HAL
open science

A Pixel Labeling Approach for Historical Digitized Books

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

To cite this version:

Maroua Mehri, Pierre Héroux, Petra Gomez-Krämer, Alain Boucher, Rémy Mullot. A Pixel Labeling Approach for Historical Digitized Books. International Conference on Document Analysis and Recognition (ICDAR), Aug 2013, Washington, DC, United States. pp.817-821, 10.1109/ICDAR.2013.167 . hal-00927126

HAL Id: hal-00927126

https://hal.science/hal-00927126v1

Submitted on 2 Dec 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 Approach for Historical Digitized Books

Maroua Mehri*†, Pierre Héroux†, Petra Gomez-Krämer*, Alain Boucher*, and Rémy Mullot*
*L3I, University of La Rochelle, Avenue Michel Crépeau, 17042, La Rochelle, France
Emails: {maroua.mehri, petra.gomez, alain.boucher, 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*—In the context of historical collection conservation and worldwide diffusion, this paper presents an automatic approach of historical book page layout segmentation. In this article, we propose to search the homogeneous regions from the content of historical digitized books with little *a priori* knowledge by extracting and analyzing texture features. The novelty of this work lies in the unsupervised clustering of the extracted texture descriptors to find homogeneous regions, i.e. graphic and textual regions, by performing the clustering approach on an entire book instead of processing each page individually. We propose firstly to characterize the content of an entire book by extracting the texture information of each page, as our goal is to compare and index the content of digitized books. The extraction of texture features, computed without any hypothesis on the document structure, is based on two non-parametric tools: the autocorrelation function and multiresolution analysis. Secondly, we perform an unsupervised clustering approach on the extracted features in order to classify automatically the homogeneous regions of book pages. The clustering results are assessed by internal and external accuracy measures. The overall results are quite satisfying. Such analysis would help to construct a computer-aided categorization tool of pages.**

Keywords*—Historical books, texture, autocorrelation, multiresolution, homogeneity, pixel labeling, consensus clustering, clustering accuracy metrics.

I. Introduction

With the recent massive digitization of cultural heritage writings performed worldwide, which aims to preserve the original documents particularly the historical documents and to distribute their content by providing adapted content-based image retrieval tools, special needs are increasing in information retrieval in digital libraries and document layout analysis. Therefore, this work is done in the context of the DIGIDOC project (Document Image diGitisation with Interactive DescriptiOn Capability)[1] which is funded by the ANR (French National Research Agency) and focuses on the acquisition step of the digitized document in order to improve and simplify their subsequent use (archiving, text recognition, document retrieval, etc.). Thus, our objective is to obtain during the production phase of the scanned document image a set of descriptors computed on it. Those descriptors will be dedicated to the acquisition, storage, analysis, and indexing of the scanned documents. In this paper, we present a part of our goal to represent a digitized document by a hierarchy of layout structure and content without any assumption on the page structure, its content and characteristics, and subsequently to define one or more signatures for each page, on the basis of a hierarchical representation of homogeneous blocks and their topology. Assigning a signature to each digitized document will help us to provide a similarity measure between the book pages.

Several studies have been proposed on document image segmentation and characterization tools. Those studies specifically target two important topics: feature extraction methods and feature space structuring methods. The feature extraction methods refer to the assignment of visual signatures to each analyzed image describing its content. Whereas, the feature space structuring methods partition the analyzed image into regions, which have homogeneous characteristics and similar properties with respect to the extracted features. For instance, clustering is a category of the feature space structuring methods. Different approaches based on a strong *a priori* knowledge [1]–[4] have been proposed for the segmentation and characterization of document physical layout content. Due to the pertinence dependence of those methods on the particular layout and document idiosyncrasies [4], the global feature extraction and analysis approaches are more suitable for complex layout documents. In the context of lacking information on the document structure and its characteristics, an alternative of segmentation methods based on texture feature extraction and analysis has been proposed recently for complex document structure analysis [5]. Those methods aim at segmenting its content into homogeneous blocks based on textural descriptors. Among the most widely used texture feature extraction and analysis methods are those derived from statistical, geometrical, model-based, and signal processing primitives [6].

II. Proposed method

We look for representing a book page by a hierarchy of homogeneous blocks defined by similar texture attributes and their topology based on the two hypotheses: Firstly, the textual regions in a digitized document are considered as textured areas while its non-text contents are considered as regions with distinct textures, and secondly text of different fonts is also distinguishable [7]. Thus, we present in this paper an automatic segmentation of historical digitized book content, based on three non-parametric tools: the autocorrelation function, multiresolution analysis and unsupervised clustering. The proposed approach is pixel-based and does not require any *a priori* knowledge on the document structure, neither about

[1]The DIGIDOC project 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

the document model, nor about the typographical parameters. Thus, our method is adapted to all kinds of books. The proposed approach is depicted in Figure 1.

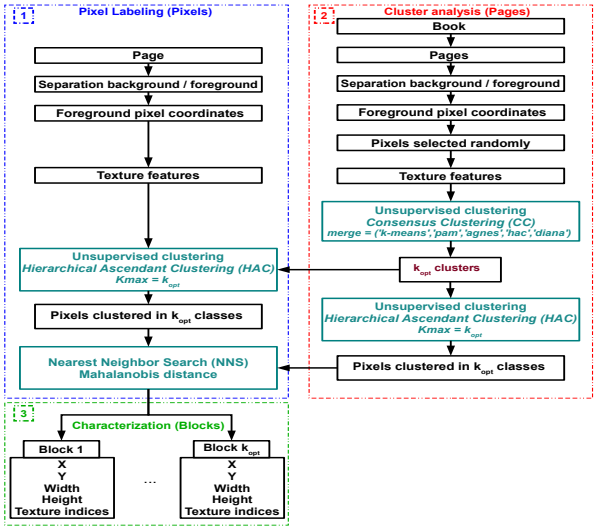


Fig. 1. Presentation of our pixel labeling approach for historical digitized book content.

Our method is composed of two main stages which are described below: Firstly, we select randomly a number of foreground pixels from a few pages of the same book and compute their autocorrelation descriptors in order to estimate the true number of clusters of homogeneous regions in the analyzed book. With the help of the Consensus Clustering method (CC) [8], we determine the exact number of clusters in our samples of foreground pixels (block 2 on Figure 1). Secondly, we compute the autocorrelation features for each page which are then used in an unsupervised clustering approach by taking into account the result of the CC (block 1 on Figure 1) in order to determine and characterize the homogeneous regions in the digitized book (block 3 on Figure 1).

III. Autocorrelation Feature Extraction

The pertinence of the segmentation experiments of historical and contemporary document images presented in [5], [9] that are based on the autocorrelation function [10] leads us to propose a layout segmentation of historical document images based on five autocorrelation indices. Those indices are based on the autocorrelation function and the directional rose [11]. The directional rose is a polar diagram derived from the analysis of the autocorrelation results, which gives good information on the principal orientation of the texture. The first three autocorrelation features have been derived from the directional rose: its main orientation, the intensity of the autocorrelation function for the main orientation, and the variance of the intensities of the directional rose [5]. In addition to the three texture features that are related to the orientation of the autocorrelation function, we use two other texture attributes also in relationship with the autocorrelation function: the mean stroke width and height. In contrast to the initial work [9], which computes the two last features in horizontal and vertical direction, we have proposed to estimate the mean stroke width and height accurately along the axis of the main angle of the directional rose in order to indicate precisely the order of magnitude of the main strokes thickness. In a previous work [12], we have introduced the five autocorrelation features. In order to improve our method efficiency, based on the assumption that the textural information of the background is superfluous, i.e. its characteristics of gray level distribution is identical, we apply a foreground pixel selection step by using a conventional non-parametric binarization method, the Otsu method [13] 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. This stage of processing, however, is beyond the scope of our work, has proven to give good results in [14]. The authors use the Otsu method in order to segment and extract text regions from a document image. Shijian and Tan [15] binarize document images by using Otsu's global thresholding 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 the computation time and memory, and an improvement in the homogeneity accuracy average compared to our previous method [12]. We are not looking for an accurate segmentation, but we aim at finding regions with similar textural content. We subsequently compute those five autocorrelation descriptors only on the foreground pixels. In order to avoid side effects, we use border replication allowing computing texture features on the whole image. The texture features, carried out in the different areas of a gray-scale page, are performed with the help of an analysis by means of a sliding window, i.e. at various sizes of analysis windows in order to adopt a multiresolution approach. The sliding window is shifted horizontally and vertically scanning the whole image. The optimal sizes of the sliding windows, respecting a constructive compromise between the computation time and segmentation quality, have been determined experimentally. Thus, we assign to each foreground pixel a vector which is composed of 20 numeric values (5 autocorrelation indices $\times$ 4 sliding window sizes). The time required to process a document image of size 1982×2750 pixels is six minutes.

IV. Estimation of the Number of Clusters

Since the autocorrelation features are computed, we need a clustering algorithm in order to group similar indices and to define different kinds of information in the digitized page content. Nevertheless, for a certain class of clustering algorithms and particularly the conventional unsupervised clustering techniques [16], [17], the number of clusters in a dataset must be specified. In [18], several techniques have been proposed to determine the correct number of clusters in a dataset. Recently, the authors of [8], [19] show the relevance of the CC method to estimate the optimal number of clusters in biological data. In our work, with the help of the CC method, we estimate the true number of clusters in a set of randomly selected pixels of few pages of a book (block 2 on Figure 1). The process of the CC starts by randomly selecting a subset of samples from the data and then clustering it by performing a specified clustering algorithm. The sampling and clustering are iterated many times to evaluate the clustering results. The consensus clustering results correspond to a consensus matrix. Due to the variations

of the clustering algorithm performances, we use the consensus merge method [8] obtained from five different clustering methods: AGglomerative NESting (AGNES) [20], DIvisive ANAlysis clustering (DIANA) [20], Hierarchical Ascendant Classification (HAC) [21], k-means clustering (k-means) [22], and Partitioning Around Medoids (PAM) [20]. The merging of clustering results between different methods provides an average 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 Δk, which is computed from the cumulative density function of the merge consensus matrix in a range of possible values of cluster number.

V. Pixel labeling

After the estimation of the optimal number of clusters k_{opt}, we perform a segmentation step with the help of non-supervised clustering techniques. The clustering methods can be divided into five categories: partitioning, hierarchical, density-based, grid-based methods, and neural network-based methods [23]. Due to the inadequacies of the CC method for large databases, i.e. the high computational time of the CC algorithm, we use HAC on the computed autocorrelation features without taking into account the spatial coordinates to search and extract homogeneous regions for each digitized book page. In [24], the authors have proven the relevance of HAC for classifying the strokes of initial letters. The process of HAC consists in merging successively pairs of existing clusters, where at each cluster grouping step, the choice of clusters pairs depends on the smallest distance, i.e. clusters are grouped if the intra-cluster inertia is minimal. By setting the maximum number of clusters to k_{opt} estimated with the CC method, HAC is applied firstly on the autocorrelation features of the selected pixels of book pages (block 2 on Figure 1) and secondly on the computed autocorrelation descriptors of each page of a book (block 1 on Figure 1). Thus, we obtain the k_{opt} clusters for both the randomly selected samples of a book and for each digitized page of the same book. Then, we perform the Nearest Neighbor Search algorithm (NNS) [25] by computing the Mahalanobis distance [26] from each cluster obtained from the HAC results 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, i.e. by selecting the minimum Mahalanobis distance. The Mahalanobis distance takes into account the dataset correlations and is particularly suited to arbitrarily shaped clusters. Experimental results have shown that the application of HAC twice for the foreground pixels of each page and also for a set of randomly selected pixels of few pages is useful to correct the tendency of NNS to process outliers differently within datasets. Finally, by using NNS we assign the same cluster identifier or label for each similar cluster extracted from the digitized book. NNS helps us to characterize the content of an entire book and to find the homogeneous regions defined by similar indices of autocorrelation on the whole book (block 3 on Figure 1).

VI. Evaluation and results

To evaluate 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. Our corpus is extracted from the Gallica digital library[2]. 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. The evaluation of segmentation and region classification requires a ground truth. Our ground truth is defined manually by using the Groundtruthing Environment for Document Images (GEDI)[3], a public domain document image annotation tool. Rectangular regions have been drawn around each selected zone and identified by different labels when regions are not similar. Our proposal reaches very satisfying results when comparing visually the segmentation results (see Figure 2). In Figure 2, we note that our approach finds homogeneous regions from the content of historical digitized books, i.e. for instance from Figure 2(a) the graphic regions (blue) and textual regions (red) are similarly labeled in two different pages of the same book. We see especially from the four figures: 2(a), 2(b), 2(c), and 2(d) that the document images have been segmented into graphic regions, which correspond to an ornament and a drop cap, and textual regions. For the printed document category (two fonts and graphics) of Figure 2(d), we note that our approach distinguishes two distinct fonts: the normal (red) and uppercase (blue) fonts. While in Figure 2(e) for the manuscript document category (only two fonts), our method discriminates the noise on the document image borders from the textual regions but it cannot separate textual regions with different sizes and fonts. In Figure 2(f) for the printed document category (only two fonts), we also note that our approach cannot discriminate two different fonts: the normal and italic fonts.

Indeed, this method of assessing the effectiveness of a segmentation method is inherently a subjective evaluation and we need to assess the effectiveness using an appropriate quantitative metric. In order to measure the performance of segmentation methods and to assess the clustering results, two kinds of clustering validation indexes have been presented: the internal and external measures. The internal or unsupervised measures evaluate the quality of the clustering by considering only the intrinsic information on the distribution of the observations in the clusters. Whereas the external or supervised measures compare the distributions of the observations in the clustering result with the ground truth [23]. As our objective is to find the homogeneous regions defined by similar indices of autocorrelation, we have defined an external evaluation metric, the homogeneity measure $H(B,G)$ in a previous work [12], which evaluates the accuracy of our methodology in terms of matching regions between the ground truth and result regions. $H(B,G)$ is defined as:

$$H(B,G) = \frac{1}{|G|} \sum_j \frac{1}{|\{b_i \in g_j\}|} C_j \tag{1}$$

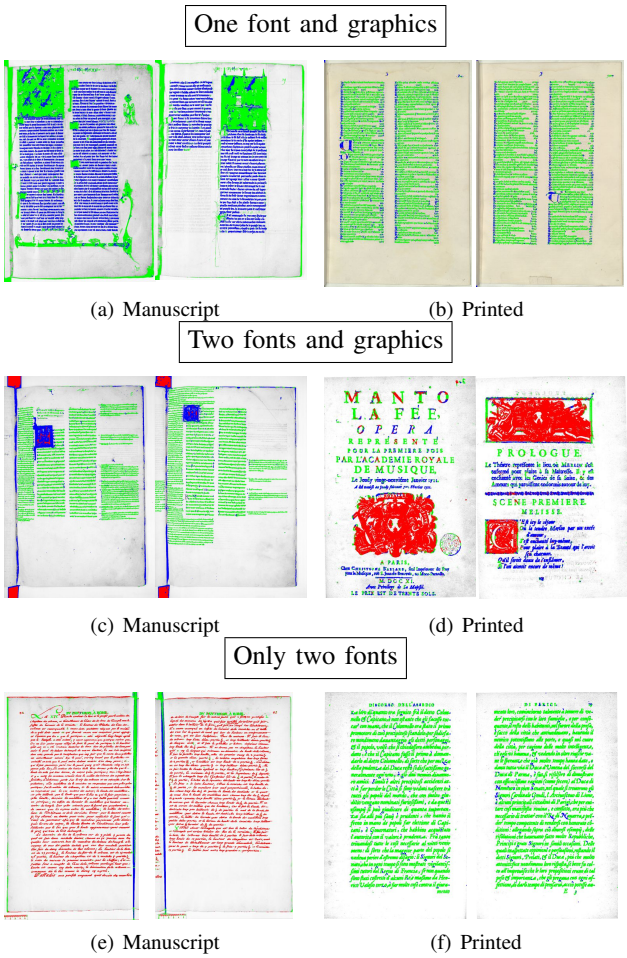
such as:

$$C_j = \max_{1 \leq k \leq k_{opt}} (|b_i, (b_i \in g_j) \wedge (l_{B_i} = k)|)$$

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 re-

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

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



(a) Manuscript (b) Printed

(c) Manuscript (d) Printed

(e) Manuscript (f) Printed

Fig. 2. Result examples of our pixel labeling approach for 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 color attributed to text or graphics may differ from one book to another.

spectively 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 the homogeneity measure (see equation (1)) are presented in Table I. We obtain 85% of mean homogeneity accuracy without taking into account the topographical relationships of selected pixels. 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 92% for the manuscript document category (one font and graphics). One assumption can be that the manuscript documents contain graphic 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 a higher homogeneity accuracy for pages containing graphics and single font text. This yields that the extracted autocorrelation features are able to distinguish textual regions from graphical ones. Whereas, in our previous work [12], we have found 80% of mean homogeneity accuracy for the printed document category (only two fonts) while in this paper, we obtain 90% of mean homogeneity accuracy. This may be explained by the fact that in this paper we introduce a step of foreground pixel selection using the Otsu method [13]. Thus, we conclude that this stage is more relevant to distinguish the foreground cluster from the background one. Overall, a slight improvement in the average of homogeneity measure is noted in addition to a significant gain in computation time with respect to our former method presented in [12].

TABLE I. HOMOGENEITY METRIC $H(B, G)$ RESULTS. μ AND σ ARE RESPECTIVELY THE MEAN VALUE AND STANDARD DEVIATION VALUES OF $H(B, G)$.

Document category	Document content	$\mu(H)$	$\sigma(H)$
Manuscript	One font and graphics	0,92	0,01
	Two fonts and graphics	0,88	0,04
	Only two fonts	0,85	0,05
	Overall	0,88	0,03
Printed	One font and graphics	0,77	0,05
	Two fonts and graphics	0,76	0,04
	Only two fonts	0,90	0,08
	Overall	0,81	0,03
Overall		0,85	0,04

An additional analysis and comparison with different internal and external clustering evaluation indices are needed in order to evaluate our segmentation and characterization method and also to validate our external evaluation metric, the homogeneity measure (see equation (1)). In this context, we compute four other clustering evaluation indices: two internal (Davies-Bouldin index [27] and Dunn index [28]) and two external (Jaccard coefficient [29] and Fowlkes-Mallows index [30]) indices.

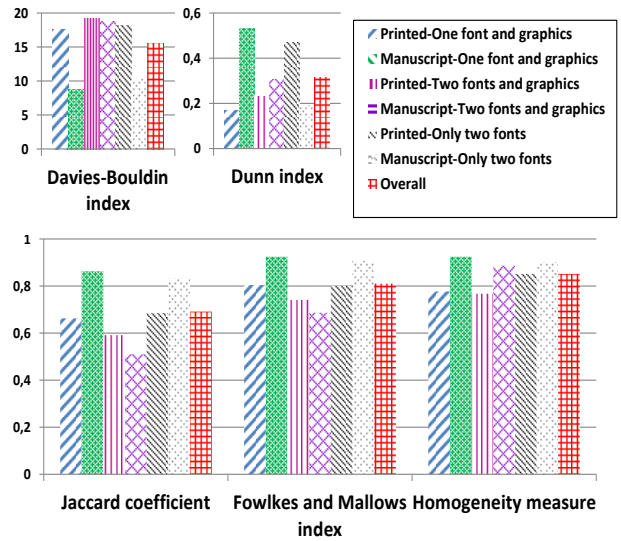


Fig. 3. Evaluation of our segmentation approach of historical digitized book content by internal and external clustering accuracy measures. The higher are these measures, the best are the results (except the Davies-Bouldin index where the lowest is the best).

In Figure 3, we show that the best clustering results are always obtained for the manuscript document category (one font and graphics) for the different clustering evaluation metrics. We also note that the second best result is observed for the printed document category (only two fonts) for the five accuracy clustering metrics. This may be explained by the fact that the autocorrelation features discriminate the noise in the document image from the textual regions (see Figure 2(e)). The Jaccard coefficient and Fowlkes-Mallows index show that

the lowest values are obtained for manuscript documents (two fonts and graphics). While the lowest outcomes for both homogeneity measure and Davies-Bouldin index are observed in the category of the printed documents (two fonts and graphics). We conclude that results of the homogeneity measure (equation (1)), are relatively in concordance with the various clustering evaluation indices presented previously and that the autocorrelation descriptors are more suitable for documents containing one font and graphics. This may be explained by the fact that the autocorrelation attributes mainly provide the main orientation of a texture (horizontal orientation for textual regions while many orientations are present in different proportions in graphic blocks).

VII. Conclusions and further work

This paper proposes an automatic layout segmentation and characterization of historical digitized book with little *a priori* knowledge, based on autocorrelation features. The robustness of the extracted features is used in a non-parametric unsupervised clustering method in order to extract the homogeneous regions defined by similar autocorrelation indices from the whole book instead of processing each page individually. The proposed method has been evaluated with promising results. Results show that the autocorrelation attributes are suitable to distinguish textual regions from graphical ones in the analyzed document. The main originalities of our new framework compared to our previous work [12] are: Firstly the integration of a new step allowing us to retrieve only the foreground pixels. This task has ensured a reduction of the data cardinality, a significant gain in the computation time and memory, and an improvement in the homogeneity accuracy average. Secondly, the integration of a new unsupervised task enabling us to automatically label content pixels with the same cluster identifier regarding to the book content. Finally, an additional analysis and comparison study with different internal and external clustering evaluation indices is presented in order to evaluate our framework and the relevance of the extracted textural features to separate text from graphics and different text fonts. The first aspect of future work will be to use recursive clustering methods in order to ensure the distinction between distinct text fonts and various graphic types. Besides, by integrating a new stage of processing after the pixel labeling task, which consists in pixel grouping by taking into consideration the topographical relationships of pixels and their labels, the classification results will be improved. Moreover, we will study and combine other statistical and frequential texture features in order to refine the segmentation.

References

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

[2] T. Pavlidis and J. Zhou, "Page Segmentation and Classification," *Graphical Model and Image Processing*, pp. 484–496, 1992.

[3] S. Khedekar, V. Ramanaprasad, S. Setlur, and V. Govindaraju, "Text - Image Separation in Devanagari Documents," in *IJDAR*. IEEE, 2003, pp. 1265–1269.

[4] S. Mao, A. Rosenfeld, and T. Kanungo, "Document structure analysis algorithms: a literature survey," in *DRR*. SPIE, 2003, pp. 197–207.

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

[6] C. H. Chen, L. F. Pau, and P. Wang, *Texture analysis in The Handbook of Pattern Recognition and Computer Vision*, 2nd ed. World Scientific, 1998.

[7] B. Julesz, "Visual pattern discrimination," in *Information Theory*. IEEE, 1962, pp. 84–92.

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

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

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

[11] 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.

[12] 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.

[13] N. Otsu, "A threshold selection method from gray-level histograms," in *Systems, Man, and Cybernetics*. IEEE, 1979, pp. 62–66.

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

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

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

[17] A. Cornuéjols and L. Miclet, *Apprentissage artificiel: Concepts et algorithmes*. 2nd Edition Eyrolles, 2010.

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

[19] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data," *Machine Learning*, pp. 91–118, 2003.

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

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

[22] 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.

[23] F. Kovács, C. Legány, and A. Babos, "Cluster validity measurement techniques," in *International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases*. World Scientific and Engineering Academy and Society, 2006, pp. 388–393.

[24] G. Nguyen, M. Coustaty, and J. Ogier, "Stroke feature extraction for lettrine indexing," in *IPTA*. IEEE, 2010, pp. 355–360.

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

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

[27] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *PAMI*, pp. 224–227, 1979.

[28] J. Dunn, "Well separated clusters and optimal fuzzy partitions," *Journal of Cybernetics*, pp. 95–104, 1974.

[29] 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*, pp. 415–425, 1991.

[30] E. B. Fowlkes and C. L. Mallows, "A Method for Comparing Two Hierarchical Clusterings," *Journal of the American Statistical Association*, pp. 553–569, 1983.