HAL
open science

A Pixel Labeling Framework for Comparing Texture Features: Application to Digitized Ancient Books

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

To cite this version:

Maroua Mehri, Petra Gomez-Krämer, Pierre Héroux, Alain Boucher, Rémy Mullot. A Pixel Labeling Framework for Comparing Texture Features: Application to Digitized Ancient Books. International Conference on Pattern Recognition Applications and Methods, Mar 2014, Angers, France. pp.553-560, 10.5220/0004804705530560 . hal-01119154

HAL Id: hal-01119154

https://hal.science/hal-01119154

Submitted on 23 Feb 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Pixel Labeling Framework for Comparing Texture Features: Application to Digitized Ancient Books

Maroua Mehri[1,2], Petra Gomez-Krämer[1], Pierre Héroux[2], Alain Boucher[1] and Rémy Mullot[1]
[1]*L3i, University of La Rochelle, Avenue Michel Crépeau, 17042 La Rochelle, France*
[2]*LITIS, University of Rouen, Avenue de l'Université, 76800 Saint-Etienne-du-Rouvray, France*
{*maroua.mehri, petra.gomez, alain.boucher, remy.mullot*}*@univ-lr.fr, pierre.heroux@univ-rouen.fr*

Keywords: Ancient digitized books, Pixel labeling, Texture, Multiresolution, Consensus clustering, Clustering and classification accuracy metrics.

Abstract: In this article, a complete framework for the comparative analysis of texture features is presented and evaluated for the segmentation and characterization of ancient book pages. Firstly, the content of an entire book is characterized by extracting the texture attributes of each page. The extraction of the texture features is based on a multiresolution analysis. Secondly, a clustering approach is performed in order to classify automatically the homogeneous regions of book pages. Namely, two approaches are compared based on two different statistical categories of texture features, autocorrelation and co-occurrence, in order to segment the content of ancient book pages and find homogeneous regions with little *a priori* knowledge. By computing several clustering and classification accuracy measures, the results of the comparison show the effectiveness of the proposed framework. Tests on different book contents (text *vs.* graphics, manuscript *vs.* printed) show that those texture features are more suitable to distinguish textual regions from graphical ones, than to distinguish text fonts.

1 INTRODUCTION

In order to provide a wider access to libraries collections which need to be protected from too frequent handling, numerous projects have been established. For instance, Google has conducted large digitization programs "Google Books Library Project" of cultural heritage with the help of several libraries. Therefore, automatic processing of ancient digitized documents has undergone tremendous growth over the last few years. Thereby, reliable ancient document interpretation systems have been the topics of major interest of many libraries and historians and the prime issue of research in document analysis community. There has been a great challenge in the refinement of the well-known approaches based on strong *a priori* knowledge. Many methods have been presented in the literature (Mao et al., 2003; Mullot, 2006) to perform this task. However, such algorithms rely on *a priori* knowledge in order to properly segment and characterize the document image content.

The problematic of this paper concerns the segmentation and characterization of ancient digitized book content with little *a priori* knowledge. Specifically, in the context of the DIGIDOC project (Document Image diGitisation with Interactive DescriptiOn Capability)[1], we aim to propose new ways of interacting with scanners as well as new tools for analyzing documents during the whole acquisition process from scanning the document to knowledge representation and management of the ancient digitized document content. LeBourgeois *et al.* (Bourgeois et al., 2004) highlight an essential necessity to conceive "intelligent" digitizers which can limit manual intervention and realize easy and high quality digitization of image documents. Thus, the main goal of our work is to develop a mapping between the scanned image and its content and subsequently use through a set of descriptors extracted and computed on it. Those descriptors will help representing 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. In such conditions, a texture analysis technique is a logical choice for solving our problem as it gives several features characterizing the textural properties of a region without using information on the document structure such as the document model and the typographical parameters (Journet et al., 2008).

[1]The DIGIDOC project is referenced under ANR-10-CORD-0020.

Among the most widely used texture feature extraction and analysis methods are those derived from statistical, geometrical, model-based, and signal processing primitives (Chen et al., 1998). Several statistical texture-based segmentation methods have been presented, *e.g.* the autocorrelation function (Petrou and Sevilla, 2006) and GLCM (Grey Level Co-occurrence Matrix) (Haralick et al., 1973; Busch et al., 2005). The extracted texture features are mainly investigated and analyzed separately in independent experiments for document analysis (Journet et al., 2008). Some works deal with the whole ancient document image (Journet et al., 2008) and others are applied to graphic images such as drop caps (Uttama et al., 2006; Coustaty et al., 2011). There have been few comparative studies on gradient, multiple channel Gabor filters, and co-occurrence features (Payne et al., 1994; Said et al., 2000; Liu et al., 2005; Ding et al., 2007; Zhu et al., 2001) for document segmentation, character recognition, and script and language identification.

In (Mehri et al., 2013a), a texture-based framework for segmentation of digitized historical books is proposed. Nevertheless, the authors present only some preliminary study of texture feature comparison (autocorrelation, co-occurrence, and Gabor) using only texture feature extraction and pixel classification tasks on simplified document images. Moreover, in (Mehri et al., 2013c), the authors present a pixel labeling approach for historical digitized books based on two non-parametric tools: the autocorrelation function and multiresolution analysis. We propose in this paper a complete book pixel labeling framework for comparing texture features based on (Mehri et al., 2013a; Mehri et al., 2013c) including the automatic estimation of the number of homogeneous and similar content regions of book pages and the pixel labeling steps. The framework is evaluated on a large corpus of historical books using the autocorrelation and GLCM texture features in order to find the homogeneous regions defined by similar texture indices from the whole book instead of processing each page individually. We compare clustering and classification performance with selectivity to the book content, *i.e.* text *vs.* graphics and also book characteristics, such as manuscript *vs.* printed, in order to summarize the pros and cons of each texture-based method for each book category.

The remainder of this paper is structured as follows. In Section 2, we describe the proposed framework of pixel labeling and characterization of the content of an entire book (Mehri et al., 2013a). The texture primitives, autocorrelation and co-occurrence, chosen for the validation and evaluation of the framework are detailed in Section 3. In Section 4, we propose a comparative analysis of the performance of the chosen texture features. Our conclusions and future work are presented in Section 5.

2 FRAMEWORK

A texture-based framework for segmentation of digitized historical books is proposed in (Mehri et al., 2013a). It is depicted in Figure 1. The proposed framework is pixel-based and does not require *a priori* knowledge on the document structure, neither about the document model, nor about the typographical parameters. Thus, it is adapted to all kinds of books. It is independent of the document layout, the typeface, the font size, the page orientation, the digitizing resolution, etc. as a texture analysis technique is introduced. The goal of the framework is to determine regions or groups of pixels which share similar properties or characteristics. These characteristics may be based on the pixel location, their surroundings, color, intensity or texture. In this work, we focus only on textural characteristics. The use of an texture-based approach has been shown to work effectively with skewed and degraded images (Journet et al., 2008).

By selecting randomly a number of foreground pixels from a few pages of the same book, their textural features are computed firstly. Then, an estimated number of homogeneous and similar content regions is computed in the analyzed book by applying the Consensus Clustering method (CC) (Simpson et al., 2010) on the extracted texture attributes (block 2 in Figure 1). Then, for each analyzed book page its texture features are extracted which are then used in a clustering approach by taking into consideration the estimation of the number of clusters given before by the CC method (block 1 in Figure 1) in order to automatically label content pixels with the same cluster identifier with respect to the book content.

Figure 1 illustrates diagrammatically the four main tasks of our proposed framework. The block 2 on Figure 1 ensures the estimation of the number of homogeneous and similar content regions from the extracted textural features on the whole analyzed book. The block 1 on Figure 1 integrates an unsupervised task enabling to automatically label content pixels with the same cluster identifier regarding to the book content in order to determine and characterize the homogeneous regions in the digitized book (block 3 on Figure 1).

The proposed framework is described by the following four tasks:
1) The texture feature extraction (Section 2.1),
2) The estimation of the number of homogeneous and similar content regions (Section 2.2),
3) The pixel clustering (Section 2.3),

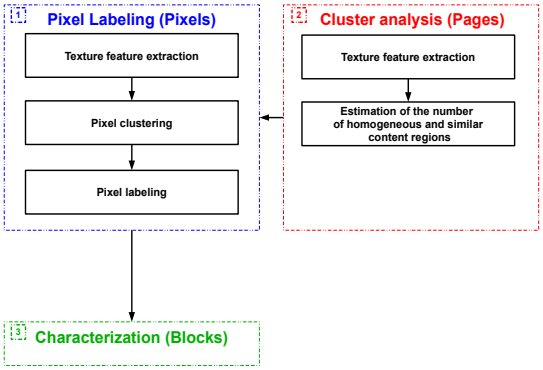

Figure 1: Stages of our pixel labeling framework of historical digitized book content.

4) The pixel labeling (Section 2.4).

2.1 Texture feature extraction

Texture feature extraction aims at representing the document image content by a set of descriptive features computed or extracted on it. Several studies (Uttama et al., 2006; Journet et al., 2008) proposed to characterize and index images of ancient documents by their content by exploring textural analysis based on statistical and spectral properties of texture. The authors of (Uttama et al., 2006) introduce a segmentation method of drop caps based on a combination of different texture analysis approaches such as the GLCM (Haralick et al., 1973) and the autocorrelation function (Petrou and Sevilla, 2006). In (Journet et al., 2008), the authors propose an extraction algorithm of texture features that is devoted to the analysis of historical 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 characterizing the content of images without assumption on the image structure or properties.

In this paper, two statistical primitives are investigated: the autocorrelation function (Petrou and Sevilla, 2006) and GLCM (Haralick et al., 1973). The extraction of textural descriptors is performed on foreground pixels of gray-level images. The texture features are performed at various sizes of analysis windows in order to adopt a multiscale approach. A multiscale approach has been proposed by selecting concentric analysis windows with distinct sizes in order to characterize the images (Journet et al., 2008; Kricha and Amara, 2011). 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 is performed from only the foreground pixels of the gray-level document images at four different sizes of sliding windows: (16×16), (32×32), (64×64), and (128×128). In order to avoid side effects, a border replication step is introduced allowing computing texture features on the whole image.

2.2 Estimation of the number of homogeneous and similar content regions

Since the goal of the proposed framework is to find homogeneous regions defined by similar texture features, a clustering approach is required in order to partition the analyzed page into regions which have similar properties with respect to the extracted features. However, the number of clusters must be known *a priori* especially for the conventional clustering techniques (Lance and Williams, 1967; MacQueen, 1967; Kaufman and Rousseeuw, 1990).

Previous work has identified a number of approaches (Ketchen and Shook, 1996) for determining the correct number of clusters in a dataset. Simpson *et al.* (Simpson et al., 2010) have recently proposed an effective method, known as the Consensus Clustering (CC), to estimate the optimal number of clusters in biological data. The idea of CC consists in performing a consensus matrix by iterating multiple runs of clustering algorithms with random and re-sampled clustering options (Monti et al., 2003). Therefore, the consensus matrix analyzes the consistency of clustering results from five different clustering algorithms: AGglomerative NESting (AGNES), DIvisive ANAlysis clustering (DIANA), Partitioning Around Medoids (PAM), k-means clustering (k-means), and Hierarchical Ascendant Classification (HAC). Therefore, by weighting the different clustering methods in order to mitigate extremes in consensus values that can be created by the sensitivity of some algorithms, a merge consensus matrix is performed.

Thus, the estimation of the number of homogeneous and similar content regions is performed by using the merged CC approach. Although, to perform this task on all pages of the analyzed book is not possible in the case of huge amount of document images, because it needs a high computational time and memory. Hence, the number of clusters is estimated by using the merged CC method in a set of randomly selected foreground pixels from few random selected pages of a book. Due to the memory constraints and the high computational time of the merged CC method, a set of 1000 randomly selected pixels of 10 pages selected randomly from the same book is firstly proposed.

2.3 Pixel clustering

Since the optimal number of clusters k_{opt} is estimated, a clustering method is required in order to characterize the content of an entire book and find the k_{opt} homogeneous regions defined by similar texture indices on the whole book. The work presented in (Nguyen et al., 2010) has shown interesting results in classifying the strokes of initial letters by using the HAC algorithm. Due to the high requirement of a too large amount of memory to perform the merged CC method on all pixels from each analyzed document image, the HAC algorithm is performed on the extracted textural features without taking into account the spatial coordinates and by setting the maximum number of clusters to the estimated one, k_{opt}, with the merged CC method. According to a hierarchical structure grouping of clusters based on the criteria of the minimum increased intra-clusters inertia, the 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.

2.4 Pixel labeling

This phase deals with labeling clusters or group of pixels with respect to the results of the pixel classification phase. The idea of this task is to assign a label to each cluster of pixels which share similar textural characteristics with respect to the obtained cluster of the selected samples of the analyzed book. Thus, the pixel labeling aims at determining and assigning the same cluster identifier for each similar cluster extracted from the digitized book.

Journet *et al.* (Journet et al., 2008) propose to perform the clustering stage by using CLustering LARge Applications (CLARA) (Kaufman and Rousseeuw, 1990), which is known to be adapted to large scale databases, in the extracted texture features computed from six pages of the same book. Then, if two pixels of two different document images have the same cluster label, they belong to the same class. However, this technique is characterized by a high computation time and memory complexity.

In (Mehri et al., 2013b), a clustering approach without taking into account the characterization step, *i.e.* the pixel labeling step. In (Mehri et al., 2013a), an unsupervised task is proposed enabling to automatically label content pixels with the same cluster identifier regarding to the book content. For the same book, each cluster (represented by a given color) represents a similar or homogeneous region. Thus, by applying HAC, the homogeneous regions are defined by similar texture indices. Then, the Nearest Neighbor Search algorithm (NNS) (Knuth, 1997) is performed in order 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 features vector to the cluster of the selected samples of a book, *i.e.* by selecting the minimum distance.

Finally, the NNS (Knuth, 1997) with the Mahalanobis distance (Mahalanobis, 1936) is applied in order to assign the same label for each similar cluster extracted from the digitized book. The 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 cluster to the one of the selected samples of a book. The Mahalanobis distance takes into account the dataset correlations and is particularly suited to arbitrarily shaped clusters, *i.e.* the minimum of the intra-cluster and inter-cluster distance is taken into consideration.

3 COMPARISON OF AUTOCORRELATION AND CO-OCCURRENCE FEATURES

In order to validate and evaluate the proposed framework of segmentation and characterization of ancient digitized books, two texture primitives are computed: the autocorrelation and co-occurrence features that are outlined below. The different texture descriptors used in this paper are reported in (Mehri et al., 2013a). Autocorrelation and co-occurrence features are extracted for several reasons: Firstly, we have made a comparative study about choosing the texture feature category, which ensures the best and constructive trade-off between best performance and lowest computation time. Secondly, the pertinence of the segmentation experiments (Journet et al., 2008) that are based on the autocorrelation function applied on document image leads us to work with the autocorrelation features in order to reach the objectives of determining homogeneous regions from the analyzed document without hypothesis on the document structure, neither on the document model nor the typographical parameters. Finally, the extraction of the these two texture features needs less parameter settings compared to the descriptors computed from Gabor filters (Jain and Zhong, 1996) for instance. A 2D Gabor filter is a linear selective band-pass filter, dependent on two parameters: spatial frequency and orientation. Moreover, Tamura features depend on the value of coarseness (Howarth

and Ruger, 2004). Indeed, without hypothesis on the document structure, neither on the document layout nor the typographical parameters, the choice of appropriate thresholds and parameters is a very difficult task.

3.1 Autocorrelation features

In this paper, we evaluate firstly the autocorrelation descriptors. The autocorrelation function (Petrou and Sevilla, 2006) is used to determine periodic patterns and can characterize similarity patterns through a number of extracted autocorrelation features. A number of autocorrelation features have been proposed in (Journet et al., 2008; Ouji et al., 2011; Mehri et al., 2013b; Mehri et al., 2013a) for segmenting ancient and contemporary documents images. In (Journet et al., 2008), the authors use the directional rose (Bres, 1994), a derivative of the autocorrelation function. The directional rose identifies significant texture orientations in the analyzed block image.

The extracted autocorrelation descriptors provide interesting information on the principal texture orientations and periodicities. Five autocorrelation features are computed, which have been reported in (Mehri et al., 2013b): 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 (Journet et al., 2008; Ouji et al., 2011; Mehri et al., 2013b).

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, to every selected foreground pixel from the digitized document image is assigned a vector which corresponds to the extracted autocorrelation indices.

3.2 Co-occurrence features

The co-occurrence attributes (Haralick et al., 1973) are the second features tested in this paper, extracted from the GLCM. The GLCM 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. By applying multi-distance and multi-direction approaches, a large number of co-occurrence descriptors can be extracted. Fourteen textural features extracted of the GLCM have been initially introduced by (Haralick et al., 1973) for texture discrimination of natural and satellite images. A number of co-occurrence feature extraction and analysis methods (Mikkilineni et al., 2005; Lin et al., 2006; Payne et al., 1994; Peake and Tan, 1997; Busch et al., 2005) have been proposed in order to segment and classify the content of document images, and to identify script and language from document images. Briefly, the GLCM matrices are obtained for a small range of distance values $d = 1,2$ and typically for the directions $\theta = \{0°, 45°, 90°, 135°\}$ (Busch et al., 2005).

Six co-occurrence features are extracted from the GLCM matrices: 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 $\{d = 1,2\}$ (Mikkilineni et al., 2005; Busch et al., 2005). 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 (Lin et al., 2006).

Extracting these co-occurrence indices using a sliding window gives a total of 56 numeric values (14 co-occurrence indices $\times$ 4 sliding window sizes for multiresolution). Therefore, to every selected foreground pixel from the digitized document image a vector is assigned which corresponds to the extracted co-occurrence indices.

4 EVALUATION AND RESULTS

In the experiment, 316 pages extracted from 13 different books are considered. Our corpus is divided into two categories: 7 printed monographs and 6 manuscripts that encompass six centuries (1200-1900) of French history. For each category, we 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. Evaluation of segmentation and region classification requires a ground truth which is performed by defining manually our ground truth with the Ground-truthing Environment for Document Images (GEDI)[2], a public domain document image annotation tool. Our corpus is composed of grayscale pages which were digitized with 300/400 dpi. The time required to process a page (1982*2750 pixels) using the autocorrelation approach is six minutes while using the co-occurrence descriptors is reduced to only one minute.

[2] `http://gedigroundtruth.sourceforge.net/`

Figure 3 shows the real coherent separating power of the extracted texture features in the context of historical digitized book with little *a priori* knowledge. For the same book, 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 book to another. The proposed framework (*cf.* Figure 1) is providing satisfying results particularly in distinguishing the textual regions from the graphical ones when comparing visually the segmentation results for both the autocorrelation and co-occurrence approaches. We note that in the case of the manuscript document category (one font and graphics), the segmentation result by the autocorrelation approach (*cf.* Figure 3(a)) is better than those performed by co-occurrence features (*cf.* Figure 3(b)), *i.e.* the graphic regions (blue) 3(a) are more homogeneous. Moreover, we show for the manuscript document category (two fonts and graphics) better results of discrimination text/graphics obtained by the autocorrelation approach (*cf.* Figure 3(e)) (graphic regions (red), textual regions (blue)) than those performed by the co-occurrence approach (*cf.* Figure 3(f)) (graphic regions (blue), textual regions (red)). In the case of the printed document category (one font and graphics), the segmentation results by the autocorrelation approach (*cf.* Figure 3(c)) show that the textural characteristics of each small letter in the beginning of each text line is different from the other text content while the clustering results by the co-occurrence approach consider them as text regions (*cf.* Figure 3(d)). This may be explained by the fact that the autocorrelation features give better information on the major orientation and periodicities of the textual texture than the co-occurrence ones. In general, the results for the two approaches are relatively similar for the documents containing only two fonts when we use the autocorrelation and co-occurrence descriptors, *i.e.* we distinguish two fonts: the normal (green) and uppercase (blue) fonts for the co-occurrence approach in Figure 3(i) and the normal (red) and uppercase (green) fonts for the co-occurrence approach in Figure 3(j). We also note that in Figure 3(g), we descriminate the normal (red) and uppercase (blue) fonts for the documents containing two fonts and graphics with the help of the autocorrelation primitives. Nevertheless, the co-occurrence approach offers important advantages such as a reduced processing time and the ease of implementation.

Indeed, this method of evaluating the performance of a segmentation method is inherently a subjective evaluation and we need to assess the effectiveness using an appropriate quantitative metric. For this reason, we perform firstly three external, or supervised clustering, evaluation indices: Jaccard coefficient (J) (Saxena and Navaneetham, 1991), Fowlkes and Mallows index (FM) (Fowlkes and Mallows, 1983), and purity per block (PPB) (Mehri et al., 2013b). Then, three classification accuracy metrics: precision (P), recall (R), and classification accuracy (CA)(Makhoul et al., 1999) are computed.

In Figure 2, it can be seen that the results obtained by the numerous clustering evaluation measures are coherent with the different classification accuracy ones for the two approaches. We obtain 85%, 70%, 70%, and 79% of mean purity per block accuracy, precision, recall, and classification accuracy respectively for the autocorrelation approach without taking into account the topographical relationships of the selected pixels with respect of 87%, 70%, 79%, and 83% respectively for the co-occurrence approach.

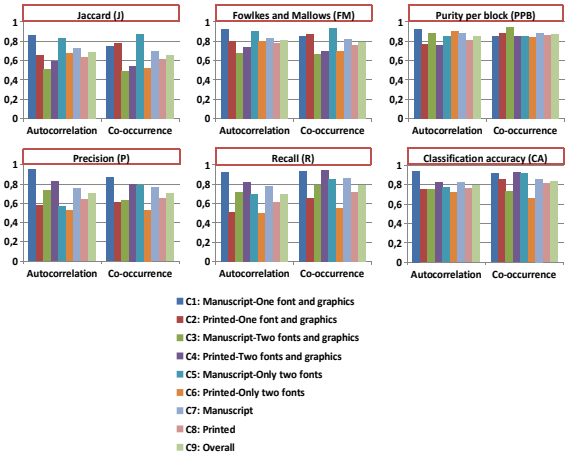

Figure 2: Evaluation of the pixel labeling framework of historical digitized book content by clustering and classification accuracy measures: three clustering accuracy metrics: Jaccard coefficient (J), Fowlkes and Mallows index (FM), and purity per block (PPB), and three classification accuracy measures: precision (P), recall (R), and classification accuracy (CA). The higher the values are, the better the results.

We note also that the best segmentation results are obtained by extracting the autocorrelation features for manuscript documents containing one font and graphics. One assumption can be that the manuscript documents contain graphic regions that are more compact and homogeneous than the printed documents and the autocorrelation features are more suitable to discriminate compact graphical regions from textual ones. Whereas, the analysis of quantitative performance noted by using the co-occurrence approach provides different results according to the computed accuracy. Nevertheless, the best performance is obtained for documents containing graphics and text by extracting the co-occurrence features. Concerning the

slight variability of the ranking of clustering performance by using the clustering and classification accuracy measures can be explained by the specificity of each clustering accuracy measure. We see that the best classification result of the precision metric (P) is obtained for the manuscript document category (one font and graphics) by using both the autocorrelation and co-occurrence approaches. Precision (P), recall (R), and classification accuracy (CA) show that the lowest values are obtained for printed documents (only two fonts) by using both the autocorrelation and co-occurrence approaches. Therefore, the obtained qualitative results are confirmed, *i.e.* the extracted autocorrelation and co-occurrence descriptors are more suitable to distinguish the textual regions from the graphical ones. We conclude that the overall results are quite satisfying since we do not integrate the topographical or spatial relationships.

5 CONCLUSIONS AND FURTHER WORK

This article presents and evaluates a framework for the comparative analysis of texture features for the segmentation and characterization of ancient book pages. The proposed framework is used to extract and compare automatically texture features. Then, a non-parametric clustering method is performed in order to determine the homogeneous regions from different pages of the same book. To validate the proposed method, we evaluate two approaches based on two different statistical categories of texture features: the autocorrelation function and the co-occurrence matrices. Our study demonstrates that the two kinds of texture features give different results according to the book content, *i.e.* text *vs.* graphics and also book characteristics, such as manuscript *vs.* printed. It shows that these texture features are more suitable to distinguish textual regions from graphical ones, than to distinguish text fonts. However, when the numerical complexity is taken into account, the co-occurrence approach would be the better choice. The first aspect of future work will be to test the framework with other texture descriptors such as the Tamura texture features, the local binary patterns, the multiple channel Gabor filters, and the wavelets. Further work also 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 most appropriate for this task.

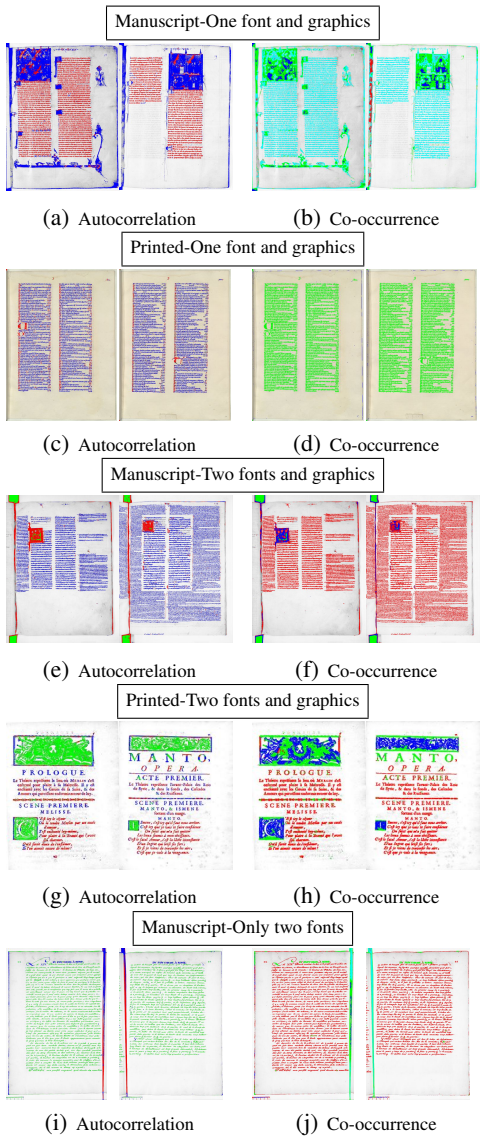

Figure 3: Result examples of the pixel labeling framework of historical digitized book content. For the same book, each cluster (represented by a given color) represents a similar or homogeneous region. Because the process is unsupervised, the colors attributed to text or graphics may differ from one book to another.

REFERENCES

Bourgeois, F. L., Trinh, E., Allier, B., Eglin, V., and Emptoz, H. (2004). Digital libraries and document image analysis. In *DIAL*, pages 2–24.

Bres, S. (1994). *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, INSA, France.

Busch, A., Boles, W. W., and Sridharan, S. (2005). Texture for script identification. *PAMI*, pages 1720–1732.

Chen, C. H., Pau, L. F., and Wang, P. (1998). *Texture analysis in the handbook of pattern recognition and computer vision*. World Scientific, second edition.

Coustaty, M., Pareti, R., Vincent, N., and Ogier, J. M. (2011). Towards historical document indexing: extraction of drop cap letters. *IJDAR*, pages 243–254.

Ding, K., Liu, Z., Jin, L., and Zhu, X. (2007). A comparative study of Gabor feature and gradient feature for handwritten chinese character recognition. In *WAPR*, pages 1182–1186.

Fowlkes, E. B. and Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *JASA*, pages 553–569.

Haralick, R. M., Shanmugam, K., and Dinstein, I. (1973). Textural features for image classification. *SMC*, pages 610–621.

Howarth, P. and Ruger, S. (2004). Evaluation of texture features for content-based image retrieval. *IVR*, pages 326–334.

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

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

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

Ketchen, D. J. and Shook, C. L. (1996). The application of cluster analysis in strategic management research: an analysis and critique. *SMJ*, pages 441–458.

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

Kricha, A. and Amara, N. E. B. (2011). Exploring textural analysis for historical documents characterization. *JC*, pages 24–30.

Lance, G. N. and Williams, W. T. (1967). A general theory of classificatory sorting strategies 1. Hierarchical systems. *CJ*, pages 373–380.

Lin, M., Tapamo, J., and Ndovie, B. (2006). A texture-based method for document segmentation and classification. *SACJ*, pages 49–56.

Liu, C. L., Koga, M., and Fujisawa, H. (2005). Gabor feature extraction for character recognition: comparison with gradient feature. In *ICDAR*, pages 121–125.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *MSP*, pages 281–297.

Mahalanobis, P. (1936). On the generalised distance in statistics. In *NISI*, pages 49–55.

Makhoul, J., Kubala, F., Schwartz, R., and Weischedel, R. (1999). Performance measures for information extraction. In *DARPA*, pages 249–252.

Mao, S., Rosenfeld, A., and Kanungo, T. (2003). Document structure analysis algorithms: a literature survey. In *DRR*, pages 197–207.

Mehri, M., Gomez-Krämer, P., Héroux, P., Boucher, A., and Mullot, R. (2013a). Texture feature evaluation for segmentation of historical document images. In *HIP*, pages 102–109.

Mehri, M., Gomez-Krämer, P., Héroux, P., and Mullot, R. (2013b). Old document image segmentation using the autocorrelation function and multiresolution analysis. In *DRR*.

Mehri, M., Héroux, P., Gomez-Krämer, P., and Mullot, R. (2013c). A pixel labeling approach for historical digitized books. In *ICDAR*, pages 817–821.

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

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus Clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *ML*, pages 91–118.

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

Nguyen, G., Coustaty, M., and Ogier, J. M. (2010). Stroke feature extraction for lettrine indexing. In *IPTA*, pages 355–360.

Ouji, A., Leydier, Y., and Bourgeois, F. L. (2011). Chromatic / achromatic separation in noisy document images. In *ICDAR*, pages 167–171.

Payne, J. S., Stonham, T. J., and Patel, D. (1994). Document segmentation using texture analysis. In *ICPR*, pages 380–382.

Peake, G. and Tan, T. (1997). Script and language identification from document images. In *DIA*, pages 10–17.

Petrou, M. and Sevilla, P. G. (2006). *Image processing: dealing with texture*. John Wiley & Sons.

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

Saxena, P. C. and Navaneetham, K. (1991). The effect of cluster size, dimensionality, and number of clusters on recovery of true cluster structure through Chernoff-type faces. *RSS*, pages 415–425.

Simpson, T., Armstrong, J., and Jarman, A. (2010). Merged consensus clustering to assess and improve class discovery with microarray data. *BMC*, pages 1471–1482.

Uttama, S., Loonis, P., Delalandre, M., and Ogier, J. M. (2006). Segmentation and retrieval of ancient graphic documents. In *GREC*, pages 88–98.

Zhu, Y., Tan, T., and Wang, Y. (2001). Font recognition based on global texture analysis. *PAMI*, pages 1192–1200.