Text Extraction and Document Image Segmentation Using Matched Wavelets and MRF Model

Sunil Kumar, Rajat Gupta, Nitin Khanna, *Student Member, IEEE*, Santanu Chaudhury, and Shiv Dutt Joshi

***Abstract*—In this paper, we have proposed a novel scheme for the extraction of textual areas of an image using globally matched wavelet filters. A clustering-based technique has been devised for estim ating globally matched wavelet filters using a collection of groundtruth images. We have extended our text extraction scheme for the segmentation of document images into text, background, and picture components (which include graphics and continuous tone images). Multiple, two-class Fisher classifiers have been used for this purpose. We also exploit contextual information by using a Markov random field formulation-based pixel labeling scheme for refinement of the segmentation results. Experimental results have established effectiveness of our approach.**

***Index Terms*—α-expansion, document image, globally matched wavelets (GMWs), matched wavelets, Markov random field (MRF), scene image.**

I. Introduction

In THIS paper, we address the problem of locating the textual data in an image. Further, we have extended text extraction scheme for the segmentation of document images. Fig. 1 shows some of the examples of different document and nondocument images (scene images) with text areas. Our text extraction scheme can identify and isolate textual regions in these kind of images. The proposed document image segmentation algorithm separates the document image into three classes, namely text, picture (or graphics), and background. Such a system finds applications in image and text database retrieval, automated processing and reading of documents, and storing the documents in digitized form.

There have been attempts in the past for extraction of textual component of an image by analyzing the edges of candidate regions or homogeneous color/gray scale components that contain the characters [8], [9], [18]. A number of page segmentation methods have been studied and compared objectively in [19]. The BESUS method [13], for example, is constructed using a number of morphology-based modules. Here text is extracted based on the spatial relationship between pairs of textlines which in turn is identified based on the similarity and distribution of connected components. The Ocè method [19] identifies connected components in the image and classifies them into different components (text, pictures, separators etc.) by a decision tree formulation with features such as width, height and number of pixels. Another significant text extraction method is proposed by Chen and Ding *et al.* [14]. In this method (termed as Tsinghua method), a bottom-up approach is followed by progressively merging image components at different levels based on the calculation of a quantitative measure [the multilevel confidence (MLC) value]. A refined version of this method was analyzed in [19], which can deal with irregular regions. Jain and Yu in [15] have surveyed some OCR and page segmentation algorithms. In this paper they have also suggested use of traditional bottom-up approach based upon connected component extraction to efficiently implement page segmentation and region identification. Wavelet transform-based text extraction systems have also been very effective. Li and Gray [7] have used distribution characteristics of wavelet coefficients for segmenting document images. Haar discrete wavelet transform (DWT) based approach has been suggested by Liang and Chen [21]. Kundu and Acharya [2] reported another scheme for segmentation of texts in the document images based on wavelet scale-space features. The method used M-band wavelet which decomposes an image into $\mathrm{M} * \mathrm{M}$ band-pass channels so as to detect the text regions easily. The real text regions are then recognized based on the intensity of the text edges in an M-band image. Etemad *et al.* [6] used a neural network to classify the output of wavelets into text and nontext regions. A technique for segmentation of an image into printed text, handwritten text and noise has been presented in [3]. Further, Markov random field (MRF) based post processing has been applied to exploit the context in the already obtained results. This leads to significant improvement of the results. Our segmentation scheme makes use of purely image based features. Thus, it has the advantage of locating text independent of scripts, font, font-size, geometric transformation, geometric distortion, and background texture.

In our scheme, we use the concept of matched wavelets [1] to develop the globally matched wavelet (GMW) filters specifically adapted for the text and nontext region [22]. We have used these filters for detecting text regions in scene images and for segmentation of document images into text, picture and background. We find the GMW filters by training matched wavelets on an image set. The key contribution of our work is that it is

Manuscript received April 5, 2006; revised March 1, 2007. This work was supported in part by Media Lab Asia, GOI, and in part by Media Lab Asia, Government of India. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tamas Sziranyi.

S. Kumar is with the IBM India Research Laboratory, Delhi-110 016, India (e-mail: sunil_narang@rediffmail.com).

R. Gupta is with Cypress Semiconductors, Bangalore-560046, India (e-mail: rajat_iitian@yahoo.com).

N. Khanna is with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47906 USA (e-mail: nitin_iitian@yahoo.com).

S. Chaudhury and S. D. Joshi are with the Department of Electrical Engineering, Indian Institute of Technology, Delhi, Hauz Khas, Delhi-110 016, India (e-mail: santanuc@ee.iitd.ac.in, sdjoshi@ee.iitd.ac.in).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2007.900098

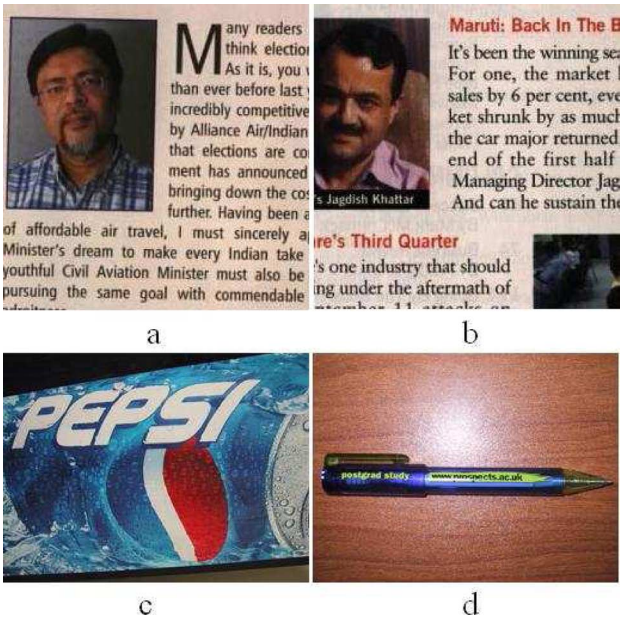


Fig. 1. (a)–(b) Examples of document images; (c)–(d) nondocument images. Note that, in (c) and (d), there is no clear distinction between background and image.

a trainable segmentation scheme based on matched wavelet filters. For high-performance systems, using application specific training image sets (e.g., license plate images, handwritten text images, printed text images), we can obtain filters customized for a particular application. Compared to other existing methods [2], [7], [10], the dimensionality, and, thus, the computation of the feature space, is considerably reduced. The filtering and the feature extraction operations account for most of the required computations; however, our method is very simple to understand, computationally less expensive, and efficient. In the latter part, we exploit the contextual information using MRF-based postprocessing to improve the results of document segmentation.

The rest of the paper is organized as follows. Section II gives the algorithm for the estimation of GMWs. In Section III, the use of GMWs in locating text regions is explained. In Section IV, the use of GMW filters for the document image segmentation has been presented. MRF-based postprocessing is presented in Section V. Section VI gives various results and comparisons. Finally, we conclude with a brief summary.

II. Estimating Globally Matched Wavelet Filters

Matched wavelet estimation for any signal is formulated as finding a closed form expression for extracting the compactly/infinitely supported wavelet which maximizes L error norm between the signal reconstructed at initial scaling subspace and successive lower wavelet subspace [1]. At an abstract level, our system use a set of trained wavelet filters matched to text and nontext classes. When a mixed document (having both text and nontext components) is passed through text matched filters, we get blacked-out regions in the detail (high pass) space corresponding to the text regions of the document and vice versa for the nontext matched wavelet filters. These blacked-out regions

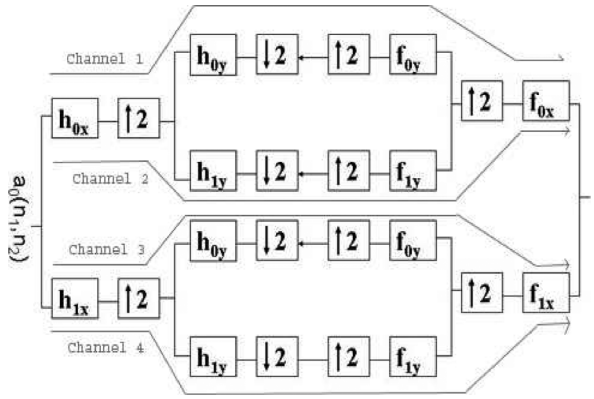


Fig. 2. Separable kernel filter bank.

in the output of text and nontext wavelet filters are used to classify various regions as either text or nontext.

In [1], an approach is proposed for estimating matched wavelets for a given image. It is further shown in [1] that estimated wavelets with separable kernel have higher peak signal-to-noise ratio (PSNR) for the same bit-rate as compared with standard 9/7 wavelet. In this section, we describe a technique for estimating a set of matched wavelets from a database of images. We term them as GMWs. These GMWs are used to generate feature vectors for segmentation. We discuss more about their implementation in subsequent subsections.

A. Matched Wavelets & Their Estimation

First, we briefly review the theory of matched wavelets with separable kernel as proposed in [1]. Consider a 2-D two-band wavelet system (with separable kernels) shown in Fig. 2. Here, x and y are the horizontal and vertical directions, the scaling filter in any direction is represented as (f_{0x}/f_{0y}), its dual is represented as (h_{0x}/h_{0y}), wavelet filter is represented as (f_{1x}/f_{1y}), and its dual is shown as (h_{1x}/h_{1y}). Further, boxes showing 2 with an upward or downward arrow represent upsampling or downsampling respectively of the signal by a factor of 2. The input to this system is a 2-D signal (an image for example) which is, for practical purposes, assumed to be continuous(being at the finest resolution) and the output of the system is another 2-D signal constructed from the sum of outputs of four channels shown in Fig. 2. The output of channel 1 is called approximation subspace or scaling subspace, whereas the output of the other three channels are called detail subspaces. This system is designed as biorthogonal wavelet system which means that it needs to satisfy following conditions for perfect reconstruction of the two- band filter bank (Fig. 2)

$$\mathrm{h}_1(n) = (-1)^n f_0(M-n) \tag{1}$$

$$\mathrm{f}_1(n) = (-1)^n h_0(M-n) \text{ where } M \text{ is any odd delay.} \tag{2}$$

The scaling function $\phi(t)$ and wavelet function $\psi(t)$ are governed by two-scale relations for the two-band wavelet system given in (7) and (8). Similar equations exist for estimating dual scaling function $\phi'(t)$ and dual wavelet function $\psi'(t)$.

The error between two signals is defined as

$$e(x) = a(x) - \hat{a}(x) \tag{3}$$

where $\mathrm{a(x)}$ is the continuous 2-D image signal and $\hat{a}(\mathrm{x})$ represents the 2-D image reconstructed from detail coefficients $d_{-1}(n)$ (sum of outputs of channel 2, 3, and 4 in Fig. 2) only.

Then, corresponding error energy is defined as

$$E = \int_{R^2} e^2(x)\,dx. \tag{4}$$

In order that the maximum input signal energy moves to scaling subspace, the energy E in the difference signal $e(x)$ should be maximized with respect to both x and y direction filters. It leads to a set of equations of the form (5) and (6). The algorithm to estimate compactly supported matched wavelet from given image with separable kernel is as below:

After fixing the filter size, say N, all rows of image are placed adjacent to each other to form a 1-D signal having variations in horizontal direction only a_{0x}. Corresponding to this 1-D signal, we estimate matched analysis wavelet filter h_{1x} using (5). Now, place all the columns of image below each other to form a 1-D signal having variations in vertical direction only a_{0y}. Corresponding to this 1-D signal, we estimate matched analysis wavelet filter h_{1y} using (6)

$$\sum_k h_{1x}(k)\left[\sum_m a_{0x}(2m+k)a_{0x}(2m+r)\right] = 0 \quad \text{for } r = 0,1,2,\ldots,j-1,j+1,\ldots,N-1 \tag{5}$$

$$\sum_k h_{1y}(k)\left[\sum_m a_{0y}(2m+k)a_{0y}(2m+r)\right] = 0 \quad \text{for } r = 0,1,2,\ldots,j-1,j+1,\ldots,N-1. \tag{6}$$

Here, the jth filter weight is kept constant to value 1. This leads to a closed form expression where the bracketed term looks like deterministic autocorrelation function of decimated input signal. These are a set of $\mathrm{N}-1$ linear equations in filter weights that can be solved simultaneously. The solution gives corresponding weights of dual wavelet filters h_{1x} and h_{1y}, i.e., the analysis high-pass filters. From it, other filters (analysis low pass, synthesis high pass, and synthesis high pass) are obtained using finite impulse response (FIR) perfect reconstruction bi-orthogonal filter bank design. It is worth noticing here that all designed wavelet filters here are pass filters. The idea to estimate an analysis wavelet filter is similar to a sharpening filter used in image enhancement. The resulting filters in (5) and (6) are observed to be high-pass filters, which is in conformity with the result of the sharpening filter in image enhancement.

Results obtained by DWT of an image using matched wavelet and standard Haar Wavelets are shown in Fig. 3. The figure shows a crisper image a_1 with matched wavelet than the a_1 image with Haar wavelet which means that more energy has passed through this subspace for matched wavelets. Moreover, d_2 image with matched wavelet is noisier than corresponding image with the Haar wavelet which means lower energy in this subspace for matched wavelet transform. Other images d_1 and d_3 are hard to perceive directly even when expanded.The result shows lower energy in detail spaces (d_1, d_2, d_3) and higher energy in the approximation space in case of matched wavelet DWT as compared to standard wavelet transforms.

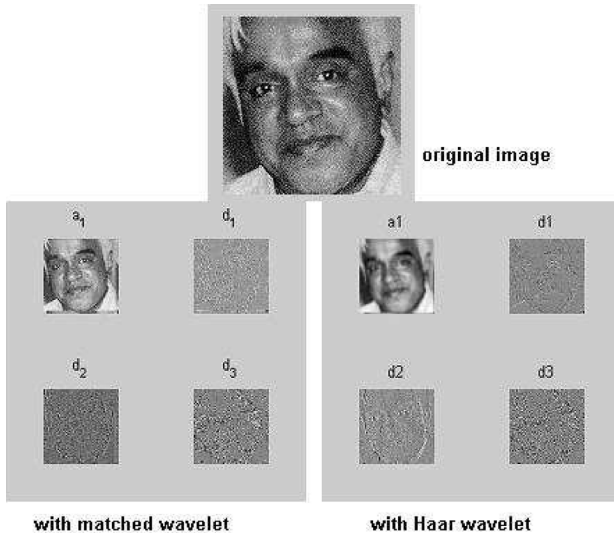


Fig. 3. (Top) DWT of an image with (bottom left) its matched wavelet and with (bottom right) traditional Haar wavelet.

B. Globally Matched Wavelets (GMWs)

Matched wavelet estimated from a given image is the optimized solution for that image only and, thus, may not represent the complete class of images. Further, for our problem, nontext regions and text regions are expected to have a variety of textural properties. For example, nontext region in an image could correspond to a landscape, background, traffic scene, sketch or computer generated texture. Hence, we need a methodology for designing a set of global filters suited for a variety of scenes and text regions. The estimation process involves use of a large number of examples representing different types of nontext and text regions. In our implementation, we have used 3000 examples of pure text and pure nontext regions. For each example, we have computed the matched wavelet filters. We found optimal size of filters to be between 4 and 8. Then, we converted the matched wavelet filter coefficient into corresponding scaling functions using two-scale relation given by

$$\Phi_1(x) = \sum_n f_{0x}(n)\sqrt{2}\phi_1(2x-n)\forall n \in \mathrm{Z} \tag{7}$$

$$\Psi_1(x) = \sum_n f_{1x}(n)\sqrt{2}\phi_1(2x-n)\forall n \in \mathrm{Z}. \tag{8}$$

Since, for our problem, nontext regions and text regions are expected to have a variety of textural properties, in terms of text fonts, image origin, etc., the simple average of the Φ functions over the two classes will not be the representative of the two classes. We establish similarity amongst the shape of Φ functions by computing the Euclidean distance between their sampled values. Depending upon this similarity measure, Φ functions are grouped into homogeneous clusters using a modified version of iso-data algorithm proposed in [25]. Homogeneous clusters are those clusters which contain Φ functions matched to regions of a single type only (text or nontext). In the present implementation of the iso-data algorithm, we start with two clusters and increase the number of clusters if

$$\exists\, i \text{ s.t. } r_i > \mathrm{thres}_1 \text{ or } r_i < \mathrm{thres}_2$$

TABLE I
HIGH-PASS GMW FILTERS

C 1	hx1	0.1083	-0.4697	0.7342	-0.4635	0.1017	0.0
	hy1	-0.0754	-0.3524	0.8617	-0.3463	-0.0800	0.0
C 2	hx1	0.0498	-0.4335	0.7837	-0.4353	0.0471	0.0
	hy1	0.0093	-0.4007	0.8132	-0.4171	0.0054	0.0
C 3	hx1	-0.1597	-0.2809	0.8906	-0.2730	-0.1595	0.0
	hy1	-0.1842	-0.2537	0.8941	-0.2528	-0.1892	0.0
C 4	hx1	0.0676	-0.4463	0.7619	-0.4506	0.0696	0.0
	hy1	0.0631	-0.4319	0.7706	-0.4566	0.0588	0.0
C 5	hx1	0.1260	-0.4487	0.6981	-0.5101	0.1357	0.0
	hy1	0.1500	-0.4849	0.6939	-0.4912	0.1333	0.0
C 6	hx1	-0.1853	0.0789	0.7674	-0.5975	-0.0637	0.0
	hy1	-0.2302	0.1152	0.7594	-0.5898	-0.0549	0.0

where $r_i = C_{1i}/C_{2i}$, $C_{ci} = \text{number of data points in cluster i which are coming from class c}$ and thres_1 and thres_2 are appropriately chosen thresholds.

Using iso-data clustering ensures that final clusters are dominated by either nontext or text samples. Thus, they are expected to be representative of either the text or nontext classes. For each cluster, globally matched wavelet filter coefficients are computed by finding their respective cluster centers. Being cluster centers, GMWs are the measure of average shape given by following formula:

$$\text{GK}_k(m) = \frac{1}{n_k}\sum_{i}^{n_k} X_i^k(m) 0 \le m \le N_\Phi \tag{9}$$

where $GX_k(\text{m})$ is the mth element of kth GMW. $X_i^k(n)$ represents the nth coefficient of the kth image of the cluster. Next, we discuss a GMW-based text detection scheme.

III. LOCATING TEXT FROM ARBITRARY BACKGROUNDS

Because of extensive training GMW-based algorithm can detect text regions in a variety of situations of text deformations, different fonts and scales. The following subsections give the detail algorithm for this purpose.

Algorithm

GMWs are the characteristic wavelets of pure nontext and text classes. We have GMW filters corresponding to the GMWs obtained in the previous section using the two scale relation. Table I shows the 12 (six each in x and y direction) high-pass analysis GMW filters obtained.

To locate text regions in an image, $\text{I}(n_1, n_2)$, pass the image through wavelet analysis filter bank (Fig. 2), with each of six GMW filter sets to obtain six transformed images. If we pass a text image through the filter bank consisting of GMW filters of text, then we get minimum energy in the detail subspace. On the other hand, if a pure picture (no text involved) is passed through this filter bank, then we get considerable energy going into detail subspace.

In order to enhance this information, we use standard deviation (SD), which is nothing but an estimate of local energy of each pixel. It is followed by Gaussian filtering to minimize effect of isolated noise in the transformed images [2]. We calculate the local standard deviation of each pixel in transformed image using

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

where $w(=9)$ is the window size and $R = w * w$, while $\bar{h}_{k_i}(x,y)$ is the mean around the (x,y)th pixel and $h_{k_i}(x,y)$ is the filtered image. Following this, the Gaussian filtering is done using

$$\text{Feat}_{k_i}(x,y) = \frac{1}{G^2}\sum_{(m,n)\in G_{x,y}}\left|\text{eng}_{k_i}(m,n)\right|$$

where $G_{x,y}$ is the $G*G$ Gaussian smoothing window and Feat_{k_i} is the final transformed 6-D feature image. Thus, at the end of it we have a stack of six transformed images. We define our feature vector at each pixel as

$$\bar{f}(x,y) = [f_1(x,y), f_2(x,y), f_3(x,y), f_4(x,y), f_5(x,y), f_6(x,y)]$$

where $f_i(\text{x},\text{y})$ is the value of $\text{pixel}(\text{x},\text{y})$ of transformed image i.

A. Classification

Having obtained the feature vectors, for classification of pixels as either text or nontext, we have used two different schemes: 1) segregation of test image pixels into two classes using k-means algorithm [24] followed by class assignment based on result of SVM trained on the ratio of cluster centers and 2) Fisher classifiers.

In the first approach, we cluster the feature vectors into two classes for a given image. For labeling the two classes, we find component-wise ratio of the cluster centers CC_1, CC_2. Now, this 6-D ratio vector CC_1/CC_2 indicate distinctive feature of the image region corresponding to the cluster center CC_1 modulo overall intensity variation. For a given set of training images (150), we use this ratio to train a SVM [16] for recognizing text region in an unknown image. For test images, we have clustered the feature vectors computed at each pixel and then classified the ratio of cluster centers to identify the text region, using SVM.

In the second approach, the Fisher classifier finds the optimal projection direction by maximizing the ratio of between class scatter to within class scatter which benefits the classification [3]. For a feature vector $\underline{X}$, the Fisher classifier projects $\underline{X}$ onto one dimension Y in direction $\underline{W}$ using

$$Y = \underline{W}^T\underline{X}. \tag{10}$$

Let $Y1$ and $Y2$ be the projections of two classes on to the optimal projection direction W_0 and let $E[Y1]$ and $E[Y2]$ be the means of $Y1$ and $Y2$, respectively. Suppose $E[Y1] > E[Y2]$; then the decision can be made as

$$C(\underline{X}) = \begin{cases} \text{class1}, & \text{if } Y > \frac{(E[Y_1]+E[Y_2])}{2} \\ \text{class2}, & \text{otherwise} \end{cases}. \tag{11}$$

Because of reduced dimensionality, the Fisher classifier is very easy to train, faster for classification and does not suffer from overtraining problems. In Section V, we have presented and compared the results obtained by using these approaches for classification.

IV. Segmentation of Document Images

In this paper, we refer to natural images (photographs for example) as scene images. The scene images that we consider, contain some written or embedded text and everything else is nontext region. In the document image, however, we consider three components: 1) text, 2) picture, and 3) background. Backgrounds are continuous tone low frequency regions with dull features although mixed with noise. Images are continuous tone regions falling in between text and background. Thus, for document images we have extended our work described in the last section (text location in general image) to segmentation of document images into three classes, *viz.* text, picture, and background. We have used the same feature vectors and classified them into three classes. For classification, we have used Fisher classifiers, first because of the advantages like ease of training (because of projecting the data on one dimension) and time efficiency. Moreover, the results of the Fisher classifier fit naturally into our MRF postprocessing step as explained in the next section [3].

The Fisher classifier is often used for two-class classification problems. Although it can be extended to multiclass classification (three classes in our case), yet the classification accuracy decreases due to the overlap between neighboring classes. Thus, we need to make some modifications to the Fisher classifiers (explained in the last section) to apply them in this case [3]. We use three Fisher classifiers, each optimized for a two-class classification problem (text/picture, picture/background, and background/text). Each classifier outputs a confidence in the classification and the final decision is made by fusing the outputs of all three classifiers.

A. Classification Confidence

Among the three classifiers, classifier 1 refers to picture and background, classifier 2 refers to picture and text, and classifier 3 refers to background and text. From the previous section, we have algorithm to evaluate the $\underline{W}_0$ values for all the three classifiers. We use these $\underline{W}_0$ values along with the groundtruthed images to find out the distribution of Y for both the classes of each classifier. For example, in our case, we obtain the distribution of Y values corresponding to picture and background in classifier 1 (represented as Y_{1a} and Y_{1c}), picture and text in classifier 2 (represented as Y_{2a} and Y_{2b}), and background and text in classifier 3 (represented as Y_{3c} and Y_{3b}), respectively. After normalizing these distributions, we fit the higher order Gaussian functions to these distributions using curve-fitting. This provides us with the approximate pdfs of these projections. Let us represent these distributions by $f_Y(y)$. The plot of these distribution functions for classifier 1 is shown in Fig. 4.

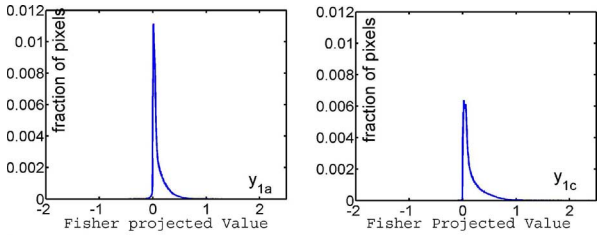


Fig. 4. Distribution of Y for image and background as obtained from classifier 1. Similar distributions are obtained for classifiers 2 and 3.

The classification confidence $C_{i,j}$ of class i using classifier j is defined as the equation shown at the bottom of the page, where i is the class label and j represents the trained classifiers. If a classifier is trained to classes 1 and 2, its output is not applicable to estimating the classification confidence of class 3. Therefore, $C_{3,j} = 0$. In other words, for a class i, $C_{i,j} \in [0, 1]$ for the two applicable classifiers and $C_{i,j} = 0$ for the third classifier. The final classification confidence is defined as

$$C_i = \frac{1}{2} \sum_{j=1}^{3} C_{i,j}. \tag{12}$$

Note that $C_i \in [0, 1]$, i =1, 2, 3. However, C_i is not a good estimate of the a posteriori probability since $\sum_{i=1}^{3} C_i = 1.5$ instead of 1. We can take C_i as an estimate of a nondecreasing function of the a posteriori probability, which is a kind of generalized classification confidence [17]. Thus, C_i represents the probability of a pixel belonging to class $i(= 1, 2, 3)$. The classification confidence maps for picture, text, and background for a sample image are shown in Fig. 15 (see Results section).

V. MRF Postprocessing for Document Image Segmentation

Using the same text extraction features for document image segmentation may lead to overlapping in the feature space. This is especially true for the picture and background classes because of lack of hard distinction between the textures of these two classes. We deal with this problem by exploiting the contextual information around each pixel. A similar approach has been used recently in [3] to refine the results of segmenting the handwritten text, printed text and noise in the document image. Results for the document image segmentation of previous section show that misclassification happens either in form of occurrence of certain isolated clusters of another class in a given class or at the boundaries of the different classes as indicated by Fig. 5. Removing this misclassification is equivalent to making the classification smoother. In this section, we present

$$C_{i,j} = \begin{cases} \frac{f_Y\left(\frac{y}{X \in \text{class}i}\right)}{f_Y\left(\frac{y}{X \in \text{class}i}\right) + f_Y\left(\frac{y}{X \in \text{anotherclass}}\right)}, & \text{if } i \text{ is applicable for classifier } j \\ 0, & \text{Otherwise} \end{cases}$$

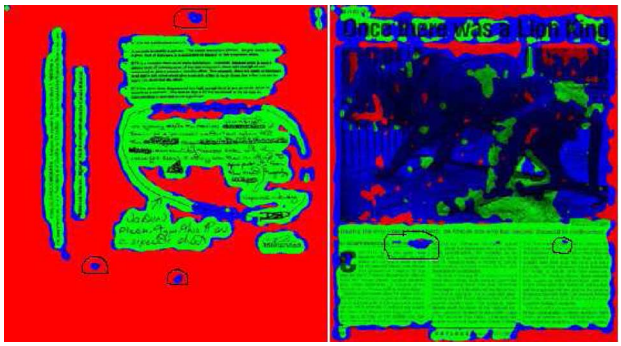


Fig. 5. Document Segmentation results obtained for two sample images from previous section. Images show that the misclassification occurs either at the class boundaries or because of the presence of small isolated clusters (some of these are circled in the images above).

MRF-based approach for classification smoothening using context around pixels.

A. MRF: Background

The problem of correcting the misclassification belongs to very general class of problems in vision and can be formulated in terms of energy minimization [11]. Every pixel p must be assigned a label in the set $L = \{\text{text, picture, background}\}$. F refers to a particular labeling of the pixels and f_p refers to the value of the label of a particular pixel. We consider the first order MRF model. This simplifies the energy function to the following form:

$$E(f) = \sum_{\{p,q\} \in N} V_{p,q}(f_p, f_q) + \sum_{p \in P} D_p(f_p) \qquad (13)$$

where N is the set of interacting pair of pixels. We consider N to be eight neighborhood of a pixel. The first and second terms in the above equation are referred to as E_{smooth} (interaction energy) and E_{data} (energy corresponding to the data term) in the literature [11]. D_p (data term) is the measure of how well a label f_p fits the pixel p given the observed data. Thus, in our case, classification confidence found in the previous section is the intuitive choice of D_p [3].

E_{smooth} makes F smooth everywhere. Although smoothing is required for removing misclassification, yet oversmoothing can lead to poor results at the object boundaries. Energy functions that do not have this problem are referred to as *discontinuity preserving*. One such example of discontinuity preserving energy function is Potts interaction penalty [11]. Geman *et al.* [23] were the first to use this model in computer vision. This is in some sense the simplest discontinuity preserving model and it is especially useful when the number of labels is small [11].

In our case, we show the significant improvement in the results by using this simple model. In this model, the discontinuities between any pair of labels is penalized equally. Mathematically, the expression for the Potts interaction penalty is

$$V_{p,q}(f_p, f_q) = \lambda . T(f_p \neq f_q) \qquad (14)$$

where λ is a constant and T is 1 when $f_p \neq f_q$ else it is 0. The value of λ controls the amount of smoothening done by the energy function. We have shown in the results section that as we increase λ we loose on the discontinuity preserving nature of the function.

Fig. 6. (a) Text on T-shirt. (b)–(c) Results using cluster center-based technique and Fisher classifiers.

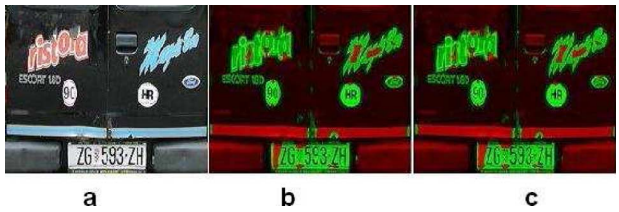


Fig. 7. (a) Text on an automobile-1. (b)–(c) Results using cluster center-based technique and Fisher classifiers.

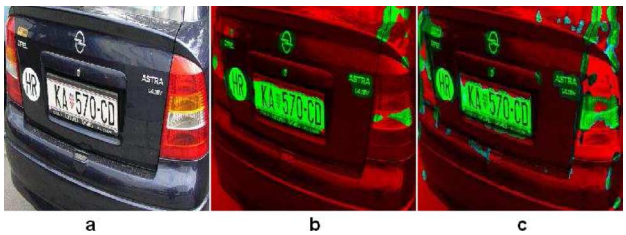


Fig. 8. (a) Text on an automobile-2. (b)–(c) Results using cluster center-based technique and Fisher classifiers.

Fig. 9. (a) Printed text with different fonts sizes. (b)–(c) Results using cluster center-based technique and Fisher classifiers.

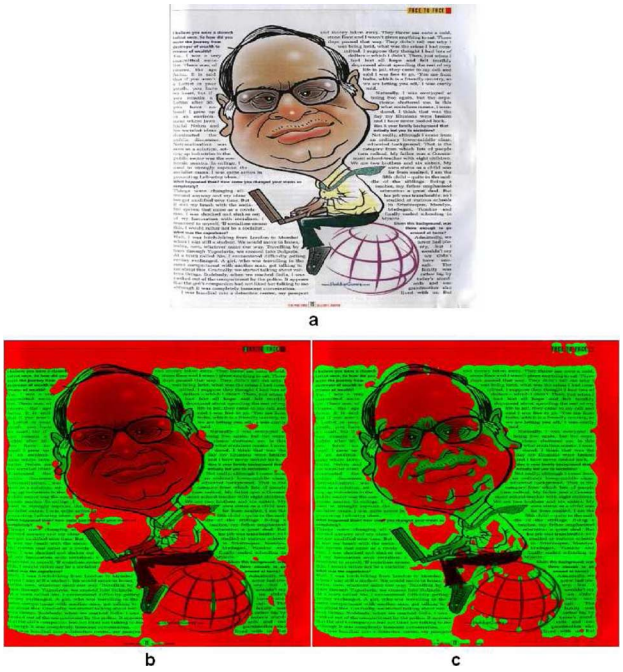


Fig. 10. (a) Printed text with irregular boundaries. (b)–(c) Results using cluster center-based technique and Fisher classifiers.

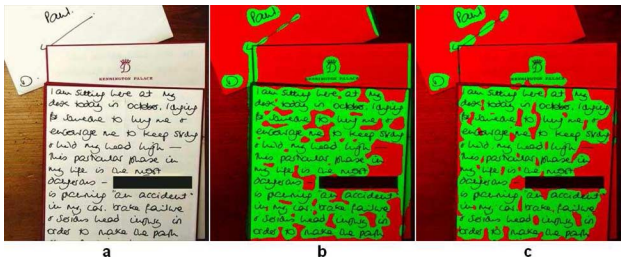


Fig. 11. (a) Handwritten text. (b)–(c) Results using cluster center-based technique and Fisher classifiers.

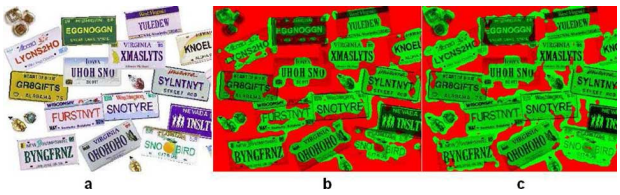


Fig. 12. (a) Randomly oriented text. (b)–(c) Results using cluster center-based technique and Fisher classifiers.

B. MRF: Optimization

The major difficulty with MRF energy minimization lies in the enormous computational costs. Even with the simple Potts energy model computing the global minima is NP-hard [11]. For energy minimization, we use the α–*expansion* algorithm proposed in [11] and [12]. For Potts interaction penalty, this algorithm gives a solution that is within a factor of 2 of the global minimum. In this algorithm, we model the image pixels as weighted graph depending on their labels and use min graph cut algorithm to associate each pixel to a label that results in minimization of the energy.

TABLE II
PRECISION RATE OBTAINED USING THE TWO CLASSIFICATION SCHEMES FOR THE PREVIOUS IMAGES

Image No.	Cluster center based classification	Fisher Classifier
Fig. 6	51.42%	49.90%
Fig. 7	68.01%	63.51%
Fig. 8	51.96%	41.2%
Fig. 9	83.05%	79.15%
Fig. 10	86.44%	73.98%
Fig. 11	88.49%	86.61%
Fig. 12	86.73%	79.62%

TABLE III
RECALL RATE OBTAINED USING THE TWO CLASSIFICATION SCHEMES FOR THE PREVIOUS IMAGES

Image No.	Cluster center based classification	Fisher Classifier
Fig. 6	83.3%	83.2%
Fig. 7	88.4%	84.5%
Fig. 8	81.7%	79%
Fig. 9	79.3%	79%
Fig. 10	82.5%	74%
Fig. 11	83%	78%
Fig. 12	60%	59%

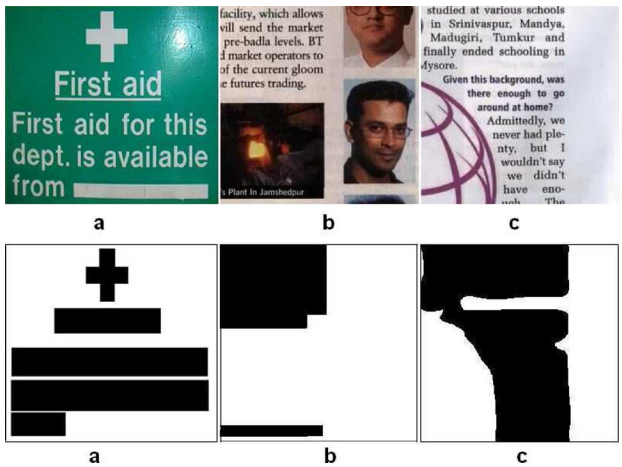


Fig. 13. (a)-(c) Sample images; (d)-(f) their corresponding groundtruths.

Inputs to the algorithm are the classification confidence maps (for image, text, background) and labelings (initial results) that we obtained in the last section using this classification confidence maps. Using the initial labelings, algorithm evaluates the interaction energy (E_{smooth}) and minimize the total energy ($E_{\text{smooth}} + E_{\text{data}}$) to obtain a new labeling. This step is repeated till no further minimization is possible, finally leaving the resulting optimized labeling.

VI. RESULTS AND DISCUSSION

We analyzed several types of images so as to demonstrate the performance of our algorithm. The images are taken from scanned pages and different websites. Some of these results are discussed in this section. First, we have discussed the results of text extraction, then results of document image segmentation have been discussed.

A. Text Extraction in General Image

In this section, we have presented the results obtained using our text extraction algorithm. We have shown the results ob-

TABLE IV
AVERAGE PERCENTAGE ACCURACY OBTAINED USING KUNDU'S METHOD AND OUR APPROACH ON A SET OF 33 IMAGES

	Cluster center based classification	Fisher Classifier	M-Band
Precision rate	76.8%	73.5%	71%
Recall rate	73.7%	67.7%	58.1%

tained by the two schemes: cluster center-based classification and Fisher classifier. We have extensively tested our algorithm, to prove its efficacy and robustness over different and diverse type of images, which includes images with handwritten text, images with printed text having different fonts sizes and shapes, billboard images, license plates, etc. In the results, regions painted **green** indicate the text area and regions painted **red** are nontext areas.

Fig. 6 is an example image taken from the ICDAR data set [20]. In this case, text regions overlap with nontext regions. Figs. 7 and 8 are the typical examples of text at the back of a vehicle. These images represents the example images where image deformation is the result of relative motion between camera and vehicle. Fig. 9 is an example where an image has multiple columns and many different fonts. Moreover, in this image, text pixels have different gray scale values. This image has been taken from the test dataset used by [2]. We have not used any *a priori* information regarding the font size etc, but this example shows that the algorithm has discriminating power between text and nontext part. Fig. 10 is the case where text boundaries are nonconvex, overlapping and irregular. Such situations commonly arise in the newspaper article texts. Along with the printed text, our algorithm also performs well on handwritten documents. We had tested the systems on many handwritten images and one such example is shown in Fig. 11. Our algorithm is based on the difference between texture of text and nontext areas. Thus, it also works successfully in the cases where text is randomly oriented at different angles as shown in Fig. 12.

From Figs. 6–12(b) and (c), we can see that the two classification schemes perform almost equally well. We further performed some experiments to strengthen this point. Text extraction is a one class classification problem. The precision and recall accuracy of this classification is computed as

$$\text{Recall rate} = \frac{\#\text{ of text pixels correctly identified}}{\#\text{ of text pixels in ground truth}} \tag{15}$$

$$\text{Precision rate} = \frac{\#\text{ of text pixels correctly identified}}{\#\text{ of text pixels detected}}. \tag{16}$$

Tables II and III show the precision and recall rates of two classifiers, respectively, for the above seven images.

This quantity has been evaluated against hand-picked groundtruth images. Fig. 13 shows some sample images and their corresponding groundtruth images made by us. In groundtruth images, regions marked black are text and regions marked white are nontext. Being a pixel level accuracy method against handpicked groundtruth, it is not an effective way to measure absolute performance of the algorithms, but it provides a good measure for comparing the two algorithms.

To give an average estimate of the performance of the classifiers and to compare the results with M-band algorithm [2], we further took a set of 33 general images. On this we found the accuracy using our methods and Kundu's M-band algorithm. The results for mean accuracy figures are presented in Table IV. The results show that our method with cluster center-based classification gives the best performance.

TABLE V
AVERAGE TIME TAKEN BY KUNDU'S METHOD AND OUR APPROACH ON A SET OF 33 IMAGES

Cluster center based classification	Fisher Classifier	M-Band
158 secs	139 secs	286 secs

We implemented our text extraction system and Kundu's M-Band algorithm in Matlab. In terms of the running time, our algorithm performs very well as compared to the M-Band algorithm. Table V shows the average time taken by each method on the same set of 33 images.

The actual time taken in each case depends on the image size. Fisher performs faster because it projects the 6-D feature vector to the 1-D space. The above results indicate that our algorithm is almost twice as fast as M-band algorithm. This improvement in the performance is primarily because of the reduced dimensionality of the feature vectors in our method. We have used very generic dataset to train our filters. However, if we develop application specific matched wavelet filters (e.g., license plate detection, printed documents, etc.), then this dimensionality can be further reduced, leading to even better performance.

B. Document Image Segmentation

In this section, we present the results for document image segmentation. Fig. 14 shows the effect of increasing λ on the postprocessing results. As the value of λ increases, number of discontinuities in the resulting image decreases as higher values of the λ leads to oversmoothing. Oversmoothing is not always desirable, specially in the cases where we have very small chunks of different classes. Such cases will demand to preserve more discontinuities. We found that results obtained with λ equal to 1 provide significant improvement in the results and it works well in most of the cases.

Here, we present the results for some of the most typical cases before and after post processing. This also aids in the comparison of the results. Results for the post processing were obtained using parameter value of 1 in the interaction term of MRF model. Fig. 15 shows the classification confidence images for the picture, text and background classes. These images lead to the resulting document image segmentation and also act as data term for the post processing step. In the classification confidence (CC) map images, whiteness of a region indicates the higher probability. Thus, we can see that on the places of presence of text, text CC map has more whiteness in comparison to picture and background CC maps and vice versa for the background and picture CC maps.

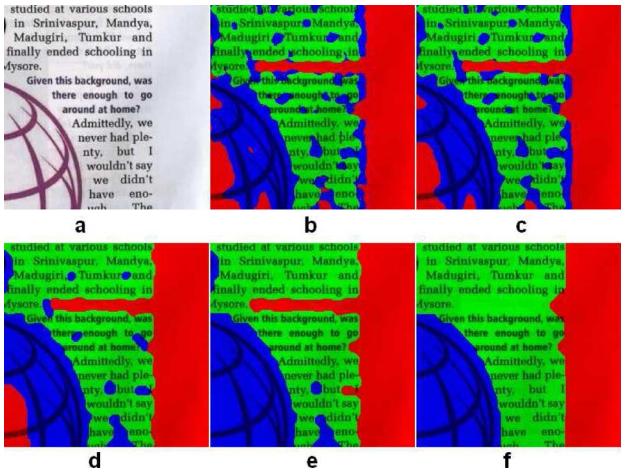


Fig. 14. (a) Original image. (b) Result without post processing. (c)–(f) Results with post processing, values of λ being 0.1, 1, 2, 10, resp. Note that, as the value of λ increases, the resulting image becomes smoother with lesser discontinuities.

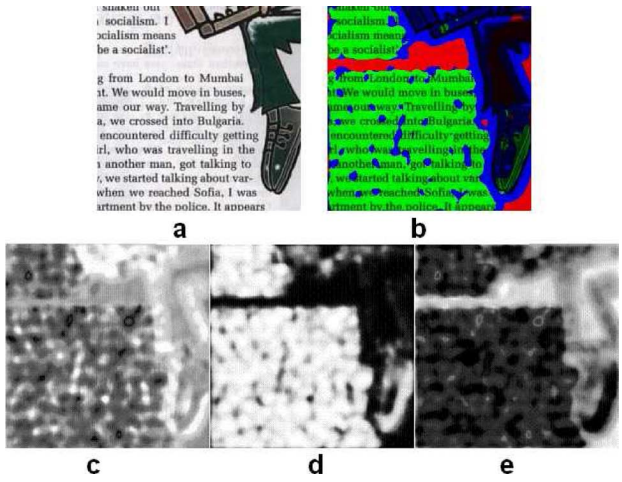


Fig. 15. (c)–(e) Classification confidence maps for picture, text, and background classes, respectively, obtained from the input image (a) and the final resulting segmented image; (b) is obtained using these classification confidence maps.

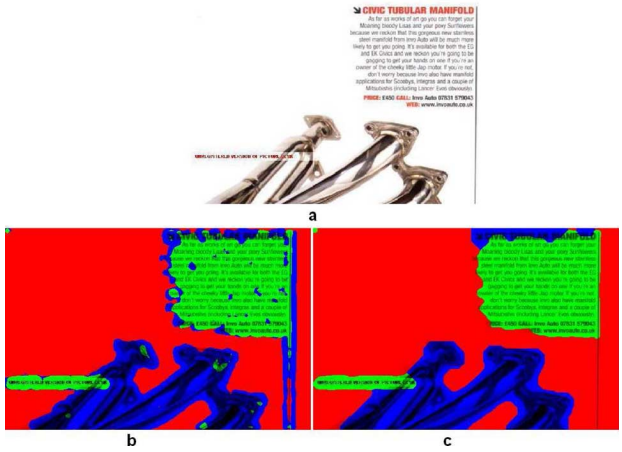


Fig. 16. Example of a general document image: (a) shows the original image, (b) is the image without postprocessing, and (c) is the final result.

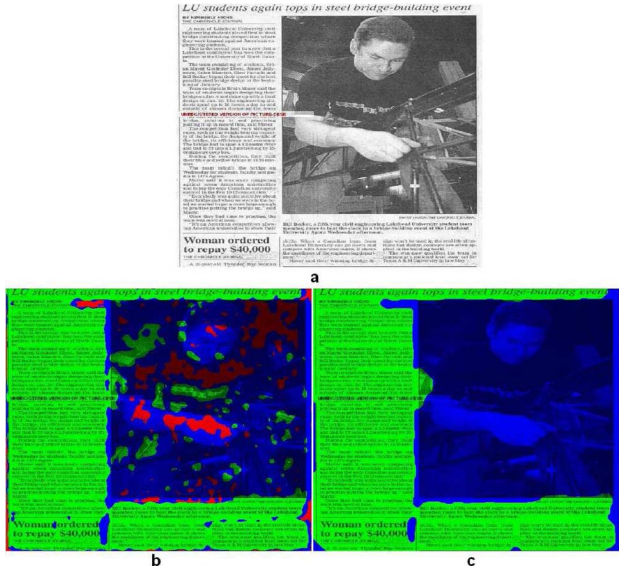


Fig. 17. Image with one of the three classes missing: (a) shows the original image, (b) is the image without postprocessing, and (c) is the final result.

Now we present some of the interesting results for document image segmentation. Here regions painted **green** represent text, regions painted **blue** represent picture and regions painted **red** represent background. Fig. 16 shows the result when text part is present in small chunks and text part is randomly scattered in different regions. Fig. 17 presents the case when one of the three classes is completely missing. Such a case is not uncommon in most document images, where either there is no picture or background. Fig. 18 is one where text and picture boundaries are irregular. Fig. 19 shows the result when text and picture parts are present in small chunks and are randomly scattered in different regions of the document. In all the above examples, the document image considered had printed text. We also tested our system on the document images with handwritten text. Fig. 20 presents such a case.

Along with above sample images, we have tested our system extensively on many other document images. To illustrate improvements in the results after post processing, we measured accuracy (using the same accuracy measure defined in the last subsection) of the results obtained against hand-picked groundtruth. Table VI shows the percentage accuracy obtained for each of the images shown above. Also, it shows the overall mean improvements in the results obtained over a set of 27 images.

This approach for segmentation is fast and robust. The later part of MRF post processing has been implemented in C. α*–expansion* algorithm completes the optimization in 2–3 iteration steps and takes few seconds (2–3) for this purpose. 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 in the image. In order to compare our system with other nonwavelet-based methods, we followed the same performance evaluation method as used in ICDAR 2005 page segmentation competition [19]. It is based on counting number of matches between the entities (text and pictures) detected by the algorithm

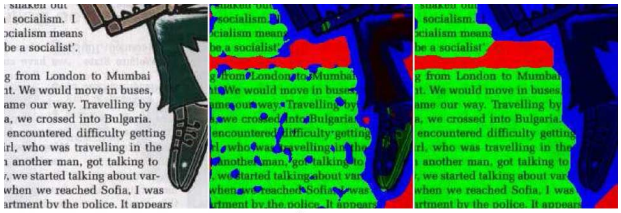


Fig. 18. Image with text and image boundaries not well defined: (a) shows the original image, (b) is the image without postprocessing, and (c) is the final result.

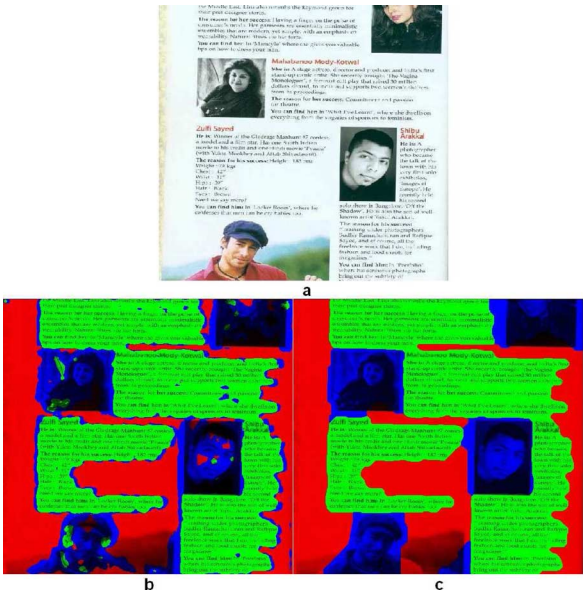


Fig. 19. Document has several small chunks of image and text parts: (a) shows the original image, (b) is the image without postprocessing, and (c) is the final result.

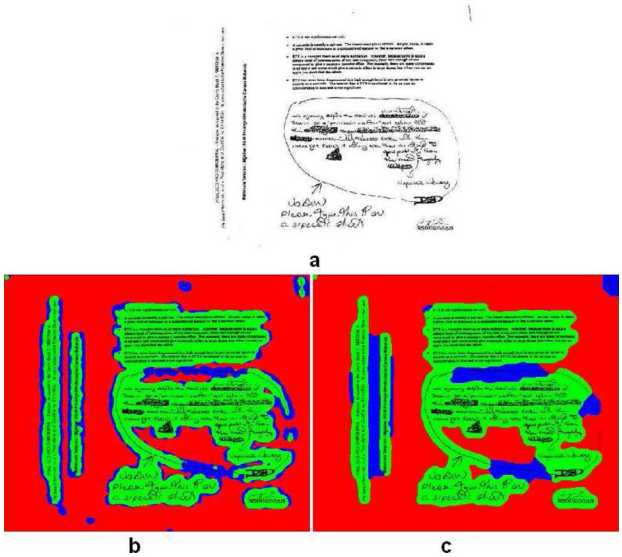


Fig. 20. Document image with handwritten text: (a) shows the original image, (b) is the image without postprocessing, and (c) is the final result.

and entities in the groundtruth. We use a global MatchScore table for all entities whose values are calculated according to the intersection of ON pixel sets of the results and the groundtruth. Detailed information about this method can be found in [4] and [19]. Here, we reproduce the formulas used in these papers for calculating two performance measures: edit cost index (ECI) and segmentation metric (SM). Other parameter used in the evaluation are same as used in aforementioned papers

$$\mathrm{ECI} = 1 - \frac{2 \times \mathrm{one2one}}{N + M} \tag{17}$$

$$\mathrm{SM} = \frac{\sum_{i \in I} N_i \times \mathrm{EDM}_i}{\sum_{i \in I} N_i} \tag{18}$$

where one2one and entity detection metric (EDM) have similar meanings as in [19] and [4]. N and M are the number of entities in groundtruth image and detected image, respectively. The dataset chosen is similar to the ICDAR 2005 dataset (but not the same since original dataset is no longer available). Groundtruths of texts and pictures were prepared separately for these images. All in all 12 text groundtruths and five pictures, groundtruths were prepared. Table VII shows entity level accuracies obtained for each of these. Table VIII shows the computed ECI and SM values scored by our system and M-band segmentation and compares them with some nonwavelet-based methods published in [19].

In Table VIII, SM-text and SM-average values are the same because it is a text classifier, and, therefore, it is tested only for text entities. The ECI value for last four methods in this table are not published. However, its value, which is an estimate of normalized post editing cost, has been published for some general graphics recognition systems in [4, p. 18] .Comparing with these results, the ECI of our method (0.3031) is nearest to the best ECI (0.2983) reported. The results also show that our method based on matched wavelet followed by MRF postprocessing gives approximately 7%–9% better performance than other wavelet and nonwavelet-based methods for text extraction. Although the results reflect that average SM value in case of our algorithm is approximately 17% better than the best nonwavelet-based method, yet this value might not be exact since the number of entities used in ICDAR 2005 competition were 5 (text, graphics, line arts, separator and noise), whereas we have defined only three entities in our experiment [text, pictures and background (or noise)]. However, since we have defined pictures as an umbrella class to include all nontext and nonbackground entities, the estimate of performance using three entities does not differ much from five entities estimate and we surmise that if the experiments conducted in [19] were repeated with three entities they would give similar average results.

VII. Conclusion

In this paper, we have presented a novel technique for locating the text part based on textural attributes using GMWs. 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. Compared to other existing methods [2], [6], the dimensionality, and, so, the computation of the feature space, is considerably reduced. We have applied our algorithm on several structured and highly unstructured images with complex backgrounds and obtained encouraging results. The results indicate that GMWs have the efficacy to discriminate between textures, and can be effectively applied for text identification. The method can be

TABLE VI
PRECISION RATE BEFORE AND AFTER POSTPROCESSING

Image No.	Before postprocessing	After postprocessing
Fig. 16	90.7%	93.8%
Fig. 17	74.4%	84.4%
Fig. 18	83.7%	90.1%
Fig. 19	79.6%	84.8%
Fig. 20	86.2%	87.5%
mean over a set of 27 images	81.4%	84.8%

TABLE VII
ENTITY LEVEL PERFORMANCE OF INDIVIDUAL IMAGES

Name	N	M	one2one	Detection	Recognition	EDM
For text						
test01.jpg	9	9	8	0.89	0.89	0.89
test02.jpg	6	9	5	0.83	0.56	0.67
test03.jpg	6	6	6	1	1	1
test04.jpg	10	15	8	0.8	0.54	0.64
test05.jpg	3	3	2	0.66	0.66	0.66
test06.jpg	2	2	1	0.5	0.5	0.5
test07.jpg	12	16	4	0.33	0.25	0.28
test08.jpg	3	2	2	0.66	0.66	0.66
test09.jpg	2	2	1	0.5	0.5	0.5
test10.jpg	5	8	2	0.4	0.25	0.31
test11.jpg	2	3	2	1	0.66	0.80
test12.jpg	2	3	2	1	0.66	0.80
For pictures						
test13.jpg	2	4	2	1	0.5	0.67
test14.jpg	1	3	1	1	0.34	0.5
test15.jpg	1	1	0	0	0	0
test16.jpg	1	6	1	1	0.17	0.3
test17.jpg	4	5	2	0.5	0.4	0.44

TABLE VIII
EVALUATION RESULTS FOR ALL ENTITIES (EDM VALUES AVERAGED OVER ALL IMAGES)

Algorithm	**SM for text only**	**SM average**	**ECI**
Matched Wavelet	62.34	59.86	0.30
M-band	51.92	51.92	0.34
BESUS method	29.62	27.74	NA
Oce method	31.36	30.69	NA
Tsinghua method	46.64	37.06	NA
Tsinghua method	53.22	42.12	NA

tuned for specific classes of texts depending on the dataset used for estimating GMWs. As an example, we are implementing a real time system for automatic license plate detection. For such an application, the dimensionality reduces even further. This leads to significant large improvement in the running time. Hence, such a system can be used for real time applications, where we have both computational and timing constraints.

Another important contribution of our work is the introduction of the segmentation of document image as a three class problem, which provides a new vision for automatic document image understanding. Thus, it opens up new research directions in the document image analysis which can provide an avenue for the development of many new applications based on document image analysis.

REFERENCES

[1] A. Gupta, S. D. Joshi, and S. Prasad, "A new approach for estimation of wavelets with separable and non-separable kernel from a given image," *IEEE Trans. Signal Process.*, to be published.

[2] M. Acharyya and M. K. Kundu, "Multiscale segmentation of document images using M -band wavelets," in *Proc. 9th Int. Conf. Comput. Analysis Images and Patterns*, pp. 510–517, ISBN 3-540-42513-6.

[3] Y. Zheng, H. Li, and D. Doermann, "Machine printed text and handwriting identification in noisy document images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 3, pp. 337–353, Mar. 2004.

[4] I. Phillips and A. Chhabra, "Empirical performance evaluation of graphics recognition systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 9, pp. 849–870, Sep. 1999.

[5] B. A. Yanikoglu and L. Vincent, "Pink Panther: A complete environment for ground truthing and benchmarking document page segmentation," *IEEE Trans. Pattern Recognit.*, vol. 31, no. 9, pp. 1191–1204, Sep. 1994.

[6] K. Etemad and R. Chellappa, "Separability based tree structured local basis selection for texture classification," in *Proc. Int. Conf. Image Processing*, 1994, vol. 3, pp. 441–445.

[7] J. Li and R. M. Gray, "Context based multi-scale classification of document images using wavelet coefficient distribution," *IEEE Trans. Image Process.*, vol. 9, no. 9, pp. 1604–1616, Sep. 2000.

[8] C. J. Park, K. A. Moon, O. W. Geun, and H. M. Choi, "An efficient extraction of character string positions using morphological operator," in *Proc. IEEE Int. Conf. Systems, Man, Cybernetics*, 2000, vol. 3, 8–11, pp. 1616–1620.

[9] Z. Yu, K. Karu, and A. K. Jain, "Locating text in complex color images," in *Proc. 3rd Int. Conf. Document Analysis and Recognition*, 1995, vol. 1, 14–16, pp. 146–149.

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

[11] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2004.

[12] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 147–159, Feb. 2004.

[13] A. K. Das and B. ChandaD, "Segmentation of text and graphics in document image: A morphological approach," in *Proc. Inf. Conf. Computational Linguistics, Speech and Document Processing*, Calcutta, India, Dec. 1998, pp. A50–A56.

[14] M. Chen and X. Ding, "Analysis, understanding and representation of chinese newspaper with complex layout," in *Proc. 7th IEEE Int. Conf. Image Processing*, Vancouver, BC, Canada, Sep. 10–13, 2000, vol. 2, pp. 590–593.

[15] A. K. Jain and B. Yu, "Document representation and its application to page decomposition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 294–308, Mar. 1998.

[16] C.-C. Chang and C.-J. Lin, Lin, Libsvm—A Library for Support Vector Machines 200 [Online]. Available: http://www.csie.ntu.edu.tw/cjlin/libsvm/

[17] X. Lin, X. Ding, and M. Chen, "Adaptive confidence transform based classifier combination for chinese character recognition," *Pattern Recognit. Lett.*, vol. 19, no. 10, pp. 975–988, 1998.

[18] D. Chen, H. Bourlard, and J. P. Thiran, "Text identification in complex background using SVM," in *Proc. IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition*, vol. 2, 8–14, pp. 621–626.

[19] A. Antonacopoulos, B. Gatos, and D. Bridson, "ICDAR 2005 page segmentation competition," in *Proc. ICDAR*, Seoul, Korea, 2005, pp. 75–80.

[20] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "ICDAR 2003 robust reading competitions," in *Proc. 7th Int. Conf. Document Analysis and Recognition*, vol. 2, 8–14, pp. 682–687.

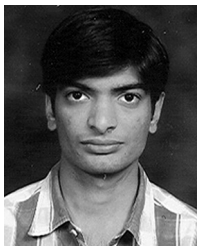
[21] C. W. Liang and P. Y. Chen, "DWT based text localization," *Int. J. Appl. Sci. Eng.*, vol. 2, no. 1, pp. 105–116.

[22] K. Sunil, K. Nitin, C. Santanau, and S. D. Joshi, "Locating text in images using matched wavelets," presented at the IEEE Int. Conf. Document Analysis and Recognition, 2005.

[23] D. Geman, S. Geman, C. Graffigne, and P. Dong, "Boundary detection by constrained optimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 7, pp. 609–628, Jul. 1990.
[24] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, Berkeley, CA, vol. 1, pp. 281–297.
[25] T. W. Ridlerand and S. Calvard, "Picture thresholding using an interactive selection method," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-8, no. 8, pp. 1264–1291, Aug. 1978.

Sunil Kumar received the B.Tech. degree from the Indian Institute of Technology in 2005.

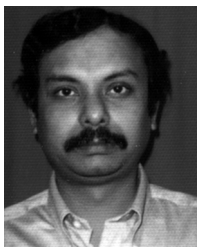
His research interests include image and signal processing. He is looking forward to a research-oriented carrier in signal processing and communications. Presently, he is with IBM-India Research Laboratory, Delhi.



Rajat Gupta received the B.Tech. degree in electrical engineering and the M.Tech. degree in information and communication technology from the Indian Institute of Technology, Delhi.

Currently, he is with Cypress Semiconductors, Bangalore. His research interests include image processing, signal processing, analog circuit design, and computer vision.

Nitin Khanna (S'06) received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Delhi, in 2005. He is currently pursuing the Ph.D. degree in electrical and computer engineering at Purdue University, West Lafayette, IN.



Santanu Chaudhury received the B.Tech. and Ph.D. degrees from the Indian Institute of Technology (I.I.T.), Kharagpur, in 1989 and 1984, respectively.

He is currently the Schlumberger Chair Professor in the Department of Electrical Engineering, I.I.T., Delhi. His research interests are in the areas of multimedia information retrieval, document image analysis and artificial intelligence.

Dr. Chaudhury was awarded the INSA medal for young scientists in 1993. He is a fellow of Indian National Academy of Engineers and National Academy of Sciences, India.

Shiv Dutt Joshi received the B.E. (Hons.) degree in electrical and electronics engineering from the Birla Institute of Technology and Science, Pilani, India, in 1981, the M.Tech. degree in communications and radar engineering, and the Ph.D. degree from the Indian Institute of Technology, Delhi, in 1983 and 1988, respectively.

He was a Lecturer with the Delhi Institute of Technology from 1988 to 1989, and he joined the Indian Institute of Technology, Delhi, as a faculty member in May 1989, where he has been a Professor since March 2000. His research interests include signal and image processing, multiscale modeling and signal/image analysis, group-theoretic approach to signal/image processing.

Dr. Joshi was a recipient of the AES award from the IEEE AES/COM Society, India Chapter, in 1986.