LAMP-TR-033 MDA 9049-6C-1250
CAR-TR-915 April 1999
CS-TR-4014

Full-Text Access to Historical Newspapers

Tapas Kanungo[1] and Robert B. Allen[2]

[1]Center for Automation Research
[2]College of Library and Information Services
University of Maryland
College Park, MD 20742
kanungo@cfar.umd.edu, rba@glue.umd.edu

Abstract

Newspapers are rich records of U.S. history. Due to the deterioration of older newspapers, the National Endowment for the Humanities is archiving 19th century newspapers on microfilm. Although microfilm is a good preservation method, it provides limited access to researchers and the general public. We are building a system to provide universal access to digital images and full-text content of historical newspapers. The system has three main components: (a) An Optical Character Recognition (OCR) module that converts digitized images into searchable text and identifies regions. (b) An Information Retrieval module that applies linguistic information to aid in segmentation, indexing, and retrieval of the noisy OCR'd text. (c) A User Interface module that allows historians and educators to query and view retrieved documents. Thus far, we have developed two OCR techniques targeted to processing historical newspapers and we have built a user interface to search the OCR output and superimpose matches on a page image from the newspaper.

This research was funded in part by the Department of Defense and the Army Research Laboratory under Contract MDA 9049-6C-1250.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
APR 1999		**00-04-1999 to 00-04-1999**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Full-Text Access to Historical Newspapers	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Language and Media Processing Laboratory,Institute for Advanced Computer Studies,University of Maryland,College Park,MD,20742-3275**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**15**	

LAMP-TR-033
CAR-TR-915
CS-TR-4014

MDA 9049-6C-1250
April 1999

Full-Text Access to Historical Newspapers

Tapas Kanungo and Robert B. Allen

Full-Text Access to Historical Newspapers

Tapas Kanungo[1] and Robert B. Allen[2]

[1]Center for Automation Research
[2]College of Library and Information Services
University of Maryland
College Park, MD 20742
kanungo@cfar.umd.edu, rba@glue.umd.edu

Abstract

Newspapers are rich records of U.S. history. Due to the deterioration of older newspapers, the National Endowment for the Humanities is archiving 19th century newspapers on microfilm. Although microfilm is a good preservation method, it provides limited access to researchers and the general public. We are building a system to provide universal access to digital images and full-text content of historical newspapers. The system has three main components: (a) An Optical Character Recognition (OCR) module that converts digitized images into searchable text and identifies regions. (b) An Information Retrieval module that applies linguistic information to aid in segmentation, indexing, and retrieval of the noisy OCR'd text. (c) A User Interface module that allows historians and educators to query and view retrieved documents. Thus far, we have developed two OCR techniques targeted to processing historical newspapers and we have built a user interface to search the OCR output and superimpose matches on a page image from the newspaper.

This research was funded in part by the Department of Defense and the Army Research Laboratory under Contract MDA 9049-6C-1250.

1 Introduction

Newspapers are rich records of U.S. history. Due to the deterioration of older newspapers the National Endowment for the Humanities is archiving 19th century newspapers on microfilm. Although microfilm is a good preservation method, it provides limited access for researchers and the general public. We propose to build a system to convert microfilmed historic newspapers into both digital image and full-text electronic archives, thereby providing universal access to their content over the web.

The extraction of text from the page images presents a wide range of research issues. To meet the research challenges of this project we have assembled an interdisciplinary team with expertise in document image analysis, user-interface design, information retrieval, system integration, cross-language retrieval, library metadata standards, history, and journalism. Some of the problems, for example the extraction of entire newspaper articles from zones of text, have not been considered in any of the related research communities.

Several other projects have developed collections of digitized newspapers, journals, and personal letters. Perhaps the best known of these is JSTOR [34] which has digitized page images of scholarly journals. While JSTOR does OCR of its content, the original material is fairly straightforward compared to the historic newspapers we plan to index. The "Valley of the Shadow" project [35], which examined primary source material from a pair of communities involved in the Civil War, did not include the full text of many news stories but had only abstracts of those stories. Our tools could greatly enrich that corpus by providing direct OCR of the newspapers.

1.1 Brief History of Newspapers as Relevant to OCR

Publick Occurrences, Both Forreign and Domestick, the first newspaper to be published in North America, was printed in Boston in 1690. This newspaper was immediately banned by the Governor of Massachusetts — non-official publications were not allowed in colonial America. After this initial setback, newspapers flourished in North America. In 1704, the *Boston Newsletter* became the first official newspaper to be published and replaced the official pamphlets, newsletters and proclamations. By the Revolution, there were 37 different newspaper titles. Many newspapers like *Pennsylvania Packet and Daily Advertiser* then started publishing daily and their titles reflected their new revenue source. Most colonial editorials tried to chronicle the historical events that led to the creation of the new nation but in the process became quite partisan. *The New York Herald*, which was established in 1835, was the first newspaper to to be advertised as a politically independent newspaper.

Technological advances in printing helped in increasing the production and improving the quality of newspapers. Prior to 1814, hand-operated wooden presses were used for printing newspapers, books, magazines, and pamphlets. The invention of the steam-driven "double-press" increased the rate of production to 5,000 copies per hour. This rate increased in 1865 with the invention of the rotary press. The quality of print was affected by the inking methods (automatic or manual), printing plates, type of paper, etc. The typesetting up to that point was done character by character. That is, each

individual character was typeset manually on a matrix and then the entire typeset page was printed. Setting individual characters led to non-uniformity in the orientation and location of characters.

A major change in newspaper print quality occurred in the late 19th century when the Linotype printing machine was invented by Ottmar Mergenthaler. This machine allowed the typesetter to typeset and cast an entire line of type using a keyboard. The Linotype improved the quality of print and also increased the rate of typesetting of newspapers. The number of American newspapers increased dramatically from 1880 to 1900 from 850 titles to 2000. In 1886 the *New York Tribune* was one of the first newspapers to use the Linotype for printing. The drawback of Linotype machines was that they did not allow kerning of letters (unless compound matrices were used) and so the italic letter *'f'* had a stunted head and tail. Monotype printing machines were invented at the same time by Tolbert Lanston in 1887 in competition with the Linotype. Monotype machines were similar to Linotype machines except that individual characters could be typeset from the keyboard. The final matrix had individual characters and so it was possible to manually introduce kerning and typeset large display characters like dropcaps and raised caps.

Page layout has changed over time. In early 1800s newspapers were typeset quite closely with very little space between lines and columns and used very small font sizes. The point size of the type used has not changed much since the early 1900s. After the 1900s older Gothic fonts gave way to newly designed fonts like Cheltenham; and headlines started appearing in Bodoni font. Line-drawings and cartoons started appearing by the 1870s. These were either carved on wooden blocks, or etched on zinc plates. The photoengraving process was invented in the 1860s in England and was perfected by Federic E. Ives of Cornell University in 1886. The photoengraving process was adapted for the rotary presses used for printing newspapers in 1880. In 1897, *The New York Tribune* was the first newspaper to start printing halftone reproductions of photographs.

In the 20th century, one of the most significant changes occurred in 1946 when the U.S. Government Printing Office developed the Intertype Fotosetter. This was a line-casting machine containing brass matrices with film negatives of characters. Images were produced on photosensitive paper from which printing plates were created. In 1950 the Photo 200 machine was invented. It had a spinning film matrix containing all characters in a font and a stroboscopic light source was used to print on photo-sensitive paper. For additional information about the history of printing, typography, and newspaper publishing see [9, 30, 8].

1.2 System Overview

Our project will develop tools to process large quantities of digitized newspaper images automatically. There are several reasons we feel that this effort is now feasible and desirable. We have access to prototype OCR research software and we have considerable expertise that will allow us to extend both OCR and IR research on these topics. In addition, the confluence of greater storage and processing capacity with acceptance of the web for distribution of content provides a suitable infrastructure for acceptance of the system.

We have begun to develop our techniques and to build a small corpus by working

from a paper copy of an original paper copy of *The Brooklyn Eagle* for November 11, 1917 (Armistice Day) and from a spool of microfilm covering that issue of the newspaper. Given the 75-year copyright duration, we wanted to find a newspaper from before 1923 and Armistice Day had a variety of news stories including news about World War I as well as descriptions of Suffragette marches. The current system includes several modules — OCR, search, and interface.

2 Research

2.1 Optical Character Recognition

THE BROOKLYN DAILY EAGLE

7 SECTIONS. NEW YORK CITY, SUNDAY, NOVEMBER 11, 1917. 68 PAGES. THREE CENTS

BRITISH GUNS ON ITALIANS' FIRST LINE; TEUTONS TAKE ASIAGO, IN TRENTINO; U. S. TO IGNORE TALK OF ARMISTICE

Austro-Germans Begin Flanking Movement Against Italian Left Wing—Capture Asiago After Hard Fighting. Opposing Armies Face Each Other Along the Piave — Great Battle Near — Venice Is in Danger.

Washington Will Ignore Lenine's Armistice Plan

Bolsheviki Forms Cabinet, With Lenine as Premier—Russian Embassy at Washington Refuses Recognition to Maximalist Regime. Revolutionists in Control of Moscow Government — Americans in Petrograd Safe, Francis Reports.

Features in Classification For the Next Army Draft

BEGIN SECOND DRAFT CLASSIFYING DEC. 15; PRESIDENT ASKS AID

Process Will Take Two Months. Crowder Orders Legal Advice Bureaus.

ACTUAL DRAFT IN FEBRUARY

NOT A PUFF!

The Russian Women's Battalion surrendered. Probably out of powder.

HAIG IN NEW DASH GAINS HALF MILE AT PASSCHENDAELE

Sweeps Forward 800 Yards on Wide Front in Flanders.

REPULSED—BERLIN

6 GRAFT ARRESTS IN FOOD SCANDAL AT THE NAVY YARD

Two Navy Men and Provision Dealer Among Those Locked Up.

U.S. LOSES THOUSANDS

ABRAHAM AND STRAUS

LOW PRICES An Old Story Here

Low prices have always been a cardinal principle with this Store.

Low prices—invariably linked with high quality.

This desirable combination is possible because we are favorably situated in several respects:—

(1) Owning our own property at low cost; avoiding high rentals, though in a central position.

(2) Paying CASH for our purchases—a habit so well known that we are always offered choice lots of merchandise at low prices FIRST by makers who wish to turn their stocks into money quickly and know we will help them.

(3) Modest profits—so that we can truthfully say that we sell at the lowest prices — for like quality—of any store in New York.

Today this question of

Figure 1: Original scanned image of the *Brooklyn Daily Eagle*.

The Optical Character Recognition system creates symbolic, searchable text from scanned images [22, 4]. While there are numerous commercial OCR products, most of them fail to recognize text in highly degraded newspaper images. The main causes for deterioration of performance are (i) joined and broken characters, (ii) salt and pepper noise, (iii) character-to-character variation due to old typesetting technology, (iv) page-bending at the spine of the newspaper, (v) the very small gaps between columns and lines of text, (vi) line separators and black outlines around advertisements, (vii) a wide variety of fonts, and (viii) paper aging and degradation.

We have built a prototype OCR system using a commercial development kit from Caere Corporation. The system currently does not provide us with segmented zones but provides us with word bounding boxes, text strings within the boxes, and OCR confidence levels. This output is currently being used by the IR system to locate the search results on the page image. The recognition accuracy level on historic documents is lower than what the system achieves on new newspapers. We are in the process of creating a benchmarking dataset and conducting an OCR accuracy evaluation using the methodology outlined in [21]. While the OCR results are not on a par with what one


THE BROOKLYN DAILY EAGLE

7 SECTIONS. NEW YORK CITY, SUNDAY, NOVEMBER 11, 1917. 68 PAGES. THREE CENTS

BRITISH GUNS ON ITALIANS' FIRST LINE; TEUTONS TAKE ASIAGO, IN TRENTINO; U. S. TO IGNORE TALK OF ARMISTICE

Austro-Germans Begin Flanking Movement Against Italian Left Wing—Capture Asiago After Hard Fighting. Opposing Armies Face Each Other Along the Piave — Great Battle Near — Venice Is in Danger.

Washington Will Ignore Lenine's Armistice Plan

Bolsheviki Forms Cabinet, With Lenine as Premier—Russian Embassy at Washington Refuses Recognition to Maximalist Regime. Revolutionists in Control of Moscow Government — Americans in Petrograd Safe, Francis Reports.

NOT A PUFF!

6 GRAFT ARRESTS IN FOOD SCANDAL AT THE NAVY YARD

Two Navy Men and Provision Dealer Among These Locked Up.

U.S. LOSES THOUSANDS

HAIG IN NEW DASH GAINS HALF MILE AT PASSCHENDAELE

Sweeps Forward 800 Yards on Wide Front in Flanders.

REPULSED—BERLIN

BEGIN SECOND DRAFT CLASSIFYING DEC. 15; PRESIDENT ASKS AID

Process Will Take Two Months, Crowder Orders Legal Advice Bureaus.

ACTUAL DRAFT IN FEBRUARY

ABRAHAM AND STRAUS

LOW PRICES An Old Story Here

Figure 2: *The Brooklyn Daily Eagle* after being processed with the line-removal algorithm.

gets on newspaper documents, the IR results have shown that it is not necessary to have very high accuracy for reasonable precision and recall (e.g., [6]).

Because our goal is to eventually retrieve articles, we need noise-removal algorithms, robust zone segmentation algorithms, and higher-level post-processing. We will approach the problem of noise filtering by first creating validated models of the microfilm degradation process and the corresponding parameter-estimation algorithms [20, 19, 16]. These models will then be used to design noise-removal and image-restoration algorithms. The impact of the noise-removal algorithm will be evaluated using a benchmarking dataset. Zone segmentation algorithms can be very sensitive to inter-column and inter-line gaps.

In the next section, we describe a morphological line-removal algorithm that filters out lines from an image before it is processed by a page segmentation algorithm. This is followed by an adaptive X-Y cut algorithm [25] to create blocks of homogeneous regions, which are then classified into text and non-text regions using a statistical decision tree.

2.1.1 Line Removal

Mathematical morphology is an image processing approach based on set theory [26, 33, 12]. Images are considered as pointsets and dilation, erosion, closing, and opening are performed on the images to extract features and find spatial relations among the detected features [17]. The standard set theory operations are also valid operations that can be performed on two images.

We first deskew the image to make the lines parallel to the vertical and horizontal axes. Next we fill in breaks in vertical and horizontal lines by performing a closing operation with vertical and horizontal structuring elements. Next, we remove speckle noise by opening the resulting image with a disk structuring element. The vertical and horizontal lines are detected by performing an opening operation with long vertical and horizontal structuring elements. The detected lines are then subtracted from the original


THE BROOKLYN DAILY EAGLE

SECTIONS. NEW YORK CITY, SUNDAY, NOVEMBER 11, 1917. 68 PAGES. THREE CENTS

BRITISH GUNS ON ITALIANS' FIRST LINE; TEUTONS TAKE ASIAGO, IN TRENTINO; U. S. TO IGNORE TALK OF ARMISTICE

Austro-Germans Begin Flanking Movement Against Italian Left Wing—Capture Asiago After Hard Fighting. Opposing Armies Face Each Other Along the Piave — Great Battle Near — Venice Is in Danger.

Washington Will Ignore Lenine's Armistice Plan

Bolsheviki Forms Cabinet, With Lenine as Premier—Russian Embassy at Washington Refuses Recognition to Maximalist Regime. Revolutionists in Control of Moscow Government — Americans in Petrograd Safe, Francis Reports.

Features in Classification For the Next Army Draft

BEGIN SECOND DRAFT CLASSIFYING DEC. 15; PRESIDENT ASKS AID

Process Will Take Two Months. Crowder Orders Legal Advice Bureaus.

ACTUAL DRAFT IN FEBRUARY

NOT A PUFF!

The Russian Women's Battalion surrendered. Probably out of powder.

HAIG IN NEW DASH GAINS HALF MILE AT PASSCHENDAELE

Sweeps Forward 800 Yards on Wide Front in Flanders.

REPULSED—BERLIN

6 GRAFT ARRESTS IN FOOD SCANDAL AT THE NAVY YARD

Two Navy Men and Provision Dealer Among Those Locked Up.

U.S. LOSES THOUSANDS

ABRAHAM AND STRAUS WEATHER TOMORROW: FAIR.

LOW PRICES An Old Story Here

Low prices have always been a cardinal principle with this Store.

Low prices—invariably linked with high quality.

This desirable combination is possible because we are favorably situated in several respects:—

(1) Owning our own property at low cost; avoiding high rentals, though in a central position.

(2) Paying CASH for our purchases—a habit so well known that we are always offered choice lots of merchandise at low prices FIRST by makers who wish to turn their stocks into money quickly and know we will help them.

(3) Modest profits—so that we can truthfully say that we sell at the lowest prices — for like quality—of any store in New York.

Today this question of

Figure 3: Automatic zone segmentation for the *Brooklyn Eagle* without first removing lines. Notice that many columns are merged.

image. For details of the algorithm the reader is referred to [15, 17, 12].

In Figure 1 we show the original image of a historical newspaper. Most of the lines are vertical or horizontal. In Figure 2 we show the result of our line removal algorithm. Notice that not all lines are removed. This is because there are still a few breaks in the lines that were not filled in by the preprocessing step.

2.1.2 Zone Segmentation

Historical newspapers have a fairly regular layout. However, the column and row gaps are typically quite narrow. Furthermore, black bounding boxes around advertisements and line separators between columns are the main causes of incorrect page segmentation. We have built an algorithm that filters the lines and bounding boxes before doing OCR. Preliminary results show that the filtering algorithm leads to proper segmentation and results in detection of words that were earlier not recognized. The segmented zones will then be classified into various types — text block, heading, title, advertisement, logo, etc. — using a statistical decision tree [14, 31, 29, 36]. The decision tree will automatically construct the rules for minimum-error classification. The construction of the decision tree requires a dataset of images with corresponding manually segmented and labeled zones. A benchmarking dataset will be created for this purpose.

Finally, to evaluate the performance of the OCR system, we will create a benchmarking image dataset with corresponding zone and character groundtruth [11, 16, 18, 21]. A statistical, stratified sampling procedure will be adopted to create a representative sample of images. The sample will be representative of the variation in font, typesetting technology, layout, degradation, microfilm, etc. This dataset will be split into two parts. One part will be used for designing noise-removal algorithms and building decision trees, and the second part will be used for independent evaluation.

Figure 4: Automatic zone segmentation for the *Brooklyn Eagle* after removing lines.

2.2 Statistical Language Processing and Information Retrieval

There have been extensive studies of the impact of degraded OCR on retrieval performance [7]. Typically, retrieval is fairly robust with word-error rates of up to about 30 or 40%. For newspapers from the late 19th and early 20th century, our results should be safely within that range.

The OCR issues merge into and interact with higher-level linguistic concerns. For instance, we need to determine how stories are continued across columns, around pictures, and onto inside pages. There are several approaches to this story-continuation problem. First, the similarity of blocks of text may be compared. Second, specific features of the text may be analyzed. For instance, a sentence fragment at the end of one section of the story must be completed in the continuation. Story continuation is itself related to article extraction. That is, how successful are we at getting intact articles and at identifying all of the articles on a page?

Once articles have been identified, we need to categorize news stories. As a benchmark, we will compare our classification of news stories with the *The New York Times Index* and we will conduct studies on the utility of the categorizations to facilitate end-user access. Furthermore, users might want to navigate through a collection by following threads of news topics. We will extend our categorization techniques to determining those threads. In a modern context, similar questions have begun to be addressed by the Topic Detection and Tracking project [1].

Beyond using linguistic information to enhance the OCR, it is the core of the retrieval process and it can also be used for linguistic research. For instance, consider the following sentence from the *Brooklyn Eagle* for November 11, 1917:

Figure 5: Prototype of the user interface. Hits of the search string "british" are indicated in the map at the top left and where they appear in the page image.

> The United States Government will listen to no armistice proposal such as is announced to be the program of Lenine and the radicals who are leading the new Revolution in Russia.

In addition, journalistic styles changed in the time periods covered by these corpora. Specifically, early news stories tended to follow a chronological order while more modern stories follow the "pyramid" structure of providing several layers of detail. We will develop automatic methods for detecting those differing styles.

We anticipate that there will be substantial interest in tracking names of people and places which appear in the newspapers. This could, for instance, be used for genealogical research. We will examine the impact of OCR errors on named-entity identification.

2.3 Metadata and Mark-Up

Several layers of metadata are required for this project. Some of these are already established. For instance, there are extensive guidelines for cataloging newspapers [32]. However, there is a need for standard descriptions of both the logical structure and physical layout of newspaper content. These goals are consistent with the mark-up standards of the Text Encoding Initiative (TEI) [13]. There is also a need to provide metadata which will enhance the performance of the OCR techniques. For instance, characteristics of the newspaper style, such as the number of columns across the page, should be useful for the recognizer.

2.4 User Interface

A newspaper is a very complex and highly detailed object. Sophisticated interfaces will be needed to allow the users to navigate at several levels of granularity — within a single page and a single newspaper but also across issues of the newspaper, and eventually across collections of newspapers.

Moreover, we need to support two very different types of users. A corpus-developer interface will be built for users who need to inspect and update the OCR and zoning. It is likely that early versions of the software will have significant numbers of errors and that hand tuning will be required to obtain a useful corpus.

The end-user interface will allow researchers and students to access the collections. This interface will allow users to search terms and phrases derived from the OCR and then to view the search hits superimposed on the page image. As shown in Figure 5, we have been experimenting with the use of thumbnail maps to provide navigation and landmarks for the full-page image.

As suggested in the discussion of IR techniques, there may be many ways to navigate through large corpora: For instance, by tracking names of people or by following threads of news stories. We will develop graphical timeline interfaces [23] to help the user keep oriented.

3 Discussion

3.1 Richer Corpora

To evaluate the performance of the OCR system and the noise-removal algorithms we require a representative sample of the newspapers. We will create testing and training datasets of scanned images that will represent the various type of fonts, typesetting technologies, layout, degradation, and microfilm. Initially, we will create representative samples of the images for the collections of newspapers we will work with. Then we will create datasets that are representative of a larger population of newspapers. One microfilm roll contains approximately two weeks of a newspaper with about 40 pages in each issue. Microfilm scanners can scan one roll at 600 dpi in approximately one hour. The creation of the scanned image collection will take approximately three months at eight hours per day.

We will digitize the Negro Newspaper Collection [24], which is available from the Library of Congress, and consists of 180 microfilm rolls. For example, our proposed indexable newspaper collection of Reconstruction Era (1863–1877) African-American newspapers would allow historians quickly to check hypotheses and search for references to specific individuals.

We will also digitize a collection of Pennsylvania Dutch German-language immigrant newspapers for cross-language research. These include the *Reading Adler*, *Der Deustsche Porcupinen und Lancaster Anzeigs-Nachrichten Deutsche Porcupinen*, and *Neue Unpartyische Readingen Zeitung und Anzeigs-Nachrichten.* Beyond these niche collections, we will also index a large-city newspaper from page images because of the wide range of issues it covers. We have chosen the *The New York Times* because of its mix of national

and international stories as well as its coverage (which is, perhaps, spotty in some cases) of the diverse ethnic groups in New York City. We will scan 15 years of *The New York Times* (approximately 400 microfilm rolls) to study issues such as the history of women and the impact of advertisements on society.

3.2 History and Journalism Research and Education

We are working with historians whose research covers the topics to be examined in this project such as the African-American experience during the Civil War and Reconstruction [3], Women's rights in the late 19th and early 20th century [27, 28], and immigrant issues. Because we expect the historians to be consistent and dedicated users of these resources, we will focus on providing support for extended interaction. For instance, we will develop techniques for them to add annotations. The researchers will be monitored in the use of the tools. Furthermore, observation sessions will be established in which the users will be asked to to "think aloud" as they interact with the tools.

Several of these historians and journalists (e.g., [2]) have particular interest in undergraduate education. While professional historians will often painstakingly examine large quantities of primary research material, students have neither the time nor the patience for that type of research. Online search and access to primary historical sources should greatly facilitate students' use of that material.

The projects will teach students the skills involved in doing primary historical research, and the courses will include a survey of United States history, a survey of U.S. women's history, and specialized courses on gender and on progressive reform in the early 20th century. The graduate assistants will not only help determine the feasibility of various projects but will also teach students the technical aspects of searching digitized newspapers.

These educational tools will be formally evaluated by dividing the students into two sections. In one section, the students will do research using traditional methods; in the second section, they will be allowed to use the research prototype. The reports prepared by these two sets of students will be graded by independent readers and the quality compared across the two groups. The results will be disseminated to other historians through the Organization of American Historians.

Inevitably, there will be errors in the OCR and linguistic processing. While the presence of these errors is not ideal for the historians, the benefits of ease of indexing outweigh the limitations of the errors. Naturally, it is a goal of our OCR research to minimize errors, but we will also provide mechanisms for quality control and corpus revisions. Furthermore, we will freely report the frequency of various types of errors so that users can be aware of them.

3.3 Further Research Issues

Additional scientific questions we plan to address include:

- Statistical stratified sampling methods [5] will be used to create a representative image dataset with groundtruth, which will be used as a benchmark for performance evaluation of OCR systems.

- Because much of immigrant history is archived in non-English-language newspapers, we will adapt the interface and search tools for cross-language information retrieval. We will work with a Pennsylvania Dutch German-language newspaper collection which is also of interest to the project historians.

- Query-by-image-content: Many image regions may not be textual, such as logos and text in unusual fonts. We will allow image-based search for such regions using QBIC-like search tools [10].

- Beyond basic text processing, the system should also be able to identify and provide access to the wide variety of material included in newspapers such as advertisements. This might be useful, for instance, in studying the changing image of women portrayed in those advertisements.

4 Acknowledgments

Numerous people have advised us on this project. Dr. Maurine Beasly from the School of Journalism provided relevant references and identified problems that journalism historians study. Drs. Leslie Rowland, Robyn Muncy, and Whit Ridgway introduced us to problems in history for which newspapers are important research material. Dr. Charles Lowry, Marietta Plank, Evelyn Remaley, and Yvonne Carrignan from the University Library gave us valuable information regarding metadata and preservation. Dr. Douglas Oard from the College of Library and Information Services raised issues regarding cross-language retrieval problems in studying immigrant newspapers. Dr. Paul Smith of the Department of Statistics has advised us about statistical sampling. We would also like to thank Dr. Azriel Rosenfeld for his comments.

Song Mao provided us with images produced by his implementation of the segmentation algorithm. Jane Acheson assisted in the development of the user interface module.

References

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. In *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218, 1998.

[2] M. H. Beasley and K. T. Theus. *The New Majority: A Look at What the Preponderance of Women in Journalism Education Means to the Schools and to the Professions.* University Press of America, 1988.

[3] I. Berlin and L. Rowland, editors. *Families and Freedom: A Documentary History of African-American Kinship in the Civil War Era.* New Press, 1997.

[4] H. Bunke and P. S. P. Wang. *Handbook on Character Recognition and Document Image Analysis.* World Scientific, Singapore, 1997.

[5] W. G. Cochran. *Sampling Techniques.* Wiley, New York, 1977.

[6] W. B. Croft, S. M. Harding, K. Taghva, and J. Borsack. An evaluation of information retrieval accuracy with simulated OCR output. In *Proc. of the Symposium on Document Analysis and Information Retrieval*, pages 115–126, Las Vegas, NV, April 1994.

[7] D. Doermann. The indexing and retrieval of document images: A survey. UMD-CS-TR-3876.

[8] E. Emery and M. Emery. *The Press and America: An Interpretative History of the Mass Media.* Prentice-Hall, Englewood Cliffs, NJ, 1978.

[9] Encyclopedia Britannica. History of publishing, history of printing, history of typography, 1981.

[10] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz. Efficient and effective querying by image content. *Journal of Intelligent Information Systems*, pages 231–262, 1994.

[11] R. M. Haralick et al. UW-CDROM-I.

[12] R.M. Haralick, S.R. Sternberg, and X. Zhuang. Image analysis using mathematical morphology. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 9:532–550, 1987.

[13] S. E. Hockey. The ACH/ACL/ALLC Text Encoding Initiative: An overview. http://www-tei.uic.edu/orgs/tei/info/teij16.html.

[14] A. Jain and B. Yu. Document representation and its application to page decomposition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20:294–308, 1998.

[15] T. Kanungo. A morphological line removal algorithm. Forthcoming technical report.

[16] T. Kanungo. *Document Degradation Models and a Methodology for Degradation Model Validation.* PhD thesis, University of Washington, Seattle, WA, 1996. http://-www.cfar.umd.edu/~kanungo/pubs/phdthesis.ps.Z.

[17] T. Kanungo and R. M. Haralick. Character recognition using mathematical morphology. In *Proc. of the Fourth USPS Conference on Advanced Technology*, pages 973–986, Washington D.C., November 1990.

[18] T. Kanungo and R. M. Haralick. An automatic closed-loop methodology for generating character groundtruth for scanned images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21:179–183, 1999.

[19] T. Kanungo, R. M. Haralick, H. S. Baird, W. Stuetzle, and D. Madigan. Document degradation models: Parameter estimation and model validation. In *Proc. of the International Workshop on Machine Vision Applications*, pages 552–556, Kawasaki, Japan, December 1994.

[20] T. Kanungo, R. M. Haralick, and I. Phillips. Non-linear local and global document degradation models. *International Journal of Imaging Systems and Technology*, 5:220–230, 1994.

[21] T. Kanungo, G. E. Marton, and O. Bulbul. OmniPage vs. Sakhr: Paired model evaluation of two Arabic OCR products. In D. Lopresti and Y. Zhou, editors, *Proc. of the Conference on Document Recognition and Retrieval VI*, pages 109–120, San Jose, CA, 1999.

[22] R. Kasturi and L. O'Gorman. *Document Image Analysis*. IEEE Press, 1995.

[23] V. Kumar, R. Furuta, and R.B. Allen. Metadata visualization for digital libraries: Interactive timeline editing and review. In *Proc. of the ACM Digital Library Conference*, pages 126–133, 1998.

[24] Library of Congress Photoduplication Service. Negro newspapers on microfilm, 1953.

[25] S. Mao and T. Kanungo. Performance evaluation of zone segmentation algorithms. Forthcoming technical report.

[26] G. Matheron. *Random Sets and Integral Geometry*. Wiley, New York, 1975.

[27] S. Michel and R. Muncy. *Engendering America: A Documentary History*. McGraw-Hill, New York, 1998.

[28] R. Muncy. *Creating a Female Dominion in American Reform 1890-1935*. Oxford, New York, 1994.

[29] L. O'Gorman. The document spectrum for page layout analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15:1162–1173, 1993.

[30] D. Pankow. Type and typesetting. In *Grolier's Encyclopedia*. Grolier's Encyclopedia.

[31] T. Pavlidis and J. Zhou. Page segmentation and classification. *CVGIP: Graphical Models and Image Processing*, 54:484–496, 1992.

[32] C. Sagendorf and D. Moore. CONSER cataloging manual: Module 33, 1996.

[33] J. Serra. *Image Analysis and Mathematical Morphology*. Academic Press, New York, 1982.

[34] S. E. Sully. JSTOR: An IP practitioner's perspective. *D-Lib Magazine*, January 1997. http://www.dlib.org/dlib/january97/01sully.html.

[35] University of Virginia. Valley of the Shadow. http://jefferson.village.virginia.edu/-vshadow2/choosepart2.html.

[36] F. M. Wahl, K. Y. Wong, and R. G. Casey. Block segmentation and text extraction in mixed text/image documents. *Computer Graphics and Image Processing*, 20:375–390, 1982.