DATA SETS FOR OCR AND DOCUMENT IMAGE UNDERSTANDING RESEARCH

ISABELLE GUYON
AT&T Bell Laboratories
955 Creston Road
Berkeley, CA 94708
isabelle@research.att.com

ROBERT M. HARALICK
Department of Electrical Engineering FT-10
University of Washington
Seattle, WA 98195

JONATHAN J. HULL
RICOH California Research Center
2882 Sand Hill Road, Suite 115
Menlo Park, CA 94025
hull@crc.ricoh.com

IHSIN TSAIYUN PHILLIPS
Department of Computer Science/Software Engineering
Seattle University
Seattle, WA 98125
yun@seattleu.edu

Several significant sets of labeled samples of image data are surveyed that can be used in the development of algorithms for offline and online handwriting recognition as well as for machine printed text recognition. The method used to gather each data set, the numbers of samples they contain, and the associated truth data are discussed. In the domain of offline handwriting, the CEDAR, NIST, and CENPARMI data sets are presented. These contain primarily isolated digits and alphabetic characters. The UNIPEN data set of online handwriting was collected from a number of independent sources and it contains individual characters as well as handwritten phrases. The University of Washington document image databases are also discussed. They contain a large number of English and Japanese document images that were selected from a range of publications.

Keywords: Data sets, databases, text images, online handwriting, offline handwriting, machine printed text, CEDAR, NIST, CENPARMI, UNIPEN, University of Washington.

1. Introduction

The availability of a data set that contains an appropriate number and selection of samples is a critical part of any experimental research project [8]. This is especially true in an image-based application such as optical character recognition (OCR)

where it is sometimes difficult for individual researchers to gather the number and type of data they need because of the costs involved. Ideally, a data set would allow a researcher to project the performance achieved experimentally to the application domain represented by the data set. This implies that the data set used for algorithm development should reflect the application domain as closely as possible.

A significant advance in the experimental integrity of OCR research has been made possible in recent years with the availability of several image data sets. These data sets allow researchers to train and test algorithms on significant numbers of data items and to compare performance on specific images. This has improved productivity since researchers can conduct experiments without first gathering data.

There are three areas of OCR research (offline handwriting, online handwriting, and machine printed text) that require specialized data sets. Offline handwriting is produced by an individual, typically by writing with a pen or pencil on paper, and is scanned into a digital format. Online handwriting is written directly on a digitizing tablet with a stylus. The output is a sequence of x-y coordinates that express pen position as well as other information such as pressure. Machine printed text occurs commonly in daily use and is produced by offset processes, laser, inkjet, or dot matrix printing. Each area has unique characteristics that affect the design of a data set. These characteristics determine how well experimental results on that data set can be generalized to another application.

Offline handwritten text data sets typically contain isolated alphanumeric characters or words. These data sets are often produced by choosing subjects to write on sample forms that are subsequently digitized. An important consideration in the design of data sets in this area is how well the subjects that produce the data and the conditions under which the data are gathered represent the eventual application. Ideally, the subjects should be chosen from the same population and the data gathered under the same conditions (e.g., data gathered in the field, same data form used, etc.) that will be used in the final application. Otherwise, significant differences in performance can occur between the data set and the data used in practice.

Databases of online handwriting also contain isolated characters and words as well as text phrases. Research in online handwriting recognition has been conducted for many years at a large number of research labs. Each project typically developed their own databases for internal use. The *UNIPEN* project has taken advantage of this existing body of data by asking individual research groups to contribute their databases to a common set. This data set will first be used for a comparative benchmark. It will then be shared with the rest of the research community. A significant issue in UNIPEN was the large number of different formats used for online handwriting. This was solved by the development of a common data format which is described in this chapter. The background, history, and future of the UNIPEN effort are also discussed.

Machine-printed text data sets often include images of complete pages. These pages can be chosen from a large number of different classes of machine printed document (e.g., scientific papers, memorandums, newspapers, etc.) and can vary widely in format (fonts, point size, column layout, etc.). A significant consideration is the document classes and variations in format represented in the data set. Also, the ground truth information supplied with such a data set should be comprehensive so that it can be used to improve the performance of OCR algorithms. This is significant since modern OCR systems are affected by many aspects of the data format which are often inter-related. This chapter describes the CDROM data sets developed by the University of Washington which include several document classes and extensive truth data.

The rest of this paper describes several data sets of offline and online handwriting as well as machine printed text. We discuss the contents of these data sets and describe how they address various research considerations.

2. Offline Handwritten Text

There are at least two classes of databases of offline handwritten text that are of interest to most researchers. Datasets that contain isolated handprinted characters and digits have been used in the development of many pattern recognition algorithms. The recognition problem represented by these data is well defined (e.g., 10, 26, 52, or 62 classes) and the recognition task is relatively easy for ordinary people. A reliable solution to this problem could have significant economic consequences since many handwritten forms could be read automatically.

Data sets can also contain isolated handwritten words and phrases. The recognition problem represented by these data is more challenging since the number of classes is larger (e.g., there are more than 40,000 legal ZIP Codes in the U.S.) and there may be few constraints on how users prepare the samples. However, a robust solution to this problem would be even more significant than for isolated character recognition since it would provide a general purpose computer input medium.

Important considerations when evaluating the usefulness of a data set for experimental purposes include the number of samples it contains, the sampling rate at which it was captured, and the pixel depth at which the data is furnished. Commonly used sampling rates include 200 or 300 pixels per inch (ppi) and pixel depths are usually one or eight bits. Such image characteristics are also important system design parameters since they influence the ultimate cost of an implementation.

2.1. Isolated Digits and Alphabetic Characters

Isolated digit and character recognition is one of the most extensively studied application domains in pattern recognition. Some of the earliest databases of digital images, beginning at least in 1966, were composed of single characters (see [7] for

a review). In recent years, several significant collections of such data have been issued.

The CEDAR data set contains nearly 28,000 examples of isolated handwritten characters and digits that were extracted from images of postal addresses [6]. Envelopes with handwritten addresses on them were sampled from the mail as it was being processed in a post office. This assured that the people that prepared the original data were not aware that their addresses would be included in the data set. The address images were scanned at 300 ppi in 8-bit gray scale and later reduced to one-bit images. Individual characters and digits were segmented from the address images by a semi-automatic process that extracted single connected components. These were displayed to a human operator who entered their truth values.

An additional group of about 21,000 digits were also extracted from ZIP Codes in the CEDAR data set by a fully automatic segmentation algorithm. These digits were manually screened for quality and placed into a *good* set if they contained no segmentation artifacts that could be considered errors such as one digit split into two. All the digits output by the segmentation algorithm are also provided in the data set. This provides some challenging examples for testing the performance of an algorithm since artifacts caused by segmentation are preserved in the individual images.

The CENPARMI data set contains 17,000 isolated digits that were extracted from images of about 3400 postal ZIP Codes [11]. The ZIP Codes were selected from dead letter mail in a working post office and thus the preparers of the data also were separated from the data collection task. The ZIP Codes were scanned at 200 ppi with a one-bit digitizer. The isolated digits were extracted from the ZIP Codes by a manual segmentation process that displayed individual images to operators who entered their identities.

The NIST data set SD3 contains one of the largest selections of isolated digits and characters that are publically available. Altogether it contains over 300,000 character images that were extracted from data collection forms filled out by 2100 individuals. Employees of the U.S. Bureau of the Census were instructed to fill the boxes on a form with a particular sequence of digits or characters and were asked to make sure that the individual digits or characters did not touch one another or the surrounding box. The forms were subsequently scanned at 300 ppi in binary mode and automatically segmented. The string of digits that should have been written in each box was used to assign a truth value to each individual image. The truth values were subsequently verified by a human operator.

An interesting issue in data set design is presented by the NIST images. The initial collection process described above was done in preparation for a competition that evaluated the performance of more than 20 different algorithms in isolated digit and character recognition. The images on SD3 were provided as a training set and were used by most of the competing organizations in developing their methods. A separate data set (called TD1) was collected to provide test data. The same

collection process was used. However, the subjects were high school students. The quality of the data varied more widely in TD1 than it did in SD3 and was on the whole more sloppy. This caused a significant drop in performance on the test data because most systems had been trained on the neater images. This experience shows that a given technique may perform very well on a specific data set. However, this does not necessarily mean that the recognition problem represented by that data has been solved. It can mean that the recognition problem has been solved for that data set only. The generalization of any results to a larger population should only be made after careful consideration and comparison of the training and test data to the real-life application.

2.2. Word and Phrase Recognition

Word and phrase recognition is a less frequently studied application of pattern recognition than digit or character recognition. However, words or phrases offer contextual constraints such as dictionaries that make it possible to model interactions between image processing operations such as segmentation and the recognition of isolated symbols given that only certain sequences of symbols occur in the dictionary.

ZIP Codes, city names, and state names are examples of handwritten word images that are available in the CEDAR data set. Approximately 5,000 ZIP Codes, 5,000 city names, and 9,000 state name images are included. These data were scanned by the process described above. However, the 300 ppi 8-bit gray scale versions of the whole word images are provided and the specific address in which each image occurred is identified. Thus, experimentation can be performed with isolated word recognition on gray scale data and a comprehensive system could be developed that used partial results from recognition of the city or state name to influence the recognition of specific digits in the corresponding ZIP Code.

A significant amount of running English text is also available in the NIST SD3 and TD1 data sets. Each collection form contained a *Constitution box* in which subjects wrote the 52-word preamble to the Constitution of the United States. This data was scanned at 300 ppi in 1-bit format as part of the normal data capture process. These images also make it possible to develop algorithms that integrate early processing with contextual analysis. The domain is constrained enough that a range of contextual constraints can be investigated and at the same time the physical format is sufficiently unconstrained (the subjects could have written the preamble anywhere in a given box) that it reasonably represents the way people would like to use handwriting to communicate with a computer.

There are three other NIST data sets (SD11-SD13) that contain examples of phrases that were written by respondents to the U.S. Census to describe their jobs. Three boxes were filled in with the title of a person's job, the work they do, and the work done by the company they are employed by. SD11-SD13 differ from SD3 and

TD1 in that the people that prepared the data were from the general population and had no idea that their writing would be scanned. Altogether, 91,500 handwritten phrases were scanned at 200 ppi in binary mode. 64,500 of the phrases were scanned from microfilm and 27,000 were scanned from the original paper versions. These data are also provided with a dictionary of *legal* entries for each box that was derived from the previous census. Thus, there is no guarantee that the letter-for-letter transcription of each image appears in the dictionary. The recognition task is to identify the dictionary entry that is the *closest* match to the phrase written in the box.

3. Online Handwritten Text

In this section, we present the design of a database for On-Line Handwriting Recognition (OLHR). The database is composed of isolated characters, words, and sentences, written in a variety of styles (handprinted, cursive or mixed). The alphabet is restricted to the ASCII keyboard set. The data were donated by 40 different institutions and therefore includes a variety of writers and recording conditions. The database size approaches 5 million characters. We provide some details about the data exchange platform which could inspire other similar efforts. The database will be distributed to the public for a small fee by the Linguistic Data Consortium (LDC) in 1996.

3.1. History of UNIPEN

On-line handwriting recognition (OLHR) addresses the problem of recognizing handwriting from data collected with a sensitive pad which provides discretized pen trajectory information. OLHR has long been the poor parent of pattern recognition in terms of publicly available large corpora of data. To remedy this problem, the UNIPEN project was started in September 1992 at the initiative of the Technical Committee 11 of the International Association for Pattern Recognition (IAPR). Two IAPR delegates (Isabelle Guyon and Lambert Schomaker) were appointed to explore the possibility of creating large OLHR databases.

In the field of OLHR there exist many privately owned databases. These databases constitute a potential resource which is much richer than any single database, because of the diversity of text entered, recording conditions and writers. Therefore data exchange is a natural way to constitute a sizable database which is representative of the tasks of interest to research and development groups.

A small working group of experts from Apple, AT&T, HP, GO, IBM and NICI laid the foundations of UNIPEN in May, 1993 and proposed that a common data format would be designed to facilitate data exchange. In the summer of 1993, the UNIPEN format was designed, incorporating features of the internal formats of several institutions, including IBM, Apple (Tap), Microsoft, Slate (Jot), HP,

AT&T, NICI, GO and CIC. The format was then tested independently by members of the working group, soon followed by many other volunteers. A second iteration of the test was organized in autumn 1993 to check the changes and additions to the format [4]. In parallel, a set of tools to parse the format and browse the data were developed at NICI (with the sponsorship of HP) and at AT&T.

In January 1994, NIST and LDC committed to help UNIPEN. NIST has been supervising the data gathering and the organization of a benchmark test. LDC will publish a CD ROM and distribute the data. There is already an FTP site at LDC where data and programs can be exchanged.

In June 1994, the instructions for participation in the first UNIPEN benchmark, limited to the Latin alphabet, were released. Potential participants were requested to donate isolated characters, words or sentences containing at least 12,000 characters. Forty institutions responded to the call for data. Further negotiations between owners of large databases succeeded in gathering larger data sets. There are nearly five million individual character samples in the database.

In February 1995, the data donators met for a one day workshop to determine how the data will be split into training and test sets. A benchmark test using these data will take place in 1995. During the test the data will remain the property of the data donators. After the test, it will become publicly available and will be distributed by LDC for a nominal fee.

The activities of UNIPEN will expand in the future according to the needs and desires of the participants.

3.2. Collecting donated samples

3.2.1. A common data format

The UNIPEN format is an ASCII format designed specifically for data collected with any touch sensitive, resistive or electro-magnetic device providing discretized pen trajectory information. Users can easily convert their own format to and from the UNIPEN format or collect data directly in that format.

The minimum number of signal channels is two: X and Y, but more signals are allowed (e.g., pen angle or pressure information). In contrast with binary formats, such as Jot [2], the UNIPEN format is not optimized for data storage or real time data transmission and it is not designed to handle ink manipulation applications involving colors, image rotations, rescaling, etc. However, in the UNIPEN format, there are provisions for data annotation about recording conditions, writers, segmentation, data layout, data quality, labeling and recognition results.

Efforts were made to make the format human intelligible without documentation (keywords are explicit English words), easily machine readable (an awk parser was developed in conjunction with the development of the format itself), compact (few keywords), complete (enough keywords), and expandable.

The format is a succession of instructions consisting of a keyword followed by arguments. Keywords are reserved words starting with a dot in the first column of a line. Arguments are strings or numbers, separated by spaces, tabs or new lines. The arguments relative to a given keyword start after that keyword and end with the appearance of the next keyword or the end of file (see Fig. 1).

Almost everything is optional, so that simple data sets can be described in a simple way. All variables are global: declared variables retain their values until the next similar declaration. Databases written in the UNIPEN format may be concatenated in a single file or they may be organized in different files and directories.

The format can be thought of as a sequence of pen coordinates, annotated with various information, including segmentation and labeling. The pen trajectory is encoded as a sequence of components. A pen-down component is a trace recorded when the pen is in contact with the surface of the digitizer. A pen-up component is a trace recorded when the pen is near the digitizer without touching it. Components are not necessarily delimited by pen-lifts and may or may not coincide with strokes. $.PEN_DOWN$ and $.PEN_UP$ contain pen coordinates (e.g. XY or XYT as declared in $.COORD$). The instruction $.DT$ specifies the elapsed time between two components. The database is divided into one or several data sets starting with $.START_SET$. Within a set, components are implicitly numbered, starting from zero.

Segmentation and labeling are provided by the $.SEGMENT$ instruction. Component numbers are used by $.SEGMENT$ to delineate sentences, words, and characters. A segmentation hierarchy (e.g. $SENTENCE\ WORD\ CHARACTER$) is declared with $.HIERARCHY$. Because components are referred to by a unique combination of set name and order number in that set, it is possible to separate the $.SEGMENT$ from the data itself.

The format also provides a unified way of encoding recognizer outputs to be used for benchmark purposes. To obtain more information about the format, it is possible to access its full definition electronically (see next section).

3.2.2. Internet connections

Without the internet, the Unipen project would not have been possible. Electronic mail has been the primary means of communication between organizers and participants. The data and the tools were exchanged by FTP.

In March 1994, UNIPEN advertised its existence on several electronic mailing lists, resulting in nearly 200 subscriptions to the UNIPEN newsletter. People interested in UNIPEN can send a request to be added to the Scrib-L mailing list. Scrib-L is a mailing list for researchers and developers in the field of handwriting. Electronic mail to:

`SCRIB-L@NIC.SURFNET.NL`

```
.VERSION                1.0
.DATA_SOURCE            ATT
.DATA_ID                Example

.COMMENT                Documentation
                         -------------
.DATA_CONTACT           Isabelle Guyon (isabelle@research.att.com).
.DATA_INFO              Latin alphabet, isolated characters. Data cleaned manually.
.SETUP                  Volunteer staff members. People sitting at a desk.
.PAD                    WACOM HD-648A LCD Digitizer.

.COMMENT                Declarations
                         ------------
.X_DIM                  4160
.Y_DIM                  3200
.H_LINE                 450 1900 2300
.X_POINTS_PER_INCH      508
.Y_POINTS_PER_INCH      508
.POINTS_PER_SECOND      200
.COORD                  X Y
.HIERARCHY              PAGE TEXT WORD

.COMMENT                Data
                         ----
  Most of the point have been removed to shorten the example.
.INCLUDE                lexicon.lex
.DATE                   9 20 93
.WRITER_ID              08171408_14804
.STYLE                  MIXED
.START_BOX
.SEGMENT PAGE           0-58
.SEGMENT TEXT_CHUNK     0-29 ? "that nothing more happened ,"
.SEGMENT WORD           0-6 ? "that"
.PEN_DOWN
707 2417
707 2424
590 2319
.PEN_UP
.DT 151
.PEN_DOWN
588 2377
586 2377
695 2393
.PEN_UP
   etc.
.COMMENT For more examples, please ftp data samples.
```

Fig. 1. Example of UNIPEN formatted data. This example is simplified. Real data are more richly annotated.

will be forwarded to all subscribers. Please refrain from sending messages which are not in the general interest of researchers in handwriting.

Scrib-L resides on the computers of the national node of the Nijmegen University Computing Centre in The Netherlands. Scrib-L subscribers represent as many as 23 countries of the world. Messages are in ASCII, max 77 columns wide, concise, and formal.

```
------------------------- Summary -----------------------------
ASCII MESSAGES (<77 chars/line) to: Scrib-L@NIC.SURFNET.NL
COMMANDS to the boss of Scrib_L:   LISTSERV@NIC.SURFNET.NL
     Subscribe:                SUBSCRIBE SCRIB-L Name (Fax:...)
     Get list of subscribers:  REVIEW SCRIB-L (COUNTRIES
     Get Archive of June'93:   SEND SCRIB-L log9306
-------------------------------------------------------------
```

An FTP site has been set up at the Linguistic Data Consortium for data and software exchange. Currently, most of the directories can be read only by data donators. When the database becomes public, more directories will be open.

To access the directories that are publicly available, proceed as follows:

```
ftp ftp.cis.upenn.edu
Name: anonymous
Password: [use your email address]
ftp> cd pub/UNIPEN-pub/documents
ftp> get call-for-data.ps (benchmark instructions)
ftp> cd ../definition
ftp> get unipen.def       (format definition)
ftp> quit
```

People having access to WWW through Mosaic will find images of UNIPEN example files as produced by the Upview program developed at NICI at:

`http://www.nici.kun.nl/unipen`

3.3. Organizing the database

3.3.1. Computing statistics

For each data set donated, a data sheet with relevant statistics was computed (see Fig. 2). These statistics serve as a basis to determine how to split the data into different subsets.

3.3.2. Defining tasks

Because the database is composed of many data sets of limited size, it is important to group the data sets that address similar tasks. This will also be important for defining a set of standard benchmarks.

```
======================================== SEGMENTATION ========================================
Type            Totals                                  Intersections
----            ------                                  -------------
                                          both            WORD            CHAR         neither
TEXT:        69 writers                      0              69               0               0
          26279 segments                     0           26279               0               0
                                          both            TEXT            CHAR         neither
WORD:        94 writers                      0              69               0              25
          69527 segments                     0           61127               0            8400
                                          both            TEXT            WORD         neither
CHAR:         0 writers                      0               0               0               0
              0 segments                     0               0               0               0
TOTAL:       94 writers
          95806 segments
         302843 characters
         588183 components
======================================== ALPHABET ========================================
                    s     l    u     d   sl   su   lu   sd    ld   ud  slu  sld  sud  lud  slud
TEXT: segments    417 11011 3048   594 1039  70  7991 1183    0    0  836    0   18    0   72
       writers     21    69   69    22   69  22    69   69    0    0   69    0   18    0   37
WORD: segments   4827 30028 7942 11186  382  22 14422  675    0    0   43    0    0    0    0
       writers     69    94   94    94   66  22    94   25    0    0   22    0    0    0    0
CHAR: segments      0     0    0     0    0   0     0    0    0    0    0    0    0    0    0
       writers      0     0    0     0    0   0     0    0    0    0    0    0    0    0    0
======================================== STYLE ========================================
                        PRINTED         CURSIVE        MIXED           _unspec           _bad
TEXT: segments           9174               0           17105              0              0
WORD: segments          26956               0           42571              0              0
WD_LU: segments         16853               0           35539              0              0
CHAR: segments              0               0               0              0              0
TOTAL: segments         36130               0           59676              0              0
======================================== LEXICON ========================================
From:                   labels          lexicon
TEXT:                    1196            1123
WORD:                    1744               0
WD_LU:                      0               0
======================================== WRITER ========================================
Total number of characters:                     302843
Total number of writers:                        94
Average number of characters/writer:            3221.73
Std. dev. of characters/writer:                 52.63
Minimum number of characters/writer:            0
Maximum number of characters/writer:            9796
========================================================================================
```

Fig. 2. Example of data sheet. The statistics of each data set donated to UNIPEN are computed and summarized in a data sheet. This example shows one of the HP data sheets. In the ALPHABET section, the symbols 's', 'l', 'u', and 'd' stand for symbols, lowercase letters, uppercase letters and digits. In the STYLE and LEXICON sections 'WORD_LU' means words contain only lowercase or uppercase letters (no symbols or digits).

Examples of tasks that were defined include:

1. Isolated characters, case separated
2. Isolated characters, mixed case
3. Isolated characters in the context of words (with a dictionary)
4. Isolated printed words
5. Isolated cursive words
6. Text (sentences)

The number of characters available for each task was computed. Tasks may overlap: some characters may be used in more than one task. There is a total of $4,422,863$ characters. For tests 1 and 2, there are $90,305$ digits, $155,494$ uppercase letters, $367,214$ lowercase letters and $85,952$ symbols. For tests 3, 4, and 6, there are respectively $320,916$, $302,251$, $341,017$ and $98,785$ characters available.

3.3.3. Defining training and test set

The next problem is to determine what size test set will give statistically significant results for each task. The remainder of the data (if any) will serve as training data. Since training data is valuable, it is important not to be wasteful when defining the test sets.

Obviously, defining a statistically significant test set size is a chicken and egg problem: before obtaining recognizer performance, it is not possible to determine statistical significance. Nevertheless, since approximate values of the error rates of particular recognizers on given tasks are known, it is possible to estimate the size of a test set using straightforward statistical arguments.

For identically and independently distributed (i.i.d.) data, if E is the error rate of a recognizer, it is possible to compute the number n of test examples such that with probability $(1-\alpha)$ the actual error rate (on an "infinite" size test set) will not be higher than βE [3]:

$$n = \left(\frac{z_\alpha}{\beta}\right)^2 \frac{(1-E)}{E} \tag{1}$$

where z_α is a tabulated coefficient which is a function of α. For typical values of the parameters, $\alpha = 0.05$ ($z_\alpha = 1.65$), $\beta = 0.05$ and $E = 1\%$ character error, the number of test examples obtained is less than $10,000$.

In reality, data are far from being i.i.d. In particular, the data usually come from a limited number of writers which necessarily introduces correlations. The data donators, who are experts in on-line handwriting recognition, advocated a minimum of 100 different writers in every test set, each writer providing at least

1000 characters. Different writers must figure in the training data and the test data for writer independent tests.

The test set size recommended by the donators is 10 times larger than what the theory predicts. At the time of writing this paper, there were still open discussions regarding this matter.

4. Machine Printed Text

The *UW-I* and *UW-II* document image databases [10] contain the following types of document image pages: *English* technical journal articles, 1147 pages *UW-I*, 623 pages *UW-II*; *Japanese* technical journal articles, 477 pages *UW-II*; *English* memorandums, 62 pages *UW-II*. Each document image page is zoned and the text zones have line by line associated character ground truth generated by a double data entry and triple verification protocol [5]. The *UW-III* document image database, which will be issued in 1996, will have line drawings and word bounding boxes for each English page in the *UW-I* and *UW-II* document image databases.

UW-I contains a large set of isolated degraded characters generated by Baird's degradation model [1]. In addition *UW-I* contains software for determining OCR accuracy by comparing OCR generated text with the ground truth text and software for degrading a document image. *UW-II* contains a document image viewer, called Illuminator provided by RAF, Inc. The databases are distributed on a CDROM media [9].

The pages from both English and Japanese journals and reports come from the University of Washington libraries. Some of the English report pages come from the University of Nevada Information Sciences Research Institute database. The memorandums come from staff members of Seattle University. The pages contain a diverse sample of document styles found in technical journals and other documents.

The document image pages are scanned at 300 ppi on either a Fujitsu 3096 document scanner or a Ricoh IS50 scanner. They consist of: binary images scanned directly from the original document pages; binary images scanned from first generation photocopies of the original document pages; binary images scanned from the second or later generation photocopies of the original document pages; gray scale images scanned from the original document pages; and synthetic noise-free binary images from LaTeX generated documents.

In addition to these images, the document image database is annotated with information about the contents of the pages. Qualitative information about the condition of each page in terms of the nature of noise present, including the page rotation angle, are presented in its page condition file. Information about the document page such as the journal from which it is taken, its page number, specification of the dominant font on the page, specification as to whether figures etc. are present on the page are in its page attribute file.

Most OCR algorithms proceed first with a segmentation of the page into "zones" which are usually rectangular areas that are semantically homogeneous. The *UW* document image data bases provide this kind of annotation. At the coarsest level, a page is decomposed into "header," "footer" and "live-matter" areas. Standard definitions of what constitutes a header, etc., from the publishing and page description world are used. The *header* is defined to be ancillary text that appears above the main body of the page. For the world of technical journals this usually includes information such as the name of the article, the journal, the authors and the page number. A similar field may be present below the main body of the page and is referred to as the *footer*. The main body of the page is referred to as the *live matter*. Information about the pixel location and size of each of these zones on the page are provided in its associated "page bounding box" annotation file.

At the next finer level, the page is decomposed into "zones". Zones can be of various types: text, figures, tables, half-tones and mathematical equations among others. Zone delineation information for each page is provided in its "zone bounding box" annotation file.

Each zone has attributes which include things such as the semantic meaning of each zone (for example, a text zone could be a section heading, reference list item or page number), the dominant font in the zone, the font-style, etc. This information is provided in the page's "zone attribute" annotation file.

Finally, there is the "ground-truth" data files. For each "non-text" zone, the zone-type (figure, displayed math, table, line drawing etc.) is given. For each text zone, the text within the zone is specified in terms of its ASCII text, line for line.

4.1. Page Attributes

For each document page in the database, there is a set of attributes that describe the top level attributes of the page. The page attributes contain the page ID, the page contents, the page layout, the font and publication information of a document page. The group of attributes associated with publication information includes: name, volume number, issue number and publication date of the journal. It also has the corresponding page number of the document page from the publication. The page attributes also include the type of language, script, font type, and the character orientation and reading direction of the document page. The font types are defined to be of two varieties: those with and without serifs. Thus fonts such as Times are part of the Serif font type and fonts such as Helvetica are part of the Sans-Serif font type. Where more than one font type is present, the dominant font type is defined to be the one which occupies the largest fraction of the page in terms of physical area.

The various page attributes and their possible values are as follows: Document ID, *8 character string*; Document language, *English, Japanese*; Document script, *Roman, Katakana, Kanji, Hiragana*; Document type, *journal, letter, memo*; Publi-

cation Information; Multiple pages from the same article, *yes, no*; Text zone present, *yes, no*; Special symbols present in text zone, *yes, no*; Displayed Math zone present, *yes, no*; Table zone present, *yes, no*; Half-tone zone present, *yes, no*; Drawing zone present, *yes, no*; Page header present, *yes, no*; Page footer present, *yes, no*; Maximum number of text columns, *1, 2, 3, 4, non-text*; Page Column layout, *regular, irregular, non-text*; Character orientation, *up-right, rotated-right, rotated-left, non-text*; Text reading direction, *left-right, right-left, top-down, bottom-up, non-text*; Dominant font type, *serif, sans-serif, non-text*; Dominant character spacing, *proportional, fixed, non-text*; Dominant font size (pts), *4-8, 9-12, 13-18, 19-24, 25-36, 37-99, non-text*; Dominant font style, *plain, bold, italic, underline, script, non-text.*

4.2. Page Condition

The page condition of a document page describes the visual condition (or qualities) of a given document page. For example, it contains information on the presence or absence of visible salt and pepper noise, or visible vertical and horizontal streaks, or extraneous symbols from other pages. It also indicates if the document page is smeared (or blurred) because of poor focusing. It contains the measured page rotation angle and its standard deviation. It contains information about how many times the document page was successively copied. Thus, if the page is scanned from a first generation copy, the "Nth copy" attribute will have the value of 1.

Quite often when a bound journal is scanned or photocopied, the portion of the page that is close to the spine of the journal is subject to perspective distortion. In regions of the document page that are close to the spine, the lines of text appear to curve (skew) towards either the top or bottom of the page. The page skewed left or right attribute value pairs indicate whether such distortion is present on the document page.

When a bound journal page is scanned or photocopied, there are sometimes sections of the page close to the spine that contain dark blotches which smear the text together. This is because the page appears darker (in grayscale) close to the spine and a uniform binarization threshold would then result in dark blotches. The page smeared left or right attribute values indicate whether such distortions are present on the document page. The page rotation angle is defined as the orientation of the lines of text relative to the horizontal. These orientations and their standard deviations are estimated using a triangulation scheme on multiple sets of manually entered groups of three points on each document page.

The page condition file has the following fields: Document ID, *8 character string*; Degradation type, *original, photocopy, fax*; Nth copy, *noise-free, original, 1, 2, ...*; Visible salt/pepper noise, *yes, no*; Visible vertical streaks, *yes, no*; Visible horizontal streaks, *yes, no*; Extraneous symbols on the top, *yes, no*; Extraneous symbols on the bottom, *yes, no*; Extraneous symbols on the left, *yes, no*; Extraneous

symbols on the right, *yes no*; Page skewed on the left, *yes, no*; Page skewed on the right, *yes, no*; Page smeared on the left, *yes, no*; Page smeared on the right, *yes, no*; Page rotation angle (in degrees); Page rotation angle standard deviation.

4.3. Zones on a Page

A document page can be geometrically partitioned into several rectangular regions, called *zones*. In general any section of text that is clearly demarcated from adjacent areas of a page by "white space" is a *text zone*.

The rules for defining zones on a page are as follows: A zone is geometrically defined as a rectangular region on a document page. A zone is confined to a single column of text. Font type, style and size are mostly homogeneous across one zone. A zone must not be nested completely within another zone. A drawing, table, or half-tone (without its caption) is a zone. A line that demarcates two sections of text or lines that make up a box that encloses a section of text is a special kind of "drawing" and is called a *ruling zone*. Another special case of a "drawing" is called the *logo zone* and usually consists of the business logo of the company that publishes the journal. Other kinds of drawings that make up distinct zones are geographic maps (*map zones*) and advertisements (*advertisement zone*). The caption of a drawing, table or half-tone is a zone. A *list* is defined as any sequence of text zones, each associated with an alphanumeric "counter" or "index tag". The index tag could be a "•" symbol or any other string for e.g. references are often indexed by a string made up of the initials of the authors and the year of publication. Every item in a list constitutes a zone. Every paragraph of text that remains unbroken (with or without in-line mathematical equations) constitutes a zone. A *drop-cap* is in a zone by itself. Every displayed mathematical equation is a zone. *Section headings* (which are distinguished in many cases from the text body by being either bold-faced or underlined) constitute a zone. The section heading could be part of a line of text. Headings that indicate that the following text is part of the abstract of a paper, or the keywords used to index the paper or the start of a list of references constitute legitimate zones. They are referred to as *abstract*, *keyword* and *reference heading* zones respectively. The *page number* of a page constitutes a separate zone even if there is no white space separating it from nearby text.

Sections of text that represent computer algorithms in "pseudo-code" are also zones (called *pseudo-code zones*). Sections of the text that form part of the title area of a journal article have special semantic meanings and are each assigned to a separate zone. These include the title (*title zones*), the names of the authors (*author zone*) (where more than one appears on a line they are all part of the same zone), the organizational affiliation of the authors (*affiliation zone*), any diplomas or educational qualifications of the authors (*diploma zone*), and memberships in academic societies (*membership zones*). Information pertaining to the date on which an article was submitted or accepted for publication has important semantic information and

constitutes an *article submission information* zone. Some articles contain a "blurb" of text that appears on the same page in order to emphasize the point the authors are trying to make. These text regions are called *highlight zones*. Some articles contain a brief summary of the contents of the page in a separate text region. Such a zone is referred to as a *synopsis zone*. The area that contains the key words used to index a paper has special semantic meaning and constitutes a *keyword zone*.

Some of the document pages contain handwritten annotations. These constitute zones which are called *handwriting zones*. No ground truth is entered for these zones. Sometimes there are extra text symbols from the opposite page that appear on a document image (this happens when photocopying from a bound journal). These symbols are not zoned.

4.3.1. Bounding box information

Bounding boxes are given relative to the page as a whole and relative to each zone on the page. Page bounding box information specifies the size and the location of the three types of special zones, i.e. page header zone, page footer zone and live matter zone. The location of the zone is specified in terms of the row and column pixel coordinates of the top left hand corner of each type of zone box. The size is specified by giving the row and column pixel coordinates of the bottom right hand corner of the zone box. These zone boxes have been delineated by hand using an interactive zone boxing tool. These zone boxes are by no means the smallest bounding rectangle but are guaranteed to contain the page area they are meant to. Where unavoidable (as a result of page rotation or too small a separation between zone boxes), there may also be an overlap area between adjacent zone boxes.

For each of the zones in the page the size and location of the zone bounding boxes is similarly specified.

4.3.2. Zone threading

The zones of each document page are grouped into several logical units. Within each logical unit, the reading order is sequential. Such a logical unit is called a *semantic thread*. Thus associated with each zone in a thread is a pointer to the next zone within the thread. Figures and their captions make up a thread. So do tables and their captions. Zones in the header or the footer areas make up threads individually. All other sets of text zones constitute the main thread of the document page. In general, the last zone within a thread has a "`nil`" pointer. In some cases, zone threading may cross pages. When consecutive document pages are presented within the database, the last zone of a given page that is in the live matter area may thread with the appropriate zone on the following page.

4.3.3. Zone attributes

Zone attributes define the properties of a zone on a document page. The zone attributes contain information on the Page ID, the Zone ID, the zone contents, the zone label, the text alignment, the font, the column number, the language and the script, the character orientation and reading direction of the zone. It also has zone threading information, which was explained in the previous section.

The text alignment attribute is defined relative to the zone bounding box for the zone. If the lines of text are aligned with the left edge of the bounding box, it is referred to as *left aligned*. There are similar definitions for right and center aligned zones. If the text is aligned with both the right and left edge of the zone bounding box it is referred to as *justified*. Text alignment within the zone, can have the values: *left, center, right, justified, justified hanging*, and *left hanging*.

Zone attributes include: Dominant font type, *serif, sans-serif, non-text*; Dominant character spacing, *proportional fixed non-text*; Dominant font size (pts), *4-8, 9-12, 13-18, 19-24, 25-36, 37-99, non-text*; Dominant font style, *plain, bold, italic, underline, script, non-text*; Character orientation, *up-right, rotated-right, rotated-left, non-text*; Text reading direction, *left-right, right-left, top-down, bottom-up, non-text*; Zone's column number, *header-area, footer-area, 1-1, 1-2, 2-2, 1-3, 2-3, 3-3, non-text*; Next Zone ID within the same thread, __, *nil*;

Other zone attributes include: Document ID, *8 character string*; Zone ID *3 character string*; Language, *English, French, German, Japanese*; Script, *Roman, Katakana, Kanji, Hiragana*; Zone content, *text, text with special symbols, text with non japanese symbols, text with special symbols non-japanese, text with rubi, text with non-japanese and rubi, text with special symbols non-japanese and rubi, math, table, halftone, drawing, ruling, bounding box, logo, map, form, advertisement, announcement, handwriting, seal, halftone with drawing, figurative text english, figurative text chinese, figurative text korean, signature, initials, new definition, block, figurative text, figurative text japanese*; Text zone label, *text body, list item, drop cap, caption, section heading, synopsis, highlight, pseudo-code, reference heading, definition, reference list, reference list item, footnote, biography, list, not clear.* Page headers and footers can have these zone labels: *page header, page footer, page number.*

Title pages of documents can have zone labels: *title, author, affiliation, diploma, membership, abstract heading, abstract body, abstract heading and body, correspondence, executive abstract heading, executive abstract body, keyword heading, keyword body, reader service, keyword heading and body, publication infomation, article submission information, not clear.*

Memos and Letter documents can also have the following zone attributes: *date, to, from, subject, cc, memo heading, complement, street address, city address, PO Box, phone number, e-mail address, fax number, telex number, laboratory, department, group, division, institution, subject matter, sender's reference, sender's name,*

recipient's name, secretary's initials, sender's title, sender's initials, opening salutation, closing salutation, home office information, founding date, enclosure.

Japanese documents can have these other label values: *subject title, illustrator, author affiliation membership, title, author affiliation.*

4.4. Ground Truth Data

The database provides the ground truth for all text zones on a document page. For non-text zones, such as the displayed mathematical formulas, line-art, figures, table zones and etc., an indication of such will be given.

In *UW-II* a zone-based ground truth method is used. The zone-based ground truth consists of the symbol strings which are contained in the text zone. This includes the standard ASCII characters as well as special symbol escape sequences for non-ASCII characters. The lines in the ground-truth data are broken at the same position of the string where the physical line is broken on the page. Tabbing or indentations are ignored. Single blank characters (spaces) are used for one or more spaces within a text line. Ground truth for Japanese text is provided in unicode. The methodology used for accurate ground truthing involves double data entry and triple verification [5]. The character substitution error rate is estimated at about 60 characters per million for *UW-I*, and 40 characters per million for *UW-II*.

5. Discussion and Conclusions

Several significant data sets for OCR and document image understanding research were surveyed. The CEDAR, CENPARMI, and NIST data sets were designed to address the needs of researchers in offline handwriting recognition.

The UNIPEN data set contains a large number of samples of online handwriting. The success of the UNIPEN project demonstrates that it is possible to exchange data on a large scale. The keys to the success of UNIPEN were to develop a common data format and to keep all decisions democratic so that they would reflect the desires of the majority of the participants. Exchanging data also raises some difficulties. One of them is that it can take a long time (the project is already two years old). Another one is that transforming many different data sets into a standard format can be time consuming.

The University of Washington data sets of machine printed text images have helped many researchers in document recognition. The large number of truthed images on the UW CDROMs have seen widespread use and have been one of the factors in improving the performance of various commercial OCR packages. This project recently provided a significant sample of scanned and truthed Japanese text.

Some interesting practical problems in the design of a data set are illustrated by these efforts. For example, in the UNIPEN project, the splitting of the data set into training and test sets while retaining a statistically significant test set size was

considered. Another open problem in data set design is in non-English language text. The University of Washington has begun to address this issue, but collections of truthed images from other languages besides English and Japanese (e.g., French, Russian, Chinese, etc.) are needed so that problems in non-English OCR can be investigated by many researchers.

Acknowledgments

Isabelle Guyon would like to thank all the persons who provided help to the UNIPEN project and in particular John Makhoul for discussions on statistical significance and the co-organizers of UNIPEN Lambert Schomaker, Stan Janet, Réjean Plamondon and Mark Liberman for sustaining several years of common effort.

References

[1] H.S. Baird, Document image defect models, in *Structured Document Image Analysis,* eds. H.S. Baird, H. Bunke, and K. Yamamoto, Springer-Verlag, Heidelberg, 1992, 546–556.

[2] D. Gerrety, JOT, A specification for an ink storage and interchange format, Technical Report draft version 0.99, Slate Corporation, San Mateo, California, 1993.

[3] I. Guyon, J. Makhoul, R. Schwartz, and V. Vapnik, What size test set gives good error rate estimates, AT&T Bell Laboratories Technical Memorandum BL0115540-951206-07, submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence, 1995.

[4] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman, and S. Janet, UNIPEN project of on-line data exchange and benchmarks, *Proc. of the 12th IAPR International Conference on Pattern Recognition*, Jerusalem, Israel, Oct. 1994, 29–33.

[5] J. Ha, R.M. Haralick, S. Chen, and I.T. Phillips, Estimating errors in document databases, *Proc. of the Third Annual Symposium on Document Analysis and Information Retrieval*, University of Nevada, Las Vegas, NV, April 1994, 435–459.

[6] J.J. Hull, A database for handwritten text recognition research, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 16, 5 (1994) 550–554.

[7] G. Nagy, At the frontiers of OCR, *Proc. of the IEEE*, 80, 7 (1992) 1093–1100.

[8] I.T. Phillips, S. Chen, J. Ha, and R.M. Haralick, English document database design and implementation methodology, *Proc. of the Second Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, NV, April 1993, 65–104.

[9] I.T. Phillips, S. Chen, and R.M. Haralick, CD-ROM Document Database Standard, *International Conference on Document Analysis and Recognition*,

Tsukuba Japan, Oct. 1993, 478–483; reprinted in *Document Image Analysis* by L. O'Gorman and R. Kasturi, Los Alamitos, CA, IEEE Computer Society Press, 1994, 198–203.

[10] I.T. Phillips, J. Ha, R.M. Haralick, and D. Dori, The implementation methodology for a CD-ROM English document database, *International Conference on Document Analysis and Recognition*, Tsukuba, Japan, October 1993, 484–487.

[11] C.Y. Suen, C. Nadal, R. Legault, T.A. Mai, and L. Lam, Computer recognition of unconstrained handwritten numerals, *Proc. of the IEEE*, 80, 7 (1992) 1162–1180.