ZfdG

Zeitschrift für digitale Geisteswissenschaften

Artikel aus:
Zeitschrift für digitale Geisteswissenschaften

Titel:
Automatic Writer Identification in Historical Documents: A Case Study

Autor/in:
Vincent Christlein

Kontakt: vincent.christlein@fau.de
Institution: Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg
GND: ORCID: 0000-0003-0455-3799

Autor/in:
Markus Diem

Kontakt: diem@caa.tuwien.ac.at
Institution: Computer Vision Lab, TU Wien
GND: ORCID: 0000-0002-5048-5128

Autor/in:
Florian Kleber

Kontakt: kleber@caa.tuwien.ac.at
Institution: Computer Vision Lab, TU Wien
GND: ORCID: 0000-0001-8351-5066

Autor/in:
Günter Mühlberger

Kontakt: guenter.muehlberger@uibk.ac.at
Institution: Digitisation and Digital Preservation (DEA), German Language and Literature, Innsbruck University
GND: ORCID: 0000-0002-7068-7261

Autor/in:
Verena Schwägerl-Melchior

Kontakt: verena.schwaegerl-melchior@uni-graz.at
Institution: Institut für Sprachwissenschaft, Karl-Franzens-Universität Graz
GND: ORCID: 0000-0002-8303-1843

Autor/in:
Esther van Gelder

Kontakt: esther.van.gelder@huygens.knaw.nl
Institution: Descartes Centre (Universiteit Utrecht) / Huygens ING (KNAW)
GND: ORCID: 0000-0002-4505-6302

Autor/in:

Andreas Maier

Kontakt: andreas.maier@fau.de
Institution: Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg
GND: ORCID: 0000-0002-9550-5284

DOI des Artikels:
10.17175/2016_002

Nachweis im OPAC der Herzog August Bibliothek:
zurzeit nicht verfügbar

Erstveröffentlichung:
05.07.2016

Letzte Überprüfung aller Verweise:
27.06.2016

GND-Verschlagwortung:
Mustererkennung | automatische Klassifikation | Handschrift (Grafologie) | Computergestütztes Verfahren |

Zitierweise:
Vincent Christlein, Markus Diem, Florian Kleber, Günter Mühlberger, Verena Schwägerl-Melchior, Esther van Gelder, Andreas Maier: Automatic Writer Identification in Historical Documents: A Case Study. In: Zeitschrift für digitale Geisteswissenschaften. 2016. PDF Format ohne Paginierung. Als text/html abrufbar unter DOI: 10.17175/2016_002.

Vincent Christlein, Markus Diem, Florian Kleber, Günter Mühlberger,
Verena Schwägerl-Melchior, Esther van Gelder, Andreas Maier

Automatic Writer Identification in Historical Documents: A Case Study

Abstracts

Die automatische Schreiberidentifizierung erlangte viel Aufmerksamkeit während des letzten Jahrzehnts. Jedoch beschränkte sich die meiste Arbeit auf zeitgenössische Vergleichsdatensätzen. Diese Datensätze beinhalten typischerweise keinerlei Rauschen oder Artefakte. In dieser Arbeit wird analysiert ob die aktuell beste Methode der automatischen Schreiberidentifizierung ähnlich gut auf historischen handgeschriebenen Daten funktioniert. Im Gegensatz zu zeitgenössischen Daten enthalten historische Daten oft Artefakte wie Löcher, Risse oder Wasserflecken, was eine zuverlässige Identifikation fehleranfällig macht. Wir führten Experimente an zwei großen Briefkollektionen mit gegebener Authentizität durch und erlangten vielversprechende Ergebnisse von 82% und 89% TOP-1 Genauigkeit.

In recent years, Automatic Writer Identification (AWI) has received a lot of attention in the document analysis community. However, most research has been conducted on contemporary benchmark sets. These datasets typically do not contain any noise or artefacts caused by the conversion methodology. This article analyses how current state-of-the-art methods in writer identification perform on historical documents. In contrast to contemporary documents, historical data often contain artefacts such as holes, rips, or water stains which make reliable identification error-prone. Experiments were conducted on two large letter collections with known authenticity and promising results of 82% and 89% TOP-1 accuracy were achieved.

1. Introduction

Similar to someone’s face or fingerprints, handwritten text can serve as a biometric identifier. In this sense, writer[1] identification is a field of research that is concerned with identifying the author of a handwritten text document, given a set of known authors.

Among the document analysis community, Automatic Writer Identification (AWI) has gained significant attention in the last decade. Several competitions with the objective of identifying a given scribe were organised at prominent conferences such as the International Conference on Document Analysis and Recognition (ICDAR) and the International Conference on Frontiers in Handwriting Recognition (ICFHR). Nevertheless, these competitions were conducted on contemporary or even artificial data sets.[2] Signature verification, a research field mainly motivated by business and forensic applications, can also be mentioned in this context.

Only recently has AWI also been applied to historical documents. For example, the project DAmalS (Datenbank zur Authentifizierung mittelalterlicher Schreiberhände)[3], which was funded

[1] Note that ›scribe‹ and ›writer‹ are used interchangeably throughout this article.
[2] See, for example, Louloudis 2011.
[3] https://homepage.uni-graz.at/de/wernfried.hofmeister/projekte/damals/.

by the Austrian Science Fund, successfully identified the correct number of scribal hands in the manuscripts of Hugo von Montfort with a high grade of certainty.[4] Similarly, Flecker et al.[5] analysed the number of scribal hands in two manuscripts. Additionally, they showed the effectiveness of their method using a corpus of 60 manuscripts written by twelve scribes consisting of about 4500 pages. Here, the authors reached up to 100% accuracy when they left out one complete manuscript and tested it against all other manuscripts.

Since we are dealing with historical documents, the focus of our work is similar; however, our dataset consists of hundreds rather than dozens of scribal hands. Actually we employ two large datasets of letters: the Clusius dataset and the Schuchardt Dataset (see figure 1). Furthermore, our evaluation is not limited to a document (manuscript) as a whole, since we also ran our evaluation on the basis of single pages as a metric unit.

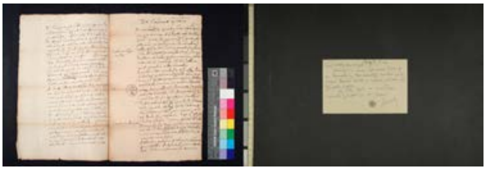


figure 1: Left: Example image from the Clusius dataset: Letter from Johannes Brambach to Carolus Clusius dated August 21, 1586 (img-id: 896664_CLUY073-001-b, VUL 101). Image used with permission of the Digital Special Collections of Leiden University Library. Right: Example image of the Schuchardt dataset: Letter from Adolf Zauner to Hugo Schuchardt, dated Februrary 27, 1912 (img-id 12977). Image used with permission of the University Library Graz, Department for Special Collections, legacy of Hugo Schuchardt.

2. Methodology

Before AWI methods can be applied, several preprocessing steps usually need to be carried out. First of all, the actual text regions of a page image must be separated from regions containing non-text elements. For example, the scanning protocols of libraries involve the addition of a color pattern and a ruler so that the color information and real scale can be reconstructed. However, for the purpose of document analysis, these parts of the page image need to be removed. Additionally, artefacts in the background of the document, such as folds or graphical illustrations, are not relevant for writer identification since we only want to analyse the text. Thus, we first detect the text areas in the document image, as described in more detail in the following section (figure 2). In the second step, the colour of the documents, or, more precisely, of the regions containing text, is reduced to 1 bit, i.e., the text regions are binarized (figure 3). As a result, the contour of the script is represented as a black line (figure 4). In the third step, features of the contour are extracted. A background model is computed from all feature descriptors of the whole dataset (or training set), and this model is in turn further used to compute global image descriptors for each page of the collection. Note that this process is

[4] Hofmeister 2009.
[5] Flecker 2014; Flecker 2015.

very similar to speaker identification in audio signals.[6] In addition to the background model, individual writer models can be computed from pages of known authorship. These models can then be used for querying the collection and assigning the correct writer to the questioned document image. Alternatively, if we have a large dataset with unknown authors and no reference models, we can group pages according to their similarity (clustering).

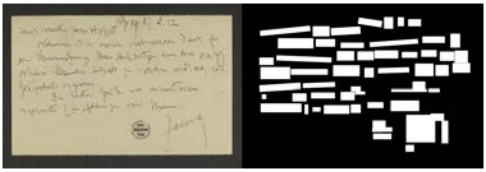


figure 2: Text detection mask (left),overlaid with the binarized input image to generate the contours (right). Local feature descriptors are extracted at the contours and aggregated to form a global image descriptor.

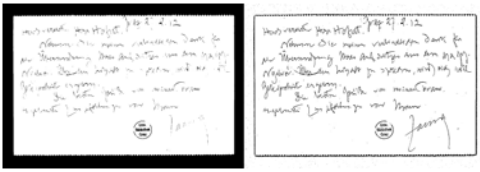


figure 3: Binarization examples obtained from Otsu’s method (left) and obtained from Bradley’s method (right).

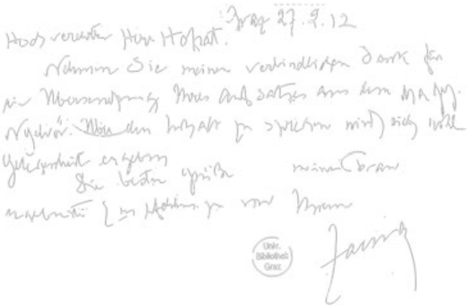


figure 4: Contour output using the image mask and the binarization result.

2.1 Text Detection

A bottom-up approach is used to analyze the page image in order to detect the text regions and to separate them from any other type of region, such as graphics or noise. This approach demonstrates more robustness with respect to noise or poorly pre-processed images than top-down approaches. First, the characters are grouped to words in the binary image using Local Projection Profiles (LPPs). Issues arising from merged ascenders and descenders between text lines are resolved using a rough text line estimation based on a first derivative anisotropic Gaussian filtering. Then, continuous local maxima are detected in the filtered image in order to

[6] Bocklet 2008.

split text lines that are merged. After these processing stages, the contour of words is known. In order to maintain processing speed and the complexity of subsequent algorithms, it is preferable to represent words using an enclosing rectangle rather than their contour. We introduced profile boxes (see figure 5) that are computed by robustly fitting lines to a word's upper and lower profile. Having detected both lines, the profile box is defined to have the mean angle of both lines, a height which is the mean distance between the lines, and a width corresponding to the maximal length of both lines. A detailed description of the text detection is presented in the work of Diem et al.[7] For the purpose of writer identification, we employ the text detection masks that are subsequently dilated (i.e. enlarged) with a rectangular shape of size 25x25.

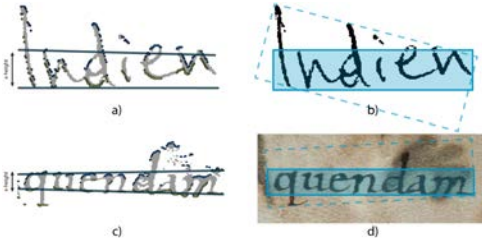


figure 5: Upper and lower profile a), c) with the corresponding upper and lower profile line. In b) and d) the resulting bounding box (dotted rectangle), minimum area rectangle (dashed line), and the proposed profile box (solid rectangle) are illustrated. Note that the profile box resembles the correct word orientation while having a minimal background.

2.2 Writer Identification

We used the writer identification method of Christlein et al.[8] This algorithm achieved state-of-the-art accuracy on all writer identification benchmarks and won the ICDAR 2015 competition in multi-script writer identification using the QUWI dataset.[9]

The following paragraph provides a brief outline of the algorithm; for more details please refer to the original work by Christlein et al. In contrast to typical allograph-based writer identification methods[10], this algorithm does not rely on keypoint locations at which feature descriptors like SIFT are extracted. Instead, feature descriptors are computed densely from the contours of the script. The contours are computed from the binarized input image. If the input image has not been binarized yet, the method of Otsu[11] is applied. At each location, rotational dependent Zernike moments up to the 11th degree are computed. Zernike moments are excellent shape descriptors and have been used in other related fields. A background model is computed from all Zernike moments of the training set using k-means. With the help of this background model, the feature descriptors of each document of the dataset are aggregated to

[7] Diem 2011.
[8] Christlein 2015.
[9] Djeddi 2015.
[10] Jain 2014; Fiel 2013; Flecker 2014; Flecker 2015.
[11] Otsu 1979.

form one global feature descriptor per document. This encoding step is achieved by computing Vectors of Locally Aggregated Descriptors (VLAD).[12] To improve accuracy, this step is repeated multiple times (up to five times) with different random initializations of k-means. The different VLAD encodings are then concatenated and jointly de-correlated and dimensionality-reduced (400 components) using Principal Component Analysis (PCA).

In order to work with historical data, we modified this approach regarding its binarization step. Otsu's binarization is a global binarization method. It finds an optimal threshold to separate foreground from background using all pixel information of the image. However, a global threshold is often suboptimal when dealing with noisy data. For example, non-uniform illumination, or the large amount of zero-pixels from the cardboard surrounding the document, generates non-optimal thresholds. Thus, we employ the local threshold-based method of Bradley et al.,[13] which can be computed efficiently. An example image where Otsu's method fails is shown in figure 3, where much script vanished due to bad contrast. However, Bradley's algorithm successfully binarized the input image.

3. Evaluation

3.1 Dataset 1: Clusius

The Clusius dataset consists of 1600 letters written to and by one of the most important sixteenth century botanists, Carolus Clusius (1526-1609). It was provided by the Huygens Institute for the History of the Netherlands (Royal Netherlands Academy of Arts and Sciences), which is creating a digital edition of Clusius' correspondence in the collaborative editing tool eLaborate.[14]

The letters were written by 330 different authors, in 6 different languages, and from 12 European countries. A unique feature of this correspondence is that the authors come from different backgrounds, including scholars, physicians and aristocrats, but also chemists and gardeners, and there are many women. This variety provides an extremely diverse glimpse into linguistic and handwriting characteristics in the second half of the sixteenth century, including the clear Latin handwriting of Clusius himself, or the almost (for us) unreadable handwriting in Viennese dialect of a lower-Austrian noblewoman. The correspondence is mainly about the exchange of plants and information, but also comprises news on politics, friends and family, court gossip, etc. An example image can be seen in figure 1 (left).

While the total number of the preserved correspondence comprises 1600 letters, only the 1175 letters that are preserved in Leiden University Library have been digitized and therefore

[12] Jegou 2012.
[13] Bradley 2007.
[14] See the editing project http://www.dwc.knaw.nl/biografie/clusius/digital-edition-of-the-clusius-correspondence/. For the life and work of Clusius and his many correspondents, see esp. Egmond 2010.

could be used for the experiment. All scribes were already identified, though some letters have co-authors, and most aristocrats had secretaries.

3.2 Dataset 2: Schuchardt

The Schuchardt dataset was provided by the project Network of Knowledge[15] dedicated to the edition of the papers and in particular the correspondence of the linguist Hugo Schuchardt (1842–1927).

This eminent scholar, who displayed an extraordinary networking capability, left more than 13,000 letters addressed to him from more than 2000 individual writers and ca. 100 institutions over a timespan of 77 years (1850-1927). The letters are in more than 20 languages.[16] A noteworthy part of this correspondence, which is preserved at the Library of the University of Graz (Special Collections), has already been edited on the website of the *Hugo Schuchardt Archiv*, [17] and several editions are still ongoing and planned. The dataset evaluated here, consisting of 13,569 single pages from 193 different scribes, is a small subset of the correspondence which has been manually categorized by correspondents in the early 1990s by Michaela Wolf.[18] An example image is shown in figure 1 (right). Three aspects of the dataset are of particular interest for writer recognition: a) the high number of already identified scribes (although only a small part of the material of Schuchardt's legacy has been chosen); b) the presence of correspondence lasting over decades, including therefore variations in the handwriting of a single scribe over time and historical changes in writing systems in a specific area (e.g., the use of German cursive) and c) graphical and scriptural variation within the documents issued by multilingual scribes depending on the languages chosen.

Although Schuchardt's correspondence has been categorized manually as mentioned above, reliable writer recognition could be used on Schuchardt's papers to attribute loose sheets and notes preserved within the section »Werkmanuskripte« which were often sent to Schuchardt by his correspondents but were separated from the letters originally containing them and do not bear any signature. In the future, the testing and improving of writer recognition on a large dataset of identified scribes of a heterogeneous collection might pave the way for its use in the framework of inventorying handwritten archives.

3.3 Evaluation Protocol and Error Metrics

The dataset images are evaluated in a leave-one-page-out scheme; from the individual test set, one query image is tested against all remaining ones, resulting in an ordered list in which the first returned image has the highest probability of having been written by the same

[15] The project directed by Bernhard Hurch (Institut für Sprachwissenschaft, Karl-Franzens-Universität Graz) is funded by the Austrian Science Funds (FWF) (project number P 24400-G15; 2012–2015).
[16] The inventory is in Wolf 1993; for information about Hugo Schuchardt and his network, see Hurch 2007 and Hurch 2009.
[17] http://schuchardt.uni-graz.at.
[18] Wolf 1993.

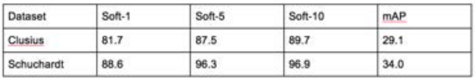
author. From the retrieval lists, we can compute the accuracy of the algorithm. As error metrics, we use the ›Soft‹ Top-k, ›Hard‹ Top-k and the mean Average Precision (mAP). The Soft Top-k denotes the precision at rank k. In other words, it describes the probability that the correct writer is among the first k retrieved documents. In contrast, the Hard Top-k rate gives the probability that the first k documents are written by the same author as the query document. The mean Average Precision is a metric commonly used in information retrieval. For each query document, the average precision of relevant documents in the retrieval list is computed. Therefore, the precision is given by the number of relevant documents in the retrieval list up to rank k divided by k.

3.4 Experiments

First, we evaluated the datasets as a whole, meaning that we did not separate the datasets into independent training and test sets. Thus, the background model stems from the same data to be evaluated. We decided to use all pages from each scribe who contributed at least two pages. For the Clusius dataset, this resulted in 2029 pages from 182 different scribes. The Schuchardt dataset has 12,846 pages written by 193 different scribes. Note that we discarded several images from both datasets that are not associated with the actual letters (or postcards).

Initial results showed that the TOP-k accuracies are very promising (table 1). In 82% and 89% of all cases, the author of the query page was identified correctly for the Clusius and the Schuchardt datasets, respectively. The high TOP-10 rates of 90% and 97% also suggest a quick detection of the correct writer in the shortlists.

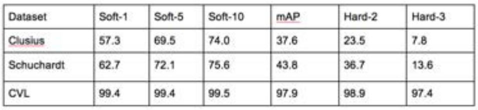
However, the rather low mAP values of 29% and 34% indicate that there are pages in the datasets which are very difficult to identify. Most likely this is related to images containing very little text, such as letters or postcards containing only the address. Also note that the number of documents per author was very unbalanced, so that the Clusius dataset has six authors who each contributed more than 50 images. Thus, document pages from these authors are likely to be identified more easily than pages from authors who appear more rarely in the dataset.



Dataset	Soft-1	Soft-5	Soft-10	mAP
Clusius	81.7	87.5	89.7	29.1
Schuchardt	88.6	96.3	96.9	34.0

table 1: Identification results in percentages using the whole datasets for the evaluation.

To allow for a fairer comparison, we conducted another set of experiments in which we picked exactly four pages from each author. The remaining pages of these authors and images from authors who contributed less than four pages were used solely to train the background model. Table 2 shows how the identification accuracy drops in comparison to the full datasets. The TOP-1 accuracy scores are about 60% and the correct author is identified only in about 75% of the cases among the first ten most similar images. Not surprisingly, this indicates once again that the more training images we have, the easier it becomes to identify the correct scribe.



Dataset	Soft-1	Soft-5	Soft-10	mAP	Hard-2	Hard-3
Clusius	57.3	69.5	74.0	37.6	23.5	7.8
Schuchardt	62.7	72.1	75.6	43.8	36.7	13.6
CVL	99.4	99.4	99.5	97.9	98.9	97.4

table 2: Identification results using four page images from each author, compared with the results using the CVL dataset.

This experiment also makes it possible to compare the results with a clean competition dataset.[19] Row three of table 2 shows the results of the method applied to the CVL datasets.[20] Apparently, writers from a contemporary and clean dataset are much easier to identify than writers from a historical dataset. The main reason lies in the feature extraction step. As described above, features are extracted densely at the contours, and errors in contour detection are therefore directly reflected in the feature descriptors. Error sources are various, such as when the text detection fails and feature descriptors are computed at the wrong contours (figure 6). Furthermore, false contours appear when the text binarization fails. Sometimes the datasets do not contain only one scribal hand but also annotations from another hand, as in the image depicted in figure 6. As a consequence, the global image descriptor encodes both scribal hands as one, which makes reliable identification nearly impossible.

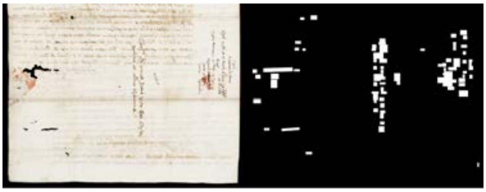


figure 6: Example of text detection failure: Letter from Simón de Tovar to Carolus Clusius, dated March 19, 1596 (img-id: CLUY030-001-c, VUL 101). Image used with permission of the Digital Special Collections of Leiden University Library.

4. Potential Use Cases and Applications

We see a number of highly interesting use cases for AWI and would like to discuss some of them:

(1) First of all, AWI offers the chance to query large collections of archival documents based on the writing style of a single person. Without having seen or transcribed the page images, a user of this type of search interface would be able to access the writings of a specific hand or scribe. This method provides completely new ways to search an archive and supports the idea that it is more important to digitize large collections of documents and to enrich them automatically, rather than combining manual metadata labeling with the digitization workflow

[19] Cf. Christlein 2015, p. 909.
[20] Kleber 2013.

itself. One could therefore think of applications in which a user could collect the writings of a person of interest within several archives using a pre-trained model of this hand, for example. A prototype implementation of such an AWI based search and retrieval tool will become available to the public via the Transkribus platform.[21] Document images that are uploaded to the platform will be processed automatically in the way described above: text regions will be detected, binarized, and descriptors for the contours will be computed for every single page image. Additionally, a user will be able to train specific AWI models for documents in which the scribe is known. Based on these two steps, a user could search thousands of page images within the platform in the following ways:

(a) *Give me all those page images which are very likely written by the same hand as my query page* (query by example). The result will be a ranked list of page images according to their distance from the query page.

(b) Another - similar - query could be: *I have trained a writer model based on several dozens or hundreds of page images. Give me all page images which are similar to this writer model (query by writer).*

(c) If several writer models are available, another query could be: *Order all page images according to the writers contained in the collection.* It must be emphasized that in all three cases, the user will need to deal with probabilities and may need to review and discard a relatively large number of false alarms but would still be able to access the writings of a specific person in a previously unknown way.

(2) A second use case deals with digital editions: As a matter of fact, many documents which stem from famous persons are not written by these famous persons themselves. To mention just one prominent example: Jeremy Bentham (1748–1832) left behind tens of thousands of pages of unpublished works, personal papers, and correspondence. These papers are currently being transcribed and published by the transcribe Bentham project[22] involving hundreds of volunteers in a crowd-sourced setting. A large portion of these papers was written by several secretaries. In order to identify an individual hand (especially in cases where the name is not known), AWI could be a support tool to make editorial work easier and could also assist editors in making an informed decision.

(3) A third use case takes a similar direction. The description of different hands is currently often based on intrinsic knowledge of the scholar and simple examples. A scholar may ›know‹ a specific hand, and describe its features, such as how certain characters are formed, but there are hardly any objective measures that can be used to identify a specific hand. Since AWI computes specific descriptors for every text region, the distance between text regions (in our case, pages or documents, but also smaller units such as blocks or lines) can be computed and may provide an objective view of a given hand.

[21] Cf. http://www.transkribus.eu/. The platform will further be developed in the H2020 Project READ (Recognition and Enrichment of Archival Documents) coordinated by the University of Innsbruck.

[22] Cf. http://blogs.ucl.ac.uk/transcribe-bentham/.

5. Discussion and Conclusion

In this article, we provided a first large-scale analysis for Automatic Writer Identification. Our results reveal that the current algorithm shows significant potential for practical use. At this stage, the technology can already be used to search through a large amount of data to give a short list of authors most likely to be the correct ones. This method can dramatically reduce the manual effort connected with searching for the writing of a specific person or for clustering a given collection. More broadly, it provides new ways to search through archive material and confirms that digitisation of large amounts of archival material provides significant benefits even without cost intensive manual metadata editing.

Further research should be dedicated to a better layout analysis and algorithms that are less error prone regarding faulty binarization. Therefore, it would be helpful to provide the computer scientists with more training and evaluation data to test their algorithms. With more data, new technologies like ›deep learning‹[23] may become possible in this field.

[23] ›Deep learning‹ is commonly associated with deep neural networks that currently achieve the highest results for image classification and object detection.

Bibliography

Tobias Bocklet / Andreas Maier / Elmar Nöth: Age Determination of Children in Preschool and Primary School Age with Gmm-based Supervectors and Support Vector Machines/Regression. In: Proc. Text, Speech and Dialogue (2008). Heidelberg, Berlin, pp. 253–260. [Nachweis im GBV]

Derek Bradley / Gerhard Roth: Adaptive Thresholding using the Integral Image. In: Journal of Graphics, GPU, and Game Tools, Vol. 12 (2007), Issue 2, pp. 13–21. [Nachweis im GBV]

Vincent Christlein / David Bernecker / Andreas Maier / Elli Angelopoulou: Writer Identification Using VLAD encoded Contour-Zernike Moments. In: Proceedings of the 13th International Conference on Document Analysis and Recognition (2015). Nancy, France, pp. 906–910. [online]

Markus Diem / Florian Kleber / Robert Sablatnig: Text Classification and Document Layout Analysis of Torn Documents. In: Proceedings of the 11th International Conference on Document Analysis and Reconstruction (2011). Bejing, China, pp. 1181–1184. [online]

Chawki Djeddi / Somaya Al-Maadeed / Abdeljalil Gattal / Imran Siddiqi / Labiba Souici-Meslati / Haikal El Abed: ICDAR2015 Competition on Multi-script Writer Identification and Gender Classification using ›QUWI‹ Database. In: Proceedings of the 13th International Conference on Document Analysis and Recognition (2015). Nancy, France, pp. 1191–1195. [online]

Florike Egmond: The World of Carolus Clusius: Natural History in the Making, 1550–1610 (Perspectives in Economic and Social History). London 2010. [Nachweis im GBV]

Daniel Fecker / Volker Märgner / Torsten Schaßan: Vom Zeichen zur Schrift: Mit Mustererkennung zur automatisierten Schreiberhanderkennung in mittelalterlichen und frühneuzeitlichen Handschriften. In: Grenzen und Möglichkeiten der Digital Humanities. Hg. von Constanze Baum / Thomas Stäcker. 2015 (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 1). text/html Format. DOI: 10.17175/sb001_008. [Nachweis im GBV]

Daniel Fecker / Abed Asi / Volker Märgner / Jihad El-Sana / Tim Fingscheidt: Writer Identification for Historical Arabic Documents. In: Proceedings of the International Conference on Pattern Recognition (ICPR). Stockholm 2014, pp. 3050–3055. [Nachweis im GBV]

Stefan Fiel / Robert Sablatnig: Writer Identification and Writer Retrieval using the Fisher Vector on Visual Vocabularies. In: Proceedings of the 12th International Conference on Document Analysis and Recognition (2013), Washington DC, USA, pp. 545-549. [Nachweis im GBV]

Wernfried Hofmeister / Andrea Hofmeister-Winter / Georg Thallinger: Forschung am Rande des paläographischen Zweifels: Die EDV-basierte Erfassung individueller Schriftzüge im Projekt DAmalS. In: Kodikologie und Paläographie im digitalen Zeitalter. Hg. von Malte Rehbein, Patrick Sahle, Torsten Schaßan. Norderstedt 2009. (= Schriften des Instituts für Dokumentologie und Editorik 2), pp. 261–292. [Nachweis im GBV]

Bernhard Hurch: Schuchardt, Hugo Ernst Mario. In: Neue Deutsche Biographie, Band 23. Berlin 2007, pp. 623–624. [online]

Bernhard Hurch: Einleitung: Prolegomena zum Briefprojekt. In: Grazer Linguistische Studien, 72 (2009), pp. 5–17. [online]

Rajiv Jain / David Doermann: Combining Local Features for Offline Writer Identification. In: Proceedings of 14th International Conference on Frontiers in Handwriting Recognition (2014), Greece, pp. 583–588. [Nachweis im GBV]

Hervé Jégou / Florent Perronnin / Matthijs Douze / Jorge Sánchez / Patrick Pérez / Cordelia Schmid: Aggregating Local Image Descriptors into Compact Codes. In: IEEE Transactions on Pattern Analysis and Machine Intelligence Vol. 34 (2012), No. 9, pp. 1704–1716. [Nachweis im GBV]

Florian Kleber / Stefan Fiel / Markus Diem / Robert Sablatnig: CVL-DataBase: An Off-Line Database for Writer Retrieval, Writer Identification and Word Spotting. In: Proceedings of the 12th International Conference on Document Analysis and Recognition (2013). Washington DC, USA, pp. 560–564. [Nachweis im GBV]

Georgios Louloudis / Nikolaos Stamatopoulos / Basilis Gatos: ICDAR 2011 Writer Identification Contest. In: IEEE International Conference on Document Analysis and Recognition (2011), pp. 1475–1479. DOI: 10.1109/ICDAR.2011.293

Nobuyuki Otsu, A Threshold Selection Method from Gray-Level Histograms. In: IEEE Transactions on Systems, Man, and Cybernetics, 9 (1), 1979, pp. 62–66. 10.1109/ICDAR.2011.293

Michaela Wolf: Der Hugo Schuchardt Nachlaß. Schlüssel zum Nachlaß des Linguisten und Romanisten Hugo Schuchardt (1842–1927). Graz: Leykam 1993. [Nachweis im GBV]

Figures

Abb. 1: figure 1: Left: Example image from the Clusius dataset: Letter from Johannes Brambach to Carolus Clusius dated August 21, 1586 (img-id: 896664_CLUY073-001-b, VUL 101). Image used with permission of the Digital Special Collections of Leiden University Library. Right: Example image of the Schuchardt dataset: Letter from Adolf Zauner to Hugo Schuchardt, dated Februrary 27, 1912 (img-id 12977). Image used with permission of the University Library Graz, Department for Special Collections, legacy of Hugo Schuchardt.

Abb. 2: figure 2: Text detection mask (left),overlaid with the binarized input image to generate the contours (right). Local feature descriptors are extracted at the contours and aggregated to form a global image descriptor.

Abb. 3: figure 3: Binarization examples obtained from Otsu’s method (left) and obtained from Bradley’s method (right).

Abb. 4: figure 4: Contour output using the image mask and the binarization result.

Abb. 5: figure 5: Upper and lower profile a), c) with the corresponding upper and lower profile line. In b) and d) the resulting bounding box (dotted rectangle), minimum area rectangle (dashed line), and the proposed profile box (solid rectangle) are illustrated. Note that the profile box resembles the correct word orientation while having a minimal background.

Abb. 6: figure 6: Example of text detection failure: Letter from Simón de Tovar to Carolus Clusius, dated March 19, 1596 (img-id: CLUY030-001-c, VUL 101). Image used with permission of the Digital Special Collections of Leiden University Library.

Abb. 7: table 1: Identification results in percentages using the whole datasets for the evaluation.

Abb. 8: table 2: Identification results using four page images from each author, compared with the results using the CVL dataset.