Handwritten text line segmentation using Fully Convolutional Network

Guillaume Renton*, Clément Chatelain*, Sébastien Adam*, Christopher Kermorvant*†, Thierry Paquet* ,
*Normandie Univ, UNIROUEN, UNIHAVRE, INSA Rouen, LITIS, 76000 Rouen, France,
†TEKLIA SAS, Paris, France

***Abstract*—In this paper, we propose a learning based method for handwritten text line segmentation in document images. The originality of our approach rely on i) the use of X-height labeling of the textline, which provides a suitable text line representation for text recognition, and ii) a variant of deep Fully Convolutional Network (FCN) based on dilated convolutions. Results are given on a public dataset and compare favorably to a standard handmade segmentation approach.**

***Index Terms*—Fully Convolutional Networks, line segmentation, Dilated Convolutions, Document Layout Analysis**

I. Introduction

Text line detection is a crucial issue of Document Layout Analysis, with potential strong impact on further text recognition performance. If this task is quite trivial in the case of printed documents, it becomes more difficult with handwritten documents mainly because handwritten lines are generally not perfectly straight. Other difficulties often occur such as the presence of connectivities between lines, the irregularity of handwritten words and characters and the intrinsic high variability of handwriting. Moreover, historical documents show even more difficulties, due to the low quality of the documents (see Fig. 1).

Line detection raises the problem of defining what is a text line. In the literature, one can observe that text lines are defined either as their baseline [1], as their bounding box [2], as the set of pixels corresponding to their handwritten components [3], or as the area corresponding to the core of the text without ascenders and descenders, also called X-Height [3]. Figure 2 shows X-height definition.

This later definition is more interesting since others representations (bounding boxes, baseline, pixels) can be recovered from X-height. It also prevents from overlapping lines, as it can be the case with bounding boxes.

In this paper, we propose a new learning-based approach for text line segmentation that relies on a deep, fully convolutional neural network. Our network has been trained with X-height labeling on different databases and obtained promising results.

This paper is structured as follow : section II describes the related works. Our approach is explained in section III and experiments are detailed in section IV.

II. Related Works

As shown in the recent and very complete survey [4], a lot of text line segmentation methods rely on algorithms which have to be hand-tuned. This tuning is a tedious task and is generally dataset-dependant. On the other side, it has be shown in many application domain that deep learning approaches [5] can provide better results than handmade algorithms. For the text line detection problem, the only works using deep neural networks are the different contributions of Moysset and al. [2], [6], [7], [8] which propose a combination of Multi Dimensional Long short Term Memory (MDLSTM) neural network combined with convolutional layers to predict a bounding box around the line. Those methods give very good results, but are limited to horizontal lines. Moreover, those

This work has been supported by the French National grant ANR 16-LCV2-0004-01 Labcom INKS

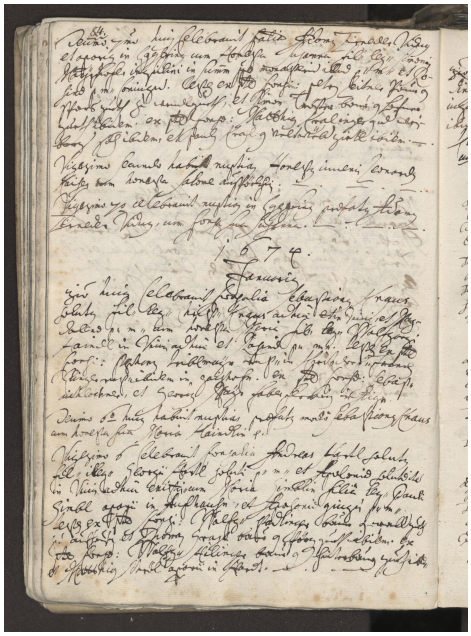

Fig. 1. Example of low quality historical document

Fig. 2. Different representations of text line

networks are pretty heavy to train.

In this work, as we have retain the X-height representation for a text line, every pixel of the document image has to be labeled as belonging to text line or not. Therefore, the text detection problem can be viewed as a semantic segmentation problem, for which interesting deep learning approaches have been proposed these last years.

In semantic segmentation, one recent interesting method is the *Fully Convolutional Network* [9], [10] (FCN). A FCN is a Convolutional Neural Networks (CNN) whose dense layers have been removed, making them able to process images from variable size. They firstly have been proposed by Long and al. in [9]. The idea behind fully convolutional networks is that a standard convolutional neural network can not take a decision for each pixel, because of the dense layers who can not keep the spatial information in the output. Thus, a fully convolutional network works as an encoder and decoder, where the encoder corresponds to the CNN without dense layers, and the decoder is an additional part which is used to build an output with the same resolution as the input.

To build the decoding part, 3 main methods have been proposed :

The deconvolution is the original method used by Long and al. [9]. It uses a convolution filter applied with a stride equals to $\frac{1}{f}$, where f is the up-sampling factor. This method has been used by [11], [10] for semantic segmentation but also in text lines segmentation by [3].

The unpooling consists in keeping a memory of the winning activation during the different pooling layers to re-inject the result at those localizations. [12] use this method while [13] use both unpooling and deconvolution in their network.

The two previous methods have shown great results in semantic segmentation, but they induce one major problem: the upsampling action can sometimes be coarse, and in the case of text lines detection, it might regroup some lines together, making the recognition impossible for both lines. For this reason, we decided to use a third method, **the dilated convolution** also known as "A trous" convolution, which is described in the next section III.

III. Proposed approach

Although FCN have shown great ability to solve semantic segmentation problems, those networks widely use pooling and striding, which drastically reduce the image resolution (32 times for VGG16 for example).

A. Motivations

Our goal here is to avoid reducing the image resolution during the training and prediction, because upsampling might lead to connected lines. Thus, the main reason for resolution to decrease comes from the pooling layers. But just removing the pooling layers can not be enough, since those pooling layers are here for two reasons: at first, they reduce the number of computations made in the network, but most important, they increase the size of the filters receptive fields. Thus, removing the pooling layers will reduce the context our network is able to see.

One solution to solve this problem could be to increase our filters size. But the number of parameters in the network is squared the filters size, so simply increasing the filters size will lead to a far greatest number of parameters. For example, a 9×9 receptive field requires 81 parameters, against only 9 parameters for a 3×3 receptive field coupled with 2 pooling layers (which would results to an equivalent 9×9 receptive field). Finally, to get the same receptive fields than VGG16, the number of parameters will explode from 9 to 4225 for each filter.

An alternative solution is to use the "A trous" algorithm, proposed by [14] to use dilated convolutions. The idea is to fill the filters with zeros, to artificially increase the size of receptive fields.

B. Dilated convolutions

Let x be the input and f the weighted filter. Then the output of a standard convolution can be computed as follow (Eq. (1)):

$$y[i] = \sum_{m=1} x[i+m]f[m] \tag{1}$$

For dilated convolution, a dilation rate r is introduced, which corresponds to the scale factor of the filter following the equation (2):

$$y[i] = \sum_{m=1} x[i+r \cdot m]f[m] \tag{2}$$

As one can see, dilated convolution is a generalization of the standard convolution, which correspond to a dilation rate of 1. A visual representation of dilated convolution is given in figure 4. The "A trous" algorithm idea comes from [14] and have firstly been used with wavelet transform. It recently have been brought up to date recently in fully convolutional networks, especially by [15], [16], [17], [18], showing interesting results.

Finally, dilated convolutions provide two advantages: first the size of the receptive fields can be controlled without reducing the resolution nor increasing the number of parameters, and secondly it also allows to reduce the number of parameters and the depth of our network, since deconvolution and unpooling induce a deeper and larger network, due to the decoder part which have to be learned. The drawback of this method is the number of computations, which increases because of the high resolution.

C. Network architecture

For our network, we decided to use a 7 layers architecture: the 2 first layers correspond to standard convolutions, with a dilation 1. Then two layers with dilation 2 and two layers with dilation 4. Those dilation rates are used to replace poolings layers, in order to keep the same receptive fields than after a 2×2 pooling layer. Those 6 layers are pretty similar to the VGG16 6 first layers. Finally, a last convolution layer is added for prediction, with dilation 1 and filters size 1. The idea behind those dilations is the fact that text line detection does not require large context to be efficient. Our network architecture is presented on figure 3.

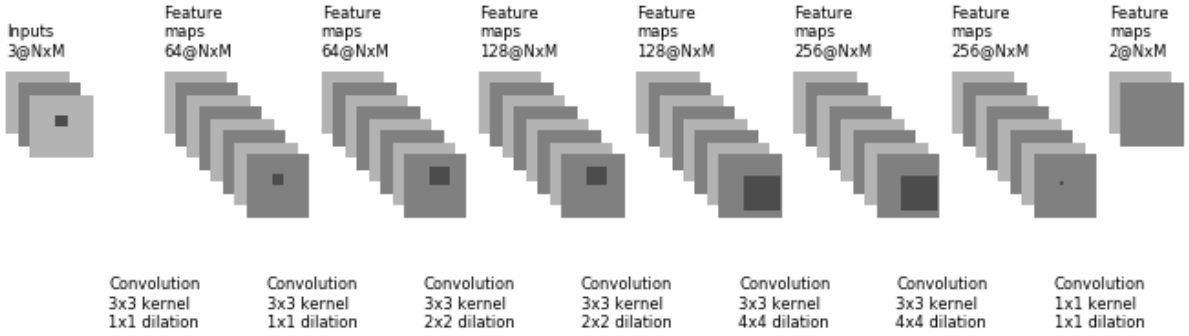

Fig. 3. Our network architecture : the input resolution is never changed and the receptive fields are increased due to the dilation

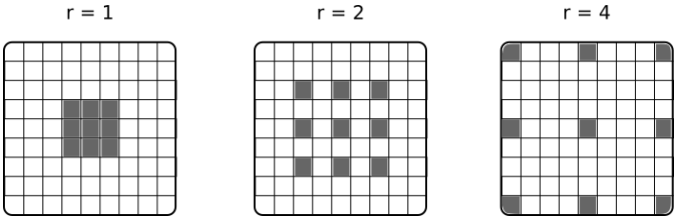

Fig. 4. Visual representation of dilated convolution

IV. Experiments

A. Data preparation

One advantage with fully convolutional networks is that they don't require fixed images size, since only convolutions are used. But some constraints still exist for those images. First, working with large images causes two problems: the number of computations increases drastically, increasing in the same way the training time. Also, the GPU memory is not enough to process whole large images. The second limitation comes from the variation of the images sizes. If this variation is too high, the network will tend to over consider biggest images over little images, due to the loss computation.

To take care of those constraint, we decided to reshape the image as follows: the largest side of the image is reduced to 608 pixels, and the other side is reduced to keep the same ratio between height and width.

B. Datasets

Our method has been tested on two different datasets. The first one is private, while the second is the dataset provided for the baseline detection competition at ICDAR 2017 (cBAD [1]), which is the successor of the ICDAR 2015 ANDAR text lines competition.

The private dataset is made of 1600 images of high resolution (3000×4000) with 1500 images for training and 100 for validation. The dataset is composed with approximately 10000 lines and is a mobile semi-synthetic dataset : real handwritten text lines from the RIMES [19] database are pasted onto real real backgrounds captured with different mobile phones. The purpose of this dataset is twofold : test the robustness of the method with a mobile capture and different backgrounds while easily obtain the location of the text lines for the evaluation. To generate this dataset, we proceeded as follow : text lines are extracted with binarization, then those lines are pasted at a random position on a random background, and finally we add a gaussian blur and compress them. The X-height of each line has been labeled at the pixel level.

The cBAD dataset is formed of 755 pages of handwritten archival documents, with 216 in training and 539 in test with high resolution too (3000×4000), but high variation in the quality. We then divided the training base in 2, with 176 remaining in training, and 40 in validation. Here also, the X-height labeling is provided for each line.

C. Training

A FCN architecture has been trained on both datasets, using Keras [20]. The training criterion is the pixel accuracy. We used a 10^{-5} learning rate with a stochastic gradient descent. Due to the variable size of our images, we did not used mini batch but worked in an online way. The convergence of the models has been respectively controlled on the validation datasets to prevent overfitting. As the networks output probability heat maps, one need a post processing in order to identify the text lines. For that, we simply choose the maximum value between line and non-line probabilities[2].

The systems are evaluated using classical Recall/Precision and F-measure criteria. We also provide the mIoU measure, computed as follows:

$$mIoU = \frac{A \cap B}{A \cup B}$$

where A is the prediction image and B is the ground truth image.

[1] https://scriptnet.iit.demokritos.gr/competitions/5/

[2] Note that the "maximum" operator is equivalent to a 0.5 threshold, which could be tuned to optimize the performance. We did not proceed to this optimization.

For the cBAD competition, the metrics are not computed at the pixel level, but regarding the position of the baseline (See [1] for details). As our approach outputs the text X-height, we extracted the baseline from the predicted core text to match the evaluation requirement.

D. Results

Since the first dataset is private, we only provide results for qualitative purpose. A visual example is given in Fig 5. However the dataset is rather easy, and does not allow to really evaluate the efficiency of the approach on difficult real-world documents.

Therefore, we decided to participate to the challenging cBAD competition. To compensate for the rather small training dataset (176 images), our architecture has been pre-trained on our private dataset (9000 images), and then fine tuned on the cBAD training dataset.

TABLE I
RESULTS OBTAINED AT THE CBAD COMPETITION USING THE COMPETITION METRICS.

	F-measure	Precision	Recall	mIoU
FCN (our approach)	0.75	0.66	0.86	0.93
Steerable filters	0.408	0.407	0.409	

Table I shows the obtained results on the cBAD testing dataset. We compare our method with the classical steerable filter approach [21], an ad-hoc method that provided good results for text line detection competition at ICDAR 2015. The results of other participants at cBAD competition will be added to the paper as soon as they are available. As one can see, our FCN approach provides a good mIoU value of 93%. Regarding recall and precision values, FCN method clearly outperforms the steerable filters approaches. Note that steerable filters predict bounding boxes and not baseline, so we picked the bottom side of the bounding box, which can be disadvantaging for the steerable filters.

Fig. 6 shows an example of a document where lines have been extracted using steerable filters and our FCN approach. Note that for the sake of visualization, the bounding boxes have been extracted for both approaches. One can observe that our FCN approach provides thinner boxes, and that it yields less under segmentation.

V. CONCLUSION

In this paper, we have presented a novel approach based on fully convolutional networks and dilated convolutions for text lines detection in handwritten documents. The approach is generic and provides very fast decision process since a whole document is segmented within a single network forward. Moreover, it shows competitive results on two datasets.

REFERENCES

[1] T. Grüning, R. Labahn, M. Diem, F. Kleber, and S. Fiel, "Read-bad: A new dataset and evaluation scheme for baseline detection in archival documents," *arXiv preprint arXiv:1705.03311*, 2017.

[2] B. Moysset, C. Kermorvant, C. Wolf, and J. Louradour, "Paragraph text segmentation into lines with recurrent neural networks," in *International Conference on Document Analysis and Recognition*. IEEE, 2015, pp. 456–460.

[3] Q. N. Vo and G. Lee, "Dense prediction for text line segmentation in handwritten document images," in *IEEE International Conference on Image Processing*. IEEE, 2016, pp. 3264–3268.

[4] S. Eskenazi, P. Gomez-Krämer, and J.-M. Ogier, "A comprehensive survey of mostly textual document segmentation algorithms since 2008," *Pattern Recognition*, vol. 64, pp. 1–14, 2017.

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[6] B. Moysset, J. Louradour, C. Kermorvant, and C. Wolf, "Learning text-line localization with shared and local regression neural networks," in *International Conference on Frontiers in Handwriting Recognition*, 2016.

[7] B. Moysset, P. Adam, C. Wolf, and J. Louradour, "Space Displacement Localization Neural Networks to locate origin points of handwritten text lines in historical documents," in *Workshop on Historical Document Imaging and Processing*, Nancy, France, Aug. 2015. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01178342

[8] B. Moysset, C. Kermorvant, and C. Wolf, "Full-page text recognition: Learning where to start and when to stop," in *International Conference on Document Analysis and Recognition*, 2017.

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

[10] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters–improve semantic segmentation by global convolutional network," *arXiv preprint arXiv:1703.02719*, 2017.

[11] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *IEEE International Conference on Computer Vision*, 2015, pp. 1529–1537.

[12] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.

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

[14] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian, "A real-time algorithm for signal analysis with the help of the wavelet transform," in *Wavelets*. Springer, 1989, pp. 286–297.

[15] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[16] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[17] ——, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *arXiv preprint arXiv:1606.00915*, 2016.

[18] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[19] E. Grosicki and H. El-Abed, "Icdar 2011-french handwriting recognition competition," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011, pp. 1459–1463.

[20] F. Chollet, "keras," https://github.com/fchollet/keras, 2015.

[21] Z. Shi, S. Setlur, and V. Govindaraju, "A steerable directional local profile technique for extraction of handwritten arabic text lines," in *International Conference on Document Analysis and Recognition*, July 2009, pp. 176–180.

Fig. 5. A result example on our private dataset.

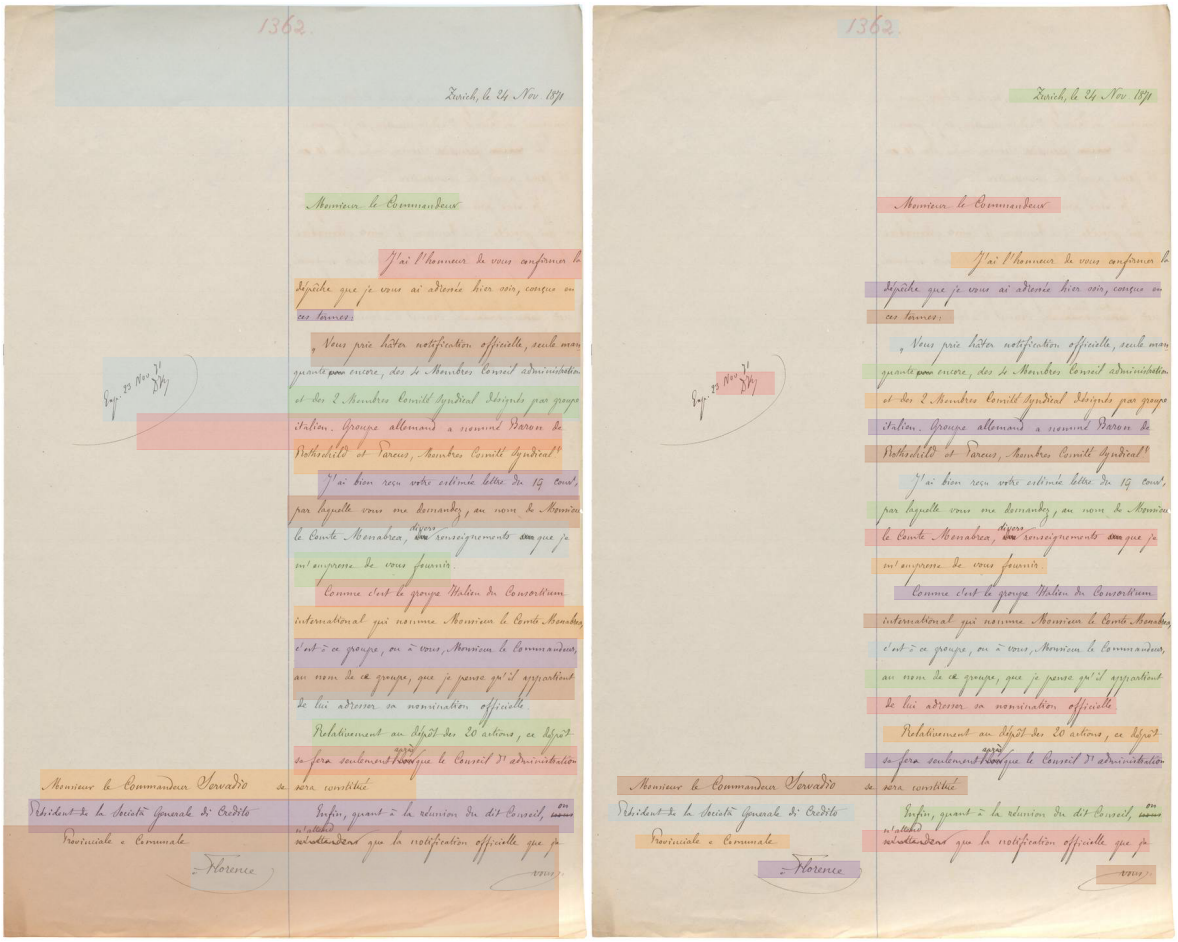

Fig. 6. Comparison between steerable filters (left) and FCN (right) on an image from the cBAD competition (better in color).