模式识别实验六

实验人：叶平

实验内容：

(a)、用 PCA 方法对实验图像设计分类器并完成训练和分类过程，统计正确分类率,

其中求解特征值和特征向量的方法分为一般方法和使用技巧[1]的方法，比较二者的运行时间和正确分类率；

(b)、用 MDA 方法对实验图像设计分类器并完成训练和分类过程，统计正确分类率。

实验图像库为 ORL 人脸图像库，共 40 人，每人 10 幅图像，其中每人的前 5 幅作为训练样本，后 5 幅作为测试分类样本，统计正确分类率。分类准则为最近邻规则。

(c)、用距离保持的降维法(DPDR)[2]进行同样的实验并与 PCA 比较。

(d)、考察 PCA 和 DPDR 的外推能力，即设 TrSet 和 TeSET 分别为训练和测试数据集，

现在 step1: 用 TrSet 获得投影阵 M，用其重建 TeSET，计算重建误差 ETE,

Step2: 用 TrSet+TeSET 获得投影阵 M+,用其重建 TeSET，计算重建误差 ETE+,

Step3: 比较 ETE 和 ETE+，你能获得何种发现？

[1] see “模式识别” 2nd Edition 2000 年

（清华出版社, Chapter 9, Section 9.9, pp. 223-228）

PCA： $XX^TU = U\Lambda$ ➔ $X^TXX^TU = X^TU\Lambda$ ➔

$X^TXV = V\Lambda$，其中 $V = X^TU$

Note： Scale reduction of XX^T to X^TX: DxD ➔ nxn

For example, (112x92)x(112x92) ➔200x200 (the size of the training set)

ORL available at http://parnec.nuaa.edu.cn

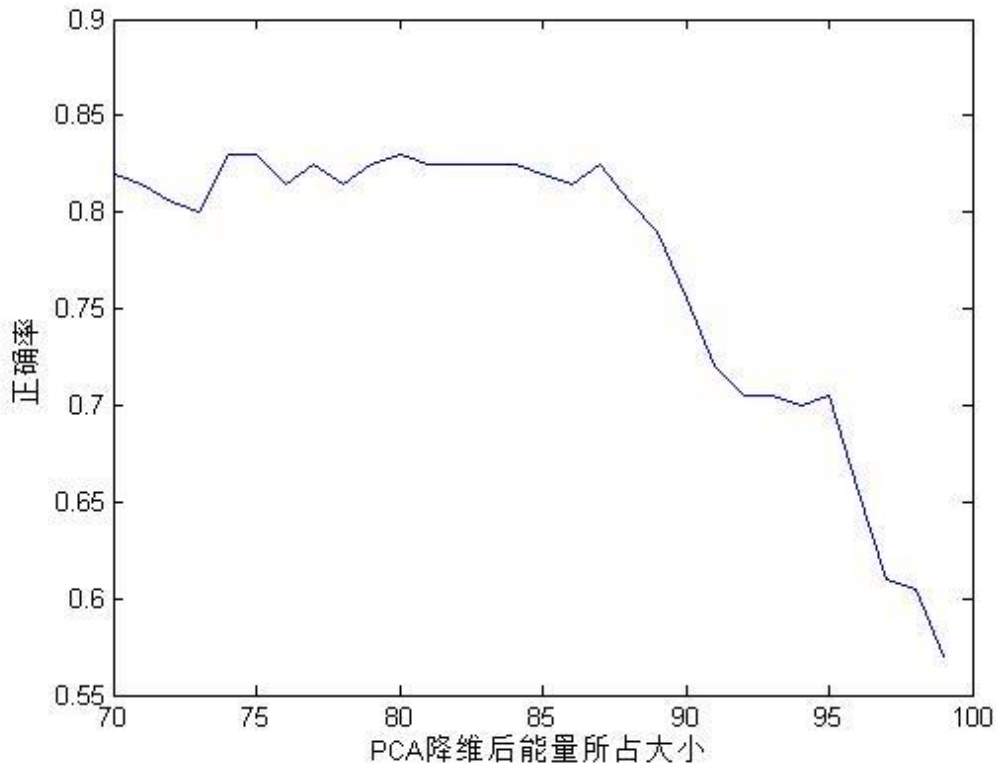
[2] Hyunsoo Kim, Haesun Park, Hongyuan Zha, Distance Preserving Dimension Reduction Using the QR Factorization or the Cholesky Factorization, available by google (scholar)

实验结果：

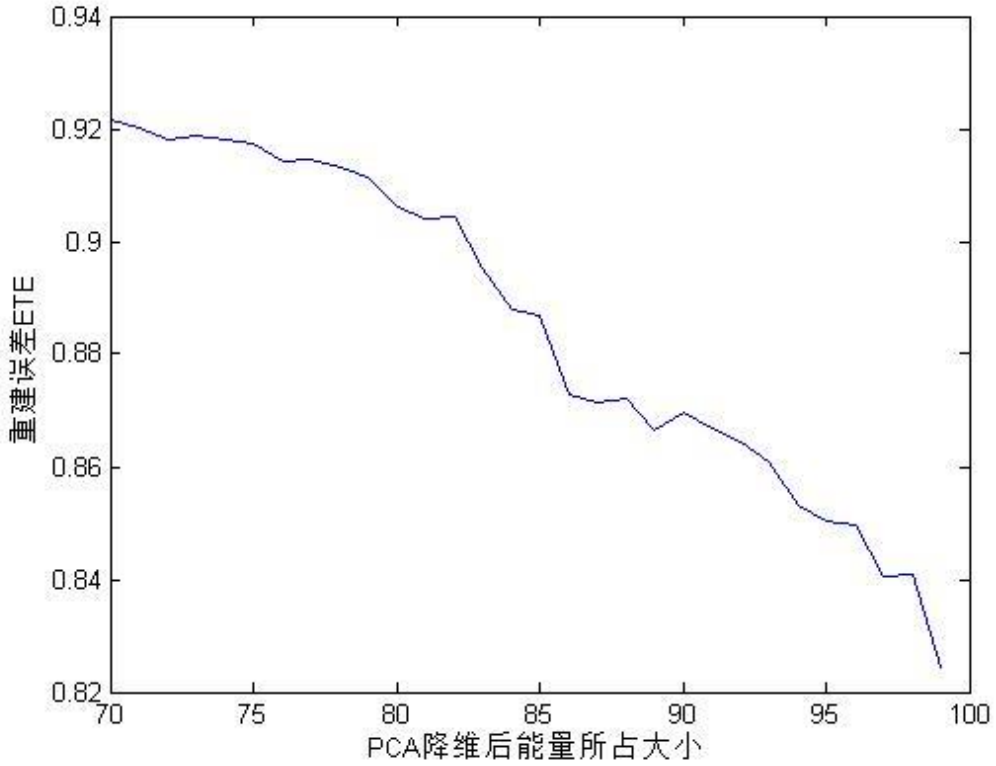
在 Command Window 下运行实验 6 文件夹下 Res.m 文件。其中 Res 函数有一个参数，当不填时运行结果如下：

从图中可得到 MDA 和 DPDR 的分类正确率，由于特征维数过大(10304 维)，MDA 如果对原数据直接处理时会产生维数灾难，使得分类正确性大大降低，所以在 MDA 处理之前，采用了 PCA 降维。此外，由于维数过大，在 DPDR 算法中进行 QR 分解时，总是报错，提示内存溢出，无法使得实验进行下去，故在 DPDR 算法进行之前，也对数据进行了 PCA 降维处理。

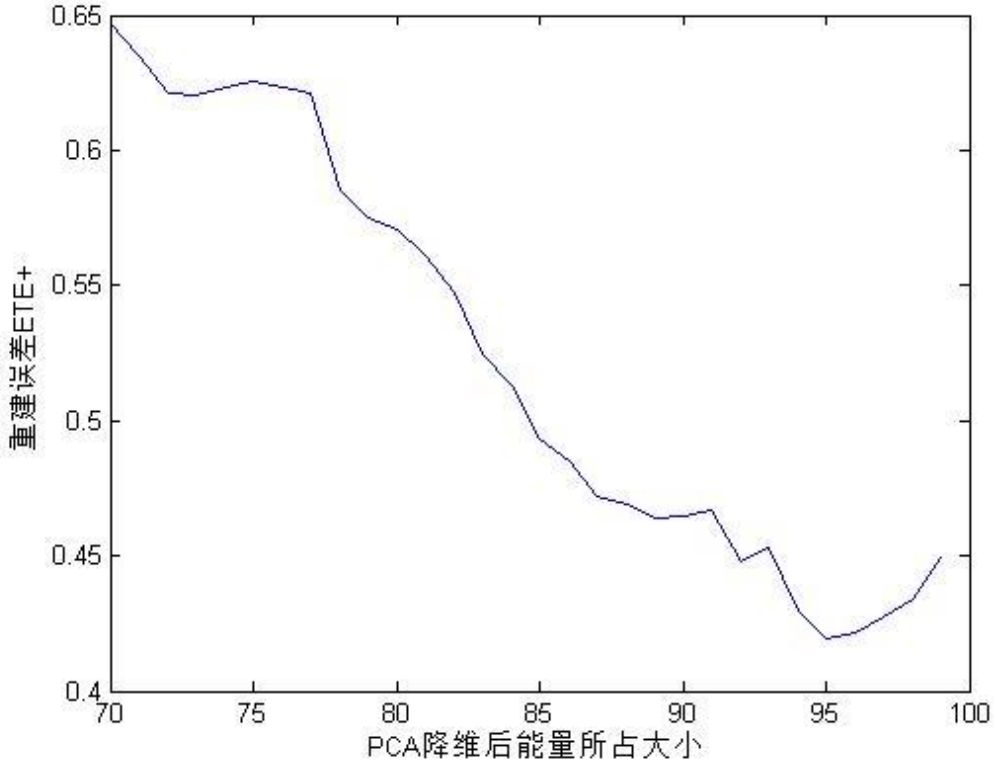
在命令行下输入 Res（1）,可以得到关于 PCA 算法的结果，如下所示：



PCA 降维后的分类正确率



PCA 的重建误差 ETE



PCA 的重建误差 ETE+

对于重建误差的度量我采取了如下公式：

$$\text{Error} = \frac{\|TeSET - TrSet\|}{\|TeSET\| + \|TrSet\|}$$

其中$\| \ \|$代表矩阵范数，由范数的性质可以知道，Error 在 0-1 之间。如果重建的 TeSET 与原 TrSet 越接近，则 Error 越趋向 0.

由上图可以看出，数据经过 PCA 和 DPDR 降维后，用 TrSet 获得投影阵 M，用其重建 TeSET，得到的重建误差 ETE+较小。随着 PCA 成分能量的增大，重建误差也随之变小。但是到了 95%-100%的时候，重建误差 ETE+还略有上升。此外，在分类中，随着 PCA 成分能量的增大，分类误差反而增大。这是由于 PCA 降维后的数据最能代表原始数据，但是它不一定会保存对分类有益的信息。