2

Video Formats and Quality

2.1 INTRODUCTION

Video coding is the process of compressing and decompressing a digital video signal. This chapter examines the structure and characteristics of digital images and video signals and introduces concepts such as sampling formats and quality metrics that are helpful to an understanding of video coding. Digital video is a representation of a natural (real-world) visual scene, sampled spatially and temporally. A scene is sampled at a point in time to produce a frame (a representation of the complete visual scene at that point in time) or a field (consisting of odd- or even-numbered lines of spatial samples). Sampling is repeated at intervals (e.g. 1/25 or 1/30 second intervals) to produce a moving video signal. Three sets of samples (components) are typically required to represent a scene in colour. Popular formats for representing video in digital form include the ITU-R 601 standard and the set of 'intermediate formats'. The accuracy of a reproduction of a visual scene must be measured to determine the performance of a visual communication system, a notoriously difficult and inexact process. Subjective measurements are time consuming and prone to variations in the response of human viewers. Objective (automatic) measurements are easier to implement but as yet do not accurately match the opinion of a 'real' human.

2.2 NATURAL VIDEO SCENES

A typical 'real world' or 'natural' video scene is composed of multiple objects each with their own characteristic shape, depth, texture and illumination. The colour and brightness of a natural video scene changes with varying degrees of smoothness throughout the scene ('continuous tone'). Characteristics of a typical natural video scene (Figure 2.1) that are relevant for video processing and compression include spatial characteristics (texture variation within scene, number and shape of objects, colour, etc.) and temporal characteristics (object motion, changes in illumination, movement of the camera or viewpoint and so on).

H.264 and MPEG-4 Video Compression: Video Coding for Next-generation Multimedia.
Iain E. G. Richardson. ISBN: 0-470-84837-5

Figure 2.1 Still image from natural video scene

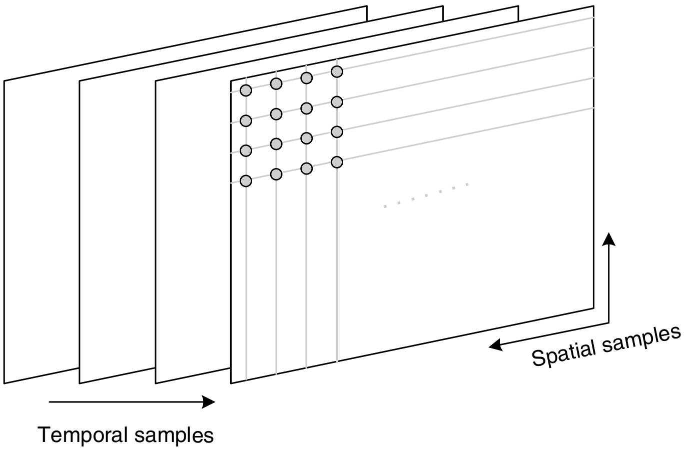

Figure 2.2 Spatial and temporal sampling of a video sequence

2.3 CAPTURE

A natural visual scene is spatially and temporally continuous. Representing a visual scene in digital form involves sampling the real scene spatially (usually on a rectangular grid in the video image plane) and temporally (as a series of still frames or components of frames sampled at regular intervals in time) (Figure 2.2). Digital video is the representation of a sampled video scene in digital form. Each spatio-temporal sample (picture element or pixel) is represented as a number or set of numbers that describes the brightness (luminance) and colour of the sample.

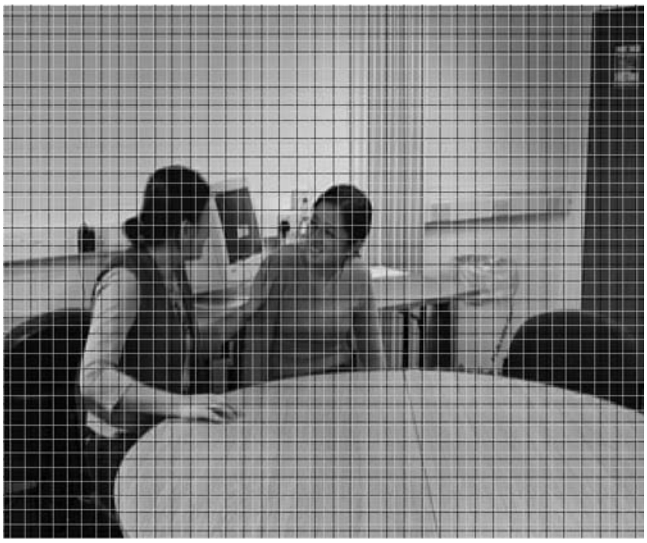

Figure 2.3 Image with 2 sampling grids

To obtain a 2D sampled image, a camera focuses a 2D 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 is separately filtered and projected onto a CCD array (see Section 2.4).

2.3.1 Spatial Sampling

The output of a CCD array is an analogue video signal, a varying electrical signal that represents a video image. Sampling the signal at a point in time produces a sampled image or frame that has defined values at a set of sampling points. The most common format for a sampled image is a rectangle with the sampling points positioned on a square or rectangular grid. Figure 2.3 shows a continuous-tone frame with two different sampling grids superimposed upon it. Sampling occurs at each of the intersection points on the grid and the sampled image may be reconstructed by representing each sample as a square picture element (pixel). The visual quality of the image is influenced by the number of sampling points. Choosing a 'coarse' sampling grid (the black grid in Figure 2.3) produces a low-resolution sampled image (Figure 2.4) whilst increasing the number of sampling points slightly (the grey grid in Figure 2.3) increases the resolution of the sampled image (Figure 2.5).

2.3.2 Temporal Sampling

A moving video image is captured by taking a rectangular 'snapshot' of the signal at periodic time intervals. Playing back the series of frames produces the appearance of motion. A higher temporal sampling rate (frame rate) gives apparently smoother motion in the video scene but requires more samples to be captured and stored. Frame rates below 10 frames per second are sometimes used for very low bit-rate video communications (because the amount of data

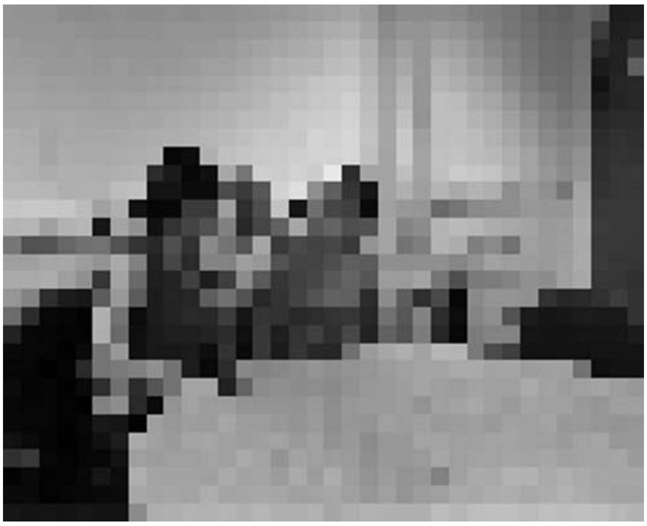

Figure 2.4 Image sampled at coarse resolution (black sampling grid)

Figure 2.5 Image sampled at slightly finer resolution (grey sampling grid)

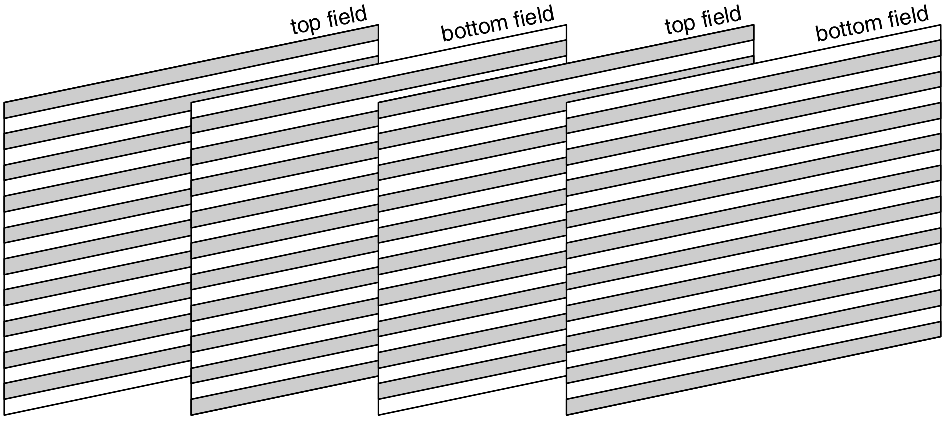

Figure 2.6 Interlaced video sequence

is relatively small) but 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; the image is smoother but jerky motion may be visible in fast-moving parts of the sequence. Sampling at 25 or 30 complete frames per second is standard for television pictures (with interlacing to improve the appearance of motion, see below); 50 or 60 frames per second produces smooth apparent motion (at the expense of a very high data rate).

2.3.3 Frames and Fields

A video signal may be sampled as a series of complete frames (*progressive* sampling) or as a sequence of interlaced fields (*interlaced* sampling). In an interlaced video sequence, half of the data in a frame (one field) is sampled at each temporal sampling interval. A field consists of either the odd-numbered or even-numbered lines within a complete video frame and an interlaced video sequence (Figure 2.6) contains a series of fields, each representing half of the information in a complete video frame (e.g. Figure 2.7 and Figure 2.8). The advantage of this sampling method is that it is possible to send twice as many fields per second as the number of frames in an equivalent progressive sequence with the same data rate, giving the appearance of smoother motion. For example, a PAL video sequence consists of 50 fields per second and, when played back, motion can appears smoother than in an equivalent progressive video sequence containing 25 frames per second.

2.4 COLOUR SPACES

Most digital video applications rely on the display of colour video and so need a mechanism to capture and represent colour information. A monochrome image (e.g. Figure 2.1) requires just one number to indicate the brightness or luminance of each spatial sample. Colour images, on the other hand, require at least three numbers per pixel position to represent colour accurately. The method chosen to represent brightness (luminance or luma) and colour is described as a colour space.

Figure 2.7 Top field

Figure 2.8 Bottom field

2.4.1 RGB

In the RGB colour space, a colour image sample is represented with three numbers that indicate the relative proportions of Red, Green and Blue (the three additive primary colours of light). Any colour can be created by combining red, green and blue in varying proportions. Figure 2.9 shows the red, green and blue components of a colour image: the red component consists of all the red samples, the green component contains all the green samples and the blue component contains the blue samples. The person on the right is wearing a blue sweater and so this appears 'brighter' in the blue component, whereas the red waistcoat of the figure on the left

Figure 2.9 Red, Green and Blue components of colour image

appears brighter in the red component. The RGB colour space is well-suited to capture and display of colour images. Capturing an RGB image involves filtering out the red, green and blue components of the scene and capturing each with a separate sensor array. Colour Cathode Ray Tubes (CRTs) and Liquid Crystal Displays (LCDs) display an RGB image by separately illuminating the red, green and blue components of each pixel according to the intensity of each component. From a normal viewing distance, the separate components merge to give the appearance of 'true' colour.

2.4.2 YCbCr

The human visual system (HVS) is less sensitive to colour than to luminance (brightness). In the RGB colour space the three colours are equally important and so are usually all stored at the same resolution but it is possible to represent a colour image more efficiently by separating the luminance from the colour information and representing luma with a higher resolution than colour.

The YCbCr colour space and its variations (sometimes referred to as YUV) is a popular way of efficiently representing colour images. Y is the luminance (luma) component and can be calculated as a weighted average of R, G and B:

$$Y = k_r R + k_g G + k_b B \tag{2.1}$$

where k are weighting factors.

The colour information can be represented as *colour difference* (chrominance or chroma) components, where each chrominance component is the difference between R, G or B and the luminance Y:

$$\begin{aligned} Cb &= B - Y \\ Cr &= R - Y \\ Cg &= G - Y \end{aligned} \tag{2.2}$$

The complete description of a colour image is given by Y (the luminance component) and three colour differences Cb, Cr and Cg that represent the difference between the colour intensity and the mean luminance of each image sample. Figure 2.10 shows the chroma components (red, green and blue) corresponding to the RGB components of Figure 2.9. Here, mid-grey is zero difference, light grey is a positive difference and dark grey is a negative difference. The chroma components only have significant values where there is a large

Figure 2.10 Cr, Cg and Cb components

difference between the colour component and the luma image (Figure 2.1). Note the strong blue and red difference components.

So far, this representation has little obvious merit since we now have four components instead of the three in RGB. However, $Cb + Cr + Cg$ is a constant and so only two of the three chroma components need to be stored or transmitted since the third component can always be calculated from the other two. In the YCbCr colour space, only the luma (Y) and blue and red chroma (Cb, Cr) are transmitted. YCbCr has an important advantage over RGB, that is 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 obvious difference between an RGB image and a YCbCr image with reduced chrominance resolution. Representing chroma with a lower resolution than luma in this way is a simple but effective form of image compression.

An RGB image may be converted to YCbCr after capture in order to reduce storage and/or transmission requirements. Before displaying the image, it is usually necessary to convert back to RGB. The equations for converting an RGB image to and from YCbCr colour space and vice versa are given in Equation 2.3 and Equation 2.4[1]. Note that there is no need to specify a separate factor k_g (because $k_b + k_r + k_g = 1$) and that G can be extracted from the YCbCr representation by subtracting Cr and Cb from Y, demonstrating that it is not necessary to store or transmit a Cg component.

$$\begin{aligned} Y &= k_r R + (1 - k_b - k_r)G + k_b B \\ Cb &= \frac{0.5}{1 - k_b}(B - Y) \\ Cr &= \frac{0.5}{1 - k_r}(R - Y) \end{aligned} \tag{2.3}$$

$$\begin{aligned} R &= Y + \frac{1 - k_r}{0.5} Cr \\ G &= Y - \frac{2k_b(1 - k_b)}{1 - k_b - k_r} Cb - \frac{2k_r(1 - k_r)}{1 - k_b - k_r} Cr \\ B &= Y + \frac{1 - k_b}{0.5} Cb \end{aligned} \tag{2.4}$$

[1] Thanks to Gary Sullivan for suggesting the form of Equations 2.3 and 2.4

ITU-R recommendation BT.601 [1] defines $k_b = 0.114$ and $k_r = 0.299$. Substituting into the above equations gives the following widely-used conversion equations:

$$\begin{aligned} Y &= 0.299R + 0.587G + 0.114B \\ Cb &= 0.564(B - Y) \\ Cr &= 0.713(R - Y) \end{aligned} \tag{2.5}$$

$$\begin{aligned} R &= Y + 1.402Cr \\ G &= Y - 0.344Cb - 0.714Cr \\ B &= Y + 1.772Cb \end{aligned} \tag{2.6}$$

2.4.3 YCbCr Sampling Formats

Figure 2.11 shows three sampling patterns for Y, Cb and Cr that are supported by MPEG-4 Visual and H.264. 4:4:4 sampling means that the three components (Y, Cb and Cr) 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 four luminance samples there are four Cb and four Cr samples. 4:4:4 sampling preserves the full fidelity of the chrominance components. In 4:2:2 sampling (sometimes referred to as YUY2), the chrominance components have the same vertical resolution as the luma but half the horizontal resolution (the numbers 4:2:2 mean that for every four luminance samples in the horizontal direction there are two Cb and two Cr samples). 4:2:2 video is used for high-quality colour reproduction.

In the popular 4:2:0 sampling format ('YV12'), Cb and Cr each have half the horizontal and vertical resolution of Y. The term '4:2:0' is rather confusing because the numbers do not actually have a logical interpretation and appear to have been chosen historically as a 'code' to identify this particular sampling pattern and to differentiate it from 4:4:4 and 4:2:2. 4:2:0 sampling is widely used for consumer applications such as video conferencing, digital television and digital versatile disk (DVD) storage. Because each colour difference component contains one quarter of the number of samples in the Y component, 4:2:0 YCbCr 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 eight bits

<u>4:4:4</u> Cb, Cr resolution: 720 × 576 samples, each eight bits
Total number of bits: 720 × 576 × 8 × 3 = 9 953 280 bits

<u>4:2:0</u> Cb, Cr resolution: 360 × 288 samples, each eight bits
Total number of bits: (720 × 576 × 8) + (360 × 288 × 8 × 2) = 4 976 640 bits

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

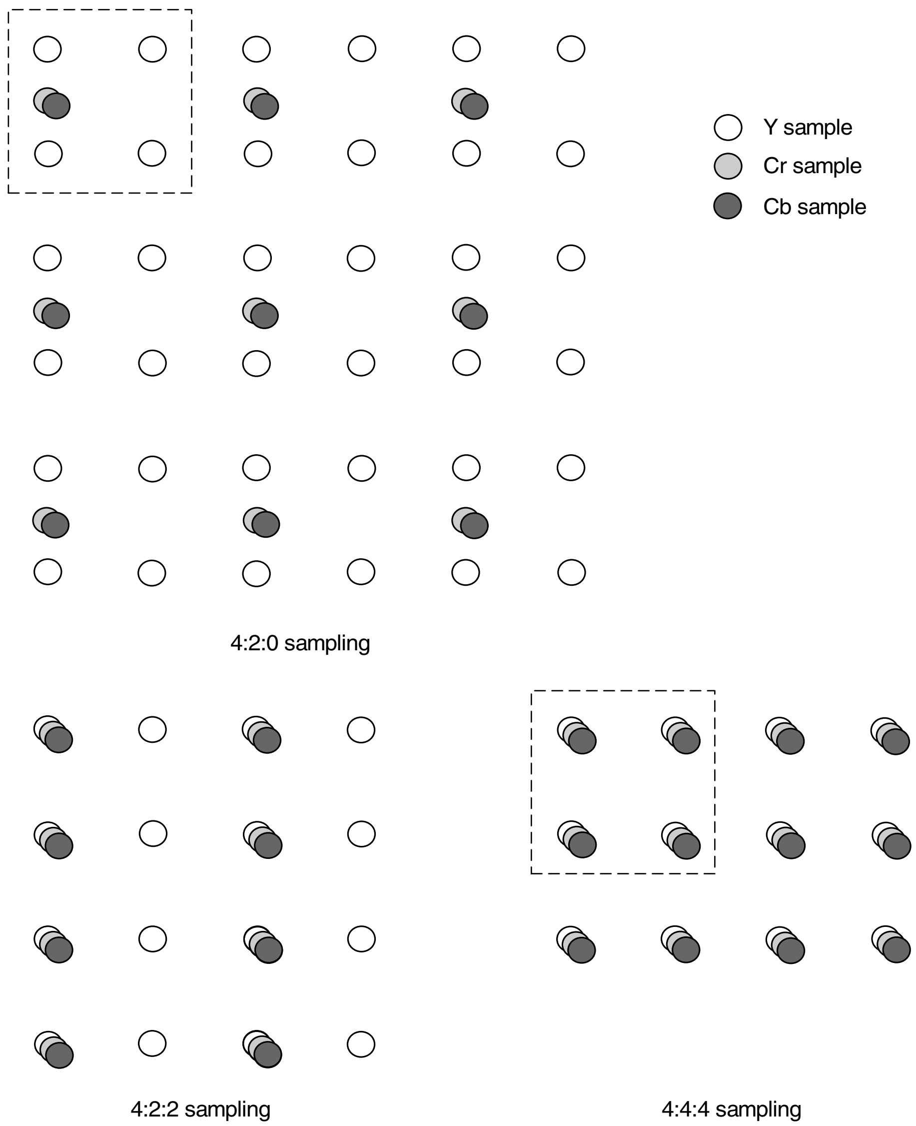

Figure 2.11 4:2:0, 4:2:2 and 4:4:4 sampling patterns (progressive)

4:2:0 sampling is sometimes described as '12 bits per pixel'. The reason for this can be seen by examining a group of four pixels (see the groups enclosed in dotted lines in Figure 2.11). Using 4:4:4 sampling, a total of 12 samples are required, four each of Y, Cb and Cr, requiring a total of $12 \times 8 = 96$ bits, an average of $96/4 = 24$ bits per pixel. Using 4:2:0 sampling, only six samples are required, four Y and one each of Cb, Cr, requiring a total of $6 \times 8 = 48$ bits, an average of $48/4 = 12$ bits per pixel.

In a 4:2:0 interlaced video sequence, the Y, Cb and Cr samples corresponding to a complete video frame are allocated to two fields. Figure 2.12 shows the method of allocating

Table 2.1 Video frame formats

Format	Luminance resolution (horiz. × vert.)	Bits per frame (4:2:0, eight bits per sample)
Sub-QCIF	128 × 96	147456
Quarter CIF (QCIF)	176 × 144	304128
CIF	352 × 288	1216512
4CIF	704 × 576	4866048

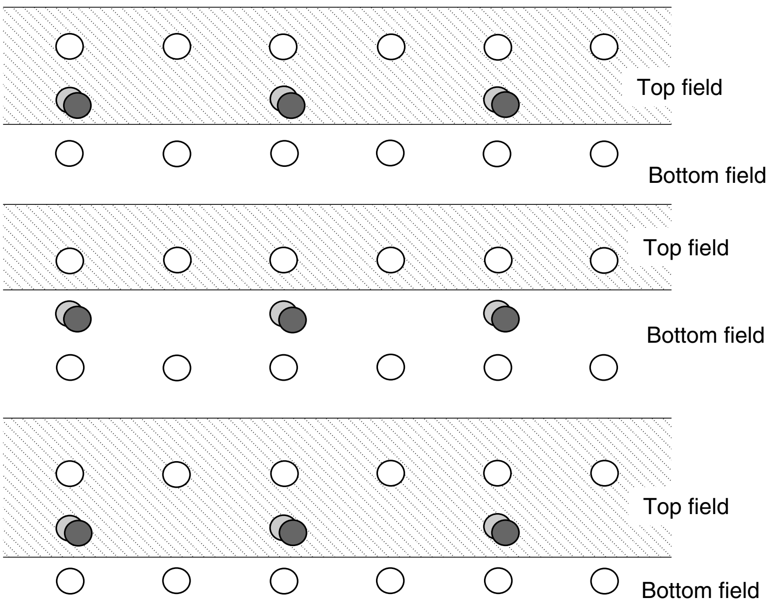

Figure 2.12 Allocaton of 4:2:0 samples to top and bottom fields

Y, Cb and Cr samples to a pair of interlaced fields adopted in MPEG-4 Visual and H.264. It is clear from this figure that the total number of samples in a pair of fields is the same as the number of samples in an equivalent progressive frame.

2.5 VIDEO FORMATS

The video compression standards described in this book can compress a wide variety of video frame formats. In practice, it is common to capture or convert to one of a set of 'intermediate formats' prior to compression and transmission. The Common Intermediate Format (CIF) is the basis for a popular set of formats listed in Table 2.1. Figure 2.13 shows the luma component of a video frame sampled at a range of resolutions, from 4CIF down to Sub-QCIF. The choice of frame resolution depends on the application and available storage or transmission capacity. For example, 4CIF is appropriate for standard-definition television and DVD-video; CIF and QCIF

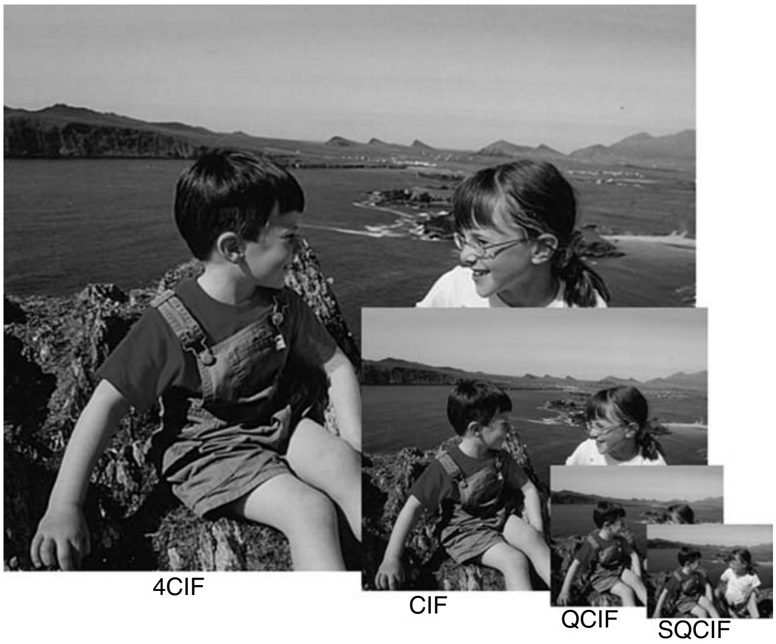

Figure 2.13 Video frame sampled at range of resolutions

are popular for videoconferencing applications; QCIF or SQCIF are appropriate for mobile multimedia applications where the display resolution and the bitrate are limited. Table 2.1 lists the number of bits required to represent one uncompressed frame in each format (assuming 4:2:0 sampling and 8 bits per luma and chroma sample).

A widely-used format for digitally coding video signals for television production is ITU-R Recommendation BT.601-5 [1] (the term 'coding' in the Recommendation title 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:Cb:Cr component signal. The parameters of the sampled digital signal depend on the video frame rate (30 Hz for an NTSC signal and 25 Hz for a PAL/SECAM signal) and are shown in Table 2.2. The higher 30 Hz frame rate of NTSC 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 to 255. Levels of 0 and 255 are reserved for synchronisation and the active luminance signal is restricted to a range of 16 (black) to 235 (white).

2.6 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

Table 2.2 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

a difficult and often imprecise art because there are so many factors that can affect the results. Visual quality is inherently *subjective* and is influenced by many factors that make it difficult to obtain a completely accurate measure of quality. For example, a viewer's opinion of visual quality can depend very much on the task at hand, such as passively watching a DVD movie, actively participating in a videoconference, communicating using sign language or trying to identify a person in a surveillance video scene. Measuring visual quality using *objective* criteria gives accurate, repeatable results but as yet there are no objective measurement systems that completely reproduce the subjective experience of a human observer watching a video display.

2.6.1 Subjective Quality Measurement

2.6.1.1 Factors Influencing Subjective Quality

Our perception of a visual scene is formed by a complex interaction between the components of the Human Visual System (HVS), the eye and the brain. The perception of visual quality is influenced by spatial fidelity (how clearly parts of the scene can be seen, whether there is any obvious distortion) and temporal fidelity (whether motion appears natural and 'smooth'). However, a viewer's opinion of 'quality' is also affected by other factors such as the viewing environment, the observer's state of mind and the extent to which the observer interacts with the visual scene. A user carrying out a specific task that requires concentration on part of a visual scene will have a quite different requirement for 'good' quality than a user who is passively watching a movie. For example, it has been shown that a viewer's opinion of visual quality is measurably higher if the viewing environment is comfortable and non-distracting (regardless of the 'quality' of the visual image itself).

Other important influences on perceived quality include visual attention (an observer perceives a scene by fixating on a sequence of points in the image rather than by taking in everything simultaneously) and the so-called 'recency effect' (our opinion of a visual sequence is more heavily influenced by recently-viewed material than older video material) [2, 3]. All of these factors make it very difficult to measure visual quality accurately and quantitatively.

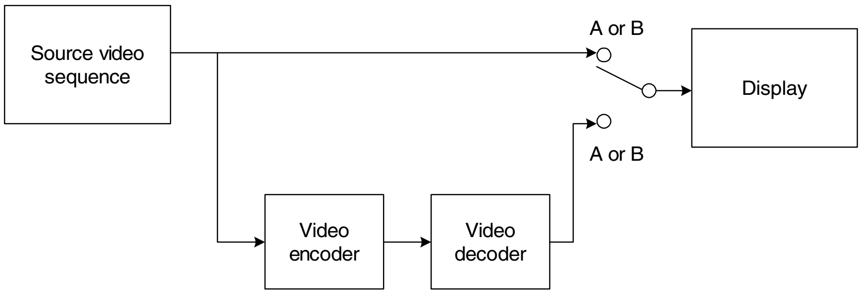

Figure 2.14 DSCQS testing system

2.6.1.2 ITU-R 500

Several test procedures for subjective quality evaluation are defined in ITU-R Recommendation BT.500-11 [4]. A commonly-used procedure from the standard is the Double Stimulus Continuous Quality Scale (DSCQS) method in which 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 'quality score' by marking on a continuous line with five intervals ranging from 'Excellent' to 'Bad'. In a typical test session, the assessor is shown a series of pairs of sequences 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. Figure 2.14 shows an experimental set-up appropriate for the testing of a video CODEC in which the original sequence is compared with the same sequence after encoding and decoding. The selection of which sequence is 'A' and which is 'B' is randomised.

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 pre-judging the impaired sequence compared with the reference sequence. At the end of the session, the scores are converted to a normalised range and the end result is a score (sometimes described as a 'mean opinion score') that indicates the *relative* quality of the impaired and reference sequences.

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

2.6.2 Objective Quality Measurement

The complexity and cost of subjective quality measurement make it attractive to be able to measure quality automatically using an algorithm. Developers of video compression and video

Figure 2.15 PSNR examples: (a) original; (b) 30.6 dB; (c) 28.3 dB

Figure 2.16 Image with blurred background (PSNR = 27.7 dB)

processing systems rely heavily on so-called objective (algorithmic) quality measures. The most widely used measure is Peak Signal to Noise Ratio (PSNR) but the limitations of this metric have led to many efforts to develop more sophisticated measures that approximate the response of 'real' human observers.

2.6.2.1 PSNR

Peak Signal to Noise Ratio (PSNR) (Equation 2.7) is measured on a logarithmic scale and depends on the mean squared error (MSE) of 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, where n is the number of bits per image sample).

$$PSNR_{dB} = 10 \log_{10} \frac{(2^n - 1)^2}{MSE} \tag{2.7}$$

PSNR can be calculated easily and quickly and is therefore a very popular quality measure, widely used to compare the 'quality' of compressed and decompressed video images. Figure 2.15 shows a close-up of 3 images: the first image (a) is the original and (b) and (c) are degraded (blurred) versions of the original image. Image (b) has a measured PSNR of 30.6 dB whilst image (c) has a PSNR of 28.3 dB (reflecting the poorer image quality).

The PSNR measure suffers from a number of limitations. PSNR requires an unimpaired original image for comparison but this may not be available in every case and it may not be easy to verify that an 'original' image has perfect fidelity. PSNR does not correlate well with subjective video quality measures such as those defined in ITU-R 500. For a given image or image sequence, high PSNR usually indicates high quality and low PSNR usually indicates low quality. However, a particular value of PSNR does not necessarily equate to an 'absolute' subjective quality. For example, Figure 2.16 shows a distorted version of the original image from Figure 2.15 in which only the background of the image has been blurred. This image has a PSNR of 27.7 dB relative to the original. Most viewers would rate this image as significantly better than image (c) in Figure 2.15 because the face is clearer, contradicting the PSNR rating. This example shows that PSNR ratings do not necessarily correlate with 'true' subjective quality. In this case, a human observer gives a higher importance to the face region and so is particularly sensitive to distortion in this area.

2.6.2.2 Other Objective Quality Metrics

Because of the limitations of crude metrics such as PSNR, there has been a lot of work in recent years to try to develop a more sophisticated objective test that more closely approaches subjective test results. Many different approaches have been proposed [5, 6, 7] but none of these has emerged as a clear alternative to subjective tests. 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) aim to develop standards for objective video quality evaluation [8]. 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 ten competing systems were tested under identical conditions. Unfortunately, none of the ten proposals was considered suitable for standardisation and VQEG are completing a second round of evaluations in 2003. Unless there is a significant breakthrough in automatic quality assessment, the problem of accurate objective quality measurement is likely to remain for some time to come.

2.7 CONCLUSIONS

Sampling analogue video produces a digital video signal, which has the advantages of accuracy, quality and compatibility with digital media and transmission but which typically occupies a prohibitively large bitrate. Issues inherent in digital video systems include spatial and temporal resolution, colour representation and the measurement of visual quality. The next chapter introduces the basic concepts of video compression, necessary to accommodate digital video signals on practical storage and transmission media.

2.8 REFERENCES

1. 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.

2. N. Wade and M. Swanston, *Visual Perception: An Introduction,* 2nd edition, Psychology Press, London, 2001.
3. R. Aldridge, J. Davidoff, D. Hands, M. Ghanbari and D. E. Pearson, Recency effect in the subjective assessment of digitally coded television pictures, *Proc. Fifth International Conference on Image Processing and its Applications*, Heriot-Watt University, Edinburgh, UK, July 1995.
4. Recommendation ITU-T BT.500-11, Methodology for the subjective assessment of the quality of television pictures, ITU-T, 2002.
5. 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*, **2668**, San Jose, 1996.
6. H. Wu, Z. Yu, S. Winkler and T. Chen, Impairment metrics for MC/DPCM/DCT encoded digital video, *Proc. PCS01*, Seoul, April 2001.
7. K. T. Tan and M. Ghanbari, A multi-metric objective picture quality measurement model for MPEG video, *IEEE Trans. Circuits and Systems for Video Technology*, **10** (7), October 2000.
8. http://www.vqeg.org/ (Video Quality Experts Group).