Video Codec Design
Iain E. G. Richardson

ISBNs: 0-471-48553-5 (Hardback); 0-470-84783-2 (Electronic)

2
Digital Video

2.1 INTRODUCTION

Digital video is now an integral part of many aspects of business, education and entertainment, from digital TV to web-based video news. Before examining methods for compressing and transporting digital video, it is necessary to establish the concepts and terminology relating to video in the digital domain. *Digital video* is visual information represented in a discrete form, suitable for digital electronic storage and/or transmission. In this chapter we describe and define the concept of digital video: essentially a sampled two-dimensional (2-D) version of a continuous three-dimensional (3-D) scene. Dealing with colour video requires us to choose a colour space (a system for representing colour) and we discuss two widely used colour spaces, RGB and YCrCb. The goal of a video coding system is to support video communications with an 'acceptable' visual quality: this depends on the viewer's perception of visual information, which in turn is governed by the behaviour of the human visual system. Measuring and quantifying visual quality is a difficult problem and we describe some alternative approaches, from time-consuming subjective tests to automatic objective tests (with varying degrees of accuracy).

2.2 CONCEPTS, CAPTURE AND DISPLAY

2.2.1 The Video Image

A video image is a projection of a 3-D scene onto a 2-D plane (Figure 2.1). A 3-D scene consisting of a number of objects each with depth, texture and illumination is projected onto a plane to form a 2-D representation of the scene. The 2-D representation contains varying texture and illumination but no depth information. A *still image* is a 'snapshot' of the 2-D representation at a particular instant in time whereas a *video sequence* represents the scene over a period of time.

2.2.2 Digital Video

A 'real' visual scene is continuous both spatially and temporally. In order to represent and process a visual scene digitally it is necessary to sample the real scene spatially (typically on a rectangular grid in the video image plane) and temporally (typically as a series of 'still'

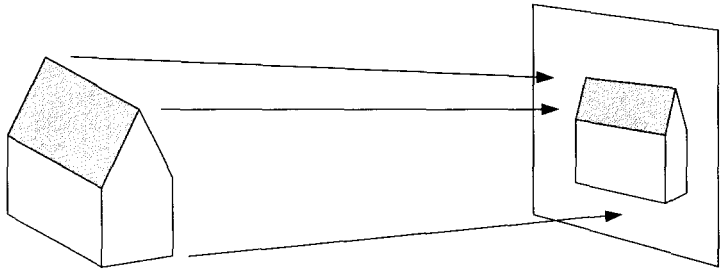


Figure 2.1 Projection of 3-D scene onto a video image

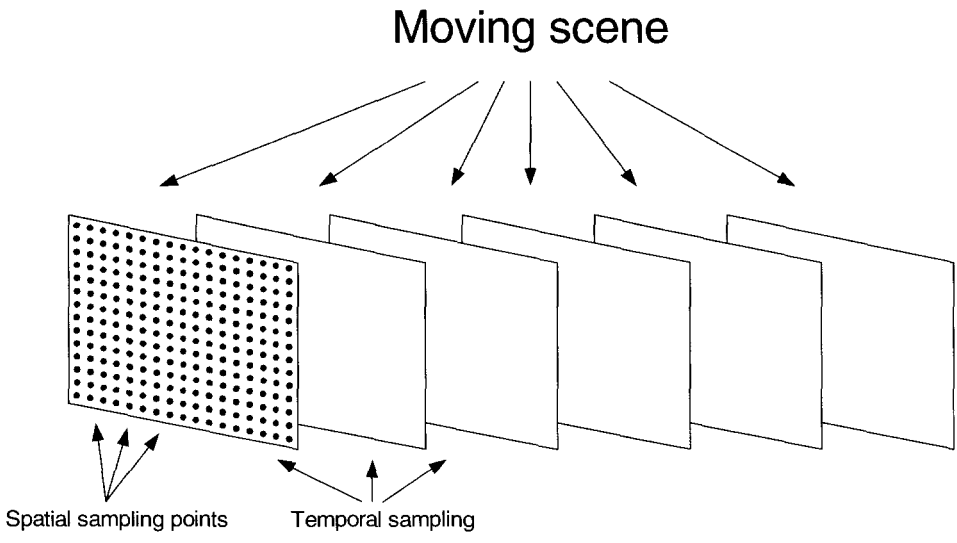


Figure 2.2 Spatial and temporal sampling

images or *frames* sampled at regular intervals in time) as shown in Figure 2.2. Digital video is the representation of a spatio-temporally sampled video scene in digital form. Each spatio-temporal sample (described as a picture element or *pixel*) is represented digitally as one or more numbers that describe the brightness (*luminance*) and colour of the sample.

A *digital video system* is shown in Figure 2.3. At the input to the system, a 'real' visual scene is captured, typically with a camera and converted to a sampled digital representation.

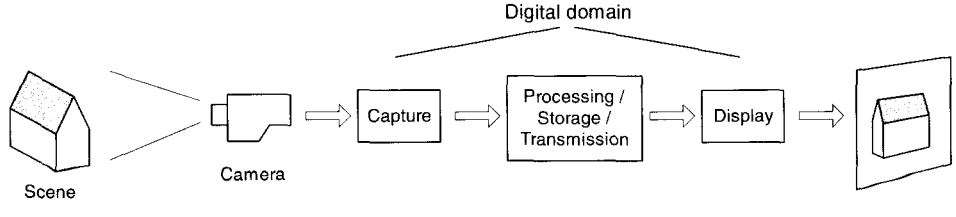


Figure 2.3 Digital video system: capture, processing and display

This digital video signal may then be handled in the digital domain in a number of ways, including processing, storage and transmission. At the output of the system, the digital video signal is displayed to a viewer by reproducing the 2-D video image (or video sequence) on a 2-D display.

2.2.3 Video Capture

Video is captured using a camera or a system of cameras. Most current digital video systems use 2-D video, captured with a single camera. The camera focuses a 2-D projection of the video scene onto a sensor, such as an array of charge coupled devices (CCD array). In the case of colour image capture, each colour component (see Section 2.3) is filtered and projected onto a separate CCD array.

Figure 2.4 shows a two-camera system that captures two 2-D projections of the scene, taken from different viewing angles. This provides a stereoscopic representation of the scene: the two images, when viewed in the left and right eye of the viewer, give an appearance of 'depth' to the scene. There is an increasing interest in the use of 3-D digital video, where the video signal is represented and processed in three dimensions. This requires the capture system to provide depth information as well as brightness and colour, and this may be obtained in a number of ways. Stereoscopic images can be processed to extract approximate depth information and form a 3-D representation of the scene: other methods of obtaining depth information include processing of multiple images from a single camera (where either the camera or the objects in the scene are moving) and the use of laser 'striping' to obtain depth maps. In this book we will concentrate on 2-D video systems.

Generating a digital representation of a video scene can be considered in two stages: acquisition (converting a projection of the scene into an electrical signal, for example via a CCD array) and digitisation (sampling the projection spatially and temporally and converting each sample to a number or set of numbers). Digitisation may be carried out using a separate device or board (e.g. a video capture card in a PC): increasingly, the digitisation process is becoming integrated with cameras so that the output of a camera is a signal in sampled digital form.

2.2.4 Sampling

A digital image may be generated by sampling an analogue video signal (i.e. a varying electrical signal that represents a video image) at regular intervals. The result is a sampled

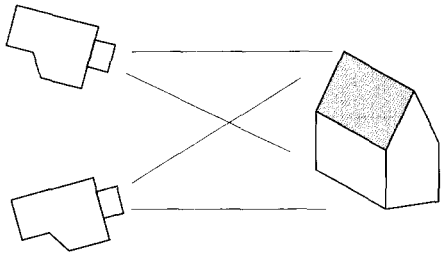


Figure 2.4 Stereoscopic camera system

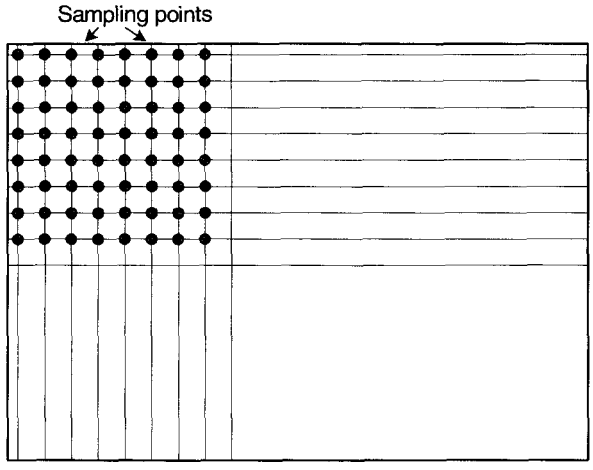


Figure 2.5 Spatial sampling (square grid)

version of the image: the sampled image is only defined at a series of regularly spaced sampling points. The most common format for a sampled image is a rectangle (often with width larger than height) with the sampling points positioned on a square grid (Figure 2.5). The visual quality of the image is influenced by the number of sampling points. More sampling points (a higher sampling resolution) give a 'finer' representation of the image: however, more sampling points require higher storage capacity. Table 2.1 lists some commonly used image resolutions and gives an approximately equivalent analogue video quality: VHS video, broadcast TV and high-definition TV.

A moving video image is formed by sampling the video signal temporally, taking a rectangular 'snapshot' of the signal at periodic time intervals. Playing back the series of frames produces the illusion of motion. A higher temporal sampling rate (*frame rate*) gives a 'smoother' appearance to motion in the video scene but requires more samples to be captured and stored (see Table 2.2). Frame rates below 10 frames per second are sometimes

Table 2.1 Typical video image resolutions

Image resolution	Number of sampling points	Analogue video 'equivalent'
352 × 288	101 376	VHS video
704 × 576	405 504	Broadcast television
1440 × 1152	1 313 280	High-definition television

Table 2.2 Video frame rates

Video frame rate	Appearance
Below 10 frames per second	'Jerky', unnatural appearance to movement
10–20 frames per second	Slow movements appear OK; rapid movement is clearly 'jerky'
20–30 frames per second	Movement is reasonably smooth
50–60 frames per second	Movement is very smooth

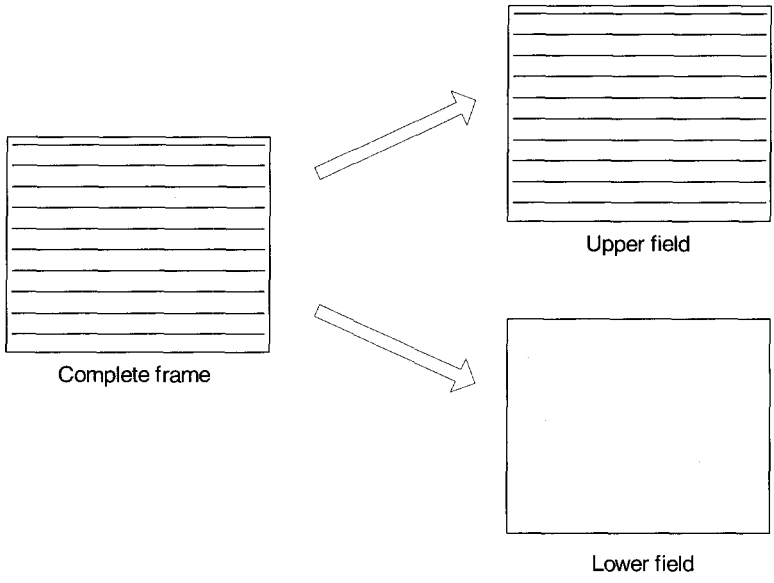


Figure 2.6 Interlaced fields

used for very low bit-rate video communications (because the amount of data is relatively small): however, motion is clearly jerky and unnatural at this rate. Between 10 and 20 frames per second is more typical for low bit-rate video communications; 25 or 30 frames per second is standard for television pictures (together with the use of interlacing, see below); 50 or 60 frames per second is appropriate for high-quality video (at the expense of a very high data rate).

The visual appearance of a temporally sampled video sequence can be improved by using *interlaced video*, commonly used for broadcast-quality television signals. For example, the European PAL video standard operates at a temporal frame rate of 25 Hz (i.e. 25 complete frames of video per second). However, in order to improve the visual appearance without increasing the data rate, the video sequence is composed of *fields* at a rate of 50 Hz (50 fields per second). Each field contains half of the lines that make up a complete frame (Figure 2.6): the odd- and even-numbered lines from the frame on the left are placed in two separate fields, each containing half the information of a complete frame. These fields are captured and displayed at 1/50th of a second intervals and the result is an update rate of 50 Hz, with the data rate of a signal at 25 Hz. Video that is captured and displayed in this way is known as interlaced video and generally has a more pleasing visual appearance than video transmitted as complete frames (*non-interlaced* or *progressive video*). Interlaced video can, however, produce unpleasant visual artefacts when displaying certain textures or types of motion.

2.2.5 Display

Displaying a 2-D video signal involves recreating each frame of video on a 2-D display device. The most common type of display is the cathode ray tube (CRT) in which the image

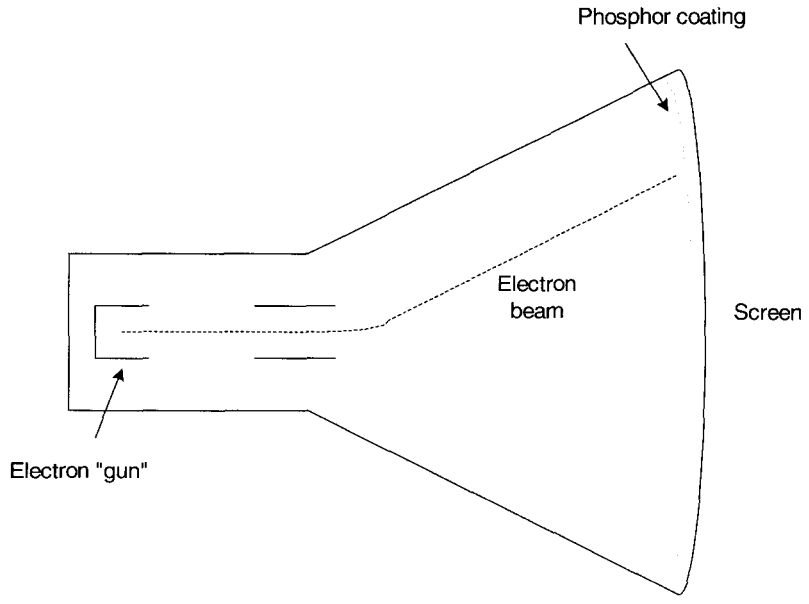


Figure 2.7 CRT display

is formed by scanning a modulated beam of electrons across a phosphorescent screen (Figure 2.7). CRT technology is mature and reasonably cheap to produce. However, a CRT suffers from the requirement to provide a sufficiently long path for the electron beam (making the device bulky) and the weight of the vacuum tube. Liquid crystal displays (LCDs) are becoming a popular alternative to the CRT for computer applications but are not as bright; other alternatives such as flat-panel plasma displays are beginning to emerge but are not yet available at competitive prices.

2.3 COLOUR SPACES

A monochrome ('grey scale') video image may be represented using just one number per spatio-temporal sample. This number indicates the brightness or luminance of each sample position: conventionally, a larger number indicates a brighter sample. If a sample is represented using n bits, then a value of 0 may represent black and a value of $(2^n - 1)$ may represent white, with other values in between describing shades of grey. Luminance is commonly represented with 8 bits per sample for 'general-purpose' video applications. Higher luminance 'depths' (e.g. 12 bits or more per sample) are sometimes used for specialist applications (such as digitising of X-ray slides).

Representing colour requires multiple numbers per sample. There are several alternative systems for representing colour, each of which is known as a *colour space*. We will concentrate here on two of the most common colour spaces for digital image and video representation: RGB (red/green/blue) and YCrCb (luminance/red chrominance/blue chrominance).

2.3.1 RGB

In the red/green/blue colour space, each pixel is represented by three numbers indicating the relative proportions of red, green and blue. These are the three additive primary colours of light: any colour may be reproduced by combining varying proportions of red, green and blue light. Because the three components have roughly equal importance to the final colour, RGB systems usually represent each component with the same precision (and hence the same number of bits). Using 8 bits per component is quite common: $3 \times 8 = 24$ bits are required to represent each pixel. Figure 2.8 shows an image (originally colour, but displayed here in monochrome!) and the brightness 'maps' of each of its three colour components. The girl's cap is a bright pink colour: this appears bright in the red component and slightly less bright in the blue component.

(a)

(b)

Figure 2.8 (a) Image, (b) R, (c) G, (d) B components

(c)

(d)

Figure 2.8 (*Continued*)

2.3.2 YCrCb

RGB is not necessarily the most efficient representation of colour. The human visual system (HVS, see Section 2.4) is less sensitive to colour than to luminance (brightness): however, the RGB colour space does not provide an easy way to take advantage of this since the three colours are equally important and the luminance is present in all three colour components. It is possible to represent a colour image more efficiently by separating the luminance from the colour information.

A popular colour space of this type is Y:Cr:Cb. Y is the luminance component, i.e. a monochrome version of the colour image. Y is a weighted average of R, G and B:

$$Y = k_r R + k_g G + k_b B$$

where k are weighting factors. The colour information can be represented as *colour difference* or *chrominance* components, where each chrominance component is the difference between R, G or B and the luminance Y:

$$\begin{aligned} \mathrm{Cr} &= \mathrm{R} - \mathrm{Y} \\ \mathrm{Cb} &= \mathrm{B} - \mathrm{Y} \\ \mathrm{Cg} &= \mathrm{G} - \mathrm{Y} \end{aligned}$$

The complete description is given by Y (the luminance component) and three colour differences Cr, Cb and Cg that represent the 'variation' between the colour intensity and the 'background' luminance of the image.

So far, this representation has little obvious merit: we now have four components rather than three. However, it turns out that the value of Cr + Cb + Cg is a constant. This means that only two of the three chrominance components need to be transmitted: the third component can always be found from the other two. In the Y : Cr : Cb space, only the luminance (Y) and red and blue chrominance (Cr, Cb) are transmitted. Figure 2.9 shows the effect of this operation on the colour image. The two chrominance components only have significant values where there is a significant 'presence' or 'absence' of the appropriate colour (for example, the pink hat appears as an area of relative brightness in the red chrominance).

The equations for converting an RGB image into the Y : Cr : Cb colour space and vice versa are given in Equations 2.1 and 2.2. Note that G can be extracted from the Y : Cr : Cb representation by subtracting Cr and Cb from Y.

$$\begin{aligned} \mathrm{Y} &= 0.299\,\mathrm{R} + 0.587\,\mathrm{G} + 0.114\,\mathrm{B} \\ \mathrm{Cb} &= 0.564\,(\mathrm{B} - \mathrm{Y}) \\ \mathrm{Cr} &= 0.713\,(\mathrm{R} - \mathrm{Y}) \end{aligned} \tag{2.1}$$

$$\begin{aligned} \mathrm{R} &= \mathrm{Y} + 1.402\,\mathrm{Cr} \\ \mathrm{G} &= \mathrm{Y} - 0.344\,\mathrm{Cb} - 0.714\,\mathrm{Cr} \\ \mathrm{B} &= \mathrm{Y} + 1.772\,\mathrm{Cb} \end{aligned} \tag{2.2}$$

The key advantage of Y : Cr : Cb over RGB is that the Cr and Cb components may be represented with a *lower resolution* than Y because the HVS is less sensitive to colour than luminance. This reduces the amount of data required to represent the chrominance components without having an obvious effect on visual quality: to the casual observer, there is no apparent difference between an RGB image and a Y : Cr : Cb image with reduced chrominance resolution.

Figure 2.10 shows three popular 'patterns' for sub-sampling Cr and Cb. 4 : 4 : 4 means that the three components (Y : Cr : Cb) have the same resolution and hence a sample of each component exists at every pixel position. (The numbers indicate the relative sampling rate of each component in the *horizontal* direction, i.e. for every 4 luminance samples there are 4 Cr and 4 Cb samples.) 4 : 4 : 4 sampling preserves the full fidelity of the chrominance components. In 4 : 2 : 2 sampling, the chrominance components have the same vertical resolution but half the horizontal resolution (the numbers indicate that for every 4 luminance

(a)

(b)

(c)

Figure 2.9 (a) Luminance, (b) Cr, (c) Cb components

samples in the horizontal direction there are 2 Cr and 2 Cb samples) and the locations of the samples are shown in the figure. 4 : 2 : 2 video is used for high-quality colour reproduction.

4 : 2 : 0 means that Cr and Cb each have half the horizontal and vertical resolution of Y, as shown. The term '4 : 2 : 0' is rather confusing: the numbers do not actually have a sensible interpretation and appear to have been chosen historically as a 'code' to identify this

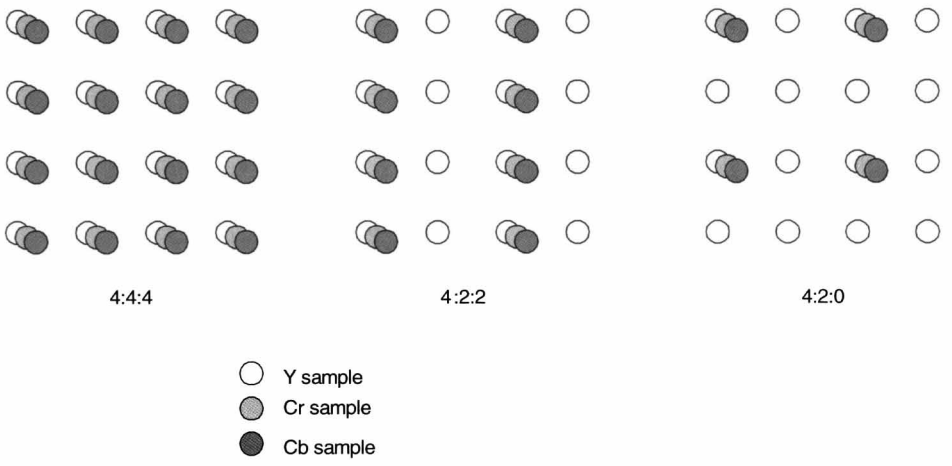


Figure 2.10 Chrominance subsampling patterns

particular sampling pattern. 4 : 2 : 0 sampling is popular in 'mass market' digital video applications such as video conferencing, digital television and DVD storage. Because each colour difference component contains a quarter of the samples of the Y component, 4 : 2 : 0 video requires exactly half as many samples as 4 : 4 : 4 (or R : G : B) video.

Example

Image resolution: 720 × 576 pixels
Y resolution: 720 × 576 samples, each represented with 8 bits

4 : 4 : 4 Cr, Cb resolution: 720 × 576 samples, each 8 bits
Total number of bits: $720 \times 576 \times 8 \times 3 = 9\,953\,280$ bits

4 : 2 : 0 Cr, Cb resolution: 360 × 288 samples, each 8 bits
Total number of bits: $(720 \times 576 \times 8) + (360 \times 288 \times 8 \times 2) = 4\,976\,640$ bits

The 4 : 2 : 0 version requires half as many bits as the 4 : 4 : 4 version

To further confuse things, 4 : 2 : 0 sampling is sometimes described as '12 bits per pixel'. The reason for this can be illustrated by examining a group of 4 pixels (Figure 2.11). The left-hand diagram shows 4 : 4 : 4 sampling: a total of 12 samples are required, 4 each of Y, Cr and Cb, requiring a total of $12 \times 8 = 96$ bits, i.e. an average of $96/4 = 24$ bits per pixel. The right-hand diagram shows 4 : 2 : 0 sampling: 6 samples are required, 4 Y and one each of Cr, Cb, requiring a total of $6 \times 8 = 48$ bits, i.e. an average of $48/4 = 12$ bits per pixel.

Figure 2.11 4 pixels: 24 and 12 bpp

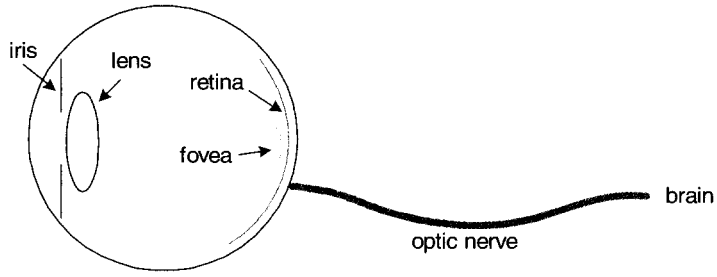


Figure 2.12 HVS components

2.4 THE HUMAN VISUAL SYSTEM

A critical design goal for a digital video system is that the visual images produced by the system should be 'pleasing' to the viewer. In order to achieve this goal it is necessary to take into account the response of the *human visual system* (HVS). The HVS is the 'system' by which a human observer views, interprets and responds to visual stimuli. The main components of the HVS are shown in Figure 2.12:

- *Eye*: The image is focused by the *lens* onto the photodetecting area of the eye, the *retina*. Focusing and object tracking are achieved by the eye *muscles* and the *iris* controls the aperture of the lens and hence the amount of light entering the eye.
- *Retina*: The retina consists of an array of *cones* (photoreceptors sensitive to colour at high light levels) and *rods* (photoreceptors sensitive to luminance at low light levels). The more sensitive cones are concentrated in a central region (the *fovea*) which means that high-resolution colour vision is only achieved over a small area at the centre of the field of view.
- *Optic nerve*: This carries electrical signals from the retina to the brain.
- *Brain*: The human brain processes and interprets visual information, based partly on the received information (the image detected by the retina) and partly on prior learned responses (such as known object shapes).

The operation of the HVS is a large and complex area of study. Some of the important features of the HVS that have implications for digital video system design are listed in Table 2.3.

2.5 VIDEO QUALITY

In order to specify, evaluate and compare video communication systems it is necessary to determine the quality of the video images displayed to the viewer. Measuring visual quality is a difficult and often imprecise art because there are so many factors that can influence the results. Visual quality is inherently *subjective* and is therefore influenced by many subjective factors that can make it difficult to obtain a completely accurate measure of quality.

Table 2.3 Features of the HVS

Feature	Implication for digital video systems
The HVS is more sensitive to luminance detail than to colour detail	Colour (or chrominance) resolution may be reduced without significantly affecting image quality
The HVS is more sensitive to high contrast (i.e. large differences in luminance) than low contrast	Large changes in luminance (e.g. edges in an image) are particularly important to the appearance of the image
The HVS is more sensitive to low spatial frequencies (i.e. changes in luminance that occur over a large area) than high spatial frequencies (rapid changes that occur in a small area)	It may be possible to compress images by discarding some of the less important higher frequencies (however, edge information should be preserved)
The HVS is more sensitive to image features that persist for a long duration	It is important to minimise temporally persistent disturbances or artefacts in an image
The illusion of 'smooth' motion can be achieved by presenting a series of images at a rate of 20–30 Hz or more	Video systems should aim for frame repetition rates of 20 Hz or more for 'natural' moving video
HVS responses vary from individual to individual	Multiple observers should be used to assess the quality of a video system

Measuring visual quality using *objective* criteria gives accurate, repeatable results, but as yet there are no objective measurement systems that will completely reproduce the subjective experience of a human observer watching a video display.

2.5.1 Subjective Quality Measurement

Several test procedures for subjective quality evaluation are defined in ITU-R Recommendation BT.500-10.[1] One of the most popular of these quality measures is the double stimulus continuous quality scale (DSCQS) method. An assessor is presented with a pair of images or short video sequences A and B, one after the other, and is asked to give A and B a 'score' by marking on a continuous line with five intervals. Figure 2.13 shows an example of the rating form on which the assessor grades each sequence.

In a typical test session, the assessor is shown a series of sequence pairs and is asked to grade each pair. Within each pair of sequences, one is an unimpaired 'reference' sequence and the other is the same sequence, modified by a system or process under test. A typical example from the evaluation of video coding systems is shown in Figure 2.14: the original sequence is compared with the same sequence, encoded and decoded using a video CODEC.

The order of the two sequences, original and 'impaired', is randomised during the test session so that the assessor does not know which is the original and which is the impaired sequence. This helps prevent the assessor from prejudging the impaired sequence compared with the reference sequence. At the end of the session, the scores are converted to a normalised range and the result is a score (sometimes described as a 'mean opinion score') that indicates the *relative* quality of the impaired and reference sequences.

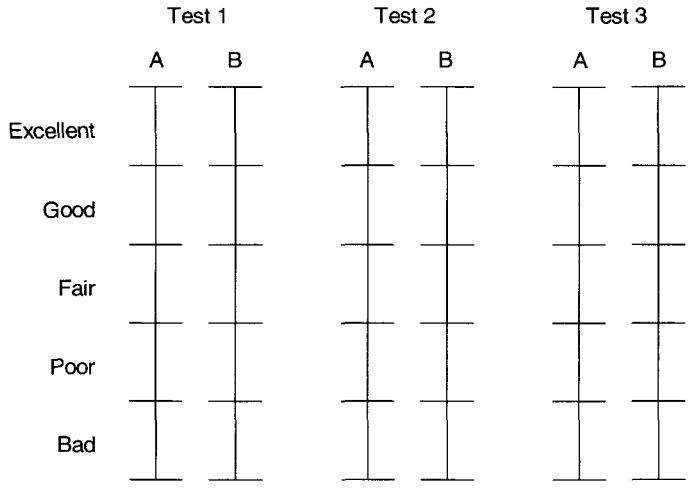


Figure 2.13 DSCQS rating form

The DSCQS test is generally accepted as a realistic measure of subjective visual quality. However, it suffers from practical problems. The results can vary significantly, depending on the assessor and also on the video sequence under test. This variation can be compensated for by repeating the test with several sequences and several assessors. An 'expert' assessor (e.g. one who is familiar with the nature of video compression distortions or 'artefacts') may give a biased score and it is preferable to use 'non-expert' assessors. In practice this means that a large pool of assessors is required because a non-expert assessor will quickly learn to recognise characteristic artefacts in the video sequences. These factors make it expensive and time-consuming to carry out the DSCQS tests thoroughly.

A second problem is that this test is only really suitable for short sequences of video. It has been shown[2] that the 'recency effect' means that the viewer's opinion is heavily biased towards the last few seconds of a video sequence: the quality of this last section will strongly influence the viewer's rating for the whole of a longer sequence. Subjective tests are also influenced by the viewing conditions: a test carried out in a comfortable, relaxed environment will earn a higher rating than the same test carried out in a less comfortable setting.

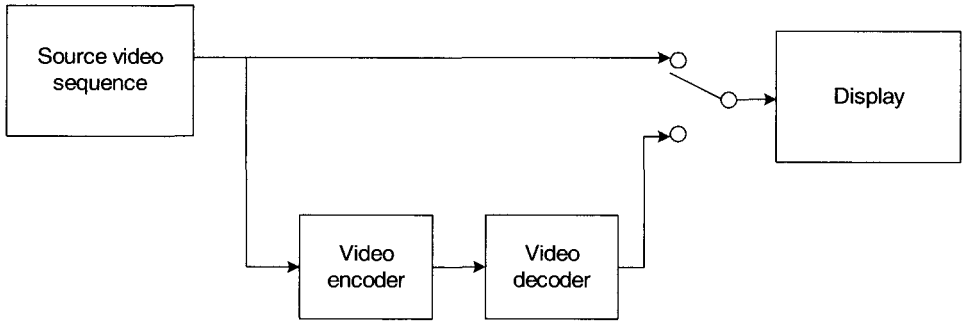


Figure 2.14 DSCQS testing system

2.5.2 Objective Quality Measurement

Because of the problems of subjective measurement, developers of digital video systems rely heavily on objective measures of visual quality. Objective measures have not yet replaced subjective testing: however, they are considerably easier to apply and are particularly useful during development and for comparison purposes.

Probably the most widely used objective measure is peak signal to noise ratio (PSNR), calculated using Equation 2.3. PSNR is measured on a logarithmic scale and is based on the mean squared error (MSE) between an original and an impaired image or video frame, relative to $(2^n - 1)^2$ (the square of the highest possible signal value in the image).

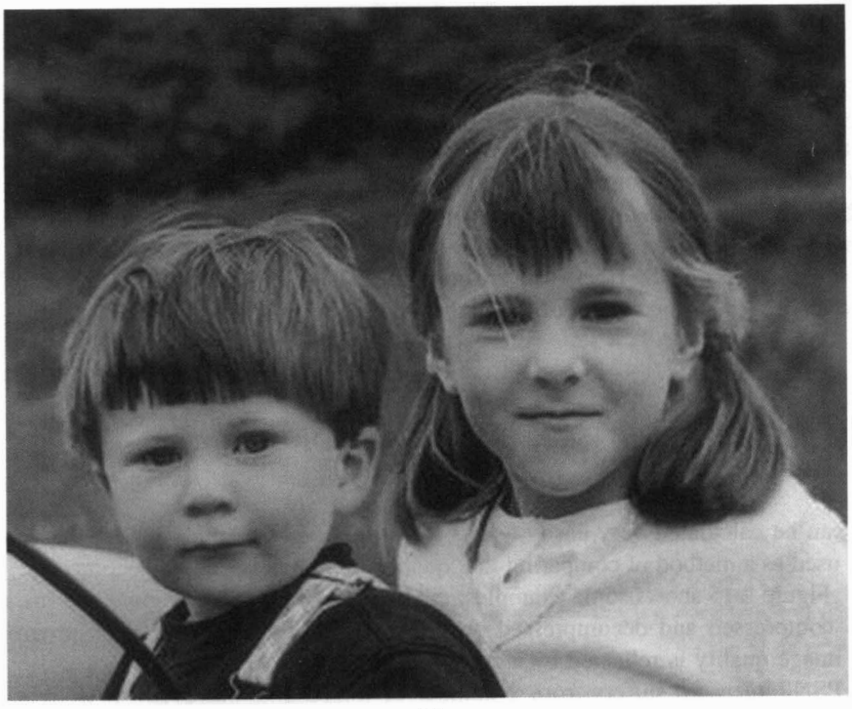
$$\text{PSNR}_{\text{dB}} = 10 \log_{10} \frac{(2^n - 1)^2}{\text{MSE}} \tag{2.3}$$

PSNR can be calculated very easily and is therefore a very popular quality measure. It is widely used as a method of comparing the 'quality' of compressed and decompressed video images. Figure 2.15 shows some examples: the first image (a) is the original and (b), (c) and (d) are compressed and decompressed versions of the original image. The progressively poorer image quality is reflected by a corresponding drop in PSNR.

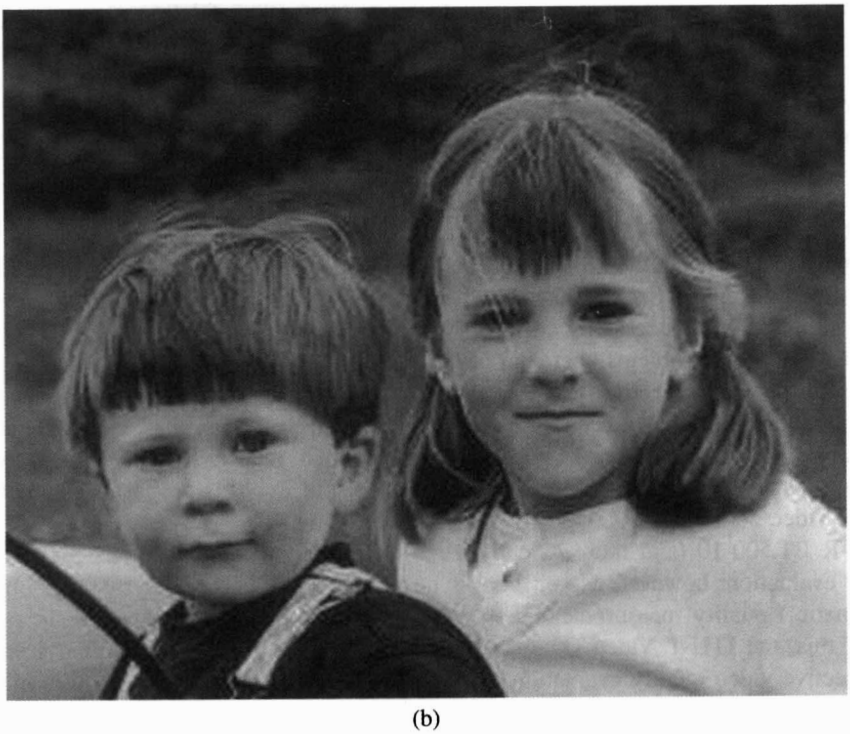
The PSNR measure suffers from a number of limitations, however. PSNR requires an 'unimpaired' original image for comparison: this may not be available in every case and it may not be easy to verify that an 'original' image has perfect fidelity. A more important limitation is that PSNR does not correlate well with subjective video quality measures such as ITU-R 500. For a given image or image sequence, high PSNR indicates relatively high quality and low PSNR indicates relatively low quality. However, a particular value of PSNR does not necessarily equate to an 'absolute' subjective quality. For example, Figure 2.16 shows two impaired versions of the original image from Figure 2.15. Image (a) (with a blurred background) has a PSNR of 32.7 dB, whereas image (b) (with a blurred foreground) has a higher PSNR of 37.5 dB. Most viewers would rate image (b) as significantly poorer than image (a): however, the PSNR measure simply counts the mean squared pixel errors and by this method image (b) is ranked as 'better' than image (a). This example shows that PSNR ratings do not necessarily correlate with 'true' subjective quality.

Because of these problems, there has been a lot of work in recent years to try to develop a more sophisticated objective test that closely approaches subjective test results. Many different approaches have been proposed,[3–5] but none of these has emerged as clear alternatives to subjective tests. With improvements in objective quality measurement, however, some interesting applications become possible, such as proposals for 'constant-quality' video coding[6] (see Chapter 10, 'Rate Control').

ITU-R BT.500-10 (and more recently, P.910) describe standard methods for subjective quality evaluation: however, as yet there is no standardised, accurate system for objective ('automatic') quality measurement that is suitable for digitally coded video. In recognition of this, the ITU-T Video Quality Experts Group (VQEG) are developing a standard for objective video quality evaluation[7]. The first step in this process was to test and compare potential models for objective evaluation. In March 2000, VQEG reported on the first round of tests in which 10 competing systems were tested under identical conditions.

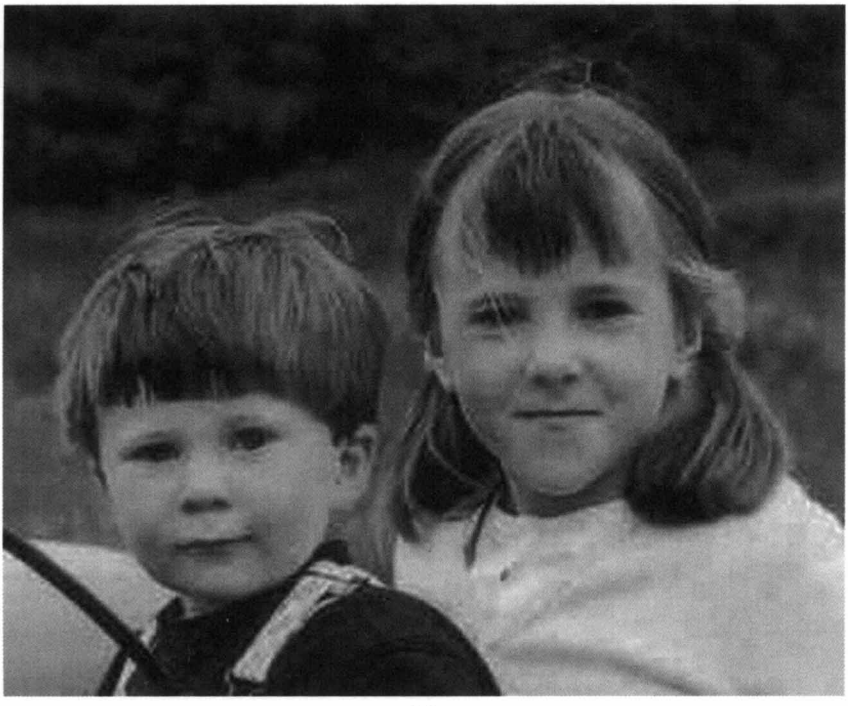


(a)

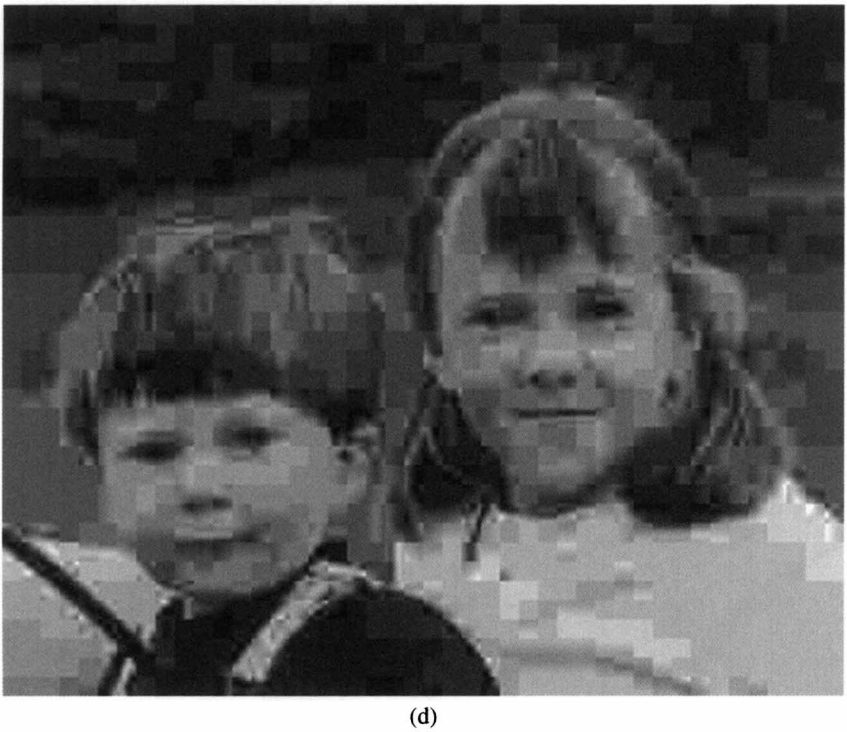


(b)

Figure 2.15 PSNR examples: (a) original; (b) 33.2 dB; (c) 31.8 dB; (d) 26.5 dB



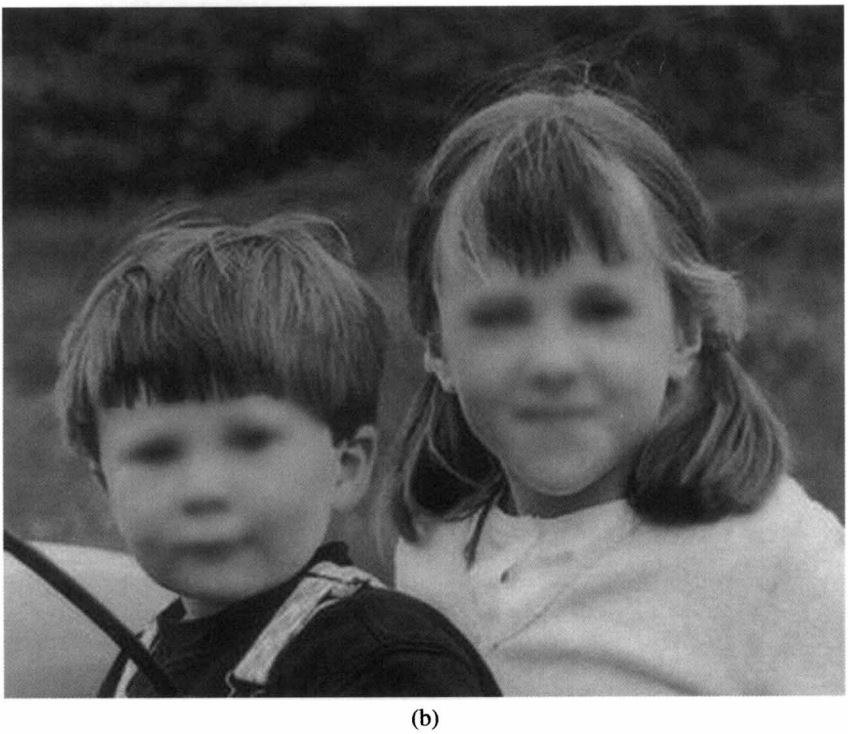
(c)



(d)

Figure 2.15 (*Continued*)

(a)



(b)

Figure 2.16 (a) Impairment 1 (32.7 dB); (b) impairment 2 (37.5 dB)

Table 2.4 ITU-R BT.601-5 parameters

	30 Hz frame rate	25 Hz frame rate
Fields per second	60	50
Lines per complete frame	525	625
Luminance samples per line	858	864
Chrominance samples per line	429	432
Bits per sample	8	8
Total bit rate	216 Mbps	216 Mbps
Active lines per frame	480	576
Active samples per line (Y)	720	720
Active samples per line (Cr, Cb)	360	360

Unfortunately, none of the 10 proposals was considered suitable for standardisation. The problem of accurate objective quality measurement is therefore likely to remain for some time to come.

The PSNR measure is widely used as an approximate objective measure for visual quality and so we will use this measure for quality comparison in this book. However, it is worth remembering the limitations of PSNR when comparing different systems and techniques.

2.6 STANDARDS FOR REPRESENTING DIGITAL VIDEO

A widely used format for digitally coding video signals for television production is ITU-R Recommendation BT.601-5[8] (the term 'coding' in this context means conversion to digital format and does not imply compression). The luminance component of the video signal is sampled at 13.5 MHz and the chrominance at 6.75 MHz to produce a 4:2:2 Y:Cr:Cb component signal. The parameters of the sampled digital signal depend on the video frame rate (either 30 or 25 Hz) and are shown in Table 2.4. It can be seen that the higher 30 Hz frame rate is compensated for by a lower spatial resolution so that the total bit rate is the same in each case (216 Mbps). The actual area shown on the display, the *active area*, is smaller than the total because it excludes horizontal and vertical blanking intervals that exist 'outside' the edges of the frame. Each sample has a possible range of 0–255: however, levels of 0 and 255 are reserved for synchronisation. The active luminance signal is restricted to a range of 16 (black) to 235 (white).

For video coding applications, video is often converted to one of a number of 'intermediate formats' prior to compression and transmission. A set of popular frame resolutions is based around the common intermediate format, CIF, in which each frame has a

Table 2.5 Intermediate formats

Format	Luminance resolution (horiz. × vert.)
Sub-QCIF	128 × 96
Quarter CIF (QCIF)	176 × 144
CIF	352 × 288
4CIF	704 × 576

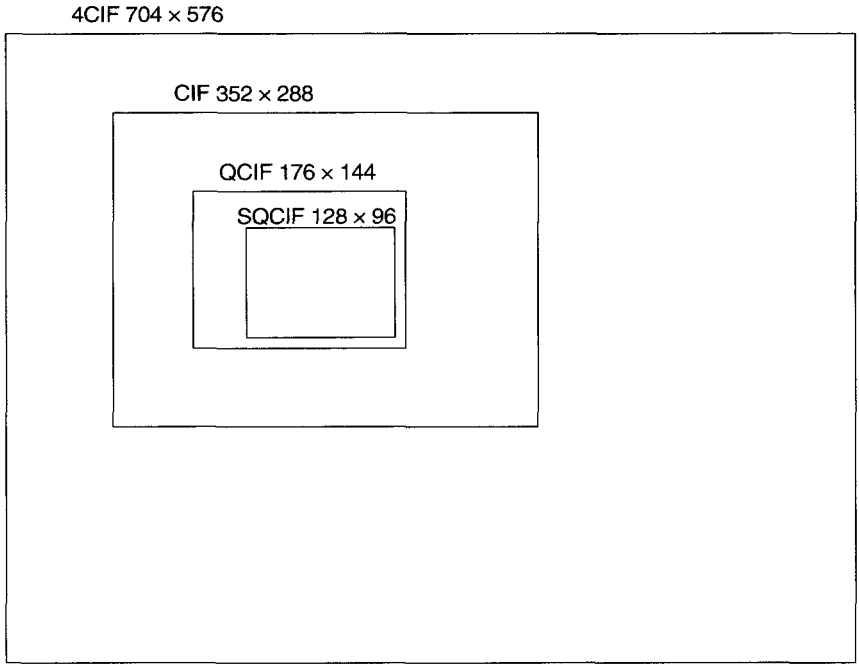


Figure 2.17 Intermediate formats (illustration)

resolution of 352 × 288 pixels. The resolutions of these formats are listed in Table 2.5 and their relative dimensions are illustrated in Figure 2.17.

2.7 APPLICATIONS

The last decade has seen a rapid increase in applications for digital video technology and new, innovative applications continue to emerge. A small selection is listed here:

- *Home video*: Video camera recorders for professional and home use are increasingly moving away from analogue tape to digital media (including digital storage on tape and on solid-state media). Affordable DVD video recorders will soon be available for the home.
- *Video storage*: A variety of digital formats are now used for storing video on disk, tape and compact disk or DVD for business and home use, both in compressed and uncompressed form.
- *Video conferencing*: One of the earliest applications for video compression, video conferencing facilitates meetings between participants in two or more separate locations.
- *Video telephony*: Often used interchangeably with video conferencing, this usually means a face-to-face discussion between two parties via a video 'link'.
- *Remote learning*: There is an increasing interest in the provision of computer-based learning to supplement or replace traditional 'face-to-face' teaching and learning. Digital

video is seen as an important component of this in the form of stored video material and video conferencing.

- *Remote medicine*: Medical support provided at a distance, or 'telemedicine', is another potential growth area where digital video and images may be used together with other monitoring techniques to provide medical advice at a distance.
- *Television*: Digital television is now widely available and many countries have a timetable for 'switching off' the existing analogue television service. Digital TV is one of the most important mass-market applications for video coding and compression.
- *Video production*: Fully digital video storage, editing and production have been widely used in television studios for many years. The requirement for high image fidelity often means that the popular 'lossy' compression methods described in this book are not an option.
- *Games and entertainment*: The potential for 'real' video imagery in the computer gaming market is just beginning to be realised with the convergence of 3-D graphics and 'natural' video.

2.7.1 Platforms

Developers are targeting an increasing range of platforms to run the ever-expanding list of digital video applications.

Dedicated platforms are designed to support a specific video application and no other. Examples include digital video cameras, dedicated video conferencing systems, digital TV set-top boxes and DVD players. In the early days, the high processing demands of digital video meant that dedicated platforms were the only practical design solution. Dedicated platforms will continue to be important for low-cost, mass-market systems but are increasingly being replaced by more flexible solutions.

The *PC* has emerged as a key platform for digital video. A continual increase in PC processing capabilities (aided by hardware enhancements for media applications such as the Intel MMX instructions) means that it is now possible to support a wide range of video applications from video editing to real-time video conferencing.

Embedded platforms are an important new market for digital video techniques. For example, the personal communications market is now huge, driven mainly by users of mobile telephones. Video services for mobile devices (running on low-cost embedded processors) are seen as a major potential growth area. This type of platform poses many challenges for application developers due to the limited processing power, relatively poor wireless communications channel and the requirement to keep equipment and usage costs to a minimum.

2.8 SUMMARY

Sampling of an analogue video signal, both spatially and temporally, produces a digital video signal. Representing a colour scene requires at least three separate 'components': popular colour 'spaces' include red/green/blue and Y/Cr/Cb (which has the advantage that the chrominance may be subsampled to reduce the information rate without significant loss

of quality). The human observer's response to visual information affects the way we perceive video quality and this is notoriously difficult to quantify accurately. Subjective tests (involving 'real' observers) are time-consuming and expensive to run; objective tests range from the simplistic (but widely used) PSNR measure to complex models of the human visual system.

The digital video applications listed above have been made possible by the development of compression or coding technology. In the next chapter we introduce the basic concepts of video and image compression.

REFERENCES

1. Recommendation ITU-T BT.500-10, 'Methodology for the subjective assessment of the quality of television pictures', ITU-T, 2000.
2. R. Aldridge, J. Davidoff, M. Ghanbari, D. Hands and D. Pearson, 'Subjective assessment of time-varying coding distortions', *Proc. PCS96*, Melbourne, March 1996.
3. C. J. van den Branden Lambrecht and O. Verscheure, 'Perceptual quality measure using a spatio-temporal model of the Human Visual System', *Digital Video Compression Algorithms and Technologies, Proc. SPIE*, Vol. 2668, San Jose, 1996.
4. H. Wu, Z. Yu, S. Winkler and T. Chen, 'Impairment metrics for MC/DPCM/DCT encoded digital video', *Proc. PCS01*, Seoul, April 2001.
5. K. T. Tan and M. Ghanbari, 'A multi-metric objective picture quality measurement model for MPEG video', *IEEE Trans. CSVT*, **10**(7), October 2000.
6. A. Basso, I. Dalgiç, F. Tobagi and C. J. van den Branden Lambrecht, 'A feedback control scheme for low latency constant quality MPEG-2 video encoding', *Digital Compression Technologies and Systems for Video Communications, Proc. SPIE*, Vol. 2952, Berlin, 1996.
7. http://www.vqeg.org/ [Video Quality Experts Group].
8. Recommendation ITU-R BT.601-5, 'Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios', ITU-T, 1995.