



Contributed Paper

Identification of Text-Only Areas in Mixed-Type Documents

C. STROUTHOPOULOS

Democritus University of Thrace, Greece

N. PAPAMARKOS

Democritus University of Thrace, Greece

C. CHAMZAS

Democritus University of Thrace, Greece

(Received July 1996; in revised form December 1996)

The identification of text areas in a document is crucial for optical character recognition (OCR), image-compression and image-storage systems. This paper presents a new method for text identification in mixed-type documents. This type of document contains text, drawings and halftones. The proposed method separates the document into text and non-text regions. Thus, the objective is to find, with confidence, the text region of the documents. The method is based on text characteristics such as size, frequency, collinearity and vicinity of connected components, while in the final stage a new texture-analysis technique is applied. For collinearity and vicinity checking, a new technique is used, that overcomes the difficulties of the application of Hough transform. The proposed segmentation method belongs to the bottom-up categories, and is more robust than other techniques. It can identify text regions in difficult cases such as skewed documents, non-rectangular text regions, or text included in drawings or halftone regions. The performance of the method was tested on a variety of images. Its effectiveness is demonstrated by several typical examples.

© 1997 Elsevier Science Ltd. All rights reserved

Keywords: Document image analysis, segmentation, page layout analysis, OCR.

1. INTRODUCTION

Most digitized documents contain halftone pictures and line drawings, along with text. An important procedure in the digital processing of documents is page layout analysis. The goal of the page layout analysis is to discover the formatting of the text and, from that, to derive the meaning associated with the positional and functional blocks in which the text is located (O'Gorman and Kasturi, 1995; Fujisawa *et al.*, 1992; Schurmann *et al.*, 1992). As a result, it labels the parts of a document as text, halftones or line drawings (Pavlidis and Zhou, 1992). Page layout analysis is an important preprocessing procedure for many other applications. For example, in optical character recognition (OCR) the seg-

mentation is a necessary pre-processing procedure. Also, the compression ratio can be improved if the text and image areas are encoded using different methods. The best archiving of mixed-type documents using block segmentation and recognition also requires text-area definition (Chauvet, 1993). Additionally, the identification of text areas is also important for other special applications such as CAD/CAM and communication systems.

There are many problems in document page-layout analysis: identification of image areas, separation of characters included in an image, identification and extraction of text-blocks, and separation of overlapping and touching characters. This paper deals with the problem of the automatic identification of text areas of a document, and the proposed segmentation method answers the question "Where in the document do we have only text?"

In the literature there are three basic segmentation approaches: top-down (or model-driven), bottom-up (or

Correspondence should be sent to: Associate Professor Nikos Papamarkos, Electric Circuits Analysis Laboratory, Department of Electrical & Computer Engineering, Democritus University of Thrace, 67100 Xanthi, Greece [E-mail: papamark@voreas.ee.duth.gr].

data-driven), and hybrid (Pavlidis and Zhou, 1992). In top-down methods, a document is segmented from large components (high-level) to smaller, more detailed, sub-components (lower-level). Most of top-down techniques are based on the *run length smoothing* (RLS) algorithm (Wong *et al.*, 1982), also called the *constrained run length* (CRL) algorithm (Wahl *et al.*, 1989), and the projection profile cuts (Wang and Shihari, 1989). The RLS method imposes a smoothing on the binary form of the document using two predetermined parameters (one for the vertical and one for the horizontal direction) defining the document blocks. By using this method, it is possible to segment even handwritten text blocks (Witten *et al.*, 1994). For the block classification, additional parameters (defined in a heuristic way) are used, leading to the necessity to train the system with documents having similar fonts or other common morphological characteristics. A primary advantage of the RLS method is that it can achieve layout analysis fast. For most page formats, this is a very effective approach. It must be noted that the method is not robust since, if the assumptions made to determine the heuristic parameters are not satisfied, the method will fail. Another disadvantage of this method is that the page must be separable into blocks by horizontal and vertical cuts [Manhattan layout (O'Gorman and Kasturi, 1995)]. Hence, for pages where text does not have linear bounds, and where graphics are intermixed both in and around text, these methods will fail.

Bottom-up methods involve the grouping of pixels as connected components (marks) and merge these components into successively larger regions. In one of them, Fletcher and Kasturi (1988), Kasturi *et al.* (1990), and Kasturi and Trivedi (1990) proposed a procedure that starts by first finding the connected components of an image, and then separating the graphics from the text using the relative frequency of the occurrence of components as a function of their areas. In the next step, they use an iterative procedure to improve the initial estimation by applying the Hough transform to all connected components. The method performs well if the initial document conforms to certain requirements about the size of characters, interline spacing, character spacing, and resolution. This method is quite complex and, as the authors have reported, computationally expensive. Another bottom-up technique proposed by O'Gorman (1993) is based on a document spectrum (docstrum) for k -nearest-neighbor marks. In this approach, the first step is to determine the marks and their centroids. The direction vectors (distances, angles) for the k -nearest-neighbor connections of each mark are accumulated in a histogram. This histogram has two major peaks: one corresponding to intercharacter spacing, and the other corresponding to the document skew. Using these peaks, a complex process groups the characters into words and text lines. This method, including both skew estimation and layout analysis, is computationally expensive and ineffective in multi-size and short text strings. Using Gabor filters, which have been used earlier for the general problem of texture segmentation (Farrokhinia, 1990), Jain and Bhattacharjee (1992) propose a method for text segmentation of

gray-level documents. The method works well with low-resolution documents and is robust to skew. Unfortunately, as the authors have reported, because of the frequency domain usage, the method is time-consuming, requiring about two minutes of a workstation CPU time for a 512×512 image. Finally, there are some methods that cannot be classified as either top-down or bottom-up. These methods are considered as hybrid. On this point, it is noted that the computational effort is a crucial criterion for the evaluation of document-segmentation techniques. Of course, the computational time must be significantly less than the time required if a character feature-extraction identification scheme is used.

This paper proposes an effective and fast bottom-up method that can classify and separate text from non-text regions for a wide class of documents. In addition to the identification of text regions, the proposed method can also identify line drawings and halftone regions. The method does not use morphological character features, and does not deal with the detailed extraction of isolated characters. The proposed technique is independent of the size and type of characters and the position of text and graphics in the document, and is tolerant to small skew errors. Its basic stages are:

- Identification of marks. Inclusion of each mark in a bounding box. Construction of the height histogram.
- Determination of the heights of the accepted boxes according to a histogram peak.
- Extension of boxes, and construction of bounding rectangles.
- Filtering of rectangles according to their base:height ratio.
- Texture feature extraction and classification.

The proposed method has two considerable advantages. The first is the substitution for the traditional and computationally expensive techniques for collinearity and vicinity checking of marks, by a new, simpler and faster one, which is based on box extensions. The other advantage is gained by the use of a new texture feature-analysis algorithm, which classifies the areas of the bounding rectangles as either text or non-text regions. The outcome of the method is the identification of areas containing only text. The remaining parts of the document might contain graphics and isolated characters.

The method was tested with many documents which contained text, line drawing and halftones. In this paper, typical examples are presented that cover special types of documents. The experimental results confirm the effectiveness of the proposed method, which has been implemented on a Pentium-100 PC using the C++ programming language. In this implementation, an average time of 5 sec per page of A4 size and 150 dpi resolution is required, without any effort to optimize the code for speed.

A General Purpose Signal Processing Package

NEW Extended Version of sig

New Menu Interface for occasional users, On-line HELP, and Command mode for experienced users.

Signal Processing Operations Include:
 Windowing, convolution, Fourier transforms, interpolation, decimation, digital filtering, linear systems, simulation, correlation, ensemble operations, spectral estimation, coherence, parametric and adaptive processing, deconvolution, identification and more...

Signal Processing Operations Include:

- Model Based Signal Processing
- Identification
- Simulation
- Linear Estimation (Kalman Filter)
- Nonlinear Estimation (Extended Kalman Filter)

Graphical Operations Include:

- Plotting
- Families of Curves
- Multiple Viewport Plots
- Many Graphics Devices

New Features Include:
 Surface (3D) plots, contour plots, multichannel signal and complex operations, 2D Fourier transforms, additional algorithms (adaptive lattice, MLM, Burg, spectral estimation)

On-line HELP Package

Menu mode for most operations and command mode for graphics

Fig. 1. Original binary image.

2. DESCRIPTION OF THE ALGORITHM

Usually, segmentation techniques tend to be iterative (O'Gorman, 1993) because they must segment documents with text of different sizes and types. The method described here for text classification of mixed-type documents, in each of its iterations, is based on the following general but reasonable assumptions about a text line:

- Text lines consist of characters (letters, numbers, symbols) of almost the same height, which are aligned in a straight line.
- The distance between neighboring characters of the same word is less than the size of their heights.
- If the characters in a text line are joined, then they can be enclosed in elongated bounding rectangles.

The number of necessary iterations is proportional to the main concentrations observed in the height histogram. These concentrations are obtained automatically by using a hill-clustering technique (Tsai and Chen, 1992; Papamarkos and Gatos, 1994) and correspond to characters of significantly different sizes. Capital letters, small letters, subscript and superscript type letters are examined together in the same iteration. Usually, at least four iterations are needed to cover the sizes of all the letters. The method is applied in documents having a binary form and, in summary, is composed of the following stages.

Step 1: In the original document, I_1 , marks are extracted and identified, and then surrounded with bounding boxes. Next, a histogram is formulated from the heights of the bounding boxes.

Step 2: The histogram's peaks are determined, using the hill-clustering method.

Step 3: For the iteration corresponding to a histogram peak value H_{max} , those boxes are accepted whose height h satisfies the relation $H_{max}/2 \leq h \leq 2H_{max}$.

Step 4: The initial rectangular shape of each bounding box is extended to give a new image, I_2 , which has touching boxes.

Step 5: The touching boxes derived in the previous step, are surrounded with bounding rectangles, and then a filtering is performed according to their base:height ratio.

Step 6: In the areas of I_1 , defined by the bounding rectangles, a texture feature-extraction technique is applied, which then classifies them into four classes: first class of text, second class of text, class of halftones and class of drawings. The first class of text corresponds to normal-type characters, while the second text class includes mainly italic-type characters or numbers. Only areas in the first two classes are considered as text areas.

Step 7: For the next histogram peak with a smaller value than the previous one, Steps 3 through 6 are repeated until no other histogram peak exists.

These steps of the entire text-identification method are analyzed below.

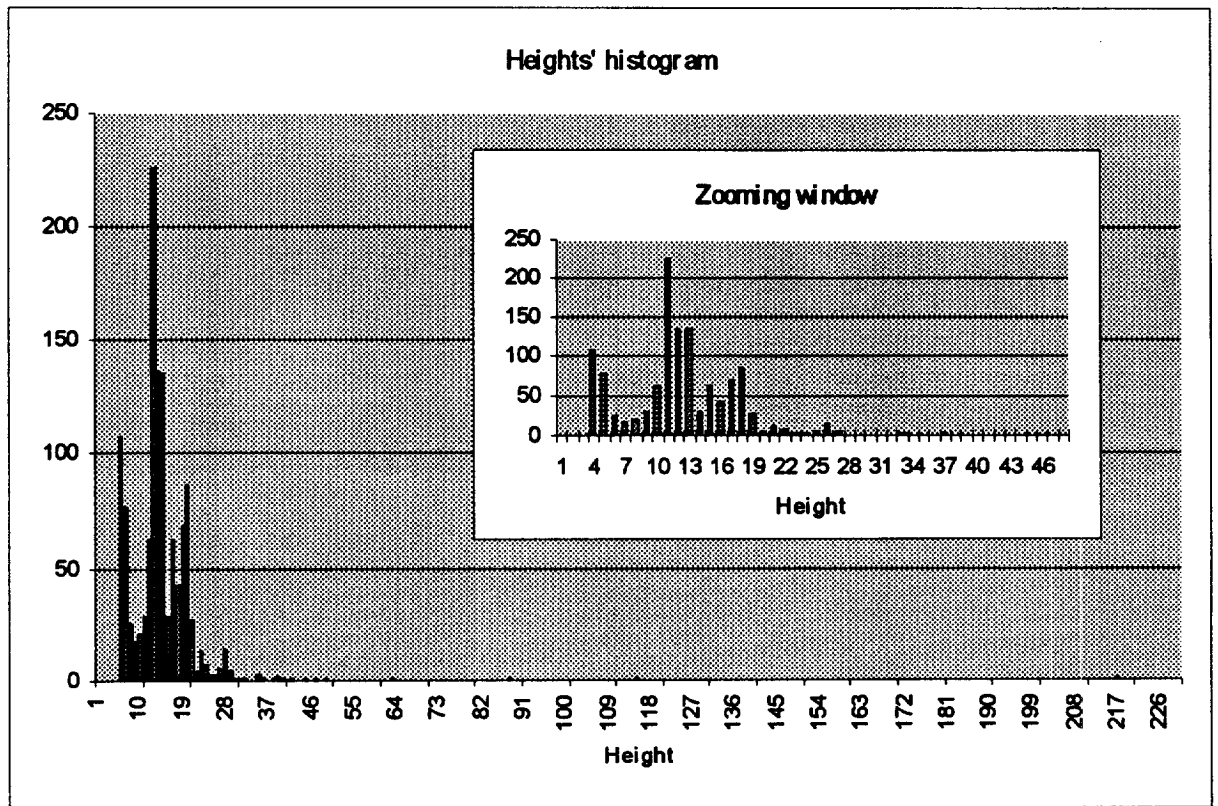


Fig. 2. Height histogram.

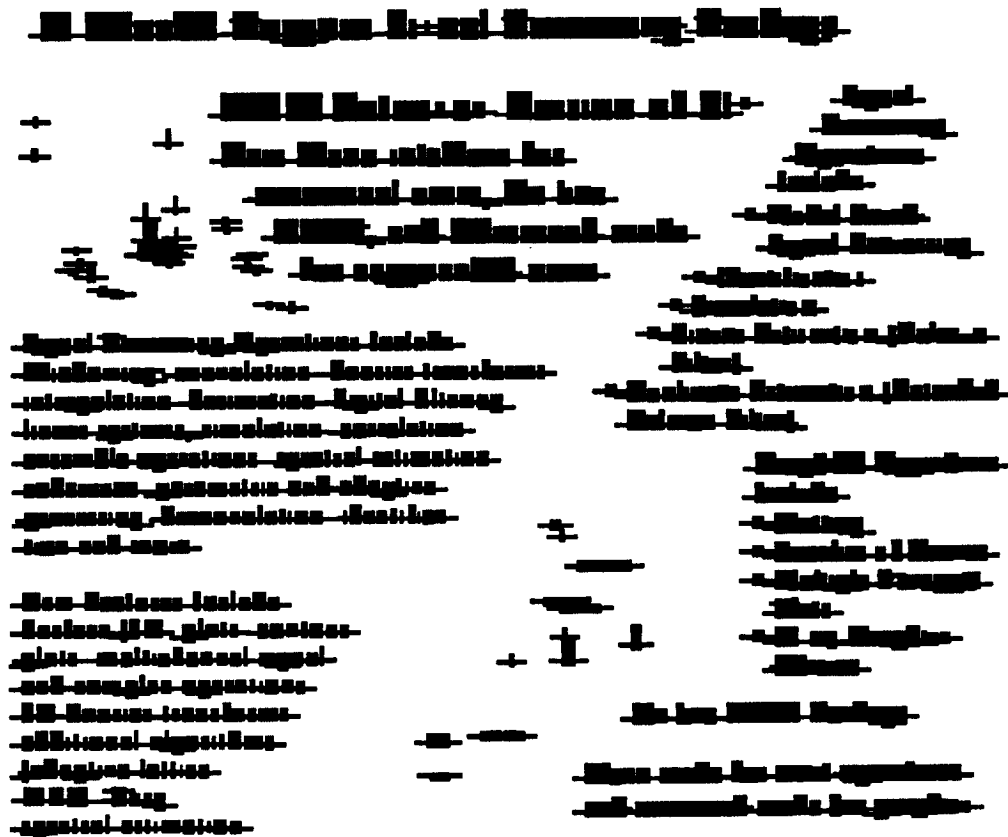


Fig. 3. Extension of the bounding boxes after the height filtering.

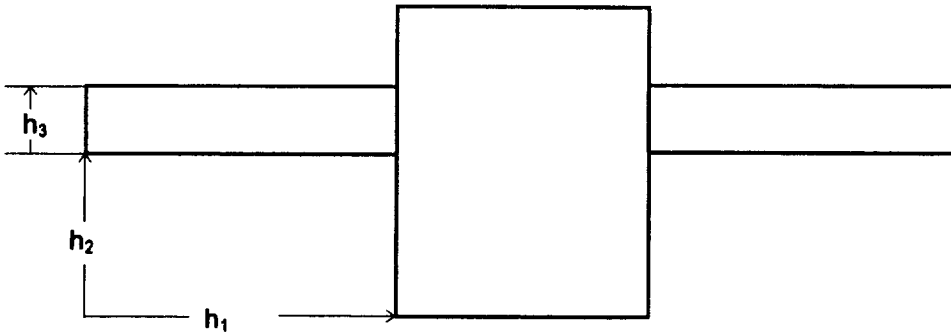


Fig. 4. Form of extension of a bounding box.

2.1. Identification of marks and formulation of the height histogram

As defined in the introduction, a “mark” is a connected group of black pixels of the document image. Therefore, a mark is a set of pixels that includes all object pixels having at least one path leading to other pixels in the set. In a text region, a mark usually corresponds to a character, or to a disconnected part of a character. For effective mark determination, a fast *contour following algorithm* (CFA) is used, which is based on the extraction of the marks’ boundary, without any morphological restriction (Pratt, 1991). All the stages of the segmentation method will be shown by applying it to the document of Fig. 1. This document is of low resolution (100 dpi), and is complex

because it contains text of different size and types, text mixed with graphics, and graphics. Next, the height histogram of these boxes is formulated in the following way.

Suppose that there are N boxes, and each box i has height h_i . Also, let $H_j, j=1,2,\dots,K$ be the function giving the K different heights. The height histogram is derived by the relation:

$$F(H_j) = \begin{cases} F(H_j) + 1, & \text{if } H_j = h_i \\ F(H_j), & \text{if } H_j \neq h_i \end{cases} \quad (1)$$

for $i=0,1,\dots,N-1$ and $j=1,2,\dots,K$.

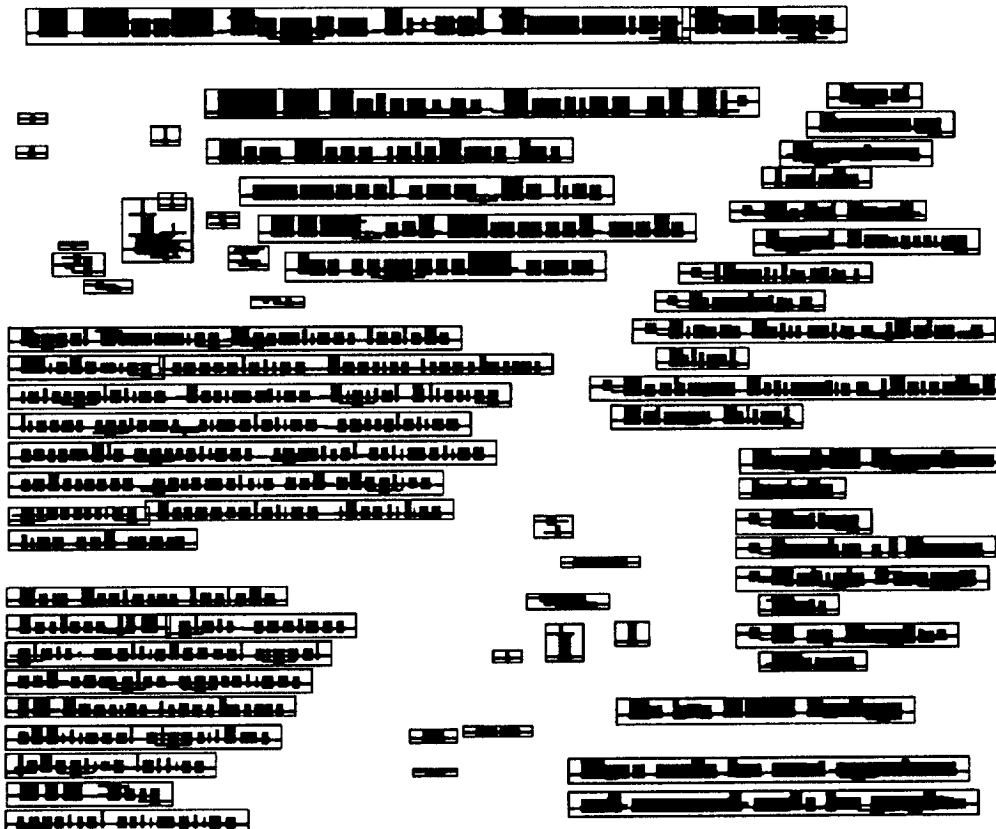


Fig. 5. Chains of overlapping boxes and the new rectangles which surround the chains.

b_8	b_7	b_6
b_5	b_4	b_3
b_2	b_1	b_0

Fig. 6. Order of pixels in a DSE.

For the document of Fig. 1, Fig. 2 shows the height histogram.

2.2. Determination of the accepted boxes according to their heights

In each iteration of the method it is accepted that the characters have almost the same height in any text area. Taking advantage of this, the algorithm continues by finding those bounding boxes whose heights appear very often in the document. To accomplish this, the histogram described above is used. To determine the distributions and the peaks of the histogram, the well-known hill-clustering method (Tsai and Chen, 1992; Papamarkos and Gatos, 1994) is used. The result of this procedure is the determination of the histogram's peaks, which correspond to the main distribution of heights in it. For a typical document, there is often a global peak for the distribution of characters of the predominant size, and, smaller peaks for the rest of the characters and noise. Therefore, $F(H_i)$ takes a global maximum value for the boxes that bound the characters of the most common size.

After the extraction of the histogram's peaks, the algorithm proceeds iteratively in a peak-by-peak manner, starting from the biggest to the smallest peak, until no other peak exists. For each iteration, corresponding to the H_{\max} histogram peak value, only those boxes with heights h_i satisfying the following condition are accepted

$$\frac{H_{\max}}{2} \leq h_i \leq 2H_{\max} \quad (2)$$

The coefficients 1/2 and 2 have been obtained by examination of the relationship between the heights of the upper and lower-case characters. Similar coefficients are used in (Fletcher and Kasturi, 1988).

2.3. Extension and connection of the bounding boxes

After the construction and filtering of the bounding boxes, they are extended to give chains of connected boxes. This procedure is important for many reasons, mainly for the determination of the text lines. Also, this procedure gives independence from small skews in documents. Figure 3 shows the remaining boxes, after the height filtering and after having them extended by adding to their left and right sides two equal rectangular extensions, called "box-hands". This procedure is depicted in Fig. 4. The value $h_1 = H_{\max}$ has been chosen appropriately to bring out the intersection of

adjacent and extended boxes which are collinear and where the distance between them is less than $2H_{\max}$. The values $h_2 = H_{\max}/2$ and $h_3 = H_{\max}/4$ have been chosen so that the boxes of tall characters (like *h*, *l*) or of characters with a tail (like *p*, *q*), can be joined with their adjacent boxes. Value h_3 is enough for the connection of bounding boxes of text lines with a small skew. A larger value of h_3 produces a larger tolerance to skew. This procedure results in the joining up of the bounding boxes of a text line, and the creation of chains of boxes which correspond to the text lines. In practice, this procedure is substituted for the complex and time-consuming use of the Hough Transform for collinearity and vicinity checking of the bounding boxes (Strouthopoulos *et al.*, 1995).

2.4. Filtering of rectangles

Considering a new binary image I_2 as the result of the previous step, each mark (which now is composed of extended-connected boxes) is surrounded with new rectangles using the CFA. It is obvious that those rectangles that enclose extended boxes of a text line are elongated. In contrast, boxes that are located too far from their adjacent boxes, or boxes that constitute a small isolated group, either do not create overlapping boxes or create chains with relatively small lengths. Figure 5 shows the chains of the overlapping boxes, and the new rectangles which surround the chains. The new rectangles that have a base:height ratio greater than 3.5 are accepted. The threshold value 3.5 corresponds to the case with only two connected characters.

2.5. Texture classification

To classify the regions of the bounding rectangles as text or non-text areas, a texture-classification technique is applied. This technique is based on 13 powerful structural features which are called *document structure elements* (DSE). A DSE is any 3×3 binary block. The order of the pixels of such a block is shown in Fig. 6.

Assign to any DSE an integer $L = \sum_{i=0}^8 b_i 2^i$, and call L the *DSE characteristic number* (DSECN). It is obvious that since $L \in \{0, 1, 2, \dots, 511\}$, there are $2^9 = 512$ different blocks of DSEs, and there is a one-to-one correspondence between DSEs and their DSECNs. For a rectangular area A , let K be the number of columns and J the number of rows. It is obvious that A has $(K-2)(J-2)$ DSEs, and for any i DSE there is a characteristic number ℓ_i . The DSEs histogram function $G(L)$ of area A is derived by the relation

$$G(L) = \begin{cases} G(L) + 1, & \text{if } \ell_i = L \\ G(L), & \text{if } \ell_i \neq L \end{cases} \quad (3)$$

for $i = 0, 1, 2, \dots, (K-2)(J-2) - 1$ and $\ell_i, L \in \{1, 2, \dots, 510\}$. Note that DSEs 0 and 511 are not considered because they correspond to pure background and object regions, respectively.

For this histogram, the probability density function $S(L)$ is equal to

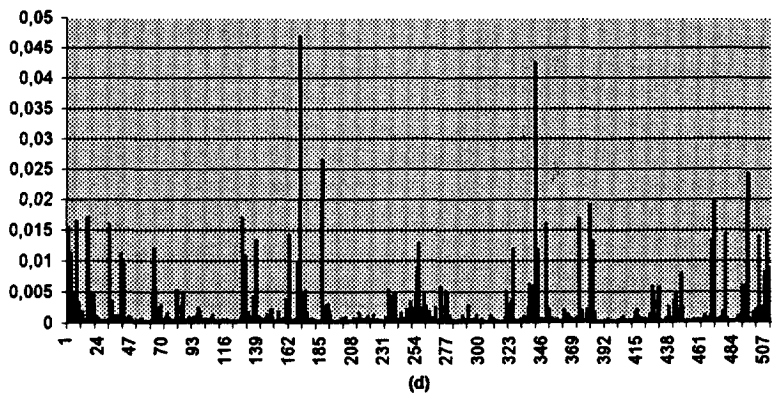
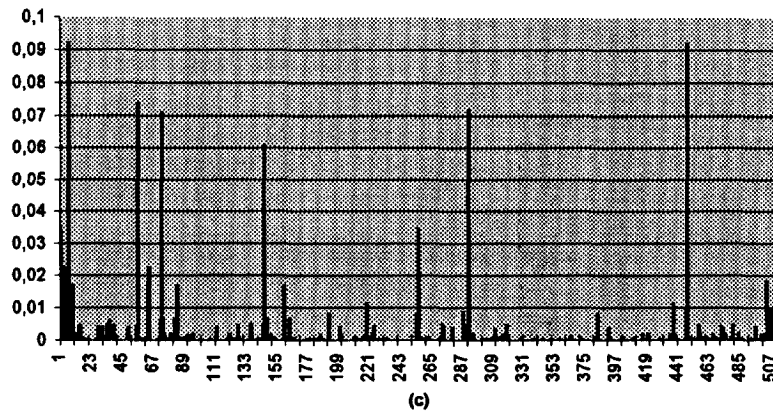
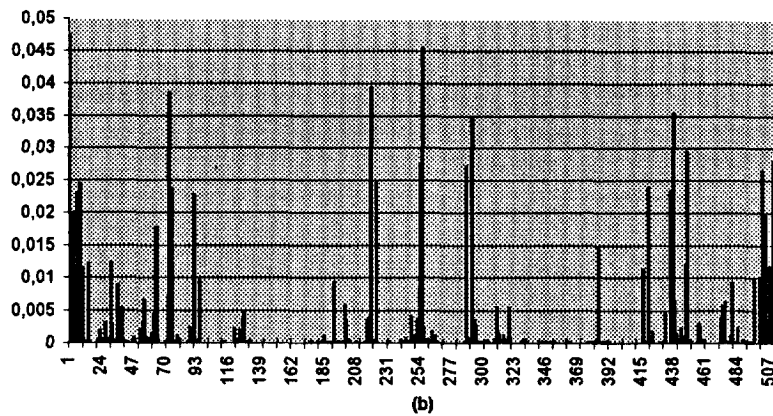
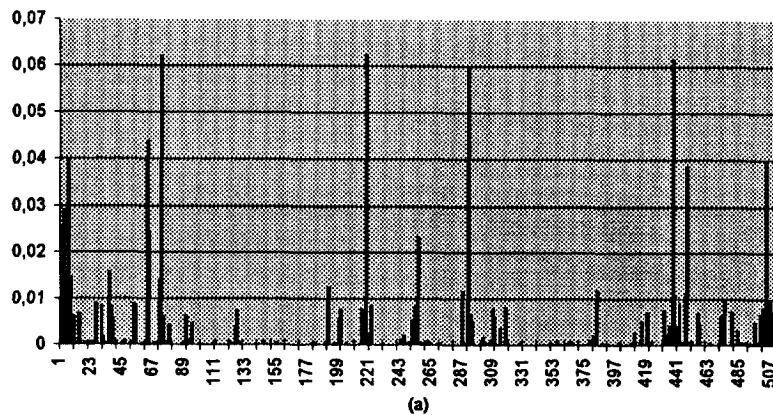


Fig. 7. Probability density functions $S(L)$ of typical regions.

$$S(L) = \frac{G(L)}{\sum_{L=1}^{510} G(L)} \quad (4)$$

Figure 7 shows the $S(L)$ functions of the four typical regions, i.e. regions corresponding to the first class of text, the second class of text, drawings and halftones.

The 510 values of $S(L)$ can be taken as texture features. However, as is explained below, by the application of a feature-reduction technique, only 13 of them are finally selected as texture features.

In the classification stage, these 13 features are used in combination with a minimum distance classifier having only four classes: text of normal characters, italic text, drawings and halftones. Specifically, if $i=1,2,3,4$ and $j=1,2,\dots,13$ are indexes for the number of classes and the number of features, respectively, the four distances D_i are estimated by the relation

$$D_i = \sqrt{\sum_{j=1}^{13} \left(\frac{f_j - \mu_{ij}}{\sigma_{ij}} \right)^2}$$

where f_1, f_2, \dots, f_{13} are the features of region A , $\mu_{i1}, \mu_{i2}, \dots, \mu_{i13}$

are the coordinates of the center of class i and $\sigma_{i1}, \sigma_{i2}, \dots, \sigma_{i13}$ the variances of class i .














Let $D = \text{minimum}\{D_1, D_2, D_3, D_4\}$. If $D < 1$ and $D = D_1$ or $D = D_2$, then A is classified as a text area; otherwise A is a non-text area. Table 1 gives the thirteen features, the DSECN L , the coordinates of each class center in the 13-dimensional feature space, the shape of each DSE and the classes containing the maximum and minimum number of DSE.

2.6. Feature reduction and clustering

An important process in a recognition system is to select a small set of appropriate features from a much bigger set. Selection of "good" features is critical to the performance of classification. To achieve this, a feature-reduction procedure was applied that permits the selection of only 13 of the 512 features. To reduce the number of 512 features, selecting a small set of "good" features which allow a region to be correctly classified as text, drawings, or halftones, a large number of such regions was examined, and an evaluation process was performed which examined the stability, separability and similarity of the features.

For the stability test, an attempt was made to identify the features which give significantly stable performance. The

Table 1. Clustering results

Feature	DSECN	Centers of classes				DSE	MAX	MIN
		1st cl. of text	2nd cl. of text	Halftones	Drawings			
01	219	0.062	0.0393	0.0011	0.0011		Normal chars.	Halftones
02	73	0.062	0.0385	0.0015	0.071		Normal chars.	Halftones
03	438	0.0614	0.0354	0.0025	0.0011		Normal chars.	Halftones
04	292	0.06	0.0345	0.0027	0.0072		Normal chars.	Halftones
05	1	0.0251	0.0473	0.013	0.0345		Italics	Halftones
06	256	0.0232	0.0454	0.0128	0.0344		Italics	Halftones
07	170	0.0	0.0	0.0467	0.0		Halftones	
08	341	0.0	0.0	0.0425	0.0		Halftones	
09	186	0.0	0.0	0.0264	0.0016		Halftones	Characters
10	495	0.0	0.0	0.0243	0.0		Halftones	
11	448	0.038	0.029	0.0025	0.0917		Drawings	Halftones
12	7	0.04	0.023	0.0047	0.0917		Drawings	Halftones
13	56	0.0	0.0	0.0001	0.0737		Drawings	Characters

A General Purpose Signal Processing Package

NEW Extended Version of Sig

New Menu interface for occasional users, On-line HELP, and Command mode for experienced users.

Signal Processing Operations Include:

- Modulo Based Signal Processing
- Identification
- Simulation
- Linear Estimation (Kalman Filter)
- Nonlinear Estimation (Extended Kalman Filter)

Graphical Operations Include:

- Plotting
- Families of Curves
- Multiple Viewport Plots
- Many Graphics Devices

Signal Processing Operations Include: Windowing, convolution, Fourier transforms, interpolation, decimation, digital filtering, linear systems, simulation, correlation, ensemble operations, spectral estimation, coherence, parametric and adaptive processing, deconvolution, identification and more...

New Features Include: Surface (3D) plots, contour plots, multichannel signal and complex operations, 2D Fourier transforms, additional algorithms (adaptive lattice, MEM, Burg spectral estimation)

On-line HELP Package

Menu mode for most operations and command mode for graphics

Fig. 8. Final separation results for the document of Fig. 2.

mean values and standard deviations for the features of each class were computed. For normalization, these values were then scaled to the same range by dividing them by their mean values. The features were divided into three groups according to their normalized standard deviation σ^* . The first group is for $\sigma^* < 0.01$ and corresponds to the high stability features. The second group contains unstable features with $0.01 < \sigma^* < 0.09$, and the third group very unstable features with $\sigma^* > 0.9$. Only the highly stable features were accepted. At this stage 12 features were rejected.

For the separability test and for every feature ℓ , belonging to the two classes i and j , the separability factor S_{avg}^ℓ was determined from the relation:

$$S_{\text{avg}}^\ell = \frac{|\mu_i^\ell - \mu_j^\ell|}{\sqrt{(\sigma_i^\ell)^2 + (\sigma_j^\ell)^2}} \quad (6)$$

where μ_i^ℓ, μ_j^ℓ are the mean values and $\sigma_i^\ell, \sigma_j^\ell$ the standard deviation values for each class. A large separability means that the feature has a good ability to distinguish between the two classes. The separability factor has been computed for any feature and for all class pairs. Generally, for a separability factor 1, the probability of successful discrimination is 50%. From experience it was found that a proper threshold value for the separability factor is 3. In the system described here, the separability test procedure rejected 436 features.

For the feature similarity analysis, and for every two features ℓ and m , belonging to the same class p , the correlation factor

$$C_{\ell m}^p = \frac{\frac{1}{n_p} \sum_{i=1}^{n_p} (\ell_i - \mu_\ell)(m_i - \mu_m)}{\sigma_\ell \sigma_m} \quad (7)$$

was estimated, where n_p is the number of elements of class p , ℓ_i and m_i the feature values of element i , $\mu_\ell, \mu_m, \sigma_\ell$ and σ_m the means and standard deviations of the features in the class p . The correlation factor measures the similarity between two features, and gives values between -1 and $+1$. A value near -1 or $+1$ means that the two features are highly correlated or inversely correlated, respectively. A value near zero indicates that the features are highly uncorrelated. Features with $|C_\ell| > 0.9$ must be rejected, and it was found that 49 features could be rejected.

After the application of the above three feature-reduction tests, only 13 features remained having the ability to classify the regions into classes. For classification, the *Nearest Means Clustering* (NMC) algorithm (Coleman and Andrews, 1979) is used which can determine the possible classes and their centers. The NMC algorithm has been applied to a large number of samples, and it has been found that except for the initial three classes (normal type text, drawings and halftones) there is also a fourth class, corresponding to regions of italic characters and numbers.

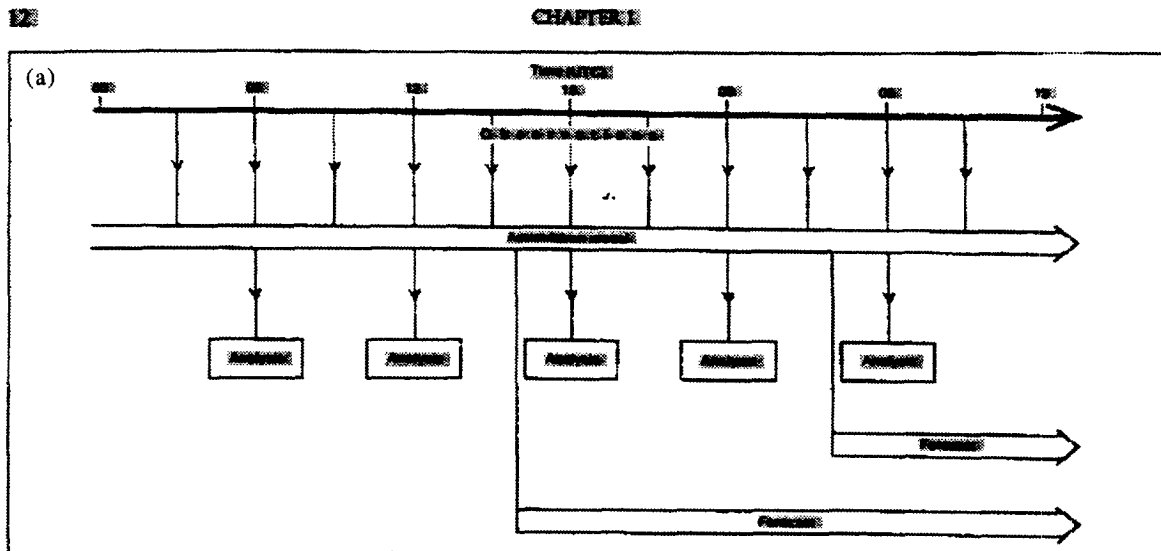


Figure 1
Continuous data assimilation

influence the model even if they are valid after the nominal analysis hour, as the model integration proceeds, later observations are left for assimilation and the system gradually assimilates further data. The assimilation of current operational system observations is essential to the best possible forecast results, providing greater accuracy for the earlier part of the forecast.

1.3.4 INITIALIZATION

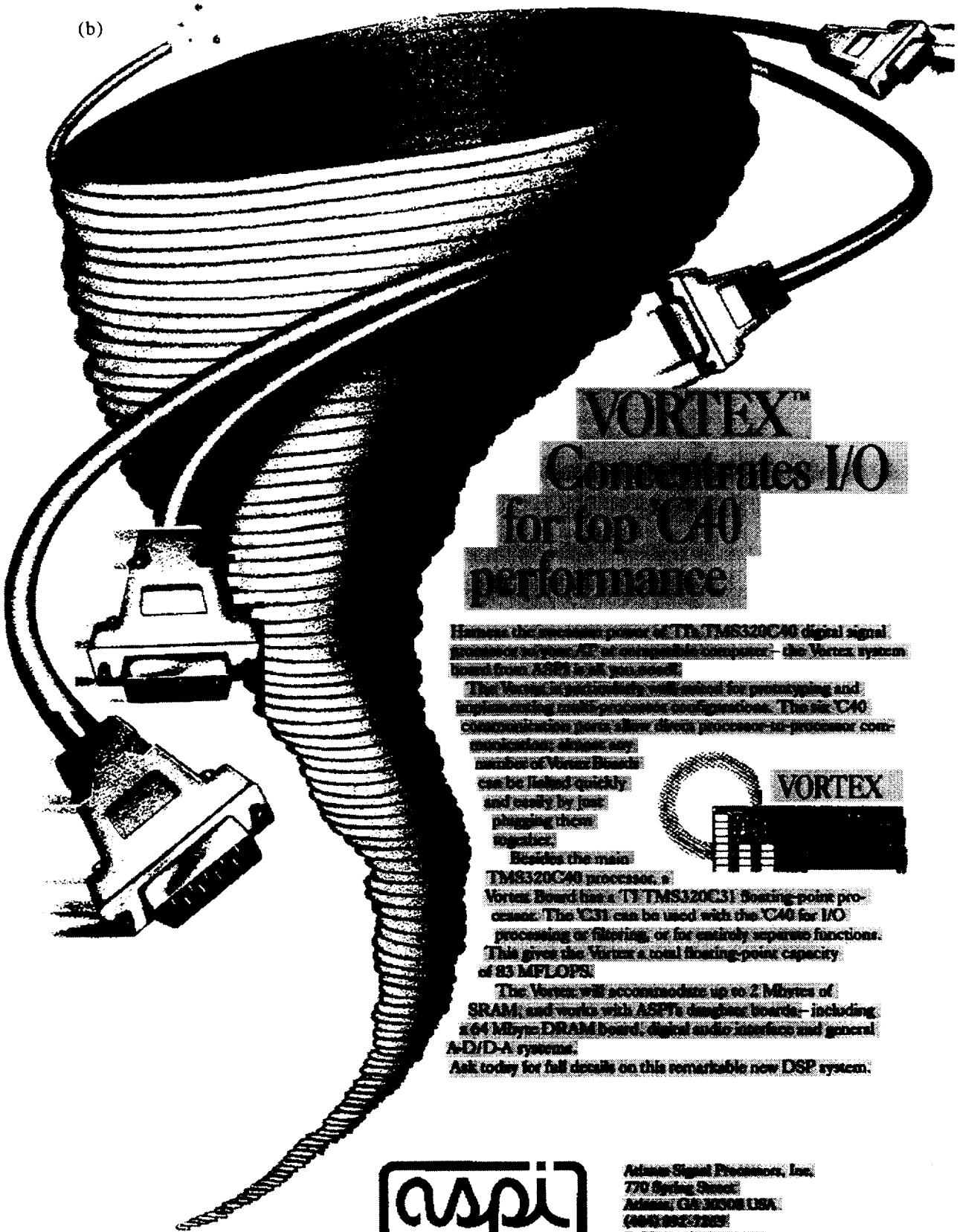
The primitive equations on which numerical models are based, generally admit high-frequency gravity wave solutions, as do the underlying Reynolds-averaged models. Both types of wave are found in the real atmosphere, but gravity waves being generally dispersive, propagate away from their source, and the atmosphere's slow response to their passage is produced by schemes based on implicit or explicit methods. Numerical models often have imbalances between the mass and wind fields which cause the forecasts to be contaminated by spurious high-frequency oscillations of much larger amplitude than those observed in the atmosphere. Although the damping terms, which are part of a numerical model, will tend to dampen these oscillations, they make the short-period forecast noisy and may be detrimental to the quality control and the assimilation cycle. For this reason an initialization step is performed after the analysis with the object of eliminating these spurious oscillations. Analysis produced by schemes based on spin-up iteration do not require a separate initialization step since balance is generally achieved during the assimilation process.

1.4 NWP PRODUCTS APPLIED TO AVIATION

World Area Forecast Centres are required to distribute grid-point data of wind and temperature at various levels, as well as information on the tropopause and the maximum wind. Suitably displayed in graphical or chart form, these data alone are of great value to aviation forecasters. However, a wide range of ancillary fields and derived data may be generated from an NWP system which provide a more complete picture of the numerical forecast. H-24 forecasts from the UK operational models serve to illustrate the range of products which can be generated from a numerical system. Most are used regularly at Bracknell in its role as a Regional Area Forecast Centre. Some are derived from fields based routinely on the GTS, but many require data at present only available at the centre. A selection of charts from the UK global and regional models is shown here, but all products could equally well have been derived from a high-resolution global model. In all cases, the model products have been projected onto a uniform grid (about 100 km grid spacing) for output purposes, which is considerably finer than the resolution of data available on the GTS in 1990. The output resolution, both in the horizontal and vertical, of course greatly affects the detail that can be identified.

Fig. 9. (a)-(c) Typical examples.

(b)



VORTEX™
Concentrates I/O
for top C40
performance

Harness the immense power of the TMS320C40 digital signal processor without ASPT or daughterboard support - the Vortex system built from ASPT is all you need!

The Vortex is particularly well suited for prototyping and implementing multi-processor configurations. The six C40 communication ports allow direct processor-to-processor communication simultaneously. **number of Vortex Boards can be linked quickly and easily by just plugging them together!**



Besides the main TMS320C40 processor, a Vortex Board has a TMS320C31 floating-point processor. The C31 can be used with the C40 for I/O processing or filtering, or for entirely separate functions. This gives the Vortex a total floating-point capacity of 83 MFLOPS!

The Vortex will accommodate up to 2 Mbytes of SRAM, and works with ASPT's daughter boards - including a 64 Mbyte DRAM board, digital audio interface and general A/D/D-A systems.

Ask today for full details on this remarkable new DSP system.



Adaptive Signal Processors, Inc.
 770 Spring Street
 Ann Arbor, MI 48106 USA
 (313) 962-3200
 FAX: (313) 962-2517

WORLD LEADERS IN DSP DESIGN TOOLS!

Fig. 9. Continued.

Three great capabilities to communicate

also communicate

If you need a design for a custom chip, we can help you. We have a team of experienced designers who can help you design a custom chip for your application. We can help you design a custom chip for your application. We can help you design a custom chip for your application.

With our new FeatureChips, you can now communicate with your customers. Our FeatureChips are designed to help you communicate with your customers. Our FeatureChips are designed to help you communicate with your customers.

For much less cost and hassle, you can offer your customers high-quality, low-cost, and high-quality data capabilities.

Free *FeatureChips* Literature Hotline: 1-800-858-0487. Ask for dept. 134

3-5300-6000, Singapore (65) 3632122, Taiwan (886) 2-719-4633, 3162-2000, United Kingdom (44) 0727-972424, ©1993 Cirrus Logic, Warren Avenue, Fremont, CA 94538, (510) 623-6900. MNP is a trademark of Microcom Systems Inc. The Cirrus Logic logo is a registered trademark of Cirrus Logic, Inc.




Fig. 9. Continued.

This class was named the second class of text. It is important to note that the objective here is to identify regions that can be successfully classified as text. Consequently, regions belonging to this new class are also labeled as text. Table 1 summarizes the results of the application of the feature-reduction and clustering procedure. As can be observed from this table, these results are reasonable and expected. For example, DSEs with characteristic numbers 170, 341, 186, 495 are the main components of halftone pictures.

Figure 8 shows the final separation results for the document of Fig. 1. It can be observed from the bottom part of this figure that good separation results are achieved, even for worse text regions.

3. EXPERIMENTAL RESULTS

The proposed method was extensively tested on a series of mixed-type documents. In addition to the authors' own test documents, the method has also been applied on some documents obtained from the University of Washington database (UW, 1993). The results appear to be very promising. To demonstrate the effectiveness of the proposed segmentation method some additional typical examples are presented.

Figure 9 presents the results obtained by applying the proposed method to six mixed-type documents. The document of Fig. 9(a) is of size 1688×2200 pixels and 300 dpi resolution, and is taken from the University of Washington database. Figure 9(b) and (c) are selected mixed-type

documents with different types of segmentation difficulties. These documents were scanned by a HP ScanJet IIC scanner at a 150 dpi resolution. All the documents in Fig. 9 contain text with fonts of different types, styles and sizes, and graphics that cannot be separated using vertical and horizontal lines. As can be observed, accurate text identification results were obtained.

As a final example, the proposed method was applied on two documents which have been artificially skewed by 5° and 10° . Figures 10(a) and (b) show the text identification results, respectively. It is noticed that the suggested method works well, even in the case of Fig. 10(b), which is a complex document that contains text and straight lines crossing each other.

4. CONCLUSIONS

Segmentation is an important issue in the automated document analysis research area. Text segmentation plays a significant role in document retrieval and storage systems. This paper proposes a new segmentation method that clusters the regions of a mixed-type document into text or non-text areas. The method belongs to the bottom-up class, and consists of the following main stages: extraction of marks, filtering of marks according to their heights, extension of mark shapes and creation of chains of connected marks, surrounding of chains by bounding rectangles, and finally, clustering the regions of the bounding rectangles as text or non-text areas. For the final stage, a technique is applied which is based on document

structure elements, and which formulates textural features. Specifically, 510 textural features were initially considered. Then, by using a feature-reduction process, only 13 of them were kept. Each of the 510 features corresponds to the

binary contents of a 3×3 mask, and its value is equal to the frequency distribution of the mask in the examination area. The application, in the training stage, of a clustering technique results in four classes: first class of text, second

(a)

VORTEX™
Concentrates I/O
for top C40
Performance

The Vortex™ DSP system is the only DSP system that can be used for both real-time and non-real-time processing. The six C40 processors provide the power to process over 100,000 samples per second. The Vortex™ DSP system can be used for both real-time and non-real-time processing. The six C40 processors provide the power to process over 100,000 samples per second.

VORTEX™

As the main processor, the Vortex™ DSP system is a Vortex Board™ (a TETRA2000) floating-point processor. The C40 can be used with the C40 for I/O processing or filtering or for creating separate functions. This gives the Vortex™ a total floating-point capacity of 100,000 OPS.

The Vortex™ will process up to 2 Million of SRAM and works with ADSP™ daughter boards—including 864 Mbyte DRAM boards, digital audio interfaces and general ADSP™ systems.

Ask today for full details on this remarkable new DSP system.



ASPI
World Leaders in DSP Applications

Adaptive Signal Processing, Inc.
770 Spring House
Andover, MA 01810 USA
(603) 532-7300
FAX (603) 532-2312

Fig. 10(a) Results for a document skewed by 5°.



Fig. 10(b) Results for a document skewed by 10%.

class of text, class of halftones and class of drawings.

In comparison with other techniques, this method has two advantages. First, for the mark-filtering stage a fast algorithm was developed and used for collinearity and vicinity checking. Second, the final stage of the method is based on the use of a texture feature analysis algorithm which can easily classify the areas of the bounding rectangles as text or non-text regions.

The proposed segmentation method was extensively tested on many mixed-type documents having significant difficulties for segmentation. It is independent of the size and type of the characters and the position of the text in the document, and it is insensitive to small tilts. It works well even in the cases where graphics and text areas cannot be

separated by vertical and orthogonal lines. Also, the method can separate out any text included in graphic regions or in mathematical equations. The results were very promising, and under normal scanning conditions no wrong classifications of a region as a text area were found. The average time for processing of a page is about 5 sec, on a Pentium-100 computer.

REFERENCES

- Chauvet, P. (1993) System for an intelligent office document analysis, recognition and description. *Signal Processing*, 32, 161-190.
- Coleman, G. B. and Andrews, H. C. (1979) Image segmentation by clustering. *Proceedings of the IEEE*, 67(5), 773-785.
- Farrokhinia, F. (1990) Multi-channel filtering techniques for texture

- segmentation and surface quality inspection. Ph.D. thesis, Dept of Electrical Engineering, Michigan State University.
- Fletcher, L. A. and Kasturi, R. (1988) A robust algorithm for text string separation from mixed text/graphics images. *IEEE Trans. Pattern Anal. Machine Intell.*, **10**, 910–918.
- Fujisawa, H., Nakano, Y. and Kurino, K. (1992) Segmentation methods for character recognition: from segmentation to document structure analysis. *Proceedings of IEEE*, pp. 1079–1092.
- Jain, A. and Bhattacharjee, S. (1992) Text segmentation using Gabor filters for automatic document processing. *Machine Vision and Applications*, **5**, 169–184.
- Kasturi, R. and Trivedi, M. (1990) *Image Analysis Applications*. Marcel Dekker, New York.
- Kasturi, R., Bow, S. T., El-Masri, W., Shah, J., Gattiker, J. R. and Mokate, U. B. (1990) A system for interpretation of line drawings. *IEEE Trans. Pattern Anal. Machine Intell.*, **12**(10), 978–991.
- O’Gorman, L. (1993) The document spectrum for page layout analysis. *IEEE Trans. Pattern Anal. Machine Intell.*, **15**(11), 1162–1173.
- O’Gorman, L. and Kasturi, R. (1995) *Document Image Analysis*. IEEE Computer Society Press.
- Papamarkos, N. and Gatos, B. (1994) A new approach for multilevel threshold selection. *CVGIP: Graphical Models and Image Processing*, **56**(5), 357–370.
- Pavlidis, T. and Zhou, J. (1992) Page segmentation and classification. *Graphical Models and Image Processing*, **54**(6), 484–496.
- Pratt, W. K. (1991) *Digital Image Processing*, 2nd edn. Wiley, New York.
- Schurmann, J., Bartneck, N., Bayer, T., Franke, J., Mandler, E. and Oberlander, M. (1992) Document analysis: from pixels to contents. *Proceedings of IEEE*, pp. 1101–1119.
- Strouthopoulos, C., Papamarkos, N. and Chamzas, C. (1995) Identification of text-only areas in mixed type documents. *IEEE Workshop on Non-linear Signal and Image Processing*, **1**, 162–165.
- Tsai, D. M. and Chen, Y. H. (1992) A fast histogram-clustering approach for multilevel thresholding. *Pattern Recognition Letters*, **13**, 245–252.
- UW (1993) English document image database. University of Washington, Seattle.
- Wahl, F. M., Wong, K. Y. and Casey, R. G. (1989) Block segmentation and text extraction in mixed text/image documents. *Computer Graphics and Image Processing*, **20**, 375–390.
- Wang, D. and Shihari, S. N. (1989) Classification of newspaper image blocks using texture analysis. *Computer Vision Graphics and Image Processing*, **47**, 327–352.
- Witten, I., Moffat, A. and Bell, T. (1994) *Managing Gigabytes*. Van Nostrand Reinhold, Amsterdam.
- Wong, K. Y., Casey, R. G. and Wahl, M. (1982) Document analysis system. *IBM Journal of Research Development*, **6**(6), 647–656.

AUTHORS’ BIOGRAPHIES

Charalampos Strouthopoulos was born in Drama, Greece, in 1962. He received his Diploma Degree in Electrical Engineering from the University of Patras, in 1985. Since 1993 he has been a Ph.D. student in the Electrical and Computer Engineering Department of Democritus University of Thrace, in the field of image segmentation and recognition. His current interests include document analysis and image processing. C. Strouthopoulos is a member of the Greek Technical Chamber.

Nikos Papamarkos was born in Alexandroupoli, Greece, in 1956. He received his Diploma Degree in Electrical and Mechanical Engineering from the University of Thessaloniki, Greece, in 1979 and a Ph.D. Degree in Electrical Engineering in 1986, from the Democritus University of Thrace, Greece. From 1987 to 1990 Dr Papamarkos was a Lecturer, and from 1990 to 1996 Assistant Professor at the Democritus University of Thrace where he is currently an Associate Professor (since 1996). His current research interests are in digital signal processing, filter design, image processing, pattern recognition and computer vision. Dr Nikos Papamarkos is a member of the IEEE and a member of the Greek Technical Chamber.

Christodoulos Chamzas was born in Komotini, Greece. He received a Diploma Degree in Electrical and Mechanical Engineering from the National Technical University, Athens, Greece, in 1974 and M.S. and Ph.D. degrees in Electrical Engineering in 1975 and 1979 from the Polytechnic Institute of New York. He was an Assistant Professor at the Polytechnic Institute of New York (1979–1982), and a member of the Visual Communications Research Department, AT&T Bell Laboratories (1982–1990) and is now an Associate Professor at the Democritus University of Thrace (1992–). His primary interests are in signal processing, image coding, multimedia and communications systems. Dr Chamzas is a member of the Technical Chamber of Greece and Sigma Xi, an Editor of the *IEEE Transactions of Communications* and a Senior Member of the IEEE.