# 3D VIDEO AND FREE VIEWPOINT VIDEO – TECHNOLOGIES, APPLICATIONS AND MPEG STANDARDS

*Aljoscha Smolic, Karsten Mueller, Philipp Merkle, Christoph Fehn, Peter Kauff, Peter Eisert, and Thomas Wiegand*

Fraunhofer Institute for Telecommunications/Heinrich-Hertz-Institut
Einsteinufer 37, 10587 Berlin, Germany
Aljoscha.Smolic@fraunhofer.hhi.de

**ABSTRACT**

An overview of 3D and free viewpoint video is given in this paper with special focus on related standardization activities in MPEG. Free viewpoint video allows the user to freely navigate within real world visual scenes, as known from virtual worlds in computer graphics. Examples are shown, highlighting standards conform realization using MPEG-4. Then the principles of 3D video are introduced providing the user with a 3D depth impression of the observed scene. Example systems are described again focusing on their realization based on MPEG-4. Finally multi-view video coding is described as a key component for 3D and free viewpoint video systems. The conclusion is that the necessary technology including standard media formats for 3D and free viewpoint is available or will be available in the near future, and that there is a clear demand from industry and user side for such applications. 3DTV at home and free viewpoint video on DVD will be available soon, and will create huge new markets.

## 1. INTRODUCTION

Digital media have influenced and changed modern society over the last 2 decades significantly. Media are more and more produced, processed, stored, and transmitted in digital formats with digital equipment. Applications, terminals and content are merging more and more. We can watch TV with our mobile phones, surf the web with the TV set, and modern home PCs are powerful multimedia workstations capable of more or less everything. On a DVD we do not buy only the movie with video and audio but also a vast amount of supplementary information. New media formats integrate all types of media, such as video, audio, computer graphics, text, images, etc. into a single file or stream format. Some of these formats further enable user interactivity with the content, i.e. the user can do something with the media in addition to just passively consuming. An important factor for this success story is the availability of international standards for digital media formats. They provide interoperability between different systems while still allowing for competition among equipment and service providers. ISO MPEG is one of the international standardization bodies that play an important role in the digital media market.

Recent research and convergence of technologies from computer graphics, computer vision, multimedia and related fields enabled also the development of new types of media, such as 3D video and free viewpoint video that expand the user's sensation far beyond what is offered by traditional media. The first offers a 3D depth impression of the observed scenery (also referred to as stereo, note that the term 3D may have different meanings in the context of this paper), while the second allows for interactive selection of viewpoint and direction within a certain operating range as known from computer graphics. Some application scenarios may be based on proprietary systems, as for instance already employed for (post-) production of movies and TV content. On the other hand there are also application scenarios that require interoperable systems, such as 3DTV broadcast or free viewpoint video on DVD. This may open huge consumer markets for 3D displays, set-top boxes, media, content, DVDs, HD-DVDs, BRDs, etc., along with the corresponding equipment for production, transmission, etc.

Therefore the MPEG committee has been investigating the needs for standardization in the area of 3D and free viewpoint video in a group called 3DAV (for 3D audio-visual) [15] in recent years. Thus far, the committee has provided an overview of relevant technologies and has shown that a number of these technologies are already be supported by existing standards such as MPEG-4. For the missing elements, new standardization activities have been launched. Some activities have already been completed, such as the new tools for the efficient and high-quality representation of 3D video objects, which have been adopted as part of the MPEG-4 Animation Framework eXtension (AFX) specification [3]. Other more challenging activities are still ongoing, such as the specification of a new standard for multi-view video coding with associated camera parameters, which will enable 3D and free viewpoint video systems as the final goal.

This paper gives an overview of the applications "free viewpoint video" and "3D video" in sections 2 and 3, highlighting the related standardization activities in MPEG. Section 4 addresses a related upcoming standard for compression of multi-view video, and finally section 5 concludes the paper and gives an outlook to the future in this area.

## 2. FREE VIEWPOINT VIDEO

Free viewpoint video (FVV) offers the same functionality that is known from 3D computer graphics. The user can choose an own viewpoint and viewing direction within a visual scene, meaning interactive free navigation. In contrast to pure computer graphics applications, FVV targets real world scenes as captured by real cameras. This is interesting for user applications (DVD of an opera/concert where the user can freely chose the viewpoint) as well as for (post-) production. Systems for the latter are already being used (e.g. for sports, movies, EyeVision, Matrix-effects).

The complete processing chain of such systems can be divided into the parts of acquisition/capturing, processing, scene representation, coding, transmission/streaming/storage, interactive rendering and 3D displays. The design has to take into account all parts, since there are strong interrelations between all of them. For instance, an interactive display that requires random access to 3D data will affect the performance of a coding scheme that is based on data prediction.

Different technologies can be used for acquisition, processing, representation, and rendering, but all make use of multiple views of the same visual scene [16], as illustrated in Fig. 1. The multiple camera signals are processed and transformed into a specific scene representation format that allows for rendering of virtual intermediate views, i.e. in between the real existing camera positions. With that the user can navigate the scene freely, meaning choosing an individual viewpoint and viewing direction. The camera setting and density imposes practical limitations to navigation and quality of rendered views at a certain virtual position. Therefore there is a classical trade-off to consider between costs (for equipment, cameras, processors, etc.) and benefits (navigation range, quality of virtual views).

Fig. 1 - Fig. 2 illustrate an example of FVV [12]. Here a 3D object is reconstructed from multiple views and represented by its 3D geometry (mesh model) and associated appearance (video textures). The 3D video object (3DVO) is dynamic (moving and deforming over time) and provides the same functionality as conventional computer graphics models (free navigation, integration in scenes) but in contrast represents a real world object.

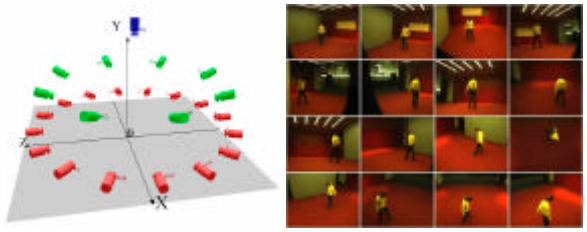

Figure 1: Multi-camera setup for 3DVO acquisition and captured multi-view video.

Fig. 1 illustrates multi-view acquisition for 3DVOs in a relatively sparse dome type setting. Accurate camera calibration information is essential to establish the 2D-3D correspondence between the image pixels and the 3D world. In most cases this is estimated before capturing using a pre-defined calibration grid and some state-of-the-art algorithms. The first multi-view signal processing step consists in segmentation of the objects of interest, i.e. those that shall be reconstructed in 3D. Although a huge effort has been put into this, segmentation is still an error prone task. It conceptually remains an estimation that can theoretically only be solved up to a residual probability. However, by proper setting of the environment this residual probability can be minimized. For instance in some application scenarios a blue-box studio environment may be used.

Having estimated the object's silhouette in each input image the 3D shape can be reconstructed using a shape-from-silhouette algorithm. The result is a voxel model of the object's 3D volume, as shown in Fig. 2 left. In a next step the object's surface can be extracted using a marching-cubes algorithm and represented as classical 3D mesh, as shown in Fig. 2 middle. This is to benefit from available graphics hardware and software APIs that are highly optimized for processing this type of data. An additional smoothing step may help to regularize the estimated 3D mesh. Finally color and texture can be projected from the available camera views onto the 3D mesh. The result is a 3DVO as shown in Fig. 2 right, a reconstruction of a real world object in 3D, represented as 3D mesh with associated textures. Such a 3DVO can be integrated into real or virtual scenes and viewed interactively from any direction.

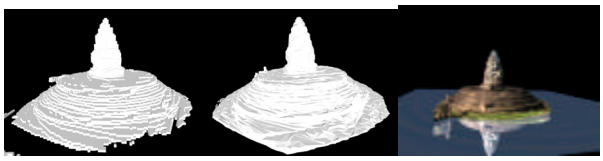

Figure 2: Reconstructed voxel model, 3D mesh model, and final 3DVO with associated textures.

Since 3D objects may appear very different from different directions, depending on light sources and reflectance properties, static texturing may lead to poor rendering results. This can be overcome by view-dependent texture mapping since multiple views of the object are available. The available textures from the cameras are weighted depending on the distance to the virtual viewpoint and blended over the object. Closer cameras contribute more than more distantly located cameras. Fig. 3 illustrates a virtual camera fly around a dynamic 3DVO represented as dynamic 3D mesh with view-dependent texture mapping. The images show virtual rendered views at 3 different times from 3 different viewpoints.

Figure 3: Virtual camera fly, rendered views at 3 different times from 3 different virtual viewpoints.

This representation for 3DVOs uses classical 3D mesh models for geometry and associated video for textures. Such a representation is already supported by MPEG-4. It is possible to create standard compliant 3DVOs that can be decoded and rendered with any appropriate MPEG-4 player. However, view-dependent texture mapping as described above is not supported in the first versions of MPEG-4. Therefore this tool was added in an update of the computer graphics part of MPEG-4 called Animation Framework eXtension (AFX) [3], as an outcome of the work in the 3DAV group.

Further, a 3DVO describes motion and deformation of a natural object over time. Therefore it is possible to constrain the reconstruction process in a way that it produces a sequence of time consistent 3D meshes. This means 3D meshes with constant connectivity over time, only the 3D position of the vertices is changing. It has been shown that predictive approaches outperform available MPEG tools for compression of such time consistent 3D meshes [11]. Therefore these algorithms are under investigation in MPEG which may lead to a further extension of the AFX standard.

An alternative to classical 3D meshes for 3D rendering is the usage of 3D point clouds or video fragments [14]. This representation uses unorganized point clouds in 3D, i.e. points with 3D coordinates but without connectivity. Additional attributes as color or normals are assigned to the points. Such a point cloud can be rendered with regard to any virtual viewpoint of the scene, by projecting the points onto the screen (splatting). The absence of connectivity is a big advantage over classical 3D meshes, and the representation itself can be regarded as a natural extension of 2D video into 3D. This makes it especially interesting for FVV, which is a reconstruction from multiple natural video signals into 3D. Compression of such data has been investigated in [10]. Point cloud representations in the way described here are not supported sufficiently in the first versions of MPEG-4. Therefore they have also been included in the new AFX part [3].

## 3. 3D VIDEO

The second new functionality provided by these new technologies is a 3D depth impression of the observed scene. In fact, this stereo functionality is not new. Extending visual sensation to the 3rd dimension has been investigated for a long time. Commercial systems (e.g. in IMAX theatres, medicine) are available. However, acceptance for large user mass markets (3DTV at home, DVDs, etc.) has not been reached yet. This may be overcome due to recent developments of 3D displays (where glasses are no longer needed) and advanced 3D rendering that supports head motion parallax viewing [16].

In principle there is no clear distinction between 3D video and FVV as described in the previous section. This classification has more historical reasons and is more related to the main focus of the involved researchers (more on free navigation or more on 3D depth impression). 3D rendering means creating 2 views, one for each eye, which if perceived by a human will create a depth impression. This is possible in principle with any of the FVV approaches described in the previous section. There are several types of 3D displays available, with and without glasses, and therefore also different types of specific 3D rendering algorithms.

Fig. 4 illustrates depth-based stereo rendering and shows an autostereoscopic 3D display, where no glasses are necessary to get a 3D impression. A video signal and a per pixel depth map is transmitted to the user. From the video and depth 2 virtual views are rendered, one slightly right and one slightly left from the original camera position, corresponding to a stereo pair for human observation [1]. These views are displayed simultaneously on the autostereoscopic 3D display, and the user perceives a 3D depth impression of the scene. The 3D display of Fraunhofer HHI shown in Fig. 4 allows for user tracking with built in camera sensors [13]. The user's eye positions are automatically tracked by the system. This is used to automatically adjust the 3D impression. Further, it supports head motion parallax viewing. Depending on the motion of the user, the rendered views are adjusted in real-time to the actual eye position. With that occlusion and disocclusion effects are supported within a limited operating range corresponding to the motion of a user sitting on a chair in front of the screen. Since rendering is done at the receiver, the depth impression can be adjusted individually by the user in the same way it is done with color or brightness using a classical TV set [1].

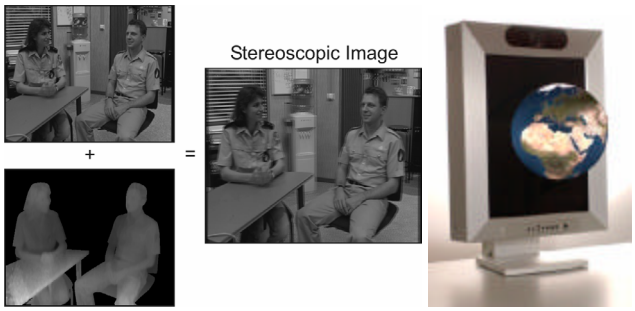


Figure 4: Depth-based stereo rendering and autostereoscopic 3D display (no glasses required).

The full 3DTV processing chain has been realized and demonstrated in the European ATTEST project [1]. The result is a backward compatible (to classical DVB) approach for 3DTV. In this context also compression of depth data has been investigated. It has been found that depth data can be very efficiently compressed using standard video codecs such as H.264/AVC [9]. From standards point of view the realization of the ATTEST concept for 3DTV only requires minor additions on the Systems level of MPEG-4. These are currently under investigation and may provide an interoperable solution for 3DTV broadcast in the very near future.

This concept for depth based 3D rendering is easily extended to N views, as shown in [17]. Depending on the user position a simple switching to the nearest original view with depth (or pair of views with disparity/depth) is possible. This extends the navigation range in front of the screen with the number of cameras used. For some application scenarios such as 3DTV broadcast this implies compression and transmission of multi-view video, which is an ongoing work item in MPEG as described below.

## 4. MULTI-VIEW VIDEO CODING

A common element of many systems described above is the use of multiple views of the same scene that have to be transmitted to the user. The straight-forward solution for this would be to encode all the video signals independently using a state-of-the-art video codec such as H.264/AVC [9]. However, in a "Call for Evidence" [4] it has been shown that specific multi-view video coding (MVC) algorithms give significantly better results compared to the simple H.264/AVC simulcast solution [6]. Improvements of more than 2 dB were reported for the same bitrate. The basic idea in all of the submitted proposals is to exploit spatial and temporal redundancy for compression. Since all cameras capture the same scene from different viewpoints spatial redundancy can be expected. A basic structure for spatio-temporal prediction including spatio-temporal B-pictures is shown in Fig. 5. Images are not only predicted from temporally preceding images but also from corresponding images in adjacent views.

Besides such spatio-temporal prediction structures that can be much more complex than the example in Fig. 5 (see [5] for details), also specific prediction tools have been proposed that can be combined with any prediction structure. This includes for instance illumination compensation, spatio-temporal direct mode, disparity/motion vector prediction, and view interpolation (see [5] for details). The latter describes prediction by warping of neighboring images using camera parameters. Camera parameters need to be

available at the decoder anyway for any application using MVC (FVV, 3D video). Therefore transmission of camera parameters (extrinsic and intrinsic) is a basic requirement for MVC. Using these parameters for prediction does not imply any overhead for transmission.

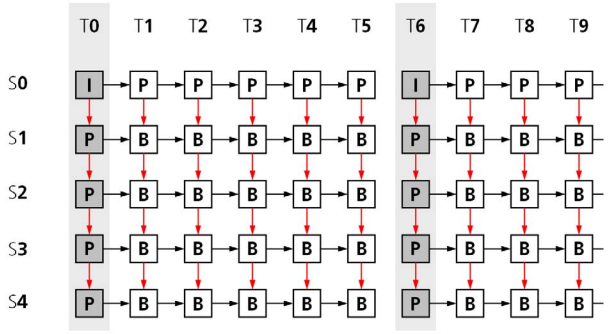


Figure 5: Spatio-temporal prediction structure for MVC, S indicates cameras, T indicates time.

Since further a "Call for Comments" [2] has shown that there is large interest from industry in systems and applications described above, MPEG decided to issue a "Call for Proposals" [8] for MVC technology along with related requirements [7]. The responses to the "Call for Proposals" have been evaluated in January 2006. All submitted proposals were extensions of H.264/AVC. The best performing proposal used hierarchical B-pictures in temporal and inter-view dimension [18]. Therefore it was decided by MPEG to make MVC an amendment to H.264/AVC which is scheduled to be available in early 2008.

## 5. CONCLUSIONS AND OUTLOOK

This paper gave an overview of technologies for 3D and free viewpoint video with a special focus on the related standardization activities in MPEG. It has shown that the technological basis for a variety of new multimedia applications is readily available and under development, including the necessary standard media formats. A lot of research has been done in this area but also more and more products such as 3D displays become available. The interest in industry is rapidly growing as well as user attention to these new types of applications. Users become more and more familiar with interactive 3D applications and systems. Computers, consumer electronics, telecommunications and related technologies converge more and more. Therefore it can be foreseen that applications and services like 3DTV at home or free viewpoint video on DVD (watch your favourite concert from your favourite viewpoint) will become reality in the near future. With that, huge markets for consumer equipment, production equipment, media, content, etc. will develop.

## 6. ACKNOWLEDGEMENTS

The authors would like to thank Stephan Wuermlin and colleagues from the Computer Graphics Lab of ETH Zurich for the help providing data and a figure for this paper.
This work is supported by EC within FP6 under Grant 511568 with the acronym 3DTV.

## 7. REFERENCES

[1] C. Fehn, P. Kauff, M. Op de Beeck, F. Ernst, W. Ijsselsteijn, M. Pollefeys, L. Vangool, E. Ofek, and I. Sexton, "An Evolutionary and Optimised Approach on 3D-TV", Proc. of IBC 2002, Int. Broadcast Convention, Amsterdam, Netherlands, Sept. 2002.
[2] ISO/IEC JTC1/SC29/WG11, "Call for Comments on 3DAV", Doc. N6051, Gold Coast, Australia, October 2003.
[3] ISO/IEC JTC1/SC29/WG11, "ISO/IEC 14496-16/PDAM1", Doc. N6544, Redmont, WA, USA, July 2004.
[4] ISO/IEC JTC1/SC29/WG11, "Call for Evidence on Multi-View Video Coding", Doc. N6720, Palma de Mallorca, Spain, October 2004.
[5] ISO/IEC JTC1/SC29/WG11, "Survey of Algorithms used for Multi-view Video Coding (MVC)", Doc. N6909, Hong Kong, China, January 2005.
[6] ISO/IEC JTC1/SC29/WG11, "Report of the subjective quality evaluation for MVC Call for Evidence", Doc. N6999, Hong Kong, China, January 2005.
[7] ISO/IEC JTC1/SC29/WG11, "Requirements on Multi-view Video Coding v.2", Doc. N7282, Poznan, Poland, July 2005.
[8] ISO/IEC JTC1/SC29/WG11, "Call for Proposals on Multi-view Video Coding", Doc. N7327, Poznan, Poland, July 2005.
[9] ITU-T Recommendation H.264 & ISO/IEC 14496-10 AVC, "Advanced Video Coding for Generic Audio-Visual Services", 2003.
[10] E. Lamboray, S. Würmlin, M. Waschbüch, M. Gross, and H. Pfister, "Unconstrained Free-Viewpoint Video Coding", Proceedings of the IEEE International Conference on Image Processing (ICIP) 2004, Singapore, October 24-27, 2004.
[11] K. Mueller, A. Smolic, M. Kautzner, P. Eisert, and T. Wiegand, "Predictive Compression of Dynamic 3D Meshes", Proceedings of the IEEE International Conference on Image Processing (ICIP) 2005, Genova, Italy, September 11-14, 2005.
[12] K. Mueller, A. Smolic, P. Merkle, M. Kautzner, and T. Wiegand, "Coding of 3D Meshes and Video Textures for 3D Video Objects", Proc. PCS 2004, Picture Coding Symposium, San Francisco, CA, USA, December 15.-17. 2004.
[13] S. Pastoor, "3D Displays", in O. Schreer, P. Kauff, and T. Sikora (Editors), "3D Video Communication", Wiley, 2005.
[14] S. Würmlin, E. Lamboray, and M. Gross, "3D video fragments: dynamic point samples for real-time free-viewpoint video", Computers and Graphics 28 (1), Special Issue on Coding, Compression and Streaming Techniques for 3D and Multimedia Data, pp. 3-14, Elsevier Ltd, 2004.
[15] A. Smolic, and D. McCutchen, "3DAV Exploration of Video-Based Rendering Technology in MPEG", IEEE Trans. on Circuits and Systems for Video Technology, Vol. 14, No. 3, pp. 348-356, March 2004.
[16] A. Smolic, and P. Kauff, "Interactive 3D Video Representation and Coding Technologies", Proceedings of the IEEE, Special Issue on Advances in Video Coding and Delivery, vol. 93, no. 1, Jan. 2005
[17] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-Quality Video View Interpolation Using a Layered Representation", SIGGRAPH04, Los Angeles, CA, USA, August 2004.
[18] P. Merkle, K. Mueller, A. Smolic, and T. Wiegand, "Efficient Compression of Multi-view Video Exploiting Inter-view Dependencies Based on H.264/MPEG4-AVC", Proc. ICME 2006, International Conference on Multimedia and Expo, Toronto, Canada, July 9-12 2006.