DIVA-HisDB: A Precisely Annotated Large Dataset of Challenging Medieval Manuscripts

Fotini Simistira*, Mathias Seuret*, Nicole Eichenberger†, Angelika Garz*, Marcus Liwicki* and Rolf Ingold*
*Document, Image and Voice Analysis Group (DIVA)
†Germanistische Mediävistik
University of Fribourg
Fribourg, Switzerland
firstname.lastname@unifr.ch

***Abstract*—This paper introduces a publicly available historical manuscript database DIVA-HisDB for the evaluation of several Document Image Analysis (DIA) tasks. The database consists of 150 annotated pages of three different medieval manuscripts with challenging layouts. Furthermore, we provide a layout analysis ground-truth which has been iterated on, reviewed, and refined by an expert in medieval studies. DIVA-HisDB and the ground truth can be used for training and evaluating DIA tasks, such as layout analysis, text line segmentation, binarization and writer identification. Layout analysis results of several representative baseline technologies are also presented in order to help researchers evaluate their methods and advance the frontiers of complex historical manuscripts analysis. An optimized state-of-the-art Convolutional Auto-Encoder (CAE) performs with around 95 % accuracy, demonstrating that for this challenging layout there is much room for improvement. Finally, we show that existing text line segmentation methods fail due to interlinear and marginal text elements.**

***Keywords*-medieval manuscripts; benchmarking; database; document analysis; layout analysis; Convolutional Auto-Encoder; DIVA-HisDB**

I. INTRODUCTION

In recent years several publicly available databases have been generated in order to assist researchers in the field of Handwriting Recognition (HWR) and DIA to compare their methods. As research in DIA advances rapidly, the type of tasks become more and more complex. While in the 1990s individual character and word databases were published (CENPARMI [1], CEDAR [2]), the difficulty increased to unconstrained text lines (lAM [3], George Washington [4]), and, finally, to historical documents with diverse scripts and languages (IAM-HistDB [5], GR-POLYDB [6]). To the best of our knowledge, no benchmark dataset comprising historical manuscripts with comparably complex layout elements, diverse scripts per page, and challenging degradations has been made available yet.

In this paper we introduce DIVA-HisDB, the first publicly available precisely annotated large dataset of challenging medieval manuscripts. The database consists of three medieval manuscripts, 50 pages each, resulting of in total 150 pages. The annotation process was jointly defined with experts from medieval studies and the Ground Truth (GT) was reviewed and refined by an expert as well. Furthermore, a layout analysis benchmark is specified and results of a state-of-the-art Convolutional Auto-Encoder (CAE) [7] are presented. The challenging nature of this dataset becomes obvious when considering that the method achieves only a result of around 95 % pixel-level accuracy. This is even further confirmed by showing that other state-of-the-art methods fail on the task of text line segmentation, even when being provided with the text region.

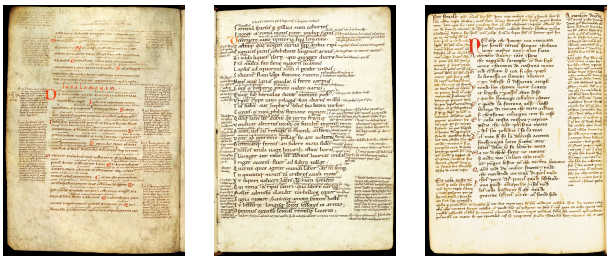
The main contributions of this paper are as follows. A complete dataset containing 150 pages is presented together with ground truth.[1]. Furthermore, experiments using deep learning for layout analysis are performed, serving as a benchmark for future studies.

The remainder of the paper is organized as follows. In Section II an overview of existing datasets is presented. In Section III a description of DIVA-HisDB is presented while an overview of the creation of the database is illustrated in Section IV. Evaluation results using several baseline methods are presented in Section V. Finally, a conclusion and an outlook of future work are given in Section VI.

II. EXISTING DATASETS

An overview of existing datasets and annotations for document analysis and recognition can be found at [9]. The U.S. ZIP code data base of CENPARMI (Concordia University) [1] contains approximately 17,000 run-length coded binarized digits. Among them are also some samples which are not easy to recognize for humans. The CEDAR database contains digital images of approximately 5,000 city names, 5,000 state names, 10,000 ZIP Codes, and 50,000 alphanumeric characters. The lAM database is a collection of handwritten English sentences from the Lancaster-Oslo/Berge (LOB) corpus and contains 1,066 forms produced by approximately 40 different writers. The George Washington database consists of 20 pages of letters in English, written

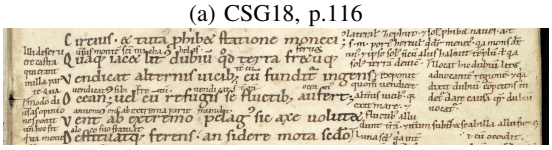
[1]Note that in [8] only 30 pages have been available and used for user studies



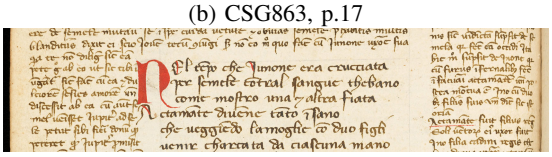
(a) CSG18, p.116 (b) CSG863, p.17 (c) CB55, p.67v

Figure 1: Samples pages of the three medieval manuscripts in DIVA-HisDB.

(a) CSG18, p.116



(b) CSG863, p.17



(c) CB55, p.67v

Figure 2: Details of Figure 1.

by George Washington and his associates in the 18th century. The database is available online at the Library of Congress[2]. The most related to DIVA-HisDB databases are the IAM-HistDB (We recently published layout analysis ground truth in [10]), containing German medieval manuscripts (13th century)[3] and GRPOLY-DB comprising documents in old Greek Polytonic script[4]. However, the current dataset is larger and contains more complex layout.

III. DESCRIPTION OF DIVA-HISDB

DIVA-HisDB[5] is a collection of three medieval manuscripts (see Figures 1 and 2) that have been chosen regarding the complexity of their layout, together with partners from e-codices[6] and the Humanities faculty in the University of Fribourg:

- St. Gallen, Stiftsbibliothek, Cod. Sang. 18, codicological unit 4 (CSG18),

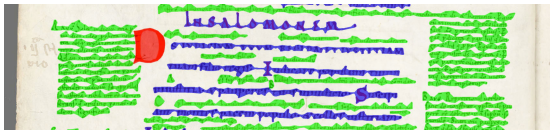
[2]http://memory.loc.gov/ammem/gwhtml/gwseries2.html

[3]www.iam.unibe.ch/fki/databases/iam-historical-document-database

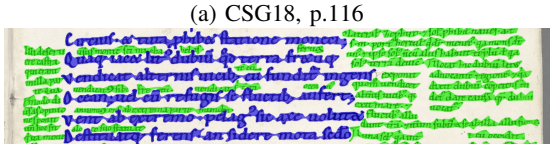
[4]http://www.iit.demokritos.gr/ nstaml/GRPOLY-DB

[5]The database will be made publicly available on the project page http://diuf.unifr.ch/hisdoc2

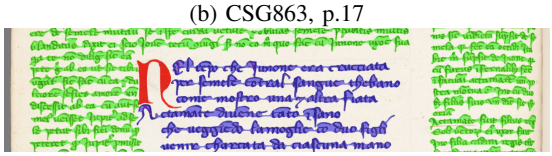
[6]http://www.e-codices.unifr.ch/



(a) CSG18, p.116



(b) CSG863, p.17



(c) CB55, p.67v

Figure 3: GT visualization of the three annotation categories (blue: main text body, green: comments, red: decorations) of the three medieval manuscripts in DIVA-HisDB.

- St. Gallen, Stiftsbibliothek, Cod. Sang. 863 (CSG863),
- Cologny-Genève, Fondation Martin Bodmer, Cod. Bodmer 55 (CB55).

The three manuscripts are representatives of a specific category of manuscript layout, i.e. complex layout containing a main text body and marginal and/or interlinear glosses, additions, and corrections. This layout category is less numerous than the most common simple layout with a main text body only. Still, a layout analysis tool trained on this more complex layouts will easily be able to deal with simple layouts. In the choice of manuscripts, we focused on the Carolingian period, represented by CSG18 and CSG863,dating from the 11th century, written in Latin language using Carolingian minuscule script and having quite similar layout features. The number of writers in these two manuscripts is unspecified. In order not to be confined to one specific period, we also included CB55, a manuscript from the 14th century that shows a different script (chancery script), language (Italian and Latin), and layout. There is only one writer in this manuscript. This choice allows for training and testing the layout analysis tool in different contexts (same script or cross-script context). The chosen manuscript pages vary from rather simple pages with only a few glosses to complex pages with many glosses, annotations, and decorations. All pages have been digitized with a resolution of 600 dpi and have a size of approximately 20 25 cm.

DIVA-HisDB consists of 150 pages in total, 50 pages from each manuscript. For the dataset, as well as for the division into training, validation, and test set (See Section V) we have selected a representative set of pages. The three different manuscripts are shortly described in the following.

A. CSG18

This is a composite manuscript; the pages selected for the dataset are part of the codicological unit 4 containing a Psalterium glossatum with comments (see Figures 1a and 2a). This manuscript part was written in St. Gall, probably in the second half of the 11th century [11]. In the earlier literature we find divergent datings: in Von Euw [12], p.502, dates the Psalterium to the late 10th century, in Scherrer [13], p.6, dates this manuscript part to the 12th century. Note that such uncertainty is quite common in manuscript studies and has to be taken into account when developing automated document classification and dating methods.

The psalms are introduced by a rubric (incipit), their beginning is marked by a red initial and a line written in display script (capitalis rustica); each psalm verse begins with a red majuscule. The comments appear as marginal glosses with interlinear reference marks (a-z and other reference marks). There are also interlinear glosses containing explanations and textual variants. The comments and glosses are written in a less calligraphic style than the main text. The manuscript part contains a drawing that is inserted in the main text body (see the description in p. 503 of [12]), but this page has not been selected for the present dataset.

B. CSG863

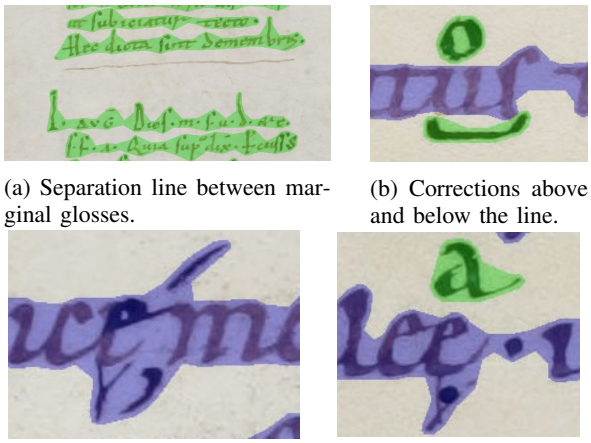
This manuscript (see Figure 1b and 2b), was written in the second half of the 11th century [11]. Euw [12], p. 505, presumes the Reichenau as place of origin, while Hoffman [11], p. 207, localizes the manuscript in St. Gall. It contains Lucan's Pharsalia libri decem. Each book begins with a red initial and two lines in display scripts (uncial, capitalis rustica); red incipits and explicits frame the books. There are marginal comments, some with interlinear reference marks, as well as interlinear glosses and corrections. Some pages (see [14], p. 169, 249, 253, 378-379, 385-389) contain interlinear musical notation (neumes). Others contain marginal drawings (see the description in [12], p. 504-505), but they were not selected for the present dataset.

C. CB55

This manuscript (see Figure 1c and 2c), known as Codex Guarneri, was written in the first half of the 14th century in Italy [15]. It contains the Inferno and Purgatorio from Dante's Divina commedia. While the manuscript is written in chancery script, it contains marginal glosses in Latin. The chapters are introduced by rubrics and red initials. Each tercet begins with a red-striped majuscule. In the glosses, the Italian reference text is sometimes repeated with a red underline.

IV. Creation of the DIVA-HisDB

For annotating the three selected medieval manuscripts we use GraphManuscribble [8], a semi-automatic tool which is based on document graphs[7] and pen-based scribbling interaction. After an initial annotation and a second correction iteration, all pages have been scrutinized and corrected by an expert. Thereafter, a final validation and correction iteration has been performed verifying that the data is consistent. DIVA-HisDB provides the GT in the PAGE [16] format.



(a) Separation line between marginal glosses.

(b) Corrections above and below the line.

(c) Correction overwriting the text.

(d) Correction above the line.

Figure 4: Annotation examples for separation lines and different type of corrections in DIVA-HisDB (main text body in blue, comments in green).

We distinguish between the main text body, the comments (marginal and interlinear glosses, explanations, corrections) and decorations (every character/sign that exceeds the size of a text line and/or is written in red). In Figure 3 the main text body is illustrated in blue, the comments in green and the decorations in red. The main goal of the annotation is the automatic layout analysis and text line segmentation. Therefore these annotation categories are based on spatial and simple visual features like color. Paleographic features like display scripts or black majuscules are not marked as a decoration, since they are too difficult to distinguish for the automatic layout analysis tool in this first step. In future work, when also script analysis will be included, these paleographic features could also be taken into account. We totally disregard elements that don't belong to a text line or are not characters/signs, like separation lines between two marginal glosses (see Figure 4a). Those elements are very rare.

Since the annotation categories are constituted for a layout analysis purpose, they don't always correspond to the paleographic description of the page. For example, they don't take into account the different time steps of the production/revision of a manuscript. Within the comments' category, we don't distinguish between the original glosses,

[7]The sparse document graph is an automatically derived structural representation of the ink-areas in the document.

Table I: DIVA-HisDB: Number of regions per annotated category (comments, decorations, main text body).

	CSG18	CSG863	CB55	total/MS	(%)
main text body	1,353	1,538	1,486	4,377	26.67
decorations	672	30	835	1,537	9.36
comments	6,260	1,656	2,584	10,500	63.97
total	8,258	3,224	4,905	16,414	

Table II: DIVA-HisDB: Surface area in pixels per annotated category (comments, decorations, main text body).

	CSG18	CSG863	CB55	surface/MS	(%)
main text body	49,114,900	111,641,860	120,034,717	280,791,477	56.94
decorations	3,945,155	369,512	4,008,359	8,323,026	1.69
comments	60,050,935	19,002,895	124,972,171	204,026,001	41.37
total	113,110,990	131,014,267	249,015,247	493,140,504	

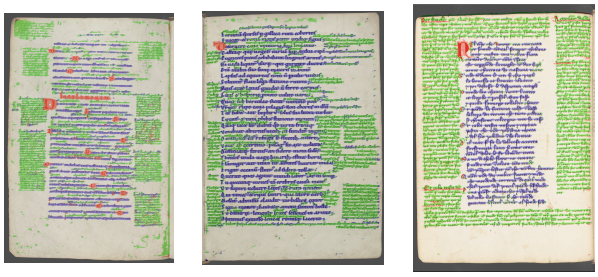
added by the writer of the main text, and the foliation added in the modern period. Also, if corrections have been made by overwriting the main text, they are not marked as posterior to the original main text (see Figure 4c); but if the corrections are made above and below the line, e.g., with a sign below the line to mark the affected letters, and the correct letter(s) above the line, they are annotated as comments (see Figure 4b and 4c).

Details of the distribution of the annotation categories appear in Table I. For the whole DIVA-HisDB 63.97 % of the total number of regions per annotated category are comments, 9.36 % are decorations and 26.67 % consists the main text body. Note that the proportions of the categories differ between the manuscripts. While the amount of textlines and comments is almost the same in CSG863, significantly less comments are present in CSG18. Additionally, by measuring the surface area (number of pixels) for each annotated category, 41.37 % of the total surface area are comments, 1.69 % are decorations and 56.94 % consists the main text body (see Table II).

V. LAYOUT ANALYSIS BENCHMARK FOR DIVA-HISDB

One of the first steps in DIA is layout analysis, i.e., identifying regions of interest in order to select which following methods to use and where to apply them. We applied a layout analysis method based on pixel-labelling in order to evaluate the difficulty of distinguishing between main text, comments, and decoration elements in DIVA-HisDB. For the benchmark we have selected 20 pages per manuscript for training, 10 pages per manuscript for validation, and 10 pages per manuscript for testing. We intentionally left out 30 pages (10 per manuscript) to allow for the organization of competitions in the near future.

For the experiments we used the *N-light-N* library presented in Seuret *et al.* [7]. We trained a CAE on the training documents in an unsupervised way to have it learn features which can then be used for classification purposes (similar to [17]). In the first layer of the CAE we used a 5×5 pixel convolution with a stride of 5 and 12 outputs. For the second layer, a 3×3 convolution with a stride of 3 generating 24 outputs was taken. In the third layer, again a 3×3 convolution with a stride of 3, but generating 72 features, was taken. The size of a patch covered by the CAE is 45x45 pixels. The CAE was trained with a learning rate of 0.001 on 10^6 equally distributed, randomly selected patches from the training set.



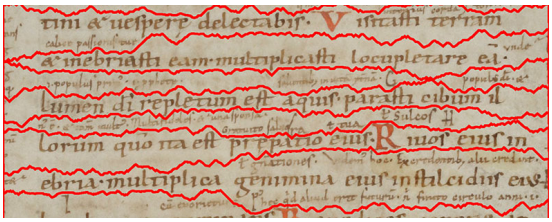
(a) CSG18, p.116 (b) CSG863, p.17 (c) CB55, p.67v

Figure 5: Layout analysis using *N-light-N* library in the three annotation categories (blue: main text body, green: comments, red: decorations) of the three medieval manuscripts in DIVA-HisDB.

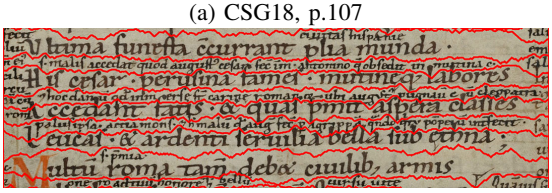
For the classification, the decoding part of the CAE is discarded and a classification layer is added on top of the CAE to create a Convolutional Neural Network (CNN). The weights of this CNN are untied, i.e., different weights can be learned for different positions in the convolutions. This additional layer has one neuron for each class, and the neurons having the highest output indicate which class has been selected. This leads to a single-class labelling of the pixels, but it is evaluated against the multi-class GT. Again the networks were trained for up to $2 \cdot 10^9$ iterations with a learning rate of 0.001. Due to the size of the document images (see Section III), pixel-labeling would become quite time-consuming. Therefore, we down-scaled the document images by factor 4 for the classification task (while for the

Table III: Accuracy (in %) using *N-light-N* library on the test set, i.e., the fraction of pixels correctly classified as belonging or not to that class.

	CSG18	CSG863	CB55
background	92.23	95.11	90.27
main text body	96.95	95.59	96.19
comments	93.18	96.96	95.73
decorations	99.09	99.93	95.04
average	95.36	96.98	94.31



(a) CSG18, p.107



(b) CSG863, p.5

Figure 6: Text line extraction results using a seam carving based method.

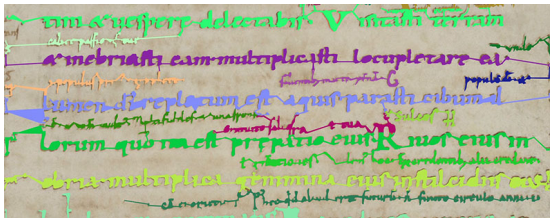
computation of the pixel-level accuracy, the full-sized GT image was taken into account).

The class-wise pixel classification results per manuscript appear in Table III. A visual representation of the results is shown in Figure 5. The overall accuracy is around 95 %. The accuracy is calculated by dividing the number of correct classifications (True Positives + True Negatives) via the total number of pixels. As can be seen, the results are quite acceptable for the main text body classification. However, for comments, which could appear anywhere on the page, the network has still much confusion. Note that the results could be improved by domain-specific post-processing (especially the background vs. comments confusion). This, however, is beyond the focus of the present paper.

Furthermore, to show the complexity of the proposed database we applied two state-of-the-art text line extraction methods:

- seam carving based method [18],
- ocropy lineseg [19][8].

We applied these two methods on CSG18, p.107 and CSG863, p.5, using DIVAServices [20]. In order to ease the process, instead of using the entire image as an input to these methods, we used only a text block containing just a column of the main text body. As such the method does not have to deal with page segmentation problems (separating the document into text blocks). In Figures 6 and 7 we can see that both methods fail to segment the manuscript correctly as there is a complex layout even in the main text body (interlinear glosses, additions, and corrections).



(a) CSG18, p.107

(b) CSG863, p.5

Figure 7: Text line extraction results using ocropy.

VI. Conclusion

This paper introduces DIVA-HisDB, a novel dataset with complex layouts providing new challenging tasks for the DIA research community. The dataset comprises 150 pages from three manuscripts together with their layout analysis GT. Furthermore, a benchmark setup is defined and a baseline experiment for layout analysis using optimized state-of-the-art CAE is performed.

While our initial results are already promising, it can be seen that the task is very challenging. To further demonstrate the challenging nature of DIVA-HisDB, we have tested state-of-the-art tools for text line segmentation. Also in those experiments a poor performance can be observed as the difficulties posed in this dataset are beyond previously published datasets.

In terms of algorithms we plan to adapt the deep learning architecture and investigate the possibility of transfer learning [21]. Furthermore, we will evaluate the approaches on the cross-script task, i.e., using one manuscript for training and the others for testing the classifier. Therefore we will also extend this database with more manuscripts containing different scribes and scripts as well as complex layouts. We will also provide the GT in a TEI format[9] along with the corresponding transcriptions. For the benchmark we intentionally left out 30 pages (10 per manuscript) to allow for the organization of competitions in the near future. The series of competitions is intended to start in 2017 and will continue with more challenging scripts, layouts, and tasks.

[8]https://github.com/tmbdev/ocropy

[9]http://www.tei-c.org/

Acknowledgment

This work was funded by the Swiss National Foundation (SNF) project HisDoc 2.0 with the grant number: 205120 150173.

References

[1] C. Y. Suen, C. Nadal, R. Legault, T. A. Mai, and L. Lam, "Computer recognition of unconstrained handwritten numerals," *Proceedings of the IEEE*, vol. 80, no. 7, pp. 1162–1180, 1992.

[2] J. J. Hull, "A database for handwritten text recognition research," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 16, no. 5, pp. 550–554, 1994.

[3] U.-V. Marti and H. Bunke, "The IAM-database: an English sentence database for offline handwriting recognition," *IJDAR*, vol. 5, no. 1, pp. 39–46, 2002.

[4] T. M. Rath and R. Manmatha, "Word spotting for historical documents," *IJDAR*, vol. 9, no. 2-4, pp. 139–152, 2007.

[5] A. Fischer, E. Indermühle, H. Bunke, G. Viehhauser, and M. Stolz, "Ground truth creation for handwriting recognition in historical documents," in *Proceedings of the 9th IAPR DAS*. ACM, 2010, pp. 3–10.

[6] B. Gatos, N. Stamatopoulos, G. Louloudis, G. Sfikas, G. Retsi nas, V. Papavassiliou, F. Simistira, and V. Katsouros, "GRPOLY-DB: An old Greek Polytonic document image database," in *ICDAR, 2015 13th International Conference on*, 2015, pp. 646–650.

[7] M. Seuret, R. Ingold, and M. Liwicki, "N-light-N: a highly-adaptable Java library for document analysis with convolutional auto-encoders," in *to appear in ICFHR*, 2016.

[8] A. Garz, M. Seuret, F. Simistira, A. Fischer, and R. Ingold, "Creating ground truth for historical manuscripts with document graphs and scribbling interaction," in *Proc. 12th int. DAS*, 2016.

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

[10] K. Chen, M. Seuret, H. Wei, M. Liwicki, J. Hennebert, and R. Ingold, "Ground truth model, tool, and dataset for layout analysis of historical documents," in *SPIE/IS&T Electronic Imaging*, 2015, pp. 940 204–940 204.

[11] H. Hoffmann, "Schreibschulen und Buchmalerei," *Handschriften und Texte des 9.-11. Jahrhunderts. Hahnsche Buchhandlung*, 2012.

[12] A. von Euw, *Die St. Galler Buchkunst vom 8. bis zum Ende des 11. Jahrhunderts*. Verlag am Klosterhof, 2008.

[13] G. Scherrer, *Verzeichniss der Handschriften der Stiftsbibliothek von St. Gallen*. Buchhandlung des Waisenhauses, 1875.

[14] G. Bobeth, *Antike Verse in mittelalterlicher Vertonung: Neumierungen in Vergil-, Statius-, Lucan-und Terenz-Handschriften*. Bärenreiter, 2013.

[15] P. Allegretti, "Catalogo dei codici italiani, Cod. Bodmer 55," in *Corona Nova*, 2003, pp. 44–47.

[16] S. Pletschacher and A. Antonacopoulos, "The PAGE (Page Analysis and Ground-Truth Elements) format framework," in *2010 20th ICPR*, 2010, pp. 257–260.

[17] K. Chen, M. Seuret, M. Liwicki, J. Hennebert, and R. Ingold, "Page segmentation of historical document images with convolutional autoencoders," in *ICDAR, 2015 13th International Conference on*, 2015, pp. 1011–1015.

[18] N. Arvanitopoulos and S. Susstrunk, "Seam carving for text line extraction on color and grayscale historical manuscripts," in *ICFHR, 2014 14th International Conference on*. IEEE, 2014, pp. 726–731.

[19] T. M. Breuel, "The ocropus open source ocr system," in *Electronic Imaging 2008*, 2008, pp. 68 150F–68 150F.

[20] M. Würsch, R. Ingold, and M. Liwicki, "SDK Reinvented: Document Image Analysis Methods as RESTful Web Services," in *Proc. 12th int. DAS*, 2016.

[21] M. Z. Afzal, S. Capobianco, M. I. Malik, S. Marinai, T. M. Breuel, A. Dengel, and M. Liwicki, "DeepDocClassifier : Document classification with deep convolutional neural network," in *13th ICDAR*. IEEE, 2015, pp. 1111–1115.