ENABLING INTRODUCTION OF STEREOSCOPIC (3D) VIDEO: FORMATS AND COMPRESSION STANDARDS

W.H.A. (Fons) Bruls, C. Varekamp, R. Klein Gunnewiek, B. Barenbrug, A. Bourge

Philips Research Laboratories, The Netherlands

ABSTRACT

After the introduction of HDTV, the next expected milestone is stereoscopic (3D) TV. This paper gives a summary of the new MPEG-C part 3 standard, capable of compressing the 2D+Z format, and shows how it can be used to serve the 1st generation of 3DTVs. Furthermore it gives directions on how this standard could be extended to serve also the generations beyond.

***Index Terms*—** 3DTV, depth, LDI, compression, MPEG-C

1. INTRODUCTION

At various events it has been demonstrated that 3D displays, auto-stereoscopic (lenticular & barrier) displays without glasses or beamers with glasses, are about ready to hit the market. It also has been shown [1] that image plus depth (2D+Z) is a suitable data format for 3D stereoscopic video, and that depth information can be compressed efficiently using existing video codecs if the depth is represented using the luminance signal.

In this paper we discuss the 2D+Z format, and how it has been standardized in MPEG-C part 3. The goal of this standard is to provide a stereoscopic viewing experience from one given viewpoint. This should not be confused with the goal of multiview coding (MVC) [2], targeting at free point (3D)TV, where a wide range of viewpoints should be covered.

In section 4 we present advanced hole filling strategies to stretch the capabilities of the standard to its limits and we discuss options to extend the standard.

2. 2D+Z FORMAT

While Figure 1 gives an example of a specific type of 3D display, different display realizations vary largely in: (a) the number of views that are represented; (b) the maximum depth that can be displayed. An input format is therefore required that is flexible enough to drive all possible variants. This can be achieved by supplying a depth or parallax value with each pixel of a regular video stream, and by generating all the required stereoscopic views at the receiver side. This is an interesting alternative to submitting all the required views which would be expensive in terms of bitrate. Submitting a small subset of views (eg. 3 views using MVC) is also not an option since it would require both disparity estimation and view interpolation at the display. The additional bitrate (compared to 2D) using MPEG-C part 3, can be limited to some few tens of %. Using MVC, every additional coded view already costs approx. 40% overhead.[2]

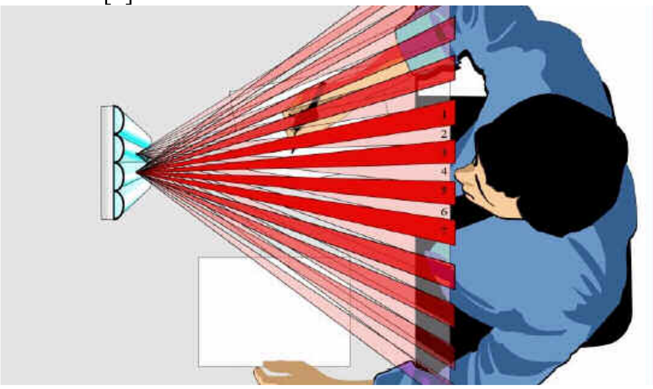

Fig.1 Example of lenticular auto-stereoscopic display requiring 9 views

2.1. RENDERING NEW VIEWS

The rendering process is important for the stereoscopic image quality. In this section we describe how a new view is created given a centre view and a depth map. A given view on the 3D display is determined by the set of rendering parameters. For the results shown in this paper we approximate the relation between the horizontal parallax of a newly rendered view and the depth using a linear relation:

$$p(x,y) = \frac{p_{\max}}{255} d(x,y) - \frac{p_{\max}}{2} \quad (1)$$

where $p_{\max}$ is the maximum horizontal (screen) parallax. Equation (1) will result in an equal amount of depth in front and behind the display surface. For the actual rendering that we use in this paper, the steps are described in [3, p.120] in combination with a technique known as supersampling [3, p.168]:

1. Signal reconstruction: Each colour and parallax sample is represented by 4 equally spaced points with the same value.

2. Signal warping using equation (1): a "back-to-front" ordering [4, p. 45] is used when warping points from the reference to the output grid. This means scanning with a decreasing x-coordinate to create a left-eye view and with an increasing x-coordinate to create a right-eye view. Filling the holes between the warped samples in the output grid, is done by repeating along the scanning direction.

3. Signal prefiltering: The warped signal is filtered using a box-filter of 4 subpixels width.

4. Signal sampling: The signal is sampled on the output grid.

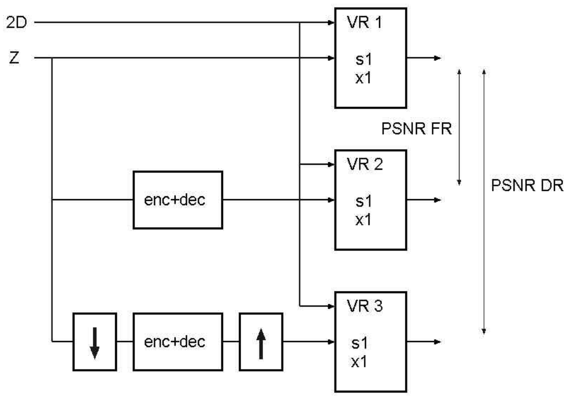

Fig.2.1 Blockdiagram determining impact of depth coding

2.2. IMPACT OF COMPRESSION ON Z

In this chapter we will investigate the effect of compression of depth or screen parallax on the rendered stereoscopic views and also investigate the option of subsampling the depth map to achieve very low bitrates.

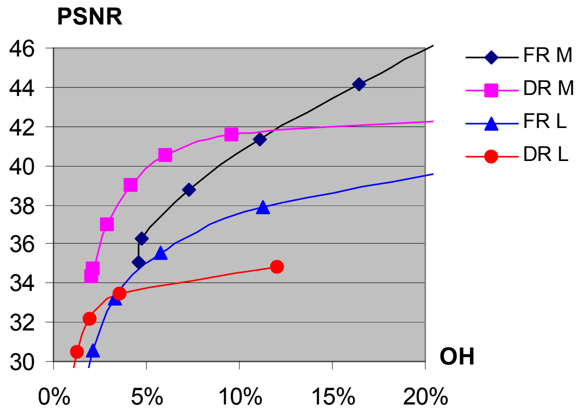

Fig.2.2 PSNR comparison according to blockdiagram fig 2.

We investigated the effect of compression by calculating the PSNR [dB] for a rendered view using $p_{\max}$ = 24 pixels (rendering parameter s1 of fig 2.1). This experiment was done for a depth map at full resolution (FR) and for the depth map at a factor two downscaled resolution (DR). Fig. 2.1 shows the blockdiagram for this experiment. VR stands for view renderer. The results for two different sequences (“M” & “L”) are given in fig. 2.2. For compression H264 was used with Qp parameters ranging from 20 to 60. The overhead (OH) was defined by:

$$\mathrm{OH} = 100 \frac{\text{bitrate depth at variable Qp}}{\text{bitrate 2D video at Qp} = 30} \quad [\%] \quad (2)$$

The results are given in fig. 2.2 and show that low bitrates yield already good quality and also show that downscaling the depth indeed is very beneficial in achieving a very low overhead.

3. MPEG-C PART 3

To enable interoperability between content providers, broadcasters and display manufacturers, standardization of a format for compressing 2D+depth video is needed.

The new MPEG-C part 3 standard is designed to meet the following important requirements:

- Low 3D overhead in terms of bitrate
- Display independence
- Backward compatibility with (2D) video standards
- Re-use of existing standards as much as possible
- Enable for adjustable depth-effect at display
- Depth range that supports comfortable viewing

The depth or screen-parallax (Z) data is compressed as conventional luminance signals using already existing (MPEG) video codecs and as an auxiliary video stream multiplexed together with the AV streams as shown in fig 3.1. This approach even allows for optional subsampling of this data in both the spatial (as used in fig 2.1) and temporal domain. This can be beneficial, depending on the particular application and its requirements, and allowing very low bitrates for the extra data of Z (<<10% of the 2D rate).

For correct 3D rendering, two further important considerations are: synchronization of 2D and depth, and normalization of depth values (dynamic range). They are both handled at MPEG-Systems level: the former through the use of so-called timestamps, the latter through a new metadata set included in the stream descriptor. Moreover, to ensure backwards compatibility with existing 2D services, the auxiliary video stream has a special stream type code 0x1E, and its own descriptor indicating the compression format (e.g. H.264/AVC), its type (e.g. 0x00 for depth), and its metadata.

In the case of depth, this metadata consists of 2 parameters K_{near} and K_{far}, defining how near to the viewer and how far from the viewer objects will maximally pop out of the display. When a parallax map is used instead of a depth map, then the correct 3D rendering is defined with 4 parameters (zero level, gain, reference monitor width and reference viewing distance). Thus, by defining the 3D metadata from a display perspective, the whole rendering process is simplified and becomes independent of the content capturing/creation process.

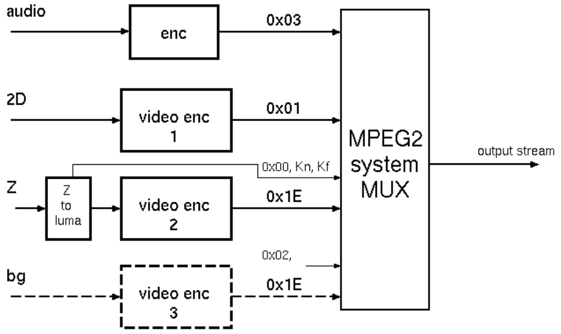

Fig.3.1 Blockdiagram MPEG-C part 3

Given the expected display capabilities of the first generation of 3D displays and quality of content conversion (limited depth quality), in practice the loss in quality due to the 2D+depth format remains small and thereby very acceptable. However, we expect that depth estimation quality will improve and that graphics will be used even more in video production than today. The new standard is very flexible and could be extended in the future with additional information in a second aux stream with a different aux video data descriptor (e.g. 0x02 for occlusion or background data), see fig 3.1.

4. IMPACT OF HOLE FILING STRATEGIES

When using 2D+Z, occlusion information is not available and the occlusion texture has to be generated by the display rendering. The warping as described 2.1 can be seen as hole-filling by repeating the last background pixel into the uncovered background. Other strategies can be applied. We evaluated (fig 4.1) different strategies by viewing the sequences on real 3D displays (auto-stereoscopic 3D displays and on shutter glasses & CRT).

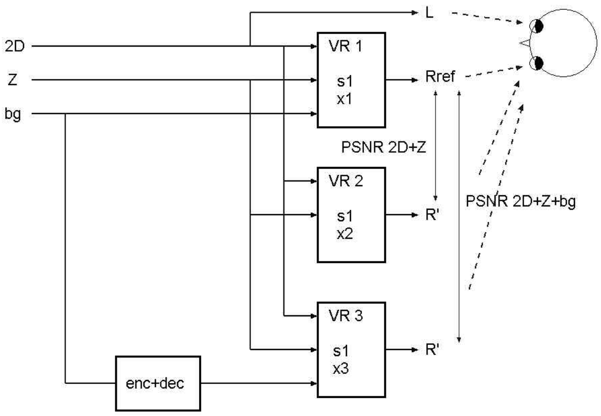

Fig 4.1 comparing 2D+Z & 2D+Z+bg

Fig 4.2 a & b give an example of the original 2D (L) and the rendered right view R' using the simple copy strategy. The R' view in itself looks relatively poor. Here the rendering parameter s1 has been enlarged to match with the expected use in future generations of displays. Also the human brain seems to fuse the views spatially and temporally together causing a kind of masking effect for the imperfect hole-filling. This matches with the viewing impressions when using real 3D displays. A more complex hole-filling strategy, based on markov random fields (MRF) [5], gives already better results as illustrated by fig 4.2.c.

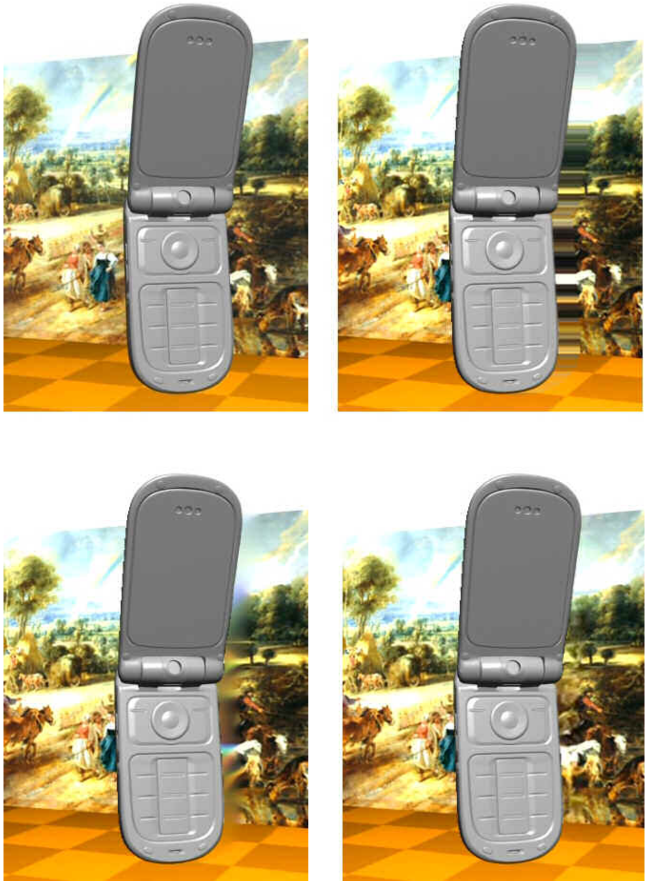

Fig 4.2 "M" sequence. a: L(=2D), b: R' x2=copy, c: R' x2=MRF, d: R' x3=coded bg @ 10% of 2D

5. EXTENDING THE 2D+Z FORMAT

As seen, hole-filling strategies can improve the visual quality, but there is a limit to it. Therefore the 2D+Z format needs to be extended for the case of stronger depth rendering. Fig 5.1 depicts in which direction we need to go to improve the quality of image based rendering. On the right hand side of Fig. 5.1 we see the Layered Depth Image format proposed by Shade et.al. [6], on the left side 2D+Z as already defined in MPEG. The next improved approximation of LDI after 2D+Z is to first represent only a single background (bg) texture layer (I_2), possibly supplemented with a bg Z (z_2).

As mentioned in chapter 3, the MPEG-C, part 3 standard is very flexible and can be extended. An interesting option is to extend it with a 2nd, besides Z, layer containing the bglayer. This allows the view-renderer (VR) to fill the holes from this bg layer. There are various scenarios on how to obtain both layers (depth and background):

- Chroma keying using a blue or green screen
- Semi-automatic extraction from 2D or stereo [7]
- from computer generated images

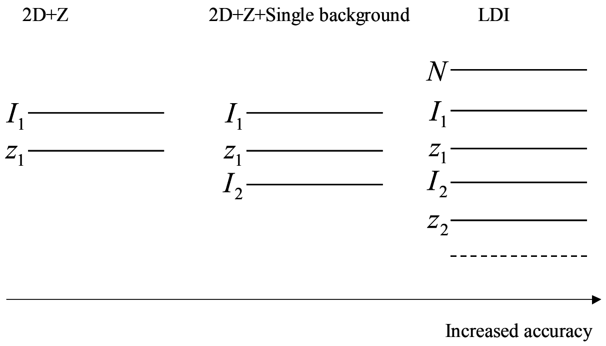

Fig.5.1 3D data formats for IBR with from left to right increased accuracy of the 3D representation.(I_1=2D).

5.1. RENDERING USING A BACKGROUND

Using the approach of section 2, signal warping is first done using I_2+ z_2 of the background and then (in a second pass) using I_1+ z_1 of the foreground. In the 2nd phase during the warping of the foreground, pixel repetition is not done for horizontal changes in depth that exceed a give threshold, thereby uncovering previously warped background texture. The final result is then prefiltered and sampled (as in steps 3 and 4 for section 2).

If only the texture of the bg (I_2) is coded then we need to approximate the corresponding bg depth (z_2) of this texture by extrapolations from Z (z_1) at the hole edges.

5.2. IMPACT OF ADDING A BACKGROUND LAYER

We compared according fig 4.2 the overall quality of newly rendered views using a background layer with the situation where no background layer is coded and hole filling is used. Since the bg layer is only used where objects are covering the bg layer, the other pixels are less important. This can be exploited and allows a significant reduction of the required bitrate for the compressed bg layer.

The resulting comparisons are illustrated by fig 4.2.

Objective comparisons are given in fig. 5.2. We see an increasing degrading of the rendered view quality when the applied screen parallax is increased for larger depth effects (larger index i). It is clearly shown that by using the bg layer, even at relatively low (~10%) additional bitrate of this layer, much larger depth effects (larger i values) become possible before the quality of the rendered views becomes a problem.

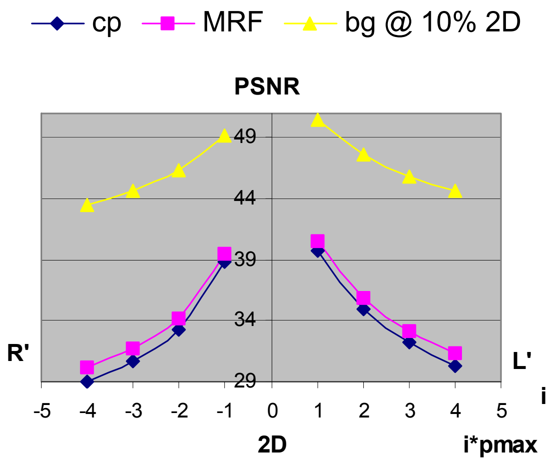

fig 5.2 comparison of 2D+Z and 2D+Z+bg (pmax=8)

6. CONCLUSIONS

With the coming availability of various types of displays, several practical ways of generating 3D content and the availability of flexible standard using 2D+Z as a basis, all ingredients are available to start with the introduction of 3D TV, thus providing the CE industry with a new impulse.

REFERENCES

[1] C. Fehn, *et al* “An Evolutionary and Optimised Approach on 3D-TV”. In Proceedings of International Broadcast Conference, pages 357-365, Amsterdam, The Netherlands, September 2002.

[2] P. Merkle, etal “Efficient compression of multi-view video exploiting inter-view dependencies based on h.264/mpeg4-avc” http://ip.hhi.de/imagecom_G1/assets/pdfs/h264_multi_view.pdf

[3] G. Wolberg. “Digital Image Warping.” IEEE Computer Society Press. 1990.

[4] L. McMillan. “An Image-based approach to three-dimensional computer graphics”. PhD thesis, University of North Carolina at Chapel Hill. 1997. (http://www.cs.unc.edu/~ibr/pubs/mcmillan-diss/mcmillan-diss.pdf).

[5] A. Efros and T. Leung. “Texture synthesis by non-parametric sampling.” IEEE Int Conf on Computer Vision, p 1033–1038, 1999.

[6] J. Shade, S. Gortler, L. He and R. Szeliski. Layered Depth Images. SIGGRAPH 98: Computer Graphics Proceedings, Annual Conference Series, July 1924, 1998.

[7] P. Harman, J. Flack, S. Fox and M. Dowley, “Rapid 2D to 3D Conversion”, Proc. SPIE, Vol. 4660, 2002