Multi-Screen Service Forum Specification

MSS.S-Y19-002

제정일: 2019년 8월 10일

360도 영상 타일 병합을 위한 분할 영역 부가정보 구성 요소 및 형식

Syntax and Semantics of Divided Region for Merging 360 Video Tiles

제출일 : 2019 년 7 월 31 일

제출기관 : 멀티스크린서비스포럼

제출인 : 류은석

서 문

1 표준의 목적

이 표준의 목적은 가상 현실(VR) 기술과 머리 장착형 영상장치(HMD), 타일(Tile) 기반 영상 분할 및 병합 기법을 사용하는 360도 영상의 적응적인 전송을 위해 영상을 분할하는 단위로 타일을 이용한다. 이 타일이 구성된 정보를 메타데이터(Metadata)로 구성하여, 영상 전송 시 대역폭을 낮추고 저지연 전송을 달성하는 데 있다.

2 주요 내용 요약

이 표준은 사용자가 머리 장착형 영상장치를 이용하여 바라보는 위치를 기반으로 영역을 특정해 고화질의 영상을 요구하기 위한 분할 영상들의 정보와 분할된 타일 정보, 고효율 비디오 부호화(HEVC)의 병렬 처리 기술 도구인 타일 분할 기법을 적용한 영상을 에지(Edge) 또는 STB 장비에서 병합하기 위한 메타데이터로 구성하는 기술 및 표준 신호체계 규격(구문과 의미론)을 기술한다.

3 인용 표준과의 비교

이 표준은 국제 표준단체(JCT-VC)의 MCTS 기술 등을 이용하는 시스템을 위한 별도의 독립적인 시그널링 표준으로서, JCT-VC의 비디오 코딩 표준과 직접적인 관련성이 없음.

Preface

1 Purpose

The standard is to use tiles as a unit to divide video for adaptive 360 video streaming using VR technology, HMD, and Tile based video extracting and merging method. The standard is to define the syntax and semantics of the tile information configured and reduce bandwidth and achieve low-latency transmission during video transmission.

2 Summary

This standard describes the syntax and semantics of the tile segmentation techniques for merging on edge or set-top box equipment, a parallel processing technology tool of HEVC, to specify areas based on the location viewed by a user, information on the segmented tile address to demand high quality video based on the position the user views using head-mounted display, and information expressed in 2D coordinates to specify FOV.

3 Relationship to Reference Standards

The standard can use the referenced video coding standard specifications such as the MCTS of the JCT-VC. But, the standard does not directly affect to or influenced by the referenced standard but specifies the signaling details independently.

목 차

360도 영상 타일 병합을 위한 분할 영역 부가정보 구성 요소 및 형식
(Syntax and Semantics of Divided Region for Merging 360 Video Tiles)

1 적용 범위

본 표준의 적용 범위는 비디오 통신에서 전달되는 신호 체계 정보를 정의하며, 메타데이터 구문(Syntax) 및 의미론(Semantics)은 세션(Session) 정보를 포함하는 고수준 구문(High-level Syntax) 프로토콜을 통해 전해질 수도 있고, 비디오 파일을 설명하는 별도의 파일로(예: DASH의 MPD) 전달될 수 있다. 본 표준은 고효율 비디오 부호화 (HEVC) 타일 정보를 위해 사용되며 다른 비디오 병렬처리 기법들(예: 슬라이스 (Slice), FMO 등)에 적용 가능하다. 또한, 머리장착형 영상장치를 이용하여 360도 영상 스트리밍을 하게 될 때 적용된다.

2 인용 표준

3 용어 정의

에지(Edge) 장비는 머리 장착형 영상장치를 통해 영상을 시청하기 위해 통신을 송수신하는 기기를 의미하며 각 에지 장비에서 본 표준의 메타데이터를 처리하여 영상 병합이 가능하다. STB(Set-Top Box)는 머리 장착형 영상장치를 보조하기 위한 기기로써, 에지 장비의 일종으로써 영상 병합 기법을 적용한 후처리(Post-processing) 과정을 수행한다.

4 약어

DASH Dynamic Adaptive Streaming over HTTP
EIS Extraction Information Sets
SEI Supplemental Enhancement Information
HEVC High Efficiency Video coding
FMO Flexible Macroblock Ordering
JCTVC The Joint Collaborative Team on Video Coding
MCTS Motion Constrained Tile Sets
MPD Media Presentation Description
MPEG Moving Picture Expert Group
STB Set Top Box

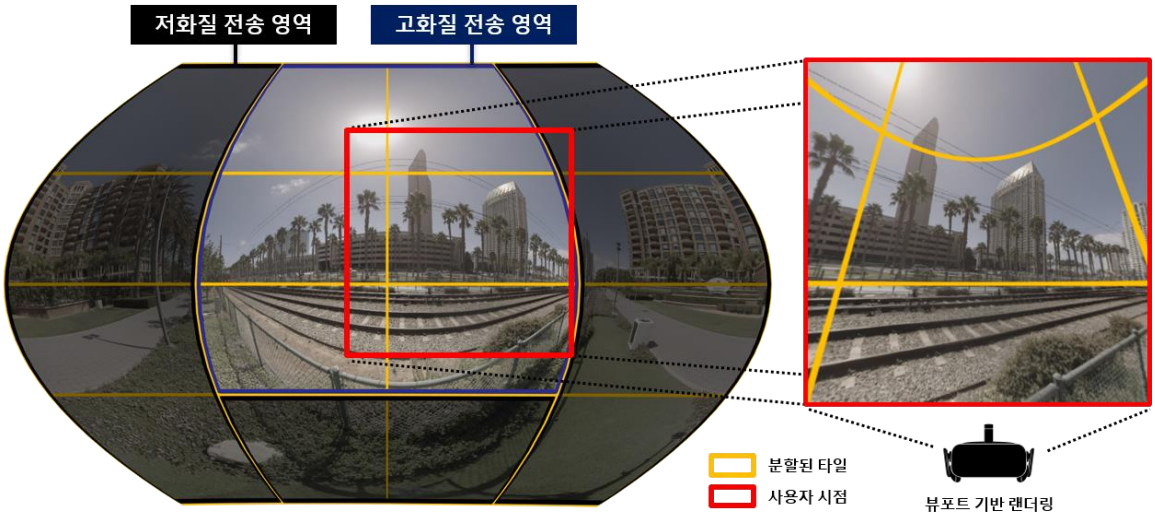
FOV Field Of View

5 분할된 영상 병합을 위한 분할 영역 메타데이터의 구성 요소 및 형식

5.1 분할 영역 메타데이터의 구성 요소 및 형식

5.1.1 비디오 프로세싱 및 렌더링 속도와 대역폭 처리

360도 영상에서 사용자의 시점에 해당하는 영역은 영상의 일부분이다. 고해상도 360도 영상을 전송하기 위해 압축된 영상 비트스트림(Bitstream)으로 받아서 이를 복호화하고 사용자가 바라보는 영역을 가상의 공간에 렌더링(Rendering)하는 기술은 고해상도로 이루어진 360도 영상 전체를 사용한다. 따라서, 비트스트림이 전송되는 대역폭은 매우 클 수밖에 없고 사용자 시점이 위치하지 않는 영역의 비트스트림을 복호화하고 렌더링을 하여 비디오 프로세싱과 렌더링 속도를 저하시키는 문제가 있다. 이를 막기 위해서 국제 비디오 표준 기법 독립적 움직임 참조를 이용한 타일 분할(MCTS)기법과 추출 정보 집합(EIS)에 대한 추가적인 향상 정보(SEI) 메시지가 사용될 수 있다. 다음 (그림 5-1)은 사용자 시점이 위치하는 타일 기반으로 렌더링하는 방법을 보여준다.



(그림 5-1) 사용자 시점기반 360 도 영상

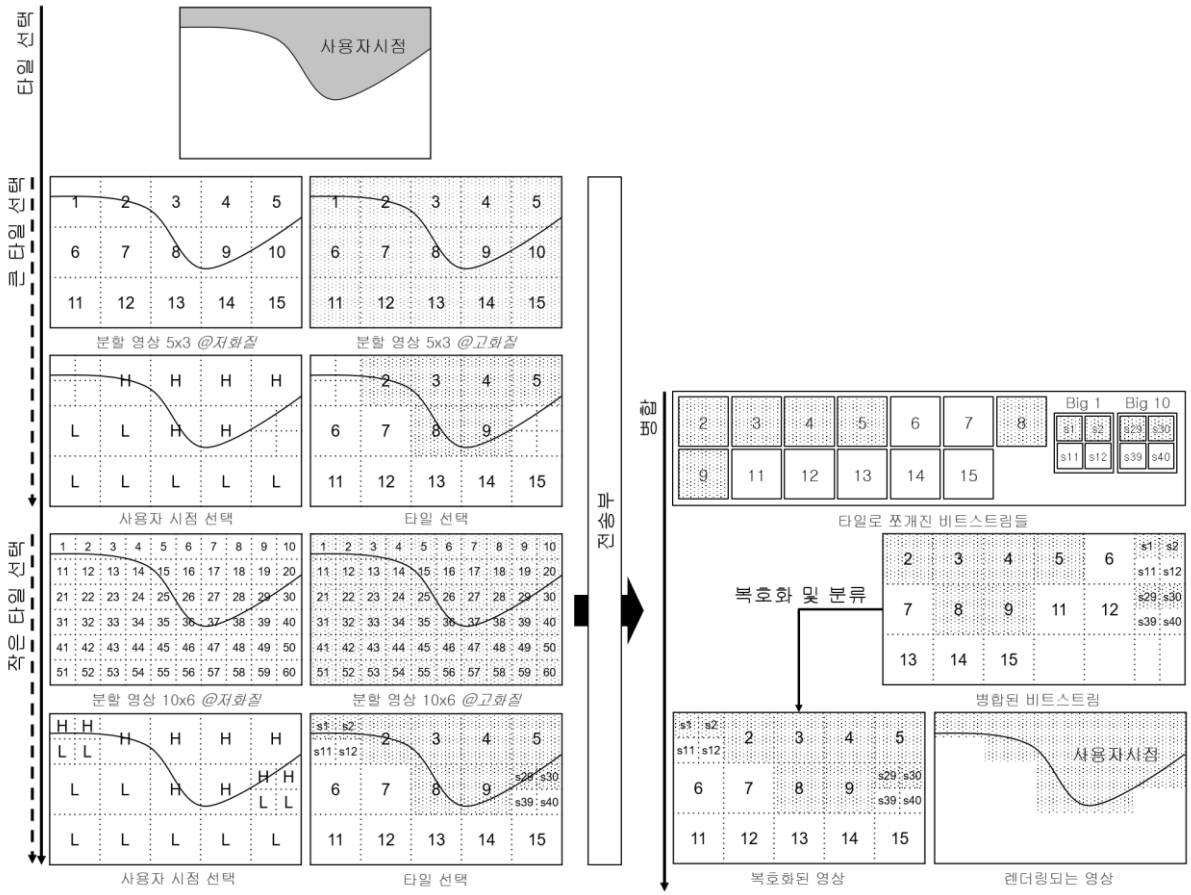
(그림 5-1)은 사용자가 시선 추적 기능과 움직임 추적 기능을 제공하는 머리 장착형 장치를 사용하여 360도 영상 스트리밍(Streaming)을 보여준다. 빨간 상자는 사용자 시점으로 현재 보여지고 있는 화면이다. 이 부분에 해당하는 타일들은 사용자가 보는 영역이므로 고화질이 요구되고, 따라서 해당하는 타일들을 고화질로 전송하고 나머지 영역을 저화질로 전송한다. 이렇게 저화질, 고화질 영역을 분리하여 전송함으로써 사용자에게 몰입감이 있는 영상을 제공하고 사용자의 주관적 화질 대비 대역폭 절감 및 복호화 연산 복

잡도 감소, 렌더링 속도 향상과 머리 장착형 장치의 주사율 증가 효과가 있을 수 있다.

5.1.2 분할된 영상의 병합

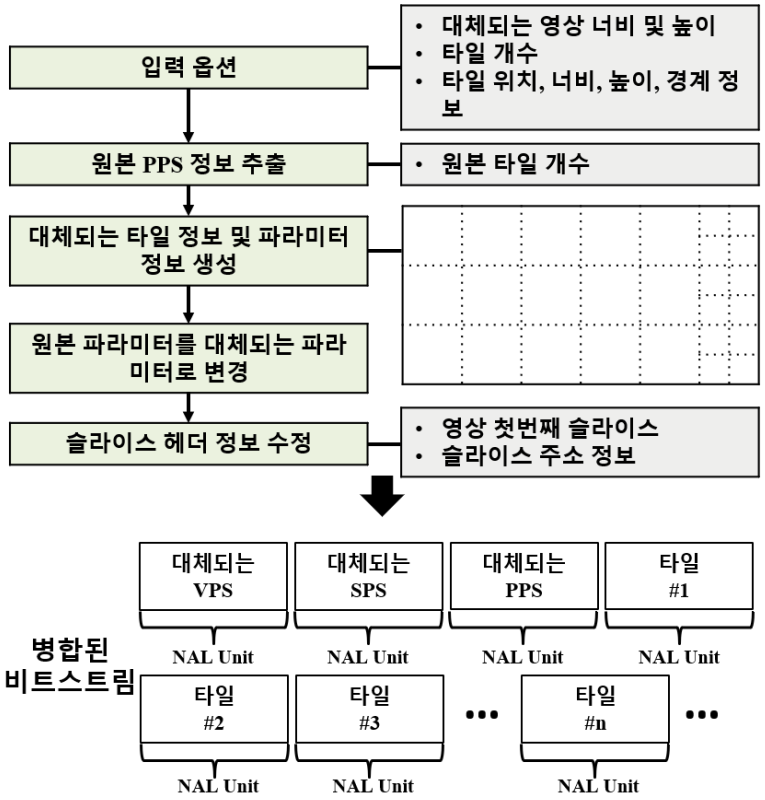
360도 분할 영상을 이용하는 스트리밍 환경에서 서버와 사용자 간 발생하는 지연 시간을 줄이는 일은 주요 과제이다. 사용자 시점 영역이 변경되었을 때, 사용자는 변경된 사용자 시점 영역의 고화질 타일로 이루어진 패킷을 다시 요청한다. 이 때, 저화질 타일로 사용자에게 몰입감 있는 환경을 일시적으로 유지하고 고화질 영상이 다시 머리 장착형 영상 장치에 재주사하는 데까지 걸리는 지연 시간을 최소화함으로써, 사용자에게 좀 더 몰입감 있는 환경을 제공할 수 있다.

본 표준 기술은 화질과 크기를 다르게 타일로 분할하여 구성된 영상들의 분할된 구조 정보를 이용한다. 분할된 영상들을 전송받기 위해 영상 분할 정보가 우선적으로 전송되며, 사용자는 사용자의 시점이 위치한 타일을 판단하여 송신 측에 타일로 분할되어 있는 영상을 요구 가능하다. 또한, 각 타일로 분할된 영상들에 대한 화질과 크기 정보는 각 분할 영상을 하나의 영상으로 병합하기 위해 생성되는 영상의 병렬 처리 기술인 슬라이스, 타일의 구조를 정하는데 이용된다. 다음 (그림 5-2)는 화질 및 크기가 서로 다른 비대칭 분할 영상들을 사용자 시점에 기반하여 하나의 영상으로 합치는 과정을 보여준다.



(그림 5-2) 사용자 시점 기반 비대칭 분할 영상의 병합 과정

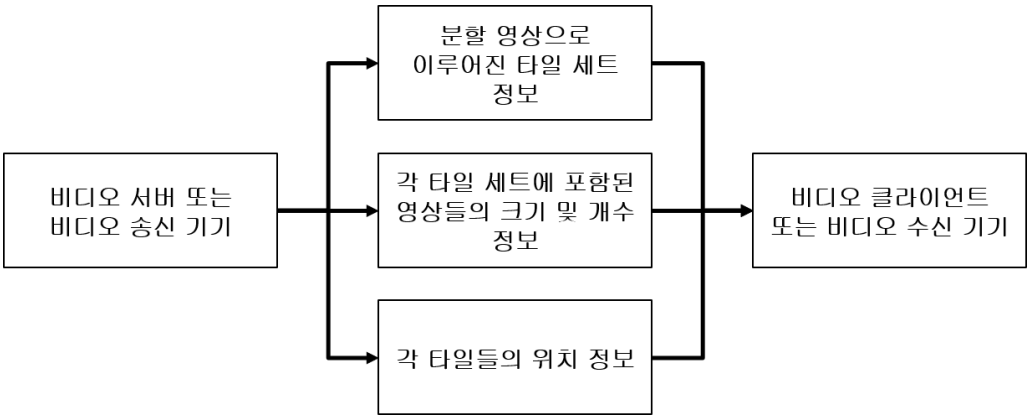
(그림 5-3)은 분할 영상을 병합하는 기능 흐름도를 나타낸다. 비대칭 분할 영상들을 병합하기 위해 먼저 미리 수신된 영상 분할 구조 정보와 사용자 시점에 위치에 따라 화질을 다르게 분할 영상들을 요구하여 패킷을 수신한다. 수신된 영상들은 하나의 영상으로 병합하기 위해 영상 분할 구조 정보에 포함된 크기에 따라 분류하여 (그림 5-3)의 기능을 수행하여 병합한다. 이렇게 하나로 병합된 영상을 복호화 하면 여러 개의 영상을 복호화 하는 과정보다 다중 처리에 이용되는 자원을 감소시킬 수 있다.



(그림 5-3) 타일 병합기의 기능 흐름도

5.2 표준 신호 체계 규격

본 표준의 핵심 신호 체계는 (그림 5-4)와 같이 비디오 수신 장치인 머리 장착형 영상장치가 360도 영상 스트리밍 서버로부터 전달받는 분할 영상으로 이루어진 타일 세트에 대한 정보, 각 타일 세트에 포함된 영상들의 크기 및 개수 정보, 그리고 각 타일들의 위치 정보를 포함한다.



(그림 5-4) 신호 체계 핵심 전달 정보

이 신호 체계는 세션 정보를 실어 나르는 상위 수준 구문 프로토콜을 통해 전해질 수도 있고, 비디오 파일을 설명하는 별도의 파일로(예: DASH의 MPD) 전달될 수 있다.
다음 <표 5-1>는 파일, 청크, 비디오 픽처 그룹별 뷰포트 신호 체계 규격이다. 표에 나온 u(n)는 통상 프로그래밍 언어에서 부호가 없는(unsigned) 'n' 비트 수를 의미한다.

<표 5-1> 파일, 청크, 비디오 픽처 그룹별 신호 체계 규격

tile_sets_info {	비트 수
version_info	u(8)
file_size	u(64)
num_tile_set	u(8)
for (i=0; i < num_tile_set; i++) {	
tile_set_id[i]	u(8)
pic_width_in_luma_samples[i]	u(16)
pic_height_in_luma_samples[i]	u(16)
max_tile_width_in_luma_samples[i]	u(16)
max_tile_height_in_luma_samples[i]	u(16)
tile_set_quality[i]	u(32)
num_tile_in_columns[i]	u(8)
num_tile_in_rows[i]	u(8)
num_tile[i]	u(16)
for (j=0; j < num_tile[i]; j++) {	
tile_id[j]	u(16)
tile_width_in_luma_samples[j]	u(16)
tile_height_in_luma_samples[j]	u(16)

tile_x_offset[j]	u(16)
tile_y_offset[j]	u(16)
}	
}	
}	

다음 <표 5-2>는 분할 영상 구조 정보에 대한 구문 의미론이다.

<표 5-2> 분할 영상 구조 정보에 대한 구문 의미론

구문	의미론
version_info	신호 체계 규약의 버전 정보, 부호 없는 8 비트의 정보로 표현된다.
file_size	파일 사이즈, 부호 없는 64 비트의 정보로 표현된다.
num_tile_set	전체 타일 세트의 개수를 의미. 부호 없는 32 비트의 정보로 표현된다.
tile_set_id	타일 세트의 구분 값(id)을 의미. 부호 없는 8 비트의 정보로 표현된다.
pic_width_in_luma_samples	타일로 구성된 영상의 휘도 너비를 의미, 부호 없는 16 비트의 정보로 표현된다.
pic_height_in_luma_samples	타일로 구성된 영상의 휘도 높이를 의미. 부호 없는 16 비트의 정보로 표현된다.
max_tile_width_in_luma_samples	영상 내 타일들 중 가장 큰 휘도 너비를 의미. 부호 없는 16 비트의 정보로 표현된다.
max_height_in_luma_samples	영상 내 타일들 중 가장 큰 휘도 높이를 의미. 부호 없는 16 비트의 정보로 표현된다.
tile_set_quality	영상 자체의 화질을 의미. 임의의 화질 정보 또는 초당 전송률로 표현될 수 있다. 부호 없는 32 비트의 정보로 표현된다.
num_tile_in_columns	영상 내 타일들을 이루는 구조의 열 개수를 의미. 부호 없는 8 비트의 정보로 표현된다.
num_tile_in_rows	영상 내 타일들을 이루는 구조의 행 개수를 의미. 부호 없는 8 비트의 정보로 표현된다.
num_tile	영상 내 타일 개수를 의미. 부호 없는 16 비트의 정보로 표현된다.

tile_id	타일 구분 값(id)을 의미. 부호 없는 16 비트의 정보로 표현된다.
tile_width_in_luma_samples	타일의 휘도 너비를 의미. 부호 없는 16 비트의 정보로 표현된다.
tile_height_in_luma_samples	타일의 휘도 높이를 의미. 부호 없는 16 비트의 정보로 표현된다.
tile_x_offset	타일의 x 좌표 위치를 의미. 부호 없는 16 비트의 정보로 표현된다.
tile_y_offset	타일의 y 좌표 위치를 의미. 부호 없는 16 비트의 정보로 표현된다.

정의된 구문과 의미론에 관한 정보들은 MPEG DASH와 같은 HTTP 기반의 영상 통신에서 각각 XML 형태로 표현이 될 수도 있다. 다음 <표 5-3>은 XML 형태로 타일 세트 개수, 타일 세트 정보, 각 타일 정보의 표현 형식이다.

<표 5-3> XML 형태로 표현된 분할 영상 구조 정보 구문

```
<tile_sets_info  num_tile_set = “6”>
    <tile_set tile_set_id = “0” pic_width_in_luma_samples = “4096” pic_height_in_luma_samples = “2048”
    max_tile_width_in_luma_samples = “768” max_tile_height_in_lujma_samples = “640” tile_set_quality =
    “50423” num_tile_in_colums = “5” num_tile_in_columns = “3” num_tile = “15”>
        <tile tile_id = “0” tile_width_in_luma_samples = “768” tile_height_in_luma_samples = “640”
        tile_x_offset = “0” tile_y_offset=”0” />
        ……
    </ tile_set>
    ……
    <tile_set tile_set_id = “3” pic_width_in_luma_samples = “4096” pic_height_in_luma_samples = “2048”
    max_tile_width_in_luma_samples = “384” max_tile_height_in_lujma_samples = “320” tile_set_quality =
    “45812” num_tile_in_colums = “10” num_tile_in_columns = “6” num_tile = “30”>
        <tile tile_id = “0” tile_width_in_luma_samples = “384” tile_height_in_luma_samples = “320”
        tile_x_offset = “0” tile_y_offset=”0” />
        ……
    </ tile_set>
    ……
    <tile_set tile_set_id = “6” pic_width_in_luma_samples = “2048” pic_height_in_luma_samples = “1024”
    max_tile_width_in_luma_samples = “384” max_tile_height_in_lujma_samples = “320” tile_set_quality =
    “26054” num_tile_in_colums = “5” num_tile_in_columns = “3” num_tile = “15”>
```

```
        <tile tile_id = “0” tile_width_in_luma_samples = “384” tile_height_in_luma_samples = “320”
        tile_x_offset = “0” tile_y_offset=”0” />
        ……
    </ tile_set>
</ tile_sets_info>
```

부 록 Ⅰ-1

(본 부록은 표준을 보충하기 위한 내용으로 표준의 일부는 아님)

필요성, 배경지식, 확장적 사용

Ⅰ-1.1 본 표준의 필요성

최근 가상 현실 기술 응용과 장비의 발달과 함께 사용자 시점 기반한 360도 영상 전송 기술 및 기기들이 상용화되고 있다. 단순히 360도 영상을 시청하는 환경에 국한되지 않고 그래픽 및 게임 콘텐츠, 의학 등 산업 전반에 걸친 응용 기술들이나 플랫폼 내에 영상을 시청하는 환경, 영상을 부분적으로 시청하는 환경, 여러 영상을 동시에 시청하는 등의 응용 기술들이 발전하고 있다. 360도 영상 분할 전송 기법은 여러 분야 및 플랫폼들에 걸쳐 통용되고 기초되는 기술이며, 따라서 분할 전송 기법에 대한 연구와 표준화가 활발하게 진행되고 있다.

360도 영상을 높은 몰입감으로 보기 위해 최소한 UHD 급의 고화질 영상이 요구되며, 이러한 UHD 동영상의 전송은 높은 대역폭을 요구한다. 높은 대역폭의 요구는 낮은 대역폭을 유지하는 무선 통신 환경만 아니라, 유선 환경에 걸쳐서도 대역폭의 제한적인 문제들을 나타낸다. 따라서, 고해상도 360도 영상의 효율적인 전송 기법들이 요구되었으며 현재는 분할 영상 전송 기법이 최근 환경에 가장 적응적인 기술로써 많은 연구가 진행되고 있다.

현재, 국제 표준화 단체인 MPEG(Moving Picture Expert Group)에서 진행 중인 표준인 OMAF(Omnidirectional MediA Format)는 HEVC(High Efficiency Video Coding)을 이용한 사용자 시점 기반 분할 영상 전송을 위해 ISOBMFF(ISO Base Media File Format) 상에 분할 기법이 적용된 영상의 구조 정보를 포함한 기본(Base) 트랙(Track)과 하나의 영상으로 병합되기 위한 구조 정보를 포함한 추출기(Extractor) 트랙을 정의한다. 베이스 트랙과 추출기 트랙은 원본 영상과 병합될 영상 정보를 포함하여, 베이스 트랙 내에 추출기 트랙이 포함되어 원본 영상과의 차이점을 통해 병합이 수행 가능해진다. 이렇게 ISOBMFF로 정의된 베이스 트랙과 분할된 영상들로 이루어진 타일 트랙들은 MPEG의 스트리밍 기술인 DASH(Dynamic Adaptive Streaming over HTTP) 상에 전송 가능한 형태로 정의되었으며, 이 구조는 MPD(Media Presentation Description) 형태로 전송되어 각 트랙들을 사용자 시점에 기반한 전송이 가능해진다. 하지만, 추출기 트랙을 포함한 전송은 사용자 시점에 기반한 모든 경우를 고려하여 서버에서 추출기 트랙들을 정의하여야 하고 각 추출기 트랙을 스트리밍 기술 상에서 처리해야하는 문제로, 실질적으로 모든 경우를 고려 불가능하여 기본적인 추출기 트랙을 가지고 다중 처리를 하는 방식 등의 제안 등이 기고되었다. 원본 영상들의 구조와 화질 정보와 같은 우선 순위 정보를 포함하여 전송할 경우, OMAF 표준의 정의를 고려하여 추가적인 정보로써 VCL(Video Coding Layer)에서 영상을 병합하여 하나의 영상만을 복호화할 수 있다. 본 표준은 이의 해결책으로 ‘분할

영상 구조 정보 전송 및 병합 과정'의 내용과 장점, 그리고 이를 통한 대역폭 절감, 저지연 전송을 설명한다.

I-1.2 배경지식

국제 비디오 코딩 표준 단체인 JCTVC(The Joint Collaborative Team on Video Coding)의 표준 기술 중 독립적 움직임 참조를 이용한 타일 분할(MCTS, Motion Constrained Tile Sets)기법과 추출 정보 집합(EIS, Extraction Information Sets)에 대한 추가적인 향상 정보(SEI, Supplemental Enhancement Information) 메시지가 있다. 또한, MPEG에서 표준화한 OMAF에서 정의한 사용자 시점에 기반한 분할 영상 전송 기법에 기반한다. 독립적 움직임 참조를 이용한 타일 분할 기법은 일시적으로 부호화기(Encoder)에서 참조 프레임(Frame)으로의 시간적 움직임 추정을 타일 내 공간으로 제한하여 시간적 및 공간적으로 독립적인 타일로 구성하고, 타일과 타일을 구별하기 위해 NAL 단위의 슬라이스를 타일과 1:1로 병합한다. 영상으로부터 분할된 타일로 구성된 정보를 담은 EIS SEI 메시지를 이용해 실질적으로 하나의 영상으로부터 타일들을 분할하고 독립적으로 복호화(Decoding)가 가능하다. 독립적으로 복호화 가능한 분할 영상들을 선택적으로 전달받아 VCL(Video Coding Layer) 또는 통신 패킷 단위에서 하나의 영상으로 병합하여 병합할 수 있다. 따라서, 영상 분할 및 전송하여 고해상도 영상의 전송 대역폭을 절감한다.

I-1.3 추가 확장적 사용

본 표준은 타일 분할 기법을 이용한 차별적 전송 기법을 다루지만, 화면 분할을 지원하는 다른 비디오 병렬처리 기법들(예: 슬라이스(Slice), FMO(Flexible Macro Block) 등)에 적용 가능하다. 또한 비트 스트림을 분할하여 전송하는 스트리밍 서비스인 MPEG DASH, Smooth 스트리밍(Streaming), HLS(HTTP Live Streaming, HTTP 라이브 스트리밍)에 적용 가능하고 MPEG에서 표준화 중인 OMAF의 일부를 수정하여 적용할 수 있다.
추가적으로 영상 분할 정보와 분할 영상을 병합하는 기능 흐름도를 따르면, 본 표준 기술의 목적인 360도 분할 영상의 병합 외에 서로 다른 영상을 병합하여 하나의 멀티뷰(Multi-view)를 생성 가능하다.