PREMIER COLLOQUE IMAGE
Traitement, Synthèse, Technologie et Applications
BIARRITZ – Mai 1984 –

MODELES DE TEXTURE POUR L'ANALYSE DE DOCUMENTS
TEXTURE MODELS FOR DOCUMENT SEGMENTATION

R.G. Casey

IBM Research Division, 5600 Cottle Road, San Jose, CA95193

RESUME

A system for automatic encoding of optically scanned documents must partition a bitmap into columns of text, image areas such as drawings and photographs, and special text blocks such as captions, headings, footnotes, etc. An attractive approach to this task is to use the texture properties of printed text, i.e., those properties that have to do with the pattern statistics of an ensemble of characters, in order to discriminate among the different regions of interest.

As part of a system presently being developed, a texture model has been derived and implemented. The model assumes that the patterns in a given region are homogeneous in thickness of strokes. Using gross measurements which can be efficiently collected from binary patterns, it provides estimates of the density and line thickness of the patterns. Changes in these characteristics can then be located in order to partition the document. Other models are under investigation for detecting halftone pictures or line drawings. Sample results with scanned documents will be presented.

SUMMARY

Un système pour l'analyse automatique des régions d'un document doit pouvoir effectuer une partition du tableau binaire provenant de la digitalization d'un document, en sous-tableaux correspondant aux diverses régions (images, colones de texte, titres, notes etc.). Cet article presente une méthode de partition qui utilise les attributs de texture de caractères. Par 'texture' on entend les propriétés statistiques d'un ensemble de caractères qui determinent, en gros, l'apparence d'une région.

Le modèle utilisé part de l'hypothèse que les échantillons des traits dans une région sont uniformes en épaisseur. Des estimations de ces épaisseurs peuvent être déduites efficacement du tableau binaire. Les changements de ces estimations d'une région a l'autre servent à identifier les différentes régions. Les résultats de l'application de la méthode sont donnés pour plusieurs exemples de documents.

MODELES DE TEXTURE POUR L'ANALYSE DE DOCUMENTS
TEXTURE MODELS FOR DOCUMENT SEGMENTATION

R.G. Casey

INTRODUCTION

A number of techniques[1-4] have previously been investigated for distinguishing between printed text and surrounding material such as photographs, line drawings, or graphic symbols on an optically scanned document. Assumed characteristics of the printing, e.g., stroke width, height, spacing, etc., serve to discriminate symbols from other page components. Typical applications include preprocessing to decide which areas of a page should be treated by optical character recognition, or to choose between several compression or enhancement algorithms.

Some of the methods operate at a fine level of detail, for example, they may separate a signature from the complimentary close of a letter. A global view of the segmentation problem also has advantages. It is ordinarily more efficient to treat the document as a collection of more-or-less homogeneous regions rather than to attempt to analyze individual objects in the field of view. In this approach the characters in a segmented region are viewed as similar in characteristics in spite of the shape variations that convey information.

The global view naturally leads us to the concept of texture on a document. Just as aerial photographs of farmland can be segmented into regions of wheat and corn on the basis of systematic variations in intensity level, so it is admissible to attempt to distinguish text regions by the use of similar information. Segmentation of documents, however, differs from aerial sensing (for example) in several ways:

(1) The input is typically binary, having been thresholded in order to highlight the character information, reduce storage requirements, and increase processing speed.

(2) Processing time must be low, in order to achieve typical text reading throughputs on the order of hundreds of characters per second.

(3) The boundaries between different regions (e.g. between text fragments printed in different fonts) is often, but not always, sharply defined by white or black separating bars.

(4) The print texture has a definite orientation, which is known beforehand.

Figure 1 is an instance in which text printed in various sizes and styles is deployed along with image information on a single page. The objective in analyzing Figure 1 is to resolve it into regions containing different classes of data, then to identify each region as either image or text and to characterize it by parameters. A human-generated segmentation is shown in Fig. 2. One use of Fig.2 by a document encoding system might be to determine which information should be processed by character recognition techniques. Since character readers typically accept only a range of fonts, the system can determine which information must be stored in image form and which should be input to a classifier.

In this paper, a model is proposed in order to account for the texture of printed characters on typical documents. The model leads to the calculation of several texture parameters, namely the strokewidth and incidence rate of the character patterns in a given region. Experimental results are included in order to demonstrate the methods. The investigation was carried out as part of the Document Analysis System[5], currently being developed as a general purpose document processing system.

A TEXTURE MODEL FOR PRINTED CHARACTERS

What is referred to here as the "texture" of a printed page of characters is largely determined by the stroke width and size of the symbol patterns, rather than by shape. Since characters are taken from finite alphabets called "fonts", and since the frequencies of occurrence of the various shapes are governed by the statistics of the language in which the text is composed, the distribution of shapes is rather well-defined once the dimensional parameters are known. In this view the segmentation problem is a converse of the recognition problem, for shape is the distinguishing attribute in the latter, and insensitivity to scale is desirable.

Accordingly, the model for printed characters posed here is parameterized by size rather than shape. Basically it is assumed that any character on a page is formed from a rectangle of length L and thickness T, by bending or by cutting and pasting (Fig. 3). Given such a pattern, T and L can be determined to quite good approximation by measurements of area and perimeter. Area and perimeter of continuous rectangles are, of course, expressed by the equations

$$A=LT, \qquad P=2(L+T),$$

respectively. These equations are easily solved for L and T as functions of the area and perimeter. These equations require slight modification for the discrete case. In addition, the analysis here will be concerned with determining average text attributes over a region rather than for individual patterns.

Suppose that when the rectangle is bent into a curve, the increase in length on one side is equal to the decrease on the other. The perimeter of the rectangle is unchanged by such a deformation, and the area varies at most slightly. Consider bending into a semicircle, for example, in which case both the area and perimeter remain fixed (Fig.3b). If the deformed rectangle is cut into sections which are pasted at the cut ends so as to form a single connected figure (Fig.3c), then both area and perimeter are still those of the original rectangle. This model of bending, cutting and pasting is sufficient to account for the formation of all except multiple component characters such as the letters i, j, or the symbol =, etc. Since, as we shall show, it is a simple matter to calculate the rectangle dimensions from its area and perimeter, it follows that we can estimate character thickness and line length from area and perimeter measurements for those character patterns that follow the rules of the model.

Further simplifying assumptions will be necessary in order to arrive at an efficient processing algorithm. It will be assumed that within a single text region the width parameter, T, is constant, although the line length, L, may vary from one pattern to another. This assumption is justified by the argument that text units, such as paragraphs or sections, are largely printed in a single font, though italics or boldface type may occur here and there. Even within a given font, of course, the widths of different strokes may differ slightly. However, since the analysis is aimed, not at achieving a precise determination of parameter values, but rather at gathering rough estimates in order to discriminate between pattern areas having gross differences, this effect will be ignored.

MODELES DE TEXTURE POUR L'ANALYSE DE DOCUMENTS
TEXTURE MODELS FOR DOCUMENT SEGMENTATION

R.G. Casey

MEASUREMENT OF AREA AND PERIMETER

The patterns to be dealt with are not continuous shapes after scanning, but rather are represented as arrays of black and white pixels. Thus the area of a pattern is measured by the number of pixels it contains. Likewise the perimeter is calculated as the number of edge pixels, i.e., those pixels that are black, but have at least one white neighbor pixel. In order to estimate area and perimeter directly, processing must first be performed in order to isolate the character patterns. On the other hand, we seek a gross method that operates on the entire pattern field in an efficient, systematic fashion. Isolating individual patterns involves detailed processing and thus a cost penalty that is undesirable during the preliminary decomposition of the document.

The algorithm proposed here superposes a rectangular grid on the document field. Within each rectangular grid section, both the number of black pixels and the number of edge pixels are counted. These two numbers express the total area and perimeter, respectively, of the patterns within the grid section.

In taking measurements in this way an extra unknown parameter is incurred, namely the number of characters within the rectangle, N. Boundary effects are also encountered due to patterns that intersect grid lines. This will be considered in a later section; for the moment it will be assumed that only complete characters lie within a grid region.

Assuming fixed strokewidth, T, the total pattern area, A, and perimeter, P, may be expressed as follows

$$A = NTA$$

$$P = 2N(T + A-2)$$

where A = the average line length of the character patterns in the grid section.

These two equations contain three unknown variables, and thus an additional relationship is required in order to solve the system. In order to do this we note first that the average line length, A , tends to vary little from one grid section to another within a given text region since it is an average over N samples in a single font. If, in addition, the differences between the fonts that constitute the different text regions are primarily due to size changes (as in newspaper headings, for example) then the ratio

$$A /T = r$$

is a constant which can be estimated a priori, providing the needed third relationship.

We thus have

$$A=rNT^2$$
$$P=2N(T+rT-2)$$

which, when solved for N and T, yields

$$T = R + (R^2 - 4\mu/r)^{0.5}$$

where $\mu = A/P$, and $R = \mu(1+r)/r$.

Parameter N is computed, once T is found, by $N = A/rT^2$.

Note in particular that the strokewidth estimate is a function of the ratio of area to perimeter in the region examined.

These calculations can be implemented efficiently on a digital computer. Typically the rows or columns of the scanned document will be packed sequentially into words of fixed length, with a '1' bit denoting that the corresponding pixel is 'black', and a '0' bit denoting white. The size of the grid region is chosen to be a multiple of the wordlength, so that one grid section corresponds to an array of words in the bitmap. Area of the pattern within a section is computed by counting the number of 1 bits in the array. Next, the bitmap is repeatedly shifted and ANDed with itself, to create an array containing only interior pixels. The 1's in this array are counted and subtracted from the area count in order to compute the perimeter.

EFFECTS OF DEPARTURES FROM THE ASSUMPTIONS

The model holds exactly only if the average line length of the N characters in a measurement rectangle is equal to a known constant, r, times the stroke width, T, and if the rectangle contains only complete characters. In this case N and T can be computed from measurements of area and perimeter. In this section we consider the errors introduced if r varies from one region to another, or if some patterns intersect the boundary between adjacent measurement rectangles.

Examining the relationships for N and T, we see that for fixed area and perimeter an increase in r produces an increase in the estimate of the number of patterns per unit area, but a decrease in calculated stroke width. Thus, if r is assigned too high a value, then T is underestimated while N is overestimated.

In order to gauge edge effects, suppose that r is fixed, but that there are N_1 complete patterns and N_2 patterns that intersect the boundary. The effect on calculations of N and T depend on the ratio of area to perimeter for the partial patterns in comparison with the value of this ratio for the complete patterns. Assume, for example, that area and perimeter divide proportionately across the boundary, i.e., for a boundary pattern having total area A and perimeter P, the portions measured within the grid region are

$$A'=kA$$
$$P'=kP$$

where the value of k may vary from one pattern to another. It follows that the overall area/perimeter ratio is unchanged, so that strokewidth T will be computed correctly, while N is computed as the number of whole patterns having area equal to that observed within the grid region. On the other hand, if area and perimeter divide nonproportionately

$$A'=kA \quad P'=mP$$

where k,m are unequal, then an error is introduced. Analysis shows that if k>m, then strokewidth is overestimated, and conversely for k<m.

The effect of this error is reduced as the relative number of whole to boundary patterns increases. For good estimation it is desirable to choose a large grid size. On the other hand, the resolution, i.e., the fineness of discrimination between regions

MODELES DE TEXTURE POUR L'ANALYSE DE DOCUMENTS
TEXTURE MODELS FOR DOCUMENT SEGMENTATION

R.G. Casey

having different textures decreases as the grid size is increased. Thus, choice of grid size represents a compromise.

EXPERIMENTS

Figure 4 shows a bitmap of a journal abstract, optically scanned at a resolution of .1 mm per pixel. Overall array size is approximately 700 rows by 1600 columns. In Figures 5a and 5b are shown the array of values obtained for N and T when the model is applied to this data using a grid size of 64 pixels and a length to width estimate (r) of 8. Note that despite edge effects the results yield reasonable estimates of T and N. In Figure 6 the capability to detect the various character sizes is shown by means of a gray scale map of Figure 5b. By local smoothing of maps it is possible to segment the text regions according to size.

We are interested not only in the estimation of strokewidth in order to discriminate different text regions, but also in the characteristics of the model when applied to other types of printed information, e.g., photographs or line drawings that might be on a page. Thus, in Fig 7 is shown a column of print containing several half-tone photographs. The mesh of a photograph, scanned into a bitmap, yields an array of small closely spaced black areas (Fig. 8), which can be distinguished from text regions by a grayscale map of the estimates of N (patterns per grid area) as shown in Fig. 9.

On the other hand, with line drawings the model is less successful. The lines used to construct the embedded drawings in Fig. 10 are on the same scale as the strokewidths of the text as a grayscale plot of parameter T shows (Fig.11). Although there is some irregularity in the distribution of values that might help with the discrimination, it is probably necessary in general to invoke other methods, e.g., that of Wahl[6] in order to dichotomize drawings from text. Note, however, that in all these examples the estimates of N and T for text regions alone are homogeneous and consistent.

CONCLUSIONS

A notion of the "texture" of lines of printed characters has been pursued with the aid of a simple model for character formation. The resulting analysis yields estimates of the density of patterns and the strokewidth in a homogeneous region from measurements that are efficiently performed on binary bitmaps. The model shows an ability to distinguish halftone image data from text, and to discriminate among regions of different type sizes. However, it is not sufficient in itself to separate line drawings from text, but requires auxiliary pattern analysis techniques as well.

ACKNOWLEDGMENTS The author is grateful to Prof. G. Stamon, whose encouragement and knowledge, provided during an 8-week stay at the author's laboratory, were essential in producing this study. Mention is due also to Dr. F. Wahl of IBM, whose work on separation of text and image on documents influenced the investigation. Finally, the support of Dr. Kwan Wong, director of the Document Analysis project in the IBM San Jose Research Laboratory, is acknowledged.

REFERENCES

1. F. M. Wahl, K. Y. Wong, and R. G. Casey, "Block segmentation and text extraction in mixed text/image documents", Computer Graphics and Image Processing, 20, 375-390 (1982).

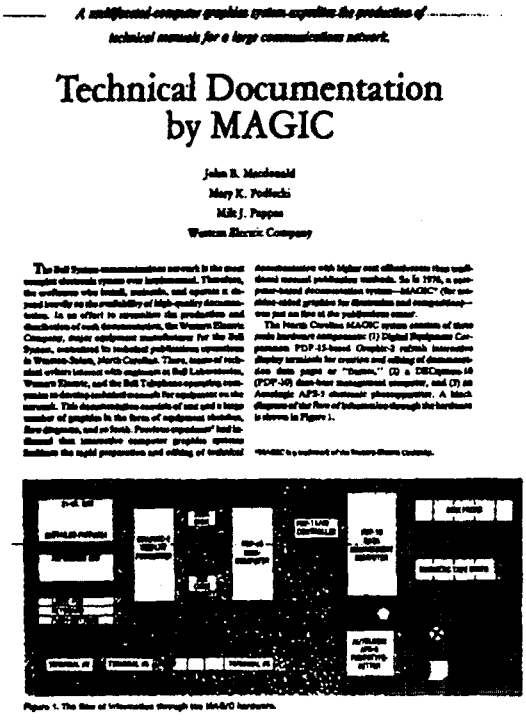
2. E. G. Johnston, "Printed text discrimination", Computer Graphics and Image Processing, 3, 83-89 (1974).

3. W. Scherl, F. Wahl, and H. Fuchsberger, "Automatic separation of text, graphic, and picture segments in printed material", in Pattern Recognition in Practice, pp213-221, E. Gelsema and L. Kanal, Eds., North-Holland, Amsterdam, 1980.

4. K. Wong and P. Stucki, "Adaptive switching of dispersed and clustered halftone patterns for bi-level image rendition", Research Report RJ2020, IBM Research Laboratory, San Jose CA95193, June, 1977.

5. F. Wahl, "A new distance mapping and its use for shape measurement on binary patterns", Computer Graphics and Image Processing, 23, 218-226 (Aug. 1983).

6. K. Y. Wong, R. G. Casey, F. M. Wahl, "Document analysis system", IBM Journal of Res. and Dev., vol. 26, n6, Nov. 1982, pp647-656.



technical manuals for a large communications network.

Technical Documentation by MAGIC

John R. Macdonald
Mary K. Podlucki
Mike J. Poppas
Western Electric Company

Figure 1. Sample document containing image as well as text in various sizes.

MODELES DE TEXTURE POUR L'ANALYSE DE DOCUMENTS
TEXTURE MODELS FOR DOCUMENT SEGMENTATION

R.G. Casey

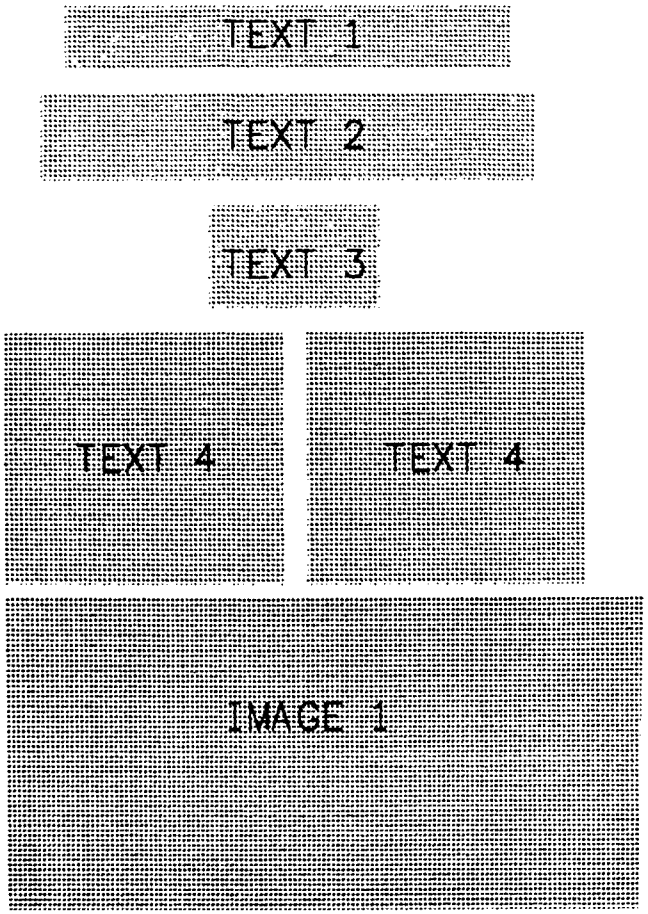


Figure 2. Manual segmentation of the document of Fig. 1

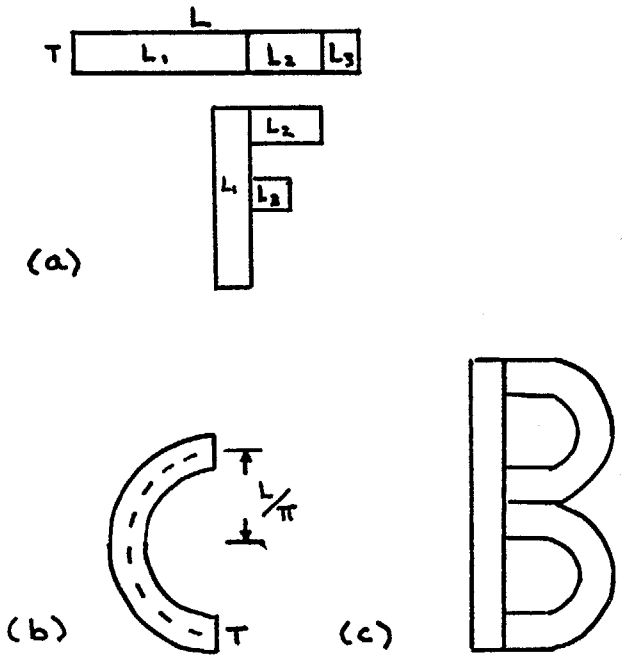


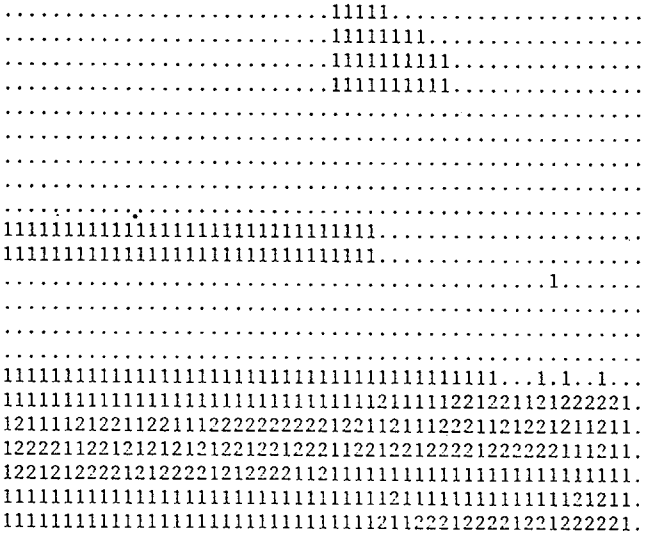
Figure 3. Model for character formation: (a) basic rectangle used for constructing characters, (b) deformation of (a) to a semicircle, (c) Cut and paste model of character formation.

J. Veldsman
Leo W. Hasvel

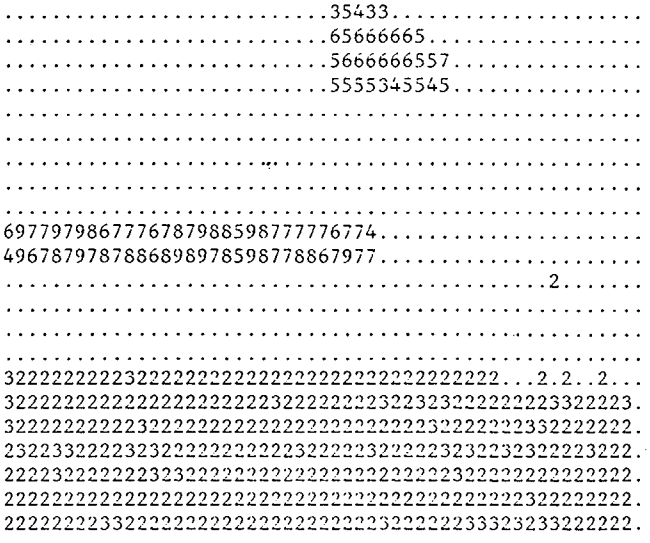
The Software-Cache Connection

This paper describes an adaptation of standard Fourier analysis techniques to the study of software-cache interactions. The cache is viewed as a "black box" boolean signal generator, where "ones" correspond to cache misses and "zeros" correspond to cache hits. The spectrum of this time sequence is used to study the dynamic characteristics of complex systems and workloads with minimal a priori knowledge of their internal organization. Line spectra identify tight loops accessing regular data structures, while the overall spectral density reveals the general structure of instruction localities.

Figure 4. Sample text bitmap



(a) N



(b) T

Figure 5. Arrays of estimates of (a) N, and (b) T for the bitmap of Fig. 4.

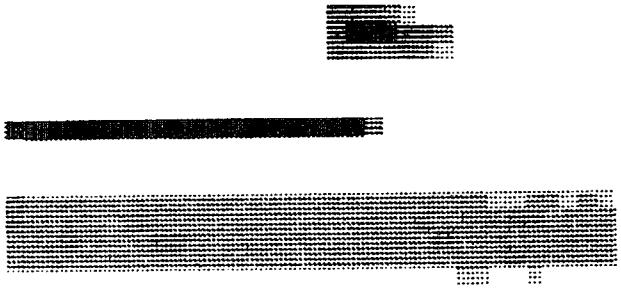


Figure 6. Grayscale map of values of T from Fig. 6b.

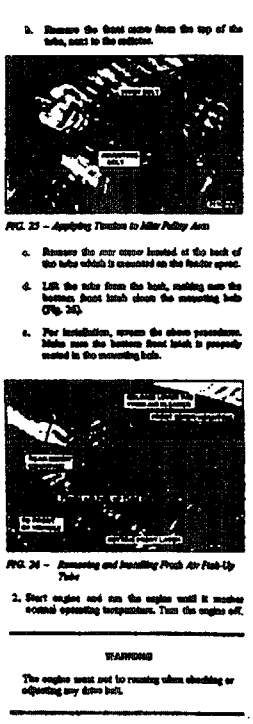


Figure 7. Text and halftone image on a document.

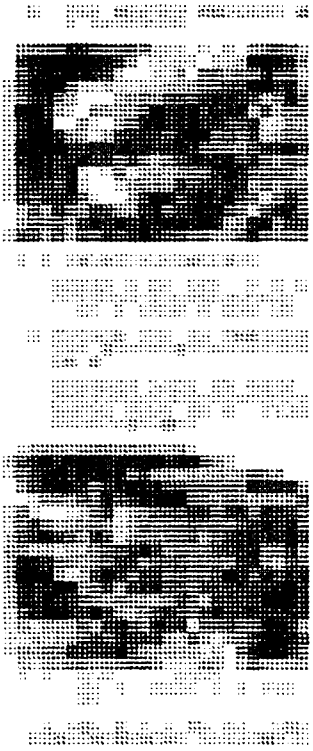


Figure 9. Grayscale map of N as estimated from the bitmap of Fig. 8.

Figure 8. Detail of image from Fig. 7, showing closely spaced halftone dots.

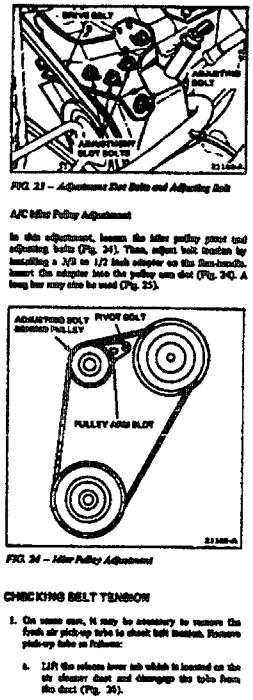


Figure 10. Line drawings and text on a document.

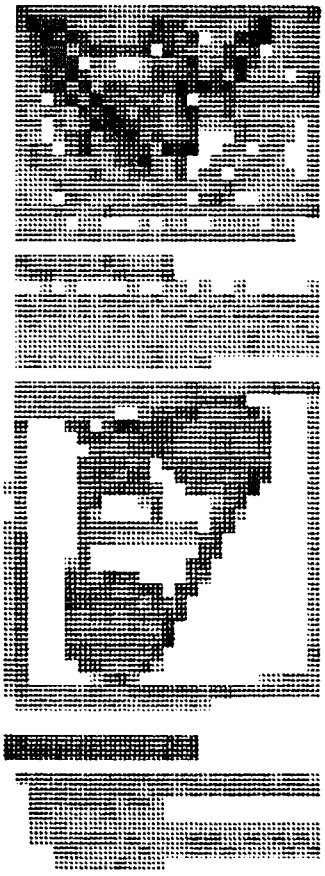


Figure 11. Grayscale map of the parameter T from Fig. 10. Neither this figure nor the map of parameter N show appreciable global differences in value that could be used for differentiating the drawing region from the text areas.