Towards the Processing of Historic Documents

Björn Gottfried and Lothar Meyer-Lerbs

Centre for Computing and Communication Technologies
University of Bremen, Germany

Abstract. This chapter describes methods required for transforming complex document images into texts. The goal is to make the contents of those documents available for search engines, which are not born-digital but converted from a physical medium to a digital format. Established optical character recognition methods fail for documents for which no assumptions can be made regarding the, probably unknown, symbols contained in the document, historic documents being the example domain par excellence. This paper, however, has a much broader goal: it outlines fundamental problems as well as a methodology in the dealing with documents containing unknown and arbitrary symbols in order to provide a basis for discussions and future work within the digital library community. In particular, future advances will more closely require the interaction of researchers concerned with such diverse topics as document digitisation, reproduction, and preservation as well as search engines, cross-language processing, mobile libraries, and many further areas. Adopting a general view on the presented issues, researchers of the aforementioned areas should be sensitised for the problems met in processing complex, especially historic documents.

1 Introduction

In the last decade several digitisation projects have been carried out. Whole books and even entire collections of libraries have been transformed into a digital format in order to provide them by what has been introduced several years ago as digital libraries. Apart from digitised content, digital libraries also include content referred to as born-digital, that is content which was created in a digital format. While the latter offers the user a sophisticated functionality to search through the content, this is impossible for printed material that just has been converted into a digital format.

In particular, in the last decade many projects have been established to digitise and archive the cultural heritage. The idea is to save the material from a loss and to distribute it through the web in order to make it available everywhere. Some of the most prominent projects include

- the project Gutenberg (`www.gutenberg.org/wiki/Main_Page`),
- the Google Books Library Project (`books.google.com/`), and
- the Million Book Project (`www.ulib.org/`).

A digitised book, however, is nothing else than a collection of images which result from scanning the according book. Therefore, most digitised collections just make available more or less large images of the contents of the books. In order to access that content,

R. Bernardi et al. (Eds.): NLP4DL/AT4DL 2009, LNCS 6699, pp. 15–28, 2011.

equally like for born-digital content, it is necessary to extract the text contained in these images. The field of document image processing is concerned with this analysis. This scientific area has been established many years ago, even before the mass digitisation projects started. The problems that arise when applying document image processing methods to document images are manifold, which is the reason for why only a small segment of scanned documents is available for search engines. Most document images are too complex regarding their layouts, fonts, and components; sometimes, even different languages are intermixed and different symbol systems are used within single documents. Established optical character recognition software is unable to successfully process such documents.

In particular, there is a large segment of books that has been published before the twentieth century. Those publications are especially difficult to access by means of search engines since they contain complex fonts, many special characters, and even symbols unknown today. Additionally, they suffer from several other problems, such as pages being yellowed, blotted, and distorted. In order to make the contents of such books available new means are required which enable the processing of historic fonts. While sophisticated image processing methods are required for this purpose, the basic idea is rather simple: for each document, which might be a single certificate, a letter, or a whole book, a document specific font is generated out of that document. This specific font derives from visual features which can be extracted out of document images. These features enable the classification of characters and of every kinds of symbols, since at this stage no assumptions are made concerning the underlying language. Referring to such visual features, which can be arbitrarily complex shape features, the font of a historic document can be arbitrarily complex itself. While the recognition of the underlying characters is not included in this process, a subsequent mapping process has just to follow for the extracted characters to be recognised with respect to a particular alphabet. In this sense, this paper describes the very first step necessary to apply other advanced technologies to such documents, for instance, to search through these documents, to deal with cross language processing issues, or to even evaluate the content at the semantic level.

This chapter is about the processing of difficult, especially historic documents. The results include methods which are about the extraction of texts from images, in particular, for those documents for which standard optical character recognition methods fail. Secondly, for each such document a document specific font is extracted, that defines for each character class a visually optimal exemplar; taking those exemplars for all characters a document can be reproduced on different media, e.g. on smart phones to access contents from everywhere, by referring to the new document specific scalable font. Thirdly, methods are investigated that are fast regarding the whole analysis process; this is important inasmuch a library would have to analyse large collections in a reasonable time. Eventually, a large compression rate is achieved since fonts are represented by means of vectors which are much more compact than the according images, which do frequently have a particular high resolution in order to make the original document persistent as precise as possible. It is shown how this approach works by analysing documents of a Fraktur type.

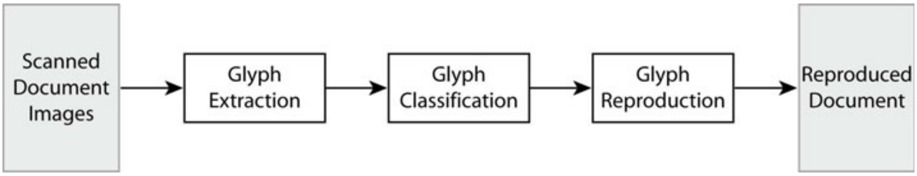


Fig. 1. The process chain

The body of this chapter is structured as follows. In Sect. 2 the methodology as a whole is outlined. It divides into three main stages: font extraction, glyph classification, and document reproduction, which are described in turn in Sects. 3 to 5. Conclusions are drawn from this work in Sect. 6, links of the presented work to different topics within the digital library community are given in Sect. 7, and a summary in Sect. 8 closes this chapter.

2 The Approach

The methodology for extracting texts out of documents containing unknown symbol sets is outlined in this section. Each document is processed as indicated in the flowchart shown in Fig. 1. Whole repositories of scanned document images are to be analysed and enter this process chain.

2.1 Glyph Extraction

In the first step, the symbols which are found on the document pages are extracted out of these document images, more precisely, out of all document images pertaining to one document, a book, a certificate or something else. From now on we use the notion of a glyph which represents the visual appearance of a symbol. A symbol, like an 'a', might be represented by different glyphs within different fonts; or, in different languages, glyphs might even represent different symbols. Furthermore, for the time being we neglect that a symbol might be represented by either exactly one or by more glyphs, which are either single or multi-piece regions in the image.

The assumption is that each document contains a font which might be quite specific to this document, because glyphs could be part of such a document which are not in use anymore today (e.g. a font of a Fraktur type). Accordingly, two different documents might contain very different glyph sets which is the reason for why those different documents should be processed separately.

The extraction of the symbols of a document requires image processing methods that are able to determine regions in images which represent exactly one glyph. Such regions are to be determined as precisely as possible, because similar glyphs are to be distinguished and the glyphs are to be reproduced in a later step in order to reproduce documents with a high quality.

2.2 Glyph Classification

The second step is about putting those glyphs into equivalence classes which represent the same symbol within the present document specific font. Those regions that represent exactly one glyph can be characterised by means of shape descriptions. Each glyph in a document needs to be represented uniquely by such a shape description; ideally, each occurrence of the same glyph would have the same shape description. In this way, the glyph extraction process takes into account the specific glyphs of the given document, namely the shapes of those glyphs. As a consequence, from each document a document specific font containing a particular glyph set can be extracted. This glyph set corresponds to an arbitrary symbol set which has been employed in the according document.

While accuracy is a fundamental requirement of the first step, efficiency is important for glyph classification. This is because taking a single document page there might be already a few thousands of glyphs on a page, the number varying with the size of the given font. That is, a huge number of glyphs is to be classified when taking a whole book. In order to process a book in a reasonable time, shape descriptions are to be investigated which allow a fast glyph classification.

2.3 Glyph Reproduction

Having extracted and described the glyphs' shapes in the previous steps, their reproduction can be based on these shapes. For this purpose, it is sufficient to reproduce their outlines and inner holes which can be represented in a SVG[1] vector format. Since the classification step results in equivalence classes for glyphs, each class can be represented by exactly one glyph in its vector format. This has the consequence that glyphs which are correctly classified but which suffer from deficits concerning how they are depicted in the original document, can be reproduced as if all visual defects have been automatically corrected. Each glyph class can be represented by a particular good exemplar.

The two main requirements for glyph reproduction are a good compression rate and a scalable glyph representation. The first requirement enables the compressed representation of large books. In this way, less memory is required for the encoded documents than for their original document images. This enables also the transfer of large documents among devices as well as to display them on devices with restricted memory resources, such as on smart phones. The latter also requires to change the scale of the font so as to make it optimally visible on a small device. This will be possible through the second requirement.

3 Extraction of Glyphs

The extraction of glyphs requires sophisticated methods which are, unfortunately, to a large degree dependent on the given image material. The latter might suffer from several different problems, such as yellowed, blotted, or distorted pages. A number of different methods have been investigated; the most important results are found in [7]. In the following we show how a specific set of image processing filters enable the extraction of glyphs from a journal series of the nineteens century containing Fraktur glyphs[2].

[1] Scalable Vector Graphics.

[2] Die Grenzboten, 28. Jahrgang, 2. Semester 1. Band, Leipzig 1869.

In order to extract glyphs from the document images (top left of Fig. 2), first of all colour images are converted to grey tone images (top right of Fig. 2). A Sigma-filter is applied in order to suppress artefacts [6]; such filters maintain edges, while the background is smoothed (bottom left of Fig. 2). Connected components which represent single glyphs are determined by applying binarisation filters, such as Sauvola [9] or Shafait et al. [11] (bottom right of Fig. 2).

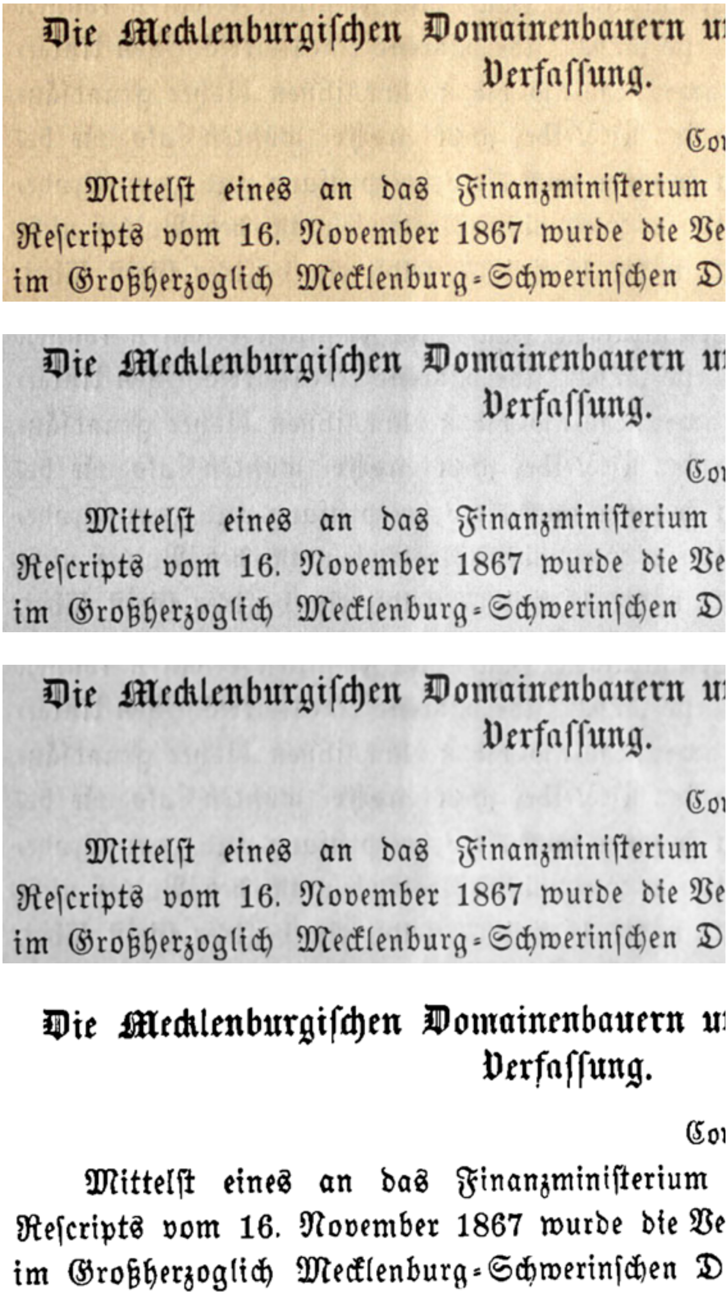
Starting from the connected components a Euklidean distance map is computed (EDM) (top left of Fig. 3). The connected components are extended by two points into each direction in order to grasp the grey values from the surroundings of each connected component. This is required for later reproducing the glyphs appropriately (top right of Fig. 3). The connected components are then cut off the denoised image and the gaps are filled with grey tone values with a bilinear approximation method; the resulting image shows the background (bottom left of Fig. 3). The extended connected components are subtracted from this background image and the result is inverted in order to get black glyphs on the white background (bottom right of Fig. 3). Deskewing algorithms can be applied to this image in order to correct the orientation of glyphs with respect to the document page, without being exposed to blurrings and other artefacts of the background. The connected components of this final image can be forwarded to the glyph classification methods.

4 Classification of Glyphs

As argued above, a fundamental constraint in the present application is efficiency. Suitable features for classifying binarised glyphs should enable fast comparisons. This is hardly possible when employing complex templates that describe shapes in a sophisticated and detailed way. By contrast, the most compact features characterise shapes by means of single numeric values. Textbook examples include the compactness of a glyph, its radius ratio, aspect ratio, convexity, and Hu moments [5]. Comparisons based on such features stick to a constant runtime complexity, since they describe glyphs independently of the number of components, which might either be contour points or all points contained within a glyph. It is therefore worthwhile to investigate whether such features are sufficiently precise.

While those single numeric features mentioned in the previous paragraph are not precise enough, it has been shown how qualitative features complement those established features while sticking to the same runtime complexity [7]. These features are based on a system of shape properties introduced in [3]. Instead of computing those features on all contour points, glyphs are first of all approximated by straight segments which frequently represent a glyph much more compact, since many glyphs contain a number of straight segments.

It is then the idea to describe a glyph shape with respect to single glyph segments. That is, the shape of a glyph extends over a specific range defined by each single segment. This latter is referred to as a segment's scope that can be succinctly described as to be left-of a segment, right-of it, on top, below and by some further directions which



Die Mecklenburgiſchen Domainenbauern u
Verfaſſung.

Co

Mittelſt eines an das Finanzminiſterium
Reſcripts vom 16. November 1867 wurde die Ve
im Großherzoglich Mecklenburg-Schwerinſchen D

Die Mecklenburgiſchen Domainenbauern u
Verfaſſung.

Co

Mittelſt eines an das Finanzminiſterium
Reſcripts vom 16. November 1867 wurde die Ve
im Großherzoglich Mecklenburg-Schwerinſchen D

Die Mecklenburgiſchen Domainenbauern u
Verfaſſung.

Co

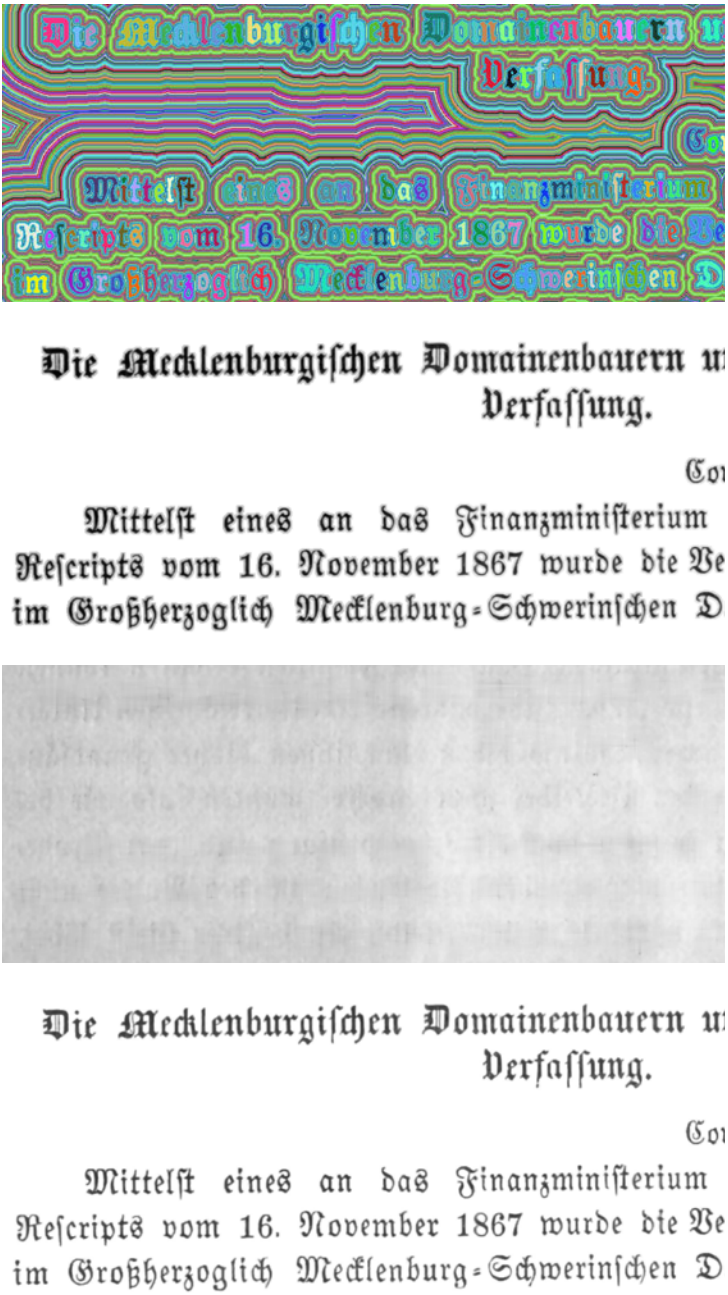
Mittelſt eines an das Finanzminiſterium
Reſcripts vom 16. November 1867 wurde die Ve
im Großherzoglich Mecklenburg-Schwerinſchen D

Die Mecklenburgiſchen Domainenbauern u
Verfaſſung.

Co

Mittelſt eines an das Finanzminiſterium
Reſcripts vom 16. November 1867 wurde die Ve
im Großherzoglich Mecklenburg-Schwerinſchen D

Fig. 2. Preprocessing steps: original, greyvalue image, noise reduction, and binarisation



Die Mecklenburgiſchen Domainenbauern u
Verfaſſung.

Co

Mittelſt eines an das Finanzminiſterium
Reſcripts vom 16. November 1867 wurde die Ve
im Großherzoglich Mecklenburg-Schwerinſchen D

Fig. 3. Next preprocessing steps: Euklidean distance map (EDM), connected components extended by two points according to EDM, bilinear averaged background, and enhanced image to be used for describing the glyphs

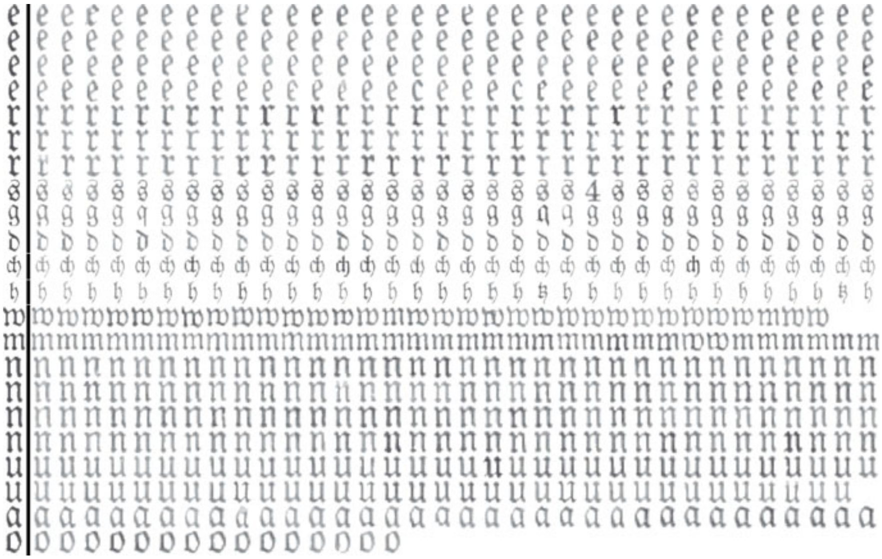


Fig. 4. Even similar glyphs could be told apart, as can be seen for the two glyph classes in the last two rows. But the compression is not optimal: among others, there are four classes for the glyph 'e', three 'r'-classes, four 'n'-classes, and two 'u'-classes.

can be combined to assemble to many different scopes. Counting their frequencies for each segment of a glyph, the scope histogram [10] is obtained. It describes the shape of a glyph in a significantly different way than other shape descriptions, why it in fact improves classification results when adding scope histograms to Hu moments and the other numeric features. Further improvements of this technique are expected by considering the orientation variance of glyphs which is neglected by all of the features used. Additionally, we aim at looking at interior contours of holes, entailing the consideration of more distinctive glyph properties which are, in our evaluation, solely taken into account by Hu's approach.

Current methods employed are computationally more expensive but result into better classification results. The horizontal and vertical profiles of single glyphs as well as a pixel correlation approach is employed [4]. The latter relates all image points to their neighbourhood with regard to their differences in grey tone values. By this means classification results are obtained with errors less than one per cent. A trade-off between this error rate and the compression rate of the obtained font representation is observed: the lower the compression rate the more precise the classification result and vice versa. Fig. 4 shows the classification results obtained for some similar glyph classes. The first column indicates the prototype glyph of a class, while all other objects in the same row pertain to the same class.

5 Reproduction of Glyphs

We aim at extracting document-specific fonts from historic document images. For their high-quality reproduction size, style and kerning information as well as subtle character

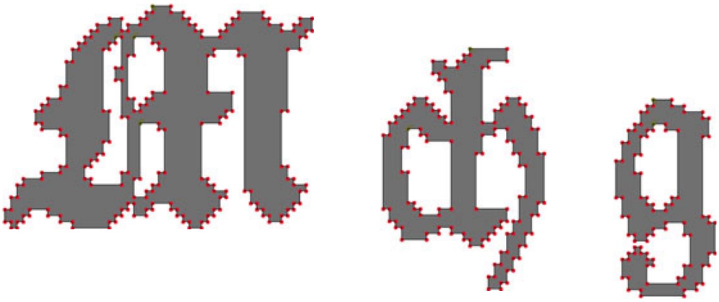


Fig. 5. Three examples for Fraktur letters: M, ch and g, generated by the described approach

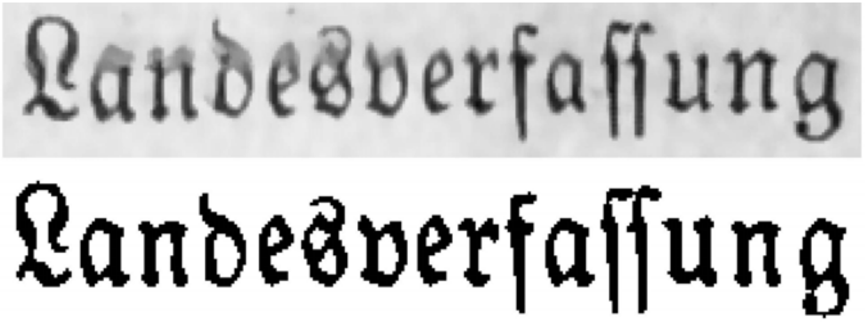
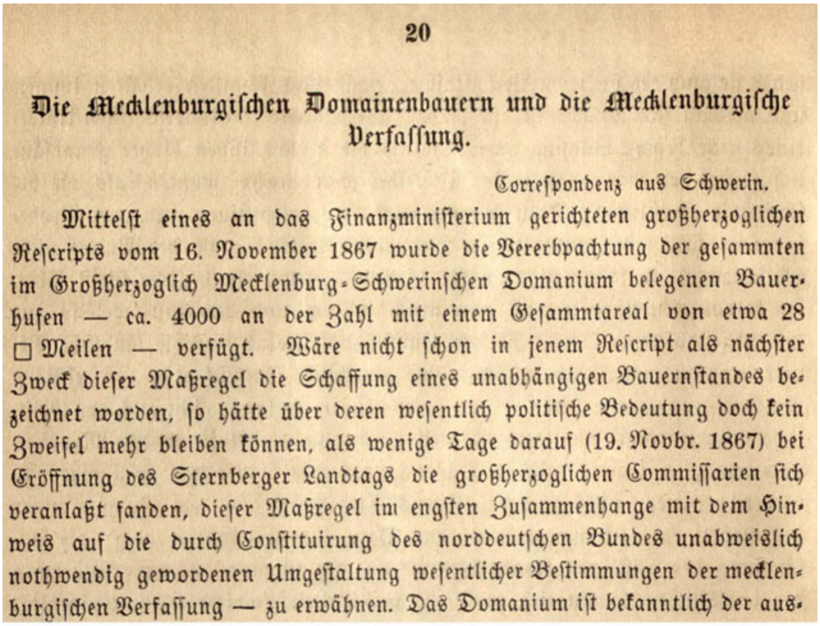

Landesverfaſſung
Landesverfaſſung

Fig. 6. Section from a document page including blurred glyphs which can be neatly reproduced

details are needed. Established optical character recognition methods are confined to correctly classifying glyphs. Here, however the idea is to assign unicode codepoints to prototype glyphs from the unicode ‘private use area’ and encode the generated fonts with unidentified glyphs. This allows reflowing of text and high speed text searches from examples – in essence a fast form of word spotting.

In detail, in order to generate a font, an edge following algorithm is applied on the binarised glyphs as well as on their holes, see Fig. 5. The obtained paths can directly be translated to SVG. In this way, it is possible to reproduce thousands of glyphs in a couple of seconds, meaning for a typical document page a processing of about ten seconds on a standard office laptop, on which no particular optimisation algorithms have been used.

Some of the advantages of the whole approach are illustrated in Fig. 6. A couple of the glyphs in this document are blurred. After the glyphs have been extracted and classified, each correctly classified glyph can be represented by a neat prototype glyph. Its SVG representation can eventually be scaled up and down arbitrarily, so that it fits the device where the document page is to be displayed. Figs. 7 and 8 in the appendix show two different examples with very different fonts and their final reproduction.

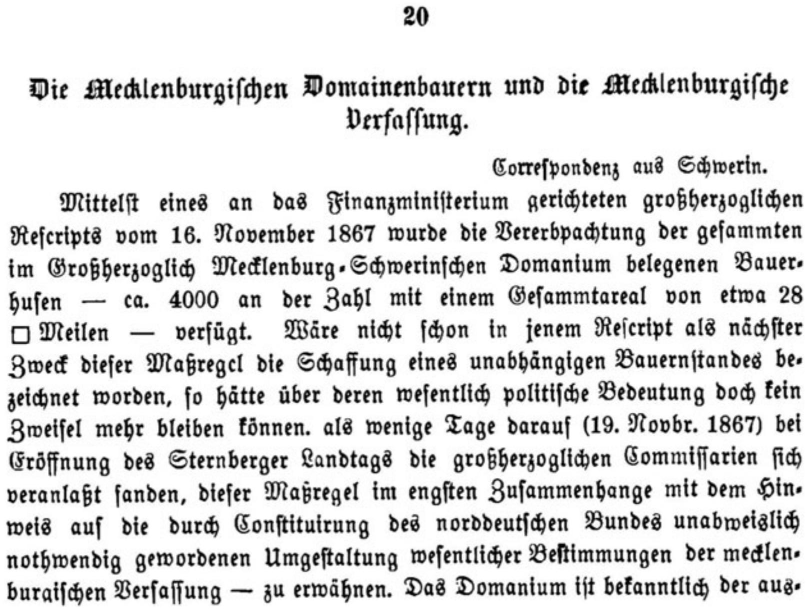


20

Die Mecklenburgiſchen Domainenbauern und die Mecklenburgiſche Verfaſſung.

Correſpondenz aus Schwerin.

Mittelſt eines an das Finanzminiſterium gerichteten großherzoglichen Reſcripts vom 16. November 1867 wurde die Vererbpachtung der geſammten im Großherzoglich Mecklenburg-Schwerinſchen Domanium belegenen Bauerhufen — ca. 4000 an der Zahl mit einem Geſammtareal von etwa 28 □ Meilen — verfügt. Wäre nicht ſchon in jenem Reſcript als nächſter Zweck dieſer Maßregcl die Schaffung eines unabhängigen Bauernſtandes bezeichnet worden, ſo hätte über deren weſentlich politiſche Bedeutung doch kein Zweifel mehr bleiben können, als wenige Tage darauf (19. Novbr. 1867) bei Eröffnung des Sternberger Landtags die großherzoglichen Commiſſarien ſich veranlaßt fanden, dieſer Maßregel im engſten Zuſammenhange mit dem Hinweis auf die durch Conſtituirung des norddeutſchen Bundes unabweislich nothwendig gewordenen Umgeſtaltung weſentlicher Beſtimmungen der mecklenburgiſchen Verfaſſung — zu erwähnen. Das Domanium iſt bekanntlich der aus-



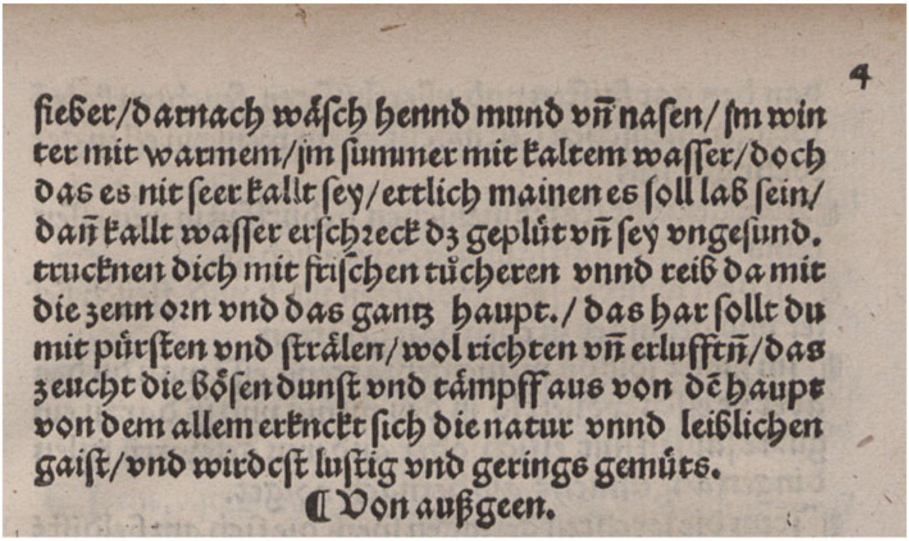
20

Die Mecklenburgiſchen Domainenbauern und die Mecklenburgiſche Verfaſſung.

Correſpondenz aus Schwerin.

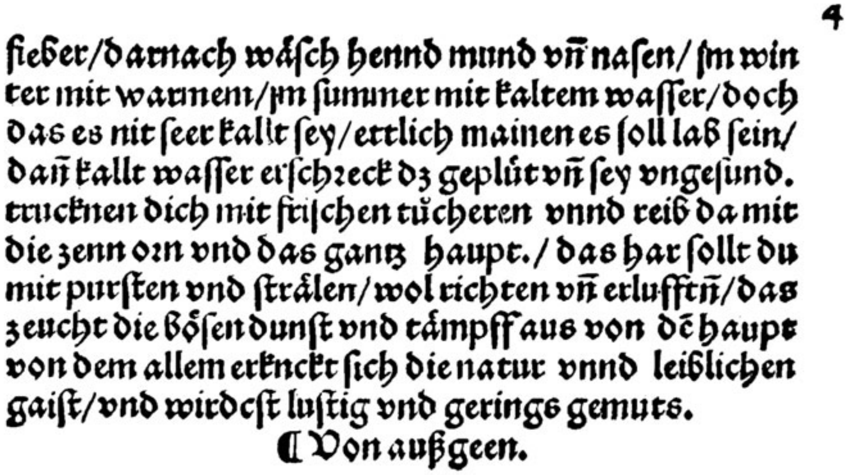
Mittelſt eines an das Finanzminiſterium gerichteten großherzoglichen Reſcripts vom 16. November 1867 wurde die Vererbpachtung der geſammten im Großherzoglich Mecklenburg-Schwerinſchen Domanium belegenen Bauerhufen — ca. 4000 an der Zahl mit einem Geſammtareal von etwa 28 □ Meilen — verfügt. Wäre nicht ſchon in jenem Reſcript als nächſter Zweck dieſer Maßregcl die Schaffung eines unabhängigen Bauernſtandes bezeichnet worden, ſo hätte über deren weſentlich politiſche Bedeutung doch kein Zweifel mehr bleiben können. als wenige Tage darauf (19. Novbr. 1867) bei Eröffnung des Sternberger Landtags die großherzoglichen Commiſſarien ſich veranlaßt fanden, dieſer Maßregel im engſten Zuſammenhange mit dem Hinweis auf die durch Conſtituirung des norddeutſchen Bundes unabweislich nothwendig gewordenen Umgeſtaltung weſentlicher Beſtimmungen der mecklenburaiſchen Verfaſſung — zu erwähnen. Das Domanium iſt bekanntlich der aus-

Fig. 7. Die Grenzboten, 28. Jahrgang, 2. Semester 1. Band, Leipzig 1869; SuUB, Bremen



4

fieber/darnach wäſch hennd mund vñ naſen/ jm win
ter mit warmem/jm ſummer mit kaltem waſſer/doch
das es nit ſeer kallt ſey/ettlich mainen es ſoll laß ſein/
dañ kallt waſſer erſchreck dz geplüt vñ ſey vngeſund.
trucknen dich mit friſchen tücheren vnnd reib da mit
die zenn orn vnd das gantz haupt./ das har ſollt du
mit pürſten vnd ſträlen/wol richten vñ erlufftñ/das
zeucht die böſen dunſt vnd dampff aus von dē haupt
von dem allem erkuckt ſich die natur vnnd leiblichen
gaiſt/vnd wirdeſt luſtig vnd gerings gemüts.
¶ Von außgeen.



4

fieber/darnach wäſch hennd mund vñ naſen/ jm win
ter mit warmem/jm ſummer mit kaltem waſſer/doch
das es nit ſeer kallt ſey/ettlich mainen es ſoll laß ſein/
dañ kallt waſſer erſchreck dz geplüt vñ ſey vngeſund.
trucknen dich mit friſchen tücheren vnnd reib da mit
die zenn orn vnd das gantz haupt./ das har ſollt du
mit purſten vnd ſträlen/wol richten vñ erlufftñ/das
zeucht die böſen dunſt vnd dampff aus von dē haupt
von dem allem erkuckt ſich die natur vnnd leiblichen
gaiſt/vnd wirdeſt luſtig vnd gerings gemuts.
¶ Von außgeen.

Fig. 8. Ellenbog, Ulrich: Ain wunderbaere jnstruction und underwysung wider die pestilentz, Memmingen, Albrecht Kunne, 1494; Bayerische Staatsbibliothek

6 Discussion

Evaluations of the presented method already show promising results. However, much progress is expected regarding further investigations into how to improve the presented methods. For example, qualitative shape descriptions can be improved wit respect to many aspects. When they reach the same classification performance than the methods currently employed, they would significantly reduce the computational complexity.

A fundamental observation, however, is that it is hardly possible to obtain the same classification results as those described in this work for arbitrary documents. Historic documents, in particular, but also other complex documents which cannot be processed by optical character recognition methods available today, would benefit from the described approach. But the difficulties met with document images concern arbitrary kinds of noise, and also, every kind of complexity concerning the document layout. What we have left out in our current evaluations are diagrams, illustrations and images which can all occur on document images and which are to be separated from the text. Furthermore, the latter sometimes runs along more columns, posing yet other challenges. This list of difficulties can be extended ever more, making it impossible to process arbitrary documents automatically.

Because of the aforementioned difficulties, we aim at developing a kind of assistance system for document processing which makes use of automatic processing methods, which, however, can be adjusted by the user within every step. As a consequence, documents with an arbitrary complex contents and layout will be successfully dealt with, although not fully automatically. This is what our approach distinguishes from others who argue in favor of a fully automatic processing approach [8].

7 Future Challenges

Apart from the discussed document processing issues, new challenges will require a tighter and interlinked cooperation among researchers coming from different areas within the digital library community. Some of these challenges are as follows, showing the place of the presented work within the digital library community:

- There might be complex documents from all sorts of areas containing every kinds of sophisticated symbols which are to be transformed into a digital format; the methodology described in this paper will be of use in these cases.
- Complex documents to be analysed could benefit from other successful document processing projects that are well accessible through library catalogues which provide features about their source of origin. The latter might give great indications about the success of specific document processing parameters. Catalogues could be enriched with meta-data about such features.
- While the parsing of content takes place at a much more abstract level of document representation, failures in document processing might have been kept undetected. Dealing with known languages during grammatical parsing, such failures could be detected when the parsing fails itself. A link back to the document processing level would inform the latter about misspellings.
- Multilingual documents are faced with the problem of special characters that are specific to a given language, such as in German or Swedish. The document specific font of a document would include all available symbols since no restrictions are made regarding the presence of more than one alphabet.
- The organisation of repositories should take into account limitations met at the document image processing level. The resulting transformations might suffer from

different problems and could be assigned with a particular grade of quality. This is of importance if the documents enter further anlysis tools.

- The context of the document image with all its characteristics could give hints about when, where, or under which circumstances the document has been written, printed, or published. This knowledge in turn can be of great value when evaluating the content itself.
- Specific problems at the level of language analysis, such as disambiguation, could benefit from background knowledge obtained with the aid of the underlying document. That is, it might be impossible to resolve ambiguities at the linguistic level, but when deriving the age of some given document, background knowledge can inform us about the possibility that specific meanings of a word or a phrase would make sense for a given period.
- Data integration at a rather basic level can benefit from and make use of the same knowledge as in the previous examples, namely concerning whatever can be derived from the document images about the period when the document has been published or from which location it is.
- Document retrieval can take place at different levels of abstraction. Digital content can be easily accessed at the symbolic level, employing all currently available means of search engines. But document images which could not be translated into a text format with a sufficient grade of quality can instead be searched by query-by-example [1] or even with the aid of query-by-sketch approaches [2].

This list is presumably not complete. But it shows the diversity of future challenges and indicates the broad spectrum of methods necessary from different areas. Scientists from those areas have to collaborate closely to manage these challenges.

8 Summary

This paper presents a method for extracting texts from images. As opposed to optical character recognition methods, no assumptions are made regarding the underlying language. Instead documents can be processed which are made of an arbitrary large and complex symbol set.

The overall goal is not confined to present that method. Rather, links to other areas within the scientific community of digital libraries are established in order to provide an agenda for future research that will deal with ever more complex challenges for dealing with and managing documents at all conceivable levels.

Acknowledgements

This work was partially funded by the DFG (Deutsche Forschungsgemeinschaft), project Venod (HE 989/10-1). The ‘Staats- und Universitätsbibliothek’ Bremen and the ‘Bayerische Staatsbibliothek’, Munich, provided many fine examples of historic printed document images.

References

1. Flickner, M., Sawhney, W., Niblack, H., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., Yanker, P.: Query by Image and Video Content: The QBIC System. Computer 28, 23–32 (1995)
2. Gottfried, B.: Shape from Positional-Contrast — Characterising Sketches with Qualitative Line Arrangements. DUV - Deutscher Universitätsverlag, Springer Science+Business Media, Wiesbaden (2007)
3. Gottfried, B.: Qualitative Similarity Measures - The Case of Two-Dimensional Outlines. Computer Vision and Image Understanding 110(1), 117–133 (2008)
4. Ho, T.K.: Random decision forests. In: ICDAR 1995: Proceedings of the Third International Conference on Document Analysis and Recognition, p. 278. IEEE Computer Society Press, Washington, DC, USA (1995)
5. Hu, M.-K.: Visual pattern recognition by moment invariants. IRE Transactions on Information Theory 8(2), 179–187 (1962)
6. Lee, J.-S.: Digital image smoothing and the sigma filter. Computer Vision, Graphics, and Image Processing 24(2), 255–269 (1983)
7. Meyer-Lerbs, L., Schuldt, A., Gottfried, B.: Glyph extraction from historic document images. In: Proceedings of the 2010 ACM Symposium on Document Engineering. ACM, New York (2010)
8. Pletschacher, S.: A self-adaptive method for extraction of document-specific alphabets. In: ICDAR 2009: Proceedings of the 10th International Conference on Document Analysis and Recognition, pp. 656–660. IEEE Computer Society, Los Alamitos (2009)
9. Sauvola, J., Pietikäinen, M.: Adaptive document image binarization. Pattern Recognition 33(2), 225–236 (2000)
10. Schuldt, A., Gottfried, B., Herzog, O.: Towards the visualisation of shape features the scope histogram. In: Freksa, C., Kohlhase, M., Schill, K. (eds.) KI 2006. LNCS (LNAI), vol. 4314, pp. 289–301. Springer, Heidelberg (2007)
11. Shafait, F., Keysers, D., Breuel, T.M.: Efficient implementation of local adaptive thresholding techniques using integral images. In: Document Recognition and Retrieval XV, San Jose, CA, p. 6 (January 2008)