Segmentation of scanned documents for efficient compression

Hei Tao Fung
Kevin J. Parker

Electrical Engineering Department
University of Rochester
Rochester, New York 14627

ABSTRACT

A scanned, complex document image may be composed of text, graphics, halftones, and pictures, whose layout is unknown. In this paper, we propose a novel segmentation scheme for scanned document images that facilitates their efficient compression. Our scheme segments an input image into *binarizable* components and *non-binarizable* components. By a binarizable component we mean that the region can be represented by no more than two gray levels (or colors) with acceptable perceptual quality. A non-binarizable component is defined as a region that has to be represented by more than two gray levels (or colors) with acceptable perceptual quality. Once the components are identified, the binarizable components can be thresholded and compressed as a binary image using an efficient binary encoding scheme together with the gray values represented by the black and white pixels of the binary image. The non-binarizable components can be compressed using another suitable encoding scheme.

Keywords : document; segmentation, compression, color, scanned, and digitized.

1. INTRODUCTION

Complex document images contain pictures, halftones, text, and other image components. They may be generated by feeding documents to a scanner, photocopier or other input devices. The complex images pose problems to conventional compression systems. For example, gray scale images cannot be sent by conventional group 3 and 4 fax devices unless they are converted by some process to halftones. Furthermore, gray scale images are conveniently compressed and transmitted by JPEG devices, but the presence of many high contrast text characters will result in degraded compressibility, or increased distortion, or both. Also, the color documents are generally

0-8194-2103-0/96/$6.00

printed as halftones. Applying JPEG [1] to encode halftone texture background regions will result in degraded compressibility. Therefore, it is highly desirable to segment pictures, text, graphics, and halftone components on scanned document images so that they can be compressed separately. For example, the background regions can be represented by one solid color. Text and graphics can be compressed by a binary image encoding scheme such as the Group 4 facsimile encoding scheme [2]. Only the continuous-tone pictures and high-resolution halftones are compressed by a continuous-tone image encoding scheme such as JPEG. The separate image components can be transmitted and stored with much greater compression than the overall document.

There are a number of segmentation algorithms for document image such as the Constrained Run Length Algorithm (CRLA) [3], the Recursive X-Y Cut (RXYC) [4], and the Segmentation by Peak And Continuous Edge method (SPACE) [5]. They have various problems. We propose a novel segmentation algorithm, which is called SMART (Segmentation by subjecting Macroblocks of Active Regions to the binarizability Test). Our algorithm will be explained in more detail in subsequent sections. It takes into account the different characteristics of a general, complex document which are not fully considered in the previously published methods. For example, in our algorithm, regions of different image components may take various shapes, as long as they do not overlap. Text may appear as brighter than the background, and multiple background regions of different gray levels can be segmented because our method works with gray levels and local groups of pixels. Tilting, which affects the region shapes, is acceptable. Also, our method is well suited for modern JPEG and fax implementations, and incorporates a degree of human visual system weighting.

2. ALGORITHM

SMART has four major processing steps (see figure 1):

1. preprocessing. Append the input image if necessary. Convert it to another color space if the input image is in color.
2. block classification. Classify small blocks of pixels as *active* or *inactive*.
3. macroblock formation. Create macroblocks as groups of 4-connected active blocks.
4. macroblock classification. Classify each macroblock as *binarizable* or *non-binarizable* using a novel *binarizability test.*

Active blocks are 8x8 blocks showing significant gray level activities. An inactive block is an 8x8 block which is not an active block. A binarizable macroblock is a region which can be represented by no more than two gray levels with acceptable perceptual quality. A non-binarizable macroblock is a region which must be represented by at least two gray levels with acceptable perceptual quality.

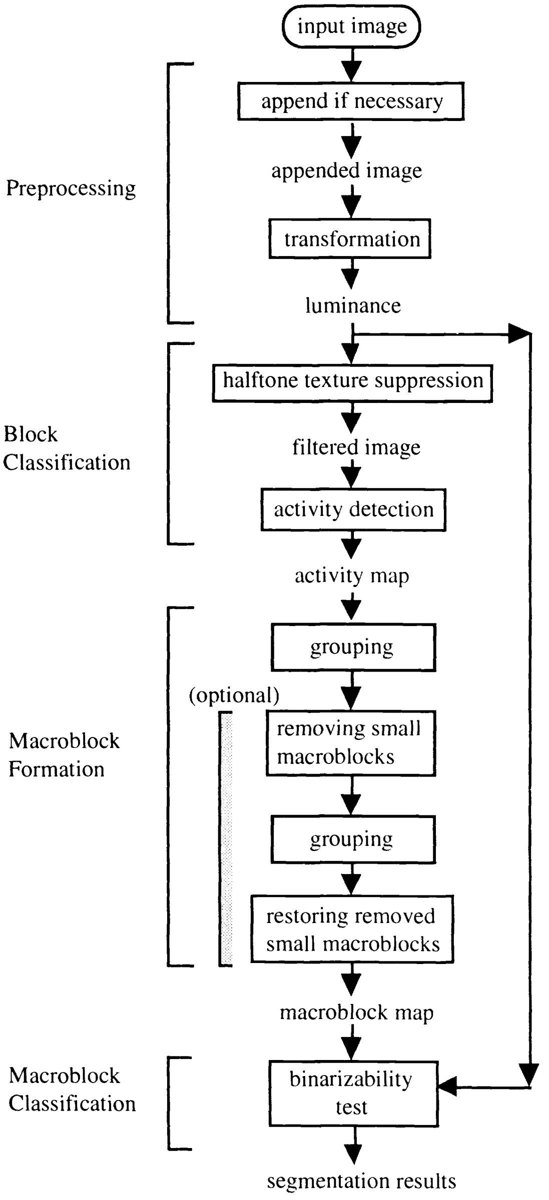


Figure 1. The segmentation processing steps for scanned document images.

2.1. Preprocessing

2.1.1. Appending the image

The image is appended on the right border or the bottom border to make its width and height to be multiples of 8. To avoid unnecessarily introducing high spatial frequencies into the blocks with appended pixels, the pixel values in the appended regions mirror the pixel values in the input image along the original borders.

2.1.2. Color transformation

According to Ohta *et. al.* [6], for the segmentation of color images, the most effective linear set of features calculated from the tristimuli R (red), G (green), and B (blue) is $(R + G + B)/3$, $R - B$, and $(2G - R - B)/2$. The first feature is the luminance. The other two carry the chrominance information. This set is an orthogonal, linear transformation from R, G, and B. Kender [7] concluded that linear transformations were preferable than nonlinear ones. On the basis of the results, we use the following transformation on the R, G, and B values:

$$Y = \frac{R+G+B}{3} \tag{1}$$

$$C1 = \frac{R-B}{2} + 128 \tag{2}$$

$$C2 = \frac{2G-R-B}{4} + 128 \tag{3}$$

$C1$ and $C2$ are normalized and offset to fit in the 8-bit range [0,255]. There is no need for transformation if the input image in monochrome.

2.2. Block classification

The image is subdivided into non-overlapping 8x8 blocks. Each block is classified as active or inactive.

2.2.1. Halftone texture suppression

The halftone texture produces periodic (or quasi-periodic) variations in pixel values and affects activity detection. It should be suppressed as noise. A median filter, which is appropriate for noise removal, is used for the halftone texture suppression. However, a median filter generally removes thin lines and corners of objects. Therefore, we adopt a special median filter which restores the original pixel value if it differs from the median filter output by more than a threshold T_1.

2.2.2. Activity detection

We define our activity measure as $\sum_{k=1}^{63} |NINT(ZZ(k) / Q(k))|$. Specifically, we take

$$If \sum_{k=1}^{63} |NINT(ZZ(k) / Q(k))| > T_2, then\, the\, block\, is\, active;$$
$$If \sum_{k=1}^{63} |NINT(ZZ(k) / Q(k))| \leq T_2, then\, the\, block\, is\, inactive. \qquad (4)$$

NINT is the nearest integer function, *k* is an index, *ZZ(k)* is the *k*th DCT coefficients in the zigzag order, *Q(k)* is the corresponding quantization table element, and T_2 is a preset threshold. *k* ranges from 0 to 63, but the activity measure excludes the value of the DC coefficient corresponding to *k* = 0.

2.3. Macroblock formation

This step converts some inactive blocks into active blocks to form larger regions of 4-connected active blocks called macroblocks. The process favors the creation of solid rectangular regions. The subprocesses are explained in the following paragraphs.

2.3.1. Grouping

Grouping consists of two parts: *expansion* and *trimming*. The *expansion* step converts some inactive blocks to active blocks according to the following criteria:

1. Any unconnected (not 8-connected) active blocks remain unconnected.
2. Circumscribed inactive blocks are converted to active blocks.
3. A region of 8-connected active blocks tends to expand into a solid rectangular macroblock provided that the first criterion is met.

The purpose is to congregate active blocks to facilitate representation and classification.

The *trimming* step trims the over-expanded macroblocks. It converts some converted inactive blocks back to inactive blocks such that any boundary of a macroblock that is touching any inactive block recedes. The purpose is to reduce the number of active blocks to be processed later.

2.3.2. Removing small macroblocks

Grouping may leave some small macroblocks being circumscribed by other macroblocks. These small macroblocks may be merged with the latter to facilitate representation and classification. There are various ways to remove the small macroblocks. We propose removing the macroblocks which have the number of 4-connected active blocks smaller than a threshold T_3.

2.3.3. Restoring removed macroblocks

After removing some circumscribed macroblocks, the grouping process previously discussed is repeated so that the macroblocks circumscribing the removed macroblocks may expand and cover the removed macroblocks. However, to avoid losing some removed macroblocks which are not covered after the second grouping, all removed macroblocks are restored. The overall effect is the reduction in fragmentation in the activity map.

2.3.4. An example

The whole process of macroblock formation is exemplified in figure 2. Figure 2(a) is an activity map, where a shaded box indicates an active block and a white box represents an inactive block. The result of the expansion step in shown in figure 2(b). A hole of inactive blocks in the top left area is filled, while the hole in the bottom area is not because the obstruction of some circumscribed small macroblocks. Note also that the blocks expand into rectangular shape unless their expansions are restricted. The trimming step deactivates some active blocks which are converted from inactive blocks so that the macroblocks are not over-expanded (figure 2(c)). Some of the small macroblocks are then removed in figure 2(d). In the second grouping, the bottom macroblock has its hole filled (figure 2(e)). At last, the removed macroblocks are restored (figure 2(f)).

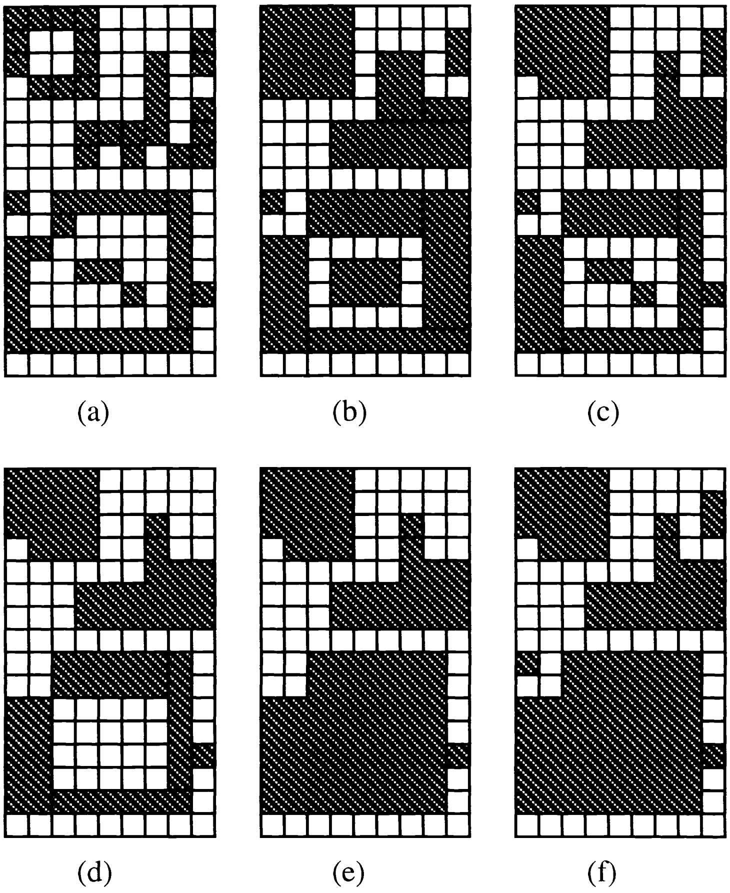


Figure 2. Macroblock formation exemplified: (a) after activity detection; (b) after expansion; (c) after trimming; (d) after removing the small macroblocks; (e) after grouping; and (f) after restoring the removed macroblocks.

2.4. Macroblock classification

To achieve efficient compression, it is desirable to classify each macroblock as binarizable or non-binarizable. In this implementation, binary images are not limited to black-and-white images. They include images of two arbitrary gray levels. However, due to noise and smearing in scanning, a scanned binary image may contain more than two gray levels.

Naturally, bimodality tests such as that in [8] can be used to distinguish binary image components from continuous-tone image components. However, they may fail in some cases where continuous-tone images exhibit bimodality. We propose a stricter bimodality test, which we called the binarizability test, as follows:

Step 1: Find the gray level histogram of a macroblock.

Step 2: Determine the two modes, m_1 and m_2, from the histogram. The modes must be separated by at least 60 gray levels (out of 256, or proportional in other bit depth systems).
Step 3: Render a pixel in a binary image 0 if $|g-m_1| \leq 30$ or $|g-m_2| \leq 30$, where g is the gray value of the corresponding pixel in the macroblock. Otherwise, render the pixel in the binary image 1.
Step 4: If there is no block of 1's of at least 4 x 4 pixels in the binary image formed in step 3, the macroblock is binarizable. Otherwise, it is non-binarizable.

3. SIMULATION RESULTS

The effectiveness of SMART is demonstrated using two images shown in figure 3 and figure 4, respectively. Although they are in color, only their luminance channels are shown. Figure 3, called Statistics, has two background regions of different gray levels. The letters have different gray levels and different relative grayness to their backgrounds. The background regions are printed as halftones. A continuous-tone image is located on the border of the two background regions. On the lighter background region, there is also a small degree of illumination variation. The image is scanned at 200 dpi. Figure 4, scanned at 200 dpi, is composed of text characters of different sizes and formats and a halftone picture of irregular shape. It is called Perseverance. The figures corresponding to the segmentation results do not reflect the actual size of their original images. Segmentation results from SMART are shown in three colors. The light-gray regions represent the background regions. The mid-gray regions are the binarizable regions. The non-binarizable regions are shown in black.

The segmentation result for Statistics is shown in figure 5. SMART can separate the two background regions by the non-binarizable macroblock which includes the border. Also, the halftone texture background regions are classified as binarizable as desired so that the background regions and text regions can be compressed by a binary compression scheme to achieve a higher compression ratio. To encode the image requires only 69,561 bits as opposed to 274,145 bits for using JPEG on the whole image. The segmentation result for Perseverance is shown in figure 6. SMART can separate the text blocks from the irregular picture region and allow binary compression on them. Using JPEG on the whole image requires 391,849 bits, while our compression with segmentation approach generates only 210,894 bits.

4. CONCLUSIONS

We have described a novel segmentation algorithm called SMART for scanned, complex document images aiming at efficient compression. Segments are classified into binarizable and non-binarizable components, where encoding schemes suitable for their kinds are used. SMART can handle image components of various shapes, multiple backgrounds of different gray levels, different relative grayness of text to the background, tilted image components, and text of different gray levels. It involves preprocessing stage, where a color space transformation may be performed, block classification into active blocks and inactive blocks, macroblock formation which congregates active blocks, and macroblock classification into binarizable and non-binarizable macroblocks. Its effectiveness in segmentation and its benefits to compression is demonstrated.

5. REFERENCES

[1] Digital Compression and Coding of Continuous-tone Still images, Part I, Requirements and Guidelines. ISO/IEC JTC1 Draft International Standard 10918-1, Nov. 1991.

[2] CCITT, "Recommendation T.6, Facsimile coding schemes and coding control functions for Group 4 facsimile apparatus," vol. VII-Fascicle VII.3, 48-57.

[3] F. M. Wahl, K. Y. Wong, and R. G. Casey, "Block segmentation and text extraction in mixed text/image documents," *Comput. Vision Graphics Image Process.* , vol. 20, 375-390, 1982.

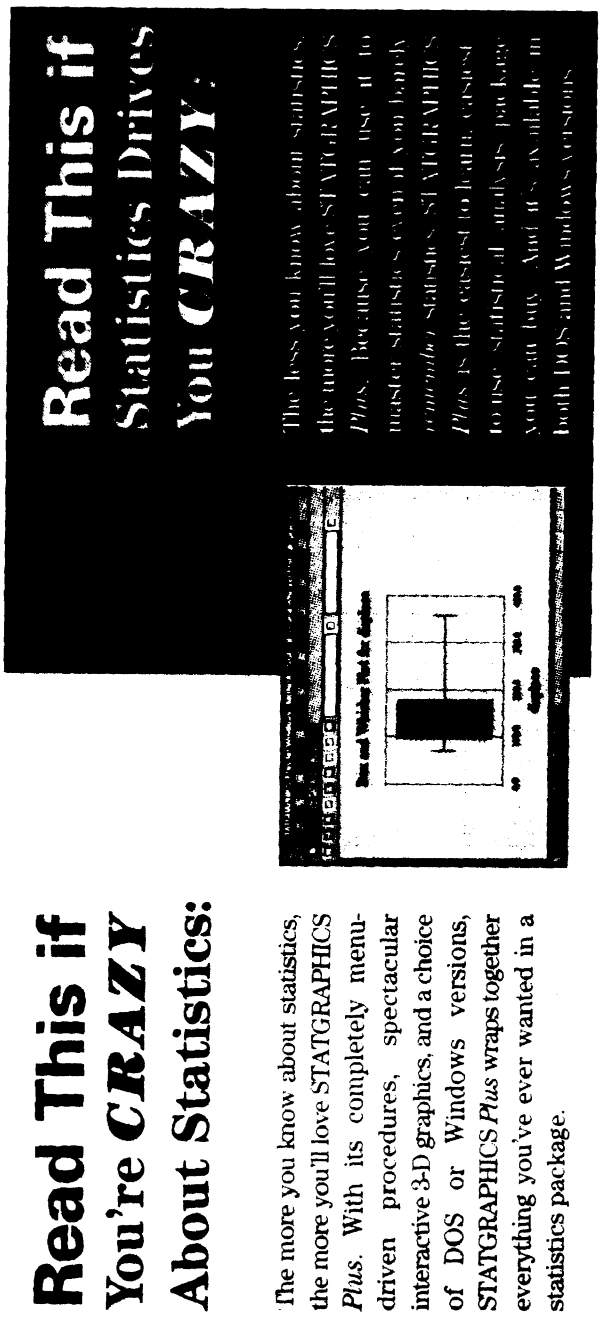
[4] G. Nagy, S. Seth, and S. D. Stoddard, "Document analysis with an expert system," Proc. Pattern Recog. in Practice, Amsterdam, June 19-21, 1985, Vol. II.

[5] S. Ohuchi, K. Imao, and W. Yamada, "A segmentation method for composite text/graphics (halftone and continuous tone photographs) documents," *Systems and Computers in Japan*, Vol. 24, No. 2, 35-44, 1993.

[6] Y. Ohta, T. Kanade, and T. Sakai, "Color information for region segmentation," Computer Graphics and Image Processing, vol. 13, pp. 222-241, 1980.

[7] J. Kender, *Saturation, Hue and Normalized Color: Calculation, Digitization Effects, and Use*, Technical Report, Department of Computer Science, Carnegie-Mellon University, 1976.

[8] A. L. Negrate, A. Beghdadi, and H. Dupoisot, "An image enhancement technique and its evaluation through bimodality analysis," CVGIP: Graphical Models and Image Processing, vol. 54, no. 1, pp. 13-22, January 1992.



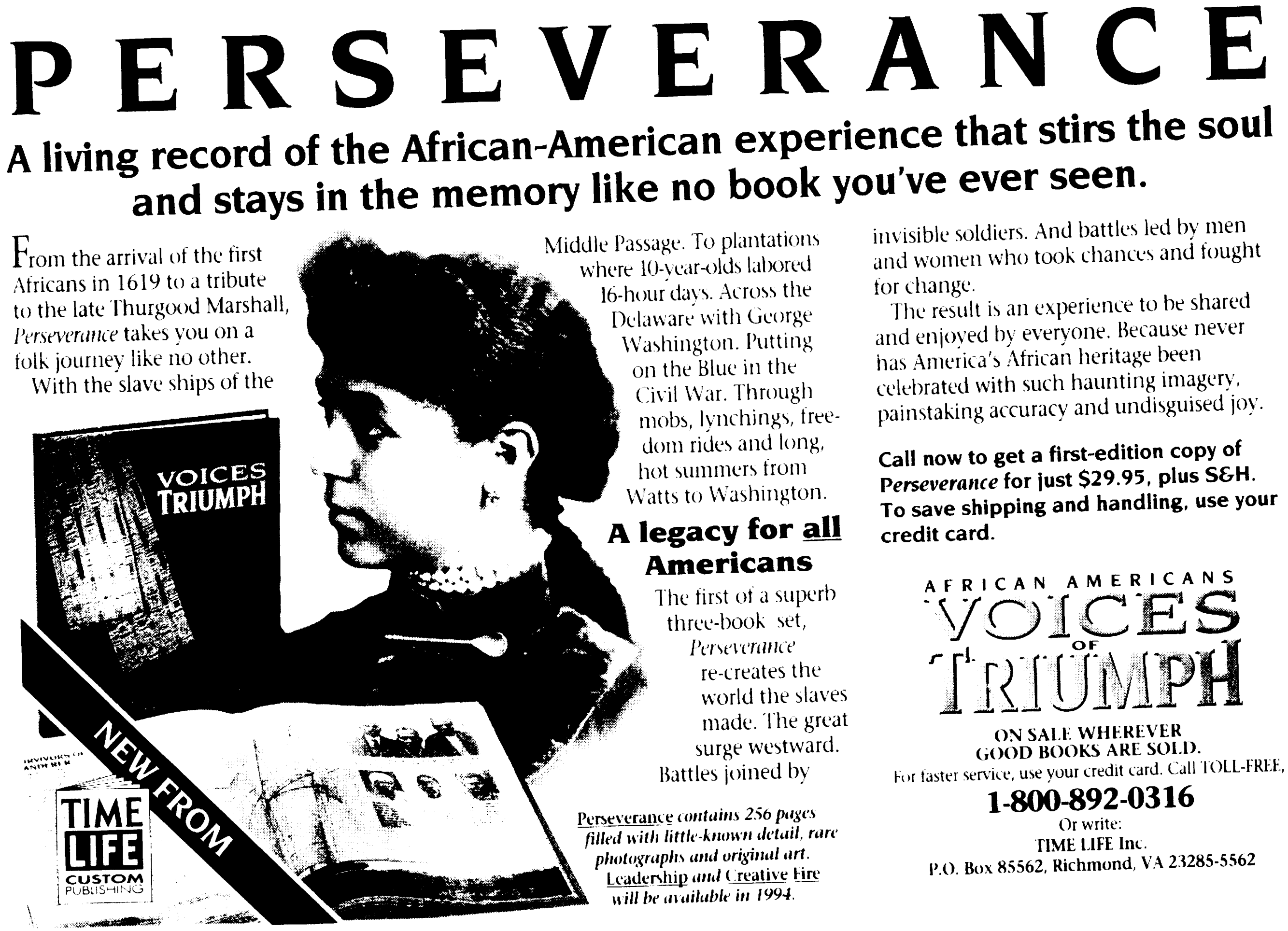
Read This if You're *CRAZY* About Statistics:

The more you know about statistics, the more you'll love STATGRAPHICS *Plus*. With its completely menu-driven procedures, spectacular interactive 3-D graphics, and a choice of DOS or Windows versions, STATGRAPHICS *Plus* wraps together everything you've ever wanted in a statistics package.

Read This if Statistics Drives You *CRAZY*:

The less you know about statistics, the more you'll love STATGRAPHICS *Plus*. Because you can use it to master statistics even if you barely *remember* statistics. STATGRAPHICS *Plus* is the easiest to learn, easiest to use statistical analysis package you can buy. And it's available in both DOS and Windows versions.

Figure 3. Statistics (638 × 1440).
© Manugistics, Inc., 1995. All rights reserved.



PERSEVERANCE

A living record of the African-American experience that stirs the soul and stays in the memory like no book you've ever seen.

From the arrival of the first Africans in 1619 to a tribute to the late Thurgood Marshall, *Perseverance* takes you on a folk journey like no other.

With the slave ships of the Middle Passage. To plantations where 10-year-olds labored 16-hour days. Across the Delaware with George Washington. Putting on the Blue in the Civil War. Through mobs, lynchings, freedom rides and long, hot summers from Watts to Washington.

A legacy for all Americans

The first of a superb three-book set, *Perseverance* re-creates the world the slaves made. The great surge westward. Battles joined by invisible soldiers. And battles led by men and women who took chances and fought for change.

The result is an experience to be shared and enjoyed by everyone. Because never has America's African heritage been celebrated with such haunting imagery, painstaking accuracy and undisguised joy.

Call now to get a first-edition copy of *Perseverance* for just $29.95, plus S&H. To save shipping and handling, use your credit card.

Perseverance contains 256 pages filled with little-known detail, rare photographs and original art. Leadership and Creative Fire will be available in 1994.

AFRICAN AMERICANS
VOICES OF TRIUMPH

ON SALE WHEREVER
GOOD BOOKS ARE SOLD.
For faster service, use your credit card. Call TOLL-FREE,
1-800-892-0316
Or write:
TIME LIFE Inc.
P.O. Box 85562, Richmond, VA 23285-5562

Figure 4. Perseverance (1048 x 1600)

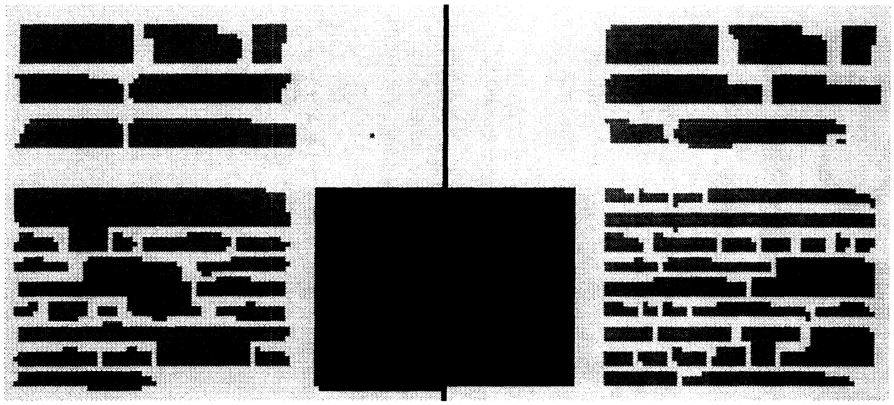


Figure 5. Segmentation of Statistics.

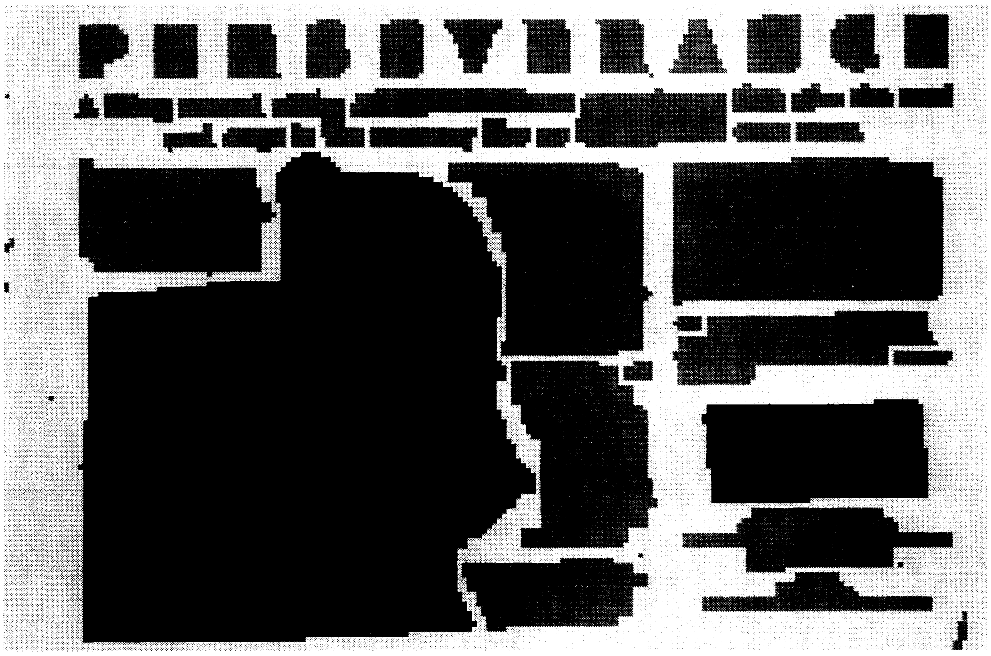


Figure 6. Segmentation of Perseverance.