Data Article

Arabic handwritten alphabets, words and paragraphs per user (AHAWP) dataset

Majid Ali Khan*

College of Computer Engineering and Science, Prince Mohammad Bin Fahd University, Khobar, Eastern Province, Saudi Arabia

A R T I C L E   I N F O

Article history:
Received 10 December 2021
Accepted 8 February 2022
Available online 13 February 2022

Keywords:
Handwritten Arabic alphabets
Handwritten Arabic words
Handwritten Arabic paragraphs
Writer identification
Arabic Text recognition

A B S T R A C T

This article presents a handwritten Arabic alphabets, words and paragraphs dataset (AHAWP). The dataset contains 65 different Arabic alphabets (with variations on begin, end, middle and regular alphabets), 10 different Arabic words (that encompass all Arabic alphabets) and 3 different paragraphs. The dataset was collected anonymously from 82 different users. Each user was asked to write each alphabet and word 10 times. A userid uniquely but anonymously identifies the writer of each alphabet, word and paragraph. In total, the dataset consists of 53199 alphabet images, 8144 words images and 241 paragraphs images. This dataset can be used for multiple purposes. It can be used for optical handwriting recognition of alphabets and words. It can also be used for writer identification (or verification) of handwritten Arabic text. It is also possible to evaluate difference in writing styles of isolated alphabets as compared to the same alphabet written as part of the word or in paragraph by the same user using this dataset. The dataset is publicly available at https://data.mendeley.com/datasets/2h76672znt/1.

© 2022 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

* Corresponding author.
E-mail address: makhan@pmu.edu.sa

https://doi.org/10.1016/j.dib.2022.107947
2352-3409/© 2022 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)
Specifications Table

| Subject                   | Computer Science                                      |
|---------------------------|-------------------------------------------------------|
| Specific subject area     | Image processing, Optical Handwritten Text Recognition, Writer Identification |
| Type of data              | Image                                                 |
| How the data were acquired| Users completed the forms (based on fixed template) with their handwriting and these forms were then scanned |
| Data format               | Raw                                                   |
| Description of data collection| Data was collected anonymously in a classroom setting. A “userid” was used to uniquely but anonymously identify the writer of each alphabet, word and paragraph. The forms were color scanned at 300 dpi and handwritten Arabic alphabets, words and paragraphs were then cropped from these scanned forms. |
| Data source location      | College of Computer Engineering and Science, Prince Mohammad Bin Fahd University |
|                          | City/Town/Region: Khobar                             |
|                          | Country: Saudi Arabia                                 |
|                          | Latitude and longitude for collected samples/data: 26.14544181406805, 50.091155268800044 |
| Data accessibility        | Repository name: Mendeley Data                        |
|                          | Data identification number: 10.17632/2h76672znt.1     |
|                          | Direct URL to data: https://data.mendeley.com/datasets/2h76672znt/1 |

Value of the Data

- This data is useful for optical handwriting recognition of Arabic text [1,2]. It can also be used for writer identification/verification of handwritten Arabic text [3–5]. This dataset contains handwritten alphabets, words and paragraphs written by the same user which makes it possible to develop writer identification models trained on alphabets and evaluated on words or sentences [6]
- The data can be used by machine learning researchers/companies to develop models for recognizing handwritten text. It can also be used by researchers/companies to develop forensic models for identifying writer of certain handwritten text.
- The existing Arabic alphabet identification datasets (Hijja [1], AHCD [7] datasets) do not provide any user information. The existing writer identification of Arabic text datasets (IFN/ENIT [8], KHAITT [9], QUWI [10] datasets) only provide handwritten words or paragraphs and do not contain alphabets. This dataset fills in this gap.
- The existing Arabic datasets cannot be used for writer identification based on Arabic alphabets or to evaluate differences in writing styles of isolated alphabets as compared to the same alphabet written as part of the word or paragraph by the same user. This dataset can be used for this purpose as it captures the user information for each alphabet, word and paragraph.

1. Data Description

The dataset contains 65 different Arabic alphabets (with variations on begin, middle, end and regular alphabets), 10 different Arabic words and 3 different paragraphs handwritten by 82 writers. Each writer was asked to write each alphabet and word 10 times. A userid uniquely but anonymously identifies the writer of each alphabet, word and paragraph. In total, the dataset consists of 53199 alphabet images, 8144 words images and 241 paragraphs images.

The writers had fluent Arabic speaking and writing background. They were not advised to use any specific type or color of writing instrument. This resulted in a dataset with several variations in type and color of writing instruments. A single writer had typically used the same type and color of writing instrument to write the alphabets, words and paragraphs.
### Fig. 1. Arabic Alphabets with variations collected in the dataset.

| Group | Alphabet | Regular | Begin | Middle | End | No. of different forms |
|-------|----------|---------|-------|--------|-----|------------------------|
| 1     | Alif     | ا       | ا     | ا      | ا   | 2                      |
| 2     | Beh      | ب       | ب     | ب      | ب   | 4                      |
| 3     | Jeem     | ج       | ج     | ج      | ج   | 4                      |
| 4     | Dal      | د       | د     | د      | د   | 2                      |
| 5     | Raa      | ر       | ر     | ر      | ر   | 2                      |
| 6     | Seen     | س       | س     | س      | س   | 4                      |
| 7     | Sad      | ص      | ص     | ص      | ص   | 4                      |
| 8     | Tah      | ط       | ط     | ط      | ط   | 2                      |
| 9     | Ain      | ع       | ع     | ع      | ع   | 4                      |
| 10    | Feh      | ف       | ف     | ف      | ف   | 4                      |
| 11    | Qaf      | ق       | ق     | ق      | ق   | 4                      |
| 12    | Kaf      | ك       | ك     | ك      | ك   | 6                      |
| 13    | Lam      | ل       | ل     | ل      | ل   | 4                      |
| 14    | Meem     | م       | م     | م      | م   | 4                      |
| 15    | Noon     | ن       | ن     | ن      | ن   | 4                      |
| 16    | Heh      | ه، ه.   | ه، ه. | ه، ه.  | ه، ه. | 5                      |
| 17    | Waw      | و       | و     | و      | و   | 2                      |
| 18    | Yaa      | ي       | ي     | ي      | ي   | 4                      |
|       |          |         |       |        |     | Total: 65              |

**Fig. 2.** Arabic words collected in the dataset.

Fig. 1 shows the set of Arabic alphabets collected in the dataset. Please note that a single alphabet is collected to represent a similarly styled group of alphabets.

Dataset contains 10 different Arabic words (that encompass all Arabic alphabets) handwritten by 82 different users. Fig. 2 shows the set of Arabic words collected in the dataset.

Dataset contains 3 different paragraphs handwritten by 82 different users. Fig. 3 shows the fixed text paragraphs collected in the dataset.

Fig. 4 shows template of the form used to collect a sample of handwritten Arabic alphabets from a user (“user001”). Fig. 5 shows template of the form used to collect a sample of handwritten Arabic words from a user (“user003”). Fig. 6 shows template of the form used to collect a sample of handwritten Arabic fixed text paragraphs from a user (“user005”).

The handwritten forms from all the 82 users were color scanned at 300dpi resulting in an image resolution of 2480 × 3507 pixels. These scanned images are provided as raw data in the folder named “raw_dataset” in the public repository.
Fig. 3. Arabic paragraphs collected in the dataset.

Fig. 4. Arabic alphabets handwritten by a user.
| User: 003 |
|-----------|

|     | صخر | المستضعفين | الأبجدية | فسيفساكمهم |
|-----|------|--------------|----------|-------------|
| a |      |              |          |             |
| a |      |              |          |             |
| a |      |              |          |             |
| a |      |              |          |             |
| a |      |              |          |             |
| a |      |              |          |             |
| a |      |              |          |             |
| a |      |              |          |             |

**Fig. 5.** Arabic words handwritten by a user.

---

| User: 005 |
|-----------|

هذا مقال عن تغير المناخ. يمكن أن يكون سبب تغير المناخ في العالم بسبب الأنشطة المختلفة. عندما يحدث تغير المناخ، درجات الحرارة يمكن أن تزيد بشكل كبير. خلال القرن الماضي، أطلقت الأنشطة البشرية كميات كبيرة من ثاني أكسيد الكربون وغازات الدفيئة الأخرى في الغلاف الجوي.

**Fig. 6.** Arabic paragraph handwritten by a user.
2. Experimental Design, Materials and Methods

The dataset from scanned images was extracted using Python scripts. The scripts are provided with the dataset in the public repository. Following is a brief description of the scripts:

- 1a_alphabet_extractor_per_alphabet.py: This script extracts alphabets from the scanned JPEG images (as shown in Fig. 4) and organizes them in a folder structure with one folder per alphabet containing that alphabet written by all the users. Each file name has the format “userid_alphabetName_variationName_index” where index increases sequentially for each extracted alphabet from a single page.
- 1a_alphabet_extractor_per_user.py: This script extracts alphabets from the scanned JPEG images (as shown in Fig. 4) and organizes them in a folder structure with one folder per user containing all the alphabets written by that user. Each file name has the format “userid_alphabetName_variationName_index” where index increases sequentially for each extracted alphabet from a single page.
- 2a_alphabets_pre_processing.py: This script is used to pre-process the extracted alphabets. It crops the alphabets from the center of image (excluding 20 pixels on each side). This was done to remove any borders surrounding the extracted alphabets. The surrounding whitespace around written alphabets was then removed. The resultant image was converted to grayscale and scaled to a height of 128 pixels (keeping the aspect ratio intact). Please note that keeping aspect ratio is important so that handwriting does not get distorted.
- 1w_word_extractor_per_user.py: This script extracts words from the scanned JPEG images (as shown in Fig. 5) and organizes them in a folder structure with one folder per user containing all the words written by that user. Each file name has the format “userid_wordName_index” where index increases sequentially for each extracted word from a single page.
- 2w_words_pre_processing.py: This script is used to pre-process the extracted words. It crops the words from the center of image (excluding 5 pixels on each side). This was done to remove any borders surrounding the extracted words. The surrounding whitespace around written words was then removed. The resultant image was converted to grayscale and scaled to a height of 128 pixels (keeping the aspect ratio intact). Please note that keeping aspect ratio is important so that handwriting does not get distorted.
- 1p_paragraph_extractor_per_user.py: This script extracts paragraphs from the scanned JPEG images (as shown in Fig. 6) and organizes them in a folder structure with one folder per user containing all the paragraphs written by that user. Each file name has the format “userid_paragraphNumber_index”.

Ethics Statements

An informed consent was obtained from all participants that the collected dataset will be used for research purposes and can be made publicly available to research community. The consent is available on each data collection form page (in “raw_dataset” folder, under each data collection form). Moreover, the participant data has been fully anonymized.

Since the dataset was collected as part of classroom assignment of a course, it did not require any prior approval from the ethics committee.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
Acknowledgments

The author acknowledges the support provided by students of “Machine Learning” class at Prince Mohammad Bin Fahd University (PMU) to help in collecting this dataset. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] N. Altwajry, I. Al-Turaiki, Arabic handwriting recognition system using convolutional neural network, Neural Comput. Appl. 8 (2020).
[2] J.H. Alkhateeb, An effective deep learning approach for improving off-line arabic handwritten character recognition, Int. J. Softw. Eng. Comput. Syst. 6 (2021) 53–61.
[3] A. Rehman, S. Naz, M.I. Razzak, Writer identification using machine learning approaches: a comprehensive review, Multimedia Tools and Applications 78 (2019) 10889–10931.
[4] M.N. Abdi, M. Khemakhem, A model-based approach to offline text-independent Arabic writer identification and verification, Pattern Recognit. 48 (2015) 1890–1903.
[5] S. He, L. Schomaker, Fragnet: Writer identification using deep fragment networks, IEEE Trans. Inf. Forensics Secur. 15 (2020) 3013–3022.
[6] M. A. Khan, N. Mohammad, G. B. Brahim, A. Bashar, G. Latif, Writer Verification of Partially Damaged Handwritten Arabic Documents based on Individual Character Shapes, Peer Comput. Sci. (submitted).
[7] H.M. Najadat, A.A. Alshboul, A.F. Alabed, Arabic handwritten characters recognition using convolutional neural network, in: 2019 10th International Conference on Information and Communication Systems (ICICS), 2019, pp. 147–151, doi:10.1109/ICACS.2019.8809122.
[8] M. Pechwitz, S.S. Maddouri, V. M’argner, N. Ellouze, H. Amiri, Ifn/enit - database of handwritten Arabic words, in: Proc. of CIFED, 2002, 2002, pp. 129–136.
[9] S.A. Mahmoud, I. Ahmad, M. Alshayeb, W.G. Al-Khatib, M.T. Parvez, G.A. Fink, V. M’argner, H. El Abed, KHATT: Arabic offline handwritten text database, in: Proceedings - International Workshop on Frontiers in Handwriting Recognition, IWFR, 2012, pp. 449–454, doi:10.1109/ICFHR.2012.224.
[10] S.A. Maadeed, W. Ayoub, A. Hassaine, J.M. Aljaam, Quwi: An arabic and english handwriting dataset for offline writer identification, in: 2012 International Conference on Frontiers in Handwriting Recognition, 2012, pp. 746–751, doi:10.1109/ICFHR.2012.256.