Datasets

12.0 RC2

We have, or have access to, some weird and wonderful datasets. They might be useful, or even suggest, individual projects. They're also available for personal projects 'just for interest'. See also CoolDatasets.com (mostly US centric), especially for a list of datasets suited to training machine learning systems.

Shuttle Radar Topography Mission

NASA's Jet Propulsion Laboratory (JPL) released the Shuttle Radar Topography Mission (SRTM) dataset, and we have a local cache. This contains spot heights (point altitude data) of the earth's surface at 30m granularity, worldwide. Sampled by the space shuttle programme in 2000, the dataset was released to the public in a sampled (low resolution) format. In 2015 the dataset was re-released in its original resolution, also with some gaps (voids) filled in with data from subsequent missions (voids were near the north and south poles, where it was impractical to fly the shuttle and telemetry equipment).

World Shorelines

The world shorelines dataset is a vector representation of the coastlines of the landmasses and major islands.

EDINA DigiMap

The UK Ordnance Survey have a number of products available free of charge for academic use. OS Places is a gazetteer (naming spacial regions, e.g. towns); there are coastline maps, contour maps, digital street maps, and the entire series of OS maps.

TrafficMaster

TrafficMaster speed sensors (the blue horn-shaped sensors you might have seen along A-roads, and gray tiles on motorway bridges) record the speed of traffic flow on the major roads across the UK and Europe. The data is presented as a live feed in a variety of formats.

Uber Movement

Uber have released anonymised traces of 2 billion Uber journeys (necessarily urban, for

The MNIST dataset and non-MNIST dataset are suitable for training and evaluating different kinds of classifier.

The MNIST database of handwritten digits, available from this page, has a training set of 60,000 examples and a test set of 10,000 examples. It is a subset of a larger set available from NIST. The digits have been size-normalized and centred in a fixed-size image. It is a good database for people who want to try learning techniques and pattern recognition methods on real-world data while spending minimal efforts on preprocessing and formatting.
— Taken from *MNIST website*

Labelled images of human faces

The Labelled Faces in the Wild and YouTube faces dataset are sets of images and video sequences in which human subjects appear.

The faces are labelled so you can tell which images contain the same person and which are different people. Again, these are suitable datasets for training and evaluating classifiers. The Labelled Faces in the Wild is a database of face photographs designed for studying the problem of unconstrained face recognition. The dataset contains more than 13,000 images of faces collected from the web. Each face has been labelled with the name of the person pictured. 1,680 of the people pictured have two or more distinct photos in the dataset. The only constraint on these faces is that they were detected by the Viola-Jones face detector.
— Taken from *Labelled Faces in the Wild website*

Musical data

music21 has some useful high-quality corpuses (Bach, folk music, hymns).

Music21 is a set of tools for helping scholars and other active listeners answer questions about music quickly and simply. If you've ever asked

tonal pitch structures or the form of minuets) if I could write a program to automatically write more of them," then music21 can help you.

— Taken from *music21 website*

Image segmentation dataset

The Berkeley Segmentation Dataset is also useful:

The Berkeley Segmentation Dataset provides an empirical basis for research on image segmentation and boundary detection. They collected 12,000 hand-labelled segmentations of 1,000 Corel dataset images from 30 human subjects. Half of the segmentations were obtained from presenting the subject with a colour image; the other half from presenting a greyscale image. The public benchmark based on this data consists of all of the greyscale and colour segmentations for 300 images. The images are divided into a training set of 200 images, and a test set of 100 images.

— Taken from *Berkeley Segmentation Dataset website*

US Government datasets

The US Government datasets contains 196,465 datasets on a broad range of topics including finance, housing, the economy and the climate.