A probabilistic contour discriminant for object localisation

John MacCormick and Andrew Blake
University of Oxford, Oxford OX1 3PJ, UK.
jmac@robots.ox.ac.uk, ab@robots.ox.ac.uk

In Proc. Int. Conf. Computer Vision 1998

Abstract

A method of localising objects in images is proposed. Possible configurations are evaluated using the contour discriminant, a likelihood ratio which is derived from a probabilistic model of the feature detection process. We treat each step in this process probabilistically, including the occurrence of clutter features, and derive the observation densities for both correct "target" configurations and incorrect "clutter" configurations. The contour discriminant distinguishes target objects from the background even in heavy clutter, making only the most general assumptions about the form that clutter might take. The method generates samples stochastically to avoid the cost of processing an entire image, and promises to be particularly suited to the task of initialising contour trackers based on sampling methods.

1 Introduction

Object localisation is one of the classic Computer Vision problems for which good solutions can often be found in specific applications, but for which more general solutions seem elusive with current technology and techniques. In this paper the task is defined as follows. There is only one class of object to be localised, and all elements of the class have a similar structure. Examples of suitable classes are faces, hands, and cars; any element of the class is called a *target*. Given an image containing one or more targets, the task is to determine the location and configuration of each. Generally the targets comprise only a small fraction of the image.

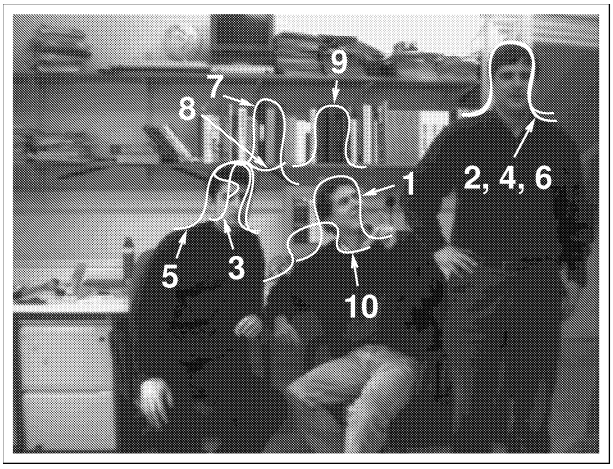
A class of targets is modelled with a configuration space of specified dimension, for example the six 3D Euclidean degrees of freedom with others for articulation or non-rigidity. A particular emphasis of this paper is on localisation for contour tracker initialisation, and this will guide our definition of successful outputs in terms of accuracy, confidence levels and speed.

We are not aware of any previous attempt to use contour outlines for object localisation as described here, but several authors have made probabilistic arguments based on different feature detection paradigms: [8] used decision theory and likelihood ratios for matching facial features, and [5, 9] both accept hypotheses based on the probability that straight lines and point features should fall in certain configurations.

The approach to object localisation presented in this paper applies to objects with complex outlines which need not contain corners or straight lines. It requires no knowledge of the background, is robust to lighting changes and works well in cluttered backgrounds. At the heart of the method is the idea of random sampling from a prior on the target's configuration. Because of the measurement regime adopted, only one-dimensional image processing is needed and only on a fraction of the image's pixels. The key to evaluating how "target-like" the samples are is a quantity called the *contour discriminant*, which is derived using a probabilistic model of the observations. The method is shown to perform very well on static images; a typical output is shown in figure 1. Note that in addition to finding the three targets successfully, the method has given us the contour discriminant of each plausible sample. These values can be used for certain inferences described later. Potential applications to contour tracker initialisation are also discussed. The nature of the output makes the method particularly well-suited to initialising trackers which use random sampling [4, 6, 7].

2 The target model and prior

The target objects in this paper are described by their outline, which is modelled as a B-spline as described in [2], for example. Any such outline is called a *contour*, and the space of possible configurations for a given target will be denoted by W. Given an image containing the target, and a hypothesised configuration $w_{\text{hyp}} \in W$, we adopt the measurement methodology of [2]: first, cast normals (called *measurement*



Ranking	$D(w)$	Ranking	$D(w)$
1	408.1	6	8.4
2	58.6	7	6.3
3	34.1	8	4.5
4	25.1	9	2.4
5	14.9	10	1.9

Figure 1: ***Finding head-and-shoulders outlines in an office scene.*** *The results of a sample of* $1,000$ *configurations are shown ranked by contour discriminant. The table shows the numerical values of the discriminant* D*; a value greater than one means a configuration is more target-like than clutter-like.*

lines) onto the image at pre-specified points around the contour; second, apply a 1-dimensional feature detector along each measurement line. A typical output is shown in figure 2. The distance from a feature to the contour is called the *innovation* of the feature. The number of features detected on the rth measurement line will be denoted N_r, and we write $\nu_{r,1}, \nu_{r,2} \dots \nu_{r,N_r}$ for the feature innovations on that line. When discussing only one measurement line, the simpler notation $\{N, (\nu_1, \nu_2 \dots \nu_N)\}$ will be used.

An essential ingredient in the proposed method is a prescription for choosing w_{hyp}. This will require a prior density $f(w)$ on W, which could be specified by the user but for our applications was learnt automatically by the following method. First obtain a video sequence of a target object undergoing typical behaviour. Initialise a contour tracker by hand, track the target throughout the sequence, and recover the implied configuration in each frame. Placing this data in suitable bins defines a distribution on W, which can then be redistributed to achieve desired global properties. For instance, in all results shown later, the prior was redistributed so that the translation component

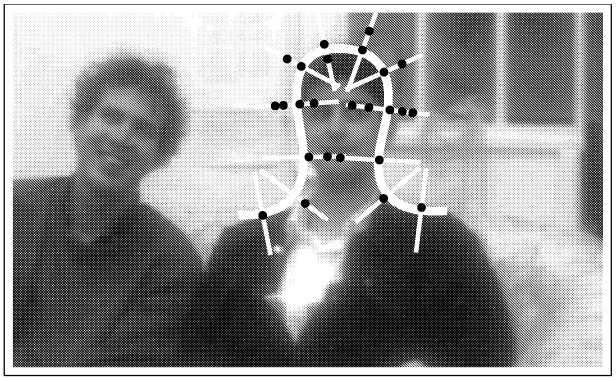


Figure 2: ***Measurement methodology.*** *The black dots are the outputs of a 1D feature detector applied to the measurement lines. This template and model were used to produce the results of figure 1.*

of the configuration is uniformly distributed over the image. The result of sampling from a prior created like this is shown in figure 3.

3 The observation model

Now we need a way of assessing the likelihood that a sample w from the prior is actually equal to the target configuration w_{true}. This requires a probabilistic model of how the measurements arise; such a model is described in this section and a useful statistic called the contour discriminant is introduced.

The first question is: *what are the measurements?* Understanding this is crucial, since throughout this paper we do not regard the grey-level values of the entire image as the measurements. This is despite the fact that (for static images at least) the image is obtained from the camera, stored in memory, and therefore in some sense "measured", before we begin analysis. Rather, this set of grey-levels is regarded as a large population which is too expensive to observe exhaustively; we will sample from it and will hopefully draw conclusions after observing only a fraction of the population. Therefore, from now we reserve the word *measurement* to mean the output of the feature detector described above — so for each measurement line, the measurements are the number of features found, N, and the feature innovations $\nu_1, \nu_2 \dots \nu_N$. The output for one measurement line will be denoted by z, so $z = \{N, (\nu_1, \nu_2 \dots \nu_N)\}$; the output for the rth measurement line in a configuration is $z_r = \{N_r, (\nu_{r,1}, \nu_{r,2} \dots \nu_{r,N_r})\}$. Finally, we write $\mathbf{z} = (z_1, z_2 \dots z_R)$, where R is the number of measurement lines on a configuration. Note that this notation specifies the measurements for a single configuration. Measurements for different configurations will be dis-

tinguished by superscripts as in $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}$.

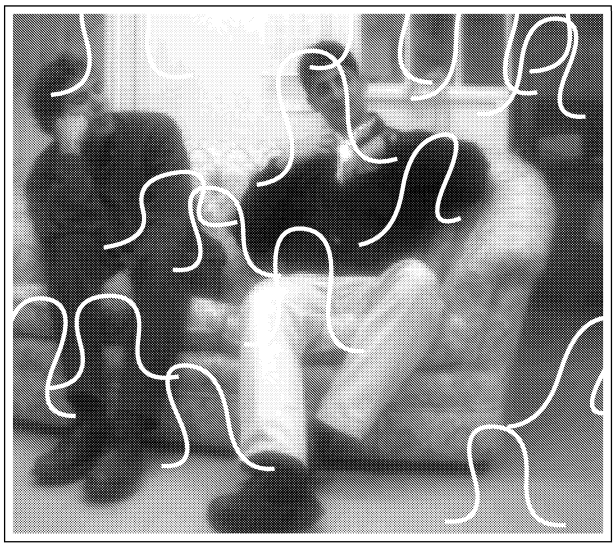


Figure 3: ***The prior density.*** *Shown here are fifteen samples from a prior for head-and-shoulders outlines constructed by the learning method described in the text. W is the 6-dimensional space of affine transformations. This prior is the one used to produce the results of figure 1. The learnt motions comprised the head tilting up to about $40°$ from the vertical in every direction, and the prior was redistributed afterwards to have the following properties: uniform over both translation parameters, and a Euclidean scaling factor obeying a normal distribution with mean 1 and standard deviation about 7%.*

3.1 Applying Bayesian decision theory

We need a decision rule to deal with the following scenario: given a fixed image and K different configurations $\{w^{(1)}, \ldots w^{(K)}\}$, which is most likely to be the target's configuration w_{true}? Or even better, we would like to know the probability that $w^{(k)} = w_{\text{true}}, k = 1, \ldots K$. Bayesian decision theory [3] answers these questions.

Consider the case where the sample consists of only one configuration, w, and we must choose between two possibilities: either $w = w_{\text{true}}$, or the features found near w were caused by clutter. We perform a measurement on the configuration, obtaining the result $\mathbf{z} = (z_1, \ldots z_R)$ where each $z_r = (N_r, \boldsymbol{\nu}_r)$. Assuming a zero-one loss function, the optimum Bayes decision rule for deciding between the two possibilities is to accept $w = w_{\text{true}}$ if

$$\text{Prob}(w = w_{\text{true}}|\mathbf{z}) > \text{Prob}(w \text{ due to clutter}|\mathbf{z}),$$

otherwise reject and conclude that the measurements at w were caused by clutter. If the prior probabilities of the two options are equal, this is equivalent to accepting $w = w_{\text{true}}$ when

$$\text{Prob}(\mathbf{z}|w = w_{\text{true}}) > \text{Prob}(\mathbf{z}|w \text{ due to clutter }).$$

This motivates the following definition. If $\mathbf{z}$ is the set of measurements for a configuration w then the *contour discriminant of* w is the likelihood ratio

$$D(w) = \frac{\text{Prob}(\mathbf{z}|w = w_{\text{true}})}{\text{Prob}(\mathbf{z}|w \text{ due to clutter})}. \qquad (1)$$

By reasoning adapted from [3], one can show the contour discriminant can be used for three main types of inferences:

- If $D(w) > 1$, then w is more "target-like" than "clutter-like". That is, it is more likely that the measurements were generated by a target in configuration w than by random background clutter.
- In any set of configurations, the single configuration most likely to equal w_{true} is the one with greatest contour discriminant.
- If a set of configurations $\{w^{(1)}, \ldots w^{(K)}\}$ is "well-separated",[1] and it is known that one element of the set is w_{true}, then the probability that $w^{(k)} = w_{\text{true}}$ is $D(w^{(k)})/\sum_j D(w^{(j)})$.

The second and third types of inference can be easily adapted to incorporate a prior $f(w)$ on W by using instead the discriminant $\hat{D}(w) = D(w)f(w)$. Note, however, that the density $f(w)$ should be expressed with respect to a measure on W which reflects the human notion of one contour being nearly the same as another. This can be done rigorously but the details are omitted.

3.2 Model description

The first step in specialising the Bayesian framework above to our application is to specify some assumptions of the model.

Non-detection probability Because some parts of the target's outline may have insufficient contrast with the background, it is possible that true boundary features will remain undetected even for the correct configuration w_{true}. We follow [1] by assigning such non-detection events a fixed probability q, independently on each measurement line.

Feature detector and model error Two sources of error cause non-zero innovations for the boundary

[1] In practical terms, this technical assumption means that no two contours in the set overlap each other significantly.

features, even for a perfect hypothesis. The first is error in the model, such as modelling a slightly non-rigid object in a configuration space W which allows only rigid transformations of the template. The second, and less significant, is uncertainty in the feature detector itself. It is assumed these two effects cause the feature detector to report feature innovations with an error whose p.d.f. is $\mathcal{E}(\cdot)$. In this paper we take $\mathcal{E}$ to be Gaussian with standard deviation σ. The examples shown later used $\sigma = 5$ pixels.

Clutter model The detection of clutter features is regarded as an i.i.d. random process on the exterior (i.e. exterior to the target object) portion of each measurement line. The probability that n clutter features are detected on a measurement line is denoted by $\pi(n)$. When necessary, the notation $\pi_l(n)$ will be used to emphasise the fact that π depends on the length l of the measurement line lying outside the target object. Regardless of the choice of π, we assume the distribution of the *position* of the clutter features is uniform over the length of the measurement line.

Interior model The detection of interior features on a measurement line is modelled in the same way as clutter is on the exterior: an i.i.d. random process whose probability of producing m features is denoted $\rho(m)$, or $\rho_l(m)$ when we want to emphasise this distribution depends on the length l of the interior portion of the measurement line. Again, the distribution of the feature positions is taken to be uniform over the length l.

A seemingly sensible choice for $\pi(\cdot)$ is a Poisson distribution, with parameter learnt from typical scenes. Certainly, *on a sufficiently small scale*, it would be hard to argue for anything else. But experiments showed that estimating a single global Poisson density parameter for the clutter distribution is a poor choice, since most images are divided into distinct regions with very different clutter feature densities. Hence a new approach is required, for which the modelled properties of clutter features are as "uniform" as possible. The examples shown later assume not only a uniform distribution for the position of the clutter measurements, but that the probability of obtaining n clutter measurements in a length l is constant for all n. In other words, we set $\pi \equiv 1$. This distribution is non-normalisable, but all the formulae below can be made rigorous for $\pi \equiv 1$ by taking a suitable limit. In particular, note that the contour discriminant (2) is invariant to the normalisation of π.

In the case of the interior features, we should certainly exploit any prior knowledge available. In the examples shown later the training was kept to an absolute minimum by learning the interior model from a single image — the same one from which the shape of the template is learnt. Suppose that m_r interior features were detected on the rth measurement line of the template, the interior portion of which has length l. Then we modelled ρ for this measurement line as a Poisson distribution with density parameter m_r/l.

3.3 The contour discriminant

Now the probabilistic model for the detection process can be stated, and from this the observation density and hence the contour discriminant will be derived. Consider just one measurement line of length L, and suppose the target boundary intersects the line at zero innovation. In the case that the target is detected, the model for the generation of features is as follows:

1. Draw a randomly from $\mathcal{E}(a)$. (So a is the reported innovation of the target.)
2. Draw m randomly from $\rho_{L/2+a}(m)$, and draw $b_1, \ldots b_m$ from $\mathrm{Rect}(-L/2, a)$. (So m is the number of interior features, and $b_1, \ldots b_m$ are their innovations.)
3. Draw n randomly from $\pi_{L/2-a}(n)$, and draw $c_1, \ldots c_n$ from $\mathrm{Rect}(a, L/2)$. (So n is the number of exterior (clutter) features, and $c_1, \ldots c_n$ are their innovations.)
4. Set $N = n + m + 1$. (So N is the total number of features found.)
5. Reorder $(a, b_1, \ldots b_m, c_1, \ldots c_n)$ as $(\nu_1, \ldots \nu_N)$ with $\nu_1 \leq \nu_2 \leq \ldots \leq \nu_N$.
6. Report $(N, (\nu_1, \nu_2, \ldots \nu_N))$.

In the undetected case, a is a hidden variable but the other steps are identical. A final piece of notation specifies whether or not the target object intersects the measurement line, and whether or not its boundary was detectable. Let P_ν denote the event that the target object boundary is present at innovation ν on the measurement line under consideration, and P' denote the event that it is not present. Also, let V be the event that the boundary of the target is visible (i.e. detectable), and V' the event that it was not visible.

It is not difficult to show that under these assumptions the observation density when the target is visible at zero innovation is given by

$$f(z = (N, \boldsymbol{\nu}) | P_0, V) = \sum_{k=1}^{N} p_{k,N} \mathcal{E}(\nu_k) \frac{(k-1)!}{(L/2 + \nu_k)^{k-1}} \frac{(N-k)!}{(L/2 - \nu_k)^{N-k}},$$

where $p_{k,N}$ is the probability, given P_0 and V, that the target object's boundary was the kth of N features found. These coefficients can be pre-calculated

by numerical integration. A similar formula holds for the density given the target was undetected, but this time it involves an integration over the hidden variable:

$$f(z|P_0,V') = \sum_{j=0}^{N} \int_{a=z_j}^{z_{j+1}} \mathcal{E}(a) \frac{j!}{(L/2+a)^j} \frac{(N-j)!}{(L/2-a)^{N-j}} \times \rho_{L/2+a}(j)\pi_{L/2-a}(N-j)\, da$$

The density when no object is present is

$$f(z=(N,\boldsymbol{\nu})|P') = \frac{N!\pi_L(N)}{L^N}.$$

The factorials in these formulae arise from the permutations of symmetric variables in uniform pdfs; in certain special cases they combine to produce binomial coefficients in the contour discriminant, which have nice intuitive interpretations.

Actually, these formulae are for the rather reductive case of only one measurement line. If we are prepared to assume the model above holds independently on the R different measurement lines, we get

$$f(\mathbf{z}|w) = \prod_{r=1}^{R} \left(q f(z_r|P_0,V') + (1-q) f(z_r|P_0,V)\right),$$

$$f(\mathbf{z}|\text{no object present}) = \prod_{r=1}^{R} f(z_r|P').$$

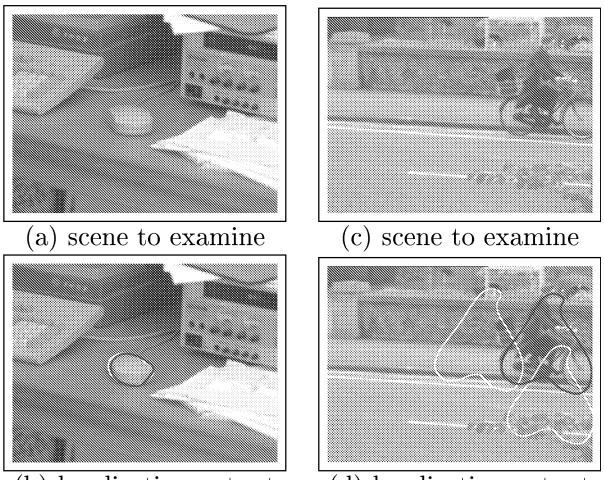
These two expressions can now be substituted in (1), giving an explicit expression for the contour discriminant:

$$D(w) = f(\mathbf{z}|w)/f(\mathbf{z}|\text{no object present}) = \prod_{r=1}^{R} \left(\frac{q f(z_r|P_0,V') + (1-q) f(z_r|P_0,V)}{f(z_r|P')} \right) \quad (2)$$

4 Localisation by random sampling

The obvious next question is: will the results of the previous section help us to find objects in cluttered static images? This will be done by randomly sampling from a prior. The main example given here shows the sampling method applied to finding head-and-shoulders outlines in an office environment. A contour template was obtained by hand-drawing round the head in figure 2, from which the parameters for the interior model were also learnt. A prior $f(w)$ was learnt by the technique described in section 2 (see figure 3). Then, the following simple procedure was applied to a completely different scene: sample from $f(w)$ 1000 times, apply a simple local maximisation to each, calculate the contour discriminant of each sample using equation (2), and report the 10 configurations with the greatest contour discriminants. Figure 1 shows the results: the three true targets were found, and 4 of the reported 10 samples were spurious. The procedure takes 1.7 seconds on an SGI O2 (R5000, 180MHz). Similar results were obtained with other scenes; two examples are given in figure 4.

It is natural to wonder whether the full complexity of the measurement model described is necessary for the intended applications, especially if there are more easily calculated alternatives which perform just as well. To test this we applied the same algorithm to the same example, using instead of the contour discriminant the density function used in [6]. This simpler discriminant has two qualitative differences to the contour discriminant: the background clutter is assumed to be a Poisson process with constant parameter λ, and the target is treated as a wire frame — in other words, the interior features are modelled in the same way as clutter. In the notation of this paper the simpler discriminant is $\prod_{r=1}^{R} \left(q\lambda + \sum_{j=1}^{N_r} \mathcal{E}(\nu_{r,j})\right)$. The performance of this discriminant was significantly worse than the contour discriminant: only two of the three targets were found and there were seven spurious hypotheses.



(a) scene to examine (c) scene to examine

(b) localisation output (d) localisation output

Figure 4: ***More localisation results.*** *In each example, 1,000 samples were drawn from the prior. The sample of highest contour discriminant is shown in black, and the 2nd and 3rd highest are in white. (In (b), the three contours are nearly identical.) In both cases, the background and lighting were different in the learnt template and examined scene.*

5 Application to tracker initialisation

One objective of this work is to use the framework of sections 2–4 for initialising contour trackers, or for reinitialising them after they have lost lock on the target. The method seems particularly suited to trackers based on sampling (such as Condensation [6], but see also [4, 7]), since the contour discriminant enables us to infer the entire posterior distribution of the target configuration rather than just a point estimate. Figure 5 shows the type of distribution which has been used in a successful preliminary implementation. In the future we hope to develop a system which will allocate resources intelligently between two modules: one for tracking based on learnt dynamics, and the other for reinitialising based on a learnt prior.

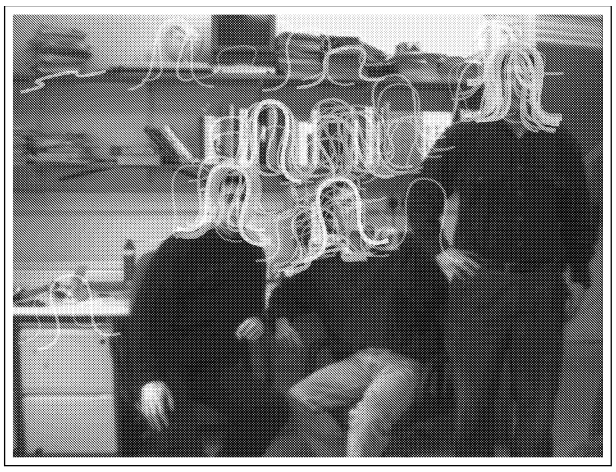


Figure 5: ***Posterior distribution for tracker initialisation.*** *The total mass of each displayed contour is proportional to the log of its contour discriminant. Such a distribution can be used to initialise trackers based on random sampling.*

6 Conclusion

The new method for object localisation uses a quantity termed the *contour discriminant*, a likelihood ratio which expresses probabilistically whether a configuration is target-like or clutter-like. Because it is based on probabilities, the contour discriminant can be used for statistical inferences. In particular, its magnitude determines the probabilities in a "clutter versus target" hypothesis test for a single configuration, which is useful for cut-off and termination criteria. Moreover, the normalised discriminants of certain sets of configurations are the probabilities for a multiple-hypothesis test on which configuration is the target. Of course, likelihood ratios are commonly used by vision researchers. The crucial aspect the contour discriminant is that it is derived from a probabilistic model of the feature detection process. This model is essential for the approach to work: it has been shown that a cruder model (with less appropriate clutter assumptions and no explicit model of features in the interior of the target) gives inferior performance.

The method requires no previous knowledge of the background, is not affected by lighting changes, and is effective in cluttered scenes. Another advantage of the approach is flexibility — the set-up time for a new class of target objects is only that required to click round a typical outline with the mouse. The method is fast because it treats the image grey-levels as a large population which is too expensive to observe completely, and instead performs simple processing on a small number of pixels in each member of a random sample.

The approach works well for targets whose configurations, as defined by their outline in B-spline representation, can be described by a reasonably simple manifold. Impressive results have been demonstrated in several examples. Implemented on a desktop graphics workstation, the method always found all the targets in the 1000 samples allowed; this takes under two seconds.

The potential applications to contour tracker initialisation were also discussed. The method seems particularly suited to initialisation of sampling trackers since it provides an estimate of the whole posterior density rather than just its mode. Future work will put this in a rigorous statistical context.

References

[1] Y. Bar-Shalom and T.E. Fortmann. *Tracking and Data Association.* Academic Press, 1988.

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

[3] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis.* John Wiley and Sons, 1973.

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

[5] W.E.L. Grimson, D.P. Huttenlocher, and D.W. Jacobs. A study of affine matching with bounded sensor error. In *Proc. 2nd European Conf. Computer Vision*, pages 291–306, 1992.

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

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

[8] T.K. Leung, M.C. Burl, and P. Perona. Finding faces in cluttered scenes using random graph matching. In *Proc. IEEE PAMI Conf.*, pages 637–644, Cambridge, June 1995.

[9] D.G. Lowe. The viewpoint consistency constraint. *Int. J. Computer Vision*, 1:57–72, 1987.