LAMIS-MSHD: A Multi-Script offline Handwriting Database

Chawki Djeddi
LAMIS Laboratory
University of Tebessa
Tebessa, Algeria
c.djeddi@mail.univ-tebessa.dz

Abdeljalil Gattal
LAMIS Laboratory
University of Tebessa
Tebessa, Algeria
ab.gattal@mail.univ-tebessa.dz

Labiba Souici-Meslati
LISCO Laboratory
Badji Mokhtar-Annaba University
Annaba, Algeria
souici_labiba@yahoo.fr

Imran Siddiqi
Department of Computer Science
Bahria University
Islamabad, Pakistan
imran.siddiqi@bahria.edu.pk

Youcef Chibani
Speech Communication and Signal Processing Laboratory
USTHB University
Algiers, Algeria
ychibani@usthb.dz

Haikal El Abed
Technical Trainers College
German International Cooperation
Riyadh, Kingdom of Saudi Arabia
haikal.elabed@giz.de

***Abstract*— This paper introduces a new offline handwriting database that was developed to be employed in performance evaluation, result comparison and development of new methods related to handwriting analysis and recognition. The database can particularly be used for signature verification, writer recognition and writer demographics classification. In addition, the database also supports isolated digit recognition, digit/text segmentation and recognition and similar related tasks. The database comprises 600 Arabic and 600 French text samples, 1300 signatures and 21,000 digits. 100 Algerian individuals coming from different age groups and educational backgrounds contributed to the development of database by providing a total of 1300 forms. The database is also accompanied with ground truth data supporting the evaluation of the aforementioned tasks. The main contribution of the database is providing a multi-script platform where same authors contributed samples in French and Arabic. It would be interesting to explore applications like writer recognition and writer demographics classification in a multi-script environment.**

I. INTRODUCTION

Research in automatic analysis of handwriting has received considerable attention in the recent years. Typical applications include handwriting recognition, word spotting, segmentation of handwriting into words/character and writer identification etc. With the increase in research in the aforementioned problem areas has increased the need for the availability of standard databases to match the real world scenarios as closely as possible. Standard databases are increasingly gaining popularity in all scientific domains and same is the case with document analysis in general and handwriting analysis in particular. The availability of such standardized databases not only serves to save researchers from collecting and labeling data sets but also allows different systems to be compared on the same databases. The same idea led to a number of evaluation campaigns and competitions that are organized regularly.

Like other scientific domains, a number of standard databases have been developed by the document recognition community, handwriting databases being the most popular ones. Well-known and widely used offline handwritten databases include CEDAR [8], IAM [10], RIMES [2], CVL [3], QUWI [7] and IFN/ENIT [6] etc. With the tremendous increase in the usage of digitizing tablets and hand held devices, online handwriting recognition and related problem areas have also been increasingly researched.

Popular online handwriting databases include IAM-OnDB [11, 12], UNIPEN [13] and IRONOFF [14]. With the exception of QUWI [7] and CVL [3] databases, all of these databases are uni-script where the writers contribute by providing there writing samples in a given script. Recently, research in analysis of handwriting in a multi-script environment has gained significant interest. Studying the writing patterns which are stable across different scripts produced by the same writer is an interesting subject and the principle motivation of developing a multi-script database presented in this paper.

The database, termed as LAMIS-MSHD is collection of Arabic and French documents including handwritten text, signatures and digits. The database mainly targets writer identification and verification in a multi-script environment and can also be effectively used to evaluate systems like handwriting recognition, script recognition and signature verification etc. The database comprises 1300 handwritten forms contributed by 100 writers and will be presented in detail in the subsequent sections. Section 2 of the paper presents an overview of well-known handwritten databases. The different statistics of the LAMIS-MSHD database are presented in Section 3 followed by a discussion on the ground truth preparation of the database. Finally, Section 5 concludes this paper with a discussion.

II. RELATED DATABASES

As discussed earlier, a number of handwritten databases have been developed by the document recognition community to allow meaningful comparison of different systems. This section provides an overview of some well-known handwriting databases.

The CEDAR database (2002) [8] developed by the University of Buffalo is considered as one of the first large handwriting databases. The database comprises a letter copied thrice by 1567 individuals representing the US population. The database was mainly developed for text dependent writer identification and verification but has also been employed for preprocessing tasks like character segmentation. The database, however, is not publically available.

The IAM Database (2002) [10] is easily the most widely used database that has long been used for evaluation of writer identification, handwriting recognition and similar related tasks. The database comprises digitized offline documents in English while the content is extracted from the corpus "Lancaster-Oslo/Bergen" (LOB). The database contains 1539 images of text written by 657 different writers. This database is publicly available and is supported by detailed ground truth data allowing evaluation of a number of segmentation and recognition tasks.

The IFN/ENIT (2002) [5, 6] is the most popular Arabic handwriting database which consists of 2200 handwritten samples contributed by 411 different writers. The text in the forms comprises names of 937 Tunisian towns/villages making a total of more than 26,000 words. The database was mainly developed for Arabic handwriting recognition but has also been used to evaluate Arabic writer identification systems.

The French offline RIMES database (2006) [2] is a collection of handwritten mails contributed by 1300 individuals each writing up to 5 mails. The complete database comprises 12,723 pages corresponding to 5605 letters of two to three pages. A number of competitions in conjunction with ICFHR 2008, ICDAR 2009 and ICDAR 2011 have also been organized using this database.

An Arabic database mainly targeting recognition of Arabic handwritten checks was developed by AlAmri et al. [4]. This database contains Arabic dates, isolated digits, numerical strings, letters, words and some special symbols. This database has been mainly used for recognition of Arabic handwritten numbers, digits and a limited vocabulary of words.

CASIA (2011) [1] is a well-known Chinese database that contains samples of isolated Chinese characters and Chinese handwritten texts. These samples have been produced by 1020 different writers using an Anoto electronic pen. Samples of the database are divided into six groups, three for isolated characters and three for handwritten texts. This database contains approximately 5090 pages and 1.35 million character samples. This database can be used for character recognition, handwritten text recognition and writer identification.

KHATT (2012) [9] is a comprehensive Arabic offline database comprising writing samples of 1000 distinct writers coming from different cultural surroundings, age groups, educational backgrounds and gender etc. Each writer filled a form of 4 pages scanned at 200, 300, 600 dpi. The database is publically available and can be used for evaluation of writer identification, text segmentation and text recognition systems.

An interesting multi-script database, QUWI (2012) [8] contains writing samples in Arabic as well as English produced by 1017 volunteers with diverse demographics. The database has been employed for offline writer identification and, gender, age and handedness classification.

A relatively recent database, CVL (2013) [3] is a collection of cursively written German and English texts. The database contains 2160 writing samples of 310 different writers and can be used for writer retrieval, writer identification, word spotting and text recognition. A summary of the databases discussed above is presented in Table 1.

Table 1 A summary of handwritten databases

Database	Number of forms	Number of writers	Public	Script
CEDAR [8]	4701	1567	No	Latin
IAM [10]	1539	657	Yes	Latin
RIMES [2]	12723	1300	No	Latin
Al-Amri [4]	656	328	No	Arabic
CVL [3]	2163	310	Yes	Latin
KHATT [9]	4000	1000	Yes	Arabic
CASIA [1]	5090	1020	Yes	Chinese
IFN/ENIT [6]	2200	411	Yes	Arabic
QUWI [7]	4068	1017	No	Arabic & Latin

In the next section, we present a detailed description of the LAMIS-MSHD, its collection, general structure and the corresponding statistics.

III. OVERVIEW OF THE LAMIS- MSHD

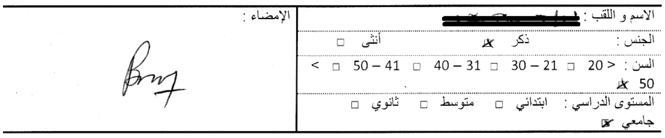
The newly developed LAMIS-MSHD (multi-script handwritten database) comprises forms with content from 13 different sources. The first category includes 20 isolated Arabic digits and 19 numerical strings each of length 10. The remaining 12 sources comprise 6 different subjects each in Arabic and French. The content of these forms and the number of corresponding words is summarized in Table 2.

The forms have been filled by 100 randomly selected volunteers in Tebessa (Algeria). These individuals included male and female Arabic and French writers with different ages and educational backgrounds. These included middle school students, secondary school students and university students. Each writer filled 13 forms making a total of 1300 forms scanned at 300 dpi.

On each of the 13 forms, the individuals provided their personal information including name, age group, gender, education level. The contributors were also asked to provide their signatures in the specified space. An instance of this information is illustrated in Figure 1. The writing samples were collected by asking the writers to copy the predefined text on each form, six in Arabic, six in French and one form with isolated digits and numerical strings. All writings were produced with a black or a blue pen.

Table 2 Summary of the content of forms

Form ID	Topic	*Content*	*Content count*
1	//	Handwritten Arabic Digits	211 Digits
2	Medicine & Health	Arabic Text	123 Words
3	Economy & Business	Arabic Text	88 Words
4	Art & Culture	Arabic Text	109 Words
5	Ecology & Environment	Arabic Text	152 Words
6	Sports	Arabic Text	128 Words
7	Data processing	Arabic Text	147 Words
8	Medicine & Health	French Text	124 Words
9	Economy & Business	French Text	109 Words
10	Art & Culture	French Text	97 Words
11	Ecology & Environment	French Text	104 Words
12	Sports	French Text	94 Words
13	Data processing	French Text	98 Words



الاسم و اللقب : [redacted]

الجنس : ذكر ☒ أنثى □

السن : 20 > □ 30 – 21 □ 40 – 31 □ 50 – 41 □ <
50 ☒

المستوى الدراسي : ابتدائي □ متوسط □ ثانوي □
جامعي ☒

الإمضاء :

Fig. 1. Writer information at the bottom of each form

The 1300 forms in the LAMIS-MSHD database contain approximately 74,700 words in Arabic and 62,600 words in French with an average of around 1375 words per writer. Moreover, the database also contains a total of 21,000 digits comprising 20 isolated digits and 19 numerical strings each of length 10. The numerical strings allow having different combinations of digits after one another and also include overlapping and connected digits (Figure 2). In addition, the database also contains 1300 genuine signatures of 100 different individuals an example of signatures being illustrated in Figure 3. A summary of the statistics of the database along with the demographic distribution of contributors is presented in Table 3 while sample images from the database are presented in Figure 4.

IV. GROUND TRUTH DATA

One of the integral components of any database is the accompanying ground truth data which in fact determines the tasks that could be evaluated using the database in question. The ground truth data for the LAMIS-MSHD was manually generated in text format. In addition to the writer information including their ID, age and gender, the ground truth data also contains at line and paragraph levels the text written by the writers allowing its usage in a variety of applications.

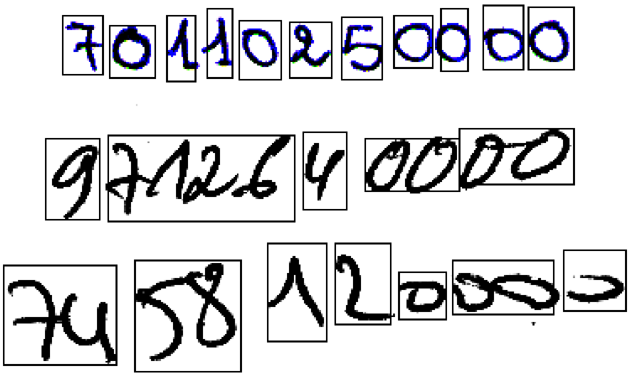


Fig. 2. Sample instances of digits in the database

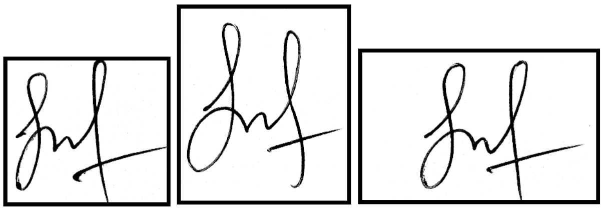
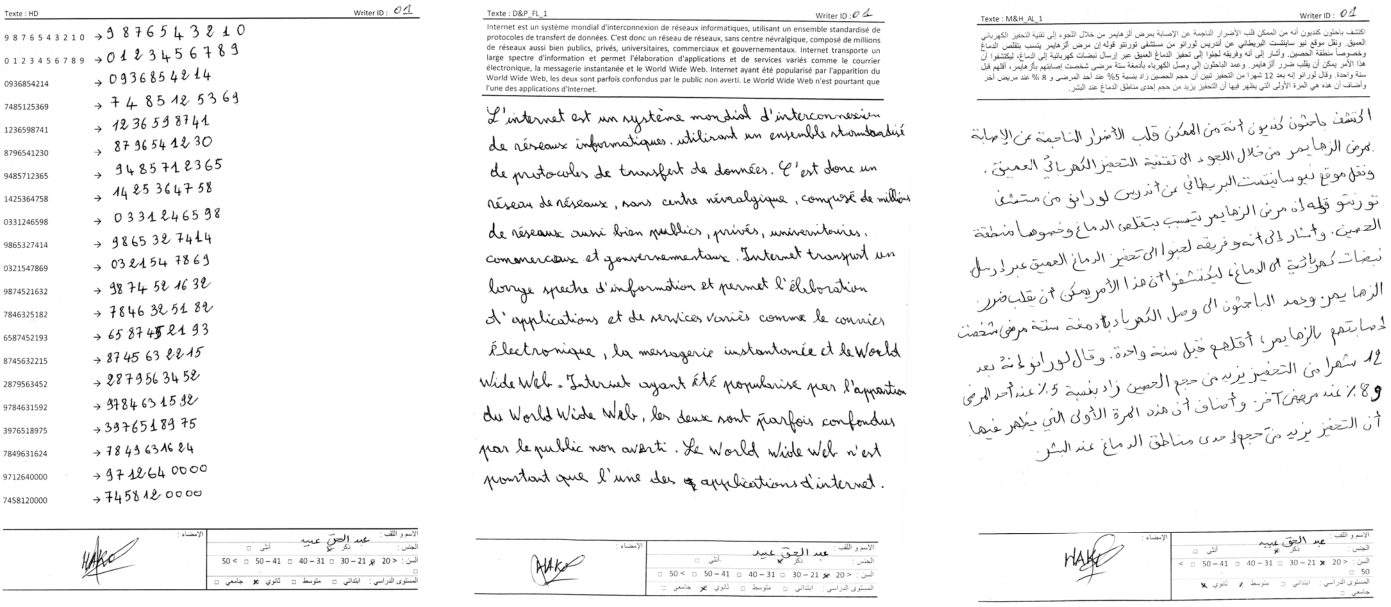


Fig. 3. Three signatures of an individual in the database

Table 3 Statistics of the database

		Total number	
Form Content	***Signatures***	1300	
	Digits	21000	
	Arabic Words	74700	
	French Words	62600	
Gender	***Male***	43	100 Writers
	Female	57	
Age group	***< 20***	28	100 Writers
	21-30	57	
	31-40	8	
	41-50	4	
	>50	3	
Educational background	***Middle school students***	2	100 Writers
	Secondary school students	13	
	University students	85	



Texte : HD

Writer ID : 01

9 8 7 6 5 4 3 2 1 0 → 9 8 7 6 5 4 3 2 1 0
0 1 2 3 4 5 6 7 8 9 → 0 1 2 3 4 5 6 7 8 9
0936854214 → 0936854214
7485125369 → 7485125369
1236598741 → 1236598741
8796541230 → 8796541230
9485712365 → 9485712365
1425364758 → 1425364758
0331246598 → 0331246598
9865327414 → 9865327414
0321547869 → 0321547869
9874521632 → 9874521632
7846325182 → 7846325182
6587452193 → 6587452193
8745632215 → 8745632215
2879563452 → 2879563452
9784631592 → 9784631592
3976518975 → 3976518975
7849631624 → 7849631624
9712640000 → 9712640000
7458120000 → 7458120000

الاسم و اللقب : عبد الحق عبيد
الجنس : ذكر ☒ أنثى □
السن : 20 > ☒ 30 – 21 □ 40 – 31 □ 50 – 41 □ 50 < □
المستوى الدراسي : ابتدائي □ متوسط □ ثانوي ☒ جامعي □
الإمضاء :

(a)

Texte : D&P_FL_1

Writer ID : 01

Internet est un système mondial d'interconnexion de réseaux informatiques, utilisant un ensemble standardisé de protocoles de transfert de données. C'est donc un réseau de réseaux, sans centre névralgique, composé de millions de réseaux aussi bien publics, privés, universitaires, commerciaux et gouvernementaux. Internet transporte un large spectre d'information et permet l'élaboration d'applications et de services variés comme le courrier électronique, la messagerie instantanée et le World Wide Web. Internet ayant été popularisé par l'apparition du World Wide Web, les deux sont parfois confondus par le public non averti. Le World Wide Web n'est pourtant que l'une des applications d'Internet.

L'internet est un système mondial d'interconnexion
de réseaux informatiques. utilisant un ensemble standardisé
de protocoles de transfert de données. C'est donc un
réseau de réseaux, sans centre névralgique, composé de millions
de réseaux aussi bien publics, privés, universitaires,
commerciaux et gouvernementaux. Internet transport un
large spectre d'information et permet l'élaboration
d'applications et de services variés comme le courrier
électronique, la messagerie instantanée et le World
Wide Web. Internet ayant été popularisé par l'apparition
du World Wide Web, les deux sont parfois confondus
par le public non averti. Le World Wide Web n'est
pourtant que l'une des ~~a~~ applications d'internet.

الاسم و اللقب : عبد الحق عبيد
الجنس : ذكر ☒ أنثى □
السن : 20 > ☒ 30 – 21 □ 40 – 31 □ 50 – 41 □ 50 < □
المستوى الدراسي : ابتدائي □ متوسط □ ثانوي ☒ جامعي □
الإمضاء :

(b)

Texte : M&H_AL_1

Writer ID : 01

اكتشف باحثون كنديون أنه من الممكن قلب الأضرار الناجمة عن الإصابة بمرض ألزهايمر من خلال اللجوء إلى تقنية التحفيز الكهربائي العميق. ونقل موقع نيو ساينتست البريطاني عن أندريس لورانو من مستشفى تورنتو قوله إن مرض ألزهايمر يتسبب بتقلص الدماغ وخصوصاً منطقة الحصين. وأشار إلى أنه وفريقه لجئوا إلى تحفيز الدماغ العميق عبر إرسال نبضات كهربائية إلى الدماغ، ليكتشفوا أن هذا الأمر يمكن أن يقلب ضرر ألزهايمر. وعمد الباحثون إلى وصل الكهرباء بأدمغة ستة مرضى شخصت إصابتهم بألزهايمر، أقلهم قبل سنة واحدة. وقال لورانو إنه بعد 12 شهرا من التحفيز تبين أن حجم الحصين زاد بنسبة 5% عند أحد المرضى و 8 % عند مريض آخر. وأضاف أن هذه هي المرة الأولى التي يظهر فيها أن التحفيز يزيد من حجم إحدى مناطق الدماغ عند البشر.

اكتشف باحثون كنديون أنه من الممكن قلب الأضرار الناجمة عن الإصابة
بمرض الزهايمر من خلال اللجوء الى تقنية التحفيز الكهربائي العميق.
ونقل موقع نيوساينتست البريطاني عن أندريس لورانو من مستشفى
تورنتو قوله ان مرض الزهايمر تتسبب بتقلص الدماغ وخصوصا منطقة
الحصين. وأشار الى أنه وفريقه لجئوا الى تحفيز الدماغ العميق عبر ارسال
نبضات كهربائية الى الدماغ، ليكتشفوا أن هذا الأمر يمكن أن يقلب ضرر
الزهايمر وعمد الباحثون الى وصل الكهرباء بأدمغة ستة مرضى شخصت
اصابتهم بالزهايمر، أقلهم قبل سنة واحدة. وقال لورانو انه بعد
12 شهرا من التحفيز يزيد من حجم الحصين زاد بنسبة 5% عند أحد المرضى
و 8% عند مريض آخر. وأضاف أن هذه المرة الأولى التي يظهر فيها
أن التحفيز يزيد من حجم احدى مناطق الدماغ عند البشر.

الاسم و اللقب : عبد الحق عبيد
الجنس : ذكر ☒ أنثى □
السن : 20 > ☒ 30 – 21 □ 40 – 31 □ 50 – 41 □ 50 < □
المستوى الدراسي : ابتدائي □ متوسط □ ثانوي ☒ جامعي □
الإمضاء :

(c)

Fig 4. Three samples of a writer in the database. (a): Isolated digits and numerical strings (b): French text (c): Arabic text

The key contribution of the newly developed database is providing a multi-script environment especially for evaluation of systems like writer identification, writer verification and, gender and (or) age classification. All of these problems have been mostly researched on writing samples in a single script. It would be very interesting to expose these systems to a true multi-script environment where the researchers could seek for stable attributes of writing which do not vary with the script. In an extreme case, the training and test samples in any of the aforementioned applications could come from different scripts. Having six samples in each script and same text copied by all writers allows the database to be used in text-dependent as well as text-independent modes. Furthermore, the database can also be employed for the traditional tasks of script recognition, different levels of segmentation and signature verification.

V. CONCLUSIONS AND FUTURE WORKS

In this paper we presented a newly developed database, the LAMIS-MSHD, that contains off-line handwritten Arabic and French handwritings contributed by 100 volunteers of different ages, gender, and educational backgrounds. Each writer filled a total of 13 forms, one comprising digits, six having French and six having Arabic text. All forms were scanned as true color images at a resolution of 300 dpi.

The ground truth of the database recorded the writer information as well as the textual content at paragraph and line levels. It is expected that that database would contribute to serve the researchers working in areas of writer recognition and writer demographic classification, signature verification and traditional tasks related to handwriting recognition. The next version of the database is likely to include an increased number of writers with annotation related to the location of lines and words. The database will also be made available publically.

REFERENCES

[1] C.L. Liu, f. Yin, D. Wang, Q. Wang, “CASIA Online and Offline Chinese Handwriting Databases, ” In Proc of the 11th International Conference on Document Analysis and Recognition (ICDAR 2011), Beijing, China, pp. 37-41, 2011.

[2] E. Augustin, M. Carré, G. E., J. M. Brodin, E. Geoffrois, and F. Preteux, “Rimes evaluation campaign for handwritten mail processing,” In Proceedings of the Workshop on Frontiers in Handwriting Recognition, pp. 231–235 , 2006.

[3] F. Kleber, S. Fiel, M. Diem, and R. Sablatnig, “CVL-Database: An Off-line Database for Writer Retrieval, Writer Identification and Word Spotting,”In Proceedigns of the 12th International Conference on Document Analysis and Recognition (ICDAR 2013), Washington, USA, pp. 560 - 564 ,2013.

[4] H. Alamri, J. Sadri, C. Y. Suen, and N. Nobile, “A novel comprehensive database for Arabic off-line handwriting recognition,” Proceedings of the 11 th International Conference on Frontiers in Handwriting Recognition, ICFHR 2008, pp. 664-669, 2008.

[5] H. El Abed and V. Märgner, “The IFN/ENIT-database - a tool to develop Arabic handwriting recognition systems,” in 9th International Symposium on Signal Processing and Its Applications. ISSPA 2007, pp.1-4, 2007.

[6] M. Pechwitz, S. S. Maddouri, V. Märgner, N. Ellouze, and H. Amiri, “IFN/ENIT - Database of Handwritten Arabic Words,” in 7th Colloque International Francophone sur l’Ecrit et le Document , CIFED 2002, pp. 129-136 , 2002.

[7] S. Al-Maadeed, W. Ayouby, A. Hassaine, and J. Aljaam, “QUWI: An Arabic and English Handwriting Dataset for Offline Writer Identification, ”In Proc of the 13th International Conference on Frontiers in Handwriting Recognition, ICFHR 2012, pp. 742-747,2012.

[8] S. Srihari, SH. Cha, H. Arora, S. Lee., “Individuality of handwriting, ”In Journal of forensic sciences. vol. 47, pp. 856-872, 2002.

[9] S.A. Mahmoud, A. Ahmad, M. Alshayeb, W.G. Al-Khatib, M.T. Parvez, G.A. Fink, V. Margner, and H EL Abed, “KHATT: Arabic Offline Handwritten Text Database,” In 13th International Conference on Frontiers in Handwriting Recognition, ICFHR 2012, pp. 447- 452,2012.

[10] U. Marti and H. Bunke, “The IAM-database: An English Sentence Database for Off-line Handwriting Recognition, ” In International Journal on Document Analysis and Recognition, vol. 5, pp. 39-46, 2002.

[11] M. Liwicki and H. Bunke, “IAM-OnDB - An on-line English sentence database acquired from handwritten text on a whiteboard”, Proc. of the Eighth International Conference on Document Analysis and Recognition, 2005.

[12] E. Indermühle, M. Liwicki, and H. Bunke, “IAMonDo database: an online handwritten document database with non-uniform contents”, Proc. of the 9th IAPR International Workshop on Document Analysis Systems, 2010.

[13] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman, and S. Janet, “Unipen project of on-line data exchange and recognizer benchmarks”, Proc. of the 12th International Conference on Pattern Recognition, 1994.

[14] C. Viard-Gaudin, P. M. Lallican, P. Binter, and S. Knerr, “The ireste on/off (ironoff) dual handwriting database”, Proc. of the Fifth International Conference on Document Analysis and Recognition, 1999.