Segmentation and Word Spotting Methods for Printed and Handwritten Arabic Texts: A Comparative Study

Mariem Gargouri Kchaou, Slim Kanoun
University of Sfax
National School of Engineers (ENIS), BP 1173
Sfax, 3038, Tunisia
gargouri.mariem@gmail.com, slim.kanoun@gmail.com

Jean-Marc OGIER
Laboratory L3i
University of La Rochelle
La Rochelle, France
jean-marc.ogier@univ-lr.fr

***Abstract*—This paper presents a comparative study for word spotting techniques according to holistic approach. So, the current work consists in experimenting word image segmentation, characterization and matching to show the most reliable techniques. The experimental process is done in the same printed and handwritten Arabic dataset. Our aim is to realize an effective system of information retrieval.**

Keywords- information retrieval, word spotting, Arabic script, text segmentation, word characterization and matching, holistic approach.

I. Introduction

Following the evolution of communication and information technologies, a great mass of printed and handwritten texts are currently digitized and put at the disposal of various organizations in the world. Thus, during the last years, several researches are done in developing systems to facilitate the automatic treatment of this great volume of documents in order to use its information wealth. The first information retrieval systems for text images are mainly optical character recognition systems (OCR) [8]. Unfortunately, these systems are limited to good quality text images. Therefore, it is crucial to set up other categories of information retrieval systems in order to extend the treatment to poor quality text images (FAX device, photocopy, handwritten texts…). For that purpose, a second system category is developed and it is focused on information retrieval in texts images without explicit recognition. The said system is called word spotting. The principle of working of such system is outlined as follows: (i) segment texts into words (ii) extract characteristic features from words images (iii) match the query image characteristic features with those of the dataset (iv) sort the obtained results (v) restore the relevant documents.

Initially, word spotting is proposed in [4] for printed text and a few years later in [6] for handwritten text. Word spotting on Latin alphabets has received more considerable attention [1] [3] [5] [7] [10-12] then other scripts, especially Arabic script. The first system for Arabic word spotting is described in Srihari et *al*. work. A precision rate of 55% and a recall rate 50% are obtained for 10 good quality recent handwritten documents [13]. We find also Leydier et *al*. work [5] with only two handwritten documents.

Word spotting methods can be classified into segmentation based techniques (analytical approach) or segmentation free techniques (holistic approach).

To avoid the segmentation problems, the most of word spotting systems uses the holistic approach [1-3] [10] [12], rather than the analytical approach [9].

In Rath et *al*. system [10], the precision rate is 65%, for 10 documents Latin of good qualities from the George Washington handwritten letters. In Anurag et *al*. system [1], for the handwritten English, an average precision of 67% was obtained for 30 words of queries. In Ataer et *al*. system [2], the success rate for 15 queries is 0.8524 for printed Ottoman documents. In Kesidis et *al*. system [3], the overall experimental results on 153 pages from a French historical book which was published in 1838 shows that the average estimation error is below 8% in terms of F-Measure. In Rusinol et *al*. system [12], for the George Washington dataset the mean average precision is 53.76% and for the Lord Byron dataset the mean average precision results is 70.23%.

In this paper, we propose a comparative study between two approaches: Anurag et *al.* approach and Manmatha et *al.* approach. In this framework, we compare the results of their related segmentation, characterization and matching techniques. Horizontal and vertical profile segmentation, geometrical moments and cosine similarity metric are used in Anuarg et *al.* approach [1]. Scale space segmentation [7], the profile projection and Dynamic Time Warping (DTW) [11] are used in Manmatha et *al.* approach.

We carried out these choices because our main objective is not to create a robust system but it is a question of comparison between the said techniques in the same Arabic dataset. In addition, we provide some amendments to the first two phases (segmentation, characterization) at the level of Manmatha et al. approach.

The remainder of this paper is organized as follows: We present in Section 2 our proposed methodology. Section 3 details the experimental setup by using

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.266

printed and handwritten texts. Finally, the conclusions can be found in Section 4.

II. Proposed Methodology

In this section, we outline our methodology by studying the following aspects according to the above techniques: segmentation, characterization and matching, as well as the proposed amendments.

A. Segmentation

In this section, we outline our methodology by studying the following aspects according to the above techniques: segmentation, characterization and matching, as well as the proposed amendments.

1) Horizontal and vertical profile features: Firstly, the horizontal profile feature of the document image is used to segment it into line images. Thereafter, the vertical profile feature of each line image is used to extract individual word images [1].

2) Scale space segmentation: Let consider *f(x,y)* a two-dimensional image, after convolving it with *L*, the corresponding result image is:

$$I(x,y;\sigma_x,\sigma_y) = L(x,y;\sigma_x,\sigma_y) * f(x,y) \quad (1)$$

$$L(x,y;\sigma_x,\sigma_y) = \frac{1}{2\pi{\sigma_x}^3\sigma_y}\left(\frac{x^2}{{\sigma_x}^2}-1\right)e^{-\left(\frac{x^2}{2\sigma_x^2}+\frac{y^2}{2\sigma_y^2}\right)} + \frac{1}{2\pi{\sigma_y}^3\sigma_x}\left(\frac{y^2}{{\sigma_y}^2}-1\right)e^{-\left(\frac{x^2}{2\sigma_x^2}+\frac{y^2}{2\sigma_y^2}\right)} \quad (2)$$

$$\eta = \frac{\sigma_x}{\sigma_y} \quad (3)$$

which $L \in R^{mXn}$ is the second order anisotropic Gaussian differential operator, and η is the multiplication factor.

After this transformation is done on the image document, horizontal and vertical profile features are used to segment it into words.

Scale space segmentation is applied for the first time on Arabic. Arabic word is a various combinations of Partial Words (PWs) where a PW is a various combinations of basic characters. This technique can solve the problem of PWs and put them together in one unit (blob) which is a word.

Following our experiments on the printed dataset described in Section 3.2.1, we estimated $\eta = 2.5$, $\sigma_x = 15$ and $\sigma_y = 6$, because they are more appropriate for the Arabic script than those proposed in [7].

We observed that the width *n* of *L* in equation (4) of the convolution influenced on the obtained segments. In the same way, we observed that the variation of the height *m* was useless, then we fixed it at 1 (*m=1*).

$$I(x,y) = \sum_{k=1}^{m}\sum_{l=1}^{n} L(k,l)f(x+k-1, y+l-1) \quad (4)$$

In table 1, we present the *n* varying effect on the resulting word image of الإقتداء which contents 5 PWs. We succeed to unify them in one unit.

TABLE I. N VARYING EFFECT

n	Resulting Image
5	
7	
9	
11	
13	
15	

B. Word spotting

After preprocessing and segmentation steps, word spotting system extracts characteristics from words images. When the user provides a query, the system returns relevant documents which are sorted by matching relevance. There exist a number of different approaches. We chose Anurag et *al.* approach and Rath and Manmatha approach.

1) Anurag et al. approach: Geometrical moments are characteristic features of the word image (see (10)), and the similarity cosine metric is the matching technique (see (11)).

$$M_{pq} = \sum_X \sum_Y x^p y^q f(x,y) \quad (5)$$

$$\bar{x} = \frac{M_{10}}{M_{00}}, \ \bar{y} = \frac{M_{01}}{M_{00}} \quad (6)$$

$$\bar{M}_{pq} = \sum_X \sum_Y (x-\bar{x})^p (y-\bar{y})^q f(x,y) \quad (7)$$

$$\sigma_x = \sqrt{\frac{\bar{M}_{20}}{M_{00}}}, \ \sigma_y = \sqrt{\frac{\bar{M}_{02}}{M_{00}}} \quad (8)$$

$$x^* = \frac{(x-\bar{x})}{\sigma_x}, \ y^* = \frac{(y-\bar{y})}{\sigma_y} \quad (9)$$

$$m_{pq} = \frac{\sum_X \sum_Y (x^*)^p (y^*)^q f(x,y)}{M_{00}} \quad (10)$$

$$SIM = \frac{\vec{q}.\vec{w}}{|\vec{q}|*|\vec{w}|} \quad (11)$$

which *f(x,y)* is a two-dimensional image, $\vec{w}$ is word image feature vector and $\vec{q}$ is query feature vector.

2) Rath and Manmatha approach: The characteristics are: Projection Profile (PP) (see (12)), Upper Word Profile (UWP) (see (13)), Lower Word Profile (LWP) (see (14)), Background to Ink Transitions (BTIT) (see (15)), and the vector combination (VC) which is the combination of these four characteristics. DTW is the matching technique (see (16) and (17)); it allows comparing two vectors which have not the same size.

$$PP(I,c) = \sum_{r=1}^{h}(255 - I(r,c)) \quad (12)$$

$$\begin{cases} UWP(I,c) = \text{undefined}, \\ \quad \text{if } \forall r \ (\text{is\_ink}(I,r,c) = 0) \\ \text{argmin}_{r=1..h}(\text{is\_ink}(I,r,c) = 1), \text{otherwise}. \end{cases} \quad (13)$$

$$\begin{cases} LWP(I,c) = \text{undefined}, \\ \quad \text{if } \forall r \ (\text{is\_ink}(I,r,c) = 0) \\ \text{argmax}_{r=1..h}(\text{is\_ink}(I,r,c) = 1), \text{otherwise}. \end{cases} \quad (14)$$

$$BTIT(I,c) = nbit(I,c)/S \quad (15)$$

which $I(r,c) \in R^{h \times w}$ is an image, r and c indicate the row and column index of the pixel.

$$D(i,j) = \min \begin{cases} D(i, j-1) \\ D(i-, j) \\ D(i-1, j-1) \end{cases} + d(x_i, y_i) \quad (16)$$

$$dist(X,Y) = D(M,N)/K \quad (17)$$

which $X = (x_1, \ldots, x_M)$ and $Y = (y_1, \ldots, y_N)$ are two series, $D \in IR^{M \times N}$ is a matrix, $d(x_i, y_i) = \sum_{k=1}^{d} (x_{i,k} - y_{i,k})^2$ is the square of the Euclidean distance, and K is the warping path.

According to [11], the standardization factor S for BTIT characteristic was estimated at 6 (see (15)). Based on our observation, we estimated this factor at 8 transitions for Arabic letters (see Figure 1).

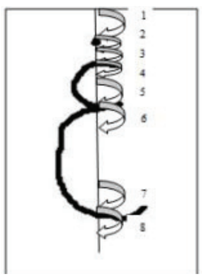

Figure 1. Background to ink transitions for an Arabic letter

Then we proposed a new method: the product **(P)** between obtained distances of these four said characteristics and it consists:

i. Calculate obtained distances between query word image and words images of the dataset by using each characteristic alone of these four characteristics.

ii. Make the product between the four obtained distances so the product result becomes the new score for each image of the dataset.

iii. Sort the new scores in order to draw the most relevant images for the query image.

For more details on Anurag *et al.* approach and Manmatha *et al.* approach, we refer to [1], [7] and [11].

III. Experimental Results and Discussion

In this section, we present our study on Arabic textual images. The dataset contains multi-fonts and multi-sizes printed texts and multi-writers handwritten texts. We study segmentation and word spotting independently.

A. Segmentation results

Two of the main problems are over-segmentation and under-segmentation. On the one hand, over-segmentation segments out a single word into multiple words. On the other hand, under-segmentation segments two or more words as a single word.

The segmentation error rate is the number of concatenated words and subdivided words divided by the whole text word number.

1) Datasets: We chose a text from "muqadima Iben Khaldoun". This text contained 208 words (see figure 2).

اعلم أن فن التاريخ فن عزيز المذهب جم الفوائد شريف الغاية إذ هو يوقفنا على أحوال الماضين من الأمم
في أخلاقهم. و الأنبياء في سيرهم. و الملوك في دولهم و سياستهم. حتى تتم فائدة الإقتداء في ذلك لمن
يرومه في أحوال الدين و الدنيا فهو محتاج إلى مآخذ متعددة و معارف متنوعة و حسن نظر و تثبت
يفضيان بصاحبهما إلى الحق و ينكبان به عن المزلات و المغالط لأن الأخبار إذا اعتمد فيها على مجرد
النقل و لم تحكم أصول العادة و قواعد السياسة و طبيعة العمران و الأحوال في الاجتماع الإنساني و لا
قيس الغائب منها بالشاهد و الحاضر بالذاهب فربما لم يؤمن فيها من العثور و مزلة القدم و الحيد عن جادة
الصدق و كثيرا ما وقع للمؤرخين و المفسرين و أئمة النقل من المغالط في الحكايات و الوقائع لاعتمادهم
فيها على مجرد النقل غثا أو سمينا و لم يعرضوها على أصولها و لا قاسوها بأشباهها و لا سبروها بمعيار
الحكمة و الوقوف على طبائع الكائنات و تحكيم النظر و البصيرة في الأخبار فضلوا عن الحق و تاهوا في
بيداء الوهم و الغلط و لا سيما في إحصاء الأعداد من الأموال و العساكر إذا عرضت في الحكايات إذ هي
مظنة الكذب و مطية الهذر و لا بد من ردها إلى الأصول و عرضها على القواعد.

Figure 2. A text of 208 words extracted from "muqadima Iben Khaldoun"

Then, we made 4 printed documents (D1, D2, D3 and D4) which were written in "Traditional Arabic" font. The sizes of D1, D2, D3 and D4 were 14, 24, 34 and 44 respectively. D5 was a printed document with 4 fonts and 4 sizes. Each 52 words were written in "Courier new" font and 44 size, "Tahoma" font and 34 size, "Arial Unicode MS" font and the 24 size, "Microsoft Sans Serif" font and 14 size, respectively. D6 was a handwritten document written by one writer. D7 was a handwritten document written by others four writers.

We used printed and handwritten recent documents, which were carefully written and the lines were well spaced. Our main aim is to study word segmentation and not line segmentation.

2) Horizontal and vertical profile features results: We used D1. The segmentation gave 384 segments, whereas the words number was 208 words. The obtained segments were all of them PWs. The subdivided words were 104. The error rate was 50%.

Even if this technique is effective for other scripts and particularly for Latin [1] or printed Ottoman [2], it segments the Arabic text in PWs and not in words.

3) Scale space segmentation results: Table 2 shows segmentation error rates using scale space theory.

For printed document, the segmentation depended not only on its size but also on its font.

We could found good results, if the whole document was written by the same font and the same size (D1, D2, D3 and D4).

But, if the document was written by different fonts and different sizes like D5, results were not so good. When we used the width 5, 44 size words were subdivided. And from the width 7, 14 size words were concatenated.

So, we could choose the width 5 or 7 for printed documents.

For handwritten documents, the document segmentation depended on each writer style. The problems in handwritten documents are the concatenation between the word of coordination "و" ("and" in English) and its successor on the one hand, and important spacing between the PW of the same word on the other hand. Despite the handwritten segmentation difficulties, our results were important and satisfactory.

So, we can choose the width 15 or 17 for handwritten documents.

In general, when the size of the core is increased, the number of concatenated words increases and the number of subdivided words decreases.

B. *Word spotting results*

1) Datasets: Two datasets of Arabic words images (PD and HD) are used. PD and HD were segmented manually in order that our results would not be influenced by the segmentation errors. PD and HD were written by the same 50 words which were extracted from "muqadima Iben Khaldoun" (see figure 3).

الأمم سياسة الدين الدنيا الكائنات التاريخ ابن خلدون مقدمة
الممالك الجيوش الآثار الرياضة السحر الخرافات قراءة
الخلافة قلم سيف حجابة الحكايات الدولة وزارة كتابة قيادة
القيروان الأغالبة الشيعة المال الصلاة الأولياء العلماء العدالة
الإسلام الحجاز الشام مصر الاسكندرية اليمن أفريقية المغرب
العمارة الأرض الجغرافيا البراءة الجهل أمراء الحرب بغداد
الوحي البشر

Figure 3. 50 words extracted from "muqadima Iben Khaldoun"

PD was a printed dataset. Each word was written by 5 fonts and 4 sizes what gives us 20 occurrences and then we had 1000 images words. The fonts were Traditional Arabic, Courier New, Tahoma, Arial Unicode MS, and Microsoft Sans Serif (see figure 4). The sizes were 14, 24, 34 and 44.

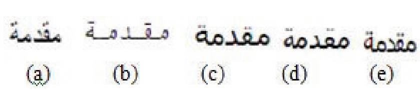

Figure 4. Words images of size 14 extracted from PD: (a) Traditional Arabic, (b) Courier New, (c) Tahoma, (d) Arial Unicode MS, (e) Microsoft Sans Serif.

HD was a handwritten dataset. Each word was written by 4 writers 5 times what gives us 20 occurrences and then we had 1000 words images (see figure 5).

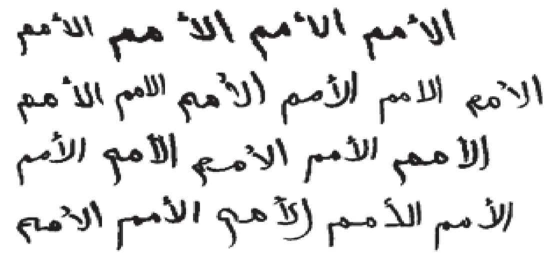

Figure 5. Images of words extracted from HD

2) Processing time: Table 3 presents processing time for word characterization and for a query execution by the two approaches. We used PD for these experiments.

Compared to a dataset containing 1000 words images, the characterization time of geometrical moments was huge, it reached the 5 minutes, whereas the characterization time of Rath and Manmatha characteristics was about a few seconds. But, the query execution time of Anurag et *al.* approach take less time than Rath and Manmatha approach, because sizes of the first characteristic vectors were smaller than those of the second, and the DTW algorithm was greedier in calculation time than the similarity cosine metric.

3) Classification accuracy: For word spotting study, we used PD and HD. Each dataset contained 1000 words. So, for each dataset, we used 900 images words for training and 100 images words for testing. That's mean, for each word, 18 images were used for training and 2 images words were used for testing.

Recall rate (Top18) is the proportion of relevant documents found compared to the whole relevant documents present in the dataset. Precision rate (Top1 and Top5) is the proportion of relevant documents compared to the whole documents provided by research.

For Anurag et *al.* approach (see table 4), we drew the following conclusions. For the printing, the best average recall rate 48% was given by the order 10. For the handwriting, we tested only the order 10. The average recall rate was 22%. On our dataset, the results of this approach were not good.

For Rath and Manmatha approach (see table 5), we drew the following conclusions. For the printing, the best average recall rate 82% was given by the product. For the handwriting, the best average recall rate 53% was also given by the product.

TABLE II. SEGMENTATION ERRORS

n		5	7	9	11	13	15	17
Printing	D1 (%)	**5,79**	70,04	x	x	x	x	x
	D2 (%)	2,41	**0,96**	0,96	32,85	x	x	x
	D3 (%)	20,28	**0**	**0**	**0**	**0**	2,41	x
	D4 (%)	x	4,34	**0**	**0**	**0**	**0**	x
	D5 (%)	39,5	**25,50**	35,00	44,00	44,00	45,00	x
Handwriting	D6 (%)	28,50	21,73	20,28	15,45	14,97	**14,00**	15,45
	D7 (%)	x	29,5	20,5	15	8	6	**3,5**

TABLE III. PROCESSING TIME

	Order							*Characteristic*					
	6	**8**	**10**	**12**	**14**	**16**	**18**	**PP**	**UWP**	**LWP**	**BTIT**	**VC**	**P**
Words characteriza-tion time(s)	60	87	119	157	207	264	314	6,77	7,00	6,89	6,36	12,35	28
Query execution time(s)	1,43	1,51	1,59	1,64	1,68	1,71	1,95	2,65	2,69	2,58	2,54	14,39	11,00

TABLE IV. PERFORMANCE RESULTS OF ANURAG ET AL. APPROACH

	Printing							*Handwriting*
	6	**8**	**10**	**12**	**14**	**16**	**18**	**10**
Top1 (%) Rank	90 2	85 3	80 4	72 6	**95 1**	70 7	75 5	30 24
Top5 (%) Rank	**67 1**	60 4	63 3	45 7	65 2	55 6	58 5	22 30
Top18(%) Rank	40 3	39 4	**48 1**	37 6	43 2	34 7	38 5	24 22

TABLE V. PERFORMANCE RESULTS OF RATH AND MANMATHA APPROACH

	Printing						*Handwriting*					
	PP	**UWP**	**LWP**	**BTIT**	**CV**	**P**	**PP**	**UWP**	**LWP**	**BTIT**	**CV**	**P**
Top1 (%) Rank	79 6	93 3	90 4	85 5	**100 1**	**100 1**	55 4	87 3	38 5	19 6	**100 1**	**100 1**
Top5 (%) Rank	72 6	88 3	87 4	84 5	**100 1**	**100 1**	42 5	48 3	44 4	21 6	**100 1**	88 2
Top18(%) Rank	53 5	63 3	51 6	54 4	76 2	**82 1**	26 6	35 3	34 4	30 5	50 2	**53 1**

IV. Conclusion and Future Work

In this paper, we presented a comparative study of various techniques of segmentation, characterization and matching, as well as, we highlighted ours contributions.

The segmentation techniques are horizontal and vertical profile features and scale space segmentation. The characterization techniques are geometrical moments, projection profile, upper word profile, lower word profile, background to ink transitions, and the vector combination. The matching techniques are similarity cosine metric and DTW. And we added a new method which is the product.

We provided our results by using the same Arabic printed and handwritten datasets.

Although the dataset is small, we could compare between these methods and draw important conclusions.

We showed that the scale space segmentation approach is more reliable than the horizontal and vertical profile features, especially for Arabic script. By varying the width of core used by the first, we could improve our results.

In terms of processing time, the projection profiles take less time than geometrical moments, but the DTW is greedier in calculation time than similarity cosine metric.

In terms of performance results, we came to reach the following conclusions. For the average precision rates, the vector combination is the most reliable one. For the average recall rates, our method product is the most reliable one.

We aim to extend our methodology to larger datasets and to incorporate and to study other different approaches in future experiments.

References

[1] B. Anurag, J. Damien and G. Venu. Script Independent Word Spotting in Multilingual Documents. The Third International Joint Conference on Natural Language Processing, January 7-12, 2008, Hyderabad, India.

[2] E. Ataer, P. Duygulu. Retrieval of ottoman documents. Multimedia Informations Retrieval (MIR) workshop, 2006, pp. 155-162..

[3] A. L. Kesidis and B. Gatos. Efficient Cut-off Threshold Estimation for Word Spotting Applications. In International Conference on Document Analysis and Recognition, ICDAR, page 279-283. IEEE, 2011.

[4] S. Khoubyari, J.J. Hull. Keyword location in noisy document images. Second Annual Symposium on Document Analysis and Informations Retrieval Las Vegas, NV, 26–18 April 1993, pp. 217–231.

[5] Y. Leydier, A. Oujia, F. Le Bourgeois, H. Emptoz. Towards an omnilingual word retrieval system for ancient manuscripts. Pattern Recognition, 2009.

[6] R. Manmatha, C. Han, et E.M. Riseman. Word spotting : A new approach to indexing handwriting. In Conference on Computer Vision and Pattern Recognition, CVPR, San Francisco, Etats-Unis, 1996, pp. 631–637.

[7] R. Manmatha, J.L. Rothfeder. A scale space approach for automatically segmenting words from historical handwritten documents. IEEE Transactions on Pattern Analysis and Machine Intelligence. 27(8), 2005.

[8] A. Nazif. A system for the recognition of the printed Arabic characters. M.Sc. Thesis. Cairo University. 1975.

[9] M.S. Praveen, K.P. Sankar, C.V. Jawahar. Character n-Gram Spotting in Document Images. In International Conference on Document Analysis and Recognition, ICDAR (2011) 941-945.

[10] T. Rath, R. Manmatha. Word image matching using dynamic time warping. IEEE Computer Vision and Pattern Recognition, Madison, USA, 2003, pp. 521–527.

[11] T.M. Rath,, R Manmatha : Word spotting for historical documents. International Journal on Document Analysis and Recognition (IJDAR) 9(2), 2007, pp. 139–152

[12] M. Rusinol; D. Aldavert; R. Toledo; Josep Llados. Browsing Heterogeneous Document Collections by a Segmentation-free Word Spotting Method. In International Conference on Document Analysis and Recognition, ICDAR 2011, pp 63-67.

[13] S. N. Srihari and H. Srinivasan and P. Babu and C. Bhole. Handwritten Arabic Word Spotting using the CEDARABIC Document Analysis System. Proc. Symposium on Document Image Understanding Technology (SDIUT-05), College Park, MD, November, 2005, pp 123–132.