Statistical models of visual shape and motion

By A. Blake, B. Bascle, M. Isard and J. MacCormick

Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, UK.

The analysis of visual motion against dense background clutter is a challenging problem. Uncertainty in the positions of visually sensed features and ambiguity of feature correspondence call for a probabilistic treatment, capable of maintaining not simply a single estimate of position and shape but an entire distribution. Exact representation of the evolving distribution is possible when the distributions are Gaussian and this yields some powerful approaches. However normal distributions are limited when clutter is present: because of their unimodality, they cannot be used to represent simultaneous alternative hypotheses.

One powerful methodology for maintaining non-Gaussian distributions is based on random sampling techniques. The effectiveness of "factored sampling" and "Markov chain Monte Carlo" for interpretation of static images is widely accepted. More recently, factored sampling has been combined with learned dynamical models to propagate probability distributions for object position and shape. Progress in several areas is reported here. First a new observational model is described that takes object opacity into account. Secondly, complex shape models to represent combined rigid and nonrigid motion have been developed, together with a new algorithm to decompose rigid from nonrigid. Lastly, more powerful dynamical prior models have been constructed by appending suitable discrete labels to a continuous system state; this may also have applications to gesture recognition.

Keywords: vision; computing; shape; sampling; estimation; gesture

1. Introduction

This paper addresses some problems in the interpretation of visually observed shapes in motion, both planar and three-dimensional shapes. Mumford (1996), interpreting the "Pattern Theory" developed over a number of years by Grenander (1976), views images as "pure" patterns that have been distorted by a combination of four kinds of degradations. This view applies naturally to the analysis of static, two-dimensional images. The four degradations are given here, together with comments on how they need to be extended to take account of three-dimensional objects in motion.

(i) *Domain warping* in which the domain of an image I is transformed by a mapping g:

$$I(\mathbf{r}) \rightarrow I(g(\mathbf{r})).$$

The three-dimensional nature of the world means that the warp g may be composed largely of "projective" or "affine" transformations. The dynamical nature of the problems addressed here will require time-varying warps $g(\mathbf{r}, t)$.

(ii) *Superposition:* objects may overlap and in certain forms of imaging this may produce linear combinations, which is fortuitous because they can be analysed by linear spectral decomposition. In images of opaque,three dimensional objects, however, far surfaces are obscured by near ones.

(iii) *Distortion and noise:* image measurements are corrupted by noise and blur:

$$I(\mathbf{r}) \to f(I(\mathbf{r}), \mathbf{n}).$$

Image degradations may be most effectively modelled as applying to certain image "features" obtained by suitable pre-processing of an image, rather than directly to an image itself.

(iv) *Observation failure:* disturbance of the observation process; often caused, in the work described here, by distracting background clutter.

A key idea in pattern theory is recognition by synthesis, in which predictions following from particular hypotheses play an important role. The predictions are generated and tested against the products of analysis of an image. Bayesian frameworks, which have gained significant influence in modelling perception [Knill et al., 1996], seem to be a natural vehicle for this combination of analysis and synthesis. In the context of machine perception of shapes we can state the problem as one of interpreting a *posterior* density function $p(\mathbf{X}|\mathbf{Z})$ for a shape $\mathbf{X}$ in some appropriate *shape-space* $\mathcal{S}$, given data $\mathbf{Z}$ from an image (or data $(\mathbf{Z}_1, \mathbf{Z}_2, \ldots)$ from a sequence of images). The posterior density must be computed in terms of *prior* knowledge about $\mathbf{X}$ and inference about $\mathbf{X}$ based on the *observations* $\mathbf{Z}$. Bayes' formula expresses this as follows:

$$p(\mathbf{X}|\mathbf{Z}) \propto p(\mathbf{Z}|\mathbf{X})p_0(\mathbf{X}), \tag{1.1}$$

in which $p_0(\mathbf{X})$ is the prior density for $\mathbf{X}$ and the conditional density $p(\mathbf{Z}|\mathbf{X})$ conveys the range of likely observations to arise from a given shape $\mathbf{X}$. All this links in directly with the four degradations above. In particular, type 1 (warping) is represented in the prior p_0. Types 3 and 4 (noise and observation failure) are incorporated into the observation density $p(\mathbf{Z}|\mathbf{X})$.

The framework for Bayesian inference of visual shape and motion that forms the basis of this paper is set out in detail in [Blake and Isard, 1998]. This paper aims to summarise that framework and introduce several new ideas. The organisation of the paper is summarised by section, as follows.

2. *Statistical modelling of shape* — how to choose a suitable shape-space $\mathcal{S}$ and a prior p_0, or to learn them from a set of examples.

3. *Statistical modelling of image observations* — how to construct an effective observation density $p(\mathbf{Z}|\mathbf{X})$ that takes into account image intensities both within the shape of interest and in the background.

4. *Sampling methods* — using random sample generation to construct an approximate representation of the posterior for $\mathbf{X}$, given that the complexity of $p(\mathbf{Z}|\mathbf{X})$ can make exact representation of the posterior infeasible.

5. *Modelling dynamics* — extending the Bayesian framework to deal with sequences of images demands priors for temporal sequences $\mathbf{X}_1, \mathbf{X}_2, \ldots$. These can either be constructed by hand or learned from examples.

6. *Learning dynamics* — the most effective way to set up dynamical models is to learn them from training sets.

7. *The Condensation algorithm* — a random sampling algorithm for interpretation of shapes in motion.

8. *Dynamics with discrete states* — extending the dynamical repertoire to modelling of motion with several modes, for example walk–trot–canter–gallop.

2. Statistical modelling of shape

This section addresses the construction of a prior model $p_0(\mathbf{X})$ for a shape. This can be done in a somewhat general way if the dimensionality of the shape-space $\mathcal{S}$ is fixed in advance to be small, for example just translations in the plane. Then extended observation of the positions of moving objects in some area can be summarised as a histogram which serves as an approximate representation of the prior p_0 [Fernyhough et al., 1996]. In higher dimensional shape-spaces, involving three-dimensional rigid motion and deformation of shape, histograms are less practical. Here we focus on Gaussian distributions.

A Gaussian distribution is specified by its mean and variance and these can be estimated from a training sequence $\mathbf{X}_1, \mathbf{X}_2, \ldots$ of shapes by taking the sample mean $\overline{\mathbf{X}}$ and the sample variance

$$\Sigma = \frac{1}{M} \sum_{k=1}^{M} (\mathbf{X}_k - \overline{\mathbf{X}})(\mathbf{X}_k - \overline{\mathbf{X}})^T.$$

Moreover, Principal Components Analysis (PCA) [Rao, 1973] can be used to restrict the shape-space $\mathcal{S}$ to explain most of the variance in the training set while keeping the dimension of $\mathcal{S}$ small, in the interests of computational efficiency [Cootes et al., 1993, Baumberg and Hogg, 1994, Lanitis et al., 1995, Beymer and Poggio, 1995, Baumberg and Hogg, 1995a, Vetter and Poggio, 1996]. An example is given in figure 1.

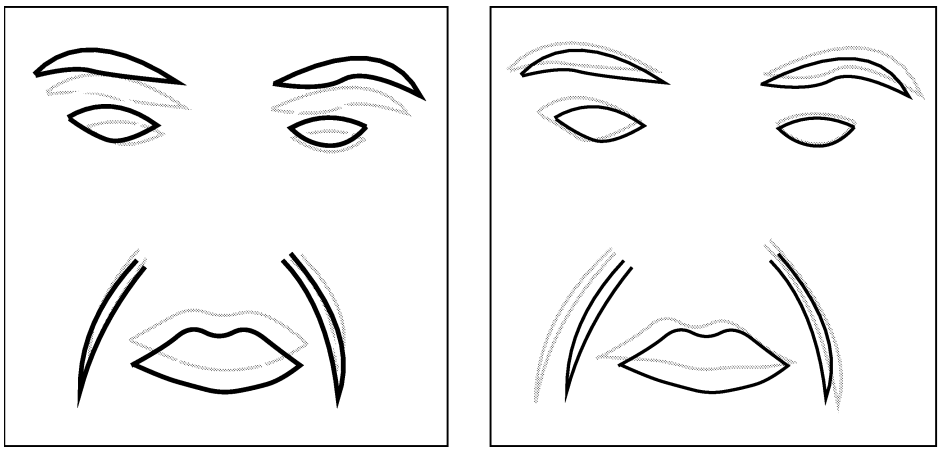
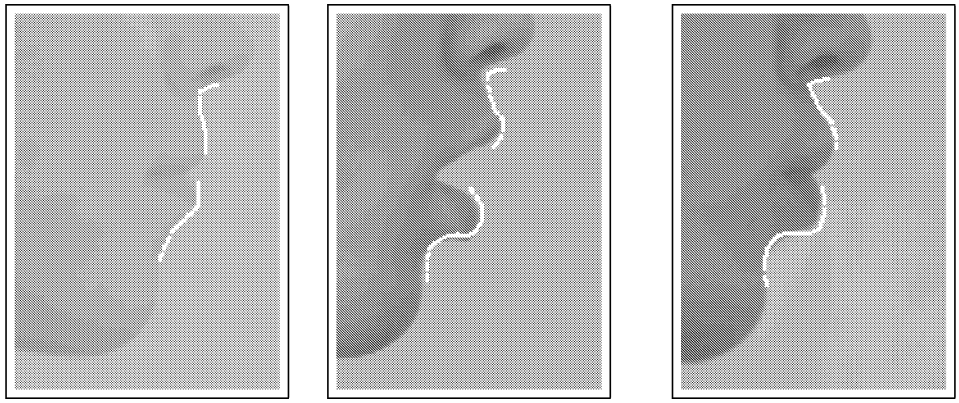


Figure 1. **PCA for faces.** *A shape-space of facial expressions is reduced here by PCA to the two-dimensional space that best covers the expressions in a certain training sequence.*

However, the resulting shape-space, though economical, is not especially easy

to interpret because Principal Components need not be meaningful. More meaningful "constructive" shape-spaces can be generated by acknowledging three-dimensional projective effects and constructing affine spaces for instance whose components are directly related to rigid body transformations [Ullman and Basri, 1991, Koenderink and van Doorn, 1991]. In addition, named deformations can be included in a basis for $\mathcal{S}$ as "key-frames" [Blake and Isard, 1994], as in figure 2.



Template $\mathbf{Q}_0$ *Key-frame: opening* $\mathbf{Q}_1$ *Key-frame, protrusion* $\mathbf{Q}_2$

Figure 2. **Key-frames.** *Lips template followed by two key-frames, representing interactively tracked lips in characteristic positions. The key-frames are combined linearly with appropriate rigid degrees of freedom, to give a shape-space suitable for use in a tracker for non-rigid motion.*

A constructive shape-space S^c can be combined with PCA to give the best of both worlds. "Residual PCA" operates on a constructive shape-space that does not totally cover a certain data-set, and fills in missing components by PCA. Then the constructive subspace retains its interpretation and only the residual components, covered by PCA, cannot be directly interpreted. This is done by constructing a projection operator E^c that maps $\mathcal{S}$ to $\mathcal{S}^c$ and applying PCA to the residual training-set vectors $\mathbf{X}_1^r, \mathbf{X}_2^r, \ldots$ where

$$\mathbf{X}^r = \mathbf{X} - E^c\mathbf{X}.$$

Full details of the algorithm are given in [Blake and Isard, 1998] and an example of its application is shown in figure 3.

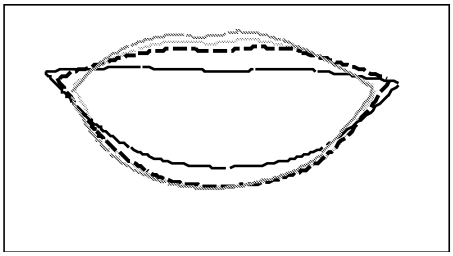


Figure 3. **Sampling from a prior for lip-shape, excluding translation** *Random sampling illustrates how a learned prior represents plausible lip configurations. Any rigid translations in the training set, due to head-motion, are separated out as a constructive shape-space in residual PCA.*

Finally, some complex issues arise when dealing with mixed rigid and nonrigid deformation. For example, one application is to track the facial motion of an actor and channel the coded motion to a graphical animation. It can be argued [Bascle and Blake, 1998] that the composition of expression and pose can be expressed bilinearly to give shape parameters

$$X_i^j = \lambda_i Y_j$$

where λ_i is the weight associated with the ith expression and Y_j is the jth component of an affine transformation. Decomposition of such products can be achieved using Singular Value Decomposition (SVD) [Barnett, 1990], as has been done elsewhere for structure and motion [Tomasi and Kanade, 1991], and shape and shading [Freeman and Tenenbaum, 1997]. The practical result is good isolation of pose from expression, as figure 4 shows.

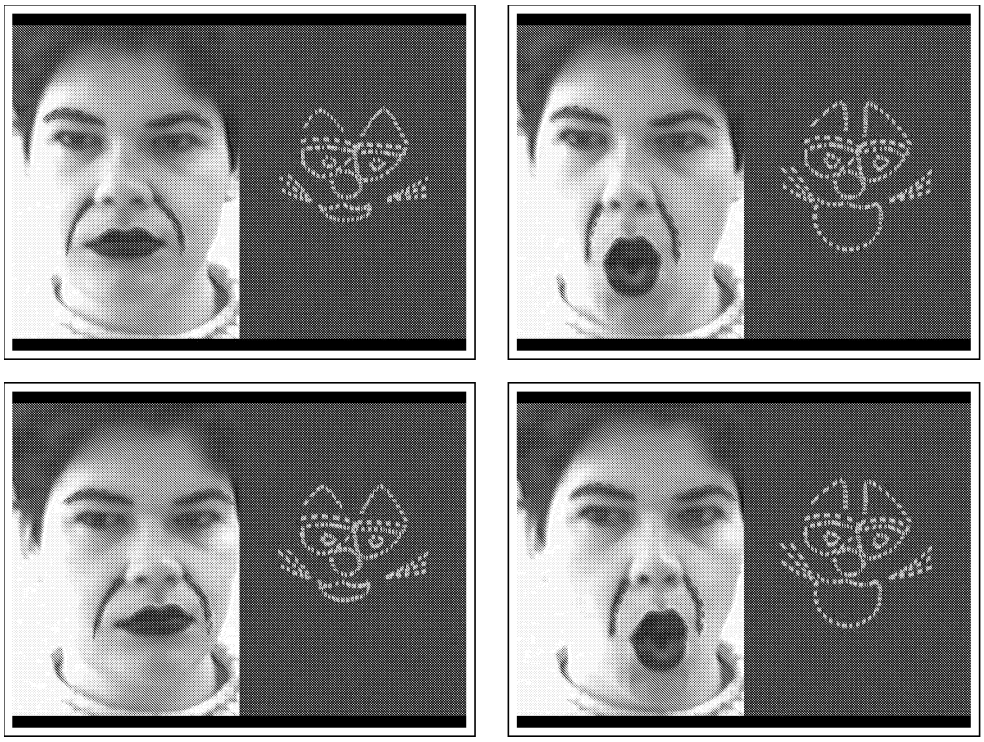


Figure 4. **Pose-invariant transmission of facial expression.** *Separation of nonrigid from rigid motion by SVD is used here to extract the facial expression of an actor. The extracted expression is displayed on this cat caricature in a fixed pose, and can be seen to be independent of the pose of the actor's head.*

3. Statistical modelling of image observations

Gaussian distributions may often be acceptable as models of prior shape, but they are adequate as observation distributions only in the clutter-free case. Typically, in our framework, observations are made along a series of spines, normal to the hypothesised shape, as in figure 5. Consider the one-dimensional problem of observing, along a single spine, the set of feature positions $\{\mathbf{z} =$

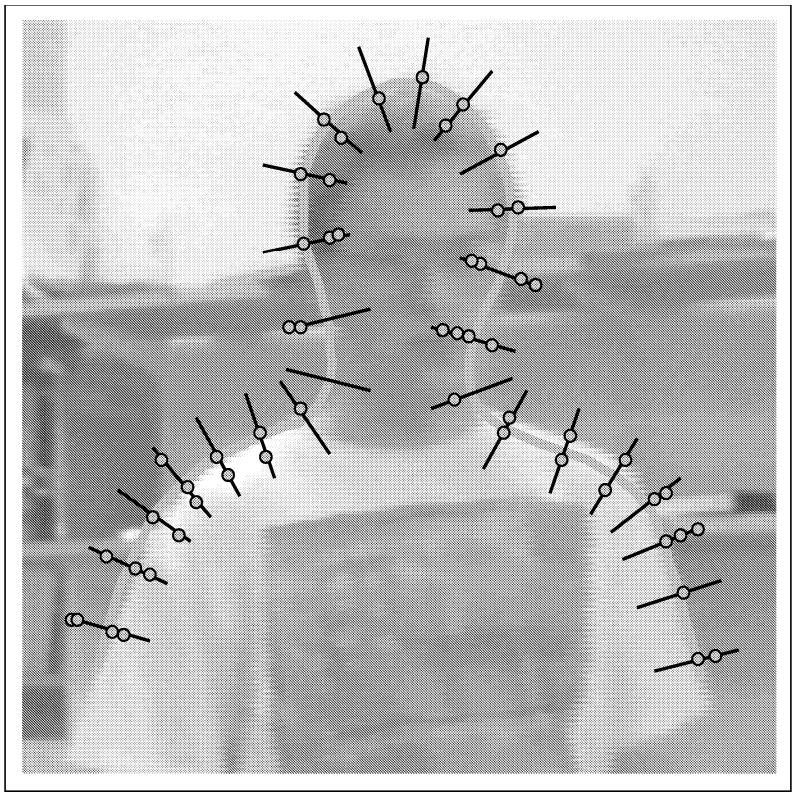


Figure 5. **Observation process.** *The thick line is a hypothesised shape, represented as a parametric spline curve. The spines are curve normals along which high-contrast features (white crosses) are sought.*

$(z_1, z_2, \ldots, z_M)\}$. Assuming a uniform distribution of background clutter, and a Gaussian model for error in measurement of the position of the true object edge, leads [Isard and Blake, 1996] to the multi-modal observation density $p(\mathbf{z}|\mathbf{X})$ depicted in figure 6. The multiple peaks in the density are generated by clutter and cannot possibly be modelled by a single Gaussian. A mixture of Gaussians might be feasible but a very efficient alternative is to use random sampling (next section).

The observation model above was based on the assumption that the observable contour is a "bent wire" resting on a cluttered background. This is not very realistic. It is highly desirable in practice to allow for object opacity and to distinguish between the textured interior of an object and its cluttered exterior. A probabilistic model that reflects this is based on the following assumptions.

Feature localisation error It is assumed that the feature detector reports object outline position with an error whose density is $\mathcal{E}(\cdot)$, taken usually to be Gaussian.

Occlusion probability The possibility that the outline is missed by the feature detector is allowed, with probability q.

Clutter model Detection of clutter features is regarded as an i.i.d. random process on the portion of each measurement line that lies outside the object. The

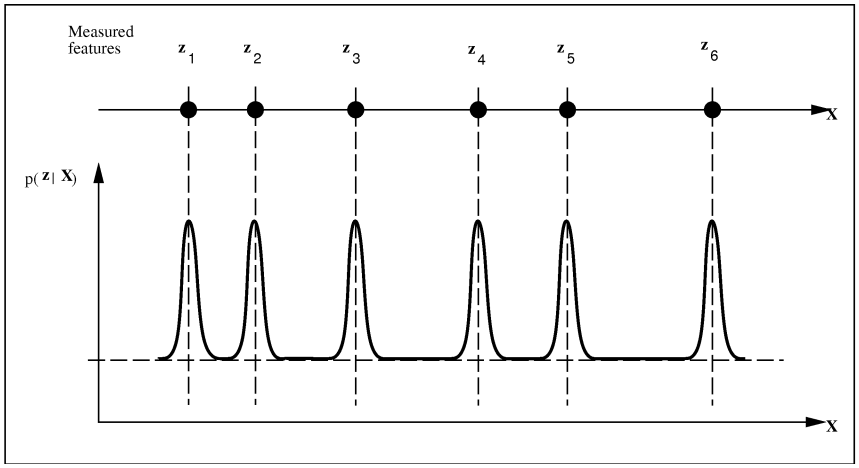


Figure 6. **Multi-modal observation density** *(one-dimensional illustration). A probabilistic observation model allowing for clutter and the possibility of missing the target altogether is specified here as a conditional density* $p(\mathbf{z}|\mathbf{X})$*. It has a peak corresponding to each observed feature.*

probability $\pi(n)$ that n clutter features are detected on a normal is generally taken to be uniform.

Interior model Interior features on a measurement line are modelled as uniformly distributed along the interior portion of the normal. The distribution $\rho(m)$ for the number m of interior features observed is taken to be Poisson with a known density parameter which is actually learned by observing instances of the object. A density $p(\mathbf{z}|\mathbf{X})$ based on these assumptions can be constructed and expressed as $p = \lambda D$ where λ is a constant and

$$D(\mathbf{X}) = p(\mathbf{z}|\mathbf{X})/p(\mathbf{z}|\text{no object present})$$

— the *contour discriminant.* This is a discriminant function [Duda and Hart, 1973] in the form of a *likelihood ratio.* It has the attraction that, in addition to conveying the relative values of $p(\mathbf{z}|\mathbf{X})$, its absolute value is also meaningful: $D(\mathbf{X}) > 1$ implies that the observed features $\mathbf{z}$ are more likely to have arisen from the object in location $\mathbf{X}$ than from clutter.

Lastly, densities $p(\mathbf{z}|\mathbf{X})$ for each normal need to be combined into a grand observation density $p(\mathbf{Z}|\mathbf{X})$, and this raises some issues about independence of measurements along an object contour. Details of the form and computation of the full observation density are given in [MacCormick and Blake, 1998]. Results of the evaluation of the contour discriminant on a real image are shown in figure 7. Analysis of the same image using the simpler "bent wire" observation model degrades the results, failing altogether to locate the leftmost of the three people. The explicit modelling of object opacity has clearly brought significant benefits.

4. Sampling methods

The next stage of the pattern recognition problem is to construct the posterior density $p(\mathbf{X}|\mathbf{Z})$ by applying Bayes' rule (1.1). In the previous section it became plain that the observation density has a complex form in clutter. This means that direct evaluation of $p(\mathbf{X}|\mathbf{Z})$ is infeasible. However iterative sampling

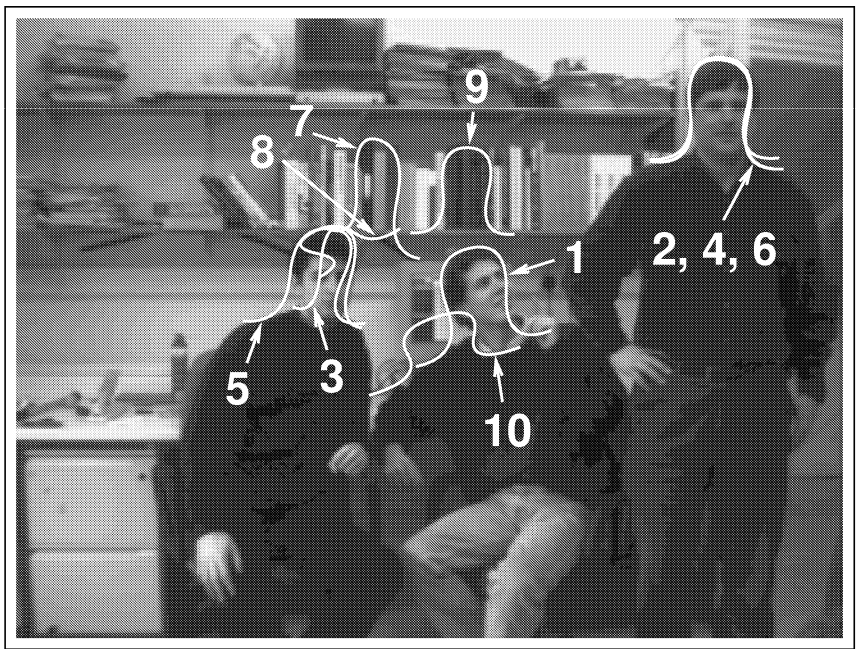


Figure 7. **Finding head-and-shoulders outlines in an office scene.** *The results of a sample of* $1,000$ *configurations are shown ranked by value of their contour discriminant. The table displays the cases in which* $D > 1$*, indicating a configuration that is more target-like than clutter-like.*

techniques can be used [Geman and Geman, 1984, Ripley and Sutherland, 1990, Grenander et al., 1991, Storvik, 1994]. The *factored sampling* algorithm [Grenander et al., 1991]. generates a random variate $\mathbf{X}$ from a distribution $\tilde{p}(\mathbf{X})$ that approximates the posterior $p(\mathbf{X}|\mathbf{Z})$. First a sample-set $\{\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(N)}\}$ is generated from the prior density $p(\mathbf{x})$ and then a sample $\mathbf{X} = \mathbf{X}_i$, $i \in \{1, \ldots, N\}$ is chosen with probability

$$\pi_i = \frac{p(\mathbf{Z}|\mathbf{X} = \mathbf{s}^{(i)})}{\sum_{j=1}^{N} p(\mathbf{Z}|\mathbf{X} = \mathbf{s}^{(j)})}.$$

Sampling methods have proved remarkably effective for recovering static objects from cluttered images. For such problems $\mathbf{X}$ is a multi-dimensional set of parameters for curve position and shape. In that case the sample-set $\{\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(N)}\}$ is drawn from the posterior distribution of $\mathbf{X}$-values, as illustrated in figure 8.

5. Modelling dynamics

In order to be able to interpret moving shapes in sequences of images, it is necessary to supply a prior distribution not just for shape but also for the motion of that shape. Consider the problem of building an appropriate prior model for the position of a hand-mouse engaged in an interactive graphics task. A typical trace in the x-y plane of a finger drawing letters is given in figure 9. If the entire trajectory were treated as a training set, the methods discussed earlier could be applied to learn a Gaussian prior distribution for finger position. The learned prior is broad, spanning a sizeable portion of the image area, and places little constraint on the measured position at any given instant. Nonetheless, it is quite clear from

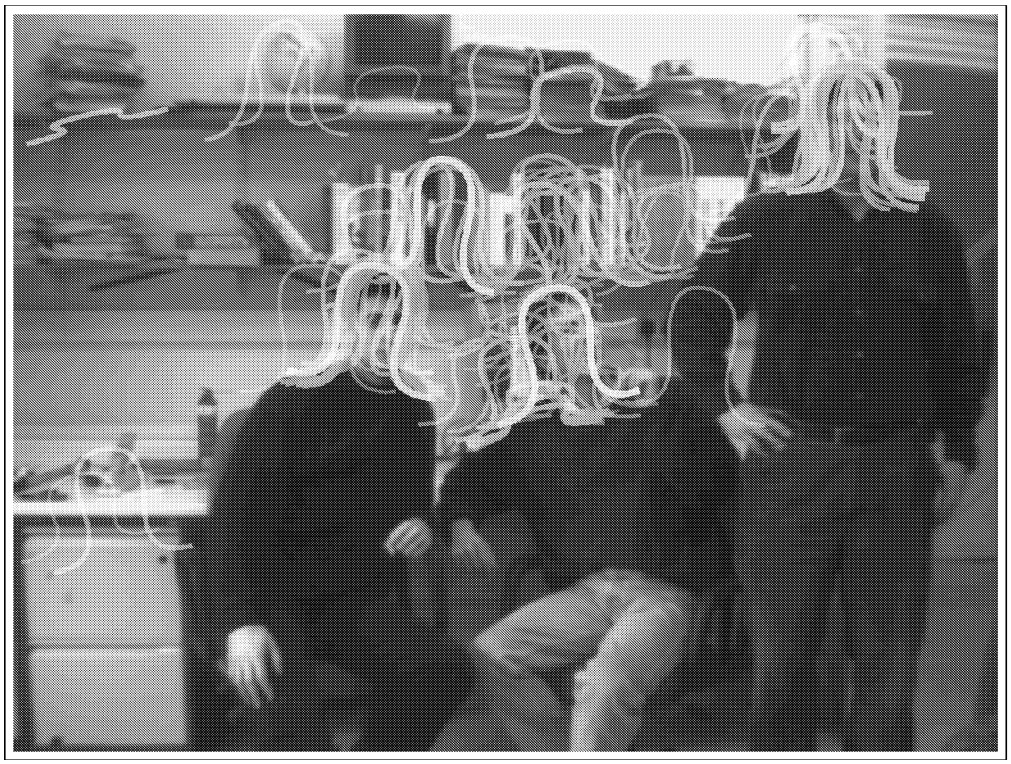


Figure 8. **Sample-set representation of posterior shape distribution** *for a curve with parameters* $\mathbf{X}$*, modelling a head outline. Each sample* $\mathbf{s}^{(n)}$ *is shown as a curve (of varying position and shape) with a mass proportional to the weight* $\pi^{(n)}$*. The prior is uniform over translation, with some constrained Gaussian variability in the remainder of its affine shape-space.*

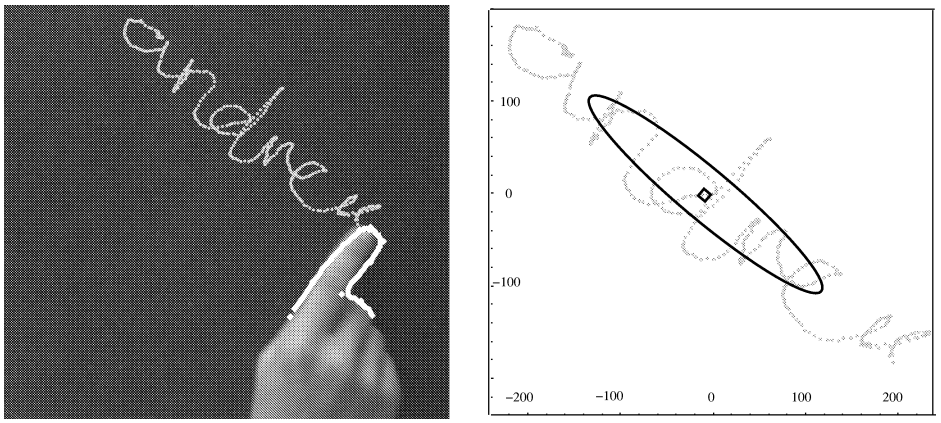


Figure 9. **The moving finger writes.** *The finger trajectory (left) which has a duration of about 10 seconds executes a broad sweep over the plane. If the trajectory is treated as a training set, the learned Gaussian prior is broad, as the covariance ellipse (right) shows. Clearly though, successive positions (individual dots represent samples captured every 20ms) are much more tightly constrained.*

the figure that successive positions are tightly constrained. Although the prior covariance ellipse spans about 300×50 pixels, successive sampled positions are seldom more than 5 pixels apart.

For sequences of images, then, a global prior $p_0(\mathbf{X})$ is not enough. What is needed is a conditional distribution $p(\mathbf{X}_k|\mathbf{X}_{k-1})$ giving the distributions of possibilities for the shape $\mathbf{X}_k$ at time $t = k\tau$ *given* the shape $\mathbf{X}_{k-1}$ at time $t = (k-1)\tau$ (where τ is the time-interval between successive images). This amounts to a "1st order Markov chain" model in shape space in which, although in principle $\mathbf{X}_k$ may be correlated with all of $\mathbf{X}_1 \ldots \mathbf{X}_{k-1}$, only correlation with the immediate predecessor is explicitly acknowledged.

For the sake of tractability, it is reasonable to restrict Markov modelling to linear processes. In principle and in practice it turns out that a 1st order Markov chain is not quite enough, generally, but 2nd order suffices. The detailed arguments for this, addressing such issues as capacity to represent oscillatory signals and trajectories of inertial bodies, can be found in [Blake and Isard, 1998, Chapter 9]. Figure 10 illustrates the point for a practical example. A second order, "Auto-Regressive Process" (ARP) is most concisely expressed by defining a state vector

$$\mathcal{X}_k = \begin{pmatrix} \mathbf{X}_{k-1} \\ \mathbf{X}_k \end{pmatrix}, \tag{5.1}$$

and then specifying the conditional probability density $p(\mathcal{X}_k|\mathcal{X}_{k-1})$. In the case of a linear model, this can be done constructively as follows:

$$\mathcal{X}_k - \overline{\mathcal{X}} = A(\mathcal{X}_{k-1} - \overline{\mathcal{X}}) + B\mathbf{w}_k \tag{5.2}$$

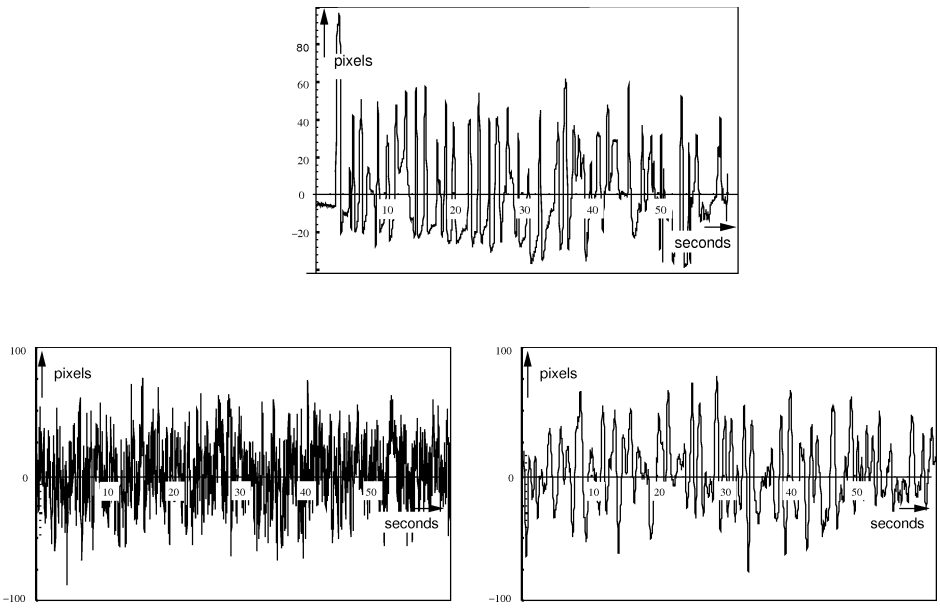


Figure 10. **Motion data from talking lips.** *Training sequence of 60 seconds duration (top). Random simulations of learned models — 1st order (left) and 2nd order (right). Only the 2nd order model captures the natural periodicity (around 1 Hz) of the training set, and spectrogram analysis confirms this.*

where

$$A = \begin{pmatrix} 0 & I \\ A_2 & A_1 \end{pmatrix}, \quad \overline{\mathcal{X}} = \begin{pmatrix} \overline{\mathbf{X}} \\ \overline{\mathbf{X}} \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 0 \\ B_0 \end{pmatrix}. \tag{5.3}$$

and each $\mathbf{w}_k$ is a vector of N_X independent random $\mathcal{N}(0,1)$ variables and $\mathbf{w}_k$, $\mathbf{w}_{k'}$ are independent for $k \neq k'$. This specifies the probable temporal evolution of the shape $\mathbf{X}$ in terms of parameters A, B, and covers multiple oscillatory modes and/or constant velocity motion. The constructive form is attractive because it is amenable to direct simulation, simply by supplying a realisation of the succession of random variates $\mathbf{w}_k$.

6. Learning dynamics

Motion parameters (A, B in this paper) can be set by hand to obtain desired effects, and a logical approach to this has been developed [Blake and Isard, 1998, chapter 9]. Experimentation allows these parameters to be refined by hand for improved tracking but this is a difficult and unsystematic business. What is far more attractive is to learn dynamical models on the basis of training sets. A number of alternative approaches have been proposed for learning dynamics, with a view to gesture-recognition — see for instance [Mardia et al., 1993, Campbell and Bobick, 1995, Bobick and Wilson, 1995]. The requirement there is to learn models that are sufficiently finely tuned to discriminate amongst similar motions. In the context here of the problem of motion tracking, rather different methods are called for to learn models that are sufficiently coarse to encompass all likely motions.

Initially, a hand-built model is used in a tracker to follow a training sequence which must be not be too hard to track. This can be achieved by allowing only motions which are not too fast, and limiting background clutter or eliminating it using background subtraction [Baumberg and Hogg, 1994, Murray and Basu, 1994, Koller et al., 1994, Rowe and Blake, 1996]. Once a new dynamical model has been learned, it can be used to build a more competent tracker, one that is specifically tuned to the sort of motions it is expected to encounter. That can be used either to track the original training sequence more accurately, or to track a new and more demanding training sequence, involving greater agility of motion. The cycle of learning and tracking is illustrated in figure 11. Typically two or three cycles suffice to learn an effective dynamical model.

In mathematical terms, the general problem is to estimate the coefficients A_1, A_2, $\overline{\mathbf{X}}$ and B from a training sequence of shapes $\mathbf{X}_1, \ldots, \mathbf{X}_M$, gathered at the image sampling frequency. Known algorithms to do this are based on the "Maximum Likelihood" principle [Rao, 1973, Kendall and Stuart, 1979] and use variants of "Yule-Walker" equations for estimation of the parameters of autoregressive models [Gelb, 1974, Goodwin and Sin, 1984, Ljung, 1987]. Suitable adaptations for multidimensional shape-spaces are given by [Blake and Isard, 1994, Baumberg and Hogg, 1995b, Blake et al., 1995], with a number of useful extensions in [Reynard et al., 1996]. One example is the scribble in figure 12, learned from the training-sequence in figure 9.

A more complex example is learning the motions of an actor's face, using the shape-space described earlier that covers both rigid and non-rigid motion. Figure 13 illustrates how much more accurately realistic facial motion can be represented

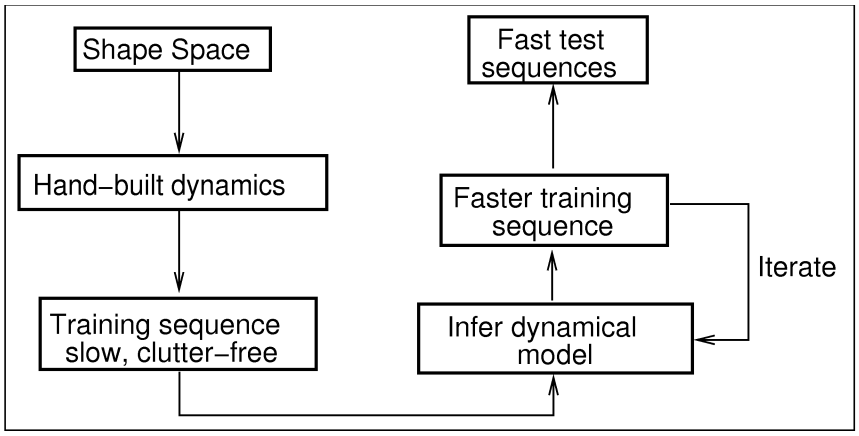


Figure 11. **Iterative learning of dynamics.** *The model acquired in one cycle of learning is installed in a tracker to interpret the training sequence for the next cycle. The process is initialised with a tracker whose prior is based on a hand-built dynamical model.*

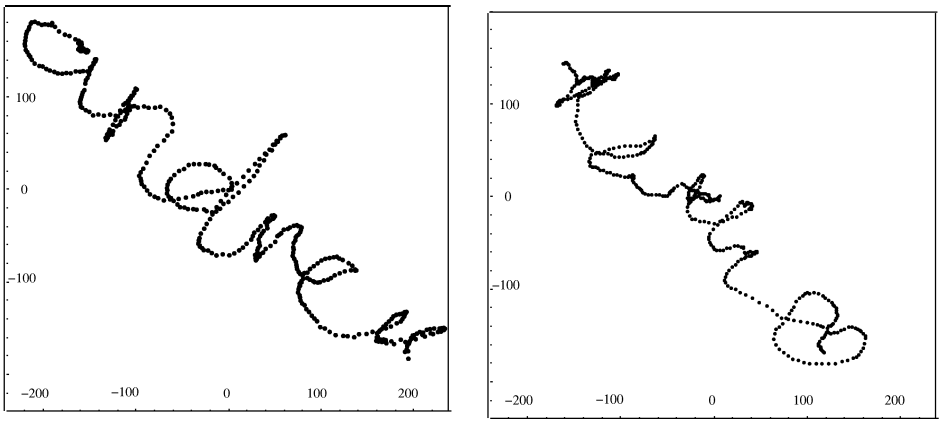


Figure 12. **Scribbling: simulating a learned model for finger-writing.** *A training set (left) consisting of six handwritten letters is used to learn a dynamical model for finger motion. A random simulation from the model (right) exhibits reasonable gross characteristics.*

by a dynamical model which is actually learned from examples.

The learning algorithms referred to above treat the training set as exact whereas in fact it is inferred from noisy observations. Dynamics can be learned directly from the observations using Expectation-Maximisation (EM) [Dempster et al., 1977]. Learning dynamics by EM is suggested by Ljung (1987) and the detailed algorithm is given in [North and Blake, 1997]. It is related to the Baum-Welch algorithm used to learn speech models [Huang et al., 1990, Rabiner and Bing-Hwang, 1993] but with additional complexity because the state-space is continuous rather than discrete. In practice, accuracy of the learned dynamics are significantly improved when EM is used, especially in the case of more coherent oscillations.

An extension of the basic algorithm for *classes* of objects, dealing independently with motion and with variability of mean shape/position over the class, is described in [Reynard et al., 1996]. The same algorithm is also used for modular

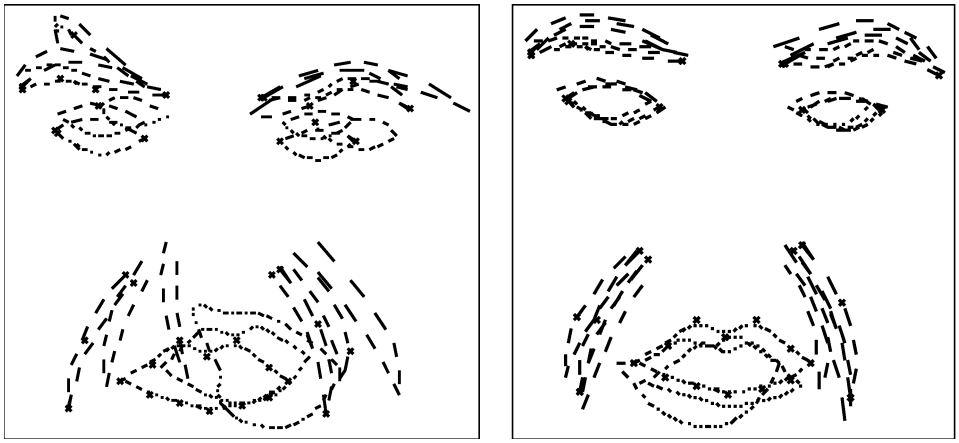


Figure 13. **Trained dynamics for facial motion.** *Hand-built dynamics, exhibited here by random simulation (left) are just good enough, when used in tracking, to gather a training sequence. Trained dynamics (right) however, capture more precisely the constraints of realistic facial motion.*

learning — the aggregation of training sets for which a joint dynamical model is to be constructed.

7. The Condensation algorithm

The CONDENSATION algorithm is a random sampling algorithm for motion tracking using statistical observations and a dynamical prior. It is based on factored sampling but extended to apply iteratively to successive images in a sequence. Similar sampling strategies have appeared elsewhere [Gordon et al., 1993, Kitagawa, 1996], presented as developments of Monte Carlo methods. The methods outlined here are described in detail elsewhere. Fuller descriptions and derivation of the CONDENSATION algorithm are in [Isard and Blake, 1996, Isard and Blake, 1998a].

Given that the estimation process at each time-step is a self-contained iteration of factored sampling, the output of an iteration will be a weighted, time-stamped sample-set, denoted $\mathbf{s}_k^{(n)}$, $n = 1, \ldots, N$ with weights $\pi_k^{(n)}$, representing approximately the conditional state-density $p(\mathcal{X}_k|\underline{\mathbf{Z}}_k)$ at time $t = k\tau$, where $\underline{\mathbf{Z}}_k = (\mathbf{Z}_1, \ldots, \mathbf{Z}_k)$, the history of observations. How is this sample-set obtained? Clearly the process must begin with a prior density and the effective prior for time-step k should be $p(\mathcal{X}_k|\underline{\mathbf{Z}}_{k-1})$. This prior is of course multi-modal in general and no functional representation of it is available. It is derived from the representation as a sample set $\{(\mathbf{s}_{k-1}^{(n)}, \pi_{k-1}^{(n)}),\ n = 1, \ldots, N\}$ of $p(\mathcal{X}_{k-1}|\underline{\mathbf{Z}}_{k-1})$, the output from the previous time-step, to which prediction must then be applied.

The iterative process applied to the sample-sets is depicted in figure 14. At the top of the diagram, the output from time-step $k-1$ is the weighted sample-set $\{(\mathbf{s}_{k-1}^{(n)}, \pi_{k-1}^{(n)}),\ n = 1, \ldots, N\}$. The aim is to maintain, at successive time-steps, sample sets of fixed size N, so that the algorithm can be guaranteed to run within a given computational resource. The first operation therefore is to sample (with replacement) N times from the set $\{\mathbf{s}_{k-1}^{(n)}\}$, choosing a given element with probability $\pi_{k-1}^{(n)}$. Some elements, especially those with high weights, may be

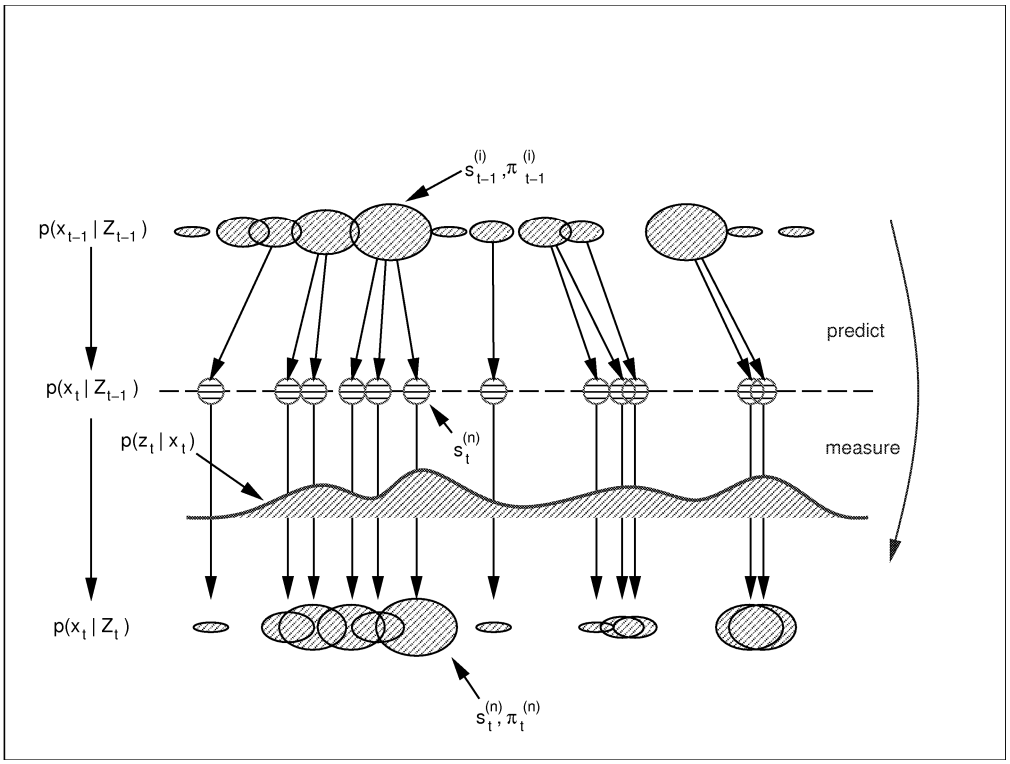


Figure 14. **One time-step in the** CONDENSATION **algorithm.** *Blob centres represent sample values and sizes depict sample weights.*

chosen several times, leading to identical copies of elements in the new set. Others with relatively low weights may not be chosen at all.

Each element chosen from the set is now subjected to a predictive step, using an ARP dynamical model as in equation (5.2). This involves sampling a value of $\mathcal{X}_k$ randomly from the conditional density $p(\mathcal{X}_k|\mathcal{X}_{k-1})$ to form a new set member $\mathbf{s}_k^{(n)}$. Since the predictive step includes a random component, identical elements may now split as each undergoes its own independent random motion step. At this stage, the sample set $\{\mathbf{s}_k^{(n)}\}$ for the new time-step has been generated but, as yet, without its weights; it is approximately a fair random sample from the effective prior density $p(\mathcal{X}_k|\underline{\mathbf{Z}}_{k-1})$ for time-step $t = k\tau$. Finally, the observation step from factored sampling is applied, generating weights from the observation density $p(\mathbf{Z}_k|\mathcal{X}_k)$ to obtain the sample-set representation $\{(\mathbf{s}_k^{(n)}, \pi_k^{(n)})\}$ of state-density for time t.

A good deal of experimentation has been performed in applying the CONDENSATION algorithm to the tracking of visual motion, including moving hands and dancing figures. Perhaps one of the most stringent tests was the tracking of a leaf on a bush, in which the foreground leaf is effectively camouflaged against the background. Results are shown in figure 15 and experimental details can be found in [Isard and Blake, 1996].

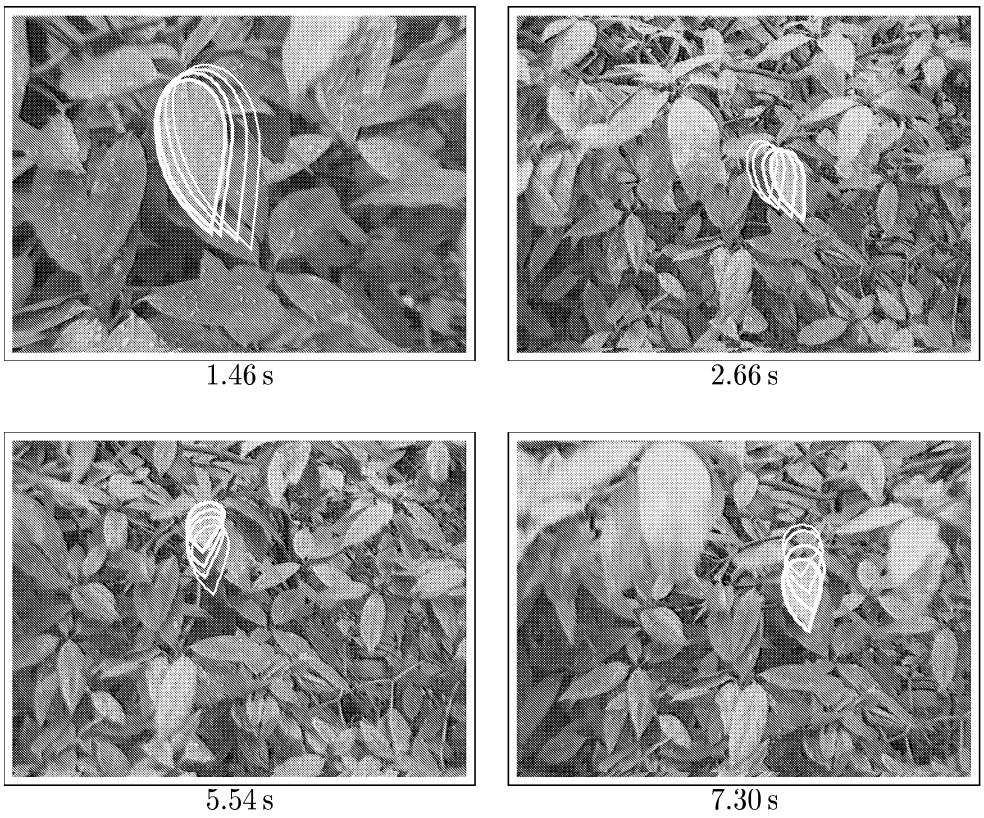


Figure 15. **Tracking with camouflage.** *Stills depict mean contour configurations, with preceding tracked leaf positions plotted at 40 ms intervals to indicate motion.*

8. Dynamics with discrete states

A recent development of the dynamical models already described is to append to the state variable $\mathcal{X}$ a discrete state y_k to make a "mixed" state

$$\mathcal{X}_k^+ = \begin{pmatrix} \mathcal{X}_k \\ y_k \end{pmatrix}, \tag{8.1}$$

where $y_k \in \{1, \ldots, N_S\}$ is drawn from a finite set of discrete states with integer labels. Each discrete state represents a mode of motion such as "stroke", "rest" and "shade" for a hand engaged in drawing. Corresponding to each state $y_{k-1} = i$ there is a dynamical model $p_i(\mathcal{X}_k|\mathcal{X}_{k-1})$ which, in the case of the drawing hand, is likely to be an ARP as in (5.2). The stroke model, for instance, might represent constant velocity motion, whereas shading would be oscillatory. In addition, and independently, state transitions are governed by

$$P(y_k = j|y_{k-1} = i) = T_{i,j},$$

a transition matrix following usual practice for Markov chains. More generally, transition probabilities could be made sensitive to the context $\mathcal{X}_{k-1}$ in state space, so that

$$P(y_k = j|y_{k-1} = i, \mathcal{X}_{k-1}) = T_{i,j}(\mathcal{X}_{k-1}).$$

For example this could be used to express an enhanced probability of transition into the "resting" state when the hand is moving slowly.

Incorporation of mixed states into the CONDENSATION algorithm is straightforward. It involves using the extended state $\mathcal{X}_k^+$ in place of the original $\mathcal{X}_k$, so that a sample $\mathbf{s}_k^{(n)}$ is now a value of the extended state. The prediction step, which generates a new sample $\mathbf{s}_k^{(n)}$ from an old one $\mathbf{s}_{k-1}^{(n)}$ requires a discrete step and a continuous one. First, the discrete state y_k for the new sample is $y_k = j$, chosen randomly, with probability $T_{i,j}$, where i is the discrete state of the old sample. Then the continuous state is chosen by sampling randomly from a continuous density, as in the original algorithm, but now one of several possible densities $p_i(\mathcal{X}_k|\mathcal{X}_{k-1})$ where again i is the discrete state of the old sample.

Experiments with a three-state model for drawing have been described in detail elsewhere [Isard and Blake, 1998b]. In addition to enhancing tracking performance, there is the bonus that the current discrete state y_k can be estimated at each time $t = k\tau$, effectively performing gesture recognition as a side-effect. One interesting variation on the mixed-state theme uses continuous conditional densities $p_i(\mathcal{X}_k|\mathcal{X}_{k-1})$ which are not ARP models. Consider the example of a moving ball, which may occasionally bounce. This could be represented using two states $\{1, 2\}$ in which $i = 1$ stands for the free ballistic motion of the ball, and $i = 2$ is the bounce event. A suitable transition matrix would be

$$T = \begin{pmatrix} 1-\epsilon & \epsilon \\ 1 & 0 \end{pmatrix}$$

in which $0 < \epsilon \ll 1$ so that ballistic motion has a mean duration τ/ϵ between bounces. The fact that $T_{2,2} = 0$ ensures that the model always returns to ballistic motion after a bounce — bouncing at each of two consecutive time-steps is disallowed. Now $p_1(\ldots|\ldots)$ is an ARP for ballistic motion but $p_2(\ldots|\ldots)$ models the instantaneous reversal of velocity normal to the reflecting surface. Details of experiments with such a model are in [Isard and Blake, 1998b] but the results are illustrated in figure 16.

9. Conclusions

A high-speed tour has been given of a framework for probabilistic modelling of shapes in motion, and of their visual observation. The key points are that visual clutter makes motion analysis hard, and demands full-blooded probabilistic mechanisms to handle the resulting uncertainty. Further, prior models of motion and of observation provide powerful constraints, especially so when the models are learned. A more detailed development is given in [Blake and Isard, 1998]. Since that account, several new modelling tools have been developed. First, the contour discriminant is a new observational model that is expressed as a likelihood ratio and takes opacity of objects into account. Second, complex models for combined rigid and nonrigid motion have been constructed, with a new algorithm for decomposing the two components. Third, extending dynamical states to include discrete labels can significantly enhance their power to constrain perceptual interpretation of shape.

Many interesting questions remain to be addressed. One is whether sampling methods for object localisation can be fused elegantly with the CONDENSATION al-

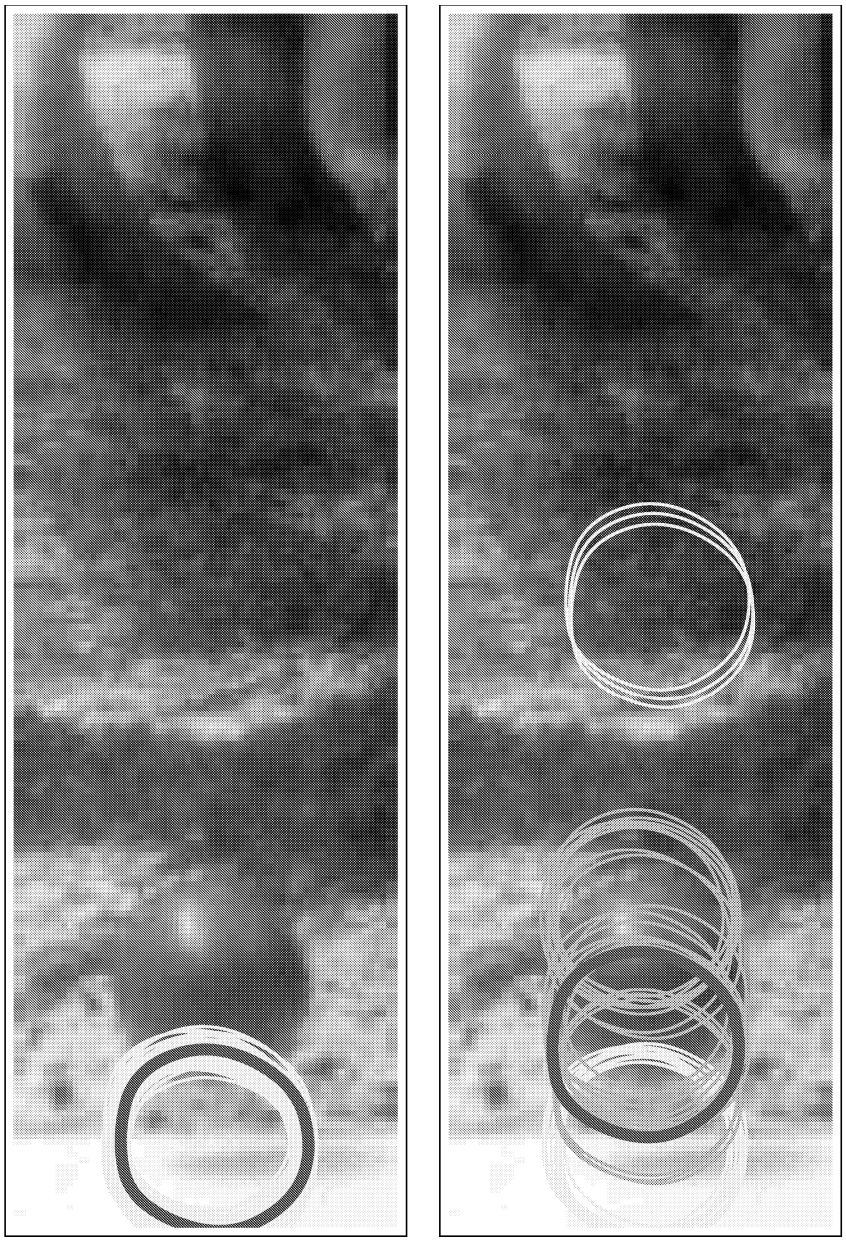


Figure 16. **Mixed states tighten constraints in dynamical models.** *A conventional, continuous-state ARP model (left) used to track ballistic motion, fails unrecoverably as the ball bounces. Introducing an explicit discrete state for the bounce allows the sample set to split, so that a significant proportion are able to track the bounce.*

gorithm, to allow robust handling of "birth" and "death" [Grenander and Miller, 1994] processes in which objects enter and leave the scene. A second is to extend mixed-state models to give reliable gesture recognition on the fly, in a manner that is integrated with the tracking process. A third is to develop algorithms, based on EM, to learn dynamical models from sequences tracked by CONDENSATION, using

the full richness of its probabilistic representation, both for continuous and mixed state systems.

Acknowledgements

The authors would like to acknowledge the support of the EPSRC and of Oxford Metrics Ltd.

References

Barnett, S. (1990). *Matrices: Methods and Applications.* Oxford University Press.

Bascle, B. and Blake, A. (1998). Separability of pose and expression in facial tracking and animation. In *Proc. Int. Conf. on Computer Vision*, 323–328.

Baumberg, A. and Hogg, D. (1994). Learning flexible models from image sequences. In Eklundh, J.-O., editor, *Proc. 3rd European Conference on Computer Vision*, 299–308. Springer–Verlag.

Baumberg, A. and Hogg, D. (1995a). An adaptive eigenshape model. *Proc. British Machine Vision Conf.*, 87–96.

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

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

Blake, A. and Isard, M. (1994). 3D position, attitude and shape input using video tracking of hands and lips. In *Proc. Siggraph*, 185–192. ACM.

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. *Artificial Intelligence*, 78, 101–134.

Bobick, A. and Wilson, A. (1995). A state-based technique for the summarisation and recognition of gesture. In *Proc. 5th Int. Conf. on Computer Vision*, 382–388.

Campbell, L. and Bobick, A. (1995). Recognition of human body motion using phase space constraints. In *Proc. 5th Int. Conf. on Computer Vision*, 624–630.

Cootes, T., Taylor, C., Lanitis, A., Cooper, D., and Graham, J. (1993). Building and using flexible models incorporating grey-level information. In *Proc. 4th Int. Conf. Computer Vision*, 242–246.

Dempster, A., Laird, M., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.*, B, 39, 1–38.

Duda, R. O. and Hart, P. E. (1973). *Pattern Classification and Scene Analysis.* John Wiley and Sons.

Fernyhough, J., Cohn, A., and Hogg, D. (1996). Generation of semantic regions from image sequences. In *Proc. European Conf. Computer Vision*, 2, 475–484.

Freeman, W. and Tenenbaum, J. (1997). Learning bilinear models for two-factor problems in vision. In *Proc. Conf. Computer Vision and Pattern Recognition*, in press.

Gelb, A., editor (1974). *Applied Optimal Estimation.* MIT Press, Cambridge, MA.

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

Goodwin, C. and Sin, K. (1984). *Adaptive filtering prediction and control.* Prentice-Hall.

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. M. (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.

Huang, X., Arika, Y., and Jack, M. (1990). *Hidden Markov Models for Speech Recognition.* Edinburgh University Press.

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

Isard, M. and Blake, A. (1998a). Condensation — conditional density propagation for visual tracking. *Int. J. Computer Vision*, in press.

Isard, M. and Blake, A. (1998b). A mixed-state Condensation tracker with automatic model switching. In *Proc. Int. Conf. on Computer Vision*, 107–112.

Kendall, M. and Stuart, A. (1979). *The advanced theory of statistics, vol 2, inference and relationship.* Charles Griffing and Co Ltd, London.

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.

Knill, D., Kersten, D., and Yuille, A. (1996). A Bayesian formulation of visual perception. In Knill, D. and Richard, W., editors, *Perception as Bayesian inference*, 1–22. Cambridge University Press.

Koenderink, J. and Van Doorn, A. (1991). Affine structure from motion. *J. Optical Soc. America A.*, 8, 2, 337–385.

Koller, D., Weber, J., and Malik, J. (1994). Robust multiple car tracking with occlusion reasoning. In *European Conference on Computer Vision.*

Lanitis, A., Taylor, C., and Cootes, T. (1995). A unified approach to coding and interpreting face images. In *Proc. 5th Int. Conf. Computer Vision*, 368–373.

Ljung, L. (1987). *System identification: theory for the user.* Prentice-Hall.

Mardia, K., Ghali, N., Howes, M., Hainsworth, T., and Sheehy, N. (1993). Techniques for online gesture recognition. *Image and Vision Computing*, 11, 5, 283–294.

MacCormick, J. and Blake, A. (1998). A probabilistic contour discriminant for object localisation. In *Proc. Int. Conf. on Computer Vision*, 390–395.

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

Murray, D. and Basu, A. (1994). Motion tracking with an active camera. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 16, 5, 449–459.

North, B. and Blake, A. (1997). Using expectation-maximisation to learn dynamical models from visual data. In *Proc. British Machine Vision Conf.*, in press.

Rabiner, L. and Bing-Hwang, J. (1993). *Fundamentals of speech recognition.* Prentice-Hall.

Rao, C. (1973). *Linear Statistical Inference and Its Applications.* John Wiley and Sons, New York.

Reynard, D., Wildenberg, A., Blake, A., and Marchant, J. (1996). Learning dynamics of complex motions from image sequences. In *Proc. 4th European Conf. Computer Vision*, 357–368, Cambridge, England.

Ripley, B. and Sutherland, A. (1990). Finding spiral structures in images of galaxies. *Phil. Trans. R. Soc. Lond. A.*, 332, 1627, 477–485.

Rowe, S. and Blake, A. (1996). Statistical mosaics for tracking. *Image and Vision Computing*, 14, 549–564.

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

Tomasi, C. and Kanade, T. (1991). Shape and motion from image streams: a factorization method. *Int. J. Computer Vision*, 9, 2, 137–154.

Ullman, S. and Basri, R. (1991). Recognition by linear combinations of models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13, 10, 992–1006.

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