**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC1/SC29/WG11**

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC1/SC29/WG11 N16532**

**October 2016, Chengdu, CN**

|  |  |
| --- | --- |
| **Source** | **Requirements** |
| **Status** | **Final** |
| **Title** | **Call for light field test material including plenoptic cameras and camera arrays** |
| **Author** | Arianne T. Hinds, Didier Doyen, Gauthier Lafruit |

# Introduction

MPEG is conducting experiments to study compression technologies for light field data for use in applications that provide an increased sense of immersion beyond what existing video coding standards can provide. A typical application is Six Degrees-of-Freedom (6-DoF) Virtual Reality (VR) where one freely walks inside a volume, surrounded by a set of fixed cameras, with a system synthesizing the proper images into the user’s VR goggles, cf. Figure 1.



Figure 1: 6-DoF VR, navigating both along the peripheral path described by the cameras, and an axial path inwards into the scene.

This call for test material seeks to increase MPEG’s existing content library with additional test material for the purpose of conducting those experiments. Within this call, certain types of camera arrangements and content types are solicited. If, however, not all requirements can be met, content providers are encouraged to submit the content anyway to the extent that they are able to meet the submission requirements for the camera configurations and content.

***Note: The conditions for usage of the test material should preferably involve as few restrictions as possible, and in particular must enable the use of the material for the standardization development work of ISO/IEC JPEG, ISO/IEC MPEG, and ITU-T VCEG. It is also highly desirable for the content to be able to be used for general research, development, and demonstration of video or image signal processing technology and for snapshots to be allowed to be included in technical publications.***

# Overview of test material and camera devices

## Static image and video sequences

Both static image and video sequence content are solicited. Static content is solicited specifically for high density multiple view images typically captured with one high-resolution SLR positioned over adjacent capture positions, as close as 1 mm from each other, with a robotic arm. Video content is solicited for fixed 2D camera arrays with inter-camera distances of at least couple of cm. High frame rates of at least 25 fps are solicited. Video content is preferred from plenoptic cameras, possibly at lower frame rates.

Both indoor and outdoor scenes are solicited.

## Natural vs. Computer-generated content

Natural content is highly preferred over computer-generated content. Depth material for all content is solicited, if available, using depth sensing devices and/or depth estimation techniques.

## Color components, depth, and metadata

In general, it is required that texture, depth, and metadata be provided separately. All texture content should be provided in the three-channel format of RGB, YUV444 or YUV420. Depth information, if available, should be provided. Metadata, when available, should be provided.

## Types of cameras and camera array arrangements

The types of cameras that are envisioned to be relevant for this call include:

1. Robotic arm cameras that capture a highly dense array of images along a predefined 2D track (e.g. 2D linear with parallel cameras, 2D linear with convergent cameras, 2D cylindrical, 2D spherical) thereby producing the equivalent of a 2D fixed camera array. In this arrangement, a single camera moves along a predefined track both in horizontal and vertical directions capturing an image of a static scene at each position of the camera. The increments by which the camera is moved by the robot arm should be in the order of 1mm to 1 cm (physically impossible to put distinct cameras at such short distances from each other, hence the use of a single camera and a static scene). High SLR resolutions are solicited, i.e. at least 1920x1080 HD resolution, but higher resolutions are preferred, e.g. 20 Mpixels or more (supporting spatial downsampling studies identifying the break-down of some algorithms below a resolution threshold)
2. Regularly spaced, fixed 2D camera arrays with a minimum of 2 rows of at least 10 cameras and up to 100 cameras in each row. Arrays of higher vertical density are preferred (to support zoom-in/out functionality studies), but not required. The spatial resolution of the cameras should be around full-HD (1920x1080) or higher. Refer to the section on additional camera specifications for 2D camera array arrangements in section [4].
3. Plenoptic cameras such as the Lytro Cinema or Raytrix, or other cameras where a microlens is inserted between the main lens and the image sensor.

The type of cameras used should be reported with data sheets, if possible. In particular, for video it’s important to report whether a global or rolling shutter camera type was used. Information about intrinsic properties of the devices should be reported if available. Even if the views were supposed to be captured over a regular grid, additional registration (alignment to a perfectly regular grid in post-processing) is preferred, but not mandatory. Color calibration over all views, reaching the same color tones over all views, is highly recommended.

# Single camera and robot arrangement

Indoor scene content with consistent lighting is recommended. High density camera arrays shall be dense enough to help us understand what is ‘the minimum camera density’ that supports proper processing in subsequent modules. One starts with an array that is known to be ‘too dense’, typically for a static scene, and then studies how ‘less dense’ the input views can become before any subsequent algorithm (e.g. depth estimation, view synthesis, etc) starts to break down.

# Fixed 2D camera array arrangement

The cameras should be in a 2D arrangement, as far as physically possible. Dense video sequences are particularly sought with a baseline distance between cameras in the order of 5cm.

In such a setup as illustrated in Figure 1, the distance between the outer cameras (DL-R) is determined by the acquisition angle 1 and the viewing distance D as:

DL-R= 2 \* tan (1/2)\*D

For 20° angle and a viewing distance of 3 m, the recommended distance between the outer cameras is DL-R= 104 cm.

It is recommended to keep the depth range of objects in the scene within certain limits. If the cameras capture a width of the scene Wsc, then the depth of the objects captured is preferred to stay within the range of DR\*Wsc, with a suggested value of 0.5 for DR.



Figure Capturing setup

**Rectification**

If rectification is necessary to provide views in linear and parallel arrangement, it is recommended to provide both original and rectified data. . Rectification should be done as well as possible using high quality algorithms.

**Camera types**

The type of cameras used should be reported with data sheets, if possible. Information about intrinsic properties of the devices should be reported if available. Global shutter cameras are recommended in video acquisitions.

**Synchronization**

Accurate temporal synchronization of the multiple cameras is required. This means that all cameras shall take images at the same time with an external global clock signal

**Resolution**

For static images, it is recommended to use high-resolution SLRs (20 Mpixels or more); a resolution of at least 1920x1080 HD is mandatory. Video sequences should have a spatial resolution of around 1920x1080 full-HD, and a frame rate of at least 25 fps.

Nevertheless, higher spatial resolutions (UHD or more) are highly desirable. The video sequences should be in progressive format, i.e. non-interlaced. Temporal resolutions should be at least 25 fps, with a multiple of 30 fps as preferred option (30 fps, 60 fps, 90 fps, 120 fps).

**White balancing, Color consistency**

Color consistency among the multiple views should be adjusted as accurately as possible, prior to capturing or during post-processing (color calibration). The multiple cameras should be white balanced (gamma correction). Most modern cameras provide functionalities for this. It can also be done, e.g., as described in [2].

**Content**

The content captured itself should be representative and challenging for coding algorithms. The scene should include sufficient texture variations, moving objects and complex scene (depth) structures. Indoor and outdoor scenes with varying lighting conditions are appreciated. Comments on the design of the test scene should be given in the annex.

**Camera calibration**

Accurate camera calibration information about extrinsic, intrinsic and sensor plane parameters is required. If the proponent cannot provide these parameters, actions should be taken to extract these parameters by self-calibration in a post-processing phase. These parameters shall be provided as specified in the annex below.

Calibration charts (checkerboards and or color charts) covering not more than 1/10th of the captured scene are allowed, and highly recommended if the proponent cannot provide calibrated/registered sequences.

**Depth data**

High resolution depth maps (at the same spatial resolution as the RGB data) are solicited. Depth maps can be estimated and/or sensed with special devices (time-of-flight, structured light, etc). Though it is expected that the depth maps should have the same spatial resolution as the RGB images, low-resolution depth maps provided by depth sensing devices, are valuable metadata to create high resolution depth maps for further processing in the pipeline. It is recommended to provide both depth sensed data (current technology typically provides such depth maps at low resolution, but the higher the resolution the better) as well as depth estimated data (stereo matching, etc) separately, to support studies about the depth map resolution trade-offs.

**Depth specification**

The necessary data for correct interpretation of depth values shall be provided. This includes near and far clipping planes (Z\_near, Z\_far), as well as definition if the given data are directly Z values (depth) or rather 1/Z (disparity). If any other type of data is provided, this shall be fully specified, including algorithms how to convert to depth or disparity values.

# Plenoptic cameras

Plenoptic cameras are defined as acquisition systems including an array of micro-lenses between the main lens and the sensor. The density of micro-lenses is supposed to be large enough to ensure a good angular sampling of the light field. The size of the sensor is supposed to be large enough to get significant difference between viewpoints to target applications such as VR with 6 degree of freedom.

**Résolution**

Video acquisitions at the highest possible frame rate the device can sustain during 10 seconds are solicited. 25 fps or more are preferred, but almost static sequences (time-lapse effect) are acceptable with durations that are proportionally increased, e.g. if only 10 fps (=25fps/2.5) can be reached, the video duration should be increased up to 25 seconds (=10s\*2.5).

Number of pixel per frame is not defined although a minimum spatial resolution is required to address use cases such as VR.

**Content**

The content captured itself should be representative and challenging for coding algorithms. The scene should include sufficient texture variations, moving objects and complex scene (depth) structures. Indoor and outdoor scenes with varying lighting conditions are appreciated. Comments on the design of the test scene should be given in the annex.

**Depth data**

High resolution depth maps (at the same spatial resolution as the RGB data) are solicited. Depth maps can be estimated and/or sensed with special devices (time-of-flight, structured light, etc). Though it is expected that the depth maps should have the same spatial resolution as the RGB images, low-resolution depth maps provided by depth sensing devices, are valuable metadata to create high resolution depth maps for further processing in the pipeline. It is recommended to provide both depth sensed data (current technology typically provides such depth maps at low resolution, but the higher the resolution the better) as well as depth estimated data (stereo matching, etc) separately, to support studies about the depth map resolution trade-offs.

**Depth specification**

The necessary data for correct interpretation of depth values shall be provided. This includes near and far clipping planes (Z\_near, Z\_far), as well as definition if the given data are directly Z values (depth) or rather 1/Z (disparity). If any other type of data is provided, this shall be fully specified, including algorithms how to convert to depth or disparity values.

**Data format and metadata**

Plenoptic camera content may be delivered in different forms. RAW data may be provided with any associated metadata relevant to exploit the content. Although whatever the application RAW data may be transformed in other data format (e.g. muti-view format) before rendering, the RAW format is anyway of interest to evaluate coding technologies. In this case metadata should include information about the lens structure, the different focal length, the vignetting.

The pre processing of RAW content into for instance a multi-view format may also be the delivery format. Metadata should be associated to get extrinsic and intrinsic parameters of the multi-view.

**Player for plenoptic video**

In order to review the content at standards meetings, a suggested player that is available free-of-charge (e.g. open source) is asked to be identified for each submission. Alternatively, the content provider may provide an executable that can be used to play back the content. Please specify the system requirements that are required for player executables (if provided) that are provided with plenoptic video content submissions.

# Contacts and other logistics

For questions or to respond to this call, please contact:

 Gauthier Lafruit: Gauthier.lafruit@ulb.ac.be

 Arianne Hinds: a.hinds@cablelabs.com

 Didier Doyen: Didier.doyen@technicolor.com

Submissions in response to this call are kindly asked to be made by the 117th MPEG meeting which is planned for 14-20 January 2017.

# References

[1] FTV Software Framework, ISO/IEC JTC1/SC29 WG11 N15349, June 2015, Warsaw, Poland

[2] Call for Contributions on 3D Video Test Material (Update), ISO/IEC JTC1/SC29 WG11 N9595, January 2008, Antalya, Turkey

# Annex 1: Camera parameter formats

Camera parameters principles

The camera parameters are subdivided into two categories:

* the camera intrinsics, i.e. the parameters of each camera individually: camera center c position in world coordinate system, principal point *p* and focal length *f* in pixel units, as shown in the top part of Figure 26 (picture from [CV1])
* the camera extrinsics, i.e. the camera rotation R and translation T, expressed in world coordinates, as shown in the bottom part of Figure 26



Figure 27: Intrinsics and extrinsics of cameras

Literature often uses the [R|c] matrix at the bottom of Figure 27 for the extrinsics. However, our depth estimation and view synthesis software uses a slightly different format, replacing c by the camera position T = - R-1 \* c, according to [N9595], which annex A is copied in the next section for completeness.

Specification of Camera Parameters

Camera parameters shall be specified as rotation matrix **R**, translation vector **t** and intrinsic matrix **A** for each camera i. Values shall be given in floating point precision as accurate as possible. More illustration and explanation is given in Annex B below.

The extrinsic camera parameters **R** and **t** shall be specified according to a right-handed coordinate system. The upper left corner of an image shall be the origin for corresponding image/camera coordinates, i.e., the (0,0) coordinate, with all other corners of the image having non-negative coordinates. With these specifications, a 3-dimensional world point, **wp**=[x y z]T is mapped into a 2-dimensional camera point, **cp** = s \* [u v 1] T, for the i-th camera according to:

s \* **cp**(i) = **A**(i) \* **R**-1(i) \* [**wp** – **t**(i)]

where **A**(i) denotes the intrinsic camera parameters, **R**-1(i) denotes the inverse of the rotation matrix **R**(i) and s is an arbitrary scaling chosen to make the third coordinate of **cp** equal to one.

The rotation matrix **R**(i) for i-th camera is represented as follows.

|  |  |  |
| --- | --- | --- |
| **r\_11**[i] | **r\_12**[i] | **r\_13**[i] |
| **r\_21**[i] | **r\_22**[i] | **r\_23**[i] |
| **r\_31**[i] | **r\_32**[i] | **r\_33**[i] |

The translation vector **t**(i) for i-th camera is represented as follows:

|  |
| --- |
| **t\_1**[i] |
| **t\_2**[i] |
| **t\_3**[i] |

The rotation matrix **R** and the translation vector **t** define the position and orientation of the corresponding camera with respect to the world coordinate system. The components of the rotation matrix **R** are function of the rotations about the three coordinate axes.

**focal\_length\_x**[i] specifies the focal length of the i-th camera in the horizontal direction units of pixels.

**focal\_length\_y**[i] specifies the focal length of the i-th camera in the vertical direction in units of pixels.

**principal\_point\_x**[i] specifies the principal point of the i-th camera in the horizontal direction units of pixels.

**principal\_point\_y**[i] specifies the principal point of the i-th camera in the vertical direction in units of pixels.

**radial\_distortion**[i] specifies the radial distortion coefficient of the i-th camera.

The intrinsic matrix **A**(i) for i-th camera is represented as follows:

|  |  |  |
| --- | --- | --- |
| **focal\_length\_x**[i] | **radial\_distortion**[i] | **principal\_point\_x**[i] |
| 0 | **focal\_length\_y**[i] | **principal\_point\_y**[i] |
| 0 | 0 | 1 |