FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

CVL Computer Vision Lab

Technical Report
CVL-TR-7

Efficient Layout Analysis of Ancient Manuscripts Using Local Features

Angelika Garz

Computer Vision Lab
Institute of Computer Aided Automation
Vienna University of Technology

September 15, 2011

Abstract

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). A binarization-free layout analysis method for ancient manuscripts is proposed, which identifies and localizes layout entities exploiting their structural similarities on a local level.

The dataset, the method is evaluated on, is an ancient manuscript originating from the 11^{th} century, discovered at St. Catherine's monastery on Mt. Sinai, Egypt, in 1975 and digitized in the course of *The Sinaitic Glagolitic Sacramentary (Euchologium) Fragments* Project. It is written Glagolica, the oldest known Slavonic alphabet.

The document layout computed allows scholars to establish the spatio-temporal origin, authenticate, or index a document. The layout entities considered in this approach include body text, embellished initials, plain initials and headings. These textual elements 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).

These features describe the structures 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 it is robust to variations in shape, illumination, writing orientation and (background) noise. Thus, it is suitable for ancient handwritten documents with varying layouts and degradation effects.

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 evaluation shows that the method is able to locate main body text in ancient manuscripts. The detection rate of decorative entities is not as high as for main body text but already yields to promising results.

Contents

Chapter 1

Introduction

Document layout analysis is referred to the segmentation of a page into homogeneous regions consisting of layout entities belonging to the same class [113]. Figure 1.1 shows the segmentation of a modern machine-printed document into layout entities. The term layout entity is used for written objects or embellishments in a document. Examples are text areas (Figure 1.1 b,e,g,h), page numbers (Figure 1.1 f), headings (Figure 1.1 a,c), initials (Figure 1.1 d) or images (Figure 1.1 i).

Layout analysis is a topic widely addressed in literature (see Chapter 2), especially for machine printed documents. This report presents a method for analyzing the layout of ancient manuscripts containing decorative entities – such as initials and headings – and main body text. Figure 1.2 illustrates the decomposition of a page into its respective layout entities for the dataset regarded in the report. The layout entities considered are initials (Figure 1.2 a-e), the main body text (Figure 1.2 f,i) and headings (Figure 1.2 g,h).

Layout analysis of ancient manuscripts induces specific challenges not present in modern machine-printed documents and historical documents from the hand-press period [4, 8, 23] (see Section 1.3). To meet these challenges, a binarization-free layout analysis method is introduced which identifies and localizes layout entities exploiting their structural similarities on a local level. Hence, the textual entities are disassembled into segments employing interest points. The structures of these segments are then described using local gradient features, leading to a part-based detection. Figure 1.3 gives an illustration of a layout entity (an embellished initial) disassembled into circular parts which represent its local structures. A set of overlapping parts describe the entity.

The approach introduced in this report is developed for a 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. The monastery is located on the foot of the Mount Sinai in Egypt and is the oldest continuous existing Christian monastery. The manuscript collection of St. Catherine's Monastery consists of more than 3,000 codices and hundreds of book-scrolls originating in the 4^{th} century onward [96]. Section 1.3.1 will give further information about the manuscript.

This report was developed within the project *The Sinaitic Glagolitic Sacramentary (Euchologium) Fragments*, an interdisciplinary project of computer science, material analysis and philology. Three codices were digitized in the course of the project in 2007 by means of multi-spectral imaging.

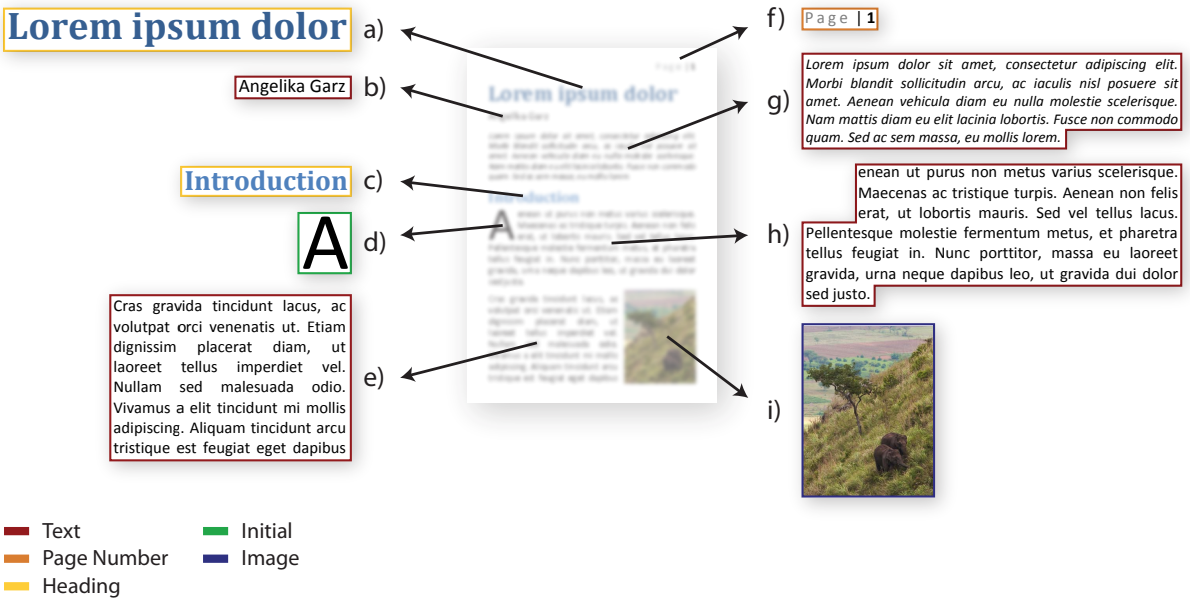


Figure 1.1: The concept of document layout analysis.

This chapter will first give the motivation for the report and will state the relevance of the topic of layout analysis of documents and especially ancient manuscripts. The next two sections will outline the scope of discussion and the objective of the report, followed by a section pointing out the main contribution of this report. Definitions of frequently used terms will be given in Section 1.2. The specifics and characteristics of ancient manuscripts and their implications on processing are described in Section 1.3, including description of the datasets regarded in this report. An overview of the methodology is provided in the subsequent section. Results will be summarized in Section 1.5, and an outline of the report completes this chapter.

1.1 Motivation

Analyzing the layout of manuscripts has two major areas of application: first, information about the layout can be employed for further processing, and second, scholars studying ancient manuscripts are supported in their studies with additional information and tools.

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 by selected areas to be analyzed. A binarization-free Ocr-system is proposed by Diem [38]. In Figure 1.4, a sample output for an Ocr system for ancient manuscripts is shown, with an image patch taken from a Glagolitic manuscript (left). On the right, the characters are annotated with their respective Ascii equivalent, correctly recognized characters are indicated by green, falsely recognized ones by red blobs.

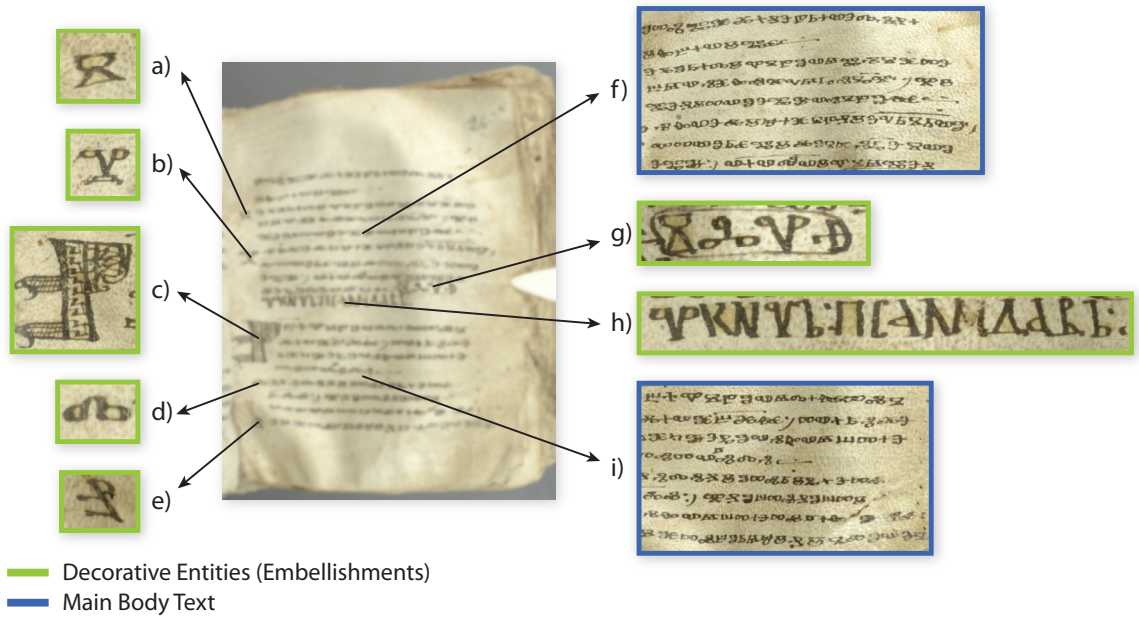


Figure 1.2: The concept of document layout analysis applied to the dataset regarded in this report.

Proposing an approach relying on local structures – and hence, parts of characters – for identification and localization of textual objects, allows to construct a Multilingual Optical Character Recognition (MOCR) system since different scripts can be distinguished by the method (see Section 2.4). A MOCR system takes texts of multiple languages as input, however, it requires the identification of the scripts prior to processing [116].

In addition to providing regions of interest for further processing, locations of the regions segmented by layout analysis and their spatial relations can be used to determine the sequence of text blocks and thus, the correct reading order [115].

Embellishments and ornamental letters detected by layout analysis can be extracted and further processed by decomposition algorithms [34,79,135] aiming at determining the letter represented by the respective initial (see Section 2.5).

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 [83,105,115]. This allows for an electronically searchable format of the document [115].

However, indexing systems based on computer vision cannot completely replace scholars studying these manuscripts, since expert knowledge is required, which is impossible to incorporate in an automated analysis system [83]. However, an automated system is able to provide assistance to an expert user.

Since the report is embedded within an interdisciplinary project including philological and paleographical processing of the manuscripts, a major task is the support of these studies with the means of computer vision and image processing. Philology stud-

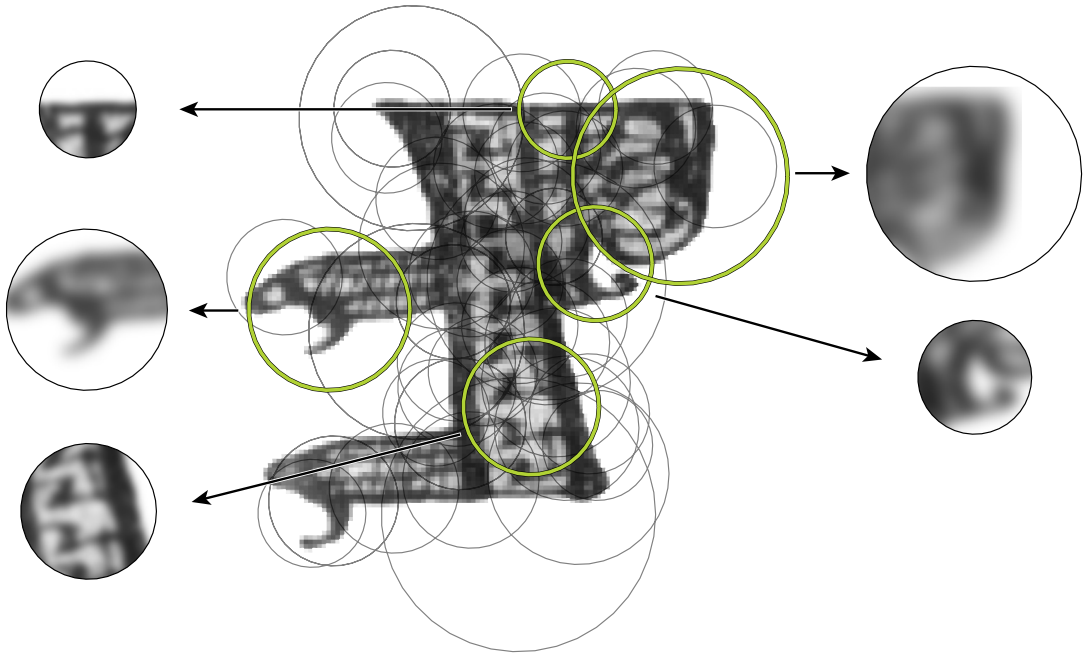


Figure 1.3: Illustration of disassembling an entity into its local structures.

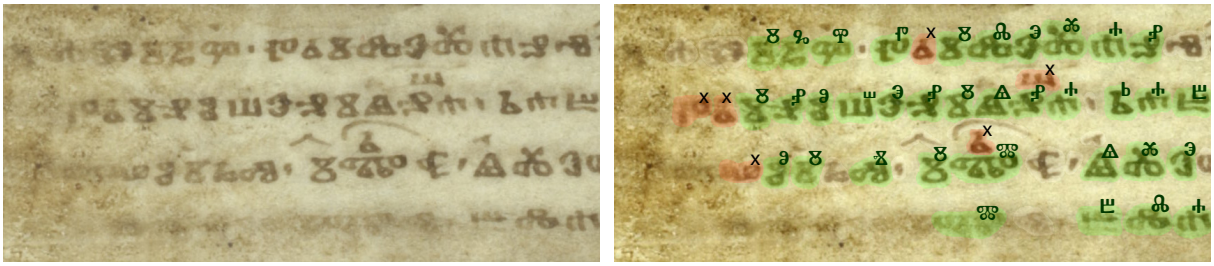


Figure 1.4: Sample output for Diem's [38] OCR system for ancient handwritten text (Figure taken from [38]).

ies historical languages, including their relationships, structure, grammars and historical evolution[1] [104, 105]. Thus, the actual content of the manuscripts is the basis their work builds upon. Paleography, on the other hand, is based on the layout of manuscripts and their historical development. Tasks of paleographers include dating historical manuscript and relate them geographically, reading, and deciphering and interpreting texts and writing systems, and further the identifying the scribes[2] [11, 88, 104, 105].

During their creation, manuscripts were edited by multiple hands. Additionally, scribes – who predominantly were monks – traveled between monasteries, and thus, manuscripts were hauled to different places whilst authoring. The work of scribes is confined to the creation of the main body text, where spaces are left out for embellishments. Ornamental letters and headings were added after writing by different scribes or illustrators.

[1] "philology, n.". OED Online. March 2011. Oxford University Press. (accessed April 21, 2011). http://www.oed.com/view/Entry/142464?redirectedFrom=philology

[2] "palaeography — paleography, n.". OED Online. March 2011. Oxford University Press. (accessed April 21, 2011). http://www.oed.com/view/Entry/136180

Layout rules, scripts and writing styles evolve during time and thus, are characteristic for each historical period and geographical localization [105].

Likforman et al. [88] point out that paleographic authentication of a manuscript benefits from document analysis since it is based on the writer and layout characteristics rather than on the content of a manuscript. Thus, the physical layout and the characteristics of the layout entities themselves as well as *"features extracted from blank spaces, line orientations and fluctuations, word or character shapes"* [88], inks used by the scribes, the texture of the writing support, or annotations added at a later date allow scholars to determine the temporal and geographical origin of a manuscript, and hence, establish its history and provenance [83].

Characteristics of embellishments such as ornamental letters provide information clues for spatio-temporal relating a document or parts of it as well as determining the circumstances of its creation. Furthermore, due to the distinctive design of the embellishments, the script, and the personal writing style, the scribe and the illustrator can be identified in one and over various manuscripts. Thus, having identified the layout entities of an entire manuscript with an automated process, scholars are enabled to retrieve and to directly work on them without the need of browsing through the document images, searching and extracting the embellishments manually.

1.1.1 Objective

The aim of this report is the development of a layout analysis system for ancient manuscripts robust with respect to their characteristics as described in Section 1.3. The goal is to split images of manuscript pages into homogeneous regions on pixel level in order to obtain a physical layout structure. Hereby, main body text, headings and initials are considered as layout entities to be extracted from document pages. The algorithm is applied to images of manuscript pages and splits the document image into homogeneous regions of interest, namely regions of the main body text, headings and initials.

The approach is developed to analyze the layout of an ancient Glagolitic manuscript probably from the 11^{th} century, which is further described in Section 1.3.1. Furthermore, two medieval manuscripts from the Austrian Stiftsbibliothek Klosterneuburg are regarded (see Section 1.3.2).

Ancient manuscripts cannot be binarized in a satisfactory way since clutter, noise and degradation traces may appear as foreground objects and faint ink vanish into the background [4, 23]. The aim is to develop a method which is independent of a prior binarization step but is directly applied to grayscale images and thus, more robust to noise and clutter.

An important objective is the robustness of the method to variations in script and writing style, since handwriting is subject to deviations e.g. in character shapes, skew, size and shifts or fluctuations of the baseline. However, despite being robust to variations, the method has to be discriminative with respect to different scripts.

In contrast to medieval manuscripts, distinctive colors for embellishments are not used in the ancient Glagolitic considered in this report. Thus, a method relying on grayscale information, independent of color segmentation is to be developed.

Introducing an approach independent of binarization, and relying on a part-based

identification of layout entities, the localization of entire objects is an issue to solve. Thus, a localization algorithm that infers the location and extend of an object based on a set of parts is a further goal of this report.

1.1.2 Scope of Discussion

A system for layout analysis of ancient manuscript is introduced that is able to identify and locate textual entities and embellishments in a document page image. The layout entities considered in this report are headings, initials and main body text. Since the regarded manuscripts do not contain images, drawings or ornamental frames, these are not included in the study.

Due to the structural similarities of their parts, decorative entities are assigned to a single class (see Section 1.3.1 for more details on this topic). A further distinction of decorative entities into their subclasses is not within the scope of this report. Features used for the identification and localization of layout entities can be used for text line segmentation within detected text blocks; however, this is not part of the report.

The analysis method proposed in this report leads to a physical structure of the manuscript pages. A further processing of assigning a functional labeling to the layout entities is not intended. In Section 2.1, physical and logical document layout structures are further explained. However, the class membership of each entity provides a rough categorization of the entities through segmenting the page by means of classification.

The images of the manuscript pages contain the page borders and surrounding areas as well as parts of the opposite page. However, the determination of the page border is beyond the scope of the approach. The processing area is manually restricted within the page border. Existing page border detection methods are based on binarization of the document image [123, 124, 129]

The method is additionally evaluated on two further manuscripts having different characteristics (described in Section 1.3.2) in order to show the flexibility of the system.

Accurate object localization based on local features is an open issue in object recognition [78, 98]. The detection method for line drawings in book collections proposed by Baluja and Covell in [9] does not include an approach for the exact localization of the drawings. The cluster-center-based approach introduced in [38] for identifying the local features belonging to one character groups the interest points, but does not provide an accurate localization on pixel-level. Additionally, this method relies on characters having approximately the same size, whereas initials and headings regarded in this report may have an arbitrary spatial extend.

1.1.3 Contribution

Local features are developed for the use of object recognition, where an exact 1:1 matching of the same object in different images is a major application [93]. In this report, local features were studied for their application in document layout analysis of ancient manuscripts. Thus, an established object recognition method is introduced in the field of document image processing. Whereas in object recognition, similar objects are to be

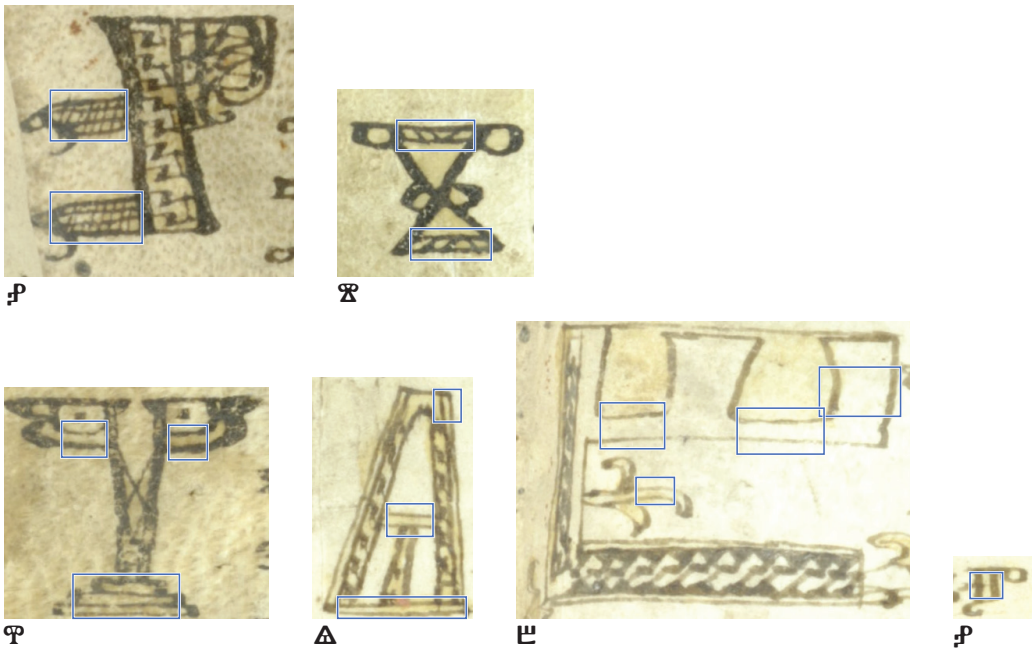


Figure 1.5: Structural similarities amongst initials (per row). Glagolitic ⰻ and ⰾ share hatches (first row), in the second row, the Glagolitic characters share outlines.

found, the method introduced relies on similar local structures of objects that are completely different but have a set of shared characteristics. Rather than performing a direct matching algorithm, or finding the same object – e.g. a horse, regardless if a Shetland pony or a Shire horse –, a machine learning approach is employed to learn structures, a specific script or a layout entity consists of – e.g. as an initial, a Glagolitic ⰾ and ⰻ may share common structures. As illustrated in the first row of Figure 1.5, these initials have hatches. Examples for outlines at different scales shared between the four initials are provided in the second row of Figure 1.5. It has to be pointed out that in case of the ⰻ, the stroke width is different to the other examples when scaled to their size.

Hence, a new layout analysis concept is introduced where layout entities are considered as parts having similar structures. These structures are exploited in order to identify objects having completely different shapes but sharing common structures.

This report introduces a localization algorithm for local features without the need for binarization. While with binarization the localization is inherently given, an approach to extend dedicated interest points to encapsulate a whole object is necessary for local features.

1.2 Definition of Terms

This section provides definitions for terms and abbreviations frequently used in this report. A complete list of abbreviations used in this report, however, is given in the List of Acronyms 6.3 in the appendix.
Terms concerning the manuscripts and philological terms are explained in the following:

***Cod. Sin. Slav.* 3n** see *Psalter*.

Embellishment Decorations of handwritten texts, ornamental patterns.

Folio A sheet in a bound book, having a front page (recto) and a back page (verso).

Glagolica Glagolitic script/alphabet created by Konstantin Kyrill in 862 AD, today named Church Slavonic [95]. It is based on the Greek script.

Glagolitic The oldest known Slavonic dialect.

Hand See *Scribe*.

Layout Entity A textual element such as a heading or a text block, or an embellishment of the document like an initial. Entities may consist of one or more elements (letters). The respective layout entities regarded in this report are described in detail Section 1.3.

Palimpsest A manuscripts on parchment or papyrus, which was reused by erasing and overwriting.

Personal Writing Style A scribe's personal style of writing, the scribe-specific shaping of characters. Different styles of writings possibly exist in one manuscript due to different scribes.

Psalter Book of Psalms. In this report the term *Psalter* is used to refer to the *Old Church Slavonic Glagolitic Psalterium Demetrii Sinaitici*.

Old Church Slavonic Glagolitic Psalterium Demetrii Sinaitici A Psalter – i.e. a manuscript written in Glagolitic script from the 11^{th} century, found at St. Catherine's Monastery on Mount Sinai, Egypt.

Scribe The writer or copyist of a manuscript, i.e. usually a person, e.g. a monk, who copied a manuscript prior to the printing epoch.

Script While one script may have different writing styles, different scripts are referred to as being distinctive system of writing, such as Glagolitic, Cyrillic, or Latin.

Writing Instrument A tool used to write the text, e.g. quill, pen, or brush.

Writing Style One script may have different writing styles, e.g. there exist two major styles for the Glagolica: the Bulgarian or round Glagolica, and the Croatian or square Glagolica. Figure 1.6 gives a comparison of two writing styles of Glagolica.

Writing Support The material, a manuscript consists of, more precisely the material, the text is inscribed on, e.g. parchment, papyrus, or paper.

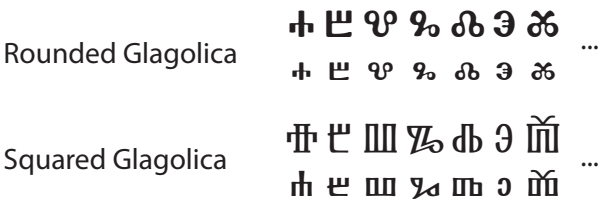


Figure 1.6: Comparison of two writing styles of the Glagolitic script

Abbreviations of the methods used in this report are given below:

DoG **Difference-of-Gaussian** An approximation of the Laplacian-of-Gaussians (LoG) that is built by successive differentiation of image representations blurred with Gaussian kernels having increasing values of their standard deviation [93]. Hence, the Difference-of-Gaussian (DoG) can be considered as a set of band-pass filters successively suppressing high-frequency structures and creating low-frequency structures. This allows for extracting blob-like structures having different scales matching the respective frequencies allowed to pass by the DoG. Thus, interest points having a specified position and spatial range can be computed. The DoG is described in detail in Section 3.1.4.

LoG **Laplacian-of-Gaussians** An image is convolved with a Gaussian kernel followed by the computation of the Laplacian operator in x and y direction. A scale-normalized LoG is generally used in order to be able to extract structures having different scales.

Sift **Scale Invariant Feature Transform** A local descriptor for characterizing image regions previously extracted by an interest point detector such as the DoG in a rotation invariant manner [93]. The descriptor is built of a three-dimensional gradient histogram which is based on the location, orientation and magnitude of gradients in the local area of an interest point. Sift is detailed in Section 3.2.3.

Svm **Support Vector Machine** A supervised machine learning algorithm for solving a binary optimization problem. It is based on the *Structural Risk Minimization* principle, where a generalized model is to be fitted to prior known dataset such that guarantees the minimum true error. Hence, the Support Vector Machine (Svm) minimizes the overall risk, which results in a good generalization performance [65].

1.3 Ancient Manuscripts

Through the centuries, ancient manuscripts decay by reason of inappropriate storing conditions, on-purpose destruction, deterioration processes, mold and moisture [4, 74], and materials they consist of, such as parchment, papyrus or paper. This results in torn pages and veined writing support having heterogeneous background intensity, artefacts due to aging, smudges, and stains [10, 88].

Ancient manuscripts include disturbing elements such as holes in the writing support, spots and ink stains [88]. Further problems are faint ink and bleed-through, which means

ink seeping through from the other side of the folio, and resulting in writing from the backside appearing on the front side. Pages suffer from scratches, crease and are corrugated, on the one hand, due to characteristics of the writing support and, on the other hand, by reason of being bound as a book [10, 74, 88]. Additionally, unequal lighting during the digitization process and uneven writing support are challenges to cope with [74, 88]. Further possible degradations of parchment are detailed in the work of Fuchs [50].

Palimpsests are manuscripts on parchment or papyrus, where the text was erased and overwritten with a new text. In case of parchment, the common procedure was to scrape off the existent text to be able to reuse the writing support. For Papyrus, the procedure was usually washing off the old ink. Writing support was reused since it was a scarce, costly product. The most famous palimpsest is the Archimedes Palimpsest[3] [139].

Pursuant to the type of manuscript – whether it has a rather unstructured layout with loose guidelines such as free flowing text of a novel, or a layout having a physical-logical structure like forms or letters –, different degrees of freedom need to be considered [4]. The following paragraphs describe layout variations more likely to occur in prevailingly unstructured manuscripts like the dataset regarded in this report.

The physical structure of ancient manuscripts is harder to extract than this of printed books since layout formatting rules of these manuscripts were looser and are not always complied with [88]. Handwritten documents may include non-rectangular layout, irregularities in locations of layout entities, or multiple scripts [88, 115]. Text line spacing of ancient handwritten documents varies between narrow spaced lines with overlapping and touching components, and double spaced lines [88]. This variance in spacing may even appear on a single page. Furthermore, these manuscripts contain interfering lines running into each other, annotations added between lines or in the page margin as well as non-constant spacing between characters [88]. Skewed text lines and blocks, shifts and fluctuations of the baseline have to be considered. Additionally, historical documents contain layout entities *"which are graphical in nature, although they represent text"* [111], such as embellished initials.

Historical manuscripts show variations in script and writing style. One reason is the way, manuscripts were created – they were edited by multiple hands as described in Section 1.1, because scribes had different tasks. A scribe is responsible for the main body text, another one adds the embellishments, and a further scribe is engaged with the headings. Another reason lies in the fact that manuscripts were moved during their creation. Scripts and writing styles are spatio-temporally characteristic and thus, the main body text is subject to changes in the hand or writing style [105]. Even the personal writing style of a scribe is inconsistent and thus, subject to irregularities such as unusual and variation of character shapes [88].

1.3.1 Old Church Slavonic Glagolitic Psalterium Demetrii Sinaitici

Having depicted the characteristics of ancient manuscripts in general, the specific challenges of the main dataset considered in this report will be described. The approach introduced in this report is applied to an ancient manuscript probably from the 11^{th} cen-

[3] http://www.archimedespalimpsest.org

tury, the *Old Church Slavonic Glagolitic Psalterium Demetrii Sinaitici*, more precisely, the *Cod. Sin. Slav.* 3N [96].

The *Psalter* is named after one of its early users, who added annotations to the main text of the manuscript. It is the second oldest Slavonic Psalter manuscript and was part of a finding of 42 manuscripts discovered at St. Catherine's Monastery on the foot of Mount Sinai in 1975 [96]. The manuscript is written in Glagolica, the oldest known Slavonic alphabet. The writing support of the manuscript is parchment.

The *Psalter* consists of 145 folia, each of which has a front page (recto, r) and a back page (verso, v). It comprises 151 psalms and secondary insertions [96]. The *Psalter* was digitized in the course of the project *The Sinaitic Glagolitic Sacramentary (Euchologium) Fragments* [96].

The manuscript has been studied in order to infer characteristics of the layout entities which are invariant and can be exploited for layout analysis. In the following, the layout entities considered in this report with their respective characteristics are provided:

Main Body Text The regular text building the continuous text consisting of the main text excluding objects such as annotations, numerations, notes, initials, headings. In the *Psalter*, it is organized in one column. The Glagolitic language and script does not conflate characters to words and thus, the horizontal spaces between characters are uniform.

Decorative Entities Characters not belonging to the main body text but having a decorative meaning such as initials or headings.

Initial An initial character that is higher and/or broader than an average letter in the main body text. It is located at the left side of the text body or within the main body text and usually highlighted with a so-called yellow wash. Moreover, initials are optionally characterized by having outlines instead of single strokes. We distinguish between two types of initials according to their meaning and decorations: embellished and plain initials. Subsequently, the term initial is used to refer to both, embellished and plain initials.

Embellished Initial A large initial covering more than two lines. It is illuminated with tendrils, bows and hatch, often with small dots in the middle of bows. The purpose of embellished initials is to indicate a new section, respectively psalm, in the text. Embellished initials may touch or overlap with other layout entities and reach into the area of the main body text. The variability in the appearance of initials denoting the same letter is high. Figure 1.11 gives two examples where the same character is represented by initials having different shapes, decorations and appearances.

Plain Initial An initial characterized by an aspect ratio different to those of the letters in the main body text, which means that either the vertical or horizontal expansion of the initial is larger. Additionally, the individual continuous strokes are longer for plain initials than for characters in the text body. This results in a less compact shape of the character. Besides, the characters of the main body text have curved strokes whereas the initials are predominated by angular shapes. Depending on the scribes personal writing style, however, plain initials located within the main body

Figure 1.7: *Psalter* – Initials denoting the same character.

text are frequently rather indicated by punctuation, a blank space larger than the usual space between characters or a continuous horizontal wave-form stroke, than by characteristics of the letter itself.

Heading Similar to initials, headings are usually highlighted with yellow wash and written in outline type. They have – such as plain initials – a different aspect ratio and mostly angular shapes. Furthermore, non-Glagolitic characters, such as Cyrillic letters, are used for headings as well as Glagolica.

Figure 1.8 compares Glagolitic ⰖⰅ and ⰐⰐ in use as embellished initial, plain initial, in headings and in the main body text. The initials and the heading characters share characteristics as described before. These characteristics include outlines, elongated strokes, or angular shapes.

In literature concerning document layout analysis, initial letters are not frequently addressed as a class (see Chapter 2). However, there exist various methods for the analysis of ornamental letters such as drop caps [70, 79, 83, 111, 114, 135], refer to Section 2.5 for more detail on these works.

Initials regarded in this report are, however, different in their appearance to the drop caps the cited authors engage in. A drop cap consists of a letter embedded in artwork, such as a line drawing with crosshatch [135], and thus, the outer shape of the initial is a geometrical shape such as a circle or square. In contrast to these drop caps, the shape of the initials regarded in this report correspond to their character's shape, the artwork is embedded in the character itself. The character itself is decoratively embellished with the shape of the character varying dependent on the scribe. Figure 2.13 in Section 2.5 compares the drop caps described above and a typical initial of the *Psalter* regarded in this report.

Figure 1.9 provides an overview of sample pages of the *Psalter* which incorporate the specifics ancient manuscripts. In the following, the challenges and problems of the

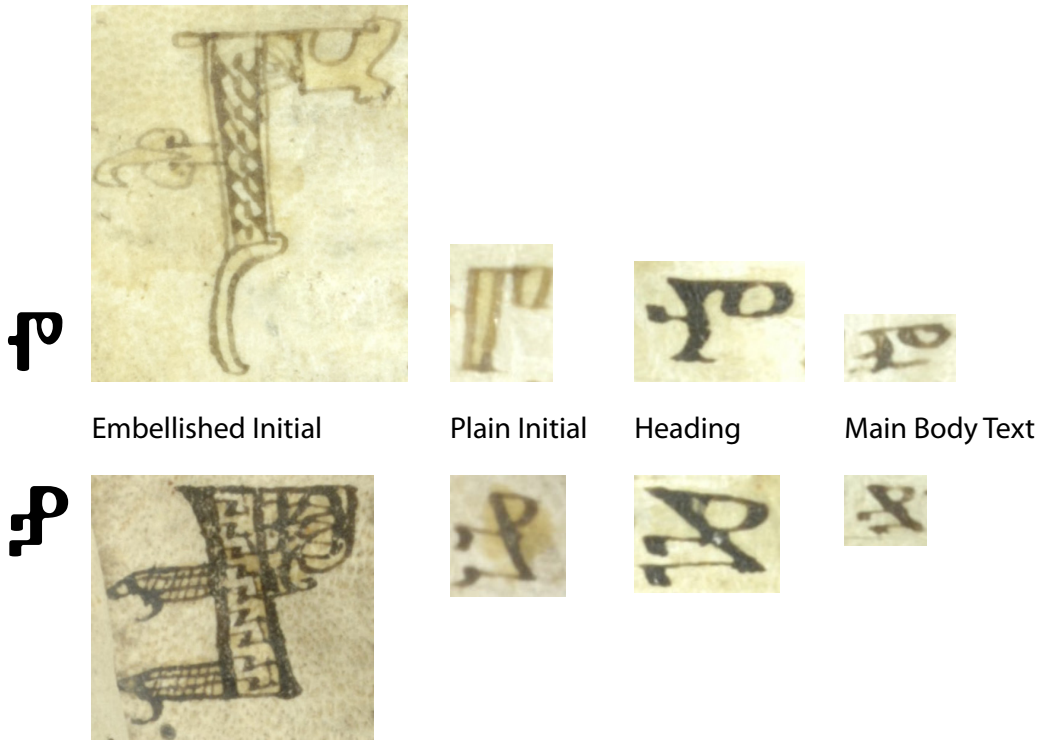


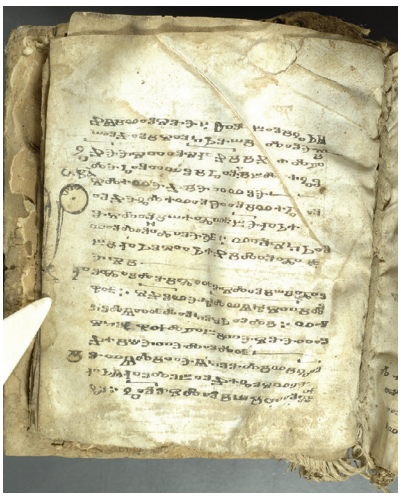
Figure 1.8: Comparison of ꝑ and a ꝓ in use as initials, in headings and in the main body text.

manuscript are described by means of giving examples.

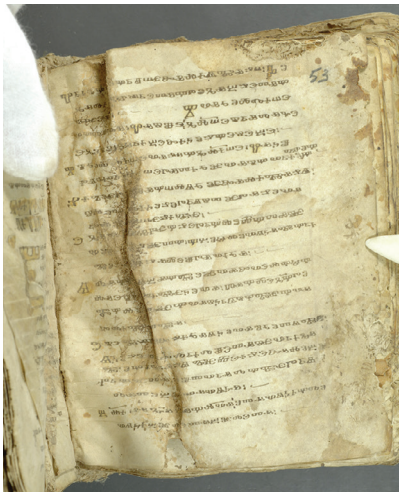
The writing supports of Folia 4v, 53r, 122r, and 139 are uneven. The lighting in Folio 84v is not optimal, parts of the page are overexposed, and thus, lacking structure. Demolition such as water stains, (4v), damage of the parchment (53r, 145r), heterogeneously textured background due to the characteristics of the writing support (100r, 145r), holes and cut outs (145r – a part of 144v is visible with a cut-out), as well as variance in the shape of the page or missing parts (84v, 122r, 140r) can be seen from the sample pages. Three-dimensional structures such as creases, folds, and wrinkles introduce artificial texture in two-dimensional images. The ink is faded out in Folia 139r, 140r, and 145r. Folio 140r is palimpsested twice – the first text is visible in the background of the whole page, the second text is visible in the top half of the page, and where the second text layer is faint, a third text is on top.

Layout rules have not strictly been applied to the manuscript, e.g. the number of text lines is prevailingly 24 (53r, 70r, 139r, 140r), however, its possible range more variant even for the given sample pages: 18 (4v), 20 (145r), 25 (84v), 26 (122r), and 27 (100r). The text body is not strictly rectangular; frequently, it is not horizontally aligned to the page, but skewed (139r, 14r, 145r), or follows the shape of the page (84v, 140r). The margin between the layout entities and the page border is variable. In case of recto-pages, the initials are close to or touch the bookbinding (122r, 145r); in one case, the initial does not fit the page (4v).

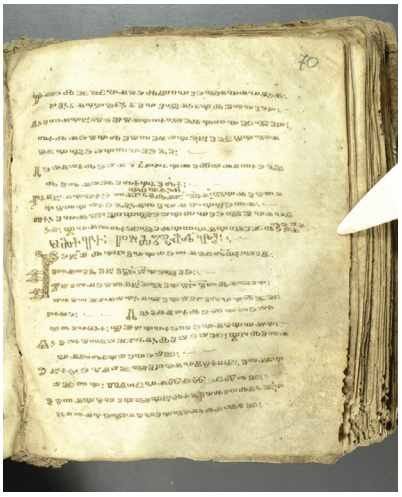
In Folio 84v, plain initials within the text body – indicated by horizontal strokes or a space in the text line – have the same appearance in terms of local structure as the main



Folio 4v

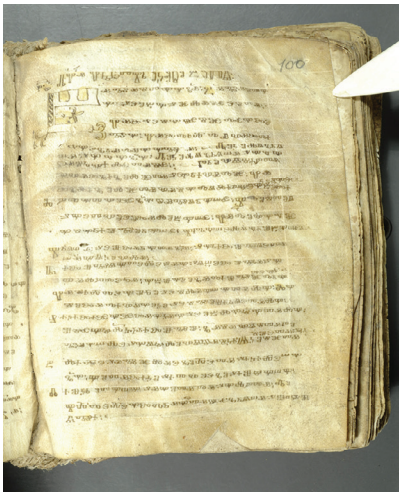


Folio 53r

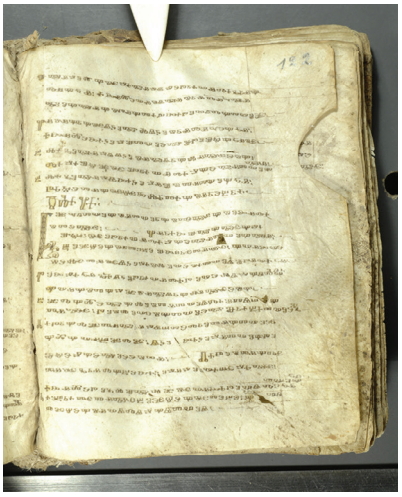


Folio 70r

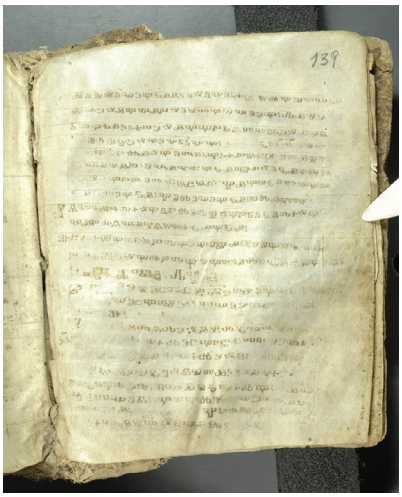
Folio 84v



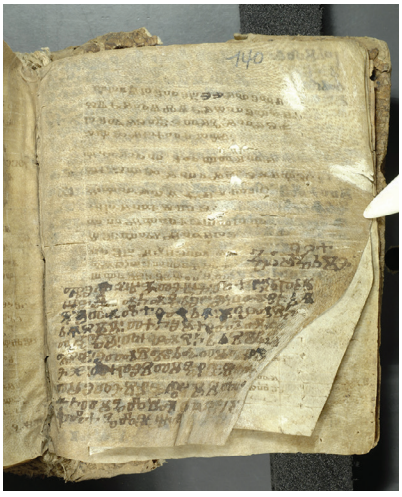
Folio 100r



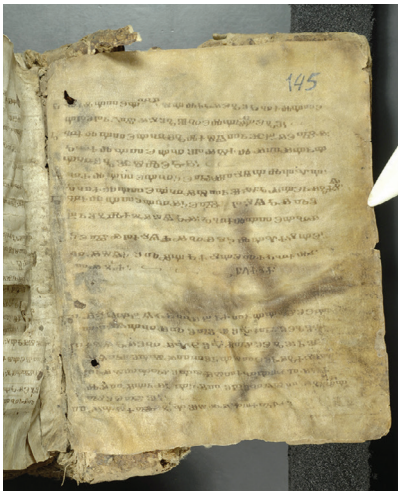
Folio 122r



Folio 139r



Folio 140r



Folio 145r

Figure 1.9: Sample pages from the *Psalter*.

body text. Plain initials usually located within the left margin of the text, are positioned as first character in the text line (84v, lines: 1, 5, 6). Embellished initials reach into the text body (100r, 122r).

The spacing between the text lines varies, even within one page (84v, 100r, 122r). The manuscript further comprises lines that converge into each other (100r) and insertions between two text lines at the end of the line (84v, 100r). Apart from skewed text blocks, lines are skewed and have baseline fluctuations (53r, 70r) or appear fluctuating due to three-dimensional distortion (4v, 70v, 122r). The text shows different character sizes (compare 4v and 84v, 140r) and high variation in the main body text characters' shapes (compare 4v, 70r, 140v, 145v).

1.3.2 Other Datasets

Additionally, the approach is applied to two medieval manuscripts from the Austrian Stiftsbibliothek Klosterneuburg in order to test the generalization ability to different scripts and layouts. The manuscripts under consideration are the following:

Codex Claustroneoburgensis 635 (*Cod. 635*[4]), a manuscript called *Ordinarium Divini Officii Secundum Consuetudienm Ecclesiae Collegiate Claustroneoburgensis*; it is originating from the 14^{th} century [58]. Similar to the *Psalter*, the manuscript consists of parchment. It comprises 120 folia written in Latin.

Codex Claustroneoburgensis 681 (*Cod. 681*[5]), a manuscript created in 1396 as a collection of medical texts [17], where the main part (Folia 13r - 159r) comprises the *Liber medicinalis*. It comprises 167 folia written in German and Latin language. In contrast to the other two manuscripts, the writing support of this manuscript is paper.

Cod. 635 and *Cod. 681* are not as decayed as the *Psalter*. However, bleeding-through ink, red colored text and staining are challenges of these documents. Both manuscripts are written in Gothic type and have strict rectangular layouts. In the first manuscript, the text is organized in one column whereas the second one has two columns. In *Cod. 635*, the text is justified to the left and right, whereas in *Cod. 681*, the text is justified to the left, resulting in a rough text line profile on the right side of a column. The manuscripts contain main body text and initials which are not decorated. The initials are written in red or blue ink and consist of bold and narrow continuous strokes. The initials are either in the left margin of the text body or embedded in the text body, which means that the initial is situated within the margin and runs several lines deep into the text. Furthermore, *Cod. 681* contains headings which are written in red ink but are not different in their local structure to the main body text.

The two manuscripts are inscribed in Western Europe calligraphic script. While in the Glagolitic script the horizontal space between characters is uniform, the Latin script distinguishes between words.

[4]Permalink: http://manuscripta.at?ID=830
[5]Permalink: http://manuscripta.at?ID=885

Figure 1.10: *Cod. 635* – Initials denoting the same character.

In contrast to the *Psalter*, the initials of the Latin manuscript are not richly embellished, and thus, have less structures which can be exploited to detect them. The initials of *Cod. 635* are written in calligraphic style and mainly consist of long strokes which change their thickness depending on the angle of the stroke. The initials are decorated with serifs. Figure 1.10 shows the possible variation in the appearance of initials in *Cod. 635*. The initials are touching text areas.

The initials of *Cod. 681* are similar to those of *Cod. 635*, however, have more variation in the stroke thickness. The initials are partly additionally decorated with dots connected to the strokes. Contrary to the strict layout, there is a high variability of actual character shapes, analogous to the initials in the *Psalter*. Figure 1.11 provides two examples for the variability in the appearance of initials in *Cod. 681*.

Figure 1.12 provides an overview of sample pages of *Cod. 635*. In contrast to the *Psalter*, the layout rules are strictly followed – the guide lines are still visible on the pages – and the writing style is consistent apart from one page (117r). Ink is bleeding through from the other side of the folia and stains are sprinkled over the pages. Initials are not richly embellished when compared to the *Psalter*, but rather plain. However, there exist a few embellishments in the text such as the head in Folia 83v and 100v, or elongated character strokes as in the first text line of Folio 107r. Additional annotations in the left margin of the text (33v, 100v) or on the bottom of the page (89v) can be found. The main body text is written in black and red ink, initials are drawn with blue and red ink. Headings cannot be found in the manuscript. The variance in the shape of the initials (e.g. compare the letter *D* in Folio 107r) is not as high as for the *Psalter*.

Figure 1.13 shows sample pages of *Cod. 681*. Similar to *Cod. 635*, the layout rules are complied with in this manuscript, annotations are added outside the text body (70v, 83v, 93v, 147v). The initials are not as decorated as in the *Psalter*; they are embedded

Figure 1.11: *Cod. 681* – Initials denoting the same character.

in the text body and frequently touch the characters of the main body text. Red ink is used for the initials for the sentence above the initial.

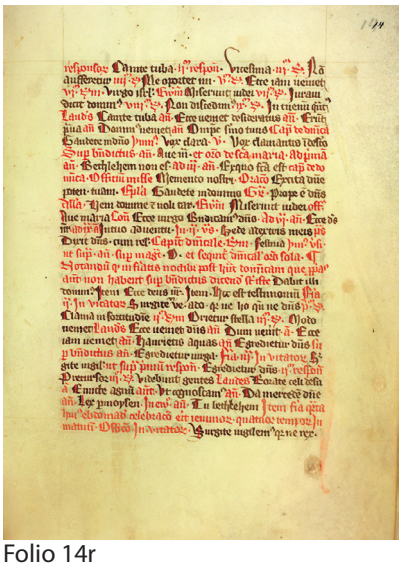
1.4 Methodology

As Antonacopoulos and Downton [4] and Bulacu et al. [23] point out, traditional approaches for the analysis of modern printed documents impose weaknesses. Due to the challenges of ancient manuscripts detailed in Section 1.3, binarization pre-processing – as used in traditional document layout analysis – produces errors since it additionally segments 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.

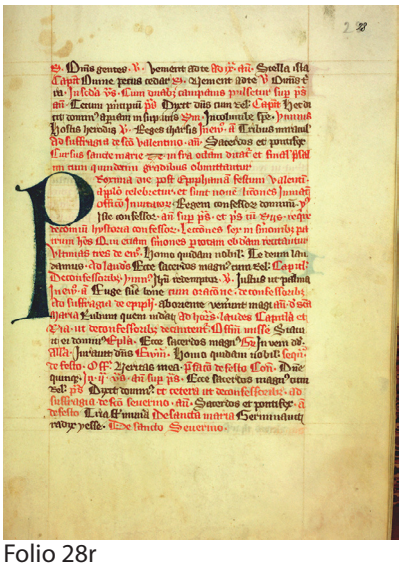
A binarization-free method for layout analysis independent of script and alphabet is proposed that takes into account the specifics of ancient manuscripts. It is a part-based method that detects and localizes layout entities based on their local structure. They are decomposed in parts employing a state-of-the-art object recognition method which identifies objects based on local features, namely 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.

Thus, a method independent from the physical and logical layout of a manuscript is introduced. This means, the method does not rely on a physical layout model, such as constraints of potential locations of layout entities or spatial relationships between them – e.g. in Glagolitic manuscripts headings are accompanied by embellished initials, where the heading is located top and right of the embellished initial. Another example is the location of plain initials – usually they are located within the left margin of the text, however, there are occurrences in the text, where spaces exist before the initial.

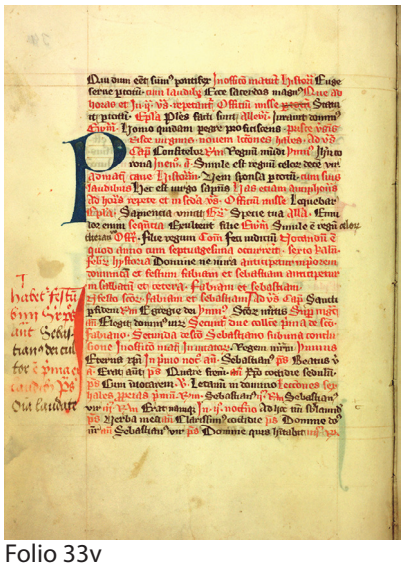
Especially in the case of embellished initials, the shape of the whole character is



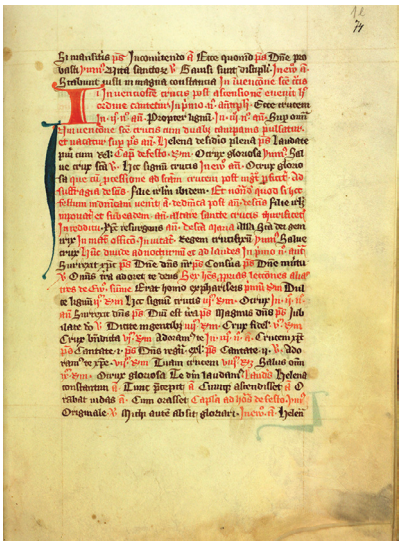
Folio 14r



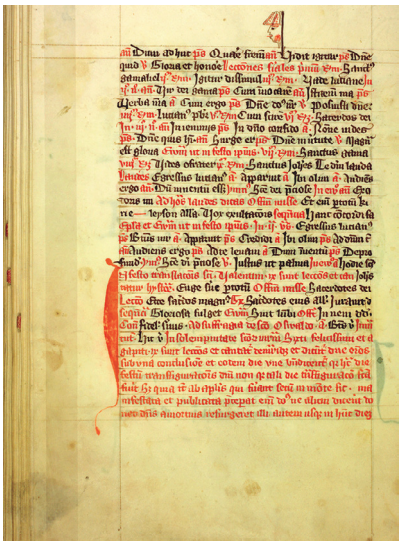
Folio 28r



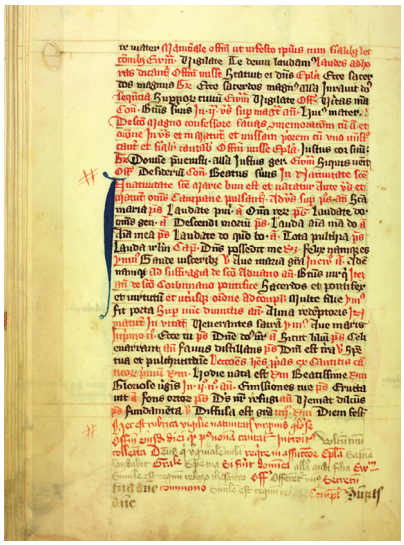
Folio 33v



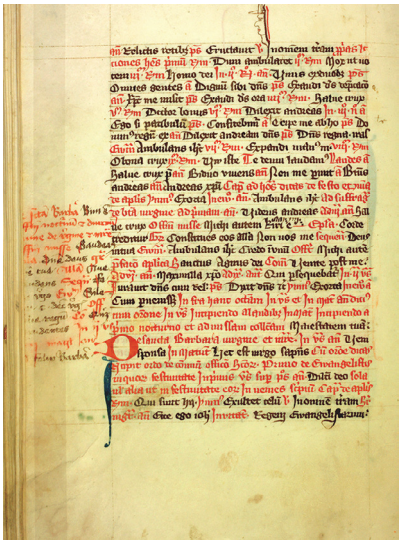
Folio 74r



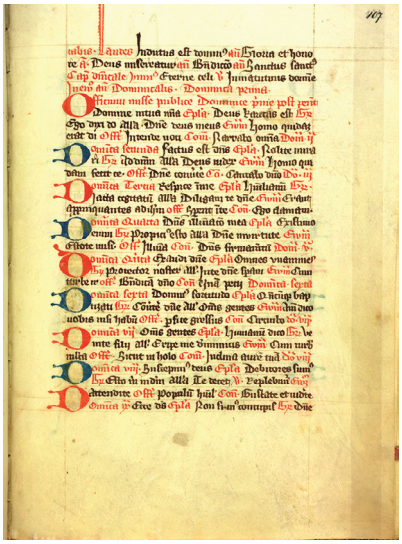
Folio 83v



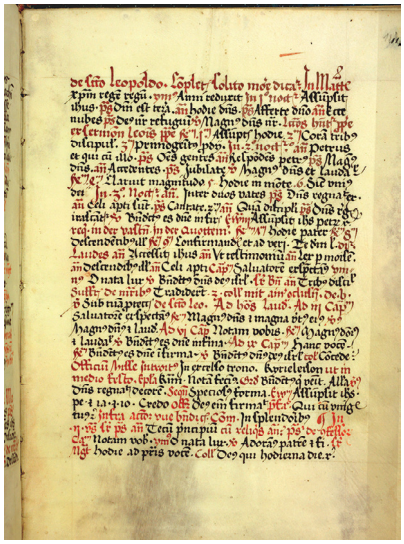
Folio 89v



Folio 100v

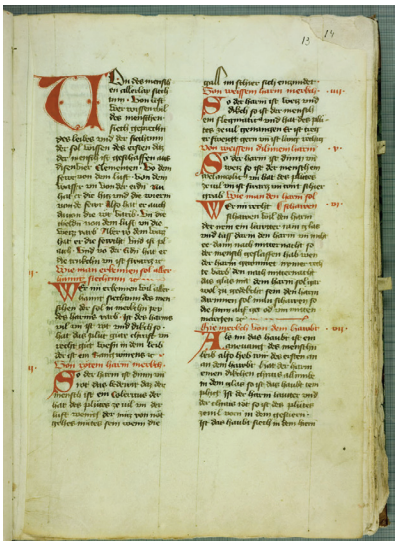


Folio 107r

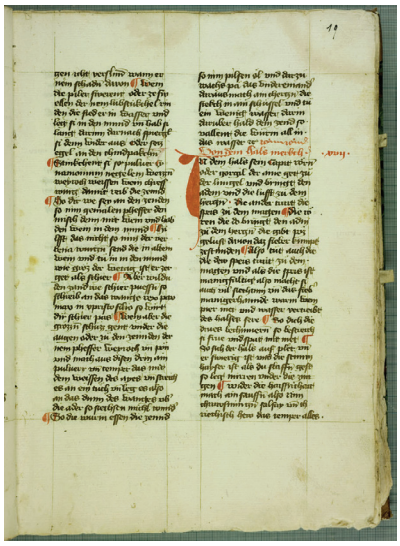


Folio 117r

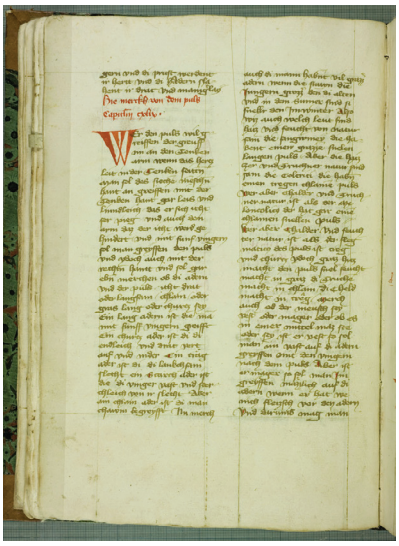
Figure 1.12: Sample pages from *Cod. 635*.



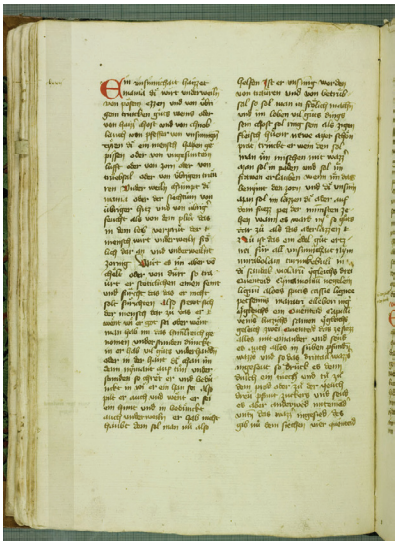
Folio 13r



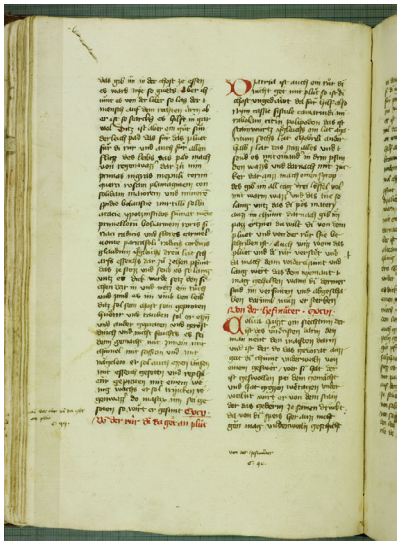
Folio 19r



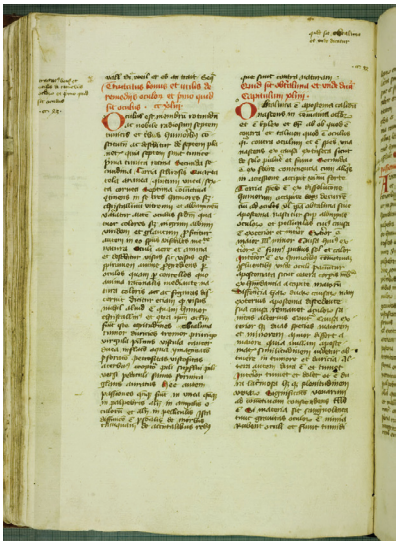
Folio 52v



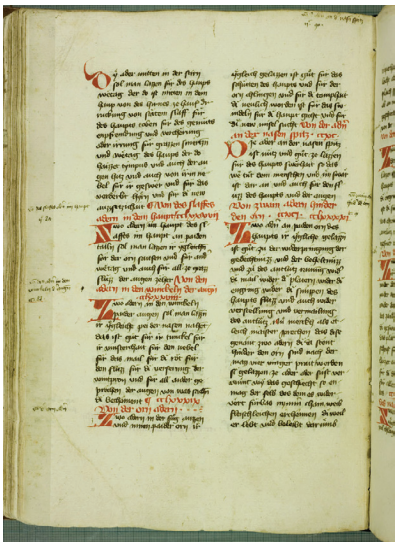
Folio 62v



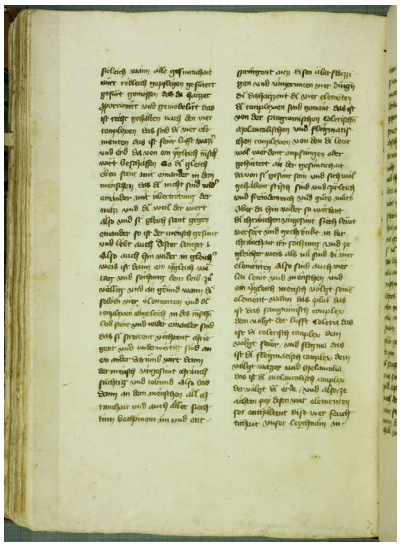
Folio 70v



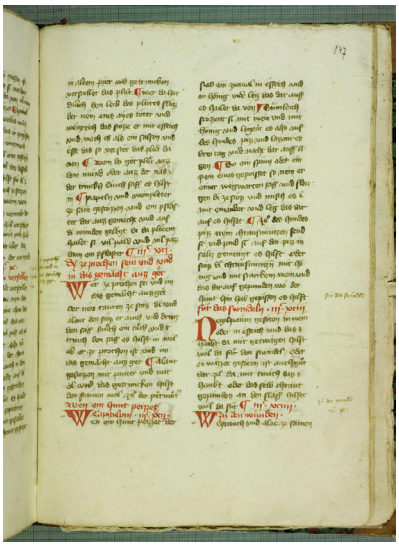
Folio 83v



Folio 93v



Folio 118v



Folio 147r

Figure 1.13: Sample pages from *Cod. 681*.

variable; however, the characteristics of the embellishments are similar. Amongst these are hatches and outlines. The layout entities considered in this report are:

Embellished Initials Decorated letters larger than the main body text, embedded in the margin of the page or at the edge region of the main body text,

Plain Initials Letters having a vertically or horizontally elongated aspect ratio and a more angular shape when compared to a character of the main body text.

Headings Characters similar to plain initials and additionally may be written in another script such as Cyrillic script.

Main Body Text 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. Hence, the writing style or script is used to differentiate between the roles of entities in the document. Further layout entities not regarded in this report are Latin page numberings added later and Glagolitic psalm numberings. The difference between main body text and psalm numberings are horizontal lines above the characters, or circles around the entire numbering.

The method introduced in this report 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 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 artefacts such as stains and clutter originating from the nature of the writing support, are detected as foreground as well. However, these artefacts 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 since 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 Svm is chosen as classifier to discriminate between two classes: regular text of the main body on the one hand and layout entities having a decorative meaning on the other hand. These entities include embellished initials, plain initials and headings; they are grouped into one class as result of their local structure correspondence as explained earlier.

A localization algorithm is then required 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.

1.5 Evaluation and Results

The methodology introduced in this report is empirically evaluated by means of manually annotated real world data. The manuscripts and experiments are selected such that the strengths and weaknesses of the method can be examined. The performance metrics employed to measure the method's accuracy, are precision and recall, and the weighted mean of these two values, the F-score. The test set for each evaluation consists of 100 randomly selected pages of the manuscript. The test set is build up with image patches containing entities of the respective classes.

The evaluation is done on pixel level for the final segmentation of the page into regions of interest, and per interest point for the evaluation of the localization algorithm.

For the interpretation of the results for the pixel-level evaluation, it has to be considered that the ratio between main body text and decorative are employed as measure metrics for the method's accuracy entities is approximately 9:1. 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.

For the *Psalter*, an F-score of 0.914 is reached for the both classes, with the main body text being detected with an F-score of 0.930, and the decorative entities being 0.629. In case of *Cod. 635*, the F-score is 0.972, where the performance of the main body text is as high as 0.978 and the F-score for the decorative entities is 0.735. The F-score for *Cod. 681* is 0.975, with 0.979 for the main body text and 0.677 for decorative entities.

Reasons for the inferior performance of the decorative entities are e.g. the embellished initials are detected and localized well if enough structural detail is present since a high density of interest points is being generated; however, single strokes and outlines produce a low number of interest points. Further reasons are features not discriminative enough from the main body text. Touching and overlapping class boundaries are an additional problem.

Even though the layouts, scripts and general appearances of the manuscripts are different, the result is comparable. Especially the performance of the *Psalter*, being more degraded than the other documents, having stains, creases and traces of decaying, has to be pointed out.

The localization algorithm is specifically important for the class of decorative entities. For strict layouts and text having a uniform appearance, the localization algorithm for rejecting misclassified interest points has no major influence. However, for the class of decorative entities, this step is crucial.

1.6 Outline of the Report

This chapter depicted the purpose and motivation of the report along with the description of the datasets. The remainder of the report is structured as follows:

Chapter 2 – State of the Art provides an overview of existing methods for document layout analysis. Hereby, traditional approaches are summarized and it is outlined why these methods are not applicable for historical documents. Then, methods with regard to historical printed documents and ancient manuscripts are discussed. A special focus is based on methods solving specific tasks in layout analysis, such as the detection and the retrieval of decorative entities, or the differentiation of fonts and writing styles with respect to layout analysis and script classification.

Chapter 3 – Methodology gives background information about existing methods employed in this report and related work. First, a summary of interest point detectors based on different principles, such as corner or blob detection, is provided. The interest point detector chosen is described in detail. Second, state-of-the-art local descriptors are reviewed, with a focus on the adopted descriptor. The third section of the chapter provides a comparison of potential classifiers and explains the chosen method in detail.

Chapter 4 – Proposed Methodology introduces the approach for document layout analysis introduced in this report. The employed existing methods and necessary adaptations to these are explained and related to the entire layout analysis system. A new localization algorithm is introduced which allows the derivation of classification results on pixel level from local features represented by interest points and their descriptors.

Chapter 5 – Evaluation and Results presents the experiments based on three datasets. First, the experiment setup is described and the performance metrics are introduced. Then, the evaluations and analysis for each manuscript are given.

Chapter 6 – Conclusion concludes this report and gives an outlook to potential improvements and extensions of the proposed method.

Chapter 2

State of the Art

This chapter reviews state-of-the-art methods for document layout analysis of historical manuscripts, and shows methods solving specific tasks in layout analysis, such as the detection and the retrieval of embellishments such as ornamental letters, or text line segmentation. A special focus is put on the analysis of ornamental letters and the identification of scripts.

First, an overview of traditional methods on document layout analysis applicable to machine-printed documents is given. The majority of these methods consider documents having a white background, which allows for binarization as first background-foreground segmentation step. Furthermore, they presume a document as having a rectangular layout, which means the layout is constrained and consists of rectangular non-overlapping (text-) blocks [115]. The documents regarded contain machine-printed font where characters do not intersect – contrary to handwriting – and hence, single characters can be segmented [115].

Then, selected methods developed for historical manuscripts and documents from the hand-press period with a special focus on documents containing initials are summarized. Hereby, binarization-based approaches and such not requiring binarization as a pre-processing step, are included in the review.

Text line segmentation can be addressed as a part of document layout analysis. In Section 2.3, methods for this purpose are surveyed with respect to ancient and historical handwritten documents.

The method proposed in this report relies on a part-based detection and discrimination of scripts and embellishments where the difference in writing styles and scripts is exploited to identify the layout elements. Thus, Section 2.4 summarizes approaches concerning the differentiation of scripts and fonts, where the emphasis is not the identification of the hand/scribe but the indexing of documents or the analysis of the layout.

Subsequently, methods processing initials – especially drop caps embedded in art work – are presented. Goals include the indexation of drop cap images and the recognition of the letter it represents.

2.1 Traditional Document Layout Analysis

In literature, methods in document layout analysis are categorized into three major classes: bottom-up, top-down and hybrid methods [18, 94, 115, 132]. In the following, these categories are described according to the literature.

Top-Down (model-driven [115] or knowledge based [132]) approaches use knowledge about the expected document layout and are initiated with the whole document image. The document image is iteratively split into major- and sub-regions specified by the document model. As prior knowledge about the structure of the document is required, the scope of application is either restricted to a limited set of similar documents or needs a complex architecture.

Bottom-up (data-driven [115, 132] or element aggregation [115]) methods imply no expectation about the nature of the document. They start from the image pixels and progressively group these to structures such as words, lines, and zones. Examples are texture-based approaches [69, 85], morphological filtering [87] or Connected Component (CC)-based methods [73, 115, 138].

Hybrid algorithms can be regarded as a combination of top-down and bottom-up approaches, they take advantage of both methods.

Chen and Blostein [28] give a characterization of features used in document analysis. They distinguish image features, structural features and textual features.

Image Features are global (of the whole image) or local (considering a region) descriptors directly extracted from the document image or its binarization. They describe e.g. texture, gradients, color or shape.

Structural Features are topological descriptions derived from the physical or logical layout, describing the relationships of objects in a page.

Textual Features are obtained from OCR or directly from the document image

Tang et al. [131, 132] give a differentiation of document processing into two phases or categories. Figure 2.1 illustrates the acquisition of knowledge of a document image, where these two categories are sequentially applied to a document. First, the document image is analyzed and the geometric or physical structure is extracted, then the logical structure is obtained by mapping the regions found in the geometric structure onto logical labels. The approach introduced in this report is of the first category.

In order to relate the method presented in this report to the categorizations described above, the respective groups applying to this method are given. The layout analysis method introduced is employed to ancient manuscript images resulting in a physical structurization of the page. However, the proposed approach is capable to do a rough labeling of the layout elements – such as decorative elements or textual regions – due to the known classes. Concerning the categorization into bottom-up, top-down or hybrid methods, the approach belongs to the data-driven or bottom-up approaches as it starts from pixel level and incorporates no model about the document except for the classes contained. And lastly, it is associated with Chen's [28] class of image features on the local level.

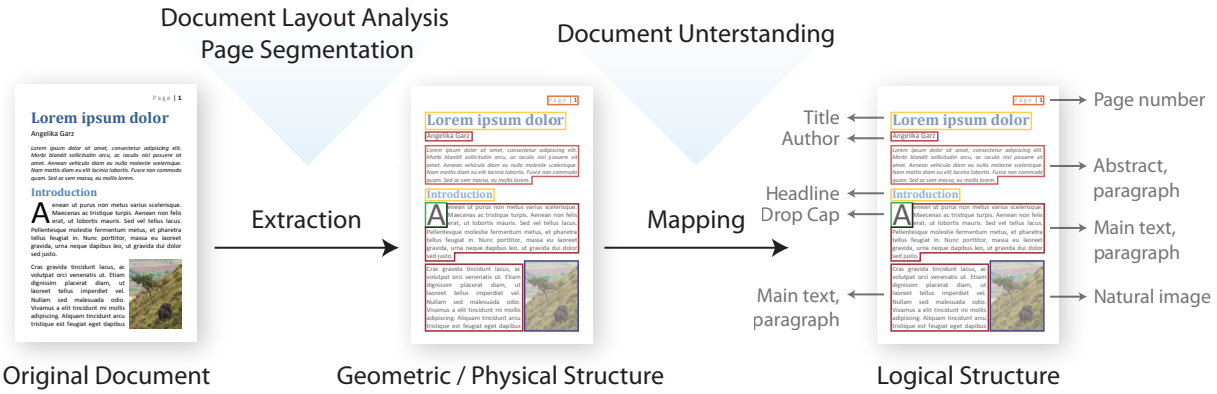


Figure 2.1: Basic model for document image analysis (Figure inspired by [131])

A literature overview of physical and logical structure document layout analysis is done by Haralick [60] on modern machine-printed documents such as business letters or scientific articles. Mao et al. [94] did a review on the same topic. Haralick [60] describes the physical or geometrical layout as the topology of homogeneous regions and their spatial relation ships, whereas the logical page structure consists of labeling these regions, sorting the text blocks conforming to the direction of writing or classifying the type of page. The work done by Nagy [109] gives a detailed survey and evaluation of 99 papers on layout analysis published prior to 2000. Okun et al. [112] survey 110 page segmentation algorithms which lead to a representation of the physical structure of a document primarily from the 1990's. The methods considered are developed for structured articles, such as scientific papers, documents with unconstrained layout such as advertisements, and as a third group, images such as maps, engineering drawings or video frames. Chen and Blostein [28] give a survey on document image classification, where the document classes of the examined works are mainly business letters, scientific articles or forms, except one paper where books from the 19^{th} century are considered.

A literature overview of layout analysis regarding historical documents is done by Ramel et al. in [115], where they further analyze why classical (binarization-based) methods for layout analysis cannot directly be applied to historical printed documents. They specify adaptions of these methods needed to apply them on theses documents. An overview of texture-based document analysis methods is done by Okun and Pietikäinen [113]. They divide approaches for document layout analysis just into two classes - texture-based and non-texture based. The papers reviewed, however, presume with one exception [30] that the background is white, not textured.

2.2 Document Layout Analysis for Historical Manuscripts and Historical Printed Books

The datasets considered in methods surveyed before, however, are printed documents rather than handwritten manuscripts. Though, as Antonacopoulos and Downton [4] state, applying approaches developed for the analysis of modern machine-printed documents on

Table 2.1: Approaches for document layout analysis for historical manuscripts

Method	Bin.[a]	P./H.[b]	Documents Considered	Goals
[83]	B	H	Medieval manuscripts (Arabic, Latin)	Retrieval of initials, illustrations, text regions or titles
[115]	B	P	Renaissance historical printed books from 14^{th} to 17^{th} century	User-driven layout analysis system, classes considered dependent on the use case
[69]	N	P	Renaissance historical printed books from 15^{th} and 16^{th} century	indexation of historical book collections, image retrieval, classes considered are text, non-text regions
[59]	N	H	Renaissance manuscript from the 15^{th} century	Segmentation system, classes considered are text, decorations and images
[23]	N	H	Handwritten index with rigid layout from 1903	Identification of certain parts of a form, text extraction
[8]	N	H	Medieval manuscripts	Annotation tool for the generation of ground truth, classes considered are text regions, decoration, background, page background and page border
[9]	N	P	Book pages from all periods	Indexation of images in large-scale book collections

[a]Denotes whether the approach is binarization-based (B) or does not rely on binarization (N).
[b]Denotes whether the documents are printed (P) or handwritten (H).

historical manuscripts imposes problems, robust methods adapted to the special challenges of these manuscripts are needed. In the following, selected related work concerning ancient manuscripts and documents from the hand-press period is given, with a focus on ornamental letters such as initials or drop caps.

Table 2.1 gives an overview of the methods surveyed in this section. For each method it is given whether it relies on binarization of the document, the documents under consideration are printed or handwritten and the respective kind of documents and their age.

Le Bourgeois and Kaileh [83] propose a document analysis system that retrieves meta data such as initials, illustrations, text regions or titles from ancient manuscript images. Their goal is to assist researchers, historians and librarians who are not experts in image analysis to analyze digitized medieval manuscripts and retrieve meta data from these. Furthermore, the system should process a wide variety of types of manuscripts. For

this, the approach is based on a bottom-up segmentation step, where the images are first binarized and then, binary features linked to shape (such as anisotropy, orientation, contour or moments) and geometry (such as object size, thickness) 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. Classification is done with supervised training using a k-Nearest Neighbor (k-NN) classifier. The approach is tested on 1361 pages of six Arabic and two Latin manuscripts. Due to the lack of ground truth, a qualitative evaluation is given. Figure 2.2 gives results of two Latin manuscripts containing two different text classes and initials.

Ramel et al. [115] present a user-driven layout analysis system for historical printed books. They suggest 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 blocks (e.g. ornamental letters, titles) according to their needs. These scenarios are stored for each kind of document image and are then applied to all of the images of the respective book. The dataset consists of pages from a Renaissance book, dating from the 14^{th} to the 17^{th} century in Latin and French. An evaluation carried out on 1,452 images of five books achieves the following discrimination results: all text blocks are correctly detected, 99 % of the ornamental letters and 90 % of the principal titles are correctly detected.

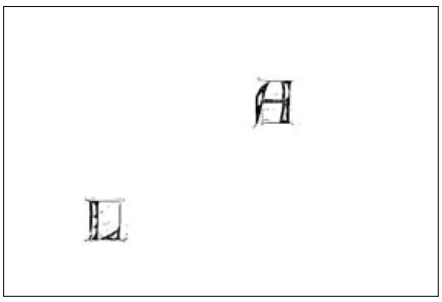
The authors of [67–69] introduce a system for the indexation of historical book collections. They adopt a texture-based method for characterizing historical printed documents from the 15^{th} and 16^{th} century for document image retrieval. Hereby, text, graphical areas – such as drop caps – and background are identified by their distinct distribution of orientations and frequencies. Similar to [113], they assume that text areas are regular periodic textures, since they consist of text lines following the same orientation and the spacing between characters and lines remain approximately the same. First, an unsupervised block classification is applied, then an user-driven indexation and semantic labeling method is applied to these homogeneous areas. The method is similar to the one proposed by Chetverikov [31], who introduced a polar transformed Auto Correlation Function (ACF)-based approach, which leads to the same results as the extended gray level difference histogram presented in [32], used to classify regions in documents. Similar to Chetverikov's method, Journet et al. use a polar transformation of the ACF applied to sliding windows passing over the document image. The ACF correlates the image patch with itself and detects periodicities and orientations in the signal. Figure 2.4 a) shows examples of polar transformations computed on different images. Their system is independent of typography, character sizes and the layout of the document but requires little variation in the text line alignment and the regularity of the text blocks. Amongst other experiments, the authors test their approach on 100 drop caps belonging to five different books, and obtain a correct classification rate of 90 %. In Figure 2.5, results obtained with the proposed approach are depicted. Differently colored pixels on the right side give the class membership.

In their work [59], 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, an improved method using circular statistics as introduced

Original Latin manuscript

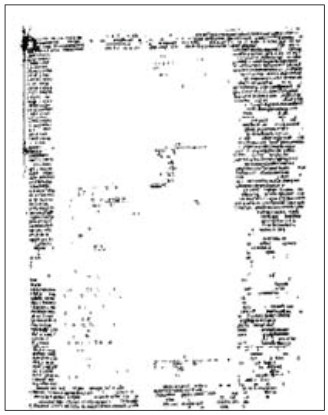
Original text from the main copyist

Text from the second copyist



Colored drop caps

Original manuscript



First writing (bold)

Second writing (fine)

Blue objects

Red objects

Figure 2.2: Results of meta data extraction of a manuscript containing initials (Figure taken from [83])

Figure 2.3: Examples of pages of historical books used as dataset used by [115] (Figure taken from [115])

by Journet et al. [69] is used to separate text, background and images, and second, visual descriptors for color and texture applied to sliding windows extract features for each block to differentiate between decorations and images. The visual descriptors are based on color histograms and a texture feature called Gradient Spatial Dependency Matrix which is inspired by the Gray Level Co-occurence Matrix (Glcm) presented by Haralick [61]. The method is evaluated on 320 pages of a Renaissance illuminated manuscript from the 15^{th} century. For text, their method provides a recall of 91.14 % with a precision of 88.4 %. Depending on the combination of features used, the best results for the differentiation of images and decorations is 57.61 % to 87.31 %. In Figure 2.6, an original page and two segmentation results - one for text an one for located decorations and images are given. Note that the three pages shown are not the same.

In [23], 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 method is developed for the index of archive of the cabinet of the Dutch Queen from the 18^{th} and 20^{th} century, which contains forms of government interventions. The rigid layout of the index is a table with fixed cell sizes which remained consistent throughout the century. The goal of the method is to split the document images into rectangular areas of interest containing the respective document elements. Figure 2.7 shows an example result of the algorithm where a color scheme indicates the different regions detected. For evaluation, the system is applied to images from the year 1903, the performance is given in means of errors; the most relevant results with respect to the system proposed in this report are given. Performance of handwritten text lines: 0.2 % are missed and 0.5 % are over-detected while decision paragraphs are missed in 0.6 % and have a higher over-detection rate of 3.9 %.

The authors of [8] introduce a semi-automatic annotation tool for the generation of

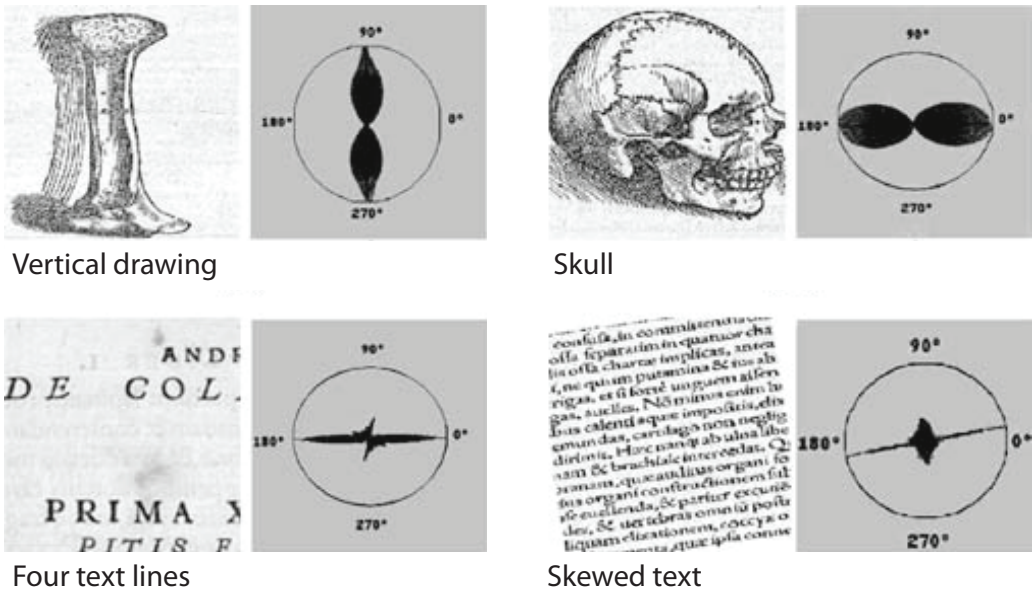


Figure 2.4: Example of the polar transformed ACF computed on different images (Figure taken from [69])

ground truth of layouts for medieval manuscripts. The tool consists of two steps, where the first segments the page into text blocks and text lines and the second is an interactive interface allowing the user to visualize, validate and annotate the results. For the first step, the layout elements a page consists of are determined using Multi-Layer Perceptron (MLP) that exploits color features. For this, the page is divided into square image patches. Each patch is classified to one of the five following classes: Text, decoration, background, page background and page border. Then, identified textual regions are segmented into text blocks and lines employing a CC algorithm. The authors do not difference between plain and highly decorated initials, and between regular text and headlines. The ground truth generated is represented in XML format.

A SIFT-based image and line drawing detection system is proposed by Baluja and Covell in [9]. 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 on 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 for drawing and images and their interest points are stored in a database. An open issue is the localization of the entire image, as the interest points just denote dedicated positions and thus, have to be amplified to cover the whole image or drawing. For evaluation, scanned pages from hundreds of books where taken. The rate of correct classifications is 83.1 % to 90 % depending on the classifier applied.

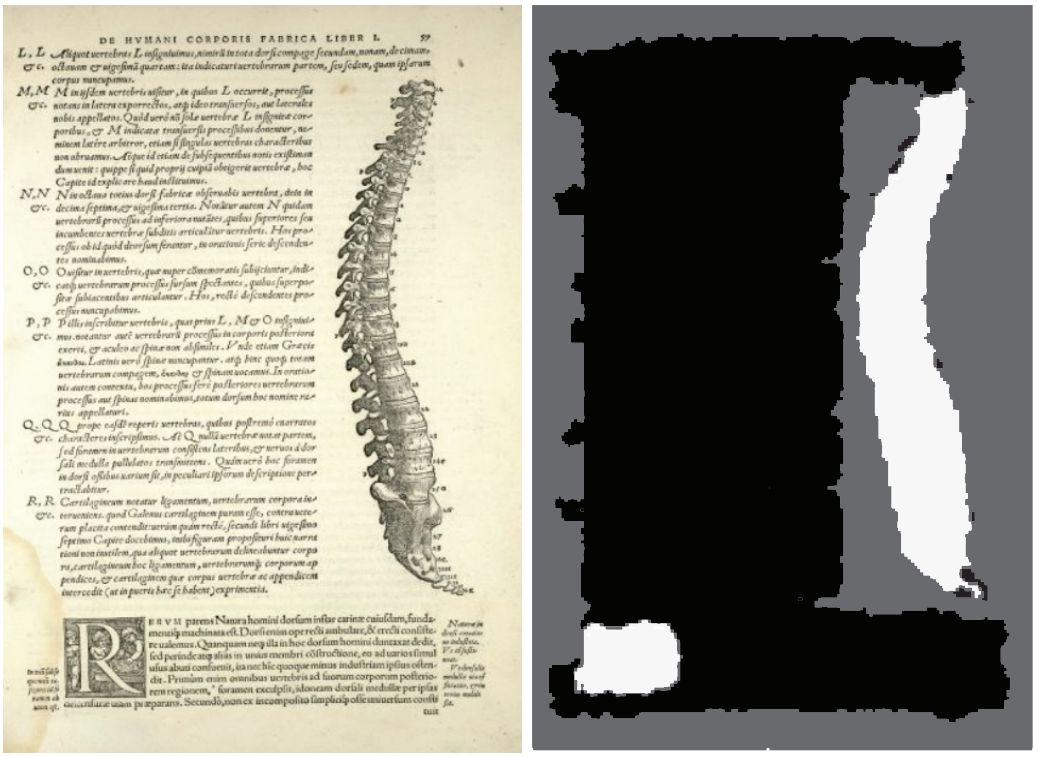


Figure 2.5: Results for the approach (Figure taken from [69])

2.3 Text Line Segmentation

Text line segmentation divides a text block into separate lines of text. A special focus is given on PP as they are commonly used in literature [5, 6, 10, 23, 125, 145].

Likforman et al. give a detailed survey about segmentation of text lines with respect to historical documents in [88]. They divide the algorithms in two classes: one separates two consecutive text lines using a line or path, and the second searches for aligned units, such as prior segmented words. Amongst other methods, they review PP, smearing, Hough-transform based and stochastic methods.

The work of Bulacu et al. [23] described in Section 2.2 is a text line segmentation method based on the number of transitions between ink and writing support. One challenge of cursive handwriting are ascenders and descenders of the characters that may overlap. Hence, applying straight cuts between text lines damages these characters. Thus, they employ a method that follows the metaphor of a raindrop starting in the center of the space between two text lines and follows the contours of the ascenders and descenders back to the intended separation line. If such a detour is impossible due to an ascender and a descender overlapping, it is cut through. The idea is illustrated in Figure 2.8 left. On the right, an example result for text line segmentation is given.

Various authors [6, 145] adapted the global PP as described above such that skewed text blocks, fluctuating, converging or merging text lines are segmented correctly. To accomplish this, they split the document into non-overlapping vertical stripes and employ the PP piecewise, see Figure 2.9 b). Hence, an initial set of candidate lines are obtained

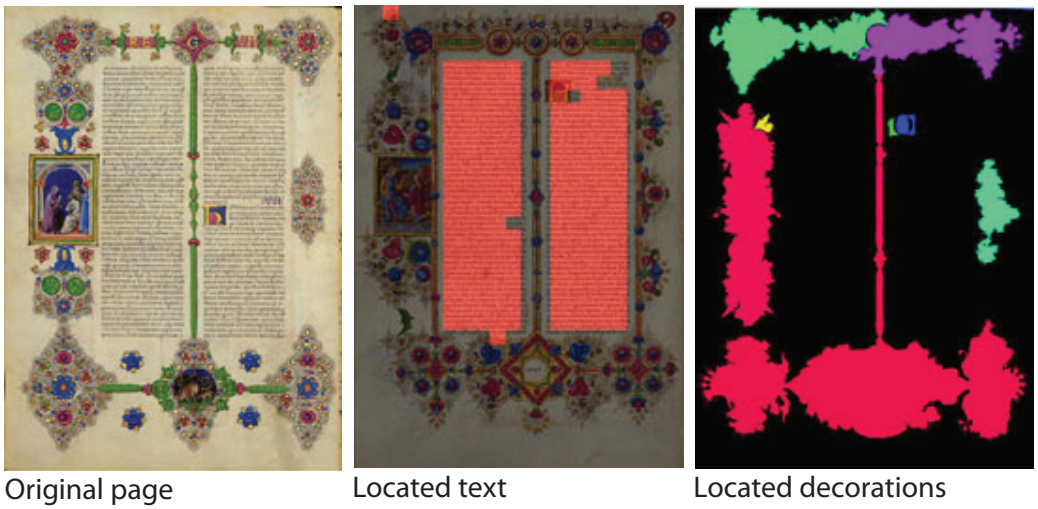


Figure 2.6: Example results (Figure taken from [59])

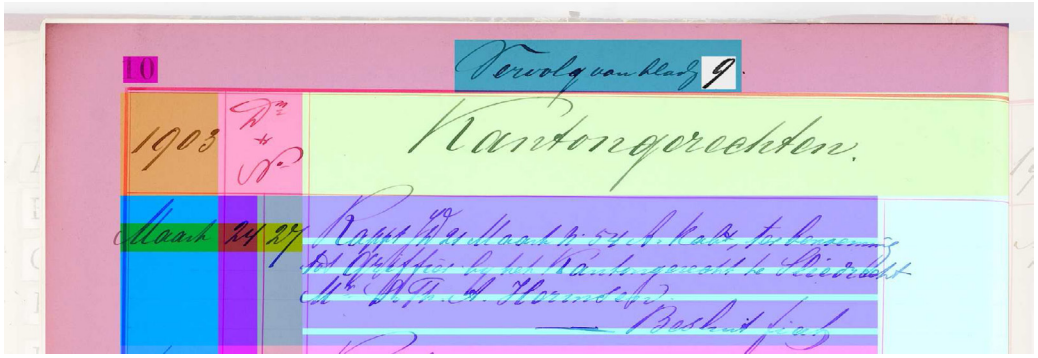


Figure 2.7: Layout analysis result where different colors indicate the regions detected (Figure taken from [23])

connecting the local minima of the PP of two consecutive stripes. Figure 2.9 c) and d) show the connecting of the local minima, where in d) a text line ends. This line set is further processed and segmented establishing a line model based on bi-variate Gaussian densities. On Arivazhagan et.al.'s test set [6], which includes 720 documents in English, Arabic and children's handwriting, 97.31 % of the text lines are segmented correctly.

However, the LPP method leads to a staircase appearance as the locations of the valleys are not directly connected but linked horizontally. Figure 2.9 shows this effect. Bar-Yosef et al. [10] extended this method to deal with any skew angle and to adapt to the course of the text lines. They introduce an Oriented Local Projection Profiles (OLPP), following the hypothesis that the orientation of a text line changes gradually, not abruptly. Hence, they employ the LPP incrementally following the text line orientation estimated locally for each stripe. Two experiments are conducted on 30 degraded ancient Hebrew manuscript pages, where in the first, the algorithm is directly applied and in the second, the images are rotated in different angles to test the robustness of the approach. The evaluation is done by visual inspection. In both experiments, 98 % of the text lines are correctly segmented.

Figure 2.8: Layout analysis result where different colors indicate the regions detected (Figure taken from [23])

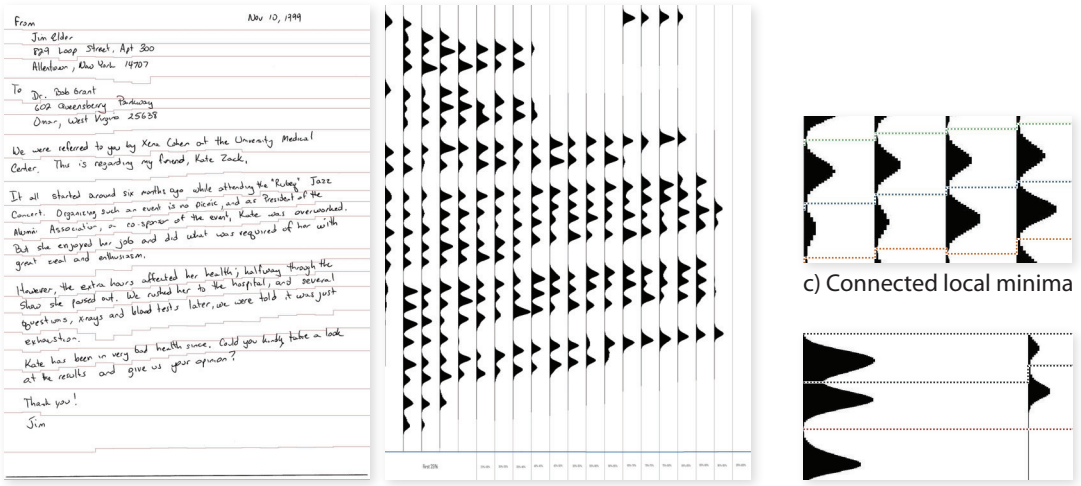


Figure 2.9: Lpp (Figure taken from [6])

2.4 Identification of Scripts for Indexation and Layout Analysis

The approach introduced in this report distinguishes headlines and regular text based on differences in the local structure of the characters. Thus, it is able to discriminate between different fonts or scripts. This section reviews state-of-the-art methods to differentiate fonts and writing styles focusing on layout analysis and script classification rather then scribe identification.

Abirami and Manjula [2] give a detailed review of script identification methods in multi-script document images. Approaches aiming at script identification can be divided into global, texture-based; local, and statistics-based classification methods. Local methods operate on the text line, word or character level whereas global approaches analyze regions such as text blocks or multiple text lines. They are further divided into block level script identification, where one whole page is considered to consist of a single script, and into word level script identification, where one page may contain multiple scripts and the classification is performed per word, and line-based script identification for local methods only. Global methods include Glcm, Gabor Filters or Wavelet Transforms. Local methods employ upward concavities, token-based approaches, topological and stroke-based

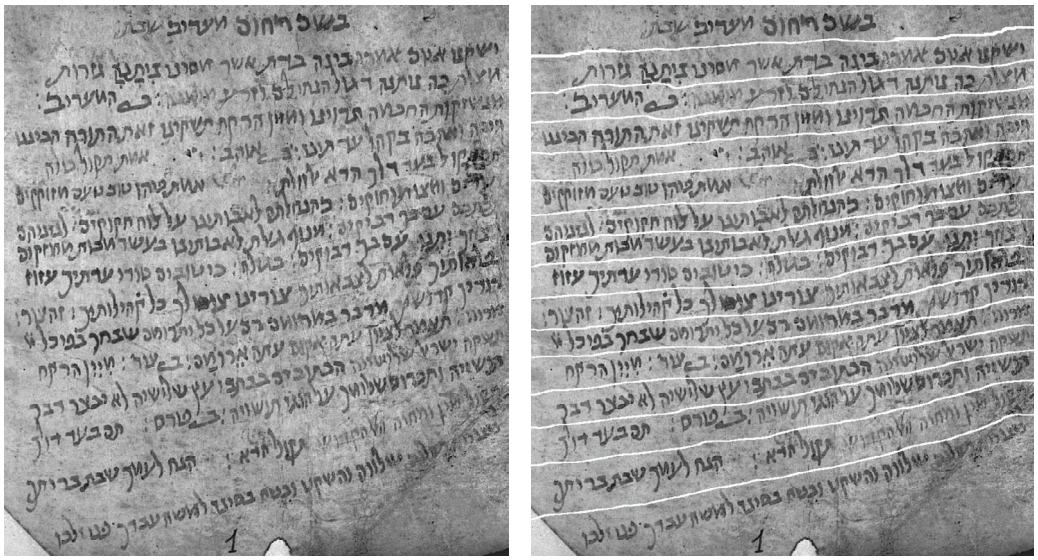


Figure 2.10: Left: ancient manuscript, right: results of the OLPP segmentation (Figure taken from [10])

features or morphological reconstruction amongst others. A further discrimination can be done into binarization-based and binarization-free approaches, just as in layout analysis. When depending on binarization, the performance of the script identification is linked to the effectiveness of the segmentation algorithm. The first two methods described depend on a binarization step while the third is applied to gray-level images. Table 2.2 gives an overview of the methods reviewed in this section.

Rashid et al. describe a binarization-based system to distinguish multiple scripts in ancient document images [116]. The dataset used are 19 manuscripts containing Greek and Latin scripts, where 12 are used for training and evaluation and seven as test set. They employ convolutional Neural Networks (NN) as discriminative learning model which works directly on CC. This implies the advantage that no prior layout analysis is required. The training of the NN is performed for each CC using a back propagation algorithm. They achieve an accuracy of 97.58 % for Latin script and 96.4 % for Greek script on CC level for their test set. In Figure 2.11, results for a Greek-Latin composite script are given, where the different scripts are encoded in colors.

In their paper [72], Khelifi et al. present an unsupervised extraction and categorization method for text regions using fractal descriptors, where similar text regions are identified by their fonts. The purpose are indexation and categorization of documents. The approach is applied to maps and to ancient manuscripts. Similar to [69, 83], their system depends on user interaction; here, Content Based Image Retrieval (CBIR) is implemented. After the text extraction – where the text is separated from background by means of color segmentation and connected components – fractal features are computed. Analogous to the approach proposed in this report, the authors presume that the font is a key characteristic differentiating between regular text, headlines, and decorative elements in ancient manuscripts and maps. The manuscripts considered are the same as in [105]. The authors give no quantitative evaluation of their approach. Figure 2.12 shows results obtained by the approach of Khelifi et.al., where the scripts are encoded with differently colored

Table 2.2: Approaches for the identification of writing styles

Method	Bin.[a]	P./H.[b]	Block/Word[c]	Documents Considered
[116]	B	P	W	Ancient manuscripts consisting of Greek and Latin scripts
[72]	B	H	B	Latin medieval manuscripts of 15 different scripts, 13^{th} – 16^{th} century, and seven Arabic styles
[105]	N	H	B	Latin medieval manuscripts of 15 different scripts, 13^{th} – 16^{th} century, and seven Arabic styles

[a]Defines whether the approach is binarization-based (b) or does not rely on binarization (n).
[b]Denotes whether the documents are printed or handwritten.
[c]Denotes whether the the analysis is done on block level or word level.

bounding boxes.

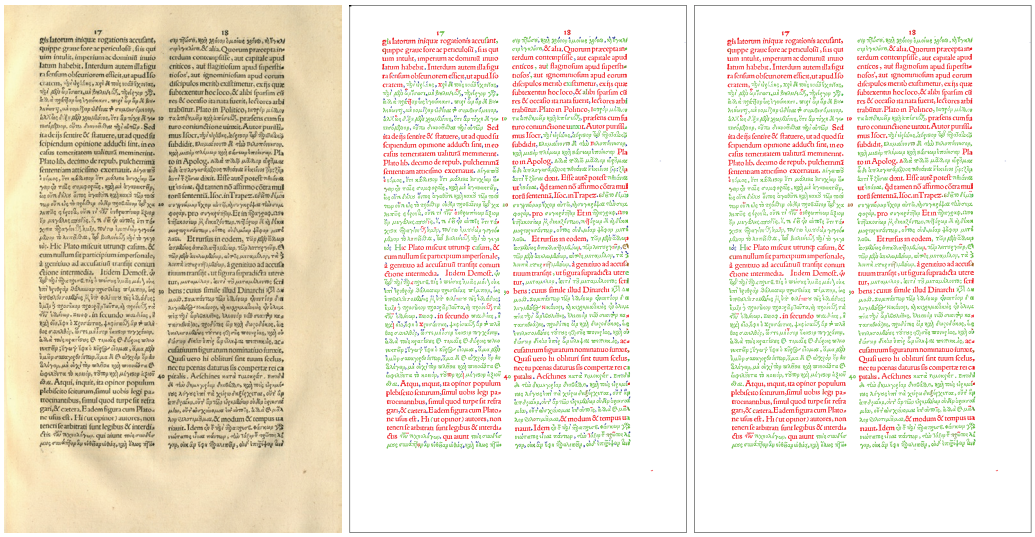
Moalla et al. [104, 105] introduce an approach for the automatic differentiation of medieval manuscripts texts based on the writing styles. The goal is to assist historians in establishing the spatio-temporal origin and the classification of manuscripts. However, the aim is not the recognition of writing styles but a query by example Cbir system. They adopt Spatial Gray-Level Dependence, which gives the co-occurrence to calculate the writing variations, and extract Haralick features to reduce the number of features. The method is applied to 15 Latin medieval text styles from the 13^{th} until the 16^{th} century and seven Arabic styles. Their evaluation show a positive identification rate of 59 % to 81 % for Medieval Latin and reaches up to 100 % for Arabic scripts.

2.5 Ornamental letters / Initials / Drop Caps

Related work concerning ornamental letters such as initials and drop caps, is mainly developed for graphical drop caps. Drop caps are initial letters at the beginning of a paragraph or chapter, which are larger than the regular text and are decorative embellishments of the document [79, 114]; they are usually embedded in the text or in the margin of the page. In contrast to the initials addressed in this report – which are characterized in Section 1.3 – in the graphical drop caps the following works regard, the respective letter is embedded in artwork. Figure 2.13 gives a comparison of the drop caps considered in the majority of the papers reviewed and a typical initial of the *Old Church Slavonic Glagolitic Psalterium Demetrii Sinaitici* regarded in this report.

The work of Bourgeois and Kaileh [83] surveyed in Section 2.2 proposes a document analysis system where initial letters are regarded. These initials are no drop caps like just described, they are not nested in artwork but the letter itself is decorated.

The following papers do not concern the extraction of the ornamental letters from the document page but the further processing. The first paper suggests an automatic retrieval



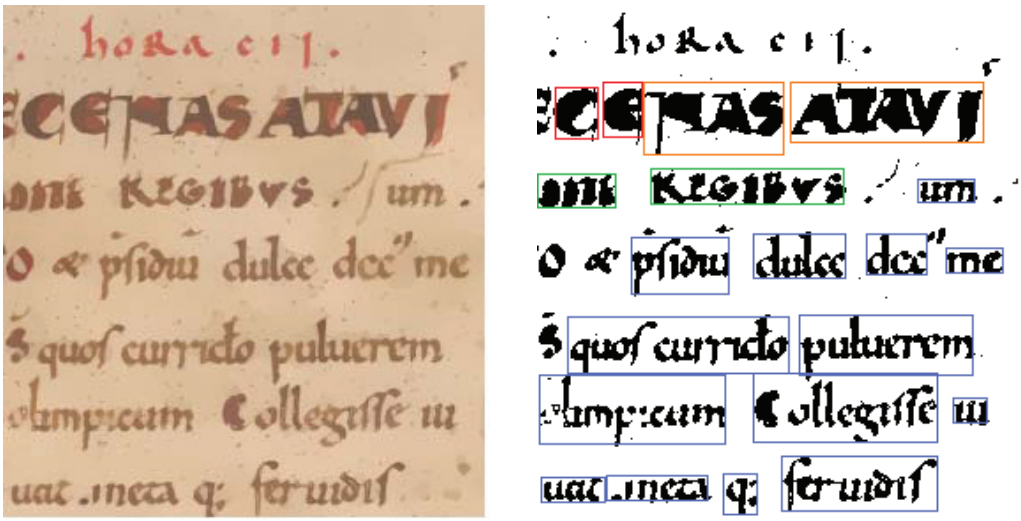
Original image CNN output Final output

Figure 2.11: Results for a Greek-Latin multi-script, where red represents Latin and green represents Greek script. In the final output, single characters incorrectly labeled in another class than their surrounding are labeled with the correct class. (Figure taken from [116])

system whereas the further papers propose a method to determine the letter nested in the drop cap. The approaches are applied to ancient documents of the hand-press period.

Pareti et al. [114] present an automatic retrieval system for graphical drop caps which was developed in the course of the MADONNE project [111]. The authors investigate methodologies of document image analysis in order to create an approach for the classification and indexing of drop caps, examples for different styles are given in Figure 2.14. They combine different signatures based on e.g. Zipf Law and topological features using interest points based on Harris detectors. For the locations of interest points, a Minimum Spanning Tree (MST) is calculated to represent the topology of the drop cap which is combined with a Pairwise Geometric Attribute (PGA) that gives the relationship of each pair of CC employing distance and orientation. They present a method to automatically extract a subset of features which best suit the application. The approach is applied to a database consisting of 200 images of drop caps from the Renaissance using a k-NN classifier. The evaluation is carried out for the Zipf Law and the MST signature. Depending on different styles of the drop caps, a recognition rate of 97 % to 100 % is achieved when a neighborhood of five is considered. The evaluation of the MST and PGA leads to an accuracy of 94 % when the signatures are optimal combined.

In [34], the authors attempt a similar approach, where they first decompose the drop cap image into several layers applying the method developed by Dubois et al. [40], then segment the layers using a Zipf Law and select CCs based on the size, the position of the center of mass and the distance to the edges of the drop cap. The letter is most likely located in the center of the drop cap and one of the largest CCs. Figure 2.15 illustrates three example drop caps in the first row, the results of the Zipf Law decomposition in the second, and the final extracted letter in the last row. The experiments are carried out



Original image Segmented image

Figure 2.12: Results for a manuscript, where the different colors encode three different script classes (Figure taken from [72])

Figure 2.13: Left: Renaissance graphical drop cap such as regarded in [34, 79, 114, 135] (Figure taken from [137], page 3), Right: Glagolitic decorative initial from *Old Church Slavonic Glagolitic Psalterium Demetrii Sinaitici*, Folio 18r

on a database of 4,500 images, where 1,500 are used as training set and the remaining as test set. To evaluate the letter extraction, two commercial Ocr systems are applied, where the recognition rate of FineReader is as much as 72.8 % and Tesseract achieves a recognition rate of 67.9 %.

The work of Coustaty et al. [34] just reviewed is inspired by the paper of Uttama et al. [135], who segment drop caps with a top-down approach which is applied in a global and a local analysis. The first aims at characterizing the areas into texture and uniform zones, where it is presumed that the letter is mostly uniform, whereas the background is textured. The local analysis is then used to group the texture zones based on their similarity or entropy. The aim of the method is to index the drop cap images. The authors give experimental results, but no quantitative evaluation of their approach.

In [79], Landré et al. focus on the segmentation and recognition of previously extracted ornamental letter images from ancient books of the hand-press period. These ornamental

Table 2.3: Approaches for the identification of writing styles

Method	Decorations Considered	Goal
[114]	Drop caps from Renaissance	CBIR for drop caps
[34]	Drop caps from Renaissance	Recognition of the letter represented by the drop cap
[135]	Drop caps from Renaissance	Indexing drop cap images
[79]	Drop caps from Renaissance	Recognition of the letter represented by the drop cap
[20]	Ornaments, 16^{th} – 18^{th} century	CBIR for fleuron ornament images
[29]	Heraldic, 16^{th} – 18^{th} century	CBIR for emblem images
[13]	Drop caps, 16^{th} – 18^{th} century	CBIR for drop caps
[36]	Drop caps, 16^{th} – 18^{th} century	CBIR for drop caps

letters consist of a character and background decorations. The goal is to determine the character, the ornamental letter image represents. First, the image is segmented to filter decorations using a multi-resolution analysis. Then, the image is binarized and reconstructed using morphological operations to extract the letter. In order to determine the actual letter represented by the drop cap, the extracted letter shape is then compared with a dictionary of capital letters using a distance defined by the Local Dissimilarity Map (LDM). The evaluation is accomplished on two datasets, where the first contains 60 images of a subset of six capital letters and achieves a recognition rate of 89.4 %, whereas the second dataset consisting of 823 images of all capital letters considered leads to an accuracy of 62.7 %. Figure 2.16 gives good and poor segmentations of drop caps by the method proposed by Landré.

Baudrier et al. [12] give an overview of CBIR systems for ornaments, such as drop caps and trademarks from the hand-press period, see Figure 2.17. They name five databases consisting of ornaments from the 16^{th} to the 18^{th} century and compare four algorithms for the retrieval of ornamental images, namely a retrieval system employed for fleuron ornament images using orientation signatures [20]. The second method [29] is a retrieval system for emblem images which determines the similarity of the images based on Harris interest points and local features consisting of Zernike moments. The authors of [13] abstain from feature extraction but directly compare images on the pixel level using a multi-resolution approach. A local measure is applied which produces a local-dissimilarity map. The last method [36] compared, relies on a compression technique, the Region Of Interest (ROI) and is employed to binarized images. A distance based on similarity of the ROI representation is used to compare two images of ornamental letters. A visual comparison of the results gained by the reviewed methods is given in the paper.

Whereas the method of [36] relies on binary images, Campana and Keogh [25] propose a distance measure for textures based on state-of-the-art video compression, where the distance is computed analyzing the compression ratios of two images. They apply their

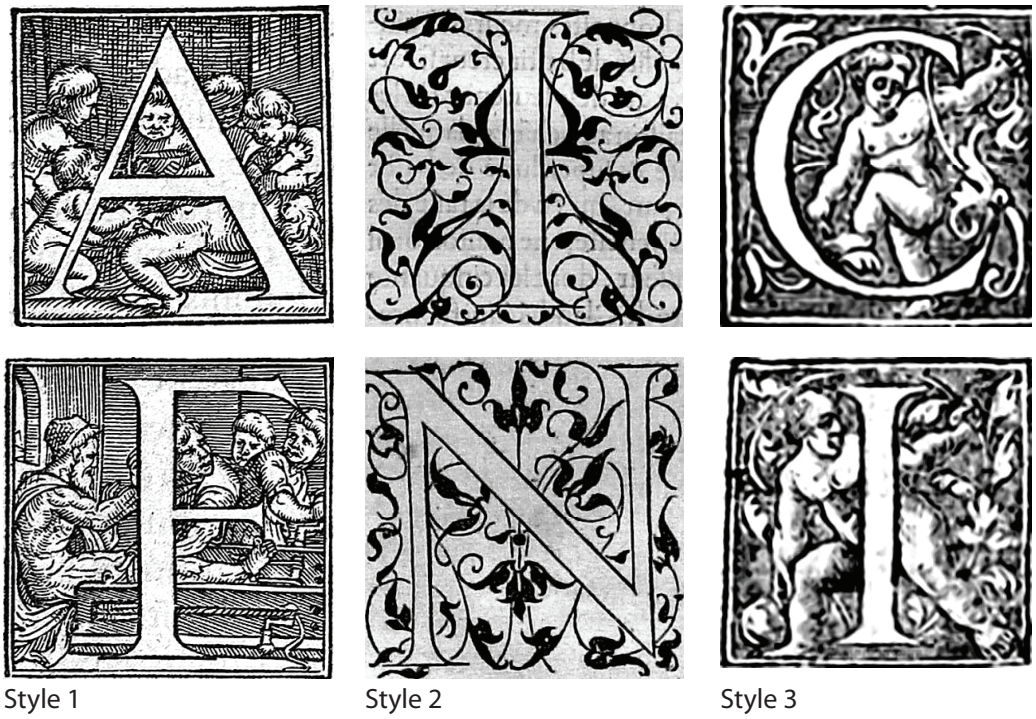


Figure 2.14: Style samples for drop caps (Figure taken from [114])

approach on various datasets available, such as a heraldic shields from the 14^{th} to 16^{th} century, the Brodatz Textures, the VisTex or the UIUCText. Depending on the dataset considered, they achieve an accuracy of up to 96.3 % when a one-nearest-neighbor classifier is used. The methods outperforms Gabor Filters and Textons on nine of 15 datasets.

2.6 Conclusion

In contrast to classical document analysis methods surveyed and altered by [115] for the purpose of applying it to historical printed books, these methods cannot be applied to the manuscripts considered in this report. First, the documents considered in this study were printed books not having the difficulties of handwritten documents enumerated in Section 1.3. Second, the documents are degraded, hence the ink is partially faded out and the background imposes difficulties to binarization approaches. Due to these reasons, approaches developed for historical printed documents which require a rectangular layout or rely on binarization are suitable for ancient handwritten documents.

As Ogier and Tombre state [111], analyzing ancient handwritten manuscripts *"leads to specific questions which are far away from usual handwriting recognition analysis as addressed in postal or banking applications, for instance"* [111]. In the processing of recent handwritten documents, the goal is the recognition of the handwritten characters, whereas for manuscripts which form part of the cultural heritage, the primary aim is to identify scripts or characterize and identify scribes e.g. in order to relate the document to a spatio-temporal origin.

Ramel et al. [115] emphasize why texture-based methods, such as Gabor filters or

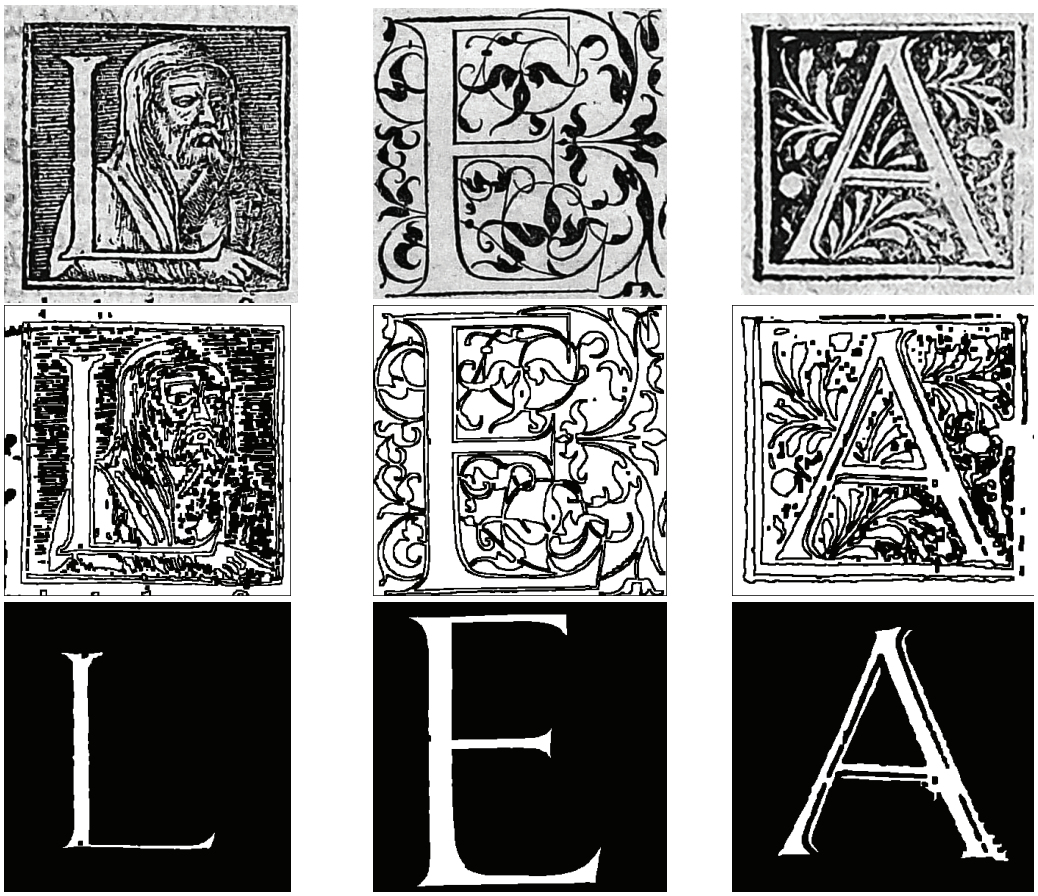


Figure 2.15: Drop caps and their extracted letters [34]

wavelets are not perfectly suited for the analysis of ancient manuscripts. They state, that due to their proximity, distinct regions – such as text areas and initials – might be incorrectly merged. Furthermore, obtaining precise contours of the layout elements, is a non-trivial task with texture-based approaches. [113] and [115] agree about the time-complexity of approaches analyzing the texture of a document.

Methods adapted to constrained handwritten layouts such as the PP proposed by [23], would fail as method to analyse the layout in the presence of background clutter, additional layout elements and the challenges of unstructured handwritten layouts such as super-imposing of layout elements, vertical proximity of text lines and fluctuating text lines as existent in the *Psalter*.

Compared to the methods identifying scripts presented in Section 2.4, the variance in the writing style of the Glagolitic manuscript considered in this report is higher. Since the aim of this report is not the identification of different writing styles or hands, the method has to be tolerant to a higher variation of character shapes.

Some methods, such as the one introduced by Bourgeois et al. [83], exploit color features to segment a document. As stated in Section 1.3, headlines and initials are frequently highlighted with a yellow wash in the Glagolitic manuscript, however, the difference in the color value is not significant enough to detect these decorative elements based on color segmentation.

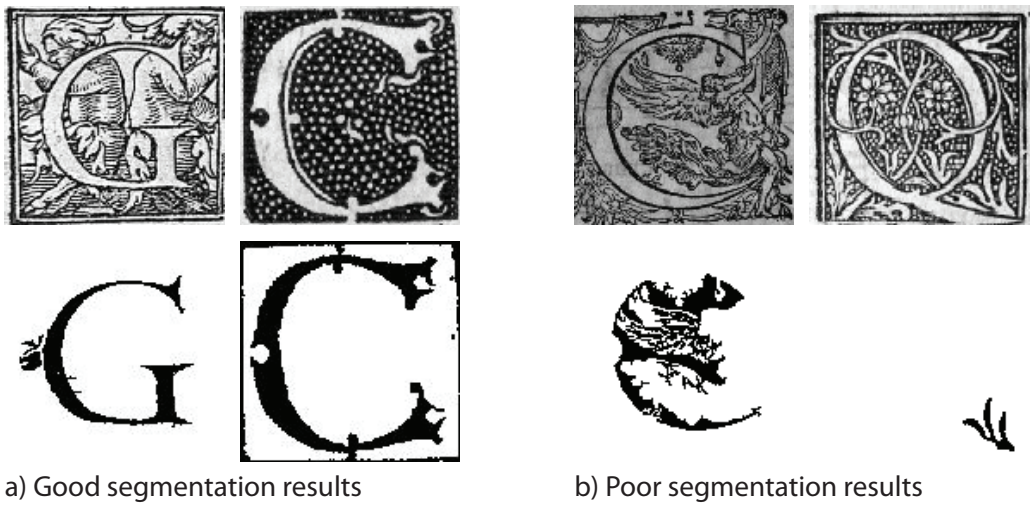


Figure 2.16: Drop caps and their extracted letters, where a) gives good, and b) poor segmentations of the contained letters (Figure taken from [79])

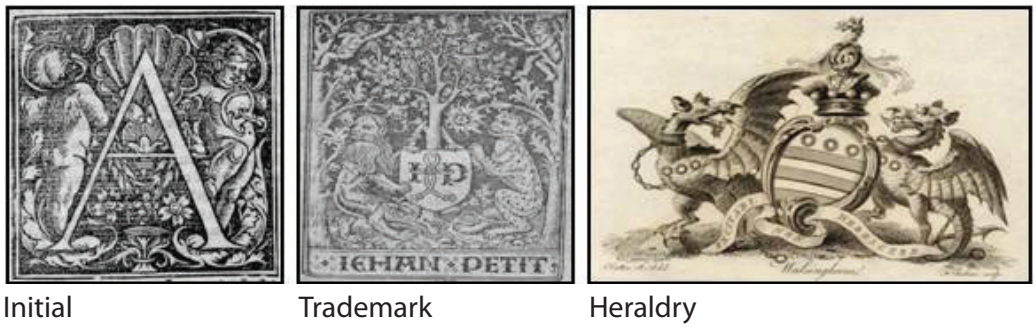


Figure 2.17: Examples for decorative elements in ancient manuscripts considered in [12] (Figure taken from [12])

Thus, a method that is robust to before mentioned challenges of ancient manuscripts and additionally takes into account the characteristics of the layout elements is proposed in this report. Considering these layout elements as objects having intra-class similarities at the local level, an approach drawing its inspiration from the field of recent object recognition methods is chosen to analyze the layout.

2.7 Summary

This chapter gave a review of state-of-the-art literature related to the topic document layout analysis. First, different possible categorizations of layout analysis approaches is given an the method introduced in this report is classified to the respective categories. An overview is given of traditional methods relying on binarization prior to layout analysis and are prevailingly designed for rectangular layouts and machine-printed text. Then, approaches being targeted on the analysis of historical manuscripts and historical documents from the hand-press period is given. Methods employing binarization and methods

working on gray-scale images are addressed. The approach proposed in this report is able to discriminate between different fonts or scripts as it distinguishes headlines and regular text based on the differences in the local structure of the characters. Thus, state-of-the-art methods to differentiate fonts and writing styles are summarized afterwards. Then, a section focusing on ornamental letters and initials describing methods aiming at the processing of these embellishments in the documents is given. The conclusion gives reasons for the choice of the methodology for this report.

Chapter 3

Methodology

The layout analysis approach proposed in this report avoids binarization since features are directly extracted from gray-level images. An approach stable in the presence of local as well as global perturbations such as rotation, illumination variation or artefacts such as noise is needed in order to be robust with respect to the challenges of ancient manuscripts described in Section 1.3. Furthermore, robustness to distortions such as variations in skew and shape is required due to the variance in the shape of handwritten characters and non-even writing support.

In their work, Baluja and Covell [9] search for images in documents which are embedded in text. Analog to their challenge, in the manuscripts taken into consideration in this report, the decorative entities are partially embedded in the text. There exist plain initials surrounded by text, embellished initials reaching into the text area, and headings are part of the text block. Hence, as for [9], global image metrics cannot be applied as they do not incorporate the required granularity for the detection of the decorative entities.

Due to the structural similarities of the layout entities on the local level described in Section 1.3.1, an approach exploiting these characteristics is suggested. These local similarities include outlines and hatches – which locally appear as outlines as well –, strokes that are elongated and have angular shapes. The inspiration is drawn from the field of object recognition methods. Local features are extracted to describe the structural characteristics of immediate regions, leading to a part-based detection of layout entities. Schmid and Mohr [120] first introduced the idea of representing an object by group of local invariant features to detect an object despite of partial occlusion.

In this report, the concept of local invariant features is used as robust image representation to detect and localize objects having different global appearances but certain local similarities. However, these similarities are no exact replications and therefore, an exact individual matching is impossible and not the purpose. Rather, a detection on the category-level [134] is deployed, based on an analysis of the statistics of the features. A machine learning algorithm is employed in order to categorize features according to their similarity.

The requirements regarding the robustness of features for the application in document images for the detection of layout elements are given below, they are similar to those defined by [105]:

Robustness: The features have to be calculated without the need of binarization and

be robust to artefacts such as noise and degradation of the manuscript.

Invariance to the writing style: The features must allow a certain variance in the shape of characters and hence in the personal writing style of scribes, however, have to be able to discriminate different scripts and writing styles independent of the actual scribe (e.g. round and square Glagolica, see Section 1.2).

Invariance to geometric transform: The features have to be invariant to size, skew and rotation of the characters and text lines on the page as they may vary depending on the writer.

Independence of the content: The features have to be independent from the language and the content of the text, since the approach is to be tested on different manuscripts.

Independence of the geometric document layout: Due to the variance in the layout of handwritten documents, features need to be independent of the geometric layout.

Granularity: On account of the proximity or superimposition of layout entities, local features independent of global metrics are needed in order to allow their detection.

This chapter describes the methods employed in the proposed approach in detail and further denotes alternative methods. In order to extract local features, an interest point detector is needed to determine the regions where the features are to be calculated. The first section of this chapter describes the interest point detectors with an emphasis on DoG. The subsequent section denotes feature descriptors with a focus on Sift. Finally, Section 3.3 gives an overview of the classification algorithm employed.

3.1 Interest Point Detector

This section gives an overview of interest point detectors, where first the term and requirements to a stable interest point are defined and then related literature is summarized.

Tuytelaars and Mikolajczyk give an in-depth summary of local invariant features detectors in their work [134]. They provide a general definition of local features, where they state that these points are structures in the image being dissimilar to their adjacent neighborhood. These local features are located close to changes in image properties such as intensity or color. In the following, the term interest point is used to refer to a local feature, where an interest point has a defined location in the image and a definite spatial extend, which is denoted as scale.

Similar to [120, 121], Tuytelaars and Mikolajczyk [134] characterize interest points as points in an image with a two-dimensional signal change. An interest point is found at locations of corners, junctions, circles, endings, dots or texture in general. Since local descriptors are computed at locations of interest points, the detection of these points is a crucial task. The performance of the descriptors is directly related to the detection of stable interest points.

In [134], the authors formulate a set of properties of stable and reliable interest points, where some of them were defined similarly by [121] for the task of interest points evalu-

ation. These properties match the requirements for features used in document analysis given in the previous chapter.

Repeatability An interest point has to be geometrically stable to local and global perturbations of the image such as a change in image scale, rotation or view point. Repeatability then means that the same interest point detected in one image is accurately detected in a geometrically transformed image [121, 134].
Schmid et al. [121] and [134] agree that repeatability is the most important property of an interest point. Tuytelaars and Mikolajczyk [134] provide two possibilities to achieve repeatability: either invariance or robustness.

- **Invariance** relates to mathematically modeling potential deformations and design detectors invariant to these. This is especially relevant for large transformations such as rotation or scale changes.
- **Robustness** means increasing the robustness to small perturbations such as noise or artefacts, e.g. through smoothing the image prior to the interest point detection.

Information content provides a measure of the distinctiveness of interest points, where the information content is characterized by the intensity patterns in the scale of the interest point, i.e. this region should show a large variation in image structure [121, 134].

Locality The spatial extend of an interest point has to be local in order to reduce the probability of corruption through occlusion and to allow for simple approximations of geometric and photometric transformations [134].

Quantity A suitable number of interest points should describe even small objects, where the density of interest points should correlate with the information content. Hence, in homogeneous regions, no interest points should be detected whereas in structured regions having high variance in intensity, the density should be high [134].

Localization accuracy measures how accurate an interest point is located at a specific 2D position and scale in the image [121, 134].

However, these properties incorporate several conflicts, where improving one property adversely affects another. *Repeatability* and *localization accuracy* are contradictory [121, 134]. Achieving robustness through smoothing prevents abrupt alterations of the descriptor with a small change in the position of the interest point [93]. Yet, smoothing impairs the localization accuracy [121]. Localization accuracy is crucial for registration and calibration tasks, whereas repeatability is more important for tasks such as object recognition, or image matching. The layout analysis system proposed as well relies on the similarities of local structures and thus, on the repeatability.

Repeatability achieved either through invariance or robustness derogates *distinctiveness (information content)*. Similar to the case of localization accuracy, image measurements discarded to achieve robustness, reduce the distinctiveness. Image measurements used to raise the degree of freedom of the mathematical model built to acquire invariance are then lost for the information content.

Similarly, *distinctiveness* and *locality* are conflicting properties since the more local an interest point is (the smaller its spatial extend is), the less structural information is contained. Hence, the less information is contained, the harder the matching or classification becomes.

Having defined the requirements to interest point detectors, in the following, contour-based detectors, corner detectors, and blob detectors are described. Contour-based detectors rely on studying the curvature of previously detected contours to find interest points, whereas other methods are directly applied to gray scale images and e.g. analyze the image intensities based on derivatives or their respective approximations or regions having high variance in intensity. The interest point detectors presented are a small selection of those proposed in literature, for a detailed overview please refer to [101, 102, 121, 134].

3.1.1 Detectors Based On Contour Curvature

Contour-based methods employ image contours or region boundaries, which first need to be detected and subsequently maximal curvature or inflexion points are extracted as interest points [93, 121]. This approach aims at making interest points more robust to perturbations such as background clutter and noise close to boundaries [93].

Shilat et al. [126] introduce an approach where first ridges and troughs are found in the image, then high curvature points and intersections of these curves as well as local minima in the image are detected as interest points. They claim that their interest points are appropriate for tracking since they are robust to occlusion since the interest points are not located at contours of an object which are more likely to be occluded.

Horoud et al. [64] employ a graph-based algorithm for grouping of line segments extracted from image contours, where they rely on geometric as well as relational structures. Interest points are extracted from intersections of these grouped lines.

Nelson and Selinger [110] use groups of image contours. Hereby, a curve-finding algorithm is applied in order to generate discontinuous segmented contours where the bounding points are locations of high curvature. Curves that intersect a square region around the curve are normalized and mapped onto it.

Mikolajczyk et al. [103] extract edges from the image employing a multi-scale Canny edge detector with Gaussian derivatives and then the edge neighborhood is determined which is related to the scale. Interest points are found on the edges where the radius of the interest point coincides with the distance of the point to the neighboring edge. Interest points in a homogeneous region are rejected due to not having a distinctive extremum in the scale space. In Figure 3.1, interest points are denoted at two different scales (scale factor of 2) as green circles; the illustrations Figure 3.1 b) and Figure 3.1 d) give the corresponding edge images. Similar interest points are found in both images.

3.1.2 Corner Detectors

Corner detectors select image locations having high curvature, i.e. strong gradients in all directions at a pre-defined image scale. Hence, they do not solely detect corners but dots and any other locations having high contrast in orthogonal orientations [93, 107].

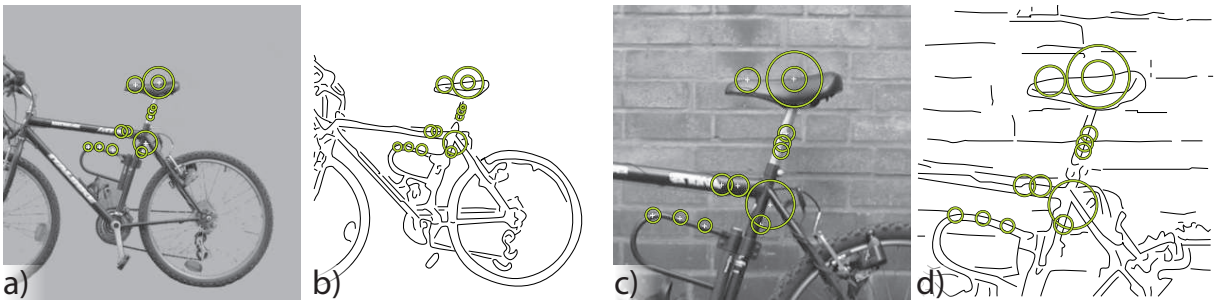


Figure 3.1: Selected interest points in images (a,c), separated by a scale factor of 2, and in the edge images (b,d). Interest points are illustrated with green circles (Figure taken from [103])

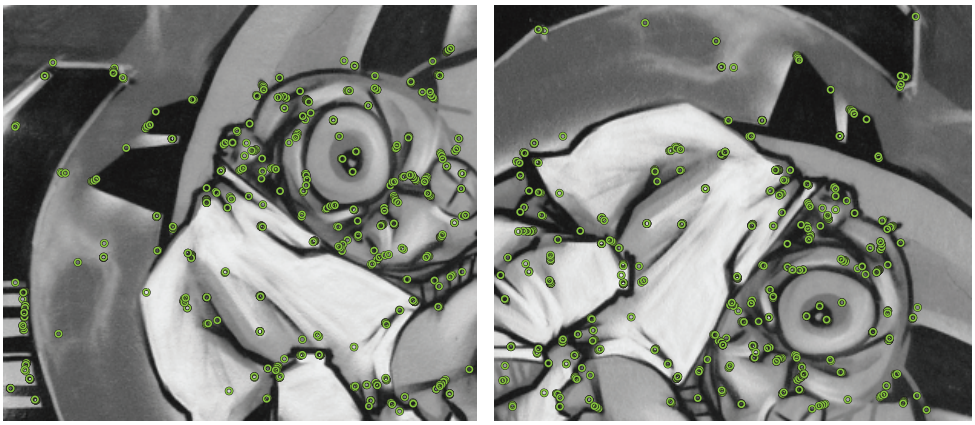


Figure 3.2: Corners found by the Harris corner detector (Figure taken from [134])

Moravec [107] first introduced local interest points for stereo matching in 1981. He defines an interest point as an image location which is repeatable and unambiguous. Hence, he states that regions having strong gradients in orthogonal directions, such as corners, are suitable for interest points. In order to find corners, Moravec's detector computes the squared sums of adjacent pixel differences along four directions separated by 45° in a local window in order to get a directional variance. An interest point is then located at a local maximum of the minimal sums in a region. As a result of the summations, the detector is sensitive to edges dissimilar to these four orientations.

Moravec's corner detector was developed further by Harris and Stephens [62] in order to improve the repeatability of the interest points in presence of edges and local image distortions such as noise. They apply an analytic expansion about the shift origin in order to avoid discrete directions. Furthermore, they change Moravec's binary rectangular window to a Gaussian weighted window to counterbalance the sensitivity to noise. Additionally, they employ a matrix related to the ACF – the second moment matrix – as new feature calculation for the corner detection. An interest point is detected at locations where the matrix has two significant eigenvalues. This so-called Harris corner detector, however, is sensitive to changes in image scale. Other ACF-based corner detectors were suggested by [47, 48, 133]

In 1997, Smith and Brady introduced Smallest Univalue Segment Assimilating Nu-

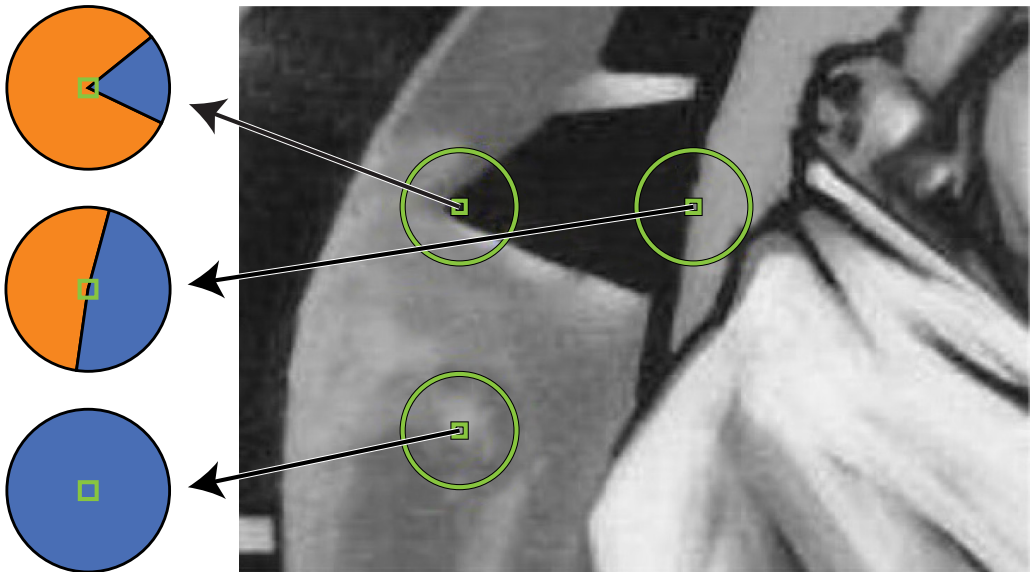


Figure 3.3: Interest points detected by the SUSAN corner detector. The neighborhood is divided into "similar" (orange) and "dissimilar" (blue). Corners are characterized by a minimum area of "similarity" (Figure taken from [134])

cleus (SUSAN) corner detector [128] able to find corners, edges and can be used for noise reduction in images while preserving structures. They use a non-linear filtering approach starting from a center pixel of a circular region having a predetermined scale. The intensity value of this center pixel is used as reference value for the filtering, and pixels are categorized according to their similarity in intensity. Each image point is then characterized by a ratio giving the homogeneity of its neighborhood. For edges, the ratio is close to 50 %, whereas for corners it is about 25 %. Thus, corners are detected at locations of local minima of the ratio using a threshold (see Figure 3.3).

Harris corner detector and SUSAN are invariant to translation and rotation, however, not to scale changes or affine transformations. To achieve scale invariance, these detectors can be applied to a scale space introduced by Lindeberg [89, 90] to extract corners at different scales.

Dufournaud et al. [41] adapted the Harris corner detector to scale changes employing a scale space with Gaussian filters. Thus, interest points are computed at different scale levels according to the scale space. The correct scale for the interest points is determined by a homography based matching algorithm. Figure 3.4 shows an example of mapping a high- image to a low-resolution image.

Mikolajczyk and Schmid [100] introduce a scale invariant detector, where they developed further the Harris corner detector. First, a multi-scale Harris corner detector is applied to spatially locate interest points, then a Laplacian operator is used to determine the characteristic scale of the corner. The scale is then chosen where the Laplacian response leads to a maximum, which is the case when the scale corresponds to the scale of the local structures in the image (see Figure 3.5). Figure 3.6 gives an example of scale-invariant interest points detected by the Harris-Laplacian, the interest points are denoted by green circles. This detector is made invariant to affine transformations using ellipses

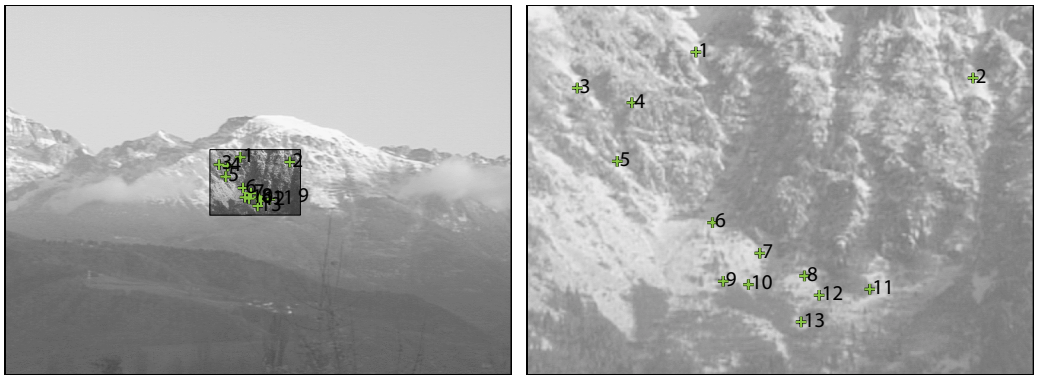


Figure 3.4: Mapping of a high- onto a low-resolution image using Dufournaud's adapted Harris detector (Figure taken from [41])

that fit the corresponding image structure instead of circles. Hereby, the affine transformation is estimated employing the second moment matrix and subsequently, the region is normalized to a circular shape. In Figure 3.7, corresponding interest points marked as green circles are found with the Harris-Affine detector in both images.

The Features from Accelerated Segment Test (FAST) detector introduced by Rosten and Drummond [117,118] builds upon the SUSAN, it detects a corner based on the fraction of pixels having higher (resp. lower) intensity value than the center pixel at a given radius. Whereas SUSAN divides the intensity values into two categories based on the similarity, for FAST, the pixels are categorized into three classes according to their brightness relative to the center pixel. Figure 3.8 illustrates the corner detection. Adjacent pixels at the given radius brighter than the center pixel are taken into account for the detection. The decision for an interest point is taken employing decision trees. The Laplacian function is applied in [86] in order to achieve scale-invariance for the FAST detector.

3.1.3 Blob Detectors

Blob detectors aim at locating blob-like structures in an image, hence they are looking for local minima or maxima which are brighter or darker than their neighborhood. Scale and shape of blobs structures are more definitely determined than those of corners, as they can be inferred from their boundaries. Due to irregularities in the shape of the region, their exact location is prone not to be as accurate as corners' locations, which can be ascribed to a single coordinate pair in the image plane. A corner is an intersection of two or more lines which, however, exists at various scales. Hence, the problem of assigning a specific scale to a corner is ambiguous [134].

Combinations of corner and blob detectors have been used to exploit their respective strengths and to improve the overall coverage of the image with interest points, hence the dependency of the method to the image content is reduced [134].

Beaudet [16] suggested calculating the Determinant-of-Hessian (DoH) matrix to detect blob-like structures. The matrix is calculated from the second-order derivates of the Gaussian smoothed image. Blob-like structures are spatially located where the determinants of the matrix are maximal and in located in scale at maxima of its trace, the LoG.

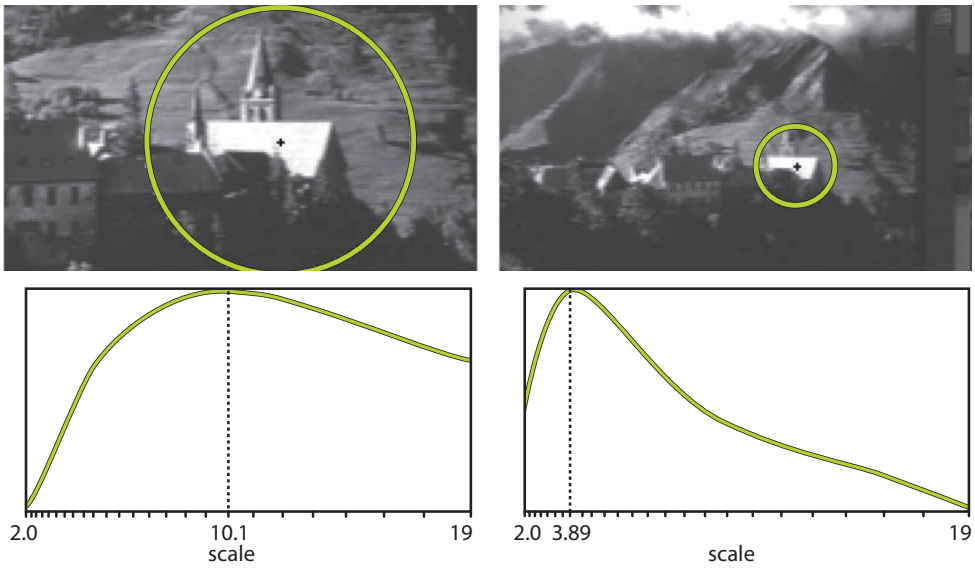


Figure 3.5: Illustration of the characteristic scales for corresponding interest points. The top row shows the images with the interest points, whereas the bottom row gives the responses of the Laplacian function over scales. It reaches a maximum where the interest points correspond. (Figure taken from [100])

The LoG with $\sigma = 7$ and filter size 43 is illustrated in Figure 3.10 as 3D Plot (left) and in normalized gray scale coding (right), where black pixels indicate negative values, white pixels positive values and gray pixels correspond to zero. An example for scale-invariant Hessian interest points is given in Figure 3.11.

Mikolajczyk and Schmid extended the DoH [100] to an affine invariant detector similar to the Harris-Affine detector (see Figure 3.5). For an example of corresponding points see Figure 3.12.

The DoG proposed by [92] detects interest points exploiting a scale space. The scale space is established by successively differencing the image convolved with a Gaussian function having an increasing scale parameter σ [93]. Interest points are detected at locations of local minima and maxima of the differential images. Their scale is estimated by detecting the local extrema over the scale space. Figure 3.13 illustrates interest points found by the DoG detector. The DoG is explained in further detail in Section 3.1.4.

In their work [14, 15], Bay et al. propose Speeded Up Robust Features (Surf), an interest point detector based on a fast approximation of the Hessian matrix. To improve the computation time, integral images are employed as approximation. For Surf, the DoH is exploited for determining both, the spatial and scale location of an interest point. To find local maxima, the discretized Gaussian second-order partial derivative is further approximated with a box-filter as shown in Figure 3.14, since their computation is very fast. Figure 3.15 illustrates interest points detected in two images having different scales.

3.1.4 Difference-of-Gaussian Interest Point Detector

The DoG detector is chosen as interest point detector in this report on account of studies by [93,101,102]. It locates stable interest points in a scale-invariant manner employing the

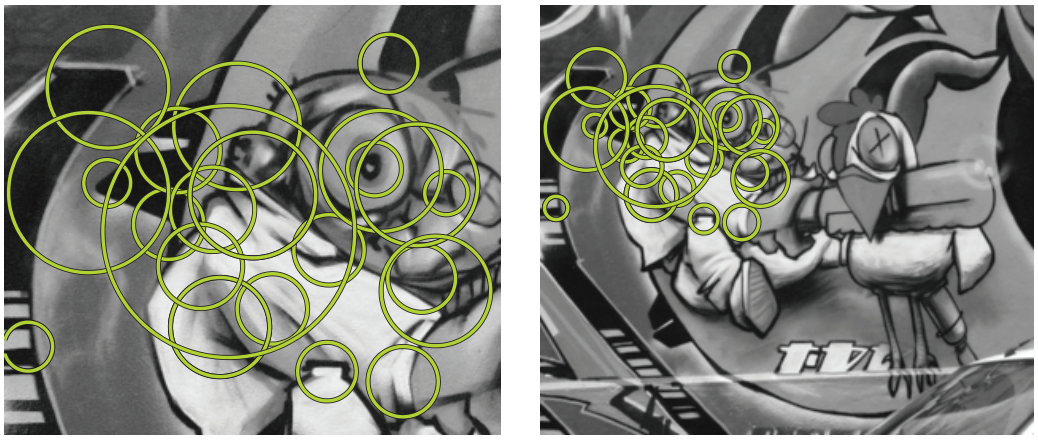


Figure 3.6: Corresponding interest points found in both image using the Harris-Laplace detector (Figure taken from [134])

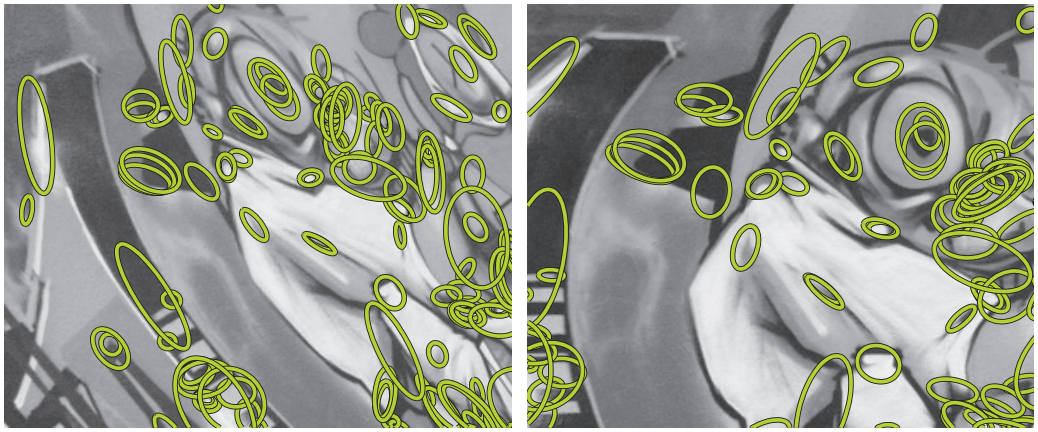


Figure 3.7: Corresponding interest points found in both image using the affine invariant Harris detector (Figure taken from [134])

scale-space [140]. The underlying idea is the observation that objects consist of different structures at different scales which can be extracted exploiting the scale-space.

Scale Space

The scale-space is a *"continuous function of scale"* [93] which models the perception of objects at different distances and hence, at different scales. Large scales allow perceiving fine details while small scales show the coarse structures of an image. Constructing the scale-space of an image means building fine-to-coarse representations of an image, where high-frequency spatial information and fine structures are successively suppressed [90,91]. Thus, an image is represented by a group of derived signals which allows having the image represented at all scales simultaneously. Figure 3.16 shows an image successively smoothed with a larger filter suppressing the details, the filter size is given for each image representation.

Lindeberg states that the coarse scales should be *"simplifications of corresponding structures at finer scales"* [91] and must not introduce new structures and artefacts in the

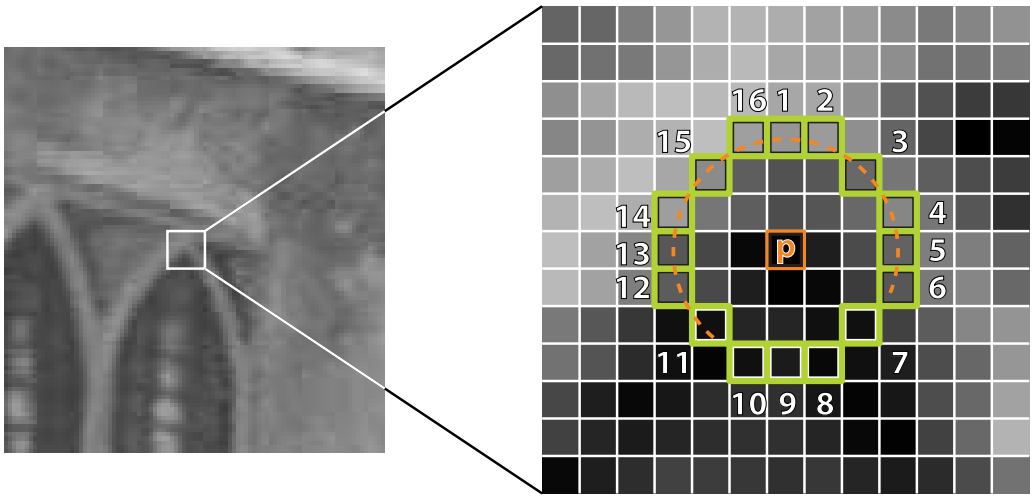


Figure 3.8: FAST uses the highlighted squares (green) at a given radius (orange arc) around the center pixel p (orange) to detect corners. Adjacent pixels brighter than p by a threshold are taken into account for the corner detection (indicated by the orange arc). (Figure taken from [118])

course of suppressing fine details. Furthermore, linearity and invariance to spatial shifts, rotation and scale transformation are necessary. The scale-space should satisfy the semi-group property, which means that a coarse-scale representation is to be computed either from any fine-scale or the original image applying the same transformation. Furthermore, local extrema are not to be enhanced by the smoothing kernel. These requirements are fulfilled by the Gaussian kernel and its derivatives [90, 91]. Hence, the scale-space $L(x, y, \sigma)$ is constructed through the convolution of the image $f(x, y)$ with a Gaussian kernel $G(x, y, \sigma)$ of increasing scale σ:

$$L(x, y, \sigma) = G(x, y, \sigma) * f(x, y) \tag{3.1}$$

where $*$ denotes the convolution in x and y direction and σ being the scale parameter, the standard deviation of the Gaussian kernel. As σ increases, the image is convolved with a larger kernel, which progressively removes high-frequency details. The Gaussian kernel $G(x, y, \sigma)$ is defined by:

$$G(x, y, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-(x^2+y^2)/2\sigma^2} \tag{3.2}$$

Since the Gaussian kernel is a symmetric kernel and hence, separable, the image can be efficiently convolved by the same 1D kernel in the horizontal and the vertical direction successively:

$$\begin{aligned} L(x, y, \sigma) &= G(x, \sigma)^T * (G(x, \sigma) * f(x, y)) \\ G(x, \sigma) &= \frac{1}{\sqrt{2\pi\sigma}} \exp^{-x^2/2\sigma^2} \end{aligned} \tag{3.3}$$

where $G(x, \sigma)^T$ denotes the transposed 1D Gaussian.

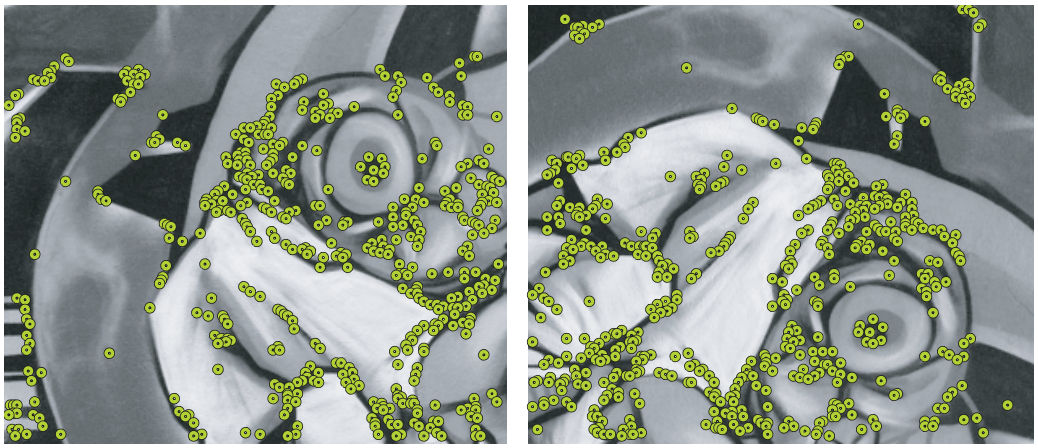


Figure 3.9: Corresponding interest points found in both image using the FAST detector (Figure taken from [134])

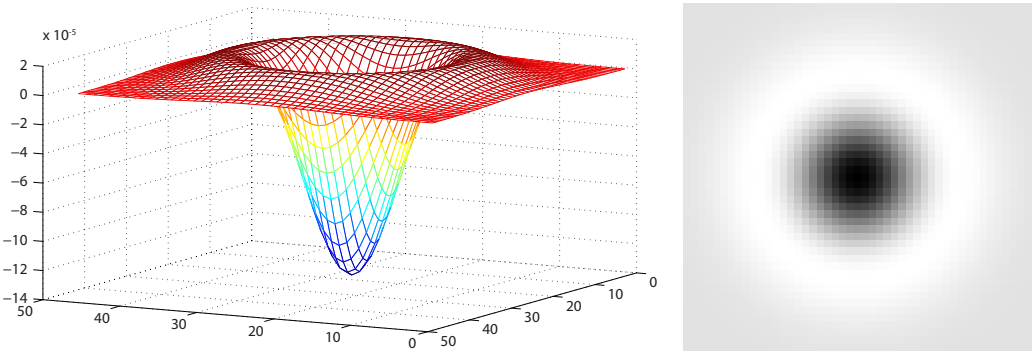


Figure 3.10: LoG as 3D Plot (left) with color coding, and in normalized gray scale coding (right)

For a more efficient computation, the image is resampled by a factor of 2 after the value of the parameter σ doubled, i.e. it has twice the initial value of σ. The resampling is done leaving out every other pixel in each row and column. A set of image representations having the same size is then called octave.

Difference-of-Gaussian

Having established the scale space, interest points are detected by means of local extrema in the scale space. Hereby, the DoG $D(x, y, \sigma)$ is employed which is computed differencing two adjacent scales. Two adjacent images in the scale space are convolved with a Gaussian kernel separated by a constant multiplicative factor k in scale space:

$$\begin{aligned} D(x, y, \sigma) &= (G(x, y, k\sigma) - G(x, y, \sigma)) * f(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma) \end{aligned} \tag{3.4}$$

As the scale space $L(x, y, \sigma)$ has to be computed for the extraction of scale-invariant feature description, the DoG is established by a subtraction of adjacent image representations in the scale space. As shown by Mikolajczyk in his work [97], the extrema of the

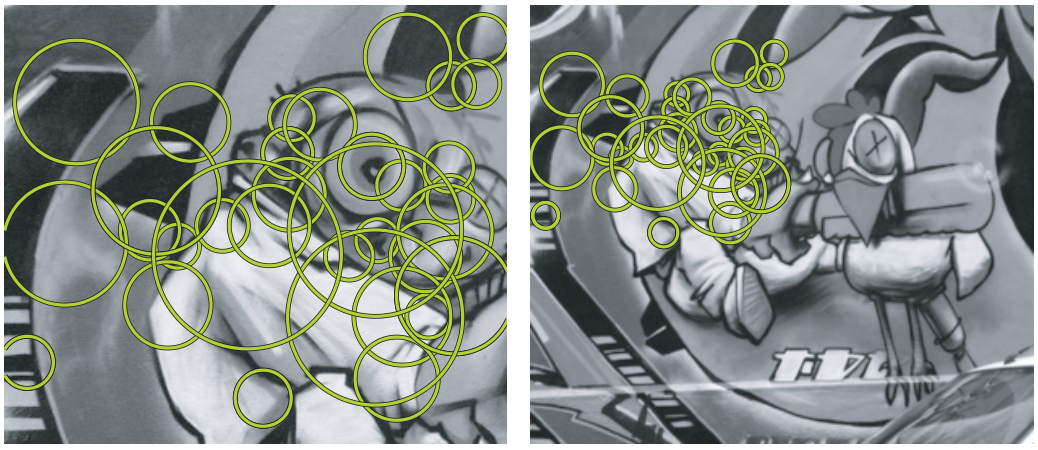


Figure 3.11: Corresponding interest points found in both images with scale change using the Hessian-Laplacian detector (Figure taken from [134])

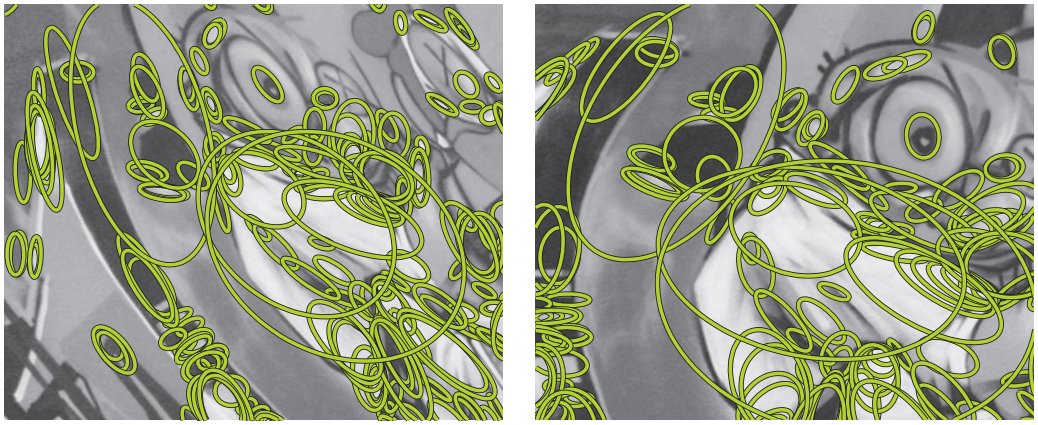


Figure 3.12: Corresponding interest points found in both images using the Hessian affine detector (Figure taken from [134])

scale-normalized LoG $\sigma^2\nabla^2 G$ produces the most stable interest points when compared to other methods such as Hessian or Harris corner detectors (see Section 3.1.2). The *"Laplacian acts as a matched filter when its scale is adapted to the scale of a local image structure"*, thus it has the highest response where the scale complies to the structure of the image (see Figure 3.5). The $\sigma^2\nabla^2 G$ corresponds to the image's derivative in scale direction as can be shown by the diffusion equation in scale space theory. Since the derivative of an image is approximated by the difference between two adjacent pixels, the $\sigma^2\nabla^2 G$ can be efficiently approximated as the difference between two representations of an image smoothed with Gaussian kernels having different scales σ.

In Figure 3.17, the computation of the DoG is illustrated. The image is repeatedly smoothed with Gaussian kernels having increasing filter sizes σ to produce the scale space representations (white). The DoG representations on the right (green) are then constructed subtracting adjacent scale space representations. Having processed an octave, the process is reinitiated with the image resampled by a factor 2.

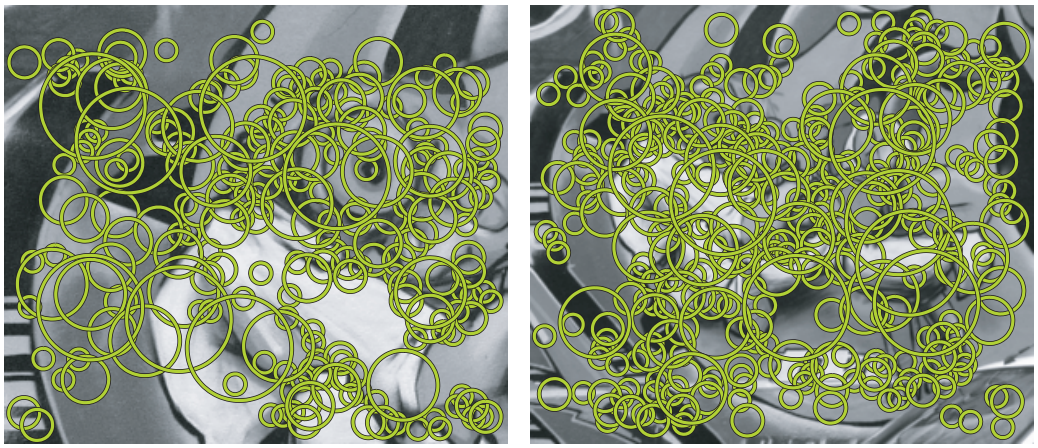


Figure 3.13: Corresponding interest points found in both image using the DoG detector (Figure taken from [134])

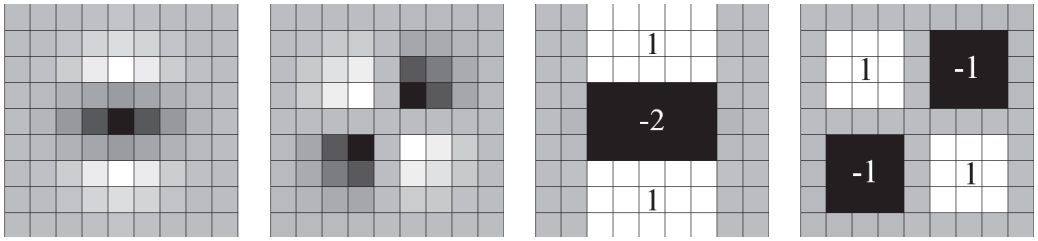


Figure 3.14: Left: The Gaussian second-order partial derivates (y-direction and xy) and their respective SURF equivalents (right). Gray regions are equal to zero. (Figure taken from [15])

Local Extrema Detection

Interest points are detected according to Mikolajczyk's conclusion [97] of minima and maxima of the scale-normalized LoG and hence, of the DoG, being the most stable interest points. Thus, in order to localize interest points spatially and in scale, each pixel of the $D(x, y, \sigma)$ is compared to its 3×3 neighborhood in the current scale and to the nine respective neighbors in the two adjacent scales. Figure 3.18 illustrates this detection. The prospected pixel is marked with an orange x and its neighbors in scale space are indicated by green circles. An interest point candidate is defined as lying at a local minimum or maximum; i.e. where all neighbors of the pixel in scale and space have a higher, or respectively lower, pixel value.

Interest Point Localization

Having a set of interest point candidates, their location and scale needs to be determined. While in [92], the interest point is located at the location and scale of the center pixel used for the local extrema detection, Lowe [93] suggests interpolating the location and scale. Hereby, a 3D quadratic function is fitted to the local neighborhood of the pixel in order to retrieve a higher repeatability as proposed by Brown and Lowe [22]. This allows for the rejection of unstable interest point candidates having low contrast using a threshold.

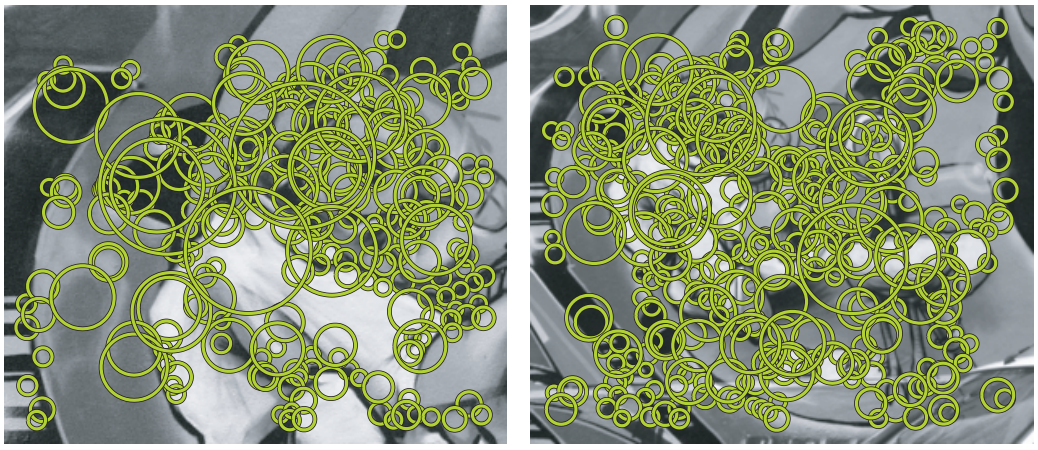


Figure 3.15: Corresponding interest points found in both image using the SURF detector (Figure taken from [134])

In contrast to the LoG, employing the DoG as interest point detector leads to local maxima close to contours or edges. However, the localization of such an interest point is poor along the edge, since the intensity value just changes in one direction. Hence, interest points located at edges are sensitive to artefacts such as noise. Employing the 2×2 Hessian matrix $\mathbf{H}$ at their scale and location, these interest point candidates are identified:

$$\mathbf{H} = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix} \tag{3.5}$$

with D being the second partial derivates which are estimated differencing neighboring pixels. For extrema located close to edges, the ratio of the principal curvatures in perpendicular directions is high – i.e. these extrema have a large principal curvature along the edge, whereas the one in the perpendicular direction is small. In order to identify interest point candidates along edges, the ratio needs to be below the threshold r. Since the ratio of the principal curvatures is important, Lowe [93] avoids calculating the eigenvalues of the matrix according to [62] and the verification of interest point candidates can be done using the measure

$$\frac{\mathrm{Tr}(\mathbf{H})}{\det(\mathbf{H})} < \frac{(r+1)^2}{r} \tag{3.6}$$

with $r = 10$ being a threshold [93], $\det(\mathbf{H})$ being the determinant of the Hessian matrix and $\mathrm{Tr}(\mathbf{H})$ is defined by:

$$\begin{aligned} \det(\mathbf{H}) &= D_{xx}D_{yy} - (D_{xy})^2 \\ \mathrm{Tr}(\mathbf{H}) &= D_{xx} + D_{yy} \end{aligned}$$

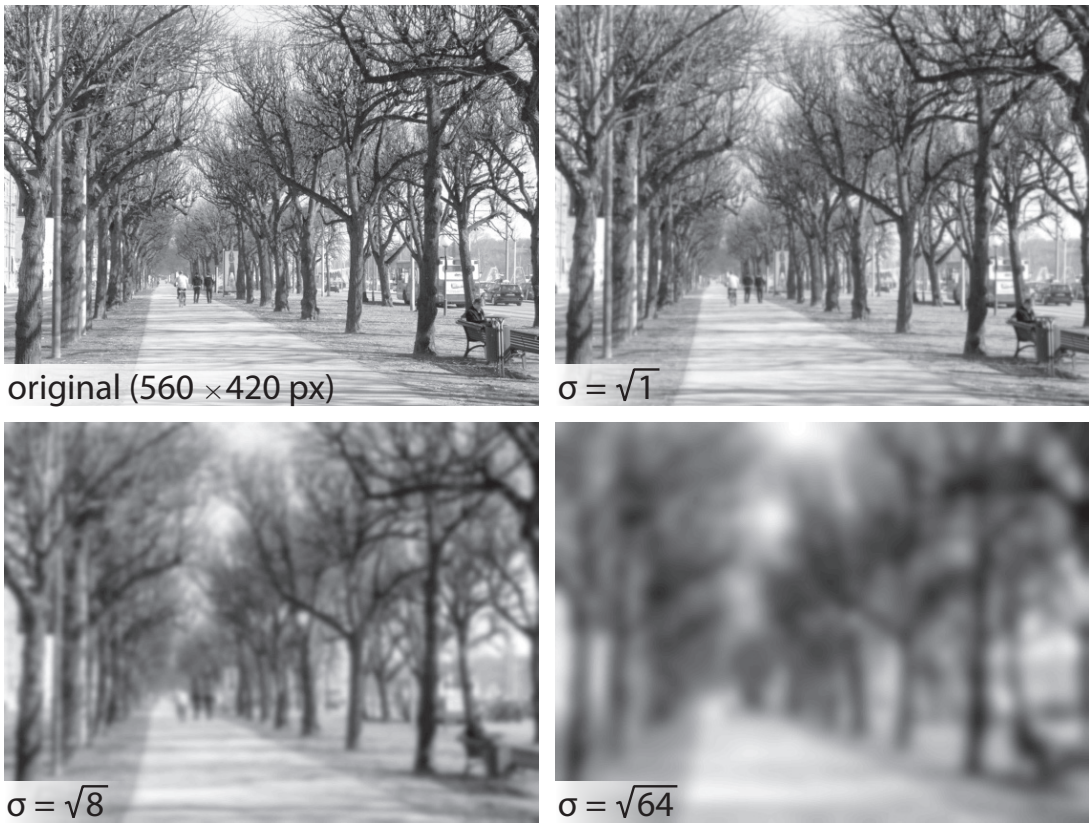


Figure 3.16: Suppressing the fine scales successively; coarse scales as simplifications of the fine scales (Figure taken from [91])

3.2 Local Descriptor

Having located interest points, measurements are computed in a region centered at their positions and transformed into a local descriptor characterizing the respective region [101, 134]. This section gives an overview of selected local descriptors.

Local descriptors describe an image region in a distinctive way independent of a set of transformations such as affine transformations and do not require a binarization or segmentation of the object. As described in Section 3.1, the advantage of using local features to global features is that local features are robust to occlusion, image clutter, artefacts, changes in viewpoint or shape of the object [93, 97, 134]. Local descriptors were first used for stereo matching, but have a broader field of application: Schmid and Mohr [120] first proposed to use local features for image retrieval. Object recognition [43, 44, 93], recognition of object categories [27, 39, 43, 98, 146], or texture recognition [81, 146] are further fields. Mikolajczyk and Schmid give an overview of local descriptors in [101].

The simplest descriptor is a normalized vector of pixel intensities in the local region around the interest point according to its scale. The similarity of two descriptors can then be computed employing cross-correlation. However, such a descriptor is sensitive to affine transformations or view point changes.

Mikolajczyk and Schmid [101] categorize local descriptors into distribution-based de-

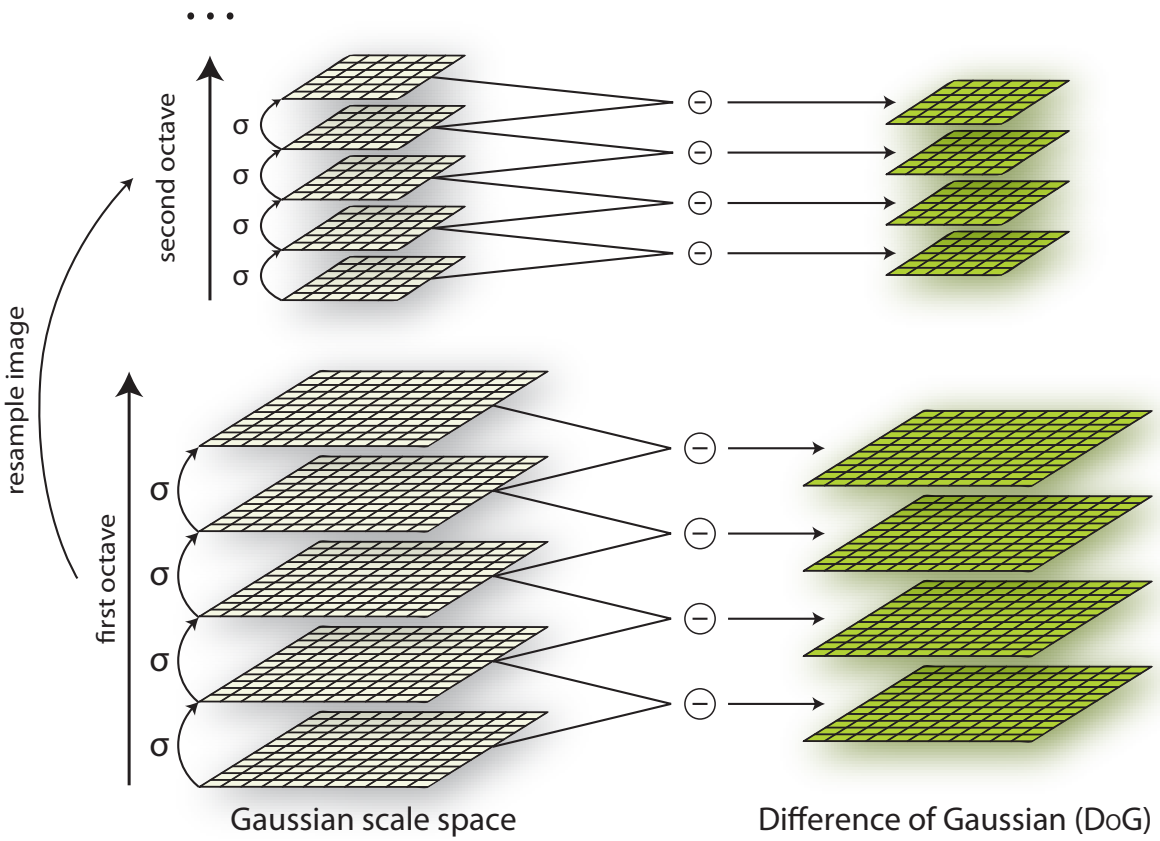


Figure 3.17: Computation of the DoG (Figure taken from [93, 134])

scriptors which use histograms to characterize an image region, spatial-frequency based descriptors which describe the local frequency distributions, and differential descriptors making use of image derivatives. In the following, the focus will be on distribution-based descriptors, since one descriptor of this group is applied in this report.

3.2.1 Distribution-Based Descriptors

These descriptors employ histograms of locally sampled data in order to characterize the local structures or shapes. Johnson and Hebert [66] introduce a histogram of local point positions for interest points in 3D space by describing coordinates of neighboring points with respect to a basis point which they call *spin image*. Lazebnik et al. [81] modified this approach for the representation of texture in 2D space. They construct a two-dimensional histogram where the intensity values of pixels and the distance from the center point of the image patch give the dimensions (see Figure 3.19). They state that their descriptor is invariant to viewpoint changes and rotation.

Zabih and Woodfill [144] propose a descriptor robust to changes in illumination. Instead of raw pixel intensities, their histogram is based on relations between pixel intensities. The relationships are encoded using binary strings. If a high-dimensional descriptor is used, an image region is distinctively represented.

Belongie and Malik [19] developed a shape descriptor called *shape context*, which

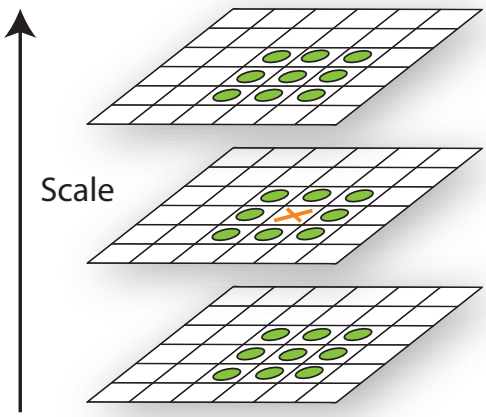


Figure 3.18: Extrema detection in the scale space (Figure taken from [93])

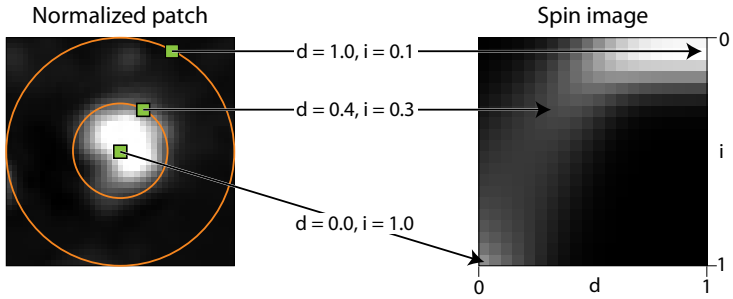


Figure 3.19: Construction of a spin image: sample points in the normalized patch (left) transformed into different locations in the spin image (right). (Figure taken from [81])

relies on the extraction of contours employing the Canny edge detection algorithm [26]. Randomly selected points are extracted along these contours. For each of these points, a histogram in log-polar space is generated, which contains the coordinates of the remaining points relative to the respective interest point. The location of the points is quantized into nine bins. This leads to a distribution of relative positions of interest points, where close interest points are emphasized.

Lowe [92,93] introduce a descriptor based on the image's gradient magnitude and gradient orientation. In order to compute them rotationally invariant, the main orientation is estimated for each interest point. Normalizing the feature vector according to the main orientation allows for a representation that is independent to rotational changes. SIFT descriptors are 128-dimensional gradient histograms. Each orientation histogram represents gradient vectors with specific orientations $(0\,°, 45\,°, 90\,°, ..., 315\,°)$. Thus, gradient vectors of an interest point are accumulated to the gradient histograms according to their orientation and their spatial location (4×4 bins). A tri-linear interpolation guarantees a robust representation of a specific image region.

Further developments of SIFT include Rotation Invariant Feature Transform (RIFT), which is a rotation-invariant generalization of SIFT proposed by Lazebnik et al. [82]. They

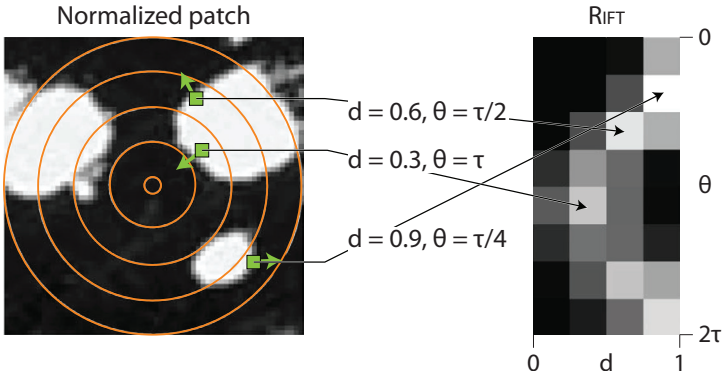


Figure 3.20: Construction of a RIFT descriptor: sample points in the normalized patch (left) transformed into different locations in the descriptor (right). (Figure taken from [80])

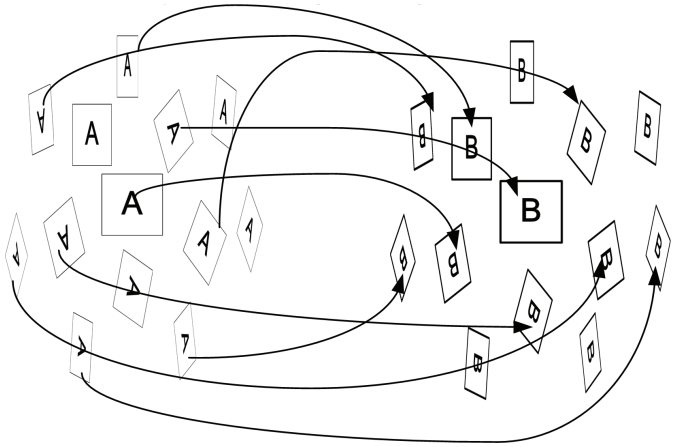


Figure 3.21: ASIFT Algorithm: pairs of rotated and tilted representations obtained from images A and B are compared by SIFT. (Figure taken from [143])

use a circular normalized image patch divided into concentric rings, where histograms are computed for each ring (see Figure 3.20 for the construction of a RIFT histogram). Color Scale Invariant Feature Transform (CSIFT) proposed by Abdel-Hakim and Farag [1] extends the SIFT for the use of color images. Morel and Yu [108, 143] suggested an affine-invariant version of SIFT, the Affine Scale Invariant Feature Transform (ASIFT), by simulating the camera axis orientations (see Figure 3.21).

Ke and Sukthankar [71] have developed an approach that is similar to SIFT, but employs Principal Component Analysis (PCA) to the normalized image gradient patch instead of weighted histograms used in SIFT. Hereby, a fixed-size image patch of 41×41 pixels centered at each interest point is extracted, rotated according to its predominate orientation. Then, a $3,042$ element feature vector is created extracting image gradients in x and y direction, and normalized in order to reduce the sensitivity to illumination changes similar to SIFT. PCA is applied in order to reduce the feature vectors dimensionality. The covariance matrix for the PCA is applied to 21,000 image patches. Having sorted the eigenvectors according to their importance, the top n are taken into account. Mikolajczyk shows [101] that PCA-SIFT is less distinctive than SIFT.

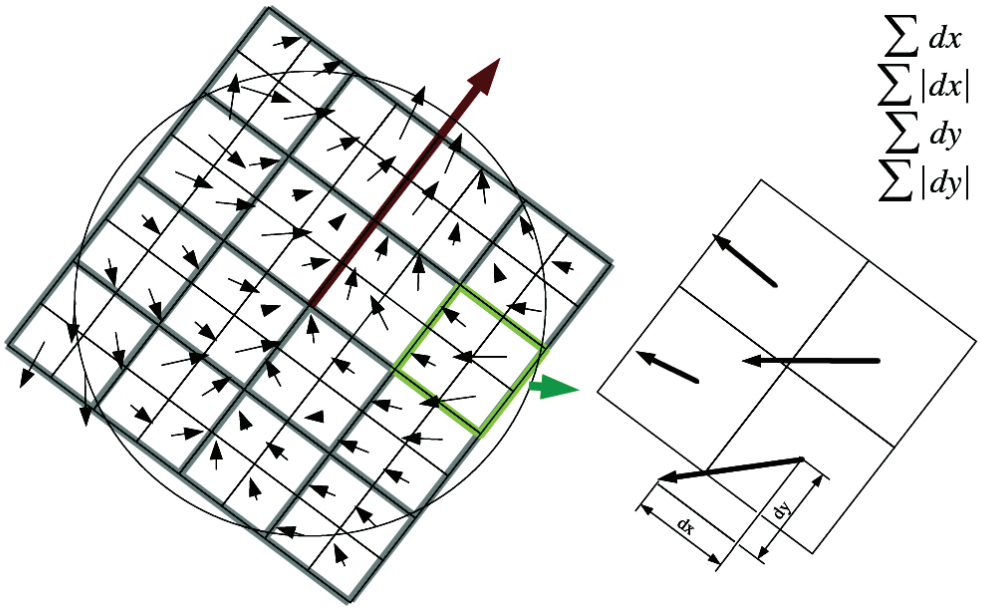


Figure 3.22: Construction of the SURF descriptor: an oriented grid hacing 4×4 square subregions is established at the interest point (left). The 2×2 grids are the actual fields of the descriptor, where sums $dx, |dx|, dy, |dy|$ are compued with respect to the grid orientation (right). (Figure taken from [14])

Mikolajczyk and Schmid [101] introduce an extension of SIFT, the Gradient Location-Orientation Histogram (GLOH), where the location grid is changed and the feature space dimensionality is reduced applying PCA in order to improve the descriptors robustness and distinctiveness. They employ a log-polar grid having two concentric rings divided in 8 angular bins each, resulting in a total number of 17 location bins as the center bin is not divided. A feature vector of dimensionality 272 – gradient orientations are quantized in 16 bins – is constructed which is then reduced by means of PCA. The descriptor is then built of the 128 largest eigenvectors.

Bay et al. [14, 15] propose SURF descriptors, where they characterize the distribution of intensity values similar to SIFT. However, instead of gradients, they employ first order 2D Haar wavelets in x and y direction exploiting integral images built for the detection of interest points as described in Section 3.1.3. Rotation invariance is achieved by normalization of the descriptor according to is predominant orientation. A square region centered at the interest point is divided into 4×4 subregions, where each bin comprises 25 Haar wavelet responses. The responses and their absolute values are summed up over each subregion in x and y direction and lead to a descriptor having a dimensionality of 64. The construction of a descriptor is summarized in Figure 3.22.

3.2.2 Other Techniques

Apart from descriptors based on the distribution of intensities, other techniques have been proposed. Differential descriptors make use of image derivatives up to a given order to characterize an interest point's neighborhood. Koenderink and van Doom [75] developed

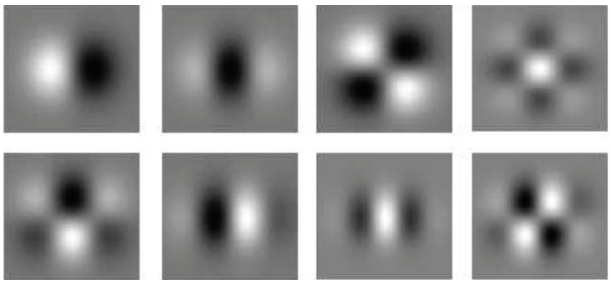


Figure 3.23: Gaussian derivatives up to the forth order. (Figure taken from [101])

the *local jet*, a method applying local spatial derivatives computed by means of convoluting the image with Gaussian derivatives (see Figure 3.23). Florack et al. [45] further developed the method in order to gain rotation invariance through combining *local jet* components. A further approach to rotational invariance is introduced by Freeman and Adelson [49], who introduce steerable filters, where derivates are steered in gradient direction. A *multi-scale local jet* exploiting the Gaussian scale-space theory was investigated by Florack et al. [46]. The Gabor filter transform proposed in [51] allows the investigation of frequencies in a local region, however, in order to characterize frequency and orientation appropriately, a large number of filters is required. Van Gool et al. [57] introduce generalized moment invariants in order to characterize multi-spectral data. Hereby, the invariants combine central moments which describe shape and intensity distribution around an interest point.

3.2.3 Scale Invariant Feature Transform

SIFT descriptors were shortly summarized in Section 3.2.1. Since SIFT descriptors are chosen for the approach introduced in this report, they will be described in more detail in this section.

SIFT is introduced by Lowe [92] and further improved in [93] for the field of object recognition in different camera views. Its features are invariant with respect to translation, scale and rotation transformations of the image, and further robust to affine transformations and illumination changes.

Orientation Normalization

To each interest point extracted by means of DoG as described in Section 3.1.4, Lowe [93] first assigns a main orientation to achieve rotation invariance, and then, a high-dimensional descriptor is computed. Assigning the predominant orientation to the interest point, it is provided with a local coordinate system where the descriptor is represented relatively to this orientation. Hence, the descriptor does not need to be computed in a rotation invariant manner.

The computation of the prevailing orientation is based on image gradients [93]. Hereby, the respective Gaussian smoothed image representation $L(x, y, \sigma)$ of the scale space is determined according to the scale σ of the interest point to achieve invariance to the image scale. For efficiency, the gradient magnitude $m(x, y)$ and the gradient orientation $\theta(x, y)$ are precomputed for the entire image representation $L(x, y)$ at scale σ by means

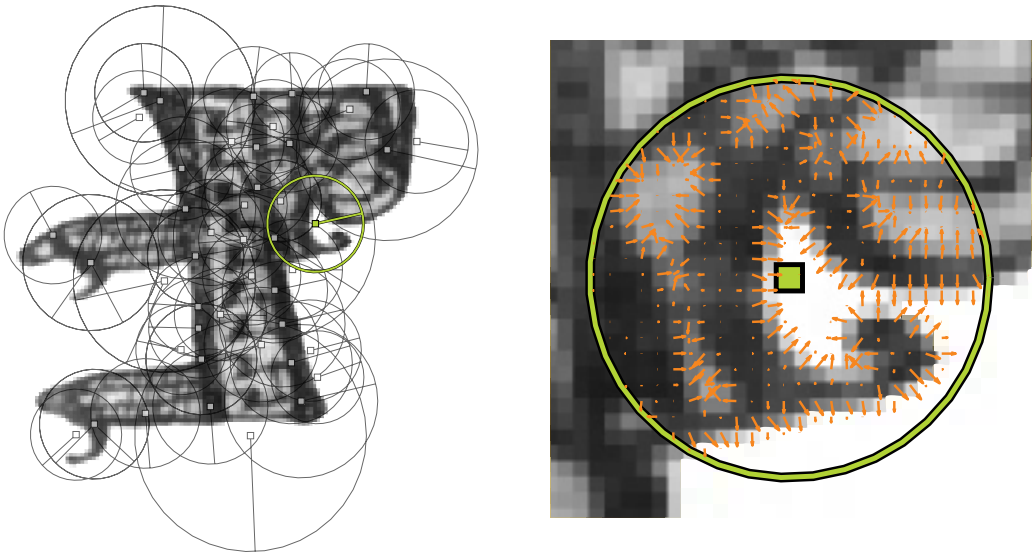


Figure 3.24: Glagolitic decorative initial with overlaid interest points (left). The right side shows a selected interest point (green) with gradients indicated by orange arrows.

of pixel differencing:

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 \; + \; (L(x,y+1) - L(x,y-1))^2} \tag{3.7}$$

$$\theta(x,y) = tan^{-1} \frac{L(x,y+1) - L(x,y-1)}{L(x+1,y) - L(x-1,y)} \tag{3.8}$$

Figure 3.24 shows a Glagolitic decorative initial overlaid with representations of interest points (left), where the locations are indicated with white squares, the scale is illustrated with a circle and the main direction is given by the line connecting the squares and the circles. The right side shows a selected interest point (green) with gradient orientations illustrated by orange arrows and magnitudes indicated by the length of the arrows.

A histogram of orientations within the local neighborhood of the interest point is built [93], where orientations are quantized in 10 ° bins, and hence leading to a histogram having 36 bins. The orientations are weighted by their gradient magnitude. In order to weight the orientations spatially, a Gaussian window having a σ being 1.5 times the scale of the interest point. Thus, orientations close to the center of an interest point are enhanced whereas the influence of those at the border of the interest point's spatial extend are alleviated. Thus, poor localization of interest points owing to artefacts such as noise or small image perturbations and affine distortions are compensated. For the same reason, the histogram is smoothed by a one dimensional Gaussian prior to the determination of the main orientation [92].

Peaks in the orientation histogram indicate prevailing orientations of local gradients, with the global maximum giving the canonical orientation. For each local peak having a value greater than 80 % of the global peak, an interest point is created with the respective orientation assigned (see Figure 3.25). In this case, multiple interest points are located

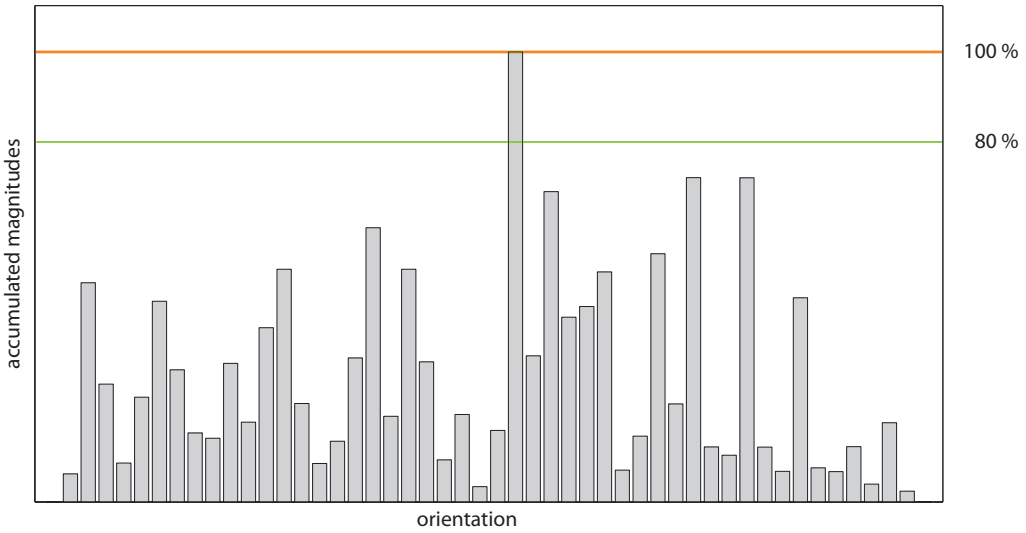


Figure 3.25: Histogram of orientations with the 100 % given as orange line and the 80 % interval denoted by the green line. The histogram has been resampled to 50 bins for visualization reasons.

at the same position, each one having a distinct main orientation. In order to achieve an accurate dominant orientation, the exact peak position is determined by an interpolation done by means of fitting a parabola to the adjacent histogram values.

Descriptor Computation

The concept of interest points detected by means of DoG and the subsequent assignment of a main orientation to the interest point provide each feature location with a local coordinate system which is embedded into the global image coordinate system according to the location and scale parameters from the DoG and the orientation of the main direction. A descriptor can then be calculated in the local region without the need of incorporating these invariances [93]. Robustness to illumination and affine transformations need to be embedded in the descriptor.

A Sift descriptor [93] is a 3D histogram of gradient location, magnitude $m(x, y)$ and the orientation $\theta(x, y)$. As for the orientation assignment, the scale of the image representation is determined by the scale σ of the interest point. The gradient magnitudes and orientations are extracted at the spatial extend of the interest point with the orientations being rotated according to the dominant orientation of the interest point. As for the determination of the dominant orientation, the gradient magnitudes are spatially weighted using a Gaussian kernel having a σ of half the scale of the interest point to counterbalance the effect of potential poor localization due to image artefacts. Figure 3.26 illustrates the process for a 2×2 descriptor, where the Gaussian window is indicated by the orange circle.

The region around the interest point is divided into a 4×4 location grid, where the gradient locations are quantized to. Thus, they approximate the spatial distribution of gradients. Gradient orientations are assigned to eight different orientation planes covering $45\,^\circ$ each $(0\,^\circ, 45\,^\circ, 90\,^\circ, ..., 315\,^\circ)$. Thus, a set of 4×4 orientation histograms having eight

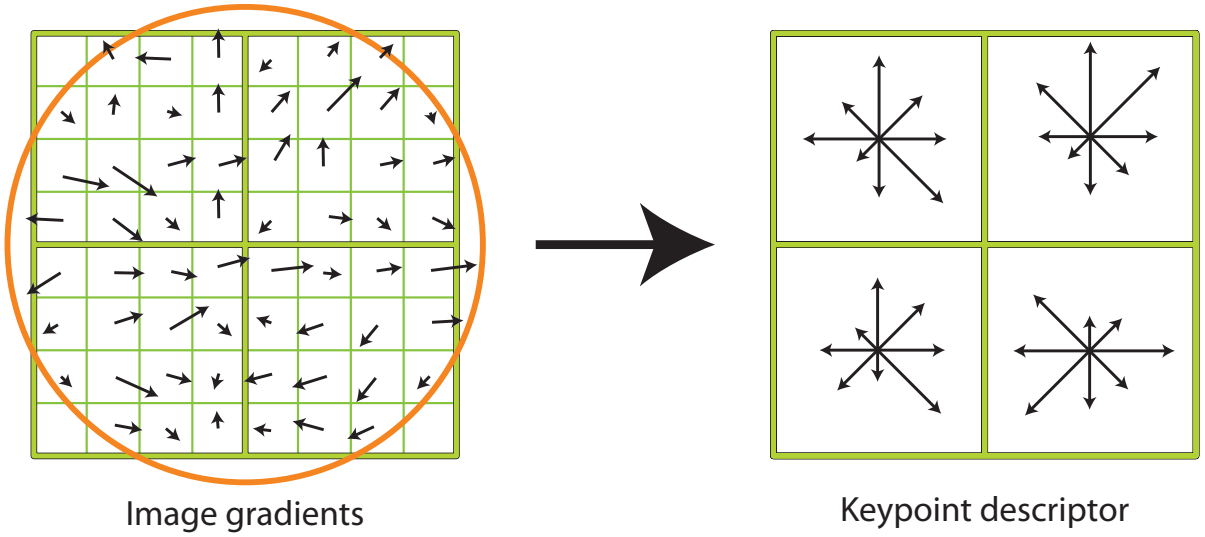


Figure 3.26: Construction of a SIFT descriptor, where for the ease of illustration, a 2×2 descriptor is shown. A sample with gradient orientations and magnitudes is shown (left), where the Gaussian window weighting the magnitudes is indicated by the orange circle. The gradients are then accumulated to an orientation histogram having bins (right). (Figure taken from [93])

bins each, are constructed, leading to a $4 \times 4 \times 8 = 128$ dimensional feature vector.

In order to quantize a gradient sample into adjacent histogram bins, trilinear interpolation is applied. Thus, abrupt changes in the descriptor with small position change or noise are avoided, as gradient samples are smoothly quantized between adjacent orientation bins [93].

Invariance to linear global brightness changes is achieved through the computation of pixel differences rather than absolute pixel values according to Equation 3.8, since they are invariant to a constant added to each pixel. Affine changes in illumination which result in change in the image contrast caused by a multiplicative operation with a constant are counterbalanced through vector normalization to unit length. Robustness with respect to non-linear illumination changes owing to shading variations on 3D surfaces or change in the camera saturation is based on the fact that they are more likely to affect gradient magnitudes than orientations. Hence, large gradient magnitudes are thresholded to reduce their influence, and the vector normalization is repeated. Thus, the emphasis of the descriptor lies in the distribution of gradient orientations rather than magnitudes.

3.3 Classification

Local features in object recognition are usually matched against each other in order to find a particular object. In this report, however, the task is not object recognition but rather object categorization, where parts of layout entities should be identified as belonging to a particular class. These parts have certain similarities rather than exactly corresponding to each other. Hence, a supervised machine learning algorithm is applied to predict a

feature's class based on the characteristics of known entities.

The classification problem in this report is a two class problem: features either belong to decorative entities or regular text. There is no need to distinguish further classes, since due to the interest point detection, no interest points are detected in background areas and all decorative entities are subsumed under one class according to their structural similarities.

3.3.1 Comparisons of Potential Classifiers

Mohanty et al. [106] compare SVM having a linear kernel, a Bayesian approach and Cross-Media Relevance Model (CMRM) for classification of 2D shape features. CMRM learns a joint probabilistic model from training samples. Test samples are then annotated with vectors of probabilities. In their experiments, SVM outperforms the Bayesian approach, however, the CMRM rise above SVM. It has to be pointed out that linear kernels are employed in this comparison.

SVM with a fifth-order polynomial kernel and NN are examined by Byvatov et al. [24] for a binary classification problem for drug/non-drug categorization. They state that SVM tend to be more robust and accurate when compared to NN.

Lee and To [84] conclude that SVM with a Gaussian Radial Basis Function (RBF) kernel performs better than Back-Propagation Network (BPN) on the evaluation of enterprise financial distress.

A comparison of several classifiers on a protein dataset is done by Bacardit et al. [7], where SVM with RBF kernel outperforms a Learning Classifier System, a rule induction system and a naive Bayes classifier. Kumar and Zhang [77] give a comparison between the following classification schemes on features from hand shape and palm texture for the task of people recognition: different naive Bayes classifiers (normal, estimated, multinomial), decision trees (C4.5, Logistic Model Tree (LMT)), k-NN, Feed-Forward Neural Network (FFN) and SVM with polynomial kernel. The SVM is shown to be more robust and accurate by their evaluation.

Justification for the Choice of the Classifier

Cristianini and Shawe-Taylor state that SVM are *"a principled and very powerful method that [...] [outperforms] most other systems in a wide variety of applications"* [35]. Wu et al. [142] and Joachims [65] stress that the SVM has a sound and solid theoretical foundation. A function which is well adapted to the training data does not may not generalize well to unknown test data [63, 122].

SVM are based on the *Structural Risk Minimization* principle, where a generalized model is to be fitted to prior known dataset such that guarantees the minimum true error [65, 122]. The true error is the possible error made on a randomly selected test sample not part of the known data. The SVM minimizes the overall risk, which results in a good generalization performance [65].

Golland et al. [56] argue that SVM are less likely to overfit the data and are more robust than other learning algorithms such as NN. SVM require a small training set and are independent of the number of dimensions of the feature space [65, 142].

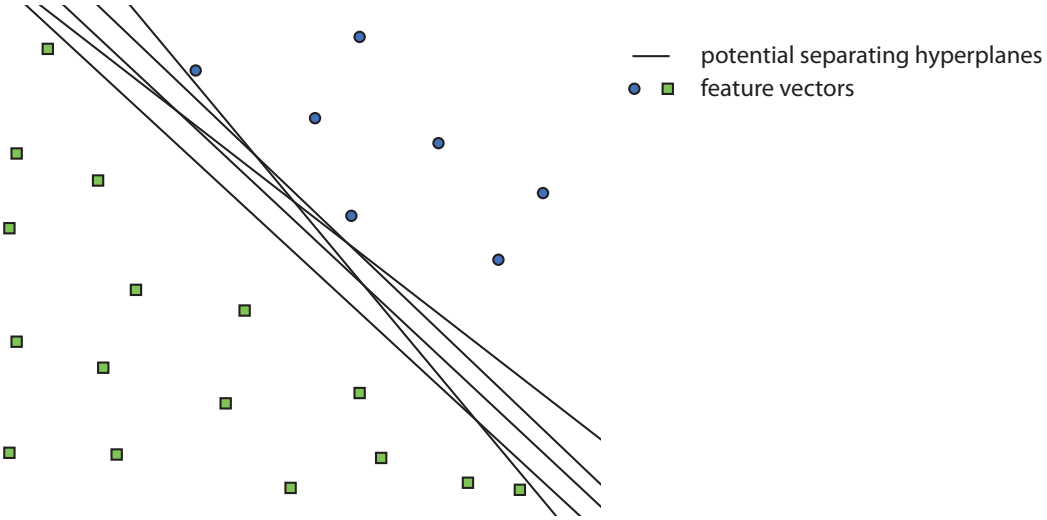


Figure 3.27: An infinite number of potential hyperplanes to separate two classes (green, blue).

3.3.2 Support Vector Machine

SVM is a supervised machine learning method introduced by Vapnik and his co-workers in [21], and further developed by Cortes and Vapnik in [33]. SVM is a binary linear classifier designed for two-class problems. For a given set of feature vectors $\vec{x_i}$ as training data with known class labels y_i: $(\vec{x_1}, y_1), \ldots, (\vec{x_n}, y_n), y_i \in \{-1, 1\}, i = 1, \ldots, n$, it generates a model $f(\vec{x})$ which predicts whether a test sample belongs to one or the other class, approximating the unknown function $f(\vec{x}) = y$ [63].

If the data is linearly separable, the classification function can be expressed geometrically as a hyperplane $f(\vec{x})$ separating the two classes. Having this model, a new data sample $\vec{x_i}$ is classified testing the sign of the function $f(\vec{x_i})$. However, for each given training set there is an infinite number of potential separating hyperplanes. In Figure 3.27, a few potential separating hyperplanes are indicated by black lines between the two classes (green, blue).

The SVM finds the optimal separating hyperplane by maximizing the margin between the hyperplane and the instances of the two classes, where the margin gives the shortest distance between the hyperplane and the class instances. The maximization of the margin leaves room for the correct classification of unknown test data and hence leads to a better generalization ability of the approach when compared to take a randomly selected hyperplane [130, 142].

The optimal separating hyperplane is uniquely determined by the so-called support vectors, which are the feature vectors on the margin of the hyperplane [130]. In order to find the maximum margin hyperplane, the following function is minimized with respect to the hyperplane

$$f(\vec{x}) = sign((\vec{w} \cdot x) + b) \tag{3.9}$$

defined by vectors $\vec{w}$ and constant b and maximized with respect to α:

$$L_p(\vec{w}, b, \alpha) = \frac{1}{2} \|\vec{w}\|^2 - \sum_{i=1}^{n} \alpha_i y_i (\vec{w} \cdot \vec{x_i} + b) - 1 \tag{3.10}$$

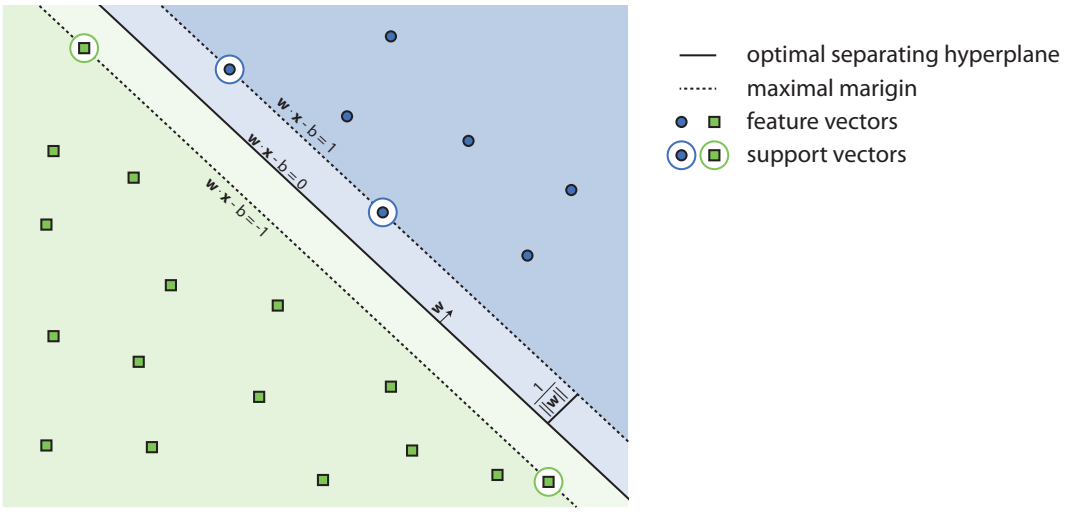


Figure 3.28: Linear separable classes divided by a hyperplane (solid black line) and the maximal margin (dotted line).

with n being the number of feature vectors, $i = 1, \ldots, n$ positive numbers, $\alpha_i \geq 0$ the Lagrange multipliers, which is $\alpha \geq 0$ for support vectors and $\alpha = 0$ for every other feature vector, and L_p the Lagrangian [142].

Figure 3.28 illustrates the optimal separating hyperplane for the classification problem. The optimal separating hyperplane is indicated by a solid black line, whereas the margins are limited by dotted black lines. The feature vectors are indicated by green squares and blue circles, and the feature vectors, the hyperplane computation is based on, are additionally surrounded by circles in the respective color.

Soft Margin Hyperplane

Cortes and Vapnik [33] suggest modifying the optimal separating hyperplane such that it allows for noise in the training set. Such noise in the training set are feature vectors being on the wrong side of the separating hyperplane. The soft margin idea extends the SVM such that it introduces a slack variable ξ_i allowing for noisy data, which attributes the possible amount of classification violation by the hyperplane. The geometric definition of ξ is the distance of a falsely classified feature vector to the hyperplane. The optimization of the hyperplane is then a tradeoff between the total cost introduced by the slack variables and a large margin [33, 142]. The Lagrangian is then the following:

$$L_p(\vec{w}, b, \alpha, \xi) = \frac{1}{2} \|\vec{w}\|^2 - C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i (\xi_i + y_i(\vec{w} \cdot \vec{x_i} + b) - 1) \tag{3.11}$$

where C denotes a constant parameter, $\xi \geq 0$ is the slack variable, and $\vec{x_i}$ is a feature vector.

Non-Linear Classification

For data which is not linearly separable, a mapping function is generated, which transfers the data into another space, the so-called *feature space* by means of the kernel trick first

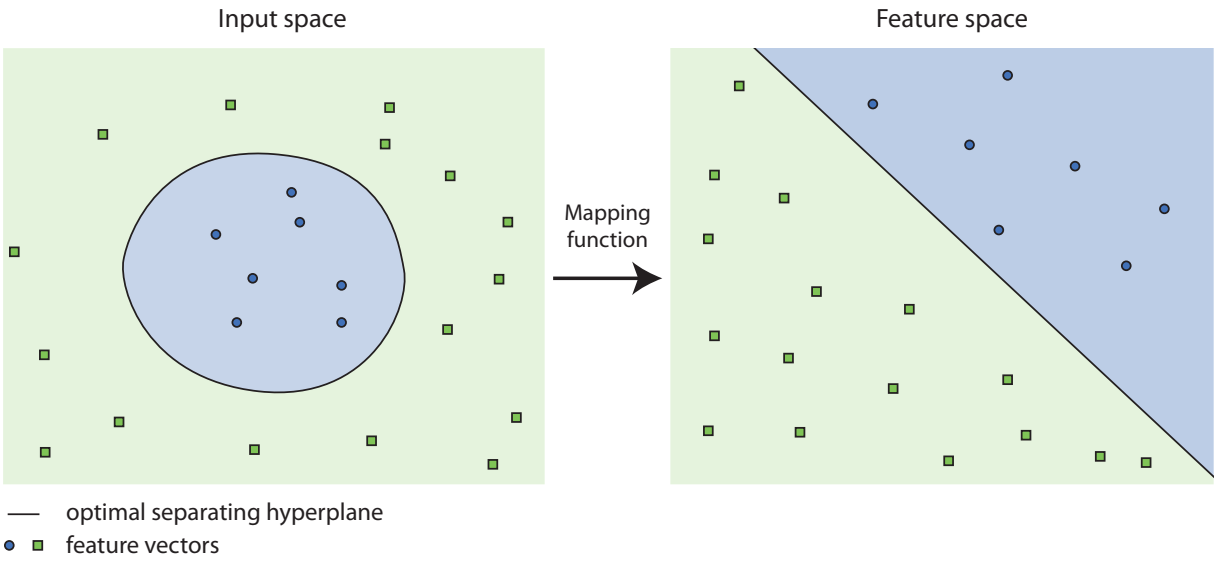


Figure 3.29: Kernel trick: mapping the input space onto a feature space through a function.

introduced by Aizerman et al. [3]. Hereby, the dot product is replaced by a non-linear kernel function and the hyperplane is transformed in to a high-dimensional space. This allows for a linear hyperplane in the feature space whereas the decision function in the input space may be non-linear with a form determined by the kernel [63,142]. Figure 3.29 illustrates the mapping from the input space to the feature space by means of the kernel trick. The original hyperplane definition in Equation 3.9 is extended to the following non-linear form [63]:

$$f(\vec{x}) = sign(\sum_{i=1}^{n} v_i \cdot k(\vec{x}, \vec{x_i}) + b) \tag{3.12}$$

The inner product can directly be computed as a function due to the kernel trick, and hence, the higher dimensional space does not need to be evaluated explicitly. The kernel function can define a variety of non-linear mappings between input- and feature space, such as a polynomial or exponential functions. In this report, the Gaussian Rbf kernel is used:

$$k(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2} \quad \gamma > 0 \tag{3.13}$$

where x_i and x_j are feature vectors and γ is a parameter.

The soft margin parameter ξ and the choice of the kernel determine the effectiveness of the Svm. The Rbf kernel has solely the parameter γ which needs to be selected such that it fits the training set, but is at the same time flexible enough to handle complex datasets.

Training

The parameter determination for ξ and γ is done in form of a cross validation, where the best combination of the parameters is determined by a grid-search for the tupel $\langle \xi, \gamma \rangle$.

Cross validation is a statistical method to estimate the accuracy of the model generated by the SVM by analysis how the model generalizes to unknown data [76].

Hereby, for a n-fold cross validation, the training set with known class labels is randomly partitioned in n subsets (folds) having approximately the same size. The learning algorithm is trained and tested n times, where the training set consists of $n-1$ folds and the remaining fold is used as test set [76]. The process is repeated for changing parameters ξ and γ, leading to a grid of classification accuracies for each tuple. The parameter combination having the best cross-validation accuracy are selected for the training of the SVM.

Even though an estimation of the accuracy based on the training set cannot be correct for any test set [76,119,141], the cross validation ensures that the RBF kernel is well-suited to the characteristics of the given classification problem [76].

3.4 Summary

This chapter gave an overview and theoretical background about the methods employed in this report. First, reasons are given for the choice of local features rather than global features. Then, requirements, the chosen features have to meet, are given. The methods employed as well as alternative approaches and related work were given. Interest point detectors were reviewed, which are needed to find regions where local features are extracted as descriptors. SIFT descriptors with interest points detected by the DoG interest point detector were chosen in this report. The last section gave reasons for the choice of a supervised learning algorithm as classifier, the SVM.

Chapter 4

Proposed Methodology

The majority of the state-of-the-art methods for layout analysis of historical documents have in common that a binarization step is required prior to the actual analysis and/or the layout is restricted to rectangular text blocks (see Chapter 2). Processing of printed historical documents from the hand-press period is more frequently addressed in literature than algorithms for ancient handwritten documents having unstructured layouts and suffering from degradation. However, ancient handwritten manuscripts cause different requirements to algorithms than printed historical documents [111].

The main dataset regarded in this report requires an algorithm to be robust with respect to background artefacts such as clutter, stains and noise, and faint ink. Hence, the proposed approach is designed binarization-free in order to be robust to these challenges. Since the considered dataset does not have a strict rectangular layout such as the documents considered in [23], a method invariant to layout inconsistencies, irregularities in script and writing style, skew, fluctuating text lines, and variable shapes of decorative entities is needed. 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 proposed was first published in [54] and further developed in [52]. It consists of two consecutive major tasks, 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 DoG described in Section 3.1.4. This interest point detector extracts blob-like regions on different scales employing a scale space.

Consecutively, a descriptor is calculated for each interest point describing an image region through gradient vectors build upon the gray-scale pixel values. The descriptor employed is Sift as detailed in Section 3.2.3 with adaptations as described in Section 4.1. This leads to local features describing parts of characters, or – depending on their scale – even whole characters or text lines.

The descriptors are then directly classified employing a kernel-based supervised machine learning algorithm. A Svm, as described in Section 3.3.2, is chosen as classifier to discriminate between two classes: main body text on the one hand and layout entities having a decorative meaning on the other hand. These decorative entities include embellished initials, plain initials and headings; they are grouped into one class as result of their

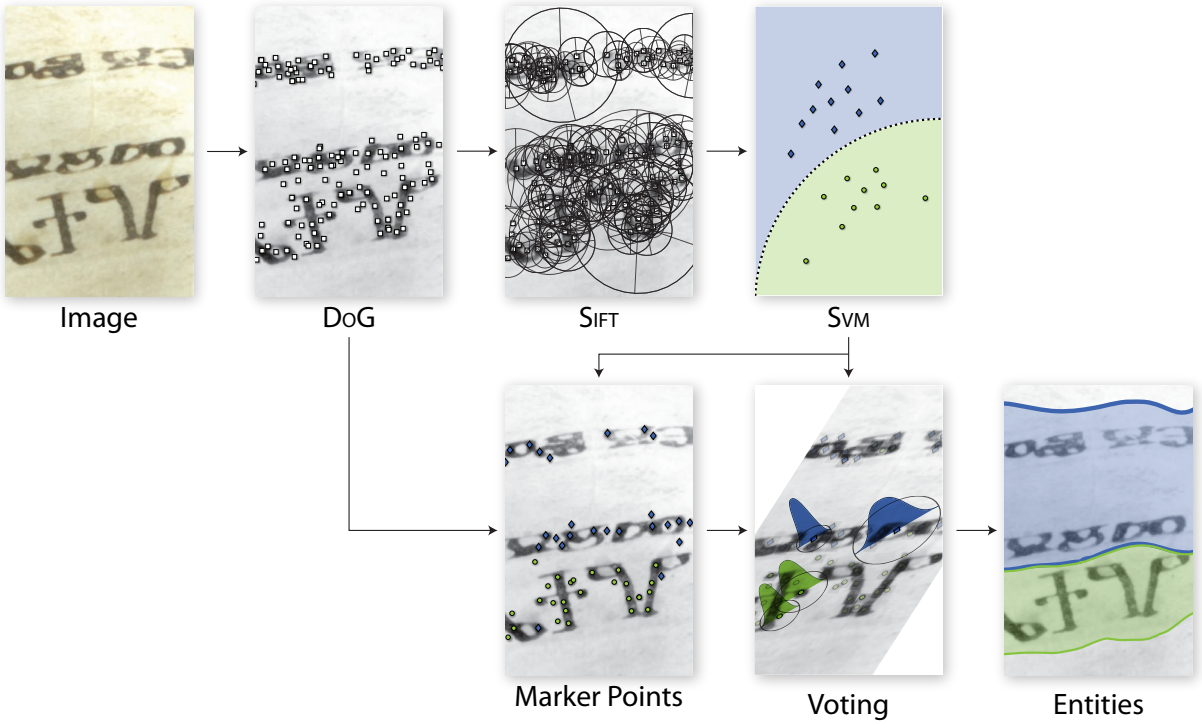


Figure 4.1: Workflow of the proposed layout analysis method showing the main tasks: feature extraction and classification (top) and localization (bottom)

local structure correspondence as explained earlier. Having classified the descriptors, one character or initial is described by multiple pre-classified points.

The next step is assigning a class label to each entity. However, the expansion of an entity cannot be directly inferred from the positions of the interest points and one entity might consist of interest points having different class labels. Hence, a localization algorithm to expand the interest points found into regions which enclose the whole layout entity and determining the class it belongs to, is required.

This chapter gives the details of the methodical steps for the layout analysis method introduced. Figure 4.1 illustrates the main tasks of approach, where the first row shows the feature extraction (described in Section 4.1) and classification steps (see Section 4.2) and the second row illustrates the localization (described in Section 4.3). As can be seen, both tasks depend on the interest points extracted by means of DoG. As a pre-processing step, the image is transformed to a gray-scale image and is normalized. Interest points are detected in the grayscale image (Figure 4.1 DoG) and the respective regions is described with Sift descriptors (Figure 4.1 Sift). The descriptors are then classified with a supervised learning algorithm (Figure 4.1 Svm). The interest points extracted and their respective classification scores are then used to establish marker points (Figure 4.1 Marker Points), which are interest points that are reliable due to their scale and classification score. A set of voting functions is applied to the interest points (Figure 4.1 Voting), and a score map is established that determines the class label on pixel level and localizes the layout entities (Figure 4.1 Entities).

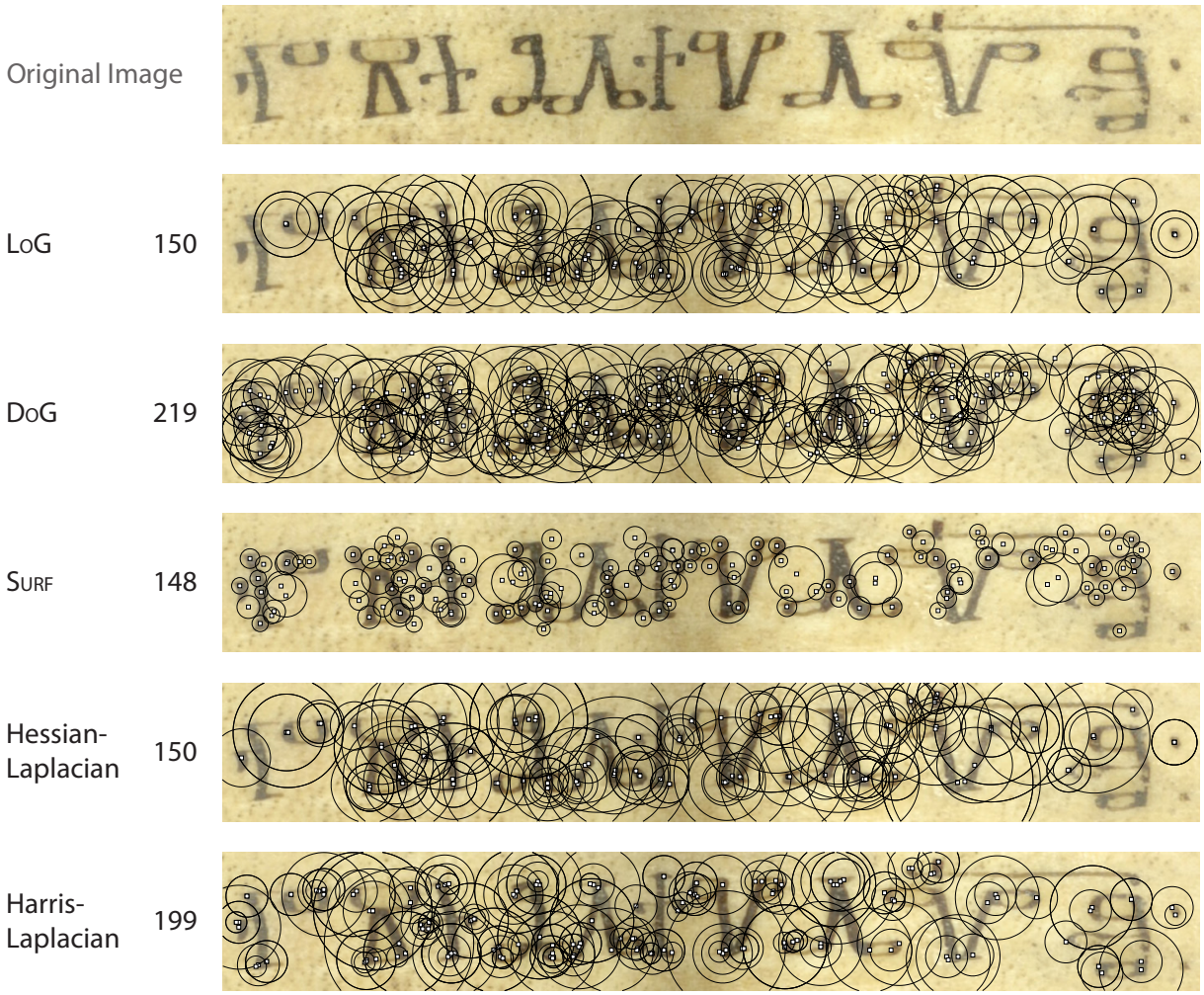


Figure 4.2: Comparison of five potential scale-invariant interest point detectors (four blob detectors: LoG, DoG, Surf, Hessian-Laplacian, one corner detector: Harris-Laplacian) on a heading of the *Psalter* dataset. The numbers in column *Points* denotes the number of Interest Points extracted with the respective interest point detector.

4.1 Feature Extraction

Though interest points can be gained applying a scale-invariant corner detection method such as the Harris corner detector combined with a Laplacian as suggested by Mikolajczyk et al. [99] or Fast [118], a corner and edge detector based on non-linear filtering, a blob-based detector is chosen on account of studies by [93,101,102]. Diem [37] performed a comparison of interest points and local features on Glagolitic manuscripts and concludes that Sift with the DoG detector is the optimal solution for these manuscripts.

Figure 4.2 gives an exemplified comparison of five scale-invariant interest point detec-

tors described in Section 3.1: the LoG implemented in *lip-VIREO*[1], the DoG as applied in the approach, the Surf as implemented in [42], the Hessian-Laplacian detector and the Harris-Laplacian implemented in *lip-VIREO*[1] with the respective number of detected interest points. The images are overlaid with visual representations of interest points. The white squares denote the locations of the interest points and the circles indicate their scale. Figure 4.2 illustrates that consistent to the conclusions of [37, 93, 101, 102], the DoG detector leads to an advantageous coverage of the letters when compared to the other detectors and, furthermore, generates a higher number of interest points.

Additionally, Sift descriptors with DoG interest points are chosen as feature system, since it is invariant to scale and rotation, which is an important aspect for ancient manuscripts, as the script size and orientation may change. A further advantage of rotation invariance especially concerns the embellishments in the *Psalter* dataset, as these entities consist of long strokes having varying orientations. An example is given in Figure 4.2 (Original Image). Since the aim is not the discrimination between characters, but the differentiation between scripts, the long strokes provide a discriminative characterization for the embellished entities regardless of their orientation.

Furthermore, Sift is 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. Thus, characters having faint ink – down to a certain contrast difference to the background, which is determined by the threshold during the computation of the interest points with the DoG – can be detected. The invariance to the 3D camera viewpoint that Sift incorporates, allows detecting the same character (on the scale of characters) or structural elements (at a smaller scale) despite deformations owing to unevenness of the writing support or variations in the script.

4.1.1 Interest Points

The DoG detects interest points at locations of local minima and maxima exploiting a scale space. The scale space is established by successively differencing the image convolved with a Gaussian function having an increasing scale parameter σ. The scale of the interest points is then estimated by detecting the local extrema over the image space and the scale space (a detailed description is given in Section 3.1.4).

The interest points extracted with this detector represent discriminative character parts such as junctions, corners, circles, arcs or endings as well as whole characters and parts of text lines. In Figure 4.3, the first representations of the DoG scale space are given for three octaves of an image. The green squares indicate locations of interest points detected in the respective image representation. In the first octave, small details are extracted, whereas higher octaves – i.e. successively larger Gaussian kernels for smoothing – lead to simplifications of the structures and larger blob-like regions. As described in Section 3.1.4, each octave consists of a set of image representations corresponding to an increasing Gaussian kernel size σ. Each structure responds to a certain octave and image representation inside this octave according to its scale. In Figure 4.3, solely the

[1]Video Retrieval Group (VIREO) at City University of Hong Kong, http://vireo.cs.cityu.edu.hk/research/project/lip-vireo.htm, accessed April 2011.

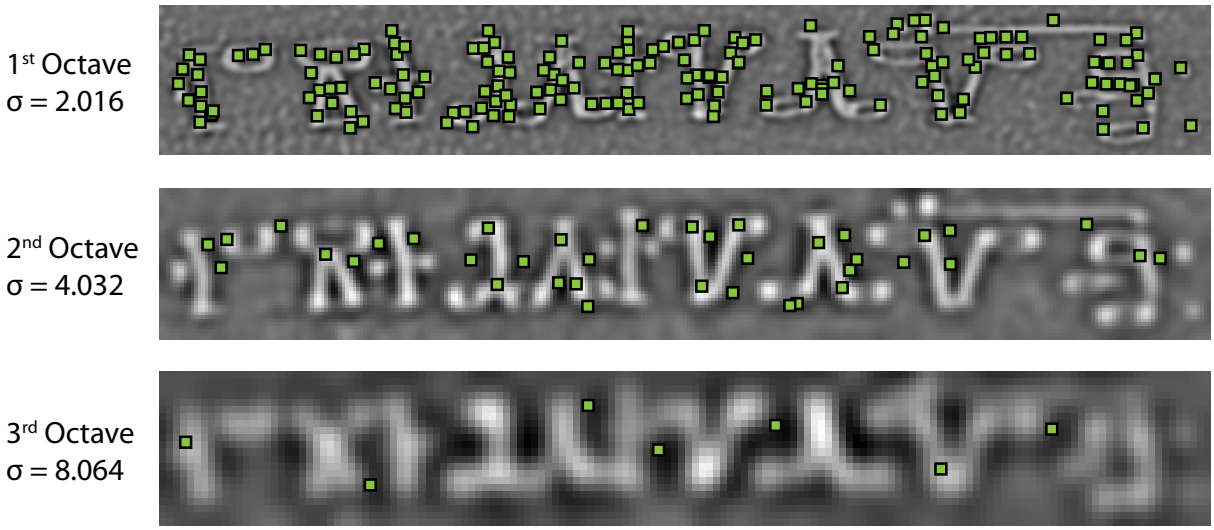


Figure 4.3: The first representations of each octave with σ given are overlaid with interest points detected in the DoG scale space which denoted by green squares.

first representations of each octave are shown and thus, not all image structures have an interest point assigned in this figure.

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

4.1.2 Local Descriptors

For each interest point detected, local features are computed using the SIFT descriptors as described in Section 3.2.3. Here, a brief summary is given. For a descriptor, 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 gradients are weighted by a Gaussian window. The region is divided into 4×4 subregions, and for each subregion, a histogram of gradients is calculated. For each interest point, a characteristic orientation is calculated and the respective coordinates and gradient orientations are rotated relatively to this main orientation in order to achieve rotation invariance. The SIFT descriptors are 128-dimensional feature vectors containing the values of the 4×4 orientation histograms having 8 bins each. Each of the bins relates to a specific main orientation $(0\,°, 45\,°, 90\,°, \ldots, 315\,°)$.

4.1.3 Modifications of the Feature System proposed by Lowe

The original local features consisting of DoG as interest point detector and SIFT as descriptor as proposed by Lowe [93] have to be modified in order to fit the requirements of the dataset and the application. The threshold for the rejection of local extrema (see Section 3.1.4, Interest Point Localization) is set to 0.02 in order to reduce the number of

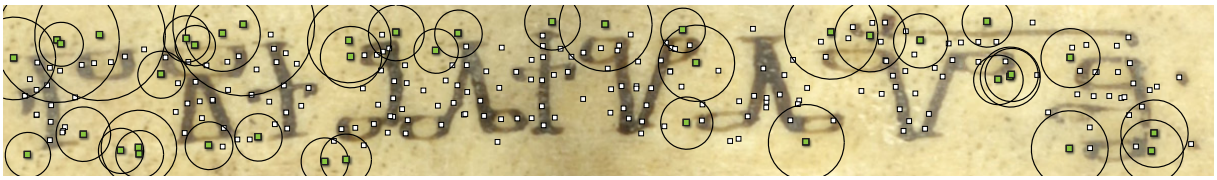


Figure 4.4: Interest points detected by the DoG detector. White squares indicate interest points with the threshold applied in this report, green squares are interest point which are additionally detected when the threshold is lowered.

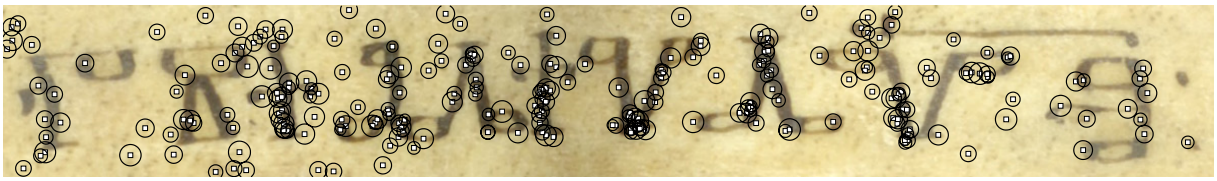


Figure 4.5: Interest points detected by the DoG detector with the first octave being the subsampled image.

interest points which represent background clutter; the threshold suggested by Lowe [93] is 0.01. In Figure 4.4, interest points extracted by the DoG detector are shown, where white squares indicate interest points extracted with the threshold 0.02 and green squares denote the interest points additionally found at threshold 0.01. As can be seen, local extrema having a larger distance to the characters are found and – especially in the lower left region – background noise is detected as local extrema.

Lowe [93] suggests subsampling the original image to double size for the first octave, resulting in a higher number of interest points. These interest points represent the highest spatial frequencies and thus, the smallest details of the image. Subsampling the image leads to interest points having a small spatial extend. As described in Section 3.1, the smaller the spatial extend, the less structural information is contained, and thus, leads to unreliable descriptors in the classification step. Additionally, the structural similarities of the layout entities do not allow for a distinction between the entities on this level. Finally, the increase of the image size enhances the background noise, the noise in the ink of the characters, and clutter in the test set. Figure 4.5 illustrates this circumstance. Interest points are found in noisy regions of the character shapes and the background. The region they describe is small and thus, does not contain much structural information. Additionally, interest points located at edges of the character, lead to a similar normalized feature vector since a larger interest point describes single strokes of an embellished initial. Thus, these interest points found in the main body text deteriorate the classification result.

In contrast to the aim of object recognition, interest points located at edges are necessary to localize the layout entities in the approach proposed. By default, the DoG is sensitive to edges, contrary to the LoG detector. Figure 4.6 gives an example detection result for both detectors, where the row indicates the number of interest points. As can be seen, interest points located at edges are necessary for the coverage of the character.

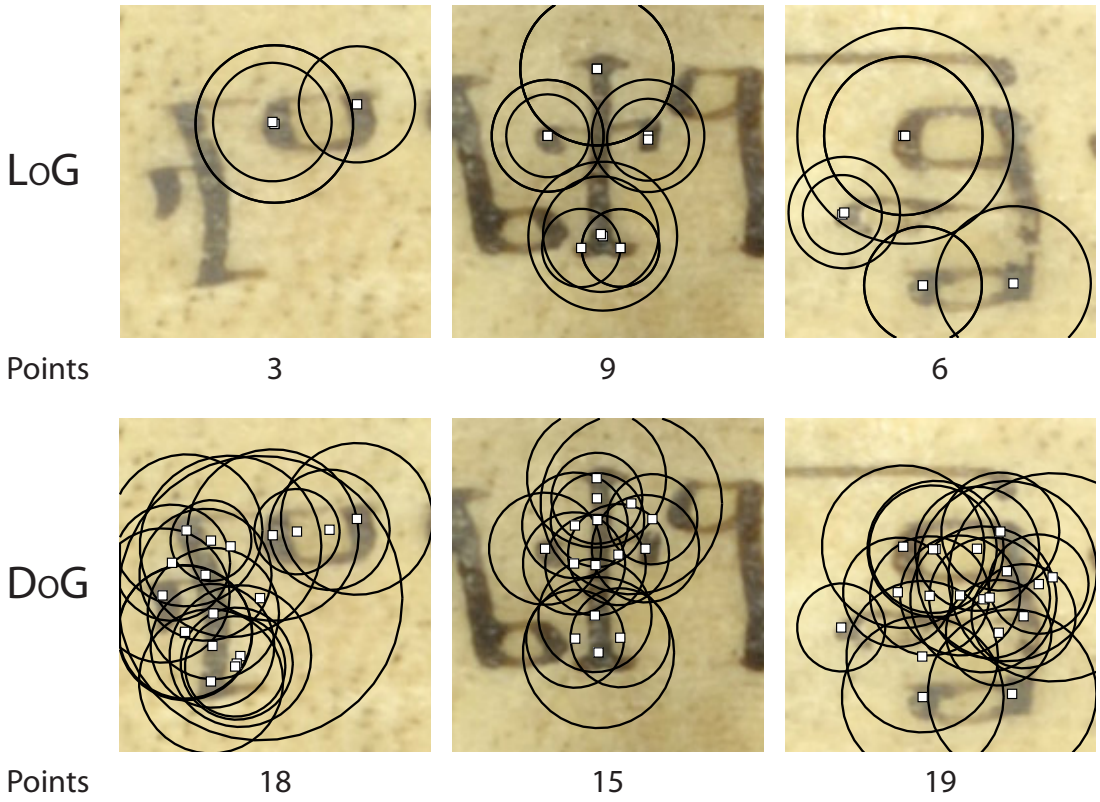


Figure 4.6: First row: Interest points detected by the LoG detector, which is not sensitive to edges. Second row: Interest points detected by the DoG detector with the configuration used for this report. The row *points* indicates the respective number of interest points.

Thus, the DoG's unwanted property of producing interest points located at edges is exploited since poor localization along the edge is not crucial for the task of layout entity localization. The embellished initials of the *Psalter* consist of long single or outlined strokes. Thus, interest points located at edges are not excluded as they contain necessary information for the classification and localization.

4.2 Classification

Having computed the Sift descriptors, the local regions around each interest point extracted of the image is mathematically characterized by a high-dimensional feature vector. As described in Section 3.1.4, interest points are only extracted at locations of local extrema, i.e. background areas without textual elements are not included. Thus, solely regions of interest have to be further processed.

As detailed in the Section 1.3.1, in case of the *Psalter* dataset, the class of decorative entities are characterized by angular shapes, elongated strokes and an aspect ratio that is vertically or horizontally stretched when compared to the main body text. Outlines and hatches are further characteristics of initials appropriate to exploit for the discrimination between main body text and embellishments (see Section 1.3.1 for more details).

Thus, two logical classes are defined: main body text and decorative entities. This means that headings and both types of initials belong to the same class due to the fact that their local structures have similar characteristics.

In order to discriminate between the feature vectors describing main body text and entities having a decorative meaning, a SVM as depicted in Section 3.3.2 is trained. A RBF kernel as formalized in Equation 3.13 is chosen in order to be able to separate non-linear data. The classifier assigns a class label to each descriptor as well as a score indicating the probability of the class assignment. These scores are used in the subsequent localization step.

Training

The SVM is trained using image patches containing one entity in case of initials, a whole heading or a certain amount of text lines. The number of entities are different for the particular datasets and will be given in Chapter 5.

The image patches are rectangular images containing the respective characters and background; they were manually extracted from the manuscripts and. Each image patch is labeled according to the class it belongs to.

Employing the RBF kernel, two parameters have to be determined for a dataset:

- γ: Determines the sensitivity of the kernel, which means it needs to be selected such that it fits the training set, but at the same time is flexible enough to handle complex datasets,
- ξ: The soft margin parameter, a slack variable which controls the flexibility of the data, which prevents overfitting the data and ensures the generalization ability of the model learned.

These parameters are determined by means of cross validation, which explores the best combination of ξ and γ applying a grid-search. The parameter tuple for the *Psalter* training set is $\langle \xi, \gamma \rangle = \langle 5, 0.5 \rangle$, determined by three-fold cross validation, where the training set is split into 3 folds. Further splitting into 5 or 7 folds increases the cross validation accuracy, since the size of the training set is increased. However, the relative overall accuracy does not change resulting in similar parameters.

In Figure 4.7, a sample classification result is given. Correctly classified interest points are depicted with white circles, while interest points having wrong class assignments are indicated with red squares. The ground truth is denoted by gray blobs, where dark blobs stand for the decorative entities and light blobs are areas of main body text.

4.3 Layout Entity Localization

Having classified the feature vectors, each descriptor has a class label assigned, where the class decision for each interest point is based upon the maximum score retrieved by the SVM, which means that the class $c \in \{\textit{decorative entity}, \textit{main body text}\}$ is determined by $max(score_{c1}, score_{c2})$.

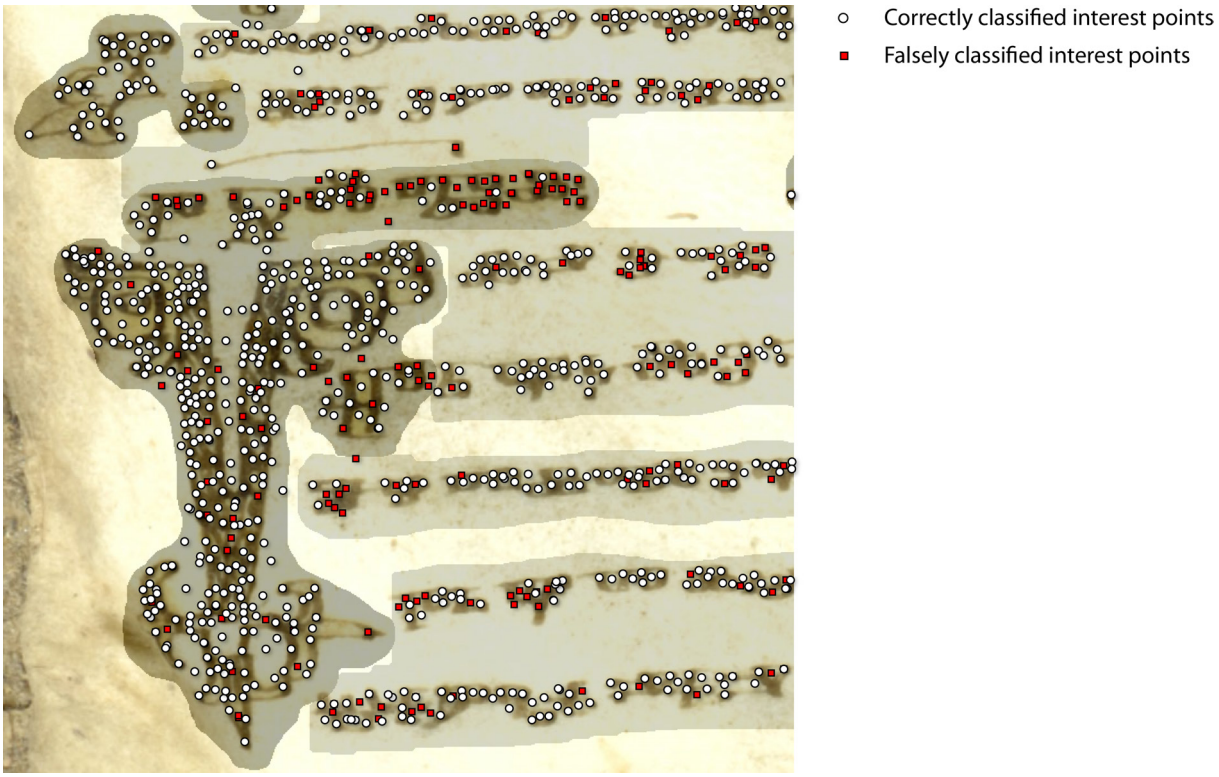


Figure 4.7: Sample classification result. White circles denote correctly classified interest points, red squares mis-classified ones.

A localization algorithm needs to be applied to the interest points in order to find regions encapsulating whole layout entities, since the classification of interest points leads to class decisions just at certain positions in the image. The location and expansion of entire entities are then deduced from the positions and spatial extends of interest points selected during the localization algorithm.

The scales and locations of interest points are exploited for the localization step. The assumption for this procedure is that an interest point represents an entity segment or even a whole entity. Thus, the scale of the interest point relates to the size of the structure it describes, e.g. an entity part. Pursuant to this assumption, a cascade localization algorithm is introduced which successively reduces the amount of mis-classified descriptors and leads to a class decision on pixel level.

Descriptors which are not unambiguously belonging to one single class might randomly be assigned to either of the classes. Such descriptors are likely to occur since not all of the structures the entities are assembled of, are unique for one of the classes. An example are rounded character segments, which are one of the discriminative characteristics of main body text, but occur in headings and initials too. 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. Straight stroke segments of characters belonging to the main body text observed at a small scale produce similar feature vectors like such segments of initials or headings at a larger scale, for example.

The suggested 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 mentioned previously 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 scores 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.

Figure 4.8 gives an overview of the stages of the localization algorithm for the case of decorative entities. Interest points are illustrated with green circles denoting their respective scale. The more intense the green color, the more interest points overlap. Figure 4.8 a)-g) show the decorative entities-class, f) relates to the main body text class. In detail, Figure 4.8

a) illustrates all decorative-entity descriptors classified by the SVM,
b) shows the marker points as selected from the second octave,
c) shows the marker points merged with overlapping interest points,
d) depicts the interest points after removing single occurrences,
e) illustrates the spatial weighting step, resulting in a score map,
f) shows the score map, with ISO-lines indicating the scores,
g) presents the final result of the localization algorithm, and
h) gives the final result for the main body text class.

In Figure 4.8 g, h), 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.

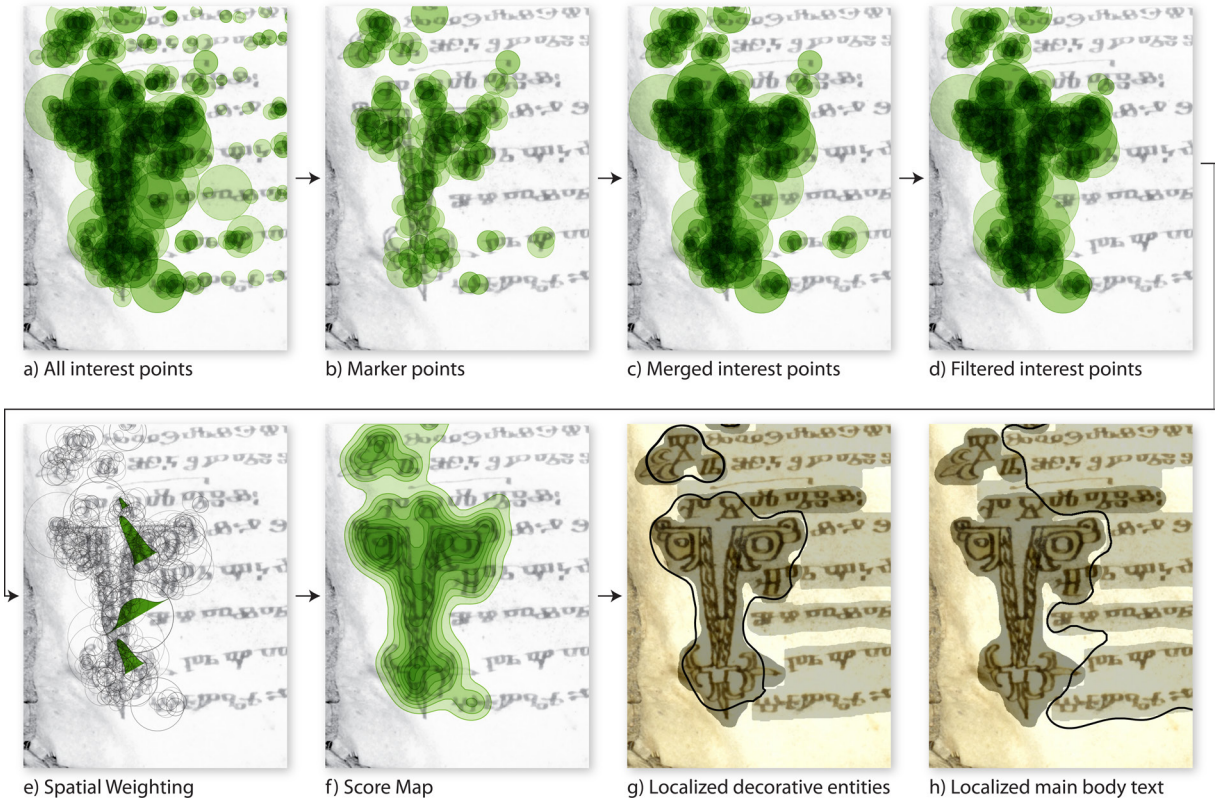


Figure 4.8: Overview of the cascading localization algorithm (*Folio 85r*)

4.3.1 Scale-Based Voting

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. An example is given in Figure 4.9 (right), where interest points having a smaller scale than 9 – which corresponds to $54\,px$ – are indicated with black circles. The interest points in the left part of the image patch depict character segments – with four exceptions –, whereas the interest points in the right half are all representing background clutter.

Large scales represent e.g. whole decorative initials, spots and stains as well as ripples of the parchment. Figure 4.9 (left) gives an example of interest points of the largest scales. These either indicate text lines (right upper and lower part of the patch) or follow the crease of the writing support (center of the patch) where they describe the fold rather than the text lines.

Owing to the distribution of the interest points' classification scores when interrelated with their scales, a linear weighting function is chosen. This weighting function implements the principle of a band-pass filter, i.e. it emphasizes a certain range of scales and lowers scales outside this range. The weight of the function is applied on the classification

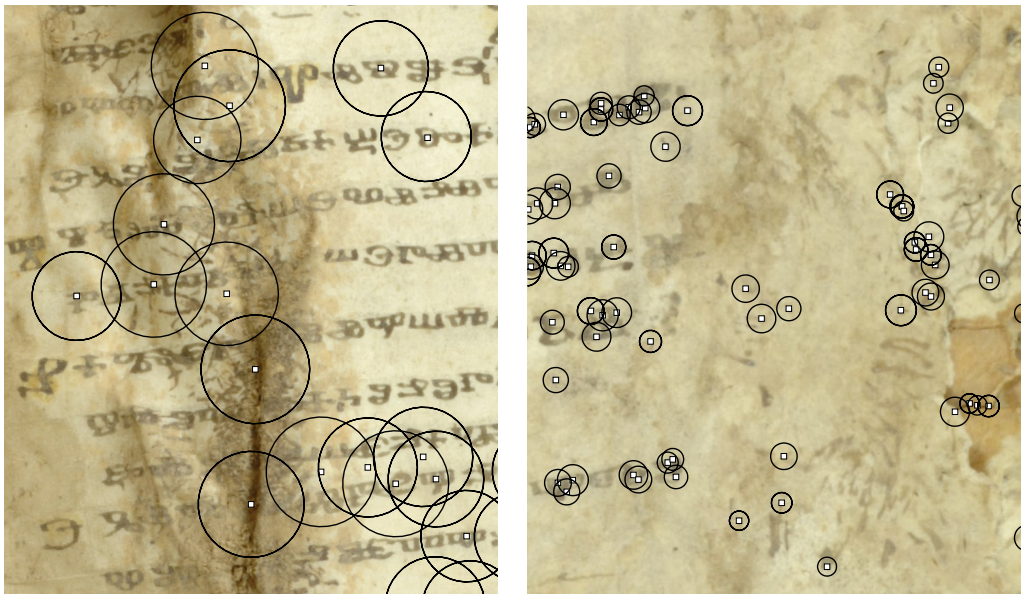


Figure 4.9: Interest points having a scale larger than $54\,px$ (left), respectively smaller than $18\,px$ (right). (The resolution of the image considered is $2848 \times 3963\,px$)

score gained from the SVM for each descriptor.

Since the mean reliability of interest points having different scales is different, the function has dissimilar slopes for small and large interest points. The inflection points of the function are determined based on scales of the second octave. Voting functions for main body text and decorative entities have to be different owing to their respective different structure scales and classification performances per scale. This means e.g. that due to their characteristics and size, larger interest points are permitted for decorative entities than for main body text.

The scales rated with the normalized weight 1 are those which represent whole characters of main body text, plain initials, headings or major components of a character such as an arc or circle, or parts of decorative initials. The scales smaller than these indicate parts of characters which are smaller than major structures. Interest points smaller than character segments are weighted with 0 or a value close to 0 as they describe clutter, and noise – of both, the background and the character shapes –, dots, and speckles of the parchment. Large scales representing whole decorative initials, spots and stains, ripples of the writing support or intra-text-line spaces respectively whole text lines are weighted with a value close to 0 too.

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.

Thus, these large and small interest points are not taken into consideration. Furthermore, interest points falling short of the scale of character segments, are randomly assigned to a class since the information content they incorporate is not sufficient to

uniquely describe one of the classes. Thus, e.g. small interest points belonging to the decorative entity class occur frequently in text areas too and vice versa. Thus, these interest points have to be assigned a weight such that the number of mis-detected interest points is reduced.

4.3.2 Marker Points

In this report, 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. Following the arguments in the previous section, a certain scale of interest points is more 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. As detailed in the previous section, scales smaller than a segment of a character such as a junction or arc, usually represent (background) clutter and noise. Scales larger than an average character of a heading or a plain initial are likely to represent stains and creases of the writing support or intra-text-line space.

Intuitively, interest points covering an entire character of one of the classes or a major character structure such as a circle, arc or junction, are most reliable to indicate the location and spatial extend of a character. The classification score as single metric for the determination of interest points is not applicable, since interest points having large scales and representing intra-text-line space of entire text lines are reliable in terms of classification score, however, they do not reliably determine a character. Thus, the scale-based voted classification scores are applied in the determination of the marker points.

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.

Aside from scale, further properties of an interest point are its main orientation (see Section 3.2.3) and the extrema value. However, a significant trend cannot be extracted from the coherence of the prevailing orientations of interest points and their classification scores. Furthermore, relying on the main orientations of the interest points leads to problems in presence of skew and rotated document pages; additionally, strokes in handwritten documents may have arbitrary orientations.

Figure 4.10 shows how the extrema values of the interest points are correlated to their classification performance. As can be seen, maxima are generally more reliable than minima. This can be traced to the fact that maxima are located on the characters while minima indicate spaces between characters or the inside for circular structures, for example. Since minima produce interest points having a large range of reliabilities, the selection of marker points is not based on this property.

Having determined the candidates for marker points by their scales, a set of filtering operations is applied in order to reject weak points.

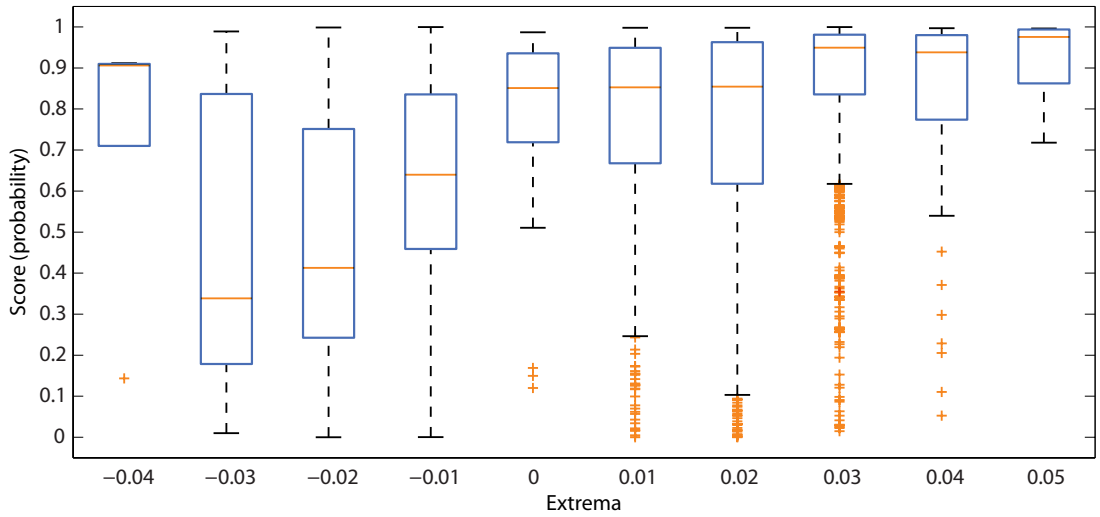


Figure 4.10: Extrema values of the interest points plotted versus their classification score

First, only those interest points having a classification score superior to 0.6 are taken into account. This threshold is chosen because the counter score is equal or less than 0.4, which means that the difference between the two scores is at least 0.2. Thus, the probability that a class label is assigned to a descriptor on random basis is reduced. This filtering step is done to reject non-reliable interest points where the class decision is ambiguous.

Subsequently, it is defined that at least two marker points must vote for an area to be considered as reliable. Thus, only marker points having a spatial overlap of their scales of at least 25 % are taken into consideration. The decision for a marker point M is based on the equation given:

$$\begin{aligned} d(x,y) &= \sqrt{\sum_{i=1}^{2}(x_i - y_i)^2} \\ r(x,y) &= \sum_{i=1}^{2} r_i \cdot 0.75 \\ M &= \begin{cases} 0 & \text{if } d(x,y) > r(x,y) \\ 1 & \text{if } d(x,y) < r(x,y) \end{cases} \end{aligned} \tag{4.1}$$

where $d(x,y)$ gives the Euclidean distance between two marker points and x_i and y_i are their respective coordinates. If the distance is smaller than the sum $r(x,y)$ of their radii r_i, both marker points are maintained, otherwise, the marker points are rejected. Applying this, isolated marker points voting for a class are not considered. The resulting marker points for an image patch of *Folio 85r* are shown in Figure 4.8 b).

4.3.3 Merging Remaining Interest Points with Marker Points

The next step is to merge the marker points M with all remaining retrieved interest points $I \backslash M$ because the localization solely based on marker points generates a sparse localization

result since interest points of a certain scale range just represent structures of these scales. Thus, structures having different scales are not completely covered with marker points.

Hence, all interest points overlapping with a marker point at least 25 % according to the computation given in Equation 4.2 are included in the set of interest points used for the following stages. More precisely:

$$Ic = \{I \backslash M \ \cup \ M \ \mid \quad d(x,y) < r(x,y) \wedge \langle x_1, y_1 \rangle \in M \wedge \langle x_2, y_2 \rangle \in I \backslash M\} \tag{4.2}$$

with Ic being the set of interest point candidates for further processing, I denoting the initial set of all interest points detected, M depicting the marker points. In Figure 4.8 c), the interest points merged are shown for the image patch of *Folio 85r*.

4.3.4 Region-Based Filtering

Having the whole set of potential interest points for the localization consisting of the marker points and the interest points overlapping with them, further filtering procedures are applied. A region-based processing is used, 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.

A threshold of 10 – giving the minimum number of interest points inside a region – is chosen in order to exclude sparse regions. While a higher threshold would reject layout entity candidates which are correctly identified, a smaller threshold would introduce weak candidates. Hence, interest points inside regions having a smaller number of points voting for it, are rejected. The assumption for this step is that a character should consist of enough structures having different scales to produce a significant number of local extrema.

The second step is based on the size of the region. Interest points inside regions smaller than an average character of the document – which is assumed to approximately correspond to the median marker point scale – are rejected. This operation is performed employing morphological opening with a circular element. A median scale is chosen as threshold owing to the fact that interest points having small scales are likely to describe clutter and noise. Thus, regions solely consisting of small interest points in a local area are not included in the set of interest point.

Rejected interest points are temporarily stored and subsequently added to the set of interest points of the other class since it is assumed that parts of characters are wrongly classified and thus, belong to the other class. The region-based stage of the localization algorithm is repeated with the new set of interest points.

The result of this region-based filtering step on the interest point candidates of image patch of *Folio 85r* is shown in Figure 4.8 d).

4.3.5 Spatial Weighting

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 $G(x, y, \sigma)$ with a σ according to the scale of the interest point.

$$map_c = \sum_{i \in c} \omega_i \cdot G\left(x_i, y_i, \sigma_i\right) \tag{4.3}$$
$$c \in \{decorative\ entity,\ main\ body\ text\}$$

where map_c denotes the score map for class c, j gives the number of interest points belonging to class c, σ_i corresponds to the radius of the i^{th} interest point in pixels and ω_i is the descriptor's classification score. The two-dimensional $G(x, y, \sigma)$ is defined by:

$$\begin{aligned} G(x, y, \sigma) &= e^{-\frac{x^2+y^2}{2\sigma_i^2}} \\ x &= \{a \mid x_i - 3 \cdot \sigma \leq a \leq x_i + 3 \cdot \sigma\} \\ y &= \{b \mid y_i - 3 \cdot \sigma \leq b \leq y_i + 3 \cdot \sigma\} \end{aligned} \tag{4.4}$$

with x_i, y_i being the coordinates of the local descriptor. The term $6 \cdot \sigma$ represents approximately 100 % of the $G(x, y, \sigma)$ distribution.

Hence, at all locations of interest points, a weighted score distribution having the same spatial extend as the interest point is generated. This step is illustrated in Figure 4.8 e). 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.

4.3.6 Post-processing

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. Background pixels have the label 0. The score maps for the image patch of *Folio 85r* are given in Figure 4.11 a) for decorative entities in green, and in Figure 4.11 b) for the main body text in blue. The more intense the green or blue color, the more interest points vote for a pixel. As can be seen, the scores overlap at several positions. The heading above the embellished initial is covered from both maps. Thus, a post-processing step is needed which includes the voting of both score maps against each other.

First, the score maps are voted against each other with the maximum probability determining the pixels class label. A sample result of such a voting is shown in Figure 4.11 c).

In a second step, the score maps are normalized and a filtering step based on a threshold t is applied in order to exclude regions of non-overlapping boundary areas of single interest points as these are likely to cover background areas. For an example, refer to Figure 4.12 a), where single interest points describing the embellished initial at a large scale, cover background area as well (center left). Furthermore, distinct decorative entities may be connected by an area of low scores if they are close enough that several of

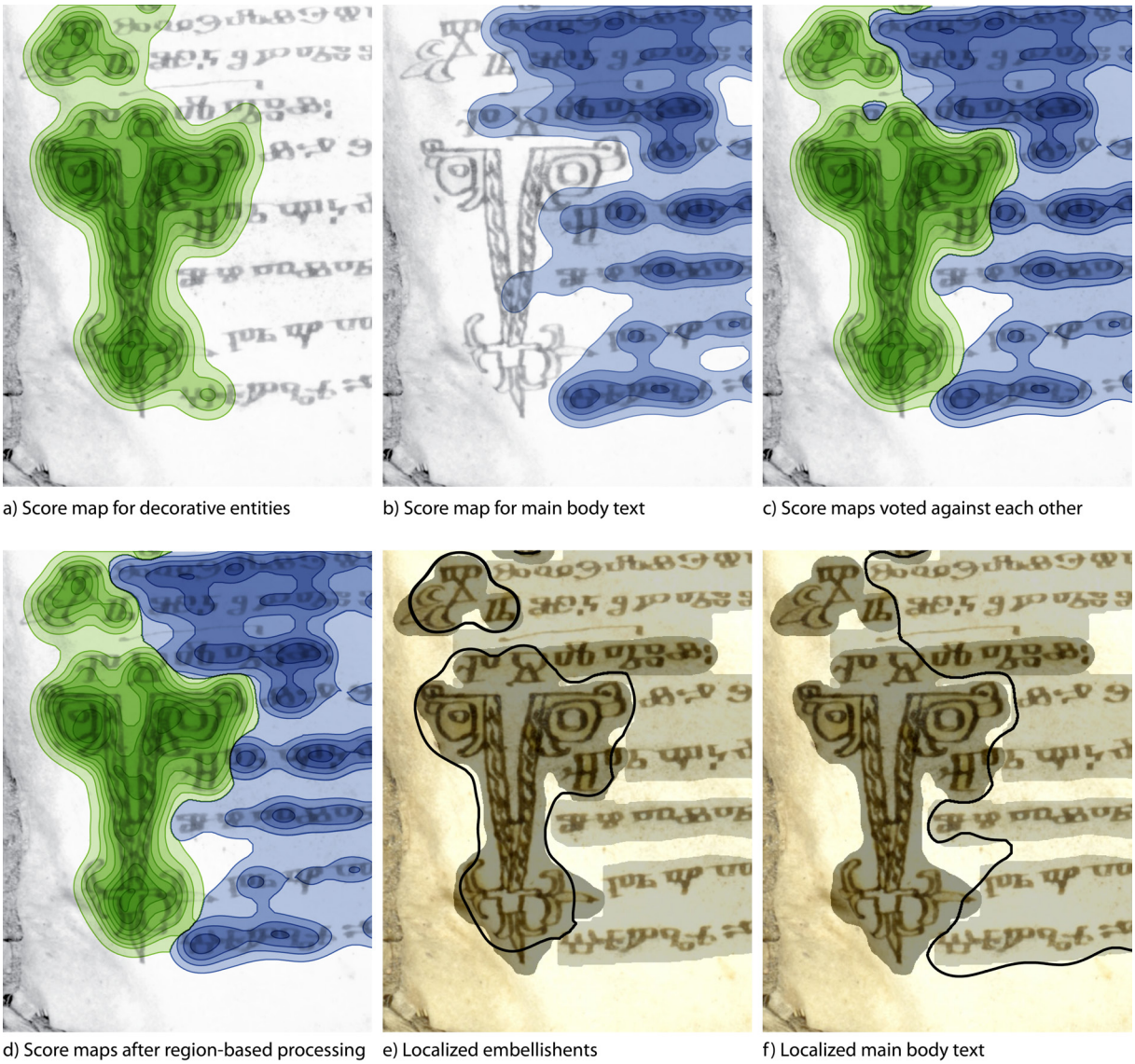
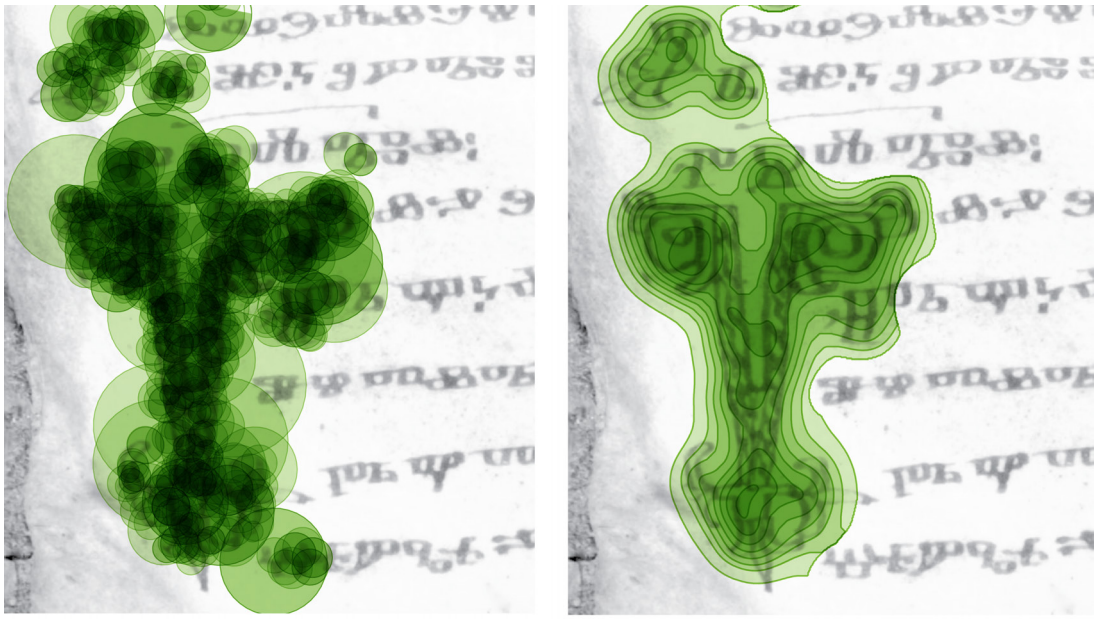


Figure 4.11: Post-processing: both score maps a) and b) are voted against each other resulting in c), a filtering approach rejecting small regions is applied leading to d). e) and f) illustrate the localized entities.

their large-scale interest points overlap. Figure 4.12 b) gives an example for a plain initial connected loosely to the heading below. The heading and the embellished initial are so close to each other that a separation based on the scores is impossible.

A pixel-based voting may lead to isolated small areas as illustrated in Figure 4.11 c) in the upper-left corner. The circular segment of the Glagolitic character ⰐⰏ in the heading above the embellished initial is assigned to the main body text class. Such isolated misclassifications occur if e.g. a character segment is characteristic for one class – and thus, produces enough interest points having a significant accumulated classification score for this character part –, while the larger structure or whole character is typical for the other class.

Hence, a region-based filtering is applied in order to reject these isolated areas not



a) Scales of single interest points overlapping with the background

b) Decorative entities are connected by areas of low scores

Figure 4.12: Need for thresholding the score maps.

large enough to be a valid character. Morphological opening is employed followed by a determination of the area of the regions with those being rejected which have a lower number of pixels than 1.5 times the area of a median marker point. The increased size of the median marker point is used since characters of the main body text appear as a member text regions rather than single characters. Plain initials though, have single occurrences and are larger than the size of a median marker point. Thus, the increased minimum size of a character prevents plain initials from being rejected. For the rejected areas it is checked if there are interest points of the other class. If their accumulated scores are above t, the pixels of the area are assigned to the alternate class label. A sample result is given in Figure 4.11 d), where the small main body text area in the heading above the embellished initial is removed and the area is assigned to the decorative entities class.

Figure 4.11 e, f) show the results for that image patch overlaid with the ground truth denoted by gray blobs. Dark areas denote decorative elements whereas lighter areas stand for the main body text. The background area is not overlaid.

4.4 Summary

Having introduced the main methodologies applied in this report in Chapter 3, namely the interest point detector DoG, the local descriptor Sift and the classifier Svm, this chapter gave the detailed description of the layout analysis system implemented.

The approach is divided into two major tasks: extraction of features, classification and subsequently localization of entire entities at pixel level. Interest points are identified and local descriptors characterizing the interest points' spatial extends are computed.

These features describe the layout entities based on their local structures. Specific similar structures having different scales characterize the classes. Requisite modifications of Lowe's [93] SIFT descriptors for adapting the features to the requirements of the task of layout analysis are given. Amongst them is exploiting the undesirable property of the DoG of being sensitive to edges. This sensitivity is a characteristic of the DoG important for the localization of embellished initials having long strokes.

The descriptors computed for each interest point are classified with a supervised learning algorithm, i.e. a SVM, which is trained on image patches of the respective classes. The classification score of a descriptor in combination with the scale and the location of the associated interest point are used in the localization step.

Since the class label of a pixel cannot be directly inferred from the positions and spatial extends of the interest points, a localization algorithm determining the final class label for each pixel is needed. This expands the single interest points to regions encapsulating whole entities. A cascading method is proposed that successively rejects unreliable interest points and generates cohesive regions based on the interest points belonging to one class.

The result of this method is a class label for each pixel of the document image which denotes if it belongs to the main body text, the decorative entities or the background. The background is simply determined by the absence of any other class label.

Chapter 5

Evaluation and Results

This chapter describes the evaluation process of the layout analysis system introduced and provides results. It is empirically evaluated by means of manually annotated real world data. The experiments described in the following section are set up in order to evaluate the strengths as well as the drawbacks of the proposed method.

Three datasets – which are described in detail in Section 1.3 – are employed for the evaluation. The first of which is an ancient manuscript from the 11^{th} century which incorporates challenges with respect to the state of the manuscript concerning degradation, the layout, and the writing style. The second dataset consists of a medieval manuscript with strict layout rules but plain initials which are not embellished. The third dataset is a medieval manuscript having a two-column layout and initials consisting mainly of long single strokes and touching text regions. The evaluation of the method on these three datasets was published in [53].

Section 5.1 will describe the experimental settings; the statistical methods applied to evaluate the method are explained in Section 5.2. Section 5.3 will give results of the method applied to the *Psalter* dataset, and the subsequent sections will provide the results for the medieval manuscripts.

5.1 Experiments Overview

The method introduced in this report is evaluated on a random sample of 100 pages of each manuscript. However, a uniform distribution of samples over the entire manuscript is intended since the writing style, the layout and the style of the embellishments may change in the course of the manuscript. This is due to the mere fact that copying a manuscript took a certain amount of time and handwriting is subject to variability as well as different scribes may have written on the same manuscript (see Section 1.3 for detailed information about ancient manuscripts).

Thus, the pages selected as test set have variations in layout (e.g. margin of the text body to the page border, number of text lines, space between two text lines), scripts, writing styles, and writing instruments. Character sizes within pages and between pages may vary. In case of these manuscripts having a stricter layout (Section 5.4 and Section 5.5), variations in writing style, in the style of the initials, and the number of text lines are

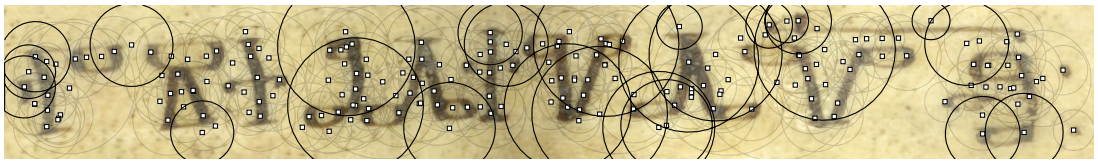


Figure 5.1: Image patch of the training set overlaid with interest points overlapping with the image border.

occurring.

Having chosen a supervised learning based classifier, the training set for each manuscript consists of image patches containing one entity in case of initials, a whole heading or a certain amount of text lines and background. The number of entities is different for the particular datasets and will be given in the respective sections. Each image patch has its class label assigned. They were manually extracted from the respective manuscripts. Due to the spatial proximity of the entities to each other, it was necessary to crop the image patches with a narrow margin. Thus, interest points overlapping with the border of the image patch do not incorporate information of the region outside the image patch, and thus, the descriptor characterizes a region smaller than the spatial extend of the interest point. However, it is presumed that the regions outside the image patch border are not relevant to the description of the layout entity, as the might just contain background information.

Figure 5.1 gives an example of an image patch of the training set overlaid with representations of interest points. The locations of the interest points are given with white squares and their spatial extend by corresponding circles. These interest points which overlap with the image border are highlighted in black.

The evaluation of the method is done at pixel level, i.e. each pixel of a manuscript page in the test set has a class label. The ground truth was manually tagged by brushing the respective regions with gray values corresponding to their class indices. Since a binarization of the manuscript is not the aim of a gray-level based method, it was not intended to generate a ground truth that exactly follows the strokes of the characters as it would be necessary for the evaluation for a binarization-based Ocr system. The goal is rather the determination of regions of interest and thus, a margin between the entities and the border of the ground truth is created. Text regions are merged to regions rather than lines if the space between two text lines does not exceed the height of a text line. An example for ground truth overlaid on an image patch of the test set is provided in Figure 5.2.

Furthermore, an exact determination of the borders of a character in degraded ancient manuscript is actually impossible since borders can be ambiguous due to various reasons e.g. regions having low contrast, faint ink, strokes petering out rather than having a determined end, or stains and damages of the writing support overlaying the character.

Smith [127] shows that the manual ground truth generation for characters in handwritten documents is subject to variability. Furthermore, ground truth that is obtained manually, is subjective and prone to error and variability, even if just one person generates the ground truth, since the judgment of a person may change over the generation of ground truth for multiple images.

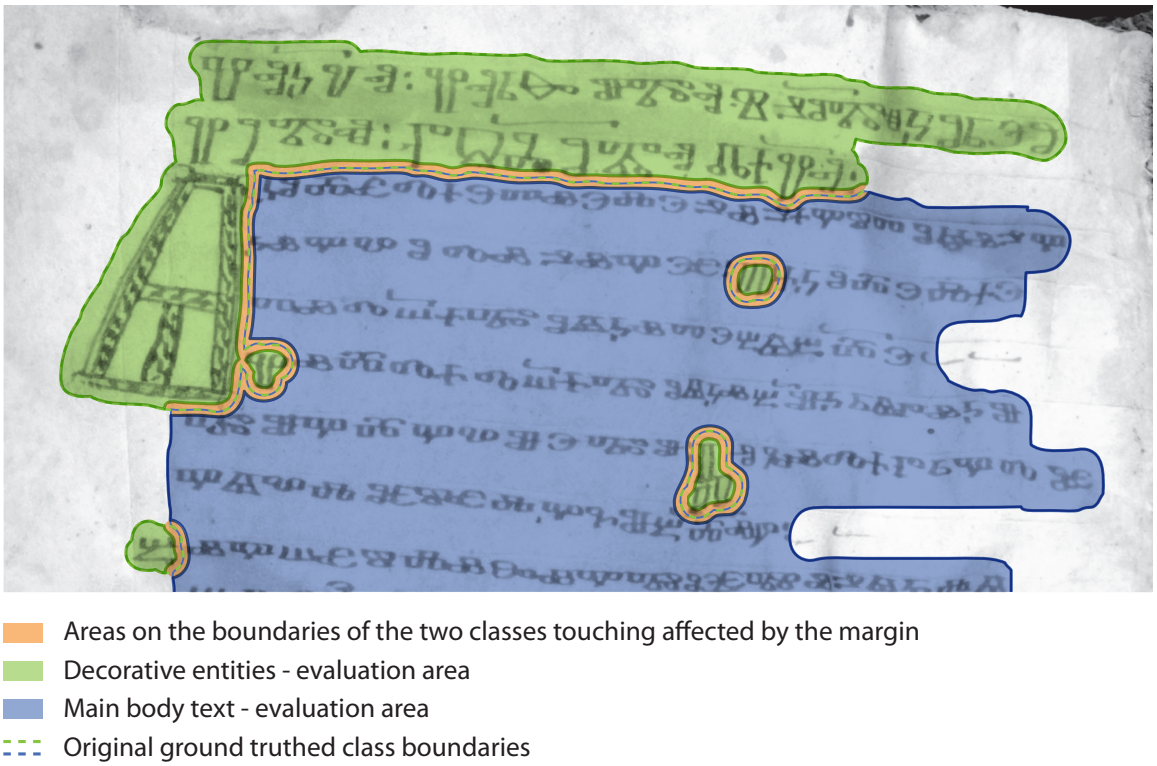


Figure 5.2: Image patch of the test set overlaid the marginized ground truth.

Since layout entities frequently touch each other or overlap, a determination of the class is not possible for the pixels in these regions. Thus, the evaluation of the localization is not carried out at positions having overlapping class labels. Therefore, 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 as described before. This noise occurs especially in border regions of overlapping classes. On the other hand – depending on the data – the classes may have fuzzy or overlapping region borders which render exact ground truth segmentation impossible.

Figure 5.2 gives an example of an image patch of the test set overlaid with the marginized ground truth. The class of decorative entities is denoted by green color, the text class by blue. In the area where the margin affects the ground truth – denoted in orange –, the original class boundaries are denoted by dashed lines.

There are two groups of evaluations given for each manuscript. First, the performance of the entire method is given, which means the final score maps are evaluated at pixel level. Second, the SIFT descriptors are evaluated based on the location of their interest points in the manuscript image. Not the scale – and thus, the spatial extend – of the interest point is taken as measurement, but the mere position of the interest points leading to a correct classification if the interest points location is within a region having the same class label as the interest point. Interest points within the intra-line space of two consecutive text lines are supposed to belong to the class of main body text since the ground truth is generated for regions of interest as described before.

5.2 Statistical Methods

The performance measure employed for the evaluation of the method is *precision* and *recall* [136], a standard measure for evaluating the accuracy. These measurements are based on following values:

True Positives (tp) pixels or interest points which are correctly classified, i.e. they correspond to the ground truth

False Positives (fp) or Type I error, pixels or interest points are located within a region of interest in the ground truth but have a non-corresponding class label – in statistical terms it means a statistical test rejects a true null hypothesis (correct class in the ground truth)

False Negatives (fn) or Type II error, pixels or interest points which are not detected by the method but in the ground truth their location is labeled to be one of the classes – in statistical terms it means the test fails to reject a false null hypothesis (not the correct class in the ground truth)

Precision and recall are computed of these values as given in Equation 5.1:

$$\mathsf{p} = \frac{\mathsf{tp}}{\mathsf{tp} + \mathsf{fp}}, \qquad \mathsf{r} = \frac{\mathsf{tp}}{\mathsf{tp} + \mathsf{fn}} \tag{5.1}$$

where r is the recall and p is the precision.

Hence, the precision specifies the proportion of correctly classified pixels/interest points to all pixels/interest points labeled as belonging to a class in the ground truth. The recall indicates the proportion of correctly classified pixels/interest points to the total number of pixels/interest points actually belonging to the respective class. The aim of a classification task is to maximize both, precision and recall. Thus, a weighted average between precision p and recall r – the F-score – is employed to give a measurement of the method's performance:

$$F_\beta = \frac{(1 + \beta^2)\mathsf{p} \cdot \mathsf{r}}{\beta^2 \mathsf{p} + \mathsf{r}} \tag{5.2}$$

with β being a weighting parameter that either increases the weight of the precision or the recall. For $\beta = 0.5$, the precision is weighted twice as much as the recall. This value depends upon the specific classification task. The parameter β is chosen to be 0.5 for the evaluations in this report since the importance of correct results retrieved is greater than to detect all members.

As regards the evaluation of interest points, not all of the interest points need to belong to the correct class since the method is robust to a certain amount of misclassifications owing to the localization algorithm introduced. Misclassifications occur when entities of different classes have structures which are similar at a certain level, e.g. large-scale structures of an initial may have similarities with structures of a character of the main body text at a smaller scale. Since mis-classified interest points are rejected during the localization algorithm, the correctly classified interest points have more importance than having all of the interest points voting for the correct class.

In case of the pixel-level evaluation, where the result of the localization algorithm – the score maps – is evaluated for each pixel, the same rule applies since entities not found generally are plain initials which are part of the text or within the left margin of the text but have similar structures as the main body text and parts of headings. There exist plain initials within the text body which cannot be distinguished from the main body text based on their local structures but just because of a space before the character that is longer than usual or special punctuation right before the character. Plain initials within the left margin of the text body, having similar characteristics as the main body text can later be detected applying a method that exploits spatial relationships between entities. Parts of headings may be classified as main body text if their structural similarity is high, however, since headings are not detected within the main body text class, and the fact that headings usually cover a whole line, the headings can be extended using that knowledge in a further step.

5.3 *Psalter*

The training set of the *Psalter* consists of image patches of the respective classes, where initials and headings build one class and main body text represents the second class. Image patches containing 18 embellished initials, 30 plain initials and 30 headings for the class of decorative entities, and 60 lines of main body text are taken as training samples for the classifier. Please note that the set of decorative entities – especially the set of embellished initials – does not cover the entire range and variety of these entities occurring in the manuscript. Thus the classification of the decorative elements completely relies on the similarity of their local structures. The same applies to the small initials, where the detection of the elements mainly depends on the angularity of the characters as well as on the length of the individual strokes of the characters.

Table 5.1, Table 5.3, and Table 5.4 give the results of the empirical evaluation of the method proposed in this report for the *Psalter* dataset. F-score, precision and recall are used as measure metrics to indicate the method's performance.

5.3.1 Final results

In Table 5.1, the results of the evaluation of the method introduced in this report are given. The final result after detecting interest points, extracting and classifying descriptors and applying the localization algorithm is a segmentation into regions of interest which are pixel-wise assessed based on the ground truth for each of the images in the test set.

For the interpretation of the results in Table 5.1 it has to be considered that the ratio between main body text and decorative entities is approximately 9.2 : 1 on 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 5.1 a-c).

Table 5.1 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. For headings and plain initials, the difference to

Table 5.1: *Psalter* – result: precision, recall and F-score for the evaluation of the score maps at pixel-level.

	Precision	Recall	F-score
a) All classes	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

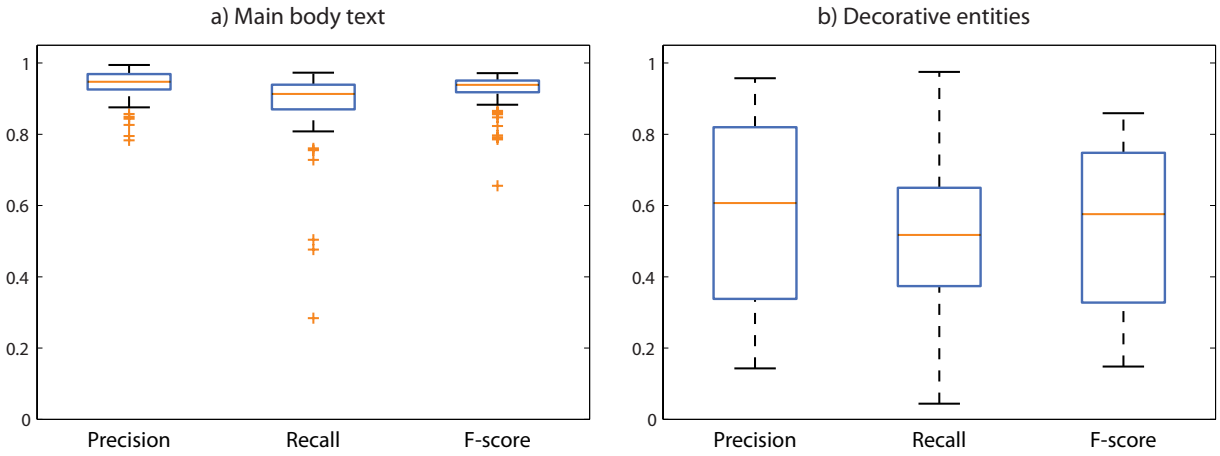


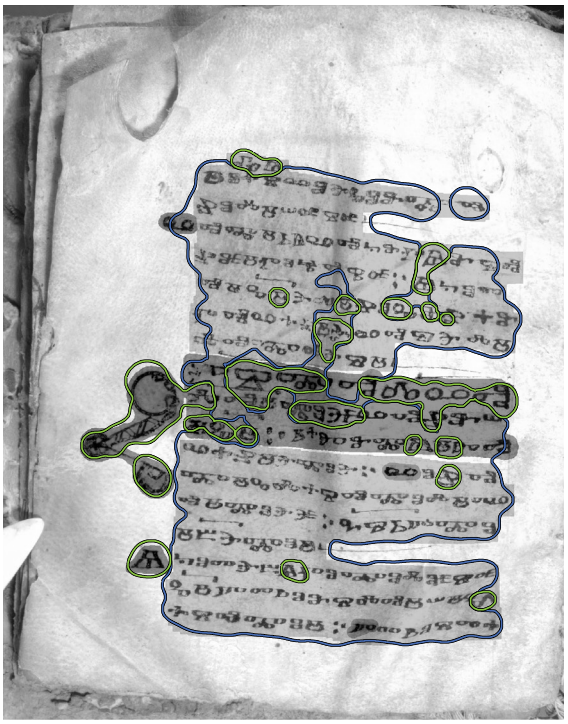
Figure 5.3: Localization evaluation per image.

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 on the one hand and plain initials or headings on the other hand is a non-trivial task. A detailed analysis will be given later in this section.

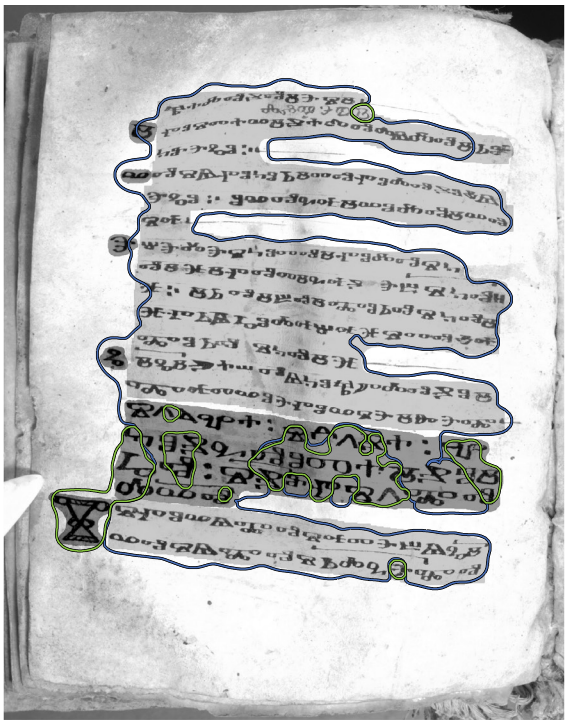
Figure 5.3 shows the variance in the results per image. On the left, precision, recall and F-score are given for the text class, and on the right for the class of decorative entities. In general, the variance in the results is less for text than for the other class. The outliers are the first or last pages of the *Psalter*. The condition of these pages is worse than for pages in the center of the document and the scribe used different writing tools and had a different writing style for these pages.

Figure 5.4 and Figure 5.5 show eight example results having different layouts, writing styles, decorative entities and degrees of faint ink. The pages were taken from one of the first to one of the last pages to additionally give an impression about possible variation in the course of the manuscript. The ground truth is given as gray blobs overlaying the image, where light gray denotes text areas and dark gray indicates decorative entities. Table 5.2 gives the F-scores, precision and recall for each folio shown in both figures.

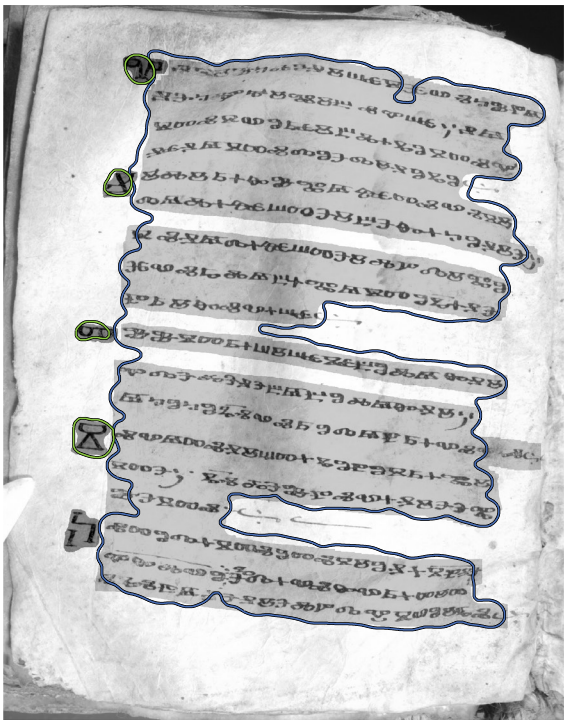
Figure 5.6 gives exemplary results for the decorative entities as headings (a-e), embellished initials (a,l-a) and plain initials (f-k,n-o,q). The names of the entities are abbreviated to H for Heading, PI for Plain Initial and EI for Embellished Initial. If an image patch contains more than one entity of each type, the number of entities is given, and in



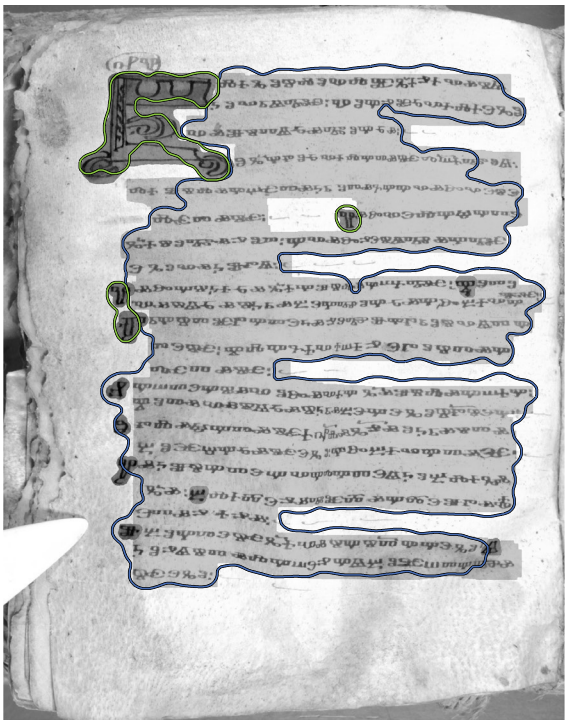
Folio 5v



Folio 10v

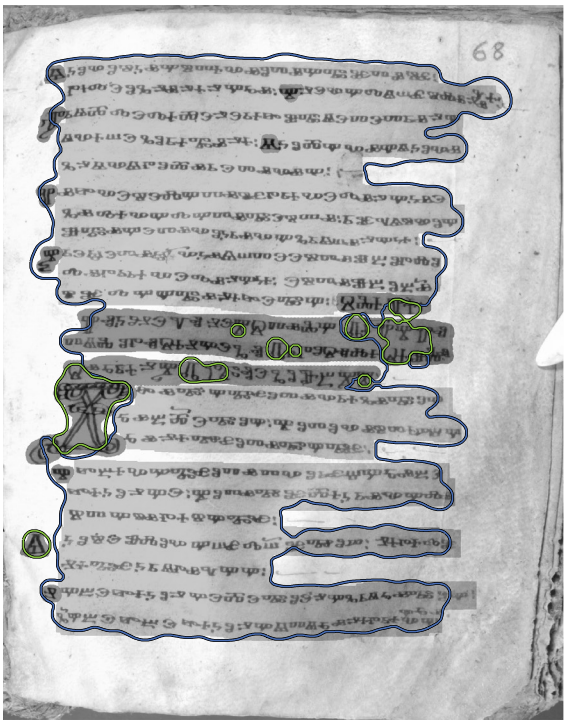


Folio 17v

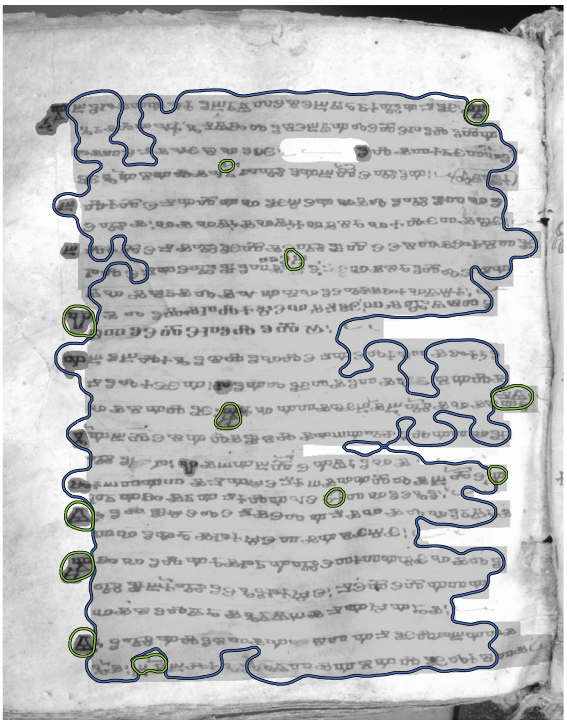


Folio 43v

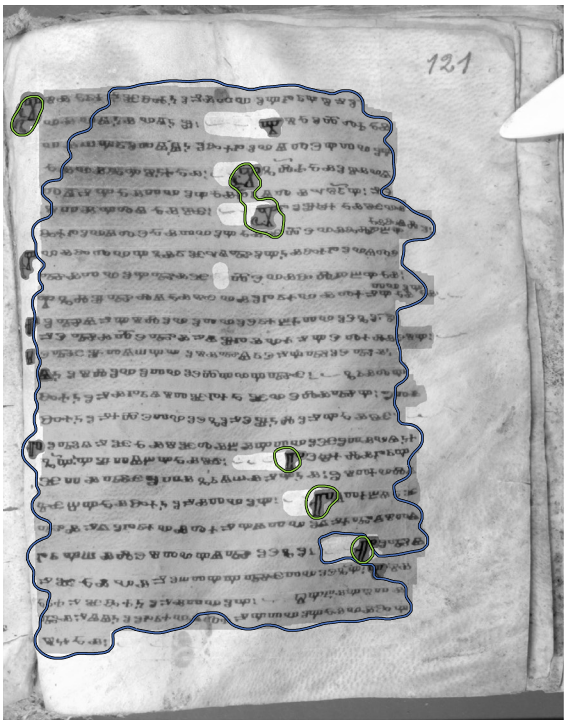
Figure 5.4: *Psalter* – Exemplary results for entire pages. The ground truth is denoted by gray blobs where light gray indicates the main body text class and dark gray stands for the decorative entities. Detected areas are surrounded by either green (decorative entities) or blue (text) contours.



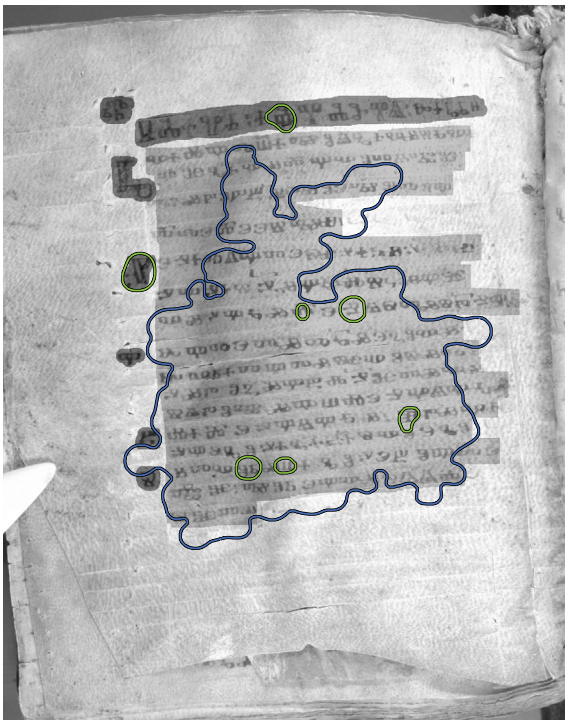
Folio 68r



Folio 92v



Folio 121r



Folio 141v

Figure 5.5: *Psalter* – Exemplary results for entire pages. The ground truth is denoted by gray blobs where light gray indicates the main body text class and dark gray stands for the decorative entities. Detected areas are surrounded by either green (decorative entities) or blue (text) contours.

Table 5.2: *Psalter* – result: precision, recall and F-score for the evaluation of the score maps at pixel-level for the folia given in Figure 5.4 and Figure 5.5.

	Main body text			Decorative Entities		
	Precision	Recall	F-score	Precision	Recall	F-score
5v	0.870	0.884	0.873	0.751	0.477	0.674
10v	0.857	0.959	0.878	0.962	0.415	0.761
17v	0.998	0.951	0.988	0.976	0.578	0.858
43v	0.986	0.931	0.974	0.982	0.629	0.883
68r	0.884	0.966	0.899	1	0.237	0.608
92v	0.989	0.891	0.968	0.529	0.436	0.507
121r	0.993	0.932	0.980	0.786	0.552	0.725
141v	0.984	0.755	0.928	0.355	0.098	0.232

case of headings, the number of lines.

Structural Information

First, the embellished initials are detected and localized well if sufficient structural detail is present. Long single strokes are not detected well by DoG and Sift, since edges do not provide reliable interest points [93]. Hence, at these strokes, the density of interest points is low and therefore, a reliable localization cannot be achieved either (see Figure 5.6 p,q) and left part of the embellished initial in Figure 5.6 o)).

Furthermore, plain initials placed in the margin of the page being smaller than an average regular character are either not covered by the marker points or do not produce sufficient interest points to be considered as an initial. For an example refer to Figure 5.6 f), first row in the left margin; there is a plain initial that is not detected.

Faint ink causes difficulties in the detection of characters or character parts, such as the upper part of the embellished initial in Figure 5.6 l) or the plain initial in the top left corner of Figure 5.6 j), where parts of the character are vanished. The remaining segments are small and they look similar to the characteristics of the main body text. For a human observer, it is easy to recognize the character since the vanishing lines of the initial are completed by human vision, which extends the visible parts of a line such that it fills in the missing part of the line. However, lacking this information introduced by the human vision and observing the existing structures, it becomes clear that the remaining parts of the character are so small that they are rejected in one of the filtering steps in the localization algorithm.

Discriminability

A second issue relates to plain initials and headings having features not discriminative 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

Figure 5.6: *Psalter* – Exemplary results for decorative entities. Detected decorative entities are surrounded by black contours.

non-trivial task. In Figure 5.6 b,c,e), headings with certain characters having similar characteristics like the main body text are shown.

Similar to headings, initial characters are – depending on the scribe – similar to the main body text and are identified as initials by characteristics such as punctuation or long spaces before the initial. Examples are the fourth character from the right side in the first row of Figure 5.6 i), the second character from the right side in the forth row of Figure 5.6 k) or the first character in the fifth row of Figure 5.6 n). A second reason concerns the proximity of another class is given below.

The Cyrillic characters are detected reliably by the method since their structural characteristics are discriminative when compared to Glagolitic letters (see Figure 5.6 a,d).

Touching and Super-Imposing Entities

The third point concerns plain initial embedded in the main body text, since they are single initials surrounded by another class. Additionally to the discriminability, the reliable detection and localization are more difficult when compared with isolated initials surrounded with background, since surrounding interest points may have a higher classification score and overlap the initial. Thus, if the characteristics of the initial are not discriminative, the character is assigned to the text class. In these cases, the determination of these entities needs to be based on other characteristics such as a larger space before the initial or a colon at the end of previous sentence (see Figure 5.4, Figure 5.5). 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. Due to the localization algorithm, that spatially weights the classification scores of interest points, abundantly embellished entities have a higher weight and thus, may superimpose the other class in boundary regions, see Figure 5.6 n-o).

5.3.2 Localization

This section gives an overview of the steps of the localization algorithm where interest points are voted. Table 5.3 and 5.4 provide the results of the respective steps in the localization algorithm. The first table summarizes the evaluation of the class of main body text; the second table shows the evaluation for the decorative entities.

The precision of the initial set of interest points for the class of main body text 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 a high number of interest points in the text regions have been assigned to the decorative entity class despite belonging to the main body text. Increasing the precision is done by reducing the fp, and hence, reducing interest points misclassified as belonging to the text class by rejecting these interest points from the class of decorative entities.

The voting step does not have a major impact on the evaluation of the location of

Table 5.3: *Psalter* – Main body text: precision, recall and F-score for the interest points.

	Precision	Recall	F-score
a) First set of interest points (Section 4.2)	0.959	0.646	0.875
b) Voted points (Section 4.3.1)	0.961	0.637	0.873
c) Marker points (Section 4.3.2)	0.962	0.941	0.958
d) Merged interest points (Section 4.3.3)	0.966	0.896	0.951
e) Final set of interest points (Section 4.3.4)	0.964	0.962	0.964

Table 5.4: *Psalter* – Decorative entities: precision, recall and F-score for the interest points.

	Precision	Recall	F-score
a) First set of interest points (Section 4.2)	0.177	0.735	0.208
b) Voted points (Section 4.3.1)	0.168	0.742	0.198
c) Marker points (Section 4.3.2)	0.636	0.719	0.619
d) Merged interest points (Section 4.3.3)	0.506	0.772	0.543
e) Final set of interest points (Section 4.3.4)	0.713	0.727	0.715

interest points since the classification score of the interest points is subject to this step and thus, the impact of this step cannot immediately be traced back in this evaluation.

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, again 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.

Similar to the evaluation of the main body text, Table 5.4 assesses the locations of the interest points assigned to the class of decorative entities by the classifier. While the precision is low, i.e. the number of interest points located in the text body falsely detected as structures of decorative entities is high, the recall is far higher. The marker points, however, are much more reliable indicators for actual decorative entities. Having 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-score for the class of decorative entities is improved from 0.208 to 0.715.

5.4 *Cod. 635*

For the training of the learning algorithm, 43 initials and 355 lines of text have been extracted from the manuscript images. This manuscript does not contain headings or different kinds of initials. As described in Section 1.3.2, the initials of this dataset are not richly embellished such as those of the *Psalter*, i.e. not consisting of as much structure,

Table 5.5: *Cod. 635* – Score maps: precision, recall and F-score for the evaluation of the score maps at pixel-level.

	Precision	Recall	F-score
a) All classes	0.969	0.984	0.972
b) Main body text	0.974	0.995	0.978
c) Decorative entities	0.765	0.6343	0.735

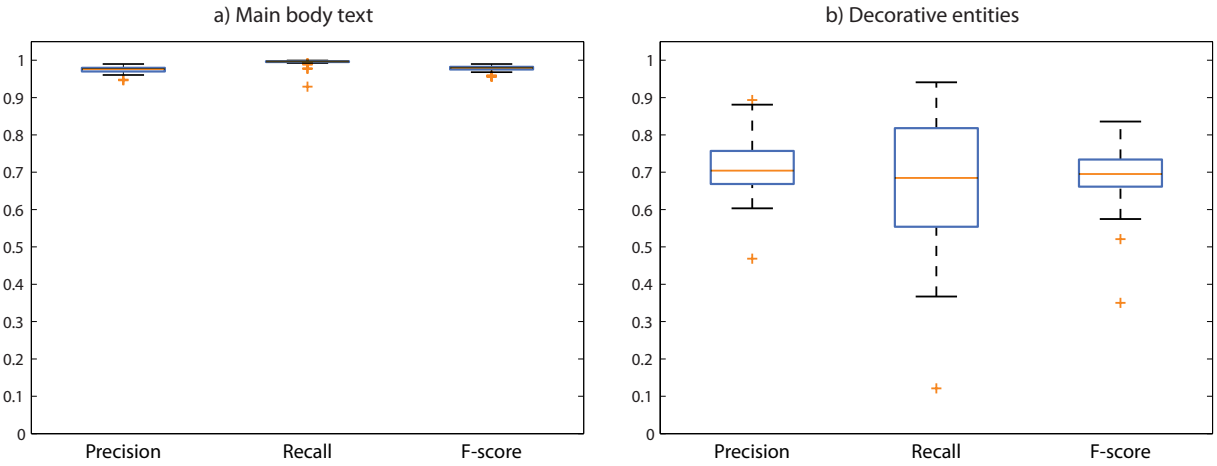


Figure 5.7: Localization evaluation per image.

producing less local minima and maxima and hence, less interest points.

Table 5.5, Table 5.7, and Table 5.8 give the results of the empirical evaluation of the method proposed for the *Cod. 635* dataset. F-score, precision and recall are used as measure metrics to indicate the method's performance.

5.4.1 Final results

In Table 5.5, the results of the pixel-level evaluation of the method for *Cod. 635* is provided. For the interpretation of the results in Table 5.5, please note that 92.2 % of the classified pixels belong to the text class. The table shows that the method performs better on this dataset than on the *Psalter*. However, as described in Section 1.3.2, this manuscript is not as degraded as the *Psalter* and has a stricter layout.

In Table 5.5 b,c), the results for the respective classes are given. The segmentation performance for the class of decorative entities is not as good as for the text class since the initials in the manuscript lack local structures.

In Figure 5.7, the variance in the performance of the method on the pages of *Cod. 635* are shown per class. The variance is higher for the decorative entities than for the text class.

Figure 5.8 gives four example pages of *Cod. 635* overlaid with the ground truth. As can be seen, the proximity of the entities affects the segmentation. The text produces a higher number of interest points than the initial – especially in case of initials representing

Table 5.6: *Cod. 635* – result: precision, recall and F-score for the evaluation of the score maps at pixel-level for the folia given in Figure 5.8.

	Main body text			Decorative Entities		
	Precision	Recall	F-score	Precision	Recall	F-score
11r	0.985	0.993	0.989	0.726	0.854	0.785
40v	0.977	0.990	0.983	0.724	0.720	0.722
60v	0.975	0.992	0.984	0.727	0.737	0.732
98v	0.969	0.991	0.980	0.607	0.456	0.521

a character such as a I or J – because it has more structures producing local minimas and maximas on different scales. The initials, however, produce a low number of interest points along the edge of the character. Thus, the classification performance is low for decorative entities located close to the text body (see Figure 5.8)Table 5.6 gives the F-scores, precision and recall for the folia shown in both figures.

Owing to the uniformity of the main body text's appearance, the contrast of the writing to the background and absence of degradation and staining – such as present in the *Psalter* –, the segmentation of the text region performs well with an F-score of 0.978 over the whole training set. The contours of the detected text regions follow the ground truth closely in absence of initials. In regions, where initials and main body text are adjacent, the boundary does not separate the two classes perfectly in all cases.

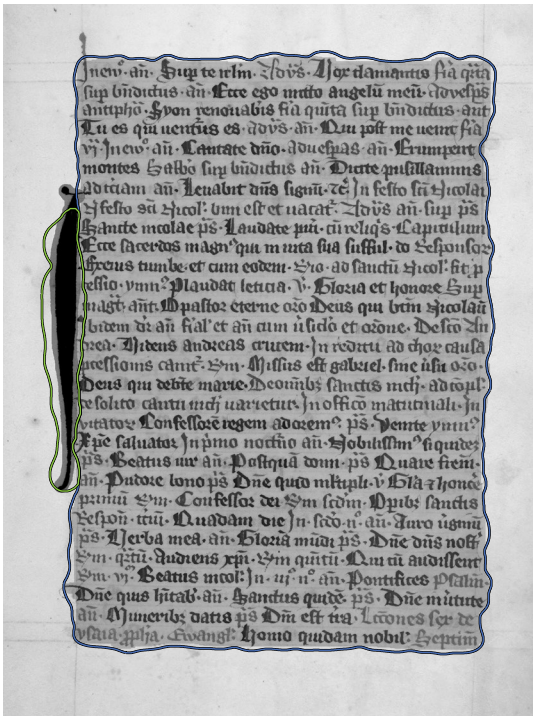
In Figure 5.8, Folio 40v, the initial on the bottom of the page touches the characters of the text, the same is true for the initials in Folio 98v. Thus, the interest points on the edge of the initial are ambiguous since approximately half of their region votes for the class of main body text. Therefore, the segmentation includes parts of the initial and a detection of the entire initial renders impossible due to the missing right edge. In Folia 11r and 60v, however, the segmentation is successful for the major part of the initial. Yet, the upper segment, which is touching the text area, is assigned to the text class in both cases.

Especially thin strokes such as the horizontal strokes of initials tend to produce an insufficient number of initials to reliably detect these areas (e.g. compare the initials B and H in Folio 40v, or both initials in Folio 60v).

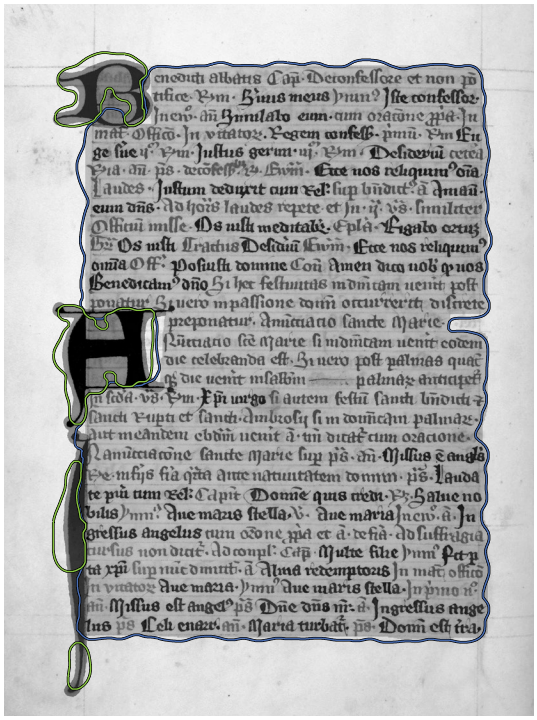
5.4.2 Localization

This section gives the evaluation of the localization algorithm for interest points on *Cod. 635* manuscript. Table 5.7 and 5.8 give an overview of the performance of the respective step of the localization algorithm. The first table summarizes the performance for the class of main body text, while the second table gives the evaluation results for the decorative entities.

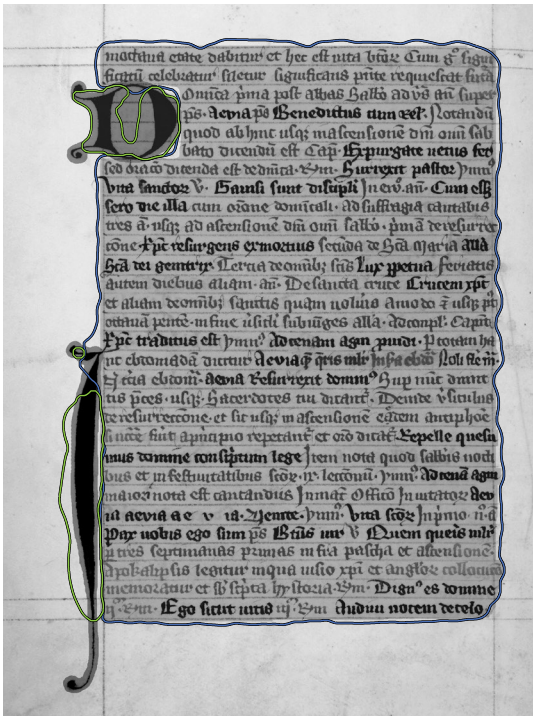
For the class of main body text, the initial set of interest points is accurate. Few interest points are misclassified in the text body (precision) and even less interest points are wrongly classified to the class of decorative entities. Thus, the localization algorithm



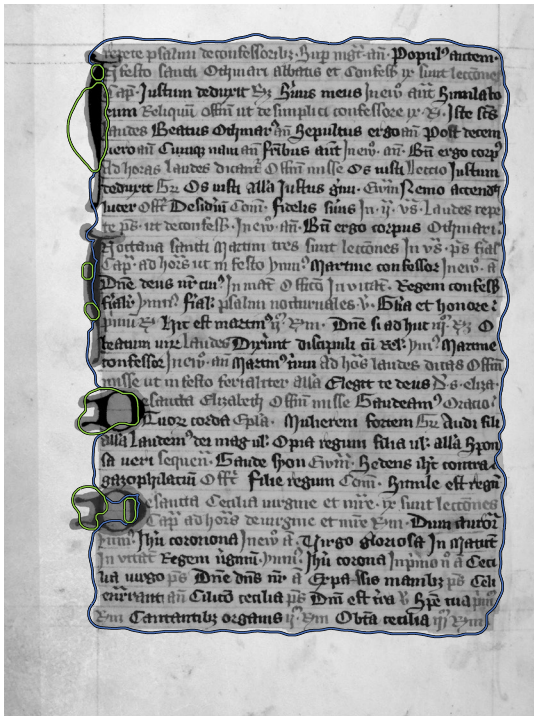
Folio 11r



Folio 40v



Folio 60v



Folio 98v

Figure 5.8: *Cod. 635* – Exemplary results for entire pages. The ground truth is denoted by gray blobs where light gray indicates the main body text class and dark gray stands for the decorative entities. Detected areas are surrounded by either green (decorative entities) or blue (text) contours.

Table 5.7: *Cod. 635* – Main body text: precision, recall and F-score for the interest points.

	Precision	Recall	F-score
a) First set of interest points (Section 4.2)	0.992	0.997	0.993
b) Voted points (Section 4.3.1)	0.992	0.997	0.993
c) Marker points (Section 4.3.2)	0.994	0.992	0.993
d) Merged interest points (Section 4.3.3)	0.994	0.995	0.994
e) Final set of interest points (Section 4.3.4)	0.994	0.999	0.995

Table 5.8: *Cod. 635* – Decorative entities: precision, recall and F-score for the interest points.

	Precision	Recall	F-score
a) First set of interest points (Section 4.2)	0.742	0.563	0.697
b) Voted points (Section 4.3.1)	0.752	0.560	0.704
c) Marker points (Section 4.3.2)	0.741	0.785	0.749
d) Merged interest points (Section 4.3.3)	0.741	0.678	0.727
e) Final set of interest points (Section 4.3.4)	0.951	0.686	0.883

has no major influence on the performance.

Contrary to the class of main body text, the localization algorithm improves the result for the interest points of the class of decorative entities. Concerning the recall, the selection of marker points is a crucial task. Interest points falsely classified as structures of decorative entities are rejected in the final filtering operation after merging the interest points. Applying this, the precision is improved from initially 0.742 to a final precision of 0.951.

5.5 *Cod. 681*

The training set for *Cod. 681* consists of 41 initials for the decorative entity class and 584 lines of text in two columns for the class of main body text. Table 5.9, Table 5.11, and Table 5.12 provide an analysis the performance of the method on this manuscript in terms of precision, recall and F-score.

Compared to the other two manuscripts, *Cod. 681* is different as regards writing support and layout. The *Psalter* and *Cod. 635* are inscribed on parchment, whereas the writing support of *Cod. 681* is paper, despite being from the same century as *Cod. 635*. The previous two manuscripts have single column layouts, and *Cod. 681* is written in a two column-layout and has layout rules strictly complied with.

5.5.1 Final results

Table 5.9 provides the evaluation of the final result of the method for *Cod. 681* on the pixel level. For the interpretation, it has to be considered that 99 % of the classified pixels

Table 5.9: *Cod. 681* – Score maps: precision, recall and F-score for the evaluation of the score maps at pixel-level.

	Precision	Recall	F-score
a) All classes	0.974	0.977	0.975
b) Main body text	0.979	0.981	0.979
c) Decorative entities	0.712	0.566	0.677

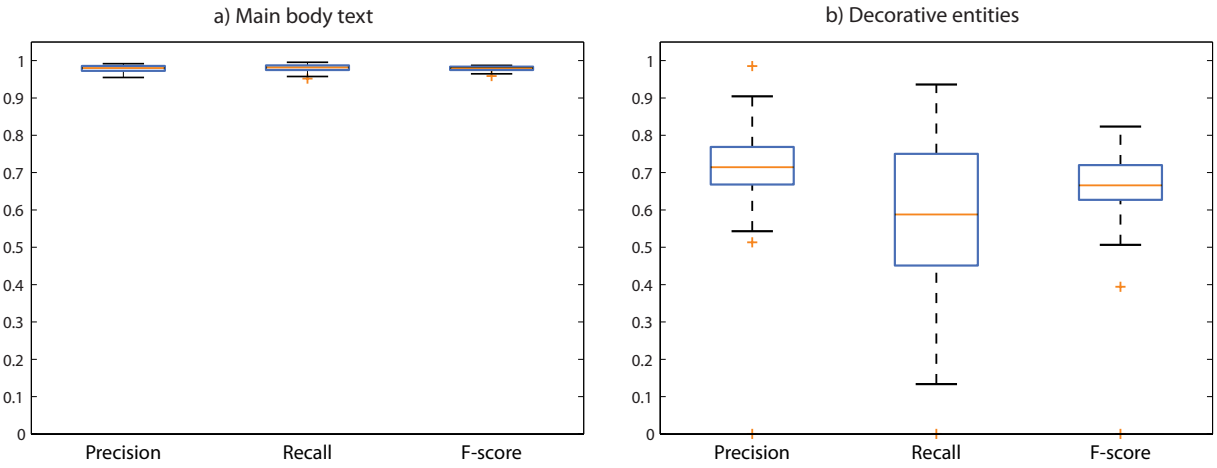


Figure 5.9: Localization evaluation per image.

belong to the class of main body text. This is due to the size of the initials as well as due to the fact that not each of the pages in the test set contains initials.

Analogous to *Cod. 635*, the method performs better on this manuscript than on the *Psalter*, due to the same reasons: stricter layout with a regular and uniform appearance of the text body. Similar to the other manuscripts, the classification performance is not as promising for the decorative entities as for the main body text. The initials of this manuscript do not provide as much structures as the those of the *Psalter*.

Figure 5.9 shows the variance in the performance of the method on the pages of *Cod. 681* per class. Similar to the two other manuscripts, the variance is higher for the decorative entities than for the class of main body text. The variance in the precision for the class of decorative entities is not as high as in the recall.

In Figure 5.10, four exemplary results for the *Cod. 681* dataset are provided. The manuscript images are overlaid with the ground truth which is denoted by gray blobs. The detection of text areas and the further segmentation of the text regions into two columns follows the contours in the ground truth. Annotations such as in Folio 66v and 120v are not entirely detected. The script is different to the one in the main body text and the training set did not contain samples of this script.

Despite different contrast levels owing to the color of the text (e.g. compare Folio 14v and 49v) – written with either black or red ink –, the performance of the approach does not decrease.

Table 5.10: *Cod. 681* – result: precision, recall and F-score for the evaluation of the score maps at pixel-level for the folia given in Figure 5.10.

	Main body text			Decorative Entities		
	Precision	Recall	F-score	Precision	Recall	F-score
14v	0.974	0.989	0.981	0.551	0.307	0.394
49r	0.977	0.984	0.981	0.650	0.724	0.685
66v	0.983	0.974	0.979	0.625	0.756	0.684
120v	0.983	0.974	0.978	0.672	0.853	0.752

Table 5.11: *Cod. 681* – Main body text: precision, recall and F-score for the interest points.

	Precision	Recall	F-score
a) First set of interest points (Section 4.2)	0.996	0.995	0.996
b) Voted points (Section 4.3.1)	0.996	0.995	0.996
c) Marker points (Section 4.3.2)	0.995	0.981	0.992
d) Merged interest points (Section 4.3.3)	0.996	0.992	0.995
e) Final set of interest points (Section 4.3.4)	0.996	1	0.997

The detection of initials faces problems if the initial is touching letters of the main body text since the interest points located at the edge touching the text regions describe text regions too. The same problem is reported for *Cod. 635* in Section 5.4.1

Tapering stroke endings such as the upper stroke of the initial Q in Folio 49 or the horizontal stroke endings of the I in Folio 49r and 66v do not generate sufficient interest points to be detected reliably. The local structure of the letter W used as initial is apparently similar to the structures of the main body text. Thus, a major part of the initial is assigned to the class of main body text in Folio 14v.

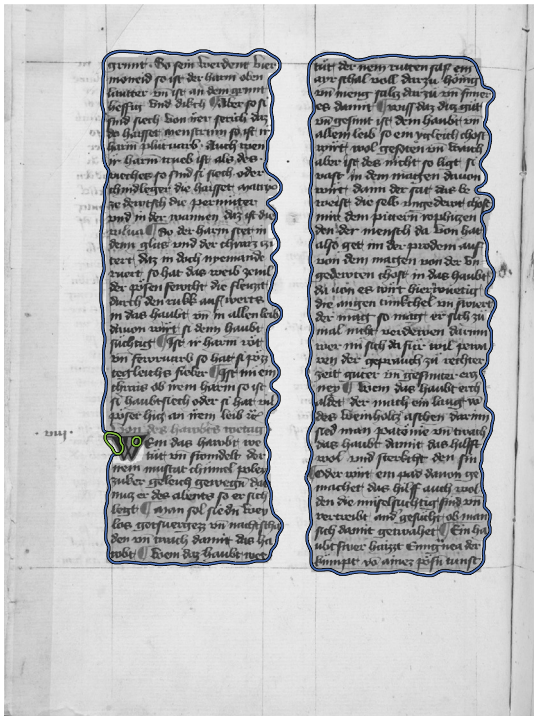
Table 5.10 gives the performance of each of the example pages in Figure 5.10. Precision, recall and F-score are given for the class of main body text and for the class of decorative entities.

5.5.2 Localization

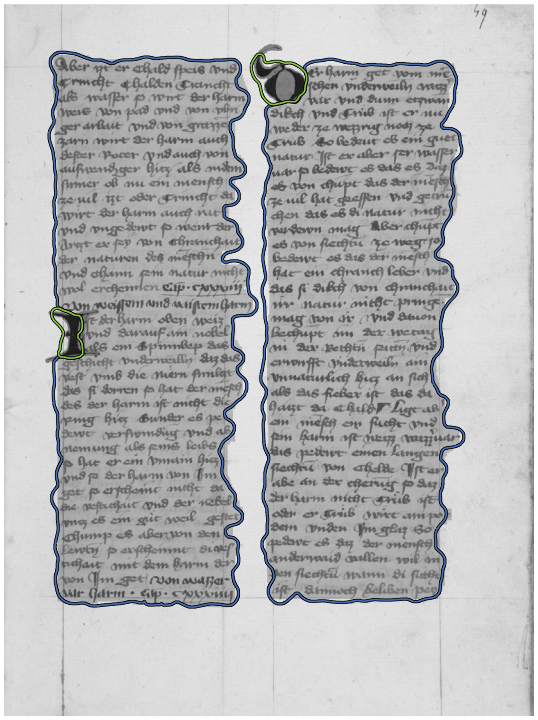
The performance of the interest points detected by the DoG algorithm and their class assignment according to the features extracted at their positions is evaluated in this section.

In Table 5.11 and 5.12, the evaluation results for each step of the localization algorithm is given. The first table summarizes the performance for the class of main body text, while the second table gives the evaluation results for the decorative entities.

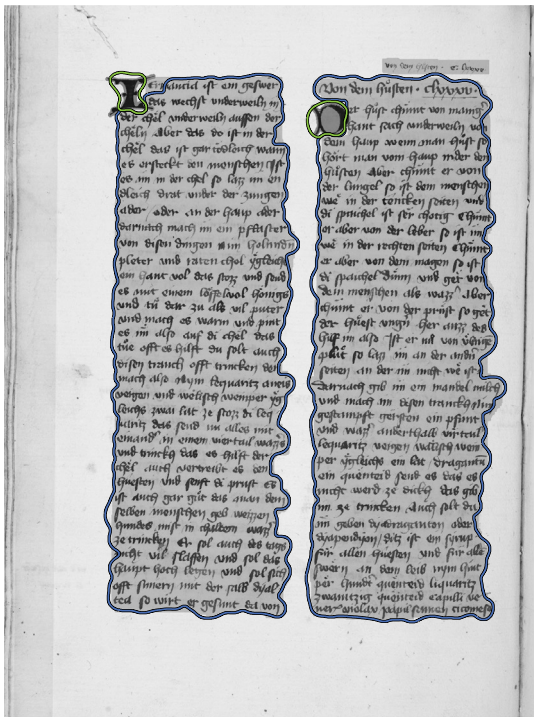
Owing to the uniform appearance of the main body text, the performance of the initial set of interest points is high and thus, the localization algorithm is not effective on these interest points.



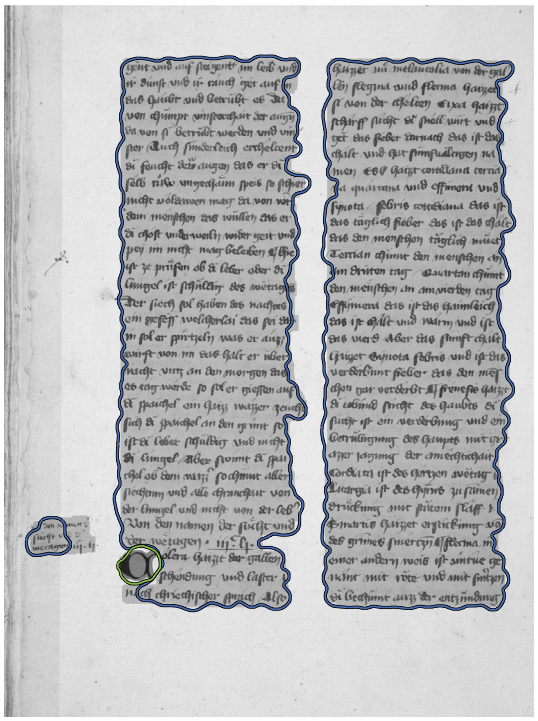
Folio 14v



Folio 49r



Folio 66v



Folio 120v

Figure 5.10: *Cod. 681* – Exemplary results for entire pages. The ground truth is denoted by gray blobs where light gray indicates the main body text class and dark gray stands for the decorative entities. Detected areas are surrounded by either green (decorative entities) or blue (text) contours.

Table 5.12: *Cod. 681* – Decorative entities: precision, recall and F-score for the interest points.

	Precision	Recall	F-score
a) First set of interest points (Section 4.2)	0.392	0.457	0.403
b) Voted points (Section 4.3.1)	0.392	0.457	0.403
c) Marker points (Section 4.3.2)	0.391	0.696	0.428
d) Merged interest points (Section 4.3.3)	0.391	0.560	0.416
e) Final set of interest points (Section 4.3.4)	0.934	0.575	0.830

However, for the class of decorative entities, the localization algorithm provides a considerable increase of performance from initials 0.403 to 0.830. The localization algorithm is most effective in rejecting fp, i.e. interest points mis-assigned to the decorative entities by the classifier. In terms of precision, the most crucial step of the algorithm is the filtering operation after merging the interest points to obtain the final set, where unreliable candidate regions for decorative entities are rejected.

5.6 Summary

In this chapter, the evaluation of the method introduced in this report was given. The method is evaluated based on manually annotated real world data, namely three ancient manuscripts from different periods of time. The manuscripts differ in degradation state – with one manuscript being more degraded than the other ones – the writing support, the layout, the degree of compliance of the layout rules, the script and the intra-manuscript variance in the writing style and script, the decoration of embellishments such as initials.

First, an overview of the experiments conducted in course of the evaluation was given. The evaluation of each manuscript was based on a test set of 100 pages of each manuscript which were randomly selected. The training set and the extraction process were described. The test set was manually annotated with ground truth. Since this process is prone to errors, a method to reduce effects of varying judgments of the person who tagged the ground truth and noise in border regions of overlapping classes is described. The method was evaluated at different steps of each algorithm, where a special focus was laid on the localization algorithm applied to interest points.

Then, the statistical measures employed to indicate the performance of the method on each manuscript, was described. The metrics used were precision, recall and $F_{0.5}$-score.

Subsequently, the results were given for each of the manuscripts. First, the final segmentation result was assessed, then the localization algorithm on the interest points. For each manuscript, sample results were shown and analyzed.

Chapter 6

Conclusion

In this report, a layout analysis approach for ancient manuscripts is introduced that exploits structural similarities of layout entities using local features. The term layout entity is referred to written or embellished objects in a manuscript or printed document, such as an initial, the main body text, a heading, or a drawing. SIFT descriptors known from the field of object recognition are employed to describe segments of layout entities in a scale-, rotation- and illumination invariant manner.

In contrast to the majority of the state-of-the-art methods for layout analysis of historical documents, which require a binarization step prior to the actual analysis, the method introduced does not need any pre-processing of the document images but is directly applied to gray-scale images. Binarization pre-processing as used in traditional document layout analysis produces errors since background clutter and noise are additionally segmented to the foreground. 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.

Contrary to modern machine-printed documents, ancient manuscripts require algorithms to be robust with respect to background artefacts such as clutter, stains and noise. Ancient handwritten documents do not have as strict layouts rules as modern machine-printed documents. 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. Furthermore, robustness to low contrast, e.g. in case of faint ink, stains and rippled pages is required.

Owing to the use of local features, the method introduced is independent of the physical and logical layout of a manuscript, i.e. it does not rely on a physical layout model, such as constraints of potential locations of layout entities or spatial relationships between them. In one of the datasets, for example, headings are accompanied by embellished initials, where the heading is located top and right of the embellished initial. Another example is the location of plain initials – usually they are located in the left margin of the text, however, there are occurrences in the text, indicated by spaces left of the initial.

Three manuscripts dating from different centuries and having different geographical origins are regarded: the *Cod. Sin. Slav.* 3N from the 11^{th} century, found on Mt. Sinai in 1975, written in Glagolitic script, the oldest known Slavonic script; the *Codex Claustroneoburgensis 635*, a manuscript written in Gothic script in Latin language originating

from the 14^{th} century; and finally, the *Codex Claustroneoburgensis 681* dating from the end of the 14^{th} century, containing text in Latin and German in a different Gothic script.

In the approach proposed, the detection of layout entities is based on intra-class similarities; e.g. in case of main body text, compact, rounded shapes are such structural similarities to exploit. Based on their characteristics, two classes are built: the main body text and layout entities having a decorative meaning such as initials and headings. In order to classify the local descriptors, a SVM with RBF kernel is used as supervised classifier. Since 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 based on reliable interest points is proposed that successively rejects weak candidates applying voting schemes. Then, a score map is established based on the scales of the interest points and the classification score of the classifier. The score map finally allows for a class decision for each pixel of the document image.

The three manuscripts regarded are different in terms of layout rules, degree of embellishments for decorative entities, degree of degradation, writing support and especially the script. However, the method does not need any parameter tuning except for minor optimizations such as the threshold for detecting interest points (Section 3.1.4) for the different manuscripts.

The approach was shown to be able to adapt to different scripts. Experiments on the *Psalter* dataset demonstrated the method's performance in presence of noise, background clutter and faint ink. Experiments on all datasets proofed that the system is capable to distinguish different scripts based on character segments and is independet of the actual script. The evaluation is based on manually annotated ground truth and is set up such that errors introduced by the classification and by the different steps of the localization algorithm are provided separately.

6.1 Disadvantages

When regarding modern printed documents, where the discriminating characteristic between the regular text and the decorative entities – such as headings – is the size of the characters, the method fails since the employed feature system is invariant to the scale. Thus, if the approach is applied to a machine-printed document whose main body text is of the same typeface as the headings, the detection of headings is unfeasible. However, the localization algorithm can be altered such that the scale is the discriminating feature between headings and main body text. Thus, if the document solely contains headings and main body text, the classification step could be omitted and the segmentation is based on interest points rather than descriptors calculated at their positions.

Due to their definition, local features have the potential drawback that they lack information about the context. The smaller the spatial extend of the interest point, the less information is contained. Spatial relationships, however, can be exploited in document layout analysis since layout rules can be incorporated in an analysis system. Examples for such relationships in the *Psalter* dataset are long spaces or specific interpunctations before plain initials in the text; headings are located above and right of embellished initials.

The cascading localization algorithm described in Section 4.3 rejects layout entity canditates if they are too small to be a valid character of the respective class or the number of interest points voting for the entity is too small. Hence, single characters such as plain initials in the margin of the text of the *Psalter* dataset which are not richly embellished, produce an insufficient number of interest points and are rejected even though they are voting for the same class. Here is a trade-off between rejecting misclassified segments of a characters which are characteristic for the other class and keeping isolated prosaic entities.

6.2 Benefits

The proposed approach does not rely on a binarization step but it is directly applied to the gray scale image. Errors introduced by binarization and loss of information content due to reduction to binary information is omitted. Thus, the information content is richer than for bi-level images. Since ancient manuscript are prone to noise and degradation due to their age, binarization is likely to introduce errors. Furthermore, the dynamic range of ancient manuscripts is low due to faint ink and detoriations. Non-binarization based methods are capable of extracting information of low contrast regions.

Employing gradient-based features, local contrasts are exploited – depending on the sensivity chosen (threshold) – such that even low dynamic range situations can be handled. Furthermore, the features employed incorporate robustness with respect to illumination changes.

Owing to the features and the machine learning approach chosen for classification, the propsed approach is robust to variations in the shape of characters, i.e. the personal writing style of a scribe. However, the method is discriminative enough to distinguish between different scripts and distinct writing styles.

The use of local features averagely having the scale of a character (the scales range from character segments to parts of text lines, or parts of embellished initials respectively, which cover more than one text line in height) renders the approach independent to the actual layout of the manuscript.

Information about the localtions and spatial relationships is not required. Thus, the method is suitable for ancient manuscripts which are subject to variations in layout and loose layout rules. Fluctuating text lines and skewed text blocks do not influence the features due to rotation invariation. Thus, the method is additionally independent of the writing direction of the manuscript.

In summary, the method proposed is suitable for ancient handwritten documents with variations in layout, writing style and being degraded. 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 well.

6.3 Future Work

Future work includes improvements of the localization algorithm for decorative entities by reducing the number of undetected elements.

Additionally, identifying the actual object class of a decorative entity – since all decorative entities are considered as one class in the method proposed – is an issue. Thus, a distinction between decorative initials, small initials and headings has to be accomplished. Hereby, spatial relationships such as the fact that headings are located above and right of an embellished initial, initials embedded in the body text are indicated by punctuation or large blank spaces before the initial, or single characters outside the main body text are plain initials despite the fact that their local characteristics are similar to the main body text.

Furthermore, a text line extraction method needs to be applied in order to determine the actual lines of text within the text regions. Employing interest points to extract local features, these interest points can 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 as only few interest points are generated for background areas between the text lines. A possible solution to this issue is to be published in [55].

List of Acronyms

ACF	Auto Correlation Function
ASIFT	Affine Scale Invariant Feature Transform
BPN	Back-Propagation Network
CBIR	Content Based Image Retrieval
CC	Connected Component
CMRM	Cross-Media Relevance Model
CSIFT	Color Scale Invariant Feature Transform
DOG	Difference-of-Gaussian
DOH	Determinant-of-Hessian
FAST	Features from Accelerated Segment Test
FFN	Feed-Forward Neural Network
GLCM	Gray Level Co-occurence Matrix
GLOH	Gradient Location-Orientation Histogram
k-NN	k-Nearest Neighbor
LDM	Local Dissimilarity Map
LMT	Logistic Model Tree
LOG	Laplacian-of-Gaussians
LPP	Local Projection Profiles
MLP	Multi-Layer Perceptron
MOCR	Multilingual Optical Character Recognition
MST	Minimum Spanning Tree
NN	Neural Networks

Ocr	Optical Character Recognition
Olpp	Oriented Local Projection Profiles
Pca	Principal Component Analysis
Pga	Pairwise Geometric Attribute
PP	Projection Profiles
Rbf	Radial Basis Function
Rift	Rotation Invariant Feature Transform
Rle	Run Length Encoding
Roi	Region Of Interest
Sift	Scale Invariant Feature Transform
Surf	Speeded Up Robust Features
Susan	Smallest Univalue Segment Assimilating Nucleus
Svm	Support Vector Machine

Bibliography

[1] Aala E. Abdel-Hakim and Aly A. Farag. CSIFT: A SIFT Descriptor with Color Invariant Characteristics. In *Proceedings of the IEEE Computer Society Computer Vision and Pattern Recognition*, volume 2, pages 1978 – 1983, 2006.

[2] S. Abirami and D. Manjula. A Survey of Script Identification Techniques for Multi-Script Document Images. *International Journal of Recent Trends in Engineering*, 1(2):246–249, May 2009.

[3] M. A. Aizerman, E. A. Braverman, and L. Rozonoer. Theoretical Foundations of the Potential Function Method in Pattern Recognition Learning. *Automation and Remote Control*, 25:821–837, June 1964.

[4] Apostolos Antonacopoulos and Andy Downton. Special Issue on the Analysis of Historical Documents. *International Journal on Document Analysis and Recognition*, 9:75–77, 2007.

[5] Apostolos Antonacopoulos and Dimosthenis Karatzas. Document Image Analysis for World War II Personal Records. In *Proceedings of the International Workshop on Document Image Analysis for Libraries*, pages 336 – 341, 2004.

[6] Manivannan Arivazhagan, Harish Srinivasan, and Sargur Srihari. A Statistical Approach to Line Segmentation in Handwritten Documents. In Xiaofan Lin and Berrin A. Yanikoglu, editors, *Document Recognition and Retrieval XIV*, volume 6500. SPIE, 2007.

[7] Jaume Bacardit, Michael Stout, Natalio Krasnogor, Jonathan D. Hirst, and Jacek Blazewicz. Coordination Number Prediction Using Learning Classifier Systems: Performance and Interpretability. In *Proceedings of the Conference on Genetic and Evolutionary Computation*, GECCO '06, pages 247–254, New York, NY, USA, 2006. ACM.

[8] Micheal Baechler, Jean-Luc Bloechle, and Rolf Ingold. Semi-Automatic Annotation Tool for Medieval Manuscripts. In *Proceedings of the International Conference on Frontiers in Handwriting Recognition*, pages 182–187, Los Alamitos, CA, USA, 2010. IEEE Computer Society.

[9] Shumeet Baluja and Michele Covell. Finding Images and Line-Drawings in Document-Scanning Systems. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 1096–1100, July 2009.

[10] Itay Bar-Yosef, Nate Hagbi, Klara Kedem, and Its'hak Dinstein. Line Segmentation for Degraded Handwritten Historical Documents. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 1161 –1165, 2009.

[11] Itay Bar-Yosef, Klara Kedem, Its'hak Dinstein, Malachi Beit-Arie, and Edna Engel. Classification of Hebrew Calligraphic Handwriting Styles: Preliminary Results. In *Proceedings of the International Workshop on Document Image Analysis for Libraries*, pages 299 – 305, 2004.

[12] Etienne Baudrier, Sébastien Busson, Silvio Corsini, Mathieu Delalandre, Jérôme Landré, and Frédéric Morain-Nicolier. Retrieval of the Ornaments from the Hand-Press Period: An Overview. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 496 –500, July 2009.

[13] Etienne Baudrier, Frédéric Nicolier, Gilles Millon, and Su Ruan. Binary-Image Comparison with Local-Dissimilarity Quantification. *Pattern Recognition Letters*, 41:1461–1478, May 2008.

[14] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc J. Van Gool. Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding*, 110(3):346 – 359, 2008.

[15] Herbert Bay, Tinne Tuytelaars, and Luc J. Van Gool. SURF: Speeded Up Robust Features. In *Proceedings of the European Conference on Computer Vision*, pages 404–417, 2006.

[16] Paul R. Beaudet. Rotationally Invariant Image Operators. In *Proceedings of the International Joint Conference on Pattern Recognition*, pages 579–583, 1978.

[17] Andrea Bedö. Medizinische Fachprosa im Stift Klosterneuburg bis 1500. Ein Beitrag zur Geschichte der Wissenschaftspflege im mittelalterlichen Niederösterreich. Ungedruckte Staatsprüfungsarbeit am Institut für Österreichische Geschichtsforschung. Wien, 1989.

[18] Abdel Belaid. Computer Aided Design of Models of Page for Their Use in Recognition of Documents. In *Workshop one Electronic Page Models*, 1997.

[19] Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape Matching and Object Recognition Using Shape Contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, 2002.

[20] Josef Bigun, Sushil K. Bhattacharjee, and S. Michel. Orientation radiograms for Image Retrieval: An Alternative to Segmentation. In *Proceedings of the International Conference on Pattern Recognition*, volume 3, pages 346 –350, August 1996.

[21] Bernhard E. Boser, Isabelle Guyon, and Vladimir Vapnik. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the Workshop on Computational Learning Theory*, pages 144–152, 1992.

[22] Matthew Brown and David G. Lowe. Invariant Features from Interest Point Groups. In Paul L. Rosin and A. David Marshall, editors, *Proceedings of the British Machine Vision Conference*, pages 656 – 665, 2002.

[23] Marius Bulacu, Rutger van Koert, Lambert Schomaker, and Tijn van der Zant. Layout Analysis of Handwritten Historical Documents for Searching the Archive of the Cabinet of the Dutch Queen. In *Proceedings of the International Conference on Document Analysis and Recognition*, volume 1, pages 357–361, 2007.

[24] Evgeny Byvatov, Uli Fechner, Jens Sadowski, and Gisbert Schneider. Comparison of Support Vector Machine and Artificial Neural Network Systems for Drug/Nondrug Classification. *Journal of Chemical Information and Computer Sciences*, 43(6):1882–1889, 2003.

[25] Bilson J. L. Campana and Eamonn J. Keogh. A Compression-Based Distance Measure for Texture. *Journal of Statistical Analysis and Data Mining*, 3(6):381–398, 2010.

[26] John Canny. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, November 1986.

[27] Peter Carbonetto, Gyuri Dorkó, Cordelia Schmid, Hendrik Kück, and Nando de Freitas. A Semi-supervised Learning Approach to Object Recognition with Spatial Integration of Local Features and Segmentation Cues. In *Toward Category-Level Object Recognition*, pages 277–300, 2006.

[28] Nawei Chen and Dorothea Blostein. A Survey of Document Image Classification: Problem Statement, Classifier Architecture and Performance Evaluation. *International Journal on Document Analysis and Recognition*, 10:1–16, 2007.

[29] Victor Chen, Agota Szabo, and Michel Roussel. Recherche d' Images Iconique Utilisant les Moments de Zernike. In *Compression et Représentation des Signaux Audiovisuels*, number 13, 2003.

[30] Jia-Lin Chenxe. A Simplified Approach to the HMM-Based Texture Analysis and its Application to Document Segmentation. *Pattern Recognition Letters*, 18(10):993 – 1007, 1997.

[31] Dmitry Chetverikov. Pattern Regularity as a Visual Key. *Image and Vision Computing*, 18(12):975–985, 2000.

[32] Dmitry Chetverikov, Jisheng Liang, Jozsef Komuves, and Robert M. Haralick. Zone Classification Using Texture Features. In *Proceedings of the International Conference on Pattern Recognition*, volume 3, pages 676 – 680, 1996.

[33] Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine Learning*, 20:273–297, 1995.

[34] Mickaël Coustaty, Jean-Marc Ogier, Rudolf Pareti, and Nicole Vincent. Drop Caps Decomposition for Indexing a New Letter Extraction Method. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 476–480, 2009.

[35] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, New York, NY, USA, 2000.

[36] Mathieu Delalandre, Jean-Marc Ogier, and Josep Lladós. *Graphics Recognition. Recent Advances and New Opportunities*, chapter A Fast CBIR System of Old Ornamental Letter, pages 135–144. Springer-Verlag, Berlin, Heidelberg, 2008.

[37] Markus Diem and Robert Sablatnig. Recognition of Degraded Handwritten Characters Using Local Features. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 221–225, 2009.

[38] Markus Diem and Robert Sablatnig. Recognizing Characters of Ancient Manuscripts. In *Proceedings of the IS&T SPIE Conference on Computer Image Analysis in the Study of Art*, volume 7531, 2010.

[39] Gyuri Dorkó and Cordelia Schmid. Selection of Scale-Invariant Parts for Object Class Recognition. In *Proceedings of the International Conference on Computer Vision*, pages 634–640, 2003.

[40] Sloven Dubois, Mathieu Lugiez, Renaud Péteri, and Michel Ménard. Adding a Noise Component to a Color Decomposition Model for Improving Color Texture Extraction. In IS&T (The Society for Imaging Science and Technology), editors, *Proceedings of the European Conference on Colour in Graphics, Imaging, and Vision*, pages 394–398, Barcelona Spain, June 2008.

[41] Yves Dufournaud, Cordelia Schmid, and Radu Horaud. Matching Images with Different Resolutions. In *Proceedings of the IEEE Computer Society Computer Vision and Pattern Recognition*, volume 1, pages 612 – 618, 2000.

[42] Christopher Evans. Notes on the OpenSURF Library. Technical Report CSTR-09-001, University of Bristol, January 2009.

[43] Robert Fergus, Pietro Perona, and Andrew Zisserman. Object Class Recognition by Unsupervised Scale-Invariant Learning. In *Proceedings of the IEEE Computer Society Computer Vision and Pattern Recognition*, pages 264–271, 2003.

[44] Vittorio Ferrari, Tinne Tuytelaars, and Luc J. Van Gool. Simultaneous Object Recognition and Segmentation by Image Exploration. In Tomas Pajdla and Jiri Matas, editors, *Proceedings of the European Conference on Computer Vision*, volume 3021 of *Lecture Notes in Computer Science*, pages 40–54. Springer Berlin / Heidelberg, 2004.

[45] Luc Florack, Bart Ter Haar Romeny, Jan Koenderink, and Max Viergever. General Intensity Transformations and Differential Invariants. *Journal of Mathematical Imaging and Vision*, 4:171–187, 1994.

[46] Luc Florack, Bart Ter Haar Romeny, Max Viergever, and Jan Koenderink. The Gaussian Scale-Space Paradigm and the Multiscale Local Jet. *International Journal of Computer Vision*, 18:61–75, 1996.

[47] Wolfgang Förstner. A Framework for Low Level Feature Extraction. In Jan-Olof Eklundh, editor, *Proceedings of the European Conference on Computer Vision*, volume 801 of *Lecture Notes in Computer Science*, pages 383–394. Springer Berlin / Heidelberg, 1994.

[48] Wolfgang Förstner and Eberhard Gülch. A Fast Operator for Detection and Precise Location of Distinct Points, Corners and Centres of Circular Features. In *Proceedings of the Intercommission Conference on Fast Processing of Photogrammetric Data*, pages 281 – 305, Interlaken, Switzerland, 1987.

[49] William T. Freeman and Edward H. Adelson. The Design and Use of Steerable Filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906, 1991.

[50] Robert Fuchs. *Pergament: Geschichte, Material, Konservierung, Restaurierung, Kölner Beiträge zur Restaurierung und Konservierung von Kunst- und Kulturgut, Band 12*, chapter Pergament - Material, Geschichte, Restaurierung, pages 9–110. Anton Siegl Fachbuchhandlung, 2001.

[51] Dennis Gabor. Theory of Communication. Part 1: The Analysis of Information. *Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering*, 93(26):429 –441, 1946.

[52] Angelika Garz, Markus Diem, and Robert Sablatnig. Detecting Text Areas and Decorative Elements in Ancient Manuscripts. In Patrick Kellenberger, editor, *Proceedings of the 12th International Conference on Frontiers in Handwriting Recognition*, pages 176–181. IEEE Computer Society, 2010.

[53] Angelika Garz, Markus Diem, and Robert Sablatnig. Local Descriptors for Document Layout Analysis. In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Ronald Chung, Riad Hammound, Muhammad Hussain, Tan Kar-Han, Roger Crawfis, Daniel Thalmann, David Kao, and Lisa Avila, editors, *Advances in Visual Computing*, volume 6455 of *Lecture Notes in Computer Science*, pages 29–38. Springer, 2010.

[54] Angelika Garz, Markus Diem, and Robert Sablatnig. Layout Analysis of Ancient Manuscripts Using Local Features. *Eikonopoiia. Digital Imaging of Ancient Textual Inheritage (Commentationes Humanarum Litterarum 129)*, 2011 (forthcoming). forthcoming.

[55] Angelika Garz, Robert Sablatnig, and Markus Diem. Layout Analysis for Historic Manuscripts Using SIFT Features. In *Proceedings of the International Conference on Document Analysis and Recognition*, Beijing, China, 2011 (forthcoming).

[56] Polina Golland, W. Eric L. Grimson, Martha Shenton, and Ron Kikinis. Deformation Analysis for Shape Based Classification. In Michael Insana and Richard Leahy, editors, *Information Processing in Medical Imaging*, volume 2082 of *Lecture Notes in Computer Science*, pages 517–530. Springer Berlin / Heidelberg, 2001.

[57] Luc J. Van Gool, Theo Moons, and Dorin Ungureanu. Affine/ Photometric Invariants for Planar Intensity Patterns. In *Proceedings of the European Conference on Computer Vision*, volume 1, pages 642–651, London, UK, 1996. Springer-Verlag.

[58] Antonia Graf. Zur Geschichte der Fleuronnéeinitiale. Unter besonderer Berücksichtigung der österreichischen Handschriften des 13. und 14. Jahrhunderts. Ungedruckte Dissertation. Wien, 1968.

[59] Costantino Grana, Daniele Borghesani, and Rita Cucchiara. Automatic Segmentation of Digitalized Historical Manuscripts. *Multimedia Tools and Applications*, pages 1 – 24, 2010.

[60] Robert M. Haralick. Document Image Understanding: Geometric and Logical Layout. In *Proceedings of the IEEE Computer Society Computer Vision and Pattern Recognition*, pages 385 –390, June 1994.

[61] Robert M. Haralick, K. Shanmugam, and Its'hak Dinstein. Textural Features for Image Classification. *IEEE Transactions on Systems, Man and Cybernetics*, 3(6):610 –621, November 1973.

[62] Chris Harris and Mike Stephens. A Combined Corner and Edge Detector. In *Proceedings of the ALVEY Vision Conference*, pages 147–151, 1988.

[63] Marti A. Hearst, Susan T. Dumais, Edgar Osuna, John Platt, and Bernhard Schölkopf. Support Vector Machines. *IEEE Journal on Intelligent Systems and their Applications*, 13(4):18 –28, 1998.

[64] Radu Horaud, Francoise Veillon, and Thomas Skordas. Finding Geometric and Relational Structures in an Image. In O. Faugeras, editor, *Proceedings of the European Conference on Computer Vision*, volume 427 of *Lecture Notes in Computer Science*, pages 374–384. Springer Berlin / Heidelberg, 1990.

[65] Thorsten Joachims. Text Categorization with Support Vector Machines: Learning with many Relevant Features. In Claire Nedellec and Celine Rouveirol, editors, *Proceedings of the European Conference on Machine Learning*, volume 1398 of *Lecture Notes in Computer Science*, pages 137–142. Springer Berlin / Heidelberg, 1998.

[66] Andrew Edie Johnson and Martial Hebert. Recognizing Objects by Matching Oriented Points. In *Proceedings of the IEEE Computer Society Computer Vision and Pattern Recognition*, pages 684–689, 1997.

[67] Nicholas Journet, Véronique Eglin, Jean-Yves Ramel, and Rémy Mullot. Text/Graphic Labelling of Ancient Printed Documents. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 1010–1014, 2005.

[68] Nicholas Journet, Rémy Mullot, Jean-Yves Ramel, and Veronique Eglin. Ancient Printed Documents Indexation: A New Approach. In *Pattern Recognition and Data Mining*, pages 580–589. Springer, 2005.

[69] Nicholas Journet, Jean-Yves Ramel, Rémy Mullot, and Véronique Eglin. Document Image Characterization Using a Multiresolution Analysis of the Texture: Application to Old Documents. *International Journal on Document Analysis and Recognition*, 11(1):9–18, October 2008.

[70] Ali Karray, Jean-Marc Ogier, Slim Kanoun, and Mohamed Alimi. An Ancient Graphic Documents Indexing Method Based on Spatial Similarity. In *Graphics Recognition. Recent Advances and New Opportunities*, volume 5046 of *Lecture Notes in Computer Science*, pages 126–134. Springer Berlin / Heidelberg, 2008.

[71] Yan Ke and Rahul Sukthankar. PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. In *Proceedings of the IEEE Computer Society Computer Vision and Pattern Recognition*, pages 506–513, 2004.

[72] Badreddine Khelifi, Nizar Zaghden, Adel M. Alimi, and Rémy Mullot. Unsupervised Categorization of Heterogeneous Text Images Based on Fractals. In *Proceedings of the International Conference on Pattern Recognition*, pages 1–4, December 2008.

[73] Koichi Kise, Akinori Sato, and Motoi Iwata. Segmentation of Page Images Using the Area Voronoi Diagram. *Computer Vision and Image Understanding*, 70:370–382, June 1998.

[74] Florian Kleber, Robert Sablatnig, Melanie Gau, and Heinz Miklas. Ancient Document Analysis Based on Text Line Extraction. In *Proceedings of the International Conference on Pattern Recognition*, pages 1–4, 2008.

[75] Jan Koenderink and Andrea van Doom. Representation of Local Geometry in the Visual System. *Biological Cybernetics*, 55(6):367–375, 1987.

[76] Ron Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 2, pages 1137–1143, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[77] Ajay Kumar and David Zhang. Personal Recognition Using Hand Shape and Texture. *IEEE Transactions on Image Processing*, 15(8):2454 –2461, 2006.

[78] Christoph H. Lampert, Matthew B. Blaschko, and Thomas Hofmann. Efficient Subwindow Search: A Branch and Bound Framework for Object Localization.

IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(12):2129–2142, 2009.

[79] Jérôme Landré, Frédéric Morain-Nicolier, and Su Ruan. Ornamental Letters Image Classification Using Local Dissimilarity Maps. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 186–190, 2009.

[80] Svetlana Lazebnik, Cordeila Schmid, and Jean Ponce. A Sparse Texture Representation Using Local Affine Regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1265 –1278, 2005.

[81] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. A Sparse Texture Representation Using Affine-Invariant Regions. In *Proceedings of the IEEE Computer Society Computer Vision and Pattern Recognition*, pages 319–326, 2003.

[82] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Semi-Local Affine Parts for Object Recognition. In *Proceedings of the British Machine Vision Conference*, 2004.

[83] Frank Le Bourgeois and Hala Kaileh. Automatic Metadata Retrieval from Ancient Manuscripts. In *Proceedings of the IAPR International Workshop on Document Analysis Systems*, pages 75–89, 2004.

[84] Ming-Chang Lee and To Chang. Comparison of Support Vector Machine and Back Propagation Neural Network in Evaluating the Enterprise Financial Distress. *International Journal of Artificial Intelligence & Applications*, 1(3):31–43, July 2010.

[85] Seong-Whan Lee and Dae-Seok Ryu. Parameter-Free Geometric Document Layout Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1240 –1256, November 2001.

[86] Vincent Lepetit and Pascal Fua. Keypoint Recognition Using Randomized Trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1465 –1479, 2006.

[87] Jiseng Liang, Jaekyu Ha, Robert M. Haralick, and Ihsin T. Phillips. Document Layout Structure Extraction Using Bounding Boxes of Different Entitles. In *Proceedings of the IEEE Workshop on Applications of Computer Vision*, pages 278 –283, December 1996.

[88] Laurence Likforman-Sulem, Abderrazak Zahour, and Bruno Taconet. Text Line Segmentation of Historical Documents: A Survey. *International Journal on Document Analysis and Recognition*, 9(2):123–138, April 2007.

[89] Tony Lindeberg. Detecting Salient Blob-Like Image Structures and Their Scales with a Scale-Space Primal Sketch: A Method for Focus-Of-Attention. *International Journal of Computer Vision*, 11:283–318, 1993.

[90] Tony Lindeberg. Scale-Space Theory: A Basic Tool for Analysing Structures at Different Scales. *Journal of Applied Statistics*, 21(2):224–270, 1994.

[91] Tony Lindeberg. *Wiley Encyclopedia of Computer Science and Engineering*, chapter Scale-Space, pages 2495–2504. John Wiley & Sons, Inc., 2008.

[92] David G. Lowe. Object Recognition from Local Scale-Invariant Features. In *Proceedings of the International Conference on Computer Vision*, pages 1150–1157, 1999.

[93] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[94] Song Mao, Azriel Rosenfeld, and Tapas Kanungo. Document Structure Analysis Algorithms: A Literature Survey. In Tapas Kanungo, Elisa, Jianying Hu, and Paul B. Kantor, editors, *Document Recognition and Retrieval X*, volume 5010/1, pages 197–207. SPIE, 2003.

[95] Heinz Miklas, editor. *Glagolitica - Zum Ursprung der Slavischen Schriftkultur*. Verlag der Österreichischen Akademie der Wissenschaften, 2000.

[96] Heinz Miklas, Melanie Gau, Florian Kleber, Markus Diem, Martin Lettner, Maria Vill, Robert Sablatnig, Manfred Schreiner, Michael Melcher, and Ernst-Georg 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.

[97] Krystian Mikolajczyk. *Detection of Local Features Invariant to Affine Transformations*. PhD thesis, Institut National Polytechnique de Grenoble, France, 2002.

[98] Krystian Mikolajczyk, Bastian Leibe, and Bernt Schiele. Multiple Object Class Detection with a Generative Model. In *Proceedings of the IEEE Computer Society Computer Vision and Pattern Recognition*, pages 26–36, 2006.

[99] Krystian Mikolajczyk and Cordelia Schmid. Indexing Based on Scale Invariant Interest Points. In *Proceedings of the International Conference on Computer Vision*, pages 525–531, 2001.

[100] Krystian Mikolajczyk and Cordelia Schmid. Scale & Affine Invariant Interest Point Detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[101] Krystian Mikolajczyk and Cordelia Schmid. A Performance Evaluation of Local Descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[102] Krystian Mikolajczyk, Tinne Tuytelaars, Cordelia Schmid, Andrew Zisserman, Jiri Matas, Frederik Schaffalitzky, Timor Kadir, and Luc J. Van Gool. A Comparison of Affine Region Detectors. *International Journal of Computer Vision*, 65(1-2):43–72, 2005.

[103] Krystian Mikolajczyk, Andrew Zisserman, and Cordeila Schmid. Shape Recognition with Edge-Based Features. In *Proceedings of the British Machine Vision Conference*, volume 2, pages 779–788, 2003.

[104] Ikram Moalla, Frank LeBourgeois, Hubert Emptoz, and Adel Alimi. Contribution to the Discrimination of the Medieval Manuscript Texts: Application in the Palaeography. In Horst Bunke and A. Spitz, editors, *Document Analysis Systems VII*, volume 3872 of *Lecture Notes in Computer Science*, pages 25–37. Springer Berlin / Heidelberg, 2006.

[105] Ikram Moalla, Frank LeBourgeois, Hubert Emptoz, and Adel Alimi. Image Analysis for Palaeography Inspection. In *Proceedings of the International Workshop on Document Image Analysis for Libraries*, pages 8 – 311, April 2006.

[106] Natasha Mohanty, Toni Rath, Audrey Lee, and Raghavan Manmatha. Learning Shapes for Image Classification and Retrieval. In Wee-Kheng Leow, Michael Lew, Tat-Seng Chua, Wei-Ying Ma, Lekha Chaisorn, and Erwin Bakker, editors, *Image and Video Retrieval*, volume 3568 of *Lecture Notes in Computer Science*, pages 591–591. Springer Berlin / Heidelberg, 2005.

[107] Hans P. Moravec. Rover Visual Obstacle Avoidance. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pages 785–790, 1981.

[108] Jean-Michel Morel and Guoshen Yu. ASIFT: A New Framework for Fully Affine Invariant Image Comparison. *SIAM Journal on Imaging Sciences*, 2:438–469, April 2009.

[109] George Nagy. Twenty Years of Document Image Analysis in PAMI. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:38–62, January 2000.

[110] Randal C. Nelson and Andrea Selinger. Large-Scale Tests of a Keyed, Appearance-Based 3-D Object Recognition System. *Vision Research*, 38(15-16):2469 – 2488, 1998.

[111] Jean-Marc Ogier and Karl Tombre. Madonne: Document Image Analysis Techniques for Cultural Heritage Documents. In *International Conference on Digital Cultural Heritage*, 2006.

[112] Oleg Okun, David Doermann, and Matti Pietikäinen. Page Segmentation and Zone Classification: The State of the Art. Technical Report LAMP-TR-036,CAR-TR-927,CS-TR-4079, University of Maryland, College Park, November 1999.

[113] Oleg Okun and Matti Pietikäinen. A Survey of Texture-Based Methods for Document Layout Analysis. In *Proceedings of the Workshop on Texture Analysis in Machine Vision*, pages 137–148, June 2000.

[114] Rudolf Pareti, Surapong Uttama, Jean-Pierre Salmon, Jean-Marc Ogier, Salvatore Tabbone, Laurent Wendling, Sébastien Adam, and Nicole Vincent. On Defining Signatures for the Retrieval and the Classification of Graphical Drop Caps. In *Proceedings of the International Workshop on Document Image Analysis for Libraries*, 2006.

[115] Jean-Yves Ramel, Stéphane Leriche, Marie-Luce Demonet, and Sébastien Busson. User-Driven Page Layout Analysis of Historical Printed Books. *International Journal on Document Analysis and Recognition*, 9(2-4):243–261, 2007.

[116] Sheikh Faisal Rashid, Faisal Shafait, and Thomas Breuel. Connected Component Level Multiscript Identifiation from Ancient Document Images. In *Proceedings of the IAPR International Workshop on Document Analysis Systems*, 2010.

[117] Edward Rosten and Tom Drummond. Fusing Points and Lines for High Performance Tracking. In *Proceedings of the International Conference on Computer Vision*, volume 2, pages 1508–1515, Oct 2005.

[118] Edward Rosten and Tom Drummond. Machine Learning for High-Speed Corner Detection. In A. Leonardis, H. Bischof, and A. Pinz, editors, *Proceedings of the European Conference on Computer Vision*, volume 3951 of *Lecture Notes in Computer Science*, pages 430–443. Springer Berlin / Heidelberg, 2006.

[119] Cullen Schaffer. A Conservation Law for Generalization Performance. In W. W. Cohen and H. Hirsch, editors, *Proceedings of the International Machine Learning Conference*, pages 259–265. Rutgers University, New Brunswick, NJ, 1994.

[120] Cordelia Schmid and Roger Mohr. Local Grayvalue Invariants for Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–535, 1997.

[121] Cordelia Schmid, Roger Mohr, and Christian Bauckhage. Evaluation of Interest Point Detectors. *International Journal of Computer Vision*, 37:151–172, 2000.

[122] Bernhard Schölkopf, Kah-Kay Sung, Chris J.C. Burges, Federico Girosi, Partha Niyogi, Tomaso Poggio, and Vladimir Vapnik. Comparing Support Vector Machines with Gaussian Kernels to Radial Basis Function Classifiers. *IEEE Transactions on Signal Processing*, 45(11):2758 –2765, November 1997.

[123] Faisal Shafait, Joost van Beusekom, Daniel Keysers, and Thomas Breuel. Page Frame Detection for Marginal Noise Removal from Scanned Documents. In Bjarne Ersboll and Kim Pedersen, editors, *Image Analysis*, volume 4522 of *Lecture Notes in Computer Science*, pages 651–660. Springer Berlin / Heidelberg, 2007.

[124] Faisal Shafait, Joost van Beusekom, Daniel Keysers, and Thomas Breuel. Document Cleanup Using Page Frame Detection. *International Journal on Document Analysis and Recognition*, 11:81–96, 2008.

[125] Vladimir Shapiro, Georgi Gluchev, and Vassil Sgurev. Handwritten Document Image Segmentation and Analysis. *Pattern Recognition Letters*, 14:71–78, January 1993.

[126] Erez Shilat, Michael Werman, and Yoram Gdalyahn. Ridge's Corner Detection and Correspondence. In *Proceedings of the IEEE Computer Society Computer Vision and Pattern Recognition*, pages 976 –981, June 1997.

[127] Elisa H. Barney Smith. An Analysis of Binarization Ground Truthing. In *Proceedings of the IAPR International Workshop on Document Analysis Systems*, DAS '10, pages 27–34, New York, NY, USA, 2010. ACM.

[128] Stephen M. Smith and J. Michael Brady. SUSAN - A New Approach to Low Level Image Processing. *International Journal of Computer Vision*, 23(1):45–78, 1997.

[129] Nikolaos Stamatopoulos, Basilis Gatos, and Thodoris Georgiou. Page Frame Detection for Double Page Document Images. In *Proceedings of the IAPR International Workshop on Document Analysis Systems*, DAS '10, pages 401–408, New York, NY, USA, 2010. ACM.

[130] Mark O. Stitson and Jason Weston. Function Estimation using Support Vector Machines. In *DIMACS Workshop Exploring Large Data Sets Using Classification, Consensus, and Pattern Recognition Techniques*, 1997.

[131] Yuan Y. Tang, Seong-Whan Lee, and Ching Y. Suen. Automatic Document Processing: A Survey. *Pattern Recognition Letters*, 29(12):1931 – 1952, 1996.

[132] Yuan Yan Tang, Chang De Yan, and Ching Y. Suen. Document Processing for Automatic Knowledge Acquisition. *IEEE Transactions on Knowledge and Data Engineering*, 6:3–21, February 1994.

[133] Carlo Tomasi and Takeo Kanade. Detection and Tracking of Point Features. *Image Rochester NY*, Technical Report CMU-CS-91-132(April):1–22, 1991.

[134] Tinne Tuytelaars and Krystian Mikolajczyk. Local Invariant Feature Detectors: A Survey. *Foundation and Trends in Computer Graphics and Vision*, 3:177–280, July 2008.

[135] Surapong Uttama, Jean-Marc Ogier, and Pierre Loonis. Top-Down Segmentation of Ancient Graphical Drop Caps: Lettrines. In *Proceedings of the 6th IAPR International Workshop on Graphics Recognition*, pages 87–96, August 2005.

[136] C. J. van Rijsbergen. *Information Retrieval*. Butterworth, 2nd edition, 1979.

[137] Andreas Vesalius. *De humani corporis fabrica libri septem*. Oporinus, 1543.

[138] Shin-Ywan Wang and T. Yagasaki. Block Selection: A Method for Segmenting a Page Image of Various Editing Styles. In *Proceedings of the International Conference on Document Analysis and Recognition*, volume 1, pages 128 –133 vol.1, August 1995.

[139] Nigel Wilson. Archimedes: The Palimpsest and the Tradition. *Byzantinische Zeitschrift*, 2:89–101, 1999.

[140] Andrew P. Witkin. Scale-Space Filtering. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1019–1022, Karlsruhe, Germany, August 1983.

[141] David H. Wolpert and R. Waters. The Relationship between PAC, the Statistical Physics Framework, the Bayesian Framework, and the VC Framework. In *Proceedings of the The Mathematics of Generalization: The SFI/CNLS Workshop on Formal Approaches to Supervised Learning*, volume XX, pages 117–214. Addison-Wesley, 1994.

[142] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey McLachlan, Angus Ng, Bing Liu, Philip Yu, Zhi-Hua Zhou, Michael Steinbach, David Hand, and Dan Steinberg. Top 10 Algorithms in Data Mining. *Knowledge and Information Systems*, 14:1–37, 2008.

[143] Guoshen Yu and Jean-Michel Morel. A Fully Affine Invariant Image Comparison Method. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1597 –1600, 2009.

[144] Ramin Zabih and John Woodfill. Non-Parametric Local Transforms for Computing Visual Correspondence. In Jan-Olof Eklundh, editor, *Proceedings of the European Conference on Computer Vision*, volume 801 of *Lecture Notes in Computer Science*, pages 151–158. Springer Berlin / Heidelberg, 1994.

[145] Abderrazak Zahour, Bruno Taconet, Pascal Mercy, and Said Ramdane. Arabic Hand-Written Text-Line Extraction. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 281 – 285, 2001.

[146] Jianguo Zhang, Marcin Marszalek, Svetlana Lazebnik, and Cordelia Schmid. Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study. *International Journal of Computer Vision*, 73(2):213–238, 2007.