Status of the Immersive Media Standard:
Related to Test Model for Immersive Video (TMIV) for 3DoF+

Presenter: Jong-Beom Jeong (uof4949@skku.edu)
Jong-Beom Jeong, SoonBin Lee, Il-Woong Ryu, Dongmin Jang, Eun-Seok Ryu

Multimedia Computing Systems Lab. (MCSL)
http://mcsl.skku.edu
Department of Computer Education
Sungkyunkwan University
Step-by-step objective of ISO/IEC MPEG Immersive Video

- MPEG-I is responsible for standardizing immersive media in MPEG and specifies the goals of step 3.
- Goal of Revitalizing VR Commercial Service by 2020
- Goal of 6DoF media support by 2022 after completing 3DoF standard by 2017

### Immersive Media Standard

**Step 1**
- Complete 3DoF standard by 2017
- Rotate head in a fixed state
- 360 video full streaming by default
- Tiled streaming if possible

**Step 2**
- Enable VR commercial servicers by 2020
- Allow head rotation and movement within a restricted area
- User-to-user conversations and projection optimization

### Immersive Media Standard

**Step 3**
- Support 6DoF by 2022
- 6DoF video will reflect user’s walking motion
- Support interaction with virtual environments
Immersive Media Standard Roadmap

Media Coding

- Genome Compression
- Descriptors for Video Analysis (CDVA)
- Neural Network Compression for Multimedia
- Color Support in Open Font Format
- Essential Video Coding
- Low Complexity Enhancement Video Coding
- Versatile Video Coding
- 6 DoF Audio
- Dense Representation of Light Field Video
- Video with 6 DoF
- Point Cloud Compression v.2
- PCC Systems Support
- Immersive Media Scene Description Interface
- OMAF v.2
- VR 360

Immersive Media Standards

- Media Orchestration
- Network-Based Media Processing
- Multi-Image Application Format
- CMAF v.2
- Partial File Format
- Web Resource Tracks
- Internet of Media Things

Systems and Tools

- Beyond Media
- Systems
- and Tools
3DoF+ CfP (w18097)

- **Background**
  - MPEG defined degree of freedom of VR as 3DoF, 3DoF+, and 6DoF
  - Limited movements for user sitting in a chair is available for 3DoF+

- **Requirements**
  - Solution will be built on HEVC with 3DoF+ metadata (included in MPEG-I part 7)
  - Both objective and subjective quality evaluation will be performed
3DoF+ CfP (w18097) (-Cont’d)

- Test Sequences

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Class</th>
<th>Resolution</th>
<th>No. of views</th>
<th>Frame count</th>
<th>Frame rate</th>
<th>Source FoV</th>
</tr>
</thead>
<tbody>
<tr>
<td>ClassroomVideo</td>
<td>A</td>
<td>4096x2048</td>
<td>15</td>
<td>120</td>
<td>30</td>
<td>360° x 180°</td>
</tr>
<tr>
<td>TechnicolorMuseum</td>
<td>B</td>
<td>2048x2048</td>
<td>24</td>
<td>300</td>
<td>30</td>
<td>180° x 180°</td>
</tr>
<tr>
<td>TechnicolorHijack</td>
<td>C</td>
<td>4096x4096</td>
<td>10</td>
<td>300</td>
<td>30</td>
<td>180° x 180°</td>
</tr>
<tr>
<td>TechnicolorPainter</td>
<td>D</td>
<td>2048x1088</td>
<td>16</td>
<td>300</td>
<td>30</td>
<td>46° x 25°</td>
</tr>
<tr>
<td>IntelKermit</td>
<td>E</td>
<td>1920x1080</td>
<td>13</td>
<td>300</td>
<td>30</td>
<td>77.8° x 77.8°</td>
</tr>
</tbody>
</table>
3DoF+ CfP (w18097) (-Cont’d)

- **Software**

<table>
<thead>
<tr>
<th>Software name</th>
<th>Location</th>
<th>Tag/branch</th>
</tr>
</thead>
<tbody>
<tr>
<td>RVS</td>
<td><a href="http://mpegx.int-evry.fr/software/MPEG/Explorations/3DoFplus/RVS">http://mpegx.int-evry.fr/software/MPEG/Explorations/3DoFplus/RVS</a></td>
<td>v3.1</td>
</tr>
<tr>
<td>WS-PSNR</td>
<td><a href="http://mpegx.int-evry.fr/software/MPEG/Explorations/3DoFplus/WS-PSNR">http://mpegx.int-evry.fr/software/MPEG/Explorations/3DoFplus/WS-PSNR</a></td>
<td>v2.0</td>
</tr>
<tr>
<td>HDRTools</td>
<td><a href="https://gitlab.com/standards/HDRTools">https://gitlab.com/standards/HDRTools</a></td>
<td>v0.18</td>
</tr>
<tr>
<td>360Lib</td>
<td><a href="https://jvet.hhi.fraunhofer.de/svn/svn_360Lib">https://jvet.hhi.fraunhofer.de/svn/svn_360Lib</a></td>
<td>360Lib-5.1-dev</td>
</tr>
<tr>
<td>HM</td>
<td><a href="https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware">https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware</a></td>
<td>HM-16.16</td>
</tr>
</tbody>
</table>

- **Target Bitrate**

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Rate 1</th>
<th>Rate 2</th>
<th>Rate 3</th>
<th>Rate 4</th>
<th>Rate 5</th>
<th>Rate 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>ClassroomVideo</td>
<td>6.5</td>
<td>10</td>
<td>15</td>
<td>25</td>
<td>40</td>
<td>65</td>
</tr>
<tr>
<td>TechnicolorMuseum</td>
<td>10</td>
<td>15</td>
<td>25</td>
<td>40</td>
<td>65</td>
<td>100</td>
</tr>
<tr>
<td>TechnicolorHijack</td>
<td>6.5</td>
<td>10</td>
<td>15</td>
<td>25</td>
<td>40</td>
<td>65</td>
</tr>
<tr>
<td>TechnicolorPainter</td>
<td>6.5</td>
<td>10</td>
<td>15</td>
<td>25</td>
<td>40</td>
<td>65</td>
</tr>
<tr>
<td>IntelKermit</td>
<td>4</td>
<td>6.5</td>
<td>10</td>
<td>15</td>
<td>25</td>
<td>40</td>
</tr>
</tbody>
</table>
Response on 3DoF+ CfP

For response of 3DoF+ CfP, 5 input documents are proposed

All of proposals introduce pruning and packing architecture

These technologies are included in test model for 3DoF+

<table>
<thead>
<tr>
<th>Number</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>m47179</td>
<td>Philips response to 3DoF+ Visual CfP</td>
</tr>
<tr>
<td>m42372</td>
<td>Description of Nokia’s response to CFP for 3DOF+ visual</td>
</tr>
<tr>
<td>m47407</td>
<td>Technical description of proposal for Call for Proposals on 3DoF+ Visual prepared by Poznan University of Technology (PUT) and Electronics and Telecommunications Research Institute (ETRI)</td>
</tr>
<tr>
<td>m47445</td>
<td>Technicolor-Intel Response to 3DoF+ CfP</td>
</tr>
<tr>
<td>m47684</td>
<td>Description of Zhejiang University’s response to 3DoF+ Visual CfP</td>
</tr>
</tbody>
</table>
Test Model for Immersive Video (TMIV)

• Encoder

  • TMIV encoder consists of three main steps:
  • (1) View optimizer
  • (2) Atlas constructor
  • (3) Video encoder & Metadata composer
Test Model for Immersive Video (TMIV) (-Cont’d)

- **Encoder**
  - Atlas constructor references the occupancy map of each views
  - Minimizes the atlas size by overlapping the patches

Example of TMIV encoder process[1]

Test Model for Immersive Video (TMIV) (-Cont’d)

• **View Optimizer**
  - Determines the basic views and the additional views
  - Select two views (=one pair) which have the largest direction deviation
  - They have the maximum value of $\theta(i, j)$ where $i, j$ are the indices the source views
  - If overlap $\geq 0.5 \times \min(FOV_i, FOV_j)$, only one basic view is selected
  - Else, multiple basic views including view m and view n are selected

![Example of the directions deviations](image)

Test Model for Immersive Video (TMIV) (-Cont’d)

- **Atlas Constructor**
  - Atlas: aggregation of patches after packing process
  - The atlas constructor is composed of three parts:
    - (1) Pruner, (2) Aggregator, and (3) Patch packer

Test Model for Immersive Video (TMIV) (-Cont’d)

- Pruner
  - De-project depth from view j to 3D space and project them to the reference view l
  - A sample is pruned if it is already covered by the previous view, determined by:
    \[ |z - z_p| < \text{RedundancyFactor} \times \min(z, z_p) \]
  - Default value of RedundancyFactor is 0.02

Example of pruning with 2 basic views and 3 additional views[1]

Test Model for Immersive Video (TMIV) (-Cont’d)

- Aggregator
  - Accumulation for each intra period
  - Implemented by logical operation OR

Patch Packer

- 8-pixel neighborhood region growing is conducted to generate a cluster
- Then, the clusters are sorted by a decreasing size order
- The number of atlases is determined by:

\[ N_{\text{Atlas}} = \lceil \frac{M\text{Pixel} \times 1024^2}{\text{AtlasWidth} \times \text{AtlasHeight}} \rceil \]

where maximum size of all atlases is expressed in Mpixels (default value: 20)

Example of clusters represented in false color on a pruned view

Example of 8-pixel neighborhood region growing
Test Model for Immersive Video (TMIV) (-Cont’d)

- Decoder
  - Decoder receives HEVC decoded atlases and the following metadata
  - TMIV decoder consists of three main steps:
    1. Video decoder and metadata parser
    2. Atlas patch occupancy map generator
    3. Renderer

Architecture for TMIV decoder[1]

Test Model for Immersive Video (TMIV) (-Cont’d)

- **Atlas Patch Occupancy Map Generator**
  - An occupancy map is generated for each atlas
  - Occupancy map contains the locations of the patches
  - Based on the occupancy map, the source views are reconstructed

Data flow of TMIV decoder[1]

Experimental Results

- HEVC anchor, TMIV anchor, and MCSL’s method are evaluated
- MCSL’s method represents TMIV-like encoder and decoder
- BD-rate, required number of decoders, encoding time are measured

Block diagram for MCSL 3DoF+ S/W platform[1]

Experimental Results (-Cont’d)

- Server
  - 2 linux servers (ubuntu 16.04) were used in experiment
  - One has 2 intel xeon E5-2687w (12 cores, 24 threads) CPUs and 128GB memory
  - Another has 2 intel xeon E5-2620 (6 cores, 12 threads) CPUs and 128GB memory

- Software
  - Versions of softwares meet the requirements of CfP
  - OpenCV 3.4.2 was used in source view pruning, reconstruction, and RVS
  - OpenGL was not used because of the building issue in linux
Experimental Results (-Cont’d)

- Experimental results for *ClassroomVideo* and *TechnicolorMuseum*
  - For the objective evaluation, WS-PSNR[1] is used
  - TMIV showed the best results on BD-rate and encoding time saving
  - MCSL’s method requires the least number of decoders among the introduced methods

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Method</th>
<th>BD-rate</th>
<th>No.of decoders required</th>
<th>Encoding time saving</th>
</tr>
</thead>
<tbody>
<tr>
<td>ClassroomVideo</td>
<td>HEVC-anchor</td>
<td>0.00%</td>
<td>30</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>TMIV-anchor</td>
<td>-80.59%</td>
<td>4</td>
<td>-86.28%</td>
</tr>
<tr>
<td></td>
<td>MCSL</td>
<td>-37.14%</td>
<td>4</td>
<td>-81.68%</td>
</tr>
<tr>
<td>TechnicolorMuseum</td>
<td>HEVC-anchor</td>
<td>0.00%</td>
<td>48</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>TMIV-anchor</td>
<td>-68.26%</td>
<td>8</td>
<td>-86.72%</td>
</tr>
<tr>
<td></td>
<td>MCSL</td>
<td>-18.90%</td>
<td>4</td>
<td>-95.21%</td>
</tr>
</tbody>
</table>

Results of HEVC anchor, TMIV anchor, and MCSL’s method

Experimental Results (-Cont’d)

RD-curve of *ClassroomVideo*

**ClassroomVideo**

- Anchor (HEVC)
- Anchor (TMIV)
- MCSL

**WS-PSNR (dB)**

- 35
- 34
- 33
- 32

**Bitrate (Mbps)**

- 6.5
- 10
- 15
- 25
- 40
- 65
Experimental Results (-Cont’d)

RD-curve of TechnicolorMuseum

- Anchor (HEVC)
- Anchor (TMIV)
- MCSL
  - R1
  - R2
  - R3
  - R4
Appendix : Atlas generated by TMIV

- Class A1(ClassroomVideo), atlas 0 (basic view), texture

Resolution = 4096x2048
Appendix : Atlas generated by TMIV

- Class A1(ClassroomVideo), atlas 0 (basic view), depth

Resolution = 4096x2048
Appendix : Atlas generated by TMIV

- Class A1(ClassroomVideo), atlas 1 (additional view), texture

Resolution = 4096x2048
Appendix : Atlas generated by TMIV

- Class A1(ClassroomVideo), atlas 1 (additional view), depth

Resolution = 4096x2048
Appendix : Packed View generated by MCSL

- Class A1(ClassroomVideo), texture