Preface

Arabic Scripts

The Internet is a source of information which is used by approximately 2 billion users worldwide (www.internetworldstats.com). Two aspects affect the way in which the Internet is used, especially in searching for documents. One is the availability of more and more scanned documents with varying amounts of metadata for searching these documents on the Internet, and the other os the appearance of more and more languages with non-Latin characters. Both aspects show the importance of developing recognition technology for all types of characters and languages to make the content of scanned images of text available to the Internet users. A worldwide-used acronym for any type of text recognition is OCR, which means optical character recognition. OCR is used not only for recognizing printed characters, but it is often also used for cursive handwriting, even when words instead of single characters are recognized. Some alternative acronyms are used for the case of handwritten words, like HWR (handwritten word recognition) but these are not in common use today.

Knowing that about 200 million people in the world use Arabic as their first language it is obvious that a growing interest of that huge group of Arabic-speaking Internet users is to search for documents in their mother tongue. In parallel to this situation is in the past few years a growing interest in Arabic word and text recognition has been observed. During that time two events have been important landmarks in Arabic text recognition technology development. In 2002 a database on Arabic handwritten words (*IFN/ENIT-database*)[1] was made available to the community and has served as a reference for competitions since 2005 (ICDAR 2005).[2] In September 2006 a summit on Arabic and Chinese Handwriting Recognition was held at College Park, MD in the USA (SACH2006),[3] where experts from both re-

[1] http://www.ifnenit.com

[2] http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=10526

[3] http://www.umiacs.umd.edu/lamp/meetings/SACH06/

Table 1 Arabic characters (www.ethnologue.com)

IPA	Value	Name	Final	Medial	Initial	Isolated	IPA	Value	Name	Final	Medial	Initial	Isolated
[ɖ]	ḍ	ḍād	ض	ضـ	ضـ	ض	[ʔ]	ʾ(a)	alif	ـا	ـ	ـ	ا
[ʈ]	ṭ	ṭāʾ	ط	طـ	طـ	ط	[b]	b	bāʾ	ـب	بـ	بـ	ب
[ƶ]	ẓ	ẓāʾ	ظ	ظـ	ظـ	ظ	[t]	t	tāʾ	ـت	تـ	تـ	ت
[ʕ]	ʿ	ʿayn	ع	عـ	عـ	ع	[θ]	th	thāʾ	ـث	ثـ	ثـ	ث
[ɣ]	gh	ghayn	غ	غـ	غـ	غ	[ʒ]	j	jīm	ـج	جـ	جـ	ج
[f]	f	fāʾ	ف	فـ	فـ	ف	[ħ]	ḥ	ḥāʾ	ـح	حـ	حـ	ح
[q]	q	qāf	ق	قـ	قـ	ق	[x]	kh	khāʾ	ـخ	خـ	خـ	خ
[k]	k	kāf	ك	كـ	كـ	ك	[d]	d	dāl	ـد	ـ	ـ	د
[l]	l	lām	ل	لـ	لـ	ل	[ð]	dh	dhāl	ـذ	ـ	ـ	ذ
[m]	m	mīm	م	مـ	مـ	م	[r]	r	rāʾ	ـر	ـ	ـ	ر
[n]	n	nūn	ن	نـ	نـ	ن	[z]	z	zāy	ـز	ـ	ـ	ز
[h]	h	hāʾ	ه	هـ	هـ	ه	[s]	s	sīn	ـس	سـ	سـ	س
[w]	w	wāw	و	ـ	ـ	و	[ʃ]	š	shīn	ـش	شـ	شـ	ش
[j]	y	yāʾ	ي	يـ	يـ	ي	[sˁ]	ṣ	ṣād	ـص	صـ	صـ	ص

Fig. 1 Example of an Arabic printed text

اَلسَلَام عَلَيكُم بِسمِ اللهِ الرَّحمٰنِ الرَّحِيمِ
شكرًا لمشَاركتكم كتَاب التعرف علَى الخط العربي

search fields presented their actual work. From that time intensive research on Arabic script recognition started and has resulted in a big step forward today.

Arabic script is the second most widespread script in the world; it is used not only for Arabic but also for the Persian, Urdu, and Pashto languages, for example. Today 14 languages use Arabic script worldwide, which shows its importance. Characteristics of Arabic script are a writing direction from right to left, characters within a word being mostly connected, 28 characters with different shapes for different positions in a word, and dots and diacritical signs above and below characters. Table 1 shows all the shapes of the 28 Arabic characters.

For different languages some additional characters may be used. Typical for Arabic script is also the variation of a word in length by elongation of the connecting lines between the characters. Figure 1 shows an example.

Table 2 Examples of ligatures

without ligatures	ligatures
مُحَمَّدٌ	مُحَمَّدٌ
اَلنبّي	اَلنبي

A further important special Arabic script style is the possibility to write characters as vertical or horizontal ligatures. These ligatures modify the shape of the characters significantly. Some examples of ligatures are shown in Table 2. All these typical Arabic script characteristics influence the processing and recognition of Arabic script in different ways and make it clear that a simple adaptation of Latin character-based processing is not possible.

This book presents the state of the art of OCR for Arabic scripts presented from most active and successful groups. The parts of the book show that a lot of work still has to be done on Arabic script recognition. But the techniques and algorithms used are of general interest; many problems are typical not only for Arabic but for many other scripts. We believe that the collection of Arabic OCR related work is also an inspiration for other scripts and vice versa.

The book is divided into four parts. Part I, Pre-processing, presents different aspects of pre-processing and feature extraction for Arabic OCR systems. Part II, Recognition, includes chapters with details about different recognition approaches. Part III collects chapters describing the important aspects of how to assess the performance of a recognition system. The final Part IV, Applications presents system solutions for selected application fields.

Part I: Pre-processing

Part I presents different approaches for the pre-processing of OCR systems for Arabic. It starts with an overview of Arabic handwriting recognition technology. Srihari and Ball present in their chapter the parts of a recognition system from pre-processing to classification. Finally they discuss application fields and challenges. Chapters 2–6 deal with pre-processing tasks of an OCR system. Bukhari, Shafait, and Breuel discuss layout analysis methods, Setlur and Govindaraju pre-processing issues, Belaid and Ouwayed segmentation of ancient Arabic documents, and Likforman-Sulem et al. features for word recognition systems.

Part II: Recognition

Chapters 7–15 present different approaches for the recognition of Arabic script. The first six chapters all use HMM-based approaches. Borovikov and Zavorin present a multi-stage approach to document analysis, Ahmed, Mahmoud, and Parvez a recognizer for printed Arabic text, Pechwitz, El Abed, and Märgner an offline handwritten

Arabic word recognizer, Dreuw, Rybach, Heigold, and Ney a large vocabulary optical character recognition system, Alkhoury, Gimenez, and Juan a Bernoulli-based handwriting recognition system, Jifroodian and Suen a handwritten Farsi word recognition system, Kessentini, Paquet, and Ben Hamadou a multi-stream Markov model recognizer.

Two chapters discuss further approaches. Graves presents a recognition system based on multidimensional recurrent neural networks, and Mozaffari discusses the application of fractal theory for document analysis and recognition. Khemakhem and Belghith discuss an OCR system based on the combination of complementary systems in Chap. 15.

Part III: Evaluation

The subject of Part III is the evaluation of recognition systems. In Chap. 16 Zavorin and Borovikov discuss data collection and annotation, and Arabic handwriting recognition competitions are described in Chap. 17 by Märgner and El Abed. In Chap. 18, Slimane et al. describe benchmarking strategies for Arabic word recognition.

Part IV: Applications

The final Part IV presents different applications using Arabic script recognition technology. In Chap. 19 Cheriet and Moghaddam present a robust word spotting system for historical Arabic manuscripts. Natarajan discusses, in Chap. 20, script-independent methods for Arabic handwriting recognition, and Kundu and Hines present an Arabic handwriting recognition system using over-segmentation in Chap. 21. Boubaker et al. discuss online Arabic databases and applications using these data in Chap. 22, and Abdelazeem et al. present, in Chap. 23, techniques for using online and offline features for Arabic handwriting recognition.

Target Audience

This book provides an overview of the state-of-the-art research in the field of OCR for Arabic scripts. Different aspects and solutions have been addressed by the authors, and we hope that this comprehensive collection of ideas, problems, and solutions motivates researchers to continue this work. In that sense this book shall serve as a reference for researchers and graduate students studying OCR technology and methodology in general and for Arabic script in particular.

Braunschweig, Germany

Volker Märgner
Haikal El Abed

http://www.springer.com/978-1-4471-4071-9

Guide to OCR for Arabic Scripts
Märgner, V.; El Abed, H. (Eds.)
2012, XX, 592 p., Hardcover
ISBN: 978-1-4471-4071-9