International Journal of Informatics and Applied Mathematics
e-ISSN:2667-6990 Vol. 2, No. 1, 48-72

Physical structure extraction of Algerian baccalaureate transcripts

Abderrahmane Kefali[1,2], Ahlem Obeizi[1], and Chokri Ferkous[1,3]

[1] Computer Science Department, University 8 Mai 1945- Guelma
[2] LabGED Laboratory, Badji Mokhtar University
[3] LabSTIC Laboratory, University 8 Mai 1945- Guelma
kefali.abderrahmane@univ-guelma.dz
ahlemobeizi@yahoo.com
ferkous.chokri@univ-guelma.dz

Abstract. In recent years, Algerian universities have become aware of the interest of electronic archiving and the digitization of archives for a better management of their documents. The development of systems enabling the analysis and understanding of archival documents became an unavoidable need. The present paper follows this trend; it proposes a system for the analysis of the physical structure of Algerian baccalaureate transcripts, stored in the universities archives. The proposed system proceeds in two phases: 1) preprocessing, in which several operations are applied in order to reduce the noise present in the input images. 2) Segmentation; It starts with the elimination of the transcript border. Then, it extracts the text lines and the blocks, based on RLSA algorithm and the projection profiles analysis. After, it proceeds to the classification of the blocks in three: textual block, table, and graphic. Finally, it recovers textual content from textual blocks and tables.

Keywords: structure of document · document understanding · segmentation · document image.

1 Introduction

Important documentary collections currently exist in libraries, societies and archives. The preservation and archiving of these documents and their access to a large number of people is today an unavoidable need. In fact, with the development of technology and the world, paper documents have been replaced in transactions by electronic documents, which allow to easily extracting and reusing information. A tendency to convert existing paper documents into electronic documents has even appeared, by digitizing paper documents, in order to take profit of electronic document advantages. Digitization is therefore the adopted solution; it allows the original document to be reproduced with sufficient quality as an electronic document for long-term preservation and communication, but it only provides images of documents, which is not always sufficient. Indeed, it is often necessary to access to the contents of digitized documents and possibly modify them. This is the purpose of Document Analysis and Recognition. Document analysis and recognition consists of converting a paper document into an electronic document based on the analysis and interpretation of document. This is the reverse process of document production. This discipline includes a set of computer techniques whose purpose is to reconstruct the content of a document from its image to facilitate the retrieval, indexing and automatic classification of documents. In order to achieve this goal we must know the studied documents to analyze their structures.

In most cases, the documents to be analyzed contain several pieces of information, not only the text or the written characters, but also the type and size of the used font, the writing color, in addition to other additional information describing the organization and the structuring of the different elements of the document. Without this additional information provided by the structure of a document, the correct reading or localization of the document would be impossible. Therefore, the understanding of a document requires the recognition of its structure in addition to its text and then any information constituting this document can be located correctly [19]. Noting that two levels of structure exist in the documents: the physical structure which describes the layout of the document, the different text blocks, their arrangement relative to each other, etc. and the logical structure which is a designation to the semantic content of the document and thus the correspondence between the physical regions and their function.

In fact, as other several administrations, the digitization and dematerialization of archives is a current trend of Algerian universities. This digitization allows to preserve the records of students, employees, etc., the invoices, purchase orders, mission orders, etc., in an electronic form but this digitization is not enough. It must be accompanied by methods and techniques that facilitating their automatic analysis and search.

The present work fits into this context. We are interested in analyzing the images of Algerian baccalaureate transcripts, which is one of the most important documents in the students file. And as we said previously, the analysis of a document is only possible through the recognition of its structures; we propose

in this paper a system aiming to the segmentation or the extraction of the physical structure of the Algerian baccalaureate transcripts.

The remaining of this paper is organized as follows. We first present an overview of the document segmentation approaches existing in the literature. In a second phase, we describe the characteristics of Algerian baccalaureate transcripts. Then, we present our proposed approach for the segmentation of the Algerian baccalaureate transcripts while detailing the various steps included, before concluding.

2 State of the art

The segmentation or the analysis of physical structure of documents has generated a great number of researchs in the literature. Several segmentation methods have been proposed and several review papers and comparative studies have been published. However, as announced [13] and [26], the proposed methods and techniques are classified into three main classes or approaches: bottom-up, top-down, and hybrid. [38] adds a fourth approach to multiscale resolution methods.

2.1 Bottom-up approach

The bottom-up approach is guided by data. The methods of this approach start with the lowest level until reaching the highest level in the document page. That means that they start at the connected components level, merge them into words, and then merge these words into lines, the lines in blocks, until the page is completely reconstituted. In this approach, several aspects have been exploited and various techniques and algorithms have been proposed.

One of the main bottom-up methods is Run Lengh Smoothing Algorithm (RLSA) proposed by Wahl et al. in 1982 [40]. This algorithm consists in blackening in a given direction, the segments of white pixels of length less than a given threshold. The segmentation is then obtained by applying the logical operator AND on the two images respectively resulting from horizontal smoothing and vertical smoothing. Thus, several authors have tried to extend RLSA algorithm in order to use it for the segmentation of more complex documents. Yamashita [43] for example proposes an RLSA smoothing algorithm with an adaptive threshold. As a result, a slightest change in the spacing between words and in the fonts size has little effect on the segmentation result. Shi and Govindaraju [33] describes a fuzzy RLSA for line segmentation of complex handwritten documents. In [32], a modified version of RLSA called spiral RLSA has been proposed to extract graphics from document images.

Another important category is that of methods using connected components. Fisher [14] combined the connected components extraction with a smoothing algorithm. The physical structure blocks are formed from the connected components and their enclosing rectangles, based on a set of features of the connected components. Another method using connected components has been presented

by Drivas in [4]. This method consists of a set of algorithms. One of these algorithms allows the segmentation and the other do the labeling of the obtained blocks in text or figure. Voronoi diagrams have also been exploited. One of the methods based on the approximate surface of Voronoi diagrams has been presented by Kise et al. in 1998 [21]. This method includes the following steps: extraction of the sampling points located on the connected components edge, noise elimination, generation of Voronoi diagram by using the obtained sampling points, and finally removal of the Superflus edges of Voronoi. Agrawal and Doermann have improved the original Voronoi algorithm with Voronoi ++ which adapts Voronoi parameters to the local spatial context [1]. Then they proposed in [2] a fuzzy version of it (with fuzzy edges) called CVS.

Other methods of segmenting technical documents are based on Clustering technique. O'Gorman [29] introduced "Docstrum" technique, which is based on the combination of bottom-up analysis and Clustering based on k-Nearest Neighbors. Faure and Vincent [17] used geometric clustering to segment historical documents. The interesting add that they have is the use of a trust value for each alignment (line of text) and a post-processing for conflict resolution in case of incoherence between two lines of text. Methods using window-based filtering rely on a scan of any size window on the entire document image. Lebourgeois [22] uses a 8 3 pixel filter. The sampled image is dilated by a horizontal structured element to gather adjacent characters. Each connected component is defined by an enclosing rectangle and by the average of lengths of black pixels ranges. If the connected component is within the range, it will be classified as text area that will be merged into blocks, otherwise it will be classified as a non-text area.

2.2 Top-down approach

The top-down approach is often used for documents with a well-defined structure. The segmentation in this approach is based on a strong prior knowledge of the document model to cut it into increasingly fine blocks. These methods start with the highest level (the entire image) until reaching the lowest level (connected components). In this approach, several algorithms may be derived.

An example of an algorithm using the top-down strategy is the famous X-Y cut algorithm [27], which is more suitable for structures of Manhattan type. The basic assumption is that the structured elements of the page are usually presented in rectangular blocks, but also that the blocks may be divided into groups so that adjacent blocks, in a group, have a dimension in common. The document is successively divided into small rectangular blocks by alternating horizontal and vertical cuts along white spaces found using a projection profile threshold. Several improvements to the X-Y Cut algorithm have been proposed. In [3], topological rules have been introduced to allow the hypothesis of rectangular blocks to be relaxed by allowing polygons through the segment tracking method. Sylwester and Seth [36] proposed to use a learning module in order to dynamize the division thresholds. In 2012, Ouwayed and Belaid [30] used projection profiles to segment multi-oriented documents. They first made an alignment of the document with rectangles. Then they calculated the projection profile of

each rectangle according to several directions. Next, they used heuristics combined with local projection profiles to detect regions with non-homogeneous text orientation and lines of text.

Some other methods try to segment the document by identifying the straight lines in the document. Authors in [25] divide the connected components horizontally into blocks based on the average height of the characters. Once this partitioning is done, they apply Hough transform on the centers of gravity of each block to detect the lines of text. In 2015, Wang et al. [42] tried to reconstruct the borders of frames in comics to segment them. Their algorithm is able to segment frames with only two apparent boundaries, but is limited to quadrangular regions. They separate the background, then they use another algorithm to adjust the quadrangles to the candidate frames. This is followed by a classification of the complexity of the frame and specific heuristics are used to complete the border of the frame.

Image background analysis has also been exploited. One of the first works based on the analysis of the white areas is the work of Spitz [34]. Its principle is to search for white flows in both vertical and horizontal directions, and to exploit them as generic delimiters of structures. Antanacopoulos [5] proposed a method, which consists of finding the longest range of pixel values in the vertical direction to blacken areas. However, it uses rectangles of different sizes to cover the image background. The extraction is done by considering the edges of the rectangles coinciding with the edges of blackened areas. Chen et al. [9] analyze white space to segment the document into text columns. Connected components are grouped into horizontal strings to create whitespace between these strings. Then, they are grouped vertically to create white row / column separators. This algorithm won the two segmentation competitions of ICDAR 2013 [16].

Other methods rely on the structure description by a grammar. Couasnon designed in 2006 a method called DMOS (Description and Modification of Segmentation), consisting of a new grammatical language and an associated analyzer [11]. The proposed language can describe any layout and the associated analyzer recognizes this layout in an image. In 2008, Lemaitre et al. [24] improved this work by adding a multi-resolution approach that made it flexible enough to segment handwritten letters and identify lines of text in administrative documents in French and Bengali. Carton et al. [8] then continued this work with an interactive learning step that may create a exhaustive set of models for a large set of data.

2.3 Hybrid approach

The hybrid approach combines bottom-up analysis to extract local primitives and top-down analysis to search for global primitives. In this approach we distinguish various subclasses of methods.

A large majority of methods collaborate several bottom-up and top-down techniques for better results. For example, in 1989, Wang [41] describes a method based on the combination of RLSA algorithm and the recursive X-Y Cut algorithm to extract homogeneous rectangular blocks from a newspaper's page. The

blocks are then classified according to the statistical textual features and decision techniques of the space. Another more recent significant work was the Multilevel Homogeneous Structure (MHS) method developed by Tran et al. [37] which won the ICDAR (International Conference on Document Analysis and Recognition) complex document segmentation competition in 2015. This method works by iteratively classifying the connected components according to the analysis of homogeneous regions at several levels and the white spaces.

Syntactic analysis which has its origins in compiling computer programs has been approached for the recognition of physical structures of documents. The system proposed in [28], begins with the segmentation of the document using the X-Y Cut algorithm. Then, syntactic analysis is applied in order to extract the physical blocks. Indeed, the syntactic analysis is done in order to generate a grammar by type of documents. This grammar expresses the predefined structure conventions of the journal's publications. A similar approach was used by Viswanathan in [39]. Thus, the pages of technical journals are analyzed according to a syntactical approach in order to hierarchically identify their spatial structure. Publication-specific knowledge is used in block segmentation. The information is meticulously encoded as block grammars used to describe the relationships between different classes of entities or blocks.

Another subclass of hybrid methods is that of split and merge methods. The method proposed in [31] allows to distinguish between bitonal and non-bitonal regions while allowing the separation between text and graphics. It begins with a descending segmentation to create an over-segmentation of the document. Then the similar and close segments are merged to form a region. Finally the separation between text and graphic is done using the black pixel density criterion. Stamatopoulos et al. [35] design a procedure to improve the performance of individual segmentation algorithms by combining their results. The procedure is based on overlapping regions produced by the algorithms. All areas less than 90% of overlap are split based on their intersection, followed by a merging from the regions with the highest overlap.

Other methods make allow to segment a document image from different representations of it. These representations are obtained for different levels of resolution of the original image. In [23] for example, a two-level analysis of resolution is performed for line segmentation of text. The first resolution allows to find the main orientation of the lines. The second is used to allow the extraction of precise features on the connectivity of the inter and intra-line components. The features extracted on these two levels are then combined by a rule-based method to extract the lines of text.

3 Characteristics of Algerian Baccalaureate transcripts

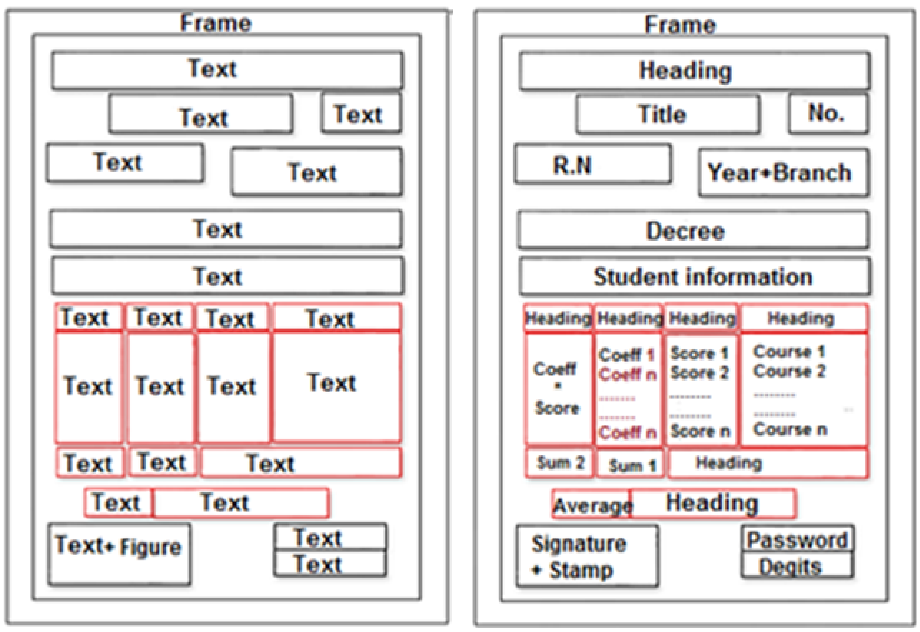
After the physical analysis of the various baccalaureate transcripts, distributed over different years (from 1997 to 2015), we noticed that almost every year the format of the transcripts changes (see Fig. 1), but the data remain the same Frame, Heading, Registration number (R.N) of the student, Student Informa-

Fig. 1. Examples of baccalaureate transcripts of various formats.

tion, Year, Branch of study, Scores table, etc.. However, the variations are at several levels; for example, at the level of paper quality (standard or special paper), the writing font, the language with which the student's information is written (Arabic or French), the stamp and the signature, the text and background colors,etc.

According to Fig. 1, it can be noted that the scores table is usually in the middle of the baccalaureate transcript and the average table is below it. It should also be noted that the frame of these two tables is a simple rectangle or a rectangle with rounded corners. Then there are several branches of study in high schools in Algeria, and the branches are different from each other by the number and the content of courses.

From this physical analysis, we extract the physical and logical structures of the Algerian baccalaureate transcript. They are shown in Fig. 2.



(a) (b)

Fig. 2. Physical and logical structure of an Algerian baccalaureate transcript, (a) Physical structure, (b) logical structure [19, 20].

4 Proposed system

We describe in this section the proposed approach for the segmentation of Algerian baccalaureate transcripts. The proposed approach consists of several pro-

cessing steps grouped into two main modules: preprocessing and physical structure extraction, as shown in Fig. 3.

The techniques and methods adopted in the system are all state of the art techniques that proved their ability and effectiveness in the literature. They have been chosen for their effectiveness and simplicity.

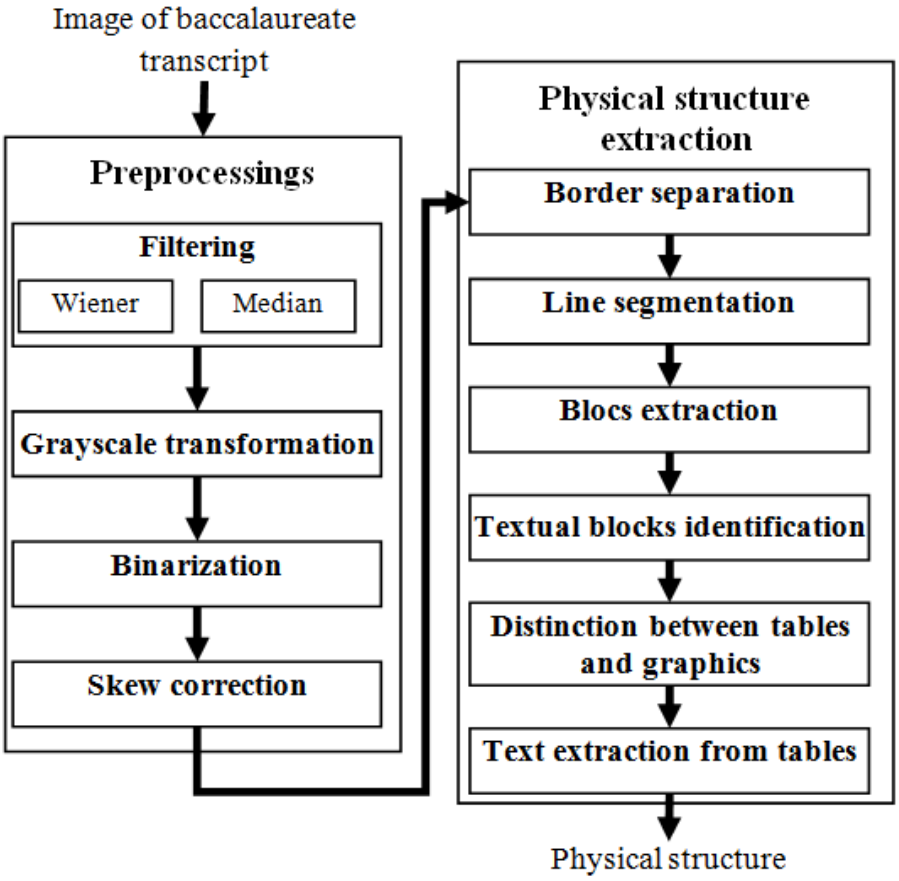


Fig. 3. Block diagram of the proposed approach.

4.1 Preprocessing

Preprocessing gathers a set of operations aimed to eliminate the noise, reduce the degradations, and preserve only the useful information of the image. This will prepare the ground for the next steps of processing. In our approach, the preprocessing includes filtering, grayscale transformation, binarization, and skew correction.

Filtering To improve the quality of a document image, several filtering may be applied. In our case, the purpose of the filtering is to make better, the details in the transcript image, in order to have good foreground/ background separation results. For this, we use two filters which are very effective for documents preprocessing:

Wiener filter: this is one of the most used filters for restoring document images [15]. This filter is effective for processing images whose small details are not enough present. Wiener filter increases the contrast between the texture and the background while smoothing the background.

Median filter: Although Wiener filter improves the quality of the document; it does not behave well when the image is strongly noisy. For this, we propose to apply another filter namely the median filter. This filter eliminates impulsive noise where the pixels become randomly scattered on the image surface by generating parasites.

Grayscale transformation this transformation is done simply by replacing the color of each pixel of the image by the average of its values of red, green, and blue.

Binarization it allows to separate the foreground from the background of the image which produces two classes of pixels: background (in white) and scene (in black). In fact, a large number of binarization techniques have been proposed. In our system we chosen to use the method of Kefali et al. described in [18]. This is a ANN (Artificial Neurons Network) based method, proposed for the binarization of old document images. In this method, the separation of image pixels to "blacks" or "whites" is performed by a Multilayer Perceptron (MLP) trained with back-propagation. Thus, the MLP does not compute or learn any threshold but runs a direct binarization by classifying the image pixels into two classes, because the binarization is a kind of two-class classification problem. However, an MLP is first defined, and then trained over a set pixels from some specific areas from different documents. For each pixel, local information (the grey values of the pixel with those of its neighbors) and global information (global mean and global standard deviation) are used as features to train the MLP.

The corresponding expected output is the binary value (black or white) of the related pixel in the ground truth image. After training, the MLP is able to output a binary value for all pixels of the image. The binarization result is shown in Fig.4.

Skew correction the techniques that we will use for segmentation of the baccalaureate transcripts are sensible to the inclination, and because some of our documents are inclined, a step of skew correction is required. However, we used a simple skew correction method based on Radon transformation [7]. The choice

(a) (b)

Fig. 4. Binarization of an Algerian baccalaureate transcript, (a) grayscale image, (b) binarized image.

of Radon transform is justified by its ability to describe the orientation of the straight lines (which are present in the baccalaureate transcripts and form tables), the simplicity of its implementation, and its independence from the presettings operations [6]. Radon transform is a tool allowing to plot the projection histogram of the pixels according to well-defined orientations. According to [12], Radon transform is defined by the following equation:

$$f\,(p,\theta)\int_{-\infty}^{+\infty} f(p.\cos(\theta)-s.\sin(\theta)+s.\cos(\theta)).d \tag{1}$$

Where θ is the projection angle, p is the coordinate of the point P on the pixels projection hyperplane and s is the coordinate of the point P along the perpendicular to this hyperplane.

Thus, Radon transform allows to concentrate the sum of pixels intensities of a straight line at a point of the transformed space. A straight line in an image is then transformed into a point of high intensity in the Radon space, as shown in Fig.5.

Since Radon transform is only applicable to a square image, we proposed to apply it to a square portion of the image with a surface equal to the width of the

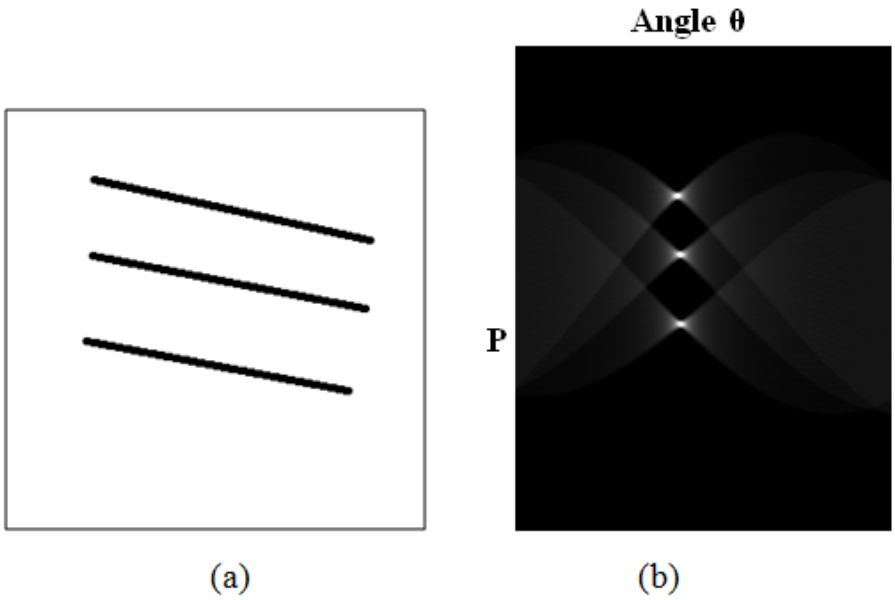


Fig. 5. Radon transform of an image, (a) image containing straight lignes, (b) Corresponding Radon space.

image width of the image. Moreover, according to the physical study that we conducted, we noticed that the inclination in our transcripts can be of negative angle, real (for example 0.6°, -3.2°, etc.), and that it never exceeds ± 15°. Thus, in order to cover all possible situations and to reduce the execution time, we proposed to customize Radon algorithm so that it takes into account the angles of real degrees between -15°and 15°.

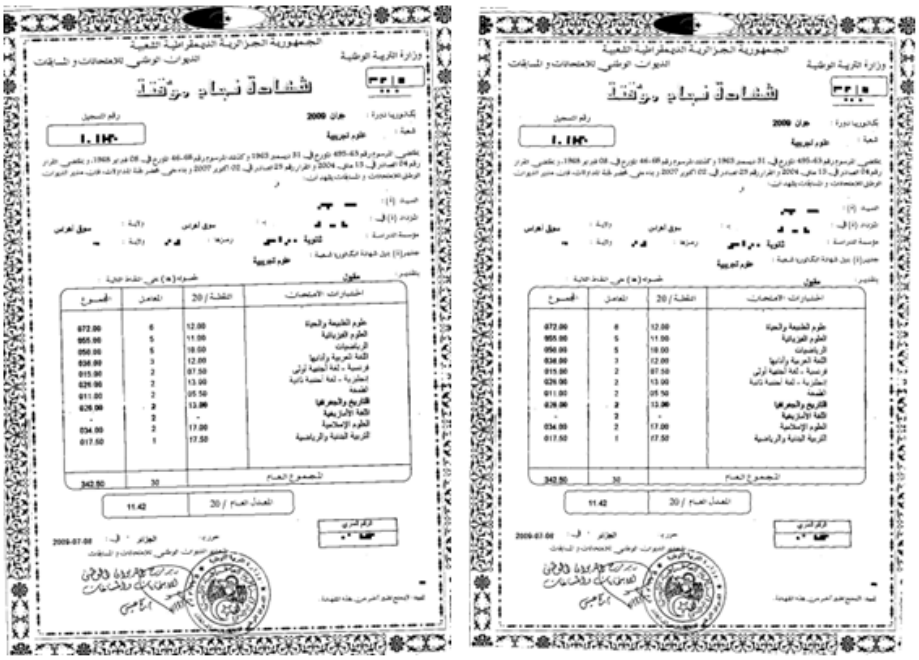
From the obtained Radon space, the inclination angle θ may be extracted easily and all that remains is to rotate the document image of angle θ. Fig. 6 illustrates the result of this step for a skwed document.

4.2 Physical structure extraction

The physical structure of the transcript is organized, as shown in Fig. 2, in blocks. There are textual blocks, composed of one or more text lines, corresponding to the student information, title, etc., and non-textual blocks. The non-textual blocks can be tables (of notes and of average), or graphics (stamp and signature, etc.).

To extract the physical structure of the baccalaureate transcripts, we precede in our approach a hybrid segmentation. First, the border of the transcript is separated from the image using a bottom-up technique because it carries no relevant information. Then, the lines and the blocks are extracted from the image of transcript without border. After that, the textual blocks are identified

and the other blocks are localized. Finally, the localized tables are segmented and their information is extracted.



(a) (b)

Fig. 6. Skew correction of an Algerian baccalaureate transcript, (a) binarized image, (b) deskwed image with an angle of 1.35°.

Border separation according to the physical study of the baccalaureate transcripts that we conducted, we noted that the most of transcripts contains different formats of the frame surrounding the document information: frame in the form of a rectangle and so it is formed of a single connected component, frame in the form of a series of stars or other geometric shapes, etc. There are also other transcripts that do not contain any borders.

To separate the border of the baccalaureate transcript, we proceed with a method based on RLSA algorithm. The idea is to gather the pixels of the frame or the border into a single unit, and to separate them from the document, the task that may be accomplished efficiently using RLSA algorithm. This latter allows to connect black pixels separated by less than n white pixels according to horizontal or vertical direction. In fact, the border takes the form of a rectangle formed of four sides (top, bottom, right, left). Separating the border therefore

requires the localization of its four sides on the document. Thus, the physical study that we did allows us to know the approximate location of the four sides of the border. The latter may differ slightly from one document to another, but they are always located in the first parts (top, bottom, right, and left) of the image with a thickness that never exceeds the value (document width / 10). To find the horizontal sides (top and bottom) of the border, for example, we apply the RLSA algorithm horizontally on the top and bottom parts of the image with a threshold n = (image width * 10%). The localization of the two vertical sides (right and left) is done in the same way but by applying a vertical RLSA on the right and left parts of the image, with a threshold n = (image width * 20%). The values of the threshold n have been chosen by experimentation so that they allow to connect the closest components of the border.

The application of RLSA algorithm on the parts of the image containing the horizontal (or vertical) sides of the frame leads to connect the black pixels of the border that are near along the horizontal (or vertical) direction. At the end the border becomes composed of a single object (Fig. 7). A connected components labeling is then performed in order to gather all the pixels composing the border into a single unit, and the border form the biggest connected component.

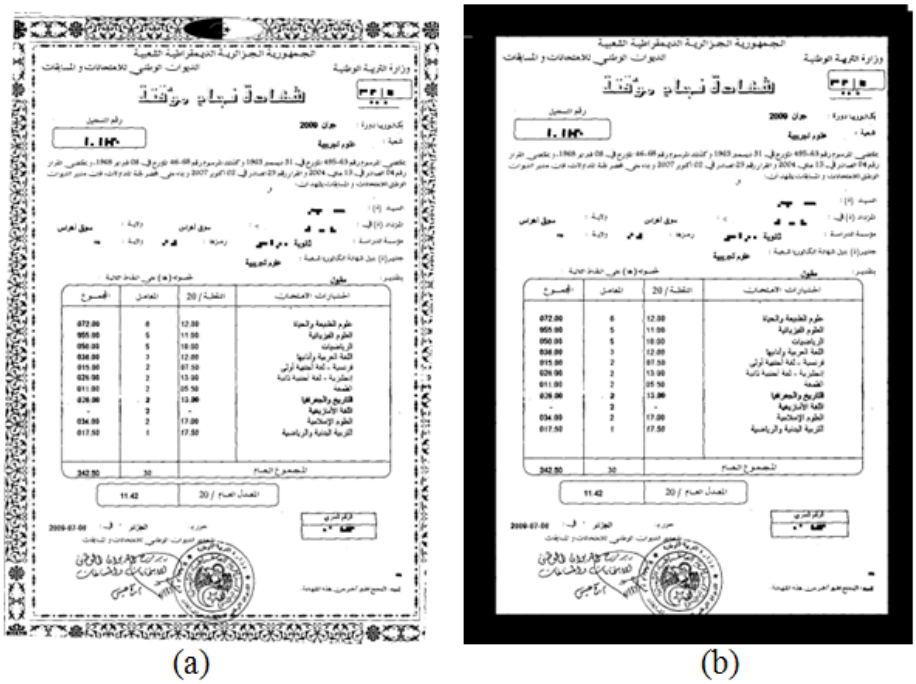
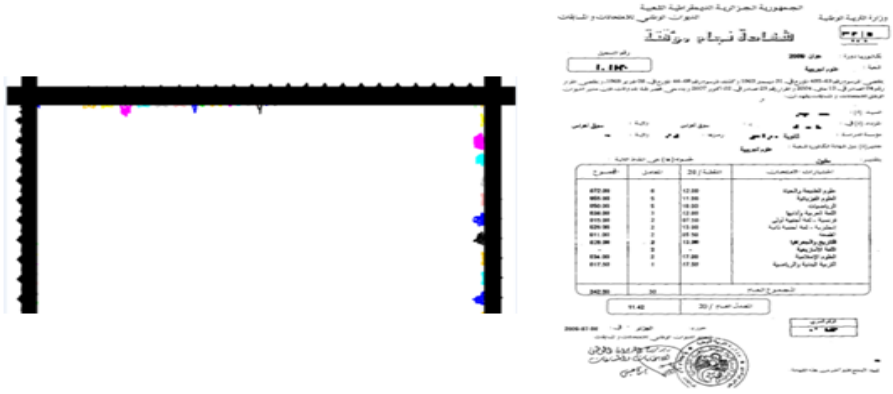


Fig. 7. Border detection, (a) binarized transcript image, (b) border detected using RLSA algorithm.



(a) (b)

Fig. 8. Border refinement and elimination, (a) connected component that do not belong to the border, (b) border removerd.

In fact, RLSA can sometimes connect some connected components close to the border with the components of the border. A refinement is then necessary to remove the pixels which do not belong to the border, from the frame. To do this, we create a new image containing only the border. Then, we redo the connected components labeling on the part of image after the border. The labeled connected components form small portions of the border but it is unclear whether they actually belong to the border or not (see Fig. 8.a). These components are then classified into border components and non-border components according to certain size and dimension criteria. Finally, the frame separated from the document image because its presence may influence the results of the following segmentation steps. The transcript without border is shown in Fig 8.b.

Lines segmentation Once the border is separated from the image, only the relevant elements of the transcript (text, tables, ...) remain in black on a white background. And since the image is well oriented (after the skew correction step), the text lines of the transcript may be extracted by a RLSA smoothing. Thus a horizontal RLSA smoothing is applied to the resulting image of the preceding steps in order to eliminate the spaces between the words close of the same text line. In addition, a vertical RLSA smoothing is applied in order to connect the diacritic marks to their corresponding words (because the transcripts are in Arabic and the diacritics marks are strongly present in Arabic script). Noting that the transcript contains several blocks (heading, title, decree, student information, scores table,...) and that two blocks may overlap horizontally. For example, from Fig. 2, the block "R.N" and the block "Year + Branch" overlap horizontally. Therefore, the horizontal RLSA threshold n must be chosen so that it connects the words of a text line of a block, and at the same time does not

Fig. 9. Horizontal and vertical RLSA for detecting lines and blocks, (a) lines detected using horizontal RLSA , (b) first extraction of blocks using vertical RLSA.

allow to connect the text lines which belong to two horizontally overlapping blocks (n = (image width * 7%)). Fig. 9.a shows the result of this step.

Blocks extraction the blocks can be extracted using RLSA algorithm together with the projection profile analysis. At first, we apply RLSA technique vertically with a predetermined threshold, on the resulting image of the previous RLSA smoothing (Fig 9.b). This second smoothing RLSA aims to connect the text lines of the same textual block or to connect the pixels, vertically close, of a non-textual block. From the remark that the interline distance is smaller than the distance between successive blocks, the smoothing threshold n must be chosen carefully. It must be large enough to allow the connection of the lines of the same block and at the same time must be insufficient to connect the blocks together. n has been chosen equal to image height / 65.

In a second time, from the smoothed image by RLSA, we proceed to the segmentation of the document into blocks based on the analysis of horizontal and vertical projection profiles. The histogram of horizontal projections is first obtained by calculating the number of black pixels in each line of the image resulting from RLSA smoothing. As the document is well aligned, the corresponding histogram of horizontal projections will consist of peaks and valleys, representing the blocks and spaces between them respectively (Fig. 10.a).

The space between two consecutive valleys describes the height of a block (or horizontally overlapped blocks). Then, we proceed to a second analysis of vertical projection profiles on each block resulting from the first analysis. This second analysis is performed to refine the block segmentation, by separating the

(a) (b)

Fig. 10. Blocks extraction based on projection profiles analysis, (a) projection profiles of the smoothed image, the valleys are in red, (b) blocks extracted.

blocks that overlap horizontally and that are considered as a single block during the first analysis. Thus, if the histogram has several peaks and valleys, this indicates that several horizontally adjacent blocks are present and to separate them it is enough to segment in the valleys. The final result of blocks extraction is illustrated in Fig. 10.b.

Textual blocks identification and lines extraction As we said before, the blocks may be textual, formed of several text lines, or non-textual. To identify the textual blocks among all the extracted blocks, we examine each block and test whether it consists of several text lines or not. In fact, we have already extracted the text lines using RLSA during the first step of segmentation, but this first segmentation may not be precise because of the presence of noise pasting two lines together, for example. This is why we propose to carry out a second extraction of text lines to confirm.

However, the presence of text lines in a block may be tested based also on the analysis of projection profiles. Thus, we calculate the histogram of horizontal projections of each block extracted from the binarized image (before segmentation). The presence of several (more than 1) global minima, with few black pixels, in the histogram may show inter-line spaces for a textual block. A text line will be between two successive minima. To ensure that this is really a textual block, we apply in a second time the analysis of vertical projection profile for each text line obtained from the first analysis. The presence of several white valleys in the histogram shows certainly interword spaces. Obviously, the absence of interline

and interword spaces confirms that it is not a textual block. Fig. 11 illustrates the horizontal projection profiles of a textual and non-textual blocks.

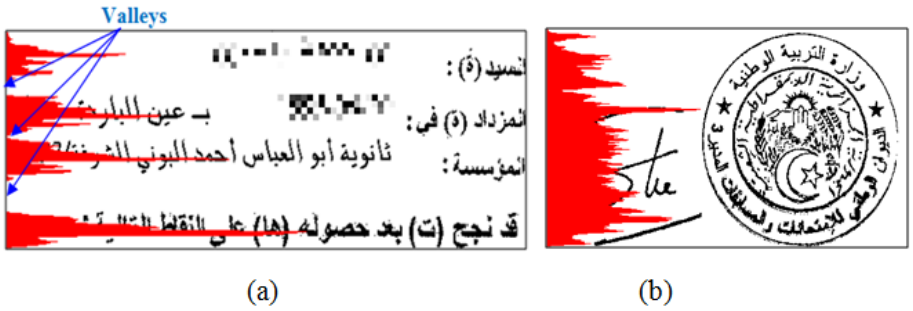


Fig. 11. Horizontal projection profile, (a) of a textual block, (b) of a non-textual block.

Distinction between tables and graphics Once the textual blocks are identified, only non-textual blocks remain. These can be tables or graphics (logos, stamp and signature). To distinguish tables from graphics we rely again on the use of Radon transform. In fact, Radon transform seems a good choice because it allows to detect the presence of straight lines. These are considered as the key element to discriminate the presence of tables. Thus, Radon transform is applied to each non-textual block extracted from the binarized image (before segmentation) in order to check for the presence of straight lines in this block. As we have

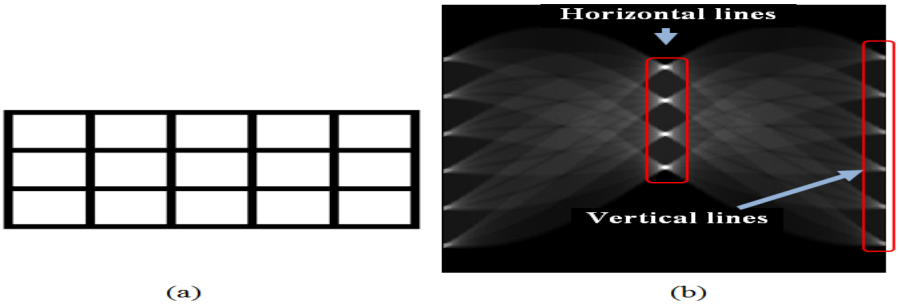
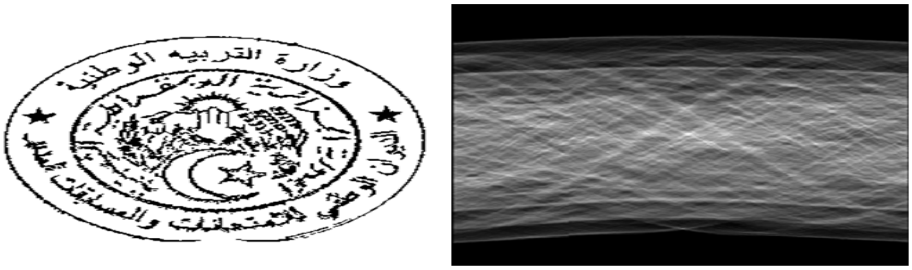


Fig. 12. Detection of a table, (a) table image, (b) its Radon transforme.

already said, Radon transform returns peaks in the form of dots, which signal the presence of straight lines. Fig. 12 shows the result of Radon transform on a table image. The Radon space thus constructed indicates the presence of 10

points of strong luminosity indicating the presence of 10 lines in the table image: 4 points corresponding to the horizontal lines and 6 points corresponding to the vertical lines.

In the case where the non-textual block does not contain straight lines (graphic), Radon transform of this block does not present any peak as a point of strong intensity (Fig. 13).



(a) (b)

Fig. 13. Detection of graphic, (a) stamp image, (b) its Radon transform.

Text extraction from tables the last step in our approach is the extraction of text from the tables identified in the previous step. To detach the text from a table, we proceed to the segmentation of this table, extracted from the binary image, in columns and then in cells, and we also make use of the analysis of projection profiles.

Segmentation into columns The segmentation of a table into columns is done by analyzing the profiles of vertical projections. But before that, it is necessary to remove the tables border. To do this, one has only to perform a connected components labeling, only on the area of the image containing the table. The largest connected component is the border of the table and it will be eliminated. Then, the histogram of vertical projections is built on the area of the table without border. The sequences of white pixels of the histogram correspond certainly to inter-column spaces. The separation of columns is done by segmenting the table in these sequences.

Columns segmentation into cells and extraction of their content The cells are segmented from the columns by analysis of horizontal projection profiles. Thus, on each column obtained, the histogram of horizontal projections is calculated. The peaks of the histogram represent the cells of the column and the valleys correspond to the spaces between the cells. The separation of the cells and the extraction of their contents is done by dividing the column in these valleys. Fig. 14 illustrates the result of this step on the scores table, where the content of cells is colored in yellow.

المجموع	المعامل	النقطة / 20	اختبارات الامتحان
072.00	6	12.00	علوم الطبيعة والحياة
055.00	5	11.00	العلوم الفيزيائية
050.00	5	10.00	الرياضيات
036.00	3	12.00	اللغة العربية وآدابها
015.00	2	07.50	فرنسية - لغة أجنبية أولى
026.00	2	13.00	إنجليزية - لغة أجنبية ثانية
011.00	2	05.50	الفلسفة
026.00	2	13.00	التاريخ والجغرافيا
-	2	-	اللغة الأمازيغية
034.00	2	17.00	العلوم الإسلامية
017.50	1	17.50	التربية البدنية والرياضية
342.50	30		المجموع العام

Fig. 14. Text extraction from the scores table.

5 Experiments and results

In this section we present the experiments conducted in order to evaluate the performance of the proposed system. We first describe the dataset used, the evaluation measures employed, and the results obtained together with a discussion of these results.

5.1 Test dateset

As we said earlier, we are interested in this work to the analysis of Algerian baccalaureate transcripts. Our test corpus used throughout our work is composed of 40 images of baccalaureate transcripts between the years 1997 to 2017. The documents in our test corpus are of different structures and formats making their processing and analysis difficult. Examples of transcripts from our test corpus are shown in Fig. 1.

5.2 Evaluation measures

The performance of our system at the block extraction level is measured in terms of Recall, Precision, and F-Measure. Noting TP, FP, and FN, the number of true positives, false positives, and false negatives respectively.

- A true positive is the result where the system correctly extracts a block.
- A false positive is the result where the system detects a block that does not really exist.

– A false negative is the result where the system fails to detect an existing block.

F-Measure was introduced firstly by Chinchor in [10]. Because of its simplicity, F-measure is considered one of the most used measures for the quantitative evaluation of segmentation. F-measure is given by:

$$FM = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{2}$$

$$with : Recall = \frac{TP}{TP + FN} \tag{3}$$

$$and : Precision = \frac{TP}{TP + FP} \tag{4}$$

5.3 Results and discussion

The block extraction results for the 40 test images as well as the average results are summarized in table 1. From Table 1, we noticed that our system has shown good performance for block extraction in terms of recall, precision, and F-Measure.

Table 1. Blocks extraction results for all images of the test corpus.(T:Transcript, P:Precision, R:Recall)

T	**#blocks**	**TP**	**FP**	**FN**	**R**	**P**	**FM**	**T**	**#blocks**	**TP**	**FP**	**FN**	**R**	**P**	**FM**
1997	11	9	0	2	0,82	1	0,90	2013	13	12	0	1	0,92	1	0,96
1998	11	10	0	1	0,91	1	0,95	2013	13	12	0	1	0,92	1	0,96
2000	12	12	0	0	1,00	1	1,00	2013	13	11	0	2	0,85	1	0,92
2000	12	12	0	0	1,00	1	1,00	2013	13	12	0	1	0,92	1	0,96
2001	12	12	0	0	1,00	1	1,00	2013	13	11	0	2	0,85	1	0,92
2001	13	12	1	0	1,00	0,92	0,96	2014	14	13	0	1	0,93	1	0,96
2001	13	13	0	0	1,00	1	1,00	2014	14	14	0	0	1,00	1	1,00
2002	13	11	0	2	0,85	1	0,92	2014	14	14	0	0	1,00	1	1,00
2005	13	11	0	2	0,85	1	0,92	2014	14	13	0	1	0,93	1	0,96
2007	13	13	0	0	1,00	1	1,00	2014	14	14	0	0	1,00	1	1,00
2008	13	12	0	1	0,92	1	0,96	2014	13	11	0	2	0,85	1	0,92
2009	14	13	0	1	0,93	1	0,96	2014	13	11	0	2	0,85	1	0,92
2010	14	12	0	2	0,86	1	0,92	2015	14	13	0	1	0,93	1	0,96
2011	14	12	0	2	0,86	1	0,92	2015	14	13	0	1	0,93	1	0,96
2012	14	13	0	1	0,93	1	0,96	2015	14	13	0	1	0,93	1	0,96
2013	13	11	0	2	0,85	1	0,92	2015	14	11	0	3	0,79	1	0,88
2013	13	12	0	1	0,92	1	0,96	2015	15	13	0	2	0,87	1	0,93
2013	13	13	0	0	1,00	1	1,00	2015	15	11	0	4	0,73	1	0,85
Average													0,91	1,00	0,95

Thus, the system was able to extract 91% of the existing blocks in the 40 transcripts images used in the tests with a perfect precision of 100%, and thus shows a high compromise between recall and precision (F-Measure = 95%). The analysis of the individual results shows us that 9 transcripts produced a value of F-Measure = 1, i.e. the physical structure of 9 among the 40 transcripts has been fully recognized. In addition, 15 transcripts have a F-Measure value greater than 95%. Finally, the somewhat low value of F-Measure with some transcripts is caused by defects in the original transcripts and not gaps in our system.

6 Conclusion

Technological innovations, including computers and informatics, produce an increasingly complex quantity of documents. This mass of documents has forced people to look for ways to exploit them easily and more effectively. New fields of research, including document analysis and recognition, and electronic archiving were born. Each of these areas integrates a multitude of research tracks whose segmentation or extraction of the physical structure of documents is one of the most important.

In this paper, we proposed a segmentation approach for a particular type of document, namely Algerian baccalaureate transcripts. These are of considerable importance in the student's file. The goal is to develop the core of a system for digitization, analysis, recognition, and retrieval of archives documents within Algerian universities.

The proposed approach may be classed within the category of hybrid methods and it consists of several stages of processing gathered in two modules. A first module incorporating various preprocessings, namely filtering, grayscale transformation, binarization, and skew correction, aimed to improve the quality of the input images and preparing them for the following steps. The second module aims to extract the physical structure of the baccalaureate transcripts by applying several steps. First, the border of the transcript is removed because it does not provide any relevant information. After that, a first lines segmentation is performed based on the RLSA algorithm. The lines are then combined into blocks by applying RLSA again and then by analyzing projection profiles. The textual blocks are next identified and their lines are extracted by alternating horizontal and vertical projection profiles. Non-textual blocks are separated into tables or graphics using Radon transform. Finally the tables are segmented into columns and cells and their contents are extracted.

Several tests have been conducted on a local image dataset to evaluate the performance of the developed system and the results obtained are encouraging. These results show the reliability and robustness of the system developed and confirm the effectiveness of the proposed segmentation approach and the validity of the choices made during the system design.

References

1. M. Agrawal, and D. Doermann, Voronoi++: A dynamic page segmentation approach based on voronoi and docstrum features, Proc. 10th International Conference on Document Analysis and Recognition (ICDAR), pp. 1011-1015, 2009.
2. M. Agrawal, and D. Doermann, Context-aware and content-based dynamic Voronoi page segmentation, Proc. 9th IAPR International Workshop on Document Analysis Systems, pp. 73-80, 2010.
3. O.T. Akindele, and A. Belaid, Page segmentation by segment tracing, Proc. 2nd International Conference on Document Analysis and Recognition (ICDAR), pp. 341-344, 1993.
4. A. Amin, and R. Shiu, Page segmentation and classification utilizing bottom-up approach, International Journal of Image and Graphics, vol. 1, No. 2, pp. 345-361, 2001.
5. A. Antonacopoulos, and R.T. Ritchings, Flexible page segmentation using the background, Proc. 12th IAPR International Conference on Pattern Recognition, vol. 3 - Conference C: Signal Processing (Cat. No. 94CH3440-5), vol. 2, pp. 339-344, 1994.
6. A. Ben Salah, Matrise de la qualit des transcriptions numriques dans les projets de numrisation de masse, doctoral dissertation, Universit de Rouen-France, 2014.
7. R.N. Bracewell, Two-Dimensional Imaging, Englewood Cliffs: Prentice Hall, vol. 247, 1995, pp. 505-537.
8. C. Carton, A. Lemaitre, and B. Coasnon, Automatic and interactive rule inference without ground truth, Proc. 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 696-700, 2015.
9. K. Chen, F. Yin, and C.L. Liu, Hybrid page segmentation with efficient whitespace rectangles extraction and grouping, Proc. 12th International Conference on Document Analysis and Recognition (ICDAR), pp. 958-962, 2013.
10. N. Chinchor, MUC-4 Evaluation Metrics, Proc. 4th Message Understanding Conference, pp. 22-29, 1992.
11. B. Coasnon, DMOS, a generic document recognition method: Application to table structure analysis in a general and in a specific way, International Journal of Document Analysis and Recognition (IJDAR), vol. 8, No. 2-3, pp. 111-122, 2006.
12. P. Courmontagne, Transforme de radon et filtrage : Application la dtection de sillages de mobiles marins, TS. Traitement du signal, vol. 15, No. 4, pp. 297 307, 1998.
13. S. Eskenazi, P. Gomez-Krmer, and J.M. Ogier, A comprehensive survey of mostly textual document segmentation algorithms since 2008, Pattern Recognition, vol. 64, pp. 1-14, 2017.
14. J. Fisher, S. Hinds, and K. DAmato, A Rule-Based System for Document Image Segmentation, Proc. 10th International Conference on Pattern Recognition, pp. 113-122, Atlantic City, USA, 1990.

15. B. Gatos, I. Pratikakis, and S.J. Perantonis, Adaptive degraded document image binarization, Pattern recognition, vol. 39, No. 3, pp. 317-327, 2006.
16. S. Eskenazi, P. Gomez-Krmer, and J.M. Ogier, A comprehensive survey of mostly textual document segmentation algorithms since 2008, Pattern Recognition, vol. 64, pp. 1-14, 2017.
17. C. Faure, and N. Vincent, Simultaneous detection of vertical and horizontal text lines based on perceptual organization, In Document Recognition and Retrieval XVI, vol. 7247, pp. 72470M, International Society for Optics and Photonics, 2009.
18. A. Kefali, T. Sari, H. Bahi, Foreground-Background Separation by Feed-forward Neural Networks in Old Manuscripts, Informatica, vol. 38, No. 4, pp. 329338, 2014.
19. A. Kefali, and S. Drabsia, Localization of scores and average in Algerian baccalaureate transcripts, Proc. International Conference on Signal, Image, Vision and their Applications (SIVA), pp. 1-6, 2018.
20. A. Kefali, A. Obeizi, and C. Ferkous, Segmentation of Algerian baccalaureate transcripts, Proc. 2nd Conference on Informatics and Applied Mathematics, Guelma - Algeria, 2019.
21. K. Kise, A. Sato, and M. Iwata, Segmentation of Page Images Using the Area Voronoi Diagram, Computer Vision and Image Understanding, vol. 70, No. 3, pp. 370-382, 1998.
22. F. Lebourgeois, Z. Bublinski, and H. Emptoz, A Fast and Efficient Method for Extracting Text Paragraphs and Graphics From Unconstrained Documents, Proc. 11th International Conference on Pattern Recognition, pp. 272-276, The Hague, 1992.
23. A. Lemaitre, J. Camillerapp, and B. Couasnon, Contribution of multiresolution description for archive document structure recognition, Proc. 9th International Conference on Document Analysis and Recognition (ICDAR), vol. 1, pp. 247-251, 2007.
24. A. Lemaitre, J. Camillerapp, and B. Coasnon, Multiresolution cooperation makes easier document structure recognition, International Journal of Document Analysis and Recognition (IJDAR), vol. 11, No. 2, pp. 97-109, 2008.
25. G. Louloudis, B. Gatos, I. Pratikakis, and C. Halatsis, Text line and word segmentation of handwritten documents, Pattern Recognition, vol. 42, No. 12, pp. 3169-3183, 2009.
26. S. Mao, A. Rosenfeld, and T. Kanungo, Document structure analysis algorithms: a literature survey, In Document Recognition and Retrieval X, vol. 5010, International Society for Optics and Photonics, 2003 pp. 197-208.
27. G. Nagy, and S. Seth, Hierarchical representation of optically scanned documents, Proc. 7th International Conference on Pattern Recognition (ICPR), pp. 347-349, 1984.
28. G. Nagy, S. Seth, and M. Viswanathan, A prototype document image analysis system for technical journals, Computer, vol. 25, No. 7, pp. 10-22, 1992.
29. L. O'Gorman, The document spectrum for page layout analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, No. 11, pp. 1162-1173, 1993.

30. N. Ouwayed, and A. Belad, A general approach for multi-oriented text line extraction of handwritten documents, International Journal on Document Analysis and Recognition (IJDAR), vol. 15, No. 4, pp. 297-314, 2012.
31. T. Pavlidis, and Z. Jiangying, Page segmentation and classification, CVGIP: Graphical models and image processing, vol. 54, No. 6, pp. 484-496, 1992.
32. R. Sarkar, S. Moulik, N. Das, S. Basu, M. Nasipuri, and M. Kundu, Suppression of non-text components in handwritten document images, Proc. International Conference on Image Information Processing, pp. 1-7, 2011.
33. Z. Shi, and V. Govindaraju, Line separation for complex document images using fuzzy runlength, Proc. International Workshop on Document Image Analysis for Libraries, pp. 2324, 2004.
34. A.L. Spitz, Recognition processing for multilingual documents, Proc. International Conference on Electronic Publishing, Document Manipulation and Typography, pp. 193-205, Gaithe rsburg, Maryland, 1990.
35. N. Stamatopoulos, B. Gatos, and S.J. Perantonis, A method for combining complementary techniques for document image segmentation, Pattern Recognition, vol. 42, No. 12, pp. 3158-3168, 2009.
36. D. Sylwester, and S. Seth, A trainable, single-pass algorithm for column segmentation, Proc. 3rd International Conference on Document Analysis and Recognition, vol. 2, pp. 615-618, 1995.
37. T.A. Tran, I.S. Na, and S.H. Kim, Hybrid page segmentation using multilevel homogeneity structure, Proc. 9th International Conference on Ubiquitous Information Management and Communication, pp. 78, 2015.
38. T.A. Tran, I.S. Na, and S.H. Kim, Page segmentation using minimum homogeneity algorithm and adaptive mathematical morphology, International Journal on Document Analysis and Recognition (IJDAR), vol. 19, No. 3, pp. 191-209, 2016.
39. M. Viswanathan, Analysis of scanned documents - A syntactic approach, In Structured Document Image Analysis, pp. 115-136, Springer, Berlin, Heidelberg, 1992.
40. F. Wahl, K. Wong , and R. Casey, Block Segmentation and Text Extraction in Mixed Text/Image Documents, Computer Vision Graphics, and Image Processing, vol. 20, pp. 375-390, 1982.
41. D. Wang, and S.N. Srihari, Classification of newspaper image blocks using texture analysis, Computer Vision Graphics and Image Processing, vol. 47, No. 3, pp. 327 - 352, 1989.
42. Y. Wang, Y. Zhou, and Z. Tang, Comic frame extraction via line segments combination, Proc. 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 856-860, 2015.
43. A. Yamashita, T. Amano, Y. Hirayama, N. Itoh, S. Katoh, T. Mano, and K. Toyokawa, A document recognition system and its applications, IBM journal of research and development, vol. 40, No. 3, pp. 341-352, 1996.