OCR from unlabelled data

Supervisor: Ben Graham

The training data for optical character recognition systems (OCR) normally consists of labelled data: pairs of (sample character, character name) with many examples of each type of character. One of the interesting properties of the technique of contrastive convergence (CD) is the ability to learn from unlabelled data. The goal of this project is to use clustering algorithms, CD training, and other techniques, to group together like characters in order to do OCR, modulo a permutation of the labels. Experience programming in Matlab or Python is required.

References:
- MNIST http://yann.lecun.com/exdb/mnist/
- Hinton, G. E. and Salakhutdinov, R. R. (2006)
  Reducing the dimensionality of data with neural networks.
  Science, Vol. 313. no. 5786, pp. 504 - 507, 28 July 2006.
  http://www.cs.toronto.edu/~hinton/science.pdf

Figure: A projection of 60,000 handwritten characters into two dimensional space. The ten colours represent ten different classes of character: the digits 0,1,...9. The projection is constructed by CD training *without* any knowledge of which characters correspond to each class.