**Source: Xiaomi**

**Title: RGBD content format and proposed device supports**

## Document for: Discussion

## Agenda Item: 9.5

# 1 Introduction

As explained in clause 2 of this contribution, several immersive applications which are discussed in the context of MeCAR and other work items are making use of the type of content called RGBD sequence. RGBD content is composed of a video sequence associated with a sequence of depth maps.

There is currently a technical gap in the technologies being discussed in the MeCAR Work Irem to enable such application making use of RGBD in a interoperable manner.

In order to enable those applications, this contribution proposes to define:

* The support of RGBD content as a media capability (including carriage)
* The corresponding required capabilities regarding RGBD support in the current MeCAR AR device categories, i.e. mandatory for Thin AR UE and AR UE and optional for XR UE.

The proposed text is written against the latest version of the MeCAR Permanent Document (v 5.1.0). Also, note that the proposed text largely leverages the content of TR 26.998 for the generic description of depth maps.

# 2 Background

# 2.1 Media Applications: Overview

In this document, we examine the use of RGBD content in different AR applications that are to be targeted by the MeCAR Work Item.

**Table 1 - Requirements form typical media applications**

|  |  |  |
| --- | --- | --- |
| **Applications**  | **Application type** | **RGBD usage**  |
| AR Video Conferencing (human-to-human communication)  | Real-time | Background removal, 2.5D user view, effects ….   |
| AR video call  | Real-time | Background removal, 2.5D user view, effects …. |
| Augmented Reality Utilities (e.g. live translation, browsing, messaging)  | Real-time/On-demand | Example: for generating volumetric video |
| Split Rendering  | Real-time | Used for 2.5D pre-rendered view of the scene   |

2.2 Media Applications: Analysis

2.2.1 Video conferencing

Video conferencing refers to the AR UE user being able to see and talk to the remote participant(s) of the teleconferencing. A type of video conferencing is the AR Video Call examined in the following paragraph.

2.2.2 AR Video Call

In AR Video Call, the user(s) are able to communicate with each other using video conferencing but with added AR functionalities, like background removal, scene blending, perspective change etc. In order to achieve these functionalities depth perception is required, typically achieved by transmitting a Depth stream, that is easy to capture (it can be even image-based generation), yet sufficing for all possible applications, since the frontal view of the user is only needed. Due to the low-latency nature of the application lightweight coding techniques are expected. This holds especially for “Device type 1: Standalone physically-constrained AR glasses”, that are energy and processing limiting.

2.2.3 Augmented reality utilities

Due to the fact that AR UEs are not a commodity yet, AR Utilities refers to applications that are currently deployed (e.g. in smartphone AR) and can be adapted to fit AR UEs. This includes image/audio-based translation, messaging, browsing, video streaming, gaming etc.

As an example from the PD, in order to support volumetric video (V3C), RGBD is needed to compose it; following is “Figure 22 - High level overview of a V3C system for one way conversational communication.” from the PD, that shows the upstream RGBD transmission:



2.2.4 Split rendering

Split rendering may include any application that offloads (part of) the rendering process to a processor connected to -but not located on- the AR UE. In this case RGBD might be sent to the AR device, in order for the Depth to be used for view correction and/or generating stereoscopic views.

As an example, following is an entry from “Table 2 – AR/MR Data type definitions” of the PD:

|  |  |  |  |
| --- | --- | --- | --- |
| **AR/MR Data category** | **AR/MR** **Data type** | **Definition** | **Media type description (Examples)** |
|  | **Split rendered 2D depth map video** | Monochrome 2D auxiliary video containing depth map in the same perspective view with primary picture. | **Type:** 2D auxiliary video with depth type defined in I.13.2.3 of ISO/IEC 14496-10 [X] and G.14.3.3 of ISO/IEC 23008-2 [Y].**Organization:** MPEG |

2.3 Media Applications: Live vs. On-Demand

On-demand content can be prepared in advance for consumption from an AR device, in such cases, the complexity of generating the content is not very important, compared to the asset size (to conserve bandwidth) and complexity of displaying the content (to conserve AR device battery). Therefore, even for on-demand applications it might make sense (depending on the device type), to perform remote rendering and transmit the resulting in a simpler representation as RGBD to the UE.

RGBD can be very useful in live application scenarios, especially UE to UE communication where generating more complex representations can induce delay and quickly deplete battery level.

# 3 Proposed changes in MeCAR PD

CHANGE #1

6.6 System capabilities

6.6.1 Support of RGBD content

6.6.1.1 General

RGBD content refers to a data sequence composed of a video sequence and a depth map sequence that share a known temporal and spatial relation. A depth map may be represented as a video frame of a video sequence in which case each pixel of this depth map sequence may represent a measure of the distance between the surface of an AR object, point (A) and the camera centre (C). Conventionally, the distance is represented by the coordinate of the point on the z-axis obtained by the orthogonal projection of the point (A) on this axis, here denoted as the point (A’). The measured distance is thus the length of the segment (CA’) as depicted in Figure 6.6.6-1.

**Figure 6.6.6-1: Pixel representation of depth images**

A depth map thus contains pixels with the distance attribute (e.g., depth). Distance is one-dimensional information and may be represented in an absolute/relative or linear/non-linear manner. Metadata to interpret the pixels of a depth map image may be provided as well as to determine the spatial and temporal relationship between pixels of the depth video sequence and pixels of the texture video sequence.

6.6.1.2 Metadata describing RGBD content

6.6.1.2.1 OpenXR extension XR\_KHR\_composition\_layer\_depth

As described in the OpenXR specification, the multi-vendor XR\_KHR\_composition\_layer\_depth extension defines “an extra layer type which allows applications to submit depth images along with color images in projection layers. The XR runtime may use this information to perform more accurate reprojections taking depth into account. Use of this extension does not affect the order of layer composition as described in Compositing”.

The depth information provided as an extra layer is thus described with the data structure and is defined as follows:

// Provided by XR\_KHR\_composition\_layer\_depth

typedef struct XrCompositionLayerDepthInfoKHR {

 XrStructureType type;

 const void\* next;

 XrSwapchainSubImage subImage;

 float minDepth;

 float maxDepth;

 float nearZ;

 float farZ;

} XrCompositionLayerDepthInfoKHR;

Member Descriptions

* type is the [XrStructureType](https://registry.khronos.org/OpenXR/specs/1.0/html/xrspec.html%22%20%5Cl%20%22XrStructureType) of this structure.
* next is NULL or a pointer to the next structure in a structure chain. No such structures are defined in core OpenXR or this extension.
* subImage identifies the depth image [XrSwapchainSubImage](https://registry.khronos.org/OpenXR/specs/1.0/html/xrspec.html%22%20%5Cl%20%22XrSwapchainSubImage) to be associated with the color swapchain. The swapchain **must** have been created with a faceCount of 1.
* minDepth and maxDepth are the window space depths that correspond to the near and far frustum planes, respectively. minDepth must be less than maxDepth. minDepth and maxDepth must be in the range [0, 1].
* nearZ and farZ are the positive distances in meters to the near and far frustum planes, respectively. nearZ and farZ **must** not be equal. nearZ and farZ **must** be in the range (0, +infinity].

6.6.1.3 Compression of RGBD information

6.6.1.3.1 Video-based depth compression

As described in clause 6.6.1.1, RGBD is essentially composed of a texture video sequence along with a depth sequence. This depth sequence is represented as a sequence of depth frame which constitute a video sequence as well.

In terms of compression, the video coding standard HEVC/H.265 has provisions to carry in a compressed form depth video sequence as auxiliary picture of type AUX\_DEPTH.

Since a depth video sequence is essentially a monochrome video sequence, monochrome profiles defined in HEVC (Monochrome, Monochrome 12, Monochrome 16) are also thus relevant for coding the depth sequence.

6.6.1.4 Processing of RGBD content

tbd

6.6.1.5 Carriage of RGBD content over RTP

6.6.1.5.1 Option 1: Single RTP stream

A first option is to carry the RGBD sequence in a single RTP stream which means that both sequences can be part of the same elementary stream. In order to do so, a possible option is to use the concept of independent layers when present in video codec. For instance with the HEVC codec, both the texture and the depth sequences may constitute different layers of the same elementary stream.

6.6.1.5.2 Option 2: Two RTP streams

In this second option, the RGBD sequence is transmitted over two concurrent RTP streams. Both video streams may be encoded with different encoding characteristics and even different codecs. The temporal relation between both sequence will be given by the timestamps.

6.6.1.6 Storage of RGBD content in ISOBMFF

6.6.1.6.1 Current State

Currently ISO Base Media File Format (ISOBMFF) supports accompanying video with depth via auxiliary tracks – with reference\_type as ’auxl’ or ‘vdep’ (or both). This allows to signal a texture video track along with a depth map video track. However, further metadata is needed to be able to interpret the depth map pixels. To this end, ISO/IEC 23002-3 “Representation of auxiliary video and supplemental information” defines such metadata, e.g. stride, near/far plane, to be carried as item of type ‘auvd’ in the auxiliary video metadata of ISOBMFF spec. These two specifications combined merely provide a partial solution for a RGBD storage and would require further specification in order to ensure interoperability for the applications targeted in MeCAR, e.g. split rendering.

6.6.1.6.2 Possible encapsulation analysis6

Even though there are already some tools in ISOBMFF for RGBD storage, there is lack of a single RGBD storage format that is a content-centric, but application and codec agnostic which would enable the encapsulation of RGBD content.

Some example depth information that must be available prior to accessing the data for immersive real time communication scenarios would be:
- Depth range
- Depth projection properties
- Depth coding properties

- Projection format of depth image

- Disparity between texture and depth image

CHANGE #2

4.2 Thin Augmented Reality User Equipment (Thin AR UE)

…

4.2.5 System capabilities

[Ed. note: The carriage in 6.6.1.3 needs to be further detailed but the following express the intent in terms of support]

The Thin AR UE shall support the carriage of RGBD content over RTP as defined in 6.6.1.3.

CHANGE #3

4.3 Augmented Reality User Equipment (AR UE)

…

4.3.3 System capabilities

[Ed. note: The carriage in 6.6.1.3 needs to be further detailed but the following express the intent in terms of support]

The AR UE shall support the carriage of RGBD content over RTP as defined in 6.6.1.3.

CHANGE #4

4.4 eXtended Reality User Equipment (XR UE)

…

4.4.3 System capabilities

[Ed. note: The carriage in 6.6.1.3 needs to be further detailed but the following express the intent in terms of support]

The XR UE should support the carriage of RGBD content over RTP as defined in 6.6.1.3.

END OF CHANGES