An Overview of 3D Video and Free Viewpoint Video

Aljoscha Smolic*

Disney Research, Zurich
Clausiusstrasse 49
8092 Zurich, Switzerland
smolic@zurich.disneyresearch.com

Abstract. An overview of 3D video and free viewpoint video is given in this paper. Free viewpoint video allows the user to freely navigate within real world visual scenes, as known from virtual worlds in computer graphics. 3D video provides the user with a 3D depth impression of the observed scene, which is also known as stereo video. In that sense as functionalities, 3D video and free viewpoint video are not mutually exclusive but can very well be combined in a single system. Research in this area combines computer graphics, computer vision and visual communications. It spans the whole media processing chain from capture to display and the design of systems has to take all parts into account. The conclusion is that the necessary technology including standard media formats for 3D video and free viewpoint video is available or will be available in the future, and that there is a clear demand from industry and user side for such new types of visual media.

Keywords: 3D video, free viewpoint video, MPEG, 3DTV.

1 Introduction

Convergence of technologies from computer graphics, computer vision, multimedia and related fields enabled the development of new types of visual media, such as 3D video (3DV) and free viewpoint video (FVV) that expand the user's sensation beyond what is offered by traditional 2D video [1]. 3DV, also referred to as stereo, offers a 3D depth impression of the observed scenery, while FVV allows for an interactive selection of viewpoint and direction within a certain operating range, as known from computer graphics. Both do not exclude each other. In contrary, they can be very well combined within a single system, since they are both based on a suitable 3D scene representation (see below). In other words, given a 3D representation of a scene, if a stereo pair corresponding to the human eyes can be rendered, the functionality of 3DV is provided. If a virtual view (i.e., not an available camera view) corresponding to an arbitrary viewpoint and viewing direction can be rendered, the functionality of FVV is provided. The ideal future visual media system will provide full FVV and 3DV at the same time. In order to enable 3DV and FVV applications, the whole processing chain, including acquisition, sender side processing, 3D representation, coding,

* Work for this paper was performed during the author's prior affiliation with the Fraunhofer Institute for Telecommunications – Heinrich-Hertz-Institut (FhG-HHI), Berlin, Germany.

X. Jiang and N. Petkov (Eds.): CAIP 2009, LNCS 5702, pp. 1–8, 2009.

transmission, rendering and display need to be considered. The 3DV and FVV processing chain is illustrated in Fig. 1. The design has to take all parts into account, 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.

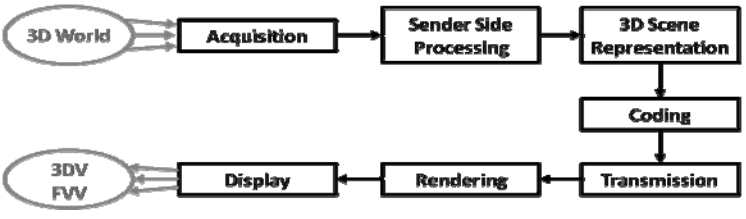

Fig. 1. 3DV and FVV processing chain

2 3D Scene Representation

The choice of a 3D scene representation format is of central importance for the design of any 3DV or FVV system [2]. On the one hand, the 3D scene representation sets the requirements for acquisition and signal processing on sender side, e.g. the number and setting of cameras and the algorithms to extract the necessary data types. On the other hand, the 3D scene representation determines the rendering algorithms (and with that also navigation range, quality, etc.), interactivity, as well as coding and transmission.

In computer graphics literature, methods for 3D scene representation are often classified as a continuum in between two extremes as illustrated in Fig. 2 [3]. These principles can also be applied for 3DV and FVV. 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 the 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 significantly enhance the realism of the models.

The other extreme in 3D scene representations in Fig. 2 is called image-based modeling and does not use any 3D geometry at all. In this case virtual intermediate views are generated from available natural camera views by interpolation. The main advantage is a potentially high quality of virtual view synthesis avoiding any 3D scene reconstruction. However, this benefit has to be paid by dense sampling of the real world with a sufficiently large number of natural camera view images. In general, the synthesis quality increases with the number of available views. Hence, typically a large amount of cameras has to be set up to achieve high-performance rendering, and a tremendous amount 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.

In between the two extremes there exists a number of methods that make more or less use of both approaches and combine the advantages in some way. Some of these representations do not use explicit 3D models but depth or disparity maps. Such maps assign a depth value to each pixel of an image (see Fig 5).

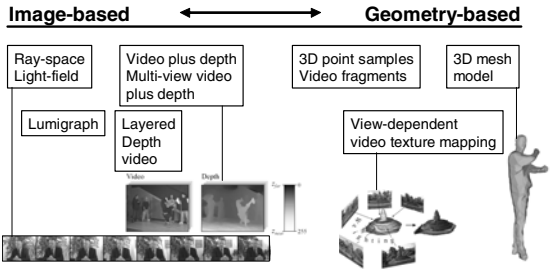

Fig. 2. 3D scene representations for 3DV and FVV

3 Acquisition

In most cases 3DV and FVV approaches rely on specific acquisition systems. Although automatic and interactive 2D-3D conversion (i.e. from 2D video to 3DV or FVV) is an important research area for itself. Most 3DV and FVV acquisition systems use multiple cameras to capture real world scenery [4]. These are sometimes combined with active depth sensors, structured light, etc. in order to capture scene geometry. The camera setting (e.g. dome type as in Fig. 3) and density (i.e. number of cameras) impose practical limitations on 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 quality (navigation range, quality of virtual views). Fig. 3 illustrates a dome type multi camera acquisition system and captured multi-view video. Such multi-view acquisition is an important and highly actual research area [4].

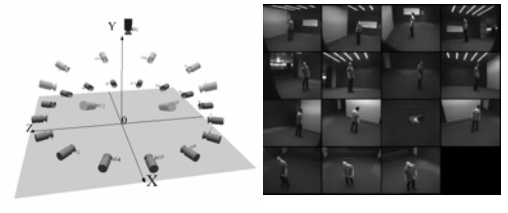

Fig. 3. Multi-camera setup for 3DVO acquisition and captured multi-view video

4 Sender Side Processing

After acquisition, the necessary data as defined by the 3D representation format have to be extracted from the multiple video and other captured data. This sender side processing can include automatic and interactive steps; it may be real-time or offline. Content creation and post processing are included here. Tasks may be divided into low-level computer vision algorithms and higher-level 3D reconstruction algorithms.

Low-level vision may include algorithms like color correction, while balancing, normalization, filtering, rectification, segmentation, camera calibration, feature

extraction and tracking, etc. 3D reconstruction algorithms include for instance depth estimation and visual hull reconstruction to generate 3D mesh models. A general problem of 3D reconstruction algorithms is that they are estimations by nature. The true information is in general not accessible. Robustness of the estimation depends on many theoretical and practical factors. There is always a residual error probability which may affect the quality of the finally rendered output views. User-assisted content generation is an option for specific applications to improve performance. Purely image-based 3D scene representations do not rely on 3D reconstruction algorithms, and therefore do not suffer from such limitations.

Fig. 4 illustrates different steps of a 3D reconstruction pipeline. This includes visual hull reconstruction, surface extraction, surface smoothing, and mesh simplification [5]. Fig. 5 illustrates video and associated per pixel depth data.

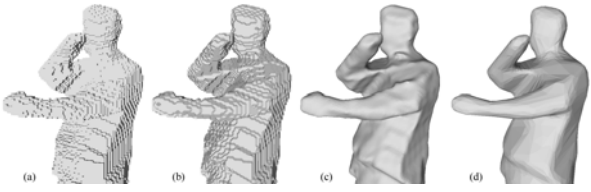

Fig. 4. Different steps of 3D reconstruction

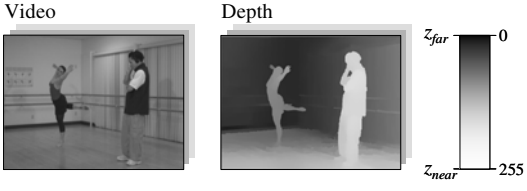

Fig. 5. Video and associated per pixel depth data

5 Coding, Transmission, Decoding

For transmission over limited channels 3DV and FVV data have to be compressed efficiently. This has been widely studied in literature and powerful algorithms are available [6]. International standards for content formats and associated coding technology are necessary to ensure interoperability between different systems. ISO-MPEG and ITU-VCEG are international organizations that released a variety of important standards for digital media including standards for 3DV and FVV. Classical 2-view stereo is already supported by MPEG-2 since the mid 90ies. Current releases of the latest video coding standard H.264/AVC also include a variety of highly efficient modes to support stereo video. This easily extends to multi-view video coding (MVC) with inter-view prediction, a recently released extension of H.264/AVC, which is illustrated in Fig. 6 [7]. It is the currently most efficient way to encode 2 or more videos showing the same scenery from different viewpoints.

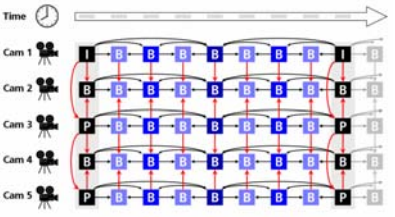

Fig. 6. Multi-view video coding (MVC)

Video plus depth as illustrated in Fig. 5 is already supported by a standard known as MPEG-C Part 3. It is an alternative format for 3DV that requires view synthesis at the receiver (see next section). Video plus depth supports extended functionality compared to classical 2-view stereo such as baseline adaptation to adjust depth impression to different displays and viewing preferences [8]. Currently MPEG prepares a new standard that will provide even more extended functionalities by using multi-view video plus depth or layered depth video [9].

Different model-based or 3D point cloud representations for FVV are supported by various tools of the MPEG-4 standard. Fig. 7 illustrates coding and multiplexing of dynamic 3D geometry, associated video textures and auxiliary data [10].

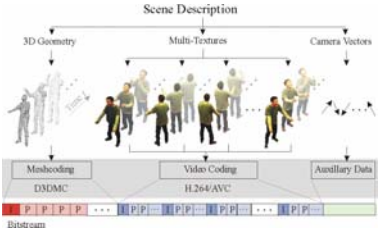

Fig. 7. Coding and multiplexing of dynamic 3D geometry, associated video textures and auxiliary data

6 Rendering

Rendering is the process of generation of the final output views from data in the 3D representation format. Fig. 8 illustrates an interactive 3D scene with a FVV object included. The scene further includes a 360° panorama and a computer graphics object. In this case rendering is done by classical computer graphics methods. The user can navigate freely and watch the dynamic scene from any desired viewpoint and viewing direction.

Fig. 9 illustrates virtual intermediate view synthesis by 3D warping from multiple video plus depth data. Any desired view in between available camera views can be generated this way to support free viewpoint navigation and advanced 3DV functionalities [11]. For instance a multi-view auto-stereoscopic display can be supported efficiently by rendering 9, 16 or more views from a limited number of multi-view video plus depth data (e.g. 2 or 3 video and depth streams).

Fig. 8. Integrated interactive 3D scene with FVV

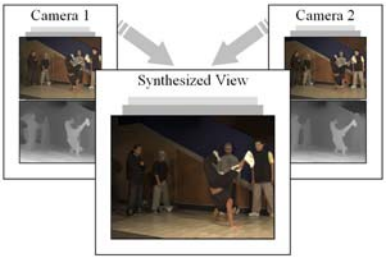

Fig. 9. Intermediate view synthesis from multiple video and depth data

7 Display

Finally the rendered output views are presented to the user on a display. FVV requires interactive input from the user to select the viewpoint. This can be done by classical devices like mouse or joystick. Some systems also track the user (head or gaze) employing cameras and infrared sensors.

In order to provide a depth impression 2 or more views have to be presented to the user appropriately at the same time using a specific 3D display. Such 3D displays ensure that the user perceives a different view with each eye at a time, by filtering the displayed views appropriately. If it is a proper stereo pair, the brain will compute a 3D depth impression of the observed scene.

Currently, various types of 3D displays are available and under development [12]. Most of them use classical 2-view stereo with one view for each eye and some kind of glasses (polarization, shutter, anaglyph) to filter the corresponding view. Then there are so called multi-view auto-stereoscopic displays which do not require glasses. Here, 2 or more views are displayed at the same time and a lenticular sheet or parallax barrier element in front of the light emitters ensures correct view separation for the viewer's eyes.

Fig. 10 illustrates a 2-view auto-stereoscopic display developed by Fraunhofer HHI, which does not require wearing glasses. 2 views are displayed at a time and a lenticular sheet projects them into different directions. A camera system tracks the

Fig. 10. Auto-stereoscopic display made by Fraunhofer HHI

user's gaze direction. A mechanical system orients the display within practical limits according to the user's motion and ensures proper projection of left and right view into direction of the corresponding eye.

8 Summary and Conclusions

This paper provided an overview of 3DV and FVV. It is meant to be supplemental material to the invited talk at the conference. Naturally, different aspects were summarized briefly. For more details the reader is referred to the publications listed below.

3DV and FVV were introduced as extended visual media that provide new functionalities compared to standard 2D video. Both can very well be provided by a single system. New technology spans the whole processing chain from capture to display. The 3D scene representation is determining the whole system. Technology for all the different parts is available, maturating and further emerging.

Growing interest for such applications is noticed from industry and users. 3DV is well established in cinemas. E.g. Hollywood is producing more and more 3D movies. There is a strong push from industry to bring 3DV also to the home, e.g. via Blu-ray or 3DTV. FVV is established as post-production technology. FVV end-user mass market applications are still to be expected for the future.

Acknowledgments. I would like to thank the Interactive Visual Media Group of Microsoft Research for providing the Breakdancers and Ballet data sets, and the Computer Graphics Lab of ETH Zurich for providing the Doo Young multi-view data set.

Background, knowledge and material for this paper were developed during my previous employment at the Fraunhofer Institute for Telecommunications – Heinrich-Hertz-Institut (FhG-HH), Berlin, Germany. Text and illustrations were developed in

collaboration with colleagues including Karsten Mueller, Philipp Merkle, Birgit Kaspar, Matthias Kautzner, Sabine Lukaschik, Peter Kauff, Peter Eisert, Thomas Wiegand, Christoph Fehn, Ralf Schaefer.

References

1. Smolic, A., Mueller, K., Merkle, P., Fehn, C., Kauff, P., Eisert, P., Wiegand, T.: 3D Video and Free View-point Video – Technologies, Applications and MPEG Standards. In: ICME 2006, International Conference on Multimedia and Expo, Toronto, Ontario, Canada (July 2006)
2. Smolic, A., Kauff, P.: Interactive 3D Video Representation and Coding Technologies. Proceedings of the IEEE, Special Issue on Advances in Video Coding and Delivery 93(1) (January 2005)
3. Kang, S.B., Szeliski, R., Anandan, P.: The Geometry-Image Representation Tradeoff for Rendering. In: ICIP 2000, IEEE International Conference on Image Processing, Vancouver, Canada (September 2000)
4. Kubota, A., Smolic, A., Magnor, M., Chen, T., Tanimoto, M.: Multi-View Imaging and 3DTV – Special Issue Overview and Introduction. IEEE Signal Processing Magazine, Special Issue on Multi-view Imaging and 3DTV 24(6) (November 2007)
5. Smolic, A., Mueller, K., Merkle, P., Rein, T., Eisert, P., Wiegand, T.: Free Viewpoint Video Extraction, Representation, Coding, and Rendering. In: ICIP 2004, IEEE International Conference on Image Processing, Singapore, October 24-27 (2004)
6. Smolic, A., Mueller, K., Stefanoski, N., Ostermann, J., Gotchev, A., Akar, G.B., Triantafyllidis, G., Koz, A.: Coding Algorithms for 3DTV - A Survey. IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Multiview Video Coding and 3DTV 17(11) (November 2007)
7. Merkle, P., Smolic, A., Mueller, K., Wiegand, T.: Efficient Prediction Structures for Multiview Video Coding. IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Multiview Video Coding and 3DTV 17(11) (November 2007)
8. Fehn, C., Kauff, P., Op de Beeck, M., Ernst, F., Ijsselsteijn, W., Pollefeys, M., Vangool, L., Ofek, E., Sexton, I.: An Evolutionary and Optimised Approach on 3D-TV. In: IBC 2002, Int. Broadcast Convention, Amsterdam, Netherlands (September 2002)
9. Smolic, A., Mueller, K., Merkle, P., Vetro, A.: Development of a new MPEG Standard for Advanced 3D Video Applications. In: ISPA 2009, 6th International Symposium on Image and Signal Processing and Analysis, Salzburg, Austria (September 2009)
10. Smolic, A., Mueller, K., Merkle, P., Kautzner, M., Wiegand, T.: 3D Video Objects for Interactive Applications. In: EUSIPCO 2005, Antalya, Turkey, September 4-8 (2005)
11. Mueller, K., Smolic, A., Dix, K., Merkle, P., Kauff, P., Wiegand, T.: View Synthesis for Advanced 3D Video Systems. EURASIP Journal on Image and Video Processing 2008, doi:10.1155/2008/438148
12. Konrad, J., Halle, M.: 3-D Displays and Signal Processing – An Answer to 3-D Ills? IEEE Signal Processing Magazine 24(6) (November 2007)