Sankhyā : The Indian Journal of Statistics
1999, Volume 61, Series B, Pt. 3, pp. 496–513

AN EFFICIENT APPROACH TO CONSISTENT SET ESTIMATION

By A. RAY CHAUDHURI
Visva Bharati, Santiniketan
A. BASU, S. K. BHANDARI, and B.B. CHAUDHURI
Indian Statistical Institute, Calcutta

SUMMARY. Determining the shape of a point pattern on the real plane is a problem of considerable practical interest and has applications in many branches of science. Set estimators of a nonparametric nature which may also be used as shape descriptors should have several desirable properties. The more important ones among them are the following: (a) The estimator should be consistent, i.e. the Lebesgue measure of the symmetric difference of the actual region and the set estimator should go to zero in probability as the number of sample points increase arbitrarily; (b) it should be computationally efficient; and (c) it should be automatic, in the sense that the method should be able to detect the number of independent disjoint components making up the true region and should not depend on this number being known. None of the currently known estimators combine all these properties.

Ray Chaudhuri et al. (1997) has introduced a shape descriptor called *s-shape* in the context of perceived border extraction of dot patterns. In this paper we develop a related idea to construct a class of set estimators which have all the three properties stated above. The emphasis of the paper is on establishing the consistency results of the proposed set estimator. It is shown that the *s-shape* is a consistent estimator not just under the uniform distribution, but also when the points are drawn according to any continuous distribution.

The method is illustrated with several examples, and the role of δ, the only parameter controlling the structure of the *s-shape* is discussed. Values of δ which appear to be intuitively and experimentally justified are proposed. A bound for the order of error is computed. Possible directions for future research are also mentioned.

1. Introduction

From an early stage of human endeavour one problem of interest is to find the shape of a point pattern. From astronomical studies to various application domains such as exploration of natural resources, urban planning, biomedical

Paper received August 1997; revised March 1999.
AMS (1991) *subject classification.* Primary 62G05; secondary 62P99, 68T10, 68U10
Key words and phrases. Consistent set estimator, dot pattern, shape description, s-shape, symmetric difference, Lebesgue measure, minimum spanning tree, digital image, mathematical morphology.

imaging, etc., the structural analysis of point set and shape recovery play an important role (R. Laurini and D. Thompson, 1992, J.C. Russ, 1995, J. Taylor, 1977). In $I\!R^2$ or $I\!R^3$ one can perceive the shape of the point set if the points are clearly visible as well as fairly densely and more or less evenly distributed. Such a point set is referred to as a *regular dot pattern*. Intuitively speaking, shape of a point pattern is the bounded region that generates the point pattern. In 1997, Ray Chaudhuri *et al.* has introduced a shape descriptor called *s-shape* in the context of perceived border extraction of dot patterns. The impression behind the *s-shape* is as follows: Let the pattern plane be partitioned by a lattice of square grids of 'appropriate' length. Consider the union the of grids containing points of the dot pattern. If the grid-length s is properly selected, this union or a 'smooth' version of this union approximates the underlying region of the pattern.

The shape description issue may be viewed as an associated problem of the more basic question of set estimation from a finite number of sample points drawn from the set. One major problem of interest in set estimation is *consistency*. An estimator is called consistent if it converges (in some appropriate sense) to the original set A as the number of points drawn from A tends to infinity. In this paper the theoretical properties of the *s-shape* as a set estimator for sets in $I\!R^2$ are studied. The spirit of the procedure is nonparametric in nature.

The paper is organized in the following order. In Section 2, we define consistency of a set estimator and discuss the existing results. Thereafter, the *s-shape* and its derivatives are formally defined. In Section 3.1, we establish the consistency of *s-shape* when the point set is generated by an uniform distribution on A. In Section 3.2 this result is extended to general continuous distributions under appropriate conditions via another theorem. In Section 4, our methods are applied to illustrate the effectiveness of *s-shape* as a set estimator. A general discussion on the proposed estimator is presented in the last section (Section 5). The advantages of *s-shape* and the directions for future work are also pointed out.

2. Definitions and Existing Results

2.1. *Set estimation and consistency.* In the following we discuss the existing results in the context of a point set in the 2-dimensional case.

DEFINITION 2.1. Let $X_1, X_2, ..., X_n$ be 2-dimensional independent and identically distributed (*i.i.d*) random vectors from a distribution which is supported on a set $\alpha \subset I\!R^2$. Let $\alpha^* \subset I\!R^2$ be an estimate of α based on $X_1, X_2, ..., X_n$. Then α_n^* is said to be a consistent estimator of α (denoted $\alpha_n^* \to \alpha$) if

$$\lim_{n\to\infty} E[\lambda(\alpha_n^* \Delta \alpha)] = 0 \qquad \ldots (2.1.1)$$

where λ is the 2-dimensional Lebesgue measure, Δ represents symmetric difference and E represents expectation.

THEOREM I. (Grenander's Theorem). *Let $\alpha \subset \mathbb{R}^2$ be a bounded Borel set whose boundary has Lebesgue measure 0. Let $\langle \epsilon_n \rangle$ be a sequence of positive numbers such that as $n \to \infty$, $\epsilon_n \to 0$ but $n\epsilon_n^2 \to \infty$. Let $\alpha_n^* = \bigcup_{i=1}^{n} \{X \mid \| X_i - X \| \leq \epsilon_n\}$ where $X_1, X_2, ..., X_n$ are i.i.d random vectors from a uniform distribution $\wp$ over α. Then $lim_{n\to\infty} E[\lambda(\alpha_n^* \Delta \alpha)] = 0$ where λ is the Lebesgue measure on $\mathbb{R}^2$ and '$\|\|$' represents the Euclidean norm (U. Grenander, 1975).*

Note that there are infinitely many different sequences of such ϵ_n's with the property that $n \to \infty, \epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$. Since the choice of ϵ_n does not depend on $X_1, X_2, ..., X_n$, Grenander's class of estimators do not have the scale equivariance property.

Another consistent set estimator based on the Minimum Spanning Tree (MST) is due to Murthy (1988). In this case, the radii ϵ_n's are made functions of $X_1, X_2, ..., X_n$ in the context of *compact regions* (T.M. Apostol, 1971)[1].

THEOREM II. (Murthy's Theorem) *Let $X_1, X_2, ..., X_n$ be i.i.d random vectors from a uniform distribution which is supported on a compact region α. Let l_n denotes the length of the MST($X_1, X_2, ..., X_n$) where $h_n = \sqrt{\frac{l_n}{n}}$. Let $\alpha_n^* = \bigcup_{i=1}^{n} \{X \mid \| X_i - X \| \leq h_n\}$. Then α_n^* is a consistent estimator of α.*

The above result is true for any continuous distribution. However, the result can not be extended to the case of union of multiple compact regions unless the number of disjoint components is known.

Note that the above two theorems basically take the union of certain circular neighborhoods centering every sample point (in Theorem I) or points over the MST of sample points (in Theorem II) as an estimate of the original set α.

2.2 *The s-shape.* Let S_n be a dot pattern containing n points in real plane. Let W_n be the *optimal* rectangle (rectangle with smallest area) with horizontal and vertical sides parallel to the rectangular coordinate axes of reference covering S_n i.e.$S_n \subset W_n \subset \mathbb{R}^2$. For a given grid side-length s_n, let $\mathcal{F}(s_n)$ denote a lattice of square grids on the real plane, with sides parallel to the coordinates axes.

Letting $g \in \mathcal{F}(s_n)$ denote a generic grid, define

$$\left.\begin{array}{l} \mathcal{G}(s_n) = \{g \mid g \cap W_n \neq \phi\} \\ G(s_n) = \bigcup\{g \mid g \in \mathcal{G}(s_n)\} \end{array}\right\} \quad \ldots (2.2.1)$$

[1] A set $\alpha \subset \mathbb{R}^2$ is said to be a compact region if α is path-connected, compact, *cls(int(α))*=α, and $\partial(\alpha)$ consists of finitely many rectifiable curves. (*cls, int* and ∂ denote respectively closure, interior and boundary).

$$\left.\begin{array}{l}\mathcal{H}(s_n) = \{g \mid g \cap S_n \neq \phi\} \\ H(s_n) = \bigcup\{g \mid g \in \mathcal{H}(s_n)\end{array}\right\} \quad \ldots (2.2.2)$$

Note that $G(s_n)$ denotes the set-union of grids over W_n while $H(s_n)$ denotes the subset of $G(s_n)$ by joining the grids each of which contains at least one point. Let P_n^H denote the number of grids in $H(s_n)$. Then the *area* of $H(s_n)$ is $\lambda(H(s_n)) = P_n^H \times s_n^2$.

DEFINITION 2.2. *The induced hull $H(s_n)$ with grid-length s_n is said to be an s-shape of S_n.*

By starting from the left-upper most grids, let the grids of $\mathcal{G}(s_n)$ be ordered in a 2-dimensional array. Then $\mathcal{G}(s_n)$ induces a *binary digital image* whose *(i,j)*-th location (pixel) represents the grid situated at *i*-th row, *j*-th column position. The *foreground* of this binary image is then generated by the grids of $\mathcal{H}(s_n)$. (For the definition of binary digital image etc., see the Appendix). As there is a one-to-one correspondence between the foreground and $\mathcal{H}(s_n)$, the foreground is also denoted by $\mathcal{H}(s_n)$.

Let this foreground be morphologically operated by a *binary closing* (R.M. Haralick, S.R. Sternberg and X. Zhuang, 1987) where the 3×3 *structuring element* having all entries equal to 1 and center of reference at the middle position is used (for definitions etc., see the Appendix). Let the resulting morphologically closed set which is superset of $\mathcal{H}(s_n)$, be denoted by $\bar{\mathcal{H}}(s_n)$.

Let $\bar{H}(s_n)$ denote the union of all grids whose corresponding *(i,j)*-th position belong to $\bar{\mathcal{H}}(s_n)$. Note that the closing 'smoothes' the set $H(s_n)$ from outside. Consider all holes as well as cusps (concavities on the boundary region) of $H(s_n)$ resulting from the empty grids. Among these, grids that can not be a part of 3×3 a lattice of grids completely in the *background*, are merged in $H(s_n)$ due to closing. For example, an empty grid having 5 non-empty grids in its 8-neighbours (east, north-east, north, north-west, west, south-west, south, south-east) becomes a part of $\bar{H}(s_n)$.

DEFINITION 2.3. $\bar{H}(s_n)$, which is a superset of $H(s_n)$ is said to be the smoothed induced hull of the latter.

In the next section, the consistency of $H(s_n)$ is first analyzed under an uniform distribution. Regarding the choice of s_n, a data driven procedure is proposed and the range where $H(s_n)$ remains consistent is established. The result is thereafter generalized to the case of arbitrary continuous distributions.

3. Consistency of the s-shape

3.1 *Points from a uniform distribution.* Consider any region α, a finite union of connected subregions in $\mathbb{R}^2$, each of which is bounded by a closed curve of finite length. Assume that the interior of α has a positive Lebesgue measure

while its boundary has Lebesgue measure 0. Also let W be the optimal (smallest area) rectangle with sides parallel to the coordinate axes such that α lies in the interior of W. Without loss of generality let the area (Lebesgue measure) of α be p, $(0 \leq p \leq 1)$ and the area of W be 1.

Let n points be chosen at random under the uniform distribution over the region α and the set of points be denoted by S_n. Let W_n be the optimal rectangle which covers this set of n points, with area $A_n \leq 1$. Consider the lattice of square grids $\mathcal{F}(s_n)$ on $\mathbb{R}^2$ with sides parallel to the coordinate axes where the grid side-length is, $s_n = n^{-\delta}\sqrt{A_n}$, $0 < \delta < 1$. Then there are approximately $n^{2\delta}$ grids in the sublattice $\mathcal{G}(s_n)$ which consists grids of $\mathcal{F}(s_n)$ intersecting W_n. For clarity of presentation, the following notations are used:

i. $P_n^T \equiv$ # of grids in $\mathcal{F}(s_n)$ intersecting α (some of them may not contain points of S_n).

ii. $T_n \equiv$ union of the above P_n^T grids intersecting α.

iii. $P_n^I \equiv$ # of grids among P_n^T completely in the interior of α.

iv. $I_n \equiv$ union of the above P_n^I grids completely in the interior of α.

v. $n_I \equiv$ # of points from S_n in the interior of I_n.

vi. $P_n^B \equiv$ # of grids in $\mathcal{F}(s_n)$ intersecting the boundary of α.

vii. $B_n \equiv$ union of the above P_n^B grids intersecting the boundary of α.

viii. $n_B \equiv$ # of points from S_n in the interior of B_n.

xi. $P_n^H \equiv$ # of grids in $\mathcal{F}(s_n)$ containing at least one point from S_n.

x. $H_n \equiv$ union of the above P_n^H non-empty grids.

Notice that $I_n \subset \alpha \subset T_n$, $B_n = T_n / I_n$ and $n_B + n_I = n$, while H_n is in fact *s-shape* of S_n. If s_n is chosen as above, we will show that the Lebesgue measure of the symmetric difference of H_n and α goes to 0 in probability as n tends to ∞ for appropriate choices of δ.

By the strong law of large numbers (SLLN) any subregion of α with positive Lebesgue measure eventually has a point chosen from it in S_n with probability 1 (*w.p.*1). Thus, as $n \to \infty$, we have in the sense of (2.1.1),

$$W_n \to W \qquad \ldots (3.1.1)$$

while with probability 1,

$$A_n \to 1 \qquad \ldots (3.1.2)$$

To prove $\lambda(\alpha \cap \tilde{H}_n) \to 0$ in probability ($\tilde{H}_n$ denotes the complement of H_n) we first look at the proportion of empty grids among the P_n^T grids intersecting the region α.

Since B_n approximates a one dimensional region with Lebesgue measure 0 (the boundary in question), and since the interior of α has positive Lebesgue measure ($=p$, the measure of α) we have

$$\lim_{n \to \infty} \lambda(B_n) = 0 \qquad w.p.1. \qquad \ldots (3.1.3)$$

In fact,

$$\lim_{n\to\infty} \frac{\lambda(I_n)}{\lambda(T_n)} = 1, \text{ and } \lim_{n\to\infty} \frac{\lambda(B_n)}{\lambda(T_n)} = 0 \qquad w.p.1 \qquad \ldots (3.1.4)$$

Similarly,

$$\lim_{n\to\infty} \frac{n_l}{n} = 1 \qquad w.p.1. \qquad \ldots (3.1.5)$$

Let $n_I = na_n$ where $\lim_{n\to\infty} a_n = 1$ $w.p.1$. By the above results, it is easily established that

$$\lim_{n\to\infty} \frac{P_n^I}{n^{2\delta}} = p \qquad w.p.1. \qquad \ldots (3.1.6)$$

Suppose that n balls are thrown at random in $P_n = \dfrac{n^\theta}{a}$ boxes, $0 < \theta < 2$, where a is a finite positive constant. Then the probability that a particular box remains empty is

$$\left(\frac{P_n - 1}{P_n}\right)^n = \left(1 - \frac{1}{P_n}\right)^n = \left(1 - \frac{a}{n^\theta}\right)^n \qquad \ldots (3.1.7)$$

Since the balls are thrown at random, this also represents the expected proportion of empty boxes. It is a standard result of calculus that

$$\lim_{n\to\infty} \left(1 - \frac{a}{n^\theta}\right)^n = \begin{cases} e^{-a}, & \theta = 1 \\ 0, & \theta < 1 \\ 1, & 1 < \theta < 2 \end{cases} \qquad \ldots (3.1.8)$$

Now, the expected proportion of empty grids among the P_n^I grids in the interior of the region α is

$$\left(1 - \frac{1}{P_n^I}\right)^{na_n} = \left(1 - \frac{n^{2\delta}}{P_n^I}\frac{1}{n^{2\delta}}\right)^{na_n} \qquad \ldots (3.1.9)$$

Since $\lim_{n\to\infty} \dfrac{n^{2\delta}}{P_n^I} = \dfrac{1}{p}$ $w.p.1$, the limit of expression in the right hand side of the above equation becomes $e^{-\frac{1}{p}}$ if $\delta = 0.5$; equals 0 if $\delta < 0.5$; and equals 1 if $0.5 < \delta < 1$.

Thus, for any choice of $\delta < 0.5$, the expected proportion of empty grids among the grids completely in the interior of the region α goes to 0 as n becomes arbitrarily large. Since the proportion of empty grids is a non-negative random variable, the proportion of empty grids also goes to 0 in probability by Markov's inequality. Also, by equation (3.1.4) as $\lim_{n\to\infty} \dfrac{P_n^B}{P_n^T} = 0$ $w.p.1$, the proportion of

empty grids among P_n^T is also 0 in the limit. Hence, $\lim_{n\to\infty} \frac{P_n^H}{P_n^T} = 1$ $w.p.1$ and H_n eventually covers α in probability. That is,

$$\lambda\left(\alpha \cap \tilde{H}_n\right) \to 0 \text{ in probability.} \qquad \ldots (3.1.10)$$

Conversely, $\lambda(H_n \cap \tilde{\alpha}) \leq \lambda((B_n \cup I_n) \cap \tilde{\alpha}) = \lambda(B_n \cap \tilde{\alpha}) + \lambda(I_n \cap \tilde{\alpha}) \leq \lambda(B_n)$.

Taking limit on both sides,

$$\lim_{n\to\infty} \lambda(H_n \cap \tilde{\alpha}) \leq \lim_{n\to\infty} \lambda(B_n) = 0 \qquad w.p.1. \qquad \ldots (3.1.11)$$

Combining (3.1.10) and (3.1.11) it is established that $\lambda(H_n \Delta \alpha) \to 0$ in probability. □

Since the above symmetric difference is a bounded random variable, from the above result the following theorem is established:

THEOREM III. *Let $X_1, X_2, ..., X_n$ be i.i.d random vectors drawn under a uniform distribution over the region α, a finite union of connected subregions in $\mathbb{R}^2$, each of which is bounded by a closed curve of finite length. Let W_n be an optimal rectangle covering $X_1, X_2, ..., X_n$ with area A_n. If $s_n = n^{-\delta}\sqrt{A_n}, (0 < \delta < 0.5)$ then the s-shape $H(s_n) = H_n$ is a consistent estimator of α.*

3.2 *Points from arbitrary continuous distributions.* Let $f\,(>0)$ be the continuous density function of the random variable over the two dimensional region α which is defined as in Section 3.1. Without loss of generality let the area (Lebesgue measure) of α be $p', (0 < p' < 1)$ and the area of W be 1.

Let $\wp(Q)$ be the probability of any subregion Q of α under the density f. Let n points be chosen from the region α at random under $\wp$ and the set of points be denoted by S_n.

Consider any $\epsilon > 0$. Then one can choose a large number m such that $\lambda(\alpha_m) > p' - \frac{\epsilon}{2}$ where

$$\alpha_m = \{x \mid x \in \alpha, \frac{1}{m} < f(x) < m\}. \qquad \ldots (3.2.1)$$

Let $\wp(\alpha_m) = p$. We assume that the boundary of the set α_m has Lebesgue measure 0. The other notations and terms used in the previous theorem are also used in this case with the only alteration that in (i.)-(v.) α is replaced by α_m. Let $H_{m,n}\,(\subseteq H_n)$ denote the union of non-empty grids intersecting α_m.

As justified in the last theorem, we have as $n \to \infty$, in the sense of (2.1.1),

$$W_n \to W \qquad \ldots (3.2.2)$$

while with probability 1,

$$A_n \to 1 \qquad \ldots (3.2.3)$$

Let a (sample) point Z be drawn from α. Then

$$P(Z \in I_n) = \int_{I_n} f(x)\,dx < m\lambda(I_n) \qquad \ldots (3.2.4)$$

For a given grid g among P_n^I,

$$P(Z \in g \mid Z \in I_n) = \frac{P(Z \in g)}{P(Z \in I_n)} > \frac{\frac{1}{m}\lambda(g)}{mP_n^I\lambda(g)} = \frac{1}{m^2 P_n^I} \qquad \ldots (3.2.5)$$

If n_I points are drawn from I_n then an upper bound of the probability of g to be empty can be found by the following equation.

$$P(g \cap S_n = \Phi \mid g \subset I_n; \#(I_n \cap S_n) = n_I) < \left(1 - \frac{1}{m^2 P_n^I}\right)^{n_I} \qquad \ldots (3.2.6)$$

By strong law of large numbers,

$$\lim_{n\to\infty} \frac{n_I}{n} = p \qquad w.p.1 \qquad \ldots (3.2.7)$$

Let $n_I = na_n$ where $\lim_{n\to\infty} a_n = p$ $w.p.1$.

For $i = 1, 2, \ldots, P_n^I$, let the characteristic function χ_i be defined as

$$\chi_i = \left.\begin{array}{ll} 1 & \text{if i-th grid is empty} \\ 0 & \text{otherwise} \end{array}\right\} \qquad \ldots (3.2.8)$$

As proportion of empty grids among P_n^I is $\frac{\sum_{i=1}^{P_n^I}\chi_i}{P_n^I}$, the *expected* proportion of empty grids among P_n^I is

$$E\left[\frac{\sum_{i=1}^{P_n^I}\chi_i}{P_n^I}\right] = \frac{\sum_{i=1}^{P_n^I}E(\chi_i)}{P_n^I} < \left(1 - \frac{1}{m^2 P_n^I}\right)^{n_I} = \left(1 - \frac{1}{m^2}\frac{n^{2\delta}}{P_n^I}\frac{1}{n^{2\delta}}\right)^{na_n} \qquad \ldots (3.2.9)$$

By (3.1.8) and (3.2.7) the limit of the expression in the right hand side of the above equation becomes $e^{-\frac{p}{m^2\lambda(\alpha_m)}}$ if $\delta = 0.5$; equals 0 if $\delta < 0.5$; and equals 1 if $0.5 < \delta < 1$.

Thus for any choice of $\delta < 0.5$, the proportion of empty grids among the grids completely in the interior of the region α_m goes to 0 for a sufficiently large n. Thus, as in Section 3.1,

$$\lambda\left(\tilde{H}_{m,n} \cap \alpha_m\right) \to 0 \text{ in probability .} \qquad \ldots (3.2.10)$$

As B_n, the union of the P_n^B grids which intersect the boundary of α, approximates a one dimensional region with Lebesgue measure 0 (the boundary in question) and interior of α has positive Lebesgue measure, we have

$$\lim_{n\to\infty} \lambda(B_n) = 0 \qquad w.p.1 \qquad \ldots (3.2.11)$$

For any given $\epsilon > 0$ and $0 < t < 1$, by (3.2.1), (3.2.10) and (3.2.11) we can choose M and N such that whenever $m \geq M$, $n \geq N$ and $\delta < 0.5$

$$\begin{aligned} P\left(\mid \lambda\left(\tilde{H}_n \cap \alpha\right) \mid < \epsilon\right) &\geq P\left(\mid \lambda\left(\tilde{H}_n \cap \alpha_m\right) \mid < \tfrac{\epsilon}{2}\right) \\ &\geq P\left(\mid \lambda\left(\tilde{H}_{m,n} \cap \alpha_m\right) \mid < \tfrac{\epsilon}{2}\right) \geq 1 - t. \end{aligned} \qquad \ldots (3.2.12)$$

As this is true for arbitrary ϵ and t, and since $\lambda(H_n \cap \tilde{\alpha}) \to 0$ in probability, $\lambda(H_n \Delta \alpha) \to 0$ in probability.

Thus from the above result the following theorem is established.

THEOREM IV. *Let $X_1, X_2, ..., X_n$ be i.i.d random vectors from a distribution $\wp$ such that its density function f is positive and continuous. Let the support of $\wp$ be α, a finite union of connected subregions in $\mathbb{R}^2$, each of which is bounded by a closed curve of finite length. Let W_n be an optimal rectangle covering $X_1, X_2, ..., X_n$ with area A_n. If $s_n = n^{-\delta}\sqrt{A_n}$, $(0 < \delta < 0.5)$ then the s-shape $H(s_n) = H_n$ is a consistent estimator of α.*

In general due to absence of small holes as well as sharp concavities, the smoothed induced hull $\bar{H}(s_n)$ is a better representation than $H(s_n)$ (see figures in Section 4). Since $\bar{H}(s_n)$ is a superset of $H(s_n)$, to establish the consistency of $\bar{H}(s_n)$ we have to concentrate only on the boundary error (limiting Lebesgue measure of $\bar{H}(s_n)$ in the complement of α). However, as $\lambda\left(\bar{H}(s_n) \cap \tilde{\alpha}\right) < 9 \times \lambda(H(s_n) \cap \tilde{\alpha})$, the boundary error may increase at most 9 times than that of $H(s_n)$. Thus, the consistency of the smoothed hull also follows.

3.3. *Error Rate.* It is crucial that the experimenter has an idea of the order of error (in terms of the Lebesgue measure of the symmetric difference) when the procedure is terminated at a particular value of n and the corresponding estimate α_n^* has been determined. Here we provide an upper bound to this error when the points are drawn under a uniform distribution. We consider the grids in the interior and boundary of α separately.

The error in the interior E_I, related to the proportion of the empty grids, is equal to

$$E_I = \lambda(I_m) \times \left(1 - c_1 \frac{1}{n^{2\delta}}\right)^{c_2 n} \qquad \ldots (3.3.1)$$

where c_1 and c_2 are positive constants.

By taking the logarithm of the R.H.S of (3.3.1), expanding $log\left(1 - c_1 \frac{1}{n^{2\delta}}\right)^{c_2 n}$, and exponentiating back the leading term is found to be $c_3 e^{-n^{(1-2\delta)}}$ where c_3 is a positive constant.

The error in the boundary E_B, satisfies

$$E_B \leq P_n^B \times n^{-2\delta} \leq \left(\frac{\text{length of } \partial(\alpha)}{n^{-\delta}} \right) \times n^{-2\delta} \leq k_1 n^{-\delta} \qquad \ldots (3.3.2)$$

where k_1 is a positive constant.

Note, the error in the boundary dominates that in the interior. Thus, the error in estimation is of order $O\left(n^{-\delta}\right)$ or smaller.

4. Experimental Results

4.1. *Digital Domain Implementation.* For visualizing the effectiveness of our proposed consistent estimators in practice, we have implemented *s-shape* in digital domain. In the absence of *isolated object pixels*, the *foreground* in a *digital image* (for definitions see the Appendix) might be considered as α. Random samples of n object pixels are taken. Their *s-shapes* $H(s_n)$, constitute α_n^* for different values of n. The *area* of $\alpha, \lambda(\alpha)$ is measured by the total number of object pixels in α.

In Fig.1(a) we have illustrated a binary digital image where the foreground has a fish like shape. The cardinality of this foreground is 63900. We draw samples of point sets [Fig.1(b), (c), (d)] for $n = 100\,(\approx 0.0015 \times \#\alpha)\,, 1500\,(\approx 0.023 \times \#\alpha)$ and $3000\,(\approx 0.047 \times \#\alpha)$ respectively. For $\delta = 0.45$, their *s-shape* are in Fig.2(a)-(c) and their smooth versions in Fig.2.(d)-(f). Corresponding figures for $\delta = 0.49$ are shown in Fig.3.

The plots of $\dfrac{\lambda(\alpha \Delta \alpha_n^*)}{\lambda(\alpha)}$ where $\alpha_n^* = H(s_n)$ against the number of sample observations for $\delta = 0.45$ and $\delta = 0.49$ are presented in Fig.4(a) and Fig.4(b) respectively. The asymptotic convergence of α_n^* is readily understood in spite of the limitation due to quantization effect and finite state computation. As the smoothed induced hull $\bar{H}(s_n)$ is in general, a better representation of the shape of a dot pattern than the ordinary *s-shape* (A. Ray Chaudhuri, B.B. Chaudhuri and S.K. Parui, 1997), we have also plotted the same for $\alpha_n^* = \bar{H}(s_n)$. The asymptotic convergence of α_n^* for $\alpha_n^* = \bar{H}(s_n)$ and its behavior with different values of δ is another interesting problem. Notice that this smoothing leads to a substantial improvement for the case $\delta = 0.49$, but not for $\delta = 0.45$ (Fig.4).

Fig.5 and Fig.6 present two other examples where our estimator is applied. The disconnected components as well as the hole are correctly recovered. The number of points in the samples of the respective images are 300 and 1500. These figures represent smoothed *s-shapes* with δ fixed at 0.49.

4.2 *Choice of* δ. It is clear that the choice of δ has considerable impact on the resulting *s-shape* (see Fig.2 and Fig.3). For smaller values of δ, the boundaries of α_n^* are cruder – so much so that the *s-shapes* for δ in the range (0, 0.45) appear to be of little practical utility. For larger values of δ, on the other hand,

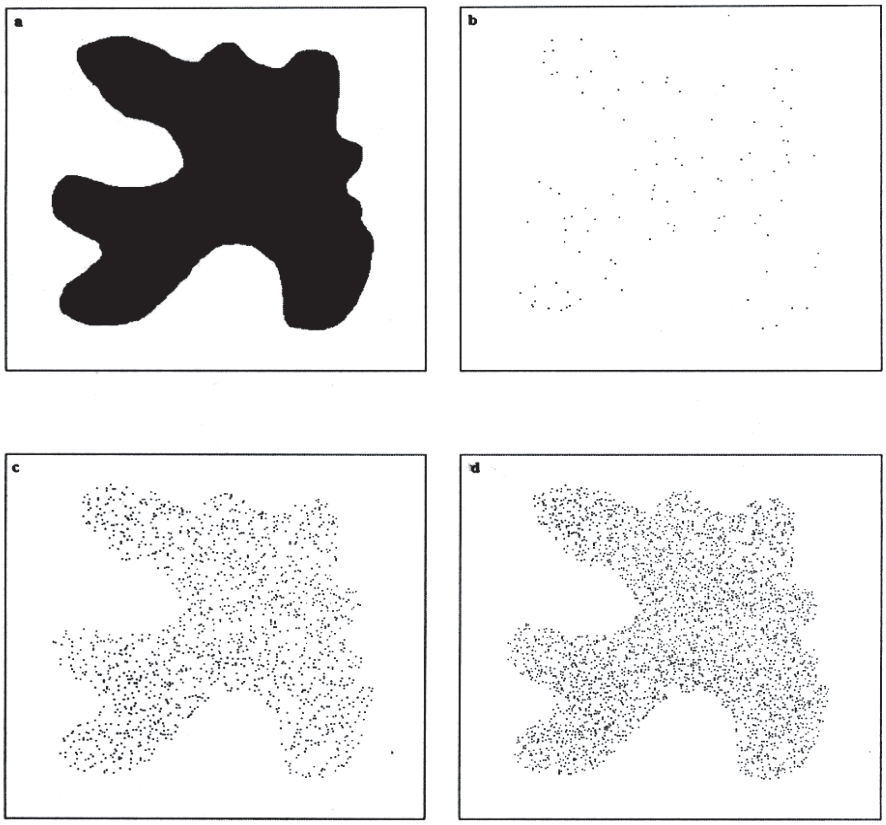

Figure 1

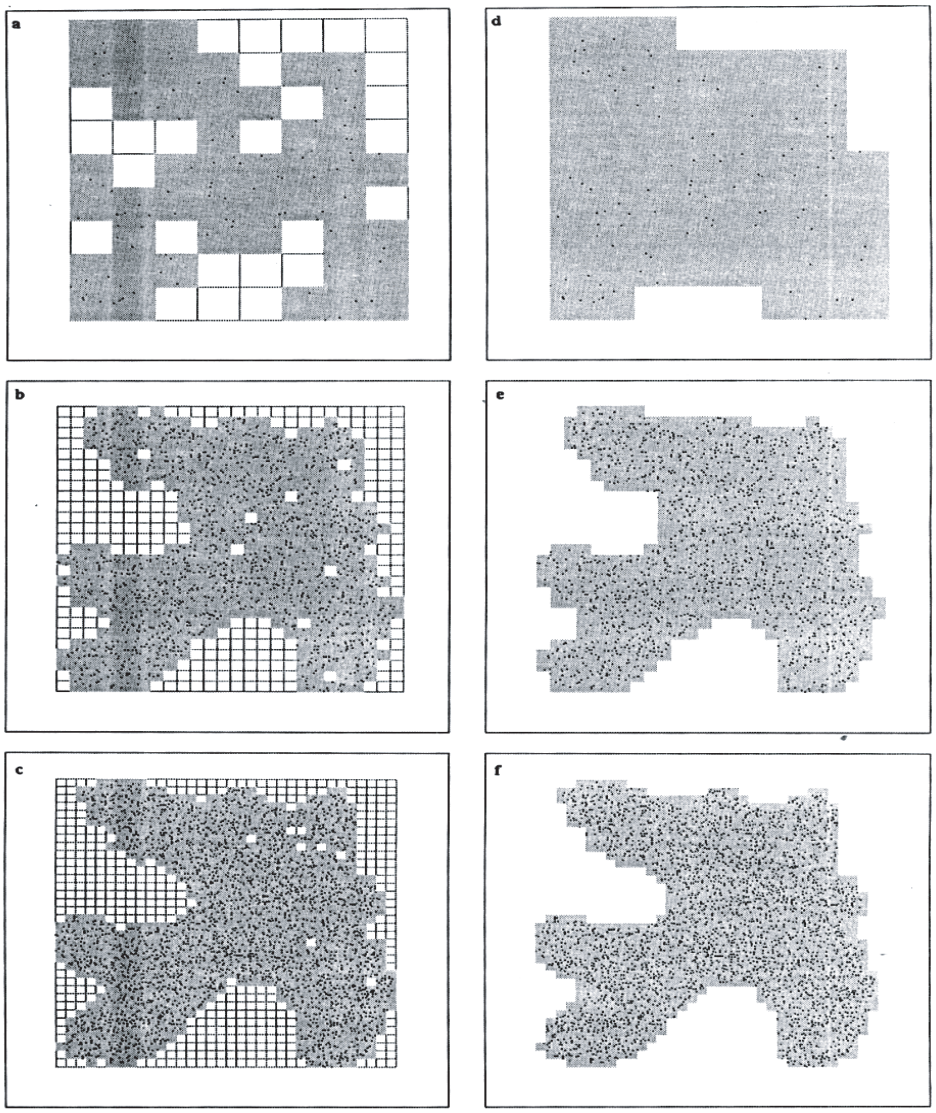

Figure 2.

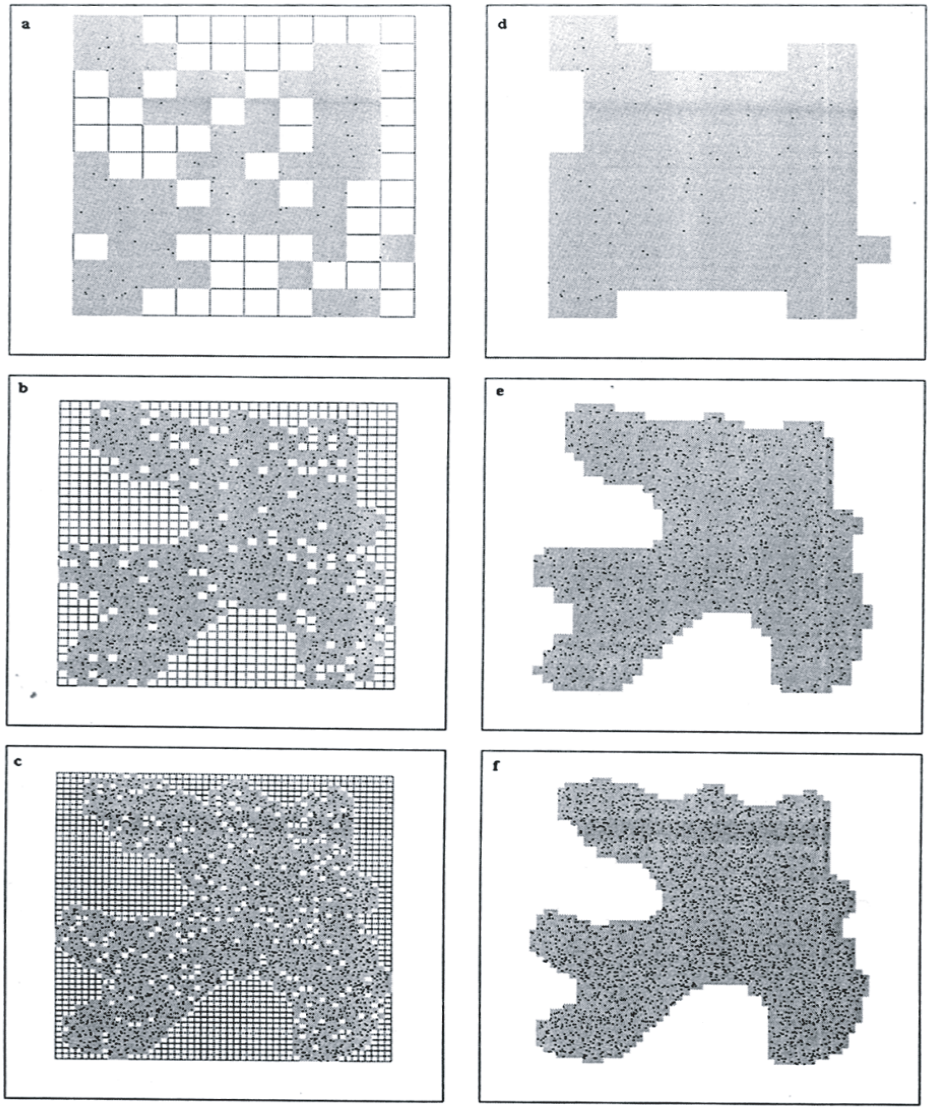

Figure 3

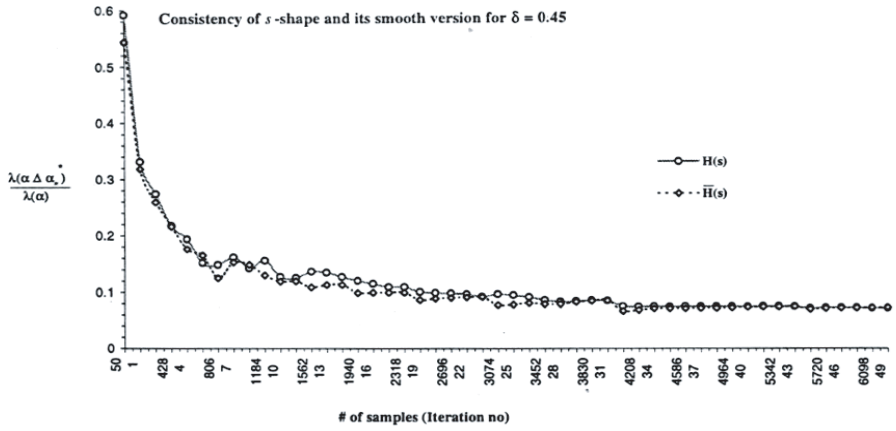
Consistency of s -shape and its smooth version for δ = 0.45
λ(α Δ α*n) / λ(α)
H(s)
H̄(s)
# of samples (Iteration no)

Fig. 4(a)

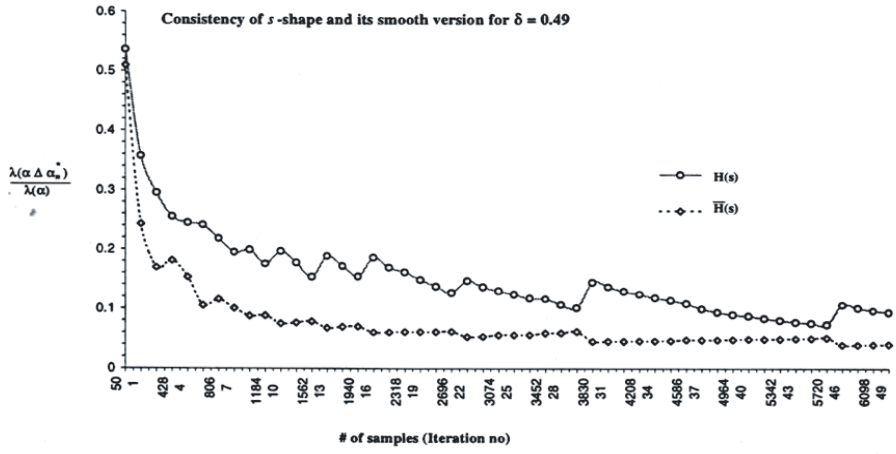
Consistency of s -shape and its smooth version for δ = 0.49
λ(α Δ α*n) / λ(α)
H(s)
H̄(s)
# of samples (Iteration no)

Fig. 4(b)

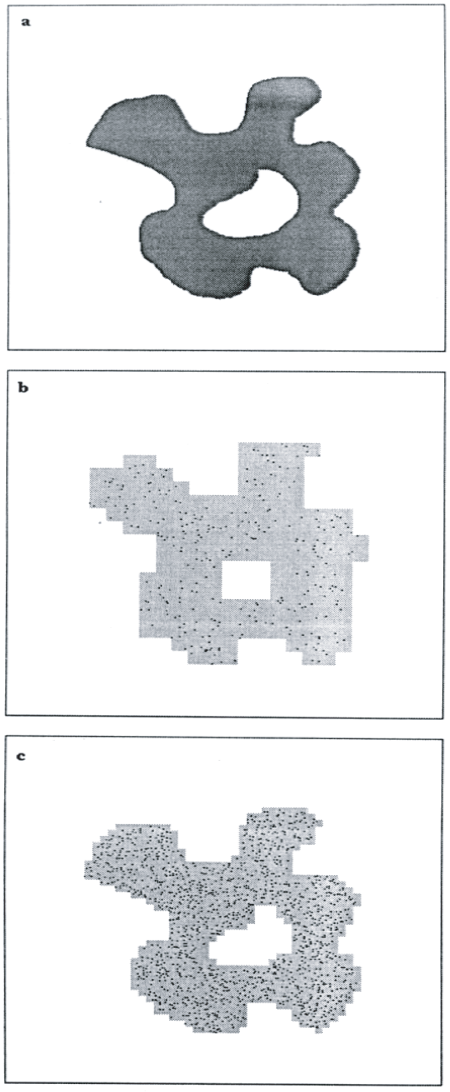

Fig. 5

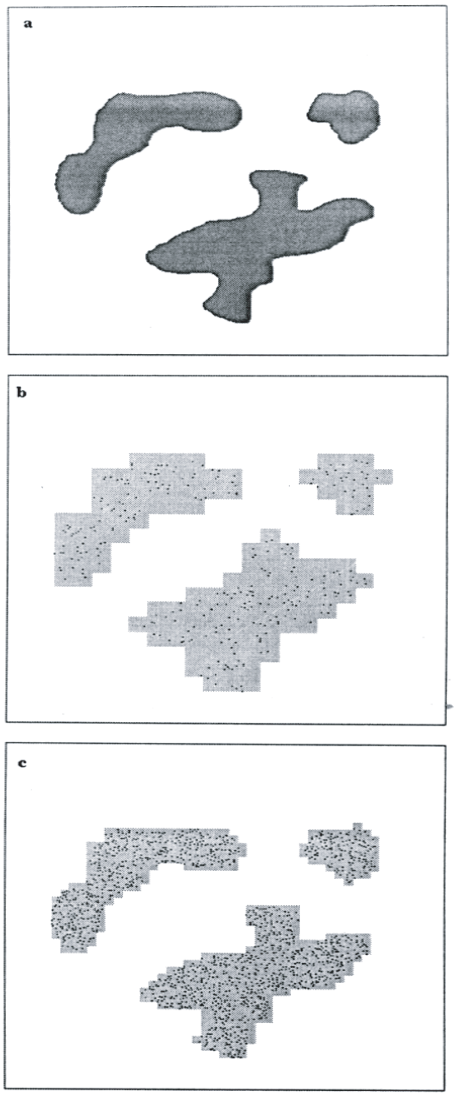

Fig. 6

the figure exhibits larger number of inconsistent holes (compare Fig.2 and Fig.3). In the particular case $\delta = 0.5$, the proportion of the area formed by the union of these holes with respect to the area of the region under estimation converges to a fixed non-zero constant so that consistency fails to hold. This suggests that 'smoothing' may be more useful for *s-shapes* obtained with values of δ close to 0.5. For a given n, it also appears that the larger values of δ lead to small values of $\lambda(\alpha_n^* \cap \tilde{\alpha})$ and smaller values of δ lead to small values of $\lambda(\tilde{\alpha}_n^* \cap \alpha)$- i.e. values of δ near opposite ends of the allowable range are more efficient in reducing opposite components of the symmetric difference.

We have experimentally found that [0.45, 0.50) can be considered as the allowable range for δ. On the whole, it appears that when single values of δ have to be recommended, then values of δ around the center of the allowable range or closer to 0.5 should be chosen (this is because larger values of δ reduce the error in the boundary, the dominant error). When coupled with smoothing, values of δ close to 0.5 should be chosen.

5. Discussion

The effectiveness of *s-shape* as a set estimator has been illustrated in the above examples and results in fair detail. From these examples it can be seen that the smoothed version of the *s-shape* can be viewed as a descriptor of the shape of the underlying region.

Our proposed consistent set estimator is totally different from the existing consistent estimators. As mentioned earlier, Grenander's consistent estimator as well as Murthy's consistent estimator are constructed by dilating sample points itself or the edges of MST of the sample points by a certain structuring disk. In our case, the optimal rectangular zone covering the sample set is partitioned by a lattice. The union of grids of the lattice which are non-empty in terms of sample points, is taken as α_n^*.

A major advantage of our estimators is their computational efficiency. It is linear in terms of cardinality of the point set. That is, for n observations the order of computational complexity is $O(n)$. The derivation is straight forward (Ray Chaudhuri, 1997). Note that for the other MST based computable estimator, the sole construction of the MST takes $O(nlog\ n)$ provided sophisticated data structures are used. Thereafter, the MST has to be dilated by a structuring disk.

The proposed estimator is fully unsupervised. The disconnected components are correctly detected and estimated as n increases. But the MST based estimator fails to do so. It needs prior knowledge about the number of disconnected components in α.

With a comparable consistent estimator in practice, one needs to have an idea of the order of error (in terms of the Lebesgue measure of the symmetric difference) when the procedure is terminated. We have provided an upper

bound to this error which may be used by practitioners as a guiding measure in determining a *stopping criterion*. Note that stopping criteria are unavailable for other existing estimators.

The consistency of *s-shape* based class of set estimators in higher dimensions ($\mathbb{R}^k$) will be established in a sequel paper. This extension, while not entirely straightforward, proceeds in similar lines to the present proofs. For example, the grid-size has to be chosen in the order of k-th root of the volume of optimal hyper-rectangle covering the dot pattern. In addition, the dependence of the error rate on dimensionality k will also be studied.

Appendix

Consider a compact bounded region R in Euclidean space $\mathbb{R}^2$. A set E in discrete integer lattice $\mathcal{T}^2$ is said to be the *binary digital image* of R if $\tilde{E} = \mathcal{T}^2/E,\ E \subseteq R$ and $\tilde{E} \cap R = \phi$.

An *object (non-object) pixels* corresponds to a discrete location *(i,j)* containing a value of 1(0) in the image. The set of object (non-object) pixels is referred to as *foreground (background)*. A binary image can be uniquely defined by the foreground (background).

Mathematical morphology is a methodology for image analysis (Haralick *et al.*, (1987, J. Serra, 1982). The principle of all basic operators is to probe the image under study with a *structuring element*. The structuring element is a set of points on which an origin is defined. To evaluate the results of a morphological operation on a image point, the structuring element is translated in such a way that its origin coincides with this image point. There are two basic operations in binary Mathematical Morphology: dilation and erosion. other operators are derived from these two operations.

Let A and B be any two bounded compact subsets in a normed space I and let t be a point in I (In our application, I is the discrete digital plane $\mathcal{T}^2$). Here B by which A is morphologically operated is considered as the structuring element.

The *dilation* and *erosion* of A and B, denoted by $A \oplus B$ and $A \Theta B$ respectively, are defined as

$$A \oplus B = \{p \in I \mid p = a + b; (a \in A) \wedge (b \in B)\}$$

$$A \Theta B = \{p \in I \mid (p + b) \in A \ \ \forall\ b \in B\}$$

Note that the dilation is the Minkowski addition of A by B; whereas the erosion is the Minkowski subtraction of $\breve{B}$ from A. ($\breve{B}$ is the reflection of B i.e. $\breve{B} = \{p \mid -p \in B\}$)

A set of transformation $\Psi(*)$ is said to be a *morphological filter* if for any two sets A and A' in the domain of transformation,

$$A \subseteq A' \Rightarrow \Psi(A) \subseteq \Psi(A') \quad (Increasing)$$

$$\Psi(\Psi(A)) = \Psi(A) \quad (Idempotence)$$

The *closing* of A by B, denoted by $A \bullet B$ is defined as $A \bullet B = (A \oplus B)\Theta B$ is a morphological filter which is also *extensive* in the sense that $A \subseteq A \bullet B$. All background structures that can not contain the structuring element are added to the set by the closing. In this sense the closing operator 'smoothes' the set from outside.

Acknowledgment. The authors appreciate the comments of two anonymous referees which have led to substantial improvement over the original version of the manuscript.

References

Apostol, T.M., (1971). *Mathematical analysis*, Addison Wesley.

Grenander, U., (1975). *Abstract inference*, John Wiley, New York.

Haralick, R.M., Sternberg, S.R. and Zhuang, X., (1987). *'Image Analysis Using Mathematical Morphology'*, IEEE Pattern Analysis and Machine Intelligence, **PAMI-9**,523-550.

Laurini, R. and Thompson, D., (1992). *Fundamental of spatial information systems*, The A.P.I.C. Series No. 37, Academic Press, London.

Murthy, C.A., (1988). *'On consistent estimation of classes in $\mathbb{R}^2$ in the context of cluster analysis'*, Ph.D. Thesis, Indian Statistical Institute, Calcutta.

Ray Chaudhuri, A., Chaudhuri, B.B. and Parui, S.K., (1997). *'A novel approach to computation of the shape of dot pattern and extraction of its perceptual border'*, CVGIP: Computer Vision And Image Understanding, Vol. 68, No. 3, 257–275.

Russ, J.C., (1995). *The image processing handbook*, C.R.C. Press, Ann Arbor.

Serra, J., (1982). *Image Analysis and Mathematical Morphology*, Academic Press Inc., New York.

Taylor, J., (1977). *Quantitative methods in geography : An introduction to spatial analysis*, Houghton Mifflin Company, Boston.

A. Ray Chaudhuri
Department of Computer and
System Sciences
Visva-Bharati, Santiniketan 731235
India
e-mail : anirban@vbharat.ernet.in

A. Basu
Applied Statistics Unit
Indian Statistical Institute
203 B.T. Road, Calcutta 700 035
India
e-mail : ayanbasu@isical.ac.in

S.K. Bhandari
Stat-Math Unit
Indian Statistical Institute
203 B.T. Road, Calcutta 700 035
India
e-mail : subir@isical.ac.in

B.B. Chaudhuri
Computer Vision and Pattern
Recongnition Unit
Indian Statistical Institute
203 B.T. Road, Calcutta 700 035
India
e-mail : bbc@isical.ac.in