Sindh Univ. Res. Jour. (Sci. Ser.) Vol.47 (1):181-186 (2015)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Multilingual Text Image Database for OCR

D. N. HAKRO[++], A. Z. TALIB, G. N. MOJAI*

School of Computer Sciences, University Sains Malaysia (USM), Penang, Malaysia

Received 8th July 2014 and Revised 19th December 2014

Abstract: Optical Character Recognition (OCR) converts image text into editable text and provides a fast method for the data entry options. Most of the languages (scripts) of the world are enriched with OCRs. The research on Latin script has reached its maturity whereas the cursive languages are also converging to the same point. All of the OCRs, Latin or non-Latin based, need text images in their respective language script for training, testing as well as data validation and elimination of errors. This paper presents a multilingual image database created from various texts containing multiple fonts collected from various sources. We have created a multilingual, multi-font, multi-size, multi-style database of text images to be used for various language scripts of most of the world. The images were created with custom built software which converts all of the texts into word, single line, and multiline and paragraph images. We have included 84 language texts for the creation of images. The database includes Holy Quran translation of 42 languages comprising of more than 25000 text pages, various dictionaries of some other languages and scanned images of some languages like Sindhi, Urdu, Pashto and Arabic. This database can provide a platform to establish a standardized image database for the OCRs of various scripts around the world. The size of the database can be increased by adding texts and language scripts. The million words can grow to billions and can benefit more researchers working on OCR around the world and this database is freely available by sending email to the authors.

Keywords: Multilingual, Text image database, character Recognition, multi-script

1. INTRODUCTION

An Optical Character Recognition is a set of algorithms which recognizes text in a scanned image and converts it into an editable text (Rani *et al.* 2011).The text image might be in any language of the world. These text images written in various scripts are recognized by applying various algorithms namely segmentation, feature extraction and recognition. In order to analyze and evaluate the OCR performance, multi-font, multi-size and multi-style text images have to be created and used (Jaiem *et al.* 2013). These text images can be used for testing and training of the OCR. Validation of samples and elimination can also be achieved with the help of these text images (Abdel Raouf *et al.* 2008). To the best of our knowledge and survey there is no multi-script text image database available that can be used for most of the scripts of the world.

In this paper we present custom-built software that can be used to convert text file into text images for most of the language scripts of the world for testing and training of their OCRs. The database created from this software is multilingual and contains multi-color, multi-font, multi-sized, multi-style text images created with the custom built application. This application is also capable of creating grey level and binary images of the text. The database material was collected from various websites and required fonts were collected and used for creation of the database. Translation of Holy Quran in 72 languages and dictionaries of various languages were downloaded and converted into various options such as font family, size, style, foreground color, background color, grey level and binary images.

2. RELATED WORK

(Abdel *et al.* 2008) developed a database from collected Arabic words comprising 6 million words with and without diacritics. The database contains only Arabic words collected from websites, newspapers, books, Holy Quran script and a thesis. Robert Parker *et.al.*, (2011) created a database of Arabic words comprising multiple editions. The fifth edition contains 1004 files of Arabic text extracted from newspapers. Pechwitz *et al.*, (2002) created a database of Tunisian city names written by 411 subjects. The subjects were asked to write the names on a form. The database can be obtained by sending email to the authors. There is any centralized organization for collecting standard words for Arabic language as reported in (Ashraf *et al.* 2002; Slimane *et al.* 2009) and most of the research on Arabic OCR are benchmarked with private datasets so far. In (Alginahi, 2013) the author reports on the only Arabic related databases, The Environment Research Institute of Michigan (ERIM) database comprises 750 pages and it is difficult to access. The Arabic Printed Text Images (APTI) database comprises 10 fonts, 4 styles, 10 various font sizes and a total of 250 million characters (Slimane

[++]Corresponding author: Dil Nawaz, email:Dill.Nawaz@gmail.com

*Institute of Mathematics and Computer Science (Bioinformatics),University of Sindh, Jamshoro, Pakistan

et al. 2009). A database created by an Egyptian organization, the Arabic Language Technology Center (Altec, 2011) comprises books, magazines and theses scanned with various resolutions with digital cameras and scanners. According to their official website (Altec, 2011) commercial version contains the full database whereas the academic version is only a part of the database. For using this database fees is required for academic or commercial use.

Center for Excellence in Document analysis and Recognition (CEDAR) created a natural scene image database OCR (Ikica and peer, 2013). The database contains 100 binary annotated characters and 1000 un-annotated, 341 rectangles annotated and 7014 cropped characters. The database can be obtained by filling license agreement for non-commercial use and emailing it to the authors. Japanese Character Image Database was created at CEDAR. The dataset can be purchased at the price of $1500 (CEDAR, 2014). Chinese online database (Wang *et al,* 2009) is one of the series of databases (Liu *et al.* 2011) created by the National Laboratory of Pattern Recognition (NLPR) for Chinese script. The databases are available in a variety of forms. The database contains handwritten samples of 1020 subjects, online and offline.

Infty Project (Infty, 2014) produced character and symbol database of English mathematical articles. The database contains 476 text pages and 688,580 samples of extracted characters. A user must be registered with certain conditions for access of this database. A Devanagri handwritten text image database (Dongre and Mankar, 2012) contains 20305 characters and 5137 numerals written by 750 writers of various ages, profession and education. The database stored text images in a TIFF format as well in binary to reduce the memory size. The database is freely available.

3. DEVELOPING A CUSTOM-BUILT SOFTWARE FOR CREATING MULTILINGUAL TEXT IMAGE DATABASE

A software was developed using MATLAB 2014a for the creation of text images for various language scripts. For this we collected text files using various scripts **(Fig.1).** We collected Holy Quran translation text for 72 languages, dictionaries for several languages, books and other materials for the remaining languages. The software takes a text file and user selects font family name, font size, font style and font weight as input and creates multiple text images containing words, lines, multiline and paragraph text images which can be easily used for testing and training of various scripts as shown in **(Fig.2)**.

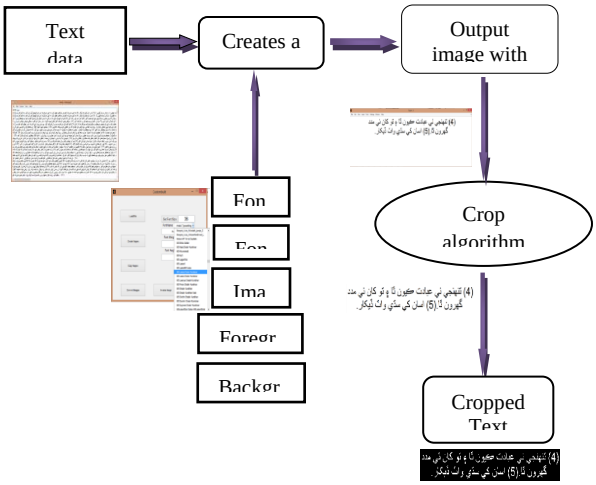


Fig. 1: Working diagram of the custom built software

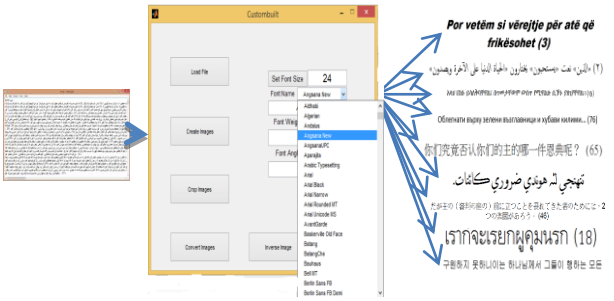


Fig 2: **Selecting options for creating images**

The output images created by the custom-built software contains some extra spaces in the background and those extra spaces are cropped with an algorithm which crops all extra spaces from the background and a text image is ready for testing and training as shown in **(Fig 3).** The image can be converted into grey level by pressing convert image button. The whole directory images are converted once the invert image button is pressed. The conversion of white background image to black is given in **(Fig 4).**

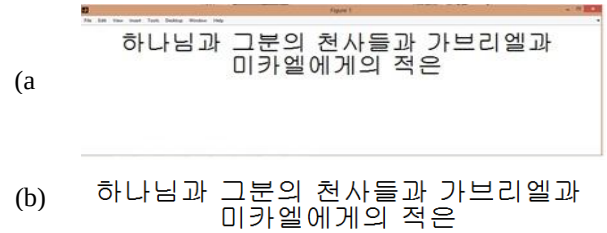


Fig 3: Cropping free spaces from a text image (a) before (b) after

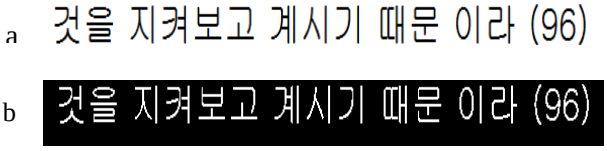


Fig 4: Image complement (inverse) (a) before (b) after

3.1 Creating Text Images for Scripts

Arabic script: Arabic script is an ancient script and widely used for writing various languages in Asia as well as Africa beside the Latin and Chinese scripts. Arabic script is adopted by languages such as Persian, Urdu, Pashto, Punjabi (Shahmukhi), Uyghur, Kashmiri, Kurdish and Sindhi which has the largest extension of the original Arabic script. Significant efforts have been made to develop OCR systems for these languages. This database is helpful in creating, testing and training text images of these languages. **(Fig. 5)** shows examples of Arabic, Persian (Farsi), Urdu, Uyghur and Sindhi text created, with white spaces cropped using the custom built software.

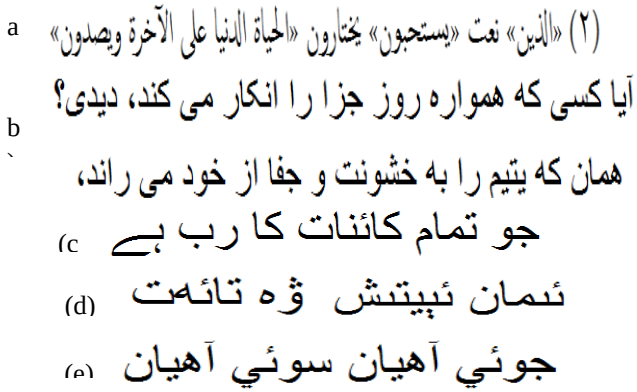


Fig 5: Examples of Arabic script language text images created with the custom built application: (a) Arabic (b) Persian (c) Urdu (d) Uyghur (e) Sindhi

3.2 Latin Scripts

Latin script is also known as Roman script which is based on the old Latin alphabet. Latin script is the largest base of writing system of many languages compared to the other scripts. Latin is the most widely used by more than 65% of the population of the world. Many of the European languages such as French, German, Spanish, Swedish and many others use Latin script for writing. Some other languages such as Malay and Turkish also follow the Latin script for writing. **(Fig. 6)** illustrates Azerbaijani, French, German, Turkish and Greek language text images

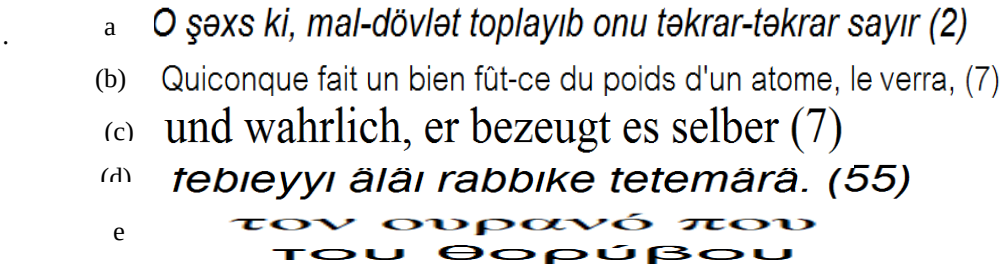


Fig 6: Examples of Latin script language text images created with the custom built application (a) Azerbaijani (b) French (c) German (d) Turkish (e) Greek

3.3 Indian Scripts

Various Indian scripts are used to write many languages such as Hindi, Malayalam, Tamil, Bengali, Oriya, Gurumukhi, Gujarati, Telugu, Kannada and Marathi and some others. The image database is equally helpful in the training and testing of these scripts. **(Fig. 7)** shows the example images created in Hindi, Malayalam, Tamil, Bengali and Kannada respectively.

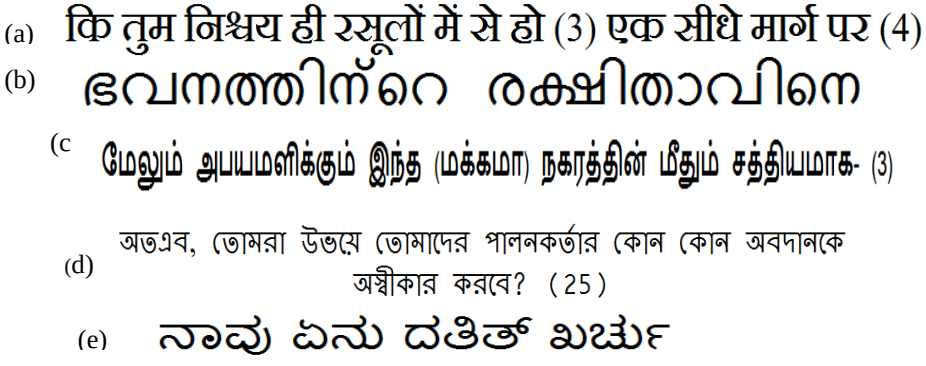


Fig 7: Examples of Indian language text images created with the custom built application (a) Hindi (b) Malayalam (c) Tamil (d) Bengali (e) Kannada

3.4 Other scripts

Some other scripts such as Amharic, Chinese, Japanese, Korean and others were also included in this image database to increase versatility and usefulness. A total of 84 language scripts were converted into images and a database was created for the text of these languages. **(Fig. 8)** illustrates the images of Amharic language text which is spoken in Ethiopia, Thai which is mostly spoken in Thailand, Chinese, Japanese and Korean respectively.

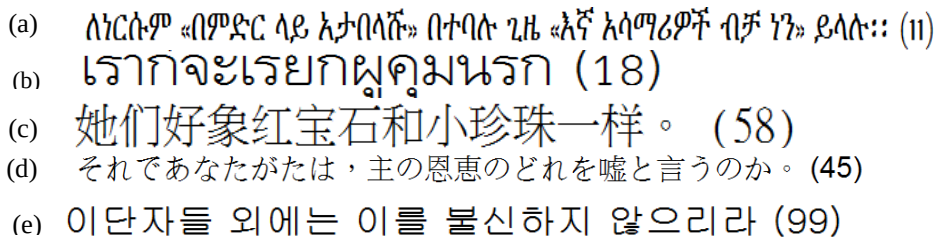


Fig 8: Examples of Other language text images created with the custom built application
(a) Amharic (b) Thai (c) Chinese (d) Japanese (e) Korean

4 Analysis of the Database

The database contains translation of 72 languages of the Holy Quran and dictionaries collected from various sources and scanned books of some of the languages. The sources included websites, books, newspapers and magazines. The Number of words and characters excluding scanned images are given in Table 1. We collected from friends and downloaded more than 2500 various fonts for these scripts and for the sake of simplicity; only one font is given in **Table 1.**

Table 1: Language script and number of words and characters

S. NO	Language script	Number of Words	Number of Char	Font	S.N	Language script	Number of Words	Number of Char	Font
1	Afrikaans	197767	416575	Afrikaans-Tulasi	43	Khannada	130957	674113	UMA_Kannada
2	Albanian	143493	989796	Albanian-IPAPhon	44	Khmer	117299	792658	Khmer-Anlongvill
3	Amazigh	105096	856829	DOS_Amazigh-1	45	Korean	397112	400168	Gulim
4	Amharic	90329	559181	Nyala , goffer	46	Kurdish	187844	1086818	Lateef
5	Arabic	213187	989796	Thabit	47	Lao	127925	817952	AloBravo
6	Armenian	145798	659383	Vrinda	48	Latin	118100	518737	Times New Roman
7	Azerbaijani	134039	856829	Azeri Latin	49	Latvian	153330	627630	Latvian-Balt
8	Basque	142532	695315	Basque-CrrCTT	50	Lithuanian	118544	627926	Lithuanian-Noxchi
9	Belarusian	169095	633843	Bierahinia	51	Macedonian	170624	635938	Mace:Aachen
10	Bengali	121157	559181	Bengali-Progoty	52	Malay	188386	1192506	Malay-ER Bukinist
11	Bosnian	144849	703352	Alegreya	53	Malayalam	101756	1024199	Karthika
12	Bulgarian	130994	593596	Bulgarian-Bookvar	54	Maltese	159246	661343	Maltese-CyrCauc
13	Catalan	180608	655500	Catalan-SPAtlantis	55	Marathi	121996	616721	Shivaji01
14	Cebuano	208072	829526	Nuevo Cebuano	56	Norwegian	131758	563064	Norwegian-Lazurski
15	Chinese	232084	245618	MingliU	57	Oriya	112980	112980	Sarala
16	Croatian	159850	595162	Croatian-Croat	58	Persian	134026	516149	Persian-Farsi 1.1
17	Czech	125965	590490	Czech-TNR Liborius	59	Polish	126766	660406	Polish-Cieszyn
18	Divehi	148818	1287518	a_haleem sh	60	Portuguese	125398	645910	Portuguese-Vera
19	Dutch	135592	654329	Dutch-ER Courier	61	Punjabi (Guru)	145488	555721	Gurakhar
20	English	87654	377093	Arial	62	Romanian	134002	633584	Romanian-Ciula
21	Esperanto	161465	624191	Esperanto-Espar	63	Russian	243498	1302594	Russian-KOI8 Courier
22	Estonian	140301	590626	ČASY EA / Abril	64	Serbian	139586	576857	Serbian-Cyrillic
23	Filipino	198494	831437	Filipino-BPGClassic	65	Sindhi	2089618	805214	MB latifee
24	Finnish	133435	664585	Finnish-Arial Ykt	66	Slovak	165890	590712	Slovak-Alice0 Lao
25	French	150470	722791	ArTarumianIshxan	67	Slovenian	156374	602582	Slovene-Great
26	Galician	155968	631895	ABeeZee	68	Somali	108054	614379	SomaliSerif
27	Georgian	125151	659408	Faruma	69	Spanish	164128	768954	IbarraReal-Regular
28	German	148128	751290	German-VPS	70	Swahili	117965	636801	Swahili-LeedsCyr
29	Greek	191184	733120	Greek-Ismini	71	Swedish	155499	692283	Swedish-Sans Uralic
30	Gujarati	155142	579076	Avantika	72	Tajik	120208	577412	Tajik Sophia
31	Haitian Cr:	185206	612246	Dosis	73	Tamil	121700	1073197	latha
32	Hausa	165360	706569	Hausa-Clarendon	74	Tatar	170994	1075120	Irma
33	Hebrew	123943	480722	Hebrew-Sil Ezra	75	Telugu	132843	640536	akshar
34	Hindi	167370	629972	Hindi-Pooja	76	Thai	34440	886561	Thai-CordiaDSE
35	Hmong	233822	815548	Latin-X	77	Turkish	104135	647365	Turkish-Lazuri
36	Hungarian	150914	625904	Hungarian-ER Kurier	78	Ukrainian	167838	623094	Ukrainian-ER Bukinist
37	Icelandic	164535	612715	Icelandic-Alice3 Lao	79	Urdu	212152	694920	Urdu-AlKatib1
38	Indonesian	146932	925168	Indonesian-Et tu,Brute	80	Uyghur	137157	927559	UKIJBasma
39	Irish	193502	707469	Latin-X	81	Uzbek	200097	1305762	XB Shafigh Uzbek
40	Italian	133447	648943	Italian-CrrCTT	82	Vietnamese	212029	668012	Hannom
41	Japanese	390887	404310	MS Gothic	83	Welsh	186858	668086	Welsh-Persian
42	Javanese	191924	829477	TuladhaJejeg_gr	84	Yiddish	165952	679093	Ain Yiddishe
Total Number of Words			**15135111**		**Total Number of Characters**			**58957990**	

The database contains 15135111 words comprising 58957990 characters excluding scanned pages collected from various sources. Font name given in Table 1 is only for reference purpose whereas other fonts including the above were also used for creating database. The images were created in words, lines, multiple lines and paragraphs. The images were used for experimenting purpose and some of the optical character recognition experiments were performed to test the database as well as segmenting of words into characters so that a base for a multilingual and multi-script optical character recognition system is developed. This research is a part of multilingual and multi-script optical character recognition.

5. Experiments performed on Database

Some of the experiments for segmentation, thinning and feature extraction were performed to test the database so that the database and segmentation and feature extraction algorithms can be benchmarked. A multi-script OCR system may be developed on the basis of these experiments and training of neural networks for various scripts; which needs extremely large number of experiments. Some of the basic experiments were performed on scripts and results are shown in **(Fig. 9)**

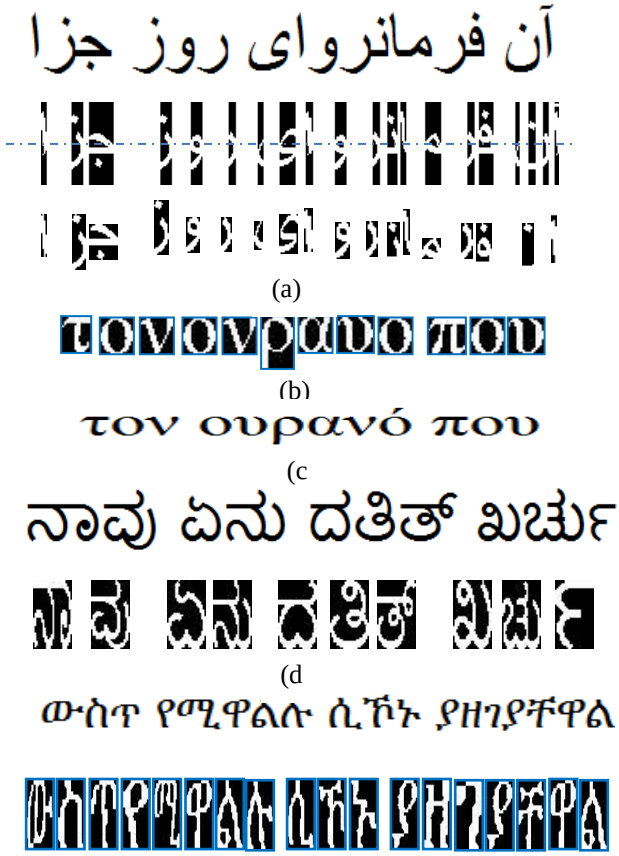


Fig 9: segmentation of (a) Persian (b) Greek (c) Kannada and (d) Amharic scripts

6. CONCLUSION AND FUTURE WORK

OCR is developed on the basis of testing and training dataset of a particular language or script. The OCRs of various scripts are developed, trained and tested on private databases. For this reason we have created a multi-font multi-size and multi-style image database of various languages to provide a single platform for testing and training of OCRs. The database contains 84 languages including Arabic script adopting languages, Latin script, Indian scrip and other script languages. The database contains 15135111 words and 58957990 characters of scripts converted into text images. This database can provide a platform to establish a standardized image database for the OCRs of various scripts around the world. The database was created by using custom built software in MATLAB 2014a and scanned pages of script text. The custom built software can generate more text of existing languages as well additional language can be employed. The size of the database can be increased by adding texts and language scripts. The million words can grow to billions and can benefit more researchers working on OCR around the world. This database is freely available by sending an email to the authors.

REFERENCES:

AbdelRaouf, A., C. Higgins, and M. Khalil, (2008), A Database for Arabic Printed Character Recognition, in Image Analysis and Recognition, Springer Berlin / Heidelberg, 567-578.

Alginahi, Y. M. (2013), A survey on Arabic character segmentation, International Journal on Document Analysis and Recognition (IJDAR), 16(2), 105-126.

ALTEC, ARABIC LANGAUGE TECHNOLOGY CENTER; http://www.altec-center.org Accessed on 09/08/2014

Ashraf, A., A.H. Colin, and K. Mahmoud, (2008), A database for Arabic printed character recognition, International Conference on Image Analysis and Recognition, ICIAR, 567–578.

Center of Excellence for Document Analysis and Recognition, CEDAR, http://www.cedar.buffalo.edu/ Accessed on 26/08/2014.

Dongre, V. J., and V. H. Mankar, (2012), Development of comprehensive Devanagri numeral and character database for offline handwritten character recognition, Applied Computational Intelligence and Soft Computing, 29-34, Article ID 871834.

Ikica, A., and P. Peer, (2013), SWT voting-based color reduction for text detection in natural scene images,

EURASIP journal on advances in signal processing, (1), 1-13.

Infty, The Research Project on Mathematical Information Processing, http://www.inftyproject.org/en/database.html Accessed on 26/08/2014.

Jaiem, F. K., S. Kanoun, M. Khemakhem, H. El Abed, and J. Kardoun, (2013), Database for Arabic Printed Text Recognition Research, in Image Analysis and Processing–ICIAP, 251-259, Springer Berlin Heidelberg.

Liu, C. L., F. Yin, D. H. Wang, and Q. F. Wang, (2011), CASIA online and offline Chinese handwriting databases, Document Analysis and Recognition (ICDAR), International Conference on 37-41, IEEE.

Parker R., D. Graff, Ke Chen, K. Junbo and Kazuaki M (2011).Arabic Gigaword Fifth edition, Linguistic Data Consortium, University of Pennsylvania.

Pechwitz, M., S. S. Maddouri, V. Märgner, N. Ellouze, and H. Amiri, (2002), IFN/ENIT-database of handwritten Arabic words, In Proc. of CIFED Vol. 2, 127-136. Citeseer.

Rani, R., R. Dhir, and G. Lehal, (2011), Identification of Printed Punjabi Words and English Numerals Using Gabor Features, World Academy of Science, Engineering and Technology 73, 392-395.

Slimane, F., R. Ingold, S. Kanoun, M.A. Alimi, J. Hennebert, (2009), A new Arabic printed text image database and evaluation protocols, Proceedings of 10th IEEE International Conference on Document Analysis and Recognition ICDAR, 946–950. Barcelona (Spain).

Wang, D. H., C. L. Liu, J. L. Yu, and X. D. Zhou, (2009), Casia-olhwdb1: A database of online handwritten Chinese characters, Document Analysis and Recognition, 2009. ICDAR09, 10th International Conference, 1206-1210, IEEE.