Lexicon Reduction for Urdu/Arabic Script Based Character Recognition: A Multilingual OCR

SAEEDA NAZ*, ARIF IQBAL UMAR**, AND MUHAMMAD IMRAN RAZZAK***

RECEIVED ON 08.04.2015 ACEPTED ON 28.05.2015

ABSTRACT

Arabic script character recognition is challenging task due to complexity of the script and huge number of ligatures. We present a method for the development of multilingual Arabic script OCR (Optical Character Recognition) and lexicon reduction for Arabic Script and its derivative languages. The objective of the proposed method is to overcome the large dataset Urdu and similar scripts by using GCT (Ghost Character Theory) concept. Arabic and its sibling script languages share the similar character dataset i.e. the character set are difference in diacritic and writing styles like Naskh or Nasta'liq. Based on the proposed method, the lexicon for Arabic and Arabic script based languages can be minimized approximately up to 20 times. The proposed multilingual Arabic script OCR approach have been evaluated for online Arabic and its derivative language like Urdu using BPNN. The result showed that proposed method helps to not only the reduction of lexicon but also helps to develop the Multilanguage character recognition system for Arabic Script.

Key Words: Urdu Optical Character Recognition, Multilingual Optical Character Recognition, Naskh, Nasta'liq.

1. INTRODUCTION

OCR is extensive and exhaustive research filed of pattern recognition and machine vision since early days of computer. The character recognition system has attained its maturity level for Latin and East Asian scripts and today, there are several OCR systems for different languages are available with accuracy reporting to 100%. But, on the other hand, researchers are struggling for the development of Arabic and other languages following Arabic script [1,2]. The character recognition for Arabic cursive script based languages is an open and active research problem. Arabic script based character recognition system is highly demanding task for document editing, sign board reading, license plate reading and mail sorting etc.

Arabic and its derivatives languages like Arabic, Punjabi, Pushtu, Sindhi, Persian, Kurdish, Jawi, and Urdu are more complex than Latin scripts [10]. The complexity of this script comes from their inherent morphology, such as context sensitive shape, cursiveness, and overlapping of word parts. In 1975 [4], the first work on Arabic character recognition was published while the efforts in field of

* Department of Information Technology, Hazara University, Mansehra, KPK, and Government Post-Graduae Girls College No. 1, Higher Education Department, Abbottabad, KPK.
** Department of Information Technology, Hazara University, Mansehra, KPK.
*** King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia.

character recognition system development for Arabic language were started in 1970 [5]. Finally, the Arabic OCR was developed in 1990s [6]. On other hand, Urdu language is also derivative language of Arabic but following the Nata'liq writing style. The first work for Urdu script was published in 2000s. The literature shows that even though Arabic script character recognition research is ongoing since 2 decades but still the character recognition system has not attained reasonable recognition rate as compare to the Latin script and East Asian languages [6].

Recent development for Arabic and its derivative languages shows that the accuracy of OCR system is affected by the language complexity as well as huge number of ligatures [7,11,13]. In this work, we analyze and compare the character set of Arabic script and its derivatives languages and present multilingual OCR concept as well as method for lexicon reduction method for Urdu script based language character recognition system.

2. ARABIC SCRIPT BASED LANGUAGES

Historically, Arabic script was written without diacritics marks. The basic character set shapes were 19 at early age. Later on in the 7th century, the diacritical marks (Hamza, dots, tota etc.) were added to the script by Iraqi governor Hajaj Bin Yousif for non-native readers. The philosophy behind the addition of new diacritical marks with base shape, is the similarity of sounds(in spoken) of base/ primary characters having same shape i.e. the first invented character carry single dot (..), second character having similar sound will carry two dots (◌̇)and likewise third and fourth characters with three dots and four dots for speaking the more similar sounds phonemes. In Urdu language, the concept of four dots were converted to horizontal line which were finally adapted into "Tota" (◌ؕ).

Associated with the Islamic movement, the Arabic character set was spread to several countries and become the writing script for other languages which do not belong to the Semitic. Some of the languages borrowed all the Arabic character set such as Persian, Urdu, Sindhi, and Pashto where as some languages borrowed few characters such as Kurdish Sorani.

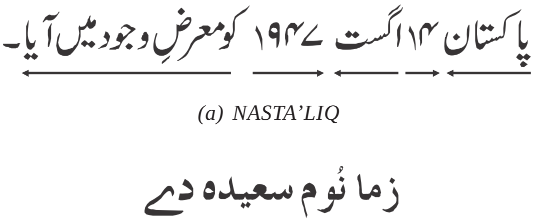
This new development require new letters, to represent a different sounds that were absent in Arabic scripts. As a consequence, several new shapes were invented in other derivative languages to represent new sounds, instead of borrowing from other scripts. The character set of Urdu, Arabic, Persian, Pashtu, Jawi and Sindhi are 38, 28, 32, 45, 33 and 52, respectively. The detail of character set is presented in Table 1. Arabicis mother but the subset of all its derivative languages. The new addition of characters in Arabic character set to deal other languages is shown in red color in Table 1.

As the Arabic and its derivatives languages share the same writing script shape, thus the basic glyph is exactly same as of Arabic. The character differ with different number of diacritical marks and their positions [1]. Additionally, there are also some other small marks known as uncommon diacritic for showing short vowels but these small marks are not using often.

The Arabic script is cursive script and written from right to left while the numerals are written from left to right. Arabic and its derivative languages follow several scripts such as Nasta'liq, Nori Nasta'liq Naskh, Diwani, Kufi, Ta'liq, Thuluth and Riq'a etc. Naskh and Nasta'liq are two commonly followed writing styles by most of Arabic script based languages [1,10,11] as shown in Fig. 1(a-b).

Arabic script is a cursive script. Arabic script based languages use joiner and non-joiner property. Due to this property and cursivenes, the characters join with its preceding or/and proceeding character on a horizontal baseline whereas the characters are not joined with each other on occurrence of non-joiner letters [7,12]. In case of Naskh writing style, the character shapes may be up to four for each character. The shape of the character depends upon the character position in the ligature (i.e. initial, middle, final and isolated form as shown in Fig. 2(a)) and Table 2. The number of shapes of each character in alphabet set is not 4in Nasta'liq writing style. The initial shape of character "bay" (ب) depends upon the association of characters on following characters as shown in Fig. 2(b). Singh [11] reported 26000 unique shapes/ligatures for Urdu Nasta'liq writing style.

Even though the huge similarity (almost same character set), the recent development for Arabic and its derivative languages shows that the researchers are focusing on the


پاکستان ۱۴ اگست ۱۹۴۷ کو معرضِ وجود میں آیا۔
زما نُوم سعیده دے

(a) NASTA'LIQ

(b) NASKH

FIG. 1. AN EXAMPLE OF SENTENCE

TABLE 1. SETS OF ALPHABETS OF ARABIC SCRIPT BASED LANGUAGES LIKE SINDHI, PASHTU, URDU, JAWI, FARSI AND ARABIC

No	Sindh	Pashto	Urdu	Jawi	Farsi	Arabic
1.	ا	ا	ا	ا	ا	ا
2.	ب	ب	ب	ب	ب	ب
3.	ٻ	پ	پ	ت	پ	ت
4.	ڀ	ت	ت	ة	ت	ث
5.	ت	ټ	ٹ	ث	ث	ج
6.	ٿ	ث	ث	ج	ج	ح
7.	ٽ	ج	ج	چ	چ	خ
8.	ٺ	ځ	چ	ح	ح	د
9.	ث	چ	ح	خ	خ	ذ
10.	پ	څ	خ	د	د	ر
11.	ج	ح	د	ذ	ذ	ز
12.	ڄ	خ	ڈ	ر	ر	س
13.	جھ	د	ذ	ز	ز	ش
14.	ڃ	ډ	ر	س	ژ	ص
15.	چ	ذ	ڑ	ش	س	ض
16.	ڇ	ر	ز	ص	ش	ط
17.	ح	ړ	ژ	ض	ص	ظ
18.	خ	ز	س	ط	ض	ع
19.	د	ژ	ش	ظ	ط	غ
20.	ڌ	ږ	ص	ع	ظ	ف
21.	ڏ	س	ض	غ	ع	ق
22.	ڊ	ش	ط	ڠ	غ	ك
23.	ڍ	ښ	ظ	ف	ف	ل
24.	ذ	ص	ع	ڤ	ق	م
25.	ر	ض	غ	ق	ک	ن
26.	ڙ	ط	ف	ک	گ	ه
27.	ڙھ	ظ	ق	ڬ	ل	و
28.	ز	ع	ک	ل	م	ي
29.	س	غ	گ	م	ن	
30.	ش	ف	ل	ن	و	
31.	ص	ق	م	و	ه	
32.	ض	ک	ن	ۏ	ی	
33.	ط	ګ	و	ه		
34.	ظ	ل	ه	لا		
35.	ع	م	ھ	ء		
36.	غ	ن	ء	ي		
37.	ف	ڼ	ی	ڽ		
38.	ڦ	و	ے			
39.	ق	ه				
40.	ڪ	ء				
41.	ک	ي				
42.	گ	ی				
43.	ڳ	ې				
44.	گھ	ی				
45.	ڱ	ئ				
46.	ل					
47.	م					
48.	ن					
49.	ڻ					
50.	و					
51.	ه					
52.	ي					

development of language based OCR rather than multilingual character recognition for Arabic scripts based languages. Based on the similarity of this script, the research community needs to be enthusiastic for the development of multilingual character recognition system.

3. LEXICON REDUCTION

Even though there is a huge similarity between Arabic script based languages, no effort has been done for the development of Multilanguage character recognition system. As we discussed above, Arabic script based languages share the almost same character set. Moreover, these languages are rich in diacritical marks and same basic shape may represent many ligature/words with different diacritical marks.

Isolated	First	Middle	Last
ب	بـ	ـبـ	ـب

(a) NASKH WRITING STYLE

ب، باجا، بش، بم، بلی، بیر، بو، بغل، بچہ، بے، بطخ، بی، بقا، بچ، بت، بد

(b) NASTA'LIQ WRITING STYLE

FIG. 2. DIFFERENT SHAPES OF "BAY"

TABLE 2. DETAIL OF DIACRITICAL MARKS IN ARABIC SCRIPT BASED LANGUAGES

Number of Dots/Hamza/Toy	Number of Charactors
One Dot	10-Arabic, Persian, Malay, Urdu
	11-Malay
	12-Pashtu, Sindhi
Two Dots	2-Urdu, Persian
	3-Arabic, Malay
	10-Sindhi
	6-Pashtu
Three Dots	2-Arabic
	6-Sindhi, Pashtu, Malay
	5-Urdu, Persian
Four Dots	6-Sindhi
Toy	3-Urdu

3.1 Theory of Ghost Character

GCT (Ghost-Character-Atomization Combination Theory) also call theory of ghost character is mainly the re-assembling of characters such as the basic glyph and diacritical marks. GCT provided a way toward the development of multilingual character recognition and multi-language machine translation system in future. This theory's proponent [8] states:

"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 plate and one font. Then I proposed the idea of Ghost Character" [8-9].

According to the GCT approach, Arabic and its derivatives languages can be represented with only 22 glyph and 22 diacritical marks as shown in **Fig. 3** [9].

3.2 Ghost Theory Based Character Recognition

Most of the prior character recognition systems are language specific. Even though, there is no multilingual OCR available, while there is very high similarity between Arabic script languages. According to Ghost character theory, the glyphs (ghost character) are exactly same for all Arabic script based languages as shown in Table 1. The main characters are different only due to placement and number of diacritical marks. Multilingual character recognition system can easily be realized by exploiting techniques of preprocessing and post processing. This system can be possible to follow the theory of ghost character theory. The character recognition system based on ghost character theory divides into following steps:

(i) Separation of dots and diacritical marks from the main word or ligature. Now there are two set of

ا ٮ ح د ر س ص ط ع ڡ ٯ ک گ ل م ں ق ہ ء ی ے ھ

FIG. 3. GHOST CHARACTER SET

characters like main words and diacritics. The main word now composes of only ghost characters (khaalikaashti) without diacritical marks and dots.

(ii) A classifier trains and recognizes the separated basic word.

(iii) A classifier trains and recognize the diacritical marks and dots associated with the recognized word.

(iv) The language specific ligature formation dictionary will use for mapping the diacritical marks to the recognized ghost character.

(v) Finally, formation of word/ligature using language specific rules.

Ghost character is an important theory which not only supportive in the multilingual OCR development but it also reduces the size of lexicon. Therefore, the efforts are needed to focus on more than one language, independent of script such as commonly used Naskh and Nasta'liq. The lexicon specific to each language can be used to form words or ligatures. The proposed process of separation of words from dots and diacritics and then mapping of dots and diacritics with the recognized words are shown in Table 3. The preprocessing, segmentation and classification perform specific to Naskh and Nasta'liq. The Fig. 4 depicted graphically a multilingual character recognition system.

3.3 Lexicon Reduction and Multilingual OCR

We can now conclude from the differences of these languages from application of multilingual character recognition system application based GCT and come to the point of writing script Naskh and Nastaliq rather than languages. Although Urdu, Arabic, Pushtu, Sindhi and Jawi languages have their own character set. These set has different characters but the basic shape/glyph are same in these languages. Some of the languages contain extra glyph due to new phonemes/sounds such as "Pay, Chay, Zhay and Gaaf" (پ ،چ ،ڑ ،ژ ،گ) in Urdu, Ki, Zi and Gi (ڼ , ځ , ښ) etc. in Pushto. Thus, Pashto alphabet set can be considered = as a superset of all Arabic script based languages.

We are taking example for representation of the total number of ligatures from the combination of basic shape of "Bay" and "Ray". Urdu has 20 possible combination of "Bay" and "Ray", like wise Arabic and Jawi have 10,

TABLE 3. SEPARATION AND MAING OF BASE CHARACTERS AND DOTS/DIACTICS

Character	Separation of Dots	Mapping of Dots	Recognition
ب -	ٮ - •	ٮ + •	- ب
پ -	ٮ - ∵	ٮ + ∵	- پ
ت -	ٮ - ··	ٮ + ··	- ت
ٽ -	ٮ - ∵	ٮ + ∵	- ٽ
ٹ -	ٮ - ط	ٮ + ط	- ٹ
ٻ -	ٮ - :	ٮ + :	- ٻ
ڀ -	ٮ - ⁘	ٮ + ⁘	- ڀ
ٿ -	ٮ - ⁘	ٮ + ⁘	- ٿ
ٺ -	ٮ - :	ٮ + :	- ٺ
ث -	ٮ - ∴	ٮ + ∴	- ث
ټ -	ٮ - ··	ٮ + ··	- ټ

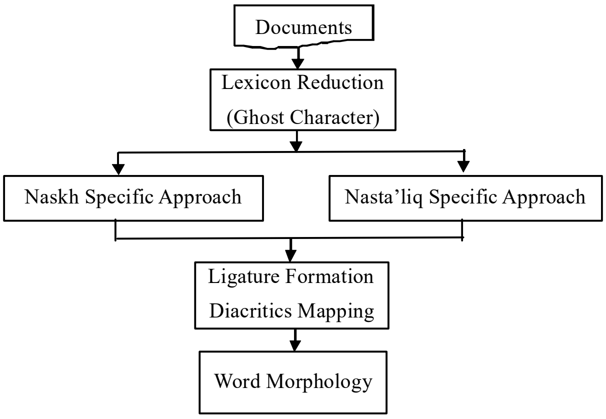


FIG. 4. THE WORKING FLOW OF PROPOSEDSYSTEM

respectively, Persian has 15 combination, Sindhi has 27 and Pashtu has 25 possible combination of ligatures. So, total of the ligatures for all languages are 107 which are shown in Table 4.

According to GCT, only one ligature/word without dots or diacritical marks is enough to represent 107 ligatures for all languages.

For development of Multilanguage OCR, we should focus on learning and recognition of basic glyphs/shapes written in Naskh or Nasta'liq.After recognition of basic ligature, the diacritical marks can be mapped with recognized word after classification. The diacritical marks mapping is dependent on the properties of languages. Each language has its own dictionary and mapping rule. Thus the ligature formation is fully based on the selected language. As all Arabic script based languages have their own rules, properties, and word but the basic shapes of the word without diacritics remain the same (Nasta'liq or Naskh). The basic shapes recognition is independent of linguistic rules and dictionary etc. but it is only dependent on Nasta'liq or Naskh writing styles. A language model is necessary for formation of word/ligature from the recognized word and diacritical marks. The dictionary of each language (Urdu, Arabic, Persian, Punjabi, Pashto, Sindhi) is used for multilingual recognition system, as shown in Fig. 5.

For validation of proposed ghost character recognition method, we have segmented the diacritical marks and base ligatures for online Arabic and its derivative languages character recognition. The diacritical marks and ghost character were trained and classified separately. For segmentation purpose, we have considered the stroke as secondary stroke if it is smaller than threshold T. We have trained three different

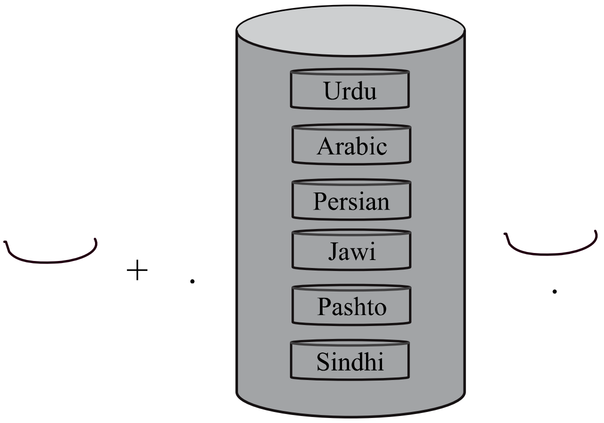


FIG. 5. DIACRITICAL MARKS AND BASED CHARACTER MAPING

TABLE 4. SIMILARITY ONE BASE LIGATURE IN DIFFERENT LANGUAGES

Language	Possible Number of Ligatures	Ligatures from Bay and Ray
Urdu	20	بر، بڑ، بز، بژ، پر، پڑ، پز، پژ، تر، ڑ، تز، تژ، ٹر، ٹڑ، ٹز، ٹژ، ثر، ثڑ، ثز، ثژ
Arabic	10	بر، بز، پر، پز، تر، تز، ٹر، ٹز، ثر، ثز
Jawi	10	بر، بز، پر، پز، تر، تز، ٹر، ٹز، ثر، ثز
Persian	15	ٹز، ٹژ، ثر، ثز، ثژ، تر، تز، تژ، ٹر، بر، بز، بژ، پر، پز، پژ
Sindhi	27	بر، بڙ، بز، ٻز، ٻڙ، ٻز، ڀر، ڀڙ، ڀز، تر، تڙ، تز، ٿر، ٿڙ، ٿز، ٿر ٿڙ ٿز، ٽر، ٽڙ، ٽز، ثر، ثڙ، ثز، پر، پڙ، پز
Pashtu	25	بر، بړ، بز، بژ، بر، پر، پړ، پز، پژ، پږ، تر، تړ، تز، تژ، تږ، ټر، ټړ، ټز، ټژ، ټږ، ثر، ثړ، ثز، ثژ، ثږ

classifiers (i.e. Classifier-I for diacritical marks, Classifier-II and Classifier-III for base glyph written in both Nasta'liq and Naskh writing style respectively). Classifier-1 is trained for diacritical marks whereas the Classifier-II and Classifier-III are trained on ghost ligatures written in Nasta'liq and Naskh writing style respectively. For evaluation purpose, we have trained classifier to six commonly used diacritical marks (one dot, two dots, three dots, hamza, toy, hey). The position of each diacritical is estimated base on the nearest character point at the time of segmentation so that it can be used mapping of diacritical marks on the ghost ligatures. Classifier-II and Classifier-III are trained on ghost ligatures and we have trained all basic characters shown in table 1 and few ligatures. For the mapping purpose, we have used grammatical rules with triples (position, ghost ligature, diacritical marks)in order to model the mapping of diacritical marks. The experiment on small dataset showed that Multilanguage character recognition system can be developed by using the concept of ghost character theory and focusing on the script rather than language. The accuracy of the proposed Multilanguage system depends upon the segmentation and mapping of diacritical marks and can be increased by adding language rules and character based recognition system.

Classifier-1 network consist of six input units (stroke size, stroke length, loops, start_shape, end_shap, diagonal) and 6 output neurons and 10 hidden neurons and learning rate .5F. Classifier-II and Classifier-III consist of 26 inputs, 166 hidden and 50 output neurons for both Nasta'liq and Naskh respectively. For the classification of ghost characters, we have used the same features extracted in [11]. We have set the learning rate to 0.5F for ghost character.

4. CONCLUSION

We discussed the script based approach rather than language dependent approach. By considering the script based approach, the future Arabic script based character recognition will be focused for Multilanguage Arabic character recognition written in Nasta'liq and Naskh writing style rather than making effort for standalone OCR for each language. We have evaluated the proposed multilingual Arabic script OCR approach for online Arabic and its derivative script using BPNN. The considering of the writing script rather than languages using ghost character theory, the result showed that proposed method not only helps in the reduction of lexicon but it also helps in the multilingual character recognition system development for Arabic Script based. Due to proposed approach, the lexicon can be reduced to at least 20%.

ACKNOWLEDGEMENT

Authors express heartfelt thanks to Prof. Attash Durrani, for his expert advice on Ghost Character Theory.

REFERENCES

[1] Naz, S., Hayat, K., Razzak, M.I., Anwar, M.W., Madani, S.A., and Khan, S.U., "The Optical Character Recognition of Urdu-Like Cursive Scripts", Pattern Recognition, Volume 47, No. 3, pp. 1229-1248, 2014.

[2] Arica, N., and Yarman-Vural, F.T., "Optical Character Recognition for Cursive Handwriting", IEEE Transaction on Pattern Anal Mach Intelligence, Volume 24, No. 6, pp. 801-813, 2002.

[3] Naz, S., Hayat, K., Razzak, M.I., Anwar, M.W., and Khan, S.Z., "Challenges in Baseline Detection of Arabic Script Based Languages", Intelligent Systems for Science and Information, Studies in Computatioanl Intelligence, Volume 542, pp. 181-196, 2014.

[4] Nazif, A., "A System for the Recognition of the Printed Arabic Characters", Master's Thesis, Faculty of Engineering, Cairo University, 1975.

[5] Al-Badr, B., and Mahmoud, S.A, "Survey and Bibliography of Arabic Optical Text Recognition", Signal Process, Volume 41, No. 1, pp. 49-77, 1995.

[6] Cheriet, M., "Visual Recognition of Arabic Handwriting: Challenges and New Directions", Arabic and Chinese Handwriting Recognition, Lecture Notes in Computer Science, Volume 4768, pp. 1-21, Springer, 2008.

[7] Naz, S., Hayat, K., Razzak, M.I., Anwar, M.W., and Akbar, H., "Arabic Script Based Language Character Recognition: Nasta'liq vs Naskh analysis", World Congress on Computer and Information Technologies WCCIT'13, Tunisia, 2013.

[8] Durrani, A., "Urdu Informatics", Center of Excellence for Urdu Informatics, National Language Authority, Islamabad, Pakistan, 2008.

[9] Durrani, A., "Pakistani: Lingual Aspect of National Integration of Pakistan", Ministry of Education Curriculum Review, Pakistan, 2009.

[10] Razzak, M.I., and Mirza, A.A., "Effect of Ghost Character Theory on Arabic Script based Languages Character Recognition", Przeglad Elektrotechniczny, [ISSN 0033-2097].

[11] Lehal, G.S., "Choice of Recognizable Units for URDU OCR", Proceedings of Workshop on Document Analysis and Recognition, pp.79-85, New York, USA, 2012.

[12] Naz, S., Umar, A.I., Shirazi, S.H., Ahmed, S.B., Razzak, M.I., and Siddiqi, I., "A Review of Segmentation Techniques for Recognition of Arabic-Like Scripts", Education and Information Technologies-Springer, Volume 20, No. 2, 2015.

[13] Ahmed, S.B., Naz, S., Razzak, M.I., Rasheed, S.F., and Khan, A.A., "Evaluation of Cursive and Non-Cursive Scripts using Recurrent Neural Networks", Neural Computation & Application-Springer, Volume 26, No. 3, April, 2015.