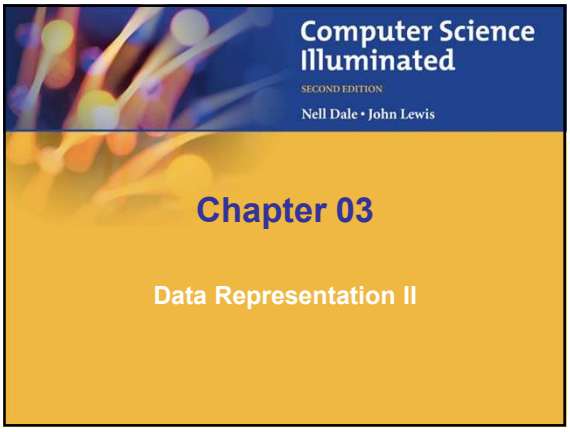


Chapter 03

Data Representation II

Chapter Goals

- Distinguish between analog and digital information.
- Describe the characteristics of the ASCII and Unicode character sets.
- Explain data compression and calculate compression ratios.
- Explain how RGB values define a color.
- Explain the nature of sound and its representation.

3-2

Data and Computers

- Computers are **multimedia** devices, dealing with a vast array of information categories. Computers store, present, and help us modify
  - Numbers
  - Text
  - Audio
  - Images and graphics
  - Video

3-3

Binary Representations

- One bit（位） can be either 0 or 1. Therefore, one bit can represent only two things.
- To represent more than two things, we need multiple bits. Two bits can represent four things because there are four combinations of 0 and 1 that can be made from two bits: 00, 01, 10,11.

3-4

Binary Representations

- In general, n bits can represent 2^n things because there are 2^n combinations of 0 and 1 that can be made from n bits. Note that every time we increase the number of bits by 1, we double the number of things we can represent.

3-5

练习一下

- a class has students up to 100;
- a school has classes up to 50；
- Question？
  - Minimum number of bits to represent each student of the class.
  - Minimum number of bits to represent each class of the school.

3-6

读“数”（连续量与离散量）

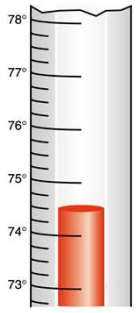


Figure 3.1
A mercury thermometer continually rises in direct proportion to the temperature

3-7

Analog and Digital Information

- Computers are finite. Computer memory and other hardware devices have only so much room to store and manipulate a certain amount of data. The goal, is to represent enough of the world to satisfy our computational needs and our senses of sight and sound.

3-8

Analog and Digital Information

- Information can be represented in one of two ways: **analog** or **digital**.

Analog data A continuous representation, analogous to the actual information it represents.

Digital data A discrete representation, breaking the information up into separate elements.

A mercury thermometer is an analog device. The mercury rises in a continuous flow in the tube in direct proportion to the temperature.

3-9

Analog and Digital Information

- Computers, cannot work well with analog information. So we **digitize** information by breaking it into pieces and representing those pieces separately.
- *Why do we use binary?* Modern computers are designed to use and manage binary values because the devices that store and manage the data are far less expensive and far more reliable if they only have to represent on of two possible values.

3-10

Representing Text

- To represent a text document in digital form, we need to be able to represent every possible character that may appear.
- There are finite number of characters to represent, so the general approach is to list them all and assign each a binary string.
- A **character set** is a list of characters and the codes used to represent each one.
- By agreeing to use a particular character set, computer manufacturers have made the processing of text data easier.

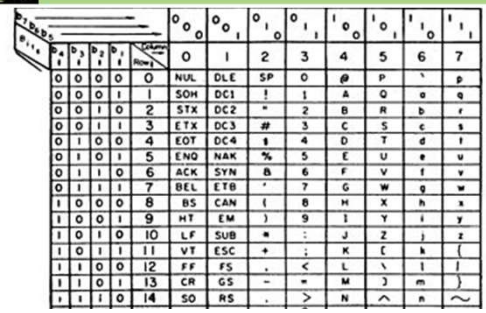
3-11

The ASCII Character Set

- ASCII stands for American Standard Code for Information Interchange. The ASCII character set originally used seven bits to represent each character, allowing for 128 unique characters.
- Later ASCII evolved so that all eight bits were used which allows for 256 characters.

3-12

The ASCII Character Set



	0	1	2	3	4	5	6	7
0	NUL	DLE	SP	0	@	P	`	p
1	SOH	DC1	!	1	A	Q	a	q
2	STX	DC2	"	2	B	R	b	r
3	ETX	DC3	#	3	C	S	c	s
4	EOT	DC4	$	4	D	T	d	t
5	ENQ	NAK	%	5	E	U	e	u
6	ACK	SYN	&	6	F	V	f	v
7	BEL	ETB	'	7	G	W	g	w
8	BS	CAN	(	8	H	X	h	x
9	HT	EM	)	9	I	Y	i	y
10	LF	SUB	*	:	J	Z	j	z
11	VT	ESC	+	;	K	[	k	{
12	FF	FS	,	<	L	\	l	|
13	CR	GS	-	=	M	]	m	}
14	SO	RS	.	>	N	^	n	~

A chart of ASCII from a 1972 printer manual

3-13

The ASCII Character Set

- Note that the first 32 characters in the ASCII character chart do not have a simple character representation that you could print to the screen.

3-14

The Unicode Character Set

- The extended version of the ASCII character set is not enough for international use.
- The Unicode character set uses 16 bits per character. Therefore, the Unicode character set can represent 256, or over 65 thousand, characters.
- Unicode was designed to be a superset of ASCII. That is, the first 256 characters in the Unicode character set correspond exactly to the extended ASCII character set.

3-15

The Unicode Character Set

Code (Hex)	Character	Source
0041	A	English (Latin)
042F	Я	Russian (Cyrillic)
OE09	ฉ	Thai
13EA	Ꮹ	Cherokee
211E	℞	Letterlike Symbols
21CC	⇌	Arrows
282F	⠯	Braille
345F	㑯	Chinese/Japanese/ Korean (Common)

Figure 3.6 A few characters in the Unicode character set

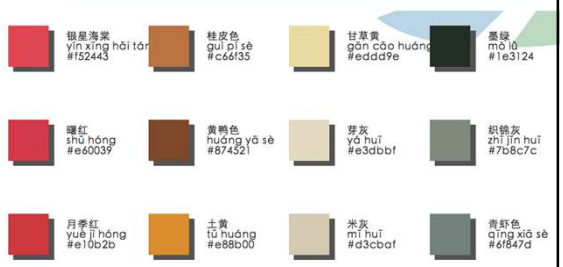
3-16

汉字字符编码举例

字符	ASCII	Unicode	UTF－8	GBK
A	41	00 41	41	—
汉	—	6C 49	E6 B1 89	BA BA

3-17

中国传统颜色



http://chinese.traditionalcolors.com/

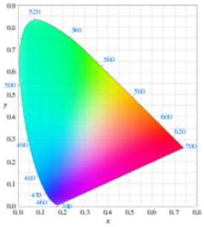
3-18

Representing Color

- Color is our perception of the various frequencies of light that reach the retinas of our eyes.
- Our retinas have three types of color photoreceptor cone cells that respond to different sets of frequencies. These photoreceptor categories correspond to the colors of red, green, and blue.

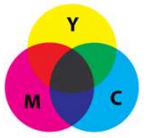
3-19

Representing Models of Color



CIE 1931 color space

Additive color mixing

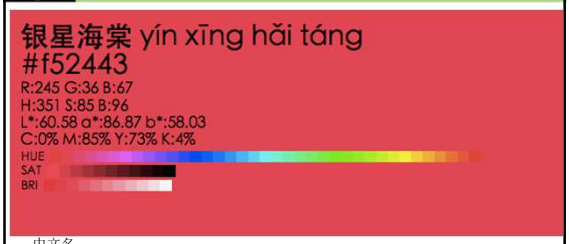


Subtractive color mixing

http://en.wikipedia.org/wiki/Color

3-20

Representing Models of Color



中文名
16进制RGB表达
RGB
HSB
CYM

3-21

Representing Color

- The amount of data that is used to represent a color is called the *color depth*.
- *HiColor* is a term that indicates a 16-bit color depth. Five bits are used for each number in an RGB value and the extra bit is sometimes used to represent transparency. *TrueColor* indicates a 24-bit color depth. Therefore, each number in an RGB value gets eight bits.

3-22

Representing Images and Graphics

RGB Value			Actual Color
Red	Green	Blue	
0	0	0	black
255	255	255	white
255	255	0	yellow
255	130	255	pink
146	81	0	brown
157	95	82	purple
140	0	0	maroon

3-23

Digitized Images

- Digitizing a picture is the act of representing it as a collection of individual dots called **pixels**.
- The number of pixels used to represent a picture is called the **resolution**.
- The storage of image information on a pixel-by-pixel basis is called a **raster-graphics format**. Several popular raster file formats including bitmap (BMP).

3-24

Digitized Images

Figure 3.12 A digitized picture composed of many individual pixels

3-25

Digitized Images

Figure 3.12 A digitized picture composed of many individual pixels

3-26

课堂练习

Xx买了部500万像素的拍照手机（输出2560×1920）

请问
（1）假设照片色彩用true color RGB raster format存储，它需要多少bytes?

（2）假设打印照片需要300dpi输出，最合适打印多大尺寸的照片？
7”（5×7英寸） 8”（6×8英寸） 10”（8×10寸）

3-27

Text Compression

- It is important that we find ways to store and transmit text efficiently, which means we must find ways to compress text.
  - keyword encoding
  - run-length encoding
  - Huffman encoding

3-28

Keyword Encoding

- Given the following paragraph,

  The human body is composed of many independent systems, such as the circulatory system, the respiratory system, and the reproductive system. Not only must all systems work independently, they must interact and cooperate as well. Overall health is a function of the well-being of separate systems, as well as how these separate systems work in concert.

3-29

Keyword Encoding

- Frequently used words are replaced with a single character. For example,

Word	Symbol
as	^
the	~
and	+
that	$
must	&
well	%
those	#

3-30

Keyword Encoding

- The encoded paragraph is

  The human body is composed of many independent systems, such ^ ~ circulatory system, ~ respiratory system, + ~ reproductive system. Not only & each system work independently, they & interact + cooperate ^ %. Overall health is a function of ~ %- being of separate systems, ^ % ^ how # separate systems work in concert.

3-31

Keyword Encoding

- There are a total of 349 characters in the original paragraph including spaces and punctuation. The encoded paragraph contains 314 characters, resulting in a savings of 35 characters. The compression ratio for this example is 314/349 or approximately 0.9.
- The characters we use to encode cannot be part of the original text.

3-32

Run-Length Encoding

- A single character may be repeated over and over again in a long sequence. This type of repetition doesn't generally take place in English text, but often occurs in large data streams.
- In run-length encoding, a sequence of repeated characters is replaced by a *flag character*, followed by the repeated character, followed by a single digit that indicates how many times the character is repeated.

3-33

Run-Length Encoding

- AAAAAAA would be encoded as *A7
- *n5*x9ccc*h6 some other text *k8eee would be decoded into the following original text nnnnnxxxxxxxxxccchhhhhh some other text kkkkkkkkeee
- The original text contains 51 characters, and the encoded string contains 35 characters, giving us a compression ratio in this example of 35/51 or approximately 0.68.
- Since we are using one character for the repetition count, it seems that we can't encode repetition lengths greater than nine. Instead of interpreting the count character as an ASCII digit, we could interpret it as a binary number.

3-34

Huffman Encoding

- Why should the character "X", which is seldom used in text, take up the same number of bits as the blank, which is used very frequently? Huffman codes using variable-length bit strings to represent each character.
- A few characters may be represented by five bits, and another few by six bits, and yet another few by seven bits, and so forth.

3-35

Huffman Encoding

- For example

Huffman Code	Character
00	A
01	E
100	L
110	O
111	R
1010	B
1011	D

3-36

Huffman Encoding

- DOORBELL would be encode in binary as 10111101101111010011001000.
- If we used a fixed-size bit string to represent each character (say, 8 bits), then the binary from of the original string would be 64 bits. The Huffman encoding for that string is 25 bits long, giving a compression ratio of 25/64, or approximately 0.39.
- An important characteristic of any Huffman encoding is that no bit string used to represent a character is the prefix of any other bit string used to represent a character.

3-37

Representing Audio Information

- We perceive sound when a series of air compressions vibrate a membrane in our ear, which sends signals to our brain.
- A stereo sends an electrical signal to a speaker to produce sound. This signal is an analog representation of the sound wave. The voltage in the signal varies in direct proportion to the sound wave.

3-38

Representing Audio Information

- To digitize the signal we periodically measure the voltage of the signal and record the appropriate numeric value. The process is called *sampling*.
- In general, a sampling rate of around 40,000 times per second is enough to create a reasonable sound reproduction.

3-39

Representing Audio Information

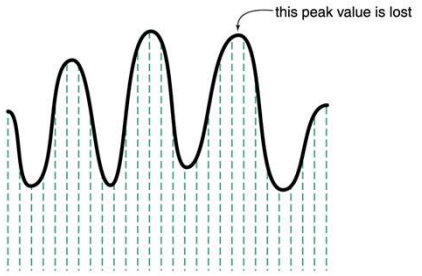


Figure 3.8 Sampling an audio signal

3-40

Representing Audio Information

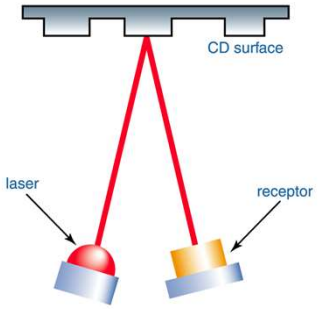


Figure 3.9
A CD player reading binary information

3-41

Audio Formats

- Audio Formats
  - WAV, AU, AIFF, VQF, and MP3.
- MP3 is dominant
  - MP3 is short for MPEG-2, audio layer 3 file ？.
  - MP3 employs both lossy and lossless compression.

  First it analyzes the frequency spread and compares it to mathematical models of human psychoacoustics (the study of the interrelation between the ear and the brain), then it discards information that can't be heard by humans. Then the bit stream is compressed using a form of Huffman encoding to achieve additional compression.

3-42

Data and Computers

- **Data compression** Reduction in the amount of space needed to store a piece of data.
- **Compression ratio** The size of the compressed data divided by the size of the original data.
- A data compression techniques can be
  - **lossless**, which means the data can be retrieved without any loss of the original information,
  - **lossy**, which means some information may be lost in the process of compaction.

3-43

作业

使用维基百科，解释以下概念。
1)ASCII
2)Color(颜色, http://zh.wikipedia.org/wiki/)

简单回答以下问题：
1) 写出字符“A”,“中”的 ASCII 码、Unicode 码、utf-8 编码。
http://wenku.baidu.com/view/f4c225340b4c2e3f572763da.html
2)黄色(yellow)的RGB编码是(, ,)
3) 从网上下载一个 BMP 格式图像,用图片编辑工具另存在 jpg、png、tiff 格式。问三种格式中,哪种格式显示质量好?相对于 BMP 格式,压缩率各是多少?
4) Winrar压缩文件是lossless, or lossy 方法？

3-44