A PROBABILISTIC ANALYSIS OF HIERARCHICAL GEOMETRIC DATA STRUCTURES*

Michael Lindenbaum
Electrical Engineering Department
Technion
32000 Haifa, ISRAEL

Hanan Samet
Computer Science Department**
University of Maryland
College Park, Maryland 20742 USA

Abstract

The size of five hierarchical geometric data structures is investigated analytically using random image models. Upper bounds on the size of the structures, as well as some lower bounds, are derived. The results are useful in predicting the storage required by the structures as well as the performance of algorithms that rely on them.

1. INTRODUCTION

Hierarchical geometric data structures employ a representational scheme which can be applied at different spatial resolutions. These structures are used for many different types of data including points, regions, lines, rectangles, surfaces, volumes, etc. The region quadtree [2,3] is one such representation for two-dimensional region data (e.g., an image). It enables the performance of many operations on images by substantially faster algorithms while maintaining a relatively compact representation [9,10].

For evaluating the space required for storing these structures and the performance of algorithms based on them, it is desired to predict their size. This is complicated by the fact that the structure of these representations depends on the data being represented. Traditional worst-case analysis is often inappropriate because the worst case tends to be both very bad and highly improbable. Thus most approaches to the analysis of hierarchical structures have been statistical in nature. Tamminen [13] considers the performance of quadtrees and binary trees under the assumption that the image consists of a single random line. Mathieu *et al.* [4] and Puech and Yahia [8] investigate the size of quadtrees and some other related questions using some assumptions on the branching probabilities of nodes in the tree. Nelson and Samet [5–7] consider the distributions of node occupancies in hierarchical geometric data structures which store a variable number of geometric data items per node. This approach is similar to hashing where each node acts like a bucket.

Although these approaches sometimes lead to remarkable agreement between theory and simulation [1,6], they have a common drawback. The explicit model of the image on which the statistical analysis is done is either exceedingly simple or it is not given at all and is just implied from other assumptions. Thus the connection between the analysis and the performance with real image data is not clear.

A nonstatistical approach was applied by Hunter and Steiglitz [2] to show that for a polygon of perimeter l, the size of the corresponding region quadtree is $O(l)$ (i.e., the number of nodes). This classic result, although depending on the data, has been observed to be sufficiently general to be used for predicting performance of a number of algorithms for different images represented by a region quadtree.

*Supported in part by NSF Grant IRI-88-02457.

**Also a member of the Center for Automation Research and the Institute for Advanced Computer Studies.

In the paper we investigate the use of a random image model consisting of M randomly drawn lines. We analyze five variants of the quadtree that can be built for data that obey this model, by determining the expected number of nodes in each of them. These variants are the region quadtree [2,3], the MX quadtree [2,9], the PM quadtree [11], the PMR quadtree [5–7], and a new representation suggested here which we call a modified PMR quadtree.

The rest of this paper is organized as follows. Section 2 gives a brief overview of quadtrees, including the definitions of the five variants that we analyze. Section 3 presents the random image model(s) and reviews some necessary results from geometric probability. Section 4 contains the statistical analysis and the result of its application to each of the afore-mentioned quadtree variants. We conclude in Section 5 with an interpretation of this analysis as well as a discussion of its application to a more general image model.

2. OVERVIEW OF QUADTREES

A quadtree is a hierarchical variable resolution data structure based on the recursive partitioning of the plane into quadrants. It can be viewed as a 4-ary tree where each node represents a region in the plane called a block, and the sons of each node represent a partition of that region into four parts. This scheme is useful for representing geometric data at a variable resolution. Quadtree variants exist for representing planar regions, collections of points, and collections of line segments, as well as more complicated objects (e.g., rectangles). Generalization of the principle to three and higher dimensions (e.g., octrees [9,10]) have also been investigated. They have many of the same basic properties.

The different variants of the quadtree can be subdivided into two categories: those based on a regular decomposition of space using pre-defined boundaries, and those where the partition is determined explicitly by the data as it is entered into the structure. For most applications, regular decomposition works at least as well as the data-based decomposition. Moreover, regular decomposition is easier to implement and analyze. In this paper, we consider only structures based on a regular decomposition. Another distinction is between quadtree variants whose maximal depth (say N) is bounded and those for whom it is not. From the structure that we discuss, only the PMR quadtree has an unbounded maximal depth.

The condition used to determine when a quadtree block should be partitioned is called a *splitting rule*. This rule is usually a function of the data that is associated with the block—i.e., the condition is evaluated using local information. The exact formulation of the rule depends on the type of the data being stored. We consider the following quadtree variants.

A region quadtree represents planar regions and its splitting rule is such that a block is split if both the depth of its corresponding node is smaller than N and if the region represented by the block is not homogeneous.

An MX quadtree represents a collection of line segments on the plane. Its splitting rule is such that a block is split if both the depth of its corresponding node is less than N and if the block contains at least one line.

A PM quadtree represents collections of line segments in the plane. Its splitting rule is such that a block is split unless the depth of the corresponding node is N, or only one line passes through the block, or all the lines that pass through the block meet at a point within the block. This is a variant of the PM_1 quadtree [11]. It differs from the PM_1 quadtree by virtue of having a bound on its depth.

A (generalized) PMR_q quadtree also represents a collection of line segments but is defined in a different way. It depends on a parameter q and is created as a dynamic result of a sequence of insertions of line segments using a splitting rule such that a block is split once if the block is both intersected by the new segment and already contains q or more segments. Note that this structure does not have a prespecified maximum depth. The resulting decomposition depends on the order in which the line segments are inserted. Moreover, the number of line segments represented by any leaf node is not guaranteed to be less than or equal to q.

We also define a new hierarchical structure called a modified PMR_q quadtree that resembles a PMR_q quadtree. It represents collections of line segments on the plane using a splitting rule such that a block is split if both the depth of its corresponding node is less than N and more than q segments pass through the block. This structure has a prespecified maximum depth. The resulting decomposition does not depend on the order in which the segments are inserted, and the number of segments stored in a leaf node whose depth is less than N is guaranteed to be less than or equal to q.

3. RANDOM IMAGE MODELS

The quadtree variants discussed in this paper represent geometric structures which are instances of a random process described as follows. Let us characterize a line by the two parameters ρ and θ. The line $L(\rho,\theta)$ consists of the points (x,y) satisfying the relation

$$L(\rho,\theta) = \{(x,y) \mid x\cos\theta + y\sin\theta = \rho\}.$$

In this case $L(\rho,\theta)$ is perpendicular to the line between the origin and the point (ρ,θ). Define R and T so that

$$R = \{(x,y) \mid \mid x \mid, \mid y \mid < 2^{N-1}\}$$

and

$$T = \{(\rho,\theta) \mid L(\rho,\theta) \cap R \neq empty\}.$$

It should be clear that T includes all the parameter pairs (ρ,θ) that represent lines that intersect the $2^N \times 2^N$ square grid.

Let us define a probability density function $p(\rho,\theta)$

$$p(\rho,\theta) = \begin{cases} \dfrac{1}{\mid T \mid} & (\rho,\theta) \in T \\ 0 & \text{otherwise} \end{cases} \qquad (1)$$

where

$$\mid T \mid = \int_T d\rho \; d\theta.$$

This distribution, called the uniform density distribution, is the only one which ensures that the probability of choosing a particular random line is independent of the coordinate system in which ρ and θ are defined (i.e., it is independent of the translation or rotation of the coordinate system [12]).

We use a random image model where each image consists of M random lines chosen independently using the probability distribution (1). The geometric structure created as an instance of the model consists of an $2^N \times 2^N$ image containing exactly M line segments whose endpoints intersect the boundary. From the continuous probability distribution it follows that no three lines can intersect in the same point as this is a zero probability event. This follows by observing that two intersecting lines define a point, say Q, which can be regarded as prespecified for the third line, say H, and the probability that a random line (i.e., H) passes through a given point (i.e., Q) is zero.

The above model is useful for predicting the sizes of all but the region quadtree which represents regions rather than lines. This leads us to the following variation of the above model, which we term a *modified* random image model. In this case, each image consists of black and white regions separated by M random lines chosen independently using the uniform probability distribution (1). The actual colors of the individual regions do not affect the total number of nodes. The resulting image can be interpreted as follows. Choose one of the regions at random and let it be black. Let all of its neighbors be white. Let all of the white node's uncolored neighbors be black. Repeat this process until all regions are colored.

Before starting the analysis, we first recall two results from geometric probability which form the basis of our results [12].

Geometric Probability Theorem 1 (GPT1): Let C_1 be a convex planar set included in the convex planar set $C \subset R$. Let L_1 and L be the perimeters of C_1 and C, respectively. Let l be a random line chosen using the uniform probability distribution (1). Then

$$p\{l \cap C_1 \neq \emptyset \mid l \cap C \neq \emptyset\} = \frac{L_1}{L}$$

Geometric Probability Theorem 2 (GPT2): Let $C \subset R$ be a convex planar set with area A and perimeter L. Let l be a random line chosen using the uniform probability distribution (1). Suppose that l intersects with C and creates a chord H with length $\mid H \mid$. Then

$$E[\mid H \mid] = \frac{\pi A}{L}.$$

4. STATISTICAL ANALYSIS OF QUADTREES

An image defined as in Section 3 is an instance of a random process. It follows that its hierarchical representation, using one of the quadtree variants, is also a random process. Moreover, the existence of a node in the tree, or its being a leaf, are random events. We make the following simplifying assumptions: given nodes u and v which exist in the tree, the events {node u is a leaf} and {node v is a leaf} are statistically independent. Assume further that the conditional probability

$$p\{\ v \textit{ has 4 sons} \mid v \textit{ exists in the tree}\ \}$$

depends only on d, the depth of v in the tree. Let P_d denote this probability. Although the statistical independence assumption does not necessarily hold for certain pairs of nodes (e.g., two nodes which are brothers), we claim that the assumption is reasonable for most pairs of nodes which correspond to small distant regions intersected by different lines.

Let S_d be the number of nodes at depth d, and let S be the total number of nodes in the tree. The expected number of nodes at each depth is given by

$$E[S_0] = S_0 = 1$$

$$E[S_1] = P_0 \cdot S_0 \cdot 4 = P_0 \cdot 4$$

and as implied from the statistical independence

$$E[S_2] = P_1 \cdot E[S_1] \cdot 4 = P_0 P_1 \cdot 4^2$$

$$E[S_3] = P_2 \cdot E[S_2] \cdot 4 = P_0 P_1 P_2 \cdot 4^3$$

$$\ldots$$

$$E[S_d] = P_{d-1} \cdot E[S_{d-1}] \cdot 4 = \left(\prod_{i=1}^{d-1} P_i \right) \cdot 4^d \qquad (2)$$

and

$$E[S] = \sum_{d=1}^{N} E[S_d] = \sum_{d=1}^{N} \left\{ \prod_{i=1}^{d-1} P_i \right\} \cdot 4^d \qquad (3)$$

Equation (3) which gives the expected number of nodes in the tree serves as the basis for our analysis. Below we focus on splitting rules for each of the quadtree variants discussed in Section 2. For each rule, we derive the corresponding probabilities P_d, and then use equation (3) to calculate the expected size of the data structure.

4.1. MX QUADTREE

An MX quadtree represents a collection of line segments on the plane. It partitions the plane into square blocks using the splitting rule such that a block is split if both the depth of its corresponding node is less than N and if the block contains at least one line. If the block does not contain a line, then it is not subdivided further and its corresponding node is a leaf. Otherwise, it is subdivided and its corresponding node has 4 sons.

4.1.1. ANALYSIS

In order to compute the probabilities P_d, we use the following argument. A node at depth d corresponds to a $2^{N-d} \times 2^{N-d}$ square. The geometric probability theorem GPT1 implies that a particular random line passes through this region with probability

$$p_d = \frac{4 \cdot 2^{N-d}}{4 \cdot 2^N} = (\frac{1}{2})^d.$$

The probability that exactly k out of M lines pass through this region is

$$p_{d,k} = \binom{M}{k} (\frac{1}{2})^{dk} \left(1-(\frac{1}{2})^d\right)^{M-k}. \tag{4}$$

The probability that one or more lines pass through this region is

$$P'_d = 1 - p_{d,0} = 1 - \left[1-(\frac{1}{2})^d\right]^M. \tag{5}$$

The existence of a node v at depth d in the tree implies that at least one line passes through the region corresponding to its father node. Thus, v exists in the tree with probability P'_{d-1} and

$$p\{\ v \text{ has 4 sons } \mid v \text{ exists }\} = \frac{p\{v \text{ has 4 sons}\}}{p\{v \text{ exists}\}} = \frac{P'_d}{P'_{d-1}} \tag{6}$$

Inserting (5) and (6) into (3) we get

$$\begin{aligned} E[S] &= \sum_{d=1}^{N} 4^d \cdot \prod_{i=1}^{d-1} P_i \\ &= \sum_{d=1}^{N} 4^d \cdot \frac{P'_1}{P'_0} \cdot \frac{P'_2}{P'_1} \cdot \frac{P'_3}{P'_2} \cdot \cdots \cdot \frac{P'_{d-1}}{P'_{d-2}} \\ &= \sum_{d=1}^{N} 4^d \cdot \frac{P'_{d-1}}{P'_0} \cdot = \sum_{d=1}^{N} 4^d \left[1 - \left[1-(\frac{1}{2})^{d-1}\right]^M\right] \end{aligned} \tag{7}$$

where we use the fact that apart from the trivial case that $M = 0$, $P'_0 = 1$.

Below we compute a bound on the number of nodes in the tree by first calculating it as a function of a parameter β defined by (8), so that it holds for every value of β. Next, we find the tightest bound by determining the value of β which minimizes the bound. We also use an additional parameter d_0 to facilitate this task. The two parameters β and d_0 are chosen to satisfy

$$M(\frac{1}{2})^{d_0} = \beta < 1 \tag{8}$$

β and d_0 will be adjusted later to obtain the tightest bound. Decompose $E[S]$ into two sums $\sum_1$ and $\sum_2$

$$\begin{aligned} E[S] &= \sum_{d=1}^{d_0} 4^d \left[1 - \left[1-(\frac{1}{2})^{d-1}\right]^M\right] \\ &+ \sum_{d=d_0+1}^{N} 4^d \left[1 - \left[1-(\frac{1}{2})^{d-1}\right]^M\right] = \sum_1 + \sum_2. \end{aligned} \tag{9}$$

$$\begin{aligned} \sum_1 &= \sum_{d=1}^{d_0} 4^d \left[1 - \left[1-(\frac{1}{2})^{d-1}\right]^M\right] \\ &\le \sum_{d=1}^{d_0} 4^d \cdot 1 \approx \frac{4^{d_0}}{1-\frac{1}{4}} \approx \frac{4}{3}\frac{M^2}{\beta^2} \end{aligned} \tag{10}$$

Taking the binomial expansion of $\sum_2$ we get

$$\begin{aligned} \sum_2 &= \sum_{d=d_0+1}^{N} 4^d \left[1 - \sum_{k=0}^{M} (\frac{1}{2})^{kd-k} \binom{M}{k} (-1)^k\right] \\ &= \sum_{d=d_0+1}^{N} \sum_{k=1}^{M} 2^{2d} (\frac{1}{2})^{kd-k} \binom{M}{k} (-1)^{k-1} \end{aligned} \tag{11}$$

changing the order of summation and separating the $k=1$ and $k=2$ cases, we get

$$\sum_2 = \sum_3 + \sum_4 + \sum_5 \tag{12}$$

$$\sum_3 = \sum_{d=d_0+1}^{N} 2^{2d} (\frac{1}{2})^{d-1} \binom{M}{1} = \sum_{d=d_0+1}^{N} 2^d \cdot 2 \cdot M \approx 4M \cdot 2^N \tag{13}$$

$$\begin{aligned} \sum_4 &= -\sum_{d=d_0+1}^{N} 2^{2d} (\frac{1}{2})^{2d-2} \binom{M}{2} \\ &= -\sum_{d=d_0+1}^{N} 4 \binom{M}{2} = -2M(M-1)(N-d_0) \end{aligned} \tag{14}$$

$$\begin{aligned} \sum_5 &= \sum_{d=d_0+1}^{N} \sum_{k=3}^{M} 2^k \binom{M}{k} (\frac{1}{2})^{(k-2)d} (-1)^{k-1} \\ &\le \sum_{k=3}^{M} \sum_{d=d_0+1}^{N} 2^k \binom{M}{k} (\frac{1}{2})^{(k-2)d} \\ &\approx \sum_{k=3}^{M} 2^k \binom{M}{k} \frac{(\frac{1}{2})^{d_0(k-2)} (\frac{1}{2})^{k-2}}{1-(\frac{1}{2})^{k-2}} \\ &= \sum_{k=3}^{M} 4 \binom{M}{k} \left(\frac{\beta}{M}\right)^{k-2} \frac{1}{1-(\frac{1}{2})^{k-2}} \\ &= \sum_{k=3}^{M} 4 \cdot \frac{M \cdot (M-1) \cdot (M-2) \cdot \cdots \cdot (M-k+1)}{k!} \frac{\beta^{k-2}}{M^{k-2}} \frac{1}{1-(\frac{1}{2})^{k-2}} \\ &\le \frac{4}{3} M^2 \frac{\beta}{1-\beta} \end{aligned} \tag{15}$$

Note that all approximations performed while calculating $\sum_1$, $\sum_3$, and $\sum_5$ can also serve as upper bounds since the approximated value is always larger than the real value. We continue by summing all contributions which are partly expected values and partly upper bounds for expected values, to get

$$\begin{aligned} E[S] &= \sum_1 + \sum_3 + \sum_4 + \sum_5 \\ &\le 4 \cdot M \cdot 2^N - 2 \cdot M^2 \cdot N + M^2 \left\{ \frac{4}{3} \frac{1}{\beta^2} + \frac{4}{3} \frac{\beta}{1-\beta} + 2\log_2 \frac{M}{\beta} \right\}. \end{aligned} \tag{16}$$

4.1.2. INTERPRETATION

The value of $E[S]$ given by (16) is an upper bound on the number of nodes in a MX quadtree. Now the value of β (and the corresponding value of d_0), may be chosen to minimize it. This method of calculation is motivated by the observation that for a large number M of lines and for a small depth $d < d_0$ of the tree, most nodes exist at this depth and may be counted.

The approximations in (10), (13) and (15) are good when $1 \ll d_0 \ll N$, but they hold as bounds for any chosen value of d_0 ($1 \le d_0 \le N$). The bound in (15) is not very tight and better bounds that depend on specific values of M should probably be used. However, we have chosen to use our approximation as it has an intuitive physical basis.

Asymptotically, the dominant contribution to the number of nodes comes from the first term in (16) which may be transformed into a more familiar form using Theorem GPT2. The expected total length of all lines in our geometric structure is

$$E[l] = \sum_{i=1}^{M} E[l_i] = M \cdot \pi \frac{(2^N)^2}{4 \cdot 2^N} = \frac{\pi}{4} \cdot M \cdot 2^N \qquad (17)$$

Substituting (17) into the first term of (16) we get

$$E[S] \approx \frac{16}{\pi} E[l]; \qquad (18)$$

In other words, the expected number of nodes in an MX quadtree is proportional to the length of the curve, a result already derived under different models [2].

4.1.3. AN APPROXIMATED LOWER BOUND

The derivation of $E[S]$ given by (9)-(16) can serve as a basis of an approximation of a lower bound on the expected number of nodes. $E[S]$ consists of the contributions of $\sum_1$, $\sum_3$, $\sum_4$, and $\sum_5$. $\sum_4$ is an exact value and $\sum_3$ is a good approximation for $M \ll 2^N$. (Note that when $M \approx 2^N$, $d_0 \approx N$ and $\sum_3 \approx 0$.) $\sum_1$ must be positive and $\sum_5$ can be easily bounded from below by $-\frac{4}{3} M^2 \frac{\beta}{1-\beta}$. Thus, $\sum_3 + \sum_4 - \frac{4}{3} M^2 \frac{\beta}{1-\beta}$. is an approximation of a lower bound on $E[S]$.

Furthermore, for large N, $\sum_1$, $\sum_4$, and $\sum_5$ are small with respect to $\sum_3$. Thus the difference between the upper bound and the approximation of the lower bound is small, and hence each of these two bounds is itself a good approximation of $E[S]$.

4.2. REGION QUADTREE

A region quadtree represents regions in the plane. It partitions the plane into square blocks using the splitting rule such that a block is split if both the depth of its corresponding node is smaller than N and if the region represented by the block is not homogeneous. If the block is homogeneous, then it is not subdivided further and its corresponding node is a leaf. Otherwise, it is subdivided and its corresponding node has 4 sons.

Assuming the modified image model described in Section 2, the expected number of nodes in a region quadtree follows directly from the derivation in Section 4.1. Consider an MX quadtree corresponding to an image which is an instance of the basic random image model (and thus contains M line segments). Consider also a region quadtree representing a second image which consists of regions separated by the line segments of the first image. It is clear that both trees have the same structure (with the only difference being the contents of each of the nodes). It follows that the statistical properties of a region quadtree corresponding to an instance of the modified random image model are exactly the same as the properties of an MX quadtree corresponding to an instance of the basic random image model. Hence, the expected number of nodes in a region quadtree is bounded from above by (16).

4.3. PM QUADTREE

A PM quadtree represents a collection of line segments in the plane. It partitions the plane into square blocks using the splitting rule such that a block is split unless the depth of the corresponding node is N, or only one line passes through the block, or all the lines that pass through the block meet at a point within the block. If the block contains a single line or all the lines pass through a common point in the block, then it is not subdivided further and its corresponding node is a leaf. Otherwise, it is subdivided and its corresponding node has 4 sons. According to our model of a random image, the probability that three or more lines intersect at a point is zero and hence this case may be neglected.

4.3.1. ANALYSIS

Let us define α as the probability that two lines intersect inside a square region, say Q, given that each of these lines passes through Q. For a square $2^{N-d} \times 2^{N-d}$ region ($0 \le d \le N$), the probability that the splitting conditions of a PM quadtree are satisfied may be written as

$$P'_d = 1 - p_{d,0} - p_{d,1} - \alpha \cdot p_{d,2} \qquad (19)$$

where $p_{d,k}$ is the probability that exactly k of M lines pass through a square region of side 2^{N-d}.

$$P'_d = 1 - \binom{M}{0}\left[1-(\tfrac{1}{2})^d\right]^M - \binom{M}{1}\left(\tfrac{1}{2}\right)^d\left[1-(\tfrac{1}{2})^d\right]^{M-1} - \alpha\binom{M}{2}\left(\tfrac{1}{2}\right)^{2d}\left[1-(\tfrac{1}{2})^d\right]^{M-2} \qquad (20)$$

Inserting (20) in (3) and using a derivation similar to that used to obtain (7), we get

$$E[S] = \sum_{d=1}^{N} 4^d \cdot \frac{P'_{d-1}}{P'_0} \qquad (21)$$

Once again we assume that P'_0 is 1. Now, let us once again choose a depth d_0 and a constant β such that

$$M(\tfrac{1}{2})^{d_0} = \beta < 1. \qquad (22)$$

Once again, β will be adjusted later to obtain the tightest bound. We now decompose $E[S]$ into two sums $\sum_1$ and $\sum_2$

$$E[S] = \sum_{d=1}^{d_0} 4^d \, P'_{d-1} + \sum_{d=d_0+1}^{N} 4^d \, P'_{d-1} = \sum_1 + \sum_2. \qquad (23)$$

A bound on $\sum_1$ is obtained as follows.

$$\sum_1 = \sum_{d=1}^{d_0} P'_{d-1} \, 4^d \le \sum_{d=1}^{d_0} 4^d \approx \frac{4}{3} \, 4^{d_0} = \frac{4}{3} \frac{M^2}{\beta^2} \qquad (24)$$

$\sum_2$ is evaluated by taking its binomial expansion to get a sum of the powers of $(\frac{1}{2})^d$ (i.e., $(\frac{1}{2})^0$, $(\frac{1}{2})^d$, $(\frac{1}{2})^{2d}$, ...). The $(\frac{1}{2})^0$ and $(\frac{1}{2})^d$ terms cancel out and we get

$$\sum_2 = \sum_{d=d_0+1}^{N} \sum_{k=2}^{M} C_k (\tfrac{1}{2})^{kd-k} \cdot 4^d, \qquad (25)$$

where

$$C_k = (-1)^{k-1}\left[\binom{M}{k} - \binom{M}{1}\cdot\binom{M-1}{k-1} + \binom{M}{2}\cdot\binom{M-2}{k-2}\alpha\right].$$

Changing the order of summation and separating the $k = 2$ case we get

$$\sum_2 = \sum_3 + \sum_4 \qquad (26)$$

$$\sum_3 = \sum_{d=d_0+1}^{N} C_2 (\tfrac{1}{2})^{2d-2} 4^d = 2(1-\alpha)M(M-1)(N-d_0) \qquad (27)$$

$$\sum_4 = \sum_{k=3}^{M} C_k 2^k \sum_{d=d_0+1}^{N} (\tfrac{1}{2})^{d(k-2)} \approx \sum_{k=3}^{M} C_k 2^k \frac{(\frac{1}{2})^{(k-2)(d_0+1)}}{1-(\frac{1}{2})^{k-2}} = 4\sum_{k=3}^{M} \frac{C_k}{1-(\frac{1}{2})^{k-2}} \left(\frac{\beta}{M}\right)^{k-2} \qquad (28)$$

Now, let us examine the coefficients C_k.

$$C_k = (-1)^{k-1}\left[\binom{M}{k} - \binom{M}{1}\cdot\binom{M-1}{k-1} + \binom{M}{2}\cdot\binom{M-2}{k-2}\alpha\right] = (-1)^{k-1}\left[\binom{M}{k} - \binom{M}{1}\cdot\frac{k}{M}\cdot\binom{M}{k} + \binom{M}{2}\cdot\frac{k(k-1)}{M(M-1)}\cdot\binom{M}{k}\alpha\right] = (-1)^{k-1}\binom{M}{k}\left[1-k+\frac{k(k-1)}{2}\alpha\right] \qquad (29)$$

It can be shown that

$$-\frac{1}{3}M^k \le C_k \le \frac{1}{6}M^k \tag{30}$$

Inserting (30) into (28) and accounting for the worst cases leads to

$$-\frac{8}{3}M^2\frac{\beta}{1-\beta} \le \sum_4 \le \frac{4}{3}M^2\frac{\beta}{1-\beta} \tag{31}$$

These bounds, which do not depend on any assumptions and hold for any value of α, demonstrate that the contribution of $\sum_4$ to $E[S]$ is $O(M^2)$.

In order to achieve a tighter bound on $\sum_4$, α is approximated as follows. Suppose that one line creates a chord of length l, say H_1, when intersected with a $2^{N-d} \times 2^{N-d}$ square. From GPT1 we have that a second random line, say H_2, chosen using the probability distribution given by (1) intersects with chord H_1 with probability

$$p\{H_2 \textit{ intersects } H_1\} = \frac{2l}{4\cdot 2^{N-d}}$$

Note that we assumed that the line H_1 is a convex set of zero width but perimeter $2\cdot l$. In order to compute this probability we should integrate over all possible chord lengths for H_1. Instead, we use an approximation by assuming that l equals its expected value which is given by GPT2 as follows.

$$l = \bar{l} = \frac{\pi A}{L} = \pi\frac{(2^{N-d})^2}{4\cdot 2^{N-d}}.$$

Under this assumption we get:

$$\alpha = \frac{2}{4\cdot 2^{N-d}}\cdot\frac{\pi\cdot(2^{N-d})^2}{4\cdot 2^{N-d}} = \frac{\pi}{8}. \tag{32}$$

Inserting this value of α in (29) yields

$$-0.13\ M^k \le C_k \le 0.026\ M^k$$

$$-1.04\ M^2\frac{\beta}{1-\beta} \le \sum_4 \le 0.21\ M^2\frac{\beta}{1-\beta}. \tag{33}$$

We checked $\sum_4$ for several values of M and found that only the $k=3$ term was significant. Furthermore, we also found that the total contribution of $\sum_4$ to $E[S]$ is negative. Collecting the contributions of $\sum_1$, $\sum_3$, and $\sum_4$ we have

$$E[S] = \sum_1 + \sum_3 + \sum_4$$
$$\le 2(1-\frac{\pi}{8})M(M-1)N + M^2\left[\frac{4}{3}\frac{1}{\beta^2} + 0.21\frac{\beta}{1-\beta} - 2(1-\frac{\pi}{8})\log_2\frac{M}{\beta}\right] \tag{34}$$

β may be chosen to minimize (34) but it must obey (22). Note that the first term in (34) is exact (it corresponds to $\sum_3$) while the second is only a bound which is not very tight. Even this nontight bound (i.e., the second term in (34)) makes a negative contribution to the total number of nodes when the number of lines exceeds some modest threshold (e.g., for $M = 16, 32, \cdots$ and $\beta = 0.5$).

4.3.2. INTERPRETATION

The number of possible line pairs in the image is $\binom{M}{2}$. Multiplying it by α yields the expected number of intersections. Approximating α by $\frac{\pi}{8}$=0.5 leads to approximately $\frac{M^2}{4}$ vertices (line intersections) in the whole image. By only considering the dominant first term in (34), which is roughly proportional to M^2, the model may be interpreted as predicting that the subdivision stops at the maximal depth N for approximately one path in the tree in the neighborhood of each vertex.

4.3.3. AN APPROXIMATED LOWER BOUND

The expected value of $E[S]$ consists of three terms $\sum_1$, $\sum_3$ and $\sum_4$. $\sum_3$ is an exact value, $\sum_1$ is positive, and $\sum_4$ is bounded from below by $-1.04\frac{\beta}{1-\beta}M^2$. Hence, $\sum_3 - 1.04\frac{\beta}{1-\beta}M^2$ is a lower bound for $E[S]$. Note that attempting to improve the bound by choosing a small value of β would fail as this requires that d_0 have a higher value which means that $\sum_3$ has a lower value. Only for a very large value of N are $\sum_1$ and $\sum_4$ negligible in comparison with $\sum_3$. This is because $\sum_1$ and $\sum_4$ are independent of N while $\sum_3$ depends on N. In this case, the difference between the upper bound and the approximation of the lower bound is small and hence each of these two bounds is itself a good approximation of $E[S]$.

4.4. MODIFIED PMR QUADTREE

A modified PMR_q quadtree represents a collection of line segments in the plane. It partitions the plane into square blocks using the splitting rule such that a block is split if both the depth of its corresponding node is less than N and more than q segments pass through the block. If the block contains q or less line segments, then it is not subdivided further and its corresponding node is a leaf. Otherwise, it is subdivided and its corresponding node has 4 sons.

We start our analysis by discussing a modified PMR_q quadtree with $q=2$. The probability that the splitting conditions are satisfied may be written as

$$P'_d = 1 - p_{d,0} - p_{d,1} - p_{d,2}. \tag{35}$$

Using the same techniques as in Section 4.3, we define β and d_0 as in (22) and decompose the sum corresponding to $E[S]$ into three sums $\sum_1$, $\sum_3$, and $\sum_4$

$$E[S] = \sum_1 + \sum_3 + \sum_4. \tag{36}$$

$\sum_3$ vanishes in this case (since $\alpha = 1$), the bounds given by (31) on $\sum_4$ hold, and thus

$$E[S] \le \frac{4}{3}M^2\left(\frac{1}{\beta^2} + \frac{\beta}{1-\beta}\right). \tag{37}$$

For example, for $M = 4, 8, 16, \cdots$ let $\beta = 0.5$ and we have

$$E[S] \le \frac{20}{3}M^2 \qquad (q=2). \tag{38}$$

The bound given by (38) means that for a PMR_q quadtree the number of nodes is proportional to the number of vertices, and does not depend on the maximal depth N. Therefore, almost everywhere, the subdivision stops before the maximal depth.

For $q > 2$ the analysis is a little more complicated but the results are essentially the same. The probability that the splitting conditions are satisfied may be written as

$$P'_d = 1 - p_{d,0} - p_{d,1} - p_{d,2} - \cdots - p_{d,q} \tag{39}$$

Inserting (39) in (3) and decomposing the sum as before using the parameters β and d_0, we may obtain a slightly lower bound for $q > 2$. For example, if $q = 3$, then it is possible to show that

$$C_k < \frac{1}{24}M^k$$

$$\sum_2 \le \frac{1}{3}M^2\frac{\beta}{1-\beta}$$

and for $\beta = 0.5$

$$E[S] \le \frac{1}{3}\left[\frac{4}{\beta^2} + \frac{\beta^2}{1-\beta}\right]M^2 = \frac{49}{9}M^2 \qquad (q=3). \tag{40}$$

It is not possible to reduce this bound by much (even for higher values of q) since the first term, $\sum_1$, remains $\frac{4}{3}\frac{M^2}{\beta^2}$.

4.5. PMR QUADTREE

A PMR_q quadtree represents a collection of line segments in the plane. It depends on a parameter q and is created as a dynamic result of a sequence of insertion of line segments using the splitting rule such that a block is split if the block is both intersected by the new segment and already contains q or more segments. The block is subdivided at most once when a new segment which intersects it is entered into the structure. Clearly, this may not be enough to ensure that the number of lines represented by each leaf is q or less. Since the maximal depth is not known in advance, a rule that splits each block until no more than q lines intersect with it may lead to an infinite tree. For example, suppose that more than q line segments intersect at a given point. The PMR quadtree splitting rule prevents

such complications. Note that for our model of a random image this problem cannot arise as the probability that three or more lines intersect at a point is zero.

Consider the following image generation process. First, create a PMR quadtree (with some q) which represents a collection of line segments using our random image model. Next, recursively subdivide every block which contains more than q line segments until every block contains at most q line segments. This process results in a different representation which must be finite since it is impossible for $q+1$ lines to intersect at the same vertex in our model. This representation is the modified PMR quadtree discussed in Section 4.4. Hence the modified PMR quadtree corresponding to an instance of the random image includes all the nodes in a PMR quadtree corresponding to the same image. It follows that the modified PMR quadtree always contains more nodes than the corresponding PMR quadtree, and thus all upper bounds derived from the expected number of nodes in a modified PMR quadtree (37–39) also hold for the PMR quadtree. Note that these bounds were independent of N, the maximal depth of the modified PMR quadtree.

5. DISCUSSION

A number of hierarchical geometric data structure were investigated using random image models. Four structures which represent collections of line segments and one which represents planar regions were presented, and an appropriate model was developed for each of them. Upper bounds on $E[S]$, the expected number of nodes, were found for each of the representations. Lower bounds and approximations were also derived for some of the cases. Given M line segments and a maximum depth N (where it is part of the definition of the representative), these bounds lead to the following asymptotic results on the expected number of nodes:

$$\begin{array}{rl} \textit{Region quadtree} & E[S] = O(M \cdot 2^N) \\ \textit{MX quadtree} & E[S] = O(M \cdot 2^N) \\ \textit{PM quadtree} & E[S] = O(M^2 \cdot N) \\ \textit{PMR quadtree} & E[S] = O(M^2) \\ \textit{Modified PMR quadtree} & E[S] = O(M^2) \end{array} \tag{41}$$

The main conclusions that can be drawn from these results are as follows. For MX (and region) quadtrees, the number of nodes is proportional to the total length of the line segments (or the region boundaries). This conclusion confirms a similar result obtained by Hunter and Steiglitz [2]. For PM quadtrees, the number of nodes is proportional to the product of the number of intersections between the line segments and the maximal depth of the tree. It also appears that in the neighborhood of most intersection points the subdivision stops only at the maximal depth (i.e., N). For PMR quadtrees and modified PMR quadtrees, the number of nodes is proportional to the number of intersection points between the line segments. It also appears that for most intersection points, the subdivision stops before the maximal depth.

For particular values of N and M, we believe that it is prudent to use the exact bounds as given in (16), (34) and (37), which depend on a parameter β. It is worth emphasizing that the values d_0 and β are not a part of the random model and are just parameters used to facilitate our calculations. In order to apply these bounds, it is required to choose a value of β which minimizes them while satisfying relation (8) (with d_0 being an integer). A useful practical approximation can be made by just letting the value $\beta = 0.5$. Of course, from a theoretical standpoint this is not realistic because it usually implies a non-integer value for d_0.

The bounds contain terms which are negative. They reduce the bounds and make them tighter. The negative terms compensate for nodes which are counted twice in other (positive) terms. For example, nodes in the PM quadtree at a depth less than d_0 are accounted for by the $\sum_1 = \frac{4}{3}\frac{M^2}{\beta^2}$ term and also by the $M^2 \cdot N$ term. The negative term $-M^2 \log_2 \frac{M}{\beta} \approx -M^2 d_0$ compensates for this situation.

These bounds hold for all values of M and N. However, they become meaningless when $M \geq 2^N$. In this case, the parameter d_0 approaches the value of the maximal depth N and $M^2 \approx (2^N)^2$ is the maximal number of nodes in the complete tree (i.e., all leaf nodes are at the maximal depth).

The bounds in this paper were computed under the assumption of a particular image model. We conjecture that the results apply also to more general images using $\pi \cdot M \cdot 2^N/4$ for $E[l]$ and $M^2/4$ for v where v is the number of vertices (i.e., endpoints of line segments and intersection points) and $E[l]$ is the expected total length of the line segments (or the boundaries of the regions). These values are discussed in the derivation of (18) in Section 4.1.2 and in Section 4.3.2. Using these values with $\beta = 0.5$ yields the following bounds on the expected number of nodes:

$$\begin{array}{ll} \textit{Region quadtree} & E[S] \leq \frac{16}{\pi} E[l] - 8 \cdot V \cdot N + 4 \cdot V \cdot (10.6 + \log_2 V) \\ \textit{MX quadtree} & E[S] \leq \frac{16}{\pi} E[l] - 8 \cdot V \cdot N + 4 \cdot V \cdot (10.6 + \log_2 V) \\ \textit{PM quadtree} & E[S] \leq 4 \cdot V \cdot N + 4 \cdot V \cdot (3.5 - 0.5 \log_2 V) \\ \textit{PMR quadtree} & E[S] \leq 24 \cdot V \\ \textit{Modified PMR quadtree} & E[S] \leq 24 \cdot V \end{array} \tag{42}$$

REFERENCES

[1] C. H. Ang, Applications and analysis of hierarchical data structures, Computer Science TR-2255, University of Maryland, College Park, MD, June 1989. 1 (1974), 1–9.

[2] G. M. Hunter and K. Steiglitz, Operations on images using quad trees, *IEEE Transactions on Pattern Analysis and Machine Intelligence 1*, 2 (April 1979), 145–153.

[3] A. Klinger, Patterns and search statistics, in *Optimizing Methods in Statistics*, J. S. Rustagi, ed., Academic Press, New York, 1971, 303–337.

[4] C. Mathieu, C. Puech, and H. Yahia, Average efficiency of data structures for binary image processing, *Information Processing Letters 26*, 2 (October 1987), 89–93.

[5] R. C. Nelson and H. Samet, A consistent hierarchical representation for vector data, *Computer Graphics 20*, 4 (August 1986), 197–206 (also *Proceedings of the SIGGRAPH '86 Conference*, Dallas, August 1986).

[6] R. C. Nelson and H. Samet, A population analysis of quadtrees with variable node size, Computer Science TR-1740, University of Maryland, College Park, MD, December 1986.

[7] R. C. Nelson and H. Samet, A population analysis for hierarchical data structures, *Proceedings of the SIGMOD Conference*, San Francisco, May 1987, 270–277.

[8] C. Puech and H. Yahia, Quadtrees, octrees, hyperoctrees: a unified analytical approach to three data structures used in graphics, geometric modeling and image processing, *Proceedings of the Symposium on Computational Geometry*, Baltimore, June 1985, 272–280.

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

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

[11] H. Samet and R. E. Webber, Storing a collection of polygons using quadtrees, *ACM Transactions on Graphics 4*, 3(July 1985), 182–222.

[12] L. A. Santalo, Integral geometry and geometric probability, in *Encyclopedia of Mathematics and its Applications*, G. C. Rota, ed., Addison-Wesley, Reading, MA, 1976.

[13] M. Tamminen, Encoding pixel trees, *Computer Vision, Graphics, and Image Processing 28*, 1(October 1984), 44–57.