AN OVERVIEW OF AVAILABLE AND EMERGING 3D VIDEO FORMATS AND DEPTH ENHANCED STEREO AS EFFICIENT GENERIC SOLUTION

Aljoscha Smolic, Karsten Mueller, Philipp Merkle, Peter Kauff, Thomas Wiegand

Fraunhofer Institute for Telecommunications - Heinrich-Hertz-Institut, Berlin, Germany

ABSTRACT

Recently, popularity of 3D video has been growing significantly and it may turn into a home user mass market in the near future. However, diversity of 3D video content formats is still hampering wide success. An overview of available and emerging 3D video formats and standards is given, which are mostly related to specific types of applications and 3D displays. This includes conventional stereo video, multiview video, video plus depth, multiview video plus depth and layered depth video. Features and limitations are explained. Finally, depth enhanced stereo (DES) is introduced as a flexible, generic, and efficient 3D video format that can unify all others and serve as universal 3D video format in the future.

Index Terms— 3D video, stereo video, video coding, MVC, depth

1. INTRODUCTION

3D video is commonly understood as a type of visual media that provides depth perception of the observed scenery. It is also referred to as stereo video. Such 3D depth perception can be provided by 3D display systems which ensure that the user sees a specific different view with each eye [1]. The stereo pair of views must correspond to the human eye positions. Then the brain can compute the 3D depth perception. Most 3D display systems require wearing specific glasses (anaglyph, polarization, or shutter) to ensure separation of left and right view which are displayed simultaneously.

Extending visual sensation to the 3rd dimension has been investigated over decades. In fact, history of 3D displays dates back almost as long as classical 2D cinematography, but only recently popularity of 3D video has been growing significantly. Awareness of and interest in 3D video is rapidly increasing, among users who wish to experience the extended visual sensation, as well as among content producers, equipment providers, and distributors, who discover new business opportunities. At the same time technology is maturating from capture to display. The market of 3D cinema is expected to continue growing rapidly over the next years. More and more cinemas are being equipped with 3D technology. Hollywood studios are making large bets on 3D movies. For instance Disney announced that all animated Disney and Pixar movies will be released in 3D in the future.

With the content being produced, 3D video is also an increasingly interesting technology for home user living room applications. 3D video content will arrive to the home by 3D-DVD/Blu-ray, Internet, 3DTV broadcast, etc. Currently, there is a great variety of different 3D display systems designed for the home user applications, starting from classical 2-view stereo systems with glasses. More sophisticated candidates for 3D vision in living rooms are multiview auto-stereoscopic displays, which do not require glasses [1]. They emit more than one view at a time but the technology ensures that users only see a stereo pair from a specific viewpoint. Today's 3D displays are capable of showing 9 or more different images at the same time, of which only a stereo pair is visible from a specific viewpoint. This supports multi-user 3D vision without glasses in a living room environment. Motion parallax viewing can be supported if consecutive views are stereo pairs and arranged properly.

As a consequence, there are a lot of different 3D video formats available and under investigation. They include different types of data, mostly related to specific types of displays. This starts from classical 2 view stereo video, extends to multiview video with more than 2 views, video plus depth, multiview video plus depth, and layered depth video. A variety of compression and coding algorithms are available for the different 3D video formats. Some of these standards and coding algorithms are standardized e.g. by MPEG, since standard formats and efficient compression are crucial for the success of 3D video applications. A generic, flexible and efficient 3D video format that would serve a wide range of different 3D video systems is highly desirable in this context. Therefore MPEG is currently investigating such a new generic 3D video standard.

This paper gives an overview of available and emerging 3D video formats. Section 2 is devoted to conventional and stereo multiview video and Section 3 describes video plus depth. These formats are quite established and standards are available. Ongoing research directions are highlighted. Sections 4 and 5 describe multiview video plus depth and layered depth video respectively. Both are advanced 3D video formats currently under development and in scope of a related MPEG activity for a new standard. Section 6 presents depth enhanced stereo as a flexible, generic, and efficient 3D video format. It can be regarded as unification of all other formats in an efficient way and may therefore serve as universal 3D video format in the future. Finally, section 7 concludes the paper.

2. CONVENTIONAL STEREO AND MULTIVIEW VIDEO

This is the most well-known and in a way most simple type of 3D video representation, which is called conventional stereo video (CSV) in the following. Only color pixel video data are involved. After capture by 2 or more cameras the 2 or more video signals may have undergone some processing steps like normalization, color correction, rectification, etc., however, no scene geometry information is involved. The video signals are meant in principle to be directly displayed using a 3D display system, though some video processing might also be involved before display.

Compared to the other 3D video formats the algorithms associated with CSV are the least complex. It can be as simple as only separately encoding and decoding the multiple video signals. Only the amount of data is increasing compared to 2D video. By

reduction of resolution (spatial and/or temporal) this can be kept constant if necessary.

Coding efficiency can be increased by combined temporal/interview prediction as illustrated in Fig. 1. MPEG-2 provided a corresponding standard already more than 10 years ago (MPEG-2 Multiview Profile). Recently, a so called Stereo SEI (Supplemental Enhancement Information) message was added to the latest and most efficient video coding standard H.264/AVC, which implements inter-view prediction similar to the principle illustrated in Fig. 1.

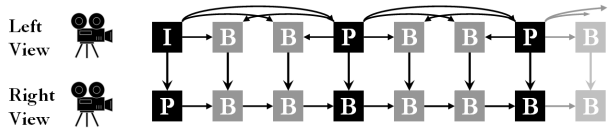

Fig. 1: Stereo coding, combined temporal/interview prediction.

For more than 2 views this is easily extended to Multiview Video Coding (MVC). A corresponding MPEG-ITU standard was released in 2008, which is an extension of H.264/AVC [2]. It can also be applied to 2 views. MVC is currently the most efficient way for stereo and multiview video coding, whereby the performance of a solution based on the H.264/AVC Stereo SEI message is similar for the stereo case.

A simple way to use existing video codecs for stereo video transmission is to apply temporal or spatial interleaving. This can in principle already be done with any existing equipment. A problem is that there is no corresponding standard available. There is no way to signal the use of interleaving to the decoder. The decoder has to know about it. A normal video decoder would decode the stereo video incorrectly.

A new approach for efficient stereo video coding which was proposed recently is derived from the so called binocular suppression theory [3]. This is illustrated in Fig. 2. Subjective test have shown that to some degree, if one of the images of a stereo pair is low-pass filtered, the perceived overall quality of the stereo video will be dominated by the higher quality image. I.e. the perceived quality will be as if both images are not low-passed.

Fig. 2: Stereo image pair with low-pass filtered right view.

Based on that effect, mixed resolution stereo video coding can be derived. Instead of coding the right image in full resolution it is downsampled to half or quarter resolution. In theory this should give similar overall subjective stereo video quality, while significantly reducing the bitrate. Taking the bitrate for the left view as given for 2D video, the 3D video functionality could be added by an overhead of 25-30% for coding the right view at quarter resolution. Such and similar approaches are currently under investigation.

A general drawback of CSV is that the 3D impression can not be modified. The baseline is fixed from capturing. Depth perception cannot be adjusted to different display types and sizes. The number of output views can not be varied (only decreased). Head motion parallax can not be supported (different perspective, occlusions & dis-occlusion when moving the viewpoint). The functionality of CSV is limited compared to the other 3D video formats described below.

3. VIDEO PLUS DEPTH

The next more complex format is a video plus depth (V+D) representation, as illustrated in Fig. 3. A video signal and a per pixel depth map is transmitted to the user. From the video and depth information, a stereo pair can be rendered by 3D warping at the decoder. Per pixel depth data can be regarded as a monochromatic, luminance-only video signal. The depth range is restricted to a range in between two extremes Z_{near} and Z_{far} indicating the minimum and maximum distance of the corresponding 3D point from the camera respectively. Typically this depth range is quantized with 8 bit, i.e., the closest point is associated with the value 255 and the most distant point is associated with the value 0. With that, the depth map is specified as a grey scale image. These grey scale images can be fed into the luminance channel of a video signal and the chrominance can be set to a constant value. The resulting standard video signal can then be processed by any state-of-the-art video codec.

In some cases such depth data can be efficiently compressed at 10-20% of the bit rate which is necessary to encode the color video [4], while still providing good quality of rendered views. However, for more complex depth data the necessary bit rate can reach the color bit rate. Recently, alternative approaches for depth coding based on so-called Platelets were proposed, which may perform better than state-of-the-art video codecs such as H.264/AVC [5].

The ability to generate the stereo pair from V+D at the decoder as illustrated in Fig. 3 is an extended functionality compared to CSV. It means that the stereo impression can be adjusted and customized after transmission. Also more than 2 views can be generated at the decoder enabling support of multiview displays and head motion parallax viewing within practical limits.

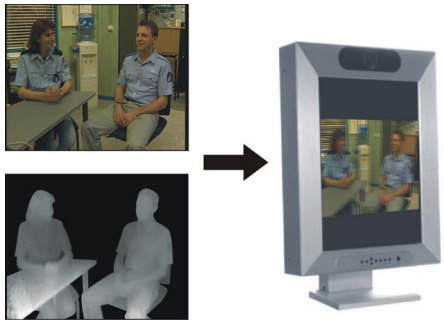

Fig. 3: Rendering of stereo video from video plus depth (V+D).

The concept of V+D is highly interesting due to the backward compatibility and extended functionality. Moreover it is possible to use available video codecs. It is only necessary to specify high-level syntax that allows a decoder to interpret two incoming video streams correctly as color and depth. Additionally, information about depth range (Z_{near} and Z_{far}) needs to be transmitted. Therefore MPEG specified a corresponding container format "ISO/IEC 23002-3 Representation of Auxiliary Video and Supplemental Information", also known as MPEG-C Part 3, for video plus depth data [6] in early 2007. This standard already enables 3D video based on video plus depth.

On the other hand the advantages of V+D over CSV are paid by increased complexity for both sender side and receiver side. View synthesis has to be performed after decoding to generate the 2nd view of the stereo pair. Before encoding the depth data have to be generated. This is usually done by depth/disparity estimation from a captured stereo pair. Such algorithms can be highly complex and are still error prone.

4. MULTIVIEW VIDEO PLUS DEPTH

Nevertheless, advanced 3D video applications exist that are not sufficiently supported by any existing standard. This includes wide range multiview autostereoscopic displays and free viewpoint video, where the user can chose an own viewpoint. Such advanced 3D video applications require a 3D video format that allows rendering a continuum of output views or a very large number of different output views at the decoder. MVC does not support a continuum and is inefficient if the number of views to be transmitted is large. V+D supports only a very limited continuum around the available original view since view synthesis artifact increase dramatically with the distance of the virtual viewpoint. Therefore MPEG started an activity to develop a new 3D video standard that would support these requirements [7]. It is based on a multiview plus depth (MVD) format, which combines multiview video with multiple depth maps.

MVD involves a number of highly complex and error prone processing steps as illustrated in Fig. 4. Depth has to be estimated for the N views at the sender. N color with N depth videos have to be encoded and transmitted. At the receiver the data have to be decoded and the virtual views have to be rendered.

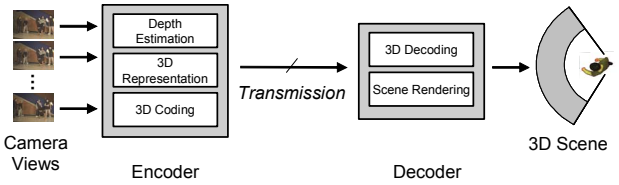

Fig. 4: MVD processing chain.

MVD can efficiently support multiview autostereoscopic displays as illustrated in Fig. 5. A display is used that shows 9 views (V1-V9) simultaneously. From a specific position a user can see only a stereo pair of them (Pos1, Pos2, Pos3). This depends on the actual position. Transmitting these 9 display views directly, e.g. using MVC, would be very inefficient. Therefore, in this example only 3 original views V1, V5, and V9 are in the decoded stream together with corresponding depth maps D1, D5, and D9. From these decoded data the remaining views can be synthesized by depth image based rendering (DIBR) [8].

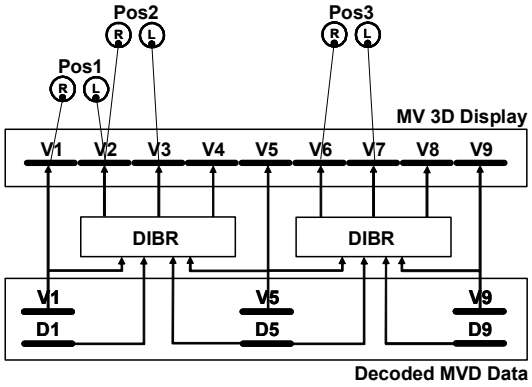

Fig. 5: MVD format and view synthesis for efficient support of multiview autostereoscopic displays.

This new standardization activity clearly targets high-quality, high-resolution, thus high-bitrate and high-complexity home user applications. General usage might be realized via a layered, scalable representation where a base layer (e.g. one color video and one depth map, perhaps at limited resolution) is accessible for low complexity devices without having to cope with the whole signal.

5. LAYERED DEPTH VIDEO

Layered depth video (LDV) [9] is a derivative and alternative to MVD. One type of LDV uses one color video with associated depth map and a background layer with associated depth map. The background layer includes image content which is covered by foreground objects in the main layer. This is illustrated in Fig. 6. Other types of LDV include one color video with associated depth as main view together with one or more residual layers of color and depth. The residual layers include data from other viewing directions, not covered by the main view. LDV supports rendering of virtual views and therefore multiview autostereoscopic displays similar to the concept illustrated in Fig. 5.

Fig. 6: Layered depth video (LDV).

LDV can be generated from MVD by warping the main layer image onto other contributing input images (e.g. an additional left and right view). By subtraction it is then determined which parts of the other contributing input images are covered in the main layer image. These are then assigned as residual images and transmitted while the rest is omitted (Fig. 6 right).

LDV might be more efficient than MVD because less data have to be transmitted. On the other hand additional error prone vision tasks are included that operate on partially unreliable depth data.

This may increase artifacts. Further over blending is not possible as with MVD data which might also reduce quality. Which one of the two methods - MVD or LDV - is favorable in which case is still to be determined, including comparison of virtual view rendering for different cases (multiview autostereoscopic displays, 2 view stereo displays).

6. DEPTH ENHANCED STEREO

The previous sections have shown that a variety of different 3D video formats exists. This situation is hampering the success of 3D video technology to some extend since it creates insecurity about compatibility of systems and content. Ideally, a generic, flexible and efficient format should be defined that would support all applications and systems. Content creation should be decoupled from the display systems, which is currently not the case.

MVD and LDV were introduced to support advanced 3D video applications such as multiview autostereoscopic displays, via virtual view rendering. In general depth-based approaches provide the flexibility of display adaptation, i.e. adjustment of the stereo baseline to the actual viewing conditions (e.g. cinema, TV, mobile). Note, that stereo content that is produced for cinema applications will look completely different on a TV-sized display.

On the other hand, there is a clear trend in industry towards conventional stereo. 3D cinema and TV content is being produced directly in this format. First home user systems are based on conventional stereo as well. It can be expected that conventional stereo will be established in the market in the near future.

We therefore propose a concept which we call depth enhanced stereo (DES) as generic 3D video format. As illustrated in Fig. 7 it extends conventional stereo. With that it provides backward compatibility. Any conventional stereo system can make direct use of the available original views. If they were produced for this type of display (e.g. cinema) best possible quality is guaranteed. Additional depth and possibly occlusion layers then provide all extended functionality (baseline adaptation, post production, N-view synthesis). Content production is decoupled from display. A scalable representation is of course possible as well, e.g. adding more than 2 views, omitting depth, occlusion, etc. With that, DES combines the important features of all other basic 3D video formats. A highest quality stereo pair is included but all advanced functionalities that rely on depth-based view synthesis are supported as well.

7. CONCLUSIONS

3D video will enter the home user mass market in the near future. Very likely initial systems will be based on conventional stereo video. However, this will not support multiview autostereoscopic displays and other advanced and necessary functionalities like baseline adaptation for different display sizes. DES is a universal 3D video format for the future. It is backward compatible to conventional stereo but also supports all advanced functionalities that rely on depth-based view synthesis. With that it decouples content creation from display and application scenario.

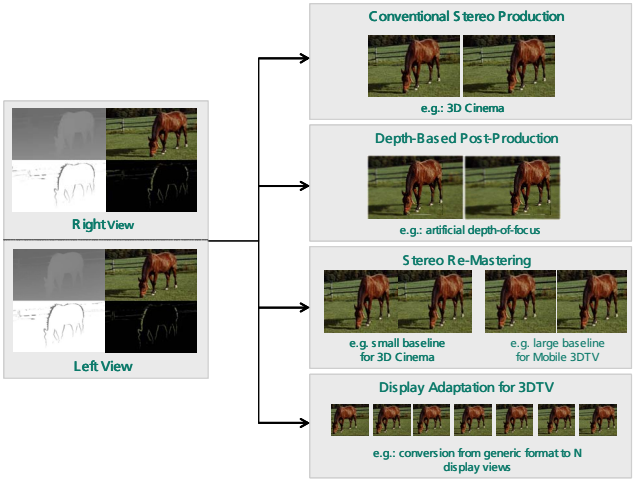

Fig. 7: Depth enhanced stereo (DES), extending high quality stereo with advanced functionalities based on view synthesis.

8. REFERENCES

[1] J. Konrad and M. Halle, "3-D Displays and Signal Processing – An Answer to 3-D Ills?," *IEEE Signal Processing Magazine*, Vol. 24, No. 6, Nov. 2007.

[2] P. Merkle, A. Smolic, K. Mueller, and T. Wiegand, "Efficient Prediction Structures for Multiview Video Coding", *Invited Paper, IEEE TCSVT*, Vol. 17, No. 11, November 2007.

[3] L. Stelmach, W.J. Tam; D. Meegan, and A. Vincent, "Stereo image quality: effects of mixed spatio-temporal resolution," *IEEE Trans. Circuits and Systems for Video Technology*, Vol. 10, No. 2, pp. 188-193, March 2000.

[4] 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.

[5] P. Merkle, Y. Morvan, A. Smolic, D. Farin, K. Mueller, P.H.N. de With, and T. Wiegand, "The Effects of Multiview Depth Video Compression on Multiview Rendering", *to appear: Signal Processing: Image Communication,* 2009.

[6] ISO/IEC JTC1/SC29/WG11, "Text of ISO/IEC FDIS 23002-3 Representation of Auxiliary Video and Supplemental Information", Doc. N8768, Marrakech, Morocco, January 2007.

[7] ISO/IEC JTC1/SC29/WG11, "Overview of 3D Video Coding", Doc. N9784, Archamps, France, April 2008.

[8] A. Smolic, K. Müller, K. Dix, P. Merkle, P. Kauff, and T. Wiegand, "Intermediate View Interpolation Based on Multiview Video Plus Depth for Advanced 3D Video Systems", *Proc. ICIP 2008, IEEE International Conference on Image Processing*, San Diego, CA, USA, October 2008.

[9] K. Müller, A. Smolic, K. Dix, P. Merkle, P. Kauff, and T. Wiegand, "Reliability-based Generation and View Synthesis in Layered Depth Video", *Proc. MMSP 2008, IEEE IEEE International Workshop on Multimedia Signal Processing*, Cairns, Australia, October 2008.