

# Towards building standard datasets for Arabic recognition

Mohamed E. M. Musa

**Abstract**— Current Machine learning and pattern recognition method needs big dataset to produce efficient and accurate recognizers. The lack of standard big Arabic dataset is one of the big challenges that faces the research in this area. This paper presents the Arabic dataset collected and annotated by SUST-ALT (Sudan University of Science and Technology-Arabic Language Technology group) to contribute in filling this gap. The datasets contain: numerals datasets, isolated Arabic letters datasets, Arabic names datasets. These datasets contain offline dataset as well as online ones. The paper also describes some published results as well as future work.

**Index Terms**—Arabic language recognition, dataset, Machine learning, Pattern recognition.

## I. INTRODUCTION

Good effort has been done so far towards building efficient automatic reading systems that enable us to get rid of the keyboard as the main entrance to computers. However, we are still far away from saying we have robust solutions for this problem. All the available OCR technologies work in restricted environments. To be sure that a human being person is entering data not an electronic agent, web-based data entering systems use rough and mixed-up writing, which should be read and reentered by that person. This is a strong evident that electronic reading systems are far from competing human reading faculty. This is the case for Latin scripts, for Arabic language we are lagging behind by at least one decade. The conclusion of this says we need to keep doing research in this area for all languages. However, for languages like Arabic we should increase and intensify our research work.

This paper describes the data set collected and preprocessed by SUST ALT (Sudan University of Science and Technology- Arabic Language Technology group)

Section II, gives some basic information about Arabic language and its alphabet. Section III, contains four subsections. Each subsection describes one dataset. Section IV, contains discussion and as usual section V, conclude the paper.

## II. ARABIC LANGUAGE

Arabic is the official language of more than twenty countries and the mother tongue of more than 300 million people [1]. Arabic is one of the six United Nations official

languages. Unlike Latin Arabic is written from right to left. The Arabic script is also used as a medium of writing for other languages like Persian. Moreover, Arabic script is the former script of the Turkish language as the Arabic script was the script of the Ottoman Empire. The Ottoman Empire produced millions of written documents. Although, most of these documents archived in Turkey, there is considerable part of them distributed around several other countries. Digitizing these documents and building efficient retrieval system for them is still a big research challenge [2].

Arabic script is cursive. However, out of 28 letters there are 6 non-cursive letters. These 6 letters cannot be connected to the succeeding letters. Thus an Arabic word may be decomposed into two sub words or more. Each Arabic letter may have up to 4 different shapes, TABLE I, shows some Arabic letters with their different shapes.

Nowadays, there are three categories for Arabic language [3]:

*Classical Arabic*: the ancient language and the language of Quran.

*Modern standard Arabic*: the universal language of Arabic-speaking world, which understood by all speakers and used by the media and the academic and official communities.

*Local Arabic dialects*: there are many local dialects which contain many words and constructs understood by the local people only.

TABLE I  
SOME ARABIC LETTER AND THEIR DIFFERENT WRITING SHALES

Name	Isolated	Initial	Medial	Final
Alif	ا			آ
Baa	ب	بـ	با	باء
jiim	ج	جـ	جا	جاء
Shin	ش	شـ	شا	شاء
Dal	د			دا
Thal	ذ			ذا
Ghayn	غ	غـ	غا	گاه
Qaaf	ق	قـ	قا	قاء
Kaf	ك	كـ	كا	كاف
mym	م	مـ	ما	ماء











### III. SUST ALT DATASETS

It may be a good idea to have huge annotated dataset that represent all types of writing to train and test new recognition systems. However, such dataset will not be useful for small scale research project. For this reason we have decided to build many datasets to be used of studying, investigating, training, and testing small proposal as well as big ones. The rest of this section outlines these dataset. All these dataset are available freely for researchers in [www.sustech.edu](http://www.sustech.edu).

#### A. Hindi numeral dataset

In first stages of OCR work for a specific language we usually try to recognize that language digits. Although, now many Hindi digit dataset exist, we build our own digit dataset to learn some lessons from doing this. However, it is also important for us to have dataset from our environment as it could be different from other environments. TABLE II, shows the basic shape for each digits and it frequencies in the dataset.








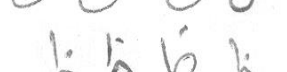
TABLE II  
HINDI NUMERAL DATASET SAMPLES AND SIZES

Digit	Hindi	samples	frequency
0	०		3680
1	१		3674
2	२		3678
3	३		3703
4	४		3700
5	५		3706
6	६		3688
7	७		3699
8	८		3702
9	९		3709

#### B. Isolated Arabic letters dataset

As illustrated in section II, Arabic letters has many shapes. However, it is useful to have a dataset for isolated letters as some applications use isolated letters. In addition, this data set will help in primary investigation of the language.

TABLE III  
SOME SAMPLES FORM THE LETTERS DATASET

Letter name	Typed form	samples
Alif	ا	
Baa	ب	
Taa	ت	
Thaa	ث	
Jiim	ج	
Thal	د	
Shin	ش	
Tha	ظ	

#### C. Arabic names dataset

The source of this dataset is the SUST graduation certificate application form, see Fig. 2. The student should write his name up to the third grandfather in this form. These documents had been collected from the SUST registrar office, scanned and segmented. Fig. 1, contains samples for the name Mohamed "محمد" in this dataset

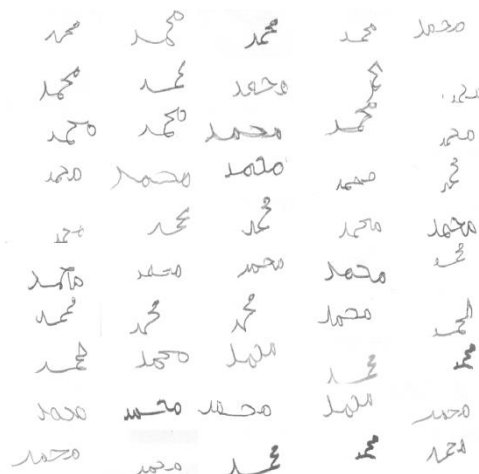


Fig. 1: Sample for the name Mohamed "محمد" in names dataset

#### D. Online Arabic dataset

Two online Arabic datasets has been established (SUSTOLAH – Sudan University of Science and Technology OnLine Arabic Handwritten data). The first one of these datasets (i.e., the dataset of the letters) contains 7827 samples of online handwriting for isolated Arabic letters. The second data (i.e., the dataset of the persons' names) contains 3097 samples of online Arabic handwriting for person's names. More than one hundred and fifty writers—from various high educational institution in Sudan— are contributed in the collection of these samples. A list of twenty person's names is specified for the collection of the samples for the dataset of the names. On the other side, the basic Arabic letters have been appointed for the collection of the samples for the dataset of the letters. Table 1 shows part of the name dataset, the table contains 8 rows for eight different Arabic names with the variation for their.

In comparison with the ADAB — an online Arabic dataset which known as a standard benchmark in the ICDAR competition of 2009 [7] —, SUSTOLAH has the following characteristics:

- SUSTOLAH involve handwritten objects of isolated Arabic letters as well as handwritten objects of cursive Arabic words. This property assigns a pedagogical research importance to SUSTOLAH.
- The datasets have an aspectual representation for the pen tips and strokes which formulate the handwriting.
- SUSTOLAH has a software tool for the collection of online Arabic handwriting as well as a verification tool. Therefore, researchers are able to create their own datasets by these tools.

#### IV. DISCUSSION AND SOME PUBLISHED RESULTS

These datasets are newly established waiting for extensive machine learning and pattern recognition research work. However, some research work has been published using them. Jadeed et al., has designed and tested a Support Vector Machine classifier for the digits dataset[4]. The accuracy of this classifier is 89% . The digits which has bad results are: zero, because it resemble the noise; two and three because they are very similar in their shape. Balola et al, has designed and tested a multi-layer perception for the isolated letters classification [5,8]. The main result of his work shows that the feature that causes the main challenge for Arabic letters classification is the usage of dots to differentiate similar letters like dal and thal ("ذ" & "ث"). Ali et al. have designed and tested a holistic classifier based on probabilistic neural network to classify Arabic names. Their experiments show that with high rejection rate the recognition rate of this classifier is very high too [6]. The online dataset is most

newest one, however, some interesting results for these data set is published in [9,10]

#### V. CONCLUSION

SUST ALT is a research group that work in Arabic language technology. One of its current interest of is the Arabic handwriting recognition. This paper describes five datasets established by this group to support the research work Machine learning and pattern recognition generally and Arabic recognition especially. The data contains: numerals, letters and names datasets.

#### REFERENCES

- [1] B. Al-Badr, S. Mahmoud "Survey and bibliography of Arabic Text recognition," Signal Processing , vol. 41, page 49-77, 1995
- [2] I. Z. Yalniz, I. S. Altıngöve, U. Gündükbay, and Ö. Ulusoy, "Ottoman Archives Explorer: A Retrieval System for Digital Ottoman archives," ACM Journal of Computing and Cultural Heritage, vol. 2, No. 3, December 2009
- [3] M. Chareit , "Visual recognition of Arabic handwriting: challenges and directions," SACH2006, LNCS 4768, pp. 1-21 2008
- [4] A. M. Gadeed and M. E. M. Musa "Handwritten Hindi Numeral s recognition using Support Vector machine" The International Arab Conference on Information Technology - Benghazi, Libya December 2010.
- [5] O. Balola and M. E. M. Musa "Multi Stages Neural Networks for Isolated Arabic Optical Character Recognition," A master in computer science thesis, Sudan University of Science and Technology, 2011
- [6] W Ali, M. E. M. Musa "Recognition of Arabic Handwritten Names Using Probabilistic Neural networks," Computer Studies Journal, Union of Arab Scientific Research Councils, Volume 1, Issue 1 (in Arabic).
- [7] M. Kherallah, N. Tagougui, Adel M. Alimi, H. El Abed, and V. Märgner, "Online Arabic Handwriting Recognition Competition," 2011 International Conference on Document Analysis and Recognition (ICDAR), Pages 1454-1458, 18-21 Sept. 2011, Beijing.
- [8] O. Balola A. Shaout and M. E. M. Musa "Two stage classifier for Arabic Handwritten Character Recognition ," International Journal of Advanced Research in Computer and communication Engineering, Vol 4, Issue 12 December 2015
- [9] H. A. Abd Alshafy, M. E. M. Musa, "Characters' boundaries based segmentation for online Arabic handwriting," In Proceedings of IEEE International Conference on Computing, Electrical and Electronics Engineering (ICCEEE13), Khartoum, Sudan, 26-28 Aug. 2013.
- [10] H. A. Abd Alshafy, M. E. M. Musa, "Online Arabic Handwriting Recognition," Doctoral thesis, Sudan University of Science and Technology.

**First Author** Mohamed Elhfiz Mustafa Musa has received his BSc and MSc from University of Khartoum in 1989 and 1996 respectively. He has received his PhD in Computer Engineering in 2003 from Middle East Technical University, Ankara, Turkey. Mr Musa was the dean of the college of computer science and Information Technology, Sudan University of Science and Technology, for eight years ( 2006-2014). He is now the director of the computer Center in Sudan University of Science and Technology

## Towards building competent dataset of Arabic recognition

Table 1: this table contains statistic for eight Arabic names; the tables shows the number of sample for each name as well as the different number of strokes for the name and their statistics

Name	Number of Patterns which take										Total Number of Patterns
	1 Stroke	2 Strokes	3 Strokes	4 Strokes	5 Strokes	6 Strokes	7 Strokes	8 Strokes	9 Strokes	10 Strokes	
عثمان	—	—	—	3	17	28	12	2	1	1	64
علي	27	35	38	6	2	—	—	—	—	—	108
فاطمة	—	—	—	1	23	26	11	6	2	—	69
مروة	—	—	—	18	29	11	7	3	—	—	68
منى	—	80	14	6	1	—	—	—	—	—	101
نفيسة	—	—	—	4	11	8	34	3	1	—	61
هبة	—	58	23	4	3	1	—	—	—	—	89
وفاء	—	—	—	47	18	6	3	—	—	—	74

جامعة السودان للعلوم والتكنولوجيا  
كلية علوم الحاسوب وتقانة المعلومات  
إستمارة تقديم داخلي لشهادة التخرج

---

جنس: ☒ أنثى ☐ ذكر

الاسم (بالعربي):	الاسم (بالحروف):	الاسم (باللغة الإنجليزية):	الاسم (باللغة الفرنسية):
عبدالله	Abdullah	Abdullah	Abdullah
فاطمة	Fatma	Fatma	Fatma
علي	Ali	Ali	Ali
عثمان	Osman	Osman	Osman

Fig. 2: Part of the application form used in name dataset collection