Automatic Enhancement and Binarization of Degraded Document Images

Jon Parker [1,2], Ophir Frieder [1], and Gideon Frieder [1]

[1]Department of Computer Science
Georgetown University
Washington DC, USA
{jon, ophir, gideon}@cs.georgetown.edu

[2]Department of Emergency Medicine
Johns Hopkins University
Baltimore, USA
jparker5@jhmi.edu

***Abstract*—Often documents of historic significance are discovered in a state of disrepair. Such documents are commonly scanned to simultaneously archive and publicize a discovery. Converting the information found within such documents to public knowledge occurs more quickly and cheaply if an automatic method to enhance these degraded documents is used instead of enhancing each document image by hand. We describe a novel automated image enhancement approach that requires no training data. The approach is applicable to images of typewritten text as well as hand written text or a mixture of both. The pair of parameters used by the approach is automatically self-tuned according to the input image. The processing of a set of historic documents stored at Yad Vashem Holocaust Memorial Museum in Israel and selected images from the 2011 DIBCO test collection illustrate the approach.**

Keywords—readability enhancement; historic document processing; document degradation

I. Introduction

The proliferation of affordable computer hardware has made it increasingly desirable to digitize items of historic significance. Archiving historic images in particular has been encouraged by the introduction of high quality scanners and a drastic reduction in the cost of storing and transmitting the images those scanners produce. The subsequent increase in availability of historic document images presents an opportunity for image enhancement techniques to improve the accessibility and processing of historic image digitization and archival efforts.

We present a novel method to automatically enhance and binarize degraded historic images. This image enhancement method is unique because it:

- Requires no human interaction. Thus, this technique can be part of a high throughput image processing and archival effort.
- Requires no training data. Thus, this technique can be applied to any dataset.
- Is trivially parallel because it is input page independent.

We illustrate our method by applying it to selected images from two different corpuses. The first corpus contains scans of historic documents that are currently stored at Yad Vashem Holocaust Memorial Museum in Israel. The second corpus contains test images from the 2011 Document Image Binarization Contest.

II. Related Work

Binarizing a potentially degraded color image is a well-studied task. In 2009 and 2011 a Document Image Binarization Contest (DIBCO) was held during the International Conference on Document Analysis and Recognition (ICDAR). Multiple algorithms were submitted to both contests. The 2009 contest compared 43 different binarization algorithms while the 2011 contest compared 18 algorithms. The 2011 DIBCO organizers provided eight images of hand written text and eight images of printed text. The images in this test collection showed various forms of degradation. The competition also provided a black and white ground truth image for each of the 16 original images. The contest summary papers [1, 2] provide short descriptions of the submitted algorithms as well as a description of the results.

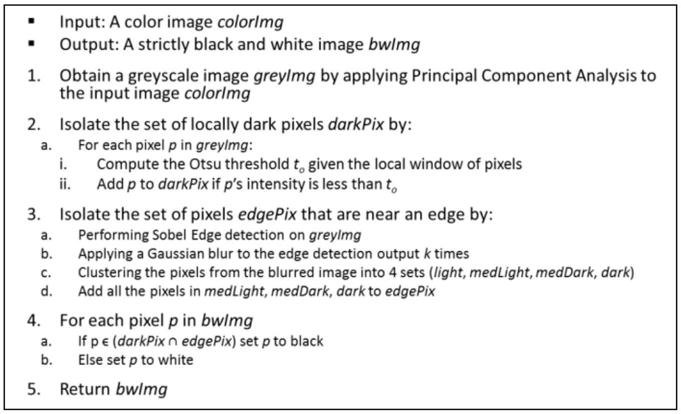
The image enhancement and binarization method introduced herein differs from some other methods in two important ways. First and foremost, the parameters of our method are set automatically. No human interaction or guidance is required to determine which parameter values to use. This differs from methods that encourage humans to adjust the parameters as necessary as in [3] as well as those methods that have sensitive parameters that must be hand set as in [4]. Automatically setting the parameters [5] also avoids the need to preset one or more global parameters as in [6, 7]. The second important distinction of this method is that it does not require training data of any kind. This distinguishes it from machine learning based methods such as [8]. An important consequence of not needing training data is that our method is input page independent. Therefore, our method can be applied to every image in a corpus in a trivially parallel manner.

III. Methodology

The method described herein converts a color or greyscale document image to a strictly black and white document image. The conversion technique is designed to simultaneously reduce the effect of document degradation and highlight the essence of the pre-degraded document. The ultimate goal is to produce a single black and white image that makes the information in the original document as legible as possible.

We empirically evaluate our method by applying it to images from Yad Vashem's Frieder Dairies. These historical, physically-degraded, museum-stored diaries were written primarily during the 1940s, survived adverse conditions during World War II, and provide a wide variety of image test cases. Pages contain typewritten text, handwritten script, pictures, multiple languages, and combinations of these. They also show differing amounts of degradation due to storage condition and paper type. This collection of images is available upon request. The diaries themselves are on permanent loan to Israel's Yad Vashem archives.

The image enhancement and binarization method presented here is based on the guiding principles that *"writing" pixels should be darker than the "non-writing" pixels nearby* and *"writing" should generate a detectable edge*. Of course, these principles are not universally true; however, they are rarely violated in practice; at least as far as observed herein. Our method is specified in Fig 1, with each of the 4 core steps discussed in detail below.



```
▪ Input: A color image colorImg
▪ Output: A strictly black and white image bwImg
1. Obtain a greyscale image greyImg by applying Principal Component Analysis to
   the input image colorImg
2. Isolate the set of locally dark pixels darkPix by:
   a. For each pixel p in greyImg:
      i.  Compute the Otsu threshold $t_o$ given the local window of pixels
      ii. Add p to darkPix if p's intensity is less than $t_o$
3. Isolate the set of pixels edgePix that are near an edge by:
   a. Performing Sobel Edge detection on greyImg
   b. Applying a Gaussian blur to the edge detection output k times
   c. Clustering the pixels from the blurred image into 4 sets (light, medLight, medDark, dark)
   d. Add all the pixels in medLight, medDark, dark to edgePix
4. For each pixel p in bwImg
   a. If $p \in (darkPix \cap edgePix)$ set p to black
   b. Else set p to white
5. Return bwImg
```

Figure 1: Pseudo code of Image Enhancement Method

A. Create a Greyscale Image

The first step towards obtaining an enhanced black and white image is to create a greyscale version of the input image. We use principle component analysis (PCA) to reduce the 3-dimensional RGB (red, green, and blue) value for each pixel to a single greyscale value.

B. Process 1: Isolating Locally Dark Pixels

The second step determines which pixels in the greyscale image are "locally dark". We use a constant sized window of pixels from the greyscale image to analyze each pixel. The window is an *n* by *n* pixel square region where *n* is always odd. As we slide this window across the greyscale image we make a "is locally dark" decision about the pixel at the center of this window. Each time the window is moved, we compute the Otsu threshold for the pixels within the window. If the center pixel is darker than the Otsu threshold we include that pixel in the set of "locally dark" pixels. For a pixel to be black in the final output image it must be flagged as "locally dark" in this step. This requirement is inspired by the general principle that "writing" pixels should be darker than the "non-writing" pixels nearby. The *winSize* parameter is set automatically using a method discussed in Section III.E.

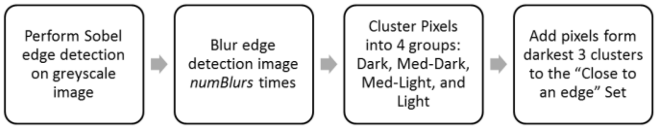

Perform Sobel edge detection on greyscale image
Blur edge detection image *numBlurs* times
Cluster Pixels into 4 groups: Dark, Med-Dark, Med-Light, and Light
Add pixels form darkest 3 clusters to the "Close to an edge" Set

Figure 2: Process Used to Isolate Pixels that are "Near an Edge"

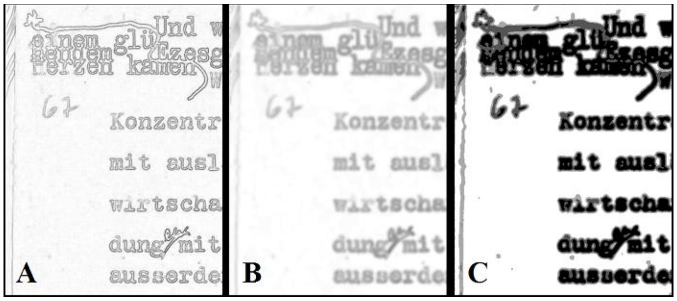


Figure 3: Intermediate Results when Isolating Pixels that a Near an Edge. (A) Edge detection output (B) Blurred edge detection output (C) Clustered output

C. Process 2: Isolating Pixels Near an Edge

The second guiding principle behind our method is that writing should generate a detectable edge. Process 2 isolates all pixels that are near detectable edges thus reflecting the second guiding principle. A summary of this pixel isolation process is depicted in Fig. 2.

We begin this process by running Sobel edge detection. The Sobel operator approximates the gradient of the greyscale image at a particular pixel. When the gradient is large, a border between light pixels and dark pixels exists. An example of the output of the edge detection step can be seen in the panel A of Fig. 3. Notice that the letters are outlined clearly in this example.

Once edge detection has been performed, we blur the resulting image one or more times. The blurring operation applies a Gaussian blur across a 5 by 5 window of pixels. The *numBlurs* parameter is set automatically using a method discussed in Section III.E. Next, the pixels within the blurry edge detection image (shown in panel B of Fig. 3) are clustered into 4 sets: dark, medium-dark, medium-light, and light pixels. An example of this clustering is shown in panel C of Fig. 3. Pixels that are assigned to the dark, medium-dark, and medium-light clusters are considered "near an edge".

D. Combining Results from Processes 1 and 2

Processes 1 and 2 generate two sets of pixels: (1) pixels that are "locally dark" and (2) pixels that are "near an edge". The final step towards creating a black and white output image is to compute the intersection of these two sets. Every pixel that is both locally dark and near an edge is set to black in the output image. If a pixel does not meet both of these criteria, it is set to white in the output image.

E. Parameter Selection

The processes discussed in sections III.B and III.C each requires one parameter: *winSize* and *numBlurs*, respectively. One of the more important aspects of this image enhancement and binarization method is that the only two parameters are determined automatically. Automatic parameter selection ensures that this method can be used with as little human interaction as possible.

The *winSize* parameter is set so that spotting like that shown in the middle panel of Fig. 4 is significantly reduced. This spotting is generally caused by noise in the image that produces phantom edges. Increasing the *winSize* parameter makes it more likely that an obvious edge is included in a window when the "is locally dark" decision is made. The inclusion of a true edge will reduce the likelihood that a false positive will occur due to noise. The net result is that spotting is less prevalent in the image.

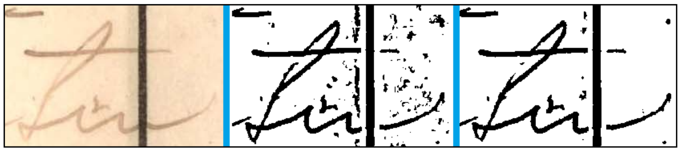
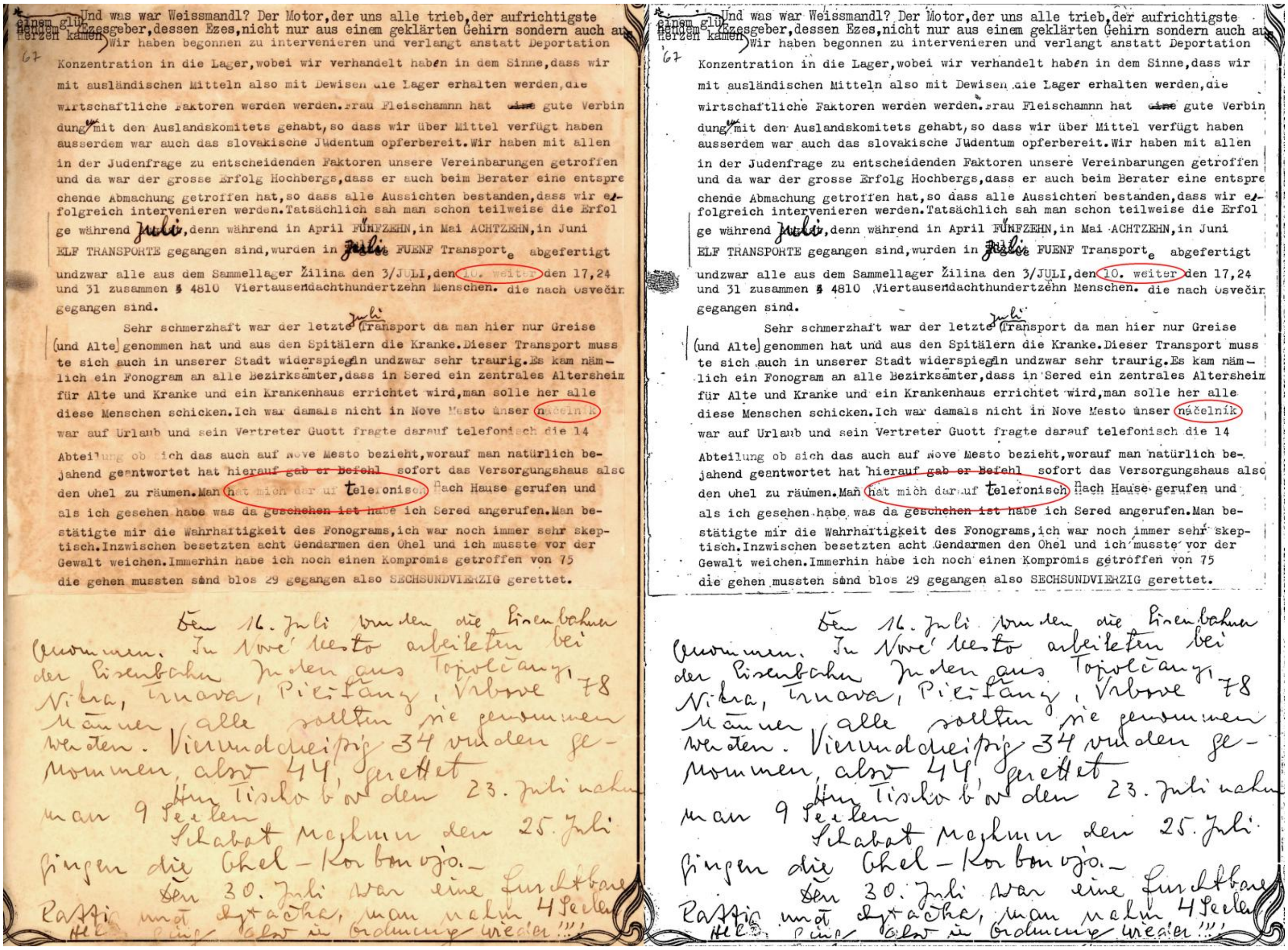


Figure 4: Increasing *winSize* parameter reduces "spotting": (left) Original (middle) *winSize* = 9, *numBlurs* = 2 (right) *winSize* = 17, *numBlurs* = 2.

The metric used to set the *winSize* parameter is designed to be sensitive to the spotting we are attempting to minimize. We increase the *winSize* parameter (from an initial value of 9) until our metric no longer changes appreciably. At this point we assume the level of spotting is also not changing appreciably.

The metric we use is deemed *the standard deviation of the standard deviations*. To compute this metric, we randomly select many (on the order of 10,000) 25 by 25 windows from an output image. We count the number of black pixels in each random window. Next, that count is converted to a set of n 0's and (625 – n) 255's where n is the number of black pixels in the corresponding window. We then compute the standard deviation of each set of 625 pixel color values. Next we compute the standard deviation of our sample of standard deviations. This metric is sensitive to spotting because the difference between a window composed of only white pixels versus a window composed of almost only white pixels is large.

The *numBlurs* parameter is set second. This parameter is gradually increased until each successive image is nearly identical to the proceeding image. A pair of images is deemed nearly identical if 99.5% of their pixels match. The *numBlurs* parameter is used mainly to enable our method to accommodate images of different resolutions.



einem glühendem Herzen kamen Und was war Weissmandl? Der Motor,der uns alle trieb,der aufrichtigste Ezesgeber,dessen Ezes,nicht nur aus einem geklärten Gehirn sondern auch au Wir haben begonnen zu intervenieren und verlangt anstatt Deportation

67 Konzentration in die Lager,wobei wir verhandelt haben in dem Sinne,dass wir mit ausländischen Mitteln also mit Dewisen die Lager erhalten werden,die wirtschaftliche Faktoren werden werden.Frau Fleischamnn hat ~~eine~~ gute Verbindungen mit den Auslandskomitets gehabt,so dass wir über Mittel verfügt haben ausserdem war auch das slovakische Judentum opferbereit.Wir haben mit allen in der Judenfrage zu entscheidenden Faktoren unsere Vereinbarungen getroffen und da war der grosse Erfolg Hochbergs,dass er auch beim Berater eine entsprechende Abmachung getroffen hat,so dass alle Aussichten bestanden,dass wir erfolgreich intervenieren werden.Tatsächlich sah man schon teilweise die Erfolge während Juli,denn während in April FÜNFZEHN,in Mai ACHTZEHN,in Juni ELF TRANSPORTE gegangen sind,wurden in Juli FUENF Transporte abgefertigt undzwar alle aus dem Sammellager Žilina den 3/JULI,den 10. weiter den 17,24 und 31 zusammen 4810 Viertausendachthundertzehn Menschen. die nach Osvečin gegangen sind.

Sehr schmerzhaft war der letzte Juli Transport da man hier nur Greise (und Alte) genommen hat und aus den Spitälern die Kranke.Dieser Transport musste sich auch in unserer Stadt widerspiegeln undzwar sehr traurig.Es kam nämlich ein Fonogram an alle Bezirksämter,dass in Sered ein zentrales Altersheim für Alte und Kranke und ein Krankenhaus errichtet wird,man solle her alle diese Menschen schicken.Ich war damals nicht in Nove Mesto unser náčelník war auf Urlaub und sein Vertreter Guott fragte darauf telefonisch die 14 Abteilung ob sich das auch auf Nove Mesto bezieht,worauf man natürlich bejahend geantwortet hat hierauf ~~gab er Befehl~~ sofort das Versorgungshaus also den Ohel zu räumen.Man hat mich darauf telefonisch nach Hause gerufen und als ich gesehen habe was da ~~geschehen ist~~ habe ich Sered angerufen.Man bestätigte mir die Wahrhaftigkeit des Fonograms,ich war noch immer sehr skeptisch.Inzwischen besetzten acht Gendarmen den Ohel und ich musste vor der Gewalt weichen.Immerhin habe ich noch einen Kompromis getroffen von 75 die gehen mussten sind blos 29 gegangen also SECHSUNDVIERZIG gerettet.

Den 16. Juli wurden die Eisenbahner genommen. In Nové Mesto arbeiteten bei der Eisenbahn Juden aus Topolčany, Nitra, Trnava, Piešťany, Vrbové 78 Männer, alle sollten sie genommen werden. Vierunddreißig 34 wurden genommen, also 44 gerettet.
Hm Tischo b'ov den 23. Juli nahm man 9 Seelen
Schabat Nachmu den 25. Juli gingen die Ohel-Korbanojo.–
Den 30. Juli war eine furchtbare Razzia und Szytacha, man nahm 4 Seelen
[illegible] also in Ordnung wieder!!!

Figure 5: Sample Result: Document M.5_193_67. Areas of interest are circled in red. Automatically set parameters: *winSize* = 13, *numBlurs* = 4

IV. Experimental Results and Discussion

Figures 5 and 6 show typical results from when our image enhancement and binarization method is applied to images in our dataset. These images were selected to illustrate some of the variety within our dataset as well as how our algorithm responds to handwritten script, typewritten text, and photographs. Areas of interest in these results are circled in red.

The three areas circled in Fig. 5 correspond to typewritten characters that are significantly lighter than their surrounding characters. Notice that these fainter characters are more legible in the enhanced document image than they are in the original document image (this is especially true when the images are viewed at their intended size). It is worth noting the phrase "hat mich darauf telefonish" is legible despite the image yellowing above "mich" and the boldly written "t" just prior to the faintly typed "uf" and almost undetectable "a".

The processed diary image on the right side of Fig. 6 shows two minor defects. In the top circle we see that only the bottom portion of the script loop is retained. A faint detectable edge is generated by the loop that is missing from the processed image. However, that detectable edge is "blurred away" when the 6 blurring operations are applied. The circle at the bottom-right highlights that the discoloration in that corner is not converted to a perfectly clean black and white image. The spotting discussed in section III.E is visible in this corner of the processed image. Note, however, that spotting is not present in most of the right hand margin – the spotting is only prevalent in the corner. This is particularly relevant because the images from Fig. 4 that introduce the spotting issue are excerpts from the original image shown in Fig 6.

The final observation to make about Fig. 6 is that the four photographs in the original image are recognizable in the final black and white image. The presence of these photographs did not hinder the ability to enhance the faint script to the right of the photos.

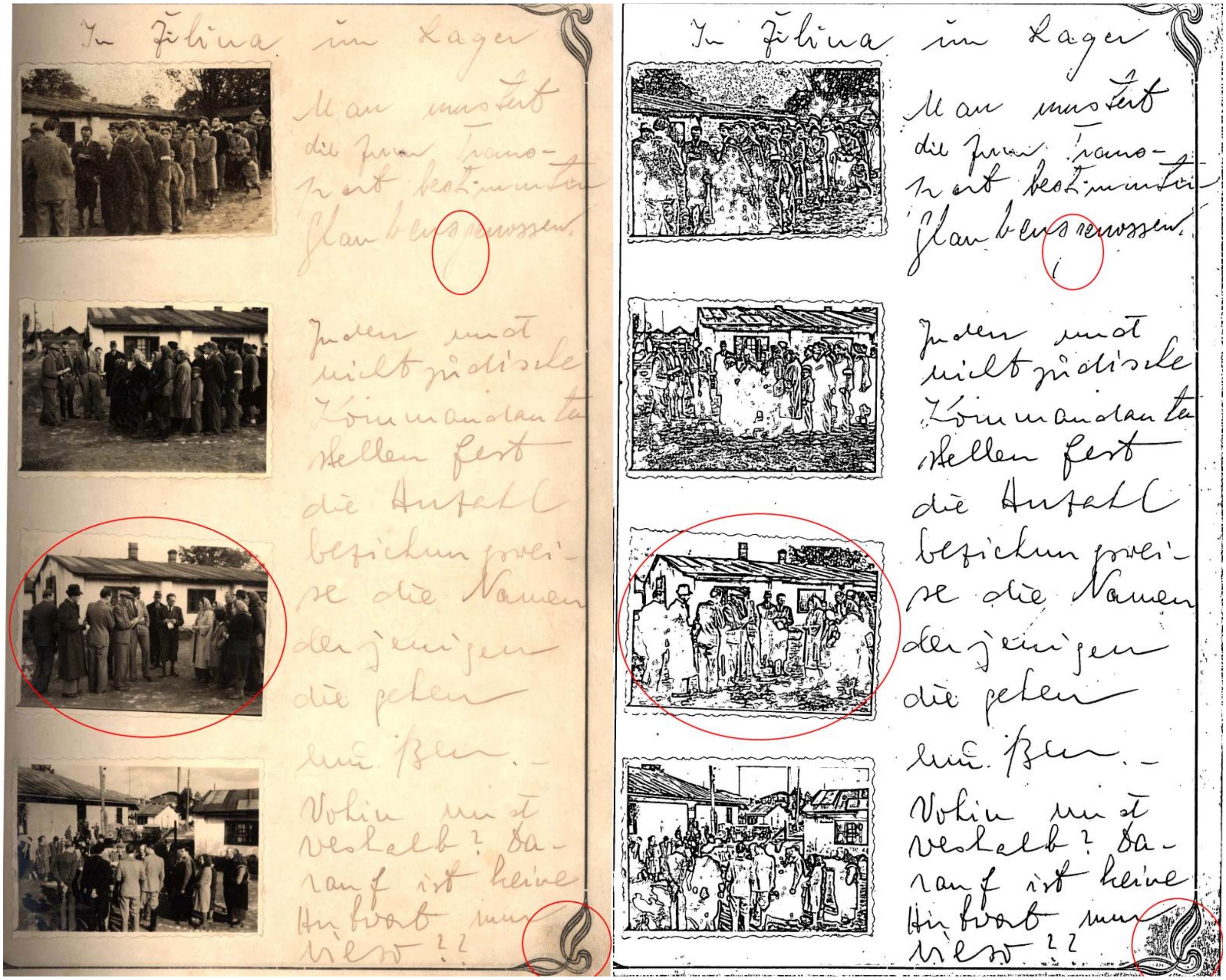

In Žilina im Lager
Man mustert
die zum Trans-
port bestimmten
Glaubensgenossen.
Juden mit
nicht jüdische
Kommandanten
stellen fest
die Anzahl
bezichungswei-
se die Namen
derjenigen
die gehen
müssen. –
Wohin und
weshalb? Da-
rauf ist keine
Antwort nur
wieso ??

Figure 6: Sample Results: Document M.5_193_25. Areas of interest are circled in red. Automatically set parameters: *winSize* = 17, *numBlurs* = 6

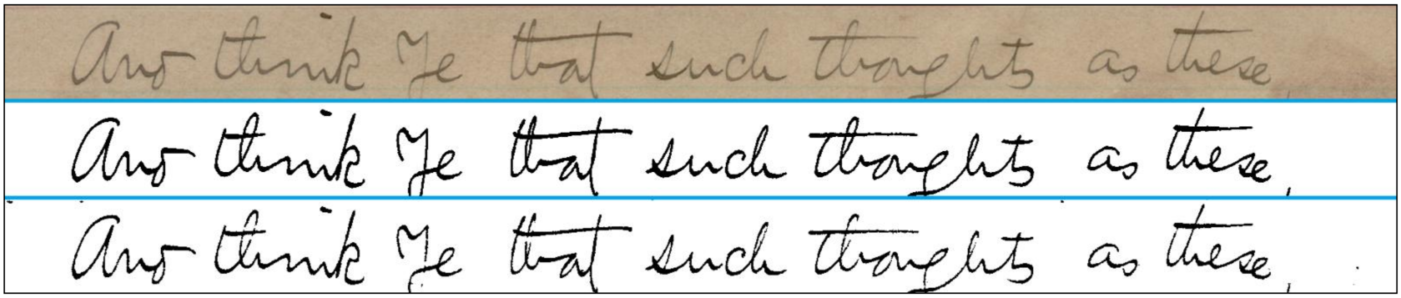


Figure 7: DIBCO 2011 HW3 Results: (top) Original image HW3 from DIBCO 2011, (middle) Ground Truth, (bottom) Results

A. The DIBCO 2011 Test set

Figs. 7 and 8 show excerpts of an original problem image (HW3 and HW2), its corresponding ground truth image, and the result of applying our method to that image. Our output image for HW3 has a precision of 0.979, a recall of 0.834, and an F-measure of 0.901. Our output image for HW2 has a precision of 0.981, a recall of 0.898, and an F-measure of 0.937. The top 3 methods from the DIBCO 2011 competition had an average F-measure of 0.927 for image HW3 and 0.944 for image HW2. Our method produces lines that are about one or two pixels thinner than the ground truth images.

V. Conclusion

The image enhancement and binarization method present here significantly improves the legibility of degraded historic images in our dataset. The main advantages of this algorithm is that it requires no human action to find parameters that yield good results nor is a training set of images needed. Avoiding the need for human interaction can significantly improve the throughput of image digitization and archival effects. Forgoing a training set enables the approach to be used on any collection. An ancillary benefit of this algorithm is that it simple to implement and easy understand. We conjecture that the enhanced images our method produces will enable improved optical character recognition performance. We plan to test this conjecture in the future.

References

[1] B. Gatos, K. Ntirogiannis, I. Pratikakis, “ICDAR 2009 Document Image Binarization Contest (DIBCO 2009),” *Document Analysis and Recognition, 2009. ICDAR '09. 10th International Conference on*, pp.1375-1382, 26-29 July 2009

[2] I. Pratikakis, B. Gatos, K. Ntirogiannis, “ICDAR 2011 Document Image Binarization Contest (DIBCO 2011),” *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pp.1506-1510, 18-21 Sept. 2011

[3] G. Agam, G. Bal, G. Frieder, O. Frieder, "Degraded document image enhancement", *Proc. SPIE 6500*, pp. C1–11, 2007

[4] C. Tomasi, R. Manduchi, "Bilateral filtering for gray and color images," *Computer Vision, 1998. Sixth International Conference on*, pp.839-846, 4-7 Jan 1998

[5] N. R. Howe, "Document binarization with automatic parameter tuning." *International Journal on Document Analysis and Recognition* (IJDAR) (2012): 1-12.

[6] T. Lelore, F. Bouchara, "Super-Resolved Binarization of Text Based on the FAIR Algorithm," *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pp.839-843, 18-21 Sept. 2011

[7] N. R. Howe, "A Laplacian Energy for Document Binarization," *Document Analysis and Recognition, International Conference on*, pp. 6-10, 2011 International Conference on Document Analysis and Recognition, 2011

[8] T. Obafemi-Ajayi, G. Agam, O. Frieder. 2010. "Historical document enhancement using LUT classification". *International Journal on Document Analysis and Recognition* (IJDAR)(March 2010), 1-17.

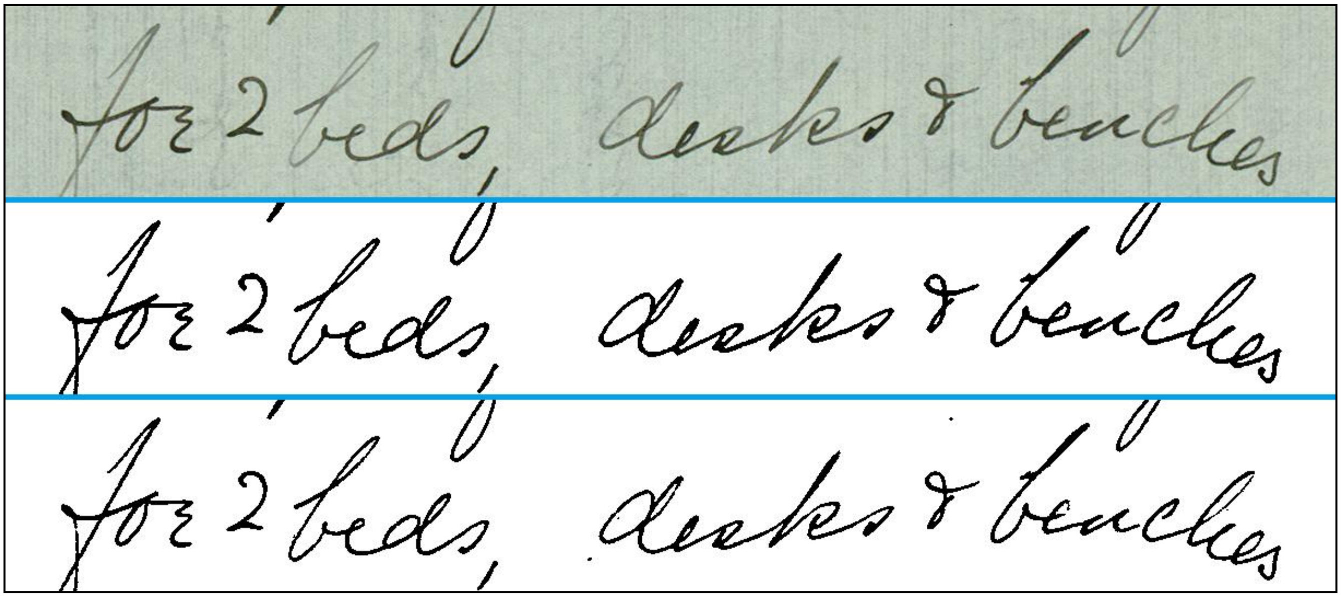


Figure 8: DIBCO 2011 HW2 Results: (top) Original image HW2 from DIBCO 2011, (middle) Ground Truth, (bottom) Results