Scale Space Technique for Word Segmentation in Handwritten Documents [*]

R. Manmatha and Nitin Srimal

Computer Science Department,
University of Massachusetts, Amherst MA 01003, USA,
manmatha,nsrimal@cs.umass.edu,
WWW home page: http://ciir.cs.umass.edu

Abstract. Indexing large archives of historical manuscripts, like the papers of George Washington, is required to allow rapid perusal by scholars and researchers who wish to consult the original manuscripts. Presently, such large archives are indexed manually. Since optical character recognition (OCR) works poorly with handwriting, a scheme based on matching word images called word spotting has been suggested previously for indexing such documents. The important steps in this scheme are segmentation of a document page into words and creation of lists containing instances of the same word by word image matching.
We have developed a novel methodology for segmenting handwritten document images by analyzing the extent of "blobs" in a scale space representationof the image. We believe this is the first application of scale space to this problem. The algorithm has been applied to around 30 grey level images randomly picked from different sections of the George Washington corpus of 6,400 handwritten document images. An accuracy of $77 - 96$ percent was observed with an average accuracy of around 87 percent. The algorithm works well in the presence of noise, shine through and other artifacts which may arise due aging and degradation of the page over a couple of centuries or through the man made processes of photocopying and scanning.

1 Introduction

There are many single author historical handwritten manuscripts which would be useful to index and search. Examples of these large archives are the papers

[*] This material is based on work supported in part by the National Science Foundation, Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623, in part by the United States Patent and Trademark Office and Defense Advanced Research Projects Agency/ITO under ARPA order number D468, issued by ESC/AXS contract number F19628-95-C-0235, in part by the National Science Foundation under grant number IRI-9619117, in part by NSF Multimedia CDA-9502639 and in part by the Air Force Office of Scientific Research under grant number F49620-99-1-0138. Any opinions, findings and conclusions or recommendations expressed in this material are the authors and do not necessarily reflect those of the sponsors.

of George Washington, Margaret Sanger and W. E. B Dubois. Currently, much of this work is done manually. For example, 50,000 pages of Margaret Sanger's work were recently indexed and placed on a CDROM. A page by page index was created manually. It would be useful to automatically create an index for an historical archive similar to the index at the back of a printed book. To achieve this objective a semi-automatic scheme for indexing such documents have been proposed in [8]. In this scheme known as *Word Spotting* the document page is segmented into words. Lists of words containing multiple instances of the same word are then created by matching word images against each other. A user then provides the ASCII equivalent to a representative word image from each list and the links to the original documents are automatically generated. The earlier work in [8] concentrated on the matching strategies and did not address full page segmentation issues in handwritten documents. In this paper, we propose a new algorithm for word segmentation in document images by considering the scale space behavior of blobs in line images.

Most existing document analysis systems have been developed for machine printed text. There has been little work on word segmentation for handwritten documents. Most of this work has been applied to special kinds of pages - for example, addresses or "clean" pages which have been written specifically for testing the document analysis systems. Historical manuscripts suffer from many problems including noise, shine through and other artifacts due to aging and degradation. No good techniques exist to segment words from such handwritten manuscripts. Further, scale space techniques have not been applied to this problem before. [1] We outline the various steps in the segmentation algorithm below.

The input to the system is a grey level document image. The image is processed to remove horizontal and vertical line segments likely to interfere with later operations. The page is then dissected into lines using projection analysis techniques modified for gray scale image. The projection function is smoothed with a Gaussian filter (low pass filtering) to eliminate false alarms and the positions of the local maxima (i.e. white space between the lines) is detected. Line segmentation, though not essential is useful in breaking up connected ascenders and descenders and also in deriving an automatic scale selection mechanism. The line images are smoothed and then convolved with second order anisotropic Gaussian derivative filters to create a scale space and the *blob* like features which arise from this representation give us the focus of attention regions (i.e. words in the original document image). The problem of automatic scale selection for filtering the document is also addressed. We have come up with an efficient heuristic for scale selection whereby the correct scale for blob extraction is obtained by finding the scale maxima of the blob extent. A connected component analysis of the blob image followed by a reverse mapping of the bounding boxes allows us to

[1] It is interesting to note that the first scale space paper by T. Iijima was written in the context of optical character recognition in 1962 (see [12]). However, scale space techniqes are rarely used in document analysis today and as far as we are aware it has not been applied to the problem of character and word segmentation.

extract the words. The box is then extended vertically to include the ascenders and descenders. Our approach to word segmentation is novel as it is the first algorithm which utilizes the inherent scale space behavior of words in grey level document images. This paper gives a brief description of the techniques used. More details may be found in [11].

1.1 Related Work

Character segmentation schemes proposed in the literature have mostly been developed for machine printed characters and work poorly when extended to handwritten text. An excellent survey of the various schemes has been presented in [3]. Very few papers have dealt exclusively with the issue of word segmentation in handwritten documents and most of these have focussed on identifying gaps using geometric distance metrics between connected components. Seni and Cohen [9] evaluate eight different distance measures between pairs of connected component for word segmentation in handwritten text. In [7] the distance between the convex hulls is used. Srihari et all [10] present techniques for line separation and then word segmentation using a neural network. However, existing word segmentation strategies have certain limitations.

1. Almost all the above methods require binary images. Also, they have been tried only on clean white self-written pages and not manuscripts.
2. Most of the techniques have been developed for machine printed characters and not handwritten words. The difficulty faced in word segmentation is in combining discrete characters into words.
3. Most researchers focus only on word recognition algorithms and considered a database of clean images with well segmented words (see for example [1]). Only a few [10] have performed full, handwritten page segmentation. However, we feel that schemes such as [10] are not applicable for page segmentation in manuscript images for the reasons mentioned below.
4. Efficient image binarization is difficult on manuscript images containing noise and shine through.
5. Connected ascenders and descenders have to be separated.
6. Prior character segmentation was required to perform word segmentation and accurate character segmentation in cursive writing is a difficult problem. Also the examples shown are contrived (self written) and do not handle problems in naturally written documents.

2 Word Segmentation

Modeling the human cognitive processes to derive a computational methodology for handwritten word segmentation with performance close to the human visual system is quite complex due to the following characteristics of handwritten text.

1. The handwriting style may be cursive or discrete. In case of discrete handwriting characters have to be combined to form words.

2. Unlike machine printed text, handwritten text is not uniformly spaced.
3. Scale problem. For example, the size of characters in a header is generally larger than the average size of the characters in the body of the document.
4. Ascenders and descenders are frequently connceted and words may be present at different orientations.
5. Noise, artifacts, aging and other degradation of the document. Another problem is the presence of background handwriting or shine through.

We now present a brief background to scale space and how we have applied it to document analysis.

2.1 Scale Space and Document Analysis

Scale space theory deals with the importance of scale in any physical observation i.e. objects and features are relevant only at particular scales. In scale space, starting from an original image, successively smoothed images are generated along the scale dimension. It has been shown by several researchers [4, 6] that the Gaussian uniquely generates the linear scale space of the image when certain conditions are imposed.

We feel that *scale space* also provides an ideal framework for document analysis. We may regard a document to be formed of features at multiple scales. Intuitively, at a finer scale we have characters and at larger scales we have words, phrases, lines and other structures. Hence, we may also say that there exists a scale at which we may derive words from a document image. We would, therefore, like to have an image representation which makes the features at that scale (words in this case) explicit. The linear scale space representation of a continuous signal with arbitrary dimensions consists of building a one parameter family of signals derived from the original one in which the details are progressively removed. Let $f : \Re^2 \to \Re$ represent any given signal. Then, the scale space representation $I : \Re^2 \times \Re_+ \to \Re$ is defined by (see [6]) letting the scale space representation at zero scale be equal to the original signal $I(\cdot; 0) = f$ and for $t > 0$,

$$I(\cdot; t) = G(\cdot; t) \star f, \tag{1}$$

where $t \in \Re_+$ is the scale parameter, and G is the Gaussian kernel which in two dimensions $(x, y \in \Re)$ is written as

$$G(x, y; \sigma) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{(2\sigma^2)}} \tag{2}$$

where $\sigma = \sqrt{2t}$. We now describe the various stages in our algorithm.

2.2 Preprocessing

These handwritten manuscripts have been subjected to degradation such as fading and introduction of artifacts. The images provided to us are scanned versions of the photocopies of the original manuscripts. In the process of photocopying, horizontal and vertical black line segments/margins were introduced. Horizontal lines are also present within the text. The purpose of the preprocessing step is to remove some of these margins and lines so that they will not interfere with the blob analysis stage. Due to lack of space, this step is not described here. More details may be found in [11].

2.3 Line Segmentation

Line segmentation allows the ascenders and descenders of consecutive lines to be separated. In the manuscripts it is observed that the lines consist of a series of horizontal components from left to right. Projection profile techniques have been widely used in line and word segmentation for machine printed documents [5]. In this technique a 1D function of the pixel values is obtained by projecting the binary image onto the horizontal or vertical axis. We use a modified version of the same algorithm extended to gray scale images. Let $f(x, y)$ be the intensity value of a pixel (x, y) in a gray scale image. Then, we define the vertical projection profile as

$$P(y) = \sum_{x=0}^{W} f(x, y) \tag{3}$$

where W is the width of the image. Fig. 1 shows a section of an image in (a) and its projection profile in (b). The distinct local peaks in the profile corresponds to the white space between the lines and distinct local minima corresponds to the text (black ink). Line segmentation, therefore, involves detecting the position of the local maxima. However, the projection profile has a number of false local maxima and minima. The projection function $P(y)$ is therefore, smoothed with a Gaussian (low pass) filter to eliminate false alarms and reduce sensitivity to noise. A smoothed profile is shown in (c). The local maxima is then obtained from the first derivative of the projection function by solving for y such that :

$$P'(y) = P(y) \star G_y = 0 \tag{4}$$

The line segmentation technique is robust to variations in the size of the lines and has been tested on a wide range of handwritten pages. The next step after line segmentation is to create a scale space of the line images for blob analysis.

2.4 Blob Analysis

Now we examine each line image individually to extract the words. A word image is composed of discrete characters, connected characters or a combination of the two. We would like to merge these sub-units into a single meaningful entity which is a word. This may be achieved by forming a blob-like representation of the image. A blob can be regarded as a connected region in space. The traditional way of forming a blob is to use a Laplacian of a Gaussian (LOG) [6], as the LOG is a popular operator and frequently used in blob detection and a variety of multi-scale image analysis tasks [2, 6]. We have used a differential expression similar to

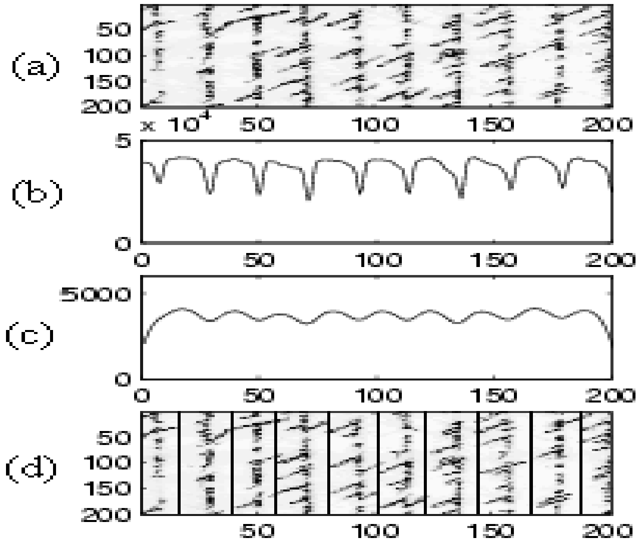


Fig. 1: (a) A section of an image, (b) projection profile, (c) smoothed projection profile (d) line segmented image

a LOG for creating a multi-scale representation for blob detection. However, our differential expression differs in that we combine second order partial Gaussian derivatives along the two orientations at different scales. In the next section we present the motivation for using an anisotropic derivative operator.

Non Uniform Gaussian Filters. In this section some properties which characterize writing are used to formulate an approach to filtering words. In [6] Lindeberg observes that maxima in scale-space occur at a scale proportional to the spatial dimensions of the blob. If we observe a word we may see that the spatial extent of the word is determined by the following :

1. The individual characters determine the height (y dimension) of the word and
2. The length (x dimension) is determined by the number of characters in it.

A word generally contains more than one character and has an aspect ratio greater than one. As the x dimension of the word is larger than the y dimension, the spatial filtering frequency should also be higher in the y dimension as compared to the x dimension. This domain specific knowledge allows us to move from isotropic (same scale in both directions) to anisotropic operators. We choose the x dimension scale to be larger than the y dimension to correspond to the spatial structure of the word. We define the anisotropic Gaussian filter as

$$G(x,y;\sigma_x,\sigma_y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-(\frac{x^2}{2\sigma_x^2}+\frac{y^2}{2\sigma_y^2})} \tag{5}$$

We may also define the multiplication factor η by $\eta = \frac{\sigma_x}{\sigma_y}$.

In the scale selection section we will show that the average aspect ratio or the multiplication factor η lies between three and five for most of the handwritten documents available to us. Also the response of the anisotropic Gaussian filter (measured as the spatial extent of the *blobs* formed) is maximum in this range. For the above Gaussian, the second order anisotropic Gaussian differential operator $L(x, y; \sigma_x, \sigma_y)$ is defined as

$$L(x, y; \sigma_x, \sigma_y) = G_{xx}(x, y; \sigma_x, \sigma_y) + G_{yy}(x, y; \sigma_x, \sigma_y) \tag{6}$$

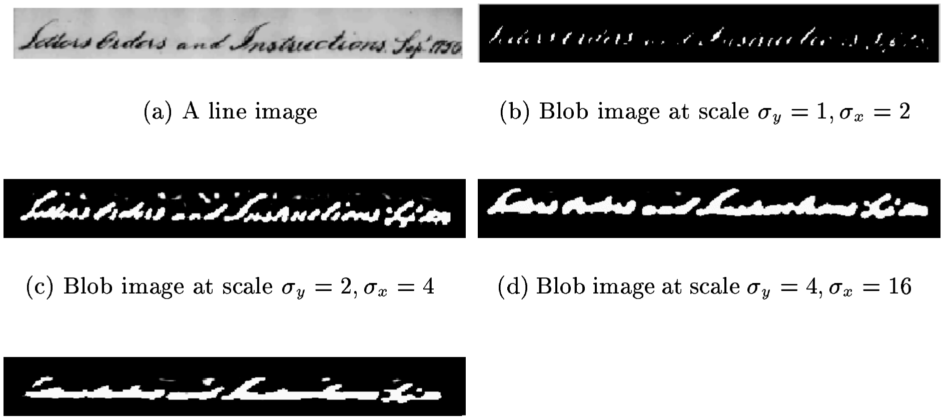
A scale space representation of the line images is constructed by convolving the image with 6. Consider a two dimensional image $f(x, y)$, then the corresponding output image is

$$I(x, y; \sigma_x, \sigma_y) = L(x, y; \sigma_x, \sigma_y) \star f(x, y) \tag{7}$$

The main features which arise from a scale space representation are blob-like (i.e. connected regions either brighter or darker than the background). The sign of I may then be used to make a classification of the 3-D intensity surface into foreground and background. For example consider the line image in Fig. 2(a). The figures show the blob images $I(x, y; \sigma_x, \sigma_y)$ at increasing scale values. Fig. 2(b) shows that at a lower scale the blob image consists of character blobs. As we increase the scale, character blobs give rise to word blobs (Fig. 2(c) and Fig. 2(d)). This is indicative of the phenomenon of merging in blobs. It is seen that for certain scale values the blobs and hence the words are correctly delineated (Fig. 2(d)). A further increase in the scale value may not necessarily cause word blobs to merge together and other phenomenon such as splitting is also observed. These figures show that there exists a scale at which it is possible to delineate most words. In the next section we present an approach to automatic scale selection for blob extraction.

2.5 Choice of Scale

Scale space analysis does not address the problem of scale selection. The solution to this problem depends on the particular application and requires the use of prior information to guide the scale selection procedure. Some of our work in scale selection draws motivation from Lindeberg's observation [6] that the maximum response in both scale and space is obtained at a scale proportional to the dimension of the object. A document image consists of structures such as characters, words and lines at different scales. However, as compared to other types of images, document images have the unique property that a large variation in scale is not required to extract a particular type of structure. For example, all the words are essentially close together in terms of their scale and can, therefore, be extracted without a large variation in the scale parameter. Hence, there exists a scale where each of the individual word forms a distinct blob. The output (blob) is then maximum at this value of the scale parameter. We will show that



(a) A line image

(b) Blob image at scale $\sigma_y = 1, \sigma_x = 2$

(c) Blob image at scale $\sigma_y = 2, \sigma_x = 4$

(d) Blob image at scale $\sigma_y = 4, \sigma_x = 16$

(e) Blob image at scale $\sigma_y = 6, \sigma_x = 36$

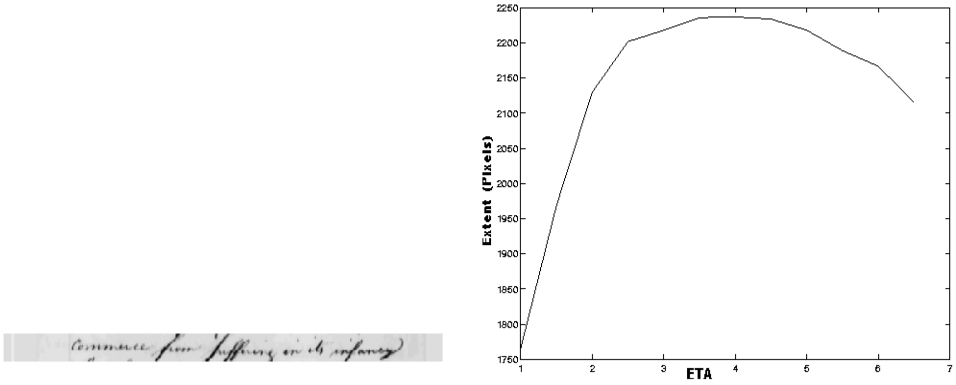
Fig. 2: A line image and the output at different scales

this scale is a function of the vertical dimension of the word if the aspect ratio is fixed.

Now, we highlight, the important differences in Lindeberg's approach to blob analysis and our work. In [6] Lindeberg determines interesting scale levels from the maxima over scale levels of a blob measure. He defines his blob measure to consist of the spatial extent, contrast and lifetime. A scale space blob tree is then constructed to track individual blobs across scales. In our analysis tracking individual blobs across scales is not the relevant issue nor is it computationally advisable because of the presence of a large number of blobs representing characters and words. Also it is impossible to determine whether an extrema corresponds to a character blob or a word blob and as mentioned earlier the variation of the best scale for a word is not large. What is important, however, is that we would like to merge character blobs and yet be able to delimit the word blobs. Therefore, we consider a blob as a connected region in space and measure its spatial extent but do not give it any volumetric significance. Spatial extent as a blob characteristic is computationally available to us and we observe that it shifts with scale giving a maximum as character blobs merge to form word blobs. This is in agreement with the intuitive reasoning that the response of the word at the correct scale of observation should be maximum as every blob has only a range of scales (lifetime) to manifest itself.

Our algorithm requires selecting σ_y and the multiplication factor η for blob extraction. We present an analysis which helped us arrive at a simple scale selection method based on the observation that the maximum of the spatial extent of the blobs corresponds to the best scale for filtering. To measure the variation in spatial extent of the blobs over scale we define ζ_i to represent the extent of a blob i. Then the total extent of the blobs A, for a line is given by $A = \sum_{i=1}^{n} \zeta_i$.

Selecting η. The parameters σ_y and σ_x try to capture the spatial dimensions of a word. An important characteristic of a word is its aspect ratio. A manual analysis of several images was carried out and it was shown that the average aspect ratio of a word in a document image is approximately $3.0 - 5.0$. We had earlier defined the multiplication factor η as $\eta = \sigma_x / \sigma_y$. An analysis of several images reveals that for constant σ_y, the maxima in extent was obtained for η lying in the range between $3 - 5$. A line image and the corresponding plot is shown in Fig. 3. In this Fig. the maximum is obtained in the region between $3.5 - 4$. This analysis along with the observation that the average aspect ratio of the word is between $3 - 5$ allows us to choose a value of η in the range $3 - 5$. Specifically, for further analysis we choose $\eta = 4$.



(a) A line image

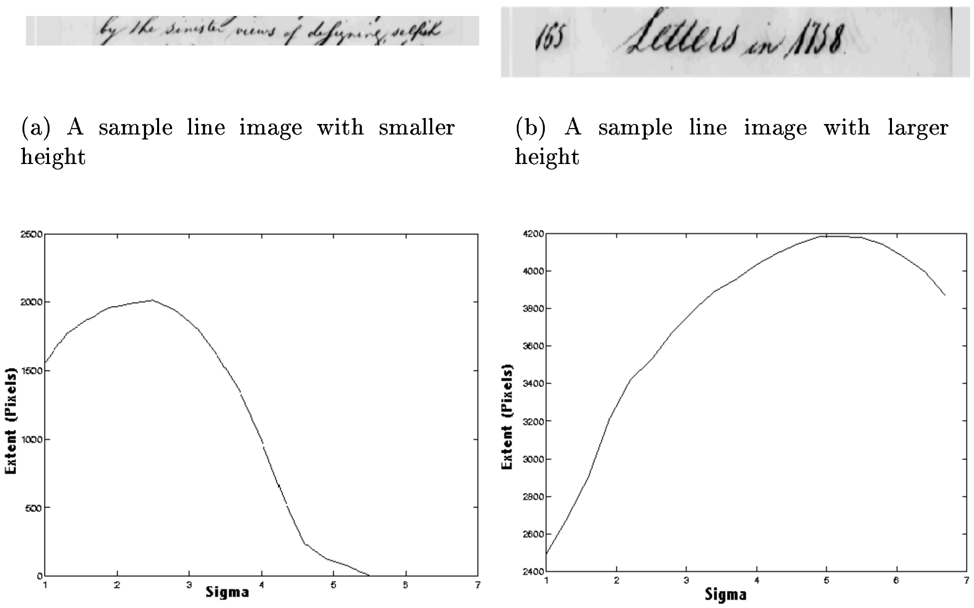
(b) Plot of extent vs η, $\sigma_y = 2$ for above image, a maximum is obtained at $\eta = 3.8$

Fig. 3: Variation of blob extent vs η with constant σ_y

Selecting σ_y. Fig. 4 shows the line images and corresponding plots of extent versus σ_y for constant η. As seen in the figures the total extent exhibits a peak which depends on σ_y. The figures also show how the peak shifts with the change in the size (height) of the characters. Experimentally it was found that σ_y (y scale) is a function of the height of the words (which is related to the height of the line). An estimate of σ_y is obtained by using the line height i.e.

$$\sigma_y = k \times \text{Line height} \tag{8}$$

where $0 < k < 1$. The nearby scales are then examined to determine the maximum over scales. For our specific implementation we have used $k = 0.1$ and sampled σ_y at intervals of 0.3. The two values were determined experimentally and worked well over a wide range of images.



(a) A sample line image with smaller height

(b) A sample line image with larger height

(c) Plot of extent vs σ_y,Maximum is obtained at $\sigma_y = 2.5$

(d) Plot of extent vs σ_y, Maximum is obtained at $\sigma_y = 6.0$

Fig. 4: Variation of blob extent vs σ_y with constant $\eta = 4$.

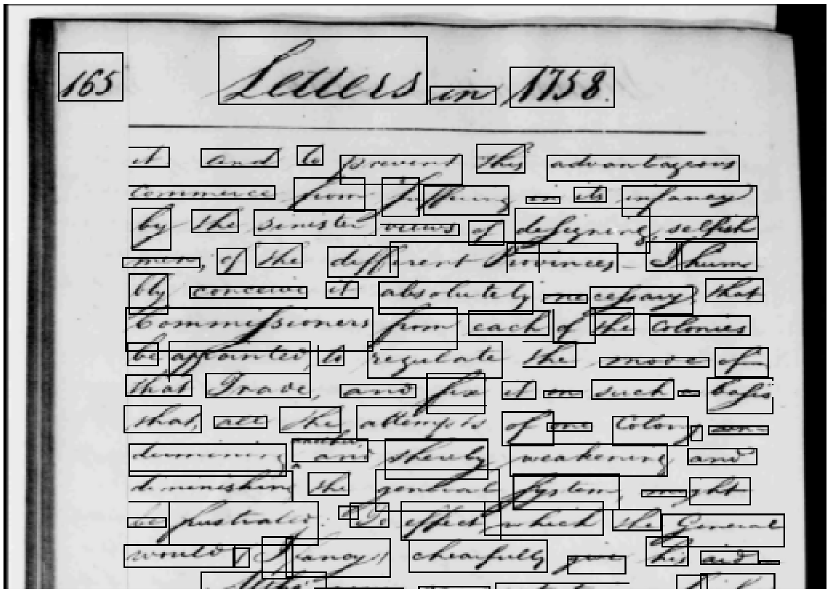
2.6 Blob Extraction and Post Processing

The blobs are then mapped back to the original image to locate the words. A widely used procedure is to enclose the blob in a bounding box which can be obtained through connected component analysis. In a blob representation of the word, localization is not maintained. Also parts of the words, especially the ascenders and descenders, are lost due to the earlier operations of line segmentation and smoothing (blurring). Therefore, the above bounding box is extended in the vertical direction to include these ascenders and descenders. At this stage an area/ratio filter is used to remove small structures due to noise.

3 Results

The technique was tried on 30 randomly picked images from different sections of the George Washington corpus of 6,400 images and a few images from the archive of papers of Erasmus Hudson. To reduce the run-time, the images have been smoothed and sub-sampled to a quarter of their original size. The algorithm takes 120 seconds to segment a document page of size 800 x 600 pixels on a PC with a 200 MHz pentium processor running LINUX. A segmentation accuracy ranging from 77−96 percent with an average accuracy around 87.6 percent was observed.

Fig. 5 shows part of a segmented page image with bounding boxes drawn on the extracted words. The method worked well even on faded, noisy images and Table


165 Letters in 1758.

it and to prevent this advantageous
commerce from suffering in its infancy
by the sinister views of designing selfish
men, of the different Provinces. I hum-
bly conceive it absolutely necessary that
Commissioners from each of the Colonies
be appointed to regulate the mode of
that Trade, and fix it on such a basis,
that all the attempts of one Colony un-
dermining and thereby weakening and
diminishing the general system might
be frustrated; to effect which the General
would I fancy, chearfully give his aid

Fig. 5: Segmentation result on part of image 1670165.tif from the George Washington collection

4 shows the results averaged over a set of 30 images. The first column indicates the average no. of distinct words in a page as seen by a human observer. The second column indicates the % of words detected by the algorithm i.e, words with a bounding box around them, this includes words correctly segmented, fragmented and combined together. The next column indicate the % of words fragmented. Word fragmentation occurs if a character or characters in a word have separate bounding boxes or if 50 percent or greater of a character in a word is not detected. Line fragmentation occurs due to the dissection of the image into lines. A word is line fragmented if 50 percent or greater of a character lies outside the top or bottom edges of the bounding box. The sixth column indicates the words which are combined together. These are multiple words in the same bounding box. The last column gives the percentage of correctly segmented words.

4 Conclusion

We have presented a novel technique for word segmentation in handwritten documents. Our algorithm is robust and efficient for the following reasons:

1. We use grey level images and, therefore, image binarization is not required. Image binarization requires careful pre-selection of the threshold and gen-

erally results in a loss of information. The threshold parameter has to be selected locally and is very sensitive to noise, fading and other phenomenon.

2. Since the images are heavily smoothed, insignificant *blobs* can easily be eliminated. Therefore, the technique is comparatively unaffected by the presence of speckles which otherwise would have greatly affected techniques requiring binarization as the first step.
3. One of the major advantages of our approach is that the scheme is largely unaffected by shine through. This is because the algorithm is based on blurring and the information is extracted in the form of blobs.
4. The algorithm makes minimal assumptions about the nature of handwriting and fonts and may be extended to word segmentation in other language documents where words are delineated by spaces. Also, the method does not require prior training.

No. of documents	Average no. of words per image	% words detected	% fragmented words + line	% words combined	% words correctly correctly
30	220	99.12	1.75 + 0.86	8.9	87.6

Table 1: Table of segmentation results

References

1. A.J. Robinson A.W. Senior. An off-line cursive handwriting recognition system. *IEEE transactions on PAMI*, 3:309–321, 1998.
2. D. Blostein and N. Ahuja. A multi-scale region detector. *CVGIP*, 45:22–41, January 1989.
3. R. G. Casey and E. Lecolinet. A survey of methods and strategies in character segmentation. *IEEE Transactions on PAMI*, 18:690–706, July 1996.
4. L. M. J. Florack. *The Syntactic Structure of Scalar Images.* Kluwer Academic Publishers, 1997.
5. J. Ha, R. M. Haralick, and I. T. Phillips. Document page decomposition by the bounding-box projection technique. In *ICDAR*, pages 1119–1122, 1995.
6. T. Lindeberg. *Scale-space theory in computer vision.* Kluwer Academic Publishers, 1994.
7. U. Mahadevan and R. C. Nagabushnam. Gap metrics for word separation in handwritten lines. In *ICDAR*, pages 124–127, 1995.
8. R. Manmatha and W. B. Croft. Word spotting : Indexing handwritten manuscripts. In Mark Maybury, editor, *Intelligent Multi-media Information Retrieval.* AAAI/MIT press, April 1998.
9. G. Seni and E. Cohen. External word segmentation of off-line handwritten text lines. *Pattern Recognition*, 27:41–52, 1994.
10. S. Srihari and G. Kim. Penman : A system for reading unconstrained handwritten page images. In *Symposium on document image understanding technology (SDIUT 97)*, pages 142–153, April 1997.
11. N. Srimal. Indexing handwritten documents, *M.S. Thesis, University of Massachusetts Computer Science Tech Report. 1999.*
12. *J. A. Weickert, S. Ishikawa, and A. Imiya. On the history of gaussian scale-space axiomatics. In J. Sporring, M. Nielsen, L. M. J. Florack, and P. Johansen, editors,* Gaussian Scale-Space Theory*, pages 45–59. Kluwer Academic Press, 1997.*