Texture feature evaluation for segmentation of historical document images

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

Texture Feature Evaluation for Segmentation of Historical Document Images

Maroua Mehri
L3I, University of La Rochelle
La Rochelle, 17042, France
maroua.mehri@univ-lr.fr

Petra Gomez-Krämer
L3I, University of La Rochelle
La Rochelle, 17042, France
petra.gomez@univ-lr.fr

Pierre Héroux
LITIS, University of Rouen
Saint-Etienne-du-Rouvray, 76800, France
pierre.heroux@univ-rouen.fr

Alain Boucher
L3I, University of La Rochelle
La Rochelle, 17042, France
alain.boucher@univ-lr.fr

Rémy Mullot
L3I, University of La Rochelle
La Rochelle, 17042, France
remy.mullot@univ-lr.fr

ABSTRACT

Texture feature analysis has undergone tremendous growth in recent years. It plays an important role for the analysis of many kinds of images. More recently, the use of texture analysis techniques for historical document image segmentation has become a logical and relevant choice in the conditions of significant document image degradation and in the context of lacking information on the document structure such as the document model and the typographical parameters. However, previous work in the use of texture analysis for segmentation of digitized historical document images has been limited to separately test one of the well-known texture-based approaches such as autocorrelation function, Grey Level Co-occurrence Matrix (GLCM), Gabor filters, gradient, wavelets, etc. In this paper we raise the question of which texture-based method could be better suited for discriminating on the one hand graphical regions from textual ones and on the other hand for separating textual regions with different sizes and fonts. The objective of this paper is to compare some of the well-known texture-based approaches: autocorrelation function, GLCM, and Gabor filters, used in a segmentation of digitized historical document images. Texture features are briefly described and quantitative results are obtained on simplified historical document images. The achieved results are very encouraging.

Categories and Subject Descriptors

I.4.6 [**Segmentation**]: Pixel classification; I.7.5 [**Document Capture**]: Document analysis

1. INTRODUCTION

The idea of scanning historical books in order to build digital libraries and platforms of scanned books becomes more and more necessary due to the emergence and development of new uses. They are related to the desire for a quicker analysis, indexing, and retrieval of the scanned documents as well as an increase of the life time of ancient books. Therefore, there is a great demand for automatic document image segmentation and characterization tools, which are a basis for analysis, indexing, and retrieval. In the context of the DIGIDOC project (Document Image diGitisation with Interactive DescriptiOn Capability)[1], we address the problem of segmenting automatically historical digitized document images. The overall goal of the DIGIDOC project is to develop a number of unsupervised and automatic techniques of feature extraction performed on the scanned document image. Those features are dedicated to the acquisition, storage, analysis, and indexing of the scanned documents, and will characterize the content of ancient book pages in terms of homogeneous regions and topological relationships by using intermediate level metadata (between image and document structure). These metadata will represent a book page by a hierarchy of homogeneous regions without any hypothesis on the document structure, neither on the document model nor the typographical parameters.

Many methods have been presented in the literature to perform this task (e.g. RLSA [44], XY-CUT [17], etc.). However, such algorithms rely on *a priori* knowledge in order to properly segment and characterize the document image content. Recently, texture feature extraction and analysis approaches [3, 22, 13, 27] have been investigated for complex document layouts in the context of missing information on the document structure such as the document model and the typographical parameters. Indeed, the segmentation methods based on texture feature extraction and analysis are an effective way and suitable alternative for solving such a prob-

[1]The DIGIDOC project is supported by ANR (French National Research Agency) and is 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`

lem as they provide a global measure of region characteristics and also segment the content of the analyzed document image into homogeneous blocks based on extracted textural descriptors. Moreover, the texture-based segmentation approach has the advantage of avoiding any hypothesis on the document structure, neither about the document model (physical structure), nor the typographical parameters (logical structure).

Texture-based approaches are based on the two following hypotheses. Firstly, the textual regions in a digitized document are considered as textured areas while their non-text contents are considered as regions with distinct textures. Secondly, text of different fonts is also distinguishable [14]. Thus, texture feature extraction is performed in order to identify automatically homogeneous regions and to ensure the distinction between distinct text fonts and various graphic types. Jain *et al.* [12] demonstrate the effectiveness of a texture-based segmentation method for a variety of document image processing issues: text-graphics separation, address-block location, etc. Among the most widely used texture feature extraction and analysis methods are those derived from statistical, geometrical, model-based, and signal processing primitives [4]. In this work, we are interested in three well-known texture-based segmentation approaches: two statistical ones, the autocorrelation function [35] and the GLCM (Grey Level Co-occurrence Matrix) [10], and a frequential approach, the Gabor filters [9].

The autocorrelation function [13, 31, 27], GLCM [28, 3], and multiple channel Gabor filters [11, 45, 24, 5] are investigated in independent experiments in order to extract texture features and segment document images. A few comparative studies of Gabor and co-occurrence features for document segmentation and script and language identification are presented in [32, 38]. More comparisons [23, 6] can be found concerning Gabor features and gradient features for character recognition. Shahabi and Rahmati [40] propose a new method for writer identification of handwritten documents by combining Gabor and co-occurrence features. Qiao and al. [36] combine Gabor wavelets and kernel-based methods for document image segmentation. Eglin *et al.* [7] propose to use the results of autocorrelation features in order to compute the Gabor descriptors for handwriting classification in ancient manuscripts. Said *et al.* [38] present a global method for handwriting identification based on the use of multi-channel Gabor filters and co-occurrence matrices.

In a previous work [27], we have introduced a framework of segmentation and characterization of the content of an entire book, based on autocorrelation features. In this work, we propose to introduce new tasks in order to ensure a significant gain in the computation time and memory, and an improvement in the performance of our approach. Then, we integrate a new unsupervised phase enabling us to automatically label content pixels with the same cluster identifier regarding to the book content. Finally, we present a comparative analysis of different texture features in a context of segmentation of historical document images. The robustness of the extracted features is evaluated in a set of simplified historical document images by computing several internal and external clustering accuracy metrics.

The remainder of this paper is structured as follows. Section 2 presents an overview of our work by describing our framework for the segmentation and characterization of the content of ancient digitized book content. In order to validate and evaluate our framework, we chose three kinds of texture primitives: autocorrelation, co-occurrence, and Gabor which are detailed in Section 3. In Section 4, we propose a comparative analysis of the performance of the chosen categories of texture features. Our conclusions and future work are presented in Section 5.

2. FRAMEWORK

In order to ensure the characterization of homogeneous regions in the digitized book content, we propose a framework based on texture features and multiresolution analysis. The proposed framework is automatic, pixel-based, and adapted to all kinds of ancient books. The proposed framework is depicted in Figure 1. It consists in two main steps. By selecting randomly a number of foreground pixels from a few pages of the same book content, we compute firstly their textural features and we estimate the true number of clusters of homogeneous regions in the analyzed book by applying the Consensus Clustering method (CC) [42] on the extracted texture attributes (block 2 in Figure 1). Then, for each analyzed book page we extract its texture features which are then used in an unsupervised clustering approach by taking into consideration the estimation of the number of clusters given before by the CC method (block 1 in Figure 1). Those two steps, which are described in the following, aim to determine and characterize the homogeneous regions in the digitized book (block 3 on Figure 1).

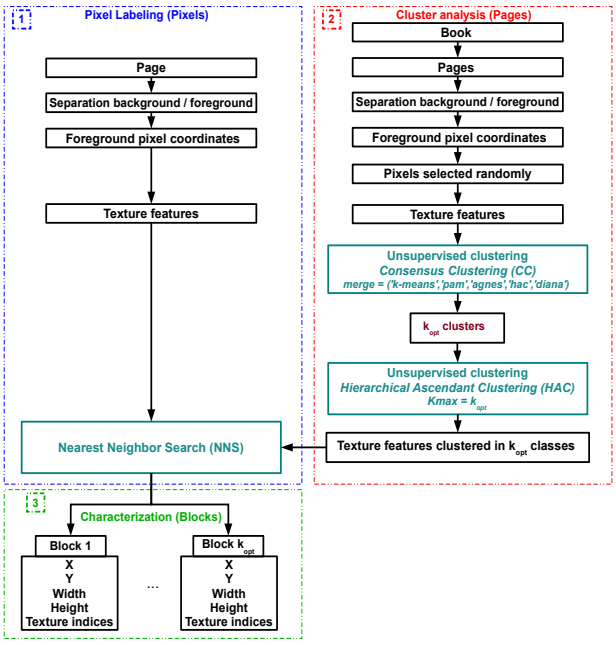

Figure 1: Presentation of our pixel labeling approach for historical digitized book content.

2.1 Estimation of the number of clusters

As our objective is to find the homogeneous regions defined by similar texture features, we need to use a clustering algorithm in order to partition the analyzed document image

into regions which have similar properties with respect to the extracted features. However, the conventional unsupervised clustering techniques [21, 25, 15] require that the number of clusters must be specified *a priori*. Therefore, the first stage of our framework is to determine the number of clusters from the extracted textural features of the whole analyzed book. Nevertheless, to perform this task on all pages of the analyzed book is not efficient in the case of large document images, because it needs a high computational time and memory. Therefore, we select randomly a number of foreground pixels from a few pages of the analyzed book. The foreground pixel selection step is performed by using the Otsu method [30] for the purpose of retrieving only pixels representing the information of the foreground (noise, text fields, drawings, etc.). By convention, white pixels are considered as background and black ones as foreground. Shijian and Tan [41] binarize document images by using Otsu's method in order to identify scripts and languages of noisy and degraded document images. Using a global thresholding approach, the Otsu method provides an adequate and a fast mean of binarization in order to extract texture features. The foreground pixel selection has ensured a reduction of the data cardinality, a significant gain in computation time and memory, and an improvement in the homogeneity accuracy average compared to our previous method [27]. We are not looking for an accurate segmentation, but we aim at finding regions with similar textural content. Then, we compute the texture features for each selected pixel from the set of random foreground pixels.

Previous work has identified a number of approaches [16] for determining the correct number of clusters in a dataset. Simpson *et al.* have recently proposed an effective method for estimating the optimal number of clusters in biological data, the Consensus Clustering method (CC) [42]. Therefore, we estimate the true number of clusters in a set of randomly selected foreground pixels from few random pages of a book by using the CC method. The idea of the CC method is to aggregate and combine the results over multiple runs of different clustering algorithms, including AGglomerative NESting (AGNES) [15], DIvisive ANAlysis clustering (DIANA) [15], Hierarchical Ascendant Classification (HAC) [21]), (k-means clustering (k-means) [25], and Partitioning Around Medoids (PAM) [15], in order to provide an averaged clustering robustness, i.e. a merge consensus matrix. The optimal number of clusters k_{opt} corresponds to the largest change in the area under the cumulative density curve which is computed from the cumulative density function of the merge consensus matrix cross a range of possible values of number of clusters.

2.2 Pixel classification and labeling

Since we estimate the optimal number of clusters k_{opt}, we can use an unsupervised clustering method to find the k_{opt} homogeneous regions defined by similar texture indices on the whole book. Due to the high requirement of a too large amount of memory to perform the CC method on all pixels from each analyzed document image which means processing more than 570,000 elements, we use HAC on the extracted textural features without taking into account the spatial coordinates and by setting the maximum number of clusters to the previously estimated k_{opt}. HAC is processed according to a hierarchical structure grouping of clusters based on the criteria of the minimum increased intra-cluster inertia. The work presented in [29] has shown interesting results in classifying the strokes of initial letters by using the HAC algorithm.

In our work, HAC is applied on the texture features of the selected pixels of book pages. This stage of processing gives k_{opt} clusters for the randomly selected samples of a book. This task, although not effective because of the high computational time of the HAC algorithm, is essential to find the k_{opt} homogeneous regions defined by similar texture indices on the whole book.

Finally, performing the Nearest Neighbor Search algorithm (NNS) [18] with the Mahalanobis distance [26] is necessary to assign the same label for each similar cluster extracted from the digitized book. NNS is used between each texture feature vector of each digitized page of the same book and the k_{opt} clusters of the selected samples of a book in order to find the closest texture feature vector to the cluster of the selected samples of a book. The squared Mahalanobis distance takes into account the dataset correlations and is particularly suited to arbitrarily shaped clusters.

3. TEXTURE FEATURES

The extraction of textural descriptors is performed on gray-level images. The texture features are performed at various sizes of analysis windows in order to adopt a multiresolution approach. The optimal sizes of the sliding windows, respecting a constructive compromise between the computation time and segmentation quality, have been determined experimentally. The extraction of textural descriptors per block is performed at four different sizes of sliding windows: (16×16), (32×32), (64×64), and (128×128). In order to avoid side effects, we use border replication allowing computing texture features on the whole image.

In order to validate and evaluate our framework of segmentation and characterization of ancient digitized books, three texture primitives are computed: the autocorrelation, co-occurrence and Gabor features that are outlined below.

3.1 Autocorrelation features

The first features tested in this paper are the autocorrelation descriptors. We investigate a non-parametric tool based on the autocorrelation function. The autocorrelation function (see equation (1)) is defined as a similarity measure between a dataset and a shifted copy of the data. It is used to find periodic patterns and can characterize similarity patterns through a number of extracted autocorrelation features. Previous works [13, 31, 27] have identified a number of autocorrelation features for segmenting ancient and contemporary documents images. Some of them use the directional rose [2], a derivative of the autocorrelation function. The directional rose (see equation (2)) reveals significant orientations of the texture in the analyzed block image. The autocorrelation descriptors highlight interesting information on the principal orientations and periodicities of the texture allowing characterizing the content of images without any assumption on the page structure and its characteristics. Thus, we compute five autocorrelation features, which have been reported in our previous work [27]: the main orientation of the directional rose, the intensity of the

autocorrelation function for the main orientation, the variance of the intensities of the directional rose, and the mean stroke width and height estimated accurately along the axis of the main angle of the directional rose, as shown in Table 1 [13, 31, 27]. Extracting these autocorrelation indices using a sliding window gives a total of 20 features (5 autocorrelation indices $\times$ 4 sliding window sizes for multiresolution). Therefore, we associate to every selected foreground pixel from the digitized document image a vector which corresponds to the results of the autocorrelation attribute extraction.

The autocorrelation function, which is computed along the horizontal and vertical axes of the analysis window I of an image according to the following equation:

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

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.

The directional rose, which is computed for each orientation by summing up the different values of the autocorrelation function, is given by:

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

Table 1: Autocorrelation features extracted from the autocorrelation function and the directional rose.

Features	Expression
Main angle	$F^{(1)}_{(x,y)} = \lvert 180 - \text{argmax}_{\Theta_i\in[0,180]}(R'^{I}_{(x,y)}(\Theta_i))\rvert$
Intensity	$F^{(2)}_{(x,y)} = R^{I}_{(x,y)}(\text{argmax}_{\Theta_i\in[0,180]}(R'^{I}_{(x,y)}(\Theta_i)))$
Variance	$F^{(3)}_{(x,y)} = \sigma^2(R'^{I}_{(x,y)}(\Theta_i))$
Mean stroke width	$F^{(4)}_{(x,y)} = \sum_{\Theta\in[10,80]} \lvert I(x,y) - T^{\Theta}_{(\alpha,0)}(I(\frac{y}{\lvert\tan(\Theta)\rvert},y))\rvert$
Mean stroke height	$F^{(5)}_{(x,y)} = \sum_{\Theta\in[10,80]} \lvert I(x,y) - T^{\Theta}_{(0,\beta)}(I(x,x*\lvert\tan(\Theta)\rvert))\rvert$

In Table 1, $\Theta_i \in [0,180]$ 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. $R'^{I}_{(x,y)}(\Theta_i)$ is a normalization of the rose of directions. σ represents the standard deviation estimator.

3.2 Co-occurrence features

The second features tested in this paper are the GLCM attributes [10]. The GLCM is a classic of statistical texture-based segmentation methods. The GLCM is an estimate of the second order probability density function of image pixels. This matrix determines the probability of occurrence of pixel pairs according to their gray levels and distance by considering the spatial relationship of pixels in the image. A GLCM element is the probability of the gray level pairs defined in a specified direction θ and separated by a particular distance of d units. The co-occurrence descriptors are then statistics computed from the GLCM. Multi-distance and multi-direction can be applied to extract a large number of co-occurrence descriptors.

Fourteen textural features extracted of the GLCM have been initially introduced by [10] for texture discrimination of natural and satellite images. A survey of document segmentation methods using texture analysis [33] presents different methods for segmenting document images and proposes a novel texture analysis approach based on the assembly of n^{th} order co-occurrence information within a processing window. This study states the texture co-occurrence technique as the best one in terms of processing time and complexity. A number of other works for co-occurrence feature extraction and analysis have also been proposed in order to segment and classify the content of document images [28, 22]. More methods based on the GLCM feature analysis have been proposed in the literature [34, 3] in order to identify script and language from document images. Briefly, the GLCMs are constructed for a small range of distance values $d = 1, 2$ and typically for the directions $\theta = \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$ [3].

We extract from the GLCM matrices six co-occurrence features: the maximum entry in the GLCM or the maximum probability, the correlation metric, the energy or the angular second moment, the entropy, the inertia or the contrast, and the local homogeneity for two distances $\{d = 1, 2\}$ [28, 3]. In addition to the twelve co-occurrence features (six for each distance), two other descriptors are computed: the mean value and the standard deviation of the energy for the two distances combined [22]. The co-occurrence features extracted from GLCM are illustrated in Table 2. Using a sliding window, we associate with every selected pixel from the digitized document image a metadata corresponding to the co-occurrence descriptor extraction. This metadata corresponds to a set of vectors composed of 56 numeric values (14 co-occurrence indices $\times$ 4 sliding window sizes for multiresolution).

In Table 2, $p_{d,\theta}(i,j)$ is the probability of the gray level pairs i and j defined in a specified direction θ and seperated by a particular distance of d units.

$$p_r(i) = \sum_{i=0}^{255} p_{d,\theta}(i,j) \quad \mu_r = \sum_{i=0}^{255} p_r(i) \quad \sigma_r^2 = \sum_{i=0}^{255} i^2 p_r(i) - \mu_r^2$$

$$p_c(j) = \sum_{j=0}^{255} p_{d,\theta}(i,j) \quad \mu_c = \sum_{j=0}^{255} p_c(j) \quad \sigma_c^2 = \sum_{j=0}^{255} j^2 p_c(j) - \mu_c^2$$

$$D(k) = \sum_{\substack{0\le i\le 255 \\ |i-j|=k}} \sum_{0\le j\le 255} p_{d,\theta}(i,j)$$

Table 2: Co-occurrence features extracted from GLCM.

Features	Expression
Maximum probability	$F_d^{(1)} = \max_{i,j} \{p_{d,\theta}(i,j)\}$
Correlation metric	$F_d^{(2)} = \sum_{i=0}^{255} \sum_{j=0}^{255} \frac{(i-\mu_r)(j-\mu_c)p_{d,\theta}(i,j)}{\sigma_r \sigma_c}$
Energy	$F_d^{(3)} = \sum_{k=0}^{255} D(k)$
Entropy	$F_d^{(4)} = -\sum_{k=0}^{255} D(k) \log_2 D(k)$
Contrast	$F_d^{(5)} = \sum_{k=0}^{255} k^2 D(k)$
Local homogeneity	$F_d^{(6)} = \sum_{k=0}^{255} \frac{D(k)}{1+k^2}$
Energy mean	$F_{d=1,2}^{(13)} = \sum_{k=0}^{510} kD(k)$
Energy standard deviation	$F_{d=1,2}^{(14)} = \sqrt{\sum_{k=0}^{510} (k - F_{d=1,2}^{(13)})^2 D(k)}$

3.3 Gabor features

The last features tested in our work are the Gabor features extracted using the multichannel Gabor filtering technique. The channel filters, known as 2-D Gabor functions, have been shown to be pertinent for segregation of textural regions of distinct spatial frequency, orientation or phase properties [1]. A 2-D Gabor filter (see equation (3)) is a linear selective band-pass filter, dependent on two parameters: spatial frequency f and orientation θ. Indeed, it consists in a Gaussian kernel function modulated by a sinusoidal plane wave. The spatial frequency f determines the distance from the Gaussian centers to the origin and the orientation θ specifies the angle from the horizontal axis, i.e. α-axis to the Gaussian centers.

Zhu *et al.* [45] propose a texture-analysis-based algorithm for automatic font recognition by extracting Gabor features. The authors show a 99, 1% of mean recognition rate. Ma and Doermann [24] propose a Gabor filter based multi-class classifier in order to identify scripts, font-faces, and font-styles. Jain *et al.* [12] show the effectiveness of the application of a multi-channel Gabor filtering-based texture segmentation approach for the segmentation and classification of document images. They choose the five following spatial frequencies: $4\sqrt{2}$, $8\sqrt{2}$, $16\sqrt{2}$, $32\sqrt{2}$, and $64\sqrt{2}$. The four directions 0, $\pi/4$, $\pi/8$, and $3\pi/4$ are widely used in the literature [11, 12, 45, 24].

One of the most serious disadvantages of the Gabor approach is its high computational cost, because the Gabor technique consists in convoluting the whole document image at each orientation and at each frequency. Anyway, to extract the Gabor features, we apply 24 Gabor filters (six different spatial frequencies f=$\{2\sqrt{2}, 4\sqrt{2}, 8\sqrt{2}, 16\sqrt{2}, 32\sqrt{2}, 64\sqrt{2}\}$ and four different orientations θ=$\{0, \pi/4, \pi/8, 3\pi/4\}$). We set the space of the Gabor filter constant $\sigma = \sigma_x = \sigma_y =$ 1. We compute then the mean value and the standard deviation of the magnitude response of each Gabor filter, as shown in Table 3. The feature vector (with dimension 48 to represent 24 channels) is constructed based on the computed mean and standard deviation of the magnitude of the transformed analyzed image by the selective Gabor filter. Thus, a total of 48 features per pixel are extracted in the analyzed sliding window. They form a 192-dimensional feature vector (48 Gabor indices × 4 sliding window sizes for multiresolution).

The Gabor transform of an image $I(x, y)$ is:

$$I_{(G_{f,\theta})}(x,y) = \sum_{\alpha \in \Omega} \sum_{\beta \in \Omega} I(x+\alpha, y+\beta) * G_{(f,\theta)}(\alpha, \beta) \quad (3)$$

Table 3: Gabor features extracted from Gabor filters.

Features	Expression
Mean	$F_{f,\theta}^{(1)} = \frac{\sum_{x=1}^{M} \sum_{y=1}^{N} I_{(G_{f,\theta})}(x,y)}{M*N}$
Standard deviation	$F_{f,\theta}^{(2)} = \frac{\sum_{x=1}^{M} \sum_{y=1}^{N} [I_{(G_{f,\theta})}(x,y) - F_{f,\theta}^{(1)}]^2}{M*N}$

In Table 3, f, θ, and σ are the spatial frequency, orientation and space constant of the Gabor envelope.

$$G_{(f,\theta)}(\alpha,\beta) = \sqrt{[G_{e(f,\theta)}(\alpha,\beta)]^2 + [G_{o(f,\theta)}(\alpha,\beta)]^2}$$

where $G_{e(f,\theta)}$ and $G_{o(f,\theta)}$ denote the spatial frequency responses of the even and odd symmetric Gabor filters.

$$G_{e(f,\theta)} = \frac{H_{1(f,\theta)}(\alpha,\beta) + H_{2(f,\theta)}(\alpha,\beta)}{2}$$

$$G_{o(f,\theta)} = \frac{H_{1(f,\theta)}(\alpha,\beta) + H_{2(f,\theta)}(\alpha,\beta)}{2j}$$

where

$$H_{1(f,\theta)}(\alpha,\beta) = \exp\{-2\pi\sigma^2[(\alpha - f\cos\theta)^2 + (\alpha - f\sin\theta)^2]\}$$

$$H_{2(f,\theta)}(\alpha,\beta) = \exp\{-2\pi\sigma^2[(\alpha + f\cos\theta)^2 + (\alpha - f\sin\theta)^2]\}$$

$$j^2 = -1$$

4. EVALUATION AND RESULTS

Historical document images are primarily characterized by a strong heterogeneity due differences in layout, typography, illustration style, and complex structures without clear layout rules. Moreover, degradation properties (yellow pages, ink stains, back-to-front interference, etc.), and scanning defects (defects of curvature, light, etc.) are added which complicate more and more any characterization or segmentation of the document image and make its processing a non-trivial task. Thus, for a first evaluation and comparison of the extracted features, we generate a new database of simplified historical document images. By adding few superpositions,

i.e. white rectangular regions on several component parts or elements of the document content, the document layout was simplified and complexity reduced. We selected four historical document images, containing graphics and text with two and three different fonts, from our corpus. From the four selected historical document images, we build a set of 25 simplified historical document images (see Figure 2).

Firstly, we computed the texture features of 25 simplified historical document images as described in Section 3. Then, we applied an unsupervised clustering step on the extracted textural attributes from our 25 simplified historical documents in order to validate and compare their robustness and pertinence. Afterwards, the clustering was performed by using an adapted HAC and by setting the maximum number of clusters equal to the one defined in our ground truth. The process of HAC consists in merging successively pairs of existing clusters, where at each cluster grouping step, the choice of cluster pairs depends on the smallest distance, i.e. clusters are grouped when the intra-cluster inertia is minimal. This linkage between clusters was performed by using the Ward criterion [43] along with the weighted Euclidean distance [20]. In the perspective of historical collection valorization and in the context of an unsupervised clustering, the authors of [29] present interesting classification results obtained by performing HAC on stroke features of initial letters. Furthermore, the authors of [19] claim that the distance computed by the Ward method has given the best results in the experiment of the HAC method.

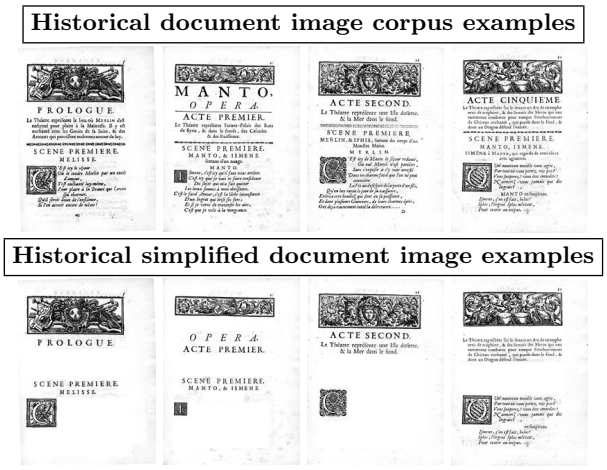

Figure 2: Historical document image corpus and simplified document image examples.

Figure 3 shows the real coherent separating power of the extracted texture features in the context of historical digitized document images with little *a priori* knowledge. The extracted textural descriptors are providing similar and satisfying results in distinguishing the textual regions from the graphical ones when comparing visually the segmentation results for the three selected approaches: autocorrelation (see Figures 3(a) and 3(b)), co-occurrence (see Figures 3(e) and 3(f)), and Gabor (see Figures 3(i) and 3(j)) approaches. While, in the case of document image containing only textual regions with distinct fonts, we demonstrate that Gabor features are the best in segregating different fonts, i.e. we distinguish two fonts: the italic (green) and uppercase (blue) in Figure 3(k) and also three fonts: the normal (green), italic (red) and uppercase (blue) fonts in Figure 3(l). This may be confirmed by the frequent use of Gabor descriptors principally to identify script and language and for character and font recognition in the literature [45, 5], because Gabor features are known to be sensitive to the stroke width.

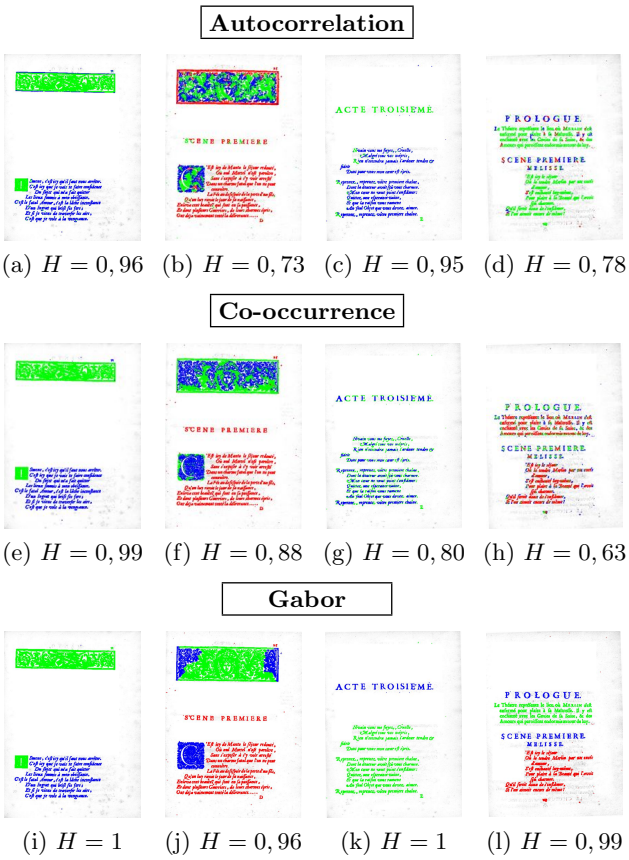

Figure 3: Result examples of texture feature evaluation on simplified historical document images. For the same page, each cluster (represented by a given color) represents a similar or homogeneous region. Because the process is unsupervised, the color attributed to text or graphics may differ from one document image to another.

Indeed, a visual evaluation of results is inherently very subjective. Hence, we need to assess the effectiveness using appropriate quantitative metrics. For this reason, we compute one internal, unsupervised clustering accuracy metric (average silhouette width index [37]) and three external, supervised, clustering evaluation indices (Jaccard coefficient [39], Fowlkes-Mallows index [8], and homogeneity measure [27]). The higher are these measures, the better are the results. The average silhouette width (SW) measures the level of compactness and separation by analyzing the distribution of the observations into clusters. The Jaccard (J) and Fowlkes-Mallows (FM) indexes compare the distributions of the observations in the clustering results and in the ground truth by measuring the probability that a pair of observations are classified together in both the clustering solu-

tion and the ground truth class. The homogeneity measure (H) evaluates the homogeneity accuracy in terms of matching regions between the ground truth and result regions, i.e. spatial overlaps of the label correspondence between the ground truth and the result regions.

The results of the internal and external clustering accuracy measures for the set of 25 simplified historical document images are presented in Table 4. Silhouette width, Jaccard coefficient, and Fowlkes-Mallows index show that the lowest values are obtained for the co-occurrence descriptors, while the lowest value of homogeneity metric is obtained for the autocorrelation attributes. This slight variability of the low ranking of clustering performance by the different clustering measures can be explained by the specificity of each clustering accuracy measure. The best clustering results are obtained by the Gabor features for almost all clustering evaluation metrics with an average silhouette width of 0.33, 95% of mean homogeneity accuracy, 72% of mean Jaccard accuracy, and 83% of mean Fowlkes-Mallows accuracy. We conclude that visual results (see Figure 3) are exactly in concordance with different quantitative metrics.

Although the Gabor features perform much superior to the autocorrelation and the co-occurrence features for font segmentation, globally the three kinds of extracted textural features are efficient to distinguish textual regions from graphical ones. Using the multi-channel Gabor approach, we perform better than the autocorrelation and the co-occurrence approaches in segmenting document image containing only textual regions with distinct fonts. However, we cannot overcome the disadvantage of the increase of the feature dimension, computation, and time cost. The time required to process a page (1982*2750 pixels) using the autocorrelation and the Gabor approaches is six and nine minutes respectively while using the co-occurrence descriptors is reduced to only one minute.

Table 4: Evaluation of the extracted features by internal and external clustering accuracy measures on historical simplified document images: silhouette width (SW), homogeneity metric (H), Jaccard coefficient (J), and Fowlkes-Mallows index (FM). $\mu(.)$ and $\sigma(.)$ are respectively the mean value and the standard deviation value of $(.)$.

	μ_a	σ_a	μ_c	σ_c	μ_g	σ_g
SW	0,24	0,21	0,23	0,10	0,33	0,06
H	0,83	0,07	0,84	0,12	0,95	0,04
J	0,58	0,15	0,56	0,17	0,72	0,20
FM	0,74	0,11	0,71	0,13	0,83	0,12

5. CONCLUSIONS AND FURTHER WORK

This article presents a framework for the texture-based segmentation of historical digitized book content with little *a priori* knowledge in order to evaluate three approaches based on different categories of texture features: the autocorrelation function, the co-occurrence matrices and the Gabor filters. Our framework consists in extracting automatically texture features by adopting a multiresolution approach and using a non-parametric unsupervised clustering technique in order to find the homogeneous regions defined by similar texture indices from the whole book instead of processing each page individually. In order to evaluate the different extracted texture features, a comparative study is conducted after extracting and clustering several texture cues from simplified historical document images. Our study demonstrates that the Gabor features show the best performance for several internal and external clustering accuracies. Furthermore, the Gabor approach is a good choice for segmenting document images containing only textual regions with distinct fonts. However, when the numerical complexity is taken into account, the co-occurrence approach would be the better choice.

Although some interesting conclusions are drawn in this paper, it is just a first evaluation and performance comparison between the autocorrelation, the co-occurrence, and the Gabor approaches on several historical simplified document images. In future research, we need to carry out all steps of our framework on a large corpus of historical books and compare clustering performance with selectivity to the book content, i.e. text vs. graphics and also book characteristics, such manuscript vs. printed, in order to summarize the pros and cons of each texture-based method for each book category. Further work needs to be done in combining various texture descriptors in order to construct an optimal texture feature set and to provide a qualitative measure of which features are the most appropriate for this task.

6. REFERENCES

[1] A. C. Bovik, M. Clark, and W. S. Geisler. Multichannel Texture Analysis Using Localized Spatial Filters. *PAMI*, pages 55–73, 1990.

[2] 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*. PhD thesis, Institut National des Sciences Appliquées de Lyon, Lyon, France, 1994.

[3] A. Busch, W. W. Boles, and S. Sridharan. Texture for Script Identification. *PAMI*, pages 1720–1732, 2005.

[4] C. H. Chen, L. F. Pau, and P. Wang. *Texture analysis in The Handbook of Pattern Recognition and Computer Vision*. World Scientific, 1998.

[5] J. Chen, H. Cao, R. Prasad, A. Bhardwaj, and P. Natarajan. Gabor Features for Offline Arabic Handwriting Recognition. In *DAS*, pages 53–58. IEEE, 2010.

[6] K. Ding, Z. Liu, L. Jin, and X. Zhu. A comparative study of Gabor feature and gradient feature for handwritten chinese character recognition. In *ICWAPR*, pages 1182–1186. IEEE, 2007.

[7] V. Eglin, S. Bres, and C. Rivero. Hermite and Gabor transforms for noise reduction and handwriting classification in ancient manuscripts. *IJDAR*, pages 101–122, 2007.

[8] E. B. Fowlkes and C. L. Mallows. A Method for Comparing Two Hierarchical Clusterings. *JASA*, pages 553–569, 1983.

[9] D. Gabor. Theory of communication. Part 1: The analysis of information. *Journal of the Institution of Electrical Engineers - Part III: Radio and*

Communication Engineering, pages 429–441, 1946.

[10] R. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. In *SMC*, pages 610–621. IEEE, 1973.

[11] A. K. Jain and S. Bhattacharjee. Text Segmentation Using Gabor Filters for Automatic Document Processing. *MVA*, pages 169–184, 1992.

[12] A. K. Jain and Y. Zhong. Page segmentation using texture analysis. *PR*, pages 743–770, 1996.

[13] N. Journet, J. Ramel, R. Mullot, and V. Eglin. Document image characterization using a multiresolution analysis of the texture: application to old documents. *IJDAR*, pages 9–18, 2008.

[14] B. Julesz. Visual pattern discrimination. In *Information Theory*, pages 84–92. IEEE, 1962.

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

[16] D. J. Ketchen and C. L. Shook. The application of cluster analysis in Strategic Management Research: An analysis and critique. *Strategic Management Journal*, pages 441–458, 1996.

[17] S. Khedekar, V. Ramanaprasad, S. Setlur, and V. Govindaraju. Text - Image Separation in Devanagari Documents. In *ICDAR*, pages 1265–1269. IEEE, 2003.

[18] D. E. Knuth. *The art of computer programming, volume 3: (2nd ed.) sorting and searching*. Addison Wesley Longman Publishing Co, 1997.

[19] H. P. Lai, M. Visani, A. Boucher, and J. M. Ogier. An experimental comparison of clustering methods for content-based indexing of large image databases. *PAA*, pages 345–366, 2012.

[20] F. Lalys, C. Haegelen, M. Mehri, S. Drapier, M. Vérin, and P. Jannin. Anatomo-clinical atlases correlate clinical data and electrode contact coordinates : Application to subthalamic deep brain stimulation. *Journal of Neuroscience*, pages 297–307, 2013.

[21] G. N. Lance and W. T. Williams. A general theory of classificatory sorting strategies 1. Hierarchical systems. *The Computer Journal*, pages 373–380, 1967.

[22] M. Lin, J. Tapamo, and B. Ndovie. A Texture-based Method for Document Segmentation and Classification. *South African Computer Journal*, pages 49–56, 2006.

[23] C. L. Liu, M. Koga, and H. Fujisawa. Gabor Feature Extraction for Character Recognition: Comparison with Gradient Feature. In *ICDAR*, pages 121–125. IEEE, 2005.

[24] H. Ma and D. Doermann. Gabor Filter Based Multi-class Classifier for Scanned Document Images. In *ICDAR*, pages 968–972. IEEE, 2003.

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

[26] P. Mahalanobis. On the generalised distance in statistics. In *Proceedings of the National Institute of Sciences of India*, pages 49–55. NISI, 1936.

[27] M. Mehri, P. Gomez-Krämer, P. Héroux, and R. Mullot. Old document image segmentation using the autocorrelation function and multiresolution analysis. In *DRR*. SPIE, 2013.

[28] A. K. Mikkilineni, P. J. Chiang, G. N. Ali, G. T. C. Chiu, J. P. Allebach, and E. J. D. III. Printer identification based on graylevel co-occurrence features for security and forensic applications. In *SSWMC*, pages 430–440. SPIE, 2005.

[29] G. Nguyen, M. Coustaty, and J. Ogier. Stroke feature extraction for lettrine indexing. In *IPTA*, pages 355–360. IEEE, 2010.

[30] N. Otsu. A threshold selection method from gray-level histograms. In *Systems, Man, and Cybernetics*, pages 62–66. IEEE, 1979.

[31] A. Ouji, Y. Leydier, and F. LeBourgeois. Chromatic / Achromatic Separation in Noisy Document Images. In *ICDAR*, pages 167–171. IEEE, 2011.

[32] J. S. Payne, T. J. Stonham, and D. Patel. Document Segmentation Using Texture Analysis. In *ICPR*, pages 380–382. IEEE, 1994.

[33] J. S. Payne, T. J. Stonham, and D. Patel. Document Segmentation Using Texture Analysis. In *ICPR*, pages 380–382. IEEE, 1994.

[34] G. Peake and T. Tan. Script and Language Identification from Document Images. In *DIA*, pages 10–17. IEEE, 1997.

[35] M. Petrou and P. G. Sevilla. *Image Processing : Dealing with texture*. John Wiley & Sons, 2006.

[36] Y. Qiao, Z. Lu, C. Song, and S. Sun. Document image segmentation using Gabor wavelet and kernel-based methods. In *ISSCAA*, pages 450–455. IEEE, 2006.

[37] P. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, pages 53–65, 1987.

[38] H. E. S. Said, T. N. Tan, and K. D. Baker. Personal identification based on handwriting. *PR*, pages 149–160, 2000.

[39] P. C. Saxena and K. Navaneetham. The effect of cluster size, dimensionality, and number of clusters on recovery of true cluster structure through Chernoff-type faces. *Journal of the Royal Statistical Society, The Statistician*, pages 415–425, 1991.

[40] F. Shahabi and M. Rahmati. A New Method for Writer Identification of Handwritten Farsi Documents. In *ICDAR*, pages 426–430. IEEE, 2009.

[41] L. Shijian and C. L. Tan. Script and Language Identification in Noisy and Degraded Document Images. *PAMI*, pages 14–24, 2008.

[42] T. Simpson, J. Armstrong, and A. Jarman. Merged consensus clustering to assess and improve class discovery with microarray data. *Boston Medical Center Bioinformatics*, pages 1471–1482, 2010.

[43] J. Ward. Hierarchical Grouping to Optimize an Objective Function. *JASA*, pages 236–244, 1963.

[44] K. Wong, R. Casey, and F. Wahl. Document Analysis System. *IBM Journal of Research and Development*, pages 647–656, 1982.

[45] Y. Zhu, T. Tan, and Y. Wang. Font Recognition Based on Global Texture Analysis. *PAMI*, pages 1192–1200, 2001.