Muhammad Imran RAZZAK[1,2], Abdulrahman A. MIRZA[1]

Information System Department, King Saud University , Saudi Arabia (1), International Islamic University, Islamabad, Pakistan (2)

Ghost Character Recognition Theory and Arabic Script Based Languages Character Recognition

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 have its own words meaning and set of alphabets. The set of Urdu alphabets is a superset of the alphabets sets for all other Arabic script based languages. Arabic script based languages character recognition is one of the most difficult task due to complexities involved in this script not exist in any other script. This paper present a novel technique 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 works for all Arabic script based languages by doing little effort for ghost character (basic skeleton) and developing dictionary for every language. Handling all Arabic script based languages has several issues like recognition rate is low as compared to system for specific languages and specific writing style i.e. Nastaliq or Naskh, but in general, this small difference of recognition rate is not a big issue for multilingual system and at the end we will get multilingual character recognition system.*

Streszczenie. *Języki arabskie są bardzo trudne do zaadaptowania w systemie automatycznego rozpoznawania znaków. W artykule opisano algorytm Ghost character umożliwiający realizację OCR większości języków arabskich. (**Algorytm Ghost character w zastosowaniu do rozpoznawania znaków języka arabskiego**)*

Keywords: Ghost Character Theory, Multilingual, Character Recognition, Arabic Script, Urdu, Persian.
Słowa kluczowe: rozpoznawanie znaków, język arabski

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 etc. 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, centre, 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.

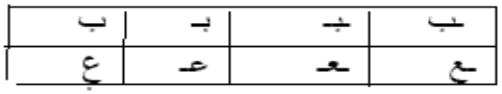

Fig 1: Different Shapes of(ب and ع)with respect to position from left to right isolated, start, mid, end

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 Naskh 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.

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

Fig 2.a. Arabic Alphabets

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

Fig 2.b. Persian Alphabets

Persian also known as Farsi is official language of Iran, Tajikistan and Afghanistan written in Arabic script (Nasta'liq style) and has alphabets 32 shown in figure 2.b. It has also large influence on Urdu, Punjabi and Sindhi and other south Asian language [8].

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 consists of 47 alphabets and ranked 11th.

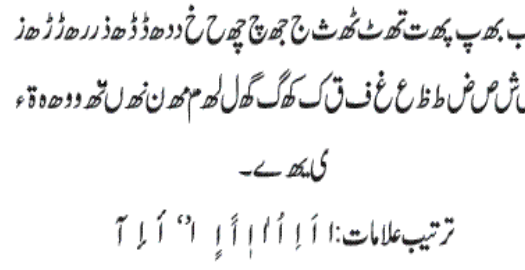

Fig 3.a. Urdu Alphabets [3]

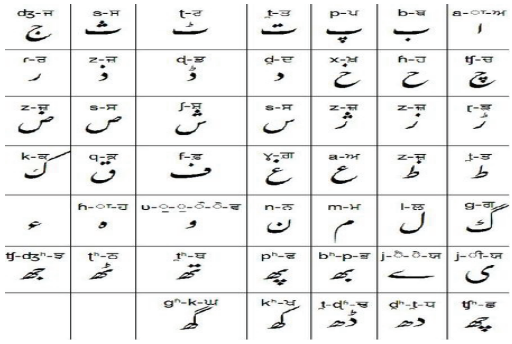

Fig 3.b. Punjabi Alphabets (Shahmukhi)

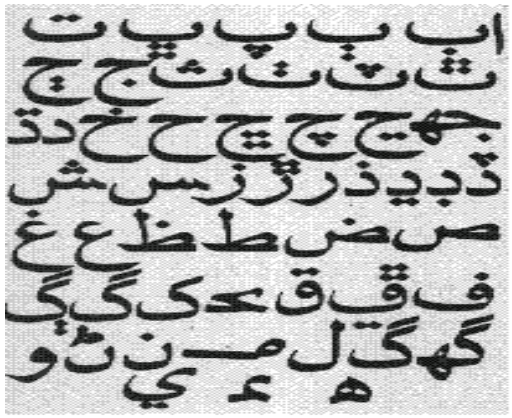

Fig 4.a. Sindhi Alphabets

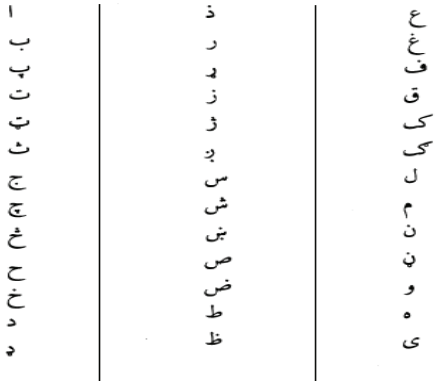

Fig 4.b. Pashto Alphabets

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 Afghanistan 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.

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

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 handwritten Arabic script is more complex than printed text, because of the variation in individual writing style. Thus recognition for handwritten Nasta'liq is much more complicated as compared to Naskh writing style due to its complex structure.

Limited research efforts have been done on Arabic script based languages character recognition especially for handwritten recognition even there is no Multilanguage character recognition system is available while there is very high similarity level between Arabic script based languages. Both segmentation base [1], [7], [10], [15-17], [19] and holistic [4-6], [11-13], [18] 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 (separate the diacritical marks form ghost character and map these diacritical marks with respect to position after recognition separately) effort proposed in the literature that leads to multilingual character recognizer.

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. [2]."

Nasta'liq and Naksh are two basic and different 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 consists of 22 basic shapes called Kashti and 22 dot (diacritical marks) [3]. Basically this idea was 700 years old when diacritical marks are applied on Quran to make easy to read for non-native

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 fig. 5.a and some of the basic shapes are added in Urdu and Persian shown in fig. 5.b [2].

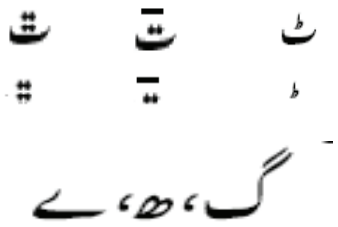

Fig 5.a. Convergence of four dots to "Tota" b. Additional shapes in Urdu and Persian.

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.

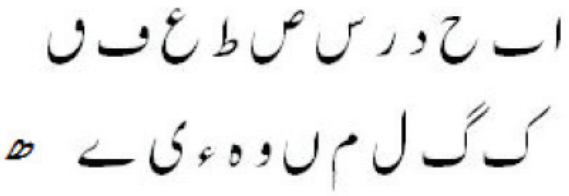

Fig 5.b. Ghost characters for Arabic script based language [2].

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 only Urdu has more than 22000 ligatures. No research efforts have been done on the side of Multilanguage character recognition system even there is minor difference between scripts followed by these languages. Most of the work done is the language specific while Multilanguage system can easily be achieved by making little more efforts on pre-processing and post processing phases. To overcome language specific character recognition with Multilanguage character recognition for Arabic script, ghost character recognition theory is presented.

All the Arabic script based languages can be written with the 22 ghost character and 22 dots and diacritical marks but 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 (glyph) are same for all Arabic script based languages with only difference in font i.e Naksh, Nasta'liq and diacritical marks followed by every language. Nasta'liq is mainly followed by Urdu, Persian, Sindhi and Punjabi and it is more complicated than Naksh i.e. "Bey" has 32 shapes shown in figure 8. Ghost character theory has great influence on Arabic script based languages character recognition even not only in language specific but also Multilanguage system. 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 from the word. Now this word consist of only ghost characters (khali kashti) and diacritical marks and diacritical marks associated with each ligature.
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.

As it is a very difficult task to classify Arabic script based languages due to complexities involved in the script, especially for handwritten text. The training of every language put a big overhead on recognition engine to classify different writing styles like Nasta'liq, Naksh by one classifier. This will increase the complexity and reduce the recognition rate.

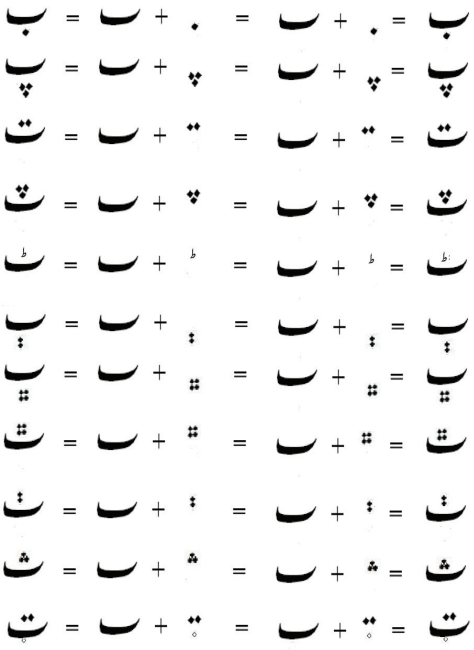

Fig 6: Recognition of 2nd ghost character letter with associated dot

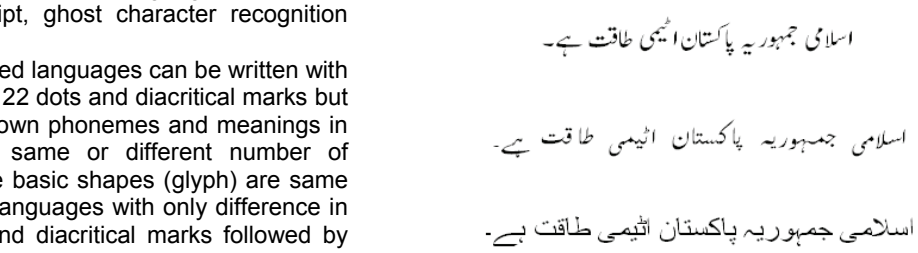

Fig 7. Urdu Samples in three different styles. Urdu Nasta'liq, Urdu Nasq, Naskh

This issue can be resolved by implementing the ghost character theory and extracting the style independent structural features like loop, cusp, end points, line shapes

etc. In the other words this can be done by developing two separate system for most using writing styles Naskh and Nasta'liq. Nasta'liq style is more complex than other style followed by Arabic script based languages shown in figure 7 and figure 8. The character appears in Nasta'liq style may also appear in Naskk etc styles with little variation. The system developed for Nasta'liq by using structural features may also work for other writing styles.

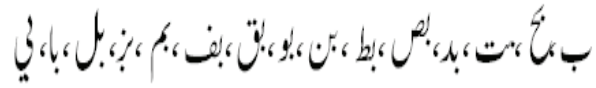

Figure 8. Different shapes of "ب" in Nasta'liq Font with respect to neighbor character

Fig 9. Feature Comparison of Nasta'liq and Naskh

It is very difficult to recognize directly, due to large variation and large data set. So the solution is to extract unique, meaningful with high class difference features from the input data to reduce the dimensionality. Generally the shape or image of word skeleton allows getting some features which are very difficult to extract from the input data. There are different kinds of features with respect to extraction mode i.e. statistical, structural, directional etc.

Basically the structural features i.e. loop, cusp, endpoint etc are intuitive aspects of writing and computed from the skeleton of the ligature. Furthermore the extraction and mapping of diacritical method is also based on the structural features especially for Arabic script based languages which are healthy in diacritical marks. Due to this reason structural features are mostly used for Arabic script based languages in literature. By deeply analyzing the both Nasta'liq and Naskh, we concluded that structural features for Urdu script written in Nasta'liq font may also work for other script written in either Nasta'liq or Naskh style. This is due to the complexities in the Nasta'liq script. The shapes in Nasta'liq are more complex and vary up to 32 with respect to its associated character and position while in Naskh shapes are only four deepening upon the position of the character.

Results and Discussions

For the testing of proposed Ghost Character Recognition Theory, we implement the proposed theory on Razzak et.al work; a fuzzy and HMM based online Urdu script based language character recognition system for both Nasta'liq and Naskh writing style [14]. Basicaly Naskh and Nasta'liq are mostly followed by Arabic script based languages. Nasta'liq is mostly followed for Urdu, Punjabi, Sindhi etc. whereas Naskh is mostly followed for Arabic, Persian etc. Thus we selected this work because of two reasons; it can recognize both Naskh and Nasta'liq writing style and recognition of diacritical marks and primary strokes are done separately. The mapping of diacritical marks and dictionary mapping is dependent upon the language selection. Each language has its own dictionary, thus the ligature formation based on diacritical marks and word formation based on the ligature is fully based on the selected language. As every language has its own rule, ligatures and word but the basic shapes are same. The recognition of basic shapes does not any need of language rules, dictionary etc. It is only depended to the writing style used i.e. Nasta'liq or Naskh etc. Whereas the ligature formation from recognized ghost character and recognized diacritical marks, and word formation from recognized ligature the language modelling is required because it is fully depended upon the language.

Dictionary D= (Urdu, Arabic, Persian, Punjabi, Pashto, Sindhi)

Ligature Dictionary for Urdu = [L1 { ……..} , L2 {……}

Li{حب،حت،حپ،حٹ،حث،جب،جت،جپ،جٹ،جث،
خب،خت،خپ،خٹ،خث،چب،چت،چپ،چٹ،چث}

……Ln{…….}.]

The mapping of diacritical marks with respect to dictionary on same ghost ligature and same no of diacritical marks is shown in figure 10.

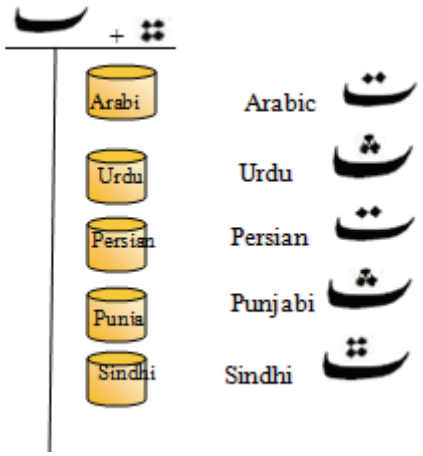

Fig 10. Combination of diacritical marks with respect to languages

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 overall ligatures are decrease.

Ligature Multilanguage = No of total ligatures by Arabic script based languages

Ligature Arabic = No of total ligatures of Arabic

Ligature Urdu = No of total ligatures of Urdu

Ligature Persian = No of total ligatures of Persian

Ligature Punjabi = No of total ligatures of Punjabi

Ligature other Arabic script based languages = No of total ligatures of other Arabic script based languages like Pashto, Sindhi etc

Ligature Multilanguage <<< Ligature Arabic + Ligature Urdu + Ligature Persian + Ligature Punjabi + Ligature other Arabic script based languages

Demerits

With the big advantage it has some disadvantages are:

Now there are Multilanguage in one classifier, thus the number of ligatures are increased. i.e. Urdu has more that 22000 ligatures.

It's a very difficult and complex task to develop classifier multi font for Arabic script based languages.

The recognition rate will be less due to multi font and large number of ligatures.

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 style by one classifier. By deeply analyzing the shapes, appearance and structure of both Nasta'liq and Naskh script with Urdu linguistics we concluded that structural features for Urdu script written in Nasta'liq font may also work for other script written in either Nasta'liq or Naskh style. This is due to the complexities in the Nasta'liq script. For testing purpose we used Razzak et al work on both Nasta'liq and Naskh styles [14]. The results shows the an efficient Multilingual character recognition for Arabic script based languages system can be developed by making little efforts on pre-processing and post processing steps.

REFERENCES

[1] 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.

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

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

[4] 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.

[5] 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.

[6] Al-Badr and R. Haralick, "Segmentation-Free Word Recognition with Application to Arabic," Proc. International Conference Document Analysis and Recognition, pp. 355-359, 1995.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] M.I. Razzak , F. Anwar, S.A.Hussain, M. Sher, "Fuzzy and HMM: A Hybrid Approach for Urdu Script Based Languages Character Recognition" Knowledge Based System (Accepted), Elsewhere

[15] 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.

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

[17] R. Haraty and C. Ghaddar, "Arabic Text Recognition," International Arab Journal Information Technology, vol. 1, pp. 156-163, 2004.

[18] 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.

[19] T. Sari, L. Souici, and M. Sellami, "Off-Line Handwritten Arabic Character Segmentation Algorithm: ACSA," Proc. International Workshop Frontiers in Handwriting Recognition, pp. 452-457, 2002.

***Authors**: Dr. Muhammad Imran Razzak, International Islamic University, Islamabad, Pakistan and Information System Department, King Saud University, Saudi Arabia imranrazak@hotmail.com; Abdulrehman A. Mirza is associate professor at Information System Department, King Saud University, Saudi Arabia.*