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

Development of a New MPEG Standard for Advanced 3D Video Applications

Aljoscha Smolic, Karsten Mueller, Philipp Merkle, Anthony Vetro

TR2009-068 November 2009

Abstract

An overview of available 3D video formats and standards is given. It is explained why none of these - although each useful in some particular sense, for some particular application - satisfies all requirements of all 3D video applications. Advanced formats currently under investigation, which have the potential to serve as generic, flexible and efficient future 3D video standard are explained. Then an activity of MPEG for development of such a new standard is described and an outlook to future developments is given.

6th International Symposium on Image and Signal Processing

Development of a new MPEG Standard for Advanced 3D Video Applications

Aljoscha Smolic[1]
Disney Research, Zurich
smolic@zurich.disneyresearch.com

Karsten Mueller, Philipp Merkle
Fraunhofer HHI
kmueller,merkle@hhi.de

Anthony Vetro
Mitsubishi Electric Research Labs (MERL)
avetro@merl.com

Abstract

An overview of available 3D video formats and standards is given. It is explained why none of these – although each useful in some particular sense, for some particular application – satisfies all requirements of all 3D video applications. Advanced formats currently under investigation, which have the potential to serve as generic, flexible and efficient future 3D video standard are explained. Then an activity of MPEG for development of such a new standard is described and an outlook to future developments is given.

1. Introduction

3D video shall be understood here as a type of visual media that provides the user with a depth perception of the observed scenery. This is achieved by specific 3D display systems that ensure that a determined different view is projected into each eye of the user. From such a proper stereo pair the brain then computes the depth perception. In fact the principle relies on a sensitive fake of the human visual system that can lead to uncomfortable sensation if the content is not produced carefully [1].

Currently 3D video is entering broad and most probably sustainable mass markets. Cinemas are being continuously upgraded to 3D, which is relatively easy for theatres which are already equipped with digital technology. Content creators in Hollywood and elsewhere are producing more and more movies in 3D and available material is being converted from 2D to 3D. Unlike in previous attempts, technology is now maturated providing excellent quality. Artists learnt their lessons as well and know how to produce 3D cinema content which is not overstressing the human visual system with too much exaggerated 3D effects. The 3D cinema chain is in place – including all other elements like coding, distribution, etc. – and creates substantial revenues already.

With that 3D video also becomes increasingly interesting for home entertainment or mobile applications. This includes movies distributed e.g. via Blu-ray disc or video-on-demand, as well as TV broadcast via various distribution channels. In particular different sports productions in 3D recently gained a lot of attention.

For home entertainment and mobile applications interoperability and compatibility become crucial issues. This can be achieved by standardized media formats for representation, coding and transmission. In particular decoupling of content creation from display technology has to be achieved. This means that production is free to use any technology and workflow as long as it produces content in a standardized format. Then any application will know how to use it. Additionally, backward compatibility to existing transport mechanisms and interfaces is a highly desirable feature for the success of new 3D video mass markets.

Currently, various types of 3D displays are available and under development [1]. 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. Different input formats and interleaving patterns are used in various solutions. Then there are so called 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. Such displays use even other input formats, interleaving and in most cases more than 2 views.

As a consequence, a lot of different 3D video formats are available, most of them related to specific display types [2]. The choice of a certain representation format is essential for the design of the whole processing chain. It widely determines capture systems, sender side signal processing, coding, and rendering. Having this determined by the display type causes inflexibility of the whole 3D chain.

We give an overview of available 3D video formats, associated coding algorithms and standards in Section 2. While these can be used to immediately introduce 3D video in the market, none of them satisfies all requirements on a 3D video format in an efficient way, which is hindering the development of mass user markets.

Another basic problem is the adaptation of the 3D video content to the actual display conditions. The 3D impression varies with the viewing position, display resolution, distance to the screen, etc. This is very similar to stereo audio which also only provides the correct impression in one specific point in the room. Therefore will stereo content, which is produced with a

[1] 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.

fixed camera setting and optimized for cinema, look different on a home TV or on a mobile device. In order to adapt the impression, stereo parameters (e.g. baseline, depth range) have to be flexible and not fixed by production.

A generic, flexible, universal and efficient 3D video format and standard, which would decouple content creation from display will therefore support view synthesis from the received and decoded data, in order to support any type of display and to adapt to specific technical conditions and user preferences. It will be efficient if it uses a minimum amount of necessary data to achieve high quality in a certain operating range (i.e. distance of virtual views from available original views). It will build on and extend available infrastructure and technology to ensure maximum compatibility.

Currently there are different formats and associated algorithms under investigation that could support these requirements. This includes multiview video plus depth (MVD) and layered depth video (LDV). Depth enhanced stereo (DES) was proposed as a specific configuration in this context. We summarize these advanced 3D video formats in Section 3.

ISO-MPEG is a major institution that creates specifications of media standards. MPEG audio and video standards enabled the digital media revolution over the last decades. This also includes a variety of standards that enable 3D video. Recently MPEG started an activity to develop a generic 3D video standard as outlined above [3]. In section 4 we give an overview of this activity and the current status. We describe the vision and requirements of the new standard, including the application areas to be supported. Selected test data and the experimental framework including reference software for depth estimation and view synthesis are outlined. The current status of experiments is presented and the applied evaluation of 3D video quality. Then the further work towards a Call for Proposals and the final international standard is outlined. Finally, we conclude this paper in section 5.

2. Available 3D video formats and standards

In this section we summarize 3D video formats and associated algorithms and standards that are already widely established. They can readily be used to implement 3D video applications and systems, however, they also suffer from certain restrictions and drawbacks.

2.1. Stereo simulcast and interleaving

The simplest choice to represent 3D video is to use 2 video signals that correspond to the human eye positions. This 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.

A simple way to use existing video codecs for stereo video transmission is to apply temporal or spatial interleaving [3]. With spatial interleaving, resolution is slightly lowered so that the data from left and right views could be packed into a single frame. There are various ways of arranging the data, e.g., a side-by-side format in which the right view is squeezed into the right side of the frame and the left view into the left side of the frame, or a top-bottom format in which left and right views are squeezed into the top and bottom of a frame, respectively. The data of each view may also be filtered using a quincunx sampling (or checkerboard format) and interleaved; the samples may also be packed into one of the other arrangements. With time multiplexing, the left and right views would be alternating in time, with a reduced temporal resolution for each view. An amendment to H.264/AVC is being developed that signals the new frame packing arrangements; temporal multiplexing was already enabled by the Stereo SEI message (see section 2.2). This signaling could be used at the receiver to de-interleave the video and render stereo to the display. Of course, legacy devices without knowledge or the interleaving format or the new signaling will not be able to perform the de-interleaving and hence such video encoded in this format is not usable for those devices. The simplicity and compatibility to existing infrastructure makes stereo interleaving formats very attractive for fast market introduction. Such approaches are being considered by various industry forums and standards organizations.

A new approach for efficient stereo video coding which was proposed recently is derived from the so called binocular suppression theory [4]. This is illustrated in Figure 1. Subjective tests 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.

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 [5].

Figure 1. Stereo image pair with low-pass filtered right view.

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

2.2. Inter-view prediction and MVC

Coding efficiency can be increased by combined temporal/interview prediction as illustrated in Figure 2. 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 Figure 2.

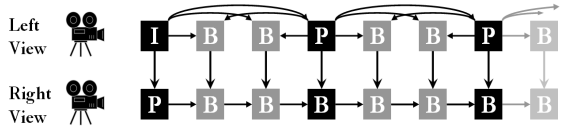

Figure 2. Stereo coding, combined temporal/ interview prediction

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 [6]. 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.

2.3. Video plus depth and MPEG-C Part 3

The next more complex format is a video plus depth (V+D) representation, as illustrated in Figure 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 in a logarithmic scale, 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 [7], 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 [8].

The ability to generate the stereo pair from V+D at the decoder as illustrated in Figure 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.

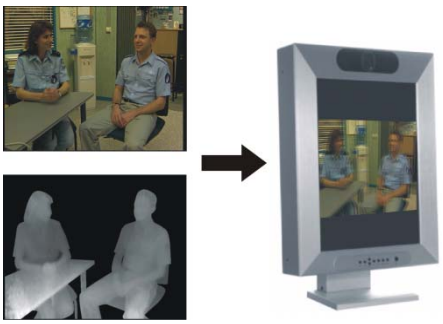

Figure 3. Rendering of stereo video from video plus depth (V+D)

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 [9] 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.

3. Advanced 3D video formats

3D video formats presented in the last section are ready to use, however, they do not satisfy all requirements in an efficient way. This includes wide range multiview auto-stereoscopic displays and free viewpoint video, where the user can chose an own viewpoint. Display adaptation is possible but very limited with V+D. 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.

The basic concept of advanced 3D video formats is illustrated in Figure 4. The decoder receives a coded representation (bitstream) of data in the advanced 3D video format. The data is decoded and then used for rendering of arbitrary views within a certain operating range. With that all requirements are satisfied. At the encoder a real world 3D scene is typically captured by multiple cameras, and a 3D representation is extracted from the multiple camera signals. Most formats under study use multiple depth maps for 3D scene representation [10], [11], but other approaches based on 3D meshes, point clouds, quads and other geometric primitives are under study as well. Sender side processing thus includes depth estimation or another kind of 3D reconstruction. Any appropriate technology and processing can be applied (including synthetic creation, animation) as long as the output is in the correct 3D video format. With that the required decoupling of content creation and display is achieved.

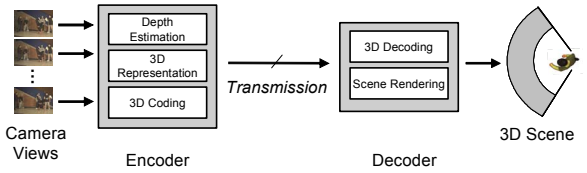

Figure 4. Advanced 3D video processing chain

The advanced 3D video processing chain involves a number of highly complex and error prone processing steps, such as depth estimation or some other 3D reconstruction at the sender, and rendering of virtual views at the receiver. In the following sections we outline some examples of such advanced 3D video formats, which can be regarded as extensions and combinations of basic formats introduced in Section 2. They have in common that they use per pixel depth maps as introduced in Section 2.3, video and other data, as in focus of the MPEG activity, which is described in Section 4.

3.1. Multiview video plus depth

Multiview video plus depth (MVD) can be regarded as extension of V+D, and fully conforms to Figure 4 [10], [11]. Efficient support of multiview auto-stereoscopic displays is illustrated in Figure 5 [12]. 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). Advanced view synthesis algorithms help to reduce rendering artifacts [10], [13].

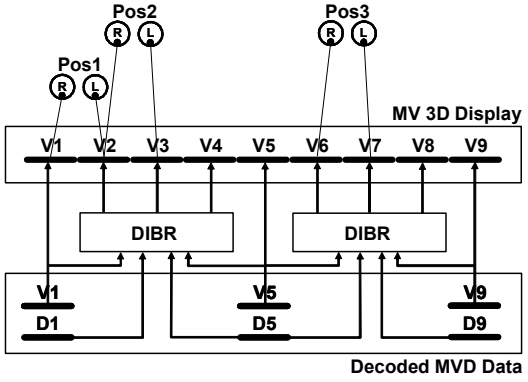

Figure 5. MVD format and view synthesis for efficient support of multiview auto-stereoscopic displays.

In order to cope with different receiver capabilities a layered, scalable representation could be used, 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.

3.2. Layered depth video

Layered depth video (LDV) [14] 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 Figure 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 multiview auto-stereoscopic displays similar to the concepts illustrated in Figures 5 and 6.

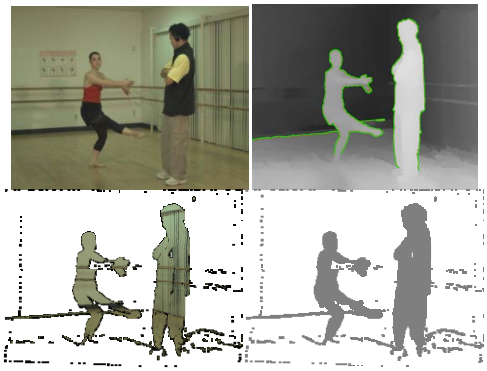

Figure 6. Layered depth video (LDV)

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 pixel-wise comparison it is then determined which parts of the other contributing input images are already covered by the main layer image. The remaining parts are then assigned as residual images and transmitted while the rest is omitted.

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 auto-stereoscopic displays, 2 view stereo displays).

3.3. Depth enhanced stereo

MVD and LDV were introduced to support advanced 3D video applications such as multiview auto-stereoscopic 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.

Therefore a concept called depth enhanced stereo (DES) was proposed as generic 3D video format [2]. As illustrated in Figure 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.

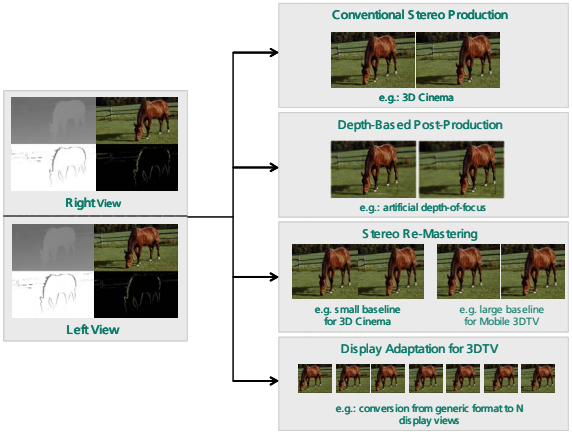

Figure 7. Depth enhanced stereo (DES), extending high quality stereo with advanced functionalities based on view synthesis.

4. MPEG activities towards a new 3D video standard

4.1. Vision and requirements

As discussed in the previous sections, MPEG has developed a suite of international standards to support various 3D services. Based on evolving market needs, MPEG is now considering a new phase of standardization. The targets of this new initiative are illustrated in Figure 8 [15].

One objective is to enable stereo devices to cope with varying display types and sizes, and different viewing preferences. This includes the ability to vary the baseline distance for stereo video so that the depth perception experienced by the viewer is within a comfortable range. Such a feature could help to avoid fatigue and other viewing discomforts.

A second target is to facilitate support for high-quality auto-stereoscopic displays. Since directly providing all the necessary views for these displays is not practical due to production and transmission constraints, the new format aims to enable the generation of many high-quality views from a limited amount of input data, e.g. stereo and depth.

A key feature of this new 3D video (3DV) data format is to decouple the content creation from the display requirements, while still working within the constraints imposed by production and transmission. Furthermore, compared to the existing coding formats, the 3DV format aims to enhance 3D rendering capabilities beyond V+D, while not incurring a substantial rate increase. Simultaneously, at an

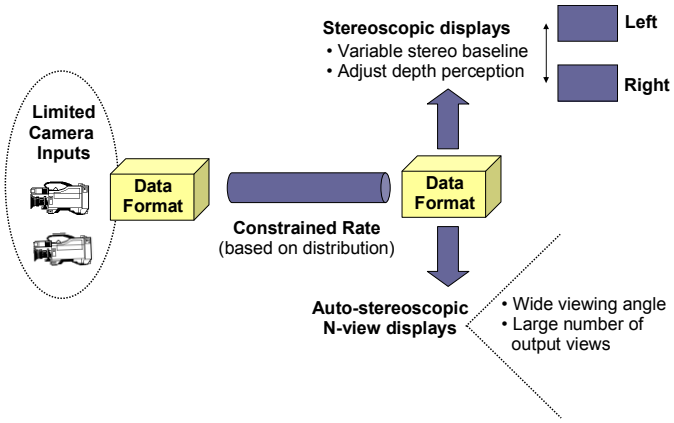

Figure 8. Target of 3D video format illustrating limited camera inputs and constrained rate transmission according to a distribution environment. The 3DV data format aims to be capable of rendering a large number of output views for auto-stereoscopic N-view displays and support advanced stereoscopic processing.

equivalent or improved rendering capability, this new format should substantially reduce the rate requirements relative to sending multiple views directly with an MVC or simulcast format. These requirements are outlined in [16].

4.2. Test data and experimental framework

The selection of an appropriate set of test data as well as the configuration of a suitable experimental framework is essential in the process of developing a new standard. The conditions for setting up test data and experimental framework are given by the vision and requirements of 3D video coding as described in the previous section.

For the new 3DV standard eleven multiview test data sets have been chosen according to the following requirements: All data sets have a linear camera arrangement, are rectified, per-pixel depth data for each view and camera parameters are provided for view synthesis and rendering. All material is progressive (no interlaced data). The test data sets cover a representative range of scene content complexity (e.g. in- and outdoor scenes), resolutions (720x540 – 1280x960 pixel), frame rates (16.7 – 30 fps), and number of cameras (3 – 80 cameras with 3.5 – 20 cm spacing between two neighboring cameras).

Unlike classic video coding standards, a more comprehensive experimental framework is required for 3D video coding approaches. Given the captured and rectified multiview video sequences on the input side and different display systems (especially stereoscopic and auto-stereoscopic multiview) on the output side, the experimental framework needs to provide reference software for both depth estimation and view synthesis. These two modules are required in the experimental framework for establishing 3D video coding processing chain. The reference software for depth estimation is necessary at the encoder side of the processing chain as scene depth information is usually not available for natural content. With this module the per-pixel depth data is generated from the multiview video input data. Unlike for classic video, such depth enhanced 3D video formats cannot be directly displayed on the decoder side, but require view synthesis for supporting different displays. The view synthesis reference software takes the video and depth information and generates the appropriate virtual views via depth-image-based-rendering (DIBR). In contrast to all previous coding approaches, interdependencies between the pure coding approach and the depth estimation and view synthesis exist and need to be considered for the overall 3D video solution.

To cope with this, one fundamental requirement regarding the experimental framework for the standardization process of 3D video coding algorithms shall ensure the best possible (ideally perfect) quality for uncompressed data. Therefore the reference software for depth estimation as well as view synthesis are integral components of the experimental framework, enabling a fair and realistic quality evaluation of different coding algorithms.

4.3. Experiments and evaluation

To evaluate depth estimation and view synthesis reference software, evaluation experiments have been carried out. For the advanced approaches in 3D Video, intermediate views are synthesized at positions, where mostly no original data is available. However, even if original data exist (e.g. due to dense camera recording for test purposes), pixel-based comparison methods like MSE-based PSNR are not useful. The classical example here is an image, shifted by one pixel and compared to its unshifted version: Although perfect in visual quality, its PSNR-value degrades dramatically [17]. Also, temporal inconsistencies, which cause flickering, are not considered. Therefore, evaluation concentrates on subjective testing, which was also provided in the context of previous standardization activities, like multi-view video coding.

For the experiments, a test room was set up with 5mx10m. Furthermore, it was darkened to avoid outside reflections. Two types of displays are used: one stereoscopic display based on polarized or shutter glasses and one auto-stereoscopic display. For the latter,

the currently available 9 view display was used. The displays were placed on tables (table height 80cm).

In the evaluation, groups of optimally 5 persons participated. The participants used chairs, as well as freely walked through the room to get viewing impressions from different viewpoints and distances. In general, a viewing distance of 3m minimum was provided. In case of the auto-stereoscopic displays, participants had to find the appropriate viewing positions for correct 3D impression. To minimize distraction from the 3D content under test caused by 3D viewing sensation, experts in 3D viewing are taken as test subjects.

In the current status of standardization, the quality of uncoded synthesized intermediate views is evaluated to judge the quality of original test data and depth maps, as well as the depth estimation and view synthesis algorithms. For this, a stereo pair with one original and one synthesized view, as well as a stereo pair of two synthesized views is presented on the stereoscopic display. In a real scenario, the first stereo pair is more realistic for stereo displays, while the second stereo pair is typical for multi-view displays.

4.4. Towards a "Call for Proposals"

Once, the depth maps and view synthesis for the test sequences have been found good enough for viewing, anchor coding for the current 3D formats, i.e. MVD and LDV will be provided. In the case of MVD, multi-view color data and associated depth maps will be coded separately using MVC to exploit inter-view dependencies. For LDV, the main and background layer for color, as well as main and background layer for depth are also coded with MVC. The anchor coding is carried out for different quality levels and will serve as a reference for future novel coding proposals.

Such proposals are sought by a "Call for Proposals" (CfP), where proponents can implement new technology and finally need to submit a coded 3D Video representation and possibly new view synthesis algorithm.

The proposals will be evaluated, using equipment and test conditions, similar to those described in 4.3. Due to the high quality range of the proposals, the Single Stimulus MultiMedia (SSMM) test method will most likely be selected. It is well known that all test methods are more or less affected by the order of presentation of the material.[2] This effect is particularly strong in the SS category test methods where no reference is present. To reduce this effect SSMM is designed to present twice any condition under test to the subjects. This allows minimizing the contextual effect.

Finally, the best 3D Video coding and view synthesis technology will be selected as reference for further development of a new 3D video coding standard. After evaluation of the Call for Proposals the collaborative phase of standard development will start. This will include improvement, extension, or replacement of algorithms in comparison to the actual reference until best possible technology is defined. Further detailed textual specifications, performance evaluations, documentations, reference software, and conformance bitstreams will be developed. Typically the collaborative phase takes 2-3 years, so that the 3DV standard can be expected in 2012 or after.

[2] This effect is known as "Contextual effect" and influences all test methods. It is caused by human short term memory. Test persons tend to be more relaxed in their judgement when two subsequently shown test sequences have similar quality. Otherwise when two subsequent conditions have highly different quality the judgement is not as fair as when two subsequent conditions have quite the same quality.

5. Conclusions

This paper reviewed the available 3D video formats, and associated standards. The merits and drawbacks of each have been discussed, and a new standards initiative has been introduced. The new 3DV format aims to support advanced stereoscopic processing as well as future auto-stereoscopic displays.

There is significant activity towards defining 3D formats for various distribution environments. It will be interesting to see whether these formats are harmonized across different domains. Hopefully, any early decisions for near-term markets will consider a migration path towards higher quality and more powerful 3D formats that are on the horizon.

6. 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] A. Smolic, K. Müller, P. Merkle, P. Kauff, and T. Wiegand, "An Overview of Available and Emerging 3D Video Formats and Depth Enhanced Stereo as Efficient Generic Solution", *Proc. Picture Coding Symposium (PCS) 2009*, Chicago, IL, USA, May 2009.

[3] A. Vetro, S. Yea, and A. Smolic, "Towards a 3D Video Format for Auto-Stereoscopic Displays", *Proc. SPIE Conference on Applications of Digital Image Processing XXXI*, Vol. 7073, September 2008

[4] 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.

[5] H. Brust, A. Smolic, K. Müller, G. Tech, and T. Wiegand, "Mixed Resolution Coding of Stereoscopic Video for Mobile Devices", *Proc. 3DTV-CON 2009*, Potsdam, Germany, May 2009.

[6] 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.

[7] 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.

[8] 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", *Signal Processing: Image Communication (2008)*, doi:10.1016/ j.image.2008.10.010.

[9] 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.

[10] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-Quality Video View Interpolation Using a Layered Representation", *ACM SIGGRAPH and ACM Trans. on Graphics*, Los Angeles, CA, USA, August 2004.

[11] P. Kauff, N. Atzpadin, C. Fehn, M. Müller, O. Schreer, A. Smolic, and R. Tanger, "Depth Map Creation and Image Based Rendering for Advanced 3DTV Services Providing Interoperability and Scalability", *Signal Processing: Image Communication. Special Issue on 3DTV*, February 2007.

[12] 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.

[13] K. Mueller, A. Smolic, K. Dix, P. Merkle, P. Kauff, and T. Wiegand, "View Synthesis for Advanced 3D Video Systems", *EURASIP Journal on Image and Video Processing*, Volume 2008, doi:10.1155/2008/438148.

[14] 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 International Workshop on Multimedia Signal Processing*, Cairns, Australia, October 2008.

[15] Video and Requirements Group, "Vision on 3D Video", ISO/IEC JTC1/SC29/WG11 N10357, Lausanne, CH, February 2009.

[16] Video and Requirements Group, "Applications and Requirements on 3D Video Coding", ISO/IEC JTC1/SC29/WG11 N10570, Maui, US, April 2009.

[17] Z. Wang and A. C. Bovik, "Mean Squared Error: Love it or leave it?", *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98-117, Jan. 2009.