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.

M. Nielsen et al. (Eds.): Scale-Space'99, LNCS 1682, pp. 22–33, 1999.

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 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

[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.