SIDOP 2nd Workshop on Signal and Document Processing
March 23-24, 2012 - Hammamet - Tunisia

Extraction and Separation of Words From bilingual printed document

Rabeb Ben Abdelbaki and Sofiene Haboubi
SIDOP Research Group
Signal, Image and Information Technologies
National Engineering School of Tunis BP 37 Belvedere Tunis, TN-1002, Tunisia
benabdelbakira@yahoo.com
sofiene.haboubi@istmt.rnu.tn

Abstract—**in this paper, we present our work about the extraction and separation of words from bilingual printed document. This approach is based on the structuring element of the morphological dilation. We report results for Arabic, Latin and bilingual Arabic-Latin scripts and we show its limitations and present the possible improvements.**

Keywords-component; Script; Arabic; Latin; Discrimination; Separation; mathematical morphology; Dilation; Structural element

I. INTRODUCTION

The character recognition system, called OCR "Optical Character Recognition" allow to find the characters forming a text, o recognize them individually and then validate them by lexical recognition of words that contain them. In other words, an OCR is the process of scanning a paper document which leads to a digital text.

Due to the oneness of the language script within the same OCR, an important problem appears when the document is no longer monolingual. In fact, if the document is multilingual, then the OCR loses its ability to read the document because of the dependence of characteristics on the structural properties of the character, style and type of writing that generally differs from a script to another. Therefore, it's imperative to identify the languages present in the document in order to redirect it to the appropriate character recognizer.

In reality, we can't speak of discrimination between scripts without involving document's segmentation. In fact, the segmentation of documents into words is an important step in the process of document recognition; this phase becomes crucial in the case of multilingual document. It is the foundation of all the following steps; it increases also the efficiency of a recognition system.

Segmentation of text into words is occurred at the step of discrimination between several scripts for the local approach which is based on words as components to be studied; it requires prior knowledge of individual words constituting the document. It is also a necessary step to discriminate between printed and handwriting document and for any system of automatic processing of multilingual documents. This task is usually a delicate and complex task in the multilingual context given the large difference between the characteristics of different scripts in the form of letters, spacing, etc... Our work aims to develop a new method for the separation and extraction of words in a bilingual printed document. We will focus in the following on stating the characteristics of Arabic and Latin scripts. And then mention some related works. After that, we will present our method of separation and extraction of words from bilingual printed documents. And finally we will interpret the results of this work.

II. MORPHOLOGICAL CHARACTERISTICS OF ARABIC AND LATIN SCRIPTS

Script is the graphical presentation of a language through signs drawn on a support. Since its appearance around the third millennium BC, it has continued to grow with the languages it represents. In this paper, we will focus on the Latin and Arabic script.

A. Arabic script

The Arabic script is a consonantal script, composed of 28 letters, excluding the "hamza", which behaves either as a full letter or as a diacritic and the symbol "~" which is written only on the support of the character "ا".

The Arab character can have up to four different forms depending on its position in the word or in the pseudo-word as

it changes its design depending on its position: initial, medial, final or isolated.

The Arabic script is a semi-cursive writing. Letters are generally linked to each other and Arabic words can be composed of one or more pseudo-words written from right to left, both in printed or handwriting form.

Several Arabic letters have the same body and differ only at the number and location of diacritical marks. These diacritics can be above or below the baseline, in different places depending on the character, but never up and down simultaneously. (Figure 1)

In Arabic, there are 15 letters, presented in Table 1, among 28 of the alphabet, which have diacritical points.

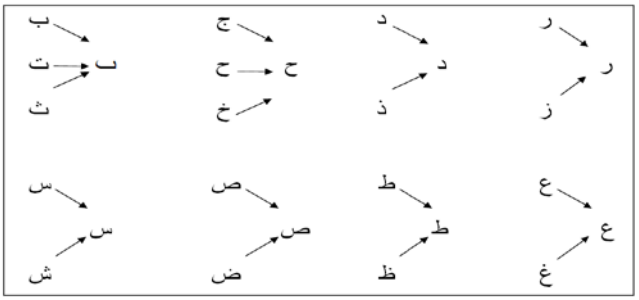

Figure 1: Characters and common body

Table 1: Letters with diacritical points

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ب	ت	ث	ج	خ	ذ	ز	ش	ض	ظ	غ	ف	ق	ن	ي

The Arabic script has no capital letters and Arabic characters include a loop that can have different forms.

Figure 2: Different forms of loops [Touj et al., 04]

In addition, Arabic script varies vertically and horizontally, because of the presence of horizontal and vertical ligatures between characters of the same word.

The Arabic word doesn't have a fixed length, it may include one or more pseudo-words called PAW (Piece of Arabic Word) each including a different number of characters.

The presence of pseudo-words in Arabic script increases the complexity of its segmentation. In fact, these PAWS induce to error segmentation's algorithms because they introduce important and variable intra-word spaces length compared to the intra-words space in Latin.

Figure 3: Examples of words containing different PAWs

B. Latin script

The Latin script uses two bicameral spellings for each character, one called lowercase, the other called uppercase or capital. In general, each grapheme possesses these two types of spelling with few exceptions changing from script to another.

The Latin alphabet has 26 basic letters. In Uppercase form, letters change shapes and sizes.

Unlike the Arabic alphabet, Latin alphabet consists of two types of graphemes, vowels and consonants. Latin characters are composed of 5 vowels presented in Table 2 and 21 consonants listed in Table 3.

Table 2: Latin's vowels

A	E	I	O	U

Table 3: Latin's consonants

B	C	D	F	G	H	J	K	L	M	N	P	Q	R	S	T	V	W	X	Y	Z

The Latin alphabet is one of the richest alphabets of national variations because of its geographic and temporal spread. Each Latin script is based on the fundamental letters of the Latin alphabet but it may have some specific letters considered as variants from the basics ones and those considered as new letters. Table 4 shows the different forms of a Latin grapheme.

Table 4: The different forms of a Latin grapheme

N°	Lettre	Formes	N°	Lettre	Formes
1	A	@ A a Ä Ā á à â ä ã å À Ã Â Á	14	N	N n Ñ ñ
2	B	B b	15	O	O o Ö Ô ó ò ô ö õ Ő Ó Ô Ò
3	C	C c Ç ç	16	P	P p
4	D	D d	17	Q	Q q
5	E	E e É é è ê ë Ë Ê È €	18	R	R r
6	F	F f	19	S	S s
7	G	G g	20	T	T t
8	H	H h	21	U	U u Ü ú ù û ü Ú Û Ù
9	I	I i Ü í ì î ï Ï Í Î Ì	22	V	V v
10	J	J j	23	W	W w
11	K	K k	24	X	X x
12	L	L l	25	Y	Y y
13	M	M m	26	Z	Z z

Identify applicable sponsor/s here. If no sponsors, delete this text box. (sponsors)

As for the Arabic script, several variants of Latin characters have diacritical signs, such as points above the body of the character, accents (acute accent, grave accent, circumflex), tilde, etc. But only two basics characters have diacritical points.

The Latin script is written from left to right, it's a non-cursive; its letters are isolated from one to another, separated by intra-words spaces in its printed form. The Latin alphabet has also many loops that can have different forms.

Table 5: Latin's loops

a	b	d	e	g	o	p	q	A	B	D	O	P	Q	R

III. RELATED WORKS

The document's words segmentation is an important phase in the document's recognition process. In fact, this phase is very crucial in the case of multilingual documents; it becomes obligatory to segment the document and identify its words individually.

This step is the foundation of all the following steps. The segmented words become the entries for the other steps of the recognition process. Despite the diversity of segmentation approaches and its richness on segmentation methods, the domain of document segmentation, especially in words, stills an open field and a powerful line of research that interests enough scientists. In fact, many researchers have focused on this axis. Our literature review led to a list of research in this area which we mention the most interesting.

- [Ma and Doermann, 03] used the Docstrum algorithm of O'Gorman for the segmentation of bilingual documents, applying it on Arabic-English, Chinese-English, English-Hindi, and Korean-English dictionary. This algorithm is a bottom-up approach based on the calculation of the k nearest neighbors for each connected component of the document. After the removal of noise, the connected components are separated into two groups according to a factor selected from the proportion of character sizes. One group consists of the characters most dominant and the other consists of the characters of titles and headers (or head) of sections. Then for each connected component, they seek the k nearest neighbors, each pair of these neighbors has an angle and an associated distance. By grouping the components through the features mentioned above, the geometric areas of physical structures of the document can be determined. The proposed method is independent of the change in orientation of the document and of the inter-words spacing. However, the value of k is dependent on the structure of the document.

- [Dhandra and al., 07] have segmented bilingual documents containing one of India's regional languages (Hindi, Kannada and Tamil) and English numbers, their method was based on the segmentation of text in different lines, then each line will be projected vertically and segmented into words based on the analysis of the valleys present in the vertical projection delimiting the different words.

- The work of [Chanda and al, 07] presents a segmentation method of bilingual documents containing Thai and English words. Their method is to encode the different lines of text depending on the position of black pixels in each line. After segmentation of the document in line, their method goes through the histogram of each vertical line and produces a 0 if it encounters two black pixels or less, if not the scan is valued at 1. The chain produced is then analyzed, if there is a set of 0 with minimum length equal to 2 * k1, mid-term is considered as the borderline for word segmentation. The value of k1 is an estimate of the white gap between two consecutive characters of a document's line.

- [Rezaee and al., 09] proposed a word segmentation method in bilingual documents containing English and Farsi scripts. Their method is based on image's directional projections and the analysis of some attributes like the gaps between words and thresholding from the peak distribution to then segment the text lines into words.

- [Da Silva and al., 11] proposed a method of word segmentation from Latin documents containing both the handwritten and printed form. This method is based on the segmentation of text on connected components and their extraction by cropping a bounding area. After extracting of the components, this method proceeds by fusion of near neighbors in the same line and having a distance between their bounding boxes less than a threshold "th" calculated by the following formula XX:

$$th = \frac{\sum_{i=0}^{k} L_i}{2k}$$

Where k is the number of frames and Li the widths of all the frames of the image.

- [Haboubi and al., 11] proposed a segmentation method for bilingual documents containing Arabic and Latin script, based on the use of mathematical morphology to delimit the different words in the text. This method uses the morphological dilation with a line structuring element. They use sequential dilation by increasing each time the size of the structuring element in order to determine a threshold that separates the spaces between words and intra-word spaces. This threshold corresponds to the dilation order where the number of connected components has a zero standard deviation.

IV. PROPOSED APPROACH

The approach developed for the word segmentation of printed bilingual documents, includes several steps (Figure 4). From a document's image, we begin with a preprocessing to prepare the scanned document to the segmentation, and then we move to the detection and extraction of text lines. After that

we analyze each line separately and we extract the different words presents in the document.

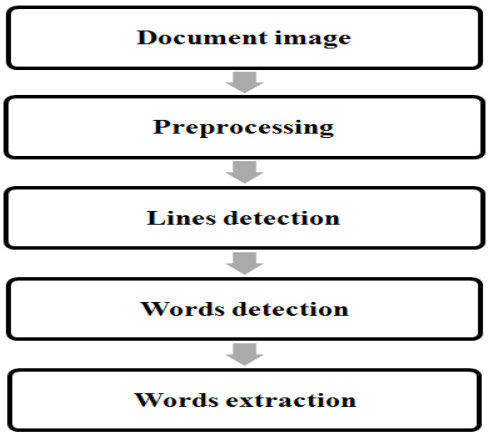

Figure 4: The process of word segmentation from a bilingual printed document

A. Pre-processing

The document image is the result of an acquisition step using a scanner. Our approach doesn't give much interest in the pre-processing because the proposed system should work with images preprocessed in advance by a dedicated pre-processing documents system such as the elimination of noise introduced sometime during the scan documents, the skew correction, deleting diacritics, etc.

However, our approach preserves the amount of information present in the text because we work with document images containing diacritical signs, given their important role in the understanding of the text, although their presence may increase the complexity of the segmentation task because of problems encountered in the detection and extraction lines. Indeed, in some writing styles, diacritics may exceed the upper or the lower limit of the line, which can error the step of detecting lines.

In our case, the pre-processing step is limited to the binarization of the document image and the values' inversion of black and white pixels in order to prepare the document to the step of detecting lines.

B. Lines detection

This phase is rather difficult in the case of bilingual documents because of the large variability between Arabic and Latin printed scripts, and it becomes more complex with the presence of diacritical marks. The morphological study of Latin and Arabic scripts shows the presence of a significant number of diacritical signs.

We have chosen to use the projection method to delimit the horizontal lines. This method corresponds to the needs of document's segmentation because we handle text documents with a simple structure. Our proposed approach goes through the image horizontally and calculates the value of black pixels in each row of the matrix representing the image. Next, we have analyzed the histogram of projections, if the number of black pixels has changed its value from 0 to a positive one then this position is the lower limit of a line. We kept, each time, the positions of the white areas that will be used for cropping the image into different lines.

C. Words detection and extraction

The document segmentation allows to segment documents at different levels, either characters, or pseudo-words, or words. This level of segmentation is the most difficult among the others, given that the segmentation has to differentiate between different types of spaces between characters, between pseudo-word and between words, which is not always obvious to a word extraction system.

The objective of the proposed approach is to segment the document image in order to separate and extract the words of a bilingual printed document. Segmentation methods segment documents into connected components; either character in the case of a Latin printed document, or pseudo-words in the case of an Arabic printed document, because the non cursive and semi cursive Latin and Arabic printed writing.

Our approach uses mathematical morphology for the elimination of intra-word spacing and the building of connected components formed by different words in the bilingual printed document. We used the morphological dilation to enlarge the image by filling the holes corresponding in our case to the intra-word spaces.

To achieve this goal, we must determine the best structural element able to stick the different characters of a word, without sticking words together. At this level, two major problems appear. The first on is the size of the structural element and the second on is its shape. The determination of these two characteristic features of the structuring element is the foundation of our work.

Choosing only one size of the structuring element for each document cannot give a performing segmentation because the intra-word spaces differ from one font to another, and depend on the size of the font. Similarly, the spaces between words depend on the text alignment, especially in the case of justified text. Moreover, a document can contain different fonts and sizes.

The shape of the structuring element solves the problem of extracting diacritics as separated words because our method doesn't eliminate these signs during pre-processing. A solution of this problem is to stick diacritics to their words. So, we opted to shapes that have a height, and we have searched the corresponding height that solves the problem with minimal line's changes.

The approach developed proceeds line by line to find each time, the size and shape of the structuring element of the dilation that can separate and extract correctly and with minimal changes the different words in a document line. Indeed, we proposed three methods to calculate the size of the structuring element and selected four specific shapes for the structuring element.

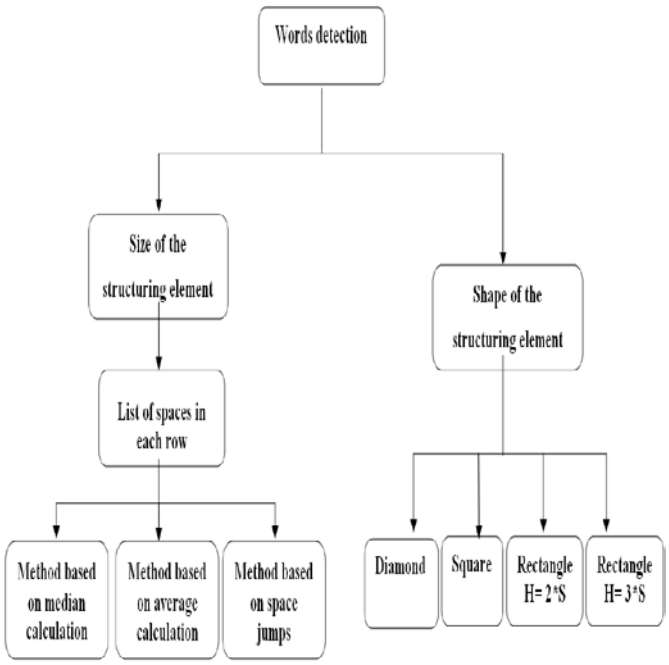

Figure 5: Proposed methods for the document segmentation

Our approach consists in testing all the combinations of methods for calculating the size of the structuring element and its forms. In fact, we fix each time the calculation method and we vary the shape. We begin by applying a combination of Latin and Arabic printed documents. If we get good results, we continue testing on mixed documents. Otherwise, we consider it unnecessary to apply the combination to bilingual documents.

- *Size's calculation of the structural element*

To calculate the size of the structuring element, we started by determining the list of spaces in each line, then, we developed three different methods to solve this problem.

➢ Identification of document spaces

Our approach proceeds by analyzing the extracted lines. This analysis is based on the calculation of the vertical projection histogram to determine the values of the different spaces in the document. We cover vertically each line and we calculate the number of black pixels presents in each column of the line's image.

The next step is to analyze the vertical histogram obtained and to determine the positions of the spaces inter and intra words and their values, if the number of black pixels becomes zero after a sequence of non-zero black pixels then this change corresponds to the presence of a space in the line. We store its position and calculate the length of this area. In fact the value of the space or its length corresponds to the distance between the position of appearance of this space and the position of the first non-zero value of the number of black pixels encountered in running through the vertical projection histogram. We obtain at the end a list composed of space values present in the considered line and a list with their positions.

➢ Methods for calculating the size of the structuring element

After the detection of spaces in each line of the document, we determine the size of the structuring element according to the three proposed methods.

o Method based on median calculation

This method proceeds by elimination of redundant spaces values presents in the considered line then sort the new list of spaces in ascending order to permit the interpretation of these values.

This list reflects the nature of spaces contained in the processed image. It begins with the relatively small areas, which actually represent the spaces between characters in a word, and reaches the largest gap present in the line. This method is based on the fact that the threshold space, able to stick the characters of a word without sticking the words together, has an intermediate value between the lower and upper bound of the new list of spaces. The median value of this list is considered as the size of the structuring element of the dilation.

o Method based on the average calculation

This second method is similar to the previous one in the determination of distinct values of spaces. It is based on the fact that the threshold value of the structural element of the dilation is proportional to the number of spaces present in the image given and their lengths. Indeed, this method sets the size of the structuring element to the average lengths of different spaces in the line introduced.

o Method based on the calculation of the difference between the values of spaces

This method is based on the detection of larger jump lengths between spaces. It works on the entire list of spaces in line to be processed. From this list, it calculates the different lengths of jumps in spaces. Then, it covers the new list by determining the greatest difference between spaces. The threshold size of the structuring element is necessarily located between the areas that have generated the biggest jump. Difference method associates the size of the structuring element the average between the two spaces relative to the largest jump determined.

This calculation method of the structuring element is to generate a list of jumps between different lengths spaces, looking for the biggest jump, find two spaces related to this jump and calculates their average. The size of the structuring element corresponds to this average.

- *Determination of the structural element's shape*

After calculating the size of the structuring element, comes the phase of the choice of suitable form which allows segmenting the bilingual printed document correctly.

The structuring element can have several forms such as square, diamond, polygon, Euclidean disc, line, point pairs, rectangle, etc.

We were interested in this approach to four specific forms of the structuring element, the first is the diamond shape, and the second is the square, the third and fourth are variants of the rectangle shape with some differences in the input parameters.

We used these shapes because of the presence of diacritical signs in the documents to be segmented. These forms have in common a height proportional to the size of the structuring element, an important feature in our approach in order to stick the different diacritical marks to their words.

- The diamond shape

For each line, the diameter of the diamond is equal to the size of the structuring element determined by one of the three calculation methods proposed later.

The following table shows the results found by combining the diamond shape with the three methods for calculating the structural element.

Table 6: Results of the diamond shape

Shape of the structural element	Method of calculation	Script	Good extraction rates
Diamond	Median	Arabic	26,23%
		Latin	59,87%
	Average	Arabic	49,72%
		Latin	82,82%
	Difference (jump)	Arabic	63,85%
		Latin	91 ,81%

- The square shape

For each line, the length of the square is equal to the size of structuring element determined by one of the three calculation methods proposed later.

The following table shows the results found by combining the square shape with the three methods of calculating the structural element.

Table 7: Results of the square shape

Shape of the structural element	Method of calculation	Script	Good extraction rates
Square	Median	Arabic	28,62%
		Latin	53,69%
	Average	Arabic	62,57%
		Latin	85,23%
	Difference (jump)	Arabic	71,93%
		Latin	93,56%

- Rectangle shape with height 2 times the size

For each line, the width of the rectangle is equal to the size of the structuring element determined by one of the three calculation methods proposed later and height is equal to two times this value.

The following table shows the results found by combining the shape rectangle of height 2 times the size calculated with the three methods of calculating the structural element.

Table 8: Results of the rectangle shape with height 2 times the size

Shape of the structural element	Method of calculation	Script	Good extraction rates
Rectangle with height 2 times the size	Median	Arabic	61,10%
		Latin	67,25%
	Average	Arabic	88,07%
		Latin	93,29%
	Difference (jump)	Arabic	94,13%
		Latin	95,84%

- The rectangle shape with height 3 times the size

For each line, the width of the rectangle is equal to the size of the structuring element determined by one of the three calculation methods proposed later and height equal to 3 times this value.

The following table shows the results found by combining the shape rectangle of height three times the size calculated with the three methods of calculating the structural element.

Table 9: Results of the rectangle shape with height 3 times the size

Shape of the structural element	Method of calculation	Script	Good extraction rates
Rectangle with height 3 times the size	Median	Arabic	72,11%
		Latin	72,35%
	Average	Arabic	92,11%
		Latin	95 ,03%
	Difference (jump)	Arabic	94,86%
		Latin	97,05%

V. INTERPRETATIONS

The following table represents the best rates achieved for each form of the structuring element.

Table 10: Best rates achieved for each form of the structuring element

Shape of the structural element	Method of calculation	Script	Good extraction rates
Diamond	Difference (jump)	Arabic	63,85%
		Latin	91 ,81%
Square	Difference (jump)	Arabic	71,93%
		Latin	93,56%
Rectangle with height 2 times the size	Difference (jump)	Arabic	94,13%
		Latin	95,84%
Rectangle with height 3 times the size	Difference (jump)	Arabic	94,86%
		Latin	97,05%

We note that the best good extraction rates are obtained for 94.86% and 97.05% Arabic to Latin. These rates are achieved by the combination of the method of calculating the structuring element's size based on the difference between the spaces values and the rectangle shape with a height equal to 3 times the size of the structuring element.

The application of this combination on the sample printed bilingual documents gave a good extraction rate equal to 94.85%. This result is explained by the adequacy of method of calculating the size of the structuring element to changes in the lengths of spaces between the lines and document and height, proportional to the size, the different distances of diacritics their words.

The figure shows a sample run of a line from a printed bilingual document with diacritics.

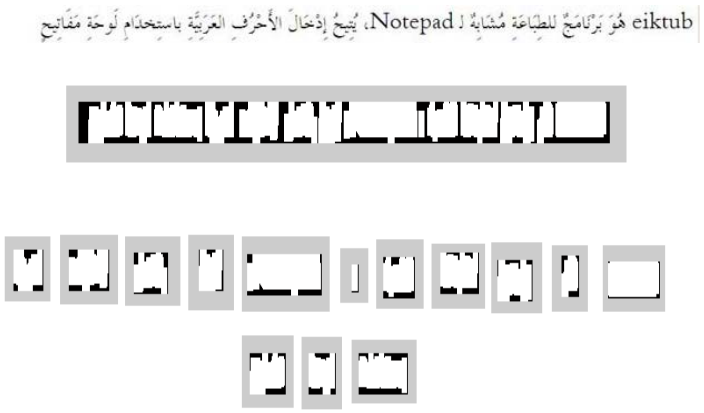

Figure 6: Example of word segmentation from a bilingual printed document

The word segmentation of the printed bilingual document gave 14 words, which correctly corresponds to the words found in the line of the document introduced.

VI. CONCLUSION AND PERSPECTIVES

The separation and extraction of words in a printed bilingual document constituted the main contribution of our recognition's area, its different stages, and the various available methods of documents segmentation into words.

After studying the Arabic and Latin scripts we have proceeded to the implementation of our approach. We have developed different methods for calculating the size of the structuring element of morphological dilation, combined with different forms and tested on samples of printed Arabic and Latin documents. After that, we have compared the results, and the best performing combination was chosen to testing printed bilingual documents, subject of our study.

Although the result obtained by this method used for the separation of words is compelling, it has some limitations:

• The sample size is 945 words for the printed Latin documents, 545 words for Arabic printed documents and 564 words for printed bilingual documents;

• This approach only deals with the printed documents;

• This approach is limited to bilingual Arabic and Latin documents.

In perspectives, we expect to enlarge the sample size to better test the performance of the proposed method. We also plan to extend our method to the processing of textual handwritten bilingual documents, to mixed bilingual documents (both handwritten and printed forms in the same document) as well as treatment of bilingual documents of any kind and even that of multilingual documents.

REFERENCES

[Touj and al., 04] Sofiene Touj, Najoua Essoukri Ben Amara, Hamid Amiri, « Reconnaissance de l'Ecriture Arabe Imprimée par Transformée de Hough Généralisée », dans Conférence Internationale Francophone sur l'Ecrit et le Document (CIFED 04) 2004.

[Ma and Doermann, 03] : Huanfeng Ma, David Doermann, « Gabor Filter Based Multi-class Classifier for Scanned Document Images », Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03) 0-7695-1960-1/03 $17.00 © 2003 IEEE.

[Dhandra and al., 07] : B.V. Dhandral, Mallikarjun Hangarge, Ravindra Hegadil and V.S. Malemathl, « Word Level Script Identification in Bilingual Documents through Discriminating Features », International Conference on Signal Processing, Communications and Networking, 2007. ICSCN '07.

[Chanda and al, 07] : S. Chanda, Oriol Ramos Terrades and U. Pal, « SVM Based Scheme for Thai and English Script Identification », Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) 0-7695-2822-8/07 $25.00 © 2007 IEEE.

[Rezaee and al., 09] : Hamideh Rezaee, Masoud Geravanchizadeh, Farbod Razzazi, « Automatic Language Identification of Bilingual English and Farsi Scripts », IEEE International Conference on Application of Information and Communication Technologies (AICT), 2009.

[Da Silva and al., 11] : Lincoln Faria da Silva, Aura Conci, Angel Sanchez, « Word-Level Segmentation in Printed and Handwritten Documents », publié dans IEEE 18th International conference on Systems, Signals and Image Processing (IWSSIP), 2011- Sarajevo.

[Haboubi and al., 11] : Sofiene Haboubi, Samia Snoussi Maddouri, Hamid Amiri, « Discrimination between Arabic and Latin from bilingual documents », publié dans IEEE International Conference on Communications, Computing and Control Applications (CCCA), 2011.