Multimedia Data Types: Text

Lecture 02

BIL464 Multimedia Systems

Mustafa Sert
Asst. Prof.
msert@baskent.edu.tr

Department of Computer Engineering, Başkent University
Ankara 06810 TURKEY

Content

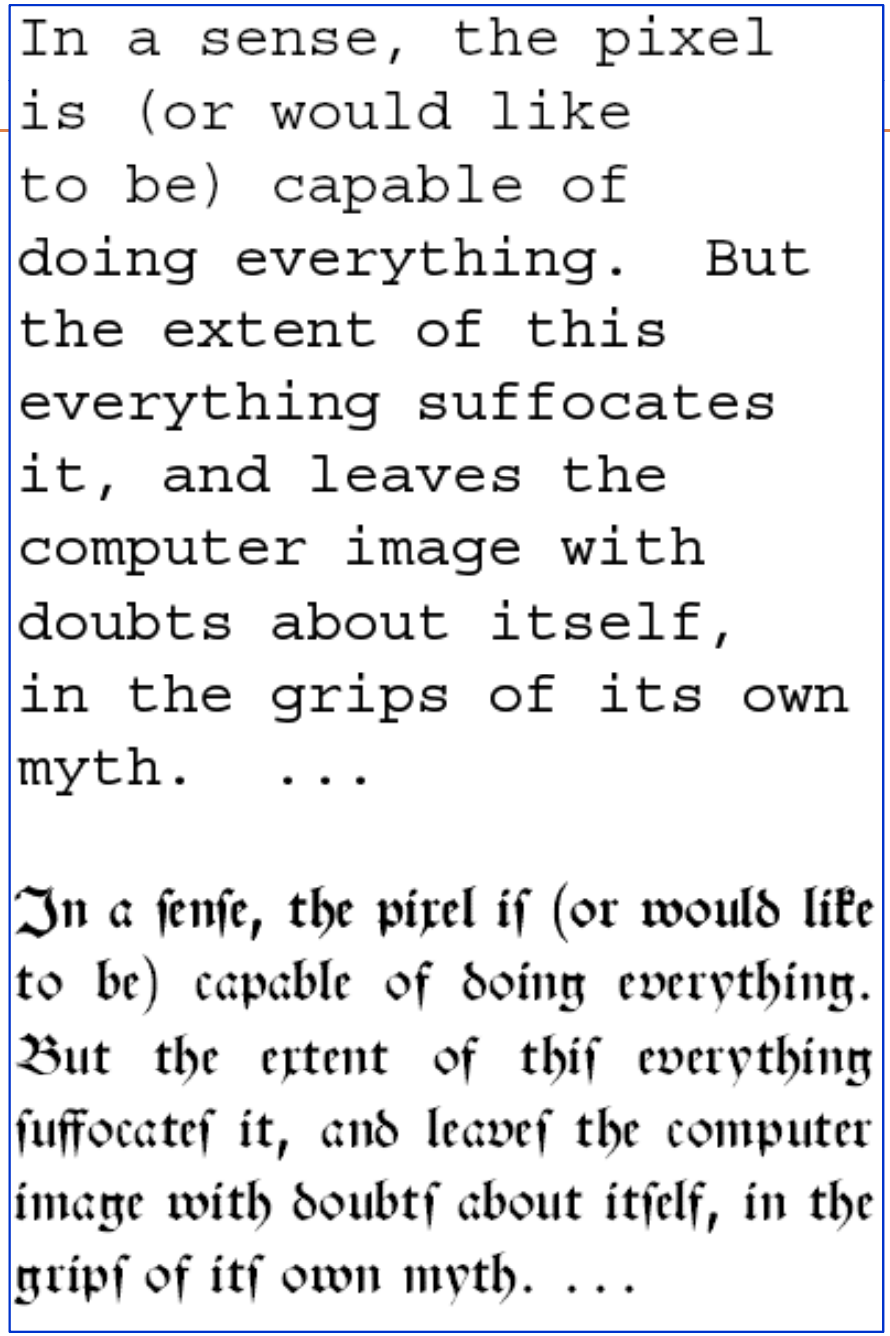
- What is Text?
- Character sets
- Typefaces and fonts
  - Outline vs. Bitmap Fonts
- Using text in multimedia

What is Text?

- Text is **vital** element of multimedia presentations
- Words and symbols in any form, spoken or written, are the most common system of communication
- It is very important to choose the suitable words and symbols in your multimedia presentation
- We will concern with another aspect of text, namely its appearance in multimedia presentations
- **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

Character Sets

- The visual appearance of a piece of text can be in many different forms
- 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: the letters A to Z, the lower case letters a to z, the digits and a number of punctuations



Character Sets – *Cont.*

- 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*.
- Clearly, if any systems want to communicate 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.

Character Sets – *Cont.*

- For instance, the earliest common character set is ASCII (American Standard Code for Information Interchange) character set.
  - The code range is 7-bit, meaning a code value can be stored in 7-bits
  - Therefore, at most 128 characters can be coded
  - It only comprises 95 printable characters
  - The values 0 to 31 and 127 are assigned to *control characters*.
  - ISO adopted ASCII as an standard (ISO 646)

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

Character Sets – *Cont.*

- Obviously, 128 values are not enough to code many of the world's languages. ISO produced a new standard ISO 8859 with 8-bit characters.
  - ISO 8859 has many parts. Each part specifies a number of character sets.
  - The lower 128 characters in all parts are identical to ASCII.

8859-1	Western Europe, Latin America, Caribbean, Canada, Africa
8859-2	Eastern Europe
8859-3	SE Europe/miscellaneous (Esperanto, Maltese, etc.)
8859-4	Scandinavia/Baltic (mostly covered by 8859-1 also)
8859-5	Cyrillic
8859-6	Arabic
8859-7	Greek
8859-8	Hebrew
8859-9	Latin5, same as 8859-1 except for Turkish instead of Icelandic
8859-10	Latin6, for Lappish/Nordic/Eskimo languages
8859-13	Latin7
8859-14	Latin8
8859-15	Latin9

Character Sets – *Cont.*

- Unicode produced a standard *The Unicode Standard, Version1.0 in 1991. The* latest version is now version 6.2 (2012, ISO/IEC 10646:2012 plus the **Turkish lira sign**)
  - Uses 2 bytes to encode each character.
  - Attempts to specify a character set to embrace all languages of the world.
  - The latest Unicode standard has more than 27484 Chinese characters.
- There have been many Chinese character set standards before the Unicode:
  - **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

Encoding

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

Binary	Oct	Dec	Hex	Glyph
010 0000	040	32	20	SP
010 0001	041	33	21	!
010 0010	042	34	22	"
010 0011	043	35	23	#
010 0100	044	36	24	$
010 0101	045	37	25	%
010 0110	046	38	26	&
010 0111	047	39	27	'
010 1000	050	40	28	(
010 1001	051	41	29	)
010 1010	052	42	2A	*
010 1011	053	43	2B	+
010 1100	054	44	2C	,
...				

Binary	Oct	Dec	Hex	Glyph
100 0000	100	64	40	@
100 0001	101	65	41	A
100 0010	102	66	42	B
100 0011	103	67	43	C
100 0100	104	68	44	D
100 0101	105	69	45	E
100 0110	106	70	46	F
100 0111	107	71	47	G
100 1000	110	72	48	H
100 1001	111	73	49	I
100 1010	112	74	4A	J
100 1011	113	75	4B	K
100 1100	114	76	4C	L
...				

Binary	Oct	Dec	Hex	Glyph
110 0000	140	96	60	`
110 0001	141	97	61	a
110 0010	142	98	62	b
110 0011	143	99	63	c
110 0100	144	100	64	d
110 0101	145	101	65	e
110 0110	146	102	66	f
110 0111	147	103	67	g
110 1000	150	104	68	h
110 1001	151	105	69	i
110 1010	152	106	6A	j
110 1011	153	107	6B	k
110 1100	154	108	6C	l
...				

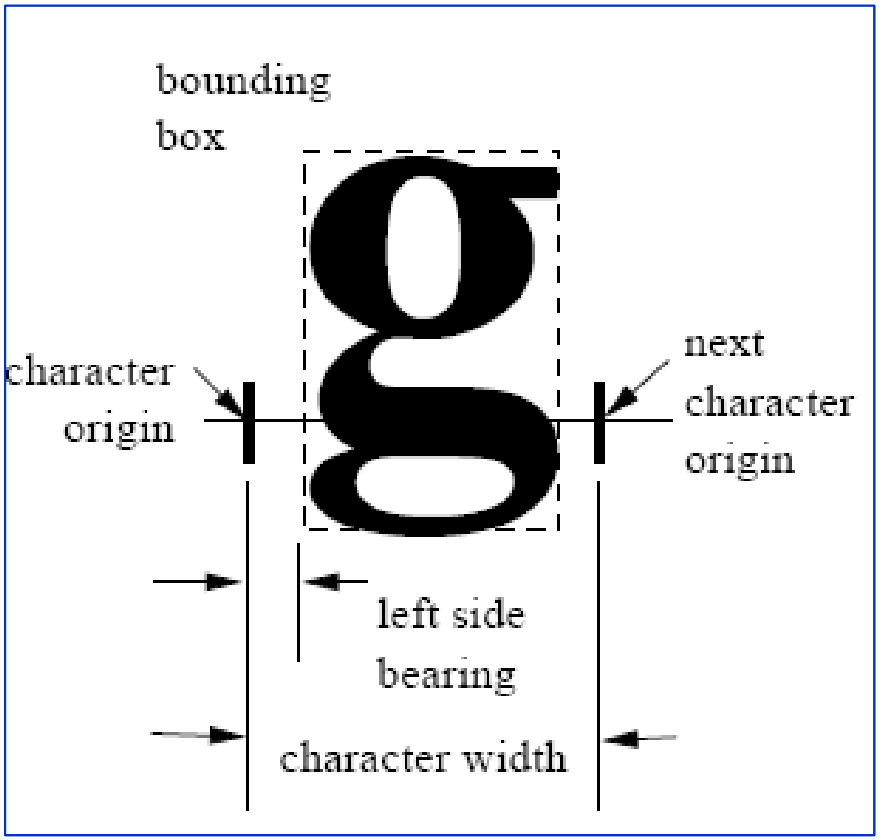
ASCII encoding

Typefaces and Fonts

- To display text, we need to have a visual representation of the characters stored as codes in the computer.
- 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 'Arial'. The font is 'Arial Narrow (Body) 24 point'.

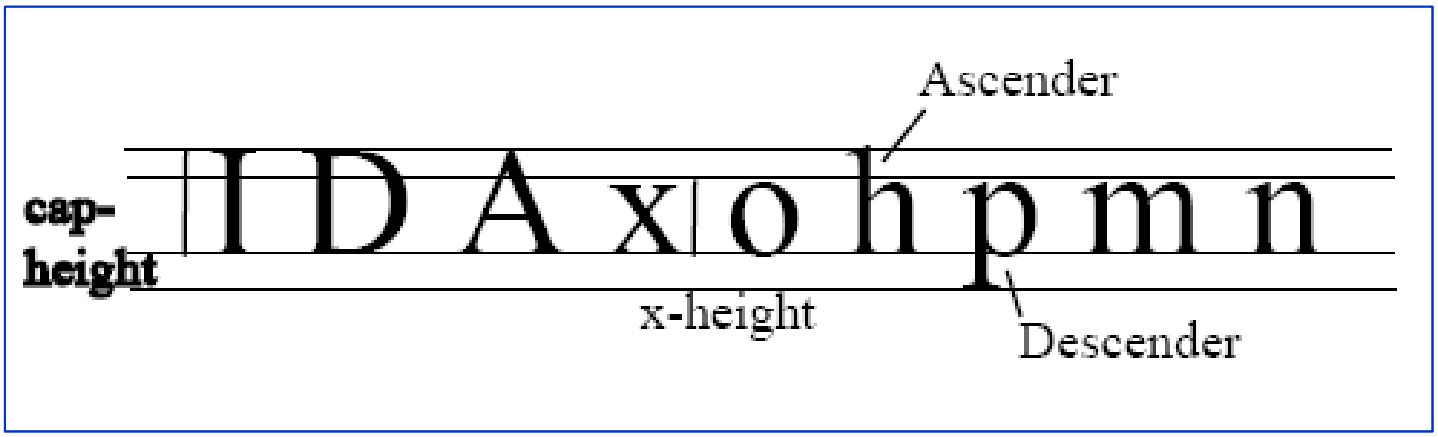
Measurements of the type

- 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*



Measurements of the type – *Cont.*

- 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*.

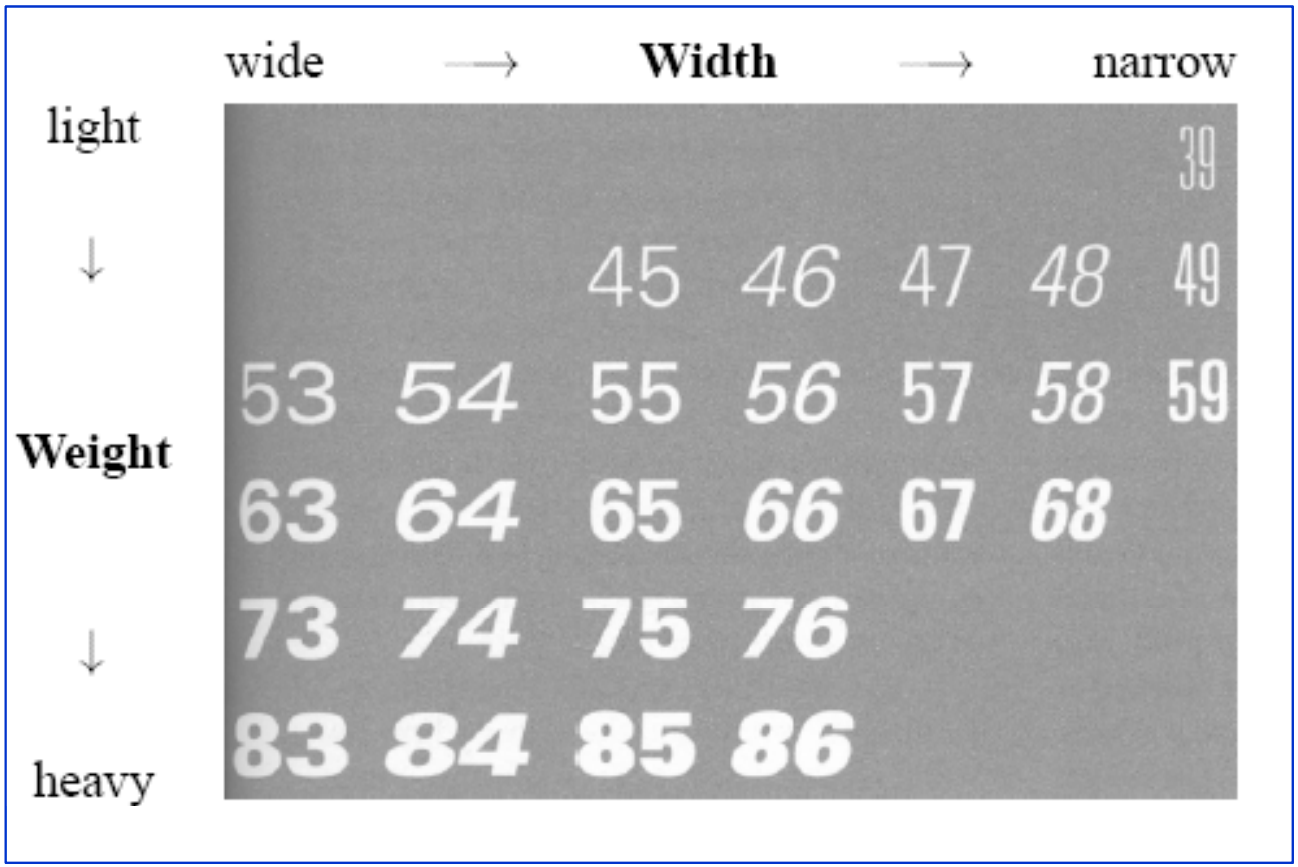


Measurements of the type – *Cont.*

- There are many fonts available. But, 5 (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:* 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.

Measurements of the type – *Cont.*

- Below table depicts the relation of width and weight for the entire range available in the family Univers.

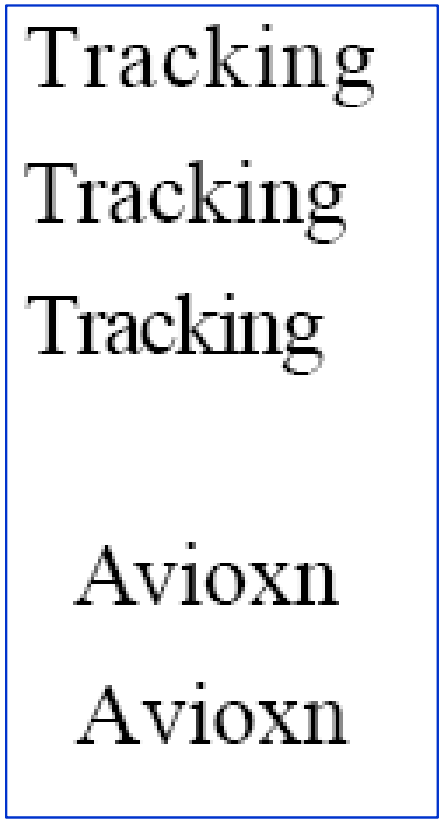


Bitmap vs. Outline Fonts

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

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. 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



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.
  - 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