CD-ROM Document Database Standard

Ihsin T. Phillips

Department of Computer Science
Seattle University
Seattle, WA 98122

Su Chen and Robert. M. Haralick

Department of Electrical Engineering
University of Washington
Seattle, WA 98195

Abstract

The paper presents the design of a comprehensive standard document database for machine-printed documents. Our effort to produce a series of carefully ground-truthed document databases to be issued on CD-ROMs is described in detail. The databases can be utilized by the OCR and document understanding community as a common platform to develop, test and evaluate their algorithms.

1 Introduction

Systems that do OCR or any aspect of Document Image Understanding must work nearly perfectly over a broad range of document conditions and types in order to be really useful. To develop algorithms for OCR or to develop algorithms for Document Image Understanding requires that the developer have a suitable database of documents which are accurately ground-truthed so that the free parameters of the algorithms can be estimated. Customers of OCR or Document Image Understanding systems likewise must have a suitable database in order that they may accurately evaluate vendor proposed systems.

Database requirements for both the developer and the customer are nearly identical. Therefore, in order to help both developer and customer, there must be the creation of a comprehensive series of databases, each specialized to a given subset of document types, or intended as additions to already created databases.

Throughout the history of OCR research, and document layout and segmentation researchers, there has been a need for common data sets on which to develop and compare the performance of the algorithms. Some efforts have been made by researchers in the OCR community to make the data sets on which their algorithms have been tested available to the research community. Unfortunately, many researchers remain unwilling to do the same. Thus it becomes impossible for a researcher to verify published results by the process of replicating the algorithm or to compare the performance of a competing algorithm that she/he is developing. Even when data sets are available, the data sets are often tuned to the algorithm that the researcher supplying the data sets has developed. Thus the researcher obtaining the data sets has no alternative but to tune her/his algorithm to the data set and this does not promote good research. It is time for a series of comprehensive standard document databases to be constructed and made them available to researchers. Such databases would serve to provide uniform platforms on which researchers could develop and compare the recognition accuracy of their OCR and document understanding algorithms.

The cost for the creation of a document image database is relatively high due to the care that must exist in the creation process and the requirement for near perfect accuracy. Therefore, it is worthwhile to consider leveraging the cost of producing the ground truth of the document images by also including in the database some software to artificially degrade and distort document images in ways which approximate the real degradations and distortions that documents undergo as they get copied, recopied, faxed, etc. Such degradation software can be used to generate controlled degraded document pages for OCR algorithm development as well as for performance evaluations and testing of the algorithms.

Our efforts to create such a series of carefully ground-truthed databases to be issued on CD-ROMS is described in this paper. The paper presents the design of a comprehensive standard document database for machine-printed documents. An English version of the database is currently under its construction at the University of Washington and it is to be completed in July, 1993.

0-8186-4960-7/93 $3.00

2 Requirement for Machine-printed Document Database

2.1 Languages

In order to be useful for developers of OCR algorithms and document understanding systems. The document databases to be constructed should reflect the full range of machine-printed documents. The document databases should also include documents in each one of the world's major language and scripts, such as Roman, Hebrew, Arabic and Farsi, Kanji, Hangul, Devnagiri and Cyrillic. Our efforts are restricted to the Roman script and English language.

2.2 Document Types

The database series should at least include the following document types:

- Articles: journals, proceedings, books, etc.;
- Business letters and memorandums;
- Newspapers/magazines;
- Maps: street maps, terrain maps, etc.;
- Forms;
- Manuscripts;
- Engineering CAD/CAM drawings;
- Advertisements.

2.3 Document Format and Quality

For each type of document, a variety of documents of various formats and quality needed to be present in the database. These documents should be drawn according to the frequency of their usage to satisfy the requirements of performance evaluation and drawn so that sufficient samples are present of each variety to satisfy requirement of algorithm developers.

2.4 Document Page Images

The database should included both greyscale and binary document images in TIFF formats, The database should also include synthesized noise-free document images as well as various degraded document images (degraded through both the real process and simulation). All image files on the CD-ROM should be compressed to save space.

2.5 Document Page Description

Page Attributes

The page attributes should include at least the language and the script, the font information, the publication information, the page condition, the page layout, and the character orientation and reading direction of the document page.

Zone Definition and Attributes

The zone definition should specify the shape, the size and the location of each zone in a document page. The zone attributes may include zone type (i.e., text zone, figure zone, form, map, table, etc.), zone label (i.e., title, page number, paragraph, author, footnote, etc.), language, script, character orientation, reading direction, font information and text alignment format.

2.6 Ground Truth

The database should provide two types of ground truths. One is the character-based ground truth and the other is the zone-based ground truth.

The character-based ground truth will include the name, the size and the position of each individual character on the page.

The zone-based ground truth will contain the character string with the line break for each text line within a text zone.

2.7 Degradation Models

It is also necessary to develop document degradation models to provide researchers with a mechanism for introducing random perturbations on noise-free ideal images. These degradation models will be developed based on the kinds of degradation found in real-life photocopying and FAX transmission, as well as, coffee stains, ink bleeding, page aging, etc.. The document degradation software will simulate these real-life degradations. This degradation software can be used to generate controlled degraded document pages for OCR algorithm development, as well as for performance evaluations and testing of the algorithms. Such synthetically degraded images give unlimited extension to the real data sets in the database without having to provide additional ground truth for the generated documents.

2.8 Software

The database should include data compression and decompression software, OCR performance evaluation software, degradation software, as well as utility tools for the database.

3 CD-ROM English Document Database: A Case Study

In the rest of the paper, we presents the design of a machine-printed English document database. The set of document pages will be selected from various technical journals and reports.

3.1 Contents and Organization Overview

The English document database has two logical compartments: the software compartment and the document compartment. The software compartment contains the data compression and decompression software, the OCR performance evaluation software, the photocopy degradation software, and the FAX degradation software. The descriptions of this compartment is given in Section 3.2.

The document compartment contains all the document page images, the page and zone attribute record files, the zoning information files and ground truth files. The descriptions of this compartment is given in Section 3.3.

Since the document database will be packaged on a CD-ROM. The file names and directory structures will be in complete compliance with ISO 9660. The general file name conventions for the document database are given in Section 3.4.

3.2 Software Compartment Contents

Compression and Decompression Software

All the scanned and simulated documents in bitmap (TIFF) format on the CD-ROM are supplied in compressed form. The compression algorithm used is the CCITT Group IV bi-level image compression standard. The user can then use the decompression program provided in the CD-ROM to uncompress the compressed data files.

OCR Performance Evaluation Software

OCR performance evaluation software will be developed and provided. Given a list of document zone IDs and OCR outputs of the zones, the algorithm will evaluate the output of OCR algorithm against the corresponding ground truth residing in the CD-ROM. A set of contingency tables for characters and misrecognized words will be computed and output by the algorithm. The user's manual of the software package is given in [4].

Photocopy Degradation Software

Software is also provided that simulates two selected document degradation models. One is Baird's degradation model [2] and the other is currently developed by researchers at the Intelligent Systems Laboratory. Given a document file (binary image file) and degradation model parameters, the user can run the photocopy degradation program to degrade the document as desired. The requirement specification of the software is given in [3].

3.3 Document Compartment Contents

Document Page Image Files

The document compartment includes a set of document image files. Each document image corresponds to one document page. The document images can be classified into the following categories:

1. Scanned greyscale images from real documents.
2. Scanned binary images from real documents.
3. Synthesized noise-free binary image.
4. Degraded noise-free binary image.

The real document pages will also include a set of documents which are taken from the set of synthesized noise-free documents and degraded through real processes – both by successively photocopying or FAX transmission. The degraded document pages will also include a set of synthetic degraded documents (for convenience) that are degraded by the same degradation software that will be provided in the database. The source document of the degraded pages will come from a set of selected pages from category 2 and 3.

Page Attribute Files

For each document page in the database, there is a set descriptive attribute which describe the various attributes of the page. Each document page type (journal, letter/memo, etc.) has its own set of attributes. For technical journals/reports, the attributes include

page condition, page bounding boxes, page contents, page layout, font information, publication information, etc. The journal page attributes definitions are given in Section 4.

The advantage of defining a set of document page attribute records per document type over a set of general page attributes for all document types is that it allows one to add another document type to the database without any change to the database design.

Zone Attribute Files

Each document page will be zoned manually according our zoning conventions [5]. Each zone in a page is associated with a set of zone attributes that describes the contents of the zone. We define a distinct zone attribute record for each document type (journal, etc.). The definitions of the zone attribute are given in Section 4.5.

Ground Truth Files

The database provides the ground truth for all text zones on all document pages. (For mathematical zones, the form of ground truth may be developed and provided. For line-art, halftone zones, etc., there will be no ground truth.) The format of the ground truth is given as character sequences: the correct character sequences (with line breaks between sequences) within the zone.

In addition, each LaTeX generated document page resides in the database, we provide a ground truth file that contains character positions of all characters on the page. The format of the ground truth is given in a character position sequence: the sequence of characters with the position (coordinates) and the size of each character in the zone. (This format only for synthesized and degraded document pages arising from the LaTeX generated documents in the database.

The ground truth files will be used by the OCR evaluation software reside in the database. The software evaluates the performance of OCR algorithms.

All special symbols will be represented in LaTeX-alike syntax. The translation table will be provided.

Bounding Box Information Files

The database also provides bounding boxes informations for the page header, page footer, live matter and each zone on a document page. The live matter of a document page is the usable area of the page between the margins [1]. Each bounding box will be represented as a rectangular region on a document page. The definitions of the bounding box information files are given in Section 4.

3.4 File Name Convention

The general file name conventions for the document database are as follows: 1) A legal file name will consist of at most 8 characters (26 capital English letters and 10 digit numbers) followed by a period and a 3-character extensions. 2) The first character of the file name must be a capital English letter.

Under our current design, the document files have additional file name constraints to make them more identifiable. For example, the first four characters of the file name represent the document page ID. The last four characters of the file name are used to identify the category of the file (scanned binary, scanned greyscale, page/zone attribute record file, page/zone bounding box record file, ground truth, etc.). The filename extension indicates the file format (.TIF for image file, .TXT for ASCII file and .TEX for LaTeX file).

4 Document Page Record Definitions

This section gives the definitions of all records that constitute record files within the document compartment of the CD-ROM.

4.1 Page Condition Record

This record includes attributes that describe the visual conditions (or qualities) of a given document page. The page condition record has the following fields:

Record Field Definitions:

- Document ID:
- Degradation type: (original)(photocopy)(fax)
- n-th copy: (noise-free)(1)(2)( )
- Visible salt/pepper noises: (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)
- Visible page rotation: (yes)(no)
- Page rotation angle (in degree):
- Page rotation angle standard deviation:

4.2 Page Attribute Record

The following record fields define the set of descriptive attributes which describe the various attributes of a journal document page.

Record Field Definitions:

- Document ID:
- Document language: (English)

  The value for this field is English for this database. The field is provided for upward compatibility with future databases that we or others might produce in languages other than English, for ex. Kanji, Arabic etc.
- Document script: (Roman)
- Document type: (journal) (letter) (memo) (news)
- Publication Information: This attributes contains information about the name, the volume number, the issue number and the publishing date of the publication. It also has the corresponding page number of the document page from the publication.
- Multiple pages from the same article: (yes)(no)

  A flag indicating whether multiple document pages from the same article are included in the database. The document pages within the same article can be retrieved by reference to the publication name, volume and issue number of the page.
- Text zone present: (yes)(no)
- Special symbol present in text zone: (yes)(no)

  The special symbols are defined as the symbols other than the standard ASCII symbols.
- 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)
- Max number of text columns: The number of equal-width text columns within of the live matter area of the document page.
- Page Column layout: (regular) (combined-columns)
- Character orientation: (up-right) (rotated-right) (rotated-left).

  This field gives the orientation of characters within the text line when the page is oriented to up-right position.
- Text reading direction: (left-right) (right-left) (top-down) (bottom-up)

  This field gives the text reading direction within a text line when a page is oriented to up-right position. For example, for a landscaped oriented page, the page needed to be rotated to in up-right position.
- Dominant font type: (Serif)(Sans-Serif)
- Dominant character spacing: (proportional) (fixed)
- Dominant font size (pts): (<< 9) (9-12) (13-18) (19-24) (25-36) (>> 36)
- Dominant font style: (plain) (bold) (italic) (underline) (other)

  Any combination of the font styles are allowed. The word 'dominant' is defined as the most frequently used font (type, style, size) in a given page.

4.3 Page Bounding Box Record

The page bounding box record defines the page header area, page footer area and live matter area. A page bounding box record has the following fields:

Record Field Definitions:

- Document ID:

- Header area upper-left corner coords:
- Header area lower-right corner coords:
- Live matter area upper-left corner coords:
- Live matter area lower-right corner coords:
- Footer area upper-left corner coords:
- Footer area lower-right corner coords:

4.4 Zone Bounding Box Record

The zone bounding box record defines each zone on a document page.

Record Field Definitions:

- Document ID:
- Zone ID:
- Zone upper-left corner coords:
- Zone lower-right corner coords:

4.5 Zone Attribute Record

This attribute record describes a set of attributes that are common to zones from a journal/report document page. The record has the following fields:

Record Field Definitions:

- Document ID:
- Zone ID:
- Zone content:

  The zone content can take on either of the following values: text, text with special symbols, displayed math, table, half-tone, drawing, form, ruling, bounding box, logo, map, advertisement, announcement, handwriting and others.

- Text zone label:

  The zone label can be one of the following values: text body, list item, drop cap, caption, abstract body, abstract heading, section heading, synopsis, highlight, pseudo-codes, reference heading, reference list item, footnote, author biography, page header, page footer, page number, article title, author, affiliation, diploma information, society membership information, article submission information, abstract heading, abstract body, footnote heading, keyword heading, keyword body and others.

- Text alignment within the zone:

  This attribute defines the text alignment within the zone. The types of text alignment are: left aligned, center aligned, right aligned, justified, justified hanging, left hanging.

- Font information: This attributes defines the dominant font type, character spacing, font size and font style within the zone.
- Character orientation:
- Text reading direction:
- Zone's column number:

  This attribute describes the zone's column location. A zone may be in the header area, footer area and column number 1 of 1, 1 of 2 and etc.

- Next zone ID within the same thread: (__) (nil)

  The zones of each document page can be grouped into several logical units. Within each logical unit, the reading order is sequential. We call such a logical unit as a semantic thread. This attribute is used to indicate the reading order among the zones that constitute a semantic thread. "nil" is used to indicate the end of the semantic thread.

References

[1] *Xerox Publishing Standards: A Manual of Style and Design*, A Xerox Press Book, Watson-Guptill Publications/New York.

[2] H. S. Baird "Document Image Defect Models", *Structured Document Image Analysis*, Springer Verlag, N. Y., 1992, p546-556.

[3] T. Kanungo, *Document Degradation Model Requirement Specification*, ISL Report, 1993, University of Washington.

[4] Su Chen, *OCR Performance Evaluation Software User's Manual*, ISL Report, 1993, University of Washington.

[5] I. T. Phillips and S. Chen, *English Document Database Zone Label Definitions and Examples*, ISL Report, 1993, University of Washington.

[6] 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*, April 26-28, pp. 65-104, 1993.