Document Image Segmentation Using Wavelet Scale–Space Features

Mausumi Acharyya and Malay K. Kundu, *Senior Member, IEEE*

***Abstract*—In this paper, an efficient and computationally fast method for segmenting text and graphics part of document images based on textural cues is presented. We assume that the graphics part have different textural properties than the nongraphics (text) part. The segmentation method uses the notion of multiscale wavelet analysis and statistical pattern recognition. We have used M-band wavelets which decompose an image into $M \times M$ bandpass channels. Various combinations of these channels represent the image at different scales and orientations in the frequency plane. The objective is to transform the edges between textures into detectable discontinuities and create the feature maps which give a measure of the local energy around each pixel at different scales. From these feature maps, a scale–space signature is derived, which is the vector of features at different scales taken at each single pixel in an image. We achieve segmentation by simple analysis of the scale–space signature with traditional *k-means* clustering. We do not assume any *a priori* information regarding the font size, scanning resolution, type of layout, etc. of the document in our segmentation scheme.**

***Index Terms*—Document segmentation, M-band wavelet, texture segmentation.**

I. Introduction

IN TODAY's world, automated processing and reading of documents has become an imperative need with the advances in communication and information technology. Efforts have been made to store the documents in digitized form, but that requires an enormous storing space, even after compression using modern techniques.

Documents can be more effectively represented by separating the text and the graphics/image part and storing the text as an ASCII (character) set and the graphics/image part as bit-maps. Document image segmentation plays an important role because this facilitates efficient searching and storage of the text part in documents, required in large databases. Consequently, several researchers have attempted different techniques to segment the text and graphics part in document images. Several useful techniques for text–graphics segmentation are given in [1], the most popular amongst these being the *top-down* and *bottom-up* approaches. The most common *top-down* techniques are *run-length smoothing* [2]–[4] and *projection profiles* [5]–[8]. *Top-down* approaches first split the document into blocks, which are then identified and subdivided appropriately in terms of columns first and then into paragraphs, text lines, and maybe also words [6], [7]. Some assume these blocks to be only rectangular [9]. The *top-down* methods are not suitable for skewed texts, as these methods are restricted to rectangular blocks, whereas the *bottom-up* methods are typically variants of the *connected components* [10]–[12] which iteratively group together components of the same type starting from the pixel level and form higher level descriptions of the printed regions of the document (words, text lines, paragraphs etc.) [13]. The drawbacks with the connected components method is that it is sensitive to character size, scanning resolution, inter-line, and inter-character spacings.

Several other approaches use the contours of the white space to delineate the text and nontext regions [14], [15]. These methods can only be applied to low-noise document images which are highly structured; that is, all objects are separated by a white background and objects do not touch each other.

Each of the above methods relies to an extent on *a priori* knowledge about the rectangularity of major blocks, consistency in horizontal and vertical spacings, independence of text, graphics, and image blocks, and/or assumptions about textual and graphical attributes like font size, text line orientation etc. So, these methods cannot work in a generic environment. It is desirable to have segmentation techniques which do not require any *a priori* knowledge about the content and attributes of the document image, or rather any such knowledge might not be available in some applications. Additionally, these methods operate on thresholded images. So, for a degraded image due to poor image-capturing conditions, the choice of an appropriate threshold is a very difficult task.

Jain and Bhattacharjee's [16] method has been able to overcome these restrictions and does not require an *a priori* knowledge of the document to be processed. The method proposed by them is mainly the same as that proposed earlier by Jain and Farrokhnia [17]. The basic assumption underlying this approach is that the text and the graphics parts are considered as two different textured regions. The document segmentation has been achieved by a texture segmentation scheme using Gabor filter as the feature extractor. One major drawback of this approach has is that the use of Gabor filter makes it computationally very expensive, as this scheme uses a subset of 20 fixed multichannel Gabor filters consisting of four orientations and five spatial frequencies, as input features for the classifier. Randen and Husøy [18] proposed a method using critically sampled infinite-impulse-response (IIR) quadrature mirror filter (QMF) banks for extracting features, thus saving considerable computational time than that required in the method in [16]. Both of the

Manuscript received March 3, 2000; revised September 11, 2002. The work of M. Acharyya was supported by the Council of Scientific and Industrial Research (CSIR), New Delhi, India, under Grant 9/93(66)2000-EMR-I. This paper was recommended by Associate Editor S. Panchanathan.

The authors are with the Machine Intelligence Unit, Indian Statistical Institute, Kolkata-700 108, India (e-mail: res9522@isical.ac.in; malay@isical.ac.in).

Digital Object Identifier 10.1109/TCSVT.2002.806812

aforementioned methods do not take into consideration the possibility of overlapped/mixed classes. Etemad *et al.* [19] have developed an algorithm for document segmentation based on document texture using multiscale feature vectors and fuzzy local decision information.

More recent research on document segmentation include works by Choi and Baraniuk, which are based on the wavelet-domain hidden Markov tree (HMT) [20]. Li and Gray [21] have developed an algorithm for segmenting document images into four classes: background, photograph, text, and graph. Features used for classification are based on the distribution patterns of wavelet coefficients in high-frequency bands. Harit *et al.* presents a new model-based document image-segmentation scheme that uses eXtensible Markup Language Document Type Definitions (XML-DTDs). Given a document image, the algorithm has the ability to select the appropriate model. A wavelet-based tool has been designed by them for distinguishing text from nontext regions and characterization of font sizes.

Some of the common difficulties that occur in documents are:

- differences in font size, column layout, orientation, and other textual attributes;
- skewed documents and text regions with different orientations;
- degraded documents due to improper scanning;
- combinations of varying text and background gray levels;
- text regions touching or overlapping with nontext regions;
- irregular layout structures with nonconvex or overlapping object boundaries;
- multicolumn document with misaligned text lines and different languages.

We develop a texture-based document image segmentation which takes care of all the above observations. This paper is based on the assumption that the text portion of the document image is comprised of one texture class and the nontext part of the other. Basically, this is a two-texture segmentation problem. It is already well known that textures can be characterized by their energies. A composite texture can be discriminated if it is possible to obtain information about the texture signal energies. The basic idea is to decompose the composite image into different frequency bands at different scales. The objective is to transform the edges between textures into detectable discontinuities. This creates the feature maps which give a measure of local energy around each pixel over small windows. We conjecture that a decomposition scheme yielding a large number of subbands would definitely improve segmentation results.

The segmentation method is based on multiscale (multiresolution) wavelet analysis. A multiscale representation provides a simple hierarchical framework for interpreting the image information. At different scales, the details of an image generally characterize different physical structures of the scene. At a coarse resolution or scale, these details correspond to larger structures, which provide the image context. The idea of scale–space refers to a family of derived signals where the fine-scale information is successively suppressed as scale increases. Features that are derived from the edges in an image respond to extractors (detectors) at different scales, and so feature extraction for an image containing features at multiple scales should incorporate multiscale information. In image processing, multiscale analysis provides a representation of an image that allows information from each scale to be analyzed separately. A feature-extraction scheme based on multiscale analysis and pattern recognition has several potential advantages over other existing feature-extraction methods, including extraction of features at different scales (i.e., features of all sizes) and robustness.

One of the salient features of document segmentation, as compared to other texture segmentation problems, is that there are large intra-class, as well as inter-class, variations in the textural features, so the multiscale nature of the document constituents which are implicitly there (e.g., characters, lines) justifies the need for a multiscale representation scheme. In this paper, we propose using an M-band wavelet decomposition of the image. The M-band wavelet transform performs a multiscale, multidirectional filtering of the images. It is a tool to view any signal at different scales and decomposes a signal by projecting it onto a family of functions generated from a single wavelet basis via its dilations and translations. Various combinations of the M-band wavelet filter decompose the image at different scales and orientations in the frequency plane. The filter extracts local frequencies of the image, which in essence gives a measure of local energies of the image over small windows around each pixel. These energies are characteristics of a texture and give the features required for classification of the various textured regions in an image.

One of the drawbacks of standard wavelets $(M = 2)$ is that they are not suitable for the analysis of high-frequency signals with relatively narrow bandwidth. Hence, the main motivation of this paper is to use the decomposition scheme based on M-band wavelets (where $M > 2$), which yield improved segmentation accuracies. Unlike the standard wavelet decomposition which gives a logarithmic frequency resolution, the M-band decomposition gives a mixture of a logarithmic and linear frequency resolution. Furthermore, as an additional advantage, M-band wavelet decompositions yield a large number of subbands which are required for good-quality segmentation.

In Section II, we give a brief overview of M-band wavelet transform. Section III describes the general texture segmentation setup, mainly as suggested in [22], and in Section IV we present the results of our experiment on different images under varying conditions. Finally, we conclude our study in Section V.

II. M-Band Wavelet Transform

The *wavelet transform* maps a function $f(x) \in L^2(R)$ onto a scale–space plane. The wavelets are obtained from a single prototype function $\psi(x)$ by scalings a and shifts b [23]–[25]. The continuous wavelet transform of a function $f(x)$ is given as

$$Wf_a(b) = \int f(x)\psi^*_{a,b}(x)\,dx. \tag{1}$$

M-band wavelet decomposition is a direct generalization of the above two-band case [26], [27]. Let $\phi(x)$ be the scaling function satisfying

$$\phi(x) = \sum_k h(k)\sqrt{M}\phi(Mx - k). \tag{2}$$

Additionally, there are $M-1$ wavelets which also satisfy

$$\psi^{(j)}(x) = \sum_k \sqrt{M}h^{(j)}(k)\psi(Mx - k). \tag{3}$$

In discrete form, these equations can be written as

$$\phi_{ik}(x) = \sum_k M^{-i/2}\phi(M^{-i}x - k) \tag{4}$$

$$\psi_{ik}^{(j)}(x) = \sum_k M^{-i/2}\psi^{(j)}(M^{-i}x - k), \quad j = 1, \ldots, M-1. \tag{5}$$

The subspaces spanned by the functions $\phi_{ik}(x)$ and $\psi_{ik}^{(j)}(x)$ can be, respectively, defined as

$$V_i = \overline{\text{span}\phi_{ik}, \forall k \in Z} \tag{6}$$

$$W_i^{(j)} = \overline{\text{span}\psi_{ik}^{(j)}, \forall k \in Z}. \tag{7}$$

It follows from (2) that the V_i subspaces have a nested property. If the scaling and the wavelet functions satisfy the orthonormality condition, then the subspaces $\{W_i^{(j)}\}$ form an orthogonal decomposition of the $L^2(R)$ function space and are related to the V_i nested subspaces by

$$V_i = V_{i+1} \oplus \left[\oplus_{j=1}^{M-1} W_{i+1}^{(j)}\right]. \tag{8}$$

A function $f(x) \in L^2(R)$ can be constructed from a discrete sequence $a(k) \in l^2(R)$ in the form

$$f(x) = \sum_k a(k)\phi(x - k). \tag{9}$$

$f(x)$ can also be expressed in terms of the sum of projections onto subspaces V_i and $W_i^{(j)}$ as

$$f(x) = \sum_k c(k)\phi_{i,k}(x) + \sum_{j=1}^{M-1}\sum_k d_j(k)\psi_{i,k}^{(j)}(x). \tag{10}$$

The expansion coefficients can be expressed as

$$c(k) = \langle f, \phi_{i,k}\rangle$$
$$d_j(k) = \left\langle f, \psi_{i,k}^{(j)}\right\rangle, \quad j = 1, \ldots M-1. \tag{11}$$

Using (2) and (3) in (11), it can be shown that

$$c(k) = \frac{1}{\sqrt{M}}\sum_l a(l)h(Mk - l) \tag{12}$$

$$d^{(j)}(k) = \sum_l a(l)h^{(j)}(Mk - l) \tag{13}$$

which is equivalent to processing the sequence $a(k)$ with a set of linear time-invariant filters of impulse responses $p_j = (1/\sqrt{M})h^{(j)}(k)$ and downsampling filter outputs by M.

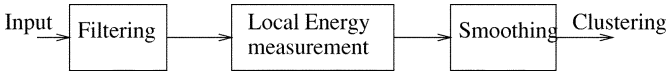


Fig. 1. Typical system setup of our work.

TABLE I
FILTER COEFFICIENTS FOR EIGHT-TAP FOUR-BAND WAVELET

No. of Taps (n)	$\psi_1(n)$	$\psi_2(n)$	$\psi_3(n)$	$\psi_4(n)$
0	-0.067371764	-0.094195111	-0.094195111	-0.067371764
1	0.094195111	0.067371764	-0.067371764	-0.094195111
2	0.40580489	0.56737176	0.56737176	0.40580489
3	0.56737176	0.40580489	-0.40580489	-0.56737176
4	0.56737176	-0.40580489	-0.40580489	0.56737176
5	0.40580489	-0.56737176	0.56737176	-0.40580489
6	0.094195111	-0.067371764	-0.067371764	0.094195111
7	-0.067371764	0.094195111	-0.094195111	0.067371764

III. MULTISCALE FEATURE EXTRACTION

A. Wavelet Scale–Space Features

The feature-extraction scheme that we have used has a multichannel filtering and a subsequent nonlinear stage followed by a smoothing filter (both these constitute the local energy estimator) as shown in Fig. 1. The objectives of the filtering and that of the local energy estimator, are to transform the edges between textures into detectable discontinuities.

B. M-Band Wavelet Filters

The filter bank, in essence, is a set of bandpass filters with frequency- and orientation-selective properties. In the filtering stage, we make use of an eight-tap, four-band, orthogonal– and linear phase wavelet transform following [26] to decompose the textured images into $M \times M(4 \times 4)$ channels (scale–space cell), corresponding to different direction and scales. The *one-dimensional* (1-D) four-band wavelet filter impulse responses denoted by ψ_r are given in Table I, and their corresponding transfer functions are represented by H_r for $r = 1, \ldots, 4$.

ψ_1 is the scaling function (low-pass filter) and the other ψ_r's correspond to the wavelet functions (high-pass filters). The M^2-channel *two-dimensional* (2-D) separable transform is obtained by the tensor product of the 1-D M-band wavelet filters, which are denoted by $\psi_{r,c}$, for $r, c = 1, \ldots, M$ with $M = 4$. In this paper, we extend the decomposition to the 2-D case by successively applying the M-band transform separably in the horizontal and vertical directions without downsampling (i.e., an overcomplete representation). The r, c^{th} scale–space cell is achieved via the filter $H_{r,c}$ for $r, c = 1, 2, 3, 4$ with $M = 4$. The size of the filter is an important factor. The filter length is increased with increased level of decomposition. The sequence of low-pass and bandpass filters of increasing width corresponding to an increased level of decomposition are expanded by inserting an appropriate number of zeros between taps of filters. So, if the filter length becomes large, it is possible that it may bias the decomposition of the image. We have chosen an eight-tap filter for suitability of the size of the image that we have considered in this study (i.e., 512×512).

The objective of the filtering is to find out about the discontinuities that exist within the image. The spectral response is

strongest along the direction perpendicular to the edge of an image, while it decreases as the direction of the filter approaches that of the the edge. Therefore, we can perform edge detection by using 2-D filtering of the image as follows:

- horizontal edges are detected by high-pass filtering on columns and low-pass filtering on rows;
- vertical edges are detected by low-pass filtering on columns and high-pass filtering on rows;
- diagonal edges are detected by high-pass filtering on columns and high-pass filtering on rows;
- horizontal–diagonal edges are detected by high-pass filtering on columns and low-pass filtering on rows;
- vertical–diagonal edges are detected by low-pass filtering on columns and high-pass filtering on rows.

A typical edge-detection filter corresponding to a particular direction covers a certain region in the 2-D spatial-frequency domain. Based on this concept, several wavelet-decomposition filters are designed which are given by the summations $\sum_{\mathrm{Reg}} H_{r,c}$, where Reg denotes the frequency sector of a certain direction and scale.

Since the filter system we are using is orthogonal and has a perfect-reconstruction quadrature mirror filter (PR-QMF) structure, that is $\sum_{r=1}^{M}\sum_{c=1}^{M}\psi_{r,c}\psi_{r,c}^{*}=1$, all frequencies in each scale–space cell are treated as equally possible by the resulting filters. The number of channels and, therefore, the number of possible filter combinations depend on the value of M. The wavelet filters denoted by $\sum_{\mathrm{Reg}} H_{r,c}$ for different directions with increasing scales can be derived as follows.

- Horizontal direction:

$$\begin{aligned}\mathrm{filt}_{\mathrm{hor}\,1} &= H_{12}\\ \mathrm{filt}_{\mathrm{hor}\,2} &= H_{12}+H_{13}\\ \mathrm{filt}_{\mathrm{hor}\,3} &= H_{12}+H_{13}+H_{14}+H_{24}.\end{aligned}$$

- Vertical direction:

$$\begin{aligned}\mathrm{filt}_{\mathrm{ver}\,1} &= H_{21}\\ \mathrm{filt}_{\mathrm{ver}\,2} &= H_{21}+H_{31}\\ \mathrm{filt}_{\mathrm{ver}\,3} &= H_{21}+H_{31}+H_{41}+H_{42}.\end{aligned}$$

- Diagonal direction:

$$\begin{aligned}\mathrm{filt}_{\mathrm{diag}\,1} &= H_{22}\\ \mathrm{filt}_{\mathrm{diag}\,2} &= H_{22}+H_{33}\\ \mathrm{filt}_{\mathrm{diag}\,3} &= H_{22}+H_{33}+H_{44}.\end{aligned}$$

- Horizontal–diagonal direction:

$$\begin{aligned}\mathrm{filt}_{\mathrm{hdiag}\,1} &= H_{12}\\ \mathrm{filt}_{\mathrm{hdiag}\,2} &= H_{12}+H_{23}\\ \mathrm{filt}_{\mathrm{hdiag}\,3} &= H_{12}+H_{23}+H_{34}.\end{aligned}$$

- Vertical–diagonal direction:

$$\begin{aligned}\mathrm{filt}_{\mathrm{vdiag}\,1} &= H_{21}\\ \mathrm{filt}_{\mathrm{vdiag}\,2} &= H_{21}+H_{32}\\ \mathrm{filt}_{\mathrm{vdiag}\,3} &= H_{21}+H_{32}+H_{43}.\end{aligned}$$

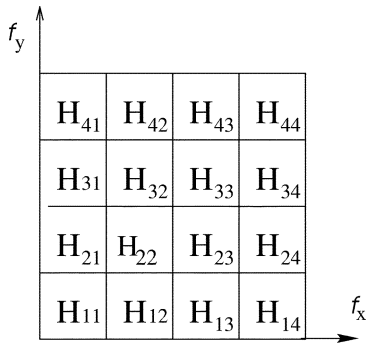


Fig. 2. Frequency bands corresponding to decomposition filters.

The $M \times M$ scale–space cells corresponding to the decomposition filters are given in Fig. 2, f_x and f_y are the frequencies in horizontal and vertical directions, respectively. These filter outputs basically give a measure of signal energies at different directions and scales $h_{k_i}(x,y)$ where $k = \mathrm{hor}, \mathrm{ver}, \mathrm{diag}, \mathrm{hdiag}$ and $v\,\mathrm{diag}$ and $i = 1, 2, 3$.

C. Local Energy Estimation

The next step is to estimate the energy of the filter responses in a local region around each pixel. The local energy estimate is utilized for the purpose of identifying areas in each channel where the band pass frequency components are strong resulting in a high energy value and the areas where it is weak into a low energy value. Although energy is usually defined in terms of a squaring nonlinearity, in a generalized energy function, however, other alternatives are also used.

We have studied several nonlinear operators. These include the *magnitude operation, average absolute deviation and standard deviation* calculated over small overlapping windows around each pixel. The local energy $\mathrm{eng}_{k_i}(x,y)$ around the x,yth pixel for several nonlinearities are formally given as follows:

magnitude operation

$$\mathrm{eng}_{k_i}(x,y) = |h_{k_i}(m,n)| \tag{14}$$

average absolute deviation

$$\mathrm{eng}_{k_i}(x,y) = \frac{1}{R}\sum_{m=1}^{w}\sum_{n=1}^{w}|(h_{k_i}(m,n) - \overline{h}_{k_i}(x,y))| \tag{15}$$

standard deviation

$$\mathrm{eng}_{k}(x,y) = \sqrt{\frac{1}{R}\sum_{m=1}^{w}\sum_{n=1}^{w}\left|\left(h_{k_i}(m,n)^2 - \bar{h}_{k_i}(x,y)^2\right)\right|} \tag{16}$$

where w is the window size and $R = w \times w$, while $\overline{h}_{k_i}(x,y)$ is the mean around the (x,y)th pixel and $h_{k_i}(x,y)$ is the filtered image.

We have experimentally observed that the standard deviation over small overlapping windows around each pixel gives better performance than the other nonlinearities mentioned above. This nonlinear operator is independent of any parameter, i.e., independent of the dynamic range of the input image and also of the filter amplification.

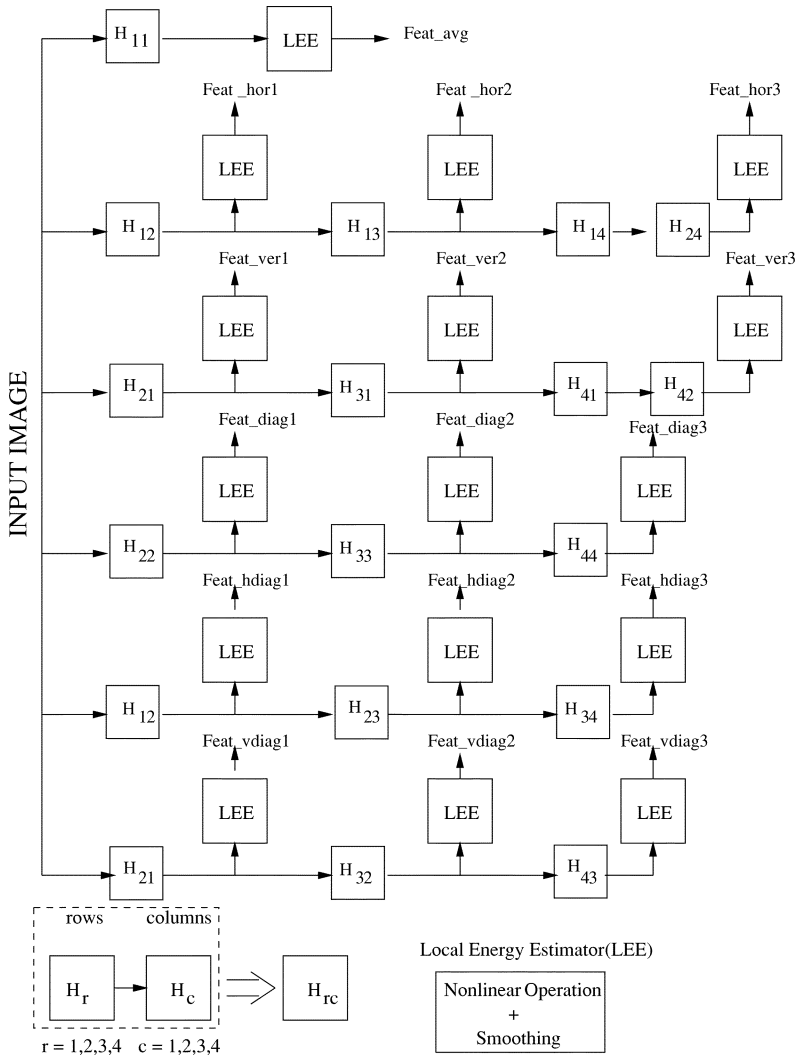


Fig. 3. Basic decomposition scheme in this paper.

The nonlinear transform is succeeded by a Gaussian low-pass (smoothing) filter is of the form

$$H_G(u,v) = \frac{1}{2\pi\sqrt{\sigma}} e^{\frac{1}{2\sigma^2}(u^2+v^2)} \tag{17}$$

where σ determines the passband width of the averaging window. Formally, the feature image $\mathrm{Feat}_{k_i}(x,y)$ corresponding to filtered image $h_{k_i}(x,y)$ is given by

$$\mathrm{Feat}_{k_i}(x,y) = \frac{1}{G^2} \sum_{(m,n)\in G_{x,y}} |\Psi(h_{k_i}(m,n))| \tag{18}$$

where $\Psi(\cdot)$ is the local energy estimator and $G_{x,y}$ is a $G \times G$ window centered at pixel with coordinates (x,y). The size G of the smoothing or the averaging window in (4) is an important parameter. More reliable measurement of the texture feature calls for larger window sizes so as to accommodate the texture periodicity completely. On the other hand, more accurate localization of region boundaries calls for smaller windows. This is because averaging blurs the boundaries between textured regions. Another important aspect is that Gaussian-weighted windows are naturally preferable over unweighted windows because the former are likely to result in more accurate localization of texture boundaries.

We found an averaging window size of 9×9 to be appropriate in most of our segmentation experiments, while the size of the window for local energy measurement was found to vary between 5×5 to 13×13, in most cases. However, in some of the images, other window sizes have also been used.

The images resulting from these operations are the features denoted by $\mathrm{Feat}_{\mathrm{hor}_i}, \mathrm{Feat}_{\mathrm{ver}_i}$, etc. for $i = 1,2,3$, as shown in Fig. 3.

D. Algorithm

The texture segmentation algorithm based on the M-band wavelet decomposition is illustrated in Fig. 3. In this figure, a set of filters H_{rc} (where $r,c = 1,2,3,4$) are used. The output corresponding to the filtering H_{rc} is obtained by convolving the image in a separable manner with H_r along rows and then with H_c along columns. The local energy estimator is denoted by LEE, and the features images (Feat_{k_i}) are obtained as shown.

This algorithm consists of the following steps.

- The input image is first decomposed into 4×4 channels by wavelet analysis without downsampling as referred

in Section III.B. In this paper, we have used an eight-tap four-band wavelet [26], so in all we get 16 decomposition channels, as discussed in Section III-B, which means the feature set comprises of 16 feature elements. Out of these 16 features, we ignore the low-frequency channel feature corresponding to $\mathrm{Feat}_{\mathrm{avg}}$, $\mathrm{Feat}_{\mathrm{hdiag}_1}$, and $\mathrm{Feat}_{\mathrm{vdiag}_1}$, since these are nothing but $\mathrm{Feat}_{\mathrm{hor}_1}$ and $\mathrm{Feat}_{\mathrm{ver}_1}$, respectively. Now we are left with 13 features.

- These outputs are subjected to the nonlinear operation followed by smoothing as discussed in Section III-C, which then form the feature vector F_q, where $q = 1, \ldots, Q$.
- We have a matrix of $Q \times S$, where Q is the number of feature elements in each vector (13 in this case) and S is the total data size (the total number of pixels in the input image, which is N^2 for an image of size $N \times N$). This step gives us the class map corresponding to the composite texture image.

E. Unsupervised Classifier

Having obtained the feature images, the main task is to integrate these feature images to achieve segmentation. We define a scale–space signature as the vector of features at different scales taken at a single pixel in an image

$$\bar{F}(x,y) = [F_1(x,y), F_1(x,y), F_2(x,y), \ldots, F_Q(x,y)]. \quad (19)$$

Suppose these scale–space signatures are considered as feature vectors in a feature space.

Let us assume that there are K texture categories present in the image. If our texture features that have already been obtained are capable of discriminating between these categories, then the patterns belonging to each category will form a cluster in the feature space which is compact and isolated from clusters corresponding to other texture categories. Pattern-clustering algorithms are ideal modes for forming such clusters in the feature space. Segmentation algorithms accept as input a set of features and use consistent labeling for each pixel. Fundamentally this can be considered a multidimensional data clustering problem. This paper is a two-class segmentation problem. A supervised version of the classifier for document image segmentation would mean that the segmentation is dependent on the knowledge of scale, scanning resolution, rotation, skewness, font size, type of layout, etc. of the document. Whereas our aim in this work has been to make the segmentation scheme independent of all the aforesaid issues and is robust, we thus need an unsupervised classifier. Also, since we emphasize the feature-extraction (representation) part, we have thus used a traditional k-*means* clustering algorithm [28]. An overview of the unsupervised k-*means* clustering algorithm is given below.

k-*means* $(x[1:Q, 1:S], K)$
$x[1:Q, 1:S]$: array of structure containing vectors, Q: Number of feature elements in a feature vector, S: Data size (number
of pixels in the image), K: Number of classes
begin
  begin (Initialization)
    Select K number of vectors arbitrarily from the array $x[1:Q, 1:S]$ and then each of these are assigned a class, these form the initial class centers C_k's.
  end
  begin
    Euclidean distance between each of the S vectors and selected K vectors are found out taking one out of S vectors at a time. A vector is assigned to the class k if it is closest to C_k. Recompute the class centers C_k taking mean of the vectors assigned to class k.
    Repeat until there is no change in the class centers.
  end
end

IV. Experiments

A. Test Images

Several document images were analyzed using our texture segmentation algorithm, so as to demonstrate the performance of our algorithm. These documents were scanned from parts of pages of the *Times of India* (TOI) and *Hindustan Times*, both of which are popular news dailies in India.

- Structured document image with nonoverlapping text and nontext regions scanned from *Times of India*.
  - Fig. 4(a) shows a document image of size 512×512.
  - Fig. 5(a) shows the same image rotated by 22.5°.
  - Fig. 6(a) shows the same image rotated 90°.
  - Fig. 7(a) shows the same image skewed by 25°.
  - Fig. 8(a) shows another image of size 512×512.
  - Fig. 8(c) shows the same image scanned at half the resolution.
- Highly unstructured images with overlapped/mixed classes of size 512×512 scanned from the *Hindustan Times*.
  - Fig. 9(a) shows a test image with document skewed and text regions with different orientations.
  - Fig. 10(a) shows a test image with nonconvex and overlapping object boundaries.
  - Fig. 11(a) shows a document image with irregular nontext region and multicolumn document with misaligned text lines and different languages.
  - Fig. 12(a) shows text portions with different orientations, as well as gray values, and different font sizes.
  - Fig. 13(a) shows text regions which overlap with nontext regions, combinations of varying text and background gray level, and text regions with widely varying font size.
  - Fig. 15(a) shows a 512×512 test image that has been used by Randen and Husøy [18].
  - Fig. 16(a) shows a 512×512 test image that has been used by Jain and Bhattacharjee [16].

B. Results

Out of the total 16 features possible in our decomposition scheme, we have found that only ten features (empirically chosen) give us the desired result. The number of features could even be reduced without any perceptible degradation of our segmentation results. In most of the cases, the number

(a) (b)

(c) (d)

Fig. 4. (a) A portion of a typical page of TOI. (b) Two-class segments from unsupervised segmentation. (c) Segmentation with regions classified excluding the picture. (d) Image segmentation after histogram thresholding.

(a) (b)

Fig. 5. (a) Same image rotated by 22.5°. (b) Segmented result.

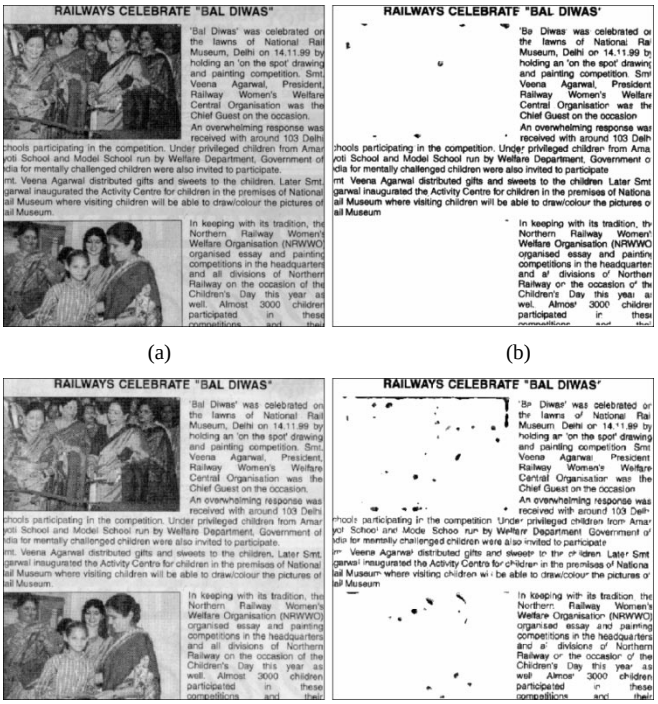
(a) (b)

Fig. 6. (a) Same image rotated by 90°. (b) Two-class segmentation.

of features was limited to between three and five. The energy of each of the 13 subbands were calculated and a ranking of these subbands were done in accordance to the magnitude of these energies. In this study, we have taken into account those subbands which have the highest values of energies, signifying

(a) (b)

Fig. 7. (a) Same image skewed by an angle 25°. (b) Corresponding image segmentation.



(a) (b)

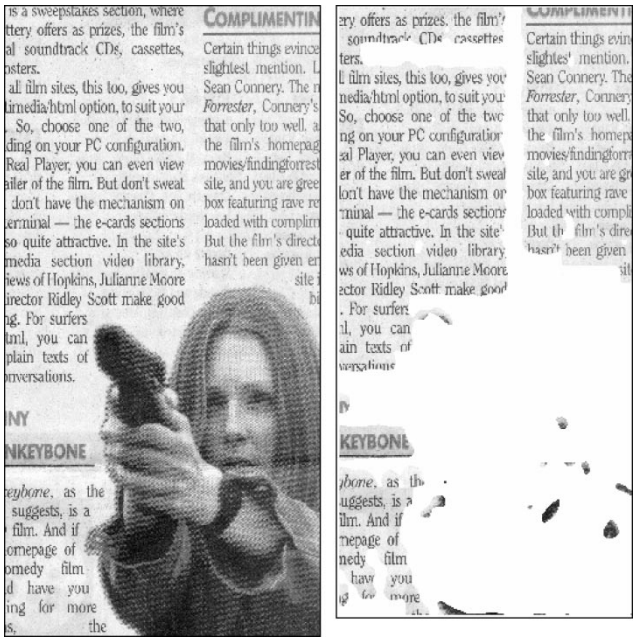
(c) (d)

Fig. 8. (a) Image of a portion of a typical page of TOI. (b) Two-class segmentation. (c) Same image scanned at half the resolution. (d) Image segmentation.

(a) (b)

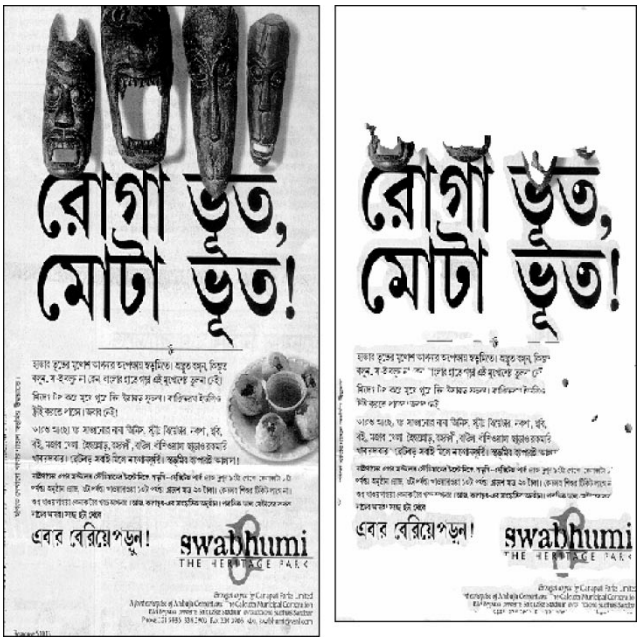
Fig. 9. (a) Test image with document skewed and text regions with different orientations. (b) Segmented result.

that these subbands would furnish more information than the others. The higher the value of energy of a subband, the more information it contains.



(a) (b)

Fig. 10. (a) Test image with nonconvex and overlapping object boundaries. (b) Segmented result.



(a) (b)

Fig. 11. (a) Test image with irregular nontext region, multicolumn document with misaligned text lines and different languages. (b) Segmented result.

Fig. 4(a) shows a 512×512 pixel scanned image of a portion of a typical page of TOI, and the successful two-class segmentation of the image [Fig. 4(b)–(d)]. In order to prove the efficacy of our algorithm, we apply this technique to segment the same image is rotated by 22.5° [Fig. 5(a)]. Since integer rotations are distortion free, we use fractional rotation deliberately to see how our algorithm behaves in such cases. The image is rotated about an axis through the center of it in a clockwise direction. The transformed coordinates in the rotated plane for a rotation of θ are given by

$$x' = x\cos\theta + y\sin\theta \quad \text{and} \quad y' = -x\sin\theta + y\cos\theta.$$

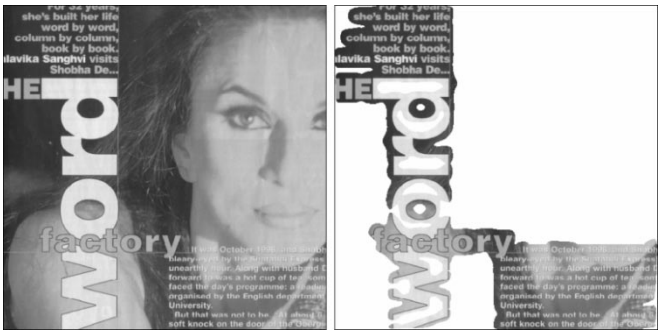
The segmentation results in Fig. 5(b) show that our method is almost invariant to rotation. We also rotate the document by 90° [Fig. 6(a)] and find that in this case, we also get an excellent segmentation result [Fig. 6(b)]; thus, we can say that the scheme is almost independent of any specific layout of the document.

The same image is also given a skewed transform about an axis through the center and the transformed coordinates for a skewness of ϕ are given by

$$x' = x + y\tan\phi \quad \text{and} \quad y' = y.$$

The image is given a skew angle of 25° [Fig. 7(a)] and the result is shown in Fig. 7(b); so we can also say that our algorithm is somewhat independent of the degree of skew. Fig. 8(a) shows a 512×512 image scanned form a typical page in the TOI and the same image scanned at half the resolution [Fig. 8(c)]. The corresponding results are provided in Figs. 8(b) and (d). From the results, we can infer that our scheme is invariant to scale also.

Thus far, we have been concentrating on structured data with nonoverlapping text and nontext regions. However, there are several instances of documents which are highly unstructured.



(a) (b)

Fig. 12. (a) Test image with text portions having different orientations and gray values as well as different font sizes. (b) Segmented result.

We have experimented on several such data for an extensive study. Fig. 9(a) shows a test image with a document skewed and text regions with different orientations; the segmentation result is shown in Fig. 9(b). To prove the efficacy and robustness of our algorithm over different and diverse types of documents, we have applied it on another test image with nonconvex and overlapping object boundaries [Fig. 10(a)]. The segmentation result [Fig. 10(b)] shows that the algorithm can efficiently identify the text and nontext regions in the documents.

Meanwhile, Fig. 11(a) shows a document image with an irregular nontext region and a multicolumn document with misaligned text lines and different languages. The segmentation result given by Fig. 11(b) clearly reveals that our scheme is quite capable of identifying the text and nontext regions in such complicated documents also.

In Fig. 12(a), the text regions overlap with nontext regions, there are combinations of varying text and background gray level, and text regions have widely varying font size. Fig. 12(b) shows the text portion which is segmented. In Fig. 13(a), the text portions have different orientations and gray values, as well as

(a) (b)

Fig. 13. (a) Images with overlapping text and nontext regions. (b) Segmented result.

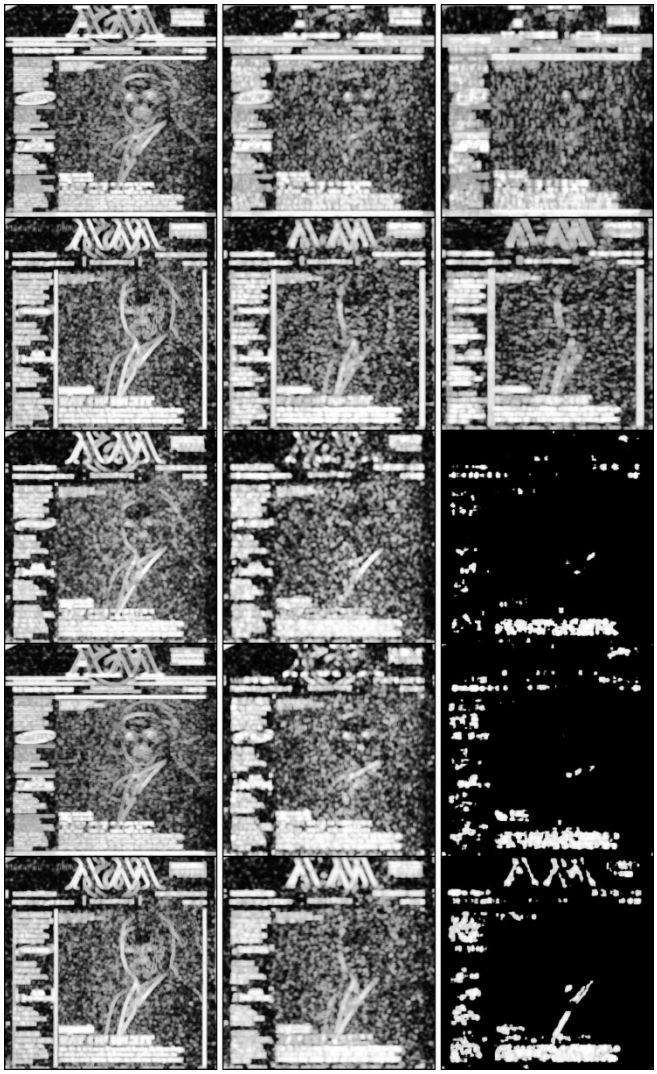


Fig. 14. Feature images for the document image in Fig. 13. Top row: $\mathrm{Feat}_{\mathrm{hor}_i}$. Second row: $\mathrm{Feat}_{\mathrm{ver}_i}$. Third row: $\mathrm{Feat}_{\mathrm{diag}_i}$. Fourth row: $\mathrm{Feat}_{\mathrm{hdiag}_i}$. Bottom row: $\mathrm{Feat}_{\mathrm{vdiag}_i}$, for $i = 1, 2, 3$.

different font sizes. The corresponding text segmentation result is shown in Fig. 13(b).

In Fig. 14, we present the feature maps corresponding to the test image in Fig. 13(a). This is to give an idea of how the image is decomposed into its constituent energies corresponding to different frequency bands and orientations. From the figure, it is well understood that the signal energies are discriminably distributed amongst different frequency bands and orientations, including horizontal, vertical, diagonal, and horizontal-diagonal and vertical-diagonal directions (here, we present the histogram-equalized versions of the original images for better visibility and understanding for the readers, although all the operations are performed on the original images).

(a) (b) (c) (d)

Fig. 15. (a) Segmentation results of the test image used in [18] using (b) our algorithm, (c) Randen and Husøy [18], and (d) Etemad *et al.* [19].

Throughout the experiment, our effort has been to segment the text part from the graphics part as accurately as possible. To compare our method with other methods, we have used the same data that has been used by Randen and Husøy [18] [Fig. 15(a)]. Using our algorithm, we find that although some of the graphics part of Fig. 15(b) is misclassified as text data, on the contrary, we get excellent results as far as text identification is concerned. The headings of two different font sizes could not be identified very accurately by Randen's method [18] [Fig. 15(c)], but have been possible by our method. The segmentation result obtained using classical wavelet packets and features suggested by Etemad *et al.* [19] [Fig. 15(d)] is also presented here for a comparative study.

For comparison purposes, we have also used the same data that has been used by Jain and Bhattacharjee [16] [Fig. 16(a)]. Using our algorithm, we find that the segmentation results [Fig. 16(c)] are more or less the same as in [18] [Fig. 16(c)]. Meanwhile, Randen and Husøy have verified that their method gives comparable segmentation results to that of the method used by Jain and Bhattacharjee. The result obtained using classical wavelet packets and features given by Etemad *et al.* is also shown in Fig. 16(d).

It is seen that there is a significant improvement in the segmentation result using $M(M > 2)$-band wavelets compared to the classical wavelet packets, where $M = 2$. This may be ex-

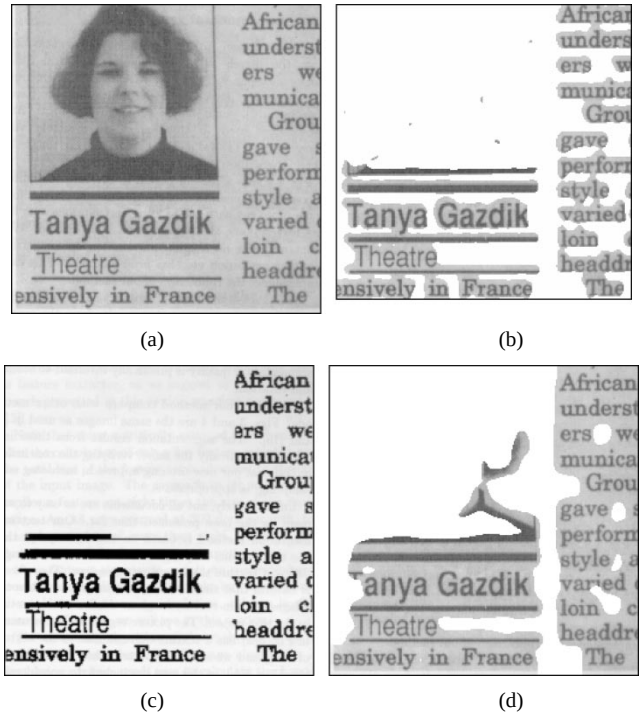


Fig. 16. (a) Segmentation results of the test image used by Jain and Bhattacharjee [16] using (b) our algorithm, (c) Randen and Husøy [18], and (d) Etemad *et al.* [19].

plained by the additional information obtained by decomposing the image into a higher number of subbands for a larger value of M. In addition, high-frequency signals with relatively narrow bandwidth are better resolved using wavelets with higher values of M. The investigation of Chang and Kuo [29] indicates that the texture features are more prevalent in the intermediate frequency band. Laine and Fan [30] have carried out studies on texture analysis based on this indication. Therefore, wavelets with a higher value of M can characterize textures more effectively in terms of its constituent signal energies because these wavelets give a logarithmic, as well as a linear, frequency resolution quite unlike the classical wavelets $(M = 2)$, which give only a logarithmic frequency resolution. This may also be attributed to the bandpass nature of higher bands that capture additional texture information. Also in our method, we have not subsampled the image, whereas in both the other methods used, for the purpose of comparison, the image has been subsampled. The suitability of our method over the subsampled methods may be explained by the fact that subsampling reduces the size of the subbands at higher levels of decomposition and can possibly bias the decomposition.

We have experimented on a variety of data having different combinations of font sizes of the text and have also considered diverse varieties of test images which are highly unstructured. In all of these data, we could reliably classify the text parts from the nontext parts.

Unlike [18], which uses a subsampling of the image for extraction of features and which introduces some edge inaccuracies, since our technique uses overcomplete representation of the image (meaning we have decomposed the image without downsampling), we get accurate edge localization, and translation invariance can also be achieved [31].

Although the method used in [18] has considerable computational savings over the approach of Jain and Bhattacharjee [16], none of the above-mentioned works have considered test data that are unstructured or that they have overlapping text and nontext regions. Whereas the work of Etemad *et al.* [19] has considered all of these data types, they have used fuzzy local-decision information for classification. This step clearly reduces the ambiguity between the various classes and hence gives better performance. We have shown that the features that we have extracted are appreciably authentic and provide good enough segmentation. However, we conjecture that incorporation of fuzziness in the features that we have extracted would have produced much better results than those reported in this paper.

It is to be noted that all of our experiments were performed with no *a priori* knowledge about the input image. We did not have any information about the font size or format of the text. While a knowledge about these can definitely improve the segmentation results, for this, we can make use of supervised segmentation.

V. Conclusion

In this paper, we have presented a new technique for segmenting the text part from the nontext part based on textural cues using M-band wavelet filters. The decomposition gives a multiscale multidirectional representation of the image and yields a large number of subbands. The filtering and the feature-extraction operations account for most of the required computations; however, our method is very simple, computationally less expensive, and efficient. It has been experimentally found that three to five features out of the 13 features are sufficient for good-quality segmentation. Hence, the dimensionality of the feature space is considerably reduced.

We have applied our algorithm on several structured and unstructured images which were decomposed into 16 channels without downsampling giving an overcomplete representation. Features were computed on the decomposed subbands using a local energy estimator over small overlapping windows around each pixel. A traditional k-*means* clustering algorithm was used for segmenting the text and nontext parts of the image, making use of the features so extracted. In contrast to most traditional methods for text–graphics segmentation, we do not assume any knowledge about the font size, scanning resolution, column layout, orientation, etc. of the input; that is, our approach is purely unsupervised. The results indicate that M-band wavelets have the efficacy to discriminate between textures, and can be effectively applied for document segmentation.

It is quite apparent that there is a need for digitization of documents for making it easily accessible via computers and networks, but it is not absolutely necessary to align the document in raster direction. It is also very important to separate the text part from the graphics part in paper documents, since it is difficult to store or retrieve the whole digitized document even if it is in compressed form. This is because if a document image consists of graphics regions, it is stored in a bit-map representation, making it practically impossible to search the text part of the document from this bit-map file. Also, if the text and graphics parts are separated and then coded separately using the

best-available coding algorithms for text and graphics, then the overall compression increase is manifold [32].

REFERENCES

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

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

[3] F. M. wahl, K. Y. Wong, and R. G. Kasey, "Block segmentation and text extraction in mixed text/image documents," *Comput. Graph. Image Process.*, vol. 20, pp. 375–390, 1982.

[4] F. Shih, S.-S. Chen, D. Hung, and P. Ng, "A document image segmentation, classification and recognition system," in *Proc. Int. Conf. Systems Integration*, 1992, pp. 258–267.

[5] O. Iwaki, H. Kida, and H. Arakawa, "A segmentation method based on office document hierarchical structure," in *Proc. Intl. Conf. Systems, Man and Cybernetics*, 1987, pp. 375–390.

[6] D. Wang and S. N. Srihari, "Classification of newspaper image blocks using texture analysis," *Compu. Vis., Graph., Image Process.*, vol. 47, pp. 327–352, 1989.

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

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

[9] S. Sural and P. K. Das, "A two step algorithm and its parallelization for the generation of minimum containing rectangles for document image segmentation," in *Proc. Int. Conf. Document Analysis and Recognition*, 1999, pp. 173–176.

[10] L. A. Fletcher and R. Kasturi, "A robust algorithm for text string separation from mixed text/graphics images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 10, pp. 910–918, Nov. 1988.

[11] T. Saitoh and T. Pavlidis, "Page segmentation without rectangle assumption," in *Proc. Int. Conf. Pattern Recognition*, 1992, pp. 277–280.

[12] F. Lebourgeois, Z. Bublinski, and H. Emptoz, "A fast and efficient method for extracting text paragraphs and graphics from unconstrained documents," in *Proc. Int. Conf. Pattern Recognition*, 1992, pp. 272–276.

[13] C. L. Tan, B. Yuan, W. Huang, and Z. Zang, "Text/graphics separation using pyramid operations," in *Proc. Int. Conf. Document Analysis and Recognition*, 1999, pp. 169–172.

[14] A. Antaonacopulos, "Page segmentation using the description of the background," *Comput. Vis. Image Understanding*, vol. 70, no. 3, pp. 350–369, 1998.

[15] T. Pavlidis and J. Zhou, "Page segmentation by white streams," in *Proc. 10th. Int. Conf. Pattern Recognition*, 1991, pp. 945–953.

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

[17] F. Farrokhnia and A. K. Jain, "A multichannel filtering approach to texture segmentation," *Proc. Computer Vision and Pattern Recognition*, pp. 364–370, 1991.

[18] T. Randen and J. H. Husøy, "Segmentation of text/image documents using texture approaches," *Proc. NOBIM-Konferansen-94*, pp. 60–67, June 1994.

[19] K. Etemad, D. Doermann, and R. Chellappa, "Multiscale segmentation of unstructured document pages using soft decision integration," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 92–96, 1997.

[20] H. Choi and R. G. Baraniuk, "Multiscale image segmentation using wavelet-domain hidden Markov models," *IEEE Trans. Image Processing*, vol. 10, pp. 1309–1321, 2001.

[21] J. Li and R. M. Gray, "Context-based multiscale classification of document images using wavelet coefficient distributions," *IEEE Trans. Image Processing*, vol. 9, pp. 1604–1616, 2000.

[22] M. Acharyya and M. K. Kundu, "An adaptive approach to unsupervised texture segmentation using M-band wavelet," *Signal Processing*, vol. 81, no. 7, pp. 1337–1356, 2001.

[23] I. Daubechies, "Orthogonal bases for compactly supported wavelets," *Comm. Pure Appl. Math*, vol. 41, pp. 909–996, 1988.

[24] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, PA: Soc. Ind. Applied Math, 1992.

[25] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, pp. 674–693, 1989.

[26] O. Alkin and H. Caglar, "Design of efficient M-band coders with linear phase and perfect reconstruction properties," *IEEE Trans. Signal Processing*, vol. 43, pp. 1579–1590, July 1995.

[27] P. Steffen, P. N. Heller, R. A. Gopinath, and C. S. Burrus, "Theory of regular M-band wavelet bases," *IEEE Trans. Signal Processing*, vol. 41, pp. 3497–3510, Dec. 1993.

[28] J. T. Tou and R. C. Gonzales, *Pattern Recognition Principles*. Reading, MA: Addison-Wesley, 1974.

[29] T. Chang and C. C. J. Kuo, "Texture analysis and classification with tree-structured wavelet transform," *IEEE Trans. Image Processing*, vol. 2, pp. 42–44, Apr. 1993.

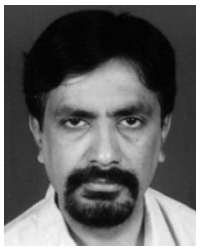
[30] A. Laine and J. Fan, "Frame representation for texture segmentation," *IEEE Trans. Image Processing*, vol. 5, pp. 771–779, May 1996.

[31] M. Unser, "Texture classification and segmentation using wavelet frames," *IEEE Trans. Image Processing*, vol. 4, pp. 1549–1560, Nov. 1995.

[32] O. J. Kwon and R. Chellappa, "Segmentation based image compression," *Opt. Eng.*, vol. 32, pp. 1581–1587, 1993.

Mausumi Acharyya received the B.Sc. degree in physics in 1988 and the B.Tech. and M.Tech. degrees in radiophysics and electronics in 1991 and 1993, respectively, all from the University of Calcutta, Calcutta, India. She has submitted her dissertation in computer science and is awaiting the Ph.D. degree.

Since 1995, she has been with the Machine Intelligence Unit, Indian Statistical Institute, Kolkata, where she is currently a Research Associate of the Council of Scientific and Industrial Research, New Delhi, India. Her current research interests are in the areas of image processing and analysis, pattern recognition, wavelets, and soft computing.



Malay K. Kundu (M'90–SM'99) received the B. Tech., M. Tech., and Ph.D. (Tech.) degrees in radio physics and electronics from the University of Calcutta, Calcutta, India.

In 1982, he joined the Indian Statistical Institute, Calcutta, as a faculty member. During 1993–1995, he was the Head of the Machine Intelligence Unit, and is now a Full Professor of the same unit. During 1988–1989, he was with the A.I. Laboratory, Massachusetts Institute of Technology, Cambridge, as a Visiting Scientist under the U.N. Fellowship Program. He has authored about 70 research papers in well-known and prestigious archival journals, international refereed conferences, and chapters in monographs and edited volumes. He is co-author of the book *Soft Computing for Image Processing* (Heidelberg, Germany: Physica-Verlag, 2000). His current research interests include image processing and analysis, image compression, machine vision, genetic algorithms, fractals, wavelets, VLSI design for digital imaging, and soft computing.

Dr. Kundu received the prestigious VASVIK award for industrial research (the Indian equivalent of the Mullard Award of the U.K.) in Electronic Sciences and Technology in 1999 and the Sir J. C. Bose Memorial Award from the Institute of Electronics and Telecommunication Engineers (IETE) of India in 1986. He is a Fellow of the National Academy of Sciences of India and the Institute of Electronics and Telecommunication Engineers of India.