Effect of Ghost Character Theory on Arabic Script Based Languages Character Recognition

MUHAMMAD IMRAN RAZZAK (imranrazak@hotmail.com)
Department of Computer Science, International Islamic University, Pakistan.
ABDEL BELAID
Director READ Group, LORIA, Nancy, France.
SYED AFAQ HUSSAIN
Chairman, Department of Computer Science, Air University, Pakistan.

ABSTRACT

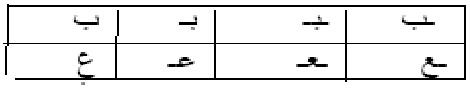
Arabic script is used by more than 1/4th population of the world in the form of different languages like Arabic, Persian, Urdu, Sindhi, Pashto etc but each language ha~~ve~~ its own words meaning. The set of ش has 58 alphabets. Arabic script based languages character recognition is difficult task due to complexities involved in this script not exist in other script. The analysis of the Arabic script is very complicated due to its use of diacritical marks associated with each character and written in many fonts and style. This script has gain very ~~less~~ intention by the researcher. This paper present a novel technique named Ghost Character Recognition Theory that will helps to develop a Multilanguage character recognition system for Arabic script ~~based languages~~ based on Ghost Character Theory. The main benefit of proposed approach is that it will work~~s~~ for all Arabic script based languages by doing effort for ghost character (basic skeleton) and developing dictionary for every language. By handling all Arabic script based languages many issues will arise like recognition rate as compared to system for specific languages, but in general it is not big issue for multilingual system and at the end we will get multilingual character recognition system.

Keywords: Ghost Character Theory, Character Recognition, Arabic Script, Urdu, Persian

1. INTRODUCTION

There are at least 26% Muslim in the world having directly or indirectly interaction with Arabic language script due to the born of Islam Arabs. Basically this script is followed in many countries ~~are~~ Arabian Peninsula, Iraq, Iran, Pakistan, Afghanistan, India, Uzbekistan, Tajikistan, Kazakhstan. Furthermore this script is ~~followed~~ by many other languages like Persian, Urdu, Punjabi, Sindhi, Pashto, Blochi, etc. Arabic script based languages especially Urdu and Arabic are used in every part of the world.

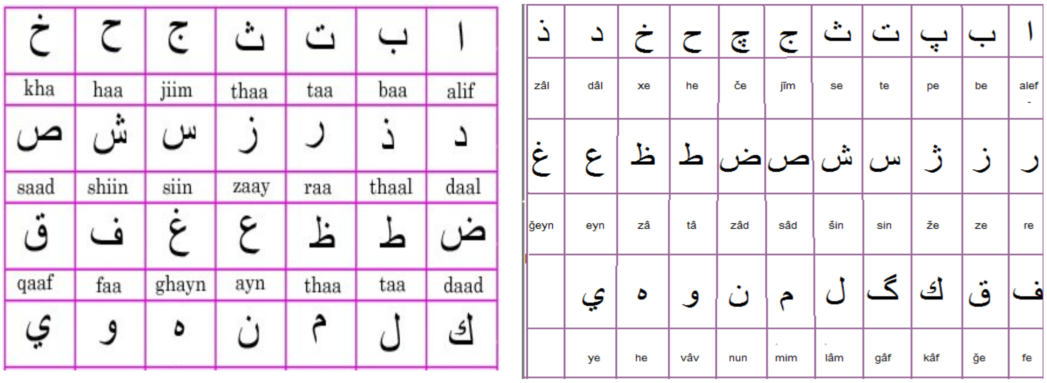
Arabic script base languages is written in cursive style from right to left in both machines printed and handwritten forms. These are the context sensitive languages and written in the form of ligatures which comprise a single or up to many different characters to form words. Most of the characters have different shapes depending on their position in the ligature e.g. the letter appeared as isolated, middle, center, end shown in figure 1. Arabic script has also uses the punctuation marks to separate sentences and have white space between ligatures and words for separation. Furthermore character overlaps each other and also contains diacritical marks (22 diacritical marks in Urdu script). While additional diacritical marks associated with ligature represent short vowels or other sounds.



ب	بـ	ـبـ	ـب
ع	عـ	ـعـ	ـع

Figure 1: Different Shapes of(ب and ع)with respect to position from left to right isolated, start, mid, end

Arabic ~~is mainly spoken in many countries are Saudi Arab, UAE, Oman, Jordan , Kuwait, Iraq etc. Arabic is the Language of Quran, a divine book on last prophet, that's why this script is used by Muslims either used directly (Arabic) or indirectly (in the form of other language like Urdu, Persian or 2nd language). It is ranked at 5th and written in Naksh style. It~~ consists of 28 alphabets shown in figure 2.a. Historically it was written without diacritical marks, latter ~~on~~ diacritical marks are added for non native by Muslim caliph. ~~Arabic has great influence on many languages especially in Muslim countries and is major source of vocabulary for many languages are Spanish, Persian, Urdu, Hindi, Punjabi, Sindhi, Pashto, Malay, Turkish, Gujarati, Kurdish, Bengali.~~



ا	ب	ت	ث	ج	ح	خ
alif	baa	taa	thaa	jiim	haa	kha
د	ذ	ر	ز	س	ش	ص
daal	thaal	raa	zaay	siin	shiin	saad
ض	ط	ظ	ع	غ	ف	ق
daad	taa	thaa	ayn	ghayn	faa	qaaf
ك	ل	م	ن	ه	و	ي

ا	ب	پ	ت	ث	ج	چ	ح	خ	د	ذ
alef	be	pe	te	se	jīm	če	he	xe	dâl	zâl
ر	ز	ژ	س	ش	ص	ض	ط	ظ	ع	غ
re	ze	že	sin	šin	sâd	zâd	tâ	zâ	eyn	ğeyn
ف	ق	ك	گ	ل	م	ن	و	ه	ي	
fe	ğe	kâf	gâf	lâm	mim	nun	vâv	he	ye	

Figure 2. a) Arabic Alphabets b) Persian Alphabets

Persian also known as Farsi is official language of Iran, Tajikistan and Afghanistan written in Arabic script (Nasta'liq style) and have alphabets 32 shown in figure 2.b. ~~It has also large influence on Urdu, Punjabi and Sindhi and other south Asian language [Lazard 1975].~~

Urdu is the 2nd most speaking language of the world but written in two main script; Arabic Script, and Devanagari script. When written in Arabic script, it is said to be Urdu and when Devanagari script is followed then its Hindi. The language scholar categorized Urdu as standard version of Hindi. Actually Urdu has different versions that depend upon regions instead of writing script [Durani 2008].Urdu is the national language of Pakistan and official language of many Indian states. Urdu written in Arabic script (Nasta'liq style) and consists of 58 basic letters shown in figure 3.a..Other languages based on Arabic script are Sindhi, Pashto Punjabi and Blochi. Punjabi is the local language of Pakistan and India. It is written in Gurmukhi and Shahmuki in Indian and Pakistani Punjab respectively. Shahmukhi is based on Arabic script and written in Nastaliq style shown in figure 3.b. Punjabi consist of 47 alphabets and ranked 11th.

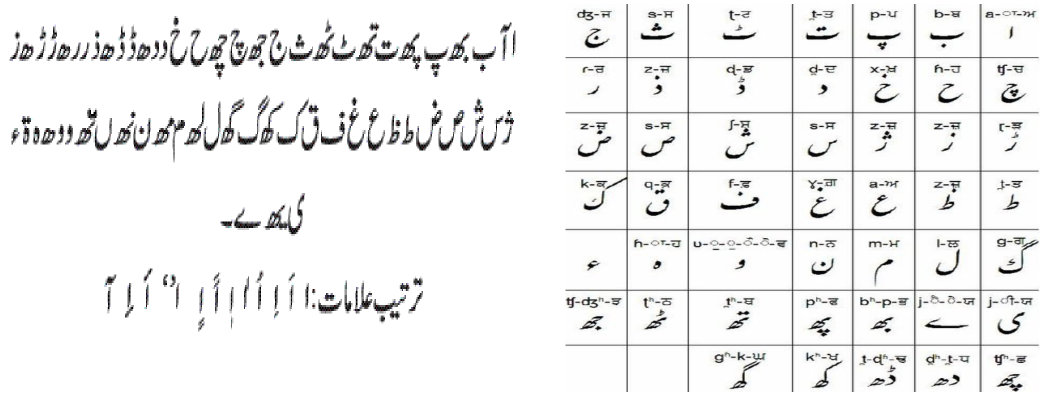


Figure 3. a) Urdu Alphabets [Durani 2008] b) Punjabi Alphabets (Shahmukhi)

Sindhi is the local language of India and Pakistan written in both Arabic and Devanagari script. It is official language in Sindh, Pakistan and some states in India. In Pakistan it is written in Arabic script and contains 52 alphabets shown in Figure 4.a. and ranked at 23. Pashto is written in Arabic script (Naskh) is spoken in Afghinstan and local language of Pakistan. It is influenced by Farsi and Avastan however most of the words are belongs to itself. It consist of 39 alphabets shown in figure 4.b. and ranked at 33.

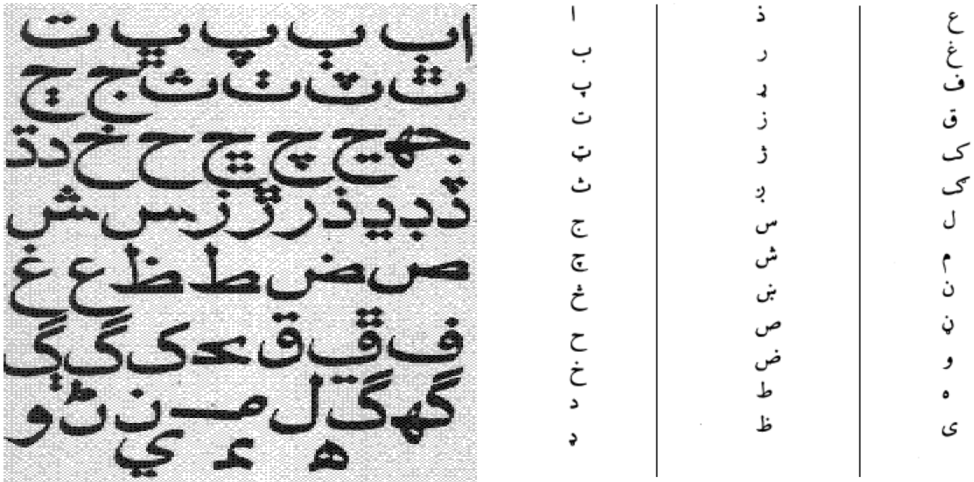


Figure 4. a) Sindhi Alphabets b) Pashto Alphabets

Urdu is the superset of all Arabic script based languages because it contain all the shapes of other languages. Local languages of Pakistan like Punjabi, Sindi, Pashto have different letter than Urdu but with the same basic shapes different diacritical marks.

2. ARABIC SCRIPT BASED LANGUAGES CHARACTER RECOGNITION

~~Character recognition is the branch of pattern recognition to imitate the computer in reading the graphical marks written by human or printed by machine so that that the machine can perform like human in reading. It has been an on-going research problem for more than four decades.~~

~~Basically character recognition is classified into three classes with respect to input namely online (handwritten), offline handwritten and offline printed recognition. In offline; input is in the form of image while in online case coordinates as well as timing information is available that make easy online character recognition little easy than offline. The offline printed character recognition is little easy task as compared to handwritten either online or offline due to large variation in writing.~~ The recognition for Arabic script based languages is much more complicated than any other language like English due to complexities of this script. The complexities are context sensitive shape, Cursiveness, Overlapping, large no of diacritical marks, segmentation of words itself and mapping of diacritical marks.

~~As h~~andwritten Arabic script is more complex than printed text, because of the variation in individual writing style. Thus recognition for handwritten Urdu is much more complicated than any other language like English due to complexities of Urdu script. The complexities are Context sensitive shape, ~~C~~ursiveness, ~~O~~verlapping , ~~each character comprises of one main strokes and multiple secondary stroke for main stroke, Baseline, Ligature , and space between the ligature~~.

Limited research efforts have been done on Arabic script based languages character recognition especially for handwritten recognition. Both segmentation base [Safabakhsh 2006, Haraty 2002, 2004, Sari 2002, Fahmy 2000, Miled 2001, Abuhaiba 1998] and holistic [Haji 2005, Meslati and Farha 2004, Adeed 2004, 2002, Khorsheed 2003, Pechwitz 2003, Dehgan 2001 and Al-Badr and Haralick 1998,1995] approaches are discussed for Arabic script based languages (both printed and handwritten) by using diacritical marks as features points with other features. There is no such (sepearte the diacritical marks form ghost character and map these diacrits marks with respect to position after recognition separately) effort proposed in the literature that leads to multilingual character recognizer.

3. GHOST CHARACTER THEORY

> *"There are some problems in Urdu ASCI code plate, when I analyzed that some symbols and all the language of Pakistan is possible from one code plate and one font. Then I proposed the idea of Ghost Character. [Durani 2008]."*

Nasta'liq and Naksh are two basic and differnt scripts that have their own fonts. Urdu is not subset of Arabic [Durani 2008]. Basically Urdu alphabets are the super set of alphabets of all Arabic script based languages written in Nasta'liq style. It more complicated than Naksh, due to different shapes of character and different position i.e. "Bay" has 35 shapes and placement [Durani].

All Arabic script based language can be written with only 44 ghost characters. Ghost character consist of 22 basic shapes called Kashti and 22 dot (diacritical marks) [Durani 2009]. Basically this idea was

700 years old when diacritical marks are applied on Quran to make easy to read for nonnative by Hajaj Bin Yousif. Before this there was no dots and diacritical marks. Arabs were using only 19 characters, and they read these dots less character by their cultural habits and had no difficulty in reading. The philosophy behind dots were; first character has one dot, 2nd character has 2 dot and 3rd has 3 dot. Persian also followed the Arabic script after Islam in Persia and some dots on character are added that were not in Arabic. Similarly in Urdu 4 nuqtas are added on ghost character, converted to line and then to Urdu letter "Tota" shown in figure 5.a and some of the basic shapes are added in Urdu and Persian shown in figure 5.b [Durani 2008].

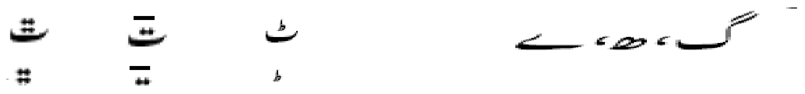


Figure 5: a. Convergence of four dots to "Tota" b. Additional shapes in Urdu and Persian.

Finally a total number of 22 ghost character are in used in Arabic script based languages are shown in figure 6. All the Arabic script based languages like Persian, Urdu, Punjabi, Sindhi, Persian Balti etc. can be written with these 22 ghost character and 22 dots and diacritical marks.

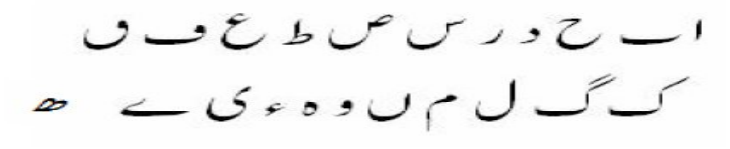


Figure 5. Ghost characters used in Arabic script based language [Durani 2008].

4. GHOST CHARACTER RECOGNITION THEORY

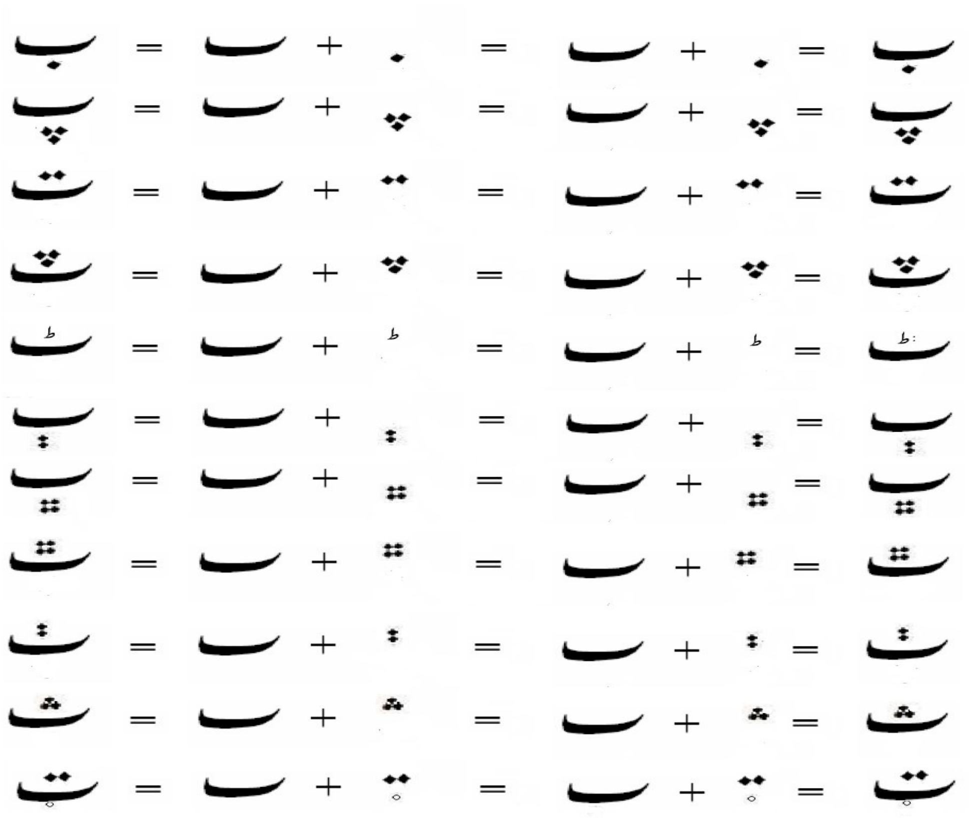
Arabic ~~script based languages~~ character recognition is very difficult task due to complicated ~~involved in this~~ script ~~and it has~~ large number of shapes even Urdu has more than 22000 ligatures. To overcome this issue ghost character recognition theory is ~~presented~~ ~~which is based on ghost character theory.~~

All the Arabic script based languages can be written with the 22 ghost character and 22 dots and diacritical marks but each language has its own phonemes and meanings of the same ligature. Thus the basic shapes(glyph) are same for all Arabic script based languages with only difference in font i.e Naksh, Nasta'liq. Nasta'liq is mainly followed by Urdu, Persian and Punjabi and it is more complicated than Naksh i.e "Bey" has 32 shapes. Idea of ghost character theory has great influence on Arabic script based languages character recognition. Ghost character theory gave an idea which made the character recognition of Arabic script easy and able to develop to Multilanguage system by doing efforts on ghost character. The ghost character recognition theory is divided into four basic steps are

1. First step is to segment the additional marks i.e dots, diacritical marks form the word. Now this word consist of only ghost characters (khali kashti) and diacritical marks and dots are separated by keeping the record of their position.
2. Recognize the separated basic shape through classifier.

3. Recognize the diacritical marks and dots associated with recognized ligature
4. Map the diacritical marks and dots on to the recognized ghost character.

The above process is shown in figure 6 for 2nd ghost character of figure 5 used in all Arabic script based languages like Arabic, Urdu, Persian, Sindhi, Punjabi, Pashto etc..



Similar shaper letters

Separated Ghost character and Dots

Recognized Ghost character and Dots separately

Mapped Dots on Ghost character

Figure 6: Recognition of 2nd ghost character letter with associated dot

~~It's a very difficult task to recognize Arabic script based languages due to complexities involved in the script, especially for handwritten text. Ghost character recognition put a big overhead on recognition engine to classify different fonts like Nasta'liq, Naksh by one classifier. This will increase the complexity and reduce the recognition rate. This issue can be resolved by extracting the font independent structural features like loop, cusp, end points, line shapes etc.~~

MERITS

The major benefit of the proposed ghost character recognition theory is that recognition system developed based on GCRT will works for all Arabic script based languages by mapping the diacritical marks and dots latter with respect to every language.. Although it is not easy to develop such system that will works for different fonts i.e Nasti'liq, Naksh. Nasti'liq and Naksh are the two most followed by these languages i.e Naksh is used for Arabic which Nasti'liq is used for Urdu, Punjabi and Persian. The over all ligatures are decrease.

Ligature$_{\text{Multilanguage}}$ = No of total ligatures by Arabic script based languages
Ligature$_{\text{Arabic}}$ = No of total ligatures of Arabic
Ligature$_{\text{Urdu}}$ = No of total ligatures of Urdu
Ligature$_{\text{Persian}}$ = No of total ligatures of Persian
Ligature$_{\text{Punjabi}}$ = No of total ligatures of Punjabi
Ligature$_{\text{other Arabic script based languages}}$ = No of total ligatures of other Arabic script based languages like Pashto, Sindhi etc

$$\text{Ligature}_{\text{Multilanguage}} \lll \text{Ligature}_{\text{Arabic}} + \text{Ligature}_{\text{Urdu}} + \text{Ligature}_{\text{Persian}} + \text{Ligature}_{\text{Punjabi}} + \text{Ligature}_{\text{other Arabic script based languages}}$$

DEMERITS

With the big advantage it has some disadvantages are

1. Now there are Multilanguage in one classifier, thus the number of ligatures are increased. i.e. Urdu has more that 22000 ligatures.
2. It’s a very difficult and complex task to develop classifier multi font for Arabic script based languages.
3. The recognition rate will be less due to multi font and large number of ligatures.

5. CONCLUSION

Every fourth person in the world is Muslim and Arabic script is used directly or indirectly by Muslims further more this script is also used by non Muslims especially in Asian countries. A large number of languages Arabic, Urdu, Persian, Punjabi, Pashto etc. are written in Arabic script. Urdu is the language that contains 58 alphabets; the basic shapes in Urdu also exit in other languages. Thus Urdu is the superset of all other Arabic script based languages. This paper presents a novel technique; ghost character recognition theory that helps to develop Multilanguage character recognition system for all Arabic script based languages. The main advantage of the proposed technique that recognition system will works for all Arabic script based languages by classifying ghost character and mapping the associated diacritical marks and dots latter with respect to selected language. Although it is not easy task to develop such system due to the complexities in Arabic script, especially for handwritten text. Furthermore it is very complex and put overhead on recognition engine to classify different fonts like Nasta'liq, Naksh by one classifier.

7. REFERENCES

Abuhaiba, M.J.J. Holt, and S. Datta, “Recognition of Off-Line Cursive Handwriting,” Computer Vision and Image Understanding, vol. 71, pp. 19-38, 1998.

Attash Durani, "Pakistani: Lingual Aspect of National Integration of Pakistan", www.nlauit.gov.pk.

Attash Durani, "Urdu Informatics" Vol. 1, pp. 102-112, pp 8-15, National Language Authority Press

A. Dehghani, F. Shabani, and P. Nava, “Off-Line Recognition of Isolated Persian Handwritten Characters Using Multiple Hidden Markov Models,” Proceeding International Conference Information Technology: Coding and Computing, pp. 506-510, 2001.

B. Al-Badr and R. Haralick, “A Segmentation-Free Approach to Text Recognition with Application to Arabic Text,” International Journal Document Analysis and Recognition, vol. 1, pp. 147-166, 1998.

B. Al-Badr and R. Haralick, “Segmentation-Free Word Recognition with Application to Arabic,” Proc. International Conference Document Analysis and Recognition, pp. 355-359, 1995.

H. Miled and N.E. Ben Amara, “Planar Markov Modeling for Arabic Writing Recognition Advancement State,” Proc. International Conference Document Analysis and Recognition, pp. 69-73, 2001.

Gilbert Lazard, “The Rise of the New Persian Language” in Frye, R. N., The Cambridge History of Iran, 1995, Vol. pp. 595–632, Cambridge: Cambridge University Press.

L. Souici, N. Farah, T. Sari, and M. Sellami, “Rule Based Neural Networks Construction for Handwritten Arabic City-Names Recognition,” Proceeding Artificial Intelligence: Methodology, Systems, and Applications, pp. 331-340, 2004.

M..M. Fahmy and S. Al Ali, “Automatic Recognition of Handwritten Arabic Characters Using Their Geometrical Features,” Studies in Informatics and Control Journal., vol. 10, 2001.

M.S. Khorsheed, “Recognising Handwritten Arabic Manuscripts Using a Single Hidden Markov Model,” Pattern Recognition Letters, vol. 24, pp. 2235-2242, 2003.

M. Pechwitz and V. Ma¨rgner, “HMM Based Approach for Handwritten Arabic Word Recognition Using the IFN/ENIT-Database,” Proc. International Conference Document Analysis and Recognition, pp. 890-894, 2003.

R. El-Hajj, L. Likforman-Sulem, and C. Mokbel, “Arabic Handwriting Recognition Using Baseline Dependant Features and Hidden Markov Modeling,” Proceeding International Conference Document Analysis and Recognition, pp. 893-897, 2005.

R. Safabakhsh and P. Adibi, “Nastaaligh Handwritten Word Recognition Using a Continuous-Density Variable-Duration HMM,” The Arabian Journal Science and Engineering., vol. 30, pp. 95-118, 2005.

R. Haraty and A. Hamid, “Segmenting Handwritten Arabic Text,” Proceeding. Int. Conf. Computer Science, Software Eng., Information Technology, e-Business, and Applications, 2002.

R. Haraty and C. Ghaddar, “Arabic Text Recognition,” International Arab Journal Information Technology, vol. 1, pp. 156-163, 2004.

S. Alma’adeed, D. Elliman, and C.A. Higgins, “A Data Base for Arabic Handwritten Text Recognition Research,” Proceeding Eighth International Workshop Frontiers in Handwriting Recognition, pp. 485-489, 2002.

T. Sari, L. Souici, and M. Sellami, “Off-Line Handwritten Arabic Character Segmentation Algorithm: ACSA,” Proc. Internationall Workshop Frontiers in Handwriting Recognition, pp. 452-457, 2002.