A. Murat Tekalp, Engin Kurutepe, and M. Reha Civanlar

3DTV over IP

End-to-end streaming of multiview video

The Internet Protocol (IP) architecture is very flexible in accommodating a wide scope of communication applications ranging from email to video over IP services. Transmission of video over IP is currently an active research and development area where significant results have already been achieved. There are already many video-on-demand services offered over the Internet [1]. Also, 2.5G and 3G mobile network operators began to use IP successfully to offer wireless video services. The next big step forward is destined to be flexible distribution of a variety of three-dimensional (3-D) video and 3-D TV services over IP networks.

While there are many alternative technologies for 3-D video representation, including holographic, volumetric, geometric (3-D mesh models), and multiview (lightfield), stereoscopic/multiview 3-D video seems to be the most mature technology at the moment. Stereoscopic video consists of two video sequences (left and right views) captured by closely located cameras (approximately the distance between two eyes). If only two views are provided, the user has no other choice but to watch the 3-D scene from the fixed viewpoint of these two cameras. In the real world, if a person moves or turns his/her head around, he/she expects to see the 3-D scene from a somewhat different viewpoint. To provide this kind of interactivity, called free-view TV, more views need to be sent, which of course increases the bandwidth requirement significantly. Different 3-D displays may have a

Digital Object Identifier 10.1109/MSP.2007.905878

different number of view requirements, which must be considered in designing a streaming solution.

3-D stereoscopic displays, which can display two or more views and may provide free-view interactivity to a single user at a time or multiple users simultaneously, include the following:

- **Polarized 3-D Projection Display:** A polarized 3-D projection display consists of a pair of projectors and a pair of polarization filters. Light from one projector is polarized in the clockwise direction and the other in the counterclockwise direction using circular polarization filters. Both projectors are precisely aligned to project onto a silver screen covered with a neutral grey reflective dielectric material to preserve the polarization of light after reflection (see Figure 1). The users wear inexpensive glasses, which have filters matching with the projectors to ensure that light from each projector is only seen by one eye. A single PC with two display outputs can drive the projectors using a virtual desktop of, e.g., $2{,}048 \times 768$ pixels where each projector displays only one-half of the extended desktop at $1{,}024 \times 768$ native resolution. Thus, left and right videos can be shown such that they exactly overlap with each other on the silver screen.
- **Time-Multiplexed Projection Display:** A time-multiplexed projection display uses a projector that can display images at twice the frame rate of a regular projector. The right and left views are then displayed successively. The user needs to use special glasses that cover one eye at a time, along with the display.
- **Autostereoscopic 3-D Laptop:** Autostereoscopic displays do not require the user to wear special glasses. Instead, they reflect each view to the respective eye using only a special lens array integrated with the display, such as an LCD display. Autostereoscopic 3-D laptops based on parallax barrier stripes are available in the market today [2].
- **Autostereoscopic Multiview Lenticular Displays:** Autostereoscopic multiview lenticular displays provide free-view interactivity to multiple users by displaying more than two views (e.g., nine views) simultaneously [3]. They use a special lens array, which can reflect different pairs of two views to the respective eyes of users within a limited angular field of view.

[FIG1] **Polarized projection stereoscopic 3-D display system.**

- **Autostereoscopic Head-Tracking Displays:** Autostereoscopic head-tracking displays provide free-view interactivity by tracking the head position of the user in real time by a head cam attached on the display [4].

An important problem with 3-D video distribution over the Internet is the large size of data to be delivered or the "bandwidth requirement" (a more suitable term is "throughput," but the use of bandwidth with an overloaded meaning is more common). State of the art in multiview video coding (MVC) is described in [5], where significant compression gains are reported over simulcast coding which compresses each view independently. However, even with the state-of-the-art compression, bit rates for multiview video (MVV) are still high: 38 dB peak signal noise ratio (PSNR) at about 5 Mb/s is a common operating point for 704×480, 30 f/s, 8-view video using MVC. Transmission of large data without appropriate congestion control not only reduces the throughput but also increases delay for other applications sharing the same links. For 3-D video broadcast, the problem becomes worse, compared to on-demand 3-D video services, because all views may have to be sent as the number of receivers increases. A well-established solution to achieve service scalability is network-level or native multicast, where the network elements, such as routers, replicate packets to be delivered to several users as needed. Although multicast has been a part of the Internet architecture since the early days, it is still not widely deployed for several reasons, including management and security problems. Without native multicast, sending two-dimensional (2-D) or 3-D video to many clients one-by-one requires a very-high-bandwidth Internet connection for the server, increasing the service cost prohibitively. As an alternative, content distribution networks (CDNs), which deploy dedicated servers at several locations on the Internet, have been proposed. The video to be distributed can be stored in each of these servers and each customer can be served from the most appropriate server. Recently, an alternative distribution scheme came into existence: peer-to-peer (P2P) distribution, where each receiver is asked to pass the packets received to other nearby (in a sense, a small number of in-between routers) clients. An analysis of the significant cost savings associated with the P2P approach over the current Internet has been presented in [1]. Several companies have already started exploiting this for 2-D video delivery. For example, Joost is planning to become a global TV network based on P2P technology over the Internet.

An end-to-end 3DTV system consists of 3-D video representation and compression, transport protocols and systems, and 3-D display client/peer, as shown in Figure 2. Different 3-D displays, discussed above, require different 3-D video representations or number of views. For example, fixed-view stereoscopic displays require only two views, while autostereoscopic displays may require eight or more views at a time to provide limited free-view functionality. A brief overview of different representations for 3-D video and their compression [5] is presented in the next section. We then present an overview of generic MVV streaming architectures and protocols. The classic unicast streaming model for 3DTV is presented and a

selective streaming architecture for single-user head tracking displays is introduced. We provide an overview of cooperative streaming architectures and protocols, including application layer multicast and P2P streaming. Strategies to deal with packet losses are summarized.

OVERVIEW OF 3-D VIDEO REPRESENTATIONS AND CODING

GEOMETRIC MODEL-BASED SCENE REPRESENTATION AND CODING

The scene geometry is often represented by a 3-D mesh model, whereas the intensity/color information is represented by a static or dynamic texture map. Compression of 3-D meshes and texture maps have been addressed in MPEG-4 under 3-D graphic compression [6]. While this representation is quite efficient for synthetic (graphics or cartoon) video, the extraction of 3-D mesh models and texture maps from real MVV is a difficult analysis problem, which limits the usefulness of this representation for real 3-D videos.

VIDEO-PLUS-DEPTH REPRESENTATION AND CODING

This approach supplements a regular video stream with a depth map providing a Z-value for each pixel, as shown in Figure 3. The depth map, which can be computed from MVV or can be acquired with special cameras, can be compressed using H.264/AVC at about 10% overhead in the bit rate. The desired multiple views are then rendered at the receiver side by using depth-image-based rendering (DIBR) [7]. Recently, a new MPEG standard that covers this approach has been published in two parts: The specification of the depth format is called ISO/IEC 23002-3 (MPEG-C) and a method for transmitting video-plus-depth within a conventional MPEG-2 transport stream is published as an amendment (Amd. 2) to ISO/IEC 13818-1 (MPEG-2 Systems).

STEREOSCOPIC/MULTIVIEW VIDEO REPRESENTATION AND CODING

There are two factors that can be exploited for efficient encoding of MVV: 1) Interview redundancy refers to correlations between the views, and 2) psycho-visual redundancy refers to the suppression theory of human visual perception of 3-D from stereoscopic video that allow subsampling of one of the views [8]. There are many research and standardization activities for MVV compression based on exploiting inter-view redundancy. Early work resulted in the MPEG-2 multiview profile [9]. Recently, new MVC methods [10], [11] based on extensions of H.264/AVC [12], [13] were introduced, and new standards are currently being developed by the Joint Video Team (JVT).

JVT MVC

JVT is developing an extension of the H.264/AVC video coding standard [12] which supports new prediction structures for MVC [10]. A reference encoder-decoder, called Joint Multiview Video Model (JMVM) [11], is publicly available, which employs hierarchical B-pictures within each view, as well as a hierarchy between views for inter-view prediction.

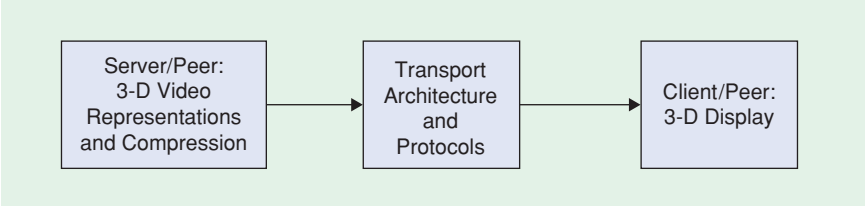
SCALABLE MVC

Scalable video coding is desirable for efficient video transport over the Internet. Hence, new scalable extensions of MVC have recently been proposed [14], [15].

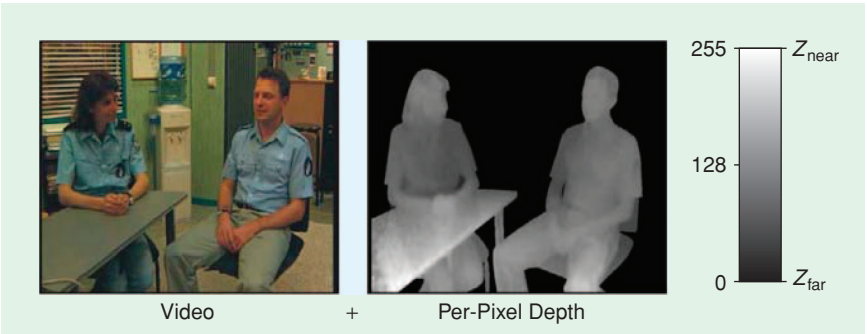
Regarding the psycho-visual redundancy, it is a common practice in monocular video compression to subsample chrominance channels, since the human vision system (HVS) is less sensitive to variations in chrominance values. Similarly, it is possible to exploit the suppression theory of human stereo perception [8], which states that humans can perceive high frequency in 3-D from one of the views even if the other view is low-pass filtered in order to maximize the overall perceived quality at a given rate or minimize bit rate at a given perceived quality. The subsampling may be in spatial resolution, temporal resolution, or quality (SNR) or a combination of these [16], [17]. Of course, spatially subsampled views will be interpolated to full resolution at the client before display. It has been shown that stereoscopic video can be encoded at about 1.2 times the bit rate of monoscopic video by unequal bit allocation between the right and left views (also called asymmetric coding) without noticeable loss of perceptual 3-D video quality [18]–[20].

REAL-TIME ENCODING AND DECODING

While real-time decoding is required for playback of streamed MVV, real-time encoding enables streaming of live MVV. There are research efforts on fast encoding and decoding implementations



[FIG2] Block diagram of an end-to-end 3DTV system.



[FIG3] 3DTV data representation using video-plus-depth [7].

in software, as well as field-programmable gate array (FPGA) implementations.

OVERVIEW OF MULTIVIEW VIDEO STREAMING ARCHITECTURES AND PROTOCOLS

The dominating transport protocol of the current Internet, transmission control protocol (TCP), has been conceived and designed for reliable delivery of non-real-time data over wired links, where the main reason for packet losses is congestion. Although user datagram protocol (UDP) is included in the original Internet protocol suite partially to allow implementations of real-time transport applications, transport of real-time multimedia can only be accomplished on a best-effort basis. The real-time transport protocol (RTP) [21] providing for several needs of real-time media transport applications does not address the quality of service (QoS) issues that need to be handled using other mechanisms such as differentiated services (DiffServ). DiffServ specifies a simple mechanism for classifying network traffic such that packets carrying critical data can be delivered with low latency and guaranteed service, while other packets can be delivered with best effort. Because of backwards compatibility and the lack of the associated value infrastructures, it is not widely deployed in today's Internet.

WITH THE INTRODUCTION OF THE WIRELESS LINKS, ANOTHER SOURCE OF PACKET LOSS HAS EMERGED DUE TO BIT ERRORS AT THE PHYSICAL LAYER BECAUSE OF FADING, INTERFERENCE, ETC.

With the introduction of the wireless links, another source of packet loss has emerged due to bit errors at the physical layer because of fading, interference, etc. since link-layer retransmission without appropriate application-layer time limit is not a suitable strategy for real-time applications. The Internet is not equipped with tools to differentiate between packet losses due to congestion and due to bit errors at the physical layer. Mechanisms such as explicit congestion notification (ECN), Wireless-TCP, and UDP-Lite may address some needs for media delivery over the wireless Internet, but deployment of these will not be immediate. Therefore, video applications expected to work in the near future should be designed to tolerate packet losses.

Another vital issue with the real-time, high-bandwidth media delivery over the Internet is congestion prevention and control. Today, the most widely used transport protocol for multimedia is RTP over UDP [21], [22], which does not contain any congestion control mechanism and, therefore, can lead to congestion collapse when large volumes of MVV are delivered. Considering the very wide-scale deployment of TCP, a congestion control technique that is "fair" to TCP applications should be employed. The recently introduced datagram congestion control protocol (DCCP) [23], running directly over IP, has built-in bandwidth usage limitation mechanisms for TCP-friendly multimedia delivery [24]. DCCP can be thought as UDP plus congestion control, connection setup, and acknowledgments, and unlike TCP it can accommodate different congestion control mechanisms. Despite the unreliable datagram flow, DCCP provides reliable handshakes for connection setup/teardown and reliable negotiation of options. Besides handshakes and feature negotiation, DCCP also accommodates a choice of modular congestion control mechanisms. There currently exist two congestion control schemes defined in DCCP, one of which is to be selected at connection startup time: 1) TCP-like congestion control (IETF RFC 4341) and 2) TCP-friendly rate control (TFRC) (IETF RFC 4342).

The former is intended for use by senders who would like to adapt to abrupt changes of the congestion window, as in regular TCP, to take full advantage of the available bandwidth in rapidly changing network conditions. The target of this approach is to send as much data as possible in a given time interval, which does not match well with real-time media streaming applications where the data rate is determined at the encoder and usually can not go above a certain value even if there is more network bandwidth available. The latter, TFRC, is an equation-based flow control mechanism that minimizes abrupt changes in the sending rate while maintaining longer-term fairness with TCP. It is, hence, more appropriate for applications that would prefer a rather smooth sending rate, including real-time streaming media applications with a small or moderate receiver buffer. In this scheme, an allowed sending rate, called the TFRC rate, is calculated using the TCP throughput equation, which is provided to the sender application upon request. The sender may use this rate information to adjust its transmission rate to get better results. Hence, the unicast server must use effective video rate adaptation methods, which will be discussed in the "Unicast Streaming of Stereo and Multiview Video" section.

There is also an experimental RFC for TCP-friendly multicast congestion control (TFMC) [25]. The multicasting paradigm aims to avoid sending duplicate packets to clients in the network in order to utilize network resources more efficiently. In network-layer multicast, packets get duplicated at multicast-enabled routers as needed and forwarded to other members of the multicast group. Although most new routers are now multicast capable, security and other concerns discourage network operators from enabling the multicast functionality; hence, network-layer multicasting is not widely deployed. Therefore, several alternative application-layer-based methods have been proposed that construct overlay networks and shift multicast functionality to hosts, which accomplishes packet duplication, forwarding, and management of distribution trees. Since peers in such overlay networks act both as receivers and senders, bandwidth load is distributed across the network instead of being concentrated at a central server. To compute the TFRC rate in a multicast scenario, each receiver computes its own TFRC rate as a function of its own measured RTT and loss rate and sends this to the server. The server then selects the minimum of these rates. However, only a limited number of selected

clients are allowed to send their TFRC rates to the server in order to prevent feedback explosion. In the case of DCCP, again each client measures its RTT and loss rates and sends them to the server, and the TCP-friendly rate is computed at the server based on the received feedback.

Currently, P2P may be the most economical overlay network solution for delivering real-time media to a large number of users simultaneously. In general, we can classify issues in P2P system design into two broad categories:

- **Topology Discovery:** This refers to determining which peer is connected to which other peer(s) over what kind of links. It needs to be accomplished with the minimum number of message exchanges and done frequently enough to account for the peers who are leaving or joining and changing channel conditions. Both central solutions (e.g., Napster), which use a peer registry, and distributed solutions (e.g., Kazaa) exist for topology discovery.
- **Forwarding:** This refers to determining which peer is going to send what data block [or forward error correction (FEC)] to which of its connected peers. Topology discovery is a well-studied problem in the P2P file-sharing literature; examples include protocols such as NICE, NARADA, ZIGZAG, etc. A multitude of forwarding techniques for P2P video streaming with varying objectives has been presented in the literature [26]. Further challenges facing an interactive P2P 3DTV distribution system are discussed in the "Cooperative Streaming for Multiview Video Distribution" section.

It is also possible to employ architectures/protocols such as Digital Video Broadcasting–Handheld (DVB-H), Terrestrial Digital Multimedia Broadcasting (T-DMB), and MediaFLO for wireless broadcast of data streams consisting of IP packets. However, advanced MVV streaming features such as dynamic rate adaptation and selective streaming would not be applicable to such one-to-many wireless broadcast architectures.

Figure 4 depicts a block diagram of a flexible 3-D video transport architecture, which supports multiple 3-D video representations and encoders, unicast, application-layer multicast (ALM) and P2P modes, and one or more wired and wireless clients. The architecture should be flexible enough to stream n views, where $1 \leq n \leq N$, to one or more clients/peers. Clients that can view only monocular video should receive a monocular stream, clients with a stereo display should receive two streams, clients with a lenticular display should receive the number of streams that they require, and clients with a head-tracking display should receive a dynamically varying number of views. The architecture should be based on open standards, such as Moving Picture Experts Group (MPEG), JVT, and Internet Engineering Task Force (IETF) standards. Examples of unicast monocular video streaming systems are Apple Darwin, GPAC [27], and VideoLAN Client/Server [28]. The next section discusses extension of such classic unicast streamers to multiview video streaming. We then introduce a new selective streaming architecture that is unique to MVV streaming and discuss extension of cooperative streaming architectures to MVV delivery.

STEREOSCOPIC VIDEO CONSISTS OF TWO VIDEO SEQUENCES (LEFT AND RIGHT VIEWS) CAPTURED BY CLOSELY LOCATED CAMERAS.

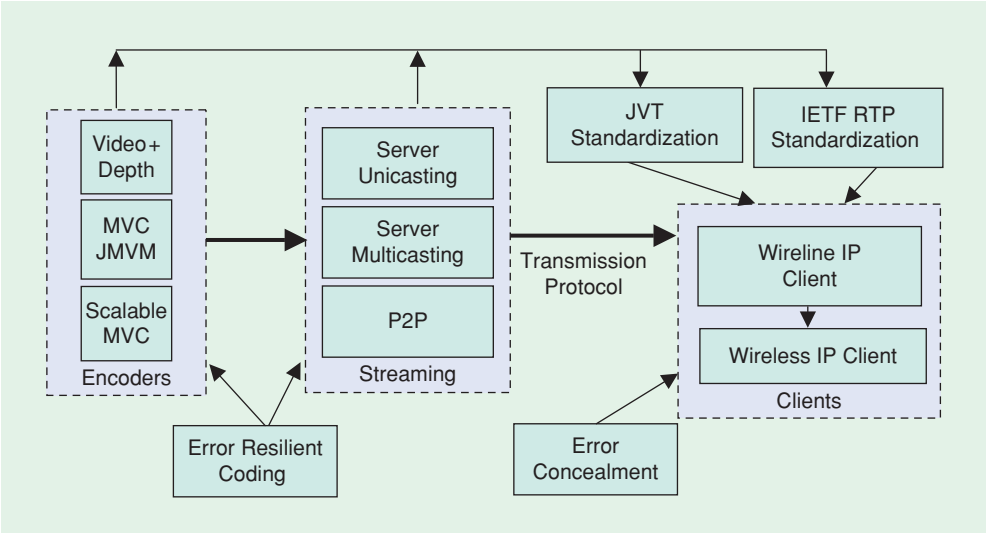
UNICAST STREAMING OF STEREO AND MULTIVIEW VIDEO

Unicast streaming servers can be based on RTP/TCP, RTP/UDP, or the newer RTP/DCCP protocol stacks. RTP/TCP may be preferred in some commercial solutions because it can easily penetrate firewalls.

STREAMING OVER UDP

An end-to-end prototype system for unicast streaming of offline encoded stereo video over RTP/UDP/IP has been recently developed [18]. A block diagram of this prototype is shown in Figure 5. The server can serve multiple clients simultaneously, which can display monoscopic or stereo streams based on their capabilities.

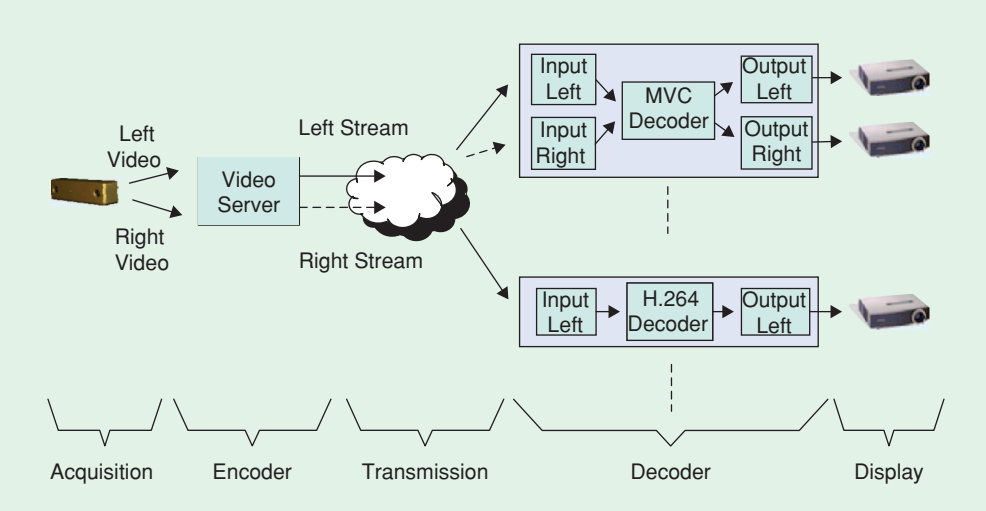
The session description protocol (SDP), with an additional session attribute to identify the right and left channels, has been used to ensure interoperability between the stereo video



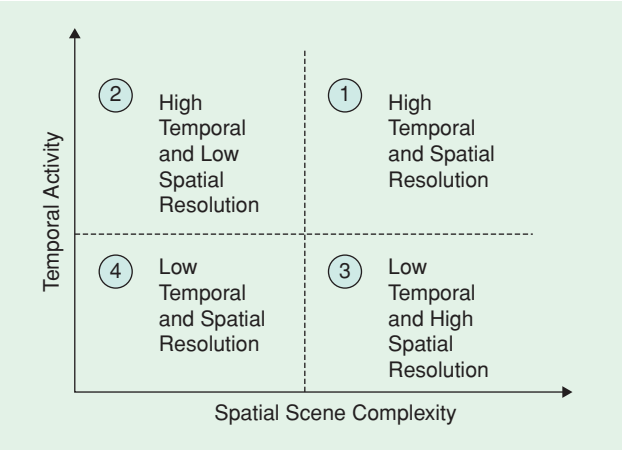
[FIG4] **Block diagram of a 3-D streaming system.**

server and clients. Three clients for different types of display systems have been implemented: 1) Client-1 supports a polarized 3-D projection display system (see Figure 1); 2) Client-2 supports an autostereoscopic 3-D laptop [2]; 3) Client-3 supports a monocular display to demonstrate backwards compatibility. The client handles packets of left and right views using two separate threads. Any MVC-compatible decoder can be used at the client. The decoder decodes and sends the decoded picture to the video output modules. The video output modules visualize the left and right frames in a synchronized manner by using the time information in the RTP timestamps.

Another 3DTV prototype system, with real-time acquisition, transmission, and autostereoscopic display of dynamic scenes, has been presented by Misubishi electric Research Laboratories (MERL) [29]. Multiple video streams are encoded and sent over a broadband network. The 3-D display shows high-resolution stereoscopic color images for multiple viewpoints without special glasses. This system uses lightfield rendering to synthesize views at the correct virtual camera positions. Both of these systems currently operate on a broadband or local area network; hence, no packet loss and video rate adaptation issues have been addressed.



[FIG5] **Block diagram of a 3-D streaming system.**



[FIG6] **Classification of GOPs according to spatial and temporal activity.**

RATE ADAPTATION FOR STREAMING STEREOSCOPIC VIDEO OVER DCCP

In streaming MVV over the Internet, the total video rate should be adapted to the available throughput and/or the TFRC rate in order to be friendly with other TCP traffic and avoid congestion. Rate adaptation of stereo and MVV differs from that of monocular video, since rate allocation between views offers new flexibilities. For example, the psycho-visual redundancy discussed earlier can be exploited for effective server-driven rate adaptation of stereo and MVV with unequal inter-view rate allocation. Several rate adaptation strategies at the server with or without client feedback are possible for stereo and MVV.

Rate adaptation and transcoding have been well-studied for monocular video [30]; however, extensions to MVV are relatively new [18], [31], [32]. It is possible to exploit the suppression theory of human stereo perception [8] to adapt the overall MVV rate to time-varying throughput [18] or the TRFC rate [31]. Dynamic rate adaptation of stereoscopic video can be achieved at almost constant perceptual quality by encoding one of the views at constant quality while varying the rate of the other view (using spatial, temporal, and/or quality subsampling) according to the network condition [18], [31]. Furthermore, the subsampling can be done adaptively to the content, where the stereo video is parsed into temporal segments, and each temporal segment of one view is encoded at lower spatial, temporal, and/or SNR resolution (hence at a lower target bit rate) depending on its low- and/or high-level content-based features. For example, the recently proposed content-adaptive stereoscopic encoder [18] classifies temporal segments into four categories according to their low-level attributes such as motion and spatial activity within the segment as follows (see Figure 6): Type 1 (high spatial and temporal activity): Do not scale the spatial and temporal formats; Type 2 (low spatial and high temporal activity): Apply spatial scaling but not temporal scaling; Type 3 (high spatial and low temporal activity): Apply temporal scaling but not spatial scaling; Type 4 (low spatial and temporal activity): Apply both temporal and spatial scaling.

Online rate adaptation can be performed either by adaptation of encoding parameters of a real-time MVC-compatible encoder or by layer extraction from an offline encoded scalable (SMVC) bit stream.

RATE ADAPTATION USING A REAL-TIME MVC ENCODER [18]

Rate adaptation can be achieved by online selection of the encoding parameters for each GoP in an MVC-compatible real-time encoder to downscale one of the views by: 1) spatial subsampling, 2) temporal subsampling, 3) scaling the quantization step size, or 4) content-adaptive scaling using a combination of the above. Simulations have shown that the average transmission rate for stereo video using this approach is about 1.2 times the bit rate of monoscopic video [18].

TWO BASIC SERVICE STRATEGIES WITH DIFFERENT REQUIREMENTS CAN BE IDENTIFIED FOR MVV STREAMING: LIVE/BROADCAST STREAMING AND ON-DEMAND STREAMING.

RATE ADAPTATION BY LAYER EXTRACTION FROM AN SMVC BIT STREAM [31]

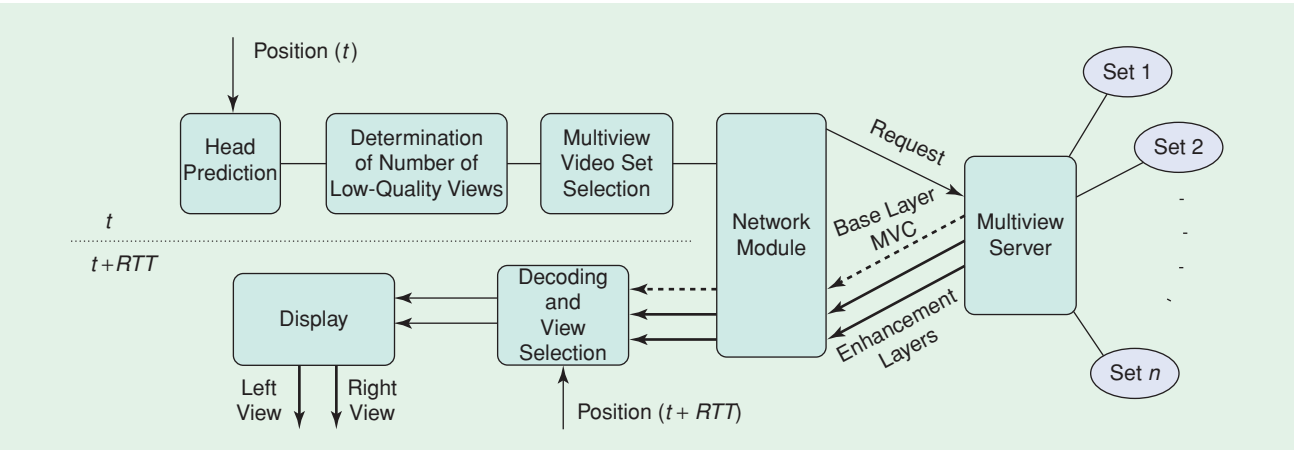
Alternatively, the video is scalable encoded offline with a predetermined number of spatial, temporal, and SNR scalability layers. Unequal bit allocation among the views is performed during bit stream extraction by selection of the number of spatial, temporal, and SNR scalability layers for each group of pictures (GOP) according to motion and spatial activity of that GOP. Experimental results demonstrating successful transmission of stereo video over DCCP with dynamic rate adaptation using the real Internet (between two cities) have been reported in [31].

In related work, rate-distortion-optimized transmission for interactive lightfield streaming has been proposed by Chang and Girod [32], where the lightfield data is transformed into blocks of wavelet coefficients; each block is then coded as a scalable bit stream and stored at the sender. Based on the estimated state of data already at the receiver, the network characteristics, and desired transmission rate, the sender performs rate-distortion-optimized bit allocation for outgoing packets to minimize the distortion of the frame rendered at the receiver. Experimental results using a statistical network model show that the proposed rate-distortion-optimized scheme reduces the required bit rate by 10% $\sim$ 25% over a heuristic scheme for a given rendering quality.

SELECTIVE UNICAST OF MULTIVIEW VIDEO FOR HEAD-TRACKING DISPLAYS

A client-driven selective MVV streaming architecture that allows a user to watch 3-D video interactively with significantly reduced bandwidth by transmitting a small number of views selected according to his/her head position has been presented in [33]. The objective of this system is to efficiently stream a set of multiview sequences, in the sense of best usage of limited network resources, by allocating most of the available bit rate to the required views assuming that there is not sufficient bandwidth to stream all views to all users all the time. Both dense multiview representations (lightfields) and wider baseline multiview sequences, together with depth information, can be employed in this architecture. The user's head position is tracked and predicted into the future to select the views that best match the user's current viewing angle dynamically. Prediction of future head positions is needed so that views matching the predicted head positions can be prefetched from the server ahead of time in order to account for delays due to network transport and stream switching. The system allocates more bandwidth to the selected views to render the current viewing angle. Highly compressed, lower-quality versions of some other views are also requested to provide protection against having to display the wrong view when the current user viewpoint differs from the predicted viewpoint. An objective measure based on the abruptness of the head movements and delays in the system is introduced to determine the number of additional lower-quality views to be prefetched. The multiview encoder makes use of MVC and scalable video coding (SVC) concepts together to obtain improved compression efficiency while providing flexibility in bandwidth allocation to the selected views.

A block diagram of the system is shown in Figure 7. Suppose that we have an MVV with N views on a server. The client side

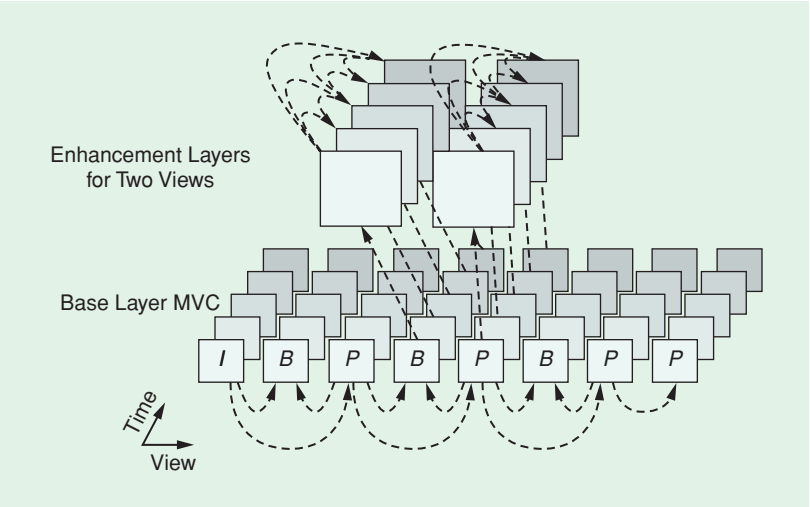


[FIG7] **Overview of head-tracking selective multiview video streaming system [33].**

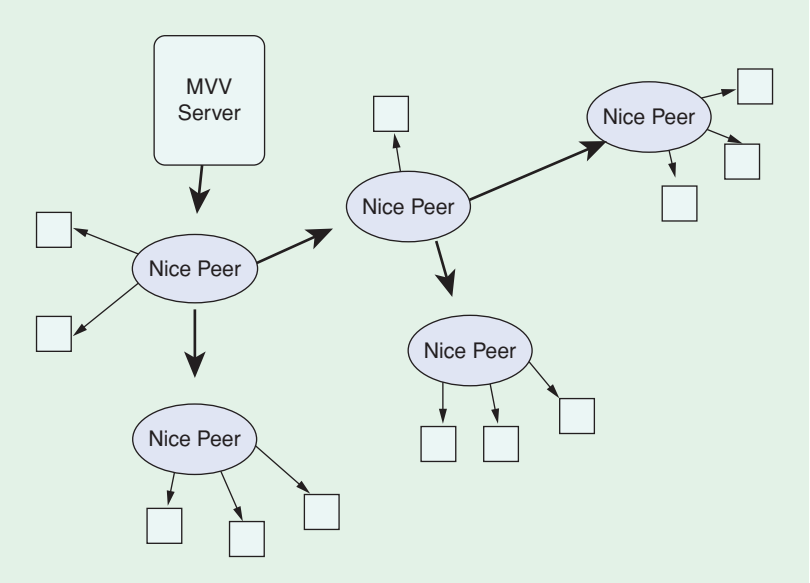
first determines the user's current head position and a Kalman-filter-based predictor predicts the user's head position d frames into the future. Then, an error measure is computed at the client to determine the number of views, $M \leq N$, to be requested from the server. The server selectively streams the MVV sequence encoded at two quality levels. As a base layer, all M views are encoded using the MVC codec at a lower bit rate. On top of this base layer, an enhancement layer is encoded for each view independently of other enhancement layers to allow random access to improve the quality of the selected views. This encoding scheme is illustrated in Figure 8. Since the total bandwidth available to the user is assumed fixed, an increased proportion of the bandwidth needs to be allocated to the base layer as M increases. This necessitates an intelligent rate allocation scheme between the base layer MVC and enhancement layer streams. We assume that that the server hosts several sets of the same MVV, with each set encoded using a different value of M and different rate allocations between the base and enhancement layers.

The client switches to the appropriate set of streams according to its bandwidth, user's predicted head position, and the current value of M. If there are no prediction errors, the received high-quality (base and two enhancement) streams are passed on to the display, which shows a high-quality view to each eye. The low-bit-rate base layer MVC enables the user to keep watching 3-D video, albeit possibly at a lower quality, when the current user head position differs from the predicted position until correct high-quality streams arrive from the server. If there is a prediction error and the wrong set of high-quality streams arrive, the system displays the low-quality version of the desired views which may be available in the base layer MVC only. According to [8], humans perceive high-quality 3-D video as long as one of the eyes sees a high-quality view. Therefore, in the presence of prediction errors, as long as at least one of the required views is delivered in high quality, the viewer might not even notice any loss of quality. If the prediction error is so severe that a required view is not delivered at all (is not among the M views in the base layer), an error concealment method is employed (e.g., the nearest available views are displayed).

It was demonstrated in [33] that selective streaming can offer up to 3 dB improvement compared to delivering standard MVC encoded 8-view video in the case of low network delay and viewer with smooth head movements. As the abruptness of head movements and the network delay increase, the advantage of a selective streaming system decreases, but a gain of about 1 dB was observed under typical conditions.



[FIG8] Multiview video encoding with MVC base layer and simulcast enhancement layers [33].



[FIG9] A sample distribution tree for a small multicast network.

COOPERATIVE STREAMING FOR MULTIVIEW VIDEO DISTRIBUTION

Cooperative streaming can be classified into two broad categories: application layer multicast (ALM) and P2P. ALM schemes postulate a central server where all content is located. Various ALM protocols organize the cooperative peers into a delivery tree rooted at the central server by using different algorithms in constructing and maintaining this tree (see Figure 9). P2P methods, on the other hand, are more flexible and do not necessarily require a central content server. Examples of P2P file sharing systems include KaZaa, Napster, and Bittorrent, where content can quickly spread from a few initial peers to thousands.

The challenges of cooperative media streaming are generally fundamentally different than file download. In addition to the

topology discovery and forwarding challenges, which have been discussed in the "Overview of Multiview Video Streaming Architectures and Protocols" section, a multiview 3-D video streaming system must also address the following issues:

1) *Path Diversity:* How to best exploit the path diversity in the overlay network is an issue for both ALM and P2P. Multiple description coding (MDC) is a promising approach that allows the use of path diversity in streaming at the expense of some coding redundancy. By sending different descriptions of coded video over different paths, the overall performance of the system can be increased in the presence of network losses [35], [36]. Multiview video (MVV) is inherently suited for MDC, where at the extreme each view could be declared as a description and can be streamed over different paths to reduce the effects of network losses. A multitree P2P streaming approach such as Chunkyspread [37] is suited for such an application with little modifications.

2) *Asymmetry of Bandwidth (BW):* Although users may have enough BW to receive streams, their out-bound BW might be much smaller, which limits the total network capacity. In addition to prioritizing peers with larger uplink (UL) BW, layered and multiple description compression techniques can be utilized to overcome this [36], [38]. Hosseini and Georganas [34] have demonstrated a 3-D video conferencing system utilizing an ALM protocol and introduced the awareness-driven video concept to overcome the up-stream bandwidth limitations for multi-user video conferencing by only delivering videos for some of the users, which is parallel to the selective streaming idea discussed earlier. Another solution exclusive for MVV streaming is requesting different views from different peers, thus reducing the load on the outbound capacity of other peers.

3) *Enticement of Peers to Stay Connected and Commit Resources:* The peers in an overlay network are more likely to disappear when compared to the routers in a multicast-capable network. The peers might fail or leave the overlay network without prior notice. Moreover, selfish users might not be willing to share their UL bandwidth. P2P file-sharing services usually attempt to overcome this issue by giving incentives for resource sharing. Peers are ranked according to their UL to downlink (UL/DL) ratio in queues for chunks (Bittorrent, eDk2000). In streaming systems, the committed UL bandwidth is more important than total UL/DL ratio. In [39], peers who commit a larger UL BW are prioritized during network join if there is limited capacity available. This both offers an incentive for users to commit as much BW as possible and enables the network capacity to grow faster. Additionally, as discussed in [40], a further incentive can be unrestricted peer selection for peers who contribute more UL capacity, whereas "free-riders" can be penalized by being able to select from a limited list of source peers. A further extension of this idea in multiview streaming is prioritization of contributing peers during view switching in a selective streaming scenario as discussed in the "Selective Unicast of Multiview Video for Head-Tracking Displays" section. A peer who commits more UL bandwidth can be provided with a wider selection of source peers during view switching such that the transition time is minimized leading to a better 3-D experience.

Two basic service strategies with different requirements can be identified for MVV streaming: live/broadcast streaming and on-demand streaming. All users in a broadcast system are synchronized and receive the same data, whereas in an on-demand system, the playback is asynchronous for different users. Therefore, in a broadcast system the packets can be simply forwarded down the delivery tree, but an on-demand system requires the peers to buffer some of the content [41], [42] so that other peers can request past packets belonging to the stream. The DONet/Coolstreaming [43] architecture provides a flexible solution for this problem by transmitting a buffer map in randomized refresh messages between peers. This provides an efficient method with low overhead to exchange data availability information between peers. In addition to keeping track of the buffering ranges of different peers, the set of views buffered by each peer must be considered as well for MVV systems, where some users might opt for selective delivery and receive a subset of all available views. We believe that the buffer map architecture provided by DONet/Coolstreaming can be readily extended for such purposes.

The requirements on the cooperative delivery system can also vary significantly according to the target display system. An N-view autostereoscopic display system needs a time-invariant set of N-views to correctly show the 3-D scene. On the other hand, a display system with viewpoint tracking can request a time-varying set of streams according to the user's viewpoint and requires a selective streaming system. Unlike a conventional video or static MVV system, such a dynamic system puts new constraints on the network protocol, including low join-latency in order to provide an interactive experience during view switching [44] and robust delivery trees. Since the peers follow users' head movements, the average duration of participation in the network for a particular view is quite short and there is a clear need for new methods that prevent downstream peers from starvation if a peer stops receiving and forwarding packets for a particular view.

PACKET-LOSS RESILIENT STREAMING AND MULTIVIEW ERROR CONCEALMENT

Streaming media applications often suffer from packet losses in the wired or wireless IP links. Congestion is the main cause of packet losses over the wired Internet. In contrast to the wired backbone, the capacity of the wireless channel is fundamentally limited by the available bandwidth of the radio spectrum and various types of noise and interference, which leads to bit errors. Most network protocols discard packets with bit errors, thus translating bit errors into packet losses. Several joint source and channel techniques have been developed for efficient transmission of monocular video streams over packet erasure channels, both in wired and wireless networks [42], [45]. Furthermore,

error concealment methods at the decoder have been considered in order to limit the damage, especially due to temporal error propagation, resulting from unpreventable packet losses.

Common approaches for reliable transmission of monoscopic video over packet networks include retransmission requests (ARQs) [46], [49] and/or FEC methods [47]–[49]. ARQ methods, which require feedback messages (ACK) that inform the sender about the reliable reception of the data, may be effective to deal with packet losses if sufficient playout (preroll) delay is allowed at the client. It may be more desirable to employ time-limited ARQ at the application layer over the UDP or DCCP protocol, which allows ARQ only within a limited period (less than the preroll delay at the client) as opposed to unlimited ARQ at the network layer (as in the TCP protocol). In cases where feedback channel cannot be used extensively, such as in broadcast and multicast services, channel coding techniques have been widely applied to combat with transmission errors. In [48] and [50], the transmission of MVV streams over packet erasure networks is examined. Macroblocks are classified into unequally important slice groups using the flexible macroblock ordering (FMO) tool of H.264/AVC. Stereoscopic video streaming using FEC techniques are examined in [50] and [51]. Frames are classified according to their contribution to the overall quality to form three layers, which are used for unequal error protection (UEP). A comparative analysis of Reed Solomon (RS) and systematic Luby Transform (LT) codes are provided via simulations to observe the optimum packetization and UEP strategies.

Several studies exist on frame loss concealment for monoscopic video, but they may not be directly applicable to stereoscopic video since human perception of errors in 3-D video is different than in the 2-D case. An error concealment algorithm that fully makes use of the characteristics of stereoscopic video is proposed in [52]. Based on relativity of prediction modes for right frames, prediction mode of each macroblock in the lost frame is chosen and finally utilized to restore the lost macroblock according to the estimated motion vector or disparity vector. A strategy for concealment of loss of block bursts in independently coded stereo video was studied in [53] assuming block-based video coding. Additional information from the corresponding view is employed to increase the quality of the reconstructed block due to high correlation between the views.

CONCLUSIONS AND FUTURE RESEARCH

Promising approaches for encoding stereoscopic and MVV have been standardized in the form of MPEG "video-plus-depth" and the JVT MVC standards. It has been shown that both approaches can encode stereoscopic video at about 1.2 times the bit rate of monoscopic video (when using unequal inter-view bit allocation in the case of MVC). A multitude of strategies have been considered for streaming such encoded MVV using RTP/UDP/IP or RTP/DCCP/IP. Video streaming architectures can be classified as 1) server unicasting to one or more clients, 2) server multicasting to several clients, 3) P2P unicast distribution, where each peer forwards packets to another peer, and 4) P2P multicasting, where each peer forwards packets to several other peers. Main current research issues in MVV streaming are: 1) determination of the best video encoding configuration for each streaming strategy—multi-view video encoding methods provide some compression efficiency gain at the expense of creating dependencies between views that hinder random access to views; 2) determination of the best rate adaptation method—adaptation refers to adaptation of the rate of each view and inter-view rate allocation depending on available network rate and video content, and adaptation of the number and quality of views transmitted depending on available network rate and user display technology and desired viewpoint; 3) packet-loss resilient video encoding and streaming strategies as well as better error concealment methods at the receiver; and 4) best P2P multicasting design methods, including topology discovery, topology maintenance, forwarding techniques, exploitation of path diversity, methods for enticing peers to send data and to stay connected, and use of dedicated nodes as relays.

ACKNOWLEDGMENT

This work is supported by EC within FP6 under Grant 511568 with the acronym 3DTV.

AUTHORS

A. Murat Tekalp (mtekalp@ku.edu.tr) received the M.S. and Ph.D. degrees in electrical, computer, and systems engineering from Rensselaer Polytechnic Institute (RPI), Troy, New York, in 1982 and 1984, respectively. He has been with Eastman Kodak Company (1984–1987) and with the University of Rochester, Rochester, New York (1987–2005), where he was promoted to Distinguished University Professor. Since 2001, he is a professor at Koc University, Istanbul, Turkey. He has chaired the IEEE Signal Processing Society Technical Committee on Image and Multidimensional Signal Processing (1996–1997). He was a Distinguished Lecturer of the IEEE Signal Processing Society (1998). He has served as an associate editor for the *IEEE Transactions on Signal Processing* (1990–1992) and *IEEE Transactions on Image Processing* (1994–1996). At present, he is the editor-in-chief of the EURASIP journal *Signal Processing: Image Communication* published by Elsevier. He has been the technical program cochair for IEEE ICASSP 2000 and the general chair of IEEE International Conference on Image Processing (ICIP) 2002. He authored the book *Digital Video Processing* (Prentice Hall, 1995). He is a Fellow of the IEEE.

Engin Kurutepe (kurutepe@nue.tu-berlin.de) received B.S. and M.S. degrees from Koc University, Istanbul, Turkey in 2004 and 2006, respectively, both in electrical and computer engineering. He then joined the Communications Systems Group at Technische Universitaet Berlin, Germany, where he is currently a Ph.D. student. His research interests concentrate on processing, compression, and transmission of conventional and MVV

M. Reha Civanlar (rcivanlar@docomolabs-usa.com) received the B.S. and M.S. degrees in E.E. from METU and Ph.D. in ECE from NCSU. He is a VP in DoCoMo USA Labs. He is on the advisory boards of Argela Technologies and Layered Media. He was a visiting professor at Koc University (2002–2006), leading a 3DTV

transport project. He headed Visual Communications Research, AT&T (1991–2002) after working at Bell Laboratories. He has over 40 patents and is an ASSP Senior Award recipient. He served as an editor for *IEEE Transactions on Communications, Multimedia, and JASP* and *Image Communication*, and was on MMSP and MDSP TCs of the IEEE Signal Processing Society. He is a Fellow of the IEEE.

REFERENCES

[1] W.B. Norton, "Video Internet: The next wave of massive disruption to the U.S. peering ecosystem," white paper, version 1.3 [Online]. Available: http://forum.stanford.edu/events/2007/cleanslate/slides/ Internet%20Video%20Next%20Wave%20of%20Disruption%20v1.3.pdf

[2] H. Isono and M. Yasuda, "Three-dimensional image display using electrically generated parallax barrier stripes," U.S. Patent No 5 315 377, May 1994.

[3] Multiview Display Technology. [Online]. Available: http://www.research.philips.com/newscenter/archive/2004/3d-display-sid.html

[4] Head Tracking Auto Stereoscopic Display Technology. [Online]. Available: http://www.hhi.fraunhofer.de/english/im/products/free2c/

[5] A. Smolic, P. merkle, K. Muller, C. Fehn, P. Kauff, and T. Wiegand, "Compression of multi-view video and associated data," in *Three Dimensional Television: Capture, Transmission, and Display*, H. Ozaktas and L. Onural, Ed. New York: Springer, 2007.

[6] *MPEG-4 Animation Framework eXtension (AFX)*, ISO/IEC 14496-16 2nd ed. [Online]. Available: www.mpeg-3dgc.org

[7] C. Fehn, "Depth-image-based rendering compression and transmission for a new approach on 3D-TV," in *Proc. SPIE Stereoscopic Displays and Virtual Reality Syst. XI*, San Jose, CA, USA, pp. 93–104, Jan. 2004.

[8] L.B. Stelmach, W.J. Tam, D. Meegan, and A. Vincent, "Stereo image quality: Effects of mixed spatio-temporal resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 2, pp. 188–193, 2000.

[9] A. Puri, R.V. Kollarits, and B.G. Haskell, "Basics of stereoscopic video, new compression results with MPEG-2 and a proposal for MPEG-4," *Signal Processing: Image Comm.*, vol. 10, no. 1–3, pp. 201–234, July, 1997.

[10] P. Merkle, K. Müller, A. Smolic, and T. Wiegand, "Efficient compression of multi-view video exploiting inter-view dependencies based on H.264/MPEG4-AVC," in *Proc. IEEE Int. Conf. on Multimedia & Expo (ICME)*, Toronto, Canada, July 2006.

[11] Joint Video Team, "Joint Multiview Video Model JMVM-4," JVT-W208, San Jose, April 2007.

[12] A. Luthra, G.J. Sullivan, and T. Wiegand, Eds., *IEEE Trans. Circuits Systems Video Technol.*, vol. 13, no. 7, July 2003.

[13] *Advanced Video Coding for Generic Audiovisual Services*, ITU-T H.264 and ISO/IEC MPEG 14496-10, 2005.

[14] N. Ozbek and A.M. Tekalp, "Scalable multi-view video coding for interactive 3DTV," in *Proc. IEEE Int. Conf. on Multimedia & Expo (ICME)*, Toronto, Canada, July 2006.

[15] M. Dröse, C. Clemens, and T. Sikora, "Extending single-view scalable video coding to multi-view based on H.264/AVC," in *IEEE Int. Conf. Image Proc. (ICIP)*, Atlanta, GA, USA, Oct. 2006.

[16] I. Dinstein, M.G. Kim, A. Henik, and J. Tzelgov, "Compression of stereo images using subsampling transform coding," *Optical Eng.*, vol. 30, no. 9, pp. 1359–1364, Sept. 1991.

[17] W. Woo and A. Ortega, "Optimal blockwise dependent quantization for stereo image coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, pp. 861–867, Sept. 1999.

[18] A. Aksay, S. Pehlivan, E. Kurutepe, C. Bilen, T. Ozcelebi, G. Bozdagi Akar, M.R. Civanlar, and A.M. Tekalp, "End-to-end stereoscopic video streaming system with content-adaptive rate and format control," *Signal Processing: Image Commun. (Special Issue on 3DTV)*, vol. 22, no. 2, pp. 157–168, Feb. 2007.

[19] N. Ozbek, A.M. Tekalp, and T. Tunali, "Rate allocation between views in scalable stereo video coding using an objective stereo quality measure," in *Proc. IEEE ICASSP*, Honolulu, Hawai, April 2007.

[20] C. Fehn, et al., "Asymmetric coding of stereoscopic video for transmission over T-DMB," in *Proc. 3DTV-CON*, Kos, Greece, May 2007.

[21] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A transport protocol for real-time applications," *IETF*, July 2003, RFC 3550. [Online]. Available: http://www.ietf.org/rfc/ref3550.txt

[22] S. Wenger, M.M. Hannuksela, T. Stockhemmer, M. Westerlund, and D. Singer, "RTP payload format for H.264 video," RFC 3984, Feb. 2005. [Online]. Available: http://www.ietf.org/rfc/ref3984.txt

[23] E. Kohler, M. Handley, and S. Floyd, "Datagram congestion control protocol (DCCP)," IETF, March 2006, RFC 4340. [Online]. Available: http://www.ietf.org/rfc/ref4340.txt

[24] S. Floyd, E. Kohler, and J. Padhye, "Profile for DCCP congestion control ID 3: TCP-friendly rate control (TFRC)," IETF, March 2006, RFC 4342. [Online]. Available: http://www.ietf.org/rfc/ref4342.txt

[25] J. Widmer and M. Handley, "TCP-friendly multicast congestion control," IETF, Aug. 2006, Rep. RFC4654. [Online]. Available: http://www.ietf.org/rfc/ref4654.txt

[26] E. Setton, J. Noh, and B. Girod, "Rate-distortion optimized video peer-to-peer multicast streaming," in *Proc. ACM P2PMMS'05*, pp. 39–48, 2005.

[27] Open Source Multimedia Framework. [Online]. Available: http://gpac.sourceforge.net/

[28] VLC Media Player. [Online]. Available: http://www.videolan.org/vlc/

[29] A. Vetro, W. Matusik, H. Pfister, and J. Xin, "Coding approaches for end-to-end 3D TV systems," in *Proc. Picture Coding Symp. (PCS)*, December 2004.

[30] S.F. Chang, and A. Vetro, "Video adaptation: Concepts, technologies, and open issues," in *Proc. IEEE*, vol. 93, no. 1, pp. 148–158, Jan. 2005.

[31] N. Ozbek, B. Gorkemli, A.M. Tekalp, and E.T. Tunali, "Adaptive streaming of scalable stereoscopic video over DCCP," in *Proc. IEEE Int. Conf. Image Processing*, San Antonio, Texas, Sept. 2007.

[32] C.-L. Chang and B. Girod, "Receiver-based rate-distortion optimized interactive streaming for scalable bitstreams of light fields," in *Proc. IEEE ICME 2004*, vol. 3, pp. 1623–1626, 2004.

[33] E. Kurutepe, M.R. Civanlar, and A.M. Tekalp, "Client-driven selective streaming of multi-view video for interactive 3DTV," to appear in *IEEE Trans. CSVT*, Oct. 2007.

[34] M. Hosseini and N.D. Georganas, "Design of a multi-sender 3D videoconferencing application over an end system multicast protocol," in *Proc. 11th ACM Intl. Conf. on Multimedia*, pp. 480–489, 2003.

[35] R. Tian, Q. Zhang, Z. Xiang, Y. Xiong, X. Li, and W. Zhu, "Robust and efficient path diversity on application-layer multicast for video streaming," *IEEE Trans. Circuits Syst. Video Techol.*, vol. 15, no. 8, pp. 961–972, 2005.

[36] E. Akyol, A.M. Tekalp, and M.R. Civanlar, "A flexible multiple description coding framework for adaptive peer-to-peer video streaming," *IEEE J. Selected Areas Signal Processing*, vol. 1, no. 2, pp. 231–245, Aug. 2007.

[37] V. Venkataraman, K. Yoshida, and P. Francis, "Chunkyspread: Heterogeneous unstructured tree-based peer-to-peer multicast," in *Proc. IEEE ICNP 2006*, pp. 2–11.

[38] Y. Cui and K. Nahrstedt, "Layered peer-to-peer streaming," in *Proc. ACM NOSSDAV'03*, pp. 162–171, 2003.

[39] D. Xu, M. Heefeda, S. Hambrusch, and B. Bhargava, "On peer-to-peer media streaming," in *Proc. IEEE ICDCS 2002*.

[40] A. Habib and J. Chuang, "Incentive mechanism for peer-to-peer media streaming," in *Proc. IWQOS 2004*, pp. 171–180.

[41] S. Jin and A. Bestavros, "Cache-and-relay streaming media delivery for asynchronous clients," in *Proc. 4th Int. Workshop on Networked Group Communication, ACM*, Boston, MA, 2002, pp. 37–44.

[42] Y. Cui, B. Li, and K. Nahrstedt, "oStream: asynchronous streaming multicast in application-layer overlay networks," *IEEE J. Selected Areas Commun.*, vol. 22, no. 1, pp. 91–106, 2004.

[43] X. Zhang, J. Liu, B. Li, and T.S.P. Yum, "CoolStreaming/DONet: A data-driven overlay network for peer-to-peer live media streaming," in *Proc. IEEE Infocom 2005*, 2005.

[44] E. Kurutepe, M.R. Civanlar, and A.M. Tekalp, "A receiver-driven multicasting framework for 3DTV transmission," in *Proc. EUSIPCO*, Antalya, Turkey, Sept. 2005.

[45] Y. Wang, S. Wenger, J. Wen, and A. Katsaggelos, "Error resilient video coding techniques," *IEEE Signal Processing Mag.*, vol. 17, no. 4, pp. 61–82, July 2000.

[46] G.J. Conklin, G.S. Greenbaum, K.O. Lillevold, A.F. Lippman, Y.A. Reznik, R.N. Inc, and W.A. Seattle, "Video coding for streaming media delivery on the Internet," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 3, pp. 269–281, 2001.

[47] M. Link, B. Girod, K. Stuhlmuller, and U. Horn, "Packet loss resilient internet video streaming," in *Proc. SPIE Visual Comm. Image Processing*, 1999.

[48] H. Cai, B. Zeng, G. Shen, Z. Xiong, and S. Li, "Error-resilient unequal error protection of fine granularity scalable video bitstreams," *EURASIP J. Applied Signal Processing*, vol. 2006, 2006.

[49] Z. Tan and A. Zakhor, "Error control for video multicast using hierarchical FEC," in *Proc. Int. Conf. on Image Processing*, Kobe, Japan, vol. 1, pp. 401–405, Oct. 1999.

[50] S. Argyropoulos, S. Tan, N. Thomos, E. Arikan, and M. Strintzis, "Robust transmission of multi-view video streams using flexible macroblock ordering and systematic LT codes," in *Proc. 3DTV-CON*, Kos, Greece, May 2007.

[51] A.S. Tan, A. Aksay, C. Bilen, G. Bozdagi Akar, and E. Arikan, "Error resilient layered stereoscopic video streaming," in *Proc. 3DTV-CON*, Kos, Greece, May 2007.

[52] L. Pang, M. Yu, G. Jiang, Z. Jiang and F. Li, "An approach to error concealment for entire right frame loss in stereoscopic video transmission," in *Proc. Int. Conf. Computational Intelligence and Security*, China, Nov. 2006.

[53] C. Clemens, M. Kunter, S. Knorr, and T. Sikora, "A hybrid approach for error concealment in stereoscopic images," in *Proc. WIAMIS*, 2004. SP