Text extraction from document images

Aditya Harsh
For the Inkredo challenge

1 Definition

This section defines the problem statement as well as an overview of the document scanning. The metrics used for evaluating the model and building the feature set are also defined below.

1.1 Project Overview

Smartphones are replacing personal scanners. They are portable, connected, powerful and affordable. They are on their way to become the new entry point in business processing applications like document archival, ID scanning, check digitization.

Mobile phone camera-based document video scanning is an interesting problem which has entered into a new era with the emergence of widely used, processing capable and motion sensors equipped smartphones. The use of these devices for document scanning provides interesting advantages over the traditional document scanning devices. They can be used to scan thick books, historical documents that are too fragile to touch, text in scenes (walls, whiteboards, etc.), and large sized documents.

.

1.2 Task

An efficient capture process should be able to:

1. Detect and segment the relevant document object during the preview phase;
2. Produce a high-quality, controlled output based on the resolution captured image.
3. Extract the text from the cropped images and calculate the accuracy using the sample out shared in dataset

1.3 Metrics

Rather than using accuracy or count of words I have used **Cosine similarity** to measure the performance of the problem, as unlike other performance metrics, cosine similarity captures the context of the document as well.

The **cosine similarity** between two vectors (or two documents on the Vector Space) is a measure that calculates the cosine of the angle between them. This metric is a measurement of orientation and not magnitude, it can be seen as a comparison between documents on a normalized space because we're not taking into the consideration only the magnitude of each word count (tf-idf) of each document, but the angle between the documents. What we have to do to build the cosine similarity equation is to solve the equation of the dot product for the $\cos\theta$.

Here, the cosine of two non-zero vectors can be derived by using the Euclidean dot product formula:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

2 Analysis

Below describes how the data was gathered, which features were selected, and which algorithms were explored. Finally, I outline the benchmark used to evaluate the performance of the trading strategy.

2.1 Dataset

The DIQ dataset contains various set where each set contains the image from which the text has to be extracted and the actual output text file of the given image in a directory named “GTOCR”.

A sample of the dataset image is depicted as follow:

A sample of the actual text of this image is depicted as:

```
Entertainment Marketing, Inc.
350 West Hubbard Street
Suite 430
Chicago. Illinois 60610
Tel 312.644.0600 Fax 312.644.0698
1 June 1993
Honorable John Faso Legislative Office Building Room 448 Albany,
NY 12248
Dear Mr. Faso:
A bill currently making its way through the New York Assembly
poses a severe threat to the promotional marketing and
advertising industries.
This measure, Assembly Bill 7139, seeks to prevent tobacco use
by minors by, among other things, banning all cigarette
advertising with the exception of ads that appear in newspapers
or magazines.
This proposal is ill-conceived and potentially destructive to my
industry, which employs thousands of people in New York State.
Studies both abroad and in the United States have determined
that cigarette advertising plays a minimal role in an
individual's decision to smoke. Peer influence, and the
influence of parents and older siblings, is by far the dominant
factor. Banning advertising, therefore, will not affect the
incidence of smoking by youths.
Besides being ill-suited to achieve its stated purpose, this
proposal is unconstitutional. Since the 1970's, the courts have
ruled that commercial speech is protected under the First
Amendment. Unfortunately, by the time the courts get around to
overturning this measure, the economic damage will already have
been done.to my industry and many others.
I hope you will take the above into consideration and oppose the
passage of Assembly Bill 7139.
Sincerely,
Christopher J. Ferraro Vice President, Marketing ENTERTAINMENT
MARKETING, INC.
CJF/sa
```

2.3 Algorithms and Techniques

As mentioned above, this problem is identified as a computer vision and optical character recognition problem. I decided to first use the **contours** approach to get only the part of the document that contains the document area and remove rest of the area.

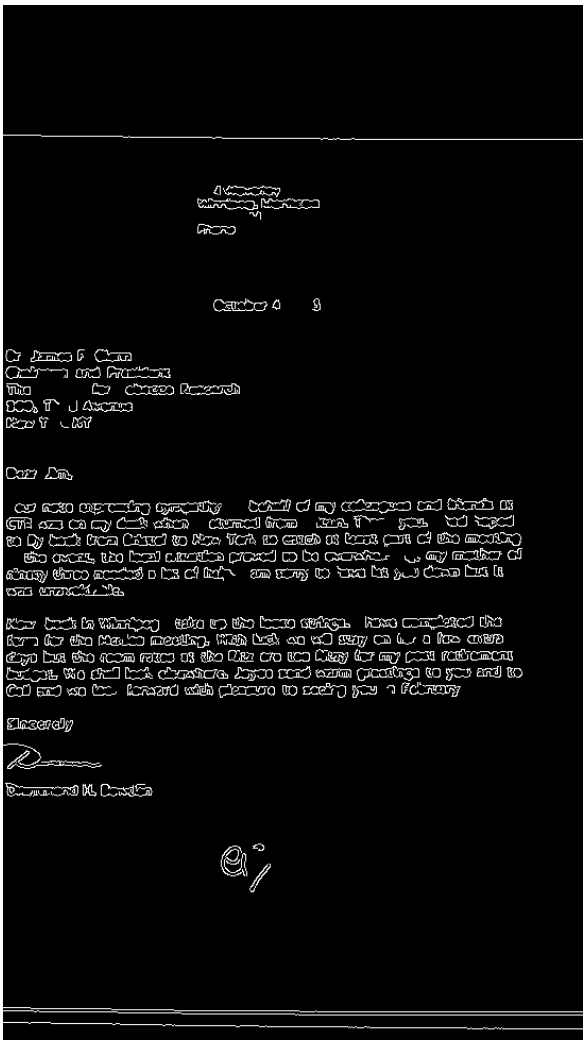
Then, I moved to using the Tesseract OCR which is a very popular engine for optical character recognition. The two approach that were chosen include:

Contour - a contour is simply the boundary of an object in an image. Various representations of contours (e.g. chain code, Fourier descriptors, and shape context) are used to recognize or categorize objects.

Tesseract with OpenCV pre-processing - any noise from the document image was removed using the python computer vision library and then the processed image was used to recognize the text using the tesseract OCR engine.

Challenges Faced

As I continued with the first approach of using contour I faced the difficulty with the boundaries of the document in images.



As we know a contour is simply the boundary of an object in an image but from the above image we can see that the document is making only the horizontal lines and not the vertical lines therefore for some images I was not getting the boundary around the document and therefore couldn't work on my idea of using contour to remove the unnecessary parts.

2.4 Benchmark

To properly test the performance of this project, I will be comparing the similarity of the documents that our algorithm returns with the original document that is correct. A satisfactory result will be one where our method will return a good cosine similarity for most of the images.

3 Methodology

3.1 Data Pre-processing

First, I converted the image to grayscale to remove some noise at pixel level but it was not giving good result. So, I continued with using the coloured image. Then, I used **dilation** and **erosion** to remove noise in the image. Dilation adds pixels to the boundaries of objects in an image, while erosion removes pixels on object boundaries. The number of pixels added or removed from the objects in an image depends on the size and shape of the structuring element used to process the image also known as kernel.

4 Results

4.1 Model Evaluation and Validation

The following table shows the cosine similarity between the actual and the predicted document for all the sets.

Set Number	Cosine Similarity
1	0.9569
2	0.9765
3	0.9563
4	0.9657
5	0.9024
6	0.9626
7	0.8789
8	0.9049
9	0.9635
10	0.9566
11	0.8475
12	0.9649
13	0.9174

We can see from the above table that the cosine similarity is more than 0.9 for almost all sets except set 7 and set 11. When analysed manually, it was found that the text in these sets were blurred and were difficult for even humans to detect. The approach was able to find most of the words in the document except some which were difficult as they were sign of the recipient.

5 Conclusion

5.1 Overview

This approach shows that when computer vision is applied with OCR engine, then it can be proven to be more efficient as the OCR doesn't do any pre-processing and simply extracts the text from the document as it which sometimes decrease the accuracy as there can be some noise present in the input images.