DIGITAL TELEVISION

Analog television was developed and standardized in the forties, mainly for over-the-air broadcast of entertainment, news, and sports. While a few upward compatible changes have been made in the intervening years, such as color, multichannel sound, closed captioning, and ghost cancellation, the underlying analog system has survived a continuous technological evolution that has pervaded all other media. Television has stimulated the development of a global consumer electronics industry that has brought high-density magnetic recording, high-resolution displays and low-cost imaging technologies from the laboratory into the living room. A vast array of video production, processing technologies make high-quality programming an everyday reality, realtime on-site video the norm rather than the exception, and video the historical medium of record throughout the world. More recently, emergence of personal computers and high speed networks has given rise to desktop video to improve productivity for businesses.

In spite of this impressive record and a large invested base, we are on the threshold of a major disruption in the television industry. After fifty years of continuous refinement, the underlying technology of television is going to be entirely redone. Digital video is already proliferating in a variety of applications such as video-conferencing, multimedia computing, and program production; the impediments that have held it back are rapidly disappearing. The key enabling technologies are: (1) mature and standardized algorithms for high quality compression; (2) inexpensive and powerful integrated circuits for the processing, storage, and reconstruction of video signals; (3) inexpensive, high capacity networks for transport of video; (4) uniform methods for storing, addressing, and accessing multimedia content; (5) evolution of computer architecture to support video I/O. The market drivers include: (1) direct consumer access for content-providers; (2) convergence of video with other information sources such as print; (3) the emergence of a fast growing consumer market for personal computing, (4) the evolution of Internet and other networks in the commercial domain, and (5) the removal of various regulatory barriers.

This article deals with the technology of digital television. We first start with how the television signal is sampled (scanning) and digitized. We then discuss techniques of compression to reduce the bit rate to a manageable level, and describe briefly the emerging standards for compression.

TELEVISION SCANNING

The image information captured by a television camera conveys color intensity (in terms of red, green, and blue primary colors) at each spatial location (x, y) and for each time instance (t). Thus, the image intensity is multidimensional (x, y, t) in nature. However, it needs to be converted to a unidimensional signal so that processing, storage, communications, and display can take place. Raster scanning is the process used to convert a three-dimensional (x, y, t) image intensity into a one-dimensional television waveform (1). The first step is to sample the television scene many times ($1/T$, where T is frame period in seconds) per second to create a sequence of still images (called frames). Then, within each frame, scan lines are created by vertical sampling. Scanning proceeds sequentially, left to right for each scan line and from top to bottom line at a time within a frame. In a television camera, an electron beam scans across a photosensitive target upon which the image is focused. In more modern cameras, charge coupled devices (CCDs) are used to image an area of the picture, such as an entire scan line. At the other end of the television chain, with raster scanned displays, an electronic beam scans and lights up the picture elements in proportion to the light intensity. While it is convenient to think of the samples of a single frame all occurring at a single time instance (similar to the simultaneous exposure of a single frame for film), the scanning in a camera and in a display results in every sample corresponding to a different point in time.

Progressive and Interlace Scan

There are two types of scanning: progressive (also called sequential) and interlaced. In progressive scanning, the television scene is first sampled in time to create frames and within each frame all the raster lines are scanned in order 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. Almost all computer displays, especially all high-end computers, are sequentially scanned.

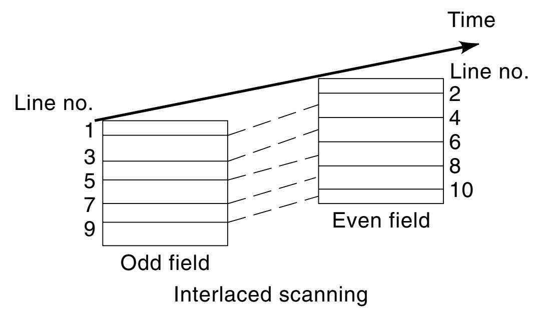


Figure 1. A television frame is divided into an odd field (containing odd-numbered scan lines) and an even field (containing even-numbered scan lines).

In interlaced scanning (see Fig. 1), all the odd-numbered lines in the entire frame are scanned first during the first half of the frame period, T, and then the even-numbered lines are scanned during the second half. This process produces two distinct images per frame 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 current TV systems (National Television System Committee [NTSC], PAL, SECAM) use interlaced scanning. One of the principal benefits of interlaced scanning is to reduce the scan rate (or the bandwidth) without significanty reducing image quality. This is done with a relatively high field rate (a lower field rate would cause flicker), while maintaining a high total number of scan lines in a frame (lower number of lines per frame would reduce resolution on static images). Interlace cleverly preserves the high-detail visual information and, at the same time, avoids visible large area flicker at the display due to insufficient temporal post-filtering by the human eye. The NTSC has 15,735 scan lines/s or 525 lines/frame, since there are 29.97 frames/s. For each scan line, a small period of time (16% to 18% of total line time), called blanking or retrace, is allocated to return the scanning beam to the left edge of the next scan line. European systems (PAL and SECAM) have 625 lines/frame, but 50 fields/s. The larger number of lines results in better vertical resolution, whereas larger numbers of frames result in better motion rendition and lower flicker.

While there is no agreement worldwide yet, high definition TV (HDTV) will have approximately twice the horizontal and vertical resolution of standard television. In addition, HDTV will be digital, where the television scan lines will also be sampled horizontally in time and digitized. Such sampling will produce an array of approximately 1000 lines and as many as 2000 pixels per line. If the height/width ratio of the TV raster is equal to the number of scan line/number of samples per line, the array is referred to as having "square pixels," that is, the electron beam is spaced equally in the horizontal and vertical direction, or has a square shape. This facilitates digital image processing as well as computer synthesis of images. One of the liveliest debates regarding the next generation television systems involves the type of scanning to be employed: interlaced or progressive. Interlaced scanning was invented in the 1930s when signal processing techniques, hardware, and memory devices were all in a state of infancy. Since all the current TV systems were standardized over five decades ago, they use interlace, and therefore, the technology and the equipment (e.g., cameras) using interlace are mature. However, interlace often shows flickering artifacts in scenes with sharp detail and has poor motion rendition, particularly for fast vertical motion of small objects. In addition, digital data compression is more easily done on progressively scanned frames. Compatibility with film and computers also favors progressive scanning.

In the future, since different stages of the television chain have different requirements, it is likely that creation (production studios), transmission, and display may employ different scanning methods. Production studios require high quality cameras and compatibility with film and computer generated material, all of very high quality. If good progressive cameras were available and inexpensive, this would favor progressive scanning at even higher scan rates ($>$ 1,000 lines/frame). However, transmission bandwidth, particularly for terrestrial transmissions, is expensive and limited, and even with bandwidth compression current technology can handle only up to 1,000 lines/frame. Display systems can show a better picture by progressive scanning and refreshing at higher frame rates (even if the transmission is interlaced and at lower frame rates) made possible by frame buffers. Thus, while there are strong arguments in favor of progressive scanning in the future, more progress is needed on the learning curve of progressive equipment. The FCC (Federal Communication Commission) in the United States therefore decided to support multiple scanning standards for terrestrial transmission, one interlace and five progressive, but with a migration path toward the exclusive use of progressive scanning in the future.

Image Aspect Ratio

The image aspect ratio is generally defined as the ratio of picture width to height. It impacts the overall appearance of the displayed image. For standard TV the aspect ratio is 4 : 3. This value was adopted for TV, as this format was already used and found acceptable in the film industry prior to 1953. However, since then the film industry has migrated to wide-screen formats with aspect ratio of 1.85 or higher. Since subjective tests on viewers show a significant preference for a wider format than that used for standard TV, HDTV plans to use the aspect ratio of 1.78, which is quite close to that of the wide-screen film format.

Image Intensity

Light is a subset of the electromagnetic energy. The visible spectrum ranges from 380 to 780 nm in wavelengths. Thus, visible light can be specified completely at a picture element (pel) by its wavelength distribution $\{S(\lambda)\}$. This radiation excites three different receptors in the human retina that are sensitive to wavelengths near 445 (called blue), 535 (called green), and 570 (called red) nm. 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.

This is the basis of the *trichromatic* theory of color which states that for human perception, any color can be synthesized by an appropriate mixture of three properly chosen primary colors $\mathbf{R}$, $\mathbf{G}$, and $\mathbf{B}$ (2). For video, the primaries are usually red, green, and blue. The amounts of each primary required are called the tristimulus values. If a color C has

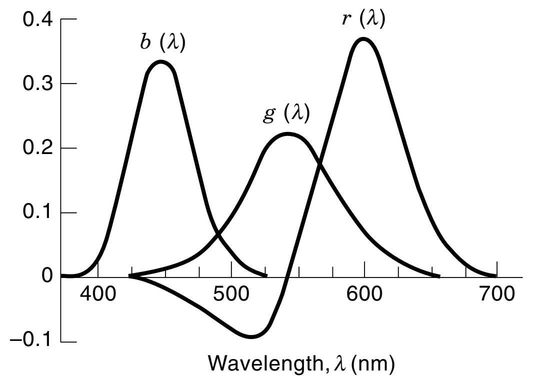


Figure 2. The color-matching functions for the 2° standard observer, based on primaries of wavelengths 700 (red), 546.1 (green), and 435.8 nm (blue), with units such that equal quantities of the three primaries are needed to match the equal energy white.

tristimulus values R_C, G_C, and B_C, then $C = R_C\mathbf{R} + G_C\mathbf{G} + B_C\mathbf{B}$. The tristimulus values of a wavelength distribution $S(\lambda)$ are given by

$$\begin{aligned} R_S &= \int S(\lambda) r(\lambda)\, d \\ G_S &= \int S(\lambda) g(\lambda)\, d\lambda \\ B_S &= \int S(\lambda) b(\lambda)\, d\lambda \end{aligned} \tag{1}$$

where $\{r(\lambda), g(\lambda), b(\lambda)\}$ are called the color matching functions for primaries $\mathbf{R}$, $\mathbf{G}$, and $\mathbf{B}$.

These are also the tristimulus values of unit intensity monochromatic light of wavelength λ. Figure 2 shows color matching functions with the primary colors chosen to be spectral (light of a single wavelength) colors of wavelengths 700.0, 546.1, and 435.8 nm. Equation (1) allows us to compute the tristimulus values of any color with a given spectral distribution, $S(\lambda)$, using color matching functions.

One consequence of this is that any two colors with spectral distributions $S_1(\lambda)$ and $S_2(\lambda)$ match if and only if

$$\begin{aligned} R_1 &= \int S_1(\lambda) r(\lambda)\, d\lambda = \int S_2(\lambda) r(\lambda)\, d\lambda = R_2 \\ G_1 &= \int S_1(\lambda) g(\lambda)\, d\lambda = \int S_2(\lambda) g(\lambda)\, d\lambda = G_2 \\ B_1 &= \int S_1(\lambda) b(\lambda)\, d\lambda = \int S_2(\lambda) b(\lambda)\, d\lambda = B_2 \end{aligned} \tag{2}$$

where $\{R_1, G_1, B_1\}$ and $\{R_2, G_2, B_2\}$ are the tristimulus values of the two distributions $S_1(\lambda)$ and $S_2(\lambda)$, respectively. This could happen even if $S_1(\lambda)$ were not equal to $S_2(\lambda)$ for all the wavelengths in the visible region.

Instead of specifying a color by its tristimulus values $\{R, G, B\}$, normalized quantities called chromaticity coordinates $\{r, g, b\}$ are often used:

$$\begin{aligned} r &= \frac{R}{R+G+B} \\ g &= \frac{G}{R+G+B} \\ b &= \frac{B}{R+G+B} \end{aligned} \tag{3}$$

Since $r + g + b = 1$, any two chromaticity coordinates are sufficient. However, for complete specification a third dimension is required. It is usually chosen to be the luminance (Y).

Luminance is an objective measure of brightness. Different contributions of wavelengths to the sensation of brightness are represented by the relative luminance efficiency $y(\lambda)$. The luminance of any given spectral distribution $S(\lambda)$ is then given by

$$Y = k_m \int S(\lambda) y(\lambda)\, d\lambda \tag{4}$$

where k_m is a normalizing constant. For any given choice of primaries and their corresponding color matching functions, luminance can be written as a linear combination of the tristimulus values, $\{R, G, B\}$. Thus, a complete specification of color is given either by the three tristimulus values or by the luminance and two chromaticities. A color image can then be specified by luminance and chromaticities at each pel.

COMPOSITE TV SYSTEMS

A camera imaging a scene generates for each pel the three color tristimulus values *RGB,* which may be further processed for transmission or storage. At the receiver, the three components are sent to the display, which regenerates the contents of the scene at each pel from the three color components. For transmission or storage between the camera and the display a luminance signal Y representing brightness and two chrominance signals representing color are used. The need for such a transmission system arose with NTSC, the standard used in North America and Japan, where compatibility with monochrome receivers required a black-and-white signal, which is now referred to as the Y signal. It is well known that the sensitivity of the human eye is highest to green light, followed by that of red, and the least to blue light. The NTSC system exploited this fact by assigning a lower bandwidth to the chrominance signals as compared to the luminance, Y, signal. This made it possible to save bandwidth without losing color quality. The PAL and SECAM systems also employ reduced chrominance bandwidths (3).

The NTSC System

The NTSC color space of *YIQ* can be generated from the gamma-corrected *RGB* components or from *YUV* components as follows:

$$\begin{aligned} Y &= 0.299R' + 0.587G' + 0.114B' \\ I &= 0.596R' - 0.274G' - 0.322B' = -(\sin 33^\circ)U + (\cos 33^\circ)V \\ Q &= 0.211R' - 0.523G' - 0.311B' = (\cos 33^\circ)U + (\sin 33^\circ)V \end{aligned} \tag{5}$$

where $U = B' - Y/2.03$ and $V = R' - Y/1.14$. (Gamma correction is performed to compensate for the nonlinear relationship between signal voltage, U, and light intensity, B [$B \cong V^\gamma$].)

The inverse operation, that is, generation of gamma-corrected *RGB* components from the *YIQ* composite color space, can be accomplished as follows:

$$\begin{aligned} R' &= 1.0Y + 0.956I + 0.621Q \\ G' &= 1.0Y + 0.272I + 0.649Q \\ B' &= 1.0Y - 1.106I + 1.703Q \end{aligned} \tag{6}$$

In NTSC, the Y, I, and Q signals are all multiplexed into a 4.2 MHz bandwidth. Although the Y component itself takes 4.2 MHz bandwidth, multiplexing all three components into the same 4.2 MHz becomes possible by interleaving luminance and chrominance frequencies, without too much "crosstalk" between them. This is done by defining a color subcarrier at approximately 3.58 MHz. The two chrominance signals I and Q are QAM (quadrature amplitude modulation) modulated onto this carrier. The envelope of this QAM signal is approximately the saturation of the color, and the phase is approximately the hue. The luminance and modulated chrominance signals are then added to form the composite signal. The process of demodulation first involves comb filtering (horizontal and vertical filtering) of the composite signal to separate the luminance and the chrominance signal followed by further demodulation to separate the I and Q components.

The Phase Alternate Line System

The *YUV* color space of PAL is employed in one form or another in all three color TV systems. The basic *YUV* color space can be generated from gamma-corrected *RGB* (referred to in equations as $R'G'B'$) components as follows:

$$\begin{aligned} Y &= 0.299R' + 0.587G' + 0.114B' \\ U &= -0.147R' - 0.289G' + 0.436B' = 0.492(B' - Y) \\ V &= 0.615R' - 0.515G' - 0.100B' = 0.877(R' - Y) \end{aligned} \tag{7}$$

The inverse operation, that is, generation of gamma-corrected *RGB* from *YUV* components, is accomplished by the following:

$$\begin{aligned} R' &= 1.0Y + 1.140V \\ G' &= 1.0Y - 0.394U - 0.580V \\ B' &= 1.0Y - 2.030U \end{aligned} \tag{8}$$

The Y, U, and V signals in PAL are multiplexed in a total bandwidth of either 5 or 5.5 MHz. With PAL, both U and V chrominance signals are transmitted with a bandwidth of 1.5 MHz. A color subcarrier is modulated with U and V via QAM and the composite signal is limited to the allowed frequency band which ends up truncating part of the QAM signal. The color subcarrier for PAL is located at 4.43 MHz. PAL transmits the V chrominance component as $+V$ and $-V$ on alternate lines. The demodulation of the QAM chrominance signal is similar to that of NTSC. The recovery of the PAL chrominance signal at the receiver includes averaging of successive demodulated scan lines to derive the U and V signals.

COMPONENT TELEVISION

In a component TV system, the luminance and chrominance signals are kept separate, such as on separate channels or multiplexed in different time slots. The use of a component system is intended to prevent the crosstalk that causes cross-luminance and cross-chrominance artifacts in the composite systems. The component system is preferable in all video applications that are without the constraints of broadcasting, where composite TV standards were made before the advent of high speed electronics.

Although a number of component signals can be used, of particular significance is the CCIR-601 digital component video format. The color Y,Cr,Cb space of this format is obtained by scaling and offsetting the Y,U,V color space. The conversion from gamma-corrected R, G, B components represented as eight-bits (0 to 255) to Y,Cr,Cb is specified as follows:

$$\begin{aligned} Y &= 0.257R' + 0.504G' + 0.098B' + 16 \\ Cr &= 0.439R' - 0.368G' - 0.071B' + 128 \\ Cb &= -0.148R' - 0.291G' + 0.439B' + 128 \end{aligned} \tag{9}$$

In these equations, Y is allowed to take values in the 16 to 235 range, whereas Cr and Cb can take values in the range of 16 to 240 centered at a value of 128, which indicates zero chrominance.

The inverse operation generates gamma-corrected RGB from Y,Cr,Cb components by:

$$\begin{aligned} R' &= 1.164(Y - 16) + 1.596(Cr - 128) \\ G' &= 1.164(Y - 16) - 0.813(Cr - 128) - 0.392(Cb - 128) \\ B' &= 1.164(Y - 16) + 2.017(Cb - 128) \end{aligned} \tag{10}$$

The sampling rates for the luminance component Y and the chrominance components are 13.5 MHz and 6.75 MHz, respectively. The number of active pels per line is 720, the number of active lines for the NTSC version (with 29.97 frames/s) is 486 and for the PAL version (with 25 frames/s) is 576. At eight bits/pel, the bit-rate of the uncompressed CCIR-601 signal is 216 Mbps.

Digitizing Video

Video cameras create either analog or sampled analog signals. The first step in processing, storage, or communication is usually to digitize the signals. Analog-to-digital converters with required accuracy and speed for video signals have become inexpensive in recent years. The cost and quality of digitization therefore is less of an issue. However, digitization with good quality results in a *bandwidth expansion,* in the sense that transmitting or storage of these bits often takes up more bandwidth or storage space than the original analog signal. In spite of this, digitization is becoming universal because of the relative ease of handling the digital signal compared to analog. In particular, enhancement, removal of artifacts, transformation, compression, encryption, integration with computers, and so forth is much easier to do in the digital domain using digital integrated circuits. One example of this is the conversion from one video standard to another (e.g., NTSC to PAL). Sophisticated adaptive algorithms required for good picture quality in standards conversion can be implemented only in the digital domain. Another example is the editing of digitized signals. Edits that require transformation (e.g., rotation, dilation of pictures, or time-warp for audio) are significantly more difficult in the analog domain. Additionally, encrypting bits is a lot easier and safer than encrypting analog signals. With digital storage, the quality of the retrieved signal does not degrade in an unpredictable manner with multiple reads as it often does with analog storage. Also, with today's database and user interface technology, a rich set of interactions is possible only with stored digital signals. Mapping the stored signal to displays with different resolutions in space (number of lines per screen and number of samples per line) and time (frame rates) can be done easily in the digital domain. A familiar example of this is the conversion of film, which is almost always at a different resolution and frame rate than the television signal.

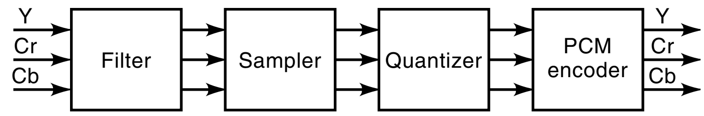


Figure 3. Conversion of component analog TV signals to digital TV signals.

Digital signals are also consistent with the evolving network infrastructure. Digital transmission allows much better control of the quality of the transmitted signal. In broadcast television, for example, if the signal were digital, the reproduced picture in the home could be identical to the picture in the studio, unlike the present situation where the studio pictures look far better than pictures at home. Finally, analog systems dictate that the entire television chain from camera to display operate at a common clock with a standardized display. In the digital domain, considerable flexibility exists by which the transmitter and the receiver can negotiate the parameters for scanning, resolution, and so forth, and thus create the best picture consistent with the capability of each sensor and display. The process of digitization of video consists of prefiltering, sampling, quantization, and encoding (see Fig. 3).

Filtering

This step is also referred to as prefiltering, since it is done prior to sampling. Prefiltering reduces the unwanted frequencies as well as noise in the signal. The simplest filtering operation involves simply averaging of the image intensity within a small area around the point of interest and replacing the intensity of the original point by the computed averaged intensity. Prefiltering can sometimes be accomplished by controlling the size of the scanning spot in the imaging system. In dealing with video signals, the filtering applied on the luminance signal may be different than that applied to chrominance signals owing to different bandwidths required.

Sampling

Next, the filtered signal is sampled at a chosen rate and locations on the image raster. The minimum rate at which an analog signal must be sampled is called the Nyquist rate and corresponds to twice that of the highest frequency in the signal. For NTSC system this rate is $2 \times 4.2 = 8.4$ MHz and for PAL this rate is $2 \times 5 = 10$ MHz. It is normal practice to sample at a rate higher than this for ease of signal recovery with practical filters. The CCIR-601 signal employs 13.5 MHz for luminance and half of that rate for chrominance signals. This rate is an integral multiple of both NTSC and PAL line rates but is not an integral multiple of either NTSC or PAL color subcarrier frequency.

Quantization

The sampled signal is still in analog form and is quantized next. The quantizer assigns each pel whose value is in a certain range a fixed value representing that range. The process of quantization results in loss of information since many input pel values are mapped into a single output value. The difference between the value of the input pel and its quantized representation is the quantization error. The choice of the number of levels of quantization involves a tradeoff of accuracy of representation and the resulting bit rate.

PCM Encoder

The last step in analog to digital conversion is encoding of quantized values. The simplest type of encoding is called pulse code modulation (PCM). Video pels are represented by eight-bit PCM codewords, that is, each pel is assigned one of the $2^8 = 256$ possible values in the range of 0 to 255. For example, if the quantized pel amplitude is 68, the corresponding eight-bit PCM codeword is the sequence of bit 01000100.

WHAT IS COMPRESSION?

Most video signals contain a substantial amount of "redundant" or superfluous information. For example, a television camera that captures 30 frames/s from a stationary scene produces very similar frames, one after the other. Compression removes the superfluous information so that a single frame can be represented by a smaller amount of finite data, or in the case of audio or time varying images, by a lower data rate (4,5).

Digitized audio and video signals contain a significant amount of *statistical redundancy,* that is, "adjacent" pels are similar to each other so that one pel can be predicted fairly accurately from another. By removing the predictable component from a stream of pels, the data rate can be reduced. Such statistical redundancy can be removed without loss of any information. Thus, the original data can be recovered exactly by inverse operation, called decompression. Unfortunately, the techniques for accomplishing this efficiently require probabilistic characterization of the signal. Although many excellent probabilistic models of audio and video signals have been proposed, serious limitations exist because of the nonstationarity of the statistics. In addition, video statistics may vary widely from application to application. A fast moving football game shows smaller frame-to-frame correlation compared to a head and shoulders view of people using video telephones. Current practical compression schemes do result in a loss of information, and lossless schemes typically provide a much smaller compression ratio (2 : 1 to 4 : 1).

The second type of superfluous data, called perceptual redundancy, is the information that a human visual system can not see. If the primary receiver of the video signal is a human eye (rather than a machine as in the case of some pattern recognition applications), then transmission or storage of the information that humans cannot perceive is wasteful. Unlike statistical redundancy, the removal of information based on the limitations of human perception is irreversible. The original data cannot be recovered following such a removal. Unfortunately, human perception is very complex, varies from person to person, and depends on the context and the application. Therefore, the art and science of compression still has many frontiers to conquer even though substantial progress has been made in the last two decades.

ADVANTAGES OF COMPRESSION

The biggest advantage of compression is in data rate reduction. Data rate reduction reduces transmission costs, and where a fixed transmission capacity is available, results in a better quality of video presentation (4). As an example, a single 6 MHz analog cable TV channel can carry between four and ten digitized, compressed, programs, thereby increasing the overall capacity (in terms of the number of programs carried) of an existing cable television plant. Alternatively, a single 6 MHz broadcast television channel can carry a digitized, compressed high definition television (HDTV) signal to give a significantly better audio and picture quality without additional bandwidth.

Data rate reduction also has a significant impact on reducing the storage requirements for a multimedia database. A CD-ROM can carry a full length feature movie compressed to about 4 Mbps. The lastest optical disk technology known as digital versatile disk (DVD), which is the same physical size as the CD, can store 4.7 GB of data on a single layer. This is more than seven times the capacity of a CD. Furthermore, the potential storage capabilities of DVD are even greater since it is possible to accommodate two layers of data on each side of the DVD resulting in 17 GB of data. The DVD can handle many hours of high quality MPEG2 video and Dolby AC3 audio. Thus, compression not only reduces the storage requirement, but also makes stored multimedia programs portable in inexpensive packages. In addition, the reduction of data rate allows transfer of video-rate data without choking various resources (e.g., the main bus) of either a personal computer or a workstation.

Another advantage of digital representation/compression is for packet communication. Much of the data communication in the computer world is by self-addressed packets. Packetization of digitized audio-video and the reduction of packet rate due to compression are important in sharing a transmission channel with other signals as well as maintaining consistency with telecom/computing infrastructure. The desire to share transmission and switching has created a new evolving standard, called asynchronous transfer mode (ATM), which uses packets of small size, called *cells.* Packetization delay, which could otherwise hinder interactive multimedia, becomes less of an issue when packets are small. High compression and large packets make interactive communication difficult, particularly for voice.

COMPRESSION REQUIREMENTS

The algorithms used in a compression system depend on the available bandwidth or storage 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) (4,5). Various issues arise in designing the compression system.

Quality

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

Table 1. Bit Rates of Compressed Video Signals

	Video Resolution (pels × lines × frames/s)	Uncompressed Bit Rate (RGB)	Compressed Bit Rate
NTSC video	(480 × 480 × 29.97 Hz)	168 Mbps	4 to 8 Mbps
PAL video	(576 × 576 × 25 Hz)	199 Mbps	4 to 9 Mbps
HDTV video	(1920 × 1080 × 30 Hz)	1493 Mbps	18 to 30 Mbps
HDTV video	(1280 × 720 × 60 Hz)	1327 Mbps	18 to 30 Mbps
ISDN videophone (CIF)	(352 × 288 × 29.97 Hz)	73 Mbps	64 to 1920 kbps
PSTN videophone (QCIF)	(176 × 144 × 29.97 Hz)	18 Mbps	10 to 30 kbps

most important aspects of video quality are spatial, temporal, and amplitude resolution. Spatial resolution describes the clarity or lack of blurring in the displayed image, while temporal resolution describes the smoothness of motion. Amplitude resolution describes graininess or other artifacts arising from coarse quantization.

Uncompressed versus Compressed Bitrates

The NTSC video has approximately 30 frames/s, 480 visible scan lines per frame and 480 pels per scan line in three color components. If each color component is coded using eight bits (24 bits/pel total), the bit rate would be approximately 168 Mbps. Table 1 shows the raw uncompressed bit rates for film, several audio, and video formats.

Robustness

As the redundancy from the video signal is removed by compression, each compressed bit becomes more important in the sense that it affects a large number of samples of the video signal. Therefore, an error either in transmission or storage of the compressed bit can have deleterious effects for either a large region of the picture or over an extended period of time.

For noisy digital transmission channels, video compression algorithms that sacrifice efficiency to allow for graceful degradation of the images in the presence of channel errors are better candidates. Some of these are created by merging source and channel coding to optimize the end-to-end service quality. A good example of this is portable video over a wireless channel. Here, the requirements on compression efficiency are severe owing to the lack of available bandwidth. Yet a compression algorithm that is overly sensitive to channel errors would be an improper choice. Of course, error correction is usually added to an encoded signal along with a variety of error concealment techniques, which are usually successful in reducing the effects of random isolated errors. Thus, the proper choice of the compression algorithm depends on the transmission environment in which the application resides.

Interactivity

Both consumer entertainment and business video applications are characterized by picture switching and browsing. In the home, viewers switch to the channels of their choice. In the business environment, people get to the information of their choice by random access using, for example, on-screen menus. In the television of the future, a much richer interaction based on content rather than channel switching may become possible.

Many multimedia offerings and locally produced video programs often depend on the concatenation of video streams from a variety of sources, sometimes in real time. Commercials are routinely inserted into nationwide broadcasts by network affiliates and cable headends. Thus, the compression algorithm must support a continuous and seamless assembly of these streams for distribution and rapid switching of images at the point of final decoding. It is also desirable that simple edits as well as richer interactions occur on compressed data rather than reconstructed sequences.

In general, a higher degree of interactivity requires a compression algorithm that operates on a smaller group of pels. MPEG, which operates on spatio-temporal groups of pels, is more difficult to interact with than JPEG, which operates only on spatial groups of pels. As an example, it is much easier to fast forward a compressed JPEG bitstream than a compressed MPEG bitstream. This is one reason why the current digital camcorders are based on motion JPEG. In a cable/broadcast environment or in an application requiring browsing through a compressed multimedia database, a viewer may change from program to program with no opportunity for the encoder to adapt itself. It is important that the buildup of resolution following a program change take place quite rapidly so that the viewer can make a decision to either stay on the program or change to another depending on the content.

Compression and Packetization Delay

Advances in compression have come predominantly through better analysis of the video signal arising from the application in hand. As models have progressed from pels to picture blocks to interframe regions, efficiency has grown rapidly. Correspondingly, the complexity of the analysis phase of encoding has also grown, resulting in the increase of encoding delay. A compression algorithm that looks at a large number of samples and performs very complex operations usually has a larger encoding delay.

For many applications, such encoding delay at the source is tolerable, but for some it is not. Broadcast television, even in real time, can often admit a delay in the order of seconds. However, teleconferencing or multimedia groupware can tolerate a much smaller delay. In addition to the encoding delay, modern data communications introduce packetization delay. The more efficient the compression algorithm, the larger is the delay introduced by packetization, since the same size packet carries information about many more samples of the video signal.

Symmetry

A cable, satellite, or broadcast environment has only a few transmitters that compress, but a large number of receivers that have to decompress. Similarly, video databases that store information usually compress it only once. However, the retrieval of this information may happen thousands of times by different viewers. Therefore, the overall economics of many applications is dictated to a large extent by the cost of decompression. The choice of the compression algorithm ought to make the decompression extremely simple by transferring much of the cost to the transmitter, thereby creating an asymmetrical algorithm. The analysis phase of a compression algorithm, which routinely includes motion analysis (done only at the encoder), naturally makes the encoder more expensive. In a number of situations, the cost of the encoder is also important (e.g., camcorder, videotelephone). Therefore, a modular design of the encoder that is able to trade off performance with complexity, but that creates data decodable by a simple decompressor, may be the appropriate solution.

Multiple Encoding

In a number of instances, the original signal may have to be compressed in stages or may have to be compressed and decompressed several times. In most television studios, for example, it is necessary to store the compressed data and then decompress it for editing as required. Such an edited signal is then compressed and stored again. Any multiple coding-decoding cycle of the signal is bound to reduce the quality of the signal, since artifacts are introduced every time the signal is coded. If the application requires such multiple codings, then a higher quality compression is required, at least in the several initial stages.

Scalability

A compressed signal can be thought of as an alternative representation of the original uncompressed signal. From this alternative representation, it is desirable to create presentations at different resolutions (in space, time, amplitude, etc.) consistent within the limitations of the equipment used in a particular application. For example, if a HDTV signal compressed to 24 Mbps can be simply processed to produce a lower resolution and lower bitrate signal (e.g., NTSC at 6 Mbps), the compression is generally considered to be scalable. Of course, the scalability can be achieved in a brute force manner by decompressing, reducing the resolution, and compressing again. However, this sequence of operations introduces delay and complexity, and results in a loss of quality. A common compressed representation from which a variety of low-resolution or higher resolution presentations can be easily derived is desirable. Such scalability of the compressed signal puts a constraint on the compression efficiency in the sense that algorithms with the highest compression efficiency usually are not very scalable.

BASIC COMPRESSION TECHNIQUES

A number of compression techniques have been developed for coding of video signals (1). A compression system typically consists of a combination of these techniques to satisfy the type of requirements that we listed in the previous section. The first step in compression usually consists of decorrelation that is, reducing the spatial or temporal redundancy in the signal (4,5). The candidates for doing this are:

1. Making a prediction of the next sample of the picture signal using some of the past and subtracting it from that sample. This converts the original signal into its unpredictable part (usually called prediction error).
2. Taking a transform of a block of samples of the picture signal so that the energy would be compacted in only a few transform coefficients.

The second step is selection and quantization to reduce the number of possible signal values. Here, the prediction error may be quantized sample at a time or a vector of prediction error of many samples may be quantized all at once. Alternatively, for transform coding, only important coefficients may be selected and quantized. The final step is entropy coding which recognizes that different values of the quantized signal occur with different frequencies and, therefore, representing them with unequal length binary codes reduces the average bit rate. We give below more details of the following techniques since they have formed the basis of most of the compression systems;

- Predictive coding (DPCM)
- Transform coding
- Motion compensation
- Vector quantization
- Subband/Wavelet coding
- Entropy coding
- Incorporation of perceptual factors

Predictive Coding (DPCM)

In predictive coding, the strong correlation between adjacent pels (spatially as well as temporally) is exploited (4). As shown in Fig. 4, an approximate prediction of the sample to be encoded is made from previously coded information that has already been transmitted. The error (or differential signal) resulting from the subtraction of the prediction from the actual value of the pel is quantized into a set of discrete amplitude levels. These levels are then represented as binary words of fixed or variable lengths and sent to the channel for

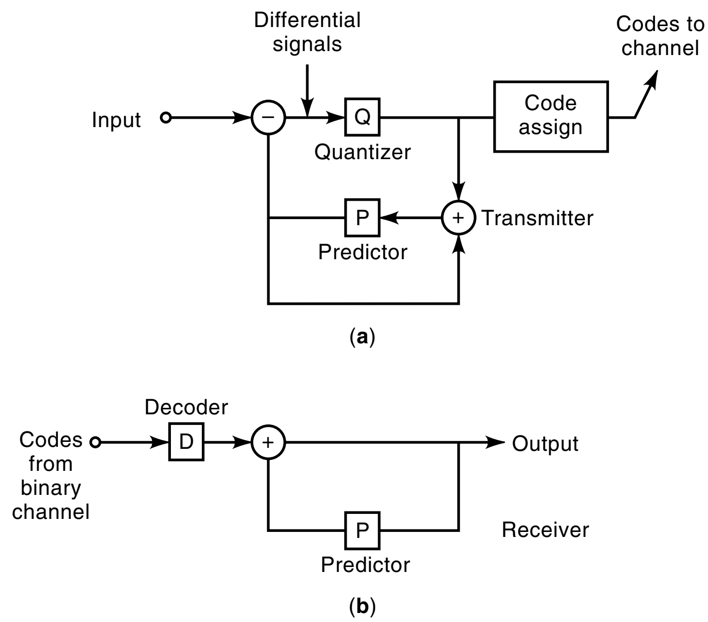


Figure 4. Block diagram of a predictive encoder and decoder.

transmission. The predictions may make use of the correlation in the same scanning line or adjacent scanning lines or previous fields. A particularly important method of prediction is the motion compensated prediction. If a television scene contains moving objects and an estimate of frame-to-frame translation of each moving object is made, then more efficient prediction can be performed using elements in the previous frame that are appropriately spatially displaced. Such prediction is called motion compensated prediction. The translation is usually estimated by matching a block of pels in the current frame to a block of pels in the previous frames at various displaced locations. Various criteria for matching and algorithms to search for the best match have been developed. Typically, such motion estimation is done only at the transmitter and the resulting motion vectors are used in the encoding process and also separately transmitted for use in the decompression process.

Transform Coding

In transform coding (Fig. 5) a block of pels are transformed by transform T into another domain called the transform domain, and some of the resulting coefficients are quantized and coded for transmission. The blocks may contain pels from one, two, or three dimensions. The most common technique is to use a block of two dimensions. Using one dimension does not exploit vertical correlation and using three dimensions requires several frame stores. It has been generally agreed that discrete cosine transform (DCT) is best matched to the statistics of the picture signal and moreover, since it has a fast implementation, it has become the transform of choice. The advantage of transform coding (4) comes about mainly from two mechanisms. First, not all of the transform coefficients need to be transmitted in order to maintain good image quality, and second, the coefficients that are selected need not be represented with full accuracy. Loosely speaking, transform coding is preferable to predictive coding for lower compression rates and where cost and complexity are not extremely serious issues. Most modern compression systems have used a combination of predictive and transform coding. In fact, motion compensated prediction is performed first to remove the temporal redundancy, and then the resulting prediction error is compressed by two-dimensional transform coding using discrete cosine transform as the dominant choice.

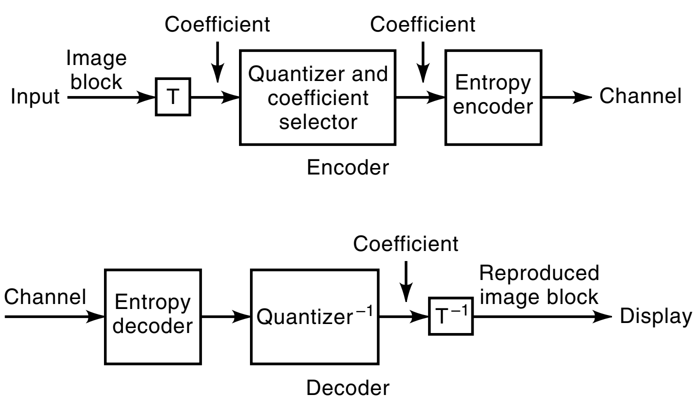


Figure 5. Block diagram of a transform coder.

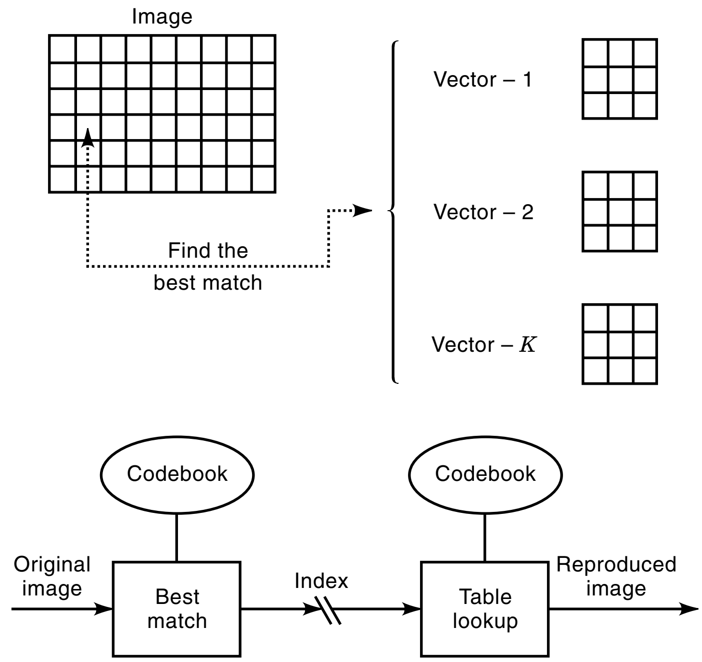


Figure 6. Block diagram of vector quantization.

Vector Quantization

In predictive coding, described in the previous section, each pixel was quantized separately using a scalar quantizer. The concept of scalar quantization can be generalized to vector quantization (5) in which a group of pixels are quantized at the same time by representing them as a code vector. Such a vector quantization can be applied to a vector of prediction errors, original pels, or transform coefficients. As in Fig. 6, a group of nine pixels from a 3×3 block is represented to be one of the k vectors from a codebook of vectors. The problem of vector quantization is then to design the codebook and an algorithm to determine the vector from the codebook that offers the best match to the input data. The design of codebook usually requires a set of training pictures and can grow to a large size for a large block of pixels. Thus, for an 8×8 block compressed to two bits per pel, one would need a 2^{128} size codebook. Matching the original image with each vector of such a large size codebook requires a lot of ingenuity. However, such matching is only done at the transmitter, and the receiver is considerably simple since it does a simple table lookup.

Subband/Wavelet Coding

Subband coding, more recently generalized using the theory of wavelets, is a promising technique for video and has already been shown to outperform still image coding techniques based on block transforms such as in JPEG. Although subband techniques have been incorporated into audio coding standards, the only image standard based on wavelets currently is the FBI standard for fingerprint compression. There are several compelling reasons to investigate subband/wavelet coding for image and video compression. One reason is that unlike the DCT, the wavelet framework does not transform each block of data separately. This results in a graceful degradation as the bit rate is lowered without the traditional "tiling effect" that is characteristic of block-based approaches. Wavelet coding also allows one to work in a multiresolution framework which is a natural choice for progressive transmission or applications where scalability is desirable. One of the current weaknesses in deploying wavelet schemes for video compression is the fact that a major component for efficient video compression is block-based motion estimation which makes the block-based DCT a natural candidate for encoding the spatial information.

Entropy Coding

If the quantized output values of either a predictive or a transform coder are not all equally likely, then the average bit rate can be reduced by giving each one of the values a different word length. In particular, those values that occur more frequently are represented by a smaller length code word (4,5). If a code with variable length is used, and the resulting code words are concatenated to form a stream of bits, then correct decoding by a receiver requires that every combination of concatenated code words be uniquely decipherable. A variable word length code that achieves this and at the same time gives the minimum average bit rate is called Huffman code. Variable word length codes are more sensitive to the effect of transmission errors since synchronization would be lost in the event of an error. This can result in several code words getting decoded incorrectly. A strategy is required to limit the propagation of errors when Huffman codes are used.

Incorporation of Perceptual Factors

The perception based coding attempts to match the coding algorithm to the characteristics of human vision. We know, for example, that the accuracy with which the human eye can see the coding artifacts depends upon a variety of factors such as the spatial and temporal frequency, masking due to the presence of spatial or temporal detail, and so on. A measure of the ability to perceive the coding artifact can be calculated based on the picture signal. This is used, for example, in transform coding to determine the precision needed for quantization of each coefficient. Perceptual factors control the information that is discarded on the basis of its visibility to the human eye. It can, therefore, be incorporated in any of the previously stated basic compression schemes.

Comparison of Techniques

Figure 7 represents an approximate comparison of different techniques using compression efficiency versus complexity as

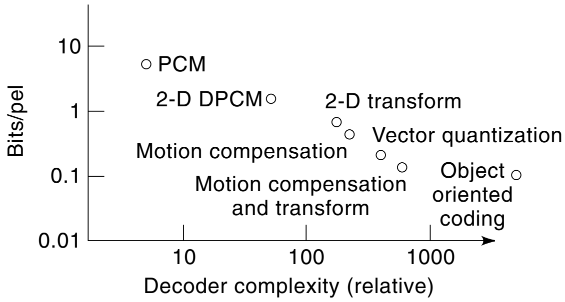


Figure 7. Bits/pel versus complexity of video decoding for several video compression algorithms.

a criterion under the condition that the picture quality is held constant at an eight-bit PCM level. The complexity allocated to each codec is an approximate estimate relative to the cost of a PCM codec which is given a value of 5. Furthermore, it is the complexity of only the decoder portion of the codec, since that is the most important cost element for digital television. Also, most of the proposed systems are a combination of several different techniques of Fig. 7, making such comparisons difficult. As we remarked before, the real challenge is to combine the different techniques to engineer a cost-effective solution for a given service. The next section describes one example of such a codec.

A COMPRESSION SCHEME

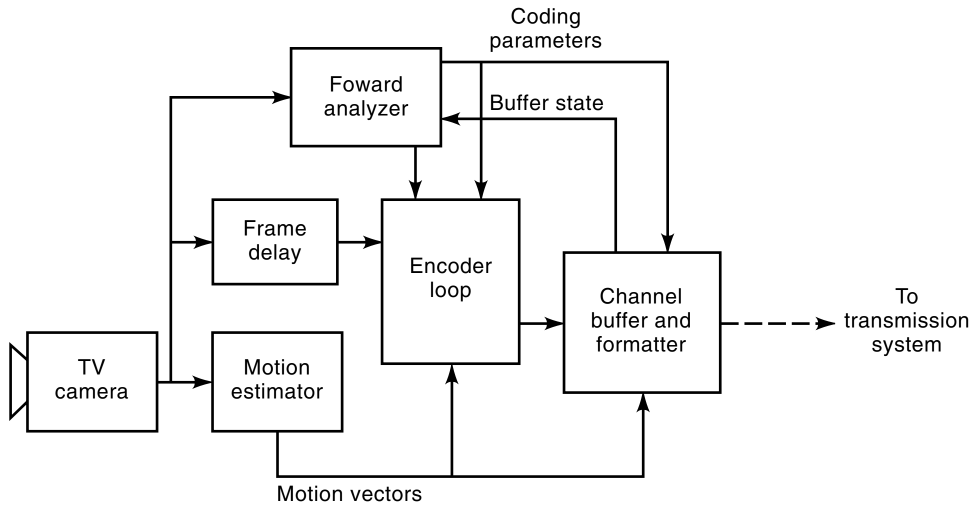
In this section we describe a compression scheme that combines the previous basic techniques to satisfy the requirements that follow.

Three basic types of redundancy are exploited in the video compression process. Motion compensation removes temporal redundancy, two-dimensional DCT removes spatial redundancy, and perceptual weighting removes amplitude irrelevancy by putting quantization noise in less visible areas.

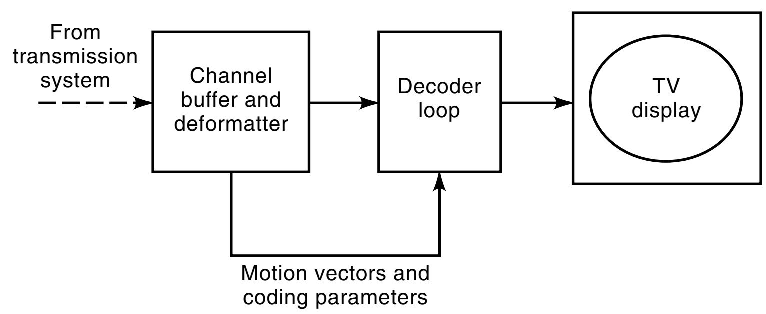
Temporal processing occurs in two stages. The motion of objects from frame-to-frame is estimated using hierarchical block matching. Using the motion vectors, a displaced frame difference (DFD) is computed which generally contains a small fraction of the information in the original frame. The DFD is transformed using DCT to remove the spatial redundancy. Each new frame of DFD is analyzed prior to coding to determine its rate versus perceptual distortion characteristics and the dynamic range of each coefficient (forward analysis). Quantization of the transform coefficients is performed based on the perceptual importance of each coefficient, the precomputed dynamic range of the coefficients, and the rate versus distortion characteristics. The perceptual criterion uses a model of the human visual system to determine a human observer's sensitivity to color, brightness, spatial frequency, and spatial-temporal masking. This information is used to minimize the perception of coding artifacts throughout the picture. Parameters of the coder are optimized to handle the scene changes that occur frequently in entertainment/sports events, and channel changes made by the viewer. The motion vectors, compressed transform coefficients, and other coding overhead bits are packed into a format which is highly immune to transmission errors.

The encoder is shown in Fig. 8(a). Each frame is analyzed before being processed in the encoder loop. The motion vectors and control parameters resulting from the forward analysis are input to the encoder loop which outputs the compressed prediction error to the channel buffer. The encoder loop control parameters are weighed by the buffer state which is fed back from the channel buffer.

In the predictive encoding loop, the generally sparse differences between the new image data and the motion-compensated predicted image data are encoded using adaptive DCT coding. The parameters of the encoding are controlled in part by forward analysis. The data output from the encoder consists of some global parameters of the video frame computed by the forward analyzer and transform coefficients that have



(a)



(b)

Figure 8. Block diagram of an encoder/decoder.

been selected and quantized according to a perceptual criterion.

Each frame is composed of a luminance frame and two chrominance difference frames which are half the resolution of the luminance frame horizontally. The compression algorithm produces a chrominance bit-rate which is generally a small fraction of the total bit-rate, without perceptible chrominance distortion.

The output buffer has an output rate of between 2 to 7 Mbps and has a varying input rate that depends on the image content. The buffer history is used to control the parameters of the coding algorithm so that the average input rate equals the average output rate. The feedback mechanism involves adjustment of the allowable distortion level, since increasing the distortion level (for a given image or image sequence) causes the encoder to produce a lower output bit rate.

The encoded video is packed into a special format before transmission which maximizes immunity to transmission errors by masking the loss of data in the decoder. The duration and extent of picture degradation due to any one error or group of errors is limited. The decoder is shown in Fig. 8(b). The compressed video data enters the buffer which is complementary to the compressed video buffer at the encoder. The decoding loop uses the motion vectors, transform coefficient data, and other side information to reconstruct the NTSC images. Channel changes and severe transmission errors are detected in the decoder causing a fast picture recovery process to be initiated. Less severe transmission errors are handled gracefully by several algorithms depending on the type of error.

Processing and memory in the decoder are minimized. Processing consists of one inverse spatial transform and a variable length decoder which are realizable in a few very large scale integration (VLSI) chips. Memory in the decoder consists of one full frame and a few compressed frames.

COMPLEXITY/COST

Since cost is directly linked to complexity, this aspect of a compression algorithm is the most critical for the asymmetrical situations described previously. The decoder cost is most critical. Figure 7 represents an approximate tradeoff between the compression efficiency and the complexity under the condition that picture quality is held constant at an eight-bit PCM level. The compression efficiency is in terms of compressed bits per Nyquist sample. Therefore, pictures with different resolution and bandwidth can be compared simply by proper multiplication to get the relevant bitrates. The complexity allocated to each codec should not be taken too liter-

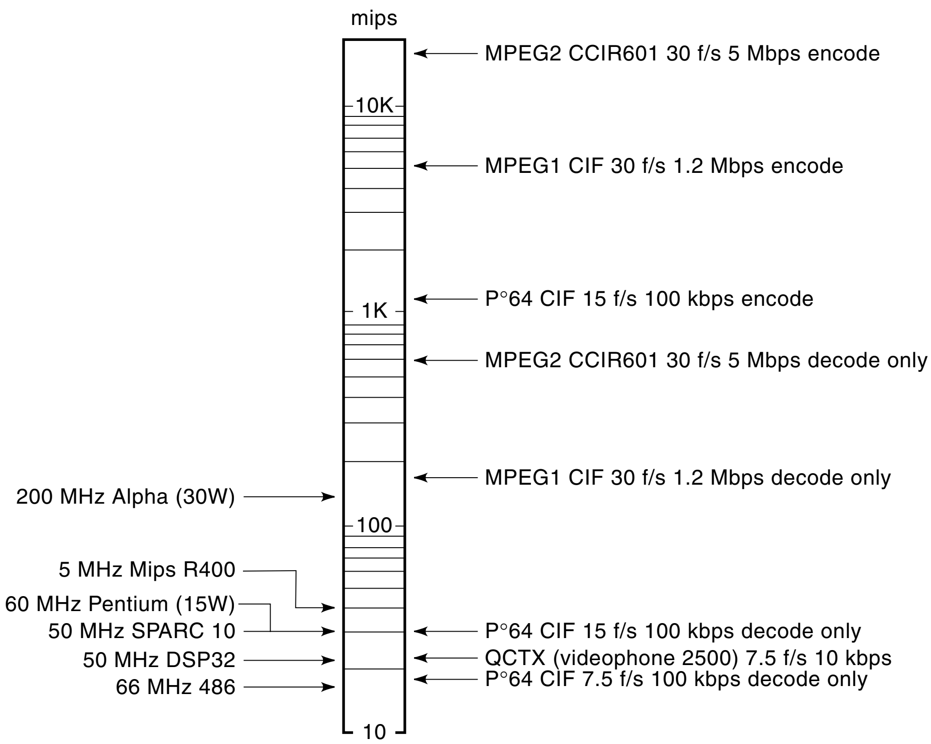


Figure 9. Computational requirements in millions of instructions per second (mips) for video encoding and decoding at different image resolutions.

ally. Rather, it is an approximate estimate relative to the cost of a PCM codec, which is given a value of 5.

The relation of cost to complexity is controlled by an evolving technology, and codecs with high complexity are quickly becoming inexpensive through the use of application-specific video DSPs and submicron device technology. In fact, very soon fast microprocessors will be able to decompress the video signal entirely in software. It is clear that in the near future a standard resolution (roughly 500 line by 500 pel TV signal) will be decoded entirely in software for even the MPEG compression algorithm. Figure 9 shows video encoding and decoding at various image resolutions.

VIDEOPHONE AND COMPACT DISK STANDARDS—H.320 AND MPEG-1

Digital compression standards (DCS) for video conferencing were developed in the 1980s by the CCITT, which is now known as the ITU-T. Specifically, the ISDN video conferencing standards are known collectively as H.320, or sometimes P*64 to indicate that it operates at multiples of 64 kbits/s. The video coding portion of the standard is called H.261 and codes pictures at a common intermediate format (CIF) of 352 pels by 288 lines. A lower resolution of 176 pels by 144 lines, called QCIF, is available for interoperating with PSTN videophones. H.263 standard is built upon the H.261 framework but modified to optimize video quality at rates lower than 64kb/s. H.263+ is focused on adding features to H.263 such as scalability and robustness to packet loss on packet networks such as the Internet.

In the late 1980s, a need arose to place motion video and its associated audio onto first generation CD-ROMs at 1.4 Mbps. For this purpose, in the late 1980s and early 1990s, the ISO MPEG committee developed digital compression standards for both video and two-channel stereo audio. The standard is known colloquially as MPEG-1 and officially as ISO 11172. The bit rate of 1.4 Mbps available on first generation CD-ROMs is not high enough to allow for full-resolution TV. Thus, MPEG-1 was optimized for the reduced CIF resolution of H.320 video conferencing. It was designed to handle only the progressive formats, later MPEG-2 incorporated progressive as well as interlaced formats effectively.

THE DIGITAL ENTERTAINMENT TV STANDARD—MPEG-2

Following MPEG-1, the need arose to compress entertainment TV for such transmission media as satellite, cassette tape, over-the-air, and CATV (5). Thus, to have available digital compression methods for full-resolution standard definition TV (SDTV) pictures such as shown in Fig. 4(a) or high definition TV (HDTV) pictures such as shown in Fig. 4(b), ISO (International Standard Organization) developed a second standard known colloquially as MPEG-2 and officially as ISO 13818. Since the resolution of entertainment TV is approximately four times that of videophone, the bit rate chosen for optimizing MPEG-2 was 4 Mbps.

SUMMARY

A brief survey of digital television has been presented in this article. Digitizing television, and compressing it to manageable bit rate, creates significant advantages and major disruption in existing television systems. The future is bright for a variety of systems based on digital television technology.

BIBLIOGRAPHY

1. P. Mertz and F. Gray, A theory of scanning and its relation to the characteristics of the transmitted signal in telephotography and television, *Bell Syst. Tech. J.,* **13**: 464–515, 1934.
2. W-T. Wintringham, Color television and colorimetry, *Proc. IRE,* **39** (10): 1951.
3. K. B. Benson (Ed.), *Television Engineering Handbook,* New York: McGraw-Hill, 1986.
4. A. Netravali and B. G. Haskell, *Digital Pictures,* New York: Plenum, 1988.
5. B. G. Haskell, A. Puri, and A. N. Netravali, *Digital Video: An Introduction to MPEG-2,* London: Chapman & Hall, 1996.

ARUN N. NETRAVALI
Bell Labs, Lucent Technologies

DIGITAL TELEVISION STANDARDS.

See TELEVISION BROADCAST TRANSMISSION STANDARDS.