Page Segmentation of Historical Document Images with Convolutional Autoencoders
(2015)

Kai Chen ; Mathias Seuret ; Marcus Liwicki ; Jean Hennebert ; Rolf Ingold

A Resume

December 4, 2018

Abstract

an unsupervised feature learning method for page segmentation of historical handwritten documents available as color images, i.e., each pixel is classified as either periphery, background, text block, or decoration.

1 Introduction

There is an increasing need to develop robust image analysis systems to retrieve textual information from documents. Page segmentation is a prerequisite step of document image analysis and understanding. It aims at splitting a page image into regions of interest and distinguishing text blocks from other regions.

In this paper, they propose a novel page segmentation approach based on unsupervised feature learning. Each pixel is represented by a feature vector. And by training a classifier with these features, they classify each pixel into one of the four classes: *periphery, background, text block, and decoration.* Using an unsupervised feature learning method instead of carefully hard-coded features to train a classifier.

We first apply an AE on small image patches to learn low-level features. Then we take the learned feature mapping functions and convolve them with larger patches. The outputs of the AE are wired to the inputs of a successive AE to learn higher level features. And Finally, the learned features are used to train a support vector machine (SVM) to predict class labels for each pixel.

2 Method

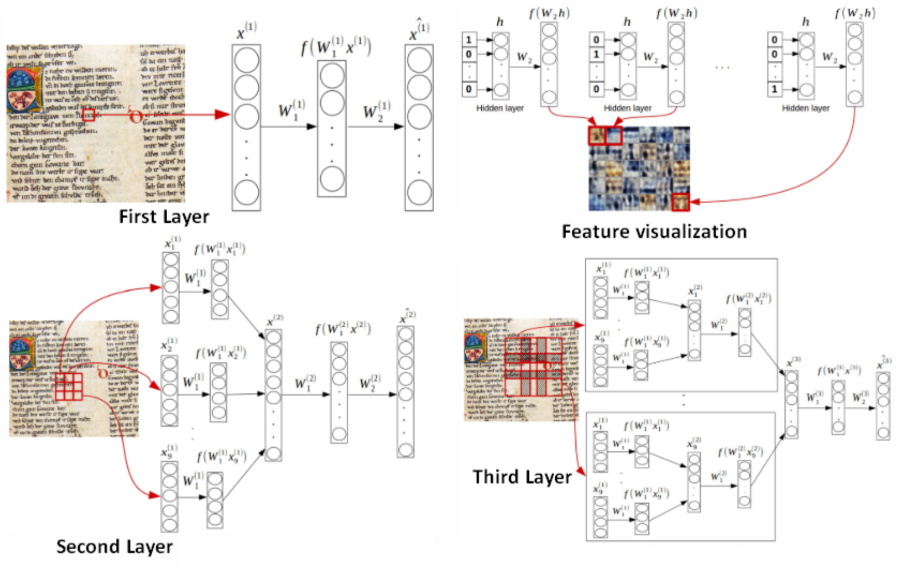
The approach differs from traditional page segmentation methods in two properties. (1) The features are learned directly from the pixels without supervision. (2) Preprocessing (i.e., binarization, connected components extraction) and prior knowledge are not needed.

Given a set of labeled training images and a trained CAE: (1) Extract features with the CAE on randomly selected image patches from the training images. (2) Concatenate the activation values of each layer of the CAE as feature vectors to train an SVM.

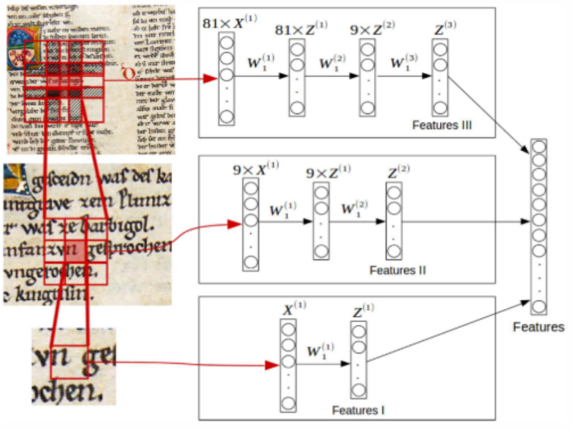
2.1 Feature learning

Features are extracted by a single-layer fully-connected neural network as an AE which tries to reconstruct the input data. Features can be discovered in the hidden layer. The number of hidden neurons is smaller than the input size to prevent learning of identity functions. The AE learns the weights W_1 and W_2 such that $f(W_2 f(W_1 x)) = \hat{x}$, where x is the input vector, the output $\hat{x}$ is similar to x and f the activation function $f(x) = \frac{x}{1+|x|}$.

In order to learn high-dimensional feature representations from unlabeled pixels, the feature learning system contains three levels:

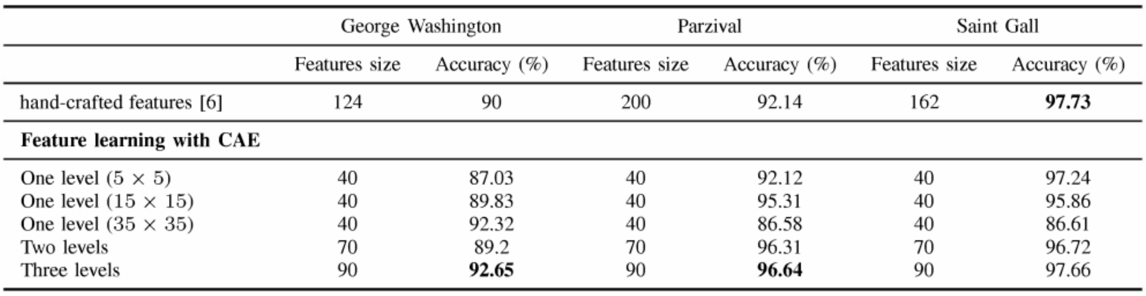


- First Level. Randomly select 500K 5 x 5 pixels image patches $P^{(1)}$ from the training set. Therefore, Giving an input vector of size $x^{(1)} \in \mathbb{R}^{75}$, With AE of 40 hidden layers, the weights $W_1^{(1)}$ and $W_2^{(1)}$ are learned using SGD.
- To learn higher level features and use the property that part of an image shares similarities with other parts, 15 × 15 pixels image patch $P^{(2)}$ is composed by 3 × 3 patches $P^{(1)}$ (from $x_1^{(1)}$ to $x_9^{(1)}$) without overlaping. This time the weights $W_1^{(2)}$ and $W_2^{(2)}$ are learned.
- Repeating the same procefure for the second level with an overlapping of 5 pixels in order to get more information at the borders. Thus $P^{(3)}$ covers 35 × 35 pixels. This time the weights $W_1^{(3)}$ and $W_2^{(3)}$ are learned.



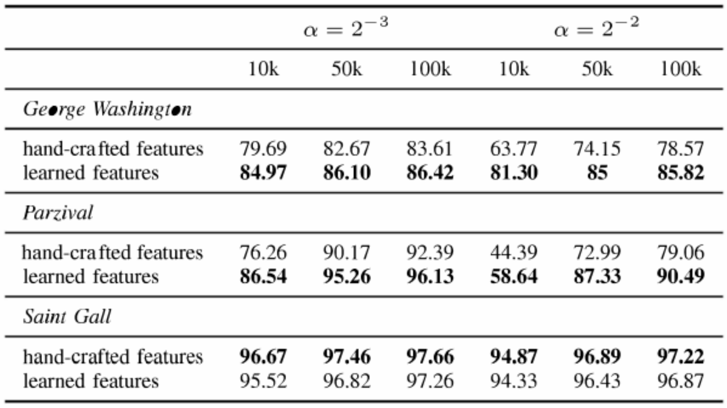
Feature extraction and classifier training The features of a given pixel are the concatenation of the nth level features $z^{(n)}$ from patches $P^{(n)}$ centered on the pixel where $z^{(n)} = f\left(W_1^{(n)} x,^{(n)}\right), n \in \{1,2,3\}$, An SVM is trained on randomly selected pixels features with their label on the training set.

3 Results



	George Washington		Parzival		Saint Gall	
	Features size	Accuracy (%)	Features size	Accuracy (%)	Features size	Accuracy (%)
hand-crafted features [6]	124	90	200	92.14	162	**97.73**
Feature learning with CAE						
One level (5 × 5)	40	87.03	40	92.12	40	97.24
One level (15 × 15)	40	89.83	40	95.31	40	95.86
One level (35 × 35)	40	92.32	40	86.58	40	86.61
Two levels	70	89.2	70	96.31	70	96.72
Three levels	90	**92.65**	90	**96.64**	90	97.66

Table II: Classification accuracy (%) on various scaling factor values and number of training samples.



	$\alpha = 2^{-3}$			$\alpha = 2^{-2}$		
	10k	50k	100k	10k	50k	100k
Ge•rge Washingt•n						
hand-crafted features	79.69	82.67	83.61	63.77	74.15	78.57
learned features	**84.97**	**86.10**	**86.42**	**81.30**	**85**	**85.82**
Parzival						
hand-crafted features	76.26	90.17	92.39	44.39	72.99	79.06
learned features	**86.54**	**95.26**	**96.13**	**58.64**	**87.33**	**90.49**
Saint Gall						
hand-crafted features	**96.67**	**97.46**	**97.66**	**94.87**	**96.89**	**97.22**
learned features	95.52	96.82	97.26	94.33	96.43	96.87

References

[1] Chen, Kai Seuret, Mathias Liwicki, Marcus Hennebert, Jean Ingold, Rolf, *Page Segmentation of Historical Document Images with Convolutional Autoencoders*