

## Four face databases in matlab format

### Algorithms

---

(Matlab 7 or higher version is required to open these files)

---

We recomend to pre-process the data using either of the following two methods:

Nomalize each vector to unit

```
%=====
[nSmp,nFea] = size(fea);
for i = 1:nSmp
    fea(i,:) = fea(i,:) ./ max(1e-12,norm(fea(i,:)));
end
%=====
```

Scale the features (pixel values) to [0,1]

```
%=====
maxValue = max(max(fea));
fea = fea/maxValue;
%=====
```

---

If you used the processed data sets on this page, we appreciate it very much if you can cite our following works:

- Deng Cai, Xiaofei He, Yuxiao Hu, Jiawei Han and Thomas Huang, "Learning a Spatially Smooth Subspace for Face Recognition", CVPR'07. [Bibtex source](#)
  - Deng Cai, Xiaofei He and Jiawei Han, "Spectral Regression for Efficient Regularized Subspace Learning", ICCV'07. [Bibtex source](#)
  - Deng Cai, Xiaofei He, Jiawei Han, and Hong-Jiang Zhang, "Orthogonal Laplacianfaces for Face Recognition", IEEE TIP 2006. [Bibtex source](#)
  - Xiaofei He, Shuicheng Yan, Yuxiao Hu, Partha Niyogi, and Hong-Jiang Zhang, "Face Recognition Using Laplacianfaces", IEEE TPAMI 2005. [Bibtex source](#)
- 

- [Yale database](#)

Contains 165 grayscale images in GIF format of 15 individuals. There are 11 images per subject, one per different facial expression or configuration: center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised, and wink.

[32x32 Data File](#): contains variables 'fea' and 'gnd'. Each row of 'fea' is a face; 'gnd' is the label. [Sample images](#)  
[64x64 Data File](#).

A random subset with  $p(=2,3,4,5,6,7,8)$  images per individual was taken with labels to form the training set, and the rest of the database was considered to be the testing set. For each given  $p$ , there are 50 randomly splits:

[2 Train](#) | [3 Train](#) | [4 Train](#) | [5 Train](#) | [6 Train](#) | [7 Train](#) | [8 Train](#) |

Each split file contains variables 'trainIdx' and 'testIdx'. The following

matlab codes can be used to generate the training and test set:

```
%=====
fea_Train = fea(trainIdx,:);
fea_Test = fea(testIdx,:);

gnd_Train = gnd(trainIdx);
gnd_Test = gnd(testIdx);
%=====
```

[Experimental results \(new!\)](#) | [old experimental results](#)

- [ORL database](#)

Ten different images of each of 40 distinct subjects. For some subjects, the images were taken at different times, varying the lighting, facial expressions (open / closed eyes, smiling / not smiling) and facial details (glasses / no glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement).

[32x32 Data File](#): contains variables 'fea' and 'gnd'. Each row of 'fea' is a face; 'gnd' is the label. [Sample images](#)  
[64x64 Data File](#)

A random subset with  $p(=2, 3, 4, 5, 6, 7, 8)$  images per individual was taken with labels to form the training set, and the rest of the database was considered to be the testing set. For each given  $p$ , there are 50 randomly splits:

[2 Train](#) | [3 Train](#) | [4 Train](#) | [5 Train](#) | [6 Train](#) | [7 Train](#) | [8 Train](#) |

[Experimental results \(new!\)](#) | [old experimental results](#)

- [PIE database, CMU](#)

A database of 41,368 images of 68 people, each person under 13 different poses, 43 different illumination conditions, and with 4 different expressions.

This dataset only contains five near frontal poses (C05, C07, C09, C27, C29) and all the images under different illuminations and expressions. So, there are 170 images for each individual.

[32x32 Data File](#): contains variables 'fea' and 'gnd'. Each row of 'fea' is a face; 'gnd' is the label. [Sample images](#)

64x64 Data File: [Pose C05](#) | [Pose C07](#) | [Pose C09](#) | [Pose C27](#) | [Pose C29](#)

A random subset with  $p(=5, 10, \dots, 130)$  images per individual was taken with labels to form the training set, and the rest of the database was considered to be the testing set. For each given  $p$ , there are 50 randomly splits:

[5 Train](#) | [10 Train](#) | [20 Train](#) | [30 Train](#) | [40 Train](#) | [50 Train](#) | [60 Train](#) |  
[70 Train](#) | [80 Train](#) | [90 Train](#) | [100 Train](#) | [110 Train](#) | [120 Train](#) | [130 Train](#) |

[Random shuffled](#): Used in the paper "Efficient Kernel Discriminant Analysis via Spectral Regression", in ICDM'08.

Among the 11,554 images, the first 8,000 images are used as the training set

and the last 3,554 images are used for testing. We also run several cases by training all the algorithms on the first 2000, 3000, ... , 8000 images in the training set.

[Pose27\(lights change\)](#): Used in the paper "Laplacian Score for Feature Selection", in NIPS' 05.

[Pose27\(lights and illumination change\)](#): Used in the paper "Graph Regularized Non-negative Matrix Factorization for Data Representation", in T-PAMI 2011.

[Another version of Pose27\(lights and illumination change\)](#): Used in the paper "Spectral Regression: A Unified Approach for Sparse Subspace Learning ", in ICDM 2007.

[Experimental results \(new!\)](#) | [old experimental results](#)

- [Extended Yale Face Database B](#)

Please refer to [this link](#) for detailed information.

For this database, we simply use the cropped images (which can be downloaded [here](#)) and resize them to 32x32 pixels. This dataset now has 38 individuals and around 64 near frontal images under different illuminations per individual.

[32x32 Data File](#): contains variables 'fea' and 'gnd'. Each row of 'fea' is a face; 'gnd' is the label. [Sample images](#)

A random subset with  $p(=5, 10, \dots, 50)$  images per individual was taken with labels to form the training set, and the rest of the database was considered to be the testing set. For each given  $p$ , there are 50 random splits:

[5 Train](#) | [10 Train](#) | [20 Train](#) | [30 Train](#) | [40 Train](#) | [50 Train](#) |

[Experimental results \(new!\)](#) | [old experimental results](#)

---

Return to [Codes and Data](#)