

Abdel Belaïd and Mohamed Imran Razzak

Contents

Introduction..... 428

Writing Systems..... 429

Middle Eastern Writing Systems..... 430

 Syriac Writing System..... 430

 Arabic Writing System..... 432

 Hebrew Writing System..... 439

Writing Recognition Systems..... 441

 Preprocessing..... 441

 Word Segmentation..... 442

 Isolated Character Recognition..... 444

 Word Recognition..... 445

Success Clues of the Topic..... 448

 Datasets..... 448

 Academic Systems..... 448

 Commercial Systems..... 450

Conclusion..... 452

Cross-References..... 453

References..... 453

 Further Reading..... 457

A. Belaïd (✉)
Université de Lorraine – LORIA, Vandœuvre-lès-Nancy, France
e-mail: abdel.belaïd@loria.fr

M.I. Razzak
Department of Computer Science and Engineering, Air University, Pakistan
College of Public Health and Health Informatics, King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia
e-mail: imranrazak@hotmail.com

Abstract

The purpose of this chapter is to give a rapid synthesis of the state of the art concerning Middle Eastern character recognition systems. This includes the presentation of the different scripts used, their characteristics, the techniques used especially for them, and the difficulties and challenges faced by researchers to treat them. The chapter will also make a rapid review of the best systems evaluated during international competitions, of published datasets specialized in these scripts, as well as of the list of existing commercial systems.

Keywords

Middle Eastern writing systems • Arabic • Hebrew • Syriac • Thaana • Word segmentation and recognition • Character recognition

Introduction

The scripts used in the Middle East mainly include Arabic, Hebrew, Syriac, and Thaana. They have a common origin dating back to the ancient Phoenician alphabet. They are all alphabetical with small character sets. As a consequence, there are a lot of morphological similarities which have to be taken into account in handwriting recognition systems.

Main Differences with Latin: Middle Eastern scripts are more complex than Latin scripts. The complexity comes from their inherent morphology, such as context-sensitive shape, cursiveness, and overlapping of word parts. Except for Thaana, they all include diacritical marks. These marks accompany the characters, either above, below, or inside, in order to represent vowel sounds. The text is written from right to left. Among these scripts, Arabic and Syriac are cursive even when they are typeset. In those scripts, the character shapes change according to the relative position with respect to the word, as opposed to Hebrew and Thaana. Rules for text rendering are specified in the block for Arabic and Syriac, whereas this is not required for Hebrew, as only five letters have position-dependent shapes and final shapes are coded separately.

Challenges: Context identification is necessary prior to recognition, since the character shape changes according to the surrounding characters. However, this is not straightforward, especially when the ligature is vertical, which can happen frequently when the text is handwritten. Moreover, the size of each ligature varies due to cursiveness and size of neighbor characters. Thus, the ligature has no symmetry in height and width, which makes the very similar contours difficult to segment and to recognize [8] (cf. Fig. 13.1).

Another source of complexity is the extensive amount of dots associated with characters. Their positions vary with respect to corresponding characters.

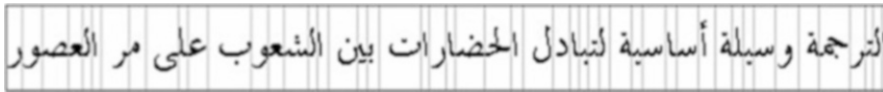


Fig. 13.1 Grapheme segmentation process illustrated by manually inserting *vertical lines* at the appropriate grapheme connection points

For example, they are often attached to characters and sometimes depicted as small lines. As a consequence, baseline detection affects segmentation heavily and obviously. Different graphemes do not have a fixed height or a fixed width. Moreover, neither the different nominal sizes of the same font scale linearly with their actual line heights nor the different fonts with the same nominal size have a fixed line height.

Finally, words are composed of sub-words, which can be considered as an advantage, but often these sub-words provide some additional confusion in segmentation because of non-heterogeneous white space separation between them.

The remainder of this chapter is organized as follows. Section “[Writing Systems](#)” presents background on the writing systems, describing their derivation tree from the origin. Section “[Middle Eastern Writing Systems](#)” presents more writing systems of the Middle East by giving details about each of them such as the scripts and the morphological peculiarities. It also includes tables that classify the scripts according to their writing aspects. Section “[Writing Recognition Systems](#)” describes the writing recognition systems. It outlines the different steps needed for designing recognition systems. Section “[Success Clues of the Topic](#)” provides some clues to the success of this research through shared databases, the academic systems and the number of commercial systems. Section “[Conclusion](#)” proposes some observations about the scripts studied and their concomitant recognition systems.

Writing Systems

According to Ghosh et al. [40], there are four classes of writing systems: logographic, syllabic, featural, and alphabetic (see Fig. 13.2), described as follows:

- The logographic script refers to ideograms representing complete words as “Han” which is mainly associated with Chinese, Japanese, and Korean writings. They are composed of several harmoniously associated strokes.
- In the syllabic system, each symbol represents a phonetic sound or syllable. This is the case of Japanese which uses both logographic Kanji and syllabic Kanas. The stroke gathering is less dense than in the logographic scripts.
- In the featural alphabet, the shapes of the letters encode phonological features of the phonemes represented. This is the case of Korean Hangul, where the five main consonants repeat the shape of the lips and tongue when producing the corresponding sound and remind the five basic elements of eastern philosophy.
- The alphabetic class consists of characters reproducing phonemes of the spoken language. Each class is subdivided into three subclasses: true alphabetic,

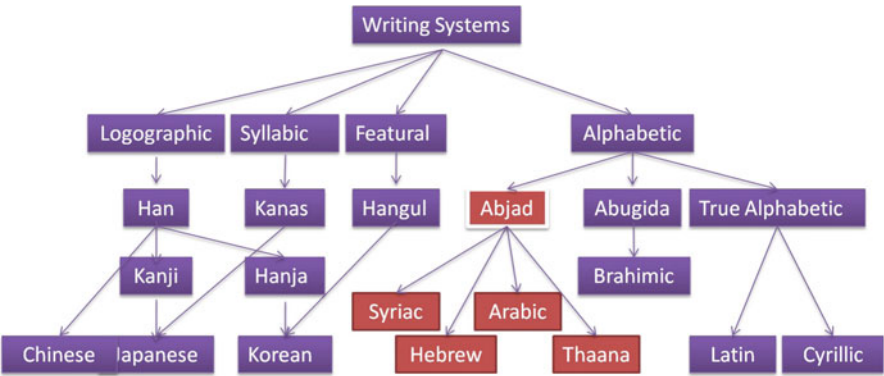


Fig. 13.2 Writing systems (From [40])

ABUGIDA, and ABJAD. The true alphabetic corresponds to Latin and Cyrillic. ABUGIDA is related to Brahmic scripts, while ABJAD refers to Arabic and Hebrew.

In fact, ABJAD (أبجد) is an acronym derived from the first four consonants of the Hebrew/Arabic/Persian alphabets: “Alif,” “Ba,” “Jeem,” and “Dal.” It expresses the fact that the vowels are not spelled out and characters are proposed “essentially” for consonantal sounds. It contrasts with ABUGIDA in which each symbol represents a syllable (consonant+vowel). This is similar in Hindi or Amharic, as well as an alphabet where each symbol represents a smaller unit of sound and the vowels are not omitted.

Middle Eastern Writing Systems

The Middle Eastern writing systems are part of the ABJAD writing system and have the same Phoenician origin. They include Arabic, Hebrew, Syriac, and Thaana. This group is divided into East Semitic (including the Akkadian), North west Semitic (Canaanite Phoenician, Hebrew, Aramaic), and Southern Semitic. Arabic is the latter group with the Ethiopian and South Arabian groups.

Syriac Writing System

The Syriac language has been primarily used since around the second century BC as one of the Semitic ABJADs directly descending from the Aramaic alphabet. It shares similarities with the Phoenician, Hebrew, and Arabic alphabets.

The alphabet has 22 letters that can be linked or not, depending of their position in the word. There are three main types of fonts: Estrangelo; Western, Jacobite or Serto script; and Eastern or Nestorian script (see Figs. 13.3–13.5). Serto and Estrangelo

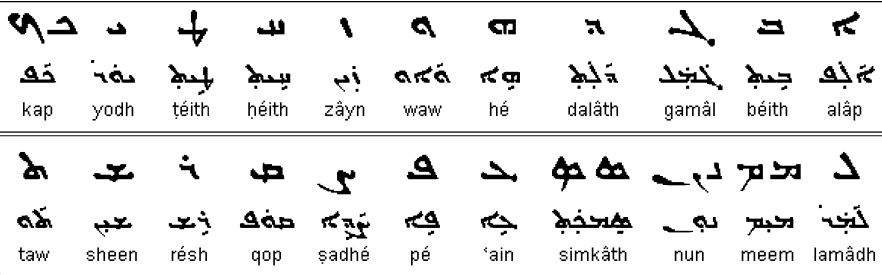


Fig. 13.3 Syriac Estrangelo (From <http://www.omniglot.com/writing/syriac.htm>)

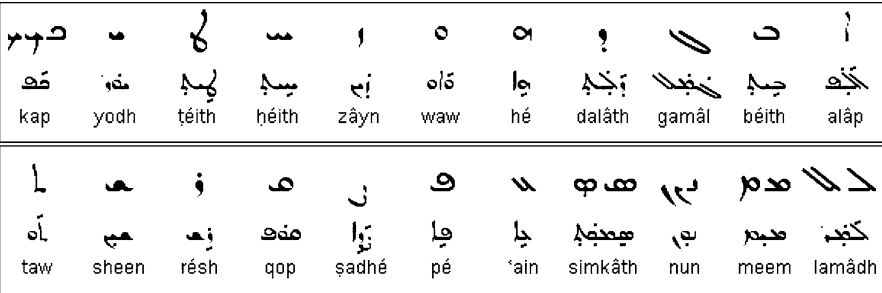


Fig. 13.4 Western, Jacobite or Serto script (From <http://www.omniglot.com/writing/syriac.htm>)

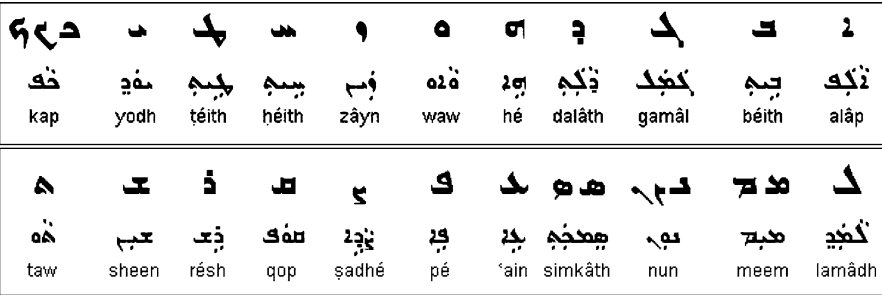


Fig. 13.5 Eastern or Nestorian script (From <http://www.omniglot.com/writing/syriac.htm>)

differ both in style and use. The Estrangelo script is used more for titles, whereas Serto is used for the body of the documents.

The letters are cursive, depending on where in the word or sub-word they are located, and thus take different shape. They can be pointed with vowel diacritics but can be written unpointed, which means that semantics are required for interpretation. Figure 13.6 shows a Syriac sample text in Nestorian script.

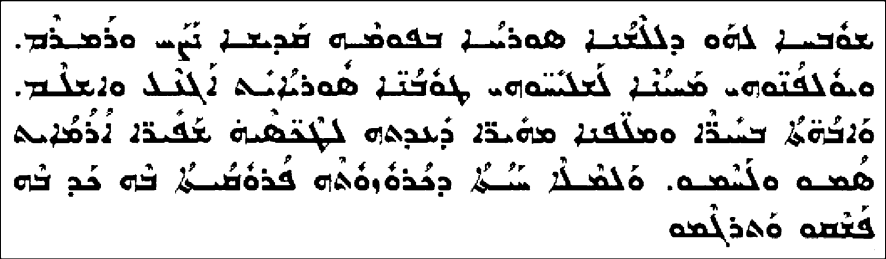


Fig. 13.6 Syriac text (From <http://www.omniglot.com/writing/syriac.htm>)

As reported in [67], some characters seem quite similar, with only small differences. This is, for instance, the case of “dalâth” and “resh” “taw” in the beginning of a word or sub-word and “alâp” at the end of a word or sub-word; “béith” and “kap” in the middle of a word or sub-word and at the beginning of a word or sub-word; and “waw,” “qop,” “lamâdh,” “héith,” and “yodh.” The characters “dalâth” and “resh” are differentiated by the presence of a single dot. Such a brittle feature does not provide high confidence in recognition. Other than the location of the dot, both “dalâth” and “resh” appear as identical. The starting character “taw” and the finishing one “alâp” have similar structures. They are both represented by a vertical line with a small stroke on the right hand side. However, the “alâp” character has a small stroke or tail that does not finish at the point where it joins the horizontal stroke. Table 13.1 traces the similarities and differences between letters.

As in other Semitic languages, Syriac words are built out of trilateral roots, which is a collection of three Syriac consonants and vowels serving as a “glue.” Figure 13.7 shows an example of words generated from the root.

Arabic Writing System

The languages written with the Arabic writing system are mentioned in Table 13.2. Its evolution is traced in the following paragraphs:

The Origin: Arabic writing appeared in the sixth century. It took its origins in the Phoenician script. It has inherited its 22 letters and the use of ligatures to connect a letter to the next. Therefore, most Arabic letters can take four forms: isolated, initial, medial, and final.

The Koranic Addition: The arrival of Islam and the idea to make the Koran a stable reference have profoundly marked the history of the Arabic writing. Several challenges have emerged, particularly in terms of lack of expressiveness of the language with just 22 letters (see Fig. 13.8). Thus, in addition to the 22 letters of the Phoenician alphabet, Arab philologists invented 6 new letters.

Table 13.1 Aramaic letters that look alike (From <http://allthingsaramaic.com/aramaic-letters-alike.php>)










Characters	Differences and resemblances
 (Beṭ) and (Kap).	Kap is more rounded than Beṭ. Beṭ has straight lines with sharp edges, and the line at the bottom right corner sticks out
 (Gamal), (Noon) and (Kap).	Gamal has a “foot” that exceeds on the right. Noon is square on the edges, without foot. Both Gamal and Noon have about half the width of other letters; for cons, Kap is larger and rounder
 (Dalat), (Resh) and (Final Kap).	Dalat has a tittle which sticks out at the top right corner, whereas Resh is rounded. Dalat is squarer than Resh; they are different from Kap (final) because they remain above the baseline
 (Heh), (Chet) and (Taw).	Heh and Chet have similar shapes, but Heh has a gap in the left vertical line, whereas Chet does not, are different from Taw which is more rounded at the top right corner
 (Waw), (Zeyn) and (Final Noon).	Final Noon can only ever occur at the end of an Aramaic word. It goes below the lower guide line. Its vertical line is generally slightly sloped, unlike Waw and Zeyn
 (Yod) and (Waw).	Has the gap at the bottom left corner, but Tet has the gap at the top left. The line of Tet extends almost into the middle of the letter, but this does not happen with Meem
 (Meem) and (Tet).	Has the gap at the bottom left corner, but Tet has the gap at the top left. The line of Tet extends almost into the middle of the letter, but this does not happen with Meem
 (Final Meem) and (Semkat).	Has the gap at the bottom left corner, but Tet has the gap at the top left. The line of Tet extends almost into the middle of the letter, but this does not happen with Meem
 (Sadeh), (Uv) and (Final Sadeh).	Has the gap at the bottom left corner, but Tet has the gap at the top left. The line of Tet extends almost into the middle of the letter, but this does not happen with Meem

Fig. 13.7 Syriac words
(From http://en.wikipedia.org/wiki/Syriac_language)

ܐܡܝܐ	– šqal: "he has taken"
ܐܡܝܩܐ	– nešqōl: "he will take"
ܐܡܝܬܐ	– šāqel: "he takes, he is taking"
ܐܡܝܩܬܐ	– šaqqeḥ: "he has lifted/raised"
ܐܡܝܩܬܐ	– ašqeḥ: "he has set out"
ܐܡܝܩܬܐ	– šqālā: "a taking, burden, recension, portion or syllable"
ܐܡܝܩܬܐ	– šeqlē: "takings, profits, taxes"
ܐܡܝܩܬܐ	– šaqūṭā: "a beast of burden"
ܐܡܝܩܬܐ	– šūqālā: "arrogance"

Table 13.2 List of languages used in the framework of the writing system

Writing system	Languages
Arabic	Azeri (Iran), Balochi, Baluchi, Beja, Berber, Dari, Fulani, Hausa, Judeo-Spanish (until the twentieth century), Kabyle, Kanuri (on occasion), Konkani, Kashmiri, Kazakh in China, Kurdish (Iran and Iraq), Kyrgyz, Malagasy (until the nineteenth century), Malay (fourteenth to seventeenth century), Mandekan, Mazanderani, Morisco, Mozarabic (now extinct), Ottoman Turkish, Pashtu, Pashto, Persian/Farsi, Punjabi (Pakistan), Rajasthani, Saraiki, Shabaki, Sindhi, Spanish (formerly before sixteenth century, a.k.a. Aljamiado), Swahili (on occasion), Tatar, Tajik (on occasion), Tausug, Urdu, Uyghur
Hebrew	Aramaic, Bukhori, Hulaula, Judeo-Berber, Judeo-Iraqi Arabic, Judeo-Moroccan, Judeo-Tripolitanian Arabic, Judeo-Tunisian Arabic, Judeo-Portuguese, Judeo-Yemenite, Juhuri, Lishan Didan, Lishana Deni, Lishanid Noshan, Shuadit, Yiddish, Zarphatic
Syriac	Garshuni, Assyrian Neo-Aramaic, Bohtan Neo-Aramaic, Chaldean Neo-Aramaic, Hertevin, Koy Sanjaq Surat, Senaya, Syriac, Turoyo
Thaana	Dhivehi

1	𐤀	𐤁	𐤂	𐤃	𐤄	𐤅	𐤆	𐤇	𐤈	𐤉	𐤊	𐤋	𐤌	𐤍	𐤎	𐤏	𐤐	𐤑	𐤒	𐤓	𐤔	𐤕
2	𐤁𐤀	𐤁𐤁	𐤁𐤂	𐤁𐤃	𐤁𐤄	𐤁𐤅	𐤁𐤆	𐤁𐤇	𐤁𐤈	𐤁𐤉	𐤁𐤊	𐤁𐤋	𐤁𐤌	𐤁𐤍	𐤁𐤎	𐤁𐤏	𐤁𐤐	𐤁𐤑	𐤁𐤒	𐤁𐤓	𐤁𐤔	𐤁𐤕
3	ا	ب	ح	د	ه	و	ز	ح	ط	ي	ك	ل	م	ن	س	ع	ف	ص	ق	ر	ش	ت
4	ܐ	ܒ	ܓ	ܕ	ܗ	ܘ	ܙ	ܠ	ܡ	ܢ	ܣ	ܥ	ܦ	ܩ	ܪ	ܫ	ܬ	ܚ	ܛ	ܝ	ܟ	ܠ
5	'	b	g	d	h	w	z	h	t	y	k	l	m	n	s	'	p/f	s	q	r	š	t
1. Aramaic ; 2. Nabataean ; 3. Arabic ; 4. Syriac ; 5. Transcription																						

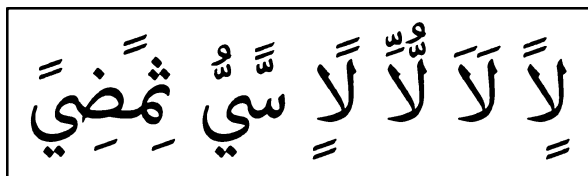
Fig. 13.8 History of the Arabic alphabet (From Wikipedia)

ا	ب	ج	د	ه	و	ز	ح	ط	ي	ك
ل	م	ن	س	ع	ف	ص	ق	ر	ش	ت
ث	خ	ذ	ض	ظ	غ					

Fig. 13.9 Arabic alphabet with the six added letters with yellow background

The Diacritics Addition: Since the Nabataean alphabet, which derives from the Arabic alphabet, was not originally fitted with diacritical marks, it remained to settle the problem of differentiation of letters which shared an identical shape. The Arabic script letters are the combination of “skeletal” letter form with dots positioned at various positions, and these dots are used as marks of differentiation. The dots are placed on top and below the contour of the letter, individually, or in groups of two or three (see Fig. 13.9, with the added six letters with yellow background). This is the inherited development of Arabic scripts. Regional languages are adopting new ways of writing to represent various sounds using combination of basic shapes plus dots (see Fig. 13.11).

Fig. 13.10 Arabic vocalization signs



Vocalization: Since Arabic script is a purely consonantal writing, ten signs were invented to clarify the pronunciation. There are three brief vowels and seven orthographic signs added above and below consonants (see Fig. 13.10).

Arabic Extension: Associated to the Islamization movement, the Arabic alphabet was spread to countries speaking languages which do not belong to the Semitic family. This led to the need for new letters, to represent a variety of sounds which were absent in standard Arabic scripts. As a consequence, new shapes were invented to represent new sounds, instead of borrowing from other scripts. Some languages have borrowed all the Arabic letters (28 consonants plus “hamza”) such as Persian, Ottoman Turkish, Urdu, and Pashto. Others have borrowed only a few, such as Kurdish Sorani, which borrowed only 21 letters. The new letters are formed by combining different dots with existing basic letters. It appears that this enhancement seems natural even instinctive, to many of those who use the Arabic script. To make some unknown phonemes in Arabic such as “p,” “v,” “g,” and “Tch,” new letters were invented by adding dots and diacritical marks to Arabic letters, by the closest pronunciation. They are summarized as follows:

- Persian and Urdu: “p” – b with three dots below
- Persian and Urdu: “ch” – j with three dots below
- Persian and Urdu: “g” – k with a double top
- Persian and Urdu: “zh” – z with three dots above
- In Egypt: “g” – j. That is because Egyptian Arabic has g where other Arabic dialects have j
- In Egypt: j – j with three dots below, same as Persian and Urdu “ch”
- In Egypt: “ch” – written as “t-sh”
- Urdu: retroflex sounds – as the corresponding dentals but with a small letter above (This problem in adapting a Semitic alphabet to write Indian languages also arose long before this: see Brahmi.)

For more clarity, Fig. 13.11 shows visual illustration of those aforementioned cases.

Vowels: The vowel notation is the main difficulty to which the introducers of the Arabic alphabet were faced. Despite a large number of consonants, the Arabic script uses only three signs to rate the three Arabic short vowels. These signs are added either above or below the consonants. Therefore, such a system, applied to languages where the vowels are numerous, gives unsatisfactory results. As an example, this may be the primary reason for the Turkish alphabet to move to Latin.

	Persian/ Farsi	Urdu	Malay	Sindhi	Kurdish	Uyghur	Arwi	Egyptian
p	پ		ف		پ		ف	
ch			چ				چ	
g				گ			گ	ج
zh	ژ						ژ	
j			ج					چ
ng			غ			ڭ		
ny			ن					
v			و		ف	و		
retroflex		ڙ						

Fig. 13.11 Adapting the Arabic alphabet for other languages: http://en.wikipedia.org/wiki/History_of_the_Arabic_alphabet

	Vowel description	I.	M.	F.	Is.
ا (a)	Slightly rounded long	ا	ا	ا	ا
و (o)	Oral rounded long	و	و	و	و
و (û)	Rounded with minimum aperture	وو	وو	وو	وو
و (ö)	Diphthong formed by the gradual opening of lips	-	وي	وي	ئوي
ه (e)	mid unrounded short	ه	ه	ه	ه
ي (i)	vowel aperture average	ي	-	-	ي
ي (ê)	close front long	ئي	ي	ي	ئي

Fig. 13.12 Kurdish Sorani vowels, where the initials are as follows: I. for initial form, M. for medium, F. for finale, and Is. for isolated

On the contrary, other languages have kept the Arabic alphabet while substituting the short vowels in Arabic letters. This is the case of Kurdish Sorani, which uses a group of two and three letters to record eight vowels (see Fig. 13.12).

Graphism Multiplication: In the Arabic writing, the letters are connected to each other. This practice leads to four different morphologies of the same letter, depending on its location in the word: initial, medial, final, or isolated, except for two letters which have only two forms. Figure 13.13 shows a few examples.

Ligatures: The Arabic script and its derivatives obey to the same ligation rules. They have three types of ligatures: contextual ligatures, language bindings, and

Fig. 13.13 Graphism multiplication, where the initials are as follows: I. for initial form, M. for medium, F. for finale, and Is. for isolated

Arabic letters	I.	M.	F.	Is.
ب	بـ	بـ	بـ	بـ
ص	صـ	صـ	صـ	صـ
ع	عـ	عـ	عـ	عـ
ز	ز	ز	ز/ز	ز

Contextual ligatures			
بعث → ب + ع + ث			
بعثت → ب + ع + ث + ت			
بعثنا → ب + ع + ث + ا			
Aesthetic ligatures			
At the beginning of words	ح + م + ل	→	حمل
	ش + م	→	شم
	ج + ن	→	جن
At the end of words	ر + ي	→	ري
	ي + ن	→	ني
	ي + ت	→	تي

Fig. 13.14 Examples of contextual and aesthetic ligatures

aesthetic ligatures. A contextual ligature is a string taking special shapes according to their position in the word, following strict grammatical rules and linked solely to writing. The language bindings are essential for writing a given language and obeying grammatical rules, which make them close to digraphs. The aesthetic ligatures are optional graphics that exist for aesthetic reasons, readability, and/or tradition. They can be replaced by their components without changing the grammatical validity or the meaning of the text. Figure 13.14 gives some examples of contextual and aesthetic ligatures.

Arabic Writing Styles: The development of Arabic calligraphy led to the creation of several decorative styles that were designed to accommodate special needs. The most outstanding of these styles are the following: Nastaliq, Koufi, Thuluthi, Diwani, Rouqi, and Naskh. An example of each one of these styles is given in Fig. 13.15.

Nastaliq	أَبْجَدُ هَوَزٍ حُطِّي كَلَمُنْ سَعْفَصْ قَرَشَتْ تَخَذْ ضَطْعُ
Koufi	أَبْجَدُ هَوَزٍ حُطِّي كَلَمُنْ سَعْفَصْ قَرَشَتْ تَخَذْ ضَطْعُ
Thuluthi	أَبْجَدُ هَوَزٍ حُطِّي كَلَمُنْ سَعْفَصْ قَرَشَتْ تَخَذْ ضَطْعُ
Diwani	أَبْجَدُ هَوَزٍ حُطِّي كَلَمُنْ سَعْفَصْ قَرَشَتْ تَخَذْ ضَطْعُ
Rouq'i	أَبْجَدُ هَوَزٍ حُطِّي كَلَمُنْ سَعْفَصْ قَرَشَتْ تَخَذْ ضَطْعُ
Naskh	أَبْجَدُ هَوَزٍ حُطِّي كَلَمُنْ سَعْفَصْ قَرَشَتْ تَخَذْ ضَطْعُ

Fig. 13.15 Examples of decorative styles

Usually, Naskh and Nastaliq are mostly followed by Arabic script-based languages. Nastaliq is mostly followed for Urdu, Punjabi, Sindhi, etc., whereas Naskh is mostly followed for Arabic, Persian, etc. (see Fig. 13.16). The difference between the two styles Naskh and Nastaliq is very significant for Urdu. Naskh style is closer to the traditional Arabic style except for some final letters. The aesthetic style is more pronounced in Nastaliq with very oblique ligatures. It also shows more pronounced variations of the letters according to their position in the word.

Ghost Character Theory: Considering the various Arabic scripts, an effort has been made to reassemble their shapes in order to facilitate their computer encoding. The National Language Authority, under Dr. Attash Durrani’s supervision, led in early 2000 to the proposition of a standard for Urdu keyboard based on the “ghost character theory.”

This theory learns that all Arabic script-based languages can be written with only 44 ghost characters. Ghost characters consist of 22 basic shapes called Kashti, shown in Fig. 13.17, and 22 dots (diacritical marks). However, each base ligature has its own phonemes and meanings in every language, with the same or different number of diacritical marks. Thus, the basic shapes (glyphs) are the same for all Arabic script-based languages with only difference in font, i.e., Naksh, Nastaliq, and diacritical marks followed by every language [32]. The “bey” contains 32 shapes shown in Fig. 13.18.

Kaf	Yod	Tet	Het	Zayin	Vav	He	Dalet	Gimel	Bet	Alef
כ	י	ט	ח	ז	ו	ה	ד	ג	ב	א
ך										
Tav	Shin	Resh	Qof	Tsadi	Pe	Ayin	Samekh	Nun	Mem	Lamed
ת	ש	ר	ק	צ	פ	ע	ס	נ	מ	ל
				ץ	ף			ן	ם	

Fig. 13.19 Hebrew alphabet (From Wikipedia)



Fig. 13.20 Typeface examples – from left to right, top to bottom: NARKISS, KOREN, AHARONI, CHAYIM, SCHOCKEN, and GILL (Reproduced (and rearranged) from [66])

Aramaic square script	אבגדהוזחטיכלמנסעפצקרשת
DSS Hebrew	אבגדהוזחטיכלמנסעפצקרשת
Paleo Hebrew	אבגדהוזחטיכלמנסעפצקרשת
Moabite Stone	אבגדהוזחטיכלמנסעפצקרשת
Ancient Hebrew	אבגדהוזחטיכלמנסעפצקרשת
Early Semitic	אבגדהוזחטיכלמנסעפצקרשת

Fig. 13.21 Different writing styles for Hebrew

Unlike Arabic, the Hebrew letters do not take different shapes depending on the surrounding letters (see Fig. 13.19). Hebrew does not have capital letters. Several typefaces are however used for printed text (see Fig. 13.20). Figure 13.21 shows different writing styles for Hebrew.

Similarities to Arabic: As in Arabic, diacritics accompany the characters, either above, below, or inside, in order to represent vowel sounds (see Fig. 13.22).

Differences with Arabic: Hebrew is not a cursive script, which means that the letters are not connected (see Fig. 13.23).

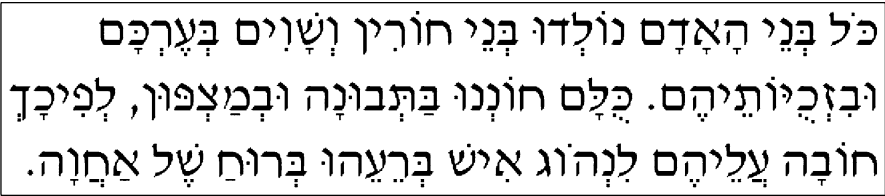


Fig. 13.22 Diacritics in Hebrew (From: <http://www.omniglot.com/writing/hebrew.htm>)



Fig. 13.23 Hebrew cursive text example (From <http://www.omniglot.com/writing/hebrew.htm>)

Writing Recognition Systems

The following lists the main steps that are needed for designing recognition systems. Within this framework, new methods that have been proposed for the aforementioned scripts as well as existing traditional methods will be explained, with a focus on how they have been adapted.

Preprocessing

This first phase for getting the word shape consists of several steps that are described in the sequel.

Dots and Diacritical Marks Removal: It is customary to remove strokes like diacritical dots or marks in the preprocessing phase, so that overall appearance of the writing does not change. They then introduce them again in the post-processing phase. This is however not a robust strategy/technique, since they cannot be equally detected because of their structural properties. Furthermore, in Urdu and Farsi, some retroflex marks correspond to small characters which need to be recognized.

Sari et al. [59] split the writing area into three bands and use upper and lower bands for locating and removing the diacritical marks. This idea is echoed by Miled et al. [50] who define a set of symbols. Zahour et al. [68] divide the image into vertical bands of uniform width, in which they proceed to classify all the connected

components. The method reuses the size where small blocks are generally classified as diacritics. Razzak et al. [56] remove the diacritical marks based on their size and position, to allow for efficient feature extraction. After that, the combination of recognized based shapes plus diacritical marks results in an efficient multi-language character recognition system. Finally, in Menasri et al. [48], several steps are followed for the diacritic mark extraction. First, a filtering step is used to exclude the “alif” and the single letters. This initial filtering is used to detect the base band which is exploited to locate the diacritics above or below. Second, another filtering of diacritics is made in these bands, referring to the position and size.

Baseline Extraction: The baseline is virtual and a foundation line on which characters are combined to form the ligatures. It provides its orientation and can be used as a guide for segmentation. A survey provided in [12] summarizes the different comprehensive approaches using the horizontal projection and detection of peaks, skeleton analysis by using linear regression as in [54] for Arabic and in [55] for Urdu, word contour representation by local minima and linear regression as in [37], or those based on the principle of components analysis [22].

In Urdu, Farsi, and some Arabic styles, specially Nastaliq writing style, the ascenders and descenders cause incorrect detection of the baseline because of their oblique orientation and long tail (see Fig. 13.24). This is why Safabakhsh and Adibi [58] proposed to eliminate certain ascenders, such as those of “Kaf” and “Gaf” (see Fig. 13.24b), and descenders of letters like “Jim,” “Che,” “He,” “Khe,” “Ayn,” and “Ghayn” (see Fig. 13.24a) to help the segmentation and the baseline extraction. For the ascenders, some features such as the overlapping with other strokes and a large change in the stroke width near its head are considered. Figure 13.25 shows correct elimination (a) and incorrect elimination (b). For the descenders, the elimination is based on the multiplication of black runs on the same line, allowing to eliminate the curve part of final characters (see Fig. 13.26). Razzak et al. [57] estimated the baseline locally on base stroke with additional knowledge of previous words for the handwritten Arabic script. As the handwritten Arabic script is more complex and may not exactly on the same baseline, the primary task was to extract global baseline that eventually helps for local baseline extraction. For more understanding, the reader is referred to Figs. 13.24–13.26 where different approaches are categorized according to the techniques used.

Word Segmentation

The purpose of segmentation is to subdivide the shape such that each pair of connected characters will be split into two. In the current state of the art, there are several different approaches. It goes without saying that for Arabic handwriting, methods based on the vertical histogram are not appropriate. Similarly, methods using the vertical thickness of the stroke are not suitable for free script. Three main

Fig. 13.24 Problems arise from descenders and ascenders: (a) “Ayn” will be considered before “Ye”; (b) “Kaf” will be considered before “Be” [58]

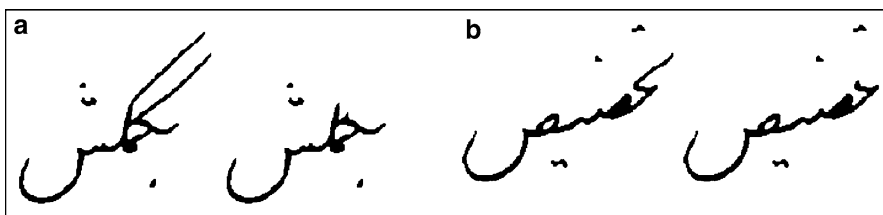
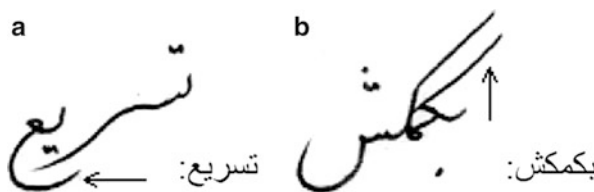


Fig. 13.25 Operation of ascender elimination algorithm: (a) correct operation; (b) incorrect operation [58]

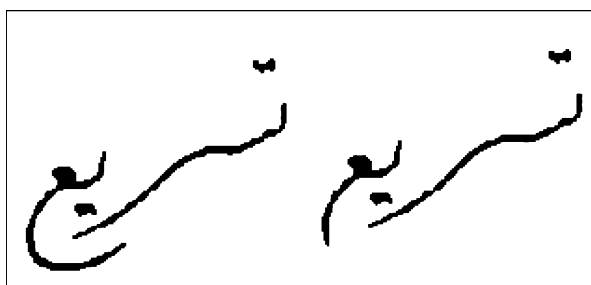
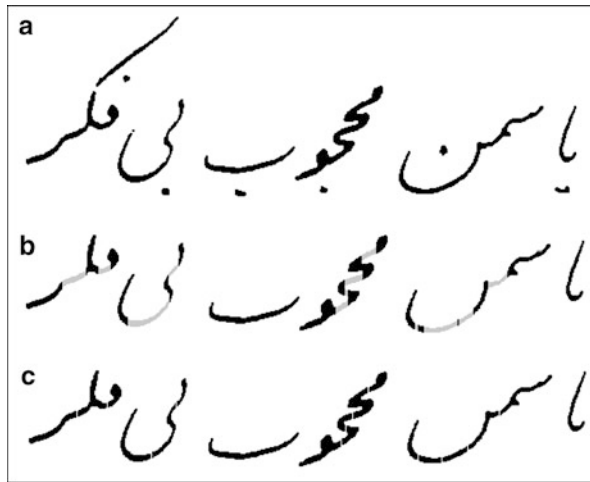


Fig. 13.26 Operation of descender elimination algorithm [58]

methods are suitable for the Arabic-based languages. They are based on regularities and singularities, local minima of the upper contour, and the right-to-left order.

Segmentation Based on Regularities and Singularities. This is proposed for Arabic handwritten words in [51]. Safabakhsh and Adibi [58] used a similar approach for Urdu. Singularities or islands correspond to the vertical parts of the image and their connections, while regularities correspond to horizontal ligatures, loops, and small curves along the baseline, obtained by a subtraction between singularities and the image. Similarities are obtained by an opening operation with vertical element whose height is larger than the pen width. The connections are reached by a closing operation with a horizontal structural element having a small

Fig. 13.27 (a) The words.
 (b) Singularities and regularities specified by *black* and *gray* colors, respectively.
 (c) Resulting segmentation [58]



width. Small loops are filled in a preprocessing phase to avoid their segmentation. Figure 13.27 shows an example of this type of segmentation.

Segmentation Using Local Minima of the Word Upper Contour. This is proposed first by Olivier et al. [52] for Arabic words. The local minima of the upper contour are first selected. Then, if these points are in favorable situation, like being at the frontiers of loops, they are considered as segmentation points. This method has been modified by Safabakhsh and Adibi [58] to extend it to Nastaliq handwritten words. In Nastaliq style, when character “Re” is connected to another character before it, it is written without any upper contour minima between it and the previous character, which cannot work here. To face this problem, the algorithm starts by detecting the “Re” and the overlapping zones. Then, similar procedures are followed to detect local minima and segmentation points. Figures 13.28 and Fig. 13.29 give examples of such a segmentation.

Segmentation Finding Right-to-Left Order. This is proposed first by Chen et al. [24] and later simplified by Safabakhsh and Adibi [58]. The idea is that once ascender and descender are removed and the order of sub-words found, the order of segments is obtained by considering the right-most segment as the first one in the sub-word. Then, one traverses the outer contour and considers the order in which segments are visited.

Isolated Character Recognition

Some researches closely examine character recognition. Alansour and Alzoubady [5] proposed a neocognitron to classify handwritten characters. The input of this

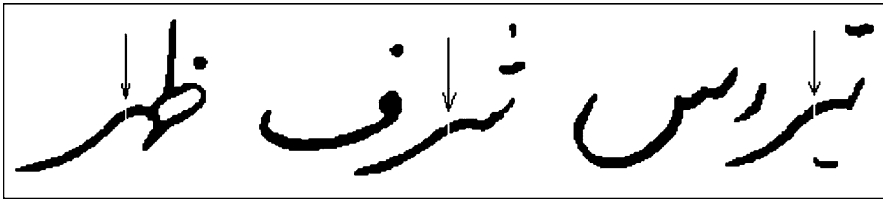


Fig. 13.28 “Re” detection algorithm for three words [58]

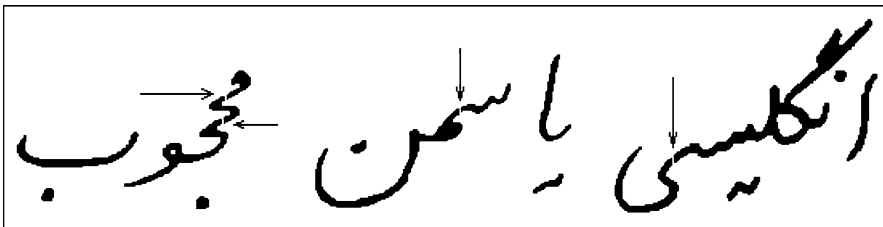


Fig. 13.29 Overlap detection algorithm for three words [58]

particular neural network (NN) is composed of structural features. The system assumes a context of a handwritten Arabic document recognition software that should be able to segment words directly into letters. Asiri and Khorsheed [14] proposed to use two different NN architectures for handwritten Arabic character recognition. For all NNs, the inputs correspond to a certain number of Haar wavelet transform coefficients. Cowell and Hussain [28] worked on isolated Arabic printed characters. A character image is normalized and a quick matching method is used. In [46], an automatic identification of hand-printed Hebrew characters is described. The squared shape of the letters led to the use of primitive-type straight via the Hough transform. These primitives are augmented by indices such as corners and end points. Then, a NN is used for recognition. Fakir et al. [35] proposed to use the Hough transform for the recognition of printed Arabic characters.

Word Recognition

While doing Arabic script recognition research, most of the proposed recognition systems can be extended to Latin, Chinese, and Japanese without considering the particular difficulties associated with different scripts. This means that the recognition system provides genericity, i.e., application independency. Today, advances produced for Arabic script, for example, will serve the derivatives languages with which they share many similarities. Thus, the usual recognition methods for recognizing and assigning them the most characteristic works of the literature will be shown concerning Arabic script-based languages recognition. Belaïd and

Choisy [17] gave a comprehensive review of the Arabic recognition approaches. The approaches are classified into four, according to the human perception of writing. They are:

- Global-based vision classifiers
- Semi-global-based vision classifiers
- Local-based vision classifiers
- Hybrid-level classifiers

Global-Based Vision Classifiers: In this holistic approach, the word is regarded as a whole, allowing correlations with the totality of the pattern. This common approach avoids the heavy task of letter location and word segmentation. The key idea involves detecting a set of shape primitives in the analyzed word and arranging them properly in the word space [6, 36, 44]. Some systems failed to evaluate the gap between two consecutive word parts (PAW or part of Arabic word) to decide the word limits. In [64], the presence of the “Alef” was reported as the first letter of many Arabic words as the most relevant feature.

Feature selection in this word vision is essentially of global nature. They have been addressed by researchers for both recognition and writer identification. Among these features can be mentioned multichannel Gabor filtering [13]; 2D DWT combined with morphological variations of writing [38]; combination of gradient, curvature, density, horizontal and vertical run lengths, stroke, and concavity features [15]; hybrid spectral-statistical measures (SSMs) of texture [7]; curvature measurements from minimum perimeter polygon [3]; fractal dimensions by using the “box-counting” method [23]; edge-based directional probability distributions and moment invariants [11]; white and black run length; and edge-hinge features [30].

Semi-global Based Vision Classifiers: More specifically, the nature of Arabic writing allows us to describe the language in fewer natural level, i.e., the PAW level. Indeed, Arabic words are built by a concatenation of several independent written parts that give another natural segmentation level. This natural segmentation allows us to refine the analysis by reducing the number of base vocabulary. It explains some approaches that are based their work on this level. By reducing the base vocabulary, it allows the possibility of extending the dictionary [18, 20].

The global features shown previously can be reported in a similar manner here for the PAW images.

Local-Based Vision Classifiers: The objective of this word vision is to focus on the interpretation of letters and or smaller entities for their interpretation. In other words, the process is to gather, bind and confront these entities to identify the word. Such an analysis level leads to the Sayre dilemma. This problem is usually eluded by the use of implicit or explicit segmentation methods.

Very recently, El Abed and Margner have organized a competition at the ICDAR conferences [33]. This competition was made mainly under the purview of feature selection, where the major extraction methods were evaluated. They are sliding

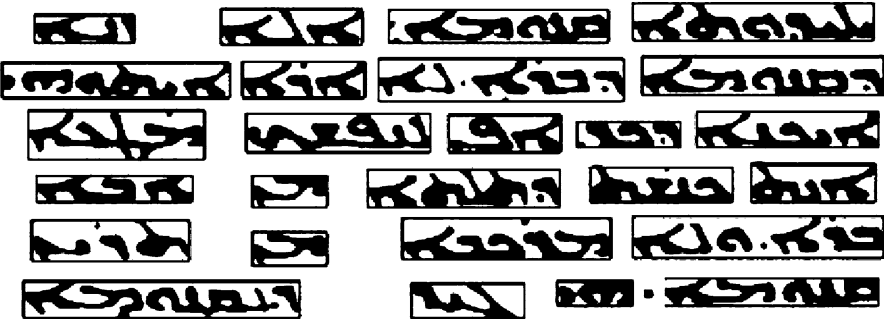


Fig. 13.30 Syriac writing: interword gaps are larger than intra-word gaps [27]

window with pixel features, skeleton direction-based features, and sliding window with local features.

In addition to these conventional features, the literature abounds with primitive extraction techniques such as analytical approaches with or without consideration of parts of words and with or without segmentation into letters. Examples are the work of Abandah et al. [2] who used mRMR to select four sets of features for four classifiers, each specialized in recognizing letters of the four letter forms. Fahmy and Al Ali proposed in [4, 34] a system with structural features. Razzak et al. [56] presented a hybrid approach using HMM and fuzzy logic for Arabic-script based languages written in both Nastaliq and Naskh styles. Fuzzy logic is used for feature extraction, clustering, and post-processing. Clocksin and Fernando proposed an analytic system for Syriac manuscripts [27]. In contrast to Arabic writing, the word segmentation simply happens as the intra-word gaps seem to be clearly smaller than interword gaps. The grammatical grammar functions almost appear as word prefixes or suffixes instead of separated words, so that a huge dictionary is inevitable to provide global word approach (see Fig. 13.30).

Miled et al. [50] proposed an analytical approach based on HMMs for the recognition of Tunisian town names. They integrate the notion of PAW in their system. They grouped letters with the same body but different diacritics to “solve” the problem of diacritic detection and classification.

Hybrid-Level Classifiers: By combining different strategies, it is possible to come near to the principle of reading – the analysis must be global for a good synthesis of information, while being based on local information that are suitable to make this information emerged [26, 62]. Such an approach combination allows us to better see how a person reads, which is to first analyze global word shapes, and search for local information only to discriminate ambiguous cases. As local-based approaches gather local information up to words, they could be hybrid ones. An important difference that exists between the two approaches is that hybrid approaches attempt a multilevel analysis of the writing, while local-based approaches gather local information.

Hybrid approaches aim at providing different levels of features and interpretation [63]. Accordingly, Maddouri et al. proposed [62] a combination of global and local models based on a transparent NN (TNN).

Contribution to Natural Language Processing: Usually, NLP is kept as the last step of the recognition system, to be able to correct the errors by providing some considerations stemming from the writing language [25]. Arabic is characterized by a complex morphological structure. It is a highly inflected language where prefixes and suffixes are appended to the stem of a word to indicate tense, case, gender, number, etc. For that reason, it seems that words are not the best lexical units in this case and, perhaps, sub-word units would be a better choice. In the literature, few researches are done in that direction. As mentioned in Cheriet [25], the question of incorporating NLP in a recognition system is a non-resolved problem. Till now, it is not definitely clear where the incorporation of NLP is more profitable for a writing recognition system. Ben Cheikh's research [19] incorporates the morphology analysis within models which are conceived to collaborate and respectively learn and recognize roots, schemes, and conjugation elements of words.

Table 13.3 categorizes other systems according to the approach, the classifier, the dataset, and the writing style.

Success Clues of the Topic

The successful achievement in any research topic is measured by the number of commercial systems, of distributed databases, and by the number of competitions held in international conferences. This section outlines some of these tell-tale signs of success of research in the Middle Eastern character recognition.

Datasets

Datasets of any printed or handwritten language are the most important part in order to measure accuracy and it gives equally for recognition system design. Table 13.4 lists the existing publicly available datasets that have been used over several years.

Academic Systems

Since 2005, a series of competitions took place at the important conferences dedicated to document analysis like ICDAR and ICFHR. During these competitions, datasets are fixed and systems are experimented on them. Table 13.5 reports the results of four competitions of Arabic handwritten word recognition systems, from 2005 to 2011. The dataset is limited to IFN/INIT with a selection of this dataset (sets d, e, f_a , f, s). All participating systems were based mainly on hidden Markov models (HMM).

Table 13.3 Comparison between techniques

System	Approach	Classifier	Dataset	Writing style	R. rate
Biologically inspired fuzzy-based expert system	Simultaneous segmentation and recognition	Fuzzy logic	Full dataset	Nastaliq, Naskh	87.3 %
Hybrid approach HMM and fuzzy logic [56]	Ligature-based approach	HMM and fuzzy logic	14,150 words	Nastaliq, Naskh	87.6 %
Arabic character recognition		Genetic algorithm and visual encoding	200 words	Naskh	97 %
Online Urdu character recognition [43]	Ligature-based approach	Back propagation neural network	240 ligatures	Nastaliq	93 %
Online Arabic handwritten recognition with templates [65]		Template-based matching	Full dataset	Naskh	68.2 %
Bio-inspired handwritten Farsi [21]		KNN ANN SVM	Farsi digits (MNIST)	Naskh	81.3 % (KNN) 94.65 % (ANN) 98.75 % (SVM)
Arabic handwritten using matching algorithm [53]	Ligature-based approach	Matching algorithm and decision tree	Arabic alphabets	Naskh	98.8 %
Recognition based segmentation of Arabic [29]	Simultaneous segmentation and recognition	HMM	Arabic	Naskh	88.8 %
Online Arabic characters by Gibbs modelling [49]		Gibbs modelling of class conditional densities	Arabic characters	Naskh	84.85 % (direct Bayes) 90.19 % (indirect Bayes)
Arabic character recognition using HMM [9]		HMM	Arabic full data set	Naskh	78.25 %

Table 13.4 Datasets for Eastern scripts

Name	Script	Component	Nature	Content
IFN/ENIT	Arabic	Handwritten words	Towns	411 writers, 26,400 samples
APTI	Arabic	Multi-font, multi-size, and multi-style text	Printed text	45,313,600 single-word images totaling to more than 250 million characters
AHDB	Arabic	Arabic words and texts	Form words	100 writers
ERIM	Arabic	Machine-printed documents	Arabic books and magazines	750 pages, 1,000,000 characters
CEDARABIC	Arabic	Handwritten documents	Document pages	10 writers, 20,000 words
ADAB	Arabic	Online	984 Tunisian town names	15,158 pen traces, 130 different writers
HODA	Farsi	Handwritten digits	Form digits	102,352 digits
CENPARMI	Farsi	Numerical strings, digits, letters, legal amounts and dates	Binary forms of handwritten digits and characters	11,000 training and 5,000 test samples while characters set includes 7,140 training and 3,400 test samples

Commercial Systems

There are a lot of commercial applications for Middle Eastern scripts, like ABBYY Finereader, Readiris, HOCR (only for Hebrew), Sakhr, Omnipage, and VERUS, whereas for online Myscript, Sakhr is publicly available. Sakhr provides 99.8 % accuracy for high-quality documents and 96 % accuracy for low-quality documents. It has supports for Arabic, Farsi, Pashto, Jawi, and Urdu, and it also supports bilingual documents: Arabic/French, Arabic/English, and Farsi/English. Moreover, it can handle image and text recognition captured by mobile devices. It can also auto-detect translation language.

US government evaluators assess Sakhr as the best available Arabic OCR. Sakhr online intelligent character recognition (ICR) recognizes Arabic cursive handwritten input through a normal pen with 85 % word accuracy.

Two groups with three systems, IPSAR, and UPV-BHMM (UPV-PRHLT-REC1 and UPV-PRHLT-REC2), were presented at ICDAR11 and tested on the APTI dataset. IPSAR is based on the HMM and follows stages: extracting a set of features from the input images, clustering the feature set according to a predefined codebook, and, finally, recognizing the characters. Very interestingly, IPSAR does not require segmentation. UPV-BHMM presented two systems UPV-PRHLT-REC1 and UPV-PRHLT-REC2. UPV-BHMM systems are based on window sliding of adequate width to capture image context at each horizontal position. The IPSAR

Table 13.5 Results of the three best systems at ICDAR 2007

Name	Classifier	Features	Set d	Set e	Set f_a	Set f	Set s
Results of the four best systems at ICDAR 2011							
JU-OCR [1]	Random forest	Explicit grapheme segmentation	75.49	63.75	64.96	63.86	49.75
CENPARMI-OCR [61]	SVM	Gradient features, Gabor features, Fourier features	99.90	99.91	40.00	40.00	35.52
RWTH-OCR [31]	HMM	Appearance-based image slice features	99.67	98.61	92.35	92.20	84.55
REGIM [42]	PSO-HMM	Karhunen-Loeve transform [33]	94.12	86.62	80.60	79.03	68.44
Results of the three best systems at ICFHR 2010							
UPV PRHLT [39]	HMM	Sliding window	99.38	98.03	93.46	92.20	84.62
CUBS-AMA	HMM		89.97	80.80	81.75	80.32	67.90
RWTH-OCR [31]	HMM	Simple appearance-based image slice features	99.66	98.84	92.35	90.94	80.29
Results of the three best systems at ICDAR 2009							
MDLSTM [41]	NN	Raw input	99.94	99.44	94.68	93.37	81.06
Ai2A [48]	GMHMM	Geometric features with sliding windows	97.02	91.68	90.66	89.42	76.66
RWTH-OCR [31]	HMM	Simple appearance-based image slice features	99.79	98.29	87.17	85.69	72.54
Results of the three best systems at ICDAR 2007							
Siemens [60]	HMM	Feature vector sequence	94.58	87.77	88.41	87.22	73.94
MIE	LDC [45]	Word length estimation	93.63	86.67	84.38	83.34	68.40
UOB-ENST	HMM [10]	Features with respect of the baseline	92.38	83.92	83.39	81.93	69.93

system provided good results for the “Traditional Arabic” and “Diwani Letter” fonts in font sizes 10, 12, and 24. However, the system UPV-PRHLTREC1 was the winner of this first competition.

During this ICDAR competition, three groups, VisionObjects, AUC-HMM, and FCI-CU-HMM, have submitted their systems. These systems were tested over the ADAB dataset. There were two levels of evaluation. The first level of evaluation

Table 13.6 Examples of commercial OCRs

System	Script	Comments	Performance
Sakhr OCR	Arabic, Farsi, Jawi, Dari, Pashto, Urdu	Arabic, Farsi, English, French	96–99 %
VERUS OCR NovoDynamics	Arabic, Farsi, Persian, Dari, Pashto		
Readiris	Arabic, Farsi, and Hebrew	<ul style="list-style-type: none">– Pro features: standard scanning support and standard recognition features– Corporate features: volume scanning support and advanced recognition features	
Kirtas KABIS employs SAKHR engine for Arabic	Arabic (Naskh, Kofi), Farsi, Jawi, Pashto, and Urdu	<ul style="list-style-type: none">– SureTurn robotic arm uses vacuum system to gently pick up and turn one page at a time	
HOCR	Hebrew	Support for all Hebrew, English, and Western European languages	
ABBYY FineReader	Arabic, Hebrew	<ul style="list-style-type: none">– Dictionary for some languages – free trial is available	99 %
Ligature-OCR	Hebrew	Omnifont reading based on stochastic algorithms and neural networks	11,000 old Hebrew books
Freeocr	Hebrew	Books and reports, selected zones	
OCR program	Yiddish	Omnifont	Line by line

was based on the subsets 1–4, and systems shows recognition rate better than 80 % on sets 1–4, whereas the second-level evaluation was performed on set 5 and set 6. The recognition rate was limited between 60.28 and 98.97 %. The system of Vision Objects was the winner of this competition (Table 13.6).

Conclusion

Middle Eastern languages have very typical scripts, especially Syriac and Arabic. Therefore, special attention must be given to the design of the complete recognition system. Based on the thorough analysis of the different techniques, the following observations can be made:

- **Feature representation:** Low-level features are language independent. Once extracted (similarly for all the scripts), the training process can arrange their

proximity to the language studied. In contrast, high-level features are language dependent and need to develop specific extraction methods to retrieve all information. Obviously, a combination of these two kinds of features should perform better where each feature level is used to complement the drawback of the other.

- **PAWs:** Contrarily to Latin script, the basic entity is not the word. Global approaches should be based on PAW. Analytical ones gain by integrating this information level. A first effect reduces the vocabulary complexity by gathering the information on an intermediate level.
- **Segmentation, in words or in letters?:** Since Arabic writing is often described as more complex than Latin, it seems obvious that a letter segmentation cannot be effective.
- **Appropriateness of approaches:** The most suitable for these scripts. Hybrid ones seem very promising. They efficiently combine different perceptive levels, allowing discrimination of words without a complete description. In comparison with global approaches, the addition of local information allows it to extend the vocabulary with less confusion. Compared to local approaches, hybrid ones can avoid the full-segmentation problems and are less disturbed by information loss.

Cross-References

- ▶ [A Brief History of Documents and Writing Systems](#)
- ▶ [Continuous Handwritten Script Recognition](#)
- ▶ [Datasets and Annotations for Document Analysis and Recognition](#)
- ▶ [Handprinted Character and Word Recognition](#)
- ▶ [Machine-Printed Character Recognition](#)
- ▶ [Text Segmentation for Document Recognition](#)

References

1. Abandah G, Jamour F (2010) Recognizing handwritten Arabic script through efficient skeleton-based grapheme segmentation algorithm. In: Proceedings of the 10th international conference on intelligent systems design and applications (ISDA), Cairo, pp 977–982
2. Abandah G, Younis K, Khedher M (2008) Handwritten Arabic character recognition using multiple classifiers based on letter form. In: Proceedings of the 5th IASTED international conference on signal processing, pattern recognition, and applications (SPPRA), Innsbruck, pp 128–133
3. Abdi MN, Khemakhem M, Ben-Abdallah H (2010) Off-line text-independent Arabic writer identification using contour-based features. *Int J Signal Image Process* 1:4–11
4. Abuhaiba ISI, Holt MJJ, Datta S (1998) Recognition of off-line cursive handwriting. *Comput Vis Image Underst* 71:19–38
5. Alansour AJ, Alzoubady LM (2006) Arabic handwritten character recognized by neocognitron artificial neural network. *Univ Sharjah J Pure Appl Sci* 3(2):1–17
6. Al-Badr B, Haralick RM (1998) A segmentation-free approach to text recognition with application to Arabic text. *Int J Doc Anal Recognit (IJ DAR)* 1:147–166

7. Al-Dmour A, Abu Zitar R (2007) Arabic writer identification based on hybrid spectral-statistical measures. *J Exp Theor Artif Intell* 19:307–332
8. Al Hamad HA, Abu Zitar R (2010) Development of an efficient neural-based segmentation technique for Arabic handwriting recognition. *Pattern Recognition* 43(8):2773–2798
9. Al-Habian G, Assaleh K (2007) Online Arabic handwriting recognition using continuous Gaussian mixture HMMs. In: *International conference on intelligent and advanced systems (ICIAS)*, Kuala Lumpur, pp 1183–1186
10. Al-Hajj R, Likforman-Sulem L, Mokbel C (2005) Arabic handwriting recognition using baseline dependant features and hidden Markov modeling. In: *8th international conference on document analysis and recognition (ICDAR)*, Seoul, vol 2, pp 893–897
11. Al-Ma-adeed S, Mohammed E, Al Kassis D, Al-Muslih F (2008) Writer identification using edge-based directional probability distribution features for Arabic words. In: *IEEE/ACS international conference on computer systems and applications*, Doha, pp 582–590
12. Al-Shatnawi AM, Omar K (2008) Methods of Arabic language baseline detection, the state of art. *ARISER* 4(4):185–193
13. Al-Zoubeidy LM, Al-Najar HF (2005) Arabic writer identification for handwriting images. In: *International Arab conference on information technology*, Amman, pp 111–117
14. Asiri A, Khorsheed MS (2005) Automatic processing of handwritten Arabic forms using neural networks. *Trans Eng Comput Technol* 7:147–151
15. Awaida SM, Mahmoud SA (2010) Writer identification of Arabic handwritten digits. In: *IWFHR10*, Istanbul, pp 1–6
16. Bar-Yosef I, Beckman I, Kedem K, Dinstein I (2007) Binarization, character extraction, and writer identification of historical Hebrew calligraphy documents. *Int J Doc Anal Recognit* 9(2–4):89–99
17. Belaïd A, Choisy Ch (2006) Human reading based strategies for off-line Arabic word recognition. In: *Summit on Arabic and Chinese handwriting (SACH'06)*, University of Maryland, College Park, 27–28 Sept 2006
18. Ben Amara N, Belaïd A (1996) Printed PAW recognition based on planar hidden Markov models. In: *Proceedings of the 13th international conference on pattern recognition*, Vienna, 25–29 Aug 1996, vol 2, pp 220–224
19. Ben Cheikh I, Kacem A, Belaïd A (2010) A neural-linguistic approach for the recognition of a wide Arabic word lexicon. In: *Document recognition and retrieval XVII*, San Jose, pp 1–10
20. Bippus R (1997) 1-dimensional and pseudo 2-dimensional HMMs for the recognition of German literal amounts. In: *ICDAR'97*, Ulm, Aug 1997, vol 2, pp 487–490
21. Borji A, Hamidi M, Mahmoudi F (2008) Robust handwritten character recognition with feature inspired by visual ventral stream. *Neural Process Lett* 28:97–111
22. Burrow P (2004) Arabic handwriting recognition. M.Sc. thesis, University of Edinburgh, Edinburgh
23. Chaabouni A, Boubaker H, Kherallah M, Alimi AM, El Abed H (2010) Fractal and multi-fractal for Arabic offline writer identification. In: *ICPR-10*, Istanbul, pp 1051–1051
24. Chen MY, Kundu A, Srihari SN (1995) Variable duration hidden Markov model and morphological segmentation for handwritten word recognition. *IEEE Trans Image Process* 4(12):1675–1688
25. Cheriet M (2006) Visual recognition of Arabic handwriting: challenges and new directions. In: *SACH 06*, College Park, Sept 2006, pp 129–136
26. Choisy Ch, Belaïd A (2003) Coupling of a local vision by Markov field and a global vision by neural network for the recognition of handwritten words. In: *ICDAR'03*, Edinburgh, 3–6 Aug 2003, pp 849–953
27. Clocksin WF, Fernando PPJ (2003) Towards automatic transcription of Syriac handwriting. In: *Proceedings of the international conference on image analysis and processing*, Mantova, pp 664–669
28. Cowell J, Hussain F (2002) A fast recognition system for isolated Arabic character recognition. In: *IEEE information visualization IV2002 conference*, London, July 2002, pp 650–654

29. Daifallah K, Zarka N, Jamous H (2009) Recognition-based segmentation algorithm for on-line Arabic handwriting. In: 10th international conference on document analysis and recognition, Barcelona, pp 886–890
30. Djeddi C, Souici-Meslati L, Ennaji A (2012) Writer recognition on Arabic handwritten documents. International Conference on Image and Signal Processing, Agadir, Morocco, June 2012, pp 493–501
31. Dreuw P, Heigold G, Ney H (2009) Confidence-based discriminative training for model adaptation in offline Arabic handwriting recognition. In: Proceedings of the international conference on document analysis and recognition (ICDAR), Barcelona, pp 596–600
32. Durani A (2009) Pakistani: lingual aspect of national integration of Pakistan. www.nlaudit.gov.pk
33. El Abed H, Margner V (2007) Comparison of different preprocessing and feature extraction methods for off line recognition of handwritten Arabic words. In: Proceedings of the 9th ICDAR 2007, Curitiba, pp 974–978
34. Fahmy MMM, Al Ali S (2001) Automatic recognition of handwritten Arabic characters using their geometrical features. *Stud Inform Control J* 10:81–98
35. Fakir M, Hassani MM, Sodeyama C (2000) On the recognition of Arabic characters using Hough transform technique. *Malays J Comput Sci* 13(2):39–47
36. Farah N, Khadir MT, Sellami M (2005) Artificial neural network fusion: application to Arabic words recognition. In: Proceedings of the European symposium on artificial neural networks (ESANN'2005), Bruges, 27–29 Apr 2005, pp 151–156
37. Farooq F, Govindaraju V, Perrone M (2005) Pre-processing methods for handwritten Arabic documents. In: Proceedings of the 2005 eighth international conference on document analysis and recognition (ICDAR), Seoul, pp 267–271
38. Gazzah S, Ben Amara NE (2007) Arabic handwriting texture analysis for writer identification using the DWT-lifting scheme. In: 9th ICDAR, Curitiba, vol.2, pp 1133–1137
39. Gimenez A, Khoury I, Juan A (2010) Windowed bernoulli mixture hmms for arabic handwritten word recognition. In: 2010 international conference on frontiers in handwriting recognition (ICFHR), Kolkata, pp 533–538
40. Ghosh D, Dube T, Shivaprasad AP (2010) Script recognition-a review. In: *IEEE Trans Pattern Anal Mach Intell* 32(12):2142–2161
41. Graves A, Schmidhuber J (2009) Offline handwriting recognition with multidimensional recurrent neural networks. *Advances in neural information processing systems* 22, NIPS'22, Vancouver, MIT Press, pp 545–552
42. Hamdani M, El Abed H, Kherallah M, Alimi AM (2009) Combining multiple HMMs using on-line and off-line features for off-line Arabic handwriting recognition. In: Proceedings of the 10th international conference on document analysis and recognition (ICDAR), Seoul, pp 201–205
43. Husain SA, Sajjad A, Anwar F (2007) Online Urdu character recognition system. In: IAPR conference on machine vision applications (MVA2007), Tokyo, pp 3–18
44. Khorsheed MS, Clocksin WF (2000) Multi-font Arabic word recognition using spectral features. In: Proceedings of the 15th international conference on pattern recognition, Barcelona, 3–7 Sept 2000, vol 4, pp 543–546
45. Kimura F, Shridhar M, Chen Z (1993) Improvements of a lexicon directed algorithm for recognition of unconstrained handwritten words. In: 2nd international conference on document analysis and recognition (ICDAR), Tsukuba, pp 18–22
46. Kushnira M, Abe K, Matsumoto K (2003) Recognition of handprinted Hebrew characters using features selected in the Hough transform space. *Pattern Recognit* 18:103–114
47. Lorigo LM, Govindaraju V (2006) Offline Arabic handwriting recognition: a survey. *PAMI* 28(5):712–724
48. Menasri F, Vincent N, Cheriet M, Augustin E (2007) Shape-based alphabet for off-line Arabic handwriting recognition. In: Proceedings of the ninth international conference on document analysis and recognition (ICDAR), Curitiba, vol 2, pp 969–973

49. Mezghani N, Mitiche A, Cheriet M (2008) Bayes classification of online Arabic characters by Gibbs modeling of class conditional densities. *IEEE Trans Pattern Anal Mach Intell* 30(7):1121–1131
50. Miled H, Olivier C, Cheriet M, Lecoutie Y (1997) Coupling observation/letter for a Markovian modelisation applied to the recognition of Arabic handwriting. In: *Proceedings of the 4th international conference on document analysis and recognition (ICDAR)*, Ulm, pp 580–583
51. Motawa D, Amin A, Sabourin R (1997) Segmentation of Arabic cursive script. In: *Proceedings of the 4th international conference on document analysis and recognition (ICDAR)*, Ulm, vol 2, pp 625–628
52. Olivier C, Miled H, Romeo K, Lecourtier Y (1996) Segmentation and coding of Arabic handwritten words. In: *IEEE proceedings of the 13th international conference on pattern recognition*, Vienna, vol 3, pp 264–268
53. Omer MAH, Ma SL (2010) Online Arabic handwriting character recognition using matching algorithm. In: *The 2nd international conference on computer and automation engineering (ICCAE)*, Beijing
54. Pechwitz M, Märgner V (2002) Baseline estimation for Arabic handwritten words. In: *Proceedings of the eighth international workshop on frontiers in handwriting recognition (IWFHR)*, Niagara-on-the-Lake, p 479
55. Razzak MI, Husain SA, Sher M, Khan ZS (2009) Combining offline and online preprocessing for online Urdu character recognition. In: *Proceedings of the international multiconference of engineers and computer scientists (IMECS) 2009*, Hong Kong, vol I. Knowledge-based systems
56. Razzak MI, Anwar F, Husain SA, Belaïd A, Sher M (2010) HMM and fuzzy logic: a hybrid approach for online Urdu script-based languages character recognition. *Knowl-Based Syst* 23:914–923
57. Razzak MI, Husain SA, Sher M (2010) Locally baseline detection for online Arabic script based languages character recognition. *Int J Phys Sci* 5(6). ISSN:1992–1950
58. Safabakhsh R, Adibi P (2005) Nastaalight handwritten word recognition using a continuous-density variable duration HMM. *Arab J Sci Eng* 30(1 B):95–118
59. Sari T, Souici L, Sellami M (2002) Off-line handwritten Arabic character segmentation algorithm: Acsa. In: *Proceedings of the eighth international workshop on frontiers in handwriting recognition (IWFHR)*, Niagara-on-the-Lake, p 452
60. Schambach M-P (2003) Model length adaptation of an HMM based cursive word recognition system. In: *7th international conference on document analysis and recognition (ICDAR)*, Edinburgh, 3–6 Aug 2003, vol 1, pp 109–113
61. Cristianini N, Shawe-Taylor J, (2000) *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, Cambridge/New York
62. Snoussi Maddouri S, Amiri H, Belaïd A, Choisy Ch (2002) Combination of local and global vision modeling for Arabic handwritten words recognition. In: *8th IWFHR*, Niagara-on-the-Lake, pp 128–132
63. Souici L, Sellami M (2004) A hybrid approach for Arabic literal amounts recognition. *AJSE Arabian J Sci Eng* 29(2B):177–194
64. Srihari S, Srinivasan H, Babu P, Bhole C (2005) Handwritten Arabic word spotting using the CEDARABIC document analysis system. In: *Symposium on document image understanding technology*, College Park, 2–4 Nov 2005, pp 67–71
65. Sternby J, Morwing J, Andersson J, Friberg Ch (2009) On-line Arabic handwriting recognition with templates. *Pattern Recognit* 42:3278–3286
66. Tamari I (1985) Curator. *New Hebrew Letter type*. An exhibition catalog in Hebrew and English. University Gallery, Tel Aviv University, Tel Aviv

67. Tse E, Bigun J (2007) A Base-line character recognition for Syriac-Aramaic. In: Proceedings of the IEEE international conference on systems, man and cybernetics, Montréal, 7–10 Oct 2007, pp 1048–1055
68. Zahour A, Likforman-Sulem L, Boussellaa W, Taconet B (2007) Text line segmentation of historical Arabic documents. In: Proceedings of the ninth international conference on document analysis and recognition (ICDAR), Curitiba, vol 1, pp 138–142

Further Reading

A good overview of Arabic handwriting recognition can be found in [47]. For a good example of a Hebrew recognition system, the reader is referred to [16].