Digital geometric methods in document image analysis

Ari Gross[a],*, Longin Jan Latecki[b]

[a]*Department of Computer Science, Computational Vision and Perception Lab, Graduate Center and Queens College, CUNY, Flushing, NY 11367, USA.*
[b]*Department of Computer Science, University of Hamburg, Vogt-Kölln-Str. 30, 22527 Hamburg, Germany*

Abstract

One of the main tasks of digital image analysis is to recognize the properties of real objects based on their digital images. These images are obtained by some sampling device, like a CCD camera, and are represented as finite sets of points that are assigned some value in a gray level or color scale. A fundamental question in image understanding is which features in the digital image correspond, under a given set of conditions, to certain properties of the underlying objects. In many practical applications this question is answered empirically by visually inspecting the digital images. In this paper, we present a mathematically comprehensive answer to this question with respect to topological properties. In particular, we derive conditions relating properties of real objects to the grid size of the sampling device which guarantee that a real object and its digital image are topologically equivalent. These conditions also imply that two digital images of a given object are topologically equivalent. Moreover, we prove that a topologically invariant digitization must result in well-composed or strongly connected sets and that only certain local neighborhoods are realizable for such a digitization. Using the derived topological model of a well-composed digital image, we demonstrate the effectiveness of this model with respect to the digitization, thresholding, correction, and compression of digital document images. © 1999 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.

Keywords: Digital topology; Digital geometry; Topologically invariant digitization; Document image analysis; Well-composed sets; Topological compression

1. Introduction

In image processing and in spatial knowledge representation, continuous objects are represented as finite sets (also called discrete sets), since only finite structures can be handled on computers. Continuous objects and their spatial relations can be characterized using geometric features. Therefore, any useful discrete representation should model the geometry faithfully in order to avoid wrong conclusions. A basic part of geometry is topology.

It is clear that no discrete model can exhibit all the features of the continuous original. Therefore, one has to accept compromises. The compromise chosen depends on the specific application and on the questions which are typical for that application. Digital geometry can be seen as an attempt to evaluate the price one has to pay for discretization. Digital topology is the theoretical basis for understanding topological features of objects in digital images, which must be related to features of the underlying continuous objects.

---

*Corresponding author.

(a) A real object   (b) Its digitization represented as   (c) Its digitization represented
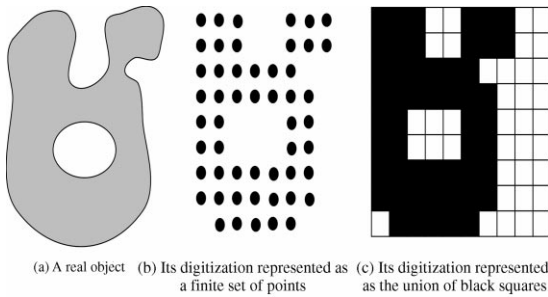   a finite set of points   as the union of black squares

Fig. 1. An object and its digital image.

In digital image processing, properties retrieved from the digital images are assumed to represent properties of the underlying real objects. Practical applications show that this is not always the case. Therefore, we want to know under what conditions do certain digital properties represent actual properties of the real object? In this paper, the authors introduce conditions that guarantee that a real object and its digital image are topologically equivalent. These results further extend the work of the authors as described in Refs. [1–5]. Due to space constraints, the basic results are presented in this paper while the detailed proofs are omitted here, but can be found in Refs. [1,5]. This paper also considers how these results in digital topology are directly useful in the area of document image processing and analysis.

In order to study the topological equivalence of a real continuous object and its digitization, which is a finite set of points, some preliminaries are in order. It is intuitively clear that the real object in Fig. 1a and the digital object in Fig. 1b have the "same topological structure". Based on the technical properties of sampling devices like a CCD camera, digital points representing sensor output are generally assumed to form a square grid and are modeled as points with integer coordinates located in the plane $\mathbb{R}^2$. By a digitization process, these points are assigned some gray level or color values. By a segmentation process, the digital points are grouped to form digital objects. For example, the digital points are grouped by thresholding gray-level values with some threshold value, i.e., the pixels whose gray-level values are greater than some given threshold value are classified as belonging to a digital object (i.e. assigned the color black). As an output of a digitization and segmentation process, we obtain a binary digital picture, with black points representing the digital object and white points representing the background.

We will identify each black point with a square centered at this point (in such a way that the squares form a uniform cover of the plane). A digital object is then represented as a union of squares which form a subset of the plane. For example, the digital set in Fig. 1b, a finite subset of $\mathbb{Z}^2$, is identified with the union of black squares in Fig. 1c, a subset of $\mathbb{R}^2$. Real objects or their projections are modeled in computer vision as subsets of the plane. Therefore, it makes sense to speak about topological equivalence between a real object (Fig. 1a and its digital image (Fig. 1c). Thus, the digitization (and segmentation) process is modeled as a mapping from continuous 2D sets representing real objects to discrete sets represented as finite subsets of $\mathbb{Z}^2$, which are identified with finite unions of squares in $\mathbb{R}^2$. Consequently, we can relate topological properties of a continuous 2D object (e.g., a projection of a 3D object) to its digital images interpreted as the union of squares centered at black points.

Serra [6] considered many kinds of digitizations. He showed that, for a certain class of planar sets, digitizations preserve homotopy. However, he proved this only for subset digitizations in hexagonal grids, where a *subset digitization* of a set A in $\mathbb{R}^2$ is the set of points in $\mathbb{Z}^2$ which are contained in A.

In this paper we derive conditions relating properties of continuous objects to the diameter of a square in the grid. If these conditions are satisfied, then the digital object obtained by this digitization (and segmentation) process is guaranteed to be topologically equivalent to the underlying continuous object.

2. Parallel regular sets

In this section we define a class of subsets of the plane representing "real objects", which we will call parallel regular sets. Let A be a planar set. We denote by A^c the complement of A, by $bd\,A$ the topological boundary of A, by $int\,A$ the topological interior of A and by $cl\,A$ the topological closure of A in the usual topology of the plane determined by the Euclidean metric. The connected components of the boundary $bd\,A$ are called *contours*. We denote by $d(x, y)$ the Euclidean distance of points x, y and by $B(c, r)$ a closed ball of radius r centered at a point c.

The following definition of parallel regular sets is based on the classical concepts in differential geometry of osculating balls and normal vectors, which we define below without using derivatives and limit points.

Definition. We will say that a closed ball $B(c, r)$ is *tangent* to $bd\,A$ at point $x \in bd\,A$ if $bd\,A \cap bd(B(c, r)) = \{x\}$.

We will say that a closed ball $iob(x, r)$ of radius r is an *inside osculating ball* of radius r to bdA at point $x \in bd\,A$ if $bd\,A \cap bd(iob(x, r)) = \{x\}$ and $iob(x, r) \subseteq int\,A \cup \{x\}$ (see Fig. 2).

We will say that a closed ball $oob(x, r)$ of radius r is an *outside osculating ball* of radius r to $bd\,A$ at point $x \in bd\,A$ if $bd\,A \cap bd(oob(x, r)) = \{x\}$ and $oob(x, r) \subseteq A^c \cup \{x\}$ (see Fig. 2).
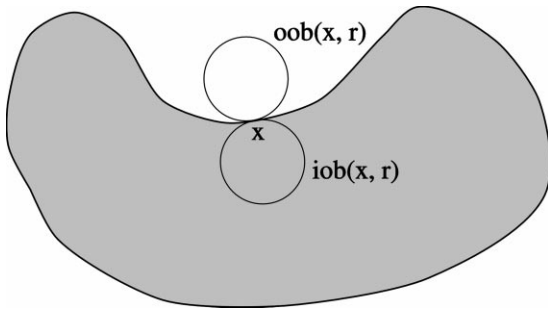
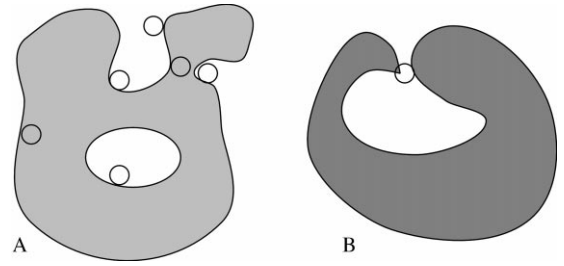Fig. 2. The inside and outside osculating balls of radius r to the boundary of the set A at point x.

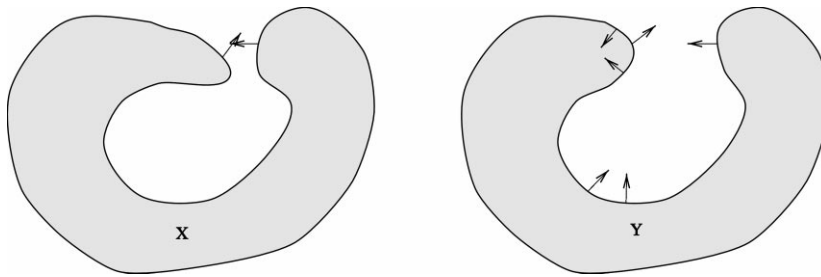Fig. 3. The set A is par(r)-regular while the set B is not par(r)-regular, where r is the radius of the depicted circles.

Fig. 4. X is not par(r)-regular, but Y is par(r)-regular.

Note that x is a boundary point, not the center, of both $iob(x, r)$ and $oob(x, r)$. For example, for every boundary point of a given ball $B(c, s)$ of radius s, there exist inside osculating balls of radii r, where $0 < r < s$. However, $B(c, s)$ itself is not an inside osculating ball for any of its boundary points. Now we define parallel regular subsets of the plane:

Definition. We assume that A is a closed subset of the plane such that its boundary $bd\ A$ is compact.

- A set A will be called *par(r, +)-regular* if there exists an outside osculating ball $oob(x, r)$ of radius r at every point $x \in bd\ A$.
- A set A will be called *par(r, −)-regular* if there exists an inside osculating ball $iob(x, r)$ of radius r at every point $x \in bd\ A$.
- A set A will be called *par(r)-regular* (or *r parallel regular*) if it is par(r, +)-regular and par(r, −)-regular. A set A will be called *parallel regular* if there exists a constant r such that A is par(r)-regular. We will sometimes call parallel regular sets (*spatial*) *objects*.

In Fig. 3 the set A is par(r)-regular while the set B is not par(r)-regular, where r is the radius of the depicted circles. Note that a parallel regular set, as well as its boundary, does not have to be connected.

In the remaining part of this section, we state some basic properties of parallel regular sets.

We have the following theorem:

Theorem 1. *A set A is par(r)-regular iff, for every two distinct points x, $y \in bd\ A$, the outer normal vectors $n(x, r)$ and $n(y, r)$ exist and they do not intersect, and the inner normal vectors $− n(x, r)$ and $− n(y, r)$ exist and they do not intersect.*

For example, in Fig. 4, set X is not par(r)-regular while set Y is par(r)-regular, where r is the length of the depicted vectors.

Definition. $B(x, r)$ denotes the closed ball of radius r centered at a point x. The *parallel set* of set $A \subseteq \text{IR}^2$ with distance r is given by

$$Par(A, r) = A \cup \bigcup \{B(x, r) : x \in bd\ A\}.$$

This set is also called a *dilation* of A with radius r. We define

$$Par(A, − r) = cl(A \setminus \bigcup \{B(x, r) : x \in bd A\})$$

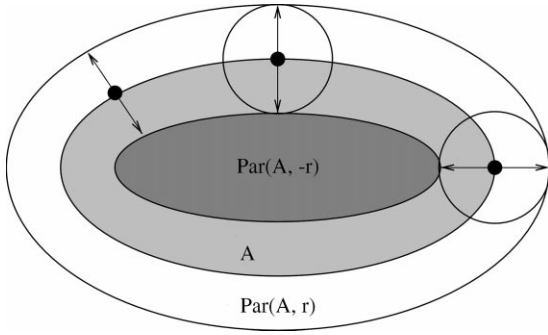For illustration, see Fig. 5. The boundaries of $Par(A, r)$ and $Par(A, − r)$ sets are often called offset curves.

Fig. 5. The set A and its parallel sets $Par(A, r)$ and $Par(A, -r)$.

It can be shown that $A = Par(Par(A, r), -r) = Par(Par(A, -r), r)$ for a par(r)-regular set A. Thus, a *par(r)-regular* set A is invariant with respect to morphological operations of opening and closing with a closed ball of radius r as a structuring element (see Ref. [6] for definitions). The following proposition motivates the name of parallel regular sets.

Proposition 1. *Let A be a par(r)-regular set. Then (see Fig.* 5)

$$Par(A, r) = A \cup \bigcup \{n(a, r) : a \in bd\,A\} \text{ and}$$

$$Par(A, -r) = cl(A \setminus \bigcup \{-n(a, r) : a \in bd\,A\}).$$

Proposition 2. *Let A be a par(r)-regular set. If x and y belong to two different components of $bd\,A$, then $d(x, y) > 2r$.*

We can use parallel sets to define the Hausdorff distance of planar sets:

Definition. *Hausdorff distance d_H of two planar sets A and B is given by*

$$d_{\mathrm{H}}(A, B) = \inf\{r \geqslant 0 : A \subseteq Par(B, r) \text{ and } B \subseteq Par(A, r)\}.$$

3. Digitization and segmentation-preserving topology

Let $\mathscr{Q}$ be a cover of the plane with closed squares of diameter r such that if two squares intersect, then their intersection is either their common side or a corner point. A *digital image* can be described as a set of points that are located at the centers of the squares of a grid $\mathscr{Q}$ and that are assigned some value in a gray level or color scale. By a *digitization process* we understand a function mapping a planar set X to a digital image. By a *segmentation process* we understand a process grouping digital points to a set representing a digital object. Therefore, the out-

put of a segmentation process can be interpreted as a binary digital image, where each point is either black or white. We assume that digital objects are represented as sets of black points. Thus, the input of a digitization and segmentation process is a planar set X and the output is a binary digital image, which will be called a *digitization* of X with diameter r and denoted $Dig(X, r)$.

In the remainder of this paper, we will interpret a black point $p \in Dig(X, r)$ as a closed (black) square of cover $\mathscr{Q}$ centered at p and the digitization $Dig(X, r)$ as the union of closed squares centered at black points, i.e., $Dig(X, r)$ will denote a closed subset of the plane.

We will treat digitization and segmentation processes satisfying the following conditions relating a planar par(r)-regular set X to its digital image $Dig(X, r)$:

ds1 If a square $q \in \mathscr{Q}$ is contained in X, then $q \in Dig(X, r)$ (i.e., q is black).
ds2 If a square $q \in \mathscr{Q}$ is disjoint from X, then $q \notin Dig(X, r)$ (i.e., q is white).
ds3 If a square q is black and $area(X \cap q) \leq area(X \cap p)$ for some square $p \in \mathscr{Q}$, then square p is black.

These conditions describe a standard model of the digitization and segmentation process for CCD cameras if we exclude digitization errors. In the following, we define some important digitization and segmentation processes satisfying the conditions ds1, ds2, and ds3 above.

Definition. Let X be any set in the plane. A square $p \in \mathscr{Q}$ is black (belongs to a digital object) iff $p \cap X \neq \emptyset$, and white otherwise. We will call such a digital image an *intersection digitization* with diameter r of set X, and denote it with $Dig_\cap(X, r)$, namely $Dig_\cap(X, r) = \bigcup \{p \in \mathscr{Q} : p \cap X = \emptyset\}$. See Fig. 6a, for example, where the union of all depicted squares represents the intersection digitization of an ellipse. With respect to real camera digitization and segmentation, the intersection digitization corresponds to the procedure of coloring a pixel black iff there is part of the object A in the field "seen" by the corresponding sensor.

Now we consider digitizations corresponding to the procedure of coloring a pixel black iff the object X fills the whole field "seen" by the corresponding sensor. For such digitizations, a square p is black iff $p \subseteq X$ and white otherwise. We will refer to such a digital image of a set X as a *square subset digitization*[1] and denote it by $Dig_\subset(X, r)$, where $Dig_\subset(X, r) = \bigcup \{p \in \mathscr{Q} : p \subseteq X\}$. In Fig. 6b, the two squares represent $Dig_\subset(X, r)$, where X is an ellipse.

---

[1] Observe that this digitization differs from subset digitization used by Serra and Pavlidis, where a square p is black iff its center point is contained in X (see Section 1).
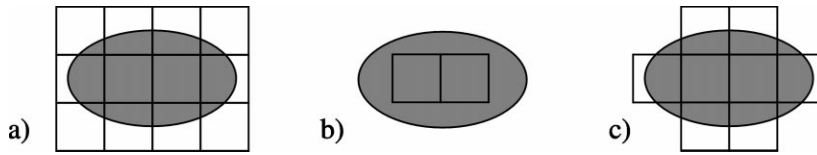
Fig. 6. (a) The union of all squares represents an intersection digitization of the ellipse. b) The two squares represent a square subset digitization of the ellipse. c) The eight squares represent a digitization of an ellipse with the area ratio equal to 1/5.

Next, let us consider a digitization and segmentation process in which a pixel is colored black iff the ratio of the area of the continuous object in a sensor square to the area of the square is greater than some constant threshold value v. An example is given in Fig. 6c, where the squares represent a digitization of the ellipse with the ratio equal to 1/5. This process models a segmentation by applying a threshold value to a gray-level digital image for all real devices in which the sensor values can be assumed to be monotonic with respect to the area of the object in the sensor square.

In the following, we briefly review the concept of homotopy equivalence.

Definition. Let X and Y be two topological spaces. Two functions $f, g: X \to Y$ are said to be *homotopic* if there exists a continuous function $H: X \times [0, 1] \to Y$, where $[0, 1]$ is the unit interval, with $H(x, 0) = f(x)$ and $H(x, 1) = g(x)$ for all $x \in X$. The function H is called a *homotopy* from f to g. Sets X and Y are called *homotopy equivalent* or of the same *homotopy type* if there exist two functions $f: X \to Y$ and $g: Y \to X$ such that $g \circ f$ is homotopic with the identity over X (id_X) and $f \circ g$ is homotopic with the identity over Y (id_Y).

Definition. We say that two topological spaces X and Y are *topologically equivalent* or *homeomorphic* if there exists a bijection $f: X \to Y$ such that f and the inverse function f^{-1} are continuous.

If two topological spaces X and Y are homeomorphic, then they are homotopy equivalent. We will use topological equivalence as a definition for topology preserving.

Definition. We will say that a digitization $Dig(X, r)$ of some set X is *topology preserving* if X and $Dig(X, r)$ are homeomorphic.

We now consider a special case of homotopy equivalence called a strong deformation retraction. Intuitively, saying that there is a strong deformation retraction from a set X to a set $Y \subseteq X$ means that we can continuously shrink X to Y.

Definition. Let X and $Y \subseteq X$ be two topological spaces. A continuous function $H: X \times [0, 1] \to X$, where $[0, 1]$ is the unit interval, is called a *strong deformation retraction*
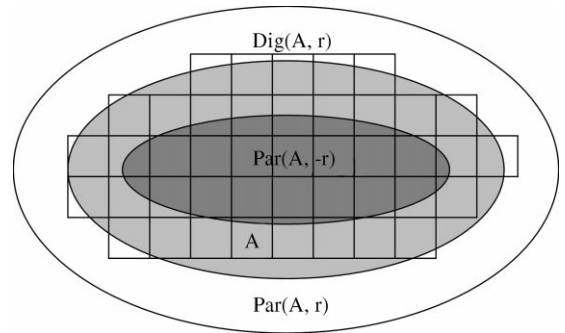

Fig. 7. $Par(A, -r)$ is a strong deformation retract of $Dig(A, r)$.

of X to Y if $H(x, 1) = x$ and $H(x, 0) \in Y$ for every $x \in X$, and $H(x, t) = x$ for every $x \in Y$ and $t \in [0, 1]$. Y is called a *strong deformation retract* of X.

Note that if Y is a strong deformation retract of X, then Y is homotopy equivalent to X. To see this, take $f: X \to Y$ to be $f(x) = H(x, 0)$ and $g: Y \to X$ to be inclusion.

Theorem 2. *Let A be a par(r)-regular set. Then $Par(A, -r)$ is a strong deformation retract of A.*

Theorem 3. *Let A be a par(r)-regular set. Then $Par(A, -r)$ is a strong deformation retract of $Dig(A, r)$ for every digital image $Dig(A, r)$ (which satisfies conditions ds1, ds2, and ds3), and $d_H(A, Dig(A, r)) \leqslant r$, where d_H is Hausdorff distance (see Fig. 7 for an illustration).*

Now we are ready to present our main theorems.

Theorem 4. *Let A be a par(r)-regular set. Then A and $Dig(A, r)$ are homotopy equivalent for every digitization $Dig(A, r)$, and $d_H(A, Dig(A, r)) \leqslant r$, where d_H is Hausdorff distance (which satisfies conditions ds1, ds2, and ds3).*

For Theorem 5, we need the following concepts:

Definition. We call a closed set A a *bordered 2D manifold* if every point in A has a neighborhood homeomorphic to a relatively open subset of a closed half-plane. A connected component of a 2D bordered manifold is called a *bordered surface*.

We suspect strongly that every *par(r)*-regular set is a bordered 2D manifold. However, a proof of this assertion would be beyond the scope of this paper. Therefore, in Theorem 5, we explicitly assume that a set *A* is a bordered 2D manifold.

Theorem 5. *Let A be a par(r)-regular bordered* 2D *manifold. Then A and Dig(A, r) are homeomorphic for every digital image Dig(A, r) (which satisfies conditions ds1, ds2, and ds3).*

An important consequence of Theorem 5 is the fact that under correct digitization resolution any two digital images of a given spatial object *A* are topologically equivalent. This means, for example, that shifting or rotating an object or the camera cannot lead to topologically different images, i.e., topological properties of obtained digital images are invariant under shifting and rotation.

Theorem 6. *Let A be a par(r)-regular bordered manifold. Then any two digitizations Dig^1(A, r) and Dig^2(A, r) of A are homeomorphic.*

Theorem 7. *Let A be a $\mathscr{C}^2$ subset of the plane (i.e., A is the closure of an open set whose boundary can be described as a disjoint finite union of twice continuously differentiable simple closed curves). Then there always exists a digitization resolution r > 0 such that every digitization Dig(A, r) of A is topology preserving.*

It can also be shown that if a set *A* is par(r)-regular, then $Dig_\cap(A, r)$ will never significantly change its local geometric properties (see Ref. [1]).

4. Digital patterns in digitizations

In this section, we show that if *A* is a par(r)-regular set, then some digital patterns cannot occur in its digitization *Dig(A, r)*. This is very useful for noise detection, since if these patterns occur, they must be due to noise. So, if in a practical application the resolution *r* of the digitization is such that the parts of the object which have to be preserved under the digitization are compatible with the square sampling grid, then our results allow for efficient noise detection. In this section, we establish two useful theorems about the digitization of parallel regular sets.

Theorem 8. *Let A be par(r)-regular. Then the pattern shown in Fig. 8 and its 90° rotation cannot occur in any Dig(A, r).*

By Theorem 8, a local 2 × 2 checkerboard pattern as shown in Fig. 8 cannot occur in any *Dig(A, r)*. In Ref. [7] a set is said to be *well-composed iff* every 8-connected
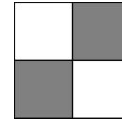

Fig. 8. This pattern and its 90° rotation cannot occur in every *Dig(A, r)*.

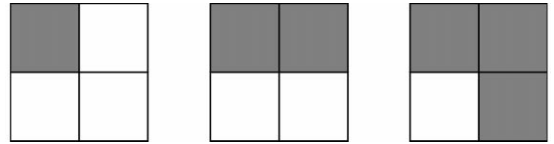Fig. 9. The only possible 2 × 2 configurations of boundary squares in *Dig(A, r)* of a par(r)-regular set *A* (modulo reflection and 90° rotation).

component is also a 4-connected component. It is further shown in Ref. [7] that this is equivalent to the set admitting no local checkerboard patterns, as those shown in Fig. 8. Thus, we have shown that a topologically invariant digitization of a parallel regular set is well-composed.

Hereafter, we will alternatively refer to a set that is well-composed as a *strongly connected* set and a set that is not well-composed as a *weakly connected* set. Our definition of strong connectivity is equivalent to the definition given in Ref. [8]. A local 2 × 2 checkerboard pattern will alternatively be referred to hereafter as an *nwc* (non-well-composed) neighborhood. Well-composed sets have very nice digital topological properties; in particular, a digital version of the Jordan Curve Theorem holds and the Euler characteristic is locally computable. These results imply that many algorithms in digital image processing can be simpler and faster (See Ref. [7]).

Since the configuration shown in Fig. 8 (and its 90° rotation) cannot occur in *Dig(A, r)* by Theorem 8, there exist only three 2 × 2 configurations of boundary squares in *Dig(A, r)* shown in Fig. 9 (modulo reflection and 90° rotation). Therefore, if we view *Dig(A, r)* as a subset of $\mathbb{R}^2$, every point in *Dig(A, r)* has a neighborhood homeomorphic to a relatively open subset of a closed half-plane. This motivates the following theorem (See Ref. [5]):

Theorem 9. *Let A be par(r)-regular. Then Dig(A, r) is a bordered* 2D *manifold.*

5. Topological invariance in digital documents

5.1. Thresholding

In the previous section, we showed that a topologically invariant digitization is rather well-behaved. In particular, it should have no checkerboard patterns, i.e. it is

well-composed and only three possible 2×2 boundary configurations are possible (modulo reflection and rotation). One of the important problems that generally needs to be solved in analyzing document images is that of finding a threshold to convert the document from a gray-level digital image to a binary one. Finding a good threshold value is important in the document domain for many subsequent applications from OCR to symbolic compression (see Refs. [3,9]).

Often image documents have the property that they are bimodal, and the two peaks of the gray-level histogram are quite distinct, but the proper threshold value between these peaks can be very hard to find. Consider the gray-level image shown in Fig. 10 of part of a digital document image. This image was captured using a scanner at 400 dpi. The apparent minimum of the gray-level histogram occurs at approximately 169. The image in Fig. 10 is shown thresholded at gray-level value 169 in Fig. 11. As is evident, this is not a particularly good threshold of the gray-level document. It seems to be considerably lower than the desired threshold value and results in many false disconnections, where components that were clearly connected in the original image have become disconnected in the thresholded image. If we desire a homeomorphic digitization, then quite clearly a digitization, which includes finding the correct threshold, is one where there are neither false connections nor false disconnections of connected components.
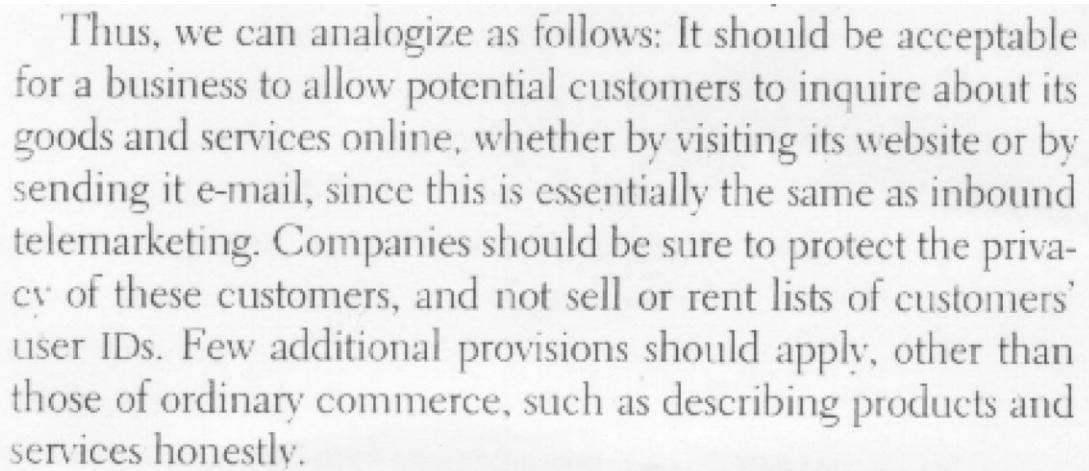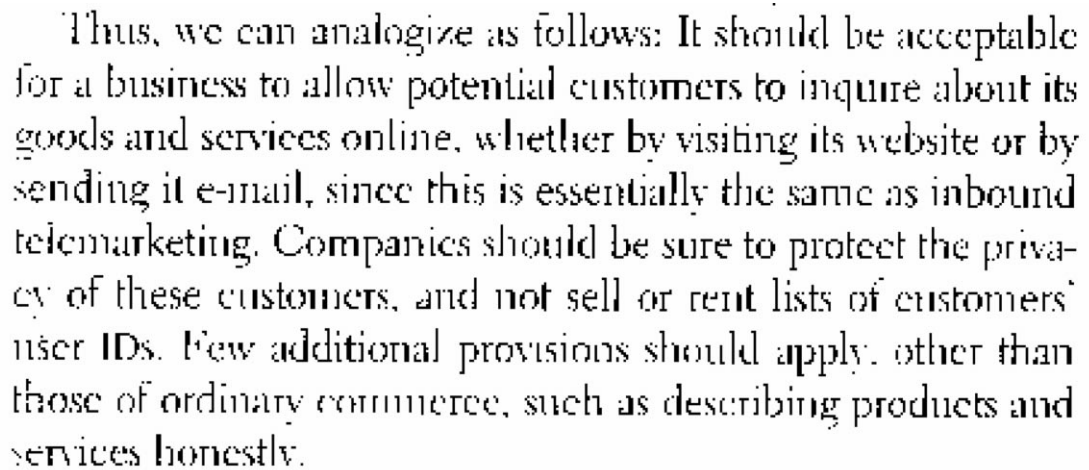

Fig. 10. A gray-level document image.

Fig. 11. The gray-level document image thresholded at the gray-level value of 169, which is the approximate minimum of the histogram of image gray-level values.

Thus, we can analogize as follows: It should be acceptable for a business to allow potential customers to inquire about its goods and services online, whether by visiting its website or by sending it e-mail, since this is essentially the same as inbound telemarketing. Companies should be sure to protect the privacy of these customers, and not sell or rent lists of customers' user IDs. Few additional provisions should apply, other than those of ordinary commerce, such as describing products and services honestly.

Fig. 12. The resulting binary image when the threshold is set to 232.

Next, let us consider the binary image shown in Fig. 12, which is once again a thresholded version of the image shown in Fig. 10. It can be seen that in this image the problem is reversed — there is primarily a problem of false connection, with letters connecting to other letters in the text. One way to view thresholding a digital document is that setting the threshold lower effectively thins out each letter, or component, while raising the threshold effectively thickens each textual component. Clearly, then, there is a tradeoff between the false connection and the false disconnection rate. Assuming the initial document was scanned at some resolution that was not completely topology preserving, as described in the previous sections, this false connection/disconnection tradeoff will almost always exist.

There is no apparent reason to assume that finding the minimum, if it can be found, in the gray-level image histogram will yield a threshold that is topologically optimal. For example, consider the text of the word "should" shown in Fig. 13a, which is a subimage of Fig. 10. Now suppose that the document was set in a different font, where the shape of the letters was exactly the same while the distance between letters (i.e., white space) was considerably smaller than in the original font and the distance between words in this new font was considerably larger. Then we might have a digitization of the word "should" as shown in Fig. 13c, which was generated by taking the bounding boxes around each gray-level letter and moving these bounding boxes as to make them adjacent, while moving all the columns containing only background pixels to the outside. Since the distribution of pixels in the image itself remained the same, the histogram for the two images also did not change. From a digital topology perspective, however, the image in Fig. 13c is very different from the one in Fig. 13a. The
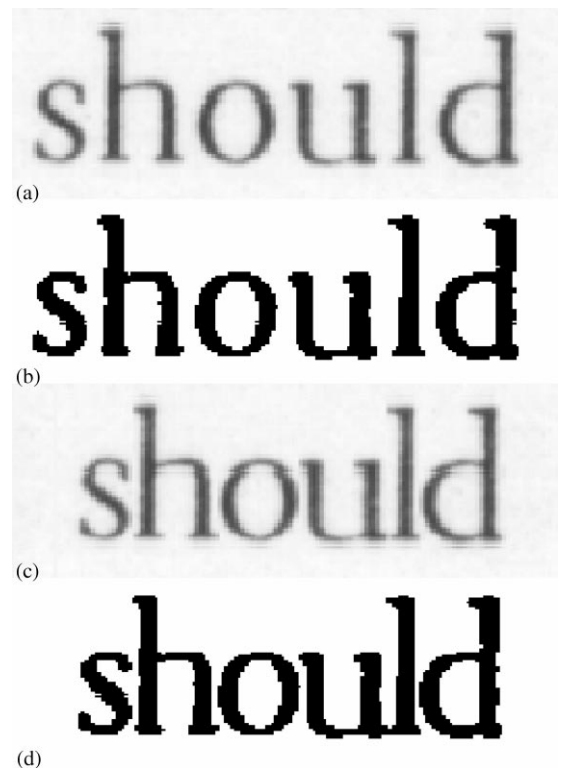

Fig. 13. Two different fonts with identical gray-level distributions and their respective binary images, thresholded at 215.

text font in Fig. 13a has wide spaces between letters, but the width of the actual letters themselves is quite narrow, often less than the two pixels required for topologically-invariant digitization. As such, we need to cut a higher

threshold, which will result in thickening the letters themselves while narrowing the distance between letters.

While the version of "should" in Fig. 13a might be par(5, +)-regular or (6, +)-regular, the generated font in Fig. 13c is in between par(1, +)-regular and par(2, +)-regular. Since the first font is more dilatable than the second, it follows that the digitization of the first font can be thresholded at a higher gray-level value than the second font, while still remaining homeomorphic to the preimage. Clearly, this is not a distinction we could make from their identical gray-level histograms.

For example, if the image in Fig. 13a were thresholded at the gray-level value of 200, then the digitization would be homeomorphic to the original image in $\mathbb{R}^2$, as shown in Fig. 13b. On the other hand, the same threshold value when applied to the text font of "should" in Fig. 13(c) results in a false connection as shown in Fig. 13(d), so that the digitization, after thresholding, is not topology-preserving. This second font, with smaller distances between letters, clearly requires a lower threshold value. Thus, in this example we have demonstrated that two images with identical gray-level histograms may require considerably different thresholds in order to insure the topological invariance of the digitization process.

It would seem that a more useful indicator as to what threshold preserves topology is to consider the weak connectivity or number of non-well-composed neighborhoods induced on the digital set by a given threshold. For example, consider the gray-level word "than" shown in Fig. 14 taken from the original document in Fig. 10. We consider what happens to this word topologically as we change the threshold. Thresholded versions of the image are shown in Fig. 15 as the threshold is set respectively to gray-level values 232, 233, 236, and 237. As one can see, a false topological connection often occurs initially as a weak connection between two sets. The binary "than", thresholded at 232, in Fig. 15a is homeomorphic to the original image in $\mathbb{R}^2$ except for the separate dot inside the h. The binary "t" thresholded at 233, shown in Fig. 15b, is

weakly connected to the "h". When the threshold is changed to 236, see Fig. 15c, the weak connection between the "t" and "h" becomes a strong connection, and at the same time, "h" becomes weakly connected to the "a". If the threshold is incremented to 237, as shown in Fig. 15d, the letters "t", "h", and "a" are now all strongly connected to each other, and there is now a non-well-composed neighborhood on the boundary of the letter "n" as it starts to connect with the "a". This example is typical of the way in which false topological connections occur as the threshold is increased.
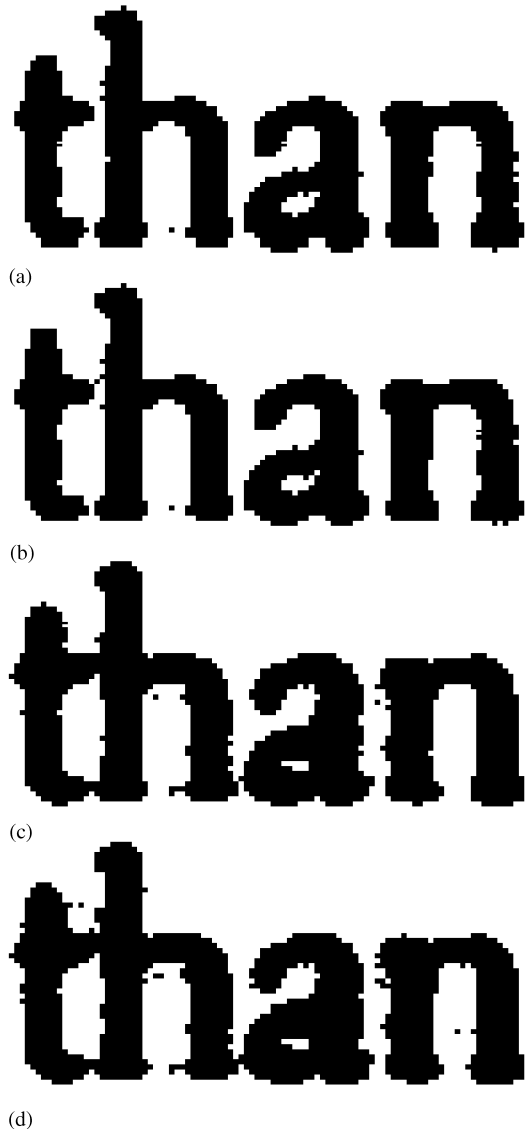

(a)

(b)

(c)

(d)

Fig. 15. A false topological connection between two digital sets often first appears as a weak connection between these sets, i.e., 8-connected but not 4-connected. As the threshold increases disjoint sets first become weakly connected then strongly connected.
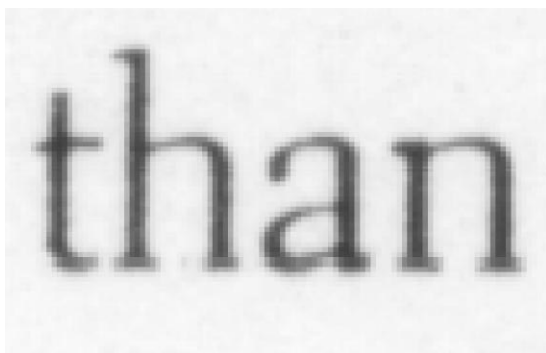

Fig. 14. A gray-level image of the word "than" taken from original document.
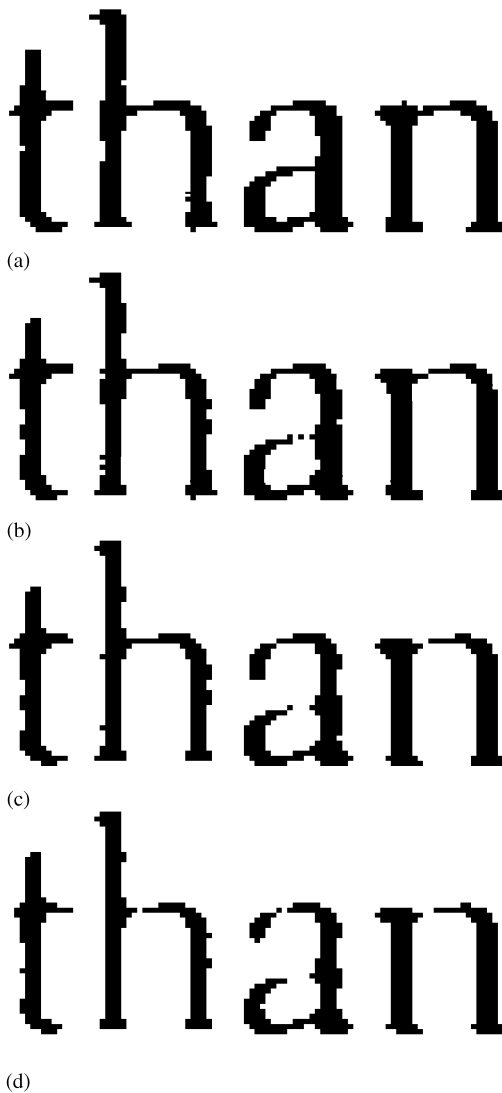
(a)

(b)

(c)

(d)

Fig. 16. A false topological disconnection between two digital sets often appears initially as a weak connection between these sets, i.e. 8-connected but not 4-connected. As the threshold decreases, connected sets first become weakly disconnected and then completely disconnected.

A similar situation frequently occurs in the reverse direction, with respect to false topological disconnections. Consider the same image, this time as the threshold is lowered to effect some false disconnections. First, the image thresholded at 174, as shown in Fig. 16a, is homeomorphic to the original. When the threshold is lowered to 169 then the "a" becomes disconnected while the "n" becomes only weakly connected, see Fig. 16b. If we lower the threshold further to 164, as shown in Fig. 16c, then the top part of the "a" becomes weakly connected, while the "n" that was weakly connected becomes completely disconnected. Finally, when the threshold is

lowered to 158 (see Fig. 16d), then top of the "a" becomes disconnected and the "h" also becomes completely disconnected (without having been previously weakly connected).

So we see that weak connectivity is often evident before a set goes from being connected to disconnected or the reverse. From our experiments, it seems that if two sets are about to connect (alternatively disconnect) due to a change in threshold, it is very likely that this will involve an interim state of weak connectivity. If we revisit the two thresholded images shown in Figs. 11 and 12, there is a clear indication from the number of weakly connected components in each case that neither threshold is topologically "optimal". In Fig. 17 there are a lot of weakly connected sets, where each locally non-well-composed neighborhood is marked for visibility purposes as a gray colored square, but these almost always occur within letters and result from sets that should have been connected but instead are only weakly connected. Similarly, the non-well-composed neighborhoods in Fig. 18 result primarily from either pairs of adjacent disconnected letters that have become weakly connected or from letters that are incorrectly self-connecting.

The histogram of the number of non-well-composed neighborhoods per threshold value is bimodal for almost all the digital documents we have tried in our experiments, which have consisted of several hundred document images scanned at different dpi's (and consisting of several different languages). The two maxima of the histogram correspond to the two gray-level values where either the rate of topological false connections or disconnections has an extrema. For the document shown in Fig. 10, the two extrema occurred at gray-level values of 86 and 246. The first extrema occurred as a result of letters that were falsely disconnecting, as can be seen in Fig. 19. Similarly, the second extrema occurred as a result of both false disconnections resulting from letters connecting or from noisy background pixels forming non-well-composed neighborhoods. Both of these images are topologically very unstable in that the underlying topological structure of the image is rapidly changing. Conversely, the minimum of weakly connected neighborhoods that occurs in between these two maxima is topologically very stable in that the rate of topological change is at a minimum, which occurs at the gray-level value of 213 as shown in Fig. 20.

Over several hundred documents that we have studied, finding the minimum of the weak connectivity histogram seems to result in a thresholding of the gray-level image into binary that is either topologically optimal, i.e. total number of false connections and disconnections is minimized, or very close to optimal. Unlike the minimum of the gray-level histogram, which is often flat or not well-defined, the minimum of the weak connectivity histogram is generally very well-defined. For example, two typical document gray-level histograms are shown in

Thus, we can analogize as follows: It should be acceptable for a business to allow potential customers to inquire about its goods and services online, whether by visiting its website or by sending it e-mail, since this is essentially the same as inbound telemarketing. Companies should be sure to protect the privacy of these customers, and not sell or rent lists of customers' user IDs. Few additional provisions should apply, other than those of ordinary commerce, such as describing products and services honestly.

Fig. 17. The image is underthresholded. The weakly connected components are almost entirely the result of false (weak) disconnections.

Thus, we can analogize as follows: It should be acceptable for a business to allow potential customers to inquire about its goods and services online, whether by visiting its website or by sending it e-mail, since this is essentially the same as inbound telemarketing. Companies should be sure to protect the privacy of these customers, and not sell or rent lists of customers' user IDs. Few additional provisions should apply, other than those of ordinary commerce, such as describing products and services honestly.

Fig. 18. The weakly connected components in this image result from false (weak) connections that occur either between adjacent disconnected letters or as a result of single letters that have incorrect (weak) self-intersections. The result is an image that is overthresholded.

Fig. 21, where the minima are either flat or non-well-defined. In contrast, the weak connectivity histograms of the same two documents are shown in Fig. 22. In both of these histograms the minima are clearly defined. In addition, in all of the experiments we have conducted, the weak connectivity minima outperform the gray-level minima considerably with respect to the topological correctness of the resulting binary image. Furthermore, when used as a filter to an OCR program (e.g. Omni-Page), OCR rates significantly improve.

The algorithm we use to compute the topology-preserving threshold is simply one of computing the number of checkerboard or non-well-composed neigh-borhoods that would result from each gray-level value and then finding the minimum that lies between the two maxima of the histogram. This algorithm runs in linear time and requires only one pass of the image. For each pixel in the image (except for last row and column), consider the 2×2 neighborhood with the given pixel in the top left corner. This 2×2 neighborhood has two diagonals, each diagonal defining an interval on the integers. Now either the two intervals defined by the two diagonals are disjoint or they intersect. If the intervals intersect then this 2×2 neighborhood will never induce a non-well-composed neighborhood and we need not consider it any further. If, however, the two intervals are
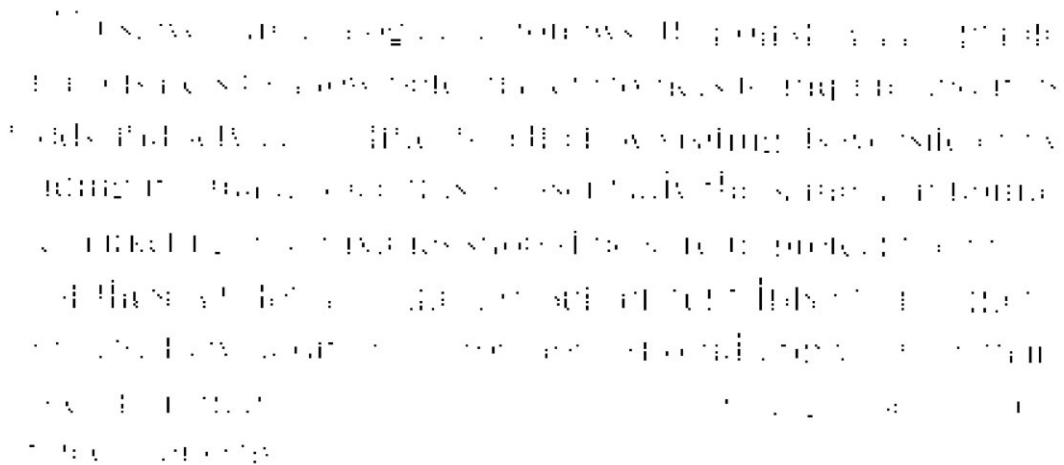
Fig. 19. Image that results from thresholding at a gray-level value of 86, which is the lower of the two maxima of the weak connectivity histogram.

Fig. 20. This image was thresholded at the minimum of the non-well-composedness histogram, which occurred at a gray-level value of 213. The resulting binary image is topologically very close to the original document.

disjoint then any threshold of the document inside the disjoint interval results in this neighborhood becoming a checkerboard pattern. Therefore, in our histogram all the bins in the disjoint interval are incremented. After traversing the entire image, having updated the histogram in the manner described, we find the two peaks in the histogram. The minimum that lies between them is selected by the algorithm as the best topology-preserving threshold.

5.2. Correction

One of the main reasons for using digital topology as a basis for document analysis is that it allows for the derivation of discrete document models, like well-composedness, that can be used to guide aspects of the processing and analysis of digital documents. For example, one of the important results in the previous section was that only three boundary configurations are possible as a result of a topologically invariant digitization. This can obviously be introduced as a dynamic constraint on the thresholding process to supplement the static method described above. We have not pursued this aspect thus far, although we certainly intend to explore it in future work. Whether or not a pixel is turned "on" or "off" during the thresholding process should probably be a function not only of the pixel's gray-level value but also a function of its local support. In Ref. [1] only 7 3×3
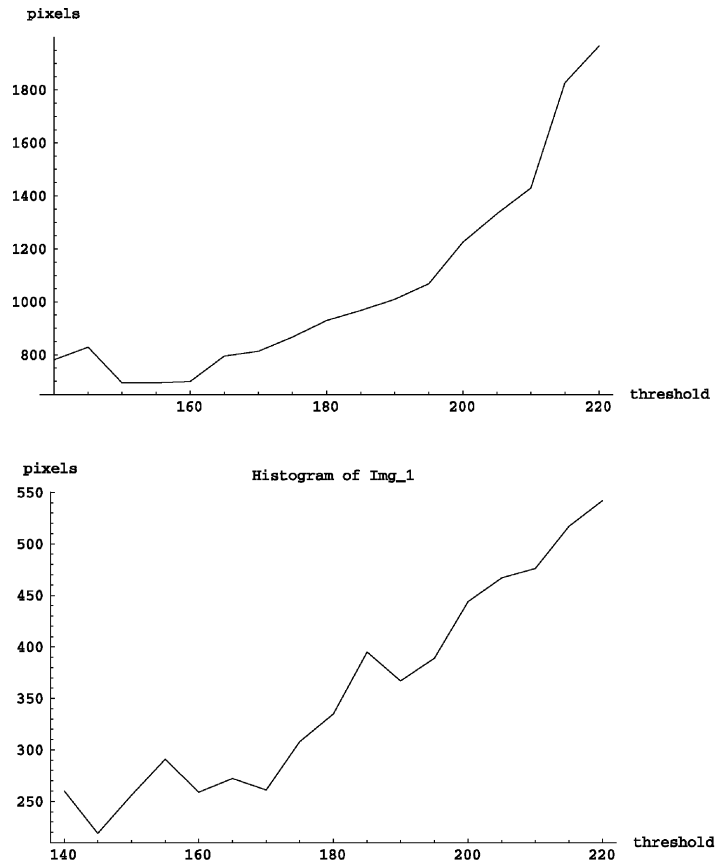
Fig. 21. Gray-level histograms for typical document images.

boundary configurations were found to be realizable under a topologically invariant intersection digitization, and subsequently we verified experimentally that almost all occurrences of boundary configurations not in this group are the result of sensor error (e.g. non-monotonicity). So certainly biasing the threshold in favor of topologically realizable neighborhoods is a logical progression. In addition, recent research indicates that the threshold should also vary dynamically with respect to the local digital geometry, e.g. convex or concave.

In this section, however, we are simply interested in converting the digital document, thresholded at a gray-level value minimizing the weak connectivity, to a digital image that is well-composed, i.e. strongly connected. The motivation for correcting the digital document in this way is essentially twofold: (i) all the non-well-composed neighborhoods are actually either connected or disconnected in $\mathbb{R}^2$ and, as such, should to be corrected in the digital image; and (ii) once we have converted the image to one that is entirely well-composed, many subsequent image processing algorithms are both faster and easier. For example, both sequential and parallel thinning algo-

rithms are greatly simplified, see Ref. [7], and guaranteed to result in a single-pixel-thick skeleton. Also, well-composed sets are quite well-behaved and satisfy some important properties. In particular, the digital Jordan Curve Theorem is satisfied, without requiring different connectivity relations for foreground and background pixels, and the Euler characteristic is locally computable.

In general, the number of pixels that actually need correction is very small compared to the number of pixels in the image. In the image shown in Fig. 10, when thresholded at a gray level of 213, only 1 in 859,308 2×2 neighborhoods was non-well-composed. Of course, this threshold was selected exactly because it minimized the weakly connected sets that would need to be resolved. A different threshold of the document could result in many more neighborhoods needing correction. If we had selected a threshold value of 169, which was approximately the minimum of the gray-level histogram, then there would have been many more neighborhoods to correct. When thresholded at 169, the resulting binary image had 111 locally non-well-composed neighborhoods.
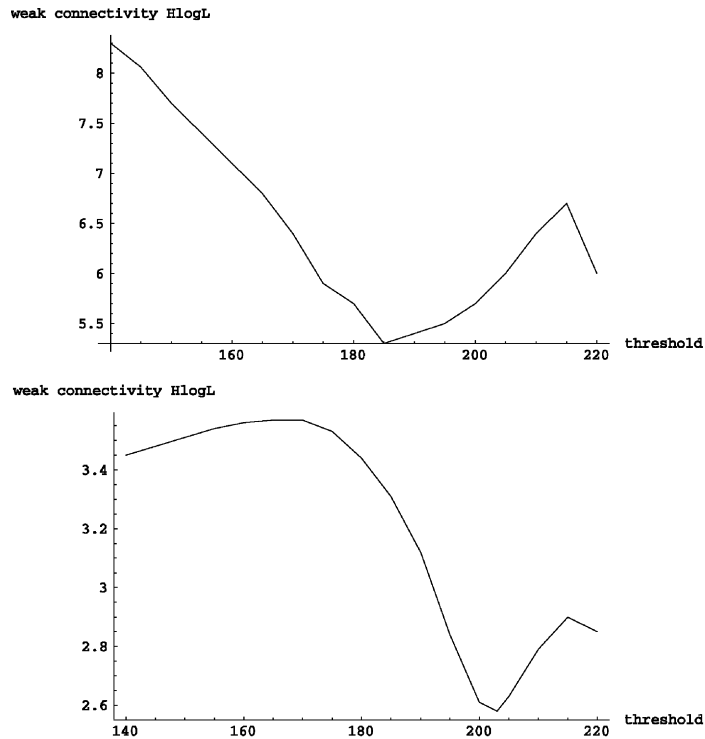
Fig. 22. Weak connectivity histograms for typical document images.

Finding neighborhoods that need to be fixed does not necessarily mean that we know how to correct them. But there seem to be two general approaches to correcting any weak connectivity on a digital document. One method is strictly based on local topology. The example shown in Fig. 23 is very typical of a weak connection. Most weak connections on the thresholded image result from two characters touching accidentally, due to insufficient sensor resolution with respect to the white space separating them. We assume that two 4-connected components that are weakly connected together are only connected together accidentally. This is corrected for by disconnecting the two weakly connected sets, i.e. changing one of the foreground pixels to a background pixel.

Another type of weak connection that occurs is shown in Fig. 24. Here the "g" should be connected, but is in fact only weakly connected since its preimage was not par(1, −)-regular, i.e. the stroke of the "g" is too narrow considering the given sensor resolution. However, since the "g" is already one connected component, even if this local neighborhood is disconnected, it is assumed that this weak connection between parts of the "g" did not occur accidentally and is resolved in favor of strong connectivity.

These are two of the rules that we currently use to resolve any weakly connected neighborhoods. These

Fig. 23. The weak connectivity between the "r" and the "a" is accidental, resulting from insufficient white space between the letters. In particular, the digitization is not homeomorphic to the preimage since the real letters "r" and "a" in $\mathbb{R}$ are not par(1, +)-regular.

Fig. 24. The weak connectivity between the two parts of the "g" is not considered to be accidental since the "g" remains one connected component regardless of how the weak connection is resolved. Consequently, this weak connectivity is resolved in favor of strongly connecting the neighborhood, i.e., one of the background pixels is changed.

rules are local as they do not require solving for parameters of the document. It is often useful, however, once a threshold has been found and the document image grouped into connected components to use aspects of the recovered parameters of the document model to further resolve any remaining topological ambiguity. For example, knowing the frequency of each type of bounding box in the document image can be useful in resolving weak connectivity. Computing a document model, however, is not within the scope of the current paper so we assume for now that any weakly connected neighborhoods can be correctly resolved using a set of local rules.

5.3. Topological compression and document readability

In this section, we develop the concept of topological compression and, at the same time, consider algorithms for computing the minimal resolution under which a document preserves readability. Transform-based compression methods such as Fourier transforms are frequency-based and, as such, are not guaranteed to preserve either topological or geometric features of the original image. But since many image analysis algorithms rely on connectivity-related features for the purpose of analysis and recognition, it seems important for the computer vision community to develop compression techniques that preserve various aspects of image structure during the compression process. We consider topological compression in this section.

Having thresholded the digital document in such a way as to minimize its weak connectivity, the image was then corrected, as described in the previous section, so that all non-well-composed neighborhoods were removed. This digital document is now strongly connected and, hopefully, homeomorphic to the original prescanned document. We now construct a pyramidal data structure, with each successive level of the pyramid consisting of an image that is half the size in each dimension of the level below it. For a description of pyramidal data structures and some related algorithms in image representation and compression, see Refs. [10] and [12]. We are interested, at each level of the pyramid, in compressing the image from $n \times n$ to $(n/2) \times (n/2)$ while preserving topology.

The basic rules we use in the algorithm are rather simple, in part, because since the digital document model is one of a well-composed set we can rely on various properties being satisfied such as the local computability of the Euler characteristic. As is classically done in pyramidal algorithms, each pixel or cell has both sons and stepsons, or alternatively, neighborhoods and overlapped neighborhoods. If a pixel on level $i + 1$ of the pyramid sees that his sons, i.e. the immediate neighborhood, are all the same color, either black or white, then he adopts that color. Since this involves no "recoloring" of

the pixels on level i, there is no topological change to the image. This case is, of course, rather trivial. Now assuming that the immediate neighborhood is not all one color, then obviously some "recoloring" is necessary.

For example, if the father pixel has in the 2×2 immediate neighborhood 3 black pixels and 1 white pixel then either value assumed will involve some recoloring of the 2×2 neighborhood. Since the compression is topological, our first priority is one of topology, and only after we are assured that topology is preserved are we concerned with preserving geometric features. Therefore, since there are always two ways to color a pixel, if only one of them preserves topology then this is the recoloring of the local neighborhood that we choose. In the case of the 2×2 neighborhood where there are three black pixels and one white pixel, if the white pixel is an isolated point, i.e. no white neighbors (perhaps it is digitization of the inside of the letter "a") then recoloring to black means that the topology is changed so the father pixel on level $i + 1$ is colored white.

A pixel is referred to as "simple" if its recoloring does not affect the image topology, see Ref. [7]. Similarly, we can refer to a 2×2 neighborhood as simple if a recoloring, either black or white, does not affect the image topology. If a neighborhood is not simple and only one recoloring of the father pixel preserves topology then we have no choice but to recolor in this manner. This is also the case if one recoloring preserves topology while the second recoloring creates a non-well-composed neighborhood since this also involves some change to the topological structure of the image. Assuming that both recolorings of the father pixel are equally topology-preserving then we select the one that is more geometry-preserving, which usually translates into using majority color. If neither recoloring is topology-preserving, then we select the recoloring most geometry-preserving and increment the variable that keeps count of how many times on a given level of the pyramid there was no possible topology-preserving recoloring of the local neighborhood. Since the underlying image is well-composed, determining whether a neighborhood is simple is straightforward.

This topological compression algorithm not only allows us to preserve the topology of the image wherever possible, but also allows us to determine the degree of topological degradation at each successive level of the pyramid. If the text font for the document was originally par(4)-regular, for example, then on the next level of the pyramid it would still be par(2)-regular, and we would not expect any topological loss of information until at least the thrid level. Generally, a given font has a certain stroke width and a certain width for the white space between strokes. There is a point in the pyramid of the binary well-composed document where topology can no longer be preserved. We usually consider a given level of the pyramid to no longer be topology-preserving if

Thus, we can analogize as follows: It should be acceptable for a business to allow potential customers to inquire about its goods and services online, whether by visiting its website or by sending it e-mail, since this is essentially the same as inbound telemarketing. Companies should be sure to protect the privacy of these customers, and not sell or rent lists of customers' user IDs. Few additional provisions should apply, other than those of ordinary commerce, such as describing products and services honestly.
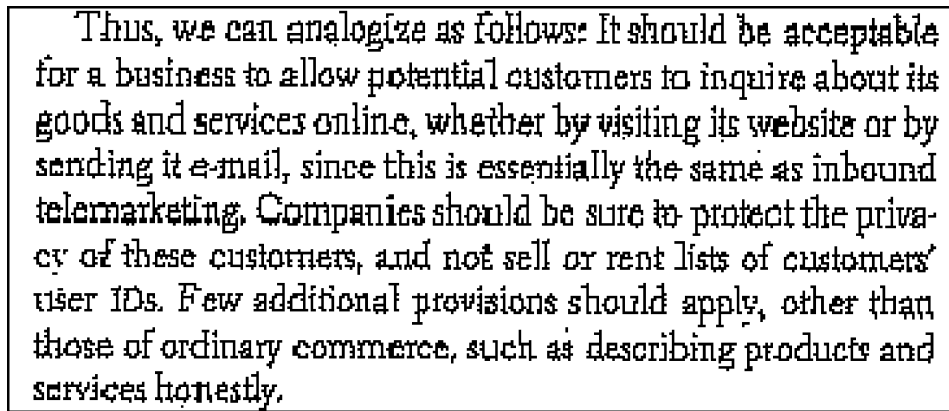
Fig. 25. Level 2 of topological compression on well-composed binary document, with only about 2% of the neighborhoods on this level of the pyramid not preserving topology.

Thus, we can analogize as follows: It should be acceptable for a business to allow potential customers to inquire about its goods and services online, whether by visiting its website or by sending it e-mail, since this is essentially the same as inbound telemarketing. Companies should be sure to protect the privacy of these customers, and not sell or rent lists of customers' user IDs. Few additional provisions should apply, other than those of ordinary commerce, such as describing products and services honestly.
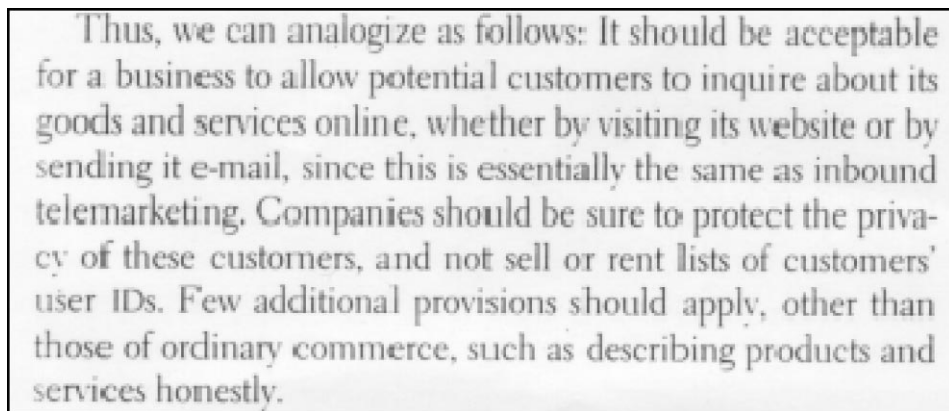
Fig. 26. Pyramid level 2 gray-level version of image preserves readability of the document.

more than 10% of the pixels on that level can no longer preserve the topology of their underlying neighborhoods.

In addition, there seems to be a correlation between preserving the topology of the digital document and preserving its readability. It appears that at an image resolution where topology can no longer be preserved neither can the readability of the document. As an example, consider once again the original document gray-level image in Fig. 10 and the thresholded (well-composed) version in Fig. 20. Assuming that these two images are the base images at level 0 of their respective pyramidal data structures, then level 2 of the binary pyramid is shown in Fig. 25. On this level of the pyramid, there were roughly 2% of the pixels that did not preserve the topology of their underlying neighborhoods. In general, this degree of topological degradation is relatively minor and indicates that the digitization at this level is approximately topology-preserving. Consequently, the corresponding level of the gray-level pyramid preserves the document readability, see Fig. 26. At the next level of the pyramid, however, the compressed version of the binary image no longer preserves topology with over 15% of the pixels on this level not able to preserve the topology of their underlying neighborhoods. The binary version of the pyramid is shown in Fig. 27. The fact that topology is not preserved at this level of the pyramid indicates that the original text of the prescanned document image was not par(8)-regular with respect to the sensor resolution of the scanner. Not surprisingly, the gray-level version of this level of the pyramid no longer preserves document readability, see Fig. 28. Thus, the algorithm described can be used both for topological compression and for computing the minimal image resolution that still preserves readability of the digital document.
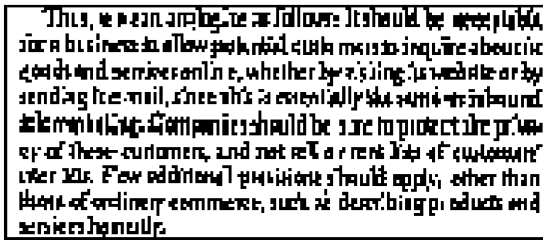
Fig. 27. Level 3 of the binary document pyramid, with over 15% of the neighborhoods on this level no longer preserving topology.
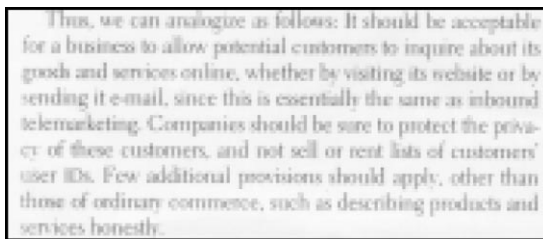

Fig. 28. Since the binary compressed document is no longer topology-preserving, level 3 of the pyramid gray-level image no longer preserves readability of the document.

6. Conclusions

In this paper, we gave conditions on the correct digitization resolution which guarantee that topology is preserved under a digitization and segmentation process. For a par(r)-regular set A, A and each of its digital images $Dig(A, r)$ are topologically equivalent. We also proved that the Hausdorff distance of sets A and $Dig(A, r)$ is less than or equal to r. These results have many consequences. For example, under correct digitization resolution any two digital images of a given spatial object A are topologically equivalent. Furthermore, $Dig(A, r)$ is well-composed, i.e. the checker board digital patterns cannot occur in $Dig(A, r)$. In Latecki et al. [27] it is shown that well-composed sets have very nice digital properties, which imply that many algorithms for digital image processing can be simpler and faster. Well-composedness can be useful for noise detection, since if the neighborhood of a boundary point contains a checker board digital pattern, it must be due to noise. For a large class of 2D objects, which includes projections of some real 3D objects, a constant r can be computed such that they are par(r)-regular.

The results derived in this paper, particularly the fact that a topologically invariant digitization results in a well-composed digital image, were applied to the domain of document image processing and analysis. We developed an algorithm to find a static threshold such that the weak connectivity of the digital document was minimized. In addition, we demonstrated that this algorithm performed very well with respect to minimizing the number of false topological connections/disconnections in the image. It was then shown how to further correct the document image to make it entirely well-composed. Finally, we developed a pyramid-based algorithm that performed topological compression and determined the minimal image resolution preserving document readability.

References

[1] A. Gross, L. Latecki, Digitizations preserving topological and differential geometric properties, Computer Vision Image Understanding 62 (1995) 370–381.

[2] A. Gross, L. Latecki, a realistic digitization model of straight lines. Comput. Vision Image Understanding 67 (1997) 131–142.

[3] A. Gross, L. Latecki, Homeomorphic digitization, correction, and compression of digital documents, IEEE Workshop on Document Image Analysis, San Juan, Puerto Rico, 20 June 1997, pp. 89–96.

[4] A. Gross, L. Latecki, Topologically-invariant methods in document image analysis, SPIE Vision Geometry VI, San Diego, CA, 28–29 July 1997, pp. 61–68.

[5] L.J. Latecki, Ch. Conrad, A. Gross, Preserving topology by a digitization process. J. Math. Imaging Vision 8 (1998) 131–159.

[6] J. Serra, Image Analysis and Mathematical Morphology, Academic Press, New York, 1982.

[7] L. Latecki, U. Eckhardt, A. Rosenfeld, Well-composed sets. Computer Vision Image Understanding 61 (1995) 70–83.

[8] E. Dougherty, J. Astola , An Introduction to Nonlinear Image Processing, vol. TT16, SPIE Optical Engineering Press, Bellingham, WA, 1994.

[9] D. Doermann, Document image understanding: integrating recovery and interpretation, Ph.D. Thesis, Univ of MD, College Park, 1993.

[10] V. Cantoni, M. Ferretti, Pyramidal Architectures for Computer Vision. Plenum, New York, 1994.

[11] X. Kong, J. Goutsias, A study of pyramidal techniques for Image Representation and compression. JVCIR 5 (1994) 190–203.

About the Author—ARI GROSS received the B.S. degree in Mathematics from Johns Hopkins University in 1982, the MS degree in Computer Science from the University of Maryland in 1986, and the Ph.D. in Computer Science from Columbia University in 1992. He is currently an Associate Professor of Computer Science at the University Graduate Center and Queens College, City University of New York, and directs research in the Computational Vision and Perception Laboratory. His research interests include digital geometry, discrete and continuous shape models, model-based image compression, and mathematical sensor modeling.

About the Author—LONGIN JAN LATECKI received his doctoral degree in 1992 and his habilitation (German postdoctoral degree) in December 1996, both from the Dept. of Computer Science, Univ. of Hamburg. He is currently working on a project "*Shape Representation in Discrete Structures*" from German Research Foundation (DFG) at the Dept. of Applied Mathematics, Univ. of Hamburg. He is a cochair of the SPIE's annual conference series "*Vision Geometry*" since 1996 (the next conference will be in San Diego, California, July 1998). His main research areas are Computer Vision: shape representation and shape similarity measures with application to image data bases, digital geometry and topology; Artificial Intelligence: representation and processing of spatial knowledge with application to robot navigation; and Computer Graphics: representation of 3D objects with triangular meshes. He is the author of two books, 15 journal articles, and over 30 book chapters and conference papers.