

The MANICURE Document Processing System

Kazem Taghva, Allen Condit, Julie Borsack,
John Kilburg, Changshi Wu, and Jeff Gilbreth

Information Science Research Institute
University of Nevada, Las Vegas

ABSTRACT

MANICURE is a document processing system that provides integrated facilities for creating electronic forms of printed materials. In this paper the functionalities supported by MANICURE and their implementations are described. In particular, we provide information on specific modules dealing with automatic detection and correction of OCR errors and automatic markup of logical components of the text. We further show that the various text formats produced by MANICURE can be used by web browsers and/or be manipulated by search routines to highlight the requested information on document images.

Keywords: OCR, Document Processing, Automatic Error Correction, Post Processing, Automatic Markup

1. INTRODUCTION

The history of information retrieval may go back as far as 1948. According to Maron,¹ 1948 signifies three important events. The first is the printing of the book *Cybernetics*² by Norbert Wiener in which it is proposed that by information processing, one can mechanically interpret intelligent behavior. The second is the work of Claude Shannon which states that information could be studied in a quantitative way.³ The third one is the technological advancement of building general purpose computers. These events subsequently led to the theory that we can build information retrieval systems that can store intelligent document representations and deliver relevant documents in response to user's queries. For this theory to flourish, we must start with a *functional* electronic version of a collection and build an intelligent representation of it for storage and retrieval. What exactly does "functional" mean? In our view, the word "functional" implies nearly correct text with well defined logical structure, and meta-knowledge (e.g. journal name, date of publication, publisher, etc.) about the documents.

For historical reasons, it seems that no scientific collection is fully available in a functional electronic form, although some may be available in printed functional form. In addition to the obvious historical reason that printing predates computer technology, it is clear that most of the authors (or organizations) did not give much thought to keeping the electronic version for storage and retrieval. For example, the information retrieval community (ACM-SIGIR) does not have electronic versions available of its annual symposium proceedings, which began in 1978. Of course, copyright issues will also make it hard to get an electronic version of any scientific collection even if it were available in a functional electronic form. Putting the copyright issues aside, we strongly believe OCR technology will play a very important role in constructing functional electronic forms of various scientific collections.

In this paper, we report on our system MANICURE (Markup ANd Image-based Correction Using Rapid Editing) that is built to facilitate the task of constructing the functional electronic forms of document collections. This system is designed to take advantage of document characteristics such as word forms, geometric information about the objects on the page, and font and spacing between textual objects to mark the logical structure of the document. In addition, the system automatically detects and corrects OCR spelling errors by using dictionaries, approximation matching, the knowledge of typical OCR errors, and frequency and distribution of words and phrases in the document. MANICURE in its automatic mode can produce functional forms of documents which are good for most text analysis and retrieval applications. For applications requiring close to 100% word accuracy and proper markup of important components such as author, title, and abstract, MANICURE in its semi-automatic mode coordinates image and text to speed up the process of correction and markup.

Further author information: Email: taghva@isri.unlv.edu; WWW: <http://www.isri.unlv.edu/info/tr/>

2. PRELIMINARIES

MANICURE is specifically designed to prepare text collections from printed materials for information retrieval applications. In this capacity, depending on the application, requirements on accuracy and text structure vary. In what follows, we will list important applications for which MANICURE is designed.

- MANICURE will try to distinguish between the actual content of the document and the superficial information which is part of its presentation. For example, the sequence of words in the printed document identifying the journal in which it appeared is presentational rather than being a part of a document's content. This extra text (and extra non-alphanumeric characters such as end-of-line hyphenation) can be both helpful and harmful. It is helpful in the sense that it identifies meta-knowledge about the document such as date of publication, but it could also be harmful if it is not properly removed or managed, such as with end of line hyphenation; this extra text will also clutter the retrieval system's index with useless information.
- Recent studies on effects of OCR errors on retrieval⁴⁻⁶ have pointed out that certain advanced functionalities of information retrieval systems, such as ranking, are not robust enough to overcome OCR errors. It is our view that the more advanced retrieval techniques and applications require a higher character accuracy rate and less graphic text. MANICURE can be used in both automatic and semi-automatic modes to produce text with higher levels of accuracy.
- Associated with each document is a list of structured data known as the *header*. The header generally contains information such as the author, the date of publication, document control number, title, and so on. Some of this information is already in the text of the document (such as the title) and some is added for the sake of record keeping (such as a document control number). The bigger the header the more it costs. The header can be used by the information retrieval system to produce a list of documents for very specific queries, such as "retrieve papers written by Olin Fnard." By properly marking some of the structured data, MANICURE provides an environment in which a part of the header can be extracted automatically, or in the case of structured based retrieval systems,⁷ can be manipulated by the retrieval engine itself.
- The logical structure of the text can be used in many retrieval applications. For example in⁸⁻¹² the individual sentences, paragraphs, sections, and section titles are analyzed as a part of the solution to these particular applications. MANICURE builds a logical representation for each document in which all these objects are marked.
- The text from printed materials can be produced in different formats for different uses. For example, the Hypertext Markup Language (HTML) (derived from SGML¹³) is a common format used for marking up documents for viewing by World-Wide Web browsers such as Chimera.¹⁴ MANICURE, in addition to providing text in HTML format, also outputs documents in another SGML-based format with detailed information about logical structure, font, subscript, superscript, and geometry information. One use of this format is to enable retrieval systems to search on the text of a document and highlight hits on the original page images.
- As MANICURE strives to mark and keep track of all the information in the document, its representation in MANICURE, in essence, gets closer to its original printed representation. This will add to our understanding of documents through markup and MANICURE can be used as an experimental environment for studying document understanding.

3. MANICURE DESIGN

MANICURE is composed of four modules and an OCR front end. The four modules consist of the parser (`doc_parse`), the logical document tagger (`autotag`), the post processing system (`ppsys`), and the semi-automatic user interface (`rummage`). In the next sections, we will explain the role of each component and relate the fundamental ideas underlying their design.

3.1. Doc_parse

The parser is an OCR dependent module which extracts necessary information from the output of the OCR device to build a physical representation of the document. The parser builds a physical representation in the form of a hierarchical tree. The leaves of the tree contain the recognized text strings together with such attributes as font information and the location of the string on the image. The interior nodes of the tree keep information pertaining to the lines, zones and pages of the input document. Since there are no formal standards that govern the output of OCR systems, this parser would need to be modified if different devices were used. If a standard such as DAFS¹⁵ becomes a reality, then a single parser could be used. Currently, the physical representation is built according to an SGML DTD and is stored as an SGML file.

3.2. Autotag

Autotag is the most sophisticated and detailed part of the MANICURE system. It starts with the physical representation produced by the parser and builds a tree representing the logical structure of the document. The leaves of this tree represent words forming the content of the document. Superfluous characters, such as end of line hyphenations, are properly managed and some typographical text, like running headers and footers, are removed. The interior nodes contain information regarding sentences, paragraphs, sections, and section titles. Autotag also extracts and marks structured data such as the title and author. This representation is also stored as an SGML file. For more information on Autotag, readers are referred to.¹⁶

3.3. PPSYS

The post processing system is devised to detect and correct OCR errors through approximation matching, by consulting the devices most common confusions, and by extracting knowledge from the complete document. Since the complete document has already been processed, information about its content can be used to correct it; an OCR device does not have access to all this information when it is processing single document pages. For example, a word on the first page may have been misrecognized but another occurrence of the same word on the next page may have been correctly recognized. By having the entire document available, we have more information for correcting the misrecognized occurrence. These ideas were originally applied in our preliminary version of the post processing system.¹⁷ The module included in MANICURE exploits analogous algorithms together with more advanced means of approximation matching. Further, the PPSYS is now a standalone system which does not require an IR index; document parsing and index information are self-contained.

The post processing module builds an inverted file from Autotag's output consisting of a document's words, their frequencies, and their proximities to other words in the document. Dictionaries and special recognizers¹⁸ are used to mark each word as correctly recognized or misspelled. The module then generates statistical phrases to correct misspellings. This correction routine is followed by approximation matching using word frequency information and OCR error information (e.g. `rn` for `m`). Preliminary evaluation shows that, depending on document quality and length, correction rate of the PPSYS lies between 15% and 50% of all misspellings in the document.

3.4. Rummage

Rummage can be invoked for semi-automatic inspection and correction of OCR errors and markup. This component can be considered the interface to the document as processed by the previous modules. A document's images together with the document's text are unified using the markup of Autotag. Figure 1 illustrates the tagging of the title and its presentation in Rummage. Further, the PPSYS has corrected some portion of the misrecognized words. Those that could not be automatically corrected can be highlighted and easily corrected using Rummage. Since each misspelling is tagged by the post processing system, Rummage can quickly run through the document, highlighting each misspelling. The user is given several options for correction: 1) select the correct misspelling from the provided list by a mouse click, 2) Re-type the correct word to replace the misspelling, or 3) pass over the word by clicking `Next`. Figure 2 shows the features of semi-automatic correction using Rummage.

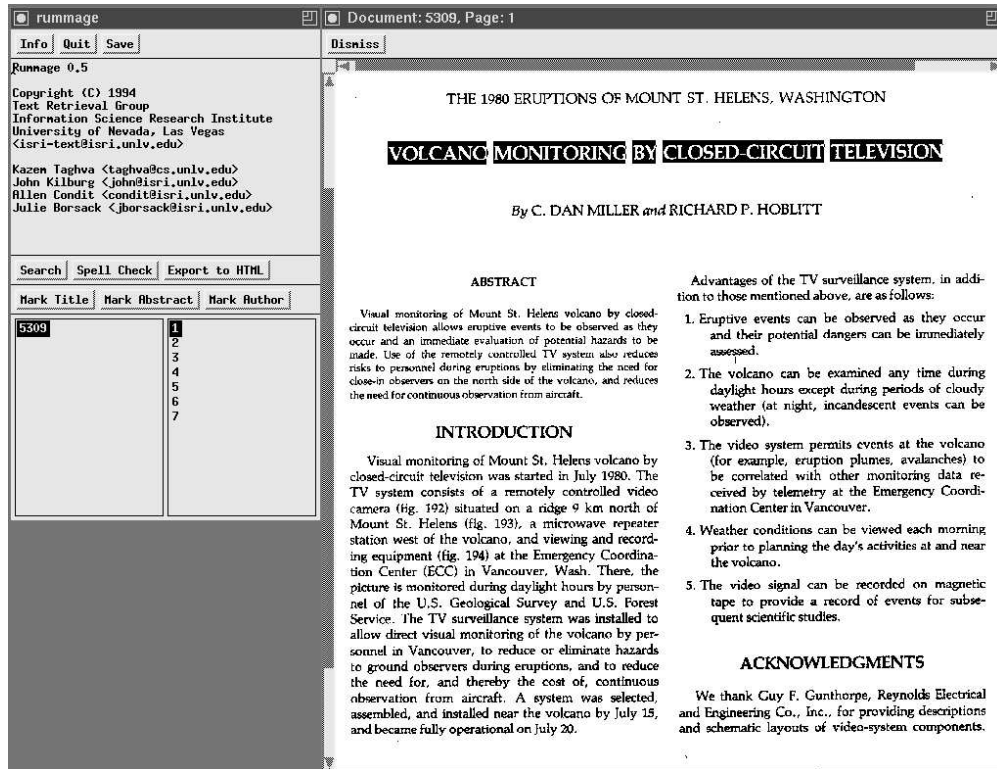


Figure 1. Title tagging as shown in Rummage.

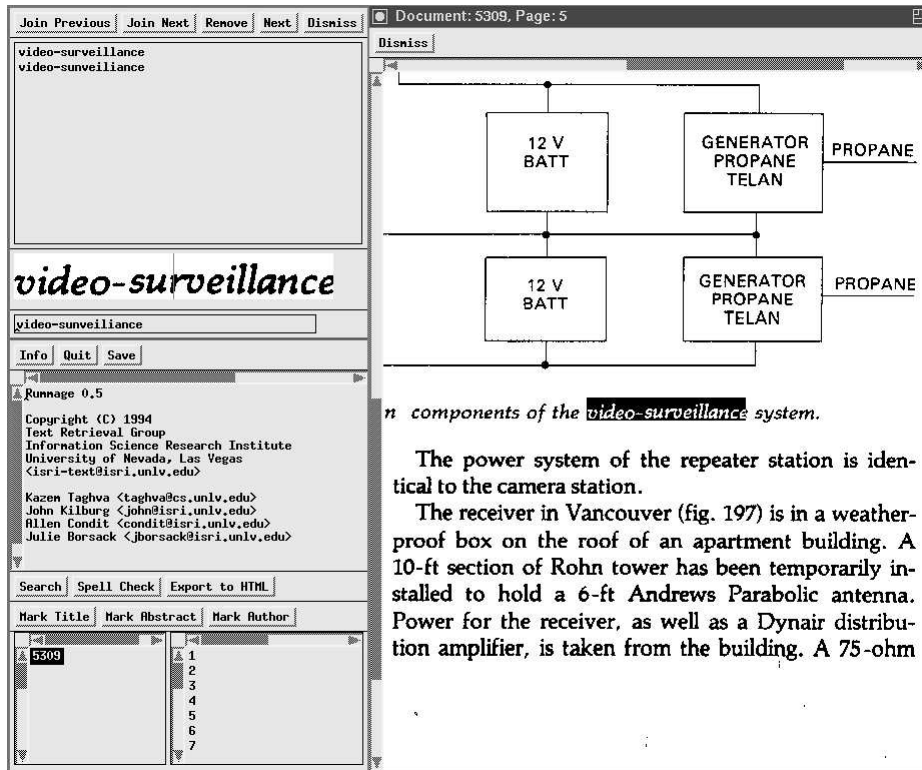


Figure 2. Semi-automatic correction features in Rummage.

4. CONCLUSION AND FUTURE WORK

The MANICURE system has evolved over the last two years and is founded upon our research of OCR and IR interaction. For example, removal of typographical text and automatic correction of misspellings are done to improve the quality of the text and overcome some of the problems we have encountered with the ranking algorithms. We strongly felt that the display of recognized documents (in the form of OCR text) would undermine the user's confidence in the system. Therefore, we designed MANICURE to provide different formats for the output text. Although HTML format can be used for browsing, the SGML-based format prepared by Autotag and used by Rummage for display, can solve the display problem in a very general way.

We are currently adding new routines to MANICURE. The first could be considered a preprocessor—a way to verify document quality before submission to the OCR device. This routine will provide some functionalities such as image rotation and simple zoning to assist in image quality control. Second, like the PPSYS where 100% automatic spelling correction is rare, we believe 100% correct markup of heterogeneous technical documents will be difficult to achieve. Thus, we believe a semi-automatic version of Autotag, guided by manual markup, would be beneficial.

REFERENCES

1. M. Maron, "Probabilistic approaches to the document retrieval problem," in *Proc. 13th Intl. ACM/SIGIR Conf. on Research and Development in Information Retrieval*, pp. 102–107, ACM Press, July 1990.
2. N. Wiener, *Cybernetics: Or Control and Communication in the Animal and the Machine*, John Wiley, New York, 1948.
3. C. Shannon, "The mathematical theory of communication," *Bell System Technical Journal*, pp. 379–423, 623–656, July and October 1948.
4. K. Taghva, J. Borsack, A. Condit, and S. Erva, "The effects of noisy data on text retrieval," *J. American Soc. for Inf. Sci.* **45**, pp. 50–58, January 1994.
5. K. Taghva, J. Borsack, and A. Condit, "Results of applying probabilistic IR to OCR text," in *Proc. 17th Intl. ACM/SIGIR Conf. on Research and Development in Information Retrieval*, pp. 202–211, (Dublin, Ireland), July 1994.
6. K. Taghva, J. Borsack, and A. Condit, "Effects of OCR errors on ranking and feedback using the vector space model," *Inf. Proc. and Management* **32**(3), pp. 317–327, 1996.
7. I. A. Macleod, "A query language for retrieving information from hierarchic text structures," *The Computer Journal* **34**(3), pp. 254–264, 1991.
8. M. Fuller, E. Mackie, R. Sacks-Davis, and R. Wilkinson, "Structured answers for a large structured document collection," in *Proc. 16th Intl. ACM/SIGIR Conf. on Research and Development in Information Retrieval*, pp. 204–213, ACM Press, (Pittsburgh, PA), June 1993.
9. M. A. Hearst and C. Plaunt, "Subtopic structuring for full-length document access," in *Proc. 16th Intl. ACM/SIGIR Conf. on Research and Development in Information Retrieval*, pp. 59–68, ACM Press, (Pittsburgh, PA), June 1993.
10. G. Salton, J. Allan, and C. Buckley, "Approaches to passage retrieval in full text information systems," in *Proc. 16th Intl. ACM/SIGIR Conf. on Research and Development in Information Retrieval*, pp. 49–58, ACM Press, (Pittsburgh, PA), June 1993.
11. J. P. Callan, "Passage-level evidence in document retrieval," in *Proc. 17th Intl. ACM-SIGIR Conf. on Research and Development in Information Retrieval*, pp. 302–310, (Dublin, Ireland), July 1994.
12. R. Wilkinson, "Effective retrieval of structured documents," in *Proc. 17th Intl. ACM-SIGIR Conf. on Research and Development in Information Retrieval*, pp. 311–317, (Dublin, Ireland), July 1994.
13. C. F. Goldfarb, *The SGML Handbook*, Oxford University Press, 1990.
14. J. Kilburg, "Chimera: An athena-based world-wide web browser," Tech. Rep. 95-01, Information Science Research Institute, University of Nevada, Las Vegas, March 1995.
15. D. I. U. D. A. Program, "Document attribute format specification (DAFS)," May 1993.
16. K. Taghva, A. Condit, and J. Borsack, "Autotag: A tool for creating structured document collections from printed materials," Tech. Rep. 94-11, Information Science Research Institute, University of Nevada, Las Vegas, December 1994.
17. K. Taghva, J. Borsack, B. Bullard, and A. Condit, "Post-editing through approximation and global correction," *International Journal of Pattern Recognition and Artificial Intelligence* **9**(6), pp. 911–923, 1995.

18. K. Taghva and J. Gilbreth, "Recognizing acronyms and their definitions," Tech. Rep. 94-07, Information Science Research Institute, University of Nevada, Las Vegas, November 1994.