International Journal of Computing and Digital Systems

Image Extracting from Ancient Arab Documents with Complex Structures

[1]Mohamed Aymen Charrada, [2]Najoua Essoukri Ben Amara

[1,2] *Advanced System in Electrical Engineering Research unit (SAGE), National Engineering School of Sousse (ENISo), University of Sousse, Sousse, Tunisia*

E-mail address: Mohamed_aymen_charrada@yahoo.fr, najoua.benamara@eniso.rnu.tn

Received October 24, 2013 , Revised November 19, 2013 , Accepted December 15, 2013 , Published 1 Jan 2014

Abstract: In this paper, we give an overview on approaches for the graphic extraction, and especially for the image extraction from printed documents. We present also our contribution to this area operating on historical Arab periodicals. The developed method works on monochrome documents and is based on the Gabor filter exploration followed by many post-processing steps. It allows firstly the text/graphic separation then it ensures the distinction between images and other graphic classes (drawings, textured titles). The various tests and experiments were performed on an image database obtained from historical documents with complex structures, collected from various newspapers coming from the National Archives of Tunisia. The obtained results and the comparison carried out with another existing approach show many interesting perspectives and prove that our approach is able to attain high level of acceptability.

Keywords: Segmentation, graphics, images, historical Arab periodicals, Gabor filters

I. INTRODUCTION

In the course of the history, humanity had the need to note and transmit their knowledge considering their limited storage capacities. For this reason, they have exploited a diversity of rigid physical supports for the conservation of information such as the stones, the wood and the papyrus. Nevertheless, the appearance of the paper has triggered a scientific and cultural renaissance because the transmission of knowledge from one period to another and from one civilization to another has become easier and more convenient. Since the appearance of the writing and the paper, the manuscript and the printed document contents have not ceased to grow up and to be enriched supplied by the development of the edition tools and the scientific and intellectual human renaissance. Thus, the structures of the produced documents become increasingly rich and complex in order to adapt to the community needs, to reduce the cost of the document edition and to benefit as well as possible from the physical space offered by the information support (paper format). In this context, the graphic, and particularly the images, have been used increasingly in the documents and especially in the periodicals jointly with the text as a visual tool for the information transmission. Indeed, images are often used as a complement to an article in order to illustrate or to prove an event or an action [1]. In this sense, it is essential in the case of the recognition and the analysis of the document physical structures to carry out the segmentation of these documents in different physical layers of information in order to apply the appropriate analysis methods to each one of these layers. In the case of documents with graphic zones, including periodicals, this step consists in separating between the graphic, and more particularly the images, and the text. However, the recognition of these images, in the case of an old and damaged document, proves to be a complex task because of the different types of deformations and damages that may affect such documents. Moreover, these documents may contain other textured physical entities other than images and text, such as drawings, textured titles, head letter, inking pads, tables..., which can alter or distort the image recognition. An example illustrating the structural richness of the treated old periodic documents is presented in Figure 1.

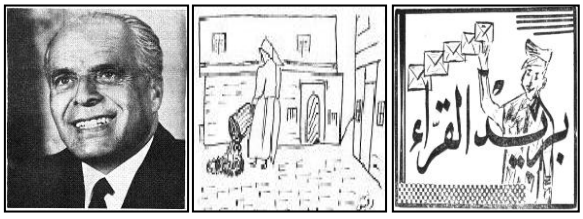
In the next section, we present an overview of the main approaches proposed for the image extraction from printed documents with complex structures, existing in the literature. Then, we describe in Section 3 our proposed approach. Finally, we present in section 4 the recorded experimental results and a comparison of our method with the approach developed by Hadjar and al. [1].

II. OVERVIEW OF IMAGE EXTRACTION

To our knowledge, there are a multitude of works on the text/graphic segmentation for old documents that are referenced in the literature, but few research works are interested in the documents with complex structures, especially Arabic periodicals and which seek to distinguish images from other classes of graphic content (drawing or textured title).

Figure 1. Example of the physical structure variability of the treated documents

Figure 2 shows an example of graphic classes extracted from the treated documents during this work.



(a) Image (b) Drawing (c) textured title

Figure 2. Example of graphic content extracted from the treated documents

Indeed, the first link in the document analysis chain is usually that which aims to segment the document images into different physical information layers in order to apply the appropriate analysis for each of these layers. In our case, this step consists in separating between the graphic zones of the document, and more precisely the images, and the part containing textual information. In fact, several algorithms have been proposed to solve this kind of problem since 1980. These algorithms have been guided by data analysis (bottom-up approaches), by model analysis (top-down approaches) or by data/model alternation (interactive or mixed approaches). Generally, the algorithm of Fletcher and Kasturi [2] is considered as a main reference for this research axis in view of it is known in the literature of the field as the first reliable and generic tool, able to treat a great variability of documents. This algorithm is considered as a fusion approach which consists to analyze and gather gradually the component connectivity. Other approaches based on the data fusion, which consist in combining the elementary objects (pixels, related objects, pixel groups) recursively following a set of block fusion rules, were proposed in literature, such as the Hadjar approach [1] which operates on Arab periodicals, the method proposed in [3] which is part of the "NaviDoMass" project and which uses a decomposition of the image in many layers followed by the application of the "Zipf" transform before carrying out a selection of related components moderated by a set of well-defined rules, and also the approach presented in [4] which consists initially in removing the image edges, making an analysis of the component connectivity and applying finally a spreading operation for the noise suppression, as well as the methods proposed in [5, 6]. Other fusion approaches use the Run Length Smoothing Algorithm (RLSA) such as the approaches proposed in [7, 8] and the method of cumulative gradients [9, 10]. These methods are simple to implement and does not require an image content model, but they require a considerable computational time and they are sensitive to noise. Other approaches are based on the cutting operation by using the recursive XY cut [11], the space analysis [12], the "Split and Merge" approach [13] and the projection profiles [14]. In fact, these methods are extremely fast but they are unreliable in the case of the degraded documents and they are limited to the constrained documents or for which the model is known. There are also many methods guided by a model or a style [15]. These methods can lead to better results on the heterogeneous documents but they require a lot of a priori knowledge. The literature presents yet more recent approaches, which are more robust and less dependent on a priori knowledge, that are mainly based on the characterization of the document texture [16, 17], the multi-scale analysis [18], the edge analysis [19],the grammatical model learning [20], the rules-based techniques [21] or the stochastic methods [22].Consequently, most of the proposed methods either require a priori knowledge related to the nature or the document structure, a high computing time and enormous resources, or they are not suitable for documents with potential defects and damages. In addition, most of the suggested works does not lead to an evaluation of their performances or to a comparison with other approaches. That is mainly explained by the fact that each author achieves a text/graphic separation method for a specified class of documents that may not be appropriate for other document classes.

Therefore, we present in the next section our contribution to the text/graphic separation and the image extraction from old periodicals in Arabic language, based on a comprehensive texture analysis of the document for an initial detection of the graphic zones followed by a local analysis using a classification technique and a neighborhood analysis for improving the results. This approach is applicable without strong a priori knowledge, as it manages in an acceptable way the defects which can affect the archival documents.

III. PROPOSED APPROACH FOR THE IMAGE EXTRACTION

In fact, the proposed approach for the image extraction from old Arabic magazines is essentially based on the use

of the Gabor filters for the characterization and the separation between the different textures. For this reason, we begin with a presentation of these filters.

A. Gabor filter

The Gabor filter is a powerful tool for texture analysis. Indeed, this is a sinusoidal function modulated by a Gaussian envelope acting as a band pass that can be used to extract a specific image frequency band. The mathematical formulation of a 2D Gabor function called "h" in the space field for a fundamental frequency U_0 throughout the X axis (i.e. $\theta = 0°$), is:

$$h(x,y) = \exp\left[-0.5\left(\frac{x^2}{\sigma_x{}^2} + \frac{y^2}{\sigma_y{}^2}\right)\right]\cos(2\pi U_0 x) \quad (1)$$

Where σ_x (respectively σ_y) is the variance of the Gaussian along the X axis (respectively the Y axis).
The filter orientation θ ($\theta \neq 0$) are obtained by carrying out a rotation of the previous equation. This change is obtained by rotating the coordinate axes as following:

$$x_\theta = x\cos\theta + y\sin\theta \quad (2)$$

$$y_\theta = -x\sin\theta + y\cos\theta \quad (3)$$

Suppose that B_r is the frequency bandwidth (in octaves) and B_θ is the angular bandwidth (in degrees) of the Gabor filter, then:

$$\sigma_x = \frac{\sqrt{2}\,(2^{B_r}+1)}{2\pi U_0\,(2^{B_r}-1)} \quad (4)$$

$$\sigma_y = \frac{\sqrt{2}}{2\pi U_0 \tan(B_\theta/2)} \quad (5)$$

In fact, the graphic areas are generally homogeneous and less rich in transitions than the textual zones. Therefore, the graphic areas are characterized by the low frequencies unlike text zones, which are much richer in high frequencies. Based on this observation, the Gabor filter is too sensitive to the text zones for the high frequencies, and it is relatively more sensitive to the graphic areas for the low frequencies, whereas the modification of the orientation makes it possible to cover all the space of the document. It remains to note that the B_r and B_θ parameter values are fixed experimentally in the following way: $B_r = 1$ octave and $B_\theta = 45°$.

By exploiting these observations, we will present in what follows our approach of the image retrieval and extraction from ancient documents with complex structures.

B. Proposed approach

In fact, Figure 3 summarizes the proposed approach for the image extraction from old Arabic newspapers.

According to Figure 3, our extraction image approach begins by applying a bench of Gabor filters to the original image.

1)Application of Gabor filters

We begin our approach by applying a bench of Gabor filters to the original image in order to filter the graphic components. This filter bench implements actually three low frequencies (1, 2 and 4 Hz), each one is coupled with four different orientations (0 °, 45 °, 90 ° and 135 °) which allows us to design a bench with 12 filters: The choice of the orientations is judged by the need to cover the principal orientations of the document content whereas the frequencies are chosen following a set of tests. In fact, the sensitivity of the Gabor filter to the detection of the segments and the discontinuities prevents it to be very effective for the text/graphic segmentation because of the presence of fine elements detectable by the Gabor filter in the graphic areas, and that for the low frequencies. To remedy this insufficiency, we used a filter bench allowing us to define a combination of frequencies and orientations which are exploited to extract the various document components, each instance of frequency and orientation defines a channel, which is used for leading, filtering and bringing out the document elements whose characteristics correspond to these values.

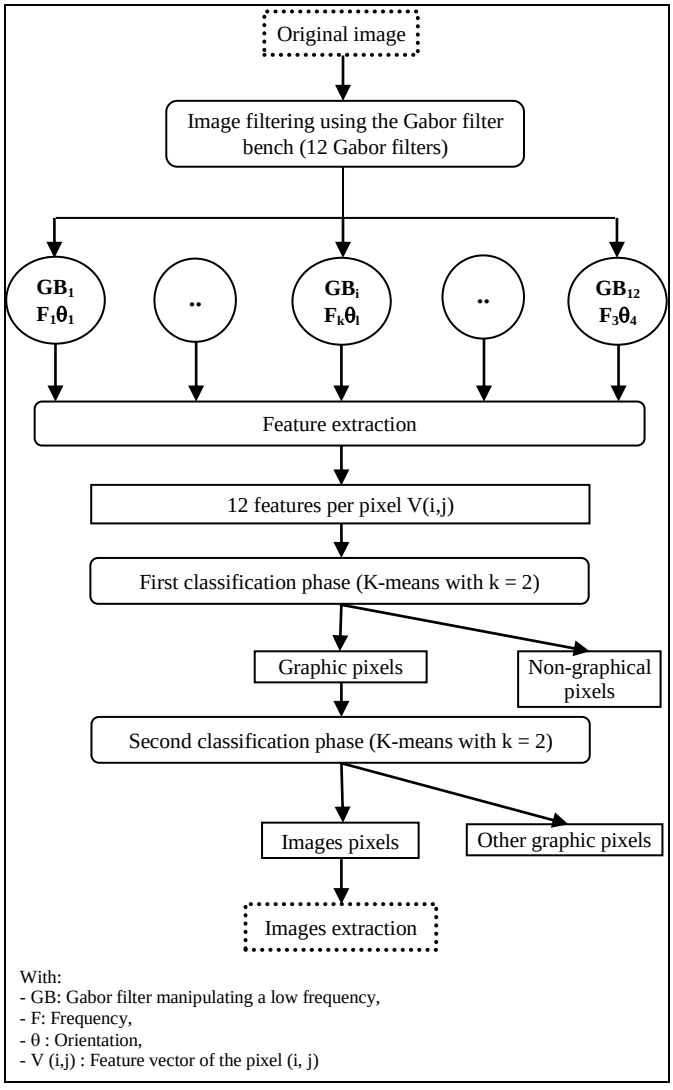


Figure 3. The process schema of our image segmentation approach using a bench of Gabor filters

Consequently, this filter bench provided us twelve different responses for each pixel of the filtered document image; these responses define a feature vector V of twelve components for each pixel. V can be written in the following form:

$$V = \{R(F_i, \theta_j), i = 1..3, j = 1..4\} \quad (6)$$

In order to prove the existence of one or more frequency breakpoints, which allow(s) to distinguish the response of Gabor filters for low and high frequency components, we exploited the grayscale histogram of the filtered image, for a given frequency, on which is applied a thresholding operation by choosing a suitable threshold that will partition the pixels of the filtered document into two classes. Table 1 shows an example of the filtering results thresholding using Gabor filters with different values of frequency and orientation.

In fact, filtering using the Gabor filter bench allows:

- To associate the majority of the graphic pixels to the graphic classes.
- To label some of textual pixels as graphic pixels and to label some pixels belonging to other classes of graphic as image pixels (problem 1).
- To label some image pixels as textual pixels or background pixels (problem 2).

TABLE I. THRESHOLDING OF THE FILTERING RESULTS USING GABORFILTERS FOR DIFFERENT VALUES OF FREQUENCY AND ORIENTATION

Original image 	**F (Hz)**	**1**	**2**	**4**
θ(rad)				
0				
Pi/4				
Pi/2				
3*Pi/4				

We present in the following part the proposed solutions to solve the two problems already mentioned.

2) Filtering result classification

Once the V (i, j)vectors are defined, a classification step is applied in order to determine the membership class of each pixel. In fact, and to solve the problem 1, we carried out a classification of the pixels in order to assign for each one the class to which it belongs: this classification stage must be able in the first place to separate between the textual pixels and those of the graphic as it must separate in a second place between the pixels belonging to the images and the pixels belonging to other types of graphic (drawing, textured title ...). To be able to classify the document pixels in three different classes, the idea was to apply the K-means clustering algorithm with K = 2 to the feature vectors of the different pixels for the first time to separate between the text and the graphic, then a second time for the separation between the pixels belonging to the images and the pixels belonging to other types of graphic. In both steps, the initialization of the cluster centers for K-means is performed using the ElAgha initialization algorithm [23].

Indeed, this step makes it possible to eliminate all textual pixels, but on the other hand it can cause, such as the filtering phase (Problem 2), the loss of some pixels belonging to the images. To solve this problem, we implemented a last post-processing phase that will correct the obtained results of the image extraction.

3) Correction of the image segmentation results

Since all the images present in the documents of our database admit a rectangular form and are surrounded by white zones of pixels and since the Gabor filter allow to detect the segments, the discontinuitiesand the rectilinear zones of transitions of grayscale levels which correspond to its orientation support, and mainly for 0° and 90°, we decided to exploit our approach of the net extraction described in [24, 25] for the generation of the horizontal and vertical gray level transition images. Once these two images are occurred, we generate the connected component image using the resulting image from the previous classification phase. For each found connected component, we look for its pixel of gravity. From this pixel, we launch a search in the four directions i.e. the

horizontal east and the horizontal west of the vertical grey level transition image, the vertical north and the vertical south directions of the vertical grey level transition image, in order to find the four grayscale transitions closest to this pixel of gravity, corresponding to the four edges of our image. Once the edges of the image are found, it is possible to correct the image extraction results by a simple addition of the missing pixels in the resulting image using the original image. Figure 4 shows an example of the correction method of our image extraction results.

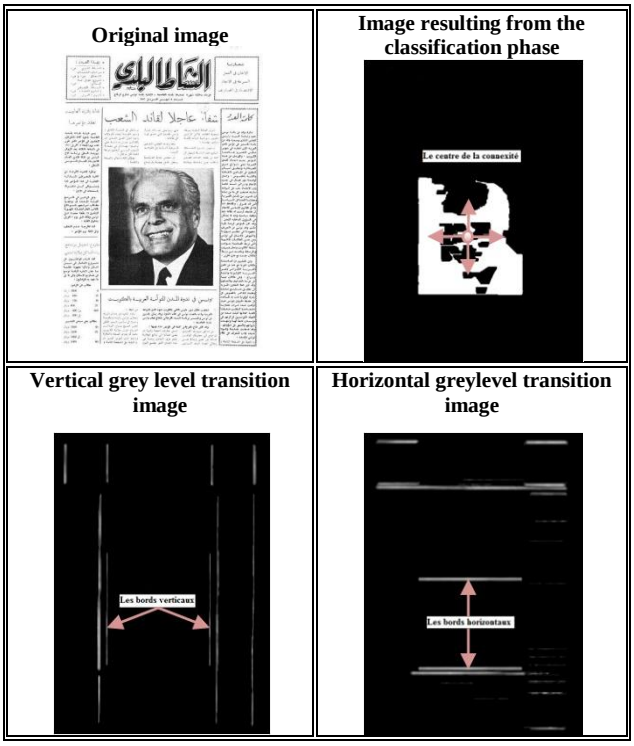


Figure 4. Process for the correction of the image extraction results

IV. EXPERIMENTS AND RESULTS

In this section, we describe first our test database and then we consider the main conducted experiments and recorded results.

A. Database

To evaluate our image extracting approach, we used a test database consisting of newspaper pages dating from 1950 to 1956, derived from two Tunisian Arab periodicals: "Al Nachat el Baladi" and "Al Madina", coming from the National archives of Tunisia (NAT). This database consists of 700 newspaper pages characterized by a great structure variability and content richness. The considered documents are in grayscales with variable sizes, scanned at a resolution of 300 dpi in format "JPEG" and cover the majority of historical document problems [26].

B. Performances of our approach

In order to evaluate our method performances, we retained firstly two discriminating evaluation parameters: the detection rate and the precision rate.

- The detection rate (TD): is the number of the correct detected items divided by the total number of the elements to detect.
- The precision rate (TP): is the number of the correct detected items divided by the total number of the detected items.

Indeed, these parameters were used in the literature to evaluate the performance of several proposed segmentation approaches [27, 28]. Our approach is tested on a set of 200 documents resulting from our database and containing in the real case 537 images to be extracted (296 images in "Al Nachat el Baladi" periodical documents and 241 images in "Al Madina" periodical documents). These images have variable sizes, whose dimensions vary between 32x18 mm for the small estimage and 154x121 mm for the largest image. The evaluation results are given in Table 2.

TABLE II. DETECTION AND PRECISION PERFORMANCES RECORDED ON OUR DATABASE

Periodical / Performances	Al Nachat el Baladi	Al Madina
TD(%)	92.22	93.77
TP(%)	94.46	94.56
Number of the detected items	289	239
Number of the correct detected images	273	226
Number of the false detected images	16	13
Number of the undetected images	23	15

With:

- Number of the detected items: it is the number of the total items detected by our algorithm and considered as images (which can be images or others).
- Number of the correct detected images: represent the number of the detected objects which really correspond to images.
- Number of the false detected images: corresponds to the number of the detected objects which are considered as images but which are not really images.
- Number of the undetected images: it is the number of the images which are not detected by our algorithm.

Figure 5 shows, on documents of our test database, extracting image results using our approach.

Figure 5. Examples of the image extraction results obtained using our approach

According to Figure 5 and Table 2, the results of image extraction recorded by our approach remain very acceptable and satisfactory. The correct detected images are generally the images which present a moderate distribution of the grayscale levels and significant levels of these transitions at the image edges, which characterize most of the images present in our database documents. Similarly, the results obtained for "Al Madina" periodical documents are slightly better than those obtained for "el Nachat Al Baladi" periodical documents because "Al Madina" documents are less damaged and contain images with qualities and sizes higher than those of the images present in "Al Nachat el Baladi" documents.

However, we found that the detection and the precision errors are mainly due to the presence of the strongly textured titles and drawings that can be treated as images and the presence of the large grayscale transition zones inside the images to be extracted which can distort the edge detection phase of our approach. Similarly, the false detected images are due primarily to the presence of the drawings, the graphs and the strongly textured titles containing lot of black pixels and having rectangular forms. Indeed, the presence of important percentages of black pixels in these items makes them candidate images. Thus, the rectangular form and the existence of enormous transitions of grayscale levels at the edges of these items cause their classification as images. In addition, the undetected images are generally images with small sizes containing important zones of white pixels especially in the surroundings of the image edges. In this case, these images are hardly detected by our algorithm and will generally be considered as drawings.

We have also tried to apply our approach to other types of documents such as documents with simple structures, graphical documents, books and composite documents containing images. This phase allows us to consolidate the performance of our approach and to expand its field of application. Figure 6 illustrates an example of the results obtained using these documents.

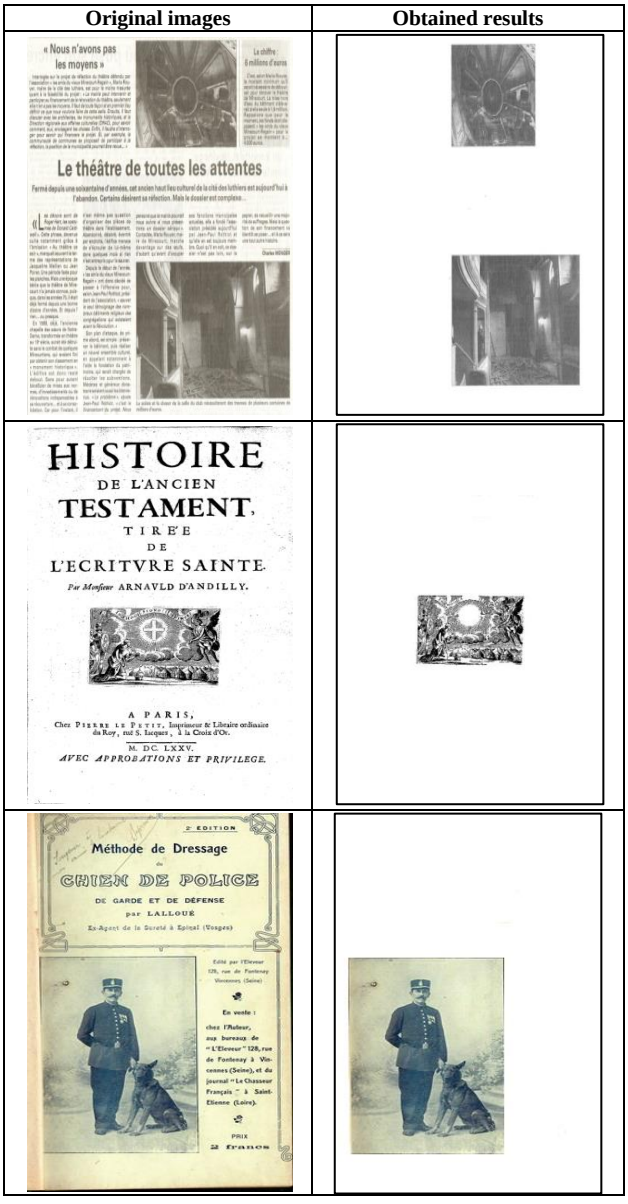


Figure 6. Results of our approach using other types of documents

The results obtained on the different types of selected documents are satisfactory. However, we have noticed some detection errors primarily for images that do not have rectangular forms, which disturbs the detection of

image edges (rectilinear areas with important grayscale transitions) using Gabor filters. Similarly, our approach has difficulties in detecting images with large areas of white pixels at their edges or at their centers which also disrupts the characterization phase (based on the use of Gabor filters) and the classification phase of our approach.

For a more objective evaluation of our approach, we compared it to that proposed by Hadjar and al. [1]. In fact, it is a bottom-up approach based on the analysis and the classification of connected components based on a set of rules and thresholds determined experimentally, and operates on Arabic periodical documents. This comparison is performed on the set of 200 documents (used in the first stage of experimentation) and the obtained results are shown in Table 3.

TABLE III. RESULTS OF THE COMPARISON BETWEEN OUR APPROACH AND HADJAR & AL. APPROACH

Approach / Performances	Our approche		Hadjar and al. Approach	
	Al Nachat el Baladi	Al Madina	Al Nachat el Baladi	Al Madina
TD(%)	92.22	93.77	83.44	84.64
TP(%)	94.46	94.56	86.36	85.71
Number of the detected items	289	239	286	238
Number of the correct detected images	273	226	247	204
Number of the false detected images	16	13	39	34
Number of the undetected images	23	15	49	37

Figure 7illustrates results comparison between our approach and the Hadjarand al. approach

The results illustrated in Table 3 and Figure 7 show that our approach is generally more suited to the nature of the used documents. Indeed, our approach takes into account the images with large areas of white pixels as it differentiates easily between the various graphic classes present in the documents of our database. In addition, we noted that our method is less sensitive to the degradation of the image quality, to the noise and more suitable for the detection of tainted or partially erased images. In fact, we noticed that the number of items detected by the two approaches are nearly equal. Nevertheless, it is clear that there is a big difference between these two approaches concerning the number of correct detected images and the number of undetected images. On the one hand, the approach of Hajarand al. does not distinguish properly between images and drawings, textured titles or other types of graphics. So, they are often detected as images, which increased the number of false detected images. On the other hand, the approach Hadjarand al. is based on the connected components analysis and the use of a set of statistical parameters such as length/width ratio and densities of black and white pixels to identify images, which involved the difficulties of detecting images with small dimensions, whose length/width ratio is too large or too small or whose white pixels density is important.

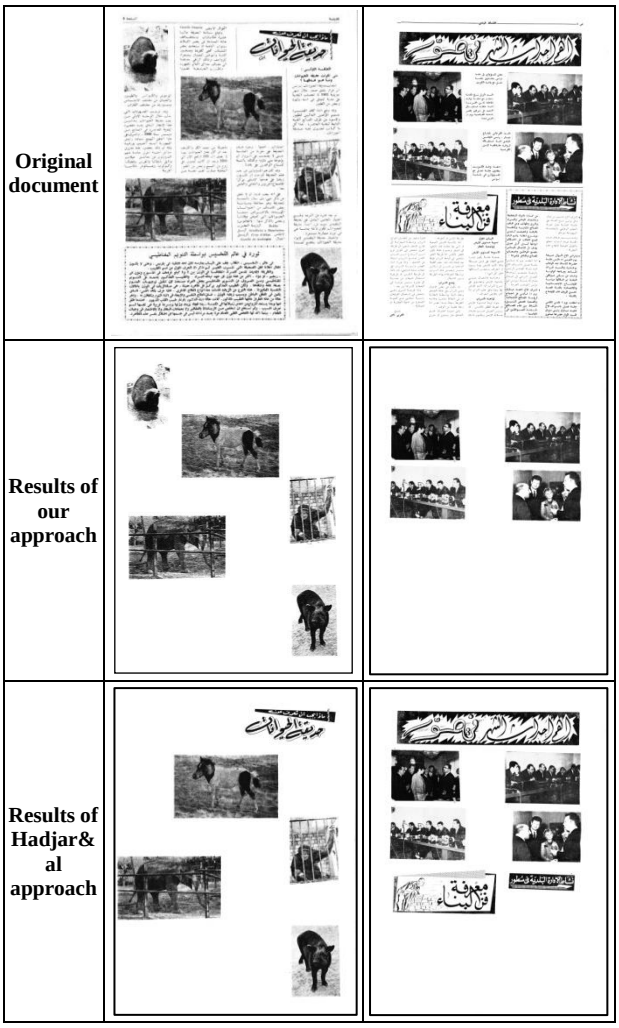


Figure 7. Comparison of image detection results between our approach and Hadjar & al. approach

V. CONCLUSIONS AND PROSPECTS

In this paper, we have been interested in old Arab periodical document structure recognition. We have presented an approach for the image detection and extraction from grayscales documents based primarily on Gabor filter exploration followed by post-processing phases adapted to the different problems present in the considered documents. All performed tests and comparisons show the effectiveness of the suggested approach. However, this approach has some problems which are mainly related to the difficulty of the Gabor filter parameterization and the high executing time compared to other tools used in the same context as the wavelets, the co-occurrence matrices and the directional gradients. On the other hand, our approach is based on some a priori knowledge about the images to be extracted and which can influence greatly the obtained results. In this context, many interesting perspectives are considered, especially in terms of historical Arab periodical

segmentation and indexing. In fact, this approach may be the kernel of a system for the segmentation and the characterization of the different physical entities present in the old Arabic periodicals.

REFERENCES

[1] K. Hadjar, "A model scalability study for Arabic document recognition in an interactive context", PhD thesis, Friburg University, Friburg- Switzerland, pp. 50-69, June 2006.

[2] L.A. Fletcher and R.Katsuri, "A Robust algorithm for text string separation from mixed text/graphics images", IEEE PAMI, vol. 10, n°6, pp. 910-918, 1988.

[3] M. Coustaty, S. Dubois, J.M. Ogier, and M. Menard, "Information Extraction from Old Images of Documents for Indexing", Proc. of the Eight IAPR International Workshop on Graphics Recognition, la Rochelle, France, pp. 303-307, July 22-23, 2009.

[4] S. Ahmed, M. Weber, M. Liwicki, A. Dengel, "Text/Graphics Segmentation in Architectural Floor Plans", Proc. of the ninth Int. Conf. on Document Analysis (ICDAR'2011), Beijing, China, pp. 734-738, September 18-21, 2011.

[5] F. Liu, Y. Luo, M. Yoshikawaf and D. Hut., "A New Component based Algorithm for Newspaper Layout Analysis", Proc. of the Sixth Int. Conf. on Document Analysis and Recognition (ICDAR'2001), Seattle- WA- USA, pp. 167-175, September 10-13 2001.

[6] R. Raveaux, J.C. Burie, and J.M. Ogier, "A colour text/graphics separation based on a graph representation", Proc. of the 19th Int. Conf. on Pattern Recognition (ICPR), Washington, DC, USA, IEEE Computer Society, p. 1-4, 2008.

[7] F. Wahl, K. Wong and G. Casey, "Block segmentation and text extraction in mixed text/image documents", Computer graphics and image processing, n° 20, pp. 375-390, 1982.

[8] N. Normand and C. Viard-Gaudin, "A background based adaptation page segmntation algorithm", Proc. of the third ICDAR, Montréal, vol.1, pp. 138-141, 1995.

[9] F. Le Bourgeois and H. Emptoz, "Document analysis in Grey Level and Typography extraction using character pattern redundancies", Proc. of the fifth ICDAR, Bangalore, pp. 177-180, 1999.

[10] C. Wolf, J.M. Jolion and F. Chassaing, "Text localisation, enhancement and binarisation in printed documents", Proc. of the third ICPR, Québec, vol.4, pp. 1037-1040, august 2002.

[11] B. Gatos, S. L. Mantzaris, K. V. Chandrinos, A. T. Sigris and S. J. Perantonis, "Integrated Algorithms for Newspaper Page Decomposition and Article Tracking", proc of the Seventh Int. Conf. on Document Analysis and Recognition (ICDAR 2003), Edinburgh-Scotland, pp. 381-386, August3-6 2003.

[12] S. Bres, J.M. Jolion and F. Le Bourgeois, "Treatement and analysis of digital images", Hermès, Paris, 2003.

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

[14] S. Khedekar, V. Ramanaprasad, S Setlur and V Govindaraju, "Text - Image Separation in Devanagari Documents", Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003), Edinburgh-Scotland, pp. 453-458, August3-6 2003.

[15] T. Kanungo and S. Mao, "Stochastic Language Models for style-directed layout analysis of document images", IEEE trans on Image processing, vol. 12, n°5, 2003.

[16] A.F. Mollah, S. Basu, and M. Nasipuri, "Text/Graphics Separation and Skew Correction of Text Regions of Business Card Images for Mobile Devices", Journal Of Computing, Volume 2, Issue 2, February 2010, ISSN 2151-9617, pp. 123-133.

[17] S. Audithan and R.M. Chandrasekaran, "Document Text Extraction from Document Images Using Haar Discrete Wavelet Transform", European Journal of Scientific Research, ISSN 1450-216X Vol.36 No.4, 2009, pp.502-512.

[18] W. Boussellaa, A. Zahour, B. Taconet, A. Alimi, "Text/graphic segmentation : Application to Arab ancient manuscripts ", Proc. of CIFED 06,Fribourg (Suisse), pp.139-144, September 18-21,2006.

[19] Q. Yuan and C.L. Tan, "Text extraction from gray scale document images using edge information," Proc. Sixth Int. Conf. on Document Analysis and Recognition (ICDAR 2001), Seattle, WA, USA, pp. 302-306, September 10-13 2001.

[20] O.T. Akindele and A. Belaid. "Construction of Generic Models of Document Structures using Inference of Tree Grammars", Proc of the 3rd Int. Conf. on Document Analysis and Recognition, Montreal, Canada, 1995, pp. 206- 209.

[21] D. Malerba, F. Esposito and O. Altamura. "Correcting the Document Layout: A Machine Learning Approach", Proc of the 7th Int. Conf. on Document Analysis and Recognition, Edinburgh, Scotland, August 2003, pp. 97-102.

[22] R. Brugger, A. Zramdini and R. Ingold. "Modeling Documents for Structure Recognition Using Generalized n-grams", Proc of the 4th Int. Conf. on Document Analysis and Recognition, Ulm, Germany, August 1997, pp. 56-60.

[23] M. El Agha, W.M. Achour,"Efficient and Fast Initialization Algorithm for K-means Clustering", I.J. Intelligent Systems and Applications, DOI: 10.5815/ijisa.2012.01.03, Junary 21-31, 2012.

[24] M.A. Charrada, A. Gardallou and N. Essoukri Ben Amara, "Gabor filters exploration for old arabic periodic segmentation", Proc of the seventh Int. French Workshops on Handwritten and document (CIFED'12), Bordeaux-France, pp.431-443, March 21-23 2012.

[25] M.A. Charrada, N. Essoukri Ben Amara, "Texture approach for nets extraction Application to old Arab newspapers images structuring", Proc for the third Int. Conf. on Image Processing Theory, Tools and Applications (IPTA'12), Istanbul-Turkey, pp.87-91, October 2012.

[26] M.A. Charrada, N. Essoukri Ben Amara, "Development of a database with ground truth for old documents analysis",Proc of the 10th International Multi-Conference on Systems, Signals and Devices (SSD'2013), March 18 - 21, 2013, Hammamet, Tunisia.

[27] B. Gatos, S.L. Mantzaris, K.V. Chandrinos, A.T. Sigris, S.J. Perantonis, "Integrated Algorithms for Newspaper Page decomposition and Article Tracking", Proc of ICDAR 2003, Edinburgh, Scotland, 3-6 Aout 2003, pp. 55-64.

[28] K.M. Summers, "Automatic Discovery of Logical Document Structure", PhD Thesis, Cornell University, pp. 210-221, 1998.

Mohamed Aymen Charrada received the computer engineering degree from the University of Sousse, Tunisia, in July 2008 and the master degree from the National Engineering School of Sousse, Tunisia, in September 2010. He is currently pursuing the Ph.D. degree in image processing at the University of Sfax.

Najoua Essoukri Ben Amar is a Professor with the department of industrial electronics of the National Engineering School of Sousse, Tunisia. She is currently serving as the director of the research unit "Advanced Systems in Electrical Engineering". She has broad researchinterests within the general areas of multimode Biometrics, pattern recognition, treatment of patrimonial documents.