

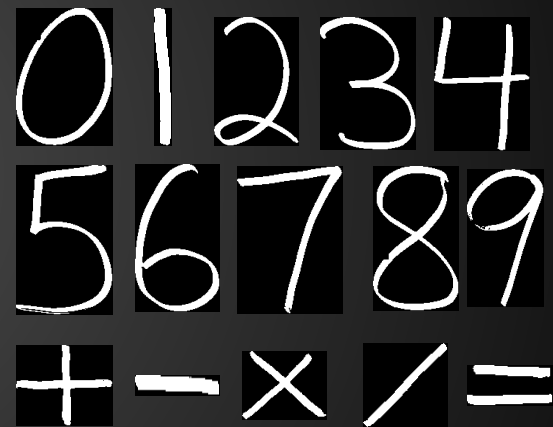
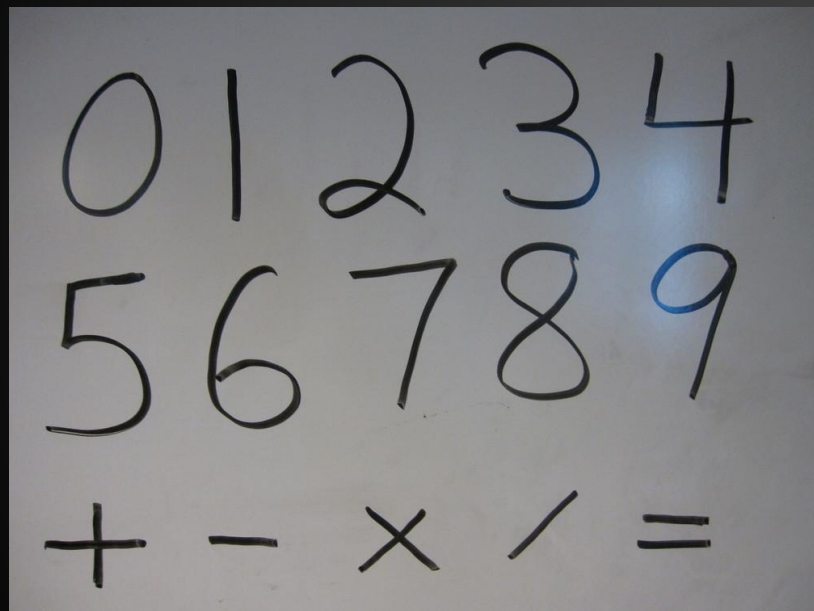
OCR to LaTeX

Chris Sadler, David McNeil, Zizhao Wang

Overview

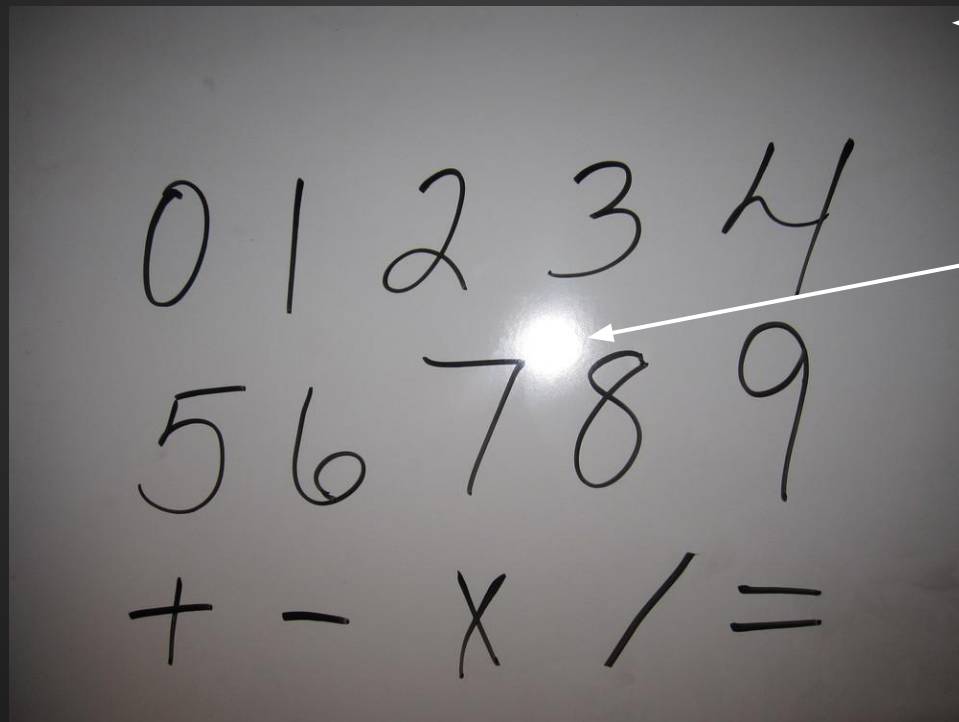
- Convert hand-drawn formula to LaTeX markup code
- Minimum requirements
 - +, -, /, *, =, and 0 through 9
- Able to classify these 15 symbols with ~91% accuracy

Image Segmentation



How do we go from a raw image to symbols we can extract features from?

Original Image

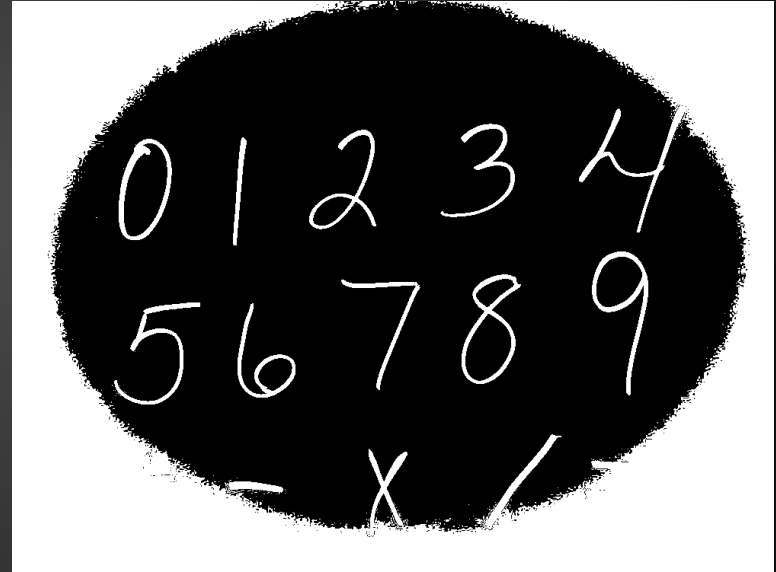


← Gradient to
very dark
color

Flash

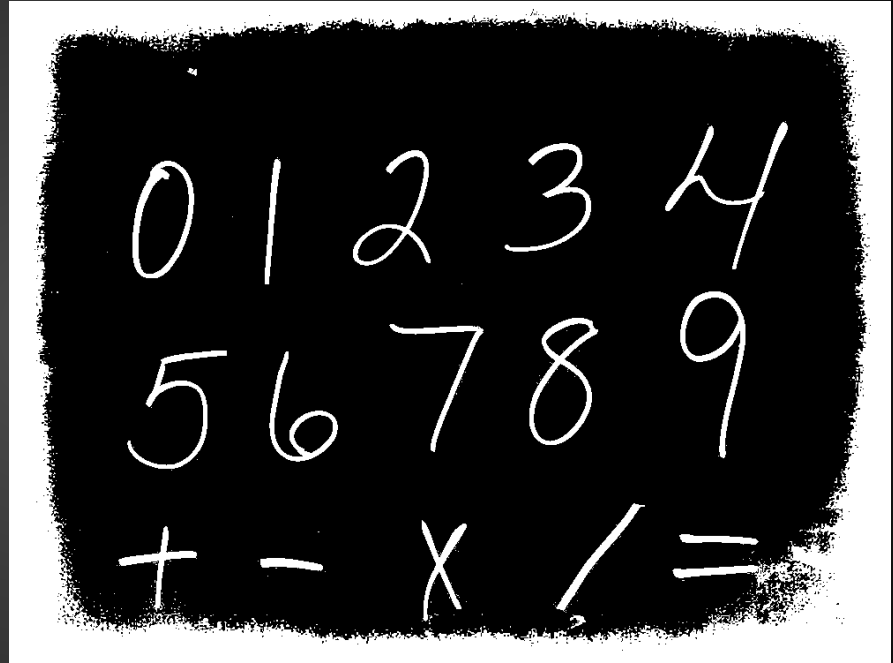
Intelligent Thresholding

- Built into matlab
- The “graythresh” function uses Otsu's method, which chooses the threshold to minimize the intraclass variance of the black and white pixels.

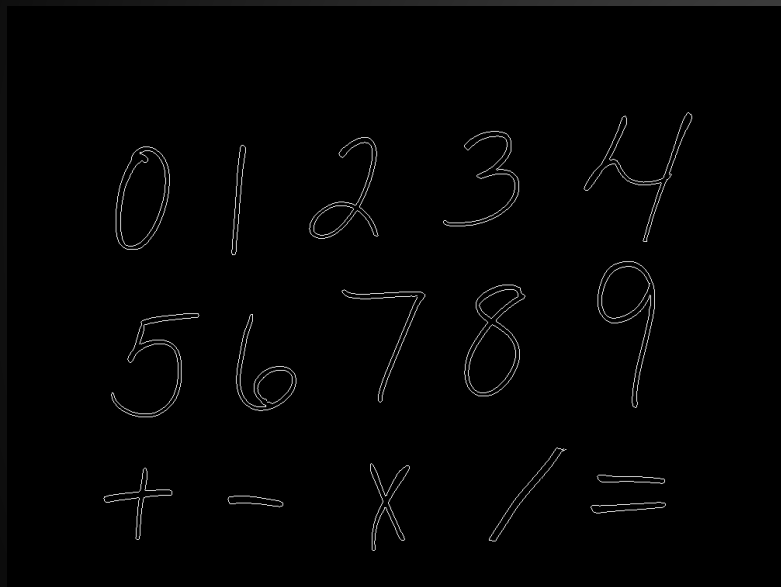


K-Means

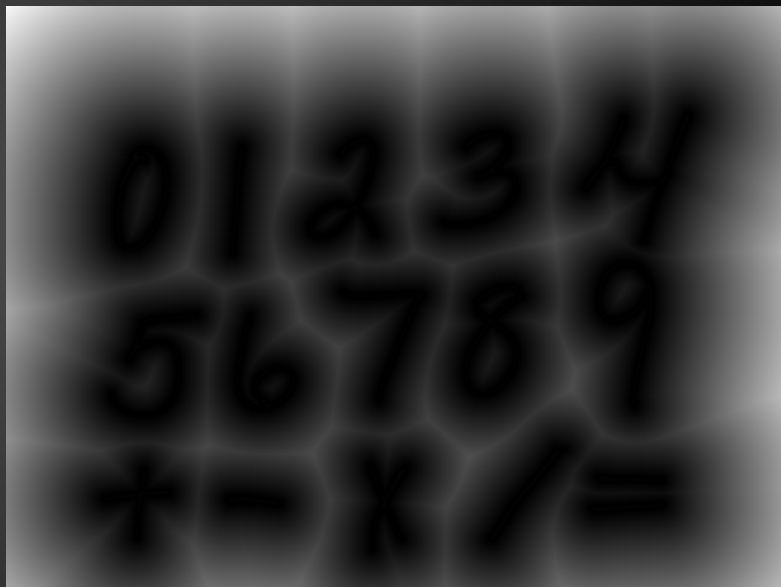
-Use unsupervised machine learning to segment image



Edges

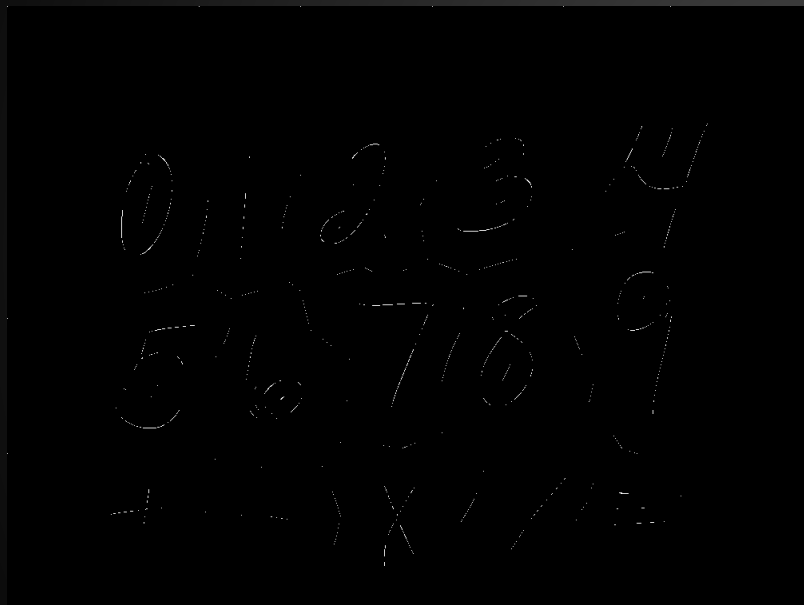


Edges using Sobel Filter

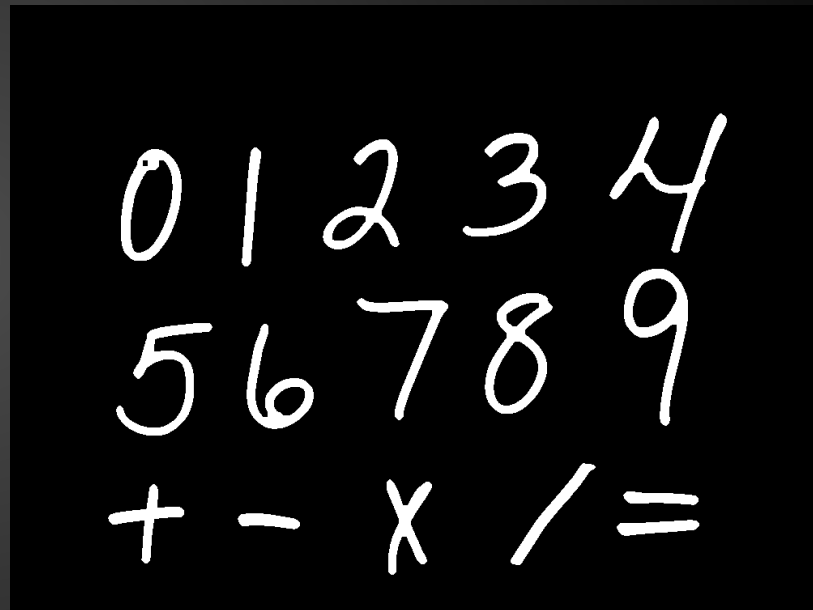


Distance to edge pixels

Edges

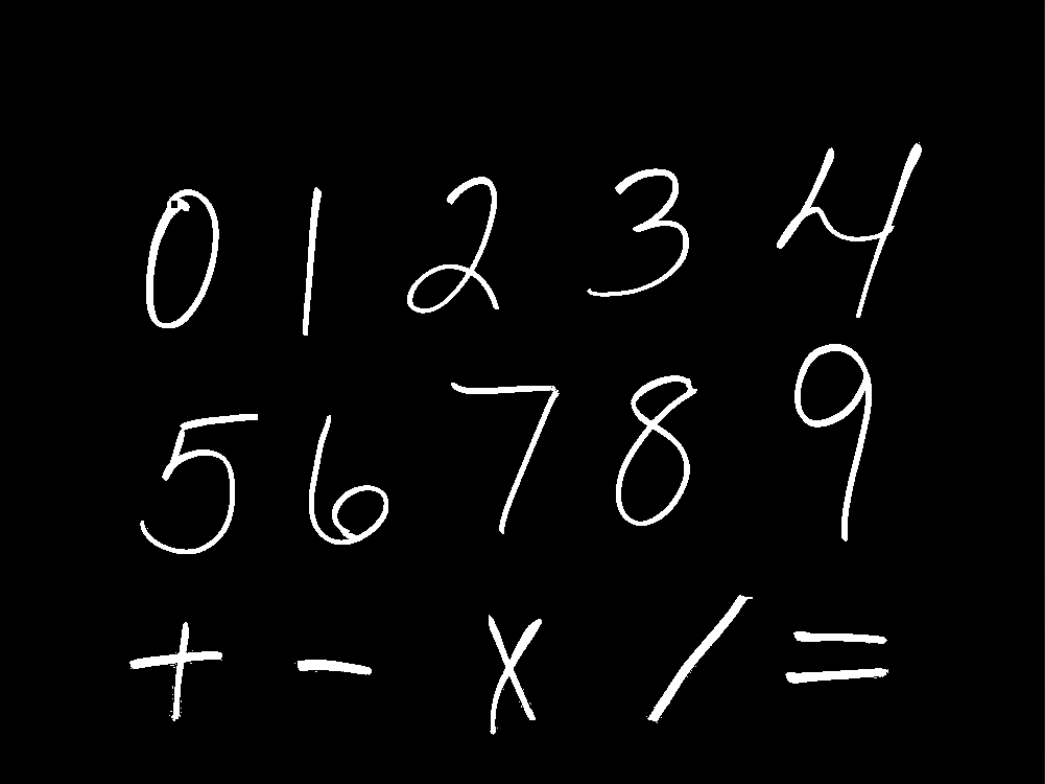


Local maximum distances to edge pixels



Dilated Image

Final Image



0 1 2 3 4
5 6 7 8 9
+ - x / =

A handwritten set of digits and mathematical operators. The first row contains the digits 0, 1, 2, 3, and 4. The second row contains the digits 5, 6, 7, 8, and 9. The third row contains the operators plus (+), minus (-), multiplication (x), division (/), and equals (=). All characters are written in a white, cursive-like script on a solid black background.

Feature Extraction

- Used a variety of features
 - Circularity, Elongation, Orientation
 - Euler Number
 - Solidity, Perimeter, Equivalent Diameter, Eccentricity
 - Centroid, Standard Deviation, Skewness, Kurtosis
 - Extent

Feature Results

- Typical feature vector:

134.6129, 1.3539, 65.4152, 1, 0.3987, 1.6035, 0.1559, 0.4928, 0.4634,
15.2121, 12.5212, -0.1778, 0.1131, 1.8016, 1.9253, 0.674, 0.3056, 0.0759,
0.1937, 0.0916, 0.1649, 0.1178, 0.1374, 0.1034, 0.1034, 0.0864

- Corresponds to:



- $\text{norm} = (\text{data} - \text{min}) / (\text{max} - \text{min})$

Training Data

- Used “The MNIST Database of Handwritten Digits”
- 10,000 image training set
- 10,000 image test set
- 91% accuracy

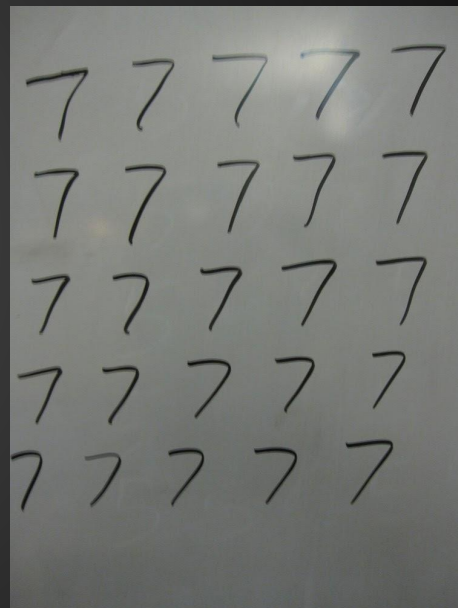
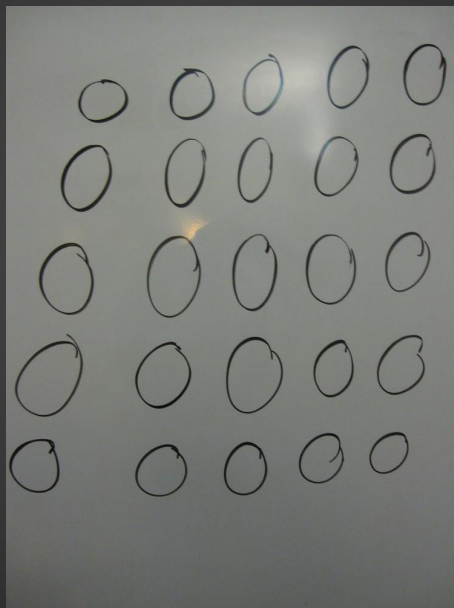


Problems with MNIST Data

- Small pictures
- Only numbers
- Square box

Our Training Set

- 331 training images
- Used in the demo



SVM Training and Optimization

- Matlab 2013a built-in SVM library
- 26 feature Multi-Class SVM
- Use the highest confidence result

Lessons

- Classification is fairly straightforward, but finding and collecting training and testing data can be difficult.
- Classification using multiclass SVMs

Results/Demonstrations

How well does it perform?