Modified Fractal Signature (MFS): A New Approach to Document Analysis for Automatic Knowledge Acquisition

Yuan Y. Tang, *Senior Member, IEEE*, Hong Ma, Dihua Xi,
Xiaogang Mao, Ching Y. Suen, *Fellow, IEEE*

Abstract—One of the key technologies related to knowledge and data engineering is the acquisition of knowledge and data in the development and utilization of information system and the strategies to capture new knowledge and data. Actually, millions of documents, including technical reports, government files, newspapers, books, magazines, letters, bank checks, etc., have to be processed every day, and knowledge has to be acquired from them. This paper presents a new approach to document analysis for automatic knowledge acquisition. The traditional approaches have two major disadvantages: (1) They are not effective for processing documents with high geometrical complexity. Specially, the top-down approach can process only the simple documents which have specific format or contain some a priori information. (2) The top-down approach needs to split large components into small ones iteratively, while the bottom-up approach needs to merge small components into large ones iteratively. They are time consuming. This new approach is based on modified fractal signature. It can overcome the above weaknesses.

Index Terms—Automatic knowledge acquisition, document analysis, modified fractal signature, Minkowski dimension, δ-parallel bodies, blanket method.

1 Introduction

As mentioned in the first issue of the *IEEE Transactions on Knowledge and Data Engineering*, one of the key technologies related to knowledge and data engineering is the acquisition of knowledge and data in the development and utilization of information systems and the strategies to capture new knowledge and data [4]. One of the most challenging topics is automatic acquisition of new data and knowledge [11]. In practice, much knowledge has to be acquired from documents, including technical reports, government files, newspapers, books, journals, magazines, letters, and bank cheques, to name a few. The acquisition of knowledge from such documents can involve an extensive amount of handcrafting on the part of the engineer. Such handcrafting is time consuming and limits the application of the information systems. Actually, it is a bottleneck of information systems. Thus, automatic knowledge acquisition from documents has become an important subject. A hopeful solution to this problem is to use a new technique—*document processing*, which belongs to a branch of artificial intelligence. It makes sense to acquire knowledge directly by analyzing and understanding documents. In the first issue in 1994 of the *IEEE Transactions on Knowledge and Data Engineering*, we introduced a survey on the techniques and problems involved in the automatic knowledge acquisition through the document processing [12].

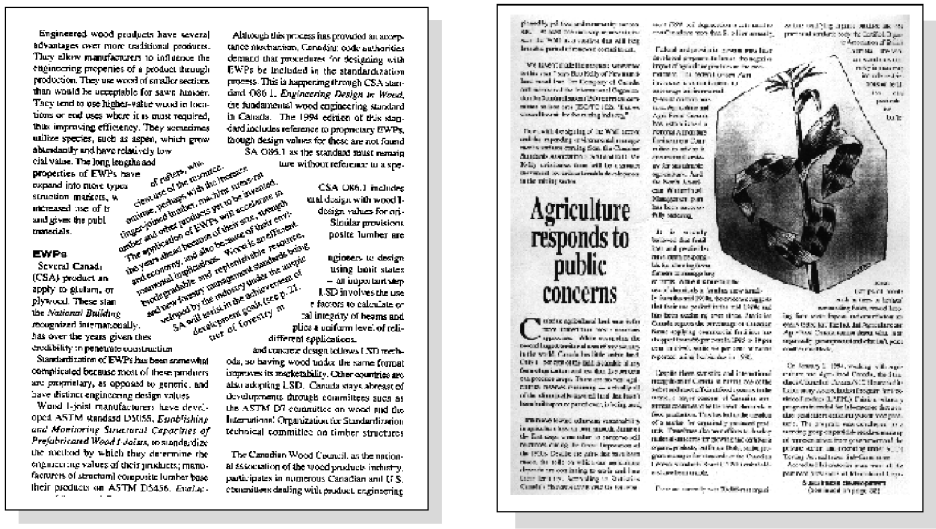
A document is considered to have two structures: geometric structure and logical structure. They play a key role in the process of the knowledge acquisition, which can be viewed as a process of acquiring the above structures. Extracting the geometric structure from a document refers to document analysis. Traditionally, two approaches have been used in document analysis, namely, top-down and bottom-up approaches [12]. Both approaches have their weaknesses:

- They are not effective for processing documents with high geometrical complexity. Specifically, the top-down approach can process only the simple documents which have specific format or contain some a priori information. It fails to process the documents which have complicated geometric structures, as shown in Fig. 1.
- To extract the geometric (layout) structure of a document, the top-down approach needs iterative operations to break the document into several blocks, while the bottom-up approach needs to merge small components into large ones, iteratively. Consequently, both approaches are time consuming.

This paper will present a new approach based on the *Modified Fractal Signature* (*MFS*) for document analysis. It does not need iterative breaking or merging, and can divide a document into blocks in only one step. It is anticipated

- *Y.Y. Tang is with the Department of Computing Studies, Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong. E-mail: yytang@comp.hkbu.edu.hk.*
- *H. Ma, D. Xi, and X. Mao may be contacted through the Department of Mathematics and the Department of Computer Science, Sichuan University, Chengdu, Sichuan 610064, Peoples Republic of China.*
- *C.Y. Suen is with the Center for Pattern Recognition and Machine Intelligence, GM-606, Concordia University, 1455 de Maisonneuve Blvd. West, Montreal, Quebec H3G 1M8, Canada.*

Manuscript received 31 Jan. 1996.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number 104036.



(a) (b)

(c)

Fig. 1. Examples of the high-complexity documents.

that this approach can be widely used to process various types of documents, including some with high geometrical complexity.

Fractals are mathematical sets with a high degree of geometrical complexity which can model many classes of time-series data as well as images. The fractal dimension is an important characteristic of fractals because it contains information about their geometric structure. Since Mandelbrot [5] proposed this technique, the subject of fractal dimension has drawn a great deal of attention from mathematicians, physicists, chemists, biologists, geologists, and electrical and computer engineers in various disciplines. Specifically, in the area of image processing, the fractal dimension has been used for image compression, texture analysis, image encoding, etc., providing a novel technique for achieving compression ratios of 10,000 to 1—or even higher [1]. Earlier results on texture analysis using fractal techniques were reported by Nguyen and Quinqueton [8] in which a one-dimensional fractal analysis along a space filling curve was used. A full two-dimensional analysis was performed by Pentland [10], and the statistics of differences of gray level between pairs of pixels at varying distances were used as indicators of the fractal properties of the texture. Maragos and Sun [6] developed a theoretical approach for measuring the fractal dimension of arbitrary continuous-time signals by using morphological erosion and dilation function operations to create covers around a signal's graph at multiple scales.

Although the fractal dimension has such wide applications, it has not touched the document processing area. This paper aims at exploring the fractal dimension in document analysis for automatic knowledge acquisition.

In the beginning of this paper, the most basic concepts of fractals and several types of fractal dimensions related to our work will be introduced. Precisely, definitions of fractals from different views will be presented in Section 2. Actually, there exist a variety of fractal dimensions, the most important one being the Hausdorff dimension, which is based on a mathematical tool—measure theory—which makes analysis easy, and is suitable for any sets. In Section 2, both the measure theory and the Hausdorff dimension will be described. The mathematical theory of the Hausdorff dimension is complete and exact, however, it is difficult to implement. In practice, the box computing dimension is convenient to apply. Therefore, it will also be discussed in Section 2. To facilitate the application of the box dimension to digital images, the Minkowski dimension will be introduced in this section. The relationship between the box

dimension and Minkowski dimension will also be supported by some theorems in Section 2.

The overall sequence of Section 2 can be illustrated as follows:

- Basic Concepts of Fractal Dimensions
  - Definitions of Fractal
  - Hausdorff Dimension
    - Hausdorff Measure
    - Hausdorff Dimension
  - Box Computing Dimension
    - Dimensions
    - Box Computing Dimension
  - Minkowski Dimension

Section 3 forms the core of this paper. It contains a description of the new approach to document analysis based on the modified fractal signature. The basic idea of this approach is that a document image can be mapped onto a gray-level function. Furthermore, this function can be mapped onto a surface to approximate its fractal dimension. In this section, computing fractal dimension by measuring the surface area will be presented first. After that, a new method for document analysis using the fractal features will be presented, followed by an algorithm of computing the fractal signature.

The overall sequence of Section 3 can be summarized below:

- Approximation of Fractal Dimension in Document Images
  - δ-Parallel Bodies
  - Blanket Technique and Modified Fractal Signature
  - Modified Fractal Signature Approach to Document Analysis
  - Algorithm

Experiments have been conducted to prove the effectiveness of this new approach in document processing, they will be reported in Section 4. Finally, conclusions will be given in Section 5.

2 Basic Concepts of Fractal Dimensions

In this section, the basic concepts of *fractals* will be introduced, and three types of *fractal dimensions*, namely, Hausdorff dimension, box-computing dimension, and Minkowski dimension, which are related to our work will also be discussed.

2.1 Definitions of Fractals

Before discussing the definitions of fractals, we will first introduce a mathematical terminology named *topological dimension*:

DEFINITION 1. *A collection $\Re$ of subsets of the space X is said to have order $m + 1$ if some point of X lies in $m + 1$ elements of $\Re$, and no point of X lies in more than $m + 1$ elements of $\Re$. Given a collection $\Re$ of subsets of X, a collection $\Im$ is said to* refine *$\Re$, or to be a* refinement *of $\Re$, if each element B of $\Im$ is contained in at least one element of $\Re$.*

Now, we define what we mean by the *topological dimension* of a space X.

DEFINITION 2. *A space X is said to be* finite-dimensional *if there is some integer m such that for every open covering $\Re$ of X, there is an open covering $\Im$ of X that refines $\Re$ and has order at most $m + 1$. The* topological dimension *of X is defined to be the smallest value of m for which this statement holds. We use notation $\dim_T X$ to represent the topological dimension of X.*

The topological dimension is a complicated and advanced mathematical topic, more details about it can be found in [7].

What are fractals? There are many definitions, because it is very difficult to define fractal strictly. Mandelbrot gave two definitions in 1982 and 1986, respectively.

1) The first definition from his original essay (1982) says:

DEFINITION 3. *A set F is called fractal set if its* Hausdorff dimension ($\dim_H F$) *is greater than the* Topological dimension ($\dim_T F$), *namely*:

$$\dim_H F > \dim_T F$$

2) In 1986, Mandelbrot defined the fractal as:

DEFINITION 4. *Fractal is a compound object, which contains several subobjects. The global characteristic of this object is similar to the local characteristics of each subobject.*

3) A more precise definition of the fractal set F can be provided below:

DEFINITION 5. *A set F is called fractal set if the following conditions are satisfied:*

- *The global characteristic of the set F is self-similar to the local characteristics of each subset, namely*:

$$\Im(F) \sim \Im(f_i), \qquad f_i \supset F,$$

- *where $\Im(.)$ stands for the characteristic of (.).*
- *The set F is infinitely separable, i.e.*

$$F = \left\{f_1^1,\ f_2^1,\ \ldots,\ f_i^1,\ \ldots,\ f_n^1\right\},$$

$$f_i^1 = \left\{f_1^2,\ f_2^2,\ \ldots,\ f_k^2,\ \ldots,\ f_n^2\right\},$$

$$\cdots\cdots\cdots$$

$$f_k^m = \left\{f_1^{m+1},\ f_2^{m+1},\ \ldots,\ f_k^{m+1},\ \ldots,\ f_n^{m+1}\right\},\ m+1 \to \infty.$$

- *Usually, the fractal dimension of the set F is a fraction, and greater than the* Topological dimension $\dim_T F$, *namely*:

$$\dim_H F > \dim_T F$$

- *In many cases the definition of F is recursive.*

2.2 Hausdorff Dimension

There exists a variety of fractal dimensions, the most important one being the *Hausdorff dimension*. Because it is based on a mathematical tool—*measure theory*—it is suitable for any sets and makes analysis of them easy.

2.2.1 Hausdorff Measure

Let U be a nonempty subset of n-dimensional Euclid space $\mathbb{R}^n$, and the diameter of U is defined as

$$|U| = sup\{|x - y| : x, y \in U\},$$

where *sup*{.} stands for the supremum of {.}. Thus, the diameter of U is the greatest distance apart of any pair of points in U. If $\{U_i\}$ is a countable collection of sets of diameter at most δ that cover F, such that

$$\Re(\delta) = \{U_i\} = \{U_i : i = 1, 2, ...\},$$

and

$$F \subset \bigcup_{i=1}^{\infty} U_i, \qquad 0 < |U_i| \le \delta,$$

we say that $\{U_i\}$ is a δ-cover of F.

Suppose that $F \subset \mathbb{R}^n$ and s is a real number, and $s \ge 0$. For any $\delta > 0$, we define

$$H_\delta^s(F) = \inf_{\Re(\delta)} \left\{ \sum_{i=1}^{\infty} |U_i|^s : \{U_i\} \text{ is a } \delta\text{-cover of } F \right\}, \qquad (1)$$

where the symbol inf{.} indicates the infimum of {.}. As δ decreases, the class of permissible covers of F in (1) is reduced. Consequently, the infimum $H_\delta^s(F)$ increases, and so approaches a limit as $\delta \to 0$. We have the following definition:

DEFINITION 6. When $\delta \to 0$, *the limit of* $H_\delta^s(F)$ *exists for any subset* F *of* $\mathbb{R}^n$, *and the limiting value can be* (*and usually is*) 0 *or* ∞. *The* s-dimensional Hausdorff measure *of* F *can be defined by*:

$$H^s(F) = \lim_{\delta \to 0} H_\delta^s(F) = \lim_{\delta \to 0} \left[\inf_{\Re(\delta)} \left\{ \sum_{i=1}^{\infty} |U_i|^s : \{U_i\} \text{ is a } \delta\text{-cover of } F \right\} \right]. \qquad (2)$$

We can clearly prove that H^s is a measure. Specifically, $H^s(\phi) = 0$, and if $E \subset F$ then $H^s(E) \le H^s(F)$. If $\{F_i\}$ is any countable collection of *Borel set*, such that

$$\bigcap_{i=1}^{\infty} F_i = \phi,$$

we have

$$H^s\left(\bigcup_{i=1}^{\infty} F_i \right) = \sum_{i=1}^{\infty} H^s(F_i).$$

Furthermore, if F is a Borel subset of $\mathbb{R}^n$, then the n-dimensional *Hausdorff measure* of F can be deduced as:

$$H^n(F) = \frac{\pi^{\frac{n}{2}}}{2^n \Gamma\left(\frac{n+2}{2}\right)} vol^n(F),$$

where

- $H^n(F)$ stands for the n-dimensional *Hausdorff measure* of F.
- $vol^n(F)$ represents the n-dimensional volume of F which is called *Lebesgue measure* of F.
  - vol^1 is length,
  - vol^2 is area,
  - vol^3 is the usual three-dimensional volume.

Consequently, Hausdorff measures generalize the familiar ideas of length, area and volume.

2.2.2 Hausdorff Dimension

Let us review (1). For any set F and $\delta < 1$, $H_\delta^s(F)$ is a nonincreasing function of s. According to (2), it can be shown that $H^s(F)$ is also a nonincreasing function of s. In fact, the stronger conclusion is that if $t > 0$ and $\{U_i\}$ is a δ-cover of F, we have

$$H_\delta^t(F) \le \sum_i |U_i|^t \le \delta^{t-s} \sum_i |U_i|^s. \qquad (3)$$

We take the infimum, that is

$$H_\delta^t(F) \le \delta^{t-s} H_\delta^s(F).$$

DEFINITION 7. *Let* $\delta \to 0$, *if* $H^s(F) < \infty$, *then* $H^t(F) = 0$ *for* $s < t$. *Therefore, there exists a critical value of* s, *such that* $H^s(F)$ *jumps from* ∞ *to* 0 *at this point. This critical value is called the* Hausdorff Dimension *of* F, *and it is symbolized by* $\dim_H F$.

Formally, we have

$$\dim_H F = \inf\{s : H^s(F) = 0\} = \sup\{s : H^s(F) = \infty\}, \qquad (4)$$

and

$$H^s(F) = \begin{cases} \infty & \text{if } s < \dim_H F \\ 0 & \text{if } s > \dim_H F \end{cases}$$

If $s = \dim_H F$, probably $H^s(F)$ is 0 or ∞, or may satisfy

$$0 < H^s(F) < \infty.$$

A Borel set is called an *s-set* if the latter condition, as shown in the above, is satisfied. More details about Hausdorff dimension can be found in [2], [3]. Of the wide variety of fractal dimensions in use, Hausdorff dimension is the oldest and probably the most important. Hausdorff dimension has the advantage of being defined for any set, and is mathematically convenient, as it is based on measures, which are relatively easy to manipulate. A major disadvantage is that, in many cases, it is difficult to calculate or to estimate by computational methods. In practice, *box-computing dimension* is convenient to apply. Therefore, our study will focus on the box-computing dimension.

2.3 Box-Computing Dimension (BCD)

In this section, we will introduce an important concept—*dimension*—followed by box-computing dimension (BCD).

2.3.1 Dimensions

Fundamental to most definitions of dimension is the idea of measurement at scale δ. For each δ, a set can be measured in a way that ignores irregularities of size less than δ, and we see how these measurements behave as $\delta \to 0$.

Suppose F is a plane curve, the measurement $M_\delta(F)$ denotes the number of sets (with length δ) which divide the set F. A dimension of F is determined by the power law obeyed by $M_\delta(F)$ as $\delta \to 0$. If

$$M_\delta(F) \sim \mathcal{K}\delta^{-s}, \qquad (5)$$

for constants $\mathcal{K}$ and s, we might say that F has dimension s, and $\mathcal{K}$ can be considered as "s-dimensional length" of F.

Taking the logarithm of both sides in (5) yields the formula:

$$\log_2 M_\delta(F) \simeq \log_2 \mathcal{K} - s \log_2 \delta,$$

in the sense that the difference of the two sides tends to 0 with δ, we have

$$s = \lim_{\delta \to 0} \frac{\log_2 M_\delta(F)}{-\log_2 \delta}. \qquad (6)$$

From the above equation, s can be regarded as a slope on a log-log scale [2].

2.3.2 Box-Computing Dimension

Box-computing dimension or box dimension is one of the most widely used dimensions. Its popularity is largely due to its relative ease of mathematical calculation and empirical estimation.

Let F be a nonempty and bounded subset of $\mathbb{R}^n$, $\Omega = \{\omega_i: i = 1, 2, 3, ...\}$ be covers of the set F. $N_\delta(F)$ denotes the number of covers, such that

$$N_\delta(F) = \left|\Omega : d_i \le \delta\right|,$$

where d_i stands for the diameter of the ith cover. This equation means that $N_\delta(F)$ is the smallest number of subsets which cover the set F, and their diameters d_i s are not greater than δ (Fig. 2).

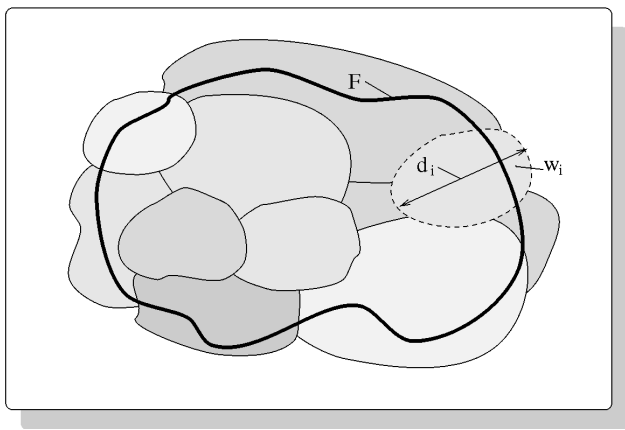


Fig. 2. Opening covers with diameters d_i s covering F.

The upper and lower bounds of the box computing dimension of F can be defined by the following formulas:

$$\underline{\dim}_B F = \underline{\lim}_{\delta\to 0} \frac{\log_2 N_\delta(F)}{-\log_2 \delta}, \tag{7}$$

$$\overline{\dim}_B F = \overline{\lim}_{\delta\to 0} \frac{\log_2 N_\delta(F)}{-\log_2 \delta}, \tag{8}$$

where the over line stands for the upper bound of dimension while the under line for lower bound. An example of the upper bound and lower bound is shown in Fig. 3.

DEFINITION 8. If both the upper bound $\overline{\dim}_B F$ and the lower bound $\underline{\dim}_B F$ are equal, i.e.,

$$\underline{\lim}_{\delta\to 0} \frac{\log_2 N_\delta(F)}{-\log_2 \delta} = \overline{\lim}_{\delta\to 0} \frac{\log_2 N_\delta(F)}{-\log_2 \delta},$$

the common value is called the *box-computing dimension* or *box dimension* of F, namely:

$$\dim_B F = \lim_{\delta\to 0} \frac{\log_2 N_\delta(F)}{-\log_2 \delta}. \tag{9}$$

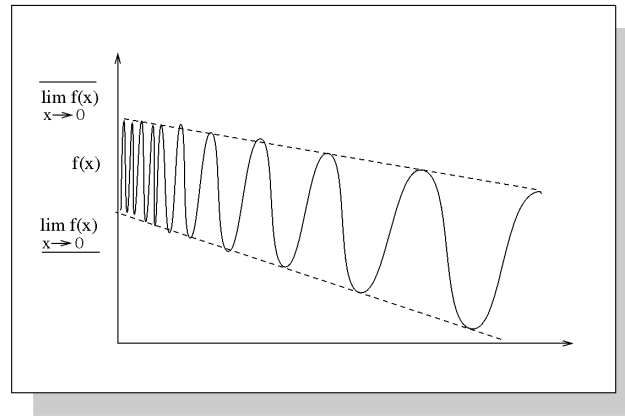


Fig. 3. Upper and lower bounds of a function.

There exist five equivalent definitions of the box computing dimension, that can be found in the following theorem:

THEOREM 1. Let F be a nonempty and bounded set in $\mathbb{R}^n$, and the upper box dimension $\overline{\dim}_B F$, lower box dimension $\underline{\dim}_B F$ and box dimension $\dim_B F$ be represented by:

$$\underline{\dim}_B F = \underline{\lim}_{\delta\to 0} \frac{\log_2 N_\delta(F)}{-\log_2 \delta}$$

$$\overline{\dim}_B F = \overline{\lim}_{\delta\to 0} \frac{\log_2 N_\delta(F)}{-\log_2 \delta}$$

$$\dim_B F = \lim_{\delta\to 0} \frac{\log_2 N_\delta(F)}{-\log_2 \delta}.$$

In the above definition, $N_\delta(F)$ can be considered as one of the following cases (Fig. 4):

(i) the minimum number of closed balls of radius δ that cover F (Fig. 4b);
(ii) the minimum number of cubes with side δ that cover F (Fig. 4c);
(iii) the minimum number of sets with diameter D that cover F such that $D \le \delta$ (Fig. 4d);
(iv) the number of δ-mesh cubs which intersect F (Fig. 4e);
(v) the maximum number of disjoint balls of radius δ with centers in F (Fig. 4f).

PROOF. This proof consists of four steps, namely:

1: (i) ⇔ (ii), 2: (i) ⇔ (iii), 3: (iii) ⇔ (iv), 4: (iv) ⇔ (v).

Step 1: (i) ⇔ (ii)

A cube with side length δ must have only a ball of radius $\frac{\delta\sqrt{n}}{2}$ that covers this cube. On the other hand, any ball of radius δ must have a cube with side length 2δ that covers this ball. Therefore, the definitions (i) and (ii) are equivalent.

Step 2: (i) ⇔ (iii)

Let $U_1, U_2, ..., U_{N_\delta(F)}$ be the sets with diameter at most δ that cover F, and $B_1, B_2, ..., B_{N'_\delta(F)}$ be the closed balls of radius δ that cover F. There exists a B_j for any U_i, such that $U_i \subset B_j$. Thus, it follows that

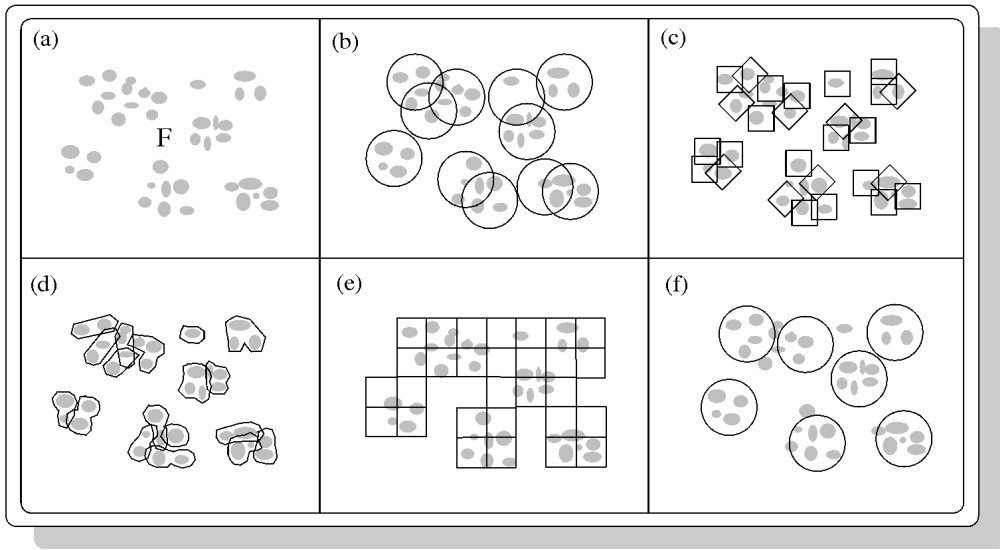


Fig. 4. Graphic illustration of the equivalent definitions of the BCD.

$$N'_\delta(F) \le N_\delta(F).$$

On the other hand, any closed ball B_j of radius δ can be regarded as a set of diameter at most 2δ. Thus, it can be found that

$$N_{2\delta}(F) \le N'_\delta(F).$$

Therefore, the definitions (i) and (iii) are equivalent.

Step 3: (iii) $\Leftrightarrow$ (iv)

Consider δ-mesh cubes in $\mathbb{R}^n$

$$[m_1\delta, (m_1 + 1)\delta] \times [m_2\delta, (m_2 + 1)\delta] \times \cdots \cdots \times [m_n\delta, (m_n + 1)\delta]$$

where $m_1, m_2, ..., m_n$ are positive integers. In the case of $n = 1$, a "cube" in $\mathbb{R}^1$ is a closed interval, while a "cube" in $\mathbb{R}^2$ is a square when $n = 2$. Suppose that

- $N'_\delta(F)$ is the smallest number of cubes in the δ-mesh that intersect F.
- $N_{\delta\sqrt{n}}(F)$ is the smallest number of sets that cover F, and the maximum diameter of them is $\delta\sqrt{n}$.

Because any cube in the δ-mesh can be regarded as a set, in which the maximum diameter is $\delta\sqrt{n}$, we have

$$N_{\delta\sqrt{n}}(F) \le N'_\delta(F).$$

On the other hand, any set of diameter with a maximum value of δ must be covered in the cubes in the δ-mesh, and the largest number of cubes is 3^n. Thus, the following inequality is true

$$N'_\delta(F) \le 3^n N_\delta(F).$$

According to the above two inequalities, we can conclude that the definitions (iii) and (iv) are equivalent.

Step 4: (iv) $\Leftrightarrow$ (v)

Let $B_1, B_2, ..., B_{N'_\delta(F)}$ be the disjoint balls of radius δ with centers in F. If $x \in F$, then the distance between x and some B_i does not extend the value of δ; otherwise, the balls of radius δ with centers in x can be added into set $\{B_1, B_2, ..., B_{N'_\delta(F)}\}$ to grow the series of balls. It is clear that $N'_\delta(F)$ number of closed balls of radius 2δ with the same centers as that of B_i can cover F. Meanwhile, the closed balls of 2δ can be considered as the sets of with a maximum diameter 4δ. Therefore, we have the following inequality:

$$N_{4\delta}(F) \le N'_\delta(F).$$

On the other hand, suppose that $U_1, U_2, ..., U_{N_\delta(F)}$ are the sets of maximum diameter at δ that cover F. Because $B_1, B_2, ..., B_{N'_\delta(F)}$ are the disjoint balls of radius δ with centers in F. Obviously, the center of each B_i must belong to some U_j, and $U_j \subset B_i$. Furthermore, B_is are disjoint each other, thus, U_js which are covered in B_i are also disjoint each other. Therefore, we have the following inequality:

$$N'_\delta(F) \le N_\delta(F)$$

According to these two inequalities, we can conclude that the definitions (iv) and (v) are equivalent. □

THEOREM 2. *Let F be a nonempty and bounded set in $\mathbb{R}^n$, and it satisfies that $1 < H^s(F)$ when $s = \dim_H F$, we have*

$$\dim_H F \le \underline{\dim_B} F \le \overline{\dim_B} F.$$

PROOF. If F can be covered in sets $B_1, B_2, ..., B_{N_\delta(F)}$, then

$$H^s_\delta(F) = \inf\left\{\sum_{i=1}^{\infty} |U_i|^s : \{U_i\} \text{ is a } \delta\text{-cover of } F\right\}$$
$$\le \sum_{i=1}^{N_\delta(F)} |B_i|^s$$
$$= N_\delta(F)\delta^s.$$

When $s = \dim_H F$, we get that

$$1 < H_\delta^s(F) = \lim_{\delta \to 0} H_\delta^s(F).$$

Thus, when $\delta \to 0$, we have

$$1 < H_\delta^s(F) \le N_\delta(F)\delta^s.$$

Taking the logarithm of both sides in this inequality yields

$$0 \le \log_2 N_\delta(F) + s \log_2 \delta.$$

From the inequality

$$s \le \frac{\log_2 N_\delta(F)}{-\log_2 \delta},$$

it follows that

$$\dim_H F \le \frac{\log_2 N_\delta(F)}{-\log_2 \delta}.$$

Furthermore, we have

$$\dim_H F \le \underline{\dim_B} F \le \overline{\dim_B} F. \qquad \square$$

2.4 Minkowski Dimension

To facilitate the application of the box dimension to digital images, we will introduce the *Minkowski dimension*, which is suitable for processing the digital images in computers.

DEFINITION 9. *Let F be a nonempty and bounded set in $\mathbb{R}^n$. For a constant s, if $\delta \to 0$, the limit of $Vol^n(F_\delta)/\delta^{n-s}$ is positive and bounded, we say that F has s dimension of* Minkowski dimension, *and is symbolized by* $\dim_M F$. *Here,* $Vol^n(F_\delta)$ is *called* Lebesgue Measure.

The relationship between the box dimension and Minkowski dimension can be provided by the following theorems.

THEOREM 3. *Let F be a nonempty and bounded set in $\mathbb{R}^n$. Then we have*

$$\underline{\dim_B} F = n - \overline{\lim_{\delta \to 0}} \frac{\log_2 Vol^n(F_\delta)}{\log_2 \delta},$$

$$\overline{\dim_B} F = n - \underline{\lim_{\delta \to 0}} \frac{\log_2 Vol^n(F_\delta)}{\log_2 \delta},$$

where F_δ stands for δ-parallel body of F, and $Vol^n(F_\delta)$ denotes n-dimensional volume of F_δ.

PROOF. If F can be covered by $N_\delta(F)$ number of closed balls of radius δ, then F_δ can be covered by balls of radius 2δ with same centers. Thus,

$$Vol^n(F_\delta) \le N_\delta(F) C_n (2\delta)^n$$

where, C_n denotes the volume of a unit ball. We take the logarithm of both sides in the above inequality, and then divide it by $\log_2 \delta$, giving

$$\frac{\log_2 Vol^n(F_\delta)}{\log_2 \delta} \ge \frac{\log_2 2^n C_n + n \log_2 \delta + \log_2 N_\delta(F)}{\log_2 \delta}.$$

Taking the limits of both sides yields

$$\begin{aligned}\overline{\lim_{\delta \to 0}} \frac{\log_2 Vol^n(F_\delta)}{\log_2 \delta} &\ge \overline{\lim_{\delta \to 0}} \left(\frac{\log_2 2^n C_n}{\log_2 \delta} + n + \frac{\log_2 N_\delta(F)}{\log_2 \delta} \right) \\ &= \overline{\lim_{\delta \to 0}} \left(n - \frac{\log_2 N_\delta(F)}{-\log_2 \delta} \right) \\ &= n - \underline{\lim_{\delta \to 0}} \frac{\log_2 N_\delta(F)}{-\log_2 \delta} \\ &= n - \underline{\dim_B} F.\end{aligned}$$

Therefore, we obtain

$$\underline{\dim_B} F \ge n - \overline{\lim_{\delta \to 0}} \frac{\log_2 Vol^n(F_\delta)}{\log_2 \delta}. \tag{11}$$

On the other hand, suppose $N_\delta(F)$ is the number of disjoint balls of radius δ, and their centers are in F. It can be shown that

$$N_\delta(F) C_n \delta^n \le Vol^n(F_\delta).$$

We take the logarithm of both sides in the above inequality, and then divide it by $\log_2 \delta$, getting

$$\frac{\log_2 2^n C_n + n \log_2 \delta + \log_2 N_\delta(F)}{\log_2 \delta} \ge \frac{\log_2 Vol^n(F_\delta)}{\log_2 \delta}.$$

Taking the limits of both sides yields

$$\begin{aligned}\overline{\lim_{\delta \to 0}} \frac{\log_2 Vol^n(F_\delta)}{\log_2 \delta} &\le \overline{\lim_{\delta \to 0}} \left(\frac{\log_2 2^n C_n}{\log_2 \delta} + n + \frac{\log_2 N_\delta(F)}{\log_2 \delta} \right) \\ &= \overline{\lim_{\delta \to 0}} \left(n - \frac{\log_2 N_\delta(F)}{-\log_2 \delta} \right) \\ &= n - \underline{\lim_{\delta \to 0}} \frac{\log_2 N_\delta(F)}{-\log_2 \delta} \\ &= n - \underline{\dim_B} F.\end{aligned}$$

Therefore, we obtain

$$\underline{\dim_B} F \le n - \overline{\lim_{\delta \to 0}} \frac{\log_2 Vol^n(F_\delta)}{\log_2 \delta}.$$

According to (10) and (11), we therefore have

$$\underline{\dim_B} F = n - \overline{\lim_{\delta \to 0}} \frac{\log_2 Vol^n(F_\delta)}{\log_2 \delta}.$$

Similar to the above proof, we obtain

$$\overline{\dim_B} F = n - \underline{\lim_{\delta \to 0}} \frac{\log_2 Vol^n(F_\delta)}{\log_2 \delta}. \qquad \square$$

From Theorem 3, we can derive Theorem 4, which is a very important theorem. It shows that the box computing dimension is equal to the Minkowski dimension in a nonempty and bounded set in $\mathbb{R}^n$.

THEOREM 4. *Let F be a nonempty and bounded set in $\mathbb{R}^n$. We have*

$$\dim_B F = \dim_M F.$$

PROOF. According to Definition 9, set F has s dimension of Minkowski dimension, which means that there exists a constant $\beta > 0$, so that

$$\lim_{\delta \to 0} \frac{Vol^n(F_\delta)}{\delta^{n-s}} = \beta.$$

Taking the limits of both sides yields

$$\begin{aligned}\log_2 \beta &= \log_2 \lim_{\delta \to 0} \frac{Vol^n(F_\delta)}{\delta^{n-s}} \\ &= \lim_{\delta \to 0} \log_2 \frac{Vol^n(F_\delta)}{\delta^{n-s}} \\ &= \lim_{\delta \to 0} \left[\log_2 Vol^n(F_\delta) - (n-s)\log_2 \delta\right].\end{aligned}$$

Furthermore, we can derive

$$\begin{aligned}\frac{\log_2 \beta}{\lim_{\delta \to 0} \log_2 \delta} &= \frac{\lim_{\delta \to 0}\left[\log_2 Vol^n(F_\delta) - (n-s)\log_2 S\right]}{\lim_{\delta \to 0} \log_2 \delta} \\ &= \lim_{\delta \to 0}\left[\frac{\log_2 Vol^n(F_\delta)}{\log_2 \delta}\right].\end{aligned}$$

It is clear that

$$\frac{\log_2 \beta}{\lim_{\delta \to 0} \log_2 \delta} = 0.$$

Hence, we get

$$s = n - \lim_{\delta \to 0} \frac{\log_2 Vol^n(F_\delta)}{\log_2 \delta},$$

and Theorem 3 shows that

$$\dim_B F = n - \lim_{\delta \to 0} \frac{\log_2 Vol^n(F_\delta)}{\log_2 \delta}.$$

Therefore, we can conclude that

$$\dim_B F = \dim_M F. \qquad \square$$

In the remaining part of this paper, the Minkowski dimension will be used for automatic analysis of documents.

3 Approximation of Fractal Dimension in Document Images

The basic idea of this new approach is that a document image is mapped onto a gray-level function. Furthermore, this function can be mapped onto a gray-level surface, and from the area of such surface, the fractal dimension of the document image can be approximated. However, directly calculating the area of the gray-level surface of the document image is an obscure task. To simplify this computation, a special equivalent technique of the Minkowski dimension technique which was mentioned in the last section is applied in this study, referred as *δ Parallel Bodies*. Using the δ parallel body of the gray-level surface of the document image, we first thicken the gray-level surface, so that it becomes a 3D parallel body. Then, we calculate the volume of that body, since the calculation of a volume is much easier than that of a gray-level surface. A diagram of this process is illustrated in Fig. 5. We call it the *Modified Fractal Signature* (*MFS*) *approach*, since the direct computation of fractal dimension of a document image is not used in this method, alternately, the volume of a δ parallel body is estimated to approximate the fractal dimension.

In this section, the preliminaries of δ parallel bodies will be presented first. To solve the problem of computing the fractal dimension of a document image, a special form of the δ parallel body, blanket, is applied, therefore, it will be stated in this section. Then a new method for document analysis using the modified fractal signature will be described. Finally, a significant algorithm will be presented.

3.1 δ-Parallel Bodies

It is worthwhile discussing another equivalent definition of the box-computing dimension, which is a rather different form.

DEFINITION 10. *δ-parallel body $F\delta$ can be defined by*:

$$F\delta = \{x \in IR^n : |x - y| \le \delta, \text{ for } y \in F\}. \tag{12}$$

It is obvious that F_δ is a set of points, where the distance between F and any element of F_δ is not greater than δ. We

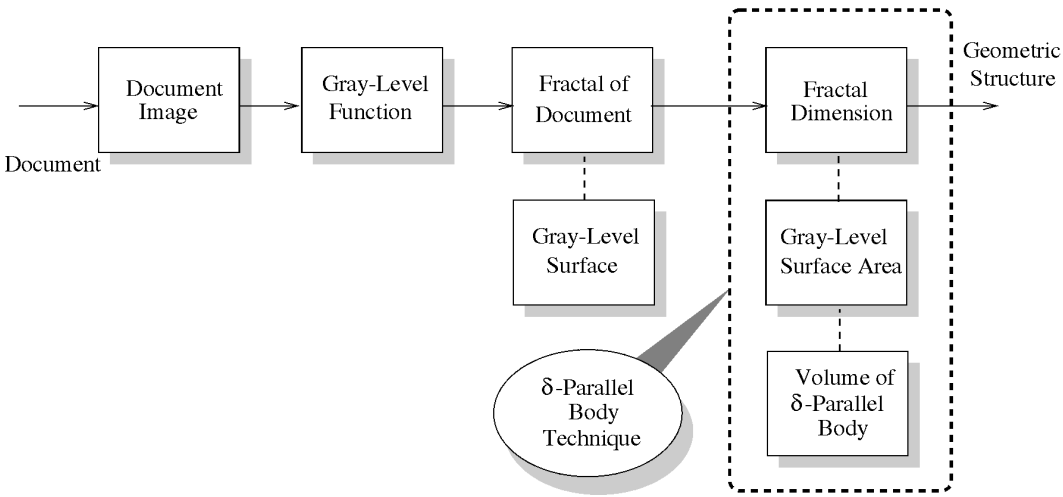


Fig. 5. Diagram of modified fractal signature approach to document analysis.

consider the rate at which the n-dimensional volume of F_δ shrinks as $\delta \to 0$. Specifically, in $\mathbb{R}^3$, we consider the following examples:

- F is a set containing only one point, i.e., F is a single point: F_δ is a ball (Fig. 6a) with volume of
$$Vol(F_\delta) = \frac{4}{3}\pi\delta^3;$$
- F is a set containing a segment of straight line with length of L: F_δ is a "cylinder" (Fig. 6b) with volume of
$$Vol(F_\delta) = \pi L\delta^2;$$
- F is a set containing a segment of curve with length of L: F_δ is a "sausage-like" (Fig. 6c) with volume of
$$Vol(F_\delta) \sim \pi L\delta^2;$$
- F is a set containing a plane with area of $\mathcal{A}$: F_δ is a "brick" (Fig. 6d) with volume of
$$Vol(F_\delta) = 2\mathcal{A}\delta.$$

In each case, the following formula holds:

$$Vol(F_\delta) \sim \beta\delta^{3-D} \tag{13}$$

where β is a constant, and the integer D is the *Fractal Dimension* of F.

Let $F = \{X_{i,j}\}$, $i = 0, 1, ..., K$, $j = 0, 1, ..., L$ be a document image with multigray level, and $X_{i,j}$ be the gray level of the (i, j)th pixel. In a certain measure range, the gray-level surface of F can be viewed as a fractal. The surface area can be used to approximate its fractal dimension.

Particularly in document processing, the gray level function F is a nonempty and bounded set in $\mathbb{R}^3$ for either text areas, graphics areas, or background areas in a document. Thus, the δ-parallel body can be applied, and a special technique that is referred to as the *Blanket Technique* is used to thicken the function F. It leads to a set F_δ, which is still a nonempty and bounded set in $\mathbb{R}^3$. According to the definition of the *Minkowski Dimension* and Theorem 4, we can conclude that if

$$\lim_{\delta\to 0} \frac{Vol^3(F_\delta)}{\delta^{3-D}} = \beta > 0,$$

then

$$D = \dim_M F = \dim_B F,$$

where β denotes a constant, and $Vol^3(F_\delta)$ stands for the volume of the blanket F_δ. Therefore, when δ is sufficiently small, we have

$$Vol^3(F_\delta) \approx \beta\delta^{3-D}.$$

Let $A(\delta)$ be the area of the surface of the blanket, it can be represented as

$$A(\delta) = \frac{Vol^3(F_\delta)}{2\delta} \approx \frac{\beta\delta^{2-D}}{2}. \tag{14}$$

In fact, $A(\delta)$ is the area of the gray level surface of the image of a document.

To simplify the representation, we use notion β instead of the notion $\frac{\beta}{2}$. Thus, (14) can be rewritten by

$$A(\delta) \approx \beta\delta^{2-D}, \text{ if } \delta \text{ is sufficiently small,} \tag{15}$$

where β denotes a constant, and D stands for the fractal dimension of a document image.

According to (15), the fractal dimension D can be computed from the area $A(\delta)$. More precisely, taking the logarithm of both sides in (15) yields:

$$\log_2 A(\delta) \approx \log_2\beta + (2 - D)\log_2\delta,$$
$$2 - D \approx \frac{\log_2 A(\delta)}{\log_2\delta} - \frac{\log_2\beta}{\log_2\delta},$$
$$D \approx 2 - \frac{\log_2 A(\delta)}{\log_2\delta} + \frac{\log_2\beta}{\log_2\delta}. \tag{16}$$

It should be noted that when δ is sufficiently small, the item $\frac{\log_2\beta}{\log_2\delta}$ in (16) approaches zero, namely,

$$\frac{\log_2\beta}{\log_2\delta} \approx 0.$$

Consequently, (16) becomes

$$D \approx 2 - \frac{\log_2 A(\delta)}{\log_2\delta}.$$

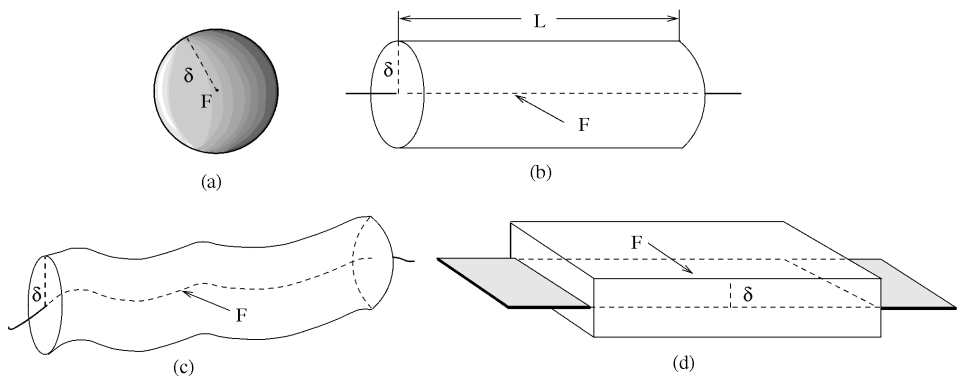


Fig. 6. δ-parallel bodies.

This is an important formula to approximate the fractal dimension of a document image.

3.2 Blanket Technique and Modified Fractal Signature (MFS)

To compute the fractal dimension, we need to measure the area of the gray level surface. In our study, a "blanket" approach is used for this purpose [9]. The idea of the blanket technique is based on the equivalent definition of the BCD shown in (12), i.e., the δ-parallel body. The basic idea of the blanket technique will be presented below:

In the blanket technique, all points of the three-dimensional space at distance δ from the gray level surface are considered. These points construct a "blanket" of thickness 2δ covering this surface. A graphical illustration is shown in Fig. 7. The document image is represented by a gray-level function $g(i, j)$. The covering blanket is defined by its upper surface $u_\delta(i,j)$ and its lower surface $b_\delta(i,j)$. Initially, $\delta = 0$ and given the gray-level function equals the upper and lower surfaces, namely:

$$g(i, j) = u_0(i, j) = b_0(i, j).$$

For $\delta = 1, 2, \ldots$, the blanket surfaces are defined iteratively as follows:

$$u_\delta(i,j) = \max\left\{u_{\delta-1}(i,j)+1, \max_{|(m,n)-(i,j)|\le 1} u_{\delta-1}(m,n)\right\} \quad (18)$$

$$b_\delta(i,j) = \min\left\{b_{\delta-1}(i,j)-1, \min_{|(m,n)-(i,j)|\le 1} b_{\delta-1}(m,n)\right\} \quad (19)$$

The image pixels (m, n) with distance less than one from (i, j) are taken to be the four immediate neighbors of (i, j). Similar expressions exist when the eight-neighborhood is desired. A point $f(x, y)$ will be included in the blanket for δ when $b_\delta(x,y) < f(x,y) < u_\delta(x,y)$. The blanket definition uses the fact that the blanket of the surface for radius δ includes all the points of the blanket for radius $\delta - 1$, together with all the points within radius 1 from the surfaces of that blanket (Fig. 8). Equation (18) ensures that the new upper surface u_δ is higher than $u_{\delta-1}$ by at least 1, and also at distance at least one from $u_{\delta-1}$ in the horizontal and vertical directions.

The volume Vol_δ of the blanket is computed from u_δ and b_δ:

$$Vol_\delta = \sum_{i,j}\left(u_\delta(i,j) - b_\delta(i,j)\right). \quad (20)$$

DEFINITION 11. *As the volume $Vol\delta$ of the blanket is measured with radius δ, the area of a fractal surface can be deduced, which is called* fractal signature (FS)

$$A_\delta = \frac{Vol_\delta}{2\delta}, \quad (21)$$

or

$$A_\delta = \frac{Vol_\delta - Vol_{\delta-1}}{2}. \quad (22)$$

In this study, the latter will be used.

The area of a fractal surface behaves according to (15), namely:

$$A(\delta) \approx \beta\delta^{2-D}, \quad \delta = 1, 2, \ldots,$$

from which the fractal dimension D can be computed. Since the dimension can be regarded as a slope on a log-log scale,

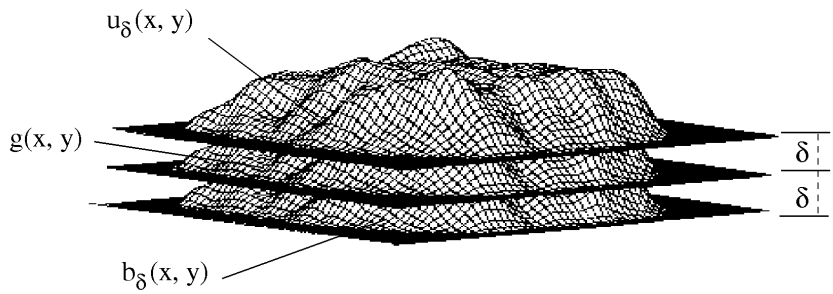


Fig. 7. Surface and its blankets.

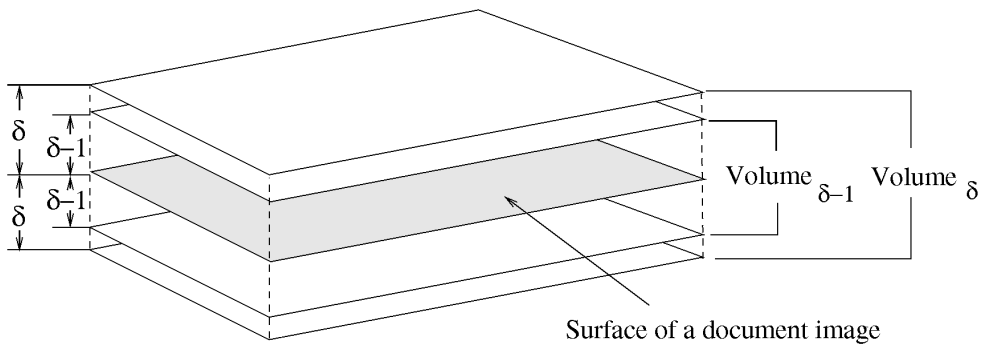


Fig. 8. Calculation of the area of a surface using volumes of two blankets δ and $\delta-1$.

to get the dimension, only two points are needed. We use two values of δ to compute the fractal dimension, namely, we take δ_1 and δ_2, then

$$A_{\delta_1} \approx \beta\delta_1^{2-D}, \tag{23}$$

$$A_{\delta_2} \approx \beta\delta_2^{2-D}. \tag{24}$$

When (23) is divided by (24), we have

$$\frac{A_{\delta_1}}{A_{\delta_2}} \approx \frac{\delta_1^{2-D}}{\delta_2^{2-D}}.$$

Taking the logarithm of the both sides yields:

$$2 - D \approx \frac{\log_2 A_{\delta_1} - \log_2 A_{\delta_2}}{\log_2 \delta_1 - \log_2 \delta_2},$$

$$D \approx 2 - \frac{\log_2 A_{\delta_1} - \log_2 A_{\delta_2}}{\log_2 \delta_1 - \log_2 \delta_2}. \tag{25}$$

Thus, the fractal dimension D has been computed.

According to the property of proportions, the following is true:

$$\text{if } \frac{a_1}{b_1} = \frac{a_2}{b_2}, \text{ then } \frac{a_1}{b_1} = \frac{a_1 - a_2}{b_1 - b_2}.$$

Therefore, from

$$\frac{\log_2 A_{\delta_1}}{\log_2 \delta_1} = \frac{\log_2 A_{\delta_2}}{\log_2 \delta_2}$$

yields

$$\frac{\log_2 A_{\delta_1}}{\log_2 \delta_1} = \frac{\log_2 A_{\delta_1} - \log_2 A_{\delta_2}}{\log_2 \delta_1 - \log_2 \delta_2}.$$

Consequently, (17) and (25) are equivalent, namely:

$$\left(D \approx 2 - \frac{\log_2 A(\delta)}{\log_2 \delta}\right) \equiv \left(D \approx 2 - \frac{\log_2 A_{\delta_1} - \log_2 A_{\delta_2}}{\log_2 \delta_1 - \log_2 \delta_2}\right) \tag{26}$$

Recall the formula of (17) is used to approximate the fractal dimension of document images, i.e.,

$$D \approx 2 - \frac{\log_2 A(\delta)}{\log_2 \delta} \qquad \delta \text{ is sufficiently small.}$$

Several points are worth noting:

- Why do the different document images have different fractal dimensions? The essential distinction of document images is their values of $A(\delta)$.
- The value of $A(\delta)$ depends on the volume $Vol^3(F_\delta)$ of the thickened blanket F_δ only.
- In summary, they can be represented as

$$D \Leftrightarrow A(\delta) \Leftrightarrow Vol^3(F_\delta).$$

Consequently, in this paper, the volume $Vol^3(F_\delta)$ of the thickened blanket F_δ is applied to identify different blocks in a document, instead of using the fractal dimension. We call such technique of approximating the fractal dimension "Modified Fractal Signature."

3.3 Modified Fractal Signature Approach to Document Analysis

In order to extract the geometric structure of a document, *Modified Fractal Signature* (*MFS*) is used in this paper.

From the definition of the FS, i.e., (21), it is clear that the fractal signature is completely determined by the area of the surface which is a mapping of the gray-level function representing a document image. Consequently, the fractal signature reflects certain characteristics of the document image.

Consider a page of document F which consists of many regions which might be texts, graphics, and background areas.

$$F = \{\Im_T, \Im_G, \Im_B\}$$

where

- $\Im_T$ represents a set of *text areas*,
- $\Im_G$ stands for a set of *graphic areas*,
- $\Im_B$ denotes a set of *background areas*.

Different regions $\Im_T$, $\Im_G$, and $\Im_B$ have different gray-level functions $g_{\Im_T}$, $g_{\Im_G}$, and $g_{\Im_B}$. The different gray-level functions have different surfaces. Furthermore, the different surfaces have different areas $A(\Im_T)$, $A(\Im_G)$, and $A(\Im_B)$ from which different fractal signatures can be estimated:

$$\Im_T \Leftrightarrow g_{\Im_T} \Leftrightarrow A(\Im_T)$$

$$\Im_G \Leftrightarrow g_{\Im_G} \Leftrightarrow A(\Im_G)$$

$$\Im_B \Leftrightarrow g_{\Im_B} \Leftrightarrow A(\Im_B),$$

where $g_{\Im_T}$, $g_{\Im_G}$, and $g_{\Im_B}$ denote the gray-level functions of the text, graphics and background regions, respectively; and $A(\Im_T)$, $A(\Im_G)$, and $A(\Im_B)$ denote the fractal signatures of the text, graphics, and background regions, respectively. For instance, the gray-level functions of the text region and background region are provided below:

1) $\Im_T$: The gray-level function for a text block is

$$g_{\Im_T}(x,y) = \begin{bmatrix} 6 & 2 & 6 & 7 & 2 & 3 & 2 & 3 \\ 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\ 7 & 2 & 7 & 3 & 2 & 6 & 2 & 4 \\ 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\ 6 & 2 & 3 & 7 & 2 & 4 & 2 & 7 \\ 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\ 4 & 2 & 6 & 4 & 2 & 7 & 2 & 8 \\ 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \end{bmatrix},$$

where the elements of values 2s denote the background, i.e., the intervals between characters and text lines. The elements with values which are not 2s indicate the pixels of the texts. This function can be mapped onto a building-surface as shown in Fig. 9a. In the same region, the area of this type of surface is greater than that of a plane.

2) $\Im_B$: The gray-level function of a block of the background is

$$g_{\Im_B}(x,y) = \begin{bmatrix} 2 & 3 & 2 & 2 & 3 & 2 & 2 & 2 \\ 2 & 2 & 1 & 2 & 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 3 & 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 1 & 2 & 3 & 2 & 2 \\ 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 & 3 & 1 & 2 & 2 \\ 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \end{bmatrix},$$

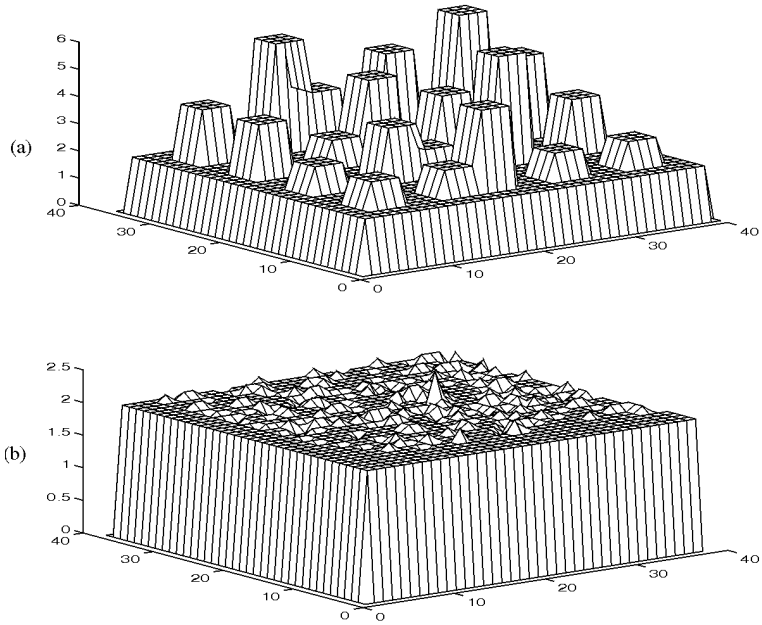


Fig. 9. Graphical example of representing gray-level functions.

where most of the elements of this matrix have the same values of 2, that means only a few changes in gray level, since this a block of the background. The elements with values which are not 2s represent noise. This function can be mapped onto a plane as shown in Fig. 9b. In the same region, the area of this surface is less than that of a building-surface.

Consequently, in the same region, the fractal signature of a plane is less than that of a building-surface, that means the fractal signature of a text block is greater than that of a background block, i.e., $A(\Im_T) > A(\Im_B)$. Thus we have the following theorem:

THEOREM. *Let $\Im_T$ be a text block of a document, and $\Im_B$ be a background block of the document. If both blocks have the same geometrical size, then the fractal signature of the text block is greater than that of the background, namely,*

$$A(\Im_T) > A(\Im_B).$$

PROOF**.** (The proof of it will be excluded, because it is very simple). □

A graphical example can be illustrated in Figs. 10a and 10b. Fig. 10a is a part of a document with texts, while Fig. 10b shows the result of using the fractal signature, the bright parts indicate the higher values of the fractal signatures which represent the text regions.

This is a very significant characteristic of the fractal signature. It can be used to identify different kinds of blocks in the document. This is the basic idea of this new approach for document analysis.

3.4 Algorithm

An algorithm has been designed to compute the fractal signature. It is based on the equivalent definition of the box computing dimension (BCD). More precisely, (18), (19), (20), and (22) are used for designing the following algorithm:

Algorithm 1 (fractal signature)
Input: a page of document image;
Output: the geometric structure of the document;

Step-1. The whole image F is divided into several non-overlapping subimages $R_k(x, y)$, and each subimage has size $N \times N$;
Step-2. For k = 1 to n do
$R_k(x, y)$ is mapped onto a gray-level function $g_k(x, y)$;
Step-3. For k = 1 to n do
Substep-1. Initially, taking $\delta = 0$, the upper layer $u_0^k(x, y)$ and lower layer $b_0^k(x, y)$ of the blanket are chosen as the same as the gray-level function $g_k(x, y)$, namely:

$$u_0^k(x, y) = b_0^k(x, y) = g_k(x, y);$$

Substep-2. Taking $\delta = \delta_1$,

(a) $u_{\delta_1}(x, y)$ is computed according to (18), i.e.:

$$u_{\delta_1}(x, y) = \max\left\{u_0(x, y) + 1, \max_{|(i,j)-(x,y)|\le 1} u_0(i, j)\right\};$$

(b) $b_{\delta_1}(x, y)$ is computed according to (19), i.e.:

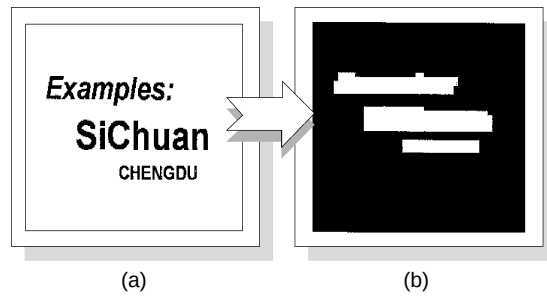


Fig. 10. Graphical example of Theorem 5.

$$b_{\delta_1}(x,y) = \min\left\{b_0(x,y) - 1, \min_{|(i,j)-(x,y)|\le 1} b_0(i,j)\right\};$$

(c) The volume Vol_{δ_1} of the blanket is computed by (20), i.e.:

$$Vol_{\delta_1} = \sum_{x,y}\left(u_{\delta_1}(x,y) - b_{\delta_1}(x,y)\right);$$

Substep-3. Taking $\delta = \delta_2 = \delta_1 + 1$,

(a) $u_{\delta_2}(x,y)$ is computed according to

$$u_{\delta_2}(x,y) = \max\left\{u_{\delta_1}(x,y) + 1, \max_{|(i,j)-(x,y)|\le 1} u_{\delta_1}(i,j)\right\};$$

(b) $b_{\delta_2}(x,y)$ is computed according to

$$b_{\delta_2}(x,y) = \min\left\{b_{\delta_1}(x,y) - 1, \min_{|(i,j)-(x,y)|\le 1} b_{\delta_1}(i,j)\right\};$$

(c) The volume Vol_{δ_1} of the blanket is computed by

$$Vol_{\delta_2} = \sum_{x,y}\left(u_{\delta_2}(x,y) - b_{\delta_2}(x,y)\right);$$

Step-4. The subfractal signature A_δ^k is computed by (22), namely:

$$A_\delta^k = \frac{Vol_{\delta_2} - Vol_{\delta_1}}{2}.$$

Step-5. Combining subfractal signatures A_δ^k, k = 1, 2, ..., n into the whole fractal signature gives:

$$A_\delta = \bigcup_{k=1}^{n} A_\delta^k.$$

4 Experiments

Experiments have been conducted to prove the effectiveness of the proposed new approach for document processing. All experiments have been conducted in both personal computer system PC/386/486, and Sun SPARCstation computer system. An HP scanner with resolution of 100-600 DPI is employed to capture the image of the documents which are scanned with multigray levels.

In our experiments, the whole image F is divided into several nonoverlapping subimages $R_k(x, y)$, and each subimage has size $N \times N$. The algorithm presented in the previous section is used. Here we concentrate our study to finding the upper and lower layers of the blanket for a given surface which represents a part of document.

Initially, taking $\delta = 0$, the upper layer $u_0^k(x,y)$ and lower layer $b_0^k(x,y)$ of the blanket are chosen as the same as the gray-level function $g_k(x, y)$. For example, a gray-level function $g_k(x, y)$ is

$$g_k(x,y) = \begin{bmatrix} 1 & 1 & 2 & 3 \\ 1 & 2 & 4 & 5 \\ 2 & 4 & 3 & 2 \\ 1 & 1 & 2 & 1 \end{bmatrix},$$

as illustrated in Fig. 11a.

Taking $\delta = 1$, the upper layer $u_1(x, y)$ is computed according to (18), i.e.,

$$u_1(x,y) = \max\left\{u_0(x,y) + 1, \max_{|(i,j)-(x,y)|\le 1} u_0(i,j)\right\}.$$

The result can be found in

$$u_1(x,y) = \begin{bmatrix} 2 & 2 & 4 & 5 \\ 2 & 4 & 5 & 6 \\ 2 & 5 & 4 & 5 \\ 2 & 4 & 3 & 2 \end{bmatrix},$$

which can be shown in Fig. 11b. The lower layer $b_1(x, y)$ is computed according to (19), i.e.:

$$b_1(x,y) = \min\left\{b_0(x,y) - 1, \min_{|(i,j)-(x,y)|\le 1} b_0(i,j)\right\}.$$

The result can be found in

$$b_1(x,y) = \begin{bmatrix} 0 & 0 & 1 & 2 \\ 0 & 1 & 2 & 2 \\ 2 & 1 & 2 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

which can be shown in Fig. 11c.

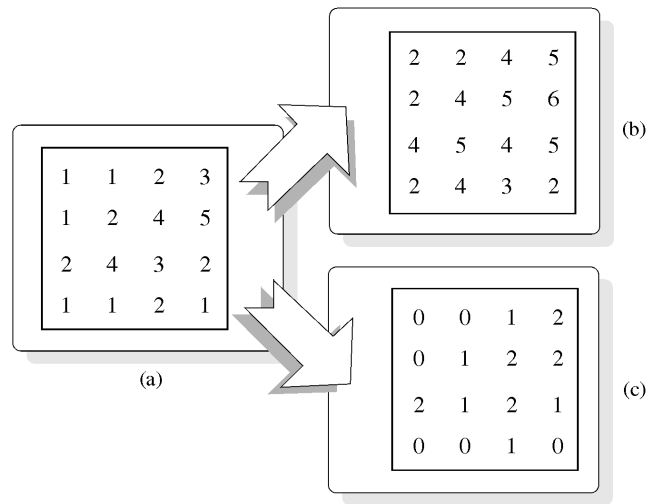


Fig. 11. An example of gray-level function and its upper and lower layers.

Two groups of documents have been tested in our experiments. The first group consists of two samples of documents where the document blocks have regular shapes. The geometric structures for these documents are not too complicated. Fig. 12a shows a portion of a page of newspaper. Fig. 12b illustrates the document blocks obtained from Fig. 12a by document analysis using the fractal signature. Furthermore, from these document blocks, the document geometric structure can be extracted and illustrated in Fig. 13. In this figure, *H*1, *H*2, and *H*3 stand for the "Headline Blocks" which represent titles of articles. *T*1, *T*2, ..., *T*6 indicate the "Text Line Blocks," corresponding to texts of different papers in the page. *G*1 means that this document contains only one graphic block.

Another example of the first group can be shown in Fig. 14.

The second group of documents used in our experiments are illustrated in Figs. 15 and 16, where the document blocks have in-regular shapes. The geometric structures for

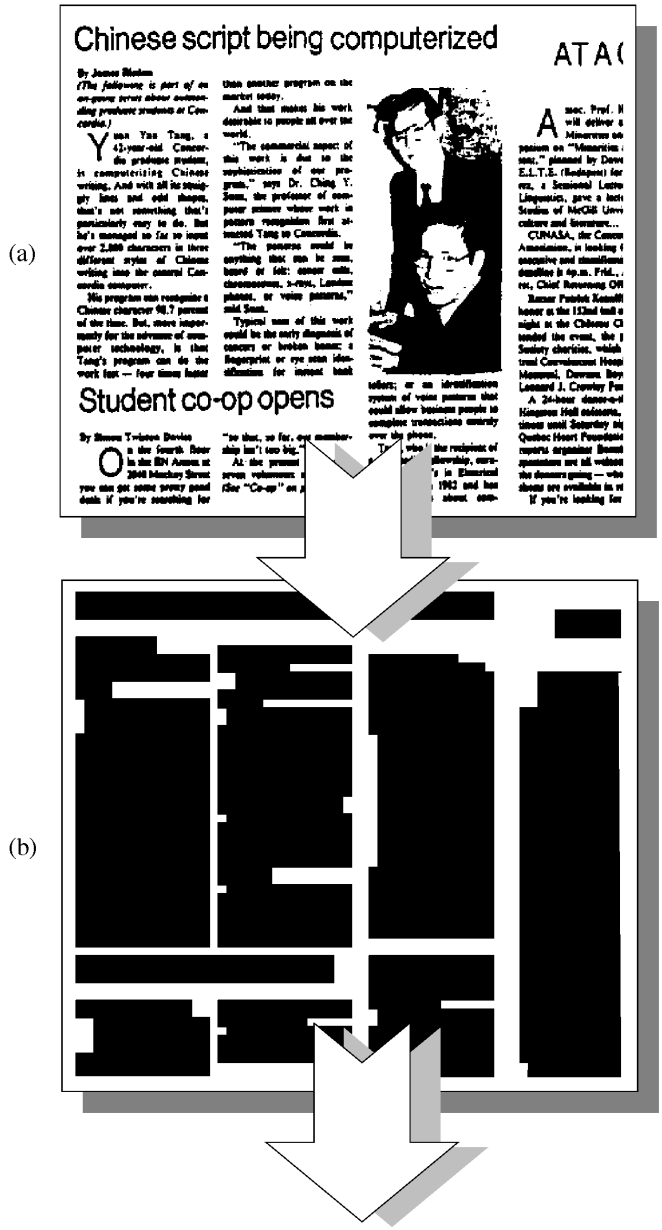


Fig. 12. A portion of a page of newspaper is broken into blocks by the MFS.

these documents are complicated. We used the traditional methods, such as projection profile cuts, run-length smoothing (RLSA) to process them will fail because the document pages cannot be divided into correct blocks. The new approach has been applied to these complicated documents, and the positive results have been obtained shown in Figs. 15 and 16.

5 Conclusions

The knowledge acquisition bottleneck has become the major impediment to the development and application of effective information systems. To remove this bottleneck, new document processing techniques must be introduced to acquire automatically knowledge from various types of documents. A document has its geometric structure which

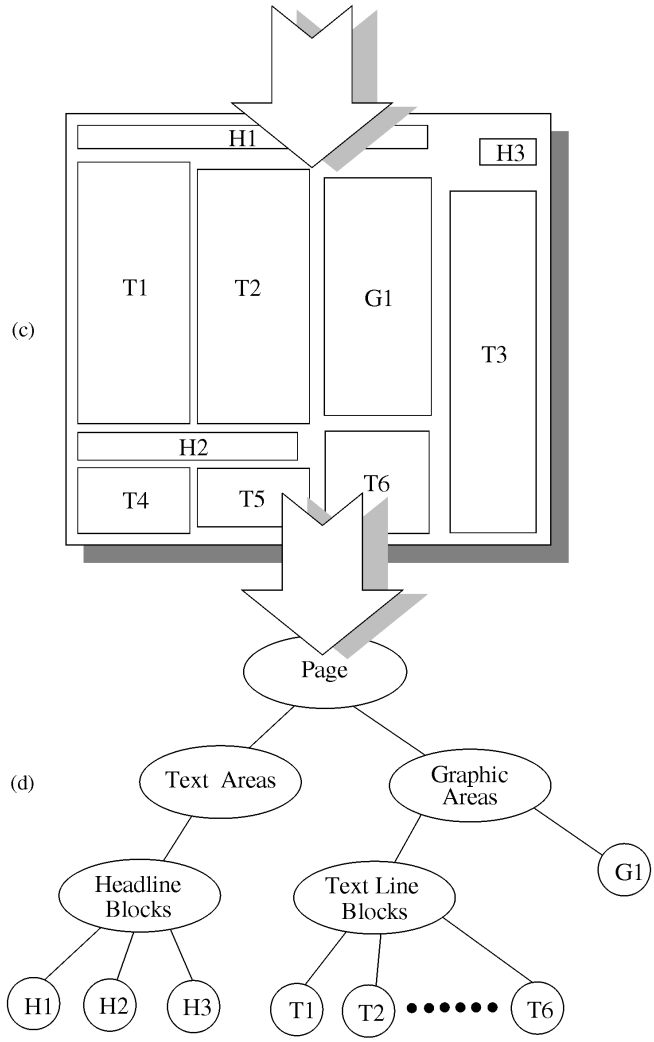


Fig. 13. Document geometric structure extracted from the blocks in Fig. 12.

plays a key role in the process of the knowledge acquisition from documents. Extracting the geometric structure from a document refers to document analysis.

Document analysis as a branch of the document processing has grown rapidly during the past decade. So far, many techniques have been developed, but all of them can be classified into two approaches, namely, top-down and bottom-up approaches. Both approaches have their weaknesses, the top-down one becomes ineffective when the documents have higher geometrical complexity; either top-down or bottom-up needs iterative operations to break a document into several blocks to extract its geometric (layout) structure. Therefore, both approaches are time consuming. To overcome these problems, this paper presents a new approach that does not need iterative breaking, it can divide a document into meaningful blocks effectively. This approach can be used to processing documents with high geometrical complexity.

The basic idea of this new approach is that a document image can be mapped onto a gray-level function. Furthermore, this function can be mapped onto a surface which can be used to approximate its fractal dimension. Fractals are

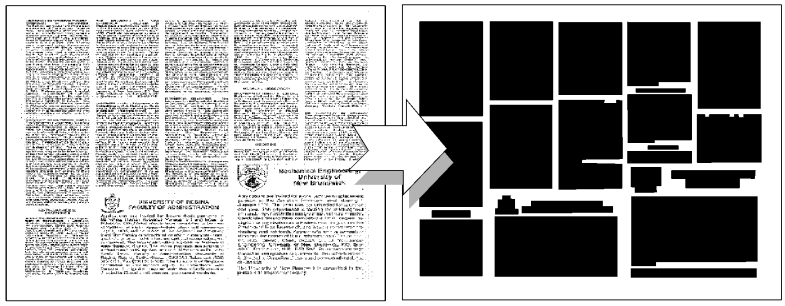


Fig. 14. An example of document analysis by the MFS.

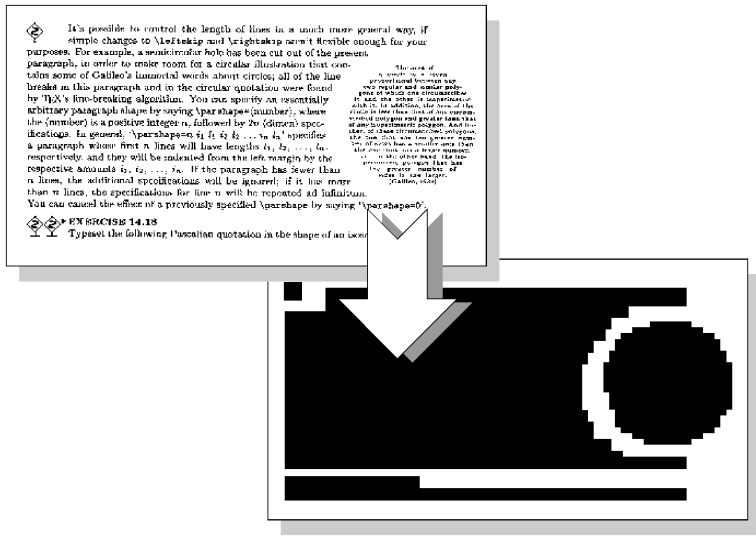


Fig. 15. Analysis of document with in-regular blocks by the MFS.

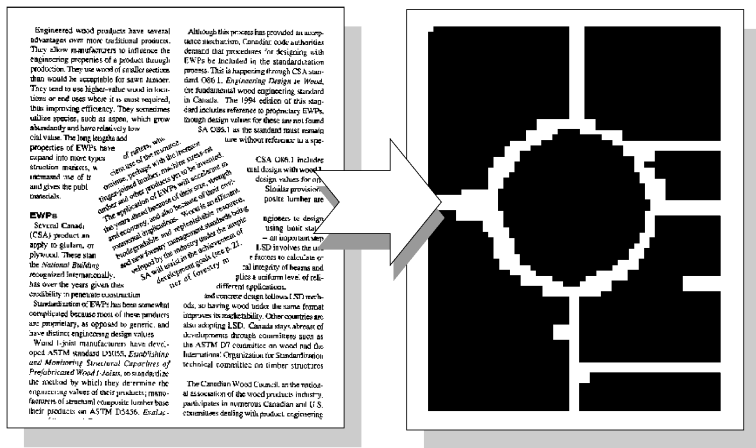


Fig. 16. Another example of document analysis by the MFS.

mathematical sets with a high degree of geometrical complexity which can model many classes of time-series data as well as images. The fractal dimension is an important characteristic of fractals that contains information about their geometrical structure. The proposed approach is based on such characteristics. More precisely, the fractal signature is completely determined by the area of the surface. The surface is a mapping of the gray-level function which represents a document image. Consequently, the fractal signature reflects certain characteristics of the document image.

ACKNOWLEDGMENTS

This work was supported by research grants received from the Research Grant Council (RGC) of Hong Kong and by a Faculty Research Grant (FRG) of Hong Kong

Baptist University. This work was also supported by the Ministry of Education of the People's Republic of China, and Sichuan University, China. We wish to express our gratitude to Professors Zhisheng You, Jiansun Nie, Hui Li, and other staff members at Sichuan University, China, for their assistance in this research project.

REFERENCES

[1] M.F. Barnsley and A.D. Sloan, "A Better Way to Compress Images," *Byte*, pp. 215-223, Jan. 1988.
[2] K.L. Falconer, *Fractal Geometry: Mathematical Foundation and Applications.* New York: Wiley, 1990.
[3] K.L. Falconer, *The Geometry of Fractal Sets.* Cambridge, England: Cambridge Univ. Press, 1985.
[4] E.A. Parrish Jr., "A Foreword to Knowledge and Data Engineering," *IEEE Trans. Knowledge and Data Eng.*, vol. 1, no. 1, pp. 5-7, 1989.
[5] B.B. Mandelbrot, *The Fractal Geometry of Nature.* New York: Freeman, 1982/1983.
[6] P. Maragos and F.K. Sun, "Measuring the Fractal Dimension of Signals: Morphological Covers and Iterative Optimization," *IEEE Trans. Signal Processing*, vol. 41, no. 1, pp. 108-121, Jan. 1993.
[7] J.R. Munkres, *Topology, A First Course.* Englewood Cliffs: N.J.: Prentice Hall, Inc., 1975.
[8] P.T. Nguyen and J. Quinqueton, "Space Filling Curves and Texture Analysis," *Proc. Int'l Conf. Pattern Recognition*, Munich, Germany, pp. 282-285, 1982.
[9] S. Peleg, J. Naor, R. Hartley, and D. Avnir, "Multiple Resolution Texture Analysis and Classification," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 6, no. 4, pp. 518-523, July 1984.
[10] A. Pentland, "Fractal-Based Description of Natural Scenes," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 6, no. 11, pp. 661-674, Nov. 1984.
[11] C.V. Ramamoorthy and B.W. Wah, "Knowledge and Data Engineering," *IEEE Trans. Knowledge and Data Eng.*, vol. 1, no. 1, pp. 9-15, 1989.
[12] Y.Y. Tang, C.Y. Suen, and C.D. Yan, "Document Processing for Automatic Knowledge Acquisition," *IEEE Trans. Knowledge and Data Eng.*, vol. 6, no. 1, pp. 3-21, 1994.

Yuan Y. Tang received the BS degree in electrical and computer engineering from Chongqing University, Chongqing, China; the MEng degree in electrical engineering from the Graduate School of Posts and Telecommunications, Beijing, China; and the PhD degree in computer science from Concordia University, Montreal, Canada.

He is presently an associate professor in the Department of Computing Studies at Hong Kong Baptist University, and a senior researcher in the Center for Pattern Recognition and Machine Intelligence, Concordia University. He is an advisory professor of Beijing University of Posts and Telecommunications and Chongqing University, China. His current interests include wavelet theory and applications, pattern recognition, image processing, document processing, artificial intelligence, parallel processing, Chinese computing, and VLSI architecture.

A senior member of the IEEE, Dr. Tang has published more than 110 technical papers and is the author/coauthor of seven books on subjects ranging from electrical engineering to computer science. He also serves as a reviewer for many journals and international conferences. He was session chairman of the Second Pacific Rim International Conference on Artificial Intelligence (1992), the International Conference on Signal Processing (1993), the International Conference on Computer Processing of Oriental Languages (1994), and the Third International Conference on Document Analysis and Recognition (1995). He is the general chairman of the 1997 International Conference on Computer Processing of Oriental Languages.

Hong Ma graduated from the Department of Mathematics, Sichuan University, Chengdu, China. From March 1984 to April 1986, as a visiting scholar, he visited the Department of Mathematics, Kobe University, Japan, where he performed research work in stochastic analysis. From April 1996 to July 1996, he worked at the Department of Computing Studies, Hong Kong Baptist University, on wavelet transformation and its application to pattern recognition. He is currently an associate professor in the Department of Mathematics and Department of Computer Science of Sichuan University and head of the Probability and Statistical Laboratory of the department. His research interests include pattern recognition, image processing, and stochastic signal processing.

Dihua Xi and **Xiaogang Mao** were students of Sichuan University who have left the institute; their biographies were not available.

Ching Y. Suen received an MSc (Eng.) degree from the University of Hong Kong and a PhD degree from the University of British Columbia, Canada. In 1972, he joined the Department of Computer Science of Concordia University, Montreal, Canada, where he became a professor in 1979 and served as chairman from 1980 to 1984. Presently, he is the director of CENPARMI, the Center for Pattern Recognition and Machine Intelligence of Concordia, and the associate dean—Research, of the Faculty of Engineering and Computer Science. During the past 15 years, he held visiting positions in several institutions in different countries.

Dr. Suen is the author/editor of 11 books on subjects ranging from computer vision and shape recognition, handwriting recognition, and expert systems to computational analysis of Mandarin and Chinese. Dr. Suen is the author of more than 250 papers, and his current interests include pattern recognition and machine intelligence, character recognition and expert systems, document processing, and computational linguistics.

An active member of several professional societies and fellow of the IEEE, IAPR, and the Royal Society of Canada, Dr. Suen is an associate editor of several journals related to pattern recognition and artificial intelligence. He founded *Computer Processing of Chinese and Oriental Languages*, an international journal of the Chinese Language Computer Society in 1983, and served as its editor-in-chief for 10 years.

During the past 15 years, Prof. Suen has served as chairman of the Character and Mark Recognition Committee of the Canadian Standards Association, which developed several Canadian standards on optical character recognition. He is the past president of the Canadian Image Processing and Pattern Recognition Society, governor of the International Association for Pattern Recognition, and past president of the Chinese Language Computer Society. He has been a consultant to numerous industrial companies, and has given more than 100 lectures at universities and in industries around the globe. He has organized many international conferences on subjects of his specialties, and was founder of the Vision Interface Conference in Canada, founder and chairman of the First International Workshop on Frontiers in Handwriting Recognition, cofounder and cochairman of the First International Conference on Document Analysis and Recognition (ICDAR), and general chairman of the Third ICDAR held in Montreal, August 1995.

Dr. Suen is the recipient of several awards, including the 1992 ITAC/NSERC award for outstanding contributions to pattern recognition, expert systems, and computational linguistics.