# CODING THEORETIC APPROACH TO IMAGE SEGMENTATION

*Unoma Ndili, Robert D. Nowak*

Dept. of Electrical and Computer Eng.
Rice University
Houston, TX 7705-1892, **U.S.A.**

*Mário A. T. Figueiredo*

Institute of Telecommunications, and
Dept. of Electrical and Computer Eng.
*Instituto Superior Técnico*
1049-001 Lisboa, **Portugal**

## ABSTRACT

This paper introduces multi-scale tree-based approaches to image segmentation, using Rissanen's coding theoretic *minimum description length* (MDL) principle to penalize overly complex segmentations. Images are modelled as Gaussian random fields of independent pixels, with piecewise constant mean and variance. This model captures variations in both intensity (mean value) and texture (variance). Segmentation thus amounts to detecting changes in the mean and/or variance. One algorithm is based on an adaptive (greedy) rectangular recursive partitioning scheme. The second algorithm is an optimally-pruned "wedgelet" decorated dyadic partitioning. We compare the two schemes with an alternative constant variance dyadic CART (*classification and regression tree*) scheme which accounts only for variations in mean, and demonstrate their performance on SAR images.

## 1. INTRODUCTION

In this paper, we propose a method to parse an $m \times n$ image $y$ into homogeneous regions in terms of mean and variance. We model the image as composed of connected regions of pixels assumed to be independent and Gaussian distributed:

$$y_{ij} = \mu_{ij} + \sigma_{ij} \, . \, z_{ij}, \qquad \text{for} \quad 1 \leq i \leq m, \;\; 1 \leq j \leq n, \tag{1}$$

where the $z_{ij}$ are i.i.d. $\mathcal{N}(0,1)$, so that $y_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma_{ij}^2)$. Both the mean $\mu_{ij}$ and variance $\sigma_{ij}^2$ can vary from region to region, but are constant within a given region; *i.e.*, we model images as Gaussian random fields with piecewise constant mean and variance. This model is intended to capture variations in both intensity (mean value) and texture (variance).

We adopt Rissanen's MDL principle [1, 9] to achieve unsupervised segmentation. MDL provides a mathematical foundation for data dependent model selection by balancing the brevity of model description to its fidelity to the data. We introduce two different methods which involve a tradeoff between speed and optimality: *adaptive recursive partitioning* (ARP), a greedy scheme which produces rectangular (not necessarily dyadic) tessellations; and *wedgelet decorated dyadic partitioning* (WDDP), an optimal multiscale analysis restricted to dyadic partitions, which allows for wedge-splits. WDDP differs from the dyadic CART algorithm proposed in [3] in that it detects changes, not only in mean, but also in variance. The ARP algorithm was developed for Poisson fields in [8]; here we adapt it to a Gaussian model of unknown mean and variance.

The paper is organised as follows: Section 2 briefly reviews the MDL principle and derives conditional densities required by our Gaussian model in the MDL criteria. Section 3 describes the two schemes: the ARP and WDDP algorithms. Section 4 presents comparative experimental results of the two schemes. We conclude in Section 5.

This work was supported by the National Science Foundation, grant no. MIP–9701692, the Army Research Office, grant no. DAAD19-99-1-0290, and the Office of Naval Research, grant no. N00014-00-1-0390.

## 2. THE MDL PRINCIPLE

Suppose we want to optimally encode and transmit a length-$n$ data sequence $X^n = \{x_1, \cdots, x_n\}$. Given a probability distribution $p(X^n|\theta)$, parameterized by $\theta$, it is well known that the optimal code-length for prefix codes is given by $-\log p(X^n|\theta)$ [1, 9]. If there are $K$ models competing to explain our data $\{p_k(X^n|\theta_k)\}_{k=1}^{K}$, the MDL criterion states that the "best" is the one minimizing the *description length* (DL). One form of this DL is obtained by *two-part coding*,

$$L(X^n) = L(\widehat{\theta}_k) - \log p(X^n|\widehat{\theta}_k), \quad k = 1, \ldots K, \tag{2}$$

where $L(\widehat{\theta}_k)$ is the DL needed to describe $\widehat{\theta}_k$, the *maximum likelihood estimate* of $\theta_k$, such that the decoder knows the model under which the data code (of length $-\log p(X^n|\widehat{\theta}_k)$) was obtained. MDL has been successfully used in several image analysis problems [4, 5, 6, 7, 8, 11].

### 2.1. MDL for Gaussian Data

In the MDL approach reviewed above, we first encode and send a parameter estimate ($\widehat{\theta}_k$), then the data itself, coded according to $p(X^n|\widehat{\theta}_k)$. In our image model, we assume that the pixels in each homogeneous region are i.i.d. Gaussian. Under this data assumption, the elements of $\widehat{\theta}_k$ are the sample mean and sample variance. Since these statistics are real-valued, they have to be quantized to finite precision in order to yield a finite $L(\widehat{\theta}_k)$. The standard solution is the well known $\frac{1}{2}\log n$ code-length for each components of $\theta_k$, which is based on asymptotic approximations [9].

To code the data under the assumed model, we don't need code-words for all possible data out-comes; in fact, once the receiver has the parameter estimate, it knows that only data out-comes that could have led to this estimate need to be coded. A code with code-words that are never used is called *incomplete* [1, 10]; to avoid incompleteness, we derive below the conditional density of the data, given the parameter estimates (or equivalently, the sufficient statistics).

#### 2.1.1. Conditional Density: Unknown Mean and Variance

Let $X^n = \{x_1, ..., x_n\}$ be the pixel values in a homogeneous region of the image (i.e., $n$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$ samples). The sample mean and sample variance, or equivalently the sufficient statistics $t_1 \equiv \sum_{i=1}^n x_i$ and $t_2 \equiv \sum_{i=1}^n x_i^2$, are sent to the decoder. To build an incompleteness-free code, we need the conditional density $-\log p(X^n|t_1, t_2)$.

The Gaussian density may be written in exponential family form with parameters $\zeta_1 = \frac{\mu}{\sigma^2}$ and $\zeta_2 = -\frac{1}{2\sigma^2}$,

$$p(X^n|\zeta_1, \zeta_2) = \left(\frac{-\zeta_2}{\pi}\right)^{\frac{n}{2}} e^{\left(\frac{n\zeta_1^2}{4\zeta_2}\right)} e^{(\zeta_1 t_1 + \zeta_2 t_2)}, \quad (3)$$

where $t_1$ and $t_2$ are the sufficient statistics. The factors in (3) can be interpreted using Neyman's factorization, $p(X^n|\zeta_1, \zeta_2) = p(X^n|t_1, t_2)p(t_1, t_2|\zeta_1, \zeta_2)$, showing that $p(X^n|t_1, t_2)$ is uniform (constant) on the constraint set $\mathcal{C}(t_1, t_2) = \{X^n : \sum x_i = t_1 \text{ and } \sum {x_i}^2 = t_2\}$; this is an $(n-1)$-sphere resulting from intersecting an origin-centered radius-$\sqrt{t_2}$ $n-$sphere, with an hyper-plane at $45^\circ$ in $R^n$.

To write $p(X^n|t_1, t_2)$ we need its normalizing constant which is the measure of $\mathcal{C}(t_1, t_2)$. The vector from the point on the hyperplane $(\frac{t_1}{n}, \cdots \frac{t_1}{n})$ to the origin has length $\frac{t_1}{\sqrt{n}}$, and is perpendicular to the hyperplane. From Pythagoras' theorem we conclude that the radius of the $(n-1)$-sphere is $r = (t_2 - \frac{{t_1}^2}{n})^{1/2}$. Using the formula for the surface area of a hyper-sphere (where $\Gamma(\cdot)$ is Euler's gamma function)

$$p(X^n|t_1, t_2) = \frac{1}{2}\,\Gamma(\frac{n-1}{2})\, r^{2-n}\, \pi^{\frac{1-n}{2}}\, I_{\mathcal{C}(t_1,t_2)}(X^n), \quad (4)$$

where $I_A(X^n)$ denotes the indicator function of set $A$.

## 3. TWO ALGORITHMS

We consider two different multi-scale tree-structured approaches to segment an image, based on our image model and the MDL criterion. The entire image is represented by the root node and final segments correspond to the leafs of the tree. Our two algorithms differ in the way this tree is grown, either from top down, in a greedy way (ARP), or using a bottom-up, optimal pruning scheme (WDDP).

To describe our algorithms, we need some notation. Given any $m_R \times n_R$ region $R$ of the image $y$, $y^R$ denotes the restriction of $y$ to $R$, $t_1^R \equiv \sum_{(ij)\in R} y_{ij}$, and $t_2^R \equiv \sum_{(ij)\in R} y_{ij}^2$. A region $R$ can be split into $H$ disjoint 'sibling' regions $R_h$, for $h = 1, \ldots, H$.

### 3.1. Adaptive Recursive Partitioning (ARP)

This algorithm recursively takes each sub-block $R$ (starting with the whole image), determining whether it is best represented as homogeneous (common mean and variance) or split into either two or four homogeneous rectangles. We take advantage of previously transmitted information, *i.e.*, if a region is split into $H$ regions, we only need to transmit sufficient statistics of $H-1$ regions, since one set of statistics may be obtained from those previously transmitted for their 'parent'. We start by assuming that the statistics $t_1^I$ and $t_2^I$ corresponding to the entire image $I$ have been transmitted. The ARP algorithm admits two possible model classes.

**Model Class 1 (No Split):** Since sufficient statistics have been transmitted, we only need to code the data. The code-length is then $L_1(R) = -\log p(y^R|t_1^R, t_2^R)$, where the probability density is given by equation (4).

**Model Class 2 (Split):** We examine all possible two-way ($H = 2$) and four-way ($H = 4$) splits and decide on the one achieving the minimum code-length. In addition to encoding the data under different models, we must also code the splitting location $k$, and sufficient statistics for the parameters of $H-1$ of the sub-regions. For illustration, let us consider the code-length involved in two-way horizontal splits of an $m_R \times n_R$ rectangular region. Since there are $J = (m_R - 1)$ possible horizontal split locations, we need $\log J$ bits to code them. For the sufficient statistics of one of the resulting regions, we use a worst case analysis, noting that $t_1^R \geq t_1^{R_h}$ and $t_2^R \geq t_2^{R_h}$ (since image data is $\geq 0$). Thus we require at most $\frac{1}{2}\log(m_R n_R)$ for each statistic. Accordingly,

$$\begin{aligned} L_2^k(R) \;&=\; -\log p(y^{R_1}|t_1^{R_1(k)}, t_2^{R_1(k)}) + \log J \\ &\quad -\log p(y^{R_2}|t_1^{R_2(k)}, t_2^{R_2(k)}) + \log(n_R m_R). \end{aligned}$$

The code-length for two-way vertical and four-way horizontal and vertical splits is computed similarly.

If $L_1(R) < \min_k L_2^k(R)$, model class 1 is chosen for the sub-block and processing is stopped, otherwise we split at $k^* = \arg\min_k L_2(k)$, and re-apply the same splitting criterion to each of the resulting sub-rectangles. The procedure stops when $L_1$ is selected for all sub-regions or we get to pixel level. Notice that this is a greedy scheme since at each level we ignore the fact that each sub-block may be further sub-divided to achieve an even shorter code-length.

### 3.2. Wedgelet Decorated Dyadic Partitioning (WDDP)

WDDP is an optimal *bottom-up* partitioning algorithm based on the CART wedgelet algorithm proposed by Donoho [3]. Our multi-scale analysis yields a wedgelet-decorated tree, optimally grown from the bottom up, following the CART approach [2]. At each level, description lengths satisfy the additivity property required by the CART theorem, allowing for sequential optimization. Wedgelet decompositions add flexibility, enabling non-dyadic partitioning and resulting in more accurate approximation of arbitrary region boundaries [3].

We restrict our analysis to dyadic partitions due to the prohibitive complexity required to search through all possible subtrees. Wedgelets [3] enable us to represent wedge-splits (line segment from a pixel-vertex on an edge of $R$ to a pixel-vertex on another edge) at different locations and orientations. To reduce our computational complexity, we construct a restricted dictionary of wedge-splits for each region using only vertices marked off at $\tau = 4$ equi-distant points on all edges of dyadic blocks of edge length greater than 4 pixels. The cardinality of our dictionary is then a constant $B = 80$ (i.e., $6\tau^2 - 4\tau$). For blocks of size 4 and less, we consider all possible pixel vertex wedges.

Given a square region $R$, a quad-split produces four "children" $R_1, ..., R_4$. The additivity property requires that the description length for $R$ be inherited as the a sum of the description length of its four children: $L(R) = \sum_{i=1}^{4} L(R_i)$. Starting with individual pixels, we perform quad-merges proceeding upwards by always inheriting the best possible cumulative description length. At each node, the description length is obtained as the best of three model classes:

**Model Class 1:** Represent $R$ as homogenous with a common mean and variance, and code-length given as:

$$L_1(R) = -\log p(y^R | t_1^R, t_2^R) + \log n,$$

where $n$ denotes the number of pixels in the image.

**Model Class 2:** Perform one of the $B$ possible wedge splits of $R$ to obtain two polygons $R_1(b)$ and $R_2(b)$ for $b = 1, \cdots, B$. In addition to coding the data in each of the resulting regions, we must encode sufficient statistics for both regions (again using worst case code-lengths), and describe the location of the wedge using $\log B$ bits. Our description length is:

$$\begin{aligned} L_2^b(R) \quad &= \quad -\log p(y^{R_1(b)} | t_1^{R_1(b)}, t_2^{R_1(b)}) \\ &- \log p(y^{R_2(b)} | t_1^{R_2(b)}, t_2^{R_2(b)}) + 2\log n + \log B \end{aligned}$$

The description length of the 'best' wedgelet representation of $R$ is $L_2(R) = \min_b L_2^b(R)$.

**Model Class 3:** In this case, $R$ is split into a set of 4 children $\{R_1, R_2, R_3, R_4\}$. The code-length for $R$ is the sum of the best description lengths of its children:

$$L_3(R) = \sum_{h=1}^{4} \min\{L_1(R_h), L_2(R_h), L_3(R_h)\}.$$

The shortest code-length for $R$ is $L(R) = \min\{L_1(R), L_2(R), L_3(R)\}$. We prune the tree if $L(R) \neq L_3(R)$.

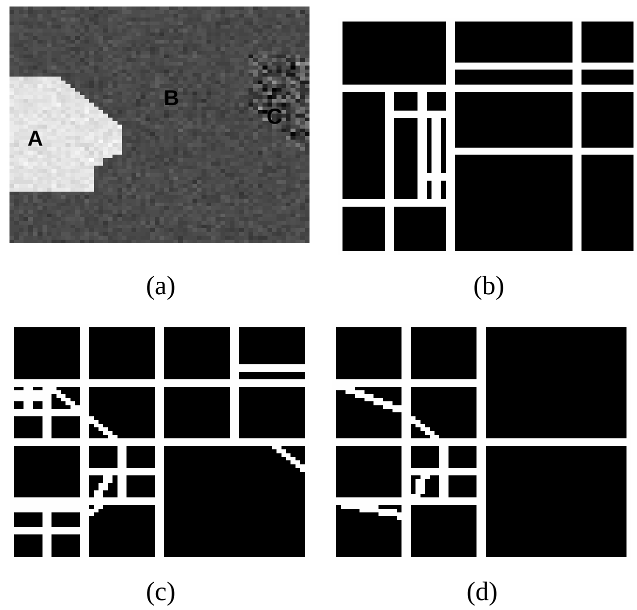


**Fig. 1**. (a) Original synthetic image. Segmentation maps obtained using the following schemes: (b) The ARP. (c) The WDDP. (d) The constant-variance dyadic CART.

## 4. RESULTS AND COMPARISONS

In Fig. 1(a), the patch on the left (A) has a different mean from B but the same variance; the region on the right (C), has the same mean as B, but a different variance. The region boundaries were chosen to illustrate the wedgelet idea. The segmentation maps produced by the ARP and WDDP algorithms, in Figs. 1(b) and 1(c), show that both methods detect the change in texture on the right patch. The ARP yields a blocky rectangular segmentation of the image; the

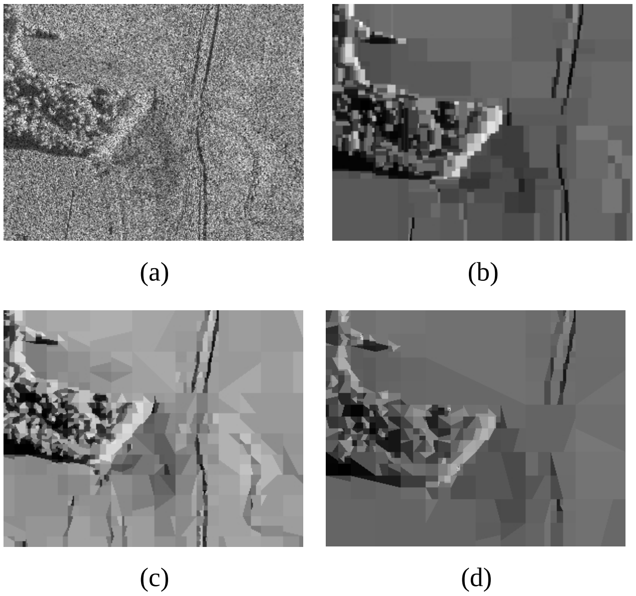


**Fig. 2**. MSTAR SAR amplitude data HB06170 (a) original image (b) ARP-segmented image (c) WDDP-segmented image (d) constant-variance WDDP segmentation

WDDP approximates better the linear boundaries by using wedgelets. Fig. 1(d) shows the segmentation map for the constant-variance wedgelet-decorated dyadic CART algorithm [3] which performs as well as the WDDP in regions with varying mean intensity, but fails to detect the change in variance in region C.

Fig. 2 shows results of all three segmentation algorithms applied to an MSTAR *synthetic aperture radar* (SAR) image, in which changes in variance are important features. Finally, Fig. 3 illustrates our two schemes (ARP and WDDP) on a SAR image of an agricultural field in Netherlands.

## 5. CONCLUSION

We have proposed an approach to segmenting images using an MDL criterion. Images are modeled as piece-wise homogeneous, with regions described as sets of independent Gaussian samples of constant mean and variance. We presented two schemes: an *adaptive recursive partitioning* (ARP) algorithm, which is greedy but not restricted to dyadic partitions; a *wedgelet decorated dyadic partitioning* (WDDP) scheme, which is optimal but restricted to dyadic partitions. ARP and WDDP are fast and unsupervised, enabling segmentation based on both intensity and variance.

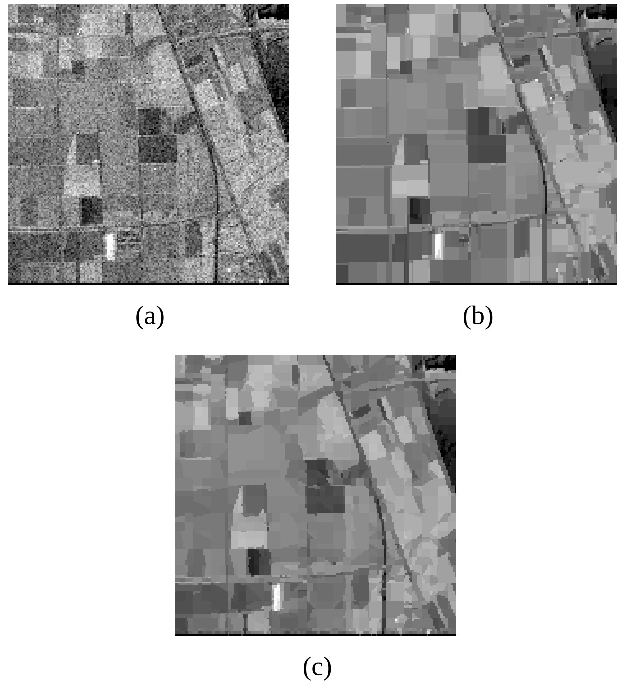


**Fig. 3**. (a) Multi-look SAR image of agricultural fields in Flevoland, Netherlands; (b) mean plot of ARP segmented image; (c) mean plot of WDDP segmented image

## 6. REFERENCES


[1] A. Barron, J. Rissanen, and B. Yu, “The minimum description length principle in coding and modeling”, *IEEE Trans. Info. Theo.*, vol. 44, pp. 2743-2760, 1998.

[2] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, CRC Press, 1983.

[3] D. Donoho, “Wedgelets: Nearly-Minimax Estimation of Edges”, *Annals of Stat.*, vol. 27, pp. 859-897, 1999.

[4] M. Figueiredo and J. Leitão, “Unsupervised image restoration and edge location using compound Gauss-Markov random fields and the MDL principle”, *IEEE Trans. Image Proc.*, vol 6, pp. 1089-1102, 1997.

[5] T. Kanungo, B. Dom, W. Niblack, and D. Steele, “A fast algorithm for MDL-based multi-band image segmentation”, in *Proc. CVPR’94*, pp. 609-616, Seattle, 1994.

[6] H. Krim and I. Schick, “Minimax description length for signal denoising and optimal representation”, *IEEE Trans. on Information Theory*, vol. 45, pp. 898-908, 1999.

[7] Y. Leclerc, “Constructing simple stable descriptions for image partitioning”, *Int. J. of Comp. Vision*, vol. 3, pp. 73-102, 1989.

[8] R. Nowak and M. Figueiredo, “Unsupervised progressive parsing of Poisson fields using minimum description length criteria”, *Proc. ICIP’1999*, Kobe, Japan, pp. II.26-II.29, 1999.

[9] J. Rissanen *Stochastic Complexity in Statistical inquiry*, World Scientific Publishers, Singapore, 1989.

[10] J. Rissanen, “Fisher information and stochastic complexity”, *IEEE Trans. on Information Theory*, vol. 42, pp. 40-47, 1996.

[11] N. Saito, “Simultaneous noise supression and signal compression using a library of orthonormal bases and the MDL criterion ”, in *Wavelets in Geophysics*, Academic Press, 1994.