AN ACCURATE MODEL FOR QUADTREES REPRESENTING NOISELESS IMAGES OF SPATIAL DATA

Enrico Nardelli and Guido Proietti

IASI, National Research Council, Viale Manzoni 30, 00185 Roma, Italy, Ph.: +39-6-77.16.420, Fax: +39-6-77.16.461,
E-Mail: {nardelli,guido}@iasi.rm.cnr.it
Univ. of L'Aquila, Dept. of Pure and Applied Mathematics, Via Vetoio, Loc. Coppito, 67100 L'Aquila, Italy

ABSTRACT

In this paper we propose and analyze a new meaningful branching sequence to generate random quadtrees representing binary images. In particular, we show that this sequence produces expected distributions of external and internal nodes much closer to real data than all previous proposed approaches in the literature to model both random binary images and quadtrees. This new model provides a good compromise in representing images belonging to various classes, more or less structured. The effectiveness of the new proposed model is shown through a comparison with respect to nodes distributions of representative real spatial data images. The introduction of this new realistic model can have a large impact on the analysis of expected performances of a large class of algorithms for spatial data processing. First experimental results show that this new model closely simulate real cases.

1. INTRODUCTION

Many different approaches have been proposed in the literature to represent noiseless images of spatial data: array representation (raster-based), runlenght codes, polygons (vector-based), bounding boxes, mapping to higher or lower dimensional spaces and so on; see [1] for a survey. A largely used data structure for its good compromise between space occupancy and time consumption in the most recurrent spatial data processing operations is the *quadtree* [2]. The term quadtree is generally used to denote a hierarchical data structure developed on the basis of a regular decomposition of the space. The hierarchical decomposition is data-driven, but always proceeds according to a regular scheme, going to deeper levels only where represented features are more densely distributed. In this way space is saved where the distribution is more scarce.

The analysis of the time and space efficiency of the quadtree data structure generally clashes with the not easy problem to give a random model defining in a satisfactory way the behaviour of a certain class of images. In fact, the intuitive and conventional random model that assigns to each pixel, independently from any other pixel, a certain probability to be black or white (i.e., the so called *pixel based* model), leads to represent images in which there is a very low degree of aggregation. Disaggregated images are not typical instances of many real image classes (e.g., landuse maps, geographical maps and also sets of geometrical objects). So, even if the probability for a pixel to be black is high, the generated random image seems always to be sparse and far away from reality. Since the quadtree data structure is not very efficient for such a kind of disaggregated images, this also affects in a significant way its performances, distorting the robustness of the theoretical approach.

Therefore, Samet [3] tried to overcome the pixel based model and suggested a new random model aiming to describe more closely the kind of images representing spatial data. The novelty in Samet's idea is in the fact that his model is "tree-oriented", in the sense that it gives a description of the way a typical quadtree is structured to represent real spatial data. According to Samet's idea of random quadtree, each leaf node is assumed to be equally like to appear at any position and level in the tree. The effectiveness of this perspective in the modeling of random images has been outlighted in [3], where it is shown that theoretical results on the average time complexity in the analysis of neighbour finding algorithms are close to statistics of real tests. Also concerning the average storage efficiency of quadtrees, Samet [4] proposed a detailed comparison among his random model and real instances taken as representative of the most common image classes.

Later, Puech and Yahia [5], proposed a more general approach, enclosing also the intuitive model given by Samet: in their definition, a random quadtree is built on the basis of a branching process that at each level assigns a certain probability for a node to be internal or external. Modifying the coefficients of the branching sequence, one is able to represent classes of images of completely different nature, and this flexibility is quite the strength of the branching model.

The introduction of the general random quadtree model representing binary images, has given the possibility to be able to produce quadtrees close to those that represent images in the reality. Starting from this objective and basing on standard images as suggested by Samet [4], we have been able to discover a very intriguing nodes

0-8186-6950-0/94 $4.00 © 1994 IEEE

distribution law that seems to link images of different classes. From this law, a recursive branching sequence has been produced. In this paper we present the sequence and show that it produces expected distributions of external and internal nodes much closer to real data than all previous proposed approaches in the literature to model both random binary images and quadtrees. Unfortunately, the recursive sequence appears hard to be exploited with respect to the quadtree level using standard analytical techniques, and then no theoretical bounds can at the moment be provided. Anyway, practical comparisons are made, that show how well the sequence works.

The paper proceeds as follows: in Section 2 we recall the definition of the quadtree structure for binary images. In Section 3 we review the definitions of the various models of image and quadtree randomness proposed up to now. In Section 4 we introduce our proposal for a new branching sequence that seems to model in a very accurate way classes of images taken as samples of the whole universe. Finally, Section 5 contains considerations for further work and concluding remarks.

2. BINARY IMAGES AND QUADTREES

When only a single feature exists in an image with respect to a background, we can think to it as constituted of black pixels (i.e., all the pixels containing the feature) and white pixels (i.e., all the pixels that do not contain the feature), and so we speak of *binary image*.

For binary images, the decomposition process carried out by the quadtree becomes intuitive. Assuming to have at disposal a binary image of size $T \times T$ (e.g., pixel elements), where T is such that there exists an integer m such that $2^m=T$, we proceed in the following way: at level m there is the whole image, of side length T. At the first stage of decomposition the image consists of four quadrants of side length $T/2$. At a second stage each quadrant is then subdivided into four quadrants of side length $T/2^2$ and so on. The decomposition stops either when a quadrant is wholly covered (it is said to be black) or wholly uncovered (it is said to be white). We shall use also the term *block* to denote a quadrant. The decomposition can go on until the pixel level, with quadrants of side length $T/2^m$. The decomposition can be represented as a tree of outdegree 4, with the root (at level m) corresponding to the whole image and each node (at level m-d) to a quadrant of side length $T/2^d$. The sons of a node are, in preorder, labeled NW, NE, SW and SE. For a given image, nodes are then black, white (leaf nodes) or grey (intermediate nodes). Correspondingly, we speak of black, white and grey blocks. Using a drawing as a sample, the ideas will be clearer:

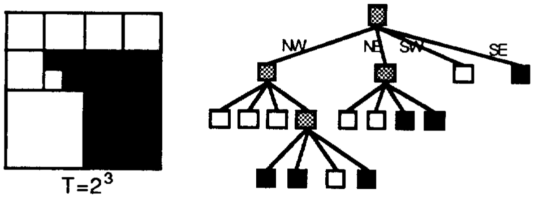

Figure 1: A binary image and its quadtree

3. PROBABILISTIC MODELS FOR BINARY IMAGES AND QUADTREES

An image can be modelled either by means of its pixel distribution or just looking to the structure of the quadtree representing it. This has produced in the literature the following two models that, under certain conditions, are equivalent [6]:

(1) *PIXEL BASED*: in this model (image oriented) each pixel is assumed to be statistically independent from any other pixel; if p is the probability for a generic pixel to contain the feature, then a level-i node is fully covered by the feature with probability $B_i=p^{4^i}$, fully uncovered with probability $W_i=(1\text{-}p)^{4^i}$ and partially covered (and then internal to the quadtree) with probability $G_i=1\text{-}p^{4^i}\text{-}(1\text{-}p)^{4^i}$.

(2) *NODE BASED*: this model (tree oriented) has been proposed by Puech and Yahia [5]. Let Q_m be the set of all class-m quadtrees, i.e. quadtrees of height less than or equal to m. Let (β_i) be a non increasing sequence of m+1 reals between zero and 1/2 ($0<\beta_i<1/2$ being $\beta_0=1/2$). A random tree of Q_m is built by using a branching process, such that the quantity $2\beta_i$ is the probability for a level i node to be external (and so by definition $2\beta_0=1$, i.e., at level 0 all nodes are external with probability 1 as they should), and then $G_i=1\text{-}2\beta_i$ is the probability for a level i node to be internal. Once a node is known to be external, we do not make assumption on its colour, i.e., the node can be indifferently white or black.

Independently from the above probabilistic assumptions, another not conventional model has been proposed by Samet [3]. There, the idea is that an external node containing a fixed pixel has equal probability of being of any size, or, in other words, a leaf node is equally likely to appear in any position and level in the quadtree. This means that in the complete quadtree of height m, and then containing $\frac{4^{m+1}\text{-}1}{3}$ nodes, there are $1,4,16,\dots,4^{m\text{-}i}$ leaf nodes at levels m, m-1, m-2,..., i respectively. In a generic sample of a random quadtree, if each leaf node is equally like to appear everywhere in the quadtree, the probability of its existence at level i is the ratio between the number of places available at level i and the total number of places, i.e., $3\left(\frac{4^{m-i}}{4^{m+1}-1}\right)$. This number also expresses how many of all leaf nodes can be found at level i. In [4] a slightly approximated value is provided for this value, namely $3/4^{i+1}$.

The equivalence between the node based model when β_i is equal to 1/2(i+1) and the random model of images as defined by Samet has been stated in [4] but not presented; here is how it works.

Proposition 1: In a random quadtree generated using a branching process with β_i=1/2(i+1), each leaf node is equally likely to appear in any position and level in the quadtree.

Proof: Because of the branching process we have that a node at level i exists as a leaf with probability:

$$L_i = \left(\frac{1}{i+1}\right) E_i$$

where E_i is the probability for a node to exist at level i. The probability E_i for a node to exist at level i depends by the probability for a node at level i+1 to exist and to be internal. This, pointed out that E_m=1 since the root always exists, produces the following recurrence:

$$E_i = G_{i+1}\,E_{i+1} = \left(1 - \frac{1}{i+2}\right) E_{i+1} = \prod_{k=i}^{m-1} \frac{k+1}{k+2} = \frac{i+1}{m+1}$$

From here, the above expression becomes:

$$L_i = \left(\frac{1}{i+1}\right) E_i = \frac{1}{i+1}\,\frac{i+1}{m+1} = \frac{1}{m+1}$$

and then the expected number of leaf nodes at level i is:

$$N_{m,i} = 4^{m-i} L_i = \frac{4^{m-i}}{m+1}$$

From this the expected number of leaf nodes for a quadtree of height m is:

$$N_m = \sum_{i=0}^{m} \frac{4^{m-i}}{m+1} = \frac{4^{m+1}-1}{3(m+1)}$$

and then the expected rate of leaf nodes at level i for a quadtree of height m is:

$$\frac{N_{m,i}}{N_m} = \frac{\frac{4^{m-i}}{m+1}}{\frac{4^{m+1}-1}{3(m+1)}} = 3\left(\frac{4^{m-i}}{4^{m+1}-1}\right)$$

perfectly in accord with the Samet's definition of random image.

□

Then, the model proposed by Samet is a significative instance of the node based model. We say significative for two main reasons: as first, it appears (as the following table borrowed from [4] shows) close to the effective node distribution of a large class of images; moreover, its meaningful resides in the fact that it is the expression of a relatively simple and structural definition.

Leaf Size	Model	Floodplain	Topography	Landuse	Pebble
1×1	26,214 (75.00)	2,468 (47.4)	14,832 (59.3)	16,112 (56.4)	27,316 (60.8)
2×2	6,553 (18.75)	1,599 (29.9)	7336 (29.3)	8,484 (29.7)	11,995 (26.7)
4×4	1,638 (4.69)	660 (12.7)	2,175 (8.70)	2,984 (10.5)	4,418 (9.83)
8×8	409 (1.17)	263 (5.05)	470 (1.88)	784 (2.62)	1,095 (3.44)
16×16	102 (0.293)	175 (3.36)	138 (0.552)	175 (0.613)	108 (0.240)
32×32	25.6 (0.073)	57 (1.09)	51 (0.204)	38 (0.133)	18 (0.040)
64×64	6.4 (0.018)	22 (0.423)	8 (0.032)	8 (0.028)	0 (0.000)
128×128	1.6 (0.005)	2 (0.038)	2 (0.008)	0 (0.000)	0 (0.000)
TOTAL	34,952	5,206	25,012	28,549	44,950

Table 1: Leaf node size distribution

To fully comprise the table, a number of remarks is needed:

1) The size of the image we are considering is 512x512, but we limit our attention (as suggested by Samet) to 128x128 size nodes;
2) In effect, it appears that the Samet's distribution differs not so slightly from real data; furthermore, a great deal of structural uniformity seems evident in all the images, apart from differences in represented data.

4. THE NEW BRANCHING SEQUENCE

We propose here to introduce a new meaningful model that allows us to closely represent real images behaviour. We start observing that in almost all different kinds of images and especially in the landuse and topography maps (incidentally, the two maps appear to have an average structure between the floodplain and the pebble maps), the leaf size percent distribution could be resumed by the following statement: the ratio between the number of leaves contained in the i-th level and the number of those contained in the i+1-th level seems to be approximately equal to i+2. So for example, it seems that at level 1 (i.e., pixel's father level) there is a half of the leaves contained at the pixel level and at the same time three times the percentage of leaf nodes at the pixel's grandfather level.

If we adopt the notations:

G_i: is the probability for a level-i node to be gray;
int_i: is the expected number of non-leaf nodes at level i;
ext_i: is the expected number of leaf nodes at level i;

with the equality:

$$G_i = \frac{int_i}{int_i + ext_i}$$

we can express the above defined leaf nodes distribution by means of the following recurrence formula:

$$\frac{ext_i}{ext_{i-1}} = \frac{1}{i+1}$$

and being: $ext_{i-1} = 4int_i(1 - G_{i-1})$,

it follows that has to be: $\dfrac{ext_i}{4int_i(1 - G_{i-1})} = \dfrac{1}{i+1}$

that becomes:

$$\frac{ext_i + int_i}{4int_i(1 - G_{i-1})} - \frac{int_i}{4int_i(1 - G_{i-1})} = \frac{1}{i+1}$$

$$\Rightarrow \frac{1}{4G_i(1 - G_{i-1})} - \frac{1}{4(1 - G_{i-1})} = \frac{1}{i+1}$$

$$\Rightarrow \frac{1 - G_i}{4G_i(1 - G_{i-1})} = \frac{1}{i+1} \Rightarrow \frac{4G_i(1 - G_{i-1})}{1 - G_i} = i+1$$

$$\Rightarrow 4G_i(1 - G_{i-1}) = i+1-iG_i - G_i \Rightarrow iG_i + 5G_i - 4G_iG_{i-1} - 1 = i$$

from which: $G_i = \dfrac{i+1}{5+i-4G_{i-1}}$ with $G_0 = 0$

From this we have that such a model can be obtained as the following instance of the node based model:

$G_i = 1 - 2\beta_i$ that is $\beta_i = \dfrac{1 - G_i}{2}$

and then: $\beta_i = \dfrac{1 - \left(\dfrac{i+1}{5+i-4(1-2\beta_{i-1})}\right)}{2}$

that becomes: $\beta_i = \dfrac{4\beta_{i-1}}{8\beta_{i-1} + i + 1}$ with $\beta_0 = \dfrac{1}{2}$

This formula appears hard to be exploited with standard analytical techniques, but we can anyway compute the sequence of β_i for a sufficient length, so that a comparison with real data can be made; in the following we provide the first 16 values of the succession:

β_0=.5; β_1=.333; β_2=.235; β_3=.16; β_4=.101; β_5=.059; β_6=.032; β_7=.015; β_8=.006; β_9=.002; β_{10}=.0009; β_{11}=.0003; β_{12}=.0001; β_{13}=.00002; β_{14}=.000007; β_{15}=.000001.

From this sequence, we can derive the expected leaf size percentage on the first seven level of a 512x512 image, to be compared with the values in table 1. Because of the branching process we have that a node at level i exists as a leaf with probability:

$$L_i = 2\beta_i E_i$$

where E_i is the probability for a node to exist at level i. The probability E_i for a node to exist at level i depends by the probability for a node at level i+1 to exist and to be internal. This, pointed out that for a quadtree of height m is E_m=1, since the root always exists, and being:

$$E_i = G_{i+1} E_{i+1} = (1 - 2\beta_{i+1}) E_{i+1}$$

produces the following recurrence:

$$L_i = 2\beta_i \prod_{k=i+1}^{m} (1- 2\beta_k)$$

that can be calculated having fixed m. From this, the expected number of leaf nodes at level i is:

$$ext_i = 4^{m-i} L_i$$

Setting m=9, we obtain the following values:

Leaf Size	# leaves	% leaves
1×1	19,783	58.5
2×2	9,891	29.25
4×4	3,297	9.7
8×8	824	2.4
16×16	164	0.4
32×32	27	0.08
64×64	3.8	0.01
128×128	0.4	0.001
256×256	0.05	0.0001
512×512	0.005	0.00001
TOTAL	33,990	100

Table 2: Leaf distribution of the new model

and then almost perfectly in accord with values reported in table 1 for the various classes of spatial data. From the table is evident that the structure generated by our sequence is more homogeneous than the Samet's one, in the sense that the nodes distribution is more balanced on the different levels.

To complete the comparison between our nodes distribution and the Samet's one, we provide now a table containing the expected number of nodes for different resolution of images:

Image size	Samet's model	New model
2×2	3	2.33
4×4	9	5.94
8×8	28	17.16
16×16	90.6	55.65
32×32	303	196.97
64×64	1,039.86	738.47
128×128	3,640.50	2,863.30
256×256	12,945	11,298.56
512×512	46,603	44,950.87

Table 3: Expected number of internal and external nodes

where the expected number of nodes can be computed with the formula:

$$A_m = 1+ \sum_{i=1}^{m} 4^i (1 - 2\beta_m)(1 - 2\beta_{m-1}). . .(1 - 2\beta_{m-i+1})$$

that has been proved in [7]. From the table the differences between the models emerge, outlighting the structure more aggregated of our proposed model.

5. CONCLUSIONS

In this paper we have proposed and analyzed a new branching sequence to generate random quadtrees representing binary images. We have shown that this sequence produces expected distributions of external and internal nodes closer to real data than all previous proposed approaches in the literature to model both random binary images and quadtrees. The effectiveness of the new proposed model with respect to real spatial data images can have a large impact on the analysis of expected performances of a large class of algorithms for spatial data processing, as for the finding of neighbours or the computing of the mean perimeter of an image. First experimental results show that this new model closely simulate real cases.

REFERENCES

[1]: H. Samet, *The Design and Analysis of Spatial Data Structures*, Addison-Wesley, Reading, MA, 1990.

[2]: H. Samet, The Quadtree and Related Hierarchical Data Structures, *in Computing Surveys*, Vol. 16, No. 2, June 1984, pp. 187-260.

[3]: H. Samet, Computing Perimeters of Images Represented by Quadtrees, in *IEEE Trans. on Pattern Analysis & Machine Intelligence*, PAMI-3 (6) 1981, pp. 683-687.

[4]: H. Samet, *Applications of Spatial Data Structures: Computer Graphics, Image Processing and GIS*, Addison-Wesley, Reading, MA, 1990.

[5]: C. Puech and H. Yahia, Quadtrees, Octrees, Hyperoctrees: a Unified Analytical Approach to Tree Data Structures Used in Graphics, Geometric Modeling and Image Processing, in *Proc. of the Symposium on Computational Geometry*, Baltimore, Maryland, June 1985, pp.272-280.

[6]: E.Nardelli and G.Proietti, A Unifying Probabilistic Model for Quadtrees Representing Binary Images, Technical Report nº 368 September 1993 of the Institute for Systems Analysis and Informatics, C.N.R., Roma.

[7]: C. Mathie, C. Puech and H. Yahia, Average Efficiency of Data Structures for Binary Image Processing, in *Information Processing Letters*, 26 (1987/88), pp. 89-93.