# 《数据科学与工程算法基础》实践报告

报告题目： 基于PCA算法对图像压缩处理

姓 名： 杨东东

学 号： 51194507017

完成日期： 2020-06-23

摘要：

图像的压缩处理是一种很常见的场景，可以在保留重要信息的情况下占用更少的存储空间或传输带宽，本文使用大家比较熟悉的PCA算法，又叫主成分分析，对一批航拍图像进行压缩。算法的本质就是找一些投影方向，使得数据在这些投影方向上的方差最大，而且这些投影方向是相互正交的，主成分分析是最为简单粗暴的一种数据降维方式。最后，对压缩效果做一些简单的统计评估。

Abstract:

Image compression processing is a very common scenario, which can occupy less storage space or transmission bandwidth while retaining important information. In this paper, a batch of aerial images are compressed using the familiar PCA algorithm, also called principal component analysis. The essence of the algorithm is to find some projection directions so that the data has the greatest variance in these projection directions, and these projection directions are orthogonal to each other. Principal Component Analysis is the simplest and roughest way to reduce the dimension of data. Finally, a simple statistical evaluation of the compression effect is made.

1. 项目概述

随着多媒体和通信技术的快速发展，多媒体信息的传输对数据的存储和传输提出了更高的要求，也给现有的有限的带宽以严峻的考验，特别是具有庞大数据量的数字图像通信，更难以传输和存储。图像压缩的目的就是把原来较大的图像用尽量少的字节来表示和传输，并且要求压缩后的图像有较好的质量。

本文采用PCA算法来对图像进行压缩处理，图像来自无人机航拍，观察数据集内容，这些图片的的原用途应该是做分类用的，本文不关心这些图片的内容，只关心压缩前后的数据对比。

1. 问题描述

主成分分析（Principal Component Analysis，PCA）， 是一种统计方法。通过正交变换将一组可能存在相关性的变量转换为一组线性不相关的变量，转换后的这组变量叫主成分。

在许多领域的研究与应用中，往往需要对反映事物的多个变量进行大量的观测，收集大量数据以便进行分析寻找规律。多变量大样本无疑会为研究和应用提供了丰富的信息，但也在一定程度上增加了数据采集的工作量，更重要的是在多数情况下，许多变量之间可能存在相关性，从而增加了问题分析的复杂性，同时对分析带来不便。如果分别对每个指标进行分析，分析往往是孤立的，而不是综合的。盲目减少指标会损失很多信息，容易产生错误的结论。

因此需要找到一个合理的方法，在减少需要分析的指标同时，尽量减少原指标包含信息的损失，以达到对所收集数据进行全面分析的目的。由于各变量间存在一定的相关关系，因此有可能用较少的综合指标分别综合存在于各变量中的各类信息。主成分分析就属于这类降维的方法。

1. 方法

图像数据集的尺寸是大小不一的，首先把图像尺寸统一标准化，尽管我觉得这一步可以省略。先遍历数据集里面的图像长和宽，得到平均值，在（940，610）附近；之后再次遍历数据集，resize图像并保存到新的目录：



图3-1 原图像尺寸



图3-2 标准化后尺寸

前期工作处理完毕，开始做压缩处理。

（1）分别求每个维度的平均值，然后对于所有的样例，都减去对应维度的均值，得到去中心化的数据。这一步也非必要，对于去中心化的数据和不去时候的数据，特征值大小会不同，但是相对大小并不会改变；

（2）求协方差矩阵C：用去中心化的数据矩阵乘上它的转置，然后除以(N-1)即可，N为样本数量；

（3）求协方差的特征值和特征向量；

（4）将特征值按照从大到小排序，选择前k个，然后将其对应的k个特征向量分别作为列向量组成特征向量矩阵。此处提供两种选择维度的手段，第一种，通过设定方差的百分比来计算将数据降到多少维，或者说保留多少主成分，第二种，直接指定保留的主成分个数，因为图像是彩色的，对图像三通道数据都做一遍PCA处理，如果指定方差百分比来设定主成分保留，可能各个通道保留的主成分数会不同，不方便后面压缩指标的计算；

（5）将样本点从原来维度投影到选取的k个特征向量，得到低维数据；

（6）通过逆变换，重构低维数据，进行复原；

（7）将处理后的图像写入新目录。

1. 实验结果

PCA处理后的效果：



图4-1原图像

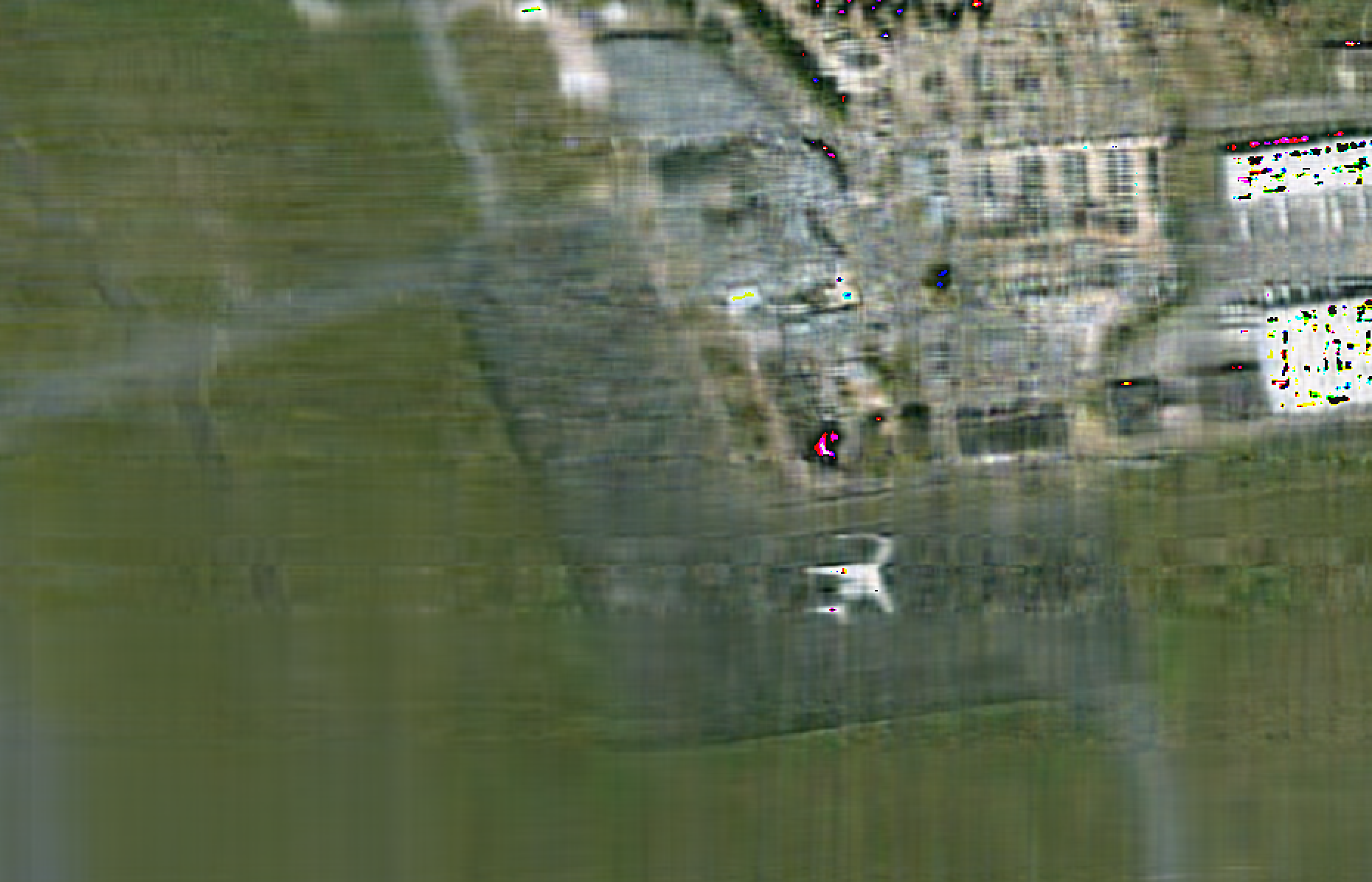


图4-2 主成分保留20

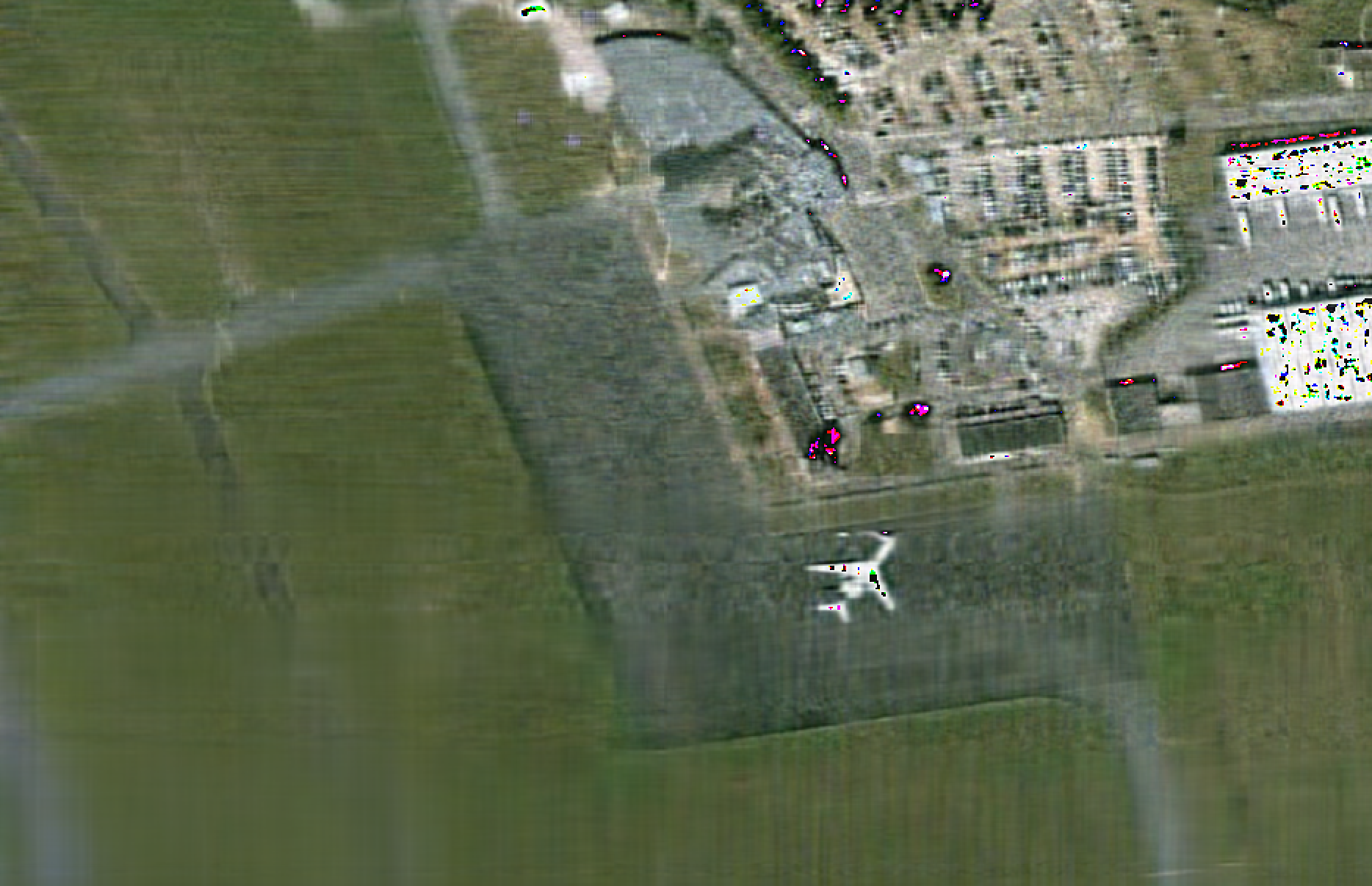


图4-3 主成分保留40



图4-4 主成分保留200

部分图像大小处理前后对比：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 图像 | 原大小 | 主成分20个 | 主成分40个 | 主成分60个 |
| 001.jpg | 78KB | 61KB | 72KB | 78KB |
| 002.jpg | 88KB | 52KB | 64KB | 73KB |
| 003.jpg | 94KB | 65KB | 77KB | 84KB |
| 数据集 | 62.1MB | 44.5MB | 54MB | 58.6MB |

不难看出，对于主成分保留个数越多的压缩图像，显示效果越好，保存原图像的信息越丰富，图像更清晰，但是图像大小的压缩不够高。反之，图像信息丢失严重，但是压缩后的图像占用空间更少。

所以，PCA算法用于图像压缩是可行的。

1. 结论

本文使用PCA算法对图像进行压缩处理，结果是符合预期的。但仍有很多不足之处，比如如何选取一个主成分保留个数可以保留原图像的一定比例的信息量，如何能兼顾压缩效果和压缩后的图像大小，这些问题需要对图像的信息进行衡量，能够比较压缩前后的图像的信息差值，并且可以在原图像信息保存率、主成分个数这两者之间建立一种映射，本文已经在程序中实现了一个简单的计算方法，但是还没形成一个完备的方案来解决这一问题，将在课后继续完成。