**深 圳 大 学 实 验 报 告**

|  |
| --- |
| **课程名称 机器学习**  **项目名称 实验一：PCA算法**  **学 院 计算机与软件学院**  **专 业 软件工程**  **指导教师 赖志辉**  **报 告 人 唐健龙 学号 2018132100**  **实验时间 2020年3月9日至2020年3月23日**  **实验报告提交时间 2020年3月19日** |

**教务处制**

# 一、实验目的与要求

**1、实现PCA算法的人脸重构，即用20,40,60,80,...,160个投影来重构图像的效果。**

**2、实现PCA算法的人脸识别，给出 10,20,30,...,160维的人脸识别识别率。**

**3、用PCA用来进行人脸图像降维，实现3个不同数据集多个子集的二维和三维空间实现数据的可视化。**

# 二、实验内容与方法

**原理：**

经典的PCA算法通过最大化数据散度以移除不必要的信息，减少特征相关性或滤除噪音，因此我们需要找到一个低维度的变换矩阵，使得原数据经过空间变换矩阵变换至给定的较低的维度，从而能够在保留主要信息、减少特征相关性和和剔除噪音的情况下，方便地进行数据的运算使用

**推导**：

假定标准正交基变换矩阵为 ，单个图片矩阵为 ，我们将其延展为行向量 ，首先通过上述处理，我们得到包含多个图片行向量的向量组 ，其中 为图片样本数，，先计算列向量内元素均值组成的矩阵

为了使得投影后的特征之间的相关性尽量小，数学上可令两个列向量的协方差为0

为了后续方便计算，首先进行中心化处理，将 的所有列向量减去对应均值，即令

为了方便讨论仍将归一化后的向量组记为 ，我们首先观察归一化后的 的协方差矩阵

可以发现，当 为对角矩阵时，恰好满足要求，但实际上我们希望的是在矩阵变换后，其协方差矩阵为对角矩阵，因此我们尝试进行变换

因此我们希望 为对角矩阵，我们继续进行变换，就有

很巧的是，根据SVD奇异值分解，我们可以将任意矩阵 转换为满足下式

其中 为主对角线全为非负数的对角矩阵， 为标准正交变换基，因此，我们只要将 进行奇异值分解，，就可以得到下式

或者

我们发现，此时令

就可以将变换后的向量组 的协方差矩阵 变为对角矩阵 ，在将该对角矩阵的特征值进行从大到小排列后，我们可以知道其特征值越大，对应的特征向量在降投影维中保留信息越重要，散度也越大，在给定降维的数字 后，我们重新取

其中 为 的第 列（ 中特征向量已按照特征值大小排序），从而可以通过将原向量组投影至该新向量组组成的 维基，从而达到降维的目的，如下所示

此处的 即为降维后的数据

若我们希望重构图像，只需要再通过 还原回原维度，如下所示

其中的 即为重构数据

# 三、实验步骤与过程

**1、实现PCA算法的人脸重构，即用20,40,60,80,...,160个投影来重构图像的效果。**

(1) 获取数据集，进行数据预处理（每张图拉成行向量并组合成向量组）

(2) 中心化数据，为防止数据运算过大导致灰度值溢出，先暂时减小灰度值倍数

(3) 通过原理中所叙述的PCA算法，计算变换矩阵和重构数据

(4) 选取不同的 进行人脸重构（需要还原倍数）

(5) 作出原图人脸和重构人脸进行比较（实验效果已置于结论处）

**2、实现PCA算法的人脸识别，给出 10,20,30,...,160维的人脸识别识别率。**

(1) 获取数据集，进行数据预处理（每张图拉成行向量并组合成向量组）

(2) 为每个图片行向量添加标签用以后续进行判断人脸识别是否正确

(3) 针对不同大小的训练集进行识别率测试

(4) 将划分出的训练集进行中心化，并通过PCA计算变换矩阵

(6) 对测试集变换后，进行识别测试，并统计识别率

(7) 作出关于投影向量数和识别率的曲线（实验效果已置于结论处）

**3、用PCA用来进行人脸图像降维，实现3个不同数据集多个子集的二维和三维空间实现数据的可视化。**

(1) 获取数据集，进行数据预处理（每张图拉成行向量并组合成向量组）

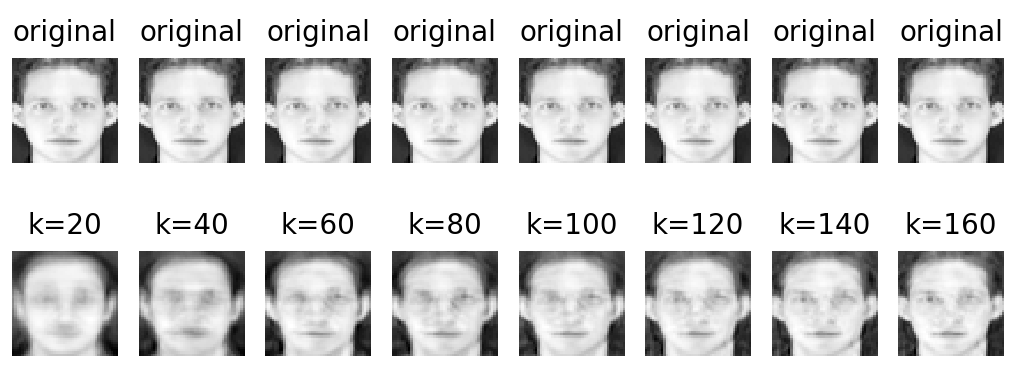
(2) 进行数据中心化，并通过PCA计算变换矩阵

(3) 选择部分样本并选择不同颜色表示不同的人

(4) 作出2D，3D散点图（实验效果已置于结论处）

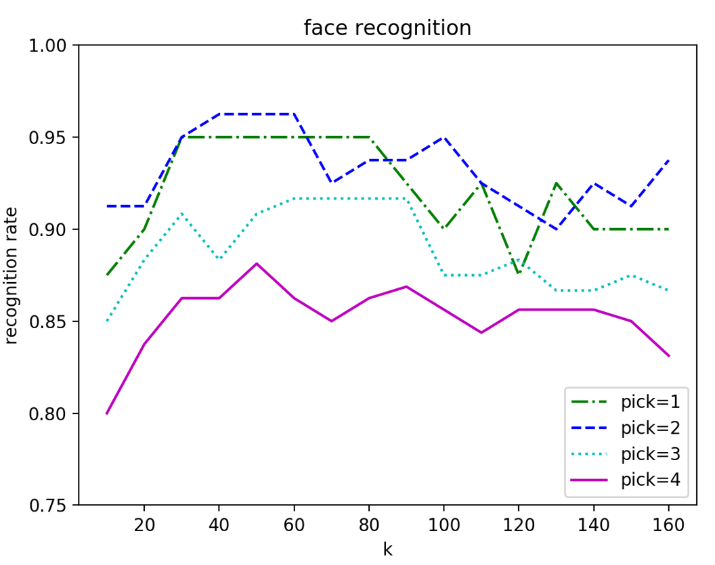
# 四、实验结论或体会

1. 人脸重构实验结果如下



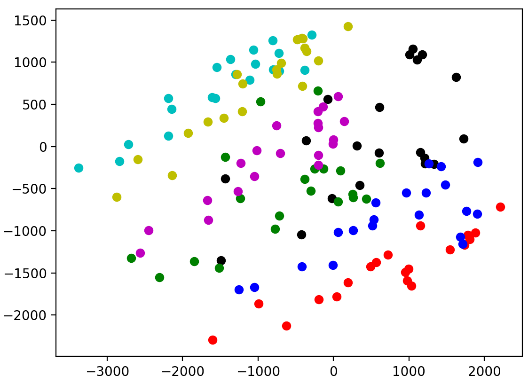
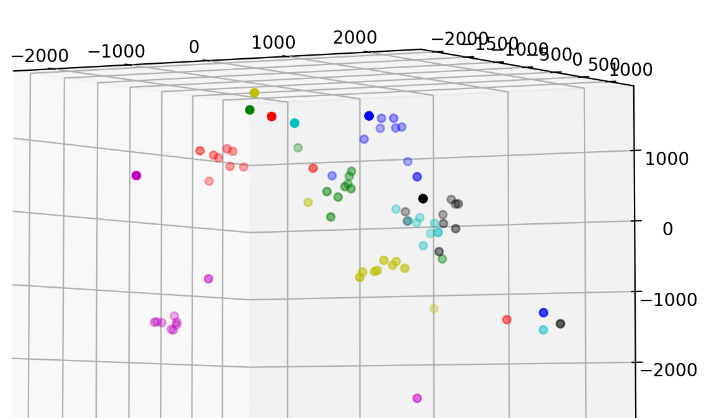
可以看到，随着投影向量数的增加，人脸的还原程度越来越高，同时，图像细节也逐渐增加，但是这也代表着，除了人脸以外的不必要特征也增加了，这就无法起到提取主特征的中心思想了，因此我们需要选择适当的投影向量数，防止过拟合或错过主特征

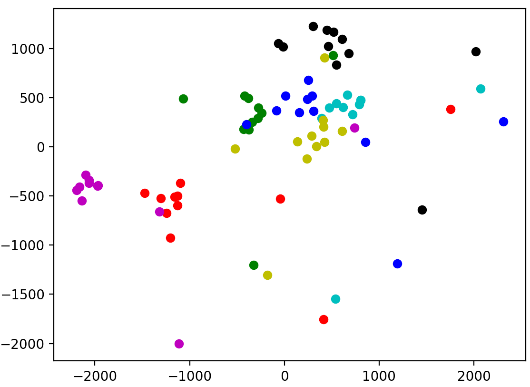
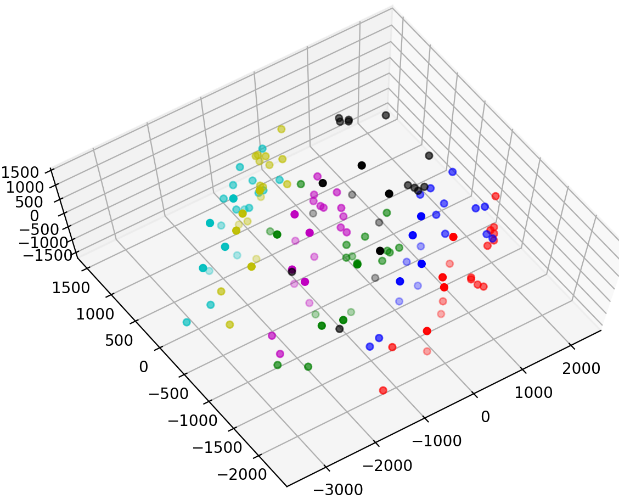
2. 人脸识别试验结果如下

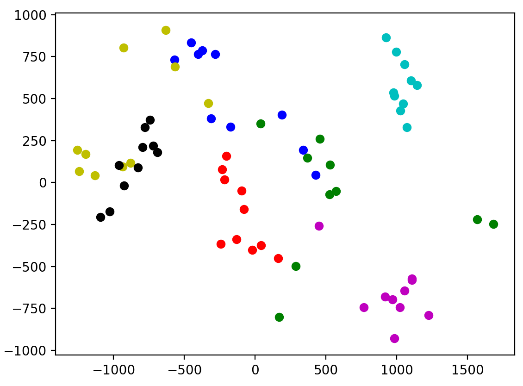
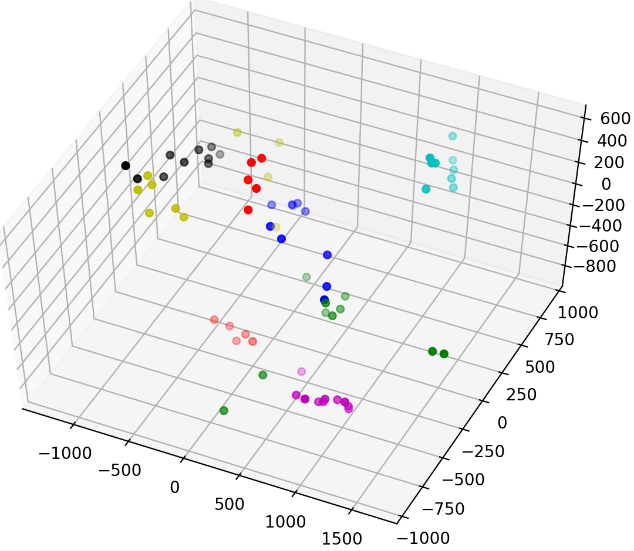


可以看到，投影向量数的增加并不能够不断地提高人脸识别的识别率，否则适得其反，原因正如上面所述，PCA专注于提取主特征，而不是充分拟合特定的训练集数据；同时，我们还发现，当抽取的测试集越来越大（此处对应到pick的大小，且这时训练集会对应地越来越小）的时候，识别率也发生下降，因此，条件允许情况下，需要保证样本数足够，否则不一定能够较好地进行降维识别

3. 人脸可视化实验结果如下

可以看到几个数据集即使在只保留2个到3个投影向量的情况下，仍然能够较好地保留原始数据的主要信息（对应于同颜色的点相对聚集），同时这也有利证明了PCA的成分分析能力出众，以至于它成为机器学习中非常重要的基础算法

这也可以证明

|  |
| --- |
| 指导教师批阅意见：  成绩评定：  指导教师签字：  年 月 日 |
| 备注： |

注：1、报告内的项目或内容设置，可根据实际情况加以调整和补充。