# DOCUMENT IMAGE ANALYSIS

Florian Kleber



- Thousands of UNIQUE documents are stored unpublished in archives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

# DOCUMENTS



# 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



#### 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- Digital copies of historical documents are needed for the digital humanities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





# 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



MSI

#### Image Enhancement



# DOCUMENT ANALYSIS AND RECOGNITION

#### TABLE CLASSIFICATION writer identification YSIS READING LYSIS **OCUMENTS** GR<sup>r</sup> UPING NAL (HTR)similar documents document categorization

#### 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 nonoverlapping 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

> orian Kleber, Markus Diam and Robert Sablatnia Computer Vision Lab Institute of Computer Aided Automation Vienna University of Technology Email: {kleber, diem, sab}@caa.tuwen.ac.at

Abstractskew estimation is a preprocessing step in doct alysis 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, t 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 th contest. In this paper it is shown, that the use of grayvalue images can outperform the results restricted to binarized image thus the proposed method avoids the binarization step which i still an open research topic in document image analysis. Th robustness of the method is also shown on a dataset comprisin 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)

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 II 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 bina-

rization 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. bleed-

through 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

with a line detection and the analysis of paragraphs.

basic FNNC clustering, the proposed approach is combined

#### 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<sup>1</sup>). 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 [11], [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 nethods and that skew estimation is still an open research topic.

<sup>1</sup>www.impact-project.eu, accessed 20.12.2013

# TitleRobustAuthorsFlorianAuthor affiliationComputeAuthor email{Kleber.

Abstract Abstract Text Robust Skew Estimation of Handwritten and Printed...

Florian Kleber, Markus Diem and Robert Sablatnig

Computer Vision Lab, Institute of Computer Aided ...

{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 ....

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

<sup>1</sup>www.impact-project.eu, accessed 20.12.2013



Document Layout Analysis vs. Understanding

#### • Extract the layout structure

- Segment a page into homogeneous regions
- Map the logical structure

#### Layout Analysis

- 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 ge
- Regions of in
- Semi structure
  - Partially defi
  - Some regular
- Loosely struct
  - Cannot be de
  - Logical object









n sieht darn klassiche Anflärichler: Alles zugleich spiele es iht. "Fökussern fökussieren". An die 400.000 Flaschen in 420 Hets s Akaden de beiden machuer soll der Absatz um 40 his 70 zwent stigen. Ziel sie der Schritt die Schweiz und nach Hallen, r. Breakeven sie sich en einma gweisen, nach Inweitlichen aber zämlich fix sigt Wilson di lacht. Viele Leute um sie hen hätten schen vor Jahren geht, tie seien Millionäre. 20 Jahre alt saler funzeilage frageracht nitterweile. Richtig gestracht rierer inter Jahr frageracht frageracht aber zienlicht ein der her die frageracht nitterweile Richtig gestracht rierer inter Jahr frageracht frageracht ist sie die Pauler funzeilen absteinisch bieher noch nicht.

#### *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

			19.27.	Marina Seconery.		
	Nummer	Datum	Aart der Akten.	Namen, Voornamen en Woonplaatsen der Partijen,	Aanteeken Registratie	aing der -Regten.
	Orde.	Akten.	Brevet. Minuten.	Aanwijzing, Legging en Prijzen der Goederen.	Datum.	Regten.
	-30	19	Hout	For descente van fonk het Henerik Beenforstarting	e.st.	14
	S S I S		10	to duyte with the strents	bok °V 122.	3. 78.
	3.0 -	20.	" Senally	Sen Misterike Com Istrenanty Can Carl Anna Canton and Part Bank Clift C. d. and comment Can Anna Kilof C. d. beljenne endel Can Anna and And Carl and Andrea Diedes, to John and the Sen Symbol Diedes, to John and the Senter to	6. str 0 26 fr 6 0 1000	123. ggz
	38.	20	· · · · · · · · · · · · · · · · · · ·	Server de la Commente de la Commente de la contra de la c	6 Alaar 12 26 / e 12 / Klo 14 3.	
/	33.	20.	- Idem	Son de stracke van Johnman Son de Son C. & Ho C. Marty - en Santanger ander Holton Ansender gekantel og for Mare Bernder om Markenen to Marage andersen bend	9 str 0 22 fo 22/2 - 0 x	×e 14.
	34.	-21.	" Acte. Noyent	Dove Johanner Verts hetelyd wan fan en er handrig to hefer he hoeven ten thantfor het Hype, thate, tet Bourg	3.11+ 6.26 /3+ 1' / (×)	1. 11.
	35.	23. e	Prokuro . , ,	Dove John annes Bila to torigis, og fjut dere Huge van Beleiter Hay syrt, om belander hand to Dugt in Lynn an Har oct- te Dugt in Lynn som to Spect, out	24 . Ei @ 26 fr 400 pr 1/2 5.	1.01
	31.	24.	" exiter	as those fantiheer Hugo ver Cheve.	3 Alt 226.	

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, ...



north over the tropical Indian



#### 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<sup>12</sup>





#### Otsu Example



Within Class Variance  $\sigma_W^2 = W_b \sigma_b^2 + W_f \sigma_f^2 = 0.4722 * 0.4637 + 0.5278 * 0.5152$ = 0.4909

#### Otsu Example

Threshold	Т=0	T=1	T=2	T=3	T=4	T=5
	8- 6-	8- 6-	8- 6-	8-	8-	8- 6-
	$\begin{array}{c} 4 \\ 2 \\ 0 \\ 0 \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}$	2- 0-012345	2- 0-012345	2- 0-012345	2- 0-012345	2 - 0 - 0 + 2 + 3 + 5
Weight Background	Wh = 0	$W_{1} = 0.222$	$W_{\rm h} = 0.4167$	$W_{1} = 0.4722$	W = 0.6389	M - 0 8889
Mean Background	$M_{\rm b} = 0$ $M_{\rm b} = 0$	$M_{\rm D} = 0.222$ $M_{\rm b} = 0$	$M_{\rm b} = 0.4107$ $M_{\rm b} = 0.4667$	$M_{\rm D} = 0.4722$ $M_{\rm b} = 0.6471$	$M_{\rm b} = 0.0309$ $M_{\rm b} = 1.2609$	$M_{\rm b} = 2.0313$
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	$\sigma^2_{W} = 3.1196$	$\sigma^2_W = 1.5268$	$\sigma^2_W = 0.5561$	$\sigma^2_W = 0.4909$	$\sigma^2_W = 0.9779$	$\sigma^2_{W} = 2.2491$

T=3

#### Otsu Results

D secundű dubium. Eln liceat emere redditű pecuniariű ad vitá sims pliciter. Potest dici ex méte doctorű comus

# Diceat emere redditű pecuniariű ad vitá lím= plícíter.Poteft díci ex métedoctorű comu=



#### 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 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, vercheter herr und freund, das neue hoft meiner zeiterhrift und sechs bewaden abringe three nachträge zum Silvester. werden Sie wohl zurnen daße ich three erlaubais neue boneskungen hinzusufigen mich to reichlich bedient und gegen den starhel gelerkt hate? verigstens unüberlegt ud unfleifig verden die hoffentlich meine annerskungen nicht finden, us mine vormelunger über 1418 and 4307 vielleicht gatheiften. StAM Marburg Grimmen dem ich für den 3n bit des weithames herzlich danke, lege

Objectives

Br 2864 L. 30 juli 1842. Hier exhalten Sie, vercheter herr und freund, das neue haft meiner zeitschrift und sechs bezondere abringe Ihrer nachträge zum Silverter. werden Sie wohl zurnen daße ich Ihrer erlaubais neue boneskungen hinzusufigen mich so reichlich bedient und gegen den starhol gelerkt habe? wenigstens unüberlegt und unfleifing werden die hoffentlich meine anmenskungen nicht finden, und meine vormulunger über 1418 und 4307 vielleicht gutheißen. Page Segmentation bouder, den ich für den 3n bis des weidhumes horzlich danke, lege

Objectives

Br 2864 1 30 juli 1842 Hier ochetter lie vercheter harr und freund das neue holl meinen zeiterhritt und sechs bewada -to?. the radifier run Silvester. worder Sie wohl zurnes dale ich three extentions neue boneskungen historie find mile in wichtich bedie. N und gagen dens startet oplants tale? and the waitelest us ualleiling ander the hallesthick mise annerthease will bridge a) none vorme lune ile 1415 a) 4307 vielloid authoiles **Baseline Segmentation** dem ich für der 3. bit der weidhimer herzlich den der 1000 1



#### Granularity





#### **Document Clustering**





Institute of Computer Aided Automation, Computer Vision Lab

#### 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, ...

Angelika Garz Computer Vision Lab Vienna Unix of Technology 1040 Vienna, Austria garz@cua.tuwien.ac.at Andreas Fischer Inst. of Computer Scie affischer@iam.unibe.a 3012 Bern, Switzerla. 3012 Bern,	Robert Sablani Computer Vision . Computer Vision . Vienna Unix. of Tech 1 040 Vienna, Aus sab@caa.tuvien.a based on Projection nary and gray-scale i the global PP such merging text lines an Recent approaches known from image Nicolaou and Gatos which follow the line page in lines. Origi [16], Indermuhle et a s, in order to find a pal lines in historical m approach directly on the starsform is comput the spaper, wi free method for line manual for the proposed method	Ing Horst Bunke i Lab Inst. of Computer Science hnology and Applied Mathematics ustria 3012 Bern, Switzerland bunke@iam.unibe.ch profiles (PP) [2], [8]–[11] for both I i images. Various authors [8], [9] adapt that skewed text blocks, converging are segmented correctly. es [12–[14] introduce seam carving [1 e retargeting for text line segmentatic is [14] use so-called local minima tracce ne spacing in order to shred the documer ginally proposed for on-line documer it al. [12] use Dynamic Programming (D) valh with the minimum cost between tv manuscripts. Asi et al. [13] apply tho on gray-scale images, where a distan uted from a Gaussian-blurred image, and ns are established using DP. we introduce an efficient binarization ne segmentation applicable to historic binarization-caused errors frequent t i mages are not inherited. Furthermon hod follows a bottom-up approach ]
Angelika Garz. Computer Vision Lab Vienna Unix, of Technology 1040 Vienna, Austria garz@caa.tuwien.ac.at Abstract—Segmenting page images into text lines. is a c clai pre-processing step for automated reading of histori documents. Challenging issues in this open research field given e.g. by paper or parchment background noise, i bleed-through, artifacts due to aging, stains, and touch text lines. In this paper, we present a novel binarizati free line segmentation method that is robust to noise opes with overlapping and touching text lines. First, inter points representing parts of characters are extracted fr documents. Challenging issues and touching text lines. First, inter points representing parts of characters are extracted fr where neighborhood is defined by the prevailing orientation the Latin manuscript images of the Saint Gall database she promising results for real-world applications in terms of baccuracy and efficiency. <i>Expwords</i> -bistorical documents, manuscripts, ancient do neuts, handwritten, text line segmentation, binarization-fre <i>L</i> INTRODUCTION Automatic segmentation of historical document page i ages is an open research field, algorithms are required to robust with respect to background artifacts such as clause to robust with respect to background artifacts due to aging, a touching or interfering lines [1]. Text line segmentation, particular, is typically needed for handwrittin grecognition to ba interfering lines [1]. Text line segmentation, bistorical documents, manuscript, and but the segmentation, particular, is typically needed for handwritten documents, due to a discus a set as the segmentation, particular, is typically needed for handwritten documents do not has strict layout rules and thus line segmentation inethods no to bartier tayout rules and thus line segmentation is denoting in the segmentation is a served in the segmentation is denoted in the segmentation in terms of bacterian to how intervinter to bacterian in a segmentation in tembods hor to barti	Robert Sablatni <i>Computer Vision</i> . <i>Computer Vision</i> . <i>Vienna Luix of Tech</i> <i>1040 Vienna, Aus</i> <i>sab@caa.tuvien.a</i> <i>asb@caa.tuvien.a</i> <i>based on Projection</i> <i>nary and gray-scale if</i> <i>the global PP such</i> <i>merging text lines an</i> <i>Recent approaches</i> <i>di Nicolaou and Gatos</i> <i>which follow the line</i> <i>sage in ines. Origi</i> <i>lines in historical m</i> <i>approach directly on</i> <i>the starting approach for this</i> <i>manuscripts. Thus,</i> <i>intorical document</i> <i>the proposed mether</i> <i>grouping parts-of-cle</i>	nig Horst Bunke 1 Lab Inst. of Computer Science hnology and Applied Mathematics stria 3012 Bern, Switzerland bunke@iam.unibe.ch bunke@iam.unibe.ch i mages. Various authors [8], [9] adapt i that skewed text blocks, converging are segmented correctly. es [12]-[14] introduce seam carving [1 e retargeting for text line segmentatic is [14] use so-called local minima tracc ne spacing in order to shred the docume ginally proposed for on-line documer it al. [12] use Dynamic Programming (D valh with the minimum cost between tv manuscripts. Asi et al. [13] apply th on gray-scale images, where a distan- tud from a Gaussian-blurred image, an sa re established using DP. we introduce an efficient binarization ne segmentation applicable to historic binarization-caused errors frequent t i mages are not inherited. Furthermon hod follows a bottom-up approach ]
Angelika Garz Computer Vision Lab Viema Unix, of Technology 1040 Vienna, Austria garz@caa.tuwien.ac.at Abstract—Segmenting page images into text lines. is a c clai pre-processing step for automated reading of histori documents. Challenging issues in this open research field given e.g. by paper or parchment background noise, i bleed-through, artifacts due to aging, stains, and touch text lines. In this paper, we present a novel binarizati free line segmentation method that is robust to noise opes with overlapping and touching text lines. First, inter points representing parts of characters are extracted fr documents. Challenging issues are identified in his density regions and touching components such as ascend and descenders are separated using seam carving. Finally, t lines are generated using seam carving finally, t lines are generated using seam carving. Finally, t lines are descented as artifacts due to aging, a locuching or interfering lines [1]. Text line segmentation, particular, is typically needed for handwritten	Robert Sablathi cc Computer Vision . computer Vision . v Vienna Unix of Tech 1040 Vienna, Aus sab@caa.tuvien.a based on Projection nary and gray-scale i the global PP such merging text lines an Recent approaches d known from image Nicolaou and Gatos which follow the line se in order to find a pai lines in historical m approach directly on the starting parts-of-of the menutority.	nig Horst Bunke I Lab Inst. of Computer Science hnology and Applied Mathematics stria 3012 Bern, Switzerland bunke@iam.unibe.ch n Profiles (PP) [2], [8]–[11] for both I images. Various authors [8], [9] adapt that skewed text blocks, converging are segmented correctly. es [12]–[14] introduce seam carving [1 e retargeting for text line segmentatic is [14] use so-called local minima tracc ne spacing in order to shred the docume ginally proposed for on-line documer tal. [12] use Dynamic Programming (D value with the minimum cost between tv manuscripts. Asi et al. [13] apply th on gray-scale images, where a distan- tud from a Gaussian-blurred image, and ns are established using DP. we introduce an efficient binarization ne segmentation applicable to historice binarization-caused errors frequent t i mages are not inherited. Furthermon hod follows a bottom-up approach I
Abstract—Segmenting page images into text lines is a c cial pre-processing step for automated reading of histori documents. Challenging issues in this open research field given e.g. by paper or parchment background noise, . bleed-through, artifacts due to aging, stains, and touch text lines. In this paper, we present a novel binarizati free line segmentation method that is robust to noise a copes with overlapping and touching text lines. First, inter points representing parts of characters are extracted fr gray-scale images. Next, word clusters are identified in hi density regions and touching components such as ascend and descenders are separated using seam carving. Finally, t lines are generated by concatenating neighboring word clusters where neighborhood is defined by the prevailing orientation the words in the document. An experimental evaluation the Latin manuscript images of the Saint Gall database she promising results for real-world applications in terms of b accuracy and efficiency. <i>Keywords</i> -historical documents, manuscripts, ancient do ments, handwritten, text line segmentation, binarization-fre <u>I. INTRODUCTION</u> Automatic segmentation of historical document page i ages is an open research field; algorithms are required to robust with respect to background artifacts such as clut tains and noise, as well as artifacts due to aging, a touching or interfering lines [1]. Text line segmentation, particular, is typically needed for handwriting recognition to be inverting to lavout inconcentencines.	<ul> <li>based on Projection any and gray-scale is the global PP such merging text lines an Recent approaches known from image</li> <li>Nicolaou and Gatos</li> <li>Nicolaou and Gatos</li> <li>Nicolaou and Gatos</li> <li>Inos. Origi</li> <li>Ifol, Indernuthle et a s, in order to find a pair</li> <li>lines in historical</li> <li>in order to find a pair</li> <li>in starting parating scamas</li> <li>In this paper, wi free method for line manuaritys. This, historical document the proposed method</li> </ul>	n Profiles (PP) [2], [8]–[11] for both l e images. Various authors [8], [9] adapt that skewed text blocks, converging are segmented correctly. es [12]–[14] introduce seam carving [1 e retargeting for text line segmentatic s [14] use so-called local minima trace ne spacing in order to shred the documer ginally proposed for on-line documer tal. [12] use Dynamic Programming (D bath with the minimum cost between to manuscripts. Asi et al. [13] apply th on gray-scale images, where a distan uted from a Gaussian-blurred image, at ns are established using DP. we introduce an efficient binarization ne segmentation applicable to historic d, binarization-caused errors frequent ti mages are not inherited. Furthermon hold follows a bottom-up approach l
I. INTRODUCTION Automatic segmentation of historical document page i ages is an open research field; algorithms are required to robust with respect to background artifacts such as clut stains and noise, as well as artifacts due to aging, a touching or interfering lines [1]. Text line segmentation, particular, is typically needed for handwriting recognition historical documents. Handwritten documents do not ha strict layout rules and thus line segmentation methods ne to be inverting to layout inconcentencing. imputations	historical document the proposed metho grouping parts-of-ch	, binarization-caused errors frequent it images are not inherited. Furthermon hod follows a bottom-up approach l abarrate interact points into text l
Automatic segmentation of historical document page i ages is an open research field; algorithms are required to robust with respect to background artifacts such as clutt stains and noise, as well as artifacts due to aging, a touching or interfering lines [1]. Text line segmentation, particular, is typically needed for handwriting recognition historical documents. Handwritten documents do not ha strict layout rules and thus line segmentation methods ne to be ingrigate to layout inconsidencies.	the proposed metho grouping parts-of-ch	hod follows a bottom-up approach
script and writing style, skew, and fluctuating text lines [ Furthermore, robustness to low contrast and rippled pay is required [2], [3]. Likforman-Sulem et al. [1] provide a detailed surv about segmentation of text lines with respect to histori documents. Well-known methods for text line segmentati in binary images include smearing [4], [5] and Hou transform [6], [7]. Another commonly used approach	e regions. Hence, it i regions beforehand, d special page layouts n Touching component locally split by mean e The experimental d is carried out on t n contains 60 pages of 9 <sup>th</sup> century written i s with ink on parchm in Figure 1. Besides y colored initial letter and margin, which wern text line spacing is height and a regula is However, there are li and descenders. Sta	tranacter interest points into text in is not necessary to extract text blo l, which can also be prone to errors f its encountered in historical documen nts such as ascenders and descenders a l evaluation of the proposed approa the Saint Gall database <sup>1</sup> [17], whi of a Latin manuscript originating from t in Carolingian script by a single writ ment. Two sample pages are illustrat es the main text body, the pages conta ers and annotations located in the out re added to the manuscript later. T is relatively large compared to the wo lar page layout is present for all page line interconnections caused by ascende atins, holes, and ink bleed-through po
This work has been supported by Austrian Science Fund (Grant P231	and descenders. Sta	ams, noies, and ink bleed-through po



### 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



neiner hierigen stellung darf ich hoffen. Ich clante mis en haft für Lachmann und eine für Mansebach zu giliger begorgung beizuleger; mit dem letzteren hats seine ale, Meusebach liest es je doch mikt eher es gebunden ist. Mit harsticher griften en die Untigen

1h

getrener

Most Haupt

methodology

North Congo Congo Con Hollow. he where the sing of topy the is a the the is , the top of a just 12 super pixel extraction

Hessisches Staatsarchiv Wandurg, Besh. 340,G

local orientation [11 Koo] nur befleckt werden durch ungesetsliche Hant hviel welcher gesellschaftlichen Region er angehör 11 Marsholan Sie Colone white the second

Quell gefallenen Staatsdieners oder Offiziers, n t denn den zum Krüppel geschossenen oder ge llanten, soferne er im Staate angestellt ist? ver der Steuerzahler. Und ist denn die Ehr stohrigen Jünglings, der noch aus der Tasche ere ezes es 67.5 1.90 112.5 1.95 197.5 paratding, daß ? Wenn die Sache nicht so ernst wäre, müß einen solchen Privatehren-Coder wirklich hellauf Benn ein Verein sich die Aufgabe stellt, diesem äßen Unfug an den Leib zu gehen, so ist das ückwünschen, denn solche mittelalterliche Hans 1 und Spielereien mit dem Menschenleben sind eine so ernste geit wie die heutige, wo es 2 un gibt, als über eine sogenannte Separate en, nicht mehr am Place. Die Ehre kann in grobe sittengesetzliche Verstöße. Der nächstbeste

meiner Gesigen stelling mary of hoffen! the clarke mis on test per Larkman in and for mancher to patiget Best besongung beizulegen ; mit dem leteren lats seine ale , Meansack west es for dock mint ene es getunden en Mut heresides profiles on die setrency local orientation  $\bigcirc$ 

meiner hesigen stelling our ich hoffen! the create mis on left juis Lashmann in and for heareback in juliger Best bejorgang beizalegen ; mit dem lateren hats seene eile , elleuresail biegt es jie doch micht eder es gedanden en Mit sorationer profiler en die S C D S C MRF optimization 0000  $\bigcirc$ 

setrency



noine hieron stellung dark ich hollow Il clarke mis en hall fin Lachmean and aix fin Mensehach in eitige besongung beizulosen; mit dem letzleren hate seine ale Meusebach hiert or je doch zikt ches es getunden ist. Mix harslicher griften en die their the estress. baseline estimation

340

(1)

Detection

0-00000-000000 10

week-and-make-decisions-based on-what they-believed was ran sweek-and-make-decisions-based on what they believed was ran 0 applied as did Sareet and rate of they are and one of 0 applied as did Sareet and 100, o Oct Sweet construction of a point of a walk of the they are an are and other to be an other to a walk of the they are an are an are and the to based on investor-demand, market conditions, and other factor, based-on-investor-defined a company's stock-to-the public sets-the-initial-price for selling a company's stock-to-the public sets the initial protective issuing company and the public the state issuing company and the public the state of the state The-tension the Wall Street brokerage touse the an intension of the Wall Street brokerage touse the an intension of the self o Result on principle offerages of each offerage offerage offerage offerages of each offerages offerages of each offerages offerages offerages of each offerages offerages offerages offerages of each offerages offerages of each offerages offerages of each offerages offerages

In the annals of Wall Street, no business had ever done a success to the annals of Wall one way tarry and Sergey wanted to do in fid-billion-collar tPO the way tarry and Sergey wanted to do in fid-billion-collar them at all. Accustomed to creaming and do in That-didn't-scare-them at all. Accustomed to creaming and days mat-clide't scare them at they were determined to have a function to the start of t

# ber to bard to ber tor bet.

o Google. Hibey were doubte operative and their elerished o If they were an interest agreed to do it their way just as they be recy tarty and Serbey agreed to do it their way just as they be recy tarty and Serbey agreed to do it their way just as they be when taking and going to tell-shear how to do the real. It cides else was going was out de the caire of the carding one our appendict of the testing of the total of the testing of testin tor Oat moto unance seemed simple compared to building a captor of the Walt Street staff seemed simple compared company a captor of the second The Walt Street station employees, and ranning a capidly, search engine, more approximation employees, and ranning a capidly, search engine, and any so-as-fay-as-they-were-concerned, ing profitable-company would find a way to maintain total control over Google and and would find a way to maintain more to them than filling and an would the a worto matered more to them than filling und an entry of the matered more to them than filling und an entry occording the door about bace and conce do the door and a the according them they were going to go running to Wait bat thit dim't mean the were going to go running to Wait Su

-THE-GOOGLE STORY 170 to bable to othe cable able able able able the the de the would have more resources to grow and realize their o

Connectives time to proceed the all contractions o (10 contrast would feel good about the deal and be more bit mores later, d-and when the contrast line big as shares later, if and when the company needed to

We der states of the states of a Contract shares and charged printing for the for the of the unerbicat all evil where Wait Street underst concentrical all evil where Wait Street underprised 1200 one and proved circuits profit-by-curaging-the-stock-on-day pen the prior sector lieted. The arts wanted no part of what

pro te a compliand totten viston. Supremety confident, they placed their trust more in mathe, energy and technology than they did in sy at enables obtained the lock. In thet, to the internet and the lock of the loce of the loc part time-understanding why Wall Street war-sell raid ad a hard the papies-toe-olo-factorie way auto the same base press in out do the job better, but the Wall Street had come that could do the job better, but the Wall Street crowd preperbat comess the way it always had putting a premi to along as no major investment house changed de de de de de la contra contr 

por of Well Street's vantage point, a company like Google-that From Wall brash, and well-known enough to attempt to set its was powerful, or going -public was rare. Wore often, businesses even reles to be hand holding, advice, and a proven way to get needed road listen to their story in order to get financings cone. investors to the offering, tather than directly out of the ness the offering, rather than directly out of the com-

bever to its thereight affect ag state and the date the Sec. Drever, and an entirely different method for distributing stock General as the state of the second of all the state of th

Majus Majus. Copulans. Ferts 2from 13. huins à Joanne Bapt. Michael Joannis Shigh Annashinghattheast inny proprietary Jam Miller coop: Supernum: Inquilin: out thing Aming mang. Michdel Mitmal Japiel in my of these for Consider her Sfondminner fittene lara tosephi with when Mary S.S. hurd is you quetin in the Spaterne Arter Congo: + Martinen the p:m: go for ghart Darbard lonjug Cil: Legitime Annins. James Contract Junius. Pom Cathol: Matthing 7. huins à francisco Jeraph Josephus Josephi strin nrentino & frindinger Ma: no Sant Cooperatore An ourling in ohn Valloning Jabriel Onlan domunice Clionberthe Confus amb gov ni amarfall. l' Legitina and a second Anna Mana Josephi Co anna freguelin: on thitter Coming Baniff et Stan amp. vir: filia legitima. Vallias Mary Colonas Rom: Cat Di: 9. Jell ar hums a form: Nicolans Matthei Chief hinnarthurtin In finding finding : 1: Miller coop: Super Colon: in Julna fining Hartholomons Anny ne pai p:m: et Mano conjug; Coloant for Min Albrin. en: filins logit: Cam Dorothea Schaffiant wormer Cafe bring grains

tables

ucy

#### degraded documents

#### Document Analysis Tasks Writer Identification & Retrieval

Institute of Computer Aided Automation, Computer Vision Lab

*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

![](_page_50_Picture_7.jpeg)

![](_page_50_Picture_8.jpeg)

#### 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

![](_page_51_Figure_7.jpeg)

![](_page_51_Picture_8.jpeg)

### 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

Vir beklagte for Breifin ningelogt mit dem Unterge, unter Auffebring tob angeforflanen Rolfeilb norf iforme bruifingbons Songe zu noknunne, mogagen Sie Elägnsin Zusüchneiping der Charaction brokly

![](_page_52_Picture_8.jpeg)

### 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)

![](_page_53_Picture_5.jpeg)

![](_page_53_Picture_6.jpeg)

## 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

![](_page_54_Picture_11.jpeg)

Word' ich zum Augenblicke soga: Verweile doch! du bist so schöh! Dann magst du mich in Fesch schlogn, Dann will ich gern zu Grude gehn! Dann mag die Todkuglocke schallen, Dann bist du deines Diensts frey Die Uhr mag stehn, der Zeige faller, Es sey die Zeit for mich vorbey!

![](_page_54_Picture_13.jpeg)

![](_page_54_Picture_14.jpeg)

### Variations of Handwritings

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

![](_page_55_Figure_5.jpeg)

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

![](_page_55_Picture_7.jpeg)

### 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. Smaline a mot don't of man and a charal to dines, Triangles, Squares, Pentopons, Hexagons, and Other 1900s, instead of remaining fixed in their places, more fully about, on a in the surface, but withat the power of isin above me now my mino now ween quence to regul rieus of things.

![](_page_56_Picture_2.jpeg)

### Handwriting Samples / Challenges

Mailuifter is an automation to the first computer within out of the intervent of the interv

201-3

![](_page_57_Picture_2.jpeg)

### Handwriting Samples / Challenges

923-6 )ann Dann ju! Legen, Egehn! Sann Schelle, Sen frey, ger felle, orby!

![](_page_58_Picture_2.jpeg)

### Bag of Words

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

![](_page_59_Figure_5.jpeg)

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

![](_page_59_Picture_7.jpeg)

#### Vocabulary Generation

![](_page_60_Figure_1.jpeg)

![](_page_60_Picture_2.jpeg)

#### Histogram Generation

![](_page_61_Figure_1.jpeg)

![](_page_61_Picture_2.jpeg)

# READ

- Recognition and Enrichment of Archival Documents
- Virtual Research Environment
- Automated recognition, transcription, indexing and enrichment of handwritten archival documents

![](_page_62_Picture_4.jpeg)

The distinction is material : for as fait income comes

- 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

![](_page_63_Picture_0.jpeg)

#### 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.

![](_page_63_Picture_4.jpeg)

![](_page_63_Picture_5.jpeg)

![](_page_63_Picture_6.jpeg)

Use Transkribus in your browser. Many of the features from

![](_page_63_Picture_8.jpeg)

# TRANSKRIBUS

# Cloud Platform offering document analysis services

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

![](_page_65_Figure_0.jpeg)

thank you

![](_page_66_Picture_1.jpeg)