

【题目】PCA 与浮游生物分类

本题的目的是让大家在实际应用中熟悉 PCA 的基本理论和应用方法，体会在听课的基础上通过自我努力扩展知识面的学习方法，感受在知识爆炸的年代进行科研和创新的艰苦和乐趣。更重要的，我们希望同学们明确这样一个观念，随机过程的确可以给我们解决问题提供理论基础，但是要想把这种作用体现出来，大量细致的工作是不可或缺的。

主成分分析（PCA）在电子信息领域有比较广泛的应用，模式识别是其中比较典型的一种。通过对待识别的原始数据进行 PCA 处理，可以降低数据的维数，减少次要因素对识别的影响，突出数据的主要特征，从而更容易构造高性能的识别算法。

PCA 在模式识别中最成功的应用莫过于人脸识别（Face Recognition）。作为最典型的人脸识别手段，特征脸（Eigenface）算法本质上就是 PCA[1][2]。从特征脸出发，以 PCA 为基本原理的模式识别方法得到了进一步的发展。[3]从计算和性能层面讨论了 PCA 类方法；[4]不但讨论了 PCA，还谈到了 PCA 的拓展——ICA 在识别中的应用；[5]涉及到了重要且有趣的问题，如何从 PCA 的结果中选取合适的特征来设计识别系统，明确指出识别与数据压缩这两种不同的目的将会导致不同的特征选择依据。[6]是一份简洁清晰的课件，便于学习 PCA 与模式识别的紧密关联。除此之外，下列网址包含丰富的文献甚至源码，对于大家了解和学习 PCA 与模式识别有帮助。

<http://www.face-rec.org/algorithms/PCA>

我们本次 Project 的问题是利用 PCA 及其相关知识对海洋中的浮游生物进行识别[7]。人脸识别已经研究了超过 20 年，取得突破和创新的机会在变小。与之相对，海洋浮游生物的识别刚刚起步。所以，我们选择海洋浮游生物作为研究对象，也符合科研选题的基本原则。识别所用的数据来自如下网站：

<https://github.com/hsosik/WHOI-Plankton>

同学需要完成的任务如下：

1. 在课上对 PCA 学习的基础上，掌握其实际计算方法，特别是使用高级语言

工具（Matlab 或 C）对实际数据进行 PCA 进行操作的能力。

2. 使用你掌握的 PCA 相关方法对数据库中的对象进行识别，最好可以对方法的优劣进行比较，弄清楚这些传统方法如果好，好在哪里？如果不好，究竟是什么原因导致效果差。
3. 你能否在学习的基础上提出新的想法来改进你的识别效果，和网站上的纪录比一比，看看能否在其中占一席之地。
4. 完成并提交 Project 报告（见“Project 报告撰写要求”）。
5. 提交 Project 工作中所用到的源代码。

【参考文献】

- [1]. Face Recognition using Eigenface
- [2]. Eigenfaces for Recognition
- [3]. Computational and Performance Aspects of PCA-based Face Recognition Algorithms
- [4]. Recognizing faces with PCA and ICA
- [5]. Analyzing PCA-based Face Recognition Algorithms Eigenvector Selection and Distance Measures
- [6]. PCA-based Object Recognition
- [7]. WHOI-Plankton- A Large Scale Fine Grained Visual Recognition Benchmark Database for Plankton Classification