Project Outline

- In the cases where text captchas are easily segment-able the problem reduces to character recognition.
- In our case we focus on hand written digit recognition. Reasons:-

1. Hand written Digit recognition is a more general problem.
2. Time limit (including characters increase no. of classes and hence much better classification techniques are required)
3. With more classes understanding how things work becomes difficult.

Overview of Algorithm

- Training phase

  Step-1 : Apply PCA to extract features from the image training dataset.

  Step-2 : Train the classifier using the features extracted.

- Testing Phase

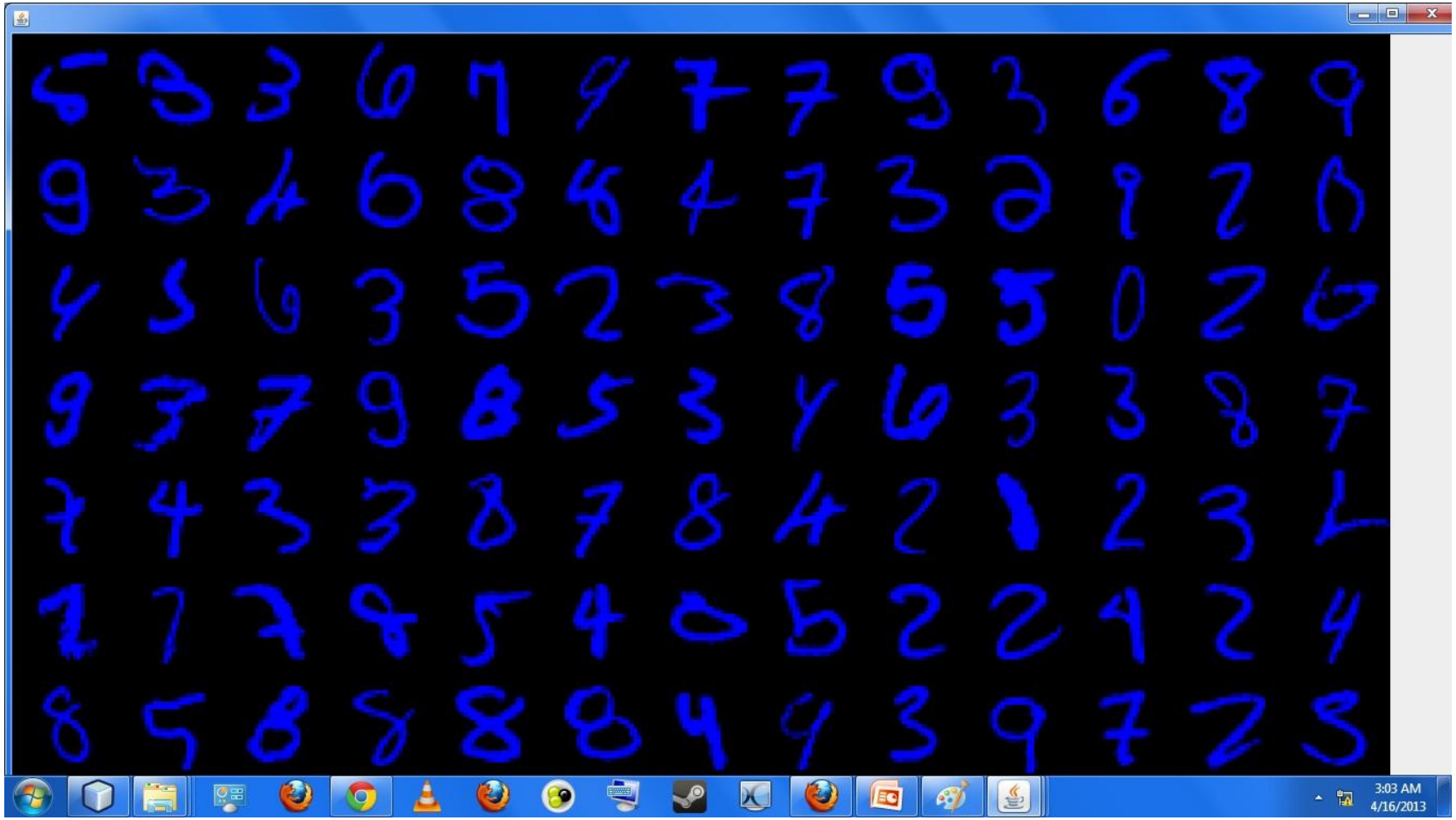
  Step-1: Find the components of test images along the features extracted in training phase.

  Step2 : Classify the vectors thus generated.

PCA

- Features were extracted so as to cover 80% variance. It required 110-125 features depending on the size of dataset.
- Images of features → x.jpg

A snapshot of dataset



Classifiers used and their error rate

- Neural Network (Multilayer Perceptron) : 3% (takes approx. 30 min. to train)
- SMO (Sequential Minimum Optimization) : 7% (takes approx. 2 min. to train)
- Naïve Bayes : 21%

Confusion Matrix(size of training set = 10000, classifier = SMO)

	0	1	2	3	4	5	6	7	8	9
0	82.0	0.0	0.0	0.0	0.0	2.0	1.0	0.0	0.0	0.0
1	0.0	124.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0
2	1.0	0.0	103.0	1.0	0.0	0.0	1.0	3.0	7.0	0.0
3	0.0	0.0	2.0	89.0	0.0	8.0	1.0	3.0	3.0	1.0
4	0.0	1.0	1.0	0.0	99.0	1.0	2.0	0.0	0.0	6.0
5	2.0	0.0	1.0	3.0	1.0	78.0	0.0	1.0	0.0	1.0
6	3.0	0.0	2.0	0.0	1.0	1.0	80.0	0.0	0.0	0.0
7	0.0	2.0	2.0	1.0	0.0	0.0	0.0	87.0	1.0	6.0
8	1.0	0.0	1.0	5.0	2.0	2.0	0.0	2.0	75.0	1.0
9	0.0	1.0	0.0	1.0	1.0	0.0	0.0	3.0	3.0	84.0

Errors

Some Drawbacks of PCA

- Fails to capture minute details like corners which are very essential information for us to classify digits

  This leads to many misclassifications like b/w 7 and 2

To touch accuracies close to 100% a much better feature extraction algorithm is needed.