DOCUMENT IMAGE ANALYSIS

Florian Kleber

DOCUMENTS

- Thousands of UNIQUE documents are stored unpublished in archives
  - e.g. Birth/marriage/death records
- Digitizing of archives is expensive
- Physical access is necessary

Image Acquisition

- Crowd scanning (ScanTent)
- MS-Imaging
- Professional mass digitalization

Recognition


URKUNDEN
Nr.: 8882_8891

CURRENT TREND

- Hessian State Archive allows to use cameras in reading rooms (6.1.2017)
- State Archive of Austria allows to use cameras in reading rooms (1.9.2017)
- Italy allows to use cameras in archives and libraries (11.09.2017)

Conference of the head of the archive administration of the German federation state that archives plan a medium-term digitization of 5-10% of archives holdings.

CROWD SCANNING

READ

DOCSCAN AND SCANTENT

LOW COST MOBILE SCANNING SYSTEM

- Digital copies of historical documents are needed for the digital humanities
- Cameras of mobile phones allow ~330 dpi for 297x210mm (DIN A4)
- Take pictures yourself in the archives
  - No waiting time
  - No digitization service needed
  - No cost intensive equipment needed

Prototype I

https://readcoop.eu/scantent/

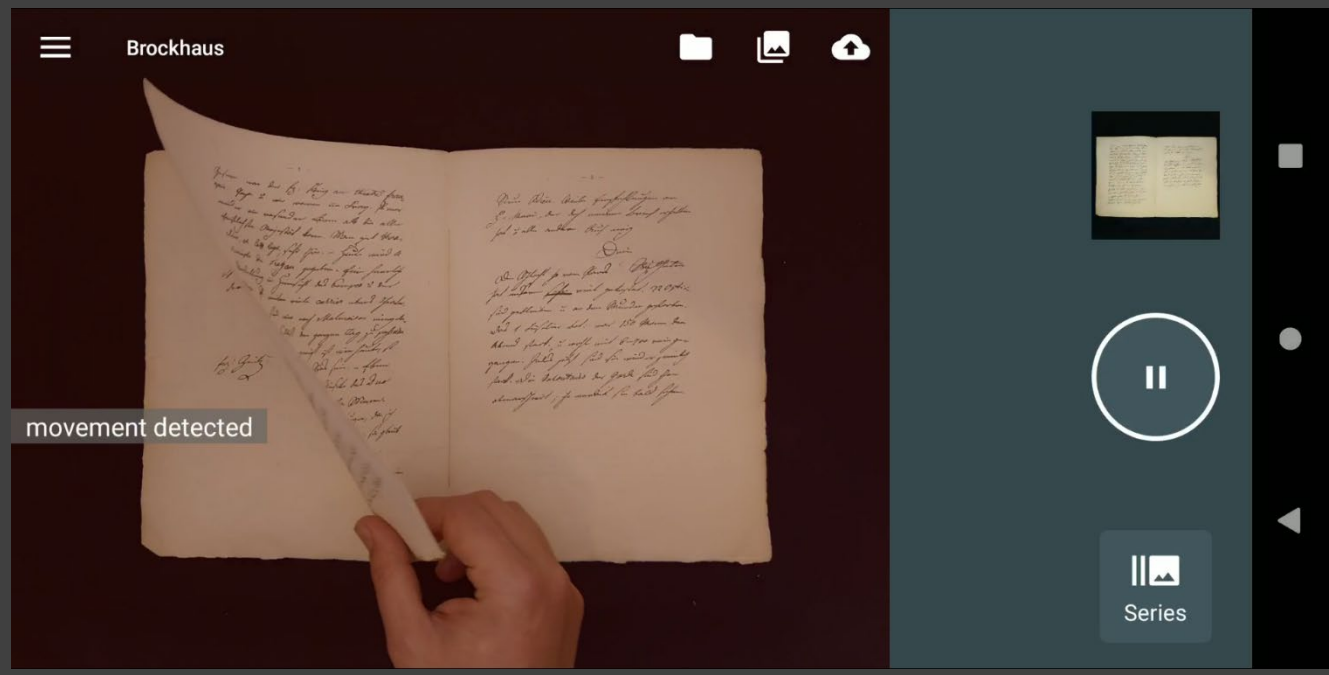
READ co·op

Transkribus ScanTent read&search The COOP Resources 0 Sign in EN

The ScanTent

Professional scanning with your smartphone

DOCSCAN



- Document scanner app
  - Focus Measure
  - Learns background model to detect page turns
  - Page detection
- Available on Google Playstore and Github (OpenSource)
  - https://github.com/TUWien/DocScan

GET IT ON Google Play

MSI

Image Enhancement

Cod. Sin. Slav. 5/N 65 recto

Independant Subspace Analysis

DOCUMENT ANALYSIS AND RECOGNITION

TABLE CLASSIFICATION

DOCUMENT ANALYSIS

writer identification

READING DOCUMENTS

(HTR)

GROUPING

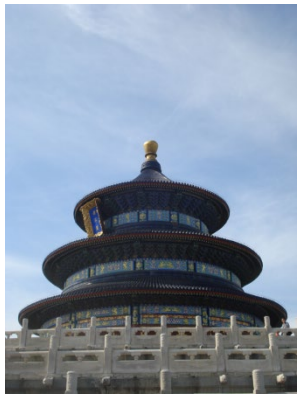
similar documents

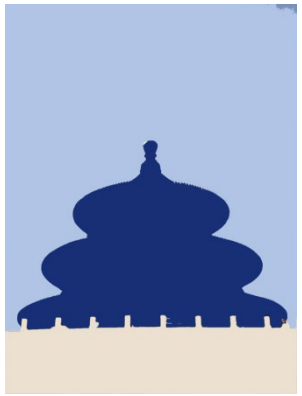
document categorization

LOGICAL LAYOUT ANALYSIS

Image Segmentation

- The objective of image segmentation is to group image pixels according to pre-defined rules.
- Segmentation is often considered to be the first step in image analysis.
  - The purpose is to subdivide an image into meaningful non-overlapping regions, which would be used for further analysis.
  - It is hoped that the regions obtained correspond to the physical parts or objects of a scene (3-D) represented by the image (2-D).





Robust Skew Estimation of Handwritten and Printed Documents based on Grayvalue Images

Florian Kleber, Markus Diem and Robert Sablatnig
Computer Vision Lab
Institute of Computer Aided Automation
Vienna University of Technology
Email: {kleber, diem, sab}@caa.tuwien.ac.at

Abstract—Skew estimation is a preprocessing step in document image analysis to determine the global dominant orientation of a document's text lines. A skew angle can be introduced during scanning, or if a document is photographed. The correction of the skew angle is necessary for further image analysis, to avoid an influence to the performance of skew sensitive methods, e.g. Optical Character Recognition (OCR) or page segmentation. The performance of current skew estimation methods is shown at the ICDAR2013 Document Image Skew Estimation Contest (DISEC), which uses a benchmark dataset of binarized printed documents with varying layouts and languages like English, Chinese or Greek. The proposed method is based on a Focused Nearest Neighbour Clustering (FNNC) of interest points and the analysis of paragraphs/lines and achieved rank 5 at the contest. In this paper it is shown, that the use of grayvalue images can outperform the results restricted to binarized images, thus the proposed method avoids the binarization step which is still an open research topic in document image analysis. The robustness of the method is also shown on a dataset comprising historical documents and on low resolution images. The method is evaluated on the DISEC dataset and three additional datasets (historical documents, low resolution documents, and machine printed documents).

I. Introduction

A digitization of documents is done on individual documents or in projects which are dealing with the mass digitalization of libraries or national archives like Google Books of Google Inc. or Improving Access to Text (IMPACT[1]). While mass digitalization projects use mainly scanners, single document images are also aquired using cameras of mobile devices (i.e. smart phones, digital cameras). Using mobile devices as aquisition device lead to a skew introduced by missing alignment mechanisms, and even in scanned documents a skew can be present [1].

A skew correction is a preprocessing step in Document Image Analysis (DIA) systems, since an introduced skew can affect the performance of DIA methods like Optical Character Recognition (OCR) or page segmentation [1], [2]. Papandreou et al. [1] state, that the threshold for the human perception is 0.1°. The first Document Image Skew Estimation Contest (DISEC) [1] was held within the International Conference on Document Analysis and Recognition (ICDAR) 2013 shows results of state-of-the-art skew estimation methods and that skew estimation is still an open research topic.

[1]www.impact-project.eu, accessed 20.12.2013

Skew estimation methods can be mainly categorized in methods based on Docstrum (nearest neighbour clustering) [3], [4], projection profiles [5], [6], [7], Hough transform [8], [9] and cross correlation [10], [11]. Fabrizio [1], the winner of the DISEC contest, uses the magnitude spectrum of the Fourier transform of clustered image regions. Carlinet and Fabrizio [1] (3rd place at DISEC) use a combination of a line detector and a clustering of connected components to which a Hough transform is applied. Hyung Il Koo [1] (2nd place at DISEC) uses a line detector and applies a maximum likelihood estimation to the lines detected. Beside a methodologoical categorisation, a classification of methods based on the document image requirements can be done, i.e. detectable angle range, document layout, type or size of fonts, image type (resolution, binary image). Epshtein [12] states, that current skew detection methods have to deal with all image classes due to images of mobile devices which are applied to mobile applications like Google Goggles or Microsoft iBing Vision. Although the benchmark dataset consists of binarized printed documents, which have varying layouts, fonts, languages and consist of different document types, e.g. newspapers, scientific books, comics, the skew angle has been restricted to ±15°.

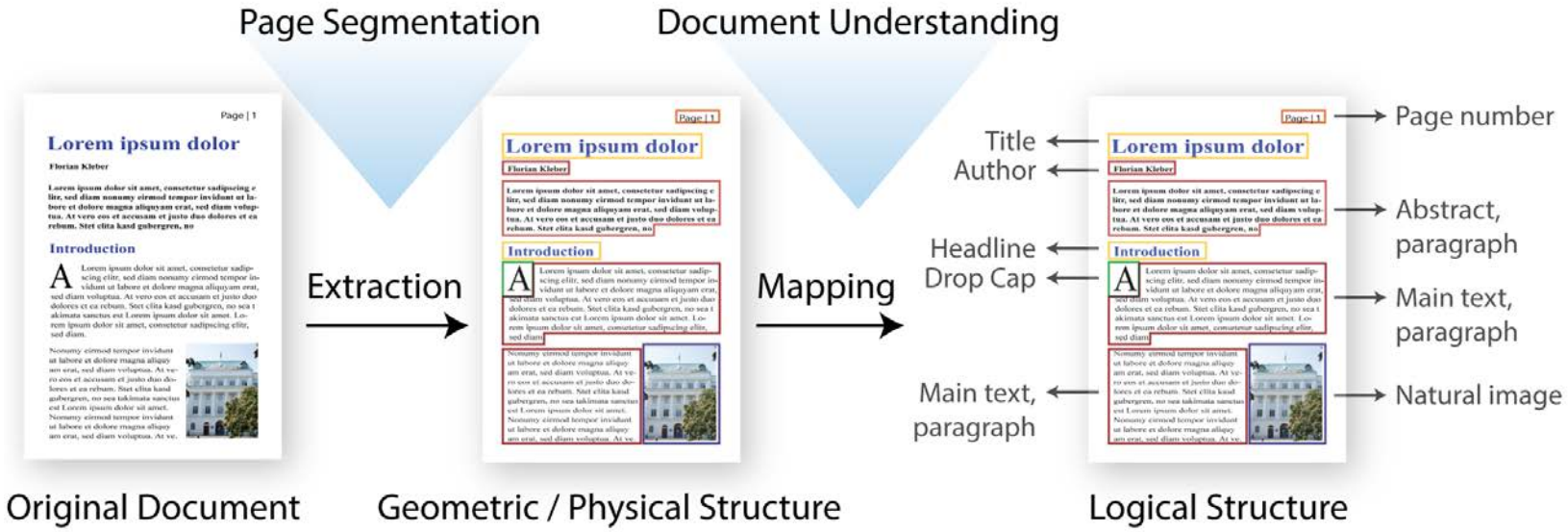
The proposed method (rank 5 at DISEC) is based on a Focused Nearest Neighbour Clustering (FNNC) of Interest Points (IP), which was originally introduced by Jiang et al. [13]. The IPs are determined as maxima of the Difference of Gaussian (DoG) [14]. At DISEC only binarized images have been used, to show results of skew estimation methods independent from the binarization method. Due to the use of a binarzation, information present in the grayvalue/color image is discarded. Thus, since the detection of IPs is binarization free, the proposed method can directly be applied to grayvalue images. It is shown, that the additional information of grayvalue images outperforms the use of binary images, especially on historical datasets comprising noise, e.g. bleedthrough text. The binarisation of historical documents is an open research topic, which has been shown during the last Document Image Binarisation Contests (DIBCO) [15]. Fig. 1 shows an example page with bleedthrough text of the dataset with historical documents. A combination of the proposed method with the gradient information for sparsely inscribed documents without an angle restriction has been presented by Diem et al. [2]. To achieve a higher accuracy compared to the basic FNNC clustering, the proposed approach is combined with a line detection and the analysis of paragraphs.

Title	Robust Skew Estimation of Handwritten and Printed…
Authors	Florian Kleber, Markus Diem and Robert Sablatnig
Author affiliation	Computer Vision Lab, Institute of Computer Aided …
Author email	{Kleber,diem,sab}@caa.tuwien.ac.at
Abstract Abstract Text	Abstract Skew estimation is a preprocessing step in document image analysis to determine the global dominant orientation of a document's text lines. A skew angle can be ….
Section Heading	1. Introduction
Section	A digitization of documents is done on individual documents or in projects which are dealing with the mass digitalization of libraries or national archives….
Footnote	[1]www.impact-project.eu, accessed 20.12.2013

Goals

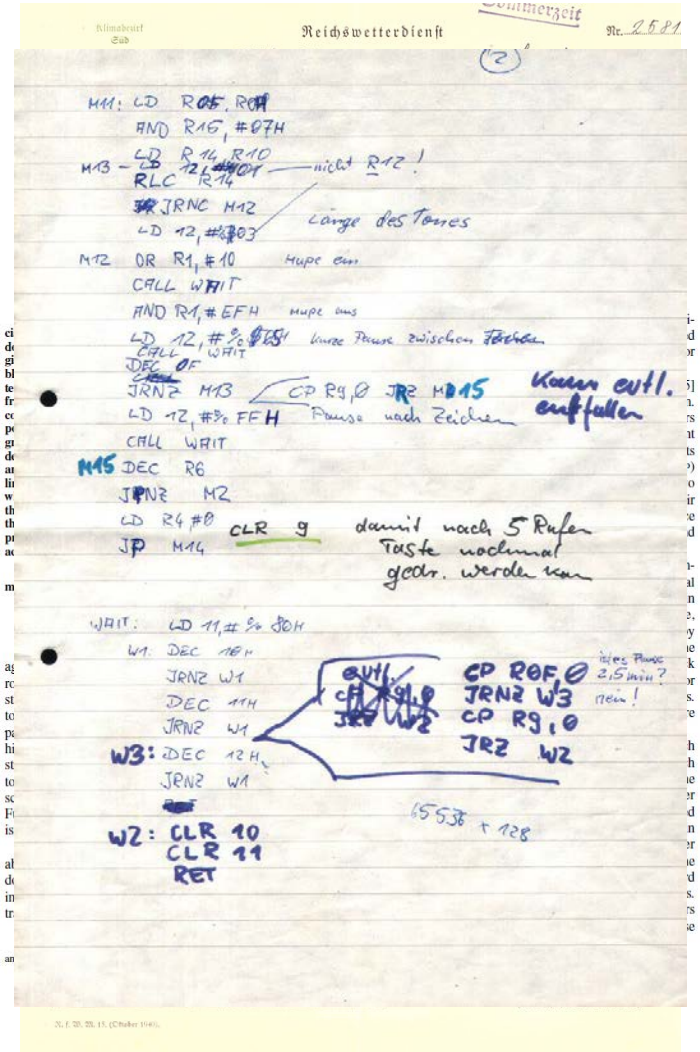
Document Layout Analysis vs. Understanding

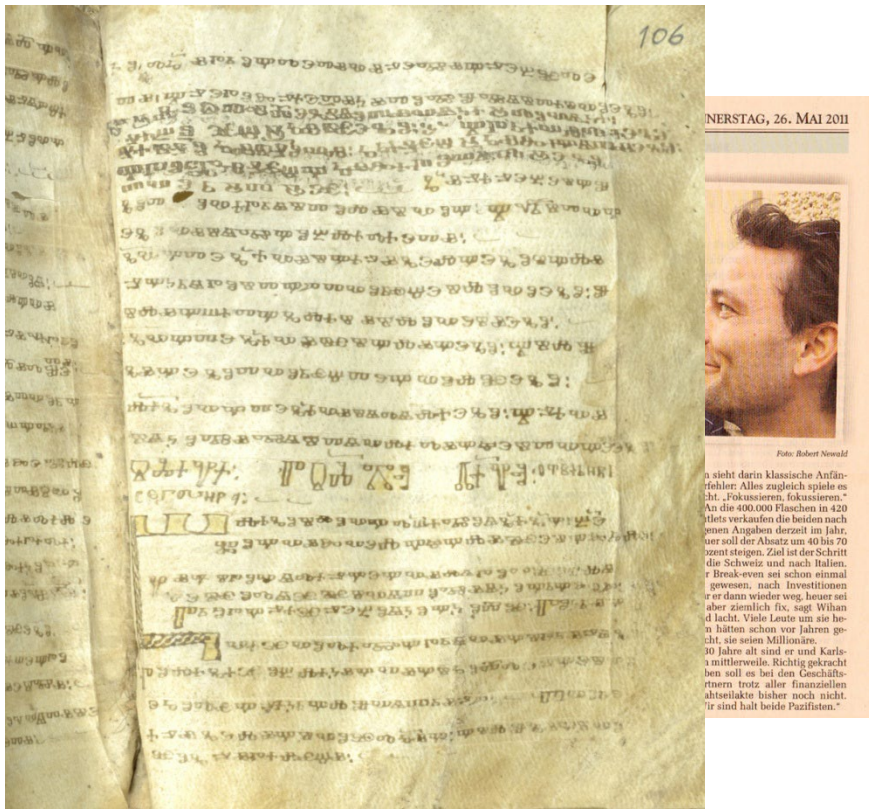
- Extract the layout structure *Layout Analysis*
  - Segment a page into homogeneous regions
- Map the logical structure *Document Understanding*
  - Label regions according to their function (e.g. author, title, abstract, footnote, ...)



Complexity: Restricted vs. Freeform Layouts

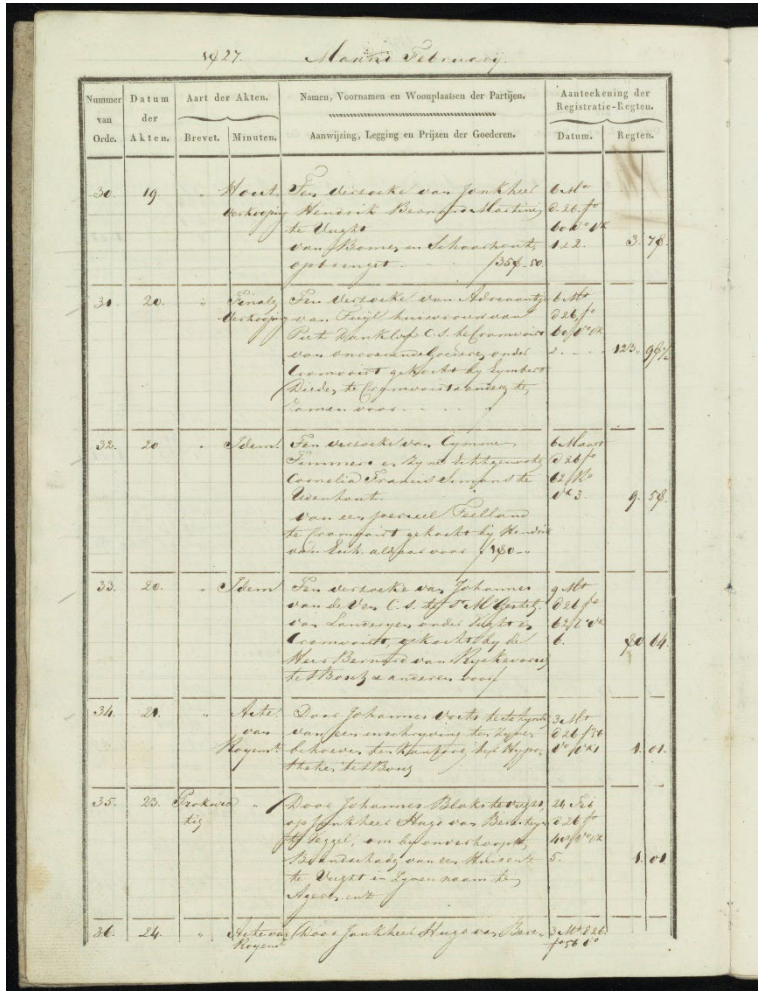
- Highly structu
  - Precisely defi
  - Defined by g
  - Regions of in
- Semi structure
  - Partially defi
  - Some regula
- Loosely structu
  - Cannot be de
  - Logical objec





Typical Applications – Form Processing

- Performance depends on
  - Form complexity
  - Form variability
- Fields are located easily
  - If their positions are fixed
  - When using different colors
- Challenges of content recognition
  - Degraded images
  - Approximate positioning of symbols
  - Variability of handwriting

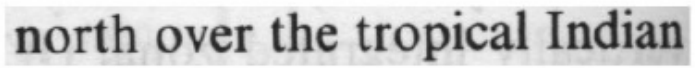


Document Analysis Tasks
Preprocessing

Document Binarization

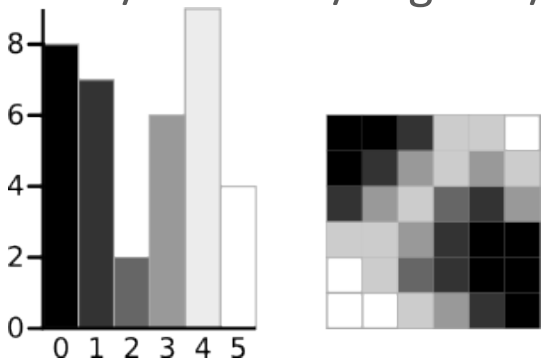
- Image Segmentation consists of 2 classes: foreground (written text) and background (paper)
- It converts a gray-scale document image into a binary document image
- Document image binarization is [was] usually performed in the preprocessing stage of different document image processing related applications:
  - OCR
  - Writer Identification
  - Layout Analysis, …

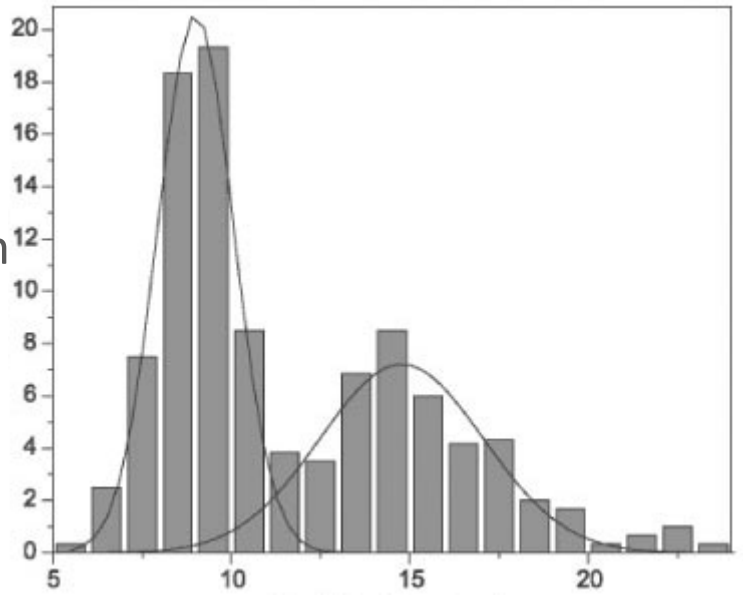
Effect of Binarization on OCR



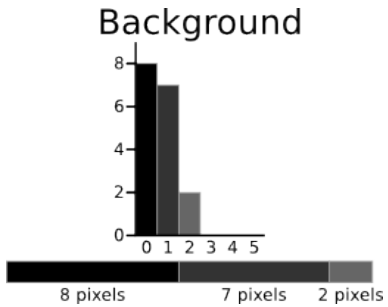
Global Binarization: Otsu

- Nobuyuki Otsu: *A threshold selection method from grey level histograms*. In: *IEEE Transactions on Systems, Man, and Cybernetics*. New York 9.1979, S.62–66. ISSN 1083-4419
- Otsu is a statistical method which assumes a bimodal histogram
  - Find a threshold that minimizes the weighted within-class variance/maximizes the between-class variance.
  - Example:
    http://www.labbookpages.co.uk/software/imgProc/otsuThreshold.htm





Otsu Example



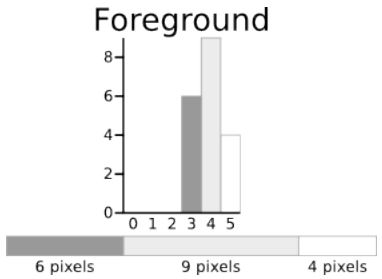
$$\text{Weight} \quad W_b = \frac{8+7+2}{36} = 0.4722$$

$$\text{Mean} \quad \mu_b = \frac{(0\times 8)+(1\times 7)+(2\times 2)}{17} = 0.6471$$

$$\text{Variance} \quad \sigma_b^2 = \frac{((0-0.6471)^2\times 8)+((1-0.6471)^2\times 7)+((2-0.6471)^2\times 2)}{17}$$

$$= \frac{(0.4187\times 8)+(0.1246\times 7)+(1.8304\times 2)}{17}$$

$$= 0.4637$$



$$\text{Weight} \quad W_f = \frac{6+9+4}{36} = 0.5278$$

$$\text{Mean} \quad \mu_f = \frac{(3\times 6)+(4\times 9)+(5\times 4)}{19} = 3.8947$$

$$\text{Variance} \quad \sigma_f^2 = \frac{((3-3.8947)^2\times 6)+((4-3.8947)^2\times 9)+((5-3.8947)^2\times 4)}{19}$$

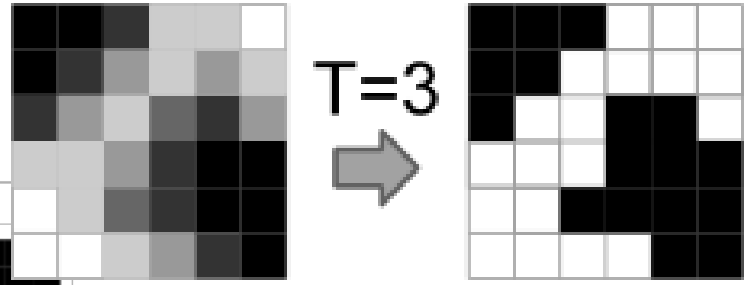
$$= \frac{(4.8033\times 6)+(0.0997\times 9)+(4.8864\times 4)}{19}$$

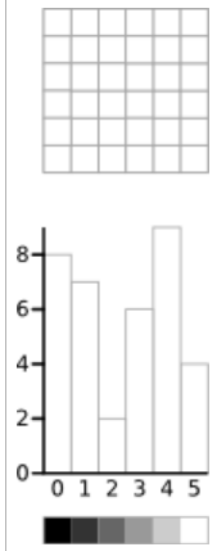
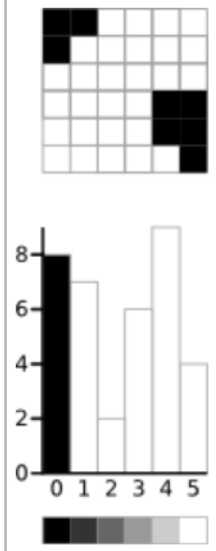
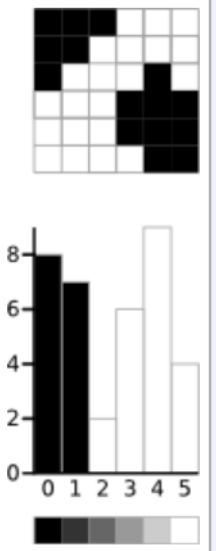
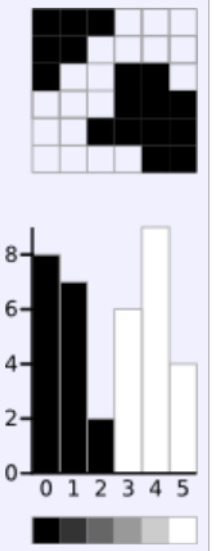
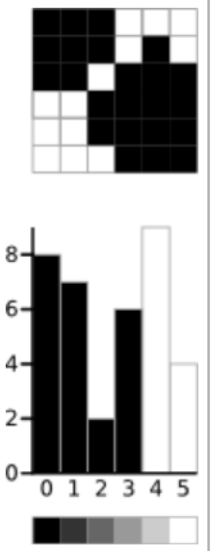
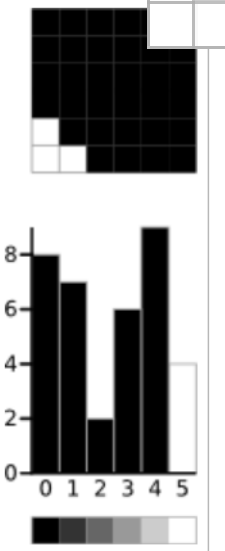
$$= 0.5152$$

$$\text{Within Class Variance} \quad \sigma_W^2 = W_b\,\sigma_b^2 + W_f\,\sigma_f^2 = 0.4722 * 0.4637 + 0.5278 * 0.5152$$

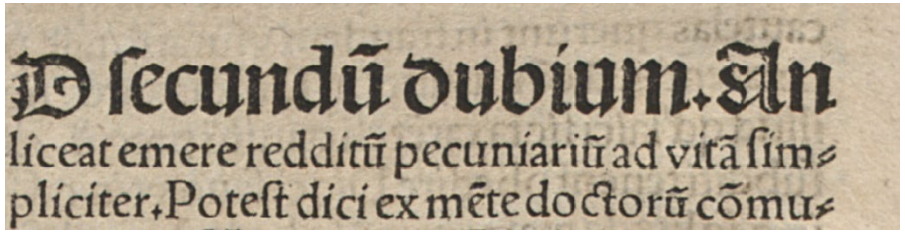
$$= 0.4909$$

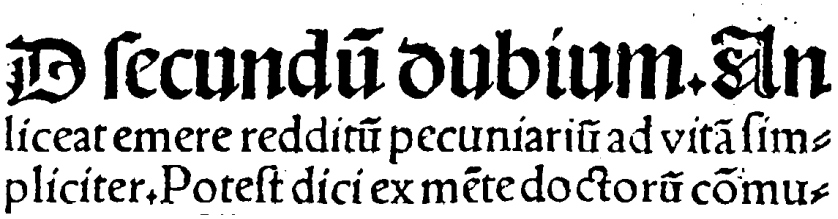
Otsu Example

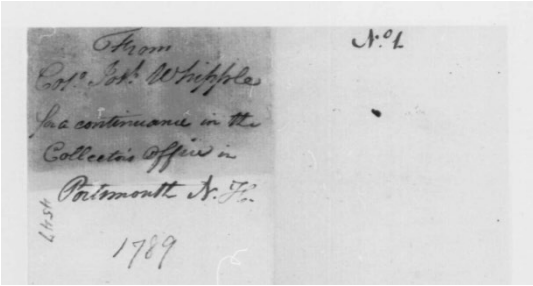


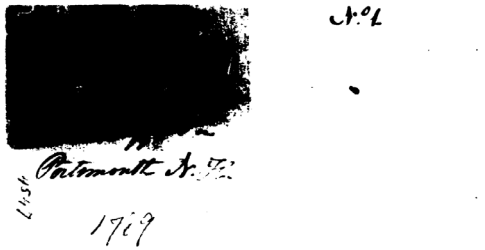
Threshold	T=0	T=1	T=2	T=3	T=4	T=5
						
Weight, Background	$W_b = 0$	$W_b = 0.222$	$W_b = 0.4167$	$W_b = 0.4722$	$W_b = 0.6389$	$W_b = 0.8889$
Mean, Background	$M_b = 0$	$M_b = 0$	$M_b = 0.4667$	$M_b = 0.6471$	$M_b = 1.2609$	$M_b = 2.0313$
Variance, Background	$\sigma^2_b = 0$	$\sigma^2_b = 0$	$\sigma^2_b = 0.2489$	$\sigma^2_b = 0.4637$	$\sigma^2_b = 1.4102$	$\sigma^2_b = 2.5303$
Weight, Foreground	$W_f = 1$	$W_f = 0.7778$	$W_f = 0.5833$	$W_f = 0.5278$	$W_f = 0.3611$	$W_f = 0.1111$
Mean, Foreground	$M_f = 2.3611$	$M_f = 3.0357$	$M_f = 3.7143$	$M_f = 3.8947$	$M_f = 4.3077$	$M_f = 5.000$
Variance, Foreground	$\sigma^2_f = 3.1196$	$\sigma^2_f = 1.9639$	$\sigma^2_f = 0.7755$	$\sigma^2_f = 0.5152$	$\sigma^2_f = 0.2130$	$\sigma^2_f = 0$
Within Class Variance	$\sigma^2_W = 3.1196$	$\sigma^2_W = 1.5268$	$\sigma^2_W = 0.5561$	$\mathbf{\sigma^2_W = 0.4909}$	$\sigma^2_W = 0.9779$	$\sigma^2_W = 2.2491$

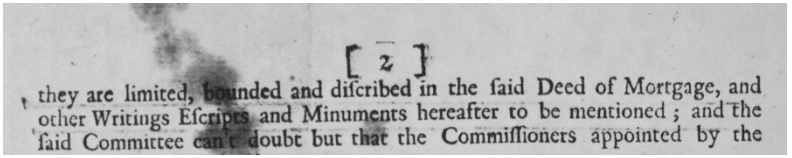
Otsu Results

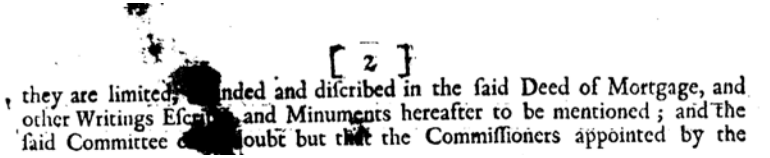












Deep Learning based Binarization

- Tensmayer and Martinez, 2017
  - Suggest FCN for image binarization, Architecture has U-Shape
  - H-DIBCO performance improvement of 0.5% compared to 97.5% (p-FM)

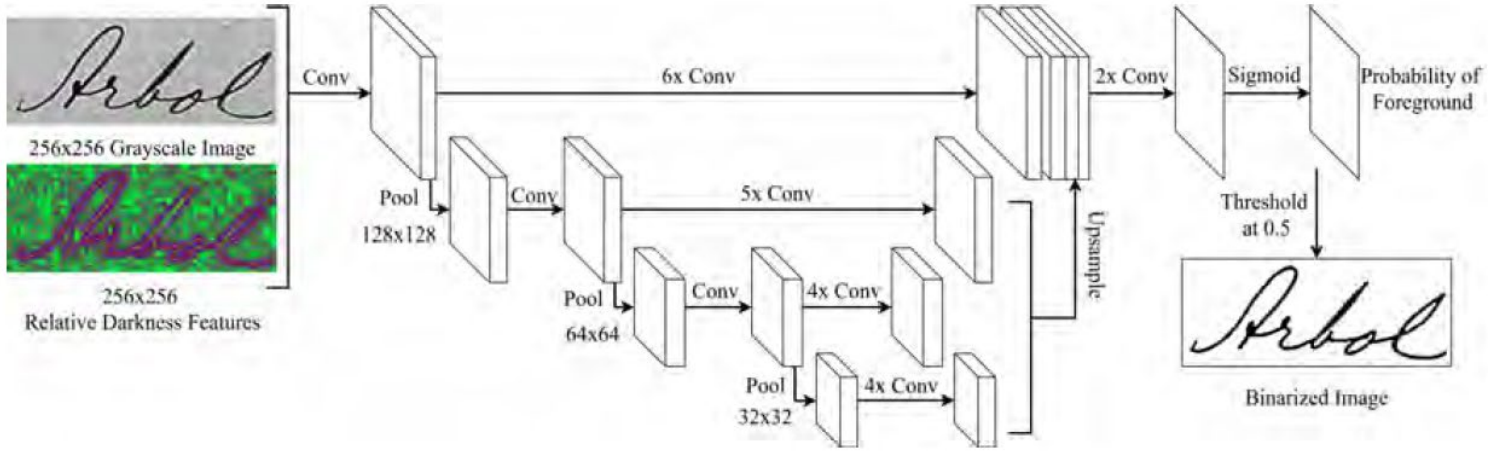
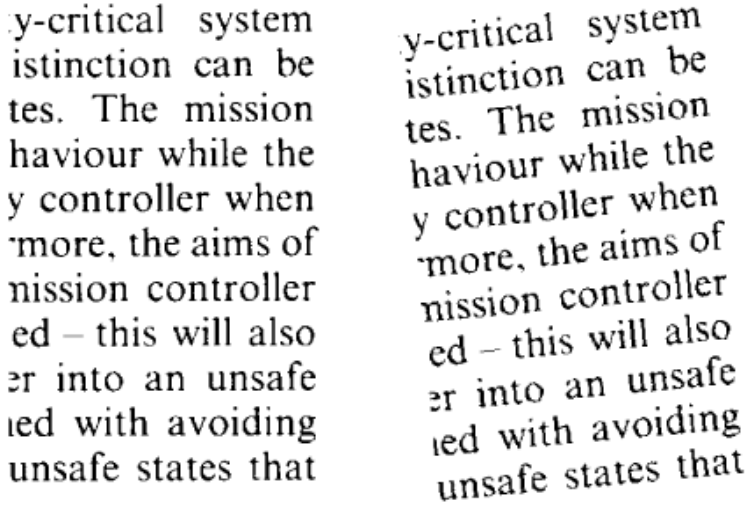


Figure 2.9: CNN architecture proposed by Tensmeyer and Martinez [TM17]. Image taken from [TM17].

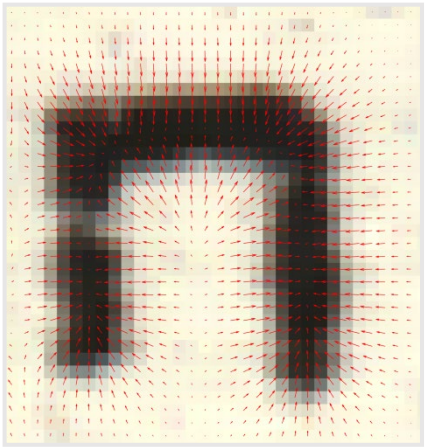
Skew Correction

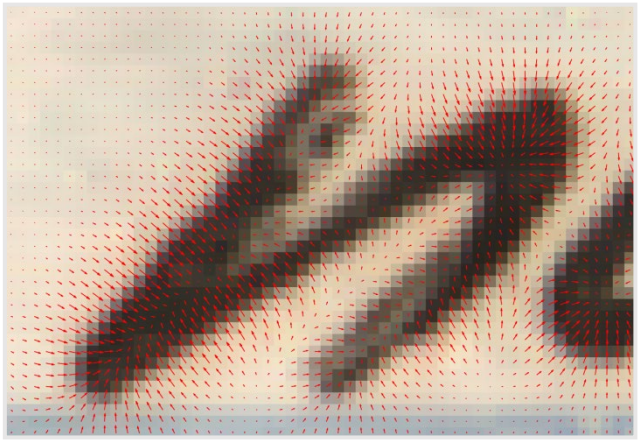
- Pre-processing step of document layout analysis and OCR methods
- *„For Humans, rotated images are unpleasant for visualization and introduce extra difficulty in text reading“* [Rafael Dueire Lins and Bruno Tenrio Avila, 2004]

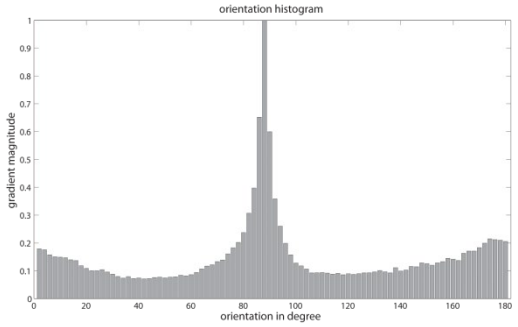

y-critical system
istinction can be
tes. The mission
haviour while the
y controller when
more, the aims of
nission controller
ed – this will also
er into an unsafe
ied with avoiding
unsafe states that

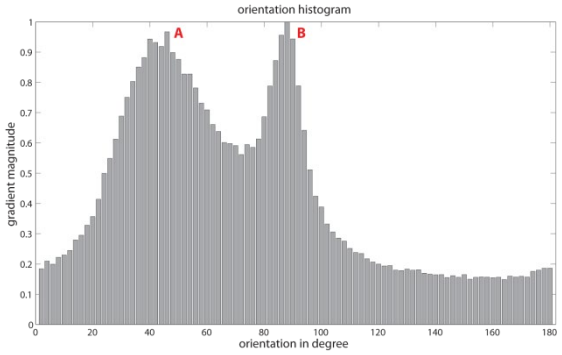
Skewed by 5°

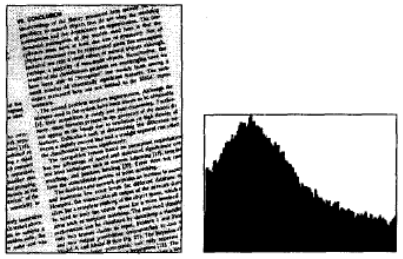
Example: Gradient and Projection Profile based











y-critical system
istinction can be
tes. The mission
haviour while the
y controller when
more, the aims of
nission controller
ed – this will also
er into an unsafe
ied with avoiding
unsafe states that

y-critical system
istinction can be
tes. The mission
haviour while the
y controller when
more, the aims of
nission controller
ed – this will also
er into an unsafe
ied with avoiding
unsafe states that

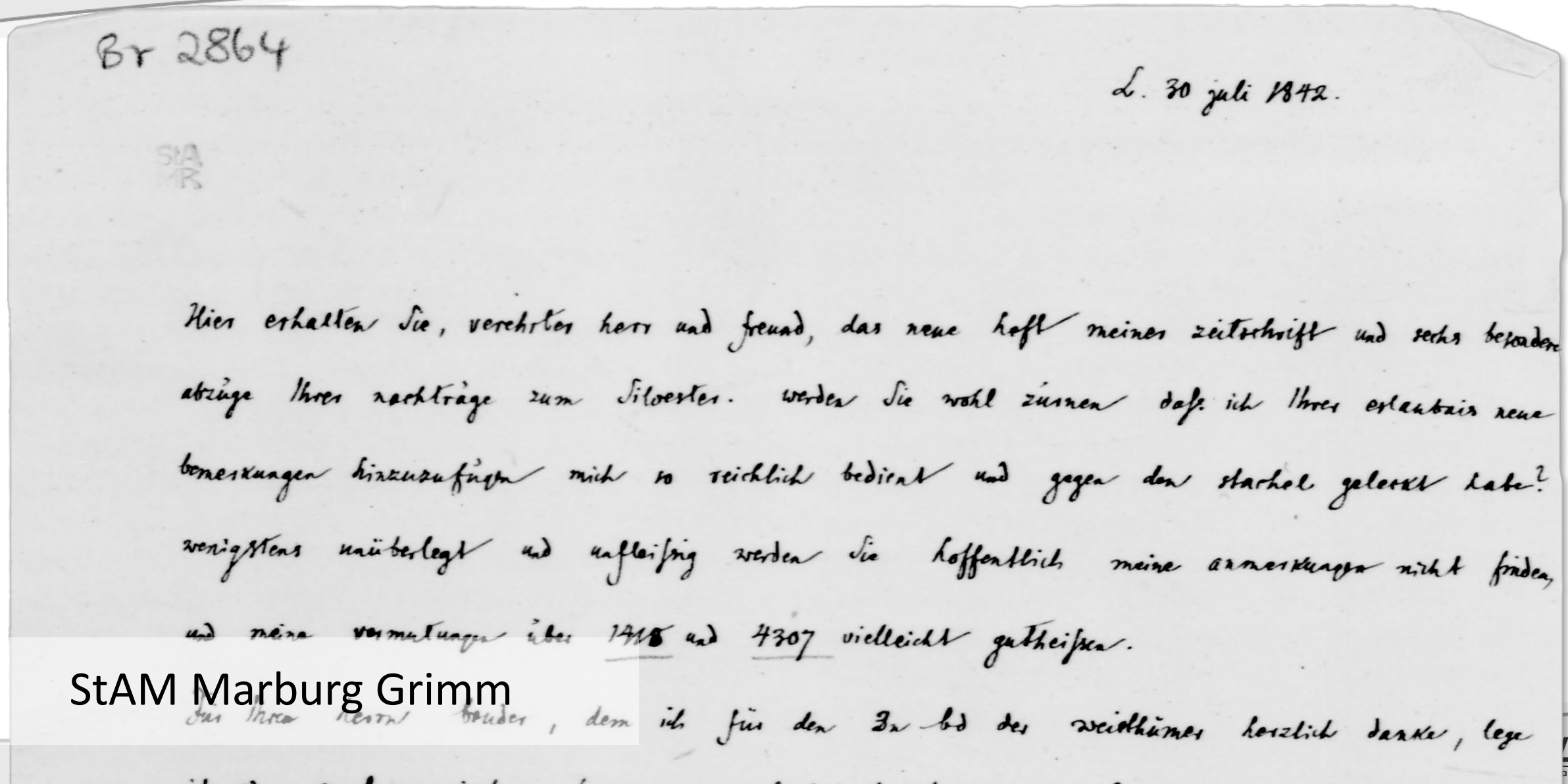
Document Analysis Tasks
Layout Analysis

Idea of Layout Analysis

- Page segmentation: split page in regions of interest (ROI)
  i.e. find homogeneous regions in a page
  - Segment into regions of one type (e.g. text, figure, formula, ...)
  - Different granularities (e.g. characters, words, text lines, paragraphs)
  - Baselines
  - Text, graphics and images separation
    Hairlines and frames detection

- Furthermore
  - Recover paragraph formatting (left, right, justified, ...)
  - Line height, line spacing, character size, ...

Objectives


Br 2864

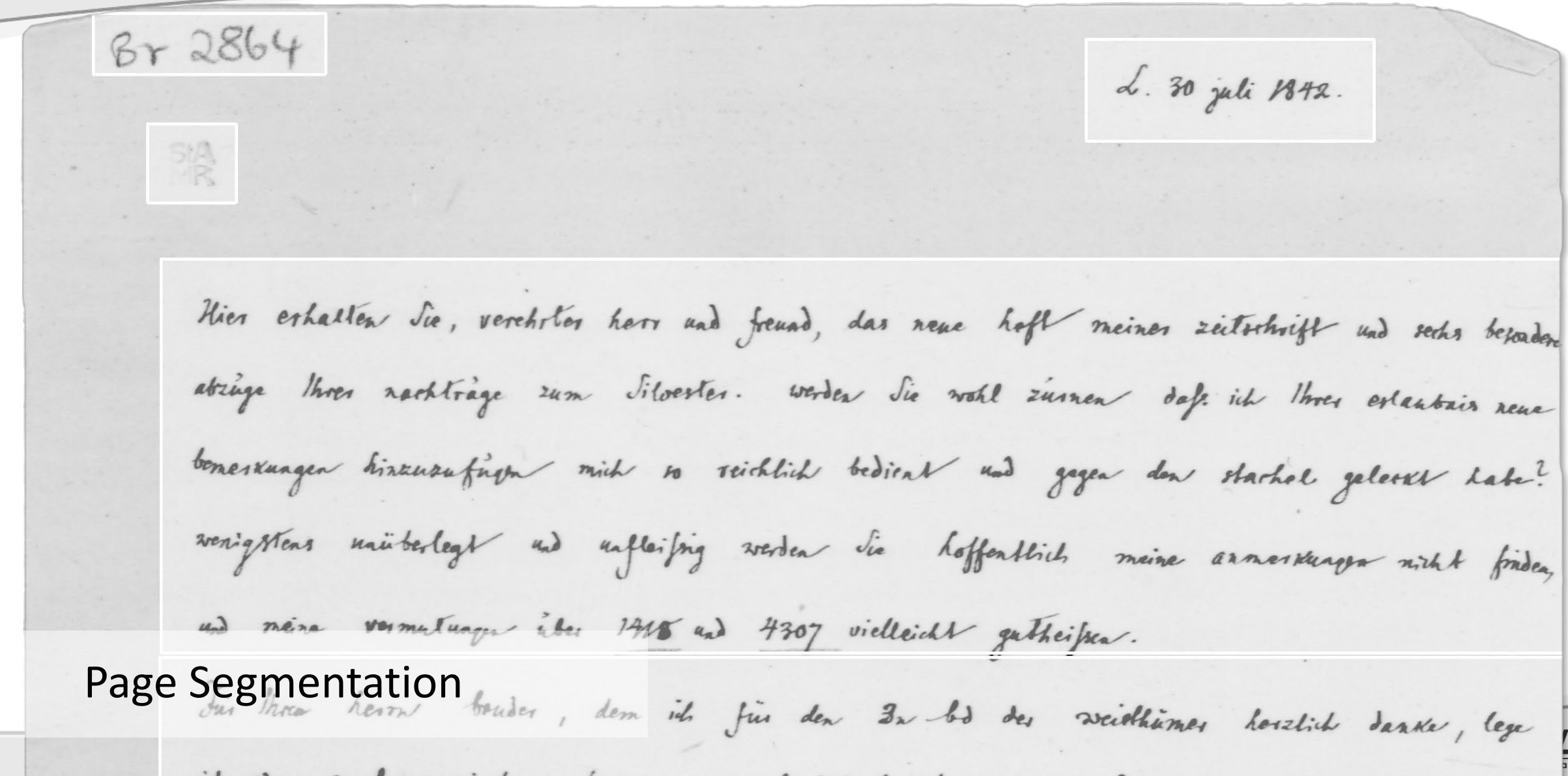
L. 30 juli 1842.

Hier erhalten Sie, verehrter herr und freund, das neue heft meiner zeitschrift und sechs besondere abzüge Ihrer nachträge zum Silvester. werden Sie wohl zürnen daß ich Ihres erlaubnis neue bemerkungen hinzuzufügen mich so reichlich bedient und gegen den stachel gelöckt habe? wenigstens unüberlegt und unfleißig werden Sie hoffentlich meine anmerkungen nicht finden, und meine vermutungen über 1418 und 4307 vielleicht gutheißen.

Ihres herrn bruder, dem ich für den 3n bd des weisthümer herzlich danke, lege

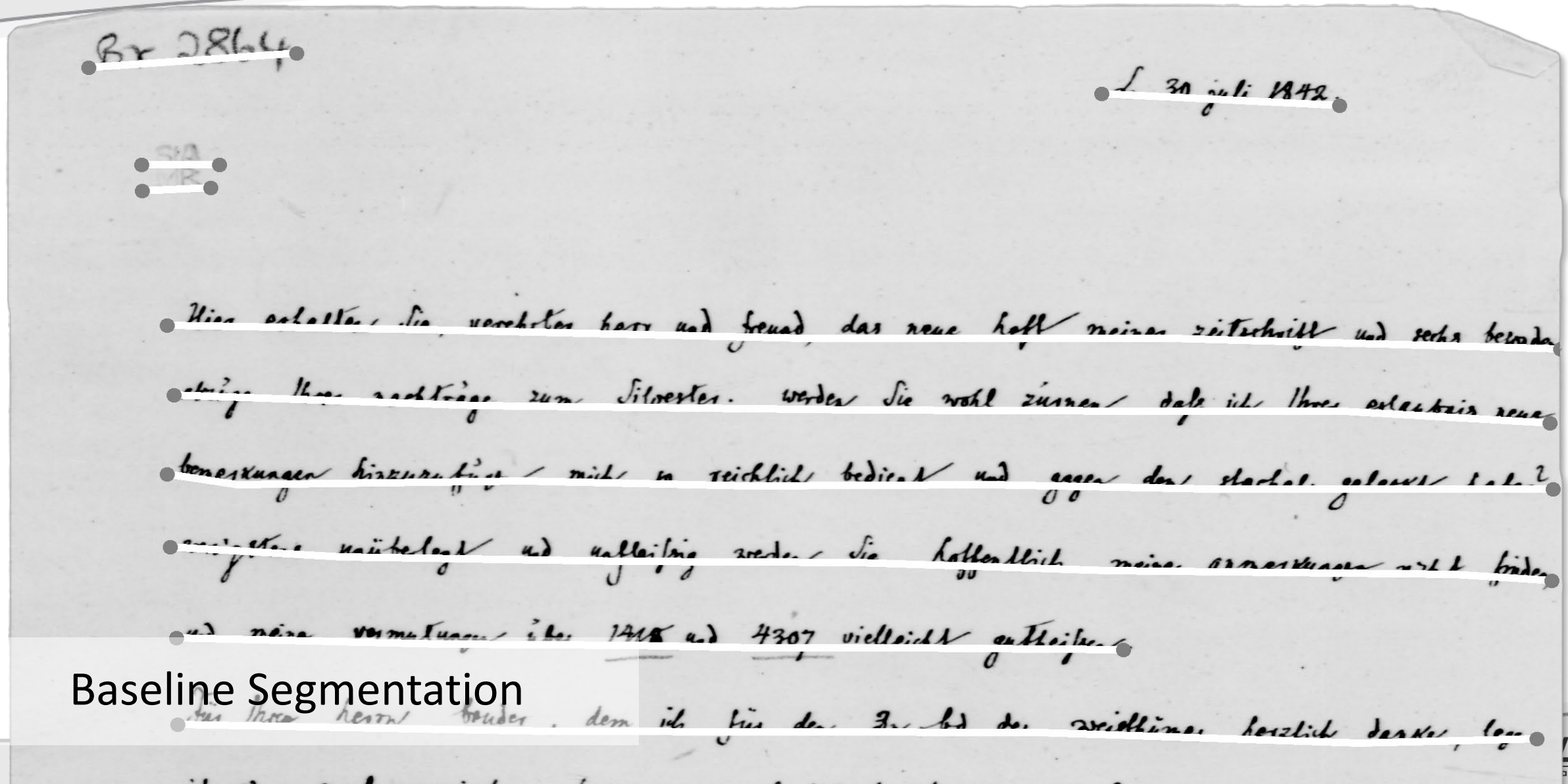
StAM Marburg Grimm

Objectives



Page Segmentation

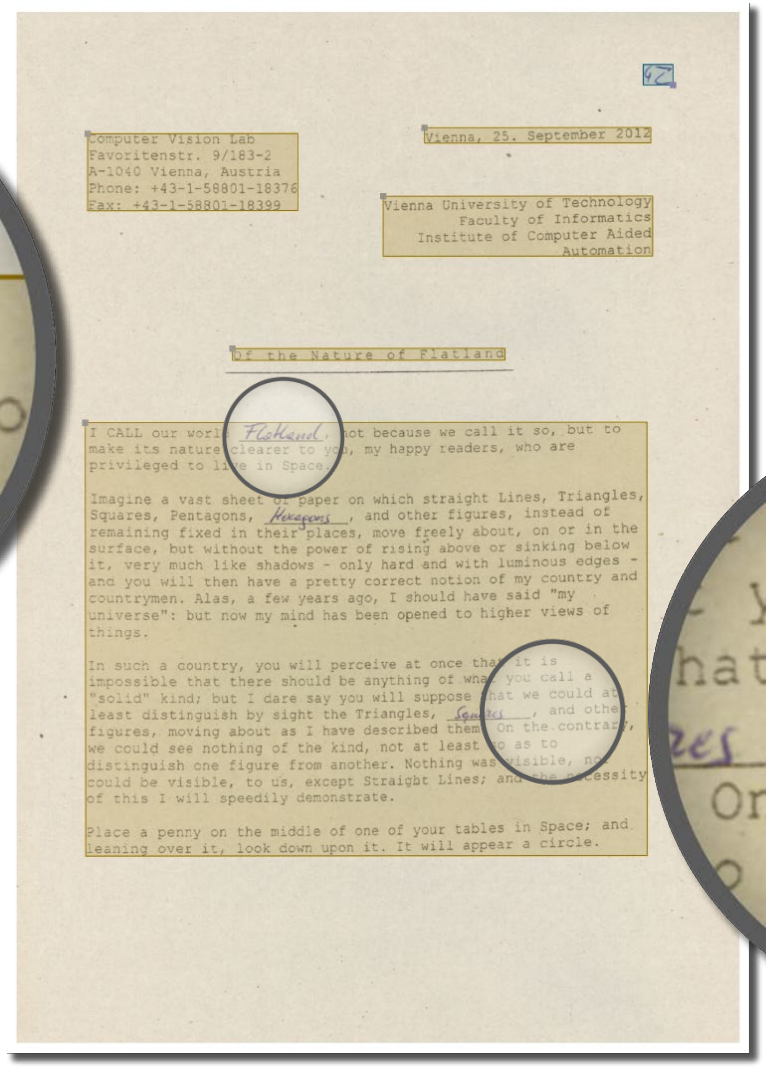
Objectives

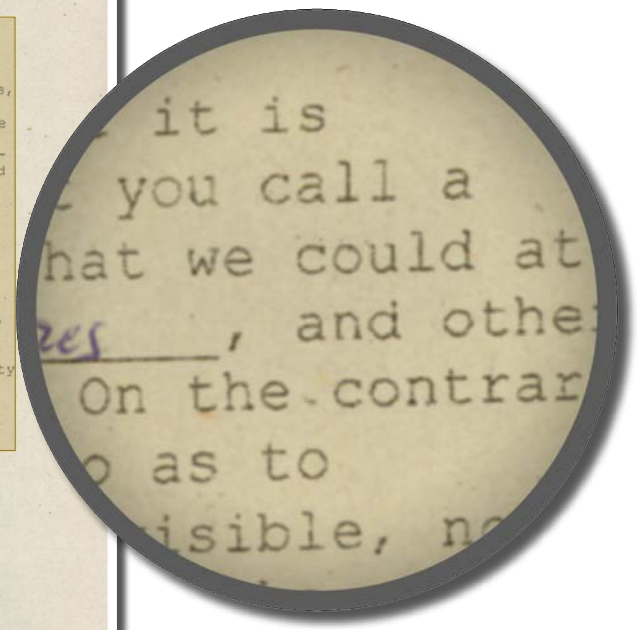


Baseline Segmentation

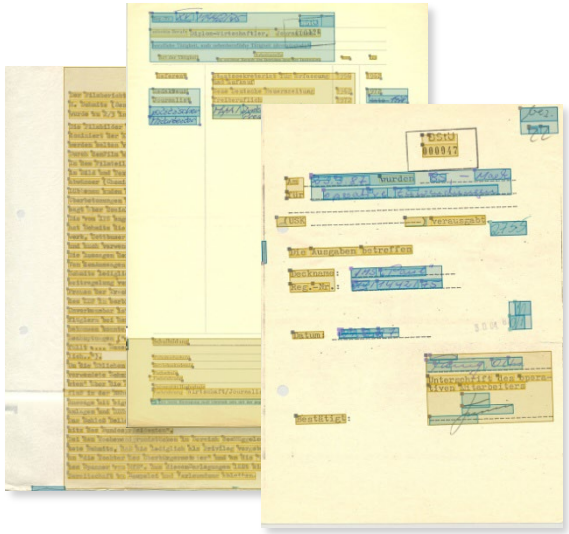

median
April
x-height
baseline

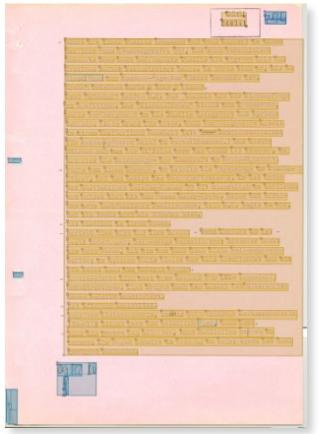
Granularity





Document Clustering





Font

Segmentation

- Strategies
  - Top-Down
  - Bottom-Up
  - Hybrid
- Segment into regions of one type (e.g. text, figure, formula, ...)
- Different granularities (e.g. characters, words, text lines, paragraphs)
- Text, graphics and images separation
- Recover paragraph formatting (left, right, justified, ...)
- Line height, line spacing, character size, ...

Binarization-Free Text Line Segmentation for Historical Documents Based on Interest Point Clustering

Angelika Garz
Computer Vision Lab
Vienna Univ. of Technology
1040 Vienna, Austria
garz@caa.tuwien.ac.at

Andreas Fischer
Inst. of Computer Science and Applied Mathematics
3012 Bern, Switzerland
afischer@iam.unibe.ch

Robert Sablatnig
Computer Vision Lab
Vienna Univ. of Technology
1040 Vienna, Austria
sab@caa.tuwien.ac.at

Horst Bunke
Inst. of Computer Science and Applied Mathematics
3012 Bern, Switzerland
bunke@iam.unibe.ch

***Abstract*—Segmenting page images into text lines is a crucial pre-processing step for automated reading of historical documents. Challenging issues in this open research field are given e.g. by paper or parchment background noise, ink bleed-through, artifacts due to aging, stains, and touching text lines. In this paper, we present a novel binarization-free line segmentation method that is robust to noise and copes with overlapping and touching text lines. First, interest points representing parts of characters are extracted from gray-scale images. Next, word clusters are identified in high-density regions and touching components such as ascenders and descenders are separated using seam carving. Finally, text lines are generated by concatenating neighboring word clusters, where neighborhood is defined by the prevailing orientation of the words in the document. An experimental evaluation on the Latin manuscript images of the Saint Gall database shows promising results for real-world applications in terms of both accuracy and efficiency.**

***Keywords*-historical documents, manuscripts, ancient documents, handwritten, text line segmentation, binarization-free**

I. Introduction

Automatic segmentation of historical document page images is an open research field; algorithms are required to be robust with respect to background artifacts such as clutter, stains and noise, as well as artifacts due to aging, and touching or interfering lines [1]. Text line segmentation, in particular, is typically needed for handwriting recognition in historical documents. Handwritten documents do not have strict layout rules and thus line segmentation methods need to be invariant to layout inconsistencies, irregularities in script and writing style, skew, and fluctuating text lines [1]. Furthermore, robustness to low contrast and rippled pages is required [2], [3].

Likforman-Sulem et al. [1] provide a detailed survey about segmentation of text lines with respect to historical documents. Well-known methods for text line segmentation in binary images include smearing [4], [5] and Hough-transform [6], [7]. Another commonly used approach is based on Projection Profiles (Pp) [2], [8]–[11] for both binary and gray-scale images. Various authors [8], [9] adapted the global Pp such that skewed text blocks, converging or merging text lines are segmented correctly.

Recent approaches [12]–[14] introduce seam carving [15] known from image retargeting for text line segmentation. Nicolaou and Gatos [14] use so-called local minima tracers which follow the line spacing in order to shred the document page in lines. Originally proposed for on-line documents [16], Indermühle et al. [12] use Dynamic Programming (Dp) in order to find a path with the minimum cost between two lines in historical manuscripts. Asi et al. [13] apply their approach directly on gray-scale images, where a distance transform is computed from a Gaussian-blurred image, and the separating seams are established using Dp.

In this paper, we introduce an efficient binarization-free method for line segmentation applicable to historical manuscripts. Thus, binarization-caused errors frequent in historical document images are not inherited. Furthermore, the proposed method follows a bottom-up approach by grouping parts-of-character interest points into text line regions. Hence, it is not necessary to extract text block regions beforehand, which can also be prone to errors for special page layouts encountered in historical documents. Touching components such as ascenders and descenders are locally split by means of seam carving.

The experimental evaluation of the proposed approach is carried out on the *Saint Gall database*[1] [17], which contains 60 pages of a Latin manuscript originating from the 9^{th} century written in Carolingian script by a single writer with ink on parchment. Two sample pages are illustrated in Figure 1. Besides the main text body, the pages contain colored initial letters and annotations located in the outer margin, which were added to the manuscript later. The text line spacing is relatively large compared to the word height and a regular page layout is present for all pages. However, there are line interconnections caused by ascenders and descenders. Stains, holes, and ink bleed-through pose

This work has been supported by Austrian Science Fund (Grant P23133) and Swiss National Science Foundation (Project CRSI22_125220).

[1]Available at http://www.iam.unibe.ch/fki/databases

Page Segmentation – Strategies

Segmentation produces a hierarchy of physical objects

- Top-Down
  - Start with entire image, recursively split to elementary shapes
- Bottom-Up
  - Start at pixel level, group to structures such as words, lines, regions
- Hybrid
  - Combine both strategies
- Deep Learning based (e.g. UMAP, ARU Net)

- Model-driven
  - Knowledge about the expected layout
- Data-driven
  - No expectation about the layout, use only data properties

meiner hiesigen stellung darf ich hoffen.

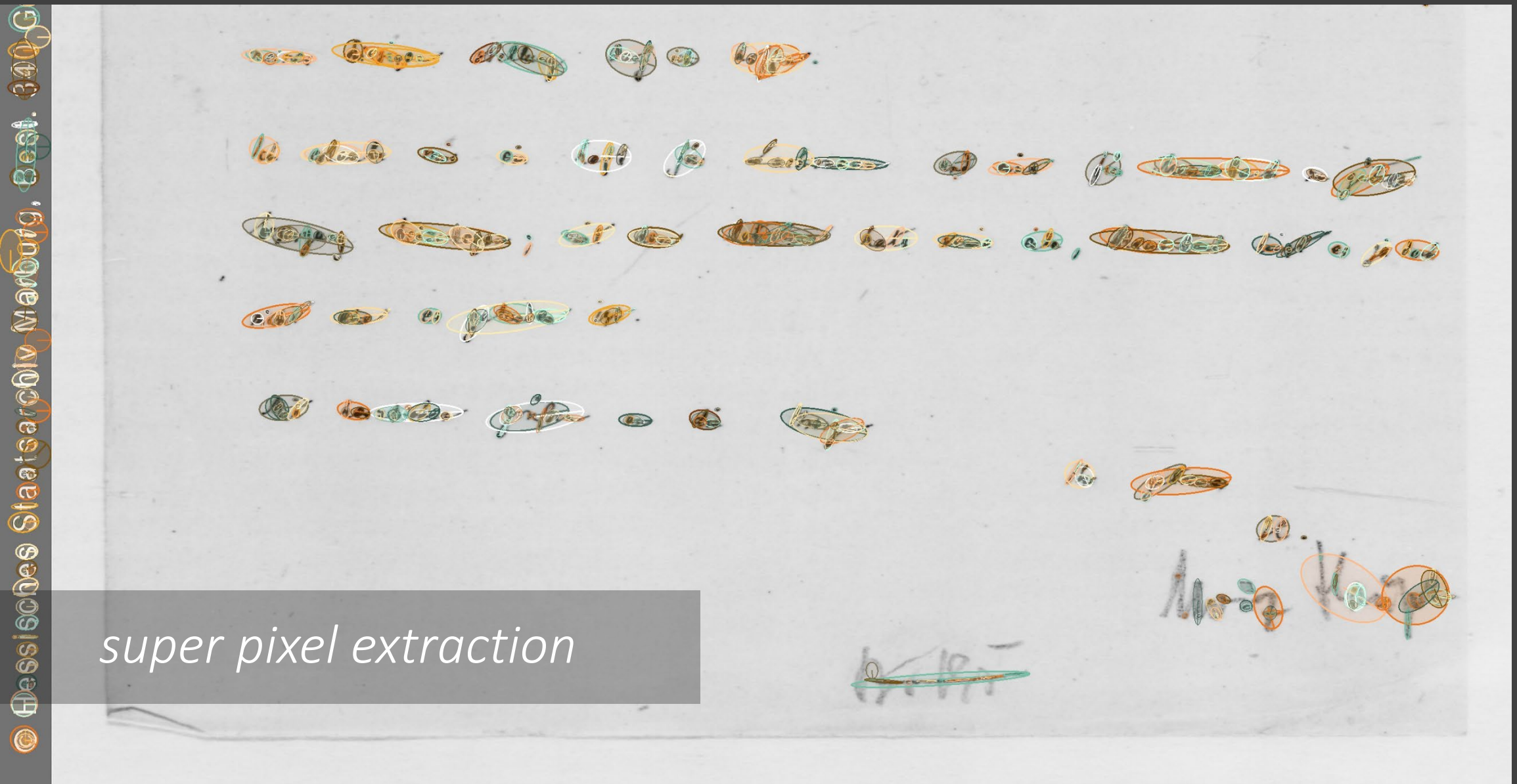
Ich erlaube mir ein heft für Lachmann und eins für Meusebach zu gütiger besorgung beizulegen; mit dem letzteren hats keine eile, Meusebach liest es ja doch nicht ehe es gebunden ist.

Mit herzlichen grüßen an die ihrigen

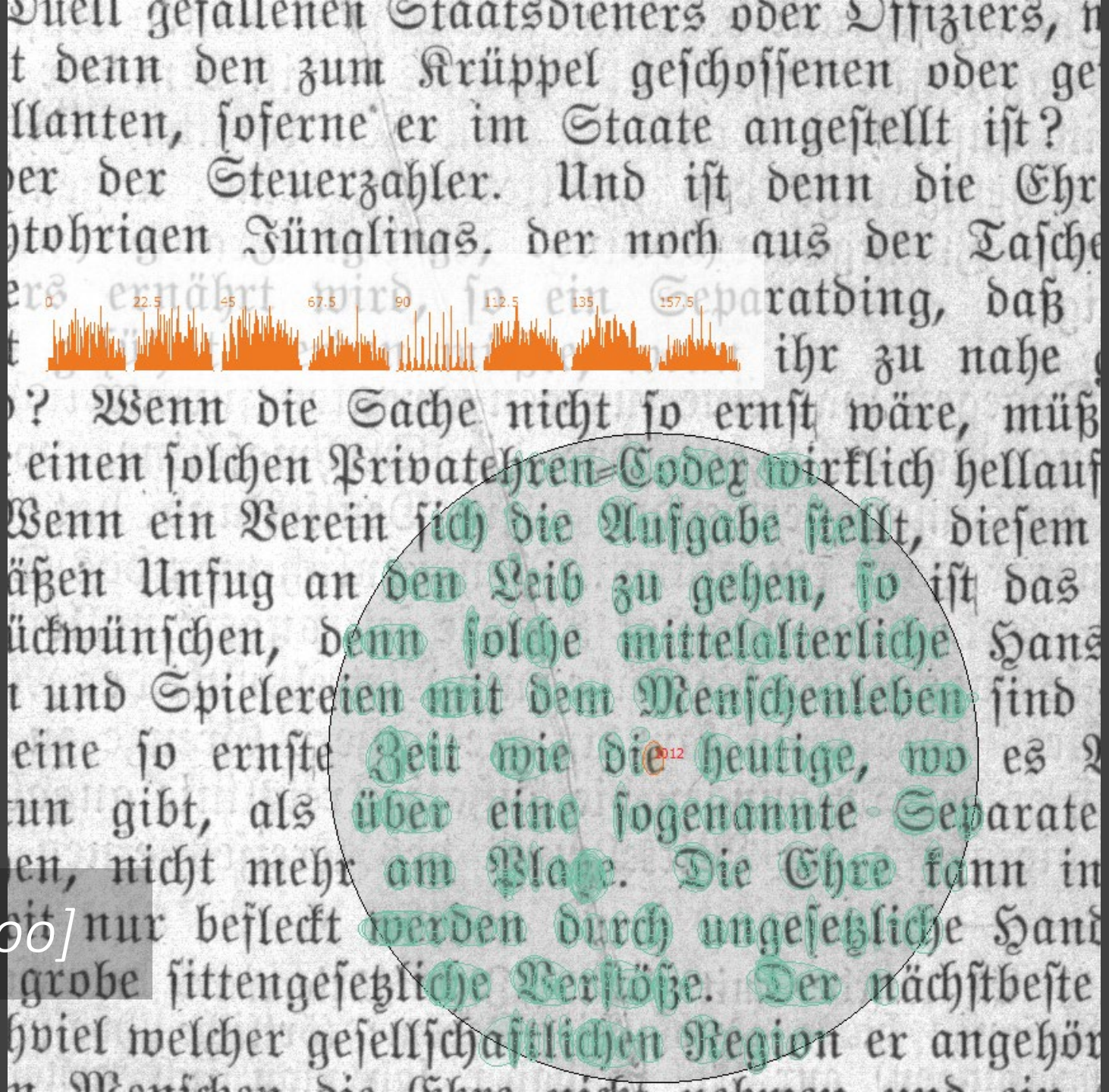
Ihr getreuer

H.

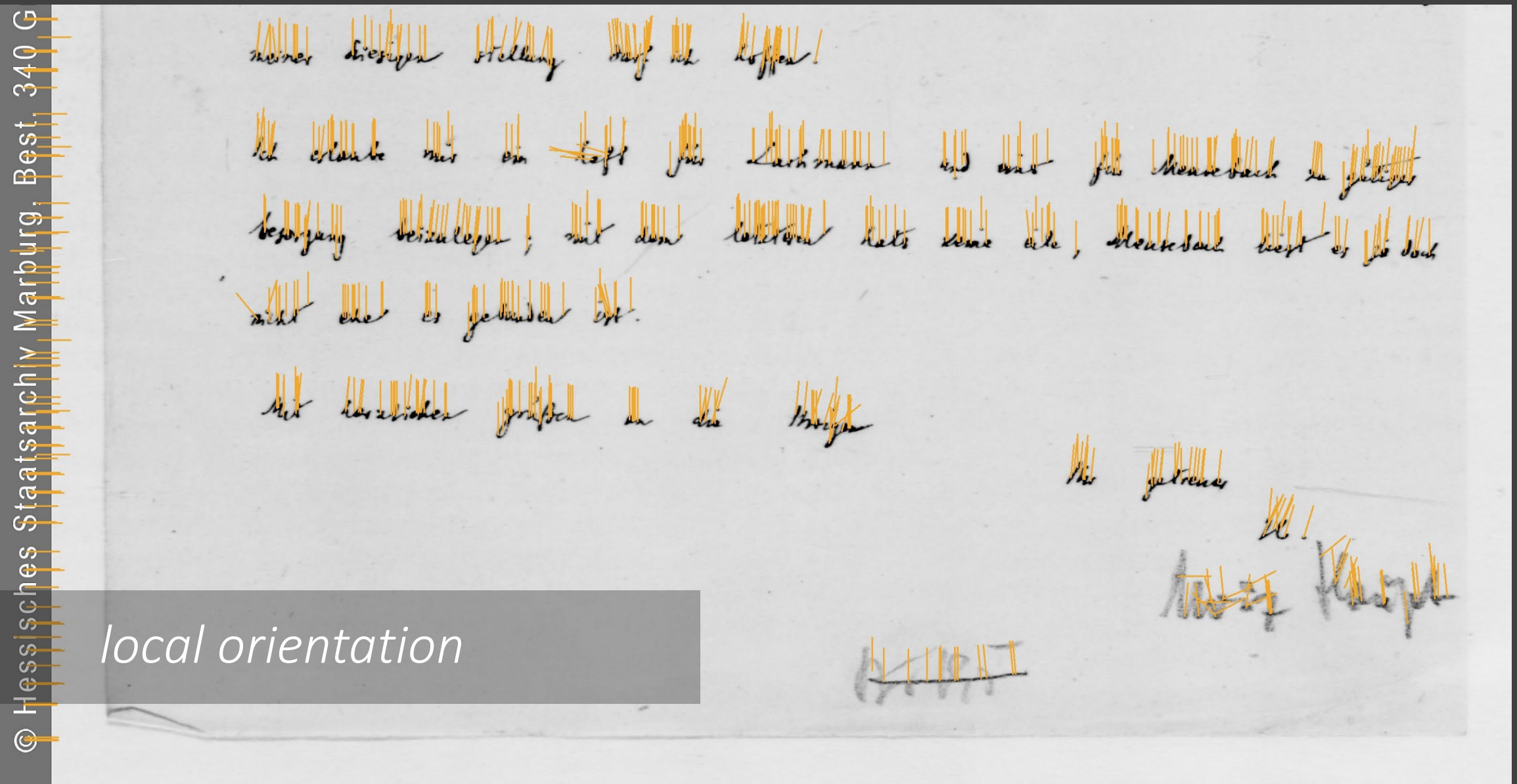
Moriz Haupt



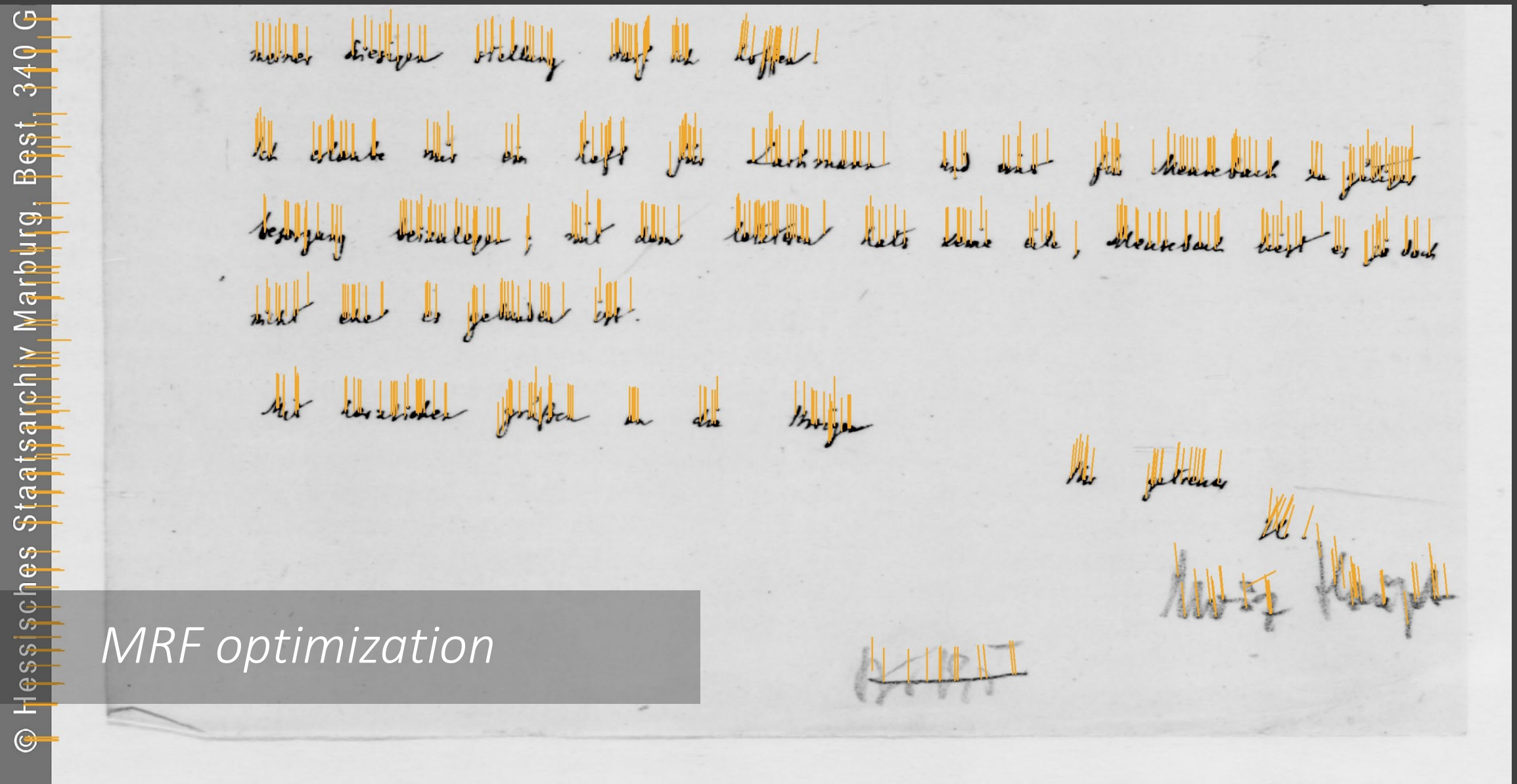
super pixel extraction



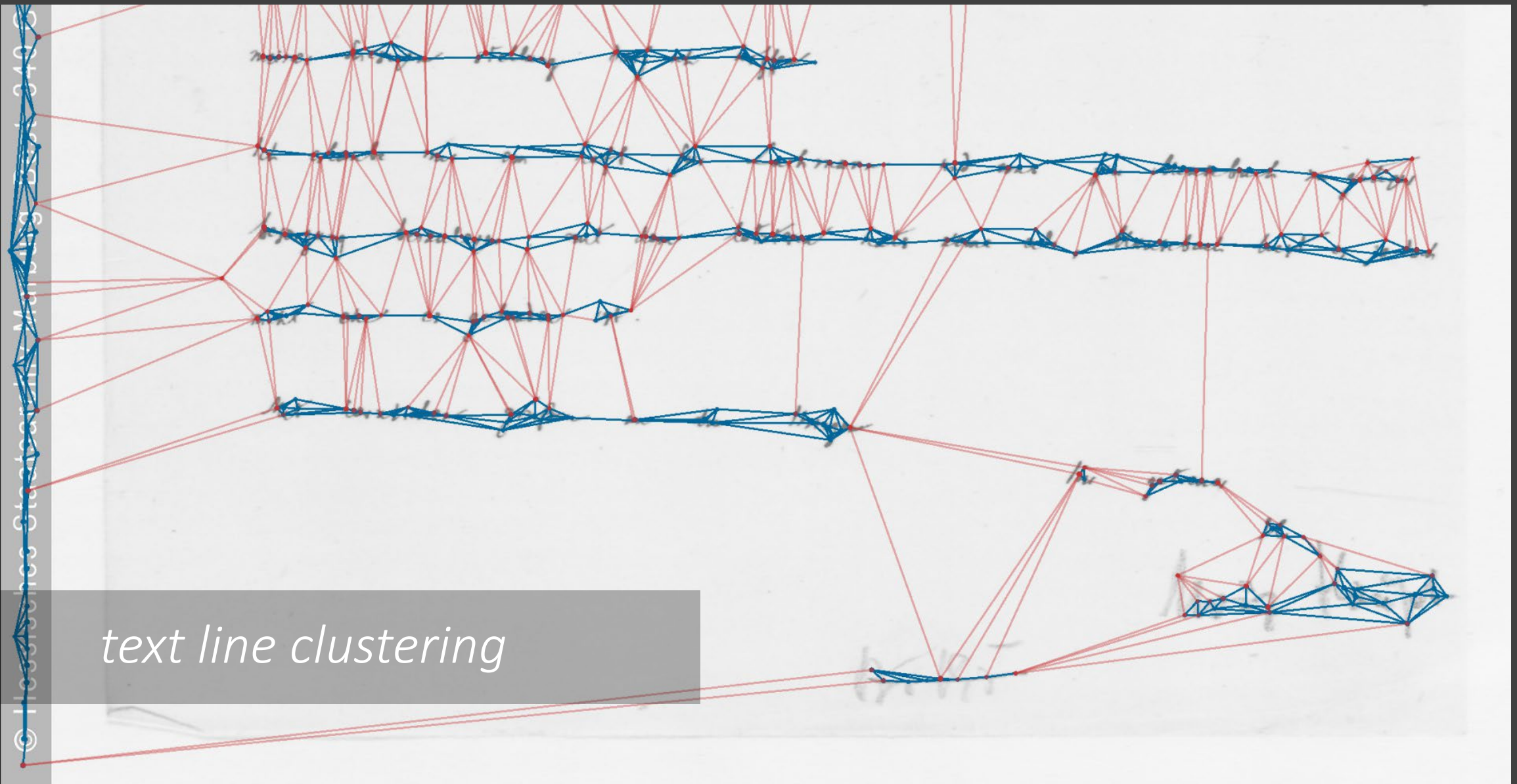
local orientation [II Koo]



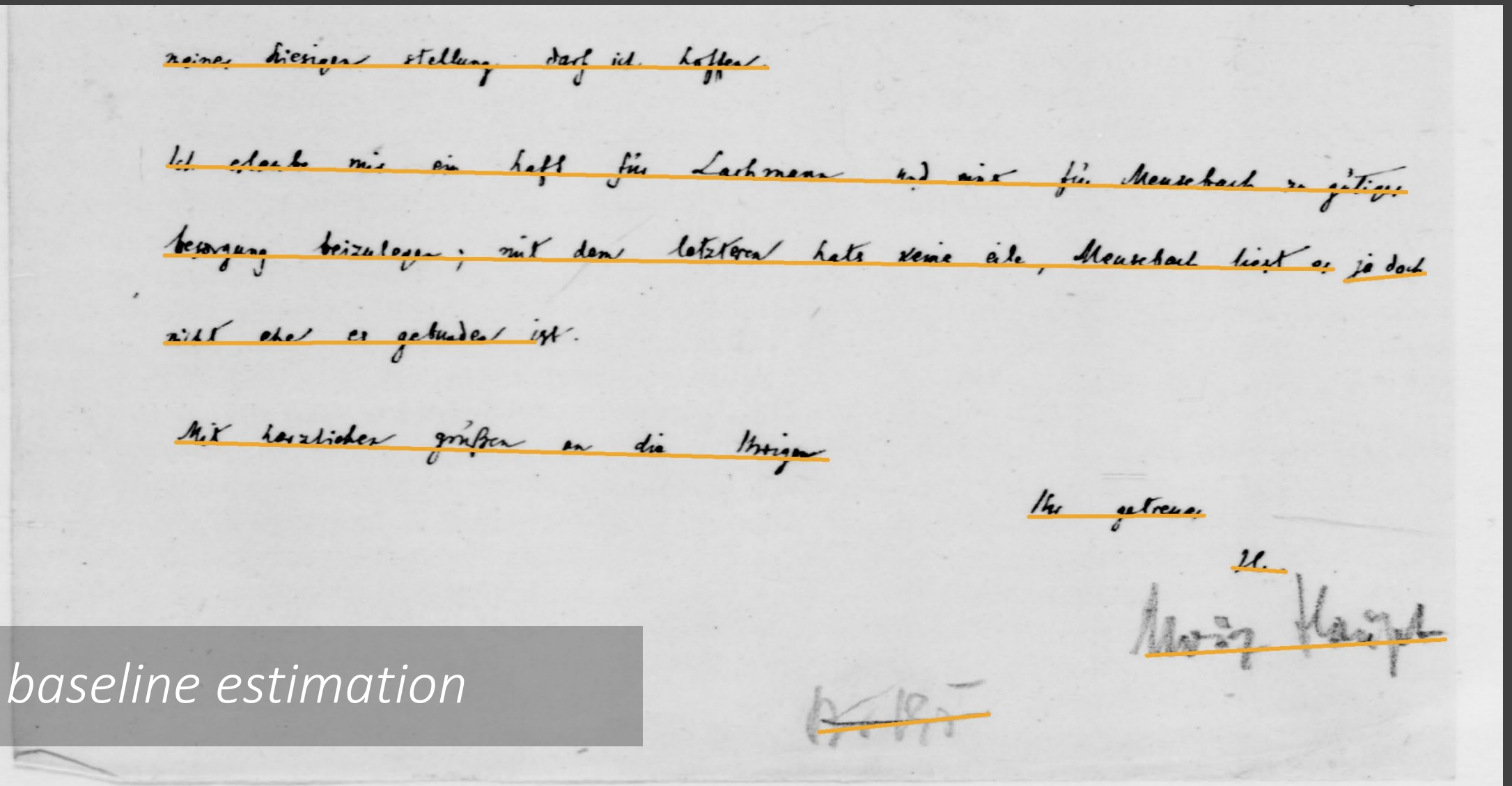
local orientation



MRF optimization



text line clustering



meine hiesigen Stellung, daß ich hoffen

ich glaube mir ein Lob für Lachmann und mir für Meinhard zu gütige

Besorgung beizulegen; mit dem letzteren hats keine Eile, Meinhard liest es ja doch

nicht eher er gebunden ist.

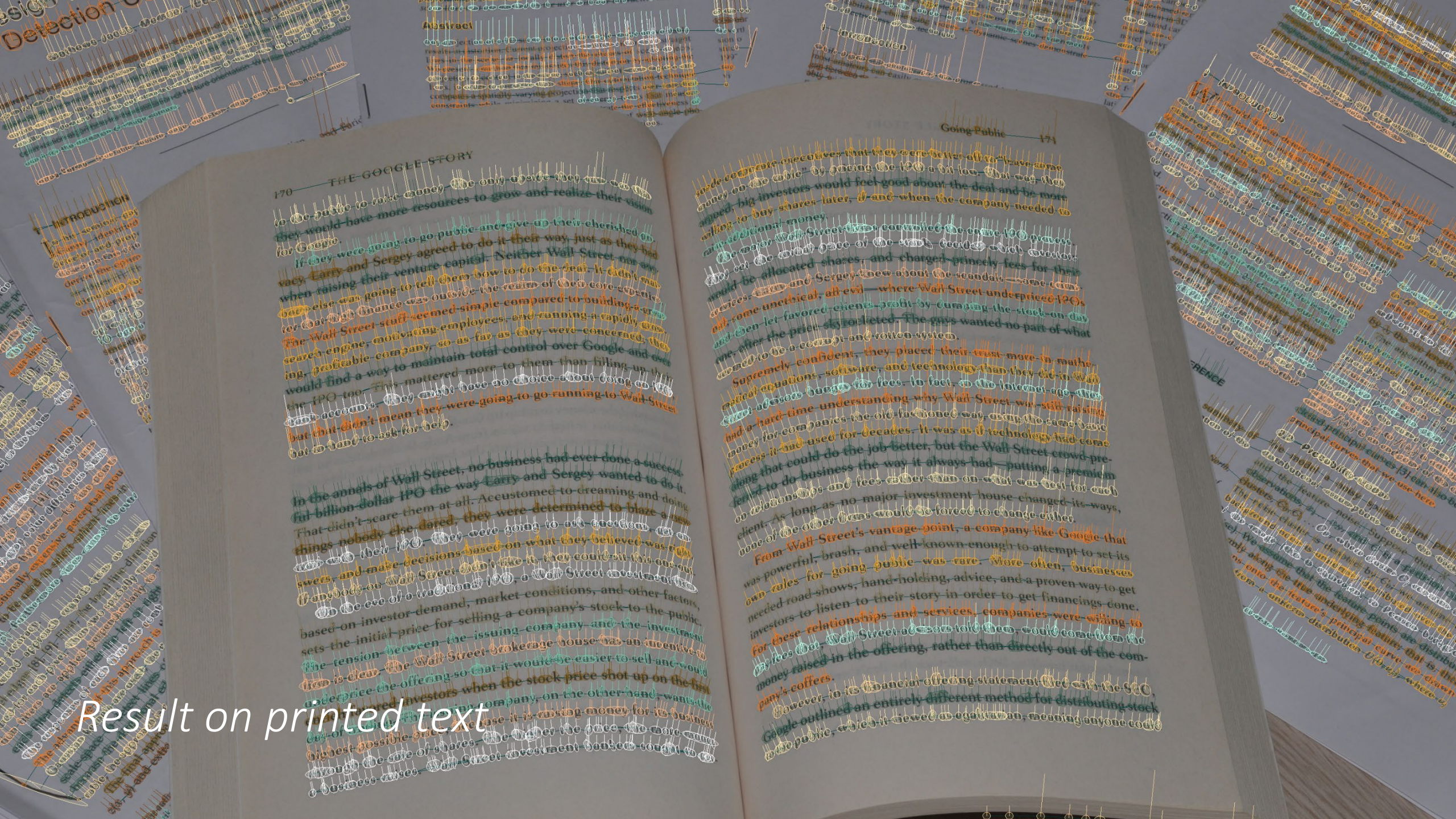
Mit herzlichen Grüßen an die Ihrigen

Ihr getreuer

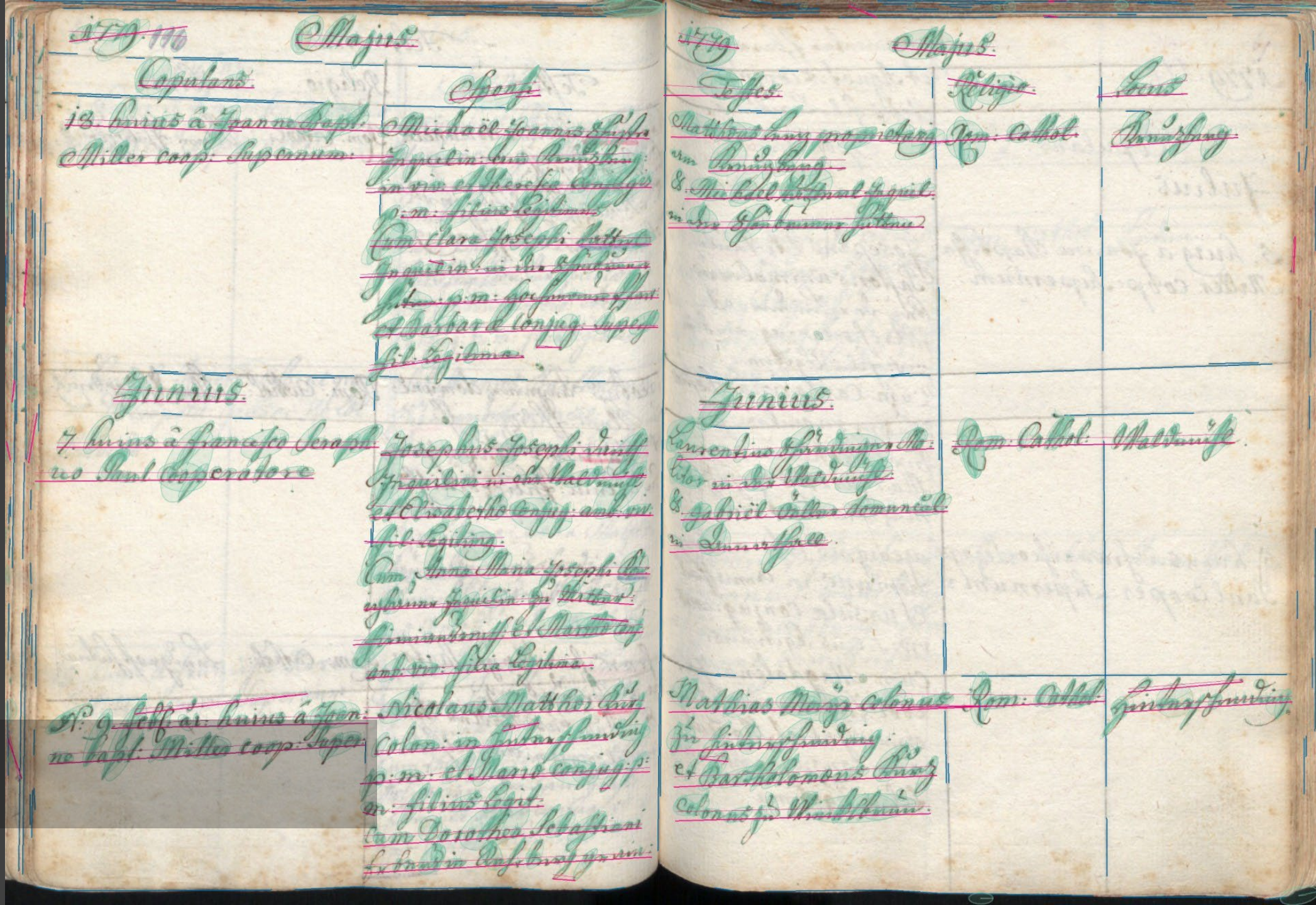
H.

Moriz Haupt

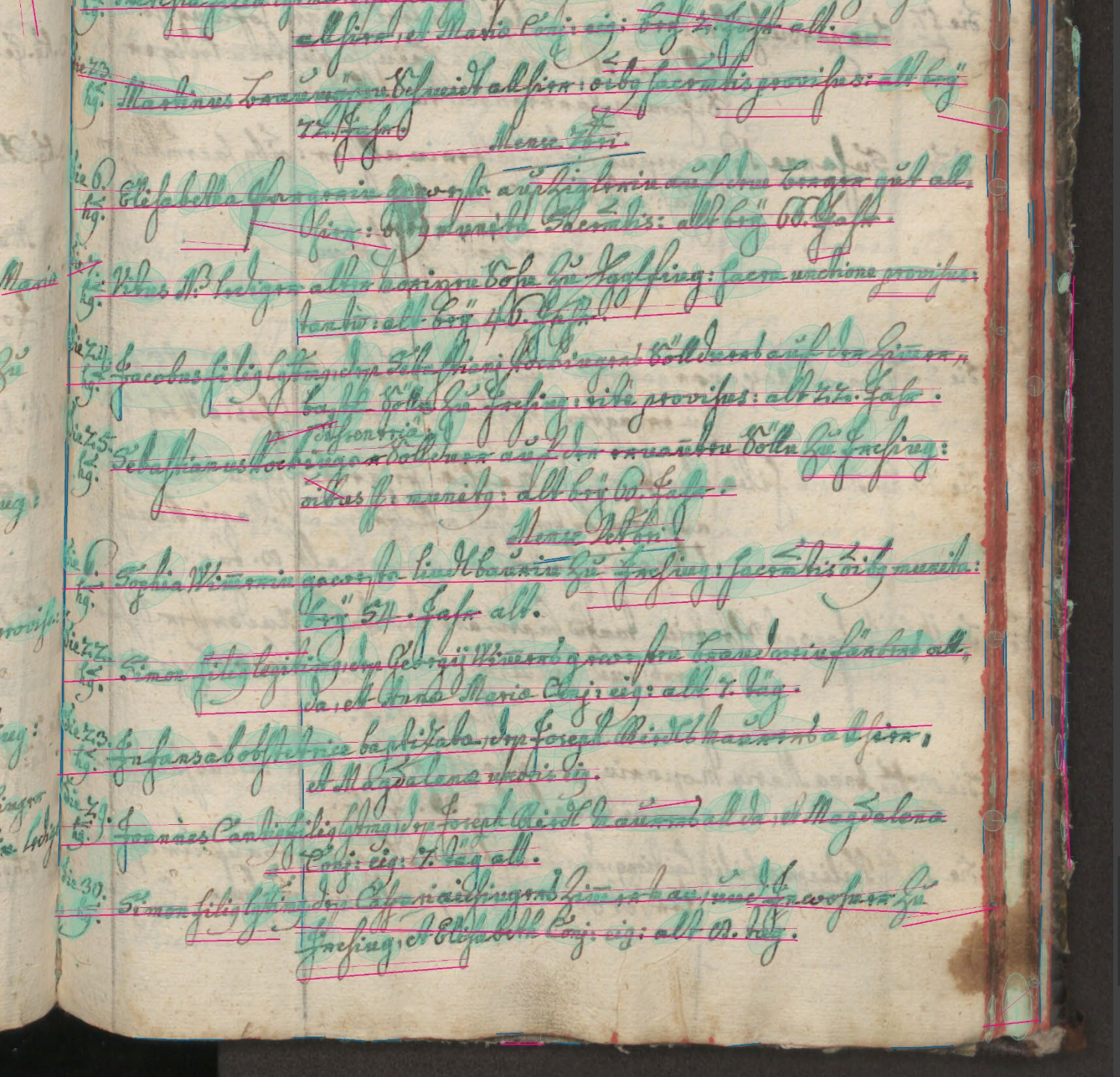
baseline estimation



Result on printed text



tables



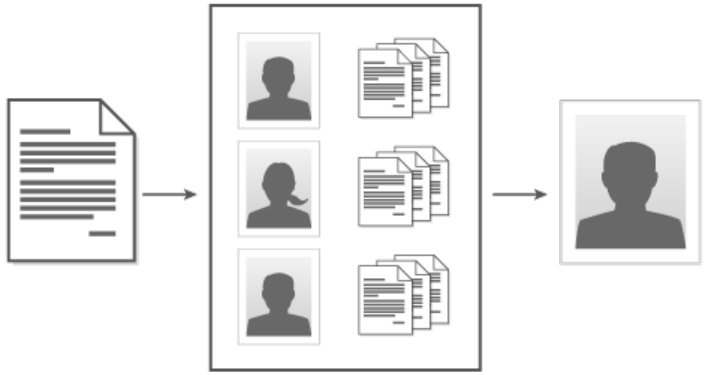
degraded documents

Document Analysis Tasks
Writer Identification & Retrieval

Writer Identification

Writer Identification is the task of determining the author of a sample handwriting from a set of writers

- Given:
  - Set of documents where the writer of each page is known
  - Document from an unknown writer
- Wanted:
  - ID of the writer of the document



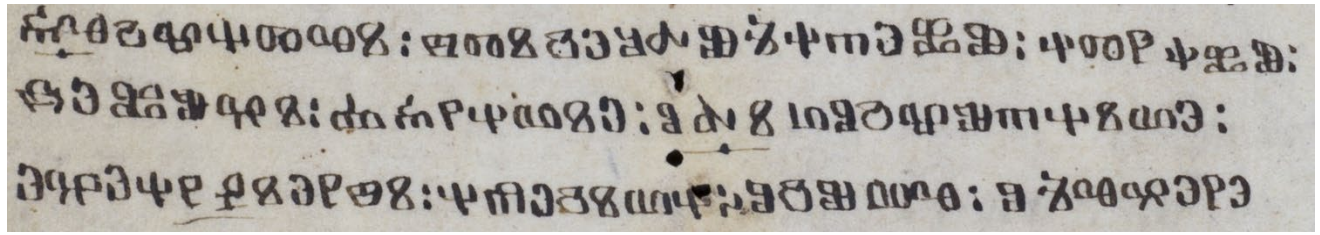
Writer Retrieval

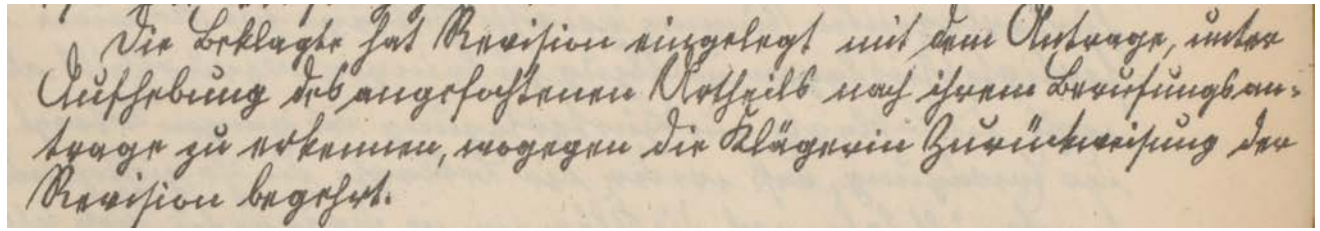
Writer Retrieval is the task to obtain all documents of one writer out of a set of documents

- Given:
  - Set of documents where the writers are not known
  - A reference document
- Wanted:
  - Ranking of the documents according to the similarity to the reference document

Use cases for Writer Identification

- Forensics (e.g. threat or ransom letters)
- Smart meeting rooms (online identification)
- Historic Document
  - Digital libraries
  - Follow trace of medieval scribes



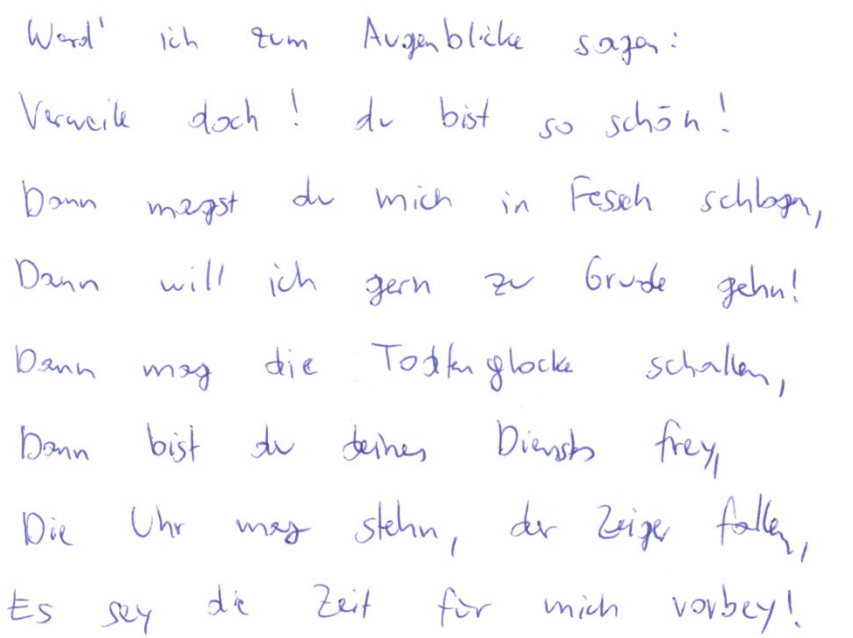


Use cases for Writer Retrieval

- Finding more documents of the same writer (e.g. unknown document of a writer in archives)
- Forensics
- Clustering of not indexed set of documents (e.g. Stasi files)
- Preprocessing for other applications (e.g Handwritten Text Recognition)

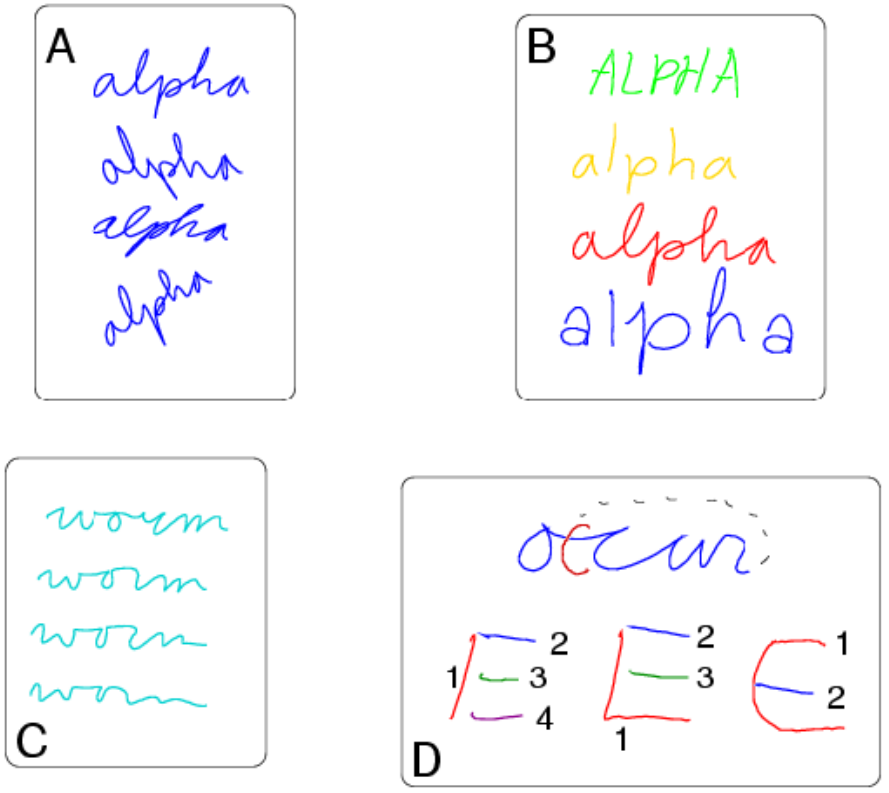
Offline vs Online Writer Identification

- Offline handwriting
  - Scanned images of handwritten documents
- Online handwriting
  - Collected in real time
    (at the same time it is produced)
  - Additional dynamic information
    - Velocity
    - Acceleration
    - Pen-pressure
    - Writing direction
    - Strokes order



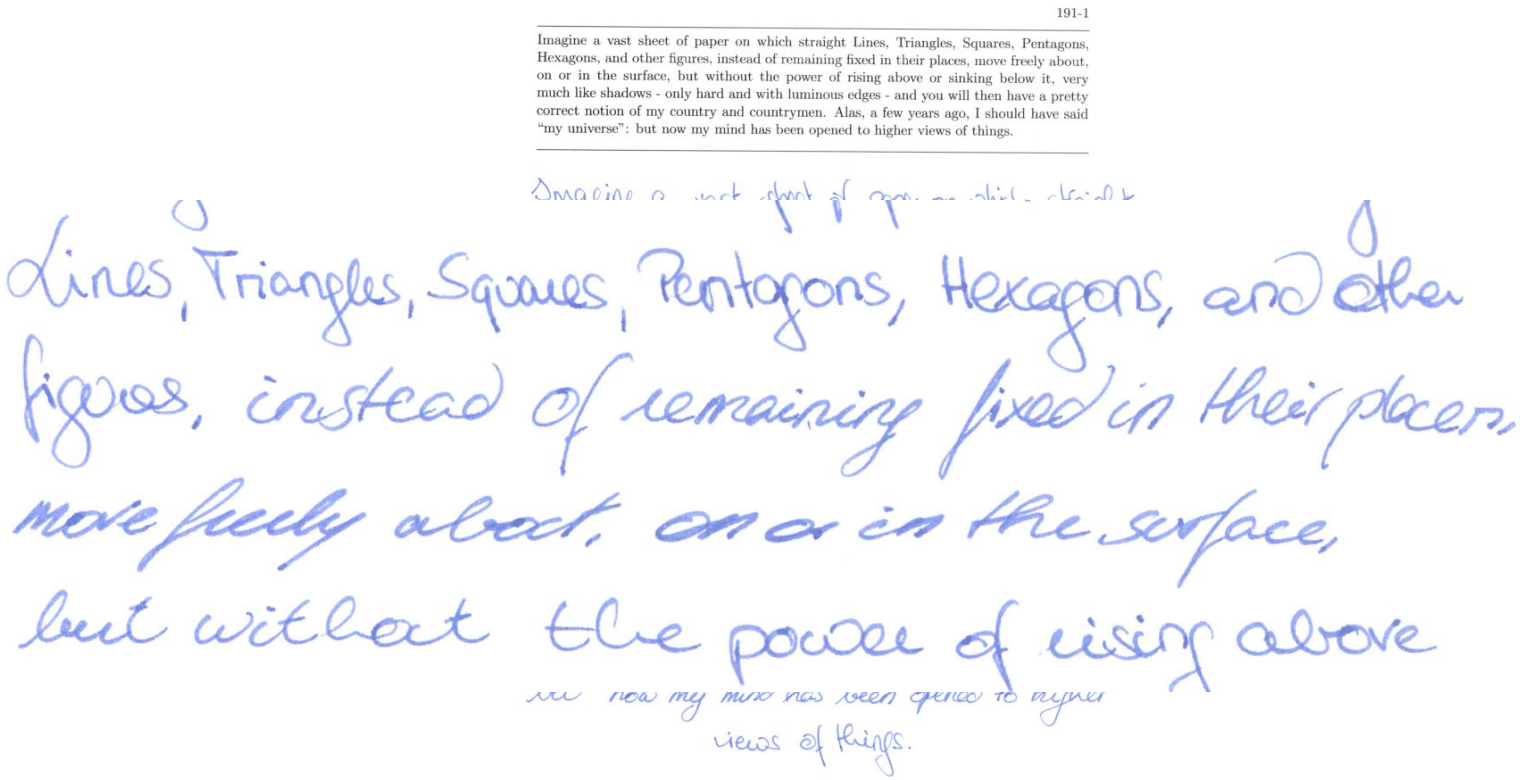
Variations of Handwritings

- *A* affine variation
- *B* allographic variation
- *C* neuro-biomechanical variability
- *D* sequence variability



Schomaker, L. & Bulacu, M. “Automatic writer identification using connected-component contours and edge-based features of uppercase Western script”, PAMI 2004, 26, 787-798

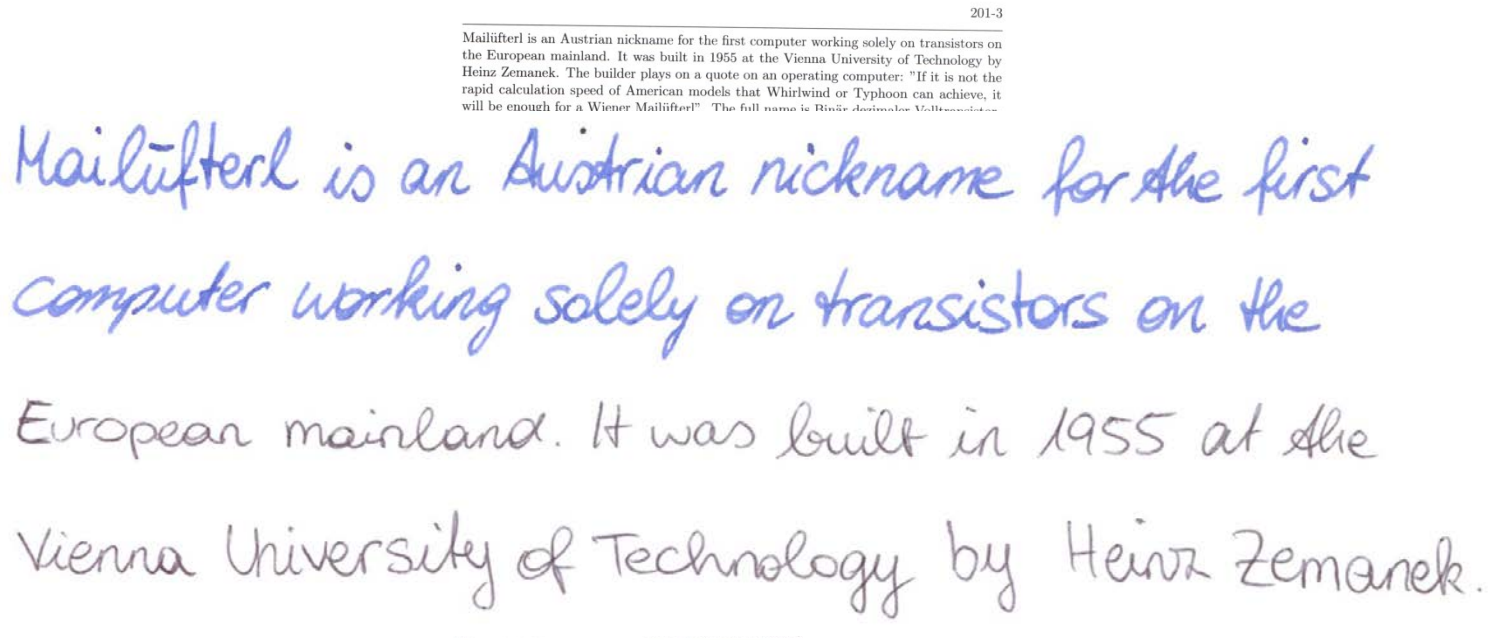
Handwriting Samples / Challenges



191-1

Imagine a vast sheet of paper on which straight Lines, Triangles, Squares, Pentagons, Hexagons, and other figures, instead of remaining fixed in their places, move freely about, on or in the surface, but without the power of rising above or sinking below it, very much like shadows - only hard and with luminous edges - and you will then have a pretty correct notion of my country and countrymen. Alas, a few years ago, I should have said "my universe": but now my mind has been opened to higher views of things.

Handwriting Samples / Challenges

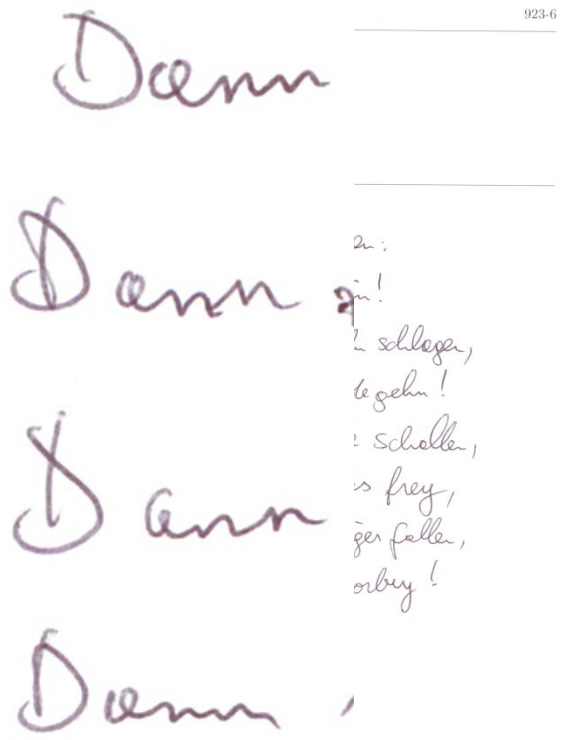


201-3

Mailüfterl is an Austrian nickname for the first computer working solely on transistors on the European mainland. It was built in 1955 at the Vienna University of Technology by Heinz Zemanek. The builder plays on a quote on an operating computer: "If it is not the rapid calculation speed of American models that Whirlwind or Typhoon can achieve, it will be enough for a Wiener Mailüfterl". The full name is Binär dezimaler Volltransistor-

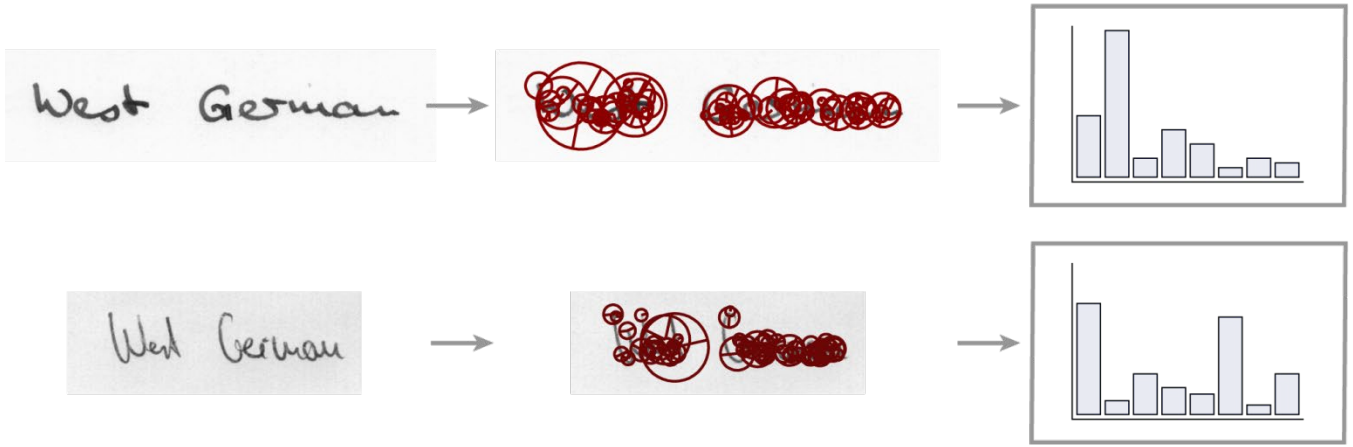
Mailüfterl is an Austrian nickname for the first computer working solely on transistors on the European mainland. It was built in 1955 at the Vienna University of Technology by Heinz Zemanek.

Handwriting Samples / Challenges



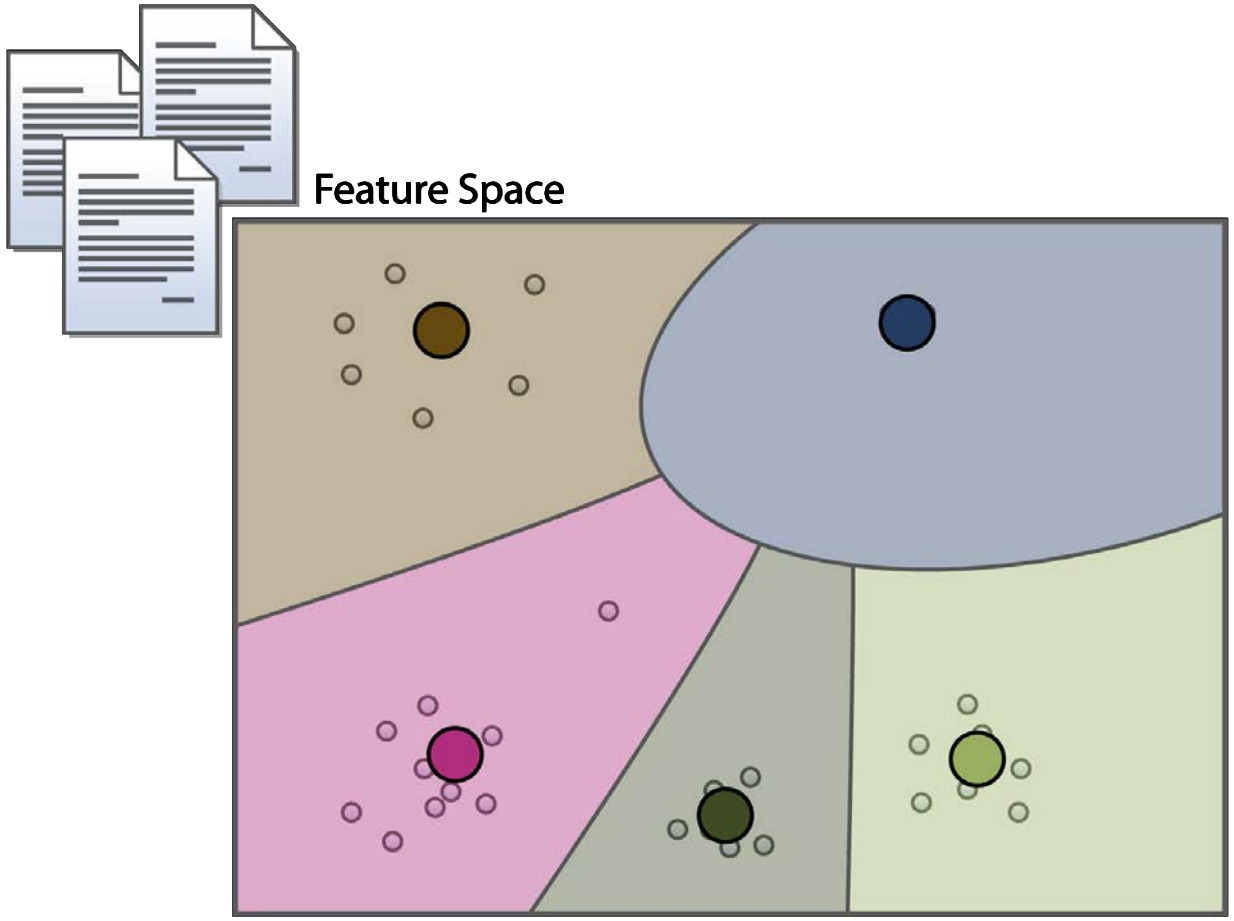
Bag of Words

- SIFT features (describes the neighborhood)
- Comparison with cluster centers of training data
- Generation of occurrence histograms
- χ^2 distance

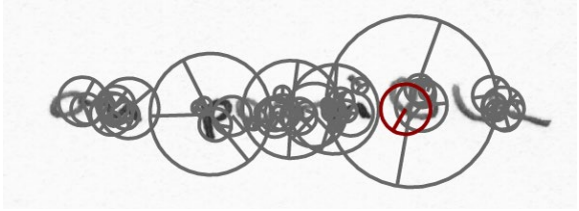


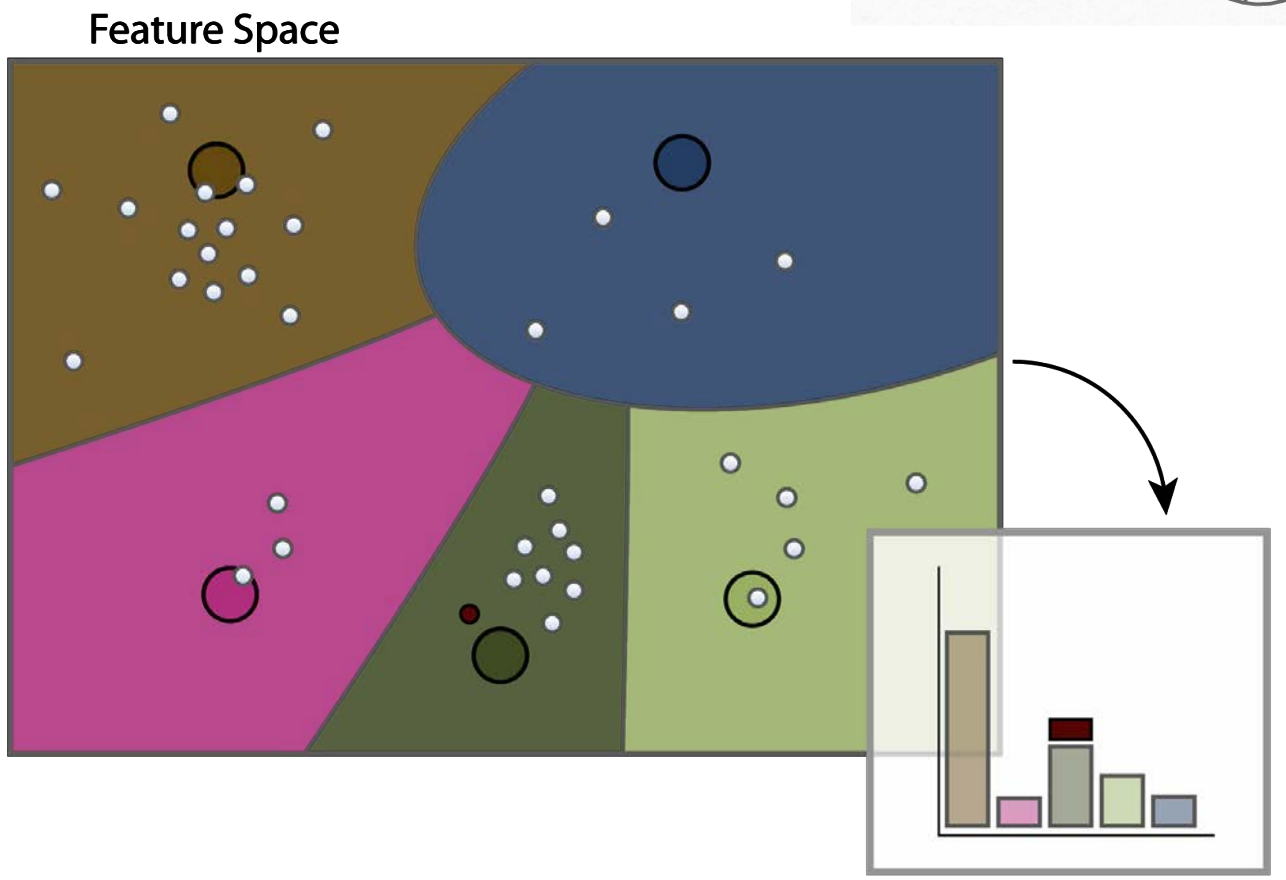
Fiel, S. & Sablatnig, R., "Writer Retrieval and Writer Identification Using Local Features"; *DAS,* 2012, 145 -149

Vocabulary Generation



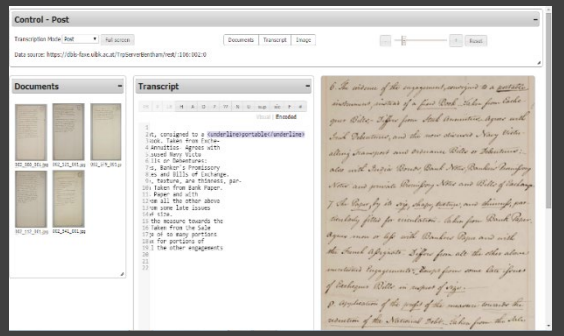
Histogram Generation

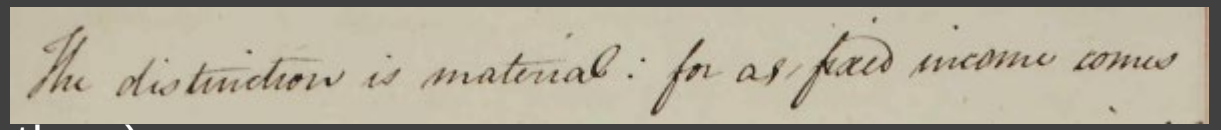




READ

- **R**ecognition and **E**nrichment of **A**rchival **D**ocuments
- Virtual Research Environment
- Automated recognition, transcription, indexing and enrichment of handwritten archival documents





- Partners (amongst others):
  - UIBK (Coordinator), NCSR (Greece), UPVLC (Spain), National Archives Finland, Zurich State Archives, ... (13 partners in overall)
- Details: 1.1.2016-30.06.2019
- Funded by: EU H2020

https://readcoop.eu/transkribus/

READ co·op

Transkribus ScanTent read&search The COOP Resources 0 Sign in EN

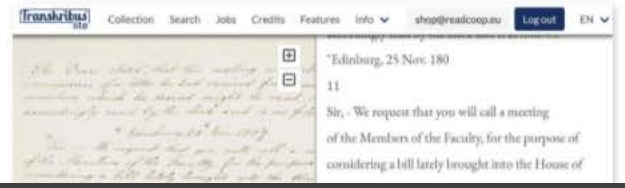
Transkribus

Where AI meets historical documents

Transkribus is a comprehensive platform for the **digitisation, AI-powered text recognition, transcription** and **searching of historical documents** – from any place, any time, and in any language.

Download Sign up for free

Transkribus Lite new



Use Transkribus in your browser. Many of the features from

TRANSKRIBUS

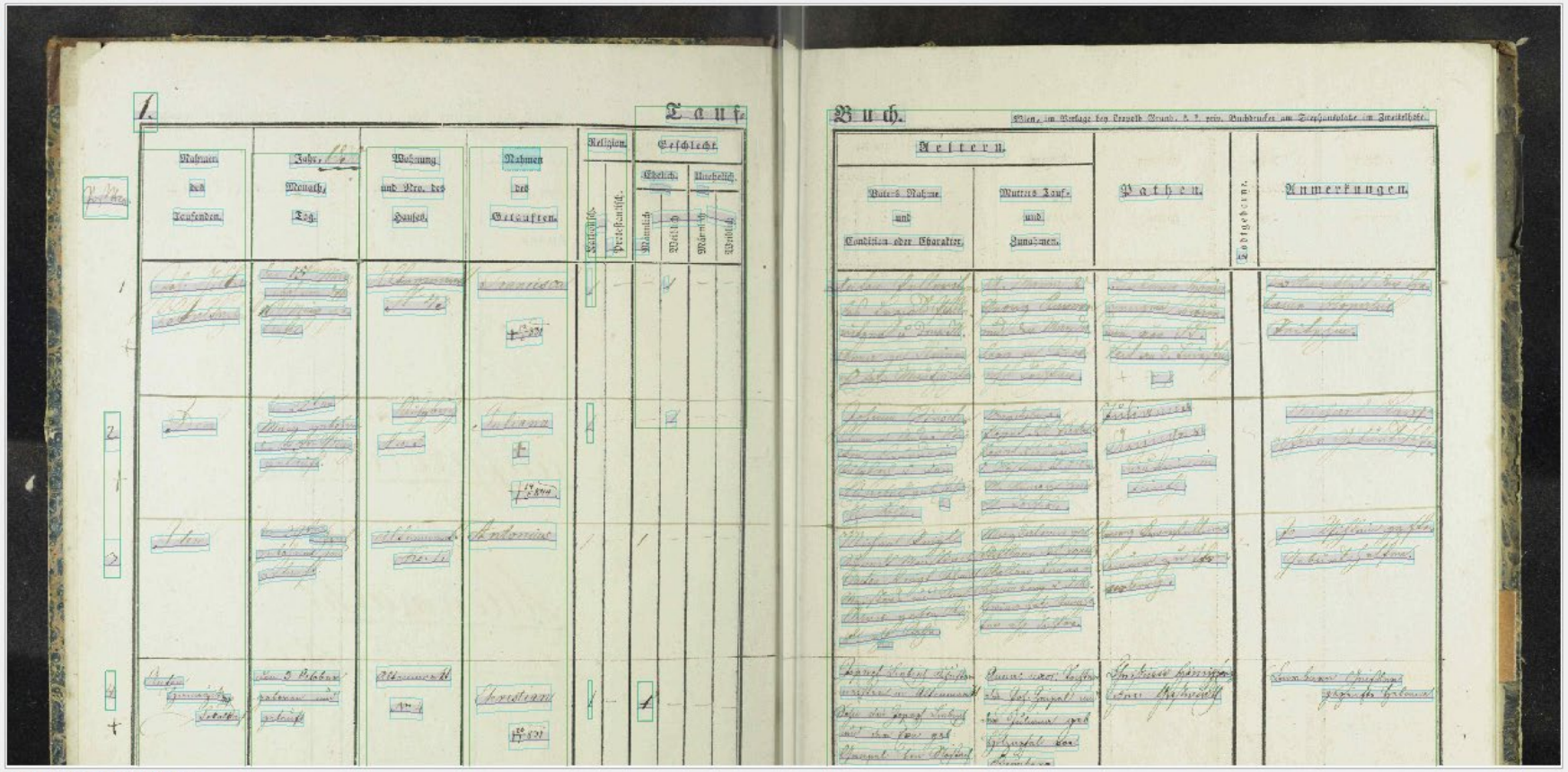
Cloud Platform offering document analysis services

- Manage collections
- Layout analysis
- Transcribe HANDWRITTEN documents automatically
- Search in documents (keyword spotting)
- https://readcoop.eu/

ection-all_for_each_doc_10, ID: 618393, Page 1, file: 0001_page_23778911_docId_595941.jpg [Image Meta Info: (Resolution:1.0, w*h: 6076 * ...

1 /200 In Progress

TR
L
BL
W
...
H
V
L
...



8-12 dtzbach
8-13 Nro 184
8-14 Altermarkt
8-15 No 30.
9-1 Nahmer
9-2 deb
9-3 Getauften.

thank you

Demo