IFN/ENIT - DATABASE OF HANDWRITTEN ARABIC WORDS

Mario Pechwitz[1], Samia Snoussi Maddouri[2], Volker Märgner[1], Noureddine Ellouze[2] , Hamid Amiri[2]

[1]Institute for Communications Technology (IFN), Technical University Braunschweig, Germany.

[2]Ecole Nationale d'Ingénieur de Tunis (ENIT), BP 37 le Belvédère 1002, Tunis.

pechwitz@ifn.ing.tu-bs.de, samia.maddouri@enit.rnu.tn, maergner@ifn.ing.tu-bs.de, n.ellouze@enit.rnu.tn, hamid.amiri@enit.rnu.tn

ABSTRACT. In this paper we are presenting a new database with handwritten Arabic town/village names. For each name the ground truth information, e.g. the sequence of character shapes, some style information, and the baseline are coded. 411 writers filled forms with about 26400 names containing more than 210000 characters. The database is described in detail. It is designed for training and testing recognition systems for handwritten Arabic words. The IFN/ENIT-database is available for the purpose of research.

RÉSUMÉ : Dans cet article on présente une novelle base de données, qui contient des noms manuscrits de villes/villages arabes. Pour chaque nom les informations de base, par exemple l'ordre des formes de caractère, les informations sur le style de l'écriture, et la ligne de base sont codées. 411 auteurs ont rempli des formulaires avec plus de 26400 noms contenant plus de 210 000 caractères. La base de données est décrite en détail, et elle est conçue pour la formation et l'essai des systèmes d'identification pour les mots arabes manuscrits. La base de données -IFN/ENIT est disponible pour la recherche.

KEYWORDS: *Arabic, Handwriting, Database, Recognition system, Arabic OCR*

MOTS-CLÉS : *Arabe, manuscript, base de données, systeme d'identification, OCR Arabe*

1. Introduction

Automatic recognition of handwritten words remains a challenging task even though the latest improvements of recognition methods and systems are very promising. Especially for the automatic recognition of Arabic handwritten words a lot of work has still to be done. It goes without saying that the most important requirement for the development and comparison of recognition systems is a large database together with ground truth (GT) information. Compared to English text where handwritten words and numbers have been publicly available for a long time (e.g. (CEDAR),(NIST)) the situation for Arabic today is quite different. In the case of Arabic handwritten words many papers use a specific, more or less small dataset of their own (see e.g. (Al-Badr *et al.,* 1995) and (Amin, 1998)), or they talk about large databases that are not available to the public (see e.g. (Kharma et al., 1999) and (Al-Ohali et al., 2000)). Reasons for this unsatisfying situation might be insufficient commercial interest, although many people are writing in Arabic, as well as the difficult, time consuming, and error prone process of generating ground truth (GT) for Arabic on character level (which is comparably easy for handwritten English words). To overcome this situation we developed the *IFN/ENIT*-database of Arabic handwritten words, which is not only essential for our research on handwritten Arabic word recognition but also a chance for other researchers to develop or to compare recognition methods and systems.

This paper is intended to give a detailed insight into the *IFN/ENIT*-database and is arranged as follows. Section 2 provides an overview of the database. Section 3 follows with more details about the construction, labelling, and verification of the database. In Section 4 information about statistics and further details of the database are presented. Section 5 describes our future plans with the *IFN/ENIT*-database. Finally, we finish with some concluding remarks.

2. Overview of the *IFN/ENIT*-database

Today the error rates of recognitions systems are only acceptable for applications with a restricted lexicon of words. That is why recognition systems are focused on a certain application such as the reading of cheque amounts or addresses, which are proven to be realistic and profitable. However the further development of recognition systems needs a large amount of data to train and test the system. Real world data, especially such from the bank or the post area, often is confidential and inaccessible for non-commercial research. As the amount of data is crucial for a reliable training of recognition systems, we decided to use similar artificial data instead of scarce real world data. As test environment we choose the names of 946 Tunisian town/villages together with the postcode. More than four hundred people, most of them selected from the narrower range of the Ecole Nationale d'Ingénieurs de Tunis (ENIT), contributed to the *IFN/ENIT*-database. Each writer was asked to fill a form with handwritten pre-selected names of Tunisian towns/villages and the corresponding postcode. An example of a filled form is shown in figure 1. Town names and numbers were extracted automatically, and GT and baseline information were added automatically as well. Finally, GT and baseline information were verified manually. As an example of the *IFN/ENIT*-database table 1 shows the images of a name, written by different writers. The GT, available for each town/village name image, will be described in section 4. The whole *IFN/ENIT*-database consists of 26459 handwritten Tunisian town/village names. The postcodes have not been processed yet.

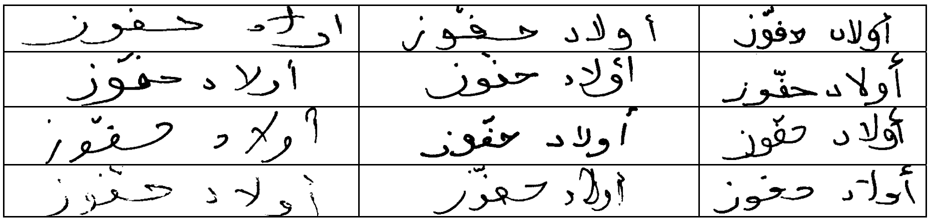


Table 1. *Examples from the IFN/ENIT-database: A town/village name written by 12 different writers.*

3. *IFN/ENIT*-database of handwritten words

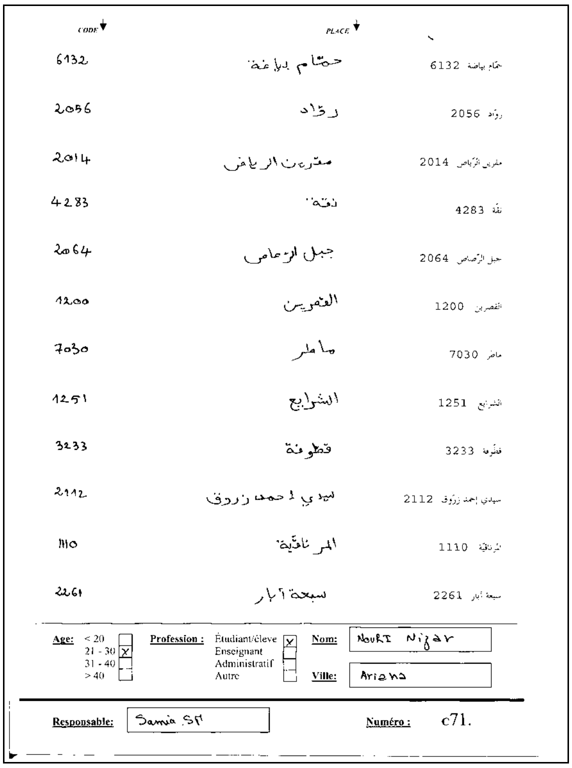
In this section we give a description of all steps we undertook to build the *IFN/ENIT*-database. The GT information will be described in section 4.

3.1. *The form*

Our aim was to collect images of handwritten town names written in a similar quality as town names of an address on a letter. The form was designed to:

- encourage writing without strong constraints
- collect writing similar to writing on a letter
- be easy to process automatically
- provide additional information about the person who filled it.

A filled example of the devised form is shown in figure 1. The form consists of three columns and a text block at the bottom. Embodied in the column on the right hand side of the form are 12 lines with printed Tunisian town/village names and their respective postcodes, which are automatically selected from the possible 946 names. The sample writers were expected to write the postcode in the left column and the town/village name in the middle column in their individual writing style. We did not print a line to write on or a box to write in, since we wanted to make the processing of the scanned data as simple as possible. To provide a light writing guidance we printed dark black rectangles on the backside of each page, which are shining through to the front side and thus mark where to write. In the scanning process these rectangles can be removed using a simple threshold. Further segmentation operations are not necessary. The names printed on each form were selected randomly with the condition that each character shape should occur at minimum more than 200 times. Therefore, those names with rare character shapes occur more often than names with frequent ones. Each word appears at least 3 times in the database. A page number is used as form identifier for the subsequent processing. In the block at the bottom additional information about age, profession, and identity of the writer is given. Each writer was asked to fill 5 forms, so one had to write 60 names.



CODE	PLACE	
6132	حمّام بياضة	6132 حمّام بياضة
2056	روّاد	2056 روّاد
2014	مقرين الرياض	2014 مقرين الرّياض
4283	نقّة	4283 نقّة
2064	جبل الرصاص	2064 جبل الرّصاص
1200	القصرين	1200 القصرين
7030	ماطر	7030 ماطر
1251	الشرايع	1251 الشرايع
3233	قطّوفة	3233 قطّوفة
2112	سيدي أحمد زروق	2112 سيدي إحمد زرّوق
1110	المرناقيّة	1110 المرناقيّة
2261	سبعة آبار	2261 سبعة آبار

Age: < 20 □ 21 - 30 ☒ 31 - 40 □ > 40 □

Profession : Étudiant/élève ☒ Enseignant □ Administratif □ Autre □

Nom: NouRI Nizar

Ville: Ariana

Responsable: Samia SF

Numéro : c71.

Figure 1. *An example of a filled form*

3.2. *Form processing*

All form pages were scanned with 300 dpi and converted to black and white (binary) images. Due to the fact that the paper was white and the words were written with a black or dark blue pen the binarisation was not a problem. While scanning a page the page number and the additional information were keyed in manually. A page slope correction was performed automatically using the dark black line at the bottom of the page as horizontal reference. An advanced projection method was performed to extract the word and the postcode images on the page automatically.

3.3. *Pre-label and pre-baseline*

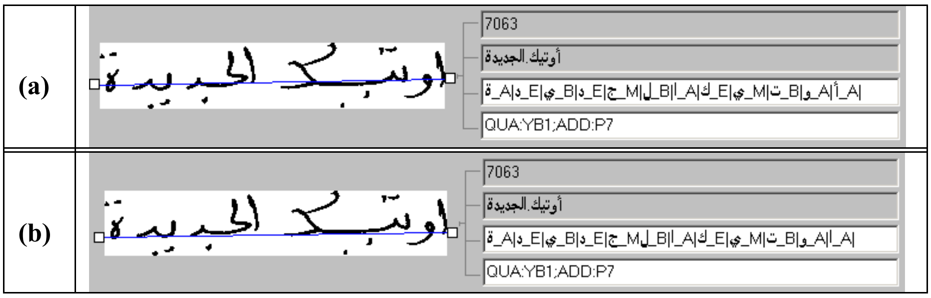
With the knowledge of the page number each word is automatically assigned a pre-label. The pre-label of the word consists of the postcode, the word in Arabic code set ISO 8859-6, and a code that describes the sequence of the Arabic character shapes. This character shape code is generated automatically using a simple glyphshaping algorithm. This is an important step for the labelling of character shapes, because the Arabic code set ISO 8859-6 does not provide shape occurrence information. To make the code unambiguous we added additional Latin characters as indexes. “B” standing for beginning, “M” for middle, “E” for end and “A” for alone/isolated character shapes. An “L” marks the “Chadda”. These descriptors are linked to the Arabic code with the “_” (c.f. table 2). The codes for each character are separated by “|” (c.f. table 4). Additionally, a pre-baseline estimation was performed on each name image automatically. This straight line should give a good estimation of the writing line of the name.

Shape position index	**Index description**	**Letter and shape position index**	**Letter shape**
B	Beginning	ح_B	ﺣ
M	Middle	ح_M	ﺤ
E	End	ح_E	ﺢ
A	Alone	ح_A	ح
L	Ligature	ح _ B ّ_L	ﺣّ

Table 2. *Labels for different shapes of letter ح*

3.4. *Verification*

Due to the variability of handwriting or simply due to writing errors, the pre-labels do not always match the handwritten word. Therefore, manual verification is needed to obtain a label that matches the handwritten word character sequence. For example, when ligatures were used and “Chadda” were not copied, the label had to be corrected. Words with writing errors or written with special ligatures, which appear only sporadically, were not included in the *IFN/ENIT*-database. During this verification procedure the automatically generated pre-baseline is also corrected. The interface for this verification procedure is shown in figure 2a. Figure 2b shows the corrected label and baseline of the image presented in figure 2a.



(a)
7063
أوتيك.الجديدة
ة_A|د_E|ي_B|د_E|ج_M|ل_B|ا_A|ك_E|ي_M|ت_B|و_A|أ_A|
QUA:YB1;ADD:P7

(b)
7063
أوتيك.الجديدة
ة_A|د_E|ي_B|د_E|ج_M|ل_B|ا_A|ك_E|ي_M|ت_B|و_A|ا_A|
QUA:YB1;ADD:P7

Figure 2. (a) *Interface of label verification. On the left you see the image of the handwritten word with the pre-baseline, on the right we have the pre-label, and in the last row some quality and additional marks are shown.**(b)** Correct label after the verification step*

4. Database details and statistics

The *IFN/ENIT*-database contains 26459 handwritten Tunisian town/village names. Table 3 provides an overview of the number of images, words, PAWs (connected **P**art of **A**rabic **W**ord), and characters contained in the *IFN/ENIT*-database.

Quantity of words in town names	Quantity of town name images	Quantity of PAW's	Quantity of characters
1	12992	40555	76827
2	10826	54722	98828
3	2599	20120	36004
4	42	188	552
Total	26459	115585	212211

Table 3. *Quantity of town name images, words, PAWs, and characters in a words*

Each handwritten town name comes with image and GT information. The following GT information is available for each name image:

- Postcode (automatically generated)
- Arabic word in ISO 8859-6 code set (automatically generated)
- Arabic word as character sequence with shape-index (automatically generated and manually verified)
- Number of words, PAW's, and characters in the town/village name (automatically generated and manually verified)
- Writer identifier, age, profession and writing quality (manually labelled)
- Baseline (automatically generated and manually verified)
- Baseline quality (manually labelled)

Table 4 gives two examples of data set entries of the *IFN/ENIT*-database.

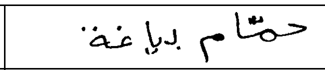
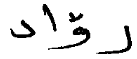
<table>
<tr><td><b>Image</b></td><td></td><td></td></tr>
<tr><td><b>Ground truth:</b></td><td></td><td></td></tr>
<tr><td>Postcode</td><td>6132</td><td>2056</td></tr>
<tr><td>Global word</td><td>حّمام بـيـاضة</td><td>روّاد</td></tr>
<tr><td>Character shape sequence</td><td>م_A|ا_E|م_M_ّL|ح_B|
ة_E|ض_B|ا_E|ي_M|ب_B|</td><td>و_A_ّL|ر_A|
د_A|ا_A|</td></tr>
<tr><td>Baseline y1,y2</td><td>70,50</td><td>46,39</td></tr>
<tr><td>Baseline quality</td><td>B1 (B1=OK; B2=bad)</td><td>B1</td></tr>
<tr><td>Quantity of words</td><td>2</td><td>1</td></tr>
<tr><td>Quantity of PAWs</td><td>4</td><td>4</td></tr>
<tr><td>Quantity of characters</td><td>9</td><td>4</td></tr>
<tr><td>Writing quality</td><td>W1 (W1=OK; W2=bad)</td><td>W1</td></tr>
</table>

Table 4. *Two data set entries of the IFN/ENIT-database*

In addition to the basic character shapes the *IFN/ENIT*-database containts 10 frequently used ligatures, which are summarized in table 5. The grey boxes in table

5 correspond to ligature shapes that are specific to handwriting and do not exist in printed letters unlike the shapes in the first four boxes.

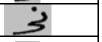
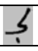
Label	Quantity	Shape	Label	Quantity	Shape
ا_Eل_B	789	لا	ح_Mل_B	363	
ا_Eل_M	1066	ـلا	ح_Mم_Mل_B	64	
إ_Eل_B	120	لإ	ح_Mن_B	100	
أ_Eل_B	357	لأ	خ_Mل_B	304	
ج_Mل_B	530		م_Mل_B	450	

Table 5. *Statistical results on different ligature shapes*

Information about age and profession was collected on each page. Actually, about 2/3 of the 411 writers were younger than 31 years and about 2/3 were students.

5. Working with the *IFN/ENIT*-database

The *IFN/ENIT*-database of handwritten Tunisian town/village names is publicly available for the purpose of research. It can be ordered by sending an e-mail to one of the authors. At the IFN Arabic recognition systems are developed based on a statistical HMM approach. An Arabic OCR system for printed Arabic words has been developed recently. This system was trained and tested on synthetically generated data (Maergner *et al.*, 2001). Now the IFN is working on an adaptation of the system to recognize handwritten words, using the *IFN/ENIT*-database. For the purpose of training such a system information about the sequence of character shapes in all words is very useful. As GT of the base line of name images is available in the *IFN/ENIT*-database, e.g. a system can be tested without a baseline finding algorithm module implemented. Up to now the IFN system uses a sliding window to collect features of a name images. With the knowledge of the position of the baseline this sliding window is shifted over the word image from right to left. Additionally, the baseline is often used as parameter to normalise a word. At the IFN tests to estimate the baseline information (Pechwitz *et al.*, 2002) have been made and several other normalisation parameters are currently being tested. The great advantage of the *IFN/ENIT*-database is that each name image comes with a manually verified baseline as a reference. At the ENIT a handwritten Arabic word recognition system based on a neural network approach (Snoussi *et al.*, 2002) uses words of the *IFN/ENIT*-database to evaluate the recognition rate.

6. Summary and Outlook

In this paper we have introduced the *IFN/ENIT*-database as a new database for handwritten Arabic words. The *IFN/ENIT*-database is publicly available for the purpose of research. 411 different writers filled 2265 forms. The large number of writers guarantees a wide variety of writing styles. 26459 valid Tunisian town/village names were extracted from the forms and the GT was added. All name images come with GT on character shape sequence level. Also the baseline and additional quality information are part of the GT. Statistics describe the *IFN/ENIT*-database in detail. With the *IFN/ENIT*-database it is possible to develop and test Arabic handwritten word recognition systems or parts of them. We, for example, have used parts of the *IFN/ENIT*-database to develop and verify baseline-finding methods (Pechwitz *et al.*, 2002). The baseline GT was essential for these tests. We kindly invite other researchers to test their systems with the *IFN/ENIT*-database.

6. Acknowledgement

Thanks are due to everyone at ENIT who participated in collecting data. Special thanks go to Emna Ben Yahia, Sameh Masmoudi, Sofien Touj, Chayma Aschi, as well as to Najoua Ben Amara and her group for their help doing label verification. This project is supported by the Volkswagen Stiftung under contract number I/75 084.

7. References

B. Al-Badr and S. A. Mahmoud, «Survey and Bibliography of Arabic Optical Text Recognition », *Signal Processing*, 41:49-77, 1995.

Y. Al-Ohali, M. Cheriet and C. Suen, «Database For Recognition of Handwritten Arabic Cheques », *Proc. of the 7'th IWFHR*, pp. 601-606, 2000

A. Amin, «Off-line Arabic Character Recognition : The State of the Art », *Pattern Recognition*, 31(5) :517-530, 1998.

CEDAR, http://www.cedar.buffalo.edu/Databases/CDROM1/

N. Kharma, M. Ahmed, R. Ward, «A New Comprehensive Database of Hand-written Arabic Words, Numbers and Signatures used for OCR Testing », *IEEE Canadian Conference on Electrical and Computer Engineering*, pp 766-768, 1999.

V. Märgner and M. Pechwitz, «Synthetic Data for Arabic OCR system development », *In Proc. of the 6.th. Int. Conference on Document Analysis and Recognition, ICDAR 2001*, pp. 1159-1163, 2001

NIST, http://www.nist.gov/srd/nistsd19.htm

M. Pechwitz and V. Märgner, «Baseline Estimation for Arabic Handwritten Words », *appears in Proc. of the 8'th IWFHR*, 2002

S. Snoussi Maddouri, H. Amiri, A. Belaid, and Ch. Choisy, «Combination of Local-global Vision Modeling for Arabic Handwritten Word Recognition », *appears in Proc. of the 8'th IWFHR*, 2002