2011 International Conference on Document Analysis and Recognition

Layout Analysis for Historical Manuscripts Using SIFT Features

Angelika Garz, Robert Sablatnig, Markus Diem
Computer Vision Lab, Institute of Computer Aided Automation, Vienna University of Technology
Vienna, Austria
Email: garz@caa.tuwien.ac.at

***Abstract*—We propose a layout analysis method for historical manuscripts that relies on the part-based identification of layout entities. A layout entity – such as letters of the text, initials or headings – is composed of a set of characteristic segments or structures, which is dissimilar for distinct classes in the manuscripts under consideration. This fact is exploited in order to segment a manuscript page into homogeneous regions. Historical documents traditionally involve challenges such as uneven writing support and varying shapes of characters, fluctuating text lines, changing scripts and writing styles, and variance in the layout itself. Hence, a part-based detection of layout entities is proposed using a multi-stage algorithm for the localization of the entities, based on interest points. Results show that the proposed method is able to locate initials, headings and text areas in ancient manuscripts containing stains, tears and partially faded-out ink sufficiently well.**

***Keywords*-Sift, part-based, layout analysis, document layout, handwritten, historical manuscripts**

I. INTRODUCTION

Contrary to modern, machine-printed documents, historical manuscripts require algorithms to be robust with respect to background artefacts such as clutter, stains and noise [1]. Historical handwritten documents do not have strict layouts rules and thus, a layout analysis method needs to be invariant to layout inconsistencies, irregularities in script and writing style, skew, fluctuating text lines, and variable shapes of decorative entities [1], [2]. Furthermore, robustness to low contrast – e.g. in case of faint ink – stains and rippled pages is required [3]–[5].

Layout analysis is the first step in the process of document understanding; it identifies and localizes regions of interest within document pages, hence it serves as input for further processing as the algorithms can be selectively applied to these regions instead of treating the whole document page. Optical Character Recognition (OCR) systems recognize and retrieve the actual character which correlates to the character written in the manuscript.

Initials can be extracted and further processed by decomposition algorithms [6], [7] that aim at determining the letter represented by the respective initial.

Gaining structural information about the manuscript pages by examining the layout, the manuscripts can be indexed and enhanced with meta data. The physical structure – spatial relationships, the number, type and modality of layout entities, positions and spatial extends –, the script, the scribe's personal writing style, the contents, author authentication and properties of the writing support may be exploited in order to generate meta data for documents [2], [8], [9].

An part-based approach independent of binarization is proposed. It detects and localizes layout entities based on their local structure. This allows detecting handwritten characters having a high variability in their shape. Owing to the use of local features, the method is independent of the physical and logical layout of a manuscript, such as constraints of potential locations of layout entities or spatial relationships between them.

The approach proposed in this paper was first applied to a Glagolitic manuscript dating from the 11^{th} century, which is part of a finding of 42 codices in St. Catherine's Monastery in the year 1975 [10]. The manuscript consists of 145 folia with a front and a back page each. It was digitized in the course of *The Sinaitic Glagolitic Sacramentary (Euchologium) Fragments Project* [10]. In the following, the layout entities considered in this paper with their respective characteristics are provided:

Main Body Text: The regular text; the characters have predominately curved strokes. The Glagolitic script does not conflate characters to words and thus, the horizontal spaces between characters are uniform.

Heading: Usually written in outline type, having a different aspect ratio and mostly angular shapes. Furthermore, non-Glagolitic characters, such as Cyrillic characters, are used for headings as well as Glagolica.

Initial: An initial character that is higher or broader than an average letter of the main body text. It is located within the left margin or within the main body text. Initials are optionally characterized by having outlines instead of single strokes. We distinguish between two types of initials: embellished and plain initials. *Embellished initials* cover more than two lines and are illuminated with tendrils, bows and hatch. The purpose of embellished initials is indicating a new section, respectively psalm, in the text. *Plain initials* are predominated by angular shapes.

The manuscript was written by four scribes who had followed different layout rules; about $75\,\%$ of the pages are written by one scribe with changing writing style and inconsistent layout rules (e.g. number of text lines, margin). Additionally, annotations have been added by a user. Approximately $50\,\%$ of the headings are written in Cyrillic

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.108

letters.

The following section gives an overview about the related work relevant to layout analysis on ancient manuscripts and historical printed books, and then in Section III, the proposed method is detailed. The subsequent section describes the evaluation and results, followed by a conclusion.

II. Related work

The data sets considered in traditional document layout analysis are printed documents rather than handwritten manuscripts. As Antonacopoulos and Downton point out in their paper [5], 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.

Ramel et al. [2] 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. Grana et al. [11] propose a system for historical manuscripts which distinguishes handwritten text, 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. In [3], 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 [4], a semi-automatic annotation tool for the generation of ground truth of layouts for medieval manuscripts is introduced. A SIFT-based image and line drawing detection system is proposed by Baluja and Covell in [12].

In contrast to state-of-the-art layout analysis methods, the proposed approach does not need a binarization step. This makes it robust to noise, background clutter and faded-out ink. As the data set considered does not have a strict, rectangular layout such as the documents considered in [3], a method invariant to skew, fluctuating text lines and differences in script and writing style is required. Color based segmentation is not suitable, as first, the decorative entities are not universally highlighted with a specific color and second, the highlight color is too similar to the background.

III. Method

The method introduced in this paper consists of two consecutive steps, where the first is the extraction and classification of features (see Figure 1 DoG, SIFT) and the second employs a multi-stage localization algorithm (see Figure 1 Marker Points, Voting). Both tasks are based on interest points computed by means of Difference-of-Gaussian (DoG).

A descriptor is calculated for every interest point (see Figure 1 SIFT) which describes parts of characters, or even whole characters or text lines. Scale Invariant Feature Transform (SIFT) are chosen as features due to their invariance to scale and rotation, which accounts for the changing script size and orientation in manuscripts. Furthermore, they are robust 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.

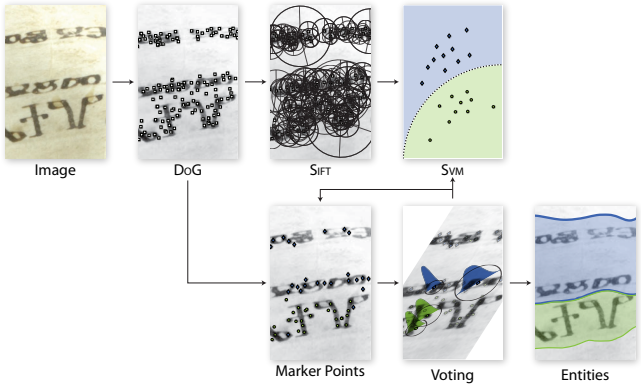

Figure 1. Workflow of the proposed layout analysis method: feature extraction and classification (top) and localization (bottom)

The descriptors are classified employing a Support Vector Machine (SVM) in order to discriminate between two classes: the main body text, and layout entities having a decorative meaning (see Figure 1 SVM).

A localization algorithm (see Figure 1 bottom row) is then needed to expand the interest points found into regions which enclose the whole layout entity, as the entity cannot be directly inferred from the positions of the interest points.

A text line segmentation method applied to the text regions in order to determine the actual lines of text is done employing the interest points found by the DoG.

The following subsections give the details of the methodical steps in the proposed algorithm.

A. Feature Extraction and Classification

Though interest points could be gained applying a scale-invariant corner detector such as the Harris corner detector combined with a Laplacian [13] or FAST [14], a corner and edge detector based on non-linear filtering, a blob-based detector is chosen on account of studies by [15], [16].

The DoG detects interest points at locations of local minima and maxima exploiting a scale space. These local extrema represent discriminative character parts such as junctions, circles, arcs or endings.

For each of the detected interest points, local features are computed using the SIFT descriptors proposed by [16]. The gradient magnitude and orientation are computed in the region of an interest point, where the level of Gaussian blur and the size of region is determined by the interest point's scale. The SIFT descriptors are 128-dimensional

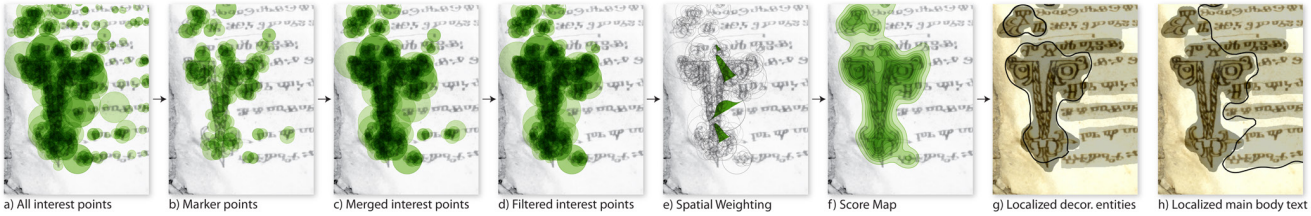

Figure 2. Overview of the localization algorithm

feature vectors containing the values of the 4×4 orientation histograms having 8 bins each.

To discriminate between the high-dimensional SIFT features describing main body text and entities having a decorative meaning, a SVM is trained. A Radial Basis Function (RBF) kernel is chosen in order to be able to non-linearly separate data.

B. Localization

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. Pursuant to this assumption, a six-step localization algorithm is introduced which successively reduces the amount of falsely classified descriptors.

Figure 2 gives an overview of the six stages of the localization algorithm for decorative entities. Interest points are indicated by green circles denoting their respective scale. The more intense the green color, the more interest points overlap. Figure 2 a)-g) show the decorative entities-class, f) relates to the main body text class. In detail, Figure 2 a) illustrates all decorative entities descriptors classified by the SVM. 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. Second, a set of marker points is established (b), which are reliable interest points indicating the position of a potential layout entity. Marker points are interest points selected from the second octave – these scales represent a certain segment size, which most reliably indicate a layout element – and have a high classification score. Second, the remaining interest points overlapping with at least one marker point at least 25% are merged to the set of candidate interest points (c). Then, a region-based processing step is employed, 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 (d). Thereafter, the interest points' scales and the previously weighted classification score are spatially weighted with a two-dimensional Gaussian distribution (e) leading to one score map per class. Figure 2 f) shows a score map, with ISO-lines indicating the scores. 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. Figure 2 g) presents the final result of the localization algorithm, and h) gives the final result for the main body text class. 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.

C. Text Line Segmentation

In order to segment the text regions further into actual lines of text, a spatial density clustering algorithm is adopted. The interest points found by the DOG are be exploited in order to segment the text lines, since interest points are mainly detected on and between characters. Thus, text lines can be identified by following the highest density of interest points since only few interest points are generated for background areas between the text lines.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [17] is applied to the coordinates of the final interest point set of the text class. Contrary to k-means clustering, the number of cluster centers does not need to be determined prior to clustering.

The DBSCAN algorithm takes two parameters – the minimum number of neighbors required to build a cluster and the neighborhood radius. The implementation of Daszykowskia [18] is used, where the neighborhood radius is analytically estimated. Since the algorithm is robust to noise, interest points lying between text lines do not deteriorate the result.

IV. EVALUATION AND RESULTS

The method proposed in this paper is evaluated on a random sample of 100 pages. The pages selected as test set have variations in the layout, scripts, writing styles, and writing instruments, and the character sizes within pages and between pages vary.

The training set consists of image patches of the respective classes; 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.

The evaluation is done on pixel-level having manually tagged ground truth. Since layout entities frequently touch each other or overlap, a determination of the class is not possible for the pixels in that region. Therefore, a $20px$ 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 as described before. This noise occurs especially in border regions of overlapping

Table I
PSALTER: PRECISION, RECALL AND $F_{0.5}$-SCORE

	Precision	Recall	$F_{0.5}$-score
Score maps at pixel-level			
a) Entire classification	0.924	0.873	0.914
b) Main body text	0.939	0.899	0.930
c) Decorative entities	0.667	0.513	0.629
Main body text – interest points			
d) First set of interest points	0.959	0.646	0.875
e) Voted points	0.961	0.637	0.873
f) Marker points	0.962	0.941	0.958
g) Merged interest points	0.966	0.896	0.951
h) Final set of interest points	0.964	0.962	0.964
Decorative entities – interest points			
i) First set of interest points	0.177	0.735	0.208
j) Voted points	0.168	0.742	0.198
k) Marker points	0.636	0.719	0.619
l) Merged interest points	0.506	0.772	0.543
m) Final set of interest points	0.713	0.727	0.715

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.

For the interpretation of the results in Table I, it has to be considered that the ratio between main body text and decorative entities is approximately 9.2:1 on pixel level (see Table I a-c).

Table I 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. This is mainly due to the fact, that for headings and plain initials, the difference to regular text is partly only the size and the angularity of the shapes as not all characters are written outlined. Even for humans who are no experts in the Glagolitic language, the differentiation between main body text and plain initials or headings is a non-trivial task.

Table I d-h) provide the results of the respective steps in the localization algorithm for the main body text class. The precision of the initial set of interest points for this class is high, which means that there are few interest points misclassified as belonging to the text class. The recall, however, which represents the fraction of interest points detected to those which should have been detected, is low. This expresses that either a high number of interest points have been classified to the decorative entity class despite belonging to the main body text.

The selection of marker points shows that the amount of interest points voting for the other class is reduced, however, adding the remaining interest points overlapping with the marker points, additionally adds interest points belonging to the decorative entity class. This happens due to touching and superimposing entities or initials within the text body which are outvoted by the interest points assigned to the class of main body text. The subsequent filtering mechanisms, however, reduce the amount of misdetected text regions.

Table I i-m) assesses the locations of the interest points assigned to the class of decorative entities by the classifier.

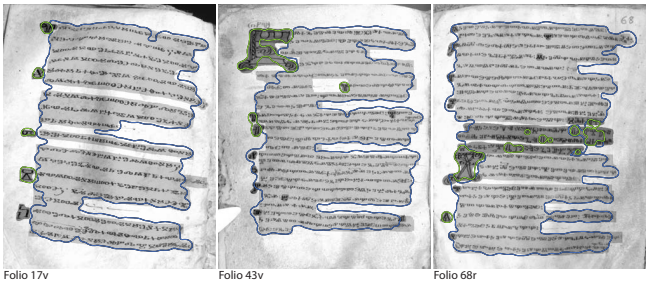

Figure 3. Exemplary results for whole pages. Green contours indicate the decorative entities detected by the algorithm, blue contours main body text.

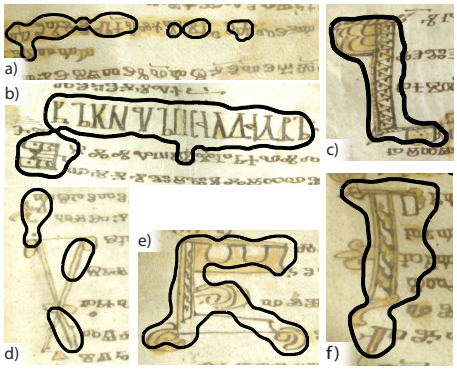

Figure 4. Exemplary results of decorative entities

While the precision is low, the recall is far higher. The marker points, however, are a much more reliable indicator of actual decorative entities. Merged with the remaining interest points and having applied the filtering steps to get the final set of interest points for the localization of decorative entities on the pixel level, the final $F_{0.5}$-score for the class of decorative entities was improved from 0.208 to 0.715

Figure 3 shows three pages of the Psalter; dark gray blobs indicate the ground truth for decorative entities and light gray blobs stand for the main body text class. Figure 4 gives exemplary results for the decorative entities as headings a-b), embellished initials c-f) and plain initials d,f).

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 since edges do not provide reliable interest points [16]. Hence, at these strokes, the density of interest points is low (see Figure 4 d,e).

A second issue relates to plain initials and headings having features not discriminative enough from the main body text, i.e. having a prevailing number of character segments characteristic for main body text, such as round, compact shapes. In Figure 4 a), a heading having characters similar to the main body text, is shown.

One aspect 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.

Additionally, class boundaries cannot always be deter-

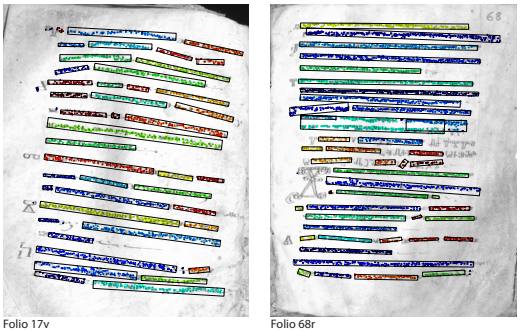

Figure 5. Exemplary results of clustered text lines indicated by black rectangles.

mined unequivocally as regions are overlapping or collide. 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 4 f).

The text line segmentation is only qualitatively evaluated since a ground truth is not available. Figure 5 provides example results for clustered text lines. The interest point locations are indicated by colorcoded dots according to their cluster, clustered lines are illustrated by minimum bounding rectangles. In Folio 68, the three lines above the embellished initial are headings. Just the parts misclassified to the main body text (see Figure 3) are included in the clustering. If the space between interest points in a text line is too large, the text lines are separated into chunks. A further problem are close text lines and annotations added between lines, since two consecutive lines are grouped into one cluster.

V. Conclusion

A part-based layout analysis method was introduced, which exploits structural similarities of layout entities employing local features. SIFT descriptors 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.

Future work includes the grouping of consecutive parts of a text line to an entire textline, quantitative evaluation of the text line segmentation method, and the verification of the grouping of the decorative entities into one cluster. A possible approach for demonstrating that the assumption of homogeneity of the local structures of this class is true, is plotting the principal components of their respective feature vectors. Additionally, identifying the actual object class of a decorative entity – since all decorative entities are considered as one class in the proposed method – is an issue.

VI. Acknowledgments

This work was supported by the Austrian Science Fund under grant P19608-G12.

References

[1] L. Likforman-Sulem, A. Zahour, and B. Taconet, "Text Line Segmentation of Historical Documents: A Survey," *IJDAR*, vol. 9, no. 2, pp. 123–138, 2007.

[2] J.-Y. Ramel, S. Leriche, M. L. Demonet, and S. Busson, "User-Driven Page Layout Analysis of Historical Printed Books," *IJDAR*, vol. 9, no. 2-4, pp. 243–261, 2007.

[3] 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*, vol. 1, 2007, pp. 357–361.

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

[5] A. Antonacopoulos and A. Downton, "Special Issue on the Analysis of Historical Documents," *IJDAR*, vol. 9, pp. 75–77, 2007.

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

[7] S. Uttama, J.-M. Ogier, and P. Loonis, "Top-Down Segmentation of Ancient Graphical Drop Caps: Lettrines," in *GREC*, 2005, pp. 87–96.

[8] F. Le Bourgeois and H. Kaileh, "Automatic Metadata Retrieval from Ancient Manuscripts," in *DAS*, 2004, pp. 75–89.

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

[10] 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*. Boyan Penev, 2008, ch. St. Catherine's Monastery on Mount Sinai and the Balkan-Slavic Manuscript Tradition, pp. 13–36.

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

[12] S. Baluja and M. Covell, "Finding Images and Line-Drawings in Document-Scanning Systems," in *ICDAR*, 2009, pp. 1096–1100.

[13] K. Mikolajczyk and C. Schmid, "Indexing Based on Scale Invariant Interest Points," in *ICCV*, 2001, pp. 525–531.

[14] E. Rosten and T. Drummond, "Machine Learning for High-Speed Corner Detection," in *ECCV*, 2006, pp. 430–443.

[15] K. Mikolajczyk and C. Schmid, "A Performance Evaluation of Local Descriptors," *PAMI*, vol. 27, no. 10, pp. 1615–1630, 2005.

[16] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[17] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *KDDM*, 1996, pp. 266–231.

[18] M. Daszykowski, B. Walczak, and D. Massart, "Looking for natural patterns in data: Part 1. density-based approach," *Chemometrics and Intelligent Laboratory Systems*, vol. 56, no. 2, pp. 83 – 92, 2001.