FD-Net: A Fully Dilated Convolutional Network for Historical Document Image Binarization

Wei Xiong[1,2](✉), Ling Yue[1], Lei Zhou[1], Liying Wei[1], and Min Li[1,2]

[1] School of Electrical and Electronic Engineering, Hubei University of Technology,
Wuhan 430068, Hubei, China
xw@mail.hbut.edu.cn

[2] Department of Computer Science and Engineering, University of South Carolina,
Columbia, SC 29201, USA

Abstract. Binarization is a key step in document analysis and archiving. The state-of-the-art models for document image binarization are variants of the encoder-and-decoder architecture, such as *fully convolutional network* (FCN) and U-Net. Despite their success, they still suffer from two challenges: (1) max-pooling or strided convolution reduces the spatial resolution of the intermediate feature maps, which may lead to information loss, and (2) interpolation or transposed convolution attempts to restore the feature maps to the desired spatial resolution, which may also result in pixelation. To overcome these two limitations, we propose a *fully dilated convolutional network*, termed FD-Net, using atrous convolutions instead of downsampling or upsampling operations. We have conducted extensive experiments on the recent DIBCO (*document image binarization competition*) and H-DIBCO (*handwritten document image binarization competition*) benchmark datasets. The experimental results show that our proposed FD-Net outperforms other state-of-the-art techniques by a large margin. The source code and pre-trained models are publicly available at https://github.com/beargolden/FD-Net.

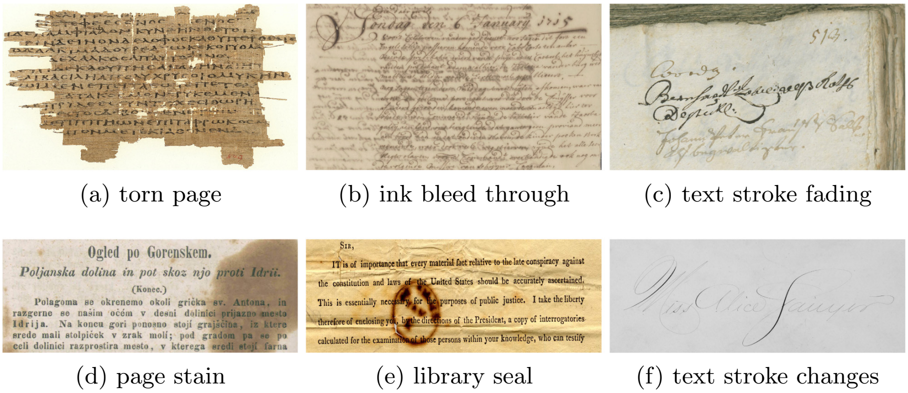
Keywords: Historical document image binarization · Document image segmentation · *Fully dilated convolutional network* (FD-Net) · Dilated convolution · Atrous convolution

1 Introduction

Document image binarization (also referred to as segmentation or thresholding) aims to extract text pixels from complex document background, which plays an important role in *document analysis and recognition* (DAR) systems. It converts a color or grayscale image into a binary one, essentially reducing the information contained within the image, and thus greatly reducing the disk storage capacity as well as network transmission bandwidth. It is widely considered to be one of

H. Ma et al. (Eds.): PRCV 2021, LNCS 13019, pp. 518–529, 2021.
https://doi.org/10.1007/978-3-030-88004-0_42

the most important pre-processing steps, and the performance of binarization will directly affect the accuracy of subsequent tasks, such as page layout analysis, machine-printed or handwritten character recognition. It also helps to resolve the conflict between document conservation and cultural heritage.


(a) torn page

(b) ink bleed through

(c) text stroke fading

(d) page stain

(e) library seal

(f) text stroke changes

Fig. 1. Historical document image samples from recent DIBCO and H-DIBCO benchmark datasets

The thresholding of high-quality images is simple, but the binarization of historical document images is quite challenging. The reason is that the latter suffers from severe degradation, such as torn pages, ink bleed through, text stroke fading, page stains, and artifacts, as shown in Fig. 1. In addition, variations in the color, width, brightness, and connectivity of text strokes in degraded handwritten manuscripts further increase the difficulty of binarization.

The *state-of-the-art* (SOTA) models for document image binarization are variants of the encoder-and-decoder architecture, such as *fully convolutional network* (FCN) [1] and U-Net [2]. These segmentation models have 3 key components in common: an encoder, a decoder, and skip connections. In the encoder, consecutive of convolutions and downsampling (e.g., max-pooling or strided convolution) are performed. This helps extract higher-level features, but reduces the spatial resolution of intermediate feature maps, which may lead to information loss. In the decoder, repeated combination of upsampling (e.g., bilinear interpolation) and convolutions are conducted to restore feature maps to the desired spatial resolution, which may also result in pixelation or texture smoothing. Therefore, after each upsampling operation, feature maps with the same level are merged by skip connections, which transfer localization information from the encoder to the decoder. In addition, sampling operations like max-pooling and bilinear interpolation are not learnable.

To overcome the aforementioned problems, we present FD-Net, a *fully dilated convolutional network* for degraded historical document image binarization. The proposed segmentation model removes all the downsampling and upsampling operations, and employs dilated convolutions (also known as atrous convolutions) instead. Therefore, the proposed segmentation model contains only convolutional and dilated convolutional layers, which are fully trainable. In this way, the spatial resolutions of all the intermediate layers are identical, but without significantly increasing the number of model parameters.

Our contributions are two folds. First, we propose a new paradigm that replaces downsampling or upsampling operations with dilated convolutions. It can achieve promising pixel-wise labeling results on various degraded historical document images. Second, we investigate hybrid dilation rate settings to alleviate the grid effect in dilated convolution.

The rest of this paper is organized as follows. Section 2 briefly reviews the related work on document image binarization. Section 3 describes our proposed fully dilated convolutional neural network in detail. The experimental results and analysis are presented in Sect. 4, and Sect. 5 concludes the paper.

2 Related Work

Existing document image binarization methods can be classified as global thresholding, local adaptive thresholding, and hybrid approaches [3].

The global thresholding method uses a single threshold to classify the pixels of a document image into two classes, namely text and background. The Otsu's [4] method is one of the best known global thresholding techniques. It uses the grayscale histogram of an image to select an optimal threshold that makes the variance within each class as small as possible and the variance between the two classes as large as possible. The Otsu's method is fast but ineffective and has poor noise immunity when dealing with low-quality images.

The local adaptive thresholding method can handle more complex cases, and automatically computes local thresholds based on the grayscale distribution within a neighborhood window around a pixel. The Niblack's [5] method uses a smoothing window mechanism, where the local threshold is determined by the mean and standard deviation of the grayscale values within the window centered at each pixel. This method is good at segmenting low-contrast text, but since only local information is considered, it is more likely to treat background noise as foreground text as well. The Sauvola's [6] and Wolf's [7] methods overcome the drawbacks of Niblack's counterpart. They are based on the assumption that the gray value of text pixels is close to 0 and the gray value of background pixels is close to 255. It makes the threshold smaller for background points with higher gray values and the same standard deviation, thus filtering out some distracting textures and noises in the background, but the binarization is still not effective in the case of low contrast between the foreground and background.

Hybrid methods for the binarization of historical document images have also been developed. Su et al. [8] present a document image binarization method using *local maximum and minimum* (LMM). The document text is segmented by constructing a contrast image and then detecting high-contrast pixels that typically lie around text stroke boundaries and using local thresholds that are estimated from the detected high-contrast pixels within a local neighborhood window. Jia et al. [9] propose a document image binarization method based on *structural symmetric pixels* (SSPs), which are located at the edges of text strokes and can be extracted from those with large gradient values and opposite gradient directions. Finally, a multiple local threshold voting-based framework is used to further determine whether each pixel belongs to the foreground or background. The contrast or edge-based segmentation methods do not work well for binarization of degraded images with complex document background, e.g., low contrast, gradients, and smudges.

Howe [10] presents an energy-based segmentation method, which treats each image pixel as a node in a connected graph, and then applies the max-flow/min-cut algorithm to partition the connected graph into two regions to determine the text and background pixels. Mesquita et al. [11], Kligler et al. [12], and Xiong et al. [13,14] propose different document enhancement techniques, followed by Howe's binarization method, to provide guidance for text and background segmentation, respectively. In addition, Chen et al. [15] and Xiong et al. [16] propose the use of *support vector machines* (SVMs) for statistical learning-based segmentation. Bhowmik et al. [17] introduce a *document image binarization inspired by game theory* (GiB). However, the main drawback of these methods is that only handcrafted features are employed to obtain segmentation results. Therefore, it is difficult to design representative features for different applications, and handcrafted features work well for one type of image, but may fail on another.

Deep learning-based binarization of degraded document images is a hot topic and trend of current research. Tensmeyer and Martinez [18] present a multi-scale FCN architecture with pseudo F-measure loss. Zhou et al. [19] also explore a multi-scale deep contextual convolutional neural network with densely connected *conditional random fields* (CRFs) for semantic segmentation. Vo et al. [20] propose a supervised binarization method for historical document images based on hierarchical *deep supervised networks* (DSNs). Calvo-Zaragoza and Gallego [21] present a *selectional auto-encoder* (SAE) approach for document image binarization. Bezmaternykh et al. [22] present a historical document image binarization method based on U-Net [2], a convolutional neural network originally designed for biomedical image segmentation. Zhao et al. [23] consider binarization as an image-to-image generation task and propose a method for historical document image binarization using *conditional generative adversarial networks* (cGANs). Peng et al. [24] propose a deep learning framework to infer the probabilities of text regions by a multi-resolution attention-based model, and then fed into a *convolutional conditional random field* (ConvCRF) to obtain the final binary images. Xiong et al. [25] present an improved semantic segmentation model, called DP-LinkNet, which adopts *hybrid dilated convolution* (HDC) and *spatial pyramid pooling* (SPP) modules between the encoder and the decoder, for more accurate binarization of degraded historical document images.

3 Proposed Network Architecture: FD-Net

The proposed fully dilated convolutional network model, referred to as FD-Net, is shown in Fig. 2. As can be seen from the figure, it consists of 3 main components, namely an encoder, a decoder, and skip connections. What distinguishes

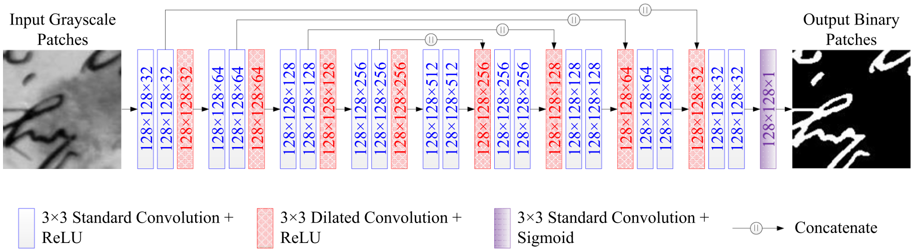


Fig. 2. The proposed FD-Net architecture

our proposed FD-Net from other SOTA models for document image binarization is that the proposed model does not contain downsampling or upsampling layers. The pooling operation or strided convolution reduces the spatial resolution of intermediate feature maps, which may lead to internal data structure missing or spatial hierarchical information loss; while the interpolation operation or strided deconvolution attempts to restore the feature maps to the desired spatial resolution, which may also result in pixelation or checkerboard artifacts. In addition, the pooling and interpolation operations are deterministic (a.k.a. not learnable or trainable).

To overcome the above problems, an intuitive approach is to simply remove those downsampling and upsampling layers from the model, but this will also decrease the receptive field size and thus severely reduce the amount of context. For instance, a stack of three 3×3 convolutional layers is equivalent to the regularization of a 7×7 convolutional layer. That's why pooling operations exist for increasing the receptive field size, and upsampling for pixel-wise prediction. Fortunately, dilated convolution can compensate for this deficiency, and it has been proven to be effective in semantic segmentation [26]. For this reason, we replace all the downsampling and upsampling layers with dilated convolutions. Therefore, the spatial resolution of all the intermediate convolutional layers in the proposed model is identical.

The encoder subnetwork comprises 4 consecutive convolutional blocks. Each encoding block includes two 3×3 standard convolutional layers followed by a 3×3 dilated convolutional layer (the dilation rate settings will be discussed in Subsect. 3.1). The number of channels or feature maps after each encoding block is doubled, so that the network can effectively learn higher-level abstract feature representations. The central block contains two 3×3 standard convolutional layers, and connects the encoder and decoder. The decoder subnetwork, similar to the encoder counterpart, also consists of 4 nearly symmetrical convolutional blocks. Each decoding block includes a 3×3 dilated convolutional layer, a merge layer that concatenates the feature maps of the decoder with those of the corresponding encoder by skip connections, and two 3×3 standard convolutional layers. The number of feature maps before each decoding block is halved. The ReLU (*rectified linear unit*) activation function is used in all the aforementioned convolutional layers. At the end, a 3 × 3 standard convolutional layer with Sigmoid activation function is adopted to generate the resulting binary image patches.

3.1 Hybrid Dilation Rate Settings

We introduce a simple hybrid dilation rate solution by setting different dilation rates to avoid the gridding effect [27,28]. In addition, choosing an appropriate dilation rate can effectively increase the receptive field size and also improve the segmentation accuracy. The purpose of our hybrid dilation rate setting is to make the final receptive field size of all the successive convolutions (including dilated ones) completely cover a specific local neighborhood. The maximum distance M between two non-zero kernel weights is defined as:

$$M_i \Leftarrow max[M_{i+1} - 2r_i, M_{i+1} - 2(M_{i+1} - r_i), r_i] \tag{1}$$

with $M_n = r_n$.

Instead of using the same dilation rates or those with a common factor relationship among all the convolutional layers, we set the dilation rates of the 3 layers in each encoding block to [1, 1, r_i], where the values of r_i used in the 4 encoding blocks are set to [2, 3, 5, 7], respectively. The dilation rates of the decoder subnetwork are set in the reverse order of the corresponding encoder subnetwork. Since spatial resolutions of all feature maps are the same, skip connections are essentially merging the features with different receptive field sizes.

3.2 Implementation Details

Given a color antique document image, it is first converted to its grayscale counterpart, then cropped to a patch size of 128 × 128 and fed into our proposed FD-Net model. The output binary patches are seamlessly stitched together to generate the resulting binary image.

We combine the Dice loss with the standard *binary cross-entropy* (BCE) loss. Combining these two metrics allows for some diversity in the loss function, while benefiting from the stability of the BCE. The overall loss function is defined as:

$$L = \underbrace{1 - \frac{2\sum_{n=1}^{N} y_n \hat{y}_n}{\sum_{n=1}^{N} y_n^2 + \sum_{n=1}^{N} \hat{y}_n^2}}_{L_{Dice}} \underbrace{- \frac{1}{N} \sum_{n=1}^{N} [y_n \log \hat{y}_n + (1 - y_n) \log(1 - \hat{y}_n)]}_{L_{BCE}} \tag{2}$$

where N is the number of image pixels, y_n and $\hat{y}_n$ are the *ground truth* (GT) and predicted segmentation, respectively.

The Adam optimization algorithm is adopted in our deep learning model. The initial learning rate defaults to 0.001, and the exponential decay rates for the first and second moment estimates are set to 0.9 and 0.999, respectively. We monitor the cost function values of the validation data, and if no improvement is seen for 10 epochs, the learning rate is reduced to half. We also use an early stop strategy once the learning stagnates for 20 consecutive epochs.

We collect 50 degraded historical document images from the *recognition and enrichment of archival documents* (READ) project[1] as training data. The Bickley Diary dataset is used for the ablation study, while the DIBCO and H-DIBCO 2009–2019 benchmark datasets are used as test data.

Table 1. Ablation study of FD-Net on the Bickley Diary dataset with varying dilation rate settings (image patch size: 128 × 128, and batch size: 32)

Network model	Dilation rates	# of 1st layer channels	Validation loss	Validation accuracy	# of model parameters
U-Net	–	32	0.0577	0.9903	8,630,177
U-Net	–	64	0.0541	0.9917	34,512,705
FD-Net	2,2,2,2	32	0.0600	0.9899	9,414,017
FD-Net	2,3,5,7	32	**0.0514**	**0.9931**	9,414,017
FD-Net	2,4,8,16	32	0.0524	0.9914	9,414,017

During the training and testing phases, the traditional color-to-gray method is performed, with no other pre-processing or post-processing. Our implementation does not apply any data augmentation techniques either.

4 Experiments

4.1 Ablation Study

In this study, we use the Bickley Diary dataset to evaluate the impact of hybrid dilation rates on the performance of our proposed FD-Net. It consists of 92 badly degraded handwritten travel diary documents, 7 of which have GT images. We crop these 7 historical document images into 1764 patches with corresponding GT ones, 20% of which to be used as validation data.

The BCE-Dice loss and accuracy metrics are used to measure the performance of our deep learning model. The loss is a sum of the errors made for each sample in the training or validation set. The loss value implies how poorly or well a model performs after each iteration of optimization, so higher loss is the worse for any model. The accuracy is usually determined after the model parameters and is calculated in the form of a percentage. It is a measure of how accurate your model's prediction is compared to the true data.

The experimental results of the ablation study are shown in Table 1. As can be seen from the table, the segmentation performance of our proposed FD-Net is basically similar to or even better than that of the vanilla U-Net architecture with the same number of layers, but the number of parameters of our FD-Net model is almost the same or much less than that of the U-Net. Among all the compared models, the FD-Net with our proposed hybrid dilation rates performs the best. The results of our ablation experiments suggest that the correct setting of the dilated convolutions can not only help maintain the receptive field size, but also further improve the segmentation accuracy.

[1] https://read.transkribus.eu/.

4.2 More Segmentation Experiments

In this experiment, we use 10 document image binarization competition datasets to evaluate the segmentation performance of our proposed FD-Net. The DIBCO 2009 [29], 2011 [30], 2013 [31], 2017 [32], 2019 [33] and H-DIBCO 2010 [34], 2012 [35], 2014 [36], 2016 [37], 2018 [38] benchmark datasets consist of 90 handwritten, 36 machine-printed, 10 Iliad papyri document images and their corresponding GT images. The 10 datasets contain representative historical document degradation, such as fragmented pages, ink bleed through, background texture, page stains, text stroke fading, and artifacts.

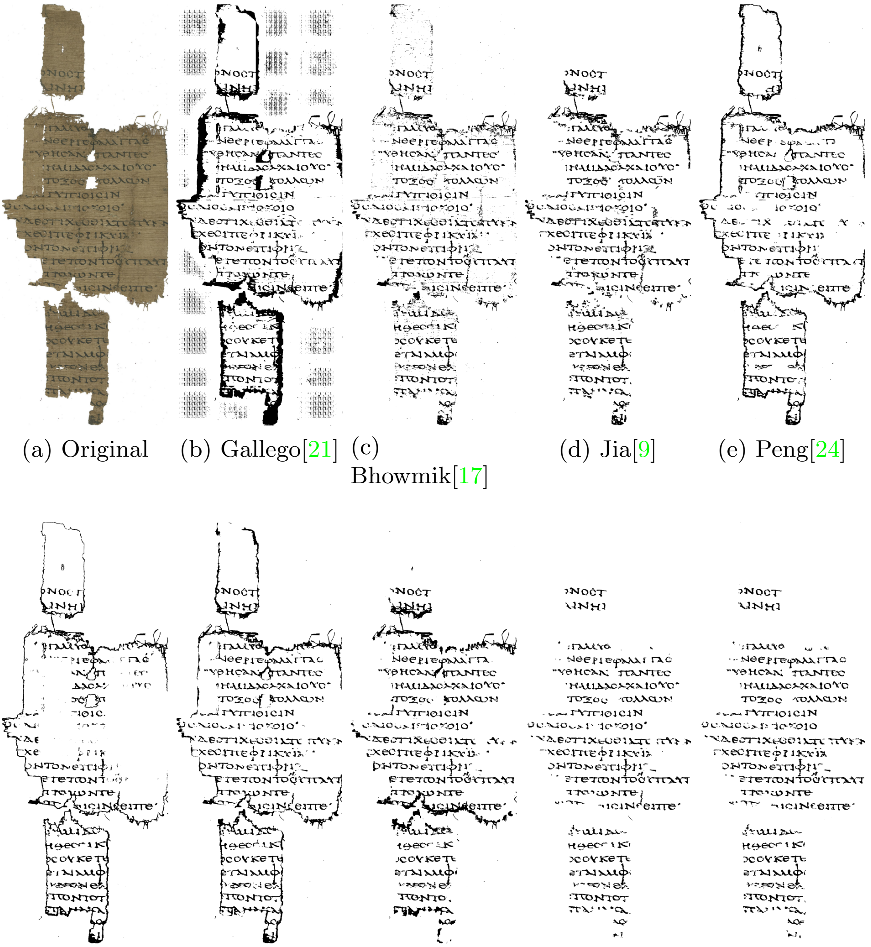
Table 2. Performance evaluation results of our proposed method against SOTA techniques on the 10 DIBCO and H-DIBCO test datasets

Method	FM↑ (%)	pFM (%)	PSNR (dB)	NRM (%)	DRD	MPM (‰)
Gallego's SAE [21]	79.221	81.123	16.089	9.094	9.752	11.299
Bhowmik's GiB [17]	83.159	87.716	16.722	8.954	8.818	7.221
Jia's SSP [9]	85.046	87.245	17.911	5.923	9.744	9.503
Peng's woConvCRF [24]	86.089	87.397	18.989	6.429	4.825	4.176
Zhao's cGAN [23]	87.447	88.873	18.811	5.024	5.564	5.536
Vo's DSN [20]	88.037	90.812	18.943	6.278	4.473	3.213
Bezmaternykh's UNet [22]	89.290	90.534	21.319	5.577	3.286	1.651
Proposed FD-Net	**95.254**	**96.648**	**22.836**	**3.224**	**1.219**	**0.201**

We adopt evaluation metrics used in DIBCO and H-DIBCO competitions to evaluate the performance of our proposed method. The evaluation metrics are FM (*F-measure*), pFM (*pseudo F-measure*), PSNR (*peak signal-to-noise ratio*), NRM (*negative rate metric*), DRD (*distance reciprocal distortion metric*), and MPM (*misclassification penalty metric*). The first 2 metrics (FM and pFM) reach the best value at 1 and the worst at 0. The PSNR measures how close a binarized image to the GT image, and therefore, the higher the value, the better. In contrast to the former 3 metrics, the binarization quality is better for lower NRM, DRD, and MPM metrics. Due to space limitations, we omit definitions of those evaluation metrics, but readers can refer to [30,33,34] for more details.

The proposed FD-Net is compared with Jia's SSP [9], Bhowmik's GiB [17], Vo's DSN [20], Gallego's SAE [21], Bezmaternykh's U-Net [22], Zhao's cGAN [23], and Peng's attention-based[24] techniques on all the 10 DIBCO and H-DIBCO datasets, and the evaluation results are listed in Table 2. It can be seen from the table that all the evaluation measures of our proposed method achieve the best results on all the 10 test datasets. Compared with Bezmaternykh's U-Net [22], the FM, pFM, PSNR, NRM, DRD, and MPM of our proposed method are 5.963%, 6.115%, 1.516dB, 2.353%, 2.067 and 1.451‰better than those of the second best technique, respectively. This also implies that our proposed FD-Net architecture outperforms U-Net, and produces more accurate segmentation.

Figure 3 further displays the resulting binary images generated by different techniques. From the figure we can see that Gallego's SAE [21] tends to produce



(a) Original (b) Gallego[21] (c) Bhowmik[17] (d) Jia[9] (e) Peng[24]

(f) Zhao[23] (g) Vo[20] (h) Bezmaternykh[22] (i) Proposed (j) GT

Fig. 3. Binarization results of all evaluation techniques for CATEGORY2_20 in DIBCO 2019 dataset

ghost text pixels in the true background region. Bhowmik's GiB [17], Peng's attention-based segmentation (woConvCRF) [24], Zhao's cGAN [23], and Vo's DSN [20] have difficulty in removing the edges of fragmented pages. Compared to Jia's SSP [9] and Bezmaternykh's U-Net [22], our proposed FD-Net can produce better visual quality by preserving most text strokes and eliminating possible noise.

5 Conclusion

In this paper, we present a fully dilated convolutional neural network, termed FD-Net, for more accurate binarization of degraded historical document images. The superior performance is attributed to its dilated convolutional architecture and skip connection, which is designed to address two major challenges faced by current segmentation models: (1) internal data structure missing or spatial hierarchical information loss, and (2) max-pooling and interpolation operations are not learnable or trainable. We conducted extensive experiments to evaluate the performance of our proposed FD-Net on the recent DIBCO and H-DIBCO benchmark datasets. Results show that the proposed method outperforms other SOTA techniques by a large margin.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China (61571182, 61601177), the Natural Science Foundation of Hubei Province, China (2019CFB530), and China Scholarship Council (201808420418).

References

1. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 640–651 (2017). https://doi.org/10.1109/tpami.2016.2572683
2. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
3. Eskenazi, S., Gomez-Krämer, P., Ogier, J.M.: A comprehensive survey of mostly textual document segmentation algorithms since 2008. Pattern Recogn. **64**, 1–14 (2017). https://doi.org/10.1016/j.patcog.2016.10.023
4. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1), 62–66 (1979). https://doi.org/10.1109/tsmc.1979.4310076
5. Niblack, W.: An Introduction to Digital Image Processing. Prentice-Hall International Inc., Englewood Cliffs (1986)
6. Sauvola, J., Pietikäinen, M.: Adaptive document image binarization. Pattern Recogn. **33**(2), 225–236 (2000). https://doi.org/10.1016/s0031-3203(99)00055-2
7. Wolf, C., Jolion, J.M.: Extraction and recognition of artificial text in multimedia documents. Pattern Anal. Appl. **6**(4), 309–326 (2003). https://doi.org/10.1007/s10044-003-0197-7
8. Su, B., Lu, S., Tan, C.L.: Binarization of historical document images using the local maximum and minimum. In: 9th IAPR International Workshop on Document Analysis Systems (DAS 2010), pp. 159–165. https://doi.org/10.1145/1815330.1815351
9. Jia, F., Shi, C., He, K., Wang, C., Xiao, B.: Degraded document image binarization using structural symmetry of strokes. Pattern Recogn. **74**, 225–240 (2018). https://doi.org/10.1016/j.patcog.2017.09.032
10. Howe, N.R.: Document binarization with automatic parameter tuning. Int. J. Doc. Anal. Recogn. **16**(3), 247–258 (2013). https://doi.org/10.1007/s10032-012-0192-x
11. Mesquita, R.G., Silva, R.M.A., Mello, C.A.B., Miranda, P.B.C.: Parameter tuning for document image binarization using a racing algorithm. Expert Syst. Appl. **42**(5), 2593–2603 (2015). https://doi.org/10.1016/j.eswa.2014.10.039

12. Kligler, N., Katz, S., Tal, A.: Document enhancement using visibility detection. In: 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018), pp. 2374–2382. https://doi.org/10.1109/cvpr.2018.00252
13. Xiong, W., Jia, X., Xu, J., Xiong, Z., Liu, M., Wang, J.: Historical document image binarization using background estimation and energy minimization. In: 24th International Conference on Pattern Recognition (ICPR 2018), pp. 3716–3721. https://doi.org/10.1109/icpr.2018.8546099
14. Xiong, W., Zhou, L., Yue, L., Li, L., Wang, S.: An enhanced binarization framework for degraded historical document images. EURASIP J. Image Video Process. **2021**(1), 1–24 (2021). https://doi.org/10.1186/s13640-021-00556-4
15. Chen, X., Lin, L., Gao, Y.: Parallel nonparametric binarization for degraded document images. Neurocomputing **189**, 43–52 (2016). https://doi.org/10.1016/j.neucom.2015.11.040
16. Xiong, W., Xu, J., Xiong, Z., Wang, J., Liu, M.: Degraded historical document image binarization using local features and support vector machine (SVM). Optik **164**, 218–223 (2018). https://doi.org/10.1016/j.ijleo.2018.02.072
17. Bhowmik, S., Sarkar, R., Das, B., Doermann, D.: GiB: a game theory inspired binarization technique for degraded document images. IEEE Trans. Image Process. **28**(3), 1443–1455 (2019). https://doi.org/10.1109/tip.2018.2878959
18. Tensmeyer, C., Martinez, T.: Document image binarization with fully convolutional neural networks. In: 14th IAPR International Conference on Document Analysis and Recognition (ICDAR 2017), pp. 99–104. https://doi.org/10.1109/icdar.2017.25
19. Zhou, Q., et al.: Multi-scale deep context convolutional neural networks for semantic segmentation. World Wide Web **22**(2), 555–570 (2018). https://doi.org/10.1007/s11280-018-0556-3
20. Vo, Q.N., Kim, S.H., Yang, H.J., Lee, G.: Binarization of degraded document images based on hierarchical deep supervised network. Pattern Recogn. **74**, 568–586 (2018). https://doi.org/10.1016/j.patcog.2017.08.025
21. Calvo-Zaragoza, J., Gallego, A.J.: A selectional auto-encoder approach for document image binarization. Pattern Recogn. **86**, 37–47 (2019). https://doi.org/10.1016/j.patcog.2018.08.011
22. Bezmaternykh, P.V., Ilin, D.A., Nikolaev, D.P.: U-Net-bin: hacking the document image binarization contest. Comput. Optics **43**(5), 825–832 (2019). https://doi.org/10.18287/2412-6179-2019-43-5-825-832
23. Zhao, J., Shi, C., Jia, F., Wang, Y., Xiao, B.: Document image binarization with cascaded generators of conditional generative adversarial networks. Pattern Recogn. **96** (2019). https://doi.org/10.1016/j.patcog.2019.106968
24. Peng, X., Wang, C., Cao, H.: Document binarization via multi-resolutional attention model with DRD loss. In: 15th IAPR International Conference on Document Analysis and Recognition (ICDAR 2019), pp. 45–50. https://doi.org/10.1109/icdar.2019.00017
25. Xiong, W., Jia, X., Yang, D., Ai, M., Li, L., Wang, S.: DP-LinkNet: a convolutional network for historical document image binarization. KSII Trans. Internet Inf. Syst. **15**(5), 1778–1797 (2021). https://doi.org/10.3837/tiis.2021.05.011
26. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2018). https://doi.org/10.1109/tpami.2017.2699184

27. Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., Cottrell, G.: Understanding convolution for semantic segmentation. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1451–1460. https://doi.org/10.1109/wacv.2018.00163
28. Wang, Z., Ji, S.: Smoothed dilated convolutions for improved dense prediction. Data Min. Knowl. Disc. **35**(4), 1470–1496 (2021). https://doi.org/10.1007/s10618-021-00765-5
29. Gatos, B., Ntirogiannis, K., Pratikakis, I.: ICDAR 2009 document image binarization contest (DIBCO 2009). In: 10th International Conference on Document Analysis and Recognition (ICDAR 2009), pp. 1375–1382. https://doi.org/10.1109/icdar.2009.246
30. Pratikakis, I., Gatos, B., Ntirogiannis, K.: ICDAR 2011 document image binarization contest (DIBCO 2011). In: 11th International Conference on Document Analysis and Recognition (ICDAR 2011), pp. 1506–1510. https://doi.org/10.1109/icdar.2011.299
31. Pratikakis, I., Gatos, B., Ntirogiannis, K.: ICDAR 2013 document image binarization contest (DIBCO 2013). In: 12th International Conference on Document Analysis and Recognition (ICDAR 2013), pp. 1471–1476. https://doi.org/10.1109/icdar.2013.219
32. Pratikakis, I., Zagoris, K., Barlas, G., Gatos, B.: ICDAR 2017 competition on document image binarization (DIBCO 2017). In: 14th International Conference on Document Analysis and Recognition (ICDAR 2017), pp. 1395–1403. https://doi.org/10.1109/icdar.2017.228
33. Pratikakis, I., Zagoris, K., Karagiannis, X., Tsochatzidis, L., Mondal, T., Marthot-Santaniello, I.: ICDAR 2019 competition on document image binarization (DIBCO 2019). In: 15th International Conference on Document Analysis and Recognition (ICDAR 2019). https://doi.org/10.1109/icdar.2019.00249
34. Pratikakis, I., Gatos, B., Ntirogiannis, K.: H-DIBCO 2010 - handwritten document image binarization competition. In: 12th International Conference on Frontiers in Handwriting Recognition (ICFHR 2010), pp. 727–732. https://doi.org/10.1109/icfhr.2010.118
35. Pratikakis, I., Gatos, B., Ntirogiannis, K.: ICFHR 2012 competition on handwritten document image binarization (H-DIBCO 2012). In: 13th International Conference on Frontiers in Handwriting Recognition (ICFHR 2012), pp. 817–822. https://doi.org/10.1109/icfhr.2012.216
36. Ntirogiannis, K., Gatos, B., Pratikakis, I.: ICFHR 2014 competition on handwritten document image binarization (H-DIBCO 2014). In: 14th International Conference on Frontiers in Handwriting Recognition (ICFHR 2014), pp. 809–813. https://doi.org/10.1109/icfhr.2014.141
37. Pratikakis, I., Zagoris, K., Barlas, G., Gatos, B.: ICFHR 2016 handwritten document image binarization contest (H-DIBCO 2016). In: 15th International Conference on Frontiers in Handwriting Recognition (ICFHR 2016), pp. 619–623. https://doi.org/10.1109/icfhr.2016.110
38. Pratikakis, I., Zagoris, K., Kaddas, P., Gatos, B.: ICFHR 2018 competition on handwritten document image binarization (H-DIBCO 2018). In: 16th International Conference on Frontiers in Handwriting Recognition (ICFHR 2018), pp. 489–493. https://doi.org/10.1109/icfhr-2018.2018.00091