Handwritten text line segmentation using Fully Convolutional Network

(2017)

Guillaume Renton,Clement Chatelain,Sbastien Adam,
Christopher Kermorvant, Thierry Paquet
Resume

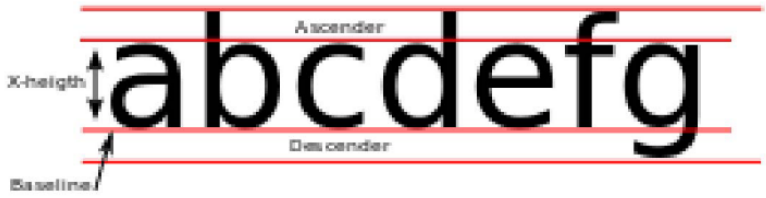
December 5, 2018

Abstract

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.

1 Introduction

Text segmentation is quite trivial for printed document, it becomes more difficult with handwritten documents, because: handwritten lines are generally not perfectly straight, the presence of connectivities between lines, the irregularity of handwritten words and characters and the intrinsic high variability of handwriting and the low quality of the documents.



Textlines Text lines are defined either as their **baseline**, as their **bounding box**, as the **set of pixels** corresponding to their handwritten components, or as the area corresponding to the core of the text without ascenders and descenders, also called **X-Height**.

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

Proposed method In this paper, the segmentation objective can be defined as labeling every pixel of the document image as belonging to text line or not. Therefore, the text detection problem can be viewed as a semantic segmentation problem, for which they propose a new learning-based approach for text line segmentation that relies on a deep, fully convolutional neural network. The network has been trained with X-height labeling on different databases.

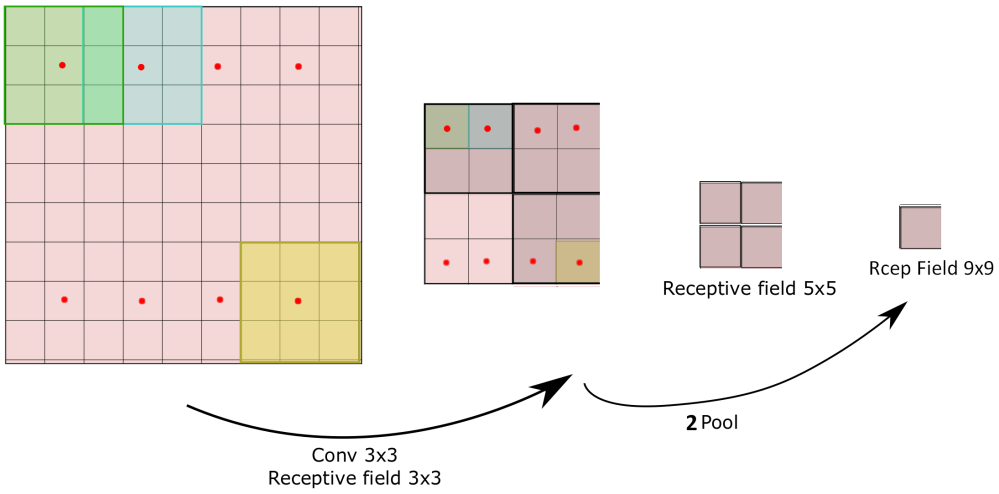
2 FCN background

A FCN is a 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]. FCN are based on CNNs that 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 FCN 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.

Decoder 3 main methods have been proposed:

- The deconvolution: It uses a convolution filter applied with a stride equals to $\frac{1}{f}$, where f is the up-sampling factor.
- The unpooling consists in keeping a memory of the winning activation during the different pooling layers to re-inject the result at those localizations.
- Given that both of the previous methods introduce 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. In this paper they use **the dilated convolution** also known as A trous convolution.

Dilated convolutions To avoid reducing the image resolution by the pooling layers during the training and prediction, because upsampling might lead to connected lines, we can remove pooling layer but given their importance for: reducing the number of computations made in the network, and most important, increasing the size of the filters receptive field. So removing them will reduce the context the network is able to see.
Another solution is increasing the filters size, but he number of parameters will increase, e.g., 9x9 filter has 9x9 receptive field with 81 parameters, but a 3x3 filter followed by 2 polling layer has the same receptive field with 9 parameters.
The solution is to use convolutions A trous, or dilated convultions.



Definition Let x be the input, f the weighted filter and m the number of parameter in the filter. Then the output of a standard convolution can be computed as follows :

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

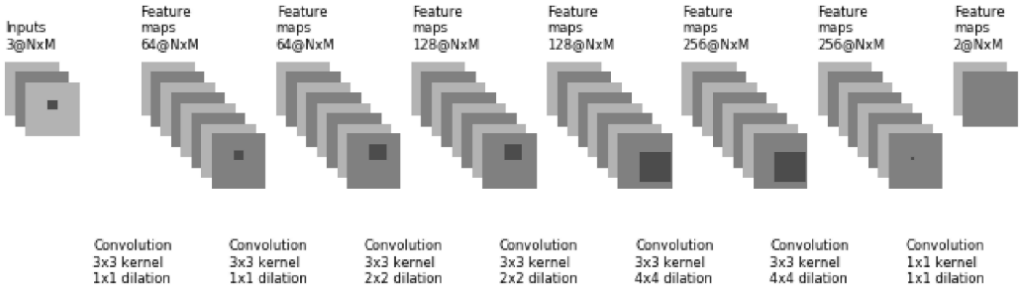
For dilated convolutions, a dilation rate r in introduced, which corresponds to the scale factor if the filter :

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

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.

3 Proposed method & Implementation details

Using 7 layer architecture



The idea behind those dilations is the fact that text line detection does not require large context to be efficien.

Data preparation Given the GPGPU memory contraints, the input images are reshaped as follows: the largest side of the image is reduced to 608 pixelsn and the other side is reduced to keep the same original ratio.

Datasets The proposed method is tested on a private dataset and on the ICDAR 2017 cBAD dataset. The model is first trained on the private dataset and then fine tuned on the cBAD dataset.

Training The used framework is keras, and the training criterion is the pixel accuracy, with a learning rate of 10^{-5} with stochastic gradient descent, in an online way without the usage of mini batch due to the variable image sizes, The systems are evaluated using classical Recall/Precision, F-measure criteria and $mIoU$.

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

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

Results

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	

References

[1] Guillaume Renton,Clement Chatelain,Sbastien Adam, Christopher Kermorvant, Thierry Paquet, *Handwritten text line segmentation using Fully Convolutional Network*