INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC1/SC29/WG11
CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC1/SC29/WG11 MPEG2014/M31954
January 2014, San Jose, US

Source: **Holografika, Huawei Technologies**
Title: **Requirements of Light-field 3D Video Coding**
Status: **Informative**
Authors: **Péter Tamás Kovács, Tibor Balogh, Jacek Konieczny, Giovanni Cordara**

---

1 Introduction

Recently, the third phase of Free-viewpoint Television (FTV) standardization activity was proposed [1], mainly targeting super multiview and free navigation applications. During the last meeting, a document describing first set of requirements and use cases was generated [2], and an AhG created, [3] in order to solicit further contributions from the industry and perform exploration experiments.

The current document presents some insight on the super multiview / light-field Holovizio display technology, already mentioned in [2], and the constraints imposed by light-field displays to the capturing setups and representation formats. Also, some amendments to the requirements are proposed, in order to support two application scenarios deemed of interest for the proponents.

2 Holovizio Light-field displays

The Holovizio [4] display technology falls into the super multiview category, as reported in [2]: multiple views are generated from a reconstruction of the light-field (LF) faring through a tri-dimensional scene.

HoloVizio 3D light-field displays are capable of providing 3D images with a continuous motion parallax on a wide viewing zone, and viewers can experience spatial vision inside this zone without wearing 3D glasses. Instead of showing separate 2D views of a 3D scene, they reconstruct the 3D light field as a set of light rays. This is achieved by using an array of projection modules emitting light rays and a custom made holographic screen. The light rays generated in the projection modules hit the holographic screen at different points and the holographic screen makes the optical transformation to compose these light rays into a continuous 3D view, as seen on Fig 1. Each point of the holographic screen emits light rays of different color and intensity to the various directions.

With proper software control, light rays leaving the screen spread in multiple directions, as if they were emitted from points of 3D objects at fixed spatial locations, appearing either behind the screen, or floating in the front of it, achieving an effect similar to holograms.

Advanced glasses-free 3D light-field displays provide hologram-like spatial visualization over an unprecedented field-of-view (up to 180 degrees), but operate with massive pixel counts (N x 100 megapixels today) and massive viewing directions (up to 200 today). Accordingly, the storage, compression, transmission and rendering of these light-fields is a major challenge, which needs to be solved to pave the way towards wide adoption of such advanced 3D technologies.

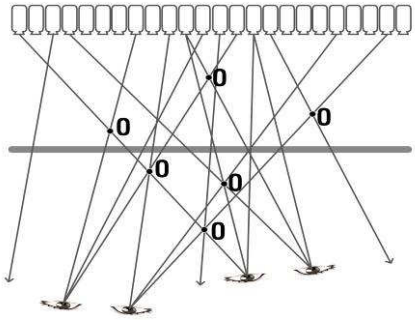

Fig. 1. Sample light-field display architecture scalable to emit 100+ viewing directions.

3 Application Scenarios

Future 3D displays will go far beyond stereoscopic 2 views or N view (N<10) multi-view displays, as we can already see from prototype and also commercially available examples.

However, the number of pixels (light rays) one can reasonably integrate into a single display is still quite limited, and display designers are thus forced to trade off horizontal viewing directions, vertical viewing directions and the spatial resolution (number of 3D pixels).

Some displays aim to reproduce full-parallax 3D light-fields (that is, having both horizontal and vertical parallax), while others omit or simplify vertical parallax in order to provide better resolution and higher number of viewing directions, resulting in wider horizontal Field Of View (FOV).We have to make a distinction between these application scenarios, as the content that needs to be encoded may seem similar but in fact they can be quite different.

Full-parallax displays typically have N x N (N<10) views, consequently narrow viewing angle in both directions, and relatively low resolution. Here the views to be transmitted can be expected to be relatively close (a few degrees apart), and thus similar to each other. These views are typically captured and parameterized on two planes (linear camera arrangement).Wide-angle LF displays on the other hand use N (N>100) viewing directions, but only in the horizontal direction (see Fig. 2). Widening the viewing direction with such a big number of views pose additional challenges for LF capturing, encoding and processing: we analyze such challenges below.

Fig. 2. 3D displays with wide viewing angle.

Need for non linear acquisition setups

We have to realize that in this case, the overlap between the leftmost and rightmost views is typically zero, as the baseline of a typical capture setup that can capture the necessary viewing angle is much wider than the scene itself. In extreme cases (180 degree viewing angle), a linear camera setup is not even sufficient or practical, and an arc camera setup is

necessary. Such a setup breaks the assumptions of parallel cameras, which in turn may require different coding strategies.

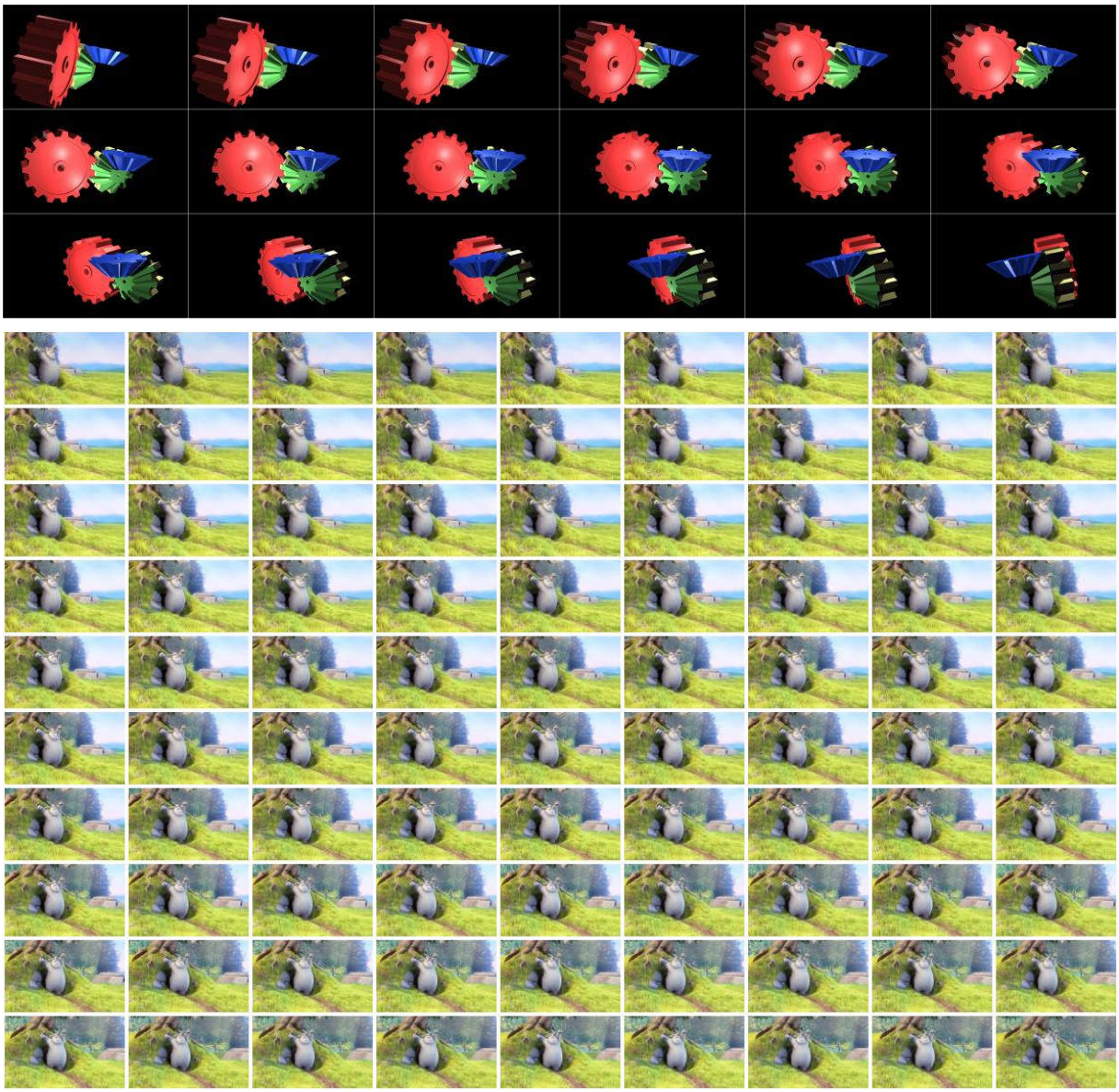

Fig. 3. Top: typical wide viewing angle content (every 10th image is shown)
Bottom: 45° 3D frame with 0.5° displacement between cameras, from Big Buck Bunny [5]

Computational complexity

Another very significant challenge to be addressed is given by the computational complexity: as 3D displays with massive pixel counts are typically driven by multiple processing units (today rendering clusters) for practical reasons, and the individual units are responsible for a small portion of the whole 3D image, the LF interpolation / rendering that is performed on one unit does not require the whole LF data. To enable decompressing only those parts which are in fact used by the rendering process (and thus save processing time by eliminating the decompression of unused parts), it is essential that the decoder enables accessing only slices of the overall compressed stream with low overhead.

Light-field displays are typically driven by multiple processing units. During the light-field conversion step, light rays captured by the cameras (views) are reordered to form a display specific light field. As different processing units generate different parts of the light field, they read different image areas from the incoming views. In fact, although there is some overlap between the inputs of neighbouring processing units, those that are responsible for distant parts of the overall light-field consume a disjoint set of pixels. The following examples illustrate which parts of the input images are actually consumed in a real setting. The simulation has been run on a HoloVizio C80 light-field display, which has 10 processing units for generating the light-field. This display has been fed with 91-camera input. Results are shown for the central camera.

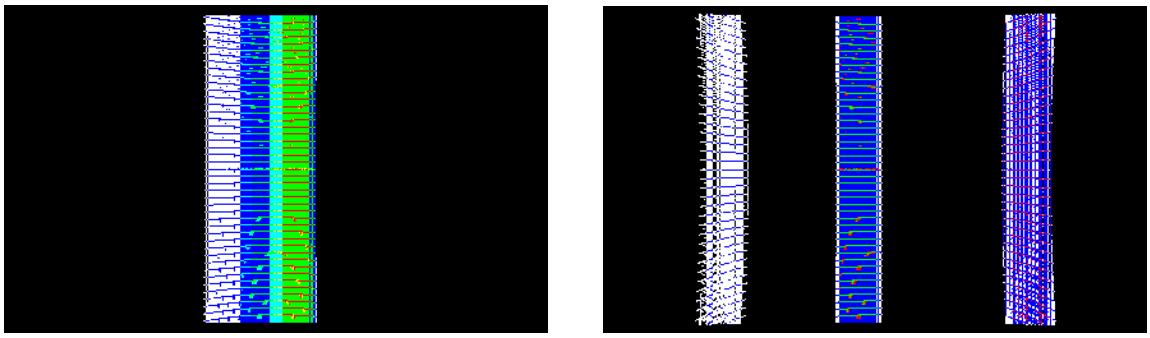

Fig. 4. Left: Pixels consumed by processing units 4, 5, 6 (Red, Green and Blue channels)
Right: Pixels consumed by processing units 0, 5, 9 (Red, Green and Blue channels)

Encoding mechanisms

For LF displays capable for larger number of views, the creation of all the views starting from a subset of captured views can be considered an inter&extrapolation task, which can be solved locally. When disparity information is available, it is beneficial to exploit such information at the encoding side, in order to obtain single geometrically correct set of motion vectors that are related to real 3D scene disparity and can be used over multiple adjacent view images. Additional motion vectors can be derived to deal with occlusions. Besides the increased compression efficiency, this can potentially speed up the generation of all the views needed by the display.

We believe that super multiview technology can have a disruptive impact on the market, enabling the real 3D viewing experience that has been object of research for long time. In particular, we would like to emphasize two applications, among those listed in [2], that, according to our vision, can revolutionize the consumer market and the common way of producing (and enjoying) video content:

- Enable light-field and super-multivew 3D televisioning on a wide range of 3D display devices, for cinema, broadcast and mobile application scenarios.
- Support 3D telepresence applications, where both encoding and decoding happens in real-time to enable smooth communication

Fig. 5. Light-field telepresence application

Dedicated coding technology is necessary in order to enable compression and transmission of such extremely high resolution content.

4 Requirements

Starting from the above mentioned considerations and targeted application scenarios, we derive some requirements, significantly overlapping those listed in [2]. For certain requirements, we propose some modifications, addressing the issues reported in previous section. For each proposed requirement, we indicate whether some modification is proposed (bolded in the proposed text below). We also propose to add a couple of new requirements, on parallel processing and random access.

4.1 Requirements for Data Format

4.1.1 Video data (proposed modifications bolded)

The uncompressed data format shall support multiple camera input and multiple camera output configurations, with a higher number of output views than input views. The input camera views along pathways different from 1D linear arrays **shall** be supported, **encompassing non-linear horizontal, horizontal/vertical and free-form (randomly placed, calibrated) camera setups**. The number of input and output views should vary between tens and hundreds of views (application examples given in Table 1) and View Synthesis should be robust against incorrectly acquired/calculated depth maps. Other input and output configurations beyond stereo **shall** also be supported.

4.1.2 Supplementary data (proposed modifications bolded)

Supplementary data shall be supported in the data format to facilitate high-quality intermediate view generation. Examples of supplementary data include depth maps, or 3D models, reliability/confidence of depth maps, segmentation information, transparency or specular reflection, occlusion data, etc. **If disparity information is available, it should be exploited to encode geometrically correct motion vectors that are related**

to real 3D scene disparity and can be used over multiple adjacent view images. Supplementary data can be obtained by any means.

4.1.3 Metadata (proposed modifications bolded)

Metadata shall be supported in the data format. Examples of metadata include extrinsic and intrinsic camera parameters **(or a shorthand notation to generate simple camera setups like linear and arc)**, scene data, such as near and far plane, **scene Region Of Interest (ROI) (3D position of intended screen plane, extents and position of the scene in all directions), to enable proper mapping to the 3D volume reconstructed by the 3D display** and others.

4.1.4 Applicability (original)

The data format shall be applicable for both natural and synthetic scenes.

4.2 Requirements for Compression

4.2.1 Compression efficiency (proposed modifications bolded)

Compression efficiency **shall** be comparable or better than the state-of-the-art video **coding** technology such as MV-HEVC or 3D-HEVC,

4.2.2 Synthesis accuracy (proposed modifications bolded)

The impact of compressing the data should introduce minimal visual distortion on the visual quality. The compression shall support mechanisms to control bitrate with proportional changes in synthesis accuracy**, and where these inaccuracies appear. For example, it may be desirable to degrade side views more in order to keep central views intact.** Increasing the ratio of output/input views and/or the input views baseline should - below a reasonable threshold - introduce minimal distortion on the visual quality of synthesized views.

4.2.3 Parallel and distributed processing (new)

The compression method shall enable parallel processing (i.e. on GPUs), as well as on parallel processing units (i.e. on multiple computers), without significant losses in the overall compression ratio compared to a single-threaded scenario.

4.3 Requirements for Decompression and Rendering

4.3.1 Rendering Capability (proposed modifications bolded)

The data format should support improved rendering capability and quality compared to existing state-of-the-art representations. The rendering range should be adjustable.

The data format should support light-field interpolation by means of ray interpolation, without relying on depth information for the synthesis.

4.3.2 Low complexity (original)

The data format shall allow real-time decoding and synthesis of views, required by any new display technology, with computational and memory power available to devices at realizable level.

4.3.3 Parallel and distributed rendering (new)

The rendering method shall enable parallel processing (i.e. on GPUs), as well as on parallel processing units (i.e. on multiple computers), which is critical to achieve real-time decoding and synthesis of views.

4.3.4 Random access (new)

The data format shall support partial decompression (random access) of some views, as well partial decompression (random access) of portions of these views. .

4.3.5 Display types (original)

The data format shall be display-independent. Various types and sizes of displays, e.g. stereo and auto-stereoscopic, super multiview, integral photography displays, etc of different sizes with different number of views shall be supported. The data format shall be adaptable to the associated display interfaces.

5 References

[1] M. Tanimoto, T. Senoh, S. Naito, S. Shimizu, H. Horimai, M. Domański, A. Vetro, M. Preda and K. Mueller, “Proposal on a New Activity for the Third Phase of FTV,” ISO/IEC JTC1/SC29/ WG11 MPEG2013/M30229, Vienna, Austria, July 2013.

[2] Mehrdad Panahpour Tehrani, Shinya Shimizu, Gauthier Lafruit, Takanori Senoh, Toshiaki Fujii, Anthony Vetro, Masayuki Tanimoto, “Use Cases and Requirements on Free-viewpoint Television (FTV),” ISO/IEC JTC1/SC29/ WG11 MPEG2013/N14104, Geneva, Switzerland, November 2013.

[3] “AHG on FTV (Free-viewpoint Television),” ISO/IEC JTC1/SC29/ WG11 MPEG2013/N140709, Geneva, Switzerland, November 2013.

[4] T. Balogh, "The HoloVizio system," Proc. SPIE 6055, Stereoscopic Displays and Virtual Reality Systems XIII, 60550U (January 27, 2006); doi:10.1117/12.650907.

[5] Big Buck Bunny, (c) copyright 2008, Blender Foundation / www.bigbuckbunny.org