# 

# 《数据科学与工程算法基础》实践报告

报告题目：基于主成分分析的图片压缩结果及相关案例研究

姓 名： 章可儿

学 号： 10182100208

完成日期： 2020.11.30

提供中英文摘要

摘要 [中文]：

这篇报告主要是关于利用主成分分析的方法对100张卫星图片进行压缩的过程和结果分析。具体实现方法有两种，对每一张图片做PCA降维和对100张图片形成的大矩阵做PCA降维，这两种方式各有优劣，在报告中也会详细体现。在实践PCA降维技术的过程中，使用python语言读取图片、处理图片并重构图片，很好地保留了图片的内容，且有一定的压缩率、一定的空间节省和较小的重构误差。使用不同的预处理方法、不同的优化方法或者选择不同的参数，都可以得到不同的压缩性能。报告的最后有关于100张图片或300张图片一起压缩的结果分析和关于在压缩效果不变的情况下降低压缩率的一点猜想。

Abstract [英语]:

This report is mainly about the process and result analysis of compressing 100 satellite images using principal component analysis. There are two specific implementation methods, PCA dimensionality reduction for each picture and PCA dimensionality reduction for a large matrix formed by 100 pictures. These two methods have their own advantages and disadvantages, which will be detailed in the report. In the process of practicing PCA dimensionality reduction technology, the python language is used to read pictures, process pictures and reconstruct pictures, which well preserves the content of the pictures, and has a certain compression rate, certain space saving and small reconstruction error. Different compression performance can be obtained by using different preprocessing methods, different optimization methods, or choosing different parameters. At the end of the report, there is an analysis of the results of compressing 100 pictures or 300 pictures together and a little conjecture about reducing the compression rate while the compression effect is unchanged.

## 一、项目概述（阐明该项目的科学和应用价值，以及相关工作进展并描述项目的主要内容）

项目基本完成目标，对100张图片有较好的压缩效果（包括人肉眼可观察到的图片效果、较高的压缩率、较低的重构误差、较大的空间节省和相关的案例研究）。在这个过程中我了解了PCA降维技术的诸多细节和应用价值。

## 二、问题描述（问题定义）

主成分分析（PCA）可用于减少矩阵（图像）的维度，并投影这些新维度以重构保留其品质但k权重较小的图像。从数据集中的单个类别中选择100张图像，然后使用PCA压缩图像。在最终报告中说明图像压缩的性能，例如重构误差，节省空间，压缩率和案例研究等。

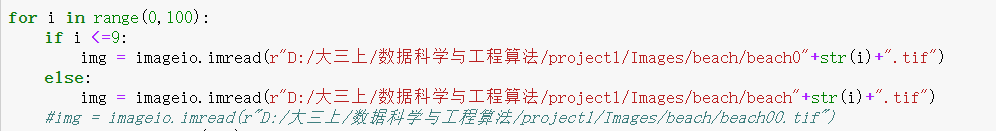
## 三、方法（问题解决步骤和实现细节）

### (1)对每一张图片做PCA降维的方法

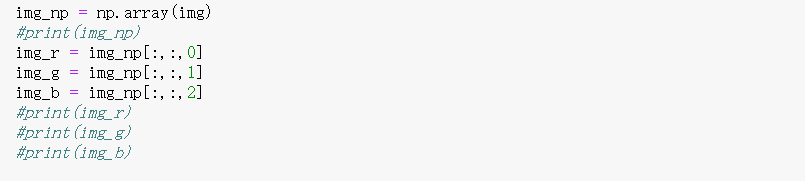
PCA降维过程中使用到的python包有numpy和imageio



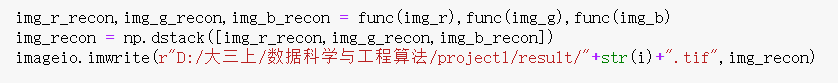
对于一类图片的100张图片，使用一个for循环读取图片（这里我选择的是beach类图片）：



将图片转化为numpy里的矩阵，因为是彩色图片，分成R、G、B三个通道的矩阵：

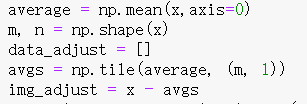


将三个矩阵分别传入func函数进行PCA降维，函数返回降维后的R、G、B矩阵，将它们重构为彩色图片并输出到文件夹：



func矩阵的实现过程是：

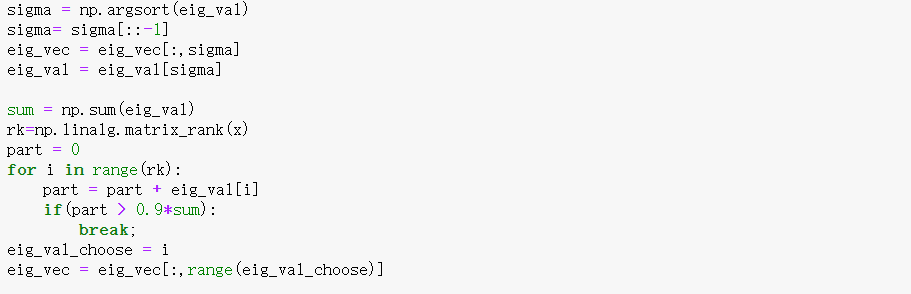
首先对矩阵去中心化



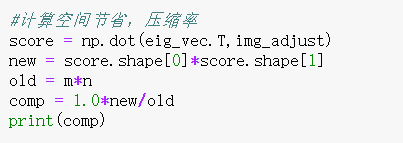
计算去中心化后的矩阵的协方差矩阵，并求出特征值和特征向量



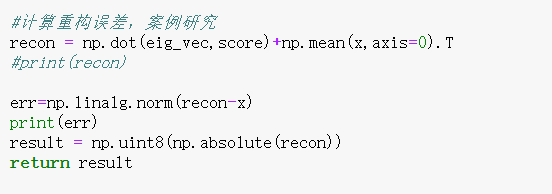
再次，将协方差矩阵的特征值和特征向量按照从大到小的顺序排列，并选择前k个主成分：（sigma是排列后的特征值，sum是特征值的总个数，当选择的特征值大于0.9\*sum时，退出循环，这里的0.9可以改成不用的值改变压缩性能）



通过新矩阵和旧矩阵的维度比计算压缩率，这里得到并输出的是1-压缩率：



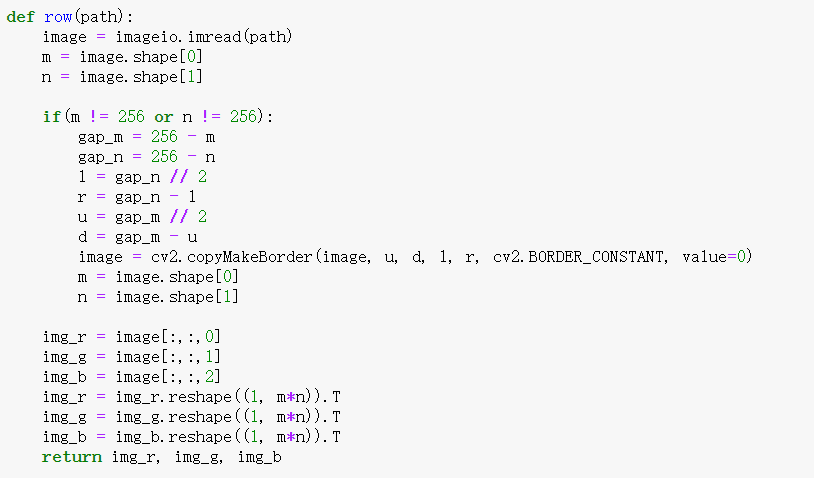
通过计算重构矩阵和原矩阵的欧式距离计算重构误差，并返回符合图像要求的重构矩阵



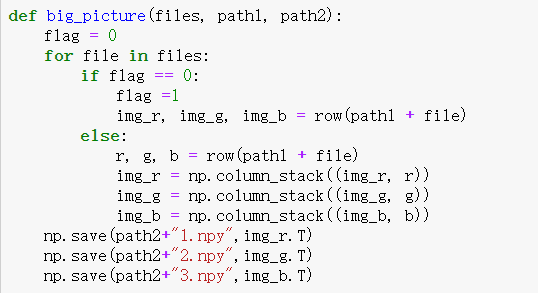
### (2)对100张图片的大矩阵做PCA降维的方法

PCA处理与(1)中方法类似，不再赘述。重点是把每一张图片拉成一个行向量构造这个大矩阵，以及降维之后再把行向量重构成图片。

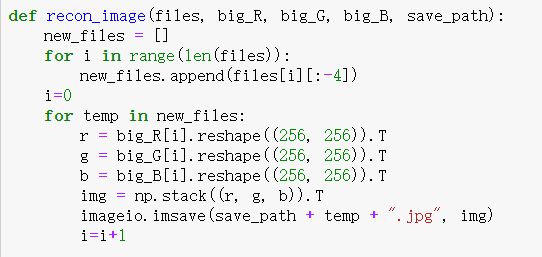
将图片拉成行向量：



行向量累加，形成大矩阵：



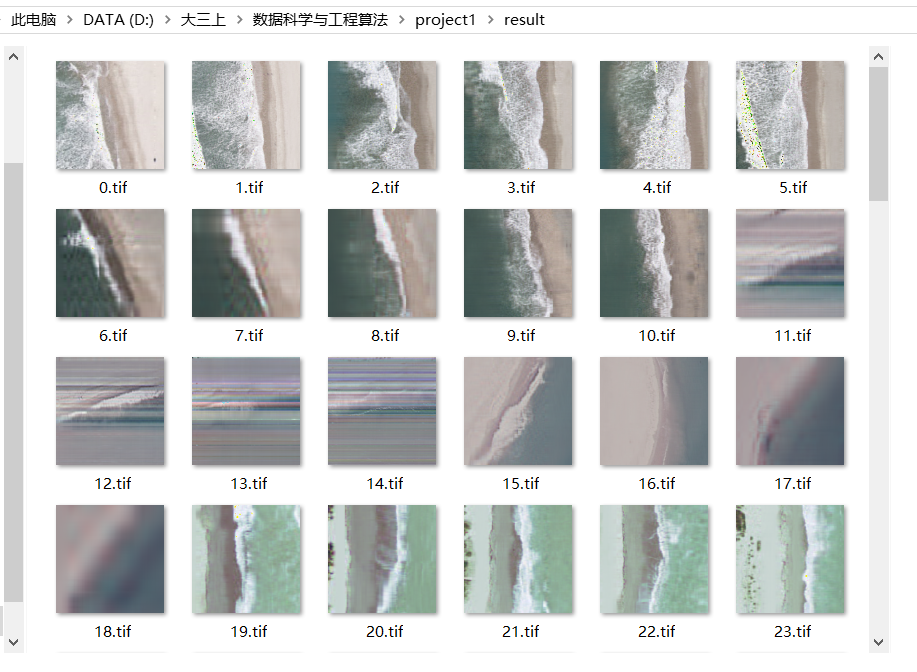
重构100张图片：

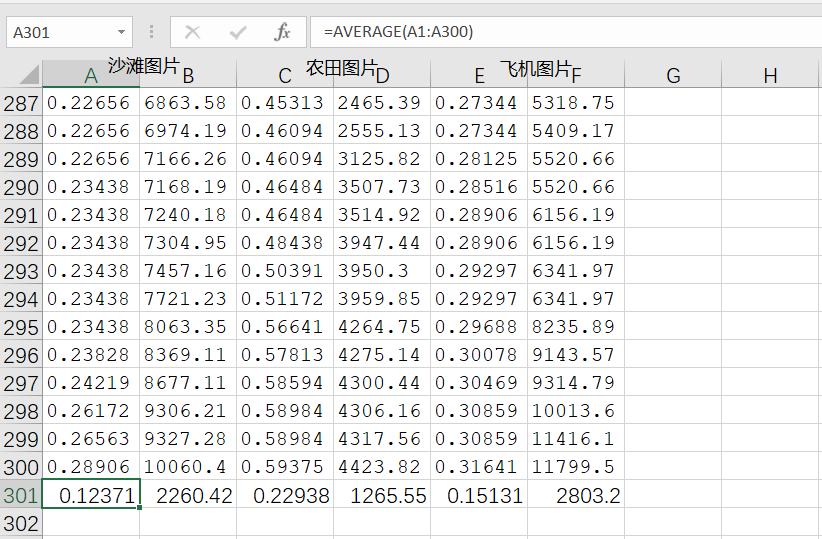


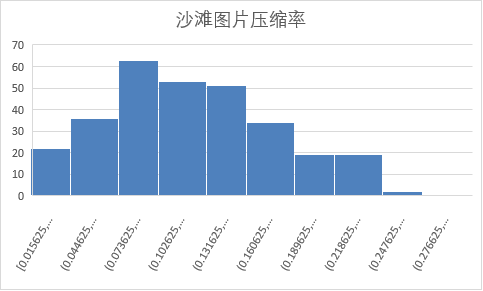
## 四、实验结果（验证提出方法的有效性和高效性）

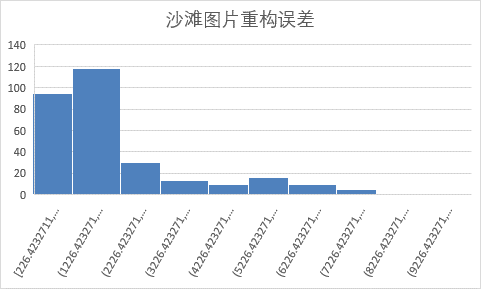
### (1)对每一张图片做PCA降维的结果

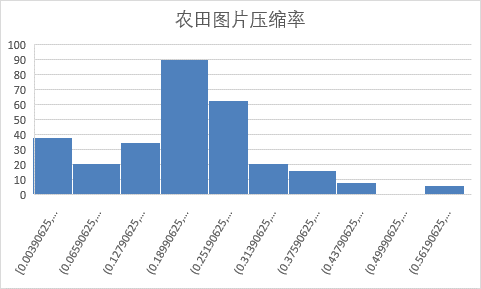
结果：result文件夹得到了压缩后的100张图片。我对三个类型的图片都尝试压缩了，得到的每一张图片的压缩率和重构误差储存在下方的表格中（根据压缩率的定义，表格中应该是1-压缩率），求出了每100张图片压缩率、重构误差的均值，并给出了这些数据分别的频率分布直方图。这是k=0.95的结果，k如果减小，压缩率会增大，空间节省增大，重构误差增大。(如果将100维的数据压成10维，压缩率为90%)。除此之外，空间节省约为19.7MB-18.7MB=1MB。

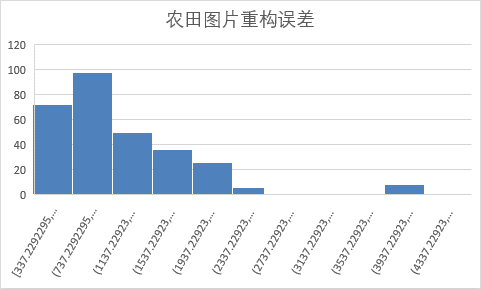


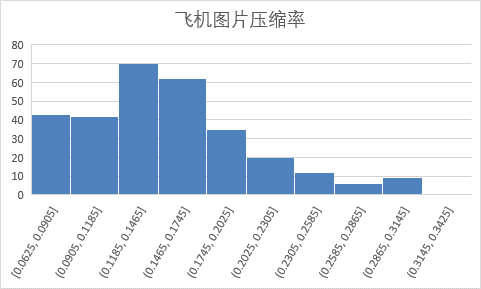


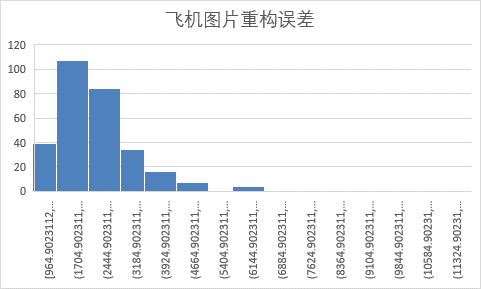




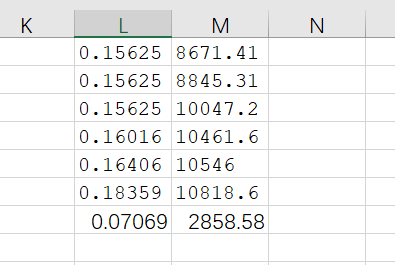








对沙滩图片，将k改为0.9后，压缩率和重构误差的均值是



可见压缩率增大，重构误差增大。

空间占用：

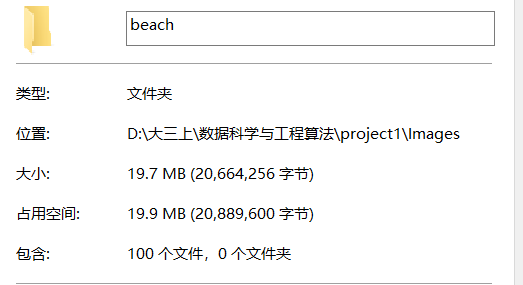


### (2) 对100张图片的大矩阵做PCA降维的结果

结果：如下图是100张沙滩图片的结果。空间节省约19MB，远高于第一种方法。

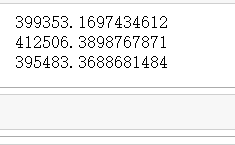




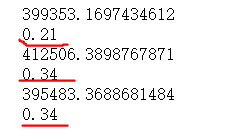


下图是R、G、B三个通道总共的重构误差，可以计算出均值：

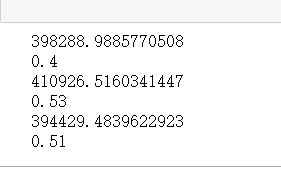
(399353.1697434612+412506.3898767871+395483.3688681484)/300=4024.473，可见平均的重构误差比第一种方法大。



下图是计算得到的压缩率，可以算出整个图片的压缩率为1-(0.21+0.34+0.34)/3=0.703。



将k设置为0.95后，得到的压缩率和重构误差如下：



压缩率、重构误差均值分别为1-(0.4+0.53+0.51)/3=0.52

(398288+410926+394429)/300=4012.14

空间节省如下图：



约为19MB

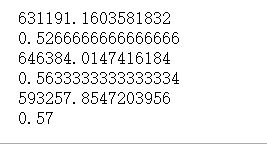
## 五、结论（对使用的方法可能存在的不足进行分析，以及未来可能的研究方向进行讨论）

总结：

|  |  |  |  |
| --- | --- | --- | --- |
|  | 压缩率均值 | 重构误差均值 | 空间节省 |
| 对每一张beach图做  K=0.95 | 0.87629 | 2260.42 | 1MB |
| 对每一张agricultural图做  K=0.95 | 0.77062 | 1265.55 | 1MB |
| 对每一张airplane图做  K=0.95 | 0.84869 | 2803.2 | 1MB |
| 对每一张beach图做  K=0.9 | 0.92931 | 2858.58 | 1MB |
| 对100张beach图做  K=0.9 | 0.703 | 4024.473 | 19MB |
| 对100张beach图做  K=0.95 | 0.52 | 4012.14 | 19MB |

**综上所述，对每一张图做PCA降维和对100张图同时做PCA降维，在k相同时，前者的压缩率更高，重构误差更低，节省空间更少；若PCA方法相同，则较高的k会得到较低的压缩率，较低的重构误差和几乎没有变化的空间节省。**

未来可能的研究方向：考虑高老师上课所说的关于100张同类图片一起做PCA降维和300张图片一起做PCA降维，修改代码，得到文件”算法-多张图片压缩1.ipynb”。它得到的压缩率、重构误差和空间节省(对300张图片除以3，以便比较)分别为0.4468、(6236.1)/3和18.68MB





可见300张图片一起做PCA降维时的压缩率远小于100张图片一起做的压缩率，猜测是因为沙滩、农田、飞机是不同类型的图片。因此得到关于得到一堆未知图片时提高压缩率的方法：对这些图片先做聚类，再压缩，可以在不影响压缩效果的情况下提高压缩率。