TEXT DATA

•Text is a vital element of multimedia presentations.
•Words and symbols in any form, spoken or written, are the most common system of communication. They deliver the most widely understood meaning to the greatest number of people— *accurately and in detail.*
•It is very important to choose the suitable words and symbols in your multimedia presentation. You will reward yourself and your users if you take the time to choose the right words.
•However, what we concern in this unit is another aspect of text, namely its appearance in multimedia presentation.
•Text is a visual representation of language, as well as a graphic element in its own right. The study of how to display text is known as *typography. It concerns the precise shape of characters, their* spacing, the layout of the lines and paragraphs, and so on.

1 Character Sets

•As we may already recognised that, the visual appearance of a piece of text can be in many different forms, the basic meaning of the text will not change. Fundamentally, a piece of text consists of letters,digits, punctuations and other symbols. These can be considered as *abstract characters*.

•Abstract characters in a particular language are grouped into alphabets. For example, The alphabet of English contains the upper case letters A to Z, the lower case letters a to z, the digits and a number of punctuations.

•To represent text digitally, it is necessary to define a *mapping* between (abstract) characters and the values that are stored in a computer system. We call this mapping a *character set*. The domain of this mapping, i.e., the abstract characters are called *character repertoire* and the values to be stored are called the *code values* or *code points*. For example, the commonest character set used in Hong Kong for Chinese is big5. Its character repertoire contains more than 13,000 characters and other symbols.

1.1 Standardisation

•Clearly, if any systems want to communicated with each other, they have to have a common language. Text is the most widely used means of communication among computer systems. Therefore, a common character set is essential.

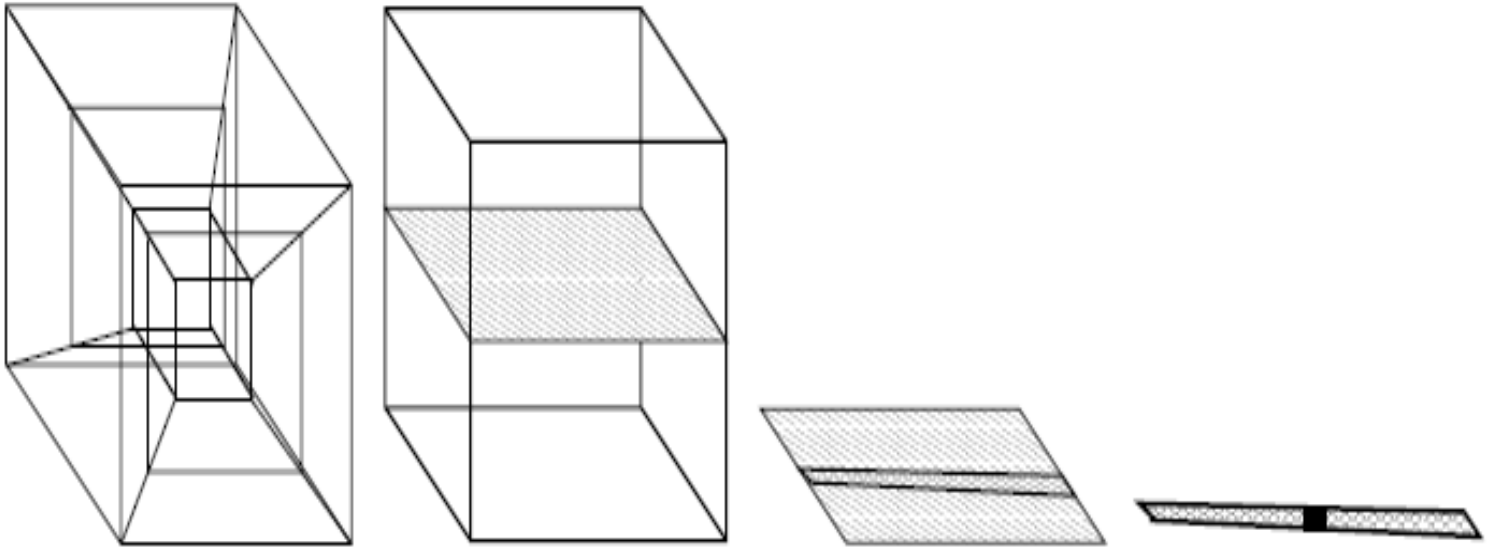
•The earliest widely accepted character set is *ASCII* which stands for American Standard Code for Information Interchange. The code range of ASCII is 7-bit, meaning that the code value can be stored in 7 bits. Therefore, at most 127 characters can be coded. However, the character repertoire of ASCII only comprises 95 printable characters. The values 0 to 31 and 127 are assigned to *control characters*. Later, ISO adopted ASCII as an standard (ISO 646).

•Obviously, 127 values are not enough to code many of the world's languages.
•ISO produced a new standard ISO 8859 with 8-bit characters. Actually, ISO 8859 has many parts. Each part specifies a number of character sets.
•The lower 127 characters in all parts are identical to ASCII.

The major shortcoming of ISO 8859 is that it uses only 8-bit values, therefore, the number of code points is limited. To overcoming this, two organisations have been working on multi-octet character set standards since the beginning of 1990s:

ISO produced a standard: ISO/IEC 10646-1:1993 *Universal Multi-Octet Coded Character Set*. It uses four bytes to encode a character, therefore, it can have at most 2^{32} code points.

The organisation of the ISO/IEC 10646 is hierarchical. The four bytes (g, p, r, c) are named group, plane, row and column.



The entire hypercube | A plane p in g | A row r in p | A character c in r

The first plane of ISO10646, $(0, 0, *, *)$, has been made compatible to Unicode. This plane is known as *Basic Multilingual Plane* (BMP).

•**Unicode** produced a standard *The Unicode*
•*Standard, Version1.0* in 1991. The latest version is version 3.0 passed at the beginning of 2000. Unicode uses 2 bytes to encode each character.
•Unicode attempts to specify a character set to embrace all languages of the world. It is gaining popular support nowadays.
•Unicode uses a process known as *CJK unification* to handle h `anz`i Chinese characters. This is needed because the number of code points is limited and the Chinese characters used in different countries and regions have many variations.
•The latest Unicode standard has more than 27484 chinese characters.

There have been many Chinese character set standards before the Unicode. The following character set standards are in common use:

- **GB2312-80** contains 6763 Chinese (simplified) characters plus other symbols.
- **big5** contains 13053 Chinese (traditional) characters plus other symbols.
- **CNS11643-1992** contains 48027 Chinese characters divided into a number of planes.
- **HKSCS** Hong Kong Supplementary Character Set (previously HK GCCS) adds 3049 Chinese characters into Big5.

1.2 Encoding

An *encoding* is another level of mapping. It transforms a code value into a sequence of bytes for storage and transmission.

Abstract	Encoded	Serialised		
		UTF-16BE	UTF-8	
Å	C5	00 C5	C3 85	Latin-1 Supplement
∬	212B	21 2B	E2 84 AB	Symbol
	F0000	DB 80 DC 00	F3 B0 80 80	Private area
A + ◌̊	61 30A	00 61 03 0A	61 CC 8A	Combining Diacritical marks

2 Typefaces and fonts

•To display text, we need to have a visual representation of the characters stored as codes in the computer. In fact, each character may be represented by many different *glyphs*.

•A *typeface* is a family of graphic characters with a coherent design and usually includes many sizes and styles.

•A *font* is a set of graphic characters with a specific design in a specific size and style.

For example, the typeface used in this paragraph is ‘Times’. The font is ‘Times Roman regular 16 point’.

•The picture on the right illustrates the lead types that are used before the computer age. In fact, it was the Chinese who invented the character block printing.

2.1 Measurements of the type

•When putting characters on to a page, we need to know some basic measurement of the types we use.
•Each character has a *bounding box*. This is the rectangle enclosing the entire character.
•Each character has an origin. It is usually place on the *baseline*. The width of the character determine where the origin of the next character will be.

- The distance between the origin and the left side of the bounding box is called *left side bearing*.
- As we all know, some of the lower case letters extend upward, like b and h, while others extend downward, like g, p and q.
- The height of the lower case letter without ascender and descender is called the *xheight*.
- The height of the upper case letters is called the *cap-height*.

There are many fonts available. *Five* attributes are often used for specifying a font:

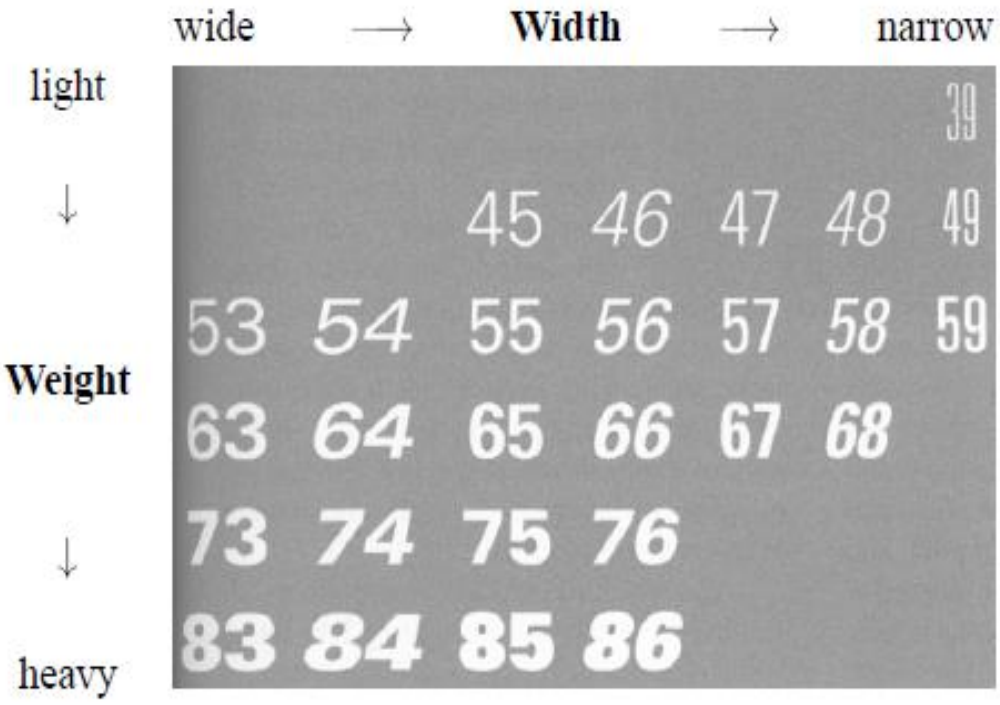
•***Family*** — fonts in the same family have a coherent design, a similar look and feel. Here are some of the common families: Times, Helvetica, Courier, Garamond, Univers

•***Shape*** — refers to the different appearance within a family. Compare the following shapes: normal (upright), sloped (oblique), *italic*, SMALL CAP

•***Weight*** — measures the darkness of the characters, or the thickness of the strokes.

•The names used to distinguish weight are not uniform between type suppliers. The commonly used names are: ultra light, extra light, light, semi light, medium, semi bold, bold, extra bold, etc.

•***Width*** — the amount of expansion or contraction with respect to the normal or medium in the family.
•***Size*** — unit is *point*. 1 inch = 72.27 point in printing industry. 1 inch = 72 point in PostScript systems.



The table above illustrates the relation of width and weight for the entire range available in the family Univers.

2.2 Classication of Typeface

- Typefaces can be classified in many ways. One classification is understood universally: *serif* and *sans serif*.
- *Serif* is the little flag or decoration at the end of a stroke.
- On printed pages, serif fonts are used for body text while sans serif fonts are used for headline because the serifs helps guide the reader's eye along the line of text.
- Multimedia presentation are displayed on low resolution screen where sans serif fonts will be far more legible.

2.3 Bitmap Fonts Versus Outline Fonts

Font formats can be divided into two main categories: *bitmap* fonts and *outline* fonts.

•Bitmap fonts come in specific sizes and resolutions. Because the font contain the bitmaps of the character shapes. The result will be very poor if they are scaled to different sizes.

•Outline fonts contain the outline of the characters. They can be scaled to a large range of different sizes and still have reasonable look.

•They need a rasterizing process to display on screen.

•Nowadays, outline fonts are much more common than bitmap fonts. There are two kinds of outline fonts: *PostScript* and *TrueType*.

•All version of Windows support TrueType fonts. Windows3.1 and Windows95 require Adobe Type Manager (ATM) to display PostScript fonts. PostScript printers have a number of resident PostScript fonts.

2.4 Measurements for Text Layout

- ***Leading*** is the distance between the baselines of two adjacent lines. Common used leadings are 14 points for 12 points text, 12 points for 10 points text.
- ***Tracking*** is the spacing between characters in text lines.
- **Loose tracking** means the space between characters are wider. Less words can be put in a line of text.

•***Kerning*** is the extra adjustment between two specific characters. Normally, characters are place one next to the other, i.e., the distance between the origins of the adjacent characters is equal to the character width. But due to the shape of the characters, the space between certain characters may look uneven, e.g., the A and v in the figure. Therefore, we need to kern the characters.

3 Using Text in multimedia

Picking the fonts to use in a multimedia presentation may be difficult. Here are some suggestions:

- For small type, use the most legible font available, decorative fonts are useless.
- Use as few different faces as possible in the same work, but vary the weight and the size and using italic or bold styles.
- In text block, adjust the leading for the most pleasing line spacing. Lines too tightly packed are difficult to read.

•Vary the size of a font in proportion to the importance of the message.
• In large size headline, do proper kerning so that the spacing feels right.
• Explore the effects of different colours and of placing the text on various backgrounds.

3.1 Cross platform issues

•When you build your multimedia project on Windows platform, and play it back on a Macintosh platform , there will be some differences.

•Fonts are perhaps the greatest cross-platform concern. If a specified font does not exist in the target machine, a substitute must be provided. Some cross-platform applications, e.g., Director, allow the developer to specify the mapping of fonts.
•Different encodings on different platform is also a big problem. Special characters may need to be converted to bitmaps in order to be display correctly on different platforms. Different systems and font manufacturers encode different symbols in the extended character set. For example, the code 165 may be a bullet () on the the Macintosh, the character for Japanese yen in Windows.

4. Exercises

1. Investigate your favourite word processor or desktop publishing package to see:

• which of the following features are available: leading, tracking, kerning, character width, font size

• how many outline fonts and bitmap fonts are available

2. Experiment with your favourite word processor to find out what is the best leading at a certain point sizes, say 10, 12 and 14, and a certain fonts, for example Times Roman, Arial and Courier, for a normal text paragraph.