Fifth generation television

By Andrew Lippman

Television is an artifact. There is nothing fundamental about its design or uses that cannot be changed in a new generation; no parameters inextricably molded into it or cows that are sacred for any reason other than past practice. Television is an artifact because its parameters date from another era in electronic design, and its uses—popular as they seem to be—derive only partly from some innate human need that cannot be satisfied any other way or is archetypical. In fact, we are at the threshold of a new generation of television systems and every aspect is open to question and change, including the lines, the frames, the innards of the set, and even its physical location in our households and offices.

The history of television is characterized by an alternation of technical breakthroughs interspersed with periods of relative systemic stability and growth as a medium for program distribution. Major innovations occur roughly every 30 years and include the invention of mechanical scanning in 1884, electronic displays and cameras in the 1920s, compatible color broadcasting in the 1950s, and little else.

High Definition Television (HDTV), which dates from the early 1970s, is the evident technical innovation of this era. As canonically defined, it is little more than a scaling of the basic parameters of existing television by a factor of two. If anything, it is the last incarnation of the first generation of television rather than the first embodiment of the next. This time, however, we may be able to break the mold of successive changes in use and technology and progress to a new generation of television systems that implies new quality as well as new program styles. What is it that allows us to do better now? What are the real opportunities? Why is a new generation of television possible or even inevitable now?

Why is a new generation of television possible or even inevitable now?

The answer to these questions lies in several areas, but predominantly, it is the inevitable intersection of computing with TV, and a profusion of non-entertainment applications ranging from desktop video simulations to Jupiter Fly-Bys. The computer changes the ground rules of system design by virtue of its ability to process both the signal and the content. In addition, it is an inherently personal device, shifting control over what is viewed from a central disseminator to the user. Its programmability is implicit, and user control of the program is well precedented and unavoidable.

What is not apparent, and perhaps requires evolution of content styles as well as technology, is the manner in which traditional entertainment viewing can evolve, given the new degrees of freedom possible in the medium. At first glance, it seems obvious that workstations will incorporate television imagery, but the way in which computing will manifest itself in consumer receivers is a little more obscure.

In the following sections, we present a few options and some arguments about their likelihood. None are apt to be absolutely correct, since there are too many societal and imaginative unknowns. But neither are they outlandish or necessarily way off base. Some of the technological imperatives are in place or about to happen. What follows is as much a research agenda as a series of conclusions.

Open architecture television

The design of an HDTV receiver or a workstation that can display High Definition Television (HDTV) is complicated by the fact that there is no universal definition of the

ANDREW LIPPMAN is associate director of the MIT Media Laboratory at the Massachusetts Institute of Technology, Cambridge, Mass.

term. In general, the international television community accepts the notion that HDTV entails a signal comprising approximately twice the spatial resolution of current television, both horizontally and vertically, and with a slightly greater aspect ratio. Much of this is based upon work initiated by NHK in Japan and reported by Fujio et al.,[1] in which a series of subjective experiments seemingly proved that maximal viewer acceptance was achieved when the screen approximates the cinematic aspect ratio of 5:3 and the scanning density approaches 1100 lines per picture height.

Subsequent to this research, the development of a picture standard with 1125 lines, an aspect ratio of 5:3, and approximately 30 MHz of baseband bandwidth was proposed for standardization of television production. Compressed versions of the signal requiring 8.4 MHz and, eventually, 8.1-MHz—called MUSE[2]—were proposed as transmission systems for this high bandwidth video system. MUSE has since been developed into a family of distribution systems, only peripherally related in their basic technological approach. Some augment NTSC transmission within the 6-MHz band[a]; others substitute a new, incompatible, 6-MHz signal.[b]

In response to this challenge, Europe launched a program in 1986[c] to develop an alternative suited to the European environment and consonant with existing European broadcasts. This system is tied to a production standard operating at a 50 Hz frame rate and with exactly twice the number of lines as PAL and SECAM broadcasts.

The result of this is the potential for a new array of television systems, each parochial in their application, in which the immediate future of television fails to achieve universal high quality program interchange or production, and the regional transmission standard divides rather than unifies.

It appears that international discord will prevent any of these systems from being universally accepted, and fundamental problems of signal conversion that originate at one frame (or field) rate to another will complicate matters rather than ameliorate them. Further, the fact that each system initially will employ interlaced scanning means that the tie to the past (when interlace was an available solution to bandwidth compression) is retained, complicating high quality production and obscuring the simple notion of a frame.

The situation is further complicated by the fact that there is no basis for assuming that a simple doubling of spatial resolution or widening of the image area will result in widespread viewer approbation. Transmission degradations appear to be as significant as line count, and extra width in the display can appear as a lack of height as well as cinematic similitude. Further, with direct view CRT displays, brightness suffers with increased resolution, thus confronting the consumer with a Hobson's choice between a sharp or a vivid picture.

In terms of computer workstations, HDTV is distinctly *not* a peripheral issue. Although many now operate at line rates close to HDTV and an increasing number are capable of video display within a window, subtle differences between defacto computer standards and consumer video can have far-reaching effects. The mere fact of a change in the shape of the CRT can impact future workstation designs, and the diminution of the difference between professional and consumer systems can be impeded by particular HDTV designs. In a day when animation and moving images are rapidly becoming the norm rather than the exception, the simple prerequisite of a videotape system interlaced to a computer is a paramount concern.

One way to rationalize the diverse interests of the various political and functional constituencies associated with television is to avoid a standard measured in lines and fields, replacing it with a format amenable to an evolving set of resolutions, sampling rates, and displays. We call such an approach Open Architecture Television.

Open architecture television is as much a matter of signal definition as receiver design. In implementation, we suggest a receiver that is more like a personal computer than a traditional television set, but it need not be so complicated. Indeed, just as there have been computers bundled with a fixed set of programs resident in internal read-only memory, one could build a receiver with little potential for either expansion or user-controllable processing.

The theme is more subtle: a new generation of consumer equipment should be amenable to an evolutionary set of scanning parameters, where scan line count is proportional to screen size and viewing distance and display characteristics are divorced from production and transmission parameters. Put another way, the 12″ television on the kitchen counter need not have as many lines as the four-foot wall-hung screen in the living room, and neither should it encompass the same processing elements. In terms of evolution, both should be capable of accepting a near-term standard used for over-the-air broadcast, as

[a] These are called EDTV systems, for Extended Definition Television. In EDTV systems, the normal receiver can decode a normal program, but a special receiver can decode augmentation information that results in a higher definition image.

[b] New use of an existing channel allocation is usually referred to as a "simulcast system" because it is presumed that a normal NTSC version of the program would be distributed through existing channels and new ones would be allocated for this new signal. The new signal occupies precisely one television channel.

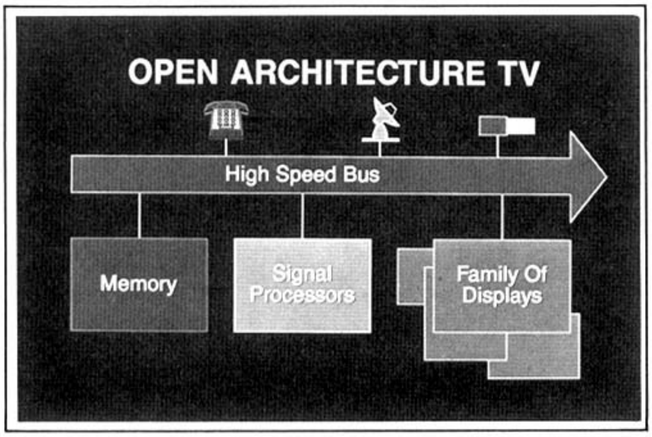
[c] Eureka-95, dedicated to the development of a uniquely European HDTV system.

well as higher fidelity signals originating locally or through proprietary distribution channels not legislatively bandwidth limited (see Fig. 1).

The key is in the signal definition itself. As noted above, frames, lines and picture elements are an artifact. In America, we chose 60 frames per second to simplify early receiver power supply design; Europe chose 50 for similar reasons. Cinema, with a carbon arc lamphouse, chose a rate dictated by the physical characteristics of sprocketed film and speeds sufficient to support optical sound tracks.

Lack of storage capability and concomitant processing mandated receivers that operate in lock step with the transmission standard, but that is no longer a limitation. It is reasonable and technologically possible to design a transmission and storage format independent of all of these parameters. This is one of the few cases where there is a technological solution to existing international system incompatibilities. We can design a video representation that favors neither 50 Hz, nor 60, and can display video at each rate on a screen operating at neither.

We suggest a digital intermediate format (DIF) that is a subband transformation of the time series of video frames. This format results from a spectral decomposition of the video frame sequence into a set of spatio-temporal components by filtering and subsampling.[3] In particular, the image sequence is filtered by a tree-structured array of separable bandsplitting filters alternately applied in horizontal, vertical, and temporal orientation. The resulting subbands are then subsampled to reflect the new information content of each band.[4–6] We have applied this to video signals for HDTV as well as for low bandwidth computer display; an example of the latter is shown in Fig. 2. Other alternative representations such as transforms, motion fields, and keyframes, or even explicit picture element representations, are candidates, but the spectral components offer several advantages:


OPEN ARCHITECTURE TV

High Speed Bus

Memory

Signal Processors

Family Of Displays

FIGURE 1. Open Architecture Television: An open architecture receiver consists of a number of extensible components that allow the set to grow into a sophisticated picture interaction system or to adapt to new program distribution formats. Internally, all signals are in an intermediate format that spans all possible spatial and temporal resolutions and is easily transcodable for high and low resolution displays.

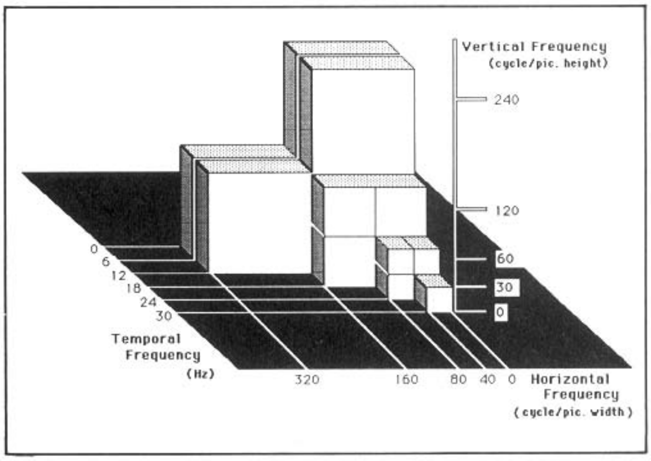


FIGURE 2. Digital Intermediate Format (shown here for low-bandwidth video compression): The internal signal representation in open architecture television is a subband decomposition of the image sequence. The three-dimensional spectrum is divided into a number of regions, called subbands, each of which is then represented with fidelity in proportion to energy content and visual sensitivity to errors in that band. Not all source material fully populates the three-dimensional space—for example, 24 fps film has no energy above 12 Hz.

■ *Direct interface to high bandwidth production standards*—The DIF can serve as a digital representation of the video signal for high bandwidth production systems. It can interface directly to video recording equipment and optical networks.

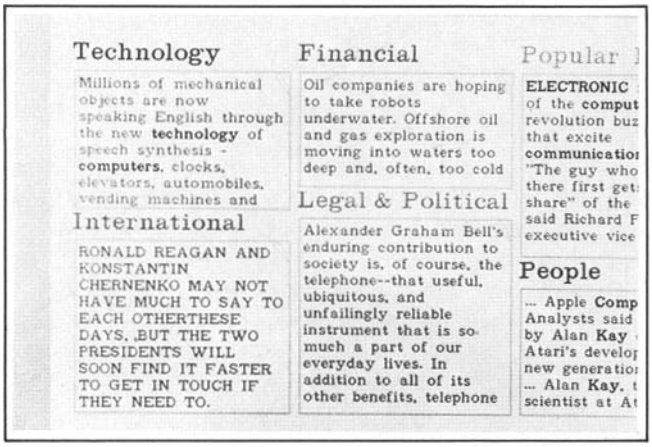
■ *Multi-standard*—DIF allows images derived from diverse standards to be simultaneously displayed.

■ *Multiple displays*—Various superficially incompatible displays can be attached to the bus, as described above.

■ *Scalability*—The DIF can be upgraded to suit a variety of system costs and architectures.

■ *Data compression*—The component representation facilitates representation of each component at different signal-to-noise ratios, thus allowing inherent compression that is not available in an explicit frame series video representation.

■ *Graceful degradation*—System overload can be handled



Technology

Millions of mechanical
objects are now
speaking English through
the new **technology** of
speech synthesis -
computers, clocks,
elevators, automobiles,
vending machines and

International

RONALD REAGAN AND
KONSTANTIN
CHERNENKO MAY NOT
HAVE MUCH TO SAY TO
EACH OTHERTHESE
DAYS, ,BUT THE TWO
PRESIDENTS WILL
SOON FIND IT FASTER
TO GET IN TOUCH IF
THEY NEED TO.

Financial

Oil companies are hoping
to take robots
underwater. Offshore oil
and gas exploration is
moving into waters too
deep and, often, too cold

Legal & Political

Alexander Graham Bell's
enduring contribution to
society is, of course, the
telephone--that useful,
ubiquitous, and
unfailingly reliable
instrument that is so
much a part of our
everyday lives. In
addition to all of its
other benefits, telephone

Popular

ELECTRONIC
of the **comput**
revolution buz
that excite
communicatio
"The guy who
there first get
share" of the
said Richard F
executive vice

People

... Apple **Comp**
Analysts said
by **Alan Kay**
Atari's develo
new generatio
... **Alan Kay**,
scientist at A

FIGURE 3. NewsPeek: In this program, print information and television material are combined to make a personalized newspaper, read either on a workstation/TV screen or printed. When read on line, the larger page image is scrolled over by touching the screen, and the reader can activate illustrations or ask for more information about any articles.

by deleting components rather than by deleting frames.

More recently, Schreiber[7] has proposed that a subband decomposition of the video signal based upon a temporal rate of multiples of 12 Hz can serve as a unifying standard for international television program exchange. In this "Friendly Family," diverse standards such as 24 fps film, 50 Hz and 60 Hz video are all transformed to the subband representation where the particular origination standard determines what bands have information content and which are empty. The subband representation provides a format whereby these seemingly incompatible standards may be processed in common and presented for display. The DIF extends this notion to include computer generated video and to accept displays operating at none of the incoming standards.

To date, the digital intermediate format exists as a set of experiments. We have simulated it using motion compensated interpolation to extend 24 Hz film to fill subbands up to 72 Hz, and we have jointly presented film and television on both 60 Hz and 72 Hz displays with little quality difference. We have also repeated these experiments without motion compensation. In this case, the signals exhibit motion blur at high frame rates, but in approximately equal amounts on all displays. In addition, we have prepared examples of bandlimiting the signal in amounts analogous to the difference between professional videotape recording and consumer cassette standards. The transition from full-band images to reduced bandwidth occurs with markedly less evident artifacts when spatiotemporal subbands are selectively deleted than when a two-dimensional decimation is used.

Program options

It is conventional wisdom that television is a "couch potato" medium. One turns the set on in the evening and occasionally glances at it through a haze of beer and conversation. Little mental or physical activity is presumed. This has never been completely true (even children don't suppress cognitive activity when watching) and it is becoming less common as local storage in the home (videocassette recorders) shifts control from the programmer to the viewer.

There is an increasing amount of what is colloquially called "zipping," "zapping," or "grazing" associated with VCR operation and the presence of a large number of cable-provided program choices. This is nascent computational viewing—the audience is creating a "designer channel" on-the-fly—but the computing agency is the viewer himself, operating through a restricted interface—either the channel control or hardwired VCR scan options. However, the program is an artifact; new types of programs can result when computing in the receiver operates on the intent rather than the waveform of the video signal. Indeed, computers that process content predate those that operate on signals—text processing manipulated words before letterforms—but video rate computers have only recently become available, and HDTV has helped propel them into television receivers.

Content processing implies a television receiver that gathers material based on user preference, presenting it in realtime or recalling it on demand. The receiver can thus include form and content controls as well as image controls, tuning the set for local news versus international events, comedy versus drama, animated versus graphic representations. When multiple simultaneous sources are available, the set can take on some of the characteristics associated with print journalism, including "paging" through video-illustrated articles, printing them on demand, or clipping them for later review or to send to a colleague. Alternatively, the program can be captioned with references from print accumulated in advance or in the course of a telecast. Such text-driven television systems have been demonstrated (see Figs. 3 and 4). Composing a user-alterable visual program such as a movie is a more difficult challenge, but it permits controls that alter the point of view and perhaps even the outcome of a drama.

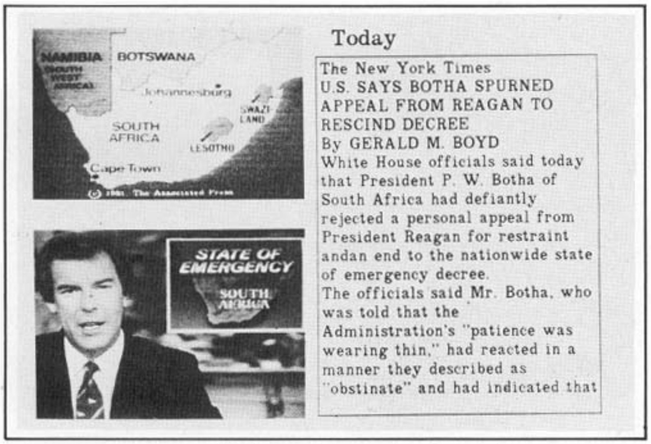
At first, this is a bilateral process, implying that the program distributor and the viewer cooperatively assemble the evening's entertainment or build the viewing experience. The television alone lacks access to the necessary side information that would allow the receiver to do the job by itself. The program can come packaged with such additional "content cues," or they can be part of the TV guide or daily newspaper. In fact, the state of the art of signal compression is such that 1.5 Mb/s images approaching broadcast quality are becoming standardized,[8] thus allowing four different television programs to be contained in a single channel or allowing four hours of programming to be transmitted in a one-hour time slot. This can be the basis for a program directed to the computer in the receiver rather than the screen, resulting in a program transmitted in one hour but watched in anywhere from 2 to 200 minutes.

We can only speculate on the popularity of such homebrew modes of operation; their success is as much a matter of imaginative program design as technological breakthrough. The key point is that the historical direct connection between the screen in the living room and the broadcasting antenna can and will be broken. Viewing time can be grossly disconnected from either broadcast time or program time, and the viewer can directly or indirectly participate in the synthesis of the event. Experiments with the synthesis of news programs with a "tell me more" button that provides related print supplements on request are promising, and audience research indicates receptivity to such ideas.[9] We suspect that this is the tip of the iceberg—the true innovations in this area are more in the domain of producers than engineers.

Image options

The television frame itself is another artifact that is no longer the boundary of image interaction. In computer-generated images, the frame is not the primitive unit—it exists only at the time a data base is actually viewed. But video has always been thought of as a succession of still pictures. The limits of our ability to process frames as picture waveforms is being approached in research labs, and the theme of worldwide activity is a higher level "model-based" image representation closer to a computer graphics set of parts than a still photograph.

The television frame itself is another artifact that is no longer the boundary of image interaction.



Today

The New York Times
U.S. SAYS BOTHA SPURNED APPEAL FROM REAGAN TO RESCIND DECREE
By GERALD M. BOYD
White House officials said today that President P. W. Botha of South Africa had defiantly rejected a personal appeal from President Reagan for restraint andan end to the nationwide state of emergency decree.
The officials said Mr. Botha, who was told that the Administration's "patience was wearing thin," had reacted in a manner they described as "obstinate" and had indicated that

FIGURE 4. Network Plus: This is the flip side of NewsPeek. During a live television broadcast—the nightly news—the closed caption transcript of the telecast is used to cue retrieval of additional information relating to the ongoing news film. This additional data can be printed or used to caption the broadcast, or the combination of the two can be used to create an illustrated edition printed on command at the end of the live program.

Nowhere is this more evident than in extremely low bandwidth video communications systems such as 64 Kb/s teleconferencing. At those rates, the image is such a reduced quality likeness that lip-synchronization is only barely possible. Innovative schemes that presume some *a priori* knowledge of the image subject involve creation of the remote image by manipulation of a three-dimensional wireframe structural model of a telephone caller that is overlaid with an image-like texture map of the face and moved in accordance with deduced parameters of motion.

KDD and GCT have demonstrated examples where the reproduced image is the Mona Lisa animating one side of a simulated telephone conversation. Two-dimensional versions of this basic idea were shown in 1981 where the motion parameters were derived at the receiving end by analysis of the voice signal alone, thus allowing re-creation of a moving picture with no transmission of video information at all. The picture was animated to replicate facial expressions that would be used to speak the words of the telephone call (see cover illustration).

The extension of these techniques to higher fidelity systems is being researched for bandwidth compression as well as pictorial enhancement. Approaches fall readily into two categories: those where the image multi-dimensional picture information is recorded at the time of image capture and those where it is derived from the two-dimensional image sequence.

Image capture systems include active and passive range sensing. Active range sensing is quite old and has been used commercially to make "solid photographs"—busts milled from a 3-D database obtained by a series of still photographs taken with structured illumination. Machine vision systems often use structured light and knowledge of camera position to identify the elements of the frame.

Passive range sensing is a more recent phenomenon and involves either stereoscopy or single lens, multiple focal length systems. The former is often used for terrain mapping and surveillance reconstruction. The latter has been more recently examined as a realtime video system where the depth information is derived from the differential depth of field obtained through a dual-aperture lenş that records a shallow depth of field image and a deep focus one placed adjacently on a single strip of film or video frame.

Other analysis tools are being discovered through greater understanding of the human visual process. Pentland has shown "shape from shading" programs where the angle of a surface is inferred from its reflectivity; and Adelson has worked with multi-frame correlations to remove noise from images and add resolution from time. Bove used depth from focus, combined with knowledge of the structure and motion of objects in the frame to create a three-dimensional image sequence from a movie and permit the viewer to display it from any point of view. While not a perfect reconstruction—there are holes and distortions left in the pictures—the path is clear; the camera is a data capture device, not solely a sampling system, and a database of the contents is becoming part of the video lexicon.

Where we go from here

We have presented the domains where artifacts in current video systems can be eliminated by modern design and innovations made possible by the incorporation of processing in the receiver and the originator of video signals. (These artifacts are not merely transmission degradations, but elements once thought inherent in the system.) We think of it in terms of entertainment video because that is the greatest challenge. The argument has been that technology will not only enable but necessitate such change. As stated, the incorporation of computing into consumer television receivers is seemingly not obvious or useful.

The workstation counterpart is usually termed "multimedia." It is equally challenging and equally open to conjecture and driven by imagination. We have had video available to workstations for 10 years through the agency of the optical videodisk, but only now, with the advent of digital compression systems and convenient small-sized optical storage media, is it ready to become prevalent. It is comforting to think that the workstation is less sensitive to price than the consumer television and will therefore gain new capabilities faster, but that is not necessarily the case.

Certainly we can count on HDTV to raise the stakes of home video and, similarly, we assert that the technology is more under control than the design of the applications. In the home, price may be the key issue, but in the workplace, functionality must be obvious. Casually watching the evening news while editing the corporate annual report may be fun, but that capability alone is probably not sufficient to build a television set into the desktop of anyone but a broadcaster.

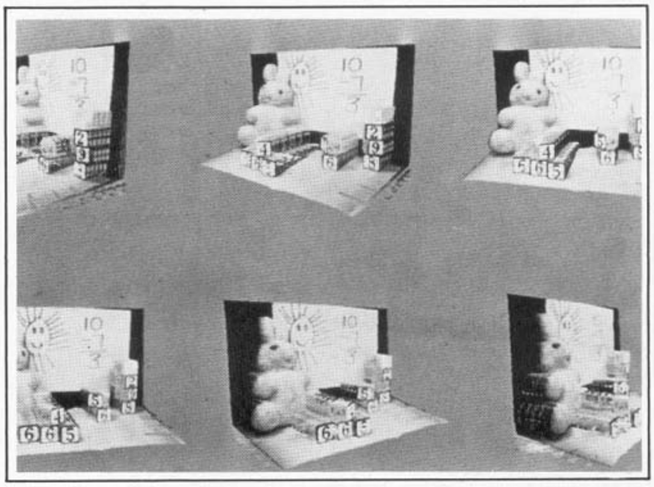


FIGURE 5. Six views from a range camera. Points from a laser range camera are projected in three-dimensional space and shown from six different points of view. When shown off the camera axis, there are parts missing—the challenge of research is to use knowledge of the scene to fill them in.

REFERENCES

1. T. Fujio et al., High definition television, Proc. of the IEEE, 73, April 1985.
2. Y. Ninomiya, MUSE coding system for HDTV broadcasting, Proc. of EURASIP Conference, Nov. 1986.
3. A.B. Lippman and W. Butera, Coding image sequences for interactive retrieval, Communications of the ACM, 32, 852–860, July 1989.
4. P.P. Vaidyanthan, Quatrature mirror filter banks, M-band extensions, and perfect reconstruction techniques, IEEE ASSP Magazine, 4–20, July 1987.
5. G. Karlsson, T. Barnwell, A procedure for designing exact reconstruction filters for tree structured subband coders, Transactions on Acoustic, Speech and Signal Processing, June 1986, pp. 434–441.
6. G. Karlsson and M. Vetterli, Subband coding of video signals for packet switched networks, Visual Comunications and Image Processing II, SPIE Proc., 845, 446–456, 1987.
7. W.F. Schreiber, Friendly family of television standards, NAB Conf. Proc., 1988.
8. Minutes of the Motion Picture Experts Group (MPEG), an ad hoc committee of the International Standards Organization, 1989.
9. W.R. Neuman, D. Gagnon, and S. Schneider, Is the mass audience ready for interactive television?, International Communication Association Conf. Proc., May 1989.