HACDB: HANDWRITTEN ARABIC CHARACTERS DATABASE FOR AUTOMATIC CHARACTER RECOGNITION

A. Lawgali, M. Angelova, A. Bouridane

School of Computing, Engineering and Information Sciences
Northumbria University, Newcastle upon Tyne, UK
ahmed.lawgali@northumbria.ac.uk

ABSTRACT

Automatic off-line Arabic handwriting recognition based on segmentation still faces big challenges. A database, covering all shapes of handwritten Arabic characters, is required to facilitate the recognition process. This paper introduces a new database for handwritten Arabic characters (HACDB), designed to cover all shapes of Arabic characters including overlapping ones. It contains 6,600 shapes of characters written by 50 writers. This database can be used for training and testing the words for their recognition after segmentation. Also, it presents the possibility for comparing different approaches and evaluate their accuracy on a common base.

Index Terms— Arabic, character, recognition, database

1. INTRODUCTION

Automatic off-line Arabic handwriting recognition still faces big challenges. This recognition has several important applications, such as the automatic sorting of postal mail, cheque processing or editing old documents. For a fair comparison of performance of Arabic handwriting systems, a standard database is needed. Most researchers have developed their systems based on set of data they themselves have gathered [1]. Some of the handwriting recognition systems yield high accuracy due to results being tested on a small database and not a standard database [2]. Most of the databases have also been developed for specific domains, such as postal addresses and financial use (for example cheques). In handwritten text, there is no control over the writer's ability or writing style to produce text with touching or overlapping characters. Therefore, to develop a recognition system, ideally, a large database for training and testing is required. A number of databases for handwritten English recognition have been developed. Hull [3] developed handwritten English text database for the centre of Excellence for Document Analysis and Recognition (CEDAR). Another large database which includes English characters (upper and lower cases) as well as number digits was provided by The National Institute of Standards and Technology (NIST) [4]. The MNIST database was developed out of the original NIST [5]. IAM database of handwritten English text is another example developed by Marti and Bunke [6]. However, there are only a limited number of databases for Arabic handwriting, which have been developed for specific domains. One of the most widely used is the IFN/ENIT database released by Pechwitz and Maergner [7] for handwritten Arabic words. The Arabic handwritten database (AHDB) developed by Alma'adeed *et al.* [8, 9] contains numbers and quantities used in cheques, as well as the most popular words in Arabic writing. Al-Ohali *et al.* [10] have built an Arabic cheque database containing subwords, Indian-Arabic digits and samples each of legal. The ADBase database has been developed by El-Sherif and Abdleazeem [11] for handwritten Arabic digits. A database for handwritten Arabic characters has developed by Asiri and Khorsheed [12]. To date, most researchers have developed their own databases to test their systems and, therefore, most of them are not available to other researchers. Also, to develop a recognition system based on segmentation, a database of characters is needed to facilitate the recognition process. This database has to cover all shapes of handwritten Arabic characters. This paper introduces the Handwritten Arabic Characters DataBase (HACDB) containing shapes of Arabic handwritten characters. The organization of the paper is as follows. Section 2 discusses and analyses the Arabic characters which might cause overlapping between characters. A new database for handwritten Arabic characters is introduced in Sections 3, 4 and 5. Section 6 gives the conclusion and some future work.

2. ARABIC CHARACTERS

Arabic script is written from right to left and is composed of 28 characters, with no capital or lower case. Each character has two or four shapes, the shape of the character depending on its position in the word, as shown in Table 1. The first column gives the number of the character, the second column is its name. The third represents the sign of an isolated character, with the fourth being its appearance at the beginning of the word. Finally, the fifth and sixth columns represent its appearance in the middle and at the end of the word, re-

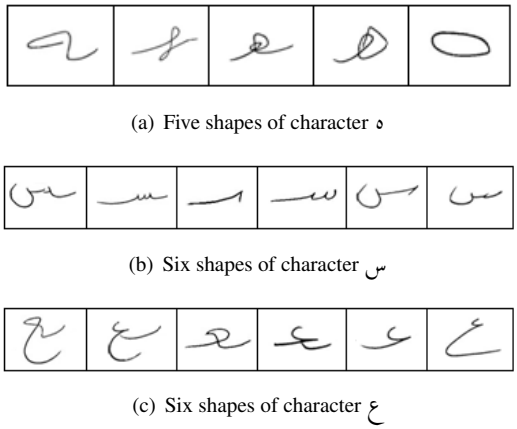
spectively. The dots play a significant role in Arabic characters. The shape of some characters is similar but the difference arises with position and number of dots, such as (ب, ت, ث), which can occur either above or below the characters. Two characters in the alphabet have three dots, three have two dots and ten have one dot. Arabic characters are composed of 28 main characters and each character has two or four shapes depending on its position in the word, thus resulting a total of 81 shapes. In fact, most of characters change slightly in shape according to their position in the word. However, the shape of some characters is similar but may become different with the number and position of dots. In handwritten Arabic characters, some characters have more than four shapes; for example, character (ه) has 5 shapes, and characters (س) and (ع) have 6 shapes, as shown in Figure 1. Therefore, there are 52 shapes of characters without dots as shown in Figure 2.

No	Name	Isolate	Beginning	Middle	End
1	Alif	ا	-	-	ـا
2	Baa	ب	بـ	ـبـ	ـب
3	Taa	ت	تـ	ـتـ	ـت
4	Thaa	ث	ثـ	ـثـ	ـث
5	Jeem	ج	جـ	ـجـ	ـج
6	Haa	ح	حـ	ـحـ	ـح
7	Khaa	خ	خـ	ـخـ	ـخ
8	Daal	د	-	-	ـد
9	Dhal	ذ	-	-	ـذ
10	Raa	ر	-	-	ـر
11	Zaa	ز	-	-	ـز
12	Seen	س	سـ	ـسـ	ـس
13	Sheen	ش	شـ	ـشـ	ـش
14	Saad	ص	صـ	ـصـ	ـص
15	Dhad	ض	ضـ	ـضـ	ـض
16	Tta	ط	طـ	ـطـ	ـط
17	Dha	ظ	ظـ	ـظـ	ـظ
18	Ain	ع	عـ	ـعـ	ـع
19	Ghain	غ	غـ	ـغـ	ـغ
20	Faa	ف	فـ	ـفـ	ـف
21	Qaf	ق	قـ	ـقـ	ـق
22	Kaaf	ك	كـ	ـكـ	ـك
23	Laam	ل	لـ	ـلـ	ـل
24	Meem	م	مـ	ـمـ	ـم
25	Noon	ن	نـ	ـنـ	ـن
26	Haa	ه	هـ	ـهـ	ـه
27	Waaw	و	-	-	ـو
28	Yaa	ي	يـ	ـيـ	ـي

Table 1. Arabic character forms

2.1. Overlapping characters

As previously stated, in handwritten Arabic text, two or more characters can be combined vertically and characterised by different shapes thus making it difficult to detect the boundaries of the characters. These particular characters such as (ا, ح, ل, م) which cause overlapping between characters have been studied and analysed. Figure 3 illustrates how some characters might cause overlapping. Figure 3(a) shows character (ا) could be combined vertically with characters (ل, ح). Character (ل) might cause overlapping in several cases. It could be overlapping with character (ا or م), and also with characters (م, ح) as shown in Figure 3(b). Figure 3(c) shows character (م) which can overlap with character (ح). These overlapped characters are difficult to segment because it is hard to detect their boundaries. Also, once segmented, the shapes of the segmented characters can be difficult to identify. For example, in Figure 4(a) the two segmented overlapping characters (ه, ل) show that the shape of the first segmented character (ل) looks like the character (ا). Also, in Figure 4(b) the three overlapping characters (ح, ه, ل) have been segmented and the shapes of the the first of two segmented characters are different from the shape of the original characters. Both examples show that the classification of overlapping characters



(a) Five shapes of character ه

(b) Six shapes of character س

(c) Six shapes of character ع

Fig. 1. Characters having more than four shapes

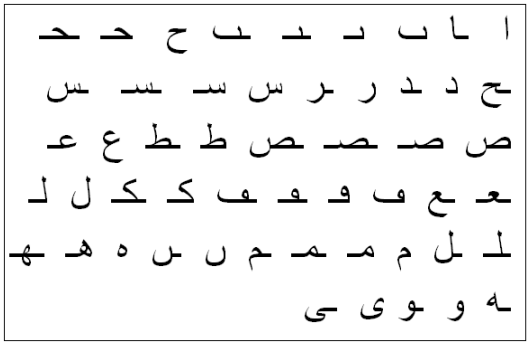
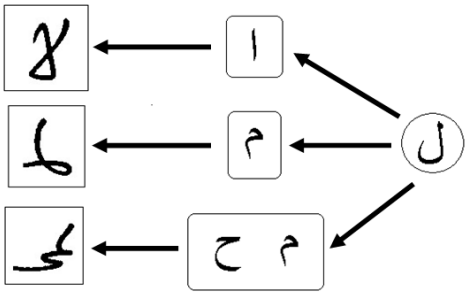


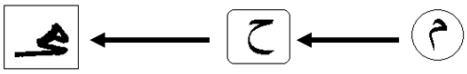
Fig. 2. The 52 different shapes of Arabic characters without dots

after the segmentation process is a difficult task. Therefore, the particular characters that might cause overlapping are considered as one shape in the database.

(a) Character (ا) with characters (ل, ح)

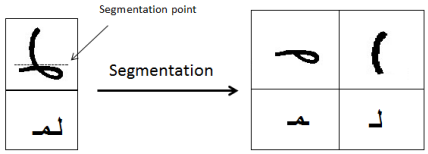


(b) Character ل with several characters

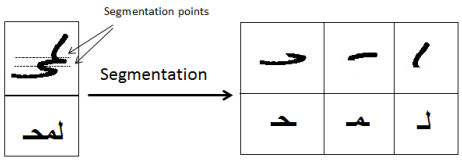


(c) Character ا with character ح

Fig. 3. Illustration of characters overlapping



(a) Segmenting two overlapped characters ه, ا



(b) Segmenting three overlapped characters ح, ه, ا

Fig. 4. Segmenting overlapped characters

3. DATA ACQUISITION

The HACDB database is designed to cover all shapes of Arabic characters and is available at (http://computing.unn.ac.uk/characterdatabase/). Due to the difficulty of segmenting overlapping characters and since existing databases do not contain such characters, they are added into the HACDB as one shape. A form for collecting data was designed with 66 small squares divided into *6*x*11* small squares. The data was collected from 50 writers, their ages ranging from 14 to 50. Each writer filled in two forms all the shapes of Arabic characters as (isolated, at the beginning of the word, in the middle and at the end of the word) also the shapes of the overlapping characters but without dots. Figure 5 shows a form filled by one writer. Although the dots play a significant role in Arabic characters, they and the marks, such as "hamza", are removed from the character list since they can affect the classification and also reduce the number of classes used. Each writer has generated two forms, where each one have 66 squares, thus resulting in 58 shapes of characters and 8 shapes of overlapping characters. A total of 132 shapes of Arabic characters written by each writer. The number of character shapes collected from 50 writers is then 6,600 shapes of characters.

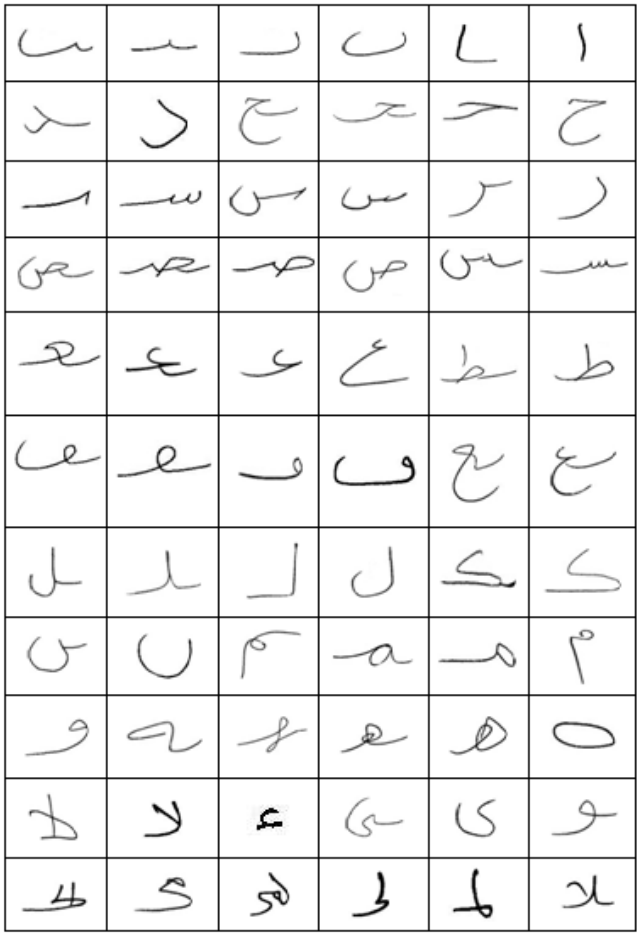


Fig. 5. A form designed by the author to collect shapes of handwritten Arabic characters

4. DATA STORING

The forms were scanned with a quality of 300 dpi. To reduce the time for processing and to do as much as possible automatically, a software based on the horizontal and vertical projection has been developed to identify and extract each shape of character to be saved as a separate image. The horizontal projection is used to identify the space between characters lines horizontally. For each line of the characters, the location of each character is detected by using the vertical histogram. Then each set of images for the same shape was saved in a separate folder with its own name. For example, an image of the shape of characters character Alif (ا) at the end of the word was saved as (Alif_E_11). This means that:

- Alif: name of the character.
- E: shape of the character at the end of the word.
- 11: the serial number of the shape of the character.

The database is divided into two versions. The first version, is used for training and testing. Therefore, it is divided into two sets (80% training and 20% testing). This version can be used for purposes such as to compare the effectiveness of some techniques for recognition of all the shapes of Arabic handwritten characters. The second version stores all shapes and can be used for training in approaches which recognize the word after segmentation.

5. PRE-PROCESSING

After scanning a character, some pre-processing tasks, such as binarization, noise removal and normalisation, were performed. The pre-processing task attempts to remove the information that have no discriminative power in the process of recognition (i.e redundant data). The images of characters are converted into a binary format: values of background pixels as 1 (white) and values of foreground pixels as 0 (black). This process is carried out by choosing an efficient thresholding method value by using Otsu's algorithm [13]. Then, the pixels that have values less than threshold are converted to 0, otherwise, to 1. If the original image is *p(x,y)*, then, it can be converted to binary format as *output(x,y)* by

$$output(x,y) = \begin{cases} 0, & \text{if } p(x,y) < \text{threshold} \\ 1, & \text{otherwise} \end{cases}$$

The small objects, not part of the handwriting, but considered as noise have also been removed. Arabic characters have different sizes. Before extracting the features of the characters, the images of characters have been resized as *128*x*128* for normalisation purposes.

6. CONCLUSION

The survey reveals that most published research use different databases for different specific aims. For example, some recognition systems achieved high accuracy because they used a small database. However, the lack of a reference database of Arabic handwriting characters which covers all shapes of Arabic characters and includes shapes of overlapping characters presents the possibility for comparing different approaches and evaluating their accuracy on a common base. The proposed HACDB database has been designed to cover all shapes of Arabic characters including overlapping ones. It contains 6,600 shapes of characters written by 50 writers. This database can be used for training and testing or it can be used for training to recognize the words after segmentation. The HACDB database is available for academic use at (http://computing.unn.ac.uk/characterdatabase/)

7. REFERENCES

[1] S. Mozaffari, K. Faez, F. Faradji, Ziaratban M., and S. M. Golzan, "A comprehensive isolated farsi/arabic character database for handwritten ocr research," *10th International Workshop on Frontiers in Handwriting Recognition (IWFHR)*, pp. 385–389, 2006.

[2] J. Alkhateeb, *Word Based Off-line Handwritten Arabic Classification and Recognition*, Ph.D. thesis, School of Computing, Informatics and Media, University of Bradford, 2010.

[3] J.J. Hull, "A database for handwritten text recognition research," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, pp. 550 –554, 1994.

[4] P. J. Grother, "Handprinted forms and character database," Tech. Rep., Technical Report and CDROM, National Institute of Standards and Technology, March 1995.

[5] L. Deng, "The mnist database of handwritten digit images for machine learning research," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 141 –142, 2012.

[6] U. V. Marti and H. Bunke, "The iam-database: an english sentence database for offline handwriting recognition," *International Journal on Document Analysis and Recognition*, vol. 5, no. 1, pp. 39–46, 2002.

[7] M. Pechwitz, S. S. Maddouri, V. Mrgner, N. Ellouze, and H. Amiri, "Ifn/enit - database of handwritten arabic words," in *In Colloque Inter. Francophone sur lEcrit et le Document, CIFED 2002*, 2002, pp. 129–136.

[8] S. Al-Ma'adeed, D. Elliman, and C. A. Higgins, "A data base for arabic handwritten text recognition research,"

in *Proceedings Eighth International Workshop onFrontiers in Handwriting Recognition,*, 2002, pp. 485 – 489.

[9] S. Al-Ma'adeed, D. Elliman, and C. A. Higgins, "A data base for arabic handwritten text recognition research," *The International Arab Journal of Information Technology*, vol. 1, no. 1, pp. 117–121, 2004.

[10] Y. AL-OHALI, M. CHERIET, and C. SUEN, "Databases for recognition of handwritten arabic cheques," *Pattern Recognition*, vol. 36, no. 1, pp. 111–121, 2003.

[11] E. EL-Sherif and S. Abdleazeem, "A two-stage system for arabic handwritten digit recognition tested on a new large database," in *International Conference on Artificial Intelligence and Pattern Recognition, AIPR-07, USA*, 2007, pp. 237–242.

[12] A. Asiri and M. S. Khorsheed, "Automatic processing of handwritten arabic forms using neural networks," in *Proceedings Of World Academy Of Science, Engineering And Technology*, 2005, pp. 147–151.

[13] N. Otsu, "A threshold selection method from gray-level hist," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.