# VML-HD: The historical Arabic documents dataset for recognition systems

**Conference Paper** · April 2017

DOI: 10.1109/ASAR.2017.8067751

CITATIONS
22

READS
97

**5 authors**, including:

Majeed Kassis
Ben-Gurion University of the Negev
**13** PUBLICATIONS **106** CITATIONS
SEE PROFILE

Ahmad Droby
Ben-Gurion University of the Negev
**19** PUBLICATIONS **57** CITATIONS
SEE PROFILE

Reem Alaasam
Ben-Gurion University of the Negev
**14** PUBLICATIONS **64** CITATIONS
SEE PROFILE

Jihad El-Sana
Ben-Gurion University of the Negev
**120** PUBLICATIONS **2,769** CITATIONS
SEE PROFILE

# VML-HD: The Historical Arabic Documents Dataset for Recognition Systems

Majeed Kassis
Department of Computer Science
Ben-Gurion University of the Negev
Beer-Sheva, Israel
majeek@cs.bgu.ac.il

Alaa Abdalhaleem,
Ahmad Droby, Reem Alaasam
Department of Computer Science
Ben-Gurion University of the Negev
Beer-Sheva, Israel
alaaabd,drobya,rym@post.bgu.ac.il

Jihad El-Sana
Department of Computer Science
Ben-Gurion University of the Negev
Beer-Sheva, Israel
el-sana@cs.bgu.ac.il

***Abstract*—In this paper we present a new database with handwritten Arabic script. It is based on five books written by different writers from the years 1088-1451. We took 680 pages from these five books, and fully annotated them on the sub-word level. For each page we manually applied bounding boxes on the different sub-words and annotated the sequence of characters. It consists of 121,636 sub-word appearances consisted of 244,553 characters out of a vocabulary of 1,731 forms of sub-words. The database is described in detail and is designed for training and testing recognition systems for handwritten Arabic sub-words. This database is available for the purpose of research, and we encourage researchers to develop and test new methods using our database.**

## I. Introduction

Word-spotting of handwritten Arabic script remains a challenging task even though the latest improvements of recognition methods and systems are very promising. Many advances has been made in the Neural Networks [1], [2], [3], [4], [5] area, which are able to consistently outperform other approaches in virtually all fields of computer vision, including word-spotting and word recognition. In traditional methods, the predominant features used in word spotting have been SIFT descriptors [6], [7], [8], [9], geometric features [10], and HOG-based descriptors [11], [12].

The most important aspect in developing and improving these recognition systems is the existence of properly annotated databases of large sizes. In comparison to the Arabic text, the Latin text has had handwritten datasets publicly available for a long time, where for Arabic, unfortunately, the availability of historical datasets is scarce, and mostly private. This lead to many of these papers in this area to use a specific, small dataset of their own or use their privately held databases that are not available to the public.

The most popular Arabic dataset is the IFN/ENIT-database [13]. It consists of town and villages names written by 411 writers, with 946 different names, totaling 26,459 handwritten names containing more than 210,000 characters. This database is largely used in contemporary research, for train and evaluation. It is a binary dataset, and was written by people recently (modern Arabic). This paper presents a gray-scale database of historically handwritten Arabic text which will provide other and new challenges for researchers.

This paper is intended to give a detailed insight on our new database and is arranged as follows: We begin with an overview of the database, then in depth information regarding the dataset and all its aspects, such as the ground truth information, the construction, and labeling. Next, we run several experiments on the dataset. Finally, we present some of our future plans then concluding remarks.

## II. Overview

The scarcity of Arabic handwritten databases have motivated us to provide the research community our own database for research purposes. As we all know, the quality of the recognition of these systems relies heavily on the quality of the training data. Many researchers train their algorithms today on vast amount of data of high variety, in order to improve the recognition rates.

We scanned more than 20 books written by hand by multiple writers in the years 1088-1451. The books where photographed using a high quality camera, namely Hasselblad H5D-60 Medium Format Digital SLR Camera from 1 meter distance. The books originate from the National Library in Jerusalem. They are stored in a high quality uncompressed TIFF format, where each image is roughly of size $6000 \times 4000$ pixels. From these books we chose five books, written by five writers. From each book we took on average 136 pages, 680 pages in total. We uploaded the images to the WebGT [14] ground truth system, which is a web-based system for ground truth generation and provides a user-friendly interface for quick annotation of degraded documents in general, and historical document images in particular. Using this system we marked the bounding of the sub-words found in each page. Then we annotated these sub-words with their corresponding sequence of characters. We've marked, in total, 121,636 sub-words of 1,731 different forms of sub-words. An example of annotated lines of a page can be seen in figure 1.

## III. Details

We aimed to generate a database of historical Arabic script which would contain high variety of sub-words, high variety of image quality, and a large number of writers. In this section,

978-1-5090-6628-5/17/$31.00 ©2017 IEEE

Fig. 1. An example of an annotated line done using Web-GT framework. We manually applied each bounding box around the sub-words present in the image, then we annotated each bounding box with its corresponding sequence of characters.

we provide detailed statistics of the dataset. Illustration of the variety in writing style can be seen in Figure 2

Complete dataset statistics can be seen in Table I, ranging from the books, pages, sub-words forms, appearances, to the number of characters in the dataset. The dataset, even though it is taken using a high quality camera, contains high variety of image quality. Many cases where the ink of the sub-word suffers from fading, yellowing, and even stains will provide a challenge for the recognition systems in terms of overcoming these issues. Please refer to Figure 3 for examples of fade, washed, and smeared text, and refer to Figure 4 well as stains, spots, and different coloring of the background.

| Type | Total |
|---|---|
| Books | 5 |
| Writing styles | 5 |
| Pages | 680 |
| Forms | 1,731 |
| Appearances | 121,636 |
| Total characters | 244,553 |

TABLE I

SOME STATISTICS OF THE DATASET RELATED TO ITS SIZE AND DISTRIBUTION. FORMS MEANS THE NUMBER OF IMAGES CONTAINING UNIQUE CHARACTER SEQUENCE. APPEARANCES ARE THE TOTAL NUMBER OF IMAGES FOUND IN THE DATASET.

For each page in the dataset, the WebGT website exports an XML file of Hadara [15] format. Each book has its own Hadara XML file which contains the coordinates of all the bounding boxes of all the images annotated of that book, as well as the sequence of characters for each bounding box applied in Arabic text. We also generated a Hadara XML file for each page in addition to the file for each book. This file contains the coordinates of the bounding boxes and the sequence of the characters for the specific page only. Finally, to alleviate any encoding issue, we also released encoded files of the same format. These files have for each letter written in Arabic script a corresponding Latin symbol which will denote such letter, and we believe all the possible options will ease the use of the dataset.

To generate the corresponding Latin symbol for each Arabic annotation, and since Latin script cursiveness is different than the Arabic script cursiveness, we add additional letters to denote their placement, and following the Arabic script connectivity rules the user can know whether they are connected to their next or previous letter. A typical letter in Arabic has up to four different forms, classified into isolated and contextual. A contextual form may be connected at the beginning of the sub-word, at the middle, or at its end. Due to this characteristic, the letters are encoded using four letters: "S" for isolated letters, "I" for initial, "M" for medial letters, and "F" for final letters. Next, for each Arabic letter, we've used a corresponding Latin encoding following the *Buckwalter* Arabic transliteration scheme which was developed at Xerox in the 1990s. It is an ASCII only transliteration scheme, representing Arabic orthography strictly one-to-one, unlike the more common romanization schemes that add morphological information not expressed in Arabic script. For an encoding example for an Arabic letters into its corresponding Latin script please refer to Table II. To annotate a sub-word we generate the corresponding Latin character prefixed as detailed before depending on its location, then the letters are separated by "|" symbol. For example عَلَيكُم will translate to "$F\_m \mid M\_k \mid M\_y \mid M\_l \mid F\_E$". This comes to allow the combination of sub-words into their corresponding words, which opens further research possibilities.

| Prefix | Form | Shape | Label |
|---|---|---|---|
| S_ | isolated | ع | S_E |
| I_ | initial | عـ | I_E |
| M_ | medial | ـعـ | M_E |
| F_ | final | ـع | F_E |

TABLE II

FOUR POSSIBLE FORMS FOR THE LETTER ع FROM LEFT TO RIGHT: ISOLATED FORM, INITIAL FORM, MEDIAL FORM, AND FINAL FORM. EACH LETTER IS PREFIXED WITH AN APPROPRIATE LETTER DENOTING IT'S LOCATION IN THE TEXT, AS SEEN IN THE LABEL COLUMN.

Finally, we generated the ground truth data for each sub-word found in the dataset. The ground truth data consists of the following fields: (1) Book number, (2) Page number, (3) Sub-word id, (4) Location coordinates, (5) Arabic annotation, (6) Latin annotation, (7) Sub-word length. Two examples of the ground truth gathered for sub-words can be seen in Table III.

| Image |  |  |
|---|---|---|
| Book number | 3158466 | 187370 |
| Page number | _006-2 | 0013-2 |
| Segment id | 183380 | 187370 |
| Location coordinates | y=1249 x=856<br>y=1249 x=945<br>y=1352 x=945<br>y=1352 x=856 | y=1212 x=649<br>y=1212 x=722<br>y=1306 x=722<br>y=1306 x=649 |
| Arabic annotation | فيها | خلا |
| Latin annotation | $F\_A \mid M\_y \mid M\_h \mid I\_f$ | $F\_A \mid M\_l \mid I\_x$ |
| | 4 | 3 |

TABLE III

TWO EXAMPLES OF GROUND TRUTH INFORMATION FOR EACH SUB-WORD: BOOK NUMBER, PAGE NUMBER, SEGMENT ID, LOCATION COORDINATES, ARABIC ANNOTATION, LATIN ANNOTATION, AND SUB-WORD LENGTH

f

Fig. 2. Illustration of the variety of the writing styles found in the dataset. About 136 pages are taken from each book written by a different writer resulting in high diversity in writing styles.

Fig. 3. Smeared, faded, and washed ink examples. These cases appear throughout the dataset and provide new recognition challenges.

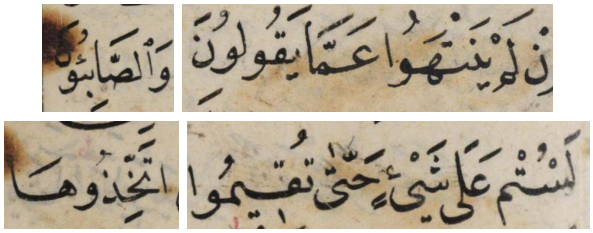

Fig. 4. Examples of stained text with variation in background color. These cases appear throughout the dataset and provide new challenges for the researchers in contrast to the binary dataset researchers currently use which provides white backgrounds.

| Book number | Top1 | Top2 | Top3 | Top4 | Top5 |
|---|---|---|---|---|---|
| 3157556 | 81.27% | 88.10% | 91.75% | 92.86% | 93.33% |
| 3158466 | 82.06% | 88.73% | 91.75% | 93.65% | 94.28% |
| 3249138 | 93.65% | 96.35% | 96.67% | 97.62% | 97.94% |
| 3368132 | 86.35% | 92.86% | 95.40% | 96.03% | 96.03% |
| 3426930 | 77.30% | 87.14% | 90.16% | 91.75% | 93.17% |

TABLE IV
HIT RATE RESULTS OF OUR METHOD ON EACH SUB-SET TAKEN FROM EACH BOOK. 2,100 SUB-WORDS PER SUB-SET.

| Method | Top1 | Top2 | Top3 | Top4 | Top5 |
|---|---|---|---|---|---|
| Radial Descriptor Graph | 83.40% | 89.84% | 92.48% | 94.00% | 95.11% |
| Radial Descriptor | 68.15% | 78.44% | 84.71% | 88.50% | 90.15% |

TABLE V
HIT RATE RESULTS OF OUR METHOD ON THE COMBINED SUB-SETS WRITTEN BY FIVE WRITERS. 10,500 SUB-WORDS IN TOTAL.

## IV. EXPERIMENTS

We've ran a couple of experiments on the new dataset, using the Radial Descriptor [16] as well as the Radial Descriptor Graph [17]. First, we ran the algorithm on a sub-set of the dataset that contains 21 kinds of sub-words of 100 appearances each, totaling $2,100$ sub-words, for each books. The results can be seen in Table IV. The second test is done on a combination of the five datasets used in the first test, totaling 10,500 sub-word images. The results can be seen in Table V.

Of course, we invite fellow researchers to test and develop new methods using the complete dataset, for all forms that has at least two appearances. The dataset has 1,651 forms of two or more appearances, totaling $121,556$ images. Using the mean average precision (mAP), one can provide clear image regarding the performance of their method. Mean average precision is a widespread measure for the performance of information retrieval systems. The metric is defined as the average of the precision value obtained after each relevant sub-word is retrieved. To access the dataset please refer to our [18], [19] websites.

## V. SUMMARY AND FUTURE WORK

In this paper we have introduced our historical Arabic database as a new database for word-spotting and recognition. The database is publicly available for the purpose of research. 680 pages taken from 5 books written by different writing using different writing styles, were annotated on the sub-word level. 121,636 sub-words, containing 244,553 characters, extracted from these annotated pages and their ground truth added, both in Arabic script as well as in Latin script. With this database it is possible to develop and test Arabic handwritten recognition systems for the first time on historical documents in gray-scale. Until today there was no historical Arabic database available for public use and no gray-scale one either.

We invite researches to test their current systems and develop new systems using our database.

## ACKNOWLEDGMENT

This research was supported in part by the Lynn and William Frankel Center for Computer Sciences at Ben-Gurion University, Israel, and we'd like to thank them for their support. We would also like to thank *Rami Amsis* and *Ghassan Nabary* for their help in creating the ground truth information for the dataset. We would like also to thank *Hussien Othman* for proofing and finalizing the ground truth information.

## REFERENCES

[1] S. Sudholt and G. A. Fink, "PHOCNet: A Deep Convolutional Neural Network for Word Spotting in Handwritten Documents," in *Frontiers in Handwriting Recognition (ICFHR), 2016 17th International Conference on*, 2016.

[2] A. Sharma *et al.*, "Adapting off-the-shelf cnns for word spotting & recognition," in *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pp. 986–990, IEEE, 2015.

[3] V. Frinken, A. Fischer, R. Manmatha, and H. Bunke, "A novel word spotting method based on recurrent neural networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 2, pp. 211–224, 2012.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European Conference on Computer Vision*, pp. 346–361, Springer, 2014.

[6] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Word spotting and recognition with embedded attributes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 12, pp. 2552–2566, 2014.

[7] M. Rusiñol, D. Aldavert, R. Toledo, and J. Lladós, "Efficient segmentation-free keyword spotting in historical document collections," *Pattern Recognition*, vol. 48, no. 2, pp. 545–555, 2015.

[8] D. Aldavert, M. Rusinol, R. Toledo, and J. Lladós, "Integrating visual and textual cues for query-by-string word spotting," in *2013 12th International Conference on Document Analysis and Recognition*, pp. 511–515, IEEE, 2013.

[9] L. Rothacker and G. A. Fink, "Segmentation-free query-by-string word spotting with bag-of-features hmms," in *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pp. 661–665, IEEE, 2015.

[10] T. M. Rath and R. Manmatha, "Word spotting for historical documents," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 9, no. 2-4, pp. 139–152, 2007.

[11] I. Rabaev, K. Kedem, and J. El-Sana, "Keyword retrieval using scale-space pyramid," in *Document Analysis Systems (DAS), 2016 12th IAPR Workshop on*, pp. 144–149, IEEE, 2016.

[12] J. Almazán, A. Fornés, and E. Valveny, "Deformable hog-based shape descriptor," in *2013 12th International Conference on Document Analysis and Recognition*, pp. 1022–1026, IEEE, 2013.

[13] M. Pechwitz, S. S. Maddouri, V. Märgner, N. Ellouze, H. Amiri, *et al.*, "Ifn/enit-database of handwritten arabic words," in *Proc. of CIFED*, vol. 2, pp. 127–136, Citeseer, 2002.

[14] O. Biller, A. Asi, K. Kedem, and I. hak Dinstein, "Webgt: An interactive web-based system for historical document ground truth generation.," in *ICDAR*, pp. 305–308, 2013.

[15] W. Pantke, V. Märgner, D. Fecker, T. Fingscheidt, A. Asi, O. Biller, J. El-Sana, R. Saabni, and M. Yehia, "Hadara–a software system for semi-automatic processing of historical handwritten arabic documents," in *Archiving Conference*, vol. 2013, pp. 161–166, Society for Imaging Science and Technology, 2013.

[16] M. Kassis and J. El-Sana, "Word spotting using radial descriptor," in *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pp. 387–392, IEEE, 2014.

[17] M. Kassis and J. El-Sana, "Word spotting using radial descriptor graph," in *Frontiers in Handwriting Recognition (ICFHR), 2014 16th International Conference on*, IEEE, 2016.

[18] "Visual media laboratory - arabic historical documents dataset." http://www.cs.bgu.ac.il/~vml, 2016. [Online; accessed 2016].

[19] M. Kassis, "The VML Arabic Historical Documents Dataset for Recognition Systems." http://www.cs.bgu.ac.il/~majeek, 2016. [Online; accessed 2016].