TUXIANG CHULI YUANLI JISHU YU SUANFA

图像处理原理、技术与算法

陆系群　陈　纯　编著

浙江大學出版社

图像处理原理、技术与算法

陆系群　陈　纯　编著

浙江大學出版社

图书在版编目（CIP）数据

图像处理原理、技术与算法 / 陆系群，陈纯编著.
杭州：浙江大学出版社，2001.8
ISBN 7-308-02777-5

Ⅰ.图… Ⅱ.①陆…②陈… Ⅲ.图像处理－计算机应用 Ⅳ.TP391.41

中国版本图书馆 CIP 数据核字（2001）第 042018 号

责任编辑：陈晓嘉
出版发行：浙江大学出版社
（杭州浙大路 38 号 邮政编码 310027）
（E-mail:zupress@mail.hz.zj.cn）
（网址:http://www.zjupress.com）
排　　版：浙江大学出版社电脑排版中心
印　　刷：浙江大学印刷厂
开　　本：787mm×1092mm 1/16
印　　张：15
字　　数：362 千
版 印 次：2001 年 8 月第 1 版 2001 年 8 月第 1 次印刷
书　　号：ISBN 7-308-02777-5/TP・215
印　　数：0001—2000
定　　价：32.50 元

前　言

图像处理技术是利用计算机来处理、分析和理解视觉信息的一项技术。

在20世纪初，运用机器来处理图片是一件非常困难的事。但随着计算机硬件、图像获取设备、显示设备的不断改进和各种高性能工作站的出现，图像处理这门新兴学科迅猛地向前发展；而信息社会的到来，又使图像处理技术进入了一个更加蓬勃发展的阶段。特别是多媒体技术、通信技术、信息存储技术和以Internet为代表的计算机网络技术的加速发展和广泛普及以及高清晰度电视(HDTV)的深入应用研究，图像处理技术研究和应用的前景将更为广阔。

图像处理技术从一开始就是一个基于线性代数、统计理论和物理学之上，具有很强理论背景的研究领域，它需要广泛的基础知识，包括计算机科学、数字信号处理、随机过程和统计数学、矩阵分析、信息论、控制论和最优化理论等。同时，图像处理又是一门与应用紧密结合的学科，应用领域极为广阔，如在计算机视觉、地理、气象、航空航天、医疗保健、刑事侦查等领域中的应用。

本书是基于经典图像处理参考文献、最新有关图像处理技术的文献及作者多年的实践经验完成的。书中对每一种原理方法作了详细的描述，并为书中每一种图像处理技术的算法提供C语言程序，给出了许多实验结果以比较不同算法之间的优缺点。本书的特点之一是不仅仅局限于理论描述，而是通过生动形象的实验结果，即使初学者对图像处理技术这门学科有一个大概的直观的了解，也为具有一定图像处理知识的科技人员提供最新的图像处理技术、图像处理过程中的经验及了解图像处理技术的最新发展方向提供帮助。本书可以作为高等院校计算机、自动化、电子工程、生物医学工程等专业高年级选修课以及研究生的教材。本书还提供了一张光盘，将书中所有程序和实验结果刻录在内，以便读者对照比较。

本书内容共分八章：

· 首先详细描述在图像处理中较常见的几种数学模型，它们各自的特点和应用范围以及图像的采样和亚采样过程。

· 考虑到对任何图像进行处理之前，都必须首先了解图像文件的输入与输出，在本书的第2章，我们介绍了几种常见的图像文件格式，如GIF、TIFF、BMP等，并给出这几种图像文件格式输入输出的C语言程序。

· 在第3章，我们将图像变换方法分为两类进行介绍。第一类是基于图像信号是平

稳信号的假设——以傅里叶变换为代表的传统的图像变换方法；第二类是将以上严格假设条件放宽的图像变换方法——以小波变换为代表的图像变换方法，它克服了传统傅里叶变换不具有时频局部性质的缺陷，是一种窗口大小不变但形状可变的短时傅里叶变换。在描述各种变换方法时，我们给出各种变换的起因、所需假设条件和较详细的推导过程，并结合实际应用讲述它们各自的优缺点。

· 紧接着的第4章是和图像变换密切相关的图像编码。我们首先从香农率失真理论出发，分别讲述无损编码(以熵编码为代表)和有损编码。对于有损编码来说，图像信号的压缩主要体现在图像信号的量化阶段。图像量化过程分为两类：标量量化，主要讨论最优非线性均方量化器——Lloyd-max 量化器；而向量量化，主要讨论著名的LBG 算法。图像编码主要利用了图像中普遍存在的冗余度，如预测编码利用了图像的统计冗余性，图像变换编码利用了图像的空间冗余度，而分形编码利用了图像的尺度冗余度。在图像变换编码中，我们还特意描述了目前静止图像压缩的标准——JPEG。我们讨论了各种算法的原理并给出相应的实验结果和程序，以使读者能够直观地看到并体验各种算法的效果，更希望读者能在上述的学习基础上结合实际需要对算法作进一步的改进。以上这些编码都是基于香农的可分离原则，即分别设计源编码和信道编码：源编码按率失真理论的下界来设计，而图像信号在传输过程中假设信道能正确无误地将信号传输到接收端。但在实际网络中，信号的传输过程并非如此，所以现在联合设计源编码和信道编码(Joint Source-Channel Coding)是研究的一大趋势。

· 第5章介绍的图像增强技术主要有两个作用：一是改善图像的视觉效果，二是使图像变得更有利于计算机处理。从作用域出发，图像增强可以分为空间域法和变换域法两大类。空间域处理技术是直接面对图像灰度作运算；而变换域法处理技术是在图像的某种变换域内，对图像的变换系数进行运算，并作某种修正，然后通过逆变换获得图像增强。在这一章中，我们详细描述了中值滤波、各种边缘提取算法等，并给出了实验结果和程序。

· 由于图像编码的最终结果是要将编码后的图像信号放在信道上传输，而信号在传输过程中难免会受到各种外界因素的干扰，因此图像恢复在目前这种源编码和信道编码分离的情况下，是一种必不可少的技术。在第6章中，我们较详细地描述了两种传统的图像恢复方法：维纳滤波和卡尔曼滤波。维纳滤波是基于图像信号是平稳信号的假设进行图像处理的，而卡尔曼滤波则可以将所处理的图像信号推广为时变信号。

· 图像分割是按照具体应用的要求和具体图像的内容将图像分割成一块块区域，目的是将我们所感兴趣的对象提取出来。图像分割技术与应用密切相关，当应用要求比较简单时，分割技术相对来说也比较简单；但如果应用要求比较复杂，那么分割技术相对来说也比较复杂。我们将图像分割技术分为三类：基于像素灰度值的分割技术、基于区域的分割技术和基于边界的分割技术。在第7章中，我们将详细讨论各种图像分割技术。

· 最后，我们集中讨论彩色图像处理技术。我们首先从彩色视觉感知出发，讨论色彩

的表示方法和各种色彩空间及它们不同的适用范围。由于现在一般的彩色监视器最多只能显示256种颜色，所以首先要进行彩色量化。在这一章末尾，我们介绍了基于块截断编码(Block Truncation Coding)和主元分析(Principal Component Analysis)的彩色图像压缩BTC-PCA算法，并给出了相应的实验结果。

以上大致介绍了本书的概况。对书中引用的一些文献和书籍，在此谨对有关作者表示衷心的感谢。由于我们水平所限，书中难免有错误及不妥之处，恳请读者批评指正。

作者E-mail地址为xqlu@cs.zju.edu.cn。

作　者

2001年5月

于浙大求是园

目　　录

第1章 绪 论

1.1 图像处理技术的发展历史及现状

人类从一出生，人眼就在不断地接受、分析和理解周围的景物，这是人类的一种本能活动。但是在计算机出现之前，不论是在科学研究领域还是在工业应用领域，用机器来处理、分析、理解视觉和其他遥感图片都是一件非常麻烦的事。在20世纪70～80年代，图像处理的研究方向主要集中于用图像变换和数学模型来表征图像信号，图像滤波和图像恢复，静止和视频图像的压缩。虽然研究的方法没有什么改变，但由于计算机硬件、图像获取设备和显示设备的不断改进，使得图像处理技术在世界各地蓬勃发展。20世纪80年代中期，各种高性能的工作站和个人电脑应用的普及使图像处理研究和应用不再仅仅是大机构或大型学术团体的"专利"。现在随着Internet的广泛普及，图像处理技术和应用前景将更为广阔。

从应用的角度来看，数字照相技术、电子影像、数字化电视机、图像数据库和多媒体技术的出现都在推动这一领域不断地向前发展。总的来说，图像处理技术将不再局限于电子工程研究领域，它已涉及到其他学科，如计算机科学、地理、医疗保健、刑事侦察等领域。另外，除了处理位于可视频谱范围的图像信号外，在过去的20年里，对射电望远镜形成的图像、红外图像、合成孔径雷达(Synthetic Aperture Radar，SAR)图像的研究都非常活跃，特别是CT和核磁共振的利用都极大地丰富了这一领域研究的内容。除了上述所讲的这些研究领域之外，图像处理技术研究人员还积极地着力于纹理和图形形状的分析与识别、运动检测与估计、图像处理并行系统、图像处理技术的软硬件研究等工作。

由于图像处理技术从一开始就是一个基于线性代数、统计理论和物理学之上，具有很强理论背景的研究领域，因此一些具有高鲁棒性的图像处理算法已经或最终将应用到消费类型的产品中去，一些较成熟的算法也已逐步形成公认的标准。如在20世纪80年代末逐步规划形成、20世纪90年代全面公布的H.263，JEPG，MPEG-1，MPEG-2等图像压缩与传输标准，使图像处理技术在产业化方面取得巨大的成功，但这些标准的建立并不意味着我们的研究工作已经走到了尽头。相反，这些标准只建立在我们对图像及视频信号基本结构非常有限的理解的基础之上，所以还需要我们不断努力，以完善现在已有的图像处理技术，勇于向已有的理论框架提出质疑，进而提出创新的理论思想，开发更好的图像处理技术。如在最近提出的JPEG2000标准中将用近年来图像变换研究的新成果——小波变

换来取代原来的DCT，这是因为小波变换克服了傅里叶变换不具有时频局部性质的缺陷，并且和DCT一样具有快速算法。

图像处理技术发展非常快，随着基础理论研究的不断推前、更新，各种新颖的图像处理技术层出不穷。就图像压缩技术来说，在1980年前后，主要依赖于传统的信息论，如Huffman编码、LZW压缩算法(GIF图像格式的压缩算法)等，这类无损压缩方法的压缩比为1～2倍，加上有损压缩技术，如DCT，压缩比可以达到30倍左右，但这些图像压缩技术是基于像素的一类压缩技术，并没有充分利用图像本身的信息。新一代图像压缩技术注重图像信号的内部特征，充分利用这些特征和人类视觉系统的特点来进行高重构图像质量、高压缩比的压缩。如图像分形压缩技术充分利用了图像内部的自相似性，以一个迭代函数来产生分形图像，从而达到图像压缩的目的。

新一代图像压缩技术中，一种重要的方法是图像子带编码。图像子带编码的基本思想是：先将图像信号的频带用一组分析滤波器分割成一个个子带信号，再针对每个不同的子带信号，按照统计特性分配不同的编码器和比特率。这样做的好处在于：压缩误差仅仅局限于各个子带信号中，互不影响，并且可以根据人眼视觉系统对各个频率区域敏感的特征，给各个不同的子带信号分配不同的比特率。这样，不但可以获得高压缩比，而且重构图像的主观视觉效果比较好。其实，小波变换、金字塔多尺度分割算法都可以归属于图像子带编码技术。

对于数字相机(Digital Camera)来说，它的未来发展趋势是智能相机(Intelligent Camera)，它的一个明显的特征是能在一个场景中自动识别我们所感兴趣的对象，并优化相机的各种参数，以最佳角度和采光效果摄取对象，并对对象进行尺度压缩和存储。

上面只简单列举了图像处理技术的几种新方法、新技术。随着信息社会的到来，图像处理技术将进入一个更加迅猛发展的阶段。特别是以多媒体技术、通信技术、信息存储技术和Internet为代表的计算机网络技术的加速发展和广泛普及以及高清晰度电视(HDTV)的深入应用研究，更加推广了图像处理技术的研究与发展。

1.2 图像的数学模型

不论是理论研究还是实际应用，都面临着这样一个问题：怎样用信号数学模型来刻画所要处理信号的特征。其原因有三点：首先，信号模型能在理论上为研究和描述信号处理过程提供一个坚实的基础，如对一幅受外界噪声干扰的图像进行恢复时，可以根据信号特征选择一个合适的图像信号模型来设计一个图像去噪系统来恢复原始图像。其次，如果能确切地知道一个信号的数学模型，就可以通过计算机来模拟构造一个信号，而不一定非要从实际生活中找到这样一个信号，这对信号传输过程极其有用；如果知道信号发送端所要发送信号的数学模型，就可以只传输几个模型参数到接收端，然后在接收端利用得到的模型参数来重构信号，从而大大压缩信号传输的量。最后也是最重要的理由，信号的数学模型对实际应用非常有帮助，如图像编码、恢复中常用的预测模型，图像识别、分类中常用的

图像统计模型等。

图像的数学模型可以解析地表征图像中信息分布的情况，现在大量文献中讨论的图像数学模型虽然各种各样，但每一种模型都有一些特定的假设和前提条件，每一种模型都是仅仅捕获了图像中信息分布的部分特征。目前，还没有一种图像模型能囊括所有不同图像中信息分布的重要特征，因此不同的图像数学模型对应于不同的图像处理应用领域和图像类型。由于不同图像类型的图像信息特征变化很大，因此要精确地对其数学模型做统一的分类是很困难的，如按图像信号在时间轴上的连续性，可以分为离散数学模型和连续数学模型；按信号的统计特性，可以分为平稳图像数学模型（即图像的统计特征不随时间的变化而变化）和非平稳图像数学模型。

在这里，我们将图像数学模型分为确定性模型和统计性模型两类。确定性模型假设图像信号可以用解析数学的形式来描述，如在图像处理中经常用矩阵形式来表示。我们在图像恢复中常用的一种降质模型就是一种确定性模型，其数学表达式为 $\boldsymbol{g}=\boldsymbol{H}\boldsymbol{f}+\boldsymbol{n}$，其中 $\boldsymbol{g}$ 代表降质图像，$\boldsymbol{f}$ 表示原始图像，$\boldsymbol{H}$ 表示线性降质矩阵算子，$\boldsymbol{n}$ 表示外界叠加噪声。统计模型往往只能捕捉到信号的一些统计特征，这类模型包括常用的高斯模型、自回归模型、隐马尔可夫模型等。下面我们就图像处理技术应用领域中经常出现的几个数学模型做一个概述。

1.2.1 正交模型

由于图像信号的行与行、列与列、图像分块与分块、图像的帧与帧之间有很强的相关性，故可以把图像信号看成是一个数学上的复杂函数。数学上为了分析一个函数，常用的一种经典的方法是将这个函数展开成一组完备正交函数的级数和。在进行图像分析和图像处理时，我们也常常把一幅 $N\times N$ 图像 $\{u_{ij}\}$ 展开成正交基函数之和，我们把这种正交展开过程称为酉变换（Unitary Transform），即

$$u_{ij}=\sum_{k=1}^{N}\sum_{l=1}^{N}v_{kl}a(i,j;k,l)\qquad 1\leqslant i,j\leqslant N \tag{1.2.1.1}$$

$$v_{kl}=\sum_{i=1}^{N}\sum_{j=1}^{N}u_{ij}a^{*}(i,j;k,l)\qquad 1\leqslant k,l\leqslant N \tag{1.2.1.2}$$

其中元素 v_{kl} 称为变换系数，而 $\{v_{kl}\}$ 则为变换后的图像信号；$\{a(i,j;k,l)\}$ 是一组完备的正交基函数，有时也称基图像，集合 $\{a(i,j;k,l)\}$ 具有如下正交性和完备性：

$$\text{正交性：}\sum_{k=1}^{N}\sum_{l=1}^{N}a(i,j;k,l)a^{*}(i',j';k,l)=\delta_{i,i'}\delta_{j,j'} \tag{1.2.1.3}$$

$$\text{完备性：}\sum_{i=1}^{N}\sum_{j=1}^{N}a(i,j;k,l)a^{*}(i,j;k',l')=\delta_{k,k'}\delta_{l,l'} \tag{1.2.1.4}$$

这里应注意，尽管实际图像具有连续形式，但在计算机上处理的总是它的离散形式。所以，这里仅考虑数字图像信号，至于图像信号如何由连续形式经过采样和量化得到数字图像信号，将在下一节图像的采样和亚采样技术和第3章图像信号量化中详细讨论。

从式（1.2.1.1）和式（1.2.1.2）中可以看出，一般酉变换的计算复杂度为 $O(N^4)$。针

对一幅 256 × 256 像素的标准图像来说，这意味着需要上亿次运算。为了减少运算次数，式(1.2.1.1)和式(1.2.1.2)中的酉变换限制为可分离形式，即满足

$$a(i,j;k,l) = a_{i,k}b_{j,l} \tag{1.2.1.5}$$

其中 $A = \{a_{i,j}\}$ 和 $B = \{b_{i,j}\}$ 为酉矩阵(即 $A^{-1} = A^{*\mathrm{T}}$，为复数域的正交矩阵)。通常在图像处理中，我们选择 $A = B$，则有下式：

$$\boldsymbol{V} = \boldsymbol{AUA}^{\mathrm{T}} \tag{1.2.1.6}$$

$$\boldsymbol{U} = \boldsymbol{A}^{*\mathrm{T}}\boldsymbol{VA}^{*} \tag{1.2.1.7}$$

因为是先对图像的列进行变换，再对图像的行进行变换，计算复杂度降为 $O(N^3)$。其实，这时计算复杂度还是很高，但在选择酉变换时，一些特殊的酉变换存在快速算法，如常见的离散傅里叶变换(DFT)、离散余弦变换(DCT)、离散正弦变换(DST)、离散小波变换(DWT)等。

酉变换的一个重要性质是变换前后图像信号的总能量保持不变：

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\left|u_{ij}\right|^2 = \sum_{k=1}^{N}\sum_{l=1}^{N}\left|v_{kl}\right|^2 \tag{1.2.1.8}$$

即帕斯瓦尔(Parseval)定理。

图像正交级数展开模型主要应用于图像数据压缩、图像恢复等。有关以上所指出的特殊酉变换，将在第 3 章图像变换中详细论述。

1.2.2 统计模型

在这一小节，我们主要讨论隐马尔可夫模型(Hidden Markov Model，HMM)，因为它是在图像处理技术中经常用到的统计模型，它在图像分割、图像识别及图像恢复中有较明显的优势。HMM 比较复杂，在讨论它之前，我们先来简单回顾一下离散马尔可夫过程(即马尔可夫链)。

假设系统有 N 个离散状态：$s_1, s_2, \cdots, s_N$。系统在任何时刻必处在这 N 个离散状态中的一个，系统在 t 时刻的真实状态为 q_t。针对离散的一阶马尔可夫链：系统在某一时刻处于某一特殊状态的概率只由系统前一时刻的状态所决定，即

$$P[q_t = s_j | q_{t-1} = s_i, q_{t-2} = s_k, \cdots] = P[q_t = s_j | q_{t-1} = s_i] \tag{1.2.2.1}$$

进一步地，我们只考虑(1.2.2.1)式右边的值与时间无关的情况，由此可以得到一组状态转移概率，即

$$a_{ij} = P[q_t = s_j | q_{t-1} = s_i] \qquad 1 \leqslant i, \quad j \leqslant N \tag{1.2.2.2}$$

可以看出，状态转移概率满足如下两个条件：

$$a_{ij} \geqslant 0 \tag{1.2.2.3a}$$

$$\sum_{j=1}^{N} a_{ij} = 1 \tag{1.2.2.3b}$$

上述模型我们也称为显式马尔可夫模型(Observable Markov Model)，其每一个状态对应于一个物理事件，模型的输出就是系统在某一时刻所处的状态。在实际应用中，一般考虑的是因果模型，即系统有初始状态

$$\pi_i = P[q_1 = s_i] \qquad 1 \leqslant i, \quad j \leqslant N \tag{1.2.2.4}$$

而上述模型过于简单，在实际应用中有许多限制。

现在我们考虑一个更复杂的模型。离散马尔可夫模型中，系统的输出与系统所处的状态一一对应，而现在我们所要考虑的系统，其输出即观察值是系统状态的概率函数，这是一个双重嵌套的随机过程。

例 假设你站在一个房间里，这个房间中间用帘子隔开，另一个人在帘子一边抛硬币，这个人并不告诉你他是怎样操作的，而只告诉你他抛硬币的结果，也就是一个输出序列 $O = O_1 O_2 O_3 \cdots O_T$ = 正 正 反 反 反 正 反 反 正 …… 正。现在我们可以用一个隐马尔可夫模型来说明这一过程。

首先我们要确定这个过程模型的状态。为此，可以假设这个人手里只有1枚硬币（也可能有多个硬币），则这个模型有两个状态分别对应于硬币的正反面。其实，这就是一个简单的离散马尔可夫过程，系统状态对应于系统的输出。系统状态之间的转移如图1.1所示。

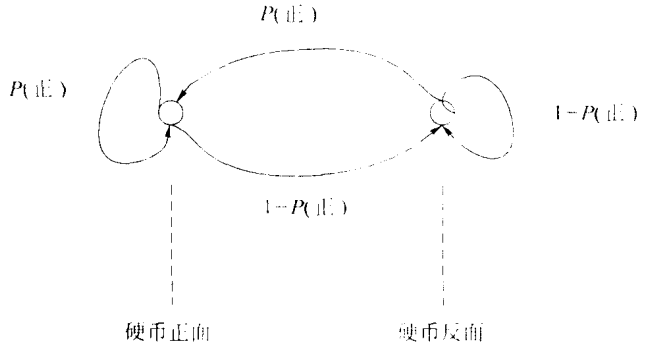

图1.1 只有一枚硬币在手的系统状态图

假设这个人手里有2枚硬币，情况就复杂多了（见图1.2）。同样是两个状态，但这两个状态分别对应2枚不同的硬币，系统状态由硬币正反面概率及两个状态之间的转移概率来描述。2枚硬币之间转换的物理机制不同于抛硬币的机制，需要用另一个随机事件来描述

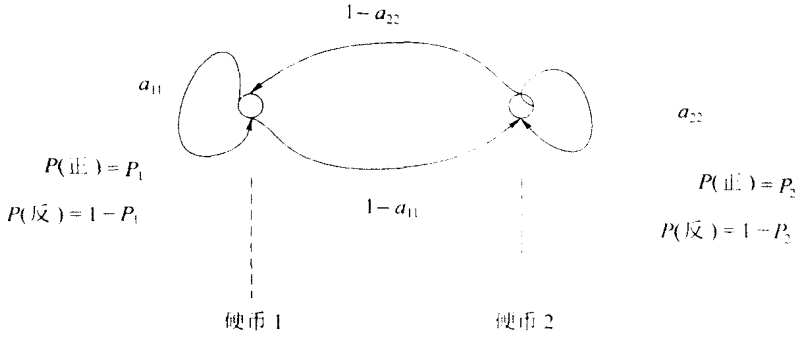

图1.2 有2枚硬币在手的系统状态图

这里需要估计的模型参数有4个：a_{11}、a_{22}、P_1、P_2；而在只有一个硬币的情况下，需要

估计的参数只有一个：P(正)。

如果考虑这个人手里有3枚硬币，情况就更复杂了(见图1.3)：模型中状态有3个；对应于3枚不同的硬币，系统所需估计的参数有9个之多。

这里牵涉到一个问题：上述3个模型中哪个模型能更好地模拟实际发生的过程？从上面的例子可以看到：自由度越大，模型越复杂，模型本身具有模拟实际过程的能力也更强。在实际应用中，常需要考虑模型的大小，模型过大，不仅计算复杂度增加，收敛性慢，模拟效果也未必好。如上述例子中，如果那个人只用1枚硬币，而我们却采用3枚硬币的模型，就未必合适。相反，如果模型较小，虽然所需估计的参数不多，但不能反映实际过程，效果同样比较差。因此，模型大小问题是HMM所要考虑的一个至关重要的问题。

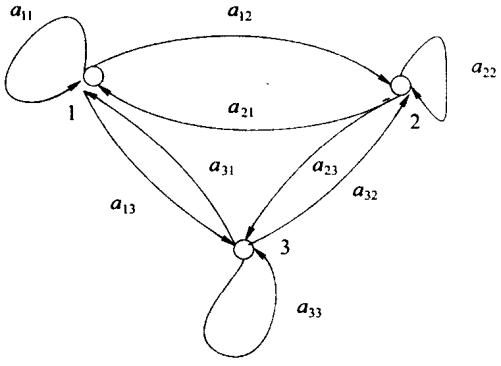

图1.3 3枚硬币的状态图

现在我们来详细讨论HMM。一个HMM所需考虑的基本参数有：

(1) N—— 模型状态个数，状态集为$\{s_1, s_2, \cdots, s_N\}$，$t$时刻系统状态为$q_t$。

(2) M—— 每个状态所能产生的观察值的个数。在上面所举的例子中，观察值即硬币的正反面。

(3) $\boldsymbol{A} = \{a_{ij}\}$—— 状态转移概率分布。$a_{ij} = P[q_t = s_j | q_{t-1} = s_i], 1 \leqslant i, j \leqslant N$。

(4) $\boldsymbol{B} = \{b_j(k)\}$—— 在状态$j$条件下，观察值$v_k$的概率分布，即$b_j(k) = P[v_k \text{ at } t | q_t = s_j], 1 \leqslant j \leqslant N, 1 \leqslant k \leqslant M$。

(5) $\boldsymbol{\pi} = \{\pi_i\}$—— 初始状态分布。$\pi_i = P[q_1 = s_i], 1 \leqslant i \leqslant N$。

HMM参数集可以简单地表示为$\boldsymbol{\lambda} = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$。利用HMM解决的实际问题大致可以归为三类：

问题一：给定观察序列$O = O_1 O_2 O_3 \cdots O_T$和模型$\boldsymbol{\lambda} = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$，怎样来计算概率$P(O | \boldsymbol{\lambda})$？

这个问题的实质是对几个竞争模型进行打分评价的过程，以便由此来选择一个最合适的模型。

问题二：给定观察序列$O = O_1 O_2 O_3 \cdots O_T$和模型$\boldsymbol{\lambda} = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$，怎样重现一个相应的状态序列$Q = q_1 q_2 q_3 \cdots q_T$？

实际上，能真正做到完全重现原来的状态序列是不可能的。因为这个模型本身是一个随机模型，只能说选择一个最接近原来状态序列的序列—— 这是一个优化过程。选择优化度量函数是一个难点。

问题三：优化模型参数集合$\lambda=(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\pi})$，使$P(O|\lambda)$最大。

优化模型过程是一个学习过程，用训练样本来训练模型可以达到优化目的。

现在分别来看看HMM的三个基本问题的解决方案。

解决问题一　给定观察序列$O=O_1O_2O_3\cdots O_T$和模型参数集合$\lambda=(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\pi})$，计算$P(O|\lambda)$？

我们首先来看看在给定的状态序列$Q=q_1q_2q_3\cdots q_T$下，产生观察序列$O=O_1O_2O_3\cdots O_T$的概率，并且假设各个观察值之间在统计意义上是独立的，也就是说，某一时刻系统在某特殊状态下产生的某观察与其他时刻无关，即

$$P(O|Q,\lambda)=\prod_{t=1}^{T}P(O_t|q_t,\lambda)=b_{q_1}(O_1)b_{q_2}(O_2)\cdots b_{q_T}(O_T) \tag{1.2.2.5}$$

而在模型λ下，产生这样一串状态序列$Q=q_1q_2q_3\cdots q_T$的概率为

$$P(Q|\lambda)=\pi_{q_1}a_{q_1q_2}a_{q_2q_3}\cdots a_{q_{T-1}q_T} \tag{1.2.2.6}$$

现在计算给定模型λ下，观察序列O和状态序列Q的联合概率，即

$$P(O,Q|\lambda)=P(O|Q,\lambda)\cdot P(Q|\lambda) \tag{1.2.2.7}$$

为了计算$P(O|\lambda)$，需要对上式在各种可能的状态序列下求和，即

$$\begin{aligned}P(O|\lambda)&=\sum_{\text{all }Q}P(O,Q|\lambda)=\sum_{\text{all }Q}P(O|Q,\lambda)P(Q|\lambda)\\&=\sum_{q_1q_2\cdots q_T}\pi_{q_1}b_{q_1}(O_1)a_{q_1q_2}b_{q_2}(O_2)\cdots a_{q_{T-1}q_T}b_{q_T}(O_T)\end{aligned} \tag{1.2.2.8}$$

上述给出了求$P(O|\lambda)$的过程，但如果直接用(1.2.2.8)式来计算$P(O|\lambda)$，所需的计算次数为$2T\cdot N^T$。因为针对每一时刻$t(t=1,2,\cdots,T)$有N种可能状态，所以所有可能的状态序列数为N^T；而针对每一种状态序列，需做$2T$次乘。精确地说，需要$(2T-1)\cdot N^T$)次乘，需做N^T-1次加。这样的计算量是非常庞大的，即便对于数值较小的N和T，也是一个天文数字，如$N=5$(5个状态)，$T=100$，所需计算次数的数量级为$2\cdot100\cdot5^{100}\approx10^{72}$。

很明显，我们需要找到一个有效算法(是指在多项式时间内能完成的算法)。常用的算法是前向-后向算法(Forward-backward Algorithm)。

一个前向变量$\alpha_t(i)$定义为

$$\alpha_t(i)=P(O_1O_2\cdots O_t,q_t=s_i|\lambda) \tag{1.2.2.9}$$

它的物理含义是：对给定模型λ，计算从开始到t时刻的观察序列为$O_1O_2\cdots O_t$而t时刻系统状态为s_i的概率。我们可以采用递归形式来计算$\alpha_t(i)$。

(1) 初始化：

$$\alpha_1(i)=\pi_ib_i(O_1)\qquad 1\leqslant i\leqslant N \tag{1.2.2.10}$$

(2) 递归计算：

$$\alpha_{t+1}(j)=\left[\sum_{i=1}^{N}\alpha_t(i)a_{ij}\right]b_j(O_{t+1})\qquad 1\leqslant t\leqslant T-1,1\leqslant j\leqslant N \tag{1.2.2.11}$$

(3) 最终结果：