

ICDAR 2005 Arabic Handwriting Recognition Competition

V. Märgner, M. Pechwitz¹, H. El Abed

Technical University Braunschweig, Institute for Communications Technology,
Schleinitzstraße 22, 38092 Braunschweig, Germany
{maergner, elabed}@tu-bs.de

Abstract

This paper describes the Arabic handwriting recognition competition for ICDAR 2005. With the presentation of the IFN/ENIT-database in the year 2002 a database with handwritten Arabic town names was made available for free to non commercial research groups. Till now more than 30 groups are working with this data worldwide.

By announcing a competition of Arabic handwriting recognition systems based on the IFN/ENIT-database, we hope to contribute to the development of Arabic handwriting recognition systems. The use of the same database by different research groups allows the comparison of different systems. We compare the systems on the most important characteristic: recognition rate, but also features like word length, writing style, and character connectivity will be discussed.

1. Introduction

Handwriting recognition is of great interest as automatic document management systems are using more and more scanning devices, OCR systems and document interpretation methods. Even though more and more printed documents of good quality are used, handwriting is still in use and often inhibits automatic processing.

Arabic OCR (AOOCR) systems are not widely in use, and Arabic handwriting recognition still lacks huge databases and comparable results. The situation is somewhat better with the release of the IFN/ENIT-database. Even though this data was captured on a special form it gives the possibility to compare different systems and algorithms.

To the best of the Authors' knowledge, there has not been any previous Arabic handwriting recognition

competition. But we know from speech recognition and OCR system development that international competitions are able to promote the development substantially.

2. The IFN/ENIT-database

The IFN/ENIT-database was developed to advance the research and development of Arabic handwritten word recognition systems. Since the presentation of this database at the CIFED 2002 conference [1] more than 30 groups started to work with that database, which is freely available (www.ifnenit.com) for non commercial research. The database in version 1.0 patch level 2 (v1.0p2) consists of 26,459 Arabic names handwritten by 411 different writers. Written are 937 Tunisian town/village names. Each writer filled some forms with pre-selected town/village names (here we use the word names only in the following) and the corresponding post code. Ground truth was added to the image data automatically and verified manually. Fig. 1 shows a dataset entry of the IFN/ENIT-database.

Image	
Ground truth:	
Postcode	8050
Global Word	الحمامات
Character shape sequence	ا ل ل م ح ا ل م ل ا ل ا ت
Baseline y1,y2	52,47
Baseline quality	B1 (B1=OK; B2=bad)
Quantity of words	1
Quantity of PAW's	4
Quantity of characters	8
Writing quality	W1 (W1=OK; W2=bad)

Figure 1: Example of a dataset entry

¹ Now with: Paradatac GmbH, Mittelweg 7, 38106 Braunschweig, Germany (mp@ifnenit.com)

Tables 1-4 show most important statistics of the IFN/ENIT-database. Table 1 shows the quantities of names, PAWs (connected Parts of Arabic Word), and characters subject to the number of words in a name.

Table 1: Quantity of words, name images, PAWs, and characters in a name

words in a name	names	PAWs	characters
1	12,992	40,555	76,827
2	10,826	54,722	98,828
3	2,599	20,120	36,004
4	42	188	552
Total	26,459	115,585	212,211

Table 2 shows a statistic of the number of PAWs in the names of the database.

Table 2: Frequency of PAWs in a name

Number of PAWs	frequency in %	Number of PAWs	frequency in %
1	2.99	6	8.24
2	15.35	7	7.32
3	17.60	8	6.04
4	24.84	>8	2.95
5	14.67		

Table 3 shows the statistic of the age and the profession of the writers who contributed to the database.

Table 3: Age and profession of the writers

age	student	teacher	technician	other	Σ
≤ 20	29%	0%	0%	0%	29%
21 - 30	35.6%	4.2%	3.9%	3.9%	47.6%
31 - 40	0.2%	3.4%	4.9%	2.0%	10.5%
> 40	0%	4.1%	5.4%	3.4%	12.9%
Σ	64.8%	11.7%	14.2%	9.3%	100%

Finally table 4 shows the size and the number of writers in the dataset d of the IFN/ENIT database.

Table 4: Statistic of the dataset d

set	number of words	number of writers
d	6735	104

Information about further details of the IFN/ENIT-database can be found at www.ifnenit.com.

The new and to the participants of this contest unknown dataset *e* consists of 6033 town/village names. Table 5 shows the frequency of the PAWs in a name of this set. A slight difference to IFN/ENIT-database can be seen. Table 6 shows a statistic of the age and the profession of the writers of the new dataset *e*. It can be seen that here more older and less students contributed to the dataset.

Table 5: Frequency of PAWs in a name of set *e*

Number of PAWs	frequency in %	Number of PAWs	frequency in %
1	4.92	6	8.50
2	16.64	7	2.57
3	26.44	8	2.02
4	23.59	>8	0.31
5	15.08		

Table 6: Age and profession of the writers of set *e*

age	student	teacher	technician	other	Σ
≤ 20	4.2%	0%	0%	4.3%	8.5%
21 - 30	2.8%	4.2%	0%	12.7%	19.7%
31 - 40	0%	8.5%	8.5%	19.6%	36.6%
> 40	0%	2.8%	7.0%	25.4%	35.2%
Σ	7.0%	15.5%	15.5%	62.0%	100%

Table 7 shows the size and the number of writers of dataset *e*.

Table 7: Statistic of dataset *e*

set	number of words	number of writers
e	6033	87

3. The competition

We invited groups working on Arabic handwritten word recognition to adopt their system to the IFN/ENIT-database and send us their system. In addition to the already freely available datasets we prepared another still unknown amount of data to be used for the competition. The data are exactly of the same type (Tunisian town/village names written on a form) but written by unknown writers. The statistic of the age of the writers differs also (see table 6). The unknown dataset contains more data from older writers than the known dataset.

4. Participating systems

The following text gives a short overview of the systems which are participating in the competition. The amount of information we received from the participating groups is very different. Some groups gave us detailed information or sent us journal papers. Other groups gave us only some short description, and one group declared all material including names of authors highly confidential. Nevertheless, we tested all systems we had. It is the first time to compare Arabic handwriting recognition systems and we hope that next time all participants will give detailed information, when it is clear that the competition is objective and useful for participants and for the community.

4.1. System ICRA

The system with the name **ICRA** (Intelligent Character Recognition for Arabic), which also means *read* in Arabic, was sent by Ahmad Abdul Kader, employed as a software architect by Microsoft within the Handwriting recognition group. His participation in this competition is as a free lancer and not as a Microsoft employee. Here a short description of the system:

- The system uses a novel idea that is inspired by the nature of Arabic writing. It is based on the concept of what is known as PAW (Part of Arabic Word).
- ICRA is a two tier recognizer.
- The 1st tier is a Neural Net based PAW recognizer that is aided by a PAW lexicon. The PAW lexicon is extracted from the master village names lexicon.
- The 2nd tier is a Neural Net based word recognizer that is aided by another lexicon. The literals (alphabet) of such a lexicon are actually the PAWs and not the characters.
- ICRA was trained on sets *a*, *b*, and *c* and tested on set *d* of the IFN/ENIT-database.
- There are no publications about this new system. The author plans to submit papers in the near future.

4.2. System SHOCRAN

The system with the name **SHOCRAN** (System for Handwritten Optical Character Recognition for Arabic Names) comes from a group of researchers in Egypt. It is declared as a confidential project, so we are not allowed to give more information.

The system is trained on all four datasets of the IFN/ENIT-database.

4.3. TH-OCR

The system with the name **TH-OCR** was developed at the State Key Laboratory of Intelligent Technology and Systems, Department of Electronic Engineering, Tsinghua University, Beijing, China, by Pingping XIU, Hua WANG, Jianming JIN, Yan JIANG, Liangrui PENG, and Xiaoqing DING.

Based on previous research work on a multilingual document recognition system for Chinese, Japanese, Korean, English, Tibetan and Uyghur languages, this research work was extended to Arabic OCR. A first step was the development of a printed Arabic document recognition system in the year 2004 [2]. The system structure of the handwritten Arabic OCR

system is similar to that of the printed Arabic OCR system, but the key technologies of handwritten character segmentation and recognition are more complex and sophisticated. The system consists of text line, word, and character segmentation, character recognition, and post-processing based on a language model.

The text line and word segmentation module adopts a hybrid top-down and bottom-up method, on the basis of connected components (CCs) classification.

The character segmentation module utilizes multiple schemes to handle the cursive character segmentation. The most important scheme is to detect potential segmentation points by using structural or geometrical configuration characteristics of the cursive script, followed by a merging of over-segmented characters.

The character recognition module is mainly based on statistical pattern recognition methods. Structural information is used for similar characters discrimination.

As the ICDAR 2005 Arabic Handwriting Recognition Competition is running on a closed dictionary set, recognition results are compared with candidate address items to find the best match. A candidate address is seen as a template, all characters are checked to find their possible correspondences in the recognition results. Two types of cost are taken into account, one is recognition cost and the other is matching cost.

The research work on handwritten Arabic OCR aims for developing a practical system to digitize handwritten Arabic documents.

4.4. UOB

The system with the name **UOB** was developed at the University of Balamand, Lebanon by Chafic Mokbel.

The UOB system is a pure HMM system developed for speech recognition at the origin. It uses a complete toolkit like HTK [<http://htk.eng.cam.ac.uk/>], which is called HCM. HCM permits the development of large HMM networks and it integrates language modeling. The properties in the HCM are published in papers in the speech recognition area e.g. [3].

The work on handwritten word recognition was started using HCM with a PhD project still under preparation by Mr. Ramy El-Hajj who developed the feature extraction module. All the work on handwritten word recognition is being done in tight collaboration with Laurence Likforman-Sulem from ENST-Paris. A paper describing the feature extraction module will be presented at ICDAR 2005 [4].

For the UOB system all four datasets were used for training. No confidence measure is implemented.

4.5. REAM

The system with the name **REAM** (**R**econnnaissance de l'**E**criture **A**rabe **M**anuscrite) comes from a group at the Laboratoire des Systèmes et de Traitement du Signal-ENIT, Tunisia. The authors of the system are Sameh Masmoudi Touj, Najoua Essoukri Ben Amara, Hamid Amiri, and Noureddine Ellouze.

The system uses a hybrid planar Markov Model to adapt to horizontal and vertical variations of the handwritten word. The approach will be presented in a journal paper in October 2005 [5].

The principal idea of this approach is the partitioning of handwritten words into five logical horizontal bands which correspond to typical Arabic parts of words like upper and lower diacritics, ascenders, descenders, and median zone. This segmentation is done in a sophisticated way using knowledge about the typical Arabic writing style. Additionally based on features of the median zone, vertical segmentation points are detected. In the next step for each type of segment a different technique of feature extraction is adopted. Finally the recognition is realized using the concept of PHMM. In the paper [5] first results of the system on parts of the IFN/ENIT-database are reported.

5. Tests

We evaluated the performance of the 5 Arabic handwriting recognition systems in two steps. In the first step we used the dataset *d* of the IFN/ENIT-database. In a second step we used the new and to all participants unknown dataset *e*.

5.1. General remarks

All participants sent us a running version of their recognition system. A first set with five input images of town/village names was used to test the basic functionality of the systems on our PC environment. All systems passed this test.

A second test with the dataset *d* of the IFN/ENIT-database which consists of 6735 town/village names failed for two systems. Nevertheless we can show results of these systems. For one system the dataset was split into smaller sets, and for another system we could use a subset of all names only.

5.2. Recognition Results

The first comparison was made on the basis of one set of the well known IFN/ENIT-database. Table 8 shows the results of the five systems in the competition, and we added the results of our system (light gray background) for comparison, which was presented at ICDAR 2003 [6] and trained on datasets *a*, *b*, and *c*. While most systems used the whole IFN/ENIT-database for training remarkable differences between the performances of the five systems can be seen. The first column shows the percentage of correct recognized city names, the second the percentage of correct result within the first five results, and the third column of the first 10.

Table 8: Recognition results in % with IFN/ENIT-database dataset *d*

System	1	1-5	1-10
ICRA	88.95	94.22	95.01
SHOCRAN	100	100	100
TH-OCR	30.13	41.95	46.59
UOB	85.00	91.88	93.56
REAM*	89.06	99.15	99.62
ARAB-IFN	87.94	91.42	95.62

* System is tested on reduced set with 1000 names

It can be seen that the properties of the systems are very different especially the system SHOCRAN, which shows 100% correct recognition, seems to be over-trained on the IFN/ENIT database.

Table 8 shows recognition results on data used for training and it follows that the behavior of the system on new data may differ very much from these results.

The following table 9 now shows the results of the systems reached with the new and unknown dataset *e*.

These results show some surprising effects. The result of system SHOCRAN confirms the impression that it was really over-trained but the same seems to be true for system REAM. For both systems the recognition rate is highly reduced.

Table 9: Recognition results in % with the new dataset *e*

System name	No.	1	1-5	1-10
ICRA	1	65.74	83.95	87.75
SHOCRAN	2	35.70	51.62	51.62
TH-OCR	3	29.62	43.96	50.14
UOB	4	75.93	87.99	90.88
REAM*	5	15.36	18.52	19.86
ARAB-IFN	6	74.69	87.07	89.77

* System is tested on reduced set with 3000 names

The systems ICRA and UOB show a normal behaviour with a slightly reduced recognition rate. The system TH-OCR shows more or less the same result than on the dataset *d*, here a better adaptation to the database should be possible.

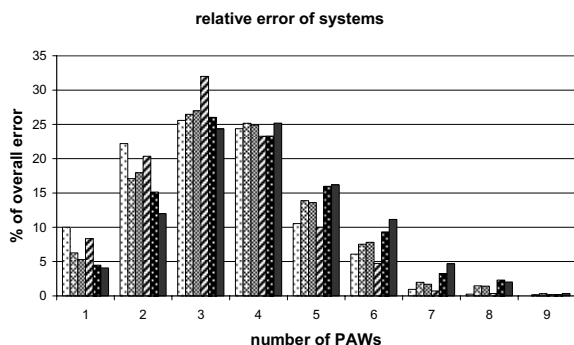


Figure 2: Relative errors of the systems dependent on the number of PAWs in a name. In each group the bars are associated with the systems 1-6 from left to right.

A comparison of the relative error rates of the different systems dependent on the number of PAWs in a name is shown in figure 2. It can be seen that for all systems the relative error follows in general the frequency of the PAWs in dataset *e* (Table 5). But some systems show more errors in long words whereas some other in short words. This may depend on different features and different normalisation.

Figure 3 shows some examples of a word written by different writers together with the results of the recognition systems. It can be seen that some very badly written words are correctly recognised with some systems while better written words are still not recognised. On a first glance systems without segmentation show better results on words with connected or broken PAWs than segmenting ones.

A comparison of the systems in more detail is not possible here as we do not have enough details of all systems under test.

6. Conclusions

The motivation of the ICDAR 2005 Arabic Handwritten Word Recognition Contest was to evaluate different systems. Five systems were evaluated and we added for comparison our system. The tests show that best results were achieved with HMM based systems and with a Neural Net approach. But other systems using HMM or NN achieved even very low recognition rates. These results again show us that it is not enough to decide which recognition method has to be used but also the features and the

normalization play a very important role in the performance of a recognition system.

System	1	2	3	4	5	6
	3	-	1	1	-	1
	1	-	-	1	-	1
	2	4	-	4	1	-
	1	1	1	1	-	1
	1	1	1	1	1	-
	1	-	-	3	1	-
	1	1	1	1	-	1
	1	-	1	1	-	1

Figure 3: Example of a town/village name written by different writers out of dataset *e* and the recognition results (1 first correct, 2 second, ..., - no correct result)

7. References

- [1] M. Pechwitz, S.S. Maddouri, V. Maergner, N. Ellouze, and H. Amiri, "IFN/ENIT-database of handwritten Arabic words", In Proc. Of CIFED 2002, Hammamet, Tunisia, October 21.-23.2002, pp. 129-136
- [2] Jianming Jin, Hua Wang, Xiaoqing Ding, Liangrui Peng, "Printed Arabic document recognition system", Proc. Of SPIE-IS&T Electronic Imaging, vol. 5676, 2005, pp. 48-55
- [3] C. Mokbel, H. Abi Akl, H. Greige, "Automatic speech recognition of arabic digits over Telephone network", Proc. of RTST, 2002
- [4] Ramy El-Hajj, Laurence Likforman-Sulem, Chafic Mokbel, "Arabic Handwriting Recognition Using Baseline Dependent Features and Hidden Markov Modeling", Proc. of ICDAR 2005, Seoul, Korea, August 29.-September 1.2005
- [5] S. M. Touj, N. E. Ben Amara, "Arabic Handwritten Words Recognition based on a Planar Hidden Markov Model", The International Arab Journal of Information Technology, vol. 2, no. 4, October 2005
- [6] M. Pechwitz, V. Maergner, "HMM based approach for handwritten Arabic word recognition using the IFN/ENIT-database", Proc. of 7th ICDAR 2003, pp. 890-89