



高立圖書

高立圖書

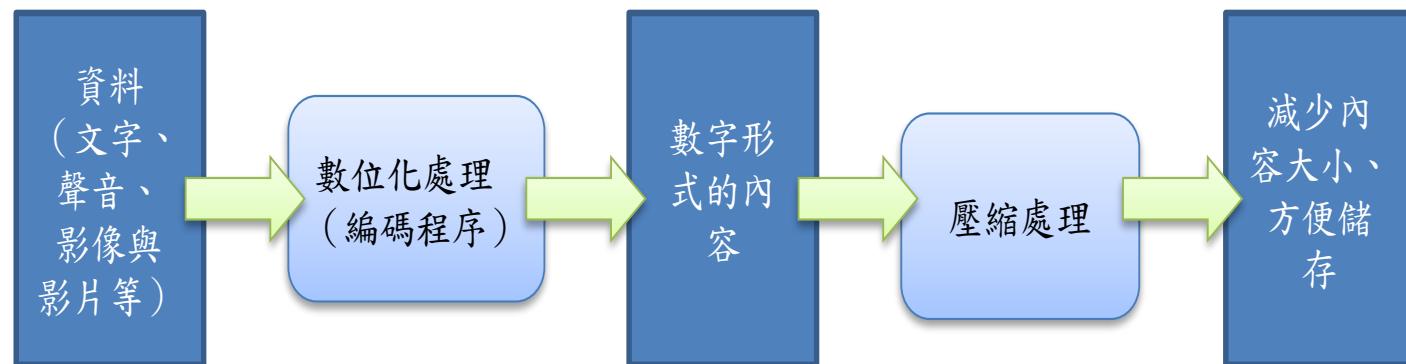
ch03

資料的表示與處理

李官陵 彭勝龍 羅壽之

學習目標

- ▶ 不同形式的資料：文字 (character)、聲音 (audio)、影像 (image) 與影片 (video) 等
- ▶ 透過編碼 (encoding) 利用不同的數字表示組成資料的基本成分
資料數位化可能產生大量的內容，導致儲存與處理不易，
資料壓縮 (compression) 能有效降低資料量



大綱

- ▶ 文字的表示與處理
- ▶ 聲音的表示與處理
- ▶ 影像的表示與處理
- ▶ 影片的表示與處理



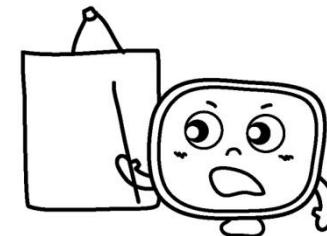
文字的表示

▶ 英文文字的編碼

- 主要由字母字元 (alphabetic character) 組成，即a~z
- 編碼方式是依據美國標準協會制訂的ASCII碼
- 基本ASCII碼利用一組7個位元的數字表示128個字元符號，95個可以顯示與列印，另外33個不可顯示
 - 例如：a = 97，b = 98
- EASCII (Extended ASCII) 使用8個位元，可表示256個字元符號，可補足其他語系所需使用的字母符號



| 十進位 | 符號 | 按鍵 | 十進位 | 符號 | 十進位 | 符號 | 十進位 | 符號 |
|-----|-----|--------|-----|-------|-----|----|-----|-----|
| 0 | NUL | Ctrl+@ | 32 | Space | 64 | @ | 96 | ` |
| 1 | SOH | Ctrl+A | 33 | ! | 65 | A | 97 | a |
| 2 | STX | Ctrl+B | 34 | " | 66 | B | 98 | b |
| 3 | ETX | Ctrl+C | 35 | # | 67 | C | 99 | c |
| 4 | EOT | Ctrl+D | 36 | \$ | 68 | D | 100 | d |
| 5 | ENQ | Ctrl+E | 37 | % | 69 | E | 101 | e |
| 6 | ACK | Ctrl+F | 38 | & | 70 | F | 102 | f |
| 7 | BEL | Ctrl+G | 39 | ' | 71 | G | 103 | g |
| 8 | BS | Ctrl+H | 40 | (| 72 | H | 104 | h |
| 9 | HT | Ctrl+I | 41 |) | 73 | I | 105 | i |
| 10 | LF | Ctrl+J | 42 | * | 74 | J | 106 | j |
| 11 | VT | Ctrl+K | 43 | + | 75 | K | 107 | k |
| 12 | FF | Ctrl+L | 44 | , | 76 | L | 108 | l |
| 13 | CR | Ctrl+M | 45 | - | 77 | M | 109 | m |
| 14 | SO | Ctrl+N | 46 | . | 78 | N | 110 | n |
| 15 | SI | Ctrl+O | 47 | / | 79 | O | 111 | o |
| 16 | DLE | Ctrl+P | 48 | 0 | 80 | P | 112 | p |
| 17 | DC1 | Ctrl+Q | 49 | 1 | 81 | Q | 113 | q |
| 18 | DC2 | Ctrl+R | 50 | 2 | 82 | R | 114 | r |
| 19 | DC3 | Ctrl+S | 51 | 3 | 83 | S | 115 | s |
| 20 | DC4 | Ctrl+T | 52 | 4 | 84 | T | 116 | t |
| 21 | NAK | Ctrl+U | 53 | 5 | 85 | U | 117 | u |
| 22 | SYN | Ctrl+V | 54 | 6 | 86 | V | 118 | v |
| 23 | ETB | Ctrl+W | 55 | 7 | 87 | W | 119 | w |
| 24 | CAN | Ctrl+X | 56 | 8 | 88 | X | 120 | x |
| 25 | EM | Ctrl+Y | 57 | 9 | 89 | Y | 121 | y |
| 26 | SUB | Ctrl+Z | 58 | : | 90 | Z | 122 | z |
| 27 | ESC | Ctrl+[| 59 | ; | 91 | [| 123 | { |
| 28 | FS | Ctrl+\ | 60 | < | 92 | \ | 124 | - |
| 29 | GS | Ctrl+] | 61 | = | 93 |] | 125 | } |
| 30 | RS | Ctrl+^ | 62 | > | 94 | ^ | 126 | ~ |
| 31 | US | Ctrl+_ | 63 | ? | 95 | _ | 127 | DEL |



文字的表示(續)

▶ 中文文字編碼

- 中文的漢字非常多，康熙字典就收錄了4萬多個漢字
- 資策會設計一套大五碼 (Big5) 繁體中文編碼標準，共收錄1萬3千多個常用的漢字，使用兩個位元組共16個位元編碼
 - 例如：電 = B971，腦 = B8A3
- Big5碼是一種內碼，僅適用於某種作業系統與應用程式內部使用
- 使用不同的中文內碼會導致資料交換不便，行政院公佈中文標準交換碼，編號為CNS 11643



文字的表示(續)

- ▶ 統整全世界文字的編碼需求，方便不同語系國家的文件瀏覽
- ▶ 國際統一碼聯盟制定萬國碼 (Universal Code, Unicode)
- ▶ 萬國碼將編碼區分17個群組，每一群組稱為一個平面 (plane)，每個平面用2個位元組編碼字元符號
- ▶ 平面0的編碼數字範圍為：U + 0000 ~ U + FFFF，後方4個十六進位的數字記錄編碼數字
- ▶ 平面0又稱基本多文種平面 (Basic Multilingual Plane, BMP)，包括各國語言基本常用的字元符號



文字的表示(續)

- ▶ 對相同的英文字母符號，萬國碼比ASCII碼使用更多的位元組
- ▶ UTF (Unicode Transformation Format) 是實現萬國碼的一種方式
- ▶ UTF-8是目前最常見的一種，數字8表示最少使用8個位元編碼
- ▶ UTF-8採用變動寬度 (variable-width) 的編碼方式
 - 依據編碼數字的大小，自動調整使用的位元組數量，至多用到4個位元組



隨堂練習

- ▶ 請回答如果需要將底下的字元符號編碼，至少需要多少位元來記錄？
 1. 數字0至9
 2. 小寫英文字母



文件的壓縮

- ▶ 在資訊處理的應用上，如果能將大量的資料內容做適當的壓縮處理，可以節省儲存的空間，傳遞資料時也會比較省時
- ▶ 壓縮有兩種基本形式：**不失真 (lossless)** 壓縮與**失真 (lossy)** 壓縮
- ▶ 壓縮效率一般用壓縮率 (compression ratio) 表示，代表壓縮後與壓縮前檔案大小的百分比
- ▶ 失真壓縮可以透過丟棄部分資料達到高的壓縮率
 - 使用者能容忍失真即可



文件的壓縮(續)

- ▶ 關鍵字編碼 (keyword encoding)
- ▶ 英文文章最常出現的關鍵字前五名依序為：the、be、to、of 與and
- ▶ 建立編碼表

| 關鍵字 | 編碼符號 |
|-----|------|
| the | # |
| be | \$ |
| to | % |
| of | ^ |
| and | & |



文件的壓縮(續)

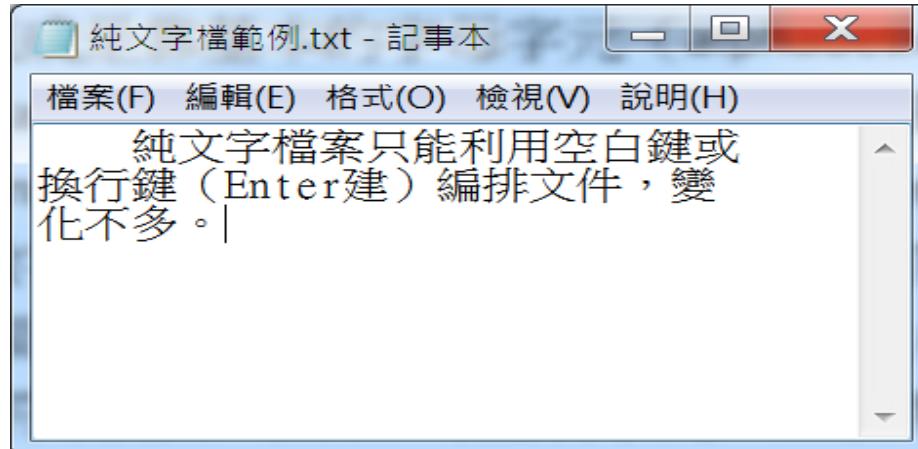
- ▶ 壓縮前：「to be, or not to be, that is the question.」
- ▶ 壓縮後 「% \$, or not % \$, that is # question.」

- ▶ 使用限制
 - 編碼符號不可與文章其他出現的字元符號相同
 - 不對長度為一個字元的關鍵字做編碼



文件的儲存

- ▶ 純文字格式：只能儲存文件的內容，無法編排成多樣格式的文件
 - 英文文件以 ASCII 碼儲存文字對應的數字碼，中文文件一般以 Big5 碼儲存
 - 常見的純文字檔案為 txt 檔案，在 Windows 系統中可以利用「記事本」程式，編輯與閱讀
 - 只能利用空白鍵或換行鍵編排文件



文件的儲存(續)

- ▶ 二進位格式：儲存文件內容外，也可以儲存編排的格式
 - 大小、字體與顏色等，讓文件可以圖文並茂
 - 由開發應用程式的業者自行定義儲存的格式
 - 整篇文件以二進位數字的內容儲存
 - 常見的格式如微軟「Word」程式使用的 doc 檔案與 Adobe 公司使用的 pdf 檔案

Word 文件可編排格式豐富的內容，包括文字的大小、字體與格式等等，也可以插入圖形。



文件的儲存(續)

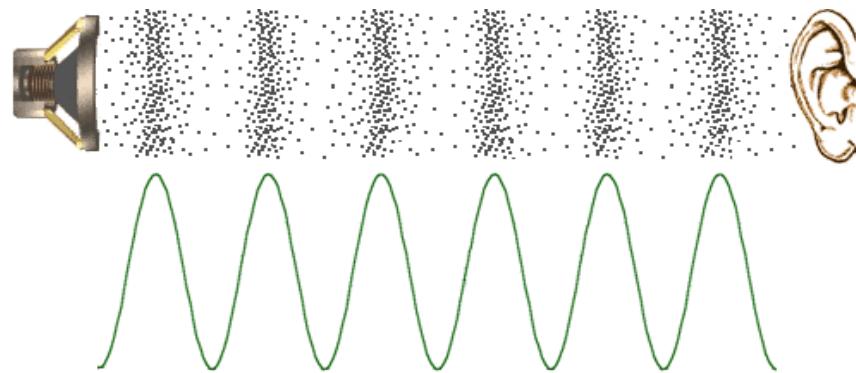
- ▶ 純文字格式化格式：利用純文字格式儲存文件內容外，也儲存文件編排的資訊
 - 常使用於全球資訊網的網頁，儲存成 html 或是 xml 檔案
 - 微軟「Word」程式新提供的 docx 檔案，也是採用 xml 格式，格式的設定靠不同的標籤來完成

```
<w:body>
  <w:p w:rsidR="009B230C" w:rsidRDefault="009B230C">
    <w:pPr>
      <w:rPr>
        <w:rFonts w:hint="eastAsia"/>
      </w:rPr>
    </w:pPr>
    <w:r>
      <w:rPr>
        <w:rFonts w:hint="eastAsia"/>
      </w:rPr>
      <w:tab/>
      <w:t>Word</w:t>
    </w:r>
    <w:r>
      <w:rPr>
        <w:rFonts w:hint="eastAsia"/>
      </w:rPr>
      <w:t>文件可編排格式豐富的內容，包括文字的</w:t>
    </w:r>
```



聲音的表示

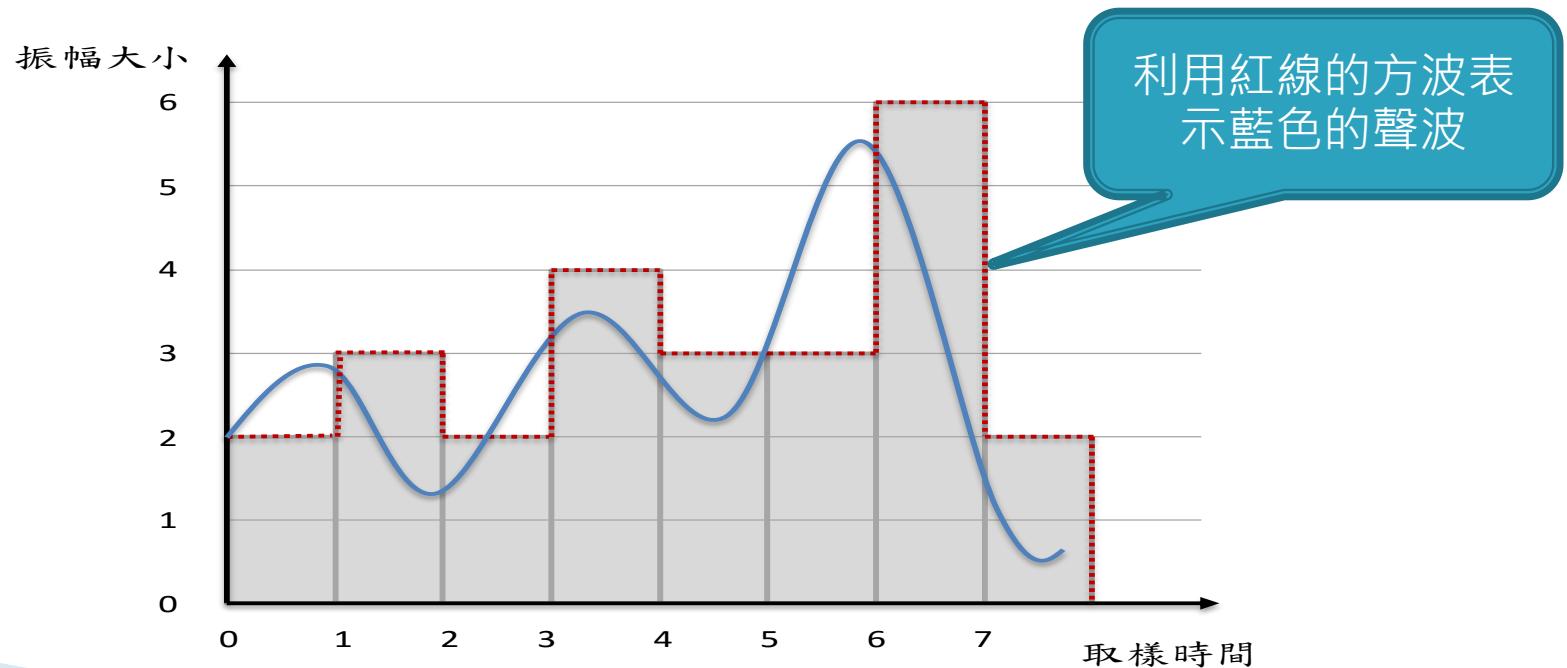
- ▶ 人可以聽見聲音是因為聲音會導致空氣的振動，進而帶動耳膜的振動
- ▶ 振動的樣式可以利用波的形狀來描述，聲波 (sound wave) 是一種隨時間連續變化的類比形式
- ▶ 利用取樣 (sampling) 的方式記錄聲波的形狀



From: <http://www.mediacollege.com/>

聲音的表示(續)

- ▶ 取樣是隔固定的时间间隔，讀取聲波資料的振幅，並以整數的數字記錄大小
- ▶ 連續變化的類比資料變成不連續變化的數位資料
 - 取樣結果：2 3 2 4 3 3 6 2



聲音的表示(續)

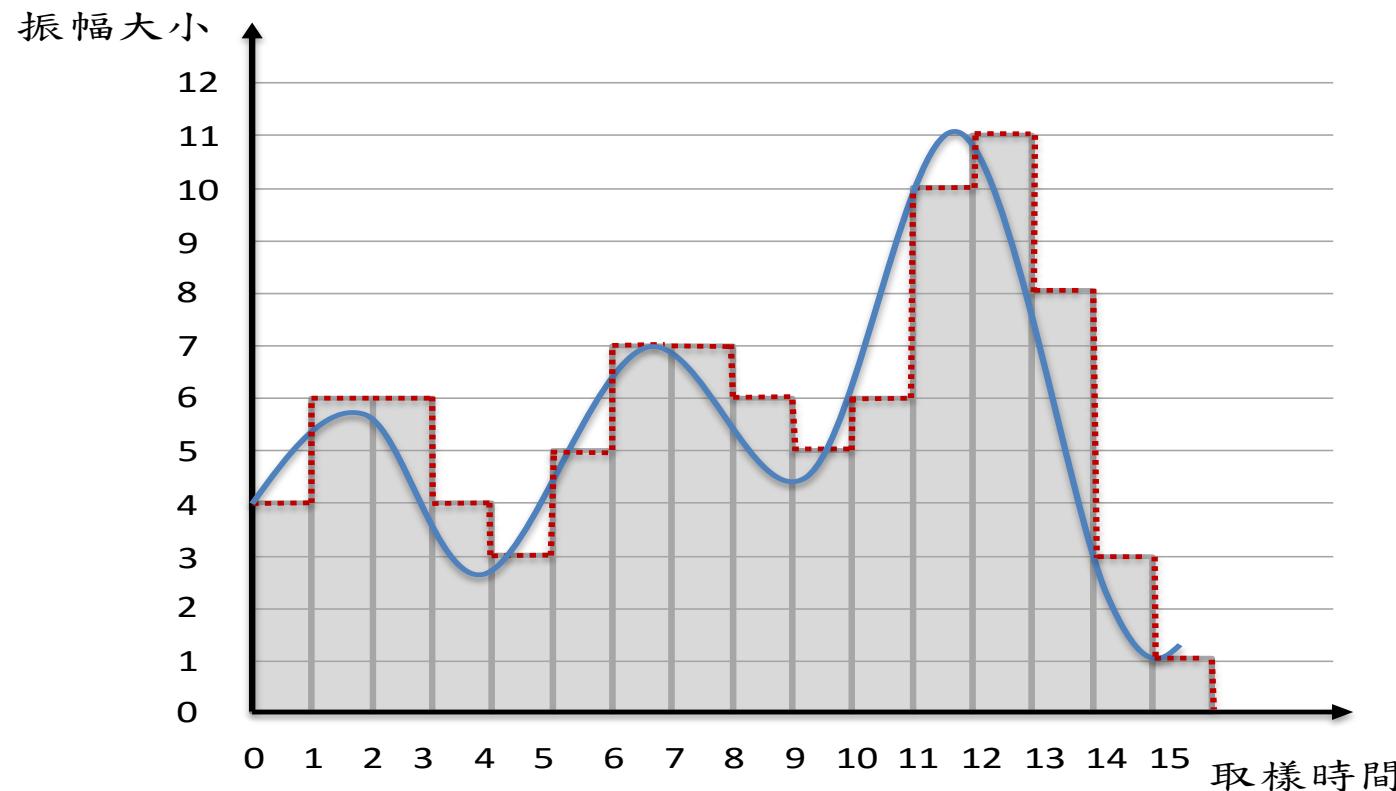
- ▶ 原先的聲波與取樣的方波不盡相同，產生失真 (distortion) 的現象
- ▶ 失真夠小的話，人的耳朵並非很靈敏，無法辨別這些微的差異
- ▶ 要如何降低取樣的失真度呢
 - 縮小取樣時間間隔，或是說提高取樣頻率
 - 讓記錄振幅大小的刻度變得更密，另一種說法是增加量化等級 (quantization level)



聲音的表示(續)

▶ 降低取樣失真度的方法

- 取樣結果：4 6 6 4 3 5 7 7 6 5 6 10 11 8 3 1



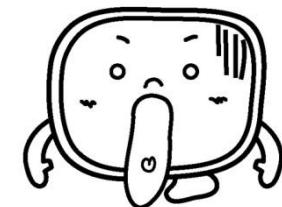
聲音的壓縮

- ▶ 將取樣的結果用二進位的數字記錄
- ▶ 最簡單的作法是用固定且最小位元數表示每個取樣的數字
- ▶ 固定長度編碼範例
 - 取樣結果：2 3 2 4 3 3 6 2
 - 編碼後：010 011 010 100 011 011 110 010



聲音的壓縮(續)

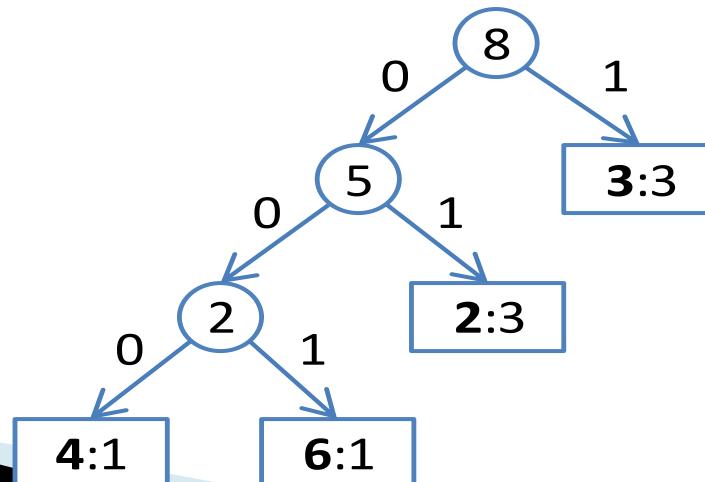
- ▶ 變動長度編碼的優勢
 - 總編碼長度變小
- ▶ 範例
 - 取樣結果：2 3 2 4 3 3 6 2
 - 編碼後：01 1 01 000 1 1 001 01



| | | | | |
|------|-----|-----|-----|-----|
| 數字 | 2 | 3 | 4 | 6 |
| 固定長度 | 010 | 011 | 100 | 110 |
| 變動長度 | 01 | 1 | 000 | 001 |

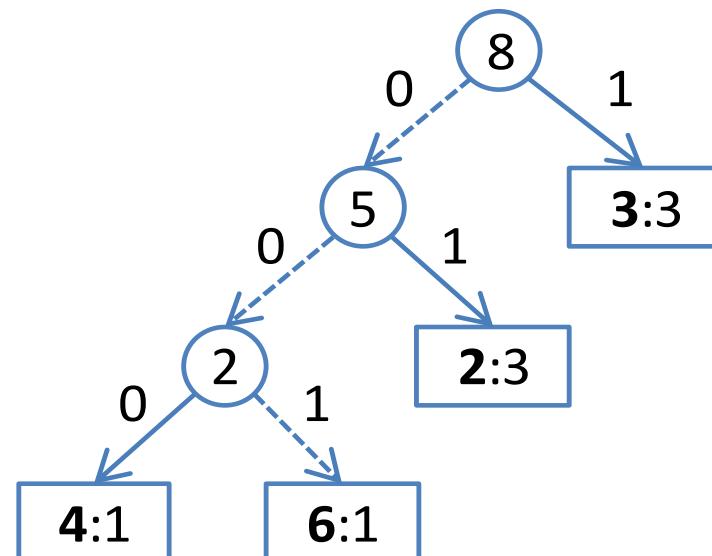
聲音的壓縮(續)

- ▶ 實現變動長度編碼的方法之一，利用霍夫曼編碼 (Huffman encoding)
- ▶ 將常出現的數字，用較少的位元表示
- ▶ 編碼過程參考霍夫曼編碼樹 (encoding tree)
 - 內部節點 VS. 外部節點
 - 外部節點的標示X:Y，表示數字X出現的次數為Y
 - 內部節點的標示X，表示連接的兩個節點的次數總和為X



霍夫曼編碼樹

- ▶ 最上方的內部節點稱為樹根 (root)
- ▶ 所有向左的分支線標示位元0，向右的分支線標示位元1
- ▶ 一個外部節點記錄的數字，編碼方式是參考由樹根走至這個節點的路徑，將路徑上的位元值列出
 - 例如：數字6的編碼為001
 - 例如：數字3的編碼為1



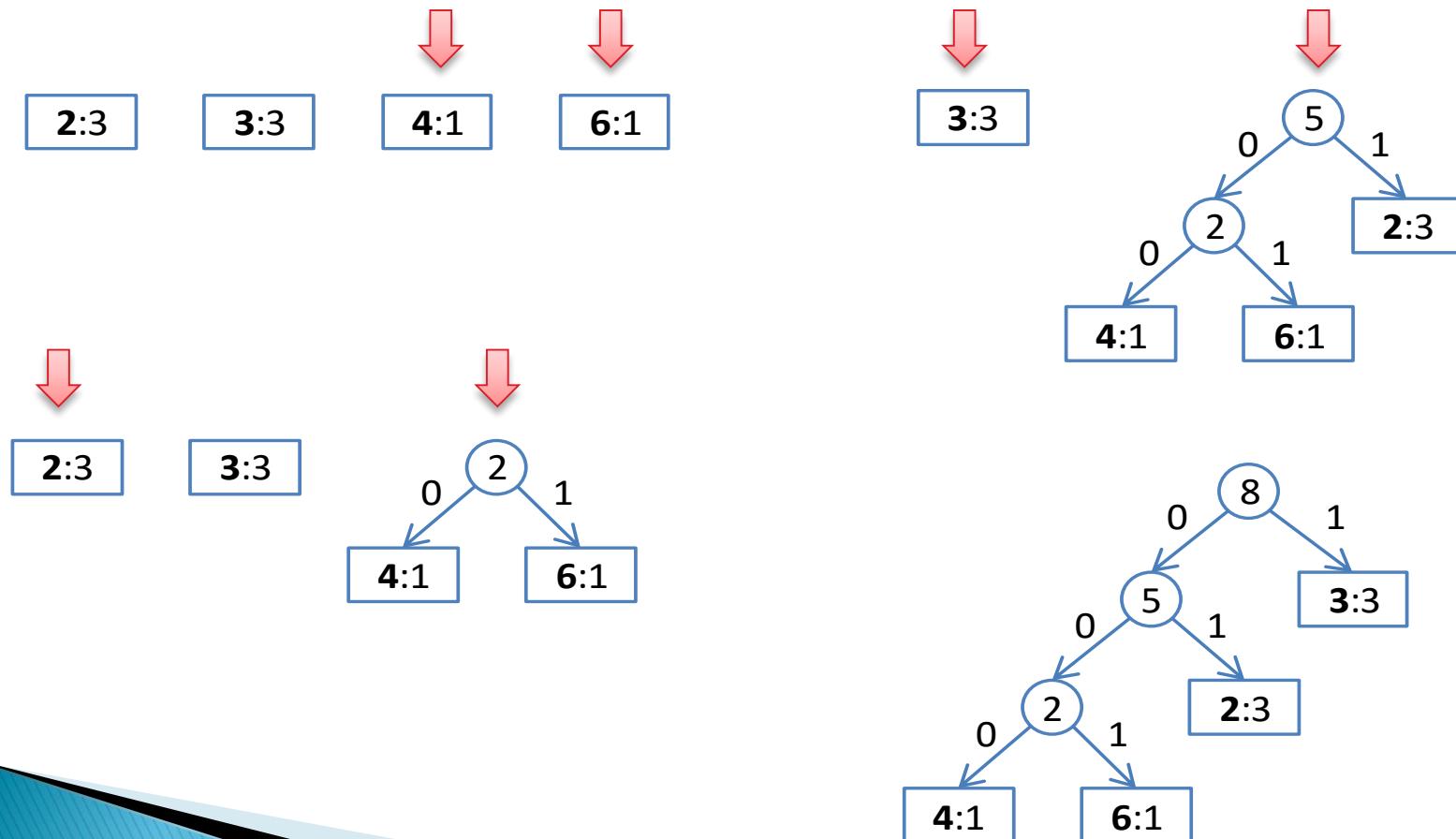
霍夫曼編碼樹 (續)

- ▶ 建構方式：
 - 步驟1：對出現的每個相異數字(或符號)，建立一個外部節點，並標示出現的次數
 - 步驟2：找出標示次數最小的兩個節點(外部節點或內部節點皆可)，建立一個內部節點連至這兩個節點(誰放左、誰放右不重要)，並標示次數總和
 - 步驟3：重複步驟2，直到所有建立的節點全部連成一棵編碼樹



霍夫曼編碼樹 (續)

► 建構範例：2 3 2 4 3 3 6 2



霍夫曼編碼樹 (續)

▶ 解碼程序

- 步驟1：移至編碼樹樹根，讀入待解碼的二進位資料
- 步驟2：若是讀入位元0則沿左邊的分支線走至下一個節點；若是讀入位元1則沿右邊的分支線走至下一個節點
- 步驟3：重複步驟2，若是走至外部節點則輸出記錄的編碼數字(或符號)，接著回到步驟1。若讀完待解碼的資料，則結束



霍夫曼編碼樹 (續)

► 解碼範例：01101000

1. 解碼資料：01101000

解碼輸出：2

2. 解碼資料：101000

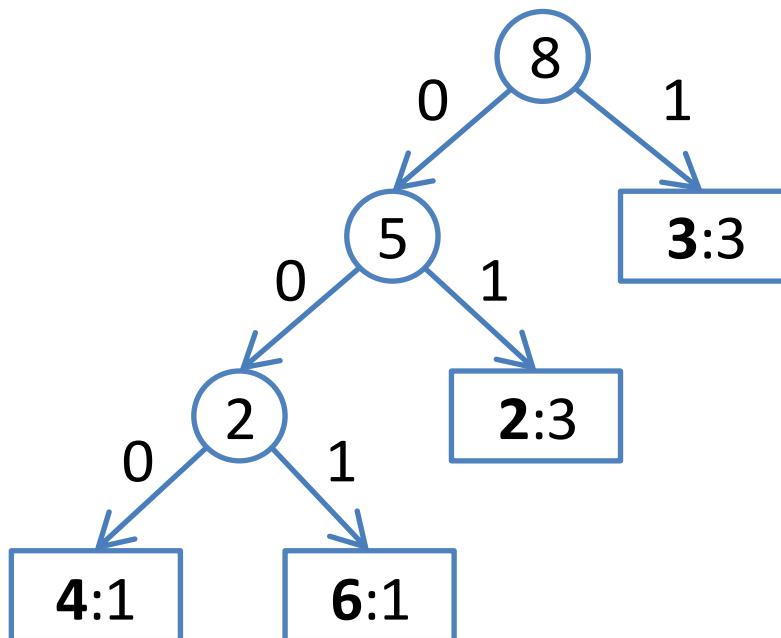
解碼輸出：3

3. 解碼資料：01000

解碼輸出：2

4. 解碼資料：000

解碼輸出：4



隨堂練習

- ▶ 請利用霍夫曼編碼，處理底下的句子：
“abcbaacda”



聲音的儲存

- ▶ 唱片CD (Compact Disk) 是最常見的儲存音樂媒介，採用的取樣頻率為 $44,100\text{ Hz}$ ，量化等級總共有 $2^{16} = 65,535$ (採用固定16位元記錄振幅大小)
- ▶ Windows電腦的聲音wav檔與CD的格式相似，均屬於未壓縮處理的聲音格式
- ▶ 壓縮的聲音檔案屬MP3最為耳熟能詳，採用失真壓縮的方式，方法是透過丟棄人耳不容易辨識的超低(低於 20 Hz)與超高頻率(高於 $20,000\text{ Hz}$)的聲音資料



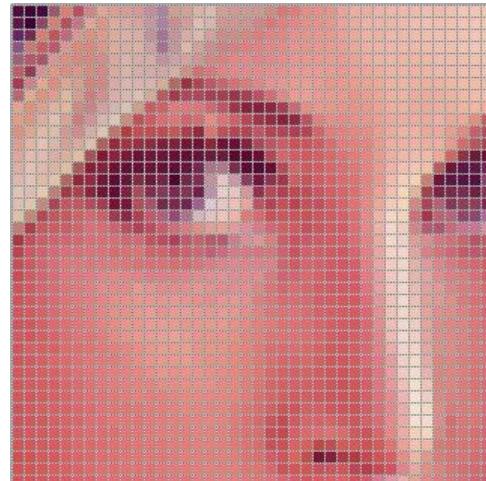
檔案大小與儲存空間

- ▶ 電腦的檔案大小用位元組為單位，以字母B表示 (Byte的意思)
 - 例如：1000 B 與 10000 B
- ▶ 位數很大時，可以配合利用K (念做kilo)、M (念做mega) 與 G (念做giga) 來表示
 - $1 \text{ KB} = 1000 \text{ B}$ 、 $1 \text{ MB} = 1000 \text{ KB}$, $1 \text{ GB} = 1000 \text{ MB}$
- ▶ KB、MB與GB的定義使用10的次方，而不是2的次方
 - 通常用來計算電腦記憶體的大小時，用2的次方 (即 $1 \text{ KB} = 1024 \text{ B}$)，但用來計算硬碟、行動碟、記憶卡與光碟容量時，用10的次方 (即 $1 \text{ KB} = 1000 \text{ B}$)



影像的表示

- ▶ 數位相機，利用感光的半導體元件CCD或是CMOS來捕捉光線
- ▶ 數位相機拍照的品質取決於感光元件佈滿的密度
- ▶ 一個感光元件記錄一個光點的訊號，進而成為照片上一個影像點或稱像素 (pixel)
 - 支援1千萬像素的數位相機，感光元件的個數規模也是1千萬等級



<http://geeksdreamgirl.com/>



高立圖書公司
GAU LIH BOOK CO., LTD.

版權所有 31

影像的表示(續)

- ▶ 將像素所表示的顏色用十進位的數字記錄，再以二進位的形式儲存
- ▶ 如果是黑白圖，則一個像素用數字0與1表示明暗，用一個位元的儲存空間即可
- ▶ 彩色圖就必須先決定希望呈現的顏色個數
 - 記錄16種顏色，每個像素以4位元儲存
 - 記錄256種顏色，每個像素以8位元儲存
- ▶ 用24個位元記錄顏色，可以表示1千6百多萬種顏色，已經接近肉眼可以辨識的所有顏色種類，因此也稱為全彩 (true color)



影像的表示(續)

- ▶ 各種顏色可以透過基本的三種顏色：紅 (Red)、綠 (Green)、藍 (Blue) 混合而成
- ▶ 每個像素所記錄的顏色可以區分三個部分：R、G與B，這種表示系統稱為RGB色彩
- ▶ 使用全彩時，R、G與B各佔用8個位元



黑白圖



16 色圖



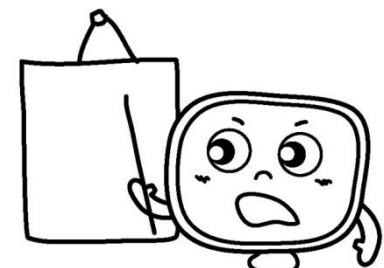
全彩圖



影像的表示(續)

▶ RGB色彩系統

| 紅色值 | 綠色值 | 藍色值 | 顏色 |
|-----|-----|-----|-----|
| 0 | 0 | 0 | 黑色 |
| 255 | 255 | 255 | 白色 |
| 255 | 255 | 0 | 黃色 |
| 157 | 95 | 82 | 紫色 |
| 255 | 130 | 255 | 粉紅色 |



隨堂練習

- ▶ 請計算一張500萬像素的照片，如果使用底下的顏色時，需要儲存的檔案大小為多少位元組
1. 256色
 2. 全彩



影像的儲存



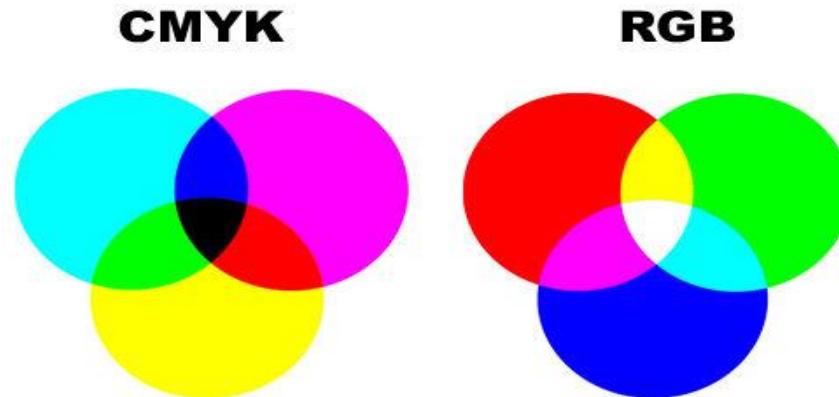
<http://6eegutierrez.wordpress.com/>

- ▶ BMP檔
 - 未壓縮導致檔案過大，不適合在網路上傳輸
- ▶ PNG檔
 - 採用不失真壓縮，在網路上是常見的影像格式
 - 支援背景透明設定
- ▶ JPG或JPEG檔
 - 採用失真壓縮，儲存時設定不同的影像品質等級，達到不同的壓縮率
- ▶ GIF檔
 - 適合影像內容是文字、幾何圖或簡單的圖示
 - 只支援256種顏色，但是支援背景透明設定
 - 支援動畫格式



影像的儲存(續)

- ▶ TIFF檔
 - 記錄CMYK的顏色資訊，方便照片的列印
- ▶ CMYK色彩系統
 - 彩色雷射印表機都使用四種顏色的碳粉匣：青 (Cyan)，洋紅 (Magenta)，黃 (Yellow) 與黑色 (black)
 - 這種調色方式最多只能呈現1百多萬種顏色



<http://www.ginifab.com/>

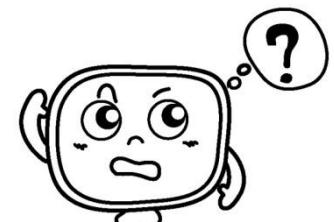


高立圖書公司
GAU LIH BOOK CO., LTD.

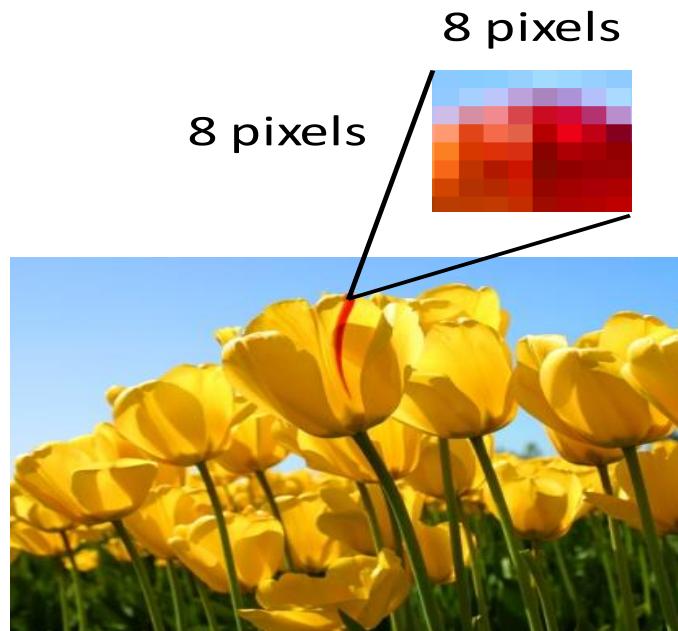
版權所有 37

影像的壓縮

- ▶ 失真壓縮：JPEG檔採用離散餘弦轉換 (Discrete Cosine Transform，DCT) 技巧
 - 影像切割成許多 8×8 的格子，每個格子代表一個像素
 - 將像素的色彩資訊轉成訊號波形的樣式
 - 基於一個波形可視為無數個不同頻率的波 (這裡的波是指數學的餘弦函數波) 所組合而成
 - 高頻的波主要來自物體邊緣的影像，如同聲音的取樣處理，使用少量化等級，減少資訊記錄量
 - 低頻波採用多量化等級，減少失真
 - 量化後的數值，透過霍夫曼編碼成二進位數字



影像的壓縮(續)



邊緣的毛髮不易察覺失真

色塊較大的區域（如樹幹、手臂）先出現失真

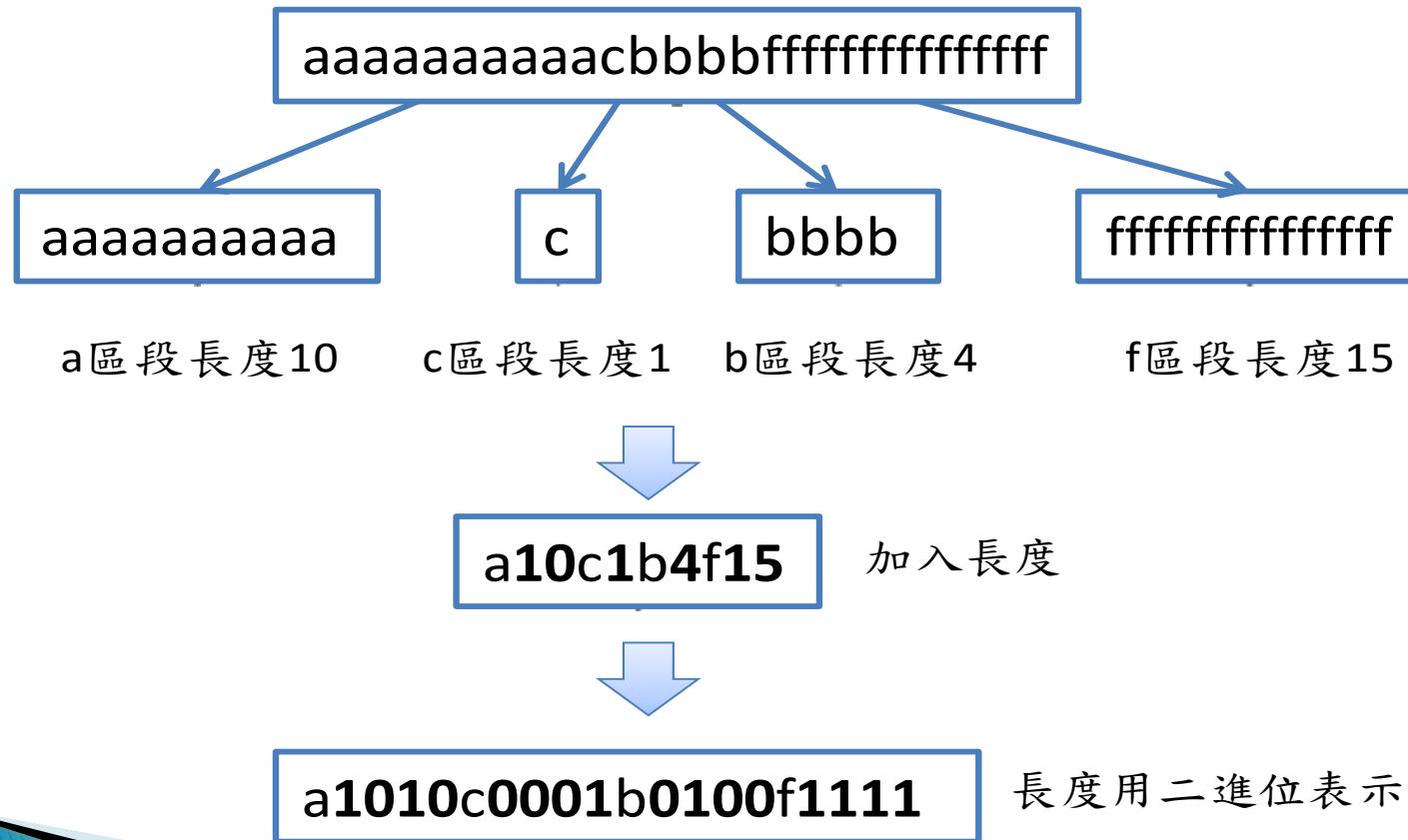
影像的壓縮(續)

- ▶ 不失真壓縮：區段長度編碼 (Run-Length Encoding , RLE)
 - 相同符號連續出現的部分稱為區段 (run)，連續出現的個數稱為該區段的長度
 - 連續出現的符號只記錄一次，後面跟著表示區段長度的數字
 - 範例：aaaabb記錄成a**4**b**2**
 - 範例：111133記錄成**1****4****3****2**
 - 適合重複性高的資料
 - 只出現一次的符號，仍需要記錄長度，反而變成浪費



RLE壓縮範例

- ▶ 對連續的字母做壓縮



RLE壓縮範例(續)

- ▶ 對連續的數字做壓縮

22222223334488888



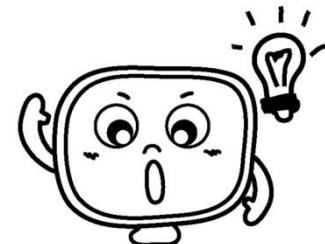
27334285

長度用粗體表示



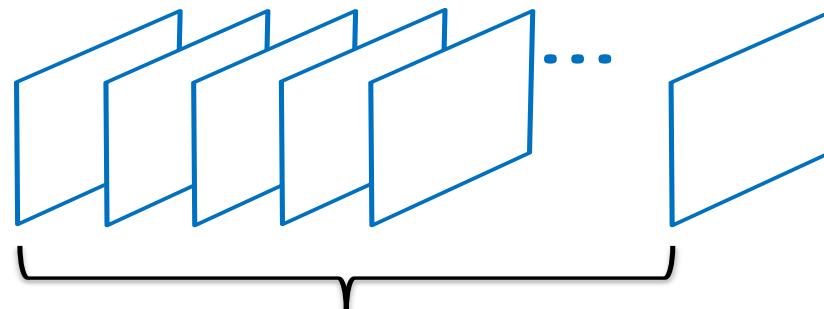
00100111001100110100001010000101

出現的數字與長度均用4個位元表示



影片的表示

- ▶ 影片是連續的影像，攝影機在固定的時間間格，拍下影像
- ▶ 播放連續影像時，由於眼睛有視覺暫留的現象，會短時間記憶前一個影像，比較下一個影像的差異後，產生影像在動的錯覺
- ▶ 一張張影像又稱為畫面 (frame)
 - 讓眼睛察覺不出影片間斷的移動，一秒播放的畫面必須達到24個畫面以上



固定間格的畫面形成影片



影片的表示(續)

- ▶ 影片畫面的解析度與畫面長與寬的像素個數有關
 - VCD影片的解析度是 352×240 像素
 - DVD影片的解析度是 720×480 像素
 - BD影片的解析度是 1920×1080 像素，俗稱高畫質 (High-Definition, HD)
 - 4K超高畫質 (Ultra High-Definition, UHD)，解析度是 3840×2160 像素 (4倍HD畫質)



影片的表示(續)

- ▶ 螢幕顯示畫面時，會將畫面水平切割成一條條的掃描線 (scan line)，個數與畫面縱向的像素相同
- ▶ 掃描線呈現的方式有兩種：交錯式 (interlaced) 與循序式 (progressive)
 - 每秒播放60個畫面為例，交錯式以60 i標示，循序式以60 p標示
- ▶ 循序表示將一個畫面全部的掃描線依序顯示
- ▶ 交錯則是先顯示第一個畫面的奇數掃描線，接著顯示第二個畫面的偶數掃描線



影片的儲存

- ▶ 高畫質的影片大小動輒好幾G，需使用高效率的失真壓縮處理
- ▶ 壓縮與解壓縮需要編解器 (Codec) 的程式：結合編碼 (coding) 與解碼 (decoding)
- ▶ MPEG是致力於視訊壓縮的專業組織，前後發表MPEG-1、MPEG-2與MPEG-4
 - MPEG-1用於VCD影片的儲存 (dat檔)，常見的MP3音樂檔格式也是源自MPEG-1
 - MPEG-2用於DVD影片的儲存 (vob檔)
 - MPEG-4適用HD高畫質影片，avi、dat、vob、mpeg、mpg、mp4、divx與xvid等檔案格式



影片的儲存(續)

- ▶ 網路影片如果每次在觀賞前需要將整份檔案下載，會浪費不少時間
 - 每個人擁有下載檔案，缺乏智慧財產的保護
- ▶ 影音串流 (video streaming) 強調邊傳邊播，播過就丟棄資料
 - Real Networks公司推出的rm或ram檔
 - 微軟的wmv檔，蘋果公司的mov檔



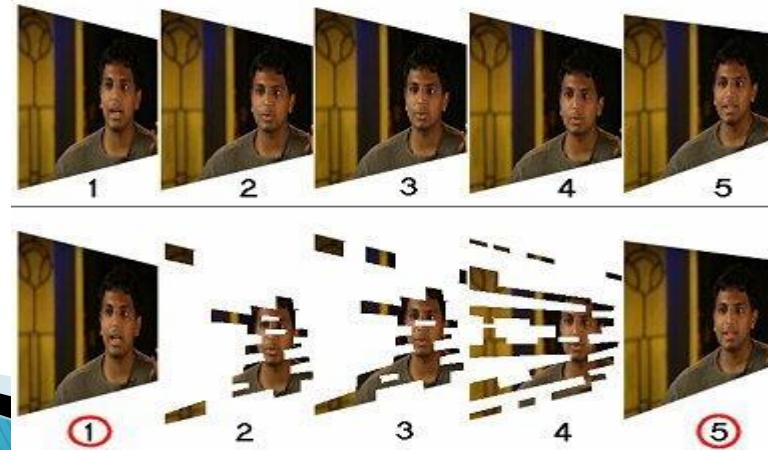
影片的壓縮

- ▶ MPEG影片在壓縮時同時考慮時間與空間上的壓縮
- ▶ 空間壓縮指單張畫面的靜止影像壓縮，利用自身畫面相鄰區域的差異性
- ▶ 時間壓縮指連續畫面的壓縮，利用前後畫面間的差異性
- ▶ 差異性壓縮原則
 - 數列”252 253 252 255”
 - 以第一個數字252為基準，將數列其他的數字均減去基準數，得到”252 1 0 3”
 - 差異數值用2個位元即可記錄，比起原先的數字均用8個位元還省空間



影片的壓縮(續)

- ▶ MPEG的時間壓縮
 - 連續的影片畫面可能來自相同的場景，背景的物體相同，差異在主要人物的姿態
 - 利用畫面影像間的差異性，記錄基準畫面與差異的部分即可
- ▶ MPEG壓縮將整個影片切割成一群一群的畫面 (Group of Pictures, GOP)
 - 每一群包含三種形式的畫面：I、P與 B畫面



<http://nickyguides.digital-digest.com/>



高立圖書公司
GAU LIH BOOK CO., LTD.

版權所有 49

影片的壓縮(續)

- ▶ I畫面：內部編碼畫面 (intracoded frame)，壓縮時只透過自身畫面的空間差異性，可做為其他畫面的基準畫面
- ▶ P畫面：預測畫面 (predicted frame)，會參考前一個I畫面或P畫面，並記錄差異的部分
- ▶ B畫面：雙向畫面 (bidirectional frame)，會參考前後的I畫面與P畫面，並記錄差異的部分

