Tifinagh Document Segmentation based on Texture Attributes

Yassine Chajri
Sultan Moulay Slimane University
Beni Mellal
Morocco

ABSTRACT
Text-graphic segmentation is a crucial step in document analysis pipeline, particularly for documents that contain a combination of textual content and graphical elements. This paper presents an approach for effectively segmenting text and graphic components in Tifinagh (alphabet used to write Amazigh language) documents. The proposed method consists of using the texture attribute to effectively detect and extract textual areas and graphical objects. Precisely, it is based on the design of two Gabor filter banks, where the first is configured with high frequencies to identify texts, and the second is designed to detect graphical components using low frequencies.

General Terms
Pattern Recognition

Keywords
Tifinagh-Segmentation-Gabor filter- K-means

1. INTRODUCTION
Writing is a means of communication that represents language through the inscription of signs on various media. Writing first appeared over 5000 years ago in different parts of the globe, including Mesopotamia, Egypt, China, and the Americas. The earliest traces of writing in temples in Iraq date back to 3500 years BCE. In Egypt as well, we find writings dated to around 3300 years BCE. The emergence of writing was driven by the needs associated with the development of civilizations and also by requirements such as the preservation of laws, the exchange and transmission of information, and the keeping of financial accounts, among others. In the beginning, writings were made on clay tablets before paper invention. Handwriting was the only solution for communication before printing press revolution. It was from this time that writing entered a new era, the era of typewritten or printed writing, which has remained dominant to this day.

Technological progress meets all needs and enables the exchange and transmission of messages and information across thousands of kilometers in seconds. This progress has contributed to highlighting new areas and research axes. The field of document analysis and recognition is one of those areas that is currently at the center of researchers' attention. It encompasses all techniques that enable the identification of text, graphics, notation, symbols and also the information extraction from all these elements.

Despite this growing interest, there are some languages whose documents have not received sufficient attention. For example, the Amazigh language, spoken by millions in North Africa and the Sahel, has gained significant status in some countries like Morocco (Amazigh language is recognized by Moroccan constitution, effective integration of Amazigh language in public policies, etc.). However, the Analysis of Amazigh documents and recognition have not reached the level of attention it should receive.

This paper's subject fits into the broader context of the Amazigh language revitalization. More precisely, we present the process of text/graphic segmentation of Tifinagh document based on the texture attribute. Texture-based segmentation involves segmenting documents into microstructures with similar texture characteristics (text and graphic). This choice is justified by two reasons:

- Firstly, the text in a document can be perceived as a texture, whereas graphical objects have a different texture.
- Secondly, the texture approach allows for extracting a set of information without any required knowledge of the context, semantics, or physical characteristics of the studied image.

This process consists in applying a bank of Gabor filters to characterize the textures in Tifinagh document in order to separate them using K-means clustering algorithm. This paper is structured as follows: the first part presents a review existing literature on document segmentation techniques. The second part explains the proposed methodology for texture-based segmentation (detail the process of applying Gabor filters to characterize textures, filters design, features extraction, etc.). The third part shows the results obtained by texture-based segmentation process and discusses the effectiveness and accuracy of the segmentation technique.

2. RELATED WORKS
Document segmentation is a crucial and determining step in all document recognition systems. It involves segmenting the document image into homogeneous blocks (text, image, background, etc.)."

The literature is rich with approaches and techniques for document analysis and segmentation which have been classified according to several criteria. Among these classifications, we find:

Layout analysis: This involves analyzing the overall layout of the document to distinguish between text and graphics based on their spatial arrangement [1].

Color-Based Segmentation: Text and graphics may have different color distributions. Segmenting based on color features allows for the separation of text from graphics [2].

Edge Detection: Text and graphics usually have different edge characteristics. Edge detection algorithms, such as Canny or Sobel are applied to locate edges and separate text and graphics based on edge information [3] [4].

Machine Learning Approaches: Supervised or unsupervised machine learning techniques, such as Support Vector Machines

(SVM), Random Forests, or clustering algorithms are used to classify text and graphics based on features extracted from images. [5] [6]

Semantic Segmentation: Utilizing semantic information, such as the meaning or context of text and graphics, can aid in their segmentation [7] [8].

Hybrid Approaches: Combining multiple segmentation techniques, such as texture analysis with color-based segmentation or edge detection, improves segmentation accuracy.

Other authors have classified these methods into two main categories (classical and texture approaches):

Classical approaches: These are based on prior knowledge and image analysis (bottom-up methods, top-down methods, and mixed or hybrid methods).

- Bottom-Up segmentation: These methods are based on an analysis that starts with low-level components (pixels) and aims to merge them to construct the logical structure of the document (Voronoï diagam, Docstrum). [9-10-11-12]
- Top-Down segmentation: The analysis starts at the global level to refine the regions. In other words, the analysis proceeds from the highest level to the lowest level (Run length smearing algorithm, X-Y Cut algorithm). [13-14-15-16-17-18-19]

Texture-based approaches: These methods allow segmenting document images into areas with the same texture characteristics and without any prior knowledge of the context, physical characteristics, or semantics of the image. In this context, we introduce some works that we find interesting:

- Lin et al.[20] presented a method for segmenting documents into three blocks (text, graphic, and space) based on the Gray-Level Co-occurrence Matrix (GLCM). For feature extraction, the authors used five Haralick properties (energy, entropy, sum of entropies, difference of entropies, and deviation). The document image is segmented into blocks, each characterized by the five Haralick features.
- Etemad et al. [21] proposed a method based on the wavelet transform. For pixel classification, they proposed a fuzzy classification system.
- Eglin [22] proposed a technique based on the analysis of compactness, entropy, and autocorrelation
- Nicolas et al. [23] proposed an approach based on Markov fields.
- Etemad et al. [24] proposed a method based on multi-resolution analysis and wavelet trees for region classification.
- Journet et al. [25] presented a technique for ancient documents segmentation. They extracted texture features for each pixel based on the compass rose, autocorrelation function, wavelet transform, and Fourier transform. They chose to calculate these features at four different resolutions.
- Wang et al. [26] introduced an approach based on the observation of black/white pixel sequences for graphic region detection.

3. METHODOLOGY

In this work, we have chosen to exploit the texture attribute in order to segment Tifinagh documents into homogeneous microstructures (text and graphics). Precisely, this approach is based on the principle of defining two banks of Gabor filters: the first one is designed for detecting textual areas by applying high frequencies, while the second is designed for determining graphic areas with low frequencies. Once the features of each pixel are extracted, the K-means algorithm is used to determine the optimal cluster assignments (text or graphic).

3.1 SYSTEM ARCHITECTURE

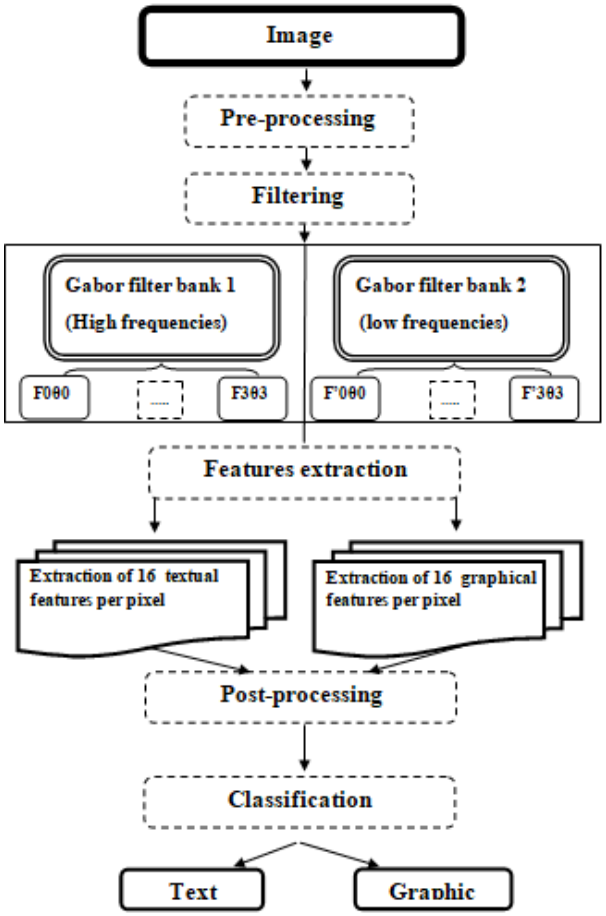

Fig 1: The main steps of the proposed approach

3.2 PRE-PROCESSING

The improvement of image quality is a crucial step in all document recognition systems because it helps overcome issues related to document quality (paper quality, humidity stains, etc.) and problems related to the document scanning process (image acquisition, scanner quality, skew, brightness, noise, etc.).

To ensure the success of Tifinagh document segmentation process, these techniques are essential [27]:

- Median filter for image noise removal
- Radon transform for skew detection and correction
- Normalization

3.3 GABOR FILTER

Gabor filters have attracted considerable attention of researchers because it is inspired by the human visual processes, which decomposes the image into a significant number of filtered images. Furthermore, these filters possess optimal localization properties in both spatial and frequency domains. A Gabor filter can be considered as a sinusoidal

function of specific frequency and orientation, modulated by a Gaussian wave.

In the case of the two-dimensional spatial domain, the Gabor filter can be written as follows:

$$h(x,y) = exp(-\frac{1}{2}(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}))cos(2\pi u_0 x)$$

Where:

- σ_x : Standard deviation of the Gaussian envelope along x direction.
- σ_y : Standard deviation of the Gaussian envelope along y direction.
- u_0 : Frequency of the sinusoidal plane wave along x-direction.
- A rotation of the x - y plane by angle θ will result in Gabor filters at orientation θ.

3.4 FILTERS DESIGN

The main idea of this approach is to design a filtering process that is particularly selective in terms of frequency and orientation in order to effectively characterize various textures.

To achieve this, we designed two filter banks. The first one uses four high frequencies ($16\sqrt{2}$, $32\sqrt{2}$, $64\sqrt{2}$, $128\sqrt{2}$) aimed at detecting textual areas. Conversely, the second one relies on four low frequencies ($1\sqrt{2}$, $2\sqrt{2}$, $4\sqrt{2}$, $8\sqrt{2}$) to determine graphic areas. Regarding the second parameter, we opted to use four orientations 0°, 45°, 90° and 135°.

3.5 FEATURES EXTRACTION

After designing the two filter banks, the Gabor features are calculated from each Gabor response matrix. For a grey-scale image $I(x,y)$ and Gabor filter $G_{u,v}(x,y)$ defined by its centre frequency f_u and orientation θ_v, we calculate their convolution $C_{u,v}(x,y)$ and the filter magnitude response $M_{u,v}(x,y)$ as follows:

$$C_{u,v}(x,y) = I(x,y) * G_{u,v}(x,y)$$

$$M_{u,v}(x,y) = \sqrt{E_{u,v}(x,y)^2 + O_{u,v}(x,y)^2}$$

Where:

$$E_{u,v}(x,y) = Re[C_{u,v}(x,y)]$$

$$O_{u,v}(x,y) = Im[C_{u,v}(x,y)]$$

This process allows extracting a significant number of features. This led us to reduce the high-frequency components as well as apply principal component analysis in order to obtain an intensity value for each pixel of the image.

Finally, based on the obtained features, the K-means algorithm is applied to group similar pixels in order to obtain microstructures with the same texture characteristics.

4. RESULTS

To provide an objective evaluation of this approach, we have prepared a set of images that include Tifinagh documents containing texts and graphic parts. Each image was processed through this system, with the orientation parameter θ being adjusted each time to determine the values that would allow for a more efficient segmentation of the document.

The values of orientation parameter had a significant impact on the accuracy of the text-graphic segmentation, as shown in the following images:

Fig 2: Text-graphic segmentation with two orientations

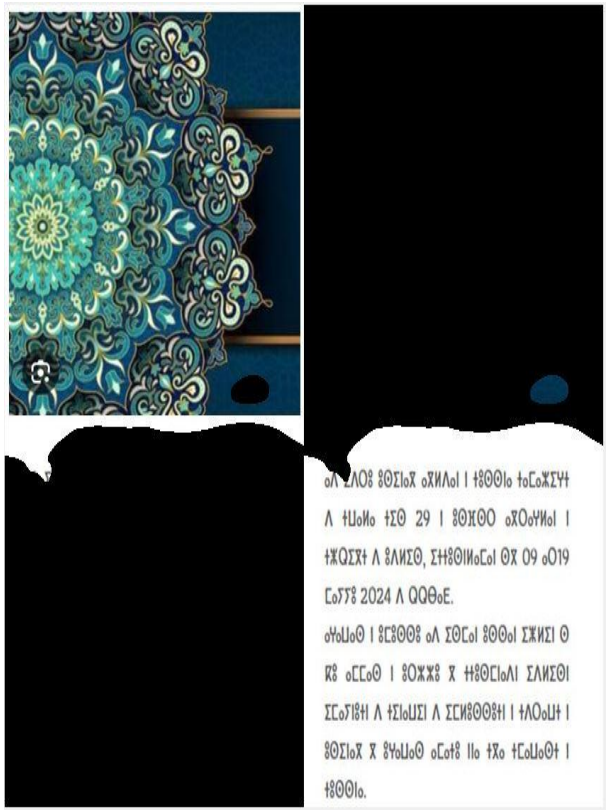

Fig 3: Text-graphic segmentation with four orientations

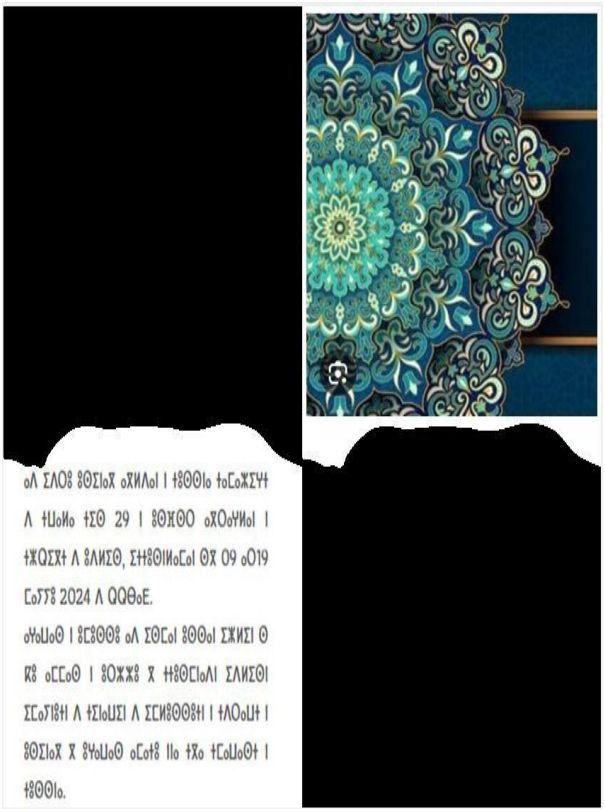

Fig 4: Text-graphic segmentation with eight orientations

The table below summarizes the results obtained by applying the proposed approach to our database of Tifinagh document images. These results are presented according to the number of orientations used in the design of the two filter banks.

Table 1. Text-Graphic Segmentation rate of Tifinagh Documents

Text-Graphic segmentation rate (%)		
Two orientations	Four orientations	Eight orientations
87.5	94.4	97.7

These results show that the text-graphic segmentation rate becomes more accurate as the number of orientations increases. But this also makes the processing time longer.

5. CONCLUSION

Text-graphic segmentation of documents is a process aimed at dividing a document into several distinct regions by separating textual elements from graphic elements. This segmentation is crucial in the context of document recognition, indexing, and information extraction. It allows for better organization of the information contained in a document for more efficient analysis or indexing. Generally, it allows documents structuring, preparation for OCR (Optical Character Recognition) and facilitating information retrieval.

In this work, this segmentation process has been applied to Tifinagh documents, which have not yet received the importance they deserve, as these are documents related to a language spoken in several countries rich in significant scientific and cultural heritage.

This process is based on an approach that involves defining two Gabor filter banks: the first is intended for the detection of textual areas using high frequencies, while the second is dedicated to the localization of graphical areas using low frequencies. Regarding the orientation parameter, we opted to use four orientations based on the results of the experiments conducted (segmentation rate and processing time).

6. REFERENCES

[1] Diem, M., Kleber, F., and Sablatnig, R. 2011. Text Classification and Document Layout Analysis of Paper Fragments

[2] Garcia, C. and Apostolidis, X. 2000. Text Detection and Segmentation in Complex Color Images.

[3] Recio, K. R. O., and Mendoza, R. G. 2019. Three-step Approach to Edge Detection of Texts.

[4] Zhou, W., Du, X., and Wang, S. 2021. Techniques for Image Segmentation based on Edge Detection. *2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*, Fuzhou, China, pp. 400-403, doi: 10.1109/CEI52496.2021.9574569.

[5] Haji, M. M., and Katebi, S. D. 2006. Machine Learning Approaches to Text Segmentation. Scientia Iranica, Vol. 13, No. 4, pp 395-403.

[6] Maia, A. L. L. M., Julca-Aguilar, F. D., and Hirata, N. S. T. 2018. A Machine Learning Approach for Graph-Based Page Segmentation. *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, Parana, Brazil, pp. 424-431, doi: 10.1109/SIBGRAPI.2018.00061.

[7] Rudresh, L. H. N. S., Otageri, S. D. S. M., and Hedge, S. S. 2018. Image understanding: Semantic Segmentation of Graphics and Text using Faster-RCNN. *2018 International Conference on Networking, Embedded and Wireless Systems (ICNEWS)*, Bangalore, India, pp. 1-6, doi: 10.1109/ICNEWS.2018.8903963.

[8] Chowdhury, S., Mandal, S., Das, A., and Chanda, B. 2007. Segmentation of Text and Graphics from Document Images. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, Curitiba, Brazil, pp. 619-623, doi: 10.1109/ICDAR.2007.4376989.

[9] O'Gorman, L. 1993. The Document Spectrum For Page Layout Analysis. IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 15, No. 11.

[10] Kise, K., Sato, A., and Matsumoto, K.. 1997. Document image segmentation as selection of voronoi edges. Proceedings of the 1997 Workshop on Document Image Analysis.

[11] Kise, K., Sato, A., and Matsumoto, K.. 1998. Segmentation of page images using the area voronoi diagram. Computer Vision and Image Understanding, vol. 70, No. 3, pp. 370–382.

[12] Kise, Z., Iwata, M., and Matsumoto, K.. 1999. On the application of voronoi diagrams to page segmentation. Proceedings of the Workshop on Document Layout Interpretation and Its Applications.

[13] Shi, K., and Govindaraju, V. 2004. Line separation for complex document images using fuzzy run length. Document Image Analysis for Libraries, 2004. Proceedings of the First International Workshop on, pp. 306–312.

[14] Sun, H. M. 2006. Enhanced constrained run-length algorithm for complex layout document processing. International Journal of Applied Science and Engineering, vol. 4, No. 3, pp. 297–309.

[15] Nikolaou, N., Makridis, M., Gatos, B., Stamatopoulos, N., and Papamarkos, N. 2010. Segmentation Of Historical Machine-Printed Documents Using Adaptive Run Length Smoothing And Skeleton Segmentation Paths. International Journal Of Image And Vision Computing 28, pp. 590–604.

[16] Ha, J., Haralick, R. M., and Phillips, N. 1995. Recursive X-Y Cut Using Bounding Boxes Of Connected Components. Proceedings Of The Third International Conference On Document Analysis And Recognition (ICDAR).

[17] Sutheebanjard, P., and Premchaiswadi, W. 2010. A Modified Recursive X-Y Cut Algorithm For Solving Block Ordering Problems. 2nd International Conference On Computer Engineering And Technology (ICCET).

[18] Chi, Z., Wang, Q., and Siu, W.C. 2003. Hierarchical content classification and script determination for automatic document image processing. Pattern Recognition, vol. 36, No. 11, pp. 2483– 2500.

[19] Chen, K., Yin, F., and Liu, C.L. 2013. Hybrid Page Segmentation With Efficient Whitespace Rectangles Extraction And Grouping. Pattern Recognition, vol. 36, No. 11, pp. 2483– 2500.

[20] Lin, M. W., Tapamo, J. R., and Ndovie, B. 2007. A Texture-based Method for Document Segmentation and Classification. Revue Africaine de la Recherche en Informatique et Mathématiques Appliquées, INRIA, vol. 6, pp. 49-56, 2007.

[21] Etemad, K., Doermann, D. S., and Chellappa, R. 1997. Multiscale segmentation of unstructured document pages using soft decision integration, IEEE Transactions on Pattern Analysis Machine Intelligence, vol. 19, No. 1, pp. 92–96.

[22] Eglin, V. 1998. Contributions à la structuration fonctionnelle des documents imprimés.

[23] Nicolas, S., Kessentini, Y., Paquet, T., and Heutte, L. 1997. Handwritten document segmentation using hidden markov random fields, ICDAR, vol. 1, pp. 212-216

[24] Etemad, K., Doermann, D. S., and Chellappa, R. 1997. Multiscale document page segmentation using soft decision integration. IEEE Transactions on Pattern Analysis Machine Intelligence.

[25] Journet, N, Mullot, R., Eglin, V., and Ramel, J. Y. 2006. Analyse d'images de documents anciens : Catégorisation de contenus par approche texture. Laurence Likforman-Sulem., SDN06, pp. 247-252.

[26] Wang, D, and Srihari, S. N. 1989. Classification of newspaper image blocks using texture analysis. Computer Vision, Graphics, and Image Processing, vol. 47, no. 3, pp. 327-352.

[27] Chajri, Y, and Bouikhalene, B. 2016. Handwritten mathematical symbols dataset Data in Brief, Vol.7, pp. 432-436, ISSN 2352-3409, doi:10.1016/j.dib.2016.02.060.