Bayes Factors for Edge Detection from Wavelet Product Spaces

F. Murtagh (1) and J.L. Starck (2)

(1) School of Computer Science, Queen's University Belfast, Belfast BT7 1NN

(2) DAPNIA/SEI-SAP, CEA-Saclay, F-91191 Gif-sur-Yvette Cedex, France

f.murtagh@qub.ac.uk, jstarck@cea.fr

ABSTRACT

Inter-band wavelet correlation provides one approach to defining edges in an image. Inter-band wavelet products follow long-tailed density distributions, and in such a context thresholding is very difficult. We show how segmentation using a Markov field spatial dependency model is a more appropriate approach to demarcating edge and non-edge regions. A key part of this work is quantitative assessment of goodness of edge versus non-edge fit. We introduce a formal assessment framework based on Bayes factors. A detailed example is used to illustrate these results.

Subject terms: wavelet transform; edge detection; segmentation; Bayes factor; BIC, Bayes information criterion; PLIC, pseudo-likelihood information criterion; likelihood; heavy-tailed distribution.

1. INTRODUCTION

Wavelet transforms of images provide localized detail signal, which is in general related to edge information. Hence taking the product of wavelet resolution scales can help to emphasize edge information. In practice a symmetric wavelet function is best for this purpose, and a redundant wavelet transform algorithm avoids aliasing difficulties.

Inter-band wavelet correlation provides one approach to defining edges in an image. Inter-band wavelet products follow long-tailed distributions, and therefore single or multiple thresholding is very difficult to achieve.

In section 2, we review previous work using wavelet transforms for edge-finding. This includes taking products of detail signal, a practice which goes at least as far back as 1970.

Section 3 overviews longtailed distributions, and indicates the problem of image thresholding – alternative viewed as scalar quantization – which we are addressing.

Section 4 introduces and discusses (i) Gaussian mixture modeling of one-dimensional data distributions, (ii) Gaussian modeling in the case of a Markov spatial dependency model, and (iii) figures of merit, or goodness of fit, in both of these contexts. We use a Bayes factor goodness of fit assessment approach, leading to approximations termed the Bayes information criterion, BIC, and the pseudo-likelihood information criterion, PLIC, in the 1D and spatial cases, respectively.

Section 5 presents experimental results illustrating how this methodology works in practice.

2. WAVELET SCALE EVOLUTION AND WAVELET SCALE CORRELATION

Malfait et al.[18] distinguish implicitly between wavelet coefficient evolution and wavelet coefficient correlation, in both cases over the sequence of resolution scales. The former includes wavelet filtering through hard and soft thresholding. More generally, this is characterized by Malfait et al. as starting with a measure of local regularity, and then dividing wavelet coefficients into those that are sufficiently "clean" according to the regularity criterion, and those that are "noisy".

An example of wavelet scale evolution is the modulus maxima approach of Mallat and Zhong,[21] who show that use of local maxima of a wavelet transform is equivalent to the Canny edge detector. Another example of wavelet scale evolution is the zero-crossings approach of Mallat,[19] which is related to the Laplacian of Gaussian (LoG) operator originally proposed by Marr and Hildreth.

The practice of taking multiscale pointwise products for determining edges goes at least as far back as Rosenfeld.[29]

For white Gaussian noise, the average number of local maxima at scale 2^{j+1} is half the number at scale 2^j. Hence increasing scales tend to smooth out noise. (See Sadler and Swami,[30] and Mallat and Hwang[20]).

The product of wavelet scales provides edge information. Chen and Tao[9] study this with decimation. Using a redundant wavelet transform avoids problems of feature aliasing and also leads to straightforward implementation. Problems of the edge's spatial resolution scale and shift in location from one resolution scale to the next are reviewed in Xu et al.[40] and Lee and Kozaitis.[17]

A variation on the theme of edge finding using wavelet scale products is used in Olivo-Marin.[25] His objective is to find peaks in molecular biology images.

3. HEAVY TAILED DENSITIES OF WAVELET PRODUCTS

In this section we find that wavelet products are heavy-tailed, and that there is no fully satisfactory way to quantize this (in order to define edges, for example).

Heavy tailed probability distributions, examples of which include long memory or $1/f$ processes (appropriate for financial time series, telecommunications traffic flows, etc.) can be modeled as a generalized Gaussian distribution (GGD, also known as power exponential, α-Gaussian distribution, or generalized Laplacian distribution):

$$f(x) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp -(\mid x \mid /\alpha)^{\beta}$$

where

– scale parameter, α, represents the standard deviation,

– the gamma function, $\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$, and

– shape parameter, β, is the rate of exponential decay, $\beta > 0$.

A value of $\beta = 2$ gives us a Gaussian distribution. A value of $\beta = 1$ gives a double exponential or Laplace distribution. For $0 < \beta < 2$, the distribution is heavy tailed. For $\beta > 2$, the distribution is light tailed.

Sadler and Swami[30] show that (i) multiscale products generally reduce correlation in the noise, and that (ii) they are heavy tailed distributions. Sadler and Swami develop a closed form PDF (probability density function) for the product $z = y_1 y_2$ where y_1 and y_2 are both zero-mean Gaussian. Correlation coefficients are tabulated for products at scales 1 through 7, with use made of the Mallat-Zhong wavelet transform. PDFs are given for products of scales 1 and 2, and of scales 1, 2 and 3, in the case of white Gaussian input, and heavy-tailed behavior is exemplified. With iid unit-variance Gaussian and iid unit-variance Laplace input, pronounced skewness for an even number of products is shown. An even number of products account for absolute gradient, whereas uneven scale products are bipolar.

Gaussian mixture modeling of heavy tailed noise distributions is feasible: a solution is provided by a weighted sum of Gaussian densities often with decreasing weights corresponding to increasing variances. Mixing proportions for small (tight) variance components are large (e.g., 0.15 to 0.3) whereas very large variance components have small mixing proportions. A signal detection test is proposed[30] based on absolute values of a 3-scale product exceeding a threshold. Location estimation of step change is also carried out, based on a Cramer-Rao bound.

Heavy tailed noise can be modeled by a Gaussian mixture model with enough terms (Blum et al.[3]). Similarly, in speech and audio processing, low-probability and large-valued noise events can be modeled as Gaussian components in the tail of the distribution. A fit of this fat tail distribution by a Gaussian mixture model is commonly carried out (Wang and Zhao[39]). As in Wang and Zhao, one can allow Gaussian component PDFs to recombine to provide the clusters which are sought. These authors also found that using priors with heavy tails, rather than using standard Gaussian priors, gave more robust results. But the benefit appears to be very small.

Chen and Karim[8] explore wavelet correlation using the Mallat scheme, often used in the context of image compression.

Tsakalides et al.[38] use a range of tests to exemplify long-tailed densities. They present considerable evidence for wavelet coefficients themselves (in their case using a 2D Haar transform) being long-tailed. In the case of the Cauchy distribution, they derive quantization levels. Related work with a similar perspective is presented in Buccigrossi and Simoncelli[4] (using the 9/7 biorthogonal wavelet transform) and elsewhere. We will not pursue the modeling of wavelet spaces further here, since our interest is related more to the behavior of wavelet product spaces.

We conclude the following: a mixture of genuine signal and flicker or pink noise constituting a heavy tail in the density implies practical difficulty in disentangling them. It may be feasible to have Gaussian components in the heavy tail corresponding to signal, and other Gaussian components in the tail corresponding to noise. However this remains an imprecise and approximate approach.

Figs. 1 and 2 illustrate long-tailed behavior and show an aspect of the the marginal density Gaussian model fitting (to be used below: the quality of this fit is given in the upper right panel of Fig. 3, corresponding to the 3-class case).

The ordinates give frequencies. The original data values are offset so that all values are positive, and mapped to [0, 255], which explains the abscissa values. In this rescaling of values, the means and standard deviations of the three classes are as follows: means – 42, 48 and 74; standard deviations – 0.8, 6.3 and 26.0. Cardinalities are, respectively, 100442, 112728 and 48974. Our initialization algorithm is as follows: (i) construct a histogram of 256 bins; and (ii) define approximately equally-sized strictly contiguous class regions. Figs. 1, 2 and 3 show that we do not find adjacent regions of the marginal density as a solution for these classes. A method such as k-means would have provided contiguous regions.

In the next section, section 4, we describe a more formal treatment of the Gaussian model fitting carried out on image marginal densities. We also look at Gaussian model fitting which takes spatial dependency into account, in the form of a Markov random field. The non-contiguous clusters found in Figs. 1, 2 and 3, and found also in many other cases relating to similar highly concentrated densities, are not fully satisfactory. However, a more important reason for us to favor Gaussian model fitting based on a Markov model is that the measures of goodness of fit found in practice are more plausible in the latter case. This will be discussed in section 5.

4. MODEL-BASED CLUSTERING

Our basic framework is that of *model-based clustering*, as described, for example, by Fraley and Raftery.[14,13] In the most basic form of this methodology, a finite mixture of Gaussian distributions is fit to the data by maximum likelihood estimation using the EM (expectation-maximization) algorithm, the number of groups can be chosen using Bayesian model selection, and if hard clustering is desired, each observation is assigned to its most likely group a posteriori.

4.1. Univariate Finite Gaussian Mixture Models

In the univariate normal finite mixture model, one-dimensional observations x_i are assumed to be drawn from G groups, each of which is Gaussian distributed. The g-th group has mean μ_g and variance σ_g^2. Given observations $x = (x_1, \ldots, x_n)$, let γ be an unobserved $n \times G$ cluster assignment matrix, where $\gamma_{ig} = 1$ if x_i comes from the g-th group, and $\gamma_{ig} = 0$ otherwise. Our goals are to determine the number of clusters G, to determine the cluster assignment of each observation, and to estimate the parameters μ_g and σ_g of each cluster. The probability density for this model is

$$f(x_i|\theta, \lambda) = \sum_{g=1}^{G} \lambda_g f_g(x_i|\theta_g), \tag{1}$$

where $\theta_g = (\mu_g, \sigma_g^2)^T$, $f_g(\cdot|\theta_g)$ is a normal density with mean μ_g and variance σ_g^2, $\theta = (\theta_1, \ldots, \theta_G)$, and $\lambda = (\lambda_1, \ldots, \lambda_G)$ is a vector of mixture probabilities such that $\lambda_g \geq 0$ $(g = 1, \ldots, G)$ and $\sum_{g=1}^{G} \lambda_g = 1$.

We estimate the parameters by maximum likelihood using the EM (expectation-maximization) algorithm.[12,23] For its application to model-based clustering, see McLachlan and Basford,[22] Celeux and Govaert[7] and Dasgupta and Raftery.[11]

The EM algorithm iterates between the E step and the M step. In the E step, the conditional expectation, $\hat{\gamma}$, of γ given the data and the current estimates of θ and λ are computed, so that $\hat{\gamma}_{ig}$ is the conditional probability that x_i belongs to the g-th group. In the M step, conditional maximum likelihood estimators of θ and λ given the current $\hat{\gamma}$ are computed.

Although the EM algorithm has some limitations (e.g. it is not guaranteed to converge to a global rather than a local maximum of the likelihood, and it requires a starting configuration), it is generally efficient and effective for Gaussian clustering problems.

4.2. Spatial Segmentation

Model fitting to the marginal density pays no attention to two-dimensional image spatial information. We can take such information into account using a hidden Markov model. Background on the approach pursued here can be found in Stanford[33] and Stanford and Raftery.[34,35]

We use Bayesian model selection to choose the number of clusters. For a review of Bayesian model selection, see Kass and Raftery[16] and Raftery.[27] Pioneering work in this area was due to H. Jeffreys, I.J. Good and (according to the latter) A. Turing. The use of the BIC in choosing clusters in a mixture or clustering model is discussed by Roeder and Wasserman[28] and Dasgupta and Raftery.[11] Applications are in Campbell et al.,[5,6] Mukherjee et al.[24] and in other articles.

We consider an unknown, true pixel state, for pixel i, as $X_i \in \{1, 2, \ldots K\}$ for K states. The observed image pixel is Y_i. In this work this is taken as a scalar (and could be taken instead as a vector for color or multiband images). Consider an indicator function, $I(X_i, X_j) = 1$ if $X_i = X_j$ and otherwise $= 0$.

We now use a Markov random field to define spatial structure on X. We take $p(X)$ as being proportional to $\exp(\phi \sum_{i,j} I(X_i, X_j))$. This is a Potts or Ising model. ϕ is a spatial homogeneity parameter, a small value implying randomness, and a large value implying uniformity. A negative value of ϕ implies dissimilarity between neighboring pixels, and is not of interest here. Our model is a hidden Markov model (HMM) because the variables X are only known through the observed Y.

Let $N(X_i)$ be the neighborhood of X_i, e.g. 3×3 pixels. Let $U(N(X_i), k)$ be the number of neighborhood pixels with state k.

From $p(X)$ we have the conditional distribution:

$$p(X_i = j \mid N(X_i), \phi) = \frac{\exp(\phi U(N(X_i)), j)}{\sum_k \exp(\phi U(N(X_i)), j)} \tag{2}$$

Having looked at the latent space, we now return to the observed data. We assume the following conditional density model connecting the observed and hidden variables: $f(Y_i \mid X_i = j)$ is Gaussian with mean μ_j and standard deviation σ_j. In the multiband case, where y is a vector, the mean vector is used, and the variance-covariance matrix. The Y_i are conditionally independent given the X_i or, alternatively expressed, dependence among the Y_i only occurs via dependence among the X_i. Call θ_k the set of parameters, (μ, σ^2) for state k. We have $f(Y \mid X) = \Pi_i f(Y_i \mid X_i) = \Pi_i f(Y_i \mid \theta_{X_i})$.

Our solution algorithm is as follows. It is based on Besag's[2] iterated conditional modes (ICM) algorithm, which reconstructs an image based on local properties modeled as an MRF. This iterative algorithm requires an initial estimate of X, $\hat{X}$, and proceeds to estimate the parameters of $p(Y_i \mid X_i)$, as well as ϕ and X. To initialize X, we note that in taking $p(Y_i \mid X_i)$ as Gaussian, then the marginal density of Y is a finite mixture of Gaussians. In the multiband case, we typically use a marginal density model on the eigen or principal component image. The EM-based modeling of the marginal density discussed in section 4.1 then applies. An alternative approach to initialization, based on wavelet products, will be investigated in section 4.

Segmentation Algorithm:

Step 0: Initialize $\hat{X}$ using a marginal segmentation.

Step 1: Update $\hat{\theta} =$ argmax $f(Y \mid \hat{X})$ based on maximum likelihood estimates of μ_j and θ_j for each class, j.

Step 2: Update ϕ using the maximum pseudo-likelihood: $\hat{\phi} = \mathrm{argmin}_{\phi}(-\log \mathrm{PL}(\hat{X} \mid \phi))$. The pseudo-likelihood is given by $\mathrm{PL}(\hat{X} \mid \phi) = \Pi_i p(\hat{X}_i \mid N(\hat{X}_i, \phi))$.

Step 3: Update $\hat{X}$: for each pixel i, $\hat{X}_i = \mathrm{argmax}_j f(Y_i \mid X_i = j) p(X_i = j \mid N(\hat{X}_i, \hat{\phi}))$.

Implementation details: In step 2, if $\hat{\phi}$ goes negative, then we reset it to zero. In all calculations, we exclude boundary pixels from consideration. Step 1 is one step of Besag's ICM (iterated conditional modes) algorithm.

4.3. Model Selection using Bayes Factors

We now turn attention to model selection. A Bayesian assessment framework provides an objective and generally-applicable approach to classification and related decision-making. The Bayes factor, developed by Jefferys in the 1930s, is the posterior odds of one model over another when the prior probabilities of the two models are equal. We describe how approximations to the Bayes factor are used in practice. In particular, we use the Bayes information criterion or BIC, and the pseudo-likelihood information criterion or PLIC. While we employ both of these criteria with Gaussian model fitting, BIC is used in the non-spatial case, and PLIC is used in the spatial case.

A model M_K is the set of parameters estimated for a given number of mixture components, K. Consider data D. The posterior probability of model M_K is

$$p(M_K \mid D) = \frac{p(D \mid M_K)p(M_K)}{\sum_{L=1}^{K\max} p(D \mid M_L)p(M_L)}$$

We can ignore $p(M_K)$ and the influence of M_L if each model is equi-likely a priori.

The integrated likelihood, $p(D \mid M_K)$, is given by

$$p(D \mid M_K) = \int p(D \mid \theta_K, M_K)p(\theta_K)d\theta_K$$

where θ_K is the set of parameters for model M_K, $p(D \mid \theta_K, M_K)$ is the usual likelihood, and $p(\theta_K)$ is the prior.

A good approximation to the integrated likelihood is given by

$$2\log p(D \mid M_K) \approx \text{BIC (Bayes Information Criterion)}$$

$$\text{BIC} = 2\log p(D \mid \hat{\theta}_K, M_K) - N\log(\dim(\theta_K)) \tag{3}$$

where $\hat{\theta}_K$ is the maximum likelihood estimator of θ_K. N is the dimensionality of the observation vectors.

An alternative derivation of BIC as a minimum description length (MDL) criterion is described by Hansen.[15]

4.4. An Information Criterion with Spatial Interaction, PLIC

In the spatial (Markov) case, the Bayes factor assessment criterion is developed not for the homogeneity parameter, ϕ, nor for the neighborhood,[32] but rather for the number of segments, K. The likelihood (first term) in the BIC, equation (3), is problematic for computational reasons.

The posterior distribution of X conditional on Y is: $f(X \mid Y) = f(Y \mid X)f(X)/f(Y) \propto f(Y \mid X)f(X)$. Since there is conditional independence between Y and X, we have that $f(Y \mid X) = \Pi_i f(Y_i \mid X_i)$ which, it has already been noted, is taken as Gaussian.

The density of x, $f(X)$, is related to all possible states, which is combinatorially explosive. Therefore the pseudo-likelihood, $\mathrm{PL}(X)$, is taken as a proxy for $f(X)$. The pseudo-likelihood, introduced in Besag,[1] restricts where the integrated likelihood is defined. We have

$$\mathrm{PL}(X, \phi) = \Pi_i p(X_k \mid N(X_i), \phi) = \Pi_i \frac{\exp(\phi U(N(X_i)), X_i)}{\sum_k \exp(\phi U(N(X_i)), k)}$$

The likelihood is made conditional on the neighborhood of pixel i. Previously we had

$$L(Y_i \mid X_i) = \sum_j f(Y_i \mid X_i = j) p(X_i = j)$$

for state or label j.

Instead, denoting X_{-i} the neighborhood of X_i not including pixel i, and with $\hat{X}$ denoting an estimate of X, we use:

$$L(Y_i \mid N(\hat{X}_{-i})) = \sum_j f(Y_i \mid X_i = j) p(X_i = j \mid N(\hat{X}_i))$$

As already noted, the first part of the right hand side term requires evaluation of a Gaussian; and the second part uses the conditional distribution defined for $p(X)$ in equation (2).

From the product of pseudo-likelihoods for all pixels, we arrive at a modified BIC, modifying equation (3). This modified criterion is termed the pseudo-likelihood information criterion, PLIC.[33–35]

5. APPLICATIONS OF WAVELET SCALE CORRELATION TO EDGE DETECTION

We used the well-known Lena test image in view of its properties, i.e. noisy, edge regions, smooth regions. Our image was of dimensions 512 × 512 and grayscale. For the wavelet transform we used the à trous redundant wavelet transform with a B_3 spline scaling function.[31,37,36] Redundancy is important to avoid aliasing of features. In the work below, we used 5 wavelet or detail resolution scales, which together with the smooth continuum provided an additive decomposition of the image:

$$Y = S + \sum_{j=1}^{5} W_j \tag{4}$$

where Y is the image, each W_j is an image of wavelet coefficients, and S is the smooth continuum. All of Y, W_j and S are of the same image pixel dimensions, in view of the transform's redundancy. Hence the product of wavelet scales 2 and 3 is given by the pixelwise product $W_2 W_3$, and the result is again an image of dimensions (in our work, here) 512 × 512.

Fig. 3 shows results of marginal clustering with use of the BIC goodness of fit criterion, and spatial segmenting with use of the PLIC goodness of fit criterion, for the products of wavelet scales 2 and 3, 3 and 4, and 4 and 5.

As evidenced in Fig. 3, the BIC value usually increases to an approximate plateau as the numbers of classes increase. However, it is also usually the case that the quality of fit can increase indefinitely. The greater the BIC value, the better the fit. PLIC values, as seen in the figure, are more diverse in behavior. Again a larger value indicates better fit. In the lower left and middle panels of Fig. 3, corresponding to the 2 × 3 and 3 × 4 product cases, the best PLIC value corresponds to a number of mixture model components equal to 2. In the lower right panel, corresponding to the 4 × 5 product case, the best PLIC value corresponds to a number of mixture model components equal to 3.

We conclude that, here, PLIC gives a better result, for the following reasons. Our earlier results in Figs. 1 and 2 have shown that BIC may well be associated with an implausible (but notwithstanding valid) non-contiguous Gaussian fit. We also have difficulty knowing where to stop in the sequence of increasing BIC values, which is unlike the case for PLIC. PLIC in addition gives us an outcome, namely that the best number of segments is 2 or 3, which is quite reasonable given the fact that our inputs consist of wavelet product – hence edge-emphasizing – images.

The cases corresponding to these best PLIC values in the lower panels of Fig. 3 are shown, respectively, in Figs. 4, 5 and 6.

We will use the result of the 3×4 product case, i.e. Fig. 5, to proceed further to derive a reasonable edge map. Fig. 7 shows an edge map derived from Fig. 5. This was done using the difference between Fig. 5 and an eroded version of the same image, using a square 2×2 kernel. Further processing steps could be availed of, e.g. deleting connected components of small size. For comparison, Fig. 8 shows a Canny result using the original image.

6. CONCLUSION

We mention in concluding some other recent publications, which share certain aspects of our approach but which differ in other ways. Crouse et al.[10] develop a two-state (high, low wavelet coefficient value) model for wavelet correlation. Our objective is more general: as shown with Fig. 6, we cannot assume that a 2-class (edge versus non-edge) fit is always best. Pižurica et al.[26] use across-scale wavelet ratios which are less easily handled than wavelet products.

The achievements of the work reported here are as follows:

- We have demonstrated how objective selection criteria may be applied to wavelet product spaces, in order to help in finding edge regions in images.

- Practical, approximate goodness of fit criteria have been developed, based on Bayes factors.
- The objective of our approach can be characterized as a methodology for avoiding the difficulties in theory and practice which are part and parcel of image (here: wavelet product) thresholding.
- From the points of view of computational efficiency and experimental results, our approach works well. Marginal density mixture modeling takes a few seconds on a Sun SparcStation 10, and spatial segmentation takes about 10 minutes, for a specified number of classes, and using a 512×512 image.

Acknowledgements

We are grateful to anonymous referees of an earlier version of this paper for their detailed and constructive comments.

REFERENCES

1. J. Besag. Statistical analysis of non-lattice data. *Statistician*, 24:179–195, 1975.
2. J. Besag. Statistical analysis of dirty pictures. *Journal of the Royal Statistical Society, Series B*, 48:259–302, 1986.
3. R.S. Blum, Y. Zhang, B.M. Sadler, and R.J. Kozick. On the approximation of correlated non-Gaussian noise PDFs using Gaussian mixture models. Conference on the Applications of Heavy Tailed Distributions in Economics, Engineering and Statistics, 1999.
4. R.W. Buccigrossi and E.P. Simoncelli. Image compression via joint statistical characterization in the wavelet domain. *IEEE Transactions on Image Processing*, 8:1688–1701, 1999.
5. J.G. Campbell, C. Fraley, F. Murtagh, and A.E. Raftery. Linear flaw detection in woven textiles using model-based clustering. *Pattern Recognition Letters*, 18:1539–1548, 1997.
6. J.G. Campbell, C. Fraley, D. Stanford, F. Murtagh, and A.E. Raftery. Model-based methods for textile fault detection. *International Journal of Imaging Science and Technology*, 10:339–346, 1999.
7. G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28:781–793, 1995.

8. Z. Chen and M.A. Karim. Forest representation of wavelet transform and feature detection. *Optical Engineering*, 5:1194–1202, 2000.

9. Z. Chen and Y. Tao. Subband correlation of Daubechies wavelet representations. *Optical Engineering*, 40:362–371, 2001.

10. M. Crouse, R. Nowak, and R. Baraniuk. Wavelet-based statistical signal processing using hidden Markov models. *IEEE Transactions on Signal Processing*, 46:886–902, 1998.

11. A. Dasgupta and A.E. Raftery. Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association*, 93:294–302, 1998.

12. A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–22, 1977.

13. C. Fraley. Algorithms for model-based Gaussian hierarchical clustering. *SIAM Journal of Scientific Computing*, 20:270–281, 1999.

14. C. Fraley and A.E. Raftery. How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, 41:578–588, 1998.

15. M.H. Hansen and Bin Yu. Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96:746–774, 2001.

16. R.E. Kass and A.E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

17. Y. Lee and S.P. Kozaitis. Multiresolution gradient-based edge detection in noisy images using wavelet domain filters. *Optical Engineering*, 39:2405–2412, 2000.

18. M. Malfait and D. Roose. Wavelet-based image denoising using a Markov random field a priori model. *IEEE Transactions on Image Processing*, 6:549–565, 1997.

19. S. Mallat. Zero-crossings of a wavelet transform. *IEEE Transactions on Information Theory*, 37:1019–1033, 1991.

20. S. Mallat and W.L. Hwang. Singularity detection and processing with wavelets. *IEEE Transactions on Information Theory*, 38:617–643, 1992.

21. S. Mallat and S. Zhong. Characterization of signals from multiscale edges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14:710–732, 1992.

22. G. McLachlan and K. Basford. *Mixture Models.* Marcel Dekker, 1988.

23. G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions.* Wiley, 1997.

24. S. Mukherjee, E.D. Feigelson, G.J. Babu, F. Murtagh, C. Fraley, and A. Raftery. Three types of gamma-ray bursts. *The Astrophysical Journal*, 508:314–327, 1998.

25. J.C. Olivo-Marin. Extraction of spots in biological images using multiscale products. *Pattern Recognition*, 35:1989–1996, 2002.

26. A. Pižurica, W. Philips, I. Lemahieu, and M. Acheroy. A joint inter- and intrascale statistical model for Bayesian wavelet based image denoising. *IEEE Transactions on Image Processing*, 11:545–557, 2002.

27. A.E. Raftery. Bayesian model selection in social research (with discussion by Andrew Gelman, Donald B. Rubin and Robert M. Hauser). In P.V. Marsden, editor, *Sociological Methodology 1995*, pages 111–196. Blackwells, 1995.

28. K. Roeder and L. Wasserman. Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, 92:894–902, 1997.

29. A. Rosenfeld. A nonlinear edge detection technique. *Proceedings of the IEEE*, 58:814–816, 1970.

30. B.M. Sadler and A. Swami. Analysis of multiscale products for step detection and estimation. *IEEE Transactions on Information Theory*, 45:1043–1051, 1999.

31. M.J. Shensa. Discrete wavelet transforms: Wedding the à trous and Mallat algorithms. *IEEE Transactions on Signal Processing*, 40:2464–2482, 1992.

32. S. Stan, G. Palubinskas, and M. Datcu. Bayesian selection of the neighborhood order for Gauss-Markov texture models. *Pattern Recognition Letters*, 23:1229–1238, 2002.

33. D.C. Stanford. *Fast Automatic Unsupervised Image Segmentation and Curve Detection in Spatial Point Patterns.* PhD thesis, Department of Statistics, University of Washington, 1999.

34. D.C. Stanford and A.E. Raftery. A consistency result for a penalized pseudolikelihood criterion for model selection with spatially dependent mixture models. Technical report, Department of Statistics, University of Washington, 1999.

35. D.C. Stanford and A.E. Raftery. Determining the number of colors or gray levels in an image using approximate Bayes factors: the pseudolikelihood information criterion (PLIC). Technical report, Department of Statistics, University of Washington, 2001.

36. J.L. Starck and F. Murtagh. *Astronomical Image and Data Analysis.* Springer-Verlag, 2002.

37. J.L. Starck, F. Murtagh, and A. Bijaoui. *Image and Data Analysis: The Multiscale Approach.* Cambridge University Press, 1998.

38. P. Tsakalides, P. Reveliotis, and C.L. Nikias. Scalar quantisation of heavy-tailed signals. *IEE Vision, Image and Signal Processing*, 147:475–484, 2000.

39. S. Wang and Y. Zhao. Online Bayesian tree-structured transformation of HMMs with optimal model selection for speaker adaptation. *IEEE Transactions on Speech and Audio Processing*, 9:663–677, 2001.

40. Y. Xu, J.B. Weaver, D.M. Healy, and J. Liu. Wavelet transform domain filters: a spatially selective noise filtration technique. *IEEE Transactions on Image Processing*, 3:747–758, 1994.

Biographies

Fionn Murtagh holds BA and BAI degrees in mathematics and engineering science, and an MSc in computer science, all from Trinity College Dublin, Ireland, a PhD in mathematical statistics from Université P. & M. Curie, Paris 6, France, and an Habilitation from Université L. Pasteur, Strasbourg, France. Previous posts have included Senior Scientist with the Space Science Department of the European Space Agency, and visiting appointments with the European Commission's Joint Research Centre, and the Department of Statistics, University of Washington. He is Professor of Computer Science at Queen's University Belfast. He is Editor-in-Chief of The Computer Journal and is a Fellow of the British Computer Society.

Jean-Luc Starck has a PhD from University Nice-Sophia Antipolis and an Habilitation from University Paris XI. He was a visitor at the European Southern Observatory (ESO) in 1993 and at Stanford's Statistics Department in 2000. He has been a Researcher at CEA since 1994, working on the Infrared Space Observatory (ISO) project. His research interests include image processing, multiscale methods and statistical methods in astrophysics. He is also author of the books *Image Processing and Data Analysis: the Multiscale Approach* (Cambridge University Press, 1998), and *Astronomical Image and Data Analysis* (Springer-Verlag, 2002).

Figure Captions

FIGURE 1: Upper left: histogram of marginal density of product of wavelet scales 4 and 5 of a 512 × 512 Lena image. Upper right, lower left, and lower right: histograms of classes 1, 2 and 3.

FIGURE 2: Overplotting of the histograms presented in Fig. 1.

FIGURE 3: Two panels, up and down, on left: product of wavelet scales 2 and 3. Two panels, up and down, in center: product of wavelet scales 3 and 4. Two panels, up and down, on right: product of wavelet scales 4 and 5. Top panels: BIC, Bayes information criterion values, for varying numbers of classes, which is based on a fit of Gaussians to the image marginal density. Bottom panels: PLIC, pseudo-likelihood information criterion, for varying numbers of classes, which is based on a fit of Gaussians to a Markov spatial dependency model. All BIC and PLIC values are scaled by a factor of 10,000, for clarity.

FIGURE 4: Using product of wavelet scales 2 and 3, two-class solution based on a spatial model. (Cf. bottom left panel in Fig. 3, which motivates our choice of a two-class solution.)

FIGURE 5: Using product of wavelet scales 3 and 4, two-class solution based on a spatial model. (Cf. bottom middle panel in Fig. 3, which motivates our choice of a two-class solution.)

FIGURE 6: Using product of wavelet scales 4 and 5, three-class solution based on a spatial model. (Cf. bottom right panel in Fig. 3, which motivates our choice of a three-class solution.)

FIGURE 7: Edge map derived from Fig. 5 by subtracting the image from an eroded version of it: see text for details.

FIGURE 8: For comparison with Fig. 7: a Canny edge map, shown histogram-equalized, from the original Lena image.

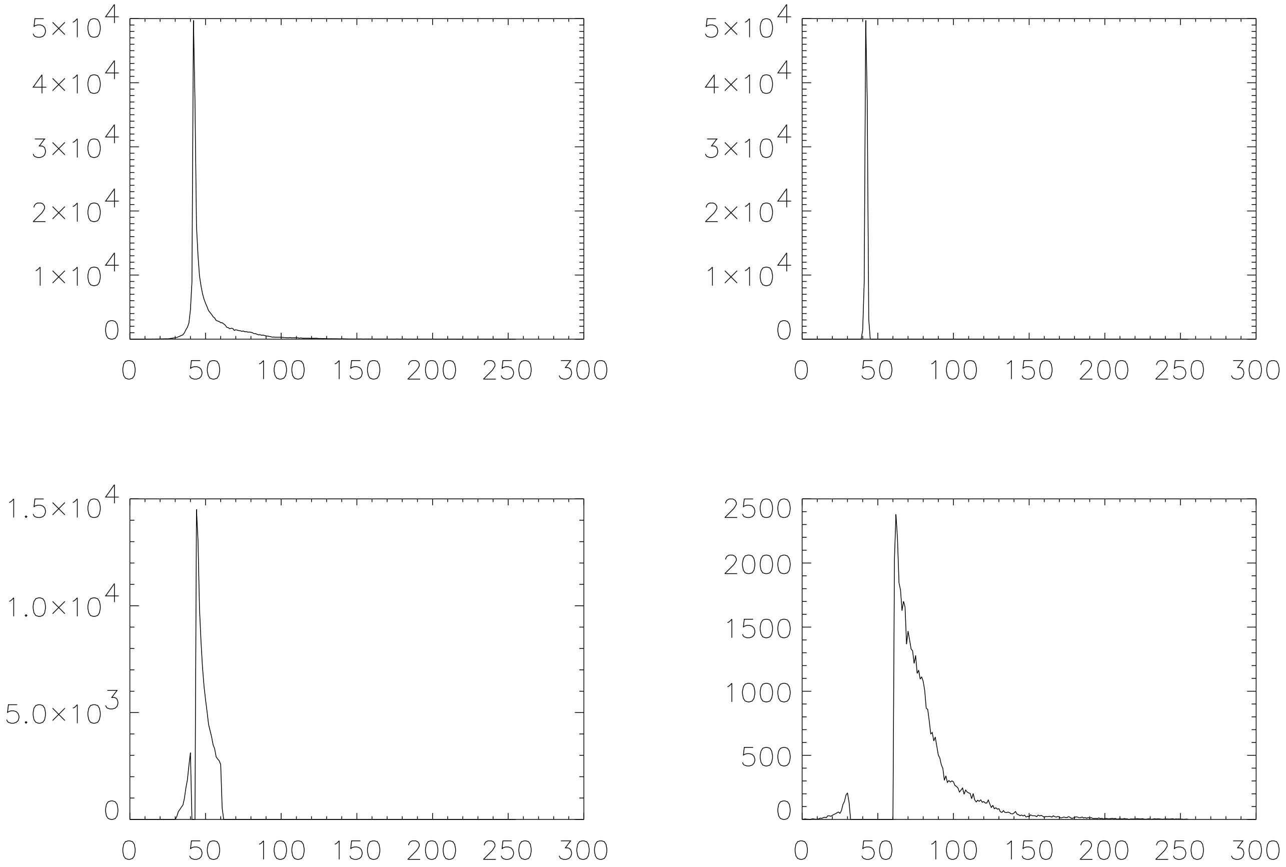

Figure 1. Upper left: histogram of marginal density of product of wavelet scales 4 and 5 of a 512 × 512 Lena image. Upper right, lower left, and lower right: histograms of classes 1, 2 and 3.

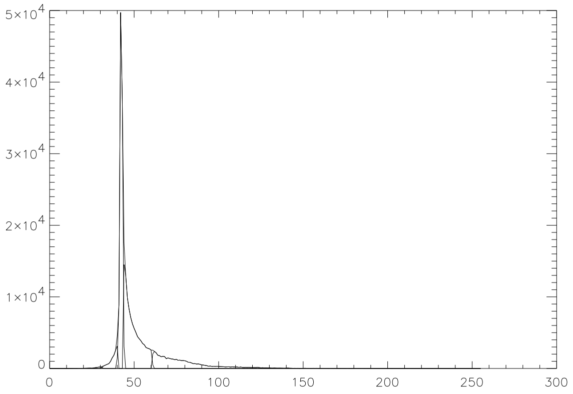

Figure 2. Overplotting of the histograms presented in Fig. 1.

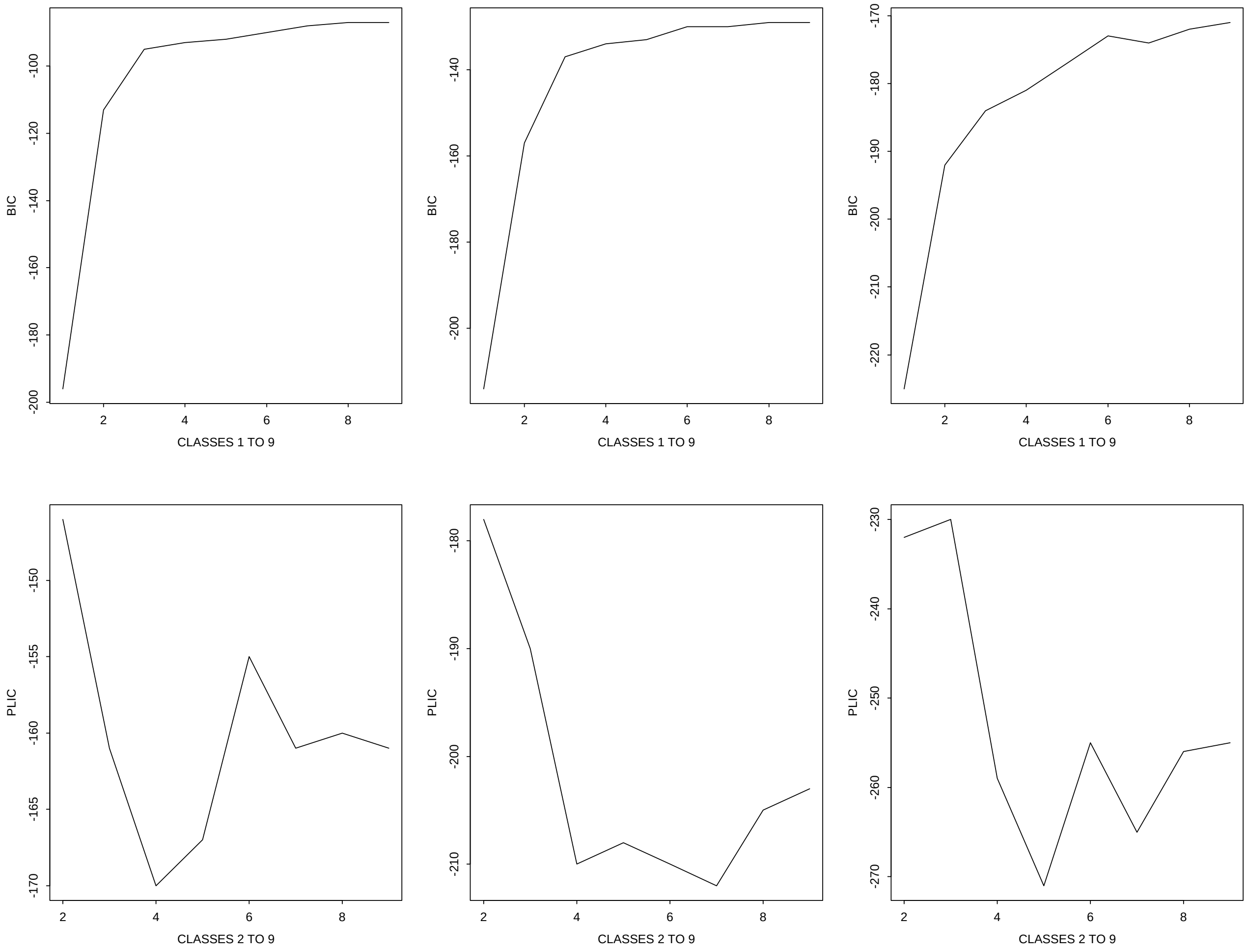

Figure 3. Two panels, up and down, on left: product of wavelet scales 2 and 3. Two panels, up and down, in center: product of wavelet scales 3 and 4. Two panels, up and down, on right: product of wavelet scales 4 and 5. Top panels: BIC, Bayes information criterion values, for varying numbers of classes, which is based on a fit of Gaussians to the image marginal density. Bottom panels: PLIC, pseudo-likelihood information criterion, for varying numbers of classes, which is based on a fit of Gaussians to a Markov spatial dependency model. All BIC and PLIC values are scaled by a factor of 10,000, for clarity.

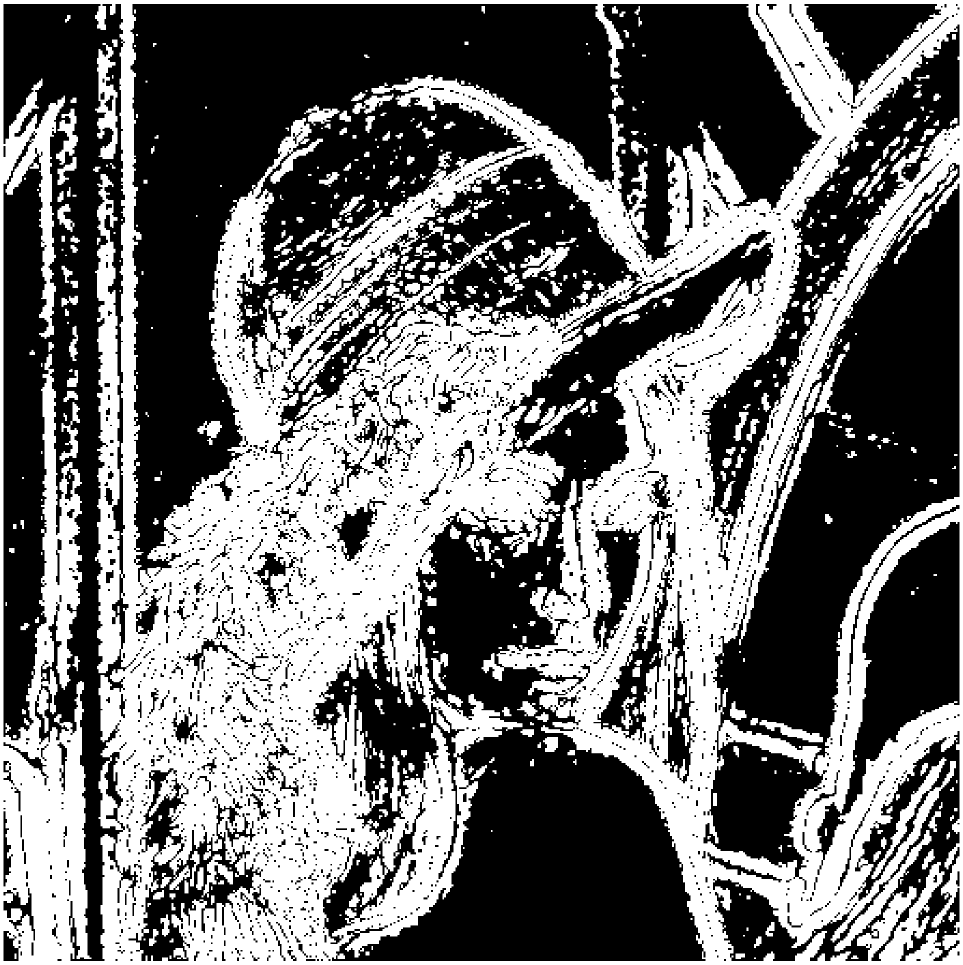

Figure 4. Using product of wavelet scales 2 and 3, two-class solution based on a spatial model. (Cf. bottom left panel in Fig. 3, which motivates our choice of a two-class solution.)

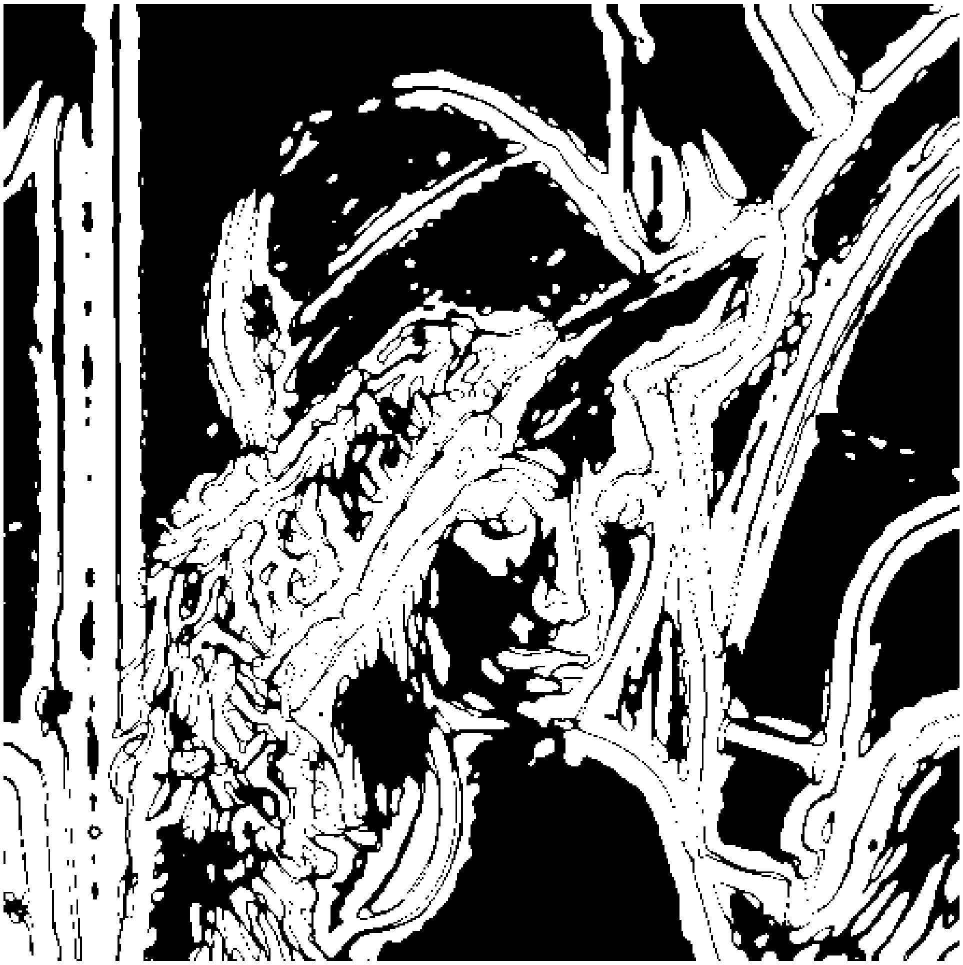

Figure 5. Using product of wavelet scales 3 and 4, two-class solution based on a spatial model. (Cf. bottom middle panel in Fig. 3, which motivates our choice of a two-class solution.)

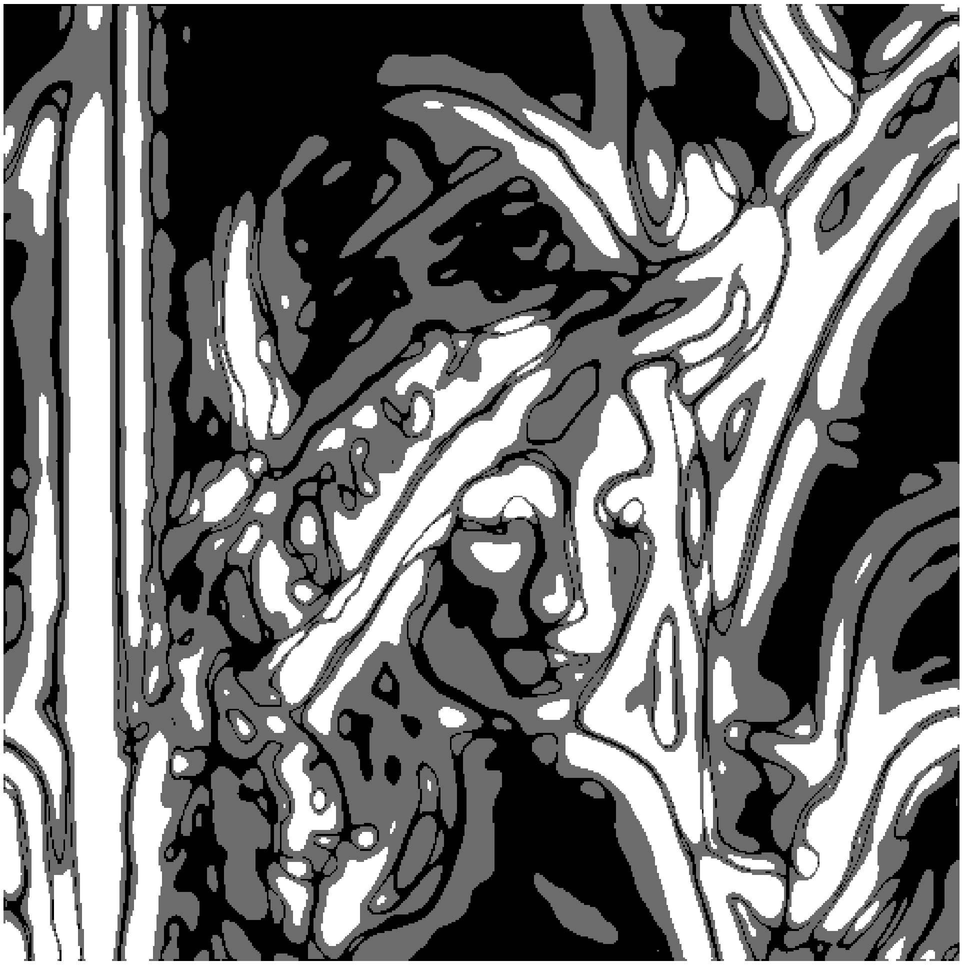

Figure 6. Using product of wavelet scales 4 and 5, three-class solution based on a spatial model. (Cf. bottom right panel in Fig. 3, which motivates our choice of a three-class solution.)

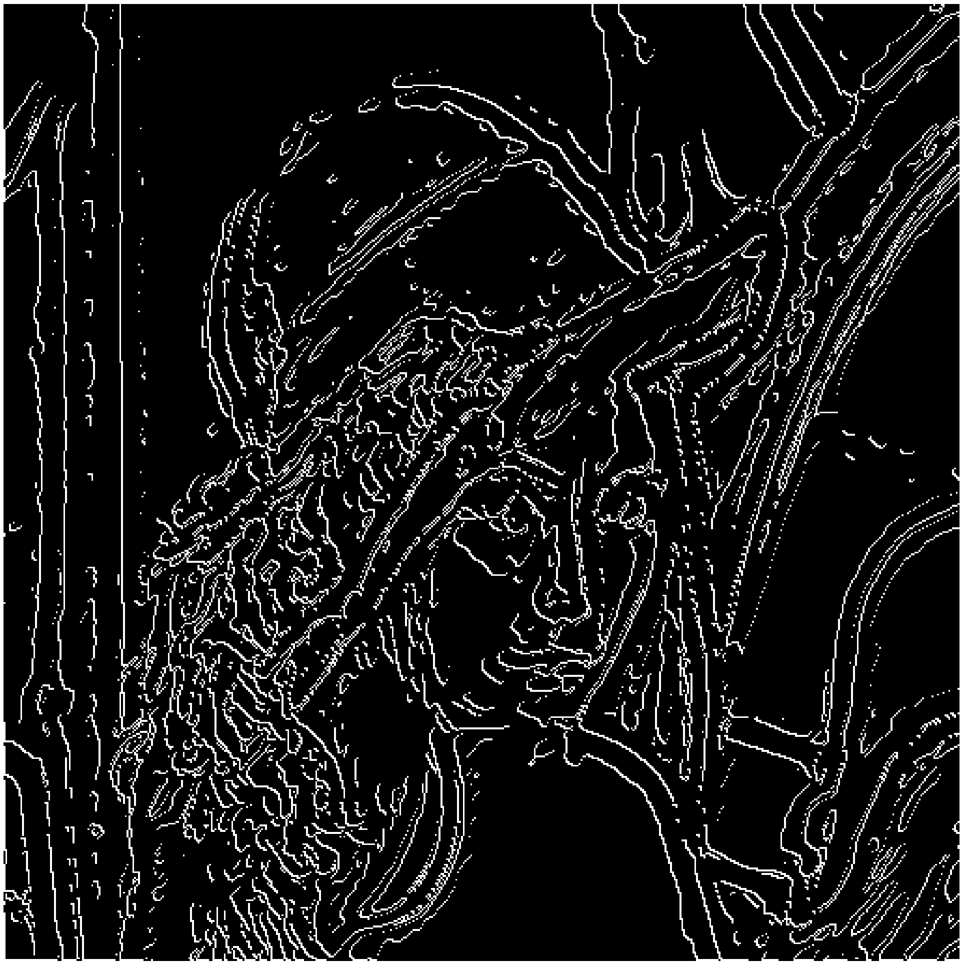

Figure 7. Edge map derived from Fig. 5 by subtracting the image from an eroded version of it: see text for details.

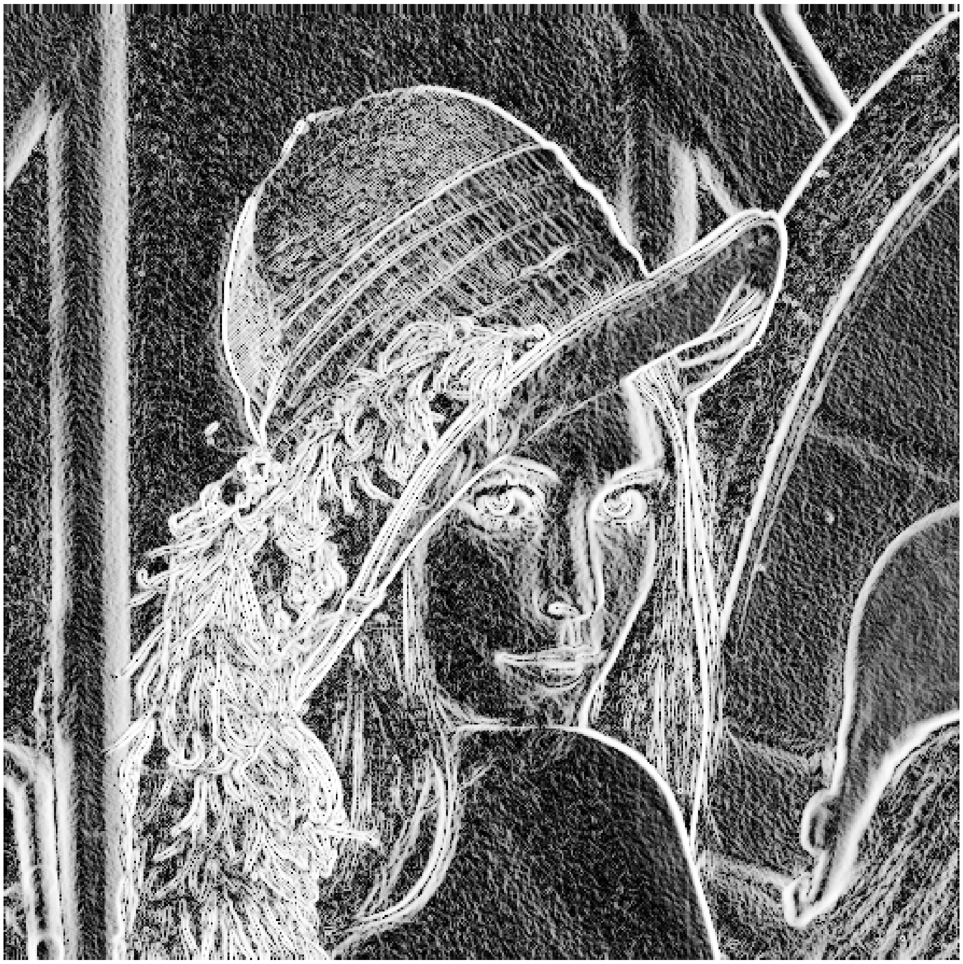

Figure 8. For comparison with Fig. 7: a Canny edge map, shown histogram-equalized, from the original Lena image.