2017 14th IAPR International Conference on Document Analysis and Recognition

Using Convolutional Encoder-Decoder for Document Image Binarization

Xujun Peng, Huaigu Cao, and Prem Natarajan
Information Sciences Institute, University of Southern California, Marina del Rey, CA, USA
Email: {*xpeng, hcao, pnataraj*}*@isi.edu*

***Abstract*—Document image binarization is one of the critical initial steps for document analysis and understanding. Previous work mostly focused on exploiting hand-crafted features to build statistical models for distinguishing text from background. However, these approaches only achieved limited success because: (a) the effectiveness of hand-crafted features is limited by the researcher's domain knowledge and understanding on the documents, and (b) a universal model cannot always capture the complexity of different document degradations. In order to address these challenges, we propose a convolutional encoder-decoder model with deep learning for document image binarization in this paper. In the proposed method, mid-level document image representations are learnt by a stack of convolutional layers, which compose the encoder in this architecture. Then the binarization image is obtained by mapping low resolution representations to the original size through the decoder, which is composed by a series of transposed convolutional layers. We compare the proposed binarization method with other binarization algorithms both qualitatively and quantitatively on the public dataset. The experimental results show that the proposed method has comparable performance to the other hand-crafted binarization approaches and has more generalization capabilities with limited in-domain training data.**

I. Introduction

Document image binarization is an active topic for researchers because it is a fundamental step to many other higher level document process and analysis tasks. By separating text from background, document binarization can provide significant amount of semantic information and facilitate the document indexing and retrieval. Although various algorithms and models have been proposed, binarization is still a challenging task due to the complexity of degradations for different types of documents. For example, historical documents can easily suffer from ink bleed, smear, paper yellowing, etc. On the other hand, most degradations of hand-held devices captured document images are uneven/bad lighting, out-of-focus blur, etc.

Generally, binarization algorithms can be categorized as two types: global thresholding based and local thresholding based methods.

One classical global thresholding method is Otsu's algorithm [1], which used a single threshold that maximizes the inter-class variance or minimizes the intra-class variance to separate foreground from background. However, this approach assumes that the image only contains two classes and can be distinguished from its histogram, which is not true for most real applications. Thus, some researchers attempted to overcome the disadvantage of Otsu's binarization. Before binarizing the degraded historical document images using a global threshold, Shi and Govindaraju normalized the grey scale level of pixels in the image [2] . Based on the similar idea, Lu and Tan mapped the original grey level of each pixel to a new domain by using polynomial functions before thresholding [3].

To address the problems of the global thresholding based methods, many pioneer approaches for local thresholding were proposed, including Niblack's algorithm [4]. This method used mean and standard deviation within a small region to determine the local threshold, where high efficiency and accuracy can be obtained under well conditioned environments, such as scanner. Despite its success, the performance of Niblack's approach is sensitive to the local contrasts. Inspired by Niblack's method, Sauvola converted Niblack's linear decision plane to the non-linear decision plane, by adapting the local variance [5]. Similarly, Wolf *et al.* also modified Niblack's method to provide better results on low-contrast images [6]. But these methods' performances are highly depending on the window parameters. In order to avoid manually adjusting the window size to the content and take advantage of Sauvola's algorithm, Lazzara and Géraud described a multiscale binarization approach in [7]. In [8], to binarize uneven lighted document images, Peng *et al.* densely calculated the optimal segmentation surface for each pixel by using sigmoid function. In [9], Su *et al.* proposed to construct an adaptive contrast map by combining local image contrast and gradient, which was subsequently binarized using both global and local thresholds.

Although better binarization results can be expected by local thresholding methods than global methods, more empirical parameters are introduced which increases the system's complexity and decreases the algorithm's efficiency. To alleviate the burden of tuning the parameters heuristically, learning based approaches are designed for document binarizations. In [10], Wu *et al.* designed a set of features for document binarization and proposed a supervised learning approach to learn the optimal parameters for these features from data. Based on the observation that ink-bleed text has different slant than frontal text, Peng *et al.* used Gabor filter bank to distinguish verso and recto text from historical documents and applied a CRF to complete the binarization task [11]. To the same end, Sehad *et al.* also suggested using Gabor filters to capture the foreground text from ancient degraded documents [12].

2379-2140/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDAR.2017.121

Recently, due to its generalization abilities, neural networks have achieved great success in the field of image processing, computer vision, document preprocessing, etc [13]. By dividing document images into small patches, Afzal *et al.* designed a 2D long short-term memory (LSTM) framework to binarize old document images [14]. In [15], Pastor-Pellicer *et al.* used a 2-convolutional-layers neural network to accomplish binarization task.

In this work, we propose a deep learning based document binarization method, where deep convolutional neural networks (DCNNs) and deep transposed convolutional neural networks (DTCNNs) are applied. Particularly, the DCNNs are employed in our work as a encoder to extract feature representations for document images and DTCNNs are used to produce the binarized image from down-sampled feature representations. By relabeling the initial binarization output from our convolutional encoder-decoder using a conditional random field (CRF), the cleaned binarized document image can be obtained.

We organize the rest of the paper as follows. In section II, we describe the proposed DCNNs based binarization approach, including our motivation, the detailed architecture of the encoder-decoder and the post-processing method. Section III covers our experimental setup and analysis. And we conclude our paper in section IV

II. Proposed Architecture

A. Motivation

Since the work by Lecun *et al.* [16] where convolutional neural networks were applied for document recognition, DCNNs have dramatically improved the performance of computer vision tasks, including image classification/recognition, object detection, image retrieval, etc.

In the last few years, especially after the success of using AlexNet for large scale image classification [17], researchers have pushed DCNNs into the more challenging areas, such as semantic image labeling and segmentation. In [18], Long *et al.* proposed a fully convolutional network (FCN) framework for semantic segmentation on Pascal VOC12 challenge set, where an image was first fed into a variation of VGG-16 network to extract features [19]. Then the down-sampled feature maps were up-sampled to the original size through transposed convolutional layers with different strides. The final finer segmented image was obtained by combining multiple up-sampled images together. Based on the similar manner, Chen *et al.* used a DCNN to densely extract features for images where the hole algorithm was applied in the DCNN [20]. To recover detailed boundary of segmented images, they adopted a fully connected conditional random field (CRF) on the sub-sampled images. Unlike the approach proposed in [18] where only a single transposed convolutional layer was used, Badrinarayanan *et al.* employed multiple transposed convolutional layers to enlarge down-sampled feature maps to accomplish the semantic segmentation task [21]. Similarly, Noh *et al.* suggested even deeper convolutional and transposed convolutional networks to initially segment 20 objects for the image, which was followed by a CRF based relabeling approach [22].

Herein, by considering that the feature representation learnt from large amount data by using DCNNs has superior capability compared to the conventional hand-crafted features, such as SURF or SIFT features, it could allow us to take advantage of DCNNs to extract powerful features from raw document images. Further more, by realizing document binarization as a two-classes segmentation problem, we can also exploit the deep convolutional/transposed convolutional neural networks to facilitate the document binarization tasks.

B. Encoder-Decoder Architecture

In Fig. 1, we briefly illustrate the overall architecture of the neural network employed in our work. In this framework, the input RGB document image is fed into a stack of residual convolutional networks, where max-pooling is applied to down-sample the feature maps during the encoding pass. In the decoding pass, the down-sampled feature representations are up-sampled through a set of max-unpooling layers along with transposed convolutional layers. The output of the proposed neural network is the confidence of text/background for each pixel through a softmax layer, which can be used for the final binarization.

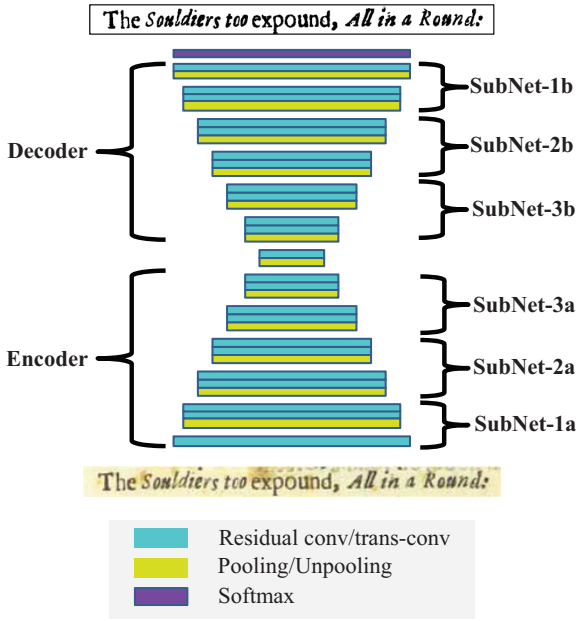
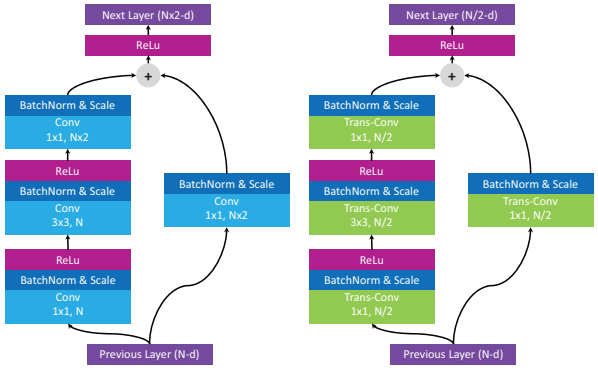


Fig. 1. The proposed encoder-decoder architecture of the DCNN binarization approach.

As Simonyan and Zisserman suggested in their work that better performance can be expected by deeper neural networks [19], in this study we propose to use a deeper neural network for our binarization task than the DNNs implemented in [21] and [22] where only shallow or conventional convolutional networks were used. However, one of the potential difficulties with very deep networks is vanishing/exploding gradient. To



(a) Residual convolutional module. (b) Residual transposed convolutional module.

Fig. 2. Residual inception modules applied in our work.

overcome this problem, we use the bottleneck-like residual neural networks (RNNs) proposed in [23] as the basic building module in the proposed framework, which passes the input of particular convolutional layer straight through to the last layer. In Fig. 2, we show the detailed structure of the residual convolutional module for the encoder and the residual transposed convolutional module for the decoder. For the residual convolutional module, in the main path, the input features are passed through three convolutional layers with kernel size 1×1, 3×3 and 1×1. And the number of filters is doubled from N to $2 \times N$ by the last 1×1 layer. And in the shortcut path, the filter number is directly increased by a 1×1 convolutional layer. Similarly, in the residual transposed convolutional module, features are passed through three transposed convolutional layers in the main path and one transposed convolutional layer in the shortcut path. The input feature depth is immediately decreased from N to $N/2$ in the first layer of this module. For all convolutional/transposed-convolutional layers inside residual module, the batch norm layer and scale layers are attached on them. Table I shows the output feature size of each residual module in our network.

In general, very deep neural networks tend to be suffered by overfitting problem during the training phase if training set's size is not big enough. In order to avoid this situation, we split the entire network into three mirrored sub-networks, which are illustrated in Fig. II. Consequently, we train the convolutional encoder-decoder in three phases, starting from SubNet-1 with all the training data and a predefined learning rate. Then the SubNet-2 and SubNet-3 are sequentially added into the system for training with the same learning rate, whilst the pre-trained sub-networks' learning rates are decreased. Eventually, three convolutional encoder-decoders with different depths are obtained.

C. Post-processing

Although the confidence map generated from the convolutional encoder-decoder network can be binarized using a simple global thresholding based approach to create the final output, it can only predict the rough position of text areas because the receptive fields in the decoder only yield smooth response. So post processing steps are required to estimate the fine-grained edges for texts.

Note that three convolutional encoder-decoders we trained have different resolution capabilities due to various minimum feature dimensions, it is a reasonable choice to combine the final responses from multiple networks to boost the binarization accuracy. Prior to the combination, the fully connected CRFs is applied on the confidence maps to suppress noises and refine the boundary of texts [24].

Formally, CRFs employ a random field $\boldsymbol{X}$ over a set of variables $\{x1, \cdots, x_N\}$ to model an image I given its features $\boldsymbol{Y}$, where N is the image size. For this random field, a complete graph $\mathcal{G}$ is created and a set of all unary and pairwise cliques $\mathcal{C}_{\mathcal{G}}$ is defined for the random field. In this work, random variables have two possible labels $\mathcal{L} = \{l1, l2\}$ which correspond to background and text. Thus, a Gibbs energy function can be applied to this random field:

$$E(\boldsymbol{x}) = \sum_i \varphi_u(x_i, Y) + \sum_{(i,j)\in\mathcal{C}_{\mathcal{G}}} \varphi_p(x_i, x_j, Y), \qquad (1)$$

where $\boldsymbol{x}$ is the class labels assigned to pixels.

In our implementation, the unary potential $\varphi_u(x_i, Y)$ is fulfilled by using confidence scores yield by the softmax layer of the convolutional encoder-decoder for each pixel.

To the pairwise potential, it has the form:

$$\varphi_p(x_i, x_j, Y) = \mu(x_i, x_j)\left\{\omega_1 \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2}\right) + \omega_2\left(-\frac{|p_i - p_j|^2}{2\theta_\beta^2} - \frac{|I_i - I_j|^2}{2\theta_\gamma^2}\right)\right\}, \qquad (2)$$

where $\mu(x_i, x_j) = [x_i \neq x_j]$ is a compatibility function, p_i and p_j are spatial positions, and I_i and I_j are color vectors of pixel i and j, respectively. The first exponential term in Eq. 2 uses both colors and positions of neighboring pixels to smooth the same class and the second exponential term in Eq. 2 removes isolated noises according their positions. These two terms' weights are controlled by ω_1 and ω_2. Parameter θ_α, θ_β and θ_γ which influence the degree of nearness and similarity between neighboring pixels can be learned from a training set.

In the testing phase, each document image is initially used as input to the three convolutional encoder-decoders we trained. Thereafter the confidence maps generated from these networks are relabeled by fully connected CRFs, which are then combined together by the majority voting scheme to produce the final binarized document image.

III. Experimental Results

In this section, we conducted the proposed convolutional encoder-decoder based document binarization approach and evaluate it on the public datasets.

TABLE I
FEATURE MAP SIZE OF EACH RESIDUAL MODULE.

SubNet-2		SubNet-2		SubNet-3	
layer type	output size	layer type	output size	layer type	output size
Conv-1	$448 \times 448 \times 32$	Conv-4	$112 \times 112 \times 64$	Conv-8	$28 \times 28 \times 128$
Conv-2	$224 \times 224 \times 32$	Conv-5	$112 \times 112 \times 64$	Conv-9	$28 \times 28 \times 256$
Conv-3	$224 \times 224 \times 64$	Conv-6	$56 \times 56 \times 64$	Conv-10	$14 \times 14 \times 256$
		Conv-7	$56 \times 56 \times 128$	Conv-11	$14 \times 14 \times 512$
		T-Conv-7	$56 \times 56 \times 64$	T-Conv-11	$14 \times 14 \times 256$
T-Conv-3	$224 \times 224 \times 32$	T-Conv-6	$56 \times 56 \times 64$	T-Conv-10	$14 \times 14 \times 256$
T-Conv-2	$224 \times 224 \times 32$	T-Conv-5	$112 \times 112 \times 64$	T-Conv-9	$28 \times 28 \times 128$
T-Conv-1	$448 \times 448 \times 2$	T-Conv-4	$112 \times 112 \times 64$	T-Conv-8	$28 \times 28 \times 128$

A. Synthetic Training Set

To train a feasible deep convolutional encoder-decoder, one of the hurdles is the lack of enough training data with manually labeled text. To address this issue, we create a synthetic dataset with large amount images for the training purpose.

In our experiment, 150 PDF files are initially collected which contain multiple pages in each file. We then convert these PDF files into three sets of gray scale images based on various DPIs (200 dpi, 300 dpi and 600 dpi). In each subsets, there are 3119 clean images are generated. The images we collected cover a wide range of topics, including technique reports, research papers, news, fictions, politics, etc. Meanwhile, in order to increase the generalization capability of the system, these images include seven languages and several different font types. In Fig. 3, the distributions of the languages in the training set are shown.

To simulate the real document images that are captured by hand-held devices or suffered by aging, we randomly add Gaussian blurs, salt and pepper noises, fade/shade/lighting effects with different degrees onto the clean images to produce the degraded images. By using this approach, a total of 200,000 synthetic images are created as our training set. The ground truth of these degraded images are obtained by simply binarizing the corresponding clean document images using a global thresholding method. Fig. 4 illustrates several synthetic images from the training corpus.

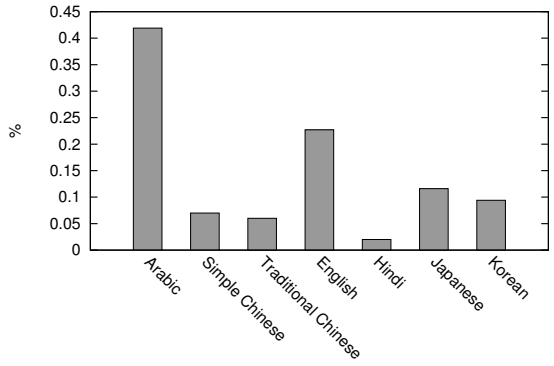


Fig. 3. Language distributions of the training set.

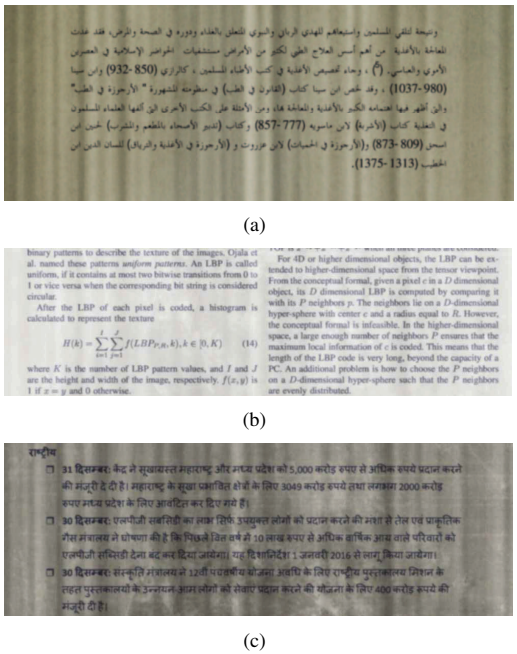


Fig. 4. Sample synthetic images in training set.

B. Training

Training is performed by three stages as described in Sec. II-B. The SubNet-1 is trained first using stochastic gradient descent (SGD) algorithm with learning rate $1.0\mathrm{e}{-6}$. Prior to feeding the training data into SubNet-1, the images are randomly rotated between $[-10°, +10°]$ and cropped to fit the size of input layer. Color normalization by subtracting the mean value over the training set from each pixel is also carried out for the cropped images. When training has converged for SubNet-1, SubNet-2 is attached to SubNet-1 where SubNet-1's learning rate is lowered to $1.0\mathrm{e}{-9}$ and SubNet-2's learning rate is set to $1.0\mathrm{e}{-6}$. The same training scheme is taken for SubNet-3. All sub-networks' weights are initialized using Xavier approach before training.

C. Evaluation Protocol

In the testing phase, we use LRDE document binarization dataset (LRDE DBD) [25] to evaluate the proposed binarization method on scanned document images. This dataset

contains 125 clean French documents extracted from PDF files with 300-dpi resolution, where images have been removed from original documents. Based on these clean images, 125 binarized images and 125 scanned documents along with their OCR groundtruth are also provided.

In our experiment, we split LRDE dataset into two subsets: tuning set and testing set. Because the scanned document images in LRDE are printed, scanned and registered to match the clean documents, the wiggling effect is inevitable in the edge areas for text which damages the tuning of the system. Thus, instead of using the scanned images for tuning in this experiment, we randomly select 100 clean document images and their corresponding binarized images for the tuning purpose and the learning rate is fixed to $1.0\mathrm{e}{-9}$ during tuning phase. The remaining 25 scanned document images are used for testing in this work. For each test image, the confidence scores through three convolutional encoder-decoders are calculated initially, then the final binarized document image is produced by using fully connected CRFs based refining and majority voting approach as described in Sec. II-C.

To measure the performance of the proposed approach, we compute the word error rate (WER) of optical character recogniton (OCR) on the binarized document images. In our implementation, we extract small text lines from the binarized document images and use Tesseract-4.0 to perform the OCR on line images. In table II, we compare the WERs of the proposed approach with other binarization approaches.

TABLE II
COMPARISON OF WORD ERROR RATE (WER) OF DIFFERENT BINARIZATION METHODS ON LRDE SET.

Binariation approach	WER (%)
Otsu [1]	15.0
Niblack [4]	15.2
Sauvola [5]	14.7
Kim [26]	15.3
Lelore [27]	15.6
TMMS [28]	30.7
Su2011 [29]	17.1
Sauvola Multiscale [7]	15.0
Proposed method	14.8

From this table, we can see that with limited tuning images, the proposed approach has a comparable performance to the other state-of-the-art binarization methods for scanned document images.

In order to assess the generalization ability of the proposed binarization approach, we test two out-of-domain document images sets using learned convoluitonal encoder-decoder and fully connected CRFs without fine-tune.

To evaluate the performance of the designed method on historical handwritten document images, we selecte H-DIBCO 2016 test set for testing [30]. In this experiment, the 10 benchmark handwritten images from H-DIBCO 2016 set, which have different types of degradations, are binarized and compared with the ground truth images. We show the F-measure and peak signal-to-noise ratio (PSNR) in table III, along with the results from other top ranked binarization approaches for H-DIBCO 2016.

TABLE III
EVALUATION RESULTS FOR DIFFERENT BINARIZATION METHODS ON H-DIBCO 2016 SET.

Method	F-Measure	PSNR
Otsu [1]	86.61 ± 7.26	17.80 ± 4.51
Sauvola [5]	82.52 ± 9.65	16.42 ± 2.87
Method in [31]	76.10 ± 13.81	15.35 ± 3.19
Method in [32]	88.11 ± 4.63	18.00 ± 3.41
Method in [33], [34]	88.72 ± 4.68	18.45 ± 3.41
Method-1 in [30]	85.57 ± 6.75	17.50 ± 3.43
Proposed method	88.07 ± 4.86	18.13 ± 3.13

From this table, we can see that even without fine-tune process, the binarization approach proposed in this paper has a comparable performance to the top methods for H-DIBCO 2016. It is necessary to note that the method-1 described in [30] used a fully convolutional network to perform the binarization but has lower F-measure score and PSNR value than our method, which shows the superior capability of the deep convolutional encoder-decoder for binarization compared to other DCNNs based approaches.

Further more, we test a set of document images captured by hand-held devices with bad/uneven lighting using the proposed binarization method. In Fig. 5, three binarizaiton examples from the proposed system are illustrated. From this figure, it can be observed that for these severely degraded document images, the proposed model can still provide reliable binarization results, which demonstrates its generalization capability.

IV. CONCLUSION

In this paper, we propose a document binarization approach which is based on deep convolutional encoder-decoder and fully connected CRFs. Unlike the conventional binarization algorithms that hand-crafted features are empirically designed for particular type of document images, we train the encoder-decoder using large amount of synthetic data and tune the system with limited amount of in-domain data. The experiments show that with limited amount of or none of in-domain data, the proposed binarization method is comparable to the other state-of-the-art approaches and has powerful generalization capability.

In the future, we will investigate the approach of integrating CRFs into the learning phase of deep convolutional encoder-decoder to boost binarization's performance.

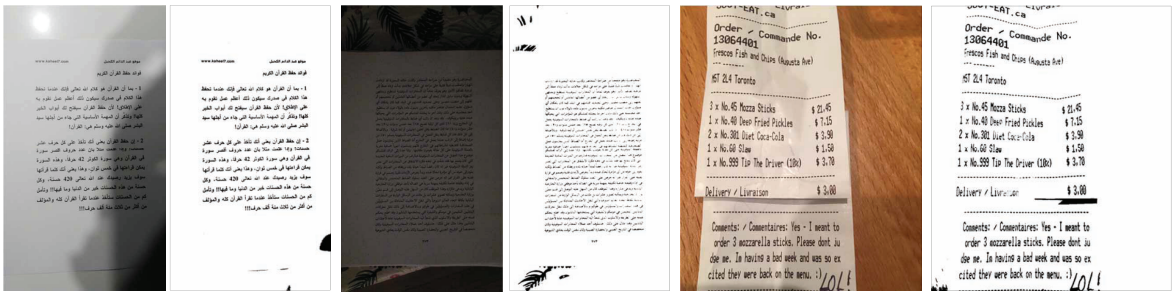
REFERENCES

[1] N. Otsu, "A threshold selection method from gray level histograms," *IEEE Trans. Systems, Man and Cybernetics*, vol. 9, pp. 62–66, Mar 1979.

[2] Z. Shi and V. Govindaraju, "Historical document image segmentation using background light intensity normalization," in *Proc. SPIE, Document Recognition and Retrieval XII*, vol. 5676, no. 1, 2005, pp. 167–174.

[3] S. Lu and C. L. Tan, "Binarization of badly illuminated document images through shading estimation and compensation," in *Proc. IEEE 9th ICDAR*, vol. 1, Sept. 2007, pp. 312–316.

[4] W. Niblack, *An Introduction to Digital Image Processing*. N.J.: Prentice Hall: Englewood Cliffs, 1986.



(a) (b) (c)

Fig. 5. Binarization results for hand-held devices captured images.

[5] J. Sauvola, T. Seppanen, S. Haapakoski, and M. Pietikainen, "Adaptive document binarization," in *Proceedings of the Fourth International Conference on Document Analysis and Recognition*, vol. 1, 1997, pp. 147–152.

[6] C. Wolf, J. M. Jolion, and F. Chassaing, "Text localization, enhancement and binarization in multimedia documents," in *Object recognition supported by user interaction for service robots*, vol. 2, 2002, pp. 1037–1040.

[7] G. Lazzara and T. Géraud, "Efficient multiscale sauvola's binarization," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 17, no. 2, pp. 105–123, 2014.

[8] X. Peng, S. Setlur, V. Govindaraju, and R. Sitaram, "Binarization of camera-captured document using a MAP approach," in *Proc. SPIE, Document Recognition and Retrieval XVIII*, vol. 7874, 2011, pp. 78 740R–78 740R–8.

[9] 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, April 2013.

[10] Y. Wu, P. Natarajan, S. Rawls, and W. AbdAlmageed, "Learning document image binarization from data," in *IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 3763–3767.

[11] X. Peng, H. Cao, K. Subramanian, R. Prasad, and P. Natarajan, "Exploiting stroke orientation for CRF based binarization of historical documents," in *12th International Conference on Document Analysis and Recognition*, 2013, pp. 1034–1038.

[12] A. Sehad, Y. Chibani, and M. Cheriet, "Gabor filters for degraded document image binarization," in *14th International Conference on Frontiers in Handwriting Recognition*, 2014, pp. 702–707.

[13] A. Rehman and T. Saba, "Neural networks for document image preprocessing: state of the art," *Artificial Intelligence Review*, vol. 42, no. 2, pp. 253–273, 2014.

[14] M. Z. Afzal, J. Pastor-Pellicer, F. Shafait, T. M. Breuel, A. Dengel, and M. Liwicki, "Document image binarization using LSTM: A sequence learning approach," in *Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing*, 2015, pp. 79–84.

[15] J. Pastor-Pellicer, S. España-Boquera, F. Zamora-Martínez, M. Z. Afzal, and M. J. Castro-Bleda, "Insights on the use of convolutional neural networks for document image binarization," in *Advances in Computational Intelligence: 13th International Work-Conference on Artificial Neural Networks, IWANN. Proceedings, Part II*, 2015, pp. 115–126.

[16] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 1097–1105.

[18] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[20] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *ICLR*, 2015.

[21] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for scene segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[22] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1520–1528.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[24] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in Neural Information Processing Systems 24*, 2011, pp. 109–117.

[25] G. Lazzara, R. Levillain, T. Geraud, Y. Jacquelet, J. Marquegnies, and A. Crepin-Leblond, "The scribo module of the olena platform: A free software framework for document image analysis," in *International Conference on Document Analysis and Recognition*, 2011, pp. 252–258.

[26] I.-J. Kim, "Multi-window binarization of camera image for document recognition," in *Ninth International Workshop on Frontiers in Handwriting Recognition*, 2004, pp. 323–327.

[27] T. Lelore and F. Bouchara, "Super-resolved binarization of text based on the fair algorithm," in *International Conference on Document Analysis and Recognition*, 2011, pp. 839–843.

[28] J. Fabrizio, B. Marcotegui, and M. Cord, "Text segmentation in natural scenes using toggle-mapping," in *16th IEEE International Conference on Image Processing (ICIP)*, 2009, pp. 2373–2376.

[29] B. Su, S. Lu, and C. L. Tan, "Binarization of historical document images using the local maximum and minimum," in *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, 2010, pp. 159–166.

[30] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos, "ICFHR2016 handwritten document image binarization contest (H-DIBCO 2016)," in *15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2016, pp. 619–623.

[31] T. Sari, A. Kefali, and H. Bahi, "Text extraction from historical document images by the combination of several thresholding techniques," *Advances in Multimedia*, vol. 2014, 2014, art. ID 934656.

[32] H. Ziaei Nafchi, R. Farrahi Moghaddam, and M. Cheriet, "Historical document binarization based on phase information of images," in *ACCV Workshops*, 2012, pp. 1–12.

[33] A. Hassaïne, E. Decencière, and B. Besserer, "Efficient restoration of variable area soundtracks," *Image Analysis & Stereology*, vol. 28, no. 2, 2011.

[34] A. Hassaïne, S. Al-Maadeed, and A. Bouridane, "A set of geometrical features for writer identification," in *19th International Conference on Neural Information Processing (ICONIP)*, 2012, pp. 584–591.