Assignment: A5

Due: 8 November 2012

You are to explore the use of PCA and k-means to find 26 clusters to classify scanned images of the 26 lower-case characters (i.e., a-z). Several aspects of this approach deserve careful attention:

- **Input vector**: choose an appropriate feature set
  - Should be d-dimensional vectors
  - Make sure that some features are correlated to see if PCA will eliminate them
- **Cluster Methods**:
  - Study performance as a function of mean versus median center calculation
  - Compare with respect to error and time performance
- **Stopping Criteria**: Study 3 of 4 at least
  - Simple algorithm (cluster centers don't change much)
  - Multiple runs to get best
  - Distance to Voronoi boundary
  - Inter- and intra-cluster distances
- **Initialization**: Study 2 of 3 at least
  - Pick k random points in the feature space
  - Pick k random points in the data set
  - Use centers of k largest spheres that can be packed in the hyper-parallelepided defined by data points
- **Data Management**:
  - Describe how you select training and testing data

In addition, the results need to be presented in a strong statistical framework; this means computing statistics (e.g., mean, variance) over several trials (how many?), and showing confidence intervals.

Finally, the analysis and interpretation are the essential parts of the report; use these to present your findings, understanding and remaining problems.

In this assignment, the major goal is to explore the use of k-means to find classes and PCA to reduce dimensions.

There is a set of sample images on the class data sub-directory.