2017 14th IAPR International Conference on Document Analysis and Recognition

Semantic Text Detection in Born-digital Images via Fully Convolutional Networks

Nibal Nayef and Jean-Marc Ogier
L3i Laboratory, Université de La Rochelle, France
{nibal.nayef, jean-marc.ogier}@univ-lr.fr

***Abstract*—Traditional layout analysis methods cannot be easily adapted to born-digital images which carry properties from both regular document images and natural scene images. One layout approach for analyzing born-digital images is to separate the text layer from the graphics layer before further analyzing any of them. In this paper, we propose a method for detecting text regions in such images by casting the detection problem as a semantic object segmentation problem. The text classification is done in a holistic approach using fully convolutional networks where the full image is fed as input to the network and the output is a pixel heat map of the same input image size. This solves the problem of low resolution images, and the variability of text scale within one image. It also eliminates the need for finding interest points, candidate text locations or low level components. The experimental evaluation of our method on the ICDAR 2013 dataset shows that our method outperforms state-of-the-art methods. The detected text regions also allow flexibility to later apply methods for finding text components at character, word or textline levels in different orientations.**

I. Introduction

There is a huge growth in the amount of multimedia data on social network media such as advertisements, holiday pictures, business cards, magazines. This has led to large data collections of heterogeneous and weakly structured content. Two popular types of images in network media are scene images with embedded text and born-digital images. Analyzing the latter class of images has not received as much attention as scene text images despite its importance. Analyzing the contents of such images is very challenging because of cluttered background, complex layout with mixed graphics and text, low resolution, variations of font type, size and color and oriented and multi-lingual text.

The textual information present in the text layer of born-digital images carries rich and precise high level semantics that would improve mining and retrieval of the web content, and would be useful in a variety of social and commercial applications. Most of prior methods for detecting or segmenting the text layer in scene or born-digital images have followed a bottom-up approach, by trying to find low level text components using local interest points (MSER), connected components or sliding windows [1], [2]. The complex background, low resolution and the variability in text scale that characterize born-digital images, make such methods not very effective in localizing all the text. Additionally, after finding initial text interest candidates, some of them may be lost in the subsequent steps of classification and grouping.

Another technical challenge that faces text detection methods is: designing text features that would achieve high accuracy in text component classification. Many methods have followed the traditional approach of using hand-crafted features for training text classifiers [3], [4]. However, such features are usually database-dependent and their computation can be highly time-consuming. Hence, the more recent approaches in text detection (mostly in scene images) have turned to the approach of automatic learning of features via deep learning techniques. However, the problem of finding initial relevant and meaningful low- or high-level text components remains challenging, specially with multi-oriented and multi-scale text which appears in advertisements and other digitally-born images in the web.

In order to overcome all of the above mentioned challenges, researchers have – very recently – turned to holistic deep learning approaches, i.e. proposal-free text detection [5], [6], [7]. The "proposals" refer to the initial text candidates or regions of interest before the learning and classification steps. In a proposal-free approach, the whole image as a pixel map is fed to a deep network, then after training, the network is able to output a corresponding heat map with classification labels of image pixels. This approach has been applied to scene text images and yielded superior performance compared to state-of-the-art methods.

In this work we have opted to follow the holistic (candidate- / proposal-free) approach for detecting text regions in born-digital images. In particular, we use Fully Convolutional Networks (FCN) [8] as a main component in our method. FCN have been developed and used very recently in computer vision for semantic segmentation, edge detection and a variety of problems, due to their holistic feature of taking full images as input.

Our method is composed of two modules: (1) The FCN-based module is trained in a holistic manner, and it generates semi-final candidate text regions through labeling text pixels, (2) The text layer formation module involves projection-based segmentation that splits the falsely connected text regions. The method is simple and generic, where the FCN module directly outputs highly accurate text regions at pixel level. Hence, our method works directly at pixel level of colored images, and gives each pixel a semantic label as text or non-text. This method has proven to be very effective for detecting text in both low- and high-resolution born-digital images.

The rest of this paper is structured as follows. Prior related work is reviewed in the next section, the description of the proposed method is detailed in Section III, and the experimental evaluation is presented in Section IV. The conclusions are discussed in the last section.

2379-2140/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDAR.2017.145

II. Related Work

Text detection and segmentation in scene images and born-digital image have been hot research topics in the recent years, with the deep learning-based approaches becoming prominent in the last two years. An excellent comprehensive survey for this topic along with its scientific issues and challenges can be found in [3]. This survey focuses on the classical approaches which are not based on deep learning. We note that much fewer works have focused on born-digital images [4], [9], [10], [11]. Those works – for born-digital images – all follow a traditional non deep learning-based approaches, and they rely on finding candidate (initial) components or interest regions as a first step.

The typical approach in deep learning-based methods is to find candidate (initial) components or interest regions, then classifying those candidates using convolutional neural networks (CNN) -based classifiers. Finding the initial candidates is done using MSER [1], CE-MSER [12], Sliding windows [2] or Edge-boxes [13]. This step can also be based on deep learning such as in the Region Proposal Network (RPN) [14]. Then, the CNN-based classifiers assign the candidates text or non-text labels, character labels etc. After this stage, a refinement step takes place where words / text-lines are formed using one of various methods: grouping by visual and geometry features, context features, graph-based and also deep learning-based [15].

In contrast to the typical approach, very few and very recent methods have followed a holistic approach so that to eliminate the need for finding initial text proposals. We review here those few methods as our method falls in this category. Zhang et al. [6] have proposed a method for multi-oriented scene text detection using two FCN networks. The first FCN takes full input images and outputs pixel heat map with text classification probabilities. They train another FCN network to predict character centroids as a refinement step of the output of the first FCN. After that, they form text-lines using many steps which also handle finding text orientation.

He et al. [5] have presented Cascaded Convolutional Text Network (CCTN) for localizing text accurately in natural scene images. Their method is composed of two cascaded steps: (1): Coarse text region detection with a VGG-based network modified to handle multi-oriented and multi-scale text with parallel convolutional layers that have different kernels. This step is also holistic, it takes a full input image and outputs a pixel-level heat map. (2): Fine text-line detection with another VGG-based network: it takes as input the cropped rough regions which resulted from step (1), and outputs pixel maps. In the ground truth images, the central textline area of text bounding boxes is labeled as positive.

Yao et al. [7] have designed an interesting method called holistic, multi-channel prediction for scene text detection. Their work is based on a variant of FCN called the HED framework, where the input is a full image with a 3-side output layer: two pixel-wise prediction maps for text regions and individual characters regions, and a map of orientation values for linking between characters. The sub-networks are trained independently, then the outputs are fused together.

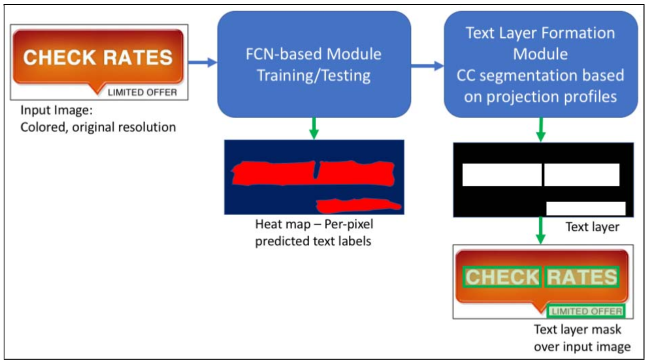

Fig. 1. Block diagram of the proposed method. The FCN module outputs a heat map in which the red pixels correspond to the positive text class. The text layer formation module splits the falsely connected detected text regions. The "text layer" output shows its comprising text regions in white color. The bottom part shows the detected text layer on top of the input test image.

These FCN-based methods have achieved the best of state-of-the-art results, specially regarding detecting oriented scene text. However, they all have a complex structure of more than one FCN network, and still face difficulties with the huge variability of text scale. The recent results show that the holistic approach is the most promising to follow and adapt for analyzing born-digital images.

III. The Holistic Text Layer Separation Approach

In this section, we present details of the proposed approach. In the training phase, colored images in their full size are fed as input to the main FCN module with their corresponding ground truth represented as binary label maps. We deal with detecting text regions as pixel-wise classification, and cast this classification as a binary semantic segmentation problem, where we have two classes with the labels: text or non-text at pixel-level.

The FCN-based network is built and trained to output the so called heat map (see Subsection III-A). This is a label image with confidence scores for each pixel being a text or non-text pixel. In the test phase, an image is fed to the trained FCN-based network, and we a get an output heat map. The second module performs connected component analysis on the output label map and uses projection profiles to split some of the resulting text regions into words and text-lines if needed. The final regions comprise the text layer of a particular test image.

Figure 1 shows a block diagram of the composing parts of our approach. The FCN-based module is trained in a holistic manner, and it generates semi-final candidate text regions through labeling text pixels. The text layer formation module involves projection-based segmentation that splits the falsely connected text regions. The method is simple and generic, where the FCN module directly outputs highly accurate text regions at pixel level. We will show how the FCN-based network is successfully trained and tested on multi-scale text. Additionally, this module could be trained on multi-oriented and multi-lingual text, because we pose no assumptions or preprocessing steps that are script dependent or orientation dependent.

This holistic FCN-based training overcomes the limitations of the approaches which rely on character or other text component detection. Such approaches cannot robustly

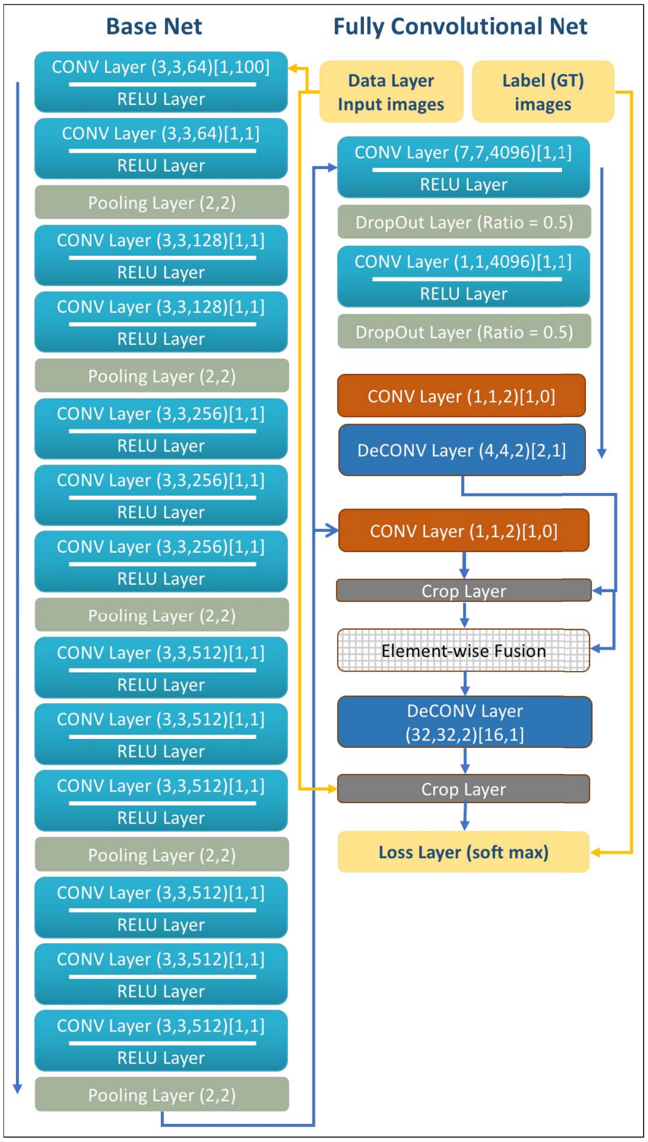

Fig. 2. The detailed architecture of the FCN-based network used in the first module of our proposed method. The architecture shows how to recover a pixel-wise prediction map that corresponds to the size of the input image through deconvolution layers. The parameters of CONV and DeCONV layers are shown as (kernel size, kernel size, num. outputs)[stride, padding]. The parameters of the POOL layers are shown as (kernel size, stride).

discriminate the large variety of text components from the complex background. Our FCN-based module learns both local and global text related features around a text pixel. Moreover, our approach does not require complicated preprocessing steps for preparing input images, but at the same time, it allows for any post-processing to be applied on the resulting generic text regions within the text layer. Any subsequent layout analysis steps could be performed on the detected text layer.

A. The FCN-based Text Region Detection

Our FCN-based module is built by modifying and adapting the architecture of the fully convolutional models for semantic segmentation of natural objects proposed recently by Shelhamer et al. [8]. The latter network was also built from the famous 16-layer VGGnet where the fully connected layers are converted to 1x1 convolutional layers to preserve spatial information.

The network design in our work is customized for text detection as shown in Figure 2 and as follows. The base net is similar to the FCN described in [8] and is composed of 13 convolutional (CONV) layers, and each of them is followed by a RELU layer. The CONV layers are separated by 5 max-pooling (POOL) layers arranged as seen in Figure 2: 1 POOL layer after the first 2 CONV-RELU layers, 1 POOL layer after the second CONV-RELU layers, and then 1 POOL after each 3 consecutive CONV-RELU layers. The first CONV layer takes image data as input.

As seen in the design in Figure 2, the network can handle any input image size of [widthxheightx3]. This input holds the raw pixel values of the image, in this case an image of width, height and three color channels R,G,B. The 3D mean is computed across the training images. The mean is subtracted from each image (training or test image) before it is fed to the network.

After the base net comes the fully convolutional part of the network where 2 fully connected layers are converted to (replaced by) equivalent 2 fully convolutional layers. As known, the fully connected layers produce non-spatial outputs, but in FCN networks, the fully connected layers are handled as convolutions with kernels that cover their entire input regions. Hence, they become capable of taking an input image of any size and output per-pixel classification maps. Each of the two fully convolutional layers is followed by a RELU layer and a dropout layer for better learning generalization.

Another CONV layer takes its input from the last dropout layer and has 2 outputs that correspond to the number of classes (text versus non-text). The first output map is produced by a deconvolution layer (DeCONV) that follows the previous layer with a kernel size of 4 and a stride of 2. This produces an output of the size of the input image. The deconvolution here is in effect an upsampling process to get back the spatial pixel locations of the image. With the proper kernel and stride sizes, the CONV output is converted to a spatial map.

In this DeCONV layer, if we used the parameters $(64, 64, 2)[32, 1]$, we could stop here and get the desired output. However, due to the large stride size of 32, we get a coarse output (the resulting text region or text pixel classification output is not accurate enough). Hence, as designed in Shelhamer et al. [8], we also use the outputs from previous layers to get a finer output as follows.

The output of the last POOL layer, is fed to a CONV layer, and then via a crop layer and a fusion function, the output of the CONV layer is fused with output of the DeCONV layer. The output map is obtained by a second DeCONV layer of a kernel size 32 and a stride 16 to recover the spatial predicted image. A crop layer is used to map the output of the convolutional layer to the input image data layer. Then the loss is computed by a Softmax layer which compares the ground truth label image to the segmentation output of the crop layer.

Overall, this design allows the network to handle variations in text scale, and in general the low resolution born-digital images. The final result of this part of our approach is a per-pixel heat-map that indicates the probability of each pixel being a text or non-text.

B. FCN-based Model Training

In the training phase, the FCN network described above is initialized as follows. The weights till the last Dropout layer are initially copied from the pre-trained model of VGG16 network as done in fine tuning training. The layers after that are all related to our text/non-text classification problem, so they are initialized with new weights by Xavier weights initialization.

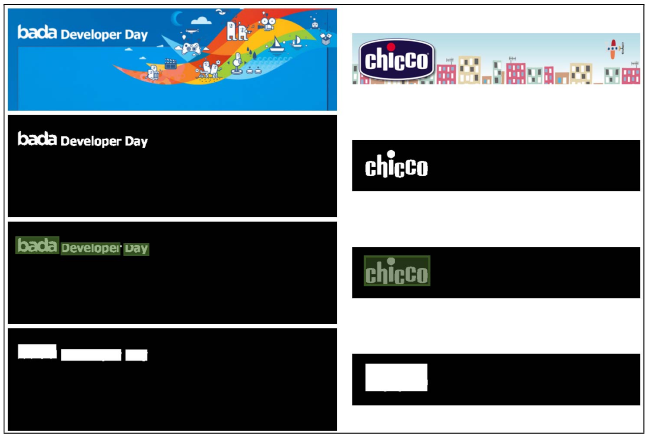

Fig. 3. Ground truth preparation. The first row shows the original images. The second row shows the pixel-level segmentation ground truth provided in the dataset. The third row shows the first variant of the ground truth preparation used in our method, where the pixels inside the green bounding boxes except text pixels are labeled as ignored. The fourth row shows the second variant where all the pixels inside the text bounding box is labeled as text.

The training process is a supervised learning one, the loss layer takes the ground truth label images to compute the loss. The ground truth is prepared in two variants as follows. In the first variant: text pixels are labeled as positive (text class), the pixels inside the bounding box of text words are labeled as ”ignored” and background pixels are labeled as negative (non-text class). In the second variant, both the text pixels and the pixels inside the bounding box of text words are labeled as positive (text class), and the rest is labeled as negative. Figure 3 shows example ground truth images of the two variants with their corresponding input images. In the second variant, a text region is considered as an object, which provides strong semantic information from the local context around the actual text pixels.

The solver parameters for carrying out the training and testing processes are detailed in Subsection IV-B. Additionally, we augmented the dataset to achieve a more robust training and prediction results. The network gives dense prediction based on the deconvolution (upsampling) and the pixel-wise loss.

In this way, our FCN-based network transforms the original image layer by layer from the original pixel values to the final class scores. The deep layers containing the learned feature maps carry knowledge about the text including low level properties and also about the local context of text components. The trained FCN is a strong text labeling model that does not make any assumptions about the orientation of the text or its script. This makes such a model applicable for analyzing any content of born-digital images.

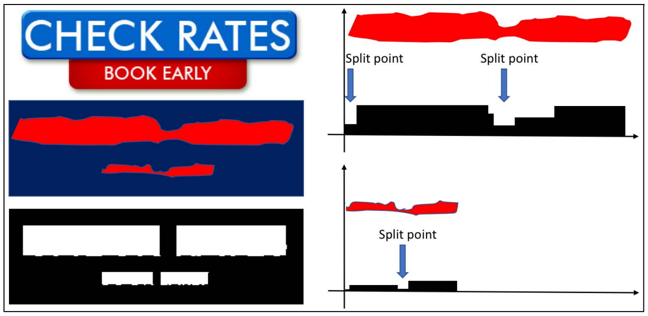

Fig. 4. The text layer formation module of the proposed method. At the top left: the original input image. The heat map resulted from the FCN module is shown second. The connected components of the heat map are analyzed via the projection profiles shown in the right column. After splitting the components at split points, we get the final output of text regions in the left bottom image.

C. Text Layer Formation

The output from the previous module of our method is a heat map that contains predicted pixel-wise text labels. The union of all those pixels comprises the text layer. The FCN network yields roughly accurate text regions that correspond to words or text lines of the input image. Even in the ground truth, the words may appear very close to each other within a textline due to the low resolution of a born-digital image.

Our goal is to separate and extract all the text in a text layer without necessarily segmenting the content into words. The complex layout of born-digital images does not allow for traditional layout analysis methods to find textlines. Once the text layer is segmented out, a post-processing method can be applied to achieve any desired level of text granularity.

In order to further enhance the text regions, we convert the heat map to a binary label image, and perform connected component analysis on it. The obtained components correspond to text regions at different granularity. For each component, we get the corresponding heat map region, and analyze its projection profile parallel to the largest principal axis. We split the region into two or more sub-regions at the low points (minima) of the projection profile. Figure 4 shows this steps of this process and the final text regions. The effect of this is getting more accurate text regions, for example splitting two attached lines or words into two separate lines or two words respectively. The union of those final regions forms the text layer of an image.

IV. Experimental Evaluation

We have implemented our method in a text classification system, and evaluated this system on a standard public database for born-digital images. In the following subsections we detail the settings of our experiments and analyze the results.

A. Database and Evaluation Metrics

We evaluated the performance of our method on the ICDAR 2013 Robust Reading Competition Challenge 1 on born-digital images [16]. Images that were originally synthesized on computers such as advertisements and web images with embedded text are called born-digital images. Many content authors choose to embed text in images, rather than encoding it explicitly in the electronic document. This hinders the process of web information retrieval [16].

The database contains 551 in total, and according to the competition settings, 410 images are assigned for training and 141 images for testing. The minimum resolution in this database is 100x100 pixels. Born-digital images are generally characterized by their low resolution such as the images embedded in webpages and email messages. However, some of them have high resolution such as advertisement images.

As for evaluation metrics, we use the standard recall, precision and f-measure metrics as done in most scene text detection works, and also in RRC competitions [16]. In such evaluation, a text word (or bounding box) detection is considered as a correct detection if the overlap ratio between the detected box and the ground truth bounding box is ≥ 0.5.

B. Experimental Setup

The train/test split of the used database is as explained above (410/141 images). The FCN-based network solving parameters are as follows. The loss is averaged each 20 iterations, and learning rate is fixed during training where are base learning rate is $1e^{-10}$. The FCN networks use a high momentum of 0.99. The iteration size is set to 1 so that to prevent gradient accumulation. The weight decay is set to 0.0005. Finally, the maximum training iterations is set to 100000.

For both the training and testing phases, the input images are fed to the network in their original variable resolutions. Interpolation surgery is applied on the upsampling (DeCONV) layers to yield heat maps that correspond to the image size. This is applied as explained in [8].

C. Results and Analysis

Two experiments are carried out while varying how the area around text pixels is labeled. As we mentioned in Subsection III-B, two variants of the ground truth are prepared. The second labeling variant – where all pixels within the text bounding box are labeled as positive – gives better results because (1) the text and non-text classes will be more balanced due to the increase in positive samples, and (2) the pixels around the actual text pixels carry local contextual cues about the text, and it is hard to distinguish them from the actual text pixels.

The resulting text label map which contains semi-final text regions is fed to the post-processing module to output the final text regions which comprise the text layer. Example results are shown in Figure 5. The figure shows that the detected text regions are mostly accurate and correspond to words or textlines in the test images. The text varies in scale in the images, but our method is capable of detecting almost all the text pixels.

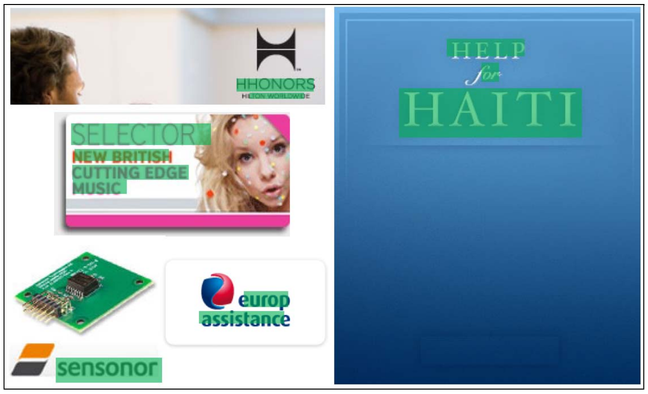

Fig. 5. Example text detection results of the proposed method. The images are 5 test images of different resolutions from the born-digital images dataset in [16]. The detection results are shown in green boxes. The green regions are in most of the cases accurate and cover words or textlines of the text.

Those final text regions are evaluated compared to the ground truth using the evaluation metrics mentioned in Subsection IV-A. Table I shows the results. As can be concluded, the method achieves high accuracy in text pixel classification despite the much larger number of pixels that belong to background. This is proven through the values of both recall and precision related to the number of correct detections of text boxes.

TABLE I. TEXT DETECTION ACCURACIES OF THE PROPOSED METHOD COMPARED TO STATE-OF-THE-ART METHODS ON THE DATASET OF BORN-DIGITAL IMAGES IN [16].

Metric / Method	Recall	Precision	F-measure
Proposed	88.38%	91.87%	**90.09%**
Pal-DAS16 [10]	87.95%	91.14%	89.51%
Pal-ICDAR15 [4]	85.44%	93.91%	89.47%
Sams [11]	89.40%	88.83%	89.11%
USTB-TexStar [17]	82.38%	93.83%	87.74%

The highest achieved results in published state-of-the-art methods – for the task of text detection in born-digital images – is 89.51% using the standard f-measure according to Chen et al. [10]. The latest results for the same task can also be found in works prior to Chen et al. in 2016 [10], such as the other work of Chen et al. in 2015 [4] and in the RRC2013 competition report [16].

As can be seen in Table I, our method performs competitively to- or better than the best state-of-the-art methods in terms of the three evaluation metrics. In our method, we do not process the detected text lines to get words, whereas the evaluation is actually done at word level and so are the reported f-measure results of state-of-the-art methods. Hence, the evaluation tool penalizes the text box that contains more than one word. In some cases (as can be seen in Figure 5), our method yields such text boxes. This could be improved by applying a post-processing or grouping methods for splitting/merging the detected text boxes into words. We argue that this would achieve even better performance of our method.

V. Conclusions and Future Directions

This paper has presented a method for separating the text layer in born-digital images in a simple holistic approach. The text pixels classification is cast as a semantic segmentation problem and is carried out using an FCN-based network. The final text regions are formed by splitting the less dense regions in the projection profiles of text connected regions.

We have shown that the choice of FCN networks fits very well the problem of detecting text of variable resolution in born-digital images. This is due to that they take input images in their full size and work directly at pixel level of colored images. As opposed to previous approaches that apply CNN classifiers on candidate text regions, our method does not require complicated preprocessing steps, and is able to directly output accurate enough text regions.

In principle, this method can be generalized to multi-oriented and multi-lingual text by extending the training data and improving the training process. As a next step, we will work on detecting multi-oriented text by both improving the architectural design on the FCN network and applying post-processing steps to compute the orientation of text regions. In another direction, we would like to investigate using bounding box regression as a post-processing step to output highly accurate text regions that can be directly used for text recognition.

VI. Acknowledgments

This work is funded by the "Agence Nationale de la Recherche (ANR)" in France under the AUDINM project.

References

[1] W. Huang, Y. Qiao, and X. Tang, "Robust scene text detection with convolution neural network induced mser trees," in *Computer Vision –ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV*, 2014, pp. 497–511.

[2] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep features for text spotting," in *European Conference on Computer Vision*, 2014.

[3] Q. Ye and D. Doermann, "Text detection and recognition in imagery: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 37, no. 7, pp. 1480–1500, 2015.

[4] K. Chen, F. Yin, A. Hussain, and C. L. Liu, "Efficient text localization in born-digital images by local contrast-based segmentation," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015, pp. 291–295.

[5] Y. Q. Tong He, Weilin Huang and J. Yao, "Accurate text localization in natural image with cascaded convolutional text network," in *arXiv:1603.09423*, 2016.

[6] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai, "Multi-oriented text detection with fully convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4159–4167.

[7] N. S. X. Z. S. Z. Z. C. C. Yao, X. Bai, "Scene text detection via holistic, multi-channel prediction," in *arXiv:1606.09002*, 2016.

[8] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Analysis Machine Intelligence (PAMI)*, vol. 39, no. 4, pp. 640–651, 2017.

[9] S. Tehsin, A. Masood, S. Kausar, and Y. Javed, "Text localization and detection method for born-digital images," *IETE Journal of Research*, vol. 59, no. 4, pp. 343–349, 2013.

[10] K. Chen, F. Yin, and C.-L. Liu, "Effective candidate component extraction for text localization in born-digital images by combining text contours and stroke interior regions." in *Workshop on Document Analysis Systems (DAS)*, 2016, pp. 352–357.

[11] S. Tehsin, A. Masood, S. Kausar, and Y. Javed, "A caption text detection method from images/videos for efficient indexing and retrieval of multimedia data," *IJPRAI*, vol. 29, no. 1, p. 1555003, 2015.

[12] T. He, W. Huang, Y. Qiao, and J. Yao, "Text-attentional convolutional neural network for scene text detection," *Trans. Imgage Processing (TIP)*, vol. 25, no. 6, pp. 2529–2541, 2016.

[13] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *Int. Journal Computer Vision*, vol. 116, no. 1, pp. 1–20, 2016.

[14] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, *Detecting Text in Natural Image with Connectionist Text Proposal Network*, 2016, pp. 56–72.

[15] Z. Zhang, W. Shen, C. Yao, and X. Bai, "Symmetry-based text line detection in natural scenes," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2558–2567.

[16] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i. Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazàn, and L. P. de las Heras, "Icdar 2013 robust reading competition," in *Proceedings of the 2013 12th International Conference on Document Analysis and Recognition*, 2013, pp. 1484–1493.

[17] X.-C. Yin, X. Yin, K. Huang, and H.-W. Hao, "Robust text detection in natural scene images." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 970–983, 2014. [Online]. Available: http://dblp.uni-trier.de/db/journals/pami/pami36.html#YinYHH14