HW4 Project Problem B for Math 420

In this problem you will analyze a collection of handwritten digits collected by the National Institute of Standards and Technology, see, e.g., the data at the page of Prof. S. Roweis:

`http://www.cs.nyu.edu/~roweis/data/mnist_all.mat`

These data consist of approximately 6000 training examples of each of 10 digits (0, 1, 2, 3, 4, 5, 6, 7, 8, 9) in the 8-bit grayscale format, along with 1000 test examples.

(i) Load the data. Make sure your sets are properly labeled. Plot and verify some examples. Represent each of these 28x28 images as a vector with 784 coefficients.

(ii) Find good compressed representations of these data sets. For your representations you can use e.g., PCA, FFT, or other numerical features defined directly from the images. How many vector components (i.e. mapped coefficients plus features) do you need to represent each of the data sets corresponding to individual digits? What are the reasons for differences in numbers of components required in these different types of representations ?

(iii) Use the performance of classification schemes such as nearest neighbors, based on your representation, as a tool to verify the success rate of your compression scheme, as well as to choose a number of coefficients that guarantees a success rate of at least 90%. Report your success percentages for each compressed digit separately, as well as globally.

(iv) Can you find some characteristics of your compressed datasets which would allow you increase the success rates of classification?