
Optical
character
®recognition

Overview

Motivation

- Lots of image data with text exists
- OCR allows us to get this text
- Applications can be built on top to utilize OCR for other tasks potentially on the fly

Problem

- Highest level: Given an arbitrary image, can we extract all of the text?
- Simpler: Given an arbitrary image of a single character can we identify which character it is?

Data & Methodology

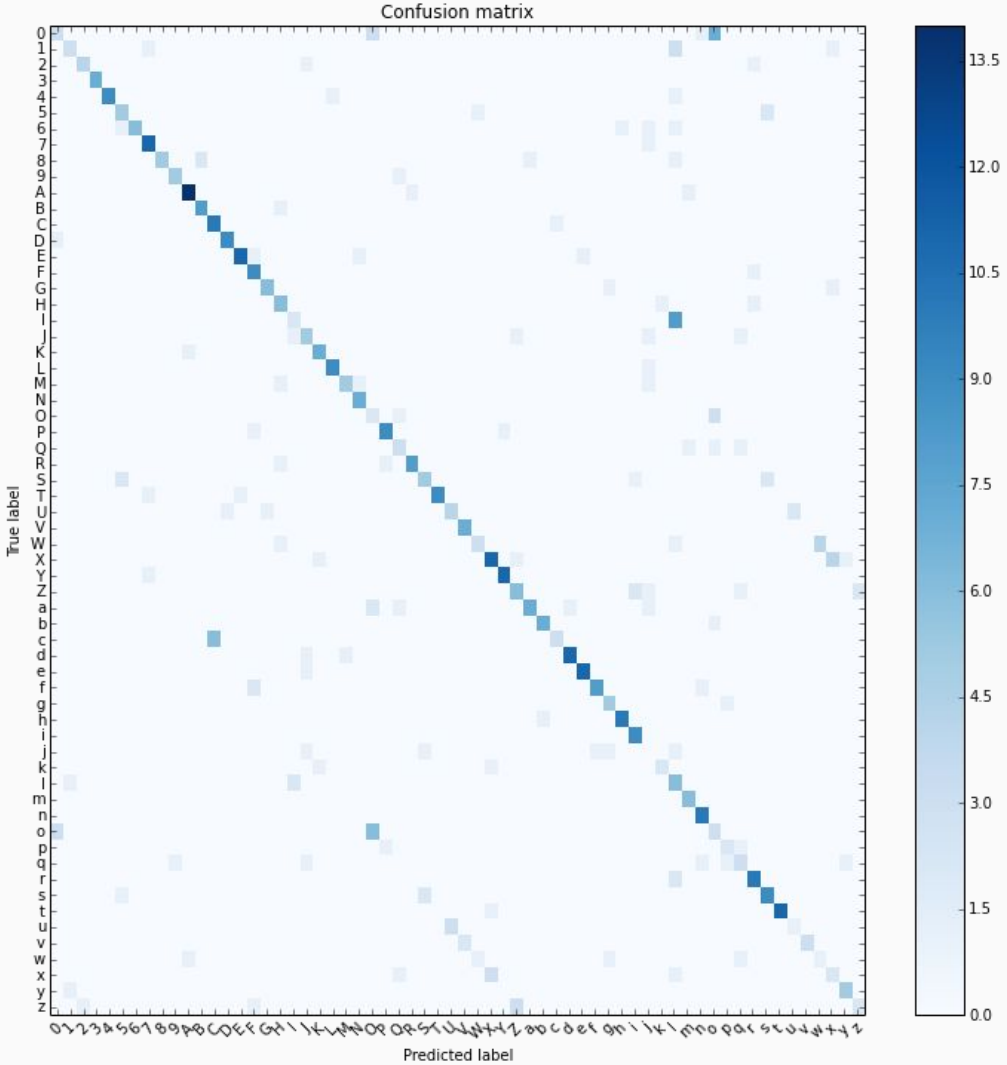
Data

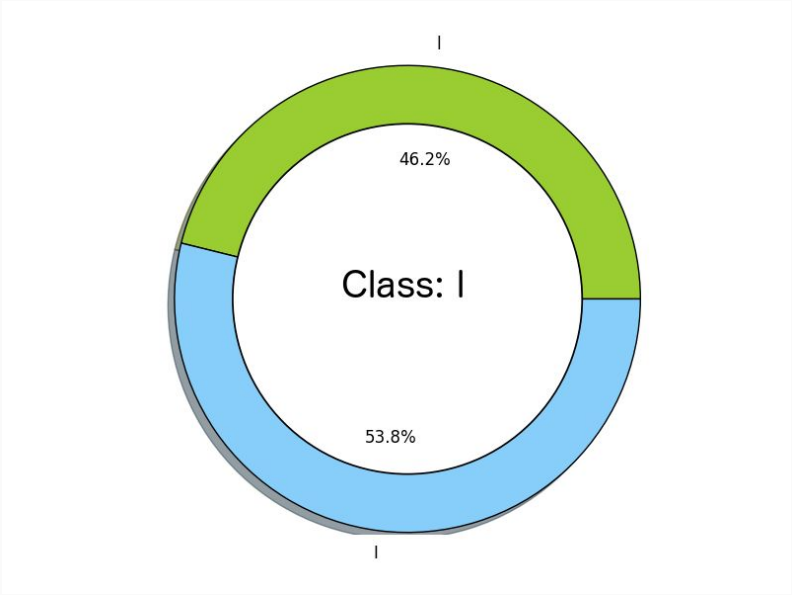
- Char 74k Dataset - Natural Images
- 62 classes (0-9, A-Z, a-z)
- 7705 images -> Used 2839
- Varying size and image quality

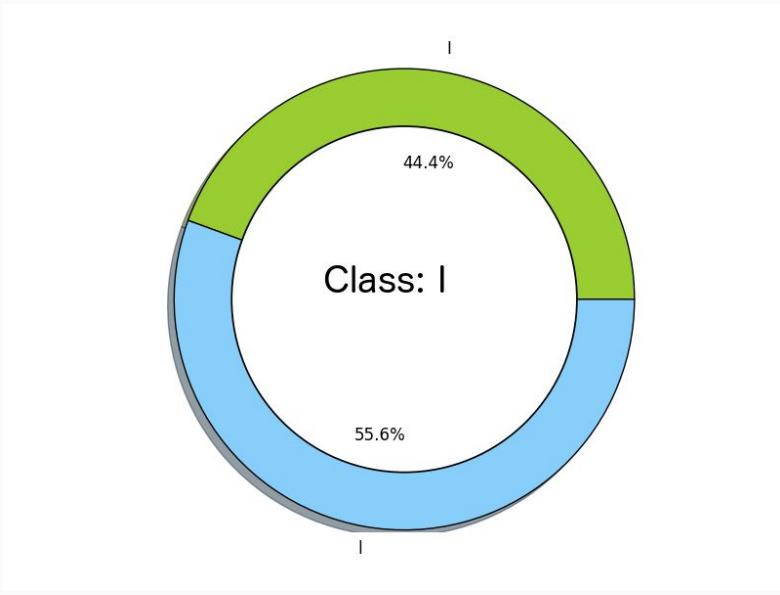
Methodology

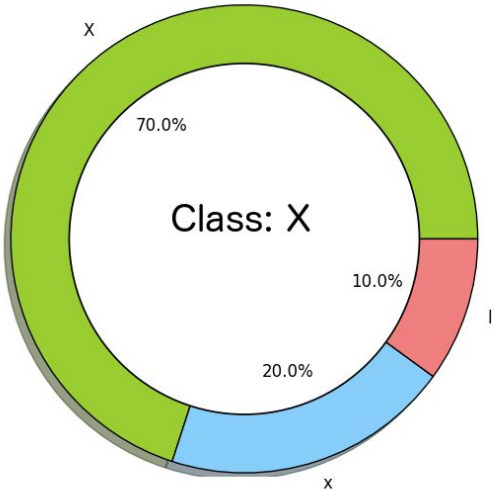
- Preprocessed greyscaled images using:
  - Binarization
  - Segmentation
  - Denoising
  - Skeletonization
  - Normalization
- Trained SVM and used Gridsearch to find the optimal parameters
- Accuracy on test set ~71%

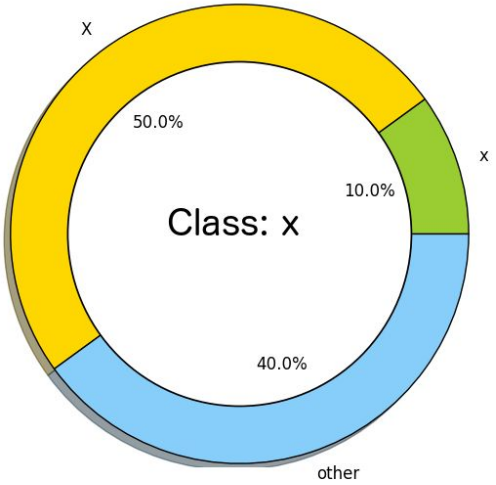
Analysis & Results


Confusion matrix
True label
Predicted label
0 1 2 3 4 5 6 7 8 9 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z a b c d e f g h i j k l m n o p q r s t u v w x y z
13.5
12.0
10.5
9.0
7.5
6.0
4.5
3.0
1.5
0.0


I
46.2%
Class: I
53.8%
I


I
44.4%
Class: I
55.6%
I


X
70.0%
Class: X
10.0%
20.0%
x


X
50.0%
Class: x
10.0%
x
40.0%
other

Conclusions

Conclusions & Future Work

- Natural images are hard to work with (filters aren’t consistent)
- Can achieve reasonable success with careful preprocessing
- Want to try analyzing the greyscaled images directly

Questions?