MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Development of MPEG Standards for 3D and Free Viewpoint Video

Aljoscha Smolic Hideaki Kimata Anthony Vetro

TR-2005-116 October 2005

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. Suitable 3D scene representation formats are classified and the processing chain is explained. Examples are shown for image-based and model-based free viewpoint video systems, 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. MPEG is currently working on a new standard for multi-view video coding. 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.

Development of MPEG Standards for 3D and Free Viewpoint Video

Aljoscha Smolic	Hideaki Kimata	Anthony Vetro
Fraunhofer HHI	NTT	MERL
Germany	Japan	USA

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. Suitable 3D scene representation formats are classified and the processing chain is explained. Examples are shown for image-based and model-based free viewpoint video systems, 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. MPEG is currently working on a new standard for multi-view video coding. 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.

Keywords: 3D video, 3DTV, free viewpoint video, MPEG, 3DAV, multi-view video coding, 3D video objects

1. INTRODUCTION

3D and free viewpoint video are new types of natural video media 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, while the second allows for interactive selection of viewpoint and direction within a certain operating range as known from computer graphics applications. Target applications include broadcast television and other forms of video entertainment, as well as surveillance. These applications are enabled through convergence of technologies from computer graphics, computer vision, multimedia and related fields, and rapid progress in research covering the whole processing chain from capturing, signal processing, data representation, compression, transmission, display and interaction. Some of these 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.

To ensure interoperability between different systems, standardized formats for data representation and compression are necessary; these interchangeable formats are typically specified by international standardization bodies such as ISO MPEG. In recent years, 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) [14]. 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 [16], [17]. 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 has been adopted as part of the MPEG-4 Animation Framework eXtension (AFX) specification [18]. 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).

2.1 Acquisition for Free Viewpoint Video

Different technologies can be used for acquisition, processing, representation, and rendering, but all make use of multiple views of the same visual scene [1], as illustrated in Fig. 2 and Fig. 3. 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 (e.g. array type as in Fig. 2 or dome type as in Fig. 3) and density (i.e. number of cameras) imposes practical limitations to navigation and quality of rendered views at a certain virtual position. For instance the setting in Fig. 2 would not allow rendering a virtual view from inside the aquarium looking towards the cameras. Therefore there is a classical trade-off to consider between costs (for equipment, cameras, processors, etc.) and benefits (navigation range, quality of virtual views).

2.2 3D Scene Representation for Free Viewpoint Video

The choice of a certain 3D scene representation format is of central importance for the design of any FVV system. On the one side it sets the requirements for acquisition and multi-view signal processing. For instance using an image-based representation (see below) implies using a dense camera setting as shown in Fig. 2. A relatively sparse camera setting as shown in Fig. 3 would only give poor rendering results of virtual views. Using a geometry-based representation (see below) in contrary implies the need for sophisticated and error prone image processing algorithms such as object segmentation and 3D geometry reconstruction. On the other side the 3D scene representation determines the rendering algorithms (with that also navigation range, quality, etc.), interactivity, as well as compression and transmission if necessary.

In computer graphics literature, methods for 3D scene representation are often classified as a continuum in between two extremes as illustrated in Fig. 1 [13]. The one extreme is represented by classical 3D computer graphics. This approach can also be called geometry-based modeling. In most cases scene geometry is described on basis of 3D meshes. Real world objects are reproduced using geometric 3D surfaces with an associated texture mapped onto them. More sophisticated attributes can be assigned as well. For instance, appearance properties (opacity, reflectance, specular lights, etc.) can enhance the realism of the models significantly.

Everyone is familiar with this type of computer graphics from games, Internet, TV, movies, etc. The achievable performance might be extremely good if the scenes are purely computer generated. The available technology for both, production and rendering has highly been optimized over the last few years, especially in the case of common 3D mesh representations. In addition, state-of-the-art PC graphics cards are meanwhile able to render highly complex scenes with an impressive quality in terms of refresh rate, levels of detail, spatial resolution, reproduction of motion, and accuracy of textures.

A drawback to this approach is the high costs for content creation. Aiming at photo-realism, 3D scene and object modeling is complex and time consuming, and it becomes even more complex if a dynamically changing environment simulating real life is being created. Furthermore, an automatic 3D object and scene reconstruction implies an estimation of camera geometry, depth structures and 3D shapes. Inherently, all these processes tend to produce occasional errors. Therefore high-quality production, e.g. for movies, has to be done user assisted, supervised by a skilled operator.

The other extreme is given by scene representations that do not use any 3D geometry at all. It is usually called image-based modeling. In this case virtual intermediate views are generated from available real views by interpolation. The main advantages are a high quality of virtual view synthesis and an avoidance of 3D scene reconstruction. However, these benefits have to be paid by dense sampling of the real world with plenty of original view images. In general, the synthesis quality increases with the number of available views. Hence, a large amount of cameras has to be set up to achieve high-performance rendering, and plenty of image data needs to be processed therefore. Contrariwise, if the

number of used cameras is too low, interpolation and occlusion artifacts will appear in the synthesized images, possibly affecting the quality.
Examples of image-based representations are Ray-Space [9], [10] or light-field rendering [26], and panoramic configurations including concentric and cylindrical mosaics [32]-[35]. All these methods do not make any use of geometry, but they either have to cope with an enormous complexity in terms of data acquisition or they execute simplifications restricting the level of interactivity.
In between the two extremes there exists a continuum of methods that make more or less use of both approaches and combine the advantages in a particular manner. For instance a Lumigraph [36], [37] uses a similar representation as a light-field but adds a rough 3D model. This provides information on the depth structure of the scene and therefore allows for reducing the number of views.
Other representations do not use explicit 3D models but depth or disparity maps. Such maps assign a depth value to each pixel of an image. Together with the original 2D image the depth map builds a 3D-like representation, often called 2.5D. This can be extended to Layered Depth Images [38] where multiple color and depth values are stored in consecutively ordered depth layers.
Closer to the geometry-based end of the spectrum we can find methods that use view-dependent geometry and/or view dependent texture [23], [31]. Surface light-fields combine the idea of light-fields with an explicit 3D model [39], [40]. Furthermore, volumetric representations such as voxels (from volume elements) can be used instead of a complete 3D mesh model to describe 3D geometry [41]-[45].
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. A complete system for efficient representation and interactive streaming of high-resolution panoramic views has been presented in [46]. Other coding and transmission aspects of such data have also been studied, for example in [10], [19], [20], [33], [47]-[52], [55]. The European IST project ATTEST has studied a complete processing chain for interactive 3D broadcast including 3DTV acquisition, data representation, joint coding of video and depth maps, auto-stereoscopic 3D displays and parallax viewing based on head tracking [56].

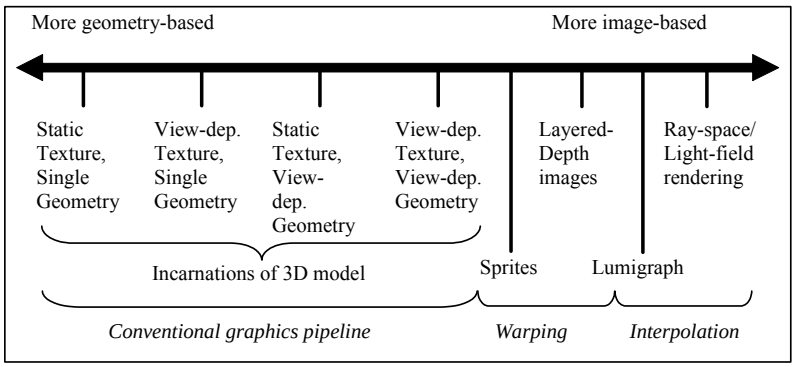

Figure 1: Categorization of scene representations [13].

2.3 Image-based Free Viewpoint Video using Ray-Space

Fig. 2 illustrates a purely image-based system called Free Viewpoint TV (FTV) developed at Nagoya University in Japan including the acquisition system and rendered virtual views [9]. A scene is captured using a dense array of synchronized cameras. The camera signals are represented in a special format called Ray-Space [10] that allows rendering the scene from any position (within practical limits) and does not rely on any geometry or depth reconstruction. Virtual intermediate views are purely generated from the available image data. Therefore rendering is rather based on signal processing than on computer graphics methods. With that Ray-Space can be regarded as the natural extension of classical 2D video to a general multiple view 3D scene representation and it may become the general video representation format in the future.

Image-based methods require a dense sampling of the scene with many cameras, to achieve a good rendering quality of virtual intermediate views. They are well suited for full complex scenes with many objects and a complex depth structure that can hardly be handled by 3D reconstruction methods, which are described in the next section. The scene in Fig. 2 contains a lot of small moving objects and has a complex depth structure. It is therefore quite difficult for methods that rely on some kind of geometry reconstruction. The Ray-Space representation provides high-quality virtual views for arbitrary viewpoints within the operating range and therefore enables general free viewpoint video applications.

However, capturing requires a tremendous effort and high quality (i.e. high image resolution) acquisition of a dynamic Ray-Space is still a difficult task. To keep the number of cameras practically reasonable, interpolation of intermediate ray data can be applied. Different methods based on table lookup, adaptive filtering and disparity-based interpolation have been evaluated experimentally in the 3DAV group. These have proven the suitability of the Ray-Space concept for general free viewpoint video applications [17].

A full dynamic Ray-Space results in an enormous data rate. Therefore efficient compression is the second key technology to make a Ray-Space system feasible besides interpolation. Compression of static light-fields has been studied for example in [19], [20]. Obviously, the images of a light-field representation include a lot of redundancy, i.e. inter-view (spatial) dependencies can be exploited for predictive coding. In case of a dynamic Ray-Space also temporal dependencies are available, so in this case a trade-off of spatial and temporal prediction has to be found. Optimum compression of dense Ray-Space and multiple view video data is still under investigation in the 3DAV group. It is expected that this work will lead to an extension of available MPEG standards (see section 4).

Figure 2: FTV acquisition and rendered virtual views.

2.4 Model-based Free Viewpoint Video

Fig. 3 - Fig. 6 illustrate another example of FVV [11]. 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.

Fig. 3 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 [15]. The first multi-view signal processing step consists in segmentation of the objects of interest, i.e. those that shall be reconstructed in 3D [23]. Although a huge effort has been put into this, segmentation is still and 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.

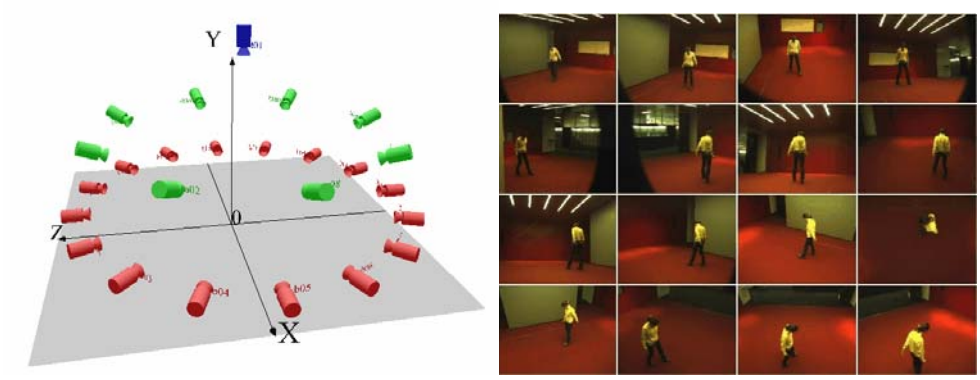

Figure 3: Multi-camera setup for 3DVO acquisition and captured multi-view video.

Having estimated the object's silhouette in each input image the 3D shape can be reconstructed using a shape-from-silhouette algorithm [24], [28]-[30]. The object's estimated silhouette in each camera image is projected from the centre of projection of each camera into 3D space as illustrated in Fig. 4. The resulting 3D volumes are called visual cones. If the 2D shape of the object is contained in each 2D shape, the 3D volume must be contained in each visual cone. Then specifically, the 3D object must be contained in the intersection volume of all visual cones. The 3D object volume can be reconstructed for instance by modelling the volume by voxels and reprojecting them into the images. A voxel belongs to the object volume if the reprojection is within the 2D silhouette in all images.

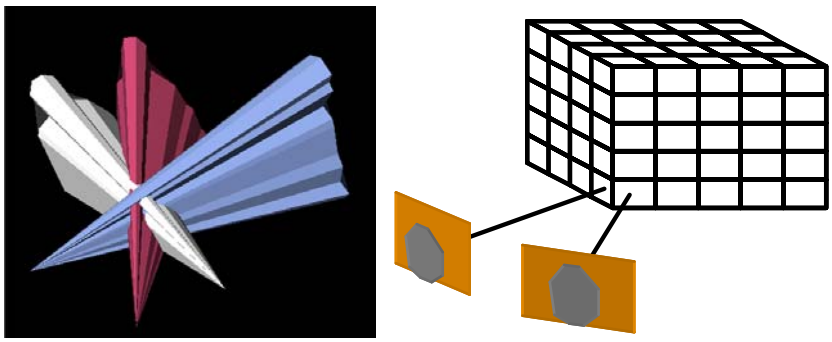

Figure 4: Visual cones and voxel reprojection.

The result is a voxel model of the object's 3D volume, as shown in Fig. 5 left. In a next step the object's surface can be extracted using a marching-cubes algorithm [21] and represented as classical 3D mesh, as shown in Fig. 5 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 [22] may help to regularize the estimated 3D mesh. Finally colour and texture can be projected from the available camera views onto the 3D mesh, as shown in Fig. 5 right. The result is a 3DVO as shown in Fig. 5 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.

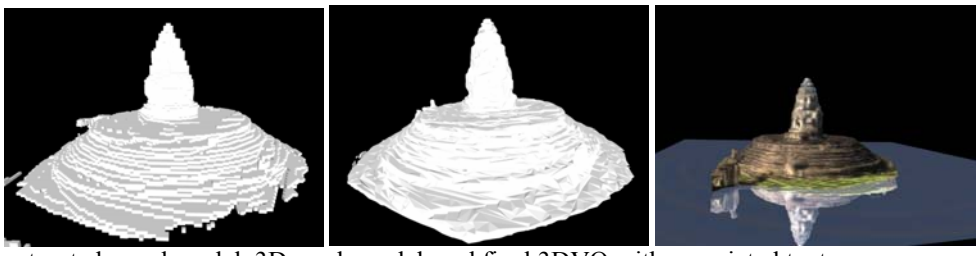

Figure 5: Reconstructed voxel model, 3D mesh model, and final 3DVO with associated textures.

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 [23], [31]. 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. 6 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. The reconstruction process can be significantly improved if a priori knowledge about the object of interest is available. In many cases this will be a human acting in some way. In [59] it is shown how usage of a skeleton model can improve 3D reconstruction.

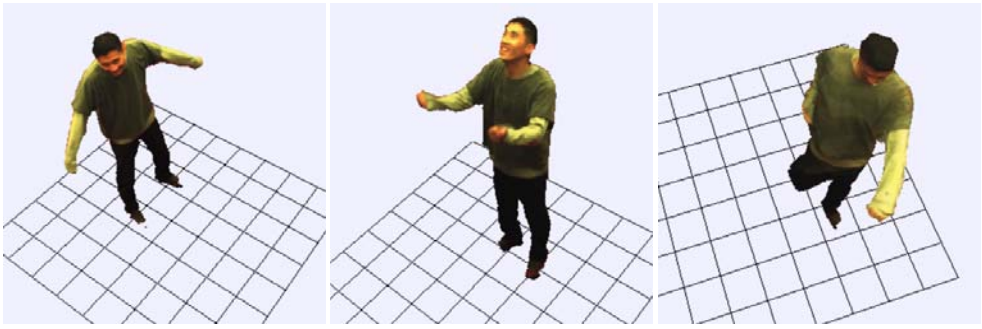

Figure 6: Virtual camera fly, rendered views at 3 different times from 3 different virtual viewpoints.

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 [57]. 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) [18], 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 [58]. 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 usage of 3D point clouds or video fragments [53], [54]. This representation uses unorganized point clouds in 3D, i.e. points with 3D coordinates but without connectivity. Additional attributes as colour 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. Fig. 7 shows an example of a 3DVO generated from a point cloud representation. Compression of such data has been investigated in [55]. 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 [18].
Other popular representation and rendering formats for FVV are based on per pixel depth information associated with the multiple views [12], which is described in more detail in the next section.

Figure 7: Left, right: original camera views, middle: virtual intermediate view of 3D video object integrated into a virtual environment [53].

3. 3D VIDEO

The second new functionality provided by these new technologies is a 3D depth impression of the observed scene. In fact, this functionality, also known as stereo, 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 no more glasses are needed) and advanced 3D rendering that supports head motion parallax viewing [1].
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). Fig. 8 shows an example of a 3DTV system. A scene is again captured by N synchronized cameras [7]. The multiple video signals are encoded and transmitted. At the receiver they are decoded, rendered and displayed on a 3D display. 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.

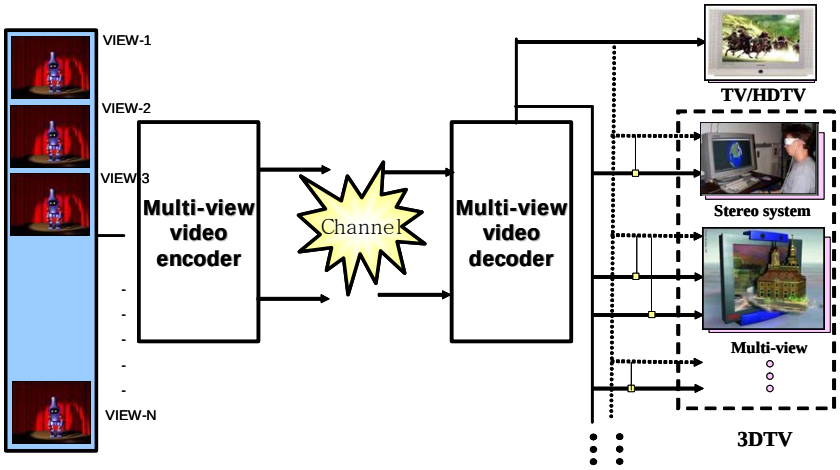

Figure 8: Example of 3DTV system.

Fig. 9 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 [61]. 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. 9 allows for user tracking with built in camera sensors [60]. The user's gaze direction is 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 colour or brightness using a classical TV set [61].

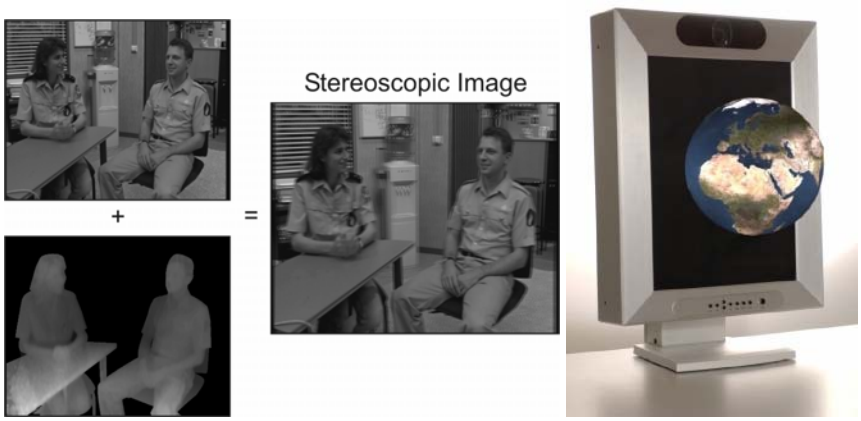

Figure 9: Depth-based stereo rendering and autostereoscopic 3D display (no glasses required).

The full 3DTV processing chain has been realized and demonstrated in the European ATTEST project [56]. 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 [62]. 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 [12]. Depending to 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 [62]. However, in a “Call for Evidence” [5] 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 [4]. A simple structure for spatio-temporal prediction is shown in Fig. 10. 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. 10 (see [4] for details) including for instance spatio-temporal B-pictures etc., 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 [4] for details). The latter describes prediction by warping of neighbouring 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.

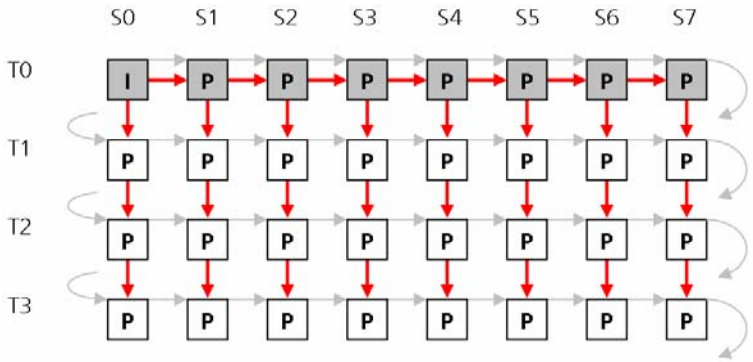

Figure 10: Spatio-temporal prediction structure for MVC, S indicates cameras, T indicates time.

Since further a "Call for Comments" [2] has shown that there is large industry interest in systems and applications described above, MPEG decided to issue a "Call for Proposals" [8] for MVC technology along with related requirements [3]. The responses to the "Call for Proposals" will be evaluated in January 2006 and will lead to a new standard for MVC.

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.

ACKNOWLEDGEMENT

The authors would like to thank a number of individuals and institutions for the help providing data, results and figures for this paper: M. Magnor, B. Goldluecke, C. Theobalt (Max-Planck-Institute for Computer Science), C. Fehn, K. Müller, P. Merkle, M. Kautzner, K. Schueuer, P. Kauff, P. Eisert, T. Wiegand (Fraunhofer HHI), Prof. T. Fujii, Prof. M. Tanimoto (Nagoya University), H.P. Pfister (MERL), E. Lamboray, S. Wuermlin, M. Waschbuesch, M. Gross (ETH Zuerich).

REFERENCES

[1] 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

[2] ISO/IEC JTC1/SC29/WG11, "Call for Comments on 3DAV", Doc. N6051, Gold Coast, Australia, October 2003.

[3] ISO/IEC JTC1/SC29/WG11, "Requirements on Multi-view Video Coding v.2", Doc. N7282, Poznan, Poland, July 2005.

[4] ISO/IEC JTC1/SC29/WG11, "Survey of Algorithms used for Multi-view Video Coding (MVC)", Doc. N6909, Hong Kong, China, January 2005.

[5] ISO/IEC JTC1/SC29/WG11, "Call for Evidence on Multi-View Video Coding", Doc. N6720, Palma de Mallorca, Spain, October 2004.

[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] W. Matusik, H. Pfister, “3D TV: A Scalable System for Real-Time Acquistion, Transmission and Autostereoscopic Display of Dynamic Scenes”, ACM Transactions on Graphics (TOG) SIGGRAPH, ISSN: 0730-0301, Vol. 23, Issue 3, pp. 814-824, August 2004.
[8] ISO/IEC JTC1/SC29/WG11, “Call for Proposals on Multi-view Video Coding”, Doc. N7327, Poznan, Poland, July 2005.
[9] Masayuki Tanimoto, “Free Viewpoint Television - FTV”, Proc. PCS 2004, Picture Coding Symposium, San Francisco, CA, USA, December 15.-17. 2004.
[10] T. Fujii, M. Tanimoto, “Free-Viewpoint TV System Based on Ray-Space Representation”, SPIE ITCom Vol. 4864-22, pp.175-189 (2002).
[11] 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.
[12] 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.
[13] S.B. Kang, R. Szeliski, and P. Anandan, “The Geometry-Image Representation Tradeoff for Rendering”, Proc. ICIP2000, IEEE International Conference on Image Processing, Vancouver, Canada, September 2000.
[14] 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.
[15] R.Y. Tsai, “A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV camera and lenses”, IEEE Journal of Robotics and Automation, Vol. RA-3, No. 4, August 1987.
[16] ISO/IEC JTC1/SC29/WG11, “Applications and Requirements for 3DAV”, Doc. N5877, Trondheim, Norway, July 2003.
[17] ISO/IEC JTC1/SC29/WG11, “Report on 3DAV Exploration”, Doc. N5878, Trondheim, Norway, July 2003.
[18] ISO/IEC JTC1/SC29/WG11, “ISO/IEC 14496-16/PDAM1”, Doc. N6544, Redmont, WA, USA, July 2004.
[19] M. Magnor, and B. Girod, “Data Compression for Light-Field Rendering", IEEE Trans. on Circuits and Systems for Video Technology, vol. 10, no. 3, pp. 338-343, Apr. 2000.
[20] M. Magnor, P. Ramanathan, and B. Girod, “Multi-View Coding for Image-based Rendering using 3-D Scene Geometry”, IEEE Trans. on Circuits and Systems for Video Technology, Vol. 13, No. 11, pp. 1092-1106, November 2003.
[21] W. E. Lorensen, and H. E. Cline, “Marching Cubes: A high resolution 3D surface reconstruction algorithm,” Proceedings of SIGGRAPH, vol. 21, no. 4, pp 163-169, 1987.
[22] G. Taubin, “Curve and Surface Smoothing Without Shrinkage”, International Conference on Computer Vision (ICCV '95), pp. 852-857, 1995.
[23] K. Mueller, A. Smolić, M. Droese, P. Voigt, and T. Wiegand, “Reconstruction of a Dynamic Environment with Fully Calibrated Background for Traffic Scenes”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 15, No. 4, pp. 538-549, April 2005.
[24] A. Laurentini, “The Visual Hull Concept for Silhouette Based Image Understanding”, IEEE Trans. on PAMI, Vol. 16, No. 2, pp. 150-162, 1994.
[25] W.H. Leung, and T. Chen, “Compression with Mosaic Prediction for Image-Based rendering Applications”, Proc. ICME2000, IEEE International Conference on Multimedia and Expo, New York, NY, USA, July 2000.
[26] M. Levoy, and P. Hanrahan, “Light Field Rendering”, Proc. ACM SIGGRAPH, pp. 31-42, August 1996.
[27] J. Li, H.Y. Shum, and Y.Q. Zhang, “On the Compression of Image Based Rendering Scene”, Proc. ICIP2000, IEEE International Conference on Image Processing, Vancouver, Canada, September 2000.
[28] M. Li, M. Magnor, and H.-P. Seidel, “Hardware-Accelerated Visual Hull Reconstruction and Rendering”, pp. 65-71, Graphics Interface'2003, June 2003.
[29] W. Matusik, C. Buehler, R. Raskar, S. J. Gortler, and L. McMillan, “Image-Based Visual Hulls”, Proc. SIGGRAPH 2000, pages 369–374, 2000.
[30] Matusik, W., Buehler, C., and McMillan, L., “Polyhedral Visual Hulls for Real-Time Rendering”, Proc. Eurographics Workshop on Rendering 2001.
[31] P. Debevec, C. Taylor, and J. Malik, “Modeling and rendering architecture from photographs: A hybrid geometry- and image based approach”, Proceedings of SIGGRAPH 1996, pp. 11-20, 1996.
[32] H.Y. Shum, and L.W. He, “Rendering with Concentric Mosaics”, Proc. ACM SIGGRAPH, pp. 299-306, August 1999.

[33] H.Y. Shum, K.T. Ng, and S.C. Chan, “Virtual Reality using the Concentric Mosaic: Construction, Rendering and Data Compression”, Proc. ICIP2000, IEEE International Conference on Image Processing, Vancouver, Canada, September 2000.
[34] R. Szeliski, and H.Y. Shum, “Creating Full View Panoramic Image Mosaics and Texture-mapped Models”, Proc. ACM SIGGRAPH, pp. 251-258, August 1997.
[35] S.E. Chen, “QuickTime VR – An Image-Based Approach to Virtual Environment Navigation”, Proc. ACM SIGGRAPH, pp. 29-38, August 1995.
[36] C. Buehler, M. Bosse, L. McMillan, S. Gortler, and M. Cohen, “Unstructured Lumigraph Rendering”, Proceedings of SIGGRAPH 2001, pp. 425-432, 2001.
[37] S.J. Gortler, R. Grzesczuk, R. Szeliski, and M.F. Cohen, “The Lumigraph", ACM SIGGRAPH '96, pp.43-54, Aug. 1996.
[38] J. Shade, S. Gortler, L.W. He, and R. Szeliski, “Layered Depth Images”, Proc. SIGGRAPH '98, Orlando, FL, USA, July 1998.
[39] D. Wood, D. Azuma, W. Aldinger, B. Curless, T. Duchamp, D. Salesin, and W. Stuetzle, “Surface Light Fields for 3D Photography”, Proceedings of SIGGRAPH 2000.
[40] W.-C. Chen, J.-Y. Bouguet, M.H. Chu, and R. Grzeszczuk, “Light Field Mapping: Efficient Representation and Hardware Rendering of Surface Light Fields”, ACM Transactions on Graphics. 21 (3), pp. 447-456, 2002.
[41] P. Eisert, E. Steinbach, and B. Girod, “Automatic Reconstruction of Stationary 3-D Objects from Multiple Uncalibrated Camera Views”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 10, No. 2, pp. 261-277, March 2000.
[42] K.N. Kutulakos, and S.M. Seitz, “A Theory of Shape by Space Carving”, University of Rochester, Tech. Report 692, May 1998.
[43] T. Matsuyama, and T. Takai, “Generation, Visualization, and Editing of 3D Video”, Proc. Symposium on 3D Data Processing Visualization and Transmission, pp.234-245, Padova, Italy, June 2002.
[44] T. Matsuyama, X. Wu, T. Takai, and T. Wada, “Real-Time Dynamic 3-D Object Shape Reconstruction and High-Fidelity Texture Mapping for 3-D Video”, IEEE Trans. on Circuits and Systems for Video Technology, Vol. 14, No. 3, pp. 357-369, March 2004.
[45] S. M. Seitz, and C. R. Dyer, “Photorealistic Scene Reconstruction by Voxel Coloring”, International Journal of Computer Vision, 35(2), 1999, pp. 151-173.
[46] C. Grünheit, A. Smolic, and T. Wiegand, “Efficient Representation and Interactive Streaming of High-Resolution Panoramic Views”, Proc. ICIP2002, IEEE International Conference on Image Processing, Rochester, NY, USA, September 22.-25. 2002.
[47] T. Hamaguchi, T. Fujii, Y. Kajiki, and T. Honda, “Real-time View Interpolation System for Multi-View 3D Display”, SPIE Electronic Imaging, Vol. 4297A, pp. 212-221, Jan. 2001.
[48] T. Kobayashi, T. Fujii, T. Kimoto, and M. Tanimoto, “Interpolation of Ray-Space Data by Adaptive Filtering”, SPIE Electronic Imaging 2000, Vol. 3958, pp. 252-259, Jan. 2000.
[49] W.H. Leung, and T. Chen, “Compression with Mosaic Prediction for Image-Based rendering Applications”, Proc. ICME2000, IEEE International Conference on Multimedia and Expo, New York, NY, USA, July 2000.
[50] J. Li, H.Y. Shum, and Y.Q. Zhang, “On the Compression of Image Based Rendering Scene”, Proc. ICIP2000, IEEE International Conference on Image Processing, Vancouver, Canada, September 2000.
[51] T. Pintaric, U. Neumann, and A. Rizzo, “Immersive Panoramic Video”, Proceedings of the 8th ACM International Conference on Multimedia, pp. 493.494, October 2000.
[52] C. Zhang, and J. Li, “Compression of Lumigraph with Multiple Reference Frame (MRF) Prediction and Just-in-time Rendering”, Proc. DCC2000, IEEE Data Compression Conference, Snowbird, Utah, USA, March 2000.
[53] S. Würmlin, E. Lamboray, O. Staadt, and M. Gross, “3D Video Recorder: A System for Recording, Processing and Playing Three-Dimensional Video”, Computer Graphics Forum 22 (2), Blackwell Publishing Ltd, Oxford, U.K., pp. 181-193, 2003.
[54] 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.
[55] E. Lamboray, S. Wümlin, 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.

[56] 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.

[57] A. Smolic, K. Mueller, P. Merkle, T. Rein, P. Eisert, and T. Wiegand, “Representation, Coding, and Rendering of 3D Video Objects with MPEG-4”, MMSP 2004, IEEE International Workshop on Multimedia Signal Processing, Siena, Italy, September 29.-October 1. 2004.

[58] K. Mueller, A. Smolić, 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.

[59] J. Carranza, C. Theobalt, M. Magnor, and H.-P. Seidel, “Free-Viewpoint Video of Human Actors”, ACM Trans. on Graphics (special issue SIGGRAPH'03), vol. 22, no. 3, pp. 569-577, July 2003.

[60] S. Pastoor, “3D Displays”, in O. Schreer, P. Kauff, and T. Sikora (Editors), “3D Video Communication”, Wiley, 2005.

[61] C. Fehn, “3D TV Broadcasting”, in O. Schreer, P. Kauff, and T. Sikora (Editors), “3D Video Communication”, Wiley, 2005.

[62] ITU-T Recommendation H.264 & ISO/IEC 14496-10 AVC, “Advanced Video Coding for Generic Audio-Visual Services”, 2003.