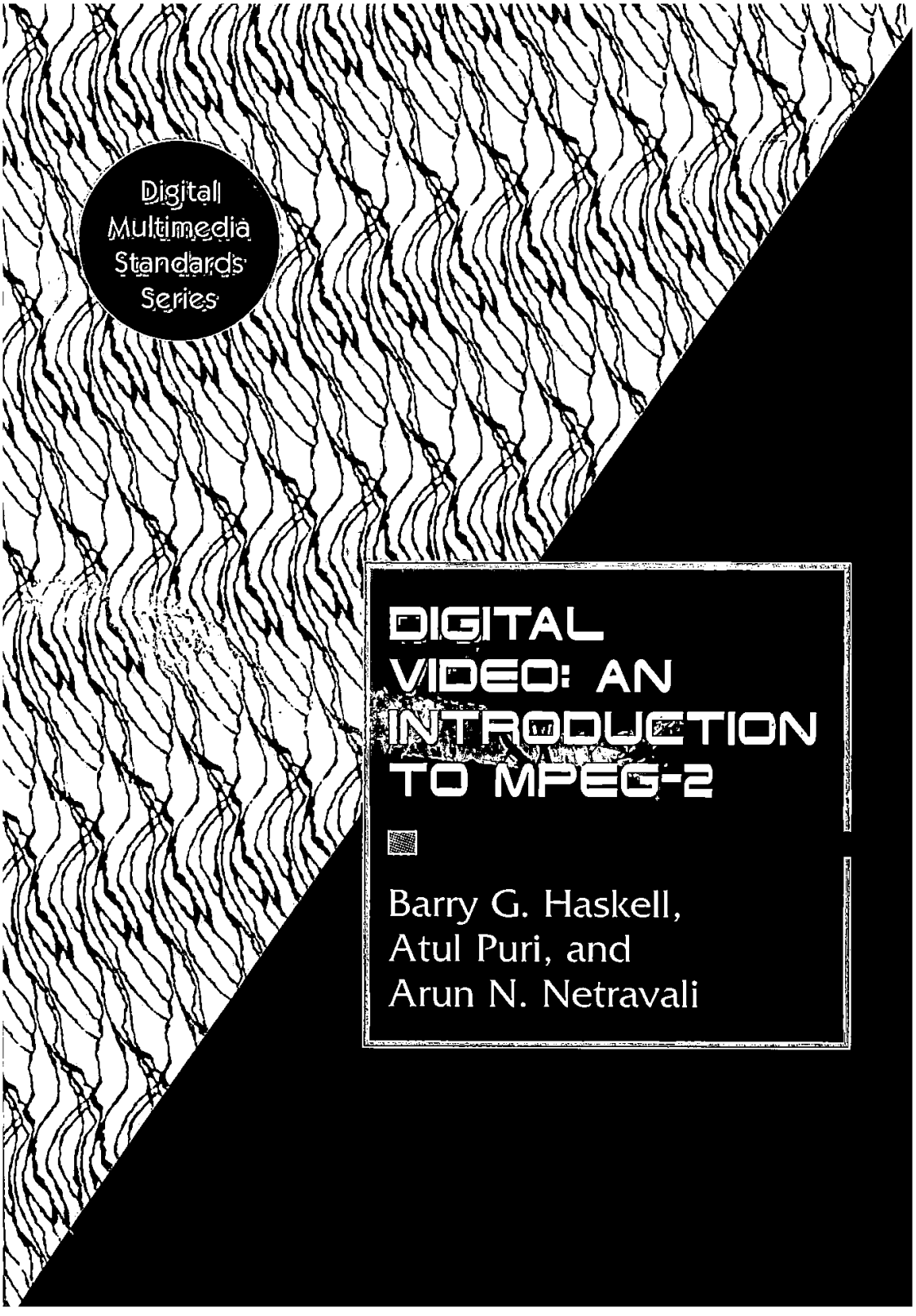

Cover design: Curtis Tow Graphics

Printed in the United States of America

Chapman & Hall
115 Fifth Avenue
New York, NY 10003

Chapman & Hall
2-6 Boundary Row
London SE1 8HN
England

Thomas Nelson Australia
102 Dodds Street
South Melbourne, 3205
Victoria, Australia

Chapman & Hall GmbH
Postfach 100 263
D-69442 Weinheim
Germany

International Thomson Editores
Campos Eliseos 385, Piso 7
Col. Polanco
11560 Mexico D.F
Mexico

International Thomson Publishing–Japan
Hirakawacho-cho Kyowa Building, 3F
1-2-1 Hirakawacho-cho
Chiyoda-ku, 102 Tokyo
Japan

International Thomson Publishing Asia
221 Henderson Road #05-10
Henderson Building
Singapore 0315

1 2 3 4 5 6 7 8 9 10 XXX 01 00 99 98 97

Library of Congress Cataloging-in-Publication Data

Haskell, Barry G.
Digital video : an introduction to MPEG-2 / Barry G. Haskell, Atul Puri, and Arun N. Netravali.
p. cm.
Includes bibliographical references and index.
ISBN 0-412-08411-2
1. Digital video. 2. Video compression -- Standards. 3. Coding theory. I. Puri, Atul. II. Netravali, Arun N. III. Title.
TK6680.5.H37 1996
621.388'33--dc20 96-14018
CIP

British Library Cataloguing in Publication Data available

"Digital Video: An Introduction to MPEG-2" is intended to present technically accurate and authoritative information from highly regarded sources. The publisher, editors, authors, advisors, and contributors have made every reasonable effort to ensure the accuracy of the information, but cannot assume responsibility for the accuracy of all information, or for the consequences of its use.

To order this or any other Chapman & Hall book, please contact **International Thomson Publishing, 7625 Empire Drive, Florence, KY 41042.** Phone: (606) 525-6600 or 1-800-842-3636. Fax: (606) 525-7778. e-mail: order@chaphall.com.

For a complete listing of Chapman & Hall titles, send your request to **Chapman & Hall, Dept. BC, 115 Fifth Avenue, New York, NY 10003.**

capacity, the features required by the application, and the affordability of the hardware required for implementation of the compression algorithm (encoder as well as decoder). This section describes some of the issues in designing the compression system.

1.3.1 Quality

The quality of presentation that can be derived by decoding the compressed multimedia signal is the most important consideration in the choice of the compression algorithm. The goal is to provide near-transparent coding for the class of multimedia signals that are typically used in a particular service.

The two most important aspects of video quality are *spatial* resolution and *temporal* resolution. Spatial resolution describes the clarity or lack of blurring in the displayed image, while temporal resolution describes the smoothness of motion.

Motion video, like film, consists of a certain number of frames per second to adequately represent motion in the scene. The first step in digitizing video is to partition each frame into a large number of *picture elements*, or *pels** for short. The larger the number of pels, the higher the spatial resolution. Similarly, the more frames per second, the higher the temporal resolution.

Pels are usually arranged in rows and columns, as shown in Fig. 1.2. The next step is to measure the brightness of the red, green, and blue color components within each pel. These three color brightnesses are then represented by three binary numbers.

1.3.2 Uncompressed versus Compressed Bitrates

For entertainment television in North America and Japan, the NTSC† color video image has 29.97 frames per second; approximately 480 visi-

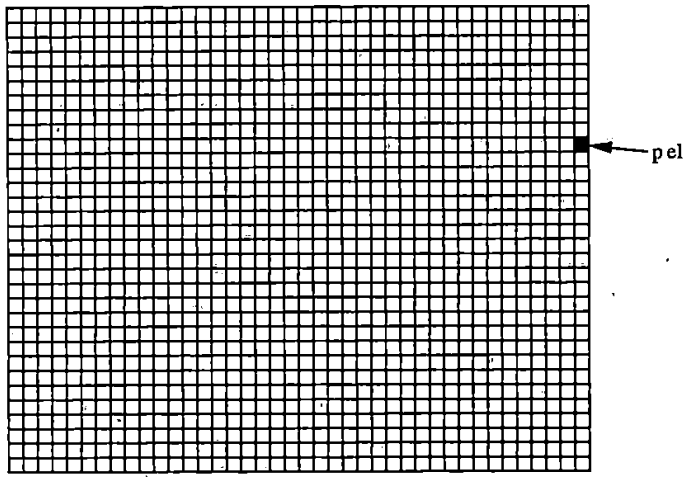

Fig. 1.2 Frames to be digitized are first partitioned into a large number of *picture elements*, or *pels* for short, which are typically arranged in rows and columns. A row of pels is called a *scan line.*

*Some authors prefer the slightly longer term *pixels.*

†National Television Systems Committee. It standardized composite color TV in 1953.

Table 1.1 Raw and compressed bitrates for film, NTSC, PAL, HDTV, videophone, stereo audio, and five-channel audio. Compressed bitrates are typically lower for slow moving drama than for fast live-action sports.

	Video Resolution (pels × lines × frames/s)	Uncompressed Bitrate (RGB)	Compressed Bitrate
Film (USA and Japan)	(480 × 480 × 24Hz)	133 Mbits/s	3 to 6 Mbits/s
NTSC video	(480 × 480 × 29.97Hz)	168 Mbits/s	4 to 8 Mbits/s
PAL video	(576 × 576 × 25Hz)	199 Mbits/s	4 to 9 Mbits/s
HDTV video	(1920 × 1080 × 30Hz)	1493 Mbits/s	18 to 30 Mbits/s
HDTV video	(1280 × 720 × 60Hz)	1327 Mbits/s	18 to 30 Mbits/s
ISDN videophone (CIF)	(352 × 288 × 29.97Hz)	73 Mbits/s	64 to 1920 kbits/s
PSTN videophone (QCIF)	(176 × 144 × 29.97Hz)	18 Mbits/s	10 to 30 kbits/s
Two-channel stereo audio		1.4 Mbits/s	128 to 384 kbits/s
Five-channel stereo audio		3.5 Mbits/s	384 to 968 kbits/s

ble scan lines per frame; and requires approximately 480 pels per scan line in the red, green, and blue color components. If each color component is coded using 8 bits (24 bits/pel total), the bitrate produced is ≈168 Megabits per second (Mbits/s). Table 1.1 shows the raw uncompressed bitrates for film, several video formats including PAL,* HDTV,† and videophone, as well as a few audio formats.

We see from Table 1.1 that uncompressed bitrates are all very high and are not economical in many applications. However, using the MPEG compression algorithms these bitrates can be reduced considerably. We see that MPEG is capable of compressing NTSC or PAL video into 3 to 6 Mbits/s with a quality comparable to analog CATV and far superior to VHS videotape.

1.3.3 Bitrates Available on Various Media

A number of communication channels are available for multimedia transmission and storage. PSTN modems can operate up to about 30 kilobits/second (kbits/s), which is rather low for generic video, but is acceptable for personal videophone. ISDN‡ ranges from 64 kbits/s to 2 Mbits/s, which is often acceptable for moderate quality video.

Table 1.2 Various media and the bitrates they support.

Medium	Bitrate
PSTN modems	Up to 30 kbits/s
ISDN	64 to 1920 kbits/s
LAN	10 to 100 Mbits/s
ATM	135 Mbits/s or more
CD-ROM (normal speed)	1.4 Mbits/s
Digital video disk	9 to 10 Mbits/s
Over-the-air video	18 to 20 Mbits/s
CATV	20 to 40 Mbits/s

LANs§ range from 10 Mbits/s to 100 Mbits/s or more, and ATM** can be many hundreds of Mbits/s. First generation CD-ROMs†† can handle 1.4 Mbits/s at normal speed and often have double or triple speed options. Second generation digital video disks operate at 9 to 10 Mbits/s depending on format. Over-the-air or CATV channels can carry 20 Mbits/s to 40 Mbits/s, depending on distance and interference from other transmitters. These options are summarized in Table 1.2.

1.4 COMPRESSION TECHNOLOGY

In picture coding, for example, compression techniques have been classified according to the groups

*Phase Alternate Line, used in most of Europe and elsewhere.
†High-Definition Television. Several formats are being used.
‡Integrated Services Digital Network. User bitrates are multiples of 64 kbits/s.
§Local Area Networks such as Ethernet, FDDI, etc.
**Asynchronous Transfer Mode. ATM networks are packetized.
††Compact Disk–Read Only Memory.

time varying as it can adequately represent the motion in a scene. Video can be thought of as comprised of a sequence of still pictures of a scene taken at various subsequent intervals in time. Each still picture represents the distribution of light energy and wavelengths over a finite size area and is expected to be seen by a human viewer. Imaging devices such as cameras allow the viewer to look at the scene through a finite size rectangular window. For every point on this rectangular window, the light energy has a value that a viewer perceives which is called its intensity. This perceived intensity of light is a weighted sum of energies at different wavelengths it contains.

The lens of a camera focuses the image of a scene onto a photosensitive surface of the imager, which converts optical signals into electrical signals. There are basically two types of video imagers: (1) tube-based imagers such as vidicons, plumbicons, or orthicons, and (2) the more recent, solid-state sensors such as charge-coupled devices (CCD). The photosensitive surface of the tube imager is typically scanned with an electron beam or other electronic methods, and the two-dimensional optical signal is converted into an electrical signal representing variations of brightness as variations in voltage. With CCDs the signal is read out directly.

5.1.1 Raster Scan

Scanning is a form of sampling of a continuously varying two-dimensional signal. Raster scanning[1–3] is the most commonly used spatial sampling representation. It converts a two-dimensional image intensity into a one-dimensional waveform. The optical image of a scene on the camera imager is scanned one line at a time from left to right and from top to bottom. The imager basically converts the brightness variations along each line to an electrical signal. After a line has been scanned from left to right, there is a retracing period and scanning of next line again starts from left to right. This is shown in Fig. 5.1.

In a television camera, typically an electron beam scans across a photosensitive target upon which light from a scene is focused. The scanning spot itself is responsible for local spatial averaging of the image since the measured intensity at a time

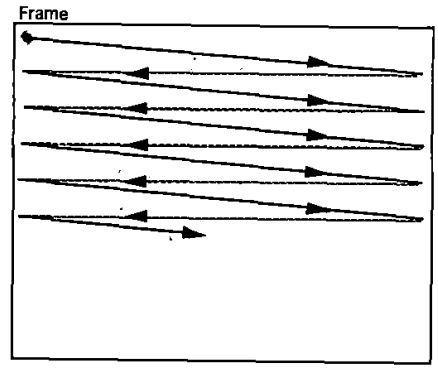

Fig. 5.1 Raster scan of an image

is actually a weighted average over the area of the spot. With the use of electron beam scanning, however, it is very hard to control the scanning spot filter function, with the result that the image is sometimes not sufficiently prefiltered. Optical scanning devices that use a beam of light are significantly better in this regard. Prefiltering can also be done in most cameras by a slight defocusing of the optical lens. With raster scanned display devices, filtering in the vertical direction is done by the viewer's eye or by adjustment of the scanning beam or optical components.

The smallest detail that can be reproduced in a picture is about the size of a picture element, typically referred to as a *pixel* (in computer terminology) or a *pel* (in video terminology). A sampled TV raster, for example, can be described as a grid of pels as shown in Fig. 1.2. The detail that can be represented in the vertical direction is limited by the number of scan lines. Thus, some of the detail in vertical resolution is lost as the result of raster scanning fall between the scan lines and cannot be represented accurately. Measurements indicate that only about 70% of the vertical detail is actually represented by scan lines. Similarly, some of the detail in the horizontal direction is lost owing to sampling of each scan line to form pels.

5.1.2 Interlace Raster Scan

The choice of number of scan lines involves a tradeoff among contradictory requirements of

bandwidth, flicker, and resolution. Interlaced[1–3,5–6] scanning tries to achieve these tradeoffs by using frames that are composed of two *fields* sampled at different times, with lines of the two fields interleaved, such that two consecutive lines of a frame belong to alternate fields. This represents a vertical–temporal tradeoff in spatial and temporal resolution. For instance, this allows slow-moving objects to be perceived with higher vertical detail while fast-moving objects are perceived at a higher temporal rate, albeit at close to half vertical resolution. Thus the characteristics of the eye are well matched because, with slower motion, the human visual system is able to perceive the spatial details, whereas with faster motion spatial details are not as easily perceived. This interlace raster scan is shown in Fig. 5.2.

From a television transmission and display aspect, the interlace raster scan represents a reasonable tradeoff in the bandwidth necessary and the amount of flicker generated. It is worth noting that the tradeoffs that may have been satisfactory for television may not be satisfactory for computer displays, owing to the closeness of the screen to the viewer and the type of material typically displayed, that is, text and graphics. Thus, if interlaced TV displays were to be used with computers, the result would be an annoying large area flicker, interline flicker, line crawling, and other problems. To avoid these problems, computer displays use noninterlaced (also called progressive or sequential) displays with refresh rates of higher than 60 frames/s, typically 72 frames/s.

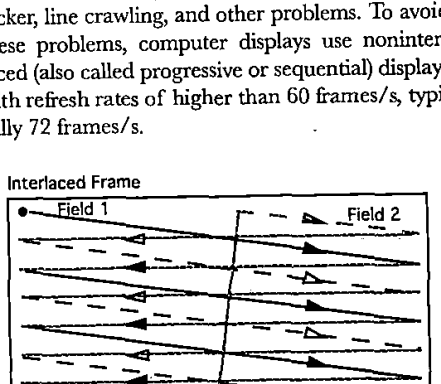

Fig. 5.2 Interlaced raster scan of an image

Table 5.1 Equivalent Line Number of Film and TV Image Formats

Image Formats	Equivalent Line Number Range
NTSC TV	175 to 310
16 mm Film	200 to 300
35 mm Film	350 to 450
HDTV	300 to 515

5.1.3 Film versus Television

Originally, television was designed to capture the same picture definition as that of 16-mm film. Actually, the resolution of 16-mm film is significantly higher than that of normal TV, but the perceived sharpness is about the same owing to a better temporal response of normal TV. To compare the performance of quite different display systems, the concept of equivalent line number is useful. In Table 5.1 we show the equivalent line numbers for film and television.

Based on what we have said so far, it may appear that a television system can achieve the performance of film. In practice, however, there are too many ways in which the performance of a TV system can become degraded. In this context, it is worth noting that *frequency response,* which is the variation in peak-to-peak amplitude of a pattern of alternate dark and bright lines of varying widths, can be used to establish the performance of an imaging system. In reality, TV and film images never really appear the same owing to the difference in shape of the overall frequency response. For instance, the frequency response of TV drops much faster, whereas the frequency response of film decays more gradually.

The response of the human eye over time is generally continuous,[1–3] that is, nerve impulses travel from the eye to the brain at the rate of about 1000 times per second. Fortunately, television does not need to provide new images of a scene at this rate as it can exploit the property of persistence of vision to retain the sensation of an image for a time interval even though the actual image may have been removed. Both TV and motion pictures employ the same principle—a scene is portrayed by showing a sequence of slightly different images (frames) taken over time to replicate the motion in a scene. For motion in a

scene to appear realistic, the rate of imaging of frames should be high enough to capture the motion without jerkiness, and the rate of display should be high enough so that the captured motion can be replicated smoothly. The display rate must also be high enough so that persistence of vision works properly. One of the nuances of human vision is that if the images being presented to it are bright, the persistence of vision is actually shorter, thus requiring a higher repetition rate for display.

In the early days of motion pictures,[1–3] it was found that for smooth motion rendition, a minimum frame rate of 15 per second (15 frames/s) was necessary. Thus old movie equipment, including home movie systems, use 16 frames/s. However, because of a difficulty in representing action in cowboy movies of those days, the motion picture industry eventually settled on the rate of 24 frames/s, which is used to this day. Note that although 24 frames/s movies use 50% more film than 16 frames/s movies, the former was adopted as a standard as it was considered worthwhile. As movie theater projection systems became more powerful, the nuance of eye mentioned earlier, which requires a higher frame rate, began to come into play, resulting in the appearance of flicker. A solution that did not require a change in the camera frame rates of movies was to use a 2 blade rotating shutter in the projector, synchronized to the film advance, which could briefly cut the light and result in double the displayed frame rate to 48 frames/s. Today, even a 3 blade shutter to effectively generate a display rate of 72 frames/s is sometimes used since projection systems have gotten even more brighter.

The frame rate used for monochrome TV in Europe was in fact chosen to be slightly higher than that of film at 25 frames/s, in accordance with electrical power there, which uses 50 cycles/s (Hz) rate. Television in Europe uses systems such as PAL and SECAM, each employing 25 frames/s. It is common practice in Europe to run 24 frames/s movies a little faster at 25 frames/s. The PAL and SECAM based TV systems are currently in widespread use in many other parts of the world as well.

The frame rate of monochrome TV in the United States was chosen as 30 frames/s, primarily to avoid interference to circuits used for scanning and signal processing by 60-Hz electrical power. In 1953, color was added to monochrome by migration to an NTSC system in which the 30 frames/s rate of monochrome TV was changed by 0.1 percent to 29.97 frames/s owing to the requirement of precise separation of the video signal carrier from the audio signal carrier. Incidentally Japan and many other countries also employ the NTSC TV system.

For reasons other than those originally planned, the frame rate chosen in the NTSC TV system turned out to be a good choice since there is a basic difference in the way TV is viewed as compared to film. While movies are generally viewed in dark surroundings that reduce the human eye's sensitivity to flicker and the smoothness of motion, TV is viewed under condition of surround lighting, in which case the human visual system is more sensitive to flicker and the smoothness of motion. Thus, TV requires a higher frame rate than movies to prevent the appearance of flicker and motion-related artifacts.

One very important point is that unlike TV, which uses interlace, film frames are sequential (also called noninterlaced or progressive).

5.1.4 Image Aspect Ratio (IAR)

The image aspect ratio is generally defined as the ratio of picture width to height* and impacts the overall appearance of the displayed image. For example, the IAR of normal TV images is 4/3, meaning that the image width is 1.33 times the image height. This value was adopted for TV, as this format was already used and found acceptable in the film industry at that time, prior to 1953. However, since then the film industry has migrated to wide-screen formats with IARs of 1.85 or higher. Since subjective tests on viewers show a significant preference for a wider format than that used for normal TV, HDTV uses an IAR of 1.78, which is quite close to that of the wide-screen film format. Currently, the cinemascope film format provides one of the highest IARs of 2.35. Table 5.2

*Some authors, including MPEG, use height/width.

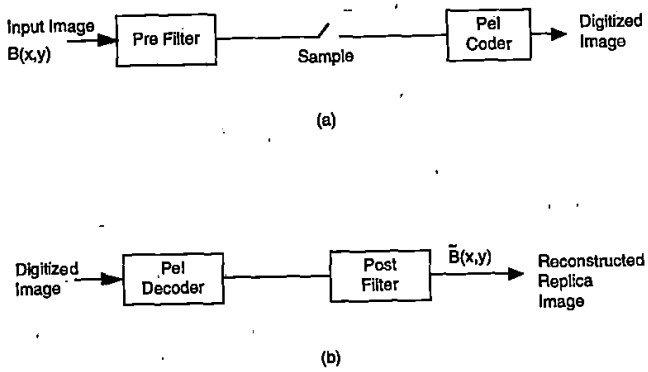

Fig. 6.1 Spatial sampling of images generally requires prefiltering prior to sampling, as shown in (a), and postfiltering following reconstruction, as shown in (b). Inadequate prefiltering leads to prealiasing, whereas inadequate post-filtering causes post-aliasing.

function, the filtered image can be computed as

$$\bar{B}(x,y) = \int_{y-\delta}^{y+\delta}\int_{x-\delta}^{x+\delta} KB(w,z)\exp\Big(-a(x-w)^2 - b(y-z)^2\Big)\,dw\,dz$$

where the intensity at each point is replaced by a weighted average over a small surrounding neighborhood. Such prefiltering could also be implemented simply by proper design of the optical components in the camera imaging system.

The postfilter is an interpolation filter. A commonly used interpolation filter is the so-called cubic B-spline $h_d(\mu, v) = S(\mu, \delta_x)S(v, \delta_y)$ where the horizontal and vertical pel spacings are δ_x and δ_y, respectively, and

$$S(z,\Delta) = \begin{cases} \frac{2}{3} - \left|\frac{z}{\Delta}\right|^2 + \frac{1}{2}\left|\frac{z}{\Delta}\right|^3 & \text{for } \left|\frac{z}{\Delta}\right| \le 1 \\ \frac{4}{3} - 2\left|\frac{z}{\Delta}\right| + \left|\frac{z}{\Delta}\right|^2 - \frac{1}{6}\left|\frac{z}{\Delta}\right|^3 & \text{for } 1 \le \left|\frac{z}{\Delta}\right| \le 2 \end{cases}$$

For positive pel values, the B-spline interpolation is always positive. The B-spline filter essentially smoothes the spatially discrete pels of an image and preserves the continuity of the first and the second derivatives in both x and y dimensions.

6.2.1 Raster Scanning

Raster scanning is a process used to convert a three-dimensional image intensity, $B(x, y, t)$, into a one-dimensional television signal waveform. First, the intensity, $B(x, y, t)$, is sampled many times per second to create a sequence of frames. Then, within each frame, sampling is done vertically to create scan lines. Scanning proceeds sequentially, left to right and from top to bottom. In a tube-type television camera, an electron beam scans across an electrically photosensitive target upon which the image is focused. At the other end of the television chain there is usually a cathode ray tube (CRT) display, in which the electronic beam scans and lights up the picture elements in proportion to the intensity, $B(x, y, t)$. While it is convenient to think of all the samples of a frame as occurring at the same time (similar to film), the scanning in a camera or display results in every sample corresponding to a different point in time.

Since the electron beam has some thickness, the scanning itself is partly responsible for local spatial averaging of the image over the area of the electron beam. This occurs both at the camera as well as the display. At the camera, as the beam scans the camera target, it converts the photons integrated (over a frame time) at each target point into an electrical signal. This results in both the spatial and temporal averaging of light and is commonly referred to as *camera integration*. Higher resolution

camera systems have a much finer electron beam, but in that case, owing to averaging over a smaller area, the amount of light intensity as well as signal-to-noise ratio is reduced. At the display, the beam thickness may simply blur the image or a thinner beam may result in visible scan lines.

There are two types of scanning: *progressive* (also called *sequential*) and *interlaced*. In progressive scanning, within each frame all the raster lines are scanned from top to bottom. Therefore, all the vertically adjacent scan lines are also temporally adjacent and are highly correlated even in the presence of rapid motion in the scene. Film can be thought of as naturally progressively scanned, since all the lines were originally exposed simultaneously, so a high correlation between adjacent lines is guaranteed. Almost all computer displays are sequentially scanned.

In analog interlaced scanning (see Fig. 6.2) all the odd-numbered lines in the entire frame are scanned first, and then the even-numbered lines are scanned. This process results in two distinct images per frame, representing two distinct samples of the image sequence at different points in time. The set of odd-numbered lines constitute the *odd-field*, and the even numbered lines make up the *even-field*.* All the current analog TV systems (NTSC, PAL, SECAM) use interlaced scanning.

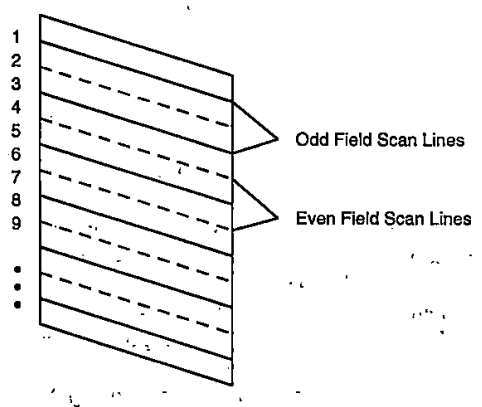

Fig. 6.2 Scanning pattern for 2:1 interlace. Odd-numbered lines (field 1) are scanned and displayed first, followed by the even-numbered lines (field 2). In this way, low spatial frequencies are displayed twice as often as high spatial frequencies, thus better matching the characteristics of human vision. Compared to scanning both the odd and even lines at the field rate, this also reduces the line rate (and bandwidth) by a factor of two.

*MPEG has a different nomenclature, as we shall see.

One of the principal benefits of interlaced scanning is to reduce the scan rate (or the bandwidth) in comparison to the bandwidth required for sequential scan. Interlace allows for a relatively high field rate (a lower field rate would cause flicker), while maintaining a high total number of scan lines in a frame (a lower number of lines per frame would reduce resolution on static images). Interlace cleverly preserves the high detailed visual information and at the same time avoids visible large area flicker at the display due to insufficient temporal postfiltering by the human eye. Interlace scanning often shows small flickering artifacts in scenes with sharp detail and may have poor motion rendition for fast vertical motion of small objects.

If the height/width ratio of the TV raster is equal to the number of scan lines/number of samples per line, the array is referred to as having *square pels*, that is, the pel centers all lie on the corners of squares. This facilitates digital image processing as well as computer synthesis of images. In particular, geometrical operations such as rotation, or filters for antialiasing can be implemented without shape distortions.

6.2.2 Color Images

Light is a subset of the electromagnetic energy. The visible spectrum wavelengths range from 380 to 780 nanometers. Thus, visible light can be specified completely at a pel by its wavelength distribution $\{S(\lambda)\}$. This radiation, on impinging on the human retina, excites predominantly three different receptors that are sensitive to wavelengths at or near 445 (called blue), 535 (called green), and 570 (called red) nanometers. Each type of receptor measures the energy in the incident light at wavelengths near its dominant wavelength. The three resulting energy values uniquely specify each visually distinct color, **C**.

7

Motion Compensation Modes in MPEG

We saw in the last chapter that exploitation of temporal redundancy commonly uses motion compensation. After the removal of temporal redundancy, spatial redundancy is reduced by transformation to the frequency domain, for example, by the Discrete Cosine Transform (DCT), followed by quantization to force many of the transform coefficients to zero.

MPEG-1[2] was developed specifically for coding progressive scanned video in the range of 1 to 2 Mbits/s. MPEG-2[4] aims at higher bitrates and contains all of the progressive coding features of MPEG-1. In addition, MPEG-2 has a number of techniques for coding interlaced video.

7.1 CODING OF PROGRESSIVE SCANNED VIDEO

The primary requirement of the MPEG video standards is that for a given bitrate they should achieve the highest possible quality of decoded video. As well as producing good picture quality under normal play conditions, some applications have additional requirements. For instance, multimedia applications may require the ability to randomly access and decode any single video picture* in the bitstream. Also, the ability to perform a fast search of video in the bitstream, both forward and backward, is extremely desirable. It is also useful to be able to edit compressed bitstreams directly while maintaining decodability. Finally, a variety of video formats and image sizes should be supported.

Both spatial and temporal redundancy reduction are needed for the high compression requirements of MPEG. Some techniques used by MPEG were described in the previous chapter.

7.1.1 Exploiting Spatial Redundancy of Progressive Video

Because video is a sequence of still images, it is possible to achieve some compression using techniques similar to JPEG.[1] Such methods of compression are called *Intraframe* coding techniques, wherein each picture of the video is individually and independently compressed or encoded. Intraframe coding exploits the spatial redundancy that exists between adjacent pels of a picture. Pictures coded using only intraframe coding are called *I-pictures.*

As in JPEG, the MPEG intraframe video coding algorithms employ a Block-based two-dimensional DCT. An I-picture is first divided into 8 × 8

*Frames and pictures are synonymous in MPEG-1, whereas MPEG-2 has both frame-pictures and field-pictures.

Blocks of pels, and the two-dimensional DCT is then applied independently on each Block. This operation results in an 8 × 8 Block of DCT coefficients in which most of the energy in the original (pel) Block is typically concentrated in a few low-frequency coefficients. The coefficients are then quantized and transmitted as in JPEG.

7.1.2 Exploiting Temporal Redundancy of Progressive Video

Many of the interactive requirements can be satisfied by intraframe coding. However, for typical video signals the image quality that can be achieved by intraframe coding alone at the low bitrates desired is usually not sufficient.

Temporal redundancy results from a high degree of correlation between adjacent pictures. The H.261 and MPEG algorithms[3,5–7] exploit this redundancy by computing an interframe difference signal called the *Prediction Error.* In computing the prediction error, the technique of motion compensation is employed to correct the prediction for motion. As in H.261, the *Macroblock* (MB) approach is adopted for motion compensation in MPEG. In unidirectional motion estimation, called *Forward Prediction,* a *Target MB* in the picture to be encoded is matched with a set of displaced macroblocks of the same size in a past picture called the *Reference* picture. As in H.261, the Macroblock in the Reference picture that *best matches* the Target Macroblock is used as the *Prediction MB.** The prediction error is then computed as the difference between the Target Macroblock and the Prediction Macroblock. This operation is illustrated in Fig. 7.1.

The position of this best matching Prediction Macroblock is indicated by a *Motion Vector* (MV) that describes the horizontal and vertical displacement† from the Target MB to the Prediction MB. Unlike H.261, MPEG MVs are computed to the nearest half pel of displacement, using bilinear interpolation to obtain brightness values between pels. Also, using values of past transmitted MVs, a *Prediction Motion Vector* (PMV) is computed. The difference MV − PMV is then encoded and transmitted along with the pel prediction error signal. Pictures coded using Forward Prediction are called *P-pictures.*

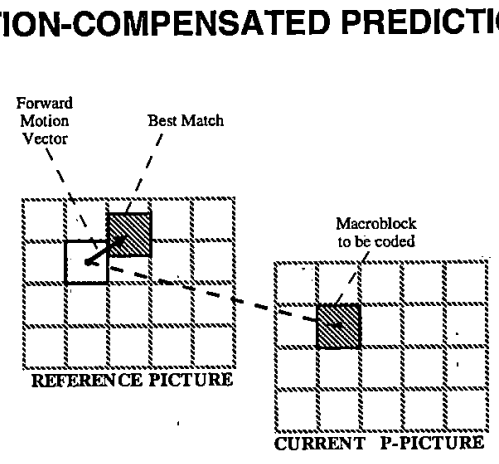

Fig. 7.1 Motion-Compensation using Unidirectional prediction. A displaced macroblock in a previous reference picture is used as the prediction of the moving-area macroblock in the current P-picture. The prediction error signal is then coded and transmitted.

*Prediction MBs do not, in general, align with coded MB boundaries in the Reference picture.

†Positive horizontal and vertical displacements imply a rightward and downward shift, respectively, from the Target MB to the Prediction MB.

transfer rates are significantly higher than a single stream's data rate, a small buffer should suffice for conventional file and operating systems support for several media streams.

The real challenge is in supporting a large number of streams simultaneously. ITV servers must process retrieval requests from several CPEs simultaneously. Even when multiple CPEs access the same file (such as a popular movie), different streams might access different parts of the file at the same time. Since disk I/O rates significantly exceed those of single streams, a larger number of streams can be supported by multiplexing a disk head among several streams. This requires careful disk scheduling so that each buffer is adequately served. In addition, new streams should not be admitted unless they can be properly served along with the existing ones.

13.2.2.1 Disk Scheduling

Database systems employ disk scheduling algorithms, such as "first come, first serve" and "shortest seek time first," to reduce seek time and rotational latency, to achieve high throughput, and to enforce fairness for each stream. However, continuous media constraints usually require a modification of traditional disk-scheduling algorithms so they can be used for multimedia servers.

One commonly used disk scheduling algorithm, called the *Scan-EDF* (Earliest Deadline First), services the requests with earliest deadlines first. However, when several requests have the same deadline, their respective blocks are accessed with a shortest-seek-time algorithm. Scan-EDF becomes more effective as the number of requests having the same deadline increase. This happens more frequently as the ITV server services a larger number of CPEs.

If data are retrieved from disks into buffers in rounds, a commonly used disk scheduling algorithm is *C-LOOK* (Circular LOOK), which steps the disk head from one end of the disk to the other, retrieving data as the head reaches one of the requested blocks. When the head reaches the last block that needs to be retrieved within a round, the disk head is moved back to the beginning of the disk without retrieving any data, after which the process repeats.

13.2.2.2 Buffer Management

Most ITV servers process multimedia stream requests from the CPE in rounds. During a round, the amount of data retrieved by the buffer could be made equal to the amount consumed by the CPE. This means that, on a round-by-round basis, data production never lags display, and there is never a net decrease in the delay of buffered data. Therefore, such algorithms are referred to as buffer-conserving. A non-buffer-conserving algorithm would allow retrieval to fall behind display in one round and compensate for it over the next several rounds. In this scenario, it becomes more difficult to guarantee that no starvation of the display will ever take place.

For continuous display, a sufficient number of blocks must be retrieved for each CPE-client during a round so that starvation can be prevented for the round's entire duration. To determine this number, the server must know the maximum duration of a round. As round length depends on the number of blocks retrieved for each stream, unnecessarily long reads must be avoided. To minimize the round length, the number of blocks retrieved for each CPE during each round should be proportional to the CPE's display rate.

13.2.2.3 Admission Control

To satisfy the CPE's real-time requirements, an ITV server must control admission of new CPEs so that existing CPEs can be serviced adequately. Strict enforcement of all real-time deadlines for each CPE may not be necessary since some applications can tolerate occasional missed deadlines. A few lost video frames or a glitch in the audio can occasionally be tolerated, especially if such tolerance reduces the overall cost of service. An ITV server might accommodate additional streams through an admission control algorithm that exploits the statistical variation in media access times, compression ratios, and other time-varying factors.

In admitting new CPEs, ITV servers may offer three categories of service:

- *Deterministic.* Real time requirements are guaranteed. The admission control algorithm must consider worst-case bounds in admitting new CPEs.

- *Statistical.* Deadlines are met only with a certain probability. For example, a CPE may subscribe to a service that guarantees meeting deadlines 95% of the time. To provide such guarantees, admission control algorithms must consider the system's statistical behavior while admitting new CPEs.
- *Background.* No guarantees are given for meeting deadlines. The ITV server schedules such accesses only when there is time left after servicing all guaranteed and statistical streams.

In servicing CPEs during a round, CPEs with deterministic service must be handled before any statistical ones, and all statistical-service CPEs must be serviced before any background CPEs. Missed deadlines must also be distributed fairly so that the same CPE is not dropped each time.

13.2.2.4 Physical Storage of Multimedia

A multimedia server must divide video and audio objects into data packets on a disk. Each data packet can occupy several physical disk blocks. Techniques for improving performance include optimal placement of data packets on the disk, using multiple disks, adding tertiary* storage to gain additional capacity, and building storage hierarchies.

Packets of a multimedia object can be stored contiguously or scattered across the storage device. Contiguous storage is simple to implement, but subject to editor fragmentation. It also requires copying overheads during insertions and deletions to maintain contiguity. Contiguous layouts are useful in mostly-read servers such as video-on-demand, but not for read-write servers. For multimedia, the choice between contiguous and scattered files relates primarily to intrafile seeks. Accessing a contiguous file requires only one seek to position the disk head at the start of the data, whereas with a scattered file, a seek may be required for each packet read. Intrafile seeks can be minimized in scattered layouts by choosing a sufficiently large packet size and reading one packet in each round. It improves disk throughput and reduces the overhead for maintaining indexes that a file system requires.

*The first two storage media are buffer RAM and main disk.

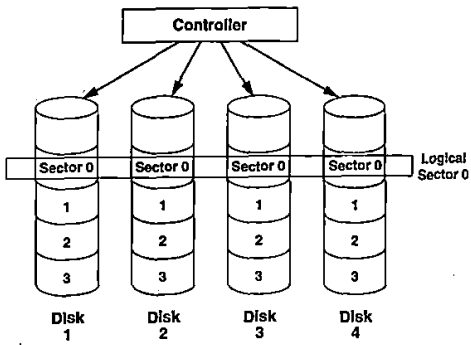

Fig. 13.7 Data is striped across several disks and is accessed in parallel.

To increase the number of possible concurrent access of a stored multimedia object, it may be scattered across multiple disks. This scattering can be achieved by using two techniques: data striping and data interleaving. RAID (redundant array of inexpensive disks) technology (see Fig. 13.7) uses an array of disks and stripes the data across each disk. Physical sector "n" of each disk in the array is accessed in parallel as one large logical sector "n." In this configuration, the disks in the set are spindle synchronized and operate in lock-step mode. Thus, the transfer rate is effectively increased by the number of disk drives employed. Striping does not improve the seek time and rotational latency, however.

In data interleaving, the packets of multimedia objects are interleaved across the disk array with successive file packets stored on different disks. A simple interleave pattern stores the packets cyclically across an array of N disks. In data interleaving, the disks in the array operate independently without any synchronization. Data retrieval can be performed in one of two ways: (1) One packet is retrieved from each disk in the array for each stream in every round or (2) data are extracted from one of the disks for a given CPE in each round. Data retrieval for the CPE cycles through all disks in N successive rounds. In each round, the retrieval load can be balanced by staggering the streams. Thus, all streams still have the same round

length, but each stream considers the round to begin at a different time.

A combination of striping and interleaving can be used to scatter a file in many disks from a server cluster. This technique is often used in constructing a scalable video that can serve a large number of CPEs accessing the same copy of the file.

The reliability of the server will become an important issue as the number of users who rely on ITV services increases. A major failure point in the server is the storage subsystem. With the increase in the number of disks, the reliability of the server decreases. Even if the mean time between failures (MTBF) of a single disk is between 20 and 100 years, as the number of disks increases, the MTBF of the disk subsystem decreases to weeks. This is because the MTBF of a set of disks is inversely proportional to the number of disks, assuming that the MTBF of each disk is independent and exponentially distributed. Further, owing to disk striping, the number of files affected per failure increases.

To provide continuous, reliable service to CPEs, the disk subsystem should employee redundancy mechanisms such as data replication and parity. However, these common techniques need to be extended for ITV servers so that in the presence of a failure all the admitted CPEs can be still serviced without interruption.

13.2.2.5 Updating ITV Servers

Copying of multimedia objects is expensive because their size is large. Copying may be reduced by considering object files to be immutable. Editing is then enabled by manipulating tables of pointers to the media component objects. Obviously, small insertions will not find pointers as valuable as large ones. Also, for deletion, if the multimedia object is an integral number of disk blocks, then the file map could simply be modified.

It is possible to perform insertion and deletions in a time that is proportional to the size of the data inserted/deleted rather than to the whole file. A common scheme is where packets must be filled to a certain minimum level to support continuous retrieval. The insertion/deletion will then consist of some number of full packets plus a remaining partially filled packet. All the packets are then inserted/deleted by modifying the file map, and then data are distributed among adjacent packets of the file to meet the required fill level.

13.2.2.6 VCR-like Functions

An ITV server must support VCR-functions such as pause/resume, fast-forward, and fast-rewind. The pause/resume function is a challenge for buffer management because it interferes with the sharing of streams among different viewers. The fast-forward and fast-rewind functions are implemented by playing back at the normal rate while skipping packets of data. This is easy to do by skipping from one I-frame to another; otherwise interframe dependencies complicate the situation.

The simplest method of fast-forward is to display only I-frames. However, this reduces flexibility. Another way is to categorize each frame as either relevant or irrelevant to fast-forward. During normal operation, both types of frames are retrieved, but during fast-forward, only the fast-forward frames are retrieved and transmitted to the CPE-buffer. It is assumed that the decoding of fast-forward blocks by CPE is straightforward. Scalable compression schemes are readily adapted to this sort of use, although the drawback here is that it poses additional overhead for splitting and recombining blocks.

13.3 USER INTERFACES[10-12]

User interface designs for ITV are more involved than for standard TV owing to the richness of the types of possible interactions with the consumer. As mentioned earlier, they need to be extremely easy to use, as well as intuitive. They will most likely vary among different applications. There will be a small number of ITV generic program types, such as movies-on-demand, shopping, and so forth, and certain "look-and-feel" guidelines will emerge for user interfaces and for ITV programming in general. Typically, multimedia interfaces are designed using a mixture of four kinds of modules:

1. Media editors
2. Multimedia object locators and browsers (navigation)
3. Authoring tools for creating program logic
4. Hypermedia object creation.

The last item is currently less relevant to ITV and is therefore not discussed here. However, it may become more widely available in the future. A number of media editors already exist to create "primitive" media objects such as text, still images, and video and audio clips. A media editor is then used for editing primitive multimedia objects to construct a composite object that is suitable as a multimedia presentation. The basic functions provided by the editor are delete, cut, copy, paste, move, and merge. The real challenge is to present these basic functions in the most natural manner that is intuitive to the user. For example, the move operation is usually represented by highlighting text and dragging and dropping it to the new location. However, the same metaphor does not work for all multimedia objects. The exact implementation of the metaphor and the efficient implementation for that metaphor will distinguish such editors from each other.

Navigation refers to the sequence in which the objects are browsed, searched, and used. Navigation can be either direct or predefined. If direct, then the user determines the next sequence of actions. If predefined, for example, searching for the next program start, the user needs to know what to expect with successive navigation actions. The navigation can also be in a *browse* mode, in which the user wants to get general information about a particular topic. Browsing is common in systems with graphical data. ITV systems normally provide direct navigation as well as browse options.

The browse mode is typically used to search for a specific attribute or a subobject. For example, in an application using dialog boxes, the user may have traversed several dialog boxes to reach a point where the user is ready to play a sound object. A browse mode may be provided to search the database for all sound objects appropriate for that dialog box. An important aspect of any multimedia system is to maintain a clear perspective of the objects being displayed and the relationship between these objects. The relationship can be associative (if they are a part of a set) or hierarchical. Navigation is undoubtedly the first generic application for ITV. Various flavors have been created, for example, mapping of ITV programs to channel numbers, and interfaces that show "picture-in-picture." Navigation is rightly the first application focus, acknowledging the real challenge ahead in helping viewers find their way in the (hoped for) proliferation of programming in the interactive world.

The interactive behavior of ITV determines how the consumer interacts with the services available. Some of the design elements are:

- Data entry dialog boxes
- A source-specific sequence of operations depicted by highlighting or enabling specific menu items
- Context-sensitive operation of buttons
- Active icons that perform ad hoc tasks (e.g., intermediate save of work).

The look and feel of a service depends on a combination of the metaphor (e.g., VCR interface) being used to simulate real-life interfaces, the windows guidelines, the ease of use, and the aesthetic appeal. An elegant service has a consistent look and feel. The interface will be more friendly if, for example, the same buttons are found in the same location in each dialog box and the functions of an icon toolbar are consistent. User interfaces for multimedia applications require careful design to ensure that they are close adaptations of the metaphors to which consumers are accustomed in other domains. Two metaphors relevant to ITV are the Aural and VCR interfaces described below.

A common approach for speech recognition (also called the Aural metaphor) user interfaces has been to graft the speech recognition interface into existing graphical user interfaces. This is a mix of conceptually mismatched media which makes the interface cumbersome. A true aural user interface (AUI) would accept speech as direct input and provide a speech response to the user's actions. The real challenge in designing AUI systems is to creatively substitute voice and ear for the keyboard and display, and to be able to mix and match them. Aural cues can be used to represent icons, menus, and windows of a GUI.* Besides computer technology, AUI

*Graphical User Interface

designs involve human perception, cognitive science, and psychoacoustic theory. The human response to visual and aural inputs is markedly different. Rather than duplicate the GUI functions in an AUI, this requires new thinking to adapt the AUI to the human response for aural feedback. AUI systems need to learn so that they can perform most routine functions without constant user feedback, but still be able to inform the user of the functions performed on the user's behalf with very succinct aural messages. This is an area of active research.

A common user interface for services such as video capture, switching channels, and stored video playback is to emulate the camera, TV, and VCR on the screen. A general TV/VCR and camera metaphor interface will have all functions one would find in a typical video camera when it is in a video capture mode. It will have live buttons for selecting channels, increasing sound volume, and changing channels. The video is shown typically as a graduated bar, and the graduations are based on the full length of the video clip. The user can mark the start and end points to play the clip or to cut and paste the clip into a new video scene. The screen duplicates the cut and paste operation visually. Figure 1.5 shows an interface with the TV/VCR metaphor. A number of video editors emulate the functionality of typical film-editing equipment. The editing interface shows a visual representation of multiple film strips that can be viewed at user-selected frame rates. The user may view every *n*th frame for an entire scene or may wish to narrow down to every frame to pinpoint the exact point for splicing another video clip. The visual display software allows index markers on the screen for splicing another video clip down to the frame level. The screen display also shows the soundtracks and allows splicing soundtracks down to the frame level.

A good indexing system provides a very accurate counter for time as well as for individual frames; it can detect special events such as a change in scenes, start and end of newly spliced frames, start and end of a new soundtrack, etc. Also, the user can place permanent index markings to identify specific points of interest in the sequence. A number of indexing algorithms have been published in the literature.[13]

Indexing is useful only if the audio/video are stored. Since audio and video may be stored on separate servers, synchronization must be achieved before playback. Indexing may be needed to achieve the equivalent of a sound mixer and synthesizer found, for example, in digital pianos. Exact indexing to the frame level is desirable for video synchronization to work correctly. The indexing information must be stored on a permanent basis, for example, as SMPTE time codes or MPEG Systems time stamps. This information can be maintained in either the soundtrack or the video clip itself and must be extracted if needed for playback synchronization. Alternatively, the composite object created as a result of authoring can maintain the index information on behalf of all objects.

13.4 CONTENT CREATION FOR ITV[14–19]

A good authoring system must access predefined media and program elements, sequence these elements in time, specify their placement spatially, initiate actions in parallel, and indicate specific events (such as consumer inputs or external triggers) and the actions (some may be algorithmic functions) that may result from them. Some of these algorithmic functions may be performed by both the CPE and the server when the application is executed. Existing commercial tools for creating interactive multimedia applications run the gamut from low-level programming languages to packages that support the relatively casual creation of interactive presentations. Two very broad categories are (1) visual tools for use by nonprogrammers and (2) programming languages. The latter might include specialized "scripting" languages as well as more traditional general-purpose programming languages, perhaps extended by object-types or libraries that add ITV functionality. Usually the media elements used by these tools are created with a variety of other tools designed for particular media types, such as for still images, video and audio segments, and animation.

The design of an authoring system requires careful analysis of the following issues:

1. Hypermedia design details
2. User interfaces

3. Embedding/linking multimedia objects to the main presentation
4. Multimedia objects—storage and retrieval
5. Synchronization of components of a composite multimedia stream.

Hypermedia applications present a unique set of design issues not commonly encountered in other applications. A good user interface design is crucial to the success of a hypermedia application. The user interface presents a window to the user for controlling storage and retrieval, connecting objects to the presentation, and defining index marks for combining different multimedia streams. While the objects may be captured independently, they have to be played back together for authoring. The authoring system must allow coordination of many streams to produce a final presentation. A variety of authoring systems exist depending on their role and flexibility required. Some of these are described below.

13.4.1 Dedicated Authoring Systems

Dedicated authoring systems are designed for single users and single streams. The authoring is typically performed using desktop computer systems on multimedia objects captured by a video camera or on objects stored in a multimedia library. Dedicated multimedia authoring systems are not usually used for sophisticated applications. Therefore, they need to be engineered to provide user interfaces that are extremely intuitive and follow real-world metaphors (e.g., the VCR metaphor). Also flexibility and functionality has to be carefully limited to prevent overly complex authoring.

13.4.2 Timeline-Based Authoring

In a time-line-based authoring system, objects to be merged are placed along a timeline. The author specifies objects and their position in the timeline. For presentation in a proper time sequence each object is played at the prespecified point in the time scale.

In early timeline systems, once the multimedia object was captured in a timeline, it was fixed in location and could not be manipulated easily. Editing a particular component caused all objects in the timeline to be reassigned because the positions of objects were not fixed in time, only in sequence. Copying portions of the timeline became difficult as well. Newer systems define timing relations directly between objects. This makes inserting and deleting objects much easier because the start and end of each object is more clearly defined.

13.4.3 Structured Multimedia Authoring

Structured multimedia authoring is based on structured object-level construction of presentations. A presentation may be composed of a number of video clips with associated sound tracks, separate music tracks, and other separate soundtracks. Explicit representation of the structure allows modular authoring of the individual component objects. The timing constraints are also derived from the constituent data items. A good structured system also allows the user to define an object hierarchy and to specify the relative location of each object within that hierarchy. Some objects may have to undergo temporal adjustment, which allows them to be expanded or shrunk to better fit the available time slot. The navigation design of the authoring system should allow the author to view the overall structure, while at the same time being able to examine specific object segments more closely. The benefit of an object hierarchy is that removing the main object also removes all subobjects that are meaningful only as overlays on the main object. The design of the views must show all relevant information for a specific view. For example, views may be needed to show the object hierarchy plus individual component members of that hierarchy, and to depict the timing relation between the members of the hierarchy.

13.4.4 Programmable Authoring Systems

Early structured systems did not allow authors to express automatic functions for handling certain routine tasks in the authoring process. Program-

mable authoring systems provide such additional capability in the following areas:

- Functions based on audio and image processing and analysis
- Embedded program interpreters that use audio and image-processing functions.

Thus the main improvement in building user programmability into the authoring tool is not only to perform the analysis, but also to manipulate the stream based on the analysis results. This programmability allows the performing of a variety of tasks through the program interpreter rather than manually. A good example of how this is used is the case of locating video "silences" which typically occur before that start of a new segment (cut). The program can be used to help the user to get to the next segment by clicking a button rather than playing the video. Advances in authoring systems will continue as users acquire a better understanding of stored audio and video and the functionalities needed by users.

The development and deployment of interactive television on a large scale presents numerous technical challenges in a variety of areas. As progress continues to be made in all these areas, trial deployments of the interactive services are necessary to ascertain consumer interest in this new form of communication. We are certainly rather naive today about the future of these services. However, it seems certain that the variety of trials now being conducted will find a subset of this technology that is popular with consumers and a solid business proposition for service providers. This will provide the definition of services and the economic drivers necessary for these services to prosper.

REFERENCES

1. D. E. Blahut, T. E. Nichols, W. M. Schell, G. A. Story, and E. S. Szurkowski, "Interactive Television," *Proc. of IEEE,* pp. 1071–1086 (July 1995).
2. L. Crutcher and J. Grinham, "The Networked Jukebox," *IEEE Trans. Circuits Systems,* pp. 105–120 (April 1994).
3. G. H. Arlen, *Networked Multimedia: A Review of Interactive Services and the Outlook for Cable-Delivered Services,* Arlen Communications (April 1993).
4. E. G. Bartlett, *Cable Television Technology and Operations,* McGraw-Hill, New York (1990).
5. D. Anderson, Y. Osawa, and R. Govindam, "A File System for Continuous Media," *ACM Trans. Computer Systems* 10(4):311–337 (November 1992).
6. A. L. Narasimha Reddy and J. C. Wyllie, "I/O Issues in a Multimedia System," *Computer,* 27(3): 69–74 (March 1994).
7. P. Lougher and D. Shepherd, "The Design of a Storage Server for Continuous Media," *Computer J.* 36(1):32–42 (February 1993).
8. P. Venkat Rangan and H. M. Vin, "Efficient Storage Techniques for Digital Continuous Multimedia," *IEEE Trans. Knowledge Data Eng.* 5(4):564–573 (August 1993).
9. B. Ozden, R. Rostogi, and A. Silberschatz, "A Framework for the Storage and Retrieval of Continuous Media Data," *Proc. IEEE Conf. on Multimedia Computing and Systems* (May 95).
10. J. Kuegel, C. Keslein, and J. Rutledge, "Toolkits for Multimedia Interface Design," *Exhibition 92,* pp. 275–285.
11. R. Hayakawa, H. Sakagami, and J. Bekimoto, "Audio and Video Extensions to Graphical User Interface Toolkits," *3rd Int'l Workshop on Operating System Support for Audio/Video,* No. 92.
12. M. Blather, R. Deneburg (Eds.), *Multimedia Interface Design,* Addison-Wesley, Reading, MA (1991).
13. A. Hampapur, R. Jain, and T. E. Weywouth, "Indexing in Video Databases," *SPIE,* 2420: 293–305 (1995).
14. J. Conklin, "Hypertext: An Introduction and Survey," *IEEE Computer* (September 1987).
15. L. Hardman, G. van Rossum, and O. Bultermans, "Structured Multimedia Authoring," *ACM Multimedia,* pp. 283–289 (August 1993).
16. MACRO MEDIA Direction Overview Manual, Macromedia Inc. (1991).
17. Script X Language Guide, Kaleida (1993).
18. MACRO MEDIA Authorwave Working Model, Macromedia, Inc. (1993).
19. Apple Computer, *Inside MacIntosh: Quick Time,* Addison-Wesley, Reading, MA (1993).

14

HIGH-DEFINITION TELEVISION (HDTV)

Analog television was invented and standardized over fifty years ago, mainly for over-the-air broadcast of entertainment, news, and sports. In North America, the basic design was formalized in 1942, revised in 1946, and shortly after, limited service was offered. Color was added in 1954 with a backward compatible format. European broadcasting is equally old; monochrome television dates back to the 1950s and color was introduced in Germany in 1965 and in France in 1966. Since then, only backward compatible changes have been introduced, such as multichannel sound, closed-captioning, and ghost cancellation. The standard has thus evolved[4] in a backward compatible fashion, thereby protecting substantial investment made by the consumers, equipment providers, broadcasters, cable/ satellite service providers, and program producers.

Television is now finding applications in telecommunications (e.g., video conferencing), computing (e.g., desktop video), medicine (e.g., telemedicine), and education (e.g., distance learning). To enable many of these applications, television is being distributed by a variety of means not imagined at the time of standardization, such as cable (fiber or coaxial), video tapes or discs (analog or digital), and satellite (e.g., direct broadcast). It has also been recognized that the television signal packaged for terrestrial broadcasting, while cleverly done in its time, is wasteful of the broadcast spectrum and can benefit from modern signal processing. In addition, the average size of television sets has grown steadily, and the average viewer is getting closer to the display.

These newer applications, delivery media, and viewing habits are exposing various limitations of the current television system. At the same time a new capability is being enabled owing to the technology of compression, inexpensive and powerful integrated circuits, and large displays. Thus, we have an unstoppable combination of technological push with commercial pull.

High-Definition Television (HDTV) systems now being designed are more than adequate to fill these needs in a cost-effective manner. HDTV will be an entirely new high-resolution and wide-screen television system, producing much better quality pictures and sound. It will be one of the first systems that will not be backward compatible.

Worldwide efforts to achieve practical HDTV systems have been in progress for over 20 years. While we do not provide a detailed historical perspective on these efforts, a summary of the progress made towards practical HDTV is as follows:

- HDTV in Japan
  - —1970's NHK started systems work; infrastructure by other Japanese companies
  - —1980's led to MUSE, 1125/60 2:1 interlaced, fits 27 MHz satellite transponder
  - —Satellite HDTV service operating; NHK's format international exchange standard

—Narrow-MUSE proposed to FCC for USA
—Future of MUSE not promising, switch to all digital system expected

- HDTV in Europe
  —In 1986, broadcasters and manufacturers started HD-MAC effort
  —HD-MAC intended backward compatible with PAL and D-MAC; satellite (DBS)
  —HD-MAC obsolete; new system with digital compression/modulation sought
- HDTV in North America
  —In 1987, broadcasters requested FCC which resulted in formation of Advisory Committee (ACATS)
  —ACATS proposed hardware testing of 23 systems, 5 survived around test time
  —All 4 digital systems in running after tests, formed Grand Alliance in 1993

This chapter summarizes the main characteristics of HDTV and its expected evolution.

14.1 CHARACTERISTICS OF HDTV

HDTV systems are currently evolving from a number of proposals made by different organizations in different countries. While there is no complete agreement, a number of aspects are common to most of the proposals. HDTV is characterized by increased spatial and temporal resolution; larger image aspect ratio (i.e., wider image); multichannel CD-quality surround sound; reduced artifacts compared to the current composite analog television; bandwidth compression and channel encoding to make better utilization of the terrestrial spectrum; and finally, digital to provide better interoperability with the evolving telecommunications and computing infrastructure.

The primary driving force behind HDTV is a much better quality picture and sound to be received in the consumer home. This is done by increasing the spatial resolution by more than a factor of 2 in both the horizontal and vertical dimensions. Thus, we get a picture that has about 1000 scan lines and more than 1000 pels per scan line.

In addition, it is widely accepted (but not fully agreed to) that HDTV should eventually use only progressive (noninterlace) scanning at about 60 frames/s to allow better fast-action sports and far better interoperability with computers, while also eliminating the artifacts associated with interlace. The result of this is that the number of active pels in an HDTV signal increases by about a factor of 5, with a corresponding increase in the analog bandwidth.

From consumers' experience in cinema, a wider image aspect ratio (IAR) is also preferred. Most HDTV systems specify the image aspect ratio to be 16:9. The quality of audio is also improved by means of multichannel, CD-quality, surround sound in which each channel is independently transmitted.

Since HDTV will be digital, different components of the information can be simply multiplexed in time instead of frequency multiplexed on different carriers as in the case of analog TV. For example, each audio channel is independently compressed, and these compressed bits are multiplexed with compressed bits from video, as well as bits for closed-captioning, teletext, encryption, addressing, program identification, and other data services in a layer fashion.

Unfortunately, increasing the spatial and temporal resolution of the HDTV signal and adding multichannel sound also increases its analog bandwidth. Such an analog signal cannot be accommodated in a single channel of the currently allocated broadcast spectrum.

Moreover, even if bandwidth were available, such an analog signal would suffer interference both to and from the existing TV transmissions. In fact, much of the available broadcast spectrum can be characterized as a fragile transmission channel. Many of the 6-MHz TV channels are kept unused because of interference considerations, and are designated as *taboo* channels.

Therefore, all the current HDTV proposals employ digital compression, which reduces the bitrate from approximately 1 Gbit/s to about 20 Mbits/s. The 20 Mbits/s can be accommodated in a 6-MHz channel either in a terrestrial broadcast spectrum or a cable television channel. This digital signal is incompatible with the current television system and therefore can be decoded only by a special decoder.

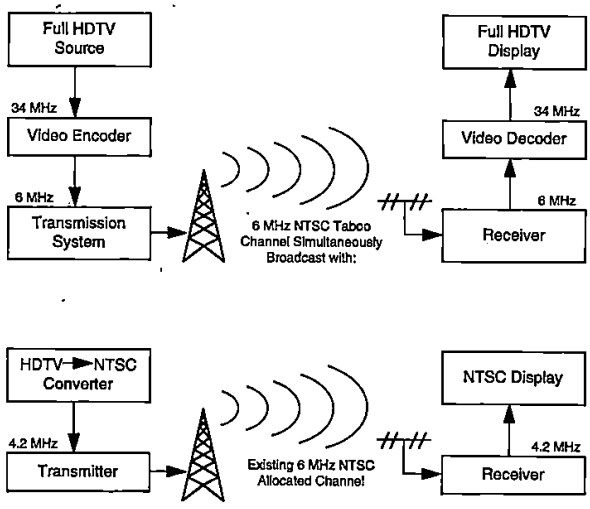

Fig. 14.1 Simulcasting of digital HDTV with NTSC analog TV. Under current channel assignments, many channels remain vacant (i.e., taboo) to avoid co-channel and adjacent channel interference. Since digital signals have lower transmission power, they can use the taboo channels in the new channel assignment.

In the case of terrestrial broadcast it is also necessary to reduce interference to and from other broadcast signals, including the analog television channels. Therefore, digital HDTV modulates the compressed audio and video bits onto a signal that has much lower power than the analog TV signal and has no high-power carriers, while requiring the same analog bandwidth as the current analog TV. In addition, most HDTV proposals simulcast* the HDTV signal along with the current analog TV having the same content (see Fig. 14.1).

A low-power, modulated digital HDTV signal is transmitted in a taboo channel that is currently unused because of co-channel and adjacent channel interference considerations. Simultaneously, in an existing analog channel, the same content (down-converted to the current television standard) will be transmitted and received by today's analog television sets. By proper shaping of the modulated HDTV signal spectrum, it is possible to avoid its interference into distant co-channel as well as nearby adjacent channel analog television signals. At the same time, distant co-channel and nearby adjacent channel analog television signals must not degrade the HDTV pictures significantly because of interference.

This design allows previously unused terrestrial spectrum to be used for HDTV. Of course, it could also be used for other services such as digital Standard Definition Television (SDTV), computer communication, and so forth. Consumers with current television sets will receive television as they do now. However, consumers willing to spend an additional amount will be able to receive HDTV in currently unused channels. As the HDTV service takes hold in the marketplace, spectrally inefficient analog television can be retired to release additional spectrum for additional HDTV, SDTV, or other services. In the United States, the FCC plans to retire NTSC after 15 years from the start of the HDTV broadcasts.

Another desirable feature of HDTV systems is easy interoperability with the evolving telecommunications and computing infrastructure. Such interoperability will allow easy introduction of different services and applications, and at the same time reduce costs because of commonality of components and platforms across different applica-

*The same program is sent on two channels.

tions. There are several characteristics that improve interoperability. Progressive scanning, square pels, and digital representation all improve interoperability with computers. In addition, they facilitate signal processing required to convert HDTV signals to other formats either for editing, storage, special effects, or down-conversions to current television. The use of a standard compression algorithm (e.g., MPEG[2,3]) transforms raw, high-definition audio and video into a coded bitstream, which is nothing more than a set of computer instructions and data that can be executed by the receiver to create sound and pictures.

Further improvement in interoperability at the transport layer is achieved by encapsulating audio and video bitstreams into fixed-size packets. This facilitates the use of forward error correction (necessary to overcome transmission errors) and provides synchronization and extensibility by the use of packet identifiers, headers, and descriptors. In addition, it allows easy conversion of HDTV packets into other constant size packets adopted by the telecommunications industry, for example, the Asynchronous Transfer Mode (ATM) standard for data transport.

Based on our discussion thus far the requirements for HDTV can be summarized as follows:

- Higher spatial resolution
  —Increase spatial resolution by at least a factor of 2 in horizontal and vertical direction
- Higher temporal resolution
  —Increase temporal resolution by use of progressive 60 Hz temporal rate
- Higher Aspect Ratio
  —Increase aspect ratio to 16:9 from 4:3 for a wider image
- Multichannel CD quality surround sound
  —At least 4 to 6 channel surround sound system
- Reduced Artifacts as compared to Analog TV
  —Remove composite format artifacts, interlace artifacts etc
- Bandwidth compression and channel coding to make better use of Terrestrial spectrum
  —Digital processing for efficient spectrum usage
- Interoperability with evolving telecommunication and computing infrastructure
  —Digital compression and processing for ease of interworking

14.2 PROBLEMS OF CURRENT ANALOG TV

Current analog TV was designed during the days when both the compression, modulation, and signal processing algorithms and the circuitry to process them were not adequately developed. Therefore, as the current TV began to get deployed for different applications, its artifacts as well as inefficiency of spectral utilization became bigger problems. Digital HDTV attempts to alleviate many of these problems. *Interline flicker, line crawl,* and *vertical aliasing* are largely a result of interlaced scanning in current television. These effects are more pronounced in pictures containing slowly moving high-resolution graphics and text with sharp edges. Most pictures used for entertainment, news, and sports did not historically contain such sharp detail, and these effects were tolerably small. The computer industry has largely abandoned interlace scanning to avoid these problems. Progressive scanning and transmission of HDTV can eliminate these effects almost entirely.

Large area flicker as seen from a close distance on large screens is the result of our ability to detect temporal brightness variations, even at frequencies beyond 60 Hz, in the peripheral areas of our vision. The cost of overcoming this problem in terms of bandwidth (i.e., by increasing the frame rate) is too large and is therefore not addressed in any current HDTV proposal. Computer displays employ much higher frame rates (up to 72 Hz) to overcome large area flicker. However, the brightness of a computer display is usually much higher than that of home TV, and this accentuates the effect, thereby requiring the larger frame rate.

Static raster visibility is more apparent on larger displays since viewers can visually resolve individual scan lines. The increased spatial resolution in HDTV will reduce this effect substantially.

Cross-color and cross-luminance are a result of the color modulation used to create the analog composite signal. HDTV will use component signals. Moreover, in digital HDTV, each color component will be compressed separately, with the compressed bits multiplexed in time. Thus, these two effects will be eliminated entirely.

With the present analog television, a number of