International Journal of Computer Applications (0975 - 8887)
Volume 181 - No.48, April 2019

Extracting Text from Telugu Color Documents by Removing Dither Patterns

Siva Rama Sastry Gumma
Assistant Professor
Department of CSE
IIIT-RK Valley, RGUKT - A.P

ABSTRACT

Preprocessing is an important step in the development of Optical Character Recognition (OCR) system. Inpreprocessing there are various modules like binarization, skew detection and correction etc. Among these modules this paper discusses about binarization module. Although there are many algorithms for binarization of a document image, there are fewer algorithms for binarization of printed color images because of printed color documents contain dither patterns, normal text, reversed text, colored text overlayed on colored background drawings and graphics appear with millions of different colors. Hence preprocessing for colored documents is a challenging task to work.For printed color documents, elimination of dither patterns using Butterworth band reject filter and text extraction in the color documents by eliminating graphics using height of the component is also presented. Results on a corpus consisting of newspapers published in Telugu show that the proposed method shows promising results.

General Terms

Extraction, Elimination, graphics, grayscale

Keywords

OCR, threshold, binarization, dither patterns, connected components

1. INTRODUCTION

Printed Color documents contain text, graphics and drawings which are colored using many colors. In many cases some portions of this text are highlighted with headings and may be overlayed on different colored background. There are also regions with larger font size when compared with normal text. A typical Newspaper image as shown in Fig. 1, and it contains headings, subheadings, colored text, graphics, normal Text and reversed Text. The main motive is to binarize the color document which removes graphics and extracts all text (normal text, colored text and reversed text). In the coming sections problems related to color document binarization are been discussed and solutions to those are proposed.

Fig. 1. Typical Color Newspaper image

2. LITERATURE SURVEY

There have been many adaptive algorithms for the binarization of a document image. All of them are basically thresholding algorithms that depend on the choice of thresholds, where the pixel is assigned the value black or white based on threshold measured compared with its original value. Although this process seems to be easy, the real difficulty is to identify the threshold. Numerous Algorithms have been discovered to identify this threshold. Out of which, Otsu [1], Niblack [2] and Sauvola [3] algorithms are some of the common ones in gray scale image binarization.

There are no standard algorithms for binarization of color documents. This is because of printed color documents are much more complex with patterns, text, drawings and graphics appearing with many of different colors. A further complication is the use of half-toning process in printing that leads to regular dot-patterns, called *dither patterns* in shaded areas. In order to identify the dither patterns frequency domain and not spatial domain is considered more suitable. As a part literature survey, some of the papers are identified which give solutions to these problems.

This method [4] uses a combination of adaptive color reduc-

tion (ACR) technique and a page layout analysis (PLA) procedure to extract text in given document image. ACR technique is used to obtain optimal number of colors (*principal* colours) in the document. Then, using the principal colors, the document image is split into the separable color planes. On each color plane, PLA procedure is applied to identify the text regions. A merging procedure is applied in the final stage to merge the text regions derived from each of the color planes and to produce the final document. The disadvantage of this approach is that selection of optimal numbers is difficult for a printed newspaper documents as they contain many colors. This method proposed in [5] introduces a new color reduction technique to decrease number of colors in the document image. This technique estimates the dominant colors in the document and re-assigns the colors of original image to reduced set. Each dominant color defines a color plane in which the connected components (CCs) are extracted. Next, in each color plane a CC filtering procedure is applied which is followed by a grouping procedure. At the end of this stage, blocks of CCs are constructed which are next redefined by obtaining the direction of connection (DOC) property for each CC. Using the DOC property, the blocks of CCs are classified as text or nontext. The identified text blocks are binarized properly using suitable binarization techniques, considering the rest of the pixels as background. The final result is a binary image which contains always black characters in white background independent of the original colors of each text block. The paper [7], discusses about region-based thresholding for color document images and other paper[8] extracting halftones from printed documents using texture analysis are also been studied as part of literature survey. From the above discussion on different methods present in the literature one could say that there is no standard technique to binarize a color document.

3. COLOR DOCUMENT BINARIZATION

3.1 Problems in Color document binarization

The main problems in processing color documents are classified into three categories-

—Colored text on colored background - Conversion of colored text with colored background to gray scale will lead to many problems. This is because of foreground and background colors may sometimes have the same gray values. As shown in Fig. 2, the foreground color is blue and background color is brown, on conversion to grayscale both colors will have same gray values. Hence the colored text information is lost. This is because of graylevel values calculated as $0.299 \times$ Red + $0.587 \times$ Green + $0.114 \times$ Blue using the standard sRGB to grayscale conversion formula. Hence the conversion will lead to loss of text.

—Dither Patterns - Periodic patterns are observed on the color document image because of printing technology used as shown in Fig. 3. Usually commercial printing technology uses CMYK screens, which are aligned in some orientations with minimum overlap error. When it is observed from normal viewing distances, the eye integrates the colored dots, producing the illusion of continuous shades of color. But on viewing them closely one can identify these periodic dither patterns. These dither patterns produces maximum noise on conversion from color to grayscale. In the coming sections the formation and elimination of these patterns are discussed in detail.

—Graphics vs Text identification - Inorder to provide a valid input for text recognition to OCR, complete text extraction is to be done by eliminating graphics.

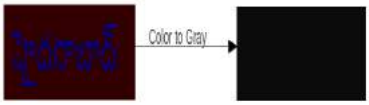


Fig. 2. Color to Gray Conversion

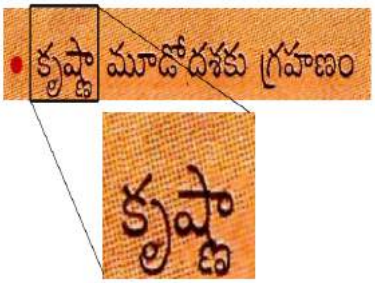


Fig. 3. Dither patterns

3.2 Problem Solving Approach

3.2.1 Solution for removal of Dither Patterns using FFTs. Commercial printers produces documents by depositing ink on paper. Hence shades of gray or color cannot be obtained them directly. To produce continuous-tone, images are typically printed as a halftone. A Halftone consists of an array of closely spaced micro-dots with varying size, all produced with same color of ink. At normal viewing distances human eye perceives colour or grayscale by integrating all the dots and the surrounding white spaces, producing a continuous shade of color. Different shades of gray or color can be simulated by varying the number of dots which are all of equal size. There were two approaches for this halftoning process: Clustered and Dispersed. Clustered-dot ordered dither is produced by grouping pixels to clusters and Dispersed-dot ordered dither is done by making position of the micro-dot is scattered.Aliasing and other visual artifacts can be reduced by employing various optimizing techniques.

Because of its simplicity, commercial printers typically use a version of clustered-dot ordered dither halftoning usually called the ‘classical screen method [6]. A glass plate etched with a grid of fine lines screen is placed between camera lens and image, to reproduce an image. The screen transforms the continuous tones in the input image into macro dots on the output. The size of the macro dot is proportional to the input gray value. Although these macro dots are bigger than micro dots, but these are visually imperceptible. Colored version of halftoning is a variation of grayscale one. Colored version uses CMYK (Cyan,Magenta,Yellow,K-Black) colors. Initially the input image is decomposed into CMYK Components. On each of component, image is converted to halftone using classical screen method as shown in Fig.4 .

These four halftones been overlaid to form a single composite halftone. The overlays are rotated relative to each other to minimize dot overlap, which reduces inter-color Moire patterns and chromatic errors. Usually Fig.5 (a) is been widely used, where Yellow screen is aligned at 0°, Cyan with 15°, Black with 45° and

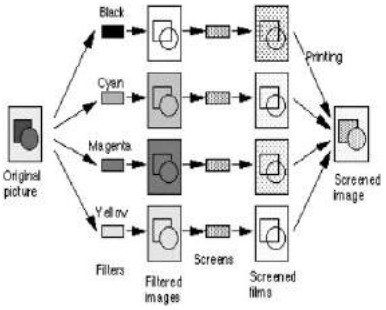


Fig. 4. Color Halftoning Process

Magenta with 75°. Color halftone as shown in Fig.3 is formed because these inter-color Moire patterns, but invisible to perceive. But when zoomed on that portion one could identify these periodic patterns.

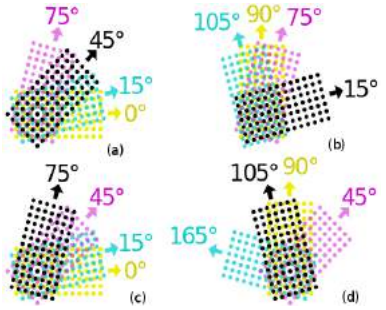


Fig. 5. CMYK Screen Orientation Angles

Clustered-dot ordered dither halftones have an interesting property. Due to optical illusion, one cannot identify this invisible texture with naked eye. In gray scale images, Dunn.et.al [8] shown that this invisible texture produces high-frequency spectral energy that is distinct from the visible information in the halftone and from other information on the page. And the halftones dervied from different images (but produced by the same screen) are effectively instances of the same texture, even though the halftones may visually appear quite different. Fig: 6 shows that grayscale image and log-magnitude specturm with high frequencies circled in black.

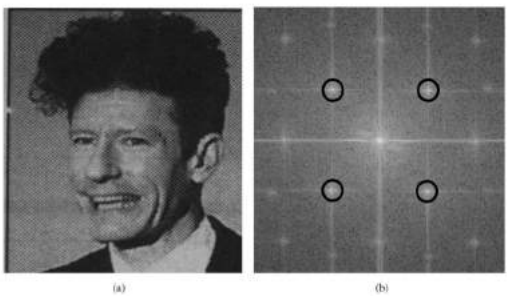


Fig. 6. (a) Halftone of a grayscale image (b) Log-magnitude spectrum with spectral energy (circled in black) at high frequencies

When coming to Color Images, which are printed using CMYK Screens in Newspapers , and are scanned the image in RGB mode with scanner. Hence there is conversion between CMYK to RGB mode. Conversion between the two-color schemes is as follows: Cyan absorbs Red and stimulates both Blue and Green (Cyan = White - Red), Magenta absorbs Green and stimulates both Blue and Red (Magenta = White - Green), Yellow absorbs Blue and stimulates both Red and Green (Yellow = White - Blue) and Black absorbs all the colors and therefore stimulates none. Because of these things when CMYK to RGB conversion takes place in Scanner, each of Red, Green, Blue color bands are produced with two colors of Cyan, Magenta, Yellow. That is, Red component is present in both Yellow and Magenta etc. Hence each color band is influenced by two colors and those two colors which have same macro dot sizes but aligned in different orientation. So the spectra of each color band contains high frequency components of two colors which are circular and centered at DC (freq =0). The Original image and Red, Green, Blue spectra are shown in Fig: 7, where high-frequency spectral components of the color halftone lie on a circle centered at DC. This suggests that a circular band filter can be used to remove these high frequency components. Because the given image Fig: 7 (a), is a squared one, the high frequency components form a circle. But for a rectangular images, the high frequency components will form an elliptical shape.
From the above discussion a band reject filter can be used to filter dither patterns for the documents. There were several band reject filters available, out of which butterworth band reject filter is used because, it has no ripples in the pass band (lesser ringing effect). The filter function for Butterworth band reject $H(u,v)$ is given by

$$H(u,v) = \left(\frac{1}{1 + \left(\frac{DW}{D^2 - D_o^2} \right)^{2n}} \right) \tag{1}$$

where
$D(u,v)$ is the distance from the center of the frequency rectangle
D_o is the radial center of the band of interest
W is the width of the band of interest

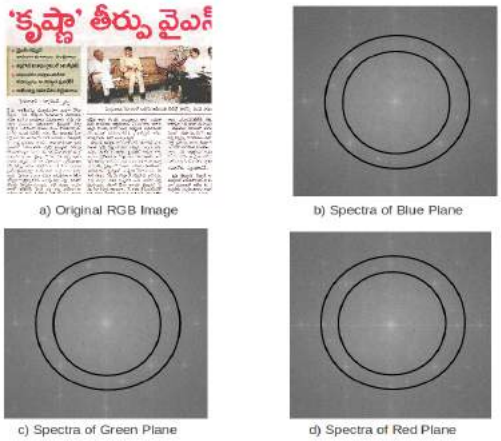


Fig. 7. Original Image with Spectra for Color Bands

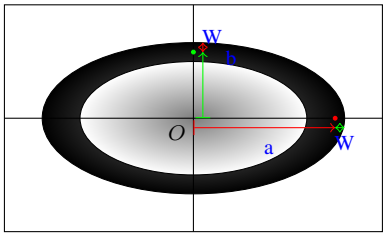


Fig. 8. Ellipse Estimation

When coming to two-dimensional geometry equations of circle and ellipse which look like

$$For \quad Circle \qquad (x-centre.x)^2+(y-centre.y)^2=r^2 \quad (2)$$

$$For \quad Ellipse \qquad \frac{(x-centre.x)^2}{a^2}+\frac{(y-centre.y)^2}{b^2}=1 \quad (3)$$

where r is the radius of the circle, and a, b are the length of semi-major and semi-minor axis of the ellipse. As one can observe closely the equation of circle Equation 2, it is also a form of ellipse with length of semi-minor axis and length of semi-major axis is equal to $r(a = b = r)$. The parameter $D(u, v)$ in the Equ 1 given by

$$D(u,v)=\sqrt[2]{\frac{(u-centre.x)^2}{a^2}+\frac{(v-centre.y)^2}{b^2}} \quad (4)$$

In the above Equ 4 $D(u, v)$ is normalized distance from centre (x, y) to the point (u, v). Now the values of W, and D_o are to be determined. W is width of the band, which could be value 10 or 20 depending on the required band width and the radial centre of interest D_o is given by $W/2$. As stated before, for a circle $(a = b = r)$, the Equ 4 holds good for all type of image dimensions. The unknowns mentioned a, b can be estimated manually by working on different images. The parameters a and b roughly depend on the dimensions of the image (a = 0.66* width, b = 0.66 * height). If the documents are same kind (produced by same halftone screens), then values remains good. But if the documents are with different kind, the estimations may change.
From the above discussion, band reject filter can remove the dither portions on the image. Dither patterns present on graphics part of the document will be removed, this can observed by Inverse-FFT and merging the R, G, B bands to form a dither removed RGB image.

3.2.2 Connected Component Algorithm for identifying Text vs Graphics. It is more important to remove graphics from a document before it gets to recognition. The simplest way to do this is to filter them using height of the component. As mentioned before, that on a document image text size is varied from the type of text (heading, subheading or normal text). Hence, to filter graphics from text, with height as a criteria can be done heuristically by finding out sizes of different text. Height of component is measured by finding out MBR (Minimum Bounding Rectangle) of the component and find $xmax - xmin$. As far as telugu text is considered there will be aksharas, vottus, matras are present and each letter can be combination of two or more of these. As shown in Fig 9 the Red color ones are vottus, Blue color are basic telugu characters and Green ones are basic character with matras.

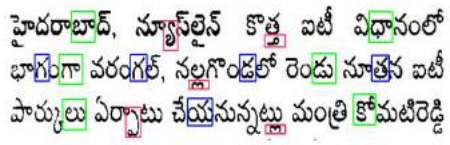


Fig. 9. Sample Telugu Text

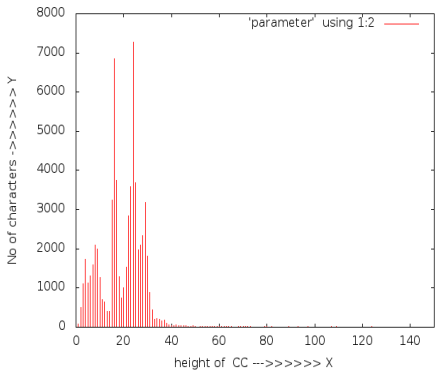


Fig. 10. Graph to Estimate Size of characters

As observed, the vottus are of variable size and so are the basic characters. Inorder to classify these, a corpus of about 50 pages of Newspaper documents is taken. The graph plotted between height of the component vs Number of Components as shown in Fig :10. One can notice that the largest peak at x = 24 is the actual size normal character, the vottus may be of size x = 14. From Fig :10 the heights are classified into the different classes as mentioned in the Table 1.
From the classification, MBR heights of components from 3 to 210 are considered and the sizes above 210 can be eliminated. After removal of graphics, only text will be retained in the document. This document can be used in the next stage of OCR to recognize them. Finally on concluding this section, removing dithered patterns and graphics will help us to extract complete text in color document. The complete algorithm of binarization of color document will be detail in next subsection.

3.3 Proposed Algorithm

From the above discussion of solving out issues in color documents, the solutions proposed are integrated in an ordered manner to binarize a color document. Conversion of an RGB into gray scale is done by *opencv* standard conversion. In case of newspapers which are with white background binarization can be done straight forwardly using Otsu [1] algorithm. The complete algorithm is as follows:

Table 1. Empirical Values for Heights of Components

Height	Description
≤ 3	noise
$4 \leq h \leq 30$	fullstop, commas, vottus, quotes, matras, normal characters
$31 \leq h \leq 50$	subheadings or words with vottus and matras
$51 \leq h \leq 78$	bold words
$79 \leq h \leq 110$	bold Text
$111 \leq h \leq 150$	bold heading with votuus and matras
$151 \leq h \leq 209$	big headings
$210 \leq h$	graphics

Algorithm 1 Algorithm for Preprocessing of Color document

Step 1: Dither patterns removal on input image
Step 2: Conversion of Dithered removed RGB image to gray scale
Step 3: Binarization of gray scale Image
Step 4: Graphics elimination

Step 1 : Dither Removal The First step is to remove dither pattern from the given input image .
Input : RGB image
Output : dither removed RGB image

Algorithm 2 Algorithm for dither removal

a: For the given Input RGB image, split into R, G, B planes (3 gray scale images)
b: On each plane obtain its FFT
c: Use Butterworths Band Reject Filter to filter high frequency components
d: Obtain inverse-FFT of that plane
e: Merge the filtered R,G and B planes to form dither removed color image

Step 2 : Conversion of dithered removed RGB image to gray scale The Second step is to convert the dither removed RGB image into GrayScale . This is done by using $opencv\ \ library$ standard, where the Grayvalue of the pixel is

$$\text{Gray Value} = 0.299 \times \text{Red} + 0.587 \times \text{Green} + 0.114 \times \text{Blue}$$

Input : dither removed RGB image
Output : gray scale image

Step 3 : Binarization of grayscale image The Third step is to conversion of gray scale image into Binary Image .This is done using standard Otsu [1] algorithm.
Input : gray scale image
Output : binary image

Step 4: Graphics Elimination
The Fourth step is to eliminate graphics present in the document. As mentioned in the Section *3.2.2*, connected component algorithm finding out MBR and height of the component, the filtering of graphics can be done
Input : binary image
Output : graphics removed binary text image

Algorithm 3 Algorithm for graphics removal

```
1: procedure GRAPHICSREMOVAL(binary_image)
2:     Apply Connected Component(CC) Algorithm on the Binary Image
3:     for Connected Component CC_i do
4:         height of CC_i = MBR[i].xmax - MBR[i].xmin;
5:         if height of CC_i is in range of [3, 110] then
6:             Place CC_i in the Output Image
7:         end if
8:     end for
9:     return Output Image
10: end procedure
```

4. IMPLEMENTATION RESULTS

On input Image is splitted into Blue, Green and Red planes which is shown in the Figure 11. Apply FFT on each plane of the input image and then apply Butterworth band reject filter on FFT and computing the inverse FFT. These results are shown in Figure 12. The step is to combine the dithered removed image into RGB image, you can see the results of this in Figure 15. After combining the images to form a RGB image, the color image is converted into GrayScale using OpenCV standard and binarizing the grayscale image using Otsu Algorithm. And then applying connected components algorithm on the Image. The results this process are shown in Figure 14. Based on the height of the components every component is classifed whether it is text or graphics, and the intermediate results are shown in Figure 15. The final resultant image after graphics removal is Figure 16.

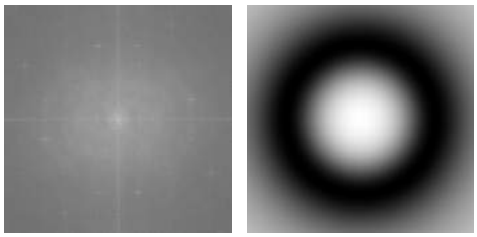
5. CONCLUSION

For colored images, the binarization is done using otsu algorithm after eliminating color halftones. The novel contribution of work is elimination of color halftones using Butterworth band reject filter, text extraction and graphics elimination. Text extraction based on a statistical analysis of the heights of the connected components. These results show that, Butterworth bandreject filter can eliminate color halftones in the documents. Since all newspapers does not use same printing technology, parameters (length of major and minor axis of ellipse) in dither removal depends on type of Newspaper.The work is performed on different telugu newspapers, each having different values say Sakshi-0.64, Eenadu-0.66, Andhrajyothy-0.71, where each value denotes percent of the size of the major and minor axis with image dimension. As far as text extraction is concerned, the results show that the heights of the text components lies in the range of 4 to 60 pixels for documents scanned at 300 dpi with text in the font sizes of 12-18 points. These

(a) Input Image (b) Blue Plane

(c) Green Plane (d) Red Plane

Fig. 11. Input RGB Image and Blue,Green and Red planes of it



(a) Fft of Blue Plane (b) Butterworth band Reject Filter

(c) Fft with filter (d) Inverse FFT

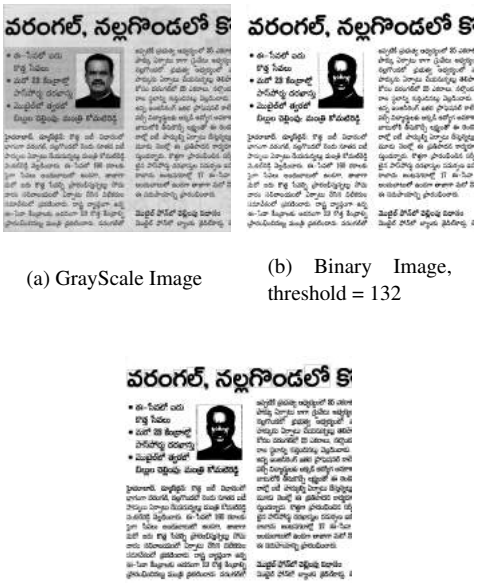
Fig. 12. FFT , band reject , FFT with filter and inverse FFT Images

results over 50 colored images show that the preprocessing improves over the existing preprocessing techniques.

(a) Inverse FFT of red (b) Dither removed image

(c) Dithered region on original image (d) After Dither Removal

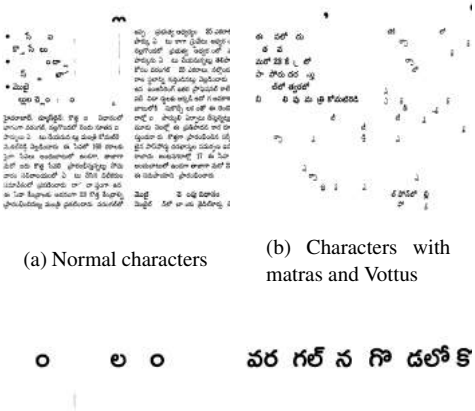
Fig. 13. Red plane Inverse FFT, dither removed image, portion of original image before and after dither removal



(a) GrayScale Image (b) Binary Image, threshold = 132

(c) Connected Components on binary image

Fig. 14. GrayScale Image, Binary Image and Connected Components of dither removed image



(a) Normal characters

(b) Characters with matras and Vottus

(c) Bold characters

(d) Bold characters with matras

Fig. 15. Different sizes of connected component images

(a) Graphics

(b) Final Text Image

Fig. 16. Graphics and Final text extracted Image

6. REFERENCES

[1] Nobuyuki Otsu A Threshold Selection Method from Gray-Level Histograms *IEEE Transactions on Systems and Cybernetics (1979)* , Vol. SMC-9 No. 1 , January 1979.

[2] W.Niblack An Introduction to Digital Image Processing. 1986.

[3] J.Sauvola,T,Seppanen, S.Haapakoski and M.Pietikainen Adaptive Document Binarization *ICDAR'97 4th Int. Conf. On Document Analysis and Recognition* pages.147- 152.

[4] C.Strouthopoulos, N. Papamarkos and A.E. Atsalakis Text Extraction in Complex Color Documents *Pattern Recognition (2002)* , pages 1743-1758, 2002.

[5] Efthimios Badekas , Nikos Nikolaou, Nikos Papamarkos Text Binarization in Color Documents *Wiley Periodicals, Inc*, Vol. 16, 262274 (2007)

[6] Dennis F. Dunn and Niloufer E. Mathew Extracting color halftones from printed documents using texture analysis *Pattern Recognition 33 (2000)* , pages 445-463, 2002.

[7] Chun-Ming Tsai Intelligent region-based thresholding for color document images with highlighted regions *Pattern Recognition 45 (2012)* pages 1341-1362

[8] D.F. Dunn, T.P. Weldon, W.E. Higgins Extracting halftones from printed documents using texture analysis *Opt. Eng 36 (4) (1997)* pages 1044-1052.