ALTID : Arabic/Latin Text Images Database for recognition research

Imen Chtourou, Ahmed Cheikh Rouhou, Faten Kallel Jaiem, Slim Kanoun
MIRACL laboratory, ISIMS, University of Sfax, Tunisia
{imene.chtourou, cheikhrouhouahmad, kallelfaten, slim.kanoun}@gmail.com

Abstract— **This paper presents a new public offline database for Arabic/Latin printed and Arabic/Latin handwriting text. The database was developed to be employed in performance evaluation, result comparison and development of new methods related to document analysis and recognition. It may be used for, script identification, font identification, writer identification and word segmentation. The printed text is scanned from 731 pages of Latin and Arabic printed documents with grayscale format and 300 dpi resolutions. After preforming a manual segmentation, we obtained 1845 Arabic text and 2328 Latin text images. The handwritten dataset includes 460 Arabic and 582 Latin text-blocks which are written by 17 individuals with different ages and educational levels. Each text image of our database is provided with a ground truth file.**

Keywords— multi-script; multi-font; offline database; handwritten document images;

I. Introduction

Research in image processing has received considerable attention in the recent years. Related applications cover script identification, font identification, writer identification and word spotting etc. Whilst the research development in the above-mentioned field, the need for standard databases to closely match the real world scenarios has been increased. Such standardized databases serve to save researchers from collecting and labeling datasets. Besides, they facilitate the comparison between different systems within evaluation campaigns and competitions.

The analysis of the literature shows that text image database differ by the language and type of script. Thus, databases can be stored in three classes: printed, handwritten and hybrid (contain printed and handwritten script in the same text).

Within the first class, there exist many printed databases such as: The UNLV dataset [1] which contain 2889 pages of scanned document images from a large variety of sources ranging from technical reports and business letters to newspapers and magazines. The dataset was specifically created to analyze the performance of leading commercial optical character recognition (OCR) systems in the UNLV annual tests of OCR accuracy [2]. The scanned images are provided at 200 and 300 DPI resolution in bitonal, grey and fax format. The ground truths with manually marked zones are provided in text format.

UW-III dataset [3] is the third in a series of UW document image databases. It contains a total of 1600 English document images randomly selected from scientific and technical journals with manually edited ground-truth of entity bounding boxes. These bounding boxes enclose page frame, text and non-text zones, text lines, and words. The type of each zone (text, math, table, half-tone,…) is also marked. The dataset is mainly used for document layout analysis.

The Media Team Oulu [4] Document Database, which is a collection of 500 scanned document images with corresponding ground truth for the physical and logical structure of the documents. It was developed by the University of Oulu Media Team. The images contain a wide range of document types including journal papers, maps, newsletters, form, music, dictionaries and can be used for comparing various tasks in Document Analysis and Recognition (DAR).

APTI [5] is a large Arabic Printed Text Images database. This database was created at 2009 and was the first public database with large vocabulary and on very low-resolution (72 dpi). It is used for the recognition of multi-font, multi-size, and multi-style Arabic text. APTI is synthetically generated using a lexicon of 113 284 Arabic words in 10 fonts, 10 sizes and 4 styles which made it suitable for the evaluation of screen-based OCR systems.

APTID/MF [6] (Arabic Printed Text Image Database /Multi-Font) may be used for word segmentation and font identification research. It contains 1845 text-blocks images which are scanned at 300 dpi resolutions in grayscale format and 27402 characters images which are in the character dataset. Such database could make a good contribution in the Arabic printed text recognition field by organizing competitions.

Last but not least, KAFD [7] (King Fahd University Arabic Font Database) is a multi-font, multi-size, multi-style, and multi-resolution database. It consists of 40 Arabic fonts. Each font in this database consists of its unique text. Each font consists of 10 font sizes (8, 9, 10, 11, 12, 14, 16, 18, 20, and 24 points) and four font styles (normal, bold, italic, and bold-italic). It is scanned in four resolutions (100, 200, 300, and 600 dpi) and two forms (page and line). KAFD database is organized into three sets (training, testing, and validation). It

978-1-4799-1805-8/15/$31.00 ©2015 IEEE

consists of 115 068 page images and 2 576 024 line images. In addition, KAFD database is made freely available to researchers. The ground truth at the page and line levels is included therefore it may be also used for multi-font Arabic Text Recognition.

For the second class, which contains handwritten databases, there exist: The CEDAR database [8]. It represents one of the first large handwriting databases. It contains a handwritten English letter copied by as many as 1500 writers. This database is mainly used in text dependent writer identification and verification. Moreover, CEDAR could be used for preprocessing tasks like character segmentation. Unfortunately, the database is not publicly available.

The IAM database [9] is the most commonly used database for writer identification as well as handwriting recognition and other related tasks. It comprises digitized offline English documents. The database contains 1539 images of text written by 657 different writers. The IAM database is publicly available with a detailed ground truth data for the evaluation.

The IFN/ENIT [10,11] database of handwritten Arabic words (Tunisian town names) is probably the most widely used database. It is made of Tunisian town/village names written by 411 writers. The text in the forms comprises names of 937 Tunisian towns/villages making a total of more than 26 000 words. The mainly use of the database is the Arabic handwriting recognition but has also been used to evaluate Arabic writer identification systems.

RIMES [12] is a collection of French handwritten mails made by 1300 individuals. Each of them writes 5 mails. The complete database comprises 12 723 pages corresponding to 5605 letters of two to three pages.

KHATT [13] is a comprehensive Arabic offline database comprising writing samples of 1000 distinct writers from different countries. Each writer filled a form of 4 pages scanned at 200,300,600 dpi. The database is made freely available to researchers world-wide to use it for research in various handwritten related problems such as text recognition, writer identification and verification, forms analysis, preprocessing, segmentation, etc.

QUWI [14] is an interesting multi-script database which contains writing samples in Arabic as well as English made by 1017 volunteers with diverse demographics. The database has been employed for offline writer identification and gender, age and handedness classification.

LAMIS-MSHD [15] is a new offline handwriting database produced by 100 different Algerian individuals. It comprises 600 Arabic and 600 French text samples, 1300 signatures and 21 000 digits. The database may be used in areas of writer recognition and writer demographic classification, signature verification and other tasks related to handwriting recognition.

There exists also hybrid class which contain printed and handwritten script in the same text like: Maurdor database [16], it is based on a corpus of heterogeneous documents. The training corpus MAURDOR 2013 includes a total of 2500 documents in English, French and Arabic within the following categories: Blank forms or completed, printed business documents which are manually commented, handwritten private correspondence sometimes may contain printed headers, commercial printed correspondence which are manually commented, other documents such as newspaper articles or maps. It is necessary to mention that this type of class is only suitable for script identification and not for font identification.

A deep observation on the state-of-the-art of existing databases shows that there is no standard corpus which includes multi-font, multi-size and multi-script printed and handwritten text. Consequently, this paper presented a detailed description of Arabic/Latin multi-font, multi-size and multi-script printed/handwritten text database.

Our database is a collection of Arabic/Latin printed text and Arabic/Latin documents including handwritten text. The database may enrich APTID/MF database. It mainly targets writer identification and verification in a multi-script environment and also can be effectively used to evaluate discrimination between Arabic/Latin script, handwriting recognition, etc.

This paper is organized as follows; section 2 of the paper presents our database with a detailed description of printed and handwritten datasets followed by statistics and ground truth preparation of our database. Finally, Section 3 concludes this paper and talk about some future work.

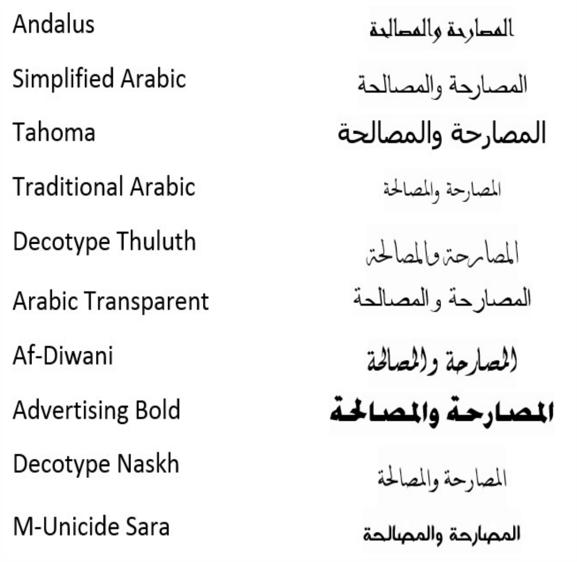
II. Overview of ALTID Database

In this section, we present our database of Arabic/Latin printed/handwritten text. The objective of our work is to enrich the APTID/MF (Arabic Printed Text Image Database/Multi-Font) database [6] with Latin multi-font and multi-size printed text as well as handwritten Arabic/Latin text.

A. Printed Text Image Dataset

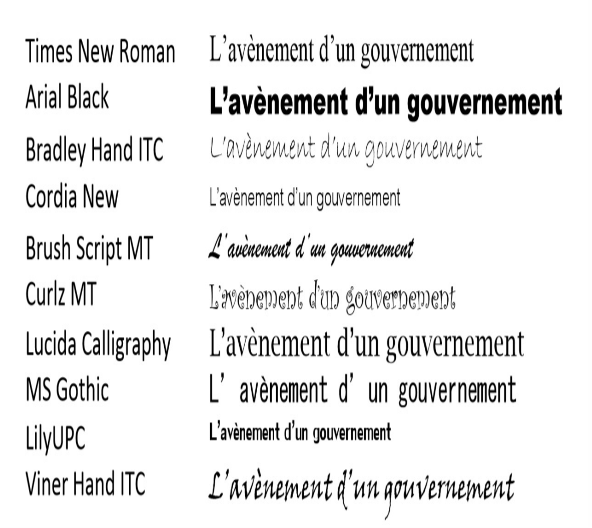
The printed text dataset include the Arabic text images dataset of APTID/MF database and a Latin text images. The APTID/MF includes 387 document pages organized in 10 fonts presented in Fig 1, and 4 sizes, segmented in 1845 text block images. The printed Arabic text images were created in height-resolution of 300 dpi. For more description of APTID/MF, we refer to [6].

In order to create the Latin printed text images dataset, we used the same process applied to create text dataset of APTID/MF. The Latin document was selected from the official site of the newspaper “le Temps”. The document pages were saved in 10 Latin fonts enumerated in Fig 2 (Times New Roman, Arial Black, Bradley Hand ITC, Cordia New, Brush Script MT, Curlz MT, Lucida Calligraphy, MS Gothic, LilyUPC, Viner Hand ITC). These set of document were written with 4 different sizes (12, 14, 16 and 18 points). These figures below present the different used Arabic and Latin fonts.



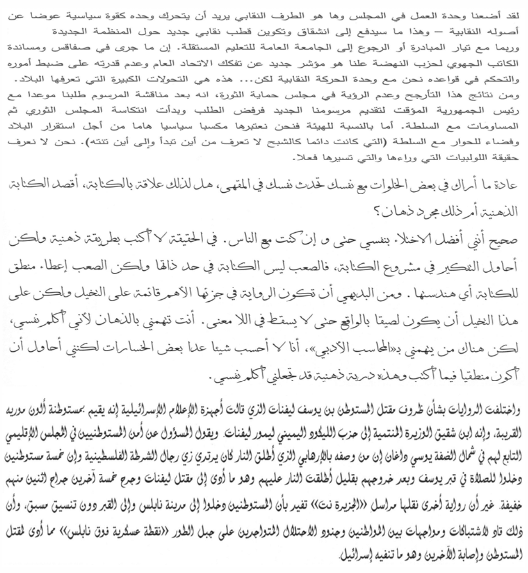
Andalus	المصارحة والمصالحة
Simplified Arabic	المصارحة والمصالحة
Tahoma	المصارحة والمصالحة
Traditional Arabic	المصارحة والمصالحة
Decotype Thuluth	المصارحة والمصالحة
Arabic Transparent	المصارحة و المصالحة
Af-Diwani	المصارحة والمصالحة
Advertising Bold	المصارحة والمصالحة
Decotype Naskh	المصارحة والمصالحة
M-Unicide Sara	المصارحة والمصالحة

Fig. 1. 10 Arabic fonts used in APTID/MF [6]



Times New Roman	L’avènement d’un gouvernement
Arial Black	L’avènement d’un gouvernement
Bradley Hand ITC	L’avènement d’un gouvernement
Cordia New	L’avènement d’un gouvernement
Brush Script MT	L’avènement d’un gouvernement
Curlz MT	L’avènement d’un gouvernement
Lucida Calligraphy	L’avènement d’un gouvernement
MS Gothic	L’ avènement d’ un gouvernement
LilyUPC	L’avènement d’un gouvernement
Viner Hand ITC	L’avènement d’un gouvernement

Fig. 2. 10 Latin fonts used in ALTID



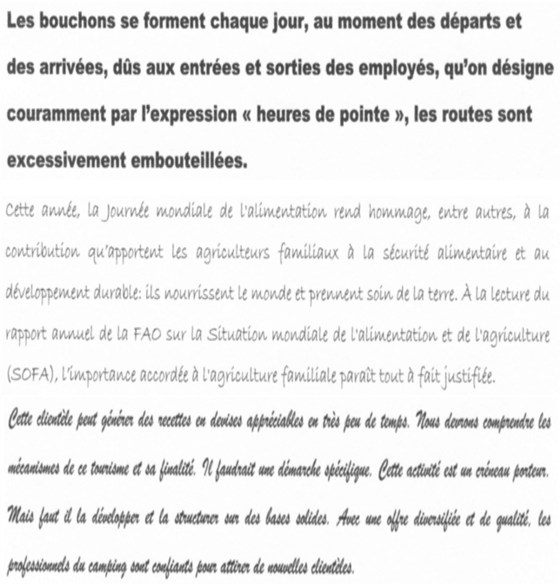
لقد أضعنا وحدة العمل في المجلس وها هو الطرف النقابي يريد أن يتحرك وحده كقوة سياسية عوضا عن
أصوله النقابية – وهذا ما سيدفع إلى انشقاق وتكوين قطب نقابي جديد حول المنظمة الجديدة
وربما مع تيار المبادرة أو الرجوع إلى الجامعة العامة للتعليم المستقلة. إن ما جرى في صفاقس ومساندة
الكاتب الجهوي لحزب النهضة علنا هو مؤشر جديد عن تفكك الاتحاد العام وعدم قدرته على ضبط أموره
والتحكم في قواعده نحن مع وحدة الحركة النقابية لكن... هذه هي التحولات الكبيرة التي تعرفها البلاد.
ومن نتائج هذا التأرجح وعدم الرؤية في مجلس حماية الثورة، انه بعد مناقشة المرسوم طلبنا موعدا مع
رئيس الجمهورية المؤقت لتقديم مرسومنا الجديد فرفض الطلب وبدأت انتكاسة المجلس الثوري ثم
المساومات مع السلطة. أما بالنسبة للهيئة فنحن نعتبرها مكسبا سياسيا هاما من أجل استقرار البلاد
وفضاء للحوار مع السلطة (التي كانت دائما كالشبح لا تعرف من أين تبدأ وإلى أين تنتهه). نحن لا نعرف
حقيقة اللولبيات التي وراءها والتي تسيرها فعلا.

عادة ما أراك في بعض الخلوات مع نفسك تحدث نفسك في المقهى، هل لذلك علاقة بالكتابة، أقصد الكتابة
الذهنية أم ذلك مجرد ذهان؟

صحيح أنني أفضل الاختلاء بنفسي حتى و إن كنت مع الناس. في الحقيقة لا أكتب بطريقة ذهنية ولكن
أحاول التفكير في مشروع الكتابة، فالصعب ليس الكتابة في حد ذاتها ولكن الصعب إعطاء منطق
للكتابة أي هندستها . ومن البديهي أن تكون الرواية في جزئها الأهم قائمة على التخيل ولكن على
هذا التخيل أن يكون لصيقا بالواقع حتى لا يسقط في اللا معنى. أنت تتهمني بالذهان لأني أكلم نفسي،
لكن هناك من يتهمني بـ«المحاسب الأدبي»، أنا لا أحسب شيئا عدا بعض الخسارات لكنني أحاول أن
أكون منطقيا فيما أكتب وهذه دربة ذهنية قد تجعلني أكلم نفسي.

واختلفت الروايات بشأن ظروف مقتل المستوطن بن يوسف ليفنات الذي قالت أجهزة الإعلام الإسرائيلية إنه يقيم بمستوطنة ألون موريه
القريبة، وإنه ابن شقيق الوزيرة المنتمية إلى حزب الليكود اليميني ليمور ليفنات. ويقول المسؤول عن أمن المستوطنيين في المجلس الإقليمي
التابع لهم في شمال الضفة يوسي داغان إن من وصفه بالإرهابي الذي أطلق النار كان يرتدي زي رجال الشرطة الفلسطينية وإن خمسة مستوطنين
دخلوا للصلاة في قبر يوسف وبعد خروجهم بقليل أطلقت النار عليهم وهو ما أدى إلى مقتل ليفنات وجرح خمسة آخرين جراح اثنين منهم
خفيفة. غير أن رواية أخرى نقلها مراسل «الجزيرة نت» تفيد بأن المستوطنين دخلوا إلى مدينة نابلس وإلى القبر دون تنسيق مسبق، وأن
ذلك قاد لاشتباكات ومواجهات بين المواطنين وجنود الاحتلال المتواجدين على جبل الطور «نقطة عسكرية فوق نابلس» مما أدى لمقتل
المستوطن وإصابة الآخرين وهو ما تنفيه إسرائيل.

(a)



Les bouchons se forment chaque jour, au moment des départs et des arrivées, dûs aux entrées et sorties des employés, qu’on désigne couramment par l’expression « heures de pointe », les routes sont excessivement embouteillées.

Cette année, la Journée mondiale de l'alimentation rend hommage, entre autres, à la contribution qu’apportent les agriculteurs familiaux à la sécurité alimentaire et au développement durable: ils nourrissent le monde et prennent soin de la terre. À la lecture du rapport annuel de la FAO sur la Situation mondiale de l'alimentation et de l'agriculture (SOFA), l’importance accordée à l'agriculture familiale paraît tout à fait justifiée.

Cette clientèle peut générer des recettes en devises appréciables en très peu de temps. Nous devrons comprendre les mécanismes de ce tourisme et sa finalité. Il faudrait une démarche spécifique. Cette activité est un créneau porteur. Mais faut il la développer et la structurer sur des bases solides. Avec une offre diversifiée et de qualité, les professionnels du camping sont confiants pour attirer de nouvelles clientèles.

(b)

Fig. 3. Examples of Printed Text: (a) Arabic text block [17], (b) Latin text block

The set of document pages are printed with two types of printer: laser printer and inkjet printer. This gives us two groups. The first one contains digitalized set of laser-printed documents scanned with an HP scanner. The second group gathers scanned inkjet-printed documents using an Epson scanner. Page images digitalization are performed using 300 dpi resolutions applying grayscale format and then stored in “JPEG” files. These images are manually segmented into text-blocks.

Fig 3 present examples of Arabic and Latin printed text respectively

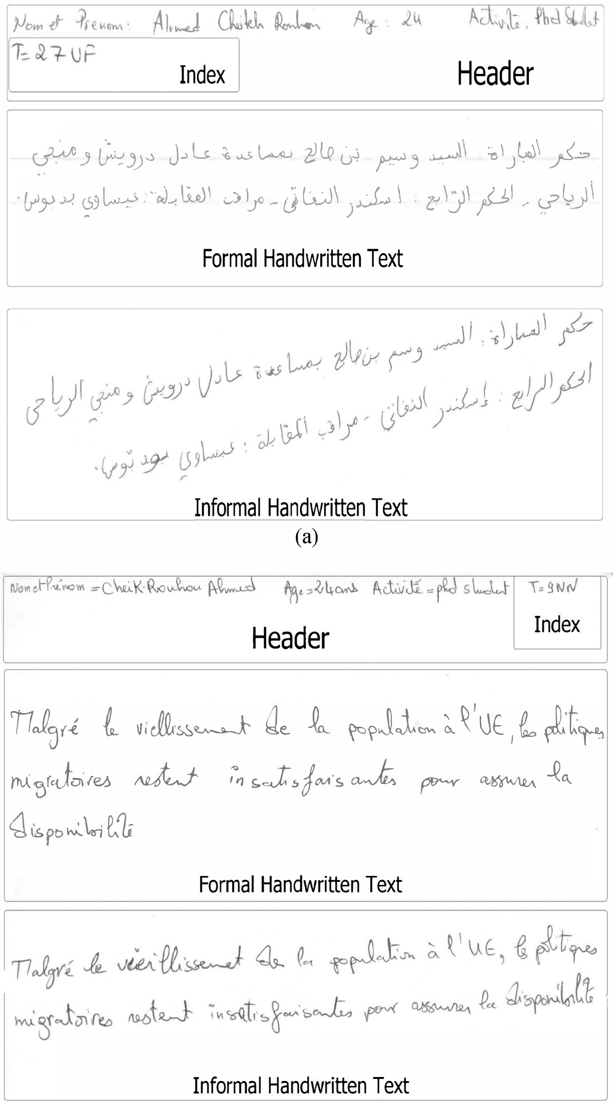
B. Handwritten Text Image Dataset

The handwritten dataset is inherited from the printed document pages. Thus, we attributed a writer for each font. As a result, we have 20 sets (10 Arabic and 10 Latin) of handwritten texts. Correspondingly, there are 17 writers as 3 among them participated in both Latin and Arabic writing task.

The volunteers are randomly selected from Sfax (Tunisia) with different age ranges, formations, activities and genders. Every participant will be assigned an index along with their names, ages and activities (Fig 4). The index is meant to organize the set from the first to the last written paragraph. Moreover, the set is affected to the volunteer by comparing their handwriting to the corresponding font/size. The resulting handwritten document may have the same disposition of words per line in the printed text.

Volunteers were asked to write each paragraph in the set in two different papers and in two different ways (Fig. 4 (a) and (b)). Further, the individual writes the text in a formalized style using a lined paper as a background template with 2cm for line spacing. Furthermore, s/he rewrites the same text naturally as informal style (freehand typing). Besides, the writers were not assisted during their contribution to our database given the massive volume of the texts in each set. In addition, they use the same pen type and color and we selected the Reynolds Blue Medium (048) Ball pen.

As far as we receive the full sets of handwritten texts, we process with the digitization of those papers with the same scanners used for the printed documents. The images were scanned at 300 dpi with grayscale output colors. Added to that, we selected the JPEG (Joint Photographic Experts Group) as bitmap format due to the high compression level and it doesn't include any other complex information. The dataset was saved and structured as following: Each writer is represented with a directory, in this one we have subdirectories for the Formal and Informal handwritten text. The paragraphs were named with their index number.


Nom et Prenom: Ahmed Chikh Rouhou Age: 24 Activité: PhD Student
T=27UF
Index
Header
حكم المباراة السيد وسيم بن صالح بمساعدة عادل درويش ومنجي
الرياحي ـ الحكم الرابع : اسكندر النفاتي ـ مراقب المقابلة: عيساوي بدبوسي.
Formal Handwritten Text
حكم المباراة: السيد وسيم بن صالح بمساعدة عادل درويش ومنجي الرياحي
الحكم الرابع: إسكندر النفاتي - مراقب المقابلة: عيساوي بودبوسي.
Informal Handwritten Text
Nom et Prénom = Cheik Rouhou Ahmed Age = 24 ans Activité = phd student
T=9NN
Header
Index
Malgré le viellissement de la population à l'UE, les politiques migratoires restent insatisfaisantes pour assurer la disponibilité
Formal Handwritten Text
Malgré le vieillissement de la population à l'UE, les politiques migratoires restent insatisfaisantes pour assurer la disponibilité
Informal Handwritten Text

(a)

(b)

Fig. 4. Samples from handwritten text: (a) Arabic text block, (b) Latin text block

C. Statistics

In total, our database contains 731 printed text pages: 387 pages from the APTID/MF database and 344 Latin text page images and 1042 handwritten text-blocks: 460 Arabic and 582 Latin text-blocks. The set of document pages are stored in 4 groups: the Arabic Printed Pages (APPage), the Latin Printed Pages (LPPage), the Arabic Handwritten Pages (AHPage) and the Latin Handwritten Pages (LHPage) datasets.

TABLE I. THE DISTRIBUTION OF PRINTED PAGE IMAGES

	12	**14**	**16**	**18**	**Total**
APPage Dataset	40	54	45	53	192
LPPage Dataset	55	72	98	119	344
Total	95	126	143	172	536

TABLE II. THE DISTRIBUTION OF HANDWRITTEN PAGE IMAGES

	Formal text	**Informal text**	**Total**
AHPage Dataset	230	230	460
LHPage Dataset	291	291	582
Total	521	521	1042

The printed page images are divided into text blocks, the segmentation phase gave us an 1845 Arabic Printed Text images, 2328 Latin Printed Text images. For the handwritten page images, each document present one text block. We have 460 Arabic Handwritten Text images and 582 Latin Handwritten Text images. In total, our database included 2328 text-blocks images. The table below show the distribution of Latin Printed Text Dataset, for the Arabic Printed Text Dataset we refer to [6].

TABLE III. THE DISTRIBUTION OF LATIN PRINTED TEXT DATASET

A laser printer and an HP scanner					
Font	***Size 12***	***Size 14***	***Size 16***	***Size 18***	***Total***
Times New Roman	28	28	28	28	112
Arial Black	30	30	30	30	120
Bradley Hand ITC	28	28	28	28	112
Cordia New	27	27	27	27	108
Brush Script MT	29	29	29	29	116
Curlz MT	29	29	29	29	116
Lucida Calligraphy	32	32	32	32	128
MS Gothic	30	30	30	30	120
LilyUPC	30	30	30	30	120
Viner Hand ITC	28	28	28	28	112
A inkjet printer and an Epson scanner					
Font	***Size 12***	***Size 14***	***Size 16***	***Size 18***	***Total***
Times New Roman	28	28	28	28	112
Arial Black	30	30	30	30	120
Bradley Hand ITC	28	28	28	28	112
Cordia New	27	27	27	27	108
Brush Script MT	29	29	29	29	116
Curlz MT	29	29	29	29	116
Lucida Calligraphy	32	32	32	32	128
MS Gothic	30	30	30	30	120
LilyUPC	30	30	30	30	120
Viner Hand ITC	28	28	28	28	112
TOTAL	**582**	**582**	**582**	**582**	**2328**

D. Ground Truth File description.

An essential component of any database is the presence of a ground truth data. For our database, we have generated a metadata files (XML file). These files present the ground-truth value of each sample of text image database, these files are described at the text-block and line level using XML file.

At the text-block level, these XML files include the following information: the text-block name (<TextImage Id = …..>) and the number of lines and words presented in the text-block (<text nbligne= …. nbword=…..>).

At the line level, these XML files include the following information: the id of line and the number of words in the line (<ligne Id= …. Nbword=…> and the id and word's value in the line (<word Id=… value=…./>).

Also, these Xml files give a presentation of the font (<Font name=…/>), the style (<Style name=…/>) and the size (<Size value=…../>) of the text-block, the type of printer used (<Imprimant name=…/>) and the name of the scanner used (<Scanner name=…./>) .

III. Conclusion

In this paper, we presented a novel database which contains off-line Arabic/Latin printed text and Arabic/Latin handwritten text. The Arabic printed text is the text image dataset of APTID/MF database, the Latin printed text was prepared using the same process applied to create APTID/MF database. Arabic/Latin handwritten texts were made by 17 volunteers with different age and educational level. Handwritten texts were written in two ways formal and informal. All texts were scanned as grayscale images at a high resolution of 300 dpi. As a future work, we will test our proposed database for script, font and writer identification. In addition, an evaluation of multi- methods such as texture analysis technique on this database will be realized for script, font and writer identification.

References

[1] http://www.isri.unlv.edu/ISRI/OCRtk

[2] S. V. Rice, F. R. Jenkins, and T. A. Nartker. The fourth annual test of OCR accuracy. Technical report, Information Science Research Institute, University of Nevada, Las Vegas, 1995.

[3] I. T. Phillips. User's reference manual for the UW English/Technical Document Image Database III. Technical report, Seattle University, Washington, 1996.

[4] J. Sauvola and H. Kauniskangas. Media Team Document Database II, a CD-ROM collection of document images, University of Oulu, Finland, 1999.

[5] F. Slimane, R. Ingold, S. Kanoun, M. A. Alimi, J. Hennebert, " A New Arabic Printed Text Image Database and Evaluation Protocols", International Conference on Document Analysis and Recognition, ICDAR 2009, pp. 946-950, 2009

[6] F.K. Jaiem, S. Kanoun, M. Khemakhem, H. El Abed, and J. Kardoun, "Database for Arabic Printed Text Recognition Research," ICIAP 2013, Part I, LNCS 8156, pp. 251–259, 2013

[7] H. Luqman, S. A. Mahmoud and S. Awaida, "KAFD Arabic font database", Pattern Recognition. vol. 47, no. 6, pp. 2231–2240, 2014.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

[15] C. Djeddi, A. Gattal, L.S.Meslati, I.Siddiqi, Y. Chibani, H.El Abed, "LAMIS-MSHD: A Multi-Script offline Handwriting Database", In International Conference on Frontiers in Handwriting Recognition, ICFHR 2014, pp. 93-97,2014.

[16] S. Brunessaux, P. Giroux, B. Grilheres, M. Manta, M. Bodin, K. Choukri, O. Galibert, and J. Kahn, "The Maurdor project - improving automatic processing of digital documents," in Proc. DAS, 2014, pp. 349–354, 2014.

[17] F.K. Jaiem, S. Kanoun, V. Eglin "Arabic font recognition based on a texture analysis".ICFHR 2014, pp. 673-677, 2014.