Benchmarking Segmentation Results Using a Markov Model and a Bayes Information Criterion

Fionn Murtagh (1), Xiaoyu Qiao (1), Danny Crookes (1),
Paul Walsh (2), P.A. Muhammed Basheer (2), and Adrian Long (2)
(1) School of Computer Science, Queen's University Belfast, Belfast BT7 1NN
(2) School of Civil Engineering, Queen's University Belfast, Belfast BT7 1NN
Corresponding author contact email: f.murtagh@qub.ac.uk

ABSTRACT

Features are derived from wavelet transforms of images containing a mixture of textures. In each case, the texture mixture is segmented, based on a 10-dimensional feature vector associated with every pixel. We show that the quality of the resulting segmentations can be characterized using the Potts or Ising spatial homogeneity parameter. This measure is defined from the segmentation labels. In order to have a better measure which takes into account both the segmentation labels and the input data, we determine the likelihood of the observed data given the model, which in turn is directly related to the Bayes information criterion, BIC. Finally we discuss how BIC is used as an approximation in model assessment using a Bayes factor.

1. INTRODUCTION

A great deal of work has been carried out on automated segmentation of texture images since Cross and Jain[2] took the ideas of Besag[1] and applied them to realistic textures. Such work is based on a Markov model of spatial context. The wavelet transform has been frequently used in order to provide an embedding of the image in a multidimensional feature space. Local energy at a range of wavelet transform bands are often used.[4] The input for segmentation is therefore a multiband image. In this work we develop a new approach to the performance evaluation of segmentation algorithms. We show how a Bayes factor, approximated by the Bayes information criterion, can be used to compare one result against another.

2. DATA, FEATURES AND SEGMENTATION

In this work, a range of synthetic images were used, each composed of texture regions (see Results section below). All image dimensions are 200×200.

A three-level (see Figure 1) Mallat wavelet transform was applied to each image, using biorthogonal 9/7 tap filters. An energy (defined as the cardinality-normalized sum of absolute values of wavelet coefficients) was determined in each wavelet band, at each level (cf. Figure 1).

Since texture is a characteristic of a local region, the wavelet transform was carried out in 65×65 sliding windows. Thus each pixel was associated with a 10-valued feature vector, where these features were provided by the set of wavelet band energies. The wavelet transform was carried out in the window centered on the given pixel.

The principal component image used below (Figure 2) was carried out on the 10-dimensional feature space, and explained 67.3% of the variance.

Two segmentation methods were applied. One did not, and the second one did, take local neighbor information into account in carrying out the segmentation. The former method was a Gaussian mixture

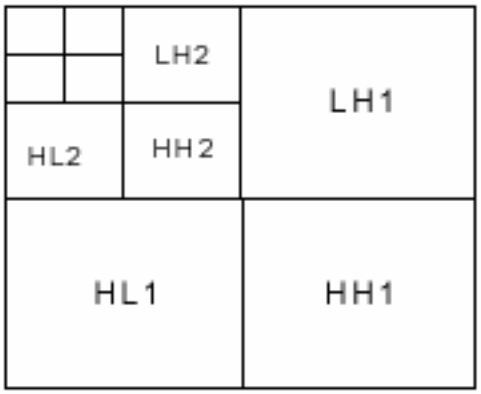

Figure 1. Ten bands, shown here, were used, following the wavelet transform. From these one energy was used to characterize each band.

model, fit to the 10-dimensional data, x_i, for each pixel i. The EM (expectation-maximization) algorithm was used. For K states, the probability density for this model is

$$f(x_i \mid \theta, \lambda) = \sum_{k=1}^{K} \lambda_k f_k(x_i|\theta_k), \tag{1}$$

where model parameters θ_k = the set of mean vectors and variance-covariance matrices, $\{\mu_k, V_k\}$; $f_k(\cdot|\theta_k)$ is a Gaussian density with mean μ_k and variance-covariance matrix V_k; $\theta = (\theta_1, \ldots, \theta_K)$; and $\lambda = (\lambda_1, \ldots, \lambda_K)$ is a vector of mixture probabilities such that $\lambda_k \geq 0$ $(k = 1, \ldots, K)$ and $\sum_{k=1}^{K} \lambda_k = 1$.

Given observations $x = (x_1, \ldots, x_{10})$, let γ be an unobserved $10 \times K$ cluster assignment matrix, where $\gamma_{ik} = 1$ if x_i comes from the k-th group, and $\gamma_{ik} = 0$ otherwise. Our goals are to determine the number of clusters K, to determine the cluster assignment of each pixel-vector, and to estimate the parameters μ_k and V_k of each cluster. In this implementation, we imposed a condition of independence, implying that the variance-covariance matrix is diagonal for each cluster.

We estimate the parameters by maximum likelihood using the EM (expectation-maximization) algorithm.[3,7] 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 is computed, so that $\hat{\gamma}_{ik}$ is the conditional probability that x_i belongs to the k-th group. In the M step, conditional maximum likelihood estimators of θ and λ given the current $\hat{\gamma}$ are computed.

So far, no local information has been used. We rewrite eqn. 1 as follows:

$$f(x_i \mid \theta, \lambda) = \sum_{k=1}^{K} \alpha_k \lambda_k f_k(x_i|\theta_k), \tag{2}$$

Weighting parameter α satisfies $\alpha_k > 0$ and $\sum_k \alpha_k = 1$. Let $N(x_i)$ be the neighborhood of x_i, taken here as a second order neighborhood of adjacent 5×5 pixels. Let $U(N(x_i), k)$ be the number of neighborhood pixels with state k. We define the weighting parameter for label or class k as:

$$\alpha_k = \frac{\exp(U(N(x_i)), k)}{\sum_j \exp(U(N(x_i)), j)} \tag{3}$$

Note that this is the density of pixel i being labeled k, conditional on the neighborhood, i.e. $p(x_i = k \mid N(x_i))$. Cf. eqn. 4 below. This probability model for x_i is the Potts or Ising model, with spatial homogeneity parameter set to 1.

To help avoid finding local optima in the EM iterations, a stochastic optimization scheme motivated by simulated annealing is used. A temperature schedule is taken as proportional to the inverse of the iteration number. The temperature, T, is then used in the multivariate Gaussian: $\exp(-\frac{1}{2T}(x_i - \mu)^t V^{-1}(x_i - \mu))$.

3. MARKOV MODELING AND BAYESIAN MODEL SELECTION

3.1. The Markov Model

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 10-valued vector. 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, taken here as adjacent 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:

$$p(x_i = j \mid N(x_i), \phi) = \frac{\exp(\phi U(N(x_i)), j)}{\sum_k \exp(\phi U(N(x_i)), j)} \tag{4}$$

Alternative notation: define energy $U(\mathbf{x}) = -\beta \sum_{i \sim j} \delta(x_i - x_j)$. The summation is over all neighboring pairs. $\delta(.)$ is a delta function, and β is the model or homogeneity parameter. The probability distribution over all possible images is: $p(\mathbf{x} \mid \beta) = 1/Z(\beta) \exp(\beta N(\mathbf{x}))$. $N(\mathbf{x})$ is the number of homogeneous cliques in the image, $\mathbf{x}$, and is equivalent to the boundary length between segments. Z is termed the partition function.

Background on the approach pursued here can be found in.[8,10,9]

3.2. Likelihood of Observed Data Given the Model

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 vector μ_j and variance-covariance matrix V_j. 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_k, V_k) for state k. We have $f(y \mid x) = \Pi_i f(y_i \mid x_i) = \Pi_i f(y_i \mid \theta_{x_i})$. This will be used below to calculate the integrated likelihood.

3.3. Potts Parameter, ϕ

Determine ϕ using the maximum pseudolikelihood: $\hat{\phi} = \text{argmin}_\phi(-\log \text{PL}(x \mid \phi))$. The pseudolikelihood is given by $\text{PL}(x \mid \phi) = \Pi_i p(x_i \mid N(x_i, \phi))$.

3.4. Bayes Information Criterion

A model M_K is the set of parameters estimated for a given number of segments, 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 ratio of posteriors, $p(D \mid M_K)/p(D \mid M_{K'})$ is referred to as a Bayes factor[6] for model M_K against model $M_{K'}$.

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. Evaluating this integral is combinatorially difficult, so an approximation which is often used is as follows:

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

where

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

where $\hat{\theta}_K$ is the maximum likelihood estimator of θ_K, $\dim(\theta_K)$ is the number of parameters estimated (constant, if we assume a fixed number of segments, as we do in each experiment below), and N is cardinality of the data (again, fixed in the experiments below).

In these circumstances, we see that the likelihood, integrated over all pixels, is the crucial term.[5,10] The results below give the log likelihood found in each case.

4. RESULTS

All figures shown here were independently histogram-equalized. Figures 2, 3 and 4 relate to three different studies. Note that in each case the most acceptable result has the largest value of the Potts spatial parameter, ϕ. In Figure 2, the principal component result is quite poor in that it gives the same label to two of the main segments. The non-locally based result is visually less good in each case, compared to the locally based result. A larger value of ϕ corresponds to wider or broader spatial influence. Therefore it is a very reasonable measure of quality of segmentation. However it is based exclusively on the segmentation labels, and it is not difficult to arbitrarily define segmentations associated with high values and which have little or no link with the input data.

In Figures 2, 3 and 4, a larger BIC also corresponds to a better result. The measure used in each case is the conditional likelihood of the observed data on the model, $L(y \mid K)$. These measures, too, are consistent with the spatial homogeneity measures.

5. CONCLUSION

The quantitative assessment criteria presented here have been shown to be consistent with visual assessment of segmentation performance. They can be used with any optimization method for inducing the segmentation. Of the segmentation approaches used here, we see that the multidimensional segmentation performs best. In particular we see that a principal component image performs poorly.

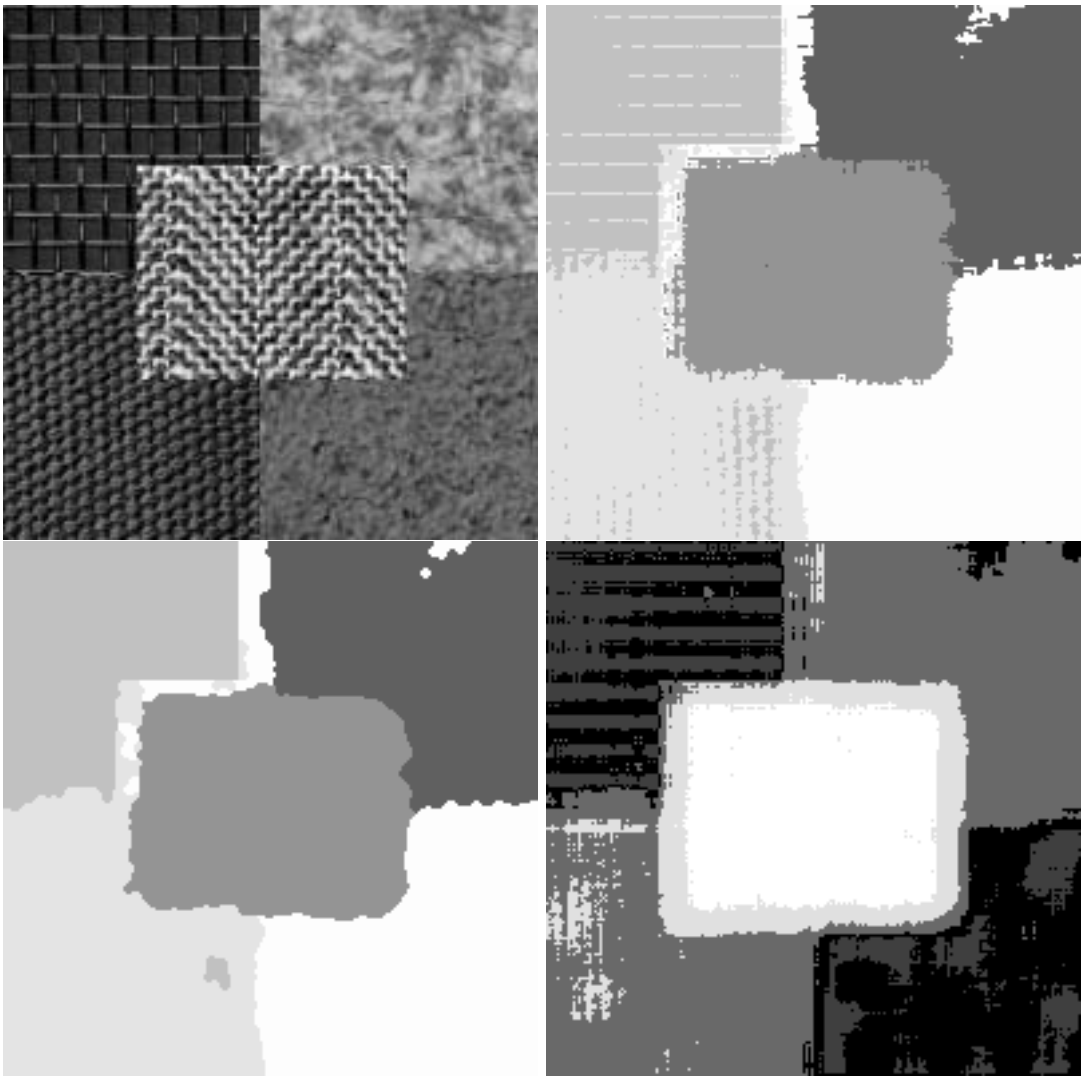

Figure 2. Clockwise from upper left: Input image. Non-locally based segmentation: $\phi = 0.7755$, BIC $= -1014900$. Locally based segmentation: $\phi = 2.7161$, BIC $= -1009171$. Segmentation of principal component image: $\phi = 0.7311$, BIC $= -1038317$.

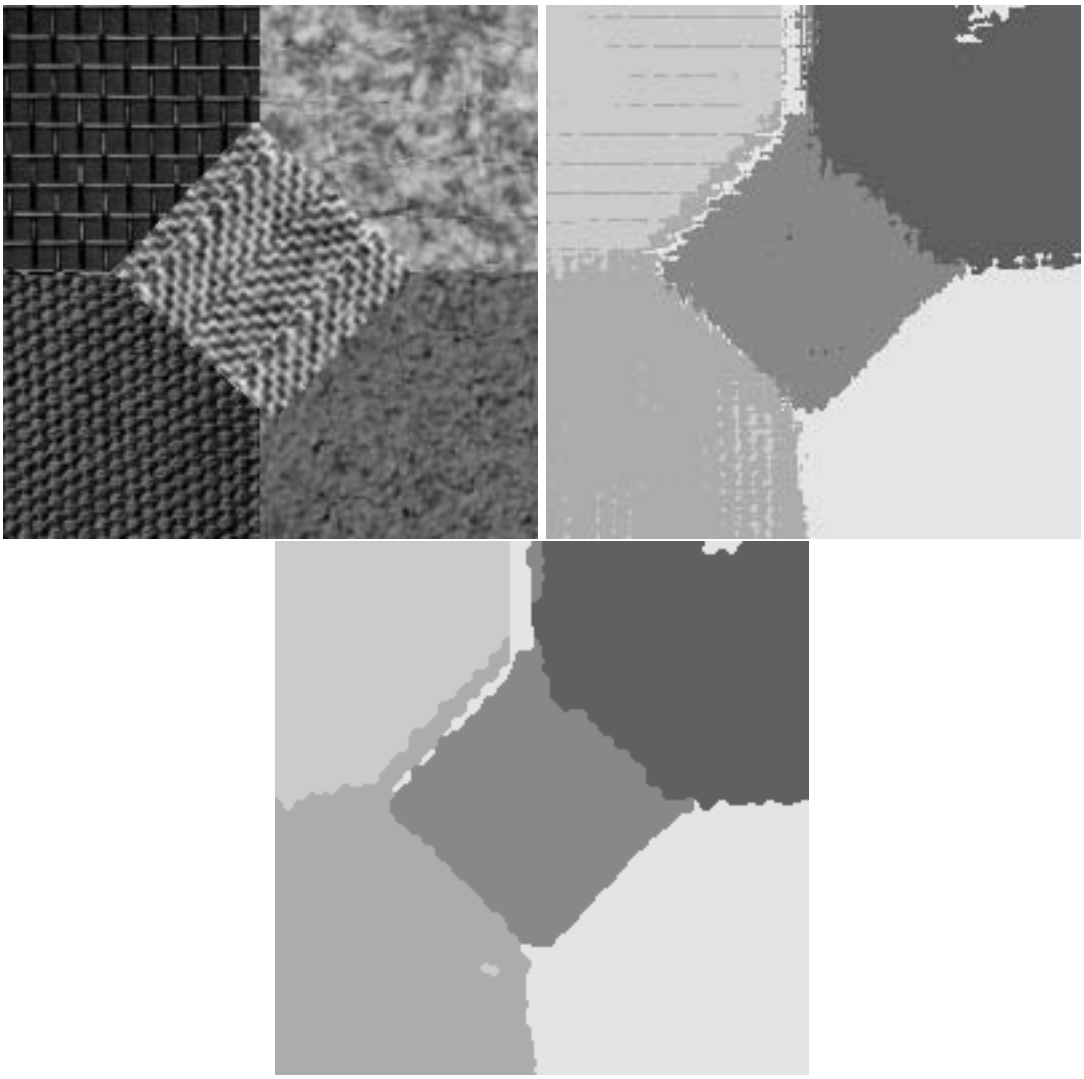

Figure 3. Clockwise from upper left: Input image. Non-locally based segmentation: $\phi = 0.7755$, BIC $= -941616$. Locally based segmentation: $\phi = 2.7244$, BIC $= -935675$.

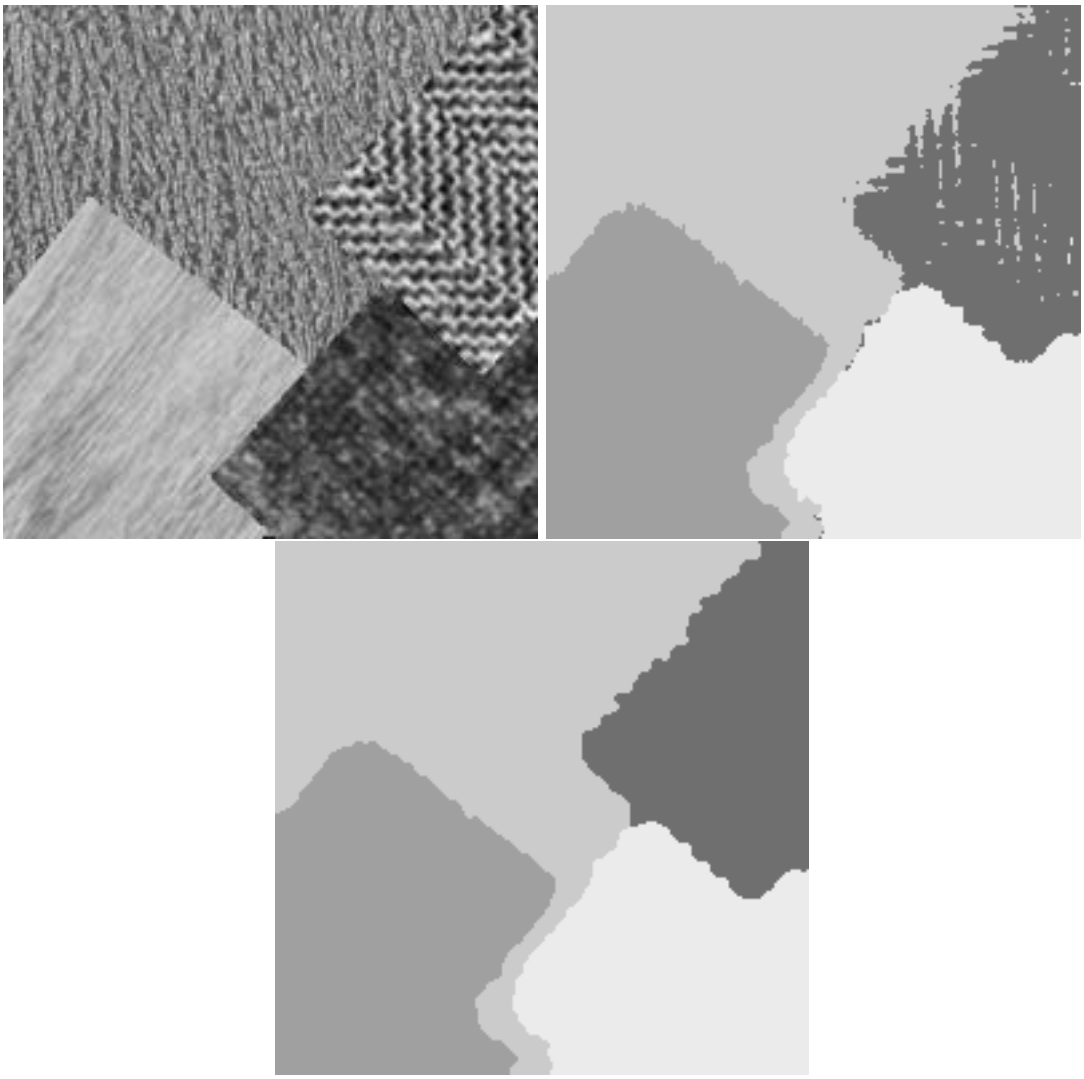

Figure 4. Clockwise from upper left: Input image. Non-locally based segmentation: $\phi = 0.9151$, BIC $= -913033$. Locally based segmentation: $\phi = 10.882176$, BIC $= -912524$.

Acknowledgement

This work was supported by the EPSRC "Virtual Sieve" project, GR/R38835.

REFERENCES

1. J. Besag. Statistical analysis of dirty pictures. *Journal of the Royal Statistical Society, Series B*, 48:259–302, 1986.
2. G. Cross and A. Jain. Markov random field texture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5:25–39, 1983.
3. 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.
4. N. Fatemi-Ghomi. *Performance Measures for Wavelet-Based Segmentation Algorithms*. PhD thesis, Surrey University, 1997.
5. C. Ji and L. Seymour. A consistent model selection procedure for Markov random fields based on penalized pseudolikelihood. *Journal of Applied Probability*, 6:423–443, 1996.
6. R. Kass and A. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.
7. G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, 1997.
8. D.C. Stanford. *Fast Automatic Unsupervised Image Segmentation and Curve Detection in Spatial Point Patterns*. PhD thesis, University of Washington, 1999.
9. 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.
10. 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.