Historical Document Image Binarization using a lightweight U-Net derivative architecture

Nikolaos Detsikas, Nikolaos Mitianoudis

Electrical & Computer Engineering Dep.
Democritus University of Thrace
Xanthi, Greece
{ndetsika, nmitiano}@ee.duth.gr

***Abstract*—Historical document image binarization has been a very important document image processing task. The task of binarization can be viewed as a pre-processing step that attempts to separate the printed/handwritten characters in the image from noise and background, assisting in the Optical Character Recognition (OCR) process. In this article, we propose a U-Net style deep learning architecture that incorporates many other developments of deep learning, including residual connections, multi-resolution connections, visual attention blocks and dilated convolution blocks. These concepts in the proposed DMVAnet have shown to improve performance in our binarization experiments. Finally, the proposed architecture is a lightweight network that performs very close or even better than state-of-the-art approaches with a fraction of the network size and parameters used by other approaches.**

***Index Terms*—image binarization, U-Net, Visual Attention, Residual Networks**

I. Introduction

Written language appears everywhere in our urban environments the ability to extract text from digital visual media is critical in many applications, such as extraction of the text from images and videos, digitisation of the written cultural heritage, Optical Character Recognition and others. Document Image Binarization is the process of separating the text from its environment in a document image. The input image is segmented into two layers of information, one for the background and one for the text, becoming essentially a binary map.

One of the most well-known approaches is Otsu's method [1], which performs automatic global image thresholding in its basic form. Sauvola [2] and Niblack [3] binarization algorithms operate by calculating local thresholds for every pixel, based on statistical information from the pixel neighbourhood. Apart from these basic techniques, more methods have been proposed in the past decade. Howe [4] binarizes the image by minimising a global energy function, based on a Markov Random Field model and performing automatic parameter tuning. Su et al. [5] construct an adaptive contrast map and based on that and the Canny edge detection map, the method detects the text stroke edges, which are used to estimate local threshold values for binarizing the image. Lelore and Bouchara [6] introduce the FAIR algorithm, which applies a modified Canny edge detection and clusters the resulting pixels. To tackle the defects of parameter selection, the process is applied twice with different parameters and the final results are merged. Nachi et al. [7] use phase congruency feature maps, based on Kovesi's phase congruency model, as well as a phase-derived denoised image in order to produce a final binarized version of the input. Mitianoudis and Papamarkos [8] address the Document Image Binarization problem by first removing the background with a long-window low-pass filtering process. The resulting image is binarized using Local Co-occurrence Mapping (LCM), that exploits common local character properties, when identifying the character pixels and Mixture of Gaussians (MoG) clustering. As a last step, a mathematical morphology step removes misclassified or noisy items. In addition, Jia at al. [9] perform a background removal process, compute a gradient map from the output and extract the structural symmetric pixels (SSPs) to calculate local thresholds for the binarization process. Finally, Bhowmik et al. [10] employ Game Theory concepts, such as two-player non-zero-sum non-cooperative game and the Nash equilibrium, in order to extract image features that are further fed to a K-means clustering step for classifying the pixels into foreground and background groups.

We focus on recent deep learning methods, evaluate their performance and suggest an innovative architecture for addressing the problem more efficiently. In [11], He and Schomaker suggest an iterative deep learning framework for improving the input images by removing noise and degradations that usually prevent efficient binarization. The framework "learns" the noise and degradations of the original image and iteratively produces a uniform variant that can be finally binarized with any binarization method. Therefore, the proposed framework acts as a deep learning augmentation pre-processing step for any binarization process. Vo et al. [12] use a multiscale hierarchical approach consisting of three Deep Supervised Networks (DSN) in order to separate text from the background noise. By using different feature scales, the model tries to optimize the classification of image pixels over large areas as well as those over the text boundaries. Zhao et al. [13] employ conditional Generative Adversarial Networks (cGANs) in order to synthesize the binarized output images from degraded document images. A two-stage generator is employed for producing binary maps, based on the learned input and ground truth images. He and Schomaker [14]

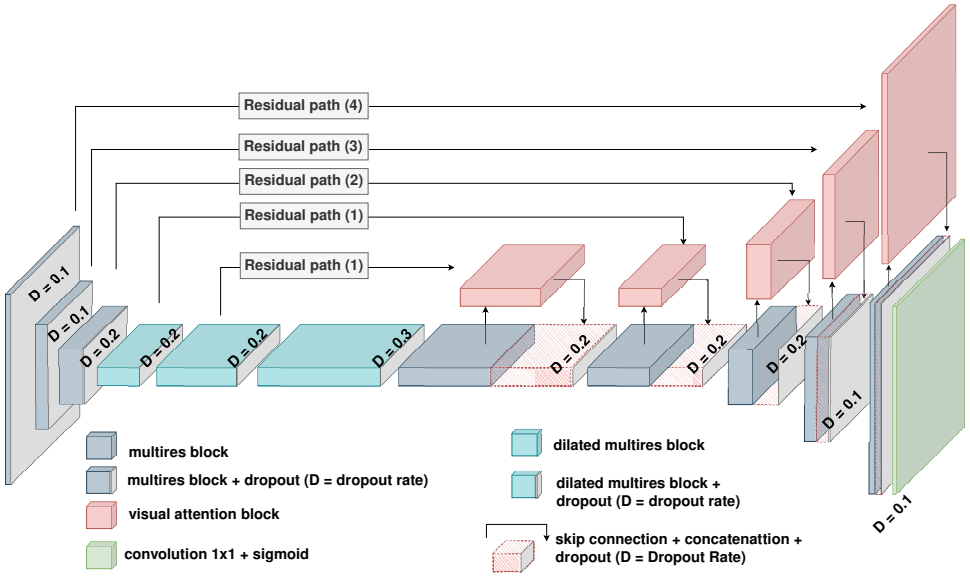

Fig. 1. The proposed Dilated MultiRes Visual Attention U-Net (DMVAnet)

describe the CT-Net, a novel T-net architecture that consists of one encoder and two decoders, the first of which performs an image enhancement task and the other a binarization task. The T-net blocks are cascaded resulting in a CT-net model, where the block connections placed along the enhancement outputs, so that each T-net of the pipeline receives a more and more enhanced input.

In this paper, we present a simple, lightweight and effective deep learning architecture that solves the binarization problem without using pre- and post- processing or ensemble of different networks. The proposed Dilated MultiRes Visual Attention U-Net (DMVAnet) architecture is a composite U-Net network that exhibits state-of-the-art performance in a single-step approach with comparatively low complexity. The proposed DMVAnet contains carefully selected features from previous networks to form a new architecture that has not been applied or tested in the context of image binarization before to the best of our knowledge. A fundamental difference of our proposal is that DMVAnet is a single network, trained only once, while related SOTA methods, either consist of multiple separate Deep Neural Networks, different pre- or post- processing steps, or have to be applied iteratively. Due to the above, they feature complex implementation steps or serve as an enhancement step that should be followed by other binarization methods. Finally, most SOTA networks feature far more complex networks with up to 32 times the parameters of DMVAnet, yielding no or minimal performance gain compared to their added complexity.

The paper is organised, as follows. In Section 2, we describe the proposed architecture with all the main structural details. Section 3 compares the proposed approach with other state-of-the-art methods on commonly-used datasets. Finally, Section 4 concludes the article and proposes steps for future work.

II. The Proposed Dilated MultiRes Visual Attention U-Net (DMVAnet)

Fig. 1 shows the complete DMVAnet architecture. The DMVAnet stems from a conventional U-Net and effectively incorporates modern deep learning architectural traits. All the selected blocks and modifications have been validated through experiments. Apart from the selected changes, other architectures (U-Net with dense connections [15], UNet++ [16] and DeeplabV3+ [17]) have been tested and led to no performance improvement.

The first novel element is that it uses residual connections along the paths of the encoder and decoder in the style of [18], in order to combat the vanishing gradient problem.

Next, the encoder uses skip connections that are scaled with visual attention blocks, with *Channel attention* and *Spatial attention* modules ([19], [20]). The details of the block are shown in Fig. II.

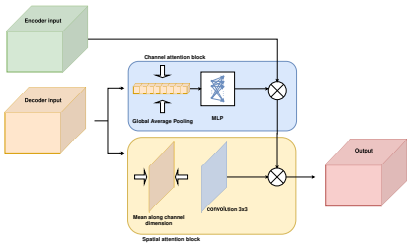

Fig. 2. The proposed Visual attention block.

In [21], Ibtehaz and Rahman proposed the replacement of *Inception blocks* with *MultiRes blocks* in order to minimise the additional memory overhead without sacrificing performance. The *MultiRes block* replaces the larger convolutions with a sequence of 3×3 convolutional layers. The block is shown in Fig. II.

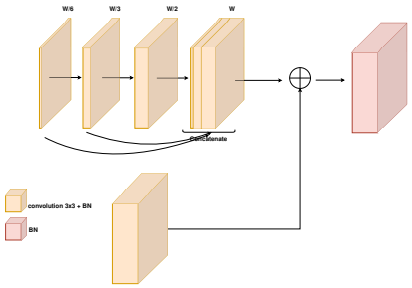

Fig. 3. The MultiRes block used in the U-Net concept.

From [21], the *Residual paths (Res paths)* are also incorporated here. Fig. II demonstrates that the *Res paths* balance the incompatibilities in the semantic information, carried by the concatenated encoder and decoder features maps.

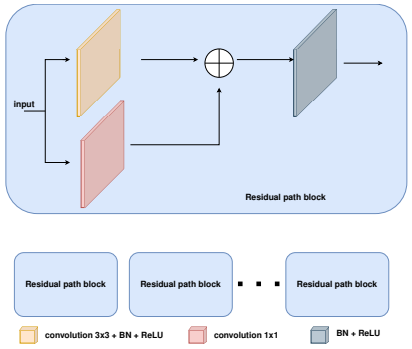

Fig. 4. The Residual path used in the MultiRes Visual Attention U-Net.

Inspired by the DeepLab network variants [17], [22], [23], we also add dilated (a-trous) convolutions to our network layers. Dilated convolutions remedy the spatial resolution loss after a series of consecutive pooling layers, by retaining the feature map dimension. The complete DMVAnet architecture is depicted in Fig. 1.

III. Performance Comparison

In this section, we evaluate the proposed DMVAnet model against state-of-the-art binarization approaches. The competitor architectures in our comparison contained both traditional image processing based methods and deep learning based methods. More specifically, the following methods were examined: Otsu [1], Sauvola [2], Su et al [5], Howe [4], Lelore et al. [6], Nafchi et al. [7], Mitianoudis et al. [8], Jia et al. [9], GiB [10], cGANs [13], DeepOtsu [11], DSN [12], PDNet [24], CT-Net-3 [14] and CTada -Net-3 [14].

In all experiments, the network training input were randomly cropped 256×256 patches of the dataset training images.

The following augmentations were added to the training samples in a cascaded manner:

- Random scale augmentation
- Random horizontal and vertical flip
- Rotation by random multiples of 90 degrees
- Random contrast, brightness, hue and saturation change

DMVAnet was trained for 150 epochs using the Adam optimizer with a polynomial decay schedule (initial rate is 0.001, power is 0.9, decay steps are the total training steps). Training and evaluation was done on an Ubuntu 20.04 PC with 64GB RAM, an Intel i9 2.5 GHz 16-Core CPU and an NVIDIA GeForce RTX 3090 GPU. The architecture was developed in Python v3.8.10 and Tensorflow v2.10.0. The developed code is available via the following url [1].

The lost function was *Dice loss* [25] which outperformed Binary Cross-Entropy, Mean Square Error, Inverse Peak Signal-to-Noise Ratio, a Differentiable F-Measure version and linear combinations. Training batch size was set to 32 (performance gains and hardware limitations trade-off). Finally, random shuffling and "save best only" training strategy were employed.

TABLE I
COMPARISON ON INDICATIVE DATASETS (DIBCO 2009, H-DIBCO 2014, DIBCO 2017 AND H-DIBCO 2018), FOLLOWING THE GUIDELINES OF THE CT-NET EXPERIMENT. DASHES INDICATE MISSING METRICS. (SOURCE [14])

DIBCO 2009

Method	FM	F_{ps}	PSNR	DRD
DeepOtsu [11]	-	-	-	-
DSN [12]	-	-	-	-
cGANs [13]	94.1	95.26	20.30	1.82
CT-Net-3 [14]	92.08	94.31	19.77	3.58
CTada -Net-3 [14]	94.18	95.80	20.50	2.56
DMVAnet	**95.7**	**96.84**	**21.42**	**1.35**
Rank	1	1	1	1

H-DIBCO 2014

Method	FM	F_{ps}	PSNR	DRD
DeepOtsu [11]	95.9	97.2	22.1	0.9
DSN [12]	96.66	97.59	23.23	0.79
cGANs [13]	96.41	97.55	22.12	1.07
CT-Net-3 [14]	**97.70**	**98.74**	**23.92**	**0.65**
CTada -Net-3 [14]	96.91	97.93	22.62	0.88
DMVAnet	97.55	98.58	23.62	0.71
Rank	2	2	2	2

DIBCO 2017

Method	FM	F_{ps}	PSNR	DRD
DeepOtsu [11]	-	-	-	-
DSN [12]	-	-	-	-
cGANs [13]	90.73	92.58	17.83	3.58
CT-Net-3 [14]	**92.72**	94.31	**19.17**	2.79
CTada -Net-3 [14]	92.65	94.73	**19.17**	2.65
DMVAnet	92.2	**95.14**	18.73	**2.6**
Rank	3	1	3	1

H-DIBCO 2018

Method	FM	F_{ps}	PSNR	DRD
DeepOtsu [11]	-	-	-	-
DSN [12]	-	-	-	-
cGANs [13]	87.73	90.60	18.37	4.58
CT-Net-3 [14]	88.90	91.45	18.84	5.58
CTada -Net-3 [14]	**92.23**	**94.97**	**20.13**	**2.70**
DMVAnet	85.9	89.45	18.16	6.99
Rank	4	4	4	4

[1] https://github.com/detsikas/DMVAnet

TABLE II
OVERALL RANKING (DASHES SHOW THAT THE METHOD IS NOT RANKED BECAUSE METRICS ARE MISSING).

Method	2009	2011	2014	2016	2017	2018	Overall
Otsu [1]	11	13	11	9	8	8	7
Sauvola [2]	10	12	12	12	7	7	8
Su et al [5]	7	11	10	11	-	-	-
Howe [4]	4	9	5	8	5	5	5
Lelore et al. [6]	5	7	8	10	-	-	-
Nafchi et al. [7]	-	-	-	-	-	-	-
Mitianoudis et al. [8]	9	10	-	-	-	-	-
Jia et al. [9]	6	8	9	5	6	6	6
GiB [10]	-	-	-	-	-	-	-
DeepOtsu [11]	-	5	7	2	-	-	-
DSN [12]	-	6	3	6	-	-	-
PDNet [24]	-	-	-	-	-	-	-
cGANs [13]	2	4	6	**1**	4	3	3
CT-Net-3 [14]	8	**1**	**1**	7	**1**	2	4
CTada -Net-3 [14]	3	3	4	4	2	**1**	2
DMVAnet	**1**	2	2	3	3	4	**1**

A. Comparisons & results

In order to compare against the other approaches, we replicate an experiment described in more detail in [14]. In each of the experiments described in [14], a DIBCO dataset is used as the evaluation dataset and the remaining are treated as training datasets. In total, the DIBCO datasets used both for training and evaluation purposes are DIBCO 2009, H-DIBCO 2010, DIBCO 2011, H-DIBCO 2012, DIBCO 2013 [26], H-DIBCO 2014, H-DIBCO 2016, DIBCO 2017 [27] and H-DIBCO 2018 [28].

The training sets also include the Bickley-diary dataset, the Persian Heritage Image Binarization dataset (PHIDB) [29] and the Synchromedia Multispectral dataset [30].

Our method is the top ranking method for DIBCO 2009. For DIBCO 2011 and H-DIBCO 2014, it is outperformed only by CT-Net-3, which is of much higher complexity. The proposed DMVAnet consists of 6.5M parameters, whereas the CT-Net-3 requires 45M parameters. In other words, the CT-Net-3 requires seven times more parameters than our proposed DMVAnet. For DIBCO 2017, our method exhibits the best pseudo f-measure and DRD values, while it is outperformed only by CT-Net-3 method variations. Finally, H-DIBCO 2018 dataset experiment renders our method fourth among the examined methods, all of which though have much higher complexity. Again, H-DIBCO 2018 presents special challenges, such as the strong bleed-through, strong paper stains and page margins not seen in other datasets. Detailed results for indicative datasets are listed in Table I.

To calculate the ranking, we rank all methods for each metric and each evaluation dataset. We add the metric ranks and calculate the total rank for each evaluation dataset by sorting the sums from lowest to highest. We performed the DIBCO 2009, DIBCO 2011, H-DIBCO 2014, H-DIBCO 2016, DIBCO 2017 and H-DIBCO 2018 experiments. Table II shows the overall ranking against the competitor methods over all evaluation datasets. Despite, the fluctuation among datasets, our method ranks first against all other experiment methods, which implies that in general the proposed method outperforms more complex networks and offers a reliable lightweight architecture for the problem of document image binarization.

IV. CONCLUSIONS AND FUTURE WORK

We have presented a single-step one-shot light-weight deep learning network for Document Image Binarization that requires neither pre- nor post- processing steps. The proposed DMVAnet combines a basic U-Net architecture with elements from modern deep learning architectures, including visual attention blocks, multi resolution blocks, residual connections and dilated convolutions that enhances its performance without inhibiting computational efficiency. The DMVAnet's performance was benchmarked with State of the Art methods on the popular (H-)DIBCO datasets and demonstrated that it exhibits better or comparable performance but with a much smaller training parameters complexity.

The DMVAnet can easily be scaled up or down in order to accommodate for different image input sizes. A change in the number of encoder/decoder layers, proportionally changes the residual path lengths. MultiRes blocks can be further extended or compressed for more fine or coarse multi-resolution aggregation. Scale changes may introduce performance overheads which can be fine-tuned by altering feature map dimensionality.

Further, continuing on the path of visual attention research, we will investigate more complex deep learning attention architectures, such as the transformer networks. Even though transformer networks had been primarily introduced for sequential data problems, such as Natural Language Processing (NLP), they process the entire input at once and take advantage of contextual information through their innate attention mechanism. Due to these properties the transformer network adaptation on image semantic segmentation task is a very promising and challenging task that should be investigated and extended with other successful and well established contemporary deep learning architectural blocks.

REFERENCES

[1] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[2] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33, no. 2, pp. 225–236, 2000. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320399000552

[3] W. Niblack, *An Introduction to Digital Image Processing*. DNK: Strandberg Publishing Company, 1985.

[4] N. R. Howe, "Document binarization with automatic parameter tuning," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 16, pp. 247–258, 2012.

[5] B. Su, S. Lu, and C. L. Tan, "Robust document image binarization technique for degraded document images," *IEEE Transactions on Image Processing*, vol. 22, no. 4, pp. 1408–1417, 2013.

[6] T. Lelore and F. Bouchara, "FAIR: A fast algorithm for document image restoration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 2039–2048, 2013.

[7] H. Nafchi, R. Farrahi Moghaddam, and M. Cheriet, "Phase-based binarization of ancient document images: Model and applications," *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 23, 05 2014.

[8] N. Mitianoudis and N. Papamarkos, "Document image binarization using local features and gaussian mixture modelling," *Image and Vision Computing*, vol. 38, pp. 33–51, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885615000360

[9] F. Jia, C. Shi, K. He, C. Wang, and B. Xiao, "Degraded document image binarization using structural symmetry of strokes," *Pattern Recognition*, vol. 74, pp. 225–240, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320317303849

[10] S. Bhowmik, R. Sarkar, B. Das, and D. Doermann, "GiB: a game theory inspired binarization technique for degraded document images," *IEEE Transactions on Image Processing*, vol. PP, pp. 1–1, 10 2018.

[11] S. He and L. Schomaker, "DeepOtsu: Document enhancement and binarization using iterative deep learning," *Pattern Recognition*, vol. 91, pp. 379–390, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320319300330

[12] Q. N. Vo, S. H. Kim, H. J. Yang, and G. Lee, "Binarization of degraded document images based on hierarchical deep supervised network," *Pattern Recognition*, vol. 74, pp. 568–586, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320317303394

[13] J. Zhao, C. Shi, F. Jia, Y. Wang, and B. Xiao, "Document image binarization with cascaded generators of conditional generative adversarial networks," *Pattern Recognition*, vol. 96, p. 106968, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320319302717

[14] H. Sheng and L. Schomaker, "CT-Net: Cascade T-Shape deep fusion networks for document binarization," *Pattern Recognition*, vol. 118, 05 2021.

[15] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2016.

[16] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support : 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, S...*, vol. 11045, pp. 3–11, 2018.

[17] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *European Conference on Computer Vision*, 2018.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[19] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," 2018.

[20] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," 2018.

[21] N. Ibtehaz and M. S. Rahman, "MultiResUNet : Rethinking the u-net architecture for multimodal biomedical image segmentation," *Neural Networks*, vol. 121, pp. 74–87, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608019302503

[22] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," 2016. [Online]. Available: https://arxiv.org/abs/1606.00915

[23] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017. [Online]. Available: https://arxiv.org/abs/1706.05587

[24] K. R. Ayyalasomayajula, F. Malmberg, and A. Brun, "PDNet: Semantic segmentation integrated with a primal-dual network for document binarization," *Pattern Recognition Letters*, vol. 121, pp. 52–60, apr 2019. [Online]. Available: https://doi.org/10.1016%2Fj.patrec.2018.05.011

[25] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer International Publishing, 2017, pp. 240–248. [Online]. Available: https://doi.org/10.1007%2F978-3-319-67558-9_28

[26] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "ICDAR 2013 document image binarization contest (DIBCO 2013)," 09 2011, pp. 1506–1510.

[27] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos, "ICDAR2017 competition on document image binarization (DIBCO 2017)," *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 01, pp. 1395–1403, 2017.

[28] I. Pratikakis, K. Zagoris, P. Kaddas, and B. Gatos, "ICFHR 2018 competition on handwritten document image binarization (H-DIBCO 2018)," 08 2018, pp. 489–493.

[29] H. Z. Nafchi, S. M. Ayatollahi, R. F. Moghaddam, and M. Cheriet, "An efficient ground truthing tool for binarization of historical manuscripts," in *2013 12th International Conference on Document Analysis and Recognition*, 2013, pp. 807–811.

[30] R. Hedjam, H. Z. Nafchi, R. F. Moghaddam, M. Kalacska, and M. Cheriet, "ICDAR 2015 contest on multispectral text extraction (MS-TEx 2015)," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015, pp. 1181–1185.