2011 International Conference on Document Analysis and Recognition

Multi Resolution Layout Analysis of Medieval Manuscripts Using Dynamic MLP

Micheal Baechler, and Rolf Ingold
University of Fribourg
Department of Informatics
1700 Fribourg, Switzerland
{*micheal.baechler, rolf.ingold*}*@unifr.ch*

***Abstract*—This paper describes a generic layout analysis system for historical documents. It presents the architecture of a pyramidal approach using three analysis levels. Each level consists of a classifier using machine learning techniques where the output of the upper level is used as a feature in the lower level. The current implementation uses a so called Dynamic Multi-Layer perceptron (DMLP), which is a natural extension of MLP classifiers. The system is evaluated on medieval documents for which a multi-layer model is used to discriminate among 10 classes organized hierarchically.**

***Keywords*-Segmention, layout analysis, Multi-Layer Perceptrons, historical documents.**

I. Introduction

In the last decade increasing activities in the digital libraries field were witnessed for the preservation of ancient documents and offering public access through internet. We are especially interested in historical manuscripts because of their complicated layouts. Their text is often spread over several columns and is often written using many different styles with variable line spacing. Moreover, such manuscripts contain decorative elements such as ornaments, illustrations, and comments in the margins and between text lines. All these elements make the layouts of the manuscripts very heterogeneous.

Nowadays, several tools assist librarians to annotate semi-automatically the physical layout and the logical structure of historical documents, as well as to transcribe them (e.g., Renaissance printed books and archival documents) [1], [2]. All these tools are dedicated to a specific set of documents and need an adaptation phase to segment other types of documents. With the increasing amount of digitalized documents, there is a need for a generic robust layout analysis approach to automatic extraction of the physical structure of the documents. Such an approach would minimize manual interventions necessary to segments text lines in order to perform the transcription of the precise textual content (as for instance, in [3]).

In order to precisely evaluate a layout analysis system, a set of ground truth layout has to be provided. Since we decided to use supervised classification techniques, such a set is also needed for the training phase. The generation of a large ground truth set of historical document layouts is a fastidious task in the development of a layout analysis system. In order to minimize human interventions in this task, we decided to develop our layout analysis using a bootstrapping approach. i.e we first develop a layout analysis system with a small initial ground truth set generated manually using a simple annotation tool. Then we use this system to generate first segmentation results from which we can generate a larger ground truth set. Thanks to this set, we can produce a new layout analysis system which performs better in terms of the segmentation quality than the first system. Then, we use the second system to enlarge the ground truth set (to improve the segmentation quality). We repeat the last two steps until we obtain the desired segmentation quality.

In this paper, we present a generic layout analysis architecture for complex documents (e.g medieval manuscripts) which is based on a pyramidal approach with several analysis levels using a series of images with increasing resolution. At each level the document image is segmented into areas with specific labels (e.g blocks of text, text lines, degraded areas, etc) and this information is forwarded to the next analysis level.

We implemented our cascade layout analysis architecture with three analysis levels in order to segment the physical structure of medieval manuscripts. Our system extracts white-space (background) in all levels, out of page in the first and the second level, decorations in the second and the third level, blocks of text in the second level and finally text lines in the third level. Encouraged by the results of our previous work in the field of document analysis we use Dynamic Multi-Layer Perceptron (DMLP)[6] and a scaled document image on each level.

This paper is organized as follows. Section 2 gives an overview of existing works in the field of layout analysis for historical documents. Section 3 describes our layout analysis system. Section 4 describes the data corpus currently used to evaluate our proposed architecture. Section 5 details our preliminary experiments. Finally, Section 6 draws some conclusions.

II. Related Works

Several works propose layout analysis techniques for detecting regions of interests in historical documents. For example, Likforman-Sulem [7] gives a very good survey of works which extract lines. She describes various ways

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.239

of describing text lines: delimiting their inter-line spaces, delimiting their strings by bounding boxes, defining their base lines, or clustering their raw components (according to some elements of the image, such as pixels or CCs). Works such as [8], [9] propose techniques for extracting lines and layout for specific kinds of historical documents.

Journet's works [10] demonstrate the possibility to extract and to compare layout elements with high semantic level only by using low level features. His experiments identify text, background and drawing parts in a manuscript. He achieves this by performing the analysis of the document texture with a multi resolutions approach.

DEBORA [1] is a complete system based on image analysis that simultaneously extracts meta data concerning the physical layout and compresses document images. The data extraction uses Connected Components (CC) analysis. DEBORA aims to assist experts in manual transcription of Renaissance books which have been printed in the sixteenth century.

AGORA project [2] gives a user-driven annotation tool to perform layout analysis of historical printed documents and to index them. Their segmentation algorithm builds a shape map and a background map. The shape map is created from the bounding box of the CCs in a page. The CCs are then classified as noise, graphics, or text, according to their size. The background map allows to extract blocks by providing white space between blocks. A user then defines a scenario for typical pages to classify the extracted blocks as desired.

III. System Description

A. Multi Resolution Analysis Architecture

In order to develop a layout analysis system for a specific kind of historical documents, we need to define a model of their layout that specifies the classes of graphical elements they contain. Thus, this model defines the labels that our system can use to segment the image documents. Since we wanted to illustrate our concept for multi resolution analysis architecture on medieval manuscripts, we defined a layout model that delimits the graphical elements in a document by bounding polygons and saves them in an XML format by organizing them in the following layers [5]:

- The decoration layer limits all decorative elements such as ornaments, drop capital, decorative initial, etc.
- The degradation layer emphasizes all damaged area such as holes, stains, seams, etc.
- The text layer limits the main text by emphasizing its blocks, lines and token in a hierarchical manner. A token corresponds to a text fragment, a word, or a number.
- The comment layer delimits annotation and inserted text between the main text lines and in the margin. As in the text layer, it contains blocks of text, text lines or tokens.

These layers delimit the foreground of the manuscript image. However, since most of digitalized document images contain an area which does not belong to the scanned document, our model emphasizes this by a bounding polygon which separates the background and the out of page area. This model defines ten labels to use for segmenting our layout. Currently, our implementation does not distinguish between the main text layer and the comment layer and does not support the token concept. Hence, our layout analysis segments a manuscript image into six classes kept in the final result. These six labels are out of page, background of page, decoration, degradation, block of text and text line.

Nowadays libraries digitalize their historical documents in high quality images. The analysis of such images is a CPU consuming task. To avoid this inconvenience, our layout analysis is based on pyramidal approach of several analysis levels. Each level classifies the scaled image pixels of the manuscript into a set of labels and forwards the information to the lower level. In order to reduce the CPU load, the pixels classified with specific labels are excluded from the consideration and their corresponding area is not reclassified by the lower levels. Thanks to these forwarded information the lower level corrects precomputed classification of the higher level and uses it to refines the classification of the rest of the image with more specific labels. Figure 1 shows the layout analysis for medieval manuscripts based on our pyramidal approach of three analysis levels. Notice that the image regions classified as out of page areas in the first level are not reclassified in the lower levels. The same is true for regions classified as decoration, out of page, or background in the second level. The choice of these labels was deduced from our experiments. This approach reduces the computation time considerably.

In the current implementation, each level consists of a DMLP which classifies scaled image pixels into a specific set of labels. DMLP directly inherits its structure and functionality from a standard MLP and it additionally restricts their topology and improves their training phases. Hence, it avoids that frequent classes influence the whole system during training, and it efficient even with only a few training samples. Finally, it is convenient for our generic purpose, because it allows to avoid the need of almost all manual configuration during training.

B. Feature Extraction

As it is common in classification techniques, the feature extraction step is crucial in order to attribute numerical vectors to image pixels. These vectors which are the inputs of DMLP combine the output of the upper level (classification results and the neuron outputs) and basic features extracted from a scaled image of the manuscript. In our current implementation for medieval manuscripts we decided to use a basic set of features for each pixel in the scaled

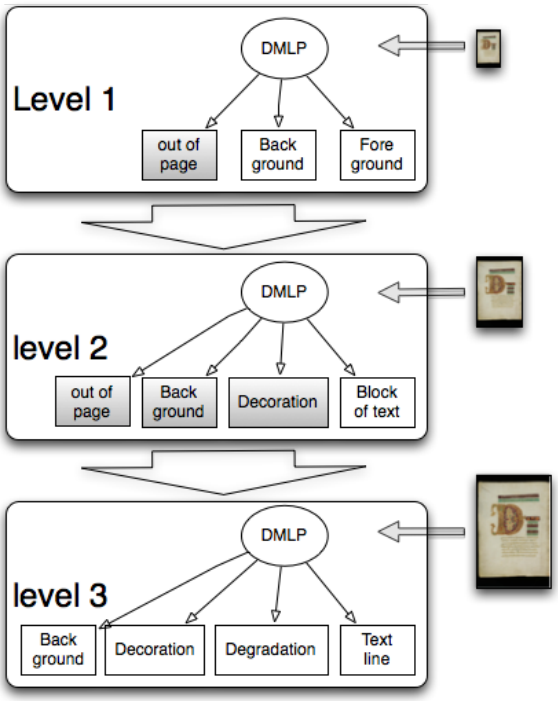


Figure 1. Our multi resolution layout analysis system for medieval manuscript[11].

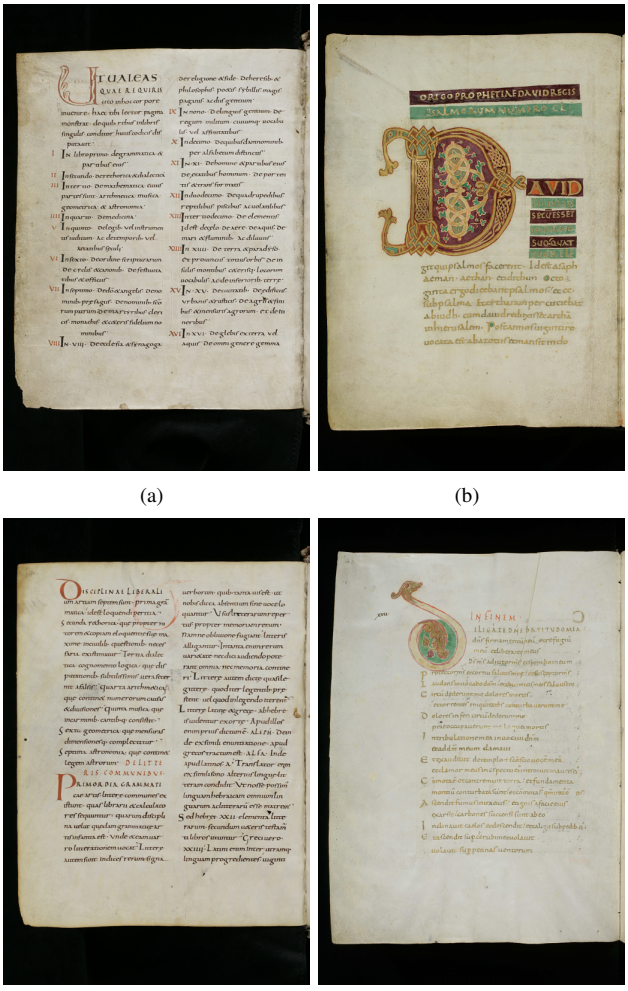
image. These features are formed by the composition of the following:

1) The pixel position is (the coordinates x, y of a pixel in the scaled image).
2) The pixel color values which consist of the three primary color components of the RGB color space.
3) The values of the neighbor pixels (the RGB color values of the pixels in a 5x5 window).
4) The outputs of the neurons produced by Dynamic MLP of a higher resolution. We enumerated the labels present in each level and also passed the label number as a feature for the lower level. Notice that for a pixel classified in the upper level, its label is passed as a feature for the corresponding pixels in the scaled image of the lower level.

We decided not to use last two features for the first analysis level. As it is common in classification the feature vector is normalized using the min-max normalization [6], so the feature vector components are mostly comprised between -1.0 and 1.0 which is suitable for DMLP.

IV. Corpus

Our pyramidal layout analysis approach is tested on two manuscripts [11] taken from the e-codices project, a virtual manuscript library from the Medieval Institute at the University of Fribourg, Switzerland. This Medieval Institute's project offers a web access to a set of medieval and modern manuscripts. Currently, their database contains 722 manuscripts from 30 different libraries, and it is regularly extended. The two manuscripts we chose were taken from Saint Gallen Stiftsbibliothek. They contain several hundreds of pages and have standard multi column layout (see Figure 2). Note that the manuscript images were scanned at a minimum resolution of 300 dpi and are therefore available in high quality (i.e., 3328x4992 pixels).



(a) (b) (c) (d)

Figure 2. Medieval manuscript examples from the Saint Gall Stiftsbibliothek's database[11].

V. Preliminary Experiments

We decided to test our layout analysis approach on medieval manuscripts using DMLPs. Each manuscript had its own DMLPs for each level. The training set for each manuscript was created by annotating manually the first ten pages of the manuscript. Two pages of the training sets are shown in Figure 2 a) and b). However, the training phases in our system began with training the DMLP of

the first level, then the one of the second level, finally the one of the last level. We assigned the same values to the parameters (max number of training cycles, number of neurons in the hidden layer connected to an output neuron) for all DMLPs during training phases. We can certainly optimize these parameters for each layer, but we preferred in these preliminary experiments to keep the training phases as simple and consistent as possible. Finally, we decided to use these three scale factors: $2^{-4}, 2^{-5}, 2^{-6}$ for the first manuscript presented in a) and c) and $2^{-4}, 2^{-6}, 2^{-7}$ for the second manuscript. These scale factors were chosen based on the visual analysis of the scaled images of the manuscripts.

We estimate that the quality of the manuscript segmentation performed by our system is relatively good (see Figure 3). It shows the segmentation result produced by the three analysis levels for the pages presented in Figure 2 c) and d). Notice that the out of page areas which were detected by the first level analysis are highlighted in black in lower levels and are not reclassified by the lower levels. This is also true for areas labeled as out of page, background and decoration in the second level.

In order to evaluate our pyramidal analysis system for medieval manuscripts, we created our test set by annotating the fifty pages which follow the training pages of each manuscript. Table I presents the classification matrix for the first manuscript which has a layout of two text columns. It shows, in our preliminary experiment, relatively good quality of the classification for the text line, out of page, and block classes. But the classification of decoration and degradation classes is worse. The classification matrix for the second manuscript shows that the classification of the block and the line classes is not as good as for the first manuscript. The classification of decoration is better than for the first manuscript. The classification quality for the second manuscript can be improved considerably by a simple post filtering process, thus we decided not to publish it yet. Notice that we preferred for our preliminary experiments not to optimize our system and not to perform any post processing, and do this in our future work.

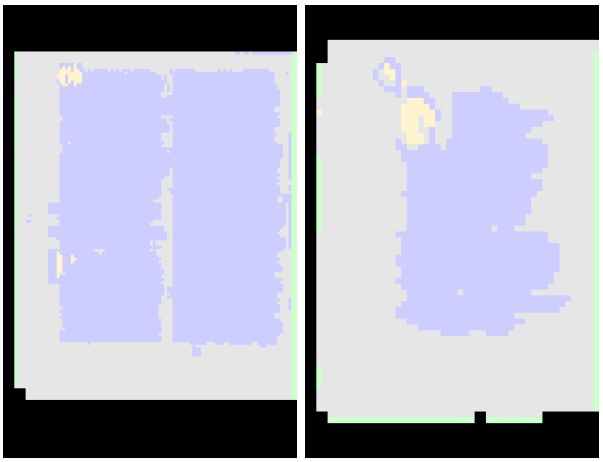
Table I
CLASSIFICATION MATRIX

	block	line	deco.	degr.	backgr.	out of p.
block	0.882	0.101	0.002	0.0	0.014	0.0
line	0.062	0.922	0.001	0.0	0.014	0.0
deco..	0.438	0.139	0.333	0.005	0.085	0.0
degr.	0.231	0.308	0.0	0.0	0.462	0.0
backgr.	0.061	0.003	0.0020	0.0	0.902	0.032
out of p.	0.0	0.0	0.0	0.0	0.002	0.998

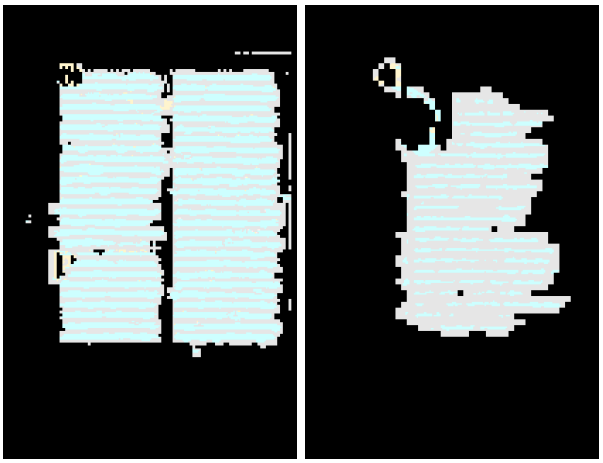
VI. CONCLUSION

In this paper, we presented a generic layout analysis architecture for the extraction of the physical structure of

(a) Level 1 segmentation



(b) Level 2 segmentation



(c) Level 3 segmentation

Figure 3. The segmentation of the two pages presented in Figure 2 c and d : a) The segmentation of the first level in which background areas are shown in gray, the foreground areas are in blue, and the out of page areas are in green. b) The segmentation of the second level in which the blocks of text are emphasized in blue, the decoration in orange, and the out of page areas in green, and background pixels presented in gray. c) The segmentation of the third level in which the background is presented in gray, the decoration in orange, and the text lines in blue.

historical documents. Our pyrimidal analysis approach consists of several levels on which images are segmented with increasing resolution. Each of these levels segments the document image into areas of specific classes (e.g block of text, text line, degraded areas, etc) and forwards this information to the next analysis level which refines the segmentation granularity. In this paper, our implementation of a multi resolution layout analysis system based on our pyrimidal architecture is limited to three analysis levels. The classification on each level is performed by Dynamic Multi-Layer Perceptron (DMLP), a machine learning approach, which uses the color features extracted from scaled document images. One advantage of our cascading layout analysis system is that it can be adapted to other manuscripts by only providing it with another set of annotated images. Currently, our preliminary experiments on two medieval manuscripts show encouraging segmentation quality for specific classes, in order to optimize our system and to perform a thorough quantitative evaluation of the segmentation quality with a heterogeneous document corpus.

Encouraged by the results of our preliminary experiments, we are planing to continue using our bootstrapping approach to reduce the amount of manual annotation needed for producing a training set even more. We will also introduce other types of features in our system. Finally, we want to evaluate our approach with several different learning algorithms.

REFERENCES

[1] F. Le Bourgeois and H. Emptoz, "Debora: Digital access to books of the renaissance," *International Journal on Document Analysis and Recognition*, vol. 9, no. 2, pp. 193–221, 2007.

[2] J. Y. Ramel, S. Busson, and M. L. Demonet, "Agora: the interactive document image analysis tool of the bvh project," *Document Image Analysis for Libraries*, pp. 145–155, 2006.

[3] A. Fischer, M. Wuthrich, M. Liwicki, V. Frinken, H. Bunke, G. Viehhauser, and M. Stolz, "Automatic transcription of handwritten medieval documents," *Virtual Systems and MultiMedia*, vol. 0, pp. 137–142, 2009.

[4] T. M. Rath and R. Manmatha, "Word spotting for historical documents," *International Journal on Document Analysis and Recognition*, vol. 9, no. 2, pp. 139–152, 2007.

[5] M. Baechler, J.-L. Bloechle, and R. Ingold, "Semi-automatic annotation tool for medieval manuscripts," *Frontiers in Handwriting Recognition, International Conference on*, vol. 0, pp. 182–187, 2010.

[6] J.-L. Bloechle, "Physical and logical structure recognition of pdf documents." Ph.D. dissertation, Faculty if Science, University of Fribourg, June 2010.

[7] L. Likforman-Sulem, A. Zahour, and B. Taconet, "Text line segmentation of historical documents: a survey," *International Journal on Document Analysis and Recognition*, vol. 9, no. 2, pp. 123–138, 2007.

[8] M. Bulacu, R. van Koert, L. Schomaker, and T. van der Zant, "Layout analysis of handwritten historical documents for searching the archive of the cabinet of the dutch queen," *International Conference on Document Analysis and Recognition*, vol. 1, pp. 357–361, 2007.

[9] V. Malleron, V. Eglin, H. Emptoz, S. Dord-Crouslé , and P. Régnier, "Extraction automatisée de lignes et de fragments textuels dans les images de manuscrits dauteur du 19ème siècle ," *MAnifestation des JEunes Chercheurs en Sciences et Technologies de l'Information et de la Communication*, Nov. 2009. [Online]. Available: http://liris.cnrs.fr/publis/?id=4507

[10] N. Journet, J.-Y. Ramel, R. Mullot, and V. Eglin, "A proposition of retrieval tools for historical document images libraries," *International Conference on Document Analysis and Recognition*, vol. 2, pp. 1053–1057, 2007.

[11] "Saint Gall, Stiftsbibliothek," Cod. Sang. 22, p. 4, p. 40, and Cod. Sang. 231, p. 4, p. 14, http://www.e-codices.unifr.ch.