4.3 Text – Data Formats and Compression

4.3.1 Text Encoding

Coding methods were developed early on for the transmission of text-based messages. Depending on its use, coding can require varying degrees of storage space. If coding is only used to preserve a message, it is then designed to be as space-saving as possible. If, however, other criteria are in the forefront, such as secure transmission or encryption for security reasons, the code used often contains redundant information, which does not contribute to the information content of the message in the true sense of the word. The most common form of displaying a message for transmission – whether it be in direct communication or with the receiver getting a time-delayed message – is a transfer in written form. In our western culture a spelling alphabet is used. Strings of letters are formed, which already contain some redundancy. For example, if one reads the series of letters in the English word "*ecember*", the receiver, who is at the same time an English speaker, is sure that it is the name for the the last month of the year – "December." To transmit messages that are written in an alphabet font with a modern means of communication, a suitable form of coding must be found for the communication medium. One historical example of letter transmission was the code that was devised to express individual letters via the positioning of the signal arm of a semaphore (or "wing telegraph"). The Morse alphabet, developed for simple, electrical or wireless communication, is another example. The character by character encoding of an alphabet is called **encryption**. As a rule, this coding must be reversible – the reverse encoding known as decoding or decryption. Examples for a simple encryption are the international phonetic alphabet (Table 4.1) or Braille script (Fig. 4.3), named after its inventor *Louis Braille* (1809 – 1852).

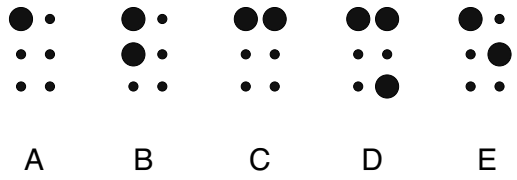
Table 4.1 The International Phonetic Alphabet

Alpha	Bravo	Charlie	Delta	Echo
Foxtrott	Golf	Hotel	India	Juliette
Kilo	Lima	Mika	November	Oscar
Papa	Quebec	Romeo	Sierra	Tango
Uniform	Victor	Whiskey	X-Ray	Yankee
Zulu				

Braille is an example of a code that is based on the binary character set $\{0, 1\}$, with every character represented by a 3×2 matrix of binary characters. The **Morse code** also uses a binary representation for the individual letters. The length of the code for one character depends on the average frequency of its occurrence (see Fig. 4.4).

Louis Braille and the Braille Writing System

Louis Braille (1809 – 1852) lost his eyesight in an accident as a child. He refused to to accept that his only access to literature would be when it was read to him and sought early in life to develop a form of writing that would make it possible for the blind to read and write.



In 1821, he published his easy to learn writing system called Braille. It had been developed from a complex "night writing" system designed for the military by artillery captain *Charles Barbier* (1767 – 1841). Commissioned by Napoleon, Barbier had invented night writing to make it possible for soldiers to communicate with each other without sound or light. However, because of its overly complicated system of dots and syllables the system writing turned out to be unsuitable for military use. Louis Braille simplified this writing by replacing the syllables with letters and reducing the number of dots per symbol from twelve to six. A letter could be easily felt with the tip of the finger, making it unnecessary to move the finger, which made rapid reading possible. Each letter of the writing system developed by Braille consists of six dots arranged in a 3x2 matrix. The encoding of the letters was done by raising certain points in the matrix so they could be felt with the fingertip.

Fig. 4.3 Louis Braille and the Braille writing system.

The first telegraph at the beginning of the 20th century still used the Morse code to transmit letters. But decoding proved to be much easier if every letter was encoded with a binary code word of a constant length. Those codes whose code words are of a constant length are called **block codes**. Frequently used characters are encoded in a block code with a code word of the same length as those code words with rarely used characters. Therefore a certain amount of redundancy is to be expected. This disadvantage is made up for by the simpler mechanical handling of the decoding. The Morse code was soon replaced by the **Baudot code**, a code developed by *Emile Baudot* (1845 – 1903) in 1880. With its 5 bits per character, it can encode two different character sets with altogether over 60 different characters. It became famous as the **International Telegraph Code No.1** (ITC-1, IA-1, CCITT-1). In addition to the 26 letters of the alphabet and the 10 digits, the Baudot code also contains control characters. These serve in formatting the font or in controlling the telegraph [15]. It is only possible to display $2^5 = 32$ characters with 5 bits – which is insufficient when it come to encoding the alphabet and the ten numerical digits. The situation was remedied with a partial double occupancy of the code words. In order to still ensure unique assignment it is possible to switch between the letter and the digit mode by means of special control characters. In 1900 another 5-bit telegraph code was introduced: the so-called **Murray code**. It later became recognized as the **International Telegraph Code No.2** (ITC-2, IA-2, CCITT-2) and is often mistakenly referred to as the Baudot code.

An important step in coding was achieved with the 7-bit telegraphic code, standardized in 1963 by the ANSI (American National Standards Institute) as **ASCII-Code** (**American Standard Code for Information Inter-**

Morse Code and Entropy

Samuel F. B. Morse and Alfred Vail began developing their electrical "writing" telegraph in 1836. The telegraph was based on the principle discovered by Hans Christian Oerstedt (1777 – 1851) in 1821 called electromagnetism – messages are encoded via changing currents, which control an electromagnet on the receiver's side. However, with the technology available at that time it was not possible to print the text that had been received. This led to both inventors finding an alternative method of text encoding. Aided by the electromagnetism of the Morse telegraph, impressions could be raised on a paper strip that was constantly moved by the mechanism of a clockwork under the electromagnet. Morse and Vail used a form of binary encoding, i.e., all text characters were encoded in a series of two basic characters. The two basic characters – a "dot" and a "dash" – were short and long raised impressions marked on a running paper tape. To achieve the most efficient encoding of the transmitted text messages, Morse and Vail implemented their observation that specific letters came up more frequently in the (English) language than others. The obvious conclusion was to select a shorter encoding for frequently used characters and a longer one for letters that are used seldom. For example, "E" – the most frequently used letter in English, is represented by only one basic character, a "dot." On the other hand, the rarely used "Q"is represented by a four character string: "dash dash dot dash." Numbers are shown with a five characters encoding and punctuation with a six character encoding. Word bord

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
-.	---	.--.	--.-	.-.	...	-	..-	...-	.--	-..-	-.--	--..

The selected encoding reflects how often the letters are used and are therefore chosen relative to the **entropy** of the encoded character. For this reason, this variation of encoding is also known as **entropy** or **statistical encoding**.

Fig. 4.4 Morse code and entropy.

change). It is still used up to today as the standard form for representing text information in computers. From the early developmental phase of the first commercial computer systems until the end of the fifties there was no standard character encoding for computers. The computers sold by IBM in 1960 used nine different letter codes alone. But as the idea of networking became more and more a reality there was a huge increase in the demand for a standard letter encoding. In 1961 *Robert Bemer* (1920 – 2004), an IBM employee, proposed ASCII encoding to the ANSI as the standard. It was adopted in 1963. The 7-bit SCII code also became an international standard in 1974, the **ISO I-646**.

However, it took another eighteen years for ASCII to actually be recognized as a general standard. This was due to IBM's newly introduced computer architecture System/360. Still independent and separate from the standardization process of ASCII, it used its own coding called **EBCDIC** (**Extended Binary Coded Decimals Interchange Code**). For reasons of compati-

Fig. 4.5 Extract from the 7 bit ASCII code.

Binary	000	001	010	011	100	101	110	111
0000				0	@	P	’	p
0001			!	1	A	Q	a	q
0010			“	2	B	R	b	r
0011			#	3	C	S	c	s
0100			$	4	D	T	d	t
0101			%	5	E	U	e	u
0110			&	6	F	V	f	v
0111				7	G	W	g	w
1000			(	8	H	X	h	x
1001			)	9	I	Y	i	y
1010			*	:	J	Z	j	z
1011			+	;	K	[	k	{
1100			,	<	L	\	l	|
1101			-	=	M	]	m	}
1110			.	>	N		n	~
1111			/	?	O	_	o	

bility, the succeeding generations of IBM System/360 continued to use the EBCDIC encoding. EBCDIC is an 8-bit encoding, an expansion of the 6-bit BCD used previously by IBM. Successive characters in the alphabet were not necessarily provided with consecutive codes since this type of encoding was still being influenced by Hollerith’s punched cards. There are different variations of EBCDIC that are incompatible with each other. Most of the time, the American variation uses the same characters as the ASCII code, yet some special characters are not included in the respective other code. IBM designed a total of 57 different, national EBCDIC codes, each containing country-specific special characters and letters. It was not until 1981 that IBM changed to the ASCII code, while in the process of developing its first personal computer.

Also the 7bits of the original ASCII encoding proved insufficient in representing all of the international character sets with their associated special characters. By adding an eighth bit some manufacturers introduced their own proprietary encoding to allow for a display of diverse special characters. But a unified standard for different international characters based on an 8 bit ASCII encoding could first be achieved with **ISO/IEC 8859** encoding. With the ISO/IEC 8859-x standard, the first 7 bits are assigned the original encoding of the 7-bit ASCII. This ensures compatibility with the previous encoding. Various national extensions of the ASCII codes are implemented via the 8th bit. Altogether there are some 15 different national standards for the 8-bit ASCII code, ISO/IEC 8859-1 through ISO/IEC 8859-16, whereby ISO/IEC 8859-12, for characters of the Indian Devanagari language, was dropped in 1997. ISO/IEC 8859-1 (Latin 1) includes special national special characters for the regions of Western Europe, America, Australia and Africa, ISO/IEC

8859-2 (Latin 2) supplements the basic characters of the 7-bit ASCII codes with other Central European special characters. ISO/IEC 8859-5 contains Cyrillic, ISO/IEC 8859-6 Arabic and ISO/IEC 8859-8 Hebrew special characters. Hebrew like Arabic has a reversed textual direction from European writing systems. For this reason, two special characters with zero width were adopted at the positions 253, or 254 respectively. These characters change the text direction from left to right or from right to left.

With Asian sets of characters such as Chinese, and its more than 10,000 ideographic characters, or Korean or Indian an 8-bit encoding is not sufficient to represent all of the characters. Moreover, the concept of what constitutes a character or letter as a "basic" unit for the encoding of a text is not the same in all languages and writing systems. In some languages, individual characters can consist of a further series of individual characters. These have a different meaning when in a group and can also change their external form when they are combined.

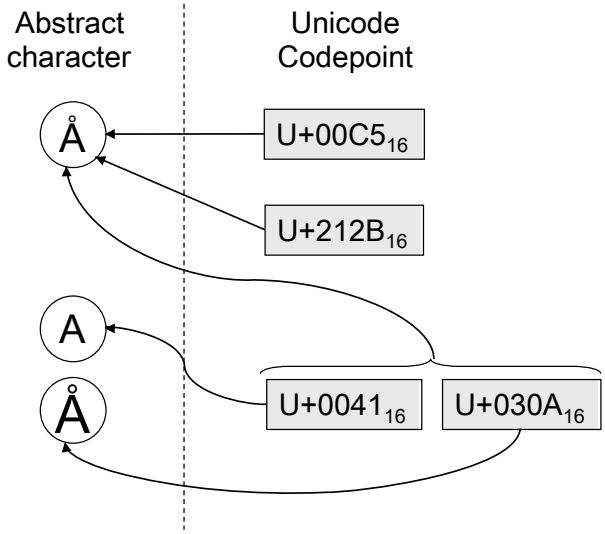
The Korean Hangul writing system combines symbols that in Korean each stand for an individual sound and in quadratic blocks each represent an individual syllable. Depending on the user as well as on the intended application individual symbols as well as syllable blocks can be represented as "characters." In Indian systems of writing every character that stands for a consonant hides an inherent vowel, which in different ways is either eliminated or replaced when individual characters are combined together into blocks. Also here, independent of user or application, individual consonants and vowels or entire consonant vowel blocks can be seen as "characters." Therefore it was necessary to develop an encoding system that could fulfill the challenges of different international writing systems in an appropriate way. The **Unicode** was designed to ensure that such a standard of encoding could be implemented for nearly all existing alphabets. Unicode encoding was introduced in 1991 and with the norm **ISO/IEC 10646 ISO/IEC 10646** and adopted as the international standard **Universal Character Set** (UCS). The nonprofit Unicode Consortium, founded in 1991, was responsible for the Unicode industry standard.

Unicode first used 16 bits for encoding multilingual characters. This was later expanded to 21 bits[1] and also included codes for Indian, Chinese, Japanese and Korean characters for which a huge supply of 2^{21} characters were available. This was to ensure that Unicode encoding in fact fulfilled the desired requirements of universality (i.e,. for every existing writing system there should be an encoding possibility) and extensibility. In addition to national alphabets of a variety of countries, additional typographic symbols and special national characters are also included.

[1] In the UTF-32 transformation, 21-bit Unicode characters are encoded with a full 32 bits.

Excursus 5: The Unicode Standard

The Unicode standard assigns a number(**code point**) and a name to each character, instead of the usual glyph.[2] It represents each individual character in an abstract manner, while the visual representation of the character is left up to the text-displaying software, e.g., the web browser. This makes sense since the graphical representation of the character may vary greatly depending on the font type chosen. Characters can be assigned to several different code points since the same characters often belongs to different writing systems.



In Unicode, a code point defines a specific character. However, it is entirely possible that a character is used in different writing systems. Structure and organization of the Unicode arranges individual writing systems within each contiguous block of codes. Some characters are assigned several code points. It is also possible for one character to be composed of several basic characters that exist separately.

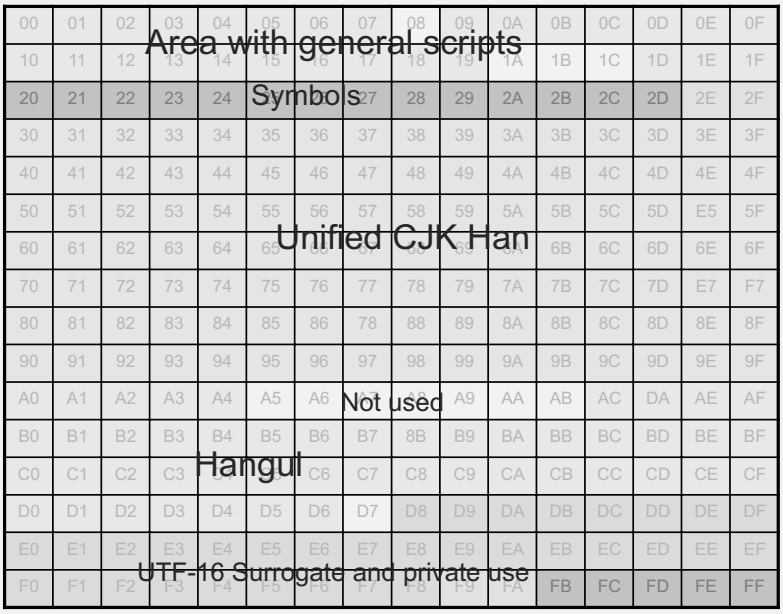
Organization of Unicode Encoding

In Unicode encoding, the first 256 characters of the Unicode are assigned characters of the ISO/IEC 8859-1 code to ensure a compatibility between the old 8-bit ASCII encoding and Unicode. Unicode characters are usually displayed in the form `U+xxxxxxxx`, whereby `xxxxxxxx` stands for a code point in hexadecimal format. Leading zeros can be omitted. The code space provided in Unicode is divided into individual planes. Each of them contains 2^{16}=65.536 code points. Of these planes, currently 17 are available for use (as a result the character space, encodable by means of Unicode, is limited to $17 \cdot 2^{16}$=1.114.112 characters), with only planes 0–1 and 14–16 in use.

Plane	Title	from	to
0	Basic Multilingual Plane (BMP)	U+0000$_{16}$	U+FFFF$_{16}$
1	Supplementary Multilingual Plane (SMP)	U+10000$_{16}$	U+1FFFF$_{16}$
2	Supplementary Ideographic Plane (SIP)	U+20000$_{16}$	U+2FFFF$_{16}$
14	Supplementary Special-purpose Plane (SSP)	U+E0000$_{16}$	U+EFFFF$_{16}$
15	Supplementary Private Use Area-A	U+F0000$_{16}$	U+FFFFF$_{16}$
16	Supplementary Private Use Area-B	U+100000$_{16}$	U+10FFFF$_{16}$

[2] In typography, a character signifies the abstract idea of a letter, whereas a glyph is its concrete, graphic display.

The first level – plane 0 with the code points 0 – 65.535 – is called the **Basic Multilingual Plane** (BMP) and includes almost all spoken languages.



The second level (plane 1), the **Supplementary Multilingual Plane** (SMP), contains rarely used and mostly historical writing systems, such as the writing system of the Old Italian language, or a precursor to Greek, the Cretan writing systems – Linear A and Linear B.
The next level or plane (plane 2), called the **Supplementary Ideographic Plane** (SIP), contains additional, rarely used, ideographic characters from the "CJK" group of Chinese, Japanese and Korean characters that are not classified in the BMP. Level 14, the **Supplementary Special-purpose Plane** (SSP), contains additional command and control characters that were not assigned on the BMP. Planes 15 and 16 accommodate privately used characters.
Within the planes, code points that belong together are grouped into blocks. Principally a Unicode block contains a complete writing system. However, a certain degree of fragmentation may be observed. In many cases new characters were later added to an already completed block and then had to be accommodated elsewhere.

Unicode UTF Transformations

Unicode codepoints can be encoded in different ways via so-called „**Unicode Transformation Formats**" (UTF). As the writing systems most often used are located within plane 0, omitting lead zeros in encoding for reasons of efficiency seems an obvious choice. The various UTF Transformations (UTF-7, UTF-8, UTF-16 or UTF-32) were developed to enable the efficient encoding of Unicode code points.
UTF-8 is the most well-known variety and implements an encoding of variable length from 1–4 bytes (based on the UTF-8 procedure, strings of up to 7 bytes long can be theoretically generated, however because of the limitation of Unicode code space, the maximum length allowance is 4 bytes). The first 128 bits of the Unicode, encompassing the characters of the 7-bit ASCII code, are only represented by one byte. The byte order of every UTF-8 encoded character starts with a preamble that encodes the length of the byte order. To ensure maximum compatibility with ASCII encoding, the 128 characters of the 7-bit ASCII code are assigned the preamble "0." If a UTF-8 encoded character consists of several bytes, the start byte always begins with a "1" and every succeeding byte with the preamble "10". With multi-byte characters, the quantity of of 1-bits in the preamble of the start bytes, gives the byte length of the entire UTF-8 encoded character. The resulting encoding scheme is therefore:

1 byte	0xxxxxxx	(7 bit)
2 bytes	110xxxxx 10xxxxxx	(11 bit)
3 bytes	1110xxxx 10xxxxxx 10xxxxxx	(16 bit)
4 bytes	1111xxxx 10xxxxxx 10xxxxxx 10xxxxxx	(21 bit)

The shortest possible encoding variation is chosen for the UTF8-encoding of a code point. The Unicode code point is always entered in the encoding scheme right-justified. The following examples illustrate the principle of UTF-8 encoding:

Character	Codepoint	Unicode binary	UTF-8
y	U+0079_{16}	00000000 01111001	**0**1111001
ä	U+$00E4_{16}$	00000000 11100100	**110**00011 **10**100100
€	U+$20AC_{16}$	00100000 10101100	**1110**0010 **10**000010 **10**101100

For all scripts based on the Latin alphabet, UTF-8 is the most space-saving method for the mapping of Unicode characters. A further variation is **UTF-16** encoding, which allocates every Unicode code point a 2 – 4-byte long bit sequence. UTF-16 is especially designed for the encoding of BMP characters. It is superior to UTF-8 encoding for texts in Chinese, Japanese or Hindi. Whereas these BMP characters are encoded in a 3-byte long bit sequence with UTF-8, a corresponding UTF-16 encoding is comprised of just 2 bytes. **UTF-32**, in contrast, assigns every Unicode codepoint a bit sequence with a constant length of 4 bytes. It therefore represents the simplest of all encoding variations as the Unicode code point can be directly translated into a 32-bit binary number. However, in regards to the characters from the BMP, UTF-32 is quite inefficient. Further encoding variations exist. Of these, the **UTF-7** encoding should be mentioned here. It was originally intended for use in communication protocols based on a 7-bit transmission standard, for example for the SMTP protocol for emails. However in email communication Base64 encoding of the MIME standard finally prevailed over UTF-7.

Further reading:

The Unicode Consortium: The Unicode Standard, Version 5.0, Addison-Wesley Professional, 5th ed. (2006)

Unicode has also been introduced in the WWW. In RFC 2070, the WWW language HTML was prepared for Unicode support. RFC 2077 also recommends the support of ISO 10646 for all new Internet protocols.

4.3.2 Text Compression

The available bandwidth of the implemented communication medium limits the amount of data transmitted. Methods were therefore developed early on to minimize the redundancy contained in a message and to use the available bandwidth as efficiently as possible. The techniques that came to be used for this purpose are called **compression** (compaction). They are commonly

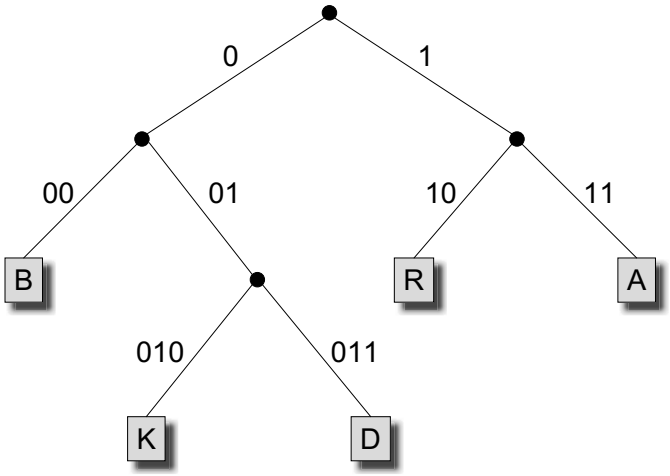


Fig. 4.6 Prefix coding in a binary tree representation.

$$B(T) = \sum_{c \in A} f(c) d_T(c).$$

The procedure developed by Huffman for the construction of an optimal prefix code works in the so-called bottom-up manner, i.e. it begins from below with a number of $|A|$ (unrelated) leaf nodes and leads to a a range of $|A|$-1 merger operations to construct a result tree. Besides bearing the letters $c \in A$, the leaf nodes also represent its frequency $f(c)$ within the encoded file. Next, the two nodes c_1 and c_2, which contain the smallest frequency data are selected. A new node c_{neu} is generated, which is marked with the sum from the two frequencies $f(c_{neu})=f(c_1)+f(c_2)$ and connected to the two nodes selected as its successor. The nodes c_1 and c_2 are taken out of the amount A, while the new node c_{neu} is added to it. By proceeding in the same way, increasingly large subtrees are generated and number of the nodes in A becomes smaller and smaller. At the end all of the nodes are connected into one single tree. Nodes with a lower frequency are then the farthest away from the root node, i.e., they are also allocated the longest used code word, while codes with greater frequency are near the root nodes and have accordingly short code words. A tree generated in this way is a direct result of the Huffman code (see Fig. 4.7).

With the help of induction it can be shown that in fact the Huffmann method generates an optimal prefix code.

Further reading

Huffman, D. A.: A Method for the Construction of Minimum-Redundancy Codes, in Proc. of the IRE, 40(9), pp. 1098-1101 (1952)

Cormen, T. H., Leiserson, C. E., Rivest, R. L.: Introduction to Algorithms, MIT Press, Cambridge MA, USA (1996)

4.4 Graphics – Data Formats and Compression

Graphics represented and processed in a computer are traditionally prepared in the form of vector graphics or bitmap graphics (also referred to as raster graphics). With **vector graphics**, lines, polygons or curves are defined by the specification of key points. A program reconstructs the geometric figure to be displayed from these key points. Additionally, the key points have specific

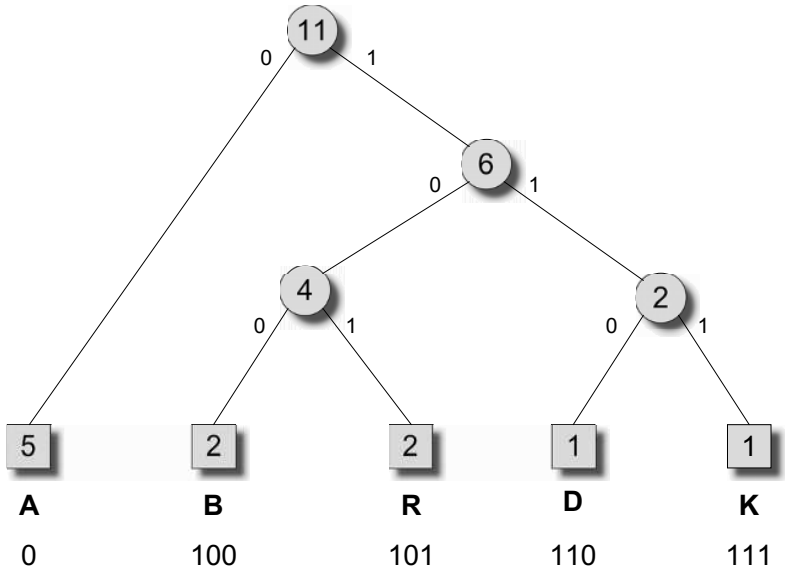


Fig. 4.7 Huffman coding represented by a binary tree.

attribute information, such as color or line thickness. The suitable program, aided by an output device, reconstructs the image to be displayed from these basic graphic elements. In a historical context, vector graphics were developed together with **plotters** – a graphic output device for computers. Plotters print by moving one or more pens across a drawing surface based on given coordinate values. An advantage of vector graphics is that the image to be represented may be scaled as desired without any quality-reduction effects.
Bitmap graphics consist of a set of numeric values, color and brightness information of individual pixels or whole picture elements. Encoding such an image means that it has to be put into the form of a grid and spatially discretized (screened), with each pixel assigned a color or brightness value. Historically, bitmap or raster graphics were associated with the development of the **cathode ray tube** (CRT) as a graphical output device. In order to display an image made up of pixels on this type of a screen the pixels are illuminated in a certain color and brightness. Raster graphics are suitable for displaying more complex images such as photographs, which cannot be created with vector graphics.
One of the disadvantages of bitmap graphics, as compared to vector graphics, is the relatively high memory consumption that normally occurs. Because bitmap graphics usually consist of a limited number of pixels, two-dimensional geometric forms can only be approximated. It is also possible that the so-called „alias effect“ can occur, i.e., rounded lines are approximated with step-like pixel sequences. In addition, information gets lost in geometric transformation, e.g., in an enlargement (scaling) or rotation of an image section. It can happen that so-called “artifacts“ appear, i.e., through the transformation of the pixels color tones can be created which were previously not there, reducing the quality of the transformed image. Although vector graphics are

also used today in the WWW, in this chapter we will limit ourselves to the most important data formats from the area of bitmap graphics.

In the efficient storage of graphic data, the following properties must be taken into account when characterizing a graph:

- **Picture Resolution**
  The picture resolution is determined by the number of pixels along the x axis and the y axis.
- **Color Depth**
  The color depth determines the number of colors a pixel has. It is given as logarithm $\log(c)$ of the actual number of possible colors c (e.g., color depth 8 corresponds to $2^8 = 256$ colors), i.e., the number of bits that are necessary to describe or specify a color uniquely. Starting at 24 bits we speak of a display of true color. Modern image processing systems even allow a 32 bit or 48 bit deep color representation.
- **Palette Size**
  Some graphics systems limit the number of available bits for specifying colors. From the outset, there is a fixed color palette with a reduced number of colors from which the image can be made. The palette colors are selected in such a way that they come as close as possible to the original colors of the image to be coded.
- **Picture Resolution (Density)**
  The picture resolution is given as the density of the individual pixels per length unit. The common unit of measure is adopted from the American system: **dpi** (dots per inch), in other words, the number of pixels per 2.54 cm. Image resolution for computer screens is about 100 dpi and more, with resolutions greater than 300 dpi normal in the area of printing. The higher the picture resolution chosen, the more detailed the display of the image, but at the same time the more storage space is needed
- **Aspect Ratio**
  The aspect ratio of an image describes the relation of the image length to the image width. It also distinguishes the aspect ratio of a single pixel (pixel aspect ratio), which also affects the aspect ratio of the total image and how it is perceived.

In a bitmap graphic, in the simplest case, the individual pixels are stored next to each other in a line and the lines sequentially in rows. Based on the depth of color to be displayed, a varying degree of storage space is needed for each pixel. While in a monochromatic image a single bit is sufficient for a single pixel, in a so-called true color display at least 24 bits are required per pixel.

Color itself is not a property of the physical world, but rather a sensory perception. The perception of color enables us to distinguish between two featureless surfaces of equal brightness. In the way it is perceived by humans, color is nothing more than light made up of multiple wavelengths.

The white light that we know is simply light made up of different frequencies whose wave lengths are in the realm of human perception at between approximately 380 nm and 780 nm (1 nm = 10^{-9}m). Ideally, colored light is generated from a radiating black body (black-body radiation). Depending on the temperature of the radiating body, the spectral composition of light differs. Therefore, different light sources are assigned **color temperatures** that correspond to the temperature of an ideal black body radiating a corresponding light spectrum. The color temperature of a standard 60 watt light bulb is 2,200° Kelvin, while the color temperature of sunlight is about 5,500° Kelvin. Low color temperatures appear to us reddish and warm while higher color temperature are perceived as bluish and cold.

Color first comes into being when white light is broken down into individual components of fixed or a similar frequency. Color is also created when, in the process of being reflected or scattered on a white light surface, certain frequencies receive preference while others are suppressed, absorbed or filtered through colored, transparent bodies. In the first case, colored light from a body (sun, light bulb) is radiated. This color is therefore also referred to as **light color**. In the second variation, color generated through reflection, absorption, scattering or filtering is called **body color**. The human eye can however, only perceive a limited number of colors in each case. At the same time, up to 10,000 colors can be distinguished simultaneously. All in all, the human eye is capable of perceiving up to 7 million different color valences, with approximately 500 different levels of brightness and about 200 different shades of color. However, this depends in each case on different parameters, e.g., background illumination and the size of the field of brightness. With optimal conditions, the value of the perceived level of brightness can increases up to 1,000. The maximum sensitivity of the human eye is dependent on the wave length and the light intensity. In daylight, the maximum sensitivity is 554 nm and with the adaption of the eyes to night conditions shifts to 513 nm. There are different mathematical **color models** for the display of color on a computer. A systematic arrangement of color had already been done by Aristotle , who ordered colors on a palette between black and white. Many scientists and artists have attempted to achieve a systematization of color throughout the centuries. The color systems developed all have the goal of arranging colors in a way that they can be described by a geometric arrangement or provide a guidelines for the mixing of new colors.

A fundamental distinction is made between additive and subtractive color models. In an **additive** color model, colors are mixed with the base color black to create new colors. The more colors are added, the more the color mixture tends toward white. Each of the colors are self-illuminating in an additive color model. A typical example of this is the raster image of a television or computer screen. The image is made up of many small dots, thus the three primary colors of the luminous red, green and blue pixels are "added" together. At a sufficient distance to one another, the adjacent red, green and blue pixels are seen by the eye as mixed and correspondingly create the

perception of color. A **subtractive color model** works in the opposite way. Fundamentally speaking, individual colors are subtracted from the base color white to create new colors. The more colors are removed, the more the color mixture tends toward black. Viewed in another way, black in a subtractive color model represents the complete absorption of the light of the color pigment. Subtractive color models are based on reflection and absorption. The colors we perceive result from the reflection of light from an external light source, such as the printed colors on a piece of paper. In a color model, colors may be subdivided into **primary colors**, **secondary colors** and **tertiary colors**, according to the degree of the mixing of the primary colors involved. The basic colors of the color model are the primary colors. If two primary colors are mixed together in equal parts the results is a secondary color. The secondary color yellow is created by mixing the primary colors red and green in the additive RGB color model. If a primary color and a secondary color are mixed together, the result is a tertiary color. The most common color models – the RGB color model, the CMY(K) color model, the HUV color model and the YUV color model – are introduced in detail in Excursus 7.

Excursus 7: What is Color? – Color and Color Systems

The ability to classify natural colors in a system and the study of such systems goes back to ancient times. Already *Aristotle* (384 – 322 BC) presents an arrangement of colors in his work entitled "*De sensu et sensato*" (about the senses). He orders colors in a strand from black to white and assigns them to the time of day. In the Middle Ages the very existence of color was debated by scholars. The Persian doctor and physicist, *Avicenna* (980 – 1037) argued about whether colors were present in the dark. Without light, colors lacked "verum esse" – their very existence. His opponent, *Alhazen* (and Ibn al-Haitham, 965 – 1040) countered with the argument that colors were still present when it was dark even though they were no longer visible to the eye. During the Middle Ages in Europe, philosopher *Roger Bacon* (1214 – 1294) addressed the question explaining that light and color only appear in combination with each other: "*Lux ... non venit sine colore*".

Many scientists and artists have since then tried to arrange colors into a color system based on specific and different objectives. While physicists see no more than the different wave lengths of light in color, a painter sees the mixture of color on the palette in terms of physiology and the affect color has on people. The purpose of a color system is to arrange colors in such a way that a guide to color mixing can be obtained from this geometry.

The knowledge that colors are nothing more than components of white light was achieved through experiments made with a **glass prism**. This insight was documented for the first time by Bohemian physicist *Marcus Marci* (1595 – 1667) in 1648 in his writings"*Thaumantias liber de arcu caelesti*" (1648) and "*Dissertatio de natura iridis*" (1650). Colors with different wave lengths are broken at different angles on a prism (**chromatic Aberration**). The colors produced by a prism cannot be further split up. Building on this theory, the English physicist *Isaac Newton* experimented with prisms and in 1672 published his results. This later became the basis for his opus, "*Opticks, or A treatise of the reflections, refractions, inflections and colours of light*" (1704). It was Newton who coined the famous phrase: "*The rays are not coloured.*" A different impression of color was instead created – one dependent on the frequency of perceived light. Long-wavelength light corresponds to the color red, while short-wavelength light correspond to the color purple. In between are the spectral colors: orange, yellow, green and blue (in that order). Light is referred to as **monochromatic light**,

if it has only a single wavelength. Newton bent the expanse of generated spectral colors into a circle subdivided into seven sectors – red, orange, yellow, green, cyan blue, ultramarine and violet blue. In the middle of this **color circle** he put white as the one color made up of all the others. He refrained from arranging colors according to their degree of brightness, i.e., from light to dark, as had been usual up to that time. In a further step, Newton placed purple (magenta) between its bordering colors in the natural spectrum – red and violet. Purple, which results from a mixture of red and violet, does not occur in the spectral decomposition of white light.

But the concept behind Newton's color theory was slow to gain interest. One hundred years later, the poet, scientist and art theorist, *Johann Wolfgang von Goethe* (1749 – 1832) disputed Newton's color circle. Newton suspected that light was composed of corpuscles, that is, small particles of different sizes. In contrast to Newton, Goethe tried to show in his *color theory*, which he incidentally considered the most important work he had written, that white light is not "assembled" and that colors result from an interchange of light and dark. Goethe explained the color separation in a prism due to an"overlapping" of light and dark, which results in a yellow and a blue border. Depending on the respective portion of light and dark, these borders become mixed as green or red – thereby creating the spectral colors. Goethe's color theory does not focus on physical color separation, but rather on the "sensory and moral effects" of color. His observations and methods regarding the effect of color are considered the beginning of modern color psychology. Goethe discovered the phenomenon of subjective color and basic principles of color vision, such as the afterimage effect and simultaneous contrast.

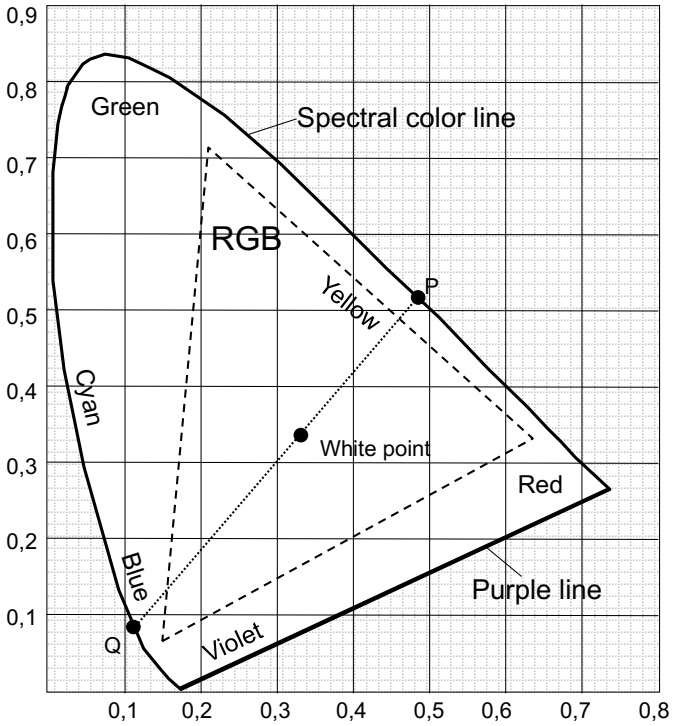
In 1802, English doctor and physicist *Thomas Young* (1773 – 1829) postulated his **trichomatic theory** (three-color vision). Young proposed that the human retina is only capable of perceiving three different base colors (based on different types of receptors). Young's trichromatic theory gained credibility in 1855 when a statistical analysis of color blindness was presented for the first time. It was shown that the recorded observations could only be explained if the assumption was made that one or two receptor types had failed in the people affected.

Then in 1859 the Scottish physician *James Clerk Maxwell* showed that in fact all colors in the spectrum could be created through a mixture of three components, provided that added together these components produced white – i.e., are located far away enough from each other in the spectrum (e.g. red, green and blue). He represented the corresponding combinations within a triangle whose corners were marked by the three primary spectral colors: red, green and blue. Each compound color was thereby in the focal point of the line that connected the primary colors to be combined. With his "theory of color vision" Maxwell was responsible for the origin and start of modern, quantitative color measurement (colorimetry).

The first person to bring attention to the difference between additive and subtractive color mixing was the German physicist *Hermann von Helmholtz* (1821 – 1894). In his handbook "Handbook on Physiological Optics" (1867), he presented the **Helmholtz coordinates**. These coordinates, named after him, were: brightness, hue and saturation. Helmholtz developed Young's trichromatic theory further in 1850 into what is known today as the Young-Helmholtz theory.

The color system presented by American painter *Albert Henri Munsell* (1858 – 1918) also met with great success. In his 1915 **color atlas** colors were categorized in relation to how they were visually perceived. He grouped all colors in three dimensions around an axis extending from black to white luminance (value). It was done in such a way as that opposite color shades (hue) mix to gray. The saturation of the respective color (chroma) is represented by the distance of the color to the central axis. He also considered the different degrees of brightness of pure spectral colors; for example that the spectral color yellow appears subjectively brighter than the spectral colors blue or red

A first truly objective assessment of color was made possible by the **Color Standard Table**, which was established by the International Commission on Illumination (Commission Internationale d'Eclairage, CIE) in 1931. The Color Standard Table was determined with the help of subjective test persons. Test subjects were required to mix the three primary colors of monochromatic light until a prescribed spectral color was achieved. A numerical value resulted for each primary color. Each given color could be described unambiguously with the three determined numerical values. Through suitable transformation and scaling these three coordinates could be mapped in a two-dimensional coordinate system. Because of the fundamental condition $x + y + z = 1$ the z-share ($z = 1 - x - y$) can be determined easily.



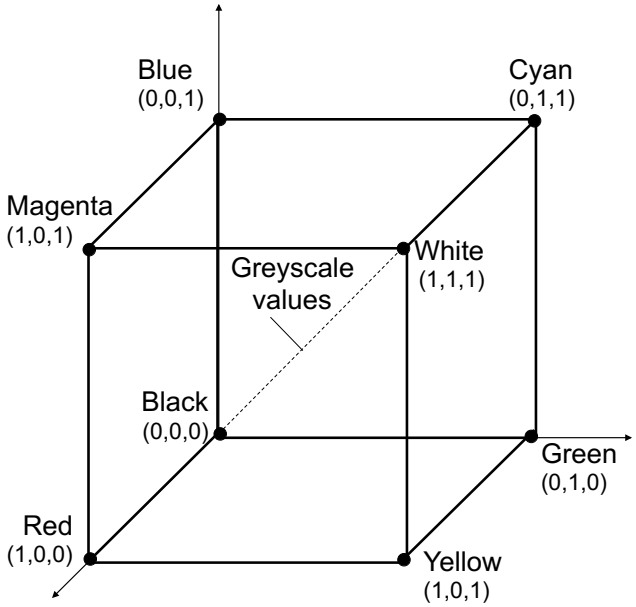
Within this coordinate system all of the colors perceived by those with normal vision lie within a horseshoe-formation whose upper border is delineated by a pure **spectral color**. The lower border is formed by the so-called **purple line** – an imaginary demarcation line between the two ends of the spectral line. The purple line does not contain any spectral colors, but only colors that can be obtained by mixing two spectral colors. The **white point** is on the inside. Proceeding from this white point, all colors perceived as the same shade can be read on one line from points P and Q on the edge of the spectral colors. On this line opposite points P and Q are **complementary colors**. The RGB color model defines the colors in the CIE Color Standard Table formed in the triangle of the primary colors red, green and blue.

The German Industrial Standard DIN 5033 defines color as "the sensation of a part of the visual field which the eye perceives as having no structure and by which this part can be distinguished from another structureless and adjoining region when viewed with just one motionless eye."

RGB Color Model (Red-Green-Blue)

RGB represents today the most widely used color model for graphic formats. It is an additive color mixing system in which respectively changing shares of the primary colors red (R, wavelength $\lambda = 700$ nm), green (G, $\lambda = 546{,}1$ nm) or blue (B, $\lambda = 435{,}8$ nm) are additively mixed to the initial black to generate new colors. G and B are lines of the mercury spectrum,

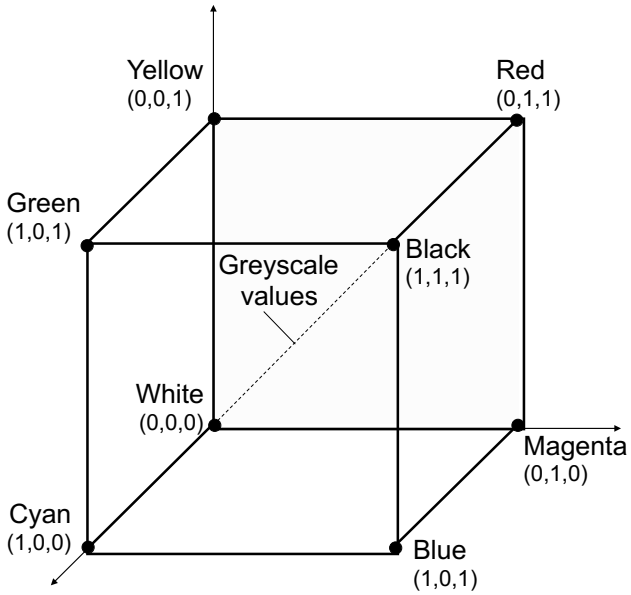
while R represents the long-wavelength end of the visible light. These three components may be regarded as linear-independent vectors spanning a three-dimensional color space. This space is illustrated by the RGB color cube.



The plane spanned by the vertices of the color cube color space is also rferred to as **Gamut** The graphics data formats use for representing a pixel in the RGB color system a color triplet of the numerical values (r,g,b). It establishes the respective color portions of the primary colors in pixels. In a 24bit true color display, for example, the triplet (0,0,0) represents the color black and (255,255,255) the color white. If all three RGB parts have the same numerical value, therefore, e.g., (66,66,66) – they are located diagonally in the RGB cube and the resulting color always results in a specific gray.

CMY (Cyan-Magenta-Yellow)

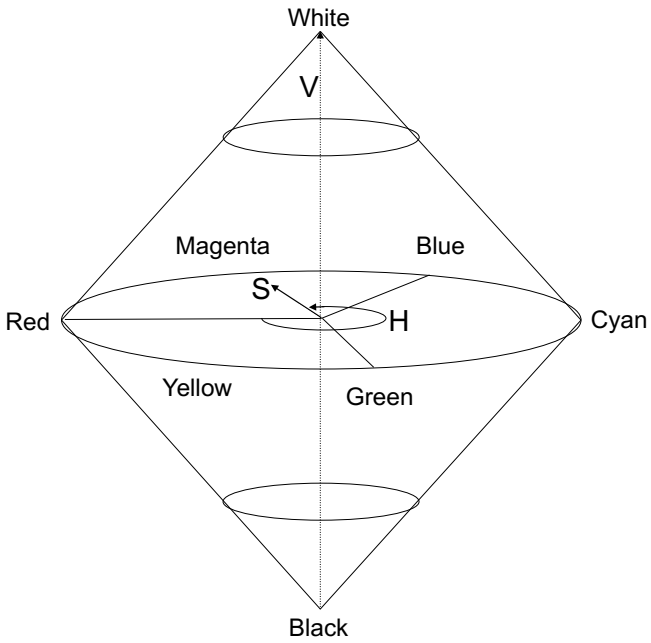
CMY is a subtractive color model that is used by printers and in photography and is based on a white surface. Almost all devices using the principle of applying color pigments to a white surface use the CMY method. If the primary surface is illuminated then each of the three primary colors used absorbs proportionally the incident light of the complementary color assigned to it: **cyan** (a greenish blue) absorbs red, **magenta** (a light violet red) absorbs green and **yellow** absorbs blue. By increasing the yellow value e.g., the portion of blue perceptible in the image lessens. If all color portions are absorbed by the incident light through a mixture of all colors, then black results.



Since, practically speaking, no perfect black can be achieved through mixing the three primary colors, the extended CMYK has prevailed. K stands for the *key color*, black. The term "key" is used instead of "black" to avoid misunderstandings concerning the letter "b," which is used for "blue." The letter "k" stands for the "key plate" used in offset printing. This is the black printing plate on whose baselines the three colored printing plates are aligned with the primary colors. In the CMYK color model, back does not serve in coloring but rather in darkening the three primary colors. CMY colors are specified as numeric triplet (CMYK colors as quadruples). Thus, in a 24-bit true color system, the CMY triplet (255,255,255) represents the color black and (0,0,0) the color white. But in many color mixing systems often only percentages are given for the proportion of the primary colors used. These are between 0% and 100%.

HSV Color Model (Hue-Saturation-Value)

The HSV color system represents a color systems where in the creation of new colors color properties vary rahter than being mixed. Thereby, **hue** determines the color tone in the literal sense, such as red, , orange, blue, etc. The specification of the hue appears as a color angle on the color wheel (e.g., 0° = red, 120° = green, 240° = blue). **Saturation** specifies the amount of the white in the selected hue. A completely saturated hue, i.e., saturation 100%, contains no white and appears as a pure hue. But if, for example, red is chosen and a saturation of 50%, the resulting color is pink. Finally, **value** indicates the degree of self-luminosity of a hue, i.e., how much light the hue emits. A hue with a high self-luminosity appears light, while a hue with low self-luminosity appears dark.



HSV closely resembles the color mixing system used by painters to achieve various degrees of color by mixing white, black or gray with a pure hue. There are a number of other, very similar, color models that change a color (hue) through the variation of two other properties. For example:

- **HSL** – Hue, Saturation, and Lightness (relative brightness),
- **HSI** – Hue, Saturation, and Intensity (light intensity),
- **HSB** – Hue, Saturation, and Brightness (absolute brightness).

Although there is a clear separation between brightness and color in this family of color systems, they have gained only minimal significance in graphic encoding and graphic compression. One reason for this is their discontinuity in the realm of color display. The color values of the angle 0° and 359° have an almost identical hue, yet their representation as a numerical value is vastly different. If, for example, both colors are to be mixed in a lossless compression, the result would be a color on the opposite side of the color wheel ((0+359)/2≈180), and the display would be strongly distorted.

YUV Color Model (Y-Signal, U-Signal, and V-Signal)

The YUV color model belongs to a family of color models distinguished from other models in its separation of image brightness from color difference. With a simple transformation the RGB components of a color image can be easily converted into its corresponding YUV counterpart. Historically, YUV color models are closely connected with the development of color television. In the transition from the black and white television to the color television it was necessary for reasons of compatibility to find a way to make it possible to continue to use the old black and white receivers as well as to allow the additional transfer of color television's color components. The brightness (luminance) (Y components) were therefore, separated from the color components (chrominance) (U and V components). The human eye has varying degrees of sensitivity in regards to brightness and color resolution. Therefore, separating the components enabled a simultaneous adjustment of the resolution of the components to human perception.

The conversion from the RGB to the YUV color model is carried out via:

$$(\mathsf{Y},\mathsf{U},\mathsf{V}) = (\mathsf{R},\mathsf{G},\mathsf{B}) \cdot \begin{pmatrix} 0,299 & -0,168736 & 0,5 \\ 0,587 & -0,331264 & -0,418688 \\ 0,114 & 0,5 & -0,081312 \end{pmatrix}$$

Within the family of these color models there are three basic models:

- **YUV** – This model is used in the PAL television standard.
- **YIQ** – This model is used in the competing NTSC color television system mainly in North America and Japan. The only difference to the YUV model is a shift in the chrominance by 33°.
- **YCbCr** – This model was derived from the YUV model especially for digital television. It distinguishes itself from the YUV model through a scaling of the individual components and an offset for chrominance.

Further reading:

Falk, D., Brill, D., Stork, D.: Seeing the Light: Optics in Nature, Photography, Color, Vision and Holography, John Wiley & Sons, New York, USA (1986)

4.4.1 Variants of Run Length Encoding for Graphics Data

Image or graphics data is normally compressed in a sequential process. To do this, the 2-dimensional image is disassembled into a 1-dimensional data stream consisting of the color information contained in the individual pixels. This process can be carried out line by line from top left to bottom right (X-axis encoding), column by column in the same order (Y-axis encoding) or even diagonally in an alternating direction (zig-zag encoding). **Run length encoding** (RLE, see Excursus 6) for graphic files are as a rule lossless. In essence, the procedure corresponds to the previously presented method for text files. Irrespective of the chosen color model, the color values of the individual pixels are indicated by several numerical values. These numerical values can take the form of a binary number and thus be presented in a continuous form as a single, long bit string. The connected groups of zeros and ones can be summarized as discussed previously. The longer the connected groups, the higher the degree of compression. A contiguous group of identical bits can only be summarized in logarithmic space, based on the original storage size. This type of encoding is known as **bit-level run length encoding**. In contrast, **byte-level run length encoding** considers the identical byte values of image information to encode and does not take into account individual bits or bit groups. The most common methods are those which encode the contiguous groups of identical bytes into a 2-byte packet, whereby the first byte indicates the number of repetitions and the second byte the relevant byte value.

On a higher level of abstraction the so-called **pixel-level run length encoding** begins. It is implemented if two or more bytes are used to store the color values of a pixel. Special tricks are used to further increase the efficiency of this method. If a complete line in an image file repeats, it is then sufficient