

**A PRELIMINARY REPORT ON UNLV/GT1:
A Database for Ground-Truth Testing
in Document Analysis
and Character Recognition**

T. A. Nartker

*Information Science Research Institute
University of Nevada, Las Vegas
Las Vegas, NV 89154*

R. B. Bradford

*Science Applications International Corporation
1710 Goodridge Dr.
McLean, VA 22102*

B. A. Cerny

*Office of Civilian Radioactive Waste Management
United States Department of Energy
1000 Independence Av. SW
Washington, D.C. 20585*

ABSTRACT

We describe a computer database being developed at the University of Nevada, Las Vegas to support experiments in the recognition and analysis of information from printed documents. The history and economic significance of the database are discussed.

It is a page-oriented database of mostly technical documents. Approximately 9300 pages are currently on-line. Methods of access are described. A set of software tools has been developed which automate much of the drudgery of performing experiments with optical character recognition (OCR) systems.

UNLV plans to encourage each succeeding researcher to add value to the database. The authors believe that GT1 will become an

increasingly valuable standard for evaluating systems and an important tool for research in document analysis. At the same time, the experimental tools described can be utilized to automate experiments with new ground-truth databases as they are added.

Introduction

In a recent paper titled "At the Frontiers of OCR" [1], G. Nagy states that "It is time for a major change of approach to character recognition research. The traditional approach, focusing on the correct classification of isolated characters, has been exhausted." Professor Nagy continues: "demonstration of the superiority of a new method under operational conditions requires large experimental facilities and databases" which are "beyond the resources of most researchers."

In this paper, we describe a database of printed documents, and a set of supporting research tools, which have been established at the University of Nevada at Las Vegas to address this problem. Although the tools and the database are installed on a file server in Las Vegas, the server is connected to the Internet and thus is accessible to researchers anywhere who can use this network.

History of Databases for OCR Research

The system we describe is believed to be the first major attempt to provide an openly accessible testing environment for OCR research. Previous efforts have been made to establish standard data sets for OCR testing. The great majority of these are character-oriented and have consisted of handprinted data. None are known to have been incorporated into a publicly accessible automated test environment.

Perhaps the earliest data set publicly distributed for OCR research was that of Highleyman in 1963 [2]. That pioneering effort included 12 row by 12 column binary matrix images of handprinted alphanumeric characters and machine-printed numerals. The handprinted data consisted of 50 samples of each of 36 alphanumeric characters, produced by 50 different people. The

machine-produced data consisted of 50 samples of each of the 10 numerals from an IBM407 line printer. The database was distributed on punch cards. Munson [3], produced an additional data set that was widely used in the early days of OCR research. This data set consisted of 12,760 samples of the 46 characters used in the FORTRAN II programming language. These handprinted characters were represented by a 24 row by 24 column binary matrix.

In Japan, considerable emphasis has been placed on the use of standard data sets for OCR testing [4]. The Electrotechnical Laboratory has developed nine such standardized databases; eight for handprinted and one for machine-printed characters. The later databases are relatively large. ETL-8 contains 152,960 characters in 160 data sets. Each data set contains 881 elementary Kanji characters and 75 Hirakana characters [5]. ETL-9 consists of 607,200 characters in 200 data sets. Each data set contains 2965 Kanji characters and 71 Hirakana characters [6].

Some of the largest databases for OCR research have been developed to support R&D for postal applications [7 and 8]. A database of 17,000 handwritten digits extracted from zip codes written on dead letter envelopes, is available from Concordia University (see C. Y. Suen, et. al. [9]). The National Institute of Standards and Technology has played an important role in developing and promulgating such databases [10 and 11].

Background

The database we describe has its origin in national efforts to deal with radioactive waste management problems. The Nuclear Waste Policy Act of 1982 set forth the responsibilities which different agencies were to accept in dealing with nuclear waste disposal problems in the United States. Among the tasks required of the U. S. Department of Energy, is the construction of a computer database system to support the site licensing proceedings of the Nuclear Regulatory Commission. The basic function of the needed system is to make all pertinent information available to all parties to the

licensing proceedings. This system is called the Licensing Support System (LSS).

The functional requirements, data requirements, and conceptual design of the LSS system were defined in a set of reports [12,13,14] published in 1988. The specifications call for all documents contained in the system to be stored both in binary image form and in ASCII text form. The ASCII text files are provided to make possible computer (full-text) searching of the database but users in general will view the image form of the pages found. Initial estimates of the amount of data to be included are of the order of 42 million pages [14]. The initial cost estimate for constructing and operating this system was \$200,000,000 over a ten year life [15].

The two main components of the LSS system are a document capture sub-system and a search and image (or storage and retrieval) sub-system. It is a surprising fact that, even though many documents produced today are originally produced on a word processor, the ASCII file corresponding to the final document is seldom available. Indeed, this was the case with LSS documents. After considerable study it was determined that the most practical means of obtaining ASCII versions of LSS documents was either to use OCR techniques or to manually rekey a hard copy version of the documents. Thus, the cost of capturing data for this system was determined to be mostly dependent upon the cost of capturing documents via OCR or upon the cost of manually rekeying the document (i.e. whichever is less).

To evaluate the cost of capturing LSS documents, and to evaluate other design parameters, DOE contractors were required to develop a realistic prototype of data capture operations and to create a prototype document database. This study confirmed that the cost of correcting OCR errors from good quality documents was less than the cost of manual rekey, even at the cheapest labor rate available. For poor quality documents, manual rekey was cheaper. Detailed results of this study have been published, see [16 and 17]. One main conclusion which resulted from this study was that a large fraction of the DOE cost to build the LSS system would be the cost of correcting the OCR output. In creating large full-text databases, recognition accuracy is the key cost factor.

The LSS Prototype Database

A set of approximately 2600 actual LSS documents (approximately 104,000 pages) was selected for use as a prototype LSS database. Some effort was expended to insure that the composition of this set reflected the composition of the final LSS database. DOE contractors were required to set up a realistic data capture environment to record and track document pages, to capture page images, and to produce ASCII text. A subset of the documents was processed by each of four different contractor facilities. Each was required to produce ASCII text corrected to 99.8% accuracy on a character basis.

To satisfy DOE requirements, some contractors processed all pages through a Calera RS9000 and provided a manual-key cleanup operation to achieve the accuracy required. Other contractors chose a manual rekey operation to obtain a first text file and then provided a manual-key cleanup operation to achieve the accuracy required. In all cases, the cost and efficiency of the conversion was monitored carefully. The information gained in building this prototype provided the important cost and operating parameters needed to complete a set of final design documents for the LSS system, see [18,19].

After careful study of the cost of operating a large document capture system, it became clear that there were many opportunities for reducing the cost of data conversion. Also, there was need for special tools to improve the usefulness of text-retrieval technologies for large (multi-gigabyte sized) databases.

As part of a cooperative agreement between UNLV and the Department of Energy, arrangements were made to transfer the LSS prototype database to UNLV. UNLV has agreed to prepare a version of this data to serve as a basis for technology improvement studies in OCR and as a tool for more general research into Document Analysis and Text-Retrieval systems.

The GT1 Database

For several reasons, an appropriately configured version of the LSS prototype database can be an ideal vehicle for testing document analysis technologies. First, because all prototype documents are actual LSS documents and because they were selected from the beginning to reflect the composition of the final LSS system, they provide a realistic definition of DOE requirements for OCR technology. In fact, given the percentage of non-original copies and the variety of document types and subject areas, they represent a moderately difficult set of test documents.

Second, the availability of page images collected under operational conditions paired with the corrected ASCII text for each page, makes this set of images/ASCII-text useful for ground-truth testing. This is the first such database at UNLV, hence the name GT1. Finally, because the size of the set is large and the source of documents is diverse, and because of the variation in font type, style, size, layout, content, etc., this database is an ideal platform for testing and comparing document character recognition and page analysis technologies.

The GT1 test-bed contains three components. First, all original hard-copy documents chosen for the LSS prototype tests, indexed by a unique document identifier, are stored in a hard-copy library on the UNLV campus and are available for manual inspection or rescan. Second, the binary image of each page, indexed by a unique page identifier, is available from an image-database stored on magnetic disk. Third, the ASCII text for each page, also indexed by page identifier, is stored on disk in an associated truth-database.

There are approximately 104,000 pages (and approximately 2600 documents) in the hard-copy library. Approximately 125 megabytes of ASCII text are printed on these pages. We have chosen to title this paper "Preliminary" because only a fraction of these are available on-line at this time.

The process of checking and converting magnetic tape files is very labor intensive. The current image-database contains 9,278 page images and the truth-database about 10 megabytes of associated ASCII text.

Composition of GT1

1. Original (Hard Copy) Documents

Approximately 84% of the documents in the GT1 library are scientific reports or technical material. Table 1 shows the percentage of documents by source document type estimated from a sample of 241 documents.

Document Source	Percent
Journal article	14
Conference paper	10
Technical report	35
Book chapter	25
Total technical documents	84
Legal documents	5
Business letters or memos	8
Other documents	3
Total	100

Table 1. Percentage of documents by document source.

Approximately 5% are legal documents and 8% letters or memos. The remaining are widely distributed including patents, theses, copies of overhead slides from presentations, and even a guide book for a field trip. Table 2 shows the subject breakdown of the technical documents estimated from the subjects covered in the 241 documents sampled. The documents identified as "mathematical modelling", were distributed over the other subject areas shown.

We estimate that less than 50% of these documents are original copies. Most of the copies are of fairly high quality. Of course, we have no way of measuring generation, but a careful

inspection supports an estimate of more than 60% original or early generation copies. Overall, we have observed a considerable range of document quality.

Document subject	Percent
Geology	6
Geophysics	7
Geomorphology	4
Economic Geology/Mineralogy	15
Hydrogeology/Geochemistry	19
Environmental Biology/Ecology	3
Tectonics	10
Vulcanology	25
Mathematical modelling	11
Total	100

Table 2. Percentage of technical documents by subject area.

2. Image-Database

All page images were produced at 300 dpi on either a Ricoh or Fujitsu scanner. Although we have no threshold or scanner setting data, all images were produced by trained operators in a realistic document capture setting [16]. Our manual comparison of original hard copy pages with the corresponding image indicates that the images are fairly accurate representations of the original page.

Each image in the image-database has been converted to TIFF format and compressed using a CCITT group 4 algorithm. Page entries in the image-database contain the condensed image as well as an ordered sequence of zone coordinates. Zones are rectangular regions on the page image. The zone sequence preserves the reading order for the page. In the current version of the database, only "Text" zones are included. Text zones are rectangular regions containing the "main body" text of the document.

Page properties are stored in an Ingres relational database. They have been chosen to permit easy access to pages containing interesting page recognition or page analysis problems.

For example, users can select pages by the minimum number of columns and by the maximum number of columns on the page. That is, if the maximum number of columns = 1, then the page contains a single column of information. If the maximum number of columns is not equal to the minimum number of columns, then portions of the page contain different numbers of columns of information.

Page Contains	Access
"main body" text	(yes/no)
non-ASCII characters	(yes/no)
non-plain styles	(yes/no)
superscripts/subscripts	(yes/no)
footnotes	(yes/no)
tables	(yes/no)
equations	(yes/no)
graphics	(yes/no)
photos	(yes/no)
maps	(yes/no)
graphs	(yes/no)
artifacts	(yes/no)
page orientation	Portrait/Landscape

Table 3. Page property indices to the image-database.

3. Truth-Database

The Truth-Database contains a computer readable version of the information printed on each page. The truth-database contains one block of correct ASCII text for each text-zone defined in the image database. This block contains a sequence of ASCII character strings. Each string contains the correct ASCII characters for the corresponding line on the image. That is, we preserve the

correspondence between lines of text on the image and ASCII strings in the truth-database. We do not, however, preserve multiple blank characters, indentation information or paragraph information.

As we mentioned, the truth-database contains only the ASCII characters from the "text" zones in the image database. All non-ASCII characters in these zones are currently represented by a tilde character (~) in the string of correct ASCII text.

Future versions of the truth-database will include representations for the non-ASCII characters in the document. Also, future versions of the image database will include zones with tables, equations, or other kinds of information. The corresponding entry in the truth-database for these zones will contain a computer readable representation of this information. Table 4 shows candidate truth-information which would enhance the value of the GT1 database. We expect each new thesis project at UNLV to add truth information for one or more of these categories.

In order to enter a new page-image and page-text pair into our database, a trained operator must inspect each, side by side, on a display to verify correspondence; must define and order the set of text zones; and also determine the page properties for inclusion in the Ingres database.

During the coming year, we plan to prepare a separate version of the GT1 ASCII text, on a separate file server, for text-retrieval research and testing.

- **Tabular zones, especially numeric tables**
- **Type size & style information (bold, italic, etc.)**
- **Non-ASCII characters**
- **Subscripts and superscripts**
- **Mathematical equations**
- **Chemical equations**
- **Graphic objects (i.e. maps, graphs, etc.)**
- **Preservation of page format information in a manner which supports comparison of the page analysis features of different systems.**

**Table 4. Candidate truth-information for Addition
to the Truth-database.**

Experimental Environment

1. Experimental Equipment

The GT1 experimental environment is shown in Figure 1. The image-database and the truth-database are stored on disk on a Unix file server which is connected to a set of PC-host computers via an ethernet backbone. All OCR devices are operated from a Unix workstation under program control. The PC's act as an image/text buffer and as an intelligent controller for OCR devices that require a PC host.

The current set of OCR devices includes the Caere Omnipage Board, the Toshiba Express Reader, the Calera RS9000, the Kurzweil 5200, and the Recognita Plus software system. Pages can be selected on the server, submitted to any or all of the devices, and the produced text from each device returned in the user's directory. The key feature is that all devices are operated "under program control."

UNLV plans to install and evaluate one copy of each different OCR device available. Experimenters will be able to submit one or more of their own algorithms to this system for evaluation or comparison with any other on-line device.

2. Experimental Tools

A set of software tools have been created to automate most of the labor intensive tasks involved in conducting recognition experiments. These tools can be classified into "front-end" selection tools, tools to submit images to and receive text from OCR devices, and "back-end" tools for comparing and analyzing the output text.

a. Front-end selection of pages/images

To simplify access to page data, an Ingres table containing page attributes can be queried. An example query might be:

```
select doc_id, pagenum from page
```

where mainbody_text = 'y' and graphics = 'n' and maxcolumns = 1

This query selects pages with no graphic content and a single column of mainbody text.

Another important front-end tool is a zoning utility which allows users to define an ordered sequence of rectangular regions (i.e. zones) on an image. Thus, images selected by query can be viewed one by one and rezoned with the zoning utility tool if desired.

b. Vendor-independent interface to Devices

To standardize access to all OCR devices, we have created a vendor-independent interface. This interface provides a single shell command, executable from a Sun workstation, which will submit a page-image to an OCR device, addressed by device name, and will return the output text-file. The command syntax is as follows:

```
ocr [-D device] [-Z zonefile] imagefile [outputfile]
```

where device = caere, calera, etc. and
-Z identifies a zone file created using the zoning utility.

New devices, including commercial OCR products and research prototypes, will be accessible from this interface as they become available.

c. Back-end processing of text

To automate comparison of device output with the correct text, a tool to synchronize and display text streams has been provided. The synchronization tool will count the number of character errors.

Current Use of the GT1 Database

The initial use of the GT1 database and experimental facilities is to thoroughly test all existing OCR technologies on a large subset of the GT1 pages. Meaningful comparisons of the capabilities of existing devices are currently not available. Projects are underway to measure the overall accuracy, sensitivity to noise and skew, sensitivity to broken and touching characters, automatic zoning capabilities, and other characteristics of the currently installed devices.

An initial study which attempts to characterize the residual problems of contemporary OCR devices is reported in [20]. This study is based on tests with 278,786 characters (on 240 pages) from GT1. A series of reports is planned, based on tests involving 2 to 10 megabytes of truth-data from GT1.

Another area of investigation planned is to study the properties of GT1 itself. Because this database was chosen to reflect the composition of the final LSS system, it provides a good definition of the DOE needs for OCR technology. Information concerning the adequacy of current technology in such a real world setting is normally hard to acquire.

Future Plans

Finally, UNLV plans to extend this experimental environment by adding other ground-truth databases. In addition to the items shown in Table 4, UNLV plans to add perfect-image databases¹, fax databases, newspaper & magazine databases, and foreign language databases to test devices which recognize foreign language text.

1 i.e. Databases of text-images generated by the T_EX document preparation system.

References

- [1] G. Nagy, "At The Frontiers Of OCR", **Proceedings IEEE**,

Spring 1992.

- [2] W. H. Highleyman, "Data for Character Recognition Studies," **IEEE Transactions on Electronic Computers**, April 1963, pp. 135-136.
- [3] J. H. Munson, "Experiments in the Recognition of Hand-printed Text: Part I - Character Recognition," In **Proceedings, Fall Joint Computer Conference**, December 1968, pp.1125-1138.
- [4] K. Toraichi, R. Mori, I. Sekita, K. Yamamoto, and H. Yamada, "Handprinted Chinese Character Database," In **Computer Recognition and Human Production of Handwriting**, World Scientific, 1989, pp. 131-148.
- [5] S. Mori, K. Yamamoto, and M. Yasadua, "Research on Machine Recognition of Handprinted Characters," **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 6, (4), July 1984, pp. 386-405.
- [6] K. Yamamoto, H. Yamada, T. Saito, and I. Sakaga, "Recognition of Handprinted Characters in the First Level of JIS Chinese Characters," **8th International Conference on Pattern Recognition**, Paris, France, October 27-31, 1986, pp. 570-572.
- [7] R. Kalberg, H. Essink, and G. Quinto, "Automatic Reading of Handwritten Postal Codes in the Netherlands," **U. S. Postal Service Advanced Technology Conference**, Washington, D. C., November 5-7, 1990, pp. 987-1001.
- [8] P. Farder, D. Hepp, B. Roster, and T. Peurach, "Pipelined Systems for Recognition of Handwritten Digits in USPS Zip Codes," **U. S. Postal Service Advanced Technology Conference**, Washington, D. C., November 5-7, 1990, pp. 539-548.

- [9] C. Y. Suen, C. Nadal, T. A. Mai, R. Legault, and L. Lam, "Recognition of Totally Unconstrained Handwritten Numerals Based on the Concept of Multiple Experts," **Proceedings of the International Workshop on Frontiers in Handwriting Recognition**, Montreal, Canada, April 2-3, 1990, pp. 131-143.
- [10] M. L. Greenough and R. M. McCabe, "Preparation of Reference Data Sets for Character Recognition Research," **NBS Report #NBSIR-75-746**, 30 June 1975.
- [11] C. L. Wilson and M. D. Garris, "Handprinted Character Database," **NIST Special Database 1, HWDB**, 18 April 1990.
- [12] DOE Office of Civilian Radioactive Waste Management, "Licensing Support System - Preliminary Needs Analysis," Feb. 1988.
- [13] DOE Office of Civilian Radioactive Waste Management, "Licensing Support System - Preliminary Data Scope Analysis," Mar. 1988.
- [14] DOE Office of Civilian Radioactive Waste Management, "Licensing Support System - Conceptual Design Analysis," May 1988.
- [15] DOE Office of Civilian Radioactive Waste Management, "Licensing Support System - Benefit-Cost Analysis," July 1988.
- [16] L. A. Dickey, "Operational Factors in the Creation of Large Full-text Databases," **DOE Infotech Conference**, Oak Ridge, TN, May 1991.
- [17] R. B. Bradford, "Technical Factors in the Creation of Large Full-text Databases," **DOE Infotech Conference**, Oak Ridge, TN, May 1991.

**[18] SAIC, "LSS Data Capture Station: Design Document,"
Final Report, DOE Contract #DE-AC01-87RW00084,
Nov. 1990.**

- [19] **SAIC, "LSS Search and Image System: Design Document,"
Final Report, DOE Contract #DE-AC01-87RW00084,
Nov. 1990.**
- [20] **T. A. Nartker, J. Kanai, and S. V. Rice, "A Preliminary Report on
OCR Problems in LSS Document Conversion," (to appear in)
Proceedings, Nuclear Waste Management Conference,
Las Vegas, NV, April 1992.**