

Visual Communications of Tomorrow: Natural, Efficient and Flexible

Janusz Konrad, *Senior Member, IEEE*

Abstract—In the last decade, we have witnessed a phenomenal growth of communication and information technologies. These technologies have greatly simplified and even enriched our daily lives; cellular telephony and the Internet are probably the most striking examples. A particularly promising, and at the same time challenging, aspect of both technologies is the transmission and use of visual information. In this paper, I overview the state of visual communication at the end of 20th century, discuss today's challenges and outline some future directions.

I. VISUAL COMMUNICATIONS TO DATE

WHEN wireline telephony was born a century ago, few could have predicted its future profound impact on society. The birth of television three decades later also drew scepticism, but the critics were proven wrong again. Both technologies have flourished ever since, fueled by human desire to communicate with one another and stay in touch with the world. The introduction of the fax machine some 20 years ago did not disappoint, although its proliferation took a few years. Today it is a ubiquitous business tool and a common home electronic appliance. The recent rebirth of the telephone in its cellular incarnation was aimed originally at the business market. Today, cellular telephony, especially its digital variety, enjoys the fastest growth of any telecommunications technology developed in the 20th century and has the widest worldwide reach.

Where does visual communications stand today at the dawn of the 21-st century? Since its inception, television has gone through a remarkable transformation, due to the introduction of color in the 1950 and constant improvements in camera and display technologies. Today, analog television (NTSC, PAL, SECAM) delivers good-quality pictures and is ubiquitous thanks to the availability of three delivery channels: terrestrial, cable and satellite. In mid-1990s digital television (DTV) was introduced with all its benefits: superior image and sound quality, bandwidth efficiency, and embedded error resilience. Based on the MPEG-2 audio-visual compression standard and efficient channel coding, this new technology allows the transmission of four DTV programs within the bandwidth of one analog TV channel (6MHz). This permits cable and satellite TV operators to quadruple channel offerings without additional spectrum allocations, thus driving transmission costs down. MPEG-2 has also been the core technology for the DVD, CD-like medium with the capacity to store up to 2 h of MPEG-2 compressed video. The remarkable compression ratio achieved (on average 40-50) is possible thanks to the extremely efficient source coding of MPEG-2 and also its variable bit-rate capability; a 4-6Mb/s bitstream is enough on average for most programs with

the exception of extremely dynamic sport events (e.g., parts of hockey or basketball games). To date, both DTV and DVD have been huge successes, and it is very likely that they will flourish in the near future.

A particularly promising, and at the same time challenging, aspect of communication and information technologies is the transmission and use of visual information.

Ironically, the video compression technology used in DTV was originally developed for high-definition television (HDTV). Although DTV is fully digital, it maintains the screen aspect ratio (4:3) and resolution (720×480 in North America and 720×576 in Europe) of its analog predecessors. On the other hand, HDTV was designed to offer cinema-like screen shape (16:9) at much higher resolutions (e.g., 1920×1152). Despite the efficiency of MPEG-2 compression, the higher information content of high-definition images requires one complete analog TV channel for the transmission of one HDTV program; the transmission of an HDTV bitstream (18Mb/s) requires about four times the spectrum of DTV transmission. Introduced in the largest cities in the United States in November 1998, HDTV services are slow to take off due to the very high cost of retrofitting the studio and broadcast equipment, as well as replacing home TV sets. Once these obstacles are overcome, the HDTV is likely to take off, although formidable competition from DTV will continue.

The ubiquity of telephony and television today is not incidental. The telephone, especially wireless, is appealing thanks to the concept of anywhere/anytime person-to-person communication, while television is attractive due to the rich visual content that it conveys through moving color imagery. In the last 30 years, intensive work has been conducted on integrating the two concepts in a point-to-point visual communication system. Stationary videophone, a low-cost, low-resolution, consumer-oriented visual communication system, was introduced on the market as early as in the 1960s, and then reintroduced in the late 1980s and 1990s, but the proposed solutions failed to capture the consumer market. Although based on the recent H.263 digital video compression format, the newest solutions offer only QCIF (176×144) or sub-QCIF (128×96) images at temporal refresh rates often much lower than 30Hz due to channel capacity constraints (e.g., 56kb/s modems). Most recently, similar, but only experimental, solutions have been demonstrated for mobile videophony, but at even lower refresh rates due to wireless channel restrictions. Such low-resolution images, which can only be viewed on small screens, combined with poor motion rendition (motion jerkiness due to low refresh rates) have been the primary factor in reluctance to adopt this technology. Other important factors have been user friendliness (some sys-

J. Konrad is with Boston University, Department of Electrical and Computer Engineering, 8 Saint Mary's St., Boston, MA 02215 (jkonrad@bu.edu).

TABLE I

COMPARISON OF VARIOUS VISUAL COMMUNICATION SERVICES OFFERED TODAY OR UNDER DEVELOPMENT. NOTE THAT THE NEWEST VIDEO COMPRESSION STANDARD MPEG-4 IS NOT REFERRED TO, BUT IT APPLIES WHEREVER H.263 DOES.

Application	Channel	Channel errors	Video compression	Resolution	Video bit rate
Videophone	PSTN	Few bit errors and packet losses	H.263(+)	176×144	10-25kb/s
Videophone	ISDN	No errors	H.261/263(+)	176×144	64-384kb/s
Videophone	Packet network	No bit errors but packet losses	H.261/263(+)	176×144	10-384kb/s
Videophone	Wireless	High bit errors and packet losses	H.263(+)	176×144	10-300kb/s
Videoconf.	Packet network	No bit errors, packet losses	H.263(+)	352×288	0.1-1Mb/s
Videoconf.	ATM	Almost no errors	MPEG-2	720×480	1-6Mb/s
DTV	Cable/satellite	Almost no errors	MPEG-2	720×480	4-12Mb/s
HDTV	Terrestrial	Few bit errors	MPEG-2	1920×1152	18Mb/s

tems required complicated setup involving a TV set), human psychology (need to stay in the field of view of the camera) and pricing (both ends need to be equipped). Although technical issues can be overcome, especially as broadband connections (asynchronous digital subscriber line, ADSL and cable modems) become ubiquitous, it is unclear whether human factor issues can be easily addressed.

In contrast to videophone's inability to capture the consumer market, videoconferencing has been relatively successful in penetrating the corporate environment. Higher image quality than that of the videophone and relative independence from human factors are perhaps the two major stimuli in the proliferation of desktop videoconferencing today. Based on the same H.263 video compression as the videophone, videoconferencing exploits higher bit rates ($n \times 64\text{kb/s}$, typically 384kb/s) to deliver CIF images at 352×288 -pixel resolution and refresh rates close to 30Hz. At this resolution images can be viewed even at the full TV screen, although from a sufficiently large distance (typically more than 6 picture heights). The relatively high bit rate required usually poses no problem on corporate intranets; however, it is likely to cause delivery "hiccups" outside of them if no quality-of-service (QoS) guarantees exist. Human factors are less of an issue in corporate videoconferencing because of the way it is used (primarily for virtual business meetings), and also because of work ethics (acceptable behavior and dress code). Today, PC-based videoconferencing is becoming a standard business tool on corporate intranets, but its success outside of them depends primarily on the available channel capacity and/or QoS guarantees.

A summary of various visual communications services either offered today or under development is presented in Table I.

II. TODAY'S CHALLENGES

Although the transition from analog to digital visual communication is already taking place, a number of issues need to be

addressed before digital visual communication becomes ubiquitous. Below, I review the most critical, in my opinion, challenges that, if unresolved, may severely impede further evolution of digital visual communication.

A. Error resilience

In analog visual communication the video content is transmitted as a contiguous waveform over a dedicated channel. For example, in analog TV a fixed 6MHz channel carries only one suitably modulated audiovisual signal. Under mild degradation of channel conditions the received images typically become more noisy. In case of strong channel degradations, images can be severely distorted or even lost, but this loss is localized in time since analog transmission is memoryless.

In digital visual communication, the use of resources is quite different. Usually, the content is broken up into smaller units, such as blocks of pixels, then compressed and encoded, and finally shipped over a fixed- or variable-capacity channel. In a fixed-capacity channel, such as plain old telephony service (POTS) line with 56kb/s modem, an integrated services digital network (ISDN) line or 6MHz TV channel with 16-quadrature amplitude modulation (QAM), a compression scheme and bit rate must be carefully selected in order to match channel characteristics while maximizing video quality. However, even if channel capacity exceeds the required bit rate, channel errors can severely degrade the performance of the system. Since most digital video compression methods are not memoryless, the impact of channel errors is more severe than in the analog case; the effects are no more localized and may extend spatially and temporally. In order to "immunize" digital visual communications to channel errors, three classes of error resilience mechanisms are presently used:

1. Methods introduced within the source and channel encoders
2. Post-processing executed at the decoder to conceal errors

3. Interaction between encoder and decoder

In the first class of methods, intentional redundancy is added to the bitstream during source and/or channel encoding. The redundancy data are carefully designed (type and location) to achieve maximum gain in error resilience and at the same time introduce least performance penalty. Examples of such techniques are insertion of resynchronization markers, reversible variable-length coding (RVLC), insertion of intra-coded blocks or frames, independent segment prediction (ISD), data partitioning, layered coding with unequal error protection (UEP), and multiple-description coding (MDC).

While the first and third class of methods require synchronism between the encoder and decoder, the concealment methods are implemented in the decoder only and in no way impact the interaction between the encoder and decoder. In other words, such methods can be devised outside of video compression standards. A variety of techniques have been proposed to date for the recovery of texture information. Most of them perform either simple interpolation (spatial or spatiotemporal) or motion-compensated temporal prediction. More advanced methods use prior information about the underlying image structure by means of regularization or projection onto convex sets (POCS) methods. The recovery of missing motion vectors is often based on simple heuristics (repetition of a corresponding vector from a previous frame) or mean/median filtering. Designed outside video compression standards, error concealment methods will eventually appear in H.263 and MPEG-4 decoders as an added value. Since theoretically all standard-compliant decoders should perform equally, it is the error concealment effectiveness that will help differentiate among them in the marketplace.

Since theoretically all MPEG-4 and H.263-compliant decoders should perform equally, their error concealment effectiveness will help differentiate among them in the marketplace.

Methods in the third class require a feedback channel from the decoder to the encoder; the decoder informs the encoder which part of the bitstream is corrupted by errors, and the latter attempts to adjust its operation to suppress the effects of such errors. If the underlying network protocol supports automatic repeat request (ARQ), a simple approach is to retransmit the lost data. This solution, however, is often unacceptable due to the incurred delays. A simple alternative is to recover synchronization and temporarily reduce the transmission rate. If rate adjustment cannot be done, another possibility is to permit errors but limit their propagation. Reference picture selection (RPS) is a method in this class that, upon learning about errors in a previously-encoded frame, selects an earlier intact frame as reference for the prediction. This approach increases buffer size but is very effective in limiting error propagation. A more sophisticated version of this approach tracks how the damaged areas in a given frame would affect future frames, and subsequently performs intra-frame coding of new-frame blocks or avoids performing prediction based on damaged areas. For an excellent review of error resilience techniques the reader should see [1].

Today, the H.263 version 3 (H.263++) and MPEG-4 stan-

dards apply some of the above error resilience schemes. In particular, by using resynchronization markers, RVLC, ISD, data partitioning, and RPS, both standards perform well under a variety of channel conditions. However, due to the explosive interest in wireless and IP-based visual communication, both characterized by severe channel errors, work on ever more resilient video compression schemes is needed. This is further discussed later.

B. Network heterogeneity

A very common problem in the deployment of visual communications infrastructure is network and transmitter/receiver heterogeneity. While the ultimate goal is a homogeneous broadband network, today's communications infrastructure is a patchwork of networks with varying channel capacity, delay, and error characteristics. Moreover, the transmitter and receiver can have different characteristics such as video compression standard compliance, screen resolution, or available processing power (software decoding).

A natural solution to network and transmitter/receiver heterogeneity problems is transcoding, that is, decoding of a data stream with one set of specifications and subsequent encoding of this very data into a stream with another set of specifications. While in the case of different compression standards used in the transmitter and receiver this is the only option, alternative mechanisms are possible when data rates or resolutions are incompatible. For example, *layered coding* facilitates gradually increasing data rates and resolutions by means of hierarchical video representation. In this case, the video data are represented in a multiresolution pyramid (e.g., by means of wavelet decomposition). First, the lowest-resolution video data are encoded. Then, at each resolution the video data are encoded differentially with respect to the preceding lower-resolution locally decoded video. All the encoded data streams are combined into one scalable video representation [2]. By transmitting and decoding several resolution levels, a hierarchy of data rates and resolutions can be achieved, thus permitting the network to select suitable data rate for particular subnetwork, and at the same time allowing the receiver to decode only the information matching its characteristics.

As discussed in the previous section, layered coding has an additional advantage in the presence of channel errors. Since typically a compressed lower-resolution layer uses less channel capacity than a higher-resolution one, it is less likely to encounter channel errors. This can be reinforced by unequal error protection among layers.

C. Ownership protection

One of the most sensitive issues concerning today's digital technology is the ownership protection. While impossible in the analog domain, lossless tapping of digital content is very easy due to the nature of the signal and the easy access to the distribution media (packet sniffing on the Internet, ease of copying a CD). Lossless data duplication poses significant challenges for content providers, and has recently spawned a major research effort.

Content protection can be addressed by means of *cryptography*; original data are encrypted by the provider using a private key, while the users decrypt the content using an associated public key. The weak point of this approach is that the decrypted data are directly vulnerable to piracy once they have been brought back to the original unprotected form.

Another solution proposed is by means of a *digital signature*, i.e., an encoded message matching the original image and appended to it. Verification procedures are based on a public algorithm and public keys. Since any modification of the image data will cause a mismatch with the signature, image tampering can be easily detected. As the signature size is usually proportional to the original data size, this solution is not very practical.

Digital watermarking of images or video is a relatively new research area that is very fast growing¹. In digital image watermarking, an invisible signal (small alteration) is hidden in the image that can be subsequently detected by the provider or the customer, depending on the application. In private-key watermarking, used for copyright protection, the provider uses a private key to alter an image distributed to customers. The provider can examine any accessible image for watermark existence using the private key. In public-key watermarking, used for content verification, the procedure is similar except that watermark casting associates a public key with the private key. The customer can use the public key and a public watermark detection algorithm to verify watermark existence. Private-key watermarking aims at the protect the provider and has been extensively studied in the literature to date. Public-key watermarking protects the customer. Despite significant research efforts and a multitude of proposed new methods, the problem remains largely unsolved; no single method proposed to date can survive a range of attacks.

III. INTO THE FUTURE

As today's challenges described above are likely to be addressed in the near future, and as the network infrastructure is upgraded to video-level bandwidths, one wonders: what's next? Clearly, videoconferencing and videophony would greatly benefit from better image quality, although at the cost of putting additional strain on network resources. However, before superior image quality of today's HDTV is used in person-to-person visual communications, other improvements are likely to take precedence. Below, some new challenges likely to shape visual communication in the 21st century are discussed.

A. Natural visual communication

Today two-dimensional (2-D) color moving imagery is a convincing visual medium in a variety of applications. However, there exist many applications, in both person-to-person and broadcast communication, that would greatly benefit from an increased degree of realism. For example, virtual visits (e.g., of homes, vacation sites), remote guidance in dangerous environments and telemedicine would be more effective were users able to move (roam) in virtual 3D space; the perception of

depth, so natural in daily life, would greatly enhance a "being there" experience.

Tomorrow's visual communications must address this issue by means of capture, transmission, and display of 3D visual cues [3]. Although holographic and volumetric 3D displays may eventually provide the ultimate 3D experience, it is unclear how to transmit the vast amounts of optical information needed by these displays. Moreover, since state-of-the-art holographic displays can present only still images and volumetric displays are bulky, both are incompatible with the anytime/anywhere concept of the 21st century visual communications.

Today two-dimensional (2-D) color moving imagery is a convincing visual medium in a variety of applications. However, there exist many applications, both in person-to-person and in broadcast-type communication, that would greatly benefit from an increased degree of realism.

A very effective alternative to holographic and volumetric systems are 3D *stereoscopic* and *multiview* (multiscopic) displays. Unlike holographic and volumetric displays, that render a 3D volume, stereoscopic displays present two views of the same scene captured from slightly different angles, and rely on the human brain to fuse those views into a meaningful 3D representation. Stereoscopy has been around for over a century, but only in the last two decades it emerged as a viable 3D technology. This has been made possible by the advances in image multiplexing/demultiplexing techniques needed to separate views intended for each eye. Three approaches are currently dominating stereoscopic imaging (Figs. 1-2):

1. *Polarization*: Views are projected through light polarizers onto a screen (superposed) and then separated by identical polarizers incorporated into eyewear (linear orthogonal or circular polarizers).
2. *Time-sequential shuttering*: Views are time-sequentially multiplexed on the screen and separated by liquid-crystal shutter (LCS) glasses that open and close in sync with the displayed images.
3. *Autostereoscopy*: Views are usually spatially multiplexed on a pixel-addressable screen and subsequently separated by an optical layer that directs left and right pixels to separate viewing zones (e.g., lenticular); no glasses are required.

Undoubtedly, autostereoscopy is the holy grail of electronic 3D displays, since no eyewear is required. However, today's approaches usually apply spatial view multiplexing, thus reducing either horizontal or vertical resolution by a factor of two. This is a very active research area with great potential. Polarization- and LCS-based systems seem to be temporary solutions only; although acceptable in broadcast applications, the eyewear required is a major obstacle in person-to-person communications.

Stereoscopic displays provide only a single 3D perspective (i.e., as if watching a scene from one single viewpoint). In order to change this viewpoint and allow a dynamic 3D perspective, multiview displays have been recently proposed. The essence of multiview displays lies in their ability to present geometrically-correct view at each viewing angle. This poses two new challenges.

¹ At the IEEE International Conference on Image Processing, ICIP-2000 the most popular topic (in terms of the number of papers presented) was digital image and video watermarking - 4 sessions.

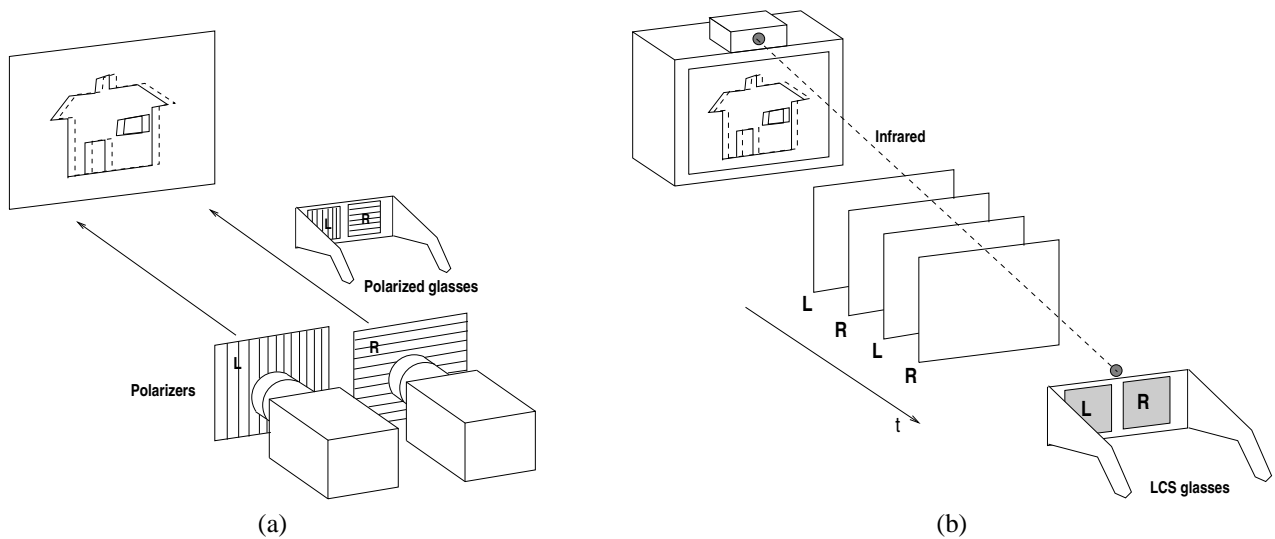


Fig. 1. Stereoscopic displays using glasses: (a) polarization-based display; and (b) liquid crystal shutter display.

The first challenge is the display itself. One approach is to detect viewer head position by means of head tracking and present proper view on a stereoscopic display (eyewear-based or autostereoscopic); as the viewpoint changes so do the presented images. Experimental systems developed to date have shown great promise; however, as was already mentioned, eyewear may not be acceptable in some applications, while autostereoscopic systems suffer from “view flipping” (in a two-zone system the eyes moving to another viewing zone are presented with the same perspective). An alternative approach, that does not require tracking, is a multiple-zone autostereoscopic display. By allowing many viewing zones, each delivering a single perspective, one can project different views into different zones, and therefore different eyes. This is a very attractive concept, except that such displays suffer from low horizontal resolution since viewing zones are multiplexed horizontally. Although both tracking- and zone-based solutions proposed to date are far from satisfactory, research into multiscopic displays is very active and promises interesting solutions in the next decade. One particularly interesting direction is the development of multiview displays with temporal instead of spatial multiplexing.

The second challenge is the acquisition and delivery of mul-

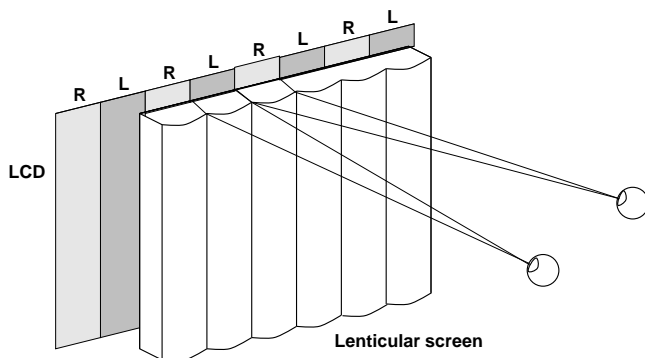


Fig. 2. An example of autostereoscopic 3D system - two-zone lenticular display. At correct distance from the screen, eyes can see independently left- and right-column pixels.

tiview data. In an ideal scenario, the viewer head motion would control the position of a camera, but this is only possible in very particular real-time applications (e.g., tele-manipulation). More typically, the 3D data are acquired with no such feedback; a fixed number of views is available for transmission. To date, experimental multi-camera 3D acquisition systems have been built and successfully tested. In order to assure view continuity, the number of cameras should be as large as possible. However, physical constraints on camera size impose a limit on camera separation. Moreover, the resulting data flow from many cameras would put significant strain on network resources. Certainly, the views are highly correlated and therefore compressible, but the MVP (multiview profile) mode of MPEG-2 [4] supporting multiview coding may not be sufficient when the number of views is large. Further work in this direction is needed [5], [6]. An alternative is to transmit fewer views and to compute intermediate (virtual) views at the receiver [7]. This solution puts less strain on the network but requires significant computing power at the receiver. Also, the quality of the computed views must be high enough in order that the experience be transparent. This is a formidable challenge especially that such computations must be performed in real time.

Although the stereoscopic and multiview displays, especially those based on autostereoscopic mechanisms, can be made portable, they are not exactly compatible with the anytime/anywhere concept of future visual communication. Within this concept the, delivery of visual data is likely to be solved by wireless transmission at broadband rates; perhaps not as soon as many expect, this problem nevertheless will be largely solved in the near future. However, it is unclear how to solve the 3D data presentation problem for a viewer on the move. Should this be a small digital communications appliance with an autostereoscopic screen, or perhaps an eyewear-based 3D display similar to those proposed for wearable computers?

B. Efficient visual communication

The enhanced realism of future visual communication will undoubtedly result in an increase in transmission requirements

whether the improvements come from stereoscopic/multiview video, from high-resolution video, or from both. Although some researchers in the networking community argue that video compression will become obsolete since sufficient bandwidth will become available to send uncompressed video, one can easily point out that as viewers' appetites for ultimate-quality images grows, high-resolution multiview point-to-point visual communication may be the next "killer" application capable of congesting the networks. Therefore, I believe that bit rate efficiency will remain one of significant challenges in the future. Although significant theoretical and practical advances have been made in monoscopic video compression to date, methods specific to stereoscopic and multiview transmission are still immature.

High-resolution multiview (3D) point-to-point visual communication may be the next "killer" application capable of congesting the networks.

A rudimentary multiview video compression (encompassing stereoscopic compression) mode has been adopted in the MPEG-2 standard as the MVP addendum. In this mode, one view is treated as a base layer while the other views are considered to be enhancement layers; the cross-view correlation is exploited by means of disparity compensation, and the bit rate is allocated unequally between the base and enhancement layers. Since the MVP prediction is bi-directional², various combinations of disparity- and motion-based predictions can be applied in forward and backward directions [4]. This flexibility jointly with unequal rate allocation constitute a promising 3D compression tool that could be used in the near future for efficient stereoscopic video delivery to the home should MVP-compatible MPEG-2 hardware become available. However, in order to maximally exploit the MVP transmission, human factor issues need to be studied more; human binocular vision possesses some very interesting properties that could be further exploited to reduce transmission costs. Despite some recent advances [5], [3], the understanding of human binocular vision remains largely untapped. One particularly critical aspect that requires attention is eye fatigue when viewing 3D images and video.

Although MPEG-2 is a very efficient video compression engine, even better compression methods can be developed for multiview data. Two factors should play a role in this. First, significant new results in video coding are published almost yearly. Second, the reliance on video compression standards in advancing state-of-the-art may be diminishing as software based compression is likely to dominate some future applications (see Section III-C).

Another compression-related issue that will remain critical for visual communication systems is error resilience. On one hand, the improving network infrastructure, with the continuously increasing capacity and lower packet loss rates, will result in higher picture quality. Consequently, in non-broadcast applications over wired or terrestrial channels we might see in the near future sufficiently short round-trip delays that allow the use of re-transmission protocols such as TCP. Error-resilient video compression may become unnecessary in such

applications. On the other hand, any form of video broadcasting (e.g., DTV, HDTV) as well as satellite transmission will need to rely on error-resilient coding since the delays are significant enough to make re-transmission impractical. Ultimately, should the bandwidth be abundant enough, simple transport-based techniques, such as forward error correction (FEC) or packet duplication, might be sufficient to yield acceptable quality levels without error-resilience mechanisms. An alternative to this *best-effort transmission* is a transmission with *quality of service* guarantees. In this case, the burden of dealing with channel errors is delegated to the network itself. If network's QoS guarantees are satisfactory for the intended video bit rate, no error resilience need be embedded into video coding. Otherwise, layered coding should be used and the most error-sensitive bits (e.g., average block intensity or motion vectors in MPEG-type coding) transmitted using QoS guarantees. A very active research area for years, networks with QoS guarantees show great promise for future visual communications.

C. Flexible visual communication

Today, passive watching of a TV program or simple video playout from the internet are no more sufficient for many viewers; they demand more flexibility and interactivity from the visual medium. One way to address these needs has been adopted in the recent MPEG-4 audiovisual compression standard [8], [9], namely object-based compression/transmission of sound and video. The basic idea is to provide the end user with a flexibility in composing the final video program; audiovisual objects are transmitted separately and can be used or replaced by other objects in the final image composition at the receiver. This new functionality is the first step towards a flexible system with which the user can interact.

However, the object-based functionality is not supported by MPEG-4 with respect to 3D stereoscopic and multiview images. Although a very exciting perspective, object manipulation in 3D viewer space is a challenging problem and requires further research on object-based video coding and description [10]. An important body of work in this area is presently being carried out within the new MPEG-7 audiovisual description standard³ [2], however it is unclear to what extent MPEG-7 will support 3D video descriptions. As pointed out earlier, frame-based multiview coding is supported by MPEG-2, but its performance for more than two views has not been thoroughly tested. Object-based multiview video coding is an even more complex topic and has been little explored to date [11]. Clearly, significant challenges and great opportunities lie ahead in this area.

It is very important to realize that although an MPEG-4 encoder can support object-based functionalities, itself it does not perform the video breakup into objects and background. Moreover, the standard does not describe how to do it. The so-called *alpha planes* or spatial segmentation maps identifying image pixel memberships in objects must be performed prior to the

²MVP is just a special case of the MPEG-2 temporal scalability mode.

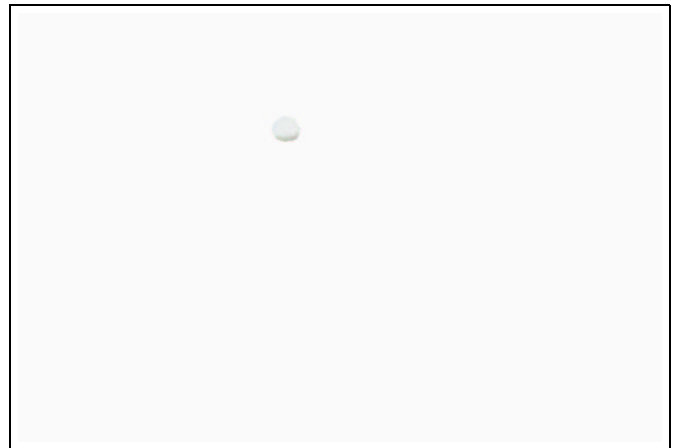
³MPEG-7 addresses many issues related to the organization of audiovisual databases and mechanisms of content extraction, very important topics for future visual assets management that are nevertheless outside of the scope of this article.

MPEG-4 encoding, and are at the discretion of the system designer. Clearly, the ability to compute alpha planes is a necessary condition for the success of the object-based modes of MPEG-4.

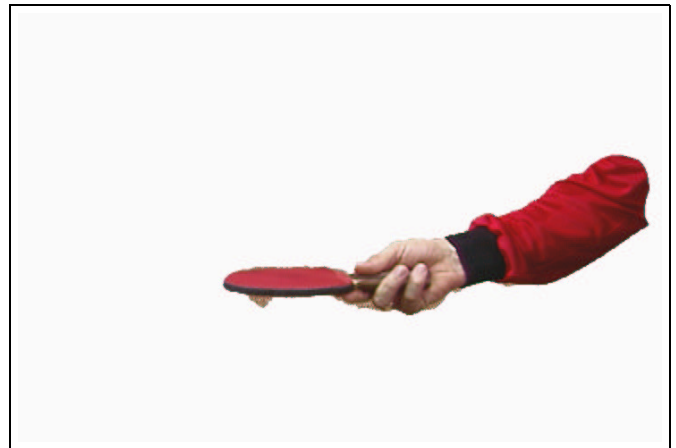
Spatial video segmentation has recently attracted a lot of interest in the research community and numerous techniques have been proposed, from simple pixel/block clustering and region growing methods through regularization/Markov/minimum-description-length models to active contours. The segmentation problem, ill-posed and even ill-defined⁴, is notoriously difficult to solve and will challenge researchers for years to come. Some recent results based on region competition and active contours [12] show an interesting research direction. This direction is particularly interesting when coupled with level sets as a solution mechanism [13]. An example of motion segmentation using region competition and level sets [14] is shown in Fig. 3. Note that the method is fully automatic and, although the result is not perfect, three distinct motions have been clearly identified (ball, arm, background); pixels not belonging to any layer do not have correspondence (occlusion areas). Despite this very encouraging result further research in this direction is need.

An important issue facing equipment manufacturers as well as consumers today is the proliferation of video compression standards; H.261, H.263, H.263+, H.263++, MPEG-1, MPEG-2, MPEG-4 have been proposed in the last decade. It is likely that new compression and delivery mechanisms will soon be developed to support new services. Video standards are developed today since specialized hardware is needed to handle the complex processing; the standards permit hardware components from different manufacturers to talk to one another. However, due to remarkable advances in the processing power of desktop computers, software decoding (and often encoding) within some standards (e.g., H.261, MPEG-1) can be executed today in real time on a modern Pentium III processor. This suggests that strict video compression standards may not be needed in certain applications, as the decoding algorithm can be transmitted in the bitstream header. In a sense, this is already happening today as users install RealNetworks, QuickTime and other players to decode the incoming bitstreams. Also, the large set of toolboxes supported by MPEG-4 suggests a slow move towards *programmable coding* rather than a single fixed algorithm.

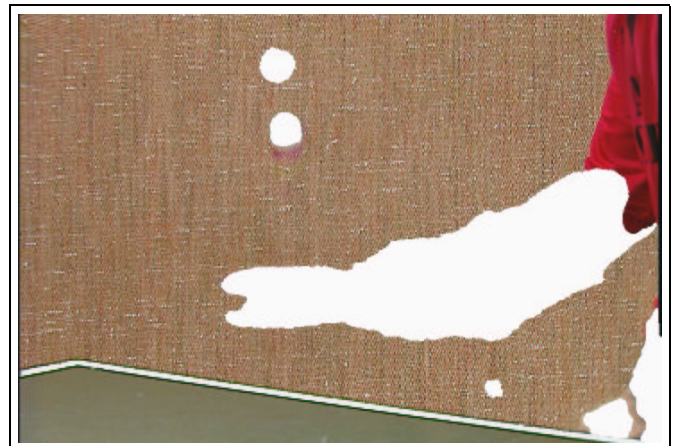
It is quite conceivable that standards-based video compression may lose its appeal in the future. Standards-based compression locks the coding performance for years until a new standard proposal is researched, adopted and suitable hardware is manufactured. As more efficient compression techniques are developed almost yearly, this cycle is too slow to keep up with the research. Programmable compression would address this issue by embedding the decoding instructions into the video bitstream; the decoder would first download and install the decoding instructions, and then use them to interpret the subsequent bitstream. For example, one could imagine that during the initial negotiation between the encoder and decoder, very much like in fax transmission, suitable compression tools and bit rate are selected given the CPU and display capabilities (resolution)



(a) Ball



(b) Hand



(c) Background

Fig. 3. An example of fully-automatic motion-based segmentation of a natural video sequence into three layers: (a) pingpong ball; (b) player's arm; and (c) background. Note that pixels not belonging to any of the three layers are labeled as occluded.

⁴What is a meaningful (for humans) segmentation of an image or video?

of the decoder.

Standards-based video compression may lose its appeal in the future since it locks the coding performance for years until a new standard proposal is researched, adopted and suitable hardware is manufactured. Programmable compression would address this issue by embedding the decoding instructions into the video bitstream.

A compromise between programmable and standards-based compression could be *toolbox-based* compression (somewhat similar to MPEG-4) where a wide variety of compression tools would be available at both the encoder and decoder, and only a high-level compression algorithm using those tools would need to be transmitted in the header. The toolboxes could be based on hardware-accelerated instructions, such as the MMX instruction set, although this solution would be perhaps closer to standards-based video compression than to programmable compression.

IV. CONCLUDING REMARKS

The future of communicating visual cues is both exciting and challenging. Remarkable progress has been achieved since the birth of television, and most of it in the last two decades. However, before a life-like, reliable and ubiquitous visual communications becomes part of our lives, a concentrated research effort in 3D displays, 3D video compression and processing, and human factors is needed. Although some researchers in the networking community argue that sufficient bandwidth will become available to send uncompressed video, it is very likely that high-resolution multiview point-to-point visual communication will be the next “killer” application capable of congesting even the fastest networks.

ACKNOWLEDGEMENT

The author would like to thank Mr. Abdol-Reza Mansouri for the preparation of the experimental results in Fig. 3.

REFERENCES

- [1] Y. Wang, S. Wenger, J. Wen, and A. Katsaggelos, “Error resilient video coding techniques: Real-time video communications over unreliable networks,” *IEEE Signal Process. Mag.*, vol. 17, pp. 61–82, July 2000.
- [2] A. Bovik, ed., *Handbook of Image and Video Processing*. Academic Press, 2000.
- [3] Special session “Digital Stereoscopic and 3D Imaging: from Research Laboratory to Mainstream Applications”, *IEEE International Conference on Image Processing, ICIP-2000*, Vancouver, Canada, Sept. 2000.
- [4] A. Puri, R. Kollaritis, and B. Haskell, “Basics of stereoscopic video, new compression results with MPEG-2 and a proposal for MPEG-4,” *Signal Process., Image Commun.*, vol. 10, pp. 201–234, 1997.
- [5] Special session “Digital Stereoscopic and 3D Video: Communication and Entertainment for the Future”, *Stereoscopic Displays and Applications, Electronic Imaging '99*, San Jose, CA, Jan. 1999.
- [6] Special issue on 3-D video technology, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 2-4, Mar.-Jun. 2000.
- [7] J. Konrad, “View reconstruction for 3-D video entertainment: issues, algorithms and applications,” in *Proc. Int. Conf. on Image Process. and its Applications*, pp. 8–12, July 1999.
- [8] Special issue on MPEG-4, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, Feb. 1997.
- [9] R. Koenen, “MPEG-4: Multimedia for the future,” *IEEE Spectrum*, vol. 36, pp. 26–33, Feb. 1999.
- [10] Special issue on object-based video coding and description, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, Dec. 1999.

- [11] M. Strintzis and S. Malassiotis, “Object-based coding of stereoscopic and 3D image sequences,” *IEEE Signal Process. Mag.*, vol. 16, no. 3, pp. 14–28, 1999.
- [12] S. Zhu and A. Yuille, “Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, pp. 884–900, Sept. 1996.
- [13] J. Sethian, *Level Set Methods*. Cambridge University Press, 1996.
- [14] A.-R. Mansouri and J. Konrad, “Multiple motion segmentation with level sets,” *IEEE Trans. Image Process.*, (submitted - April 2000).