AIP

Midterm Project Report

Vishay Raina
14860

Abstract

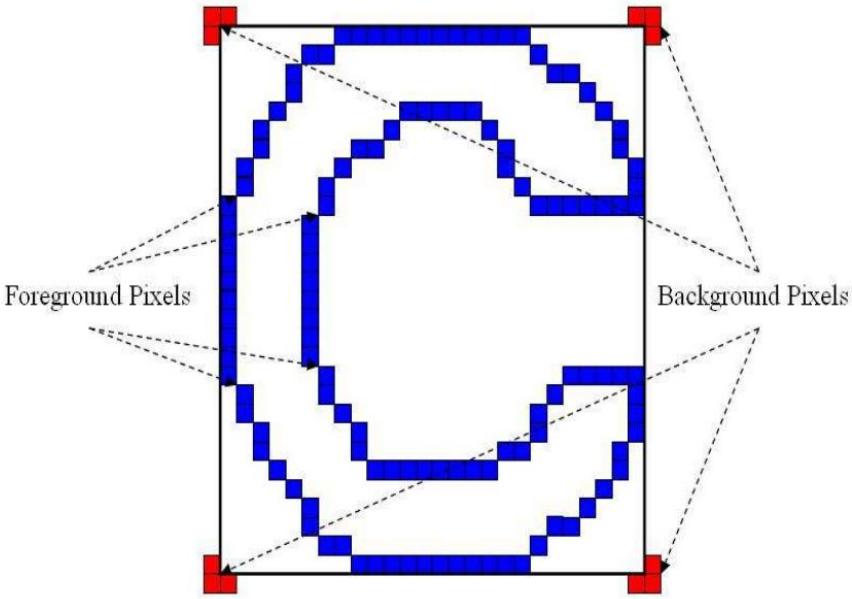
In this project, two different approaches for document binarization were used to binarize a dataset of magazine documents. One approach is an unsupervised approach, based on edge detection followed by connected component algorithm and rejection of non text regions based on some sensible assumptions and the other is a supervised learning approach using a CNN. The binarized images were compared with the ground truth using a pixel wise comparison and then computing the f-measure.

Unsupervised Approach

Edge detection: The proposed method uses an edge-based connected component approach to automatically obtain a threshold for each component. Canny edge detection is performed individually on each channel of the color image and the edge map is obtained by combining the three edge images as follows:

$$E = E_R \vee E_G \vee E_B$$

Connected Components: An 8-connected component algorithm follows the edge detection step to find the contours in the image and the associated bounding box information is computed.



Elimination: The aspect ratio is constrained to lie between 0.1 and 10 to eliminate highly elongated regions. The size of the EB should be greater than 15 pixels but smaller than 1/5th of the image dimension to be considered for further processing.

Since the edge detection captures both the inner and outer boundaries of the characters, it is possible that anEB may completely enclose one or more EBs . So if a contour contains one or two EBs inside it, then it may be a character (eg. 8) and the inner EBs can be ignored, but if a contour contains more than two EBs inside it then it is a carrier of text characters and it can be ignored whire the inner EBs are retained.

Threshold estimation: The foreground intensity is computed as the mean gray-level intensity of the pixels that correspond to the edge pixels. Local background intensity is estimated considering the median intensity of the 12 background pixels at the periphery of the corners of the bounding box.
Depending on whether the fore-ground intensity is higher or lower than that of the background, each binarized output is suitably inverted so that the foreground text is always black and the background always white.

Supervised approach

This approach basically consists of querying each pixel of the image to classify it as foreground or background. This is done by taking a window around the pixel and passing it through a CNN to distinguish between these two categories.

Network: conv(3,3,32)→maxpool(2,2)→conv(3,3,32)→maxpool(2,2)→dropout(0.25)→fc(128) →dropout(0.5)→f c(2)

Assumption: the region surrounding the pixel of interest contains enough information to discriminate between these two cases.

Evaluation: Leave one out cross validation was applied to a dataset of 10 pages from a corpus of scanned musical notes. From the 9 pages, 2 million training samples were randomly chosen (5% of the data). And 90% data was used for testing and 10 % for validation to determine the number of epochs to be trained.

Dataset:

LRDE Document Binarization Dataset (LRDE DBD)

This is a dataset is composed of 125, full-document images,of resolution : 3272 x 2516
and Binarized ground truth images to perform an evaluation of binarization algorithms.

Evaluation

The dataset was processed through the unsupervised approach to generate the binarized images. These were compared pixel by pixel with the ground truth and f1 measure was calculated:
Mean F1 Score : : **0.5445**
Standard deviation : : **0.1089**

In the supervised approach, the model that was used was:

conv2d(3,3,32)→batch_norm()→max_pool(2,2)→dropout(0.25)→fc(128)→batch_norm →dropout(0.5)→fc(2)

The number of trainable parameters was 263,746.
From 10 images, of size 3272 x 2516, 3 million samples were taken randomly. 10% of it was used for validation to optimize the number of epochs by using early stoppage based on the increase in the value of validation loss.
The model would usually stop training after 2 epochs.
The loss function used was binary_crossentropy loss and the optimizer used was Adam.
25% of the training data used for testing. Trained on 2025000 samples, validated on 225000 samples and tested on 750,000 samples the model gives:
F1 Score : : **0.9830**

Conclusion

Both are interesting and unique approaches to the same problem, while the deep learning method might be more accurate, but is really time consuming. While the unsupervised method only takes few seconds to binarize a document, the CNN takes almost an hour to train and to binarize an image of size 3272 x 2516, ie. 8,232,352 data points takes a huge time when compared to the unsupervised method.

Original image

Binarized by unsupervised method