A structural signature based on texture for digitized historical book page categorization

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

A Structural Signature Based on Texture for Digitized Historical Book Page Categorization

Maroua Mehri*†, Pierre Héroux†, Julien Lerouge†, Petra Gomez-Krämer*, and Rémy Mullot*
*L3i, University of La Rochelle, Avenue Michel Crépeau, 17042, La Rochelle, France
Emails: {maroua.mehri, petra.gomez, remy.mullot}@univ-lr.fr
†LITIS, University of Rouen, Avenue de l'Université, 76800, Saint-Etienne-du-Rouvray, France
Emails: {pierre.heroux, julien.lerouge}@univ-rouen.fr

***Abstract*—The work conducted in this article presents a structural signature based on texture for the characterization and categorization of digitized historical book pages. The proposed signature does not assume *a priori* knowledge regarding page layout and content, and hence, it is applicable to a large variety of ancient books. By integrating varying low-level features (e.g. texture) characterizing the different page components (*i.e.* different text fonts or graphic regions) on the one hand, and structural information describing the page layout on the other hand, the proposed signature provides a rich and holistic description of the layout and content of the analyzed book pages. More precisely, the signature-based characterization approach consists of two stages. The first stage is extracting automatically homogeneous regions. Then, the second one is proposing a graph-based page signature, which is based on the extracted homogeneous regions, reflecting its layout and content. This signature ensures the implementation of numerous applications for managing effectively a corpus or collections of books (e.g. information retrieval in digital libraries according to several criteria or page categorization). To illustrate the effectiveness of the proposed page signature, a detailed experimental evaluation has been conducted in this article for assessing two possible categorization applications, unsupervised page classification and page stream segmentation.**

***Keywords*—Historical books, Categorization, Texture, Graph.**

I. Introduction

Over the last few years, there has been tremendous growth in digitizing collections of cultural heritage documents. Thus, many challenges and open issues have been raised, such as information retrieval in digital libraries or analyzing page content of digitized historical books (DHBs). Therefore, with the support of the French National Research Agency[1], we are working on a project named DIGIDOC[2]. The ultimate goal of the DIGIDOC project is developing new ways of interacting with scanners by assisting the digitization operator to adjust automatically the best set of parameters (e.g. resolution, lightening, color calibration), detecting errors in the digitization process (e.g. blur, skewed or folded pages), providing appropriate assistance for document indexing (e.g. by recognizing automatically page types or breaks in a sequence of pages), *etc.* There is an absolute need to design "smart" digitizers which can limit manual intervention and perform easy and high quality digitization of document images (DIs) [1]. Therefore, to achieve better interaction with scanners, we need to design a computer-aided categorization tool, able to index or categorize DHB pages according to several criteria (e.g. layout structure, graphical properties or typographical characteristics of their content).

Several scientific works in contemporary document image analysis (DIA) have described several relevant approaches enabling multiple forms of indexing and classification based on content analysis of DIs. The current systems for categorizing digitized DIs are based on several criteria for the textual content by applying optical character recognition (OCR) or by using the interest point detection approach [2]. Nevertheless, the transposition of these tools for historical DIA, that are dedicated initially for contemporary DIA, is not a straightforward task. Grana *et al.* [3] stated that, despite the OCR-based methods have yielded reliable results for contemporary DIA, analyzing historical document images (HDIs) by separating textual regions from the graphical ones is still more challenging. Indeed, these tools for performing the historical DIA tasks have poor performance due to many particularities of HDIs (e.g. large variability of possible page layouts and/or contents, noise and degradation).

As a matter of fact, the work conducted in this article proposes a structural signature based on texture for DHB page characterization and categorization. The proposed signature does not assume *a priori* knowledge regarding the layout and content of the analyzed DHB pages, and hence, it is applicable to a large variety of ancient books. It integrates varying low-level features characterizing the different HDI content components (*i.e.* different text fonts or graphic regions) on the one hand, and structural information describing the HDI layout on the other hand. This rich and holistic representation of the content and layout of the analyzed DHB page can be adapted to the user preferences and specified criteria through the extracted varying levels of information (e.g. by selecting only the information characterizing the HDI layout and/or content or by retrieving any useful information available for a subsequent use). The proposed page signature allows the implementation of several applications for managing effectively a corpus or collections of DHBs. To name a few, we may underline the following applications based on the defined page signature in this work:

- recognizing the analyzed page type to ensure an automatic adjustment of the quality of the page scanning process with respect to the page signature and the subsequent use,
- modeling a computer-aided categorization tool, able to index, compare or classify DHB pages or DHBs

[1] http://www.agence-nationale-recherche.fr/en/

[2] http://www.agence-nationale-recherche.fr/en/anr-funded-project/?tx_lwmsuivibilan_pi2[CODE]=ANR-10-CORD-0020

according to several criteria (e.g. HDI layout and/or content) or to retrieve pages which have particular layout and/or content (e.g. empty or cover DHB pages),

- identifying specific pages, such as the transition pages in a DHB (e.g. title pages of chapter), which require a particular indexing process, to generate automatically a table of contents/summary of the analyzed DHB,
- detecting pages having scanning failure occurring during the digitization process (e.g. blur, skewed or folded pages), *etc.*

The remainder of this article is organized as follows. In Section II, the proposed signature for DHB pages characterization and categorization is detailed. Section III describes firstly the experimental protocol. Then, qualitative results and an assessment of the different steps of the proposed approach used to generate this signature are presented. Subsequently, an analysis of the obtained results is discussed. Afterwards, a thorough evaluation has been conducted for assessing two possible signature-based applications, unsupervised book page classification and book page stream segmentation, to illustrate the potential of the proposed signature. Finally, our conclusions and future work are presented in Section IV.

II. Proposed Signature-Based Approach

In this article, we propose a structural signature which characterizes the layout and content of DHB pages. To illustrate the potential of the proposed signature, two possible signature-based applications (unsupervised page classification and page stream segmentation) are proposed for DHB page categorization.

A. Characterization

We proposed in our previous work [4], a pixel-labeling framework which automatically extracts texture descriptors for discriminating the different classes of the foreground layers. The proposed framework is supported by the fact that pages of the same book usually present strong similarities in the organization of the HDI information (*i.e.* layout) and in the graphical and typographical features (*i.e.* content) throughout the DHB pages. Indeed, the texture information which is often repeated and recurrently present in many DHB pages, can be deduced by exploiting the regularities of the associated textures through the whole DHB pages. The texture features are then used in an unsupervised classification approach to determine the number of DHB content types that are defined by similar textural descriptors. The originality of the proposed framework lies in the texture feature analysis that is used on an entire book instead of processing each page individually. The proposed framework does not require *a priori* knowledge of the layout, typographical parameters or graphical properties of the analyzed DHB pages.

Based on the results obtained from the pixel-labeling framework presented in [4], the first contribution of this article consist in determining automatically homogeneous texture regions or groups of foreground pixels which share similar characteristics for DHB content characterization. Once, the homogeneous regions have been extracted, the next stage of the proposed approach consists in providing a graph-based signature. This signature represents the layout and content of the analyzed DHB page. Figure 1 illustrates the detailed schematic block diagram of the proposed approach used to generate the graph-based signature for DHB page characterization. The four first steps of the proposed approach: *Step* 1 (*cf.* Section II-A1), *Step* 2 (*cf.* Section II-A2), *Step* 3 (*cf.* Section II-A3) and *Step* 4 (*cf.* Section II-A4) have been thoroughly detailed in [4], [5]. The proposed approach, used to generate the graph-based signature for DHB page characterization, is composed of the following seven tasks:

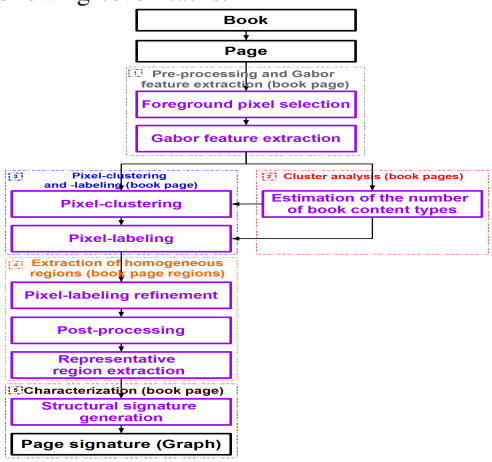


Fig. 1. Detailed schematic block diagram of the proposed approach used to generate the graph-based signature for DHB page characterization.

1) Pre-processing and Gabor feature extraction (Step 1*):* In our previous work [5], some of the well-known texture-based approaches were compared for HDI segmentation. We concluded that the Gabor features are the highest performing descriptors for font and text/graphic segmentation. Thus, Gabor filters are used in the proposed framework, which is initially evaluated with other descriptors deduced from the auto-correlation function in [4], by setting the same default Gabor parameters as proposed in [5]. The Gabor features are only extracted from the selected foreground pixels of the transformed image by the selective Gabor filter at different sizes of sliding windows to adopt a multi-scale approach.

2) Estimation of the number of DHB content types (Step 2*):* An unsupervised classification step is subsequently performed on the Gabor features, which are extracted from a sub-sampling in the entire book. Hence, a number of foreground pixels from pages of the same DHB are selected randomly, and their Gabor descriptors are then analyzed in order to estimate automatically the number of book content types in the analyzed DHB, using a merge consensus clustering technique.

3) Pixel-clustering and labeling (Step 3*):* The extracted Gabor features for each DHB page are afterwards used in an unsupervised classification approach by taking into account the estimation of the number of book content types (*Step* 2). The idea consists in integrating an unsupervised task that automatically labels content pixels with the same cluster identifier as the DHB content. For each DHB page image, the foreground pixels belonging to the same cluster are labeled according to the cluster label obtained from the initial clustering, which is performed at the DHB scale (*Step* 2).

4) Pixel-labeling refinement (Step 4*):* The pixel-labeling task consists in labeling independently each foreground pixel based on analyzing texture features on different sizes of sliding windows. Nevertheless, due to the presence of noise, the foreground pixels will be prone to incorrect labeling.

However, based on the fact that the neighboring pixels have higher probability to belong to the same page content type, the mis-labeling errors can be corrected. Thus, to refine the pixel-labeling results, the topological relationships between the selected foreground pixels are integrated using the spatial multi-scale analysis of majority votes.

5) Post-processing (Step 5): Since the pixel-labeling refinement step (*Step* 4) has been performed, our output data consists of a refined pixel-labeled HDI. Nevertheless, we aim in this step to identify few groups of pixels sharing similar textural properties to extract homogeneous regions. Thus, a task of extracting and analyzing the connected components (CCs) with an adaptive run-length smoothing algorithm (ARLSA) are carried out to connect the extracted CCs and subsequently to define regions. Then, the majority voting technique is performed to assign a label to each extracted CC, based on the majority label among the region pixels.

6) Extraction of representative homogeneous regions (Step 6): As already mentioned, due to the characteristics of HDIs linked to the presence of noise and degradation, many extracted CCs correspond usually to noise. Thus, a selection of the most representative or significant CCs is required (*i.e.* the largest CCs). The idea of the selection of representative CCs is to keep a sub-set of the initially extracted CCs in the *Step* 5, by at least retrieving 95% of the total number of foreground pixels. This step aims at removing the small isolated CCs induced by the presence of noise, while keeping the relevant information. Therefore, by extracting the significant CCs, the representative homogeneous regions are determined and labeled. This step ensures both the size reduction of the proposed graph-based signatures and speeding up their handling by graph algorithms whose computational complexity is exponential in the number of vertices of the involved graphs.

7) Generation of a structural signature per page (Step 7): Leveraging on the numerous advantages offered by using a structural representation instead of a statistical or ontology one [6], [7], a structural signature is proposed in this work. The proposed signature is used to represent a DHB page with the set of representative homogeneous regions extracted in the *Step* 6. It is in the form of a directed attributed graph. The graph vertices correspond to the extracted representative homogeneous regions. Each vertex is described by a 238-D feature vector, representing the varying low-level features:

- 192 Gabor attributes (A_G^v) to characterize the typographical and graphical characteristics of the extracted representative homogeneous region [5],
- 46 shape, geometric and topological attributes (A_{SGT}^v) (e.g. centroid position, number of pixels, gray-level average, contour area and perimeter, Hu, spatial, central and central normalized moments) to describe the shape and spatial properties of the extracted representative homogeneous region.

Furthermore, a set of edges is built based on the topological relationships connecting the extracted representative homogeneous regions. An edge is built between two vertices, if $F_e^{s,d} \geq Th_e$, where $F_e^{s,d}$ and Th_e denote the edge force and threshold, respectively. The $F_e^{s,d}$ (*cf.* equation 1) characterizes the gravitation force between two graph vertices: source (G_v^s) and destination (G_v^d). The Th_e has been experimentally determined, and is equal to 0.1. The $F_e^{s,d}$ models the interaction existence and level between two extracted representative homogeneous regions by drawing an arrow pointing to the G_v^s (*cf.* Figures 2(b) and 2(d)). There is no interaction between two small regions unless they are close to each other, however a large region can have multiple interactions with more distant regions.

$$F_e^{s,d} = \frac{N_{G_v^d}}{(ED_{G_v^{s,d}})^2} \quad (1)$$

In equation 1, $N_{G_v^d}$ denotes the number of pixels of the G_v^d. $ED_{G_v^{s,d}}$ denotes the Euclidean distance between the two graph vertices G_v^s and G_v^d.

Besides the edge force attribute (A_F^e) used to characterize the topological relationships between two extracted regions, two other topological edge attributes (A_T^e) are also computed: the absolute differences between the two extracted region centroids in the x- and y-axis.

B. Categorization

Since the characterization of the DHB page layout and content is performed using the proposed graph-based signature, the categorization task of the DHB pages can be carried out by comparing the different graph-based signatures. As a consequence, the similarities of DHB page layout and/or content can be deduced. Then, the DHB pages can be categorized, and DHB pages with similar layout and/or content pages can be grouped. The proposed DHB page signature allows the implementation of numerous applications for managing effectively a corpus or collections of DHBs.

Among the possible applications of the proposed DHB page signature, cited earlier, a thorough evaluation has been conducted in this work for assessing:

1) Unsupervised DHB page classification,
2) DHB page stream segmentation (*i.e.* to identify the transition pages in a DHB such as the title pages of chapter).

To categorize and group DHB pages with similar layout and/or content, the obtained graph-based DHB page signature can be compared using a graph dissimilarity. In our experiments, we use the graph edit distance (GED). The GED is used to measure the (dis)similarity between the obtained graph-based DHB page signatures [8]. The GED deals with the computation of the minimum-cost sequence of the basic graph editing operations (e.g. insertion/deletion and substitution of vertices or edges) to transform a graph to another one. The GED has to be set up based on the costs of the elementary edit operations (insertion/deletion, substitution). These costs are functions of the label of vertices/edges. The weight of each feature composing the label has been set after a statistical analysis of the feature variations in order to give the same importance to texture features and shape/geometric/topological descriptors. In this work, the computational complexity of the GED is reduced, since we have a limited number of vertices in the obtained graphs (*i.e.* up to 11 vertices).

Therefore, the evaluation of the proposed page signature has been carried out based on firstly computing a distance matrix, whose elements represent the dissimilarity between the compared graphs. The dissimilarity corresponds to the GED,

normalized with respect to the graph size. Indeed, for a fixed number of edit operations needed to transform one graph into another, the dissimilarity is higher if the graphs are small, and lower if the involved edit operations only affect a tiny portion of a large graph. Then, by analyzing the elements of the resulting distance matrix (M^g), two following applications are targeted.

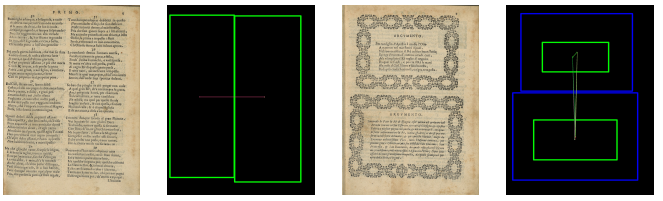
Firstly, an unsupervised classification task using the hierarchical agglomerative clustering algorithm, is performed on all elements of M^g ($m^g_{i,j}$). Since we deal with an unsupervised classification task, we aim to separate the involved DHB pages into 2 clusters. One cluster representing frequent pages having similar layout and/or content and the other one illustrating pages having particular layout and/or content. Secondly, by only analyzing the $m^g_{i,i+1}$ elements of M^g, the different pairs of the successive DHB pages can be grouped or retrieved according to a pre-defined threshold GED value. This task aims to retrieve the transition pages in the involved DHB (*i.e.* identify different series of successive pages having distinct layout and/or content). It is worth noting that this task is considerably important, since it can detect pages having scanning failure occurring during the digitization process (e.g. blur, skewed or folded pages). Moreover, by identifying these transition pages (e.g. title pages of chapter), a particular indexing process can be carried out to assist a user in generating a table of contents/summary (*i.e.* DHB page stream segmentation).

III. Evaluation and Results

Our experimental corpus contains a DHB of 322 ground-truthed one-page color HDIs[3]. The analyzed DHB has been collected from Gallica[4], and its pages have been digitized at 300 dpi and saved in the TIFF format. The analyzed DHB consists of 81 pages containing graphical and textual regions (*i.e.* pages that have particular layout and/or content) and 241 pages containing only textual regions (*i.e.* the most common or frequent pages that have similar layout and/or content).

A. Characterization

To evaluate the performance of the proposed signature-based approach for DHB page characterization, qualitative results and an assessment of its different steps are presented in Figure 2 and Table I, respectively.



(a) *Input* 1 (b) *Step* 7(*Output* 1) (c) *Input* 2 (d) *Step* 7(*Output* 2)

Fig. 2. Illustration of the obtained graph-based signatures of DHB pages.

TABLE I. Evaluation of the different steps of the proposed approach for DHB page characterization.

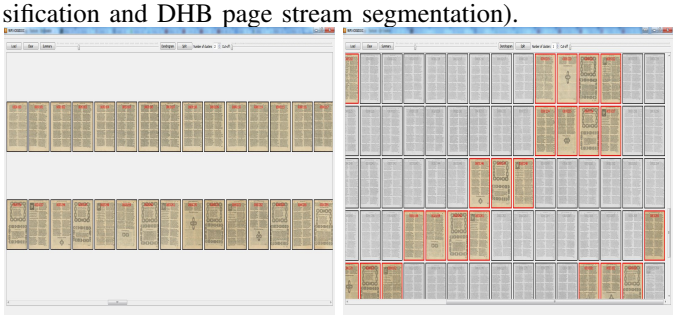
		Step 3	*Step* 4	*Step* 5			*Step* 6
CA	μ	0.977	0.983	0.987	J_{AR}	μ	0.952
	σ	0.066	0.085	0.076		σ	0.174

[3]http://gallica.bnf.fr/ark:/12148/bpt6k132294p/f5.planchecontact.r=.langFR
[4]http://gallica.bnf.fr

The success of the proposed approach is demonstrated by visual inspection of the segmented HDIs (*i.e.* homogeneous regions are determined by identifying the graphic regions (blue) and textual regions (green)). Then, the pixel-based classification accuracy (*CA*) is computed to evaluate quantitatively the obtained results of the following steps of the proposed approach for DHB page characterization: the pixel-clustering and labeling step (*Step* 3), the pixel-labeling refinement step (*Step* 4) and the post-processing step (*Step* 5) [5]. Another accuracy metric is calculated, the Jaccard index (J_{AR}), for assessing the step of extraction of representative homogeneous regions (*Step* 6) [9]. $\mu(.)$ and $\sigma(.)$ are the mean and standard deviation of $(.)$, respectively. The higher the mean values, the better the results. High performances of the computed accuracy metrics are obtained for the different steps of the proposed signature-based approach (*i.e.* more than 95%). Moreover, a slight gain in the average value of *CA* is obtained from one step to the next, in order to achieve the aim of identifying homogeneous regions with an average value of J_{AR} equal to 95%.

B. Categorization

A GUI tool (*cf.* Figure 3) is designed in this work to illustrate graphically the performance of the two evaluated signature-based applications (*i.e.* unsupervised DHB page classification and DHB page stream segmentation).



(a) Unsupervised page classification (b) Page stream segmentation

Fig. 3. Screen shots illustrating graphically the performance of the two evaluated signature-based applications.

First, we can see in Figure 3(a) the separation of the DHB pages into 2 clusters. One cluster representing frequent pages having similar layout and/or content and the other one illustrating pages having particular layout and/or content. Each cluster is represented in a separate line. In our experiments and particularly in the DHB on which the evaluation has been performed, we note that the clustering achieves a distinction between pages having similar layout and/or content (*i.e.* double columns of text), and those having particular layout and/or content (*i.e.* textual and graphical regions). Second, we can see in Figure 3(b) the different detected transition DHB pages. Only DHB pages having GEDs above a pre-defined threshold GED value are retrieved. The shaded DHB pages are considered as non-transition pages, while the DHB pages with red borders are considered as the transition pages (*i.e.* they have layout and/or content that differ from the following page). Using the developed computer-aided tool for characterization and categorization of DHB pages in this work, users are able to vary the threshold GED in order to increase or decrease the number of transition pages. The proposed tool for characterization and categorization of DHB pages provides an integrated user-centered GUI which is specifically engineered

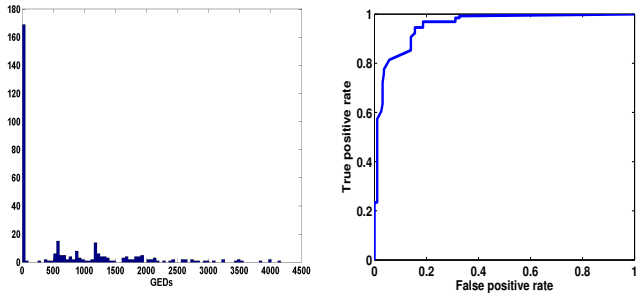
to make it easy the identification of the transition pages in the DHB under consideration according to the user requirements.

1) Unsupervised DHB page classification: To get an insight into the classification accuracy, a confusion matrix is computed (*cf.* Table II). The confusion matrix illustrates one cluster containing the most common pages in the involved DHB (*i.e.* pages containing only text) on the one hand, and those considered as particular pages in the involved DHB (*i.e.* pages containing text and graphics) on the other hand. The following classification accuracy measures are computed: precision (P), recall (R), F-measure (F) and classification accuracy (CA). P_i and R_j denote the individual cluster precision and recall, respectively. For the cluster representing the most common pages in the involved DHB (*i.e.* pages containing only text), 91%(P) and 94%(R) are obtained. On the other side, for the cluster representing the pages that have particular layout and/or content (*i.e.* pages containing text and graphics), we find 85%(P) and 77%(R). Thus, we show that the proposed approach tends to miss-classify more the pages containing textual and graphical regions than those containing only textual regions, due to the complexity of the layout and content of the particular DHB pages (*cf.* Figure 2(c)). Nevertheless, the overall result is quite encouraging, since we obtain 87%(F) and 90%(CA). This confirms that the proposed signature ensures the unsupervised DHB page classification according to the DHB page content.

TABLE II. EVALUATION OF THE PROPOSED SIGNATURE FOR UNSUPERVISED DHB PAGE CLASSIFICATION.

		Ground truth		
		Class1	***Class2***	
Clustering outcomes	***Cluster1***	69	12	$\leftrightarrow P_1 = 0.85$
	Cluster2	20	221	$\leftrightarrow P_2 = 0.91$
		$\updownarrow R_1 = 0.77$	$\updownarrow R_2 = 0.94$	

2) DHB page stream segmentation: By analyzing the $m^g_{i,i+1}$ elements of the normalized distance matrix M^g, the different pairs of the successive DHB pages can be grouped according to different GED values. As a matter of fact, the pairs of the successive DHB pages that have lower GED values, have certainly similar layout and/or content (*i.e.* non-transition pages). On the other side, the other pairs that have higher GED values correspond to pages have different layout and/or content (*i.e.* transition pages such as the title pages of chapter). By analyzing the composition of the involved DHB, 102 pairs of the successive DHB pages are identified as pairs of transition pages. From these pairs, 128 DHB pages are considered as transition pages. By drawing the histogram of the computed GED values between each pair of successive DHB pages (*cf.* Figure 4(a)), on peak is showed with lower values of GED (*i.e.* the lower the GED value, the more similar the pages in terms of layout and content). This histogram peak corresponds to the number of detected pairs of successive DHB pages which have similar layout and/or content (*i.e.* double columns of text) and can be identified as non-transition pages according to the obtained GED value. Thus, this confirms that the proposed graph-based signature used for page stream segmentation is robust and relevant. Figure 4(b) illustrates the ROC curve by varying the GED threshold values illustrating the good performance of the proposed signature for the identification of the transition DHB pages. This strengthens our previous results and confirms that the proposed signature ensures the identification of the transition pages in a DHB such as the title pages of chapter and subsequently it allows the DHB page stream segmentation.



(a) (b)

Fig. 4. Evaluation of the proposed page signature for DHB page stream segmentation.

IV. CONCLUSIONS AND FURTHER WORK

This article proposes a generic signature for DHB page characterization and categorization with no hypothesis concerning the page layout and its content. Based on characterizing each DHB page with a set of homogeneous texture regions with varying low-level features, a graph-based signature is designed for each DHB page. It ensures the characterization of the DHB page layout and content. Then, by comparing the different graph-based signatures, the DHB pages with similar layout and/or content pages can be grouped. The proposed signature has been evaluated on two possible applications, unsupervised book page classification and book page stream segmentation, and it has achieved promising results.

The first aspect of future work will be to use the proposed signature on a larger corpus. This study is ongoing and will evaluate the signature more adequately, with more convincing experimental results. We will then assess other possible applications (e.g. finding pages or particular content components in a DHB that match specific criteria defined by a user). Furthermore, we will investigate a finer unsupervised book page classification with different values of the number of clusters. We also intend to analyze the impact of different feature weighting schemes in the cost of the edit operations when computing the GED.

REFERENCES

[1] F. LeBourgeois, E. Trinh, B. Allier, V. Eglin, and H. Emptoz, "Document images analysis solutions for digital libraries," in *DIAL*, 2004, pp. 2–24.

[2] O. Augereau, N. Journet, A. Vialard, and J. P. Domenger, "Improving classification of an industrial document image database by combining visual and textual features," in *DAS*, 2014, pp. 314–318.

[3] C. Grana, G. Serra, M. Manfredi, D. Coppi, and R. Cucchiara, "Layout analysis and content enrichment of digitized books," *MTA*, pp. 1–22, 2014.

[4] M. Mehri, P. Héroux, P. Gomez-Krämer, and R. Mullot, "A pixel labeling approach for historical digitized books," in *ICDAR*, 2013, pp. 817–821.

[5] M. Mehri, M. Mhiri, P. Héroux, P. Gomez-Krämer, M. A. Mahjoub, and R. Mullot, "Performance evaluation and benchmarking of six texture-based feature sets for segmenting historical documents," in *ICPR*, 2014, pp. 2885–2890.

[6] D. Conte, P. Foggia, C. Sansone, and M. Vento, "Thirty years of graph matching in pattern recognition," *IJPRAI*, pp. 265–298, 2004.

[7] S. Jouili, M. Coustaty, S. Tabbone, and J. M. Ogier, "NaviDoMass: structural-based approaches towards handling historical documents," in *ICPR*, 2010, pp. 946–949.

[8] H. Bunke and K. Riesen, "Towards the unification of structural and statistical pattern recognition," *PRL*, pp. 811–825, 2012.

[9] S. Brunessaux, P. Giroux, B. Grilheres, M. Manta, M. Bodin, K. Choukri, O. Galibert, and J. Kahn, "The Maurdor project: improving automatic processing of digital documents," in *DAS*, 2014, pp. 349–354.