Invited Paper 22.1
46th Session of the ISI

RECURSIVE METHODS IN IMAGE PROCESSING

P.J. Green and D.M. Titterington
University of Durham and University of Glasgow, UK

1. Introduction

Problems of image processing have been discussed in the engineering literature for many years but, more recently, statisticians have become involved and their journals are now showing indications of this involvement; see Besag (1986), for instance. So far, most of the literature has concerned the restoration of a single frame on which a scene is depicted. The frame is usually two-dimensional and partitioned into a large number, N, of (usually) rectangular pixels; the data consist of the colours, grey-levels or intensities observed on the pixels. In general terms, the objective of any restoration algorithm is to make some sort of inference about the true scene, given the data. Usually the inference takes the form of a restored image, made up of an assignment of colours, grey-levels or intensities to the pixels, and it is anticipated that the resulting restoration will be a better representation of the true scene than was provided by the original observed records.

Of course, image processing is concerned with many other problems besides the direct restoration of a degraded image. Much of what follows may be applicable to such matters as classification, segmentation, tomographic reconstruction and so forth, but for clarity we shall continue to use the terminology of restoration.

So far, most of the work has concentrated on the analysis of a single frame. The objective of this paper is to make a start on the development of methods for the treatment of sequences of images which are assumed to be sequentially related. Although such a framework would include, trivially, the case of replicated observations of the same true scene, our principal target is to treat a sequence of scenes that are not identical but that are temporally correlated.

Applications are multitudinous: they include sequences of overlapping frames of remotely-sensed data, sequences of ultrasonic images of living organisms, in which pulsations and other, more general, types of motion occur, films of moving vehicles, and many other manifestations in medicine, the biological sciences and elsewhere.

One specific example, that might be handled successfully using some of the models proposed here, concerns a sequence of images recorded by gamma camera showing the temporal progress of a radioactively tagged pharmaceutical from the blood stream through the kidneys into the bladder.

The analysis of a sequence of frames, rather than a single one, is one special feature of this paper. Another is the decision to consider recursive methods. One approach would be to gather all the data, from the total of T frames, for instance, and then undertake the analysis. Instead, we shall develop methods that analyse the data frame by frame, in the order in which they are obtained. Behind this decision lies the hope of being able to process the data in real time, thereby creating the restoration of one frame before the next one is considered. Clearly, if such a technique is to be feasible for the treatment of movie film, the stage-by-stage analysis will have to be very fast. As a result, we shall see that "approximations" to normative procedures are almost inevitable and we are, no doubt, still some way from establishing a well-valid-

ated method.

The recursive methods could also be applied to replicated, statistically independent pictures of a constant scene, although an artificial ordering would have to be imposed on the individual frames.

The plan of the paper is as follows. In Section 2 we establish the basic structure of our models for the sequence of true scenes and observed records thereof, and we identify the two problems of creating restorations of the true scenes and estimating unknown parameters in the models. In some existing methods, the latter problem is often ignored. The formulation then leads to two, more specific types of model, both related to models already in use in the analysis of single frames. These are Gibbs distributions (discussed in Section 3) and dynamic linear models (outlined in Section 4). Section 5 describes possibilities for further work.

We shall provide references, to existing work on the use of Gibbs distributions and dynamic linear models, in the corresponding sections. Clearly, the analysis of image sequences is closely related to motion analysis, and the literature on that topic contains reference to yet further general methods. Buxton, Buxton and Stephenson (1984), for instance, follow the motion of objects within a sequence of images by detecting the movement of the edges of the objects; see also Fang and Hwang (1984), Yasumoto and Medioni (1986), Sethi and Jain (1987) and Bresler and Merhav (1987). These authors typically consider the motion of specific features of the scene, whereas our models are aimed at the whole scene.

2. Notation and Basic Structure

The fundamental structure of our problem consists of a sequence of unobservables $\{x_t: t = 1, 2, \ldots\}$, of which x_t represents the true scene depicted in frame t and a corresponding sequence of observed $\{y_t: t = 1, 2, \ldots\}$, where y_t is the record on frame t. The subscript t indicates the correct sequential relationships among the frames. Normally, t will have a largest value, T, corresponding to the end of the data. For modelling purposes, it is sometimes convenient to commence the x sequence with x_0. Both x_t and y_t will be regarded as vectors, with the pixels arranged in some convenient, fixed order, e.g., lexicographically.

We shall also use the following notation:

$$x_{\leq t} = \{x_s: s \leq t\}$$
$$x_{<t} = \{x_s: s < t\},$$

with a corresponding notation for the y's.

In general terms, the problem of restoration is to make some inference about $x_{\leq T}$, given $y_{\leq T}$. If we adopt the recursive approach, then inference about x_t will be based on $y_{\leq t}$, t = 1, ..., T.

2.1 A Markov model

In what follows, the letter p will be used generically to mean "probability density function" or "probability mass function" (for discrete random variables). Any such function will usually involve parameters. For the time being, mention of these will either be omitted, or be made generically, in terms of the letter θ.

Suppose we are at the point of considering frame t. We shall specify a model for the joint distribution of the observed $y_{\leq t}$ and the missing $x_{\leq t}$, namely,

$$p_\theta(x_{\leq t}, y_{\leq t}) .$$

Such a model provides the following (normative) approaches to the objectives of restoration and estimation.

(i) Restoration. This should be based on

$$p_\theta(x_{\leq t}|y_{\leq t}) = p_\theta(x_{\leq t},y_{\leq t})/p_\theta(y_{\leq t}) . \tag{1}$$

(ii) Estimation. If we interpret "normative" estimation to mean likelihood-based estimation, then we should use the likelihood corresponding to the observed data, namely,

$$p_\theta(y_{\leq t}) = \int p_\theta(x_{\leq t},y_{\leq t})dx_{\leq t} . \tag{2}$$

Note that, if the domain of x_t is discrete, then the integral in (2) should be replaced by a summation.

We make the following assumptions:

$$p_\theta(y_t|x_{\leq T},y_{<t}) \equiv p_\theta(y_t|x_t) \tag{3}$$

$$p_\theta(x_t|x_{<t}) \equiv p_\theta(x_t|x_{t-1}) . \tag{4}$$

Identity (4) corresponds to a Markov assumption about the sequence of true scenes, whereas (3) reflects an independence between the noise associated with a particular frame and any other random variable underlying scenes and previous records. As a result, we have (omitting mention of θ),

$$p(x_{\leq t},y_{\leq t}) = \{\prod_{s=1}^{t} p(y_s|x_s)p(x_s|x_{s-1})\}p(x_o) .$$

Although (1) shows how inference may be made about $x_{\leq t}$ from $y_{\leq t}$, we have stated our aim to restore the frames recursively. Thus, of particular interest is

$$\begin{aligned} p(x_t|y_{\leq t}) &= \int_{x_{t-1}} p(x_t,x_{t-1}|y_{\leq t})dx_{t-1} \\ &\propto \int_{x_{t-1}} p(x_t,x_{t-1},y_{\leq t})dx_{t-1} \\ &\propto p(y_t|x_t) \int_{x_{t-1}} p(x_t|x_{t-1})p(x_{t-1}|y_{<t})dx_{t-1} . \end{aligned} \tag{5}$$

Assumptions (3) and (4) also have implications for

$$p(x_t|x_{<t},y_{\leq t}). \tag{6}$$

This is the probability function for x_t given all the preceding true scenes and all the past and present records. If (3) and (4) hold, then (6) becomes

$$p(x_t|x_{<t},y_{\leqslant t}) \propto p(x_t|x_{t-1})\, p(y_t|x_t) . \qquad (7)$$

Although, strictly speaking, (7) is unusable, it does stimulate a recursive procedure if x_t is to be restored after $x_{<t}$ have been treated. To be specific, suppose $\hat{x}_{<t}$ denotes the restored versions of $x_{<t}$. Then, from (7)

$$p(x_t|\hat{x}_{<t},y_{\leqslant t}) \propto p(x_t|\hat{x}_{t-1})\, p(y_t|x_t) \qquad (8)$$

might be used as the sources of a restoration of x_t. This discussion is developed in the next section. Although the recursion is temporal and x_t denotes a whole frame of pixel values, this approach is similar to one cycle of Besag's ICM procedure for restoring a single frame by scanning through the individual pixels and simplifying the formula with the help of a (spatially locally dependent) Markov-random-field (MRF) assumption rather than the temporal Markov assumption.

2.2 Recipes for restoration

As remarked in Section 2.1, the source of restorations will be either $p(x_t|y_{\leqslant t})$, as defined in (5), or $p(x_t|\hat{x}_{<t},y_{\leqslant t})$, from (8). In principle, there are two ways in which such a distribution might yield a single restoration. If we write, temporarily, $p(x_t)$ for either (5) or (8), then we can either compute a deterministic summary parameter, such as its mean or mode, and use that for $\hat{x}_t$, or simulate a value of x_t from $p(x_t)$. The former method is similar to the decision-directed (DD) approach to unsupervised learning problems, whereas the latter is analogous to the technique known as learning with a probabilistic teacher (PT); see, for instance, Chapter 6 of Titterington, Smith and Makov (1985).

There is an element of PT in the use of simulated annealing by Geman and Geman (1984) to approximate the mode of the posterior distribution of x.

In Section 4, we shall see that, in principle, calculation of the appropriate DD restoration is not difficult. In practical terms, things are not so straightforward.

2.3 Recipes for estimation

At this juncture, we reinstate θ to the notation. As remarked in Section 2.1, likelihood inference should be based on $p_\theta(y_{\leqslant t})$, as given in (2). In many contexts, the parameter, θ, can be partitioned into $\theta = (\theta_1, \theta_2)$, where θ_1 defines the distribution of $y_{\leqslant t}$, given $x_{\leqslant t}$, and θ_2 is associated with the distribution of $x_{\leqslant t}$. One can interpret this in two ways. In the former, and as hinted in Section 2.1, one can treat the problem as an incomplete-data problem, with the y_t's observed and the x_t's missing, so that θ_1 defines a conditional distribution and θ_2 the marginal distribution for the x_t's. Alternatively, the Bayesian would regard θ_2 as defining the prior for the x's, whereas θ_1 are the parameters in a "likelihood" for the data, y; see Geman and Geman (1984) and Besag (1986). So far as θ_2 is concerned, maximisation of (2) in the Bayesian

context would therefore constitute an empirical Bayes procedure. Such a procedure is approximated, in the image-restoration context, by Geman and McClure (1985).

For the time being, we shall concentrate on the frequentist interpretation and view the problem as one involving missing data. As a result, we can consider using the EM algorithm of Dempster, Laird and Rubin (1977) to estimate θ at stage t. This would require a sequence of double-step iterations to generate the maximiser, $\hat{\theta}_t$, of $p_\theta(y_{\leq t})$, as the limit of a sequence $\hat{\theta}_t^{(r)}$, r = 1, 2, ... The double step is as follows. (We assume that r iterations have already been completed.)

Define

$$\ell(x_{\leq t}, y_{\leq t}; \theta) = \log_e p_\theta(x_{\leq t}, y_{\leq t}) .$$

E-step: Calculate $Q(\theta) \equiv E\{\ell(x_{\leq t}, y_{\leq t}; \theta) | y_{\leq t}, \hat{\theta}_t^{(r)}\}$

M-step: Find $\theta = \hat{\theta}_t^{(r+1)}$ to maximise $Q(\theta)$.

The sequence $\{p_{\hat{\theta}_t^{(r)}}(y_{\leq t}): r = 0, 1, \ldots\}$ is typically increasing, and convergence to the globally maximised likelihood occurs under certain conditions.

A full-blooded EM procedure at each frame is out of the question, because of the familiar slow convergence of the algorithm and because of complications incurred by the mutual correlations among the x_t's.

The simplest version of the problem occurs if each x_t is scalar (one pixel per frame!) and takes its values from a finite set. If the x_t's were mutually independent, then the observed y_t's would form a sample from a finite mixture distribution (provided assumption (3) holds), and θ_2 would represent the mixing weights. In many such cases, the E- and M-steps are explicit; see Chapter 4 of Titterington et al (1985). If, however, the x_t's are mutually correlated, the M-step may still be explicit but the E-step is complicated. In a sense it is still explicit, but it requires time-consuming forward and backward passes through the data. For the case where the x_t's form a Markov Chain, the example is discussed briefly as Example 4.3.10 of Titterington et al (1985); see also Baum, Petrie, Soules and Weiss (1970) and, for details of the E-step, Pickett and Whiting (1987).

Although we discount the idea of attempting a full EM algorithm at each stage, we propose a one-step recursive approximation to the EM algorithm at each frame, as follows. From our assumptions, (3) and (4), we have

$$\begin{aligned} p_\theta(x_{\leq t}, y_{\leq t}) &= p_\theta(x_t, y_t | x_{<t}, y_{<t})\, p_\theta(x_{<t}, y_{<t}) \\ &= p_\theta(y_t | x_t)\, p(x_t | x_{t-1})\, p_\theta(x_{<t}, y_{<t}) \\ &= \prod_{s=0}^{t} \exp\{\ell_s(\theta)\}, \end{aligned}$$

where $\ell_s(\theta) = \log_e\{p(y_s|x_s)p(x_s|x_{s-1})\}$, s = 1, ..., t, and $\ell_o(\theta) = \log_e\{p(x_o)\}$. Thus,

$$\ell(x_{\leq t}, y_{\leq t}, \theta) = \sum_{s=0}^{t} \ell_s(\theta) .$$

Our recursive procedure takes the following form. Define $L_t(\theta)$ and $\tilde{\theta}_{t-1}$, recursively, by

$$L_t(\theta) = E\{\ell_t(\theta)|y_{\leq t},\tilde{\theta}_{t-1}\} + L_{t-1}(\theta) \qquad (9)$$

and $\tilde{\theta}_t = \arg\max_{\theta} L_t(\theta)$, $t = 1, \ldots$.

For the case where the x_t are independent, see Titterington and Jiang (1983), Titterington (1984a) and Titterington et al (1985, Chapter 6). In some simple problems of that case, reassuring asymptotic properties can be established (Titterington, 1984a). In the case of non-independent x_t, the asymptotic properties are, as yet, unknown.

The complication of the E-step results from the (one-dimensional) temporal correlation among the x_t. In the single-frame case, in which x_t would be the value associated with pixel t, correlations are spatial. Exact formulation of the E-step in this case seems not only complicated but impossible, as pointed out by Kay and Titterington (1986). So far, for the single-frame example, parameter estimation has not been dealt with to a totally satisfactory extent. For the Gibbs-distribution formulation (see Section 3), methods and theory are discussed in Besag (1974, 1986), Geman and Graffigne (1986), Derin (1987), Derin and Elliott (1987) and Possolo (1986). For the dynamic linear model structure discussed in Section 4, see, for instance Tekalp, Kaufman and Woods (1985, 1986).

2.4 Simultaneous, recursive restoration and estimation

In many incomplete-data problems, maximisation of the complete-data likelihood, $p_\theta(x_{\leq t},y_{\leq t})$ is easy. This suggests that, if restorations, $\hat{x}_{\leq t}$, exist, then θ might be estimated by maximising

$$p_\theta(\hat{x}_{\leq t},y_{\leq t}) \quad .$$

It is well known that, if the $\hat{x}_{\leq t}$ have been generated by a DD-type method, then the resulting estimators of θ, θ_t^*, say, tend to be biased; see for instance, Little and Rubin (1983), Titterington (1984b) and Woodward et al (1984), who point out that the use of robust estimators can remove much of the bias. Less bias is likely if PT methods are used to "impute" the x_t's although, for the single-frame problem, DD restorations are likely to provide visually more helpful segmentations of the image (Symons, 1981, Sclove, 1984).

It is not hard to formulate recursive versions of simultaneous restoration and estimation. For a simple case, examples are given in Titterington and Jiang (1983). Detailed examination of their implementation on the present problem is, however, beyond the scope of this paper; see also Kiiveri (1986).

3. Models based on Gibbs distributions

A class of processes that has been shown to be useful in the modelling of single true scenes is that of Markov random fields with local spatial dependence; see Geman and Geman (1984), Besag (1986), and Derin and Elliott (1987). We start our brief discussion of such a formulation for image sequences with assumptions (3) and (4), and the requirement of local dependency both within a frame and between successive frames.

The connection with Gibbs distributions provided by the Hammersley-

Clifford theorem (see, for example, Besag (1974)) will be applied to the conditional distributions $p(x_t|x_{<t})$. Let I denote the set of all pixels of x_t, and suppose we specify a graph with I as the set of vertices. A clique is a set of vertices such that every distinct pair of vertices in the set are neighbours with respect to the graph. We extend our notation for x and y to allow a second subscript to indicate a pixel or set of pixels. Then the conditional distributions must have the representation

$$p(x_t|x_{<t}) = \{Z_t(x_{<t})\}^{-1}\exp\{-\sum_c V_c(x_{t,c};x_{<t})\}$$

where Z_t is a normalising constant, and $\{V_c\}$ a set of potential functions indexed by cliques c. By virtue of assumption (4), these potentials must depend on $x_{<t}$ only through x_{t-1}, and, we will suppose, only on a subset thereof, say x_{t-1,c^*}. Thus

$$p(x_t|x_{<t}) = \{Z_t(x_{t-1})\}^{-1}\exp\{-\sum_c V_c(x_{t,c};x_{t-1,c^*})\}$$

from which it follows that

$$p(x_{t,i}\,|\,x_{<t},x_{t,I\setminus i}) = p(x_{t,i}\,|\,x_{t-1,\pi i},x_{t,\partial i}) \qquad (10)$$

where $\pi i = \bigcup_{c\ni i} c^*$ and $\partial i = \bigcup_{c\ni i} c$ are, typically, small neighbourhoods of pixel i.

Local dependence would also be assumed in the 'noise' distribution (3): for example, in the extreme case of independent records with no blurring, and assuming that x_t and y_t are located on the same pixel grid I,

$$p(y_t|x_t) = \prod_{i\in I} p(y_{t,i}|x_{t,i}) \,. \qquad (11)$$

From (10) and (11) it follows that

$$p(x_{t,i}\,|\,y_{\leq t},x_{<t},x_{t,I\setminus i}) \propto p(y_{t,i}|x_{t,i})p(x_{t,i}|x_{t-1,\pi i},x_{t,\partial i}). \qquad (12)$$

Such a representation of the local conditional posterior distributions offers the prospect of efficient algorithms based on local computations, much in the spirit of Geman and Geman (1984) and Besag (1986).

4. The Linear Gaussian model

In one important special case, the ideas of Section 2 lead to explicit and exact algorithms with no necessity for approximation or stochastic sampling. It is a case of some practical interest, and under additional assumptions we shall see that it can lead to algorithms that are very fast indeed.

4.1 Model assumptions

The Markov assumption (4) on the sequence of true scenes (x_t) will be satisfied if we assume that this sequence follows a first-order matrix autoregression of the form

$$x_t = Gx_{t-1} + \eta_t \,. \qquad (13)$$

Here G is a fixed square matrix and (η_t) is a sequence of innovations: random vectors which we will take to have the multivariate Normal distribution $N(\mu_\eta,V_\eta)$, with η_t independent of $x_{<t}$. Suppose we start with $x_o \sim N(\mu,V)$.

In a similar vein, we suppose that the noisy records (y_t) are generated

by

$$y_t = Hx_t + \varepsilon_t \tag{14}$$

where H is a fixed matrix, and $\varepsilon_t \sim N(\mu_\varepsilon, V_\varepsilon)$, independent of $x_{\leq T}$ and $y_{<t}$. This supposition ensures that our basic assumption (3) is satisfied.

When both of these linear Gaussian assumptions hold, the entire aggregate $(x_{\leq T}, y_{\leq T})$ has a joint Gaussian distribution: we exploit this to derive explicit algorithms.

This model has parameters $(\mu, V, G, \mu_\eta, V_\eta, H, \mu_\varepsilon, V_\varepsilon)$. Of these, G and V_η govern the spatial and temporal continuity in the true scenes (x_t), while H determines the blurring and other deterministic degradation in the observed records, and V_ε controls the variance and correlation in the noise. In an application, some of these parameters may have known values, or be estimated from training data; others will need to be estimated during the course of (recursive) restoration of the sequence (x_t).

4.2 The Kalman Filter

Models of the form (13,14) are of course well known in the control and time-series literature, where they are known as general state-space models or dynamic linear models; see, e.g. Jones (1966), Harrison and Stevens (1976). A good general reference is Anderson and Moore (1979). In such a setting, the dimensions of x_t and y_t would typically be much smaller than is necessary in the image-processing context, and H in (14) would represent the design matrix of a regression model, with values typically known, and changing with t.

Whatever the setting, the consequence of the linear Gaussian assumptions is that not only does the conditional posterior (7) take a simple form, but the integral in (5) has a convolution structure. The result is that, conditional on $y_{\leq t}$, x_t has a Gaussian distribution with

$$E(x_t | y_{\leq t}) = m_t$$

$$\mathrm{var}(x_t | y_{\leq t}) = V_t$$

where m_t and V_t satisfy the recurrence relations

$$V_t = \left[(GV_{t-1}G^T + V_\eta)^{-1} + H^T V_\varepsilon^{-1} H \right]^{-1} \tag{15}$$

$$m_t = (Gm_{t-1} + \mu_\eta) + V_t H^T V_\varepsilon^{-1} (y_t - \mu_\varepsilon - H(Gm_{t-1} + \mu_\eta)). \tag{16}$$

This is known as the Kalman filter; see e.g., Harrison and Stevens (1976). It constructs not only the posterior mode/expectation, conditional on past data, but also the variance of the posterior distribution, so that confidence statements can be made.

This use of the dynamic linear model and the Kalman filter for a sequence of images can be regarded as an extension of their application to a single frame, treated as a sequence of individual pixels, or complete rows or columns, visited recursively. For applications to a single frame, see for example Nahi (1972), Katayama and Kosaka (1978), Katayama (1979), Biemond and Gerbrands (1979, 1980), Wu (1985) and, particularly relevant to the context of this paper, Biemond, Rieske and Gerbrands (1983). There are also close connections with Markov mesh models (Abend, Harley and Kanal, 1965; Kanal, 1980).

At least when all the parameters are assumed known, the recurrence should be initiated by setting $m_o = \mu$ and $v_o = V_o$, and running (15) and (16) from t=1. A possible alternative is to set $V_o = \infty$, formally: then m_o is eliminated from (16), and m_1 is just a least squares estimate based on y_1. We would often wish to assume that, unconditionally, the process (x_t) is stationary, so that its parameters satisfy

$$\mu = G\mu + \mu_\eta \quad \text{and} \quad V = GVG^T + V\eta \quad .$$

Whether or not this holds, however, the recurrence (15) for the posterior variance converges: we do not need to place constraints on the eigenvalues of G. Our limited experience, with practically sensible values for the parameters, suggests that in fact convergence occurs quite quickly: thus, although (15) and (16) represents a non-stationary recurrence (corresponding to the increase in information about x_t in $y_{\leq t}$ as t increases), it may in fact be possible to approximate this by a stationary recurrence. This and other approximations and simplifying assumptions are likely to be needed in practice: the explicit form of (15) and (16) is deceptive about the amount of computing entailed. For example, to handle 512×512 images means inverting matrices with 2^{36} elements in equation (15).

4.3 Parameter estimation

In practice, at least some of the parameters $(\mu, V, G, \mu_\eta, V_\eta, H, \mu_\varepsilon, V_\varepsilon)$ may be unknown. As in Section 2, we will use θ to denote the unknown parameter, generically, and consider estimation of θ simultaneously with restoring (x_t) using its posterior distribution given $y_{\leq t}$. It will usually be appropriate at least to assume particular functional forms for any unknown variance matrices, so that θ will consist of only a moderate number of independent components.

There are a number of challenging estimation problems in the present context. We will follow the spirit of recursive restoration and allow only a single pass through the data, temporally. Thus at time t, given $y_{\leq t}$, we both restore x_t (and give a measure of its uncertainty) and update our estimate of θ. In the presence of unknown parameter values, the Kalman filter is no longer strictly optimal, of course. Further, we are not allowing the revision of an earlier restoration x_s, $s < t$, in the light of information gained about θ at time t. This takes us still further from optimality, but the prohibition we have imposed will be realistic in much real-time image processing.

We therefore consider the recursive procedure defined by (9) as applied to our model (13,14). We omit $\ell_o(\theta)$ as containing rather little information about θ.

Suppose the Kalman filter recursion (15,16) is used to compute m_t and V_t before updating θ, so that

$$\begin{aligned} m_t &= E(x_t | y_{\leq t};\ \tilde{\theta}_{t-1}) \\ V_t &= \mathrm{var}(x_t | y_{\leq t};\ \tilde{\theta}_{t-1}) \ . \end{aligned} \tag{17}$$

Then $\theta = \tilde{\theta}_t$ should be chosen to maximise

$$L_t(\theta) = \sum_{s=1}^{t} E(\ell_s(\theta) | y_{\leq s};\ \tilde{\theta}_{s-1}) \ . \tag{18}$$

Omitting μ_η and μ_ε for simplicity (the general case only adds notational difficulty), we can write

$$E(\ell_s(\theta)|y_{\leqslant s};\tilde{\theta}_{s-1}) = -\tfrac{1}{2}\log\det V_\varepsilon - \tfrac{1}{2}\log\det V_\eta$$
$$-\tfrac{1}{2}E\{\mathrm{tr}[V_\varepsilon^{-1}(y_s - Hx_s)(y_s - Hx_s)^T]\,|\,y_{\leqslant s};\tilde{\theta}_{s-1}\}$$
$$-\tfrac{1}{2}E\{\mathrm{tr}[V_\eta^{-1}(x_s - Gx_{s-1})(x_s - Gx_{s-1})^T]\,|\,y_{\leqslant s};\tilde{\theta}_{s-1}\}\ . \qquad (19)$$

Most of the expectations required can be simply expressed using (17). It remains necessary to find

$$E(x_{s-1}x_s^T|y_{\leqslant s};\tilde{\theta}_{s-1}) \quad\text{and}\quad E(x_{s-1}x_{s-1}^T|y_{\leqslant s};\tilde{\theta}_{s-1})\ .$$

Some rather tedious algebra (c.f. Shumway and Stoffer, 1982, Appendix) leads to

$$E(x_{s-1}|y_{\leqslant s};\tilde{\theta}_{s-1}) = m^*_{s-1} + V^*_{s-1}G^TZ(m_s - Gm^*_{s-1}) = \tilde{m}_{s-1},\ \text{say}$$

$$\mathrm{cov}(x_s,x_{s-1}|y_{\leqslant s};\tilde{\theta}_{s-1}) = V_sZGV^*_{s-1} = V_{s,s-1},\ \text{say}$$

$$\mathrm{var}(x_{s-1}|y_{\leqslant s};\tilde{\theta}_{s-1}) = V^*_{s-1} - V^*_{s-1}G^TZGV^*_{s-1}$$
$$+ V^*_{s-1}G^TZV_sZGV^*_{s-1} = \tilde{V}_{s-1},\ \text{say}$$

where $m^*_{s-1} = E(x_{s-1}|y_{\leqslant s-1};\tilde{\theta}_{s-1})$

$$V^*_{s-1} = \mathrm{var}(x_{s-1}|y_{\leqslant s-1};\tilde{\theta}_{s-1})$$

and $Z = (GV^*_{s-1}G^T + V_\eta)^{-1}$. Note that the asterisks label quantities from the Kalman filter computed <u>after</u> updating the parameters.

Substituting into (18), we find

$$L_t(\theta) = -\frac{t}{2}\log\det V_\varepsilon - \frac{t}{2}\log\det V_\eta$$
$$-\tfrac{1}{2}\mathrm{tr}[V_\varepsilon^{-1}\{A_t - HB_t^T - B_tH^T + HC_tH^T\}]$$
$$-\tfrac{1}{2}\mathrm{tr}[V_\eta^{-1}\{C_t - GE_t^T - E_tG^T + GF_tG^T\}] \qquad (20)$$

where $A_t = \sum_{s=1}^{t} y_sy_s^T$, $B_t = \sum_{s=1}^{t} y_sm_s^T$, $C_t = \sum_{s=1}^{t}(m_sm_s^T + V_s)$

$$E_t = \sum_{s=1}^{t}(m_s\tilde{m}_{s-1}^T + V_{s,s-1})$$

and $$F_t = \sum_{s=1}^{t}(\tilde{m}_{s-1}\tilde{m}_{s-1}^T + \tilde{V}_{s-1}).$$

This expression is easily minimised over choice of G, H, V_η and V_ε, subject to any linear constraints of parameterisation. In particular, the 'non-parametric' estimates, free of any such constraints, take the simple forms

$$
\begin{aligned}
G &= E_t F_t^{-1} \\
H &= B_t C_t^{-1} \\
V_\varepsilon &= t^{-1}\{A_t - B_t C_t^{-1} B_t^T\} \\
\text{and } V_\eta &= t^{-1}\{C_t - E_t F_t^{-1} E_t^T\} .
\end{aligned} \tag{21}
$$

4.4 Practical implementation of the Kalman Filter for image sequences

As mentioned earlier, operation of the Kalman filter ostensibly involves manipulations of very large matrices, and this is also true of the method of parameter estimation in subsection 4.3.

There are modest savings to be made by using a stationary approximation to the Kalman filter, but this is incompatible with updating of parameter estimates. Another possibility is replacement of the linear operators concerned by 'local' approximations of limited bandwidth. For example, if G, H and $V_t H^T V_\varepsilon^{-1}$ have only a few non-zero terms in each row, the recurrence (16) can be computed very quickly. While such approximations, in conjunction with assuming V_t stationary in t, can be algebraically tenable, they are unappealing as they involve interaction between assumptions on (13) and on (14). Again, this does not tie in well with recursive parameter estimation.

One general approach that seems promising is available if G, H, V_ε, V_η and V_o can be simultaneously diagonalised (by similarity transforms). Then (15) can be diagonalised, and the result used in (16) to implement the Kalman filter entirely with scalar operations on the various eigenvalues, in parallel. But what prospect is there of simultaneous diagonalisation being possible? Suppose now that both (x_t) and (y_t) are located on a two-dimensional rectangular grid of dimensions m by n. It is convenient to index mn×mn matrices, with both rows and columns corresponding to pixels in lexicographic order, by pairs of double subscripts, e.g. $(M_{ijk\ell})$, where $1 \leq i, k \leq m$ and $1 \leq j, \ell \leq n$. Consider such a matrix, with the special form

$$
\begin{aligned}
M_{ijk\ell} = {} & \alpha I[i=k,\ j=\ell] + \beta_1 I[i=k,\ j=\ell\pm 1] \\
& + \beta_2 I[i=k\pm 1,\ j=\ell] \\
& + \gamma I[i=k\pm 1,\ j=\ell\pm 1] .
\end{aligned}
$$

where $I[\ ]$ denotes the indicator function. For example, assuming that H had such a form would specify that blurring extended only to the eight neighbouring pixels (four orthogonal and four diagonal) and had certain symmetry properties. All such matrices have the same eigenvectors (Ord, 1975, Besag, 1978). This is easily seen by noting that they can be expressed as sums of Kroneker products of tridiagonal Toeplitz matrices and using a result in Bellman (1960, p.230).

The eigenvalues and eigenvectors involve only sines and cosines at Fourier frequencies, so the Kalman filter could be implemented efficiently if G, H, V_o, V_η and V_ε all had this structure. This may be acceptable: no special treatment is needed at the edges of the image, and the innovations η_t and ε_t are first-order spatial moving average processes. However, it does seem rather restrictive.

A wider range of possibilities results if we make assumptions of toroidal

stationarity instead. This means that all of the matrices have block-circulant structure, i.e. $M_{ijk\ell}$ depends only on (k-i) mod m and (ℓ-j) mod n. It identifies opposite edges of the image to give two-dimensional periodicity, but removes the bandwidth restriction above. All block-circulant matrices are diagonalised by the (two-dimensional, discrete) Fourier transform, a unitary operation (see, e.g., Hunt, 1973). Thus the entire Kalman filter recursion can be carried out in the Fourier domain using only scalar operations. The only assumptions are spatial stationarity and periodicity of all of the stochastic processes involved. In practice we believe that this mis-treatment of the edges of the image will not have a severe effect: various tricks borrowed from time series analysis might be used if necessary. Imposing periodicity in addition to stationarity is equivalent to replacing a Toeplitz matrix by its circulant approximation, as discussed by Gray (1972) and Sherman (1985). Using the Fast Fourier Transform or its generalisations leads to very rapid computation.

4.5 Parameter estimation by Fourier methods

Under assumptions of toroidal stationarity, the estimation method of subsection 4.3 can also be handled by Fourier methods. Define U to be the unitary matrix with $U_{ijk\ell} = (mn)^{-\frac{1}{2}}\nu^{-ik}\omega^{-j\ell}$ where ν and ω are respectively m^{th} and n^{th} complex roots of unity. Providing we interpret the transpose symbols in (20) as complex conjugate transposes, the value of this expression is unchanged if we replace each of the matrices M = G, H, V_η, V_ε, A_t, B_t, C_t, E_t and F_t by UMU*. In the case of a block-circulant M, UMU* is the diagonal matrix whose entries are the unnormalised, two-dimensional discrete Fourier coefficients of an arbitrary row of M. Of course, A_t, B_t, etc., depend on the data and are not block-circulant, but it is clear from the form of the trace terms in (20) that only the diagonal elements of UA_tU^*, etc., are required. Such diagonal elements are simply sums, over s = 1 to t, of the squared moduli of the Fourier coefficients of y_s.

The approximate maximum likelihood estimator can therefore be found quite simply. For example, free of any constraints except toroidal stationarity, we obtain estimates again given by expression (21), where each matrix is now interpreted as the diagonal part of its Fourier transform. More interestingly, we can estimate G and H subject to parametric constraints. Because of the form of (20) the results are weighted least squares estimates (based on complex data). For example, if we revert to a single subscript for pixels, and let v_i, a_i, b_i, c_i and h_i denote the i^{th} diagonal elements of UMU* for M = V_ε, A_t, B_t, C_t and H, then from (20) we see that $\{h_i\}$ should be chosen to minimise

$$\sum_i v_i^{-1}(a_i - c_i^{-1}|b_i|^2) + \sum_i v_i^{-1}c_i|h_i - c_i^{-1}b_i|^2 .$$

Thus if H is assumed symmetric (blurring symmetric, not necessarily isotropic), h_i is estimated as $(2c_i)^{-1}(b_i + b_i^*)$. If H was parameterised to have only its first p Fourier components non-zero, then h_i is estimated as $c_i^{-1}b_i$ for i = 1, ..., p, and zero otherwise. Assumptions of simple form in the original non-Fourier domain (e.g. finite range blur) would lead to explicit least squares computations regressing $c_i^{-1}b_i$ on appropriate basis vectors, with weights $v_i^{-1}c_i$.

The mean vectors μ_η and μ_ε can be allowed to be non-zero, and estimated in a similar fashion. If we impose toroidal stationarity on the mean structure as well, these will each have only one non-zero Fourier component but more generally it may be appropriate to allow non-stationarity in the mean (see Hunt, 1977). For any Fourier component of μ_η and μ_ε allowed to be non-zero, the corresponding components of $\{m_s\}$ and $\{y_s\}$ are mean-corrected before being used to define A_t, etc., leading to estimates of G, H, V_η and V_ε much as before. The estimates of μ_η and μ_ε result from simple averaging.

4.6 Other aspects of the state space model

There are a number of ramifications of the ideas in this section that it seems appropriate to mention briefly.

Many real image sensing and recording processes operate non-linearly. Hunt (1977) argues that a model of the form

$$y = s(Hx) + \varepsilon$$

will often be adequate, perhaps after non-linear transformation of x_t or y_t. Here the function s, representing for example a film characteristic curve, operates component-wise. Hunt's paper, which addresses only the single-image problem, uses this replacement for our (14) together with a Gaussian prior on x and circulant approximations to Toeplitz matrices to derive an iterative algorithm for maximum a posteriori restoration of x using the Fourier transform.

A possible approach to approximating the recursive estimation method of subsection 4.3 uses ideas analogous to those of Titterington and Jiang (1983). The expectation in (19) would be replaced by an imputation procedure, in which a random deviate from the $N(m_t, V_t)$ distribution would be used to represent x_t. This leads to a convenient rank-one update of the matrices C_t and F_t so that estimates can be obtained from (21).

Finally, we should mention the restoration of x_t and estimation of θ given future as well as past data. This problem is no longer recursive, but leads to some similar computations in the linear Gaussian case. Shumway and Stoffer (1982) consider this in detail for an ordinary multivariate time series: their approach to the exact EM estimation of θ using the Kalman filter involves iterating by alternately scanning forwards and backwards through the sequence. This would doubtless be extremely expensive in our image-processing context.

5. Discussion and further work

It will be clear that we have not attempted a complete treatment of the topic of our title. Many ramifications of the approach we have described remain unexplored, and this section serves to record the future work we propose.

The Kalman filter approach described in Section 4 has been implemented, together with recursive parameter estimation, and our presentation in Tokyo will include some illustrations of this in action.

We intend to develop specific methods based on (12), probably incorporating stochastic imputation to obtain practical approximations. This seems the best prospect for progress where the linear Gaussian assumptions are untenable.

None of the methods described is well tuned to the classic computer-vision problems of motion of solid objects. The suggestion by Kelly (1986) of increasing the dimension of x_t to include velocity information seems a promisin approach, and one which we intend to investigate.

Acknowledgement

The authors are grateful to Julian Besag for helpful discussions and encouragement.

BIBLIOGRAPHY

Abend, K., Harley, T.J. and Kanal, L.N. (1965). Classification of binary random patterns. IEEE Trans. Inf. Thy., IT-11, 538-544.

Anderson, B.D.O. and Moore, J.B. (1979). Optimal filtering. Prentice Hall.

Baum, L.E., Petrie, T., Soules, G. and Weiss, N. (1970). A maximisation technique occurring in the statistical analysis of probabilistic functions of Markov Chains. Ann. Math. Statist., 41, 164-171.

Bellman, R.E. (1960). Introduction to matrix analysis. New York: McGraw-Hill.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). J.R. Statist. Soc. B, 36, 192-236.

Besag, J. (1978). Contribution to discussion of paper by Bartlett. J.R. Statist. Soc. B, 40, 165-166.

Besag, J. (1986). On the statistical analysis of dirty pictures (with discussion). J.R. Statist. Soc. B, 48, 259-302.

Biemond, J. and Gerbrands, J.J. (1979). An edge-preserving, recursive, noise-smoothing algorithm for image data. IEEE Trans. Syst. Man. Cybernet., SMC-9, 622-627.

Biemond, J. and Gerbrands, J.J. (1980). Comparison of some two-dimensional recursive point-to-point estimators based on a DPCM image model. IEEE Trans. Syst. Man. Cybernet., SMC-10, 929-936.

Biemond, J., Rieske, J. and Gerbrands, J.J. (1983). A fast Kalman filter for images degraded by both blur and noise. IEEE Trans. Acoust. Speech Signal Proc., ASSP-31, 1248-1256.

Bresler, Y. and Merhav, S.J. (1987). Recursive image registration with application to motion estimation. IEEE Trans. Acoust. Speech Signal Proc., ASSP-35, 70-85.

Buxton, B.F., Buxton, H. and Stephenson, B.K. (1984). Parallel computations of optic flow in early image processing. Proc. IEEE., Ser. F, 131, 593-602.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). J.R. Statist. Soc. B, 39, 1-38.

Derin, H. (1987). Estimating components of univariate Gaussian mixtures using Prony's method. IEEE Trans. Pattern Anal. Mach. Intell., PAMI-9, 142-148.

Derin, H. and Elliott, H. (1987). Modelling and segmentation of noisy and textured images using Gibbs random fields. IEEE Trans. Pattern Anal. Mach. Intell., PAMI-9, 39-55.

Fang, J.Q. and Hwang, T.S. (1984). Some experiments on estimating the 3-D motion parameters of a rigid body from two consecutive image frames. IEEE Trans. Pattern Anal. Mach. Intell., PAMI-6, 545-554.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell., PAMI-6, 721-741.

Geman, S. and Graffigne, C. (1986). A consistency theorem for Markov random fields. Technical report, Brown University.

Geman, S. and McClure, D.E. (1985). Bayesian image analysis: an application to single photon emission tomography. Proc. Amer. Statist. Assoc., Stat. Comp. Sect., 12-18.

Gray, R.M. (1972). On the asymptotic eigenvalue distribution of Toeplitz matrices. IEEE Trans. Inf. Thy., IT-18, 725-730.

Harrison, P.J. and Stevens, C.F. (1976). Bayesian forecasting (with discussion). J.R. Statist. Soc. B, 38, 205-247.

Hunt, B.R. (1973). The application of constrained least squares estimation to image restoration by digital computer. IEEE Trans. Computers, C-22, 805-812.

Hunt, B.R. (1977). Bayesian methods in nonlinear digital image restoration. IEEE Trans. Computers, C-26, 219-229.

Jones, R.H. (1966). Exponential smoothing for multivariate time series. J.R. Statistic. Soc. B, 28, 241-251.

Kanal, L.N. (1980). Markov mesh models. In Image Modelling. New York: Academic Press.

Katayama, T. (1979) Restoration of noisy images using a two-dimensional linear model. IEEE Trans. Syst. Man. Cybernet., SMC-9, 711-717.

Katayama, T. and Kosaka, M. (1978). Smoothing algorithms for two-dimensional image processing. IEEE Trans. Syst. Man. Cybernet., SMC-8, 62-66.

Kay, J.W. and Titterington, D.M. (1986). Image labelling and the statistical analysis of incomplete data. Proc. 2nd Int. Conf. Image Processing and Applications. Conf. Publ. No. 265, London: IEE, p. 44-48.

Kiiveri, H. (1986). Contribution to discussion of paper by Besag. J.R. Statist. Soc. B, 48, 293-294.

Kelly, F.P. (1986). Contribution to discussion of paper by Besag. J.R. Statist. Soc. B, 48, 287.

Little, R.J.A. and Rubin, D.B. (1983). On jointly estimating parameters and missing values by maximizing the complete data likelihood. Amer. Statist., 37, 218-220.

Nahi, N.E. (1972). Role of recursive estimation in statistical image enhancement. Proc. IEEE, 60, 872-877.

Ord, J.K. (1975). Estimation methods for models of spatial interaction. J. Amer. Statist. Assoc., 70, 120-126.

Pickett, E.E. and Whiting, R.G. (1987). On the estimation of probabilitistic functions of Markov chains. Proc. Conf. Model Oriented Data Analysis, Eisenach, DDR, Springer, to appear.

Possolo, A. (1986). Estimation of binary Markov random fields. Technical Report no. 77, Dept. of Statistics, University of Washington.

Sclove, S.C. (1984). Reply to Titterington (1984b). IEEE Trans. Pattern Anal. Mach. Intell., PAMI-6, 657-658.

Symons, M.J. (1981). Clustering criteria and multivariate normal mixtures. Biometrics, 37, 35-43.

Sethi, I.K. and Jain, R. (1987). Finding trajectories of feature points in a monocular image sequence. IEEE Trans. Pattern Anal. Mach. Intell., PAMI-9, 56-73.

Sherman, P.J. (1985). Circulant approximations to Toeplitz matrices and related quantities with application to stationary random processes. IEEE Trans. Acoust. Speech and Signal Proc., ASSP-33, 1630-1632.

Shumway, R.H. and Stoffer, D.S. (1982). An approach to time series smoothing and forecasting using the EM algorithm. J. Time Series Anal., 3, 253-264.

Tekalp, A.M., Kaufman, H. and Woods, J.W. (1985). Fast recursive estimation of parameters of a space-varying auto-regressive image model. IEEE Trans. Acoust. Speech Signal Proc., ASSP-33, 469-472.

Tekalp, A.M., Kaufman, H. and Woods, J.W. (1986). Identification of image and blur parameters for the restoration of noncausal images. IEEE Trans. Acoust. Speech Signal Proc., ASSP-34, 963-972.

Titterington, D.M. (1984a). Recursive parameter estimation using incomplete data. J.R. Statist. Soc. B, 46, 257-267.

Titterington, D.M. (1984b). Comment on a paper by Sclove. IEEE Trans. Pattern Anal. Mach. Intell., PAMI-6, 656-657.

Titterington, D.M. and Jiang, J-M. (1983). Recursive estimation procedures for missing-data problems. Biometrika, 70, 613-624.

Titterington, D.M., Smith, A.F.M. and Makov, U.E. (1985). Statistical Analysis of Finite Mixture Distributions. Wiley.

Woodward, W.A., Parr, W.C., Schucany, W.R. and Lindsey, H. (1984). A comparison of minimum distance and maximum likelihood estimation of a mixture proportion. J. Amer. Statist. Assoc., 79, 590-598.

Wu, Z. (1985). Multidimensional state-space model Kalman filtering with applications to image restoration. IEEE Trans. Acoust. Speech Signal Proc.,

ASSP-33, 1576-1592.

Yasumoto, Y. and Medioni, G. (1986). Robust estimation of three-dimensional motion parameters from a sequence of image frames using regularization. IEEE Trans. Pattern Anal. Mach. Intell., 8, 464-471.

SUMMARY

In this paper we discuss statistical methods for the analysis of a sequence of digital images. Such a sequence usually represents the development of a true scene through time, and the observed data are degraded, for example by blur or noise. We concentrate on recursive procedures, that might be applied in real time. Specific examples include models based on Gibbs distributions or the Kalman filter. In addition to restoring the image, we suggest methods for recursively estimating the parameters in both the models for the sequence of true scenes and the degradation process.

RESUMÉ

Dans cette communication on discute les méthodes statistiques pour l'analyse d'une séquence d'images numériques. Ordinairement une telle séquence représente le développement d'une scène vraie dans le temps et les données observées sont dégradées, par example par le flou on par le bruit. On se concentre ice sur les procédures récursives que l'on peut appliquer en temps réel. Des examples spécifiques renferment des modèles fondées sur les distributions de Gibbs ou sur le filtre de Kalman. Au dessus de restaurer l'image, on suggeste des méthodes pour l'estimation recursive des parametres de la modèle pour la séquence des scènes vraies et du processus de dégradation.