Old document image segmentation using the autocorrelation function and multiresolution analysis

Maroua Mehri, Petra Gomez-Krämer, Pierre Héroux, Rémy Mullot

Old document image segmentation using the autocorrelation function and multiresolution analysis

Maroua Mehri*†, Petra Gomez-Krämer*, Pierre Héroux†, and Rémy Mullot*
*L3I, University of La Rochelle, Avenue Michel Crépeau, 17042, La Rochelle, France
Emails: {maroua.mehri, petra.gomez, remy.mullot}@univ-lr.fr
†LITIS, University of Rouen, Avenue de l'Université, 76800, Saint-Etienne-du-Rouvray, France
Email: pierre.heroux@univ-rouen.fr

***Abstract*—Recent progress in the digitization of heterogeneous collections of ancient documents has rekindled new challenges in information retrieval in digital libraries and document layout analysis. Therefore, in order to control the quality of historical document image digitization and to meet the need of a characterization of their content using intermediate level metadata (between image and document structure), we propose a fast automatic layout segmentation of old document images based on five descriptors. Those descriptors, based on the autocorrelation function, are obtained by multiresolution analysis and used afterwards in a specific clustering method. The method proposed in this article has the advantage that it is performed without any hypothesis on the document structure, either about the document model (physical structure), or the typographical parameters (logical structure). It is also parameter-free since it automatically adapts to the image content. In this paper, firstly, we detail our proposal to characterize the content of old documents by extracting the autocorrelation features in the different areas of a page and at several resolutions. Then, we show that is possible to automatically find the homogeneous regions defined by similar indices of autocorrelation without knowledge about the number of clusters using adapted hierarchical ascendant classification and consensus clustering approaches. To assess our method, we apply our algorithm on 316 old document images, which encompass six centuries (1200-1900) of French history, in order to demonstrate the performance of our proposal in terms of segmentation and characterization of heterogeneous corpus content. Moreover, we define a new evaluation metric, the homogeneity measure, which aims at evaluating the segmentation and characterization accuracy of our methodology. We find a 85% of mean homogeneity accuracy. Those results help to represent a document by a hierarchy of layout structure and content, and to define one or more signatures for each page, on the basis of a hierarchical representation of homogeneous blocks and their topology.**

Keywords*—Segmentation, autocorrelation, directional rose, multiresolution, consensus clustering.

I. INTRODUCTION

A document has a structure which provides additional information. Without document structure, it would be difficult to index and retrieve correctly the information contained in the document. In this context, document structure analysis remains a fundamental step and crucial stage in any indexing and retrieval system such as optical character recognition (OCR) and graphic recognition modules. Several researches have been carried out to characterize the document layout with the result of structuring it into three different levels: the physical layout, the intermediate functional structure, and the logical structure analysis [1]. Firstly, the physical level defines both the typography and document organization. The typography sets the style of different kinds of informative regions (fonts, colors, lines, frames, etc.) and the form layout (line spacing, alignment, etc.). The document organization describes the layout of all the visual elements (characters, words, lines, blocks, columns, and non-text regions) which compose it and the topographical relationships between those elements (hierarchy, inclusion, neighborhood position). Secondly, the intermediate functional structure is a physical interpreted one which is adapted to the recognition of the logical structure. Finally, the logical level concerns the interpretation of different parts that compose the document and specifies the logical relationship between them [2].

In this work, we focus only on the first segmentation level i.e. the physical level. Several works have been presented on document image segmentation and characterization tools. There are two main kinds of page segmentation and characterization methods: the first one includes data-driven, model-driven, and hybrid methods and the second one is based on texture analysis.

A. Data-driven, model-driven, and hybrid methods

Data-driven, model-driven, and hybrid methods are devoted to contemporary document recognition and are significantly widespread in the literature because those methods are based on a strong *a priori* knowledge such as the repetitiveness of document structure in a corpus. This family of document structure analysis and page segmentation methods, mixing *a priori* knowledge and image analysis, can be classified in three main categories: data-driven, model-driven, and hybrid methods [1], [2].

The first class of methods, known as data-driven segmentation, does not include (or little) knowledge of the model. Those approaches are based on low-level data mining of pixels (color, position, etc.). They rely on the study of the space between pixels and the grouping of pixels in order to segment the different elements of the page content into blocks. For example, the Run Length Smearing Algorithm (RLSA) [3] studies the spaces between black pixels in order to link together neighboring black areas. There are certain limitations of data-driven segmentation methods: Firstly, they are based on the definition of complex criteria and rules. Secondly, those methods are sensitive to noise and not robust to slanted texts. Thus, the data-driven approaches are suitable for documents whose areas are clearly demarcated and rectangular. Furthermore, the pertinence of this segmentation category depends on the particular layout and idiosyncrasies of the documents.

The second class of methods (model-driven methods) is guided by a model of the document. Often used for well-defined and invariant structured documents, those approaches are based on strong *a priori* knowledge to guide the segmentation and recognition. For example, the XY-CUT algorithm [4] consists in computing the horizontal and vertical projection profiles (corresponding to the sum of the pixels along the horizontal axis and the vertical axis) of the whole document image and in iteratively splitting them into smaller ranges until a condition about hollow projections (interline spaces) has been satisfied. This requires the definition of criteria for cutting (and possibly fusion). Although the model-driven approaches are generally faster, they are not well-adapted to complex layout documents.

Finally, the hybrid methods are often non-sequential and combine data-driven and model-driven algorithms. For example, the hybrid method [5] uses a split-and-merge strategy. However, by combining tools from data-driven and model-driven approaches, hybrid methods can deal with a wide variety of documents and cope with complicated page segmentation problems, but many parameters and thresholds must be adjusted.

B. Segmentation methods based on texture analysis

To handle the drawbacks of the methods described above, new alternatives for document structure analysis based on texture have been developed, which ensure the segmentation and the characterization of information contained in a document. Those methods are pixel-based and do not require neither a document model nor *a priori* information relative to the semantic and physical characteristics of the document class. Thus, texture-based segmentation methods characterize generally complex documents aiming at segmenting the different elements of the page content into homogeneous blocks. A texture is defined as a spatially uniform distribution of local gray-value variations. In the literature [6], texture is defined as a suitable measure for the analysis of the block contents of the physical layout. Texture analysis methods have been used in image segmentation in order to extract textural characteristics. Texture-based segmentation methods can be classified into four categories: methods based on probabilistic models, and geometric, frequential, and statistical methods.

Markov random fields [7] and fractals [8] are both the most commonly used tools based on probabilistic models. This category of texture-based segmentation methods is complex to implement. There are many difficulties in the learning phase and a long computation time is required. Markov random fields are perfectly adapted to documents with high variability in terms of the layout and the quality of the scanned document, which yields good performances in handwritten documents. However, Markov random fields are not robust since the learning phase is only valid for one type of document at a time. Fractal dimensions compute measures of texture roughness and repeatability of a pattern. Fractals are considered as a useful tool for image segmentation when the image characteristics tend to be predictable and repetitive and in which the objects to segment tend to be irregular or different from the background.

The second class of texture analysis based segmentation methods is known as geometric methods. Those methods are used to describe intricate patterns to retrieve and to characterize the notion of a texton. Among the classics of geometric methods, moment-based texture segmentation [9] is one of the well-known methods. Anyway, moment-based texture segmentation is not sufficient to discriminate all types of texture and the algorithm needs a non-linear transformation of the images. Another geometric method [10] aims at extracting three classes: text, background and design from postal images based on six features derived from wavelet transforms. Even if the proposed algorithm has a good recognition rate, one of the features must be adjusted manually and the efficiency (computation time)

of the algorithm is limited.

Frequential methods like Gabor filters, Fourier transforms, and wavelets are widely used in indexing and segmentation of natural images. For instance, the unsupervised texture segmentation algorithm [11] is used to segment an input image into regions of homogeneous texture based on a bank of Gabor filters. Gabor filters have the advantage of reducing the computational complexity and are suitable for document texture analysis. One of the limitations of such an algorithm based on a fixed set of Gabor filters is that many parameters must be fixed [12]. Another frequential approach [13] combines the wavelet and the Fourier transform to index image databases. Although their algorithm is faster and more robust than the separate use of the discrete Fourier and wavelet transforms, the computation time is directly dependent on the level of wavelet decomposition. The method described in Ref. [14], combines kernel-based methods and a Gabor wavelet to segment document images scanned from popular newspapers and journals. According to the authors, the analysis of multiresolution and multiorientation properties of an image is ensured, but the effectiveness and computational complexity of the algorithm is no longer preserved and a proper post-processing is needed to improve the segmentation result.

The last class is statistical methods that have the advantage of being simple to implement and their effectiveness is proved. The GLCM (Grey Level Co-occurrence Matrix) [15] is one of the classics of texture analysis based statistical segmentation methods. By computing some indices on the GLCM [16], the texture regularity and repetitiveness are characterized. The authors of [17] propose a document page segmentation method using a neuro-fuzzy methodology. Another approach is the extraction algorithm of texture features [18] that is devoted to the analysis of documents. The computed texture features are based on frequencies and the autocorrelation function. This method gives good information on the principal orientations and periodicities of the texture allowing to characterize the content of images without any assumption on the image structure or properties. Although their results are promising, their algorithm is computationally expensive because it is carried out for each pixel and the size of the analysis window is a critical parameter that is difficult to determine. The authors of [19] introduce a novel lettrine segmentation method based a combination of different texture analysis approaches: GLCM [15], autocorrelation function [18], etc.

The authors of [20] claim that multiresolution in document image analysis and pyramid methods [21] in image processing allow to perceive texture at different scales and provide rich information like the characteristics of gray level distribution. There are two ways to implement multiresolution algorithms: The first solution consists in fixing the window size and changing the size of the image. The second solution is to keep the size of the original image and simply vary the size of the analysis window.

C. Scope

This work is a part of the DIGIDOC project (Document Image diGitisation with Interactive DescriptiOn Capability)[1]. Generally, the DIGIDOC project aims at simplifying and improving the archiving, processing, comparison and indexing of digitized old document images. Specifically, the objective of the DIGIDOC project is to integrate a module in scanners that will provide in addition to the scanned image a set of descriptors computed on it. Those descriptors dedicated to the acquisition, storage, analysis, and indexing of the scanned documents, will adapt the quality of the scanned document with respect to its content and to the subsequent use of the document image. This would help ensuring better interaction with scanners and offers new tools for document analysis.

The work presented in this paper is a part of our goal to provide a similarity measure between pages by defining one or more signatures, e.g. a graph of homogeneous regions, for each digitized page. Hence, a set of metadata characterizing the physical structure of pages in terms of homogeneous areas and topological relationships has to be proposed. In this context, we present a method to distinguish similar regions of the analyzed document without expressing any hypothesis about its physical structure and its logical structure. According to the literature, texture analysis based segmentation methods are the most appropriate choice if no assumption on document structure (model) or the typographical parameters (font size) should be made. Therefore, we present a fast and automatic layout segmentation to determine homogeneous areas in old document images based on texture analysis. Our proposal consists firstly in characterizing the content of old documents by extracting the autocorrelation features in the different areas of a page and at several resolutions. Then, we show that is possible to find automatically the homogeneous regions defined by similar indices of autocorrelation without knowledge about the number of clusters using adapted hierarchical ascendant classification (HAC) and consensus clustering (CC) approaches.

This paper is organized as follows. We detail in Section II our method for the segmentation of old documents and the characterization of their content by

[1]The DIGIDOC project is funded by the ANR (French National Research Agency), referenced under ANR-10-CORD-0020. For more details, see http://www.agence-nationale-recherche.fr/en/anr-funded-project/?tx_lwmsuivibilan_pi2[CODE]=ANR-10-CORD-0020

extracting five autocorrelation features in the different areas of the digitized page and at four resolutions. To assess our method, we present in Section III, the pertinence of our proposition with segmentation experiments which show homogeneous regions defined by similar indices of autocorrelation and we quantify the relevance of our layout analysis method with a new evaluation metric, the homogeneity accuracy. Our conclusion and future work are presented in Section IV.

II. SEGMENTATION USING THE AUTOCORRELATION FUNCTION AND MULTIRESOLUTION

We propose a fast automatic layout segmentation of old document images based on non-parametric tools: the autocorrelation function and multiresolution analysis.

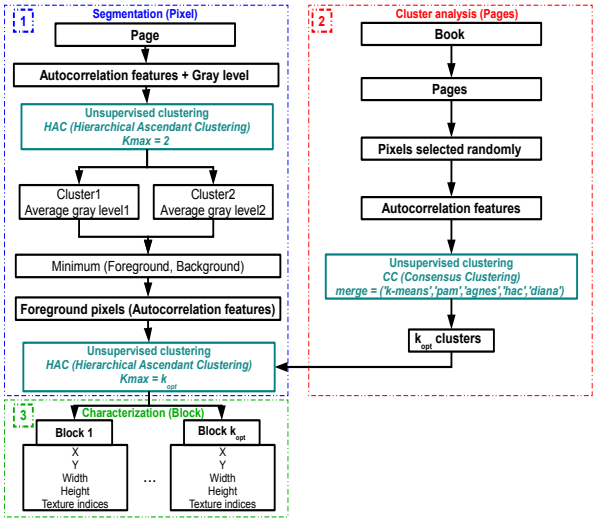

Fig. 1. Presentation of our automatic layout segmentation of old document images.

Our method focuses on the extraction of texture information without binarizing the image. We avoid a binarization of the image because it causes a loss of information specifically textural information. The pertinence of the segmentation experiments of old documents [18] that are based on the autocorrelation function leads us to work with autocorrelation features. The authors use the autocorrelation results to construct a rose of directions [22]. In order to obtain the orientation information, three features have been derived from the directonal rose: its main orientation, the intensity of the autocorrelation function for the main orientation, and the standard deviation of the intensities of the directional rose [18]. In order to extract texture information, we chose to compute five descriptors based on the autocorrelation function carried out in the different areas of a gray-scale page and at different resolutions. The proposed approach is pixel-based and independent of the model or type of document image. It does not require any *a priori* knowledge on the content or font styles of the document. Thus, our method is adapted to all kinds of document. The proposed approach is illustrated in Figure 1. It is composed of two main stages which are detailed below: Firstly, the computation of autocorrelation features for each page which are then used in an unsupervised clustering approach (block 1 on Figure 1) in order to determine and characterize the homogeneous regions in the document (block 3 on Figure 1). Secondly, we estimate the true number of clusters of the homogeneous regions defined by similar autocorrelation indices by performing the consensus clustering method applied to a number of pixels chosen randomly from a few pages of a book (block 2 on Figure 1). The clustering approach is more significant when we compute it on an entire book instead of processing each page of a book individually as our goal is to compare and index the content of digitized book.

A. Autocorrelation feature computation

The first stage of our proposal (see Figure 1) is to compute the autocorrelation features. We propose a feature vector composed of five texture indices all extracted from the autocorrelation function and based on a non-parametric tool, the directional rose [22]. The extraction of those five texture attributes helps us to characterize the content of the digitized document. The goal of our feature extraction process is to propose a set of metadata characterizing the physical structure of pages in terms of homogeneous areas and topological relationships. The texture features are computed at various sizes of analysis windows in order to adopt a multiresolution approach. We decide to keep the size of the original image and varying the size of the analysis window. The sliding window is shifted horizontally and vertically scanning the whole image. In order to avoid side effects, a quick and easy way to compute texture features on the whole image, is border replication. The extraction of autocorrelation descriptors per block is carried out at four different sizes of sliding windows: (16×16), (32×32), (64×64), and (128×128). Therefore, we obtain 20 numeric values (5 texture indices $\times$ 4 sliding window sizes) for each selected pixel from the digitized document image (see Figure 2). We will detail in the following the five texture indices.

The textual regions in the document are considered as textured areas while non-text contents in the document, such as blank spaces, graphics, and noise, are considered as regions with different textures. Therefore, we use the directional rose derived from the autocorrelation function to characterize homogeneous regions in a digitized document. The autocorrelation function

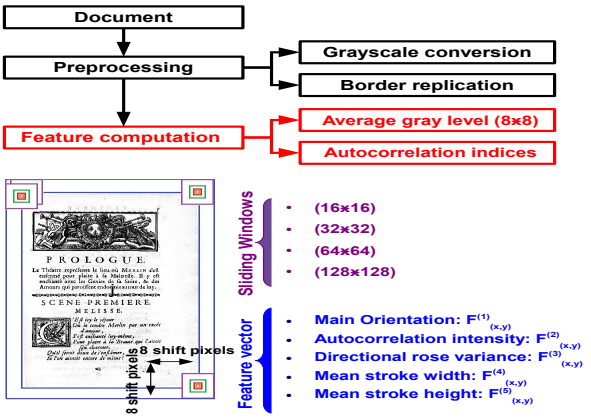

Fig. 2. Feature extraction.

$R^{I(\alpha,\beta)}_{(x,y)}$, considered as a statistical and global measure, is computed along the horizontal and vertical axes of the analysis window I of an image according to the following equation:

$$R^{I(\alpha,\beta)}_{(x,y)} = \sum_{\alpha\in\Omega}\sum_{\beta\in\Omega} I(x,y)I(x+\alpha,y+\beta) = FFT^{-1}([FFT\,[I(x,y)]\,FFT^{*}\,[I(x,y)]]) \quad (1)$$

where $I(x+\alpha,y+\beta)$ is the translation of the analysis window of an image $I(x,y)$ by α and β pixels along the horizontal and vertical axes respectively, defined on the plane Ω. FFT, $(.)^{*}$, and $(.)^{-1}$ denote respectively the Fast Fourier transform, the complex conjugate, and the inverse transform.

From the autocorrelation function, the directional rose can be deduced, that has been initially proposed by [22]. The rose of directions is a polar diagram based on the analysis of the autocorrelation results. In order to identify the main orientation of the analyzed image, the rose of directions is computed for each orientation by summing up the different values of the autocorrelation function (see equation (1)):

$$R^{I}_{(x,y)}(\Theta_i) = \sum_{D_i} R^{I(\alpha,\beta)}_{(x,y)} \quad (2)$$

where $\Theta_i \in [0,180]$ is the selected orientation of the set of the possible orientations D_i, which is represented by a straight line passing through (x,y) and the angle Θ_i. In order to select only the relative variations of all contributions for each direction, the authors of [18] present a normalization of the rose of directions. The definition of the relative sum $R'^{I}_{(x,y)}(\Theta_i)$ is:

$$R'^{I}_{(x,y)}(\Theta_i) = \frac{R^{I}_{(x,y)}(\Theta_i) - R^{I}_{min}}{R^{I}_{max} - R^{I}_{min}} \quad (3)$$

with $R^{I}_{max} \neq R^{I}_{min}$ and where R^{I}_{min} and R^{I}_{max} represent respectively the minimum and maximum value of $R^{I}_{(x,y)}(\Theta_i)$, computed both on the analysis window of an image $I(x,y)$.

In order to illustrate the relevance of discriminating textual regions from graphical ones in the analyzed document and to determine the main orientation of a texture, we present in Figure 3 the results of the directional rose for four different textures. It can be observed that the shape of the rose is different for each type of texture. For textual regions such as (c), the shape of the rose depends on the orientation of the text and the main information. The horizontal orientation (0° and 180°) is clearly identifiable in (g). For the drawing (d), the directional rose (h) is deformed.

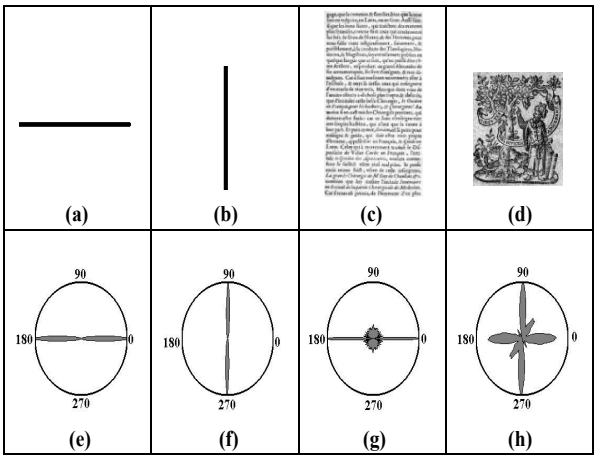

Fig. 3. Examples of directional roses. {**(a)**,**(b)**,**(c)**,**(d)**} are the original images and {**(e)**,**(f)**,**(g)**,**(h)**} are respectively their roses of directions.

A study of the variance of the shape of the rose applied to old grayscale documents has shown that the variety of textures does not define a homogeneous model of the rose of directions [18]. The computation of the rose helps us to extract significant and relevant indices. The authors of [18] define three texture features linked to the orientation information in order to analyze the digitized document and to describe their contents. The first texture feature $F^{(1)}_{(x,y)}$ corresponds to the main angle of the rose of directions extracted from its maximal intensity. It is normalized by the deviation from the horizontal angle in order to avoid handling circular data. It is given by:

$$F^{(1)}_{(x,y)} = \left\| 180 - \underset{\Theta_i\in[0,180]}{\operatorname{argmax}} \left(R'^{I}_{(x,y)}(\Theta_i)\right) \right\| \quad (4)$$

Likewise the second texture feature $F^{(2)}_{(x,y)}$ corresponds to the intensity of the autocorrelation function

for the main orientation (equation (4)), which is computed on the non-normalized value of the autocorrelation function (equation (2)). This feature evaluates the level of anisotropy of the analysis window of an image $I(x, y)$ since the directional rose ensures the association of gray levels of pixels in a specific direction. It is computed as:

$$F^{(2)}_{(x,y)} = R^{I}_{(x,y)}(\underset{\Theta_i \in [0,180]}{\operatorname{argmax}}(R'^{I}_{(x,y)}(\Theta_i))) \quad (5)$$

The third texture index $F^{(3)}_{(x,y)}$ characterizes the overall shape of the rose. $F^{(3)}_{(x,y)}$ is the variance of the intensities of the rose, except for the orientation of the maximal intensity. If $F^{(3)}_{(x,y)} \ll 1$, it means that the main orientation is significantly more prevalent than the other orientations. Otherwise, if the variance of the rose intensities is high, it signifies that the rose is deformed and a large number of orientations are present in different proportions (graphic blocks). Hence, the third texture attribute is:

$$F^{(3)}_{(x,y)} = \sigma^2(R'^{I}_{(x,y)}(\Theta_i)) \quad (6)$$

where $\Theta_i \in [0, 180] \setminus \operatorname{argmax}(R'^{I}_{(x,y)}(\Theta_i))$ and σ represents the standard deviation estimator.
$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(R'^{I}_{(x,y)}(\Theta_i))^2 - \frac{n}{n-1}(\mu)^2$ where μ, σ and n are respectively the mean value, standard deviation and 179 orientation values.

In addition to the three texture features that are related to the orientation information of the autocorrelation function, we introduce two other texture attributes also in relationship with the autocorrelation function: the mean stroke width and height [23] of an image. In contrast to the initial work which computes the features in horizontal and vertical direction [23], we propose estimating the mean stroke width and height accurately along the axis of the main angle of the directional rose in order to indicate the order of magnitude of the main strokes thickness.

The next texture index corresponds to the estimation of the mean stroke width $F^{(4)}_{(x,y)}$. It is computed from a derivative of the autocorrelation function along the axis of the directional rose's main angle Θ (see equation (4)) if $\Theta \in [10, 80]$, otherwise the mean stroke width is estimated along the horizontal axis. If the growth rate of the sequence defined in the equation (7) becomes lower than 10%, we estimate the mean stroke width, otherwise we continue to compute the sequence (equation (7)) until we reach the horizontal borders of the analyzed sliding window. Thus, we define $F^{(4)}_{(x,y)}$ as:

$$F^{(4)}_{(x,y)} = \sum_{\Theta \in [10,80]} \left\| I(x,y) - T^{\Theta}_{(\alpha,0)}(I(\frac{y}{|\tan(\Theta)|}, y)) \right\| \quad (7)$$

where $T^{\Theta}_{(\alpha,0)}(I(.,.))$ is the translation of the analysis window of an image I by α pixels along the axis of the main angle Θ of the directional rose and $\Theta = F^{(1)}_{(x,y)}$.

The computation of the last texture attribute is similar to that of the fourth texture index $F^{(4)}_{(x,y)}$. $F^{(5)}_{(x,y)}$ is an estimation of the mean stroke height computed along the axis of the directional rose's main angle Θ if $\Theta \in [10, 80]$, otherwise the mean stroke height is estimated along the vertical axis. If the growth rate of the sequence defined in the equation (8) becomes lower than 10%, we estimate the mean stroke height, otherwise we continue to compute the sequence (equation(8)) until we reach the vertical borders of the analyzed sliding window. $F^{(5)}_{(x,y)}$ is defined as:

$$F^{(5)}_{(x,y)} = \sum_{\Theta \in [10,80]} \left\| I(x,y) - T^{\Theta}_{(0,\beta)}(I(x, x * |\tan(\Theta)|)) \right\| \quad (8)$$

where $T^{\Theta}_{(0,\beta)}(I(.,.))$ is the translation of the analysis window of an image I by β pixels along the axis of the main angle Θ of the directional rose and $\Theta = F^{(1)}_{(x,y)}$.

B. Unsupervised clustering approach

After the computation of the autocorrelation features of all pages, we propose a non-parametric unsupervised clustering method (block 1 on Figure 1) which aims at determining the homogeneous regions. The homogenous regions are assumed to have similar autocorrelation indices. They are obtained without any knowledge on the number of clusters using hierarchical ascendant classification (HAC) and consensus clustering (CC). Firstly, we perform HAC in order to discriminate the foreground cluster which is considered as the most representative and significant pixels. But, the great variability of the old document content, like the various fonts and different types of drawing (ornaments, illuminations, drop caps, initial letters, frames, stamps, etc.), remains a difficulty in correctly predicting the optimal number of clusters. Thus, we propose to perform CC applied to a number of pixels chosen randomly from a few pages of a book in order to estimate the true number of clusters. Finally, we apply a third clustering step using HAC in order to classify all foreground pixels of the document in homogenous regions.

First of all, we perform a segmentation step with the help of non-supervised techniques which aims at extracting two clusters. One represents the information of the foreground (noise, text fields, drawings, etc.) and the other represents the background. The authors of [24] apply HAC on stroke features in order to classify the strokes of initial letters with interesting classification results. We chose to perform an adapted HAC using the Ward criterion [25], applied only on the average gray

level of the analyzed sliding window and setting the maximum number of clusters to two. In the following, we consider only the foreground cluster in order to deal with representative and significant pixels and to limit the amount of data.

After the extraction of the foreground pixels, a second clustering step is performed using a non-supervised technique. The objective is to determine the exact number of clusters and to find the homogeneous regions defined by similar autocorrelation indices. Typically, conventional unsupervised clustering techniques like partitioning and hierarchical methods [26] can not determine the ideal number of clusters. However, using the CC method [27], the ideal number of clusters can be obtained. The CC consists in calculating a consensus matrix obtained by iterating multiple runs of clustering algorithms with random and re-sampled clustering options. The consensus matrix analyzes the consistency of clustering results from five different clustering algorithms: agglomerative hierarchical clustering (agnes) [28], divisive analysis clustering (diana) [28], partitioning around representative objects (pam) [28], k-means clustering (k-means) [29], and hierarchical cluster analysis (hac) [30]. The authors of [27] show that hierarchical clustering methods are highly sensitive to outliers while partitioning ones are relatively insensitive. Therefore, they propose to use a new merged consensus clustering approach by applying a weighted averaging of the clustering results to estimate the true number of clusters.

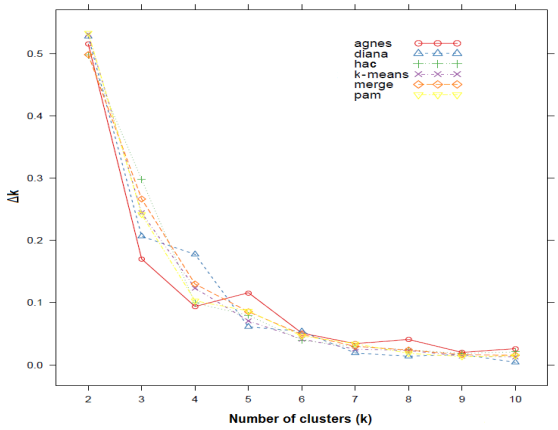

Fig. 4. Consensus clustering: Plot of Δk changes in area under the cumulative density curve for the consensus matrix for each clustering experiment against number of clusters k.

The clustering approach is more significant when we compute it on an entire book instead of processing each page of a book individually as our goal is to compare and index the content of a digitized book. Thus, in order to estimate the true number of clusters k, we use the consensus merge method [27] by determining k for five different clustering methods (agnes, diana, pam, k-means, hac) only for a set of randomly selected pixels of few pages of a book. By weighting the different clustering methods, we mitigate extremes in consensus values that can be created by the sensitivity of some algorithms. Finally, the optimal number of clusters corresponds to the largest change in area under the cumulative density curve Δk for the merge consensus matrix. An example of Δk is shown in Figure 4. The optimal number of clusters $k_{opt} = 2$ is estimated by finding the peak in the Δk values of the merge curve.

Once the number of clusters is known, we perform a third clustering step in order to classify all foreground pixels of the document in the homogenous regions. HAC is applied using the optimal number of clusters k_{opt} and the Ward criterion. Figure 5 illustrates the final result of the three clustering steps where two clusters were obtained. From (b) and (c), we can see that the page has been segmented into graphic regions (blue), which correspond to an ornament and a drop cap, and textual regions (red). (d) represents the proposed ground-truth (graphic regions (blue), textual regions (font1-red, font2-green)).

III. EXPERIMENTAL RESULTS

To assess our method, we present in this section the results of our algorithm on heterogeneous content. Moreover, we define a new evaluation metric, the homogeneity measure, which aims at evaluating the segmentation and characterization accuracy of our method.

A. Evaluation Corpus

In the context of a collaboration with the BnF within the DIGIDOC project, we had access to Gallica digital library[2] for old documents. The heterogeneity of pages, the database size and the degradation of certain documents of this corpus are examples representing specific issues and scientific challenges. This corpus includes many particularities of old documents, for instance a great variability of the page layout: complicated layout, random alignment, specific fonts, presence of embellishments (ornaments, illuminations, drop caps, initial letters, frames, etc.), variations in spacing between the characters, words, lines paragraphs and margins, and the superimposition of informations layers (stamps, handwritten notes, noise, back-to-front interference). For a first evaluation of our approach, we have selected 316 pages from 13 books of two categories: 7 printed monographs and 6 manuscripts that encompass six centuries (1200-1900) of French history.

[2]http://gallica.bnf.fr

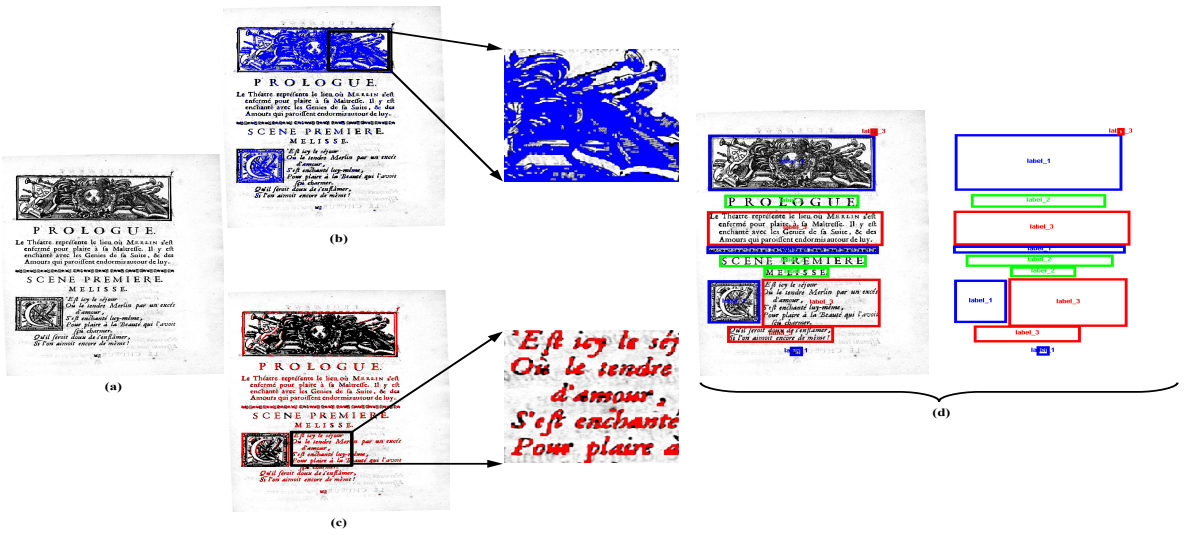

Fig. 5. Example of segmentation result: **(a)** original grayscale image, **(b)** cluster representing the graphics, **(c)** cluster representing the text, and **(d)** ground-truth.

For each category, we have decided to select three types of page content: 110 pages containing only two fonts, 100 pages containing graphics and single font texts, and 106 pages containing graphics and text with two different fonts.

B. Method evaluation

Our method reaches very satisfying results when comparing visually the segmentation results (see Figure 5). Indeed, this method of assessing the effectiveness of a segmentation method is inherently a subjective evaluation and we need to assess and evaluate the effectiveness using an appropriate quantitative metric. The lack of appropriate quantitative measures for segmentation quality and the difficulty of defining criteria for specific application-dependent segmentation are the shortcomings that limit researchers in an objective unsupervised evaluation of their results.

Evaluation of segmentation and region classification requires a ground-truth which is performed using the Ground-truthing Editor (GEDI)[3], a public domain document image annotation tool, providing us spatial boundaries of regions with labels. GEDI is considered as an annotation toolkit, which allows us to define manually bounding boxes drawn around each selected zone in order to define in page regions with different labels.

Several methods have been presented to measure the performance of segmentation methods [31]. For instance, the authors of [32] evaluate their segmentation using Jaccard coefficient. However, this coefficient is not suitable to assess the accuracy of our method because our goal is not an accurate pixel-based segmentation, but we are interested in finding the homogeneous regions defined by similar indices of autocorrelation. Thus, by defining our ground-truth, we define manually rectangular regions drawn around each selected zone and precise different labels when regions with different fonts (see Figure 5-(d)). Furthermore, each rectangular region is characterized by its location in the page, its height, its width, and a label. Then, we define an evaluation metric based on an homogeneity measure, which aims at evaluating the accuracy of our methodology in terms of matching regions. This metric is based on spatial overlaps of the ground-truth rectangle and the segmentation result. It takes into account the label correspondence between the ground-truth and result regions, based on the set of pixels they contain. Our evaluation accuracy is defined by the following equation:

$$H(B,G) = \frac{1}{|G|}\sum_{j}\frac{\max_{1\leq k\leq k_{opt}}(|b_i,(b_i\in g_j)\wedge(l_{B_i}=k)|)}{|\{b_i\in g_j\}|} \quad (9)$$

where $|.|$ is the number of pixels in the given block. $B = \{b_1, b_2, ..., b_i, ..., b_n\}$ and $G = \{g_1, g_2, ..., g_j, ..., g_m\}$ are respectively the sets of result blocks and rectangular regions of the ground-truth. $L_B = \{l_{B_1}, l_{B_2}, ..., l_{B_i}, ..., l_{B_n}\}$ corresponds to a set of labels obtained with our clustering methodology.

The results of our homogeneity measure (see equation (9)) are presented in Table I. We find a 85% of

[3]http://gedigroundtruth.sourceforge.net/

	Document category	Document content	Number of pages	$\mu(H)$	$\sigma(H)$	Max(H)	Min(H)
H(B,G)	Manuscript	One font and graphics	50	0,94	0,03	0,99	0,83
		Two fonts and graphics	56	0,84	0,05	0,97	0,64
		Only two fonts	50	0,87	0,05	0,96	0,72
	Printed	One font and graphics	50	0,84	0,14	1,00	0,60
		Two fonts and graphics	50	0,80	0,05	0,92	0,62
		Only two fonts	60	0,80	0,10	0,98	0,49

TABLE I. HOMOGENEITY METRIC. $\mu(H)$ AND $\sigma(H)$ ARE RESPECTIVELY THE MEAN VALUE AND STANDARD DEVIATION VALUES OF THE HOMOGENEITY MEASURE.

mean homogeneity accuracy. The overall results are quite satisfying especially for the manuscript document category which contains textual (one and two fonts) and non-textual regions. The mean homogeneity accuracy is 94% for the manuscript document category (one font and graphics). One assumption can be that the manuscript documents contain drawing regions that are more compact and homogeneous than the printed document ones. By comparing the average of homogeneity measure for different document categories, we observe that a greater $H(B, G)$ for pages containing graphics and single font texts. This yields that the extracted autocorrelation features are able to distinguish textual regions from graphical ones. We conclude that the autocorrelation descriptors are suitable to distinguish textual regions from graphical ones of the analyzed document without expressing any hypothesis either about its physical structure or its logical structure.

IV. CONCLUSIONS AND FUTURE WORK

We proposed in this paper a method for the segmentation and characterization of old document images without any *a priori* knowledge. The segmentation is based on a feature vector that is composed of texture indices, all based on the autocorrelation function and the directional rose. The texture features are automatically extracted in the different areas of a page and at several resolutions. The robustness of the extracted features is used in a non-parametric unsupervised clustering method which aims at determining the homogeneous regions which are assumed to be defined by similar indices of autocorrelation. Moreover, the number of clusters does not need to be known in advance as it is determined automatically.

The proposed method has been evaluated with promising results on 316 pages of old documents. We conclude that the used autocorrelation features allow a good discrimination of the foreground layers of the document, particularly of two classes: text and graphics. Results show that the autocorrelation descriptors are suitable to distinguish textual regions from graphical ones of the analyzed document.

The first aspect of future work will be to improve the classification by integrating a rejection mechanism taking into account a compactness criterion, e.g. the Mahalanobis distance, to classify all the pixels of a digitized document. Furthermore, we will elaborate frequential and statistical texture features in order to refine the segmentation and ensure the distinction between different text fonts and various graphic types. Besides, our clustering method performs classification on all pages of the document which can be very costly in case of large documents e.g. books. Thus, incremental clustering methods will be elaborated in order to reduce computation times.

This work gives reason to several perspectives. For instance, by representing a document with a hierarchical layout structure, a signature can be defined for each page. This page signature can be used in similarity measure for pages for the categorization of pages and retrieval.

V. ACKNOWLEDGMENTS

The authors would like to thank Geneviève CRON of the BnF for providing access to the Gallica digital library.

REFERENCES

[1] S. Mao, A. Rosenfeld, and T. Kanungo, "Document structure analysis algorithms: a literature survey," in *Document Recognition and Retrieval (DDR 2003)*. SPIE, 2003, pp. 197–207.

[2] R. Mullot, *Les documents écrits: De la numérisation à l'indexation par le contenu*. Hermès, 2006.

[3] K. Wong, R. Casey, and F. Wahl, "Document Analysis System," in *IBM Journal of Research and Development*. IBM Research Division, 1982, pp. 647–656.

[4] S. Khedekar, V. Ramanaprasad, S. Setlur, and V. Govindaraju, "Text - Image Separation in Devanagari Documents," in *International Conference on Document Analysis and Recognition (ICDAR 2003)*. IEEE, 2003, pp. 1265–1269.

[5] T. Pavlidis and J. Zhou, "Page Segmentation and Classification," in *Graphical Model and Image Processing (CVGIP 1992)*. Elsevier Science, 1992, pp. 484–496.

[6] B. Allier, J. Duong, A. Gagneux, P. Mallet, and H. Emptoz, "Texture Feature Characterization for Logical Pre-labeling," in *International Conference on Document Analysis and Recognition (ICDAR 2003)*. IEEE, 2003, pp. 567–571.

[7] S. Nicolas, Y. Kessentini, T. Paquet, and L. Heutte, "Handwritten document segmentation using hidden Markov random fields," in *International Conference on Document Analysis and Recognition (ICDAR 2005)*. IEEE, 2005, pp. 212–216.

[8] R. Ferrell, S. Gleason, and K. Tobin, "Application of fractal encoding techniques for image segmentation," in *International Conference on Quality Control by Artificial Vision VI (QCAV 2003)*. SPIE, 2003, pp. 69–77.

[9] M. Tuceryan, "Moment Based Texture Segmentation," in *Pattern Recognition Letters*. Elsevier Science, 1994, pp. 659–668.

[10] K. Varshney, "Block-segmentation and Classification of Grayscale Postal Images," in *Report in School of Electrical and Computer Engineering, Cornell University*, 2004.

[11] A. Jain and S. Bhattacharjee, "Text Segmentation Using Gabor Filters for Automatic Document Processing," in *Machine Vision and Applications*. Springer, 1992, pp. 169–184.

[12] S. Raju, P. Pati, and A. Ramakrishnan, "Text Localization and Extraction from Complex Color Images," in *International Symposium on Visual Computing (ISVC 2005)*. Springer, 2005, pp. 486–493.

[13] C. Sabharwal and S. Subramanya, "Indexing image databases using wavelet and discrete Fourier transform," in *Symposium on Applied Computing (SAC 2001)*. ACM, 2001, pp. 434–439.

[14] Y. Qiao, Z. Lu, C. Song, and S. Sun, "Document image segmentation using Gabor wavelet and kernel-based methods," in *International Symposium on Systems and Control in Aerospace and Astronautics (ISSCAA 2006)*. IEEE, 2006, pp. 450–455.

[15] R. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," in *Systems Man and Cybernetics (SMC)*. IEEE, 1973, pp. 610–621.

[16] M. Hall-Beyer, "GLCM Texture: A Tutorial," in *National Council on Geographic Information and Analysis Remote Sensing Core Curriculum*, 2000.

[17] L. Caponetti, C. Castiello, and P. Górecki, "Document page segmentation using neuro-fuzzy approach," in *Applied Soft Computing*. Elsevier Science, 2008, pp. 118–126.

[18] N. Journet, J. Ramel, R. Mullot, and V. Eglin, "Document image characterization using a multiresolution analysis of the texture: application to old documents," in *International Journal of Document Analysis and Recognition (IJDAR 2008)*. Springer-Verlag, 2008, pp. 9–18.

[19] S. Uttama, P. Loonis, M. Delalandre, and J. M. Ogier, "Segmentation and Retrieval of Ancient Graphic Documents," in *International conference on Graphics Recognition: Ten Years Review and Future Perspectives (GREC 2006)*. Hong Kong, China: Springer-Verlag, 2006, pp. 88–98.

[20] A. Lemaitre, J. Camillerapp, and B.Coüasnon, "Multiresolution Cooperation Improves Document Structure Recognition," in *International Journal of Document Analysis and Recognition (IJDAR 2008)*. Springer-Verlag, 2008, pp. 97–109.

[21] S. Contassot-Vivier, G. L. Bosco, and N. C. Dao, "Multiresolution approach for image processing," in *Erasmus ICP-A-2007*, 1996.

[22] S. Bres, "Contributions à la quantification des critères de transparence et d'anisotropie par une approche globale: application au contrôle de qualité de matériaux composites," Ph.D. dissertation, Institut National des Sciences Appliques de Lyon, Lyon, France, 1994.

[23] A. Ouji, Y. Leydier, and F. LeBourgeois, "Chromatic / Achromatic Separation in Noisy Document Images," in *International Conference on Document Analysis and Recognition (ICDAR 2011)*. IEEE, 2011, pp. 167–171.

[24] G. Nguyen, M. Coustaty, and J. Ogier, "Stroke feature extraction for lettrine indexing," in *International Conference on Image Processing Theory Tools and Applications (IPTA 2010)*. IEEE, 2010, pp. 355–360.

[25] J. Ward, "Hierarchical Grouping to Optimize an Objective Function," in *Journal of the American Statistical Association (JASA 1963)*. JSTOR, 1963, pp. 236–244.

[26] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. 2nd Edition Wiley-Interscience, 2001.

[27] T. Simpson, J. Armstrong, and A. Jarman, "Merged consensus clustering to assess and improve class discovery with microarray data," in *Boston Medical Center Bioinformatics (BMC Bioinformatics 2010)*. BioMed Central, 2010, pp. 1471–1482.

[28] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, 1990.

[29] J. B. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," in *Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, 1967, pp. 281–297.

[30] G. N. Lance and W. T. Williams, "A general theory of classificatory sorting strategies 1. Hierarchical systems," in *The Computer Journal*. Oxford University Press, 1967, pp. 373–380.

[31] S. Wontaek, M. Agrawal, and D. Doermann, "Performance Evaluation Tools for Zone Segmentation and Classification (PETS)," in *International Conference on Pattern Recognition (ICPR 2010)*. IEEE, 2010, pp. 503–506.

[32] F. Ge, S.Wang, and T. Liu, "New benchmark for image segmentation evaluation," in *Journal of Electronic Imaging*. SPIE, 2007, pp. 1–16.