Multi-Screen Service Forum Specification

MSS.S-Y18-004

제정일: 2018년 7월 23일

시선 기반의 360 비디오 처리를 위한 눈동자 움직임 판단 메타데이터의 구성 요소 및 형식

Syntax and Semantics of Eye Gaze Prediction and Tracking for 360 Video Processing

제출일 : 2018. 7. 23.

제출기관 : 멀티스크린서비스포럼

제출인 : 가천대 류은석 교수

서 문

1 표준의 목적

이 표준의 목적은 최근 부상하고 있는 가상 현실(Virtual Reality; VR) 기술과 머리장착형 영상장치(Head-Mounted Display; HMD), 시선 추적(Eye-Tracking) 기술을 사용하는 시선 기반 360 영상 전송을 위해 눈동자 움직임을 판단할 수 있는 영상 특성 정보를 메타데이터(Metadata)로 구성하여, 영상 전송 시 대역폭을 낮추고 신속한 시선 판단으로 저지연 전송을 달성하는 기술을 설명함에 있다.

2 주요 내용 요약

이 표준은 시선추적이 가능한 머리장착형 영상장치를 통한 타일 기반 360 분할 영상을 전송할 때, 사용자의 시선이 위치하거나 향할 확률이 높은 영역을 특정하기 위해 영상의 특성 정보를 메타데이터로 구성하는 기술 및 표준 신호 체계 규격(구문과 의미론)을 기술한다.

3 인용 표준과의 비교

3.1 인용 표준과의 관련성

이 표준은 국제 표준단체 The Joint Collaborative Team on Video Coding (JCT-VC)의 Motion Constrained Tile Sets (MCTS) 기술 등을 이용하는 시스템을 위한 별도의 독립적인 시그널링 표준으로써, JCT-VC 의 비디오 코딩 표준과 직접적인 관련성이 없음.

Preface

1 Purpose

The standard defines syntax and semantics of the eye gaze prediction and tracking for 360 video processing using Virtual Reality (VR) technology, Head-Mounted Display (HMD), and eye-tracking. The purpose of the standard is to describe the eye gaze prediction which is composed of characteristic information of the video for achieving low-delay transmission with low-speed bandwidth and low-speed visual determination.

2 Summary

The standard includes the specifications for signaling characteristic information of video to specify a region where a user's line of sight is located or is likely to face when transmitting tile-based 360 videos through a head-mounted display capable of eye-tracking. The standard signaling specifications including syntax and semantics.

3 Relationship to Reference Standards

The standard can use the referenced video coding standard specifications such as the Motion Constrained Tile Sets (MCTS) of the Joint Collaborative Team on Video Coding (JCT-VC). But, the standard does not directly affect to or influenced by the referenced standard but specifies the signaling details independently.

목 차

시선 기반의 360 비디오 처리를 위한 눈동자 움직임 판단 메타데이터의 구성 요소 및 형식 (Syntax and Semantics of Eye Gaze Prediction and Tracking for 360 Video Processing)

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 표준의 SEI, VUI, 또는 슬라이스 헤더(Slice Header) 등의 패킷 단위에서 전해질 수도 있고, 비디오 파일을 설명 (Descript)하는 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(Flexible Macro Block) 등)에 적용 가능하다.

2 인용 표준

해당 사항 없음

3 용어 정의

해당 사항 없음

4 약어

DASH Dynamic Adaptive Streaming over HTTP
EIS Extraction Information Sets
HEVC High Efficiency Video coding
HTTP Hyper Text Transfer Protocol
JCTVC The Joint Collaborative Team on Video Coding
MCTS Motion Constrained Tile Sets
MPD Media Presentation Description
MPEG Moving Picture Experts Group
SEI Supplemental Enhancement Information
UHD Ultra High Definition
VUI Video Usability Information

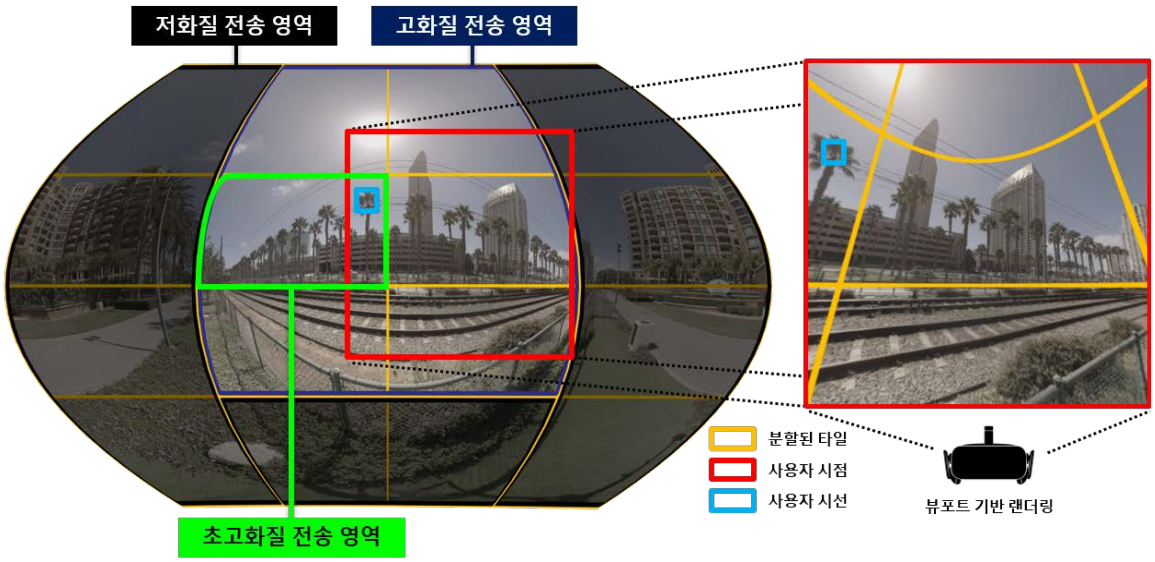
5 시선 기반의 360 비디오 처리를 위한 눈동자 움직임 판단 메타데이터의 구성 요소 및 형식

본 표준은 시선 기반의 360 비디오 처리를 위한 눈동자 움직임 판단 메타데이터의 구성 요소 및 형식 표준에 관한 신호 체계 규격을 구체적으로 설명한다.

5.1 눈동자 움직임 판단 메타데이터의 구성 요소 및 형식

5.2.1. 비디오 프로세싱 및 렌더링 속도와 대역폭 처리

360 영상에서 사용자의 시점에 해당하는 영역은 영상의 일부분이다. 고해상도 360 영상을 전송하기 위해 압축된 영상 비트스트림으로 받아서 이를 복호화하고 사용자가 바라보는 영역을 가상의 공간에 렌더링(Rendering)하는 기술은 고해상도로 이루어진 360 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 중 MCTS와 EIS SEI 메시지가 사용될 수 있다. 그 방법은 그림 5-1과 같다.



(그림 5-1) 사용자 시점 및 시선 기반 360 비디오 전송

그림 5-1은 사용자가 시선 추적 기능과 움직임 추적 기능을 제공하는 머리장착형 장치를 사용하여 360 영상 스트리밍(Streaming)을 할 때를 가정한 것이다. 빨간 상자는 사용자 시점으로 현재 보여지고 있는 화면이다. 이 부분에 해당하는 타일들은 사용자가 보는 영역이므로 고화질이 요구되고, 따라서 해당하는 타일들을 고화질로 전송하고 나머지

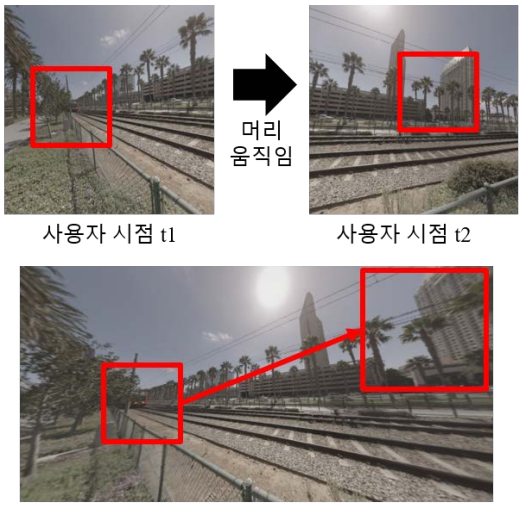
영역을 저화질로 전송한다. 파란 상자는 시선 추적이 된 사용자 시선이 위치한 영역이다. 이 부분에 해당하는 타일은 초고화질로 전송한다. 이렇게 저화질, 고화질, 초고화질 영역을 분리함으로 사용자에게 몰입감이 있는 영상을 제공하고 사용자의 주관적 화질 대비 대역폭 절감 및 복호화 연산 복잡도 감소, 렌더링(Rendering) 속도 향상과 머리장착형 장치의 주사율 증가 효과를 가질 수 있다.

5.2.2. 눈동자 움직임 판단 개선 및 전송 지연 시간 처리

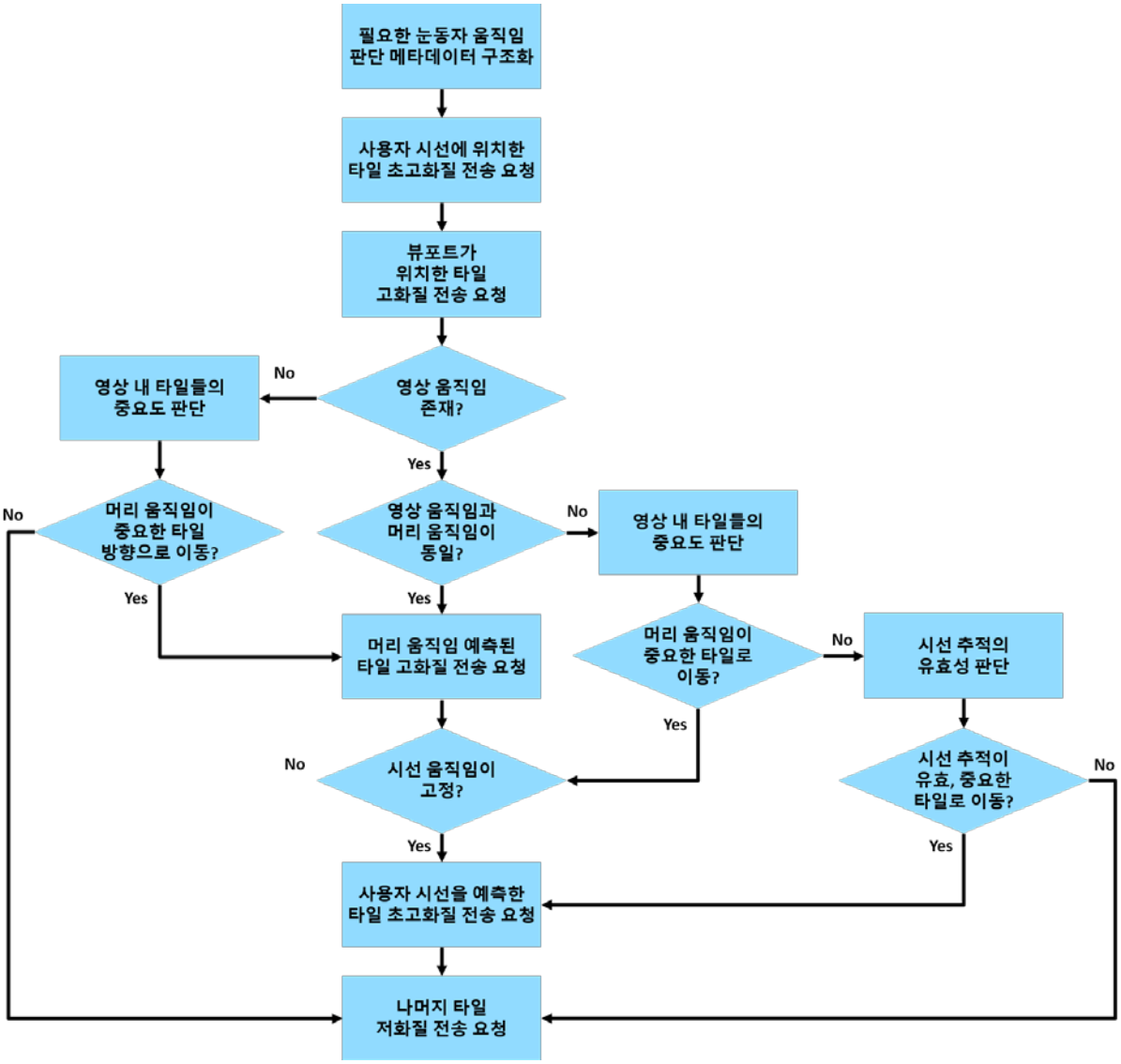
360 분할 영상을 이용하는 스트리밍 환경에서 서버와 사용자 간 발생하는 지연 시간을 줄이는 일은 주요 과제이다. 사용자 시점 영역이 변경되었을 때, 사용자는 변경된 사용자 시점 영역의 고화질 및 초고화질 타일로 이루어진 패킷을 다시 요청한다. 이 때, 저화질 타일로 사용자에게 몰입감 있는 환경을 일시적으로 유지하고 고화질 및 초고화질 영상이 다시 머리장착형 영상장치에 재주사 하는 데까지 걸리는 지연 시간을 최소화함으로써, 사용자에게 좀 더 몰입감 있는 환경을 제공할 수 있다.

본 표준 기술은 영상의 특성 정보인 영상의 움직임과 타일 별 중요도 및 시선 추적의 유효성 판단을 가능하게 하는 정보를 이용한다. 영상의 움직임이 존재하는 경우, 시선이 고정되거나 예측하는 방법은 영상의 움직임과 머리의 움직임, 시선의 움직임과 모두 상관관계가 있다. 영상의 움직임과 머리의 움직임이 동일하다면 영상 내 물체를 주시하는 중이라 판단할 수 있고 머리 움직임이 도달하는 회전 량 또는 이동 량을 이용하여 고화질 타일을 예측하여 전송 요청한다. 또한, 시선이 불안정하거나 급격한 이동 중에 있지 않고 영상 움직임과 상관관계를 가진다면 위와 동일한 이유로 시선의 예측을 판단 가능하고 예측되는 타일을 전송 요청한다. 영상의 타일 별 중요도를 정하여 판단하는 방법은 머리나 시선이 중요도가 높은 타일로 이동할 확률이 높으므로 머리나 시선이 이동할 것이라 생각하고 예측하여 전송 요청한다. 위와 같이 영상 움직임, 머리 움직임, 시선 움직임 간 상관관계와 타일 별 중요도를 정하는 방법은 전송 지연 시간을 낮춘다. 5.2.1과 마찬가지로 영상 타일 분할 전송 기술을 적용하고 5.2.2는 영상 내 타일을 선택하여 전송 요청을 판단하는 기술로 적용된다.

앞서 소개한 5.2.1과 5.2.2의 기법들을 적용한 순서도는 아래 그림 5-3과 같다.



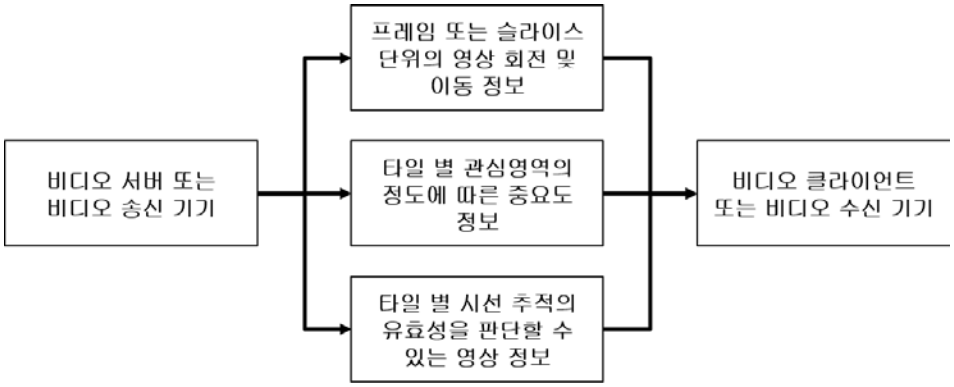
(그림 5-2) 머리 움직임과 시선 움직임을 고려한 시선 판단 기법



(그림 5-3) 본 표준의 기법 적용 순서도

5.2 표준 신호 체계 규격

본 표준에 정의된 신호 체계를 통해 따라서, 본 표준의 핵심 신호 체계는 (그림 5-4)와 같이 비디오 수신 장치인 머리 장착형 영상장치가 360 비디오 스트리밍 서버에 전달하는 영상의 이동 및 회전 정보와 타일 별 관심영역의 정도에 따른 중요도 정보, 시선 추적의 유효성을 판단할 수 있는 정보이다.



(그림 5-4) 신호 체계 핵심 전달 정보

이 신호 체계는 세션 정보를 실어 나르는 상위 수준 구문 프로토콜을 통해 전해질 수도 있고, 비디오 표준의 SEI, VUI, 또는 슬라이스 헤더 등의 패킷 단위에서 전해질 수도 있고, 비디오 파일을 설명하는 별도의 파일로(예: DASH의 MPD) 전달될 수 있다.

<표 5-1>은 H.264 AVC나 H.265 HEVC와 같은 국제 비디오 표준에서의 SEI 메시지 페이로드 (payload) 구문의 형식(tile_characteristics)를 보여주고 있다. 만일 표준 구문이 SEI의 188번으로 정해진 경우 다음과 같다. 검은색 글씨는 기존의 표준 구문이며, 굵은(Bold) 글씨는 새로 추가되어야 할 내용이다.

<표 5-1> SEI 페이로드 구문의 형식

sei_payload(payloadType, payloadSize) {	Descriptor
if(payloadType = = 0)	
buffering_period(payloadSize)	
……..	
else if(payloadType = = 189)	
tile_characteristics(payloadSize)	
……..	

다음 <표 5-2>와 <표 5-3>은 본 표준에서 비디오 픽처별 뷰포트 신호 체계 규격과 파일, 청크, 비디오 픽처 그룹별 뷰포트 신호 체계 규격을 다루고 있고, <표 5-4>는 이의 구문 설명을 한다.

표에 나온 u(n)는 통상 프로그래밍 언어에서 부호가 없는 (unsigned) ‘n’ 비트 수를 의미하며, ‘v’로 표시된 부분은 변화 가능한 비트수(표준에서는 varies로 읽힘)를 의미한다. 또한, i(n)은 부호가 있는 (signed) ‘n’ 비트 수를 의미한다.

<표 5-2> 비디오 픽처별 신호 체계 규격

tile_characteristics (payloadSize) {	**비트수**
rotation_flag	u(1)
translation_flag	u(1)
if(rotation_flag) {	
rotation_x	i(32)
rotation_y	i(32)
rotation_z	i(32)
}	
if(translation_flag) {	
translation_x	i(32)
translation_y	i(32)
translation_z	i(32)
}	
tile_num	u(8)
for (i=0; i < tile_num; i++) {	
tile_region_of_interest_list[i]	u(4)
tile_sharpness_list[i]	u(8)
tile_complexity_list[i]	u(8)
}	
}	

<표 5-3> 파일, 청크, 비디오 픽처 그룹별 신호 체계 규격

tile_characteristics {	**비트수**
version_info	u(8)
file_size	u(64)
for (j=0; j < file_size; j++) {	

poc_num	u(32)
rotation_flag	u(1)
translation_flag	u(1)
if (rotation_flag) {	
rotation_x	i(32)
rotation_y	i(32)
rotation_z	i(32)
}	
if (translation_flag) {	
translation_x	i(32)
translation_y	i(32)
translation_z	i(32)
}	
tile_num	u(8)
for (i=0; i < tile_num; i++) {	
tile_region_of_interest_list[i]	u(4)
tile_sharpness_list[i]	u(8)
tile_complexity_list[i]	u(8)
}	
}	

<표 5-4> 본 표준 구문에 대한 의미론

구문	의미론
version_info	신호 체계 규약의 버젼 정보, 부호 없는 8비트의 정보로 표현된다.
file_size	파일 사이즈, 부호 없는 64 비트의 정보로 표현된다.
poc_num	HEVC와 같은 비디오 표준에서의 POC (Picture Order Count) 정보를 의미함, 기존의 H.264 AVC 표준에서의 프레임 번호와 유사한 의미. 부호 없는 32 비트의 정보로 표현된다.
rotation_flag	영상의 회전 여부에 대한 플래그, 부호 없는 1비트의 정보로 표현된다. 0: 영상이 회전 중이지 않음. 1: 영상이 회전 중에 있음.
translation_flag	영상의 이동 여부에 대한 플래그, 부호 없는 1비트의 정보로 표현된다. 0: 영상이 이동 중이지 않음.

	1: 영상이 이동 중에 있음.
rotation_x	이전 프레임으로부터의 영상 회전 정도이며 구체 모델에 대한 x 축 회전을 의미한다. 부호 있는 32 비트의 정보로 표현된다.
rotation_y	이전 프레임으로부터의 영상 회전 정도이며 구체 모델에 대한 y 축 회전을 의미한다. 부호 있는 32 비트의 정보로 표현된다.
rotation_z	이전 프레임으로부터의 영상 회전 정도이며 구체 모델에 대한 z 축 회전을 의미한다. 부호 있는 32 비트의 정보로 표현된다.
translation_x	이전 프레임으로부터의 영상 이동 정도이며 전역 좌표축을 기준으로 이동 정도를 의미한다. 부호 있는 32 비트의 정보로 표현된다.
translation_y	이전 프레임으로부터의 영상 이동 정도이며 전역 좌표축을 기준으로 이동 정도를 의미한다. 부호 있는 32 비트의 정보로 표현된다.
translation_z	이전 프레임으로부터의 영상 이동 정도이며 전역 좌표축을 기준으로 이동 정도를 의미한다. 부호 있는 32 비트의 정보로 표현된다.
tile_region_of_intrest_list [ ]	타일 내 관심영역이 존재하는 정도를 의미하며, 부호 없는 4 비트의 정보로 표현된다.
tile_sharpness_list [ ]	해당 타일 화질의 선명도를 의미하며, 부호 없는 8 비트의 정보로 표현된다.
tile_complexity_list [ ]	해당 타일 내 물체 움직임와 물체가 존재하는 정도를 의미하며, 부호 없는 8 비트의 정보로 표현된다.

이상 정의된 구문과 의미론에 관한 정보들은 MPEG DASH와 같은 HTTP 기반의 영상 통신에서 각각 XML 형태로 표현이 될 수도 있다. 다음 <표 5-5>는 XML 형태로 정보 모드, 추가 사용자 정의 모드 플래그, 뷰포트 개수 정보, 타일 개수 정보, 전송되는 뷰포트 번호 정보, 전송되는 타일 번호 정보를 표현한 한 예이다.

<표 5-5> XML 형태로 표현된 타일 정보 구문

```
< tile_characteristics>
    <rotation_flag = “0” translation_flag =”1” translation_x = ”2000” translation_y = “3000”
    translation_z = “-2500” tile_num=“6” tile_region_of_interest_list =”15 1 3 2 2 2”
    tile_sharpness_list =”230 150 100 50 96 32” tile_complexity_list=”23 60 31 45 44 240”>
</ tile_characteristics >
```

부 속 서 A

(본 부속서는 표준 내용의 일부임)

본 표준의 필요성, 배경지식, 확장적 사용

A.1 본 표준의 필요성

최근 가상 현실 기술 응용과 장비의 발달과 함께 Fove사의 시선 추적 기반 머리장착형 영상장치와 같은 시선 추적이 가능한 기기들이 상용화되고 있다. 현재 시선 추적이 가능한 연구 분야 뿐만아니라 그래픽 및 게임 콘텐츠, 의학 등 산업 전반에 걸쳐 응용 기술들이 상용화되고 있다. 360 영상과 시선 추적 기술은 여러 분야에 걸쳐 통용될 수 있는 기술이고 360 영상 압축 및 전송이나 시선 추적 판단 및 멀미 저감하기 위한 연구와 표준화가 활발하게 진행되고 있다.

몰입형 실감 미디어인 360 영상의 이용이 급속도로 증가하고 있는데, 이러한 360 영상을 높은 몰입감으로 보기위해 최소한 UHD 급의 고화질 영상이 요구된다. UHD 동영상의 전송은 높은 대역폭을 요구하게 되고, 따라서 고해상도 360 영상의 효율적인 전송 기법으로 시선 추적을 기반한 전송 기법을 요구한다.

아직까지 표준화되거나 표준화된 전송 기술들을 이용하여 360 영상 전송을 할 경우에는 세계 최고 기술 수준은 약 500ms, 국내 최고 기술 수준은 약 3000ms의 관심영역 스위칭 지연 시간을 가지고 있어 지연이 심해 상용화되기 어렵고 이를 해결하기 위한 관심영역 추정 및 판단, 예측 전송 기법을 요구하고 있으며 이러한 문제를 해결한다면 고해상도 360 영상 전송 기술을 응용한 서비스가 널리 보급될 것이라 예상한다.

현재, 저대역폭 및 저지연 전송을 위한 시선 추적 기법을 360 영상 콘텐츠에 적용하기 위해 겪는 어려운 문제는 (1) 생리학적 또는 병리학적 안구 문제로 인해 사용자의 시선의 위치의 변화 및 안구의 움직임을 조절하는 반사기능이 다르다는 점이다. 따라서, 안구를 추적한 시선 정보는 불안정한 상태를 가지게 된다. 더불어, 안정적인 시선과 관심영역을 측정하기기 위하여 (2) 망막의 중심와(Fovea Centralis)에 상이 뚜렷하게 맺혀 물체에 시점을 위치하는 순간(Fixation)은 사용자 개인별로 다르고 판단하기 힘들다는 점이다. 마지막으로는 머리장착형 장치를 착용한 사용자의 회전이나 영상 내 물체의 움직임, 영상 자체가 이동 또는 회전하는 등의 문제로 해당하는 (3) 안구 운동이 도약 안구 운동(Saccade)인지 원활 추종 운동(Smooth Pursuit)인지 판단하기 어렵다는 점이 있다.

본 표준은 이의 해결책으로 ‘눈동자 움직임 판단을 위한 영상의 특성 메타데이터’의 내용과 장점, 그리고 이를 통한 (1) 타일 기반 360 영상 전송 기법을 통한 대역폭 절감,

(2) 저지연을 위한 움직임 판단을 설명한다.

A.2 배경지식

국제 비디오 코딩 표준 단체인 The Joint Collaborative Team on Video Coding(JCTVC)의 표준 기술 중 독립적 움직임 참조를 이용한 타일 분할(Motion Constrained Tile Sets) 기법과 Extraction Information Sets (EIS) Supplemental Enhancement Information (SEI) 메시지가 있다. 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) 가능하다. 따라서, 영상 분할 및 전송하여 고해상도 영상의 전송 대역폭을 절감한다.

A.3 추가 확장적 사용

본 표준은 스케일러블 비디오와 뷰포트 및 거리정보를 통한 차별적 전송 기법을 이야기하고 있지만, 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, 마이크로소프트(MS)사의 Smooth 스트리밍(Streaming), 애플(Apple)사의 HLS (HTTP Live Streaming; HTTP 라이브 스트리밍)에 적용 가능하고 MPEG에서 표준화중인 뷰 합성 기술을 속도 향상을 위해 타일 기반으로 알고리즘을 적용할 수 있다.