A novel comprehensive database for Arabic and English off-line handwritten digits recognition

Ahmed Subhi Abdalkafor[1], Waleed Kareem Awad[2], Khattab M. Ali Alheeti[3]
[1]Career Development Center, University of Anbar, Iraq
[2,3]Department of Computer Networking Systems, University of Anbar, Iraq

Article Info

Article history:

Received Feb 10, 2020
Revised Apr 11, 2020
Accepted Apr 25, 2020

Keywords:

Arabic
Cropping
Database
Handwritten
Word arabic digits

ABSTRACT

The recognition of Arabic handwritten is received at the same interest as other Latin languages. In Optical Character Recognition (OCR), handwriting Arabic recognition is considered as one of the critical and difficult tasks in the various scientific area. The main issues of this matter were due to the lack of public Arabic handwriting databases and the cursive nature of Arabic writing. In this paper, a new benchmark database is built for the Arabic and English off-line handwritten digits Recognition. The original form is divided into three groups: Arabic digits, English digits, and word Arabic digits which written five times by 100 different academic staff and students of university writers. Our database contains 14500 images; divided into two subsets of training and testing to help researchers through evaluating and comparing results obtained from their systems.

Corresponding Author:

Khattab M. Ali Alheeti,
Department of Computer Networking Systems,
College of Computer Science and Information Technology,
University of Anbar, Iraq.
Email: co.khattab.alheeti@uoanbar.edu.iq

1. INTRODUCTION

Pattern recognition has become an attractive filed, due to the wide uses for many applications that humans use frequently in their daily lives. Handwriting recognition is one of the methods of computerization to imitate or understand human behaviors in recognizing ways to convert handwritten or printed text into automatically encrypted numbers [1-3].

Many researches have focused on the recognition of Chinese, Japanese and other languages. In contrast, A little researches focusing on Arabic texts due to several challenges in terms of grammar, writing style, lack of access to a database in terms of Arabic as opposed to and other factors that make it difficult to identify this language [4-7]. Recently, Arabic Digit handwritten has a very high level of interest, it entire in wide applications in a different area, such as verification of bank-check verification, entry data applications criminal evidence and all types of vehicles plates to register the violations The recognition of digits is one of the issues in pattern recognition so it is adopted as a measure to compare different classification algorithms [8-9]. The researchers focused on neural network algorithms for wide use in recognition, prediction, and others [10]. In the recognition filed they have been implemented by using different databases to evaluate the proposed works. [11-13].

There are few numbers of databases have been proposed especially in Arabic digits to evaluate the classification algorithms. This proposed work focuses on some previous studies such as Alayba and et al. have been explained construct models of Word2Vec from large Arabic corpus [14]. However, these models are obtained from ten newspaper in various Arabic area. Grother in [15] proposed the EMNIST database that merges the lowercase and uppercase into a single 26-class task of digits handwritten. On the other hand, LeCun [16] proposed the database named MNIST for English handwritten digits recognition all images are gray-scale and the size 28 by 28 pixels the overall of these images are 60000 images. R. Vijaya Kumar Reddy et al. [17] proposed different techniques of neural network approach to identify the Handwritten Hindi characters. They assessed the execution utilizing Convolutional Neural Network (CNNs) with improvement strategies and Deep Feed Forward Neural Network (DFFNN). The study identifies the set of database images that were collected from different users then train and test these techniques or strategies on it [18]. H. Alamri et al [19] proposed a new inclusive database for identifying Arabic handwriting. The submitted database looks at the first database that hiding place a different set of the Arabic language. It includes dates, isolated letters, strings of numerical, isolated Indian numbers, a set of (70) significant words and a set of distinct symbols. In 2011, I. Ahmad et al. [20] proposed a new comprehensive database of Arabic handwritten text (AHTD). This developed database consists of text written by (1000) writers from different countries. The submitted database includes images of the written text with different accuracy.

The basic database contains metadata describing the text written on the page, paragraph, and line levels [20-22]. This paper is structured as follows: The next section presents the characteristic of Arabic and English Digits. Then, the proposed system that involves four phases is presented in Section 3. Finally, Section 4 presents the conclusion of the study.

2. CHARACTERISTIC OF ARABIC AND ENGLISH DIGITS

Arabic digits consist of 10 digits and written from right to left. Each digit has various shapes also in the word Arabic digits depending on the position of the character in the word and the writer himself. Conversely, the digits of the English language were written from left to right and not consists of complexity as the Arabic language [23].

3. METHODOLOGY

The evaluation process of any classification system had been heavily based on the database. As we know, we have a problem with type and number of databases whether Arabic or English that utilized to test identification systems in research area. In addition, there are few numbers of databases have been suggested especially in Arabic digits to test the classification algorithms. In this research, a new benchmark is generated for the Arabic and English off-line handwritten digits recognition. In more detail, the original form is separated into main three sets: Arabic digits, English digits, and word Arabic digits.

The methodology of the new benchmark is comprised of four phases which are acquisition images from the various volunteers, data extraction phase, pre-processing phase, and put the image into folders. All these phases are explained in the Figure 1. The lifecycle of the proposed system is shown in Figure 1 that explained the generation process from the collected image to the creation process of the new benchmark. Thus, all these phases are considered a very important part of this paper; all of them will explain in detail each follows the subsection of this paper.

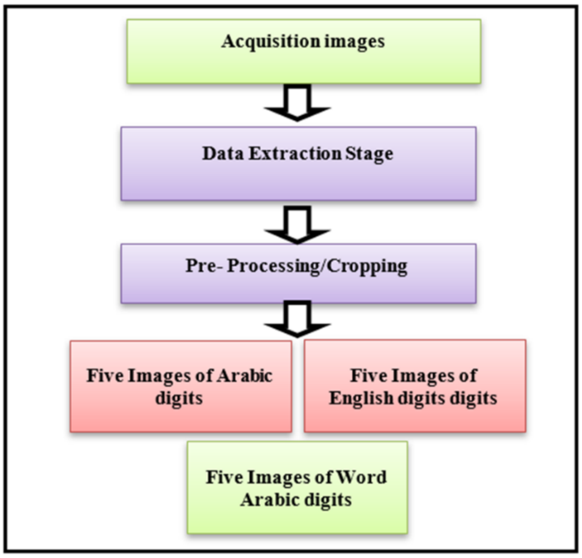
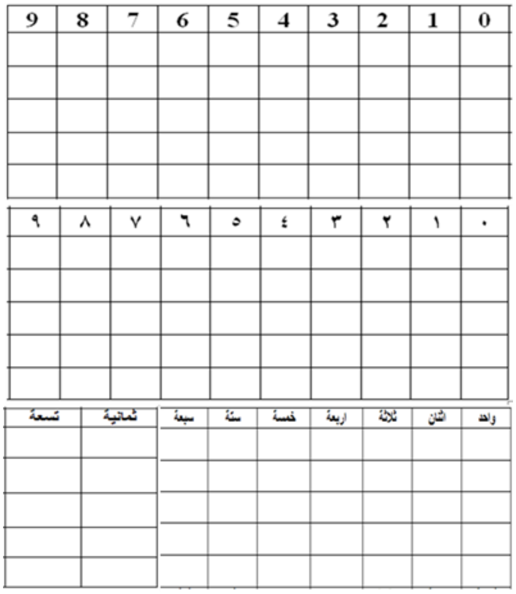

Acquisition images
Data Extraction Stage
Pre- Processing/Cropping
Five Images of Arabic digits
Five Images of English digits digits
Five Images of Word Arabic digits

Figure 1. The block diagrams

3.1. Data acquisition

The basic step to building any database is to find the appropriate data source [19]. The original form as shown in Figure 2 consists of three digits groups including (Arabic digits, English digits, and word Arabic digits) where each digit has been printed and the writers have been asked to write each digit in five empty boxes. The proposed database is collected from 100 writers, the average age range between 10 to 40 years from different levels of education that including the academic staff, employees, students of university and others. Little writers do not enter the schools of education, this status provides the proposed database with a diversity of shapes and patterns for each digit.



9	8	7	6	5	4	3	2	1	0

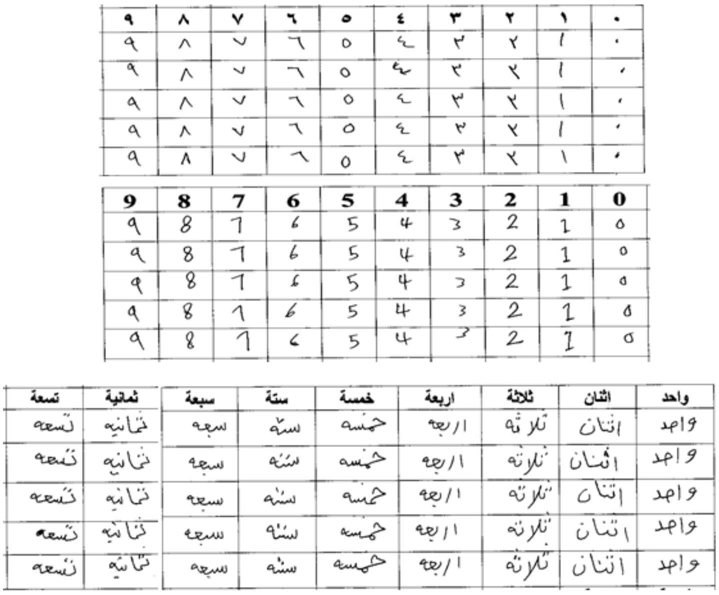
٩	٨	٧	٦	٥	٤	٣	٢	١	٠

تسعة	ثمانية	سبعة	ستة	خمسة	اربعة	ثلاثة	اثنان	واحد

Figure 2. Original form before scanning

3.1.1. Form scanning

The original form was scanned using a high-resolution scanner. The scanner scans 300 pages using 400 dpi and the output is PNG format of the scanned process. Figure 3 illustrated the original form after the scanning method.



٩	٨	٧	٦	٥	٤	٣	٢	١	٠
٩	٨	٧	٦	٥	٤	٣	٢	١	٠
٩	٨	٧	٦	٥	٤	٣	٢	١	٠
٩	٨	٧	٦	٥	٤	٣	٢	١	٠
٩	٨	٧	٦	٥	٤	٣	٢	١	٠
٩	٨	٧	٦	٥	٤	٣	٢	١	٠

9	8	7	6	5	4	3	2	1	0
9	8	7	6	5	4	3	2	1	0
9	8	7	6	5	4	3	2	1	0
9	8	7	6	5	4	3	2	1	0
9	8	7	6	5	4	3	2	1	0
9	8	7	6	5	4	3	2	1	0

تسعة	ثمانية	سبعة	ستة	خمسة	اربعة	ثلاثة	اثنان	واحد
تسعه	ثمانيه	سبعه	سته	خمسه	اربعه	ثلاثه	اثنان	واحد
تسعه	ثمانيه	سبعه	سته	خمسه	اربعه	ثلاثه	اثنان	واحد
تسعه	ثمانيه	سبعه	سته	خمسه	اربعه	ثلاثه	اثنان	واحد
تسعه	ثمانيه	سبعه	سته	خمسه	اربعه	ثلاثه	اثنان	واحد
تسعه	ثمانيه	سبعه	سته	خمسه	اربعه	ثلاثه	اثنان	واحد

Figure 3. Form after scanning

3.2. Data extraction phase

In this study, we applied a software of Adobe Photoshop program to extract each image of digits from the five boxes, Figure 4 illustrated this stage. After this phase, our database has saved in three main an isolated folder and each folder contains 10 subfolders, each subfolder contains five images to produce 14500 digits of images.

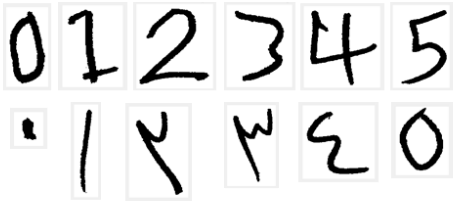


Figure 4. Extraction images

3.3. Pre-processing stage

This process is considered a necessary method for any recognition system. One of the disadvantages resulting from the scanning process is the presence of spaces around the body of an image thus will generate a database of high-storage space and this will be affected the functioning of the system in terms of time. Therefore, a cropping manner was applied to delete all the increases and spaces around the body of the image [23-25]. This manner led to getting a low-storage database and reduce the time during the processing of the system. Thus, improving the performance of the recognition system. Figure 5 showed a universe of discourse manner.

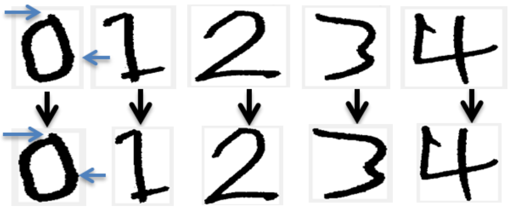


Figure 5. Cropping method

After completely these methods, these images put in three isolated folders. Firstly, the folder contains five images of Arabic digits and secondly folder five English digits and Arabic word digits in the third folders all of them were written by 100 writers and the total is 15000 images. The statistics of the proposed database are illustrated in Table1.

Table 1. Statistics of the proposed database

Variables	Value
Number of Writers	100
First Folder (Arabic digits)	500
Second Folder (English digits)	500
Third Folder (word Arabic digits)	450
Bit Depth	24
Number of images	14500
Training Sets	10150
Testing Set	4350
Type of image	(.PNG)

4. CONCLUSION

In this paper, a new database is presented for Arabic handwritten recognition and English digits for any recognition systems. It collected from 100 writers filled an original form that has five boxes. However, more than 14500 images are extracted and implemented some suitable methods via cropping method for reducing some images of noise that may exist after the scanning process. A Novel Comprehensive Database is divided into 70% for training and 30% for testing.

REFERENCES

[1] L. M. Lorigo & Govindaraju, V., "Offline Arabic handwriting recognition: a survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 5, pp. 712-724, 2006.

[2] A.S, Abdalkafor, and A. Sadeq. "Arabic Offline Handwritten Isolated Character Recognition System Using Neural Network," *International Journal of Business and ICT*, vol. 2, no. 3, pp. 41-50, 2016.

[3] W. Al-Ani1, et al., "An overview of wireless sensor network and its applications," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS*), vol. 17, no. 3, pp. 1480-1486, 2013.

[4] T. S. Gunawan, et al., "M. Development of English Handwritten Recognition Using Deep Neural Network," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 10, no. 2, 562-568, 2018.

[5] J. Sadri, Ching Y. Suen, and Tien D. Bui, "Application of support vector machines for recognition of handwritten Arabic/Persian digits," *In Proceedings of Second Iranian Conference on Machine Vision and Image Processing*, vol. 1, pp. 300-307. 2003.

[6] Assegie, T. A., et al., "Handwritten digits recognition with decision tree classification: a machine learning approach," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 5, pp. 4446-4451, 2019.

[7] Jihad, A. A., & Abdalkafor, A. S., "A Framework for Sentiment Analysis in Arabic Text," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 16, no. 3, pp. 1482-1489, 2019.

[8] A. S. Abdalkafor, "Survey for Databases On Arabic Off-line Handwritten Characters Recognition System," *1st International Conference on Computer Applications & Information Security (ICCAIS)*, pp. 1-6, 2018.

[9] A. S. Abdalkafor, "DIFRS-Database for Fingerprint Recognition System Using Ink-On-Paper Technique," *Journal of Engineering and Applied Sciences*, vol. 13, no. 17, pp. 7401- 7407, 2018.

[10] A. S. Abdalkafor, N. M. Aiman, O. N. Mustafa, "Predicting The Success Rates Of Schools Using Artificial Neural Network," *Journal of Theoretical & Applied Information Technology*, vol. 96, no. 19, pp. 6339- 6348, 2018.

[11] B. Tu, J. Wang, X. Kang, G. Zhang, X. Ou and L. Guo, "KNN-Based Representation of Superpixels for Hyperspectral Image Classification," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 11, pp. 4032-4047, 2018.

[12] A. S. Abdalkafor, et al., "A Novel Method Based On Priority For Enhancement Round-Robin Scheduling Algorithm," *Journal of Theoretical and Applied Information Technology*, vol. 96, no. 13, pp. 4092-4102, 2018.

[13] S. Routray, A. K. Ray, C. Mishra, & G. Palai, "Efficient hybrid image denosing scheme based on SVM classification," *Optik*, vol. 157, pp. 503-511, 2018.

[14] A. M. Alayba, V. Palade, M. England and R. Iqbal, "Improving Sentiment Analysis in Arabic Using Word Representation," *2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*, pp. 13-18, 2018.

[15] P. Grother, K. Hanaoka, "NIST special database 19 handprinted forms and characters database," *National Institute of Standards and Technology*, 2016.

[16] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE,* vol. 86, no. 11, pp. 2278–2324, 1998.

[17] R. Vijaya Kumar Reddy, U. Ravi Babu, "Handwritten Hindi Character Recognition using Deep Learning Techniques," 2019.

[18] A. S. Abdalkafor, "Designing Offline Arabic Handwritten Isolated Character Recognition System Using Artificial Neural Network Approach," *International Journal of Technology*, vol. 3, pp. 528-538, 2017.

[19] H. Alamri. J. Sadri. C. Y. Suen. and N. Nobile, "A novel comprehensive database for Arabic off-line handwriting recognition," *In Proceedings of 11th International Conference on Frontiers in Handwriting Recognition, ICFHR*. vol. 8, pp. 664-669, 2008.

[20] A. Mahmoud. I. Ahmad. M. Alshayeb and G. Al-Khatib, "A Database for Offline Arabic Handwritten Text Recognition," *International Conference Image Analysis and Recognition. ICIAR 2011*, pp. 397-406, 2016.

[21] A. Iqbal and A. Zafar, "Offline Handwritten Quranic Text Recognition: A Research Perspective*," 2019 Amity International Conference on Artificial Intelligence (AICAI)*, pp. 125-128, 2019.

[22] H.A Al-Muhtaseb, S.A Mahmoud, and R.S Qahwaji, "Recognition of off-line printed Arabic text using hidden markov models," *Signal Process*, vol. 88, no. 12, pp. 2902-2912, 2008.

[23] A. S. Abdalkafor, E. T. Allawi, K. W. Al-Ani and A. M. Nassar, A Novel Database for Arabic Handwritten Recognition (NDAHR) System. *In 2019 IEEE 2nd International Conference on Computer Applications & Information Security (ICCAIS)*, pp. 1-6, 2019.

[24] J. H. Alkhateeb, "Off-Line Arabic Handwritten Isolated Character Recognition," *International Journal of Engineering Science and Technology. Technol*, vol. 7, pp. 251-257, 2015.

[25] H. Xiangyu, C. Hutao, C. Pingyuan and L. Enjie, "Autonomous orbit determination for the probe around small body," *in Journal of Systems Engineering and Electronics*, vol. 15, no. 3, pp. 327-332, 2004.