Arabic Newspaper Page Segmentation

Karim Hadjar and Rolf Ingold

DIUF, University of Fribourg
Chemin du Musée 3, 1700 Fribourg, Switzerland
{karim.hadjar, rolf.ingold}@unifr.ch

Abstract

The aim of layout analysis is to extract the geometric structure from a document image. It consists of labeling homogenous regions of a document image. This paper describes the performance of segmentation algorithms and their adaptation in order to treat complex structured Arabic documents such as newspapers. Experimental tests have been carried out on four different phases of newspaper image analysis: thread recognition, frame recognition, image text separation, text line recognition, and line merging into blocks. Some promising experimental results are reported.

1. Introduction

In the field of document recognition many improvements have been made during the last decade. In fact, there is a variety of proposed algorithms for geometric layout analysis of document images: morphology or smearing based approaches, projection profiles, texture-based analysis, analysis of the background structure, and others [2, 11]. As we know layout information describes the manner in which specific components of the documents, such as blocks of text and individual words, are spatially organized within the document image.

In the literature layout analysis methods have been firstly focused on simple document structures [8]. But these methods showed some limitations when dealing with complex structured documents such as newspapers or magazines. The major difficulty of such kind of documents is the variability of layout between newspapers and even different issues of the same newspaper. Some recent works show a great interest in complex layout analysis [3, 4, 5]. Currently known approaches rely on document models [10] and interactive incremental learning [6] which is one of the main goals of the CIDRE[1] project. None of these methods mentioned above have been applied on the Arabic case.

The Arabic language is known to be a difficult language: the alphabet is much richer that the Latin one, the form of the letter changes depending on its position inside the word, the words are written from right to left. Therefore we have tried to test the performance of the well known algorithms of segmentation and adapt them in order to treat Arabic newspapers. In this paper, we try to give an overview of the methods with their adaptations.

This paper is organized as follows: in section 2 we present a brief introduction of the characteristics of the Arabic language. In section 3 we present the different algorithms used for page segmentation and their adaptation. In section 4 we present our experimental results on Arabic newspaper page segmentation and finally section 5 brings up the conclusion and future work.

2. Characteristics of the Arabic language

Arabic is spoken by almost 250 million people and is the (or one) official language of 19 countries. There are two main types of written Arabic: classical Arabic the language of the Quran and classical literature and modern standard Arabic the universal language of the Arabic-speaking world which is understood by all Arabic speakers. Each Arabic speaking country or region also has its own variety of colloquial spoken Arabic.

Arabic belongs to the group of Semitic alphabetical scripts [7] in which mainly the consonants are represented in writing, while the markings of vowels (using diacritics) is optional.

The Arabic alphabet contains 28 letters as shown in figure 1, words are written in horizontal lines from right to left, numerals are written from left to right as illustrated in figure 2.

ا ب ت ث ج ح

خ د ذ ر ز س

ش ص ض ط ظ ع

غ ف ق ك ل م

ن ه و ى

Figure 1. The 28 letters of the Arabic alphabet

[1] CIDRE stands for Cooperative and Interactive Document Reverse Engineering and is supported by the Swiss National Fund for Scientific Research, code 2000-059356.99-1.

Figure 2. Numerals

Most letters change form depending on whether they appear at the beginning, middle or end of the word [1], or on their own as shown in figure 3. We also notice the diacritics attached to the letters that can be 1, 2 or 3 diacritics points located either above or under the letter. For the vowels we distinguish 3 long vowels letters alif (ا), yaa (ي), waaw (و) and short vowels as illustrated in figure 4.

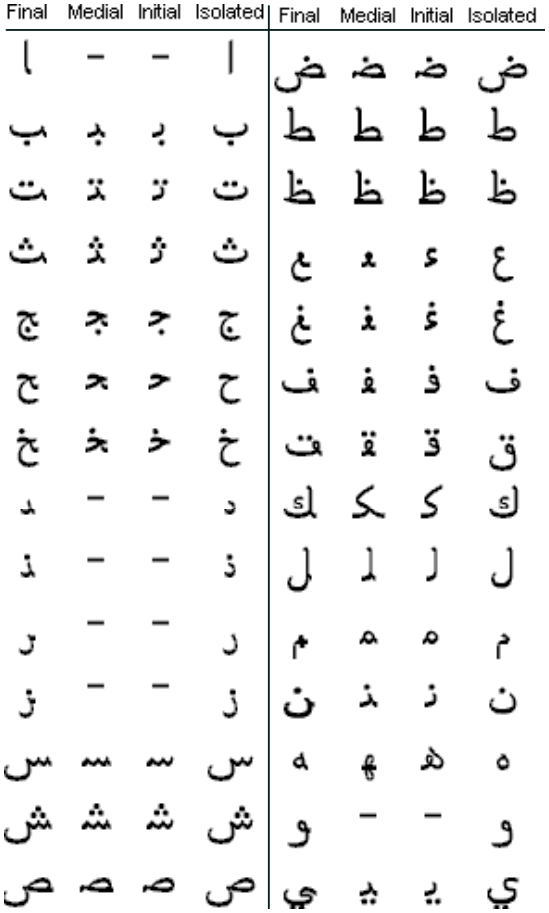

Figure 3. Letters and their appearance inside the word

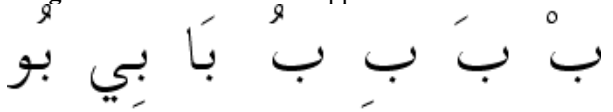

Figure 4. Arabic vowels

The vowels are used in order to ensure that the text is read aloud without mistakes. Such kinds of text are Quran, poetry and textbooks for foreign learners.

The disposition of diacritics and the vowels and their numbers increase the complexity of the Arabic language. In fact there are many cases to be treated and sometimes we face an ambiguity especially when diacritics of the first line and those of the second line are near each other or merged. In figure 5 we see that two diacritics under the last character of the first line are merged with the diacritic of one character of the second line.

Figure 5. Problem of separation of diacritics points

3. Algorithms

In computer vision, the aim of image segmentation is to separate the given image into homogenous region. Each region is indicated by a meaningful property. In fact the following steps are done: thread extraction, frame extraction, image text separation, text line extraction and line merging into blocks. We have modeled the algorithm as a set of tools that can be used separately, for example if we need thread extraction we use the appropriate tool.

The algorithm uses a bottom-up approach based on connected components for image, thread and frames extraction. Connected components are also used for text line extraction after the use of the *Run Length Smearing Algorithm* (RLSA) [12]. The line merging into blocks is done according to rules taking into account the characteristics of the Arabic language

3.1 Thread extraction

Threads are an essential part of the layout structure of newspaper. In fact they serve as separators between columns of text or between different entities. These entities can be articles or a group of articles. We extract them by taking into account connected component exceeding a certain width/height or height/width ratio δ.

3.2 Frame extraction

Frames are special kind of paragraphs we can find inside newspapers. In fact they are paragraphs surrounded by rectangles. In our algorithm we extract a connected component bigger than a given threshold θ, then we extract 4 rectangles north, south, east and west from this component. We compute for each of the previous rectangles the ratio (black pixel density)/(black + white pixel density). If the ratio of the rectangles is greater than a given threshold λ, then we have a frame candidate.

3.3 Image text separation

For the detection of images, the connected components are calculated. Every connected component is described

by its rectangular bounding box. In our algorithm, we extract a connected component bigger than a given threshold α, and then we compute for this connected component found the black or white density pixel. If the black or white density pixel is bigger respectively than a given threshold β or γ, then we have an image candidate.

3.4 Text line extraction

In the literature many methods are used for line extraction. We have used the basic RLSA (Run-length Smearing Algorithm) [12]. It merges any two black pixels which are less than a threshold apart, into a continuous stream of black pixels. The method is first applied row-by-row and then column-by-column, yielding two distinct bit maps. The two results are then combined by applying a logical AND operator between both images.

For text line extraction we have applied RLSA horizontally, and then we have extracted the connected components; the threshold has been chosen manually according to the text size.

The lines of the paragraphs are correctly extracted but within titles there are either diacritics points or certain characters not correctly merged with the lines found. In fact, it depends on the position of diacritics above or under the character and also of the *shadda*, treated as one of the diacritics points, as shown by the top arrow in figure 6. We have also noticed that the titles are not segmented in text line.

Figure 6. Problems with RLSA in text line extraction

Thus for the titles, the merge process has been extended into three directions: from right to left, from the text line found towards up and down direction for the diacritics and the *shadda*. The diacritics points can be found above or under the character, a special case for the shadda which is always located above the character. For these cases we've used three thresholds: μ for the merge of horizontal text line, φ and ψ for the diacritics either above or under the characters.

3.5 Line merging into blocks

In order to merge the lines, obtained in the line extraction phase, into blocks we have used some rules taking into account the characteristics of the Arabic language. Since the words are written from left to right, paragraphs in Arabic correspond to the transposed form of the Latin ones (see figure 7).

Figure 7. Paragraph disposition in Latin and Arabic.

In order to merge the text lines, we have used the following algorithm: (All the thresholds represented in Greek letters are the result of a statistical analysis made on a variety of sample documents).

1. A text line is extracted from the set of text lines. This one is represented as a rectangle, named main rectangle (r_M) from which we create two rectangles, north (r_N) and south (r_S), distanced by χ pixels from (r_M). The χ is determined by computing the distance average between lines within paragraphs.

2. Rectangles intersecting r_N, r_S are searched inside the set of text lines.

3. We merge the obtained text lines, unless we have the following conditions (a and b and c) or d verified:
   a. If the intersected rectangles have the same height as the main rectangle (for the core of the paragraph);
   b. And If $[r_N.width + r_N.x + ratio] - [r_M.width + r_M.x] <$ threshold (ς) for the first line of the paragraph;
   c. And If $[r_S.width + r_S.x] - [r_M.width + r_M.x + ratio] <$ threshold (τ) for the last line of the paragraph;
   d. There is also a special case for the last line of the paragraph which is find centered as illustrated in figure 8. This case is treated as follows: if the rectangles r_N and r_S have the same height and if $[r_S.width + r_S.x] - [r_M.width + r_M.x + 2*ratio] <$ threshold (ξ).

4. Go back to 1 for next text line.

4. Results

The algorithms presented in the previous section have been tested on a set of pages from ANNAHAR, a Lebanon Arabic newspaper (see figure 9) and ALHAYAT an independent Arabic-language daily newspaper published in London (see figure 10). The evaluation has been performed on TIFF images generated from PDF files. The choice was made for two reasons: first, for a scientific reason, in order to evaluate the methods on noise free images, and second, for a practical reason, because the developed method may be used to perform layout analysis of encrypted documents.

The viewing of the results of the experiments is done using the XMIllum environment [9]. Figure 11 shows the results of the segmentation of the document shown in figure 9 using an appropriate XSLT stylesheet under XMIllum. In fact, user can toggle on/off the view of the different layers: image text separation, threads, text line extraction and line merging into blocks.

The first prototype has been tested on four different phases of newspaper image analysis: thread recognition, frame recognition, image text separation, text line recognition, and line merging into blocks. On 50 pages of ANNAHAR and ALHAYAT newspaper, the following results have been obtained as shown in table 1.

Table 1. The results of the recognition.

%	thread	frame	image	text line	Line merging into blocks
Annahar	50,539	99,819	97,598	96,516	95,217
Alhayat	99,452	90,987	91,178	92,368	91,438

For threads of ANNAHAR newspaper and for images of ALHAYAT newspaper the low recognition rate is due essentially to the presence respectively of threads with textures and images surrounded with textures which aren't treated within the algorithm. For frames the decrease rate of ALHAYAT newspaper compared to the one obtained for ANNAHAR newspaper is essentially due to a certain type of frame (a non closed rectangle). For the text line the recurrent error is that there is an ambiguity especially when diacritics of the first line and those of the second line are near to each other or merged as shown in figure 5 and figure 12. In figure 8 we see a set of text lines correctly extracted from the paragraph. The recognition rate of line merging into blocks is based on the results of the text recognition. Therefore, the errors of the text line recognition are propagated to the next processing phases.

لندن – أب – دعـا رئيس الوزراء
الاسرائيلي أرييل شارون في مقابلة مع
صحيفة "التايمس" البريطانية الى وضع
ايران على لائحـة الدول التي يجب
توجيـه ضربة اليها بعد الانتهاء من
الضربة العسكرية المحتملة للعراق.
– التتمة في الصفحة ١٦ –

Figure 8. Text line extraction of Arabic paragraph

Figure 9. Page sample of ANNAHR Arabic newspaper

Figure 10. Page sample of ALHAYAT Arabic newspaper

Figure 11. Page segmentation results using XMIllum viewer.

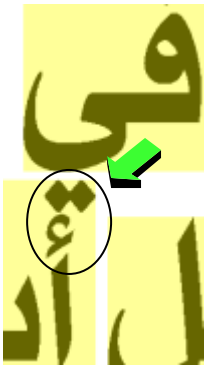

Figure 12. Diacritics points near characters

5. Conclusion

This paper presents our approach for segmenting Arabic newspaper images. Encouraging experimental results are reported concerning the adaptation of the well known segmentation algorithms to the Arabic case. Future work is the improvement of the adaptation, the right separation of the diacritics points between lines.

Despite the encouraging results obtained we believe that automatic layout analysis is not well suited for complex structured documents such as newspapers. The major difficulty of such kind of documents is the tremendous variability of the layout. The next step is to add user interaction in order to correct the results of the system and include incremental learning of contextual thresholds in order to avoid repetitive mistakes.

6. References

[1] N.E. Ben Amara, "Sur la problématique et les orientations en reconnaissance de l'écriture arabe", *CIFED'02*, Hammamet (Tunisie), October 2002, pp. 1-10.

[2] R. Cattoni, T. Coianiz, S. Messelodi and C.M. Modena, "Geometric layout analysis techniques for document image understanding a review", *Technical report, IRST*, Trento, Italy, 1998.

[3] B. Gatos, S.L. Mantzaris, A. Antonacopoulos, "First international newspaper segmentation contest", *ICDAR'01*, Seattle (USA), September 2001, pp. 1190-1194.

[4] B. Gatos, S.L. Mantzaris, K.V. Chandrinos, A. Tsigris and S.J. Perantonis, "Integrated algorithms for newspaper page decomposition and article tracking", *ICDAR'99*, Bangalore (India), September 1999, pp. 559-526.

[5] K. Hadjar, O. Hitz and R. Ingold, "Newspaper Page Decomposition using a Split and Merge Approach", *ICDAR'01*, Seattle (USA), September 2001, pp. 1186-1189.

[6] K. Hadjar, O. Hitz, L. Robadey and R. Ingold, "Configuration REecognition Model for Complex Reverse Engineering Methods: 2(CREM)", *DAS'02*, Princeton, NJ (USA), August 2002, pp. 469-479.

[7] H. Haouala and M.C. Fehri, "Arabic document analysis and recognition a case study: official journal of the Tunisian republic", *CESA'98*, Hammamet (Tunisie), 1998, pp. 9-12.

[8] R.M. Haralick, "Document image understanding: Geometric and logical layout", *Proc. Internet. Conf. On Computer Vision and Pattern Recognition*, 1994, pp. 385-390.

[9] O. Hitz, L. Robadey and R. Ingold, "An architecture for editing documents recognition results using xml technology", *DAS'2000*, Rio de Janeiro (Brazil), December 2000, pp. 385-396.

[10] J. Hu, R. Kashi, D. Lopresti, G. Nagy and G. Wilfong, "Why table ground truthing is hard", *ICDAR'01*, Seattle (USA), September 2001, pp. 129-133.

[11] G. Nagy, "Twenty Years of Document Image Analysis in PAMI", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No 1: January 2000, pp. 38-62.

[12] K.Y. Wong, R.G. Casey and F.M. Wahl, "Document analysis system", *IBM Journal of Research and Development*, November 1982, pp. 647-656.