Journal of Computer Science 7 (4): 512-518, 2011
ISSN 1549-3636

Multilevel Classifier in Recognition of Handwritten Arabic Characters

Rawan Ismail Zaghloul, Enas Faisal AlRawashdeh and Dojanah Mohammad Kadri Bader
Department of Management Information Systems, Al-Balqa'a Applied University,
Amman College, Amman, Jordan

Abstract: Problem statement: Arabic offline handwriting recognition is considered one of the most challenging topics. This is probably caused by the fact that Arabic recognition system faced many problems during the development stage. It faces the usual problems of character recognition in general, in addition to the problems that are specific to Arabic language only. The aim of this study was to build a classifier to solve Arabic text ambiguity; to be used in text recognition applications. **Approach:** A multilevel classifier based on pattern recognition techniques, is proposed. The proposed classifier was implemented using MATLAB and also tested with a large sample of handwritten datasets. **Results:** Pattern recognition techniques are used to identify Arabic handwritten text. After testing, the recognition rates reached {93, 84, 89 and 85%} for the isolated letters, letters at the beginning, at the middle and at the end of the word respectively. **Conclusion:** Even that the Arabic letters change their shape depending on their position in a word, the proposed classifier, using the powerful set of features, is proved to be effective in the recognition of Arabic letters.

Key words: Pattern recognition, binary image, Arabic alphabet, Arabic language, isolated letters, Arabic script, OCR system, radon transform, Optical Recognition System (OCR)

INTRODUCTION

Character recognition systems have never achieved a read rate that is 100%. Because of this, a system which permits rapid and accurate correction of rejects is a major requirement. The success of any Optical Recognition System (OCR) device to read accurately without substitutions is not the sole responsibility of the hardware manufacturer. Much depends on the quality of the items to be processed. Through the years, the desire has been to increase the accuracy of reading, that is, to reduce rejects and substitutions, to reduce the sensitivity of scanning to read less-controlled input, to eliminate the need for specially designed fonts (characters) and to read handwritten characters (online/offline) (Ergin *et al.* , 2010).

Arabic language has its own features that affect the recognition process of the Arabic letters. The reading of Arabic might be ambiguous; because Arabic script is fundamentally cursive most letters have slightly different shapes depending on whether they occur at the beginning, middle, or at the end of the word as depicted in Fig. 1.

In recent years, many researchers have addressed the recognition of Arabic text, including Arabic numerals (Ahranjany *et al*., 2010) and state of the art and future trends of Arabic text recognition. In addition,

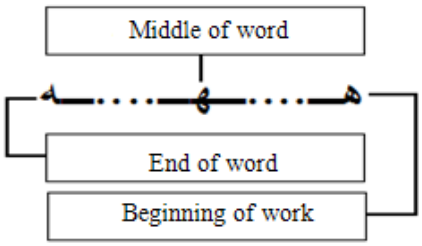

Fig. 1: Shapes of letter hah

several researchers have reported the recognition of Persian (Farsi) handwritten digits (Mahmoud and Awaida, 2009). However, the reported recognition rates for Arabic letters need more improvements to be practical (Sabri *et al*., 2009).

Abandah *et al*. (2009) suggested an approach for text feature extraction to achieve high recognition accuracy of handwritten Arabic letters. Mozaffari *et al*. (2009), proposed an algorithm for text feature extraction of Faris handwritten letters.

MATERIALS AND METHODS

Arabic text characteristics: Arabic language is one of the most ancient languages and spoken by many people

Corresponding Author: Rawan I. Zaghloul, Department of Management Information Systems, Al-Balqa'a Applied University, Amman College, Amman, Jordan

in areas around the globe. Arabic script and language have resisted any major change for centuries now. Text written or words used more than 1000 years ago are still being used and understood by schoolboys around the Arab world. Nevertheless, with the advent of computer age and information technology, efforts have been directed to adapt the Arabic script for ease of use with the new tools. Arabic alphabets have special characteristics that it composed of 28 characters, they present a lot of similarities and composed of many loops and cusps, characters are connected even when typed or printed and the most challenging point that Arabic characters change their shape depending on their position in the word (Slimane *et al.*, 2010), as shown in Fig. 2.

General outline of the proposed approach: The implemented Arabic OCR system involves four image processing techniques which are the image acquisition, the preprocessing, the feature extraction and the classification, as illustrated in Fig. 3. Preprocessing is the process of compensating a poor quality original and/or poor-quality scanning. The image is then ready for segmentation and feature extraction. However, as a recognition-based character feature extraction technique is used, a feedback loop is linked between the output of the classification stage and the input of the feature extraction stage.

The proposed model: The model starts by applying the segmentation to the scanned preprocessed character, according to the result of the segmentation stage the classifier determines the next feature to be extracted. Our model depends on the following extracted features:

- The existence of secondary parts
- The number of secondary parts
- The type of secondary parts
- The position of the secondary parts whether above or under the body of the character
- The existence of loops in the body of the character
- The Radon transform for the body part

Segmentation: Refers to the process of partitioning a digital image into multiple segments. The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries (lines, curves) in images. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain visual characteristics. Letter segmentation is the first step; in order to determine the body and the secondary parts of the letter as shown in Fig. 4.

Determining the number of secondary parts: Utilizing segmentation results, we can determine the number of objects in the image. For example, the letter "Tha" has four object labels one for the body and three for the secondary parts, as depicted in Fig. 5.

So, we can conclude that NS = NL-1, where *NS* is the number of secondary parts and *NL* is the number of object labels.

Arabic Alphabet	At the End	At the Middle	At the beginning	Isolated
Alef		ـا		ا
Ba	ـب	ـبـ	بـ	ب
Ta	ـت	ـتـ	تـ	ت
Tha	ـث	ـثـ	ثـ	ث
Jeem	ـج	ـجـ	جـ	ج
Ha	ـح	ـحـ	حـ	ح
Kha	ـخ	ـخـ	خـ	خ
Dal			د	د
Thal			ذ	ذ
Ra			ر	ر
Zain			ز	ز
Seen	ـس	ـسـ	سـ	س
Sheen	ـش	ـشـ	شـ	ش
Sad	ـص	ـصـ	صـ	ص
Dad	ـض	ـضـ	ضـ	ض
Tah			ط	ط
Thah			ظ	ظ
Ain	ـع	ـعـ	عـ	ع
Ghain	ـغ	ـغـ	غـ	غ
Fa	ـف	ـفـ	فـ	ف
Oaf	ـق	ـقـ	قـ	ق
Kaf		ـكـ	كـ	ك
Lam	ـل	ـلـ	لـ	ل
Meem	ـم	ـمـ	مـ	م
Noon	ـن	ـنـ	نـ	ن
Hah	ـه	ـهـ	هـ	ه
Waw			و	و
Ya	ـي	ـيـ	يـ	ي

Fig. 2: Shapes of Arabic letters according to their position in the word

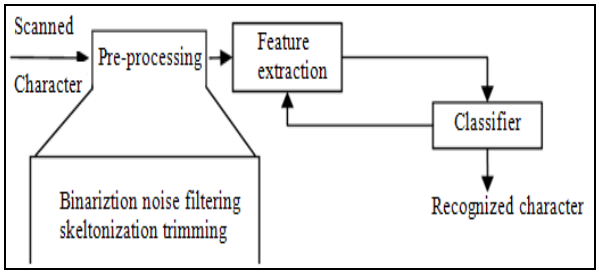

Fig. 3: Arabic text recognition block diagram

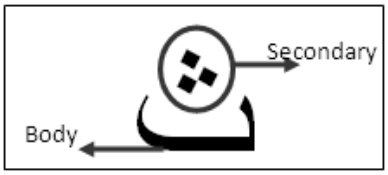

Fig. 4: The result of the segmentation on letter "Tha"

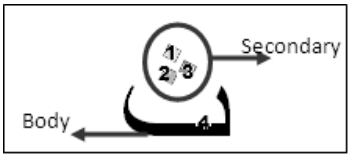

Fig. 5: The result of labeling on letter "Tha"

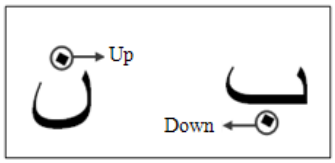

Fig. 6: The position of dot according to the body of letters "ba" and "noon"

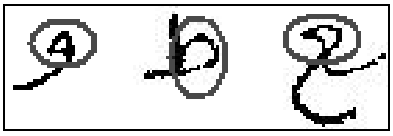

Fig. 7: Loop shapes in letters "Waw, Tah and Ain" from left to right respectively

The type of the secondary part: Does the secondary pattern present dot "." or hamza "ء"? You can differentiate between them by computing the correlation between the two patterns as illustrated in the equation bellow (Szekely and Rizzo, 2009):

$$\gamma = \frac{\Sigma_m \Sigma_n (A_{mn} - \bar{A})(B_{mn} - B)}{\sqrt{(\Sigma_m \Sigma_n (A_{mn} - \bar{A})^2)(\Sigma_m \Sigma_n (B_{mn} - \bar{B})^2)}} \quad (1)$$

Where $\bar{A}$ and $\bar{B}$ are the mean for A and B patterns respectively.

The position of the secondary parts: Another important feature in Arabic letters is to determine the position of the secondary parts according to the body. As illustrated in Fig. 6, both letters have one secondary part. However, the secondary part is above the letter "noon ن", while it is under the body in the letter "ba ب".

Centroid is used to describe objects after segmentation. We suggest the use of the centroid of the character's body to determine the position of the secondary parts according to the centroid point, where the centroid is determined by computing: {Mean (X), Mean (Y)}, where X and Y are the coordinates of the character's body.

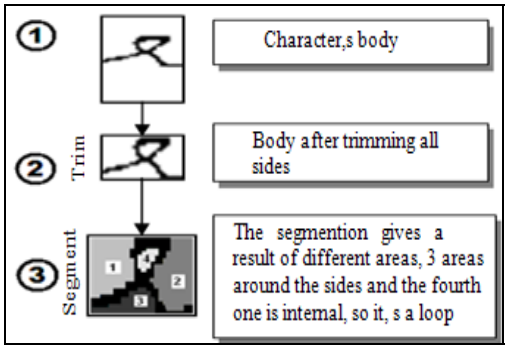

Fig. 8: Loop Detection for letter "Ain in the middle of the word"

Loop detection: Many algorithms are used for detecting shapes, curves and motions in the field of image processing and computer vision such as hough transform, however hough transform has several shortcomings, including high computational cost, low detection accuracy and possibility of missing objects (Surhone *et al.*, 2010). Moreover, Loops in Arabic handwritten letters aren't necessarily exact circles or ovals it may has some irregular shape as shown in Fig. 7.

In this study a model for Loop detection is suggested as illustrated in Fig. 8.

Body similarity: If the letters have the same features (secondary parts and their position and loop existence), then we have to distinguish between the letters according to the shape. There are many comparison algorithms that can be applied at this stage like finding the least mean square error, peak signal to noise ratio, or comparing according to the energy.

The comparison is done by using the radon transform (The Radon transform is the projection of the image intensity along a radial line oriented at a specific angle) for angle = "45".

Proposed algorithm: The proposed classifier work as follows:

- Divide the letters into regions in order to determine the body and secondary parts
- Classify the letter whether it has a secondary or not, so if it doesn't have secondary, try to classify according to the loops or shapes
- If it has secondary so, try to find how many they are
- If the number of secondary parts = 1 then it could be one of the following cases:
    - Hamza (ئ ،أ ،ك), so you have to classify according to the shape

  - Dot (noon ن, fa ف, ghain غ, thah ظ, dad ض, kha خ , jeem ج, ba ب), so you have to classify according to the position of the secondary whether it is upper (up) or lower (down) the body of the letter, then check for loops and finally check according to body similarity
- If the number of secondary parts = 2 dots (Qaf ق, Ta ت, Ta marbota ة, Ya ي), then classify according to the position of the secondary whether it is upper (up) or lower (down) the body of the letter. So if they are lower the body then it is the letter "ya", otherwise check for loops so if it has no loops then it is "Ta", otherwise check according to body similarity to determine it is "Qaf" or "Ta marbota:
- If the number of secondary parts = 3 dots (sheen ش, tha ث)
- If it has no secondary then check for loops and try to compare according to body similarity. (More details are exposed in Fig. 10)

RESULTS

The experiments were carried out in a collection of Arabic handwritten alphabets which collected from a large number of people to test the suggested model. The samples database was split into 2 subsets; one is used to build the comparison pools (25%) and the other for testing (the remaining 75%). Each sample contains the 28 Arabic characters, they can be written in different styles (isolated form, at the beginning, middlem and at the end of the word). So, character sets are expanded to approximately 100 classes. Figure 9 shows one of samples. Other experiments were executed with different pool and testing samples and comparable results were achieved. All the experiments are implemented using Matlab 7.0 software.

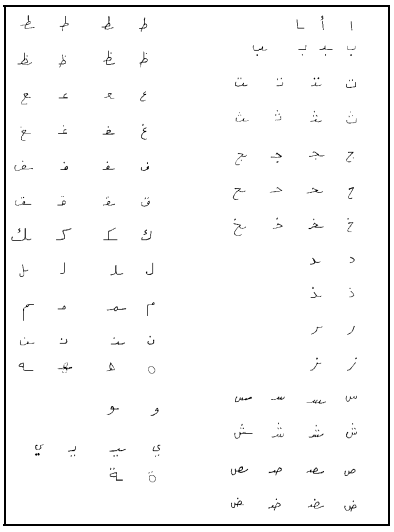

Fig. 9: Sample of scanned handwritten dataset

In order to analyze the effectiveness of the suggested features, each feature type is tested separately.

Therefore, the features of Loop detection, secondary positions and the number of the secondary parts shows excellent results around 99% detection, while the detection rates of the body similarity is as listed in the Table 1-8.

Table 1: Detection rates for characters without secondary parts

Letters	Isolated	Beginning	Middle	End
Alef	99%	98%	98%	98%
Ha	50%	50%	85%	50%
Dal	95%	92%	60%	50%
Ra	95%	92%	77%	77%
Seen	98%	50%	98%	50%
Ain	90%	50%	80%	98%
Lam	95%	98%	98%	98%
Hah	98%	98%	98%	98%

Table 2: Detection rates for characters with hamza above the body

Letters	Isolated	Beginning	Middle	End
Alef	98%	98%	98%	98%
Kaf	98%	95%	98%	98%

Table 3: Detection rates for characters with one dot above the body

Letters	Isolated	Beginning	Middle	End
Kha	50%	50%	85%	50%
Thal	90%	90%	70%	60%
Zain	94%	94%	85%	98%
Dad	95%	60%	70%	98%
Thah	98%	85%	98%	98%
Ghain	98%	85%	74%	98%
Fa	98%	95%	85%	50%
Noon	90%	85%	88%	80%

Table 4: Detection rates for characters with two dots above the body

Letters	Isolated	Beginning	Middle	End
Ta	99%	99%	99%	99%
Qaf	99%	90%	99%	99%

Table 5: Detection rates for characters with three dots above the body

Letters	Isolated	Beginning	Middle	End
Tha	99%	99%	99%	99%
Sheen	99%	90%	90%	92%

Table 6: Detection rates for characters with two dots under the body

Letters	Isolated	Beginning	Middle	End
Ya	99%	99%	99%	100%

Table 7: Detection Rates for Characters with One Dot under the Body

Letters	Isolated	Beginning	Middle	End
Ba	99%	99%	99%	99%
Geem	99%	66%	99%	99%

Table 8: Average recognition rates

Isolated	Beginning	Middle	End
93%	84%	89%	85%

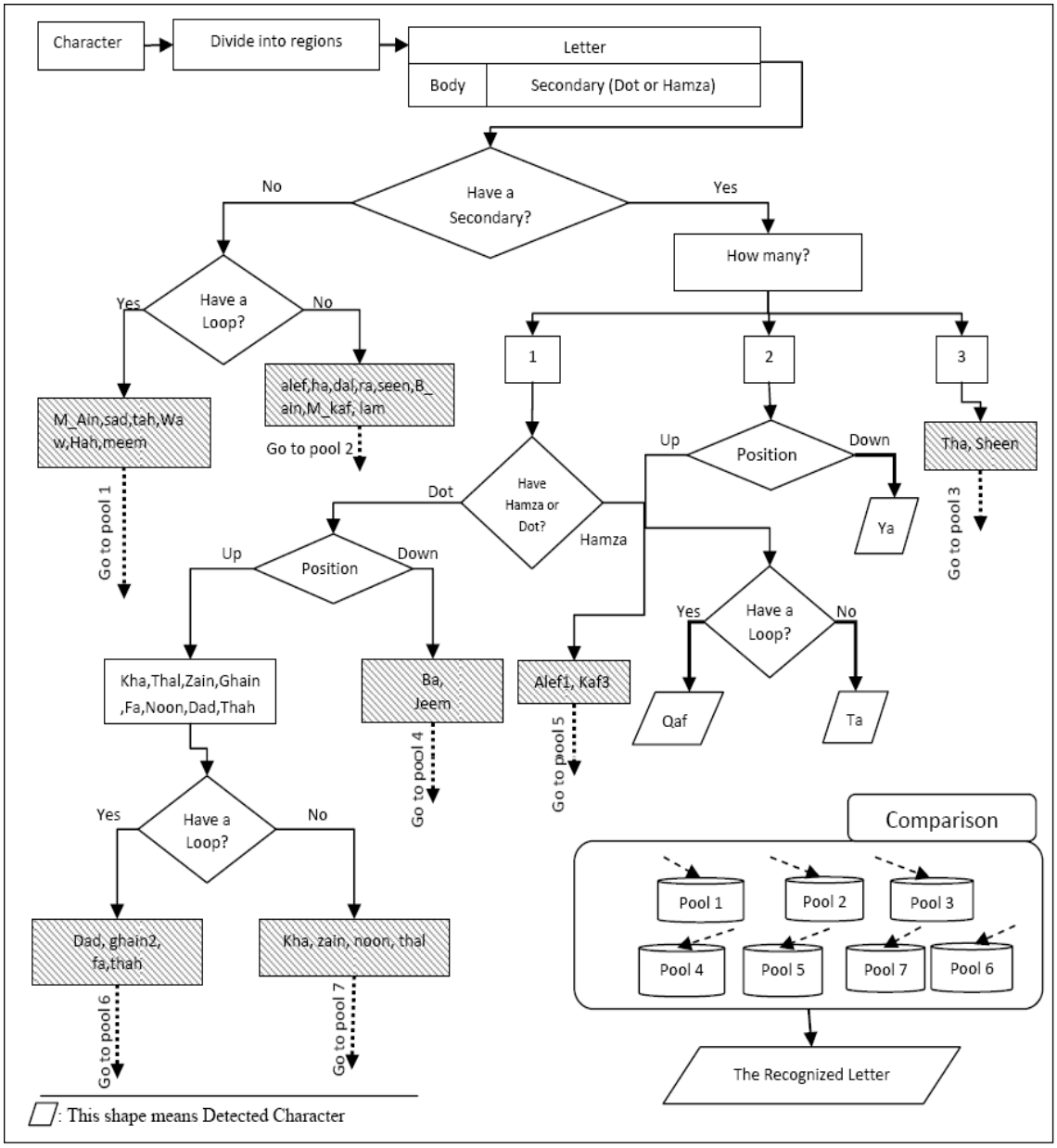

Fig. 10: Block diagram for the proposed model

DISCUSSION

Text mining is the most important research to extract text feature, but it is more difficult to extract text feature of offline handwritten Arabic text.

Abandah *et al*. (2009), present an approach for extraction text feature to achieve high recognition accuracy of handwritten Arabic letters. This approach exploits the classification potential of the secondary components of Arabic letters and overcomes some of their handwritten variations. This approach extracts moment features not only from the whole letter, but also from the main body and the secondary components. The results presented in this research show that better recognition accuracies are achieved when features are selected from the mixture of moment features (Abandah and Anssari, 2009).

Abandah *et al*. (2009), design a model for feature extraction from a large database of handwritten Arabic letters. The Arabic letters have multiple forms depending on the letter's position in the word. The

researchers dealing with processing of unconstrained handwritten Arabic cursive scripts must overcome many difficulties such as unlimited variation in human handwriting, similarities of distinct character shapes, character overlaps and interconnections of neighboring characters (Abandah *et al*., 2009). Mozaffari *et al*., 2009. This study describes the result of the ICDAR 2009 competition for handwritten Farsi/Arabic character recognition. To evaluate the submitted systems, author used large datasets containing both binary and grayscale images. Many different groups downloaded the training sets; however, finally 4 systems successfully participated in the competition. The systems were tested on two known databases and one unknown dataset. Due to the similarity between some digits and characters in Farsi and Arabic, each recognizer was tested for digit and character sets separately. For benchmarking, only the recognition rates, as the most important characteristic, are considered. Since participants used different software and even operating systems, the relative recognition speed is not compared in this competition. The competition results show a remarkable progress in Farsi/Arabic recognition systems. However, the obtained results are still less than English character recognition systems. System ECA was the winner of the digits recognition part with %95.9 recognition rate. In the character recognition competition, system MDLSTM reached %91.85 accuracy and outperforms the other systems (Mozaffari *et al*., 2009). Ahranjany *et al*., proposed a new method for recognizing arabic/farisi handwritten digits. They suggest the use of neural networks for automatic extraction of input pattern's features and fuse the results of boosted classifiers to compensate the recognizers' errors. The results of their experiments reveal a very high accuracy classifier. They achieved a recognition rate of 99.17% (Ahranjany *et al*., 2010).

CONCLUSION

The ultimate goal of designing a handwriting recognition system with an accuracy rate of 100% is impossible, because even human beings are not able to recognize every handwritten text without any doubt.

The recognition of Arabic handwritten Letters is hard not only because of the hand writing ambiguity but also because the similarity between letters according to their position in a word.

Classification stage introduces one of the most serious problems in the development of cursive script OCR system including Arabic language scripts. In order to overcome this problem, we use a set of features to be extracted. That is, letters are divided into body and secondary parts then a set of features were detected such as the number and type of secondary parts, the position of the secondary parts whether above or under the body of the character, the existence of loops and the Radon transform for the body part.

The set of sampled images, which contains large number of handwritten character samples are fed into the system to test. The samples present the different shapes of Arabic letters that is taken from different persons with different writing styles. Using the powerful set of extracted features, that is proved to be effective during the recognition of Arabic letters. The recognition rates reached {93, 84, 89 and 85%} for the Isolated letters, letters at the beginning, at the middle and at the end of the word respectively.

From testing results, it was concluded that the system had more trouble identifying letter Ha "ح". This is may be caused by the fact that this letter has a little bit similarity with letter Ain.

Apart from the previous shortcomings the overall system is proved to be stable and successful. In addition, this recognition-based model seems to be more suitable to other cursive scripts including Farisi handwriting.

REFERENCES

Abandah, G. and N. Anssari, 2009. Novel moment features extraction for recognizing handwritten arabic letters. J. Comput. Sci., 5: 226-232. DOI: 10.3844/jcssp.2009.226.232

Abandah, Gheith, Khedher and Mohammed, 2009. Analysis of handwritten arabic letters using selected feature extraction techniques. Int. J. Comput. Proc. Languages, 22: 1-25.

Ahranjany, S.S., F. Razzazi and M.H. Ghassemian, 2010. A very high accuracy handwritten character recognition system for Farsi/Arabic digits using Convolutional Neural Networks. Proceedings of the IEEE 5th International Conference on Bio-Inspired Computing: Theories and Applications, Sept. 23-26, IEEE Xplore, Changsha, pp: 1585-1592. DOI: 10.1109/BICTA.2010.5645265

Ergin, F.G., B.B. Watz and K. Erglis, 2010. Poor-contrast particle image processing in microscale mixing. Proceedings of the 10th Biennial Conference on Engineering Systems Design and Analysis, July 12-14, Istanbul, Turkey, pp: 649-653. DOI: 10.1115/ESDA2010-24900

Mahmoud, S.A. and S.M. Awaida, 2009. Recognition of off-line handwritten Arabic (Indian) numerals using multi-scale features and support vector machines vs. hidden Markov models. Arabian J. Sci. Eng., 34: 429-444.

Mozaffari, S. and H. Soltanizadeh, 2009. ICDAR 2009 handwritten Farsi/Arabic character recognition competition. Proceedings of the 10th International Conference on Document Analysis and Recognition, July 26-29, IEEE Xplore, Barcelona, pp: 1413-1417. DOI: 10.1109/ICDAR.2009.283

Slimane, F., R. Ingold, S. Kanoun, A.M. Alimi and J. Hennebert, 2010. Impact of character models choice on arabic text recognition Performance. Proceedings of the 12th International Conference on Frontiers in Handwriting Recognition, Nov. 16-18, Kolkata, India, pp: 670-675.

Surhone, L.M., M.T. Tennoe and S.F. Henssonow, 2010. Randomized Hough Transform. 1st Edn., VDM Verlag Dr. Mueller AG and Co. Kg, Germany, ISBN-10: 6134695823, pp: 92.

Szekely, G.J. and M.L. Rizzo, 2009. Brownian distance covariance. Ann. Applied Stat., 3: 1236-1265. DOI: 10.1214/09-AOAS312