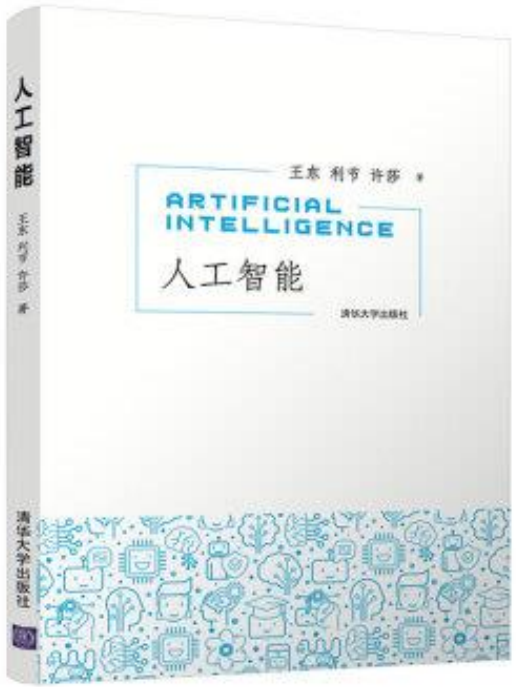
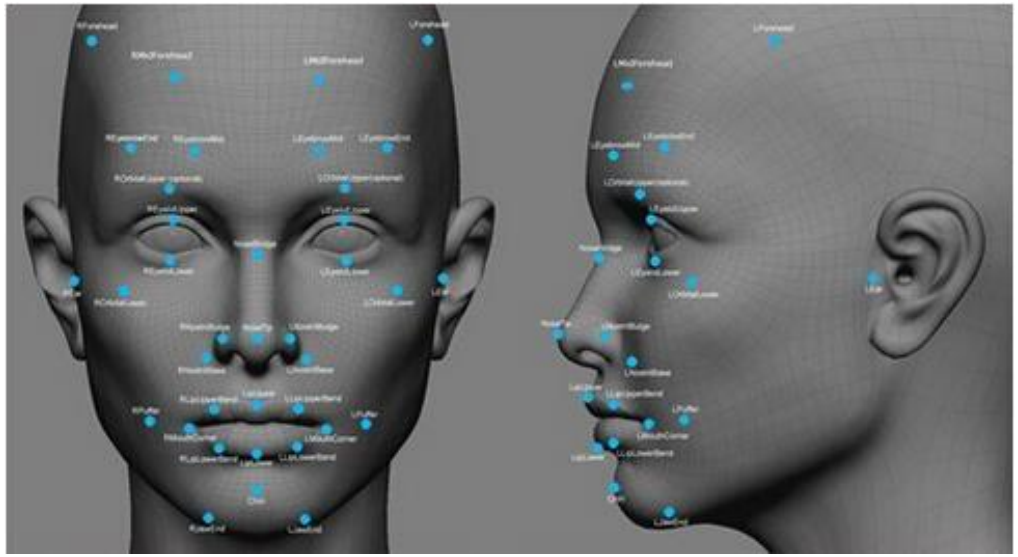
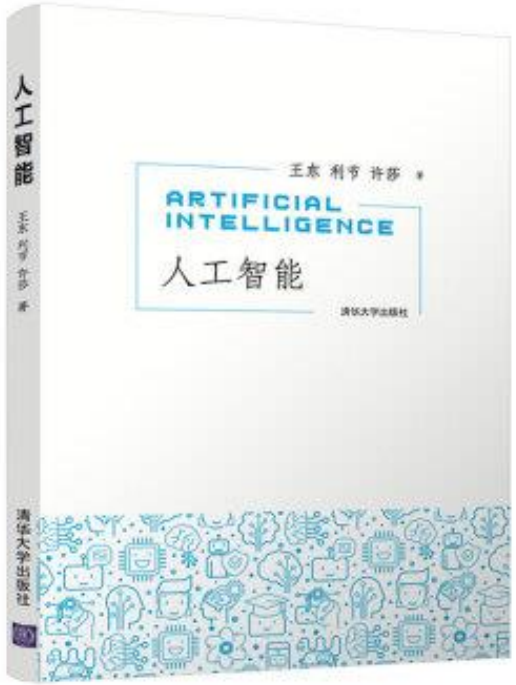

人工智能
王东 利节 许莎 著
ARTIFICIAL
INTELLIGENCE
人工智能
清华大学出版社
清华大学出版社



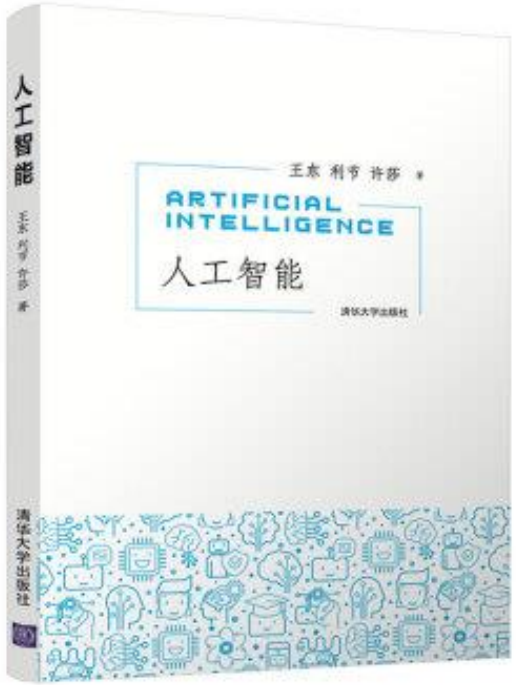
认识你的脸

利节



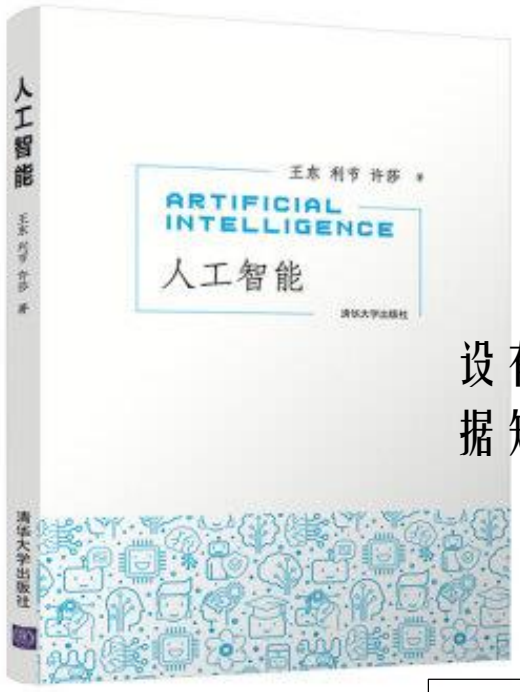
目　录

- 主成分分析（PCA）
- Eigen Face



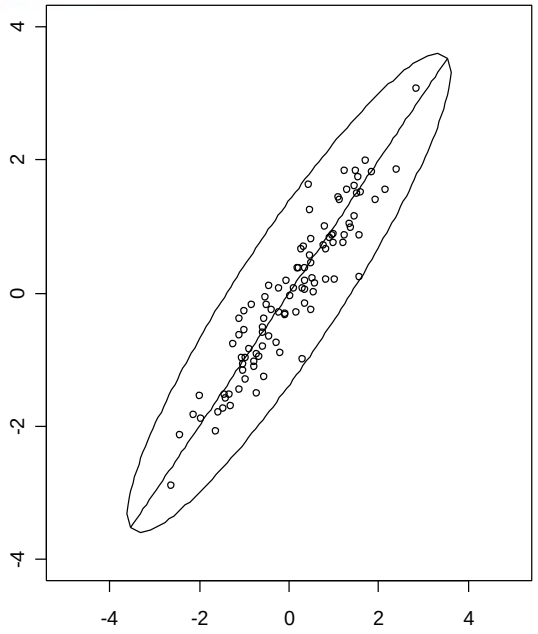
目　录

- 主成分分析（PCA）
- Eigen Face



均值和协方差
特征值和特征向量

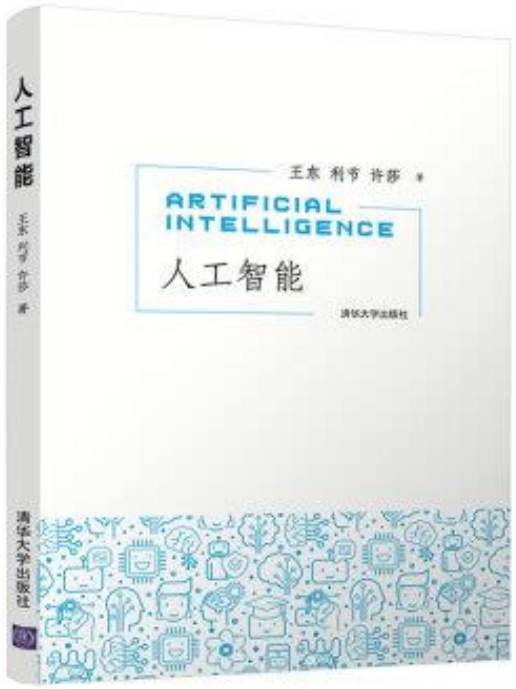
设有n个样本，每个样本观测p个指标（变量）：X_1，X_2，…，X_n，得到原始数据矩阵：



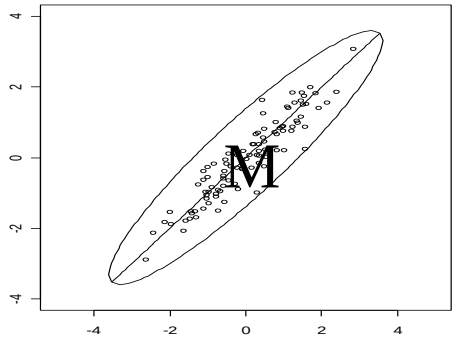
$$\mathbf{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{pn} \\ x_{21} & x_{22} & \cdots & x_{pn} \\ \vdots & \vdots & & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{bmatrix}_{p\times n}$$

$$\uparrow \quad \uparrow \quad \uparrow$$

$$\mathbf{X}_1 \quad \mathbf{X}_2 \quad \mathbf{X}_n$$



1. 样本均值



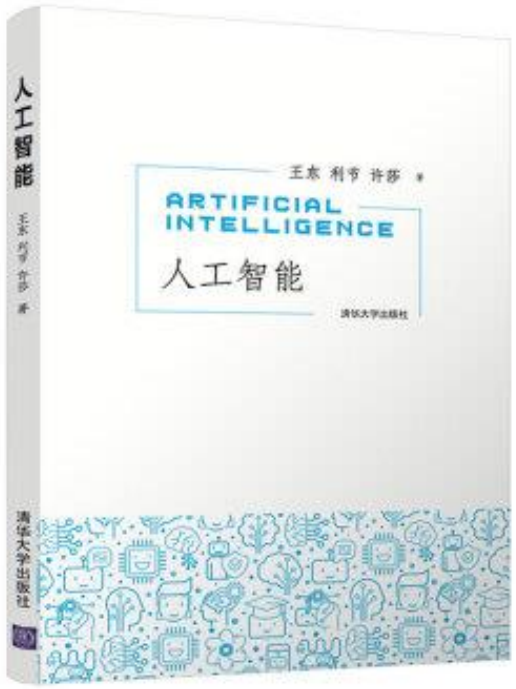
$$\mathbf{M} = \frac{1}{n}(\mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_n).$$

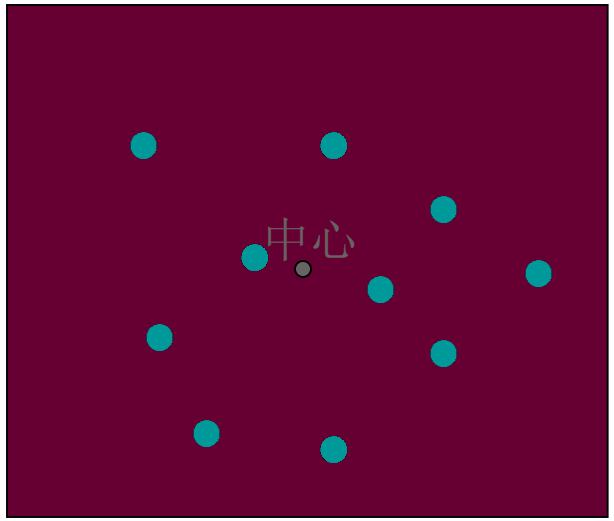
显然,样本均值是数据散列图的中心.

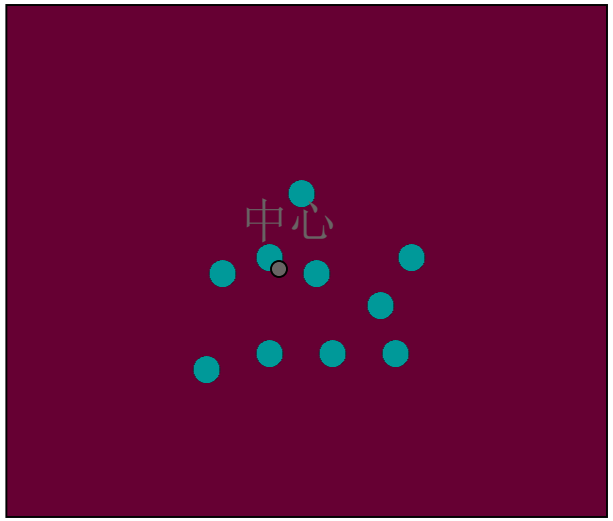
$$\mathbf{X}_k = \mathbf{X}_k - \mathbf{M}$$

于是 $p*n$ 矩阵的列B具有零样本均值,称为平均偏差形式

$$\mathbf{B} = \left[\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n\right]$$





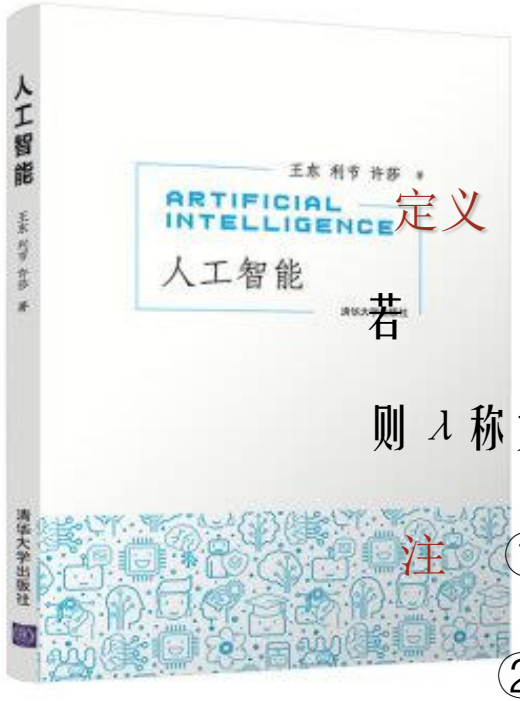


2. 样本协方差

注意：协方差
是对称矩阵且半正定

$$\mathbf{S} = \frac{1}{n-1}\mathbf{B}\mathbf{B}^T$$

协方差的大小在一定程度上反映了多变量之间的关系，但它还受变量自身度量单位的影响.



特征值与特征向量

定义 A 为 n 阶方阵，λ 为数，X 为 n 维非零向量，若

$$AX = \lambda X$$

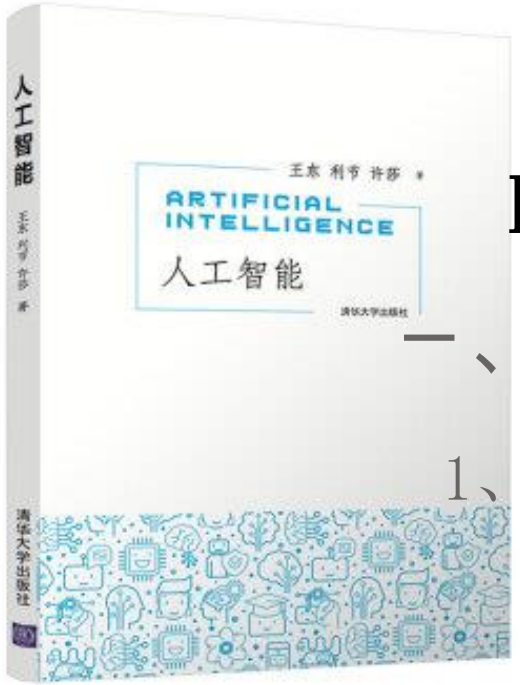
则 λ 称为 A 的**特征值**，X 称为 A 的**特征向量**.

注 ① 特征向量 $X \neq 0$，特征值问题只针对与方阵；

② λ, X 并不一定唯一；

③ n 阶方阵 A 的特征值，就是使齐次线性方程组 $(\lambda I - A)x = 0$ 有非零解的 λ 值，即满足 $|\lambda I - A| = 0$ 的 λ 都是方阵 A 的特征值.

定义 称以 λ 为未知数的一元 n 次方程 $|\lambda I - A| = 0$ 为 A 的**特征方程**.



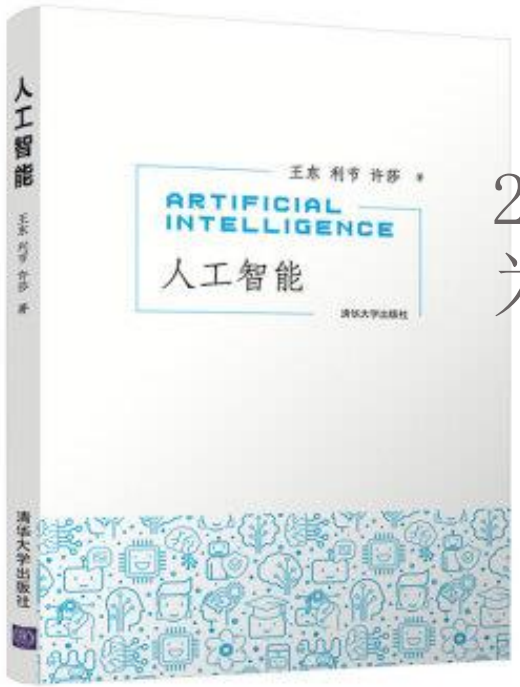
PCA的性质

一、两个线性代数的结论

1、若A是p阶实对称阵，则一定可以找到正交阵U，使

$$U^{-1}AU=\begin{bmatrix}\lambda_1 & 0 & \cdots & 0\\ 0 & \lambda_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \lambda_p\end{bmatrix}_{p\times p}$$

其中 $\lambda_i, i=1.2.\cdots p$ 是A的特征根。



2、若上述矩阵的特征根所对应的单位特征向量为 $\mathbf{u}_1,\cdots,\mathbf{u}_p$

令 $$\mathbf{U}=(\mathbf{u}_1,\cdots,\mathbf{u}_p)=\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix}$$

则实对称阵 $\mathbf{A}$ 属于不同特征根所对应的特征向量是正交的，即 $\mathbf{U}'\mathbf{U}=\mathbf{U}\mathbf{U}'=\mathbf{I}$

$$\mathbf{U}'\mathbf{U}=\mathbf{U}\mathbf{U}'=\mathbf{I}$$

§3.4 PCA的性质(续)

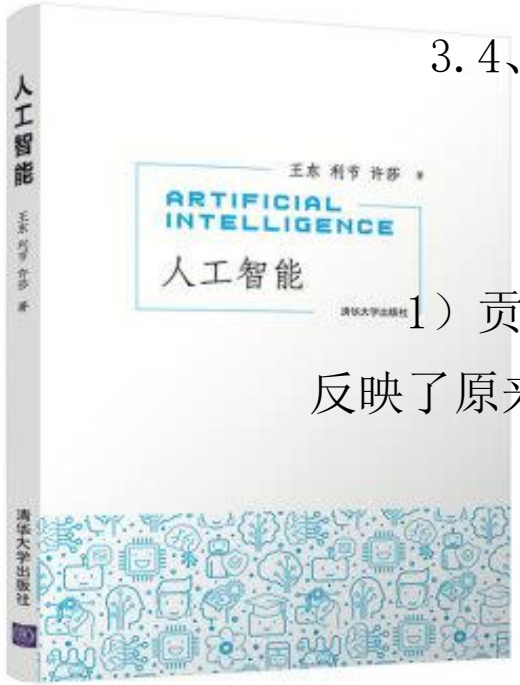
3、均值 $E(\mathbf{U}^T x) = \mathbf{U}^T M$

4、方差为所有特征根之和

$$\sum_{i=1}^{p} Var(F_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_p^2$$

说明主成分分析把P个随机变量的总方差分解成为P个不相关的随机变量的方差之和。

协方差矩阵Σ的对角线上的元素之和等于特征根之和。



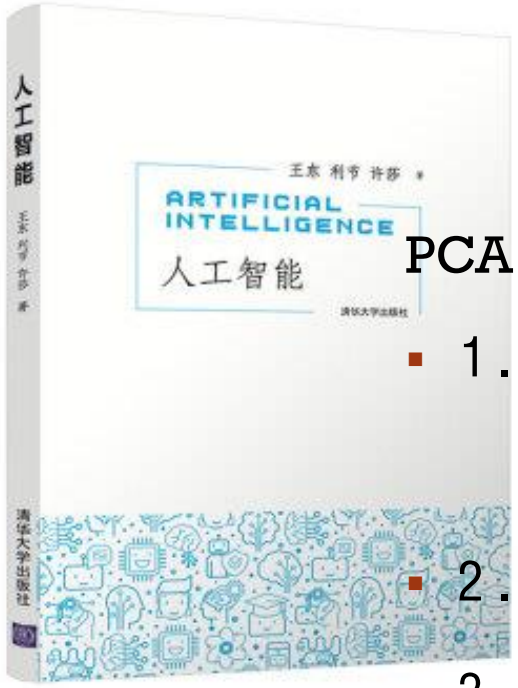
3.4、精度分析

1）贡献率：第i个主成分的方差在全部方差中所占比 $\lambda_i \Big/ \sum_{i=1}^{p} \lambda_i$ ，称为贡献率，反映了原来P个指标多大的信息，有多大的综合能力 。

2）累积贡献率：前k个主成分共有多大的综合能力，用这k个主成分的方差和在全部方差中所占比重

$$\sum_{i=1}^{k} \lambda_i \Big/ \sum_{i=1}^{p} \lambda_i$$

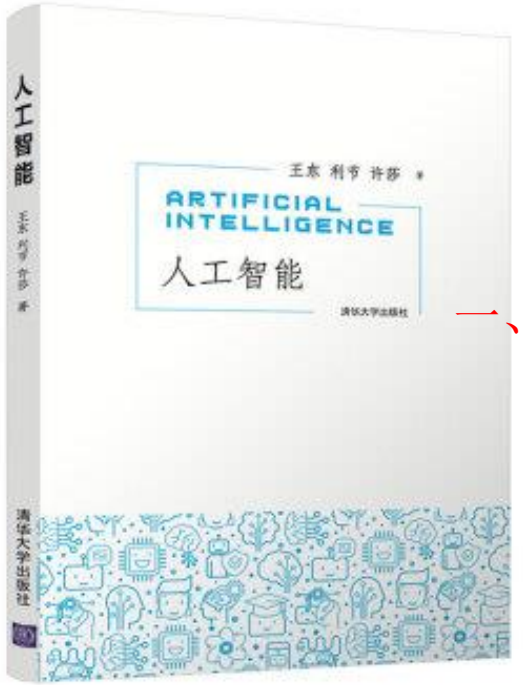
来描述，称为累积贡献率。 $\sum_{i=1}^{k} \lambda_i \Big/ \sum_{i=1}^{p} \lambda_i$



PCA 常用统计量：

- 1.特征根 λ_i
- 2.各成分贡献率
- 3.前各成分累计贡献率 $\frac{\lambda_i}{\sum \lambda_i}$
- 4.特征向量 各成分表达式中标准化原始变量的系数向量，就是各成分的特征向量。

我们进行主成分分析的目的之一是希望用尽可能少的主成分F_1，F_2，…，F_k（k≤p）代替原来的P个指标。到底应该选择多少个主成分，在实际工作中，主成分个数的多少取决于能够反映原来变量80%以上的信息量为依据，即当累积贡献率≥80%时的。最常见的情况是主成分为2到3个。主成分的个数就足够了



主成分分析的步骤

一、基于协方差矩阵

$$X_1 = (x_{1l}, x_{2l}; \cdots, x_{pl})'(l = 1\ 2; \cdots, n)$$

$$\hat{\Sigma}_x = \left(\frac{1}{n-1} \sum_{l=1}^{n} (x_{il} - \bar{x}_i)(x_{jl} - \bar{x}_j) \right)_{p \times p}$$

第一步：由X的协方差阵Σ_x，求出其特征根，即解方程 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ ，可得特征根 $|\boldsymbol{\Sigma} - \lambda \mathbf{I}| = 0$ 。

第二步：求出分别所对应的特征向量U_1，U_2，…，U_p，

$$\mathbf{U}_i = \left(u_{1i},\ u_{2i},\cdots,\ u_{pi}\right)^T$$

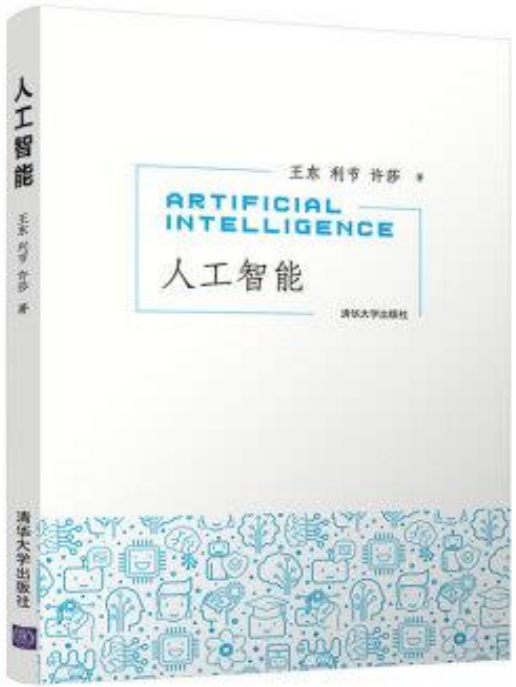
第三步：计算累积贡献率，给出恰当的主成分个数。

$$F_i = \mathbf{U}^T{}_i\mathbf{X},\ \ i = 1,2,\cdots,\ \ k(k \le p)$$

第四步：计算所选出的k个主成分的得分。将原始数据的中心化值:

代入前k个主成分的表达式，分别计算出各单位k个主成分的得分，并按得分值的大小排队。

$$\mathbf{X}_i^* = \mathbf{X}_i - \overline{\mathbf{X}} = \left(x_{1i} - \overline{x}_1,\ \ x_{2i} - \overline{x}_2,\cdots,\ \ x_{pi} - \overline{x}_p\right)^T$$

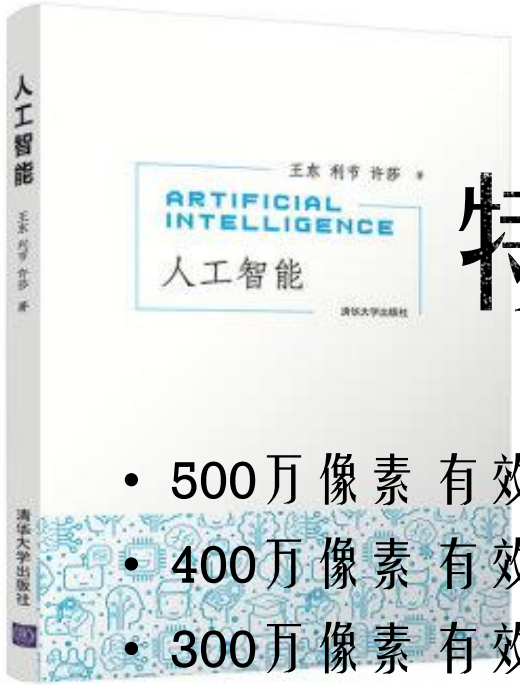


目　录

- 主成分分析（PCA）
- Eigen Face

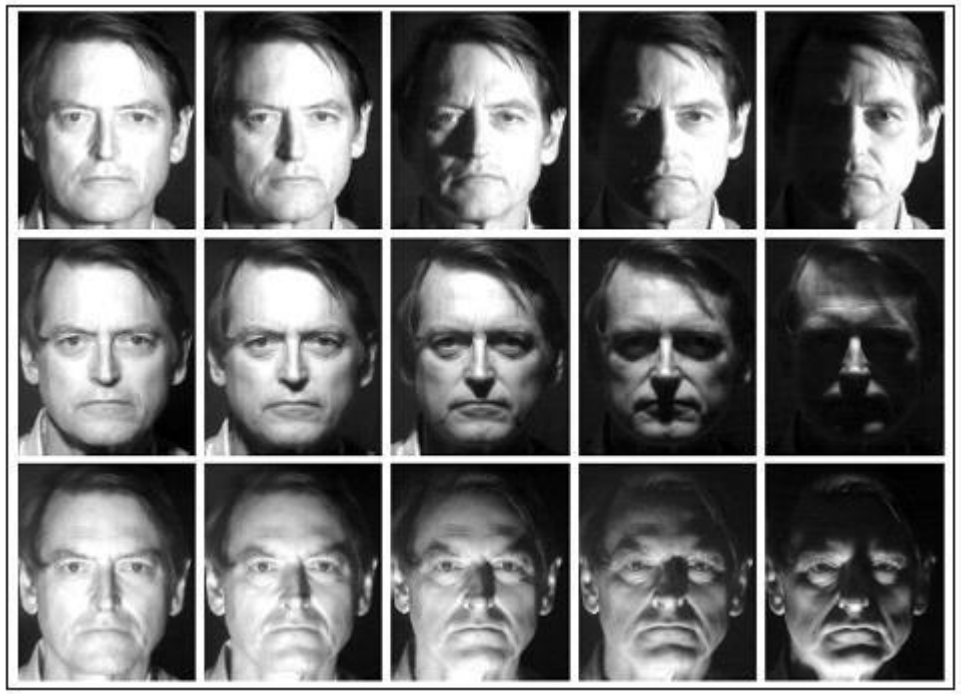
特征提取

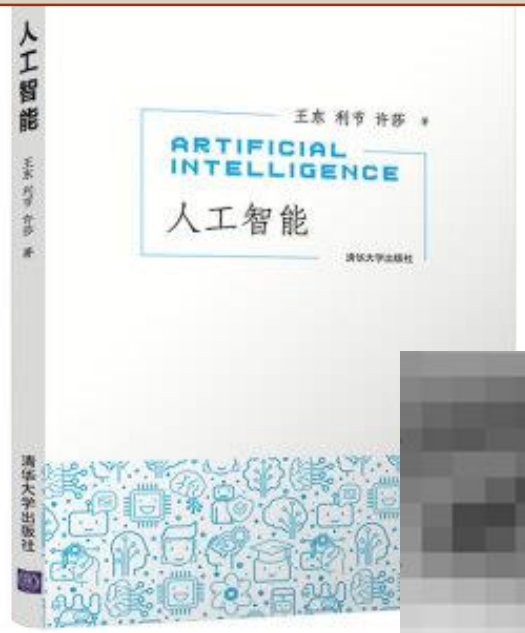
像素（px）是计算机处理图像的基本单位，用像素表示图像是所有图像处理的第一步，让我们感受一下像素表示下的人脸图像。



特征提取

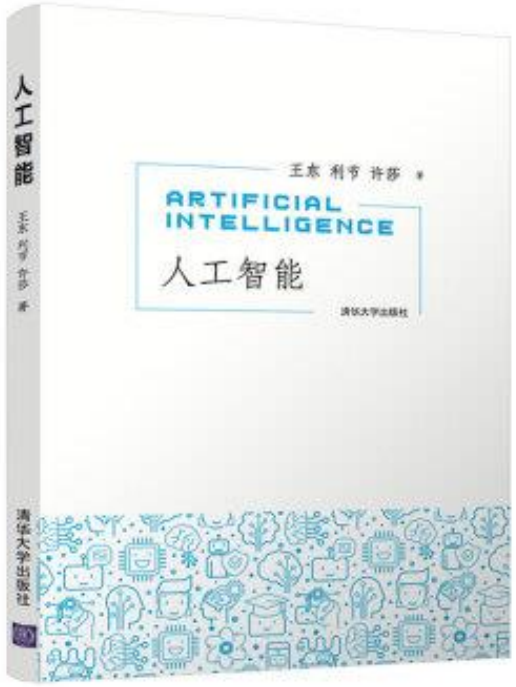
- 500万像素 有效4915200，像素2560*1920
- 400万像素 有效3871488，像素2272*1704
- 300万像素 有效3145728，像素2048*1536
- 200万像素 有效1920000，像素1600*1200
- 130万像素 有效1228800，像素1280*960
- 80万像素 有效786432，像素1024*768
- 50万像素 有效480000，像素800*600
- 30万像素 有效307200，像素640*480



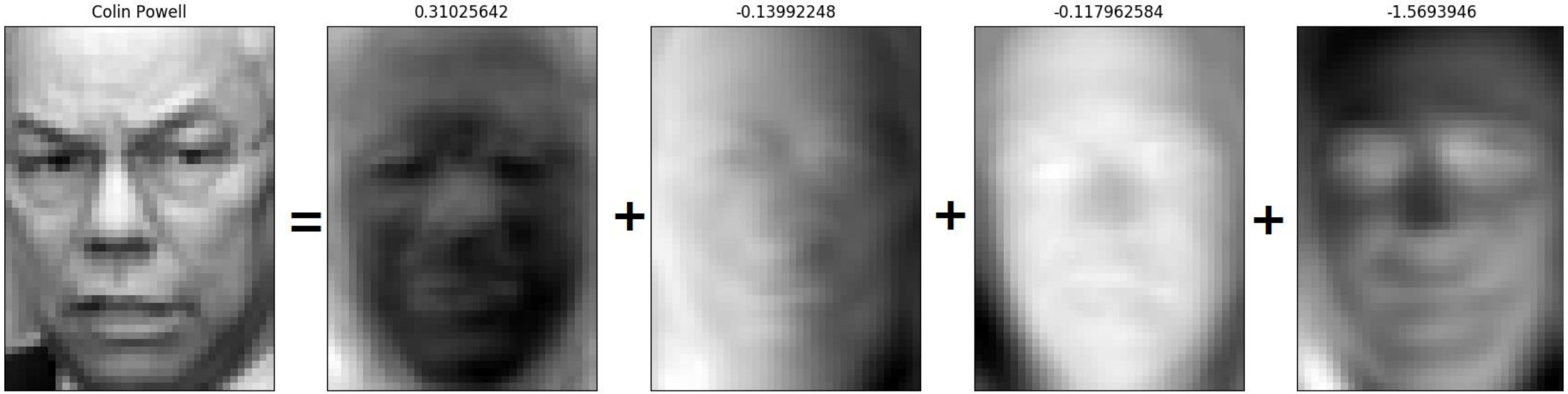


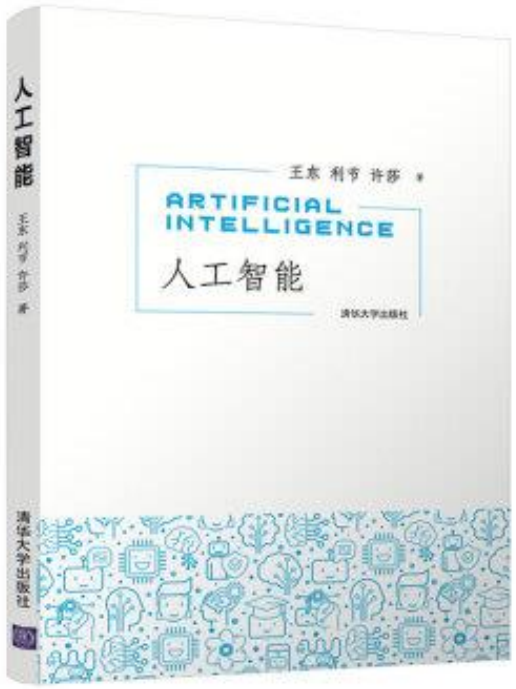
特征提取

计算机需要从这些象素中认出人脸，必须脱离象素，而寻找更全局的特征。



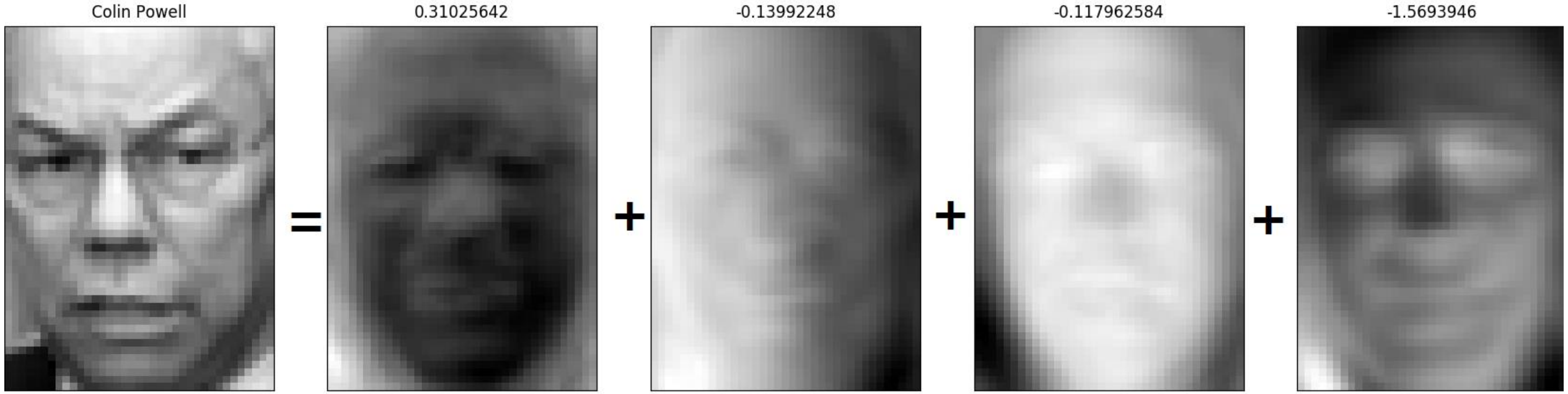
这些典型人脸照片并不是真正的人脸照片，而是代表了某一方面人脸特性的“特征照片”，只有将这些特征照片按一定权重加和起来（称为加权和）才能组成一张真正的人脸照片。这些“特征照片”称为特征脸。

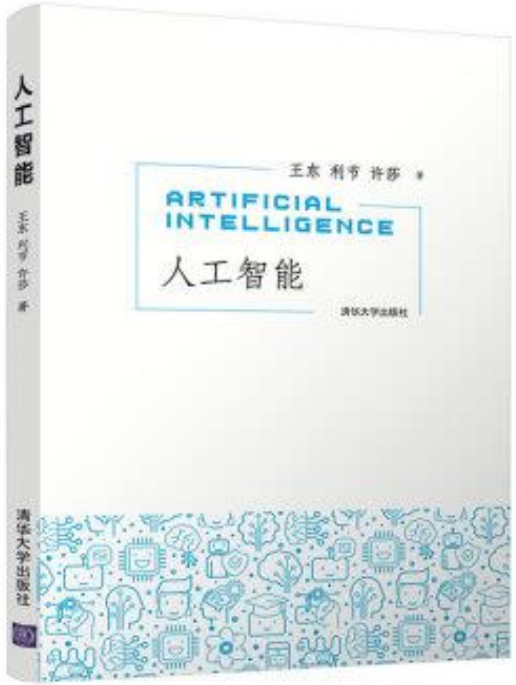




如何得到这些特征脸呢?

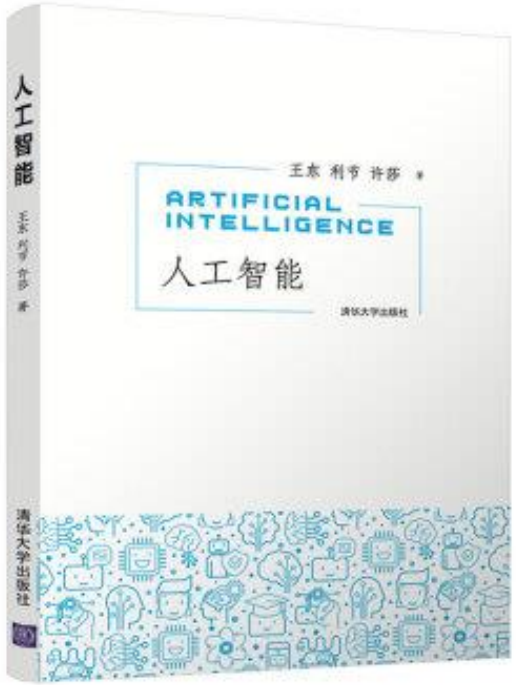
首先找到一张特征脸，记为**V1**，让它尽可能近似所有人脸；然后在所有人脸中减去基于**V1** 的近似，即得到**V1** 近似后的余量，或称为“残差”。注意该残差也是一张图片。再找到一张特征脸，让它尽可能近似所有残差图片，得到**V2**。依此类推，每次用一张新的特征脸来近似前面所有特征脸近似后的残差，从而使近似程度越来越高。



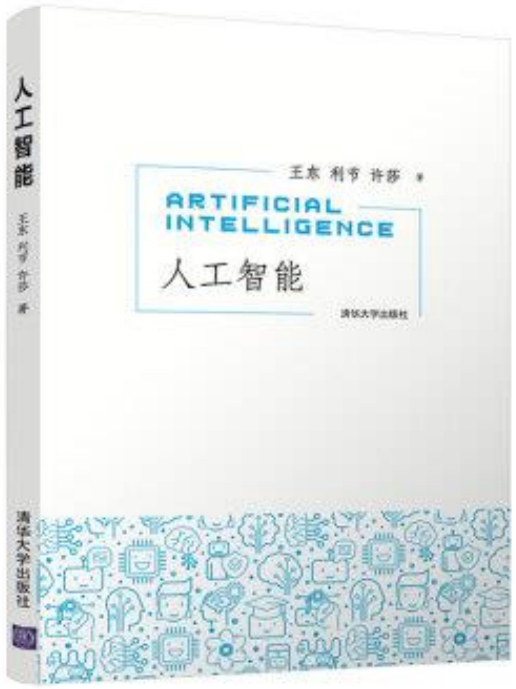


上述分步近似法在机器学习中称为
主成分分析（**Principle Component Analysis, PCA**）

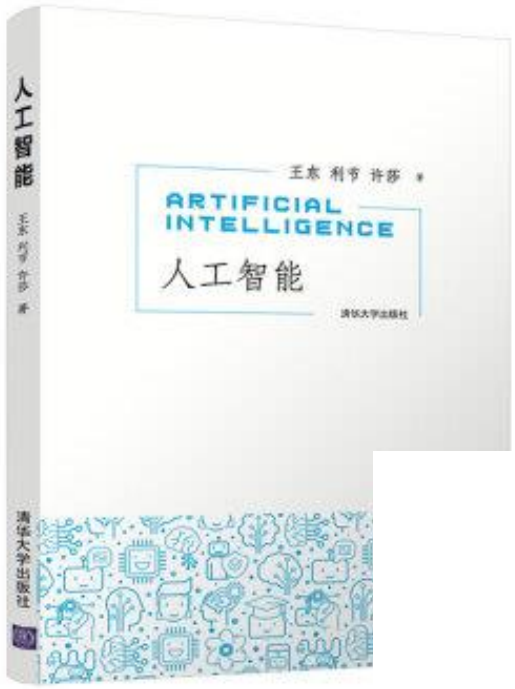
这些特征脸被称为**eigen-face**



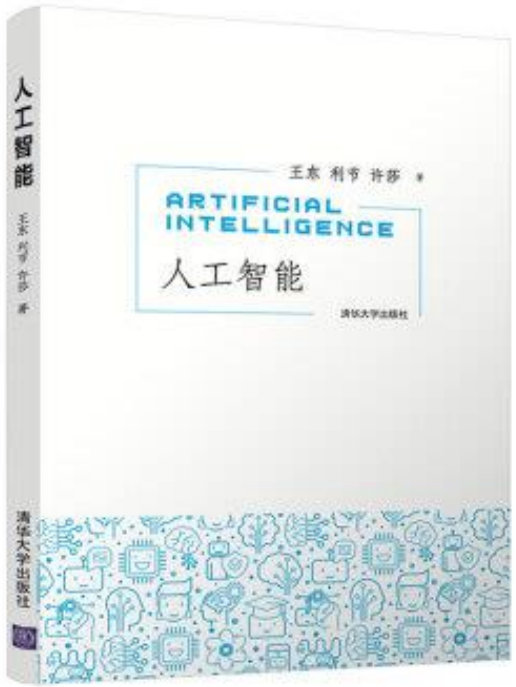
EigenFace(特征脸)在人脸识别历史上应该是具有里程碑式意义的，其被认为是第一种有效的人脸识别算法。**1987**年 **Sirovich and Kirby** 为了减少人脸图像的表示（降维）采用了**PCA**（主成分分析）的方法，**1991**年 **Matthew Turk**和**Alex Pentland**首次将**PCA**应用于人脸识别，即将原始图像投影到特征空间，得到一系列降维图像，取其主元表示人脸，因其主元有人脸的形状，估称为“特征脸”。



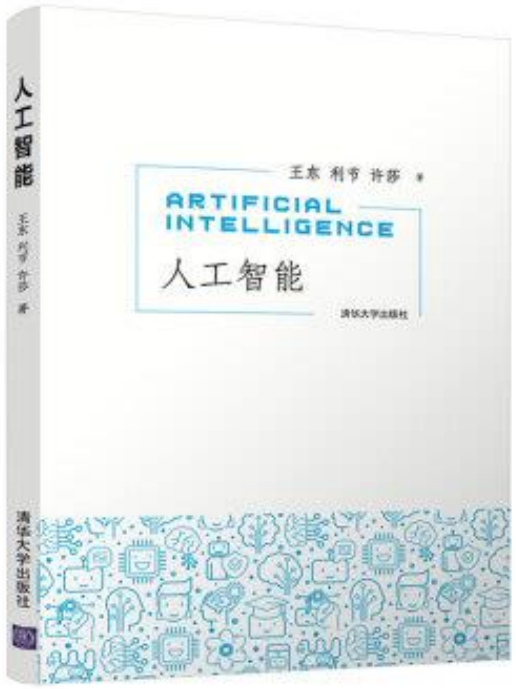
EigenFace是一种基于统计特征的方法，将人脸图像视为随机向量，并用统计方法辨别不同人脸特征模式。**EigenFace**的基本思想是，从统计的观点，寻找人脸图像分布的基本元素，即人脸图像样本集协方差矩阵的特征向量，以此近似的表征人脸图像，这些特征向量称为特脸。



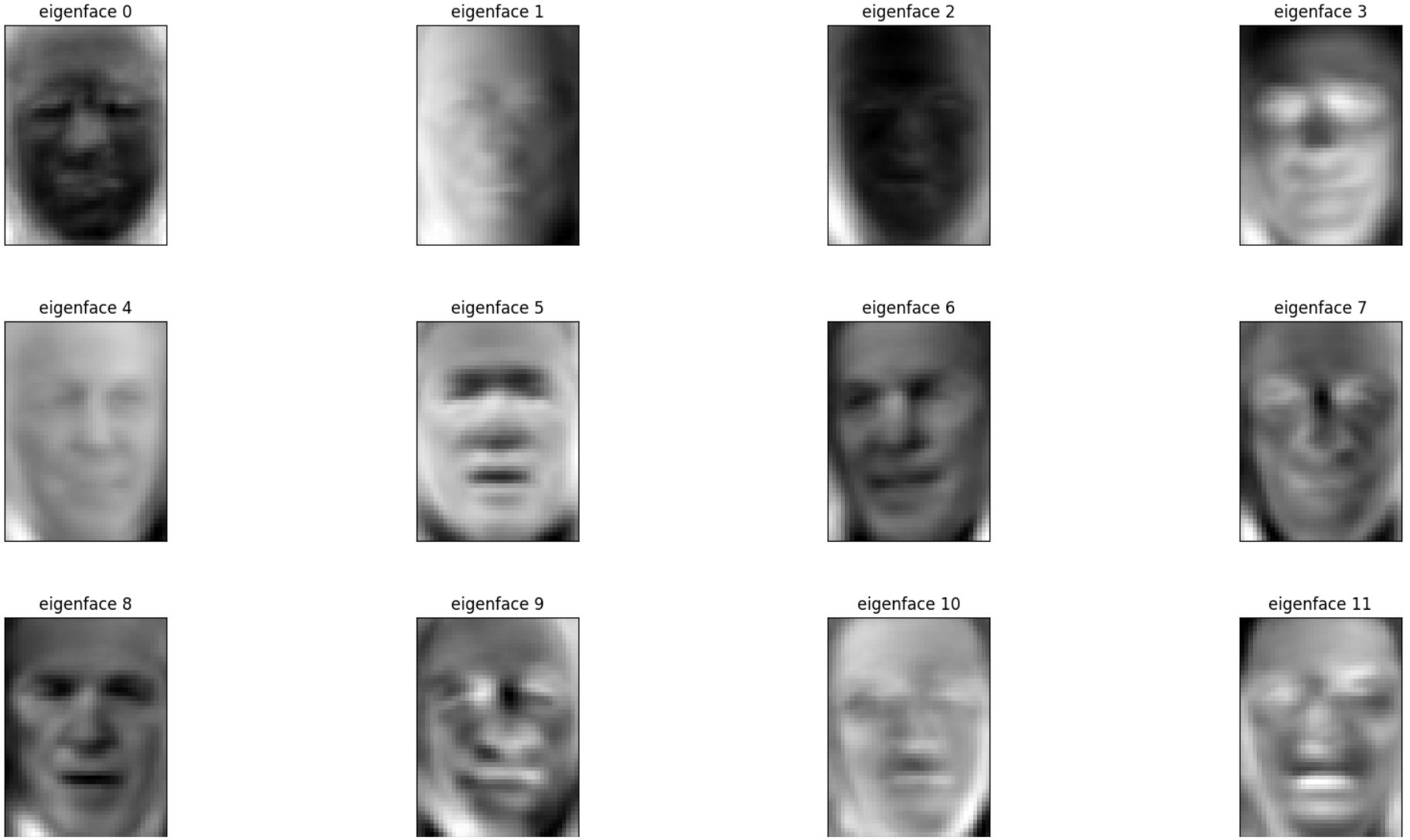
eigen-face的流程图

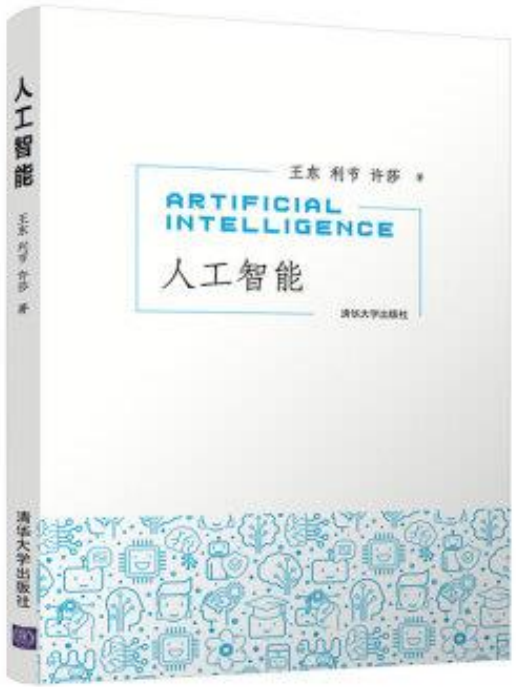


每次得到的特征脸在PCA 中称为一个主成份（Principle Component, PC）。对一张照片进行近似时，每个特征脸上的权重事实上是该照片在相应主成份上的投影大小。特征脸方法是PCA 在人脸数据上的应用，基于该方法得到的特征（即各特征脸上的权重）也称为PCA 特征（PCA Feature）。

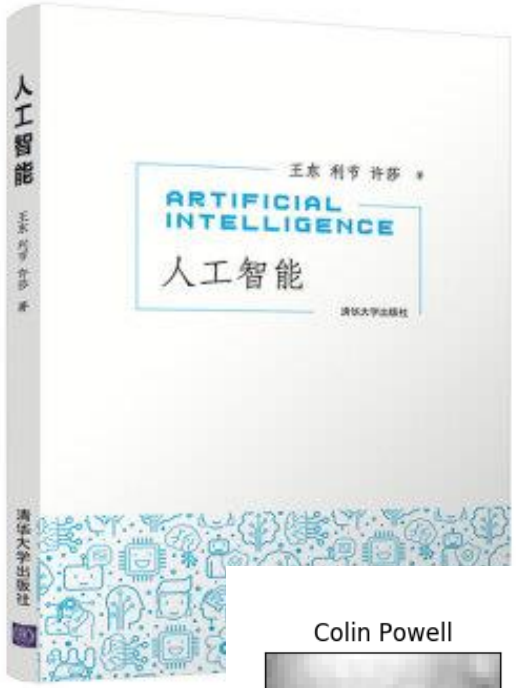


一个基于**LFW** 数据集训练的**PCA** 模型中前十二个主成份对应的特征脸。可以看到，不同特征脸代表了人脸图片的不同典型特征。

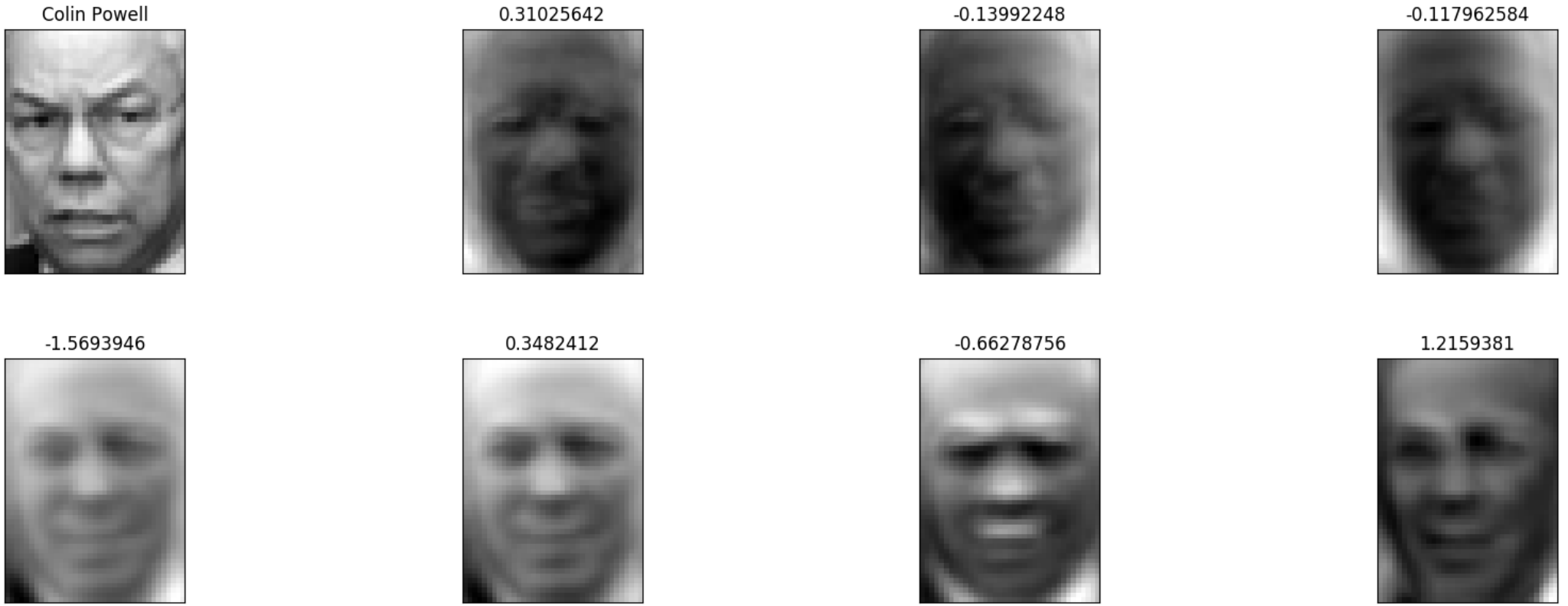


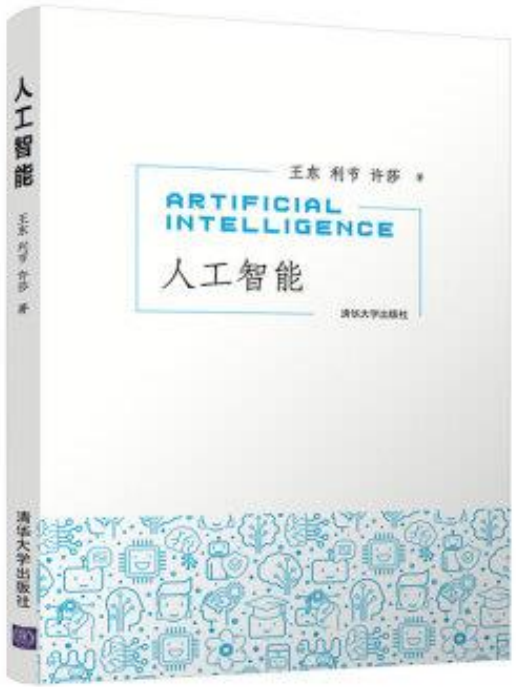


值得再次强调的是，基于**PCA** 得到的特征脸不是同一层次的，每张新的特征脸描述的都是前面所有特征脸组合后仍然无法描述的残差。



人脸近似过程可以理解为一个素描过程。在画一幅素描时，画家通常先画一个人脸的基本轮廓和基本部件，这相当于按权重画出第一张特征脸***V1***，之后对眼鼻等主要部件进行细致描绘，相当于按权重画出第二张特征脸***V2***，接着对头发、眉毛、嘴角等更细微的部分细描，相当于按权重画出第三张特征脸***V3***，... 如此迭代求精，最后即可得到一幅完整的人脸素描像。从第一张特征脸开始，逐次加入其它特征脸，渐渐生成脸部细节。




人工智能
王东 利节 许莎 著
ARTIFICIAL
INTELLIGENCE
人工智能
清华大学出版社
清华大学出版社

The end !