FAKULTÄT
FÜR !NFORMATIK

Faculty of Informatics

Technical Report
CVL-TR-12

Document Image Analysis Preprocessing of Low-Quality and Sparsely Inscribed Documents

Florian Kleber

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

February 28, 2014

Acknowledgements

First of all, I would like to thank my doctoral advisor Prof. Dr. Robert Sablatnig. Besides all his technical and human support he gave me the possibility to work in an interesting field in research and the facility to work as a research assistant at the Institute of Computer Aided Automation, Computer Vision Lab at the Vienna University of Technology. He is the one who's responsible for the successful ending of this work. Thanks a lot Robert, I know what you've done for me.

I would like to thank Dr. Basilis Gatos for reviewing my thesis. Special thanks to Allan Hanbury for proofreading my thesis and Martin Kampel for his support during my studies.

I wish to express my gratitude to Prof. Dr. Heinz Miklas, who gave me insights into the humanities and historical manuscripts.

I would like to say "thank you" to all my colleagues at the CVL, for all scientific and non scientific discussions. Special thanks to Rainer, Sebastian, Michi, Andi, for inspiring discussions during playing darts and Fabian for his calm human behaviour, while we were playing darts in his office. I also would like to thank Melanie, where i knew, that the way was always worth going: "*Das ist Russisch*".

Special thanks to Markus and Stefan (although he never did his master party) who always motivated my work and have always been a great backup. Additionally, I want to thank Markus and Stefan for fruitful discussions, where the *Montagsbier*, the 3 monkeys, prime numbers and a lot of other things have been invented, mostly, during the *Montagsbier*. Thanks!

I would also like to thank Sabine for her patient support during my studies. Finally, I want to express my gratitude to my family, especially my mother, who supported me during all the years.

Ich erfuhr Dinge, die ich nie als notwendig zu Wissen erachtet habe...

Kettcar

Abstract

The mass digitalization of libraries, national archives or museums needs an automated processing of the acquired image data for a further preparation (indexing, word spotting) and improving the access to the content, thus a document analysis. Projects and institutions that are dealing with the digitalization of documents are amongst others the manuscript research center of Graz University (Vestigia), Improving Access to Text (IMPACT), or projects like Google Books of Google Inc.

Document preprocessing is one of the most important steps of document image analysis and is defined as noise removal and binarization, thus foreground/background separation. An additional preprocessing step is the skew estimation of documents which can be based on binarized images or on original grayvalue image. Uncorrected documents can affect the performance of Optical Character Recognition (OCR) and segmentation (layout analysis) methods. Document classification can be used for automated indexing in digital libraries by classifying all e.g. "Table of Contents" pages or allows a document retrieval on large document image databases. By classifying document types, a-priori knowledge (position of text boxes) can be incorporated into the document image analysis system, thus facilitating higher-level document analysis. While binarization and skew estimation are defined as classical preprocessing steps, form classification is added as a preprocessing step within this thesis. The research within this thesis deals with this three preprocessing steps for ancient and historical documents with sparsely inscribed information (printed or written text). Historical documents can be degraded (e.g. faded out ink or noise like background stains) or fragmented due to their storage conditions. The methods are evaluated using state of the art metrics and are compared to methods of current document image analysis contests regarding binarization and skew estimation.

Kurzfassung

Aufgrund einer steigenden Digitalisierung von den Beständen von Bibliotheken, Handschriftenabteilungen (altertümliche Manuskripten), oder per Hand ausgefüllte Formulare gibt es die Notwendigkeit der automatischen Verarbeitung von digitalen Bildern von Dokumenten. Projekte wie Google Books of Google Inc. oder IMPACT (Improving Access to Text) benötigen automatisierte Systeme der Dokumentenanalyse.

Zu den Vorverarbeitungsschritten in der Dokumentenanalyse von Bildern gehören die Binarisierung (Einteilung in Vordergrund und Hintergrund) und die Detektion der Dokumentausrichtung. Eine Formularklassifikation erlaubt die Extraktion von Formularfeldern aufgrund der MetaInformation (Position der Formularfelder) von bekannten Formulartypen. Binarisierung als auch die Korrektur der globalen Ausrichtung sind wesentliche Vorverarbeitungsschritte für die Layoutanalyse als auch der Zeichenerkennung (OCR). Eine Formularklassifikation erlaubt einerseits das Sortieren von Dokumenten und ist ebenfalls ein Vorverarbeitungsschritt für die Layoutanalyse (z.B. Form Dropout). Diese Dissertation beschäftigt sich mit den drei genannten Dokument-Vorverarbeitungsschritten, wobei vor allem schlecht erhaltene (historische, altertümliche) Dokumente als auch Dokumente mit geringem Inhalt (wenige Worte) betrachtet werden. Die entwickelte Methodik kann dabei zum Beispiel auf Dokumentfragmente angewendet werden, wodurch eine Rekonstruktion von “zerrissenen” Dokumenten ermöglicht wird. Die erforschten Methodiken werden mit State of the Art Metriken evaluiert und mit Methoden die im Rahmen von Contests präsentiert wurden verglichen.

Contents

CHAPTER 1

Introduction

To preserve cultural heritage and human knowledge in the form of written texts and printed books, libraries, national archives and projects like Google Books of Google Inc. [36] started a mass digitalization. Additional projects, like Improving Access to Text (IMPACT[1]) and manuscript research centers (e.g. Vestigia - The Manuscript Research Centre of Graz University[2]), have the aim to digitize and improve the access to historical documents. The acquired image data needs an automated processing (Document Image Analysis (DIA)) and additionally in the case of historical documents a digital restoration [36]. The research topics of this thesis comprise DIA preprocessing, specifically document binarization, document skew estimation and form classification.

1.1 Motivation

Document preprocessing is the first step of DIA systems and is defined as noise removal and binarization [118, 171]. Additional preprocessing steps of DIA systems are a skew estimation [3, 12, 112] and document classification, e.g. form classification [26]. The skew estimation can be based on binarized images or on original grayvalue images. Uncorrected documents can effect the performance of Optical Character Recognition (OCR) and segmentation [135]. Document classification can be used for automated indexing in digital libraries by classifying all "Table of Contents" pages or allows a document retrieval on large document image databases [26]. Chen and Blostein state that "*document classification is used to tune Optical Character Recognition (*OCR*) parameters, or to choose an appropriate* OCR *system for a specific type of document*" [26]. By classifying document types a-priori knowledge can be incorporated into the DIA system, thus facilitating higher-level document analysis [26]. While binarization and skew estimation are defined as classical preprocessing steps, form classification is added as a preprocessing step within this work due to the definition of Chen and Blostein [26]. Document

[1] `www.impact-project.eu/`, accessed 22.09.2013
[2] `www.vestigia.at/`, accessed 22.09.2013

binarization is a research topic for ancient manuscripts and historical documents due to degradations and the achieved results of state of the art methods are summarized in the Document Image Binarization Contest (DIBCO) (see Section 2). The research on skew estimation is summarized in the first Document Image Skew Estimation Contest (DISEC) which uses a dataset of entire (mainly fully) inscribed printed documents. Thus, the methods of DISEC are not evaluated regarding fragmented documents (sparsely inscribed) and handwritten content.

Ancient Manuscripts can be degraded due to their storage conditions: faded-out ink or noise like background stains can arise from environmental effects like mold or humidity [80]. Exemplarily, the Missale Sinaiticum, a manuscript from the 11th century, has been exposed to water during a fire-fighting at St. Catherines Monastery [81]. Single folios of this manuscript have also been fragmented. A different example of historic valuable fragmented documents are the records of the Stasi. The Stasi was the Ministry for State Security of the German Democratic Republic (GDR, East Germany). The documents were fragmented in 1989 when Stasi officers tried to destroy secret files shortly before the fall of the Berlin Wall [124]. In total, about 600 million snippets of Stasi documents were discovered after the fall of the Berlin Wall in addition to complete Stasi documents. The Fraunhofer Institute for Production Systems and Design Technology (IPK) Berlin is investigating methods for the reconstruction and has developed a system for the reassembling of torn Stasi-files [124, 158]. For an efficient matching additional features like the skew, and the document foreground (text lines, etc.) based on a binarization of fragments are used beside shape information. The collapse of the historical archive of the City of Cologne [31], lead also to fragmented manuscripts (a total of 18 shelf kilometers of books have been destroyed) which have to be restored. Within this thesis the binarization of ancient manuscripts or historic documents, as well as a skew estimation which can be applied to paper fragments (sparse content) are investigated. Additionally, the classification of fragmented and reconstructed form documents is studied.

Howe defines binarization as a subjective and ill-posed problem [69]. However, binarization is a preprocessing step for OCR, layout analyis, document classification, and e.g. skew estimation. One goal of binarization is to reduce the image to a black and white representation, since "*most documents are produced using monochromatic ink on paper, and their meaning is embodied solely in the distribution of the ink*" [70] which is represented by a binarized image. A further advantage of binarized images is the reduced storage for large archives. Global binarization methods, such as Otsu [134], can be used for scanned documents which are well preserved and have a typical bimodal grayvalue distribution. Background variations lead to improved methods such as Niblack [123] and Sauvola [155], which are milestones in document image binarization [69].

The ongoing research of binarization methods is summarized within the DIBCO [53, 141, 142, 144, 145]. The DIBCO is held within the Framework of the International Conference on Document Analysis and Recognition (ICDAR). In conjunction with the International Conference on Frontiers in Handwriting Recognition (ICHFR) a Handwritten-Document Image Binarization Contest (H-DIBCO) is established. The evaluation metrics of the contests are summarized in Section 2.4 and the results of the contest are summarized in Section 2.5.1. Degraded ancient manuscripts can have a high variation in the contrast of the image. In addition background clutter can produce errors if global methods are applied. Beside ancient documents,

Figure 1.1: Images of the DIBCO benchmark datasets with binarization problems (a)-(d).

printed carbon copies can also contain noise (e.g. historic valuable records of the Stasi [158]). The benchmark sets of the contests contain images of handwritten and printed test representatives of potential problems, which are defined as follows:

(a) Bleed-through text

(b) Background variation

(c) Different text stroke widths

(d) Low contrast

(e) Different paper structure as well as lined/checked paper

(f) Distortions/background clutter

(g) Image compression artefacts

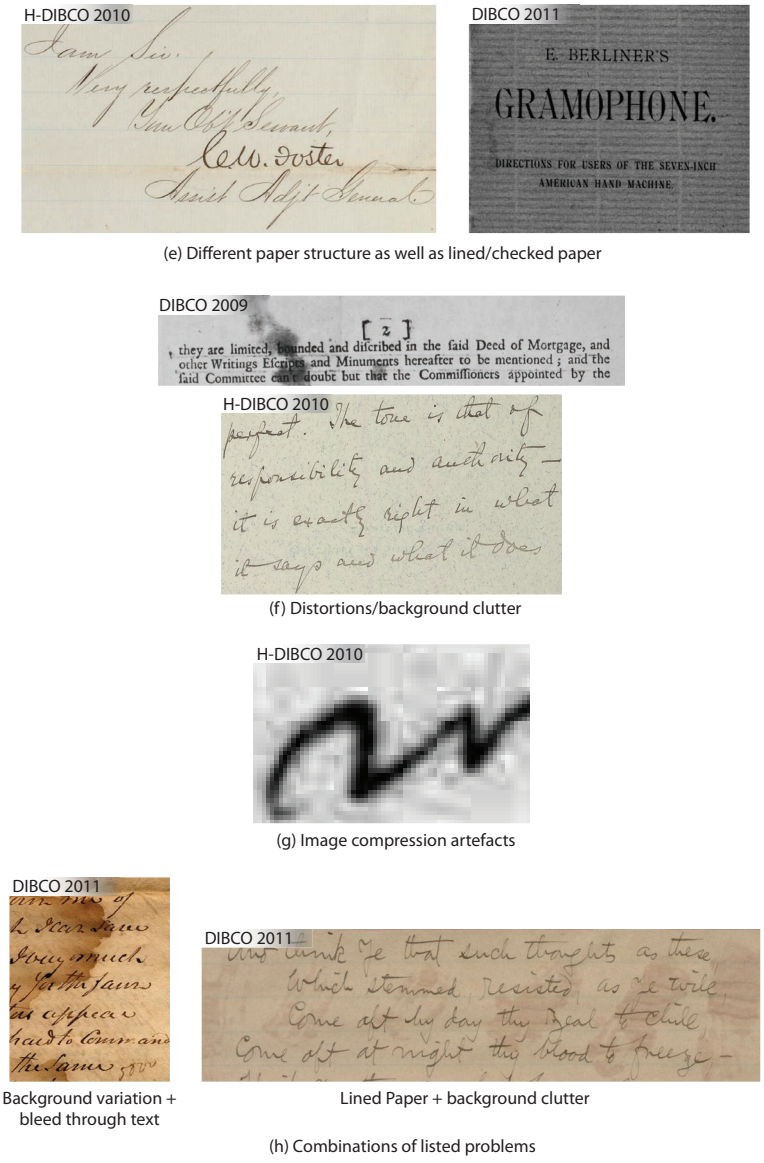


Figure 1.2: Images of the DIBCO benchmark datasets with binarization problems (e)-(h).

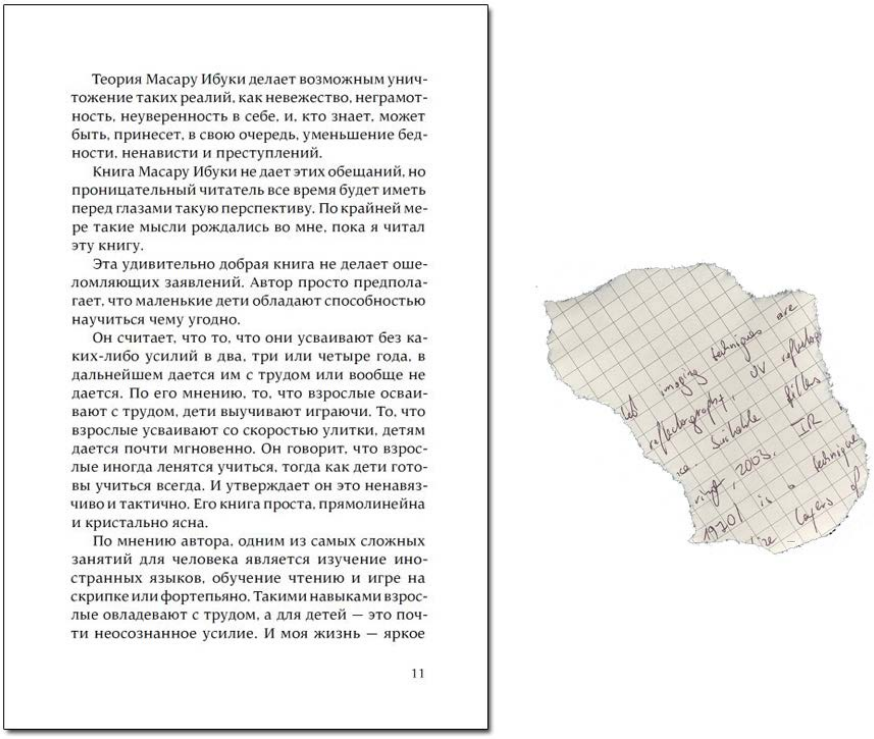
(h) Combinations of the listed problems

Figure 1.1 and Figure 1.2 show images of the DIBCO benchmark set related to the defined problems (a)-(g).

The results of state-of-the-art binarization methods at the DIBCO and H-DIBCO show "*that there remains room for improvement in the quality of automatic binarization*" [70]. Especially bleed-through text (see Figure 1.1 a) and a paper structure with Gaussian noise in combination with low contrast text (see Figure 1.2 e) are currently challenging problems as shown

by the results of methods submitted to DIBCO and H-DIBCO. The research contribution of the thesis within the topic document binarization is the exploitation of a Gaussian scale space to avoid the estimation of text specific parameters. It is shown that by propagating information through the scale space a parameter free binarization can be established. Additionally, evaluation metrics are discussed.

Skew estimation is used as a preprocessing step of DIA systems. A skew corrected page is more pleasant for visualization [105] and is needed for OCR [135] and layout analysis [33, 83]. State of the art methods submitted to the first DISEC within the framework of the ICDAR 2013 are restricted to binary input images. The skew angle of the test images is within a range of $\pm 15\,^{\circ}$. However, Ephstein [44] states that current skew estimation methods must be able to deal with no restriction on the skew angle due to OCR methods in mobile applications (Google Goggles, iBing Vision) that use images of mobile devices (e.g. smart phones). Further, the skew of sparsely inscribed documents (e.g. fragments or images with a few words) must be detected correctly.


Теория Масару Ибуки делает возможным уничтожение таких реалий, как невежество, неграмотность, неуверенность в себе, и, кто знает, может быть, принесет, в свою очередь, уменьшение бедности, ненависти и преступлений.

Книга Масару Ибуки не дает этих обещаний, но проницательный читатель все время будет иметь перед глазами такую перспективу. По крайней мере такие мысли рождались во мне, пока я читал эту книгу.

Эта удивительно добрая книга не делает ошеломляющих заявлений. Автор просто предполагает, что маленькие дети обладают способностью научиться чему угодно.

Он считает, что то, что они усваивают без каких-либо усилий в два, три или четыре года, в дальнейшем дается им с трудом или вообще не дается. По его мнению, то, что взрослые осваивают с трудом, дети выучивают играючи. То, что взрослые усваивают со скоростью улитки, детям дается почти мгновенно. Он говорит, что взрослые иногда ленятся учиться, тогда как дети готовы учиться всегда. И утверждает он это ненавязчиво и тактично. Его книга проста, прямолинейна и кристально ясна.

По мнению автора, одним из самых сложных занятий для человека является изучение иностранных языков, обучение чтению и игре на скрипке или фортепьяно. Такими навыками взрослые овладевают с трудом, а для детей — это почти неосознанное усилие. И моя жизнь — яркое

11

Figure 1.3: Example page of the DISEC and a skewed fragment with handwritten text.

Figure 1.3 shows an example page of the DISEC and a paper fragment with handwritten text. Due to the size of a fragment the inscribed content is restricted to a couple of words in contrast to the test images of DISEC. Within this thesis a skew estimation for sparsely inscribed documents which exploits also the grayvalue information, is presented. An additional challenge is the irregular shape of the fragments which can result in an irregular length of text

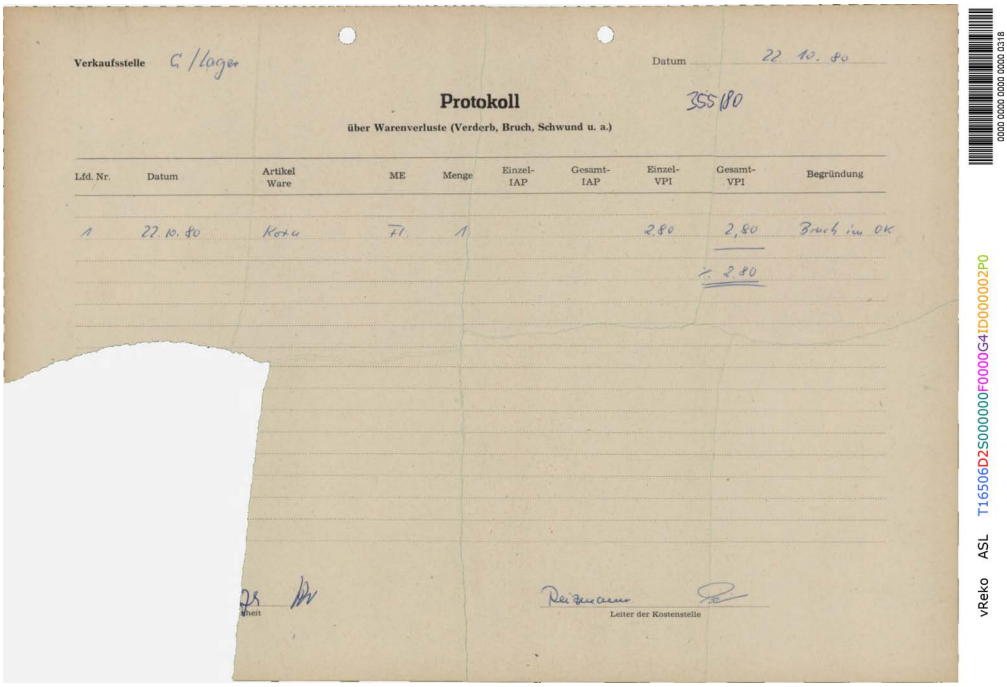


Figure 1.4: Reassembled form document of a real world example of the Fraunhofer IPK reconstruction system.

lines and the varying content (printed, handwritten, graphics, mixed). A main contribution of the proposed method in this thesis, is a skew estimation on grayvalue images, which avoids the binarization step. The accuracy of gradient based orientation measures is evaluated as well as Focused Nearest Neighbour Clustering (FNNC) methods based on interest points. State of the art methods are described in Section 2.2.

Document classification is defined by Chen and Blostein by assigning "*a single-page document image to one of a set of predefined document classes*" [26]. Therefore a *Document Space* (set of all expected input documents) and the *Document Classes* (subset definition of the document space) are defined. Form classification restricts the document classes to different types of form documents. The document space can also comprise documents which do not belong to any defined form class, which adds a reject class to the document (form) classes [9]. Due to the syntactical knowledge (defined structure) of a form type semantic information can be extracted: the classification of form documents allows automated extraction of filled-in data in form processing systems [9, 41]. The retrieval and classification of forms allows grouping and indexing of entire records due to the knowledge of the composition of records. A form class like e.g. *Index* can be at the beginning of a record and infers about the rest of the records content [85]. State of the art methods use either binary images (e.g. NIST tax forms database [127]) or grayvalue images of entire form images. Within this thesis a form classification and retrieval for degraded and reconstructed form documents is discussed and evaluated on images of the Stasi records.

Figure 1.4 shows an image of a reassembled Stasi form document of the Fraunhofer IPK

Berlin reconstruction system. It can be seen that lines are broken due to the fragments boundaries and even entire parts of the form document can be missing. In contrast to current form processing systems the proposed method must be able to deal with degraded form documents as well as with form structure variations within a single form class (template of certain Stasi forms can vary over the time). The form classification will support archivists to group and index single documents to entire records, since specific form documents like the *Table of Contents* can give conclusions about the missing documents of a record. Hand-written filled in data can affect (global) form features and the occurence of unknown form types can cause additional errors in form processing systems [9]. State of the art form classification methods are summarized in Section 2.3.

1.2 Problem Statement and Aim of the Work

DIA have a preprocessing stage necessary for the further analysis like OCR, layout analysis and document clustering. Within this thesis preprocessing is a binarization (foreground/background separation), the task of determining the document skew and a form classification for form analysis and document clustering. Thus, the problem statement can be summarized as follows:

Binarization Foreground/Background separation of low-quality documents

Document Skew Determine the introduced skew of sparsely inscribed documents

Form Classification Determine a form class of low-quality documents

The main aim of the methods developed is the support of reconstruction systems of fragmented historic or ancient documents. Such systems use the layout information (based on the binarization) and the document image skew as features for the matching process. The clustering/grouping of reassembled or partly reassembled pages to entire files is based on the information of certain form types, like a table of content. The binarization developed must be able to deal with the kind of distortions (see also Section 1.1) present in historical documents. Additionally, fragmented objects vary in size and the amount of the inscribed content. The main aim of the skew estimation compared to state of the art methods is the ability to detect the skew of sparsely inscribed documents with mixed contents (handwritten/printed/images). The form classification has to deal with distortions like broken and missing lines.

1.3 Methodological Approach and Innovative Aspects

In this section the main approaches are summarized and the innovative aspects compared to state of the art methods are detailed.

Current state of the art binarization methods are mainly based on edge detection and an estimation of the stroke width (see Section 2). Regarding the document image binarization contests, certain classes of noise are represented. Figure 1.5 shows an example of an image of a carbon copy and the binarized image with the approach of Su [168]. It can be seen, that Gaussian noise cannot be handled due to the edge sensitivity of the method. The estimation of

the stroke width can lead to holes in the foreground, if the present stroke width is different from the estimated size.

Figure 1.5: Binarization example of the method of Su (2011) of an image with Gaussian noise.

Thus, a scale space is implemented within this thesis to avoid an estimation of the stroke width to treat different text sizes. Noise such as background clutter is suppressed due to the continous smoothing from finer to coarse scales, since coarse scales represent homogeneous regions in the image. The thesis presents a foreground estimation based on different scales to apply a weighting scheme which suppresses noise without losing low contrast text. The proposed method is compared with state of the art methods with the metrics presented at DIBCO.

State of the art skew estimation methods are based on binarized document images (see DISEC). The method proposed in this thesis is based on the text's gradients in combination with a FNNC of interest points, thus exploiting the grayscale information of the image. Figure 1.6 shows a part of a document image containing a character. The upper part shows a binarized image and the lower part of Figure 1.6 shows the grayvalue image with the corresponding gradient information. Due to the effects of the binarization (sharp edges) it is shown that exploiting the grayvalue information can reduce the error from $2.35\,^{\circ}$ to $0.19\,^{\circ}$. The research analyzes also the possible accuracy of gradient based orientation measures. The combination of both methods is able to handle also slanted handwritten text and fragments with at least 2 words, thus sparsely inscribed documents. The detectable angle range of the proposed method is not restricted.

Form classification methods are designed to handle form documents of daily life, thus non degraded document images. The research within this thesis addresses the classification of form documents with broken lines and missing line information. Figure 1.4 shows a reconstructed form document. Due to the fragment borders, lines are broken. The method developed is based on shape features of the sampled line information of a binary image of the form document. In the training, the shape features for each form type are clustered to create a dictionary and based on the occurence histogram, form documents are classified by comparison with the occurence histogram of the form templates.

The main innovative aspects in this thesis are a parameter free binarization regarding the stroke width and foreground estimation to reduce noise without the loss of low-contrast text. The form classification is robust against broken lines and uses shape features instead of defined (and restricted) line crossings. The skew estimation presented can be used for mixed documents

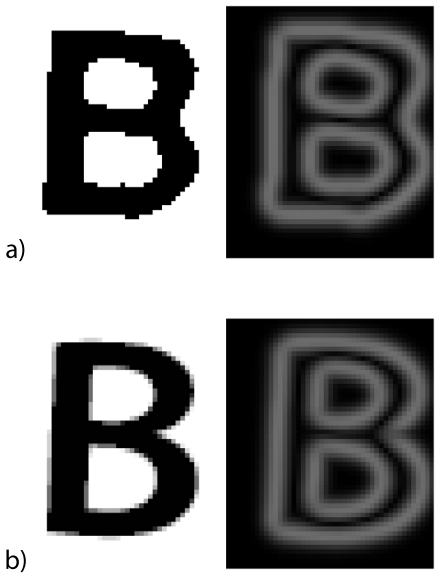


Figure 1.6: Example image of a character and the gradient information of the a) binarized image and the b) grayvalue image

(handwritten and printed text) and exploits the grayvalue information of the document image in contrast to state of the art methods.

1.4 Structure of the Work

Section 2 reviews state of the art methods for binarization, skew estimation and form classification and retrieval. A comparison of related work is drawn at the end of this section based on current contests. Additionally, state of the art databases for evaluation are presented. Section 3 describes the investigated methods on these 3 preprocessing topics. The binarization approach is detailed in Section 3.1 while the form classification is presented in Section 3.3. A skew estimation method which exploits also the grayvalue information with no restriction on the detectable angle range is shown in Section 3.2.

Evaluation measures for binarization, skew estimation and form classification are analyzed and presented in Section 2.4. Additionally the evaluation of the three presented preprocessing methods for low quality and sparsely inscribed documents and the used datasets are described subsequently to each methodology. A conclusion is drawn in Section 4.

CHAPTER 2

State of the Art

In this chapter an overview on DIA preprocessing methods is given. The main focus is based on document image binarization, skew estimation and form classification, the main preprocessing steps for layout analysis and OCR methods [53, 55, 135]. Additionally form classification and retrieval is a main step of document clustering systems [152]. Section 2.1 deals with state of the art methods of global and adaptive binarization methods and summarizes recent efforts of binarization techniques based on the DIBCO. A survey of skew estimation methods is presented in Section 2.2, while Section 2.3 summarizes approaches on form identification. Also the winning methods of the first document image skew estimation contest [135] are presented in Section 2.2. A summary and comparison based on the results of current contests, DIBCO, H-DIBCO [53, 141, 142, 144, 145] and DISEC [135], are presented in Section 2.5.

2.1 Binarization

The objective of image segmentation is to group image pixels according to constituent regions or objects [58]. On document images this problem consists of two classes: foreground and background. According to [69] the "*binarized image should be perceptually similar*". For the binarization of documents global and adaptive binarization methods exist. While the same single threshold is applied on every pixel by global algorithms, adaptive methods define local regions in which separate threshold values are calculated. Current binarization methods use grayvalue images as input (see [141, 142, 144, 145]). Color images can be converted with the standard conversion $I(x, y) = 0.3R(x, y) + 0.59G(x, y) + 0.11B(x, y)$, where R, G and B are the Red, Green and Blue channel of the color image [62]. For an $m \times n$ grayvalue image $I(x, y)$ with intensity values between 0 and 1 and a threshold $T(x, y)$ each image pixel is classified in foreground (labeled as 1) and backgound (labeled as 0) resulting in the thresholded image $I_{th}(x, y)$:

$$I_{th}(x, y) = \begin{cases} 1 & \text{if } I(x, y) > T(x, y) \\ 0 & \text{if } I(x, y) \leq T(x, y) \end{cases} \tag{2.1}$$

where $T(x, y) = T_g = constant$ if a global threshold is applied. Adaptive methods have the characteristic that the value of T depends on the local gray value characteristics. Global thresholds are suitable for images with a bimodal gray value distribution, where adaptive methods can handle documents with e.g. non-uniform illumination [55]. Recent developments (see DIBCO and H-DIBCO) show that binarization methods estimate the background or combine multiple binarization methods to achieve a better segmentation. A comparison of binarization methods for historical archive documents is presented by He et al. [62]. The methods presented comprise Niblack, Sauvola (see Section 2.1.2) and a color segmentation method [63]. In the following, state of the art methods of image binarization are categorized in global and adaptive methods, methods based on background estimation and methods that use a combination of binarization methods.

2.1.1 Global Methods

Documents which are digitized with a defined setup (e.g. scanner which uses a constant illumination) and a defined minimum contrast between background and foreground (no faded out text) can use a pre-defined constant threshold value T [101].

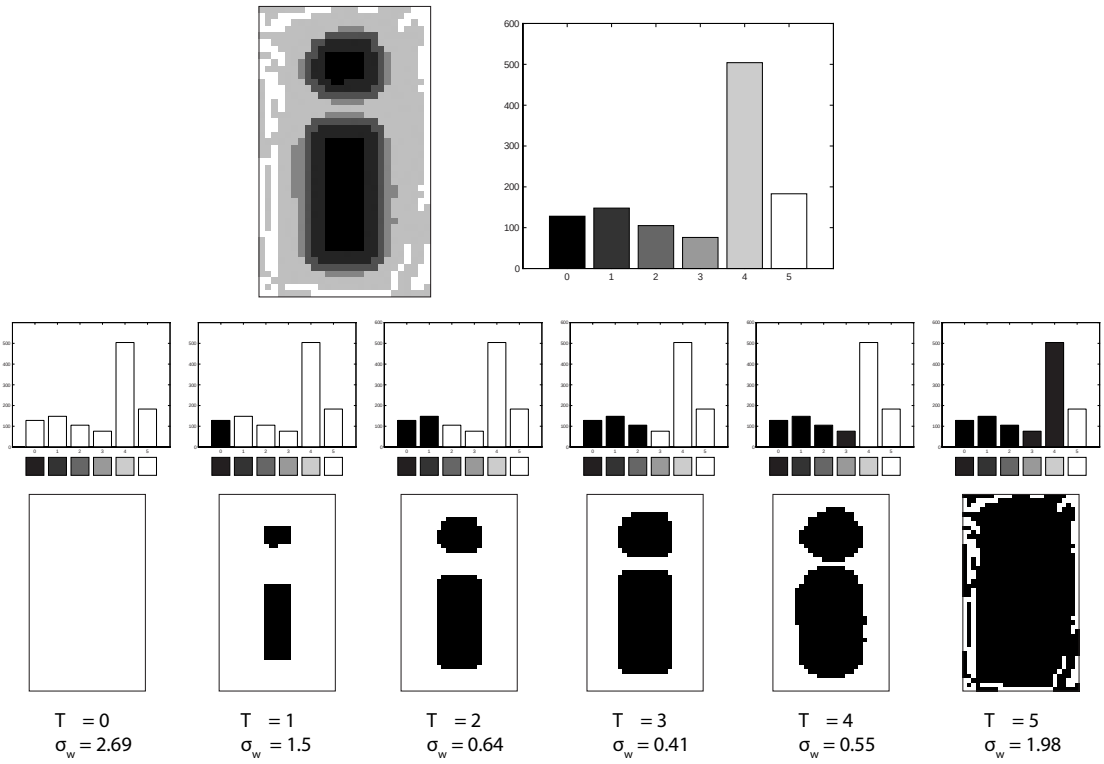


Figure 2.1: Grayvalue image of the character "i" and the corresponding grayvalue histogram. The results of all possible thresholds and the associated intra-class variance are shown to illustrate the result of Otsu's method [134].

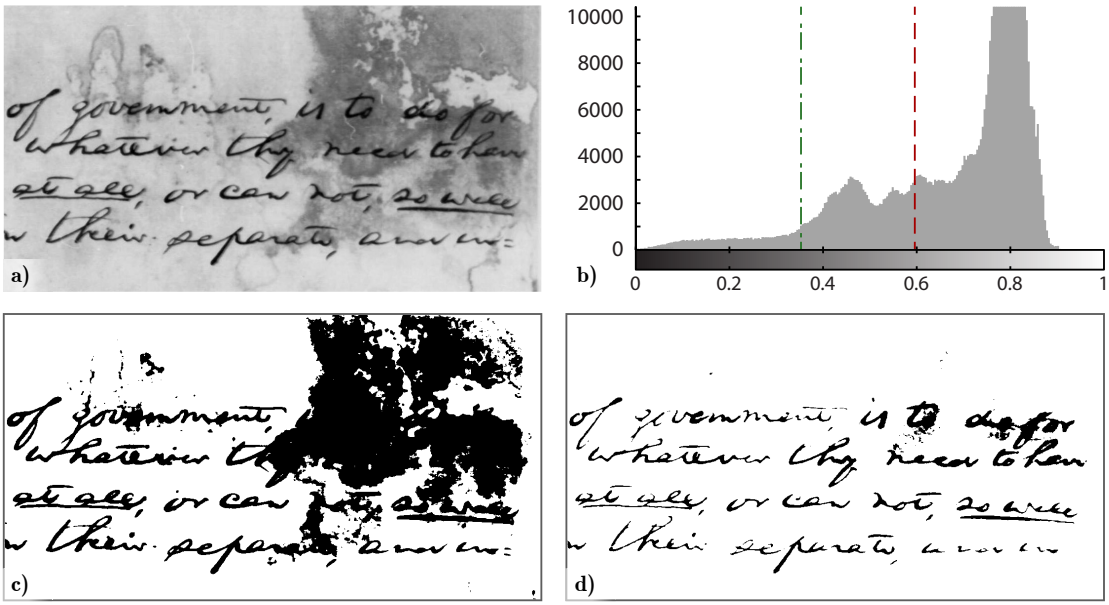


Figure 2.2: (a) Image of the DIBCO 2009 dataset (b) histogram with Otsu threshold (dashed) and manual threshold (dashed-dotted) (c) Otsu threshold image (d) manually thresholded image

A global threshold, which analysis the distribution of the gray values, is introduced by Otsu [134]. Otsu's thresholding method [134] assumes a bimodal histogram and minimize the intra-class variance. Global methods can be used for e.g. scanned documents (constant illumination) with a uniform background. Historical and degraded documents need adaptive algorithms due to the low contrast of faded out text and the presence of background clutter.

Figure 2.1 shows a gray value image of the character "i" and the corresponding histogram. The image consists of 6 different grayvalues. To illustrate the methodology of Otsu, the image is thresholded at every possible grayvalue T and the binarized results with the associated intra-class variance σ_w [134] values are shown. It can be seen that at threshold $T = 3$ the classification into foreground and background leads to the smallest intra-class variance $\sigma_w = 0.41$, which will be the final result of Otsu's method.

Figure 2.2 a) shows an image of the DIBCO 2009 dataset with background variations and the corresponding histogram. It can be seen in Figure 2.2 c) that the result of Otsu's method classifies parts of the background as foreground values due to the grayvalue distribution. But it is shown in Figure 2.2 d) that a manual global threshold value leads to a better binarization result. Thus, Otsu determines the optimal global threshold value only for images with a bimodal distribution by definition.

Additionally selected global thresholding methods are proposed by Kittler and Illingworth [79] (based on the probability of the classification error), Fan et al. [47] (based on 2D temporal entropic thresholding) and Xia et al. [181] (entropic thresholding based on the gray-level spatial correlation histogram).

2.1.2 Adaptive Methods

Adaptive methods define local regions $R_{x,y}$ and calculate a separate threshold value $T(x, y)$ for each region. Current techniques [55, 123, 168] make a rectangular subdivision of the grayvalue image depending on the character size.

Niblack [123] defines a threshold T based on the mean m and variance s within a local rectangular window by:

$$T = m + k \cdot s$$

where k is a negative parameter defining the amount of the print object boundary taken as a part of the given object [94] (constant over the entire image). According to Gatos [55] and Wolf et. al [179] the window size has to cover at least $1 - 2$ characters and in [17, 173] k is set to -0.2 and the window size is 15×15 pixels. Milewski and Govindaraju [120] state that document images with noise result in "noise, jagged edges and broken character segments".

An adaption of Niblack's algorithm is published by Sauvola and Pietikainen [155] where the threshold T within a local rectangular window is defined by:

$$T = m \left[1 + k \left(\frac{s}{R} - 1\right)\right]$$

where m is the mean, R is the dynamic range of the standard deviation and s is the variance of the grayvalues in the local window. Sauvola sets the parameter k to 0.5 [153]. In comparison to Niblacks method Sauvola can handle background noise.

Wolf et al. [179] participated in the DIBCO 2009 competition and achieved rank 5. The proposed binarization algorithm is an adaption of Sauvola where the contrast and the mean gray level of the image is normalized. The main application of the method are multimedia documents and video frames. The threshold value using the normalized mean gray level is calculated by:

$$T = m - k\alpha\,(m - M)\,,\ \alpha = 1 - \frac{s}{R},\ R = max(s)$$

where m is the mean gray value, s is the standard deviation, M is the minimum graylevel of the image and R is the maximum of the standard deviation of all local windows. k is set to 0.5.

Multiscale approaches are published by Tabbone and Wendling [171] or e.g. Dorini and Leite [40]. Tabbone and Wendling apply a Laplacian for a segmentation of the image and a nonlinear filter to remove noise. A multiscale statistical test of homogenity regions identify stable regions which gray value distributions are used as a model to compute a threshold value. Dorini and Leite use a self-dual multiscale morphological toggle operator which "*replaces the original value of each pixel with the most similar between its scaled dilation and erosion*" [40]. The method is tested on ill-illuminated images.

Fabrizio et al. [46] use a morphological toggle operator [161]. Similar to Dorini and Leite [40] the pixel is marked as background if the eroded value is closer to the actual pixel value and otherwise as foreground (actual pixel value is closer to the dilated pixel value). To avoid salt and pepper noise pixels can also be classified into a third class which represents homogeneous regions. Homogeneous regions are classified into foreground and background based on their boundaries. The proposed method achieved rank 2 at DIBCO 2009.

Su et al. [145] (winner DIBCO 2013) use a combination with an exponential function of the local image contrast and the local image gradient. A second combination of the local image contrast is done with the edge map. As a last step the image is binarized based on the local edge map and the estimated stroke width. This work is based on previous methods which have been submitted to H-DIBCO 2010 [141] (rank 1) and DIBCO 2011 [142] (rank 2).

Howe [69] defines an energy function (fitness function) which is minimized by the ideal binarization. The energy function accumulates the costs for assigning a pixel to the foreground, background at a term that defines the cost, if a pixel has a different label compared to the neighbours. The formulation of the energy function "*corresponds to a Markov random field and allows a more specific expression for the energy*" [69]. To define the label costs a Laplacian of the image intensities is used (illumination invariant). To allow discontinuities in the final result a Canny edge detector is used to define these regions. To solve the energy function a graph cut implementation is used. The proposed method achieved rank 3 at DIBCO 2011 [142].

Howe [70] uses the method described in [69] as a base method and studies the automatic selection of the parameters of the base method. It is stated that the result of a binarization method can be improved by choosing the paramter values according a given image class (certain class of distortions). A stability heuristic criterion is introduced which allows to set the 2 "important" parameters of the method described in [69]. As a result it is stated that the "*tuning algorithms given come close to maximize the potential of the base binarization algorithm*" [70]. The method has been submitted to DIBCO 2013 [145] and achieved rank 2.

2.1.3 Methods based on Background Estimation

A different class of adaptive algorithms especially used for ancient manuscripts are methods which estimate the background. A background estimation allows to compensate a "*variable background intensity caused by non-uniform-intensity, shadows, smear, smudge and low contrast*" [55]. Gatos et al. [55, 56] use a Sauvola threshold for a rough foreground estimation. Based on the result of Sauvola they calculate a background surface estimation where foreground pixels are interpolated by a mean value of the surrounding background pixel. For the final thresholding the background image is subtracted from the original image to examine the pixel contrast, and an adaptive threshold function based on a sigmoid function is defined. To enhance the result, a preprocessing is done by applying an adaptive low-pass Wiener filter. As a post-processing a shrink and swell filter is applied to remove noise and correct gaps, breaks or holes [55].

Bolan Su et al. [53, 141, 145] are the winner of DIBCO 2009 (S. Lu, B. Su and C.L. Tan) H-DIBCO 2010 and DIBCO 2013 competition. Lu et al. [111] estimate the document background using a one dimensional polynomial smoothing (Savitzky-Golay smoothing). Afterwards a global polynomial smoothing is applied to avoid the estimation of text regions (foreground) as in Su et al. [168] and Gatos et al. [55]. The background image is compensated using the background image and thresholded as described in Su et al. [168] using the text stroke edge image based on the contrast image. The proposed method is the winner of DIBCO 2009.

Su et al. [168] use a normalized gradient image - called contrast image - which is based on the local maximum and minimum of a 3×3 window. They state that the normalization "compensates for the effect of the image contrast/brightness variation" [168]. To estimate the foreground/background similar to Gatos et al. [55], a *simple* threshold method (Otsu) is applied

to the contrast image. The final threshold is defined by $m+s/2$ where m is the mean value of the estimated foreground and s is the standard deviation within a local window. The window size is based on the mean strokewidth which is determined using a horizontal projection and counting the distances between two stroke edges. The proposed method outperforms the one published in [111].

2.1.4 Methods based on the Combination of Different Binarizations

Recent developments (see DIBCO contests [143,145], [121]) show that a combination of different binarization methods leads to better results. This can be done by applying different thresholding methods to the same image and afterwards select the best result [118], or a feature vector is created and classified. Gatos et al. [57] and [169] use the binarization results of different algorithms and calculate additional features from the original grayvalue image based on e.g. the edge information or a contrast map. Below, current methods, which are based on the combination of different binarization techniques, are described.

Gatos et al. [57] binarize an input image with an odd number of thresholding methods. For the final binary image a pixel is marked as foreground pixel if the majority of binarization methods have classified the pixel as foreground. To improve the binarization result a foreground filling is applied to regions defined by the edge information (see [57]). Edges are detected using a Canny edge detector. As a preprocessing step a Wiener filter is applied and a shrink and swell filter as post-processing step similar to [55].

Su et al. [169] use n different binarization methods, where two result images are sucessively combined. Image features, namely a contrast map and the intensity value are extracted from the grayvalue image and used to classify uncertain pixel values. Depending on the local mean contrast and intensity values of the background and foreground within a local neighbourhood, uncertain pixel values are classified [169].

Messaoud et al. [118] state that "*digital images belonging to different books of the same database are generally different*". Thus, different binarization methods with different parameters must be used for each book. As a result Messaoud et al. [118] propose a system with an automatic selection of a binarization method and its parameters. Therefore, within a training phase a subset of each book is selected and each document is classified into one of 4 noise classes: bleed-through, high similarity between background and foreground, variable background and all other images. Based on the noise class the input parameters are selected for the binarization methods and the resulting binarized images are ranked for each method using a metric which combines the measures used in DIBCO contests. After the selection a validation phase is performed on a different subset of documents. Messaoud et al. [118] conclude that 25% of the images are sufficient for the training phase and that the choice of the method leads to the best binarization method in 83% to 90% of the tested collections.

Moghaddam et al. [121] propose an unsupervised ensemble of experts framework which combines the outputs of different experts - in this case binarization methods (called experts). On the one hand the method can use the output of different methods or, on the other, the output of one method with different parameters. The first case is tested with the result images of all participants of the H-DIBCO 2012 [144] which leads to a result with 3% improvement. The latter is tested with the grid-based Sauvola method [121] (3% improvement) and the algorithm proposed by

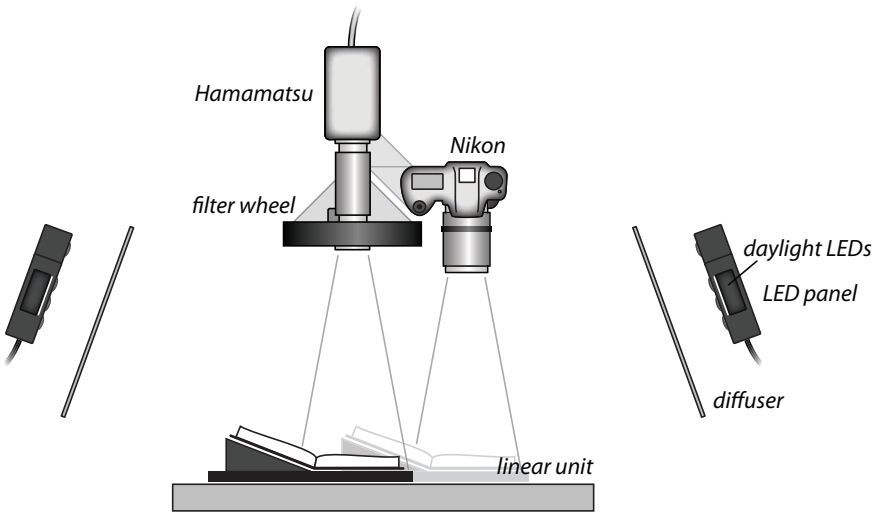


Figure 2.3: Schematic setup of the CVL MS acquisition system with UV illumination.

Howe [70] (1% improvement). The selection process is based on a confidence map which is used to create an endorsement graph. The endorsement graph shows the coherence between the experts. Coherent experts are called school of experts and it is assumed that experts of a school have a high endorsement on each other. Based on the schools of experts and the endorsement values higher than a threshold, the selection is done. The method has been submitted to DIBCO 2013 [145] and achieved rank 3.

2.1.5 Binarization using Multi-spectral Images

An alternative to the methods mentioned is to use multispectral imaging and to exploit information in the non-visible wavelengths of the reflected and emitted light of historical documents. Based on the assumption that the optoelectronic transfer function of the imaging system is linear [157], the cameras response r of an image pixel is given by the following equation [157]:

$$r = \int_{\lambda_{min}}^{\lambda_{max}} I(\lambda)R(\lambda)O(\lambda)S(\lambda)d\lambda \tag{2.2}$$

where λ is the wavelength, $I(\lambda)$ is the illumination energy that reaches the observed object, $R(\lambda)$ is the color reflectance of the object, $O(\lambda)$ describes the properties of the optical system and $S(\lambda)$ is the responsiveness of the cameras sensor. Depending on the filters and illumination used, different spectral representations of cultural heritage objects (manuscripts) can be obtained. Figure 2.3 illustrates a possible MultiSpectrum (MS) acquisition setup [86].

The imaging techniques used for the aquisition of ancient manuscripts are known as UV fluorescence/reflectography and IR reflectography [60, 86, 138]. Based on the properties of the writing material (e.g. iron-gall ink) and the writing carrier (e.g. parchment) the irradiated UV light is either reflected (UV reflectography) or absorbed resulting in a “light source” emitting

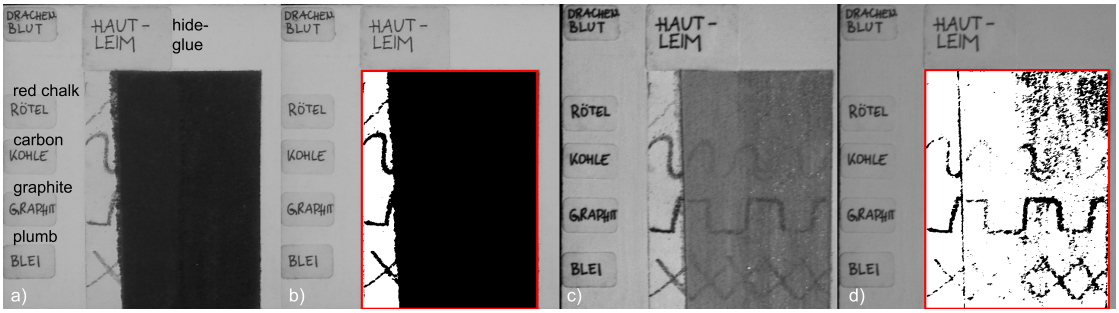


Figure 2.4: Test panel with different writing materials which are covered by hide glue. a) RGB image b) Global threshold of the test pattern c) IR image d) Global threshold of the IR image.

radiation in the VIS part of the electromagnetic spectrum (UV fluorescence) [86, 138]. Iron-gall inks used in ancient manuscripts have the property that they do not fluorescence in contrast to the parchment used as writing carrier [86]. Thus, UV fluorescence can be used to enhance the contrast between the writing and the carrier material [42, 60, 138, 151] by exploiting optical properties of different materials. In Pentzien et al. [138] it is stated that IR radiation "*is less scattered than visible light*". By observing the reflected IR radiation (IR reflectography) it is possible to differentiate between different text layers (e.g. palimpsest text vs. newer text) [138].

Figure 2.4 shows a test panel where patterns are drawn with different writing materials. The patterns are covered with a painting layer (hide glue) such that the patterns are not visible to the human eye. It can be seen in Figure 2.4 b) that a global binarization (Otsu) applied to the painting area can only segment the patterns in the area which is not covered by the additional painting layer. If the panel is captured in the IR range of the light, the contrast of the drawn patterns is visible and can be segmented using a global binarization (see Figure 2.4 d) without any further contrast enhancement. Thus additional information is revealed in the multispectral images, which can be exploited for the binarization of document images. Lettner [96, 98] shows the possibility to use the information within different wavelengths to enhance the result of the binarization. Hollaus [68] uses MultiSpectral Imaging (MSI) to enhance the contrast of images by applying statistical analysis like Principle Component Analysis (PCA) and Linear Discrimant Analysis (LDA) [67, 68].

2.2 Skew Detection

Document skew detection estimates the main global orientation of a document and is a preprocessing step for DIA systems [3, 12, 112]. Chen [27] defines skew as a documents "*dominant (most frequently occuring) text baseline direction*" [27]. Figure 2.5 a) shows an example document of a skewed page, where the global skew angle is introduced due to a scanning process. A photograph of an ancient manuscript where the text lines have different skew angles is shown in Figure 2.5 b). Different skew angles can occur on handwritten documents due to variations of the writer or due to environmental effects which effects the carrier material (e.g. the parchment/paper).

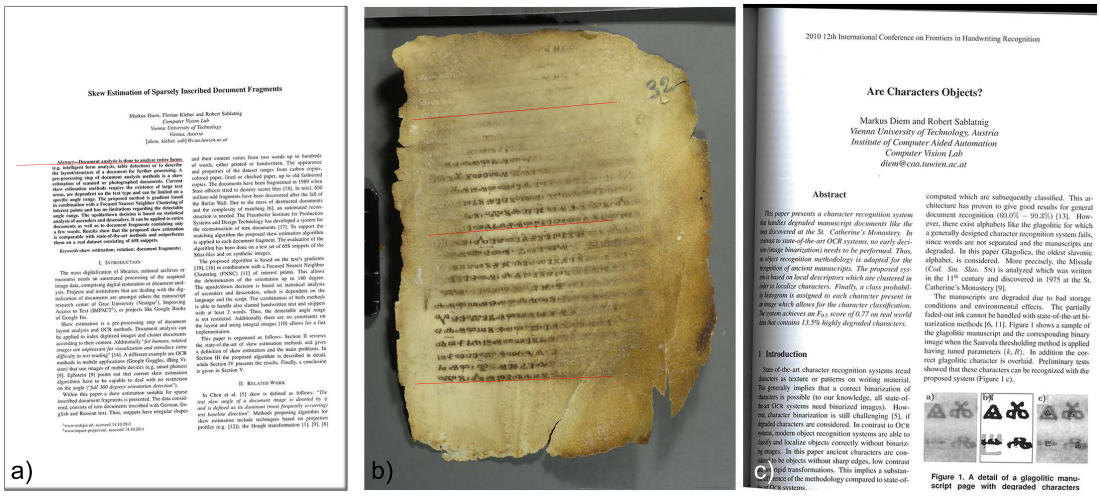


Figure 2.5: Example of a) a skewed page with a global skew angle, b) an ancient manuscript with different skewed text lines and c) skewed text lines of a scanned page due to the book binding. Red Lines indicate the skew angle.

The manuscript page in Figure 2.5 b) had been exposed to water, which lead to a distortion of the parchment. A global skew orientation will determine, as stated in Chen [27], the dominant text baseline direction. Different skewed text lines appear also if e.g. a page of a bound book is scanned (see Figure 2.5 c). To correct local distortions, dewarping methods [20, 45, 52, 186] can be applied. Okun et al. [131] has defined 3 types of skews: "*a global skew, a multiple skew when certain blocks have a different slant than others, and a non-uniform text line skew, when the orientation fluctuates within a line*" [131].

Thus a skew can be introduced by scanning devices like flatbed scanners (manually or due to an automatic feeding device) [10] or document images taken with a camera [44]. Additionally printer with a wrongly aligned paper in the feeder can produce skewed texts. Handwritten texts can be skewed if no ruling scheme is applied or due to the variation of the writer. A distortion of the carrier material (paper, parchment) can introduce a skew as illustrated in Figure 2.5 b). A document skew correction is done as a preprocessing step for DIA systems, since uncorrected documents affect the performance of OCR, line extraction, and page segmentation [2,12,44,78]. Also due to perception reasons a document image is skew corrected, since "*skewed images are difficult for visualisation and reading*" [10]. Problems of skew estimation methods are summarized in Kavallieratou et al. [78]:

- *Restriction to detectable angle range*
- *Restrictions on type or size of fonts*
- *Dependence on page layout.* [...]
- *A specific document resolution is required.*

- *High computational cost*
- *Limitation to specific applications*
- *Large text areas are required*
- *Most of the proposed algorithms are appropriate for machine-printed pages and fail when they deal with handwritten documents.* [...]

However, Epshtein [44] states that current skew detection algorithm have to deal with all of the estimated problems, particularly the restriction of the detectable angle range, since applications for mobile devices (e.g. Google Goggles, Microsoft iBing Vision, etc.) have to deal with photographs of documents or text with unconstraint conditions. A summary of skew estimation methods is presented in Cattoni et al. [23] and Hull [71]. Amin and Wu [3] and Aradhya et al. [7] categorize skew detection methods on:

- *Docstrum (*K-NN *clustering)*
- *Projection Profile (Hough transform)*
- *Fourier transform*
- *Cross Correlation*
- *Other methods*

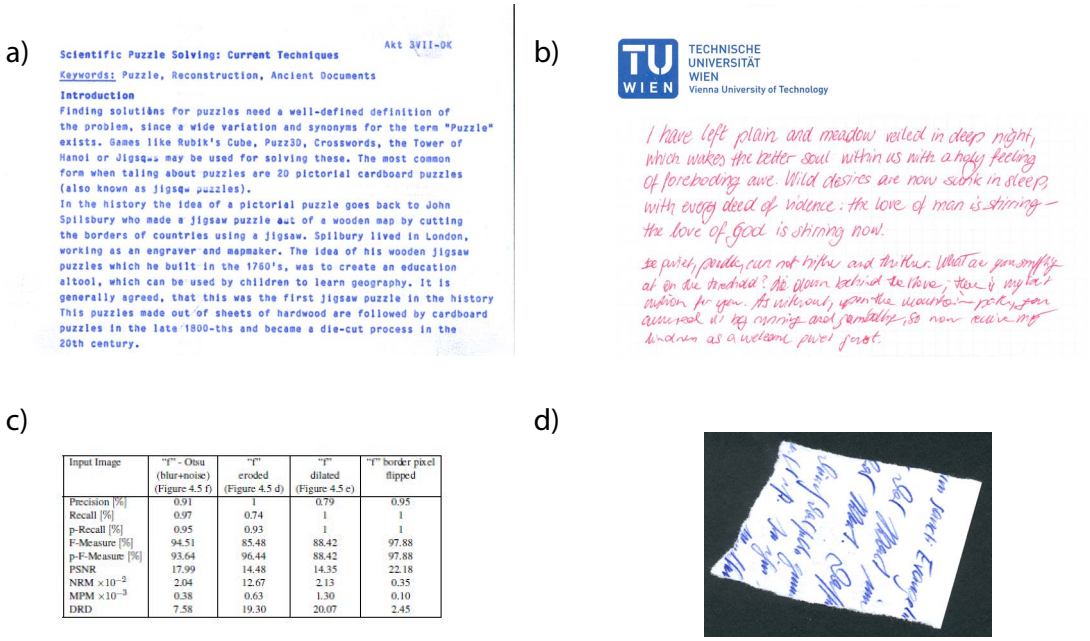
State of the art skew detection methods based on the methods defined by [3, 7] are summarized in the subsequent sections. An additional category are morphological approaches (e.g. [113]) which are also presented in Section 2.2.4. Within the framework of ICDAR 2013, the first international DISEC has been organized. This shows that skew detection is an active research field within the document analysis community. The contest comprises binary images of figures, tables and writings in several languages and scripts. It is stated by the organizers that it is still common for big archives to scan the documents in black and white.

Figure 2.6 shows examples of different document types ranging from printed or handwritten text to documents with sparse text and tabular structure.

2.2.1 Projection Profile (Hough transform) based Skew Estimation

The general idea of projection profiles to determine the skew of a document is described by maximizing the number of co-linear black pixel [71]. Figure 2.7 shows the horizontal projection profile of a document image, and the same document image skewed by $5\,^{\circ}$.

It can be seen that the unskewed document produces higher peaks with smaller deviations according to the text lines. Within the skewed document two or more text lines can overlap within a single projection line, leading to reduce the gaps between the peaks. The method is sensitive to the content of document images (e.g. pictures, etc.) and to the length and the alignment of the text lines. Figure 2.8 shows a detail of a two column document page presented



Input Image	"f" - Otsu (blur+noise) (Figure 4.5 f)	"f" eroded (Figure 4.5 d)	"f" dilated (Figure 4.5 e)	"f" border pixel flipped
Precision [%]	0.91	1	0.79	0.95
Recall [%]	0.97	0.74	1	1
p-Recall [%]	0.95	0.93	1	1
F-Measure [%]	94.51	85.48	88.42	97.88
p-F-Measure [%]	93.64	96.44	88.42	97.88
PSNR	17.99	14.48	14.35	22.18
NRM $\times 10^{-2}$	2.04	12.67	2.13	0.35
MPM $\times 10^{-3}$	0.38	0.63	1.30	0.10
DRD	7.58	19.30	20.07	2.45

Figure 2.6: Different document types: a) printed document b) handwritten document with a printed logo c) tabular document d) handwritten document containing only a few words.

in Figure 2.5 a). It can be seen that the text lines of the two columns are miss-aligned, which has the same effect on the projection profile histogram as a skewed page.

Kanai and Bagdanov [77] have developed a projection profile based skew estimation for JBIG images. JBIG is a lossless image compression standardized as ISO/IEC 11544 which is developed for binary images and is used in fax machines. The compression scheme allows to determine feature points, so called (white) pass codes, which are used for building the projection profile histogram. Lee [93] states "*that pass codes occur at locations corresponding to bottom of strokes (white pass) or bottom of holes (black pass)*". Thus, the feature points describe the lower baseline of texts using Roman alphabets. To determine the skew the optimization function proposed by Postl [140] is used. The number of text lines influence the accuracy of the proposed method, as well as different contents of a document page, e.g. images, have a significant influence on the skew.

Lu et al. [112] use a horizontal and vertical white run histogram for the determination of the skew. Based on the skew of the document image one of the two histograms (horizontal or vertical) consists of 2 peaks which are related to the interline spacings and the white runs within single characters. For estimating the skew, further analysis is taken on the white run histogram with 2 bins. The peak corresponding to the interline spacing is detected using a threshold determined by statistical analysis (interclass variance is maximized). The starting and ending points of the detected white runs (defining interline spacings) define the lower and the upper baseline of text lines, thus the orientation of the page. The least square algorithm is used to estimate the baselines, and the median of all baseline orientations defines the skew. The

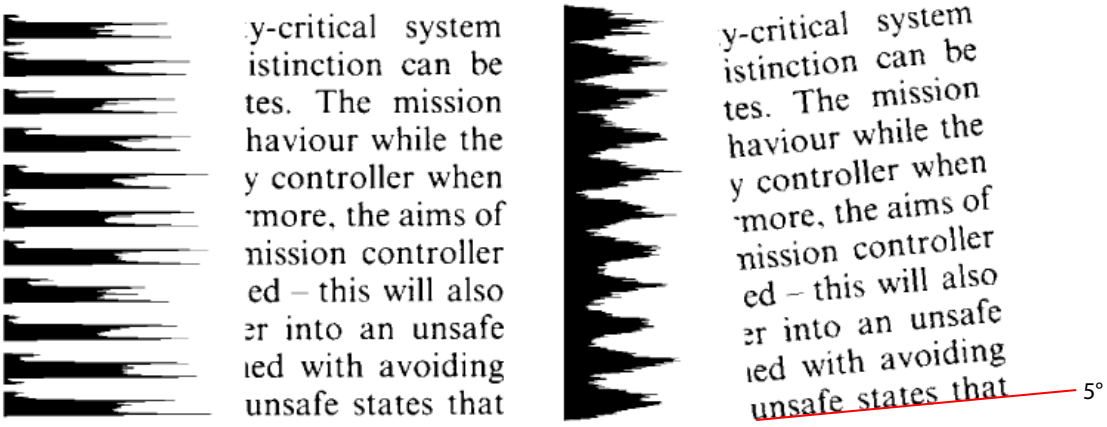


Figure 2.7: Projection profile of a correct aligned document image and skewed by $5\,^{\circ}$, courtesy by Hull [71].

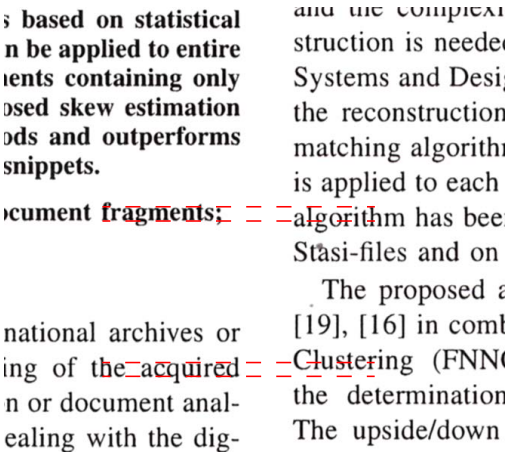


Figure 2.8: Detail of 2 column page of Figure 2.5 a), where the dashed lines show the text line missalignment between two columns.

upside-down orientation is based on statistical analysis of the ascenders and descenders.

Kavallieratou et al. [78] calculate the Wigner-Ville distribution of the horizontal projection profile of a document image. The Wigner-Ville distribution of the horizontal projection profile with the highest intensity designates the skew angle. To minmize the computational effort, the skew is estimated in several steps: first a rough estimation in the interval of $-84\,^{\circ}$ to $+84\,^{\circ}$ is done by calculating the projection profile respectively the Wigner-Ville distribution for every $12\,^{\circ}$. After the first estimation the search interval can be reduced to $-6\,^{\circ}$ to $+6\,^{\circ}$ with an interval step of $1\,^{\circ}$. The final skew angle is determined within the interval $\pm 0.5\,^{\circ}$ of the second skew angle estimation with an interval step of $0.1\,^{\circ}$. The "*Wigner-Ville distribution of the histograms represents their time-frequency distribution*" [78] where time is related to the page height. The

Figure 2.9: Synthetic test image with Gaussian blur and noise added.

applicability of the Wigner-Ville distribution is discussed in [78]. To overcome the problems of multi-column pages, e.g. presented in Figure 2.8, and to reduce the computational cost, only a part of a page is selected.

A Hough based skew estimation is presented by Amin and Fischer [2]. The document image is binarized using an adapted version of Otsu's method and all Connected Component (CC) are represented by rectangular boxes. The CC are analyzed according their size and a grouping (clustering) is performed which defines regions (e.g. paragraphs, captions). Grouped regions are divided into vertical segments and only the center points of the rectangles of the last row of each region are considered as feature points. By applying the Hough transform to the feature points the CCs describing the bottom text line are determined. The skew angle of the bottom row is calculated by applying the least square method to the feature points of the CC. The final skew angle is the averaged skew angle of all groups.

Manjunath Aradhya et al. [7] apply the Hough transform to determine the skew angle of printed text. Therefore the average height and width of single characters is estimated based on bounding boxes of CC. As a preprocessing step all characters with ascenders, descenders and uppercase characters are filtered, since they are larger then the average height. The remaining characters are represented by their bounding box, wheras the upper and lower coordinates of the bounding boxes match the upper and lower baseline of the text lines. The bounding boxes are considered as filled boxes, to which a thinning algorithm is applied. The Hough transform is applied to the thinned boxes to determine the skew angle.

A skew estimation based on the Muff transform (modified Hough transformation), which is a bounded parameterization for straight lines introduced by Wallace [176], is presented in Yuan and Tan [185]. A Laplacian of Gaussian is applied to the grayscale document image to detect straight edges. In comparison to the previously presented methods straight lines from non-textual objects (e.g. line drawings, separating lines of multi-column texts, dark borders from photocopies) are considered. Detected lines are filtered according perpendicular or parallel lines. The median slope of the perpendicular, parallel or all lines determines the final skew.

Jiang et al. [75] use a coarse skew angle estimation to reduce the search space from $\pm 90\,^{\circ}$ to $\pm 3\,^{\circ}$. Feature points, called detection points, are calculated by dividing the document image page in stripes, whereas the leftmost pixel are defined as detection points. The angle of the vector of detection points between neighbouring stripes determines the coarse skew angle. The Hough transform is applied to the selected detection points for the final skew angle.

Epshtein [44] presents a method with almost no restrictions which can be used for document images taken by mobile devices. Figure 2.9 shows a synthetic test image which contains just 2 printed words, where Gaussian blur and noise are added. The proposed method is able to handle such sparse inscribed document images.

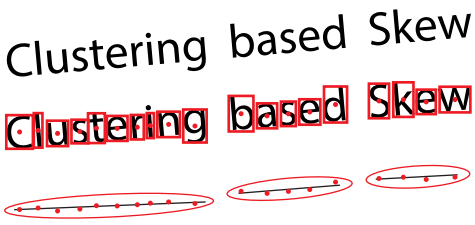


Figure 2.10: Illustrative Example of Nearest Neighbour Clustering Skew Estimation.

The image is binarized and all CCs are filtered according geometric properties (e.g. size, aspect ratio). A mask is created from the filtered image by applying a dilation. Based on the pixel values of the mask and the filtered binary image the Hough accumulator is established. If a pixel is defined as background in the filtered image, but is set in the mask the corresponding cell of the Hough accumulator is incremented. For pixels defined as foreground the corresponding cell is decremented by a defined value. The maximum of the projected Hough accumulator onto the θ axes determines the final skew angle. It is stated that "*the lines coinciding to the white space between lines will get most of the votes in the Hough accumulator*". It is shown that the proposed method has less catastrophic errors (see Skew Evaluation Measures 2.4.2) than the traditional Hough based skew estimation and projection profile based methods (see Section 3.2.6).

2.2.2 Clustering based Skew Estimation

Clustering based methods segment single objects in a document image (e.g. characters) and cluster them to document specific structures like words or text lines. One of the characteristics of the clustered structure is the skew based on the direction vector of clustered elements [113]. Figure 2.10 shows exemplarily the idea of nearest neighbour clustering approaches. Feature points of CCs are clustered and local slope lines are calculated (e.g. least square method). The orientation of all local skew lines can be added to an orientation histogram and the maximum determines the skew of a document. In the following clustering based approaches are summarized.

Okun et al. [131] applies an image compression to a binarized document image using the OR-rule. It is stated that the image structure is preserved and the result image is similar to a run length smoothed/smeared image. Thus single characters are grouped to words or text lines which can be seen as the first clustering part. A fuzzy logic classifier is used to classify each CC to one of the classes *text, character, graphic or line* based on the shape. Based on the classification the skew angle of each element *text* or *line* is estimated by means of the first eigenvector of the covariance matrix. For the final skew detection four different methods are presented: The first method uses an orientation histogram of all determined angles in the previous step. The maximum in the orientation histogram determines the skew angle. The second approach clusters co-linear CCs and describes the skew of the clustered elements again by the first Eigenvector of all CCs within the cluster. The skew angle is weighted by the number of CCs in each cluster. The final skew is determined by an orientation histogram as presented in the first approach.

Clustered CCs lead to a more stable estimation of the skew [131]. The third method uses skews of the first and second method and the skew taken is the one with the higher maximum in the corresponding orientation histograms. The last method combines the orientation histograms of method 1 and 2 in a single histogram. The global maximum determines the skew. It is shown by Okun et al. [131] that the combination of the 2 histograms (method 4) leads to the best results.

Jiang et al. [76] propose a method based on FNNC. The centroids of all CCs are considered as feature points. For each feature point p_i the k nearest-neighbours are searched and for all pairs of the neighbouring feature points a local skew line is calculated. The angle of the local skew line with the minimal perpendicular distance to the current feature point p_i is accumulated into an orientation histogram. The peak in the orientation histogram indicates the skew angle of the document.

A K-NN clustering approach is presented by Lu and Tan [113, 114]. The proposed method uses a binarized image and defines bounding rectangles for all CCs. Based on the ℓ_1 norm (Manhattan distance) of the centroids of the bounding rectangles of 2 CCs, the gap distance (see [113]) and the dimension of the bounding boxes, a nearest neighbour clustering ($k = 2$) is performed. To achieve strings of length k the pairwise clustering is extended to clusters containing k elements. Each neighbouring CCs within the string fullfill the condition of the nearest neighbour clustering. A slope line is defined for all clusters. The mean or median slope of all slopes determines the final skew.

Liolios et al. [106] skew estimation is based on a line clustering of CCs. In comparison to Lu and Tan [113] the Euclidean distance between bounding rectangles of the centroids of CCs is used. The line based approach clusters all elements which Euclidean distance is smaller than 5 times the average width of the CCs. Lu and Tan [113] restricted the number of elements to a specific k. As a preprocessing step all CCs are filtered according their size to remove elements such as punctuations and images. To determine the final skew "*a least square fit is performed through the mass centers of the components*" [106] of each line cluster and the weighted average slope is calculated from all line clusters.

A skew determination using the slope of the upper and lower baseline of text lines is proposed by Avila and Lins [10]. The clustering of CCs of a binarized image is performed similar to Liolios et al. [106] to obtain groups of CCs representing text lines. Using least square line fitting of the middle top points/bottom points of the bounding boxes of all CCs of a line, the upper/lower baseline of textlines is estimated. The maximum of the orientation histogram of all text lines determine the final skew. The upside/down decision is based on the statistical analysis of the ascenders and descenders.

2.2.3 Cross Correlation based Skew Estimation

Gatos et al. [54] use the correlation information of 2 or multiple vertical lines. The binary images are smoothed using the run length smoothing algorithm and 2 vertical lines are defined at $1/3$ and $2/3$ from the left border (margin of the image). It is assumed that due to the smoothing the text lines yield to "uniform" horizontal (skewed) lines. For skewed text lines, the interline spacing remains constant, whereas the position of the text line is vertically shifted. A cross-correlation matrix of the vertical lines is build, and the maximum of the vertical projection of the matrix determines the skew angle. To increase the accuracy and robustness multiple text lines are used

in comparison to the method proposed by Yan [182]. It is stated that the presented approach is more efficient than the Hough transform since only a reduced number of image pixels (defined by the vertical lines) is used for the computation.

A cross correlation based on randomly selected regions is introduced by Chen and Ding [25]. A horizontal and vertical cross correlation is performed to distinguish horizontal and vertical text layouts which is common in Chinese or Japanese documents [25]. Based on statistics of the horizontal and vertical cross correlation the direction of the text layout is estimated and the content of the randomly chosen window is classified into a region containing *text* or an *image*. The selection of randomly chosen windows additionally reduces the computational effort. The skew is estimated by the peak value in the horizontal or vertical cross correlation.

2.2.4 Fourier transform based Skew Estimation and Other Methods

A skew estimation done by exploiting the Fourier spectrum is presented by Postl [140] and extended by Peake and Tan [137]. The skew angle of the document is related to the "*direction for which the density of the Fourier Spectrum is the largest*" [170]. The dominant peaks in the Fourier spectrum are colinear and are caused by the line spacing of the text lines. Peake and Tan perform a peak pair detection and calculate the angle with reference to the vertical axes. To avoid the main influence of graphics or charts the image is divided into rectangular blocks (in contrast to Postl [140]). The skew angles are accumulated to an orientation histogram and the main peak determines the document orientation. For the final skew the median value of all angles within a certain value of the main peak is taken. It is stated that the accuracy of the proposed method is within $\pm 0.5^\circ$ and is tested on documents containing only text and text with graphics. The Fourier spectrum can also show colinear peaks corresponding to the character/word spacing. Also vertical lines can cause a different direction (with the highest density) compared to the text lines [170].

Fabrizio et al. [135] cluster all document regions using a K-NN after a preprocessing step. The clustered regions are described by the convex hull. To determine the final orientation the magnitude spectrum of the Fourier transform of all clustered regions convex hulls is exploited. The proposed method achieved rank 1 at the DISEC contest 2013.

A skew detection based on texture direction (gradient direction) is applied by Sauvola and Pietkainen [154] and Sun and Si [170]. Sauvola and Pietkainen determine the main local orientation of subimages based on an edge image which is smoothed with a Gaussian filter. The local orientations based on Chaudhuri [24] and Rao [146] are accumulated into an orientation histogram (called direction histogram). The orientation histogram shows two peaks (landscape vs. upright format) which are extracted to determine the skew. Sun and Si [170] accumulate the gradient orientation of the entire image into the orientation histogram. The histogram is smoothed using a median filter and the maximum determines the orientation of the scanned document image. It is stated that gradient based methods can lead to an error for italic/slanted text.

Egozi and Dinstein [43] calculate the slope of single text lines and use an orientation histogram of the text line angles to determine the skew. The centroids of all CCs of a binarized document image are used as feature points. It is assumed, that a text line can be represented by a straight line which is "*corrupted by Gaussian noise*" [43]. Instead of a least square line fitting

applied to the feature points, a statistical mixture model of straight lines corrupted by Gaussian noise is used. The line parameters are estimated by the Expectation Maximization (EM) method. It is stated that the proposed method has the advantage that the feature points must not be clustered according the line structure.

Bar-Yosef et al. [12] propose a skew estimation based on the distance transform. The distance transform is calculated on the binarized image and the gradients of the distance transform are used to estimate the skew since it is stated that "*the dominant orientation is perpendicular to the orientation of the text lines*" [12]. The orientation of the gradients is accumulated into an orientation histogram, where the maximum peak denotes the current skew of a document. Based on the properties of the distance transform, the approach can be considered as background analysis, which is "*less sensitive to text degradation, and are generally independent of text properties*" [12].

A morphological based approach is presented by Das and Chanda [32]. Instead of smearing a binary input image using e.g. Local Projection Profiles (LPP) to form textlines [34] from single characters or words, a morphological closing (line structuring element) and opening (square structuring element) is used. The opening is performed to remove effects caused by ascenders or descenders. Base lines are determined using a vertical scanning to identify base line pixels. Based on geometrical analysis of the base lines after labeling, small and curved lines are removed. The skew of each baseline is estimated using the endpoints. The final skew for the document is defined as the median value of the skew values of all base lines. It is stated that the algorithm "*expects documents mostly filled with text lines*" [32]. The method is tested on documents containing English, Bangla and Devnagari texts and is considered for skewed documents within a skew angle range of $\pm 3\,^{\circ}$.

A block-based edge detector is used by Hyung Il Koo [91, 135] to extract specific types of straight lines in edge maps. The detected lines can originate from "*text-lines, boundaries of figures, tables, vertical and horizontal separators as well as any combination*" [91]. Based on the lines detected the final skew is estimated using a maximum-likelihood estimation. The proposed method achieved rank 2 at the DISEC Contest 2013.

Carlinet and Fabrizio [135] extract lines using a Line Segment Detector (LSD) proposed in von Gioi et al. [175]. LSD detects edges based on a line representation which groups aligned orientation pixels and has the ability to eliminate false positives (see [175] for a detailed description). The information from the LSD detected lines are merged with convex hulls of objects originating from a clustering of CCs if the document has not enough structure (like line separators or frames). The final skew angle is determined by a Hough transform. At the DISEC contest 2013 the described method achieved rank 3.

2.3 Form Classification and Retrieval

Document type (DOCTYPE) classification, thus form classification is a preprocessing step in DIA [152] for information extraction based on a-priori knowledge of the form layout/structure. The goal of document processing systems (e.g. office automation) is to automatically extract and understand the form content [28,129]. A survey of document image classification is presented in Chen and Blostein [26]. The extraction of form data allows also a manipulation of the data [41].

Form understanding systems (form recognition) [22] comprises form dropout [184] (line and preprinted text removal, stroke reconstruction [38]), item extraction [65], OCR and contextual processing of filled-in information based on the form classification done [22,116]. Forms can be described as "*structured documents that are used for information gathering, storage, retrieval, approval, and distribution*" [116].

The use of forms printed on paper is still common, since it allows a direct manipulation independent of any electronic device and is common for most people [149]. Examples like transfer slips, annual tax declarations and service applications are used in paper form and "*handled daily in business and government organizations*" [163]. Mandal et al. [115] state that the automatic processing of forms with a high count of use (e.g. tax return) will save time, cost and allows for an efficient storage in databases [115]. After the digization document processing systems have to classify the form type to correctly extract the information present on the form.

Figure 2.11 shows exemplarily a form document processing system. If several variants of forms are mixed, a form classification has to be performed [9]. The form modeling describes the feature extraction based on the form representation/definition which is stored in a database for all reference model forms (form templates). Depending on the classification system either a blank form [116] or filled-in forms can be used to create the form model reference database. An input document is digitized and serves as form document image. Based on the form modeling [116] the features describing the input form are extracted. The form classification stage determines the form class and "*allows to select the appropriate reading model*" [28]. Based on the reading model the processing of the form allows to recognize the form content for further processing.

The representation of forms is based on the form modeling and thus on the definition of forms. Duygulu and Atalay [41] define a form as a structured document which consists of [41]:

- "*horizontal and vertical layout lines: straight and continuous;*"
- "*preprinted data: machine printed characters, symbols and pictures;*"
- "*user filled-in data: machine typed, hand-printed or handwritten characters.*"

Most of the state of the art form classification methods use the line structure (crossing types, hierarchical structure) of forms as features solely [19,41,102,163]. Contrary to this information of preprinted data or preprinted and filled-in data can be used [9, 64, 149]. Bart and Sarkar [14] try to determine the repeating structure of a document and use a probabilistic model. In this case only the information of the text is used for the form modeling. Lin et al. [102] states that the frame structure of most of the business forms can be defined by 3 primitives, namely forms with rectangular frames, forms with triangular frames and horizontal lines (Primitive A, Primitive B and Primitive C, see Figure 2.12) and by a combination of these. Figure 2.12 shows the form primitives and possible combinations. A weaker definition is given by Mandal et al. [115] who defined a form of 2 primary types which consists of boxes ($F1$) or horizontal lines ($F2$) ("*markers above which the information is filled-in*" [115]) and a combination of those (see Figure 2.12 d). Arlandis et al. [9] states that form classification is not trivial due to filled-in information which alter global features and similar form variants (see [9]). Methods which are mainly using the line information must be able to deal with "*noises similar to a line, disappeared lines, broken lines or partially disappeared lines*" [19].

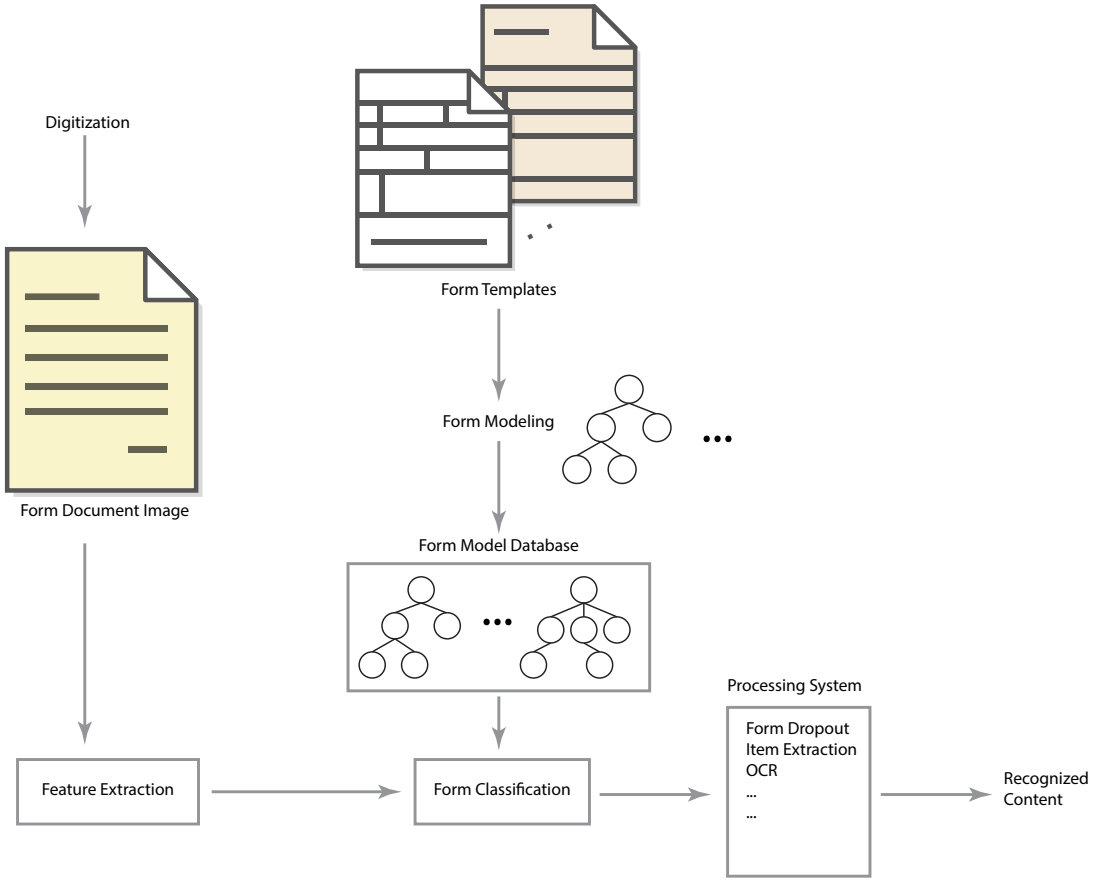


Figure 2.11: Form Document Processing System

In the following, state of the art form classification methods are classified in global image based features, hierarchical descriptions, and local and structural based features. Global image based features use the entire document image and calculate characteristics like pixel density or e.g. projection profiles. Hierarchical descriptions use either the cell structures of forms or the pre-printed/filled-in data, whereas local and structural features determine discriminative local regions or properties like e.g. line crossings.

Form retrieval is defined in Liu and Jain [108] as the following question: "*Given a form image database and a query image, how do we retrieve form images in the database with the same or similar layout structure as the query?*" [108]. According to [41] form retrieval can be used for scale changes, distortions introduced to the scanning process or minor form changes. A different application is the analysis of unsorted scanned documents [124]. Since e.g. forms like "Table of Contents" contains information about single files a form retrieval system is used to analyze the reconstructed document images. The retrieval system is based on the form classification proposed in this work (see Section 2.3).

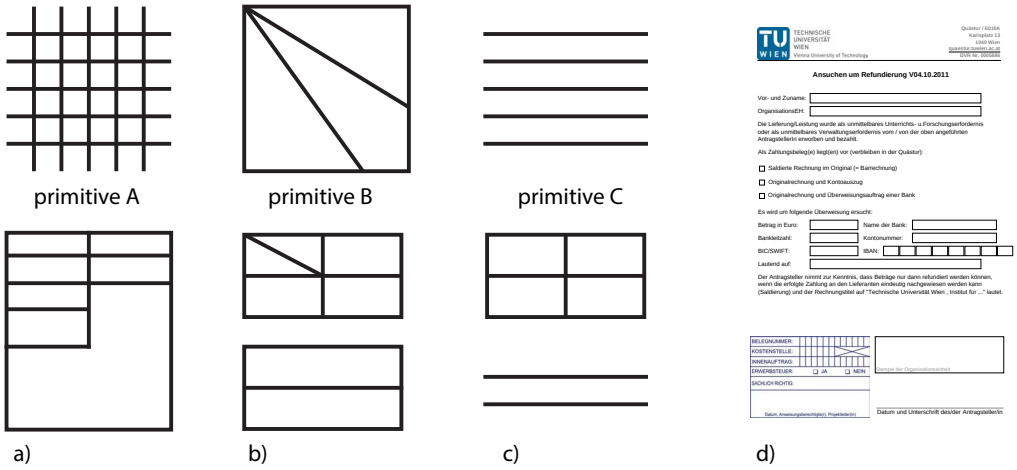


Figure 2.12: Form primitves A,B and C as defined by [102]. a) one or more primitives of A b) two primitives of A whereas one contains a B primitive c) primitive C is outside a combination of primitives A. Courtesy by [102]. d) real world example form, courtesy by TU Wien.

Document form processing systems are summarized in [22, 38, 116]. A first step for the recognition of form documents after the classification is a form registration to handle distortions [65,73]. Field extraction methods can be based on form models [65] or template-free recognition systems [15, 66]. Additionally color [180] or probabilistic graphical models [139] can be used for form recognition.

2.3.1 Global Image Based Features for Form Modelling

Ohtera and Horiuchi [130] present a form classification system for faxed forms. It is stated that FAXOCR systems can use a unique form id and markers for the registration of forms. To be able to handle forms without these features the Hough histogram of vertical and horizontal lines is used for classification. As a preprocessing step characters are separated and the skew is estimated by exploiting the Hough-space $H(\theta, \rho)$, since a rotation is transformed into a shift on the axis defining the angle θ. Additionally the parallel transform is corrected. The form classification is performed by comparing the Hough-histogram of vertical and horizontal lines of the input form image and the model form image. The system is tested with 10, respectively 30 different commercial forms and has a classification rate of 100%. The skew is restricted to $\pm 9\,^{\circ}$ since it is stated that this is the restriction of 14 commercial systems in Japan.

He et al. [61] use the power spectrum of the Discrete Fourier Transform (DFT) of horizontal and vertical projections of binary form images (see Figure 2.13 for an example of form document projection profiles) as feature vector for classification. Since the printed and handwritten text represent the high frequency part of the image (details) "*the low-pass frequency components of the Fourier transform are used to describe the overall image*" [61]. If two different form documents are vertically symmetric (horizontal symmetry axis) the vertical projection profile and the resulting power spectrum of the Fourier transform have a similar structure. An example

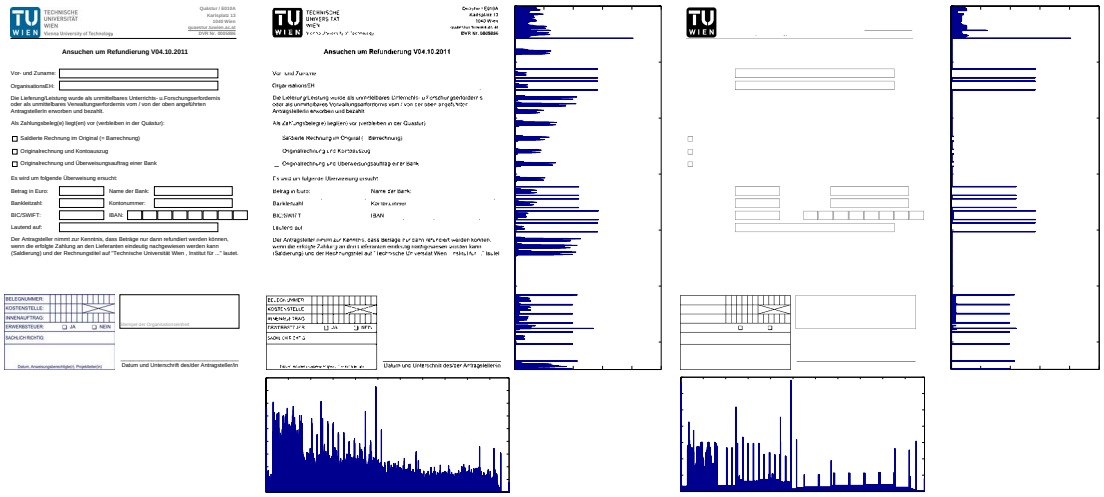


Figure 2.13: Example form document (courtesy by TU Wien): horizontal and vertical projection profiles of the binarized form document with text and filtered text.

is shown in [61]. To classify symmetric document forms correctly, the horizontal and vertical information of a size normalized image is used for classification. To minimize the effect of noise and variations of filled-in items multiple forms of the same class are used for the training of a Backpropagation Neural Network (NN). He et al. [61] tested the approach for 3 different kinds of form documents and used 10 training samples for each class. After the training a set error of 0.005 is reached.

Liolios et al. [107] have developed a form classification system that can also determine the skew of a form document. They use the "*Power Spectral Density (*PSD*) of the forms horizontal projection profile, as a shift invariant feature vector*" [107] (see Figure 2.13 for an example of form document projection profiles). To reduce the dimensionality of the feature vector the Karhunen-Loeve transform [147] is used. The PSD is defined as the Fourier transform of the covariance function [159] and shows the average power distribution as a function of the frequency. The feature vector has a dimensionality of 1025 (1025 frequencies) and is reduced to 128 dimensions using the Karhunen-Loeve transform. To be invariant to a skew angle introduced the template forms (prototypes) are rotated with steps of $0.2\,^\circ$ within $\pm 10\,^\circ$ and a feature vector is calculated for every angle. A codebook is generated using the feature vectors of the prototypes and a Learning Vector Quantization. To classify a form document the Euclidian distance of the feature vector to the class centroids in the codebook is calculated as similarity measure. The system was tested for 26 different form types and reached a minimum accuracy of 98.9%.

Mandal et al. [115] use a combination of global image based and structural features for form classification. As global shape features the 2nd and 4th order moments of the horizontal and vertical projection profiles (see Figure 2.13 for an example of form document projection profiles) are used to determine a subset of the form prototype database. The subset is chosen

on a defined distance measure based on the moments and a statistical threshold. For the final classification relative line positions and line crossings (see Section 2.3.3, Figure 2.14) are used as structural features. Based on these, relationship matrices are defined as proposed in [48] and used for classification. Mandal et al. [115] tested the proposed approach with 40 different types of forms and have an accuracy of 96%. Additionally it is shown, that the defined pre-selection can be performed with an error of 0.25% (3 forms have not been selected).

An approach using an image pyramidal decomposition based on the pixel density of a form image (binary image) is presented by Heroux et al. [64]. Thus, the form document image is described by a feature vector with a dimension of 341 resulting from a 5 level cut pyramid representing the pixel density. K-NN and a Multi-Layer Perceptron (MLP) is used for classification. The method is tested with 27 different types of form documents and 2 up to 5 pyramid levels. At level 5 both classifiers ($k = 1$ for K-NN, 1 hidden layer and 27 neurones forMLP) reach an accuracy of 100%. For a 2 level pyramid the MLP has an error rate of 4.21% compared to 3.16% for K-NN for a total of 570 forms. Heroux et al. [64] present also a form classification based on a tree comparison of the form content (see Section 2.3.2). In Clavier [28] the possibility of a combination of both classification methods is presented (see Section 2.3.2).

2.3.2 Methods based on Hierarchical Descriptions for Form Modelling

Lin et al. [102] use the relationship of adjacent boxes. They use a binarized image of the document form and the line information, whereas in general a form document is described by a combination of 3 basic shapes (see Figure 2.12). Virtual edges are introduced to close all open box frames. Figure 2.12 c) describes a form document which is a combination of primitive C outside primitive A. In this case 2 virtual vertical lines are created to constitute a closed frame of primitive C. After the "closing", each frame is labeled from left to right and from top to bottom using a consecutive number and the primitive type, and a horizontal and vertical graph is created to represent the form document. For the matching of document forms with form prototypes stored in a database, the horizontal and vertical graph is represented as a pair of 1-D strings created by topological sorting. The classification is performed using a string matching. The method is tested with 100 different form types.

Duygulu and Atalay [41] already assume a closed document form which consists of box frames (smallest frame is called "block") and has rectangular border frame. The geometric structure is transformed to a hierarchical structure using the XY-tree method which is proposed by Nagy [122]. Geometric varying form documents which have the same logical structure are handled by defining ambigous blocks. Form identification and retrieval is done by calculating the distance of tree T_1 (to be classified) to tree T_2 (form prototype stored in the database), which is defined as cost for the transformation of T_1 to T_2. The retrieval rate of forms with the same logical structure and geometric modifications is appr. 70%, whereas the retrieval rate of form documents with similar subtypes (geometric modifications) is between 35% and 75% depending on the number of retrieved documents.

Heroux et al. [64] extract a hierarchical structure based on the relationships between the line information and the contents of the form document (text, table and graphics). For that purpose a layout analysis is performed to find homogenous blocks (e.g. paragraphs), whereas text blocks are further divided into text-lines. The entire form document is modelled by a hierarchical tree

representing the contents. To classify a form document the trees of a form document and a form model are compared (see [64]) and features are extracted (number of nodes and overlap rates) which are used to find the nearest model tree. The method is tested on a form database with 1420 forms and 26 different form documents (120 forms belong to unknown form models). The evaluation shows that the proposed methodology can identify unknown form classes with an accuracy of 100%. An overall recognition rate of 87.13% (10 forms/class in the training set) to 99.23% (40 forms/class in the training set) is reached.

2.3.3 Methods based on Local and Structural Features

Fan et al. [48] proposed a method based on line structure features. Nine different types of line crossings are defined (see Figure 2.14) and are detected for every form document image. Based on the line crossing detection a *Line Crossing Relationship Matrix* is defined, where an entry at position (i, j) defines the crossing type of the i^{th} horizontal line with the j^{th} vertical line. Additionally a horizontal and vertical *Line Distance Relationship Matrix* is calculated based on

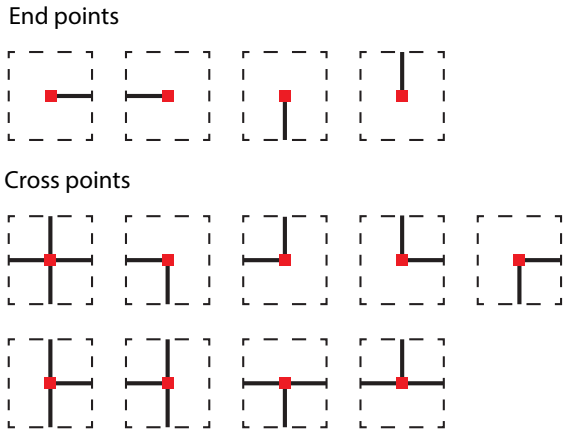


Figure 2.14: Feature Points defined by end points and crossing types of lines [48, 65].

the line length of horizontal and vertical lines. Four length types depending on the size of the form are defined. Matching form model candidates are determined by the number of horizontal and vertical lines which must be below a certain threshold. The classification is based on a combination of the similarity measure of the 3 relationship matrices. The accuracy of the form classification is 100% and 95% if lines are randomly deleted and new lines are inserted. The method is robust against rotation and scaling.

A combined approach of global image based features and line structural features as described above (Fan et al. [48]) is presented by Mandal et al. [115]. The method is described in Section 2.3.1.

Sako et al. [149, 150] introduces a document form classification based on the constellation matching of keywords. Their method extracts keywords of model forms (templates). The keywords (word strings) and the location of the keywords of a single form document are stored in a keyword dictionary. The method is named constellation matching, since "*words are like fixed*

stars in a constellation" [149]. To classify a form, the detected keywords are matched with the keywords in the dictionary (a-priori knowledge) and a matching score is calculated using Dynamic Programming (DP). A final score depending on the number of detected keywords and their locational information is determined [149]. The model form with the highest score is chosen as form template. Due to the score calculation this method can also be used for form retrieval. The keywords of the model forms are chosen according the keyword stability and the keyword uniqueness. The method is tested with 107 different form documents and a form database consisting of 671 samples. The accuracy of the proposed method is 97.1% and a correct rejection rate of 2.9%.

Arlandis et al. [9] proposed a form classification system for forms with minor changes such as e.g. a differing form type number, page number or geometric changes in multi-page forms. As a preprocessing step a skew estimation is performed. To create the form model database the system determines θ-landmarks which define discriminant areas for each document form class. It is stated that θ-landmarks represent "*salient visual features related to a location*" [9]. A chosen θ-landmark of a reference form document must have a high dissimilarity to all other form classes. To measure the dissimilarity a distance function based on correlation can be used. For the form classification a similarity measure based on the determined θ-landmarks is defined. The size of the θ-landmarks is depending on the size of the characters (defining the height, 24 pixel), and the width (64 pixel) is choosen based on experiments. The method is tested with 7 different form classes and a test set of 753 document form images consisting of known and unknown forms. Due to the high computational cost for defining the θ-landmarks a hierarchical method with a pre-classification based on e.g. global features is suggested.

Saund [152] uses a graph lattice for the representation of form documents and a BOW approach for classifcation. Crossing and end points (called junction/termination types) as shown in Figure 2.14 are defined, which are all possible combinations (13) for "*rectilinear line-art*" [152]. It is shown on the NIST tax forms database [37] that solely the count of the extracted junction types is not discriminative enough as a classification feature. Thus, a data graph consisting of the junction/termination points as nodes and the links between them is constructed. Nodes defined by the junctions/terminations defined in Figure 2.14 are called Primitives. Subgraphs are defined by a combination of 2 or more primitives and represent more complex features. In the training discriminative subgraphs are choosen for each form type and Common-Minus-Difference (CMD) is used as a similarity measure. The method is tested with a subgraph size of 2, 3 and 4. A graph lattice is used as data structure, which allows an "*efficient computation of subgraph matches [...] and effective construction of more complex features from smaller ones*" [152]. For subgraph sizes of 1-3 and 1-4 the proposed method has a classification accuracy of 100% on all 11,185 images of the NIST SpecialDatabase2 and SpecialDatabase6.

2.4 State-of-the-Art Evaluation Metrics

This section summarizes evaluation metrics used for binarization, skew estimation and form classification methods. The measures comprise the metrics of the contests DIBCO and DISEC and metrics presented in current state of the art work. The following subsections present the

metrics divided for the preprocessing topics binarization, skew estimation and form classification.

2.4.1 Binarization Evaluation Measures

Binarization methods are either evaluated on a pixel basis compared to a Ground-Truth (GT) (binarized image) or based on semantics of the image which is done by recognizing characters using an OCR. Within DIBCO and H-DIBCO [53, 141, 144, 145] only pixel based evaluation measures are used. The drawback of pixel based measures is the fact that certain metrics (e.g. FM, NRM) do not account for the type of distortion of the binarized information (see Section 2.4.1.1). The advantage is that no recognition system has to be trained. This is especially a drawback if handwritten documents are used. In Sections 2.4.1.1 and 2.4.1.2 popular metrics are presented. The pixel based metrics which are used in DIBCO and H-DIBCO are summarized in Table 2.5 in Section 2.5.1 with respect to the year the contest was held. A subjective evaluation can be performed by humans by a visual inspection [126], e.g. counting the number of broken symbols [173]. Since there is no "*satisfactory precision by humans*" [126] quantitative measures are preferred for scientific evaluations.

2.4.1.1 Pixel Based Evaluation Measures

Pixel based evaluation measures compare the binarization result respectively a GT image on a pixel by pixel basis. The GT image is either labeled manually or can also be done with semi-automatic procedures as proposed by Ntirogiannis et al. [125], if real images are used. The latter constructs a skeletonized GT image and uses a dilation which is constrained by the edge image to create the final GT image. Synthetic images provide the GT image by definition, but since they "*do not reflect the degradation encountered in real documents*" [125, 126] the use of images from real documents like ancient manuscripts is preferred. Figure 2.15 shows an example test image of DIBCO 2011 and the corresponding GT.

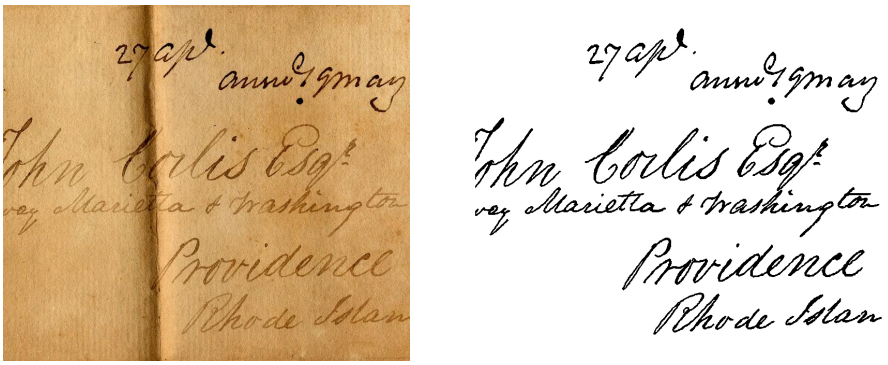


Figure 2.15: Test image of DIBCO 2011 and the corresponding GT.

Recall, Precision and F-measure classifiy pixel as True Positives (TP), False Positives (FP) and False Negatives (FN), which is illustrated in Figure 2.16, showing exemplified a GT of the

binarized character *A* (left) and a possible binarization result (right). Each pixel is marked to show the definition of TP, FN and FP.

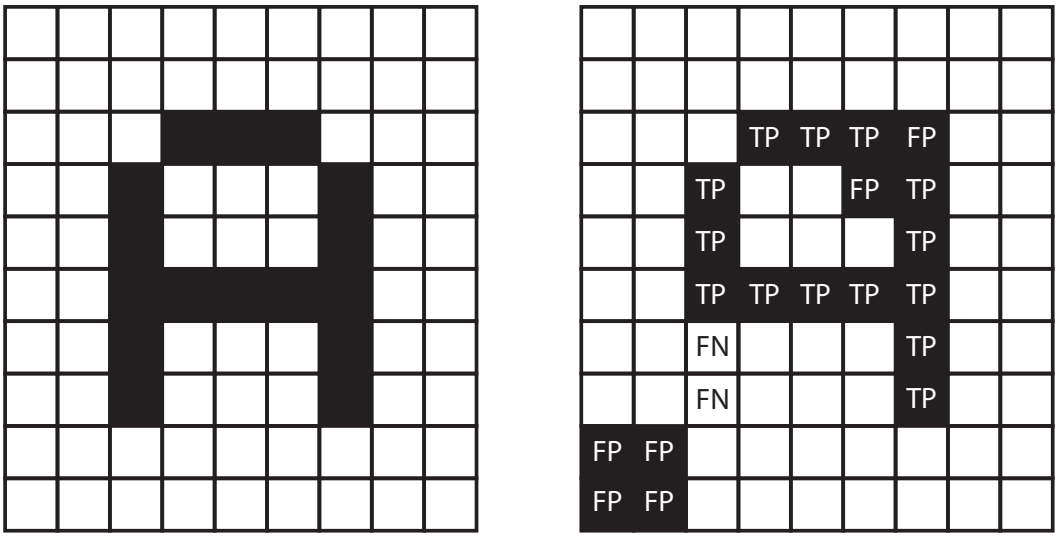


Figure 2.16: Ground Truth of the binarized character A (left). Binarization result (right).

TP denote pixels which are marked as foreground pixels in the GT image and in the binarized image (correctly classified as foreground/text). Pixels which are denoted as foreground in the GT image and classified as background in the binarization result are FN. Pixels detected as foreground in the binarization result but belong to the background in the GT are called FP. According to these definitions Recall and Precision are defined by Equation 2.3.

$$Recall = \frac{TP}{TP + FN} \qquad Precision = \frac{TP}{TP + FP} \tag{2.3}$$

Recall is the fraction of correctly classified foreground pixels divided by the the total number of foreground pixels. A high recall can indicate that the method binarizes low contrast text (no missing parts of characters or words) correctly. In contrast, precision is the number of correctly classified pixels divided by the total number of pixels classified as foreground. Thus, a method with a high precision shows that the method is prone to noise. A combination of precision and recall is defined by the F-measure, which is the harmonic mean of precision and recall. The F-measure (balanced F-Score) is then defined as follows:

$$F - measure = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision} \tag{2.4}$$

Recall, Precision and F-Measure take the number of flipped pixels into account independent of the position. Thus, at H-DIBCO 2010 [141] the pseudo-Recall and the pseudo-F-measure proposed by [125] is introduced. Compared to F-Measure and Recall only the skeleton pixels of foreground objects in the GT image have an impact on the final values since "*each character*

has an unique shilouette which can be represented by its skeleton" [141]. This is illustrated in Figure 2.17, which shows the GT image (a) and the skeleton of the GT image, $SG(x, y)$ in (b).

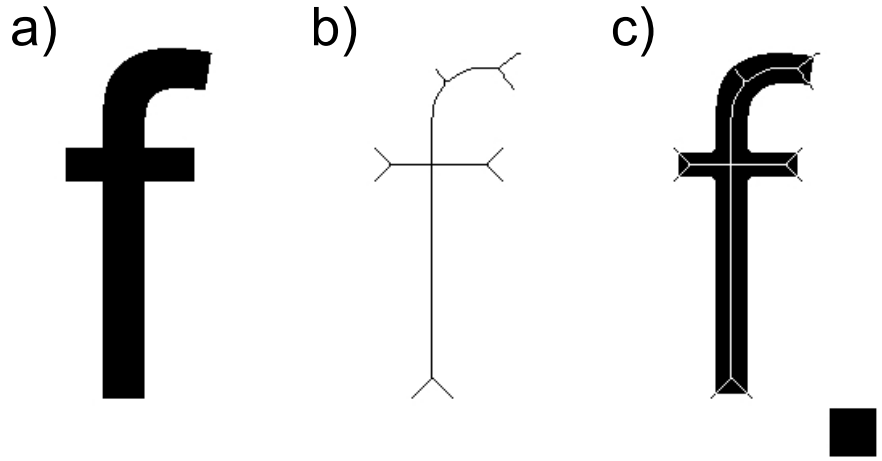


Figure 2.17: a) GT Image b) Skeleton of GT c) Binarized Image: only white pixels have an impact on the pseudo-Recall.

Figure 2.17 c) shows the binarized image $B(x, y)$ and the overlay of the GT skeleton where white pixel denote $SG(x, y) \cdot B(x, y)$. Thus, p-Recall and p-FM are defined by Equation 2.5 and Equation 2.6.

$$p - Recall = \frac{\sum_{x=1,y=1}^{x=M,y=N} SG(x, y) \cdot B(x, y)}{\sum_{x=1,y=1}^{x=M,y=N} SG(x, y)} \tag{2.5}$$

$$p - FM = \frac{2 \cdot p - Recall \cdot Precision}{p - Recall + Precision} \tag{2.6}$$

The Peak Signal-to-Noise Ratio (PSNR) is the ratio between the difference value C of foreground and background (255) and the mean squared error of the binarized image $B(x, y)$ and the GT image:

$$PSNR = 10 \log \left(\frac{C^2 \cdot M \cdot N}{\sum_{x=1}^{M} \sum_{y=1}^{N} (B(x, y) - GT(x, y))^2} \right) \tag{2.7}$$

PSNR measure has a logarithmic scale and is also used to measure the quality of lossy image compression methods [74, 110] and is also used in Stathis et al. [167]. In Lu et al. it is stated that the PSNR does not match the distortion perceived of the human visual system since the mutual relation of pixels is not considered [110]. In DIBCO 2009 and H-DIBCO 2010 the Negative Rate Metric (NRM) is used which is the arithmetic mean of the false negative rate NR_{FN} and the false positive rate NR_{FP}. Thus, NRM is defined by Equation 2.8:

$$NRM = \frac{NR_{FN} + NR_{FP}}{2} \tag{2.8}$$

where NR_{FN} and NR_{FP} is defined by:

$$NR_{FN} = \frac{FN}{FN + TP} \qquad NR_{FP} = \frac{FP}{FP + TN} \tag{2.9}$$

The false positive rate NR_{FP} takes also the TN into account, which are all pixels correctly classified as background [164]. Thus, the NR_{FP} can be interpreted as a *noise* measure in relation to the background. The Misclassification Penalty Metric (MPM) is used until DIBCO 2011. The MPM "*evaluates the predicition against the* GT *on an object-by-object basis*" [53] where pixel classified as FN or FP are weighted with the distance from the ground truth objects border.

$$MPM = \frac{MP_{FN} + MP_{FP}}{2} \tag{2.10}$$

where MP_{FN} and MP_{FP} are defined as follows, with d^i_{FN} denoting the distance of the i-th FN pixel and d^j_{FP} is the distance of the j-th FP pixel:

$$MP_{FN} = \frac{\sum_{i=1}^{N_{FN}} d^i_{FN}}{D} \qquad MP_{FP} = \frac{\sum_{j=1}^{N_{FP}} d^j_{FP}}{D} \tag{2.11}$$

D is defined as the sum over all pixel to contour distances. As distance measurement d the chessboard distance is used, which defines the distance between pixel p_1 with position (x_1, y_1) and pixel p_2 with position (x_2, y_2) as $d = max(|x_1 - x_2|, |y_1 - y_2|)$.

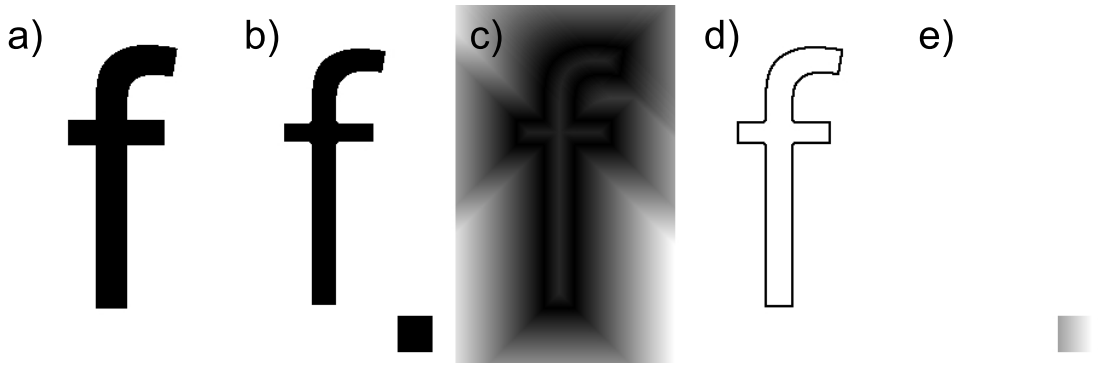


Figure 2.18: a) GT Image b) Binarized Image c) Pixel Distance to the GT's border Pixel d) d^i_{FN} e) d^i_{FP}

Figure 2.18 visualizes the calculation of the MPM, where a) shows the GT image and b) shows a possible binarization result. The distance image to the GT border pixel is shown in Figure 2.18 c). The distance values of all FN pixels are illustrated in Figure 2.18 d), whereas the distance values of all FP pixels are shown in e). Thus, the sum of the distances of all FN (MP_{FN}) and all FP (MP_{FP}) normalized by D have an impact on the MPM metric. It is stated that a "*low* MPM *score denotes that the algorithm is good at identifying an object's boundary*" [53].

Since DIBCO 2011 the Distance Reciprocal Distortion Metric (DRD) metric is introduced, which has mainly replaced the NRM and MPM at H-DIBCO 2012 and DIBCO 2013. The

DRD is a measure of the visual distortion [110] in binary images which correlates with the human visual perception [142]. Lu et al. [110] states that due to the sharp contrast of binarized images the "*distance between pixels is found to play an important role in human perception of distortion*" [110] and thus the reciprocal is defined as metric to measure the distortion. The DRD metric is defined by Equation 2.12:

$$DRD = \frac{\sum_{k=1}^{S} DRD_k}{NUBN} \tag{2.12}$$

with $NUBN$ is the number of non uniform 8×8 blocks (not entirely black or white), S is the number of flipped pixels and DRD_k is the distortion of the k-th flipped pixel and is defined by Equation 2.13:

$$DRD_k = \sum_{i,j} [|B_k(i,j) - GT_k(x,y)| \times W_{Nm}(i,j)] \tag{2.13}$$

where W_{Nm} is a 5×5 weight matrix, where each element is the Euclidean distance to the center element normalized by the sum of all distances (for a detailed desciption see [110]). The calculation is based on the fact that "*the distortion (flipping) of one pixel is more visible when it is in the field of view of the pixel in focus. The nearer the two pixels are, the more sensitive it is to change one pixel when focusing on the other*" [110].

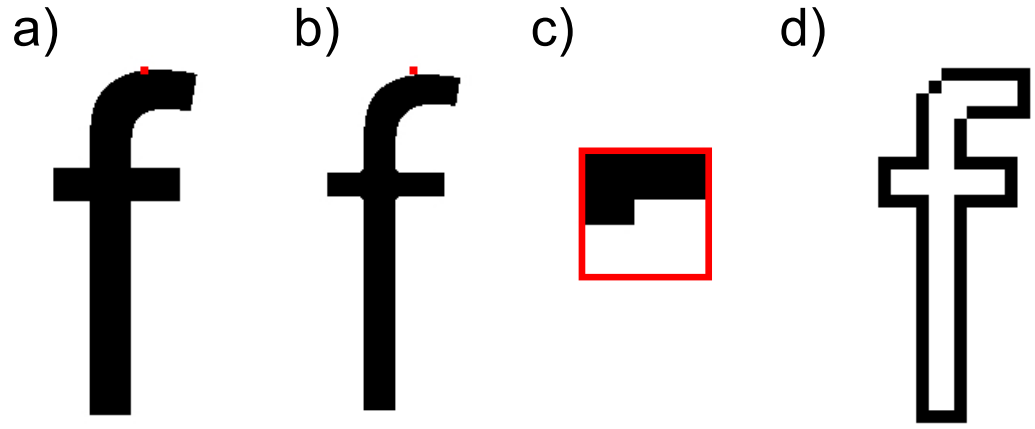


Figure 2.19: a) GT Image b) Binarized Image c) 5×5 pixel window centered at the first flipped Pixel k (GT=1, Binarized Image=0) - white Pixels have an impact on DRD_k d) NUBN

Figure 2.19 a), b), c) shows the calculation of DRD_1 where the center of the red 5×5 square is the first flipped pixel ($k = 1$). All non uniform blocks (8×8) of the GT image (Figure 2.19 a) are shown in Figure 2.19 d).

Barney Smith et al. [164] use a Normalized Cross Correlation (NCC) as an additional pixel based evaluation metric. The NCC is defined by Equation 2.14:

$$NCC = \frac{\sum_{x=1}^{N} \sum_{y=1}^{M} (GT(x,y) - \overline{GT}) \cdot (B(x,y) - \overline{B})}{\sqrt{\sum_{x=1}^{N} \sum_{y=1}^{M} (GT(x,y) - \overline{GT})^2 \sum_{x=1}^{N} \sum_{y=1}^{M} (B(x,y) - \overline{B})^2}} \tag{2.14}$$

The NCC has not been used in DIBCO but in individual publications for the evaluation of binarization methods [164, 166]. In Smith et al. [164] it is stated that the NCC is used in the context of image registration. Additional metrics which are mentioned in Ntirogiannis et al. [126] are the chi-square metric [11] and the geometric-mean accuracy [136]. Ntirogiannis et al. [126] propose an evaluation metric designed for the analysis of binarized documents which extends recall and precision by a weighted measure around the GT border and incorporates also the stroke width to overcome the drawbacks of the metrics precision introduced, NRM, MPM, NRM or PSNR [126] (see also Section 2.4.1.3). Thus, a pseudo-F-Measure F_{ps} based on a pseudo recall R_{ps} and pseudo precision P_{ps} (different from $p - Recall$ 2.5 and $p - FM$ 2.6) is introduced:

$$F_{ps} = \frac{2 \cdot R_{ps} \cdot P_{ps}}{R_{ps} + P_{ps}} \tag{2.15}$$

where R_{ps} and P_{ps} is defined by:

$$R_{ps}(x,y) = \frac{\sum_{x=1,y=1}^{x=M,y=N} B(x,y) \cdot G_w(x,y)}{\sum_{x=1,y=1}^{x=M,y=N} G_w(x,y)} \quad P_{ps}(x,y) = \frac{\sum_{x=1,y=1}^{x=M,y=N} G(x,y) \cdot B_w(x,y)}{\sum_{x=1,y=1}^{x=M,y=N} B_w(x,y)} \tag{2.16}$$

M and N denote the image dimensions. G_w is the weighted ground truth image, and B_w defines the weighted binarized image. The weight map of image G_w results from the distance map of the GT border (maximum values of foreground objects are at the position of the skeleton) normalized for segments which connect two anti-diametric points (normal to the skeleton). To create the weight map for B_w each foreground object is extended by its average stroke width. Again the distance map is applied to the extended regions and normalized (similar to G_w but taking the background skeleton into account). For a detailed definition of the weighting see [126]. It is stated that the proposed metric has a "*better correlation with OCR*" [126] based evaluation and overcomes problems of distance based measures like MPM (high penalty of FP with a high distance to the GT; in contrast to the proposed P_{ps}) and DRD (low penalty of FP near foreground objects/characters; in contrast to R_{ps}) [126].

2.4.1.2 OCR Based Evaluation Measures

In contrast to pixel based metrics, OCR based evaluation metrics apply a recognition system on the binarized image and compare the recognized text on word [59, 128] or character basis [126]. Ntirogiannis et al. [126] states that state of the art OCR engines are focused on machine printed text and can lead to insufficient results for handwritten text in historical documents. The performance of the OCR depends also on the base method of the engine and thus provides no "*direct evaluation of the* OCR" [126]. A state of the art OCR engine is e.g. ABBYY Finereader [50]. An additional drawback is that OCR engines must be trained regarding the font used.

An OCR based evaluation is done by O'Gorman [128] who defines a recognition rate as an accuracy metric for the binarization as follows:

$$accuracy = \frac{\#words\ with\ all\ characters\ correctly\ recognized}{\#words} \tag{2.17}$$

Gupta et al. [59] adapted the proposed recognition rate of O'Gorman and defined accuracy by:

$$accuracy = \frac{\sum_{j=1}^{J} min(groundtruth(j), OCR(j))}{\sum_{j=1}^{J} min(groundtruth(j)} \tag{2.18}$$

where J is the number of unique words, $groundtruth(j)$ is the count of the unique word j in the GT text and $OCR(j)$ is the count of the word j in the recognized text. The adaption of the accuracy defined by O'Gorman is done since words like *the* can be recognized if only parts of words like *their* are detected or column breaks are not detected correctly [59].

Gatos et al. [55, 56] uses the Levenshtein distance [99] between the GT text and the recognized text. The Levenshtein distance is a distance measure between two strings and defines the minimum number of character edits (deletion, insertion and substitution) to convert the first string into the second one.

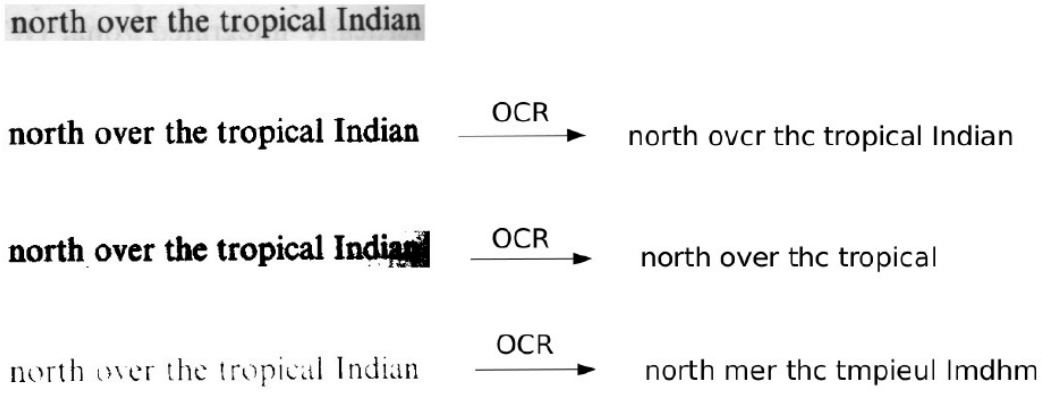


Figure 2.20: Effect of Binarization on OCR, courtesy of Faisal Shafait [162]

Figure 2.20 [162] shows possible effects of the binarization on the OCR result. The evaluation of binarization methods using an OCR system "*has been criticized as being a metric of how well the binarization output fits with the remainder of the* OCR *processing, and not a direct measure of the binarization algorithm itself*" [165].

2.4.1.3 Interpretation of Pixel Based Binarization Metrics

F-Measure (Recall, Precision) and PSNR take only the number of flipped pixel (FP and FN) into account independent of their position. In Ntirogiannis et al. [126] examples of images are shown which achieve a higher F-Measure although more broken characters and noise exist compared to a binarized result with missing pixels at the border region of characters. It is also stated that distance-based measures (MPM or DRD) "*can overpenalize a binarized image with noise far from the text preserving the textual information*" [126]. Figure 2.21 shows exemplarily 2 characters, where missing parts are marked black. The first example shows a *c* with missing

pixels at the border at the same amount of missing pixels at the center resulting in a broken character, thus changing the characters topology. The second example shows an *m* with the same effect. F-Measure and PSNR will achieve the same performance rate due to the same amount of missing pixels. MPM and DRD will penalize the *c* with missing pixels at the border (due to their definition, see Section 2.4.1.1) although in the second case the topology is changed. The $p-Recall$ 2.5 and $p-FM$ 2.6 (see Section 2.4.1.1) can overcome the mentioned problems.

Figure 2.21: Two examples of characters with the same number of missing pixels (regarding each character) at different positions, courtesy of [126].

A different example is shown in Figure 2.22. In Figure 2.22 a) the synthetic GT image of the character *f* is shown, which is blurred and Gaussian noise added b), c). To show the behaviour of the metrics used at DIBCO 4 different cases are considered: the synthetic image c) (Gaussian blur and noise) is thresholded with Otsu, the original GT image a) is eroded e) and dilated f), and noise is added to the border g).

Table 2.1 shows the results of the metrics Precision, Recall, p-Recall, F-Measure, p-F-Measure, PSNR, NRM , MPM and DRD. The measures DRD and MPM penalize the eroded and dilated *f* in contrast to the binarized image with blur and noise (Figure 2.22 d) or the image with the flipped border pixel (Figure 2.22 f). This shows that although the topology is not changed a penalty is applied since the border pixel are completely changed in the eroded or dilated version.

It is also shown that the p-F-Measure penalizes the dilated *f* (88.42%) compared to the eroded *f* (96.44%). A different example is shown in Figure 2.23 where the GT image of character *f* (Figure 2.23 a) is cut in 2 halves, such that the number of pixels is appr. equal in both halves (Figure 2.23 b) and c). The results of the metrics are shown in Table 2.2.

The metrics Recall, Precision, F-Measure and PSNR achieve approximately the same result since the same number of pixels are flipped. In contrast to the eroded and dilated image also the measures MPM and DRD have approximately the same penalty since the border is "*equally*" distorted. The $p-FM$ of the upper half (73.03%) is less penalized than the lower half (59.63%) since the main structure of the skeletonized *f* is in the upper half.

Barney Smith and An [165] have also analyzed the effect of GT on image binarization and it is shown that "*the three edge tolerant metrics all favor a thin* GT *and will penalize algorithms that produce wider strokes*" [165].

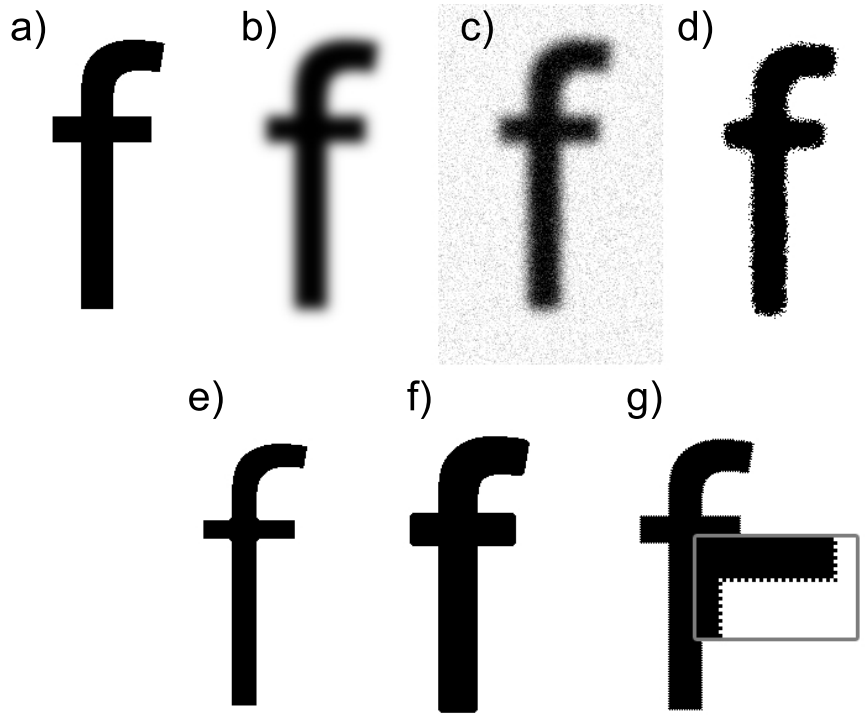


Figure 2.22: a) GT image, b) GT image with Gaussian blur σ=5, c) Gaussian noise added ($m = 0, \sigma = 0.01$) to the blurred GT image, d) Otsu Threshold of c), e) Eroded GT image (disc, $size = 4$), f) Dilated GT image (disc, $size = 4$), g) Border Pixel of "f" flipped alternating. See Table 2.1 for Evaluation Measures.

Input Image	"f" - Otsu (blur+noise) (Figure 2.22 d)	"f" eroded (Figure 2.22 e)	"f" dilated (Figure 2.22 f)	"f" border pixel flipped (Figure 2.22 g)
Precision [%]	91	1	79	95
Recall [%]	97	74	1	1
p-Recall [%]	95	93	1	1
F-Measure [%]	94.51	85.48	88.42	97.88
p-F-Measure [%]	93.64	96.44	88.42	97.88
PSNR	17.99	14.48	14.35	22.18
NRM $\times 10^{-2}$	2.04	12.67	2.13	0.35
MPM $\times 10^{-3}$	0.38	0.63	1.30	0.10
DRD	7.58	19.30	20.07	2.45

Table 2.1: Evaluation Measures on synthetic images d) e) f) g) of Figure 2.22.

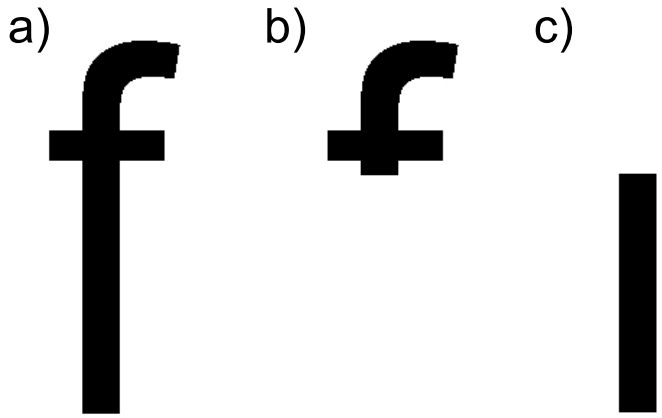


Figure 2.23: a) GT image b) upper half of the GT image c) lower half of the GT image. The GT image is cutted, such that the number of segmented pixels in both halves is appr. equal (see Table 2.2.

Input Image	"f" - upper half see Figure 2.23 b)	"f" - lower half see Figure 2.23 c)
Precision [%]	1	1
Recall [%]	0.50	0.49
p-Recall [%]	0.57	0.42
F-Measure [%]	66.75	66.57
p-F-Measure [%]	73.03	59.63
PSNR	10.98	10.97
NRM $\times 10^{-2}$	24.94	25.05
MPM $\times 10^{-3}$	8.74	7.86
DRD	45.02	45.11

Table 2.2: Evaluation Measures on a synthetic image which is cutted in two halves (see Figure 2.23).

2.4.2 Skew Evaluation Measures

Skew estimation methods are evaluated on (binary) document images which are skewed by a-priori known angles. For the DISEC the skew angle is restricted to $\pm 15^\circ$ and a set of 175 images is used where each image is randomly rotated with 10 different angles. 200 examples from the resulting 1750 skewed images provide a test dataset, whereas the rest (1550 images) build the benchmark dataset [135]. For the evaluation in the DISEC the angle difference $E(j)$ between the determined skew angle of a method and the GT of an image j is calculated. All methods are ranked by the following three metrics:

- Average Error Deviation (AED) 2.19

- TOP80 AED - the first 80% of the sorted differences $E(j)$ (ascending order) are taken into account 2.19
- Average percentage of Correct Estimations (CE) - all images with with a difference of $E(j) \leq 0.1$ are taken into account 2.21

Thus, AED, TOP80 and CE are defined as follows:

$$AED = \frac{\sum_{j=1}^{N} E(j)}{N} \tag{2.19}$$

where N is size of the benchmark dataset.

$$TOP80 = \frac{\sum_{j=1}^{M} E(j)}{M} \tag{2.20}$$

where M is $0.8 \cdot N$ (1240 for the DISEC benchmark dataset with a size of 1550 images). By ranking the methods using the TOP80 criterion categories of test images for which the method has not been designed are excluded (e.g. technical sketches). The CE criterion determines the percentage of images (respectively the size of benchmark dataset n), which have an angle difference $E(j)$ smaller than $0.1\,^{\circ}$.

$$CE = \frac{\sum_{j=1} NK(j)}{N} \text{ where } K(j) = \begin{cases} 1 & \text{if } E(j) < 0.1 \\ 0 & \text{otherwise} \end{cases} \tag{2.21}$$

In Papandreou [135] it is stated that $0.1\,^{\circ}$ is choosen since greater angles are " *visible to a human observer*" [135]. For the final ranking at DISEC the rank of AED, TOP80 and CE are accumulated. Additional metrics which are used in other publications are the mean, standard deviation, and the median of the angle differences $E(j)$ [12, 44]. Epshtein [44] has also defined catastrophic errors, which are "*cases where the detected text orientation differs from the ground truth by more than* $\pi/10$". It has to be mentioned that the method of e.g. Epshtein [44] and the proposed method have no restrictions on the skew angle compared to the DISEC dataset. Thus, the measure of catastrophic errors can be used to show also the mean and the variance error if all catastrophic images are skipped. This can be compared to the TOP80 criterion which skips the images categories (20%) for which the method is not designed.

2.4.3 Form Classification Evaluation Measures

For the form classification, accuracy a is employed as a performance measure. The accuracy is computed by

$$a = \frac{tp + tn}{tp + fp + fn + tn} \tag{2.22}$$

with tp being the sum of true positives (a true positive is a document labeled as belonging to the positive class), fp being the sum of false positives (documents incorrectly labeled as belonging to the positive class), fn being the sum of false negatives (documents which are not labeled as

	$form_p$	$!form_p$
$form$	tp	fn
$!form$	fp	tn

Table 2.3: Confusion matrix containing the two labels "form document" and "non form document"; the indices p denote the predicted class label.

	a_0	a_1	$\cdots$	$\mathbf{a_i}$	$\cdots$	a_n
c_0				fp		
c_1				fp		
$\vdots$				$\vdots$		
$\mathbf{c_i}$	fn	fn	$\cdots$	tp	$\cdots$	fn
$\vdots$				$\vdots$		
c_n				fp		

Table 2.4: Confusion matrix with $n = 10$ (9 different form document classes and one class containing all non form documents); a_i are predictions of class i, and c_i are the true class labels.

positive class but should be) and tn being the sum of true negatives (documents correctly labeled belonging to the negative class). Table 2.3 shows the confusion matrix with a *form document* denoting the positive class and a *non form document* denoting the negative class.

This definition is applicable if only 2 class labels "form document" and "non form document" are considered. In the following the definition is extended to 10 classes (9 different form document classes and one class containing all non form documents). Thus *non form documents* are treated as a different type of a form document, and true positives tp_i, false positives fp_i, and false negatives fn_i of a given class i are defined by:

$$\begin{aligned} tp_i &\ldots \langle a_i, c_i \rangle \\ fp_i &\ldots \langle a_i, c_{j \neq i} \rangle \\ fn_i &\ldots \langle a_{j \neq i}, c_i \rangle \end{aligned}$$

where $i, j \in 0...n$ and $n = 10$ to represent all classes (form documents and non form documents). To illustrate the definitions, a confusion matrix with corresponding labels is given in Table 2.4.

Furthermore precision p_i and recall r_i are calculated for each class i in order to allow for drawing conclusions about the nature of errors and class confusions. Given the previously defined true positives tp_i, false positives fp_i, and false negatives fn_i; precision p_i, recall r_i, and

accuracy a_i of a class i are defined as:

$$p_i = \frac{tp_i}{tp_i + fp_i}$$
$$r_i = \frac{tp_i}{tp_i + fn_i}$$
$$a_i = \frac{tp}{tp_i + fp_i + fn_i}$$

2.5 Comparison and Summary of Existing Approaches

The state of the art in document image binarization is evaluated within DIBCO and H-DIBCO. In Section 2.5.1 the results of the binarization methods of the contests are summarized. The evaluation measures used are described in detail in Section 2.4.1. An additional evaluation survey of binarization methods for historical documents is summarized in Stathis et al. [167]. Results of form classification methods are discussed in Section 2.5.3. Section 2.5.2 presents the results of the DISEC 2013. Evaluation measures to compare results of skew estimation methods are summarized in Section 2.4.2.

2.5.1 Analysis of Binarization Methods

As summarized in Section 1.1 the first DIBCO was held in 2009 [53]. Until 2013 a DIBCO contest was organized every year within the ICDAR and a H-DIBCO in conjunction with the ICFHR. Table 2.5 shows the evaluation measures used for every contest.

	DIBCO 2009	H-DIBCO 2010	DIBCO 2011	H-DIBCO 2012	DIBCO 2013
Precision	+	+	+	+	+
Recall	+	+	+	+	+
F-Measure	+	+	+	+	+
p-F-Measure	-	+	-	+	+
PSNR	+	+	+	+	+
NRM	+	+	-	-	-
MPM	+	+	+	-	-
DRD	-	-	+	+	+

Table 2.5: Evaluation Measures of Document Binarization Contests (2009-2013).

It can be seen that F-Measure and Peak-Signal-to-Noise Ratio are used in all contests. At the last 2 contests the DRD and the pseudo-F-Measure are introduced while the NRM and the MPM are omitted.

Images from the contests and their distortions are shown exemplarily in Figure 1.1 and Figure 1.2 (see Section 1.1). The number of participants and the results of the first 10 places (Score, F-Measure and PSNR) are summarized in Table 2.6 (DIBCO 2011 [142]), Table 2.7 (H-DIBCO

Rank	Score	Method	F-Measure (%)	PSNR (%)
1	309	T. Lelore and F. Bouchara	80.86	16.13
2	346	B. Su et al.	85.20	17.15
3	429	N. Howe	83.21	17.84
4	470	I. Ben Messaoud et al.	-	-
5	489	N. Tanaka	-	-
6	515	V. Papavassiliou and F. Simistira	-	-
7	532	R. Neves and C.A.B. Mello	-	-
8	600	T.H. Ngan Le at al.	-	-
9	610	Abdelaali Hassaine et al.	-	-
10	620	M. Zayed	-	-
15	715	Kleber et al.	-	-

Table 2.6: Results of DIBCO 2011 (Rank 1-10, 18 participants)

Rank	Score	Method	F-Measure (%)	PSNR (%)
1	172	N.Howe	89.47	21.8
2	340	T. Lelore and F. Bouchara	92.85	20.57
3	412	B. Su et al.	91.54	20.14
4	435	O. Nina	90.38	19.3
5	494	Y. Yazid et al.	91.85	19.65
6	501	M. Ramirez-Ortegon et al.	89.98	19.44
7	531	H. Nafchi et al.	91.16	19.32
8	570	H. Nafchi et al.	90.20	19.07
9	575	A. Okamoto et al.	89.69	19.0
10	601	L. Ma et al.	89.48	18.74

Table 2.7: Results of H-DIBCO 2012 (Rank 1-10, 24 participants)

2012 [144]) and Table 2.8 (DIBCO 2013 [145]). Since the evaluation metrics used for the final ranking changed within the contests, the final score is not comparable for all contests. The F-Measure of Rank 1 to 3 is within 89% to 92% for the datasets used in the years 2009, 2010, 2012 and 2013. Only in the year 2011 (DIBCO 2011, see Table 2.6) the F-Measure of the first 3 ranks is within 80% to 85%. The challenges of the images used within DIBCO 2011 consist mostly of bleed-through text, bleed-through text in combination with background stains and images with a Gaussian background noise (comparable to noise existent in carbon copies). The results lead to the question of the best result a binarization method can achieve without a recognition step. Edge based methods cannot differentiate between low contrast parts of a text compared to low contrast noise or e.g. the ruling of a paper. Combination based methods with a machine learning approach for the final decision are able to differentiate between low contrast text belonging to the bleed-through text and the foreground text.

Binarization methods must be able to detect and suppress bleed-through text. Although the

Rank	Score	Method	F-Measure (%)	PSNR (%)
1	322	B. Su et al.	92.12	20.68
2	342	N. Howe	92.7	21.29
3	362	R. Moghaddam et al.	91.81	20.68
4	408	T. Lelore and F. Bouchara	91.69	20.54
5	636	M. Ramirez-Ortegon et al.	90.92	19.32
6	642	Y. Hassaine et al.	89.77	19.26
7	646	R. Moghaddam et al.	89.79	18.99
8	688	H. Nafchi et al.	88.95	18.74
9	716	Y. Hassaine et al.	89.46	19.05
10	725	R. Neves et al.	89.29	18.5

Table 2.8: Results of DIBCO 2013 (Rank 1-10, 23 participants)

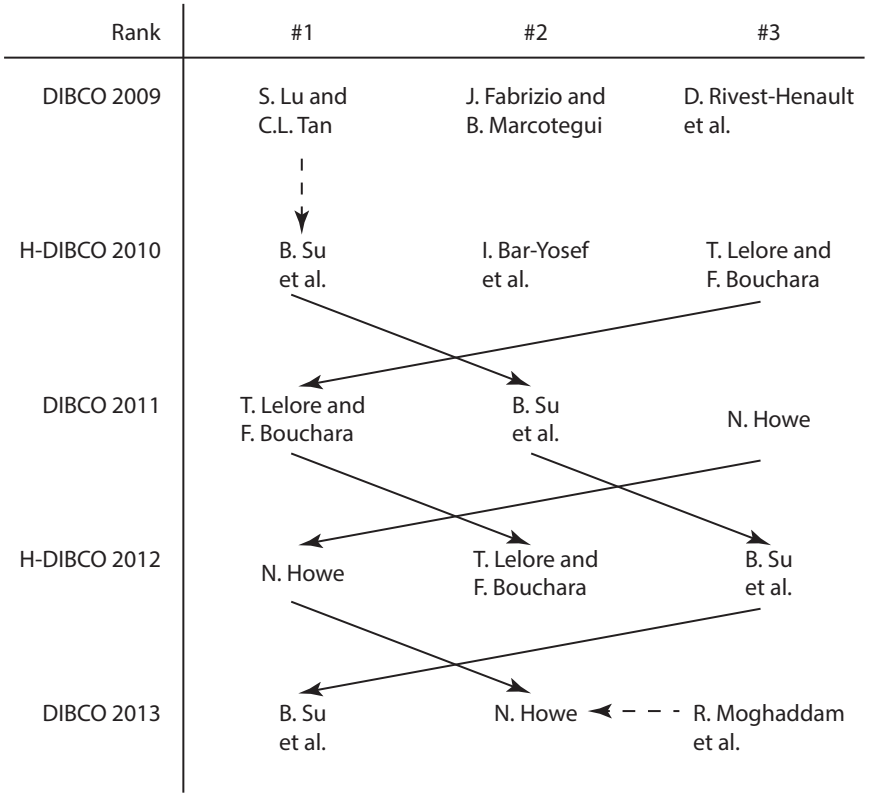


Figure 2.24: DIBCO Overview and Relations of the first 3 Ranks.

last DIBCO contests showed that this problem can be solved with the submitted algorithms, an important step in state of the art methods is the applied post-processing.

The fundamentals of the submitted state of the art binarization methods can be summarized

in keywords as follows:

- Image contrast
- Image gradient
- Edge detection/Edge Map
- Minimization of a global energy function
- Combination of binarization methods

Figure 2.24 shows the first 3 ranks of each year of DIBCO and the relations between the submitted methods. Arrows indicate that a method is based on the method submitted the previous year, or a combination of it (dashed arrow).

Exemplarily, N. Howe participated in DIBCO 2011 with a method based on the minimization of a global energy function [69] and achieved rank 3. In the following year he adapted his method [69] and introduced an automated parameter selection [70] which achieved rank 1 in H-DIBCO 2012. Thus the methods are connected with an arrow. A different example is the method submitted by Su et al. [145] which can be traced back until the year 2010. Common problems of state of the art methods are related to bleed-through text or palimpsest text, as well as e.g. washed-out ink. Bleed-though text as shown in Figure 1.1 has a low contrast compared to the foreground text but still text characteristics regarding the texture. Errors can occur if the foreground text has almost the same contrast as the bleed-through or palimpsest text. Analyzing state of the art methods shows 2 trends in binarization: extended post-processing and a combination of different binarization methods [142, 144, 145].

2.5.2 Analysis of Skew Determination Methods

The first DISEC [135] was held 2013 in conjunction with the ICDAR. Figure 2.25 shows exemplarily two pages comprising a technical sketch consisting of straight lines and only some text labels and a scan of a book page with a chapter caption. The benchmark images (175 documents) "*contain various sizes of document images, any kind of mixed content, vertical and horizontal writing, multi-sized fonts and multiple number of columns in the same document*" [135] in different languages (English, Chinese, Greek, Japanese, Bulgarian, Russian, Danish, Italian, Turkish and ancient Greek languages). The skew angle for the contest was restricted to $-15\,^{\circ}$ to $+15\,^{\circ}$.

Additionally all images are binarized since it is stated by the competition organizers that it is common for big archives to scan documents in black and white. Table 2.9 shows the results of the first 10 ranks.

The winning method proposed by J. Fabrizio exploits the magnitude spectrum of the Fourier transform of the convex hull of clustered regions in the image. Thus the method is based on a local Fourier transform in combination with clustering. Rank 3 submitted by E. Carlinet and J. Fabrizio uses the approach of J. Fabrizio (rank 1) in combination with a Line Segment Detector (LSD) [175]. The final skew is based on the Hough transform of the combined image. Rank 2 is based on the orientation of detected lines [91] (see Section 2.2.4). It can be seen, that the winning

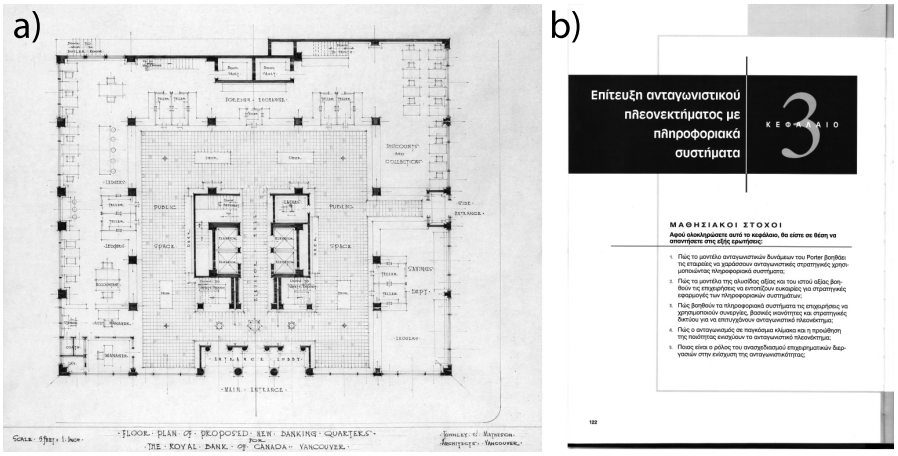


Figure 2.25: DISEC 2013 example images, where a) shows a technical sketch and b) a s can of a book page.

Rank	Method	AED	TOP80	CE	S
1	J. Fabrizio	0.072	0.046	77.48	3
2	H. I. Koo and N. I. Cho	0.085	0.051	71.23	6
3	E. Carlinet and J. Fabrizio	0.097	0.053	68.32	10
4	C. Dalitz	0.184	0.057	68.90	12
5	M. Diem, F. Kleber and R. Sablatnig	0.103	0.058	65.42	15
6	X. Wu, Y. Tang and H. Wang	0.730	0.061	65.74	20
7	K.H. Steinke	0.227	0.069	58.84	21
8	O. Naydin	0.184	0.089	50.39	25
9	X. Wu, Y. Tang and H. Wang	0.750	0.078	57.29	28
9	V. Roy	0.768	0.073	58.32	28

Table 2.9: Results of DISEC 2013 (Rank 1-5, 12 participants)

methods use line information as a feature for a stable and robust skew information. Either the lines are extracted from tables, separators, or originate from entities like text lines. Additionally almost all of the images have at least 2 or 3 paragraphs of text or consist of sketches, comics, ... Thus, in general the entire page consists of printed information. Handwritten pages are not present in the dataset.

2.5.3 Analysis of Form Classification Methods

Although the International Association for Pattern Recognition Technical Committee 11 (Reading Systems) collected databases for Layout Analysis (PRImA) or Mixed Content Documents (Tobacco 800) amongst others, no database for form classification is provided. One (not freely) available database which can be used for form classification is the NIST tax forms database [127]

(e.g. used by Saund [152]). It consists of 12 different United States tax forms available as binarized images (no gray value images available) and a total of 11,185 images.

On the NIST database the best classification result of 100% is achieved by Saund [152] using rectilinear line art and construction of a graph lattice (see Section 2.3.3). Several authors [19, 48, 61, 115, 130] use individual databases and tested different applications of form classification. Ohtera and Horiuchi [130] used 10 Japanese commercial forms with unique ids and registration markers for a FaxOCR system. He et al. [61] has only 3 different kinds of form document, whereas Mandal et al. [115] has 40 different form types (total database of 400 forms) and Liolios et al. [107] has 26 different form types (total database consists of 2284 forms with a subset of 1200 forms taken from the NIST database). Byun et al. [19] uses 6 different form models and has an overall database size of 246 forms. Fan et al. [48] uses 30 different kinds of forms and for testing they generate a synthetic dataset by applying shift, rotation and scaling to the form documents. Additionally the forms used within the experiments are selected to show the specific application of the presented methods. Exemplarily, Mandal et al. [115] emphasized on similar form documents, Ohtera and Horiuchi [130] on form documents used within a FaxcOCR system and e.g. Fan et al. [48] on geometric distortions and also the deletions of lines (fragmented line information). For a detailed description of the results see Section 2.3.

State of the art methods show that the use of features based on the line information (junctions) are stable and can be used by hierarchical and BOW approaches for form classification. Saund [152] also states that "*prior work emphasizes sophisticated learning algorithm applied to simple, easy-to-compute features*".

2.6 Summary

State of the art methods for three DIA methods are presented: binarization, skew estimation and form classification. All presented methods are categorized in subgroups based on the methodology. Additionally, state of the art evaluation metrics are presented and discussed in Section 2.4. The presented metrics are used in Chapter 3 for the evaluation of the proposed methods and comparison with state of the art. Finally, existing methods are presented and analyzed in Section 2.5.1. This comprises also a summary of scientific databases for the presented DIA topics. For the binarization and skew estimation the results of current contests are shown.

CHAPTER 3

Methodology

This chapter describes the methodology for the DIA preprocessing steps binarization, skew estimation and form classification and retrieval. The research is focused on methods for documents with sparsely inscribed information and low contrast (faded out text, historical documents). The main contribution of the binarization is the use of a scale space approach to eliminate parameters like stroke width which is used by current state of the art methods (see Chapter 2) and to emphasize text regions. The skew estimation uses the gray value information of the input image to achieve accurate results compared to methods using binarized images. The form classification introduces shape features for the classification and retrieval task which are stable compared to the description of defined junctions. The evaluation of each method is following the description of the methodology.

3.1 Binarization

Current state of the art binarization methods (see Section 2.1) analyze the gradients of the gray value image to detect edge information, but "*go beyond locally adaptive thresholding* [...] and perform a modelling of the ink and background classes to enable more accurate individual pixel classification" [70]. Su [168] uses the local maximum and minimum to define the contrast and e.g. Howe [69, 70] uses an edge detection (Laplacian of the image intensity, Canny edge detector) in combination with a global energy function. To classify the foreground additional properties like the global stroke width [168] and the background [56] using efficient thresholds are estimated. DIBCO have also shown a trend in combining different thresholding methods in combination with machine learning and exploiting different properties of binarization methods (e.g. over- vs. under-segmentation).

In the context of text recognition based on segmentation Sayre has formulated the paradox that "*a word cannot be segmented before being recognized and cannot be recognized before being segmented*" [156]. Fischer et al. [51] calls this paradox a "chicken-and-egg" problem since on the one hand "*letters in a word are connected and cannot be segmented reliably before*

recognition, and on the other hand, the recognition of the letters requires a segmentation" [51]. Maronze et al. [117] separates the paradox in *localization guiding recognition* and *recognition guiding localization* and state that DIA systems must "*be able to incorporate both kinds of information (structural and content-based) during processing*" [117]. The Sayre paradox applies also for document binarization, if *segmented* is interpreted as the detection of the foreground (i.e. text). State of the art binarization methods must be able to handle differences in the contrast of text, thus correctly determining faded out text. Contrarily, noise and structures like the ruling of a page must be suppressed. Thus, foreground regions must be recognized as text. This leads also to the necessary definition of foreground in the context of document binarization. Within the DIBCO's the GT defines only text as foreground. Structures like the ruling of a page, bleed-through text, noise like ink stains have to be suppressed in the final binarized image. Within this thesis any inscribed or printed content on writing material (paper, parchment, papyrus) is defined as foreground for the binarization classification problem. In general, additional information like pre-printed lines (ruling of a paper) can be reliably determined [34, 83, 183] (see also Section 3.3.1) and broken characters (i.e. ascenders or descenders) restored [1, 183] in the preprocessing step of a DIA system.

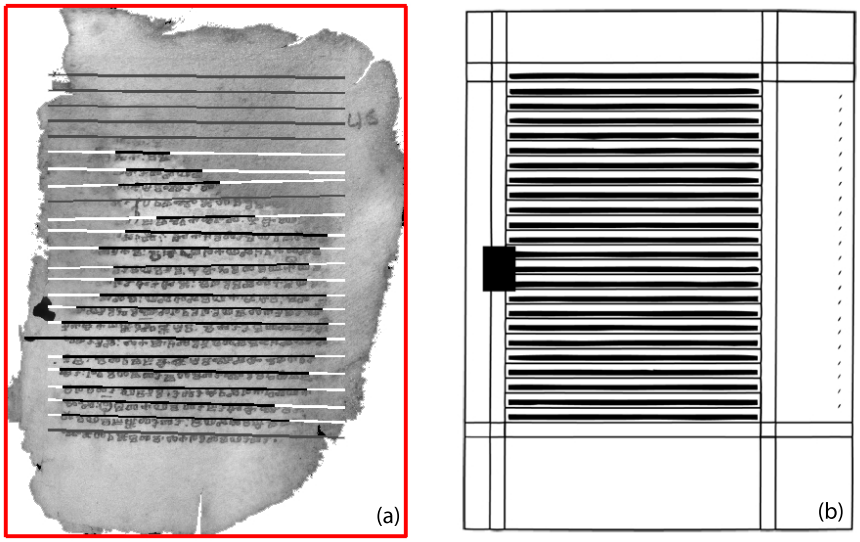


Figure 3.1: Ruling estimation of the Missale (Sacramentarium) Sinaiticum, folio 45 recto.

To resolve the Sayre paradox in the context of binarization an estimation of the text/foreground region must be performed. Thus "*localization is guiding recognition*" and allows to emphasize low contrast text without additionally boosting noise. The localization of text regions can be performed using a-priori knowledge of the layout [89], text detection (i.e. text confidence maps [4, 133] with different feature sets like Histogram of Oriented Gradients (HOG) [39], Local Ternary Patterns (LTP) [172] or MACELBP [4]), or the proposed method within the scale-space approach (see Section 3.1.2).

Within the project *The Sinaitic Glagolitic Sacramentary (Euchologium) Fragments* (funded

by the Austrian Science Fund under grant P19608-G12) the Missale Sinaiticum (Cod. Sin. slav. 5/N, 11^{th} century) [119] belonging to the classical Old Church Slavonic (OCS) canon has been digitised. It contains handwritten Glagolitic [119] text and was found in 1975 at St. Catherine's monastery on Mt. Sinai. The ruling scheme of a manuscript is used to gain information about the scribe (hand) of the manuscript, its spatiotemporal (see e.g. Leroy for the Greek tradition [95]) origin, and for layout analysis. Figure 3.1 (b) shows the a-priori knowledge ruling scheme of the manuscript under investigation.

The folios of the manuscript are degraded due to water influence, and thus parts of the text have vanished or have a low contrast (see Figure 3.1 (a)). In Kleber et al. [89] the a-priori knowledge of the ruling scheme is exploited to estimate ruling baselines if parts of the text are lacking. Figure 3.1 (a) shows the estimated ruling, where black lines indicate the detected text lines and white/gray lines are extrapolated lines due to the a-priori information of the ruling scheme. The method is based on a text line detection which is basically shown in Figure 3.2.

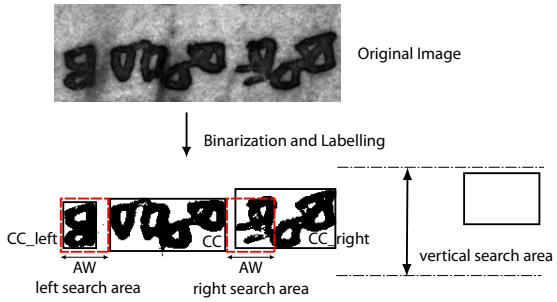


Figure 3.2: CC grouping to build text lines.

The detection of the ruling scheme (see Kleber et al. [89]) has a preprocessing stage, which comprises a skew estimation (Kleber and Sablatnig [87]), a standard image binarization, and a noise removal. After this preprocessing step the text components (words, characters, etc.) are segmented and finally grouped to extract the text lines (see Figure 3.2). Due to the knowledge of the a-priori ruling scheme a robust estimation of the text lines can be performed (see [89] for a detailed description and the results). Note that text with low contrast must not be classified as foreground by the binarization method. A readability enhancement based on a spectral and spatial analysis of the multivariate image data by MSC is done in [97] by emphasizing text regions. This is done by applying a weight mask which is derived from the ruling scheme. For a detailed description see Lettner, Kleber and Sablatnig [97].

It is shown on the example of the Missale Sinaiticum that detecting and emphasizing text regions can be reliably used for an enhancement of the readability or as a basis for binarization [89, 97]. Although a stable estimation of the text lines/text region is performed, the method is restricted to the a-priori knowledge of the ruling scheme of the Missale Sinaiticum (see Figure 3.1 (b)) and cannot be applied in a general manner. Thus a general foreground estimation is developed and used in the scale space binarization. Additionally, the method is parameter free with regard to the estimation of the stroke width which can also vary within a single document.

Section 3.1.1 describes general properties of the scale space and also the main advantages

within document binarization. The binarization method with the propagation of the foreground information over different scales is described in Section 3.1.2.

3.1.1 Scale Space

The estimation of textural properties of text in a global context leads to incorrect estimations (i.e. stroke width) if different fonts or font sizes are used (see Figure 3.5). Thus, "*the only reasonable approach is to consider representations at all scales simultaneously*" [104] by using a scale space representation. Witkin [178] discussed the blurring properties of one dimensional signals [92], illustrated in Figure 3.3, and Koenderink [90] extended the scale space to more dimensional functions. The scale space theory has also been studied by Lindeberg [104] who has also shown "*the transfer of continuous concepts to discrete algorithms*" [92].

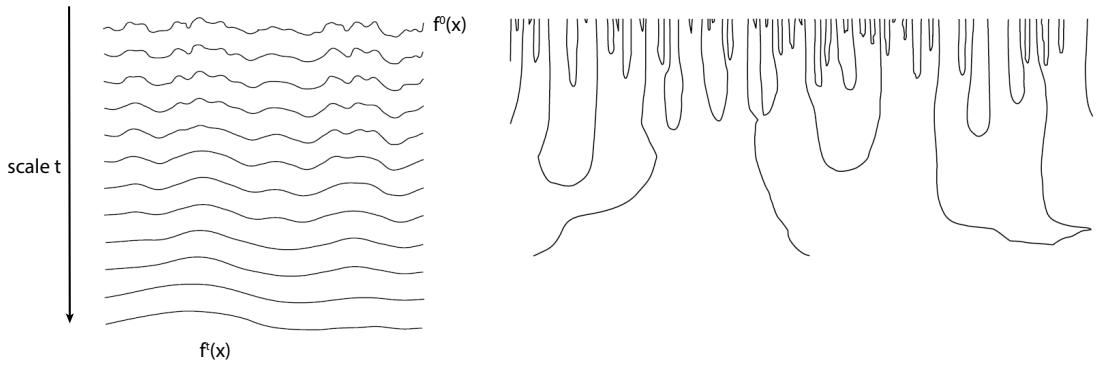


Figure 3.3: Sequence of Gaussian smoothing's of a signal $f^0(x)$ with increasing σ from top to bottom (left), zero crossings of 2nd derivative (right), courtesy of Witkin [178].

A requirement of a scale space is the successive suppression of fine scale structures at coarser scales without introducing any new structures [104]. This is also shown in Figure 3.3 where fine scale information is successively suppressed with an increasing scale. Mathematically, "*the basic idea of scale space method is to embed an image in an one-parameter family*" [189]. A scale space $L(x, y, t)$ of an image $f(x, y)$ is gained by convolving $f(x, y)$ with Gaussians $G(x, y, t)$:

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

where $*$ denotes the convolution and $t = \sigma^2$ the scale parameter. The Gaussian filter is defined by:

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

Lindeberg [103] has shown that the Gaussian filter kernel is the only low-pass filter, which satisfies the following conditions (scale space axioms):

- linear and shift/rotation invariance
- semigroup property

- continous signals ("no new local extrema or zero crossings are introduced with increasing scale parameter" [103])
- non-enhancement of local extrema (causality)

The causality property of the scale space implies that "*the intensity of maxima decrease and those of minima increase during the blurring process*" [92]. This means that faded out text (i.e. text with a low contrast) will successively disappear in coarser scales. Additionally the shape of a foreground object changes due to the scale space smoothing which concerns the problem "*how to relate structures at different scales. This subject is termed deep structure by Koenderink (1984)*" [103].

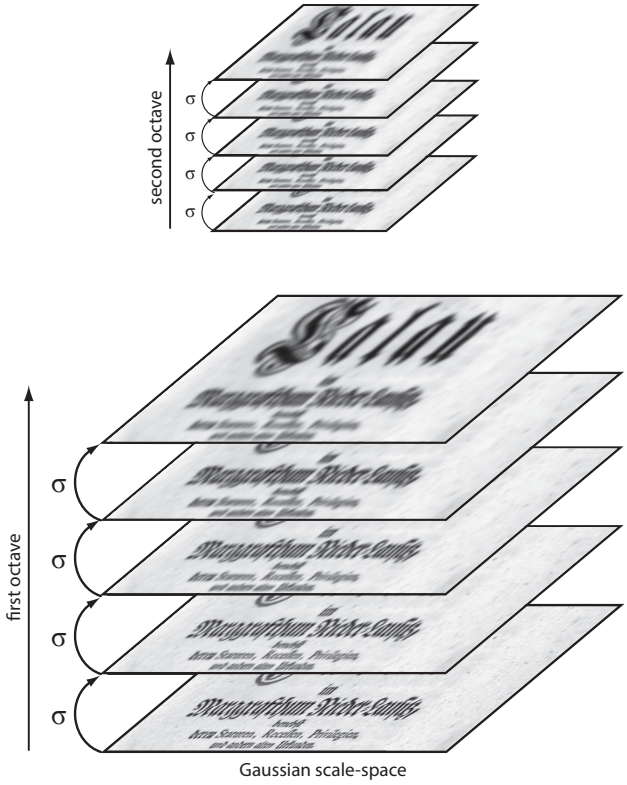


Figure 3.4: Scale space of an image of the DIBCO 2009 dataset.

Thus, structures of a coarse scale represent simplified structures of finer scale levels, which allows to use a coarse scale (see Section 3.1.2) as a foreground estimation that suppresses noise with a high frequency such as background clutter. Due to the fact that no new structures are introduced, no errors can be propagated from coarse to finer scales which allows to use the scale space for a binarization approach. Figure 3.4 shows the scale space of an image of the dataset of DIBCO 2009. The document image is successively smoothed which suppresses fine scale information. In the context of different font sizes, *small* text fuses to text lines (similar

to LPP), while the structure of *larger* text is still preserved. The results of the binarization contest in 2009 and 2010 show that algorithms using (text) stroke edge regions as candidate pixels performed best for ancient manuscripts. To avoid distortions arising from the variation of the background e.g. Gatos et al. [55] and Lu et al. [111] estimate the background for a contrast compensation. Both methods estimate parameters such as the local window size based on e.g. the mean character height or the stroke width. Problems may arise if different text sizes and stroke widths occur on the same folio/page. Figure 3.5 shows a result of the algorithm of Su et al. if a wrong stroke width is applied and the result of the proposed scale space method.

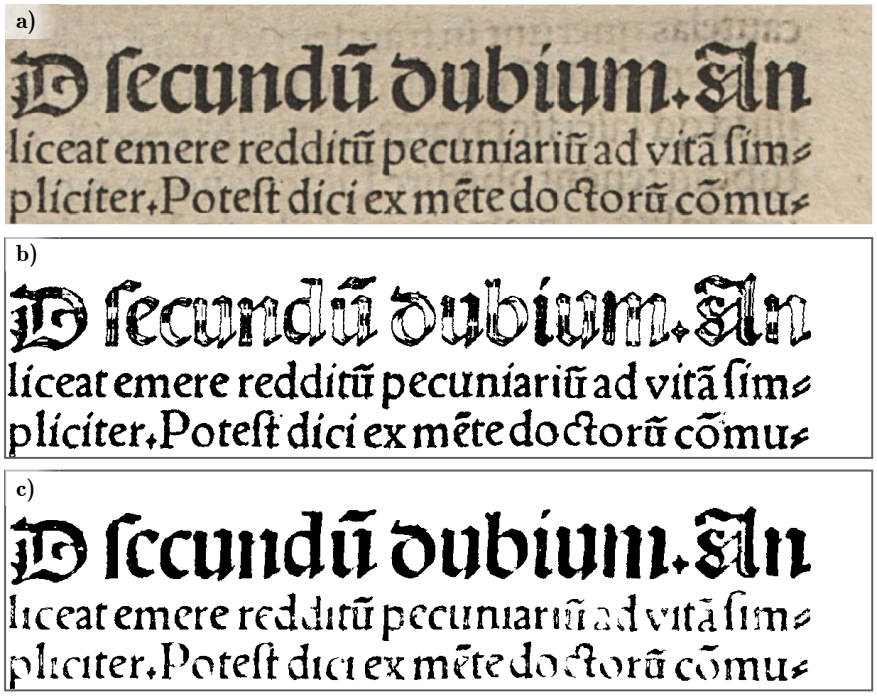


Figure 3.5: (a) Image of the DIBCO 2009 dataset (b) Su et al.'s approach (c) proposed approach

Thus, a scale space allows fixed parameters due to the different scales of the image. Propagating information from coarse to fine scales makes the algorithm independent to the text size. The continuously Gaussian smoothing of the image suppresses noise with a high variation. As a result, the coarse scales are used as a foreground estimation. The use of integral images allow an efficient implementation of the algorithm [72]. Section 3.1.2 describes the scale space binarization [84].

3.1.2 Scale Space Binarization

The basic approach of the scale space binarization [84] is illustrated in Figure 3.6. To allow for an efficient implementation of the Gaussian scale space octaves are calculated, where the scale factor σ is doubled within each octave (see Figure 3.4) and the image is resized to half of its

resolution in x and y, i.e. $1/4$, between two octaves [148]. It can be seen that only the script with a font size corresponding to the estimated stroke width is correctly binarized in the original scale. The scale space allows a binarization independent of the stroke with where the foreground information is propagated from coarser scales (corresponding to larger font sizes) to finer scales (corresponding to smaller font sizes). The fine structures of the binarized image are defined by the last scale and a foreground weighting allows to suppress noise.

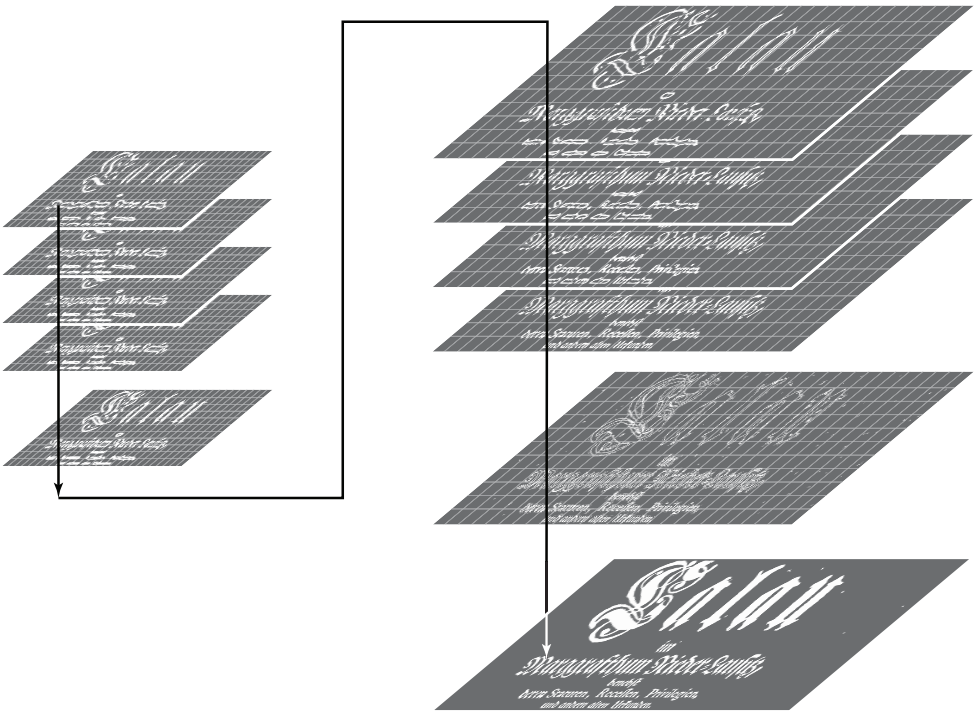


Figure 3.6: Scale information propagation for binarized images

The scale space of a document image $f(x, y)$ is constructed as proposed by Lindeberg [103]. The information of the scale space is propagated from coarse to fine scales as follows: Let $L(x, y, t_{i+1})$ (parent image) denote a binarized image of the scale space with scale t_{i+1} and $L(x, y, t_i)$ (child image) denotes the binarized image of the next finer scale. The scale parameter is $t = \sigma^2$, and it is increased by σ between two successive scales. Both images are binarized using the Su et al. approach with a constant stroke width and a constant size of the neighborhood window (5 px). Each pixel (x, y) in $L(x, y, t_i)$ is compared with the associated pixel of the parent image $L(x, y, t_{i+1})$. If the pixel (x, y) is segmented in the parent image but not in the current image a new threshold $thr(x, y)$ is applied to the pixel (x, y) in $L(x, y, t_i)$. This is illustrated in Figure 3.6. It can be seen that with the current parameters only the outer contours of the largest two fonts are classified as foreground. By propagating the foreground information from

coarse scales to fine scales, a new threshold can be applied to this regions, which is computed as follows:

In contrast to the threshold candidates of Su et al. (local areas that contain a defined number of high contrast pixel proportional to the stroke width) the threshold candidates are redefined by segmented areas $R_1 \dots R_n$ of the binarized parent image $L(x, y, t_{i+1})$. The threshold thr of each area $R_i, i \in \{1 \dots n\}$ is defined by

$$thr_{R_i} = E_{mean} \tag{3.3}$$

where E_{mean} is the mean of all segmented binary high contrast pixel (BinaryContrastChild, BCChild) in $L(x, y, t_i)$ with $(x, y) \in R_i$. The foreground information propagation to fine scales which allows to redefine regions as foreground can be formulated as shown in Algorithm 3.1:

1: $L = scaleSpace(img)$;
2: $i = size(L)$;
3: $parent = binarizeSu(L(i-1))$;
4: **for** $k = size(L) - 2 \rightarrow 1$ **do**
5: $child = binarizeSu(L(k))$;
6: **for all** (x, y) such that $parent(x, y) = 255$ **and** $child(x, y) = 0$ **do**
7: $thr = mean((parent$ **and** $BCChild) \cdot L(k), R)$;
8: $child = L(k) < thr$;
9: **end for**
10: **for all** (x, y) such that $parent(x, y) = 0$ **and** $child(x, y) = 255$ **do**
11: $weightImg = L(k-2)$;
12: $normalize(weightImg)$;
13: $invert(weightImg)$;
14: $thr = thr \cdot weightImg$;
15: $child = L(k) < thr(x, y)$;
16: **end for**
17: $parent = child$
18: **end for**

Algorithm 3.1: PseudoCode of the Scale-Space Binarization.

Pixels that are segmented in $L(x, y, t_{i+1})$ and not in the finer scale belong to homogeneous regions, which are not segmented in the current scale if a wrong stroke width is estimated. Figure 3.7 shows an image and the binarized images from two subsequent scales with the effect described.

Due to properties of the scale space (see Section 3.1.1) noise cannot be introduced in coarse scales. Hence, noise is not propagated through the scale space. Even areas with a dark background are not segmented due to the threshold which is calculated by the pixels defined in the binary high contrast image of $L(x, y, t)$. Pixels that are segmented in the finer scale and not in the coarse scale belong to finer structures. Noise with a high frequency is therefore detected in fine scales, thus with a small scale parameter t ($= \sigma^2$). To suppress noise in fine scales a foreground estimation is performed using the calculated scale space. The scale must be chosen

Figure 3.7: (a) current scale (b) parent image (c) current segmentation (d) scale space binarization

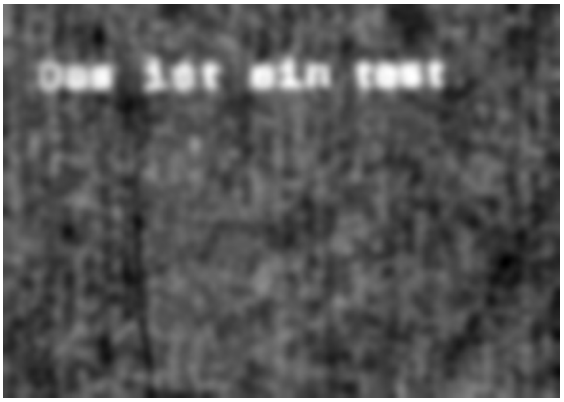


Figure 3.8: weight image (foreground estimation)

according to the smallest font size to avoid a negative weighting of text regions. The proposed approach uses the image at scale $L(x, y, t_{i-2})$ (see Algorithm 3.1, line 11-14) as foreground estimation. If a pixel is segmented in the current scale and not in the parent image the current threshold as defined by Su et al. is weighted with the foreground image. Figure 3.8 shows the foreground estimation of the image at the last scale. It can be seen that background noise is suppressed (background is a homogeneous area).

3.1.3 Results of Scale Space Binarization

The proposed method is tested on the DIBCO and H-DIBCO datasets (2009-2012). The number and type of images for each DIBCO are summarized in Table 3.1. The test images contain handwritten and printed text of historic documents and are provided together with the GT by the DIBCO organizers. The GT of the dataset is produced by an semi-automatic procedure, which calculates the skeleton of the image binarized by an adaptive method, followed by a dilation which is limited by the edges of strokes detected by the Canny-Edge Detector (see Section 2.1 [125]).

	printed	handwritten
DIBCO 2009	5	5
H-DIBCO 2010	-	10
DIBCO 2011	8	8
H-DIBCO 2012	-	14

Table 3.1: Number and type of DIBCO and H-DIBCO evaluation images.

In addition to the DIBCO images the method is evaluated on 3 synthetic images with noise and varying text size and a real world image of a carbon copy with printed text to illustrate the behavior on Gaussian noise. The GT of the synthetic images is defined by the original image without noise, and the GT of the carbon copy is defined manually. For the evaluation of the scale space binarization the metrics of DIBCO, which are also presented in Section 2.4.1, are used: FM, P-FM, PSNR and DRD. The evaluation measures for a series of pictures are accumulated pixel wise. Thus, the e.g. FM is based on a precision/recall, where TP, FN and FP are accumulated over all images. An imagewise evaluation can be done by averaging the measures of each single image. However, using an imagewise evaluation disregards the size of the single images if no weighting scheme is applied. As a result, the pixel based evaluation is used.

Method	F-Measure [%]	p-FM [%]	PSNR	DRD
Otsu	87.7538	87.4363	13.361	25.2177
Sauvola and Pietikainen	68.4178	74.5194	8.5726	71.049
Su et al.	33.87	54.2703	6.4699	114.1268
Fabrizio and Marcotegui	81.2771	81.227	10.7229	44.6971
Wolf	67.3515	90.0042	11.0195	40.3919
proposed approach	96.8046	95.1937	25.9444	3.1428

Table 3.2: Binarization results of synthetic image database, pixel wise evaluation.

Figure 3.9 shows an image of a carbon copy, a synthetic image with background noise, the results of the binarization of the proposed approach and the result of Su et al. It can be seen that

Method	F-Measure [%]	p-FM [%]	PSNR	DRD
Otsu	83.0921	83.516	23.2997	125.4465
Sauvola and Pietikainen	72.2764	74.7438	10.8566	192.6858
Su et al.	58.3462	67.6551	9.1918	266.0979
Fabrizio and Marcotegui	82.1963	82.7101	21.8788	232.2692
Wolf	73.667	84.9727	14.2552	59.6656
proposed approach	90.6425	92.2599	26.2545	4.274

Table 3.3: Binarization results of synthetic image database, image wise evaluation.

the background noise in both images is suppressed due to weighting with the foreground estimation of the proposed approach. The synthetic image (see Figure 3.7) contain a black rectangle, which is correctly binarized by the proposed approach. The binarization of Su et al. binarize only the contour of the rectangle due to the estimation of the stroke width. Table 3.2 shows the result of the proposed approach on the synthetic image database compared to the methods of Otsu, Sauvola and Pietikainen, Fabrizio and Marcotegui and Wolf [53]. The proposed approach has an FM of 96.80% compared to Wolf with a FM of 90.00%, which is the second best method. It is shown that the proposed approach can handle noise with a high frequency. Table 3.3 shows the binarization results of the synthetic image database, if the evaluation is done image wise. The results differ due to the different size of the images. Table 3.4 shows the results of the scale space binarization on the DIBCO and H-DIBCO datasets (2009-2012). The performance of state of the art methods is summarized in Section 2.5.1. The performance on the single DIBCO dataset arises from the foreground estimation, where the weighting with the foreground let thin strokes disappear in the last scale.

Method	F-Measure [%]	p-FM [%]	PSNR	DRD
proposed approach DIBCO 2009	83.3893	89.5267	15.9599	6.0474
proposed approach H-DIBCO 2010	81.006	84.8692	16.7148	5.0401
proposed approach DIBCO 2011	78.9627	84.7054	14.6399	7.7138
proposed approach H-DIBCO 2012	83.6823	86.9363	17.0644	5.7949

Table 3.4: Binarization results of the proposed approach - DIBCO

Figure 3.10 shows the errors of the proposed approach on an example image of the DIBCO 2009. The wrong segmentation arises from the edge of the background variation, which leads to lower scores for the DIBCO dataset. However, the shown result is the final result of the binarization without any post-processing step. Thus, if a blob analysis is performed as proposed by [35] the foreground objects can be classified into text and noise as a post-processing step to improve the final results.

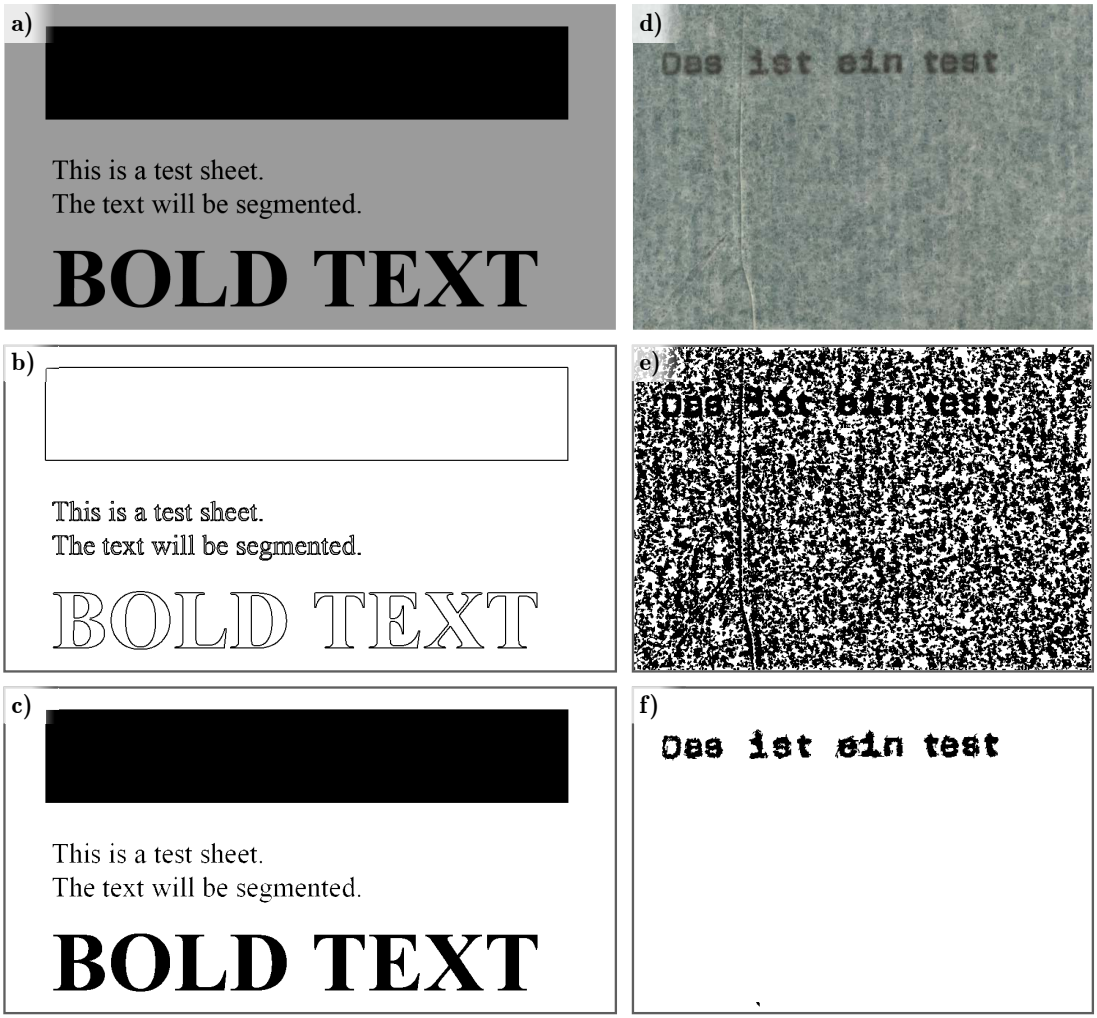


Figure 3.9: (a) synthetic test image (b) Su et al. (c) proposed method (d) carbon copy (e) Su et al. (f) proposed method

Table 3.4 shows the results on the DIBCO datasets (2009-1012). Although, e.g. Su et al. perform best at DIBCO 2009 the proposed method performs best on the DIBCO 2009 dataset combined with the dataset containing a carbon copy with background noise and synthetic images with background noise. The results of Otsu, Sauvola and Pietikainen, Fabrizio and Marcotegui, Wolf and the proposed scale space binarization are summarized in Table 3.5 for the combined dataset. The scale space binarization has a FM of 89.95% compared to a FM of 86.22% of Fabrizio and Marcotegui, which is the second best method.

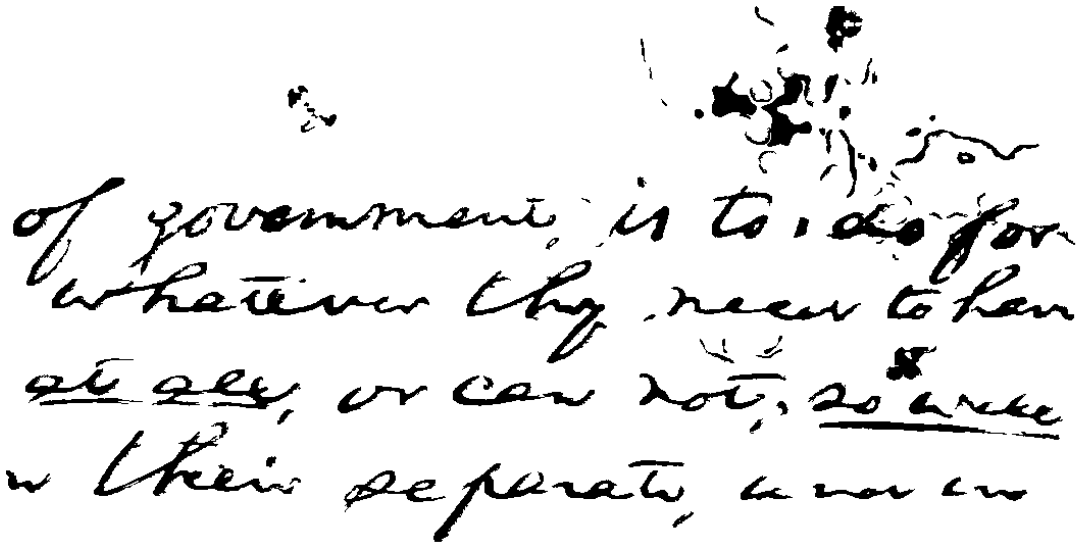


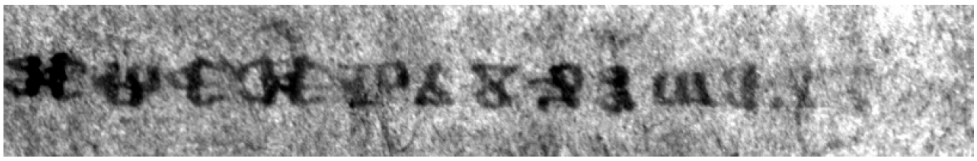
Figure 3.10: result of the proposed approach of an image of the DIBCO 2009 dataset

Method	F-Measure [%]	p-FM [%]	PSNR	DRD
Otsu	78.2989	79.0125	12.2918	20.8401
Sauvola and Pietikainen	50.5003	52.062	6.9969	72.1111
Su et al.	67.8986	82.1502	11.7685	24.0166
Fabrizio and Marcotegui	86.2281	85.9866	14.8533	11.309
Wolf	81.793	91.0154	15.0446	10.0333
proposed approach	89.959	91.455	17.1053	5.4921

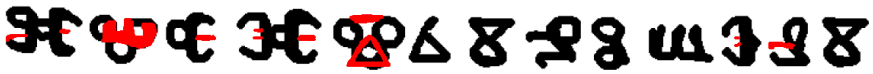
Table 3.5: Binarization results of DIBCO 2009 and synthetic image database

3.1.4 Summary and Critical Reflection of the proposed Binarization

The proposed binarization uses a Gaussian scale space to avoid the estimation of parameters like the stroke width. It is shown that due to the different scales, text with different stroke widths can be reliably binarized. The method has been evaluated on the DIBCO and H-DIBCO datasets and with the use of the evaluation metrics defined by DIBCO and H-DIBCO. The main contribution of the proposed method is to show that parameters like the stroke width can be avoided by introducing a scale space. It is also shown by the combination of different test sets, that an e.g. single image can significantly change the results. This means that the size of evaluation databases for binarization must be enlarged. As future work, a classification of text regions must be involved without the a-priori knowledge of the documents layout, which can improve the binarization result. Additionally, the a-priori knowledge of the script texture must be



a) Detail of a historical Manuscript



b) GT manually tagged by an individual

Figure 3.11: (a) Detail of an ancient manuscript (b) manually tagged GT by an individual with errors (red)

incorporated into the binarization. This can be exemplified by Figure 3.11, which shows a detail of an ancient manuscript with Glagolitic characters [88]. The GT has been labeled by individuals who have no knowledge about the alphabet. Due to the high degradation and the fact that the individuals consider only the gray value, errors are introduced into the characters (red markers, Figure 3.11 b). If the text is labeled by a Slavonic expert who has a meta knowledge about the alphabet, the character is correctly labeled. Thus, binarization methods have to consider the content of document images. The evaluation shows also the impact of a pixel based evaluation compared to an image wise evaluation.

3.2 Skew Estimation of Sparsely Inscribed Documents

State of the art skew estimation methods are performed on binarized document images (see Section 2.2). Also the latest DISEC 2013 used binary document images as input images. The use of binarized images has the drawback that errors can be introduced during this preprocessing step. Binarization methods may introduce additional objects (blobs) due to noise or classify foreground as background which results in missing text parts. Binarization as a preprocessing step in DIA is still an open research topic, which is shown in Section 2.1.

Figure 3.12 shows a detail of a form document and the binarization effect. Lines are present in technical sketches, form documents and also in *plain* text documents as underlines and can be used for a stable estimation of the skew. The binarization effect shows that lines appear as a close approximation to straight lines due to rendering effects (see Bresenham [18]). An example of the effects of a binarized character is shown in Figure 1.6 (see Section 1.3) which introduces an error of $2.16\,^{\circ}$. The proposed method [87] exploits the gray value information for the calculation of interest points, which represent foreground objects at different scales, and the gradient information. The clustering of the defined interest points are used for a stable estimation of the global text alignment to avoid errors which can be caused due to slanted text. The gradient based method accurately detects the main orientation and is robust with respect to sparse document content. The combination of the two methodologies with different characteristics allows

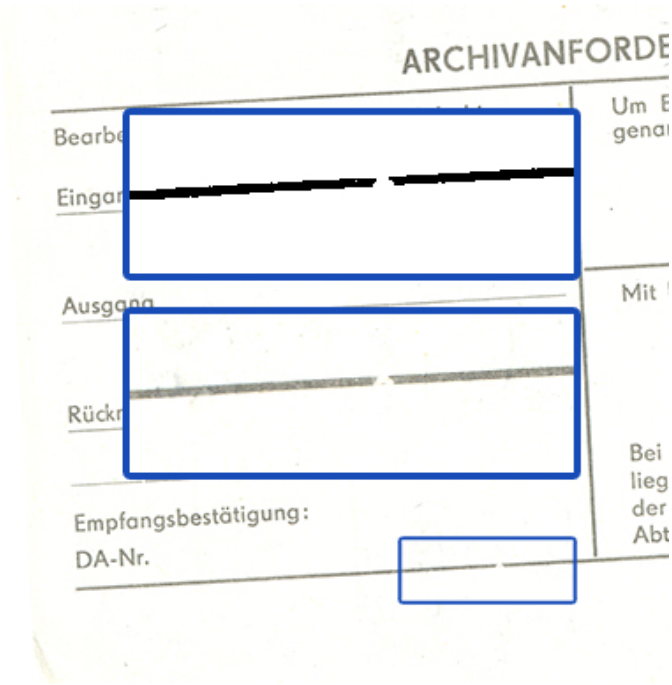
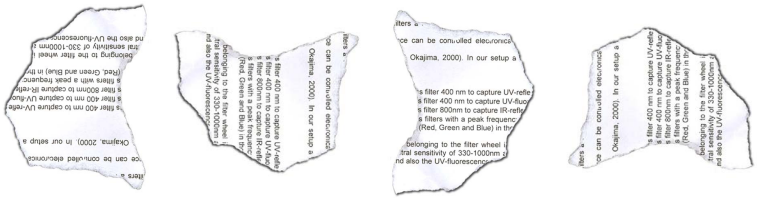


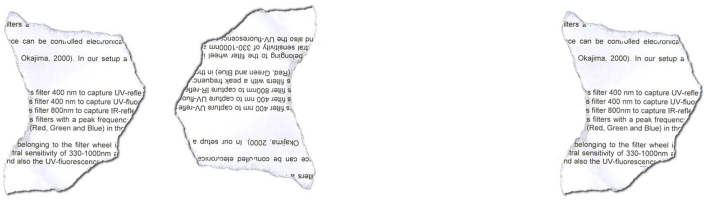
Figure 3.12: Detail of a form document with exemplified binarization effects.

a stable skew estimation for sparsely inscribed documents as well as a different document content like handwritten or printed text (within DISEC only printed pages are present). Methods based on gradients are presented by Sun and Si [170] and Omar et al. [132]. In contrast, the FNNC [76] is robust with respect to slanted handwritten text and is able to detect the angle up to $180\,^{\circ}$. The final upside/down decision is based on the statistical analyzes of ascenders and descenders (see Section 3.2.4). Figure 3.13 summarizes the possible orientations after the single methods and their combinations.

Since state of the art skew estimation methods are designed to determine a global skew angle, documents with curled text lines and different orientations are not treated. Section 2.2 uses the definition of Chen [27] who defines a global skew as a documents "*dominant (most frequently occurring) text baseline direction*" [27]. By definition, the baseline direction is the direction of single words and thus characters, which can be seen as a composition of strokes in mainly two directions: horizontal and vertical for printed text and horizontal and a slanted direction for handwritten text (dependent on the font). Based upon this fact the orientation of sparsely inscribed documents (e.g. one word) can be robustly estimated up to $\pm 90\,^{\circ}$. The correction of the slant can be done on statistic measures based on the gradient histogram. However, results show that the FNNC clustering of interest points is more stable and the combination of both methods solve also the $\pm 90\,^{\circ}$ decision. The next two Sections 3.2.1 and 3.2.2 detail the characteristics of both methods and Section 3.2.3 explains the combination.



Possible orientations after the gradient orientation method



Possible orientations after the combination with FNNC Final Orientation after the Upside/down decision

Figure 3.13: Possible orientations of the single skew estimation methods and their combinations.

3.2.1 Gradient Orientation Measure

The gradient orientation estimation is a pixel based method. Its key concept is that script comprises mostly vertical or horizontal strokes. This assumption can be verified if solely printed text is considered. However, for handwritten text with a slant, the modal angle corresponds to the slant and not to the text line angle. Nevertheless, this methodology has the advantages of considering additional information like ruling lines or underlines, provides an accurate angle estimation (median error $< 0.08\,°$; PRIMA 2009 dataset, see Section 3.2.6) and can also deal with document fragments or document images with less than two words content (which is defined as sparsely inscribed), compared to the second method.

Figure 3.14 shows 2 sparsely inscribed (2 up to 4 words) document images/fragments with printed and handwritten text. The document image with printed text (Figure 3.14 b) is taken from the synthetic test set of Epshtein [44] where a Gaussian blur and noise is added. The second example (Figure 3.14 a) shows a document fragment with slanted handwritten text. Epshtein [44] states that state of the art skew estimation methods have to deal with no restriction on the detectable angle range and must also deal with sparsely inscribed documents as exemplarily shown in Figure 3.14. The limitations of the method based on the gradient vector is detailed in the following paragraphs.

The only preprocessing of the gray value image is a smoothing with a Gaussian kernel ($\sigma = 12$, see Table 3.7) to suppress the gradient information of noise and background clutter. Compared to a binarization a Gaussian smoothing cannot introduce any new structures. The

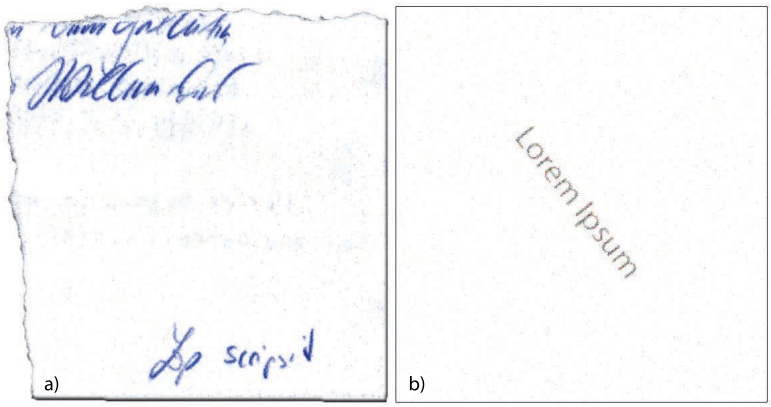


Figure 3.14: a) Synthetic test set image comprising only 2 words with Gaussian blur and noise added [44] b) Document fragment (4 words and 2 words horizontally cut).

gradient vectors of the image $I(x, y)$ are computed by:

$$I_x(x, y) = I(x + 1, y) - I(x - 1, y) \tag{3.4}$$

$$I_y(x, y) = I(x, y + 1) - I(x, y - 1) \tag{3.5}$$

$$m(x, y) = \sqrt{I_x(x, y)^2 + I_y(x, y)^2} \tag{3.6}$$

$$\theta(x, y) = \tan^{-1} \frac{I_y(x, y)}{I_x(x, y)} \tag{3.7}$$

where $I(x, y)$ represents the image, $m(x, y)$ denotes the gradient magnitude of a pixel (x, y) and $\theta(x, y)$ is the gradient vector's angle. The gradient vectors are illustrated in Figure 3.15 which shows a detail (two characters) of a document page with printed text. It is also shown, that the information of the gradients are robust with respect to noise due to the preprocessing which performs a Gaussian smoothing.

Figure 3.15 illustrates that printed Latin characters are assembled of horizontal and vertical strokes resulting in two main orientations regarding the distribution of the gradient orientation. To make a directional gradient analysis, all gradients are accumulated into an orientation histogram $H_n(\phi)$ that consists of n bins representing the global skew domain ϕ $[-\frac{\pi}{2}\ \frac{\pi}{2})$. The number of bins n define the angle resolution $r = n/180$. A too high angle resolution regarding the orientation histogram $H_n(\phi)$ will result in a *wider* distribution. A resolution of $1\,^\circ$ ($n = 180$ bins) is chosen, since experiments have shown that the distribution of the inscribed contents leads to an unambiguous peak. The other two quadrants are neglected since a gradient difference of π stands for a black-to-white instead of a white-to-black transition. The orientation histogram $H(\phi)$ is weighted by accumulating the gradient magnitude $m(x, y)$ of each pixel to the bin that corresponds with the pixel's angle $\theta(x, y)$. Thus, pixels with a low gradient magnitude (weak edge) are weighted less than those having strong edges. In addition, numerical artifacts and noise is reduced if the gradients are linearly interpolated. The documents main orientation is then defined as the maximum of the orientation histogram.

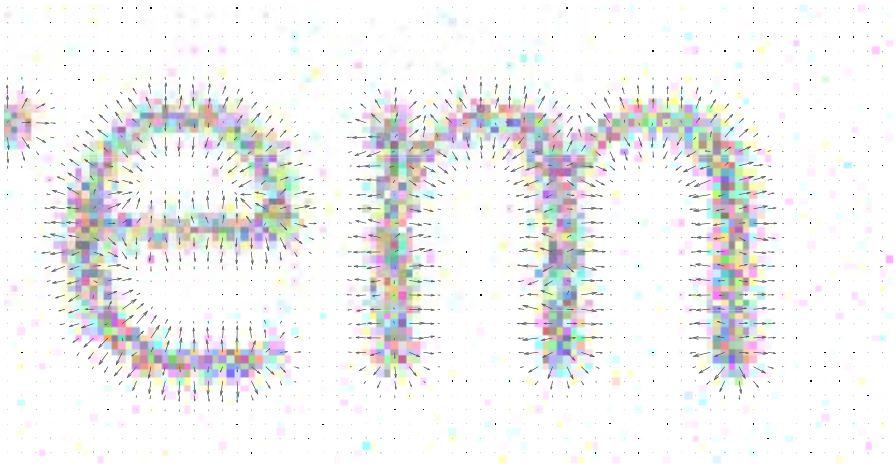


Figure 3.15: Gradient vectors of noisy text. For illustration a vector is assigned to every 2nd pixel.

To show the accuracy of the gradient orientation measure, the method is evaluated on a synthetic dataset comprising a line, which is rotated from $-90\,^\circ$ to $+90\,^\circ$ with steps of $0.1\,^\circ$ and smoothed with different Gaussian kernels, referred as line dataset. Figure 3.16 shows exemplarily an image of the synthetic generated line which has been rotated by $-34\,^\circ$ and smoothed with a Gaussian with $\sigma = 12$ (see Table 3.7. The second image shows the gradient magnitude $m(x, y)$ and the third image the gradient orientation $\theta(x, y)$. For the final angle, the median of all angles of the highest bin in the orientation histogram is chosen. Table 3.6 shows the mean,

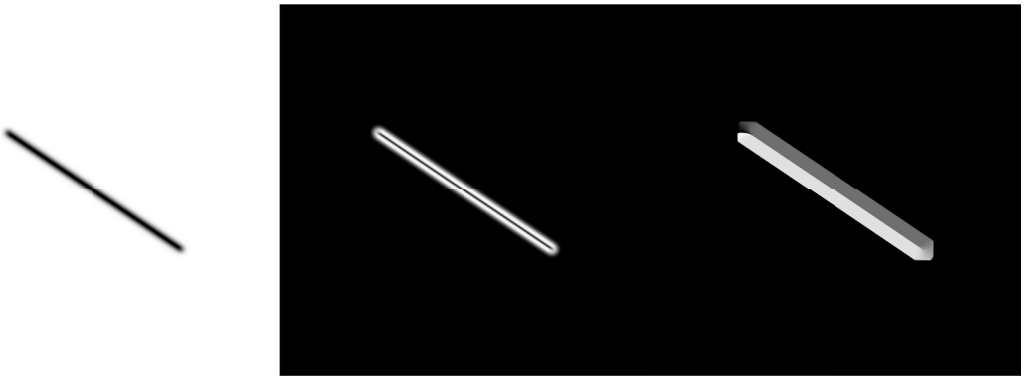


Figure 3.16: Line with an orientation of $-34\,^\circ$ which is smoothed with a $\sigma = 12$. The second image shows $m(x, y)$ and the third image illustrates $\theta(x, y)$.

the median error and the CE of the line dataset. It is shown that the mean and median error decreases with a higher σ from $0.12\,^\circ$ ($\sigma = 2$) to $0.007\,^\circ$ ($\sigma = 14$). Thus, the gradient orientation measure achieves accuracies, which are higher than $0.008\,^\circ$ (mean error) for a σ higher than 12. Figure 3.17 shows the angular error of the line dataset over the entire angle range for

sigma	2	4	6	8	10	12	14
mean error [°]	0.12	0.05	0.01	0.01	0.01	0.008	0.007
median error [°]	0.11	0.05	0.01	0.007	0.006	0.006	0.005
CE [%]	42.47	99.88	99.88	99.88	99.88	99.88	99.88

Table 3.6: Gradient orientation measure results on the line dataset.

$\sigma = 12$. Two errors are accumulated: the first on has a periodic pattern with a frequency of $1\,^\circ$. It is introduced since the final angle is chosen as the median value of all angles of the highest bin in the orientation histogram. Thus, the smallest error occurs at the exact angle related to the corresponding bin. Angles in the range between two bins have a higher error because of the uneven distribution of the angles between two bins (see also Figure 3.20). The second error with a frequency of $45\,^\circ$ is introduced due to aliasing effects. To show the accuracy of the gradient

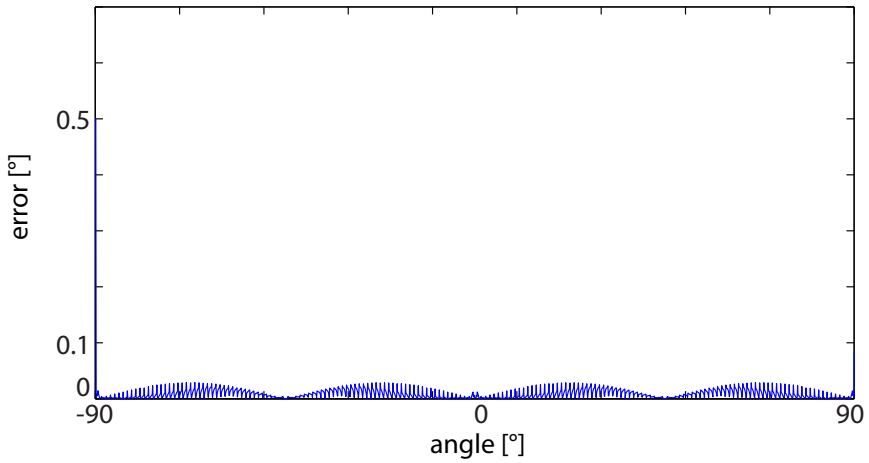


Figure 3.17: Angular error of the line dataset ($\sigma = 12$).

orientation measure for printed text documents and that the same errors apply for printed text, a synthetic test set comprising a text page of the PRIMA 2009 dataset is rotated from $-90\,^\circ$ to $+90\,^\circ$ with steps of $0.1\,^\circ$ and smoothed with different Gaussian kernels, referred as single page dataset. Figure 3.18 shows exemplarily an image of the synthetic rotated document page (rotation angle is $-34\,^\circ$) which is smoothed with a Gaussian with $\sigma = 12$. The second image shows the gradient magnitude $m(x, y)$ and the third image the gradient orientation $\theta(x, y)$. Table 3.7 summarizes the gradient error of the single page dataset for different σ. The mean and the median error decreases with a higher σ which is the same effect shown on the line dataset. However, the error is higher for the same σ due to the distribution of the angles. The gradient orientation histogram of a page of the single page dataset skewed with $-34\,^\circ$ is shown in Figure 3.19. The highest peak is at the correct bin, corresponding to $-34\,^\circ$. By determining the main orientation by choosing the median value of all angles of the highest bin, the same angular errors occur as shown within the line dataset. The angular error of the single page dataset is visualized

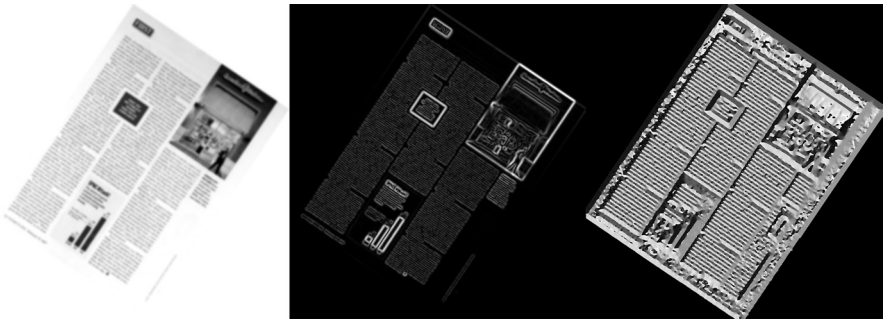


Figure 3.18: Document page with an orientation of $-34\,^\circ$ which is smoothed with a $\sigma = 12$. The second image shows $m(x, y)$ and the third image illustrates $\theta(x, y)$.

sigma	2	4	6	8	10	12	14	16	18
mean error $[^\circ]$	0.46	0.25	0.25	0.20	0.17	0.14	0.13	0.11	0.10
median error $[^\circ]$	0.40	0.21	0.20	0.17	0.15	0.13	0.12	0.10	0.09
CE $[\%]$	10.82	21.37	25.37	29.42	31.81	38.42	41.69	48.75	51.69

Table 3.7: Gradient orientation measure results on the single page dataset, varying σ.

in Figure 3.20, where also a detail of the angular error at $-34\,^\circ$ is shown. Again, the minimal angular error appears at the angular value which exactly corresponds to the bin due to the distribution of the angles. Figure 3.20 shows also the distribution of the highest bin and its neighbors

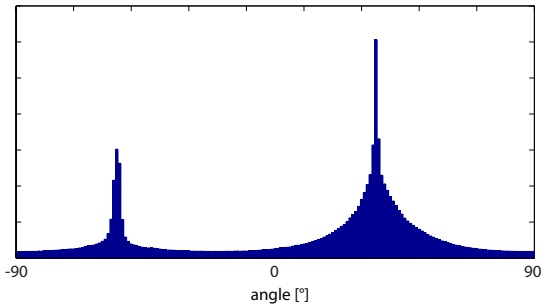


Figure 3.19: Gradient orientation histogram of a document page with an orientation of $-34\,^\circ$ ($\sigma = 12$).

of the correct angle $-34\,^\circ$ and the angular neighborhood. For all further tests a σ value of 12 is chosen as a tradeoff between accuracy and the computational effort. In order to improve the result, the distribution of the highest bin and its neighbors is taken into account. Thus, the ratio of the left neighbor bin and the highest bin is denoted by r_l and the ratio of the right neighbor bin and the highest bin is denoted by r_r. Finally the angles of the three highest bins are sorted

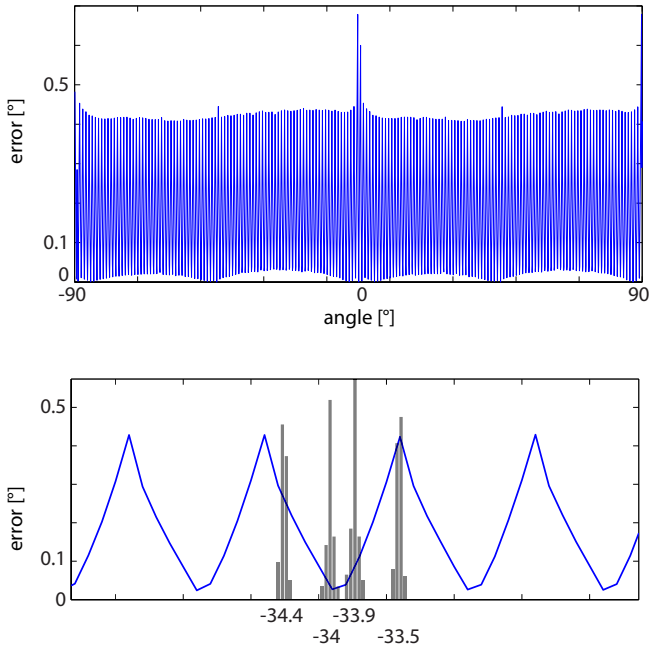


Figure 3.20: The angular error of the single page dataset and a detail of the angular error at $-34\,^{\circ}$.

	mean error $[^{\circ}]$	median error $[^{\circ}]$	CE $[\%]$
median highest bin	0.14	0.13	38.42
median 3 highest bin according r_l and r_r	0.050	0.039	94.72
spline interpolation	0.17	0.19	29.92

Table 3.8: Gradient orientation errors ($\sigma = 12$) on the single page dataset.

according their values. All angular values from the highest bin are taken into account. The number of angles from the neighborbins taken into account depends on the ratio r_r and r_l (i.e. $r_l\%$ of the angles of the left bin and $r_r\%$ of the angles of the right bin). The final orientation is determined by the median of the determined angle set, which is equally distributed. This leads to a mean error of $0.05\,^{\circ}$, a median error of $0.039\,^{\circ}$ and a CE of 94.72%. Thus, the mean error is improved by $0.09\,^{\circ}$ and the median error by $0.091\,^{\circ}$. The CE improves from 38.42% to 94.72%. Table 3.8 summarizes the gradient orientation measure errors on the single page dataset for the different methods determining the final skew angle (median of the highest bin, median of the highest bin and its neighbors according r_l and r_r, spline interpolation). Previous research [82] used a second order polynomial spline sp interpolation, which is fit into the maximal bin and its neighbors:

	mean error [°]	median error [°]	CE [%]
line dataset	0.008	0.006	99.88
line dataset binarized	0.037	0.017	93.28
PRIMA 2009 dataset	0.087	0.066	68.90
PRIMA 2009 dataset	0.158	0.124	42.72

Table 3.9: Gradient orientation errors ($\sigma = 12$) on gray value and binarized datasets.

$$sp = \sum_{i=0}^{2} a_i x^i \tag{3.8}$$

The maximum peak of the spline sp is defined as global orientation, which is gained by analyzing the first derivative. An example of an orientation histogram with a detail image of the corresponding text and the spline interpolation to determine the global orientation are shown in Figure 3.21. However, Table 3.8 shows that the spline interpolation has the worst result with a mean error of $0.17\,^\circ$ on the single page dataset.

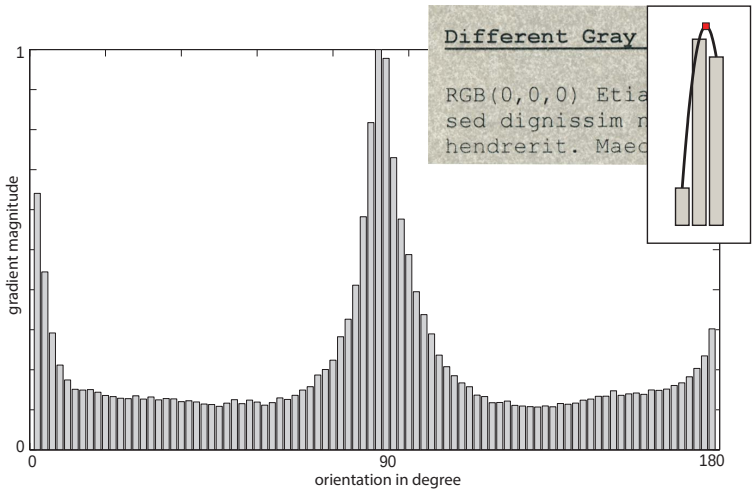


Figure 3.21: Orientation histogram of printed text with spline interpolation.

Table 3.9 shows that the gradient orientation measure is significantly influenced, if binarized datasets are used due to the loss of the gray value information.

To demonstrate the performance of the gradient orientation measure, the method has been tested on all single characters (Times New roman, 12 pts, 300 dpi, upper and lower case) of the alphabet. Table 3.10 shows the results for the alphabet printed in lower and upper case. The maximal error of $43\,^\circ$ and $39.9\,^\circ$ occur for the characters *o* and *x*. Without regarding these 2 characters the mean error for the lower case alphabet is $0.2\,^\circ$. For the upper case alphabet,

	mean error [°]	median error [°]	CE [%]
alphabet, lower case	3.39	0.02	57.69
alphabet, upper case	1.41	$0.4 \cdot^{1} 0^{-5}$	65.38

Table 3.10: Gradient orientation errors ($\sigma = 12$) on single characters of the alphabet.

the maximal error of $10\,^{\circ}$ and $21\,^{\circ}$ occur for the characters *C* and *L*. If the 2 characters are not considered, the mean error of the upper case alphabet is $0.16\,^{\circ}$. The results of the gradient orientation measure on images of single characters, show the ability to determine the correct orientation for sparsely inscribed documents.

It can be seen in Figure 3.22 that horizontal lines (e.g. ruling lines, horizontal ruling, or lines from tables) enhance the corresponding peaks for vertical gradients. A skew estimation based on the direction of the gradients has also been proposed by Sun [170] and Sauvola and Pietikainen [154] (see Section 2.2). The methods of Sun and Sauvole and Pietikainen restrict the skew angle (like the restriction in DISEC) which restricts also the search space for the maximum peak in the histogram. A skew estimation with no limitation of the global skew angle has to analyze the full range of the orientation histogram. Peaks that are not representing the main orientation of a document can be introduced due to a straight torn border (with an arbitrary angle) of a document fragment or document content like slanted text. Figure 3.22 shows an

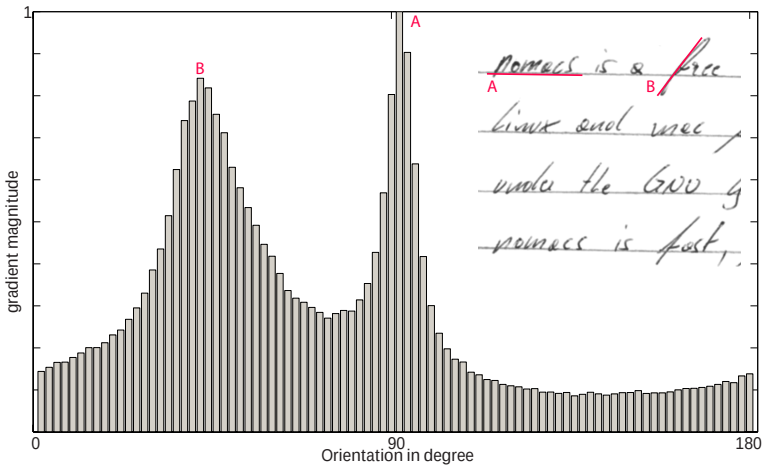


Figure 3.22: Orientation histogram of slanted text.

orientation histogram of handwritten slanted text on a lined sheet of paper. Two main peaks are visible in the orientation histogram: the peak A with an orientation of appr. $90\,^{\circ}$ corresponding to horizontal text strokes and the horizontal ruling, and peak B with an orientation of appr. $135\,^{\circ}$ corresponding to the slant of the handwritten text. The correct maximum in this example is given by peak A which is *supported* by the horizontal ruling. On a blank sheet of paper (no preprinted ruling) peak B is dominating which can lead to wrong results. Additionally, if no restriction

on the skew angle is considered, the gradient orientation method cannot reliably determine the orientations quadrant since also in the ideal case 2 peaks (horizontal and vertical strokes) are present in the orientation histogram (see Figure 3.21). Thus the skew detection is restricted to $[0\ \pi/2]$.

Previous research [33, 34, 82] used the analysis of the orientation histogram and a quadrant estimation to provide a $360\,^\circ$ skew estimation. The highest peak of the orientation histogram relative to its neighbors is taken into account rather than the global maximum. For this purpose, the histogram is smoothed with a Gaussian ($\sigma = 3$) which discards small local maxima. Then the local maxima $lm(x_i)$ are detected in the smoothed histogram and again allocated in the original histogram. Finally, the peak is detected by:

$$p = \max(lm(x_i) - median[lm(x_{i-j})\ lm(x_{i+j})]) \tag{3.9}$$

where $l(x_i)$ is the i-th local maximum, p is the resulting peak and j determines the interval of the local neighborhood. The quadrant estimation in [34, 82] is based on a blob analysis of the binarized and smoothed image. For the binarization Otsu [134] is applied in [34, 82] and the smoothing is done using LPP. For each blob the minimum area rectangle is the basis for the quadrant estimation, where each blob represents a word due to the previous smoothing. The minimum area rectangles are first rotated relative to the main orientation. Then they are accumulated into a w (*width*) and h (*height*) histogram depending on their angle. A weight based on the angle, size and aspect ratio is assigned to each rectangle. Thus, rectangles having a relative orientation of $45\,^\circ$ have a lower weight than those with $1\,^\circ$. If the resulting h bin is higher than the w bin, the document needs to be corrected by $90\,^\circ$.

Figure 3.23: Document fragment; the word blobs of the document fragment gained using LPP and the corresponding minimum area rectangles; the corrected document fragment.

Figure 3.23 shows a document fragment and the word blobs of the smoothed image (LPP of the binarized image). The orientation of the minimum area rectangles (red) determine the correct quadrant for the skew estimation. The parameter evaluation of the number of bins n and of the smoothing σ is done in [34, 82]. Using the gradient orientation approach with the quadrant estimation presented leads to an average error of 1.95% and a median error of 0.37% (database of 658 images) [34]. The quadrant estimation leads to an error of 7.67% (50 document

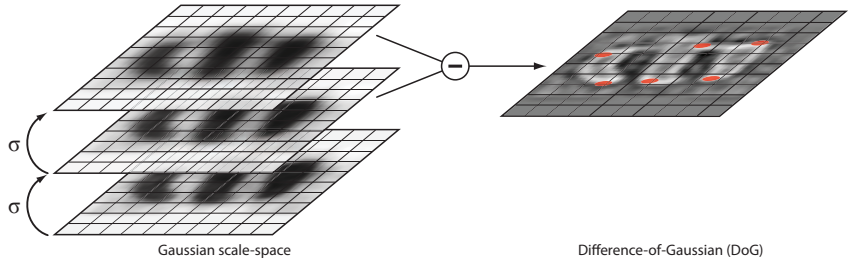


Figure 3.24: Interest point detection for the local FNNC skew lines.

fragments are not correctly rotated) as a consequence of binarization errors and errors of the relative orientation estimation (orientation of the minimum area rectangles of word blobs).

The results show that the gradient orientation estimation can be used for the skew estimation of sparsely inscribed documents due to the stroke characteristics of characters for printed and handwritten documents. The final orientation can be determined up to $90\,^{\circ}$. Due to the main drawbacks of the quadrant estimation (binarization and alignment of the minimum area rectangles) and the peak determination in the orientation histogram (e.g. slanted text) the proposed gradient orientation measure is combined with a FNNC of interest points representing *characters*. The combination of both methods avoids a binarization and a further analysis of the classified foreground.

3.2.2 Focused Nearest Neighbor Clustering

The second method presented is introduced by Jiang et al. [76]. This method focuses on the words' skew which allows for a skew estimation up to $180\,^{\circ}$. Even though it is not as accurate as the previously described method, it is more robust if handwritten documents are observed.

The FNNC is based on local skew lines that are fitted into small subsets of points. In contrast to Jiang et al. [76] the use of DoG [109] interest points for the FNNC is proposed, since they are fast to compute and more robust than centroids of CC due to the binarization. The detection of IPs by determining maxima in a DoG image of a handwritten character is illustrated in Figure 3.24. To represent roughly the centers of characters rather than junctions or corners (see Figure 3.24) a larger scale in the DoG pyramid is chosen. Experiments show that the second scale in the third octave of the Gaussian scale space is a good choice for the feature point selection. Thus, the image is resized by $\frac{1}{4}$ and subsequently smoothed with a Gaussian kernel having $\sigma = 2$. Since the interest points represent characters rather than edges or junctions the word's orientation is observed and the slant angle is not considered. As a result, the method can be applied to slanted handwritten text.

After the detection of N IPs, the k nearest neighbors $n_1, n_2, ..., n_k$ of each interest point p_i are observed, where $i \in \{1 \cdots N\}$. Each pair of neighbor points n_o, n_p with $o, p \in \{1 \cdots k\}$ and $o \neq p$ define $\binom{2}{k}$ local skew lines. The local skew line l_i of p_i is the skew line with the

smallest distance d:

$$d = \frac{|n_{l_i} * (p_i - n_o)|}{|n_{l_i}|} \tag{3.10}$$

with n_{l_i} being the normal vector of l_i and d is the distance of the connecting line to p_i. If the points n_o, n_p are close, the accuracy of the local skew line suffers from small deviations of the interest points. In order to find a robust local skew line, the longest line connecting p, n_o, n_p is chosen. Figure 3.25 shows a cluster with k=7 interest points. The gray line shows a rejected local skew line with distance d to p_i. In addition, the final local skew line is illustrated.

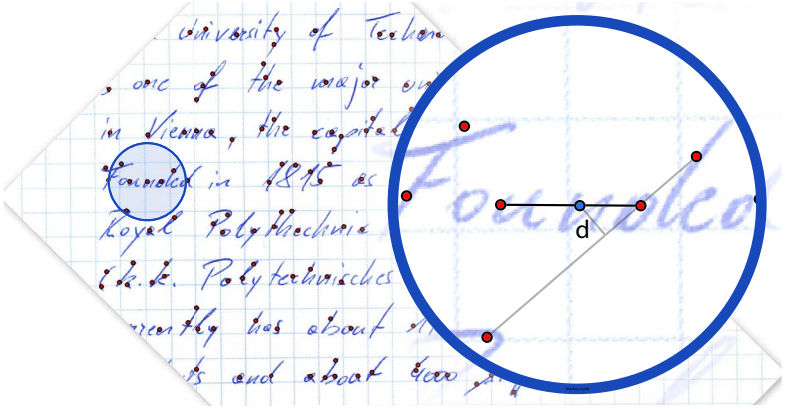


Figure 3.25: Nearest neighbors with $k = 7$, rejected local skew line having distance d and selected local skew line of the FNNC.

The document's dominant angle is determined by accumulating the local skew lines' angles to an orientation histogram. Due to the restriction to $\pm 15\,^\circ$ a size of 90 bins is chosen, with an angular resolution of $1\,^\circ$. Figure 3.26 shows the orientation histogram of an example image of the PRIMA 2009 dataset with a GT angle of $-11.73\,^\circ$. The maximum bin determines the main orientation with a resolution of $1\,^\circ$.

All local skew lines that contributed to the maximum bin and its neighbours are taken into account. The angles are sorted in an ascending order according to their distance value d. The final angle is determined by the median of the angles with the 100 smallest distances.

3.2.3 Method Combination

As shown in previous sections, the FNNC method and the gradient orientation estimation are designed for documents with different types of inscribed content. Hence, the FNNC performs better if handwritten text is supplied while the gradient method is more accurate than FNNC and considers background information.

The orientation histograms based on the FNNC and the gradient orientation are accumulated, whereas a weighting scheme is established: a lower weight is applied to the method, which is assumed to fail. Therefore, the weighting scheme is based on the knowledge of the orientation histogram shape. A method fails if the orientation histogram has an equal distribution in contrast to an orientation histogram where the gradients, respectively the line orientations

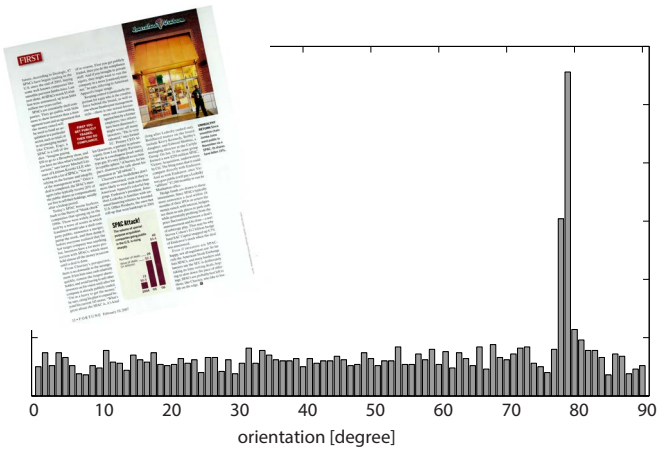


Figure 3.26: FNNC Orientation histogram of an example image of the PRIMA 2009 dataset. GT angle is $-11.73\,^{\circ}$.

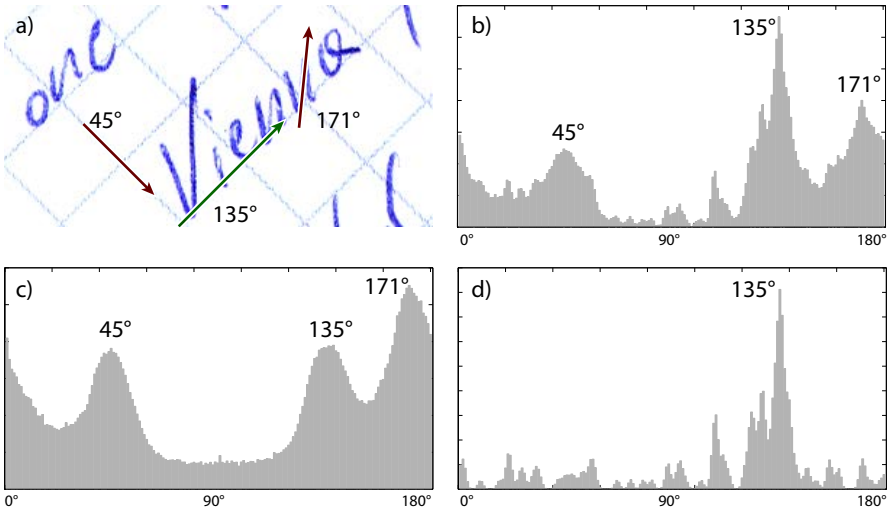


Figure 3.27: (a) Text region on checked paper (b) combined histogram (c) gradient histogram (d) FNNC histogram

of the FNNC method accumulate into the same orientation (representing one sharp peak in the histogram). Thus, an integral measure is defined as weight for each orientation histogram:

$$\omega_o = 1 - \frac{\sum H_n(x)}{n \cdot \max H_n(x)} \tag{3.11}$$

where ω_o is the weight of the orientation histogram $H_n(x)$ and n is the number of bins. If $\min H_n = \max H_n$ or if less than 5 interest points were detected in an image ω_o is set to 0 for that histogram. Figure 3.27 shows the histogram combination. It can be seen that the gradient

orientation method (c) considers the background ruling but has a higher peak resulting from the writing's slant. The FNNC method (d) completely disregards the slant and the background information due to the fact that the detected interest points belong to the handwritten text. After the weighting and the combination of the orientation histograms wrong peaks in the histogram (e.g. belonging to the slant or the ruling) are suppressed (b).

The combined method provides a skew estimation with a decision up to $180\,^{\circ}$. The final up/down decision must be performed statistical analysis of the script.

3.2.4 Up/Down Orientation

Without a text recognition the document up/down orientation can only be based on statistical features of the inscribed text. Caprari [21] and Cullen et al. [30] use the distribution of frequency of ascenders and descenders of Roman letters and Arabic numerals. Statistics of German and English text show that the occurrence of ascenders is dominating [100] (see Table 3.11). Thus,

Stroke	Letters	Frequency English	Frequency German
Descender	j,p,q,y	4.15%	1.12%
Ascender	b,d,f,h,k,l,t	27.92%	24.19%
Neither	a,c,e,i,m,n,o,r, s,u,v,w,x,z	67.93%	74.69%

Table 3.11: Character frequency in English and German [100, 174]

Openess	Letters	Frequency English
To left	a,d,j,q	11.64%
To right	b,c,e,f,g,h,k,p,r,y	37.65%
Neither	i,l,m,n,o,s,t,u,v,w, x,z	50.71%
To left	J,Q	0.53%
To right	C,E,F,G,K,L,P,R	34.07%
Neither	A,B,D,H,I,M,N,O, S,T,U,V,W,X,Y,Z	65.4%

Table 3.12: Openess of characters defined by Aradhye [8].

after the skew correction based on the combination of the gradient orientation and the FNNC method, a line histogram is analyzed to determine the ascender/descender frequency. Since the line histogram is sensitive to the correct skew, Caprari [21] divides the entire page into stripes. It is analyzed that the best results are gained when a document is divided into 6 stripes. The orientation cannot be estimated reliably in the case of block letters or if other scripts and languages are used [8].

Aradhye [8] propose a method which analyses the *openess* of characters. He distinguishes between characters that are open to the left and to the right (see Table 3.12). Although the method proposed by Aradhye [8] can determine the orientation of uppercase letters (block letters) a blob analysis of single characters must be performed. Thus the method cannot be applied to handwritten and printed text. As a result the orientation of ascenders and descenders is analyzed for the final upside/down decision.

3.2.5 Skew Correction by Line and Paragraph Analysis

The determined angle of the combined approach can be corrected by applying a line and text paragraph analysis. Thus, the image is rotated according the final angle of the combined approach and Local Projection Profiles (LPP) [13] are applied. The resulting image can be interpreted as a smoothed image, where single characters are connected to text lines.

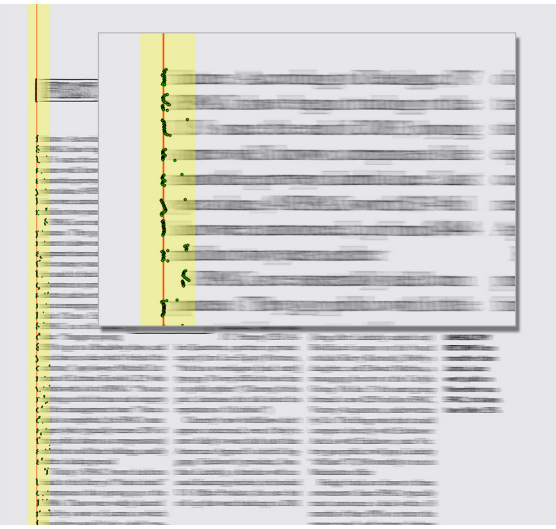


Figure 3.28: Paragraph lines based on the LPP to refine the FNNC angle.

A horizontal differential operator is applied to the LPP image to determine *left* edge points (transition from background to foreground). A vertical sliding window with a width of 100 pixels (experimentally determined) is shifted over the entire image, and a line is fitted to all points determined by the horizontal differential operator within the sliding window. Figure 3.28 shows an example image of the PRIMA 2009 dataset, the sliding window (rectangle) at the position of the beginning of the paragraph and all points within the sliding window. A line is fitted by minimizing $\sum \rho(d_i)$, where d_i is the distance between the i^{th} point and the fitted line, and $\rho(d)$ is the Welsch distance [177]:

$$\rho(d) = \frac{C^2}{2}\left(1 - exp\left(-\left(\frac{d}{C}\right)^2\right)\right) \tag{3.12}$$

where C is a constant and defined as $C = 2.9846$. Figure 3.28 shows the line fitted of the sliding window. The orientation of the paragraph lines are normalized respectively the final FNNC angle. After the paragraph analysis, N additional *paragraph lines* l_p are detected.

In addition to the orientation of the paragraph lines fitted, the orientation of lines detected in a binarized image is done. The line detection is based on analyzing horizontal and vertical Directional Single Connected Chains (DSCC) as proposed by Zheng et al. [188]. Figure 3.29 shows the detected lines in a binarized image. Thus, after the line detection, L additional lines are detected.

Figure 3.29: Detected lines in the binarized image using DSCC.

Finally, all $N + L$ lines detected (determined by analyzing the paragraph and the line detection) with an orientation difference larger than $1\,^{\circ}$ (angular resolution of the FNNC and Gradient Measure) compared to the FNNC angle are filtered. The corrected skew is statistically determined by the median of the orientation of all remaining lines. It is shown in Section 3.2.6 that the skew correction leads to a higher accuracy for documents containing lines and printed text paragraphs

3.2.6 Results Skew Estimation

The datasets used for the evaluation have different characteristic image types: binary images, gray value images with a low resolution (appr. 70 dpi), gray value images of machine printed documents, gray value images of historical documents and gray value images of sparsely inscribed documents (fragments). The different characteristics show the behavior of the gradient orientation measure, the FNNC and the combined method for the skew estimation on this datasets. In Section 3.2.1 it is already shown that binarized images have a significant influence on the gradient orientation measure due to the loss of information. Section 3.2.6.1 describes the different datasets and Section 3.2.6.2 summarizes the results of the skew estimation.

3.2.6.1 Datasets

The method is evaluated on 6 different datasets. The first one is the original DISEC 2013 benchmark dataset, which consists of binarized images. The second test set contains the same images as the DISEC 2013 benchmark dataset as gray value images. Additional test sets are generated from the images from the PRIMA 2009 and 2011 contest, which contains on the one hand mainly printed text, and historical documents on the other. All datasets used are available at `http://caa.tuwien.ac.at/cvl/research/skew-database/`. In addition a synthetic dataset which has also been used by Epshtein [44] and a dataset which consists of 658 document fragments from the Stasi files is prepared. The datasets address skew estimation problems regarding different spatial resolutions, modern printed layouts and historical documents and sparsely inscribed documents. The characteristics of each dataset are described in detail in the following paragraphs.

DISEC 2013 Dataset The DISEC 2013 benchmark dataset [135] consists of 155 binarized images. Each image has been randomly rotated with 10 different angles resulting in 1550 images. The images "*contain various sizes of document pages, any kind of mixed content, vertical and horizontal writing, multi-sized fonts and multiple number of columns in the same document*" [135] and text in different languages is available. The skew angle is restricted to $\pm 15\,^{\circ}$.

A second test set has been generated using the gray value images from the DISEC 2013. The images have the same skew as the binarized one in the benchmark dataset, but are only available at a resolution of appr. 70 dpi. Thus, the results between the binarized images and the gray value images are not directly comparable to the results of DISEC 2013. Figure 3.30 shows an example image of the dataset illustrating the difference of the spatial resolution. However, the dataset can be used to show the robustness of a skew estimation method regarding different spatial resolutions.

PRIMA 2009 Skew Dataset The PRIMA 2009 dataset has been generated from the images from the ICDAR 2009 Page Segmentation Competition [6]. The dataset consists of 55 color images with printed English text and different layouts with multicolumn text, multi-sized fonts and embedded images (e.g. newspapers, scientific papers). Each image is randomly rotated ten times, whereas the skew angle is in between the range of $\pm 15\,^{\circ}$, comparable to DISEC 2013. Thus, the entire dataset consists of 550 images. Each image is provided with a binary mask to discard the edge information of a single page, since skew estimation methods should estimate the angle based on the content and not on the edge information of a page.

PRIMA 2011 Skew Dataset The PRIMA 2011 dataset has been generated from the images from the ICDAR 2011 Historical Document Layout Analysis Competition [5]. The dataset comprises 100 color images of historical documents. Since some pages are scanned from books, parts of text lines can be curved due to the bookbinding. Additionally, some text lines can have different orientations. To create the GT, 3 individuals have been asked to manually annotate the main orientation of each page independent from each other. The angle annotated for each page and individual is visualized in Figure 3.31. For each page, the minimum difference and

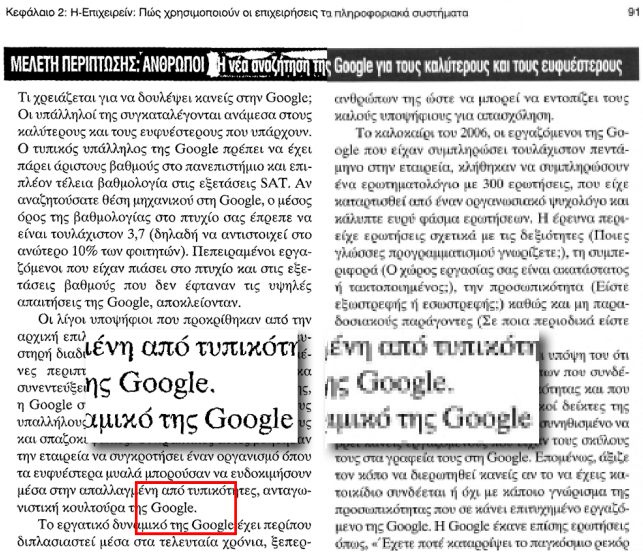


Figure 3.30: Resolution difference of the binarized images (left column) and the gray value images (right column) of DISEC 2013.

the maximum difference between the 3 manually estimated angles has been taken into account to select a subset of the entire dataset. Figure 3.32 shows the angle differences sorted for each image in an ascending order. Images where the minimum angle difference between two individuals is smaller than $0.1\,^\circ$ are selected for the PRIMA 2011 skew dataset, resulting in a test set size of 72 images. The final images are rotated with the median of the 3 annotated angles to create a GT set which is aligned (main orientation) with an accuracy of $0.1\,^\circ$. The threshold of $0.1\,^\circ$ is equivalent to the human perception regarding the skew of a text page [135]. To create the final test set, each image is randomly rotated 10 times between the range of $\pm 15\,^\circ$ (equivalent to the angle range of DISEC 2013) resulting in a final dataset size of 720 images. Similar to the PRIMA 2009 skew dataset a mask is provided for each image to discard the edge information of the cropped image. The dataset is challenging for methods using binarized images due to e.g. bleed-through text and noise. Additionally, text lines can be distorted and are not perfectly aligned compared to modern printed documents. An example image of the PRIMA 2011 skew dataset is shown in Figure 3.33.

Epshtein dataset The dataset of Epshtein consists of 8 images with horizontal printed text with a text count of 2 words up to 250. Figure 3.34 shows 3 representative example documents of the synthetic dataset, which illustrate the different layouts and text sizes used.

To simulate real world examples taken with a camera, a Gaussian blur with $\sigma = 1.5$ and then Gaussian noise $\sigma = 0.05$ is added to the images (see also Figure 3.35). The images are rotated from 0 to π with a step of 0.05 radians as suggested in [44]. It can be seen that strong edges are not present in the synthetic test images and that the proposed method must be able to deal with noisy images even if solely two words are present. Although, the synthetic test set is reproduced

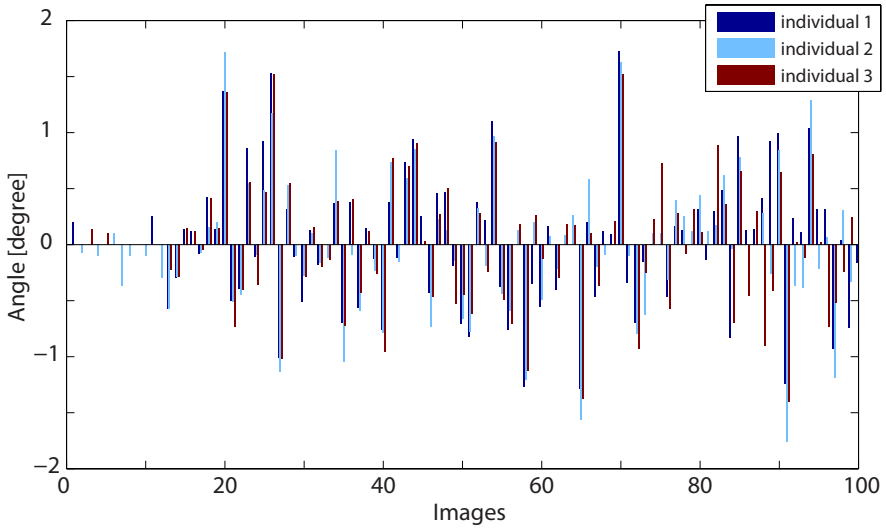


Figure 3.31: The GT-angles manually annotated by 3 individuals.

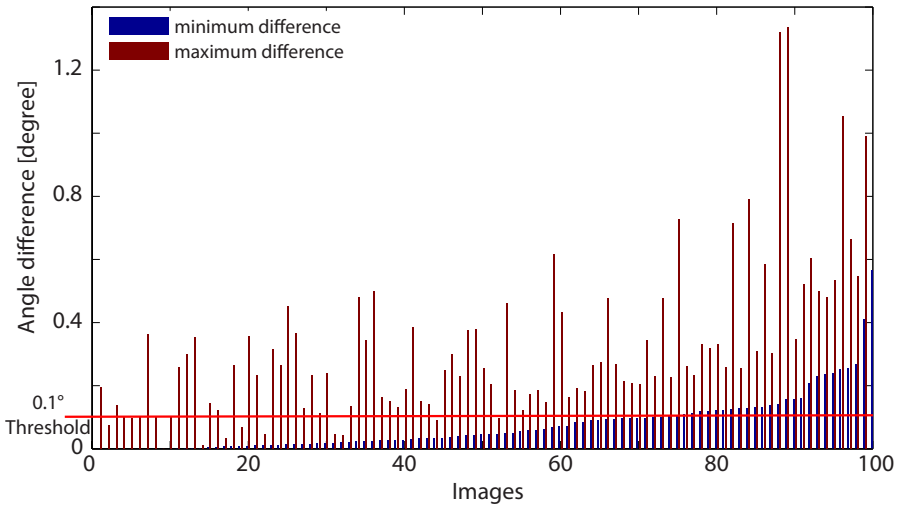


Figure 3.32: The sorted GT-angle differences of three individuals according to the minimum distance.

from Ephsteins paper, the blur and Gaussian noise parameters are unknown (not presented in the paper [44]), and thus leading to not directly comparable results. The presented noise parameters are adapted to achieve visually similar images as presented by Epshtein.

Stasi Document Fragments The Stasi documents can be categorized as historic documents. The document fragments have irregular shapes from poststamp size up to a size of a full DIN

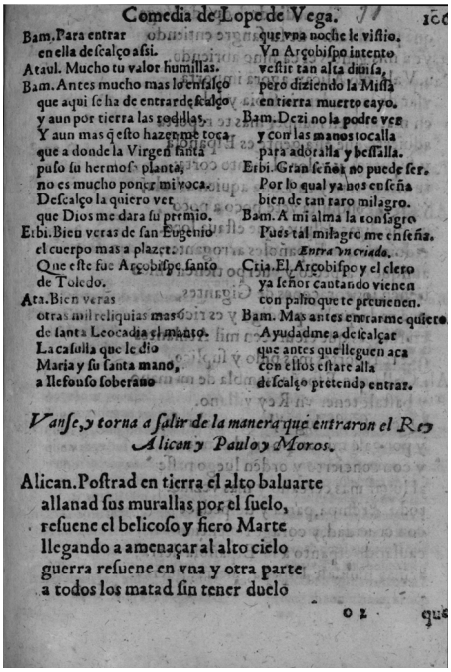


Figure 3.33: PRIMA 2011 example image with bleedthrough text .

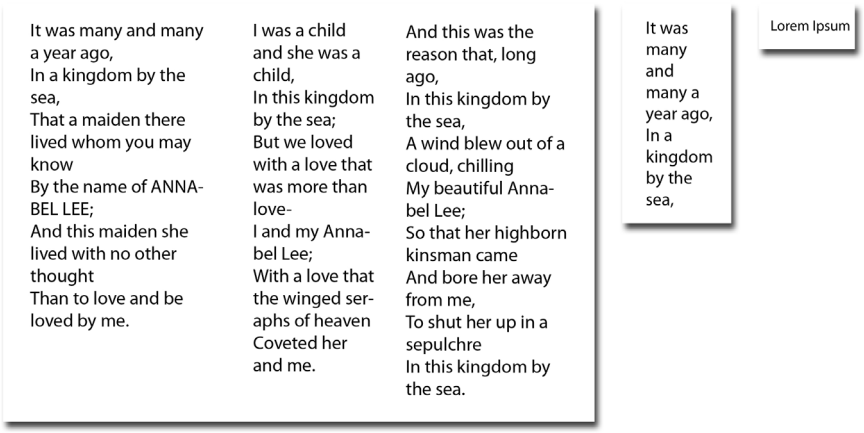


Figure 3.34: 3 Example documents of the Synthetic dataset of Epshtein [44].

A4 page, different paper types (checked, lined or blank paper) and the content ranges from two up to hundreds of handwritten or printed words. The size distribution of the fragments is shown in Figure 3.36.

Figure 3.37 shows exemplarily a well preserved fragment comparable to the dataset, since original fragments cannot be shown due to privacy reasons.

The GT of the Stasi test set has been manually tagged by defining the baseline of printed

Figure 3.35: Synthetic test image with Gaussian blur and noise added.

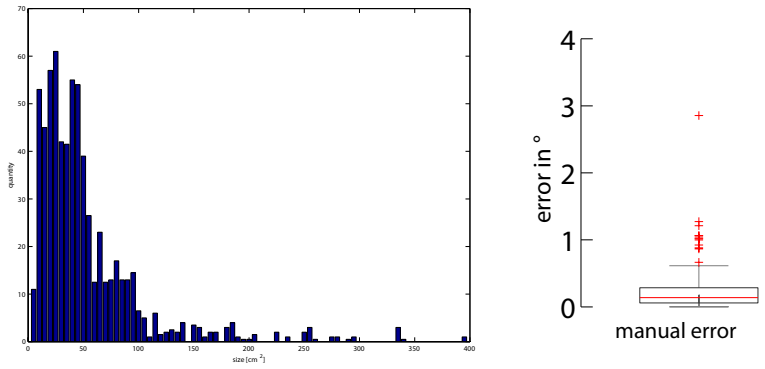


Figure 3.36: Stasi document fragments size distribution of the test set consisting of 658 snippets (left). GT Cross validation of the GT error of the manually tagged Stasi test set (right).

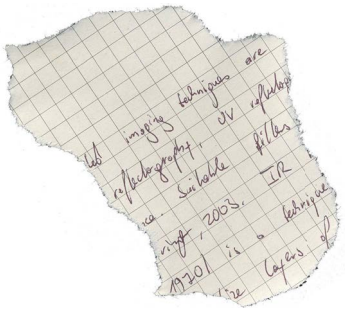


Figure 3.37: Exemplary document fragment.

text. In the case of handwritten text one global orientation is defined as GT (direction with the most words aligned). If the handwritten text is based on a ruling, the alignment of the horizontal preprinted ruling is defined as global angle. To analyze the error of the manually labeled Stasi test set, a cross validation is performed on 150 images, so that the variation of different operators can be analyzed. The resulting median error of the rotational angle, which was annotated by dragging a line in a given image corresponding to the baseline of the text/ruling, is $0.14\,^\circ$ ($q_{0.75}$: $0.29\,^\circ$). The maximal error of $2.86\,^\circ$ can be traced back to the fact that some fragments do not have an obvious main direction (e.g. if handwritten lines are not parallel to each other). Figure 3.36 (right) shows the boxplot of the cross validation to visualize the statistical error of the Stasi

	mean error [°]	median error [°]	CE [%]
gradient orientation	0.360	0.165	35.16
FNNC	0.103	0.090	65.42
combined method	0.235	0.172	34.32
combined method with skew correction	0.185	0.105	48.64
J. Fabrizio	0.072	-	7.48
H. I. Koo and N. I. Cho	0.085	-	71.23
E. Carlinet and J. Fabrizio	0.097	-	68.32

Table 3.13: Skew estimation results on the DISEC dataset ($\sigma = 12$, dataset binarized).

	mean error [°]	median error [°]	CE [%]
gradient orientation	0.087	0.066	68.90
gradient orientation, binarized dataset	0.158	0.124	42.72

Table 3.14: Skew estimation results on the PRIMA 2009 dataset ($\sigma = 12$).

test set. The following sections describe the test sets in more detail and present the results.

3.2.6.2 Results of Datasets with an angle restriction

The DISEC, PRIMA 2009, PRIMA 2011 dataset have an angle restriction of $\pm 15\,^\circ$. Table 3.13 shows the results of the FNNC approach, the gradient orientation measure and the combined method with the skew angle correction on the DISEC dataset. The result of the FNNC is the method submitted to the DISEC (rank 5). Additionally, the results of the first 3 ranks of the DISEC are shown. The FNNC method outperforms the gradient orientation measure due to the fact that only binary information is available. Thus, the combined method which uses the angles of the gradient orientation method performs with a mean error of $0.235\,^\circ$ compared to a mean error of $0.103\,^\circ$ of the FNNC. Table 3.14 shows the difference of the gradient orientation measure on gray value images compared to binarized images (using Su et al.) on the PRIMA 2009 dataset. It is shown that the gradient orientation measure has a high accuracy with a mean error of $0.08\,^\circ$ compared to a mean error of $0.15\,^\circ$ on the binarized version.

The results of the gray value images from the DISEC contest with a low resolution of approximately 70dpi are shown in Table 3.15. On *low* resolution images, the FNNC approach outperforms the gradient orientation measure. The proposed method outperforms also the method of H.I. Koo and N.I. Cho [135]. In comparison to the DISEC dataset, the gradient orientation measure has better results on the PRIMA 2009 dataset. Additionally the skew correction reduces the mean error from $0.08\,^\circ$ to $0.05\,^\circ$. The CE of the combined method is 83.09% compared to 68% (FNNC and gradient orientation measure). The reason for the enhancement of the CE is based on the characteristic of the PRIMA 2009 dataset which contains printed documents with lines and text columns with a *perfect* alignment. The results on the dataset comprising historical documents is shown in Table 3.17. The evaluation shows that the gradient orientation method

	mean error [°]	median error [°]	CE [%]
gradient orientation	0.464	0.329	20.70
FNNC	0.234	0.180	31.74
combined method	0.356	0.289	22.83
combined method with skew correction	0.323	0.253	23.09
H. I. Koo and N. I. Cho	0.334	0.200	28.70

Table 3.15: Skew estimation results on the DISEC dataset ($\sigma = 12$, gray value 70 dpi).

	mean error [°]	median error [°]	CE [%]
gradient orientation	0.087	0.066	68.90
FNNC	0.083	0.068	68.72
combined method	0.109	0.085	58.36
combined method with skew correction	0.057	0.030	83.09

Table 3.16: Skew estimation results on the PRIMA 2009 dataset ($\sigma = 12$).

	mean error [°]	median error [°]	CE [%]
gradient orientation	0.251	0.081	54.16
FNNC	0.180	0.149	32.91
combined method	0.211	0.112	47.50
combined method with skew correction	0.200	0.129	42.77

Table 3.17: Skew estimation results on the PRIMA 2011 dataset ($\sigma = 12$).

outperforms the FNNC. The combined method is less accurate due to the presence of lines which are not perfectly aligned with the text orientation. The results show the ability to use the proposed approach also on historical (handwritten) documents which can introduce binarization errors.

3.2.6.3 Results of Datasets without angle restriction

The proposed method can also be applied to datasets without an angle restriction. The quadrant decision ($90\,^\circ$ interval) is based on the FNNC. A dataset which consists of a single random page of the PRIMA 2009 dataset has been generated by rotating the page from $-90\,^\circ$ to $+90\,^\circ$ with steps of $0.1\,^\circ$. The results are shown in Table 3.18. The combination of the FNNC and the gradient orientation measure can estimate the skew up to the up/down decision of a page ($180\,^\circ$). To be able to solve the skew estimation without any angle restriction the up/down decision is solved by analyzing the occurrence of ascenders and descenders as presented in Section 3.2.4. As stated in Section 3.2.1, the gradient orientation measure can solve the skew estimation up to $90\,^\circ$ (horizontal or vertical), which is the main orientation of the strokes of Latin characters.

	mean error [°]	median error [°]	CE [%]
gradient orientation	0.042	0.031	92.44
FNNC	0.067	0.059	75.18
combined method	0.042	0.031	92.43
combined method with skew correction	0.046	0.031	91.22

Table 3.18: Skew estimation results on a single page of the PRIMA 2009 dataset ($\sigma = 12$), rotated from $-90\,^\circ$ to $+90\,^\circ$ with steps of $0.1\,^\circ$.

Figure 3.38 shows the error of the gradient orientation measure on the PRIMA 2009 dataset. The $90\,^\circ$ errors are solved by the combination with the FNNC method. Two datasets are presented

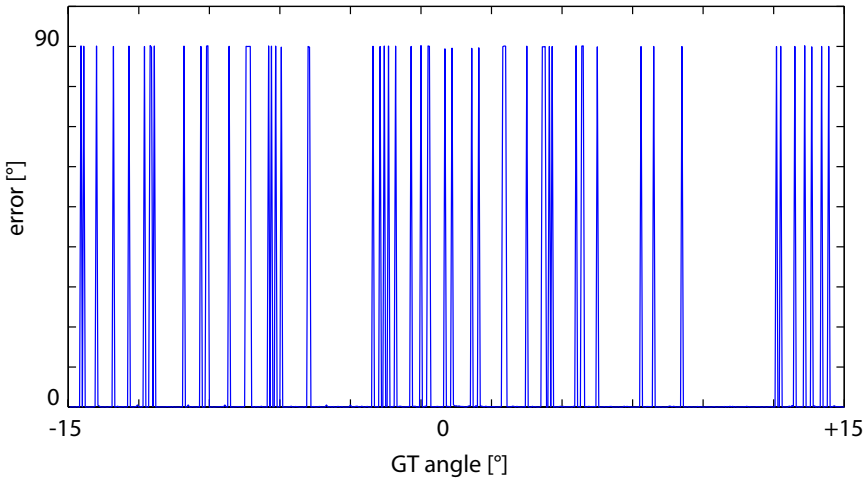


Figure 3.38: Gradient orientation measure errors.

in Section 3.2.6.1 comprising data that is rotated up to $360\,^\circ$. Table 3.19 shows the results on the synthetic dataset of Epshtein [44] which is compared to the results of Epshtein's method and the Bar-Yosef et al. method [12]. The proposed method outperforms the skew estimation presented by Bar-Yosef et al. [12]. Table 3.19 additionally shows that the proposed method has no catastrophic errors (errors $> \pi/10$) and a mean error of $0.64\,^\circ$ (median is $0.35\,^\circ$). The dataset shows also the ability of the proposed method to estimate the skew of sparsely inscribed (2 words) and noisy (no sharp edges) documents.

Table 3.20 shows the results of the proposed method and Bar-Yosef's et al. method applied to 658 document fragments. Since Bar-Yosef's method needs a binarized image, all fragments are thresholded using the Su et al. [168] method. Additionally a mask has been generated to ignore the border region of the fragments. Hence, the resulting binarization is improved. The dataset comprises also document fragments with handwritten text (see Section 3.2.6.1).

The proposed method has a mean error of $1.75\,^\circ$ and 120 catastrophic errors ("*cases where*

	median	mean	variance	catastrophic
proposed method	$0.35\,^\circ$	$0.64\,^\circ$	± 0.88	0
Bar-Yosef et al. [12]	$0.37\,^\circ$	$0.67\,^\circ$	± 2.05	2
Ephstein [44]	-	$0.497\,^\circ$	-	0

Table 3.19: Synthetic images (504), see [44]

	median	mean	variance	catastrophic
proposed method	$0.56\,^\circ$	$1.75\,^\circ$	± 4.57	120
Bar-Yosef et al. [12]	$0.62\,^\circ$	$4.24\,^\circ$	± 9.10	106

Table 3.20: Document fragments (658)

the detected text orientation differs from the ground truth by more than $\pi/10$" [44]), compared to a mean error of $4.24\,^\circ$ and 106 catastrophic errors. Catastrophic errors result from slanted text if solely a few words are present on a fragment. In these cases, the FNNC detects insufficient interest points for a statistically significant orientation estimation.

The proposed method outperforms the skew estimation presented by Bar-Yosef et al. [12]. The method cannot be compared to methods submitted to the DISEC on the presented datasets, since the datasets have no restriction on the skew angle.

3.2.7 Summary and Critical Reflection of the proposed Skew Estimation

The proposed skew estimation is a combination of the FNNC and the gradient orientation measure: The gradient orientation measure can be applied on the original gray value image without preprocessing except a Gaussian smoothing. Thus no binarization, which can introduce errors on historical document images, is needed. The use of gradients to determine the skew is also proposed by Sauvola and Pietkainen [154] and Sun and Si [170]. The main contribution of this thesis regarding the gradient orientation measure is the analysis of the accuracy depending on the Gaussian smoothing and the type of content. It has been shown that the distribution of the angles with respect to the GT angle is taken into account. Additionally, the use of binarized images has a significant influence on the result due to the loss of information. The main advantage of the gradient information is the high accuracy (if gray value images are used) and the possibility to use the measure for handwritten text and sparsely inscribed documents. It is shown that the information of single (printed characters) is sufficient for the gradient orientation measure. The FNNC method allows to estimate the skew up to $180\,^\circ$. Additionally the FNNC is robust against handwritten slanted text. Thus, the combination allows to correct the result of the gradient orientation measure by reducing the search space for a maxima in the gradient orientation histogram. It is shown that the combined method can be applied to document fragments with a high accuracy. The proposed skew correction based on the analysis of paragraphs and lines can enhance the final result based on the types of documents. It is shown that the result is improved on printed documents with an accurate alignment of text columns. If handwritten (his-

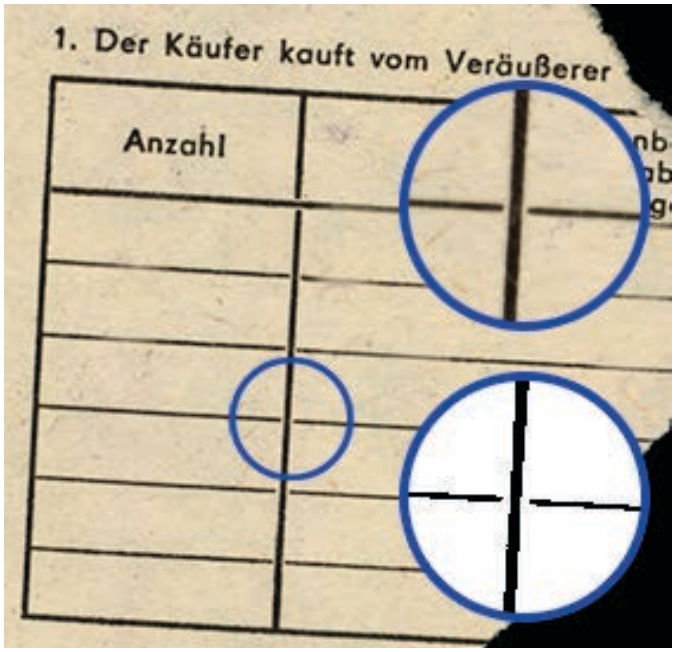


Figure 3.39: Detail of a form document. The horizontal strokes have no junction with the vertical stroke.

torical) documents are used, the final result is less accurate due to the alignment of handwritten text. The main contribution of the FNNC method is the research on handwritten documents and the robust estimation of the FNNC with the gradient orientation measure. Thus, the proposed method shows that the skew of a document can be estimated on the original gray value images without an angle restriction. If binarized images are used, the method has drawbacks due to the loss of information.

3.3 Form Classification based on Binary Information of Line Structure

The research for the form classification and retrieval is based on the binary line information of form documents. Saund [152] has shown that the count of primitive junction types as defined by Fan [48] (see Section 2.3, Figure 2.14) has not sufficient information to distinguish the different categories of the NIST tax form database. Thus, collections of primitive junction types are used to create distinctive features. It is also stated that "*effective handling of image noise and variability is a key issue in document recognition*" [152]. Within a preprocessing stage the line and junction extraction is performed with a reliability of 95%.

Figure 3.39 shows a part of a form document and a detail containing an X-junction. It can be seen that the horizontal lines are not connected to the vertical line using a global threshold (Otsu). Depending on the quality of the form document image and the preprocessing different

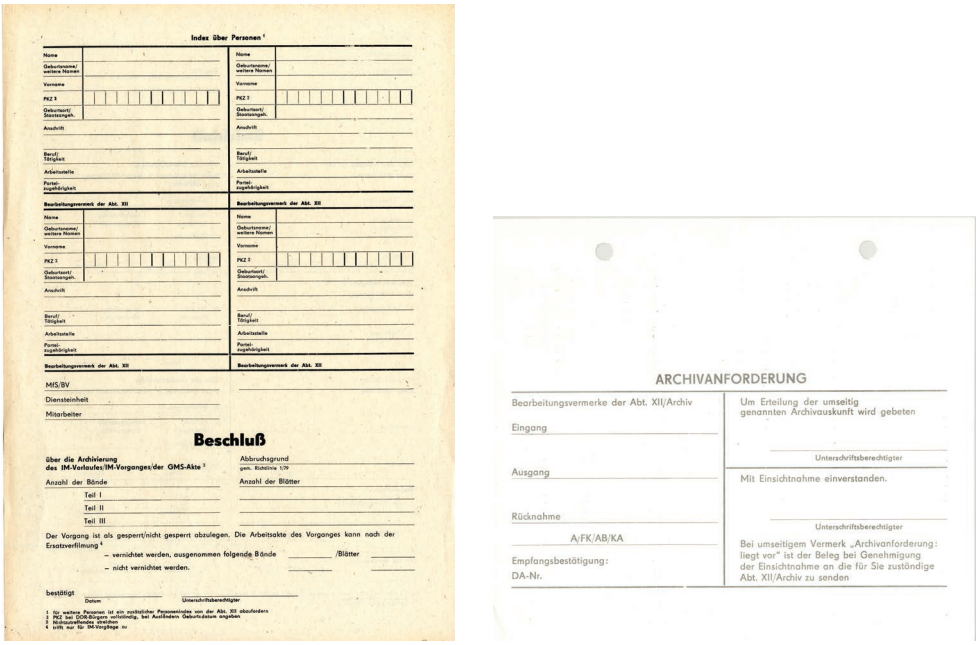


Figure 3.40: Form samples.

junction types are classified (e.g. X vs. T junction). By sampling the line information and the use of shape features of the binary line information stable features are extracted for the classification and recognition. Due to the representation of forms as histograms of line structures (shapes) form template variations can be correctly classified. Thus, forms having the same line or similiar line structure and only changes within the text cannot be distinguished. Since broken or missing lines result only in minor changes in the feature histogram, degraded documents can be correctly classified. Figure 3.40 shows exemplarily 2 forms occurring in the Stasi dataset. The size of the form ranges from approximately DIN A6 to DIN A4.

Section 3.3.1 describes the preprocessing of the documents, while Section 3.3.2 explains the shape feature extraction and the classification using a BOW approach. The proposed approach is compared to the method presented by Arlandis et al. [9] which can classify similar filled-in forms. The method is extended to train different form classes with multiple form images from the same class and is summarized in Section 3.3.4.

Dominant line structures (line endings, crossings, T-junctions, ...) of a form type are determined and represent a dictionary for each form class. Based on the dictionary, a feature histogram for a form can be calculated which allows a classification of the form type by comparing the histogram with the form-class histograms. Using different scales allows to describe local as well as global structures of forms.

3.3.1 Preprocessing

In the preprocessing step the skew is estimated using a combination of a gradient based approach and a FNNC of interest points as described in Section 3.2 and in [36]. The gradient based methodology is stable for forms with solid lines, while the FNNC is used if dotted lines or text determine the main orientation of a form. After the correction of the skew the image is binarized using the scale space approach presented in Section 3.1.

Based on the binary image a vectorisation algorithm is performed to detect lines by analyzing horizontal and vertical run lengths [34, 187]. Analyzing run lengths allows to remove handwritten and printed text and to close small gaps occurring due to ascenders and descenders of text. Additionally a pixel accurate detection of the line within the image is done by the vectorization method. In the following paragraph the line detection is summarized.

Horizontal and vertical Directional Single-Connected Chain (DSCC) are calculated according to [187]. A horizontal (vertical) DSCC consists of vertical (horizontal) neighboring pixel run lengths. The definition of a vertical pixel run length R_i according to [187] is:

$$\begin{aligned} R_i(x_i, y_{s_i}, y_{e_i}) = \{(x,y) | \forall p(x,y) = 1, x = x_i, y \in [y_{s_i}, y_{e_i}] \\ p(x_i, y_{s_i-1}) = p(x_i, y_{s_i+1}) = 0\} \end{aligned} \tag{3.13}$$

where $p(x,y)$ represents the pixel value (1 = foreground, 0 = background), x_i, y_{s_i} and y_{e_i} denote the current row, the starting and end point of the vertical run length. The definition of a horizontal DSCC is defined respectively the perpendicular direction to the vertical DSCC. Junctions (DSCCs connected to more than one pixel run length R_i) are deleted, and thus after the horizontal and vertical DSCC analysis blobs with multi-connected DSCCs are fragmented.

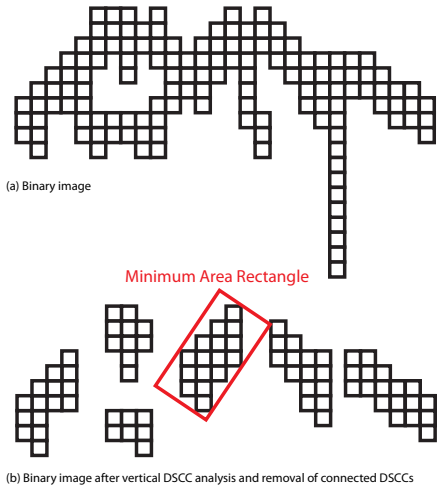


Figure 3.41: Example for a vertical DSCC analysis of an binary image

Figure 3.41 a) shows exemplified a binary image with multi-connected DSCC and the remaining fragmented lines. It is assumed that lines or parts of lines have a length that is greater

than the width of a character, the aspect ratio is less than 0.5 (lines have elongated shape), and the size (pixel area) is greater than the size of a character. To determine the aspect ratio the minimum area rectangle of a DSCC is calculated (see Figure 3.41 (b)). In addition the ripple-rating of a DSCC is calculated. The ripple rating is defined as the area of the DSCC in proportion to the area of the minimum area rectangle. The parameters are defined according the text size and validated through experiments. After removing all DSCCs that do not fullfil the described characteristics the remaining parts are merged. This is done by analyzing the orientation and the gap between two DSCCs. The orientation is defined as the angle between the axis of abscissae and the major axis of the DSCC. The gap between two line candidates is defined dynamically as $1.5 \times linelen$, where $linelen$ is the length of the longer line candidate. The deviation of the orientation of two candidates has to be less than $4°$. The thresholds $(4°, 1.5 \times linelength)$ is evaluated empirically.

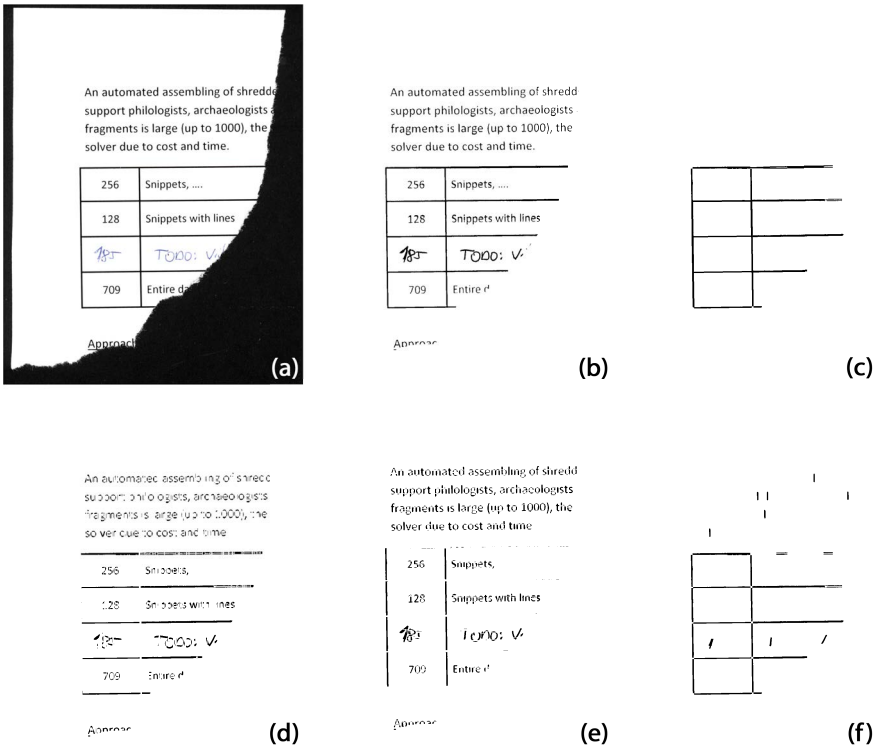


Figure 3.42: (a) Document fragment (b) segmented image (c) merged line image (d) horizontal lines (e) vertical lines (f) filtered line image

Figure 3.42 a) shows an example of a fragment with printed and handwritten text (parts of the text are underlined) and the steps of the line segmentation algorithm (b)-(f).

For the detection of dotted lines a template matching is done to determine dots. Based on the dots a nearest neighbor clustering is done for dotted line detection.

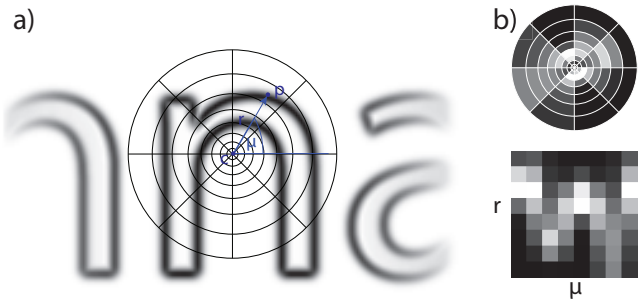


Figure 3.43: Shape Context Features proposed by Belongie et al. [16].

3.3.2 Structural Features

The structural line features to describe lines and line crossings are based on shape context features proposed by Belongie et al. [16]. The shape context features of Belongie are used in relation to object recognition for 2D and 3D objects and exemplified on e.g. handwritten digits (2D) and 3D Objects from the Columbia COIL dataset [16]. Figure 3.43 shows the shape context of point c which is defined as histogram of the distribution of neighboring points (Figure 3.43 b). Belongie et al. represent a shape by uniform sampled points of the contour.

Sample points closer to the actual shape context point have a higher influence due to a binning which is uniform in the log-polar space [16]. The robustness of shape context features is empirically evaluated. The proposed shape context of Belongie et al. is adapted and simplified for the defined structural features to describe junctions of form documents.

The detected lines in the preprocessing step are the basis for the feature computation. Two sets of lines are defined: solid lines $l_s = \{l_{s1}, \ldots, l_{si}\}$ and dotted lines $l_d = \{l_{d1}, \ldots, l_{dj}\}$, where i is the number of solid lines and j is the number of dotted lines. All solid lines l_s and all dotted lines l_d are sampled equally at a distance of ds pixel. The sampling distance ds defines the coarseness of the line structure and is set to 10 pixel (spacing distance of dotted lines). Thus, all lines $l_{s,d}$ are represented by a set of sample points $P = \{p_1, \ldots, p_k\}$, $p_i \in \mathbb{R}^2$ of k points.

For each point p_i an orientation histogram $H_i(\phi)$ is defined as line structure (shape feature):

$$NP = \{p_j \mid \|p_j - p_i\| < r\} \tag{3.14}$$

NP are defined as all Neighbor Points p_j within the radius r. The radius defines the scale of the shape and thus the geometric complexity. All neighbour points in NP are represented by their polar coordinates (L, ϕ) relative to p_i (center point of the current shape feature):

$$L_j = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2} \tag{3.15}$$

$$\phi_j = \arctan \frac{y_j - y_i}{x_j - x_i} \tag{3.16}$$

where x and y are the image coordinates of $p_{i,j}$.

Figure 3.44 shows a detail of a lined form with one vertical line. At each (sampling) point the line structure (shape) is calculated within a circular shape-region. The radius defines the

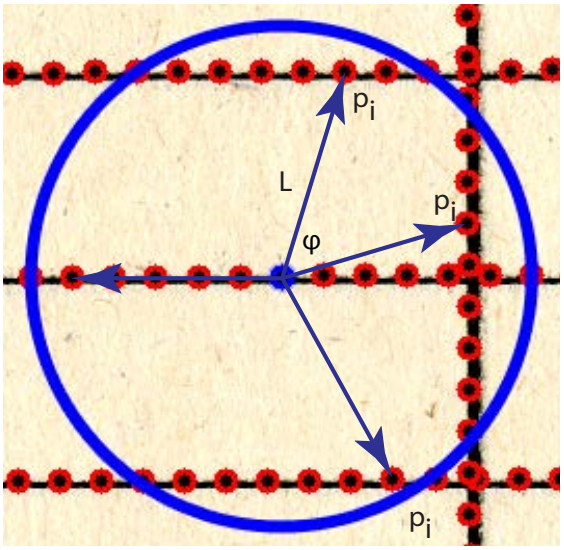


Figure 3.44: Sampled lines with a shape region with a scale of 120 pixels.

scale of the shape. The representation of a neighbor point p_j of p_i by its polar coordinates is represented in Figure 3.44. All points of NP - defined by their angle and distance relative to the shape center - are accumulated into an orientation histogram $H_i(\phi)$ which is weighted by the distance L_j of every neighbor point. The weighted orientation histogram is defined as line structure.

Closer points to the center have less influence on the shape, thus weighting the orientation by the distance leads to more stable results. If the current point p_i belongs to a dotted line l_d, the orientation histogram is weighted by negative distances to distinguish between solid and dotted lines:

$$H_i(\phi) = -H_i(\phi) \; if \; p_i \in l_d \tag{3.17}$$

Different scales of shape regions are applied to represent local as well as global line structures. The distances L are normalized by the base scale (smallest scale). Figure 3.46 shows a detail of a form with 2 shape regions with a scale of 80 pixels (base scale) and 3 shape regions with a scale of 120 pixels, and its structural features (weighted orientation histogram $H_i(\phi)$). The final line structure feature has a dimension of 24 angular bins (every 5 degrees) which locally describes the shape of binary solid lines, dotted lines and line junctions, robust against distortions like gaps and broken lines.

Compared to Mandal et al. [115] and Fan et al. [49] the line features are not restricted to a certain number of crossings (e.g. 9 [115]). Shapes with a scale smaller than the line spacing can be assumed as the shape primitives defined by Fan et al. [49]. The lines are sampled every 10 pixels to reduce the computational effort. The proposed features are robust against broken lines, since missing points do not affect the shape. Figure 3.45 shows a junction, the current point p_i (red) and the corresponding structural features. All blue points represent the sample points p_j

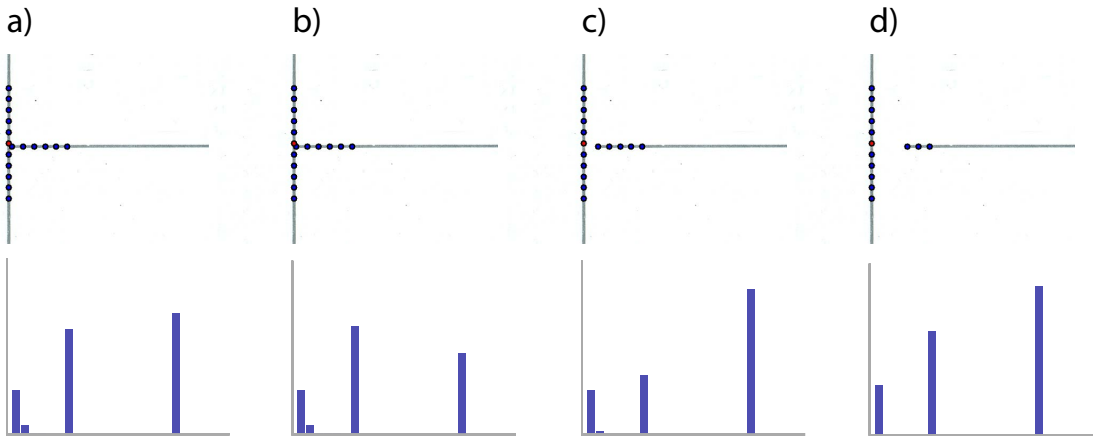


Figure 3.45: T-Junction as a detail of a form with the sampled line points and the structural features below: The red point shows the current point p_i; the blue points are all sample points p_j within in the search window of p_i. In a) the junction is not connected, whereas b) has a connected junction. c) and d) show a broken junction with gaps.

within the search window of p_i. It can be seen that the feature is robust against broken junction or gaps with different sizes.

3.3.3 Classification using BOW

The classification and retrieval is based on the bag of keypoints (motivated by BOW) approach proposed by Csurka et al. [29] who summarizes the classification as following [29]:

- *Detection and description of image patches*
- *Assigning patch descriptors to a set of predetermined clusters (a vocuabulary)*
- *Constructing a bag of keypoints.*
- *Applying a multi-class classifier, treating the bag of keypoints as the feature vector.*

The image patches correspond to the structural features described in the previous section. Instead of determining keypoints like in traditional Content Based Image Retrieval (CBIR) methods each sampling point is used. For the classification the Euclidean distance is tested for classification and retrieval tasks. The predetermined clusters, identifying the appearance of the structural features of a form class (dictionary), are defined by clustering the structural features using k-means (which has also been used by Csurka et al. [29]) and the cluster centers w_i form the words of the dictionary of size i. This is done on a labeled training dataset consisting of forms of the same class. The words of all dictionaries represent frequent structures of all form types. Figure 3.47 (upper part) illustrates the codebook generation. The blue dots represent the

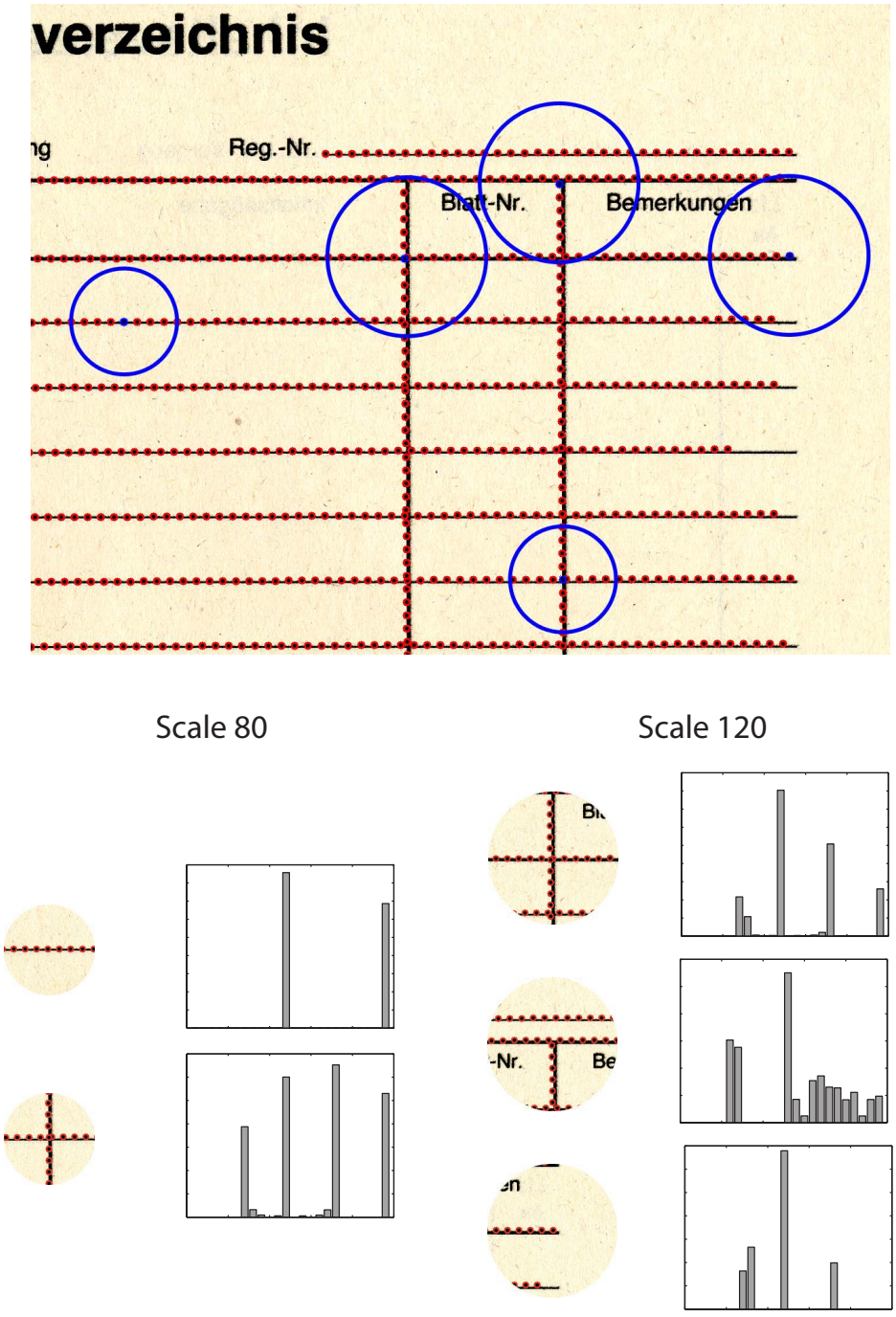


Figure 3.46: Structural shape features for sampled form lines.

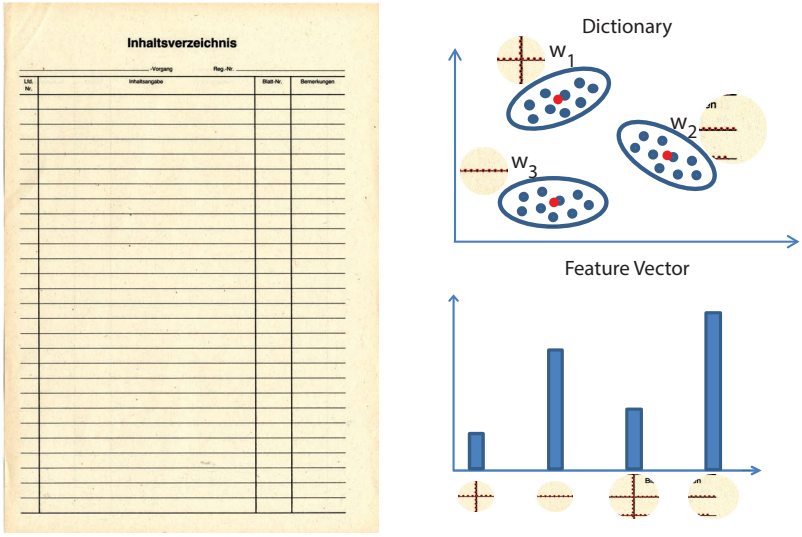


Figure 3.47: Dictionary (feature space with clustered *words*) and a feature vector of a form.

form structures in the feature space, and the cluster centers (ellipse center point, red dots) build the final dictionary. Thus, each form type is represented by a feature vector.

The structural features of a form s_j are calculated and assigned to the cluster center w_i (word) with the smallest (euclidian) distance $\min_i \|s_j - w_i\|$; thus building a histogram of occurrences of the cluster centers (words). Figure 3.47 (lower part) shows a typical form and the pre-determined codebook (frequent structures). Based on the codebook the histogram of occurrences describes the final feature vector.

For the classification the occurrence histogram (feature vector) of the unknown document is compared with every occurrence histogram of the trained form classes. The class with the smallest distance defines the form type. As a distance measurement the Euclidean distance is evaluated. For the form retrieval an arbitrary form is chosen and the feature vector is created and compared with each feature vector of the documents in the dataset using the Euclidean distance. The distances are sorted which defines a ranking for the similarity of the chosen form document. As parameter for the classification the dictionary size (number of clusters defined for the k-means algorithm) has to be chosen. Tests have shown (see Section 3.3.5) that a dictionary size of 120 leads to the best results for the current form types.

3.3.4 Identification of Similar Filled-In Forms (Arlandis et al.)

The form classification using structural features and a BOW approach for classification is compared to the method proposed by Arlandis et al. [9]. The most discriminant regions of document form images are detected in a training phase and a distance based classifier is used for the classification. It is stated that the method can correctly classify similar filled-in forms. Additionally "*very similar*" forms can be distinguished. Since only sub-images are used for classification, the method is also able to handle distorted form documents since global distortions do not affect the

classification. Additionally the distance function to determine the sub-images and to classify a document form image can be adapted. In the following paragraph, the method of Arlandis et al. is summarized.

The discriminant regions are defined as sub-images of a document form image and are called δ-landmarks. For each form class c one reference image I^c is used and the discriminant power $r^c_{x,y}$ of subimages $I^c_{x,y}$ centered at x, y of size w with respect to other form classes is defined by [9]:

$$r^c_{x,y} = \min_{\substack{c' \neq c \\ -w_x \leq i \leq w_x \\ -w_y \leq j \leq w_y}} d(I^c_{x,y}, I^{c'}_{x+i,y+j}) \tag{3.18}$$

All sub-images with a dissimilarity $r^c_{x,y}$ higher than a defined threshold T are defined as δ-landmarks for a form class c and are called class model L^c. The threshold T and the size of the δ-landmarks are chosen based on experiments and the characteristics of the form document classes. The similarity $S^c(J)$ of a document J respectively class c is defined as [9]:

$$S^c(J) = \frac{1}{\|L^c\|} \sum_{I^c_{x,y} \in L^c} \frac{s^c_{x,y}(J)}{r^c_{x,y}} \tag{3.19}$$

where

$$s^c_{x,y}(J) = \min_{\substack{-w_x \leq i \leq w_x \\ -w_y \leq j \leq w_y}} d(I^c_{x,y}, J_{x+i,y+j}) \tag{3.20}$$

It is stated that "*the distribution of the dissimilarities of a document to the others in its own class, S^c, should not overlap with the distribution of the dissimilarities to documents of different classes $S^{c'}$*" [9]. Based on the distance measurement and the dissimilarity distribution a reject option for unknown documents is implemented. As a distance function the sum of absolute differences of binarized sub-images is used. Figure 3.48 shows the selected $delta$-landmarks marked with a red square for the form class "Treffbericht". For a detailed description see Arlandis et al. [9].

The summarized method uses a single reference image to determine the δ-landmarks for a class model L^c. To be able to handle local distortions such as broken junctions and noise the method is extended by Selendi [160] to use M reference images and thus M different sub-images $I^{c_m}_{x,y}, m \in M$ for a form class c. The use of multiple reference images allow to define the inter-class discriminant power $r^c_{inter\,x,y}$ of a class c [160]:

$$r^c_{inter\,x,y} = \min_{\substack{c' \neq c \\ m \in M \\ -w_x \leq i \leq w_x \\ -w_y \leq j \leq w_y}} d(I^{c_m}_{x,y}, I^{c'}_{x+i,y+j}) \tag{3.21}$$

and the intra-class discriminant power $r_{intra\,x,y}$ [160]:

$$r^c_{intra\,x,y} = \max_{\substack{m,n \in M \\ m \neq n}} \min_{\substack{-w_x \leq i \leq w_x \\ -w_y \leq j \leq w_y}} d(I^{c_m}_{x,y}, I^{c_n}_{x+i,y+j}) \tag{3.22}$$

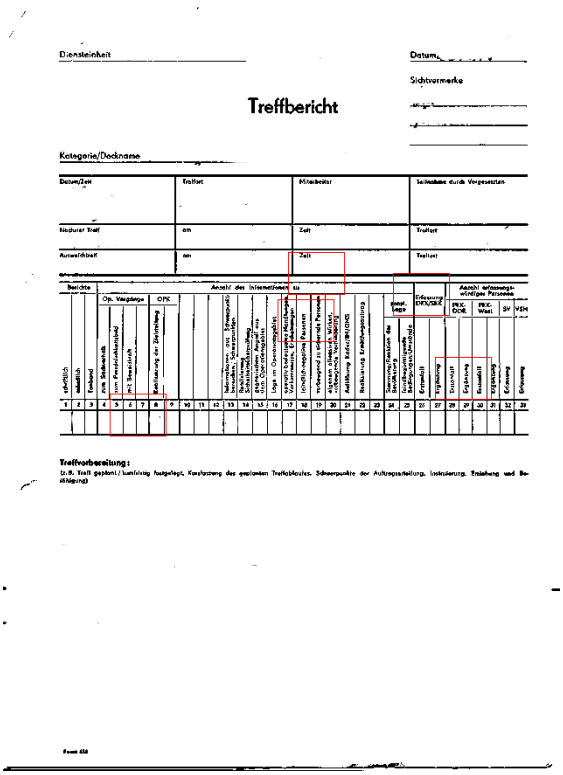


Figure 3.48: Selected δ-landmarks (red squares) of the form class "Treffbericht".

According to the inter-class and intra-class discriminant power the total discriminant power is defined by Equation 3.23:

$$r^c_{x,y} = r^c_{inter\,x,y} - r^c_{intra\,x,y} \tag{3.23}$$

The inter-class discriminant power is the minimum distance when all sub-images $I^{c_m}_{x,y}$, $m \in M$ of a class are compared to all images of other classes. The intra-class discriminant power is the maximum distance of all sub-images $I^{c_m}_{x,y}$ of the same class. If the adapted version which uses multiple reference images is used, the selection of δ-landmarks can be adapted to more stable image-regions. The evaluation presented in Section 3.3.5 shows that the performance increases if multiple reference images for each form class are used.

3.3.5 Results Form Classification

The form classification [85] as well as the line detection [83] of the preprocessing step are evaluated in this section. For the evaluation of the line detection a test set of 27 images is manually labeled. The test images comprise cases with underlined text, images with forms (tables) and images that have no lines. The images are binarized with the method presented in Section 3.1 and based on the binarized image the GT is manually labeled. Table 3.21 shows the

correct classification rate (true positive) as well as the number of pixels falsely classified as lines (false positive). In the final image small gaps ($\leq 20px$) of neighboring lines (same orientation)

Number of images:	27
Correct as line classified pixel (tp):	91.58%
Falsely as line classified pixel (fp):	15.05%

Table 3.21: Evaluation of the preprocessing (line detection).

are closed which leads to the value of 15.05% of false positives. Additionally handwritten text can comprise strokes which are falsely classified as lines. Form classification methods have to deal with incorrectly detected lines (arising from e.g. handwritten text) and incorrectly closed gaps (e.g. see Figure 3.39).

For the form classification a form dataset consisting of 9 different form types with varying size (DIN A6 to DIN A4) and a *non-form* class from the Stasi (see Section 1) dataset has been created. Figure 3.40 in Section 3.3 shows exemplarily 2 form types with a size of appr. DIN A4 and DIN A5. All images have been scanned with a resolution of 300 dpi. The training dataset has at least 4 training forms for every class, the entire test dataset consists of 489 documents, comprising 287 forms and 202 non-form documents. The distribution of the number of documents within a form class represents the number of the chosen form types within a single Stasi IM-record (unofficial collaborator file) and is shown in Table 3.22.

The proposed method is evaluated regarding the size of the structural features and the size of the dictionary of the BOW approach. The size of the structural features corresponds to the complexity of the shapes, whereas the size of the dictionary corresponds to the number of different structural features. A dictionary size of 120 words leads to the best results. Codebooks with fewer words combine different structural features within a single cluster, and a higher number of structural features causes empty bins in the occurrence histogram (feature vector). For the scale of the structural features a combination of different scales has the best result. A scale size of 120 pixels comprises *single* junction types similar as the one defined by Fan et al. [48].

Table 3.22 shows the result of the classification with a dictionary size of 120, and 3 different scales (size of the shape region) of the structural features comprising 120, 580 and 840 pixels. The classification regarding only forms has an overall precision of 87.11% and the precision of the classification including the non-form class is 80.98%. Misclassified documents of the non-form class are documents which have a similar structure compared to forms (e.g. lined paper can be classified as form *"Table of Contents"*, if the lined structure of the paper is segmented).

Table 3.23 shows the classification result with the same dictionary size of 120 words, if only a single scale (120 pixel) is applied. It can be seen that the overall precision drops from 87.11% to 80.35%, since global structures are not represented (leading to ambiguous feature vectors). Table 3.24 summarizes the classification results regarding different scales, whereas Table 3.25 summarizes the classification results regarding different dictionary sizes at the same scales. A smaller dictionary size combines similar structures (clusters in the feature space) resulting in less descriptive feature words, and thus in an accuracy of 83.62% (without non-forms) compared to an accuracy of 87.11%.

Table 3.22: The rows of the confusion matrix show the GT labels (9 different form types and a class which contains no form documents), while the columns represent predicted labels (e.g. 2 forms of the type 0 ("Table of Contents") are falsely classified as form type 4). A scale size of 120, 580 and 840 (size of the shape regions) pixels, and a dictionary size of 120 words is set. Overall accuracy: 87.11% (without non-form class) rsp. 80.98%.

	predicted										
	0	1	2	3	4	5	6	7	8	no form	#
0	47	3	1	2	2	·	12	·	·	·	67
1	·	39	3	·	·	·	·	·	·	·	42
2	·	·	70	·	·	1	·	·	·	·	71
3	·	·	·	26	·	·	·	·	·	·	26
4	·	·	·	1	9	2	·	·	·	·	12
5	·	·	·	6	·	13	·	·	·	·	19
6	·	·	·	·	·	·	31	·	·	·	31
7	·	·	·	3	·	1	·	1	·	·	5
8	·	·	·	·	·	·	·	·	14	·	14
no form	7	·	·	19	9	16	2	1	2	146	202

Table 3.23: Classification with a single scale size of 120 pixels (size of the shape regions), and a dictionary size of 120 words is set. Overall accuracy: 80.35% (without non-form class) rsp. 77.91%.

	predicted										
	0	1	2	3	4	5	6	7	8	no form	#
0	42	6	3	·	·	·	13	1	·	2	67
1	·	37	5	·	·	·	·	·	·	·	42
2	·	·	60	·	5	6	·	·	·	·	71
3	·	·	·	26	·	·	·	·	·	·	26
4	·	·	·	1	8	3	·	·	·	·	12
5	·	·	·	6	·	13	·	·	·	·	19
6	·	·	·	·	·	·	30	1	·	·	31
7	·	·	·	2	1	1	·	1	·	·	5
8	·	·	·	·	2	·	·	·	12	·	14
no form	5	2	1	19	7	11	3	1	1	152	202

The proposed method is compared to the form classification of Arlandis et al. [9], which is summarized in Section 3.3.4. The method of Arlandis et al. is chosen since it is capable of

Table 3.24: Classification results regarding scales (dictionary size of 120).

scales [pixel]	accuracy (incl. non-forms) [%]	accuracy (w/o non-forms) [%]
120	77.91	80.35
120, 580	78.12	81.53
120, 580, 840	80.98	87.11

Table 3.25: Classification results regarding dictionary size (scales 120, 580, 840).

dictionary size	accuracy (incl. non-forms) [%]	accuracy (w/o non-forms) [%]
100	78.73	83.62
120	80.98	87.11
140	80.78	87.46

filled-in forms with handwritten text and can also identify *similar* form documents (e.g. table of contents and index). As parameters the δ-landmark size and the δ-landmark shift is evaluated. In [9] it is stated that the size of the $delta$-landmarks has an impact on the discriminant power dependent on the form type variations. In contrast to the original work of [9] the method is also evaluated with the extended approach, where multiple form documents of the same class can be used in training step. Table 3.28 shows the results of the original approach with different δ-landmark sizes, with a fixed landmark shift and a discriminant power of 0.5. The same evaluation is done regarding the δ-landmark size but with the extended approach (multiple training images) in Table 3.29. The size of the $delta$-landmarks has no significant influence on the classification since the form documents differ in their general appearance. As stated in [9] the size has only a "*strong influence on the discriminant power of the δ-landmarks when the document classes are very similar*". The results show that the discriminant power defined by the intra-class and inter-class discriminant power leads to results with an accuracy over 80% compared to the original approach with appr. 75%. Table 3.29 shows also the impact of the discriminant power threshold for the selection of the δ-landmarks. Also the overall accuracy is higher with a lower threshold (less discriminative landmarks), since more non-form documents are classified as form documents. The missing entries occur if the method does not find any landmarks. Figure 3.49 shows the selected landmarks within the form *Treffbericht* marked with red boxes corresponding to the size of the landmarks. It can be seen that a different size leads to different landmarks based on the discriminant power.

To compare the best results of Arlandis with the proposed approach, the confusion matrix of the best result achieved by Arlandis et al. is shown in Table 3.26 (threshold of 0.1) and in Table 3.27 (threshold 0.5). The method achieved an overall accuracy of 83.03% (threshold 0.5) and an accuracy of 86.09% with a discriminant power threshold of 0.1.

Table 3.27 shows the confusion matrix of the result achieved by Arlandis with the same δ-landmark size and shift, but with a threshold of 0.5 for the discriminant power (selection of relevant landmarks) to show the influence of the threshold. It can be seen that with a higher threshold (less discriminative landmarks) more non-form documents are classified as form doc-

Table 3.26: Classification with the extended method of Arlandis et al. [9]. The best result is achieved with a $delta$-landmark size of width 24 px, height 8 px, a δ-landmark shift of 50 px and a threshold of 0.1 for the discriminant power (selection of the landmarks).

	predicted										
	0	1	2	3	4	5	6	7	8	no form	#
0	66	·	·	·	·	·	·	·	·	1	67
1	·	42	·	·	·	·	·	·	·	·	42
2	·	·	71	·	·	·	·	·	·	·	71
3	·	·	·	18	·	·	·	·	·	8	26
4	·	·	·	·	·	·	·	·	·	12	12
5	·	·	·	·	·	·	·	·	·	19	19
6	·	·	·	·	·	·	25	·	·	6	31
7	·	·	·	2	1	1	·	4	·	1	5
8	·	·	·	·	2	·	·	·	·	14	14
no form	·	·	·	6	·	·	1	·	·	195	202

Table 3.27: Classification with the extended method of Arlandis et al. [9]. $delta$-landmark size of width 24 px, height 8 px, a δ-landmark shift of 50 px and a threshold of 0.5 for the discriminant power (selection of relevant landmarks).

	predicted										
	0	1	2	3	4	5	6	7	8	no form	#
0	66	·	·	·	·	·	·	·	·	1	67
1	·	42	·	·	·	·	·	·	·	·	42
2	·	·	71	·	·	·	·	·	·	·	71
3	·	·	·	23	·	·	·	·	·	3	26
4	·	·	·	1	·	·	·	·	·	11	12
5	·	·	·	8	·	·	·	·	·	11	19
6	1	·	·	·	·	·	28	·	·	2	31
7	·	·	·	·	·	·	·	4	·	1	5
8	·	·	·	·	2	·	·	·	·	14	14
no form	2	12	1	13	·	·	2	·	·	172	202

uments. Thus, the overall accuracy drops from 86.09% to 83.03%.

Table 3.30 compares the results of Arlandis et al. with the proposed approach. It can be seen that the accuracy of the proposed approach for the classification without non-forms outperforms the approach of Arlandis et al. (87.11% vs. 86.41%). A typical form image that is correctly

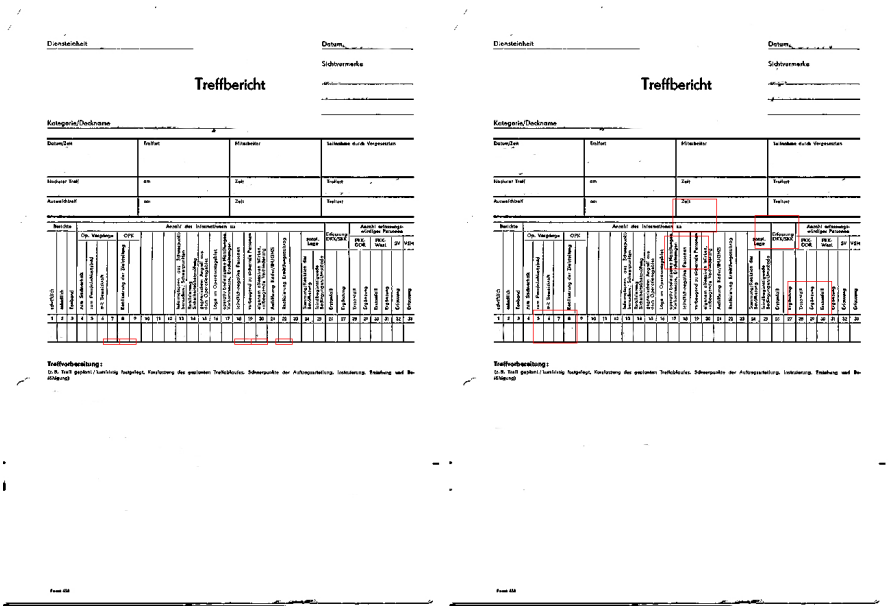


Figure 3.49: Delta landmarks of width 24 and height 8 (left) and width 64 and height 48 (right) with the highest discriminative power.

Table 3.28: Classification results regarding δ-landmark size and training with one reference image. δ-landmark shift is set to 50 px; discriminant power threshold=0.5.

width [px]	accuracy [%] height = 24px	accuracy [%] height = 48px
32	74.23	70.35
48	76.69	82.53
64	76.48	75.26
80	75.05	76.28

Table 3.29: Classification results regarding δ-landmark size and training with multiple reference images. δ-landmark shift is set to 50 px.

	threshold=0.5		threshold=0.1	
width [px]	accuracy [%] height = 24px	accuracy [%] height = 48px	accuracy [%] height = 24px	accuracy [%] height = 48px
32	-	84.25	-	84.66
48	83.44	79.14	82.41	79.35
64	80.16	83.84	79.75	84.05
80	84.05	83.03	84.05	82.82

classified by the proposed approach and not by Arlandis et al. is shown in Figure 3.50. The form is posted on a larger page. Since the structural features are not affected by this type of

Table 3.30: Results of the proposed approach vs. Arlandis [9] (Overall accuracy and accuracy without non-forms)

	overall accuracy (inc. non-forms [%]	accuracy (w/o non-forms) [%]
Arlandis et al. threshold=0.5	83.03	86.41
Arlandis et al. threshold=0.1	86.09	83.62
proposed approach	80.98	87.11

transformation the document is correctly classified, whereas Arlandis et al. does not allow this type of change (local position of the landmark changes). Additionally, if a class has no correctly classified form types, no δ-landmarks are found. This can be seen in the confusion matrix (see Table 3.26 and Table 3.27).

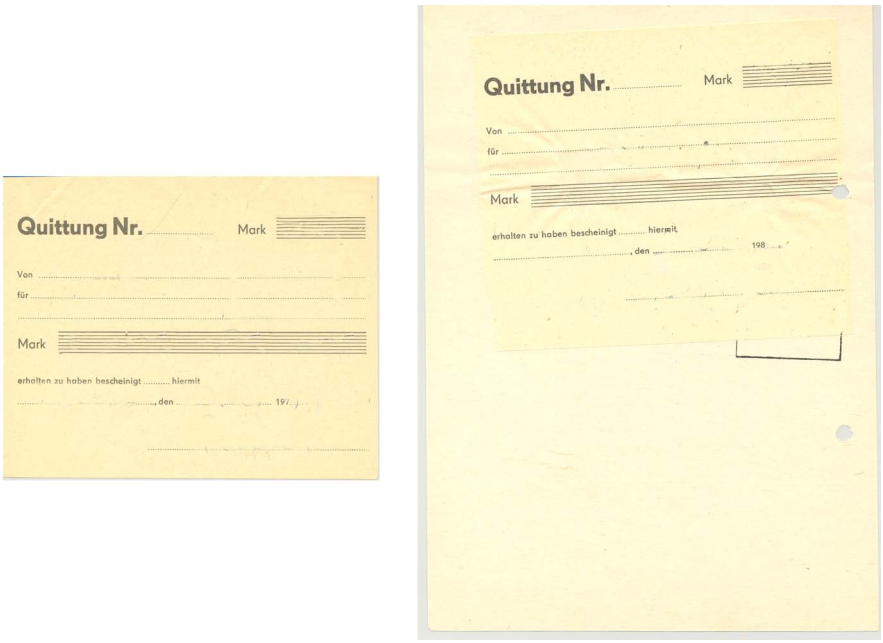


Figure 3.50: Training form *Quittung* document (left) and a correct classified test document (right).

It is also shown in Table 3.30 that the overall accuracy (with non-form documents) of Arlandis et al. is better than the proposed approach (86.09% vs. 80.98%). This is based on the fact, that the proposed approach classifies documents with a similar structural appearance as a form document, which is rejected by Arlandis et al. . Figure 3.51 shows a document which is rejected by Arlandis et al. but classified as *Suchauftrag* by the proposed method.

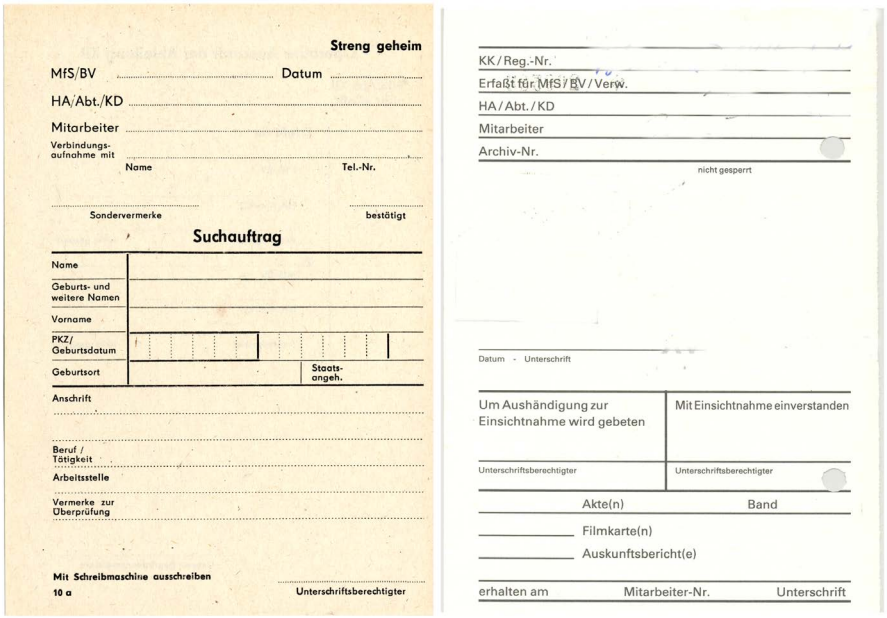


Figure 3.51: Training form *Suchauftrag* document (left) and a falsely classified test document (right). It can be seen that the structural features are similar of both documents are similar.

3.3.6 Summary and Critical Reflection of the proposed Form Classification

The proposed form classification is based on the line structure of form documents. In contrast to current state of the art methods, junctions are not defined by their type, but represented by their shape of interest points of sampled form lines. The main junctions are then defined by the form types using a BOW approach. The main contribution of the form classification is the definition of shape features, which are robust against broken lines. Thus, no preprocessing is necessary in contrast to approaches that analyze the structure of form documents. The method is evaluated on a real world dataset comprising historical documents. It is shown that the proposed approach outperforms the approach of Arlandis et al. which is adapted to the used dataset (see Section 3.3.4). The definition of the shape features allows also to search form documents with a similar structure in a document database. However, this has a negative impact for the form classification, since similar documents regarding their structure cannot be distinguished. Thus, as a future work, the approach has to be combined with the a-priori knowledge of preprinted text.

CHAPTER 4

Conclusion and Future Work

In this thesis three topics of DIA are investigated: document binarization, document skew estimation and form classification. The first topic, document image binarization, covered a Gaussian scale space binarization with a foreground estimation. It is shown, on the basis of a historical dataset with a defined ruling, that the layout information can be used as foreground estimation. The foreground is estimated to suppress background noise. The main contribution to document image binarization is to eliminate parameters like e.g. the stroke width by introducing a Gaussian scale space. The foreground information of different scales is propagated to the scale with the original resolution, to treat objects of different sizes. Thus, the main advantage is the independence to scale dependent parameters. The evaluation is based on the metrics presented at DIBCO and H-DIBCO. In addition to synthetic images, the contest datasets are used for the evaluation. As future work, a classification of text regions must be involved without the a-priori knowledge of a documents layout. The detection of the text region improves the binarization result by suppressing noise outside text regions and enhancing low contrast text within the determined text region. To localize text regions, binarization free text recognition methods have to be evaluated. Based on the results, a weighting scheme can be established as a foreground estimation.

In Section 3.2 the proposed skew estimation is presented. The main research contribution of the skew estimation is the investigation of a binarization free method with no angle restriction. Additionally, the proposed method is suitable for handwritten documents and sparsely inscribed documents. The method is based on a combination of FNNC and a gradient orientation methods. The thesis covers the analysis of the accuracy of the gradient orientation and the comparison of grayvalue images respectively binarized ones. The gradient information has a high accuracy regarding the orientation of printed and handwritten (Latin) text on grayvalue images. It is shown that the information of single (printed characters) is sufficient for the gradient orientation measure. The only preprocessing applied to the gradient orientation method is a Gaussian smoothing. Thus no binarization, which can introduce errors on historical document images, is needed.

The FNNC can solve the skew estimation on an angle range up to $180\,^{\circ}$ and is robust against slanted text. The main research contribution of the FNNC method is the behavior of the method on handwritten documents. Thus, the combination allows to correct the result of the gradient orientation measure by reducing the search space for a maxima in the gradient orientation histogram. It is shown that the combined method can be applied to document fragments with a high accuracy. The proposed skew correction based on the analysis of paragraphs and lines can enhance the final result based on the types of documents. It is shown that the result is improved on printed documents with an accurate alignment of text columns. To propose a skew estimation without an angle restriction, the up/down orientation decision is based on the statistical analysis of ascenders and descenders of English and German Latin text. The method is evaluated on the DISEC dataset and two datasets generated from the PRIMA images with an angle restriction of $\pm 15\,^{\circ}$. The PRIMA images comprise printed document images as well as historical document images. Additionally a synthetic test set based on Epshtein (see [44]) and a real world dataset consisting of 658 document fragments of the Stasi-files are used. It is shown that the orientation can be reliably calculated without any restrictions on the detectable angle range. As future work, the up/down decision must be investigated for document images comprising different languages and scripts.

The last DIA preprocessing topic of this thesis, form classification, is presented in Section 3.3. The research of the thesis is focused on the representation of form documents based on shape features of the line structure. The form document is represented by a histogram of structural features of lines (solid and dotted) which are trained offline for every form class. The creation of a dictionary for every form class is based on the BOW approach. The main contribution of the form classification is the definition of shape features, which are robust against broken lines. The method is evaluated on a real world dataset comprising historical documents. It is shown, that the proposed approach outperforms the approach of Arlandis et al. which is adapted to the dataset used (see Section 3.3.4, overall accuracy of 87.11%). The definition of the shape features allows also to search form documents with a similar structure in a document database. However, this has a negative impact on the form classification, since similar documents due to their structure cannot be distinguished. As future work and to improve the classification accuracy, the information of preprinted text has to be incorporated into the vocabulary representation. The combination will allow to classify also forms without or only sparse line information. Additionally forms differing only in the layout of the preprinted data can be correctly classified by combining layout features with the line information. Future tests will also comprise reconstructed documents with missing parts to analyze the robustness of the proposed method to missing line parts.

APPENDIX A

Acronyms and Symbols

DIA	Document Image Analysis
CC	Connected Component
SVM	Support Vector Machine
K-NN	k-Nearest Neighbor
OCR	Optical Character Recognition
PCA	Principle Component Analysis
DSCC	Directional Single-Connected Chain
H-DSCC	Horizontal-Directional Single-Connected Chain
V-DSCC	Vertical-Directional Single-Connected Chain
TP	True Positives
TN	True Negatives
FP	False Positives
FN	False Negatives
P-FM	pseudo-F-Measure
FM	F-Measure
P-R	p-Recall
PSNR	Peak Signal-to-Noise Ratio

NRM	Negative Rate Metric
MPM	Misclassification Penalty Metric
N	Number
GT	Ground-Truth
DRD	Distance Reciprocal Distortion Metric
DIBCO	Document Image Binarization Contest
H-DIBCO	Handwritten-Document Image Binarization Contest
ICDAR	International Conference on Document Analysis and Recognition
ICFHR	International Conference on Frontiers in Handwriting Recognition
MSI	MultiSpectral Imaging
BOW	Bag of Words
FNNC	Focused Nearest Neighbour Clustering
NCC	Normalized Cross Correlation
DFT	Discrete Fourier Transform
NN	Neural Network
PSD	Power Spectral Density
MLP	Multi-Layer Perceptron
DP	Dynamic Programming
DOCTYPE	Document type
DISEC	Document Image Skew Estimation Contest
EM	Expectation Maximization
LPP	Local Projection Profiles
BB	Bounding Box
LDA	Linear Discrimant Analysis
CBIR	Content Based Image Retrieval
CMD	Common-Minus-Difference
AED	Average Error Deviation

CE Correct Estimations

TOP80 TOP80 Average Error Deviation

HOG Histogram of Oriented Gradients

LTP Local Ternary Patterns

MACELBP MACeLBP

MSC Multivariate Spatial Correlation

Bibliography

[1] B. Allier, N. Bali, and H. Emptoz. Automatic accurate broken character restoration for patrimonial documents. *International Journal for Document Analysis and Recognition (IJDAR)*, 8(4):246–261, 2006.

[2] A. Amin and S. Fischer. A Document Skew Detection Method Using the Hough Transform. *Pattern Analysis and Applications*, 3(3 2000):243–253, 2000.

[3] A. Amin and S. Wu. Robust skew detection in mixed text/graphics documents. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 247 – 251 Vol. 1, aug. 2005.

[4] M. Anthimopoulos, B. Gatos, and I. Pratikakis. Detection of artificial and scene text in images and video frames. *Pattern Analysis and Applications*, 16(3):431–446, 2013.

[5] A. Antonacopoulos, C. Clausner, C. Papadopoulos, and S. Pletschacher. Historical document layout analysis competition. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520, 2011.

[6] A. Antonacopoulos, S. Pletschacher, D. Bridson, and C. Papadopoulos. ICDAR 2009 Page Segmentation Competition. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 1370 –1374, jul. 2009.

[7] V.N. Manjunath Aradhya, G. Hemantha Kumar, and P. Shivakumara. Skew Detection Technique for Binary Document Images based on Hough Transform. *Int. Journal of Information Technology*, 3(1):194–200, 2006.

[8] Hrishikesh B. Aradhye. A generic method for determining the up/down orientation of text in roman and non-roman scripts. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 187–191, Washington, DC, USA, 2005. IEEE Computer Society.

[9] J. Arlandis, J.-C. Perez-Cortes, and E. Ungria. Identification of Very Similar Filled-in Forms with a Reject Option. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 246 –250, jul. 2009.

[10] B. Tenório Ávila and R. Dueire Lins. A fast orientation and skew detection algorithm for monochromatic document images. In *Proceedings of the ACM Symposium on Document Engineering (DocEng)*, pages 118–126, New York, NY, USA, 2005. ACM.

[11] E. Badekas and N. Papamarkos. Automatic evaluation of document binarization results. In Alberto Sanfeliu and ManuelLazo Cortés, editors, *Progress in Pattern Recognition, Image Analysis and Applications*, volume 3773 of *Lecture Notes in Computer Science*, pages 1005–1014. Springer Berlin Heidelberg, 2005.

[12] I. Bar-Yosef, N. Hagbi, K. Kedem, and I. Dinstein. Fast and Accurate Skew Estimation Based on Distance Transform. In *Proceedings of the International Workshop on Document Analysis Systems (DAS)*, pages 402 –407, sep. 2008.

[13] I. Bar-Yosef, Nate Hagbi, Klara Kedem, and Itshak Dinstein. Line Segmentation for degraded handwritten historical documents. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 1161–1165. IEEE Computer Society, 2009.

[14] E. Bart and P. Sarkar. Information extraction by finding repeated structure. In *Proceedings of the International Workshop on Document Analysis Systems (DAS)*, pages 175–182, New York, NY, USA, 2010. ACM.

[15] Y. Belaid, A. Belaid, and E. Turolla. Item searching in forms: Application to French tax form. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, volume 2, pages 744 –747 vol.2, aug. 1995.

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

[17] I. Blayvas, A. Bruckstein, and R. Kimmel. Efficient computation of adaptive threshold surfaces for image binarization. *Pattern Recognition*, 39(1):89–101, 2006.

[18] J.E. Bresenham. Algorithm for computer control of a digital plotter. *IBM Systems Journal*, 4(1):25–30, 1965.

[19] Y. Byun, S. Yoon, Y. Choi, G. Kim, and Y. Lee. An Efficient Form Classification Method Using Partial Matching. In Markus Stumptner, Dan Corbett, and Mike Brooks, editors, *AI 2001: Advances in Artificial Intelligence*, volume 2256 of *Lecture Notes in Computer Science*, pages 291–300. Springer Berlin / Heidelberg, 2001.

[20] H. Cao, X. Ding, and C. Liu. Rectifying the bound document image captured by the camera: a model based approach. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 71 – 75 vol.1, aug. 2003.

[21] R. S. Caprari. Algorithm for text page up/down orientation determination. *Pattern Recognition Letters*, 21(4):311–317, 2000.

[22] R. Casey, D. Ferguson, K. Mohiuddin, and E. Walach. Intelligent forms processing system. *Machine Vision and Applications*, 5:143–155, 1992.

[23] R. Cattoni, T. Coianiz, S. Messelodi, and C. M. Modena. Geometric layout analysis techniques for document image understanding: a review. Technical report, ITC-IRST Technical Report 9703-09, 1998.

[24] B.B. Chaudhuri, P. Kundu, and N. Sarkar. Detection and gradation of oriented texture. *Pattern Recognition Letters*, 14(2):147 – 153, 1993.

[25] M. Chen and X. Ding. A robust skew detection algorithm for grayscale document image. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 617 –620, sep. 1999.

[26] N. Chen and D. Blostein. A survey of document image classification: problem statement, classifier architecture and performance evaluation. *International Journal for Document Analysis and Recognition (IJDAR)*, 10(1):1–16, 2007.

[27] S. Chen, R.M. Haralick, and I.T. Phillips. Automatic text skew estimation in document images. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, volume 2, pages 1153 –1156 vol.2, aug. 1995.

[28] E. Clavier, E. Trupin, M. Laurent, S. Diana, and J. Labiche. Classifiers combination for forms sorting. In *Proceedings of the 15th International Conference on Pattern Recognition (ICPR)*, volume 1, pages 932–935 vol.1, 2000.

[29] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[30] J. F. Cullen, D. G. Stork, P. Hart, and K. Ejiri. Method for detecting inverted text images on a digital scanning device. us patent 6574375 b1., 1996.

[31] A. Curry. Archive Collapse Disaster for Historians. *Spiegel online international*, accessed 04th march 2009. `http://www.spiegel.de/international/germany/0,1518,611311,00.html`.

[32] A. Kumar Das and B. Chanda. A fast algorithm for skew detection of document images using morphology. *International Journal for Document Analysis and Recognition (IJDAR)*, 4(2):109–114, 2001.

[33] M. Diem, F. Kleber, and R. Sablatnig. Analysis of Document Snippets as a Basis for Reconstruction. In Kurt Debattista, Cinzia Perlingieri, Denis Pitzalis, and Sandro Spina, editors, *10th International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST)*, pages 101–108, St. Julians, Malta, 2009.

[34] M. Diem, F. Kleber, and R. Sablatnig. Document Analysis Applied to Fragments: Feature Set for the Reconstruction of Torn Documents. In D. Doermann, V. Govindaraju, D. Lopresti, and P. Natarajan, editors, *Proceedings of the International Workshop on Document Analysis Systems (DAS)*, pages 393–400, Boston, USA, June 2010.

[35] M. Diem, F. Kleber, and R. Sablatnig. Text Classification and Document Layout Analysis of Torn Documents. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 1181–1184, Beijing, China, 2011. IEEE Computer Society CPS.

[36] M. Diem, F. Kleber, and R. Sablatnig. Skew Estimation of Sparsely Inscribed Document Fragments. In *Proceedings of the International Workshop on Document Analysis Systems (DAS)*, Goldcoast, Australia, 2012.

[37] D. Dimmick, M. Garris, and C. Wilson. Structured forms database. *Technical Report Special Database 2, SFRS, National Institute of Standards and Technology*, 2001.

[38] D.S. Doermann and A. Rosenfeld. The processing of form documents. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 497 –501, oct. 1993.

[39] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *British Machine Vision Conference (BMVC)*, 2009.

[40] Leyza Baldo Dorini and Neucimar Jeronimo Leite. A Multiscale Morphological Binarization Algorithm. *Computer Vision, Imaging and Computer Graphics. Theory and Applications*, 68:283–295, 2010.

[41] P. Duygulu and V. Atalay. A hierarchical representation of form documents for identification and retrieval. *International Journal for Document Analysis and Recognition (IJDAR)*, 5(1):17–27, 2002.

[42] R.L. Easton, K.T. Knox, and W.A. Christens-Barry. Multispectral Imaging of the Archimedes Palimpsest. In *32nd Applied Image Pattern Recognition Workshop, AIPR 2003*, pages 111–118, Washington, DC, October 2003. IEEE Computer Society.

[43] A. Egozi and I. Dinstein. An EM Based Algorithm for Skew Detection. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 277 –281, sep. 2007.

[44] B. Epshtein. Determining document skew using inter-line spaces. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 27–31, 2011.

[45] H. Ezaki, S. Uchida, A. Asano, and H. Sakoe. Dewarping of document image by global optimization. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 302 – 306 Vol. 1, aug. 2005.

[46] J. Fabrizio, B. Marcotegui, and M. Cord. Text Segmentation in Natural Scenes Using Toggle-Mapping. In *Proceedings of the International Conference on Image Processing (ICIP)*, pages 2373–2376, 2009.

[47] J. Fan, R. Wang, L. Zhang, D. Xing, and F. Gan. Image sequence segmentation based on 2d temporal entropic thresholding. *Pattern Recognition Letters*, 17(10):1101–1107, 1996.

[48] Kuo-Chin Fan, Yuan-Kai Wang, and Mei-Lin Chang. Form document identification using line structure based features. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 704 –708, 2001.

[49] L. Fan and C. L. Tan. Binarizing document image using coplanar prefilter. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 34 –38, 2001.

[50] ABBYY FineReader. www.finereader.com, accessed 09, 2010.

[51] A. Fischer, M. Wuthrich, M. Liwicki, V. Frinken, H. Bunke, G. Viehhauser, and M. Stolz. Automatic transcription of handwritten medieval documents. In *15th International Conference on Virtual Systems and Multimedia (VSMM)*, pages 137–142, 2009.

[52] B. Gatos, A. Antonacopoulos, and N. Stamatopoulos. Handwriting Segmentation Contest. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, volume 2, pages 1284 –1288, sep. 2007.

[53] B. Gatos, K. Ntirogiannis, and I. Pratikakis. ICDAR 2009 Document Image Binarization Contest (DIBCO 2009). In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 1375 –1382, jul. 2009.

[54] B. Gatos, N. Papamarkos, and C. Chamzas. Skew detection and text line position determination in digitized documents. *Pattern Recognition*, 30(9):1505 – 1519, 1997.

[55] B. Gatos, I. Pratikakis, and I.J. Perantonis. Adaptive degraded document image binarization. *Pattern Recognition*, 39(3):317–327, 2006.

[56] B. Gatos, I. Pratikakis, and S. J. Perantonis. An adaptive binarization technique for low quality historical documents. In *Proceedings of the International Workshop on Document Analysis Systems (DAS)*, pages 102–113, 2004.

[57] B. Gatos, I. Pratikakis, and S.J. Perantonis. Efficient Binarization of Historical and Degraded Document Images. In *Proceedings of the International Workshop on Document Analysis Systems (DAS)*, pages 447 –454, sep. 2008.

[58] R. Gonzalez and R. Woods. *Digital Image Processing*. Prentice Hall, 2008.

[59] M. R. Gupta, N. P. Jacobson, and E. K. Garcia. OCR binarization and image preprocessing for searching historical documents. *Pattern Recognition*, 40(2):389–397, 2007.

[60] M. Hain, J. Bartl, and V. Jacko. Multispectral Analysis of Cultural Heritage Artefacts. In *Measurement Science Review*, volume 3, pages 9–12, 2003.

[61] G.-H. He, Z.-M. Xie, and R. Chen. Automatic classification of form features based on neural networks and fourier transform. In *International Conference on Machine Learning and Cybernetics, 2008*, volume 2, pages 1162–1166, 2008.

[62] J. He, Q.D.M. Do, A.C. Downton, and J.H. Kim. A comparison of binarization methods for historical archive documents. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 538 – 542 Vol. 1, aug. 2005.

[63] J. He and A.C. Downton. Colour map classification for archive documents. In S. Marinai and A. R. Dengel, editors, *Proceedings of the International Workshop on Document Analysis Systems (DAS)*, volume 3163 of *Lecture Notes in Computer Science*, pages 241–251. Springer Berlin Heidelberg, 2004.

[64] P. Heroux, S. Diana, A. Ribert, and E. Trupin. Classification method study for automatic form class identification. In *Pattern Recognition.*, volume 1, pages 926 –928 vol.1, aug 1998.

[65] T. Hirano, Y. Okada, and F. Yoda. Field extraction method from existing forms transmitted by facsimile. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 738 –742, 2001.

[66] J. Hirayama, H. Shinjo, T. Takahashi, and T. Nagasaki. Development of template-free form recognition system. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 237–241, 2011.

[67] F. Hollaus, M. Gau, and R. Sablatnig. Acquisition and enhancement of multispectral images of ancient manuscripts. In Regina Franken-Wendelstorf, Elisabeth Lindinger, and Juergen Sieck, editors, *Cultur and Computer Science - Visual Worlds and Interactive Spaces*, pages 187–197, 2013.

[68] F. Hollaus, M. Gau, and R. Sablatnig. Enhancement of multispectral images of degraded documents by employing spatial information. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 15–149, 2013.

[69] N. Howe. A Laplacian energy for document binarization. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 6–10, 2011.

[70] N.. Howe. Document binarization with automatic parameter tuning. *International Journal on Document Analysis and Recognition (IJDAR)*, 16(3):247–258, 2013.

[71] J. J. Hull. Document Image Skew Detection: Survey and annotated Bibiliography. In Jonathan J. Hull and Suzanne L. Taylor, editors, *Document Analysis System II*, pages 40–64, 1998.

[72] M. E. Hussein, F. Porikli, and L. S. Davis. Kernel integral images: A framework for fast non-uniform filtering. In *International Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1–8, 2008.

[73] L.A.D. Hutchison and W.A. Barrett. Fast registration of tabular document images using the Fourier-Mellin transform. In *Proceedings of the International Workshop on Document Image Analysis for Libraries*, pages 253 – 267, 2004.

[74] Q. Huynh-Thu and M. Ghanbari. Scope of validity of PSNR in image/video quality assessment. *Electronics Letters*, 44(13):800–801, 2008.

[75] H.-F. Jiang, C.-C. Han, and K.-C. Fan. A fast approach to detect and correct skew documents. *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, 3:742–746, 1996.

[76] X. Jiang, H. Bunke, and D. Widmer-Kljajo. Skew detection of document images by focused nearest-neighbor clustering. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 629 –632, sep. 1999.

[77] J. Kanai and A. D. Bagdanov. Projection profile based skew estimation algorithm for JBIG compressed images. *International Journal for Document Analysis and Recognition (IJDAR)*, 1(1):43–51, 1998.

[78] E. Kavallieratou, N. Fakotakis, and Kokkinakis G. Skew angle estimation for printed and handwritten documents using the Wigner-Ville distribution. *Image and Vision Computing*, 20:813–824, 2002.

[79] J. Kittler and J. Illingworth. Minimum error thresholding. *Pattern Recognition*, 19(1):41–47, 1986.

[80] F. Kleber, M. Diem, F. Hollaus, M. Lettner, R. Sablatnig, M. Gau, and H. Miklas. Technical Approaches to Manuscript Analysis and Reconstruction. In Patricia Engel, Jospeh Schiro, Rene Larsen, Elissaveta Moussakova, and Istvan Kecskemeti, editors, *New Approaches to Book and Paper Conservation-Restoration*, pages 533–558, Horn, Austria, 2011. Verlag Berger.

[81] F. Kleber, M. Diem, M. Lettner, M. Vill, and R. Sablatnig. The Sinaitic Glagolitic Sacramentary Fragments. In Andreas Bienert, Gerd Stanke, and James Hamsley, editors, *Proc. of the EVA 2008 Berlin Conference, Electronic Imaging and the Visual Arts*, Berlin, Germany, 2008.

[82] F. Kleber, M. Diem, and R. Sablatnig. Torn Document Analysis as a Prerequisite for Reconstruction. In Robert Sablatnig, Martin Kampel, and Martin Lettner, editors, *15th International Conference on Virtual Systems and Multimedia (VSMM 2009)*, pages 143–148, Vienna, Austria, 2009.

[83] F. Kleber, M. Diem, and R. Sablatnig. Document Reconstruction by Layout Analysis of Snippets. In *SPIE 2010*, San Jose, CA, USA, 2010.

[84] F. Kleber, M. Diem, and R. Sablatnig. Scale Space Binarization Using Edge Information Weighted by a Foreground Estimation. In *Proceedings of the 11th International Conference on Document Analysis and Reconstruction (ICDAR 2011)*, pages 854–858, Beijing, China, 2011. IEEE Computer Society CPS.

[85] F. Kleber, M. Diem, and R. Sablatnig. Form Classification and Retrieval using Bag of Words with Shape Features of Line Structures. In *Document Recognition and Retrieval XXI*, 2014.

[86] F. Kleber, M. Lettner, M. Diem, M. Vill, R. Sablatnig, H. Miklas, and M. Gau. Multispectral Acquisition and Analysis of Ancient Documents. In M. Ioannides, A. Addison, A. Georgopoulos, and L. Kalisperis, editors, *Proc. of the 14th International Conference on Virtual Systems and MultiMedia (VSMM 2008), Dedicated to Cultural Heritage - Project Papers*, pages 184–191, Limassol, Cyprus, 2008. Archaeolingua.

[87] F. Kleber and R. Sablatnig. Skew Detection Technique Suitable for Degraded Ancient Documents. In *Proceedings of the 36th Conference on Computer Applications and Quantitative Methods in Archaeology (CAA)*, pages 320–325, Budapest, Hungary, 2008.

[88] F. Kleber, R. Sablatnig, M. Gau, and H. Miklas. Ruling Estimation for Degraded Ancient Documents Based on Text Line Extraction. In Robert Sablatnig, James Hemsley, Paul Kammerer, Ernestine Zolda, and Johann Stockinger, editors, *Proc. of 2nd EVA 2008 Vienna Conference, Digital Cultural Heritage - Essential for Tourism*, pages 79–86, Vienna, Austria, 2008. OCG.

[89] F. Kleber, R. Sablatnig, M. Gau, and H. Miklas. Ancient Document Analysis Based on Text Line Extraction. In *Proc. of the International Conference on Pattern Recognition (ICPR)*, Tampa, Florida, USA, 2008.

[90] J. Koenderink. The structure of images. *Biological Cybernetics*, 50:363–370, 1984.

[91] H. I. Koo and N. I. Cho. Skew estimation of natural images based on a salient line detector. *Journal of Electronic Imaging*, 22(1):013020–013020, 2013.

[92] A. Kuijper. *The deep structure of Gaussian scale space images, ISBN 9039330611.* PhD thesis, Utrecht University, the Netherlands, 2002.

[93] D.-S. Lee and J. J. Hull. Group 4 compressed document matching using endpoints. In *3rd IAPR Symposium on Document Analysis Systems*, pages 29–38, Nagano, Japan, 1998.

[94] G. Leedham, Chen Yan, K. Takru, Joie Hadi Nata Tan, and Li Mian. Comparison of some thresholding algorithms for text/background segmentation in difficult document images. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 859 – 864, aug. 2003.

[95] J. Leroy. Les types de reglure des manuscrits grecs, Paris, 1976.

[96] M. Lettner. *On the Combination of Spatial and Spectral Features for Image Restoration*. PhD thesis, Vienna University of Technology, Institute of Computer Aided Automation, Computer Vision Lab, 2010.

[97] M. Lettner, F. Kleber, R. Sablatnig, and H. Miklas. Contrast Enhancement in Multispectral Images by Emphasizing Text Regions. In Koichi Kise and Hiroshi Sako, editors, *Proc. of 8th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 225–232, Nara, Japan, 2008.

[98] M. Lettner and R. Sablatnig. Higher Order MRF for Foreground-Background Separation in Multi-Spectral Images of Historical Manuscripts. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems DAS*, pages 317–324, Boston, USA, 2010. ACM.

[99] V. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707, 1966.

[100] R. E. Lewand. *Cryptological Mathematics*. The Mathematical Association of America, 2005.

[101] D. Li, J. Elton, and S. Steger. Methods and apparatus for auto image binarization, us patent 20100158373 a1, 2010.

[102] J.-Y. Lin, C.-W. Lee, and Z. Chen. Identification of business forms using relationships between adjacent frames. *Machine Vision and Applications*, 9:56–64, 1996.

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

[104] T. Lindeberg. *Scale Space*. Encyclopedia of Computer Science and Engineering. John Wiley and Sons, Hoboken, New Jersey, 2009.

[105] R. D. Lins and B. Tenorio Avila. A new algorithm for skew detection in images of documents. In *Image Analysis and Recognition*, volume 3212 of *Lecture Notes in Computer Science*, pages 234–240. Springer Berlin Heidelberg, 2004.

[106] N. Liolios, N. Fakotakis, and G. Kokkinakis. Improved document skew detection based on text line connected-component clustering. *Proceedings of the International Conference on Image Processing*, 1:1098–1101, 2001.

[107] N. Liolios, N. Fakotakis, and G. Kokkinakis. On the generalization of the form identification and skew detection problem. *Pattern Recognition*, 35(1):253 – 264, 2002.

[108] J. Liu and A.K. Jain. Image-based form document retrieval. In *Proceedings of the Fourteenth International Conference on Pattern Recognition (ICPR)*, volume 1, pages 626–628, 1998.

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

[110] H. Lu, A.C. Kot, and Y.Q. Shi. Distance-reciprocal distortion measure for binary document images. *IEEE Signal Processing Letters*, 11(2):228–231, 2004.

[111] S. Lu, B. Su, and C. L. Tan. Document image binarization using background estimation and stroke edges. *International Journal for Document Analysis and Recognition (IJDAR)*, 13:303–314, December 2010.

[112] S. Lu, J. Wang, and C. L. Tan. Fast and Accurate Detection of Document Skew and Orientation. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, volume 2, pages 684 –688, sep. 2007.

[113] Y. Lu and C. L. Tan. Improved nearest neighbor based approach to accurate document skew estimation. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 503 – 507 vol.1, aug. 2003.

[114] Y. Lu and C. L. Tan. A nearest-neighbor chain based approach to skew estimation in document images. *Pattern Recognition Letters*, 24(14):2315 – 2323, 2003.

[115] S. Mandal, S.P. Chowdhury, A.K. Das, and B. Chanda. A hierarchical method for automated identification and segmentation of forms. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 705 – 709 Vol. 2, aug.-1 sept. 2005.

[116] J. Mao, M. Abayan, and K. Mohiuddin. A model-based form processing sub-system. In *Proceedings of the 13th International Conference on Pattern Recognition (ICPR)*, volume 3, pages 691–695, 1996.

[117] A. O. Maroneze, B. Coueasnon, and A. Lemaitre. Introduction of statistical information in a syntactic analyzer for document image recognition. In Gady Agam and Christian Viard-Gaudin, editors, *Document Recognition and Retrieval XVIII (DRR)*, volume 7874 of *SPIE Proceedings*, pages 1–10, San Francisco, 2011. SPIE.

[118] I.B. Messaoud, H. Amiri, H.E. Abed, and V. Margner. Document preprocessing system - automatic selection of binarization. In *10th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 85 –89, march 2012.

[119] H. Miklas. Zur editorischen Vorbereitung des sog. Missale Sinaiticum (Sin. slav. 5/N). In H. Miklas, V. Sadovski, and S. Richter, editors, *Glagolitica - Zum Ursprung der slavischen Schriftkultur*, pages 117–129. (OAW, Phil.-hist. Kl., Schriften der Balkan-Kommission, Philologische Abt. 41), 2000.

[120] R. Milewski and V. Govindaraju. Binarization and cleanup of handwritten text from carbon copy medical form images. *Pattern Recognition*, 41(4):1308 – 1315, 2008.

[121] R. F. Moghaddam, F. F. Moghaddam, and M. Cheriet. Unsupervised ensemble of experts (eoe) framework for automatic binarization of document images. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 703–707, 2013.

[122] G. Nagy and S. Seth. Hierarchical representation of optically scanned documents. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 347–349, 1984.

[123] W. Niblack. *An Introduction to Image Processing*. Prentice-Hall, 1986.

[124] B. Nickolay and J. Schneider. Virtuelle Rekonstruktion "vorvernichteter" Stasi-Unterlagen. Technologische Machbarkeit und Finanzierbarkeit - Folgerungen für Wissenschaft, Kriminaltechnik und Publizistik. In *Schriftenreihe des Berliner Landesbeauftragten für die Unterlagen des Staatssicherheitsdienstes der ehemaligen DDR*, volume 21, pages 11–28. Johannes Weberling and Giselher Spitzer, 2007.

[125] K. Ntirogiannis, B. Gatos, and I. Pratikakis. An Objective Evaluation Methodology for Document Image Binarization Techniques. In *Proceedings of the The Eighth IAPR International Workshop on Document Analysis Systems (DAS).*, pages 217–224, Washington, DC, USA, 2008. IEEE Computer Society.

[126] K. Ntirogiannis, B. Gatos, and I. Pratikakis. Performance evaluation methodology for historical document image binarization. *IEEE Transactions on Image Processing*, 22(2):595–609, 2013.

[127] National Institute of Standards and Technology. NIST Special Database 2 - Structured Forms Reference Set of Binary Images. http://www.nist.gov/srd/nistsd2.cfm, accessed 10/2013.

[128] L. O'Gorman. Experimental comparisons of binarization and multi-thresholding methods on document images. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition (ICPR)*, volume 2, pages 395–398 vol.2, 1994.

[129] Lawrence O'Gorman and Rangachar Kasturi. *Document Image Analysis*. IEEE Computer Society Press, 1995.

[130] R. Ohtera and T. Horiuchi. Faxed Form Identification using Histogram of the Hough-Space. In *Proceedings of the 17th International Conference on Pattern Recognition, (ICPR)*, volume 2 of *ICPR '04*, pages 566–569, Washington, DC, USA, 2004. IEEE Computer Society.

[131] O. Okun, M. Pietikäinen, and J. J. Sauvola. Document skew estimation without angle range restriction. *International Journal for Document Analysis and Recognition (IJDAR)*, 2(2-3):132–144, 1999.

[132] K. Omar, A. Ramli, R. Mahmod, and M. Sulaiman. Skew Detection and Correction of Jawi Images using Gradient Direction. *J. Techn.*, 37:117–126, 2002.

[133] M. Opitz. Text detection and recognition in natural scene images. Master's thesis, Vienna University of Technology, Computer Vision Lab, 2013.

[134] N. Otsu. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):62–66, January 1979.

[135] A. Papandreou, B. Gatos, G. Louloudis, and N. Stamatopoulos. DISEC 2013 - Document Image Skew Estimation Contest. *International Conference on Document Analysis and Recognition*, pages 1476–1480, 2013.

[136] R. Paredes, E. Kavallieratou, and R. Dueire Lins. Icfhr 2010 contest: Quantitative evaluation of binarization algorithms. In *2010 International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 733–736, 2010.

[137] G.S. Peake and T.N. Tan. A General Algorithm for Document Skew Angle Estimation. In *Proceedings of the International Conference on Image Processing ICIP*, pages 230–233, 1997.

[138] S. Pentzien, I. Rabin, O. Hahn, J. Krüger, F. Kleber, F. Hollaus, M. Diem, and R. Sablatnig. Can modern technologies defeat nazi censorship? In Jong-Il Park and Junmo Kim, editors, *Computer Vision - ACCV 2012 Workshops*, volume 7729 of *Lecture Notes in Computer Science*, pages 13–24. Springer Berlin Heidelberg, 2013.

[139] E. Philippot, A. Belaid, and Y. Belaid. Use of pgm for form recognition. In *Proceedings of the International Workshop on Document Analysis Systems (DAS)*, pages 374–378, 2012.

[140] W Postl. Detection of linear oblique structures and skew scan in digitized documents. In *Proc. of the 8th Int. Conference on Pattern Recognition (ICPR)*, pages 687–689, Paris, France, 1986.

[141] I. Pratikakis, B. Gatos, and K. Ntirogiannis. H-DIBCO 2010 - Handwritten Document Image Binarization Competition. *International Conference on Frontiers in Handwriting Recognition*, pages 727–732, 2010.

[142] I. Pratikakis, B. Gatos, and K. Ntirogiannis. ICDAR 2011 Document Image Binarization Contest DIBCO. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 1506–1510, 2011.

[143] I. Pratikakis, B. Gatos, and K. Ntirogiannis. Icdar 2011 document image binarization contest (dibco 2011). In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 1506–1510, 2011.

[144] I. Pratikakis, B. Gatos, and K. Ntirogiannis. ICFHR 2012 competition on handwritten document image binarization (h-dibco 2012). In *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 817 –822, 2012.

[145] I. Pratikakis, B. Gatos, and K. Ntirogiannis. DIBCO 2013 - Document Image Binarization Contest. *International Conference on Document Analysis and Recognition*, pages 1102–1107, 2013.

[146] A. Ravishankar Rao. *A taxonomy for texture description and identification*. Springer-Verlag New York, Inc., New York, NY, USA, 1990.

[147] K. Ramamohan Rao and P. Yip, editors. *The Transform and Data Compression Handbook*. CRC Press, Inc., Boca Raton, FL, USA, 2000.

[148] I. Reisner-Kollmann, A. Reichinger, and W. Purgathofer. 3d camera pose estimation using line correspondences and 1d homographies. 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 6454 of *Lecture Notes in Computer Science*, pages 41–52. Springer Berlin Heidelberg, 2010.

[149] H. Sako, N. Furukawa, M. Fujio, and S. Watanabe. Document-Form Identification Using Constellation Matching of Keywords Abstracted by Character Recognition. In *Document Analysis Systems*, pages 261–271, 2002.

[150] H. Sako, M. Seki, N. Furukawa, H. Ikeda, and A. Imaizumi. Form reading based on form-type identification and form-data recognition. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 926–930, 2003.

[151] E. Salerno, A. Tonazzini, and L. Bedini. Digital image analysis to enhance underwritten text in the Archimedes palimpsest. *International Journal for Document Analysis and Recognition (IJDAR)*, 9(2-4):79–87, 2007.

[152] E. Saund. A graph lattice approach to maintaining dense collections of subgraphs as image features. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 1069–1074, 2011.

[153] J. Sauvola and M. Pietikainen. Page segmentation and classification using fast feature extraction and connectivity analysis. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, volume 2, pages 1127 –1131 vol.2, aug. 1995.

[154] J. Sauvola and M. Pietikainen. Skew angle detection using texture direction analysis. In *In Proc. of the 9th Scandinavian Conference on Image Analysis*, pages 1099–1106, 1995.

[155] J. Sauvola and M. Pietikainen. Adaptive document image binarization. *Pattern Recogn.*, 33(2):225–236, 2000.

[156] K. M. Sayre. Machine recognition of handwritten words: A project report. *Pattern Recognition*, 5(3):213 – 228, 1973.

[157] F. Schmitt, H. Brettel, and J. Hardeberg. Multispectral Imaging Development at ENST. In *Proceedings of the International Symposium on Multispectral Imaging and High Accuracy Color Reproduction*, pages 58–64, 1999. F. Schmitt, H. Brettel, and J. Y. Hardeberg, (Chiba, Japan).

[158] J. Schneider and B. Nickolay. The Stasi puzzle. *Fraunhofer Magazine Special Issue*, pages 32–33, 2008.

[159] R. A. Schowengerdt. *Remote Sensing: Models and Methods for Image Processing*. Elsevier, 3rd edition, 2007.

[160] S. Selendi. Identification of weather data forms. Technical report, Vienna University of Technology, Computer Vision Lab, 2013.

[161] J. Serra. Toggle mappings. In J.C. Simon, editor, *From Pixels to Features*, pages 61–72. Elsevier Science Inc., 1989.

[162] F. Shafait. *Geometric Layout Analysis of Scanned Documents*. PhD thesis, Department of Computer Science, Technical University of Kaiserslautern, 2008.

[163] S. Shimotsuji and M. Asano. Form identification based on cell structure. In *Proceedings of the 13th International Conference on Pattern Recognition (ICPR).*, volume 3, pages 793 –797 vol.3, aug 1996.

[164] E. H. Barney Smith. An analysis of binarization ground truthing. In *Proceedings of the International Workshop on Document Analysis Systems (DAS)*, pages 27–34, New York, NY, USA, 2010. ACM.

[165] E. H. Barney Smith and Chang An. Effect of "ground truth" on image binarization. In *Proceedings of the 2012 10th IAPR International Workshop on Document Analysis Systems*, DAS '12, pages 250–254, Washington, DC, USA, 2012. IEEE Computer Society.

[166] E. H. Barney Smith, L. Likforman-Sulem, and J. Darbon. Effect of pre-processing on binarization. In *Document Recognition and Retrieval (DRR)*, pages 1–10, 2010.

[167] P. Stathis, E. Kavallieratou, and N. Papamarkos. An evaluation survey of binarization algorithms on historical documents. In *International Conference on Pattern Recognition.*, pages 1–4, 2008.

[168] B. Su, S. Lu, and C. L. Tan. Binarization of historical document images using the local maximum and minimum. In *Proceedings of the International Workshop on Document Analysis Systems (DAS)*, pages 159–166, New York, NY, USA, 2010. ACM.

[169] B. Su, S. Lu, and C. L. Tan. Combination of Document Image Binarization Techniques. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, Beijing, China, 2011.

[170] C. Sun and D. Si. Skew and Slant Correction for Document Images Using Gradient Direction. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 142–146, Washington, DC, USA, 1997. IEEE Computer Society.

[171] S. Tabbone and L. Wendling. Multi-scale binarization of images. *Pattern Recognition Letters*, 24(1-3):403 – 411, 2003.

[172] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Transactions on Image Processing*, 19(6):1635–1650, 2010.

[173] O.D. Trier and T. Taxt. Evaluation of binarization methods for document images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(3):312 –315, mar. 1995.

[174] J. van Beusekom, D. Keysers, F. Shafait, and T.M. Breuel. Example-Based Logical Labeling of Document Title Page Images. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, volume 2, pages 919 –923, sep. 2007.

[175] R.G. von Gioi, J. Jakubowicz, J. M Morel, and G. Randall. Lsd: A fast line segment detector with a false detection control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):722–732, 2010.

[176] R.S. Wallace. A modified hough transform for lines. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 665–667, 1985.

[177] Roy E. Welsch and Edwin Kuh. Linear Regression Diagnostics. Technical Report 923-77, Massachusetts Institute of Technology, April 1977.

[178] A. P. Witkin. Scale-space filtering. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence.*, volume 2 of *IJCAI'83*, pages 1019–1022, San Francisco, CA, USA, 1983. Morgan Kaufmann Publishers Inc.

[179] C. Wolf, J.M. Jolion, and F. Chassaing. Text Localization, Enhancement and Binarization in Multimedia Documents. *International Conference onPattern Recognition, (ICPR)*, 2:1037–1040, 2002.

[180] W. S. Wong, N. Sherkat, and T. Allen. Use of colour in form layout analysis. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 942 –946, 2001.

[181] Y. Xiao, Z. Cao, and T. Zhang. Entropic thresholding based on gray-level spatial correlation histogram. In *19th International Conference on Pattern Recognition (ICPR)*, pages 1–4, 2008.

[182] H. Yan. Skew correction of document images using interline cross-correlation. *CVGIP: Graphical Models and Image Processing*, 55(6):538 – 543, 1993.

[183] J.Y. Yoo, M.K. Kim, S.Y. Ban, and Y.B. Kwon. Line removal and restoration of handwritten characters on the form documents. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 128–131, aug. 1997.

[184] B. Yu and A. K. Jain. A Generic System for Form Dropout. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(11):1127–1134, 1996.

[185] B. Yuan and C. L. Tan. Skew estimation for scanned documents from noises. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 277 – 281 Vol. 1, aug. 2005.

[186] Z. Zhang and C. L. Tan. Recovery of distorted document images from bound volumes. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 429 –433, 2001.

[187] Y. Zheng, C. Liu, X. Ding, and S. Pan. Form frame line detection with directional single-connected chain. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 699–703, 2001.

[188] Yefeng Zheng, Huiping Li, and D. Doermann. A model-based line detection algorithm in documents. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 44 – 48 vol.1, aug. 2003.

[189] Y. Zhihui, C. Wenjuan, and Z. Mengmeng. Deep structure of gaussian scale space. In *International Conference on Computer Science and Software Engineering*, volume 6, pages 381–384, 2008.

Contact Information:
Eisentürgasse 11
3500 Krems
e-mail: florian.kleber@gmx.at

Florian Kleber

Personal Data

Date of Birth	20.12.1979
Place of Birth	St. Pölten
Nationality	Austria
09/2006-06/2007	Civilian Service (*Zivildienst*) in lieu of military service (work as paramedic at the Austrian Red Cross Society in Krems/Donau)

Education

04/2008	Graduation, *Mag.rer.soc.oec.* (comparable with Master of Social and Economic Sciences), Vienna University of Technology, passed with distinction/honours
2007-2008	Study of Computer Science Management at Vienna University of Technology, 1040 Wien
12/2006	Graduation, *Dipl.-Ing.* (comparable with Master of Science), Vienna University of Technology, Thesis: *High Resolution Image Scan and Mosaicing*
2000-2006	Study of Computer Science at Vienna University of Technology, 1040 Wien
06/2000	School Leaving Examination (*Matura*), passed with distinction/honours

Education (continued)

1995-2000	Secondary College for Electronics (*HTBLVA*) - Special Training Focus Technical Computer Science, 3100 St. Pölten

Professional Work

since 03/2009	Senior Researcher at the Computer Vision Lab, Vienna University of Technology. He gained experience as a project collaborator in several projects at the Computer Vision Lab, VUT, dealing with the multi-spectral acquisition and restoration of ancient manuscripts. He is involved in lecturing at Vienna University of Technology, amongst others, Document Analysis. His research interests are Cultural Heritage Applications and Document Analysis Applications.
07/2007-03/2009	Project collaborator at the Insititute of Computer Aided Automation, Pattern Recognition and Image Processing Group, Vienna University of Technology for *The Sinaitic Glagolitic Sacramentary-Fragments* project, funded by the Austrian Science Fund under grant P19608-G12
10/2007-02/2008	Tutor at the Institute of Computer Aided Automation, Pattern Recognition and Image Processing Group, Vienna University of Technology
01/2006-04/2006	Digital Restoration of High Resolution Multispectral Images of a Palimpsest Manuscript (*Plautus, Codex Ambrosianus*) at the Institute of Computer Aided Automation, Pattern Recognition and Image Processing Group, Vienna University of Technology, Favoritenstraße 9, 1040 Vienna
2002	Tutor at the Institute of Computer Graphics and Algorithm of the Vienna University of Technology

Personal Interests

Rowing, Skiing, Biking, Photography