2021년 IT21
Global Conference
Human in SW, SW in Human

Session 2-5

Immersive Video 부호화 표준기술

김재곤 교수 (한국항공대학교)

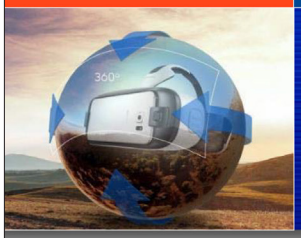
MPEG에서 표준화 중인 MIV(MPEG Immersive Video) 표준 기술과 표준화 현황을 소개한다. 먼저 몰입형 미디어 서비스를 제공하는 Immersive Video의 개념과 압축의 필요성 및 그 접근 방법을 살펴본다. MIV에서 3DoF+ 비디오 압축 표준기술을 개발하기 위하여 개발 중인 참조SW 코덱인 TMIV(Test Model for Immersive Video)의 뷰 최적화, 아틀라스 생성, 뷰 합성 등의 표준기술을 간략히 소개한다. 또한, 6DoF 비디오로의 확장을 위한 Future MIV 표준화 전망을 살펴본다.

약 력

1992 ~ 2007	한국전자통신연구원(ETRI) 팀장/선임연구원
2001 ~ 2002	Columbia University 방문연구원
2015	UC San Diego 방문교수
2009 ~ 현재	한국방송·미디어공학회 이사/논문지편집위원
2007 ~ 현재	한국항공대학교 항공전자정보공학부 교수

관심분야 비디오 부호화 표준, 비디오 신호처리, Immersive Video, 딥러닝

S2 Mixed Media



Immersive Video 부호화 표준기술

2021. 6. 10.

김 재 곤

한국항공대학교
KOREA AEROSPACE UNIVERSITY

Contents

I. Immersive Media

- 360 video/VR
  - Has become popular as a new media type that gives immersive experience
- Immersive media (video)
  - Immersion will be enabled by providing the viewer with the freedom to experience visual content with 6DoF of the movement of the user
    - Full parallax that is coherent to the movement of the user's viewing position and point of view, as well as to the object motion in the scene
  - Immersive Media is an umbrella term for VR, AR, mixed reality, and 360-degree video; where the physical world is emulated through a digitally simulated world

MPEG Activities on Immersive Media

- Standardization on immersive media
  - Provide enablers for production, coding, transmission, and consumption of immersive media and for new user experiences
- MPEG-I
  - *Coded Representation of Immersive Media* referred to as MPEG-I
  - Goal and key aspects
    - Develop the standard using coding technologies to transmit and storage immersive media over networks
    - Architectures, systems, video, audio, point clouds, as well as metrics, metadata and interfaces for network-based processing of immersive media content
  - The 116th Chengdu MPEG meeting, Oct. 2016

MPEG-I Overview

- ISO/IEC 23090 (MPEG-I) – 23 parts with 2 phases

Phase 1 (1a, 1b)	Part 1 : Immersive Media Architectures	TR
	Part 2 : Omnidirectional MediA Format (OMAF) • 1a OMAF for 360, 2017 • 1b Extension of OMAF for 3DoF+, 2020 (2nd Ed.)	FDIS, Oct. 2020
Phase 2 (2a, 2b)	Part 3 : Versatile Video Coding (VVC)	FDIS, July 2020
	Part 4 : Immersive Audio	Exploration
	Part 5 : Visual Volumetric Video-based Coding (V3C) and Video-based Point Cloud Compression (V-PCC) (2nd Ed.)	DIS, Jan. 2020
	Part 6 : Immersive Media Metrics	CD, Jan. 2021
	Part 7 : Metadata for Immersive Media (Systems)	FDIS, July 2021
	Part 8 : Network-based Media Processing (NBMP)	DIS, Jan. 2020
	Part 9 : Geometry-based Point Cloud Compression (G-PCC)	DIS, Apr. 2020
	Part 10: Carriage of Point Cloud Data	DIS, Apr. 2020
	Part 11: Implementation Guidelines for Network-based Media Processing	PDTR, Jan. 2020
	Part 12: Immersive Video	**FDIS, July 2021**
	Part 23: Conformance and Reference Software for MPEG Immersive Video	**CD, Oct. 2021**

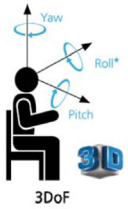
MPEG-I Overview

- Visual volumetric content
  - Point cloud
    - A set of individual 3D points associated with a number of attributes
  - Immersive video content
    - A real or virtual 3-D scene captured by multiple real or virtual cameras
- ISO/IEC 23090-5 – V3C and V-PCC
  - Visual volumetric content is coded using projection followed by 2D compression
  - 4 fundamental elements – atlas, geometry, occupancy map, attributes
- ISO/IEC 23090-12 – Immersive Video
  - MIV syntax and semantics are used with reference to V3C
    - Common syntax – occupancy, geometry/attribute, atlas data, etc.
    - Specific syntax (MIV extension)
      - Pruning graph, camera parameters, MIV view parameters list, etc.

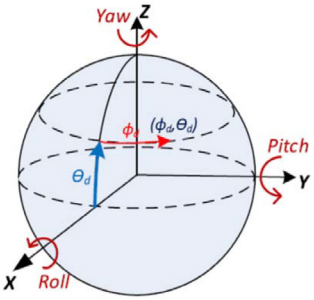
※ Visual Volumetric Video-based Coding (V3C)

Immersive Media

- Different levels of immersiveness
  - Defined w.r.t. an increasing degree of freedom in terms of movements of the observer within the immersive media scene
- 3DoF – 360 video
  - Allowing the user to look around in all directional from a fixed viewpoint
  - Three rotational and un-limited movements around the X, Y and Z
    - Rolling – tilting side to side on X-axis
    - Pitching – tilting forward/ backward on Y
    - Yawing – turning left/right on Z-axis
    - Yaw and pitch – the shit of a point on the sphere by azimuth ϕ_d and elevation θ_d

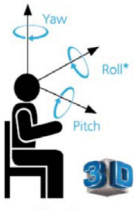


3DoF

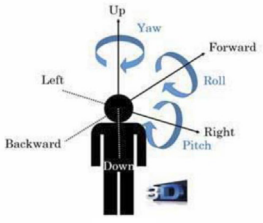


Level of Immersiveness

- 6DoF
  - Provide immersive experiences with free navigation in 6DoF in wider volume (e.g., walking zone)
  - 3DoF with additional relocation of the viewpoint, by translational movements of the user around the original viewpoint
    - User can move up/down, left/right and forward/backward
  - Adding *motion parallax to 360 video*
    - Where the relative positions of objects move, based on viewer motion
      - Changes in both orientation (yaw, pitch, roll) and spatial position

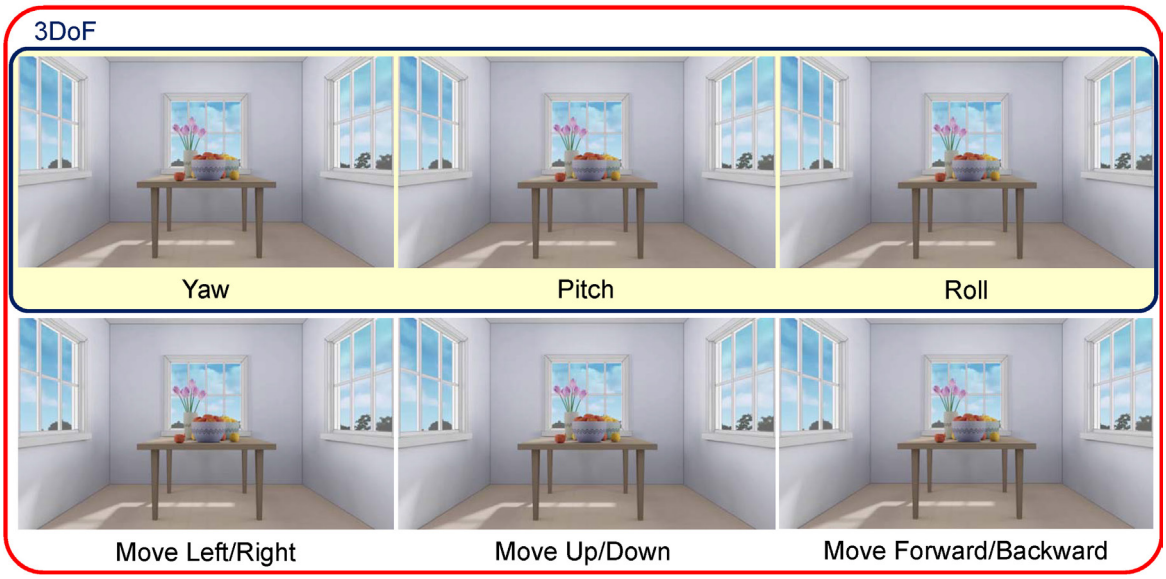


3DoF



6DoF

Level of Immersivness – DoF



http://blog.lytro.com/understanding-the-six-degrees-of-freedom/

II. Overview of Immersive Video Coding

- Overview of Immersive Video Coding
  - Standardization progress
  - Immersive Video – Test Sequences
  - Overview of Immersive Video Compression

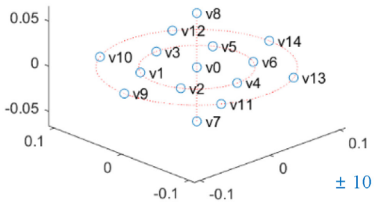
1. MIV Overview

- MPEG Immersive Video (MIV) – Immersive video compression
  - Evaluation of CfP responses, Mar. 2019
  - Current status
    - TM for immersive video (TMIV8)
    - DIS of immersive video (MPEG-I Part 12)

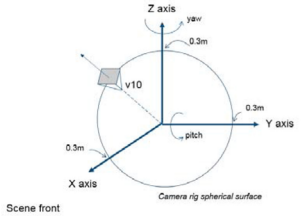
- Test sequences
  - Contain texture/depth for a scene simultaneously captured from many different camera positions
    - Real or virtual cameras
  - Along with metadata describing the camera positions
  - Each camera either capture 360 video (in ERP) or 2D video

2. Test Sequences for CTC

- Four computer graphic sequences



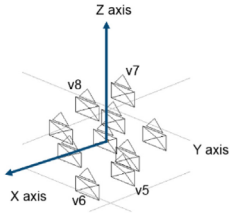
± 10 cm



radius 30 cm

Name	Projection	Input resolution	FoV	# of views	Frame count
Classroom video	ERP	4096 × 2048	360° × 180°	15	120
TechnicolorMuesum	ERP	2048 × 2048	180° × 180°	24	300

Test Sequences



± 30cm

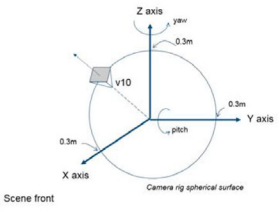
V0	V1	V2	V3	V4
V5	V6	V7	V8	V9
V10	V11	V12	V13	V14
V15	V16	V17	V18	V19
V20	V21	V22	V23	V24

80 × 80 cm (20 cm)

Name	Projection	Input resolution	FoV	# of views	Frame count
InterdigitalHijack	ERP	4096 × 2048	180° × 90°	10	300
OrangeKitchen	Rectilinear	1920 × 1080	53.1° × 31.4°	25	97

Test Sequences

- TechnicolorMuseum

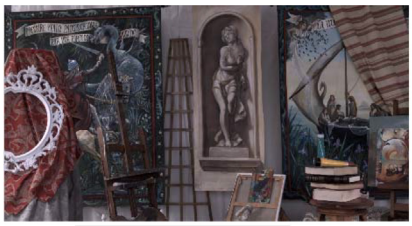


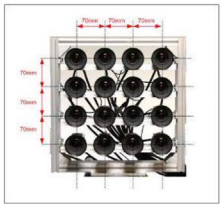
Test Sequences

- TechnicolorMuseum

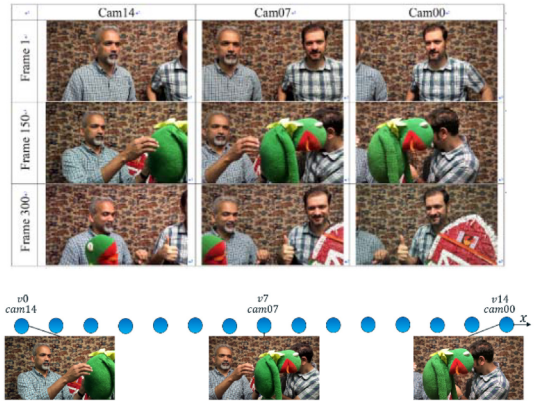
Test Sequences

- Three natural sequences





21 × 21 cm (7 cm)

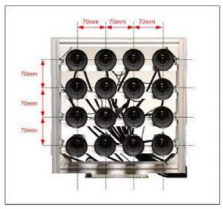


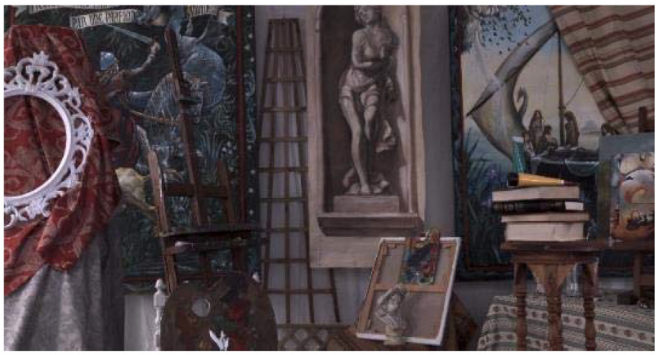
44.1 cm (3.675 cm)

Name	Projection	Input resolution	FoV	# of views	Frame count
TechnicolorPainter	Perspective	2048 × 1088	46º × 25º	16	300
IntelFrog	Perspective	1920 × 1080	63.6° × 38.5°	13	300

Test Sequences

- TechnicolorPainter





Test Sequences

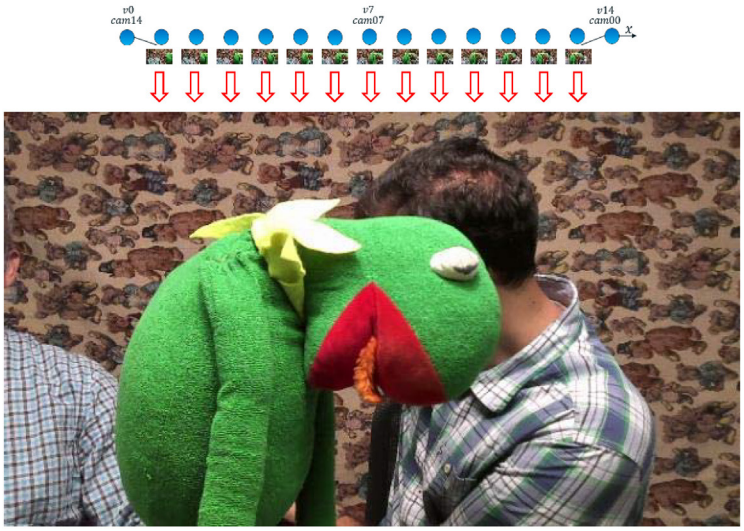
- TechnicolorPainter

Original view from Camera Position

View synthesis result from Original Camera Position

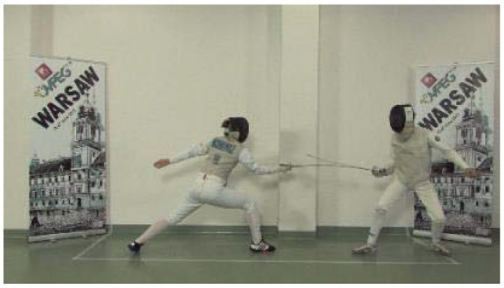
Test Sequences

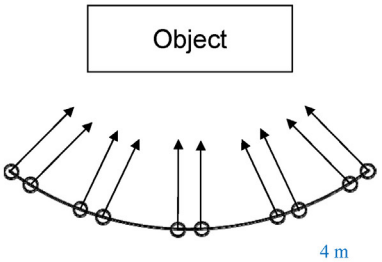
- IntelFrog



Test Sequences

- PoznanFencing
  - 5 stereopairs placed on arc (1m gap between them)
  - Angle between neighboring stereopairs: 15 degrees

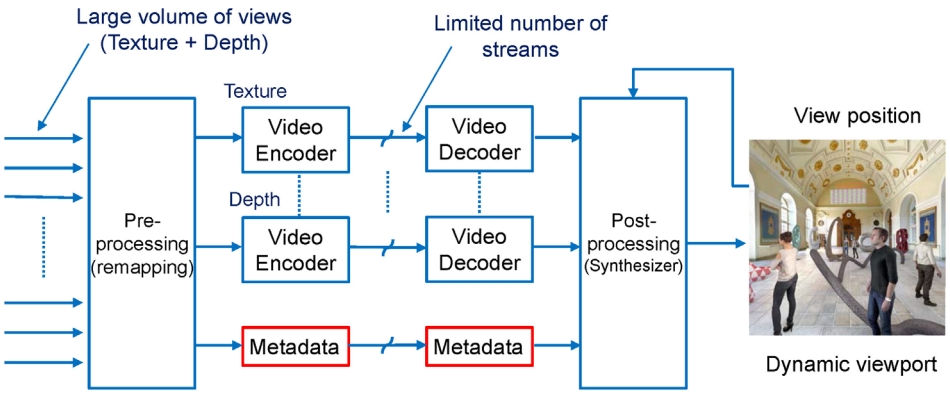




Name	Projection	Input resolution	FoV	# of views	Frame count
PoznanFencing	Perspective	1920 × 1080	63° × 48°	10	250

3. Overview of Immersive Video Compression

- Overall compression architecture



Concept of Immersive Video Compression

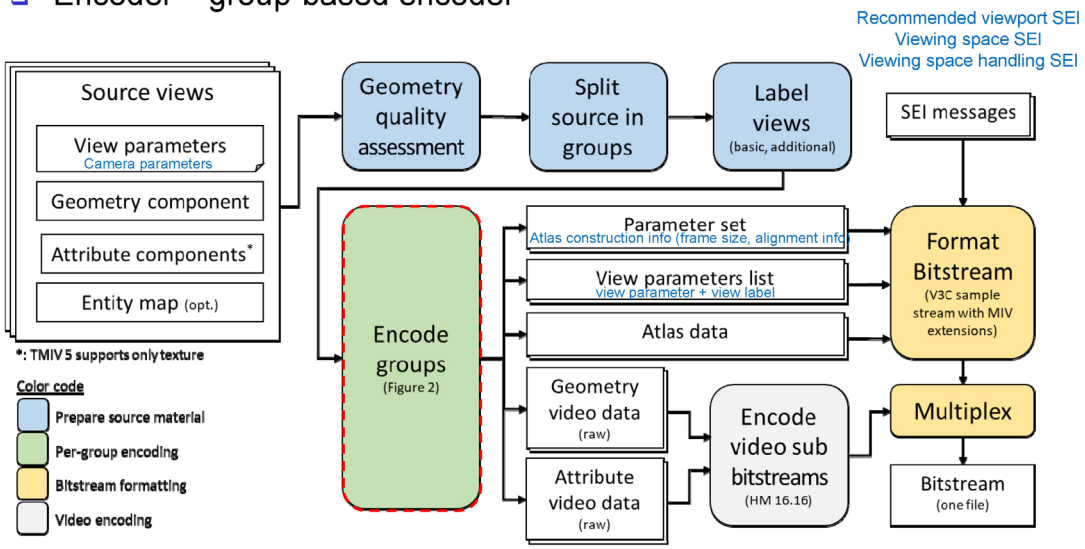
- Immersive video sequences have multiple views and their depth value
  - Very large volume
    - E.g., TechnicolorMuseum
      - 2048 x 2048 x 2 (texture & depth) x 24 views x 30 frames x 10 bits x 3 (YCbCr) = 16.8 Gbits/sec
- Reducing the inter-view redundancy
  - Remove these redundancy at each view (pruning), and just encode few basic views + residual parts
    - Residual parts (patches) of all views are packed in one frame (atlas)
    - Pruned value can be restored from basic views by synthesis
      - Need metadata for where each patch come from
  - Pixel rate reduction
    - 4096 x 2048 x 2 (texture & depth) x 2 (basic & packed image) x 30 frames x 3 (YCbCr) x 10 bits + additional metadata
    - Almost 1/6 of original volume

III. Test Model for Immersive Video

- Test Model for Immersive Video (TMIV 8)
  - Architecture and algorithms
    - Encoder
    - Decoder
  - Experiment results

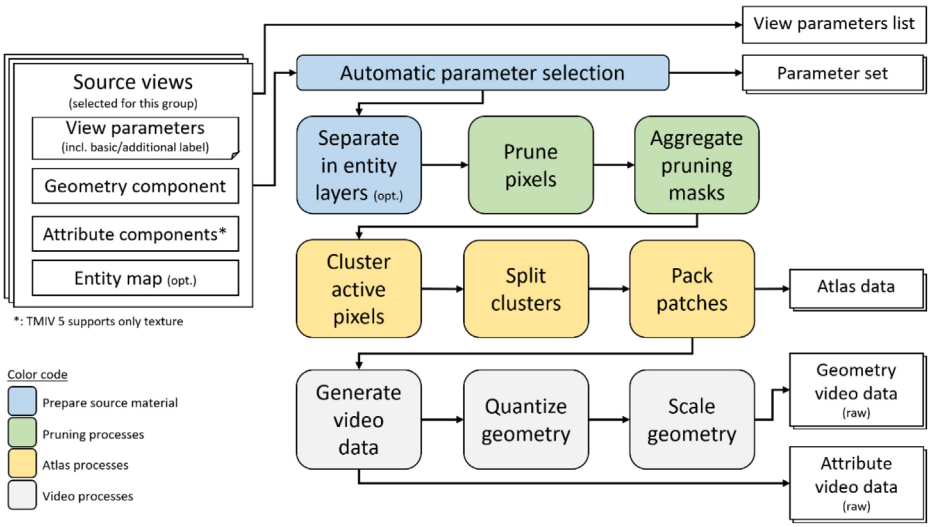
1. Architecture – Encoder

- Encoder – group-based encoder



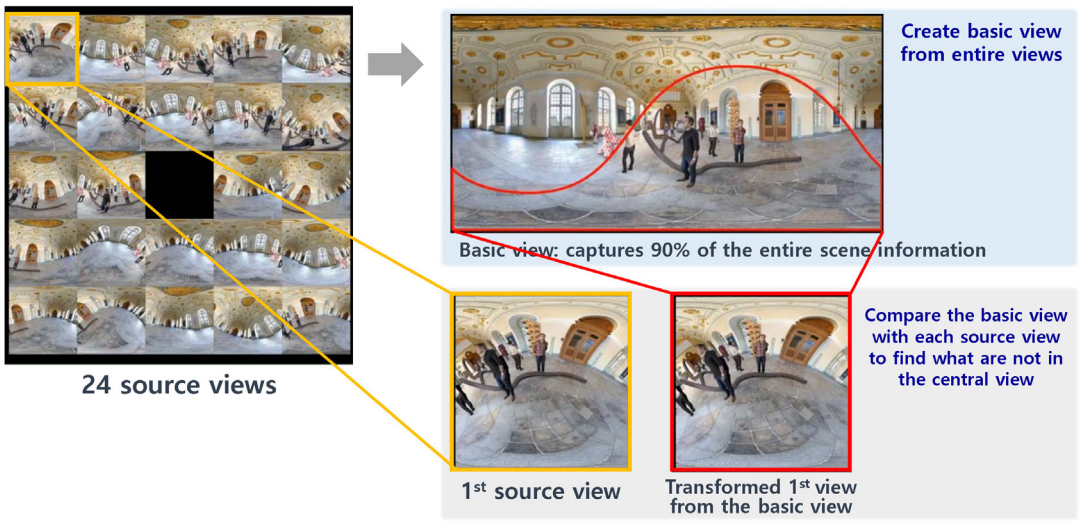
Architecture – Encoder

- Single-group encoder



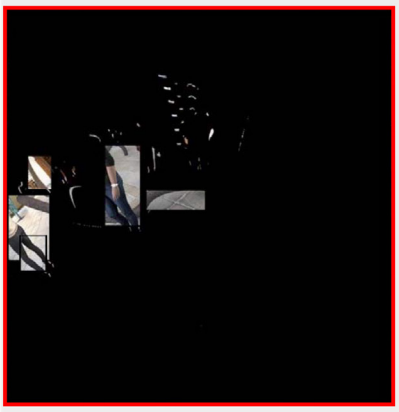
Overview of Atlas Generation

- Pruning Step & Packing Step



Overview of Atlas Generation

- Compare the basic view with each source view to find what are not in the central view



Transformed 1st view from the central view

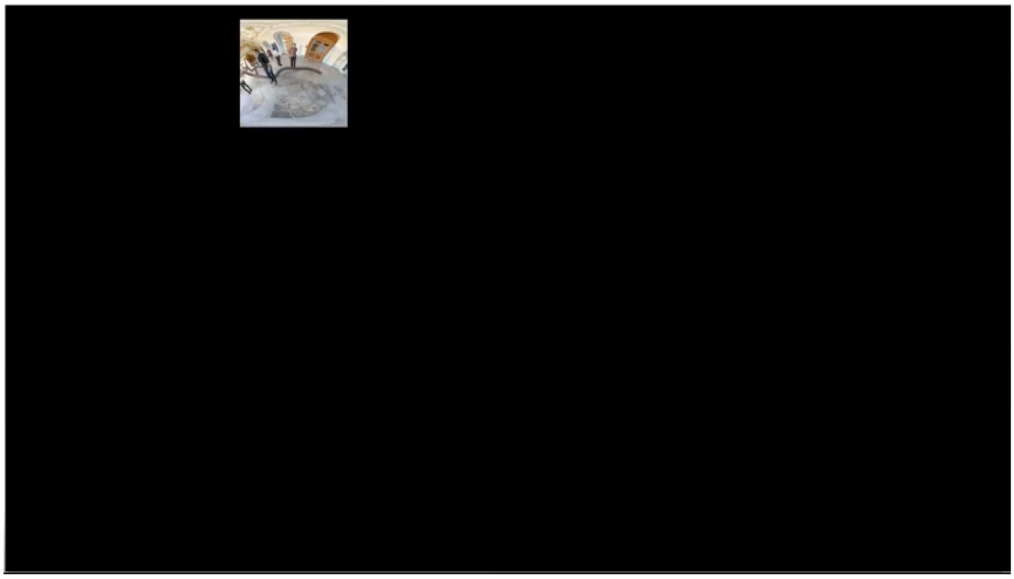
(red: area that is not in the central view)

Patches

Texture atlas

Overview of Atlas Generation

- For Museum sequence



Encoder Processes

- Preparation of source views – grouping
  - Geometry (depth map) quality assessment
  - Splitting source views in groups
  - View labeling – basic/additional views
- Encoding of each group separately – atlas
  - Automatic parameter selection
  - Pixel pruning
  - Atlas construction – clustering and packing
  - Video process
    - Geometry quantization & down-scaling
- Formatting of bitstream – metadata
  - V3C sample stream with MIV extensions and SEI
- VVC or HEVC encoding of video sub bitstreams – texture/depth
- Multiplexing into a single MIV-compliant bitstream

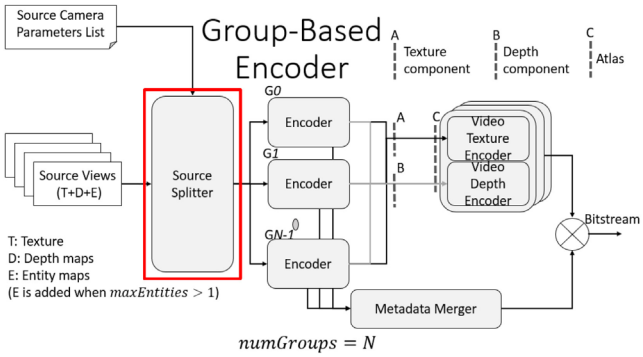
1) Preparation of Source Views

- Geometry (depth map) quality assessment
  - Signaling the quality of the geometry
    - casme_depth_low_quality_flag
  - An example of usage
    - An indicator to distinguish between natural and synthetic sequence in a patch dilation
  - Assessed automatically based on the first frame of the geometry
    - Each input view is re-projected to the position of all remaining views
    - Check if the re-projected pixel value is consistent to the collocated pixel or any of its neighbors (3x3)
      - If the number of inconsistent pixels between any pair of input views is higher than a threshold – the quality is supposed to be low

※ CASME – CASPS (Common Atlas Sequence Parameter Set) MIV extension

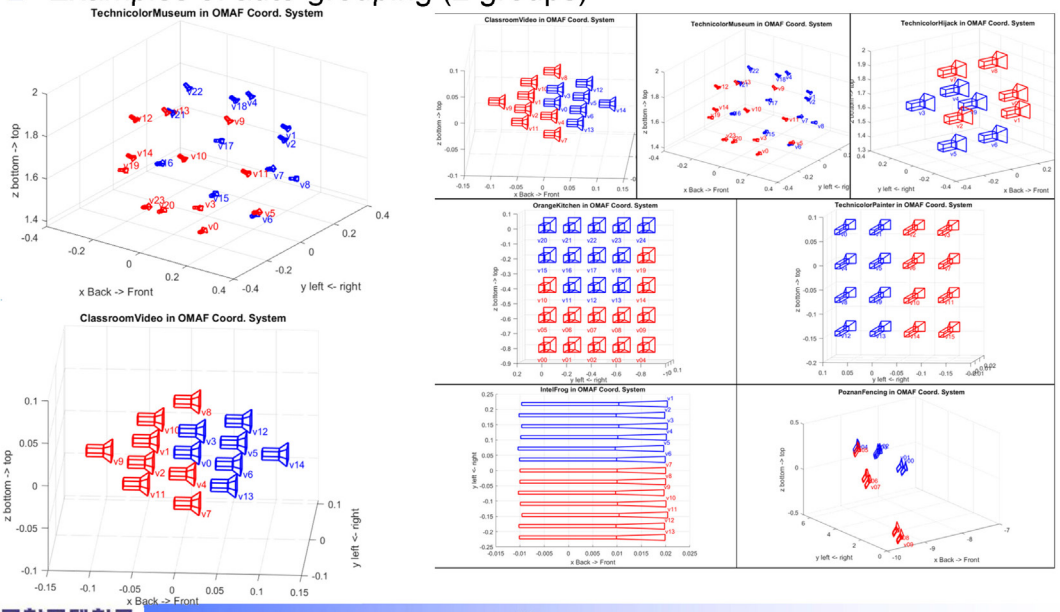
Preparation of Source Views

- Splitting source views in groups
  - Helps rendering local coherent projections of important regions
    - Having fewer samples of the regions in a single group
  - Divide source views into few groups
    - Based on the view parameters list, the number of groups (set by a user)
  - Encoded independently (in parallel)



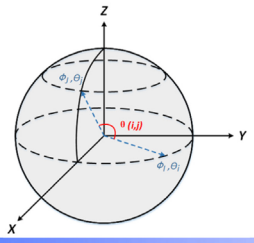
Preparation of Source Views

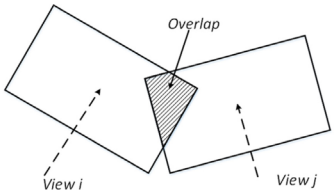
- Examples of auto-grouping (2 groups)



Preparation of Source Views

- View labeling
  - Selects one or multiple basic views from the source views in a group
    - The rest are labeled as additional views
  - Based on camera parameters list
    - Position $[x, y, z]$, orientation $(yaw, pitch, roll)$ $[\alpha_d, \beta_d, \gamma_d]$, projection type
  - Two steps for selecting the view
    - Step 1 – Determination of the number of basic views
      - Direction deviation (θ), FOV, distance, overlap between views
    - Step 2 – Selection of the basic views
      - Distance to a central view position, FOV overlap



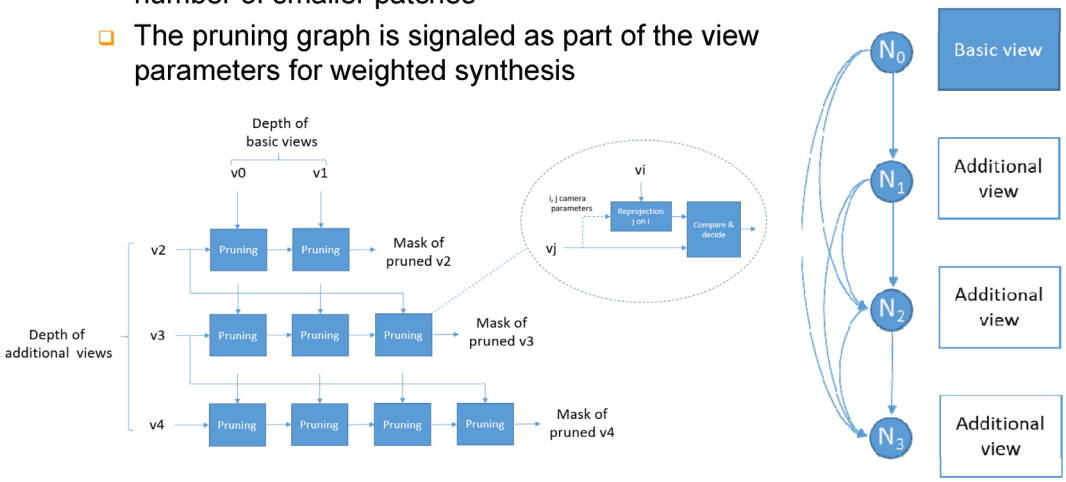


2) Pixel Pruning

- Pruner
  - Creation of the masks, which indicates the part of the input view to be kept further in the pipeline
    - Difference between the basic view and the input view
  - Basic view's mask is filled with the value '255' – preserved
  - Additional view's mask is filled with the value 'pruned (invalid)' (0) or 'preserved (valid)' (255)
    - Obtained by re-projecting each depth pixel value of the additional view onto each basic view, and validate or invalidate the pixel
    - Done in a ladder type process

Pixel Pruning

- Pruning graph – defines hierarchy of view pruning
  - Select the least overlapping additional view as the upper order
  - Prefer fewer larger patches instead of large number of smaller patches
  - The pruning graph is signaled as part of the view parameters for weighted synthesis



Pixel Pruning

- Aggregator
  - The mask is reset at the beginning of each intra period
  - An accumulation is done at the pixel level, across the different frames of the intra period

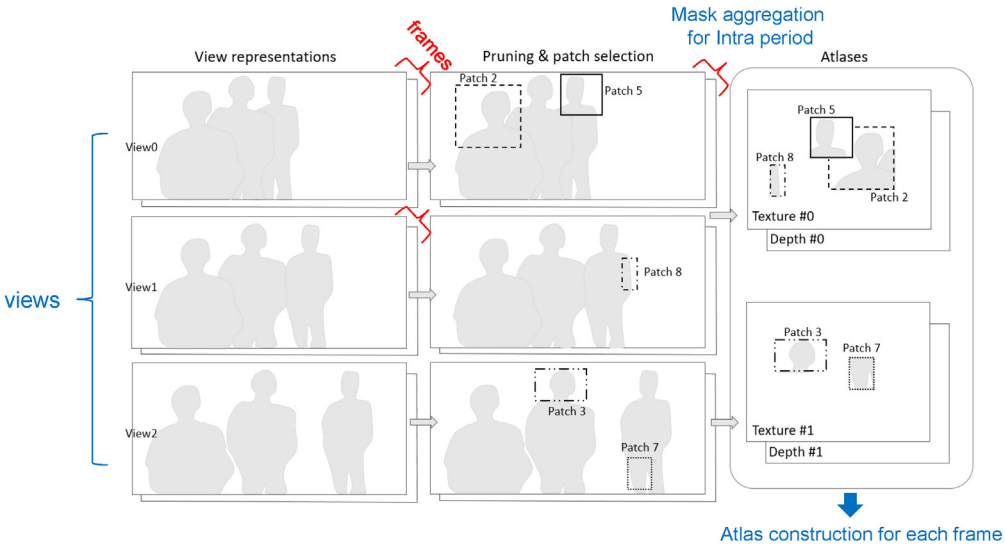
AggregatedMask @ frame i

AggregatedMask @ frame i + k

Accumulation of non-null samples between the frame i and $i + k$
Contours are getting thicker on the changing parts of the geometry accounting for the motion

Pixel Pruning

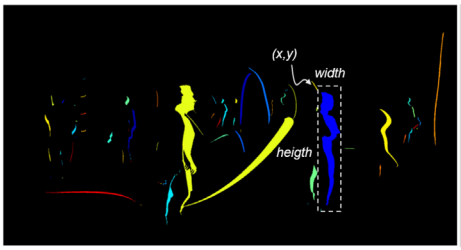
- Aggregator



3) Atlas Construction

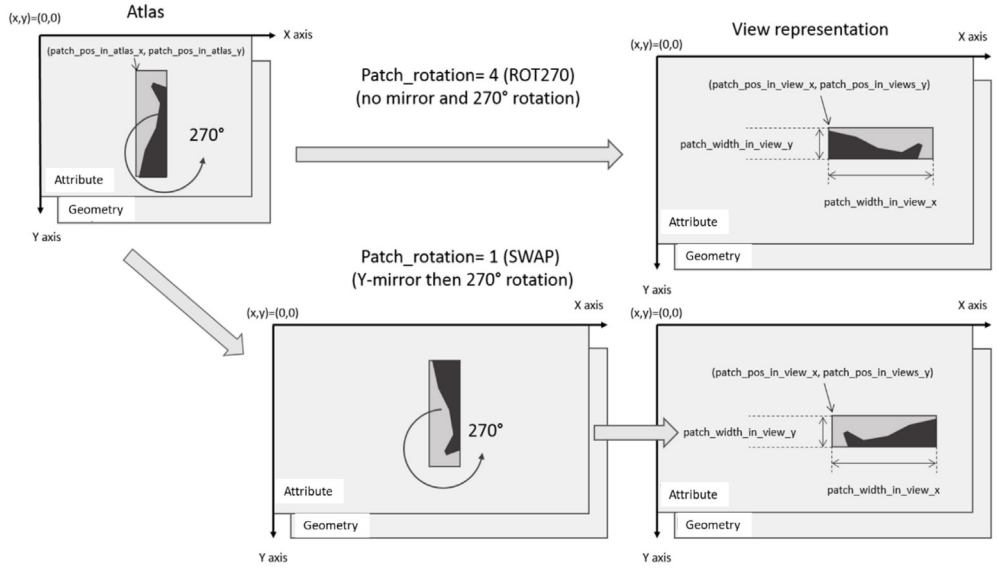
- Clustering active pixels
  - Cluster
    - A set of connected mask pixels of 1s obtained by the mask aggregator
    - Connection criteria – the presence of at least one other pixel among the 8 neighbors
  - Parameters of each cluster
    - x, y positions of the top left corner of the bounding box
    - Width and height of the bounding box
  - The clusters are then sorted by a decreasing size order

1	2	3
4		5
6	7	8



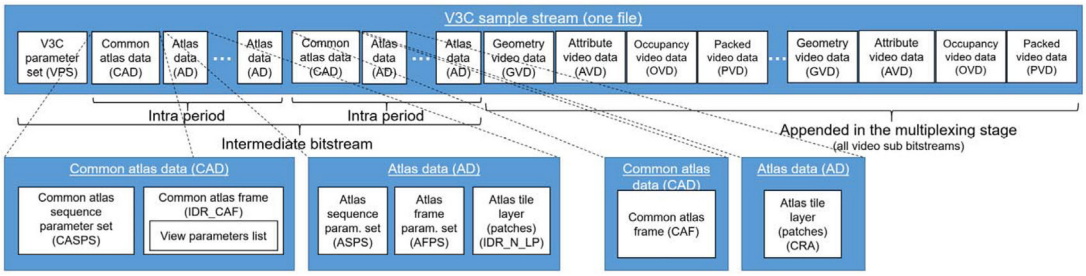
Atlas Construction

- Atlas data



Bitstream

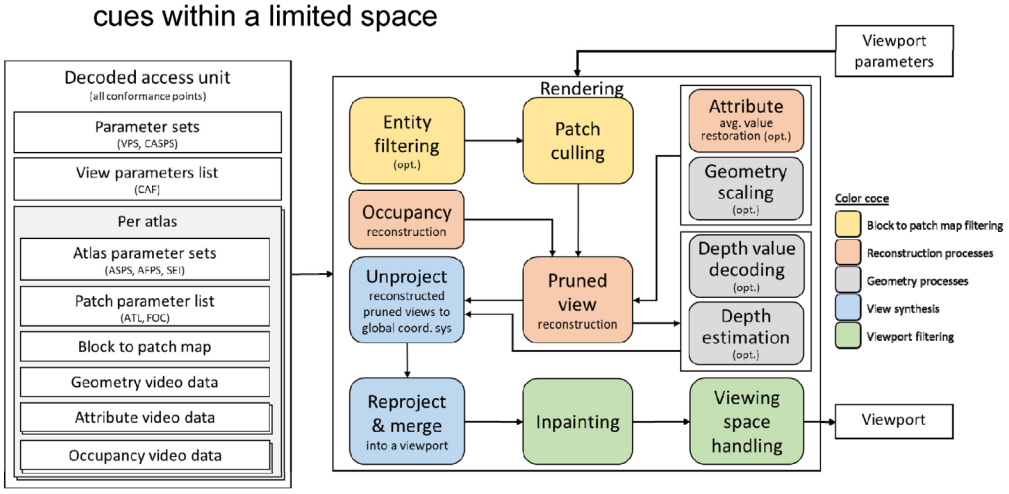
- Bitstream formation and multiplexing
  - TMIV encoder output – a V3C sample stream with MIV extensions



V3C sample stream with MIV extensions

2. Architecture – Decoder

- Decoder side
  - Most of the process is done by the reverse order of encoder
  - Output of the decoder is the desired viewport – enabling motion parallax cues within a limited space



3. Experiment Results

- TMIV encoder – TMIV5.0
  - *InterdigitalHijack*

v0_depth_4096x2048_yuv420p16le.yuv
v0_texture_4096x2048_yuv420p10le.yuv
v1_depth_4096x2048_yuv420p16le.yuv
v1_texture_4096x2048_yuv420p10le.yuv
v2_depth_4096x2048_yuv420p16le.yuv
v2_texture_4096x2048_yuv420p10le.yuv
v3_depth_4096x2048_yuv420p16le.yuv
v3_texture_4096x2048_yuv420p10le.yuv
v4_depth_4096x2048_yuv420p16le.yuv
v4_texture_4096x2048_yuv420p10le.yuv
v5_depth_4096x2048_yuv420p16le.yuv
v5_texture_4096x2048_yuv420p10le.yuv
v6_depth_4096x2048_yuv420p16le.yuv
v6_texture_4096x2048_yuv420p10le.yuv
v7_depth_4096x2048_yuv420p16le.yuv
v7_texture_4096x2048_yuv420p10le.yuv
v8_depth_4096x2048_yuv420p16le.yuv
v8_texture_4096x2048_yuv420p10le.yuv
v9_depth_4096x2048_yuv420p16le.yuv
v9_texture_4096x2048_yuv420p10le.yuv

⇒

ATL_SC.bit
ATL_SC_TG_c00_2048x1088_yuv420p10le.yuv
ATL_SC_TG_c01_2048x1088_yuv420p10le.yuv
ATL_SC_TT_c00_4096x2176_yuv420p10le.yuv
ATL_SC_TT_c01_4096x2176_yuv420p10le.yuv

Atlas data (.bit)
+
2 geometry atlases (downscaled)
+
2 attribute atlases

Experiment Results

- TMIV encoder (*InterdigitalHijack*)

TMIV_A17_C_tex_c00_4096x2176_yuv420p10le.yuv

Atlas with a basic view + patch

TMIV_A17_C_tex _c01_4096x2176_yuv420p10le.yuv

Packed patches of all views

10 source views
↓
2 atlases
(1 group, 1 basic view)

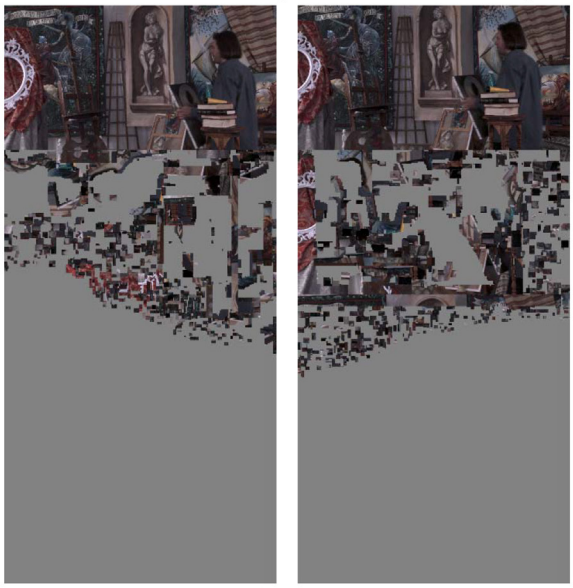
Experiment Results

- TMIV encoder (*TechnicolorMuseum*)

24 source views
↓
2 atlases
(1 group, 2 basic views)

Experiment Results

- Atlas constructor (*TechnicolorPainter*)



16 source views
↓
2 atlases
(2 groups, 1 basic view per each group)

IV. Future MIV EE

- EE1. IVDE(Immersive Video Depth Estimation) depth generation
  - MIV anchor based on the depth maps generated by the IVDE
    - IVDE 4.0 will be used for depth map generation
- EE2. MV-HEVC and 3D-HEVC
  - Selecting the most relevant codec for the verification tests
- EE3. Multiple texture patches per geometry patches
  - Use of multiple texture patches per geometry patch to synthesize view-dependent effects of non-Lambertian surfaces
  - Non-Lambertian effects are common in most natural scenes and in high quality synthetic content

V. Summary

- Provisions of immersive media experience
  - New media type in 5G era – volumetric video
    - 360 video/VR, 6DoF, Point Cloud, Light Field
- Activities on standardization for immersive media – MPEG-I
  - Provide enablers for production, coding, transmission, and consumption
- Compression of immersive video is an essential to handle a huge volume of immersive video
  - Pixel rate reduction and 2D codec – HEVC, VVC
- CEs on further enhancements
  - Improvements on atlas generation for coding efficiency
  - Metadata
- VVC will be a new codec for 6DoF
  - 6DoF Exploration Experiments

Thank you for your attention!