A Word Image Matching: Case of the Official Bulletin of Kingdom of Morocco

Anass Smaili, Ali Lasfar, Mohamed Sbihi
LASTIMI
EST SALE
Rabat, Morocco
anass.smaili@yahoo.fr

ABSTRACT: *The technique of optical character recognition has a particular research interest that laboratories continue to improve and develop. This technique is a good alternative to entering text, the availability of various tools involved in the singularity of this service. Undoubtedly, the development of such tools will provide a considerable gain in time and costs.*

The basis of the Arabic language, prompted us to work hard for a successful control of its optical character recognition, otherwise it is necessary to develop another approaches, therefore an approach based on finding a word-image in a text-image. Finally, the objective of this work is to conduct research word-image in text image of the Moroccan official bulletin.

Keywords: OCR, Arabic Language Processing, Character Recognition, Character, Matching

Received: 19 August 2012, Revised 10 October 2012, Accepted 18 October 2012

1. Introduction

In the work environment where Arabic is the main language, Morocco government has huge library resources. Therefore, exploitation of these resources to fill the database requires, in this case, a digitization operation and conversion to text by using an OCR system or typing information directly into a computer, a task that is very tedious and requires a word-processing specialist team.

There are many tools bound for Optical Arabic Character Recognition (OACR). These programs often provide a satisfactory result but not in all cases. Thereafter, a checkout and a review are requires of this recognition operation. In short, offer a system that let you search a word-image in a text-image is the main goal of this study.

We will start by presenting our motivation, an overview of the Arabic language, the research on the OACR, the related work and then we propose the word-image system searching. In conclusion, we will propose the outlook for such research.

2. The motivation

The official bulletin, general edition, is the official document published by the State that contains all legislative and regulatory texts in arabic. Moroccan Secretariat of the Government launched an operation to digitize the official bulletin in PDF format since 1913. This operation becomes an encouraging factor to exploit this legal library resource.

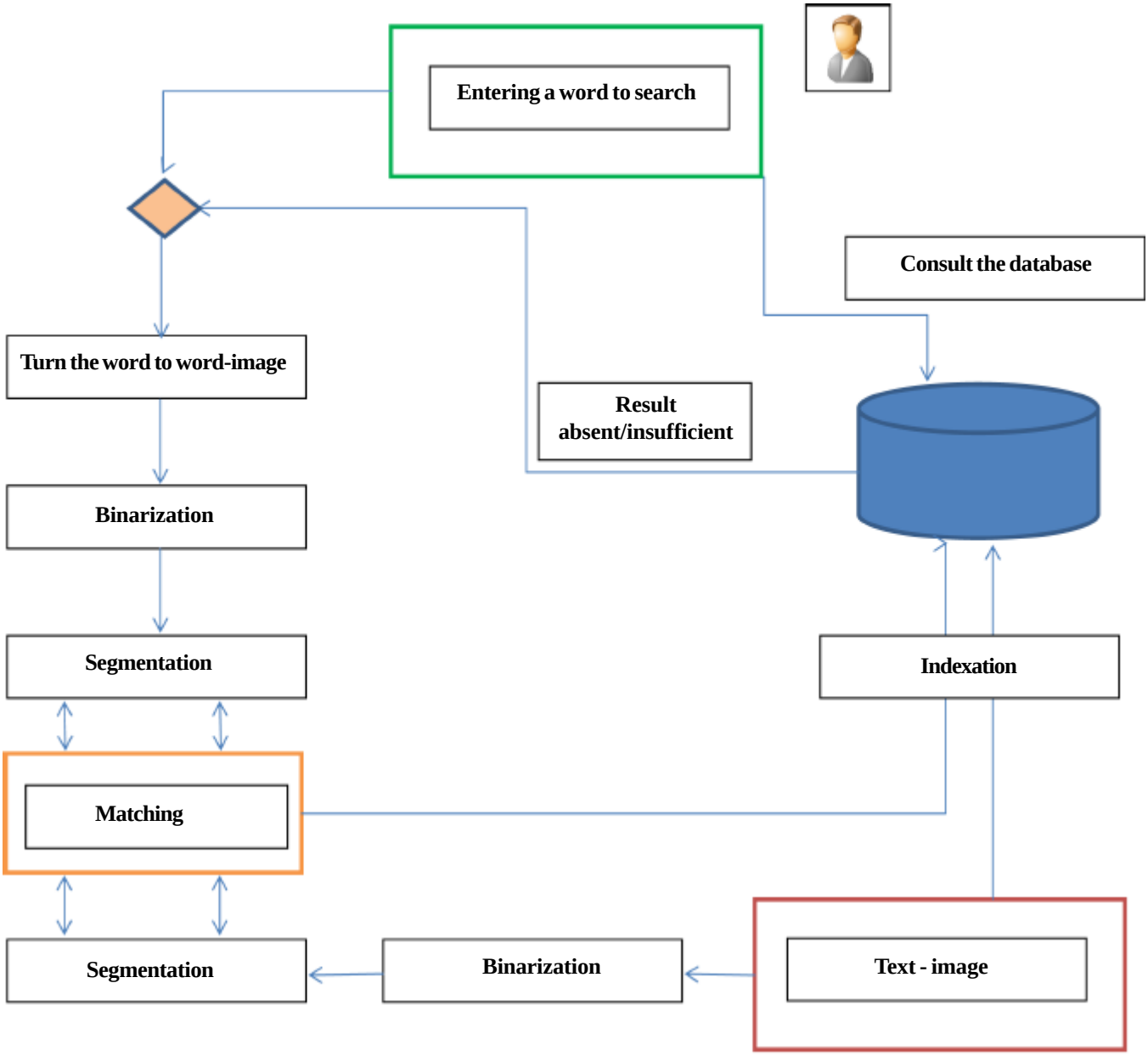

Figure 1. A Word-image's searching system

Using an optical arabic character recognition system did not give a satisfactory results (Figure 2), because of this issue, we stopped and we cannot be able to go further and give value to this law heritage.

However, many requests from users those want to be able to localize the information in the various issues of the official bulletin. As for ourselves, we focused our effort on this requirement to put forward a system that will give the user a way to search in the text-image. The text-image is a digitized page of the official bulletin (Figure 1).

The context of this paper is to propose a system which, through a database containing the various pages of the official bulletin in Arabic, will make the search for a word-image in a text-image possible.

3. Arabic language

The Arabic language is constructed from an alphabet of 28 letters. The characteristic of its characters is that each one can have

2792 الجريدة الرسمية عدد 4330 – 29 جمادى الأولى 1416 (25 أكتوبر 1995)

مرسوم رقم 2.95.610 صادر في 8 جمادى الأولى 1416 (4 أكتوبر 1995) بالموافقة على مقرر مجلس جماعة سلا - بطانة بعمالة سلا القاضي بتفويت هذه الجماعة قطعتين أرضيتين من الأملاك الجماعية الخاصة لفائدة الدولة (الأملاك الخاصة).

الوزير الأول ،

بناء على الظهير الشريف رقم 1.76.583 الصادر في 5 شوال 1396 (30 سبتمبر 1976) المعتبر بمثابة قانون يتعلق بالتنظيم الجماعي ، كما وقع تغييره أو تتميمه ؛

وعلى الظهير الشريف الصادر في 17 من صفر 1340 (19 أكتوبر 1921) في شأن الأملاك البلدية ، كما وقع تغييره أو تتميمه ؛

وعلى القرار الصادر في فاتح جمادى الأولى 1340 (31 ديسمبر 1921) بتحديد كيفية إدارة شؤون الأملاك البلدية ، كما وقع تغييره أو تتميمه ؛

وبعد الاطلاع على مقرر مجلس جماعة سلا - بطانة خلال دورته العادية المنعقدة بتاريخ 28 من ربيع الأول 1415 (6 سبتمبر 1994) ؛

وباقتراح من وزير الدولة في الداخلية وبعد استشارة وزير المالية والاستثمارات الخارجية ؛

رسم ما يلي :

المادة الأولى

يوافق على المقرر الصادر عن مجلس جماعة سلا - بطانة في 28 من ربيع الأول 1415 (6 سبتمبر 1994) بتفويت هذه الجماعة قطعتين أرضيتين من الأملاك الجماعية الخاصة موضوعا الرسمين العقاريين رقمي 36020 ر و 31955 ر لفائدة الدولة (الأملاك الخاصة) مساحتهما على التوالي أربعة آلاف وستة وعشرون مترا مربعا (4026 م²) وألف ومائة وتسعة وثلاثون مترا مربعا (1139 م²) تقعان بسلا - بطانة.

وقد رسمت حدود القطعتين الأرضيتين بوضوح في المخطط المضاف إلى أصل هذا المرسوم.

المادة الثانية

ينجز التفويت الموافق عليه بموجب هذا المرسوم بثمن إجمالي قدره مليونان وخمسمائة وثلاثة وثمانون ألفا وخمسمائة درهم (2.582.500 د) أي على أساس خمسمائة درهم للمتر المربع الواحد (500 د/م²).

المادة الثالثة

يسند إلى رئيس مجلس جماعة سلا - بطانة تنفيذ ما جاء في هذا المرسوم الذي ينشر بالجريدة الرسمية.

وحرر بالرباط في 8 جمادى الأولى 1416 (4 أكتوبر 1995).

الامضاء : عبد اللطيف الفيلالي.

وقعه بالعطف :

وزير الدولة في الداخلية ،

الامضاء : إدريس البصري.

قرار لوزير الفلاحة والاستثمار الفلاحي رقم 1732.95 صادر في 23 من محرم 1416 (22 يونيو 1995) بالموافقة على البرنامج الخاص بإنجاز أشغال التجهيز الداخلي للقطاعين الثاني والثالث من الدائرة السقوية لسبو الأوسط.

وزير الفلاحة والاستثمار الفلاحي ،

بناء على الظهير الشريف رقم 1.69.25 الصادر في 10 جمادى الأولى 1389 (25 يوليو 1969) المعتبر بمثابة قانون الاستثمارات الفلاحية ولاسيما الفصل 13 منه ؛

وعلى المرسوم رقم 2.84.320 الصادر في 22 من شعبان 1404 (24 ماي 1984) القاضي بتحديد القطاعين الثاني والثالث للمنطقة السقية بحوض سبو المتوسط (أقاليم فاس وتاونات وسيدي قاسم) الخاضعة لأحكام الظهير الشريف المشار إليه أعلاه ؛

وبعد الاطلاع على محضر اللجن المحلية للاستثمار الفلاحي لعمالة زواغة مولاي يعقوب وإقليمي سيدي قاسم وتاونات خلال اجتماعاتها المنعقدة على التوالي بتاريخ 24 يونيو 1994 و 12 و 28 يوليو 1994 ؛

وبعد استشارة وزير الدولة في الداخلية ،

قرر ما يلي :

المادة الأولى

يشمل برنامج سنوات 1994–1995 و 1995–1996 و 1996–1997 الخاص بإنجاز أشغال التجهيز الداخلي القطاعين الثاني والثالث البالغة مساحتهما 6.578 هكتارا الدائرة السقوية لحوض سبو الأوسط بولاية فاس وإقليمي سيدي قاسم وتاونات.

ويتناول البرنامج إنجاز أشغال تتمثل في قلب التربة وتسوية الأرض.

ويمكن اتخاذ التدابير الآتية لانجاز الأشغال المذكورة :

- منع عمليات الحرث والسقي ؛
- إتلاف المزروعات الموجودة.

المادة الثانية

ينشر هذا القرار بالجريدة الرسمية.

وحرر بالرباط في 23 من محرم 1416 (22 يونيو 1995).

الامضاء : حسن أبو أيوب.

قرار لوزير الاسكان رقم 1852.95 صادر في 4 صفر 1416 (3 يوليو 1995) بتفويض الامضاء

وزير الاسكان ،

بناء على الظهير الشريف رقم 1.95.40 الصادر في 27 من رمضان 1415 (27 فبراير 1995) بتعيين أعضاء الحكومة ؛

وعلى الظهير الشريف رقم 1.57.068 الصادر في 9 رمضان 1376 (10 أبريل 1957) في شأن تفويض إمضاء الوزراء وكتاب الدولة ونواب كتاب الدولة كما وقع تغييره وتتميمه ولاسيما الفصلين الأول والثاني منه ؛

Figure 2. Page from the official bulletin number 4330

four different forms depending on its position in the word [10] (Figure 3).

Distinctness of its fonts and the multiplicity of its writing styles are among the criteria that improve the Arabic language but make difficult the character recognition process.

For the official bulletin, general edition, we found that the font has not changed too much and keep pretty much the same writing style.

4. Optical Arabic Character Recognition

The difficulty of the Arabic language [5] [9] [10] [17] [18] induces us to see the various studies and research on the ORAC.

We started by seeing the work of character recognition that does not cover Arabic language. V. K. Govindan and A. P. Shivaprasad [3] did a review on Methodologies for character recognition in 1990.

B. Al-Badr, S. Mahmoud [24] did a study on the ORAC in 1995, making clear the difficulties that researchers meet to recognize the arabic character.

A study of ORAC systems [13] [14] is made to determine the main problem to success the recognition operation is related to the segmentation. Arabic sentence structure is hard to detect. A first Arabic character recognition system is proposed in 1998 by A. Cheung and others [1].

Figure 2. (a) Text-image extracted from the official bulletin (b) recognition result

	isolated (i)	end (e)	middle (m)	beginning (b)
alif	ﺍ	ـﺎ	ـﺎ	ﺍ
ba	ﺏ	ـﺐ	ـﺒـ	ﺑـ
ta	ﺕ	ـﺖ	ـﺘـ	ﺗـ
tha	ﺙ	ـﺚ	ـﺜـ	ﺛـ
jim	ﺝ	ـﺞ	ـﺠـ	ﺟـ
ha	ﺡ	ـﺢ	ـﺤـ	ﺣـ
kha	ﺥ	ـﺦ	ـﺨـ	ﺧـ
dal	ﺩ	ـﺪ	ـﺪ	ﺩ
dhal	ﺫ	ـﺬ	ـﺬ	ﺫ
ra	ﺭ	ـﺮ	ـﺮ	ﺭ
zan	ﺯ	ـﺰ	ـﺰ	ﺯ
siin	ﺱ	ـﺲ	ـﺴـ	ﺳـ
shiin	ﺵ	ـﺶ	ـﺸـ	ﺷـ
sadd	ﺹ	ـﺺ	ـﺼـ	ﺻـ
dad	ﺽ	ـﺾ	ـﻀـ	ﺿـ
tahn	ﻁ	ـﻂ	ـﻄـ	ﻃـ
zah	ﻅ	ـﻆ	ـﻈـ	ﻇـ
ayn	ﻉ	ـﻊ	ـﻌـ	ﻋـ
ghayn	ﻍ	ـﻎ	ـﻐـ	ﻏـ
fa	ﻑ	ـﻒ	ـﻔـ	ﻓـ
qaf	ﻕ	ـﻖ	ـﻘـ	ﻗـ
kaf	ﻙ	ـﻚ	ـﻜـ	ﻛـ
lam	ﻝ	ـﻞ	ـﻠـ	ﻟـ
miim	ﻡ	ـﻢ	ـﻤـ	ﻣـ
noon	ﻥ	ـﻦ	ـﻨـ	ﻧـ
ha	ﻩ	ـﻪ	ـﻬـ	ﻫـ
waw	ﻭ	ـﻮ	ـﻮ	ﻭ
ya	ﻯ	ـﻰ	ـﻴـ	ﻳـ
lamalif	ﻻ	ـﻼ	ـﻼ	ﻻ
tamarbot	ﺓ	ـﺔ		

Figure 3. Different forms of Arabic characters

In 2000 P. Ahmed and Y. AL-Ohali [4] gave an introduction about the status of the recognition of character Arabic and the challenges to rise to be able to achieve a higher satisfaction rate of recognition.

In 2002, M. S. Khorsheed [6] presented a review on uniqueness of Arabic language at the hands of recognition systems. L. M Lorigo and V. Govindaraju took a close look at the Arabic writing. They determine the way to advance in this area and improve the recognition process [7].

Z. Shaaban contributed in 2008 through a new approach that helps to make the character recognition based on neural networks [11].

A. M. AL-Shatnawi and others [2] gave an overview of the Arabic character recognition systems and indicate that studying the nature of the Arabic text is essential before designing a recognition system.

New approaches to the ORAC based on dissimilarity measures calculated on the basis of some polygonal attributes are developed [15] to increase the precision degree of recognition and reach a more reliable result.

5. Related work

The previous section showed us the different research on the ORAC. E. Ataer and P. Duygulu [12] understood the weakness of Arabic optical character recognition and they thought about using the word-image to index documents writing by the Ottoman character.

In 2004, L. Yue, T. Chewlim has done the same thing for Chinese language in order to search a word-image in digitized document.

Andreev and N. Kirov [20] used the Hausdorff distance in order to match the word-image and a text-image. This technique is used to offer a search tool in the Bulgarian document.

6. Proposed system

With a diversified language like Arabic, we opted to design a system that will allow searching with indexing [21] [22] in the Arabic official Bulletin (Figure 1). This system will achieve to index progressively digitized PDF document of the Official Bulletin.

The future user of this system (Figure 4) will be able to type a word that goes to the database first, if the result is not satisfactory or the search word is not yet indexed, the word will turn to word-image that will undergo a binarization operation (0.1) as a matching process will start to find the word-image in the text-image.

7. Conclusion and outlook

We presented the various works on Arabic optical character recognition and the difficulties that the researchers met to establish a good system of recognition. The problem of character segmentation is the focal point for a successful result.

Languages such as Chinese, Bulgarian experienced many research works in the word-image matching. Index documents based on this method can satisfy the need to find information for a user.

We tried to propose a system that will give control to the user to fill out and search on the text-image database of the official bulletin in Arabic.

Currently, we could segment a page of the Official Bulletin and match the word-image and the text-image to produce a search result. Matching word-image image and text-image will be the subject of a forthcoming publication that deals the various algorithms that we studied and our choice.

A semantic approach [23] is included in the system to bring out the results that have a connection with the user query. The prior existence of another database that contains summaries of the official bulletin in text format will allow us to integrate this semantic aspect to our system. This approach will be the subject of a forthcoming publication.

References

[1] Cheung, A., Bennamoun, M., Bergmann, N. W. (1998). A recognition-based Arabic optical character recognition system, *In*: IEEE International Conference on Systems, Man and Cybernetics, 4189-4194.

[2] AL-Shatnawi, A. M., AL-Zawaideh, F. H., AL-Salaimeh, S., Omar, K. (2011). Offline Arabic Text Recognition – An Overview, World of Computer Science and Information Technology Journa, 1 (5) 184-192.

[3] Govindan, V. K., Shivaprasad, A. P. (1990). Character recognition - A review, *Pattern Recognition* - PR, 23 (7) 671-683.

[4] Ahmed, P., Al-ohali, Y. (2000). Arabic character recognition: Progress & chalenges, J King Saud Univ., 12, Comp. & Infor. Sci., 85-116.

[5] Märgner, V., El Abed, H. (2009). Arabic Word and Text recognition – Current Developments, the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt.

[6] Khorsheed, M. S. (2002). Off-Line Arabic Character Recognition – A Review, *Pattern Analysis & Applications*, 5(1) 31-45.

[7] Lorigo, L. M., Govindaraju, V. (2006). Offline Arabic handwriting recognition: a survey, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28 (5) 712–724.

[8] AlKhateeb, J. H., Ren, J., Ipson, S., Jiang, J. (2008). Knowledge-based baseline detection and optimal thresholding for words segmentation in efficient pre-processing of handwritten Arabic text. Fifth international conference on information technology: new generations. IEEE computer society. 1158-1159.

[9] Argner, V., El Abed, H. (2008). Databases and Competitions: Strategies to Improve Arabic Recognition Systems. 82-103.

[10] Amin, A. (1997). Arabic character recognition, *In*: Bunke, H.,Wang, P. S. P., editors, Handbook of Character Recognition and Document Image Analysis_ chapter, 15, 397-420 World Scientic.

[11] Shaaban, Z. (2008). A New Recognition Scheme for Machine-Printed Arabic Texts based on Neural Networks. *In*: Proceedings of World Academy of Science, Engineering and Technology, 31, Vienna, Austria.

[12] Ataer, E., Duygulu, P. (2007). Matching ottoman words: an image retrieval approach to historical document indexing, CIVR, the 6th ACM international conference on Image and video retrieval, 341 – 347, ACM, NY, USA.

[13] Kanungo, T., Marton, G. E., Bulbul, O. (1998). Performance Evaluation of Two Arabic OCR Products. *In*: Proc. of AIPR Workshop on Adv. in Comp. Assist. Recognition., SPIE (W. DC). 3584

[14] Jumari, K., Ali, M. A. (2002). A Survey And Comparative Evaluation Of Selected Off-Line Arabic Handwritten Character Recognition Systems. *Jurnal Teknologi*, 36 (1-18), Jun.

[15] Chaker, I., Benslimane, R. (2011). nouvelle approche pour la reconnaissance des caractères arabes imprimés, *Revue méditerrannéenne des Télécommunications*, 1 (2) 87-92.

[16] Ahmed M. Zeki, Mohamad S. Zakaria. (2009). Challenges in Recognizing Arabic Characters'Information Technology. ITSim. International Symposium.

[17] Chan, J., Ziftci, C., Forsyth, D. (2006). Searching off-line arabic documents. *In*: proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[18] Bilal Khan, Khaled S. Alghathbar, Muhammad Khurram Khan, Abdullah M. AlKelabi, Abdulaziz AlAjaji. (2010). Using Arabic CAPTCHA for Cyber Security, *Communications in Computer and Information Science*, 122, 8-17.

[19] Yue, L., Chewlim,T. (2004). Chinese word searching in imaged documents, *International Journal of Pattern Recognition and Artificial Intelligence*, 18 (2) 229- 246.

[20] Andreev, A., Kirov, N. (2008). Some variants of Hausdorff distance for word matching. Review of the National Center for Digitization, 12, 3–8.

[21] Kamel, Ibrahim, Alaa Eltalmas, Muneer Abubaker. (2005). On searching Arabic records in electronic libraries. *International Cataloguing and Bibliographic Control*, 34 (2) 23-26.

[22] Eldos, T. (2003). Arabic Text Data Mining: A Root-Based Hierarchical Indexing Model, *International Journal of Modelling and Simulation*, 23, 158–166.

[23] Boucher, A., Le, T. (2005). Comment extraire la sémantique d'une image ?. *In*: 3rd International Conference: Sciences of Electronic, *Technologies of Information and Telecommunications*, SETIT 2005, Tunisia. p. 295-306. March.

[24] Al-Badr, B., Mahmoud, S. (1995). Survey and bibliography of Arabic optical text recognition, *Signal Process*. 41, 49–77.