

# Unsupervised Categorization of Heterogeneous Text Images Based on Fractals

Badreddine KHELIFI; Nizar ZAGHDEN, Mohamed Adel ALIMI  
And Remy MULLOT

*REGIM: Research Group on Intelligent Machines,  
University of Sfax, ENIS, Department of Electrical Engineering.  
BP W - 3038, Sfax, Tunisia*

*khelifi\_badreddine@ieee.org; nizar.zaghden@ieee.org; adel.alimi@ieee.org;  
remy.mullot@univ-lr.fr*

## Abstract

*This paper deals about text extraction from heterogeneous documents for categorizing documents and indexing tasks. The purpose of this work is to find similar text regions basing on their fonts. First text regions are extracted, and then font matching is performed using fractal descriptors. Experiments are done for both maps and ancient documents.*

## 1. Introduction

Despite the progress in compression techniques and storage technologies, which allows saving huge quantity of paper format documents, thousands of documents remain in unused dead format all over the world. Techniques to render handling and managing these documents easier are needed. The text information is a major key which describes the content of documents; that's why text extraction is an active axle of research in various fields. Nevertheless, the more complex is the document, the harder recognition is needed. Many of studies have been dedicated to optical character recognition (OCR), but they usually neglect the font identification and the importance of the information that can give. In fact, in documents, the font style of the text can be used to differentiate between their roles. For example in maps, towns and roads are not written with the same font; likely in ancient documents where titles, subtitles and core have not the same font style. The basic idea is that the user starts with a query text example and retrieves a set of similar items. The query can be done to find textual

zones in the same document having the same font, or documents containing the same font.

Actually, Font recognition is a fundamental issue in the identification and analysis of documents. There are several techniques that have been proposed to solve this problem [1, 2, 3]. Some of them are related to the characteristics of language [4] but others are content independent [5]. The methods have been used to identify fonts in text rich documents, and in this work we try to identify them in heterogeneous content documents.

This paper has been organized as follows. The second section describes the adopted method used for categorizing fonts in heterogeneous images, giving details of different steps. The third section presents the results and the discussions of the experiments. Finally, concluding remarks are given and jointly focusing the attention to further investigation on robustness to noise, scale and rotation invariance, and to potential out comings to improve or affect the quality of recognition and query steps.

## 2. Methodology

The proposed methodology for maps and ancient documents categorization is illustrated in figure 1, and fully described in this section.

### 2.1. Preprocessing

The main idea for combining different clusters issued from different origins is that in both cases, text is affected with noise. We can't also forget the complexity of ancient and map documents and that it is

very helpful to develop a system which can do well with different kinds of degraded images. In map texts, we can really conclude that text clusters can't be extracted without any noise resulted from the presence of background and graphics, which in many cases needs steps of pretreatments before extracting features from text images. The ancient documents images are real scans of old books, hence an 'original' version without defects does not exist; Consequently, to characterize them well, we need to enhance these images.

The use of an adaptive Wiener filter based on statistics estimated from a local neighborhood around each pixel has proved efficient for the above goals.

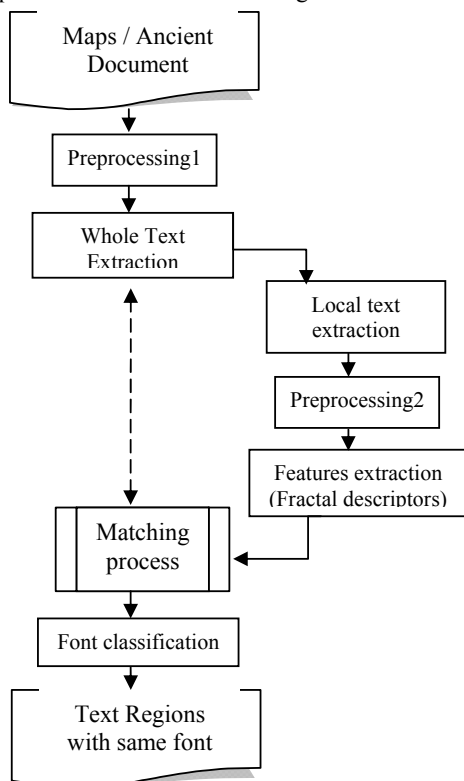


Figure 1. Flow chart of the proposed method

We notice that two preprocessing steps are performed according to the proposed method. The first one is described above, and is based on Wiener filter. While the second aims to prepare the extracted text to the treatments. It includes, local region text denoising then its centering and its normalization. In fact, eliminating noisy pixels is important to well achieve the normalization process, since it depends on the dimension of the bounded box given after centering (where extra white pixels are removed).

## 2.2. Text extraction

We distinguish two steps for text extracting: first the text separation from other data (graphics and background), second the extraction of local grouped text candidate regions.

Many attempts to separate text from other data in complex documents are done [6, 7, 8]. We use the method of [9] which applied to our documents (see Figure 2) which provides an acceptable text layer which contains textual zones in different orientations, with different size and fonts. This method is based on color characteristics to extract text from geographic maps and isolate it from other data. It is an enhanced approach which is applied on technical documents of engineering, and it is based on texture and the tracking of connected components. First, different color layers are extracted to have homogenous images. Then, the results are binarized, and text region candidates are grouped. The fact of grouping these zones helps to eliminate symbols accompanying strings. After that, an SVM classifier is performed to verify the membership of every connected component to text entities or no.



Figure 2. Text extraction from map example

The word grouping based on connected components is done likely in [9], and more details can be found there.

### 2.3. Features extraction

The CDB (Counting Densities per Box) given in [5] is calculated and used as a feature describing each extracted text region.

The fractal dimension is a useful method to quantify the complexity of feature details present in an image. In this paper, we compute fractal dimension according to the CDB method of [5] as it is asserted to give good results compared to other font recognition methods.

The used method for estimating fractal dimension is derived from the box counting method. We consider that the image of size  $M \times M$  pixels has been scaled down to a size  $s \times s$  where  $M/2 > s > 1$  and  $s$  is an integer. Then we have an estimation of  $r = s / M$ . The  $(x, y)$  space is partitioned into boxes  $(i, j)$  of size  $s \times s$ . On each box, we calculate the density of black pixels  $n_r(i, j)$ ;  $N_r = \sum n_r(i, j)$  represents the total contribution of the image.

Then  $N_r$  is counted for different values of  $r$ , and we can estimate fractal dimension from the least square linear fit of  $\text{Log}(N_r)$  against  $\text{Log}(1/r)$ .

$$D = \frac{\text{Log}(N)}{\text{Log}(1/r)}$$

Where :  $N$ : the number of boxes intersecting objects  
 $r = S/M$ ;  $S$ : size of boxes  
 $M$ : width of the object

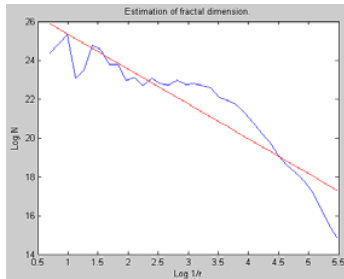


Figure 3. Fractal dimension estimation

### 2.4. Font classification

The font extraction gives us important information for document categorizing and indexing. In fact, for ancient documents, the font is a key to differentiate the latrine from the content; it gives also information about eventual titles and paragraphs.

In geographic maps, each kind of text information is represented sometimes with different color, but usually with a specific font (cities names, country name, Rivers...).

The idea of this step is inspired from [10]. Once fractal descriptors are calculated for extracted local text zones, we perform a font classification step. This Matching is achieved with the method of [1] according to Bayes criteria.

In fact in our method, we compare every text block index with others in the same image. Our object is to find the best text images which resembles to the first one. This method is also applied for different blocks issued from our heterogeneous database. The main contribution of this text categorization is that we deal with an unsupervised classification. This choice gives us a better liberty related to our application since we work with various types of text styles and that we can't train our system with all properties of textual information.

### 3. Discussion of the experimental results

In our work, we make experiment on maps given from geological department database. We test also our work on ancient documents from database used in [11, 12].

The first experiments show that we reach good results in text extraction as well in text classification. Figure 2 shows the extraction of the text layer. This layer contains text regions with different orientations and styles (size, font and language), but containing also, some extra noise (circles). The goal of this stage is to localize the block of characters belonging to the same class after being already pretreated the whole of the image.

In figure 4, we can observe three classes of fonts, which are represented with three colors (Red, Green and Blue). Nevertheless, some regions are not classed. This could be explained by the presence of noise (small circles) and the change of orientation (for two regions).

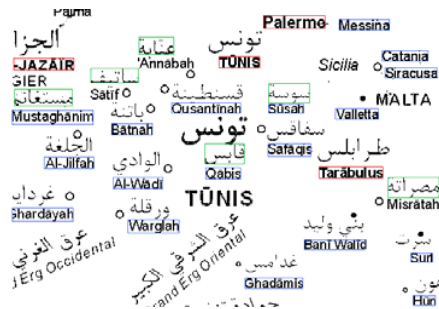


Figure 4. Extracted classes

The approach presented in this paper has been designed to overcome the difficulties presented in the

text [13] issued from both maps and historical documents by categorizing text blocks after doing as possible the denoising of the text surface.

The same work is done for ancient documents, and figure 5 shows the three classes extracted from the three existing fonts. They correspond to three separated parts on the document.

#### 4. Conclusion and perspectives

This subject concerns the characterization of the text contents of the corpuses with the aim of an indexation and of navigation. It is indeed important to treat the indexation by the analysis of the textual contents. The systematic point that we consider here is the too strong heterogeneousness of the envisaged corpuses and the state of the tools of recognition of the ancient texts and also text issued from maps. In this paper, we propose an approach of indexation or navigation in this very heterogeneous corpus on the basis of search for homogeneities in pages.

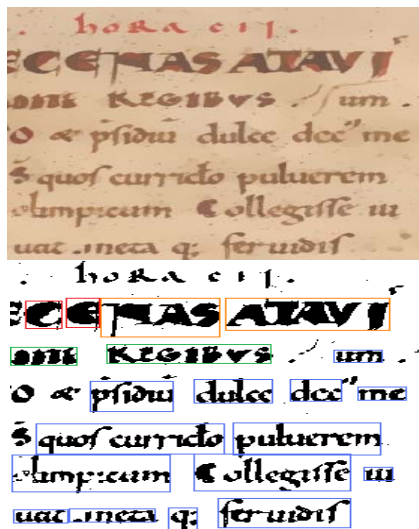


Figure 5. Results of ancient document treatment

In this work, the detection of text zones is done for every direction, but the font recognition is well achieved only for horizontal zones. Further investigations on other features related to the character itself, which are rotation invariant would improve the recognition step, and so the indexing process by looking for same characters with the same font.

The main design criteria for the matching procedure should be then quoted such as generality, extensibility, invariance (scale, rotation), robustness (to noise and distortion), and computational complexity.

#### References

- [1] H.M. Sun. "Multi-Linguistic Optical Font Recognition Using Stroke Templates". *International Conference on Pattern Recognition* 06(2):889-892, 2006.
- [2] L. Shijian, C.L. Tan. "Script and Language Identification in Noisy and Degraded Document Images". *Pattern Analysis and Machine Intelligence* 08(30):14-24, 2008.
- [3] A. L. Spitz, "Determination of Script and Language Content of Document Images," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 97(19/3):235-245, 1997.
- [4] A.M Alimi, "A Neuro-Fuzzy Approach to Recognize Arabic Handwritten Characters", *Proc. International Conference on Neural Networks, ICNN'97, Houston, TX, USA*, pp. 1397-1400, 1997.
- [5] N. Zaghden, S. BenMoussa, M.A. Alimi. "Reconnaissance des fontes arabes par l'utilisation des dimensions fractales et des ondelettes", *Colloque International Francophone sur l'Ecrit et le Document (CIFED 06)*, Fribourg(Suisse) (06) :277-282, Septembre 18-21, 2006.
- [6] N. Journet, V. Eglin, J-Y. Ramel, R. Mullot. "Text/Graphic labelling of Ancient Printed Documents". *International Conference on Document Analysis and Recognition*, Séoul (Corée), 05(02):1010-1014, 2005.
- [7] S. Khedekar, V.Ramanaprasad, S. Seltur., V. Govindaraju. "Text - Image Separation in Devanagari Documents". *Document Analysis and Recognition*, 03: 1265-1269, 2003.
- [8] Y. Zhan, W. Wang, W. Gao "A Robust Split-and-Merge Text Segmentation Approach for Images", *International Conference on Pattern Recognition* 06(2):1002-1005, 2006
- [9] B. Khelifi, N. Elleuch, S. Kanoun, A. M. Alimi. "Enhanced Color Based Method for Text Extraction from Maps", *International conference MM-CCA*, Jordan, 2007.
- [10] L. Qin, W. Wang, Q. Huang, W. Gao. "Unsupervised Texture Classification: Automatically Discover and Classify". *International Conference on Pattern Recognition* 06(2):433-436, 2006.
- [11] I. Moalla, F. Le Bourgeois, H.Emptoz, M.A Alimi. "Image Analysis for Palaeography Inspection". *2nd IEEE International Conference on Document Image Analysis for Libraries, Lyon, France*, 06:303-310, April 28-29, 2006.
- [12] C. L. Tan, R. Cao, P. Shen, "Restoration of archival documents using a wavelet technique", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 02(24/10): 1399-1404, October 2002.
- [13] W. Maghrebi, L. Baccour, M.A. Khabou, A.M. Alimi, "An Indexing and Retrieval System of Historic Art Images Based on Fuzzy Shape Similarity", *Mexican International Conference on Artificial Intelligence MICAI 2007, Lecture Notes in Computer Science*, pp. 623-633, 2007.