2018 16th International Conference on Frontiers in Handwriting Recognition

A deep Convolutional Encoder-Decoder Network for Page Segmentation of Historical Handwritten Documents into Text Zones

Panagiotis Kaddas[1,2] Basilis Gatos[1]

[1]*Computational Intelligence Laboratory, Institute of Informatics and Telecommunications*
National Center for Scientific Research Demokritos
GR-153 10, Agia Paraskevi, Athens, Greece
{*pkaddas, bgat*}*@iit.demokritos.gr*

[2]*Department of Informatics and Telecommunications, University of Athens, GR-157 84, Athens, Greece*

***Abstract*—Recent research activity for page segmentation and pixel-labeling problems focuses strongly on deep Neural Network architectures. In this paper, we present a Convolutional Encoder-Decoder based method for the segmentation of historical handwritten images into distinct text zones. This is achieved by labeling each pixel of the image to one of the predefined classes (main body, comments, decorations, periphery, background). Traditional methods make use of prior knowledge of documents and rely on data-oriented features and experimental rules. We propose a method using Convolutional Encoder-Decoder pairs and we show that deep architectures fit properly to our problem. Experiments on different public datasets demonstrate the effectiveness of the proposed method that outperforms previous techniques in many cases.**

***Keywords*-historical document image processing; page segmentation; deep convolutional neural networks;**

I. Introduction

Segmentation of historical handwritten documents into several text zone categories is a crucial step in Handwritten Text Recognition (HTR) and Document Understanding tasks. Some factors like unconstrained layout, writing style, local skew and document degradations make segmentation of these collections a challenging problem compared to page segmentation of machine printed document images. Our goal is to develop a robust pixel-labeling method able to detect all regions of interest in a historical handwritten document image. This is achieved by classifying each pixel to one of the predefined classes (main body, comment, decoration, background, periphery) [1], as given in Figure 1.

Extensive research has been done on unsupervised page segmentation methods [2], [3], [4], [5]. These methods are considered to be data-oriented because they rely on experimental rules and prior knowledge of the document corpus. In contrast, supervised techniques [1], [6], [7], [8], [9] apply machine learning algorithms on smaller processing units like connected components and superpixels in order to automatically learn generic and discriminative features and patterns.

Convolutional Neural Networks (CNN) and Convolutional Encoder-Decoder (CED) architectures have recently achieved state-of-the-art perfomance in various fields like image classification [10], [11] and pixel-labeling of natural images [12], [13] or historical handwritten documents [14]. Spatial information can be learned efficiently by a CNN or a CED [15] without using rule-based features. Therefore, these networks fit properly to our pixel-labeling problem.

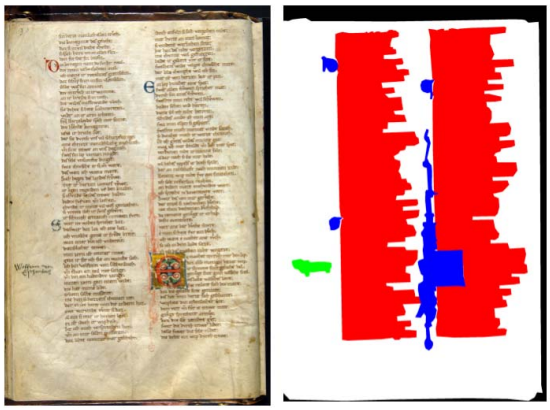

Figure 1: Page Segmentation into distinct regions using pixel-labeling. Periphery, background, main body, decoration, and comment regions are represented with black, white, red, blue and green color respectively.

In this work, we train a deep network using a CED architecture. It consists of five Encoder-Decoder pairs and takes an RGB image of arbitrary size as input and outputs a labeled image. We conducted experiments concerning the effect of a simple pre-processing step. Evaluation on public historical handwritten datasets show that the proposed method achieves superior results in many cases, when compared to state-of-the-art techniques.

The rest of the paper is organized as follows. Section II presents some of the related works, Section III introduces the proposed CED, Section IV demonstrates our experimental results and Section V presents the conclusion of this work.

II. Related Work

In this section some relative and representative works are reviewed. Our focus is on the page segmentation of

978-1-5386-5875-8/18/$31.00 ©2018 IEEE
DOI 10.1109/ICFHR-2018.2018.00053

historical handwritten documents into text zones. Layout inconsistencies, writing style irregularities and low quality of such documents are challenges that developed methods try to deal with.

Layout analysis based on the detection of table rule lines and page margins is presented by Bulacu et. al in [2]. Contour tracing is also used in order to preserve ascenders and descenders by generating curvlinear segmentation paths between text lines. In [3], image binarization and Laplacian of Gaussian are used to extract connected components. Then, each connected component is classified as text or non-text using basic features like bounding box coordinates, stroke width, estimated distance between text lines and an energy minimization method.

Extraction of main text zones and text lines using prior knowledge of the corpus structure is presented in [5]. A vertical black run profile is combinbed with vertical white runs in order to split the document into columns. Then, a horizontal refinement of these columns is applied. Finally, text lines are extracted using a Hough transform based method. Despite that all these unsupervised methods achieve high performance on the datasets that they use, they lack of generality due to their many hand-crafted rules that require prior knowledge of the dataset.

In contrast to unsupervised techniques, methods that rely on supervised algorithms are more efficient and robust for page segmentation problems. Nicolas et al. [16] use stohasitc and contextual models in order to learn spatial variability and combine some prior knowledge about the global structure of the document. The goal is to extract distinct regions of the manuscripts like main body, header, footer, page number and marginal annotations. The method is applied at a finer analysis level in order to split the document into background, erasures, diacritics, and textual components.

In [7], the authors propose a feature learning technique for the segmentation method of Arabic historical documents. A multilayer perceptron (MLP) is trained to classify each connected component as one of two classes (main body and side-notes). The input of the MLP consists of features extracted from connected cmponents (normalized height, foreground area, relative distance, orientation and neighborhood information).

Chen et al. [17], train an autoencoder that tries to reconstruct itself. Then, a classifier is applied to predict a label for each superpixel. In [6], the same authors refine their results by applying a superpixel algorithm (SLIC) as a pre-processing step. Post-processing of this technique using Conditional Random Fields (CRF) [8] was used in order to model spatial and contextual information of the superpixels. CRF technique excels in terms of both accuracy and time. Recently, in [1], a simple Convolutional Neural Network on superpixels i trained using only one convolution layer followed by a fully connected layer and result into neurons representing the probability for each class. Also in this approach, superpixels are used as processing unit.

Jobin et al. [9] propose two methods for pixel-labeling of historical handwritten documents through superpixels using weights from a pre-trained network. They first apply convolutions on the image and a descriptor is extracted either through fully connected layers (*FC-CNN*) or through a Fischer-vector encoder (*FV-CNN*). Finally, a Support Vector Machine (SVM) classifier is used to label each superpixel.

Finally, Xu et al. [14] present a fully convolutional network (FCN) which produces a coarse pixel-level segmentation of historical handwritten documents. A post-processing step is applied based on connected components analysis and overlapping cases are identified using size and spatial information.

III. Proposed Method

This Section describes the proposed CED architecture used to classify each pixel of the image to one of the predefined classes, following the experimental protocol of [1]. Also, in our case we have 5 classes (main body, comments, decorations, periphery, background) (Figure 1). This labeling process results into the page segmentation of documents into distinct zones, combining both spatial and texture information.

A. Pre-processing

An important advantage of CED networks is that simple pre-processing steps are sufficient in order to improve the learning procedure. In this work, we use as input a 3-channel RGB image and we apply a *Local Contrast Normalization (LCN)* algorithm [18] for pre-processing. The effect of LCN technique is examined further in Section IV. A normalization is applied locally using a 9×9 Gaussian weighting window. Each channel is processed separately. When normalizing, the local mean value is subtracted from each pixel and then the result is divided by the local standard deviation. With the LCN technique, differences in and between feature maps are highlighted and spatial variability is emphasized.

B. CED Architecture

A general architecture of a CED network is shown in Figure 2. Our network is similar to the VGG-based [11] Encoder-Decoder proposed in [13] and used for semantic segmentation of natural images. An input (RGB) image I_{in} is resized to a fixed size ($640\times416\times3$) and forwarded to the first *Encoder* of the network.

An *Encoder* consists of a stack of batch-normalized (BN) [19] convolutional layers, which are fed into the element-wise rectified non-linearity unit (ReLU) $max(0, x)$. Then a sub-sampling step is applied (Max Pooling) for dimensionality reduction. Optionally, a dropout layer [20] can be applied to the output in order to prevent over-fitting. In this work, we apply dropout only at the training phase. Each Encoder has a symmetric *Decoder*, forming an *Encoder-Decoder* pair.

In the *Decoder*, up-sampling is applied using pooling indices extracted from the respective encoder in order to preserve boundary details. The last decoder is connected to a logistic regression layer using the softmax function.

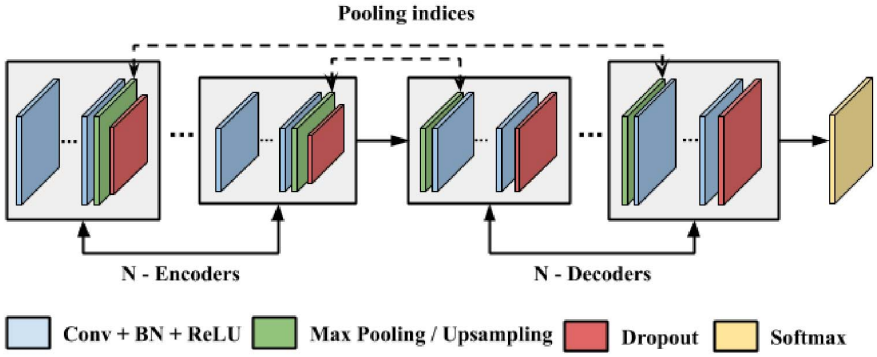

Figure 2: A schematic of a Convolutional Encoder-Decoder

The output of the softmax layer is a new image P of shape $(640 \times 416 \times C)$, where C is the number of classes ($C = 5$). Each pixel of P is a $(1 \times 1 \times C)$ vector representing the probability distribution of each class at this position. Image I_{out} of the predicted labels is calculated by detecting the maximum probability for each pixel, such that

$$I_{out}(x, y) = \underset{c}{\operatorname{argmax}}\, P(x, y, c), \tag{1}$$

where $c \in [1, 5]$. Finally, I_{out} is resized to the dimension of I_{in}, but with a single channel, using nearest-neighbor interpolation.

The network consists of five Encoder-Decoder pairs. The first two (outer) have 2 convolutional layers, while the other have 3. A 3×3 kernel and a stride equal to 1 are used for all convolutions. Layer depth is $64, 128, 256, 512, 512$ with respect to the pair that belongs to. Max pooling and up-sampling are applied using a 2×2 window with no overlaps.

IV. Experiments

A. Datasets

Six public historical handwritten datasets are used through our experiments (also used in [1], [6], [8], [9]) (Table I).

The first collection (*Set-1*) consists of the annotated *G. Washington, Parzival, St. Gall* datasets [21]. These documents were created in 18th, 13th and 9th century respectively. Some of the characteristics of this collection is that it contains gray-scale and RGB document images written in English or Latin by multiple writers. Text annotations contain different zones given at a region level like text columns or paragraphs.

The second collection (*Set-2*) contains the *CB55, CSG18, CSG863* [22] datasets. Documents in this collection have complex layout and were written in 11th or 14th century. The number of writers is unknown. Provided text annotations [22] are at line level. Note that periphery and background class are not separated and considered as one.

Table I: Datasets used in our experiments. TR, VA, TE denote training, validation and test sets respectively.

Datasets	Image Size (pixels)	TR	VA	TE
G.Washington	2200×3400	10	5	5
Parzival	1664×2496	22	2	13
St.Gall	2000×3008	20	10	30
CB55	4872×6496	20	10	10
CSG18	3328×4992	20	10	10
CSG863	3328×4992	20	10	10

B. Evaluation Metrics

Evaluation metrics used in this work were originally introduced in [12]. Methods presented in [1], [6], [8], [9] are also evaluated using this protocol.

Let n_{ij} be the number of pixels of class i predicted as of class j , C is the number of classes and $t_i = \sum_j n_{ji}$ is total of pixels that belong to class i. The following metrcis are defined:

- Pixel Accuracy:

$$PA = \frac{\sum_i n_{ii}}{\sum_i t_i} \tag{2}$$

- Mean Accuracy:

$$MA = \frac{1}{C} \times \sum_i \frac{n_{ii}}{t_i} \tag{3}$$

- Mean Intersection over Union (Mean IU):

$$mIU = \frac{1}{C} \times \sum_i \frac{n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}} \tag{4}$$

- Frequency weighted Mean IU:

$$fwIU = \frac{1}{\sum_k t_k} \times \sum_i \frac{t_i n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}} \tag{5}$$

C. Training

We train our CED network using Adam optimizer [23]. Also, L2 norm and weight decay (5×10^{-5}) are applied on every weight. Dropout is applied on the three deeper *Encoder-Decoder* pairs, where each neuron is discarded

Table II: Evaluation of different CED variants (in percentage).

CED Variant	*Set-1*				*Set-2*			
	PA	MA	mIU	fwIU	PA	MA	mIU	fwIU
RGB-CED	85	77	61	80	94	60	53	89
LCN-CED	**97**	**86**	**82**	**95**	**96**	**77**	**68**	**92**

Table III: Comparative Evaluation Results (in percentage) of the proposed CED.

Method	*G. Washington*				*Parzival*				*St. Gall*			
	PA	MA	mIU	fwIU	PA	MA	mIU	fwIU	PA	MA	mIU	fwIU
Local MLP [6]	87	89	75	83	91	64	58	86	95	89	84	92
CRF [8]	91	90	76	85	93	70	63	88	97	88	84	94
CNN [1]	91	91	77	86	94	75	68	89	98	90	87	96
FV-CNN [9]	95	93	81	91	**97**	**76**	**71**	**94**	**99**	**91**	**88**	**98**
Proposed (LCN-CED)	**96**	**94**	**83**	**92**	94	75	69	90	98	90	87	97
Method	*CB55*				*CSG18*				*CSG863*			
	PA	MA	mIU	fwIU	PA	MA	mIU	fwIU	PA	MA	mIU	fwIU
Local MLP [6]	83	53	42	72	83	49	39	73	84	54	42	74
CRF [8]	84	53	42	75	86	47	37	77	86	51	42	78
CNN [1]	86	59	47	77	87	53	41	79	87	58	45	79
FV-CNN [9]	95	73	64	91	92	72	60	89	94	71	61	91
Proposed (LCN-CED)	**96**	**75**	**67**	**92**	**96**	**80**	**69**	**92**	**96**	**75**	**66**	**92**

with a probability of 0.5 at each iteration. We use a constant learning rate of 10^{-4}. The goal is to minimize the cross-entropy loss between the predicted probabilities of each class and the one-hot encoded ground-truth labels. The model is trained until convergence is reached and we keep the model that performed better on the validation set, based on the *mean IU* metric.

D. Experimental Results

In order to examine the effect of *LCN* pre-processing technique, we conduct an experiment (Table II) by comparing this model (*LCN-CED*) to the one with no pre-processed images (*RGB-CED*). We train both variants on each set (*Set-1, Set-2*), using all the respective training images. As shown in Table II, the *LCN-CED* model performs significantly better in contrast to *RGB-CED* for all metrics. This is expected because local spatial variability is highlighted, enabling the CED model to learn discriminative features faster and recognize these differences efficiently.

Finally, we compare the proposed *LCN-CED* model to other state-of-the-art techniques presented in [1], [6], [8] and [9]. Results are given in Table III. Note that these methods train a different network on each one of the six datasets. On the contrary, we followed the more realistic scenario of training just two models using all the training images of each set (*Set-1*, *Set-2*). This results into models with greater generalization ability because more robust features are learned through documents of different nature. The reason that we do not train one model is that the provided annotations for the two sets are different. Ground-truth images for *Set-1* are provided at a coarse level (paragraphs, text columns), while ground-truth images for *Set-2* are given at a finer level (lines). As a result, testing our CED on these two sets is considered as two different pixel-labeling problems.

We can see that in *G. Washington, Parzival* and *St. Gall* datasets (*Set-1*), our CED network is competitive to the network proposed in [9] and gives higher results than all the other methods, while in *CB55, CSG18* and *CSG863* datasets (*Set-2*) the CED model is superior to all other methods and proves the fact that the it can deal with complex datasets, despite that no pre-trained weights were used for initialization as in [9]. Moreover, in *Set-2*, the results of our *LCN-CED* are of smaller variance for all three datasets. Some segmentation results of the proposed method are given in Figure 3.

Our CED is implemented using the open-source machine learning python framework TensorFlow[1]. Also, an NVIDIA GTX-1080 GPU is used, enabling as to train our model in less than 4 hours.

V. Conclusion

In this work, we propose a Convolutional Encoder-Decoder (CED) network for the page segmentation of historical handwritten documents into distinct text zones through pixel-labeling. While most of the research focuses on superpixel processing, we show that a simple local contrast normalization technique is sufficient as a pre-processing step and enables our model to learn spatial and texture variability from the whole image instead of image patches. No prior knowledge of the dataset is required and there are no data-oriented rules. Also, in contrast to other techniques, the output of the our network needs no post-processing. Experiments on six public datasets show that the proposed model outperforms other techniques in many cases. It is worth to notice that we did not train a model

[1]https://www.tensorflow.org/

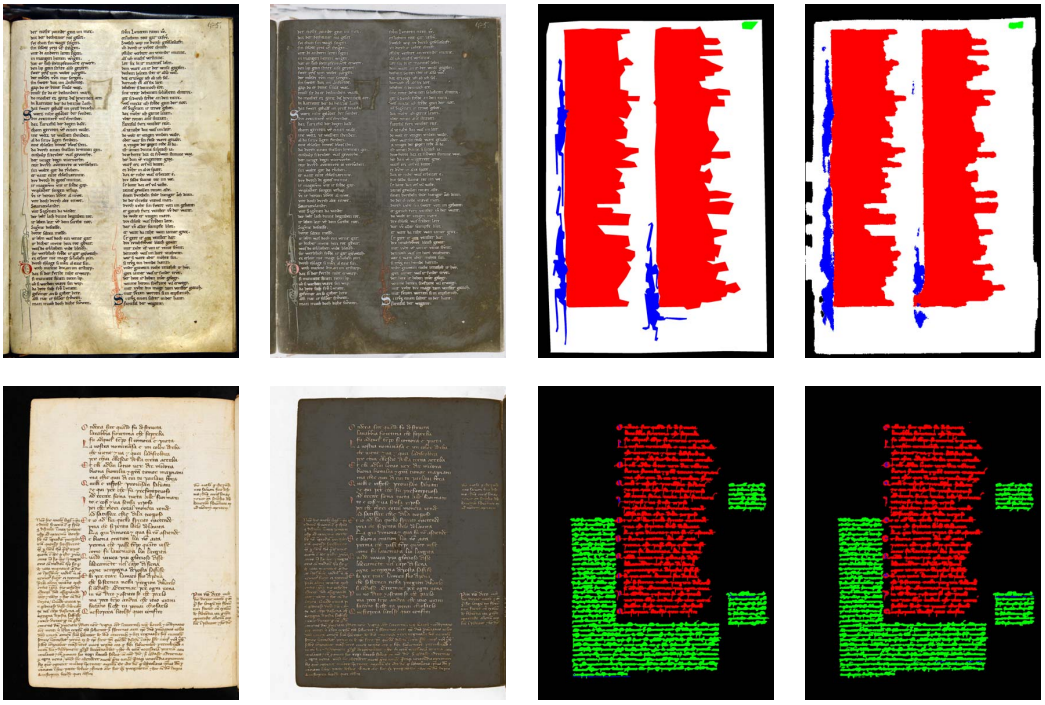

Figure 3: Segmentation results of the proposed CED. From top to bottom: Example images from (1) Parzival and (2) CB55 datasets. First column shows the original images. Second column shows the result of the LCN pre-processing step. Third column shows ground-truth labels and last column shows the predicted labels.

for each dataset as in [1], [6], [8], [9], but we trained just one model per set that includes three datasets. This reflects the generalization ability of the proposed framework.

ACKNOWLEDGMENT

This work has been supported by the program of Industrial Scholarships of Stavros Niarchos Foundation[2] as well as the European Unions H2020 grant READ (Recognition and Enrichment of Archival Documents) (Ref: 674943)[3].

REFERENCES

[1] K. Chen, M. Seuret, J. Hennebert, and R. Ingold, "Convolutional neural networks for page segmentation of historical document images," in *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 01, November 2017, pp. 965–970.

[2] M. Bulacu, R. van Koert, L. Schomaker, and T. van der Zant, "Layout analysis of handwritten historical documents for searching the archive of the cabinet of the dutch queen," in *9th International Conference on Document Analysis and Recognition (ICDAR)*, 2007, pp. 357–361.

[3] R. Cohen, A. Asi, K. Kedem, J. El-Sana, and I. Dinstein, "Robust text and drawing segmentation algorithm for historical documents," in *2nd International Workshop on Historical Document Imaging and Processing (HIP)*, 2013, pp. 110–117.

[2]https://www.snf.org/en/

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

[4] A. Asi, R. Cohen, K. Kedem, J. El-Sana, and I. Dinstein, "A coarse-to-fine approach for layout analysis of ancient manuscripts," in *14th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2014, pp. 140–145.

[5] B. Gatos, G. Louloudis, and N. Stamatopoulos, "Segmentation of historical handwritten documents into text zones and text lines," in *14th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2014, pp. 464–469.

[6] K. Chen, C. L. Liu, M. Seuret, M. Liwicki, J. Hennebert, and R. Ingold, "Page segmentation for historical document images based on superpixel classification with unsupervised feature learning," in *12th IAPR Workshop on Document Analysis Systems (DAS)*, April 2016, pp. 299–304.

[7] S. S. Bukhari, T. M. Breuel, A. Asi, and J. El-Sana, "Layout analysis for arabic historical document images using machine learning," in *13th International Conference on Frontiers in Handwriting Recognition (ICHFR)*, 2012, pp. 639–644.

[8] K. Chen, M. Seuret, M. Liwicki, J. Hennebert, C.-L. Liu, and R. Ingold, "Page segmentation for historical handwritten document images using conditional random fields," in *15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2016, pp. 90–95.

[9] K. V. Jobin and C. V. Jawahar, "Document image segmentation using deep features," in *6th National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*, December 2017.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[12] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell. (T-PAMI)*, vol. 39, no. 4, pp. 640–651, 2017.

[13] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell. (T-PAMI)*, vol. 39, no. 12, pp. 2481–2495, 2017.

[14] Y. Xu, W. He, F. Yin, and C. L. Liu, "Page segmentation for historical handwritten documents using fully convolutional networks," in *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 01, November 2017, pp. 541–546.

[15] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, 1998, pp. 2278–2324.

[16] S. Nicolas, T. Paquet, and L. Heutte, "Complex handwritten page segmentation using contextual models," in *Second International Workshop on Document Image Analysis for Libraries (DIAL)*, April 2006, pp. 46–59.

[17] K. Chen, M. Seuret, M. Liwicki, J. Hennebert, and R. Ingold, "Page segmentation of historical document images with convolutional autoencoders," in *13th International Conference on Document Analysis and Recognition (ICDAR)*, August 2015, pp. 1101–1105.

[18] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?" in *12th IEEE International Conference on Computer Vision*, September 2009, pp. 2146–2153.

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML'15, 2015, pp. 448–456.

[20] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[21] K. Chen, M. Seuret, H. Wei, M. Liwicki, J. Hennebert, and R. Ingold, "Ground truth model, tool, and dataset for layout analysis of historical documents," in *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 9402, February 2015.

[22] A. Garz, M. Seuret, F. Simistira, A. Fischer, and R. Ingold, "Creating ground truth for historical manuscripts with document graphs and scribbling interaction," in *Proc. 12th Int. Workshop on Document Analysis Systems (DAS)*, April 2016, pp. 126–131.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.