Immersive Video Activities in MPEG-I: Current Status and Upcoming Challenges

Joel Jung

Orange Labs

WORKSHOP: Computational Imaging with Novel Image Modalities

May 27-28 2019, INRIA, RENNES, FRANCE

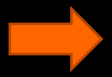
Introduction – 2D video

2D video coding in on track:

- Capture is easy, content is available.
- Use cases are known, no surprise.
- Improvements are incremental, predictable and «easy» (in the sense: depends on allowed CPU ressources).
- The only cloud in the horizon is the multiplicity of standards and the royalty policies.

Problem of 2D video:

- Increasing resolutions, frame rate, dynamic range only slightly improves the quality of experience.
- Current 2D is a very basic and preliminary way to represent the reality.
- It does not make dream.



But, improvements in 2D video coding are still needed because transmission of 2D video is a challenge in many situations.

In MPEG-I, most of the ressources are still spent in 2D coding

Introduction – Immersive video

Immersive video is a different story:

- Capture is challenging, can be done by differents means: no single representation format
- Natural and realistic content is missing
- Displays are not yet mature for mass market deployment, progress is promising
- Use cases are multiple, but fragmented in multiple area
- Improvements are fast and huge, but impact very different aspects of the immersive framework

Immersive video basic goal is similar as 2D video goal: representing the world
Immersive video is the continuity of 2D video.
The only difference is the level of expectation: immersion and quality of experience are requested

Immersive video is in a pre-pre-pre-preliminary status!

- It makes dizzy because not technically ready.
- When digging in the technical solutions, feeling of vertigo is possible for those in a hurry or having short term vision.

Immersive video raises multiple technical challenges: perfect time to work on it!

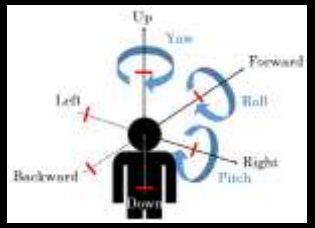
What is the goal and what is currently addressed in MPEG-I Visual?

Goal of immersive video studies

Towards a better representation of the reality for a better immersion with 6 DoF

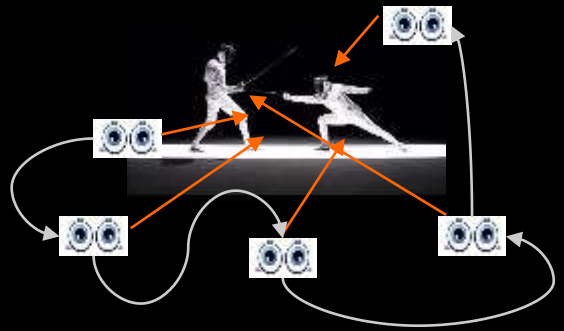
What is 6DoF (6 degrees of freedom)?

- Ability to move in the 6 directions (3 rotations + 3 translations)
- No notion of distance



Target in MPEG-I visual:

- Provide the correct pixel corresponding to the exact motion of the user (which is with 6DoF)
- Does not necessarily mean moving everywhere in the scene and navigate behind objects (this is capture dependant)
- Natural content, not only CG, full video scenes (sport events for instance)
- High quality of experience / rendering quality
- Reasonnable pixel rate, and bit-rate



MPEG-I Visual:

"how to design light-field compression and synthesis to achieve 6DoF immersion/experience?"

Outline

Introduction

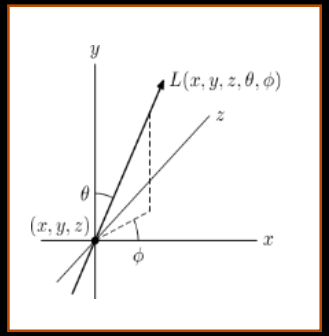
Capture of the light-field
- Highlight on some mutli-camera capturing devices

Current status of 6oF immersive video in MPEG-I Visual
- Depth estimation and synthesis
- Compression

Weaknesses of the current MPEG-I Visual approaches and perspectives

Conclusion

Light-field and rendering



General definition of the light-field:

“light flowing in every direction through each point of space”

Observation:

Assuming the light-field is available, immersive video becomes mostly a compression problem:

- All pixels are available, for any point of view, from any point of view.
- No more motion sickness, no shift between what is displayed and the expectation of the brain.

Immersive video is all about displaying the light-field!

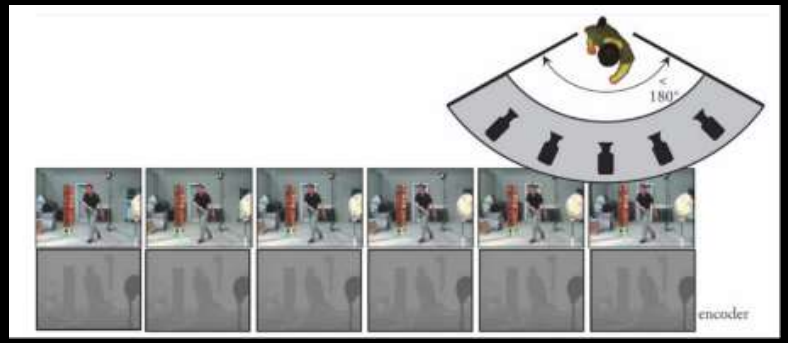
So why don’t we just display it?

Capturing devices and formats

- Lenslet image (or integral image, or plenoptic image)
- It is captured by lenslet (or plenoptic) cameras

- Point cloud image
- It is captured by traditional cameras and depth sensors

Capturing devices and formats



- Mulii-view + depth video
- It is captured by arrays of traditional cameras and depth is computed

- It is an omni-directional video (or 360 video)
- It is captured by omni-directional cameras arrays

Convergent camera arrays in 2019

Experimental mobile setup

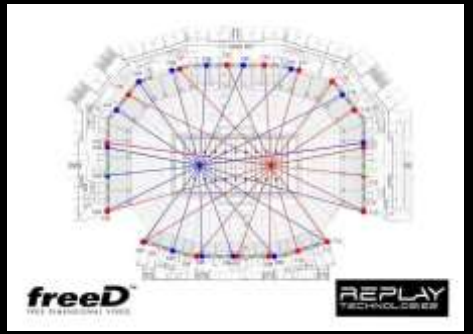
Stanford array of 128 cameras

Technicolor array of 16 cameras

Fraunhoffer array of 9 cameras

2017: Intel True View cameras:

- 360-degree replays and player perspectives for the NFL season
- 11 NFL stadiums equipped
- 38 ultrahigh-definition cameras (5K)
- Intel Core i7 servers that process 1 terabyte of data per 15- to 30-second clip.



Camera layout

Divergent omni-directional cameras in 2019

GoPro Fusion
5.2K@30fps

Samsung Gear
360
4K@24fps

Garmin Virb
5.7K@30fps

Ricoh Theta V
4K@30fps

Insta 360 pro
4K@100fps
8K@30fps

GoPro Omni
8K@25fps

Kandao Obsidian
8K@30fps
4K @60fps

MPEG-I Visual activity

All these formats are representations of sub-sampled light-field, captured by different means.

The light-field cannot be captured.

The problem of immersive video is not a compression only problem.

Need to render a light-field as dense as possible, without occlusions/holes under constraint of sparse capture, and transmission/compression constraints.

This involves view synthesis, which involves (often) depth estimation.

MPEG-I Visual addresses compression and synthesis (and depth estimation) for omni-directional and convergent capture and format

Outline

Introduction

Capture of the light-field
- Highlight on some mutli-camera capturing devices

Current status of 6oF immersive video in MPEG-I Visual
- Depth estimation and synthesis
- Compression

Weaknesses of the current MPEG-I Visual approaches and perspectives

Conclusion

MPEG-I Visual current status: depth estimation

- FTV (Free Viewpoint TV) group as been addressing 3D for a while, progressing at a very slow pace
- MPEG-I Visual is the successor of FTV group
- The activity has been significantly boosted with the arrival of new actors:
  - 2016: Technicolor, Univ. of Brussels, Orange Labs
  - 2017: Philips, Intel, Nokia

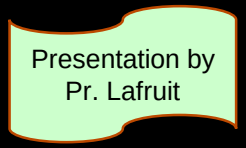
Status on depth estimation:

- Little progress so far, actors are focussed on compression and synthesis
- Several non-MPEG tool look much more promising than current DERS (Depth Estimation Reference Software). In practice, it is a bit more balanced when:
  - Depth maps quality is assessed according to the quality of the synthesis
  - Realistic content is considered
- The group has defined test conditions to evaluate tools or improvements properly, based on view synthesis

MPEG-I Visual current status: view synthesis

In 2017, VSRS (View Synthesis Reference Software), designed by FTV group is outperformed by:

- RVS (University of Brussels and Philips)



- VVS (Orange Labs)
  - Selection of information safe to warp
  - Depth warping and conditional depth merging
  - Texture backward warping, texture and depth merging
  - Temporal inpainting and filtering

  Designed to specifically handle compressed texture and depth maps with typical artifacts
  Designed to optimize the subjective quality

Omnidirectional test material

Philips content

Technicolor content

Set of several 360 cameras

Perspective test material

1D 10x1

PoznanFencing

2D 5x5

ULBUnicorn

2D 5x4

ETRIChef

Array 5x5

OrangeShaman

Array 5x5

OrangeKitchen

Arc array (14x3)

OrangeDancing

Perspective test material

2D 4x4

TechnicolorPainter

1D 14x1

IntelKermit

1D and 2D multiview content
All test material is provided with depth maps (computed)

MPEG-I Visual current status: view synthesis

Under MPEG-I Visual Test Conditions:

- VVS2.0 is 1.6dB (163% Bd-rate) better than VSRS4.2
- VV2.0 is 0.5dB (53% BD-rate) better than RVS3.1 (and 50% faster)
- VVS2.0 and RVS3.1 are more complex than VSRS4.2

Sequence	synth PSNR / total bitrate	synth PSNR diff
TechnicolorPainter	84.1%	1.32
ULBUnicornA	315.6%	2.17
ULBUnicornB	122.9%	1.29
OrangeShaman	185.1%	1.01
OrangeKitchen	366.2%	2.26
OrangeDancing	253.1%	1.45
ETRIChef	147.5%	0.66
IntelKermit	0.0%	3.36
PoznanFencing	0.0%	1.05
average	**163.8%**	**1.62**

VVS versus VSRS

Sequence	synth PSNR / total bitrate	synth PSNR diff
TechnicolorPainter	1.6%	0.03
ULBUnicornA	104.6%	1.42
ULBUnicornB	72.2%	0.88
OrangeShaman	-11.9%	-0.15
OrangeKitchen	37.2%	0.56
OrangeDancing	30.7%	0.34
ETRIChef	24.9%	0.16
IntelKermit	84.8%	0.52
PoznanFencing	139.8%	0.46
average	**53.8%**	**0.47**

VVS versus RVS

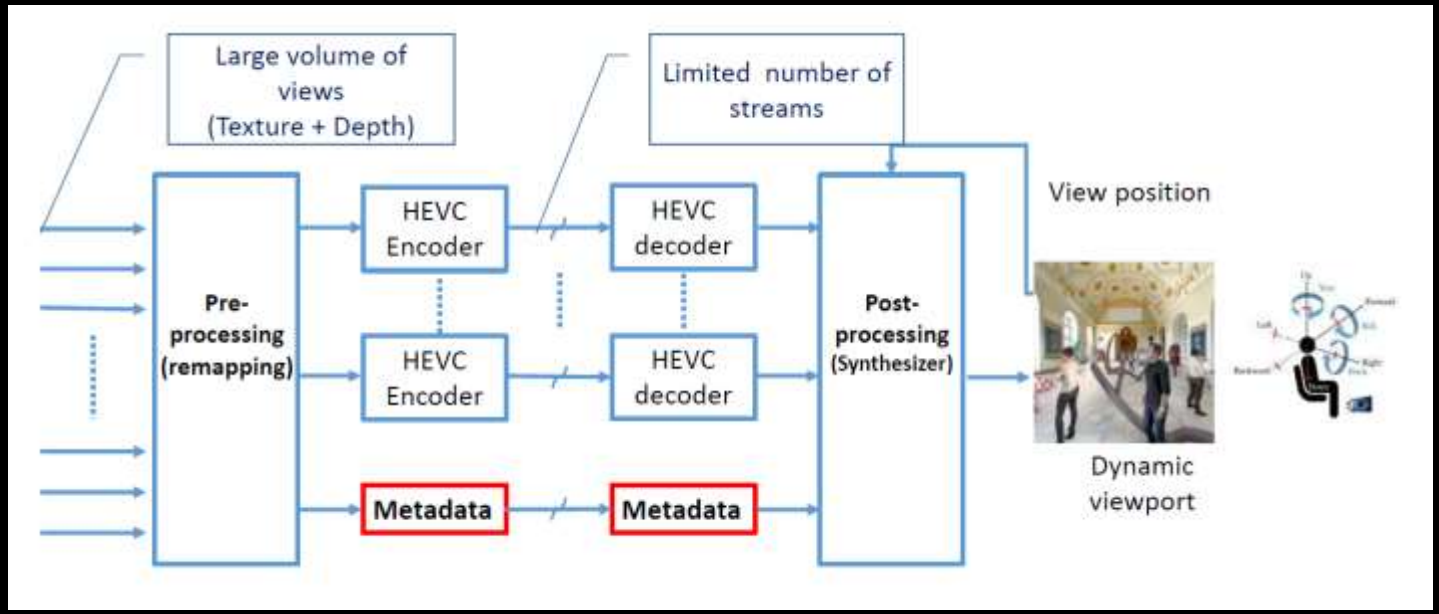
Cumulative depth estimation and view synthesis improvements: about 3dB in 18 months.

MPEG-I Visual current status: compression

- 2 anchors for the current 6DoF activity
  - MIV: Metadata for Immersive Video (draft international standard on Jan. 2020)
    - For some reasons called 3DoF+. It is 6DoF in a limited volume: MIV is a first attempt towards video coding for 6DoF.
    - Call for Proposal released on January 2019, 5 CfP responses, interesting approaches
    - The codec remains HEVC with some metadata
  - MV-HEVC + VVS
    - High level syntax changed compared to HEVC: one additional frame is used for motion estimation (inter-view reference)

It is not known yet which one is the most efficient

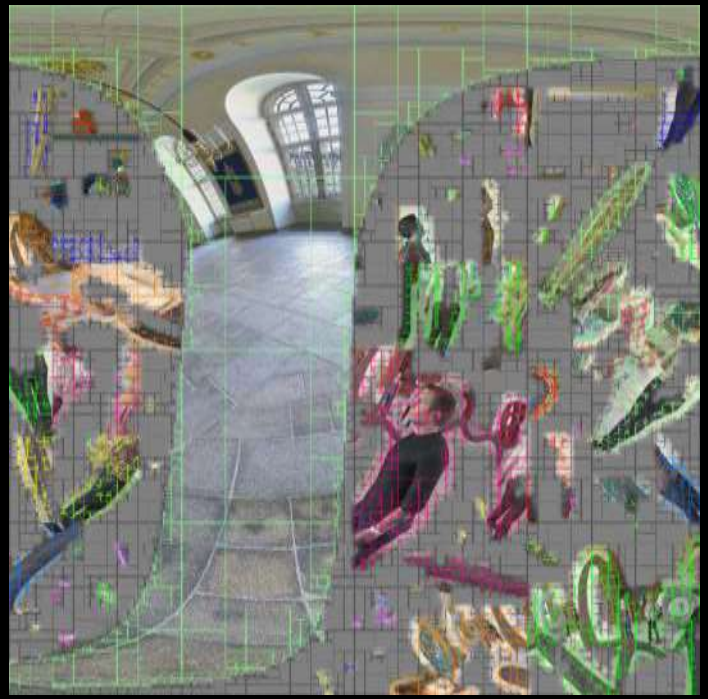
MPEG-I Visual current status: MIV call for proposal



General principle

Philips response to the CfP

- Iteratively (hierarchically) prune the source views with view synthesis in the loop,
- Create a block tree representation of pruned views and packed views,
- Pack blocks in order of relevance,
- Transmit the block tree as metadata.



Example of packed view with
Block tree overlay

Nokia response to the CfP

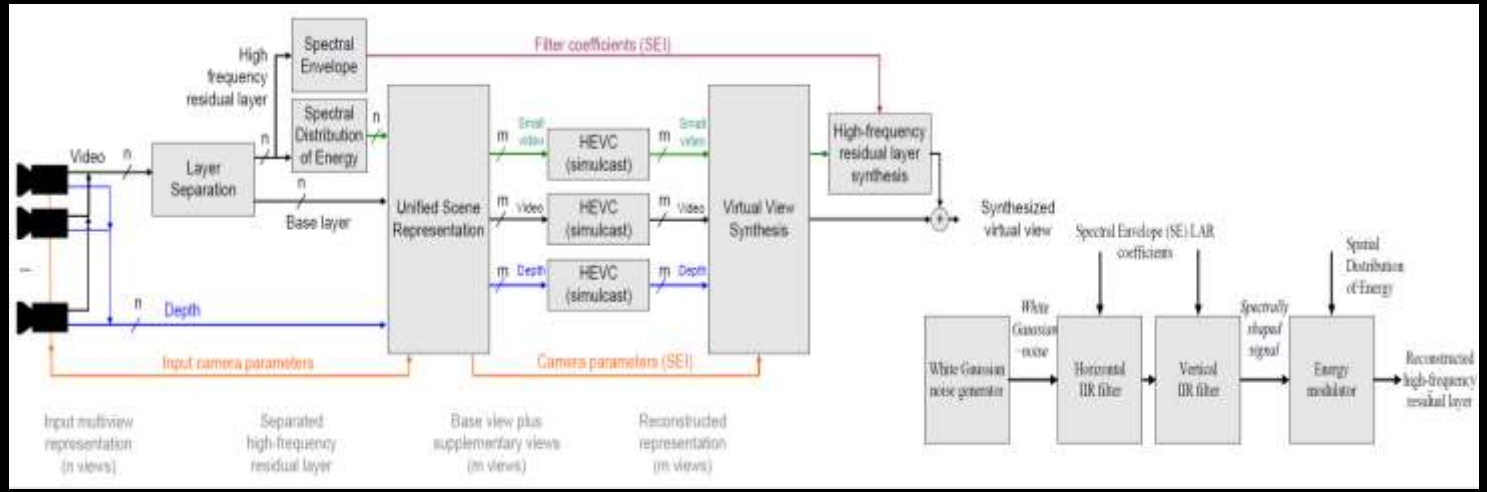
- The 3D volume is reconstructed as a point cloud, considering captured depth and colour images
- The 3D space points are projected to 2D planar surfaces (projection planes), different from capture planes
- Projected images (shards) are packed into two rectangular images (mosaics); one for colour and another for depth.
- Metadata that carries the information about each shard is sent. Mosaics are encoded with HEVC.

Shards

Point cloud rendering of Technicolor-Museum

Poznan University/ETRI response to the CfP

- Employ a different scene representation, called Unified Scene Representation (USR).
- USR obtained by view synthesis performed to create a different, optimized set of views.
- Uses a 2 layers approach, video is split into layers in the spatial frequency domain
- Both layers are transmitted to the respective decoders and after decoding are summed together in order to produce reconstructed video.

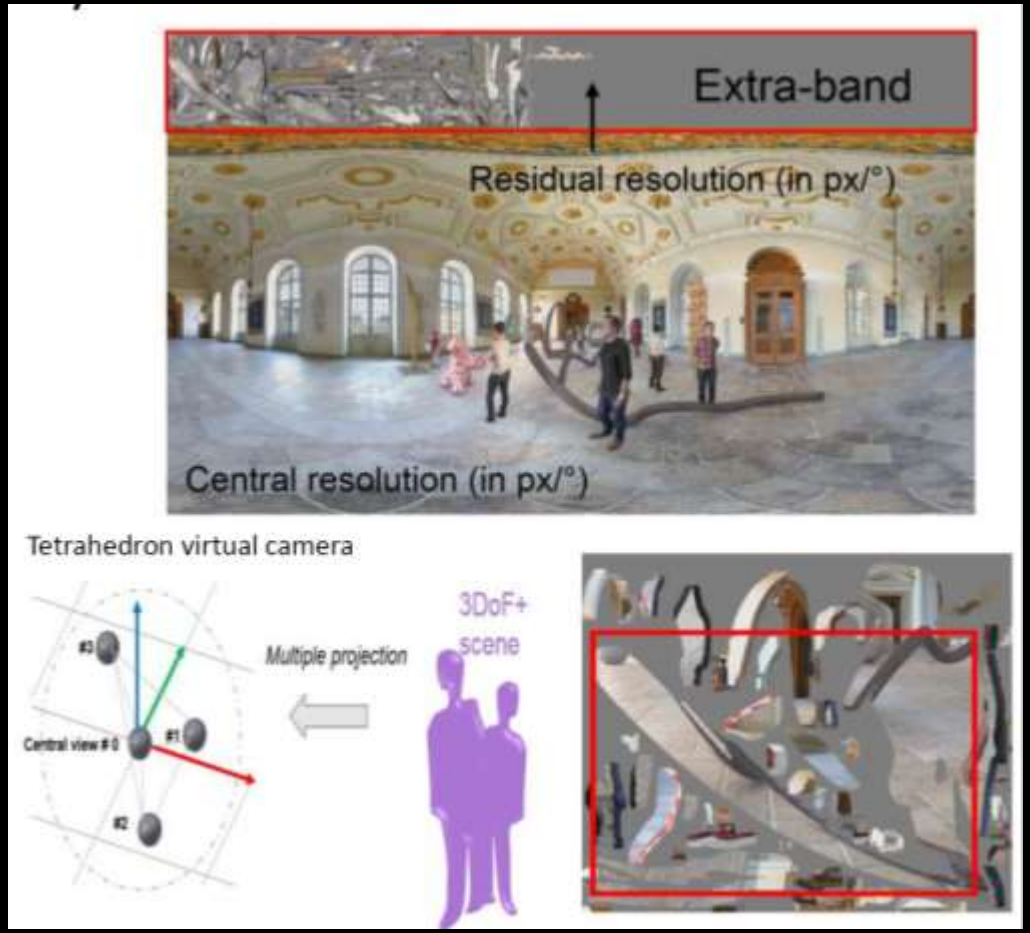


Technicolor/Intel response to the CfP

- The scene is re-projected on a number of virtual “transport views”
- The re-projection is based on a View Synthesizer which operates a re-sampling of the points located in 3D space.
- 4 transport cameras disposed in a tetrahedron shape

Principle:

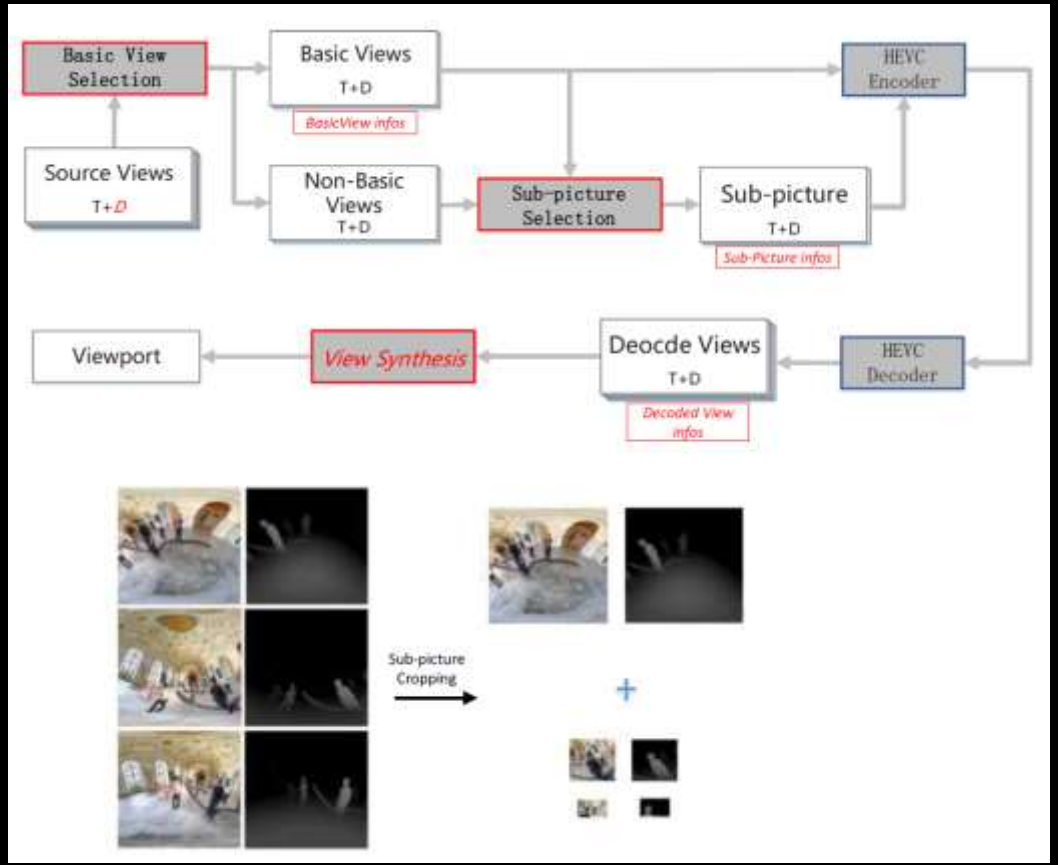
- Redundancy removal between the source views.
- One global view taking most part of visual material, typically more than 90%, and in a center position
- other views will convey the residual parts.



Zhejiang University response to the CfP

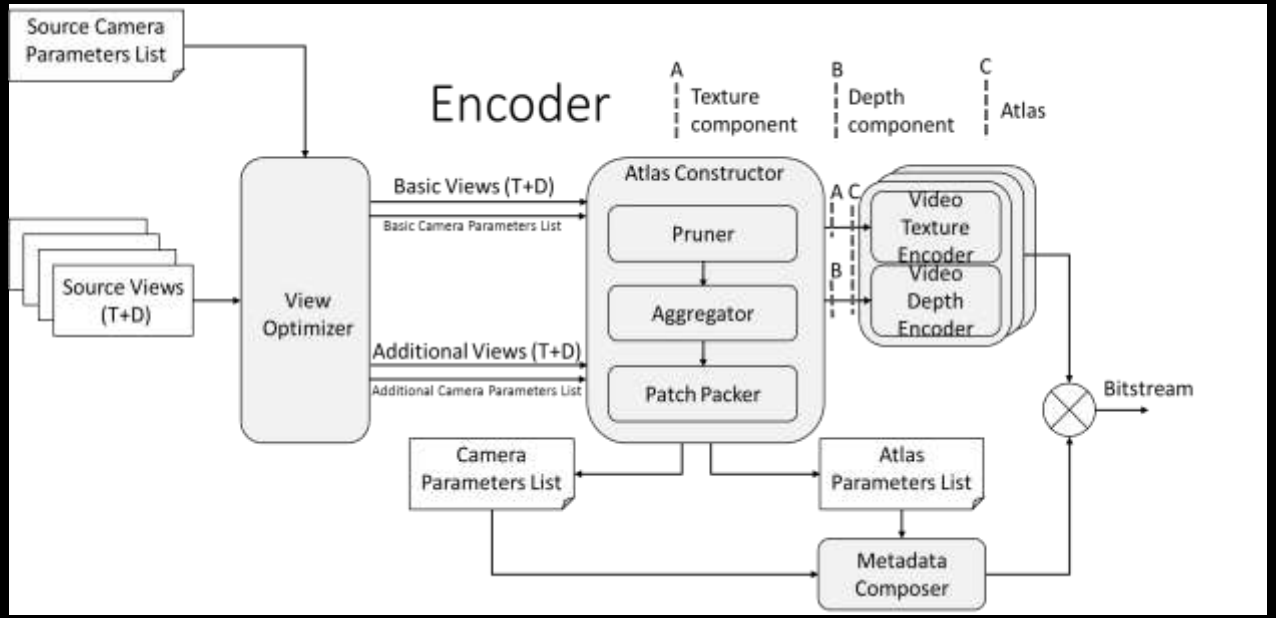
- Adaptive selection of basic views and complementary sub-picture(s).
- Extract sub-pictures, sent as metadata
- HEVC used for encoding all basic views and complementary sub-pictures.

- After decoding, all the basic views and sub-pictures and corresponding metadata information are used to synthesize views/viewports at source or intermediate positions.



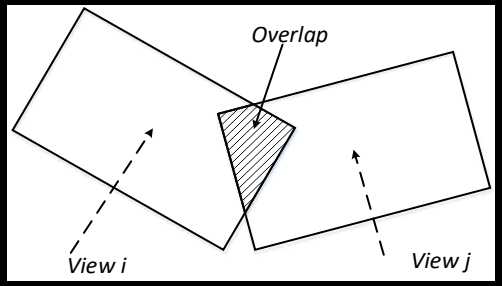
Current MIV standard – encoder

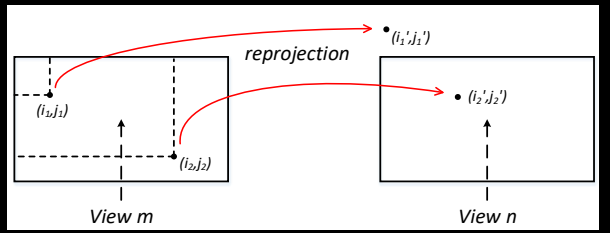
- When considering the different sequences, bit-rates, pixel rates, subjective and objective quality, none of the approaches was significantly better than all others
- A test model v0 is constructed, from the 5 responses (1st version released in May 2019)



N18470, Test Model for Immersive Video, Salahieh, Kroon, Jung, Domański Geneva, March 2019

Current MIV standard – view optimizer



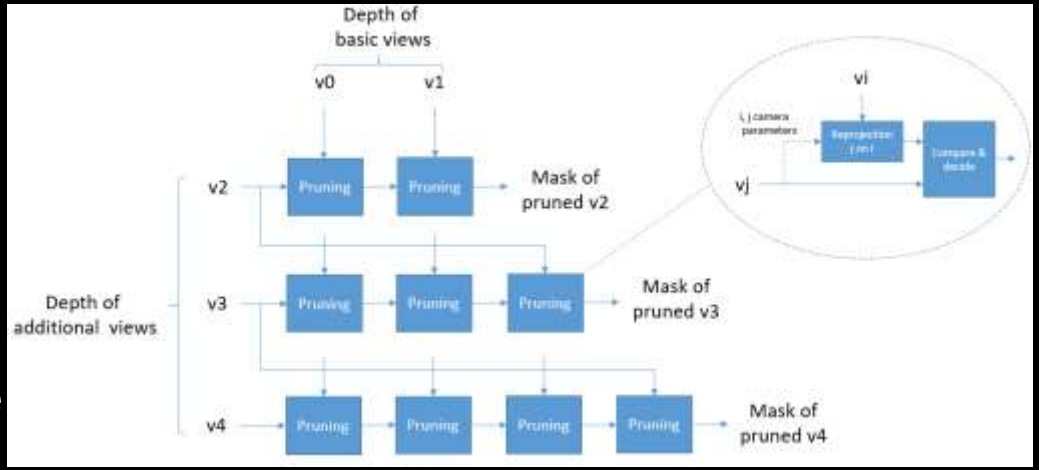


View optimizer: selects one or multiple views from the source views, called basic views

- Find a pair of views (view m, view n) that has the largest direction deviation, with largest sum of field of views, and largest distance between each other
- Decision based on the overlap obtained when projecting from one view to another

- If 1 basic view is needed
  - Basic view is the source view the closest to the central camera position of the source capturing system (using camera parameters list).
- If several basic views are needed,
  - The pair (m, n) is selected
  - A view k which has the largest direction deviation with view m and view n becomes basic if it has less than 50% overlap with m or n (repeated process)

Current MIV standard – atlas constructor (1/3)



Pruner:

Creates a mask that indicates the part of the additional view to be kept

- Validation of a pixel by re-projecting each depth pixel value of the additional view onto each basic view and validate or invalidate the pixel accordingly (ladder process)
- A sample of the additional view is pruned if it is already “covered” by a sample of a previous view (basic or additional)

Aggregator:

- The mask is reset at the beginning of each intra period.
- An accumulation is done for each mask’s pixel

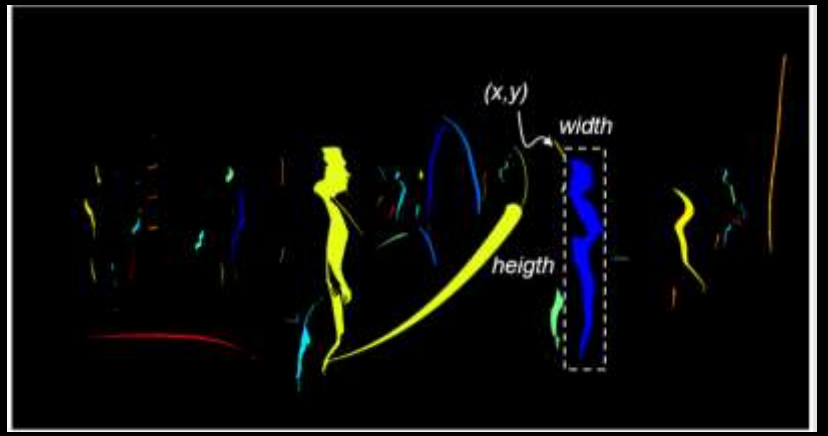
Example of aggregated mask

Current MIV standard – atlas constructor (2/3)

Patch packer

- Rectangle clusters are computed around patches
- Sequentially packing of each cluster into “atlases”

  (MaxRect algorithm: makes use of the existing “Used Space” first, by examining the space which is efficiently occupied “Filled space”



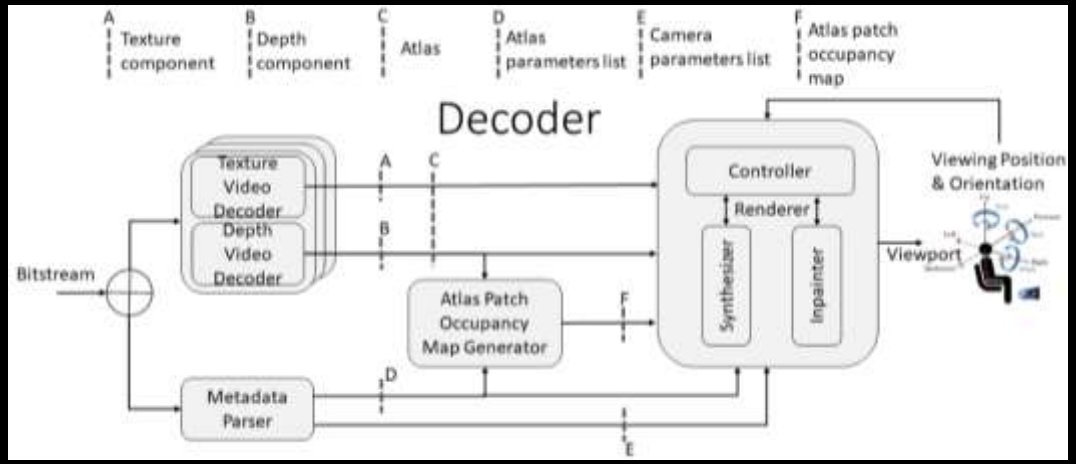
Patch packer provides: patch list for each atlas with all necessary information to recover at the decoder side:

- The location (x,y) in the atlas and the atlas id
- The location (x’,y’) in the original view representation, and its dimensions (w,h)
- The related CameraId, which itself refers to the de-projection parameters for that view in the decoder
- A possible rotation by 90°

Current MIV standard – atlas constructor (3/3)

Example of atlases:

Current MIV standard – decoder



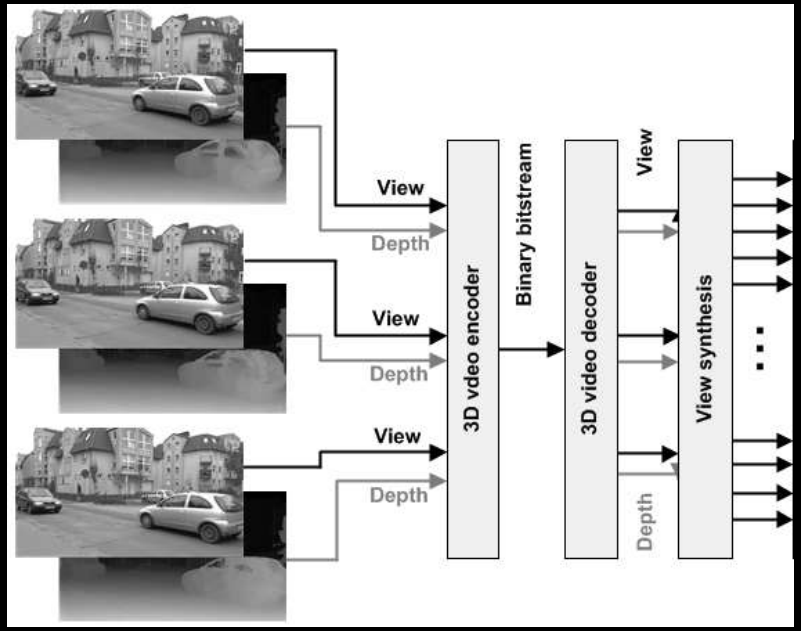
Includes:

- An HEVC decoder
- A metadata parser
- RVS4.0 to synthesize the views from the decoded atlases

  (includes an inpainting module from Poznan Univ. of Technology)

MPEG-I Visual current status: MV-HEVC + VVS

- Simple encoding of all views (texture + depths)
- Synthesis of intermediate viewpoints from decoded textures and depths



MPEG-I Visual 6DoF framework

Common test conditions are defined

No more 3DoF+ / 6DoF CTCS, those are nearly aligned towards immersive video CTCs

- 2 anchors: both supposed to be improved (TMIV / MV-HEVC+VVS)
- Test material: 12 sequences
  - 3 omni-directional, 9 perspective
  - 6 natural, 6 computer generated
- Defined list of views to be synthesized

To compare TMIV versus MV-HEVC+VVS
To evaluate possible improvements of TMIV or MV-HEVC+VVS
To evaluate any new coding+synthesis approach

Outline

Introduction

Capture of the light-field
- Highlight on some mutli-camera capturing devices

Current status of 6oF immersive video in MPEG-I Visual
- Depth estimation and synthesis
- Compression

Weaknesses of the current MPEG-I Visual approaches and perspectives

Conclusion

Weaknesses of the TMIV approach

TMIV limits might appear when:

- Using estimated depths (not computer generated)
- Considering more complex scenes: current omnidirectional content has limited motion, with moving objects on less than 1/3rd of the 360 sphere

Increasing the distance between cameras (so occlusions)

Or/and avec more motion inside a view

Will drastically increase the number of atlases to encode, making it difficult to meet the targeted pixel rate

Weaknesses of the MV-HEVC+VVS approach

The pixel rate is high

One additional frame to store in the frame buffer for motion compensation

Perspectives

A 6DoF codec has to be designed in a way to be not only compatible, but also friendly with the synthesis

MIV and PCC are good first examples (they share metadata needed by the synthesis)

Not clear today if the best basis is MIV based of MV-HEVC based, for natural content (low quality depths), with larger baseline

May be a mix of both is a solution:

- Improved patch extraction and representation
- Improve atlas coding (HEVC definitely not adapted for such content)
- Add some interview prediction
- Add some specific tools for coding the depth contents
- Improve the view basic selection (captured or synthesized)

Perspectives

May be a more generic framework is a solution:

- Think about the necessity of sending depth
- Think about the necessity of sharing information between the decoder and the synthesis algorithms
  (joint approach for decoding / “dept estimation” or “partial depth estimation / synthesis”, etc)
- Add new metadata for better collaboration between the codec and the synthesis

May be a nice opportunity for CNN based approaches:

- Some relevant approaches for depth estimation and view synthesis have recently emerged
- Are they robust enough to be considered in practical real life cases and various test conditions?

May be another way is to get rid of traditional DIBR approaches.

The framework for testing different approaches is set.

Outline

Introduction

Capture of the light-field
- Highlight on some mutli-camera capturing devices

Current status of 6oF immersive video in MPEG-I Visual
- Depth estimation and synthesis
- Compression

Weaknesses of the current MPEG-I Visual approaches and perspectives

Conclusion

Conclusion (1/3)

2D video is on track and mature.

Current 2D representation of the world is an extremely basic representation

Immersive video has just been initiated, with a lot of challenges and expectations

A new level of “representation of the world” is in preparation

It is challenging: no more only about increasing resolution, frame rate, or number of views

It is challenging: several formats exists, and content is difficult to capture

No killer use case is needed: multiple use cases will appear along with the technology

Achieving a perfect immersive representation of the reality is not a use case

It is a logical/normal evolution, so it is the natural ultimate goal

We have to provide algorithmic blocks that will progressively enable full immersion.

The MPEG-I visual goal is simple:

how to design light-field compression and synthesis to achieve 6DoF immersion/experience?”

Conclusion (2/3)

6DoF refers to a motion that includes 3 rotations and 3 translations (so any motion)
Light-field is the light flowing in every direction through each point of space: it cannot be captured

The captured sub-sampled light-field is encoded and used to recover the light-field
Compression is no more the only player: synthesis and consequently depth estimation are to be considered

MPEG-I visual has recently made some progress on depth estimation and significant progress on view synthesis (total 3dB in 18 months):
Still insufficient to provide correct quality of experience

MPEG I visual has recently made significant progress in coding, with 2 anchors:
TMIV, based on patch extraction, HEVC + metadata coding, and view synthesis
MV-HEVC+VVS, based on HEVC coding, with inter-view prediction, and view synthesis

Conclusion (3/3)

Both anchors have severe bottlenecks
Both anchors offers huge perspectives of improvement to answer the 6DoF challenge
Most likely an approach taking advantage of both schemes can be the most efficient

Common test conditions are ready to evaluate in a fair and practical way:
- Improvements of these 2 anchors
- Other approaches not considered so far (CNN based for instance, etc)
- More original approaches, with better frameworks (depth estimation, view synthesis, and codec better working in a more collaborative way)

MPEG has defined 2 codecs for 2 representations of the light-field (PCC / MIV)
- Close from a WD point of view, different from a reference SW and use case point of view.
- They target similar goal: high quality rendering of the immersive world

Searching for a joint approach when designing the phases 2 of these 2 tracks would make a lot of sense

Thank you for your attention