文字的表示

- 要以數位形式表示文字文件，我們必須要能表示每一個可能出現的字元。
- 要表示的字元的數目是有限的，所以一般的方法是列出所有要表示的字元，並為每一字元指定其特定二進制數元串。
- 一個字元集（character set）簡單說是一個字元及其編碼的對應表單。藉由同意使用特定的字元集，電腦製造商已經讓文字資料的處理變得容易。

ASCII字元集

- ASCII 是 American Standard Code for Information Interchange 的縮寫，代表美國標準資訊交換碼。ASCII字元集最初用7個位元來表示每一個字元，總共可表示128個個不同的字元。之後ASCII 逐步發展成將一個位元組 (byte) 的所有8個位元都用來表示字元，所以可代表 256 個字元。
  - 這個 8位元版本的字元集正式名稱為 Latin-1 擴充ASCII字元集。這個擴充ASCII字元集包括不同口音字母(如 Ä)以及數個額外的特殊符號。

ASCII字元集 (前 128 個字元)

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

圖 3.5 ASCII 字元集

•ASCII字元集中的前面 32 個字元，由於它們都無法有簡單的字元可以表示，所以無法列印至螢幕。這些字元為了表示如換列 (LF, Line Feed)、Enter (CR, Carriage Return)、Escape (ESC)、進位、定位點等無法顯示、無法列印的字元。這些字元通常都藉由程式在處理資訊時以特殊的方式被詮釋。

Unicode字元集

- Unicode字元集的每一個字元使用了16個位元。因此，相較於擴充ASCII字元集可以表示256個字元，Unicode字元集則可以表示2^{16} (超過65,000)個字元。

Unicode字元集

圖 3.6 一些以 Unicode 字元集表示的字元

碼 (十六進制)	字元	來源
0041	A	英文 (拉丁的)
042F	Я	俄文 (西里爾字母的)
OE09	ก	泰文
13EA	Ꮺ	卻洛奇族文 (Cherokee)
211E	℞	像字母的符號
21CC	⇌	箭號
282F	⠯	布拉耶點字
345F	㑟	中文/日文/韓文 (共通的)

Unicode 被設計成 ASCII 字元集的超級集合 (superset)。也就是說，Unicode字元集中的前面256個字元完全與擴充過之ASCII字元集相符合 。

http://www.blind.org.tw/braile.htm

文字壓縮

- 尋找有效率儲存文字以及在一部電腦與另一部電腦之間有效率傳輸文字的方式是很重要的，尤其是針對圖片內容(後面會教到)，這可藉由文字壓縮來達成。以下為 3 種常見的文字壓縮方法：
  - 關鍵字編碼 (keyword encoding)
  - 遊程長度編碼 (run-length encoding)
  - 霍夫曼編碼 (Huffman encoding)

文字壓縮

- 關鍵字編碼 (keyword encoding)
  - 它將使用頻繁的單字以一個未出現於文章的單一字元符號取代。

單字	符號
as	^
the	~
and	+
that	$
must	&
well	%
these	#

範例

- 原文
  - 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.
- 編碼後的段落如下： (壓縮率：314/349 ≅ 0.9)
  - The human body is composed of many independent systems, such ^ ~ circulatory system, ~ respiratory system, + ~ reproductive system. Not only & all systems work independently, they & interact + cooperate ^ %. Overall health is a function of ~ %- being of separate systems, ^ % ^ how # separate systems work in concert.

文字壓縮

- 遊程長度編碼 (run-length encoding)
  - 在某些情況下，某個單一字元可能於長序列中一再重複。這種重複類型並不是局限於英文文字，也經常發生在大資料流中，比如DNA序列、圖片的像素資料流。
  - 一種利用這些條件來編碼的技術稱作遊程長度編碼。有時候也稱為**循環編碼** (*recurrence coding*)。
  - 遊程長度編碼中，一個序列的重複字元是由一個**旗標字元** (*flag character*)，緊接著這個重複字元，再緊接著一個指示重複多少次的單一數元來取代。

遊程長度編碼(續)

- AAAAAAA 會被編碼成： *A7
- *n5*x9ccc*h6 some other text *k8eee

  會被解碼成下列的原始文字：

  nnnnnxxxxxxxxxccchhhhhh some other text kkkkkkkkeee
- 原始字串含有51個字元，編碼後的字串則含有35個字元，因此這個範例的壓縮率是35/51大約是 0.68。
- 因為我們使用 一個 數元 來計算重覆次數，當重覆長度超過 9 個時，會有 2 種解讀的方式(如 *n30 是代表 nnn0 還是 30 個 n)，造成困擾。
  - 其實目前計數字元是以ASCII 數元來詮釋，這可以被取代成計數字元以二進制數字來詮釋，這樣就可編碼重覆 255 次的字元序列。

霍夫曼編碼 (Huffman encoding)

- 是以它的發明者霍夫曼博士來命名。
  - 為什麼很少使用於文字的字元“X”要與使用很頻繁的空白字元 “” 佔用相同的位元數目？
  - 一些字元可能以五個位元表示，另外一些字元則以六個位元表示，還有一些以七個位元來表示，等等。
  - 霍夫曼使用可變長度位元的字串來表示每一個字元。
  - 假使我們使用較少位元來表示經常出現的字元，且保留較長位元的編碼給不常出現的字元，則所代表的整個文件就會變小。

霍夫曼編碼 (Huffman encoding)

- 假設我們使用下列的霍夫曼編碼來表示一些字元：

霍夫曼碼	字元
00	A
01	E
100	L
110	O
111	R
1010	B
1011	D

則單字DOORBELL以二進位制編碼為：
1011 110 110111101001100100

- 如果我們使用固定位元數目（8位元）的字串來表示每一個字元，則原始字串的二進制形式將是64位元 (8個字元乘以 8位元/字元)。這個字串的霍夫曼編碼長度是25個位元，得到壓縮比為25/64，大約0.39。
- 作為一個霍夫曼編碼的重要特徵是，沒有用來表示一個字元的位元串是任何其他用來表示一個字元的位元串的前綴字首(prefix)。

音訊資料的表示

- 一般而言，取樣速率每秒大約40,000次便足夠建立可用的再生聲音。如果取樣速率遠低於該值，則人類耳朵會開始聽到失真的聲音。

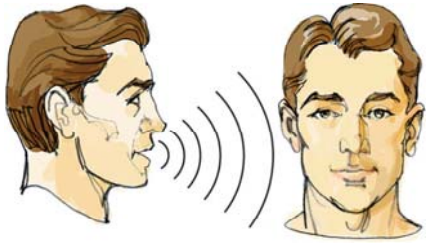

圖 3.7 聲波震動我們的耳膜

音訊資料的表示

•要將聲音訊號數位化，我們必須週期地測量該訊號的電壓，並紀錄適當的數字數值。此過程稱為*取樣（sampling）*。
•一般而言，取樣速率每秒大約40,000 次便可收集夠多資訊，可效果尚佳的再生原來聲音。

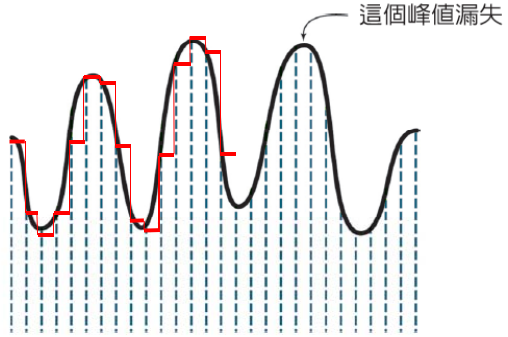

圖 3.8 音訊訊號的取樣

音訊資料的表示

▪光碟片（Compact disk, CD）以數位方式儲存聲音資訊。 在CD的表面具有微小的凹坑，用來表示二進制的位元。以低強度的雷射光瞄準碟片，如果表面很平坦，雷射光會強烈反射；但如果表面有凹坑，則僅有很小反射。

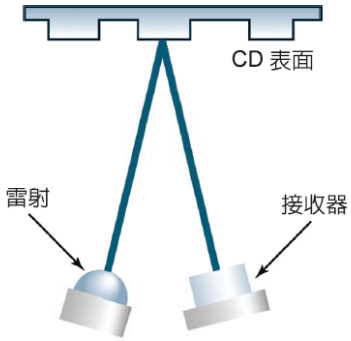

圖 3.9 CD 播放器讀取二進制資料

音訊格式

- 幾種受歡迎的音訊資訊格式，包括WAV、AU、AIFF、VQF以及MP3等。
  - MP3是MPEG-2音訊第三級檔案 (MPEG-2, audio layer 3 file)的縮寫。
  - MP3 同時使用漏失型與非漏失型壓縮。首先它分析音訊傳播頻率寬度，再與人類心理聲音學（研究耳朵與大腦之間的關係）的數學模型作比較，然後丟棄人類無法聽到的資訊。之後用霍夫曼編碼的形式壓縮得到的位元流，以達到更多的壓縮效果。
- 這些都是將類比訊號取樣並加以儲存電壓數值的音訊壓縮方式。
- 現在最佔優勢的音訊資料壓縮格式是MP3，它受歡迎的主要原因是由於比其他可用的格式具有更強大的壓縮比。

影像與圖形的表示：顏色的表示

- 顏色是我們對於各種頻率的光線到達我們眼睛視網膜的感知結果。
  - 我們的視網膜有三種顏色光感知器圓錐形細胞，它們對頻率的不同集合有反應。這些光感知器類別相當於顏色的紅色、綠色以及藍色，人類眼睛對於其他顏色的感知結果則可藉由這三種顏色的各種數量組合來獲得。

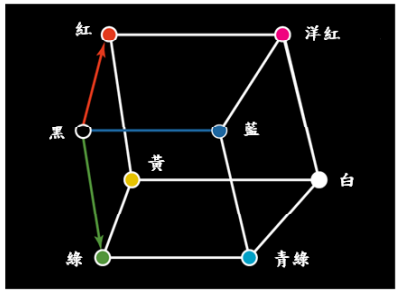

圖 3.10 三維顏色空間

顏色的表示

- 用來表示顏色的資料數量稱為**顏色深度** (*color depth*)。它通常以位元數目來表示它的顏色深度。
- **高彩** (*HiColor*) 表示16位元的顏色深度
  - 3 個 5位元 用來表示 RGB 的貢獻度，1 個位元用來表示透明度 (transparency)
  - 常用來做背景選項
- **全彩** (*TrueColor*) 是表示24位元的顏色深度。
  - 全彩能提供 1670 萬種 (2^{24}) 不同顏色，提供了比人類眼睛可以區分還要更多種的顏色。

影像與圖形的表示(續)

- 電腦中的顏色是以 RGB（紅-綠-藍）值來表示。這三個介於 0 – 255 的數值用來指示這三個主要顏色的相對貢獻度。
- 例如，RGB數值(255,255,0)是表示紅色與綠色有最大貢獻度，藍色則有最小貢獻度，這種組合結果造成該顏色為黃色。

- 全彩RGB數值以及它們所表示的顏色範例：

RGB 數值			真實顏色
紅色	綠色	藍色	
0	0	0	黑色
255	255	255	白色
255	255	0	黃色
255	130	255	粉紅色
146	81	0	咖啡色
157	95	82	紫紅色
140	0	0	褐紫紅色

- 老式監視器及某些應用軟體則僅能顯示 256 色，以調色盤展現：

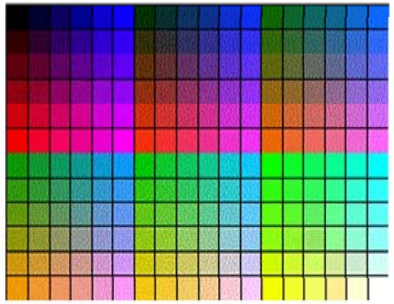

圖 3.11 限定顏色的調色盤

影像及圖形的數位化

- 像素 (pixel)
  - 圖像的數位化是一種以聚集個別點來表示圖像的動作，此種個別點稱為像素。
- 解析度 (resolution)
  - 用來表示圖像的像素數目。
- 光柵圖形格式 (raster-graphics format)
  - 以一個像素接著一個像素的準則來儲存影像資訊。
  - 幾個受歡迎的光柵圖格式包括位元映射（ bitmap，BMP）、GIF (Graphics Interchange Format) 以及JPEG等。

影像及圖形的數位化

圖 3.12 由許多個別像素所組成的數位化圖像 (*Courtesy of Amy Rose*)

影像及圖形的數位化

- **位元映射檔** (*bitmap file*) 是最直截了當的圖形表示法之一。
  - 由左至右，由上而下，一個一個像素的 RGB 值串連
  - 可用遊程長度編碼來壓縮。
- GIF格式 (Graphics Interchange Format) 是由CompuServe於1987年所發展出來的，它限制可用的顏色數目為256。
  - 可儲存一系列 GIF 影像成為動畫

影像及圖形的數位化

- JPEG格式是被設計來開發我們眼睛的本性。
  - 由於人類對於某段距離內亮度以及顏色逐漸變化時的察覺能力比當它們快速變化時來得強大，因此，JPEG格式所儲存的資料是短程距離的平均顏色色度。
  - JPEG格式被認為用在相片顏色影像時有較為優異的表現。
- PNG (發音為 “ping”) 格式代表可攜式網路圖形 (Portable Network Graphics)。
  - PNG影像通常可完成比GIF更大的壓縮率，同時提供更寬範圍的顏色深度。

圖形的向量表示法

- 向量圖形 (vector graphics)
  - 是另外一種表示影像的技術。它取代了光柵圖形中將顏色指定成像素的方式，改以線條以及幾何形狀來描述影像。
  - 向量圖形是一系列描述線條方向、線條厚度、以及線條顏色的指令，這類型的檔案大小通常較小，因為並不是每一個像素都必須被表示到。
  - 向量圖形可以以數學計算的方式重定大小，而且這些變化可以隨時依據需要加以計算。 適合用於製圖、卡通等設計圖形。
  - 然而，向量圖形不適合用來表示真實世界影像。
  - 現今網路上最受歡迎的向量圖形格式為 Flash。Flash影像以二進制格式儲存，且需要用一種特殊的編輯器來建立。
  - 發展中的新向量圖形格式：可變尺度向量圖形 (Scalable Vector Graphics, SVG)

視訊的表示：視訊壓縮器/解壓縮器

- Codec 是壓縮器/解壓縮器（COmpressor / DECompressor）的縮寫。

- 視訊壓縮器/解壓縮器 (video codec)
  - 指用來壓縮影片以讓它可以在電腦上或透過網路來播放的方法或工具。常見的有 MPEG、Real Video 等
  - 大部分的壓縮器 / 解壓縮器是區塊導向，意即視訊的每一個畫面都被分成矩形的區塊。
  - 幾乎所有的視訊壓縮器/解壓縮器都使用漏失型壓縮來最小化視訊的龐大資料量。

兩種視訊壓縮類型：時間性與空間性

- 時間性壓縮 (temporal compression)
  - 如果大部分兩張畫面的影像間沒有變化，那麼為什麼我們要浪費空間來複製所有相似的資訊呢？
  - 尋找連續畫面間的差異。
  - 時間性壓縮在視訊畫面間只有稍微變化時，是很有效的。
- 空間性壓縮 (spatial compression)
  - 移除畫面內多餘的資訊。
  - 空間性的視訊壓縮經常將相同顏色的像素群聚成區塊(矩形區域)，不再儲存每一個像素，改為儲存顏色與該區域的座標。

MGM vs. Grokster

- 使用檔案分享網站的 P2P 系統來交換含版權音樂的 MP3 檔案是否違法？
- 美國唱片工業協會 (RIAA) 改要求網際網路服務商 (ISP) 提供使用者名單，這是否侵犯使用者隱私權？
- 2003 年 MGM 控告 Grokster (擁有 KaZaA檔案分享網站) 助長版權侵犯，你覺得 Grokster 是否應負法律責任？
  - 美國最高法院認為 Grokster 應該為透過他們的廣告等營業項目“誘導” 侵犯版權而負責
  - 台灣也有 foxy 的官司 (http://taiwanpatent-ip.blogspot.com/2009/04/foxy-60.html)