第七次作业（大作业）

---

*ORL Database*是一个有名的人脸数据库。里面有40个ID的人脸，每个ID有10张图。本次作业是一个综合的project. 要求自己写PCA,马氏距离，k-means代码，不调用相关的库。

数据集的文件组织：有40个文件夹(s1,s2,...,s40)，每个文件夹是一个ID，每个文件夹（ID）有10张人脸(1.pgm, 2.pgm,..., 10.pgm)。

训练和测试集划分：训练集由每个文件夹中的1.pgm，2.pgm,...,6.pgm组成，共240张图。测试集由剩下160张图组成。

说明：这里的训练集也同时作为gallery，测试集是query。意思是，1）我们会用训练集来计算pca，同时在评测阶段，我们把训练集的图片作为检索库。2）举个例子来说，我们会用160张图的每一张作为query，依次去对照这240张图（gallery），看看哪张图最像，然后返回这张最像的图。

问题1：PCA降维。

每张图的大小是112* 92，我们把它拉成一个向量112* 92=10304维度，240张图就组成240 * 10304的矩阵。我们降维到160* 40维度的矩阵，也就是特征10304维度降到了40维度。同时保存PCA的投影矩阵和均值。

我们利用训练集计算的投影矩阵和均值，对测试集的每一张图（10304维度）降维到40维度。得到160*40矩阵。160是图像的个数，40是维度。

问题2：人脸识别和匹配。

我们根据计算得到的训练集（240 * 40）和测试集（ 160 * 40）进行匹配。注意到训练集在这里将作为gallery库。

1）用欧式距离匹配。

我们用每个测试集图（query）的40维特征去对比gallery库中的40张图。计算哪张图欧氏距离最近，返回哪一张图。计算返回的图和该张图的ID是否一样。最终，我们计算160张图中，正确检索的图的个数(collect_num)。计算准确率collect_num/160.

2）用马氏距离匹配。

类似1），这里我们使用马氏距离，对训练每个ID求均值和协方差，假设每个ID的协方差都不一样。利用马氏距离归类于某个ID。

问题3：聚类分析。

聚类分析，我们利用k-means。在这一问，我们不分训练和测试，直接利用400张图进行聚类分析。我们分别利用降维前和降维后的特征进行聚类，并利用第四次作业的聚类评测标准进行评测。