ISO/IEC JTC 1/SC 29/WG 04

MPEG Video Coding

Convenorship: CN

Document type:	Output Document
Title:	Common test conditions for MPEG immersive video
Status:	Approved
Date of document:	2024-05-25
Source:	ISO/IEC JTC 1/SC 29/WG 04
Expected action:	None
Action due date:	None
No. of pages:	25 (without cover page)
Email of Convenor:	yul@zju.edu.cn
Committee URL:	https://isotc.iso.org/livelink/livelink/open/jtc1sc29wg4

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC 1/SC 29/WG 04 MPEG VIDEO CODING

ISO/IEC JTC 1/SC 29/WG 04 N 0495

April 2024, Rennes, FR

Title	**Common test conditions for MPEG immersive video**
Source	**WG 04 MPEG Video Coding**
Status	**Approved**
Serial Number	**23997**
Authors	**Adrian Dziembowski (PUT), Bart Kroon (Philips), Joel Jung (Tencent)**

1 Introduction

Common test conditions are desirable to conduct coding experiments in a well-defined environment and ease the comparison of the outcome of experiments. This document specifies the common test conditions for ISO/IEC 23090-12 MPEG immersive video (MIV) related activities. This document replaces [8]. The common test conditions are defined to evaluate the coding efficiency, subjective quality, pixel rate and user experience of immersive video solutions. The technical approach is following these steps:

1. Compress test content,
2. Synthesize intermediate views from decoded views and metadata (when available),
3. Render viewports of real/virtual pose traces with a limited or a wider movement,
4. Evaluate coding efficiency and parallax effect, considering both decoded views and synthesized views.

The bitstream is viewer-independent, meaning that neither the position nor the orientation of the viewer is considered when compressing the test content. The range of supported possible viewer positions is constrained and known.

Two anchors are used that are based on the latest Test Model for MPEG immersive video (TMIV) [1, 2]. The first one, the *MIV main anchor*, is a configuration of TMIV and VVenC encoder[1], encoding some source views completely while taking only patches of others. The second one, the *MIV decoder-side depth-estimating (DSDE) anchor*, is a configuration of TMIV + VVenC + Immersive video depth estimation (IVDE) [4], restricted to encoding an automatically-determined subset of source views without geometry information, and applying depth estimation in between decoding and rendering.

In addition, the *best reference* is the best-known method to render synthesized views using the full source material (without coding). The views are synthesized with the TMIV renderer with view-weighting synthesizer (VWS).

[1] https://github.com/fraunhoferhhi/vvenc

2 Test material

This section lists the test material that is used by the common test conditions. The configuration files are attached to the TMIV reference software [2].

Non-MPEG members may request access to the test material. Some test material is publicly available on the MIV website[2]. All test material is available from the following location on the MPEG content server:

```
/MPEG-I/Part12-ImmersiveVideo/ctc_content/
```

The test material is provided as a set of raw sequences, one per view and component (texture or depth). Texture and depth maps sequences characteristics are reported in Annex A. The sequences have a common format as defined in the Call for MPEG-I Visual Test Materials [6] determining texture and depth representations, filenames, and metadata. The views are numbered according to the ordering of the metadata files. Video data is named as follows:

v{i}_{t}_{w}x{h}_yuv{f}p{b}le.yuv

where:

- *i* is a unique positive integer index used to identify the camera that captured the video sequence. The indexing should preferably start from zero.
- *t* denotes the property of the video that is represented by the video stream. The following types are allowed: (a) texture (b) depth.
- *w* is the width (total number of pixels in a row) of a frame in the Y (luma) channel of the video sequence.
- *h* is the height (total number of pixels in a column) of a frame in the Y (luma) channel of the video sequence.
- *f* is the YUV sub-sampling format used for the video sequence, e.g., 420.
- *b* is the bits per channel of the YUV sequence.

Examples:

- v0_depth_2048x1088_yuv420p16le.yuv
- v2_texture_2048x1088_yuv420p10le.yuv

Table 1 provides the list of sequences. The test material is organized into two categories: computer-generated content and natural content with estimated depth, and several classes:

- Class A – omnidirectional scene captured by spherical cameras (ERP representation),
- Class B – omnidirectional scene captured by semi-spherical cameras (ERP representation),
- Class C – semi-omnidirectional scene captured by semi-spherical cameras (ERP representation),
- Classes D and J – scene captured by camera array (perspective content),
- Class E – scene captured by linear multicamera system (perspective content),
- Classes L and W – scene captured by converging cameras (perspective content).

There are mandatory sequences, which results are required for non-informative proposals, and optional sequences. Optional sequences are challenging content that are deliberately difficult to handle. They are not meant for evaluation or promotion of the test model. The split between mandatory and optional sequences is given in Section 4, as it depends on the anchor.

[2] https://mpeg-miv.org

Table 1: List of sequences. Descriptions of listed sequences are available in Annex A.

Computer-generated content	
Class A:	
A01	ClassroomVideo
Class B:	
B01	Museum
B02	Chess
B03	Guitarist
Class C:	
C01	Hijack
C02	Cyberpunk
Class J:	
J01	Kitchen
J02	Cadillac
J03	Mirror
J04	Fan
Class W:	
W01	Group
W02	Dancing

Natural content	
Class D:	
D01	Painter
D02	Breakfast
D03	Barn
Class E:	
E01	Frog
E02	Carpark
E03	Street
Class L:	
L01	Fencing
L02	CBABasketball
L03	MartialArts

3 Software tools

The referenced tools are listed in Table 2, with source code location, documentation, and release tag.

Table 2: List of used tools

Tool name		Location	Release
TMIV	[2]	https://gitlab.com/mpeg-i-visual/tmiv	v20.0
VVenC		https://github.com/fraunhoferhhi/vvenc	v1.11.1
VVdeC		https://github.com/fraunhoferhhi/vvdec	v2.3.0
IV-PSNR	[3] [7]	https://gitlab.com/mpeg-i-visual/ivpsnr	v5.0
IVDE	[4]	https://gitlab.com/mpeg-i-visual/ivde	v8.0

3.1 VVenC

The VVenC implementation of VVC is used for all anchors. The expert mode (vvencFFapp) based on VVC test model is used, with the random access “slow” configuration. The configuration file is attached to the TMIV software.

3.2 IV-PSNR

The PSNR for Immersive Video (IV-PSNR) metric is a full-reference metric based on the PSNR. It includes two major changes: the pixel shift, that considers that edges of the objects in the synthesized view may be shifted due to rounding errors, and the global color shift, that considers that different input views may have various color characteristics. IV-PSNR software produces, in addition to the IV-PSNR score, the WS-PSNR score for both perspective and omni-directional contents.

3.3 IVDE

Immersive Video Depth Estimation (IVDE) is a depth estimation method that can be used to create geometry data for a 6DoF scene representation from views acquired by multiple perspective or omnidirectional cameras. Depth is estimated for segments instead of individual pixels, and thus the size of segments can be used to control the trade-off between the quality of depth maps and the processing time. Larger segments can be used to attain fast depth estimation, or finer segments can be used to attain higher quality.

IVDE further includes a feature extractor, used with the decoder-side-depth-estimating anchor, to produce a list of blocks to be skipped by IVDE.

4 Anchor definition

Two anchors are considered to encode the multi-view sequences:

- MIV main anchor: coding with TMIV + VVenC, packing some source views completely while taking only patches of others,
- MIV decoder-side depth-estimating anchor: encoding with TMIV + VVenC, packing a subset of source views completely but without their geometry information, followed by decoding, depth estimation and rendering with TMIV + IVDE.

In addition, there is a non-anchor reference condition named the best reference that directly renders from all source views without coding, using the best-known method, which currently is the TMIV renderer with the view-weighting synthesizer (VWS).

An algorithmic description of TMIV is provided in [1]. VVenC is configured to encode each video sub-bitstream using the VVC random access profile. All configuration parameters for each of the anchors is provided by this document and its attachments.

CTC-specific configuration files and detailed run instructions are provided with the reference software [2]. CTC-specific IVDE configuration files for IVDE are provided with IVDE software.

The collaborative anchor generation [5] is performed only for mandatory sequences (cf. Tables 4 and 5). If a proponent decides to include optional sequences, they must generate the anchor as well.

4.1 Coding of the anchor views

For each anchor, a sequence of 65 consecutive frames is encoded. The start frames for each sequence are reflected in Table 3. Specific details for each anchor are given in the following sub-sections.

Table 3: Start frames for each sequence.

Id	Sequence	Start frame
A01	ClassroomVideo	23
B01	Museum	100
B02	Chess	60
B03	Guitarist [1]	0
C01	Hijack	0
C02	Cyberpunk	0

J01	Kitchen	0
J02	Cadillac	0
J03	Mirror	0
J04	Fan	0
W01	Group	0
W02	Dancing	0
D01	Painter	40
D02	Breakfast	0
D03	Barn	0
E01	Frog	135
E02	Carpark	115
E03	Street	167
L01	Fencing	0
L02	CBABasketball [2]	0
L03	MartialArts	0

[1] for Guitarist sequence only odd views are used: v1, v3, v5, etc.

[2] for CBABasketball sequence, only first 15 views are used: v0 – v14.

4.2 Coding for the MIV main anchor

For each video sequence, four rate points are considered. The set of QPs for the texture is sequence dependent, to target the 5 to 50 Mbps bitrate range. The set of texture QPs is supplied in Table 4. To each texture QP corresponds one single geometry QP. The mapping from texture QP (q) to geometry QP (q') is the same for all sequences, and is given by:

$$q' = \max(1, [-14.2 + 0.8q]) \quad (1)$$

whereby $[\cdot]$ indicates the rounding to nearest integer operation.

Table 4: List of texture QPs for the MIV main anchor.

Sequence		RP1	RP2	RP3	RP4
A01	ClassroomVideo	26	30	38	51
B01	Museum	29	40	47	51
B02	Chess	18	27	35	45
B03	Guitarist	22	24	29	39
C01	Hijack	19	24	34	49
C02	Cyberpunk	21	24	29	39
J01	Kitchen	18	26	33	41
J02	Cadillac	22	31	41	51

J03	Mirror	26	33	42	51
J04	Fan	32	38	45	51
W01	Group	26	31	37	46
W02	Dancing	20	24	28	40
D01	Painter	24	32	43	51
D02	Breakfast	25	30	35	43
D03	Barn	25	30	35	42
E01	Frog	29	34	40	46
E02	Carpark	23	28	37	47
E03	Street	21	25	32	41
L01	Fencing	23	28	39	51
L02	CBABasketball	24	27	31	43
L03	MartialArts	24	27	31	43

Besides the four rate points defined above, there is also the "RP0" rate point, in which no video compression (i.e., no VVenC coding) is performed.

The anchor encodes all source views. The anchor bitstreams include decoded picture hashes for automatic consistency checking.

The mandatory sequences are B02, B03, J02, J04, W01, D01, E01, and L02. All other sequence from Table 1 are optional.

Coding for the MIV main anchor has the following steps:

1. Encode the MIV bitstream using the TMIV encoder.
2. Encode the resulting attribute video data using the VVenC encoder (only for RP1 – RP4).
3. Multiplex the bitstreams using the TMIV multiplexer (only for RP1 – RP4).
4. Decode and render the MIV bitstream using the TMIV decoder.

4.3 Coding for the MIV decoder-side depth-estimating anchor

The encoder parameters for the MIV decoder-side depth-estimating anchor are the same as for MIV main anchor except:

- maxBasicViewFraction = 1.0,
- outputAdditionalViews = false,
- dynamicDepthRange = false,
- haveGeometryVideo = false,
- geometryScaleEnabledFlag = false,
- maxAtlases = 4

The mandatory sequences are J01, W01, D01, D03, L01, and L02. All other sequence from Table 1 are optional.

The set of texture QPs is supplied in Table 5. Note that there is no geometry to encode, hence there are no depth QPs. Besides the four rate points defined below, there is also the "RP0" rate point, in which no video compression (i.e., no VVenC coding) is performed.

Coding for the MIV decoder-side depth-estimating anchor has the following steps:

1. Encode the MIV bitstream using the TMIV encoder.
2. Encode the resulting attribute video data using the VVenC encoder (only for RP1 – RP4).
3. Multiplex the bitstreams using the TMIV multiplexer (only for RP1 – RP4).
4. Decode (but not render) the MIV bitstream using the TMIV decoder.
5. Estimate depth maps for each view using IVDE.
6. Render using the TMIV renderer.

Table 5: List of texture QPs for the MIV DSDE anchor.

Sequence		RP1	RP2	RP3	RP4
A01	ClassroomVideo	29	33	41	50
B01	Museum	39	47	49	51
B02	Chess	23	29	35	42
B03	Guitarist	30	36	40	42
C01	Hijack	18	23	29	35
C02	Cyberpunk	25	30	35	39
J01	Kitchen	26	32	38	43
J02	Cadillac	26	33	40	48
J03	Mirror	25	33	40	48
J04	Fan	29	31	37	46
W01	Group	28	33	39	46
W02	Dancing	28	33	38	42
D01	Painter	23	30	36	41
D02	Breakfast	21	27	31	36
D03	Barn	21	26	30	35
E01	Frog	29	32	37	44
E02	Carpark	21	25	31	35
E03	Street	21	24	28	35
L01	Fencing	21	25	28	34
L02	CBABasketball	19	23	25	29
L03	MartialArts	19	23	25	29

Synthesis of intermediate views

Both for objective and subjective testing, a range of frames of each sequence is synthesized at source positions. For the synthesis, all decoded atlases are used as input of the view synthesis algorithm.

Proposals are not required to code views corresponding to all anchor-coded views but are required to be able to reconstruct source views and generate viewports for any intermediate view position in the

designated range for each test sequence. Each sequence definition includes a virtual view named "viewport" that defines the field of view and resolution of the viewports.

The format of each synthesized view is an omnidirectional image with equirectangular projection with the same angular resolution (pixels / degree) for ERP or semi-ERP test materials, and a linear perspective projection for linear perspective input content. The synthesis result is 10-bit YUV 4:2:0 format for subjective and objective evaluation. Inpainting of invalid pixels is used for both subjective and objective testing.

5 Evaluation of proposals

Objective and subjective results on mandatory sequences only are required for an adoption of a proposal. Additional results obtained on optional sequences can be provided as additional information.

Bitrate matching between the anchor and the proposal is encouraged to facilitate the objective and subjective comparison. The proponent is allowed to have a single QP change to better match the QP of the anchor.

A proponent may choose to base their proposal on another TMIV version than the one defined in Table 2. However, such a choice should be made with caution, to ensure that a proposed tool do not interfere with changes that occurred between used and current TMIV releases. Moreover, the used TMIV version (if different from Table 2) should be explicitly written in the proposal. If such a proposal is considered relevant, then it can be tested on the latest version of TMIV in a collaborative manner.

5.1 Subjective quality evaluation

For subjective viewing, each sequence is synthesized according to a set of pose traces. A pose trace specifies for each frame the position and orientation of the viewport to synthesize. Each pose trace is stored as a comma-separated table with position (X, Y, Z) and orientation (Yaw, Pitch, Roll) columns and exactly one row per frame of the sequence. The format of each synthesized view is an image with perspective projection with 1920 × 1080 pixels resolution, at most 90-degree field of view and 10-bit YUV 4:2:0 color format. The purpose is to mimic natural viewing on a head-mounted display while using offline tools and a 2D monitor.

Because of the large difference in visual comfort between a viewer that voluntarily initiates head motion versus a viewer watching the same viewport on a 2D monitor, pose traces have a small amount of motion. For each sequence there are three pose traces – named *X*p01, *X*p02 and *X*p03 – which are meant to represent a diversity of natural head movement compliant with the overall dimension of the capture rig. The pose traces are attached to the TMIV reference software [2]. The TMIV decoder is configured to extend the video to 260 frames by mirroring the 65-frame sequences. It is meaningful to define the pose traces according to the conditions of capture, and typically to define the related path within the volume of the camera rig. It is convenient to formulate this range as a volume in 3D space.

For adoption of a proposed method, the proponent must provide any pose trace of the proposed method, during a viewing session, and make it clear what the bitrate and pixel-rate differences with the anchor are.

The mp4 pose traces are 1920 x 1080 viewports converted from the YUV pose trace. The following command line shall be used to generate the mp4 pose traces:

```
ffmpeg \
  -f rawvideo -pix_fmt yuv420p10le -s:v 1920x1080 -r ${rate} -i {input}.yuv \
  -c:v libx264 -crf 10 -pix_fmt yuv420p {output}.mp4
```

whereby ${rate} is the frame rate in Hz, e.g., 25 or 30. It is recommended to use ffmpeg 4.2 or newer.

5.2 Objective evaluation

A "synthesized view" corresponds to a source view that is reconstructed through synthesis (view interpolation) by the anchor using the decoded bitstream. All source views are synthesized for objective evaluation, and it is not considered if the source view is fully, partially, or not at all present in the bitstream.

The proposal should be compared with the anchor coding results, by reporting the metrics using the reporting template provided in [5]. This includes a tab per sequence, a summary sheet, and an analysis sheet per metric. BD-rates, BD-PSNRs, and averages are automatically calculated to ensure consistent reporting.

For all test classes, the IV-PSNR software will be used to compute both the WS-PSNR and IV-PSNR based BD-rate and BD-PSNR values calculated for synthesized source views. The comparison of proposals with the anchors will be expressed in terms of BD-rate and BD-PSNR computed on rate-distortion curves.

The BD-rate and BD-PSNR values for anchor and proponent are obtained from:

- The average over each source view and specified frames (Table 3) of the metric between the intermediate view synthesized from decoded atlases and the original/non-compressed source views,
- The total bitrate required to encode the views (including depths) for all frames.

For WS-PSNR and IV-PSNR the average is computed in mean square error (MSE) space.

Because TMIV makes use of floating-point operations, it is important to report the compiler and operating system that are used for evaluation. Preferred compilers are GCC 7 or newer and VC15 or newer. The TMIV software includes a manual with build instructions.

When the BD-rate or BD-PSNR computation returns a zero value, no average over all sequences will be calculated for this metric (instead "---" is printed). The reporting template includes rate-distortion curves for each metric to study and report the reason for the lack of overlap. The result of the "RP0" is shown as a horizontal line presenting the practical limitation of the TMIV encoding of each sequence (without the influence of video compression).

5.3 Pixel rate evaluation

Objective evaluation criteria include pixel rate, which is included in the reporting template. Contributions are required to provide pixel rate for each tested sequence. Proponents should report results which they believe are the most optimal compromise between pixel rate and quality. To provide a meaningful reference for pixel rate values, the following constraints are defined:

The pixel rate test condition constraints:

- The combined maximum luma sample rate across all decoders is maximally 1,069,547,520 samples per second (e.g., 32 MP @ 30 fps, corresponding to HEVC Main 10 profile @ Level 5.2)
- Each decoder instantiation is constrained to a maximum luma picture size of 8,912,896 pixels (e.g., 4096 x 2048, corresponding to HEVC Main 10 profile @ Level 5.2).
- The maximum number of simultaneous decoder instantiations is four.

These conditions are orthogonal to bitrate conditions. All anchors satisfy the pixel rate test condition constraints: the test model automatically determines suitable atlas frame sizes based on these constraints.

With proper motivation, a proponent may choose another pixel rate or release the pixel rate constraint completely. In such a case, the proponent shall generate the anchor accordingly.

5.4 Runtime evaluation

Runtimes should be reported for anchors and proposals (corresponding cells in the reporting template are mentioned):

- Atlas generation including all preprocessing but without video encoding, cells M224 to M227 and AA224 to AA227.
- Video encoding of all atlases of all components, cells M4 to M103 and AA4 to AA103, M114 to M213 and AA114 to AA213.
- Rendering including video decoding and all postprocessing, cells N232 to N356 and AB232 to AB356.

The reference software includes measurement of CPU runtime, excluding loading from disk and writing to disk. Proposals should include a similar runtime measurement.

It is reminded that the proponent should fill in the runtimes for both anchor and proposed method, so that the delta between anchor and proposal runtimes has a meaning. The anchor and proposed methods should be run on the same system with a similar load. On a compute server with fluctuating load, the anchor and proposal may be run concurrently. On a workstation the conditions may be run successively, with no other processes active.

6 References

[1] A. Dziembowski, G. Lee, Test model 20 for MPEG immersive video, ISO/IEC JTC 1/SC 29/WG 04 N 0514, April 2024, Rennes.

[2] TMIV reference software, public url: https://gitlab.com/mpeg-i-visual/tmiv, MPEG-internal url: https://git.mpeg.expert/MPEG/Video/MIV/Software/TMIV.

[3] Software manual of IV-PSNR for Immersive Video, ISO/IEC JTC 1/SC 29/WG 04 N 0411, October 2023, Hannover.

[4] Manual of Immersive Video Depth Estimation, ISO/IEC JTC 1/SC 29/WG 04 N 0058, May 2020, Online.

[5] A. Dziembowski, B. Kroon, J.Y. Jeong, Report of MPEG immersive video CTC anchor generation, ISO/IEC JTC 1/SC 29/WG 04 N 0496, April 2024, Rennes.

[6] Vinod Kumar Malamal Vadakital, Call for MPEG Immersive Video Test Materials, ISO/IEC JTC 1/SC 29/WG 04 N 0170, January 2022, Online.

[7] A. Dziembowski, D. Mieloch, J. Stankowski, and A. Grzelka, "IV-PSNR – the objective quality metric for immersive video applications," IEEE Tr. on Circuits and Systems for Video Technology, 2022, doi: 10.1109/TCSVT.2022.3179575.

[8] A. Dziembowski, J. Jung, B. Kroon, Common test conditions for MPEG immersive video, ISO/IEC JTC 1/SC 29/WG 04 N 0406, October 2023, Hannover.

Annex A: test sequences characteristics

Computer-generated content

ClassroomVideo

The general characteristics of the ClassroomVideo sequence are summarized in Table 6. Source view positions are according to a hexagonally-packed circular disc with an additional top and bottom view, as shown in Figure 1.

Table 6: Characteristics of the ClassroomVideo sequence

Category – Short name	A01
Input contributions	WG 11 M42415, WG 11 M 42756 and WG 11 M 42944
Length & frame rate	120 frames (30 fps)
Number of source views	15
Texture format	YUV 4:2:0 10 bits
Depth format	YUV 4:2:0 16 bits
Depth range	[0.8m, ∞), normalized disparity
Source view resolution	4096 × 2048
View FoV & mapping	360° × 180° ERP
Global FoV	360° × 180°

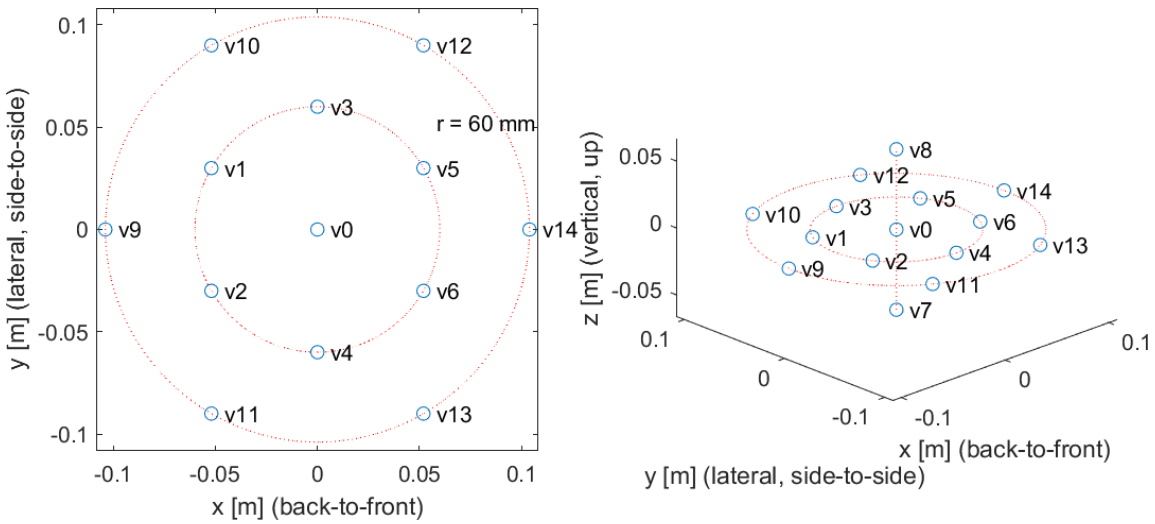

Figure 1: Visualization of the view positions of the ClassroomVideo sequence

The viewing space volume is a spheroid centered at source view v0 e.g. (0, 0, 0) meter position, with equatorial radius 104 mm and polar distance 60 mm:

$$\frac{x^2 + y^2}{(104\ \text{mm})^2} + \frac{z^2}{(60\ \text{mm})^2} = 1$$

Museum

The general characteristics of the Museum sequence are summarized in Table 7. The cameras are disposed on a spherical surface of 30 cm radius, and divergent in the direction of the sphere radius. Figure 2 provides the (X, Y, Z) coordinates and the spherical dimension, with an example using the 11th view. The metadata file comprising source and intermediate view positions is attachment A12 to this output document.

Table 7: Characteristics of the Museum sequence

Category – Short name	B01
Input contribution	WG 11 M42349
Length & frame rate	300 frames (30 fps)
Number of source views	24
Source view resolution	2048 × 2048
Texture format	YUV 4:2:0 10 bits
Depth format	YUV 4:2:0 16 bits
Depth range	[0.5 m, 25 m], normalized disparity
View FoV & mapping	180° × 180° ERP
Global FoV	360° × 180°

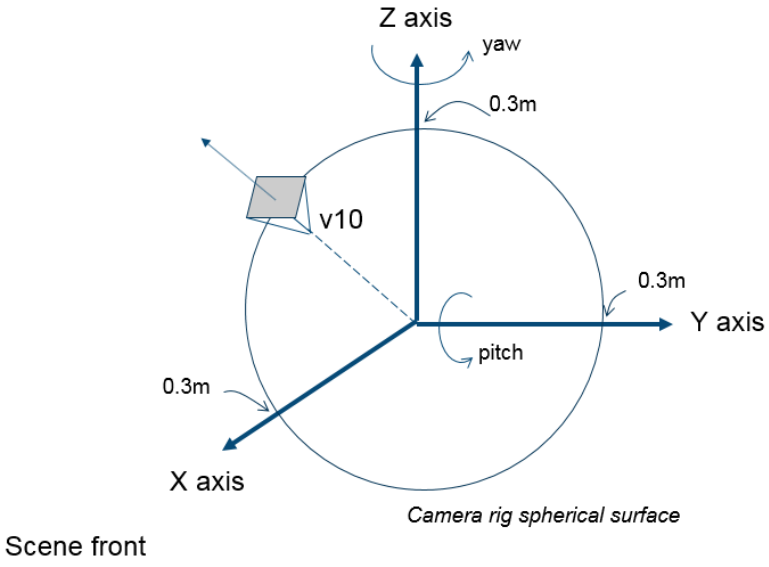

Figure 2: Coordinate system as used by 3D Audio and OMAF, with view 10 of the Museum sequence superimposed

The viewing space volume is a sphere centered at position [0, 0, 1.65] meter with a 300 mm radius:

$$\frac{x^2 + y^2 + (z - 1.65\ \mathrm{m})^2}{(300\ \mathrm{mm})^2} = 1$$

Chess

The general characteristics of the Chess sequence are summarized in Table 8. In total there are ten source cameras, laid out in a sphere-like arrangement as illustrated in Figure 3. One camera in the constellation captures the top of the scene and another the bottom. The remaining eight cameras are pointing outwards to capture the rest of the scene. The radius of the spherical camera constellation is 30 cm. This sequence also comes with a ground truth pose-trace viewport video.

Table 8: Characteristics of the Chess sequence

Category – Short name	B02
Input contributions	WG 11 M50787
Length & frame rate	300 frames (30 fps)
Number of source views	10
Texture format	YUV 4:2:0 10-bits
Depth format	YUV 4:2:0 16-bits
Depth range	[0.1 m, 500 m], normalized disparity
Source view resolution	2048 × 2048
View FoV & mapping	180° × 180° ERP
Global FoV	360° × 180°

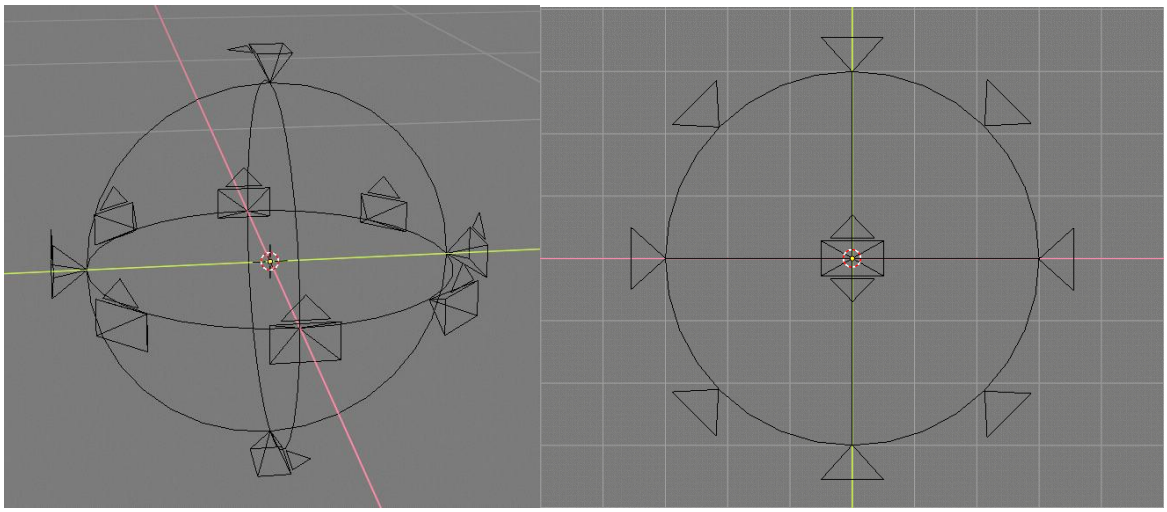

Figure 3: visualization of the camera constellation for Chess

The viewing space volume is a sphere centered at position [-0.5, -0.5, 1.0] meter with a 300 mm radius:

$$\frac{(x + 0.5\text{m})^2 + (y + 0.5\text{m})^2 + (z - 1\text{m})^2}{(300\ \text{mm})^2} = 1$$

Guitarist

The general characteristics of the Hijack sequence are summarized in Table 9. Figure 4 provides a visualization of the virtual camera rig.

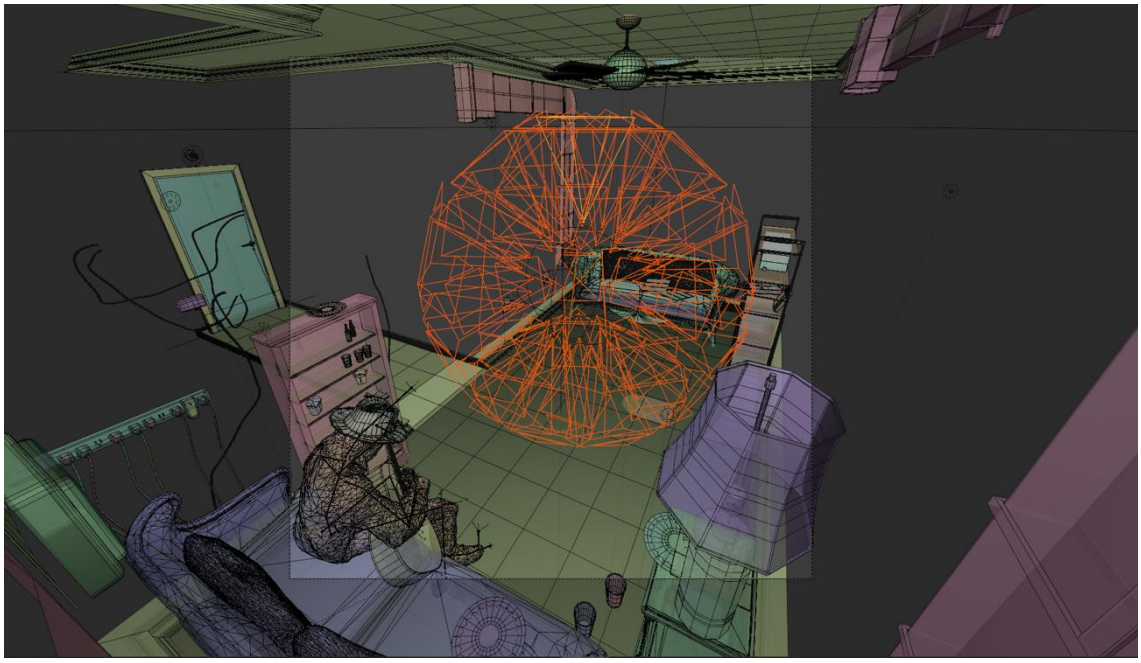

Figure 4: Visualization of the view positions of the Guitarist sequence

Table 9: Characteristics of the Guitarist sequence

Category - Name	B03
Input contributions	M58080
Length & frame rate	300 frames (30 fps)
Number of source views	46
Texture format	YUV 4:2:0 10-bit
Depth format	YUV 4:2:0 16-bit, normalized disparity in (0.1 m, 500.0 m) range
Source view resolution	2048 × 2048
View FoV & mapping	180° × 180° ERP
Global FoV	360° × 180°

Hijack

The general characteristics of the Hijack sequence are summarized in Table 10. Figure 5 provides a visualization of the virtual camera rig in bias, top and front view respectively.

Table 10: Characteristics of the Hijack sequence

Category – Short name	C01
Input contribution	WG 11 M42349
Length & frame rate	300 frames (30 fps)
Number of source views	10
Source view resolution	4096 × 2048
Texture format	YUV 4:2:0 10 bits
Depth format	YUV 4:2:0 16 bits
Depth range	[0.5 m, 25 m], normalized disparity
View FoV & mapping	180° × 90° ERP
Global FoV	180° × 90°

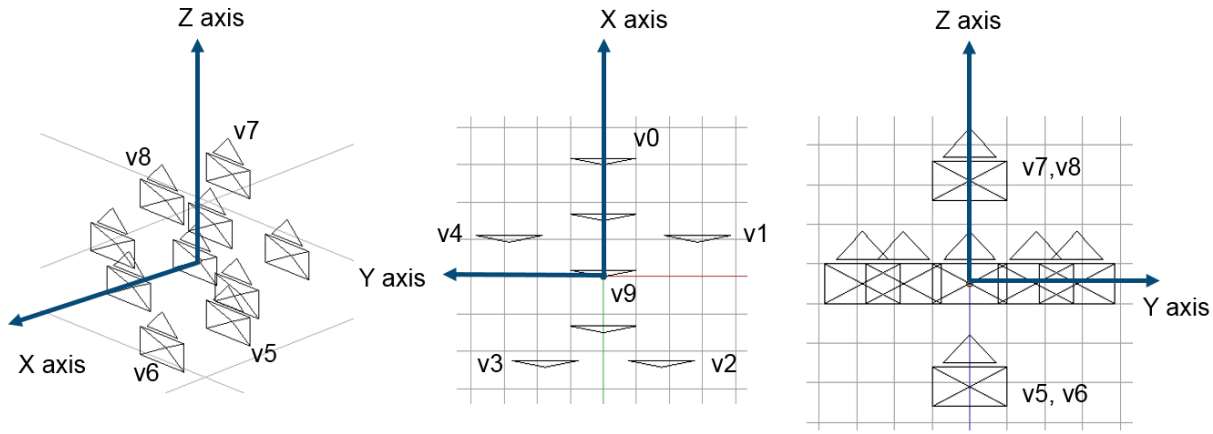

Figure 5: Visualization of the view positions of the Hijack sequence

The viewing space volume is a sphere centered at position [0, 0, 1.65] meter with a 300 mm radius:

$$\frac{x^2 + y^2 + (z - 1.65\ \mathrm{m})^2}{(300\ \mathrm{mm})^2} = 1$$

Cyberpunk

The general characteristics of the Hijack sequence are summarized in Table 11. Cameras were arranged in the same way, as for Hijack sequence.

Table 11: Characteristics of the Cyberpunk sequence

Category – Short name	C02
Input contribution	M58433
Length & frame rate	100 frames (30 fps)
Number of source views	10
Source view resolution	2048 × 2048
Texture format	YUV 4:2:0 10 bits
Depth format	YUV 4:2:0 16 bits
Depth range	[0.95 m, 77 m], normalized disparity
View FoV & mapping	180° × 180° ERP
Global FoV	180° × 180°

Kitchen

The general characteristics of the Kitchen sequence are summarized in Table 12. The captured views form a 5 × 5 planar array.

Table 12: Characteristics of the Kitchen sequence

Category – Short name	J01
Input contribution	WG 11 M43318
Length & frame rate	97 frames (30 fps)
Number of source views	25 (5x5)
Source view resolution	1920x1080
Texture format	YUV 4:2:0 10 bits
Depth format	YUV 4:2:0 10 bits
Depth range	[2.24 m, 7.17 m], normalized disparity
View FoV & mapping	53.1° × 31.4° Rectilinear
Lens	32 mm
Camera spacing	20cm x 20cm

The viewing space volume is a spheroid centered at position [0, -0.4, 0.4] meter, covering a vertical square of side equal to 0.8m and developed in the forward axis by 0.35m max.

Cadillac

The general characteristics of the Cadillac sequence are summarized in Table 13. The rig is a planar rectangular rig of 5 by 3 cameras with a slight tilt upward.

Table 13: Characteristics of the Cadillac sequence

Category – Name	J02
Input contribution	WG 11 M57186
Length & frame rate	100 frames (30 fps)
Number of source views	15
Source view resolution	1920x1080
Texture format	YUV 4:2:0 10 bits
Depth format	YUV 4:2:0 16 bits
Depth range	[1 m, 14 m], normalized disparity
View FoV & mapping	66° horizontal Rectilinear
Camera spacing	5 x 3 rectangular rig spaced by 20cm horizontally and vertically 1st row: v0 to v4 2nd row: v5 to v9 3rd row: v10 to v14

The viewing space volume is a spheroid encompassing the 15 cameras.

Mirror

The general characteristics of the Mirror sequence are summarized in Table 14. The rig is a planar rectangular rig of 5 by 3 cameras with a slight tilt downward.

Table 14: Characteristics of the Mirror sequence

Category – Name	J03
Input contribution	WG 11 M55710
Length & frame rate	100 frames (30 fps)
Number of source views	15
Source view resolution	1920x1080
Texture format	YUV 4:2:0 10 bits
Depth format	YUV 4:2:0 16 bits
Depth range	[1.5 m, 8.0 m], normalized disparity
View FoV & mapping	70° horizontal Rectilinear
Camera spacing	5 x 3 rectangular rig spaced by 20cm horizontally and vertically 1st row: v0 to v4 2nd row: v5 to v9 3rd row: v10 to v14

The viewing space volume is a spheroid encompassing the 15 cameras.

Fan

The general characteristics of the Fan sequence are summarized in Table 15. The rig is a planar rectangular rig of 5 by 3 cameras with a slight tilt downward.

Table 15: Characteristics of the Fan sequence

Category - Name	J04
Input contribution	WG 11 M54732
Length & frame rate	97 frames (30 fps)
Number of source views	15
Source view resolution	1920x1080
Texture format	YUV 4:2:0 10 bits
Depth format	YUV 4:2:0 16 bits
Depth range	[0.35 m, 12.5 m], normalized disparity
View FoV & mapping	50° horizontal Rectilinear
Camera spacing	5 x 3 rectangular rig spaced by 10cm horizontally and vertically 1st row: v0 to v4 2nd row: v5 to v9 3rd row: v10 to v14

The viewing space volume is a spheroid encompassing the 15 cameras.

Group

The general characteristics of the Group sequence are summarized in Table 16 and source view positions illustrated in Figure 6.

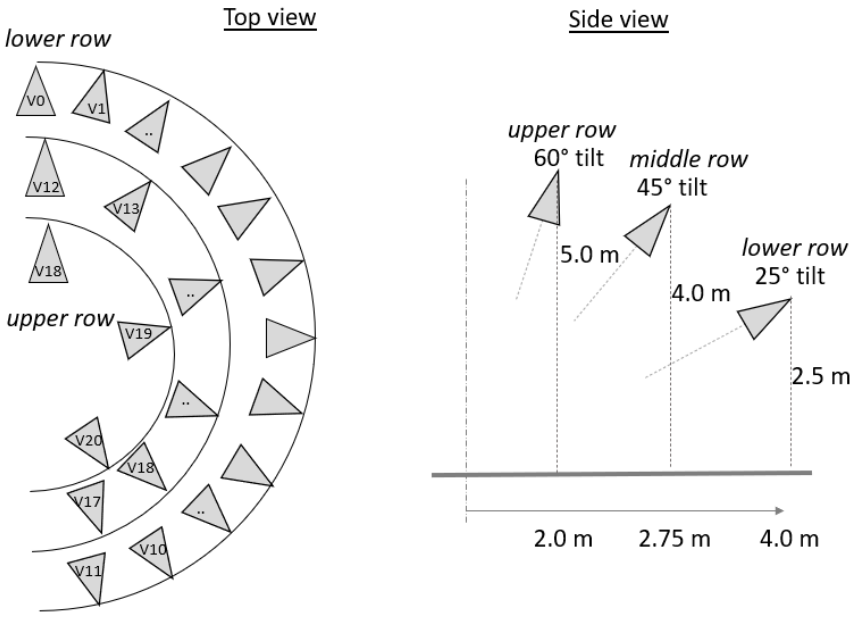

Figure 6: Visualization of the view positions of the Group sequence

The captured views form a partial dome of 21 views made of lower, middle and upper arc shape row, roughly span over 180° and looking inward to a central scene.

Table 16: Characteristics of the Group sequence

Category - Name	W01
Input contribution	WG 11 M54731
Length & frame rate	99 frames (30 fps)
Number of source views	21
Source view resolution	1920x1080
Texture format	YUV 4:2:0 10 bits
Depth format	YUV 4:2:0 16 bits
Depth range	[1.5 m, 25 m], normalized disparity
View FoV & mapping	75° × 48° Rectilinear
Camera spacing	12 cameras span on #180° of arc radius 4.0m, height 2.5 m tilt 25° 6 cameras span on #180° of arc radius 2.75m, height 4.0 m tilt 45° 3 cameras span on #180° of arc radius 2.0m, height 5.0 m tilt 60°

The viewing space volume is a flat volume which would be wrapped around this partial half dome.

Dancing

The general characteristics of the Dancing sequence are summarized in Table 17. Figure 7 provides a visualization of the virtual camera rig.

Table 17: Characteristics of the Dancing sequence

Category - Name	W02
Input contribution	WG 11 M43318, M57751
Length & frame rate	300 frames (30 fps)
Number of source views	24
Source view resolution	1920x1080
Texture format	YUV 4:2:0 10 bits
Depth format	YUV 4:2:0 16 bits
View FoV & mapping	90° horizontal rectilinear
Camera spacing	3 x 8 vertically stacked arcs : 1st row: v0 to v7 2nd row: v8 to v15 3rd row: v16 to v24 vertical spacing = 0.3m horizontal spacing (identical on each arc): v0 -> v1 = 0.62m v1 -> v2 = 0.64m v2 -> v3 = 0.66m v3 -> v4 = 0.34m v4 -> v5 = 0.68m v5 -> v6 = 0.69m v6 -> v7 = 0.70m
zNear	1.2 m
zFar	14.2 m

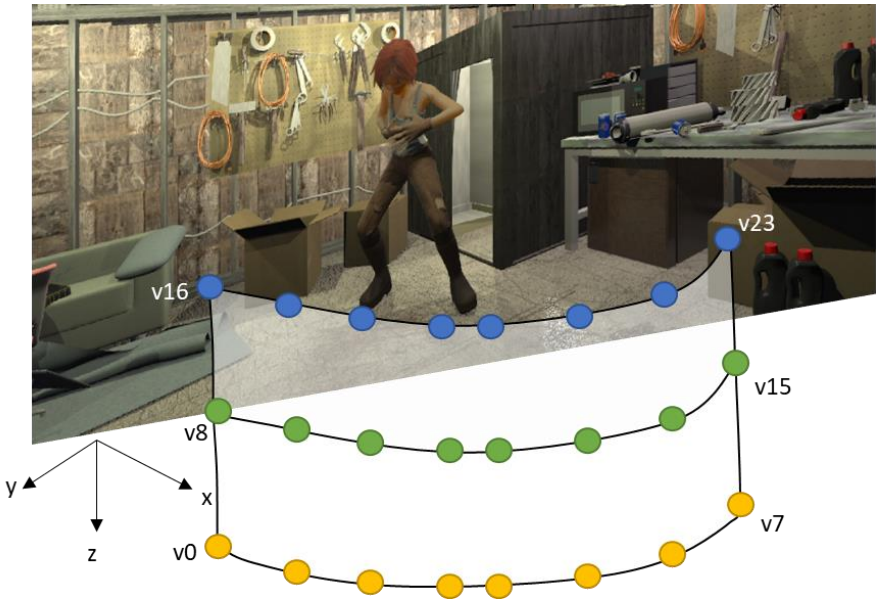

Figure 7: Visualization of the view positions of the Dancing sequence

Natural content

Painter

The general characteristics of the Painter sequence are summarized in Table 18. The refined depths proposed in [m47445] are used.

Table 18: Summary of the Painter sequence

Category – Short name	D01
Input contributions	WG 11 M40010, WG 11 M40011, WG 11 M43366 and WG 11 M47445.
Length & frame rate	300 frames (30 fps)
Number of source views	16 (4x4)
Source view resolution	2048 × 1088
Texture format	YUV 4:2:0 10 bits
Depth format	YUV 4:2:0 16 bits
Depth range	[1 m, 10 m], normalized disparity

The viewing space volume is a spheroid centered at position [0, -0.35, -0.35] meter, covering a vertical square of side equal to 20cm and developed in the forward axis by 25cm max.

Breakfast

The general characteristics of the Breakfast sequence are summarized in Table 19 form a 5x3 planar array.

Table 19: Summary of the Breakfast sequence

Category – Short name	D02
Input contributions	M56730, M63015
Length & frame rate	97 frames (30 fps)
Number of source views	15 (5x3)
Source view resolution	1920 × 1080
Texture format	YUV 4:2:0 10 bits
Depth format	YUV 4:2:0 16 bits
Depth range	[1.8 m, 15 m], normalized disparity
View FoV	66° × 40°
Camera spacing	20 cm horizontally, 23 cm vertically

Barn

The general characteristics of the Barn sequence are summarized in Table 20 form a 5x3 planar array.

Table 20: Summary of the Barn sequence

Category – Short name	D03
Input contributions	M56730
Length & frame rate	97 frames (30 fps)
Number of source views	15 (5x3)
Source view resolution	1920 × 1080
Texture format	YUV 4:2:0 10 bits
Depth format	YUV 4:2:0 16 bits
Depth range	[2.5 m, 25 m], normalized disparity
View FoV	66° × 40°
Camera spacing	20 cm horizontally, 23 cm vertically

Frog

The general characteristics of the Frog sequence are summarized in Table 21. The captured views form a 15x1 line following left to right scan order. The refined depths proposed in [m47445] are used; these depths do not exist for extreme view positions v0 and v14 and therefore only the views from v1 to v13 are used.

Table 21: Characteristics of the Frog sequence

Category – Short name	E01
Input contribution	WG 11 M43748, WG 11 M44914 and WG 11 M47445
Length & frame rate	300 frames (30 fps)
Number of source views	13 (13x1)
Source view resolution	1920x1080
Texture format	YUV 4:2:0 10 bits
Depth format	YUV 4:2:0 16 bits
Depth range	[0.3 m, 1.62 m], normalized disparity
View FoV & mapping	63.65° × 38.47° Rectilinear
Lens	2.16 mm
Camera spacing	3.675 cm

The viewing space volume is a rectangle centered at position [0, 0, 0] meter with a 15cm width, 44.1cm length, and no z component.

$$rect\left(\frac{x-7.5cm}{15\ cm}, \frac{y}{44.1cm}, 0\right)$$

Carpark

The general characteristics of the Carpark sequence are summarized in Table 22. The captured views form a 9x1 line and are numbered v0 to v8 following left to right scan order.

The scene unit (u) is unknown, but in the same order as a meter. The reported camera spacing does not match with the camera extrinsics.

Table 22: Characteristics of the Carpark sequence

Category – Short name	E02
Input contribution	WG 11 M51598
Length & frame rate	250 frames (25 fps)
Number of source views	9 (9x1)
Source view resolution	1920x1088
Texture format	YUV 4:2:0 10 bits
Depth format	YUV 4:2:0 16 bits
Depth range	[3.45 u, 276 u], normalized disparity
View FoV & mapping	63° × 48°
Lens	4.5 mm
Camera spacing	13.75 cm

Street

The general characteristics of the Street sequence are summarized in Table 23. The captured views form a 9x1 line and are numbered v0 to v8 following left to right scan order.

The scene unit (u) is unknown, but in the same order as a meter. The reported camera spacing does not match with the camera extrinsics.

Table 23: Characteristics of the Street sequence

Category – Short name	E03
Input contribution	WG 11 M51598
Length & frame rate	250 frames (25 fps)
Number of source views	9 (9x1)
Source view resolution	1920x1088
Texture format	YUV 4:2:0 10 bits
Depth format	YUV 4:2:0 16 bits
Depth range	[3.45 u, 276 u], normalized disparity
View FoV & mapping	63° × 48°
Lens	4.5 mm
Camera spacing	13.75 cm

Fencing

The general characteristics of the Fencing sequence are summarized in Table 24. The captured views form a 10x1 linear arc and are numbered following left to right scan order.

Warning: Fencing textures have changed from WG 11 N 19484 to WG 11 N 19679.

Warning: Fencing depth maps have changed from WG 11 N 19484 to WG 11 N 19679.

Warning: Fencing length has changed from N0372 to N0406

Warning: Fencing depth maps have changed from N0372 to N0406

Table 24: Characteristics of the Fencing sequence

Category – Short name	L01
Input contribution	WG 11 M38247
Length & frame rate	65 frames (25 fps)
Number of source views	10
Source view resolution	1920x1080
Texture format	YUV 4:2:0 10 bits
Depth format	YUV 4:2:0 16 bits
Depth range	[3.0 m, 7.0 m], normalized disparity
View FoV & mapping	63° × 48°
Lens	4.5 mm
Camera spacing	5 stereopairs (baseline: 22 cm) placed on arc (radius: 4 m), angle between neighboring stereopairs: 15 degrees, total angle of the system: 60 degrees

CBABasketball

The general characteristics of the CBABasketball sequence are summarized in Table 25. The captured views form a 34x1 linear arc.

Table 25: Characteristics of the CBABasketball sequence

Category – Short name	L02
Input contribution	M58275
Length & frame rate	97 frames (30 fps)
Number of source views	34
Source view resolution	2048x1088
Texture format	YUV 4:2:0 10 bits
Depth format	YUV 4:2:0 16 bits
Depth range	[100 u, 1000 u], normalized disparity

MartialArts

The general characteristics of the MartialArts sequence are summarized in Table 26. The cameras were arranged in stereopairs placed on an arc, on two different elevations, as shown in Figure 8.

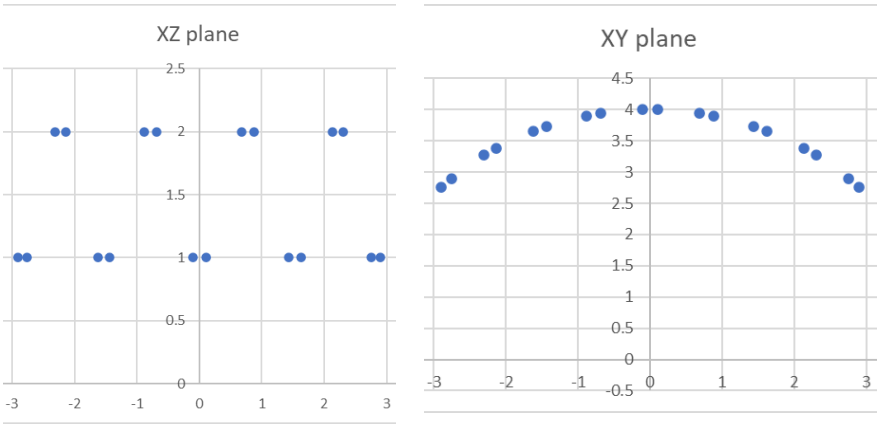

Figure 8: Visualization of the view positions of the MartialArts sequence

Table 26: Characteristics of the MartialArts sequence

Category – Short name	L03
Input contribution	M61949
Length & frame rate	97 frames (25 fps)
Number of source views	15
Source view resolution	1920x1080
Texture format	YUV 4:2:0 10 bits
Depth format	YUV 4:2:0 16 bits
Depth range	[0.8 u, 500 u], normalized disparity
Camera spacing	Arc radius: 4 m Stereopair baseline: 0.8 m Vertical distance between neighboring stereopairs: 1 m Horizontal distance between neighboring stereopairs: 1 m