多媒体技术

1.3 多媒体技术研究的主要内容

1.3.1 什么是多媒体技术

- 电视计算机teleputer
  - HDTV、常规电视数字化、交互式电视、影视节目制作
  - VCD、DVD
  - 智能家居
- 计算机电视compuvison

1.3.2 多媒体技术研究的主要内容

- 常规电视数字化
  - 采用数字式视频、数字化音频以及MPEG压缩编码方法
  - 汤姆逊消费电子公司
- 交互式电视技术ITV
  - 节目间交互：VOD、near VOD
  - 节目内交互

1.3.2 多媒体技术研究的主要内容

- VCD
  - 多媒体压缩编码技术
  - 美国C-Cube公司先后生产了JPEG静态图像压缩解压缩处理器和MPEG-I解码器，在Comdex展会上受到欢迎，后来用到VCD中。
    - 万燕公司
    - 姜万勐、孙燕生

1.3.2 多媒体技术研究的主要内容

- 智能家居
  - 通过物联网技术将家中的各种设备连接到一起
  - 如音视频设备、照明系统、窗帘控制、空调控制、安防系统、数字影院系统、影音服务器、影柜系统、网络家电

1

1.3.2 多媒体技术研究的主要内容

2016年微软：智能冰箱

利用计算机视觉技术，背后是深度学习算法，识别物体

1.3.2 多媒体技术研究的主要内容

- 多媒体技术的基础
  - 媒体
    - 媒体的性质与相应的处理方法
    - 与心理学有关
    - 感觉相乘效应
    - 每一种媒体的采集、存储、传输和处理
  - 数据压缩
    - 文本，视频的特性
    - MPEG，JPEG，H.26L等

1.3.2 多媒体技术研究的主要内容

关键技术

- 数据存储技术
  - 数字化的多媒体信息经过压缩后仍有大量数据，提出两方面要求:
    - 大容量存储技术;
    - 足够的数据传送带宽和支持多媒体的实时处理功能。
  - 在大容量、高速度和低价格的存储器未解决之前，只读光盘、U盘是最受欢迎的、理想的多媒体存储介质。
  - SSD
  - Intel全新的存储技术3D XPoint

1.3.2 多媒体技术研究的主要内容

关键技术

- 多媒体的压缩编码和解码技术
  - 一幅图：分辨率512*512，RGB三个颜色，每个颜色用8比特表示;
  - PAL制式：25帧/s，NTSC制式：30帧/s;
  - 512*512*8*3*25（30）= 180M/s
  - 过去的总线ISA：150KB/s，1.2M/s
  - VCD的标准MPEG-1的固定传输率1.5Mbps
    - 帧内压缩、帧间压缩

1.3.2 多媒体技术研究的主要内容

关键技术

- 多媒体信号的实时处理技术
  - 电视扫描光栅一行是64微秒，正程52.2微秒，逆程11.8微秒，分辨率512个像素，一个像素0.1微秒;
  - 实时：0.1微秒的时间处理完一个像素。

1.3.2 多媒体技术研究的主要内容

关键技术

- 多媒体信号的实时处理技术
  - **R5(中心像素)=R1G1 + R2G2 + R3G3 + R4G4 + R5G5 + R6G6 + R7G7 + R8G8 + R9G9**

$$R=\begin{bmatrix} R1 & R2 & R3 \\ R4 & R5 & R6 \\ R7 & R8 & R9 \end{bmatrix} \quad G=\begin{bmatrix} G1 & G2 & G3 \\ G4 & G5 & G6 \\ G7 & G8 & R9 \end{bmatrix}$$

1.3.2 多媒体技术研究的主要内容

- 多媒体软硬件平台技术
  - 硬件
    - 光盘驱动器、声音适配器、图形显示卡、扫描仪、打印机数码相机、带震动感的鼠标、交互式键盘遥控器
  - 软件
    - 操作系统
    - 编辑创作软件： PS、会声会影、Cool edit
    - 专用软件：Unity 3D

1.3.2 多媒体技术研究的主要内容

关键技术

- 专用芯片
  - 视频信号和音频信号数据实时压缩和解压缩处理需要进行大量复杂计算，普通计算机无法胜任，因此，多媒体专用芯片技术是多媒体发展的关键技术。
    - 美国C-Cube公司的CL-550芯片可用30f/s的速度完成静止图像JPEG算法;
    - CL-450芯片实现动态图像MPEG1算法的实时解压缩;
    - Intel公司的i750芯片组可实时完成DVI视频图像的编码和解码算法。

1.3.2 多媒体技术研究的主要内容

关键技术

- 专用芯片
  - 用于多媒体技术的专用芯片常见的两种类型
    - 一种是有固定功能的芯片，目标是提高图像数据的压缩率；
    - 一种是具有可编程的处理器，目标是提高图像的运算速度。
      - GPU
      - CUDA/OpenGL/OpenCL

1.3.2 多媒体技术研究的主要内容

- 多媒体信息管理与处理技术
  - 多媒体数据库
    - 扩展现有的关系数据库
    - 建立面向对象数据库系统
    - 三个方面
      - 体系结构，数据模型和用户接口
  - 多媒体信息的分析与处理
    - 基于内容的检索
      - 图像、视频、音频

1.3.2 多媒体技术研究的主要内容

关键技术

- 信息的管理
  - 文件系统管理方式
    - 多媒体信息以文件的形式存储在计算机中，操作系统的文件管理功能可以实现信息存储管理等。
    - 对于不同格式的文件采用相应的软件进行打开、编辑、修改。
    - 当多媒体信息较少时，浏览查询方式快捷，当多媒体数量和种类较多时，管理不方便。

1.3.2 多媒体技术研究的主要内容

关键技术

- 信息的管理
  - 扩充关系数据库方式
    - 用专用字段存放全部多媒体文件。
    - 多媒体数据分段存放在不同字段，调用时再重新构建。
    - 文件系统与数据库相结合，多媒体数据以文件系统存放，用关系数据库存放媒体类型、应用程序名、媒体属性、关键词等。

1.3.2 多媒体技术研究的主要内容

关键技术

- 信息的管理
  - 面向对象数据库方式
    - 将面向对象程序设计语言与数据库技术结合，通过引入对象、类、方法、消息、封装、继承等概念，有效的描述各种对象以及内部结构联系，因此，适合于描述多媒体信息。

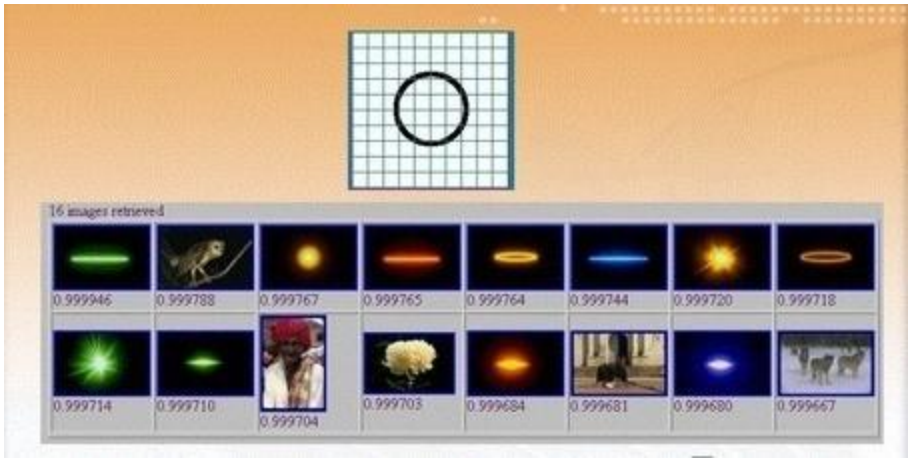
1.3.2 多媒体技术研究的主要内容

关键技术

- 多媒体信息的检索
  - 基于内容的多媒体信息检索技术，通过对多媒体对象的内容及上下文语义环境进行检索，如对图像中的颜色、纹理，视频中的场景、片断，音乐的旋律、音调、音质等进行分析和特征提取，基于这些特征进行相似性匹配。
  - MPEG-7标准称为“多媒体内容描述接口”（Multimedia content description interface），是一种多媒体内容描述的标准。

1.3.2 多媒体技术研究的主要内容

关键技术



Sketch Recognize Demo

zdd
zddhub@gmail.com
http://sr.opensse.com

1.3.2 多媒体技术研究的主要内容

- 网络多媒体与应用技术
  - 超媒体被称为天然的多媒体信息管理方法
  - 多媒体系统一般都是基于网络的分布应用系统
  - 多媒体网络系统
    - IP上的多媒体
  - 典型的应用
    - 基于计算机的会议系统
    - 计算机支持的协作工作
    - 视频点播
    - 交互式电视

1.3.2 多媒体技术研究的主要内容

- 网络多媒体与应用技术
  - 实时性和同步性
  - **HTML5**
    - 移动互联网，超文本标记语言的第五次重大修改，支持**MPEG-4**、**H.264**及**WebM**等影音编码
    - http://www.html5tricks.com/demo/jiaoben1144/index.html
  - **WebGL**
    - **JavaScript API**，用于在兼容的**Web**浏览器中呈现交互式**3D**和**2D**图形，免去开发网页专用渲染插件
    - https://www.html5tricks.com/demo/html5-webgl-galaxy/index.html

1.4 小结

本章重点

- 多媒体技术概念
- 多媒体的三大关键特性
  - 信息载体的多样性、交互性和集成性
- 多媒体是技术与应用发展的必然产物
- 多媒体改善了人类信息的交流，缩短了人类信息交流的路径
- 多媒体技术研究内容

多媒体技术学术资料

一、 期刊

- ACM Transactions on Graphics
- IEEE Transactions on Image Processing
- IEEE Transactions on Visualization and Computer Graphics
- ACM Transactions on Multimedia Computing, Communications and Application
- Computer Graphics Forum
- IEEE Transactions on Multimedia
- Computer Animation and Virtual Worlds
- Multimedia Tools and Applications
- The Visual Computer
- 计算机学报、软件学报、计算机研究与发展、计算机辅助设计与图形学学报

多媒体技术学术资料

- 二、 会议
- ACM International Conference on Multimedia
- IEEE Visualization Conference
- IEEE Virtual Reality
- ACM SIGMM International Conference on Multimedia Retrieval
- ACM/Eurographics Symposium on Computer Animation
- Data Compression Conference
- IEEE International Conference on Multimedia& Expo
- Pacific Graphics: The Pacific Conference on Computer Graphics and Applications
- International Conference on Multimedia Modeling
- 中国多媒体大会、中国虚拟现实大会、中国计算机图形学大会等
- 全国多媒体课件大赛

第二章 媒体及媒体技术

2.1 媒体的种类和特点

2.1.1 常见的媒体元素

- 文本
  - 字符代码的识别是计算机文字处理程序的基础
    - 英文：ASCII
    - 当字节中的最高位b_7=0时，其余7位(b_6~b_0) 产生0~127之间的128个字符的代码，分成两类：
      - 控制字符，也称非打印字符，其代码范围是0~31和127
      - 可打印字符，也称显示字符，其代码范围是32~126
    - 当字节中的最高位b_7=1时，其余7位(b_6~b_0) 产生128~255之间的128个字符的代码
      - 对不同的计算机系统、程序、字体或图形字符，扩展ASCII码指定的字符通常是不同的

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

2.1.1 常见的媒体元素

- 汉字
  - Hanzi，Hantsu，汉字
  - Ideographic character，表意字符，中文字符
  - Kanji，日文中的叫法
  - Hanja，朝鲜文中的叫法
  - CJK，中日韩通用字符集
  - Unihan

2.1.1 常见的媒体元素

- 汉字
  - 字符编码
    - 给字符集中的每一个字符指定一个编号或存放位置，称为“码位”
    - 然后给它分配一个数字，称为“字符代码”，简称“代码”
  - 特点
    - 汉字数量大，目前统计超8万，常用汉字3500

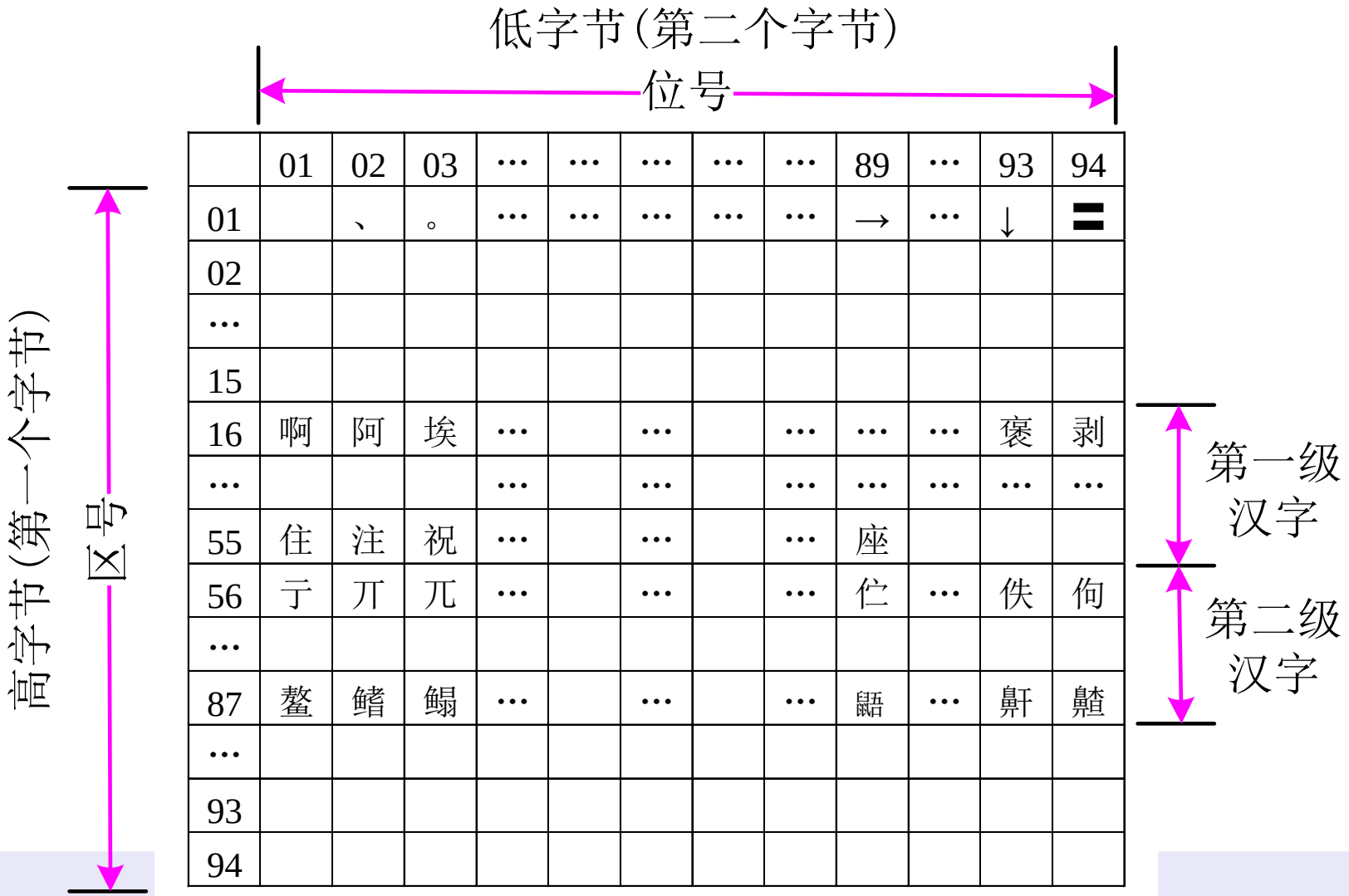
2.1.1 常见的媒体元素

- GB 2312标准
  - 是我国第一个汉字编码标准(1980)
    - 收录字符数
      7445个 = 6763个汉字 + 682个全角字符
    - 编码空间或码位空间
      - 编码方法可表达的字符数目
      - 用两字节(16位)编码，最大编码空间为65 536个字符
      - 双字节字符集(double-byte character set，DBCS):
        用两字节(16位)代码表示字符构成的字符集
    - GB 2312标准使用双字节的编码方法

2.1.1 常见的媒体元素

- GB 2312标准
  - 区位码
    - 7445个字符组成94 × 94方阵，每一行称“区”，编号0l ~ 94；每一列称为“位”，编号0l ~ 94
    - 编码空间：94 × 94 = 8836个字符
    - 6763个汉字分成两级
      - 常用汉字3775个，按汉语拼音字母排序
      - 不太常用汉字3008个，部首笔画排序
    - 1-9区：西文字母、数字、日文假名、图形符号
    - 16-87区：汉字
      - 16-55区：一级汉字，40*94-5=3755个
      - 56-87区：二级汉字，32*94=3008个
    - 10-15区和88-94区：用户自定义

2.1.1 常见的媒体元素



	01	02	03	…	…	…	…	…	89	…	93	94
01		、	。	…	…	…	…	…	→	…	↓	〓
02												
…												
15												
16	啊	阿	埃	…		…		…	…	…	褒	剥
…				…		…		…	…	…	…	…
55	住	注	祝	…		…		…	座			
56	亍	丌	兀	…		…		…	伫	…	佚	佝
…												
87	鳌	鳍	鳎	…		…		…	鼍	…	鼾	齾
…												
93												
94												

2.1.1 常见的媒体元素

- GB 2312标准
  - 国标码(国家标准代码/汉字交换码)
    - 区位码表示字符所在位置的编号，对这些编号的编码即国标码
    - 国标码由两个字节组成，其高字节和低字节的最高位(b_7)都为0，与7位标准ASCII码类似
    - 编码方法
      - 先将十进制表示的区码和位码转换为十六进制表示的区码和位码，再将这个代码的高字节(第一个字节)和低字节(第二个字节)分别加上20H(10 0000B)，就得到国标码
      - 每个字节加20H原因是为避开ASCII码表的32个控制字符

2.1.1 常见的媒体元素

- GB 2312标准
  - 国标码(国家标准代码/汉字交换码)
    - “啊”字：区码和位码合起来构成的区位码是1601，变成十六进制表示的区位码是0x1001，高字节(10)和低字节(01)分别加上20后就得到“啊”的国际码0x3021

2.1.1 常见的媒体元素

- GB 2312标准
  - 机内码(内码)
    - 将国标码的高字节和低字节的最高位(b_7)都变成1，或者说每个字节都加0x80
    - 解决ASCII码和国标码在同时使用时产生的二义性
    - “啊”的区位码是1601，国标码是0x3021，机内码是0xB0A1

2.1.1 常见的媒体元素

- **GB 2312标准**
  - 机内码(内码)

表 2-7 16 区字符的位号

	+0	+1	+2	+3	+4	+5	+6	+7	+8	+9	+10	+11	+12	+13	+14	+15
0		啊	阿	埃	挨	哎	唉	哀	皑	癌	蔼	矮	艾	碍	爱	隘
16	鞍	氨	安	俺	按	暗	岸	胺	案	肮	昂	盎	凹	敖	熬	翱
32	袄	傲	奥	懊	澳	芭	捌	扒	叭	吧	笆	八	疤	巴	拔	跋
48	靶	把	耙	坝	霸	罢	爸	白	柏	百	摆	佰	败	拜	稗	斑
64	班	搬	扳	般	颁	板	版	扮	拌	伴	瓣	半	办	绊	邦	帮
80	梆	榜	膀	绑	棒	磅	蚌	镑	傍	谤	苞	胞	包	褒	剥	

表 2-8 16 区字符的机内码(与表 2-7 对应)

机内码	+0	+1	+2	+3	+4	+5	+6	+7	+8	+9	+A	+B	+C	+D	+E	+F
B0A0		啊	阿	埃	挨	哎	唉	哀	皑	癌	蔼	矮	艾	碍	爱	隘
B0B0	鞍	氨	安	俺	按	暗	岸	胺	案	肮	昂	盎	凹	敖	熬	翱
B0C0	袄	傲	奥	懊	澳	芭	捌	扒	叭	吧	笆	八	疤	巴	拔	跋
B0D0	靶	把	耙	坝	霸	罢	爸	白	柏	百	摆	佰	败	拜	稗	斑
B0E0	班	搬	扳	般	颁	板	版	扮	拌	伴	瓣	半	办	绊	邦	帮
B0F0	梆	榜	膀	绑	棒	磅	蚌	镑	傍	谤	苞	胞	包	褒	剥	

2.1.1 常见的媒体元素

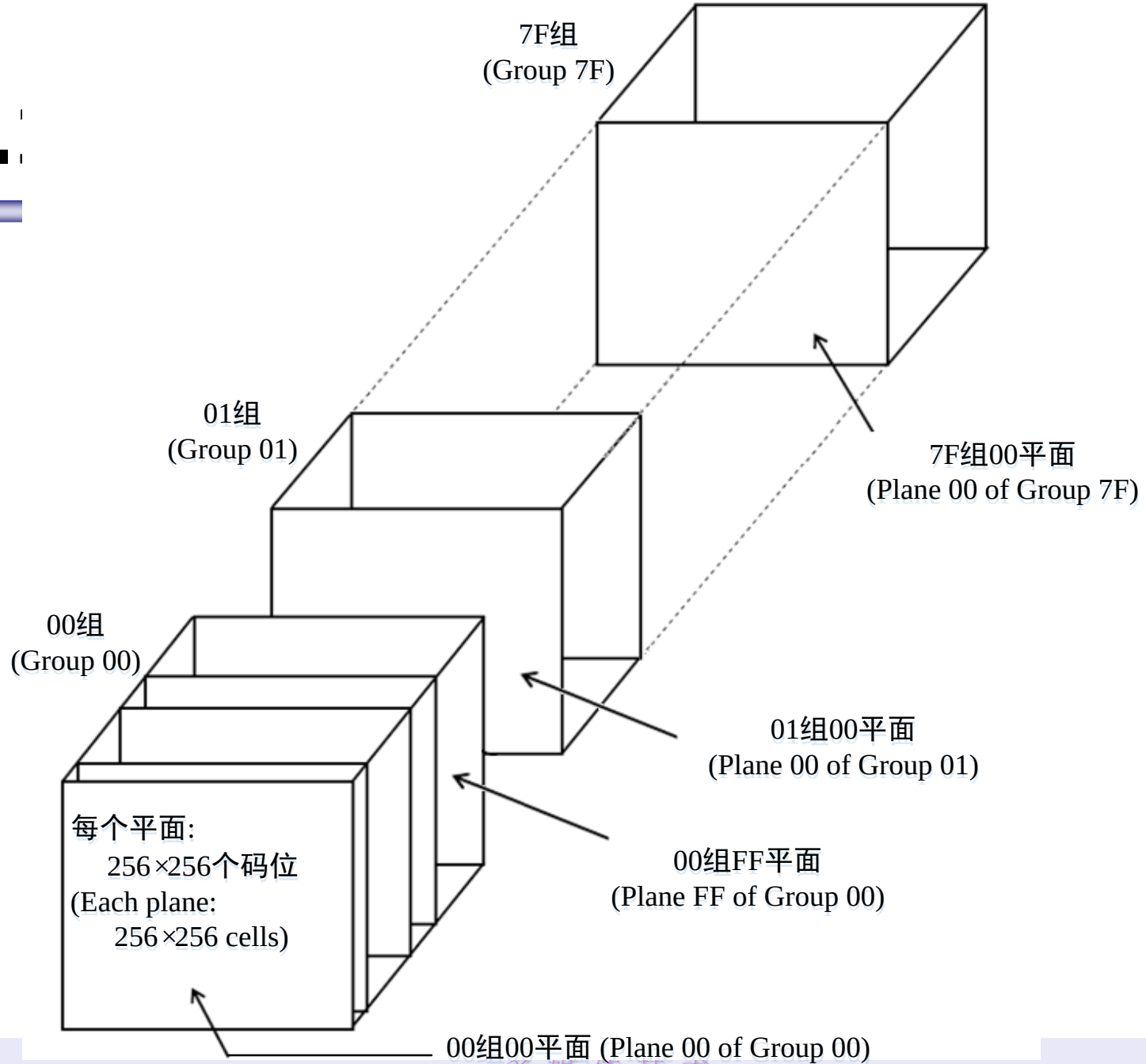
- GB 18030-2005标准
  - 替代先前字符编码标准
    - GB 13000.1-1993
    - GBK-1995
    - Big5-2003
    - GB 18030-2000
  - 支持Unicode的CJK统一汉字，70244个汉字
  - 增加少数民族文字的编码

2.1.1 常见的媒体元素

- **ISO/IEC 10646标准**
  - **1989年开始开发的单一字符集——“通用字符集(Universal Character Set，UCS)”字符编码标准**
  - 字符集：包含已知语言的所有字符以及大量的图形、印刷、数学和科学符号
    - 拉丁语、希腊语、斯拉夫语、希伯来语、阿拉伯语、亚美尼亚语、格鲁吉亚语
    - 中、日、韩的象形文字，或称表意文字。
    - 古埃及的象形文字和不常见的汉字

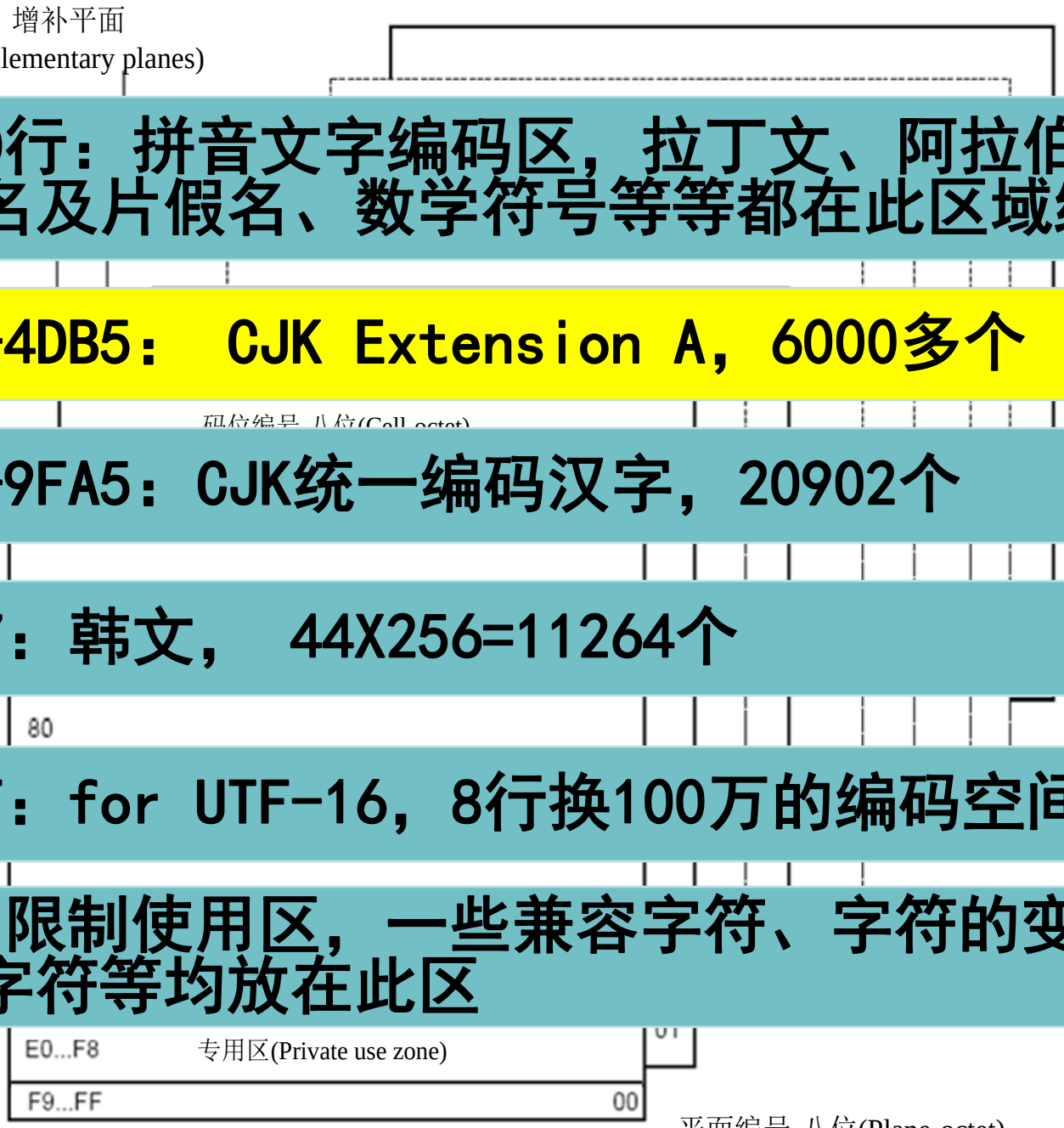
2.1.1 常见的媒体元素

- ISO/IEC 10646标准
  - 字符集的编排结构
    - 共有128个组(group)，每组有256个平面(plane)，每个平面由256行(row) × 256个码位(cell)组成
    - 合计码位：128 × 256^3 =2 147 483 648个码位
  - 两种编码方案：UCS-4
    - 每个字符的码位用4个字节表示，分别表示组、平面、行和码位
    - 如：00 00 00 30H 表示数字“0”
    - 如：00 00 00 41H 表示字母“A”
  - 两种编码方案：UCS-2


7F组
(Group 7F)
01组
(Group 01)
00组
(Group 00)
7F组00平面
(Plane 00 of Group 7F)
01组00平面
(Plane 00 of Group 01)
00组FF平面
(Plane FF of Group 00)
每个平面:
256×256个码位
(Each plane:
256×256 cells)
00组00平面 (Plane 00 of Group 00)

2.1.1 常见的媒体元素

- ISO/IEC 10646标准
  - 基本多文种平面(Basic Multilingual Plane，BMP)
    - 00组00平面，在此平面上用行、列两个八位表示一个字符
    - 如：00 30H表示“0”，00 41H表示“A”
    - 各种文字的最常用字符，包括CJK字符



00-4D行：拼音文字编码区，拉丁文、阿拉伯文、日文的平假名及片假名、数学符号等等都在此区域编码

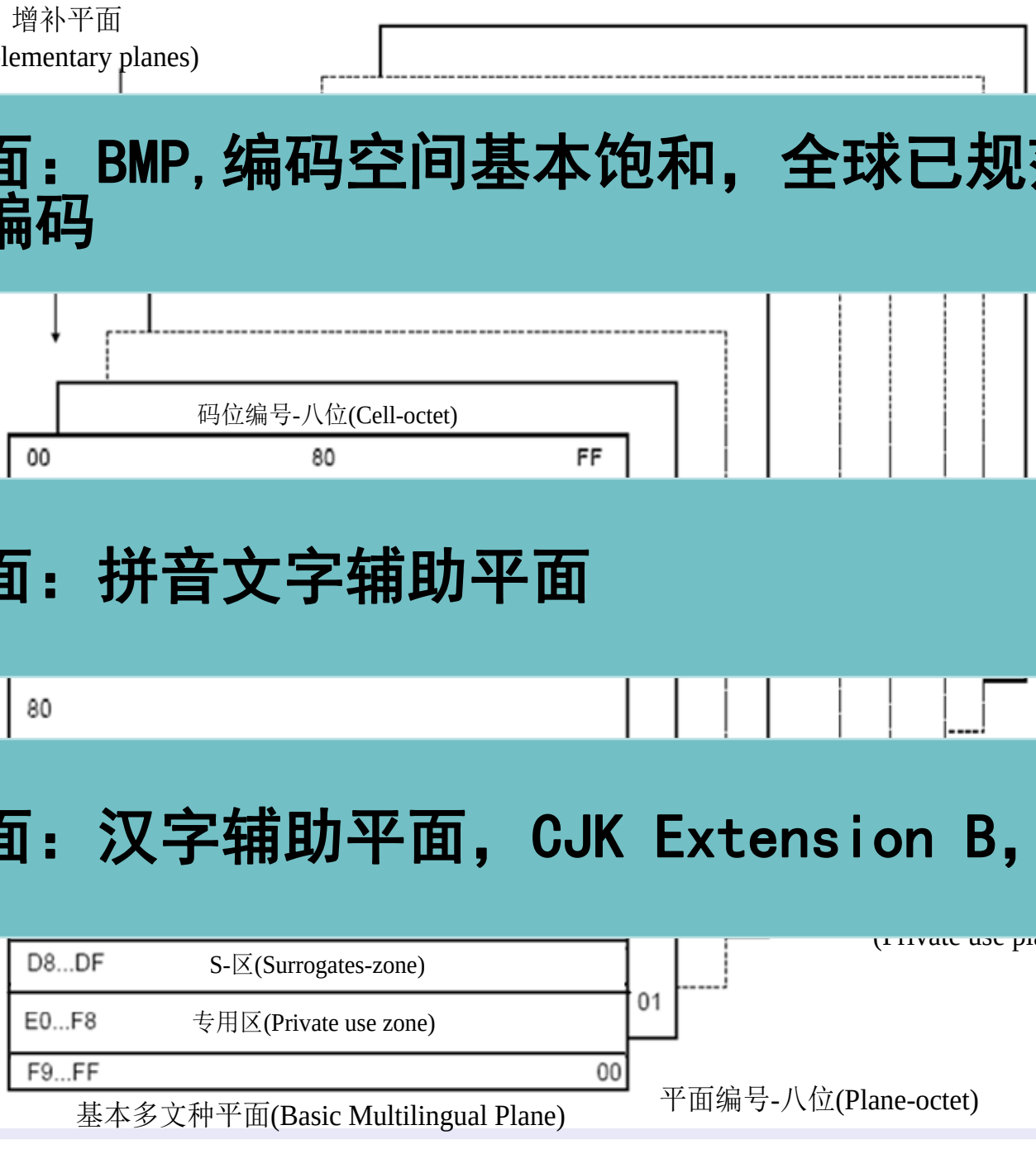
3400-4DB5： CJK Extension A，6000多个

4E00-9FA5：CJK统一编码汉字，20902个

AC-D7：韩文， 44X256=11264个

D8-DF：for UTF-16，8行换100万的编码空间

E0-：限制使用区，一些兼容字符、字符的变形显现形式、特殊字符等均放在此区



00平面：BMP，编码空间基本饱和，全球已规范语种的基本文字编码

01平面：拼音文字辅助平面

02平面：汉字辅助平面，CJK Extension B，4万多

2.1.1 常见的媒体元素

- Unicode标准
  - UCS：ISO/IEC开发的单一字符集
  - Unicode：Unicode联盟开发的单一字符集
  - 世界不需要两个不兼容的字符集
    - 从Unicode 2.0开始保持两个标准码表的兼容性，共同调整未来的扩展工作。

2.1.1 常见的媒体元素

- Unicode标准
  - 编码空间：17个平面(平面0，…，16)
    - 用5位(bit)表示平面的编号
    - 每个平面包含256行 × 256码位/行=65536个码位，用十六进制数表示，用于表达有含义的抽象字符
    - 在理论上说，编码空间可容纳17 × 256 × 256 =1 114 112字符，编码空间的范围为0 ~ 0x10FFFF

2.1.1 常见的媒体元素

- Unicode标准

安排在BMP上的字符的码位用16位表示，与UCS-2一致，最多可表示65536个字符，基本满足各种语言的使用

平面编号	码位范围(HEX)	说明
0	0000–FFFF	基本多文种平面(Basic Multilingual Plane，BMP)
1	10000–1FFFF	多文种增补平面(Supplementary Multilingual Plane，SMP)
2	20000–2FFFF	表意词增补平面(Supplementary Ideographic Plane
3～13	30000–DFFFF	当前未分配
14	E0000–EFFFF	特殊用途增补平面(Supple e Plane，SSP)

其他平面上的字符码位用四个字节表示，与UCS-4一致

2.1.1 常见的媒体元素

- Unicode标准
  - 2009年10月发布了Unicode 5.2.0版本，定义了107 361个字符的码位
  - 可计算得到Unicode 5.2.0定义的中日韩CJK统一汉字

Unicode 5.2增加的汉字放在9FA6-9FCB的位置

字符集名称	编码范围(十六进	字符数
CJK 统一汉字	4E00～9FCB	20940
CJK 统一汉字扩充 A	3400～4DFB	6652
CJK 统一汉字扩充 B	20000～2A6DF	42720
CJK 统一汉字扩充 C	2A700～2B73F	4160
CJK 兼容汉字	F900～FAFF	512
CJK 兼容汉字补充	2F800～2FA1D	542
CJK 部首/康熙字典部首	2F00～2FDF	224
CJK 字根补充	2E80～2EFF	128
CJK 笔画	31C0～31EF	48
CJK 描述字符	2FF0～2FFF	16

2.1.1 常见的媒体元素

一	丁	丐	七	丄	丅	丆	万	丈	三	上	下	丌	不	与	丏
4E00	4E01	4E02	4E03	4E04	4E05	4E06	4E07	4E08	4E09	4E0A	4E0B	4E0C	4E0D	4E0E	4E0F
丐	丑	丒	专	且	丕	世	丗	丘	丙	业	丛	东	丝	丞	丢
4E10	4E11	4E12	4E13	4E14	4E15	4E16	4E17	4E18	4E19	4E1A	4E1B	4E1C	4E1D	4E1E	4E1F
北	两	丧	丩	两	严	並	丧	丨	丩	个	丫	丬	中	丮	丯
4E20	4E21	4E22	4E23	4E24	4E25	4E26	4E27	4E28	4E29	4E2A	4E2B	4E2C	4E2D	4E2E	4E2F
丰	丱	串	弗	临	丵	丶	丷	丸	丹	为	主	丼	丽	举	丿
4E30	4E31	4E32	4E33	4E34	4E35	4E36	4E37	4E38	4E39	4E3A	4E3B	4E3C	4E3D	4E3E	4E3F
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
9F80	9F81	9F82	9F83	9F84	9F85	9F86	9F87	9F88	9F89	9F8A	9F8B	9F8C	9F8D	9F8E	9F8F
龐	龑	龒	龓	龔	龕	龖	龗	龘	龙	龚	龛	龜	龝	龞	龟
9F90	9F91	9F92	9F93	9F94	9F95	9F96	9F97	9F98	9F99	9F9A	9F9B	9F9C	9F9D	9F9E	9F9F
龠	龡	龢	龣	龤	龥	*									
9FA0	9FA1	9FA2	9FA3	9FA4	9FA5	9FA6	9FA7	9FA8	9FA9	9FAA	9FAB	9FAC	9FAD	9FAE	9FAF
9FB0	9FB1	9FB2	9FB3	9FB4	9FB5	9FB6	9FB7	9FB8	9FB9	9FBA	9FBB	9FBC	9FBD	9FBE	9FBF
9FC0	9FC1	9FC2	9FC3	9FC4	9FC5	9FC6	9FC7	9FC8	9FC9	9FCA	9FCB	9FCC	9FCD	9FCE	9FCF

中日韩汉字Unicode编码表

2.1.1 常见的媒体元素

- **2016年发布Unicode 9.0.0，CJK统一汉字总数81774**

表 2-17 Unicode 9.0 标准中的汉字[①]

字符集(script)名称	编码范围(十六进制数)	字符数
CJK 统一汉字	4E00～9FD5	20950
CJK 统一汉字扩充 A	3400～4DB5	6582
CJK 统一汉字扩充 B	20000～2A6D6	42711
CJK 统一汉字扩充 C	2A700～2B734	4149
CJK 统一汉字扩充 D	2B740～2B81D	222
CJK 统一汉字扩充 E	2B820～2CEA1	5763
CJK 兼容汉字	F900～FAD9	477
CJK 兼容汉字补充	2F800～2FA1D	542
CJK 部首/康熙字典部首	2F00～2FD5	214
CJK 字根扩展	2E80～2EF3	116
CJK 笔画	31C0～31E3	32
表意文字描述	2FF0～2FFB	12
汉语注音	3105～3120	4
注音扩展	3A10～3ABA	27

2.1.1 常见的媒体元素

- UTF编码
  - 在Unicode标准中，表示字符位置的码点(编号)长度是固定的
  - 为节省文件的存储空间，尤其是以7位ASCII字符为主的西文，并考虑到要与先前的字符编码兼容：

    Unicode定义了1字节(8位)、2字节(16位)和4字节(32位)的三种Unicode转换格式(Unicode Translation Format或Universal Character Set Transformation Format，UTF)，分别称为UTF-8、UTF-16和UTF-32编码，用来表示一个字符的代码

2.1.1 常见的媒体元素

- **UTF-8 (8-bit UCS/Unicode Transformation Format)**
  - 代码长度可变的转换格式，也是ISO/IEC 10646字符的转换格式;
  - 字符编码用由8位构成的1、2、3或4个字节(码元)表示;
  - UTF-8可表示Unicode标准中的任何字符

2.1.1 常见的媒体元素

- UTF-8 (8-bit UCS/Unicode Transformation Format)
  - 1字节码：用于U+0000 ~ U+007F的字符，对ASCII字符集中的128个字符进行编码，可使原来处理ASCII字符的软件无需或只做少量修改就可继续使用
  - 2字节码：用于U+0080 ~ U+07FF的1920个字符，对拉丁文、希腊文、西里尔字母、亚美尼亚语、希伯来文和阿拉伯文等文种的字符进行编码
  - 3字节码：用于U+000800 ~ U+00D7FF和U+00E000 ~ U+00FFFF的61 440个字符，对大部分常用字符，包括CJK统一汉字进行编码
  - 4字节码：用于U+010000 ~ U+10FFFF的1048576个字符进行编码

2.1.1 常见的媒体元素

表 2-18 用 UTF-8 转换 Unicode 字符编码的方法①

Unicode 码点	UTF-8 编码			
	字节 1	字节 2	字节 3	字节 4
0000～007F (128 个代码)	0xxxxxxx (ASCII 字符集，最高位为 0)			
0080～07FF (1920 个代码)	110yyyxx (110 开始，转换码为 2 字节)	10xxxxxx (10 开始)		
0800～FFFF (61440 个代码)	1110yyyy (1110 开始，转换码为 3 字节)	10yyyyxx (10 开始)	10xxxxxx (10 开始)	
10000～10FFFF (1048576 个代码)	11110zzz (11110 开始，转换码为 4 字节)	10zzyyyy (10 开始)	10yyyyxx (10 开始)	10xxxxxx (10 开始)

编码规则：字节1的高位使用固定的0、110、1110或11110，分别表示转换码的长度为1、2、3或4个字节。这样就可根据字节1高位的数值，判断UTF-8转换码的Unicode字符