Transport Methods in 3DTV - A Survey

Gozde B. Akar¹, A. Murat Tekalp², C. Fehn³, and R. Civanlar⁴

¹ Dept. of Elec. Eng., M.E.T.U., Ankara, Turkey
² College of Engineering, Koç University, Istanbul, Turkey
³ Fraunhofer-HHI, Berlin, Germany
⁴ NTT DoCoMo USA Labs, Palo Alto, California, USA

Abstract

Three dimensional (3D) movies have quite a long history. However, technologies needed for successful deployment of 3D television (3DTV) systems are becoming available only recently. This survey outlines previous attempts to transport 3D video signals, and technologies that may be useful in designing 3DTV systems of tomorrow. Current approaches for digital broadcast and streaming over the Internet are discussed in more detail, including the 3D video representations and encoding methods; rate adaptation/error correction methods and transport protocols. Streaming over the Internet provides a flexible architecture to support different display technologies, visual data representations, and encoding options, as well as different transport protocols for unicast and multicast 3DTV distribution. Some exemplary digital 3DTV broadcast and streaming system demonstrations are also reviewed. This survey is intended as a starting reference for 3DTV transport researchers.

Keywords: multi-view video, video-plus-depth representation, stereoscopic displays, 3DTV broadcasting, 3DTV streaming over IP.
1. INTRODUCTION

Although the ultimate goal in 3D video and TV may be dynamic holography, most systems available today create the 3D viewing experience via stereoscopy, that is, by showing a scene from slightly different angles to the left and right eyes of a viewer. Three dimensional television (3DTV) systems can be designed to support fixed-view stereoscopy, represented by only two views, where all viewers see the 3D scene from the same viewing angle, and/or free-view stereoscopy, represented by multiple views, to enable viewers to watch the 3D scene from different angles within a limited viewing range on 3D displays that support this functionality.

The history of 3D visual motion imagery can be traced back to 1903, when the first stereoscopic 3D movie was created. This could only be watched from a fixed viewing angle by one viewer at a time with a modified stereoscope. In 1922, the first full length stereoscopic movie was shown simultaneously to a large group of viewers using the anaglyphic process. Hollywood started 3D movie production in big numbers in the 1950s [1][2]. While 3D cinema does not have to deal with transport issues, the broadcast industry had to deal with efficient transport of content in addition to issues related to effective and inexpensive 3D displays for successful deployment of 3DTV. Hence, transmission of 3D video signals is a vital and challenging component of a 3DTV system. Over the years a consensus has been reached among experts that the introduction of 3DTV can only be a lasting success if it is backwards compatible with the conventional 2D television, supports different numbers of users with affordable 3D display technologies, requires low additional transport/transmission overhead, and if the perceived visual quality and viewing comfort is better than the conventional 2DTV [3].
Development of a backwards compatible and efficient 3DTV transport technology requires thorough consideration of the end-to-end system, including the 3D display technology, visual data representation and rendering, as well as capture and encoding methods. Different display technologies may mandate different visual data representations, which in turn may affect the optimal compression and transport strategies. The evolution of 3DTV transport technology follows the path of analog broadcast, digital broadcast, and most recently streaming over the Internet Protocol (IPTV). Analog and digital broadcasting of 3DTV are reviewed in Section 2 together with the required visual data representation, rendering/display, and capture and encoding methods. Streaming over IP provides a more flexible means of 3DTV. The models and issues related to 3DTV transport over IP networks are discussed in Section 3. Discussion and conclusions are provided in Section 4.

2. 3DTV BROADCAST

This section reviews analog and digital 3DTV broadcast technologies. The analog transmission technologies are included to provide a historical perspective.

2.1. Analog Transmission

The first known experimental television 3D broadcast in the U.S. was on April 29, 1953 when a trial live broadcast of the series SPACE PATROL was run in Los Angeles at the National Association of Radio and Television Broadcasters 31st Annual gathering. The ABC affiliate station KECA-TV also aired the show but viewers without a pair of special polarization lenses saw only a blurred mess [1]. The first “non experimental” 3D television broadcast occurred about 30 years later. On December 19, 1980, a 3D feature
film, “Miss Sadie Thompson”, and a 3D short starring the Three Stooges were aired over SelecTV, a Los Angeles Pay-TV system. This broadcast was made possible by a new company, 3D Video Corporation, who developed a working and practical system for presenting 3D on conventional television sets using the anaglyph format [4]. This first broadcast on SelecTV was B&W, i.e. the color of the original film was removed prior to conversion with the 3DTV process. With the success of this first broadcast, SelecTV asked 3D Video Corporation to convert another film, this time preferably in color. This has been achieved on April 10, 1981, when SelecTV broadcasted MGM’s musical classic, “Kiss Me Kate.” Following these early efforts, several other features and shorts were produced and broadcasted worldwide in subsequent years.

In Europe, one of the earliest 3D television activities were the experimental 3DTV broadcast programs that were transmitted in 1982 in several European countries [3]. These programs, which were aired in a simple red/green anaglyph format, were initiated by H.-J. Herbst of the Norddeutscher Rundfunk (NDR), Hamburg. In cooperation with Philips Research Laboratories, Eindhoven, two popular-scientific 3D series were produced. Together with some further 3D productions by other European TV stations, these transmissions received an extremely favorable response. More than 40 million red/green viewing spectacles were sold. The expectation, however, that this could be ‘the TV of the future’ was disillusioned by the poor visual quality attainable with the anaglyphic method, especially if transmitted via an unmodified standard TV system.

A second activity that has given further stimulation to the 3DTV research and development efforts was the 3DTV demonstration in 1983 at the International Audio and Video Fair in Berlin [3]. Because of the high public interest in the earlier, anaglyphic 3DTV transmissions, the German broadcasters asked the Institut für Rundfunktechnik (IRT) in
Munich to also showcase the high-quality 3DTV system developed at IRT at that time. This stereoscopic 3D system was based on a standard PAL (Phase Alternating Line) distribution chain that was operated in two-channel mode. For display, two video projectors with orthogonal polarization filters were used; in order to separate the views, the users had to wear matching polarization glasses. The presentations were so successful that they were continued at the Audio and Video Fairs 1985 and 1987. Unfortunately, because the transmission system required a custom TV receiver, its use remained limited to large venue demonstrations at exhibitions, symposia, conferences, and workshops.

In June of ‘91, John Wayne's only 3D film, Hondo, was broadcast on a network of 151 stations in the US. This was the closest event to a “network” 3D broadcast to that date.

2.2. Digital Transmission

With the ongoing transition from analogue to digital TV services, additional hope for 3D television arose in the early 1990s; and especially in Europe, a number of European-funded projects (e.g., COST 230, RACE DISTIMA, RACE PANORAMA, ACTS MIRAGE, ACTS TAPESTRIES) were set-up with the aim to develop standards, technologies and production facilities for 3DTV [5]. Other groups, which focused on human factors requirements for high-quality stereoscopic television, also joined the efforts [6]. Motivated by this revived interest in broadcast 3DTV, the Moving Pictures Expert Group (MPEG) developed a new compression technology for stereoscopic video as part of the successful MPEG-2 standard [7]. The chosen approach, called the MPEG-2 Multi-view Profile (MVP), can be regarded as an extension of the temporal scalability tool. It encodes the left-eye view as a base layer in conformance with the MPEG-2 Main Profile (MP) – thus providing backwards-compatibility with the ‘conventional’ 2D digital TV receivers. The right-eye
view is encoded as an enhancement layer using the scalable coding tools with additional prediction from the base layer. While MPEG-2 has become the underlying technology for digital standard-definition and high-definition 2DTV broadcasts worldwide, the Multi-view Profile (MVP), unfortunately, has not found use in commercially available services.

Some promising attempts have been made to integrate stereoscopic TV and digital HDTV (High-Definition Television) into a new, high-quality 3D entertainment medium. Live broadcasting of stereoscopic HDTV was tried in Japan during the Nagano Winter Games in 1998 [8]. In cooperation with other organizations, TAO distributed right-eye and left-eye HDTV images of the events in Tokyo at a bitrate of 45 Mbits/s each via N-Star to NHK Broadcasting Center and Chiyoda Broadcasting Hall, both in Tokyo, where they were projected onto a large screen. These live 3D images included those from hockey games and other events, impressing the audience with their powerful sense of reality.

A similar experiment was conducted in Korea/Japan during FIFA World Cup 2002 [9]. Using a terrestrial and satellite network, a compressed stereoscopic 3D HDTV signal was multicast to seven predetermined demonstration venues, which were approved by host broadcast service (HVS), KirchMedia, and the FIFA. More specifically, the right-eye and left-eye HDTV images were compressed in the so-called side-by-side format (horizontally decimated by a factor of two and rearranged into a single standard video field) using the MPEG-2 Main Profile@High Level codec at a bitrate of 40 Mbps and transmitted in an MPEG-2 Transport Stream (DVB-ASI) over a DS-3 network as specified in ITU-T Rec. G.703 [10]. At the receiving venues, the stereoscopic images recovered by the 3D HDTV receiver were decoded, expanded horizontally, and displayed on large screens with polarized beam projectors.
Recently, research on 3DTV has moved its focus from the classical fixed-view stereoscopic television concept towards more flexible 3D visual data representation formats. The Australian 3D company DDD, for example, is marketing multi-view autostereoscopic 3D displays, where the required views (typically 8 or 9 views) are generated from the so-called “video-plus-depth” representation, which is a combination of monoscopic color video and associated per-pixel depth maps [11]. They propose a system that encodes the depth data in a proprietary, very low bitrate format, which is then transmitted in the ‘private’ or ‘user data’ of an MPEG-2 Transport Stream [12]. Required views are rendered at the receiver side by using depth-image-based rendering (DIBR). A similar approach was followed by the European IST project ATTEST (Advanced Three-Dimensional Television Systems Technologies). Again, the “video-plus-depth” representation was used for transmitting 3D visual information. In contrast to the DDD concept, standard MPEG technologies were used for the compression of the depth information as well [13][14]. Within the project, it was shown that by using the newest MPEG video coding standard H.264/AVC (Advanced Video Coding), it is possible to compress typical depth data to bitrates of around 200-300 kbit/s. Thus, compared to a conventional 2D digital TV broadcast, the overhead required for the 3D visual information is in the area of only 10% (assuming a bitrate of around 3 Mbit/s for a typical 2DTV transmission over DVB-C/S/T).

First demonstration of a complete 3DTV system based on the ATTEST developments, whose block diagram is depicted in Figure 1, was shown by Fraunhofer HHI at the International Broadcast Convention (IBC) in 2004 [15]. The distribution side of the demonstration consisted of a DTV-Recorder-Generator (DVRG) from Rhode & Schwartz, which was used for the real-time replay of an offline-generated MPEG-2 Transport Stream. By means of a connected DVB-T sender (SFQ), the 3D television signal was aired at the booth. The transmitter power was adjusted such that reception of the signal could be ensured
within an area of a few square-meters. The MPEG-2 TS contained two 3D programs each comprised of an MPEG-2 coded color video stream as well as an associated H.264/AVC coded depth-image sequence. The synchronization of the two bitstreams was assured by conforming to the tools described in the MPEG-2 Systems specification and their respective amendments. The receiver side consisted of a conventional desktop PC with a PCI DVB-T card from TechnoTrend. The received MPEG-2 Transport Stream was demultiplexed in software and the respective synchronized video bitstreams were decoded in real-time and forwarded to a 3D renderer module which generated “virtual” stereoscopic views for display on the Fraunhofer HHI autostereoscopic Free2C 3D display. This was the first demo of a 3D TV service based on the “video-plus-depth” 3D data representation format using a real DVB-T transmission.

The “video-plus-depth” representation has been standardized within MPEG (Motion Pictures Experts Group) as a result of work initiated by Philips and Fraunhofer HHI. Only the representation has been standardized – by means of metadata which conveys the meaning of the graylevel values in the depth imagery – and some additional metadata to signal the existence of an encoded depth stream [2]. The actual compression of the per-pixel depth information, on the other side, has not been defined explicitly such that every conventional MPEG video codec (e.g., MPEG-2 or H.264/AVC) can be used for this purpose. The new standard has been published in two parts: The specification of the depth format itself is called ISO/IEC 23002-3 (MPEG-C), and a method for transmitting “video-plus-depth” within a conventional MPEG-2 Transport Stream has become an amendment (Amd. 2) to ISO/IEC 13818-1 (MPEG-2 Systems). Both standards have been finalized at the MPEG meeting in Marrakech, Morocco (January 2007). Technical details can be found in the standardization documents [54], [55].
3. 3DTV OVER IP NETWORKS

The Internet Protocol (IP) is proving to be very flexible in accommodating a wide range of communication services as can be seen from the ongoing replacement of classical telephone services by voice over IP applications. Transmission of video over the Internet is currently an active research and development area, where significant results have already been achieved. There are already video-on-demand services, both for news and entertainment applications, offered over the Internet. Also, 2.5G and 3G mobile network operators started to use IP successfully to offer wireless video services. Looking at these advances, the transport of 3DTV signals over IP packet networks seems to be a natural choice. The IP itself leaves many aspects of the transmission to be defined by other layers of the protocol stack and, thus, offers flexibility in designing the optimal communications system for various 3D data representations and encoding schemes.

3DTV streaming architectures (see Fig. 2) can be classified as: i) server unicasting to a single client, ii) server multicasting to several clients, iii) peer-to-peer (P2P) unicasting, where each peer forwards packets to another peer, and iv) P2P multicasting, where each peer forwards packets to several other peers. Multi-view video streaming protocols can be RTP/UDP/IP [16], which is the current state of the art, or RTP/DCCP/IP [18], which is the next generation protocol. Multicasting protocols can be supported at the network-layer or application layer. In the following, we first give an overview of the streaming protocols in Section 3.1. Then, data representation, encoding, and rate adaptation for multi-view video will be discussed in Section 3.2. Section 3.3 reviews methods for packet loss resilience. Examples of demonstration systems are reviewed in Section 3.4.
3.1 Streaming Protocols

Today, the most widely used transport protocol for media/multimedia is the Real-time Transport Protocol (RTP) over UDP [16]. However, RTP/UDP does not contain any congestion control mechanism and, therefore, can lead to congestion collapse when large volumes of multi-view video are delivered. The Datagram Congestion Control Protocol (DCCP) [18] is designed as a replacement for UDP for media delivery, running directly over the Internet Protocol (IP) to provide congestion control without reliability. DCCP can be thought as TCP minus reliability and in-order packet delivery, or as UDP plus congestion control, connection setup, and acknowledgements.

The Datagram Congestion Control Protocol (DCCP) is a transport protocol that implements bi-directional unicast connections of congestion-controlled, unreliable datagrams. Despite of 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 exist two congestion control schemes defined in DCCP currently, one of which is to be selected at connection startup time. These are TCP-like Congestion Control [19] and TCP-Friendly Rate Control (TFRC) [20]. TCP-like Congestion Control, identified by Congestion Control Identifier 2 (CCID2) in DCCP, behaves similar to TCP’s Additive Increase Multiplicative Decrease (AIMD) congestion control, halving the congestion window in response to a packet drop. Applications using this congestion control mechanism will respond quickly to changes in available bandwidth, but must tolerate the abrupt changes in the congestion window size typical of TCP. On the other hand, TFRC, which is identified by CCID3, is a form of equation-based flow control that minimizes abrupt changes in the sending rate while maintaining longer-term fairness with TCP. It is hence appropriate for
applications that would prefer a rather smooth sending-rate, including streaming media applications with a small or moderate receiver buffer. In its operation, CCID3/TFRC calculates an allowed sending rate, called the TFRC rate, by 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 in order to get better results.

There is also an experimental RFC for TCP-Friendly Multicast Congestion Control (TFMC) [56]. In order to compute the TFRC rate in a multicast scenario, each receiver computes their own TFRC rate as a function of their 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 their RTT and loss-rates and send them to the server, and the TCP-friendly rate is computed at the server based on the received feedback.

Hence, future 3DTV over IP services is expected to employ the DCCP protocol with effective video rate adaptation to match the TFRC rate. Multi-view video source rate adaptation strategies are discussed in the following.

3.2 Multi-View Video Encoding and Rate Allocation/Adaptation

For streaming applications, multi-view 3D video can be represented and encoded either implicitly, in the so-called “video-plus-depth” representation, or explicitly in raw form. Representation and encoding of “video-plus-depth” data was briefly discussed in Section 2.2. There are various approaches for representation and encoding of multi-view raw video, which provide a trade-off between random access, ease of rate adaptation, and compression efficiency. These approaches include simulcast coding, scalable simulcast coding, multi-
view coding [21]-[26], and scalable multi-view coding [27][28]. A complete treatment of 3D and free viewpoint video representations and their compression is given in [29].

In streaming multi-view video over the Internet, the video rate must be adapted to the available throughput and/or the TFRC rate in order to avoid congestion to be friendly with other TCP traffic. The rate adaptation of stereo and multi-view video differs from that of monocular video, since rate allocation between views offers new flexibilities. According to the suppression theory of human visual perception of 3D from stereoscopic video, if the right and left views are transmitted and displayed with unequal spatial, temporal and/or quality resolutions, the overall 3D video quality is determined by the view with the better resolution [57]. Therefore, rate adaptation of multi-view video may be achieved at constant perceived 3D video quality by adaptation of the spatial, temporal and/or SNR resolution of one of the views while encoding/transmitting the other view at full rate. Several open loop and closed loop rate adaptation strategies for stereo and multi-view video at the server and client side are studied for UDP and DCCP protocols. In the closed loop rate adaptation, each client estimates some function of the received signal and feeds it back to the transmitter. The transmitter determines an optimized rate for the next transmission based on the received feedback. In the open loop rate adaptation, the transmitter does not use any feedback from the receiver.

In [30]-[31], open-loop rate adaptation strategies for stereo and multi-view video at the server side for UDP and DCCP protocols are studied. In [30], rate adaptation has been achieved by downscaling one of the views using i) spatial sub-sampling, ii) temporal sub-sampling, iii) scaling the quantization step-size, and iv) content-adaptive scaling. In content-adaptive video scaling, the right video is first divided into temporal segments (shots or sub-shots) using well-known temporal segmentation methods. The temporal segments (shots) are
then classified into four categories as determined by their low-level attributes such as the amount of motion and spatial detail within the segment. Shots with high temporal activity (high motion) need to be encoded at full temporal resolution for a smooth viewing experience. On the other hand, if a somewhat stationary shot is being encoded, the temporal sampling rate can be reduced to a lower value without any loss of perceptual quality. Likewise, shots with high spatial detail should not be reduced to lower spatial resolutions, while downsampling can be applied to shots with low spatial detail. Experimental results show that content-adaptive approach for temporal and spatial down-sampling of one of the views yields better compression with higher perceptual quality.

In [31], the video is encoded off-line with a predetermined number of spatial, temporal and SNR scalability layers. Content-aware bit allocation among the views is performed during bitstream extraction by adaptive selection of the number of spatial, temporal, and SNR scalability layers for each group of pictures (GOP) according to the motion and the spatial activity of that GOP. If the GOP has low spatial detail which is shown by the spatial measure, only the base SNR layer is extracted. For high spatial detail, both base SNR and FGS (Fine Granular Scalable) layers are extracted. Similarly, for a low motion GOP, only quarter temporal resolution is extracted, whereas for a high motion GOP half temporal resolution is extracted. The required bitrate reduction is only applied to one of the views. In the experiments, the sequences are encoded off-line with 3 temporal layers per view and with single FGS layer on top of the base quality layer. The results show that adaptive selection of temporal levels and quality layers provides better rate-distortion performance compared to static cases.

In [32] and [33], closed loop strategies have been proposed where rate adaptation is done at the server side by feedback from the user. In [32], a client-driven multi-view video
streaming system is simulated that allows a user to watch 3D video interactively with significantly reduced bandwidth requirements by transmitting a small number of views selected according to his/her head position. 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 requested 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 in order to render the current viewing angle. The proposed system makes use of multi-view coding (MVC) and scalable video coding (SVC) concepts together to obtain improved compression efficiency while providing flexibility in bandwidth allocation to the selected views. Rate-distortion performance of the system has been demonstrated under different conditions. In [33], this idea is extended to a multicast scenario where each view is streamed to a different IP-multicast address. A viewer's client joins appropriate multicast groups to only receive the 3D information relevant to its current viewpoint. The set of selected videos changes in real time as the user’s viewpoint changes. The performance of the approach has been studied through network experiments.

The transmission of another promising new 3D data representation format [35] is assessed by Chang and Girod [34] who developed a rate-distortion optimized scheme for interactive lightfield streaming. In their approach, the lightfield data set is transformed into blocks of wavelet coefficients; each block is then coded as a scalable bitstream and stored at the sender. To render a frame, the receiver issues a request for relevant data. Based on the request, the estimated state of the data already at the receiver, the network characteristics, and the desired transmission rate, the sender performs rate-distortion optimized bit allocation as a convex optimization process, customizing the outgoing packets to minimize the distortion of the frame rendered at the receiver for the given transmission
rate. Experimental results with a statistical network model show that the proposed rate-
distortion optimized scheme reduces the required bitrate by 10% ~ 25% over a heuristic
scheme at the same render quality.

3.3 Error Correction and 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 lead to bit errors. Most network protocols discard packets with bit errors; thus,
translating bit errors into packet losses. Therefore, the wireless channel is the “weakest link”
of future multimedia networks and, hence, requires special attention, especially when
mobility gives rise to fading and error bursts. In particular, joint source and channel coding
techniques have been developed for the efficient transmission of video streams over packet
erasure channels, both in wired and wireless networks [50][51]. Furthermore, error
concealment methods at the decoder must be considered in order to limit the damage,
especially due to temporal error propagation, resulting from unpreventable packet losses.

Common error correction approaches for reliable transmission of monoscopic video
over packet networks include retransmission requests (ARQ) as in [36] or forward error
correction (FEC) as in [37][38][39]. ARQ methods, which require feedback (ACK)
messages that inform the sender about the reliable reception of the data, may be effective to
deal with packet losses if sufficient playout (pre-roll) 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 pre-roll 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. Advanced channel coding techniques for transmission of 3D data have been reported in [40][41][42]. In [41], the transmission of multi-view video encoded streams over packet erasure networks is examined. Macroblock classification into unequally important slice groups is achieved using the Flexible Macroblock Ordering (FMO) tool of H.264/AVC. Systematic LT codes [52] are used for error protection due to their low complexity and advanced performance. The optimal slice grouping and channel rate allocation is determined by an iterative optimization algorithm based on dynamic programming. The optimization procedure starts by determining the channel protection of each frame. Then macroblocks are classified into slice groups and optimal channel protection for this classification is found. The next step is to calculate the expected distortion of allowable neighboring macroblock classifications with the restriction that a single packet can be exchanged between successive groups. The last step includes comparison of the distortion of the ancestor classification and lowest average distortion of all descendant classifications. Based on this comparison either the previous steps are repeated or the algorithm is terminated. Even though it has been shown that the proposed algorithm performs significantly better than multiview coding schemes using one slice group, no results for the computational complexity is given.

Stereoscopic video streaming using Forward Error Correction (FEC) techniques are examined in [42]. Frames are classified according to their contribution to the overall quality, and then used as layers of the video. Since losing I-frame causes large distortions due to motion/disparity compensation and error propagation, I-frames should be protected the most. Among P-frames, left frames are more important since they can be encoded without the help of right frames. According to this prioritization of the frames, three layers
are formed. These three layers of stereoscopic video are used for unequal error protection (UEP). A comparative analysis of RS and systematic LT codes are provided via simulations to observe the optimum packetization and UEP strategies.

When dealing with low bitrate videos, packet losses may lead to the loss of an entire frame of video. Several studies exist in the literature on frame loss concealment algorithms for monoscopic video, but these methods may not be directly applicable to stereoscopic video [58]-[63]. The reconstruction of lost information is principally based on a priori knowledge about the characteristics of the error and the lost data. Many strategies are based on interpolation of the surrounding image data into the lost region. While in a monoscopic scenario interpolation techniques yield satisfactory results, they are not sufficient for example in a stereoscopic scenario because the information on depth is not preserved. Human perception of errors in 3D video data is different than in the 2D case. In [44] it has been shown that even a small degradation in one of the views result in a significant perceptual distortion.

In [43], an error concealment algorithm that fully makes use of the characteristics of a stereoscopic video sequence based on a relativity analysis is proposed. Based on the relativity of prediction modes for right frames, the 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. Experimental results show that the proposed algorithm can restore the lost frame with good quality, and that is efficient for the error concealment of entire lost right frames in a stereoscopic video sequence.

To estimate the capabilities of error concealment for stereoscopic image data, a strategy for concealment of block bursts in independently coded images was studied in [44] and [45] assuming block based video coding and loss of consecutive blocks. To increase
the quality of the reconstructed block, additional information from the corresponding view is utilized. As in a stereoscopic sequence, samples of both views correspond to each other through the 3D geometry of the scene and camera properties, the two views are highly correlated. Due to this high correlation between the views, information about the corresponding region is highly useful for the reconstruction of the lost block. First, corresponding pixel pairs (matches) around the erroneous region are identified using feature matching and principles of epipolar geometry. To reduce the negative effect of outliers, i.e., badly localized matches, robust estimation of the transformation parameters is used [45].

3.4 3D Video Streaming Demonstrations

Several studies were reported in the literature for end-to-end 3D video streaming systems. In [46], a 3DTV prototype system, with real-time acquisition, transmission, and auto-stereoscopic display of dynamic scenes has been presented by MERL. This system is composed of a multi-projector 3D display, an array of cameras and network connected PCs. Multiple video streams are individually encoded and sent over a broadband network to the display. The 3D display shows high-resolution stereoscopic color images for multiple viewpoints without special glasses. This system uses light-field rendering to synthesize views at the correct virtual camera positions.

In [47], a streaming solution which is based on depth-image-based rendering is proposed. An efficient content delivery architecture based on resource sharing in groups of collaborating network hosts is also proposed.

Recently, an end-to-end prototype system for point-to-point streaming of stereoscopic video over UDP was demonstrated at the IST 2006 event [48]. A block-diagram of the
prototype system is shown in Figure 3. Multiple clients have been developed by modifying the VideoLAN client for different 3D displays. The prototype system operates over a LAN with no packet losses. The server employs the protocol stack RTP/UDP/IP, and can serve multiple clients simultaneously. The session description protocol (SDP) is used to ensure interoperability with the clients. Three clients have been implemented for different types of display systems: Client-1 supports the autostereoscopic Sharp 3D laptop, Client-2 supports a monocular display to demonstrate backwards compatibility, and finally Client-3 supports an in-house polarized 3D projection display system that uses a pair of Sharp MB-70X projectors as shown in Figure 4.

Another demonstration that was built on the concept of transmitting “video-plus-depth” (disparity) information over IP was the VIRTUE 3D videoconferencing system [49]. The innovative approach combined 3D video and VR techniques in a mixed reality application, providing immersive telepresence as well as a natural conferee representation in a shared collaboration space. The generated 3D data was encoded using MPEG-4 technologies and streamed over a packet-switched network using RTP/UDP with the payload formats defined for MPEG-4 audio/visual streams.

4. DISCUSSION AND CONCLUSIONS
A comprehensive survey of the state-of-the-art in transport techniques that are potentially applicable to 3DTV transmission has been presented. A particular emphasis is given to packet networks using the Internet Protocol, which plays an integrating role for all media transport. It is understood that 3DTV transmission may create the largest resource demands faced by the network infrastructure up to now. While the transport solutions must address backwards compatibility issues with the existing digital TV standards and infrastructure,
and, hence, can only support a limited set of 3D data representations and rendering technologies, the streaming over IP solutions offer more flexibility to support different 3D displays, 3D data representations, and compression options. As a result, while we reviewed a specific 3D data representation, the “video-plus-depth” representation and image based rendering technology for digital 3DTV broadcast, we aimed to review the general state of the art in streaming protocols, multi-view video compression standards, and rate adaptation, packet loss protection methods, and research directions for streaming 3DTV over IP.

The current and future research issues for 3D TV transmission can be summarized as: i) The choice of the joint transport and coding because the gains obtained in one of them can easily be nullified by the other one, ii) determination of the best rate adaptation method – adaptation refers to adaptation of the rate of each view as well as 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; iii) error resilient video encoding and streaming strategies utilizing the 3D structure.

ACKNOWLEDGMENTS

This work has been supported by European Commission within FP6 under Grant 511568 with the acronym 3DTV.

The authors would like to thank Cagdas Bilen and Anıl Aksay from the Middle East Technical University, Selen Pehlivan, Engin Kurutepe, Burak Gorkemli, and Goktug Gurler from Koç University, and Nukhet Ozbek from Ege University for their help and support.
REFERENCES


FIGURES

Figure 1: Block diagram of the 3DTV over DVB-T demo [53].
Figure 2: Block diagram of a 3D streaming system.
Figure 3: A block-diagram of the end-to-end stereoscopic video streaming test-bed.
Figure 4: Stereoscopic projection display system.