Segmentation of text/image documents using texture approaches

Trygve Randen and John Håkon Husøy
Høgskolen i Stavanger
P.O. Box 2557 Ullandhaug, 4004 Stavanger, Norway
E-mail: t.randen@ieee.org
URL: http://www.hsr.no/~tranden

Abstract

A fast and flexible method for segmenting text and graphics of document images is presented, assuming that text-regions have different texture properties than non-text-regions. As features we use measures of the local sub-band energies in the output of a computationally efficient critically sampled filter bank. First we develop an un-supervised approach, and we get good results, but even better segmentation is desired. We therefore extend our approach with a novel supervised segmentation scheme. The experimental results from the supervised method show the expected improvement over the results from the un-supervised method.

Keywords: Pattern recognition, texture segmentation, document segmentation.

1 Introduction

Traditionally, paper has been the main source for publication of newspapers, periodicals, reports, books etc., and an enormous amount of printed material is available on paper. The digital computer and computer networks have made it possible to search for and retrieve electronically stored documents in seconds, no matter where in the world they are stored. This is far from the reality for documents stored as paper copies. Therefore there is a considerable interest in digitizing paper documents.

However, it is not trivial to automatically digitize a document in a suitable manner. The straight-forward approach is to scan the whole document, and store it as a bit-map representation. Unfortunately this approach requires a huge amount of storage capacity (even with modern image compression techniques) and does not enable search for text in the documents. This approach has been applied for the proceedings of the IEEE 1994 International Conference on Acoustics, Speech & Signal Processing (ICASSP'94) and several other conferences. The proceedings are available on a CD-ROM, stored as TIFF bitmaps, and only the titles and the abstracts are available in a searchable form.

A more efficient way of representing documents is to separate the text and the graphics, and store the text as ASCII text and the images as bit-maps. Therefore, several approaches to text-graphics segmentation have been proposed. The most popular approaches may broadly be classified as either *top-down* or *bottom-up* approaches [1]. One of the most common top-down techniques is the *run length smoothing* [2, 3, 4] (also known as the *constrained run length* method). With the run length smoothing algorithm, the document image is processed along vertical and horizontal runs (scan-lines). If the number of white pixels between two black pixels is less than a threshold, then the white pixels are converted to black. Thereby, white gaps within and between words are converted to black, while white gaps between columns remain white, and similarly for lines. Thus, the run length smoothing method is sensitive to font-size, character spacing, line and column spacing and to some extent to the orientation of the document.

Another common top-down technique is the *recursive projection profiles* method [4, 5, 6] (also known as the *recursive X-Y cuts* method). The basic idea underlying the projection profiles method is that the numbers of black pixels in the columns and rows of the image are projected onto the horizontal and vertical axes, respectively. Thus column or line spacings, respectively, generates detectable valleys in the projection profiles, and it is possible to extract the column structure of the document. The recursive projection profiles method is therefore restricted to rectangular blocks, and consequently not suited for skewed text.

The bottom-up methods are typically variants of the *connected components* method [7]. The connected components method basically extracts connected components, like a character or a connected part of a drawing, performs some kind of analysis on each connected component, and then merges connected components that are close to each other. The drawbacks with the connected components method are that it is character size dependent, sensitive to inter-line and inter-character spacings and sensitive to the digitizing resolution.

In addition to the already mentioned restrictions with these more or less ad hoc methods, they all operate on thresholded images. If the image capturing conditions are difficult, the determination of an appropriate threshold may be difficult.

One interesting ability of humans, is the ability to tell what is text and what is not text in a document viewed from a distance too far away for letter discrimination. This suggests that textural clues are important for humans in separating text regions from non-text regions. In this paper we present an approach to segmentation of text and graphics in scanned documents, based on the assumption that the text in a document may be viewed as one texture, while the graphics is a different one. Using this assumption, we segment the documents with a texture segmentation scheme using a filter bank and a local energy function in the sub-bands as the feature extractor. We propose both an un-supervised and a supervised segmentation scheme. While most traditional text-graphics segmentation schemes assumes some a priori knowledge of the input, our un-supervised scheme is independent of document layout, typeface, font size, scanning resolution etc. However, the un-supervised segmentation scheme shows a somewhat limited performance for difficult documents. By sacrificing some of the flexibility with the un-supervised classifier, we show that the segmentation is improved by using a supervised classifier.

An approach similar to our un-supervised approach [8], uses the Gabor filter and a local energy function as the feature extractor. The Gabor filter is the most frequently used filter for multichannel texture segmentation, and the rationale presented for the use of the Gabor filter is its relation to models for the early vision of mammals as well as its joint optimum resolution in time and frequency [9, 10]. One major drawback with the Gabor approach is its computational complexity. The approach presented in [8] uses a subset of eight filters, from the total set of 20 filters suggested in [11], but even when this restricted subset is used, a total data amount of eight times the input has to be used as features for the classifier.

It is well known that sub-sampling of the various frequency bands, in such a way that the total number of samples in the sub-bands equals the number of signal samples in the input image, can be performed without losing any information, as long as the filters are designed appropriately [12]. This fact has been used for a while in image sub-band coding, see for example Husøy [13]. We have previously shown that such a filter bank may be successfully used for a general texture segmentation approach [14], with approximately the same segmentation quality as with the Gabor filters, and with *considerable* computational savings. In this paper we show that good document segmentation results may be obtained using such a filter bank.

In Sec. 2 we propose an un-supervised segmentation scheme, and in Sec. 3 we propose a novel supervised segmentation scheme. In both cases we give some experimental results which illustrate the performances of the suggested approaches.

2 Un-supervised segmentation

We use a critically sampled filter bank with the filter 'F_2_1_09' and the filter structure 'C' along with the sigmoidal nonlinearity and Gaussian smoothing filter as feature extractor, as we suggest in [14]. The approach presented in this section was first published in [15].

Note that using a critically sampled filter bank implies the calculation of only a few feature samples, and the total size of the feature set is identical to the size of the input image. The approach in [8], in contrast, implies a feature set eight times the size of the input image. An A4 page scanned at 300 dots per inch, for example, consists of more than eight million pixels, thus the size of the feature set is important.

If our features give rise to correct un-supervised segmentation, we will not have to make any assumptions about font size, scanning resolution, column layout, orientation etc., as we will see in this section.

The clustering algorithm we have chosen to use is the k-means algorithm. The k-means algorithms, in the version we have used, tries to calculate the k most typical cluster centers, $\mu_1 \cdots \mu_k$, using basically the following algorithm:

1. Initial random selection of $\vec{\mu_1} \cdots \vec{\mu_k}$.
2. Assign each feature vector to its nearest $\vec{\mu_i}$, using an Euclidean distance measure.
3. Set the $\vec{\mu_i}$'s equal to the means of their assigned feature vectors.

4. Repeat steps 2 and 3 as long as the sum of the squared distances from the feature vectors to the nearest cluster centers decreases.

2.1 Experimental results

We have chosen to use parts of pages scanned from the march 2., 1994 issue of the local Stavanger newspaper "Stavanger Aftenblad" (SA) as our test images. For comparison purposes, we have also included some images used in [8, 16]. All the images used in this paper are available via anonymous FTP from `ftp.hsr.no` or on the web `http://www.hsr.no/tele-signal-group/`.

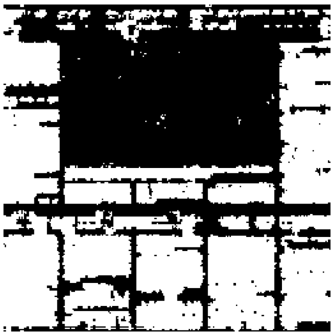
Fig. 1 shows a part of page 8 in SA, and the quite successful segmentation of the document. Except for some problems with the heading and spurious dots in the text regions, the segmentation is as desired. We point out that no a priori information on font size, scanning resolution etc. has been assumed. This example shows the robustness of our method.

To further emphasize the flexibility of our approach Fig. 2 shows the segmentation results from the same page, rotated by 55° and at half the resolution, respectively. This figure shows one important result: the segmentation quality is practically invariant to scaling and rotation.

How does our method compare with other methods? Figs. 3 and 4 are the same images as used in [8] and [16]. The segmentation results from these images are practically the same, verifying the conclusion in [14] that our new filtering approach, including subsampling, is appropriate.

Unfortunately, not all documents are as easy to segment as the ones presented thus far. Our text segmentation method is based on the assumption that the text regions may be considered as constituting a different texture than the picture regions. Therefore it is natural that the method performs poorer when the images contain textured regions similar to the texture in the text region. The picture region in the document in Fig. 5(a) has a texture quite similar to the texture of text, and we see, as expected, that the segmentation, Fig. 5(b), in this case has included a considerable part of the graphics in the text category.

a



b

c

Figure 1: The 1024 × 1024 pixel scan of parts of page 8 from SA (a), the segments (b), and (a) with regions classified as picture excluded (c).

2.2 Comments

It is important to be aware that all our experiments in this section were done with no assumptions on the input image and with no post-processing. Several kinds of post-processing techniques are possible, and should improve the results, especially if we know more about the format of the input. For example, if the input is known to be restricted to standard newspaper columns, a post-processing technique looking for rectangular blocks of the particular width in the segmented image may be employed.

3 Supervised segmentation

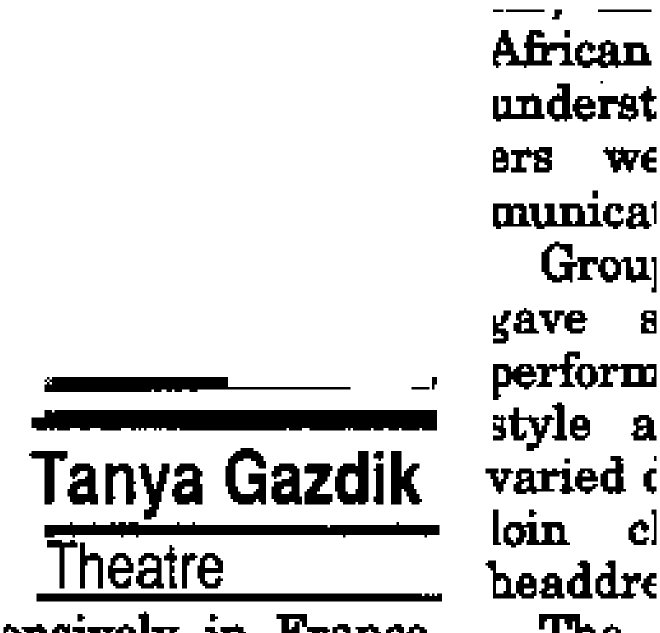
Motivated by the problem with the un-supervised segmentation approach in the previous section, we propose a somewhat different approach, with supervised

a

b

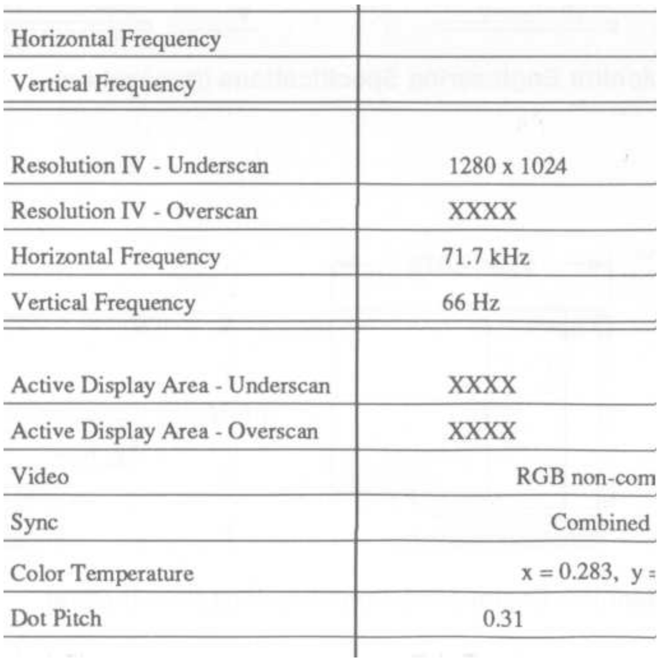
Figure 2: Segmentation of page 8 from SA rotated by 55° (a), and segmentation of the page scanned at half the resolution, 512 × 512 pixels (b). Both images are obtained by excluding regions classified as picture.

a



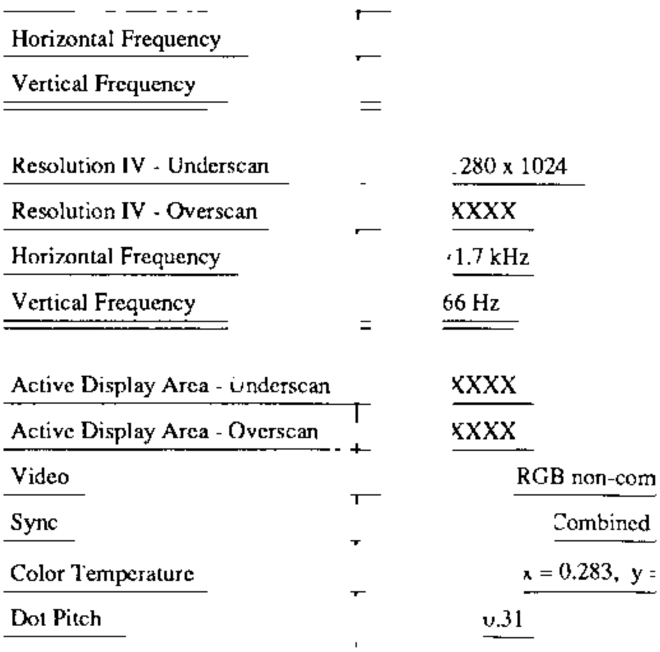
b

Figure 3: Test image used by Jain and Bhattacharjee (a), and segmentation of the image (b).



Horizontal Frequency	
Vertical Frequency	
Resolution IV - Underscan	1280 x 1024
Resolution IV - Overscan	XXXX
Horizontal Frequency	71.7 kHz
Vertical Frequency	66 Hz
Active Display Area - Underscan	XXXX
Active Display Area - Overscan	XXXX
Video	RGB non-com
Sync	Combined
Color Temperature	x = 0.283, y =
Dot Pitch	0.31

a



Horizontal Frequency	
Vertical Frequency	
Resolution IV - Underscan	.280 x 1024
Resolution IV - Overscan	XXXX
Horizontal Frequency	·1.7 kHz
Vertical Frequency	66 Hz
Active Display Area - Underscan	XXXX
Active Display Area - Overscan	XXXX
Video	RGB non-com
Sync	Combined
Color Temperature	x = 0.283, y =
Dot Pitch	0.31

b

Figure 4: Test image used by Jain and Bhattacharjee (a), and segmentation of the image (b).

a

b

Figure 5: Page 2 from SA (a), and the segmentation of the page (b).

segmentation. A supervised approach will of course mean a limitation when compared to the flexibility of the un-supervised approach, but in many applications it may be assumed that the font-size, scanning resolution, and page orientation are fixed. This will probably be the case if for instance a newspaper decides to digitize all issues from a given year.

In addition to the un-supervised approach, a supervised approach was also presented in [8]. The specific application where the supervised approach was applied, was segmentation of bar-code from text. Their supervised approach was based on training a classifier using exemplar features from bar-code and text. However, in a general text segmentation application, it is not possible to assume that the non-text class is well defined, like the bar-code. It will therefore not be possible to train a classifier on the non-text regions.

3.1 Training approach

Although it is not possible to assume that the non-text class has fixed properties, it is very well possible to assume that the text class has fixed and well defined properties. Many approaches to supervised text segmentation using texture techniques neglects this fact.

To classify we need to calculate some kind of measure of similarity with the prototype text class. We experimented with various techniques for determining prototype features and measuring distance, and settled on using the weighted Euclidean distance measure, given by

$$d = (\vec{x} - \vec{\mu})^T \, \Sigma^{-1} \, (\vec{x} - \vec{\mu})$$

where $\vec{\mu}$ is the estimated mean of the text feature vectors and Σ is the matrix with the estimated feature variances on the main diagonal. This is a new approach to supervised segmentation in a texture context.

Having all features scaled to approximately the range $0 \cdots 255$, we experimentally determined a distance of 50 from the mean to be an appropriate decision limit for text. It is important to keep in mind that techniques like Bayes decision rule [17] may not be employed to determine an optimal threshold for classification as text, because we don't know the non-text properties.

3.2 Experimental results

We have chosen to use two documents for our supervised experiments. First, we'll show the results with page 8 from "Stavanger Aftenblad" (SA), a page which were successfully segmented by the un-supervised approach. Then we'll show how the method performs for a more difficult page, page 2 of SA.

Page 8 of SA, Fig. 1 (a), contains four different kinds of text, normal text, boldface text in the article introduction, and headings of two sizes. We have tried to classify the document according to each of these text types. Remember that we are using a single codebook vector for each of the four kinds of text. As mentioned, a feature vector is classified as text if it is less than a distance of 50 from the prototype feature vector. If a feature vector satisfies this for several of the text classes, it is classified to the class with the closest prototype.

How good is the separation between text and non-text regions? We get an impression of this by looking at the distances to the text prototype from the text region versus the distances from the picture region. Fig. 6(a) shows the histogram of the distances from some of the normal text feature vectors, and Fig. 6(b) shows the histogram of the distances from the picture feature vectors. We see that the histograms are well separated. From the histograms we understand why a text distance threshold of 50, as marked in the figures, is appropriate. A higher threshold would classify more

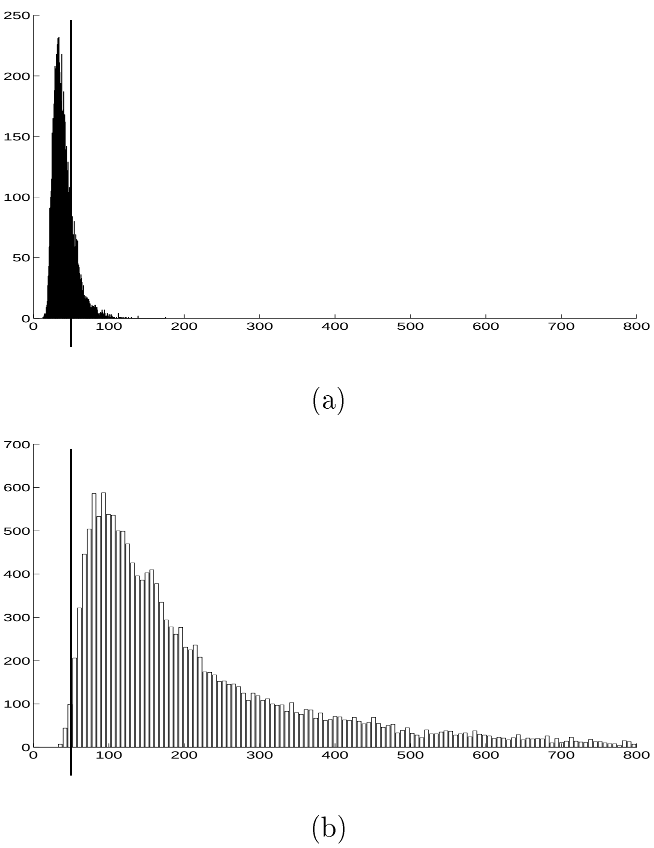


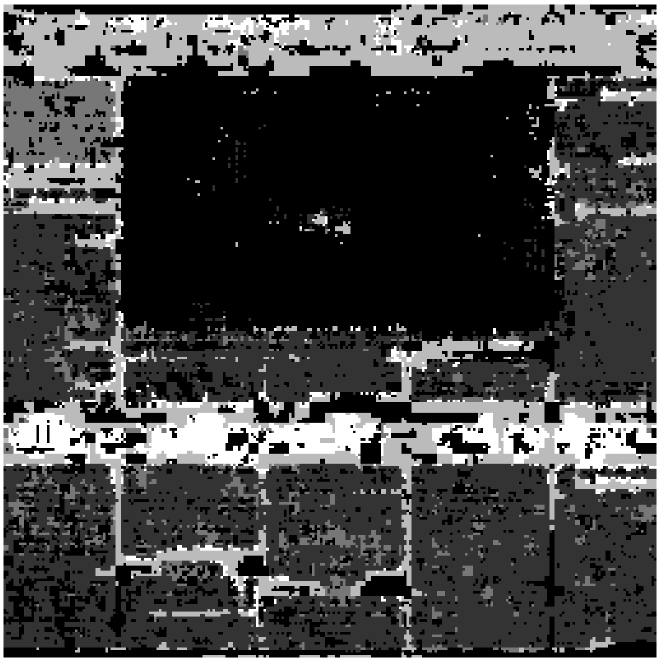
Figure 6: Histograms for text (a) and image (b) for page 8 of SA.

picture regions as text, and a lower would classify more text regions as non-text.

The segmentation results from page 8 are given in Fig. 7. As we see, the segmentation is mainly correct, but with some small regions misclassified. These misclassified regions should be possible to detect and remove with various post-processing techniques.

Promising results from segmentation with the supervised approach applied to page 8 of SA were expected, since we also got good results with the un-supervised approach to this page. The only improvement for page 8 is that the headings are now more reliably classified. The main motivation for developing a supervised approach was, however, that we wanted improvement of the results for the more difficult documents, like page 2 of SA. In Fig. 6 we saw the histograms of the distances from the text prototype for page 8 of SA. The corresponding histograms for page 2 are given in Fig. 8. As we could expect, we see that the histograms overlap more. However, a text classification distance threshold of 50 still seems reasonable.

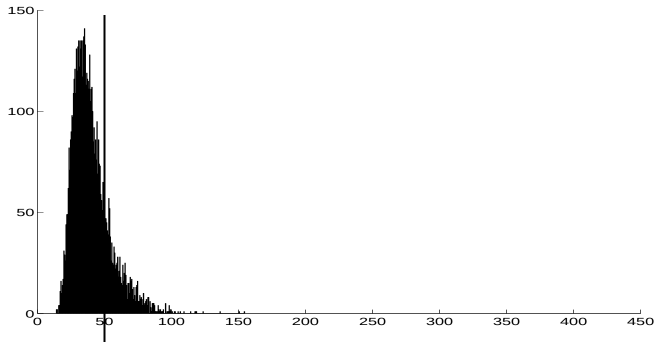
The segmentation results for page 2 are given in



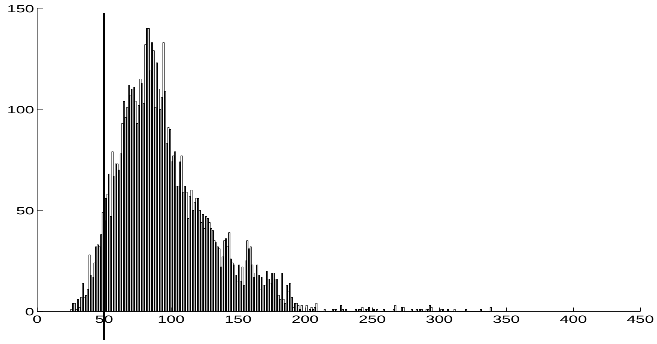
a

b

Figure 7: Segmentation of page 8 from SA using the supervised approach, (a) shown as regions (black is non-text regions, dark gray is normal text, medium gray is boldface text and light gray and white are large and small headings, respectively) and (b) the input image masked through the segmentation results (non-text regions masked out).



(a)



(b)

Figure 8: Histograms for text (a) and image (b) for page 2 of SA.

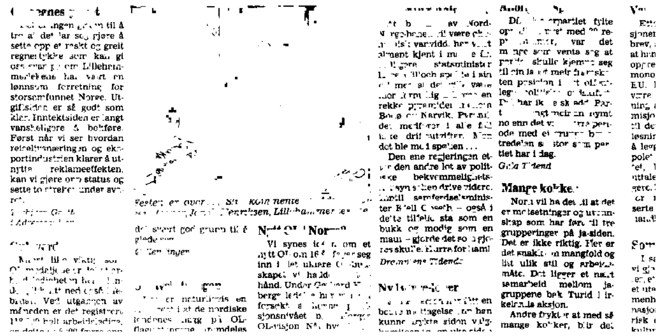
Fig. 9. We see that only small parts of the graphics region are classified as text, while large parts were classified as text in the un-supervised approach. Thus the results are much better with the supervised approach than with the un-supervised approach.

4 Summary and conclusions

Storage of documents in digital form is a useful tool in making the contents easily available, and if important documents are made available for search and retrieval by computers, it will be of great importance in the work of scientists, engineers, journalists etc. In digitizing existing paper documents, it is of great importance to be able to separate the text from the graphics, both because it is then possible to store the text more effectively than the graphics, and because it is practically impossible to search for text in a bit-map file.

In this paper we have presented two new approaches to segmentation of text and graphics in digitized document images, both based on texture segmentation with computationally effective critically sampled fil-

a



b

Figure 9: Segmentation of page 2 from SA using the supervised approach, shown as regions (a) and the input image masked through the segmentation results (non-text regions masked out) (b) .

ter banks. Our first approach used an un-supervised classifier. While most traditional text-graphics segmentation schemes assume some a priori knowledge of the input, this approach uses no a priori knowledge of the input, and still achieves satisfying results in many cases. The second approach was a supervised classification approach, yielding near perfect segmentation, even for difficult documents where the un-supervised approach had poor segmentation.

The approaches presented in this article may be applied either as an addition or a replacement to the more traditional approaches, depending on the application and the complexity of the input. However, it should be noted that the output from our algorithms should be subject to some kind of post-processing before application of an optical character recognition algorithm. This will remove spurious misclassified dots in the image. Possible post-processing methods are median filtering, removal of isolated regions below a certain size threshold, rectangle assumptions, looking for rectangles of a certain width, as in newspapers, etc.

References

[1] S. N. Srihari, "Document image understanding," in *Proc. IEEE Computer Society Fall Joint Computer Conf.*, (Dallas, TX), pp. 87–96, Nov. 1986.

[2] 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*, vol. 20, pp. 375–390, 1982.

[3] P. Chauvet, J. Lopez-Krahe, E. Taflin, and H. Maître, "System for an intelligent office document analysis, recognition and description," *Signal Processing*, vol. 32, no. 1-2, pp. 161–190, 1993.

[4] D. Wang and S. N. Srihari, "Classification of newspaper image blocks using texture analysis," *Computer Vision, Graphics and Image Processing*, no. 47, pp. 327–352, 1989.

[5] G. Nagy, S. Seth, and M. Viswanathan, "A prototype document image analysis system for technical journals," *Computer*, vol. 25, no. 7, pp. 10–22, 1992.

[6] M. Krishnamoorthy, G. Nagy, S. Seth, and M. Viswanathan, "Syntactic segmentation and labeling of digitized pages from technical journals," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 15, no. 7, pp. 737–747, 1993.

[7] L. A. Fletcher and R. Kasturi, "A robust algorithm for text string separation from mixed text/graphics images," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 10, Nov. 1988.

[8] A. K. Jain and S. Bhattacharjee, "Text segmentation using Gabor filters for automatic document processing," *Machine Vision and Applications*, vol. 5, no. 3, pp. 169–184, 1992.

[9] D. Gabor, "Theory of communication," *J. Inst Elect. Engr.*, vol. 93, pp. 429–457, 1946.

[10] J. G. Daugman, "Uncertainty relation for resolution in space, spatial-frequency, and orientation optimized by two-dimensional visual cortical filters," *Journal of Optical Society of America*, vol. 2, no. 7, pp. 1160–1169, 1985.

[11] A. K. Jain and F. Farrokhnia, "Unsupervised texture segmentation using Gabor filters," *Pattern Recognition*, vol. 24, no. 12, pp. 1167–1186, 1991.

[12] P. P. Vaidyanathan, "Quadrature mirror filter banks, M-band extensions and perfect-reconstruction techniques," *IEEE ASSP Magazine*, vol. 4, pp. 4–20, July 1987.

[13] J. H. Husøy, "Low complexity subband coding of still images and video," *Optical Engineering*, vol. 30, pp. 904–911, July 1991.

[14] T. Randen and J. H. Husøy, "Multichannel filtering for image texture segmentation," *Optical Engineering*, vol. 33, pp. 2617–2625, Aug. 1994.

[15] T. Randen and J. H. Husøy, "Segmentation of text/image documents using texture approaches," in *Proc. NOBIM-konferansen-94*, (Asker, Norway), pp. 60–67, June 1994.

[16] M. Tuceryan and A. K. Jain, "Texture analysis," in *Handbook of pattern recognition and computer vision* (C. H. Chen, L. F. Pau, and P. S. P. Wang, eds.), ch. 2, pp. 235–276, Singapore: World Scientific, 1993.

[17] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York, NY, USA: Wiley, 1973.