Page Retrieval System in Digitized Historical Books Based on Error-tolerant Subgraph Matching

Maroua Mehri, Pierre Héroux, Julien Lerouge, Rémy Mullot

Page Retrieval System in Digitized Historical Books Based on Error-tolerant Subgraph Matching

Maroua Mehri*†, Pierre Héroux†, Julien Lerouge† and Rémy Mullot‡
*LATIS Laboratory, Sousse University, National Engineering School of Sousse, 4023, Sousse Erriadh, Tunisia
†LITIS Laboratory, Normandie University, Avenue de l'Université, 76800, Saint-Etienne-du-Rouvray, France
‡L3i Laboratory, University of La Rochelle, Avenue Michel Crépeau, 17042, La Rochelle, France
Emails: maroua.mehri@gmail.com, {pierre.heroux, julien.lerouge}@univ-rouen.fr and remy.mullot@univ-lr.fr

***Abstract*—Developing smart ways of interacting with scanners is one of the emerging needs identified by numerous digitization professionals. To achieve better interaction with scanners, the research community in historical document image analysis is particularly interested in providing reliable tools for computer-aided indexing and retrieval of historical document images. Thus, we propose in this article a method able to retrieve from a digitized historical book, pages having layout and/or content which meet the user-defined query. Amongst the user-defined queries we focus on the transition pages (e.g. title pages of chapter, end-of-chapter and end-of-act) and pages containing a particular content component or a group of patterns (e.g. ornaments, illustrations and drop caps) in our work. The method adopted in this work is firstly based on using low-level features (texture, shape and geometric descriptors) to represent each page in the form of a graph-based signature. Then, a set of costs is estimated using an error-tolerant subgraph isomorphism algorithm in order to measure the similarity between the user-defined query formulated in terms of a pattern graph and the different subgraphs of the book page signatures and to find book pages similar to the user-defined query. To illustrate the effectiveness of the proposed method, a thorough experimental study has been conducted with quantitative observations obtained from a large number of queries having different contents and structures.**

***Index Terms*—Page retrieval, Low-level features, Graph-based signature, Error-tolerant subgraph matching.**

I. Introduction

Since the last decade of the twentieth century, numerous large-scale digitization projects of cultural heritage documents have been conducted by several museums and libraries around the world. A number of research projects have been set up with the support of public funding to provide reliable indexing and retrieval systems in digital libraries. For instance, in the context of the BVH project an interactive historical document layout analysis and segmentation tool, named AGORA, was designed to ensure new powerful technological capabilities that enable users to search among titles, authors, dates and other different queries relative to the digitized books to retrieve a particular book or book page block (e.g. graphical or textual parts) [1]. In the context of the DMOS project, Coüasnon [2] proposed a generic document recognition method based on a grammatical language and an associated parser able to deal with noise in damaged military forms of the 19^{th} century, found in French archives. In the context of the MADONNE project, Journet *et al.* [3] proposed firstly an unsupervised book content characterization method based on extracting signatures which represent textural characteristics of book pages. Then, to illustrate the pertinence of the extracted signatures an experimental evaluation was conducted for evaluating two possible image retrieval applications. The first application is based on comparing pages by considering the pixel partition and the spatial organization of blocks of texts, of drawings and of background as the criteria for measuring the similarities between pages. The second application is based on comparing elements of content by searching images by the content based on historical drawings of old documents (e.g. drop caps).

In this article, a page retrieval system is proposed. The proposed system aims at retrieving from a digitized historical book, pages having layout and/or content which meet the user-defined query. The transition pages (e.g. title pages of chapter, end-of-chapter and end-of-act) and pages containing a particular content component or a group of patterns (e.g. ornaments, illustrations and drop caps) are few examples of the user-defined queries which have been explored in this article. The proposed method is firstly based on using low-level features (texture, shape and geometric descriptors) to represent each page in the form of a graph-based signature. Then, it uses an error-tolerant subgraph isomorphism algorithm to measure the similarity between the user-defined query formulated in terms of a graph and the different subgraphs of the book page signatures and to find book pages similar to the user-defined query. The main contribution of our work consists in evaluating the capability and the robustness of a graph-based signature to retrieve from a digitized historical book, pages having layout and/or content which meet the user-defined query. The proposed page retrieval system is based on using a previously proposed subgraph matching technique on a challenging type of document images.

The remainder of this article is organized as follows. Section II presents briefly the main phases of the proposed method used to generate a graph-based signature for book page content and layout characterization. Section III presents the error-tolerant subgraph isomorphism algorithm. Section IV first describes the experimental corpus and then details the experiments carried out to evaluate the use of the proposed graph-based signature for page retrieval in digitized historical books. Finally, our conclusions and future work are presented in Section V.

II. Graph-based signature

In this section, we will present succinctly the main phases of the proposed method used to generate a graph-based signature for book page content and layout characterization which were introduced by Mehri *et al.* [4].

Researchers state that there is still a great need for robust pattern recognition and analysis techniques supporting the particularities of historical documents (e.g. large variability of page layout and/or content, noise, degradation) and able to ensure rigorous description, classification and indexing of historical document collections without *a priori* knowledge regarding document layout and/or content [5]. Given these constraints, the use of the structural approaches has become an appropriate choice for document image representation [6], [7]. By using a structural approach, a document is represented by a data structure (e.g. graph) to model content elements and their relationships in a document image.

Thus, a structural signature based on low-level features is used in this work to characterize the digitized page content and layout. This structural signature is represented in the form of a region adjacency graph (RAG) in which vertices correspond to content elements on the analyzed book page (textual blocks and graphical regions) and whose edges represent the topological relationships connecting the different content elements. The book page regions are represented by graph vertices. Each vertex is labeled with a 22D feature vector describing the region or content element using low-level features (texture, shape and geometric descriptors), characterizing the content element. The low-level features used to characterize the content elements are twofold:

- *Shape and geometric*: 2 shape and geometric descriptors have been extracted: the number of pixels composing the area of the content element and its eccentricity (*i.e.* ratio of the height to the width of the bounding rectangle of the content element). These descriptors are used to describe the shape of the content elements;
- *Texture*: 20 auto-correlation descriptors have been extracted by means of a multi-scale analysis technique. These descriptors represent the mean values of the auto-correlation indices computed from all selected foreground pixels composing the content element at four different sizes of rectangular overlapping processing windows [8].

To complete our structural representation in the form of a directed attributed graph, a set of edges connecting the different vertices, which are associated with the different content elements or homogeneous regions of a document image is added. These edges represent the most significant interactive relationships between the different content elements composing a document. The criterion of interaction between two different content elements composing a document, which is called the edge force, is modeled based on the principle of Newton's law of universal gravitation [4]. Therefore, to measure the interaction level between two element contents of a document, an edge force is firstly computed. Then, if the value of the edge force exceeds an empirically determined threshold, an oriented edge between the two involved vertices will be first added and then the absolute differences between the two region centroids in the x and y-axis besides to the value of the edge force will be defined in the list of the edge attributes.

Figure 1 illustrates qualitatively the obtained graph-based signatures of book pages. In this work, a structural representation of each book page is obtained by creating a graph in which a vertex is associated with an extracted representative homogeneous region. The extraction of representative homogeneous regions is performed by using a bottom-up segmentation method which was introduced in [4]. The used bottom-up segmentation method in this work is based on analyzing texture features and connected components with an adaptive run-length smoothing algorithm.

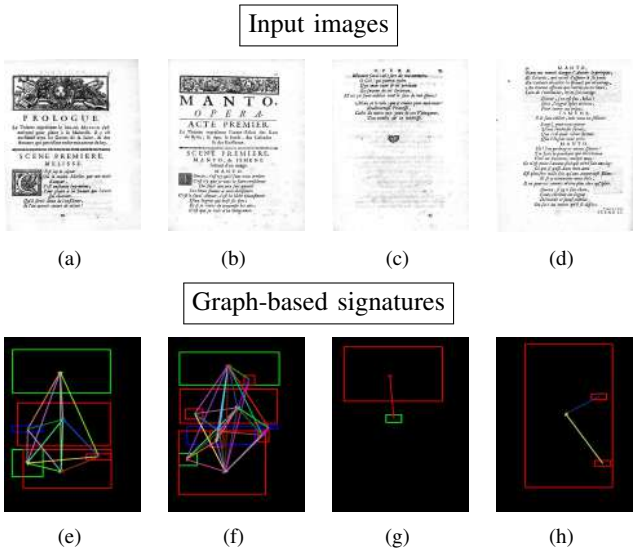


Fig. 1. Illustration of some examples of graph-based signatures of book pages. (e), (f), (g) and (h) depict the obtained graph-based signatures corresponding to the input images (a), (b), (c) and (d), respectively.

III. Error-tolerant subgraph matching

Retrieving pages having layout and/or content which meet the user-defined query is the goal of this work. Particularly, we are interested in finding the transition pages in a book (e.g. title pages of chapter, end-of-chapter and end-of-act) in order to help the extraction of its table of content or finding pages containing a particular content component or a group of patterns (e.g. ornaments, illustrations and drop caps) in order to investigate and analyze parts of historical document images or to acquire all drawings in one or more ancient books to build an extensive database of drawings. Having graph-based signatures, a user-defined query can be formulated with a pattern graph. The idea is to identify whether or not the defined pattern graph is isomorphic to one or many subgraphs in the target graphs (*i.e.* book page signatures). In this context, a graph-matching algorithm is obviously required to measure the similarity level between the user-defined query formulated in terms of a pattern graph and the different subgraphs of the

book page signatures. Then, it is possible to order all book pages based on the similarity criterion, and finally, to present to the user book pages similar to the user-defined query.

Nevertheless, there is awareness that maybe there are non-relevant information when using a RAG to characterize the content and the layout of a book page due to the segmentation errors (occurred when analyzing texture features and connected components with an adaptive run-length smoothing algorithm for extracting representative homogeneous regions) and also to the particularities of historical documents (e.g. noise and degradation). Furthermore, it is considered that the user-defined query can contain specific layout structure and/or typographic/graphical characteristics (e.g. an italic textual block on the right side of a drop cap), which must meet the structural representations of the target book pages. Considering the particularities of historical documents, many regions in the involved document images may not have been extracted accurately. Indeed, we are facing the problem of under-segmentation or over-segmentation, which could lead to non-detection of one or more content elements.

Consequently, the graph-matching algorithm used to measure the similarity level between the user-defined query and the different subgraphs must tolerate differences regarding the values of the vertex and the edge attributes in the graphs and should be robust to both structural and attribute distortions. The measure proposed by Lerouge *et al.* [9] meets the requirements mentioned above. This measure which is modeled as an objective function of a binary linear program, is based on the search for the subgraph in the target that minimizes a matching cost with the pattern graph (*i.e.* an optimization problem). The matching cost to minimize is the sum of the costs of the elementary matching, vertex to vertex and edge to edge. The elementary matching cost corresponds to an increasing function of the difference between the vertex attributes or the edge attributes.

IV. Evaluation and results

In this section, we will detail our experimental protocol and the experiments carried out to evaluate the use of a graph-based signature for page retrieval in digitized historical books. First, our experimental corpus will be detailed in Section IV-A. Then, we will specify the configuration of the function costs associated with the different elementary edit operations used to match user-defined query and the graph-based signature (*cf.* Section IV-B). Subsequently, we will present the used performance measures in Section IV-C. Finally, we will discuss the obtained results in Section IV-D.

A. Experimental corpus

Experiments in this work are conducted by using two historical books (*Book 1* and *Book 2*) which have been collected from two different digital libraries: Gallica[1] and *Centre d'Etudes Supérieur de la Renaissance*[2] (CESR). Gallica is a digital library of the French national library "bibliothèque nationale de France" (BnF)[3]. CESR is a French training and research centre of François Rabelais university. *Book 1* and *Book 2* are two printed monographs. The pages of *Book 1* and *Book 2* which are one-page gray-scale images have been digitized at 300 dpi and saved in the *TIFF* format.

- The first book (*Book 1*)[4] has been collected from Gallica[1]. It is composed of 64 images, including 6 title pages of chapter (*cf.* Figure 2(a)), 13 end-of-chapter pages (*cf.* Figure 2(b)), 2 end-of-act pages (*cf.* Figure 2(c)), and the remaining of *Book 1* pages are sequences of pages mainly containing a textual content (the most common or frequent pages that have similar layout and/or content), *cf.* Figure 2(d)).
- The second book (*Book 2*)[5] has been collected from CESR[2]. It is composed of 297 images, including 5 title pages of act (*cf.* Figure 2(e)), 66 title pages of chapter (*cf.* Figure 2(f)), 2 pages containing only drawings (*cf.* Figure 2(g)), and the remaining of *Book 2* pages are sequences of pages mainly containing a textual content (the most common or frequent pages that have similar layout and/or content), *cf.* Figure 2(h)).

The choice of these two historical books is based on the fact that it fits the needs of historians and digitization professionals by firstly providing a computer-assistance tool for generating automatically a table of content (by detecting the chapter breakups), and by secondly developing a reliable system ensuring the automatic identification of pages containing graphics such as drop caps or an arrangement of several graphic elements such as drop caps and ornaments.

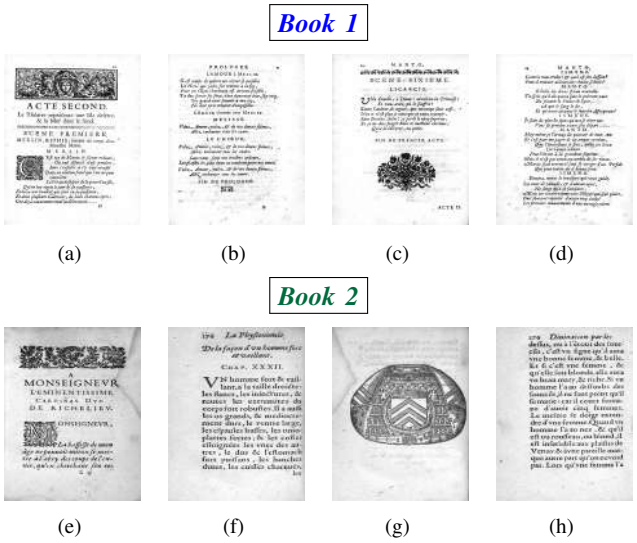


Fig. 2. Examples of *Book 1* and *Book 2* pages.

First of all, 52 user-defined queries (28 and 24 for *Book 1* and *Book 2*, respectively) have been set up to evaluate the use of a graph-based signature for page retrieval in digitized

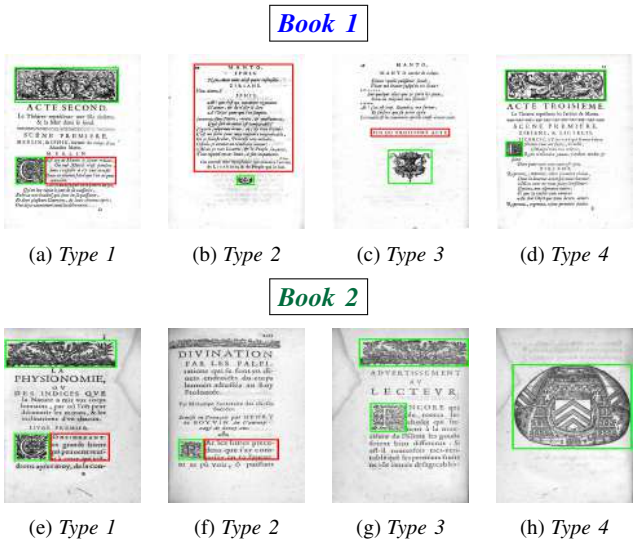
[1]http://gallica.bnf.fr
[2]http://cesr.univ-tours.fr/
[3]http://www.bnf.fr/fr/acc/x.accueil.html
[4]http://gallica.bnf.fr/ark:/12148/bpt6k840383d/f1.planchecontact.r=
[5]http://www.bvh.univ-tours.fr/Consult/index.asp?numfiche=82

historical books. These user-defined queries have been formulated in the form of simple or composite pattern graphs. Indeed, these user-defined queries reflect the need to retrieve particular pages containing a particular content element or an arrangement of several content elements. Then, we identify book pages considered relevant for each of these user-defined queries. The different user-defined queries evaluated in this article are presented in Table I and Figure 3. The selected content elements characterizing the user-defined query are marked with colored rectangles (red for the textual content and green for the graphical one).

TABLE I
USER-DEFINED QUERIES EVALUATED IN THIS ARTICLE.

Query	Description
28 graph-based queries related to the ***Book 1***	
Type 1	6 graph-based queries having 3 vertices: an ornament, a drop cap and a textual block on the right side of the drop cap to identify the title pages of chapter (*cf.* Figure 3(a)).
Type 2	2 graph-based queries having 2 vertices: a large textual block and a small size graphic area at the bottom of the page to identify the end-of-chapter pages (*cf.* Figure 3(b)).
Type 3	2 graph-based queries having 2 vertices: a textual block and an illustration below this textual block to identify the end-of-act pages (*cf.* Figure 3(c)).
Type 4	6 graph-based queries having 2 vertices: a drop cap and an ornament (*cf.* Figure 3(d)).
Type 5	6 graph-based queries having a single vertex: a drop cap.
Type 6	6 graph-based queries having a single vertex: an ornament.
24 graph-based queries related to the ***Book 2***	
Type 1	4 graph-based queries having 3 vertices: an ornament, a drop cap, and a textual block on the right side of the drop cap to identify the title pages of act (*cf.* Figure 3(e)).
Type 2	4 graph-based queries having 2 vertices: a drop cap and a textual block on the right side of the drop cap (*cf.* Figure 3(f)).
Type 3	4 graph-based queries having 2 vertices: a drop cap and an ornament (*cf.* Figure 3(g)).
Type 4	2 graph-based queries having a single vertex: an illustration (*cf.* Figure 3(h)).
Type 5	5 graph-based queries having a single vertex: a drop cap.
Type 6	5 graph-based queries having a single vertex: an ornament.

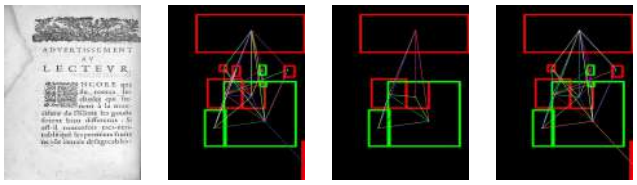
Next, for each page of our experimental corpus three different graph-based signatures have been generated as described in Section II (*cf.* Figure 4). These three graph-based signatures differ in the maximum number of the retained vertices in order to investigate the influence of the maximum number of vertices of the target graph corresponding to the user-defined queries on the performance. First, we keep all the vertices in the target graph (*cf.* Figure 4(b)). Then, for the two other graph-based representations, we only retain the vertices corresponding to the n most representative regions in a book page (*i.e.* by selecting maximum n vertices with the most number of pixels from the graph-based signature illustrated in Figure 4(b)). The two other graph-based representations are obtained with values of n equal to 5 (*cf.* Figure 4(c)) and 10 (*cf.* Figure 4(d)). Therefore, in total 156 (*i.e.* $156 = (28+24)*3$) user-defined queries have been evaluated in this article.



(a) *Type 1* (b) *Type 2* (c) *Type 3* (d) *Type 4*

(e) *Type 1* (f) *Type 2* (g) *Type 3* (h) *Type 4*

Fig. 3. Illustration of the user-defined queries evaluated in this article.



(a) Input image (b) Signature (c) 5 vertices (d) 10 vertices

Fig. 4. Illustration of three graph-based signatures differing the maximum number of the retained vertices. (b) depicts a generated graph-based signature corresponding to the input image (a). (c) and (d) illustrate two generated graph-based signatures by selecting only 5 and 10 vertices with the most number of pixels among those in (d), respectively.

B. Cost function configuration

In our experiments, we use the *GEM++*, which is a tool for solving substitution-tolerant subgraph isomorphism to estimate the matching cost between a pattern graph and a target graph [9]. The matching cost must be set by the cost functions associated with the substitution or deletion elementary edit operations of vertices and edges. The configuration of the cost functions used in our experiments will be described in the following.

Based on a statistical study, the standard deviation (σ_i) of each extracted feature composing the attributes of the vertices and the edges is estimated (*cf.* Section II). Then, the matching cost between two vertices or two edges is calculated as the Euclidean distance between the vertex or edge attributes after normalizing each dimension by a $1/\sigma_i$ and weighting with $1/N$, where N denotes the dimension of feature vector of vertices or edges. The values of N are equal to 22 and 3 for vertices and edges, respectively. Normalizing the values of vertex and edge attributes by a $1/\sigma_i$ aims to provide equivalent weight to each extracted feature in the feature vector, whereas the $1/N$ weight aims to give to a pair of vertices a weight identical to that of the edges. The cost associated with the creation of vertices and edges has been fixed empirically at a

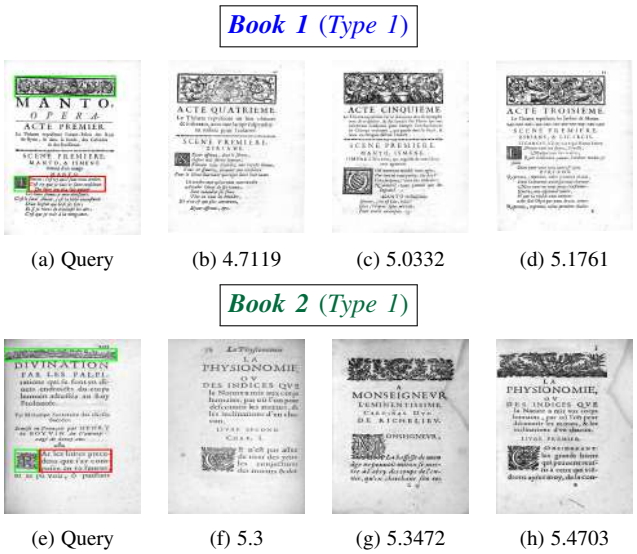
constant value equal to 3 if the normalized Euclidean distance between two attribute values is less than 3 times the value of σ_i in order to prioritize a substitution to a deletion.

C. Performance measures

After configuring the cost functions and measuring the matching costs between each user-defined query (formulated in terms of a pattern graph) and the different book page signatures (formulated in terms of target graphs) by means of the *GEM++* tool, a list of book pages sorted in ascending order of matching costs has firstly returned in order to determine the most similar retrieved pages according to the user-defined query. Indeed, the lower the values of the computed matching costs between the user-defined query (formulated in terms of a pattern graph) and the different book page signatures (formulated in terms of target graphs), the more the retrieved pages are similar to the user-defined query. Then, we have gone through this ranked list by evaluating to each element the value of the precision and the recall. Afterwards, in order to evaluate the obtained ranked retrieval results a precision-recall curve has been plotted for each of the user-defined queries. To compute the average of these curves on the set of the evaluated user-defined queries in this article the curves of the interpolated precision as a function of the recall have been subsequently plotted. Having an interpolated precision value for all the recall values, it is then possible to average the curves obtained for each user-defined query. In this paper, the *AUC* performance is computed to evaluate the proposed page retrieval system in digitized historical books.

D. Results

A visual comparison of the result images of the retrieved pages from *Book 1* and *Book 2* given by an example of *Type 1* user-defined query is illustrated in Figure 5. Figures 5(a) and 5(e) illustrate two examples of *Type 1* user-defined query for *Book 1* and *Book 2*, respectively. Figures 5(b), 5(c) and 5(d) depict the first three result images of the retrieved pages from *Book 1* given by an example of *Type 1* user-defined query. Figures 5(f), 5(g) and 5(h) illustrate the first three result images of the retrieved pages from *Book 2* given by an example of *Type 1* user-defined query. The obtained matching cost values by using the *GEM++* software are presented at the bottom of each result image. By visual inspection of the obtained results, we observe that the proposed page retrieval system provides satisfying results in identifying title pages of chapter (*i.e.* pages containing an ornament, a drop cap and a textual block on the right side of the drop cap) in digitized historical books. Nevertheless, we note that the proposed page retrieval system returns a book page containing only a drop cap (*cf.* Figure 5(f)) in the case of an example of *Type 1* user-defined query (graph-based queries having 3 vertices: an ornament, a drop cap, and a textual block on the right side of the drop cap to identify the title pages of act) for *Book 2* (*cf.* Figure 5(e)). This can be explained by the fact that in certain cases, the graph-based signature confuses the uppercase text and the graphical components (e.g. ornament) and the query and result pages are visually too similar pages.



(a) Query (b) 4.7119 (c) 5.0332 (d) 5.1761

(e) Query (f) 5.3 (g) 5.3472 (h) 5.4703

Fig. 5. Examples of user-defined query results obtained by the proposed page retrieval system based on a graph-based signature.

Since the way of comparing visually the effectiveness of a proposed method is inherently a subjective evaluation, we have chosen to compute the *AUC* as a measure of performance evaluation of the proposed page retrieval system in digitized historical books. Table II summarizes the results of 52 user-defined queries (28 and 24 for *Book 1* and *Book 2*, respectively) by computing the *AUC* average for each type of user-defined query and by varying the maximum number of the retained vertices in order to investigate the influence of the number of vertices on the performance. Figure 6 illustrates the curves of the interpolated precision-recall regarding the maximum number of the retained vertices in the target graphs.

By varying the maximum number of the retained vertices of the target graphs, we note that the proposed page retrieval system is robust since the average performance loss in terms of the *AUC* measure remains low ($\approx 6\%$ and $\approx 0.7\%$ for *Book 1* and *Book 2*, respectively and $\approx 3.3\%$ for two-book average). However, the performances of *Type 5* and *Type 6* user-defined queries for *Book 1* remain stable by varying the maximum number of the retained vertices in the target graphs. This can be justified by the simplicity of the user-defined queries since they have a single vertex representing either a drop cap or an ornament. On the other hand, for more complex user-defined queries such as *Type 1* queries, we note a loss of performance equal to 15.9% by tuning the maximum number of the retained vertices in the target graphs (*i.e.* from structural representations of target graphs retaining all vertices to those retaining only 5 vertices with the most number of pixels from the graph-based signatures). We also show a drop in the average *AUC* value equal to 14.1% when evaluating the proposed page retrieval on *Book 2* comparing to *Book 1*. This can be justified by the produced bias when generating the graph-based signatures due to the particularities related to the digitization of the *Book*

2 pages (uneven lighting due to a folding). In Figure 7, we note that the ornament has not been extracted accurately due the presence of the noise caused by uneven lighting due to a folding.

Furthermore, we note that the composite user-defined queries having three vertices for both books have the best performance. This strengthens the importance of using a signature in the form of a directed attributed graph for the user-defined queries to highlight the interactive relationship between the different content elements composing a book page. In conclusion, the overall results are encouraging since we obtain a two-book *AUC* average equal to 0.74 (*cf.* Figure 6). The experimental results have shown the accuracy and the robustness of the proposed page retrieval system in digitized historical system.

TABLE II
AUC PERFORMANCES.

	5 vertices	10 vertices	All vertices
Book 1			
Type 1	0.911	0.819	0.752
Type 2	0.692	0.712	0.627
Type 3	0.870	0.870	0.870
Type 4	0.924	0.909	0.793
Type 5	0.915	0.915	0.911
Type 6	0.927	0.927	0.927
Average (*Book 1*)	**0.873**	**0.859**	**0.813**
Book 2			
Type 1	0.758	0.746	0.740
Type 2	0.740	0.763	0.740
Type 3	0.733	0.706	0.717
Type 4	0.562	0.556	0.555
Type 5	0.574	0.573	0.566
Type 6	0.705	0.716	0.715
Average (*Book 2*)	**0.679**	**0.676**	**0.672**
Average (two books)	**0.776**	**0.768**	**0.743**

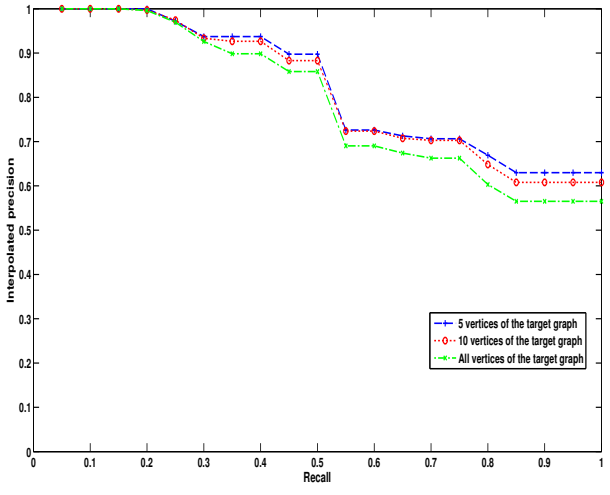


Fig. 6. Curves of the interpolated precision-recall.

(a) Input image (b) Pixel-labeled (c) Graph

Fig. 7. Example of a graph-based signature for a *Book 2* page.

V. Conclusions and Further Work

This paper evaluates the use of a graph-based signature and an error-tolerant subgraph matching algorithm to retrieve pages in digitized historical books. Many simple and composite user-defined queries based on the proposed signature have been evaluated. These user-defined queries meet the user's need in terms of identifying the transition pages (e.g. title pages of chapter, end-of-chapter and end-of-act) and pages containing a particular content component or a group of patterns (e.g. ornaments, illustrations and drop caps) in digitized historical books. The obtained experimental results show the robustness of the proposed page retrieval system in digitized historical books.

Our further works will be in line with those described in this article. The first aspect of future work will be to evaluate the proposed page retrieval system on other digitized historical books in order to help improve its scalability. In addition, we intend to study the influence of graph matching costs on the performance. Furthermore, we will develop a graphical user interface tool for content-based image retrieval by introducing several advanced human-computer interaction techniques to optimize the way in which the historians and the digitization professionals can define the user-defined queries.

References

[1] J. Y. Ramel, S. Leriche, M. L. Demonet, and S. Busson, "User-driven page layout analysis of historical printed books," *IJDAR*, pp. 243–261, 2007.

[2] B. Coüasnon, "DMOS, a generic document recognition method: application to table structure analysis in a general and in a specific way," *IJDAR*, pp. 111–122, 2006.

[3] 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.

[4] M. Mehri, P. Héroux, J. Lerouge, P. Gomez-Krämer, and R. Mullot, "A structural signature based on texture for digitized historical book page categorization," in *ICDAR*, 2015, pp. 116–120.

[5] A. Antonacopoulos, S. Pletschacher, D. Bridson, and C. Papadopoulos, "ICDAR 2009 page segmentation competition," in *ICDAR*, 2009, pp. 1370–1374.

[6] J. Liang, D. Doermann, M. Ma, and J. K. Guo, "Page classification through logical labelling," in *ICPR*, 2002, pp. 477–480.

[7] H. Bunke and K. Riesen, "Recent advances in graph-based pattern recognition with applications in document analysis," *PR*, pp. 1057–1067, 2011.

[8] M. Mehri, P. Héroux, P. Gomez-Krämer, and R. Mullot, "Texture feature benchmarking and evaluation for historical document image analysis," *IJDAR*, pp. 1–35, 2017.

[9] J. Lerouge, M. Hammami, P. Héroux, and S. Adam, "Minimum cost subgraph matching using a binary linear program," *PRL*, vol. 71, pp. 45–51, 2016.