HBA 1.0: A Pixel-based Annotated Dataset for Historical Book Analysis

Maroua Mehri, Pierre Héroux, Rémy Mullot, Jean-Philippe Moreux, Bertrand Coüasnon, Bill Barrett

HBA 1.0: A Pixel-based Annotated Dataset for Historical Book Analysis

Maroua Mehri*†, Pierre Héroux†, Rémy Mullot‡, Jean-Philippe Moreux§, Bertrand Coüasnon¶ and Bill Barrett‖
*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
§Digitization service, Bibliothèque nationale de France, Paris, 75706, France
¶Intuidoc, IRISA, Avenue du Gnral Leclerc, 35042, Rennes, France
‖Brigham Young University, 84602, Provo, UT, USA
Emails: maroua.mehri@gmail.com, pierre.heroux@univ-rouen.fr, remy.mullot@univ-lr.fr,
jean-philippe.moreux@bnf.fr, bertrand.couasnon@irisa.fr and barrett@cs.byu.edu

***Abstract*—This paper introduces HBA 1.0, a representative pixel-based annotated dataset which is released at the ICDAR2017 Competition on Historical Book Analysis (HBA2017). The HBA 1.0 dataset is composed of 4,436 real scanned ground truthed historical document images from 11 books (5 manuscripts and 6 printed books) in different languages and scripts published between the 13th and 19th centuries. The HBA 1.0 dataset contains 2,435 and 2,001 manuscript and printed pages, respectively. The ground truth of the HBA 1.0 dataset contains more than 7,58 billion annotated pixels. The HBA 1.0 dataset addresses a thriving topic of major interest of many researchers in different fields including (historical) document image analysis, image processing, pattern recognition and classification. The HBA 1.0 dataset and its ground truth can be used to evaluate the capabilities of image analysis methods to discriminate the textual content from the graphical ones on the one hand, and to separate the textual content according to different text fonts (e.g. lowercase, uppercase, italic) on the other hand. Evaluation results of a state-of-the-art pixel-labeling method on the HBA 1.0 dataset are reported and discussed in this paper in order to provide a benchmark/baseline for future evaluation studies and to showcase the intended use of the HBA 1.0 dataset.**

***Index Terms*—Historical book collection, Layout analysis, Pixel-labeling, Annotated document images, Ground truth, Pixel level.**

I. Introduction

Providing reliable computer-based access and analysis of cultural heritage documents has been flagged as a very important need for the library and the information science community, spanning educationalists, students, practitioners, researchers in book history, computer scientists, historians, librarians, end-users and decision makers. More specifically, there is a consistent and clear need for robust and accurate document image analysis (DIA) methods that deal with the idiosyncrasies of historical document images (HDIs) [1]. Indeed, historical DIA remains an open issue due to the particularities of HDIs, such as the superimposition of information layers (e.g. stamps, handwritten notes, noise, back-to-front interference, page skew) and the variability of their contents and/or layouts. Moreover, analyzing HDIs and characterizing their layouts and contents under significant degradation levels and different noise types and with no *a priori* knowledge about the layout, content, typography, font styles, scanning resolution or DI size, *etc.* is not a straightforward task. As a consequence, researchers specialized in historical DIA keep proposing novel reliable approaches and rigorous techniques for historical DIA, segmentation and characterization.

Nevertheless, many important issues arise to provide an informative benchmarking of historical DIA methods such as the lack of a common dataset of HDIs and the lack of the appropriate quantitative evaluation measures. Moreover, many researchers have addressed the need of a good dataset. Antonacopoulos *et al.* [1] considered a dataset as a good one if it is realistic (*i.e.* it must be composed of real digitized document images), comprehensive (*i.e.* it must be well characterized and detailed for ensuring in-depth evaluation) and flexibly structured (*i.e.* to facilitate a selection of sub-sets with specific conditions). Although the issue of the realistic dataset availability and the broadband access to researchers for the performance evaluation of contemporary document images have been discussed and solved by Antonacopoulos *et al.* [1], representative datasets of HDIs with their associated ground truths are currently hardly publicly accessible for HDI layout analysis. Finding a large corpus of HDIs having many annotated HDIs with various content and layout characteristics is still a challenging issue for HDI layout analysis. This is mainly due to the intellectual and industrial property rights. Another challenge facing founding a representative dataset of HDIs concerns the definition of its objective and complete associated ground truth. Defining an objective ground truth is still a complex and burdensome task due to the mentioned particularities of HDIs. These characteristics complicate the definition of the appropriate and objective ground truth and the characterization of HDIs [1].

Therefore, we introduce in this paper HBA 1.0, a representative pixel-based annotated dataset which is released at the ICDAR2017 Competition on Historical Book Analysis - HBA2017[1]. The images composing the HBA 1.0 dataset

[1]http://icdar2017hba.litislab.eu/

were collected from the French digital library Gallica[2]. The HBA 1.0 dataset is publicly available for scientific use and on request subject to the agreement from the French national library "Bibliothèque nationale de France" (BnF)[3].

The main contributions of this paper are as follows. A representative and complete pixel-based annotated dataset containing 4,436 real scanned ground truthed HDIs from 11 books and more than 7,58 billion annotated pixels is presented. To the best of our knowledge, there is no benchmark pixel-based annotated dataset comprising a sufficiently large collection of HDIs and a representative ground truth data that can be easily exploited by many researchers working in different fields including (historical) DIA, image processing, pattern recognition and classification. Furthermore, experiments have been carried out with a standardized benchmark protocol along with a state-of-the-art pixel-labeling method on the HBA 1.0 dataset in order to provide a benchmark for future studies.

The remainder of this article is organized as follows. Section II reviews the existing datasets and annotations for historical DIA and recognition with a particular focus on those related to historical document layout analysis. Section III describes the HBA 1.0 dataset. Section IV details the evaluation protocol used to assess a state-of-the-art pixel-labeling method on the HBA 1.0 dataset by outlining the ground truth, the experimental protocol, the accuracy metric and the obtained evaluation results. Finally, our conclusions and future work are presented in Section V.

II. State-of-the-art datasets

A large number of datasets and annotations has recently been devoted for document analysis and recognition [2]. Nevertheless, a limited number of public datasets of HDIs and their associated ground truths are freely available for layout analysis, historical handwriting recognition or word spotting (e.g. George Washington[4], Parzival[5], Saint Gall[6], RODRIGO[7], Montesquieu's and Flaubert's manuscripts[8], Vesalius's manuscripts[9], ESPOSALLES[10], BH2M[11], IAM-HistDB[12], GRPOLY-DB[13], HBR[14], DIGIDOC-Texture[15], DIVA-HisDB[16]) [1]. Those datasets are being used in the context of different research projects to deal with handwritten documents of inheritance by developing innovative techniques and proposing different approaches. It is obviously necessary to note the unavailability or lack of a standard public large dataset of HDIs and its associated ground truth. Moreover, most available datasets contain only handwritten HDIs. Table I presents a list of existing public datasets and annotations dedicated to HDI layout analysis.

TABLE I
Datasets and annotations dedicated to HDI layout analysis.

Dataset	Characteristics	Number of pages
George Washington[4] [3]	-Two writers -18th century -English language -Longhand script -Ink on paper	20
Parzival[5] [4]	-Three writers -13th century -Medieval German language -Gothic script -Ink on parchment	47
Saint Gall[6] [5]	-Single writer -9th century -Latin language -Carolingian script -Ink on parchment	60
RODRIGO[7] [6]	Single writer	853
Montesquieu's and Flaubert's manuscripts[8] [7]	Handwritten HDIs from the 18th and 19th centuries	500
Vesalius's manuscripts[9] [8]	Rare DHBs	85
Unspecified [9]	Damaged military form pages of the 19th century	88,745
ESPOSALLES[10] [10]	Marriage license book which was written between 1617 and 1619 by a single writer	202
BH2M[11] [11]	Manuscripts from the 17th century	174
IAM-HistDB[12] [5]	-Parzival[5] -Saint Gall[6] -George Washington[4]	-74 -60 -20
GRPOLY-DB[13] [12]	Machine-printed and handwritten old Greek documents	399
HBR[14] [1]	Printed documents of various types in 25 languages from the 17^{th} century to the early 20^{th} century	100
DIGIDOC-Texture[15] [13]	Images selected from several books of Gallica[2]	1,000
DIVA-HisDB[16] [14]	3 medieval manuscripts	150

III. HBA 1.0 description

The HBA 1.0 dataset is composed of 4,436 real scanned ground truthed one-page HDIs from 11 books (5 manuscripts and 6 printed books) in different languages and scripts published between the 13th and 19th centuries. The HBA 1.0 dataset contains 2,435 and 2,001 manuscript and printed pages, respectively. The selected documents are gray-scale/color images which were digitized at 300/400 dpi and saved in the *TIFF* format which provides a high resolution of digitized images. The images composing the HBA 1.0 dataset were

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

[3]http://www.bnf.fr/fr/acc/x.accueil.html

[4]http://memory.loc.gov/ammem/gwhtml/gwhome.html

[5]http://www.iam.unibe.ch/fki/databases/iam-historical-document-database/parzival-database

[6]http://www.iam.unibe.ch/fki/databases/iam-historical-document-database/saint-gall-database

[7]https://www.prhlt.upv.es/page/projects/multimodal/idoc/rodrigo

[8]http://www.bovary.fr/folios_liste.php?type=f&id=4&mxm=0101030105&recueil=1&page=25&nb=24

[9]http://www.bvh.univ-tours.fr/Consult/index.asp?numfiche=56&url=/resrecherche.asp?ordre=titre-motclef=andre\%20vesale-bvh=BVH-epistemon=Epistemon

[10]http://dag.cvc.uab.es/the-esposalles-database

[11]http://dag.cvc.uab.es/the-historical-marriages-database

[12]http://www.iam.unibe.ch/fki/databases/iam-historical-document-database/

[13]http://users.iit.demokritos.gr/~nstam/GRPOLY-DB/

[14]http://www.primaresearch.org/datasets

[15]http://litis-digidoc.univ-rouen.fr/texture/DIGIDOC-Texture.tar.gz

[16]http://diuf.unifr.ch/main/hisdoc/diva-hisdb

collected from the French digital library Gallica[2]. The HBA 1.0 dataset is publicly available for scientific use and on request subject to the agreement from the French national library "Bibliothèque nationale de France" (BnF)[3], but as it remains exclusive property of the BnF. It is released at the ICDAR2017 Competition on Historical Book Analysis (HBA2017) and hosted on a server maintained by the HBA2017 competition organizers[1]. Table II presents the 11 books of the HBA 1.0 dataset: noitemsep,nolistsep,itemindent=0pt,leftmargin=0.1in

- *Book 1* has been collected from Gallica[17] (*cf.* Figure 1(a)). It is titled "Plutarchus, Vitæillustrium virorum", written in Latin by Italian copyist and published in 1743-1774.
- *Book 2* has been collected from Gallica[18] (*cf.* Figure 1(b)). It is titled "Justinien, Institutes", written in French by at least two copyists and published in 1342.
- *Book 3* has been collected from Gallica[19] (*cf.* Figure 1(c)). It is titled "Girart d'Amiens, Meliacin ou le Cheval de fust", written in French by at least three copyists and published in 1285.
- *Book 4* has been collected from Gallica[20] (*cf.* Figure 1(d)). It is titled "Chronique, histoires de la Bible, Vies de saints, Sermons de MAURICE DE SULLY", written in French and published in 1201-1300.
- *Book 5* has been collected from Gallica[21] (*cf.* Figure 1(e)). It is titled "Memoire relatif à la carte du Guipuscoa", written in French and published in 1758.
- *Book 6* has been collected from Gallica[22] (*cf.* Figure 1(f)). It is titled "Il mondo nuovo, del sig. Giov. Giorgini da Jesi", written in Italian and published in 1596.
- *Book 7* has been collected from Gallica[23] (*cf.* Figure 1(g)). It is titled "Manto la Fée, opéra", written in French by Mennesson and published in 1711.
- *Book 8* has been collected from Gallica[24] (*cf.* Figure 1(h)). It is titled "Le Mirouer de la redemption de l'umain lignage", written in French and published in 1478-1480.
- *Book 9* has been collected from Gallica[25] (*cf.* Figure 1(i)). It is titled "Cy commencent le Procès de Belial à l'encontre de Jhésus", written in French by Jacques de Teramo and published in 1481.
- *Book 10* has been collected from Gallica[26] (*cf.* Figure 1(j)). It is titled "Voyage pittoresque de la Grèce", written in French by Marie-Gabriel-Florent-Auguste de Choiseul-Gouffier and published in 1782-1822.
- *Book 11* has been collected from Gallica[27] (*cf.* Figure 1(k)). It is titled "La Chartreuse de Parme", written in French by Stendhal and published in 1839.

[17]http://gallica.bnf.fr/ark:/12148/btv1b8446958b/f1.planchecontact.r
[18]http://gallica.bnf.fr/ark:/12148/btv1b8447185s/f1.planchecontact.r
[19]http://gallica.bnf.fr/ark:/12148/btv1b8447872k/f1.planchecontact.r
[20]http://gallica.bnf.fr/ark:/12148/btv1b90075392/f1.planchecontact.r
[21]http://gallica.bnf.fr/ark:/12148/btv1b55005693r/f1.planchecontact.r
[22]http://gallica.bnf.fr/ark:/12148/bpt6k132294p/f1.planchecontact.r
[23]http://gallica.bnf.fr/ark:/12148/bpt6k840383d/f1.planchecontact.r
[24]http://gallica.bnf.fr/ark:/12148/btv1b7300026v/f1.planchecontact.r
[25]http://gallica.bnf.fr/ark:/12148/btv1b73000367/f1.planchecontact.r
[26]http://gallica.bnf.fr/ark:/12148/btv1b8449081d/f1.planchecontact.r
[27]http://gallica.bnf.fr/ark:/12148/btv1b8623296m/f1.planchecontact.r

TABLE II
COMPOSITION OF THE HBA 1.0 DATASET.

	Number of pages	Book type	Image type
Book 1	730	Manuscript	Color
Book 2	486	Manuscript	Color
Book 3	350	Manuscript	Color
Book 4	813	Manuscript	Gray-scale
Book 5	56	Manuscript	Color
Book 6	322	Printed	Color
Book 7	64	Printed	Gray-scale
Book 8	403	Printed	Gray-scale
Book 9	341	Printed	Color
Book 10	440	Printed	Color
Book 11	431	Printed	Color

The 11 URL links of the books of the HBA 1.0 dataset mentioned above will bring you to the French digital library Gallica[2] where only low resolution images are publicly available online. The 11 books of the HBA 1.0 dataset follow very different formatting rules (e.g. manuscript/printed, layout, language, script, character fonts, font sizes). The characteristics of the HBA 1.0 dataset are primarily: strong heterogeneity, with differences in layout, typography, illustration style, historic fonts, complex layouts (e.g. dense printing, irregular spacing, varying text column widths, marginal notes), ink shining through and historical spelling variants, *etc.* In addition to this specificity, the issues affecting document image layout analysis, such as the degradation properties (e.g. yellow pages, ink stains, back-to-front interference) and scanning defects (e.g. defects of curvature and light) are adequately covered. It is worth noting that the HDIs of the HBA 1.0 dataset were selected so as to be as realistic as possible, in order to reflect that there is still much room for improvement in HDI layout analysis due to the mentioned particularities of HDIs. Figure 1 illustrates few samples of book pages of the HBA 1.0 dataset.

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

(e) ***Book 5*** (f) ***Book 6*** (g) ***Book 7*** (h) ***Book 8***

(i) ***Book 9*** (j) ***Book 10*** (k) ***Book 11***

Fig. 1. Sample book pages of the HBA 1.0 dataset.

IV. Historical book analysis

The research community is continuing to propose automatic systems able to discriminate and/or recognize the content type (text or graphic) and/or the type of fonts (e.g. lowercase, uppercase, italic). Indeed, providing relevant information about the content type improves the optical character recognition (OCR) performance and learns how to tune the OCR parameters automatically for different content types. Therefore, two DIA challenges can be evaluated by using the HBA 1.0 dataset. The first challenge is interested in raising issues related only to how image analysis methods are performed for discriminating the textual content from the graphical ones. However, the second challenge evaluates the capabilities of image analysis methods to firstly distinguish between text and graphic, and secondly to separate the textual content according to different text fonts (e.g. lowercase, uppercase, italic). Therefore, we consider when using the HBA 1.0 dataset a HDI layout analysis challenge as a pixel-labeling task (*i.e.* each pixel is classified as one of the predefined classes). Indeed, in the first challenge a binary classification task can be evaluated, while in the second challenge a multi-class classification task can be assessed. In the following, we detail the evaluation protocol by presenting the defined ground truth, the experimental protocol, the used accuracy metric for performance evaluation and the evaluation results of using a state-of-the-art pixel-labeling method on the HBA 1.0 dataset [15].

A. Ground truth construction

The annotation process of the HBA 1.0 dataset was defined and reviewed manually. First, the ground truth has been manually outlined using rectangular regions drawn around each selected zone. The regions have been ground-truthed by zoning each content type (*i.e.* each rectangular region has been classified into text or graphics). Different labels for regions with different fonts have been also assigned for evaluating the participating methods to separate various text fonts. Then, the foreground pixels have been retrieved from the analyzed historical document image. Afterward, only the foreground pixels defined on the outlined rectangular regions have been selected. Each selected foreground pixel has been labeled according to the content type deduced from the label of the corresponding rectangular region. A set of classes has been defined for each book of the dataset. The ground truth of the HBA 1.0 dataset is currently available at pixel level. We have ground truthed the HBA 1.0 dataset by annotating each foreground pixel. The ground truth of each selected foreground pixel has been defined by means of a label indicating the content type or the content class of the analyzed HDI. Different labels for the selected foreground pixels with different fonts were also assigned for evaluating image analysis methods to separate various text fonts. Table III presents the set of classes used in the annotation process of the HBA 1.0 dataset. Figure 2 illustrates few examples of the defined ground truth of the 11 books of the HBA 1.0 dataset. Each selected foreground pixel is marked by a color that symbolizes the corresponding content type (*i.e.* the predefined class). Table IV details the distribution of the annotation classes of the HBA 1.0 dataset. The ground truth of the HBA 1.0 dataset contains more than 7,58 billion annotated pixels. The proportions of the whole annotated pixels per book are 13.01%, 19.25%, 5.07%, 30.98%, 0.15%, 2.25%, 0.32%, 9.77%, 3.63%, 14% and 1.56% for *Book 1-11*, respectively. For the whole HBA 1.0 dataset, 82.01% of the total number of the annotated pixels represent textual content (*i.e.* the sum of the five following classes, *Class 2-6*), while 17.99% are considered as graphical content (*i.e. Class 1*). 80.33%, 0.79%, 0.27%, 0.36% and 0.26% of the total number of the annotated pixels represent *Class 2-6*, respectively. The proportions of the content classes differ between the 11 books of the HBA 1.0 dataset.

TABLE III
Classes of the HBA 1.0 ground truth.

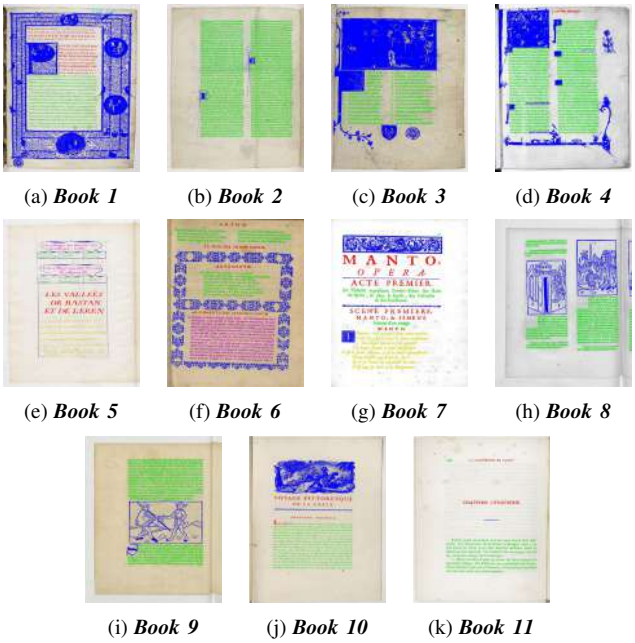
	Description
Class 1	Graphics
Class 2	Main text body
Class 3	Capitalized text
Class 4	Handwritten text
Class 5	Italic text
Class 6	Footnote text

TABLE IV
Composition of the HBA 1.0 ground truth.

	Class 1	*Class 2*	*Class 3*	*Class 4*	*Class 5*	*Class 6*
Book 1	56,516,548	896,927,247	14,540,765	18,498,248	0	24,914
Book 2	60,003,754	1,398,226,275	1,404,820	11,053	0	0
Book 3	31,934,853	351,840,866	403,257	0	0	0
Book 4	339,913,377	1,995,762,437	11,871,050	877,544	0	0
Book 5	1,621,450	715,376	248,437	47	7,773,935	1,086,691
Book 6	14,420,477	148,455,093	1,510,113	30,698	250,128	5,578,507
Book 7	4,789,604	3,484,690	3,418,605	0	12,876,345	25,931
Book 8	115,817,914	624,525,928	204,231	245,909	0	0
Book 9	32,524,625	242,691,099	299,411	0	0	0
Book 10	705,550,447	312,477,072	23,248,336	667,350	6,426,830	12,858,856
Book 11	674,124	114,602,887	2,632,159	189,764	31,345	172,929
Overall	1,363,767,173	6,089,708,970	59,781,184	20,520,613	27,358,583	19,747,828
		6,217,117,178				
		7,580,884,351				

B. Experimental protocol

The HBA 1.0 dataset aims at evaluating methods which would automatically annotate an important number of book pages, based on a limited number of manually annotated pages of the same book. The provided limited number of manually annotated pages constitutes the training image dataset. It is, therefore, ensured that each class of content type is represented in the set of the training pages for each book. It is worth pointing out that the content classes in the HBA 1.0 dataset vary from one book to another book and have very different headcounts. Indeed, the textual content is predominant in monographs, compared to the graphical content. Moreover, among the textual content a great majority represent the body text while other character fonts are more marginal. This is compounded by the difficulty of a image analysis methods to perform on different types of content in historical books published at different eras such as printed books from the 19^{th} century or manuscripts from the 13^{th} century. This imbalanced headcounts between classes varies from one book to another book in the HBA 1.0 dataset. There is surely a great deal



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

(e) ***Book 5*** (f) ***Book 6*** (g) ***Book 7*** (h) ***Book 8***

(i) ***Book 9*** (j) ***Book 10*** (k) ***Book 11***

Fig. 2. Sample ground truthed book pages of the HBA 1.0 dataset.

of imbalanced headcounts between classes in the same book. The class headcounts in the training dataset are not thereby be similar to the test dataset. Indeed, the minority classes are adequately represented in the training dataset in order to ensure an appropriate learning task. However, unlike the minority classes, the majority classes are clearly less represented in the training dataset in comparison with the class headcounts in the test dataset. These requirements have been satisfied in the selection of the training pages of the HBA 1.0 dataset.

Each book of the HBA 1.0 dataset is composed of a set of training images and a set of test images. The training dataset contains a reduced number of book pages, along with their ground truth. The training images are representative of different contents and layouts of the book pages. On the other side, the test dataset is composed of images representing the remainder book pages. The numbers of the training images of all 11 books are 22, 42, 56, 30, 27, 42, 24, 45, 20, 26, and 32, respectively. Table V details the distribution of the annotation classes of the training pages of the HBA 1.0 dataset. The ground truth of the training pages of the HBA 1.0 dataset contains more than 520 million annotated pixels. The proportions of the annotated pixels in the training dataset per book are 8.04%, 22.63%, 12.37%, 16.41%, 1.07%, 4.24%, 2.09%, 13.95%, 2.81%, 15.14% and 1.25%. For the training subset of the HBA 1.0 dataset, 70.17% of the number of the annotated pixels in the training pages represent textual content (*i.e.* the sum of the five following classes, *Class 2-6*), while 29.83% are considered as graphical content (*i.e. Class 1*). 66.45%, 1.21%, 0.17%, 1.36% and 0.98% of the number of the annotated pixels in the training pages represent *Class 2-6*, respectively.

TABLE V
COMPOSITION OF THE GROUND TRUTH OF THE TRAINING DATASET OF THE HBA 1.0 DATASET.

	Class 1	*Class 2*	*Class 3*	*Class 4*	*Class 5*	*Class 6*
Book 1	17,925,932	22,200,446	1,224,080	520,322	0	0
Book 2	16,466,443	101,276,671	58,252	0	0	0
Book 3	16,855,612	47,523,355	0	0	0	0
Book 4	20,582,668	64,498,393	322,549	45,957	0	0
Book 5	947,011	715,376	235,232	47	2,854,301	800,595
Book 6	2,505,817	15,587,432	231,791	29,800	54,275	3,661,336
Book 7	3,873,163	1,947,797	1,468,246	0	3,541,202	25,931
Book 8	19,030,136	53,340,325	0	245,909	0	0
Book 9	3,452,965	11,155,159	23,659	0	0	0
Book 10	53,597,870	22,296,931	1,804,226	1,361	596,908	547,428
Book 11	47,388	5,412,435	952,468	30,633	31,345	58,703
Overall	155,285,005	345,954,320	6,320,503	874,029	7,078,031	5,093,993
		365,320,876				
		520,605,881				

C. Evaluation results

To provide a benchmark for future studies, a baseline pixel-labeling method presented by Mehri *et al.* in [15] is applied on the HBA 1.0 dataset. The assessed pixel-labeling method is based on integrating an unsupervised task that automatically labels content pixels with the same cluster identifier as the book content. For each book 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 book scale to estimate automatically the number of book content types. The results of using the baseline pixel-labeling method on the HBA 1.0 dataset are compared with the ground truth information and the confusion matrix is computed. From the confusion matrix, the per-pixel classification accuracy rate (CA) is calculated for each book page by dividing the number of correct labeled pixels via the total number of foreground pixels. The per-pixel classification accuracy is computed for each book of the HBA 1.0 dataset. This shows that whether an image analysis method behaves uniformly among all the books or if, conversely, it achieves a different level of performance for different books. Table VI illustrates the obtained per-pixel classification results using the baseline pixel-labeling method on few books of the HBA 1.0 dataset. The classification results of the first challenge of our experimental protocol are presented in Table VI. We note an average classification accuracy over five books of the HBA 1.0 dataset equal to 75.9%. The evaluated pixel-labeling method has achieved quite promising results.

TABLE VI
PIXEL-LEVEL CLASSIFICATION ACCURACY (CA IN %).

	Book 5	*Book 6*	***Book 7***	*Book 9*	*Book 11*
CA	0.8016	0.8288	0.7186	0.6970	0.7511

V. CONCLUSION AND FUTURE WORK

In order to address an important lack in providing a publicly available representative pixel-based annotated dataset for historical DIA, this paper introduces HBA 1.0, a novel freely accessible dataset with more than 7.58 billion annotated pixels. Using the HBA 1.0 dataset, we aim to show how low-level analysis methods will perform for discriminating the textual content from the graphical ones, and separating the textual

content according to different text fonts. First results obtained with using a baseline system on the HBA 1.0 dataset have shown that our dataset allows a consistent evaluation and a fair comparison of low-level image processing methods for historical DIA on the one hand, and that there is room for improvement in regard to the design of more reliable systems of layout analysis of cultural heritage documents on the other hand. The first aspect of future work will be to refine and complete the proposed ground truth to ensure the evaluation of page content classification at block level. Furthermore, we plan to evaluate deep learning architectures for layout analysis using the HBA 1.0 dataset.

ACKNOWLEDGMENT

The authors would like to acknowledge the BnF[3] for providing access to the Gallica digital library[2].

REFERENCES

[1] A. Antonacopoulos, C. Clausner, C. Papadopoulos, and S. Pletschacher, "ICDAR 2013 Competition on Historical Book Recognition (HBR 2013)," in *ICDAR*, 2013, pp. 1459–1463.

[2] E. Valveny, "Datasets and annotations for document analysis and recognition," *Handbook of Document Image Processing and Recognition*, pp. 983–1009, 2014.

[3] T. M. Rath and R. Manmatha, "Word spotting for historical documents," *IJDAR*, pp. 139–152, 2007.

[4] A. Fischer, A. Keller, V. Frinken, and H. Bunke, "Lexicon-free handwritten word spotting using character HMMs," *PRL*, pp. 934–942, 2012.

[5] H. Wei, K. Chen, R. Ingold, and M. Liwicki, "Hybrid feature selection for historical document layout analysis," in *ICFHR*, 2014, pp. 87–92.

[6] N. Serrano, F. Castro, and A. Juan, "The RODRIGO database," in *ICLRE*, 2010, pp. 2709–2712.

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

[8] J. Y. Ramel, S. Busson, and M. L. Demonet, "AGORA: the interactive document image analysis tool of the BVH project," in *DIAL*, 2006, pp. 145–155.

[9] B. Coüasnon, J. Camillerapp, and I. Leplumey, "Access by content to handwritten archive documents: generic document recognition method and platform for annotations," *IJDAR*, pp. 223–242, 2007.

[10] V. Romero, A. Fornés, N. Serrano, J. A. Sánchez, A. H. Toselli, V. Frinken, E. Vidal, and J. Lladós, "The ESPOSALLES database: an ancient marriage license corpus for off-line handwriting recognition," *PR*, pp. 1658–1669, 2013.

[11] D. Fernández-Mota, J. Almazán, N. Cirera, A. Fornés, and J. Lladós, "BH2M: the Barcelona historical handwritten marriages database," in *ICPR*, 2014, pp. 256–261.

[12] B. Gatos, N. Stamatopoulos, G. Louloudis, G. Sfikas, G. Retsinas, V. Papavassiliou, F. Sunistira, and V. Katsouros, "GRPOLY-DB: an old Greek polytonic document image database," in *ICDAR*, 2015, pp. 646–650.

[13] M. Mehri, P. Héroux, P. Gomez-Krämer, and R. Mullot, "Texture feature benchmarking and evaluation for historical document image analysis," *IJDAR*, pp. 325–364, 2017.

[14] F. Simistira, M. Seuret, N. Eichenberger, A. Garz, M. Liwicki, and R. Ingold, "DIVA-HisDB: A precisely annotated large dataset of challenging Medieval manuscripts," in *ICFHR*, 2016, pp. 471–476.

[15] M. Mehri, P. Gomez-Krämer, P. Héroux, A. Boucher, and R. Mullot, "A texture-based pixel labeling approach for historical books," *PAA*, pp. 325–364, 2017.