Wavelet Based Page Segmentation

Puneet Gupta gupta@cfar.umd.edu
Neeti Vohra nvohra@cise.ufl.edu
Santanu Chaudhury santanuc@ee.iitd.ernet.in
Shiv Dutt Joshi sdjoshi@ee.iitd.ernet.in

Department of Electrical Engineering
Indian Institute of Technology
Hauz Khas, New Delhi 110016

ABSTRACT

The process of page segmentation produces a description of the spatial extent and position of various components on the document page. In this paper, we present an approach for segmentation of a general document page image using wavelets. This method uses orthonormal wavelet decomposition to extract the attributes of the document spread over different scales. We have devised a scheme for the parameterisation of the font-size of text and also for distinguishing between text and non-text regions in the document. Based on these, a segmentation algorithm has been implemented and evaluated it through extensive testing.

Keywords : Page Segmentation, Wavelets, Multiresolution Analysis, Logical Page Decomposition

1 Introduction

The need for automated reading and processing of documents has been on a rise with advances in information and communication technology. Since the electronic counterparts of paper-based documents have obvious advantages in terms of storage, retrieval and updating, research in the area of document image understanding has been one of the areas of concentration of various research groups. Making legacy paper documents accessible by electronic means requires the development of efficient techniques for extraction of information from document images. Pixel based image data of a document page is mapped onto areas of semantic significance through a process of image segmentation. In this paper, we propose a new wavelet based approach for page segmentation.

Various schemes of page segmentation have been proposed by researchers. One of the most well known approach for page segmentation is the one of connected components aggregation [11]. Connected component based algorithms fall in the category of bottom-up segmentation algorithms. The components of same type are iteratively grouped together to form progressively higher-level descriptions of the printed regions of the document (words, lines, paragraphs, etc.) [6]. Other approaches use the description of white space to identify homogeneous regions. Antonacopoulos *et al.* [12] [1] [13] use the contours of the white space to delineate the boundary of text and image regions in the page. In most of the other approaches in the literature, researchers make assumptions about the general layout of the document page to be segmented. Some assume that the text or image blocks may only be rectangular [7], while others may assume that sentences in text are all evenly spaced. Others assume that the document belongs to a specific category, such as a newspaper [9] or a technical article [10]. However, making such assumptions restricts the applicability of the page segmentation scheme to a limited number of document classes. These methods cannot work in a generic environment as they work on an assumption that an apriori knowledge of the document model is available [8].

In this paper, we present a scheme that segments the page into its logical components independent of the layout of the document. Our algorithm does not assume the availability of any information about the layout and can, in general, be applied to any document type (check, memo, journal, newspaper etc.). Wavelets have interesting properties like Multiresolution Analysis(MRA) and the ability to provide local information in spatial as well as temporal domain. In the process of document-segmentation, we have used these properties to realise specific objectives. First, the nature of the energy distribution over different spatial scales has been used for separation of text from pictures. The ability of wavelets to separate out information at different scales enables us to characterise the font-sizes in text and yields a segmentation on the basis of font-size. This forms the basis for logical segmentation of text regions in the document page into titles, subtitles and general text. Scale property based approach for logical segmentation of a document image is one of the prime contributions of this work. Related work has been done by Doermann *et al.* where they use a multiscale segmentation of unstructured documents using soft decision integration [4]. Although their algorithm segments out the image and the text regions, it does not incorporate any feature that can be used to further segment out the text on the basis of font-size of the text. So, a complete logical decomposition of the document page is not attained by their segmentation algorithm.

The organisation of the paper is as follows. In Section 2, we present an overview of our proposed segmentation scheme. The description of the predicates that have been used in the segmentation are given in Section 3. The details of the segmentation algorithm have been explained in Sections 4 through 6. Section 7 discusses the results of implementation of our segmentation algorithm. Section 8 concludes this paper.

2 Overview of Our Segmentation Scheme

The task of page segmentation calls for a systematic approach so as to yield semantically meaningful segments. The strategy that we have proposed includes:

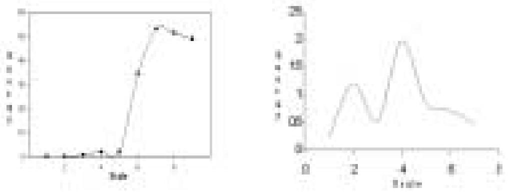
- Identification of text and non-text regions within the document
- White space separation
- Segmentation of text regions on the basis of font-size

The approach involves extraction of information based the distribution of energy over different scales called the **scalogram**, of wavelet transform. Predicates such as the value of third and fourth order central moments and location of the scale with maximum energy are obtained from the scalogram and used in the segmentation algorithm.

3 Scalogram and Segmentation Predicates

The scalogram represents the variation of the energy with the scale. The energy associated with each scale can be computed by the following expression.

$$e_j = \sum_k (d_{j,k})^2 \qquad (1)$$



Scalogram for Text Scalogram for Non-Text

Figure 1: Scalograms for Text and Non-Text Regions

where e_j is the energy at jth resolution and d_js are the Discrete Wavelet Transform(DWT) coefficients at resolution j, k being the spatial variable.

3.1 Predicates for Identification of Text and Non-Text Regions

In case of text, it is quite intuitive to expect that the structure of the alphabet will become most apparent at a certain scale. Therefore, the scalogram is expected to show a distinct peak at that scale. However, for pictures(non-text), the scalogram may or may not peak. The shape of the scalogram may not be consistent for different images and even if there exists a scale at which the scalogram peaks, the scale itself may vary for different sample pictures. Depending upon the distribution of pixel values over the picture, the details contained in the information may become apparent at any arbitrary scale or at more than one scale. Figure 1 shows example scalograms for text and non-text document pages. In general, the scalogram of a picture is flatter than the scalogram for text. Moreover, the scalogram of text is often skewed towards its predominant scale, unlike scalogram of an picture. The third order central moment for a probability distribution function(pdf) is a measure of the degree of skewness of the pdf. The fourth order central moment of the pdf measures the degree of flatness of the pdf. We have, therefore, employed the third and fourth order central moments of the scalograms normalised with respect to the total energy summed over all the scales.

We have computed the values of the third and fourth order central moments for documents containing only text or only pictures. It was observed that for text, the third order central moment is negative but for images, the value of third order central moment is typically random. The value of fourth order central moment was less than a threshold value(= 30) for text and greater for non-text. These thresholds were determined on the basis of a dataset of over 150 samples from each category. The consistency in the value of the third and fourth order central moments was as high as 98.4%, and hence the choice of these thresholds is justified.

3.2 Shift Invariant Property for Font-Size Characterisation

Consider the wavelet expansion of a signal

$$f(t) = \sum_i \sum_j a_{i,j} 2^{i/2} \psi(2^i t - j) \qquad (2)$$

Scaling the image by a factor of 2^p,

$$g(t) = f(2^p t) = \sum_i \sum_j b_{i,j} 2^{i/2} \psi(2^i t - j) \qquad (3)$$

where

$$b_{i,j} = \int f(2^p t) 2^{i/2} \psi(2^i t - j) dt \qquad (4)$$

Let $\mathcal{T} = 2^p\mathrm{t} \Rightarrow \mathrm{d}\mathcal{T} = 2^p\mathrm{dt}$ or

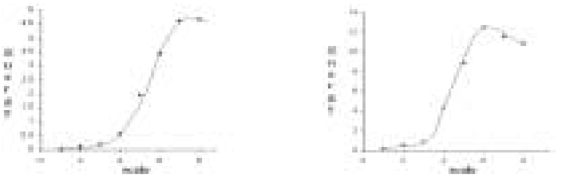
$$b_{i,j} = \int f(\mathcal{T}) 2^{i/2-p} \psi(2^{i-p}\mathcal{T} - j) d\mathcal{T} = 2^{-p/2} a_{i-p,j} \quad (5)$$

Total energy ($\mathcal{E}_b$) at i*th* scale for the scaled signal is given by

$$\mathcal{E}_b = \sum_i |b_{i,j}|^2 = \sum_i 2^{-p} \times |a_{i-p,j}|^2 = 2^{-p}\mathcal{E}_a \qquad (6)$$

where $\mathcal{E}_a$ is the energy of the original signal.

As apparent from the last equation, the energy of the scaled image at the i*th* scale is related to the original energy value at that scale by a magnitude scaling and peak translation. Scalograms for sample documents with font sizes 10 and 20 respectively are shown in Figure 2. It can be seen that the peak of the scalogram for the first document is at a higher scale (peak scale = 8) as compared to the scale at which the scalogram of the second documentpeaks (peak scale = 6). Such a difference in the location of the peak of the scalogram has been used for the characterisation of font-sizes in the segmentation approach that we have proposed.



Scalogram(Font Size = 10) Scalogram(Font Size = 20)

Figure 2: Scalograms for Documents of Different Font-Sizes

4 Identification of Text Regions

In order to extract the 'relevant' portions of the page image, the first step in the analysis is the separation of the part of the page that contains text from that which contains pictures. We have used scalogram for carrying out the preliminary separation between text and non-text regions in the document. The results obtained are then refined using a histogram based analysis. As discussed earlier, we have employed the third and fourth order central moments of the normalised scalograms as the distinguishing criterion between scalograms of the text and non-text regions using the quad-tree based segmentation.

4.1 Quad-Tree Based Segmentation Scheme

The algorithm that segments the page into text and non-text regions is based on the *Quad-Tree Segmentation* approach. The steps of the algorithm followed are shown in Figure 3. The steps follow as:

1. Compute the DWT of the entire document image and obtain its scalogram.
2. Normalise the scalogram with respect to the total energy and obtain the corresponding *scalogram pdf.* Compute the third(**O3**) and fourth(**O4**) order central moments of the scalogram pdf.
3. If **O3** is negative and **O4** is less than threshold (T=30), label the region as text. Otherwise, label the region as non-text. Cross check with the histogram based threshold (average gray scale>60). If labelling is not consistent, relabel the region suitably.
4. If the region is non-text, divide it into four equal regions, each of half its own size (top-left, top-right, bottom-left and bottom-right).
5. For each region, compute the DWT.
6. Repeat Steps 2-4 till either all the sub-regions have been identified as text regions or the size of sub-region reduces to the minimum allowable size of 64 × 64.
7. Repeat the last two steps for the remaining three regions.

After all the regions and sub-regions have been analysed, we carry out a *connected component merging* of non-text regions. This yields the boundary of the non-text regions, thus enabling us to identify the text regions that will pass through the next stage of segmentation.

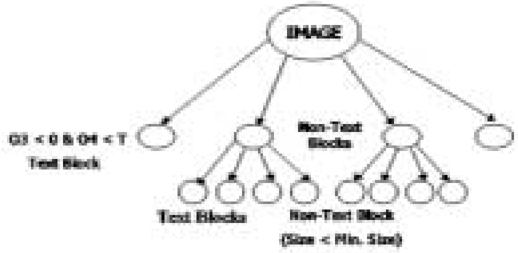
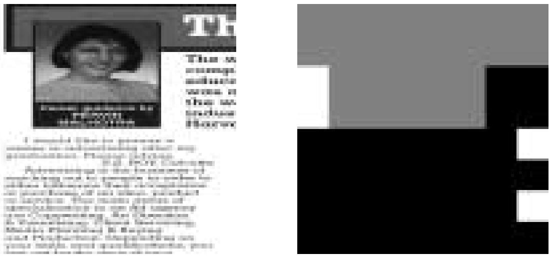


Figure 3: Quad-Tree Segmentation Algorithm



Original Image   Quad-Tree Segmented Image

Figure 4: Result of Refined Quad-Tree Segmentation

4.2 Refinement of Segmentation using Histogram

For a gray scale image, the histogram depicts the frequency of occurrence of specific gray levels over the range [0,255]. In a normal gray scale text-image, there are primarily two representative gray scales, one corresponding to the background and the other corresponding to the foreground or the text. For images however, the distribution of gray scales is more uniform. Unlike the case of text-images, the histogram for an image is expected to be smoother. As experiments revealed, in some cases, the values of third and fourth order moments for a non-text region may turn out to be similar to those for text. In order to reduce the probability of error, the scalogram based results were cross checked using a histogram based feature called the average gray scale value. The value of average gray scale has been found to be lower than 40 and that for images was consistently higher than 90. So a average gray scale 60 was chosen as threshold. The result of quad-tree segmentation, after incorporating this refinement step, can be seen for a sample document page under consideration in Figure 4.

5 Separation of White Space

Once the text regions have been recognised, the portions within the text which actually contain information must be separated from those which contain only the background (white space). Intuitively, background information is analogous to a constant signal and hence have no information contained in the wavelet coefficients other than the one at the lowest scale. On this basis, such regions can easily be identified from those that contain text or images. For the purpose of white space separation, we follow the following steps.

1. Select a spanning window size, say 8 × 8.
2. Compute the DWT of the portion of block and obtain its DWT coefficients.
3. If the DWT coefficient at the lowest or coarsest scale is non-zero with all the other coefficients at all other scales being zero, label the region as white space, otherwise label as non-white region.
4. Move the spanning window to obtain the next available block. Repeat steps 2 and 3 till all blocks have been labeled.
5. Obtain the blocks which belong to the background portion of the image. These are then merged together using *connected component analysis* and the boundary of the merged portion is then determined.

This approach is quite simple to implement and is also reliable. The white regions (obtained after the connected component merging) are logically removed from the domain of analysis during the segmentation for the rest of the image.

6 Font-Size based Segmentation of Text

After we have separated white space from the text regions, we carry out font-size based segmentation of the text using the scalogram. The following algorithm involves a **region growing** approach for the same. Starting with a small window size, we shall compute the DWT and hence the scalogram peak for the portion of image spanned by this window. The window is grown till segment boundaries are identified. This algorithm is stated more formally and precisely as follows.

1. Begin with the first text region obtained from the processing steps. Obtain the dimensions of this region.
2. Select a base window size for text analysis, say 16 × 16.

3. Begin from the top-left corner of the region. Compute the scalogram peak and the corresponding scale, for that window. Note that for the purpose of this computation, we first carry out a *texture creation* step for the portion of the image that has been spanned by the window. If the size of the text region is smaller than the base size, use the entire region for analysis.

4. Grow the window size by a given fraction. Calculate the peak scale again. Continue this process till the value of the peak scale stabilizes. This is an indication that the window contains text belonging to the same font-size. On continuing this process further, the peak scale value may get destabilized again. At this moment, we can infer that the window now includes some text that does not belong to the previous font-size. So, a region boundary can be identified.

5. Compute the DWT again with the base window size, starting from the place where the last boundary was identified. Repeat Step 4 and continue till the entire region has been spanned.

6. Repeat steps 3 through 5 for the other text regions obtained after white space separation(Section 5).

This algorithm enables us to segment a textual region into logical components without using arbitrary thresholds on dimensions of connected component or height of the lines [3]. After these steps have been carried out on all the text regions, the complete segmentation of the document page results.

7 Results and Discussions

We have implemented the proposed algorithm and tested it on a large database of images. The success of segmentation, however, is subject to the choice of segmentation parameters like the minimum window size for quad-tree segmentation. For our implementation, these heuristics have been chosen after a rigorous experimentation over different sets of parameters. The algorithm was first evaluated on pure text images. This algorithm involved white-space separation from the document, followed by segmentation of text based on the font size.

The segmentation algorithm for text-page segmentation was tested on images from various categories of documents like advertisements, newspaper articles, technical papers and magazine pages. The success of segmentation was evaluated by visual inspection of the segmented image. The lighter the shade of gray in the segmented images, larger the font size that it depicts. The white regions correspond to the white regions in the original image. As can be seen, the segmentation is quite satisfactory. The success rate of this scheme is around 88% over a set about 150 text pages. Note that the training set and test set of images were independent. However, a few errors do show up in the segmentation (see the merging of the phrase ‘By Dileep V Mavalankar’ - towrads the top-centre of the sample in Figure 6 - with the text block under it). Such errors arise because of the choice of the base size of the growing window and the neighborhood window size for connected component merging. And the overall quality of segmentation is quite good by visual standards even for such samples. The encouraging results of text segmentation leads us to the evaluation of the complete quad-tree segmentation. The parameters on the basis of which we have distinguished between the text and image regions have been discussed in Section 4.1 and Section 4.2. The minimum window size to which we split the original image (size 512 × 512) was 64 × 64. It can be seen that even for fairly complex samples like the one in Figure 8 have been segmented quite well. The text regions have been distinguished even on the basis of font-size.

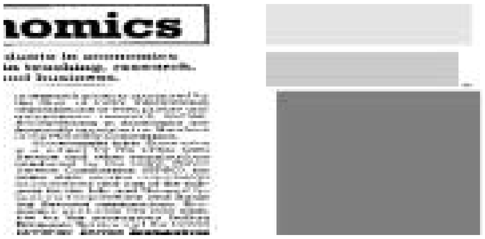


Figure 5: Segmentation Result of Text Document (1)

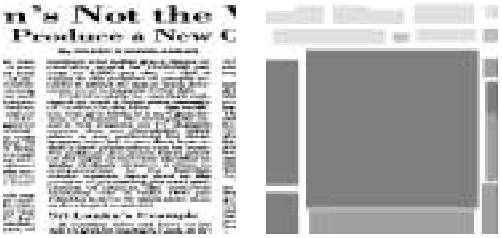


Figure 6: Segmentation of Text Document (2)

8 Conclusion

Each of the techniques mentioned in Section 1 relies to a certain extent, on the prior knowledge about the generic document layout or structure. However, making such assumptions about the textual or graphic

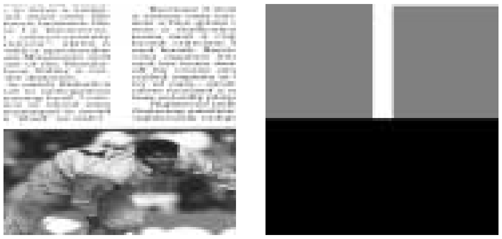


Figure 7: Segmentation of General Document (1)

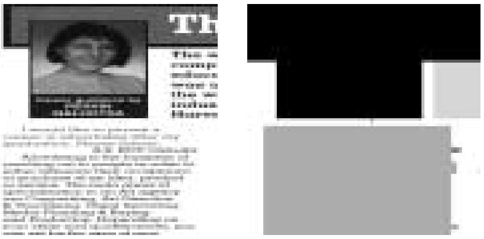


Figure 8: Segmentation of General Document (2)

attributes limits the applicability of the page segmentation scheme. In this paper, we have presented a generic scheme that segments the page into its logical components independently of the layout of the document, and is therefore applicable to a variety of document classes.

In our scheme, we first identify the text and non-text regions from the document page. This is followed by the separation of white space from the identified text regions. We then, carry out segmentation of the text regions on the basis of font-size. As indicated by the implementation results, this type of segmentation is efficient and does not assume any specific layout or shape of the image or text regions in the document page. The samples on which the scheme has been tested contains samples from various types of documents and the results for each of these have been equally encouraging.

Our approach can also be thought of as a step further in the approach followed by the Doermann *et al.* [4] in their multiscale segmentation of an unstructured document based on wavelets. We have not only been able to segment the document page into regions of different types, but also convey the semantic information about the relationship between the document's physical structure to its logical structure embedded within the text blocks.

References

[1] Antaonacopulos A., ‘Page Segmentation Using the Description of the Background’, Computer Vision and Image Understanding, Vol. 70, No. 3, June, pp. 350-369, 1998

[2] Wahl F.M., Casey R.G., ‘Block Segmentation and Text Extraction in mixed Text/Image Documents’, Computer Graphics Image Process, Vol. 20, pp. 375-390, 1982

[3] Jain A.K., Zhong Yu, ‘Page Segmentation Using Texture Analysis’, Pattern Recognition, Vol. 29, No.5, pp. 743-770, 1996

[4] Etemad K., Doermann D., Chellappa R., ‘Multiscale Segmentation of unstructured Document Pages Using Soft Decision Integration’,IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 1, January 1997

[5] Jain A.K., Yu B., ‘Document Representation and its Application to Page Decomposition’, IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 20, No. 3, March 1998

[6] Tan C. L., Yuan B., Huang W., Zhang Z., ‘Text/Graphics Separation using Pyramid Operations’, International Conference on Document Analysis and Recognition 1999, Bangalore, pp.169-172

[7] Sural S., Das P.K., ‘A Two Step Algorithm and its Parallelisation for the Generation of Minimum Containing Rectangles for Document Image Segmentation’, International Conference on Document Analysis and Recognition 1999, Bangalore, pp.173-176

[8] Jain A.K., Yu B., ‘Page Segmentation Using Document Model’, International Conference on Document Analysis and Recognition 1997, Munich, Vol.1, pp.173-176

[9] Wang D., Srihari S. N.,‘Classification of Newspaper Image Blocks Using Texture Analysis’,CVGIP, Vol. 47, pp. 327-352,1989

[10] Vishwanathan M., Nagy G., ’Characteristics of Digitized Images of Technical Articles’, SPIE, Vol. 1, 661, pp. 6-17, 1992

[11] Fletcher L., Kasturi R., ‘A Robust Algorithm for Text String Separation From Mixed Text/Graphics Images’, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 10, pp.. 910-918, 1988

[12] Antonacopoulos A., Ritchings R., ‘Flexible Page Segmentation Using the Background’, Proceedings of the 12th International Conference on Pattern Recognition, pp. 339-344, Jerusalem, 1994

[13] Pavlidis T., Zhou J., ‘Page Segmentation by White Streams’, Proceedings of the 10th International Conference on Pattern recognition, pp. 945-953, Saint-Malo, France, 1991