Debono, C. J. and Assuncao, P. A. A. (2015). *Xjenza Online*, 3:183–188.

Xjenza Online - Journal of The Malta Chamber of Scientists
www.xjenza.org
DOI: 10.7423/XJENZA.2015.2.09

Review Article

3D Video Coding and Transmission

C. J. Debono*[1] **and P. A. A. Assuncao**[2]
[1] *Department of Communications and Computer Engineering, University of Malta, Msida, Malta*
[2] *Instituto de Telecomunicacoes, IPLeiria, Portugal*

Abstract. The capture, transmission, and display of 3D content has gained a lot of attention in the last few years. 3D multimedia content is no longer confined to cinema theatres but is being transmitted using stereoscopic video over satellite, shared on Blu-Ray^{TM}disks, or sent over Internet technologies. Stereoscopic displays are needed at the receiving end and the viewer needs to wear special glasses to present the two versions of the video to the human vision system that then generates the 3D illusion. To be more effective and improve the immersive experience, more views are acquired from a larger number of cameras and presented on different displays, such as autostereoscopic and light field displays. These multiple views, combined with depth data, also allow enhanced user experiences and new forms of interaction with the 3D content from virtual viewpoints. This type of audiovisual information is represented by a huge amount of data that needs to be compressed and transmitted over bandwidth-limited channels. Part of the COST Action IC1105 "3D Content Creation, Coding and Transmission over Future Media Networks" (3D-ConTourNet) focuses on this research challenge.

Keywords: 3D video transmission, multi-view video coding, quality of services

1 Introduction

Multimedia communications has been improving over the years, starting from the broadcasting of black and white television to today's ultra high definition colour transmission and stereoscopic video. These improvements, together with the availability of more services and use of different devices to view the content, including mobile equipment, require more and more data to be transmitted, increasingly demanding more bandwidth from the telecommunication networks. Recent surveys (CISCO, 2014) expect that video traffic will reach around 79% of all the consumer generated Internet traffic in 2018.

To date most of the 3D multimedia experiences have been limited to cinema viewing and controlled environments. This is mainly attributed to the high investments needed to develop these environments and bandwidth demands. However, technologies across the whole chain from capture to transmission to displays have been advancing at a high rate and stereoscopic video has become available for home consumption with content transmitted over satellite, Blu-Ray^{TM}, and Internet technologies (Vetro, Tourapis, Müller & Chen, 2011). In general, viewing this type of video requires the use of special glasses to filter the content towards the correct eye of the viewer to obtain the 3D perception. However, the experience of the viewer can be further improved by transmitting more camera views of the same scene and use displays which do not need glasses. If depth data is added to the multi-view stream, virtual views can be generated using Depth-Image-Based Rendering (DIBR) at the display allowing the user to determine a personal viewing angle, known as Free-viewpoint TeleVision (FTV) (Ho & Oh, 2007). All the data generated has to generally pass through a limited bandwidth channel and therefore adequate coding must be performed.

Transmission of 3D and immersive multimedia services and applications over heterogeneous networking technologies includes broadcasting channels, wideband backbone links, bandwidth-constrained wireless networks, among others (Lykourgiotis et al., 2014). At transport level, three main system layers have been considered in the recent past, as the most adequate for 3D media delivery: MPEG-2 systems, Real-time Transport Protocol (RTP) and ISO base media file format (Schierl & Narasimhan, 2011). However, these legacy technologies are now facing new challenges as a result of fast

**Correspondence to*: C. J. Debono (carl.debono@um.edu.mt) evolution towards future media networks. For instance, 3D multimedia streaming requires flexible adaptation mechanisms capable of delivering subsets of 3D data according to network constraints or users' preferences and robust coding and transmission schemes are necessary to cope with error-prone channels and dynamic networking such as Peer-to-Peer (P2P) (Gurler & Tekalp, 2013). In this context, the challenge of achieving an acceptable level of Quality of Experience (QoE) has been evolving from a technological perspective (Cubero et al., 2012) by including an increasing number of human factors (Taewan, Sanghoon, Bovik & Jiwoo, 2014) and acceptance in different usage scenarios (Wei & Tjondronegoro, 2014).

The COST Action IC1105 "3D Content Creation, Coding and Transmission over Future Media Networks" (3D-ConTourNet) aims at bringing together researchers from all the spectrum of the 3D technology chain to discuss current trends and research problems. It also provides, through dissemination of findings, information for stakeholders on the state-of-the-art technology and services. This article deals with the 3D video coding and transmission part of this COST Action.

The paper is divided into five sections. The next section gives information related to the available 3D video formats. Section 3 deals with the coding of 3D videos while section 4 focuses on the transmission of the 3D content. At the end, a conclusion is given.

2 3D Video Formats

2.1 Stereoscopic Representations

The most cost effective way to transmit 3D videos is using stereoscopic representation. This only needs to transmit two views, one intended for the left human eye and the other one for the right eye. The transmission is done sequentially. These two views can be transmitted at a lower resolution in the same space dedicated for a high definition television frame and positioned either side-by-side or in top-and-bottom fashion. In (Zhang, Wang, Zhou, Wang & Gao, 2012), the authors propose the transmission of one single video plus the depth information. In this case the second view is generated at the display using DIBR techniques. In all cases, the video can either be viewed using a normal television by decoding one of the views or in 3D using any type of stereoscopic display.

2.2 Model-based Representation

This approach considers the video as a sequence of 2D projections of the scene. It uses closed meshes, such as triangle meshes (Theobalt, Ziegler, Magnor & Seidel, 2004), to represent generic models. Adaptation through scaling of the segments and deformation of surfaces is then applied to better represent the objects in the scene. The input streams are mapped into texture space transforming the 3D model into 2D. The texture maps of each camera view are encoded at every time stamp using 4D-SPIHT (Theobalt et al., 2004; Ziegler, Lensch, Magnor & Seidel, 2004) or similar methods. Semantic coding can also be used for model-based representations, where detailed 3D models are assumed to be available (Kaneko, Koike & Hatori, 1991). The drawback of semantic coding schemes is that it can only be used for video having known objects.

2.3 Point-sample Representation

2D video can be mapped to 3D video polygon representation using point sample methods. Such a technique is applied in Würmlin, Lamboray and Gross (2004), where a differential update technique uses the spatio-temporal coherence of the scene captured by multiple cameras. Operators are applied on the 3D video fragments to compensate for the changes in the input and are transmitted with the video stream. The 3D video fragment is defined using a surface normal vector and a colour value. This method also needs the transmission of camera parameters and identifiers together with the coordinates of the 2D pixels.

2.4 Multi-view Video Representation

This representation considers the capturing of a scene from multiple cameras placed at different angles. This generates a huge amount of data proportional to the number of cameras capturing the scene. To reduce this huge data and provide for better scalability Multi-view Video Coding (MVC) can be used (Vetro, Tourapis et al., 2011). Furthermore, the Multi-View (MV) representation is an extension of the High Efficiency Video Coding (HEVC) standard. An overview of HEVC can be found in Sullivan, Ohm, Han and Weigand (2012).

2.5 Multi-view Video Plus Depth Representation

The Multi-view Video plus Depth (MVD) format includes the transmission of depth maps with the texture video. The depth information adds geometrical information that helps in achieving better encoding and view reconstruction at the displays. This format supports the use of less views, as intermediate views can be constructed at the display, ideal for wide angle and autostereoscopic displays (Vetro, Yea & Smolic, 2008). This format will probably be the main format for transmission of 3D videos for HEVC coded content.

3 3D Video Coding

3.1 Stereoscopic 3D Video Coding

The current way of transmitting 3D video is using stereoscopic technology. This mainly involves the capture of the scene using two cameras similar to the human vision system. These sequences are then separately presented to the left and right eye of the viewer. In this case, the video is either coded by means of simulcasting, where each view is compressed using H.264/AVC or HEVC, or by placing the two images, one from each stream, in a single high definition frame. In the latter, known as frame compatible format, the resolution is decreased, but is an efficient way of coding since the bandwidth required is similar to the single-view transmission.

3.2 Multi-view Video Coding

This coding scheme allows for a more efficient way to transmit multiple views compared to simulcasting each individual view. This is done by exploiting the redundancies available between camera views. Thus, H.264/MVC and MV-HEVC use spatial, temporal and inter-view predictions for compression. An overview of the MVC extension to the H.264/AVC can be obtained from Vetro, Weigand and Sullivan (2011). The multiview video can be coded using different structures; the most commonly used in literature being the low latency structure and the hierarchical bi-prediction structure. The low latency structure, shown in Figure 1 for 3 views, uses only previously encoded blocks for its predictions in the time axis. Bi-prediction is still applied in between views, but this is done at the same time instant and therefore the decoding does not need to wait for future frames and needs a smaller buffer. On the other hand, the hierarchical bi-prediction structure uses future frames in the encoding as shown in Figure 2. This implies that a larger buffer is needed and the decoding has to wait for the whole group of pictures to start decoding. The advantage of this structure is that it provides a better coding efficiency and therefore less data needs to be transmitted.

3.3 Video-plus-depth Coding

Even though current multi-view encoders can provide very high compression ratios, transmission of the multiple views still needs huge bandwidths. However, to satisfy the need of a high number of views to generate an immersive 3D experience, a lower number of views can be transmitted together with the depth data. The missing views can then be interpolated from the transmitted views and depth data. This can be done using a synthesis tool such as DIBR with the geometry data found in the depth maps. The texture and depth videos can be encoded using the 3D video coding extensions discussed above and then multiplexed on the same bit stream. Otherwise, they can be jointly encoded such that redundancies inherent in the texture and the depth videos can be exploited for further coding efficiencies. An example of such a coding method is found in Müller et al. (2013) and is now an extension of the HEVC standard. The HEVC extension for 3D (3D-HEVC) improves the coding efficiency by exploiting joint coding of texture images and the corresponding depth maps (Tech et al., 2015).

3.4 Research Trends in Video Coding

Although a lot of work has been done in 3D video coding, more research is still needed to provide for fast, more efficient and cheap encoders. This can be done by reducing further the redundancies in the videos, applying more parallel algorithms, simplifying processes, catering for scalability due to the different display resolutions, applying more prediction schemes, and other ideas. The 3D-ConTourNet COST Action members are discussing these issues and are working to address these in order to get better 3D video transmission closer to the market.

4 3D Video Transmission

Three-dimensional video delivery is mainly accomplished over broadcasting networks and the Internet, where the IP protocol prevails in flexible platforms providing IPTV and multi-view video streaming. In broadcasting and IPTV services, 3D video streams are encapsulated in MPEG-2 Transport Streams (TS) and in IPTV TS that are further packetised into the Real-Time Protocol (RTP) and User Datagram Protocol (UDP), which provide the necessary support for packet-based transmission and improved QoS. Since TS and RTP provide similar functionalities at systems level, this type of packetisation introduces some unnecessary overhead, which is particularly relevant in multi-view video due to the increased amount of coded data that is generated. In the case of Internet delivery, HTTP adaptive streaming is becoming more relevant, since it allows low complexity servers by shifting adaptation functions to the clients, while also providing flexible support for different types of scalability under user control, either in rate, resolution and/or view selection, besides improved resilience to cope with network bandwidth fluctuations.

Since the term “3D video” does not always correspond to a unique format of visual information, the actual transport protocols and networking topologies might be different to better match the compressed streams. For instance, enabling multi-view 3D video services may require more bandwidth than available if all views of all video programs are simultaneously sent through existing networks. However, as mentioned above, if DIBR is used, a significant amount of bandwidth may be saved, because the same performance and quality might be kept by simply reconstructing some non-transmitted views at the receivers, from their nearby left and right views. Such possibility is enabled by the MVD format, which allows reconstruction of many virtual views from just few of them actually transmitted through the communications network.

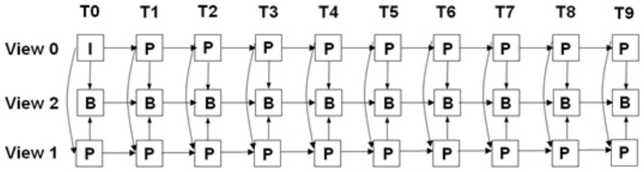

Figure 1: The low latency MVC structure. I represents an Intra coded frame, P represents a predicted frame, and B represents a bi-predicted frame.

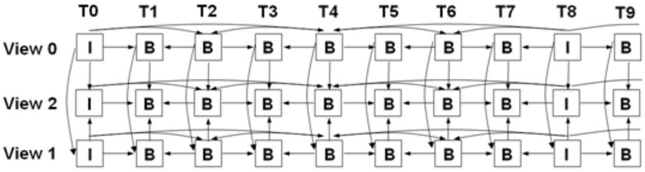

Figure 2: The hierarchical bi-prediction MVC structure.

Interactive streaming also poses specific transmission requirements in 3D multi-view video. In non-interactive services, multiple views can be sent through a single multicast session shared simultaneously by all clients, while interactivity requires each view to be encoded and transmitted separately. This allows users to freely switch between views by subscribing to different multicast channels. Multipath networks, such as P2P, can also provide the necessary support for interactive multi-view 3D video streaming by assigning the root of different dissemination trees to different views, which in turn can even be hosted in different servers (Chakareski, 2013). In the case of mobile environments, there are quite diverse networking technologies that might be used to provide immersive experiences to users through multi-view video, but the huge amount of visual data to be processed and the limited battery-life of portable devices is pushing towards cloud-assisted streaming scenarios to enable deployment of large-scale systems where computational power might be provided at the expense of bandwidth (Guan & Melodia, 2014).

Figure 3 summarises the main protocol layers used in 3D video broadcasting and streaming services. In the left side, the traditional DVB, including satellite, terrestrial and cable is shown. Basically, the Packetised Elementary Streams (PES) are encapsulated in TS before transmission over the DVB network. An extension of the classic 2D MPEG-2 Systems was defined to support multi-view video, where different views may be combined in different PES to provide multiple decoding options. The right side of Figure 3 shows a typical case of IP broadcasting and/or streaming of 3D multi-view video. Multi-Protocol Encapsulation (MPE) is used to increase error robustness in wireless transmission (e.g. DVB-SH), while Datagram Congestion Control Protocol (DCCP) may be used over Internet. In this case, MPEG-2 TS encapsulation may not be necessary. In the case of multi-view streaming using RTP, either single-session or multisession may be used to enable a single or multiple RTP flows for transport of each view. The underlying communication infrastructure can be quite diverse (e.g. cable, DVB, LTE). Like in classic 2D video transmission, dynamic network conditions fluctuation, such as available bandwidth, transmission errors, congestion, jitter, delay and link failures are the most significant factors affecting delivery of 3D video across networks and ultimately the QoE.

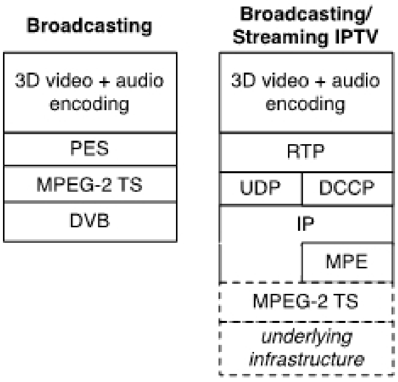
Broadcasting
3D video + audio encoding
PES
MPEG-2 TS
DVB
Broadcasting/ Streaming IPTV
3D video + audio encoding
RTP
UDP
DCCP
IP
MPE
MPEG-2 TS
underlying infrastructure

Figure 3: Generic protocol stack for 3D video services.

However, the increased amount of coded data and high sensitivity of 3D video to transmission errors requires robust coding techniques and efficient concealment methods at the decoding side because the perceived QoE in 3D video is known to be more sensitive to a wider variety of quality factors than in classic 2D (Hewage, Worrall, Dogan, Villette & Kondoz, 2009). Two robust coding techniques suitable for such purposes are scalable 3D video coding and Multiple Description Coding (MDC). In both of them several streams are produced and transmission losses may only affect a subset of them. In the case of scalable 3D video coding, there is one main independent stream (base layer) that should be better protected against transmission errors and losses while the other dependent streams, or layers, can be discarded at the cost of some graceful degradation in quality. In MDC, each stream is independently decodable and can be sent over different paths to avoid simultaneous loss. This is particularly efficient in multipath transmission over P2P streaming networks (Ramzan, Park & Izquierdo, 2012).

4.1 Research Trends in 3D Multimedia Transmission

Current research trends in 3D and multi-view transmission span over several key interdisciplinary elements, which aim at the common goal of delivering an acceptable QoE to end-users. Heterogeneous networks comprising hybrid technologies with quite diverse characteristics and the increasing dynamic nature of 3D multimedia consumption (e.g. mobile, stereo, multi-view, interactive) pose challenging research problems with regard to robust coding, network support for stream adaptation, scalability and immersive interactive services, packet loss and error concealment. Hybrid networks and multipath transmission in P2P is driving research on MDC of 3D multimedia combined scalability and P2P protocols. While MDC is certainly better for coping with dynamic multipath networks, scalability might offer the most efficient solution for pure bandwidth constraints. Network-adaptation by processing multiple streams in active peer nodes is also under research to ensure flexibility and acceptable QoE in heterogeneous networks with different dynamic constraints and clients requiring different sub-sets of 3D multimedia content. The problem of accurate monitoring of QoE along the delivery path has been an important focus of the research community, but no general solution has yet been devised, so much more research is expected in the near future in this field. Synchronisation of the video streams across the different network paths is another open problem which can lead to frequent re-buffering and jittering artifacts. Overall, joint optimisation of coding and networking parameters is seen as the key to accomplish high levels of QoE, validated through widely accepted models.

5 Conclusion

An overview of the most important elements of 3D video coding and transmission was presented with emphasis on the technological elements that have open issues for further research and development. 3D video formats have evolved from simple stereo video to multi-view-plus-depth, which leads to a huge amount of coded data and multiple dependent streams. The need for robust transmission over future media networks using multiple links, providing in-network adaptation functions and coping with different client requirements was also highlighted as necessary for achieving and acceptable QoE. As an active multidisciplinary field of research, several promising directions to carry out further relevant investigations were also pointed out.

References

Chakareski, J. (2013). Adaptive mutiview video streaming: Challenges and Opportunities. *IEEE Communications Magazine*, *51*(5), 94–100.

CISCO. (2014). *Cisco visual networking index: forecast and methodology, 2013-2018.*

Cubero, J. M., Gutierrez, J., Perez, P., Estalayo, E., Cabrera, J., Jaureguizar, F. & Garcia, N. (2012). Providing 3D video services: The challenge from 2D to 3DTV quality of experience. *Bell Labs Technical Journal*, *16*(4), 115–134.

Guan, Z. & Melodia, T. (2014). Cloud-Assisted Smart Camera Networks for Energy-Efficient 3D Video Streaming. *Computer*, *47*(5), 60–66.

Gurler, C. G. & Tekalp, M. (2013). Peer-to-peer system design for adaptive 3D video streaming. *IEEE Communications Magazine*, *51*(5), 108–114.

Hewage, C., Worrall, S., Dogan, S., Villette, S. & Kondoz, A. (2009). Quality Evaluation of Color Plus Depth Map-Based Stereoscopic Video. *IEEE Journal of Selected Topics in Signal Processing*, *3*(2), 304–318.

Ho, Y. S. & Oh, K. J. (2007). Overview of multi-view video coding.

Kaneko, M., Koike, A. & Hatori, Y. (1991). Coding of a facial image sequence based on a 3D model of the head and motion detection. *Journal of Visual Communications and Image Representation*, *2*(1), 39–54.

Lykourgiotis, A., Birkos, K., Dagiuklas, T., Ekmekcioglu, E., Dogan, S., Yildiz, Y., … Kotsopoulos, S. (2014). Hybrid broadcast and broadband networks convergence for immersive TV applications. *IEEE Wireless Communications*, *21*(3), 62–69.

Müller, K., Schwarz, H., Marpe, D., Bartnik, C., Bosse, S., Brust, H., … Wiegand, T. (2013). 3D High-Efficiency Video Coding for Multi-view Video and Depth. *IEEE Transactions on Image Processing*, *22*(9), 3366–3378.

Ramzan, N., Park, H. & Izquierdo, E. (2012). Video streaming over P2P networks: Challenges and opportunities. *Signal Processing: Image Communication*, *27*(5), 401–411.

Schierl, T. & Narasimhan, S. (2011). Transport and Storage Systems for 3-D Video Using MPEG-2 Systems, RTP, and ISO File Format. *Proceedings of the IEEE*, *99*(4), 671–683.

Sullivan, G. J., Ohm, J.-R., Han, W.-J. & Weigand, T. (2012). Overview of the high efficiency video coding (HEVC) standard. *IEEE Transactions on Circuits and Systems for Video Technology*, *22*(12), 1649–1668.

Taewan, K., Sanghoon, L., Bovik, A. C. & Jiwoo, K. (2014). Multimodal Interactive Continuous Scoring of Subjective 3D Video Quality of Experience. *IEEE Transactions on Multimedia*, *16*(2), 387–402.

Tech, G., Chen, Y., Muller, K., Ohm, J.-R., Vetro, A. & Wang, Y.-K. (2015). Overview of the Multiview and 3D Extensions of High Efficiency Video Coding. *Circuits and Systems for Video Technology, IEEE Transactions on*, *PP*(99), 1.

Theobalt, C., Ziegler, G., Magnor, M. & Seidel, H. P. (2004). Model-based free-viewpoint video acquisition, rendering and encoding.

Vetro, A., Tourapis, A., Müller, K. & Chen, T. (2011). 3D-TV content storage and transmission. *IEEE Transactions on Broadcasting, Special Issue on 3D-TV Horizon: Contents, Systems, and Visual Perception*, *57*(2), 384–394.

Vetro, A., Weigand, T. & Sullivan, G. J. (2011). Overview of the stereo and multiview video coding extensions of the H.264/MPEG-4 AVC standard. *Proceedings of the IEEE*, *99*(4), 626–642.

Vetro, A., Yea, S. & Smolic, A. (2008). Towards a 3D video format for auto-stereoscopic displays.

Wei, S. & Tjondronegoro, D. W. (2014). Acceptability-Based QoE Models for Mobile Video. *IEEE Transactions on Multimedia*, *16*(3), 738–750.

Würmlin, S., Lamboray, E. & Gross, M. (2004). 3D video fragments: dynamic point samples for real-time freeviewpoint video. *Computers & Graphics*, *28*(1), 3–14.

Zhang, Z., Wang, R., Zhou, C., Wang, Y. & Gao, W. (2012). A compact stereoscopic video representation for 3D video generation and coding.

Ziegler, G., Lensch, H. P. A., Magnor, M. & Seidel, H. P. (2004). Multi-video compression in texture space using 4D SPIHT.