2011 International Conference on Document Analysis and Recognition

Automatic Estimation of the Legibility of Binarised Historic Documents for Unsupervised Parameter Tuning

M. Stommel
Artificial Intelligence Group
TZI, University of Bremen
Bremen, Germany
mstommel@tzi.de

G. Frieder
Engineering Management and Systems Engineering
The George Washington University
Washington, DC, USA
gfrieder@gwu.edu

***Abstract*—Document enhancement tools are a valuable help in the study of historic documents. Given proper filter settings, many effects that impair the legibility can be evened out (e.g. washed out ink, stained and yellowed paper). However, because of differing authors, languages, handwritings, fonts and paper conditions, no single filter parameter set fits all documents. Therefore, the parameters are usually tuned in a time-consuming manual process to every individual document. To simplify this procedure, this paper introduces a classifier for the legibility of an enhanced historic text document. Experiments on the binarisation of a set of documents from 1938 to 1946 show that the classifier can be used to automatically derive robust filter settings for a variety of documents.**

***Keywords*-Document enhancement, historic documents, legibility estimation**

I. Introduction

Readers of historic documents are often confronted with damaged or degraded pages where the text has become difficult to read. Stains and washed-out ink impair the contrast of the text and hinder the efficient study of larger numbers of documents.

A certain improvement can be achieved by the application of digital image processing techniques: Shading filters achieve a homogeneous background intensity. Smoothing filters remove noise, sharpening filters accentuate details and contours. A binarisation might be conducted to separate text from background. By adjusting the filter parameters properly, these methods achieve good results with most types of documents. The downside is that historic documents vary strongly in their appearance, so the optimal parameter settings are specific to individual documents. A manual parameter tuning for every document is time-consuming and may compensate the advantage of improved legibility.

This paper proposes a method that finds robust parameter settings for a variety of historic documents without human supervision in each case. This is done by automatically exploring the parameter space of a filter and measuring the effect on the legibility of a certain document. Once the parameter space is known, an appropriate filter setting can be easily chosen. For each point of the parameter space, the legibility of the filtered document is classified automatically. This is done on the basis of a binary image that marks supposed text pixels. The classifier itself is trained in a supervised way. Supervision is however only needed once for a specific filter, as opposed to once per document in the traditional way. Good experimental results have been achieved for the diaries of Dr. A. A. Frieder [1] which show all the mentioned legibility impairments.

II. Related Work

In Optical Character Recognition (OCR), Document Retrieval or Structure Analysis, a large number of features have been proposed to characterise letters or distinguish text from background.

Difficulties arise from the visual properties of the documents in our data base [1] which are characterised by mixed handwritten, typewritten and printed text, multiple languages, authors and fonts, overlayed text, lined paper, washed out and bleeding ink and toner, different text colours, newspaper articles, collages, photos, and partial damages.

OCR features address the recognition of typewritten letters. Usually variants of geometric moments, parametric curves or Fourier descriptors [2] are used to represent the bitmap, contour or skeleton of a letter. For recognition, machine learning algorithms are supported by font [3] and language models [4]. However, even in the restricted case of typewritten texts, OCR does not handle historic documents well [5] because it relies on a correct text segmentation [6]. Except from simple pixel based features (e.g. [5]) most OCR features are highly shape selective. For the classification of the legibility this seems unfavourable because the method should generalise over handwritten and printed characters.

To improve text segmentation, Agrawal and Doermann [7] remove noise from binary images by constraining the aspect ratio of thinned line segments. Such features seem to be much less selective than OCR features. Moghaddam et al. [8] work with a level-set method to detect bleed-through.

The distinction of text and background is often studied in document structure analysis or image retrieval. Jung et al. [9] distinguish between regional approaches based on colour, texture or gradient, and shape-based approaches

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.30

where connected components are analysed. Text line orientation is a useful additional feature (e.g. [10]). The lower selectivity of most features (e.g. geometric moments) make these methods more robust against mixed handwritten and typed text. However, since the focus of these methods is text localisation, the legibility of a text region is usually not analysed explicitly. A pixel based measure is however proposed by Kuk and Cho [11]. Based on manually defined multi-scale filter responses, the authors rate a pixel as text, near text or background. An energy minimisation is performed to relabel pixels in order to create a binary image.

III. Automatic Exploration of the Parameter Space

The search for an optimal filter setting ϕ_{opt} of a given filter can be expressed as solving the equation

$$\phi_{\text{opt}} = \arg\max_{\phi} \Lambda(\psi(\phi, d)), \tag{1}$$

where the function Λ judges the legibility of a document d based on a feature description ψ of the filtered image. The optimisation itself is performed without manual interaction. However, the training of the function Λ is done in a supervised manner once for a specific filter. In the following the performance of the approach is demonstrated for the well known thresholding filter. To account for complex document structures, the legibility function is applied to overlapping subwindows of the document.

A. Legibility Estimation

Following Kuk and Cho [11], we define Λ as a ternary function that results 0 if a filter result does not represent legible text (black or white areas, photos, blobs or scattered points), 1 for fragmented or merged letters, and 2 for clearly outlined or optimally filtered text. In contrast to the cited work, we define Λ over an image region and use different features. The descriptor ψ is computed over binarised image regions of 100×100 pixels size. The regions are big enough to capture whole letters or words but at the same time smaller than the spatial frequency of the background gradient. We combine OCR features for historic documents [5] with features from noise removal in binary images [7]. The resulting feature descriptor is 14-dimensional. The first seven elements are the number of black connected components, white components, and black pixels in a grid cell, each normalised to the cell size, then the mean area and diameter of black components, the most frequent principal orientation, and the mean ratio of area to length of the black components. The dimensions 8 to 13 repeat these features for an eroded version of the input image using a 3×3 structuring element.

A soft margin SVM classifier [12] is trained to learn the legibility function Λ from a set of sample images. To this end, 1620 binary images are created by selecting 10 regions from every document in a set of 18 documents and applying 9 binarisation thresholds. The images are manually annotated, randomised and split into approximately equally sized training and test sets (each one covering all 18 documents).

Table I
Confusion matrix for the test of the legibility function using a 14-dimensional feature descriptor.

Recognised legibility	Annotated legibility		
	good	medium	bad
good	158	31	9
medium	19	60	15
bad	5	60	428
Precision	0.70	0.64	0.92
Recall	0.87	0.40	0.95

For the 14-dimensional feature descriptor, an accuracy of 86% for the training set and 82% for the test set is achieved (cf. tab. I). The results as well as projections of the feature space document a certain overlap of the good and medium class. A restriction to the first three features decreases the accuracy by 1% for the test set. A more detailed analysis reveals that the shape features in the higher dimensions of the descriptor primarily influence the variance among samples already classified correctly using only three features.

Figure 1 shows some of the test samples ordered by classification result. The samples classified as good contain mostly clearly outlined letters but sometimes also letters with small holes, or straight lines at the edges of a photo. Samples classified as medium contain mostly letters with thin strokes, broken lines and sometimes parts of photos. They are often visually close to the good quality samples. The image regions classified as bad contain primarily bright samples with occasional clutter, as well as dark parts of photos.

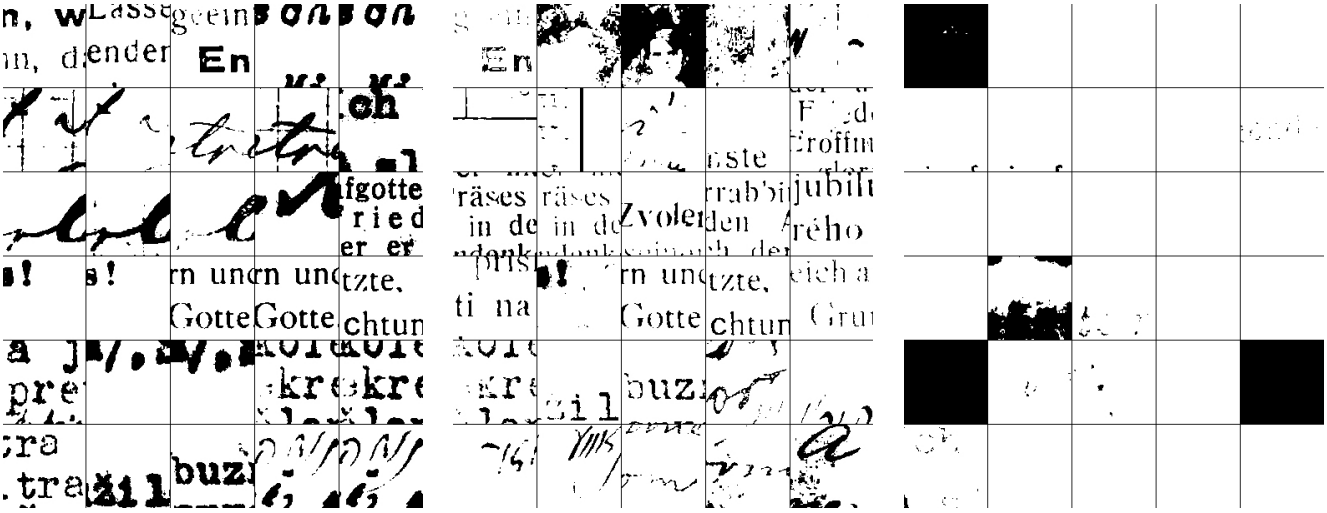
In summary, the classification results seem to be reasonably good. Since the intercoder reliability (the agreement of different people labelling the groundtruth, measured e.g. as Krippendorff's alpha) has not been measured yet, it is unclear if better results can be obtained realistically.

B. Parameter Optimisation and Filtering

To analyse the derived legibility function Λ, we compute its value for the complete range of binarisation thresholds. Figure 2 plots the results for a few examples.

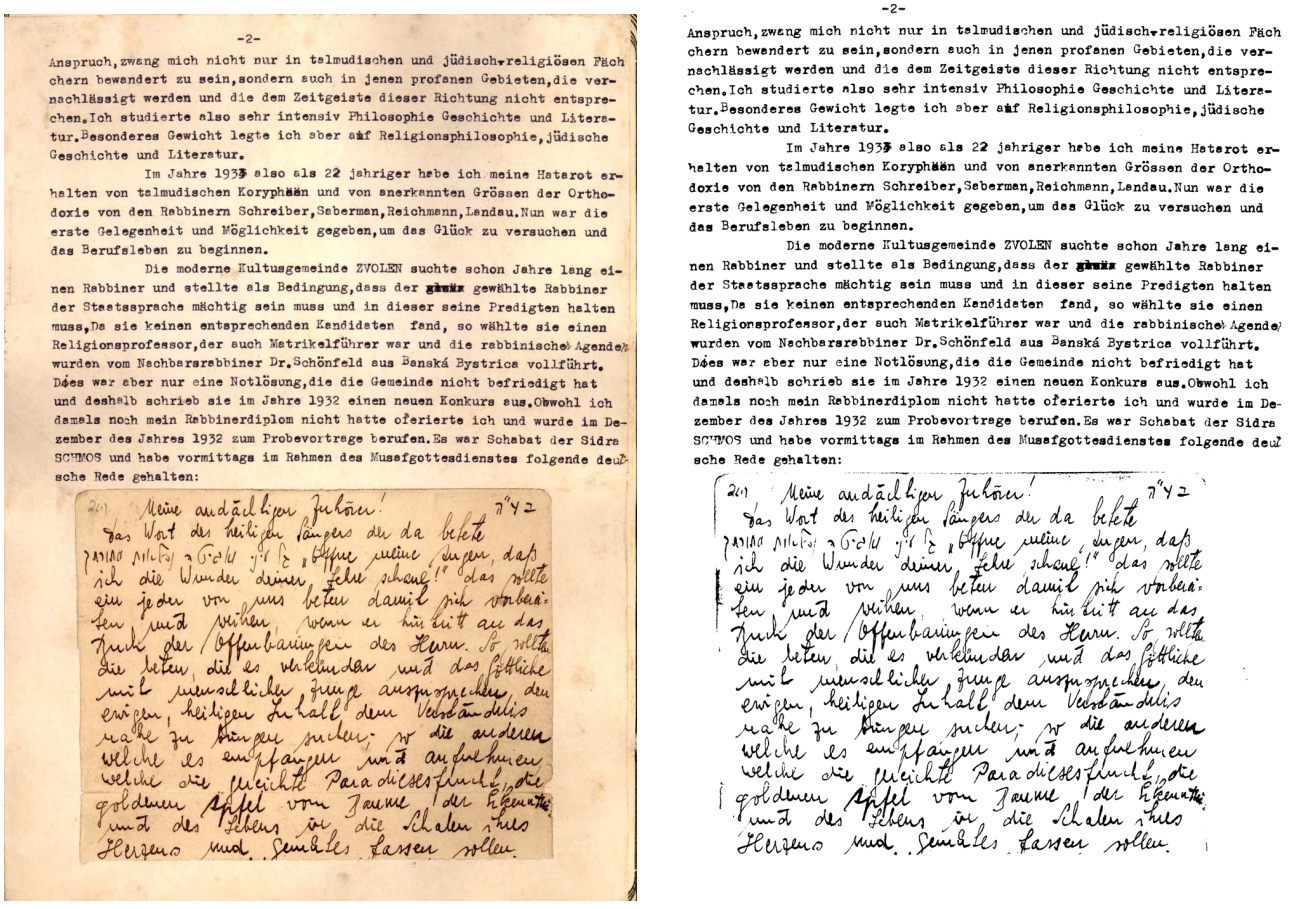
The first plot (a) shows the legibility for text region. The output of the function has a single broad maximum indicating a wide range of viable parameters. The center of gravity of the function is computed and used to threshold the original region. The result is a perfectly legible binarised image section.

Plot (b) shows a less robust example where the parameter interval 157–226 has not been classified as legible. In this case the center of gravity provides a good binarisation result. The span from the minimum to the maximum value classified as legible is as big as in the previous example.



a) Binary images classified good b) Binary images classified medium c) Binary images classified bad

Figure 1. Test samples and output of the legibility classifier. For every class, 30 binary images are shown.



-2-

Anspruch, zwang mich nicht nur in talmudischen und jüdisch-religiösen Fäch chern bewandert zu sein, sondern auch in jenen profanen Gebieten, die vernachlässigt werden und die dem Zeitgeiste dieser Richtung nicht entsprechen. Ich studierte also sehr intensiv Philosophie Geschichte und Literatur. Besonderes Gewicht legte ich aber auf Religionsphilosophie, jüdische Geschichte und Literatur.

Im Jahre 1935 also als 22 jähriger habe ich meine Heterot erhalten von talmudischen Koryphäen und von anerkannten Grössen der Orthodoxie von den Rabbinern Schreiber, Seberman, Reichmann, Landau. Nun war die erste Gelegenheit und Möglichkeit gegeben, um das Glück zu versuchen und das Berufsleben zu beginnen.

Die moderne Kultusgemeinde ZVOLEN suchte schon Jahre lang einen Rabbiner und stellte als Bedingung, dass der gewählte Rabbiner der Staatssprache mächtig sein muss und in dieser seine Predigten halten muss. Da sie keinen entsprechenden Kandidaten fand, so wählte sie einen Religionsprofessor, der auch Matrikelführer war und die rabbinische Agende wurden vom Nachbarsrabbiner Dr. Schönfeld aus Banská Bystrica vollführt. Dies war aber nur eine Notlösung, die die Gemeinde nicht befriedigt hat und deshalb schrieb sie im Jahre 1932 einen neuen Konkurs aus. Obwohl ich damals noch mein Rabbinerdiplom nicht hatte oferierte ich und wurde im Dezember des Jahres 1932 zum Probevortrage berufen. Es war Schabat der Sidra SCHMOS und habe vormittags im Rahmen des Musafgottesdienstes folgende deutsche Rede gehalten:

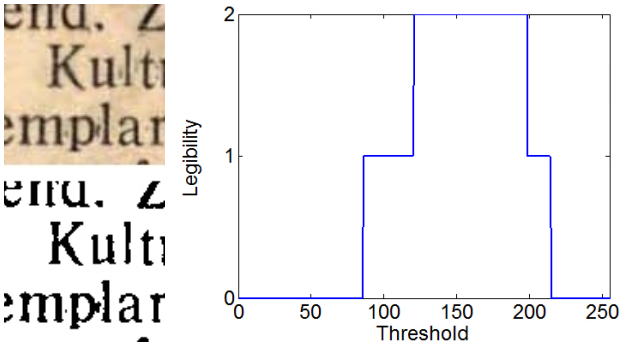
-2-

Anspruch, zwang mich nicht nur in talmudischen und jüdisch-religiösen Fäch chern bewandert zu sein, sondern auch in jenen profanen Gebieten, die vernachlässigt werden und die dem Zeitgeiste dieser Richtung nicht entsprechen. Ich studierte also sehr intensiv Philosophie Geschichte und Literatur. Besonderes Gewicht legte ich aber auf Religionsphilosophie, jüdische Geschichte und Literatur.

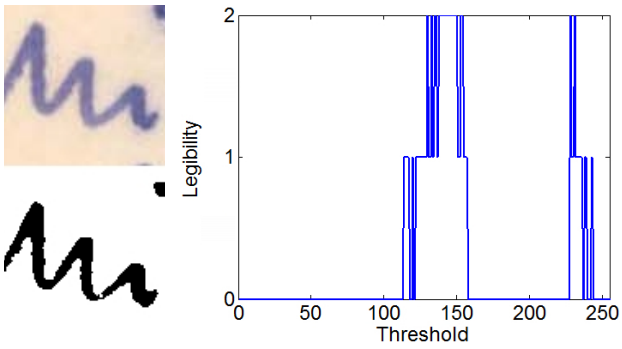
Im Jahre 1935 also als 22 jähriger habe ich meine Heterot erhalten von talmudischen Koryphäen und von anerkannten Grössen der Orthodoxie von den Rabbinern Schreiber, Seberman, Reichmann, Landau. Nun war die erste Gelegenheit und Möglichkeit gegeben, um das Glück zu versuchen und das Berufsleben zu beginnen.

Die moderne Kultusgemeinde ZVOLEN suchte schon Jahre lang einen Rabbiner und stellte als Bedingung, dass der gewählte Rabbiner der Staatssprache mächtig sein muss und in dieser seine Predigten halten muss. Da sie keinen entsprechenden Kandidaten fand, so wählte sie einen Religionsprofessor, der auch Matrikelführer war und die rabbinische Agende wurden vom Nachbarsrabbiner Dr. Schönfeld aus Banská Bystrica vollführt. Dies war aber nur eine Notlösung, die die Gemeinde nicht befriedigt hat und deshalb schrieb sie im Jahre 1932 einen neuen Konkurs aus. Obwohl ich damals noch mein Rabbinerdiplom nicht hatte oferierte ich und wurde im Dezember des Jahres 1932 zum Probevortrage berufen. Es war Schabat der Sidra SCHMOS und habe vormittags im Rahmen des Musafgottesdienstes folgende deutsche Rede gehalten:

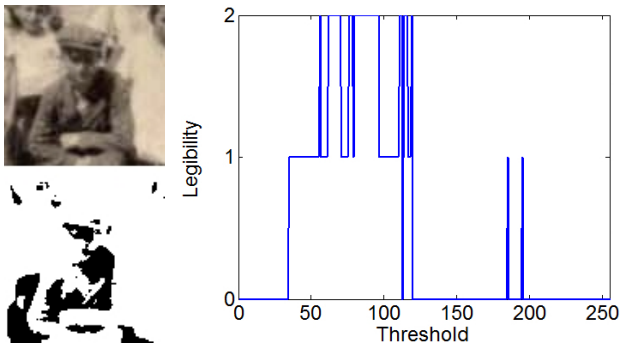
Figure 4. Original and binarised version of document "M.5_191_2" (test set)



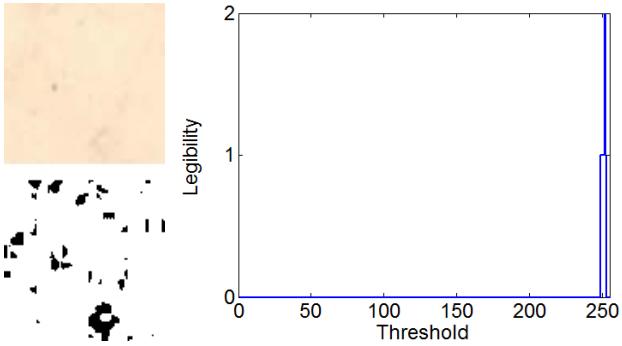
a) Perfectly recognised text sample



b) Partly misclassified text sample

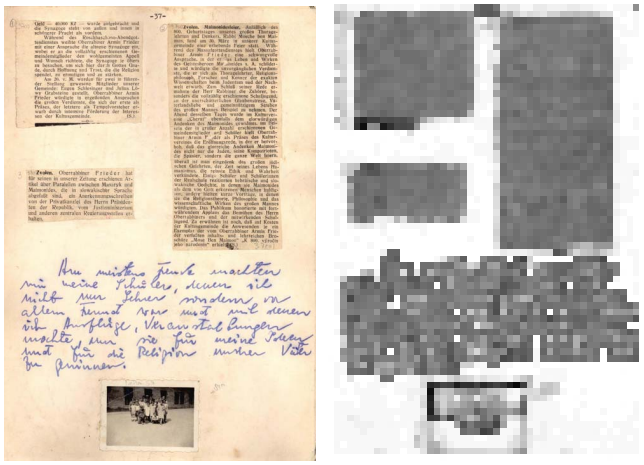


c) Part of a photo

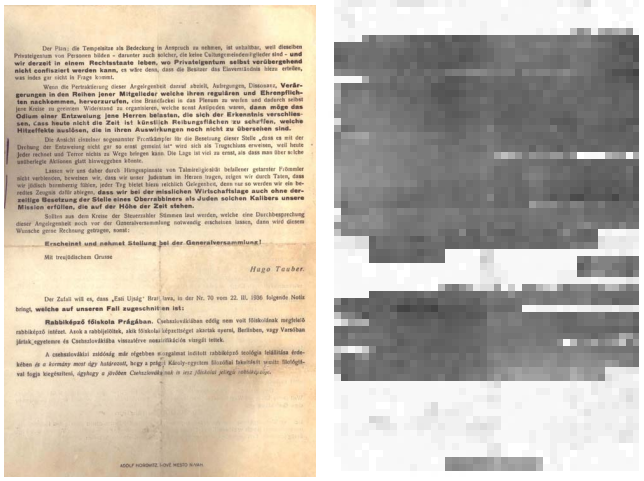


d) Background area

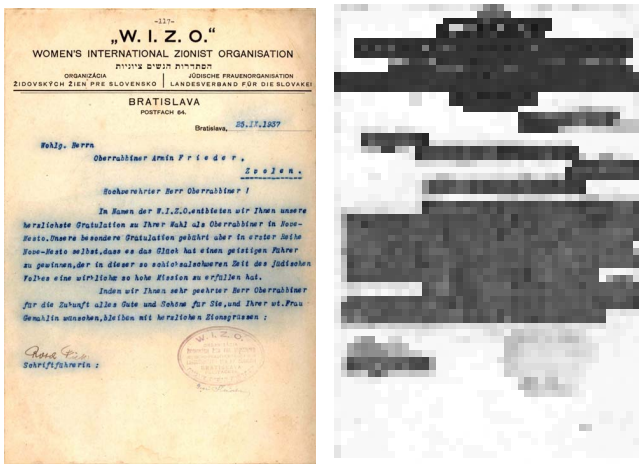
Figure 2. Dependency of the legibility on the binarisation threshold for different image material. The bitmaps on the left show the original gray value image, as well as a binarised version thresholded at the center of mass of the legibility curve on the right.



a) Sample document “M.5_191_38”



b) Sample document “M.5_191_112”



c) Sample document “M.5_191_122”

Figure 3. Text and background estimation based on the range of thresholds classified as good. Dark gray indicates a big range of good binarisation results and therefore a high text probability.

Plot (c) shows the output for a part of a photo. For a wide range of parameters the statistical properties of the resulting connected regions are indistinguishable from text. However, since filtering is not expected to transform photos into legible text, one would apply different methods to graphical regions anyway.

Plot (d) shows a typical result for a background region. The function identifies only a very small range of values as suitable for binarisation. The corresponding binary images usually show noise, JPEG-artefacts and paper edges.

The derived legibility function therefore seems well suited to find a good binarisation threshold. In noisy cases, the center of gravity serves as a robust candidate. While graphical areas cannot be distinguished from text areas, empty background regions can be detected by observing the parameter range classified as legible.

Based on these observations, a test system for the binarisation of whole documents is built. To this end, the document is partitioned into overlapping blocks suitable for the trained classifier. For every block, the legibility is computed over possible threshold parameters and the center of gravity is used for filtering. To avoid the emphasis of compression artefacts, the threshold is set to zero if the range of good parameters is too small. This results in white non-text/non-photo areas. A 3×3 median filter is applied to take advantage of the redundancy in adjacent blocks. To avoid visible borders between the filtering results of adjacent blocks, thresholds for single pixels are bilinearly interpolated between four neighbouring blocks. Close to text regions, the background detection is suppressed to avoid fading of the text as a result of the interpolation.

The method is evaluated by visual inspection of the filtering results for a test set of 93 documents (details are given in our technical report [13], fig. 4 shows an example). In 80% of the documents, the method is able to find appropriate binarisation thresholds that lead to a good general impression. In no case is the threshold chosen too low. Stains or lined paper have been solved well. However, the method is not suited to compensate for the known drawbacks of thresholding, so unevenly saturated letters often cause thinned binarisation results.

IV. Conclusion

This paper presents a method for the automatic optimisation of filter parameters in the enhancement of historic documents. To this end, a classifier for the legibility of binarised text is introduced. Since the method does not explicitly perform character recognition, it is applicable to both typewritten and handwritten text. An accuracy of 82% has been achieved in our experiments. We show that the classifier is suited to automatically explore the space of filter parameters in order to identify stable parameter sets. Test results on 93 documents encourage further experiments with more sensitive binarisation methods.

References

[1] A. A. Frieder, "The Diaries of Rabbi Dr. Avraham Abba Frieder," 2010, available online at: http://ir.iit.edu/collections/frieder_diaries_README.html. Original is kept as a permanent loan in the Archives of Yad Va Shem, under reference numbers M.5 191,192,193 and 194.

[2] O. D. Trier, A. K. Jain, and T. Taxt, "Feature Extraction Methods For Character Recognition - A Survey," *Pattern Recognition*, vol. 29, no. 4, pp. 641–662, 1996.

[3] A. Kae and E. Learned-Miller, "Learning on the Fly: Font-Free Approaches to Difficult OCR Problems," *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 571–575, 2009.

[4] M. Wüthrich, M. Liwicki, A. Fischer, E. Indermühle, H. Bunke, G. Viehhauser, and M. Stolz, "Language Model Integration for the Recognition of Handwritten Medieval Documents," *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 211–215, 2009.

[5] S. Pletschacher, J. Hu, and A. Antonacopoulos, "A new framework for recognition of heavily degraded characters in historical typewritten documents based on semi-supervised clustering," *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 506–510, 2009.

[6] K. Wang and S. Belongie, "Word Spotting in the Wild," *European Conference on Computer Vision (ECCV)*, 2010.

[7] M. Agrawal and D. Doermann, "Clutter Noise Removal in Binary Document Images," *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 556–560, 2009.

[8] R. F. Moghaddam, D. Rivest-Hénault, I. Bar-Yosef, and M. Cheriet, "A unified framework based on the level set approach for segmentation of unconstrained double-sided document images suffering from bleed-through," *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 441–445, 2009.

[9] K. Jung, K. I. Kim, and A. K. Jain, "Text information extraction in images and video: a survey," *Pattern Recognition*, vol. 37, pp. 977–997, 2004.

[10] A. Clavelli and D. Karatzas, "Text Segmentation in Colour Posters from the Spanish Civil War Era," *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 181–185, 2009.

[11] J. G. Kuk and N. I. Cho, "Feature based binarization of document images degraded by uneven light condition," *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 748–752, 2009.

[12] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer Verlag, 1995.

[13] M. Stommel and G. Frieder, "Evaluation of the Legibility Estimation for 93 Historic Documents," Center for Computing and Communication Technologies, University Bremen, Germany, Tech. Rep. 57, 2011.