CS231A Course Project Proposal
Robust Text Reading in Natural Scene Images

Tao Wang, David Wu
Computer Science Department
Stanford University
353 Serra Mall, Stanford, CA 94305
{twangcat, dwu4}@stanford.edu

Abstract

This project aims to build an end-to-end system which is able to detect and recognize English text in natural scene images with accuracy that is compatible with the state-of-art results.

1. Introduction

Text extraction from images has been a long-studied computer vision problem. Although state-of-art methods have near-perfect accuracies on scanned typewritten documents in English, none of them works well enough on text in natural images to compete with human accuracy. Unlike the quasi-binary setting in scanned documents, various factors contribute to the difficulty of natural scene text recognition. Examples include font and color diversity, background clutter, text-background ambiguity and scale variation. The aim of this project is to build a scene text reading system which is robust to these adverse conditions and consistently outputs good accuracies on natural images.

2. Why is Scene Text Recognition Interesting

Apart from being an unsolved challenging task, scene text recognition is itself an interesting research topic in computer vision. As a special case of object recognition, results of scene text recognition could provide useful insights for more general recognition/classification tasks. Many new object classification algorithms are indeed benchmarked on text-like datasets such as MNIST. In addition, taking text infomation into account could be another avenue to improve scene understanding.

3. Datasets

The most commonly used dataset for scene text recognition is the ICDAR 2003 Robust Reading dataset, which will be the main dataset that my system is trained and tested on. In addition to that, many authors also use synthetic images to inject more variety to the datasets. Since images containing text on natural scene backgrounds can be easily synthesized, this project will also adopt this approach. Other datasets such as Chars74k and Weinman will be used as auxiliary datasets.

4. Approach

Like many previous works on scene text recognition, we adopt a multi-stage approach. We will first train a text detector which will attempt to detect possible regions of text. This is essentially a binary classifier to find regions of interests (ROIs) for the character level classifier The next step is to train a character level classifier which has high accuracies on patches containing one character each (including a "non-character" class and a "space" class). The classifiers in both stages can be trained using the methods introduced in the class: Bag-of-Words, SIFT, or even learnt features. For text detection in particular, we also would like to compare our own approach with the Stroke Width Transform (SWT) which is the current state-of-art known to produce the best result on the ICDAR dataset. We then slide the character classifer across the ROIs at different scales to obtain a confidence map for each character class at each location. For text detection in particular, we also would like to compare our own approach with the Stroke Width Transform (SWT) which is the current state-of-art known to produce best result on the ICDAR dataset. We then use a probablistic model such as CRF which incorporates some character-level language model to output the most likely text. With the help of a large synthetic dataset, we would also like to scale our system up on GPUs to achieve higher accuracy.

5. Evaluation

As a convention in the scene text recognition community, result evaluation can be done on the character level and word level, which compare the output string of the system with the ground truth, and find the ratio of matching characters or words respectively. Results can also be evaluated on an image level, in which we find the ratio of images in which the output matches the ground truth exactly. Apart from end-to-end measures, character level accuracies on isolated letters and pixel level accuracies can be used to evaluate the performance of the character classifier and and the text detector module individually.