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

ISO/IEC JTC1/SC29/WG11 MPEG2015/N15348
June 2015, Poland, Warsaw

Source	**Requirements**
Status	**Approved**
Title	**Call for Evidence on Free-Viewpoint Television: Super-Multiview and Free Navigation**
Editors	**Gauthier Lafruit, Krzysztof Wegner, Masayuki Tanimoto**

1 Introduction

4k/8k UHDTV (Ultra High-Definition TV) offers viewing at the highest resolution in visual media. However, it transmits only a single view and users cannot change the viewpoint. It is still far away from our viewing experience in the real world.

MPEG has been developing video coding standards enabling some forms of 3D vision. Previous stereo and multiview coding standards, such as developed in the Multi-View Coding (MVC) activities with MV-HEVC, have focused on the compression of camera views “as is”, all rendered without means to facilitate the generation of additional views.

Depth-based 3D formats (3D-HEVC) have been developed to address this shortcoming: with the use of depth image based rendering techniques, the generation of additional views from a small number of coded views was enabled, supporting auto-stereoscopic 3D display applications with tens of output views from a couple of input camera feeds. These standards were issued during the first and second phase of FTV and assume a linear and narrow baseline arrangement of camera inputs.

Recently, more advanced 3D displays – called Super-Multi-View (SMV) displays – are emerging, which render hundreds of linearly or angularly arranged, horizontal parallax ultra-dense views, thereby providing very pleasant glasses-free 3D viewing experience with wide viewing angle, smooth transition between adjacent views, and some “walk-around feeling” on foreground objects. Such displays require both a very dense and a very wide baseline set of views which results in vast number of input views that need to be transmitted. Moreover transmitted and displayed views are highly correlated, which was observed in previously considered applications. As such it puts very demanding requirements to the codec for both compression efficiency and throughput in order to handle hundreds of wide baseline, ultra-dense views.

The next challenge for MPEG is FTV (Free-viewpoint TV) [1]-[3]. FTV enables users to view a scene by freely changing the viewpoints as we do naturally in the real world. FTV is the ultimate 3DTV with an infinite number of views and ranks as the top of visual media. It provides a very realistic glasses-free 3D viewing without eye fatigue. FTV will have a great impact on various fields of our life and society.

FTV enables SMV and Free Navigation (FN) applications [4]. However, SMV displays need huge amounts of data. If compressed with the use of currently available inter-view prediction techniques, the data rate increases linearly with the number of views and beyond economically viable transmission bandwidths.

Literally surrounding the scene with a non-linear, ultra-dense array of several hundreds of cameras offers FN functionalities around the scene, similar to the Matrix bullet effect, further extending the aforementioned "walk-around feeling" to the full scene. Of course, sparse camera arrangements in large baseline setup conditions would be preferable (drastic cost reductions in content generation as well as transmission), however, smooth transitions need to be synthesized. A key challenge therefore resides in the development of novel technology that supports the generation of additional views at the decoder side, which were not already present in the encoded bitstream. Rendering of zoomed-in/out virtual views to support real "fly through the scene" functionalities at reasonable capturing and transmission costs is required.

Evidently, such 2D rendered Free Navigation (FN) applications might be combined with SMV for full-immersive viewing on 3D displays. In this case, each FN virtual viewpoint request will synthesize hundreds of linear or angularly adjacent viewpoints to feed the SMV display. This will probably require processing acceleration to cover real-time scenarios, but hardly imposes new algorithmic challenges compared to FN.

The realisation of all aforementioned FN and SMV systems requires technologies that are not currently available in MPEG. Therefore, companies and organizations that have developed compression technologies that they believe to perform better than 3D-HEVC (that has originally been designed in different application conditions), are kindly invited to bring such information to MPEG in response to this Call for Evidence (CfE).

If proposed technology significantly outperforms currently available MPEG technology, MPEG plans to issue the Call for Proposals (CfP), subsequent to this CfE, to develop standards that allow increased compression performances beyond 3D-HEVC in FN and SMV application scenarios.

The timeline for this Call for Evidence has been fixed as follows:

- Test sequences and preliminary 3D-HEVC anchors are available: 2015-06-15
- Final 3D-HEVC anchors are available: 2015-07-10
- Submission of contributions (descriptive document): 2016-02-22
- Decoded sequences, bitstreams and binary decoders are made available for the 114th MPEG meeting by 2016-02-01 (three weeks prior to the meeting)
- Evaluation of the responses at the 114th MPEG meeting (2016-02-22 – 2016-02-26)

2 Target application scenarios

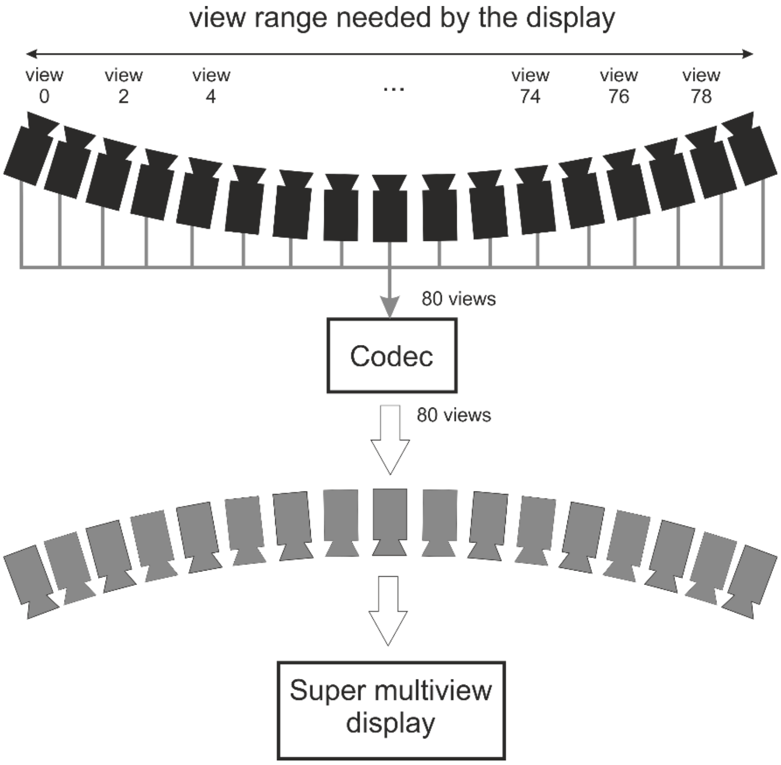
Two main applications are targeted: Super-Multiview Display (SMV) and Free Navigation (FN).

Though there exist commonalities between SMV and FN, these two categories are evaluated in a different way: SMV aims at high compression exploiting the essential information embedded in all camera views, while improved view synthesis is an additional cornerstone for FN in large baseline arbitrary camera arrangements.

Submitters are encouraged - but not required - to submit results for all application scenarios. However submitters are required to provide results for all sequences in a given category (SMV, FN).

2.1 Application Scenario #1: Super-Multiview Display (SMV)

Reference framework: The source is a large number of views as required by the display. The number of views is typically 80 views or more, and they are arranged in a dense 1D array (linear or arc), covering a wide view range.



Challenge: The reference framework requires high bandwidths for transmission due to the large amount of captured data. The challenge in SMV consists in improving the coding efficiency to such level that the total bitrate for transmitting hundreds of views stays within realistic and economically viable bandwidth requirements (e.g. data rates corresponding to the transmission of tens of views with today's technology). This may require the development of new coding paradigms with better view prediction and view synthesis techniques.

Objective: The main objective in this application scenario is to substantially reduce the data rate required to reconstruct the full set of input views at the receiver compared to existing MPEG state-of-the-art compression standards. The codec may directly transmit all of the input views, use intermediate representations like previously developed multiview plus depth representation with depth estimation and view synthesis embedded in the codec, or may use any other representation which leads to recreation of full set of input views at the receiver. At the decoder it is required to recreate the full set of input views.

2.1.1 Test Sequences

Sequences to be used for evaluation are summarized in Table 1. Exact position of views (view number) to be considered as input and output of the codec are provided in Table 2.

Table 1. Summary of the sequence to be used.

No.	Provider's Name	Seq. Name	Number of Views	Resolution (pel)	Frame rate (fps)	Length	Cam Arrangement (1D parall, 1D arc, 2D parall, 2D arc, Sphere, Arbitrary)	condition to use (if any)	URL of sequence (ID, PW)
1	Nagoya University	Champagne Tower	80	1280 x 960	29.4114	10 sec 300 frames	linear 1D parallel	See M15378 and provider webpage	http://www.fujii.nuee.nagoya-u.ac.jp/multiview-data no password required
2	Nagoya University	Pantomime	80	1280 x 960	29.4114	10 sec 300 frames	linear 1D parallel	See M15378 and provider webpage	
3	Holografika	Big Buck Bunny flowers	91	1280x768	24	5 sec 121 frames	45 degree arc convergent	See M36500	http://mpeg3dvideo.holografika.com/BigBuckBunny_Flowers/Arc/ Login: mpegftv Password: lf3dvideo
4	Holografika	Big Buck Bunny butterfly	91	1280x768	24	5 sec 120 frames	45 degree arc convergent	See M36500	http://mpeg3dvideo.holografika.com/BigBuckBunny_Butterfly/Arc/ Login: mpegftv Password: lf3dvideo

Table 2. View positions to be transmitted for each sequence to be used.

No.	Seq. Name	Views positions to be transmitted
1	Champagne Tower	0-79
2	Pantomime	0-79
4	Big Buck Bunny flowers	5-84
5	Big Buck Bunny butterfly	5-84

Table 2a. Frame range to be transmitted for each sequence to be used.

No.	Seq. Name	Frame range to be transmitted
1	Champagne Tower	0-299
2	Pantomime	0-299
4	Big Buck Bunny flowers	0-120
5	Big Buck Bunny butterfly	0-119

2.1.2 Anchors and bitrates

In the anchor configuration all of the input views are directly compressed with the use of 3D-HEVC encoder, and then transmitted and decoded for display.

As a 3D-HEVC encoder MPEG reference software namely HTM version 13.0 has been used. Software can be found at https://hevc.hhi.fraunhofer.de/svn/svn_3DVCSoftware/tags/HTM-13.0

Following source code modification have been made in order to allow 3D-HEVC for compression of FTV material.

```
In TLibCommon/TypeDef.h, line 57,

#define HEVC_EXT 2 // 3D-HEVC mode // define 2 for 3D-HEVC mode.
```

```
In TAppEncoder/TAppEncTop.cpp Line 1336:

Int* aiIdx2DepthValue = (Int*) calloc(uiMaxDepthValue+1,
sizeof(Int)); //Fix Fix Owieczka +1 Added
```

```
In TLibEncoder/TEncCavlc.cpp line 444:

WRITE_CODE(uiNumDepthValues_coded, 9,
"num_depth_values_in_dlt[i]"); // Fix Fix Owieczka 8->9
```

```
In TLibDecoder/TDecCavlc.cpp line 470:

READ_CODE(9, uiNumDepthValues, "num_depth_values_in_dlt[i]");
//Fix Fix Owieczka 8->9
```

```
In /Lib/TLibDecoder/TDecCAVLC.cpp,

line 235: // assert (uiCode <= 15); // comment it out

line 771: // assert (uiCode <= 15); // comment it out
```

```
In TLibCommon/TComSlice.h,

line 3039: // assert (psId < m_maxId); // comment it out
```

3D-HEVC is used in its most efficient mode HEVC_EXT 2 where sub CU level tools are enabled.

Input 80 views are divided into two halfs counting 40 views each. Each half is separately encoded by HTM 13 encoder.

3D HEVC encoder was configured as follows:

- Inter-view coding structure
    - P-..-P-I-P-..-P inter-view prediction. Each view is predicted only based on previously encoded view (fig. 1).

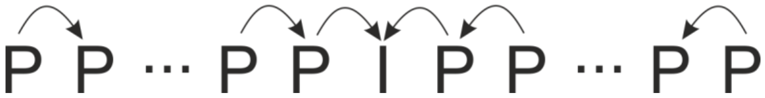


Figure 1. Inter-view prediction structure used for anchor generation

- Temporal prediction structure
    - GOP size 8
    - Intra period every 24 frames (random access at roughly each second)
- 8-bit input data have been used
- HEVC codecs have been configured with 8-bit internal processing

Detailed configuration files of 3D-HEVC can be found in the attachment.

In anchor configuration all of the input views were encoded at 4 rate-point as specified in Table 3.

Table 3. Specification of rate-point for SMV case.

No.	Seq. Name	Maximum bitrate for all data required for recreation of required number of output views [kbps] (QP values used for anchor bitstreams)			
		Rate-point 1	Rate-point 2	Rate-point 3	Rate-point 4
1	Champagne Tower	5433.9 (37)	2553.8 (43)	1590.2 (47)	1204.9 (50)
2	Pantomime	7585.8 (37)	3188.7 (43)	1919.6 (47)	1393.3 (50)
3	Big Buck Bunny flowers	5513.0 (35)	3298.5 (40)	1905.0 (45)	1156.3 (50)
4	Big Buck Bunny butterfly	1464.1 (37)	xxx (40)	799.4 (44)	279.1 (50)

2.1.3 Submissions

Submissions to the call shall:

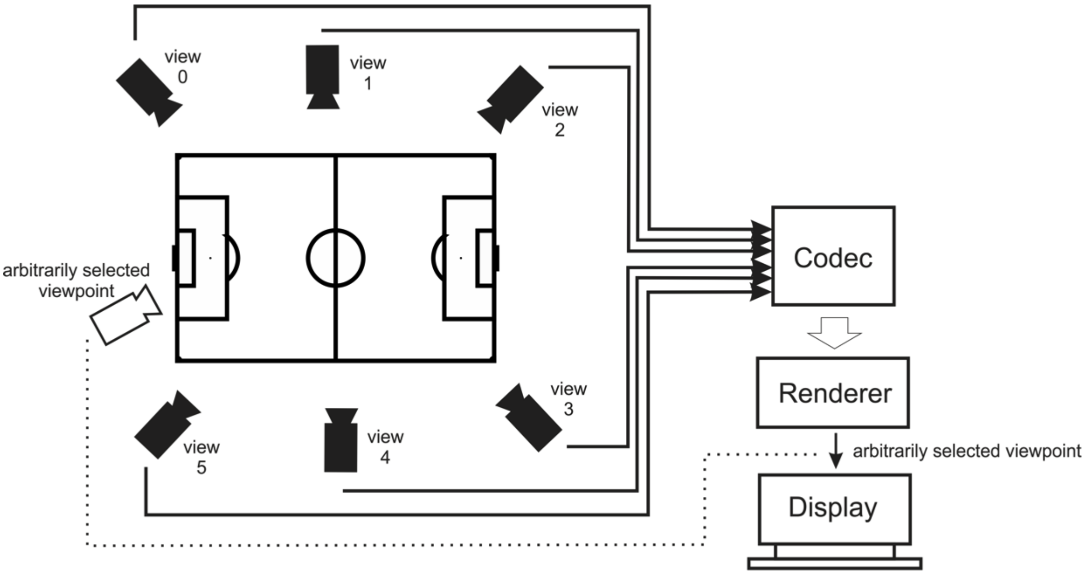
- Be encoded at bitrates no larger than those defined in Table 3.
- Include all necessary data into a single bitstream (including possible depth data, and all supplementary data) for each test sequence.
- Reproduce all of the input views at the decoder as defined in Table 2.
- Allow for random access at intervals not more than 1 second.

Participants shall provide:

- Input contribution to the 114th meeting describing the submission with rough indication about normative processing involved.
- Decoder executable (including all processing tools necessary to re-creation of 80 views required to be transmitted).
- Bitstream for each rate-point.
- Video clip in YUV 4:2:0 files for every output view (according to Table 2) for all rate-point. In total this corresponds to $4 \cdot 80 = 320$ video clips.
- PSNR values for each output view at each rate-point in attached CfE excel sheet.
- Runtime of the coding process (including all processing necessary to re-creation of 80 views required to be transmitted).

2.2 Application Scenario #2: Free Navigation (FN)

Reference Framework: The source is a sparse number of views (e.g., 6 - 10) with arbitrary positioning and wide baseline distance between each view. The input views, along with all supplemental data such as depth, are transmitted. The output will render arbitrary view positions in the 3D space.


view 0
view 1
view 2
arbitrarily selected viewpoint
Codec
Renderer
view 3
view 5
view 4
arbitrarily selected viewpoint
Display

Challenge: The reference framework is comprised of sparse views with large baselines that are not very highly correlated (relative to the ultra-dense views required for Super Multiview displays). The key challenge in the Free Navigation scenario resides in keeping the quality of the view rendering from arbitrarily positioned views notably high. Existing depth-based 3D formats are not able to satisfy such flexible rendering needs with high quality and robustness.

Objective: The main objective in this application scenario is to substantially improve rendering quality at arbitrary virtual view positions in 3D space. It is expected that this may be achieved through an alternative representation format (different from simulcast HEVC and 3D-HEVC), in which case compression efficiency must also be considered. While the emphasis is on the rendering and view synthesis quality, it should be clarified that there is no intention to standardize post-processing tools subsequent to the decoder in the processing chain. However, a more appropriate representation/coding model may be required.

2.2.1 Test Sequences

Sequences to be used for evaluation are summarized in Table 4. Only seven views of each sequences will be used for evaluation. Exact view positions to be transmitted are provided in Table 5.

Table 4. Summary of the sequence to be used.

No.	Provider's Name	Seq. Name	Number of Views	Resolution (pel)	Frame rate (fps)	Length	Camera Arrangement	Depth Data Available	URL of sequence (ID, PW)
1	UHasselt	Soccer-Linear 2	8	1600x1200	60	22 sec 1250 frames	1D parallel	Yes	http://wg11.sc29.org/content//MPEG-04/Part02-Visual/FTV_AhG/UHasselt_Soccer
2	UHasselt	Soccer-Arc 1	7	1920x1080	25	22 sec 550 frames	120 deg. Corner, arc	Yes	https://wg11.sc29.org/content/MPEG-04/Part02-Visual/FTV_AhG/UHasselt_Soccer
3	Poznan University of Technology	Poznan Blocks	10	1920x1080	25	40 sec 1000 frames	100 deg. arc around the scene	Yes	ftp://multimedia.edu.pl/ftv Password provided upon request please email kwegner@multimedia.edu.pl
4	Holografika	Big Buck Bunny Flowers noBlur	91	1920x1080	24	5 sec 121 frames	45 deg, arc	Yes, ground truth depth	http://mpeg3dvideo.holografika.com/BigBuckBunny_Flowers_noblur/Arc/ Login: mpegftv Password: lf3dvideo

Table 5. View positions to be transmitted for each sequence to be used.

No.	Seq. Name	Views positions to be transmitted
1	Soccer Linear 2	1-7
2	Soccer Arc 1	1-7
3	Poznan Blocks	2-8
4	Big Buck Bunny Flowers	6,19,32,45,58,71,84

Table 5a. Frame range to be transmitted for each sequence to be used.

No.	Seq. Name	Frame range to be transmitted
1	Soccer Linear 2	0-599
2	Soccer Arc 1	0-249
3	Poznan Blocks	0-249
4	Big Buck Bunny Flowers	0-120

2.2.2 Virtual view positions

Number of required virtual views to be rendered between each pair of transmitted views are provided in Table 6. Exact camera parameters for each individual virtual viewpoint are attached to CfE.

Table 6. Number of virtual views between each pair of cameras.

Sequence	Number of required virtual views between each pair of cameras
Soccer Linear 2	14
Soccer Arc 1	22
Poznan Blocks	12
Big Buck Bunny Flowers	12

2.2.3 Anchors and bitrates

In the anchor configuration all of the input views along the corresponding depth data are compressed using 3D-HEVC encoder, and then transmitted. After decoding, the requested view positions are being synthesized using VSRS version 4.0.

As a 3D-HEVC encoder MPEG reference software namely HTM version 13.0 has been used. Software can be found at https://hevc.hhi.fraunhofer.de/svn/svn_3DVCSoftware/tags/HTM-13.0

3D HEVC encoder was configured as follows:

- Inter-view coding structure
    - P-..-P-I-P-..-P inter-view prediction. Each view is predicted only based on previously encoded view (fig. 1).

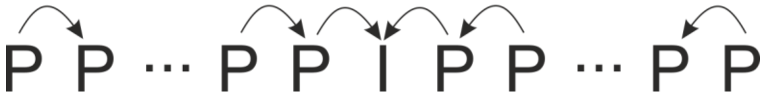


Figure 1. Inter-view prediction structure used for anchor generation

- Temporal prediction structure
    - GOP size 8
    - Intra period every 24 frames (random access at roughly each second)
- 8-bit input data have been used
- HEVC codecs have been configured with 8-bit internal processing
- VSO off

Detailed configuration files of 3D-HEVC can be found in attachment.

In anchor configuration seven input views along the corresponding depth data were encoded at 4 rate-points as specified in Table 7.

Virtual views in anchor configuration are rendered by MPEG view synthesis reference software version 4.0 (VSRS 4.0). The software is available at MPEG SVN repository http://wg11.sc29.org/svn/repos/Explorations/FTV

Detailed description of configuration for VSRS.

Table 7. Specification of rate-point for FN case.

No.	Seq. Name	Maximum bitrate for all data required for recreation of required number of output views [kbps]			
		Rate-point 1	Rate-point 2	Rate-point 3	Rate-point 4
1	Soccer Linear 2	1901.7 (30/)	611.5 (37/)	238.1 (44/)	168.3 (47/)
2	Soccer Arc	5462.9 (30/)	2284.2 (37/)	899.1 (44/)	592.0 (47/)
3	Poznan Blocks	5927.8 (30/)	3187.95 (35/)	1559.6 (40/)	Xx (45/)
4	Big Buck Bunny Flowers	1769.7 (37/)	1267.8 (40/)	816.9 (44/)	561.2 (47/)

2.2.4 Submissions

Submissions to the call shall:

- Be encoded at bitrates no larger than those defined in Table 7.
- Include all necessary data into a single bitstream (including possible depth data, and all supplementary data) for each test sequence.
- Be able to create any requested view at any position in 3D space.
- Create all requested virtual views at predefined virtual camera positions
- Allow for random access at intervals not more than 1 second.

Participants shall provide:

- Input contribution to the 114th meeting describing the submission with rough indication about normative processing involved.
- Decoder executable.
- Renderer executable (if it is not a part of the decoder).
- Bitstream for each rate-point.
- Video clip in YUV 4:2:0 files for every virtual view position rendered from the original uncompressed data (texture and depth data or any other data used by the proposed renderer). In total this corresponds to (14+22+12+12)·(7-1) = 360 video clips.
- Video clip in YUV 4:2:0 files for every virtual view position rendered from coded data at each rate-point. In total this corresponds to (14+22+12+12)·(7-1)·4 = 1440 video clips.
- Video clips in YUV 4:2:0 files for every decoded view used for rendering virtual views. In total this corresponds to 7·4 = 28 video clips
- PSNR values for each output view at each rate-point in attached CfE excel sheet.
- Runtime of the encoding, decoding and rendering process.

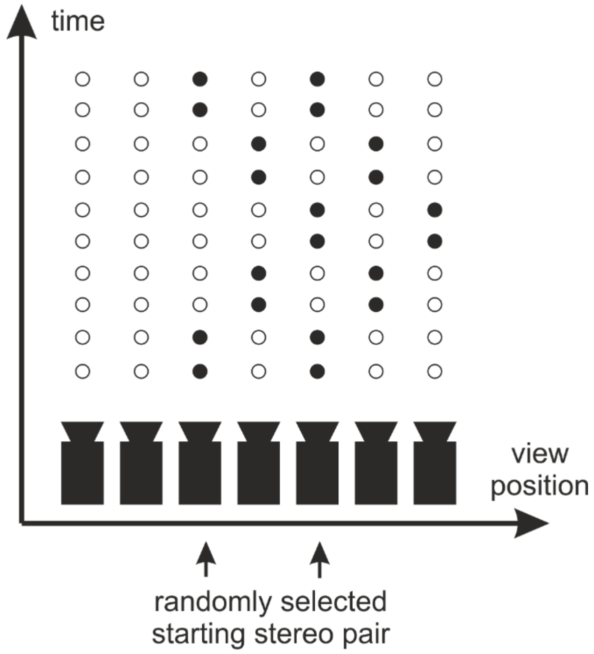
3 Evaluation procedure

Submissions will be evaluated through subjective testing performed during the 114th meeting on stereoscopic monitor. Submissions for both targeted application scenarios will be evaluated separately. A detailed description of evaluation procedure can be found below.

3.1 Super-MultiView (SMV) evaluation procedure

Evaluation of submission for super-multiview scenario will be conducted on stereoscopic monitor. Video clips of decoded views will be combined to create sweeps though all of the transmitted views. Starting position of the sweep will be selected randomly by the test chair.

Sweeps will be constructed at a speed of one frames per view. The baseline distance for the stereo pairs used to create sweeps can be found in Table 8



It is expected that the results of the stereoscopic viewing of the created sweep produce the most meaningful results in terms of overall evaluation of the proposals.

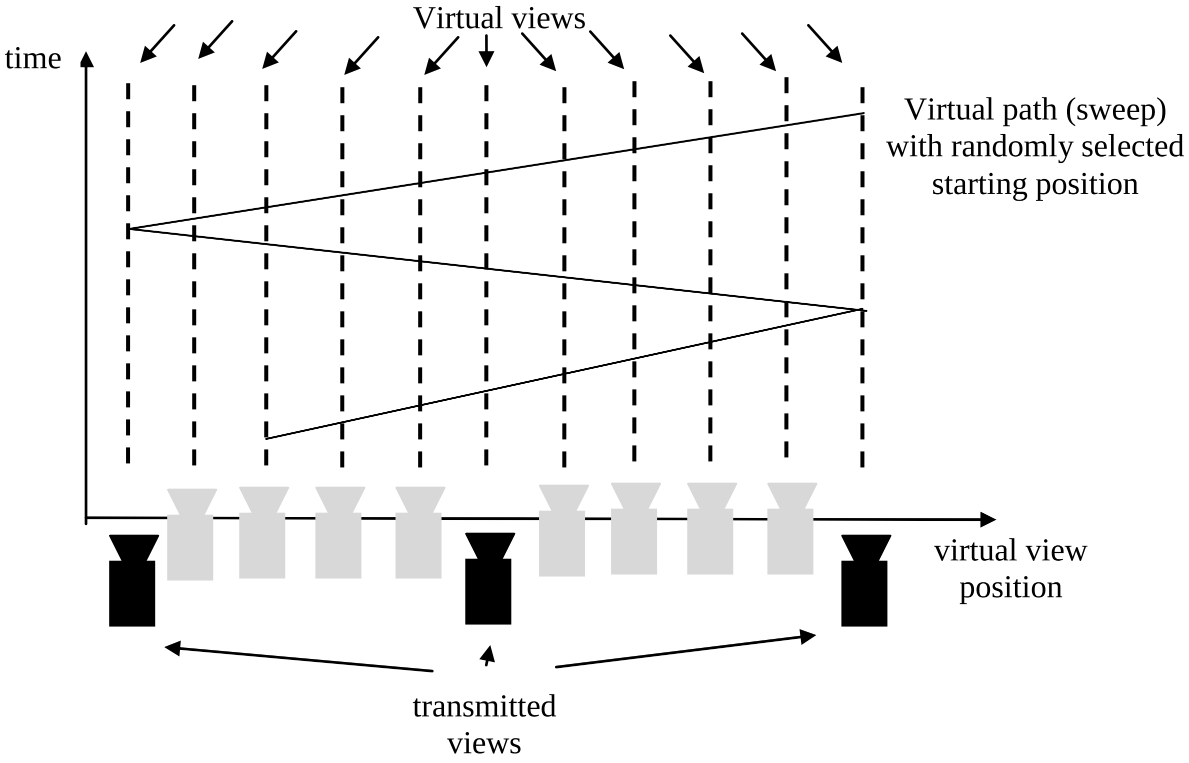
Table 8. Specification of baseline distances and view sweep speed .

No.	Seq. Name	View sweep speed	Stereo baseline distance in terms of view positions
1	Champagne Tower	1	1 (e.g. view 0 1)
2	Pantomime	1	1 (e.g. view 0 1)
4	Big Buck Bunny flowers	1	1 (e.g. view 0 1)
5	Big Buck Bunny butterfly	1	3 (e.g. view 0 3)

Created sweeps will be evaluated subjectively according to BT.500 [5].

3.2 Free Navigation (FN) evaluation procedure

Evaluation of submission for free navigation scenario will be conducted on classical 2D monitor. Video clips of virtual views created based on transmitted data will be combined to create virtual path (sweep) though out all of the virtual views. The starting position of the sweep will be selected randomly by the test chair. Virtual paths will be constructed at a speed of one frame per view.



4 Logistics

Prospective contributors to the Call for Evidence should contact:

Prof. Dr.-Ing. Joern Ostermann

E-mail: mailto:mostermann@tnt.uni-hannover.de

5 References

[1] Masayuki Tanimoto, "FTV (Free-viewpoint Television)", APSIPA Transactions on Signal and Information Processing, Vol. 1, Issue 1, e4 (14 pages) (August 2012). doi: 10.1017/ATSIP.2012.5.

[2] W. Matusik, H. Pfister, "3D TV: A Scalable System for Real-Time Acquisition, Transmission and Autostereoscopic Display of Dynamic Scenes", ACM SIGGRAPH, ISSN: 0730-0301, Vol. 23, Issue 3, pp. 814-824, August 2004.

[3] "FTV Seminar Report," ISO/IEC JTC1/SC29/WG11 MPEG2014/N14552, Sapporo, Japan, July 2014.

[4] "Use Cases and Requirements on Free-viewpoint Television (FTV)," ISO/IEC JTC1/SC29/WG11 MPEG2014/N14178, San Jose, US, January 2014.

[5] International Telecommunication Union Radio Communication Sector; Recommendation ITU-R BT.500-11