A Step Towards Efficient Bayesian Signal Reconstruction

J.W. Dickson IBM UK Scientific Centre, Winchester, England.

Abstract

This paper presents a theoretical basis for a set of optimal filters for the reconstruction of piecewise-continuous one-dimensional signals, drawing from Bayesian networks and Kalman filters. Results are presented for synthetic and real data, using both the optimal filters and a sub-optimal implementation. The results compare well with linear space invariant filtering or facet fitting approaches, and present a basis for the design of image restoration algorithms.

1 Introduction

This paper addresses the problem of reconstructing a piecewise continuous signal which has been corrupted by noise. This is poorly handled by linear space invariant (LSI) filters as the signal spectrum around a discontinuity is different from the spectrum of a continuous portion of signal. This has led researchers to explore the use of non-linear or adaptive filters.

The approach taken is to derive a particular class of optimal adaptive filters for one-dimensional signals, based on the theory of Bayesian networks [1] and Kalman filters [2, 3]. The signal models on which the optimal filters are based can correspond either to facet-fitting models or deformable membrane models, depending on the choice of parameters. It is recognised that these optimal filters will not be tractable for the ultimate goal of image restoration, so we also investigate the use of a sub-optimal algorithm. We apply optimal and sub-optimal algorithms, using different signal models, to both synthetic and real one-dimensional data; the results in several cases show improvements over traditional facet-fitting or LSI filtering techniques.

This approach brings the power of Bayesian theory to a stage where that theory can be applied to some more concrete problems in an efficient manner. That was the aim in carrying out this work and we see it as one of the important contributions of this paper.

2 Bayes' Rule and Sensors

The purpose of sensing is to combine information from observations with *prior* knowledge of the state of some external object in order to obtain more accurate *posterior* knowledge. We denote the state we wish to determine by a variable h and the observation set by o; the prior knowledge is written probabilistically as a distribution p(h) and the posterior knowledge as a conditional distribution p(h|o). Furthermore we describe the sensor in probabilistic terms by the BMVC 1992 doi:10.5244/C.6.20

conditional distribution p(o|h) and the prior knowledge of the observations by p(o); we can then apply Bayes' rule [4]:

$$p(h|o) = \frac{p(h)p(o|h)}{p(o)} \tag{1}$$

In many practical applications we may only wish to find the most likely value of h for a particular observation set o; in these cases we may ignore the (constant) denominator and merely maximise the numerator on the RHS of equation (1). This is the approach taken in this paper.

In order to make the sensor model tractable it is useful to be able to separate the individual members o_i of the observation set o. This can be done if the individual observations can be said to be *conditionally independent*, conditioned upon h [1]. We can then write:

$$p(o_1 \dots o_n | h) = \prod_{i=1}^n p(o_i | h)$$
 (2)

A particularly simple and common example of this is where the o_i are a sequence of measurements of a constant value h; typically the arithmetic average of these measurements would be taken as the most likely estimate.

3 Recursive Probabilistic Signal Models

In order to make maximum use of the conditional independence assumption, this paper develops recursive probabilistic models for discrete 1D signals. The assumptions that underlie these models are:

- 1. signal values are corrupted independently by the sensor;
- the original, uncorrupted values depend only on their direct neighbours' values (Markov assumption).

If we interpret these assumptions in probabilistic terms, then for each discrete value h_i of the uncorrupted signal, we have a single observation o_i and the following distribution functions:

- a prior $p(h_i)$;
- a sensor noise model $p(o_i|h_i)$;
- and a markovian signal model $p(h_{i-1}|h_i)$, $p(h_{i+1}|h_i)$.

The datasets o_i , $o_{j < i}$ and $o_{k > i}$ are assumed to be conditionally independent, conditioned on h_i , so that:

$$p(o|h_i) = p(o_i|h_i)p(o_{j < i}|h_i)p(o_{k > i}|h_i)$$
(3)

where each of the last two terms can be expanded recursively as:

$$p(o_{j < i}|h_i) = \oint_{h_{i-1}} p(o_{j < i-1}|h_{i-1}) p(o_{i-1}|h_{i-1}) p(h_{i-1}|h_i) dh_{i-1}$$
(4)

Here the first term in the integral is the same expression as the LHS, evaluated at step i - 1, the second term is the sensor model at step i - 1; these two are combined and then the results carried forward to step i via the third term in the integral.



Figure 1: Kalman filter block diagram; \mathbf{u} is the system noise, \mathbf{w} is the observation noise.

4 The Kalman Filter

It has been shown ([5], chapter 3) that the Kalman filter [2, 3] is a special case of the class of filter formulated in the previous section, where all the distribution functions are assumed to be Gaussian and the signal can be described by linear state-space equations. Here we will show where the equivalence lies without making a formal proof.

In its usual formulation the discrete Kalman filter step consists of two phases. In the *prediction* phase data up to step i - 1, in the form of a state estimate vector $\dot{\mathbf{x}}(i-1|i-1)$ and a mean squared error (MSE) estimate matrix $\mathbf{P}(i-1|i-1)$, are used to obtain estimates of the state and MSE at the next step i from a state transition matrix $\mathbf{F}(i)$ and a system noise matrix $\mathbf{U}(i)$; the new state estimate is used to obtain an estimate of the corresponding observation $\dot{\mathbf{z}}(i|i-1)$ via the measurement matrix $\mathbf{H}(i)$.

$$\mathbf{\hat{x}}(i|i-1) = \mathbf{F}(i)\mathbf{x}(i-1|i-1)$$
(5)

$$\mathbf{P}(i|i-1) = \mathbf{F}(i)\mathbf{P}(i-1|i-1)\mathbf{F}^{T}(i) + \mathbf{U}(i)$$
(6)

$$\mathbf{z}(i|i-1) = \mathbf{H}(i)\mathbf{x}(i|i-1) \tag{7}$$

In the *update* phase, the new measurement vector $\mathbf{z}(i)$, along with its expected MSE matrix $\mathbf{R}(i)$, are combined according to filter gain $\mathbf{W}(i)$ with the earlier data to give new estimates of system state and MSE:

$$\mathbf{P}^{-1}(i|i) = \mathbf{P}^{-1}(i|i-1) + \mathbf{H}^{T}(i)\mathbf{R}^{-1}(i)\mathbf{H}(i)$$
(8)

$$\mathbf{W}(i) = \mathbf{P}(i|i)\mathbf{H}^{T}(i)\mathbf{R}^{-1}(i)$$
(9)

$$\hat{\mathbf{x}}(i|i) = \hat{\mathbf{x}}(i|i-1) + \mathbf{W}(i) \left[\mathbf{z}(i) - \hat{\mathbf{z}}(i|i-1)\right]$$
(10)

These equations are summarised in the diagram in figure 1, showing the models of the system and sensor.

The equivalence between this formulation and the probabilistic one from equation (4) is as follows:

• observations o_i and $\mathbf{z}(i)$ are equivalent, as are state values h_i and $\mathbf{x}(i)$;

- the expression $p(o_{j \le i}|h_i)$ is a Gaussian distribution in h_i with mean $\hat{\mathbf{x}}(i|i-1)$ and covariance matrix $\mathbf{P}(i|i-1)$;
- $p(o_i|h_i)$ has mean $\mathbf{H}(i)\mathbf{x}(i)$ and covariance matrix $\mathbf{R}(i)$;
- and finally $p(h_{i-1}|h_i)$ has mean $\mathbf{F}(i)\mathbf{x}(i-1)$ and covariance $\mathbf{U}(i)$.

The prediction phase is the evaluation of $p(o_{j < i}|h_i)$ from $p(o_{j < i}|h_{i-1})$, while the update phase is the combination of $p(o_{j < i}|h_i)$ with $p(o_i|h_i)$ to obtain $p(o_{j < i+1}|h_i)$.

In order to apply this model to the reconstruction of continuous signals, where we can use later observations $o_{k>i}$ as well as past observations $o_{j<i}$, we add a backward "prediction" phase from step i + 1. This corresponds to $p(o_{k>i}|h_i)$, also evaluated recursively, thus giving all three terms from equation (3).

5 Piecewise Continuous Signals

In order to accommodate piecewise continuous signals we need to extend the model. Preserving the earlier assumptions, we introduce discontinuities by modifying the state transition relationship:

$$p(h_{i-1}|h_i) = \frac{c_i}{c_i+1} p_c(h_{i-1}|h_i) + \frac{1}{c_i+1} p_d(h_{i-1}|h_i)$$
(11)

where p_c describes the distribution resulting from a continuous step as before, p_d describes the distribution from a discontinuous step and c_i denotes the odds for the particular step being continuous.

To integrate this into the Kalman filter approach, both p_c and p_d must be Gaussian distributions. In the implementations a discontinuity is simulated by using a system noise matrix \mathbf{U}_d with large covariance values; thus equation (11) corresponds to splitting a single filter into two concurrent filters at each step. As a result we need 2^{n-1} filters for *n* observation points; the problem is no longer tractable. Two different heuristics to overcome this problem are discussed in a later section.

The other consequence of equation (11) is that each of the filters needs to be ranked by its posterior probability. This too can be done recursively, so that at each step *i* the distributions $p(o_{j < i}|h_i)$ and $p(o_{k>i}|h_i)$ are represented by a weighted sum of Gaussian distributions. The weights are calculated in the update phase of the Kalman filter, where the Gaussian $p(o_i|h_i)$ is combined with the incoming predictions. This combination corresponds to the multiplication of the distributions, which in turn corresponds to the multiplication of two Gaussian functions (as in equation (4)). The product of two Gaussian functions with means μ_1 , μ_2 and covariance matrices C_1 and C_2 respectively is:

$$\mathcal{N}_{\boldsymbol{\mu}_{1},\mathbf{C}_{1}}(\mathbf{x}) \times \mathcal{N}_{\boldsymbol{\mu}_{2},\mathbf{C}_{2}}(\mathbf{x}) = \mathcal{N}_{\boldsymbol{\mu}',\mathbf{C}'}(\mathbf{x}) \times \mathcal{N}_{0,\mathbf{C}''}(\boldsymbol{\mu}_{1}-\boldsymbol{\mu}_{2})$$
(12)

where

$$\begin{aligned} \mathbf{C}'^{-1} &= \mathbf{C}_1^{-1} + \mathbf{C}_2^{-1} \\ \boldsymbol{\mu}' &= \mathbf{C}' \left(\mathbf{C}_1^{-1} \boldsymbol{\mu}_1 + \mathbf{C}_2^{-1} \boldsymbol{\mu}_2 \right) \\ \mathbf{C}'' &= \mathbf{C}_1 \left(\mathbf{C}_1^{-1} + \mathbf{C}_2^{-1} \right) \mathbf{C}_2 \end{aligned}$$

In this case,

$$\mathcal{N}_{\boldsymbol{\mu}_1, \mathbf{C}_1}(h_i) = p(o_{j < i} | h_i) \text{ and } \mathcal{N}_{\boldsymbol{\mu}_2, \mathbf{C}_2}(h_i) = p(o_i | h_i)$$

As may be seen, the first term on the RHS of equation (12) is of the same form as the terms on the LHS; it is a normal distribution whose parameters are given by the Kalman filter equations ((5) to (10)). The second term on the RHS of equation (12) is a Normal distribution in the difference of the means of the two terms on the LHS; the value of this term is used to rank the individual Gaussian functions that make up $p(o_{j \le i}|h_i)$ and $p(o_{k \ge i}|h_i)$.

Thus the posterior likelihood assigned to each individual Kalman filter is a combination of the prior likelihoods c_i for each step being continuous and the results from equation (12), which is a hypothesis test between $p(o_{j \le i}|h_i)$ and $p(o_i|h_i)$.

We can also make use of the hypothesis test to determine the probability of discontinuities. We denote the state of a link from step i - 1 to step i by the symbol l_i , taking the value 0 for a discontinuity and the value 1 for a continuous step. Then we can evaluate:

$$p(o|l_i) = \oint_{h_i} p(o_{j < i}|h_i, l_i) p(o_i|h_i) p(o_{k > i}|h_i) dh_i$$
(13)

which, combined with the prior odds c_i , allows us to calculate the posterior odds for l_i .

6 Algorithms for Reconstruction

The results in this paper were obtained from two different algorithms:

- 1. a near-optimal algorithm which carries forward all the ranked hypotheses as per equation (11), pruning only those whose likelihood is less than 10^{-3} times the most likely;
- a sub-optimal algorithm which locally labels each step as continuous or discontinuous and only carries forward the currently-labelled hypothesis.

The second algorithm is a decentralised version of the Highest Confidence First (HCF) algorithm described in [6]. The HCF algorithm makes a single, globally most promising decision at each iteration. The algorithm used here, which could be called Local Highest Confidence First (LHCF) [7], finds the most promising decision over each neighbourhood and takes each of these locally maximal decisions at each step. It might be expected that the performance of the LHCF approach will deteriorate in cases where the decisions made locally have a more global effect.

Four different signal models were used:

- piecewise constant: $\mathbf{F}(i) = [1]$, $\mathbf{U}_c(i) = [0]$, $\mathbf{U}_d(i) = [\sigma_d^2]$
- piecewise constant with system noise (piecewise almost constant): $\mathbf{F}(i) = [1]$, $\mathbf{U}_c(i) = [\sigma_c^2]$, $\mathbf{U}_d(i) = [\sigma_d^2]$



Figure 2: Step size 1.5 in noise $\sigma = 0.707$; (a) piecewise constant, near-optimal algorithm; (b) piecewise constant, LHCF algorithm; (c) piecewise almost constant, near-optimal algorithm, (d) piecewise almost constant, LHCF algorithm

- piecewise linear: $\mathbf{F}(i) = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \mathbf{U}_{c}(i) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \mathbf{U}_{d}(i) = \begin{bmatrix} \sigma_{1d}^{2} & 0 \\ 0 & \sigma_{2d}^{2} \end{bmatrix}$ • piecewise almost linear: $\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \sigma_{1d}^{2} & 0 \\ 0 & \sigma_{2d}^{2} \end{bmatrix}$
- $\mathbf{F}(i) = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \mathbf{U}_c(i) = \begin{bmatrix} 0 & 0 \\ 0 & \sigma_{2c}^2 \end{bmatrix}, \mathbf{U}_d(i) = \begin{bmatrix} \sigma_{1d}^2 & 0 \\ 0 & \sigma_{2d}^2 \end{bmatrix}$

7 Experimental Results

We have prepared synthetic data to illustrate the performance of the algorithms and the models with well-controlled inputs. In each case, piecewise-constant or piecewise-linear data was corrupted by additive Gaussian noise and the results processed by the algorithms.

In order to obtain a more realistic assessment of the performance of this approach, we generated 1D datasets from slices through real images. In these cases the actual parameters of the image and noise are not known exactly and have been estimated from plots of the data.

Figure 2 shows the effect of applying the two order zero models to a signal containing a step in noise. The corrupted signal is shown by the dashed line while the reconstructed signal is shown by the solid line. The near-optimal algorithm shows reasonable competence with both models, while the LHCF



Figure 3: Roof function in noise; (a) piecewise linear, near-optimal algorithm; (b) piecewise linear, LHCF algorithm; (c) piecewise almost linear, near-optimal algorithm, (d) piecewise almost linear, LHCF algorithm

algorithm only performs well with the less rigid signal model; this is as we might expect, since the rigid signal model makes local decisions extremely error-prone. The near-optimal algorithm, using the more rigid model, was able to detect a step size of 1.0 in the same noise.

Figure 3 shows the effect of applying the two first-order models to a roof signal in noise. Once again the near-optimal algorithm shows reasonable competence for both models. The LHCF algorithm appears quite competent with the rigid model, while with the flexible model it fails to detect the discontinuity; however, its competence with the rigid model is likely to derive from the fact that there is only one discontinuity in the dataset.

Finally, figure 4 shows the result of applying the two flexible models to a slice through a real image. As expected the near-optimal algorithm shows better competence than the LHCF one, though the differences are minimal, particularly with the lower pair of plots. The data clearly would not fit either of the rigid models shown in the earlier results. Note also that the discontinuities in the data are not sharp, as per the signal model, but blurred by some form of point spread function.

The results shown here should be compared with either facet-fitting or LSI filtering techniques.

From the results of applying the rigid (facet-like) models to the simulated data, we can see some reasons for the problems encountered with the use of parametric patches to segment images. While the near-optimal algorithm is



Figure 4: Real data; (a) piecewise almost constant, near-optimal algorithm; (b) piecewise almost constant, LHCF algorithm; (c) piecewise almost linear, near-optimal algorithm, (d) piecewise almost linear, LHCF algorithm

successful in fitting such patches, the sub-optimal algorithm exhibits the fragmentation which is often observed. This is likely to be an unavoidable problem with such models because of the strong dependencies between non-local segmentation decisions. In cases where the optimal approach is not desirable or tractable, sophisticated algorithms will be required to obtain reasonable results.

The flexible models used here provide the closest comparisons with LSI filtering techniques; in fact where there are no edges the sub-optimal filter is equivalent to optimal LSI filtering. Figure 5 shows the application of a standard LSI filter and the corresponding LHCF filter to a step in noise. The use of LSI filters for piecewise continuous data is inevitably a compromise, as outlined in the introduction. Thus while these filters may provide either sharp boundaries or smooth surfaces in between, they cannot provide both simultaneously. The filtering technique developed here is capable of providing them both; where the underlying signal model is sufficiently flexible it can provide these relatively efficiently.

8 Further Work

The task of reconstructing or segmenting images using Bayesian theory has been tackled by several researchers in the past [8, 9, 10]. In many cases the results have proved to be unwieldy.

It is clear that the recursive algorithms and models presented in this paper



Figure 5: Comparison between LSI and LHCF approaches: Step size 1.0 in noise $\sigma = 0.707$; (a) piecewise almost constant, LSI filter; (b) piecewise almost constant, LHCF algorithm

cannot be directly generalised to the processing of a 2D dataset; the conditional independence assumptions required cannot be justified. However we can show that the recursive models derived here can be expressed equivalently as differential models, which can then be applied to 2D datasets. Though this in itself is not a novel idea, we expect the fusion of the two approaches to provide the inspiration and justification for a more rigorous application of differential models to image restoration and segmentation.

We are currently applying the theory described in this paper to the completion of edge maps, as produced by an edge detector (eg [11]). These edge maps often have gaps where the grey-level contrast is insufficient or where the grey-level surface does not match the edge detector's expectations (for example at junctions). The aim is to use the Kalman filtering technique to estimate the orientation at a terminator so the edges can be extrapolated, and to use the hypothesis testing techniques to determine the best completion for each terminator. This requires the application of some sophisticated search control techniques and decision-theoretical methods. Only preliminary results are available at this time, but these appear encouraging.

9 Conclusions

We have shown how Bayesian theory can be used to design recursive filters for one-dimensional signals. For the special case of Gaussian distribution functions and linear state-space signal models we have shown efficient implementations of this Bayesian theory in terms of Kalman filters, and successfully applied the results to the reconstruction of signals corrupted by additive noise.

A requirement for a more general application of this work is the viability of sub-optimal algorithms; we have compared the performance of one such algorithm with the performance of the optimal filter, showing that in many cases the results are comparable.

The approach described in this paper is currently being applied to a range of problems. Some difficulties still remain:

• the choice of suitable sub-optimal algorithms for both 1D and 2D appli-

cations;

- at present the approach does not cater for images which have been blurred as well as corrupted with noise; it is not clear how a blurring model (point spread function) could be integrated with the recursive approach;
- we have not investigated calibration methods; clearly we will need a more rigorous approach to choosing parameters than the current "eyeballing" techniques.

References

- Judea Pearl. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann Publishers, Inc., 1988.
- [2] Y. Bar-Shalom and T.E. Fortmann. Tracking and Data Association. Academic Press, 1988.
- [3] Arthur Gelb. Applied Optimal Estimation. MIT Press, 1974.
- [4] T. Bayes. An essay towards solving a problem in the doctrine of chances. *Phil. Trans.*, 3:370-418, 1763.
- [5] J. W. Dickson. Image Structure and Model-Based Vision. PhD thesis, Oxford University, Department of Engineering Science, 1990.
- [6] Paul B. Chou and Christopher M. Brown. The Theory and Practice of Bayesian Image Labeling, September 1988.
- [7] M.J. Swain and L.E. Wixson. Efficient estimation for Markov random fields. In Image Understanding and Machine Vision, 1989.
- [8] B.F. Buxton, H. Buxton, and A. Kashko. Optimization, Regularization and Simulated Annealing in Low-Level Computer Vision. In Ian Page, editor, Parallel Architectures and Computer Vision Workshop. OUP, 1987.
- [9] S. Geeman and D. Geeman. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis Machine Intell.*, 6:721-741, 1984.
- [10] J.F. Silverman and D.B. Cooper. Bayesian Clustering for Unsupervised Estimation of Surface and Texture Models. *IEEE Trans. Pattern Analysis Machine Intell.*, 10(4), July 1988.
- [11] J.F. Canny. Finding Edges and Lines. Technical Report 720, Massachusetts Inst. Technol., 1983.
- [12] A. Papoulis. Probability, Random Variables, and Stochastic Processes. McGraw-Hill, 1984.