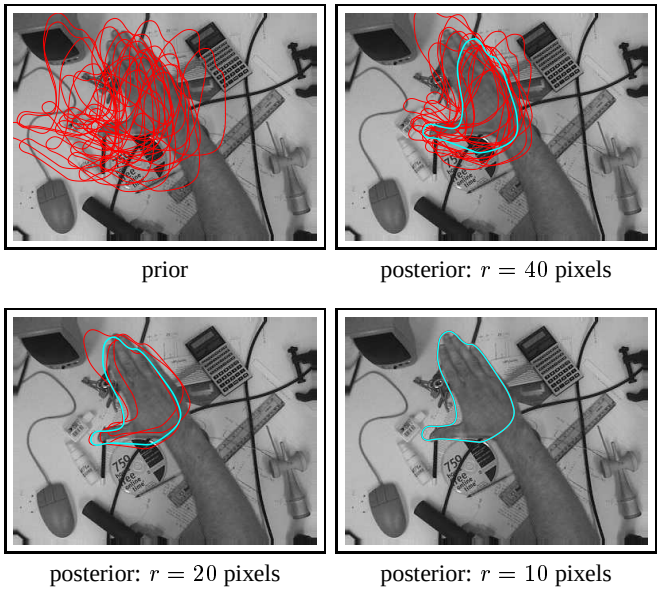
As a final check, it is interesting to consider the likelihood ratio for two configurations, one correctly positioned over the target, and one way out over background as in figure 1. In such cases, treating pixels as independent typically produces ridiculously large likelihood ratios. Even using Gaussian masks ($r = 20$), which we know not to be independent, gives a likelihood ratio of $1 : 10^{55}$ in this case — implausibly large. However, this falls considerably with $\nabla^2 G$ masks, as expected given the independence of their output over foreground and background, to a far more plausible $1 : 10^4$

Sampling from the posterior The full joint likelihood function $p(Z|X)$ is constructed as a product (3), in which the offset ρ for each support segment is obtained from its offset function $\rho_k(X)$:

$$p(Z|X) = \prod_{k=1}^{K} p_k(z_k|\rho_k(X)). \qquad (6)$$

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 (6), and similarly for any support always in the foreground.

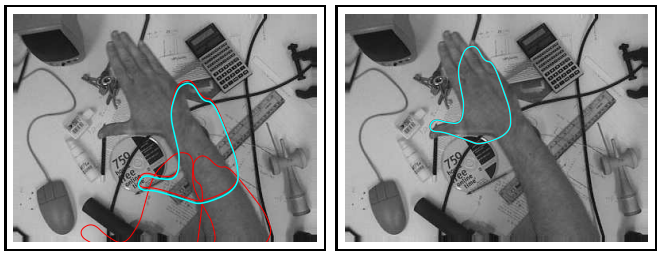
The practical application of Bayesian correlation is to problems involving the localization of objects. For example, to locate a hand against a cluttered background, a prior $p_0(X)$ is chosen over the space of Euclidean similarities. Samples from the posterior, at several scales, are shown in figure 8. The broad prior is focused down to a posterior distribution which is narrower at finer scales. It is not clear from figure 8 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 finer scales break down, as figure 9 shows, while at coarse scale, sampling from the posterior continues to operate correctly. That suggests a role for coarser scales



prior | posterior: $r = 40$ pixels

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

Figure 8: **Random samples from the posterior.** $p(X|Z)$. The prior $p_0(X)$ is a broad distribution of Euclidean similarities (planar rigid motion plus size-scaling). At each scale r, the posterior mean $\mathcal{E}[X|Z_r]$ (blue) is close to the true configuration; $\mathrm{Var}[X|Z_r]$ decreases with r, as expected. Particle set size is $N = 240$ here. (For clarity, not all particles are shown.)



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

Figure 9: **A broader prior "overloads" factored sampling.** Now the experiment of figure 8 is repeated, but with a prior 1.5 times as broad, causing sampling at these finer scales to break down. (Again, $N = 240$.)

in guiding or constraining finer ones, if only a Bayesian sampling mechanism can be found to do it, and that is the subject of the next section.

6 Layered sampling

In section 5, the practical problem of "overloading" was demonstrated, that occurs when image observations are made

at fine spatial scale, in Bayesian correlation. Layering, introduced here, is a powerful general strategy for reducing computational complexity of factored sampling when the observation likelihood function $f(X)$ is narrow. Layered sampling proves effective in dealing with the problem of multi-scale overloading.

Importance resampling Layered sampling uses what we term "importance resampling", in which the particles representing some prior distribution $p_0(X)$ are replicated and re-weighted (but, unlike conventional importance sampling [12], none are generated in new configurations). Particles are replicated to a degree that is proportional to the value of some weighting function $g(X)$, denoted $\sim g$ in the top half of figure 10. Following the re-distribution, likelihood weights are

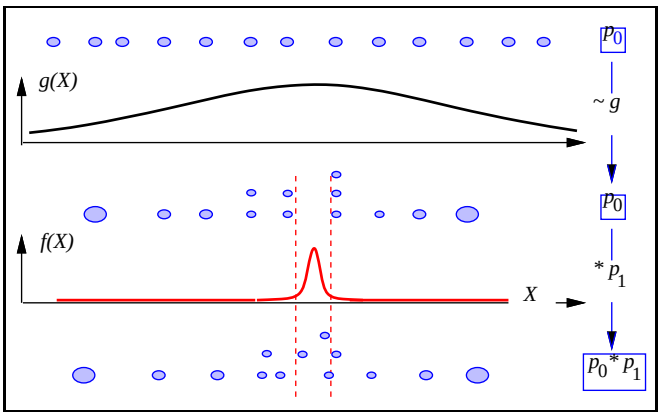


Figure 10: **Resampling followed by convolution.** A simplified example to illustrate that importance resampling ($\sim g$) on its own may not repopulate the sparsely sampled support of the likelihood f. A subsequent random step, with distribution p_1, is needed.

adjusted to compensate, so that the particle-set continues to represent the same underlying prior p_0. The re-sampling operation is denoted by a $\sim$ operator with a weighting function g, as in the following example of factored sampling (4) with an extra, intermediate, weighted resampling stage:

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

In terms of particle-sets, the resampling operation $\sim g$ is defined as follows

$$\{(s_i, \pi_i),\ i = 1..M\} \rightarrow \{(s_{i(j)}, 1/g(s_{i(j)})),\ j = 1..N\}$$

where each $i(j)$ is sampled with replacement from $i = 1..M$ with probability proportional to $\pi_i g(s_i)$. (Note that in the original factored sampling example (4), we had $M = N$.)

A key property of the resampling operation $\sim g$ is that it is an *asymptotic identity*: random resampling from its input and output particle sets, respectively, produce random variables whose distributions converge to one another, weakly as $N \rightarrow \infty$.

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 not reduced (figure 10). Adding independent random variables with density p_1 to each particle has the effect of diffusing apart identical copies of particles generated in the resampling step and filling the gaps. The combined operation is no longer an asymptotic identity — particles at the output are distributed asymptotically according to the density $p_0 * p_1$.

The layered sampling algorithm Layered sampling is applicable when importance resampling functions $f_1, \ldots, f_M$ are available, in which $f_M = f$ 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

$$p_0 = p'_0 * p'_1 \ldots * p'_{M-1} \tag{7}$$

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 [11], 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 Gaussian model $p_d(X(t)|X(t-1))$ for the likely displacement over one time-step. With this decomposition of the prior, the sampling process (4) can be replaced by a sequence of layers:

$$\begin{aligned} &\boxed{p'_0} \xrightarrow[N]{} \bigcirc \\ &\xrightarrow[N]{\sim f_1} \bigcirc \xrightarrow{* p'_1} \bigcirc \ldots \xrightarrow[N]{\sim f_{M-1}} \bigcirc \xrightarrow{* p'_{M-1}} \bigcirc \\ &\xrightarrow{\times f_M} \bigcirc \xrightarrow[N]{\sim} \bigcirc, \end{aligned} \tag{8}$$

where $* p$ denotes the particle-set operation $(s_i, \pi_i) \rightarrow (s_i + Y_i, \pi_i)$, $i = 1, \ldots, N$, and Y_i are random variables drawn independently from $p(.)$. 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.

Proof of asymptotic correctness The diagrammatic form of specification of the particle filter facilitates the proof of asymptotic correctness — that each particle in the output set is drawn from a distribution that converges (weakly) to the posterior, as $N \rightarrow \infty$. Asymptotically (using the identity property of $\sim$), (8) can be rewritten, deleting resampling links, to give

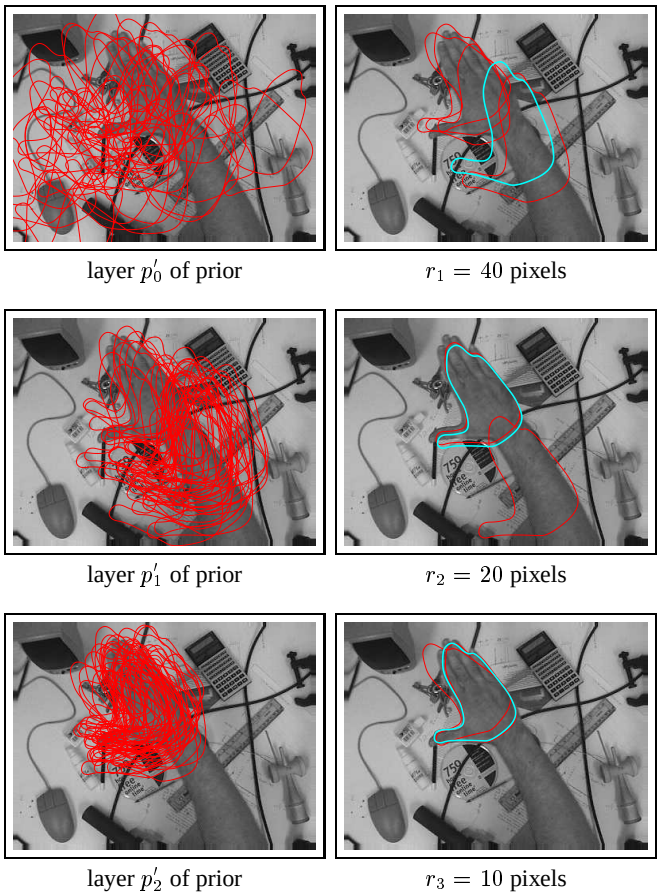
$$\begin{aligned} &\boxed{p'_0} \xrightarrow[N]{} \bigcirc \xrightarrow{* p'_1} \bigcirc \ldots \xrightarrow{* p'_{M-1}} \bigcirc \\ &\xrightarrow{\times f_M} \bigcirc \xrightarrow[N]{\sim} \bigcirc \end{aligned}$$

and now all the p'_m convolutions can be composed to give p_0, as in (7), and since $f_M = f$, the process reduces to the original factored sampling (4). There remains the issue of how to choose the decomposition of p_d. A good argument can be made (details omitted) that, in order to minimise N, successive spatial scales should be in fixed ratio; further work is needed to generalise this.

7 Results

Layered sampling is applied here to the problem of multi-scale localization and pose determination.

Sampling across scales 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 11, for the hand-finding problem that caused the overloading of single-scale sampling in section 5. The Gaussian prior p_0 is split, as a sum of Gaussian variables, into 3 factors $p_0 = p'_0 * p'_1 * p'_2$, each factor to be used before respective scales r_1, r_2, r_3 which diminish in fixed ratio for maximum efficiency, as mentioned earlier. 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, via importance resampling before the next finer scale. This avoids any need for a formal assumption of independence across scales.



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 11: **Layered sampling across spatial scales** The experiment of figure 9 is repeated, but now with layered sampling, from coarse to fine scale. The overload evident in figure 9 is rectified here, with similar computational effort ($N = 240$ particles, $N/3 = 80$ particles per layer).

Occlusion One of the attractions of correlation 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 dealt with in the Bayesian Correlation 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 12.

prior | $r_3 = 10$ pixels

Figure 12: **Layered sampling with occlusion** An experiment like the one in figure 11 but now with the object suffering unpredicted occlusion (intermediate scales not displayed).

Pose variations Bayesian correlation is capable of handling a configuration space $\mathcal{X}$ that incorporates varying 3D pose, as the demonstration of figure 13 shows. The foreground distribution is learned using pooling over $\mathcal{X}$ and foreground subdivision, as discussed in section 5.

Figure 13: **Pose variations** A foreground distribution was trained on 3 training images. Test images here show the posterior from broadly distributed priors, under variation of pose.

Motion tracking Random sampling lends itself to serial Bayesian inference, for example over multiple scales as above. Serial inference can also proceed over time, in order to analyse motion sequences. Edge-based temporal analysis [11] requires fairly precise initial alignment whereas, in Bayesian correlation, use of intensity information allows a degree of automatic initialisation, as in figure 14, even against camouflage, and despite the vigour of the motion. We are not yet sure, though, whether Bayesian correlation can align as precisely as edge-based analysis can.

Finally, an example is shown of motion analysis for deforming objects. A person walking across a room is tracked (figure 15) *without background subtraction*. Instead, distracting background clutter is dealt with by the learned foreground/background models embedded in the observation

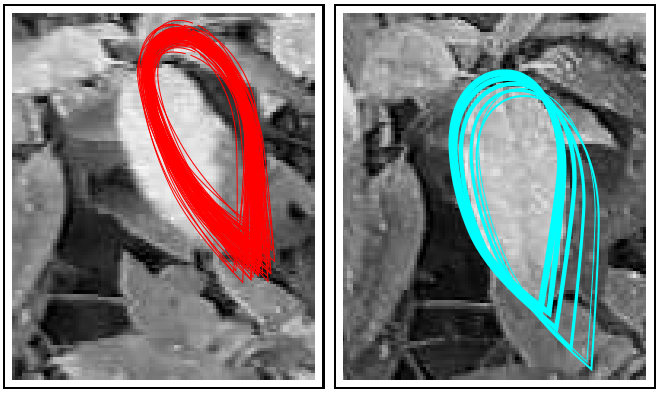


Figure 14: **Motion analysis:** leaf blowing vigorously, in camouflage. A prior (a) is chosen that is too badly misaligned for edge-based tracking. Bayesian correlation nonetheless initialises correctly and, 1120 ms later, is still tracking (b) — trail of mean shapes shown. (Data and learned motion model as in [11]; $r = 10$ pixels and $N = 1500$ samples.)

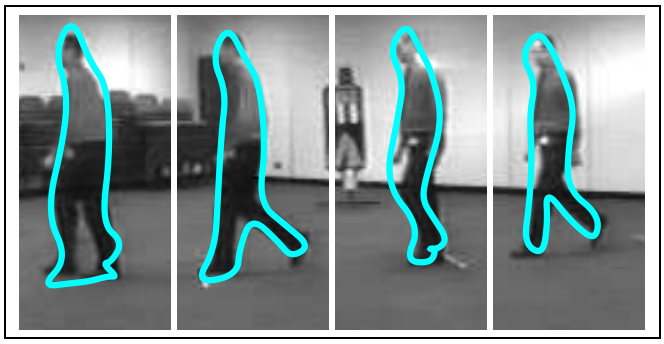


Figure 15: **Deformable motion.** A deformable contour model with 8 free parameters is used to track a walking person, over about 3 seconds. (Scale: $r = 15$ pixels with $N = 1500$ samples. See version of this paper on our web-site for a movie.)

likelihood. Consequently, the method is not limited to backgrounds that are stationary, or moving in some easily predictable fashion. The computational load consists principally of: evaluating the likelihood (6), of which offset functions $\rho_k(X)$ are the main burden; image processing to obtain the z_k. The image processing could be done using pyramid hardware [5]. The offset functions (at scale $r = 40$) can be evaluated, for $N = 500$, at frame-rate, on a desk-top workstation (SGI Octane). Bayesian correlation at video frame-rate should therefore be quite feasible.

8 Conclusions

Bayesian correlation is a synthesis of cross-correlation matching with probabilistic sampling. Its key, original elements are: the development of likelihood functions for correlation; learning of foreground and background distributions, with particular attention to statistical independence and "mixed" receptive fields; probabilistic multi-scale analysis by means of "layered sampling".

The approach has been widely 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 discussions with D. Mumford, S. Mallat, G. Hinton, B. Buxton, A. Zisserman and P. Torr.

References

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

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

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

[4] Blake, A., and Isard, M. *Active contours*. Springer, 1998.

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

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

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

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

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

[10] Hubel, D., and Wiesel, T. Receptive fields and functional architecture of monkey striate cortex. *J. Physiol. Lond. 195* (1968), 215–244.

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

[12] Isard, M., and Blake, A. ICondensation: Unifying low-level and high-level tracking in a stochastic framework. In *Proc. 5th European Conf. Computer Vision* (1998), 893–908.

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

[14] Mumford, D. Pattern theory: a unifying perspective. In *Perception as Bayesian inference*, D. Knill and W. Richard, Eds. Cambridge University Press, 1996, 25–62.

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

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

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

[18] Witkin, A., Terzopoulos, D., and Kass, M. Signal matching through scale space. In *5th National Conference on AI* (1986).

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

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