PROCEEDINGS

IS&T / SPIE
Electronic
Imaging
SCIENCE AND TECHNOLOGY

Document Recognition and Retrieval XXI

Bertrand Coüasnon
Eric K. Ringger
Editors

5–6 February 2014
San Francisco, California, United States

Sponsored by
IS&T—The Society for Imaging Science and Technology
SPIE

Volume 9021

IS&T
imaging.org

SPIE

PROCEEDINGS
IS&T / SPIE
Electronic
Imaging
SCIENCE AND TECHNOLOGY

Document Recognition and Retrieval XXI

Bertrand Coüasnon
Eric K. Ringger
Editors

5–6 February 2014
San Francisco, California, United States

Sponsored by
IS&T—The Society for Imaging Science and Technology
SPIE

Published by
SPIE

Volume 9021

Proceedings of SPIE 0277-786X, v. 9021

The papers included in this volume were part of the technical conference cited on the cover and title page. Papers were selected and subject to review by the editors and conference program committee. Some conference presentations may not be available for publication. The papers published in these proceedings reflect the work and thoughts of the authors and are published herein as submitted. The publishers are not responsible for the validity of the information or for any outcomes resulting from reliance thereon.

Please use the following format to cite material from this book:
Author(s), "Title of Paper," in *Document Recognition and Retrieval XXI*, edited by Bertrand Coüasnon, Eric K. Ringger, Proceedings of SPIE-IS&T Electronic Imaging, SPIE Vol. 9021. Article CID Number (2014)

ISSN: 0277-786X
ISBN: 9780819499387

Copublished by
SPIE
P.O. Box 10, Bellingham, Washington 98227-0010 USA
Telephone +1 360 676 3290 (Pacific Time) · Fax +1 360 647 1445
SPIE.org
and
IS&T—The Society for Imaging Science and Technology
7003 Kilworth Lane, Springfield, Virginia, 22151 USA
Telephone +1 703 642 9090 (Eastern Time) · Fax +1 703 642 9094
imaging.org

Printed in the United States of America.

Paper Numbering: Proceedings of SPIE follow an e-First publication model, with papers published first online and then in print and on CD-ROM. Papers are published as they are submitted and meet publication criteria. A unique, consistent, permanent citation identifier (CID) number is assigned to each article at the time of the first publication. Utilization of CIDs allows articles to be fully citable as soon as they are published online, and connects the same identifier to all online, print, and electronic versions of the publication. SPIE uses a six-digit CID article numbering system in which:

- The first four digits correspond to the SPIE volume number.
- The last two digits indicate publication order within the volume using a Base 36 numbering system employing both numerals and letters. These two-number sets start with 00, 01, 02, 03, 04, 05, 06, 07, 08, 09, 0A, 0B … 0Z, followed by 10-1Z, 20-2Z, etc.

The CID Number appears on each page of the manuscript. The complete citation is used on the first page, and an abbreviated version on subsequent pages. Numbers in the index correspond to the last two digits of the six-digit CID Number.

Conference Committee

Symposium Chair

Sergio R. Goma, Qualcomm Inc. (United States)

Symposium Cochair

Sheila S. Hemami, Northeastern University (United States)

Conference Chairs

Bertrand Coüasnon, Institut National des Sciences Appliquées de Rennes (France)
Eric K. Ringger, Brigham Young University (United States)

Conference Program Committee

Gady Agam, Illinois Institute of Technology (United States)
Sameer K. Antani, National Library of Medicine (United States)
Elisa H. Barney Smith, Boise State University (United States)
William A. Barrett, Brigham Young University (United States)
Kathrin Berkner, Ricoh Innovations, Inc. (United States)
Hervé Déjean, Xerox Research Centre Europe Grenoble (France)
Xiaoqing Ding, Tsinghua University (China)
David Scott Doermann, University of Maryland, College Park (United States)
Oleg D. Golubitsky, Google Waterloo (Canada)
Jianying Hu, IBM Thomas J. Watson Research Center (United States)
Ergina Kavallieratou, University of the Aegean (Greece)
Christopher Kermorvant, A2iA SA (France)
Laurence Likforman-Sulem, Telecom ParisTech (France)
Xiaofan Lin, A9.com, Inc. (United States)
Marcus Liwicki, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Germany)
Daniel P. Lopresti, Lehigh University (United States)
Umapada Pal, Indian Statistical Institute (India)
Sargur N. Srihari, University at Buffalo (United States)
Venkata Subramaniam, IBM India Research Laboratory (India)
Kazem Taghva, University of Nevada, Las Vegas (United States)
George R. Thoma, National Library of Medicine (United States)
Christian Viard-Gaudin, Université de Nantes (France)
Berrin Yanikoglu, Sabanci University (Turkey)
Richard Zanibbi, Rochester Institute of Technology (United States)
Jie Zou, National Library of Medicine (United States)

Additional Paper Reviewers

Alireza Alaei
Sukalpa Chanda
Rajiv Jain
Le Kang
Jayant Kumar
William B. Lund
Varun Manjunatha
Palaiahnakote Shivakumara

Session Chairs

1 Handwriting
Eric K. Ringger, Brigham Young University (United States)

2 Form Classification
Gady Agam, Illinois Institute of Technology (United States)

3 Invited Presentation I
Bertrand Coüasnon, Institut National des Sciences Appliquées de Rennes (France)

4 Text Recognition
Sameer Antani, National Library of Medicine (United States)

5 Handwritten Text Line Segmentation
Elisa H. Barney Smith, Boise State University (United States)

6 Invited Presentation II
Eric K. Ringger, Brigham Young University (United States)

7 Layout Analysis
Daniel P. Lopresti, Lehigh University (United States)

8 Information Retrieval
Xiaofan Lin, A9.com, Inc. (United States)

9 Data Sets and Ground-Truthing
Bertrand Coüasnon, Institut National des Sciences Appliquées de Rennes (France)
Eric K. Ringger, Brigham Young University (United States)

Panel Discussion: Data Sets and Ground-Truthing
Bertrand Coüasnon, *Moderator,* Institut National des Sciences Appliquées de Rennes (France)
Eric K. Ringger, *Moderator,* Brigham Young University (United States)

Introduction

On behalf of the Document Recognition and Retrieval XXI 2014 (DRR XXI) Program Committee, welcome to the Twenty-first Document Recognition and Retrieval conference being held in San Francisco, California, USA. DRR is held annually as part of the IS&T/SPIE Symposium on Electronic Imaging. It is one of the leading international conferences on document recognition, with a presence for related research on information retrieval and text mining.

This year we received 37 paper submissions. 28 papers were accepted, for an overall acceptance rate of 76%. Of the accepted papers, 21 were selected for oral presentation (57%), and 7 were selected for poster presentation (19%). We want to sincerely thank the Program Committee members and additional referees for helping us create a strong technical program. This year's program includes excellent tracks on Handwriting, Form Classification, Text Recognition, Handwritten Text Line Segmentation, Layout Analysis, Information Retrieval, and Data Sets and Ground-Truthing.

For the Best Student Paper Award, 8 authors have applied. We are grateful to Elisa H. Barney Smith (chair) and the award committee for carrying out the difficult task of choosing the winning paper. The winner will be announced in the EI Symposium-wide award ceremony on Wednesday morning of the conference. Google has provided $500 for the Best Student Paper Award for the third year, and we are truly grateful for their continued support of the conference.

This year we have two very interesting invited presentations. Ashok Popat and Ray Smith of Google Research will give a joint presentation on "OCR for Google Books" where many challenges arise from the scale and the diverse nature of the scanned corpus. Alexei A. Efros from the University of California, Berkeley, will give a talk entitled, "What makes Big Visual Data Hard?" and speak about problems encountered in collecting and using large visual data sets, based on his extensive research in computer vision.

We hope that you all have an excellent experience at DRR XXI!

Bertrand Coüasnon
Eric K. Ringger

Contents

HANDWRITTEN TEXT LINE SEGMENTATION

LAYOUT ANALYSIS

INFORMATION RETRIEVAL

DATA SETS AND GROUND-TRUTHING

INTERACTIVE PAPER SESSION

Writer Identification on Historical Glagolitic Documents

Stefan Fiel, Fabian Hollaus, Melanie Gau, and Robert Sablatnig

Computer Vision Lab
Vienna University of Technology
Vienna, Austria

ABSTRACT

This work aims at automatically identifying scribes of historical Slavonic manuscripts. The quality of the ancient documents is partially degraded by faded-out ink or varying background. The writer identification method used is based on image features, which are described with Scale Invariant Feature Transform (SIFT) features. A visual vocabulary is used for the description of handwriting characteristics, whereby the features are clustered using a Gaussian Mixture Model and employing the Fisher kernel. The writer identification approach is originally designed for grayscale images of modern handwritings. But contrary to modern documents, the historical manuscripts are partially corrupted by background clutter and water stains. As a result, SIFT features are also found on the background. Since the method shows also good results on binarized images of modern handwritings, the approach was additionally applied on binarized images of the ancient writings. Experiments show that this preprocessing step leads to a significant performance increase: The identification rate on binarized images is 98.9%, compared to an identification rate of 87.6% gained on grayscale images.

Keywords: Writer identification, historical documents

1. INTRODUCTION

This work deals with the automated identification of writers of ancient manuscripts. Currently, paleographers are performing this task mainly manually in order to localize, date or authenticate historical writings.[1,2] A large number of historical documents has been digitized in the past decade and has been made accessible to a growing number of users - for example see[3] or.[4] By automating the task of writer identification, it can be applied to a vast amount of historical documents and thus become a valuable tool for paleographers.

The historical manuscripts investigated in this work originate from the 10th to 11th centuries and are written in Glagolitic, the oldest Slavic script.[5] Writings belonging to five different manuscripts have been examined: Three books have been imaged at Mt. Sinai[6] and the remaining writings were photographed in libraries in Austria and Italy. Scholars found that the investigated manuscript leaves were written by seven scribes, whereby two manuscripts were written by several different hands, while another manuscript, which today is stored in two parts in different libraries, was written by one single scribe. In Figure 1 three examples of the manuscripts investigated are shown.

Recently, we proposed a writer identification method[7] that has been designed for Latin texts. The aim of the current work is to evaluate the applicability of this approach to the Glagolitic writings examined. To the best of our knowledge this is the first time that a writer identification approach is applied to Glagolitic handwritings. Compared to modern handwritings, the scribe identification task is complicated by the circumstance that the writings are partially in bad condition, since characters are faded-out and the documents are degraded by background clutter.

While the majority of the writer identification approaches deals with modern handwritings,[2] recently several approaches have been applied to historical writings: Bensefia et al.[8] use graphemes as features for the identification task. The approach is applied on a database consisting of modern handwritings, as well as on a data set, which contains documents written by 39 scribes and originating from the 19th century. The results obtained on the

Further author information:
Stefan Fiel, Fabian Hollaus, Melanie Gau, and Robert Sablatnig: E-mail: {fiel, holl, mgau, sab}@caa.tuwien.ac.at
Telephone: +43 (0)1 58801 18386

Document Recognition and Retrieval XXI, edited by Bertrand Coüasnon, Eric K. Ringger,
Proc. of SPIE-IS&T Electronic Imaging, Vol. 9021, 902102 · © 2014 SPIE-IS&T
CCC code: 0277-786X/14/$18 · doi: 10.1117/12.2042338

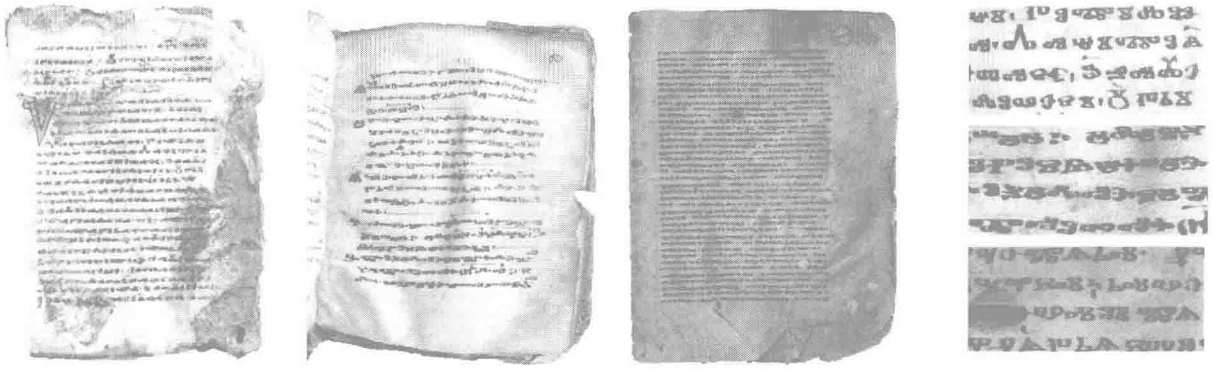

Figure 1. Examples taken from the dataset. (From left to right) Euchologium Sinaiticum folio 22 recto (Writer 3), Psalterium Demetrii Sinaitici folio 10 recto (Writer 2), Codex Clozianus folio 8 verso (Writer 7). Portions taken from the same images, from top to bottom: Writer 3, Writer 2, Writer 7.

historical database are significantly lower compared to the results gained on modern documents, which can be attributed to the presence of noise and slant amongst others.[8]

Bulacu and Schomaker[2] propose a writer identification system that combines textural and allographic features. The approach is tested on 70 medieval English documents that were written by 10 different scribes. The authors show that the combination of the feature groups leads to an increased performance.

Brink et al.[9] suggest to use the width of the ink trace along with directionality measurements as features for writer identification purposes. The approach is evaluated on two historical handwriting datasets of English and Dutch documents. Similar to the approaches mentioned above the dataset consists of manually cropped manuscript images.

Yosef et al.[1] note that historical documents are generally in a poor condition, which impedes a proper binarization. Therefore, the authors propose a multi-stage binarization approach that is especially designed for ancient manuscripts. Afterwards, several selected letters are automatically found and used for the writer identification, which is based on style analysis. For the classification task, K-nearest neighbors and Linear Bayes classifier are compared, whereby the latter mentioned is better suited for classification.

While the approaches mentioned above are applied on binarized document images, Bres et al.[10] perform writer identification on grayscale images. Therefore, the Hermite transformation is used for denoising and identification of handwritings. The system has been tested on 1438 historical documents stemming from 189 different writers, whereby the writings are written in different languages and alphabets.

Recently, Wolf et al.[3] proposed an unsupervised approach that is related to writer identification: In order to detect historical documents that have been taken from the same book, but have been dispersed, a semi-automatic clustering method is suggested. The approach is based on graphical models and image similarities are determined by using a Bag of Words approach. By using their technique about 1000 new connections between so far unrelated manuscript folios could be found and verified by scholars. Although the authors note that their approach is not entirely suited for writer identification, this result proves the valuable support that image processing techniques can provide for the paleographer.

On modern handwritings Li and Ding[11] proposed Grid Microstructure Features for Writer Identification. First the edges of the handwriting are extracted and on each pixel of the border the neighborhood is described. A feature vector is generated with the probability distribution of the different pixel pairs. This method won the "ICDAR 2011 Writer Identification Contest".[12] Jain and Doermann[13] are using K-adjacent segment features in a Bag of Words framework for writer identification. The relationship between sets of neighboring edges in an image is represented.

The current work is structured as follows. In Section 2 the methodology is introduced and in Section 3 the performance of the system is evaluated. Finally, in Section 4 a conclusion is drawn.

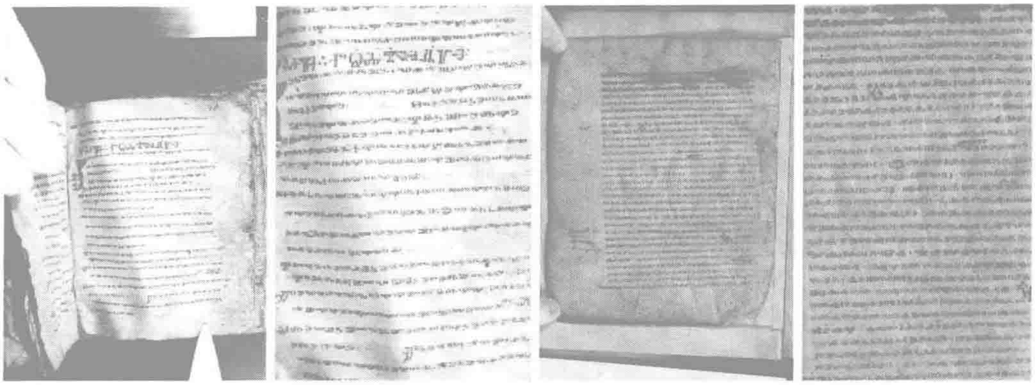

Figure 2. Input and result images of the cropping procedure. (From left to right) Psalterium Demetrii Sinaitici folio 47 recto (Writer 1). Corresponding result image. Codex Marianus folio 2 verso (Writer 6). Corresponding result image.

2. METHODOLOGY

As preprocessing step the document images are cropped to regions containing solely text. It should be noted that the manuscripts have been imaged in different places and their conditions are varying considerably. Thus, the background is heterogeneous - for example on several images there is a color chart in the background - or on other pages only a minor region of the parchment contains legible text, whereas the remaining characters are faded out. In order to remove the background, but also parchment regions containing no characters or text with a considerably low contrast, the following simple approach is used for the extraction of the main text block:

Text lines are found by using a text line detection method similar to the one proposed by Yosef et al.:[14] First, Local Projection Profiles (LPP) are applied to the image considered. Afterwards, the LPP image is filtered with a Gaussian column kernel and zero crossings of its first derivative are found. In a final step, a non-extremum suppression is performed in order to remove false positives. The zero crossings found are located at local minima and maxima and the minima encode the text lines. Afterwards, text lines with a length smaller than a predefined threshold are rejected and the bounding box containing the residual text lines is used for the cropping of the text region. Two examples for the cropping procedure are given in Figure 2. It can be seen that the main text block of both manuscripts is successfully extracted, but it has to be mentioned that only rectangular regions are extracted from the images. In order to reject for example decorative elements - like initials - a more sophisticated layout analysis technique would be more appropriate.

For the task of writer identification the method by Fiel and Sablatnig[7] is used. This method is designed to work without binarization, on pages with uniform background, modern handwriting, and if there is not only handwriting on the page, with a segmentation of areas which contain handwriting. In contrast, the manuscripts examined are corrupted by background clutter, faded-out ink, and water stains. Since experiments[7] have shown that when applying the method on binarized images it can keep up with other state of the art methods, it is also applied to binarized images of the dataset. This method is described shortly afterwards.

The Glagolitic writings are suffering from the aforementioned degradations, and thus the images are binarized by employing a binarization approach suited for historical documents that has been proposed by Su et al.[15] The algorithm starts with a calculation of a contrast image, whereby the pixels in this contrast image encode normalized intensity differences between the maximum and minimum gray levels within a local neighborhood. Afterwards, high contrast pixels, which are located at stroke boundaries, are found by applying a global Otsu[16] threshold. Those high contrast pixels are used in the final binarization step: A pixel is classified as a foreground pixel, if the following two requirements are met: First, the pixel should be in the near of a predefined number of high contrast pixels. Second, the pixel intensity must be smaller or equal than the mean intensity of the high contrast pixels in a local neighborhood window. An example for the binarization of a Glagolitic writing is given in Figure 3. It can be seen that the majority of the characters is successfully segmented, although the parchment portion is corrupted by clutter and the foreground to background contrast is varying.

The method for writer identification is based on the Fisher Kernels, introduced by Perronnin and Dance[17] and improved by Perronnin et al.,[18] which are calculated on Visual Vocabularies. The first step is the application

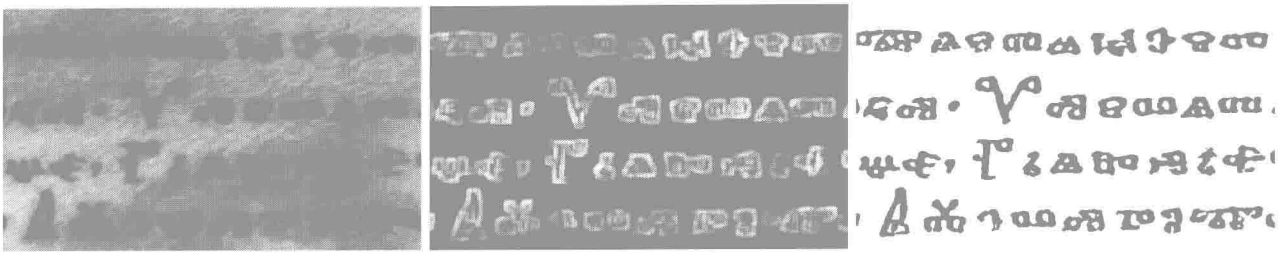

Figure 3. Binarization of a manuscript portion. (Left) Input image. (Middle) Contrast image. (Right) Binarization result.

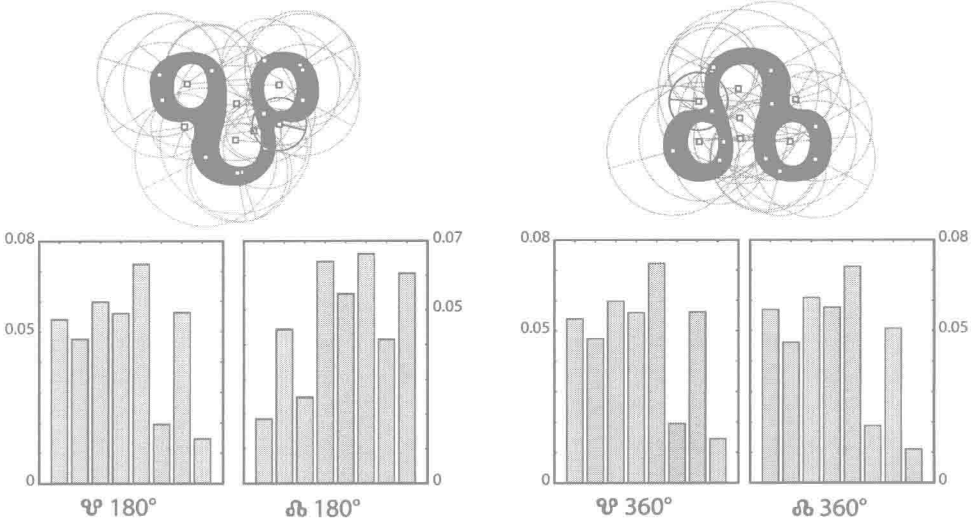

Figure 4. Two Glagolitic characters Ⱌ and Ⰴ and their corresponding SIFT features. The second row are the generated histograms of the two marked features (colored blue). In the bottom row on the right the SIFT features are calculated rotational invariant, thus they both generate a similar histogram. On the left side the features are calculated rotational dependent, which makes the two characters distinguishable. By courtesy of Diem and Sablatnig.[20]

of the Scale Invariant Feature Transform (SIFT) from Lowe.[19] These features have been modified by mirroring the angle of the keypoint, if the angle is larger than 180 degrees, proposed by Diem and Sablatnig[20] for character recognition of Glagolitic characters. Figure 4 shows two Glagolitic character and their SIFT features. It can be seen that the interest points are located in the middle of the circles and at the corners. Also the down sampled histograms, rotation invariant on the right side and rotationally dependent on the left side, of the highlighted SIFT features are shown. When the features are calculated rotation invariant, they are not distinguishable. In contrast, when the features are calculated rotational dependent they generate different histograms, allowing the distinction whether the features are located at the upper or the lower profile of the writing. It has been shown that the upper and lower profile of a writing is a discriminative feature for writer identification. Since the background contains a lot of noise the contrast threshold for the SIFT features has been set to a higher level and also the edge threshold is lowered to reduce the number of edge-like features - compared to the parameters used in.[7]

After the calculation of the SIFT features the visual vocabulary is generated. This is done on a separate training set to ensure the independence of the writers in the evaluation set. For performance reasons a Principal Component Analysis (PCA) is applied on the features of the training set to reduce the dimensionality from 128 to 64. The visual words are represented by means of Gaussian Mixture Models (GMM). The amount of the GMMs has to be set in advance. Experiments have shown that for this task the best number of GMM is 100. It is assumed that the generation process of all SIFT features in all images of the training set can be modeled by a probability density function.[18] The parameters of this density function can be estimated by an Expectation-Maximization(EM) algorithm. The EM algorithm estimates iteratively the three parameters of the

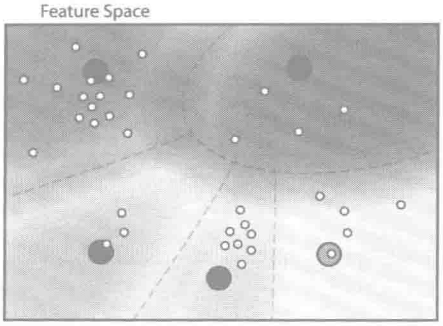

Figure 5. Schematic partitioning of the feature space with k-means (dashed lines) and with GMMs (colors). Since GMMs do not have strict borders the features space can be described more precisely.

different Gaussians. The advantage of using GMM for visual vocabularies instead of k-means, which is normally used for Bag of Words, is that the feature space can be described more precisely (see Figure 5). When using k-means for clustering strict borders are introduced and the distance of the feature to the cluster center and influences from other cluster centers are not taken into account. Since the GMM does not have any strict borders it overcomes these problems.

After estimating the parameters for the GMM the feature vector for each image can be generated. This is done by calculating the SIFT features $\mathcal{X} = \{x_t, t = 1 \dots T\}$ for one image and then applying the PCA with the parameters calculated for the training set. Afterwards the Fisher Kernel is applied. It is computed by[18]

$$\mathcal{G}_k^{\mathcal{X}} = \frac{1}{\sqrt{w_k}} \sum_{t=1}^{T} P(k|x_t)(\frac{x_t - \mu_k}{\sigma_k})$$

where $\mathcal{G}_k^{\mathcal{X}}$ is the feature vector for one specific distribution k. w_k are the weights of the k-th distribution, μ and σ are the means respectively the variation of the particular distribution. The results for all distributions are then concatenated, resulting in a feature vector of the image with ND-dimensions, where N is the number of distributions and D the dimension of the SIFT features (for this task $N = 100$ and $D = 64$ since we apply the PCA). As proposed by Perronnin et al. the vector is additionally raised by the power of 0.8. They also showed that the cosine distance is a natural measure of similarity for the Fisher Vector.

3. EVALUATION

The dataset for the evaluations contains 361 images with Glagolitic writing on it. Seven different writers have been identified on this pages.[5] Table 3 shows the distribution of the writers in the dataset - along with the associated manuscripts. It can be seen that by far the most documents in the dataset were written by Writer 1 (207) whereas Writer 6 has only 3 pages. With 58 documents Writer 5 has the second most documents in the dataset, followed by Writer 3 with 40 and Writer 7 with 24. Writer 2 has 22 documents. The documents of Writer 5 are from two different manuscripts and the documents of Writer 7 are stored at two different locations. Additionally, the approximated average character height (in pixel) and the corresponding standard deviation (std) are provided in Table 3, in order to enable a comparison between the character resolutions of each writer. The character height per writer is approximated by averaging the median height of the segmented characters within a binarized page. It should be noted that this approximation is dependent on the performance of the binarization method applied. It can be seen that character resolution of the pages belonging to Writer 7 is considerably lower compared to the remaining writers. Additionally, the character height is partially varying within a page due to decorative elements - such as initials - and warping effects. Exemplar pages with varying character heights can be seen in Figure 1. For generating the visual vocabulary 8 document fragments of unknown writers are used.

Table 1. Distribution of writers in the dataset.

Writer Id:	1	2	3	4	5	6	7
Psalterium Demetrii Sinaitici:	207	22	40				
Euchologium Sinaiticum:					7		
Missale Sinaiticum:				7	51		
Codex Marianus:						3	
Codex Clozianus:							24
Total :	207	22	40	7	58	3	24
Average character height:	27.7	31.5	32.3	34.4	28.8	29.3	21.4
Std. of character height:	2.5	5.0	3.8	9.5	4.2	1.5	3.4

For the evaluation of the writer identification the method of the ICDAR 2011 Writer Identification Contest[12] has been used. Each document in the database is taken as reference document and the distance to all other documents are calculated. Then two criteria, namely the soft and the hard criterion, are evaluated. For the soft criterion the first N most similar documents are observed. If one of these documents is from the same writer as the reference document then it is considered as a correct hit. The result is the percentage of the correct hits for all documents. For the hard criterion all N documents have to be from the same writer. The soft criterion is evaluated with N equals 1, 2, 5, and 10. For the hard criterion N is 2, 3, and 4. It has to be noted, that since Writer 6 has only 3 documents in the dataset the hard criterion with N equals 4 cannot be satisfied. These evaluation methods have been carried out on the cropped dataset and also on the binarized cropped dataset.

Table 2. Evaluation of the soft criterion (in %)

	Top 1	Top 2	Top 5	Top 10
grayscale dataset	87.6	89.8	92.5	93.6
binarized dataset	98.9	98.9	99.7	100.0

Table 2 shows the evaluation of the soft criterion on the dataset. It can be seen that the method performs better on the binarized dataset. The maximal difference is 11.3% for the Top 1, which means that the writer of 40 additional documents is identified correctly when using binarized data. Since the images have a non-uniform background, SIFT features are not only located on the writing itself, but also on the background, holes in the document, and border of water stains. Also because only a simple cropping method is used parts of the book cover can occur in the images and SIFT features are also calculated there. These features influence the performance on the non binarized dataset. In contrast, by binarizing a document image parts of the image which are not text are deleted.

Table 3. Evaluation of the hard criterion (in %)

	Top 2	Top 3	Top 4
grayscale dataset	80.9	78.1	75.3
binarized dataset	98.1	95.3	93.6

Table 3 shows the evaluation of the hard criterion. Again, the method performs better on the binarized dataset. The maximal difference is 18.3% for N equals 7. Like when using the soft criterion features, which are not located on the writing itself influence the performance on the non binarized dataset. This time the influence is greater, since all 4 most similar documents have to be from the same writer. Another possible effect is that documents with water stains are more likely to be considered as similar because of features which are located on the boarder of these stains.

Figure 6 shows plots of the different distances for all documents to two reference documents. The distances are calculated on the binarized dataset. For the left plot the reference document is page number 10 and for the right plot page number 353. The vertical red lines indicate a change of the writer. The reference document has a

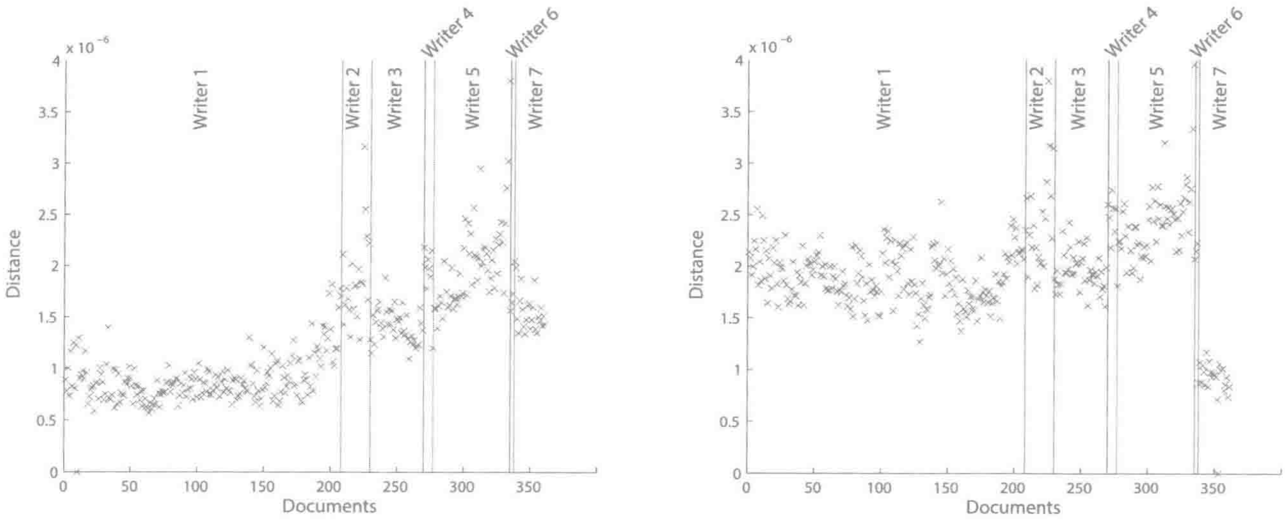

Figure 6. Distances of all images to two reference documents for the evaluation on the binarized dataset (left: page number 10, right: page number 353). The vertical red lines indicate a change of the writer.

distance of 0 to itself. For all other documents the distance is calculated using the cosine distance. It can be seen that documents written by the same writers as the particular reference document have a smaller distance than documents from other writers. There are also some outliers in the distances of other writers, which are pages where the binarization does not give exact results and parts of characters are lost or the border of the characters cannot be determined exactly.

Table 4. Mean precision of each writer with varying N. Cells which have no text are skipped because the corresponding writers do not have enough documents in the dataset for evaluation.

Top-N	Writer Id						
	1	2	3	4	5	6	7
1	100	90.1	100	100	96.6	100	100
2	100	88.6	97.5	100	96.6	100	100
3	100	90.1	96.7	100	93.7		100
5	99.9	83.4	97	100	93.1		100
10	99.9	77.7	94.3		91		100
15	99.9	61.2	91.5		86.8		100
20	99.8	51.4	89.4		82.1		98.8

Table 4 shows the mean precision per writer with varying N. For all documents of one writer the percentage of the documents in the first N in the ranking is calculated. For Writer 4 and Writer 6 the evaluation was skipped if N was higher than the number of their documents in the dataset.

The documents of Writer 1 have a big influence on the result of the evaluation because, as stated above, the largest contingent of documents was written by this scribe. Thus, a further evaluation has been carried out. This time the documents of Writer 1 are not taken as reference document, but the corresponding images remain in the dataset, so these documents can still occur in the most similar pages of the other writers.

Table 5 and Table 6 are showing the results of the evaluation on the databases where the documents of Writer 1 are not considered as reference document. Compared to the results with Writer 1 every evaluation has worse results. On the binarized dataset the performance drops 2.5% for the soft evaluation, for the hard evaluation the maximal drop is 7.9%. As expected, the fact that Writer 1 has by far the most documents in the dataset has an influence on the results, but at least for the evaluation on the binarized dataset the influence is tolerable. When

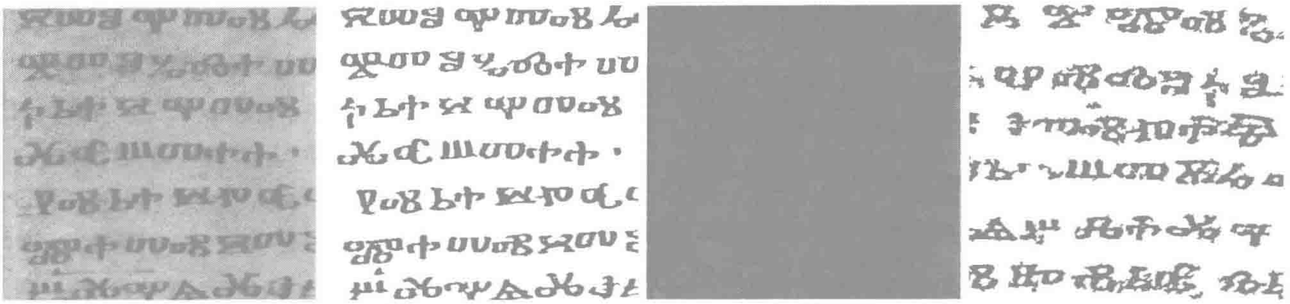

Figure 7. Two portions of Codex Clozianus. (From left to right) Folio 8 recto, stored at the City Museum in Trent, Italy. Corresponding binarization result. Folio 3 verso, stored at the Ferdinandeum museum in Innsbruck, Austria. Corresponding binarization result.

regarding the evaluation on the grayscale dataset the difference is significant. For the soft Top 1 evaluation the performance drops by 16.7% compared to the soft evaluation on the grayscale dataset with Writer 1, but compared to the results on the binarized dataset the difference is 19.9% worse. For the hard evaluation this gap increases. That means on the grayscale images the method has performed better on the images from Writer 1 than on the other images. The reason for this is, that the manuscript where the images of Writer 1 originate from has, compared to the other manuscripts, the most uniform background and also the ink is not faded out or blurred. Due to the binarization these factors do not influence the results on the binarized dataset, here the background is removed and the characters are clearly visible.

Table 5. Evaluation of the soft criterion (in %) without the pages of Writer 1 as reference documents.

	Top 1	Top 2	Top 5	Top 10
grayscale dataset	70.8	76.0	82.5	85.1
binarized dataset	97.4	97.4	99.4	100.0

Table 6. Evaluation of the hard criterion (in %) without the pages of Writer 1 as reference documents.

	Top 2	Top 3	Top 4
grayscale dataset	55.2	48.7	42.2
binarized dataset	95.5	89.0	85.7

One example for the correct identification of manuscripts written by the same scribe is given in Figure 7. The two portions shown - respectively the corresponding folios - are both belonging to a manuscript named Codex Clozianus, but the folios are stored in different libraries - namely libraries in Innsbruck, Austria and Trent, Italy. It can be seen that the writing on folio 8 recto has a higher contrast to the remaining background than the text on folio 3 verso. Additionally, the latter mentioned folio is corrupted by background stains. These circumstances lead to a poor binarization result. Nevertheless, the method is capable of determining that both folios were written by the same scribe, as can be seen in Figure 6 (right): In this plot, folio 8 recto of the Codex Clozianus is used as reference document and the distances to the remaining folios of the same codex - including folio 3 verso - are smaller than the distances to the other writings in the dataset.

4. CONCLUSION

This paper presented the first application of writer identification on Glagolitic documents. This is done by using the Fisher Vector on visual vocabularies. First SIFT features are calculated on the document image and with help of GMM, which has been generated in advance on a training set, the Fisher Vector can be generated. The documents are then ordered by their similarity to a reference document and the evaluations are carried out using the nearest neighbors. As input image a dataset with 361 images with Glagolitic writing is used, which were preprocessed by cropping the area which contains text. On the complete dataset the best performance, with 98.9% correct first nearest neighbor, was achieved by applying a binarization to the image. Since the method