USING LOCAL FEATURES FOR EFFICIENT LAYOUT ANALYSIS OF ANCIENT MANUSCRIPTS

Angelika Garz, Robert Sablatnig, Markus Diem

Vienna University of Technology, Institute of Computer Aided Automation, Computer Vision Lab
1040 Vienna, Austria
email: garz@caa.tuwien.ac.at
web: www.caa.tuwien.ac.at/cvl

ABSTRACT

A binarization-free layout analysis method for ancient manuscripts is proposed, which identifies and localizes layout entities exploiting their structural similarities on the local level. Hence, the textual entities are disassembled into segments, and a part-based detection is done which employs local gradient features known from the field of object recognition, the Scale Invariant Feature Transform (SIFT), to describe these structures. Layout analysis is the first step in the process of document understanding; it identifies regions of interest and, hence, serves as input for other algorithms such as Optical Character Recognition (OCR). Moreover, the document layout allows scholars to establish the spatio-temporal origin, authenticate, or index a document. The layout entities considered in this approach include the body text, embellished initials, plain initials and headings.

1. INTRODUCTION

Document layout analysis of ancient manuscripts induces specific challenges not present in modern machine-printed documents and historical documents from the hand-press period [5, 2, 1]. Layout analysis is referred to the segmentation of a page into homogeneous regions consisting of layout elements belonging to the same class. Layout entities are objects such as initials, the main body text, or headings.

Through the centuries manuscripts decay due to inappropriate storing conditions, deterioration processes, mold and moisture [9, 1], resulting in torn pages and heterogeneous background intensity having artifacts due to aging, smudges, and stains [12, 4]. Further problems are seeping ink from the other side of the page (bleed-through), ink stains, and pages suffer from scratches, crease and are corrugated [9, 12, 4]. Furthermore, uneven lighting during the digitization process and fading out of ink are challenges to cope with [12, 9]. Hence, binarization pre-processing as used in traditional document layout analysis produces binarization errors, as they additionally segment background clutter. This especially applies to document images having a low dynamic range, which is the case if the ink is faded-out or the paper is stained and, therefore, the contrast between characters and background diminish.

The physical structure of ancient manuscripts is harder to extract than this of printed books as layout formatting rules of these manuscripts were looser and are not always complied with [12]. Handwritten documents may include narrow spaced lines, interfering lines running into each other, non-constant spacing between characters and lines, non-rectangular layout, variable locations of layout entities, or multiple scripts [12, 17].

Layout analysis identifies regions of interest and, hence, serves as input for other algorithms such as Optical Character Recognition (OCR) to retrieve the (ASCII) characters which correlate to the characters in the manuscript. A binarization-free OCR-system is proposed by Diem [7]. Initials can be extracted and further processed by decomposition algorithms [6, 10, 18] that aim at determining the letter represented by the respective initial. Gaining information about the layout, the manuscripts can be indexed and enhanced with meta data. Moreover, amongst other characteristics, the layout allows scholars to determine the temporal and geographical origin of a manuscript as the script and writing style of every historical period and place is characteristic. Hence, scholars are supported in dating and authenticating a document [16, 15].

The data set, the method is evaluated on, is an ancient manuscript probably originating from the 11^{th} century, discovered at St. Catherine's monastery on Mt. Sinai, Egypt, in 1975 [14]. It is written in Glagolica, the oldest known Slavonic alphabet. The manuscript in question, the Old Church Slavonic Psalter of Demetrius (*Cod. Sin. Slav.* 3N), consists of 145 folios having a front page (recto, r) and a back page (verso, v). The manuscript was digitized in the course of *The Sinaitic Glagolitic Sacramentary (Euchologium) Fragments Project* [14].

A binarization-free method for layout analysis independent of script and alphabet is proposed that takes into account the specifities of ancient manuscripts. It is a part-based method that detects and localizes layout entities based on their local structure. They are decomposed in parts employing a state-of-the-art object recognition method which identifies objects based on local features, namely Scale-Invariant Feature Transform (SIFT). This allows detecting handwritten characters having a high variability in their shape – depending on the scribe and the time and place where it was written. Especially in the case of embellished initials, the shape of the whole character is variable; however, the characteristics of the embellishments are similar. Amongst these are hatches and outlines. The layout entities considered in this paper are

- embellished initials, which are decorated letters larger than the regular text, embedded in the margin of the page or at the edge region of the main body text,
- plain initials, which are letters having a vertically elongated aspect ratio and a more angular shape when compared to a character of the main body text.
- headings, which are similar to plain initials and additionally may be written in another script such as Cyrillic script.
- main body text, where letters are characterized by a compact, rounded shape.

Due to the structural similarities of embellished and plain initials and headings, these entities are considered as one class and the main body text as a second class. Further layout entities not regarded in this paper are Latin page numberings added later and Glagolitic psalm numberings. The only difference between main body text and psalm numberings are horizontal lines above the characters.

The following section gives an overview about the related work relevant for layout analysis on ancient manuscripts and historical printed books, and then in Section 3, the proposed

THIS WORK WAS SUPPORTED BY THE AUSTRIAN SCIENCE FUND UNDER GRANT P19608-G12.

method is detailed. The subsequent section describes the evaluation and results, followed by a conclusion.

2. RELATED WORK

The data sets considered in traditional document layout analysis are printed documents rather than handwritten manuscripts. Though, as Antonacopoulos and Downton point out in their paper [1], applying approaches developed for the analysis of modern machine-printed documents on historical manuscripts imposes problems, robust methods adapted to the special challenges of these manuscripts are needed.

Bourgeois and Kaileh [11] propose a document analysis system that retrieves meta data such as initials, illustrations, text regions or titles from ancient manuscript images. The approach is based on a bottom-up segmentation step, where binary features linked to shape and geometry as well as color features from the original image are extracted for each connected component. The components are then merged into more complex elements such as text or decoration.

Ramel et al. [17] present a user-driven layout analysis system for historical printed books. They propose a two-step method that first creates a mapping of connected components to a so-called shape map and a mapping for background areas. The result of this first step is a list of segmented blocks. Then, this initial document representation is presented to users, who interactively build scenarios to label, merge or remove the blocks (e.g. ornamental letters, titles) according to their needs.

Whereas the two methods surveyed above incorporate a binarization, the following approaches directly work on grey-scale images.

In their work [8], Grana et al. propose a segmentation system for historical manuscripts, which distinguishes handwritten text, (floral) decorations and images. The method consists of two steps: first, circular statistics are used to separate text, background and images, and second, visual descriptors for color and texture applied to sliding windows extracts features for each block to differentiate between decorations and images.

In [5], Projection Profiles (PP) based on the number of transitions between ink and writing support are used as the main analysis method for structured handwritten documents in combination with color filtering, contour tracing and run-length extraction. The goal of the method is to split the document images into rectangular areas of interest containing the respective document elements.

A SIFT-based image and line drawing detection system is proposed by Baluja and Covell in [3]. They developed their approach for the indexation of these images in a large-scale book-scanning system. The documents, the approach is applied to, include historical printed books, manuscripts, newsprint and modern printed books. The method first extracts SIFT descriptors of the whole page and then classifies the descriptors using multiple classifiers trained with AdaBoost. Whereas the descriptors corresponding to the text class are dismissed, the descriptors and their interest points are stored in a database.

3. METHOD

In contrast to the majority of state-of-the-art layout analysis methods, the proposed approach does not need a binarization step. This makes the method robust to noise, background clutter and faded-out ink. As the considered data set does not have a strict, rectangular layout such as the documents considered in [5], a method invariant to skew, fluctuating text lines and differences in script and writing style is required. Color based segmentation is not suitable, since first, the decorative entities are not universally highlighted with a specific color and second, the highlight color is too similar to the background.

The method introduced in this paper consists of two consecutive steps, where the first is the extraction and classification of features and the second employs a cascading localization algorithm. Both tasks are based on interest points computed by means of Difference-of-Gaussian (DoG).

Applying an interest point detector in a scale-invariant manner, the foreground of a document is disassembled into segments, where every interest point represents a part of a character or initial dependent on its scale. Please note that background artifacts such as stains and clutter originating from the nature of the writing support, are detected as foreground as well. However, these artifacts are rejected in the classification or localization step.

Consecutively, a descriptor is calculated for every interest point. This leads to local features describing these parts of characters, or – depending on their scale – even whole characters or text lines. SIFT are chosen as features as they are invariant to scale and rotation, which is an important aspect for ancient manuscripts, as the script size and orientation may change. Furthermore, they are invariant to illumination changes, which allows for variations in the background intensity due to uneven or heterogeneously textured writing support, and changing intensity of the ink. The invariance to the 3D camera viewpoint SIFT incorporates, allows detecting the same character despite deformations owing to unevenness of the writing support or variations in the script.

The descriptors are then classified employing a kernel-based supervised machine learning algorithm. A Support Vector Machine (SVM) with a Radio Basis Function (RBF) kernel is chosen as classifier to discriminate between two classes: the main body text on the one hand, and layout entities having a decorative meaning on the other hand. These entities include embellished initials, plain initials and heading. They are grouped into one class as result of their local structure correspondence as explained earlier.The RBF kernel is chosen in order to be able to separate non-linear data.

Having classified the descriptors, a class label for each interest point is known. However, a localization algorithm needs to be applied in order to find regions encapsulating whole layout entities, as the classification of interest points leads to class decisions just on certain locations in the image, but not for image areas.

The scales and locations of interest points are exploited for the localization step. The presumption for this procedure is that an interest point represents an entity segment or even a whole entity. The scale of the interest point hence relates to the size of the entity part. Pursuant to this presumption, a cascade localization algorithm is introduced which successively reduces the amount of falsely classified descriptors.

Descriptors which are not non-ambiguously assignable to one class might be classified to either class. Such descriptors are likely to occur as not all of the elements, the entities are assembled of, are unique for one of the classes. A further reason is the scale invariance, which on the one hand is important due to the reasons given previously, but on the other hand, adds potential misclassifications. Examples are rounded shapes which are one of the discriminative characteristics of main body text, but occur in headings and initials too.

The proposed localization algorithm incorporates six consecutive steps successively reducing the number of mismatched descriptors.

Scale-Based Voting: The first step is a scale-based voting, where the classification scores obtained from the SVM are weighted according to the scale of their interest points. The underlying presumption is based on the observation that interest points of a certain scale are most reliable.

Marker Points: Second, a set of marker points, which are reliable interest points indicating the position of a potential layout entity, is established. Marker points are interest points having a certain scale and a high classification score.

Merging: Then the remaining interest points overlapping with at least one marker point are merged to the set of candidate interest points.

Filtering: The fourth step is region-based processing, where overlapping interest points set up a region. Interest points of regions including less than 10 interest points or regions smaller than an average character of the document are rejected.

Spatial Weighting: Thereafter, the interest points' scales and the previously weighted classification score are spatially weighted with a two-dimensional Gaussian distribution leading to one score map per class.

Post-Processing: The final step after voting the score maps pixel-wise against each other with the highest probability determining the final class label of the pixel, is a second region-based processing to reject isolated areas not large enough to be a valid character.

In the first stage, the interest points are voted based on their scale using a voting function penalizing diminutive and large scales (diminutive respectively large scales in this context means interest points smaller respectively larger than a whole character of the regular text or a heading). The small interest points are e.g. background clutter, dots, small structures of characters and speckles of the parchment. The large scales represent e.g. whole decorative initials, spots and stains as well as ripples of the parchment.

Owing to the distribution of classification scores when interrelated with their interest points' scales, a linear weighting function is chosen. This weighting function implements the principle of a band-pass filter and, thus, emphasizes a certain range of scales and lowering scales outside this range. The weight of the function is applied on the classification score gained from the SVM for each descriptor.

Applying this voting scheme, interest points representing entire characters – in the case of main body text – are stressed, whereas the impact of interest points having other scales is diminished. In case of the decorative entities, interest points indicating entire heading characters or plain initials are highlighted.

Three octaves are established for the scale-space, and thus, the interest points cover are range from small structures such as dots or stroke endings to large structures enclosing a whole embellished initial. As mentioned before, a certain scale of interest points is likely to reliably locate characters. Reliability is defined in terms of the capability of an interest point to indicate the location and expansion of a whole character rather than in terms of classification score.

In account of the characteristics of the scale, the classification score distributions and interrelations, the selection of candidates for so-called *marker points* is done based on the scale range of the second octave. The aim of marker points is indicating possible locations and extends of layout entities. Since all subsequent filtering steps are based on these initial marker point candidates, entities can solely be localized at positions where marker points are detected. Hence, the determination of these marker points is crucial for the performance of the localization algorithm.

The next step is to merge the marker points with all remaining retrieved interest points because the localization solely based on marker points generates a sparse localization result. Thus, all interest points overlapping with a marker point at least 25 % are included in the set of interest points used for the following stages.

Then, a region-based processing is applied, where a region is defined by overlapping interest points. Regions are candidates for layout entities. They are filtered based on two aspects: the number of interest points setting up a region and the size of the region.

Having determined the final set of interest points with their weighted classification scores, a so-called *score map* having the same size as the input image is established for each respective class. Hereby, each interest point's classification score is spatially weighted with a two-dimensional Gaussian distribution according to the scale of the interest point.

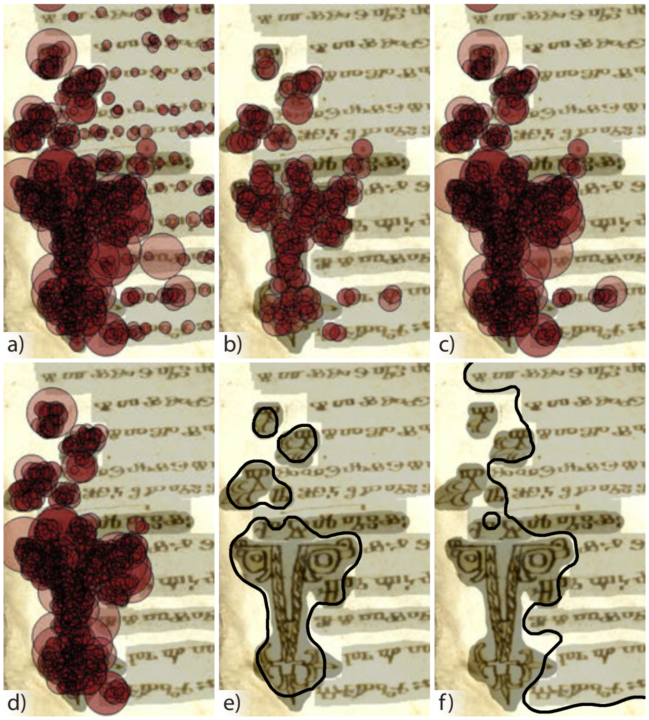

Figure 1: Exemplary results for decorative entities

Hence, at all locations of interest points, a weighted score distribution having the same spatial extend as the interest point is generated. For each pixel in the score map, the values of overlapping interest points are accumulated. This results in one score map for each respective class representing the accumulated score for each pixel indicating the expectation belonging to the particular class.

The final step in localizing layout entities is post-processing on the score maps generated in the previous phase. The two score maps are spatial distributions of probabilities for the class of each pixel. After voting the score maps pixel-wise against each other with the highest probability determining the final class label of the pixel, a second region-based processing to reject isolated areas not large enough to be a valid character is done.

Figure 1 gives an overview of the different stages of the localization algorithm. Gray blobs denote the ground truth, at which dark gray blobs indicate decorative entities and light gray blobs stand for the main body text class. The circles mark the interest points with their respective scales. Figure 1 a)-e) show the decorative entities-class, where a) illustrates all decorative entities descriptors classified by the SVM, in b), the marker points are shown as selected from the second octave, c) depicts the marker points after removing single occurrences, d) shows the marker points merged with overlapping interest points and e) presents the final result of the localization algorithm. In f), the final result for the main body text class is given.

4. EVALUATION AND RESULTS

The proposed method is evaluated on a random sample of 100 pages of the Psalter having varying layouts, scripts and writing styles. The training set consists of image patches of the respective classes, where 18 embellished initials, 30 plain initials and 30 headings are taken as training samples for the class containing the decorative entities, and 60 lines of main body text. Please note that the set of embellished initials does not cover the range and variety of these entities occurring in the manuscript.

Table 1: $F_{0.5}$-score, Precision and Recall for the different stages.

	$F_{0.5}$-score	Precision	Recall
a) Entire classification	0.9135	0.9243	0.8731
b) Main body text	0.9304	0.9388	0.8985
c) Decorative entities	0.6293	0.6670	0.5132
d) Main body text initial set of interest points	0.8726	0.9616	0.6370
e) Main body text interest points after the localization algorithm	0.9522	0.9709	0.8843
f) Decorative entities initial set of interest points	0.1982	0.1675	0.7419
g) Decorative entities interest points after the localization algorithm	0.4987	0.4569	0.7868

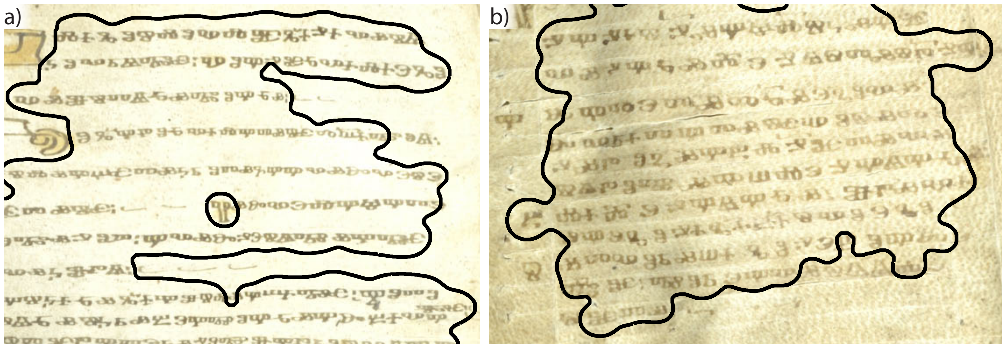

Figure 2: Exemplary results for the main body text

The evaluation is done on pixel-level, having manually tagged ground truth. The evaluation of the localization is not carried out at positions having overlapping class labels. Thus, a 20*px* margin is added to the blobs in each ground truth image (having a mean resolution of 2850 × 3150). This technique is motivated by two considerations: On the one hand, manually tagged ground truth is tainted with noise. This noise occurs especially in border regions of overlapping classes. On the other hand – depending on the data – the classes may have a fuzzy or overlapping region border which renders exact ground truth segmentation impossible.

Table 1 gives the evaluation results for the method, where $F_{0.5}$-score, precision and recall are employed as measure metrics for the method's accuracy. For the interpretation of the results, it has to be considered that the ratio between main body text and decorative entities is approximately 9:1 on the pixel level. Hence, the performance of the detection and localization of main body text has a higher influence on the entire classification result than the performance for the decorative entity class (compare Table 1 a-c). Rows a-c) of Table 1 give the performance on pixel level for the score maps, whereas rows d-g) show the results for the localization algorithm on the interest points, i.e. before the Gaussian weighting of the interest points.

Table 1 b, c) gives the results for the respective classes. The results for the decorative entities are not as promising as those for the main body text. When evaluating the results visually, the reasons for this are multiple. Figure 2 shows two results for the main body text, where b) is a page with faded-out ink. Figure 3 gives exemplary results for the decorative entities as headings (a-c), embellished initials (b,d-i) and plain initials (e-g,j,k).

First, the embellished initials are detected and localized well if enough structural detail is present. Long single strokes are not detected well by the DoG and SIFT, as edges do not provide reliable interest points [13]. Hence, at these strokes, the density of interest points is low and therefore, a reliable localization cannot be achieved either (see Figure 3 e,h).

Furthermore, plain initials placed in the margin of the page being smaller than an average regular character, and, hence, are either not covered by the marker points or do not produce enough interest points to be considered as an initial. For an example refer to Figure 3 j), in the left margin, there is a plain initial that is not detected.

A second issue relates to plain initials and headings having features not discriminative enough from the main body text. This concerns entities having a prevailing number of character segments characteristic for main body text, such as round, compact shapes. Hence, even for humans who are no experts in the Glagolitic language, the differentiation between main body text on the one hand and plain initials or headings on the other hand is a non-trivial task. In Figure 3 a), a heading having characters similar to the main body text, is shown.

The third point concerns plain initial embedded in the main body text, as they are single initials surrounded by another class, the reliable detection and localization is obfuscated when compared with isolated initials surrounded with background. In these cases, the determination of these entities is based on other characteristics such as a larger space before the initial or a colon at the end of previous sentence (see Figure 2). These characteristics cannot be exploited by the proposed method.

Additionally, class boundaries cannot always be determined unequivocally as regions are overlapping or collide. Hence, if the distance between an embellished initial and the main body text is smaller than distances occurring between internal segments of an initial, e.g. outlines or hatches, the features of both classes are included in the descriptors of this area and, thus, are ambiguous. Entities abundantly embellished produce more interest points than plain entities, and hence, due to the localization algorithm, that spatially weights the classification scores interest points, abundantly embellished entities have a higher weight and, thus, may superimpose the other class in boundary regions, confer Figure 3, f).

Table 1 d-g) give results from the evaluation of the classified descriptors, where d) and f) give the accuracy before the voting scheme and e) and g) after applying it on the interest points. The voting scheme reduces the number of incorrectly classified interest points.

5. CONCLUSION

A part-based layout analysis method is introduced, which exploits structural similarities of layout entities employing local

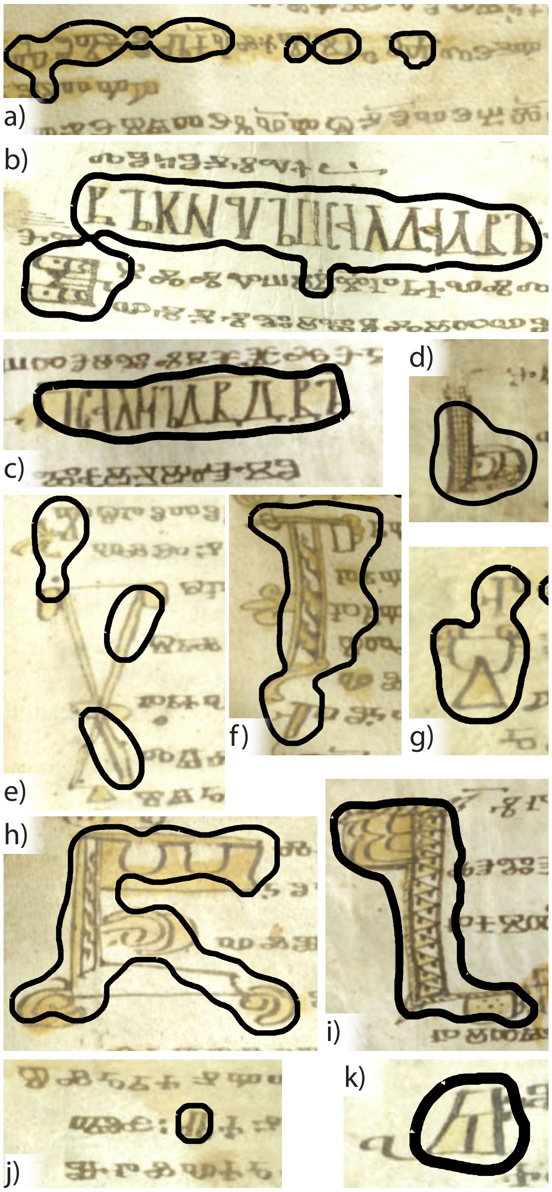

Figure 3: Exemplary results for decorative entities

features. Scale Invariant Feature Transform descriptors known from the field of object recognition are employed to describe segments of layout entities in a scale-, rotation- and illumination invariant manner. Hence, this approach does not rely on a binarization step but is directly applied to the gray scale image, and furthermore is robust to variations in shapes, illumination and writing orientation as well as (background) noise. Thus, it is suitable for ancient handwritten documents having varying layouts and being degraded. The detection of the layout entities is then based on intra-class similarities, in case of main body text, compact, rounded shapes are such structural similarities to exploit. As the whole entity cannot directly be inferred from the mere positions of the interest points, a localization algorithm is needed that expands the interest points according to their scales and the classification score to regions that encapsulate the whole entity. Hence, a cascading algorithm is proposed that successively rejects weak candidates applying voting schemes. The evaluations show that in case of decorative entities such as initials and headings, the detection and localization still poses challenges. For the main body, however, the identification works.

REFERENCES

[1] A. Antonacopoulos and A. Downton. Special Issue on the Analysis of Historical Documents. *IJDAR*, 9:75–77, 2007.

[2] M. Baechler, J.-L. Bloechle, and R. Ingold. Semi-Automatic Annotation Tool for Medieval Manuscripts. *ICFHR*, pages 182–187, 2010.

[3] S. Baluja and M. Covell. Finding Images and Line-Drawings in Document-Scanning Systems. In *ICDAR*, pages 1096 –1100, July 2009.

[4] I. Bar-Yosef, N. Hagbi, K. Kedem, and I. Dinstein. Line Segmentation for Degraded Handwritten Historical Documents. In *ICDAR*, pages 1161 –1165, 2009.

[5] 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. In *ICDAR*, volume 1, pages 357–361, 2007.

[6] M. Coustaty, J.-M. Ogier, R. Pareti, and N. Vincent. Drop Caps Decomposition for Indexing a New Letter Extraction Method. In *ICDAR*, pages 476–480, 2009.

[7] M. Diem and R. Sablatnig. Recognizing Characters of Ancient Manuscripts. In *IS&T SPIE Conf. on Computer Image Analysis in the Study of Art*, volume 7531, 2010.

[8] C. Grana, D. Borghesani, and R. Cucchiara. Automatic Segmentation of Digitalized Historical Manuscripts. *Multimedia Tools and Applications*, pages 1–24, 2010.

[9] F. Kleber, R. Sablatnig, M. Gau, and H. Miklas. Ancient Document Analysis Based on Text Line Extraction. In *ICPR*, pages 1–4, 2008.

[10] J. Landré, F. Morain-Nicolier, and S. Ruan. Ornamental Letters Image Classification Using Local Dissimilarity Maps. In *ICDAR*, pages 186–190, 2009.

[11] F. Le Bourgeois and H. Kaileh. Automatic Metadata Retrieval from Ancient Manuscripts. In *DAS*, pages 75–89, 2004.

[12] L. Likforman-Sulem, A. Zahour, and B. Taconet. Text Line Segmentation of Historical Documents: A Survey. *IJDAR*, 9(2):123–138, April 2007.

[13] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, 60(2):91–110, 2004.

[14] H. Miklas, M. Gau, F. Kleber, M. Diem, M. Lettner, M. Vill, R. Sablatnig, M. Schreiner, M. Melcher, and E.-G. Hammerschmid. *Slovo: Towards a Digital Library of South Slavic Manuscripts*, chapter St. Catherine's Monastery on Mount Sinai and the Balkan-Slavic Manuscript Tradition, pages 13–36. Boyan Penev, 2008.

[15] I. Moalla, F. LeBourgeois, H. Emptoz, and A. Alimi. Contribution to the Discrimination of the Medieval Manuscript Texts: Application in the Palaeography. In H. Bunke and A. Spitz, editors, *DAS*, volume 3872 of *Lecture Notes in Computer Science*, pages 25–37. Springer Berlin / Heidelberg, 2006.

[16] I. Moalla, F. Lebourgeois, H. Emptoz, and A. Alimi. Image Analysis for Palaeography Inspection. In *DIAL*, pages 8 pp. – 311, April 2006.

[17] J.-Y. Ramel, S. Leriche, M. L. Demonet, and S. Busson. User-Driven Page Layout Analysis of Historical Printed Books. *IJDAR*, 9(2-4):243–261, 2007.

[18] S. Uttama, J.-M. Ogier, and P. Loonis. Top-Down Segmentation of Ancient Graphical Drop Caps: Lettrines. In *GREC*, pages 87–96, August 2005.