

Document Image Binarization - Heavily Degraded Swedish and Icelandic Texts

Jonathan Kurén, Simon Leijon, Petter Sigfridsson, Martin Sundberg, and Hampus Widén

Uppsala University, Uppsala, Sweden

Abstract. Document image binarization is the process of converting a document into a bi-level document image. Document images commonly suffer from various degradations over time, rendering document image binarization a daunting task. The aim of this project was to investigate different methods for document image binarization, that could be used for documents that suffered from various degradations. Two different methods were tested, U-Net and Deepotsu. For each method two different datasets were used, a dataset with only bleedthrough images and a dataset with no bleedthrough images. When testing on degraded images it is shown that the models work decently for their specific purpose but struggles heavily when the background of the image is dark.

Keywords: Binarization · Neural Network · Image Analysis

1 Introduction

Document image binarization is the process of converting a document into a bi-level document image. The main goal of the process is the segmentation of the foreground text and the background of the document image, where the background is usually represented in white pixels and the text as black pixels. The field of image binarization is an actively researched area and is important for tasks such as optical character recognition, document layout analysis or simply for the purpose of enhancing and digitally storing degraded ancient manuscripts.

Document images commonly suffer from various degradations over time, rendering document image binarization a daunting task. Typically, a document image can be heavily degraded due to ink bleed-through, faded ink, wrinkles, stains, missing data, contrast variation, warping effect, and noise due to lighting variation during document scanning. Though document image binarization has been extensively studied, their performance for heavily degraded ancient manuscripts is significantly low. This project aim at developing, evaluating and improving document binarization methods, especially tailored for heavily degraded Swedish and Icelandic texts.

Current techniques for binarization can roughly be divided into two groups, local and global threshold methods. Global methods rely on a single threshold for the

whole document whereas local methods assign a threshold for a small region of the document images. There also exist deep learning binarization methods that makes use of deep learning or the combination of deep learning and threshold methods. This project will focus on methods involving deep learning.

2 Related work

Within the document image binarization field a multitude of algorithms exists for preprocessing, binarization and post-processing. Several studies [3–6] have been made in attempts to summarize the area and its current tools available. A major difficulty when performing binarization is the existence of various problems such as noise, low contrast between back- and foreground, bleed-through and degradation in general [3]. Preprocessing aims to solve this through different methods such as noise removal, background estimation and grayscale conversion.

As for the binarization itself there exists a variety of different methods ranging from traditional threshold methods to edge based methods, conditional random fields (CRF) and deep learning methods. One of the more frequently used traditional methods is Otsu [7] which is a parameter free global threshold method. Other popular choices include Niblack [8] and Sauvola [9] which are both adaptive (local) threshold methods.

Post-processing can be done to further improve the result of the binarization [3]. Typically, once the binarization mask has been produced it can either be improved by comparing to known characteristics about masks or after the actual binarization has happened by for example comparing the result to the grayscale image.

U-net is a pixel classification methodology developed by Ronneberger et al. based on convolutional neural networks [1]. The original use case for the methodology was in the area of biological image segmentation. However, this method has proven highly useful for document image binarization, winning DIBCO 2017 with a fairly large margin. The network architecture has two paths that forms the shape of a U where the downward path (encoding) is in line with traditional CNNs where downsampling is performed with convolutional/pooling layers and an activation function. The upward path (decoding) consists of upsampling the downsampled result to obtain an image in the same resolution as the input. This network architecture allows for classifying each pixel independently and as such predicting a binarized image with a resolution that equals that of the input image. An important aspect of U-net is that the network needs to learn the invariance of flipping, shifting, rotating, gray value variations etc. to properly predict different images. An appropriate method to use, especially in the case that there is not much training data, is data augmentation. The original authors points this out as an essential part of the process of training the U-net.

To solve the problem of document image binarization the DeepOtsu project divided it into two parts, enhancement and binarization [2]. The enhancement part handles the degradation of the document images and is built as a U-Net neural network. This will make the images clean and uniform. The neural network works on the images iteratively by using the output as input to fine-tune the result. Once the images are enhanced they are binarized with the Otsu method, which uses a global threshold which works well on uniform images.

3 Aim

The aim of this project was to investigate different methods for document image binarization, that could be used for documents that suffered from various degradations. These degradations include faded ink, stains, ink bleed-through, contrast variations and wrinkles. Even though document image binarization is a broadly studied subject, the performance for heavily degraded documents is low. This project uses scanned documents from The National Archives and The Regional States Archive in Uppsala that have been damaged by water and fire making them hard to read. The goal is to find a method to make these documents readable using image processing techniques and machine learning.

4 Methods

For U-Net and DeepOtsu, when inputting the data to the networks the images were divided into smaller images of 128x128 pixels since this is the size the model was made to work with.

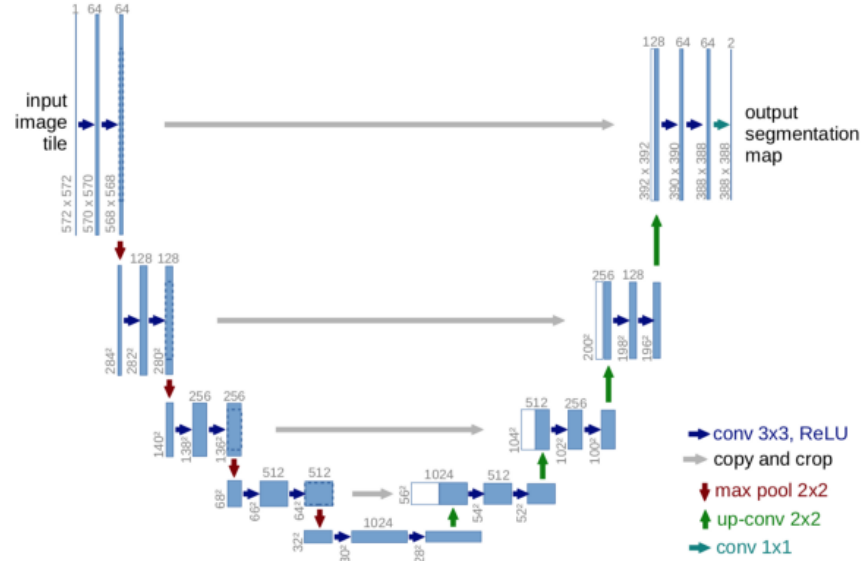
4.1 Otsu's method

Otsu's method is a global threshold method used to binarize an image. The method takes a grayscale image as input and calculates the threshold value which decides whether a pixel in the image should be classified as white or black. Once the threshold value has been calculated the pixels which are brighter than the threshold are painted white, and the pixels which are darker than the threshold are painted black.

4.2 U-Net

The implementation of U-Net is inspired by Edward Roes version of U-Net [10]. During the encoding part of the network there is a repeating pattern of two convolutions, each followed by a rectified linear unit (ReLU) and then downsampling by a max pooling operation before next repetition. During the decoding part the repeating pattern consists of an upsampling layer, a concatenation and then two convolutions, each followed by a ReLU. The final part of the network is a sigmoid function.

Fig. 1. U-net Architecture [1]



4.3 DeepOtsu

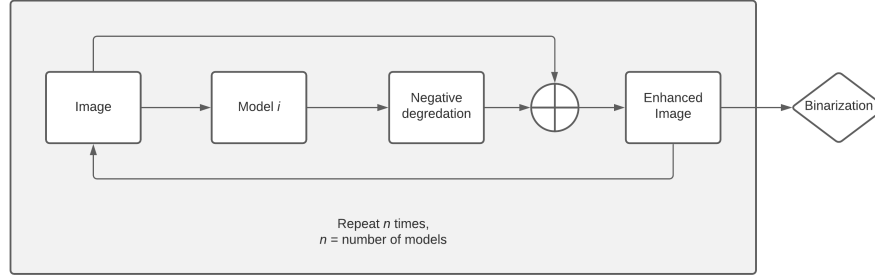
Our DeepOtsu method is based on the original DeepOtsu developed by Sheng He. The method is based on several deep neural networks (namely U-Nets) where the output from one neural network acts as the input to the next. A key difference to the standalone U-Net method and the U-Nets utilized in DeepOtsu is that we do not wish to learn a binarized version of the ground-truth. Rather, we wish to learn the degradations of the image, this is because if the binarization also happened in the neural network the output could not function as an input to the next neural network. After having passed through the image through several (4-5) networks all degradations ideally have been detected and removed from the image, giving us a clean enhanced version of the original image. The final step is then to perform binarization, where the chosen method is Otsu, a global threshold method. The end result is a binarized image with less degradations.

A key difference between our version and the original DeepOtsu is that we include BatchNormalization layers whereas the original does not. Furthermore the number of convolutional layers in between each MaxPooling layer differs.

To enhance the images with DeepOtsu both the **stacked refinement** and the **recursive refinement** methods described in [2] were tested. Stacked Refinement - the image is passed through different models each iteration. The stacked refinement implementation can be seen in Figure 2. Recursive Refinement - the image is fed through the *same* model each iteration. This can basically be seen

as stacked refinement but with the same model each iteration instead of separate ones.

Fig. 2. DeepOtsu Stacked Refinement.



5 Experiments

5.1 Evaluation metrics

The following evaluation metrics were used to measure the performance of the different binarization methods:

A F-Measure

The F-Measure is calculated using the precision and the recall and calculates the accuracy of the result. A high F-Measure indicates high accuracy of the prediction.

B PSNR

PSNR measures the peak error between the ground truth and the binarization. A high ratio indicates a high similarity between the images. [11]

C MPM

MPM measures the similarity between the prediction and the Ground truth on an object by object basis. The further the misclassified pixels are from ground truth objects border the score is penalized less. This leads to MPM being good at showing how well the algorithm is at identifying an objects border. [11]

D DRDM

Distance-Reciprocal Distortion measure (DRDM) measures the distortion of a prediction compared with the ground truth. This is done using a weighted matrix with its weights being determined by the reciprocal of a distance measured from the center pixel. This method matches the subjective ability of the human visual perception. [12]

5.2 Data Sets

The test data set is provided by The National Archives and consists of images of heavily degraded documents in Swedish and Icelandic from the 1700-1800s. This data set is significantly more damaged than the training data which has been used.

For the training, the data sets which have been used are Cuper, Bickley, Irish, Nabuco and Icdar. All of these data sets consists of handwritten documents with varying amount of degradations.

To improve the usefulness of the training with regards to the goal, a data set consisting of hand picked documents with heavy bleedthrough was constructed from all the data sets as well as a training data set consisting of the documents without bleedthrough. The idea behind this was that more specifically trained models might perform better on documents with specific types of degradations. Obviously a model trained on only bleedthrough images should perform better on images with heavy bleedthrough. The model without bleedthrough images is expected to perform on images where the lines are very thin because of degradations. It is expected that the models will perform worse on the DIBCO datasets, but since this is not the aim of the project we are fine with this.

5.3 DIBCO results

The following section demonstrates the proposed methods on benchmark datasets and compares it to existing methods. Models named NoBT refers to it being trained using a dataset with no bleedthrough ink, similarly a model named BT refers to it being trained using a dataset with bleedthrough. These results are included to provide a comparison between the developed models and other commonly used methods. An important note is that the goal of these experiments was not necessarily to produce a general model that outperforms across general benchmarks. The goal was rather to use state-of-the-art methods to produce models that can binarize difficult Swedish and Icelandic texts. With this background in mind it is still useful to highlight the general performance of the developed models on the different years of DIBCO data.

According to the DRD metric, U-Net NoBT performs the best across the data sets. However, Sauvola performs best on DIBCO 2010 while DeepOtsu BT performs the best on DIBCO 2011. Focusing on F-measure the developed models are beat by significant margin on DIBCO 2014&2010 data sets while they still produce decent results. The PSNR measure generally follows the F-measure ranking with a few exceptions such as DeepOtsu BT beating Otsu on DIBCO 2014 and DeepOtsu NoBT on DIBCO 2012 data.

Table 1. DIBCO 2009 results

Models	Measures			
	F-Measure (\uparrow)	PSNR (\uparrow)	MPM (\downarrow)	DRD (\downarrow)
U-Net NoBT	0.855	17.086	0.004	0.330
U-Net BT	0.859	17.114	0.002	0.440
DeepOtsu NoBT	0.825	15.282	0.013	0.340
DeepOtsu BT	0.877	17.761	0.001	0.446
Otsu	0.786	15.313	0.013	0.554
Niblack	0.496	8.89	0.082	0.399
Sauvola	0.825	15.860	0.003	0.465

Table 2. DIBCO 2010 results

Models	Measures			
	F-Measure (\uparrow)	PSNR (\uparrow)	MPM (\downarrow)	DRD (\downarrow)
U-Net NoBT	0.690	15.594	0.00262	0.228
U-Net BT	0.745	16.962	0.00156	0.216
DeepOtsu NoBT	0.710	14.979	0.012	0.322
DeepOtsu BT	0.715	16.587	0.00179	0.299
Otsu	0.855	17.538	0.00167	0.306
Niblack	0.439	8.927	0.084	0.342
Sauvola	0.689	15.479	0.00209	0.192

Table 3. DIBCO 2011 results

Models	Measures			
	F-Measure (\uparrow)	PSNR (\uparrow)	MPM (\downarrow)	DRD (\downarrow)
U-Net NoBT	0.740	14.042	0.0270	0.458
U-Net BT	0.847	16.834	0.00411	0.612
DeepOtsu NoBT	0.702	12.663	0.0533	0.515
DeepOtsu BT	0.825	16.339	0.00949	0.446
Otsu	0.819	15.683	0.0163	0.625
Niblack	0.491	8.787	0.0860	0.507
Sauvola	0.809	15.474	0.00730	0.827

Table 4. DIBCO 2012 results

Models	Measures			
	F-Measure (\uparrow)	PSNR (\uparrow)	MPM (\downarrow)	DRD (\downarrow)
U-Net NoBT	0.776	17.067	0.0332	0.130
U-Net BT	0.755	17.643	0.0017	0.210
DeepOtsu NoBT	0.805	16.558	0.0101	0.218
DeepOtsu BT	0.774	18.237	0.0011	0.230
Otsu	0.777	15.627	0.0272	0.186
Niblack	0.449	9.185	0.0744	0.214
Sauvola	0.753	16.488	0.00194	0.263

Table 5. DIBCO 2013 results

Models	Measures			
	F-Measure (\uparrow)	PSNR (\uparrow)	MPM (\downarrow)	DRD (\downarrow)
U-Net NoBT	0.829	17.591	0.00312	0.129
U-Net BT	0.760	18.227	0.00265	0.135
DeepOtsu NoBT	0.806	16.270	0.0126	0.165
DeepOtsu BT	0.856	18.670	0.00227	0.163
Otsu	0.800	16.622	0.0146	0.194
Niblack	0.456	8.840	0.0824	0.191
Sauvola	0.811	16.706	0.00372	0.225

Table 6. DIBCO 2014 results

Models	Measures			
	F-Measure (\uparrow)	PSNR (\uparrow)	MPM (\downarrow)	DRD (\downarrow)
U-Net NoBT	0.850	17.674	0.00103	0.0896
U-Net BT	0.820	17.679	0.00228	0.251
DeepOtsu NoBT	0.837	16.689	0.00864	0.254
DeepOtsu BT	0.852	18.764	0.00102	0.301
Otsu	0.916	18.714	0.00127	0.305
Niblack	0.540	9.332	0.0784	0.280
Sauvola	0.802	16.304	0.00178	0.381

5.4 Model Training

The training was done on the Alvis cluster, which is a part of the Swedish National Infrastructure for Computing. Either a NVIDIA T4 or a NVIDIA TESLA V100 was used, depending on what was available. The training for one U-Net model took approximately 5 hours and used 300 000 images with size 128x128.

The training for one DeepOtsu model took 50 hours and used 200 000 images with size 128x128.

5.5 Degraded images with U-Net

When testing the U-Net models on the images from The National Archives the results are decent. In Figure 3 it can be seen that both models are able to binarize the upper portion of the image well, but not the lower portion. The suspected reason for this is the dark background of the image. Both models perform similarly on this image which might indicate that the problem is a lack of training data with a dark background.

In Figure 4 it can be seen that the No Bleedthrough model performs much better than the Bleedthrough model. This is in accordance with our idea of the No Bleedthrough model doing well on images with thin lines. The parts of the image that the Bleedthrough model is not able to binarize is the thin orange text which looks like bleedthrough.

In Figure 5 the same problem as in Figure 3 can be seen, where the dark background makes the binarization difficult. The Bleedthrough model performs slightly better on this image which goes against the idea of the No Bleedthrough model performing better on images with thin lines. However when the background is dark no model performs well enough where you can read the text in the binarization so the comparison falls flat.

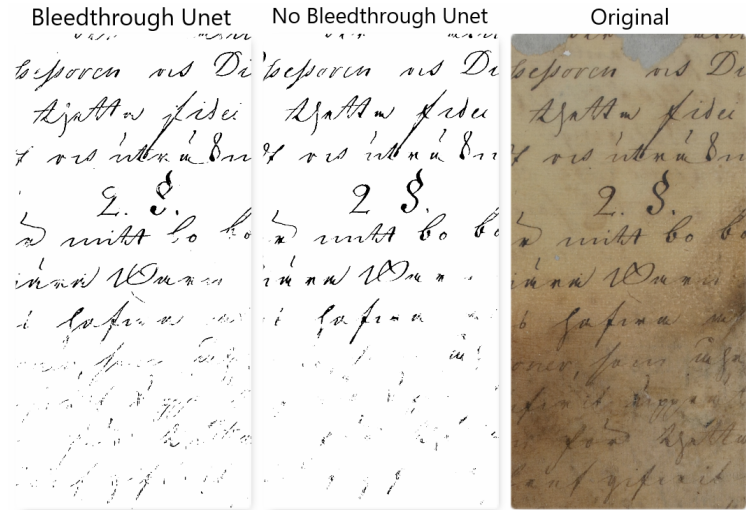


Fig. 3.

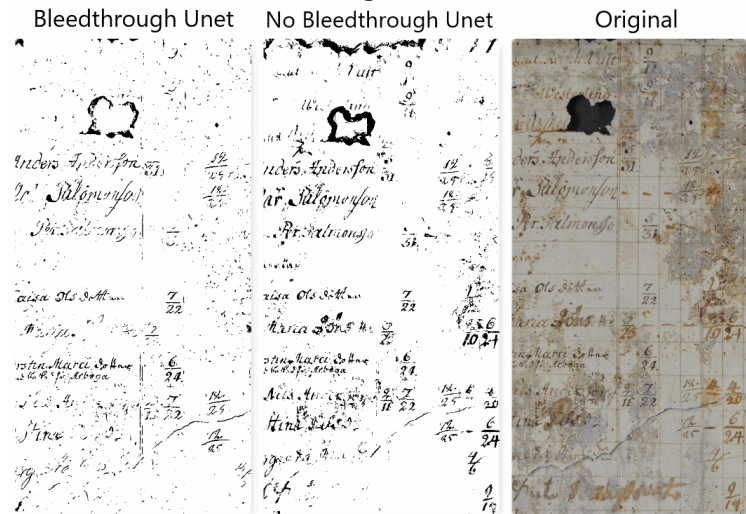


Fig. 4.

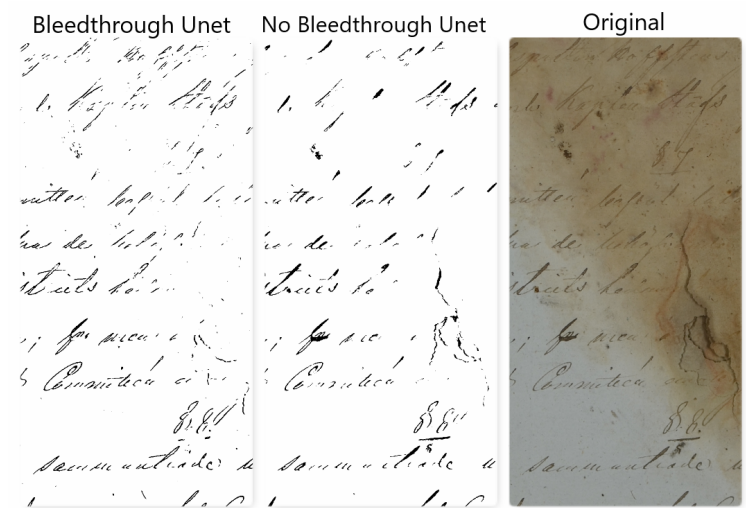


Fig. 5.

5.6 Degraded images with DeepOtsu

In Figure 6 it is apparent that the DeepOtsu models have the same problem as the U-Net models when it comes to dark backgrounds. This further reinforces the theory that it is the lack of training data that is the problem. A difference between the DeepOtsu models is how they behave when the background is dark. The No Bleedthrough model creates black areas meanwhile the Bleedthrough model keeps it white. The reason for this is unknown, but this might explain why the No Bleedthrough model have worse results when testing on the DIBCO data sets.

Figure 7 shows that the No Bleedthrough model performs slightly better than the Bleedthrough model on images with thin lines. Compared to the No Bleedthrough U-Net model, DeepOtsu performs slightly worse and includes much more black dots all over the images which makes it harder to read.

Figure 8 follows the same pattern as before where the dark area cannot be binarized well for both models. However if we compare DeepOtsu with U-Net, it is clear that U-Net performs much better on this image. When the background is dark the DeepOtsu models are not able to find any text, meanwhile U-Net still manages to find some parts of the text.

6 Discussion

6.1 Original DeepOtsu

Comparing the performance of the developed models (DeepOtsu BT/DeepOtsu NoBT) with the performance reported in the original paper of DeepOtsu [2] illustrates that the developed models underperform in the general case. This was to be expected since the intention was not to produce a general-purpose model. Even so, the models are not useless in the general case. The F-measure for DIBCO 2011 DeepOtsu BT is 0.825 vs. original DeepOtsu scoring a value of 0.934. This is of course a significant difference and other years show similar contrast.

6.2 Problems

A potential problem which was encountered is that the result differed depending on which images the training began with. This could also be due to how the weights in the networks are initialized. The difference in the result was significant and could be due to the difference in degradation in the training data sets. This made it hard to definitively say whether the increase/decrease in performance was due to a change of parameters or if it was simply a “lucky”/”unlucky” run.

Other issues regarding training concerned overfitting in the sense of increasing the amount of epochs too much, which decreased performance. The overfitted model produced less clear binarized documents compared to model trained with less epochs.

One way of increasing the performance of the models, w.r.t to training, is to experiment with the tuning of the hyperparameters, in particular the reduced learning rate. Experimenting with other activation functions than ReLu could also lead to an increase in performance.

Furthermore, there were also problems regarding the training data. One of these problems that was encountered was that the ICDAR data set makes up for a large part of the training data. The problem is that the ICDAR ground truths are created by a computer and in these ground truths, the more heavily damaged parts of the documents are removed and replaced with white areas. When training the model with these documents and their ground truths the model will learn to remove parts which are damaged rather than extracting the text from them.

Additionally, after the binarization of the documents from The National Archives it was clear that the models struggled with darker areas of the documents. This affects the score of the result negatively. This problem might be solved by adjusting how the enhancement works or by splitting the document into bright and dark parts. Another solution would be to augment the training data by creating

documents with dark areas.

Finally, to improve the binarization of the documents from The National Archives further, the model could be improved by training on data sets more similar to the test data. The training data used in this project had different degradations compared to the documents from The National Archives and were not as damaged. By using previous work in binarization, ground truths for the documents from The National Archives could be produced. These ground truths would not be perfect but could still be used to train a model to learn about severe degradations and their corresponding ground truths. This would however, run the risk of encountering the same issue as with the ICDAR data set and the generated ground truths would have to be manually reviewed before used as training data. At least to make sure that it is clear what the model is learning.

7 Conclusion

The goal of this project was to investigate different methods for document image binarization. The documents in question were of Swedish and Icelandic text from The National Archives and The Regional States Archive. Two different machine learning models using convolutional neural networks (CNNs) were used: U-net and DeepOtsu. U-Net is a single CNN method which was used to immediately binarize the image. DeepOtsu on the other hand uses multiple neural networks (U-nets) to remove degradation from the input image, at the end the simple binarization method Otsu is used to reach a binarized image. For both of the models two models were created, one which focused on bleed-through and another which didn't.

The experiments shows that there does not exist a single method which always outperforms the others. Furthermore the two DeepOtsu methods (DeepOtsu BT and DeepOtsu NoBT) both performed worse than what was reported in the original DeepOtsu paper [2]. However, their performance is better in multiple cases when compared to other binarization methods and perform quite well on the target images.

Current problems could be analyzed further and considered for future work. One such problem is the dependency on which images the neural network trained on first and/or it's initial weights. Multiple runs with the same dataset and parameters would lead to different result, making it hard to predict the behaviour of the model. Another problem is that we didn't have access to enough images which were similar to the documents from The National Archives, making it impossible for the neural network to learn how to handle such degradations.

References

1. Ronneberger, Olaf, Philipp Fischer and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation." MICCAI (2015).

2. He, Sheng and Lambert Schomaker. “DeepOtsu: Document Enhancement and Binarization using Iterative Deep Learning.” *Pattern Recognit.* 91 (2019): 379-390.
3. Tensmeyer, Chris and Martinez, Tony. “Historical Document Image Binarization: A Review” *SN Computer Science* (2020).
4. Bawa RK, Sethi GK. A review on binarization algorithms for camera based natural scene images. In: *International conference on advances in computing, communications and informatics*. ACM; 2012. p. 873–78
5. Sezgin M, Sankur B. Survey over image thresholding techniques and quantitative performance evaluation. *J Electronic Imaging*. 2004;13(1):146–66.
6. Ismail SM, Abdullah SNHS, Fauzi F. Statistical binarization techniques for document image analysis. *J Comput Sci.* 2018;14(1):23–36.
7. Otsu N. A threshold selection method from gray-level histograms. *Trans Syst Man Cybern.* 1979;9(1):62–6.
8. Niblack W. *An introduction to digital image processing*. Birkerod: Strandberg Publishing Company; 1985
9. Sauvola J, Pietikäinen M. Adaptive document image binarization. *Pattern Recognit.* 2000;33(2):225–36.
10. Edward Roe, Aged Document Binarization Using the U-Net Architecture, https://medium.com/@er_95882/aged-document-binarization-using-the-u-net-architecture-fa171cba6bd2. Last accessed 16 Dec 2021
11. B. Gatos, K. Ntirogiannis, I. Pratikakis. Icdar 2009 document image binarization contest (dibco 2009). in *Document Analysis and Recognition, 2009 International Conference on*. IEEE, 2009, pp. 1375–1382
12. Lu, Haiping Kot, Alex Shi, Y.Q.. (2004). Distance-Reciprocal Distortion Measure for Binary Document Images. *Signal Processing Letters, IEEE.* 11. 228 - 231. 10.1109/LSP.2003.821748.