

第1章 概 论

多媒体通信是指通常可由人类感官感知的多种类型信息的结合。在此背景下，视听信息（语音/音频/声音/图像/视频/图形）显得尤为重要，但是就信号处理的复杂度而言，对视听信息的处理又面临很大挑战。多媒体信息的数字表示为数据的交换、分发、获取提供了更大的自由度，其中的通信包括人与人之间、人与设备之间，或设备与设备之间的信息交换。对视听数据进行再现，总是希望获得良好的质量，但是，用于数据传输的比特数量通常是有限的，因此压缩技术成为一项关键技术。本章将介绍多媒体信号压缩和传输中所涉及的基本概念和术语，并对视听信号源数字表示的获取、常见格式和数据量进行概述。

1.1 概念和术语

通信系统设计的基本原则是，以尽可能低的代价传递信息，同时为用户提供足够高的质量，使用户最大限度地受益。多媒体通信中，由于数据量大而对传输速率和处理速度要求较高的信息类型是最为重要的，包括音频(audio)、图像(image)、视频(video)和图形(graphics)等信源。对于这些类型的数据，通信系统(网络和设备)往往会在极限状态下工作。因此，在开发具有更强处理能力和更大存储空间的设备以及具有更高容量的网络时，视听信号的处理、存储和传输已经成为最为重要的驱动因素之一。多媒体信源有的是由传感器采集产生的(比如，能够完成模数转换的相机、话筒等)，有的是通过合成获得的(比如，图形、合成的语音或声音等)。为了将信息再现以供人们使用，通常需要进行其他转换，将信息搭载到相应的物理媒介上(比如，通过扬声器产生声波)。

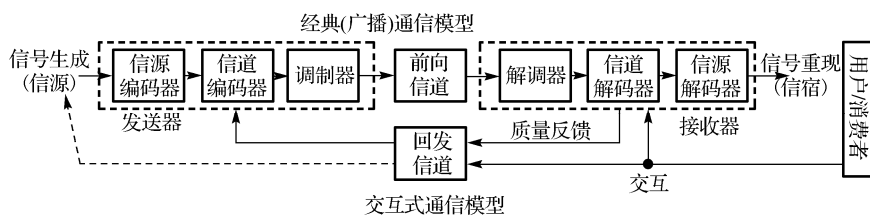


图 1.1 经典通信系统与反馈/交互信道的概念

单向通信系统的经典概念如图 1.1 的上半部分所示。该模型适用于不存在回发信道的应用环境，比如广播。左侧的设备(发射器)与右侧的设备(接收器)是相辅相成的。信源编码器与解码器、信道编码器与解码器，采用协议规范互相配合工作，设法补偿信道损失和信道容量的局限。这种系统的目标是以尽可能高的质量传递尽可能多的信息。而这是通过优化信源编码器、信道编码器和调制器实现的：信源编码器将信源压缩，以尽可能少的比特数表示信源；信道编码器保护码流免于可能产生的损失；调制器以最合适的方式在物理信道中发送符号载体。考虑到信道特性(比如，由于干扰和噪声造成的信道损耗)以及可用的物理带宽，有必要对上述各模块进行联合优化。信道具有某一容量(capacity)，表示该信道能在发射器和接收器之

间进行无损失传输时的最大信息速率。在这种意义上,信源编码的主要任务是,将通常存在于多媒体信源原始信息中的冗余去除(remove redundancy),并以较低的码率将其压缩。信道编码加入冗余(add redundancy),这是为了当出现信道损失时,为码流提供差错保护。信源产生的码率不可以超过信道容量,如果出现这种情况,要么会产生信息损失,要么需要采用更有效的方法对信源进行压缩(见 1.1.1 节)。而后者可能是更好的选择,因为它受到编码器的控制,即便在引入较多失真时,也可以产生比较平稳的降质,而不会造成随机的信道损失。

在通信的经典概念中,接收器会被动地、确定性地解码一切经信道接收到的信号。该模型适用于广播系统,这种系统没有必要对某一用户的干预做出反应,但是,在点对点通信系统中,情况便不同了。经典的通信系统模型通过回发信道和交互组件加以扩展,如图 1.1 的下半部分所示。用户本身能够与系统链路中的任何环节进行交互并产生影响,其中包括信号的产生环节。而且,接收器的任何一个环节都可以自动地产生反馈,例如,请求重传丢失的信息,或者通知发射器当前信道的状况等。更“智能”的系统甚至包括信号内容分析等器件,用来协助用户完成交互以及搜索多媒体数据。

除传统的广播信道外,如今往往需要通过特征差异非常大的异构信道(heterogeneous channels)完成多媒体数据流的传输(例如,基于互联网的物理传输)。尤其在无线和移动传输中,由于信道特征的瞬时波动,需要频繁而且适当地对传输进行调整。从广义的角度来看,信道可以是网络或者存储设备,其中不仅可以在传输链路的末端(服务器和客户端)加入存储设备,还可以在代理服务器中设置临时存储设备,用于快速访问可能会被多次存取的内容。当出现传输错误时,在客户端进行信息恢复,要远比传统的信道和信源解码复杂得多,比如,需要采用包括内容分析在内的先进的差错隐藏方法。

归根到底,可以根据为用户带来的价值,即质量与成本之比(质量/成本)来评价一个多媒体通信系统的性能。服务质量(Quality of Service, QoS),这里也可以将其理解为感知质量,得到了广泛的使用。传输速率、延迟和损耗等网络参数,信源编码/解码的压缩性能,以及与信道特性的相互关系都会影响 QoS。向用户收取的费用取决于完成这些任务的设备成本。因此,在满足用户需求的前提下,最好采用复杂度较低的系统。

本书以多媒体信号的编码表示(coded representation)和传输(transmission)为重点进行讨论。与此相关的概念和术语将在本节的余下部分进行简要的介绍。

1.1.1 信源编码的信号表示

通过多媒体信号压缩(multimedia signal compression),在获得尽可能高的视觉质量的条件 下,可以实现最简洁的表示。采集到信号以后,将其转换成具有有限样点个数与幅度等级的数字表示。这一步骤将会对最终的质量产生影响。如果在获取信号时,并不知道某个候选信道能够传输的码率范围或某种应用所需的分辨率,则可取的做法是,以尽可能高的质量采集信号,必要时可在以后对信号进行下采样,以获得质量较差的信号。

冗余(redundancy)(信号样点之间存在的相似性,或某些信源状态出现得较为频繁)是信号本身的属性,它使信源编码器(source coder)能在不造成信息损失的条件下,降低码率。如果针对的用户是人,考虑到感知特性(perceptual properties),由于为用户提供的质量没有必要比其所能(或所希望)感知到的更加精细,因此可以进一步对信源编码方法进行调整。某些发生在信源编码和解码过程中的失真可能不会被感知到,或者失真很小而能够容忍,或者需要进

行修复(比如, 去除传感器噪声)。这些是原始信号中**无关紧要的(irrelevance)**部分, 编码时最好能将其去除^①。然而, 如果信道容量很小, 仅去除冗余或不重要部分可能不足以降低码率。那就难免会引起其他失真。这些失真在本质上是不同的, 例如, 空间或时间分辨率的降低, 或者编码/量化噪声。还可以考虑与**内容相关的(content-related)**属性。例如, 对用户来说更加重要的部分或片段, 可以为其提供较好的质量。

编码信息通常以二进制数[**比特(bits)**]的形式进行表示。码率的单位是**比特/样点(bit/sample)**^②, 或**比特每秒(bit per second)**[**比特/秒(b/s)**], 其中后者是由**比特/样点(bit/sample)**的值乘以**样点/秒(samples/s)**(采样率)的值得到的。**压缩率(compression ratio)**是评价信源编码方案性能的一项重要指标。它表示的是原始未压缩信源和压缩后的信源所需码率之比。比如, 对数字高清电视而言, 如果原始未压缩的信源需要至少 800 Mb/s ^③的码率, 若压缩后的码率须达到 4 Mb/s , 那么所需的压缩率就是 $800:4=200$ 。如果压缩的信号流以文件的形式存储在计算机磁盘上, 可以根据文件的大小, 评价压缩性能。当将文件大小转换为码率时, 需要注意的是, 通常文件的大小是以 **K(千)字节(Kbyte, KB)**, **M(兆)字节(MByte, MB)**等为单位的, 其中 **1 字节(Byte)**由 **8 比特(bit)**组成, $1\text{K 字节(KByte)}=1\,024 \text{ 字节(Byte)}$, $1\text{M 字节(MByte)}=1\,024 \text{ K 字节(KByte)}$ 。

典型多媒体信号编码和解码系统的基本框图, 如图 1.2 所示。

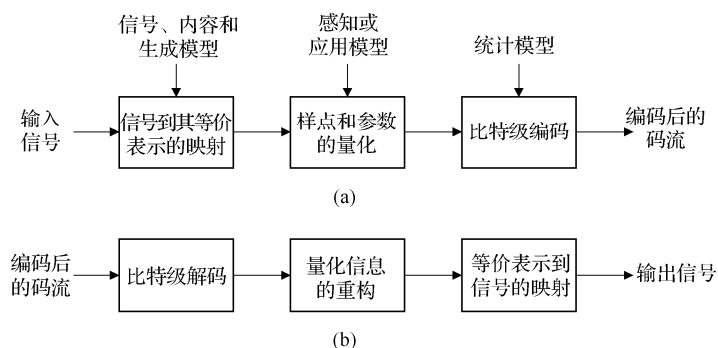


图 1.2 多媒体信号编码系统的基本原理。(a) 编码器; (b) 解码器

- 第一步是信号样点到其等价表示的映射, 理想情况下, 通过反向映射可以完全恢复信号样点。一般来说, 在等价表示空间中的表示要比在原始信号空间中的表示更加稀疏, 所以后续步骤可以变得更简单并能更好地对信号进行优化。稀疏性指的是等价表示需要更少的样点, 或者样点的幅度集中在较少的幅值上。**预测(prediction)**和**变换(transform)**是实现这种映射的重要方法。当采用线性变换完成映射时, 映射又被称为**去相关(de-correlation)**, 这是因为, 映射通常可以去除在等价表示空间中样点之间的线性统计相关性。可以采用统计模型对这种系统进行优化, 同时还需要考虑信号的生

① 当针对的用户是机器时, 信号中哪些部分是无关紧要的, 由信号的使用目的决定。例如, 信源编码引起的失真可能会导致对内容的分析失败, 这是无法接受的。

② 图像信号还使用**比特/像素(bit/pixel, b/p)**, 其中“像素”作为**图像元素(picture element)**的缩写被广泛使用, 即图像样点。为了在用词上对各种类型的信号没有偏向性, 对于图像和视频, 我们也使用“样点”一词。

③ Mb/s : 10^6 比特每秒, kb/s : 10^3 比特每秒。

成(例如, 获取过程的特性)和语义的内容。如果映射函数是可变的(正如具有局部变化特征的多媒体信号通常所需的), 则可能需要产生额外的**边信息**(side information), 在解码时需要这些边信息来完成逆映射。

- 第二步是**量化**(quantization), 它将信号、信号的等价表示或边信息参数映射成离散字符集中的值(如果之前的值并不是离散字符集中的值, 或者需要进行压缩效果更明显的有损压缩)。在量化时, 还应考虑观察者的感知局限或其他非相关性准则(与预期的应用相关), 这样利用给定数量的离散(量化值)状态便可以尽可能多地保持重要信息。一般来说, 量化值状态的个数较少, 自然会使每个样点需要的比特数也较少, 但是也会使失真增加。
- 最后一步是**比特级编码**(bit-level encoding), 这一步的目的是以更低的码率表示离散信息状态, 而通常并不改变这些信息状态。编码的优化主要是以统计原则为基础进行的, 比如各个状态出现的频率。各个状态实现唯一性表示(比如, 有限离散字符集中的字符)所能达到的最低码率称为**熵**(entropy)。

除**码率**(rate)和**失真**(distortion)外, 其他调整信源编码算法的参数还包括**延迟**(latency)和**复杂度**(complexity)。这4个参数是相互影响的。码率与失真之间的关系是由**率失真函数**(rate distortion function)决定的(见2.5.5节和4.2.1节), 当给定失真的上限时, 利用率失真函数可以得到码率的理论下限值。增加编码/解码算法的复杂度往往能够使率失真性能得到提高(就是说, 在保持失真不变的条件下, 获得更高的压缩率), 例如, 根据各种信源的不同统计特性, 采用更复杂的方式实现编解码算法。另外, 增加延迟也可以提高压缩性能, 比如, 如果编码器能够预知当前的决策对后续编码过程所产生的影响, 则将有利于提高编码器的性能。

码率和延迟能够直接利用许多参数进行定量评价, 比如, 均值、峰值和变化量等, 而对失真的定量评价将需要较复杂的度量方式, 其中需要考虑用户的感知, 包括均方误差、信噪比(SNR)等参数被广泛采用, 同时这些参数也较容易进行计算(见3.3节)。对复杂度的定量评价也是个挑战, 因为同一编/解码器的各种组件可能会很容易地在不同的平台上实现, 但也可能会很困难, 并且复杂度的判断需要考虑多种因素, 比如, 逻辑门的数量、芯片尺寸、功耗、内存使用情况、软件/硬件友好性、并行运算能力等。

1.1.2 传输的优化

信源编码器与信道之间的关系对整体感知服务质量也是至关重要的。信源编码器去除信号中的冗余, 而信道编码器向码流中**加入冗余**(add redundancy), 以达到在有损环境下保护和恢复码流的目的。在接收端, 信道解码器去除由信道编码器加入的冗余, 并利用这些冗余进行纠错, 而信源解码器补充由信源编码器去除的冗余, 利用接收到的信息尽可能好地对信源进行重构。由此看来, 信源编码/信道解码与信道编码/信源解码的作用是相似的。实际上, 最复杂的部分通常处于去除冗余的一端, 因为这部分的目标是对完备表示中的**重要信息**(relevant information)进行估计。事实上, 为了获得最优性能, 需要对信源编码和信道编码进行联合优化。比如, 利用信道编码, 向码流中对用户来说不太重要的部分加入冗余, 是没有什么意义的。信道上的传输还包括调制, 在通信技术中, 为了逼近信道容量的极限, 通常会联合使用信道编码和物理层调制^[Proakis, Salehi, 2007]。

在设计多媒体系统时, 利用某一模型把信道看成一个“黑匣子”, 往往是比较有利的。这与误差/损失特性、传输速率、延迟(延时)等对网络质量来说最重要的参数是密切相关的。

当可以保证网络的最低质量时,则几乎能够以最优的方式,使信源编码和网络传输相匹配。这通常需要通过协商协议来实现。如果网络的质量无法保证,可以引入专门的机制使服务器端和客户端相互匹配。这包括基于网络质量评估的专用差错保护机制或重传协议。为了提高传输质量,引入时延也是一种切实可行的方法,例如,优化传输调度,在信息传输到接收端之前采用信息临时缓冲,或者当传输中可能发生突发丢包时,对码流进行置乱/交错(见 9.3 节)。在极端情况下,接收器能够请求重传丢失或已损坏的信息,直到最终收到该信息,但这样做将会增加传输延迟。信道编码部分也同样如此,如果允许采用具有较长时延 (higher latency) 或较高复杂度 (higher complexity) 的系统,则预计将会获得更好的质量。然而,这并不是对所有应用都适用。比如,实时会话服务(比如,视频电话)只允许低时延,对于需要快速反应的所有交互应用,也同样如此。对于移动设备,其电池的容量是最为关键的,因此,一般应维持较低的整体系统复杂度。

从物理上来说,抽象的信道通常是由若干具有不同特性的网络路径组成的链路。此时,整个链路中最薄弱的环节决定了信道的性能。在异构网络中,媒体流具有适应网络特性变化的能力,是非常重要的。在某些情况下,可能需要将码流转码(transcode)成一种不同的形式,使其能够更好地适应特定的网络特性。从这个意义上说,能够产生具有很强适应能力并可以独立于编码过程而被截断的可伸缩码流(scalable streams)的信源编码方法是非常具有优势的(见 5.4.2 节)。另外一种方法是基于内容多描述(multiple descriptions)的编码方法(见 5.4.3 节),这种方法把内容的多个描述经过并行的传输路径传送给接收器。

1.2 信号源与信号的获取

多媒体系统通常处理的是数字表示的信号,而自然信号(natural signals)的获取和产生可能并不是由数字设备直接完成的,例如,转换中可能会涉及电磁(话筒)、光学(镜头)、化学(胶片)媒介。模数转换本身由采样和量化两个步骤组成,采样是将空间-时间连续的信号映射为离散的样点,量化是将幅度连续的信号映射为数值。采集装置,比如配备 CCD 或 CMOS 传感器的数码相机,在输出端本身就能提供采样的信号和量化的表示。多媒体信号也可以通过合成产生,比如,图形、合成声音、人造音效等,还可以通过将自然声音和合成声音混合而产生。即使采集到了自然信号,原本存在于外部(三维)世界的部分信息也会丢失,这是由于:

- 采集设备的频率带宽或分辨率是有限的;
- 采集设备的“非遍及性”,采集设备位于三维外部世界的某一位置,所以只能在这一特定的观察点或收听点获取信号,即使使用多个相机或话筒,采集到的三维空间信息也是不完整的(incomplete),这是因为传感器不可能布置得无限密集。

下面以相机系统为例,讨论外部世界与信号的不完全投影之间的关系。用“世界坐标系”表示位置 $\mathbf{W}=[W_1, W_2, W_3]^T$, 其中,在某一特定位置 $\mathbf{W}$, 某一特定时间 t 的成像特性[又称为全视函数(plenoptic function)] $I(\mathbf{W}, \boldsymbol{\Theta}, t)$, 由来自角度方向 $\boldsymbol{\Theta}=[\theta_1 \ \theta_2 \ \theta_3]^T$ 的光线颜色和强度给出[Mcmillan, Bishop, 1995]。这可以通过位于位置 $\mathbf{W}$ 、光轴(观察线)方向为 $\boldsymbol{\Theta}$ 的相机获取。对于这种“光场”表示,一台相机只能在位置 $(\mathbf{W}, \boldsymbol{\Theta})$ 处获得一个样点,若有其他相机,则能够使这个样点变得更加完整。另外,其他视图还可以通过内插或者外推产生(见 MCA^①, 第 7 章)。

① MCA 表示 *Multimedia Content Analysis* 一书,后文同。——编者注

针孔相机 (pin-hole camera) 模型如图 1.3 所示。这里，世界坐标系以相机所在的位置为原点，其中，相机的光轴沿 W_3 (深度) 轴方向，垂直于图像平面。相机的二维图像平面 (例如，胶片或电子传感器) 在时间 t 接收到由三维外部世界投射到图像平面的投影。在针孔相机模型中，可以保证只有一束 (one) 来自外部世界的光线 (light ray) 可以到达图像平面中的一个点，这样便能够保证图像是聚焦的^①。外部世界坐标系中， $\mathbf{W}_p = [W_{1,p}, W_{2,p}, W_{3,p}]^T$ 位置处的一点 P 与其在图像平面中的投影——点 $\mathbf{t}_p = [t_{1,p}, t_{2,p}]^T$ 之间的关系，由中心投影方程 (central projection equation) 确定，其中 F 是相机的焦距，

$$t_{i,p} = F \frac{W_{i,p}}{W_{3,p}}, \quad i=1,2 \quad (1.1)$$

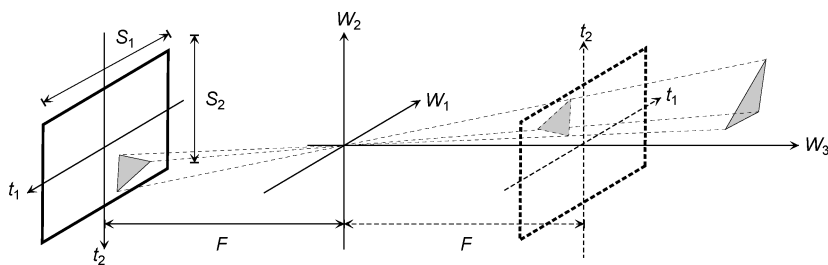


图 1.3 针孔相机模型

世界坐标系的原点位于“针孔”的位置，即相机的焦点。图像平面本身位于 $W_3 = -F$ ，其中心位于 $W_1 = 0$ 、 $W_2 = 0$ 处； $W_2 = 0$ 处的指向为 t_1 轴方向的图像平面，与 $-W_1$ 平行； $W_1 = 0$ 处的指向为 t_2 轴方向的图像平面，与 $-W_2$ 平行。当光线穿过位于 $W_3 = F$ 的镜像平面时，同样的投影方程依然成立，因此，两个坐标轴 t_i 不需要关于相应的 W_i 镜像对称。图像平面本身的宽度为 S_1 ，高度为 S_2 。关于光轴对称的水平视角和垂直视角为

$$\varphi_i = \pm \arctan \frac{S_i}{2F}; \quad i=1,2 \quad (1.2)$$

在电子相机和扫描仪中，二维图像平面通常是按照由左及右/从上到下的顺序逐行进行扫描和采样的。然后，宽度为 S_1 、高度为 S_2 的图像平面被映射到间距为 $T_i = S_i/N_i$ 的 $N_1 \times N_2$ 个离散采样位置。这种采样过的并且空间上有界的图像可以用矩阵表示。一般情况下，将图像左上角的样点作为坐标原点 (0,0)，这个样点也是矩阵左上角的元素^②。

可以用一个包含 $N_1 \times N_2$ 个样点的矩阵对此进行说明，如图 1.4 所示，

$$\mathbf{S} = \begin{bmatrix} s(0,0) & s(1,0) & s(2,0) & \cdots & \cdots & s(N_1-1,0) \\ s(0,1) & s(1,1) & & & & s(N_1-1,1) \\ s(0,2) & & \ddots & \ddots & & \vdots \\ \vdots & & & & & s(N_1-1, N_2-2) \\ s(0, N_2-1) & s(1, N_2-1) & \cdots & \cdots & s(N_1-2, N_2-1) & s(N_1-1, N_2-1) \end{bmatrix} \quad (1.3)$$

① 从物理上来说，这是不可能做到的，因为小孔本身必然具有一定直径。在带有镜头的相机中，只有在镜头聚焦时，来自外部世界同一点的多条光线才能会聚于图像平面中的某一点。

② 需要注意的是，中心投影方程中使用的二维坐标系，其原点位于光轴穿过图像平面的位置 (可能位于现有样点之间的位置)，相对于这一坐标系，此坐标系存在一个偏移量。

在某些情况下(例如,当矩阵中不同行中的值需要参与同一算数运算时),从图像左上角的样点开始,对逐行扫描的图像矩阵中的所有样点进行重排,并将其组成一个向量,会更加方便,

$$\mathbf{s} = [s(0,0) \ s(1,0) \ \cdots \ s(N_1-1,0) \ s(0,1) \ s(1,1) \ \cdots \ \cdots s(0,N_2-1) \ \cdots \ s(N_1-1,N_2-1)]^T \quad (1.4)$$

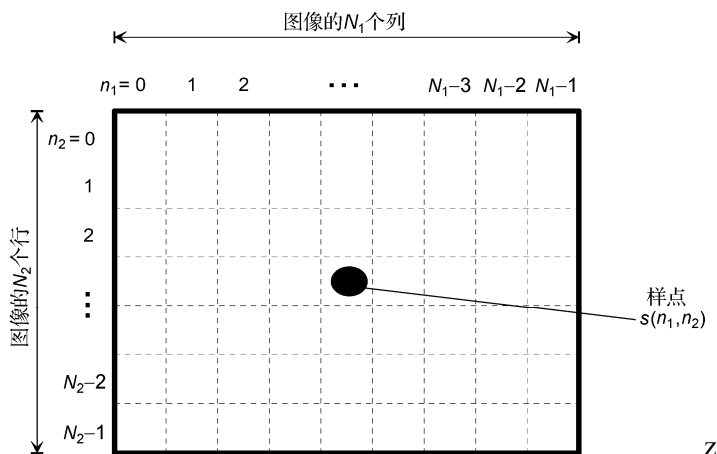


图 1.4 离散图像矩阵的定义

一直以来,图像采样都是首先在时间维度进行的,使用相机拍摄静止图像时,以单个样点进行采样,使用摄像机拍摄影片时,以图像样点序列进行采样。假如采集到的图像之间的时间间隔为 T_3 ,则每秒采集到的图像数量[也称之为帧率(frame rate)]便是 $1/T_3$ 。在电子(模拟)视频摄像机中,采用行结构完成垂直的空间采样,而每行中的信号,在水平方向上^①以及幅度上(未量化)依然是连续的。在数字成像中,信号最终在水平维度上也进行采样,并被量化成用数值表示的幅度值。

在模拟视频技术中(仍存在于一些数字视频相机中)就已经采用了隔行扫描(interlaced)采集(见 2.3.3 节)。在这种情况下,一帧(frame)图像包含两幅场图像(field picture),分别包含偶数行和奇数行。采集到的两幅场图像之间存在半个帧周期的时间间隔。当整帧图像被同时采集时(电影摄像机正是这样做的),这种采集方式被称为逐行(progressive)扫描采集。

在立体和多视图显示中,三维空间感是通过将不同的图像(由两台稍有距离的相机采集,具有立体视差)展示给观察者的左眼和右眼而获得的。多视图显示比立体显示更优越,当观察者在显示器前移动(运动视差)时,它能够通过改变视图来改善空间感。为了获得用于立体和多视图显示的内容,需要利用两个或多个相机对场景进行采集,或者提供深度信息,以合成的方式产生额外的视图,但这会进一步增加存储和传输所需的数据速率(见本书 7.6 节和 MCA, 第 7 章)。

1.3 多媒体信号的数字表示

信号的模数转换过程包括采样(见 2.3 节)和量化(见第 4 章)。得到的“原始”数字格式用

① 需要注意的是,在模拟视频采集和显示中,各行是按时间顺序进行扫描的,因此,原则上对整幅图像来说,并不存在唯一的采集时间。当然,只有在图像中存在运动场景的情况下,讨论这个问题才有意义。

称为脉冲编码调制 (Pulse Code Modulation, PCM) 来表示。在数字多媒体信号处理应用中, 原始格式通常被当作原始参考。在视觉和听觉信号中采用的这些格式, 其具体特性将会在本节进行解释。

1.3.1 图像与视频信号

当对图像和视频信号进行数字化表示时, 如果样点的每个分量用 B 比特来表示^①, 那么利用 PCM 则能够表示 2^B 个不同的灰度级。目前, 消费级照相机、摄像机和扫描仪仍然主要采用 $B=8$ 比特 (256 个灰度级) 表示, $B=10$ 比特甚至更高的表示精度是今后发展的趋势。在专业和科学应用领域中, 每个分量往往用 $B=16$ 比特表示。电子相机通常需要通过非线性映射将亮度映射为幅值, 从而为较暗的亮度值提供较精细的幅值。这称为采集设备的伽马传输特性 (gamma transfer characteristic), 由下面的传输方程 (近似地) 描述:

$$V = c_1 \cdot \Phi^\gamma + c_2 \quad (1.5)$$

其中, Φ 是归一化到最高灰度级的光通量, V 是信号的幅度, c_1 是相机的灵敏度, c_2 是偏移量^②。

在对数字表示进行处理时, 往往不考虑这些非线性特性, 但是在高保真呈现 (打印、显示) 或精确分析中, 这些非线性特性就变得非常重要了。当图像信号以线性幅度尺度进行映射时 (计算机图形学应用中, 通常采用这种做法), 采用 16 甚至更多个比特来表示每个分量是很有必要的。表示图像传感器获取的数据所需的最大比特数, 主要取决于噪声的大小。若传感器噪声水平超过了预期的量化噪声 (见 4.1 节), 则较低重要性位的值是随机的。由于新一代相机传感器 (如 CMOS) 往往具有较小的噪声, 所以使用较大的位深变得越发有意义, 同时通过传感器本身就可以实现较高的幅度动态范围。需要注意的是, 即使是采用 16 比特表示的幅度范围, 与视觉上可分辨的自然光照水平的有效范围 (涵盖的亮度范围超过 10^9) 相比, 仍然相对较低。人眼的虹膜、膈肌和感光受体等生理结构都具备调节功能; 相机也一样, 可以根据光照强度调节相机的快门时间。在高动态范围 (High Dynamic Range, HDR) 成像中, 对同一场景在不同的曝光量下进行多次采集, 并根据场景中某一部分的亮度, 将多次采集的场景合并形成一幅图像, 相当于人为地扩展了这幅图像的位深。然而, HDR 仅在静态场景及相机位置固定的情况下效果较好。否则, 必须进行配准, 如果配准效果不佳, 将会造成伪影。HDR 成像是一个典型的例子, 它存储的图像并不是直接由传感器数据得到的, 而是经过另外的信号处理过程而获得的。对于其他类型信号, 比如医学图像数据, 图像的幅值往往表示组织或骨骼的密度和吸收特性, 在存储图像之前对其进行处理是非常常见的。

除与亮度/颜色相关的数据外, 还可以获取其他形式的图像, 比如深度数据 (例如, 利用距离传感器获得或者利用立体照相装置对深度进行估计) 以及所谓的阿尔法 (透明度) 图, 通常利用这些数据, 根据不同的图像完成场景的融合/合成 (见 MCA, 第 7 章)。

为了采集和显示彩色图像, 用光的三原色——红 (R)、绿 (G)、蓝 (B) 来表示彩色图像是

① 分量分别为: 颜色、亮度或色度, 见下文。

② 例如, ITU-R Rec. BT.709 在两个区间内对传输特性做出了规定, 其中, 较低的光通量区间具有线性特性:

当 $1 \geq \Phi \geq 0.018$ 时, $c_1 = 1.099$; $c_2 = -0.099$; $\gamma = 0.45$;

当 $0 \leq \Phi < 0.018$ 时, $c_1 = 4.500$; $c_2 = 0.000$; $\gamma = 1.00$ 。

最常用的方法,这与人眼的色彩灵敏度是基本吻合的(见 3.1.3 节)。这些分量是分开获取和采样的。当用三个单独的传感器采集颜色时(高端相机正是这样做的),彩色样点的总数是单色图像的 3 倍,而三种颜色可以视为在样点位置 $\mathbf{n}$ 处的一个向量 $\mathbf{s}(\mathbf{n})=[s_R(\mathbf{n}),s_G(\mathbf{n}),s_B(\mathbf{n})]^T$ 。在多光谱表示中,颜色的真实表示甚至可能需要更多的分量,另外,向量中的其他分量也可以是其他数据,比如同一样点位置 $\mathbf{n}$ 处的深度或透明度。

具有单个传感器芯片的相机,采用光学彩色滤镜阵列以空间多路复用的方式采集彩色图像,其中通常采用贝尔模板(Bayer pattern),贝尔模板中绿色样点的数量分别是红色样点和蓝色样点数量的 2 倍(见图 1.5)。这样做是因为人眼的最大灵敏度与绿色的光谱范围几乎是匹配的^①。需要注意的是,这种相机生成的全分辨率 RGB 图像是通过内插处理而获得的,并不是直接获取的。

显示时,也采用与贝尔模板类似的空间多路复用方式。某些高端显示器人为地生成黄色样点(主要利用红色样点和绿色样点,通过插值处理实现),并用其替换掉一半的绿色样点,以获得更加自然的颜色再现。

B	G	B	G	B	G	B	G
G	R	G	R	G	R	G	R
B	G	B	G	B	G	B	G
G	R	G	R	G	R	G	R
B	G	B	G	B	G	B	G
G	R	G	R	G	R	G	R

图 1.5 单芯片传感器 RGB 采样所采用的贝尔模板

原始的 RGB 表示并不能直接用于解释物理光谱及其相关的颜色映射。为此,[Cie, 1931]中定义了基于所谓**标准观察者**(standard observer)的伪 XYZ 色彩空间。它允许规定参考亮度和显示颜色,并能将其映射为**色度图**(chromaticity diagram),色度图对色彩范围进行解释,色彩可以根据色调和饱和度通过混合各个原色而获得^②。

彩色图像和视频通常以一个**亮度**(luminance)分量 Y 和两个**色度**(chrominance)(色差)分量表示。对于 RGB 表示与亮度/色度之间的转换,根据具体的应用领域,存在不同的规定。比如,在标清电视分辨率视频中,通常采用如下的映射规则(按照在 ITU-R rec. BT.601 中的定义):

$$Y = 0.299 \cdot R + 0.587 \cdot G + 0.114 \cdot B$$

$$C_b = \frac{0.5}{0.866} \cdot (B - Y); \quad C_r = \frac{0.5}{0.701} \cdot (R - Y) \quad (1.6)$$

对于高清(High Definition, HD)视频,通常使用在 ITU-R rec. BT.709 第 2 部分中定义的映射规则^③:

$$Y = 0.2126 \cdot R + 0.7152 \cdot G + 0.0722 \cdot B$$

$$C_b = \frac{0.5}{0.9278} \cdot (B - Y); \quad C_r = \frac{0.5}{0.7874} \cdot (R - Y) \quad (1.7)$$

对于超高清(Ultra High Definition, UHD)显示设备,可以预见色域范围会大大扩展。因此,在

① 还需要注意,在贝尔模板中,绿色分量采用五株采样(见 2.3.2 节),这使得绿色分量在水平和垂直方向上的采样频率是红色和蓝色的两倍。

② 见 MCA, 4.1.1 节。用 XYZ 值表示数字图像,能够为具有各种不同输出特性的设备提供比较通用的颜色映射,从这个角度来看,采用这种表示方式通常是具有优势的。

③ 这样做的其中一个目的就是扩展色域,即能够从 YC_bC_r 表示中产生可能的色度范围。但是,由于发光荧光粉的性质决定了 CRT 可显示颜色的范围要受到较多的限制,所以式(1.6)中的映射规则主要对 CRT 进行了优化,而如今的显示器,比如 LCD 能提供更宽的色域,因此,式(1.7)和式(1.8)中的规则能够更好地支持这类显示器。

ITU-R rec. BT.2020 中, 针对原始 UHD 格式, 定义了如下映射规则:

$$Y = 0.2627 \cdot R + 0.6780 \cdot G + 0.0593 \cdot B$$

$$C_b = \frac{0.5}{0.9407} \cdot (B - Y); \quad C_r = \frac{0.5}{0.7373} \cdot (R - Y) \quad (1.8)$$

与原始的 RGB 分量所涵盖的幅度范围相同, Y 分量涵盖的幅度范围也是 $0 \sim A$, 色度分量涵盖的幅度范围是 $-A/2 \sim A/2$ 。把幅度范围限制在 $A_{\min} \sim A_{\max}$, 可以避免运算负担过重, 其中对于亮度, $A_{\min}$ 是指黑色电平, $A_{\max}$ 是指标称峰值电平(相应地对于色度, $A_{\min} = -A_{\max}$ 是指最小的负值, $A_{\max}$ 是指最大的正值)。这可以通过如下的表达式进行调整

$$Y' = Y \frac{A_{\max} - A_{\min}}{A} + A_{\min}; \quad C'_{b/r} = C_{b/r} \frac{A_{\max} - A_{\min}}{A} \quad (1.9)$$

最后, 由于到目前为止所讨论的问题仍然都假设幅值是连续的(虽然幅值是受限的), 所以需要对样点进行量化, 以利用有限数量的比特对样点进行表示。这里采用均匀量化(见 4.1 节)。为简单起见, 假设归一化 $A=1$, 利用下面的运算, 能够得到 B 比特量化精度的无符号整型数^①。

$$DY' = \lfloor Y' \cdot 2^B \rfloor; \quad DC'_{b/r} = \lfloor (C'_{b/r} + 0.5) \cdot 2^B \rfloor \quad (1.10)$$

如果采用的是 RGB 彩色空间, 可以对三原色分别进行式(1.9)和式(1.10)中对 Y 分量所进行的裁剪和舍入运算。需要注意的是, 在转换成量化的 YC_bC_r 表示之后, 并不能保证在 RGB 空间中量化的 B 比特表示能够得到无损的复原(反之亦然)。然而由于在 RGB 空间中颜色的变化是受限的, 所以从压缩的角度看, 转换到 YC_bC_r 空间是很有益处的, 无论在直观感受上还是在统计意义上, 都能让越重要的颜色被表示得越精确。 YC_bC_r 表示还可以减少 3 种 RGB 颜色主分量之间的相关性, 比如, 可以使结构细节(纹理)集中于 Y 分量。与灰度分量的亮度和对比度相比, 人类视觉系统对色度的空间分辨率并不那么敏感, 针对人类视觉系统的这种特性, 对色度成分进行亚采样可以获得进一步压缩的数据(见 3.1.3 节)。然而, 在某些极端情况下, 比如彩色图像中包含尖锐的边缘, 这种亚采样的影响还是很大的。包含合成内容的图像多数属于这种情况, 相机获取的图像较少属于这种情况, 因为边缘总会受到光学系统的影响而产生轻微的模糊。

对于隔行采样的视频, 往往仅在水平方向上对色度成分进行亚采样, 对于逐行采样的视频, 通常在水平方向和垂直方向上都进行亚采样。采样格式通常采用 “ $N_Y:N_{C1}:N_{C2}$ ” 的形式进行表示, 它可以说明各分量样点数量之间的相对关系^②。例如,

- 如果所有三个分量采用的样点数量相同, 则用 “4:4:4” 来表示;
- 如果两个色度分量仅在水平方向上进行采样因子为 2 的亚采样, 这样的采样结构用 “4:2:2” 来表示; “4:1:1” 表示仅在水平方向上进行采样因子为 4 的亚采样;

① 式(1.9)中采用的典型值如下: 对于亮度, $A_{\min} = 1/16$, $A_{\max} = 15/16$, 对于色度, $A_{\max} = 7/16 = -A_{\min}$ 。

② 这些表示方法背后的依据要回溯到模拟视频, 但这些表示方法目前仍然被广泛采用。 N_Y 表示亮度样点的数量与后面两个值所代表的色度样点数量的比值, 其中假设每一行中都存在亮度分量。 N_{C1} 和 N_{C2} 分别表示偶数行和奇数行(第 1 个偶数行是第 0 行)中的色度样点的数量(两种色度成分的样点数量是相同的)。事实上, 仅包含 3 个数字的表示方式并不能独立地在水平和垂直两个方向上完全涵盖三个分量所有可能的采样比。对于 RGB 信源, 用 N_Y 等效地表示绿色分量, 而用两个色度分量等效地代替红色和蓝色分量。

- 如果两个色度分量在水平和垂直两个方向上都进行亚采样, 即水平方向和垂直方向上色度样点的数量是亮度样点数量的一半, 这样的采样结构用“4:2:0”表示。

各种信源采样格式标准还规定了色度成分的亚采样样点相对于亮度样点的位置。

数码相机采集静止图像通常使用的分辨率如表 1.1 所示, 除此之外, 市场上还存在很多其他分辨率格式。3:2 的图像宽高比——**宽度:高度 (width:height)**是最普遍被采用的, 它与传统的横向照片格式是相同的; 还有一些格式是源于早期计算机显示器 4:3 的宽高比; 如今, 由于大部分显示器和投影仪都采用 16:9 的宽高比, HD 和 UHD 视频的图像宽高比和分辨率也随之出现, 并应用于静止图像(见表 1.2)。通常, 各种格式中的图像在宽度和高度方向上包含的样点数量都是 16、32 或 64 的倍数, 这对基于块的压缩算法, 比如变换编码, 是非常有利的。因此, “6 百万像素”的格式并不见得得有 6 000 000 个样点。表 1.1 中还列出了样点的总数, 以及当采用 $B=8$ 比特精度表示图像时, 每个分量所占的比特数。

表 1.1 数字静止图像的分辨率格式

	640×480 (VGA)	1024×768 (XGA)	1536×1024 (1.5MPix)	2048×1536 (3MPix)	3072×2048 (6MPix)	4256×2848 (12MPix)	5472×3648 (20MPix)
样点数量	307 200	786 432	1 572 864	3 145 728	6 291 456	12 121 088	19 961 856
图像宽高比	4:3	4:3	3:2	4:3	3:2	3:2	3:2
每个分量的比特数 量, 8 比特精度	2 457 600	6 291 456	12 582 912	25 165 824	50 331 648	96 968 704	159 694 848

一般来说, 表示原始数据所需的数据量, 可以根据以下因子中至多 5 个进行相乘运算而求得: 宽度、高度、分量的个数、每个分量的比特深度、色度亚采样格式(若适用)以及图像的数量(当数据是图像序列或多视图视频时)。

对于视频表示, 除存储一个视频所需的比特总数外, **每秒的比特数 (number of bits per second)**也是很重要的, 这个值可以通过将每幅图像所需的比特数与每秒钟播放的图像数量相乘求得。普遍采用的视频格式的一些参数如表 1.2 所示。在某些情况下, 样点的数量、帧率等是可以变化的。

标清电视分辨率的数字电视格式起源于欧洲 625 行的模拟电视信号(或美国和日本的 525 行)。这些模拟信号以 13.55 MHz 的采样率对亮度进行采样。在把**垂直消隐间隔 (Vertical Blanking Interval, VBI)**去除之后, 还剩 575 (480) 个有效行。表 1.2 中列出的数字格式仅存储有效样点和水平消隐间隔中少量**多余样点 (surplus samples)**带来的很小额外开销。日本和美国传统上采用大约 60 场/秒 (30 帧/秒)^①的模拟格式, 而在欧洲, 模拟电视格式 (PAL, SECAM) 采用的是 50 场/秒 (25 帧/秒)。就帧率和场率而言, 定义 HD 格式的数字标准是更加灵活的, 对于逐行扫描, 允许的帧率包括 24、25、29.97、30、50 和 60 帧/秒, 对于隔行扫描, 允许的场率包括 50 场/秒和 60 场/秒。计算机显示器或移动设备采用的格式还包括 Common Intermediate Format (CIF)、VGA、Quarter VGA (QVGA) 和 Wide-screen VGA (WVGA), 但是这其中越来越多的格式预计将被高清格式所取代。目前, **超高清 (Ultra HD, UHD)**格式已经出现。与 HD1080 相比, UHD 格式在水平和垂直方向上包含的样点数量通常是其 2 倍(称为“4K×2K”), 或 4 倍(称为“8K×4K”)。尽管已经采用了逐行扫描采样, 但是帧率在未来还有可能进一步得到提高(72 帧/秒甚至更高)。

① 它们通常被称为 NTSC 格式。模拟 NTSC 格式精确的帧率是 $30 \times (1000/1001) \approx 29.97$ 帧/秒。

表 1.2 数字视频信源格式及其性质

	格式						
	QVGA/ QWVGA	CIF/ SIF	VGA/ WVGA	ITU-R BT.601 (SDTV)	ITU-R BT.709 ^① (1080×)	SMPTE 296M (HD 720p)	ITU-R BT.2020 (4K×2K/8K×4K)
采样频率(Y) (MHz) ^②	—	—	—	13.55	74.25 ... 148.5	74.25	—
采样结构*	P	P	P	I	P/I	P	P
样点/行(Y)	320 (432)	352	640 (864)	720	1920	1280	3840/ 7680
行数(Y)	240	288 (240)	480	576 (480)	1080	720	2160/ 4320
色彩采样格式	4:2:0	4:2:0	4:2:0	4:2:2	4:2:2	4:2:2	4:4:4/ 4:2:2/ 4:2:0
图像宽高比	4:3 (16:9)	4:3	4:3 (16:9)	4:3	16:9	16:9	16:9
帧率(Hz)	10~30	10~30	24~60	25 (30)	24~60	24~60	24~120
PCM 比特深度(b)	8	8	8	8	8~10	8~10	10~12
数据速率(未压缩的)							
每帧(KByte)	112.5	148.5	450	810	4 050	1 800	30 375
1KByte=1024×8 b	(151.8)	(123.8)	(607.5)	(675)	...	...	...
每秒(Mb/s)	9.22	10.13	88.44	165.9	5 063 796.2	2 250 353.9	145×10 ³ 2 986
	...	...	...	...	...	...	...
	37.3	36.5	298.6		2 488	1 106	1435×10 ³

* P/I: 逐行/隔行

QVGA 和 HDTV 之间所支持的采样图像区域(假设样点大小/密度是相同的), 如图 1.6 所示。增加样点数量, 可以提高分辨率(空间细节), 也可以让场景在一个更大的区域上显示。例如, 人脸的特写镜头很少出现在影院的电影中。显示在影院荧幕上的影片, 可以让观察者的眼睛仔细观察荧幕不同位置处的场景, 而在标准清晰度的电视屏幕甚至更小尺寸的屏幕上, 这种能力就非常有限, 因而场景多以特写镜头的方式呈现。

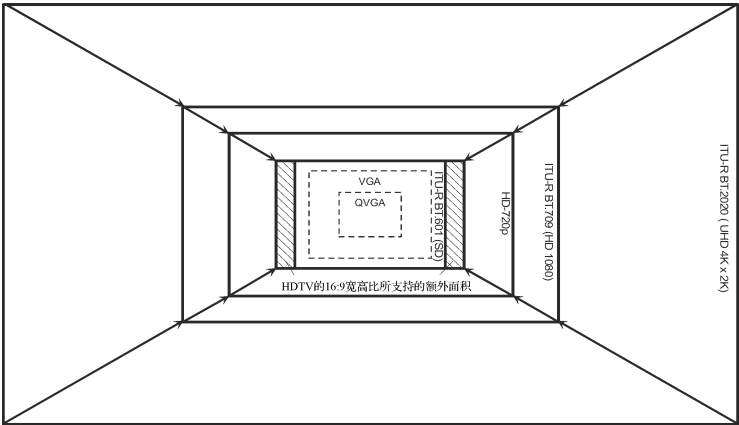


图 1.6 数字视频/图像格式^③的尺寸和宽高比

① BT.709 第 2 部分对 “common image format” 做出了规定, 并在数字高清视频领域得到了广泛的使用。
② 像 BT.2020 这种全数字格式规范, 不能列出实际采样频率。
③ 受空间限制, 没有给出 8K×4K 格式。

在医学和科学应用中,使用的数字图像的分辨率远高于电影作品中所采用的分辨率,高达 $10\,000\times 10\,000=100\,000\,000$ 的分辨率都很常见。

1.3.2 语音和音频信号

对于音频信号,其数字表示所需的数据速率,受到采样率、精度(比特深度)以及声道的数量等参数的影响最大。这些参数高度依赖信号的特性,以及对质量的要求。在语音信号的量化中,采用的是使用对数幅值压缩的非线性映射^①,对于低幅值,即使仅使用8比特/样点进行量化,也可以获得与采用12比特量化一样低的量化噪声。若想获取具有CD音质的音乐信号,至少需要采用16比特线性表示。对于一些特殊应用,往往还采用比CD常用的更高的比特深度和更高的采样速率。在语音和宽带音频的数字表示中,一些典型的参数如表1.3所示。所需的码率(每秒)是通过计算采样速率、比特深度和音频通道个数三者的乘积而求得的。存储所需的比特数,是码率与音轨时长的乘积。

1.3.3 压缩技术的必要性

由于对原始未编码的信号进行表示需要的数据量是极大的,尽管随着通信技术的发展,可用的传输带宽得到了进一步提高,但是图像、视频和音频应用永远都存在着对数据压缩的强烈需求。一般而言,以往的经验已经说明,多媒体数据流量增长的速度比信道容量增加的速度要快,并且数据压缩传输的成本本身就较低。假如信道的容量足够大,如果把信号的分辨率提高,那么就服务用户的质量而言,信道便得到了更有效的利用。利用更有效的压缩,甚至可以在不提高数据速率的条件下,实现对信道更加有效的利用。另外,某些类型的通信信道(尤其在移动传输中)本身就是稀缺且昂贵的,并且由于物理极限性,对带宽利用率(信道单位带宽每秒能够可靠传输的比特数)有更加严格的限制。但是,还必须与压缩算法的实现复杂度进行权衡,高复杂度会导致设备成本增加以及耗电量增加,这对移动设备是尤为关键的。

表 1.3 音频信号的采样与原始码率

	频率范围[Hz]	采样率[KHz]	信道个数与 PCM 分辨率[比特/样点]	最大 PCM 码率[kb/s]
电话语音	300~3 400	8	8~16	64~128
宽带语音	50~7 000	16	8~16	128~256
CD 音频	10~20 000	44.1	2×16(立体声)	1410
多声道 DVD	10~22 000	48	(5+1)×16	4.6×10^3
DVD 音频	10~44 000	最高为 96	2×24	最高为 4.6×10^3
多声道音频 (5+1, 7+2, M+N)	10~44 000	最高为 96	(M+N)×24	最高为 18.4×10^3
环绕声音频	10~44 000	48	(最高为 200)×16	最高为 153.6×10^3

1.4 习题

习题 1.1

时长为150分的一部影片,采用逐行HD格式进行数字表示(每帧 1920×1080 个亮度样点, $2\times 960\times 1080$ 个色度样点,60帧/秒,8比特/样点PCM)。

^① 称为 μ 律和A律量化,见4.1节。

- a) 如果以原始格式存储整部影片，需要多大的硬盘存储容量？
- b) 如果采用 HEVC 压缩算法对这部影片进行编码，若其压缩率是 PCM 的 200 倍。现在要把码流存储在硬盘上，码流文件的尺寸是多少？
- c) 为了在卫星信道上传输这部影片，采用信道编码以提供差错保护。这样做会使码率增加 10%。那么在信道上传输该码流时，所需的传输带宽(Mb/s)是多少？

习题 1.2

通过 LTE 网络传输智能手机拍摄的视频，假设数据速率是 720 kbit/s。用于宽带语音传输的数据速率是 20 kb/s。时域采样率是 30 帧/秒。

- a) 由于视频编码器输出的不是恒定速率的码流，在通过信道传输之前，需要对信息进行缓冲。如果由缓冲引起的最大延迟是 100 ms，那么所允许的最大缓冲区大小(以比特为单位)是多少？
- b) 可以通过调节量化器的设置(离散步长“Q 因子”)改变编码器的输出码率。作为经验法则，Q 因子每降低一个步长将会使码率提高 1.1 倍。已知编码一幅图像所使用的比特数为 19 280 比特，为了避免缓冲器的上溢和下溢，Q 因子能够降低和提高的步长分别为多少？