Document Image Binarization Using Dual Discriminator Generative Adversarial Networks

Report submitted in partial fulfilment of the requirement for the award of the degree of

Bachelor of Computer Science and Engineering
In the Faculty of Engineering and Technology

Jadavpur University

By

Rajonya De, Roll No. 001610501031

Anuran Chakraborty, Roll No. 001610501020

Under the guidance of

Dr. Ram Sarkar

Department of Computer Science and Engineering

Jadavpur University
Kolkata – 700 032
2020

ACKNOWLEDGEMENT

We are indebted to our project guide, **Dr. Ram Sarkar** for his constant support, prompt help, support and unwavering encouragement which helped us a lot in completing the project. Under his guidance, we have been able to learn many new concepts regarding the topic at hand.

We are also thankful to **Prof. Mahantapas Kundu**, Head of the Department of Computer Science and Engineering, Jadavpur University for allowing us to carry out research in the department.

Rajonya De Anuran Chakraborty

ABSTRACT

For document image analysis, image binarization is an important preprocessing step. Also, binarization can help in improving the readability of old and historical manuscripts. Such documents are generally degraded due to various reasons such as bleed-through, faded ink, or stains. Achieving good binarization performance on these documents is a challenging task. In this work, a deep learning based model for document image binarization has been proposed, comprising a Dual Discriminator Generative Adversarial Network (DD-GAN) which uses Focal Loss as generator loss. The DD-GAN consists of two discriminator networks - one looks for the global similarity i.e. on the whole image, and another one explores the image in small patches i.e. local similarity. At the final stage, simple thresholding is performed on the generated images. The method has been tested on five recent DIBCO datasets. It has been found that the method is robust and it provides results comparable with state-of-the-art methods.

Keywords: *Binarization, Deep Learning, Generative Adversarial Network, Document Image, Focal Loss.*

Dr. Ram Sarkar
Project Guide

I. INTRODUCTION

Document image binarization, the task of classifying a pixel as foreground or background, is an important preprocessing step in document image analysis. The documents, especially historical manuscripts, often suffer from degradation due to aging effect, ink stains, bleed-through, faded ink, etc. which make the binarization process a challenging task. Early binarization techniques [1][2][3][4] used various thresholding techniques either based on single or multiple thresholds to classify the pixels. Recently, deep learning frameworks [5][6] have also become popular for binarization of document images. The success of these frameworks has been broadly attributed to their ability to effectively capture the spatial dependency among the pixels in an image. A fully convolutional neural network (FCNN) [7] is ideal for the semantic labeling of pixels. In [8], document image binarization is formulated as an energy minimization problem, whereby an FCNN is trained for semantic segmentation of pixels that provides a labeling cost associated with each pixel. Other frameworks such as the holistically-nested edge detector (HED) [9] and U-Net [10] also provide significant performance gain. In [11], the authors use a recurrent neural network (RNN) based model for binarization.

II. PROPOSED MODEL

A deep learning based method for image binarization has been proposed in this work. The method uses a Dual Discriminator Generative Adversarial Network (DD-GAN) which consists of a U-Net network as the generator and two independent discriminators – a Global Discriminator which looks at the higher-level features and the overall document, and another Local Discriminator which analyses the low-level features in the local patches of the document. The architecture of DD-GAN is described hereafter.

A. Architecture

Traditionally, GANs consists of two networks, a generator that is trained to generate target images in the output domain, and a discriminator trained to distinguish between real (in the output domain) and fake (generated) images. These two networks are trained together and the generator tries to generate images plausible enough to fool the discriminator.

The proposed GAN consists of 3 networks, a **Generator** which aims at generating the binarized image, a **Global Discriminator** which classifies whether an image is real or fake by taking as input the whole image, and a **Local Discriminator** which considers local patches of the image and classifies the patches as real or fake. The architecture of the overall model is given in Fig 1. The main intuition behind the use of two discriminators is that the global discriminator classifies an image by looking at the overall image and higher-level features of the image (such as image background, texture) while the local discriminator looks at lower level features (such as text strokes) while classifying the images. Thus, we eliminate the pooling layers in the local network as it is difficult to obtain such precise localized details when the images are scaled down [6]. Moreover, to capture these low-level intricacies, the network is empirically observed to require more layers. For global discriminator, it is observed that computing features at multiple scales lead to improved performance. Hence, pooling is an essential component of this network. Moreover, document images characteristically possess limited variations for texture or background unlike in other computer vision domains. Thus, it is observed that a lesser number of layers serve our purpose for this task. For this reason, in the proposed method, in global discriminator, less number of layers is used compared to that of the local discriminator.

Generator The architecture of the generator network follows that of a standard U-Net [10] used in pix2pix GAN [12]. Since binarization can be thought of as a classification problem where a pixel is classified as a background or a foreground, the task of the generator is to take the input image and classify each of the pixels present therein as mentioned.
Generator Loss Function The loss function used for the generator is the focal loss [13]. Focal loss has been chosen as it counters the class imbalance problem, which is relevant here due to the presence of a lot more background pixels than foreground pixels, thus preventing any bias while training the network.

Global Discriminator The global discriminator consists of two convolutional layers with batch normalization and LeakyReLU as the activation function, followed by a max pooling layer and 3 fully connected layers. The architecture is shown in Fig 2.
Local Discriminator The local discriminator consists of 4 convolutional layers with batch normalization and LeakyReLU as activation function, followed by 4 fully connected layers. The architecture is shown in Fig 3. It is to be noted that pooling layers are absent in the local discriminator to minimize any loss of spatial information due to pooling, as the local discriminator aims to capture more intricate details.
Discriminator Loss Function Both the discriminators use the Binary Cross Entropy (BCE) loss.
Total GAN Loss The total GAN loss is calculated as shown in equation (1).

$$L_{total} = \mu\left(L_{global} + \sigma L_{local}\right) + \boldsymbol{\lambda} L_{gen} \qquad \ldots(1)$$

Where L_{total} is the total loss, L_{global} is the global discriminator loss, L_{local} is the local discriminator loss (averaged over all patches of an image), and L_{gen} is the generator loss. μ, σ, and λ are parameters. For our experimentation μ=0.5, σ=5, and λ=75. The value

of σ is taken to be greater than 1 so that the local discriminator contributes more to the loss than the global. λ is kept to a high value for the generator to train well.

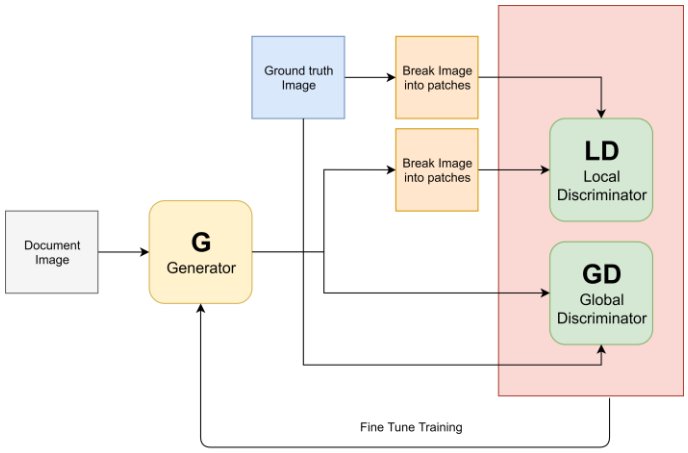


Fig. 1. The architecture of the overall model used for binarization.

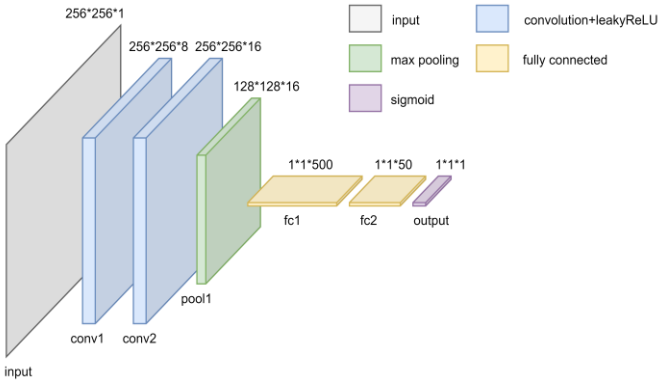


Fig. 2. The architecture of the global discriminator.

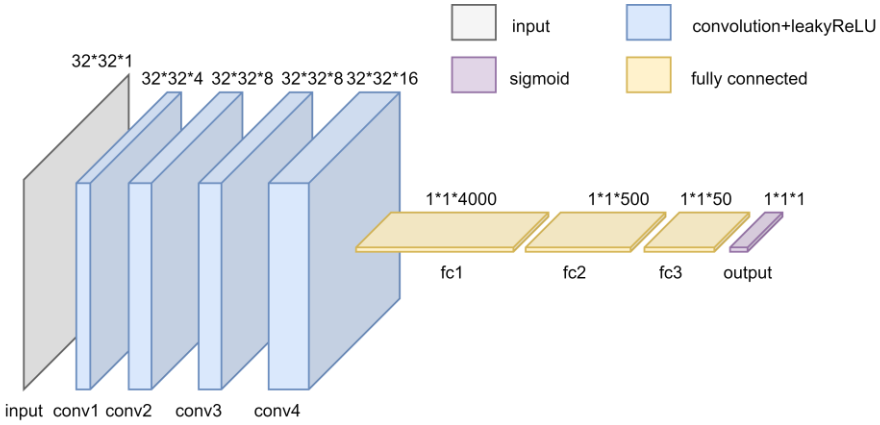


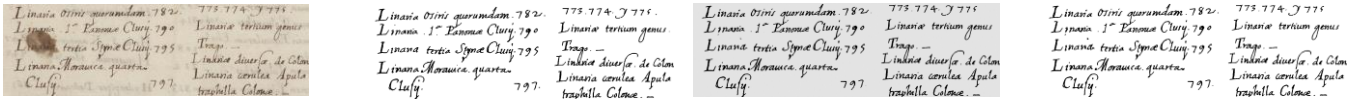
Fig. 3. The architecture of the local discriminator.

B. Training

The GAN is trained by feeding the document images (converted to grayscale) and the corresponding ground truth images to the generator. The discriminator classifies whether an image is real or fake (i.e. generated by the generator). To the global discriminator, a whole image (generated or ground truth) is fed as input. Simultaneously, the corresponding image is split into non-overlapping patches of size 32×32 pixels, and each patch is separately fed into the local discriminator for classification. The two discriminators are thereby trained independently in parallel. Further, the generator loss also considers each discriminator loss separately for its input images, during training (see eq. 1).

C. Testing

Document images from the test set are first converted to grayscale and then fed as an input to the trained GAN which generates images hereafter referred to as the generated image. The GAN performs most of the tasks of cleaning such as removing bleed through, lightening the stain marks, darkening text strokes that were faint, such that the generated image presents more or less a bimodal distribution. Hence, applying a global threshold of 127 on the generated image provides the binarized image with reasonably good accuracy. Fig. 4 shows an original image, generated image, and the image obtained by applying the threshold on a DIBCO 2018 image.



(a) (b) (c) (d)

Fig. 4. A sample image from DIBCO 2014 dataset showing (a) the original image (b) the ground truth image (c) the image generated by DD-GAN (d) image obtained after applying a global threshold of 127 on the generated image

III. EXPERIMENT

In this section, we present the experimental results of our proposed model for document image binarization.

A. Datasets

There are several standard datasets for document image binarization, such as (H-)DIBCO datasets which are made publicly available through competitions organized by the ICDAR committee. We have used the DIBCO 2013 [14], DIBCO 2017 [15] and H-DIBCO 2014 [16], H-DIBCO 2016 [17], H-DIBCO 2018 [18] datasets for our experiment. Since each image is considerably large to be fed as input to a neural network, each image has been broken into overlapping patches of size 256×256 with a stride of 128 in both horizontal and vertical directions. However, during evaluation, the entire images have been considered for each of these datasets. When testing the model on a particular dataset, all the remaining datasets have been aggregated to form the training set. For example, to evaluate the model on DIBCO 2017 dataset, all the images of DIBCO 2013, H-DIBCO 2014, H-DIBCO 2016, and H-DIBCO 2018 have been collectively used for training while the images in DIBCO 2017 have been used to finally evaluate the performance of the trained model. The same procedure is followed for other datasets.

B. Results

In this section, we discuss the performance of our proposed model. Four metrics namely F-measure (FM), pseudo F-measure (P-FM), distance reciprocal distortion metric (DRD) [19], peak signal-to-noise ratio (PSNR) (formula given in Eqn. (2) are used to evaluate performance.

$$PSNR = 10log_{10}\left(\frac{255^2}{MSE}\right) \quad ..(2)$$

where MSE is the Mean Squared Error between pixels.

Additionally, we compare the performance of our model with other state-of-the-art methods in Tables I-V. The results of the top-ranked methods in the respective competitions are also specified (as Comp #rank).

TABLE I
RESULTS ON DIBCO 2013 DATASET

Method	**FM**	**P-FM**	**PSNR**	**DRD**
Comp #1	92.12 (8)	94.19 (8)	20.68 (9)	3.10(8)
Comp #2	92.70 (7)	93.19 (9)	21.29 (4)	3.18(9)
Comp #3	91.81 (9)	92.67(10)	20.68(10)	4.02(11)
GiB [20]	91.14(11)	94.75 (7)	19.58(11)	2.77(7)
TM [6]	93.17 (6)	**96.81 (1)**	20.71 (8)	2.21(5)
HDSN [5]	94.40 (3)	96.00 (5)	21.40 (3)	1.80(3)
Jia [21]	93.42 (5)	96.05 (4)	20.78 (6)	2.03(4)
Howe[22]	91.34(10)	91.79(11)	21.29 (5)	3.18(10)
Pix2Pix GAN[12]	93.63 (4)	95.18 (6)	20.78 (7)	2.62(6)
Zhao et.al.[23]	**95.28 (1)**	96.47 (3)	22.23 (2)	**1.39 (1)**
DD-GAN	*95.10 (2)*	*96.52 (2)*	***22.43 (1)***	*1.75 (2)*

TABLE II
RESULTS ON H-DIBCO 2014 DATASET

Method	**FM**	**P-FM**	**PSNR**	**DRD**
Comp #1	**96.88(1)**	97.65(2)	22.66(2)	0.90(2)
Comp #2	96.63(3)	97.46(5)	22.40(4)	1.00(4)
Comp #3	93.35(10)	96.05(10)	19.45(11)	2.19(10)

GiB [20]	94.00(9)	96.48(8)	19.93(10)	0.90(3)
TM [6]	91.96(11)	94.78(11)	20.76(8)	2.72(11)
HDSN [5]	96.66(2)	97.59(3)	**23.23(1)**	**0.79(1)**
Jia [21]	94.98(8)	97.18(7)	20.56(9)	1.50(8)
Howe[22]	96.49(4)	97.38(6)	22.24(5)	1.08(6)
Pix2Pix GAN[12]	95.15(7)	96.19(9)	21.55(7)	1.61(9)
Zhao et.al.[23]	96.41(5)	97.55(4)	22.12(6)	1.07(5)
DD-GAN	*96.27(6)*	***97.66(1)***	*22.60(3)*	*1.27(7)*

TABLE III
RESULTS ON H-DIBCO 2016 DATASET

Method	**FM**	**P-FM**	**PSNR**	**DRD**
Comp #1	87.61(10)	91.28(11)	18.11(10)	5.21(10)
Comp #2	88.72(8)	91.84(9)	18.45(8)	3.88(7)
Comp #3	88.47(9)	91.71(10)	18.29(9)	3.93(8)
GiB [20]	**91.15(2)**	93.03(6)	19.18(3)	**3.20(2)**
TM [6]	89.52(7)	93.76(3)	18.67(7)	3.76(5)
HDSN [5]	90.10(4)	93.57(4)	19.01(4)	3.58(3)
Jia [21]	90.48(3)	93.27(5)	**19.30(2)**	3.97(9)
Howe [22]	87.47(11)	92.28(7)	18.05(11)	5.35(11)
Pix2Pix GAN [12]	89.73(6)	92.09(8)	18.95(5)	3.76(5)
Zhao et.al.[23]	*91.66(1)*	*94.58(2)*	*19.64(1)*	*2.82(1)*
DD-GAN	*89.98 (2)*	***95.23 (1)***	*18.83 (6)*	*3.61 (4)*

TABLE IV
RESULTS ON DIBCO 2017 DATASET

Method	**FM**	**P-FM**	**PSNR**	**DRD**
Comp #1	**91.04(1)**	**92.86(1)**	18.28(2)	**3.40(1)**
Comp #2	89.67(5)	91.03(6)	17.58(7)	4.35(7)
Comp #3	89.42(7)	91.52(5)	17.61(5)	3.56(4)
GiB [20]	89.75(4)	90.85(7)	17.60(6)	4.19(6)
Jia [21]	85.34(8)	86.06(8)	16.25(8)	8.18(8)
Pix2Pix GAN[12]	89.53(6)	91.73(4)	17.91(3)	3.55(3)
Zhao et.al.[23]	*90.73(3)*	*92.58(3)*	*17.83(4)*	*3.58(5)*
DD-GAN	*90.98 (2)*	*92.65 (2)*	***18.34 (1)***	*3.44 (2)*

TABLE V
RESULTS ON H-DIBCO 2018 DATASET

Method	**FM**	**P-FM**	**PSNR**	**DRD**
Comp #1	**88.34(1)**	**90.24(2)**	**19.11(1)**	**4.92(2)**
Comp #2	73.45(4)	75.94(4)	14.62(4)	26.24(7)
Comp #3	70.01(5)	74.68(5)	13.58(5)	17.45(5)
Comp #4	64.52(6)	68.29(6)	13.57(6)	16.67(4)
Comp #5	46.35(7)	51.39(7)	11.79(7)	24.56(6)
Zhao et.al.[23]	*87.73(2)*	*90.60(1)*	*18.37(2)*	*4.58(1)*
DD-GAN	*75.53 (3)*	*78.05 (3)*	*14.64 (3)*	*14.61 (3)*

The best results obtained for a particular metric over all the methods are marked in bold. The rank of the score obtained by DD-GAN and the other methods for each metric is also specified in parenthesis. As can be seen from the tables, the performance of the proposed model is comparable to state-of-the-art methods while also surpassing the many methods considered here for comparison. The good performance of the model on most of the datasets serves as a proof of robustness. However, as we can see from Table V the method does not perform well on the H-DIBCO 2018 dataset. An original image from H-DIBCO 2018 dataset and its binarized output image is shown in Fig. 5. The poor performance can be mainly attributed to the fact that for the 2018 dataset, many of the images contain background pixels at the edges which are not part of a manuscript. However, the other datasets do not contain such images, and therefore, the GAN fails to learn to distinguish between those pixels and classifies them as foreground despite being able to clean the central manuscript portion of the image.

(a) (b)

Fig. 5. A sample image of H-DIBCO 2018 dataset showing (a) the original image (b) the output binarized image

TABLE VI
ABLATION STUDY ON DIBCO 2013 H-DIBCO 2018 DATASET

Dataset	*Method*	*FM*	*P-FM*	*PSNR*	*DRD*
2013	*DDGAN*	***95.10***	***96.52***	***22.43***	***1.75***
	Global Only	*93.79*	*94.57*	*21.55*	*2.21*
	Local Only	*94.56*	*95.82*	*22.21*	*1.97*
2018	*DDGAN*	***75.53***	***78.05***	***14.64***	***14.61***
	Global Only	*69.55*	*71.33*	*13.32*	*22.52*
	Local Only	*72.56*	*75.48*	*14.23*	*17.54*

From Table VI, it can be seen that the proposed method with two discriminators (one local and one global) performs better than each of the discriminators separately. This can mainly be attributed to the fact that with two discriminators two levels of features are captured which helps in better training of the generator.
Further, our model was partly inspired by Pix2Pix GAN [12], which too uses a single discriminator, and has an optimized architecture for common image-to-image translation tasks. This network too has also shown consistent poor performance compared to our proposed model, as can be seen in Tables I-V.

IV. CONCLUSION

In this work, a deep learning based model, named as DD-GAN, has been developed for document image binarization. As both global and local thresholding based approaches have their own pros and cons, hence researchers prefer a hybrid approach for the said purpose. Going by the trend, the proposed model has utilized both the global and local information about the pixel distribution. In doing so, DD-GAN uses two discriminator networks to capture both global and local information. Besides, the focal loss has been used to handle the class imbalance of the pixels as document images generally contain more background pixels then foreground pixels. The model has been evaluated on five recent datasets of DIBCO series and it has been found that DD-GAN provides better or analogous results when compared with some state-of-the-art methods. In the future, we would like to apply this model to some other datasets. Besides, some other loss functions can be explored, and the value of the λ (in equation 1) can be made dynamic.

REFERENCES

[1]N. Otsu, "THRESHOLD SELECTION METHOD FROM GRAY-LEVEL HISTOGRAMS.," *IEEE Trans Syst Man Cybern*, vol. SMC-9, no. 1, pp. 62–66, 1979, doi: 10.1109/tsmc.1979.4310076.
[2]J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognit.*, 2000, doi: 10.1016/S0031-3203(99)00055-2.
[3]N. Phansalkar, S. More, A. Sabale, and M. Joshi, "Adaptive local thresholding for detection of nuclei in diversity stained cytology images," in *ICCSP 2011 - 2011 International Conference on Communications and Signal Processing*, 2011, doi: 10.1109/ICCSP.2011.5739305.
[4]B. Gatos, I. Pratikakis, and S. J. Perantonis, "Improved document image binarization by using a combination of multiple binarization techniques and adapted edge information," in *Proceedings - International Conference on Pattern Recognition*, 2008, doi: 10.1109/icpr.2008.4761534.
[5]Q. N. Vo, S. H. Kim, H. J. Yang, and G. Lee, "Binarization of degraded document images based on hierarchical deep supervised network," *Pattern Recognit.*, 2018, doi: 10.1016/j.patcog.2017.08.025.
[6]C. Tensmeyer and T. Martinez, "Document Image Binarization with Fully Convolutional Neural Networks," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2017, doi: 10.1109/ICDAR.2017.25.
[7]J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, doi: 10.1109/CVPR.2015.7298965.
[8]K. R. Ayyalasomayajula, F. Malmberg, and A. Brun, "PDNet: Semantic segmentation integrated with a primal-dual network for document binarization," *Pattern Recognit. Lett.*, 2019, doi: 10.1016/j.patrec.2018.05.011.
[9]S. Xie and Z. Tu, "Holistically-Nested Edge Detection," *Int. J. Comput. Vis.*, 2017, doi: 10.1007/s11263-017-1004-z.
[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015, doi: 10.1007/978-3-319-24574-4_28.
[11] F. Westphal, N. Lavesson, and H. Grahn, "Document image binarization using recurrent neural networks," in *Proceedings - 13th IAPR International Workshop on Document Analysis Systems, DAS 2018*, 2018, doi: 10.1109/DAS.2018.71.
[12] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, doi: 10.1109/CVPR.2017.632.
[13] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, doi: 10.1109/ICCV.2017.324.

[14] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "ICDAR 2013 document image binarization contest (DIBCO 2013)," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2013, doi: 10.1109/ICDAR.2013.219.
[15] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos, "ICDAR2017 Competition on Document Image Binarization (DIBCO 2017)," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2017, doi: 10.1109/ICDAR.2017.228.
[16] K. Ntirogiannis, B. Gatos, and I. Pratikakis, "ICFHR2014 Competition on Handwritten Document Image Binarization (H-DIBCO 2014)," in *Proceedings of International Conference on Frontiers in Handwriting Recognition, ICFHR*, 2014, doi: 10.1109/ICFHR.2014.141.
[17] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos, "ICFHR 2016 handwritten document image binarization contest (H-DIBCO 2016)," in *Proceedings of International Conference on Frontiers in Handwriting Recognition, ICFHR*, 2016, doi: 10.1109/ICFHR.2016.0118.
[18] I. Pratikakis, K. Zagori, P. Kaddas, and B. Gatos, "ICFHR 2018 competition on handwritten document image binarization (H-DIBCO 2018)," in *Proceedings of International Conference on Frontiers in Handwriting Recognition, ICFHR*, 2018, doi: 10.1109/ICFHR-2018.2018.00091.
[19] H. Lu, A. C. Kot, and Y. Q. Shi, "Distance-reciprocal distortion measure for binary document images," *IEEE Signal Process. Lett.*, 2004, doi: 10.1109/LSP.2003.821748.
[20] S. Bhowmik, R. Sarkar, B. Das, and D. Doermann, "GiB: A game theory inspired binarization technique for degraded document images," *IEEE Trans. Image Process.*, 2019, doi: 10.1109/TIP.2018.2878959.
[21] F. Jia, C. Shi, K. He, C. Wang, and B. Xiao, "Degraded document image binarization using structural symmetry of strokes," *Pattern Recognit.*, 2018, doi: 10.1016/j.patcog.2017.09.032.
[22] N. R. Howe, "Document binarization with automatic parameter tuning," *Int. J. Doc. Anal. Recognit.*, 2013, doi: 10.1007/s10032-012-0192-x.
[23] J. Zhao, C. Shi, F. Jia, Y. Wang, and B. Xiao, "Document image binarization with cascaded generators of conditional generative adversarial networks," *Pattern Recognit.*, 2019, doi: 10.1016/j.patcog.2019.106968.