Robust Text Reading in Natural Scene Images

Anshuman Majumdar, Shivin Yadav and Shubham Vijayvargiya

Door Drishti

1 Introduction

The problem involves detection of text regions and recognition of the extracted text in natural scene images. Retrieval of text from scene images is a quite challenging task. The first step is isolation of the textual regions in an image. The detected text is then fed into a recognition system which gives the textual information as output. Finally, we get an end-to-end text recognition system for natural scene images.

2 Previous Work

There has been numerous works in the past in this domain. Here we focus on some of the selected works and try to implement them for the purpose of our project.

2.1 Text Localization

There are two different approaches have been used for text localization from complex images:

(i) Region based approach - It is basically divided in two sub categories - edge based and connected component (CC) based methods. The edge based method mainly focuses on the high contrast between text and background. In this method, firstly text edges are identified in an image and are merged. Finally, some heuristic rules are applied to discard non-text regions. Connected component based method considers text as a set of separate connected components,each having distinct intensity and color distributions. The edge based methods are robust to low contrast and different text size whereas CC based methods are somewhat simpler to implement, but they fail to localize text in images with complex backgrounds.

(ii) Texture based methods - Text in images have distinct textural properties which can be used to differentiate them from the background or other non text regions. This method is based on the concept of textural properties. In this method, Fourier transforms, Discrete cosine transform and Wavelet decomposition are generally used. The main drawback of this method is that it is highly complex in nature but, on the other hand, it is more robust than the CC based methods in dealing with complex backgrounds.

2.1.1 Text Detection using Maximally Stable Extremal Regions (MSER)

The MSER feature detector works well for finding text regions in unstructured scenes. It works well for text because the consistent color and high contrast of text leads to stable intensity profiles. Although the MSER algorithm picks out most of the text, it also detects many other stable regions in the image that are not text.

2.1.2 Text Detection using Stroke Width Transform (SWT)

A constant stroke width of textual content is utilized to recover regions that are likely to contain text. The main idea is to compute the stroke width for each pixel. The method suggested here differs from previous approaches in that it does not look for a separating feature per pixel, like gradient or color.

2.2 Text Recognition

This is basically either a Optical Character Recognition OCR system or an intelligent, trained learning model that can be used to find the ground truth of the text from text images.

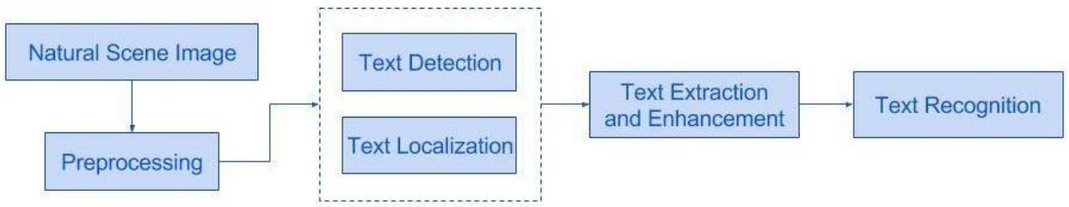
2.2.1 Reading Text in the Wild with Convolutional Neural Networks (CNNs)

Here the text recognition system is in the form of a deep convolutional neural network which takes the whole word image as input to the network. Evidence is gradually pooled from across the image to perform classification of the word across a huge dictionary. The CNN model is trained purely on synthetic data.

3 Project Architecture

3.1 Method Workflow

The basic work flow of text detection and recognition method can be seen in the figure below. Preprocessing steps are used to prepare the image for text detection algorithms. Then, MSER and SWT features are used to get the text regions from the images. The text regions are then cropped out as patches using their bounding boxes and fed to the text recognition system. In the heart of the text recognition system is a CNN model which is used to recognize the given text from the extracted patches.



3.2 MSER Feature Extraction

MSER is a method for blob detection in the images which uses stable connected components of some gray level sets in the image. MSER depends on thresholding of the image, considering some threshold value. The pixels below that threshold value are considered 'white' and all those equal or above are considered 'black'. MSER detects the objects but these objects can also contain parts of the background. Those are removed in the canny edge detection process. While MSER has been identified as one of the best region detectors due to its robustness against view point, scale, and lighting changes, it is sensitive to image blur. Thus, small letters cannot be detected or distinguished in case of motion or defocus blur by applying plain MSER to images of limited resolution. To cope with blurred images we propose to combine the complementary properties of Canny edges and MSER. The outline of extremal regions can be enhanced by applying the precisely located but not necessarily connected Canny edges. We remove the MSER pixels outside the boundary formed by the Canny edges. This is achieved by pruning the MSER along the gradient directions computed from the original gray-scale image. Since the type of the letter (bright or dark) is known during the MSER detection stage, the gradient directions can be adapted to guarantee that they point towards the background.

3.3 SWT Method

SWT tries to capture the only text effective features and using geometric signature of text to filter out non-text areas. As a result, SWT gives you reliable text regions that is language neutral. So the non text regions from text are further classified using stroke width transform. The Stroke Width Transform is a local image operator which computes per pixel the width of the most likely stroke containing the pixel. The output of the SWT is an image of size equal to the size of the input image where each element contains the width of the stroke associated with the pixel. A a stroke is defined to be a contiguous part of an image that forms a band of a nearly constant width. The output of the SWT is an image where each pixel contains the width of the most likely stroke it belongs to. The next step of the algorithm is to group these pixels into letter candidates. Two neighboring pixels may be grouped together if they have similar stroke width. For this the classical Connected Component algorithm is modified by changing the association rule from a binary mask to a predicate that compares the SWT values of the pixels.

3.4 Text Line Formation and Word Separation

Text lines are important cues for the existence of text, as text almost always appear in the form of straight lines or slight curves. To detect these lines, we first pairwise group the letter candidates using the following rules. As letters belonging to the same text line are assumed to have similar stroke width and character height, two letter candidates are paired if the ratio of their stroke width medians is lower than 1.5 and their height ratio is lower than 2 (taking upper and lower case letters into account). Additionally, two CCs should not be paired if they are very distant. Subsequently, text lines are formed based on clusters of pairwise connected letter candidates. A straight line is fitted to the centroids of pairs of letter candidates within each cluster and the line that intersects with the largest number of text candidates is accepted. The process is iterated until all text candidates have been assigned to a line, or if there are less than three candidates available within the cluster. A line is declared to be a text line if it contains three or more text objects.

3.5 Text Recognition using CNNs

The second stage of our framework produces a text recognition result for each proposal generated from the detection stage. We take a whole-word approach to recognition, providing the entire cropped region of the word as input to a deep convolutional neural network. Our method for text recognition also follows a whole word image approach. We take the word image as input to a deep CNN, however we employ a dictionary classification model. Recognition is achieved by performing multi-class classification across the entire dictionary of potential words. The final fully-connected layer performs classification across the dictionary of words, so has the same number of units as the size of the dictionary we wish to recognize.

4 Libraries and Tools

We will be using the following libraries and tools for carrying out this project -

- MATLAB with MatConvNet and VLFeat
- C++ with OpenCV

5 Image Dataset

We will be using the following datasets to conduct our experiments -

- COCO-Text dataset
- ICDAR-2015 incidental text dataset

6 Work Distribution

The work will be distributed as follows among the team members -

- Anshuman Majumdar - SWT Method
- Shivin Yadav - MSER Feature Extraction
- Shubham Vijayvargiya - Text Recognition using CNNs

7 Estimated Project Completion

We are expecting to complete the text detection and localization part till the second evaluation.

8 References

[1] Chen, Huizhong, et al. "Robust Text Detection in Natural Images with Edge-Enhanced Maximally Stable Extremal Regions." Image Processing (ICIP), 2011 18th IEEE International Conference on. IEEE, 2011.
[2] Gonzalez, Alvaro, et al. "Text location in complex images." Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE, 2012.
[3] Li, Yao, and Huchuan Lu. "Scene text detection via stroke width." Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE, 2012.
[4] Neumann, Lukas, and Jiri Matas. "Real-time scene text localization and recognition." Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012.
[5] J. Canny, A computational approach to edge detection, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 8, pp. 679698, 1986.
[6] Epshtein, Boris, Eyal Ofek, and Yonatan Wexler. "Detecting text in natural scenes with stroke width transform." Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010.
[7] Jaderberg, Max, et al. "Reading text in the wild with convolutional neural networks." International Journal of Computer Vision 116.1 (2016): 1-20.
[8] Chen, Huizhong, et al. "Robust text detection in natural images with edge-enhanced maximally stable extremal regions." 2011 18th IEEE International Conference on Image Processing. IEEE, 2011.