Page Segmentation for Historical Handwritten Document Images Using Conditional Random Fields

(2016)

A Resume

November 30, 2018

Abstract

Using CRFs to deal with problem of local and contextual information - Segmentation of historical documents.

1 Introduction

Page segmentation is an important initial step for document image analysis and understanding. The goal is to split a document image into regions of interest.

In order to achieve better segmentation, contextual information is needed. By considering the labels of neighboring image patches helps to decide the label of a given patch. Instead of labeling pixel by pixel individually, the labes Y are predicte jointly, such that $\hat{Y} = \arg\max_{Y \in \mathcal{Y}} P(Y|X)$ where $\mathcal{Y}$ is all the possible labeling over the pixels.

2 Method

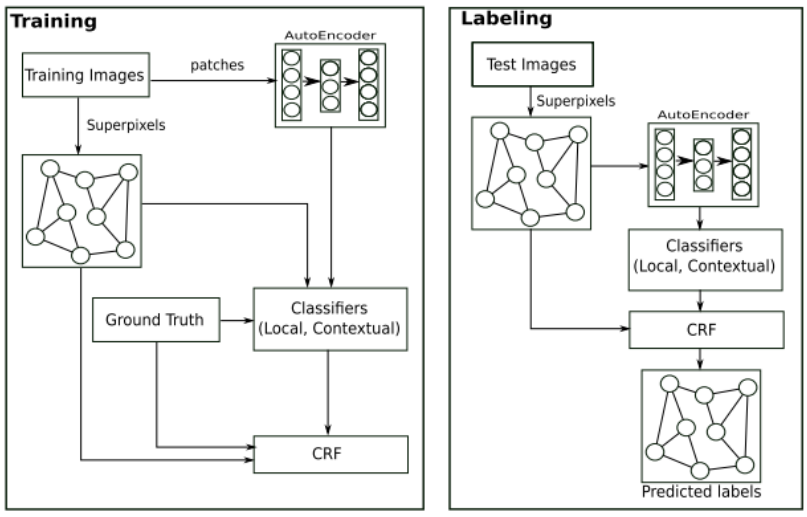

Figure 1: Page segmentation workflo

The proposed method consists of two steps. In the first step, features are learned from pixels with an unsupervised learning method. In the second step, an image is represented as a graph, where

each node is represented by a superpixel. Then the superpixels are labeled into different classes with a CRF model.

Feature learning Using a neural network with one hidden layer as an autoencoder (Concolutionnal AE). The AE learns to reconstruct its input data. Features can be discovered in the hidden layer. Concretely, the AE learns the weights $W1$ and $W2$, such that $\sigma\left(W_2\sigma\left(W_1x\right)\right) = \hat{x}$, where x is the input vector, the output $\hat{x}$ is similar to x soft-sign function, such that $(x) = x$ and () is the activation function is $\sigma(x) = \frac{x}{1+|x|}$.

They use three layer of autoencoders, **First layer**, 10 millions 5 5 pixels image patches $\mathcal{P}^{(1)}$ hidden neurons of the AE is set to 40. **Second layer** takes the results of the first layer and constructs, an input of 15x15 pixels for $\mathcal{P}^{(2)}$ composed if 3x3 patches of $\mathcal{P}^{(1)}$, For the **Third layer**, the same procedure is applied, $\mathcal{P}^{(3)}$ this time covers 45x45 pixels.

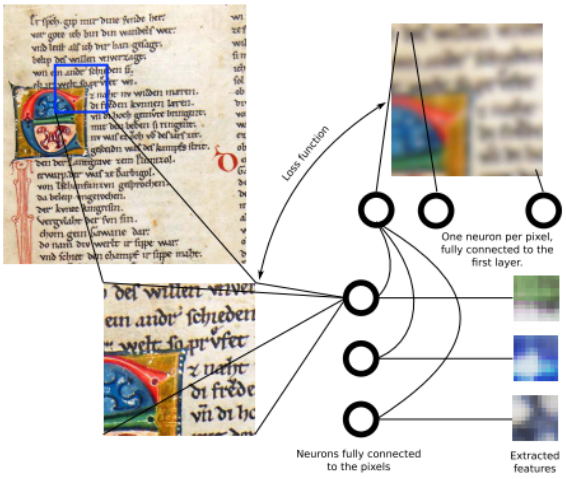

Figure 2: Extracted features

Conditionnal random field For a given image, let X be the observation over a set of nodes S, such that $X = \{x_i\}, i \in \{1, 2, \ldots, n\}$. Where x_i is the observation of node si, $si \in S$, and $|S| = n$. Let Y be a set of random y_i is the label of node s_i and L is the label set. We define variables over the nodes, such that$Y = \{y_i\}, y_i \in \mathcal{L}$, where y_i is the label of node s_i and L is the label set. We define $\mathcal{L} = \{P, B, T, D\}$ for periphery, background, text block, eight nearest neighbors.

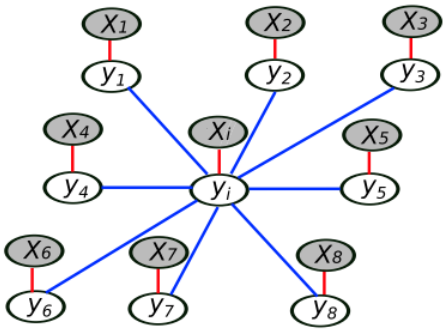

Figure 3: The topology of the CRF model. Each node yi represents a label of a superpixel connects to its eight nearest neighbors. Each observation node xi represents the features of the superpixel. Edges represent the similarity between any pair of nodes.

MLP Two kind of feature functions are defined: local feature function and contextual feature function. Using a discriminative classifiers to model these feature functions. The Multilayer Per-

ceptron (MLP) is used to create the feature functions due to its performance. However, it could be replaced by other classifiers:

Local feature function $f_l(\cdot)$: using the autoencoder features and ground-truth label of the superpixels on the training images to train an MLP. For a given node s_i, the MLP outputs a vector of scores $c_j^i, j \in \{1, 2, \ldots, |\mathcal{L}|\}$, c_i is the label set. Each c_j^i presents the probability of label l_j associated to the node s_i.

Contextual feature function $f_c(\cdot)$: Using an MLP, taking as input the output of the local feature function of the nine nodes of its neighbors, to output a vector of scores $d_j^i, j \in \{1, 2, \ldots, |\mathcal{L}|\}$. Each d_j^i presents the probability of label l_j assigned to the node s_i given the label configurations of its neighbors. The weight $_c$ of the contextual function $f_c()$ is a vector of size $|\mathcal{L}|$.

Inference For a given image X, in order to find the optimal label configuratin without computing the normalization factor for the whole image document, they apply the Iterated Conditioal Model to find an approximate solution of :

$$\hat{Y} = \arg\max_{Y \in \mathcal{Y}} P(Y|X; \lambda)$$

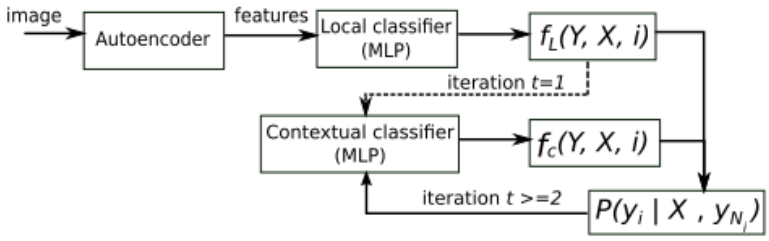

Figure 4: Inference

Side note: Simple Linear Iterative Clustering) SLIC is the state of the art algorithm to segment superpixels which doesnt require much computational power. In brief, the algorithm clusters pixels in the combined five-dimensional color and image plane space to efficiently generate compact, nearly uniform superpixels.

How does SLIC work? The approach is really simple actually. SLIC performs a local clustering of pixels in 5-D space defined by the L, a, b values of the CIELAB colorspace and x, y coordinates of the pixels. It has a different distance measurement which enables compactness and regularity in the superpixel shapes, and can be used on grayscale images as well as color images.
SLIC generates superpixels by clustering pixels based on their color similarity and proximity in the image plane. A 5 dimensional [labxy] space is used for clustering. CIELAB color space is considered as perpetually uniform for small color distances. It is not advisable to simply use Euclidean distance in the 5D space and hence the authors have introduced a new distance measure that considers superpixels size.

References

[1] Kai Chen et all, *Page Segmentation for Historical Handwritten Document Images Using Conditional Random Fields*