Printed Arabic Text Database (PATDB) for Research and Benchmarking

AMIN G. AL-HASHIM and SABRI A. MAHMOUD
Department of Information and Computer Science
King Fahd University of Petroleum and Minerals
KFUPM 1510, Dhahran 31261
SAUDI ARABIA
e-mail: alhasha@kfupm.edu.sa , smasaad@kfupm.edu.sa
Web: http://www.ccse.kfupm.edu.sa/~alhasha/ , *http://faculty.kfupm.edu.sa/ICS/smasaad/*

Abstract: - This paper presents the details of a comprehensive database of Printed Arabic text for Arabic text recognition research. It consists of scanned images of different forms of Arabic printed text (viz. book chapters, advertisements, magazines, newspapers, and reports) scanned with 200, 300, and 600 dpi resolutions. A total of 6954 pages are scanned. The database may be utilized by Arabic printed text recognition research community. It may be used as a benchmark database where researchers can evaluate their algorithms and results compared with published work of other researchers using the same database. To the best of our knowledge, there is no public comprehensive printed Arabic text database that is freely available. Hence, this database may address this deficiency in Arabic printed text recognition research. This database will be made freely available to interested researchers.

Key-Words: - Arabic Printed Text Database, OCR Datasets, Arabic Text Recognition, OCR

1 Introduction

This paper presents the details of a comprehensive database of Printed Arabic text for Arabic text recognition research. It consists of scanned images of different forms of Arabic printed text (viz. book chapters, advertisements, magazines, newspapers, and reports) scanned with 200, 300, and 600 dpi resolutions. A total of 6954 pages are scanned. The database may be utilized by Arabic printed text recognition research community. It may be used as a benchmark database where researchers can evaluate their algorithms and results compared with published work of other researchers using the same database. To the best of our knowledge, there is no public comprehensive printed Arabic text database that is freely available. Hence, this database may address this deficiency in Arabic printed text recognition research. This database will be made freely available to interested researchers.

Having a public large-scale comprehensive database of Arabic text ready for researchers and developers can contribute to the development of the ATR field in different aspects. Researchers need not prepare their own data. In addition, It will enable researchers and developers to collaborate remotely and compare algorithms and results [1, 2]. This paper describes a database that aims to serve Arabic ATR researchers and developers. The database, which includes Arabic printed text, is named as PATDB (Printed Arabic Text DataBase).

The rest of the paper is organized as follows: Section 2 surveys the printed Arabic text databases available in the literature. A description of the PATDB contents and attributes is given in Section 3. The process used to develop the PATDB is detailed in Section 4. Section 5 expresses the details of the PATDB and the meta information of PATDB records. Finally, conclusions are presented in Section 6.

2 Literature Review

To the best of our knowledge, there is no public, large-scale, and comprehensive printed Arabic text database that is freely available. A small number of bounded, commercial and/or special purpose Arabic text databases are only available. The most popular printed Arabic text databases in the literature of which the writers aware are: ERIM database and DARPA corpus.

2.1 ERIM Database

ERIM database was created by Environmental Research Institute of Michigan from a set of machine-printed Arabic books and magazines. It contains over 750 pages that consists of approximately 1,000,000 characters and over 200 distinct ligatures. Pages were scanned with a resolution of 300 dots per inch (dpi). The database is divided into 3 distinct sets, namely, training, statistics and testing set. It is available on a CD ROM for US$ 40 [6].

ERIM database covers only two aspects of real life written communications, namely, books and magazines. However, many other aspects of real life written communications exist such as letters and newspapers. This coverage limitation is considered a disadvantage when developing a general-purpose ATR application. In addition, one may consider the fact that it is not freely available as another disadvantage.

2.2 DARPA Corpus

DARPA (Defense Advanced Research Projects Agency) Arabic corpus was created by Scientific Application International Company (SAIC) for the US Department of Defense [7]. DARPA contains 345 pages of images (around 670,000 characters) with ground-truth. Images were scanned with a resolution of 600 dpi. Images are zones of a single column of text that vary in quality. The corpus was collected from book chapters, magazine articles, newspapers and computer generated documents having only 4-fonts.

DARPA corpus has the same disadvantage of ERIM database. It does not cover all the aspect of real life written communications such as letters and advertisements. In addition, it is not currently freely available.

3 Database Overview

The set of document pages, that are included in the PATDB, are selected from various printing forms (viz. advertisements (2.1%), book chapters (21.3%), magazines (59.2%), newspapers (3.0%) and reports (9.9%)). Section 5 illustrates the details of each category of the selected document pages.

Based on our analysis, the needed attributes of the Arabic database are the followings:

- Simple, i.e., easy to use;
- Extensible, i.e., eligible for adding new document types without modifications;
- Standard, i.e., any addition to the database will follow a set of pre-defined rules;
- Public;
- Comprehensive, i.e., covers different types of documents;
- Uniform, i.e., allows different sub-tasks to access different types of data in the same way;
- Reflects the physical document hierarchy as well as the logical structure;
- Stores the scanned images and their ground-truth information (corresponding text, style information, etc.) separately;
- May include a set of tools for manipulating the database and simulating the natural paper degradation models such as paper aging, multi-coping, skewing and pepper-and-salt noise.

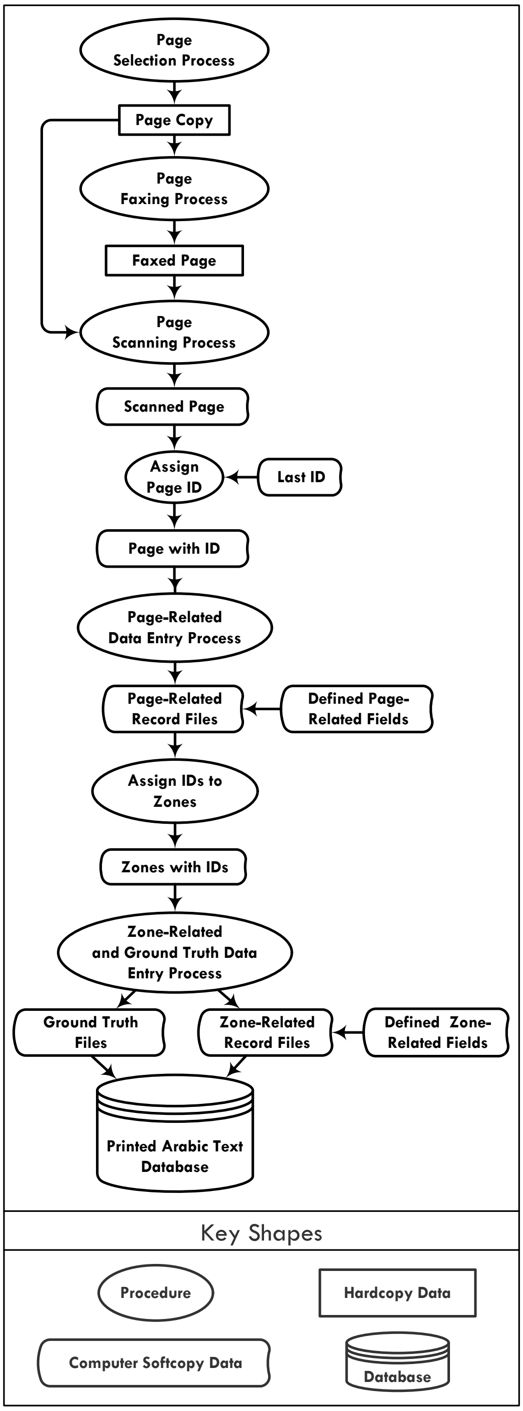


Fig.1: PATDB high-level implementation process

TABLE 1: DISTRIBUTION OF PAGE IMAGES ACROSS PATDB

Category	Color Format / Resolution									Total
	Black & White			Grayscale			Color			
	200 dpi	300 dpi	600 dpi	200 dpi	300 dpi	600 dpi	200 dpi	300 dpi	600 dpi	
ADS	22	22	14	22	22	14	22	22	14	174
BOOK	111	111	111	111	111	111	0	0	0	666
BOOK (fax)	111	187	111	187	111	111	0	0	0	818
MAG	536	536	536	536	536	536	284	284	336	4120
NEWS	35	35	35	35	35	35	0	0	0	210
REP	161	161	161	161	161	161	0	0	0	966
Total	976	1052	968	1052	976	968	306	306	350	6954

The construction process of the PATDB, which is addressed in Section 4, takes into account the above-mentioned attributes.

4 Database Construction Process

The construction process of the PATDB follows an implementation methodology similar to the one carried out by [3]. Figure 1 outlines this implementation methodology. As shown in the figure, the process starts by selecting an appropriate page. The selection process depends on a set of pre-defined criteria. The criteria are specific to each category type. After selecting an appropriate document page, the page may be faxed before scanning to represent faxed images with lower resolution. After that, the page is scanned and given a unique number. This unique number acts as the page identifier and appears in all the files that describe this page. Then, the page is attributed and zoned. The attributes and ground-truth value of each identified zone are keyed. All the resulting files (page image, attribute files and ground-truth value files) are named according to a pre-defined naming convention, stated in Section 5.4, before uploading them to the database.

5 Database Specification

This section addresses the specifications of the PATDB. The different image formats used to store the scanned page images in the PATDB are expressed in Section 5.1. Section 5.2 presents the details of the storage requirements of the PATDB. PATDB meta information is detailed in Section 5.3. Section 5.4 defines the naming conventions of the files of PATDB.

5.1 Page Image Formats

The document pages, that are stored in the PATDB, are scanned and stored in three different formats: (1) black & white (binary) format with color depth of 1-bit per pixel; (2) grayscale format with color depth of 8-bit (1-byte) per pixel (0 to 255 gray levels); and (3) color (or RGB) format with color depth of 24-bit (3-byte) per pixel. The resolutions at which the document pages are scanned are 200, 300 and 600 dots per inch (dpi).

The uncompressed TIFF file format (file-extension .tif) is chosen to store the images of the scanned document pages in the PATDB. The TIFF format is selected because it can store complex information for the CMYK color model and can use JPEG compression techniques.

All the page images in the PATDB are scanned using HP Scanjet N8400 series scanner. The faxing of the page images is done within our campus from a low-end Panasonic fax to another Panasonic fax.

5.2 Storage Requirement

The PATDB contains 6954 page images with their associated metadata files (records). These records describe the scanned pages and the zones of these pages. Three records describe the page images, namely, a page condition record, a page attribute record and a page bounding box record. Similarly, three records describe each identified zone on the page image, namely, a zone bounding box record, a zone attribute record and a zone ground-truth value record. The detailed description of these records is stated in Section 5.3. Table 1 shows the number of images available in the PATDB.

5.3 Meta Information

This section provides the metadata information of the page-related records (page condition record, page attribute record, and page bounding box record) and zone-related records (zone bounding box record, zone attribute record, and zone truth-value record). The metadata is adopted from [4] to reflect the PATDB.

TABLE 2: PAGE CONDITION RECORD

Attribute	Possible Values
Document ID	
n-th copy	0, 1, 2, …
n-th fax	0, 1, 2, …
Resolution	200 (fax), 300, 600, …
Scanning type	black & white, grayscale, color
Color depth (in bits)	1, 2, 4, 8, …
Degradation type	original, page aging, …
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
Visible page rotation	yes/no
Page rotation angle (in degree)	
Page rotation angle standard deviation	

5.3.1 Page Condition Record

The purpose of the set of attributes that constitute the page condition record is to describe the various visual conditions of a given page image. These attributes and their possible values are listed in Table 2. As shown in the table, these attributes are self-explanatory. The '*n-th copy*' attribute represents the number of consecutive copies applied on the page. Similarly, the '*n-th fax*' attribute presents how many times the page is faxed consecutively before scanning.

5.3.2 Page Attribute Record

The various properties of a given page are described through the attributes of the page condition record. The list of attributes is given in Table 3.

The value of the '*document language*' attribute is '*Arabic*' for PATDB. It is used for databases that may be produced in other languages later.

The '*publication information*' attribute holds the information about the source from which the page is taken.

The '*multiple pages from the same article*' attribute indicates whether an article span more than one page. This attribute can be used with the '*publication information*' attribute to retrieve a full article that exists on more than one page.

The '*max number of text columns*' represents the number of text columns in the body area of the given page.

The '*character orientation*' attribute provide the orientation of the characters within the text line when the page oriented to up-right position. Similarly, the '*text reading direction*' attributes

TABLE 3: PAGE ATTRIBUTE RECORD

Attribute	Possible Values
Document ID	
Document language	Arabic, Farsi, Dari, Azeri, Urdu, Uygur, Tajik, Pashto, Kurdish, English
Document script	Arabic, Latin
Document type	newspaper, book, report, magazine, advertisement, letter
Publication information	
Multiple pages from the same article	yes/no
Text zone present	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
Max number of text columns	
Page column layout	regular, combined-columns
Character orientation	up-left, up-right, rotated-right, rotated-left
Text reading direction	left-right, right-left, top-down, bottom-up
Dominant font type	Traditional Arabic, Arial, Simplified Arabic, Arabic Transparent, Times New Roman, Andalus, Courier New, Microsoft Sans Serif, Tahoma, …
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

presents the text reading direction within a text line when the page is oriented to up-right position.

5.3.3 Page Bounding Box Record

The boundary of the header area, body area and footer area of a given document page is defined in the page bounding box record. For each area, the coordinates of the upper-left and lower-right corners are included in the record. Table 4 shows the six attributes of these three areas.

TABLE 4: PAGE BOUNDING BOX RECORD

Attribute	Possible Values
Document ID	
Header upper-left corner coordinates	(X,Y)
Header lower-right corner coordinates	(X,Y)
Live matter area upper-left corner coordinates	(X,Y)
Live matter area lower-right corner coordinates	(X,Y)
Footer upper-left corner coordinates	(X,Y)
Footer lower-right corner coordinates	(X,Y)

5.3.4 Zone Bounding Box Record

The zone bounding box record of a given zone in a given document page contains its identification number and its upper-left and bottom-right bounding box (x, y) coordinates within the document page.

5.3.5 Zone Attribute Record

The zone attribute record describes the common characteristics of an identified zone in a given document page. Table 5 presents the set of attributes that constitute the zone attribute record.

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

The zones of each document page can be grouped into several logical units. Within each logical unit, the reading order is sequential. This logical unit is called a semantic thread. So, the '*next zone ID within the same thread*' 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.

5.3.5 Zone Truth-Value Record

The zone truth-value record is available in the case of text zones only. It contains the ground-truth value of the given text zone.

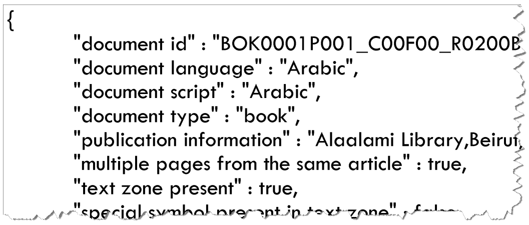
TABLE 5: ZONE ATTRIBUTE RECORD

Attribute	Possible Values
Document ID	
Zone ID	
Zone content	text, text with special symbols, displayed math, table, half-tone, drawing, form, ruling, bounding box, logo, map, advertisement, announcement, handwriting, others
Text zone label	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, other
Text alignment within the zone	left aligned, center aligned, right aligned, justified, justified hanging, left hanging
Dominant font type	Traditional Arabic, Arial, Simplified Arabic, Arabic Transparent, Times New Roman, Andalus, Courier New, Microsoft Sans Serif, Tahoma
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
Character orientation	up-left, up-right, rotated-right, rotated-left
Text reading direction	left-right, right-left, top-down, bottom-up
Zone's column number	
Next zone ID within the same thread	

All the page-related and zone-related records, except the zone truth-value record, are stored as ASCII text files (extension .txt) with a simple representation using JSON (JavaScript Object Notation). JSON bases on name/value pairs idea [5]. In PATDB, the name represents the attribute name while the value is one of the possible values of this attribute.

The zone truth-value records are UNICODE text files (extension .txt) that holds only the ground truth-value of a given text zone.

A part of a text file represented using JSON for a page attribute record of a given book page image is given as an example in Figure 2.



```
{
        "document id" : "BOK0001P001_C00F00_R0200B
        "document language" : "Arabic",
        "document script" : "Arabic",
        "document type" : "book",
        "publication information" : "Alaalami Library,Beirut,
        "multiple pages from the same article" : true,
        "text zone present" : true,
        "special symbol present in text zone" : false
```

Fig.1: A Sample of a Page Attribute Record File

5.4 File Naming and Database Hierarchy

Table 6 and Table 7 show the naming conventions of the different categories of the scanned document pages and the scanning types, respectively.

TABLE 6: CATEGORIES ABBREVIATIONS

Abbreviated Name	Full Name
ADS	Advertisements
BOOK	Book Chapters
MAG	Magazines
NEWS	Newspapers
REP	Reports

TABLE 7: SCAN-TYPE ABBREVIATIONS

Abbreviation	Meaning	Color Depth (bit/pixel)	Number of Colors
BW	Black & White (binary)	1	2
GS	Grayscale	8	256 gray shades
CL	Color	24	Millions of colors

All the items (page images and records, etc.) that make up the PATDB are placed and named according to the following criteria:

- The first level of the database contains a list of directories. Each one of these directories represents a category of document files, e.g., magazines. These directories are named according to the 'Categories Abbreviations' table. Therefore, any newly acquired document file should be added to one of these categories. If the document file cannot fit to any of these directories, then a new directory is created after recording its name and abbreviation in the 'Categories Abbreviations' table.
- The name of the document file within its corresponding directory is formatted as follows: *CategoryNamennnn.FileExtension*, where CategoryName is the abbreviated category name according to the 'Categories Abbreviations' table; nnnn is a 4-digit unique number; and FileExtension is the extension of the file being stored. This document file acts as the source from which the ground-truth value of the images is taken.
- For each added document file, a directory is named after the document file name excluding the file extension. This directory is created in the corresponding category directory. It includes the following items for each page of the document file:

1. A scanned image is named as follows: *DocNamePppp_CccFff_RrrrrScanTypelll.tif*, where DocName is the document name without file extension; ppp is a 3-digit page number; cc is a 2-digit copy number; ff is a 2-digit fax number; rrrr is a 4-digit resolution level in dot per inch

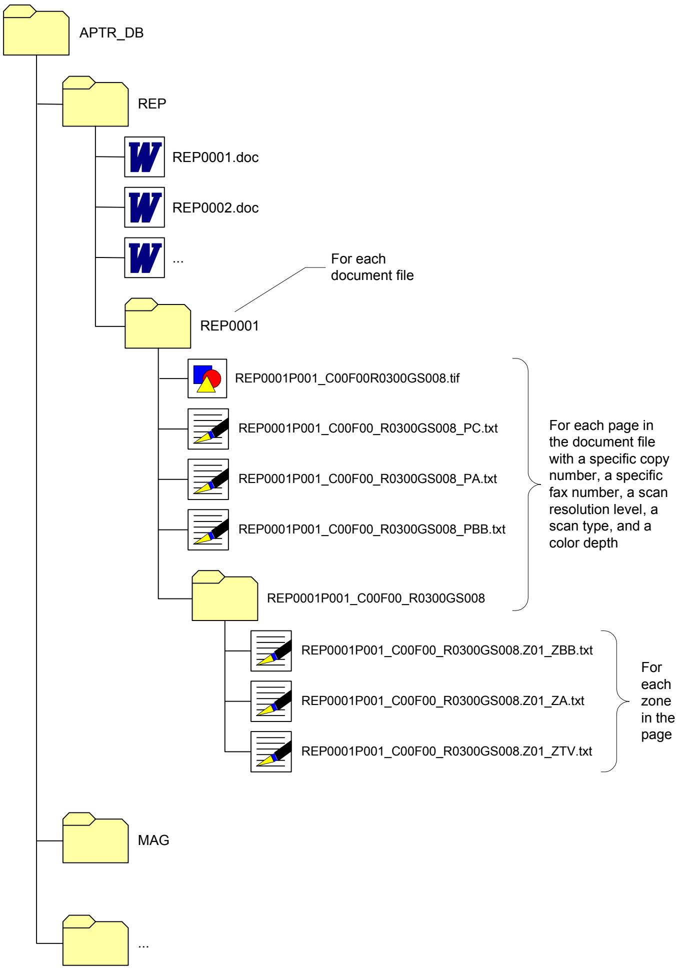

APTR_DB
REP
REP0001.doc
REP0002.doc
...
For each document file
REP0001
REP0001P001_C00F00R0300GS008.tif
REP0001P001_C00F00_R0300GS008_PC.txt
REP0001P001_C00F00_R0300GS008_PA.txt
REP0001P001_C00F00_R0300GS008_PBB.txt
For each page in the document file with a specific copy number, a specific fax number, a scan resolution level, a scan type, and a color depth
REP0001P001_C00F00_R0300GS008
REP0001P001_C00F00_R0300GS008.Z01_ZBB.txt
REP0001P001_C00F00_R0300GS008.Z01_ZA.txt
REP0001P001_C00F00_R0300GS008.Z01_ZTV.txt
For each zone in the page
MAG
...

Fig.2: PATDB Hierarchy

(dpi); ScanType is an abbreviated san type according to the 'scan types abbreviations' table; and lll is a 3-digit color depth in bits.

The copy number, referred to by cc, represents the number of consecutive copies that are done on the page before scanning it. Similarly, the fax number, referred to by ff, represents the number of consecutive faxes that are done on the page before the scanning process.

2. Three text files representing the page condition record, page attribute record, and page bounding box record are named as follows, respectively:
DocNamePppp_CccFff_RrrrrScanTypelll_PC.txt,
DocNamePppp_CccFff_RrrrrScanTypelll_PA.txt, and
DocNamePppp_CccFff_RrrrrScanTypelll_PBB.txt,
where DocName, ppp, cc, ff, rrrr, ScanType and lll are as defined earlier.
3. A directory for the page zones is named as follows:
DocNamePppp_CccFff_RrrrrScanTypelll,
where DocName, ppp, cc, ff, rrrr, ScanType and lll are as defined earlier.
This directory includes three text files for each identified zone in the page. These text files represent the zone bounding box record, zone attributes record and zone ground-truth value record and they are named as follows, respectively:
DocNamePppp_CccFff_RrrrrScanTypelll.Zzz_ZBB.txt,
DocNamePppp_CccFff_RrrrrScanTypelll.Zzz_ZA.txt, and
DocNamePppp_CccFff_RrrrrScanTypelll.Zzz_ZTV.txt,
where DocName, ppp, cc, ff, rrrr, ScanType and lll are as defined earlier and zz is a 2-digit zone number.

Figure 3 presents the above representation graphically. The figure shows in details a sample case from the report category (REP). The other types are similarly constructed.

6 Conclusion

The details of a comprehensive database of Arabic printed text are presented. It consists of scanned images of book chapters, advertisements, magazines, newspapers, and reports that are scanned at 200, 300, and 600 dpi resolutions. A total of 6954 pages are scanned. The database may be utilized by Arabic printed text recognition research community. Thus providing researchers and developers working on the field of automatic Arabic text recognition with benchmark database of printed Arabic text which enables them to collaborate remotely and compare algorithms and results. The database will soon be made freely available for researchers and developers. The database covers most of the written communications faced in real life. This database addresses one of the obstacles of conducting automatic Arabic printed text recognition research as to authors knowledge no such database is freely available to researchers.

References:

[1] J. J. Hull, *A Database for Handwritten Text Recognition Research*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 16 (5), May 1994, pp. 550-554.

[2] V. Märgner and M. Pechwitz, *Synthetic Data for Arabic OCR System Development*, ICDAR, Proceedings of the Sixth International Conference on Document Analysis and Recognition, 2001, pp. 1159-1163.

[3] I. T. Phillips et al., *The Implementation Methodology for CD-ROM English Document Database*, Proceedings of the Second International Conference on Document Analysis and Recognition, 20-22 Oct. 1993, pp. 484 - 487.

[4] I. T. Phillips, S. Chen and R. M. Haralick, *CD-ROM Document Database Standard*, Proceedings of the Second International Conference on Document Analysis and Recognition, 20-22 Oct. 1993, pp. 478-483.

[5] Introducing JSON. http://www.json.org/ (02-Feb-2008)

[6] www.erim.org (22-Jan-2008)

[7] R. Davidson and R. Hopely, *Arabic and Persian OCR Training and Test Data Sets*, Proc. of Symp. on Document Image Understanding Technology, 30 April-2 May 1997.