FBMF Technical Report

미래방송미디어표준포럼 기술보고서

FBMF-TR-009 제정일: 2021년 11월 29일

몰입형 비디오 기술 및 표준화 동향 (기술보고서)

A Trend of Technology and Standardization for Immersive Video

(앞 표지)

FBMF

미래방송미디어표준포럼

Future Broadcast and Media Standards Forum

기술보고서 초안 검토 위원회	미래미디어분과위원회
기술보고서안 심의 위원회	운영위원회

	성명	소 속	직위	위원회 및 직위	기술보고서번호
기술보고서(과제) 제안	윤국진	ETRI	책임	미래미디어 분과장	FBMF-TR-009
	위정욱	KETI	책임	미래미디어분과 간사	FBMF-TR-009
	김현식	KETI	책임	미래미디어분과 위원	FBMF-TR-009
	박민규	KETI	선임	미래미디어분과 위원	FBMF-TR-009
	김제우	KETI	센터장	운영위원회 간사	FBMF-TR-009
기술보고서 초안 작성자	윤국진	ETRI	책임	미래미디어 분과장	FBMF-TR-009
	정준영	ETRI	연구원	미래미디어분과 위원	FBMF-TR-009
	정준영	ETRI	책임	미래미디어분과 위원	FBMF-TR-009
	김준수	ETRI	선임	미래미디어분과 위원	FBMF-TR-009
	곽상운	ETRI	선임	미래미디어분과 위원	FBMF-TR-009
사무국 담당	김제우	KETI	센터장	운영위 간사	FBMF-TR-009

본 표준 발간 이전에 접수된 지식재산권 확약서 정보는 본 표준의 '부록(지식재산권 확약서 정보)'에 명시하고 있으며, 이후 접수된 지식재산권 확약서는 미래방송미디어표준포럼 웹사이트에서 확인할 수 있습니다.
본 표준과 관련하여 접수된 확약서 외의 지식재산권이 존재할 수 있습니다.

발행인 : 미래방송미디어표준포럼 의장
발행처 : 미래방송미디어표준포럼
06130, 서울특별시 강남구 테헤란로 7길 22 신관 1108호
Tel : 02-568-3556, Fax : 02-568-3557
발행일 : 2021.11

서 문

1 기술보고서의 목적

기술보고서의 1차적인 목적은, 최근 전세계적으로 관심이 높아지고 있는 몰입형 미디어 서비스를 위하여 요구되는 주요 비디오 중심의 요소기술과 관련 주요 산업체 및 표준화 동향을 분석함으로써, 관련 산업 종사자 및 표준화 담당자가 전반적인 사항을 쉽고 빠르게 파악할 수 있도록 하는 데 있다.

또한 기술보고서의 2차적인 목적은, 몰입형 비디오 관련 기술이 국내 표준화 추진 가능성 및 추가 표준화 아이템을 분석하기 위한 기초 자료를 제공하는 데 있다.

2 주요 내용 요약

기술보고서의 5장에서는 몰입형 미디어 서비스 동향을 소개한다. 5.1절에서는 몰입형 미디어의 발전 방향인 메타버스 서비스에 관해 설명하고, 5.2절에서는 메타버스 기반의 몰입형 미디어 서비스 동향에 관해 기술한다.

기술보고서의 6장에서는 몰입형 비디오 기술 개발 동향을 소개한다. 6.1절에서는 몰입형 미디어 서비스에서 요구되는 AR/VR/XR 기술 개발 동향에 대해 설명하며, 6.2절에서는 LF 영상 기술 동향에 대해 살펴본다. 6.3절에서는 몰입형 콘텐츠 획득을 위한 Volumetric 콘텐츠 기술 개발 동향을 요약한다.

기술보고서의 7장에서는 몰입형 비디오 표준화 동향에 관해 기술한다. 7.1절에서는 현재 진행중인 MPEG-I Visual 표준화 동향 및 전망에 대해 살펴보며, 7.2절에서는 ETSI ISG ARF 표준화 동향 및 전망을 4개의 워크아이템 중심으로 설명한다.

3 인용 기술보고서와의 비교

3.1 인용 기술보고서와의 관련성

해당사항 없음

3.2 인용 표준과 본 기술보고서의 비교표

해당사항 없음

Preface

1 Purpose

The main purpose of this technical report (TR) is to summarize the base technologies and related standardization regarding immersive video like Metaverse services, and to provide related issue and considerations for immersive media industry.

An additional purpose of this TR is to provide a base material for analyzing the technical availability for a domestic immersive video standard.

2 Summary

Chapter 5 of this TR introduces the trend for immersive media services. Section 5.1 introduces the Metaverse service which is the final evolution form for immersive media, and section 5.2 introduces trends of the Metaverse based immersive media service.

Chapter 6 introduces trends of development for the immersive video technologies. Section 6.1 and 6.2 describe trends for development of AR/VR/XR technology and LF image that can be used in immersive media services, respectively. Section 6.3 briefly introduces development of Volumetric contents.

Chapter 7 introduces the standardization of the MPEG-I Visual and ETSI ISG ARF. Section 7.1 explains the current MPEG-I Visual standardization activity and prospect. Section 7.2 describes the ETSI ISG ARF standardization trends and prospects, focusing on 4 work items.

3 Relationship to Reference Standards

N/A

목 차

몰입형 비디오 기술 및 표준화 동향
(Technology and Standardization of Immersive Video)

1 적용 범위

기술보고서는 몰입형 미디어 관련 주요 산업체의 기술 개발 동향 및 표준화 기구의 동향을 종합하여 요약/정리하고 주요 이슈 및 고려사항을 제시함으로써, 관련 산업 종사자 및 표준화 담당자가 전반적인 사항을 쉽고 빠르게 파악할 수 있도록 한다.

2 인용 표준

해당사항 없음

3 용어 정의

3DoF
3차원 좌표계에서 원점을 중심으로 사용자 머리 회전(rotation)인 Ywa, Pitch, Roll이 허용된 움직임을 3DoF 혹은 3 자유도라 함

3DoF+
3차원 좌표계에서 원점을 중심으로 사용자 머리 회전이 허용된 3DoF에 더해 제한적인 범위 내에서 사용자의 상하/좌우/앞뒤 움직임 허용된 것을 3DoF+라 함

6DoF
3차원 좌표계에서 원점을 중심으로 사용자 머리 회전(rotation)인 Yaw, Pitch, Roll이 허용된 것을 3DoF(3 자유도)라 하며, 3 자유도에 더하여 사용자의 상하/좌우/앞뒤 움직임까지 광범위한 범위에서 허용된 것을 6DoF(6 자유도)라 함

메타버스
메타버스란 가상·추상을 의미하는 메타(Meta)와 현실 세계를 의미하는 유니버스(Universe)의 합성어로 현실을 초월한 디지털 세상을 의미함

복셀(Voxel)
볼륨을 구성하는 가장 작은 단위

암시적 볼륨 공간

객체가 존재하는 3차원 공간을 복셀의 집합으로 표현하는 방법으로, 객체 안쪽에 있는 복셀은 음수 값을 바깥에 있는 복셀은 양수 값을 가지도록 구성함. 여기서 복셀의 값이 0이 되는 표면(surface)이 객체가 갖는 표면 정보가 되며, 이 때문에 암시적 이라는 용어가 사용됨

플랜옵틱

플랜옵틱 기술은 완전한을 의미하는 라틴어(Plenus)와 광학(Optic)의 합성어로, 빛 정보를 고차원으로 획득한 후 연산을 통해 사람이 인식할 수 있는 다양한 입체 영상을 만드는 기술

HMD(Head mounted Display)

사용자의 머리에 장착한 디스플레이 장치를 통해 영상을 표시하는 장치로서, 자이로 센서를 함께 탑재하여 비행기 조종 시뮬레이션 등에 주로 사용되었으며, 360VR 영상의 시청 및 VR 게임을 위한 필수 장비임

4 약어

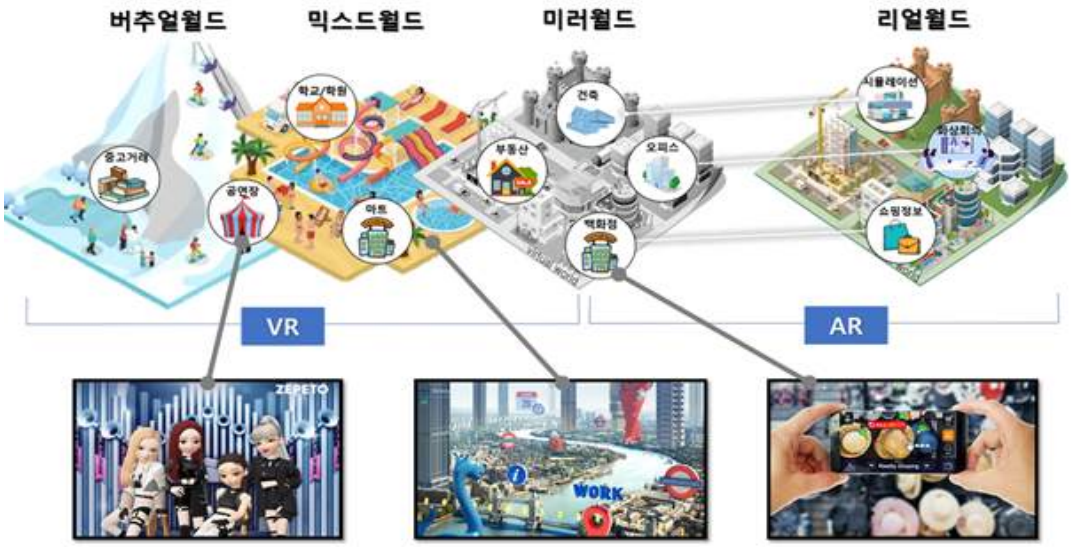
AI	Artificial Intelligent
API	Application Programming Interface
APS-C	Advanced Photo System type-C
AR	Augmented Reality
ARAF	Augmented Reality Application Format
ARF	Augmented Reality Framework
ARML	Augmented Reality Markup Language
AWS	Amazon Web Service
CD	Committee Draft
CDN	Content Delivery Network
CE	Core Experiment
CfP	Call for Proposal
CfTM	Call for Test Materials
CG	Computer Graphics
CTC	Common Test Conditions
DASH	Dynamic Adaptive Streaming over HTTP
DERS	Depth Estimation Reference Software
DoF	Degrees of Freedom
ERP	Equi-Rectangular Projection
ETSI	European Telecommunications Standards Institute
FDIS	Final Draft International Stand

FPS	Frame per Second
G-PCC	Geometry-based Point Cloud Compression
GNSS	Global Navigation Satellite System
GoP	Group of Picture
GPS	Global Positioning System
GPU	Graphics Processing Unit
GTD	Ground Truth Depthmap
HEVC	High Efficiency Video Coding
HMD	Head Mounted Display
HTTP	Hyper Text Transport Protocol
IDEA	Immersive Digital Experience Alliance
IP	Internet Protocol
ISG	Industry Specification Group
IVDE	Immersive Video Depth Estimation
IV-PSNR	Immersive PSNR
JSON	JavaScript Object Notation
JVET	Joint Video Exploration Team
LF	Light Field
MIV	MPEG Immersive Video
MPEG	Moving Picture Experts Group
MPI	Multi-Plane Image
MR	Mixed Reality
MSI	Multi-Sphere Image
NFT	Non-Fungible Token
OGC	Open Geospatial Consortium
PSNR	Peak Signal-to-Noise Ratio
QP	Quantization Parameter
TCP	Transmission Control Protocol
TMIV	Test Model of MIV
UDP	User Datagram Protocol
UHD	Ultra High Definition
V3C	Volumetric Video-based Coding
V-PCC	Video-based Point Cloud Compression
VR	Virtual Reality
VVC	Versatile Video Coding
WD	Working Draft
WS-PSNR	Weighted to Spherically uniform PSNR
XR	Extended Reality

5 몰입형 미디어 서비스 동향

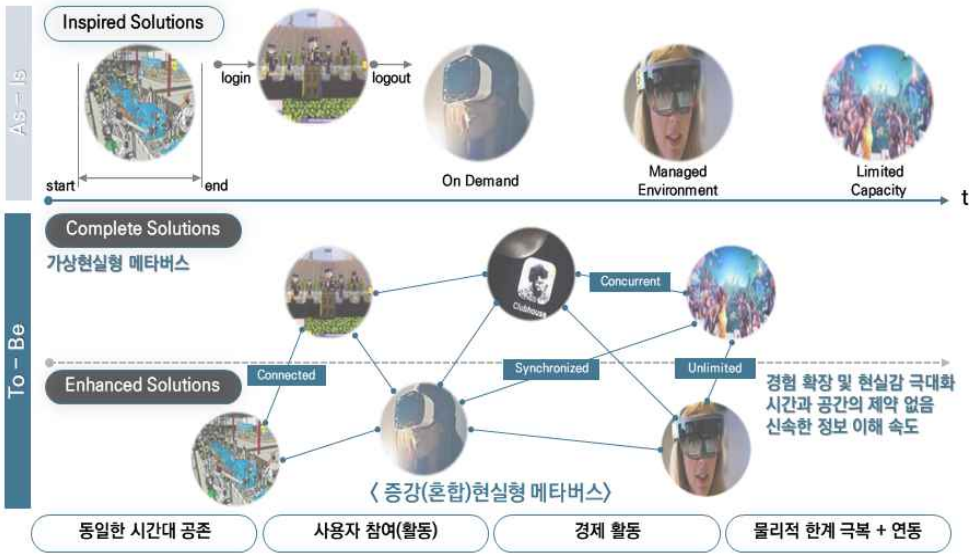
5.1 몰입형 미디어 서비스 발전 방향

코로나19로 인해 비대면 서비스에 대한 수요가 확대됨에 따라 오프라인 서비스들의 온라인 전환과 전 산업에 걸쳐 플랫폼을 통한 온라인 소비문화가 빠르게 확산되었다. 최근에는 온라인 공간에서 오프라인과 같은 유사한 경험을 하고자 하는 사람들의 수요가 증가하면서 미디어 서비스도 지금까지 우리가 경험했던 전통적인 형태의 서비스 및 소비 행태에서 탈피하여, 급변하는 환경에 발맞추어 끊임없이 발전하는 기술의 변화를 통해 기존의 시공간의 한계를 적극적으로 극복하는 동시에 보다 능동적인 사용자 참여가 가능한 서비스로 발전하고 있다. 특히, 대용량 멀티미디어 데이터 전송 기술과 초고속, 초저지연, 초연결의 특성을 가지는 5G 이동통신 서비스를 중심으로 시각적 정보를 표현하고 처리하는 영상 처리 및 표현 기술이 함께 발전함에 따라 가상현실 및 증강현실 기술 등이 적용된 실감 콘텐츠의 개발과 높은 몰입감(immersive), 상호작용(interactive), 지능화(intelligence)를 통해 기존 미디어 서비스의 한계를 적극적으로 극복하여 높은 현실감과 사용자 경험의 영역을 극단적으로 확장시키는 몰입형 서비스에 대한 수요는 더욱 증대되고 있다. 이런 상황 속에서 최근에는 XR 기술의 재부상과 그 경제적 잠재력에 대한 긍정적 전망, 코로나19가 가속화시킨 새로운 디지털 사회와 문화에 대한 욕구가 결합하며 가상과 현실세계가 상호작용하며 그 속에서 사회・경제・문화 활동이 이루어지며 가치가 생산되는 메타버스(Metaverse)에 대한 관심이 크게 부각되고 있다. 메타버스란 가상・추상을 의미하는 메타(Meta)와 현실 세계를 의미하는 유니버스(Universe)의 합성어로, 현실을 초월한 디지털 세상을 의미하는 것으로, 가상으로 창작되나 개개인이 아바타로 참여하여 현실에서와 같은 사회・경제 활동을 영위하거나, 현실에는 존재하지 않는 가상의 객체를 현실에 동기화하여 부가가치와 기능을 향상시키며 동일한 시간대에 공존하고 동기화되어 현실 세계의 한계를 극복하고 그 속에서 새로운 가치를 창출하는 세상이다.



(그림 5-1) 몰입형 메타버스 서비스 개념도

메타버스 서비스는 '92년에 처음으로 등장한 '메타버스(가상세계)' 개념이 사회・문화・기술・경제적 요인에 따라 '디지털 전환'을 주도하는 핵심 범용기술로 주목받기 시작하여 최근에는 XR・홀로그램, 블록체인・NFT(non-fungible token) 등을 활용하여 기존과 완벽히 다른 사회・경제적 경험을 제공함과 동시에 이용자 스스로 자신의 세계관을 창조하는 환경까지 구현하여 앱을 넘어 현실-가상세계를 아우르는 시・공간의 제약이 없는 몰입형 서비스로 발전하고 있다.



(그림 5-2) 메타버스 기술의 발전 전망

메타버스 서비스는 게임, 엔터테인먼트・분야에서 시장에 진입하였으나 게임과 엔터테인먼트를 넘어 융합의 영역으로 확장되고 있다. BTS의 다이너마이트도, 블랙핑크의 팬사인회도 메타버스 플랫폼에서 개최되었으며, 호두랩스는 여행과 어학교육을 융합한 메타버스를 선보이고 있다. SK와 롯데홈쇼핑은 채용설명회를 메타버스 플랫폼으로 진행하였다. 하루가 멀다하고 쏟아져 나오는 수많은 사례가 메타버스의 무한한 융합의 잠재력을 보여주고 있다.

<표 5-1> 메타버스 속성

메타버스 속성	내용
멀티버스 연결	현실과 가상이 분리되고 서로 다른 메타버스가 연결되어 데이터와 에셋을 사용할 수 있는 개방성 제공
자율적 성장	플랫폼 기업에 지배·종속되지 않고 생산적 참여자에 의해 영역이 확장되고 고도화되는 자율적 성장 생태계 추구
경제 생태계	현실세계와 같이 신원을 확인하고 디지털 저작물·서비스를 소유·저작·소비할 수 있는 경제활동 가능
크로스 플랫폼	사용자 이용환경(PC·모바일·디바이스) 제약을 최소화하여 다양한 세대가 참여하는 디지털 격차 최소화

<Entertainment> <Collaboration>

<Community> <Culture/ART>

(그림 5-3) 메타버스 서비스 분야

최근에는 스마트폰, 헤드마운트 디스플레이(XR 헤드셋) 등 '메타버스' 접속을 위한 하드웨어 인터페이스의 경량화(글래스·렌즈), 독립화(스마트폰 등 없이 작동)로 더욱 빠르고 간편한 메타버스(가상세계) 접속, 안정적인 연결 및 사용자 간 원활한 소통 환경 제공과 더불어 가상화폐, 블록체인 등 메타버스에서 유통 가능한 금융 수단과 접목되어 제품(아이템)의 현실과의 연동(가상-현실 환전 등)이 점차 가능해짐에 따라 메타버스 서비스는 전 산업과 사회 분야로 확산·적용되며 더욱 급성장할 것으로 전망된다. 이처럼 메타버스 서비스의 발전 변화의 폭과 깊이는 매우 크고 향후, 메타버스 안에서 보내는 시간이 더욱 증가함에 따라 메타버스의 영향력이 게임, 생활·소통 등 B2C 분야를 넘어 B2B, B2G 등 경제 전반으로 영향력이 확대될 전망이다.

(그림 5-4) 메타버스 서비스 혁신

5.2. 메타버스 서비스 동향

5.2.1 메타버스 서비스 플랫폼

■ 제페토

제페토제트에서 출시한 AR Virtual 플랫폼인 제페토(‘18)는 카메라 애플리케이션인 스노우에서 파생된 서비스로서, AI・AR・3D 기술을 통해서 사용자 맞춤형 3D 아바타를 생성하고, 가상 공간인 ‘월드’에서 새로운 체험(월드에서의 커뮤니티 활동, 게임, 이벤트 참여 등)하면서 가상 세계만의 새로운 생태계를 구축하고 있다. 국내 서비스 중 가장 많은 2억 명의 글로벌 이용자를 확보하였으며, 10대가 80% 이상이며 블랙핑크의 ‘블핑하우스’는 팬들의 명소로 알려져있다. 명품브랜드 구찌의 신상품을 볼 수 있는 가상 공간인 ‘구찌 빌라’가 제페토에 설치된 바 있으며, 나이키와 협업한 운동화 아이템은 500만 개가 팔렸다.

(그림 5-5) 제페토제트의 제페토

■ 로블록스

미국 로블록스사에서 2006년에 출시한 ‘사용자가 게임을 프로그래밍하고, 다른 사용자가 만든 게임을 즐길 수 있는 온라인 게임 소셜 플랫폼 및 게임 제작 시스템’인 로블록스는 사용자는 자체 게임 엔진인 "로블록스 스튜디오"(Roblox Studio)를 사용하여 자신만의 게임을 만들고 자신이 만든 게임을 다른 사용자가 플레이할 수 있으며, 게임 프로그래밍 언어 루아 5.1.4를 사용하는 객체 지향 프로그래밍을 사용하고 있다. 월간 이용자 수는 약 1억 5천 만명, 월별 누적 이용시간은 30억 시간 이상으로 미국의 16세 미만 어린이의 절반 이상의 유저층을 형성하고 있으며, 주요 비즈니스 모델은 로벅스로 1로벅스가 0.01 달러에 거래되고 있다.

(그림 5-6) 로블록스사의 로블록스

■ 마인크래프트(Minecraft)

2014년 MS사에 인수된 모장(Mojang) 스튜디오에서 2011년 출시한 마인크래프트('11)는 사용자가 아바타를 활용해 블록으로 구조물, 기능 등 콘텐츠를 제작할 수 있는 샌드박스 형식의 비디오 게임으로, 네모난 블록으로 이루어진 가상 세계에서 혼자, 혹은 여럿이 생존하면서 건축, 사냥, 농사, 채집, PvP, 회로 설계, 또는 직접 게임을 제작하는 등 정해진 목표 없이 자유롭게 즐길 수 있는 게임 플랫폼이다. 평균 이용자 1억 2천만 명 수준으로, 최근 가상 캠퍼스를 만들어 수업, 가상졸업식 등을 진행하였으며, IP를 확장하여 장난감, 소설, 영화, 교육용 도구 등을 출시하고 있다.

(그림 5-7) MS사의 마인크래프트

■ 포트나이트

포트나이트(Fortnite; ‘17)는 에픽게임즈에서 출시한 온라인 3인칭 비디오 서버이벌 슈팅 게임 플랫폼으로, 게임 뿐만 아니라 파티로얄모드에서는 콘서트, 영화 상영 등이 가능하고, 힙합 뮤지션 ‘트래비스 스캇의 가상 현실 콘서트’, ‘영화 테넷의 트레일러’, ‘BTS의 Dynamite 뮤직 비디오 안무 버전 최초 공개’, ‘아리아나 그란데 콘서트’ 등을 진행하였으며, 3억 5천 만명 이상이 이용하고 있다.

(그림 5-8) 에픽게임즈사의 포트나이트

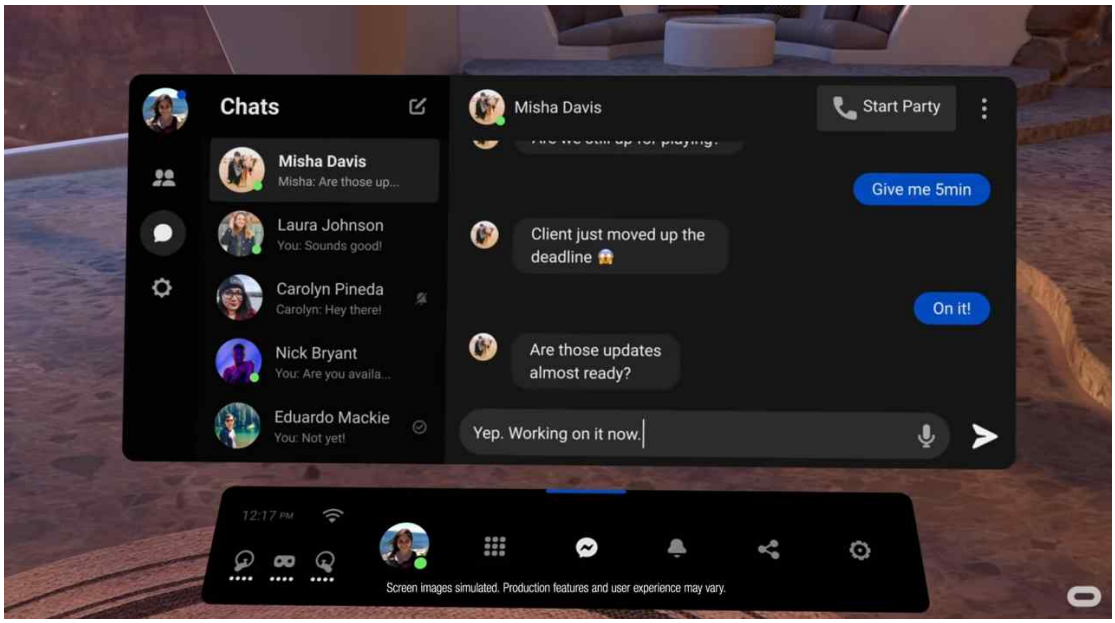
■ Facebook Horizon

Facebook Horizon(‘20 beta version)은 페이스북에서 만든 가상현실(VR) 세계 플랫폼으로, 타인과 대화하고 자신만의 공간을 만들거나 새로운 게임을 만드는 등의 다양한 활동을 VR 플랫폼에서 체험할 수 있다. 즉, 페이스북을 가상 공간으로 옮겨온 것과 같은 서비스로서, 2021년 현재 북미에서 ‘월드빌더’라는 기능을 통해 각자가 자신의 가상공간을 만들 수 있는 서비스를 베타 테스트하고 있다.

■ Facebook Infinite Office

Facebook Infinite Office(‘20)는 페이스북에서 페이스북 커넥트(Facebook Connect)에서 발표한 가상 사무실 환경 플랫폼으로, VR기기를 통해 가상현실 화면에서 사용자 정의 가능한 여러 화면을 갖춘 작업 공간이 등장하고 여기에서 작업을 하게하고, 현재 오큘러스 브라우저(Oculus Browser)를 기반으로 하고 있어 데스크톱급 웹 경험을 제공하고 있으며, 또한 전체 가상현실 뿐 아니라 실제 주위 상황도 확인할 수 있다.

(그림 5-9) Facebook의 Facebook Horizon



(그림 5-10) Facebook의 infinite Office

■ 모여봐요 동물의 숲

모여봐요 동물의 숲(Nintendo; '20)은 동물의 숲 시리즈 닌텐도 스위치용 생활 시뮬레이션 게임으로, '아무것도 없는 무인도에 초기 주민 둘과 함께 이주해 처음부터 섬을 개척해 나가는 게임'으로 마인크래프트나 심즈 시리즈와 각종 건설 경영 시뮬레이션 게임이 부럽지 않은 나만의 가상 공간을 만들 수 있다. 뿐만 아니라 네트워크를 통해서 이용자들에게 사적 모임과 음악인들이 콘서트를 열 수 있는 가상 공간을 제공하며, 유명 패션 브랜드인 마크제이콥스와 발렌티노가 자사 디자인 의상을 아바타용 의상으로 제공하고 있다.

(그림 5-11) Nintendo의 모여봐요 동물의 숲

■ 가상 컨퍼런스/이벤트

Life logging과 Virtual Worlds가 결합된 형태로 영국 Hopin, Teooh 등의 기업이 제공하는 가상 Conference/Events가 대표적으로서 가상과 현실 목표를 동일하게 설정·연동하거나 사용자들의 모든 활동이 life logging으로 연계되어 사후 성과 측정이 가능한 서비스이다.

(그림 5-12) Teooh의 Virtual Events 예

5.2.2 메타버스 디지털 휴먼

■ 에스파(Aespa)

에스파(aespa; ‘20)는 SM의 신인 걸그룹 '에스파'를 모양만을 흉내낸 게임 캐릭터와는 다르게 실제 가수가 메타버스 기반에서도 활동하는 형태의 서비스를 제공하고 있다.
무대에서 가수들과 아바타들이 함께 집단 군무를 펼치고, SNS에는 인공지능으로 무장한 아바타들이 팬들과 소통하는 형태의 서비스 제공을 계획하고 있다.

(그림 5-13) SM의 에스파(Aespa)

■ 루이(Rui)

루이(Rui; ‘21)는 디오비 스튜디오의 버츄얼 휴먼으로, 7명의 얼굴 데이터를 수집한 후 인공지능 기계학습 기술을 활용해 만든 가상 얼굴을 이용하여, 실제 촬영한 동영상(몸체)에 가상의 얼굴을 합성하는 방법으로 제작된 디지털 휴먼으로 현재 유튜버 가수로 활동 중이며, 2021년 현재 한국관광공사 명예 홍보대사로 활동하고 있다.

■ 김래아(Reah Keem)

김래아(‘21)는 LG전자의 가상 인간(디지털 휴먼)으로, 인공지능(AI) 기술을 기반으로 목소리를 입히고 움직임을 구현한 버추얼 인플루언서로, CES2021에서 LG전자의 언론 발표회에서 '김래아'가 3분간 프레젠테이션을 진행하였으며, 인스타그램 계정도 운영하고 있다. SNS를 바탕으로 패션, 뷰티 브랜드 등과 협업할 예정이다.

(그림 5-14) 디오비스튜디오의 루이 리

(그림 5-15) LG전자의 김래아

■ 삼성 Neon 디지털 아바타

삼성 Neon 디지털 아바타('20)는 물리적인 하드웨어는 없으나, 딥러닝을 기반으로 가상 공간 내에서 인간과 매우 흡사한 아바타를 구현하여 인간처럼 자연스럽게 대화를 나누고 감정과 지능을 표현 가능한 수준이다.

(그림 5-16) 삼성이 개발한 디지털 아바타 Neon

■ 로지

로지('20)는 콘텐츠 크리에이티브 그룹 싸이더스 스튜디오 엑스가 지난해 선보인 버츄얼 인플루언서로 실제 사람처럼 자연스러운 모습으로 화제를 모았다. 키, 나이, 취미, 관심사, 심지어는 MBTI 유형까지 세세하게 기획 하였으며, 인스타그램에서는 4개월 동안 일상을 공유 하였는데 아무도 로지가 가상의 인물임을 의심한 사람은 없을 정도로 실제와 유사하다. 현재는 약 5만 명의 팔로워를 얻은 SNS 스타로 제로웨이스트, 업사이클링 등 다양한 주제로 게시물을 올리고 있으며, 뷰티 등 유료 광고도 일부 게시하고 있다. 최근에는 실제 톱스타들을 밀어내고 각종 브랜드의 모델을 꿰차기 시작했으며, 전기자동차와 생명보험사, 4~5성급 호텔 등의 모델로 선정되었다.

(그림 5-17) 로지 인스타그램

6 몰입형 비디오 기술 동향

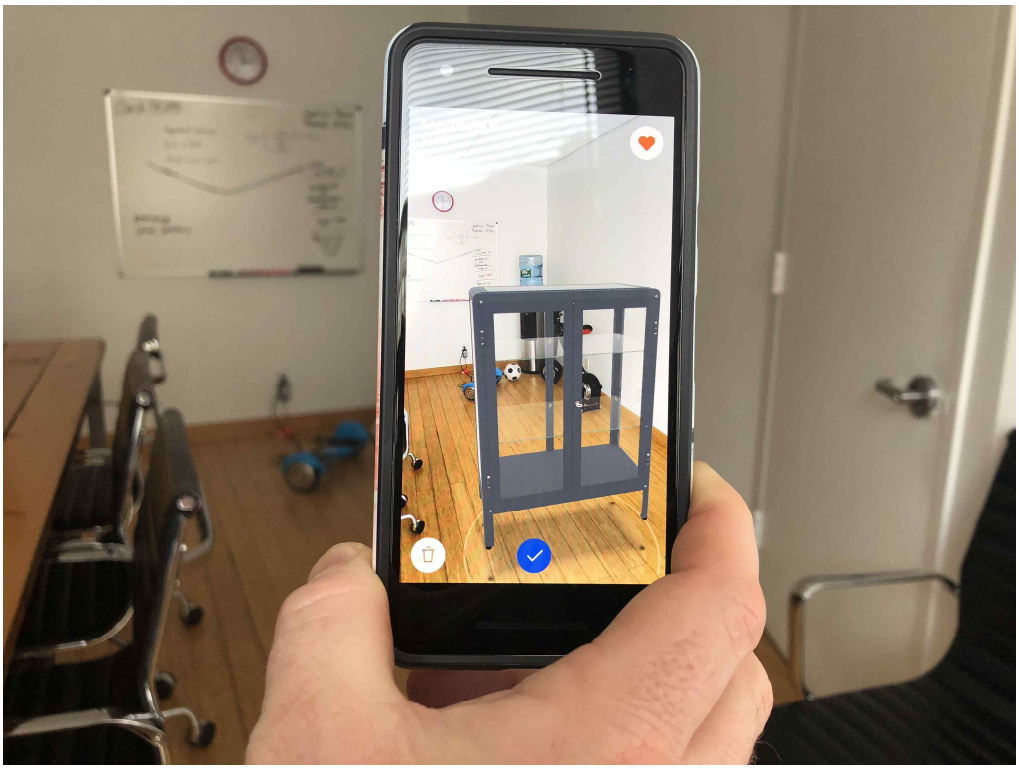
몰입형 미디어는 공간과 시간의 제약을 극복하고 극대화된 현실감 및 몰입감을 제공하는 미디어로 UHD(Ultra High Definition) 이후 차세대 미디어의 하나로 각광받고 있다. 시청자에게 자연스러운 몰입감을 제공하기 위해서는 움직임에 부합하는 영상을 보여주는 운동시차, 실제와 같은 입체감 등이 필수 요소라 할 수 있다. 본 장에서는 몰입감을 제공할 수 있는 기술인 AR/VR/XR 기술, LF 영상 기술, Volumetric 콘텐츠 기술 동향에 대해 기술한다.

6.1 AR/VR/XR 기술

6.1.1 AR/VR/XR 개발 플랫폼

■ ARCore(Google)

구글에서 개발한 전용 디바이스를 요구하지 않는 모바일 기기에서의 AR 앱 개발을 위한 SW 프레임워크인 ARCore('17)는 모션추적, 환경인식, 조명추정 등의 기능을 사용하여 스마트폰의 카메라를 통해서 가상의 콘텐츠를 실제 세계와 통합함으로써 AR를 구현할 수 있다. AR world map을 통해서 환경과 대상 객체 사이의 상호작용을 지원하면서 가상 객체의 배치, 주변 사물, 평면 등의 표면 각도 등의 기능을 제공하고 있다.



(그림 6-1) Google사의 ARCore를 활용한 AR 앱

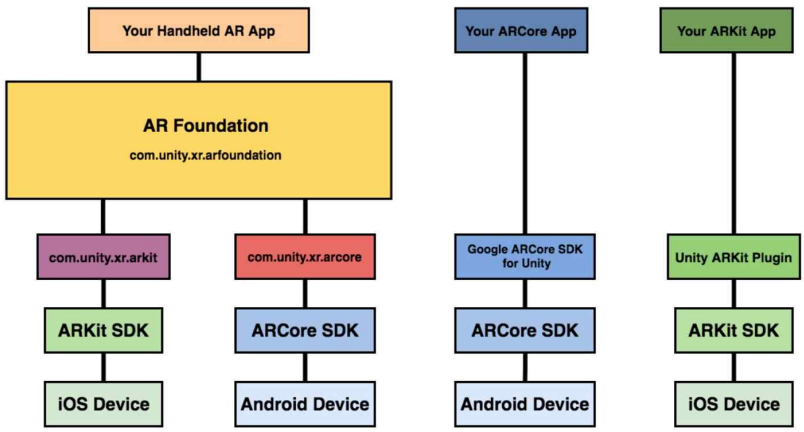
■ ARKit(Apple)

ARKit('17)은 Apple에서 출시한 AR 앱 제작을 위한 SW 프레임워크로서, iPhone 등 Apple 제품에서 AR 기술을 구현할 수 있게 한다. 구글의 ARCore와 마찬가지로 모션 추적, 환경 인식 등의 기본 기능에 Scene Geometry, human occlusion, Depth API, 얼굴 추적, 위치 앵커 등 다양한 기능들이 추가되어 강력한 AR 앱 개발 플랫폼으로 진화하고 있다.

(그림 6-2) Apple사의 ARKit을 활용한 AR 앱

■ AR Foundation(Unity)

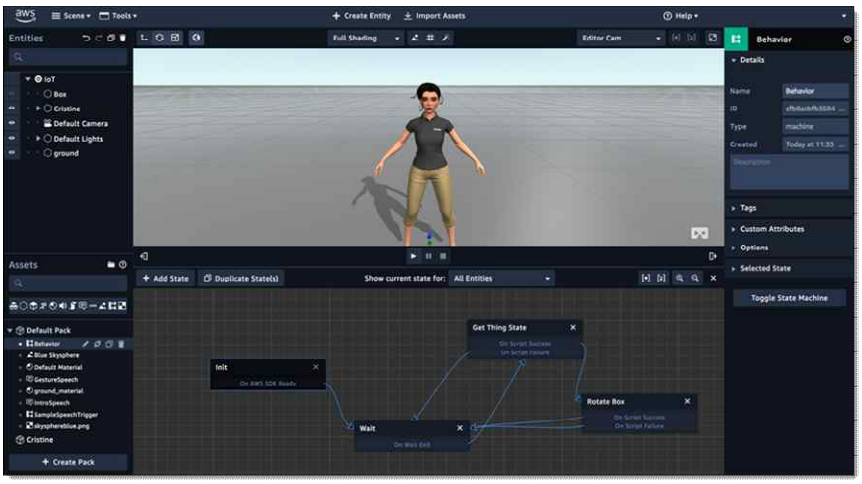
유니티에서 개발한 AR SW 개발 프레임워크인 AR Foundation('19)은 ARKit, ARCore, Magic Leap, HoloLens의 핵심 기능을 비롯하여 고유한 Unity 기능을 포함하는 하나의 통합된 워크플로우를 통해 멀티플랫폼용 AR 앱 개발을 지원하고 있다.



(그림 6-3) Unity의 AR Foundation vs. ARCore, ARKit

■ Sumerian

Sumerian('18)은 Amazon에서 개발한 AR SW 개발 플랫폼으로, 프로그래밍이나 3D 그래픽 전문 지식 없이도 VR/AR 및 3D 앱 개발이 가능한 웹브라우저 기반 툴 셋트로서, 다양한 단말을 위한 크로스 플랫폼을 지원하고 있다. 웹 기반의 편집기를 통해 시각적 스크립팅 도구를 사용하여 대상과 캐릭터(Sumerian 호스트)가 사용자 작업에 응답하는 방법을 제어하는 로직을 구현하고, 또한 Amazon Lex, Polly, AWS Lambda, AWS IoT 및 Amazon DynamoDB와 같은 AWS 서비스와 연동을 지원한다.



(그림 6-4) AWS의 Sumerian 웹 기반 편집기

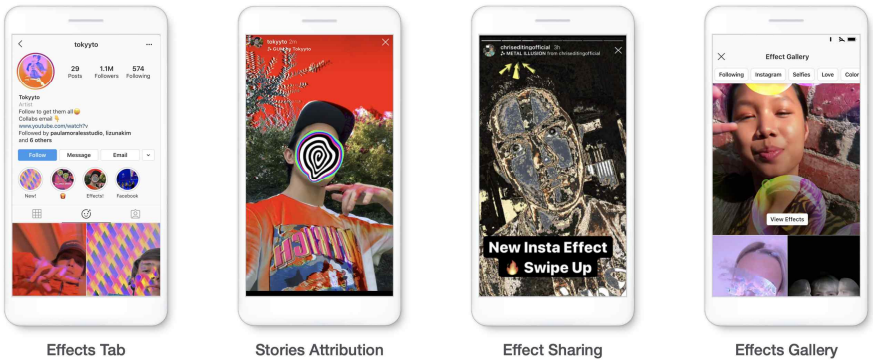
■ Real World Platform

Real World Platform('18)은 Niantic Labs에서 개발한 AW SW 개발 플랫폼으로, 컴퓨터 비전 기술을 통해 카메라의 현실 피사체의 종류, 위치, 특성을 분석하고 이 결과를 AR 콘텐츠에 적용할 수 있다. 스마트폰과 PC, 혹은 PC와 가상현실 기기 등 다양한 규격의 기기에서 동시 적용 가능한 크로스 플랫폼을 지원한다.

(그림 6-5) Niantic Labs의 리얼월드 플랫폼

■ Spark AR(Facebook)

Spark AR('19)는 페이스북에서 개발한 AR 콘텐츠 제작 전용 툴로서, 페이스북 패밀리앱(Facebook, Instagram 등)에 적용 가능한 AR 효과를 개발, 출시, 관리할 수 있는 플랫폼이다. Spark AR 홈페이지에서 AR 효과를 제작하고 공유 가능하며, 특히 인스타그램은 다양한 아티스트들이 제작한 AR 효과를 모은 "효과 갤러리"를 운영중이다.



(그림 6-6) Facebook의 Spark AR Effect 예

■ Meta Human Creator(Epic Games)

Meta Human Creator('21)은 에픽게임즈에서 개발한 클라우드 기반 스트리밍 애플리케이션으로, 게임 개발자와 실시간 3D 콘텐츠 크리에이터가 디지털 휴먼을 제작하는데 몇 주 또는 몇 개월이 소요되는 시간을 한 시간 미만으로 단축하는 동시에 최고 수준의 품질을 유지할 수 있도록 지원하는 언리얼 엔진 픽셀 스트리밍 기반의 애플리케이션이다. 폭넓은 데이터베이스에서 제작에 사용할 페이스 프리셋을 선택·설정하고 추가적으로 캐릭터를 변경하면 라이브러리의 실제 샘플이 자연스럽게 블렌딩되어 사실적인 디지털휴먼의 제작이 가능하다. 또한, Unreal 엔진의 스트랜드 기반으로 제작된 다양한 헤어스타일이나 저사양 플랫폼을 위한 헤어 카드가 제공될 뿐만 아니라, 다양한 의상과 다른 비율의 18가지 체형 선택이 가능하다.

■ Spatial(Spatial Systems)

Spatial Systems에서 출시한 AR 기반의 원격협업 지원 크로스 플랫폼인 Spatial('19)은 자체 인공지는 알고리즘을 이용해 각 사람의 2D 이미지로부터 3D 아바타를 생성하고, 사용자들 사이에서 웹 페이지, 미디어 콘텐츠, 3차원 객체, 화이트보드 등을 동일 공간에 있는 것처럼 구현하여 원격회의가 가능한 솔루션이다. 스마트폰, PC는 물론 HoloLens, Oculus, 매직리프, Nreal의 증강현실 글래스까지 다양한 기기에서 사용가능하다.

(그림 6-7) 에픽 게임즈의 Meta Human Creator

(그림 6-8) Spatial Systems의 원격 회의 플랫폼

6.1.2 AR/VR/XR 디바이스

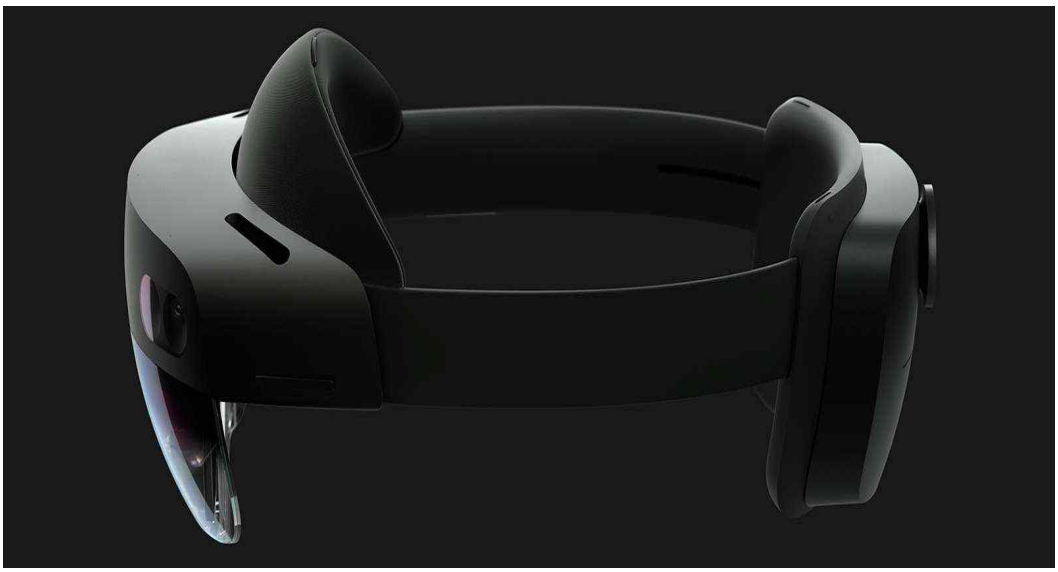
■ Oculus Quest2(Facebook)

Oculus Quest2('20)은 페이스북 자회사인 Oculus사에서 출시한 XR 디바이스로서, 자체 전용 OS 및 AP(Qualcomm Snapdragon XR2)를 탑재, 외부 PC나 콘솔없이 자체적으로 VR 콘텐츠의 실행 및 플레이가 가능한 독립형 VR 기기이다. Oculus Insight-out tracking을 통한 6DoF 움직임 추적 적용, 무선 네트워크를 통한 사용 지원 및 하드웨어 사양이 우수하며, 편의성 등이 전작보다 개선되었다. 기기 전면에 4개 카메라를 탑재하여 외부 환경 인식 및 움직임 범위 설정, 패스스루 기능을 통해 외부 상황 확인이 가능하다.

(그림 6-9) Oculus Quest2 XR Headset

■ Hololens 2(Microsoft)

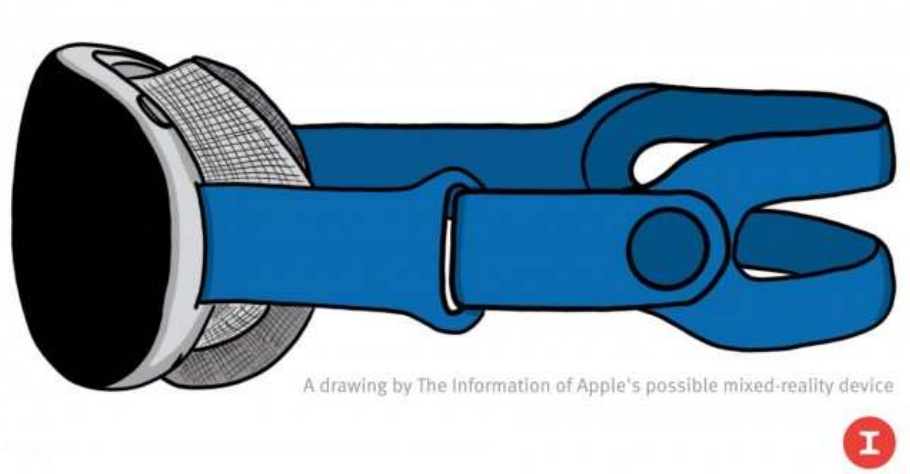
Hololens 2('20)는 마이크로스프트사에서 출시한 AR/MR 디바이스로서, 자체적으로 윈도우 10 OS를 탑재하여 스마트폰이나 PC 연결 없이 MR 콘텐츠를 제공할 수 있는 기업용 기기로서, 기존 대비 2배 이상 넓은 시야각을 보여주며, AI가 내장된 심도 센서 및 홍체인식을 통한 조작을 지원한다. 초경량 탄소섬유 소재 사용 및 무게중심 이동을 통해 착용감을 증가시켰으며, 클라우드 Azure를 통해 홀로그램으로 구현된 작업내역을 팀원들에게 공유할 수 있다.



(그림 6-10) Microsoft Hololens 2

■ AR Glass(Apple)

AR Glass('22 예정)는 애플에서 22년 중반 출시 예정인 AR/MR 디바이스로서, 스포츠 고글 형태로 예상되며, 듀얼 8K 디스플레이와 AR/VR 콘텐츠를 지원할 예정으로, 15개의 외부카메라 사용 미 "공간감 오디오 기술" 등을 적용할 것으로 예상된다.



(그림 6-11) Apple AR Glass 예상도

■ KAT Walk C(KAT VR)

KAT Walk C('20)는 KAT VR에서 출시한 VR Treadmil (VR을 사용하는 게임 등을 보다 더 생생하게 체험하기 위한 장비로서, 360도 전방향으로 걷기, 제자리 뛰기나 앉기 등과 같은 다양한 움직임을 구현) 제품으로, 보급형 VR Treadmil 제품을 지향한다. VIVE, Oculus, PIMAX, PS VR등 모든 주요 HMD 호환이 가능하다.

(그림 6-12) KAT VR의 KAT Walk C

■ Cybershoes(Cybershoes Inc)

Cybershoes('19)는 오스트리아 스타트업 사이버슈즈가 개발한 VR 신발 제품으로, 일반 슬리퍼와 동일한 형태의 VR 워킹 장비이다. 단순한 디자인의 신발 바닥과 발목에 고정시키기 위한 스트랩으로 구성되며, 신발 아랫면에 장착된 롤러를 통해 이용자의 발이 앞뒤로 움직이는 속도에 따라 게임 내 이동속도를 계산한다. 내부에 탑재된 방향추적센서를 통해 발끝이 향한 방향으로 이용자의 움직임을 재현하고, Z-스케일링 및 점프 동작 재현이 가능하다.

(그림 6-13) Cybershoes

■ Ekto One(EKTO VR)

Ekto One('20)은 피츠버그의 스타트업인 EKTO VR사에서 개발한 VR 신발 제품으로, 경량 탄소섬유로 제작되었고, 자동으로 작동하는 브레이크가 달린 스케이드와 같은 제품이다. 착용자가 움직이는 방향으로 회전하는 바닥에 내장된 회전 디스크를 통해 방향을 인식하고, 발이 바닥에 닿으면 앞으로 걸어가는 동안 바퀴가 다리를 뒤로 당겨서 전진/후진 등의 이동을 재현한다.

(그림 6-14) EKTO One

6.2 LF(Light Field) 영상 기술

완전 입체 영상을 지원할 수 있는 기술로는 Light Field 기술이 있으며, 이를 미디어에 접목하기 위한 LF 영상 기술 개발이 국내외적으로 활발히 진행되고 있다. 본 절에서는 LF 영상 서비스에 필요한 획득, 송수신 및 재현 기술의 개발 동향에 대해 살펴보고자 한다.

6.2.1 LF 영상 획득

6.2.1.1 마이크로렌즈 어레이 기반 플렌옵틱 카메라

LF를 획득하는 대표적인 장치로 마이크로렌즈 어레이 기반 플렌옵틱 카메라를 들 수 있다. 이 범주에 해당하는 카메라들은 대개 통상의 이미지 렌즈와 이미지 센서 사이에 마이크로렌즈 어레이를 설치하는 방식으로 구현되는데, 이러한 설계로 인해 같은 위치에서 출발하여 서로 다른 방향으로 진행하는 광선들이 이미지 센서 상의 서로 다른 위치에 도달하게 된다. 결과적으로 4D light field를 샘플링하여 2D 평면에 패킹한 형태의 이미지를 획득하게 되며, 이 이미지를 재구성하면 원하는 시점 위치에서 원하는 거리에 초점이 맞는 이미지를 얻어낼 수 있다.

■ Lytro light field camera

미국의 Lytro는 2012년에 일반 사용자들이 손에 들고 다닐 수 있는 정도로 작은 플렌옵틱 카메라를 출시하였고, 2014년에는 2세대 제품인 Lytro Illum이라는 제품을 출시하였다. 특히, Illum의 경우 40Mpixel 센서를 사용하여 전작의 낮은 화질 문제를 상당 부분 해결하였고 조작성과 내장 및 PC 소프트웨어 면에서도 상당히 개선되었다. 그러나 매크로 촬영을 하는 경우를 제외하면 시점이동 효과를 드라마틱하게 느끼기 힘들고, 촬영 후 초점 조절을 위해 사진 화질을 크게 희생해야 하므로, 특수한 목적이 있는 사용자에게만 어필할 수 있다는 한계가 있었다. 2018년에 Lytro는 영업을 종료하였으며, 이에 따라 현재는 상기 제품들이 생산되지 않는다.

(그림 6-15) Lytro에서 출시한 1세대/2세대 light field camera

■ Raytrix

독일의 Raytrix는 산업/연구용 3D LF 카메라를 판매하고 있는데, 일반적인 머신비전 카메라에 사용되는 마운트의 렌즈를 활용 가능하며, 8MP 센서를 사용하여 2K 정도의 유효 해상도를 얻을 수 있는 R8 모델부터 Sony의 150MP 센서를 탑재한 R150까지 필요에 따라 선택이 가능하다. 산업용 3D 계측이나 마이크로스코피가 주요 활용 용도이나, 실감형 미디어 획득에도 응용이 가능한데, 예를 들어 벨기에의 ULB에서 다수의 R8 카메라를 배열한 구조체를 활용하여 획득한 multiview lenslet 데이터셋을 MPEG에 제공한 바 있다.

6.2.1.2 6DoF 지원 VR 영상 획득 카메라

실제 공간을 가상공간에 동일한 스케일로 재현하여 수십 cm 수준의 6 자유도 움직임 (3축 회전과 3축 평행이동의 조합)을 지원하고자 하는 경우, 움직임 범위에 상응하는 규모의 멀티 카메라 구조물을 사용하는 것이 일반적이다. 이러한 카메라들은 같은 점에서 출발하여 서로 다른 카메라 방향으로 진행하는 광선을 획득한다는 의미에서 light field의 단면들을 획득하는 것으로 해석할 수 있다. 보통 하드웨어적 제약조건에 의해 angular radiance를 조밀한 간격으로 측정하지 못하기 때문에 공간의 기하정보 추정, 표면 특성에 대한 추정 등에 기반하여 측정되지 않은 light field 정보를 생성하게 된다.

■ 스테레오 전방위 파노라마 영상 획득 카메라

Insta360, Z CAM, Kandao 등은 최상급 VR 카메라 모델로써 스테레오 전방위 파노라마 영상 획득을 지원하는 카메라를 출시하였다. 제품 출시 경향을 보면, 저조도 환경이나 짧은 셔터 속도가 요구되는 상황에서도 높은 영상 품질을 얻을 수 있도록 점차적으로 센서의 크기가 증가되고 있는 추세로, 최근에는 개별 카메라의 센서 크기가 APS-C 포맷까지 증가되었다.

(그림 6-16) Lytro에서 출시한 1세대/2세대 light field camera

스테레오 전방위 파노라마 획득 카메라는 어안렌즈를 활용한 광각 카메라가 원형으로 배열된 형태이며, 모노 전방위 파노라마만을 지원하는 하위 모델에 비해 이웃 카메라와의 영상 중복이 훨씬 크도록 설계되어 있다. 비교적 가까이에 위치한 객체들도 2개 이상의 카메라에 동시에 촬영되기 때문에 입체감 있는 파노라마 영상을 생성할 수 있으며, 원본 영상을 활용하면 3차원 기하정보 추정 및 이에 기반한 6DoF 자유도 지원도 가능하다. 이 범주의 카메라들은 애초에 양안 영상 획득을 목적으로 만들어진 물건이기 때문에 양안

간격의 스케일에 맞춰 컴팩트하게 설계되어 있고, 이에 따라 휴대성이 좋은 장점이 있으나 6DoF 자유도가 지원되는 공간의 크기가 매우 협소하다는 한계를 가진다.

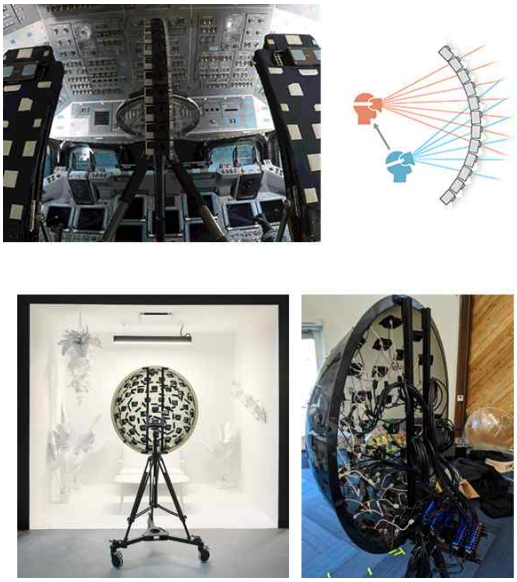
■ Lytro Immerge

미국의 Lytro는 2017년에 95개 카메라를 평면에 지그재그형의 격자 형태로 배열한 멀티 카메라 시스템과 이를 활용하여 촬영된 6DoF 자유도 지원 VR 콘텐츠인 'Hallelujah'를 공개하였다. 전체 카메라를 회전시키며 5개 방향에 대한 영상을 획득한 뒤에 기하정보 추정을 통해 복원된 3D 구조를 이어 붙여 전방위 공간을 생성하는 방식으로, 동시 촬영되지 않은 영상을 이어붙이기 위한 촬영 공간 상의 제약 또는 추가적 계산 부담이 가해지는 대신, 동일 수의 카메라를 전방위로 분산시킨 경우와 비교했을 때, 상대적으로 조밀하게 light field를 획득할 수 있다.

(그림 6-17) Lytro Immerge 카메라

■ Google light field 카메라

미국의 Google은 SIGGRAPH 2018에 전동 회전 스테이지에 장착한 카메라를 활용하여 정지된 공간에 대한 light field 획득하는 시스템 및 이를 통해 획득한 콘텐츠의 압축 및 렌더링 방법을 발표하였다. Light field 획득 시스템은 직경 70cm의 수직 방향 원호를 따라 16개의 GoPro 카메라를 배열한 뒤, 전동 스테이지로 카메라들을 회전시켜 임의의 간격으로 시점 이미지를 획득할 수 있도록 구현되었다. 위도 방향으로는 수동으로 카메라들의 수직방향 위치를 조절할 수 있는데, 카메라들의 위도를 바꾸어가며 여러 번 경도 방향으로 회전시키면 매우 조밀한 간격의 시점 이미지들의 획득이 가능하다. 개별 시점의 원본 해상도는 2704 x 2028이며, 한 공간에 대해 16 x 72개 시점 이미지를 획득하는 경우 총 6Giga 픽셀 수준의 데이터를 획득하게 된다. 많은 수의 시점 이미지를 바탕으로 매끄러운 물체에서의 반사 등을 사실적으로 재현할 수 있음을 보였으나, 동영상 촬영이 불가하다는 한계가 있었다.



(그림 6-18) Lytro Immerge 카메라

이어서 Google은 SIGGRAPH Asia 2019에 light field 동영상 촬영을 위한 카메라 구조를 발표하였다. 직경 92cm의 아크릴 반구 표면에 46개의 4K 액션 카메라를 배열한 형상으로, 1/2 프레임 이내의 동기화 오차를 가지는 다시점 동영상을 획득 가능하다. 저가의 카메라를 사용한 프로토타입 수준의 구조체이므로 비교적 적은 비용으로 구축이 가능하나, 카메라 자체의 성능한계로 인해 고품질의 원본영상을 획득할 수는 없다. 이 카메라 구조체를 활용하여 획득한 콘텐츠는 SIGGRAPH 2020에 발표된 layered mesh 기반 이머시브 light field 비디오 생성 연구에 활용되었다.

■ Facebook 6DoF VR 카메라

미국의 Facebook은 2017년에 2종의 서라운드 360 카메라를 발표하였다. 이 중 X24 모델은 24대의 FLIR사 카메라를 다면체 형태로 배열한 직경 10 인치의 구형 구조체인데, Facebook은 이 카메라로 획득한 콘텐츠를 활용한 6DoF VR 데모 영상도 같이 공개하였다. 프랑스의 Technicolor는 X24 구조를 참조하여 제작한 다시점 CG 콘텐츠를 MPEG에 제공하였으며, 이는 이머시브 미디어를 위한 메타데이터 (Metadata for Immersive Media) 표준화를 포함한 MPEG-I 활동에 활용되고 있다.

(그림 6-19) Lytro Immerge 카메라

2018년 하반기에 Facebook은 또다른 6DoF VR 카메라인 Manifold를 발표하였는데, SIGGRAPH 2019에 발표한 논문에서 해당 카메라에 대한 기술적 세부사항을 설명하였다. Facebook Manifold는 직경 100 cm의 구형 카메라 구조체로, RED Hellium 8K 센서와 슈나이더에서 커스텀 제작한 180도 화각 어안렌즈를 장착한 시네마급 카메라 16대가 사용되었다. Manifold에는 다양한 부분에 대한 최적화 설계가 이루어졌는데, 카메라간 간격을 최대한 균등하게 만드는 설계, 이웃 영상과의 중첩이 일부 방향에 집중되지 않도록 만드는 설계, 그리고 다수의 고성능 카메라에서 나오는 열을 효율적으로 방출할 수 있도록 하는 열 흐름 설계가 포함된다.

6.2.2 LF 영상 송수신

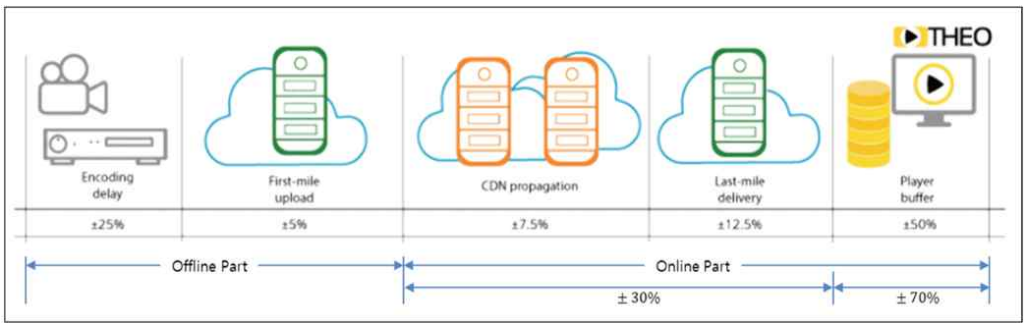
LF 영상 송수신 시스템은 네트워크를 고려하지 않은 실시간 재생 시스템과 네트워크 환경을 고려한 적응형 스트리밍 시스템 구현 방향으로 연구가 진행되고 있다. 특히, 공용 네트워크망을 통해 서비스하기 위한 연구들이 활발하게 진행되고 있다. 대표적으로 MPEG-DASH 표준에 부합하는 Media Presentation Description(MPD) 설계, 인공지능 알고리즘을 이용한 124~322Mbps 데이터 전송률 달성, 기존 대용량 동적 적응형 미디어 스트리밍 서비스에서 사용된 기법들을 적용 가능성 유무에 관한 연구가 있다.

LF 영상은 물리공간 내 여러 방향의 빛 정보를 한꺼번에 센서에서 획득하고, 이를 그대로 재현함으로써 사용자에게 초실감 경험을 제공한다. 이를 위해서는, 여러 시점에서 획득된 영상 정보를 토대로 사용자의 시청 환경 및 시점 변화에 대해 즉각적인 전환이 이루어져야 하기에 실시간에 가까운 초저지연의 전송 기술을 요구한다. 본 절은 다양한 LF 영상 서비스를 위해 적응적 저지연 전송 기술을 적용함으로써 고려해야 할 사항들에 대해 살펴본 후 기술별 개발 동향을 설명한다.

6.2.2.1 LF 영상 서비스를 위한 적응적 저지연 전송 요구사항

■ 전송 지연 요소

사용자가 불편함을 느끼지 않으면서 사용자 인터랙션에 의해 요청되는 해당 LF 영상을 실시간으로 제공해야 한다. (그림 6-20)은 LF 영상 스트리밍을 위해 요구되는 공통의 송수신구조를 나타내며 전송 지연에 영향을 주는 요소들은 다음과 같다.

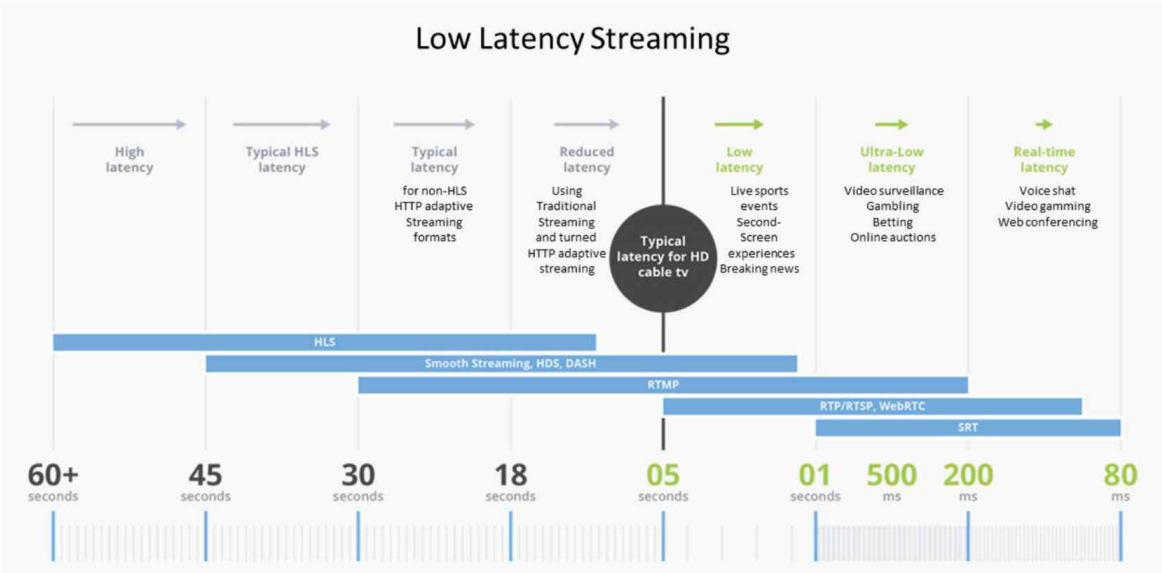


(그림 6-20) LF 영상 전송을 위한 공통 송수신 구조

- LF 영상 부호화 방법(ex. I-frame 코딩 등) 및 구조, 패키징 포맷
- LF 영상 파일 배포를 위한 CDN(Content Delivery Network)에서의 캐쉬 설정, 운용절차 및 전송 파라미터 설정
- LF 영상 전송 지연 최소화를 위한 전송 프로토콜 또는 스트리밍 매커니즘
- LF 영상 전송 스트림 버퍼링, 디코딩 및 후처리

■ 응용 서비스별 전송 지연

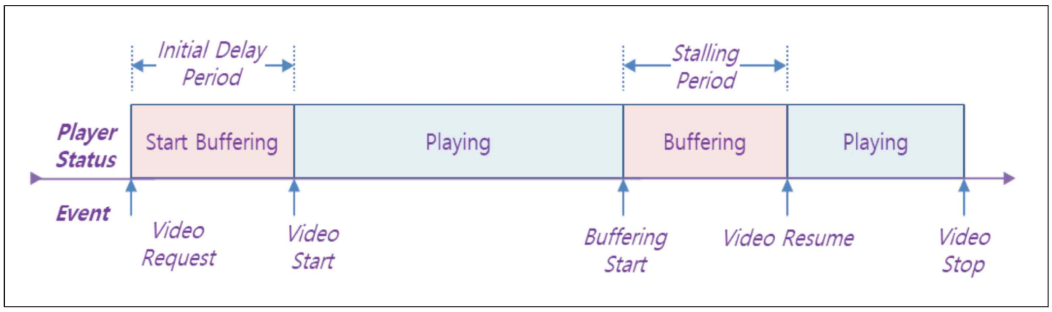
(그림 6-21)은 허용되는 전송 지연에 따른 서비스 유형을 나타낸다. 일반적인 단방향 서비스가 아닌 사용자 인터랙션이 포함된 양방향 서비스의 경우 5초~1초 사이의 저지연(low latency) 조건이 필요하며, 실시간에 가깝게 느껴지는 비디오 감시 등의 서비스를 위해서는 1초 이하(Ultra low latency), 완전한 실시간이 요구되는 화상회의 및 게임 등은 0.2초 이하의 초저지연(Real-time latency) 조건이 요구된다. 더 나아가 사용자의 인터랙션 또는 뷰포트를 반영하여 서비스하기 위한 몰입형 미디어는 수십 ms 이하로, 거의 지연이 없는 전송이 가능해야 한다.



(그림 6-21) 전송 지연에 따른 서비스 유형

■ 클라이언트에서의 지연 최소화

클라이언트에서의 지연을 증가시키는 요소는 크게 사용자가 재생을 시작한 후, 해당하는 영상이 화면에서 재현될 때까지의 초기 지연(Initial delay)과 시청 도중 인터랙션에 의한 장면 전환 또는 시청환경 등을 반영한 리버퍼링(Rebuffering)에 의한 지연으로 나눌 수 있다. 이러한 지연을 해결하기 위해서는 서버로부터 전송되는 세그먼트(최소 디코딩 지원) 및 뷰포트 이동 등에 따른 리버퍼링/디코딩 이후 프레임 간 Re-ordering 등의 최소화 과정이 요구된다.



(그림 6-22) 클라이언트에서 지연 요소

6.2.2.2 LF 영상 전송 기술 연구 개발 동향

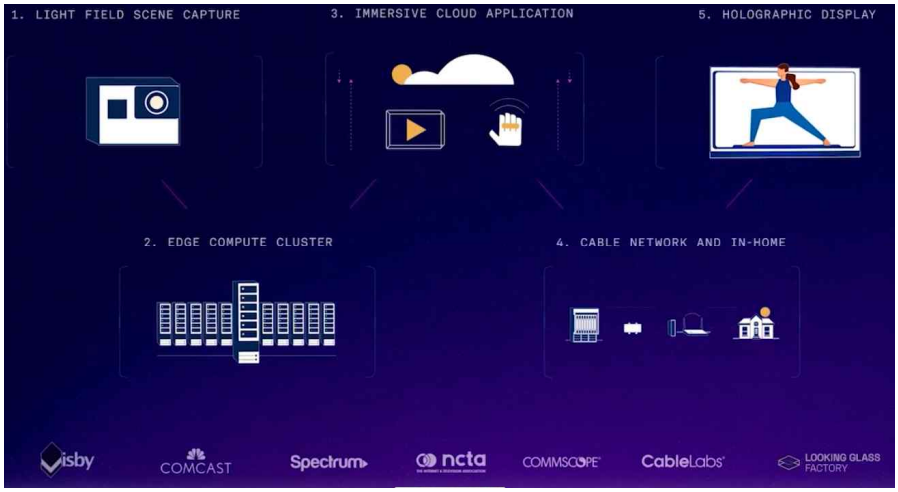
■ 10G 케이블 기반 LF 영상 스트리밍

IDEA(Immersive Digital Experiences Alliance)는 몰입형 미디어(ex. Volumetric 및 LF 영상)의 종단 간 전달을 위하여 상호 운영 가능한 인터페이스 및 저장/전송 포맷을 개발하기 위한 비영리 산업 연합으로, LF 영상을 통하여 초실감 경험을 제공하기 위한 “Streaming a LF over a 10G[1] Cable Network” 기술을 2020년 10월 SCTE 케이블 기술 박람회에서 발표하였다. 10G 케이블 네트워크 기반의 몰입형 VoD 서비스 가능성을 제시하였으며, 서버-클라이언트 기반 사용자 위치 이동 또는 인터랙션에 따른 실시간 몰입형 미디어 3차원 재현 가능성을 선보였다.

(그림 6-23) IEDA members

LF 영상은 양안시차를 이용하는 입체감뿐만 아니라 사용자의 자연스러운 시점이동에 따른 전후/좌우/앞뒤 가상시점 및 가변 초점을 제공함으로써 초실감 서비스를 제공하기 위한 하나의 미디어 형태로 인식된다. 하지만 초실감 경험을 제공하기 위해서는 대용량 데이터가 요구됨에 따라 획득-송수신-재현에 이르기까지 해결하기 위한 다양한 솔루션 개발 및 서비스 플랫폼이 요구된다. (그림 6-24)는 LF 영상 획득 시스템, LF 영상 전처리를 위한 에지 컴퓨팅 시스템, 사용자 위치 또는 제스처기반 LF 영상 클라우드 서버, 전송된 LF 영상 디코딩 및 사용자 인터랙션 정보 전송을 위한 클라이언트, 3차원 실감 영상 재현을 위한 45시점 3D 디스플레이로 구성되는 서비스 플랫폼 구조를 나타낸다.

1) 10G: 소비자 수요보다 다양한 장치 및 네트워크 성능을 발전시킴으로써 새로운 몰입형 디지털 경험과 신기술 제공이 가능한 IP 플랫폼



(그림 6-24) 10G 기반 LF 영상 송수신 플랫폼 구조

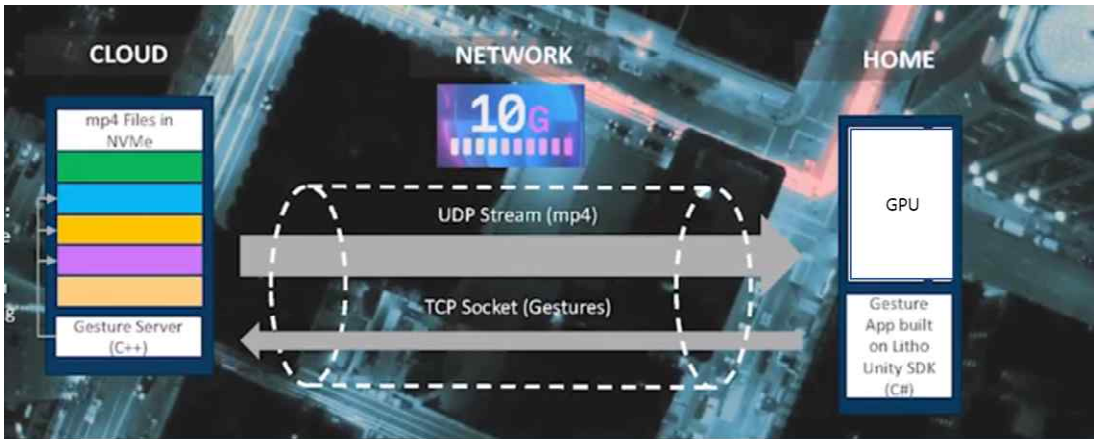
• LF 영상 획득 및 인코딩

수렴형 LF 획득 구조체를 통한 100 시점 LF 영상 획득 및 각 시점 전처리를 통하여 45 시점(5x9) 타일드(Tiled) 영상을 생성한다. 타일드 영상은 사용자의 위치 또는 제스처에 의하여 클라우드 서버에서 클라이언트에 전송되는 해당 시점을 나타낸다. 타일드 영상의 해상도는 8K(8192x8192)이며 GPU 기반 비실시간 인코딩을 수행한다.

(그림 6-25) 100시점 LF 영상 획득 예

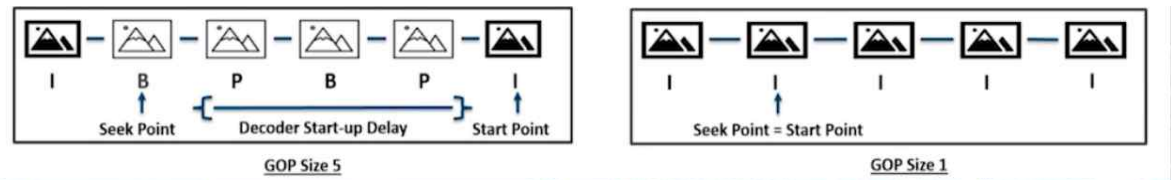
• 사용자 인터랙션에 따른 LF 영상 전송 및 스위칭 지연 최소화

인코딩된 각 타일드 스트림은 MP4파일 형식으로 클라우드 서버에 저장된다. (그림 6-26)은 사용자 인터랙션에 따른 MP4 파일 전송 구조를 나타낸 것으로, 사용자 위치 또는 제스처 정보는 TCP/IP를 통하여 서버에 전달되며, 이를 토대로 클라우드 서버는 전송 지연을 최소화하기 위하여 해당 시점을 포함하는 MP4 파일을 UDP로 전송한다.



(그림 6-26) 사용자 인터랙션에 따른 MP4파일 전송 구조

이때 MP4 파일의 전환에 따른 스위칭 지연을 최소화하기 위하여 모든 프레임을 I-frame 부호화(GoP size=1)를 수행한다. (그림 6-27)은 시점전환에 따른 MP4 파일 시작 대기 시간을 나타낸 것으로, 일반적으로 스위칭 지연시간은 I-frame 수신에 따라 달라진다. 또한, 디코딩 이후 프레임간 Buffering re-ordering 시간을 최소화함으로써 전체적으로 미디어 기반의 스위칭 지연 시간을 최소화한다.



(그림 6-27) 스위칭 지연 시간 최소화를 위한 LF 영상 부호화

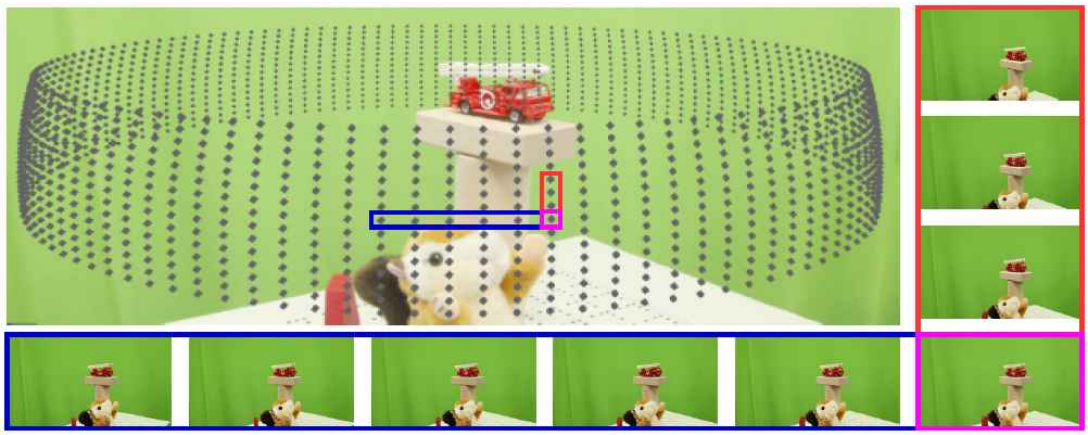
- 사용자 인터랙션 인식 및 LF 영상 3D 렌더링

사용자 제스처 인식을 위하여 블루투스기반 리소(Litho) 장치를 사용하며, 인식된 제스처 정보는 클라우드 서버에 전송한다. 전송된 해당 시점의 MP4 파일은 GPU 기반 디코딩 이후 3차원 재현을 위하여 무안경 3D 디스플레이에 특성화된 픽셀 다중화를 수행한다. (그림 6-25)는 리소 장치 및 인터랙션에 따른 3차원 재현 예를 나타낸다.

(그림 6-28) 리소 장치(좌) 및 사용자 인터랙션에 따른 3D 렌더링 예

■ MPEG-DASH를 이용한 대화형 LF 영상 스트리밍 송수신 시스템

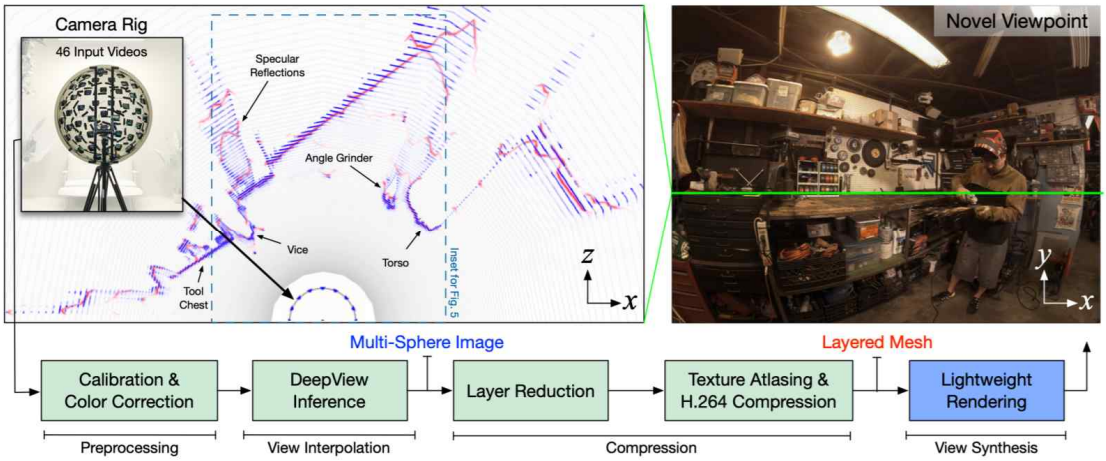
Hasselt 대학은 Virtual Reality Software and Technology (VRST)'18 학회에서 MPEG-DASH 표준을 적용한 정적(Static) LF 영상 스트리밍 서비스 솔루션을 발표하였다. LF 영상 전송을 위한 MPD를 정의하고 네트워크 스케쥴링, 세그먼트 렌더링, 비디오 메모리 캐싱 방법을 제안하였다. 4K / 5K LF 이미지를 입력으로 받아서 H.264로 압축하여 세그먼트 구성하였으며 전체 LF 이미지를 보내는 것이 아니라 렌더링을 위해 필요한 LF 이미지 부집합(sub-set)을 가용 대역폭과 사용자의 시점 위치 및 방향 조건 등에 따라 적응적으로 클라이언트에 전송하여 데이터양을 최소화하였다. 10~100Mbps 대역폭, 15msec 이하 지연시간을 가지는 네트워크 환경에서 대화형 LF 영상 스트리밍 서비스 가능성을 보였다.



(그림 6-29) 렌더링을 위해 필요한 LF 이미지 부집합 구성 예

■ 인터넷망에서 스트리밍 가능한 종단 간 6DoF 비디오 시스템

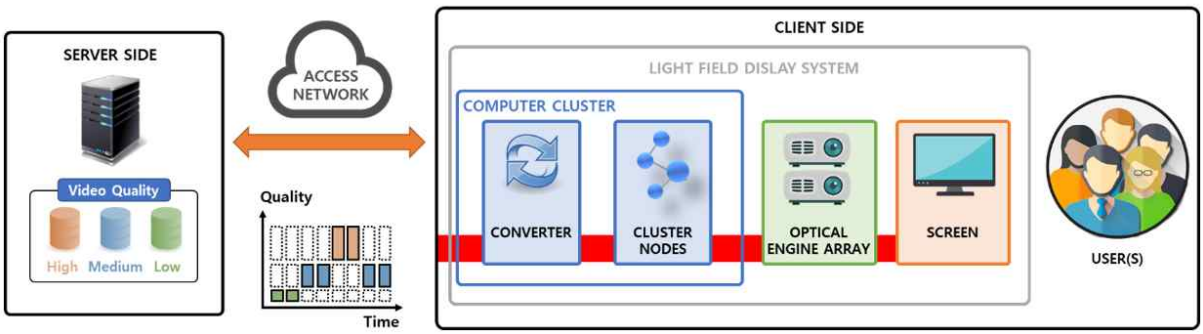
구글은 SIGGRAPH 2020에서 기계 학습 알고리즘 DeepView를 사용하여 인터넷망에서 스트리밍 가능한 종단 간 6DoF 비디오 시스템을 발표하였다. 카메라 장비에는 초당 30 프레임으로 실행되는 46개의 동기화된 4K 카메라로 구성하였으며 그 결과 머리를 움직일 수 있는 너비 70cm의 220도(해상도 1도당 10픽셀) LF 영상을 생성하였다. DeepView 알고리즘은 2D 평면을 통해 장면을 표현하는 대신 구형 쉘(shell) 컬렉션을 사용하여 훨씬 적은 수의 셀로 재처리하여 LF 이미지를 생성한다. 총 3가지 LF 영상의 (Dog, Flames, Car) 이미지와 메쉬(mesh)는 각 H.265 및 Draco geomerty 압축 라이브러리를 사용하여 압축하였으며 최종 데이터 전송률은 124~322Mbps를 달성하였다.



(그림 6-30) 구글에서 제안한 LF 비디오 시스템

■ 동적 적응형 LF 영상 스트리밍 연구

기존 멀티뷰를 포함한 대용량 동적 적응형 미디어 스트리밍 서비스를 위해 연구되었던 방식들에 대한 LF 영상 스트리밍 서비스에서의 적용 가능성 유무를 사용자 체감품질을 측정하여 판단하는 연구들이 진행 중이다. Field of View (FOV) 조절, 보간 기법 (Interpolation techniques)을 사용하여 중간 뷰 생성 기법, 사용자의 위치를 추적을 통한 Tile-based Spatial Relationship Description (SRD)에서 사용되는 알고리즘 등을 통한 데이터 전송량 저감 기법들이 LF 영상 스트리밍 서비스에도 모두 적용 가능함을 보였다. 동적 적응형 LF 영상 스트리밍 시스템에 대한 사용자 체감품질 측정 결과는 사용자가 스톨링 이벤트보다 공간 및 각 해상도 변화에 따른 품질 저하를 더 선호함을 보여준다.

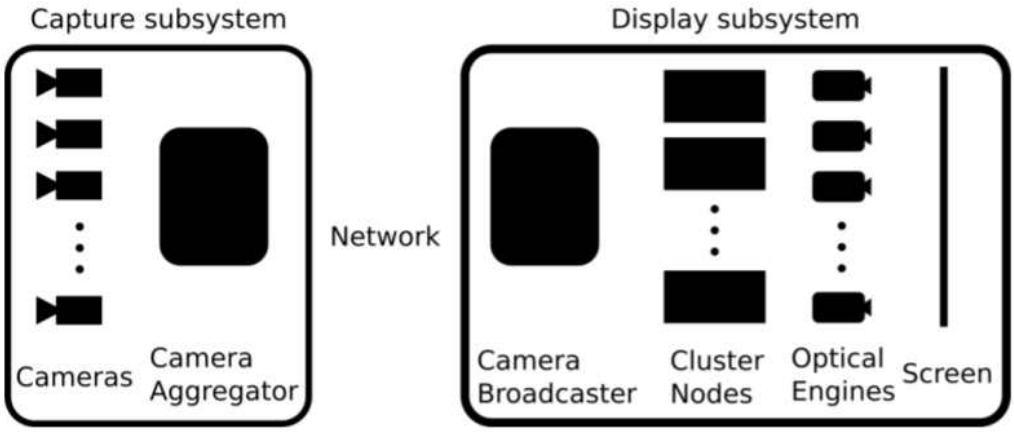


(그림 6-31) 동적 적응형 LF 비디오 시스템

■ 실시간 LF 3D 텔레프레즌스(Telepresence) 스트리밍 시스템

Holografika는 2018년 European Workshop on Visual Information Processing (EUVIP) 2018에서 카메라로부터 획득된 객체를 실제와 같은 크기로 시각화하는 단방향 실시간 LF 3D 텔레프레즌스 시스템을 발표하였다. 96개의 카메라로부터 1280x1024 픽셀 이미지 캡처 광(optical) Local Area Network(LAN)을 이용해 전송하며 높이 180cm, 넓이 100cm, FOV 180도, 25 FPS, 2D 등가 해상도 720x1280 픽셀, 각 해상도 0.9도를 가지는 디스플레이를 통해 시연하였다. 카메라 영상 획득부터 디스플레이에 LF 영상이 재생되기

까지 측정된 총 시스템 지연시간은 100msec였다. 다만, 시스템 지연시간에 네트워크 지연시간은 포함되지 않았다. 실제 서비스를 위해 LF 영상 압축 및 네트워크 지연시간 등 추가 고려사항들에 관한 연구를 진행 중이다.



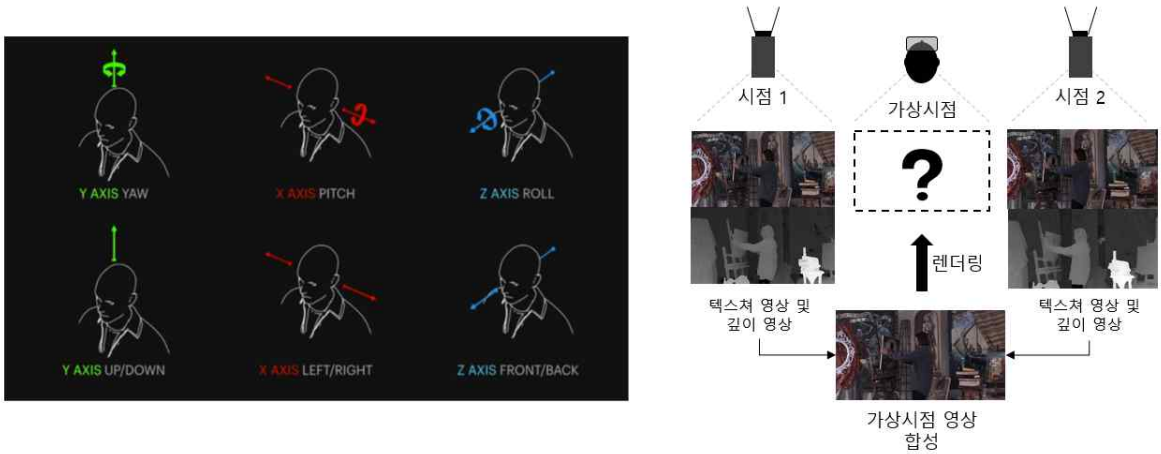
(그림 6-32) 텔레프레즌스 시스템 구성도

6.2.3 LF 영상 재현

6.2.3.1 HMD 기반 전방위 LF 영상 동적 가상시점 합성 및 재현

■ 6DoF 지원을 위한 가상시점 합성 기술

전방위 LF 영상의 몰입감 높은 재현을 위해서는 HMD 사용자의 회전 및 병진 운동에 따라 그에 맞는 양안시차와 운동시차를 제공하는 것이 중요한 요소이다. 그러나 일반적으로 LF 영상의 장면 공간 내에서 모든 시점 위치의 영상을 사전에 획득/처리/송수신하여 재현하는 것은 불가능하기 때문에, 제한적으로 획득된 유한한 수의 입력 영상으로부터 사용자 시점 위치에 해당하는 가상의 시점 영상을 합성하여 재현하는 기술이 필수적이다. (그림 6-33)은 사용자의 6DoF 움직임 및 가상시점 합성 기술의 개념을 나타낸다.



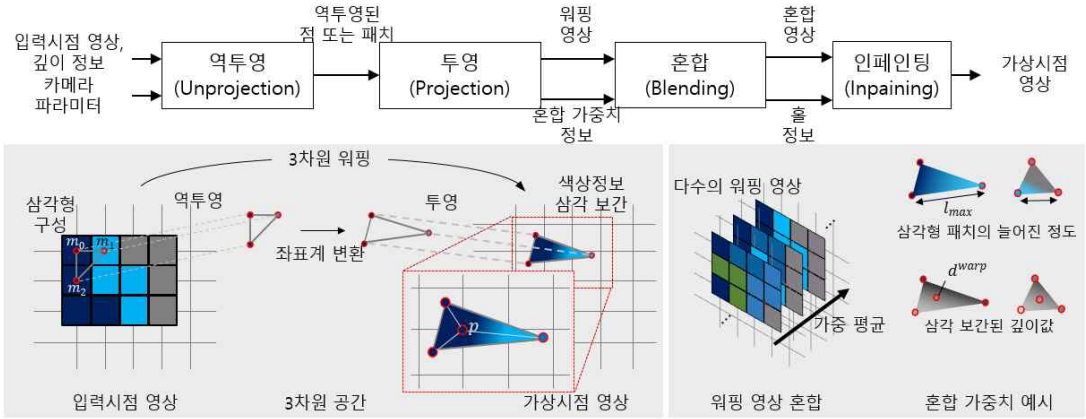
(그림 6-33) 사용자의 6DoF 움직임(좌) 및 가상시점 합성 기술(우)의 개념

■ 깊이 영상 기반 렌더링 기술

LF 영상의 가상시점 합성을 위한 대표적인 접근 방법의 하나로서 깊이 영상 기반 렌더링(depth image-based rendering, DIBR)이 있다. 입력 영상에서 제공되지 않는 임의 위치에서의 시점 영상을 합성하기 위해, 먼저 각 입력 시점 위치에서의 깊이 영상을 추정하고 이를 기반으로 각 영상을 가상시점 위치로 3차원 워핑(warping)한 후, 워핑된 영상들을 가중치를 이용하여 혼합함으로써 가상시점 영상을 구성하는 것을 기본으로 한다.

• 패치 단위의 워핑 기술

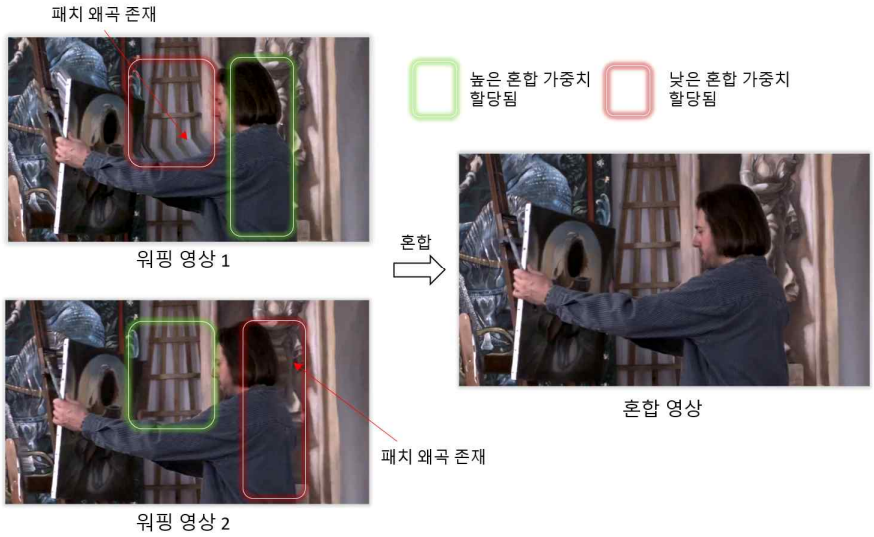
고전적인 3차원 워핑 기술은 입력 영상에서의 각 화소(pixel)단위로 이루어졌으나, 이러한 방식은 시점 변화가 큰 가상시점 영상 합성의 경우 크랙(crack)과 같은 홀을 발생시킬 수 있기 때문에, 최근에는 패치(patch) 단위의 워핑기술이 많이 사용된다. 몰입형 미디어를 위한 MPEG-I 표준화에서는 가상시점 합성을 위한 참조기술로 이러한 패치 단위워핑이 적용된 깊이 영상 기반 렌더링 기술을 사용하고 있다. 패치 단위의 워핑은 먼저 각 입력 영상에서 인접한 화소들을 이용하여 패치를 구성하고 각 패치 단위로 3차원 워핑을 수행한 후, 워핑된 패치 내부 색상을 꼭지점에 해당하는 화소의 색상을 이용하여 보간하는 방식이다. (그림 6-34)는 삼각형 패치 단위의 워핑이 적용된 가상시점 합성 알고리즘 구조의 예를 나타낸다.



(그림 6-34) 패치 단위 워핑을 사용하는 가상시점 합성 알고리즘 구조 예

• 워핑 영상 혼합 기술

워핑된 다수의 영상은 기설정된 가중치를 통해 혼합되어 한 장의 혼합 영상을 구성한다. 혼합 가중치 설정에 따라 가상시점 영상의 품질이 크게 달라질 수 있으며, 개별 워핑 영상에 존재하는 패치의 왜곡이나 아티팩트(artifact) 등이 제거될 수 있다. 대표적인 혼합 가중치 설정 방법으로는 워핑된 패치의 늘어짐(elongation) 정도를 이용하는 방식이 있다. (그림 6-35)는 패치의 늘어짐 정도를 이용하여 혼합 가중치를 할당함으로써, 혼합 과정에서 패치 왜곡을 제거하는 예를 나타낸다.



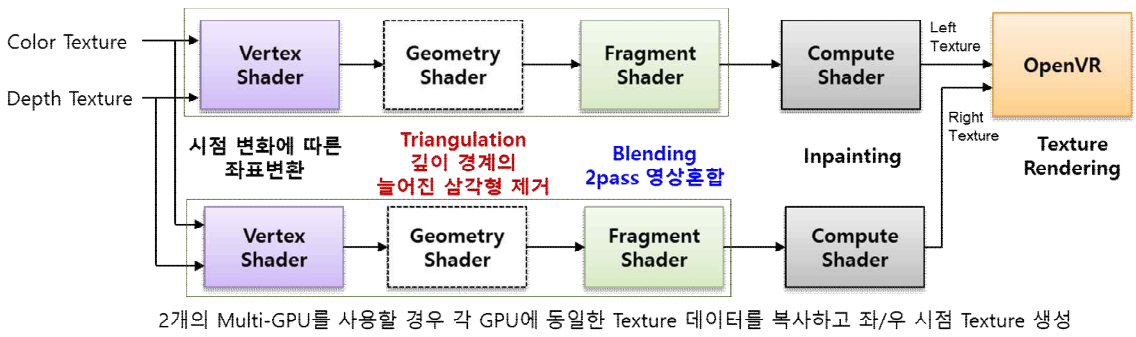
(그림 6-35) 혼합 과정을 통한 패치 왜곡 제거 예

- 인페인팅 기술

유한한 수의 LF 영상으로 장면 전체의 정보를 완전히 표현할 수 없기 때문에, 가상시점 위치에 따라 혼합 후에도 영상 정보가 없는 홀이 존재할 수 있다. 이런 홀 영역은 인페인팅(inpainting) 기법 등을 적용하여 채우게 되는데, 대표적인 인페인팅 기법으로는 홀 인접 영역의 화소 값을 그레디언트(gradient) 방향으로 확산시키는 방식 또는 영상 내에 존재하는 유사 패치를 탐색하여 예제(exemplar)로 이용하는 방식 등이 있다.

■ GPU기반 합성 및 재현 고속화

HMD 사용자의 움직임에 따른 시점 영상을 지연 없이 제공하기 위해서는 가상시점 합성 및 재현 프로세스를 GPU를 이용하여 고속화하는 것이 필수적이다. (그림 6-36)은 대표적인 GPU 기반 그래픽스 파이프라인인 OpenGL을 이용하여 가상시점 합성 및 렌더링 과정을 고속화한 예를 나타낸다. OpenGL의 쉐이더(Shader) 기능을 통해 3차원 워핑 및 혼합, 홀채움 과정이 고속화될 수 있으며, 출력되는 텍스쳐 포맷의 양안 가상시점 영상은 OpenVR 등을 통해 HMD 디스플레이에 직접 렌더링이 가능하다.

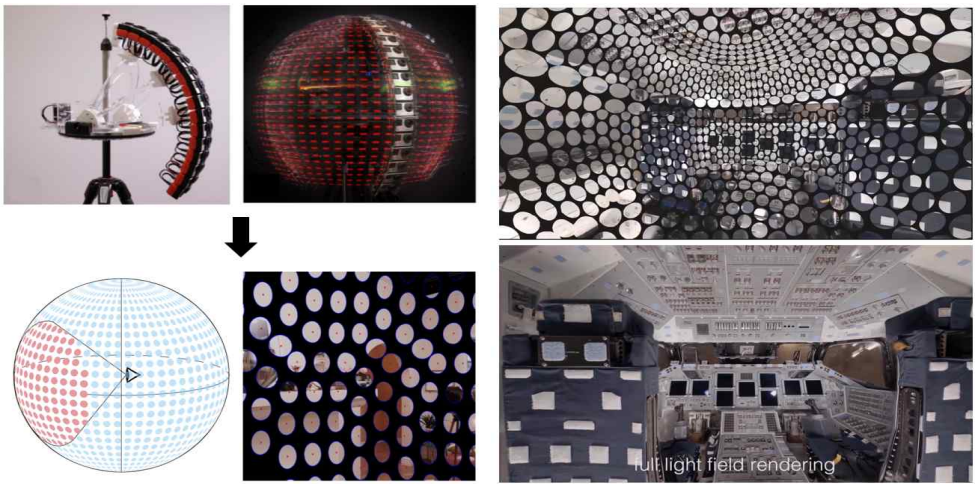


2개의 Multi-GPU를 사용할 경우 각 GPU에 동일한 Texture 데이터를 복사하고 좌/우 시점 Texture 생성

(그림 6-36) GPU 그래픽스 파이프라인 기반 가상시점 합성 및 렌더링 구조

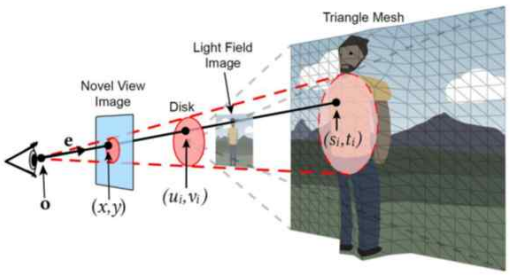
• 재현 고속화를 위한 디스크 기반 렌더링 기술

Google 사에서는 자사에서 개발 중인 LF 영상의 획득, 처리, 재현까지의 전체 파이프라인을 소개하였다. 특히 제안된 회전형 획득 구조체로부터 얻어진 전방위 LF 영상을 효과적으로 고속 재현하기 위한 디스크 기반 렌더링 기술을 제안하였다. 해당 기술은 기본적으로 (그림 6-37)과 같이 전방위에 대해 매우 조밀하게 획득된 LF 정지 영상으로부터 가상시점의 위치 및 방향에 맞는 디스크 영역만을 가져와 렌더링하는 방식으로, 직경 60cm 구 형태의 시점 공간 내에서 6DoF 지원이 가능하며, 상용 PC 수준의 HW환경에서 최대 90fps의 재현 속도를 지원한다.

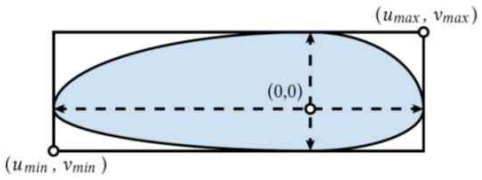


(그림 6-37) Google의 회전형 LF 영상 획득 및 재현 기술 개념

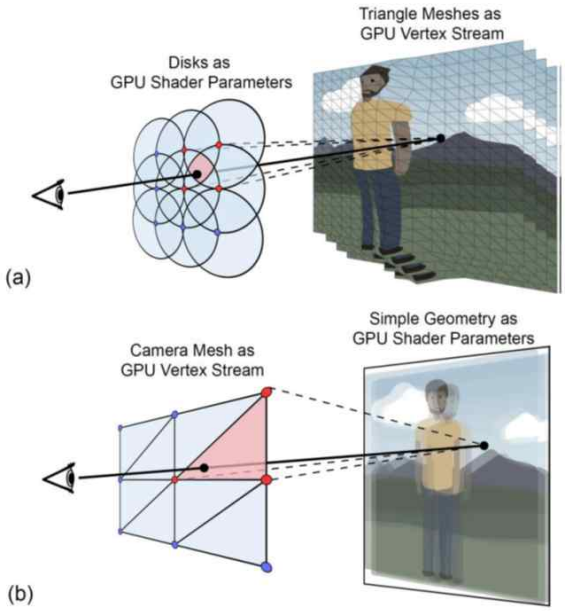
(그림 6-38)에 디스크 기반 렌더링 기술의 개념 및 기존 방식과의 비교를 나타낸다. 가상시점의 투영 중심으로부터 출발하는 광선(ray)의 방향성을 기준으로, 획득된 방향과 유사한 가상시점의 광선 묶음을 디스크 형태로 가져온다. 이때 디스크의 모양은 왼쪽 아래 그림과 같이 타원(oval)형을 제안하고 있으며 디스크 간 중첩 영역에 대해서는 디스크 중심으로부터의 거리에 따라 선형적으로 감소하는 가중치를 이용하여 혼합한다. 또한, 오른쪽 그림과 같이, 이 기법은 뷰 당 기하(per-view geometry) 방식을 사용하기 때문에, 모든 카메라 메쉬(mesh)가 장면 기하 전체를 쉐이더 파라미터화 하여야 하는 기존 방식보다 효율적인 렌더링이 가능하다.



디스크 기반 렌더링 개념



디스크의 구성



디스크 기반 렌더링(a)과 기존 방식(b)의 비교

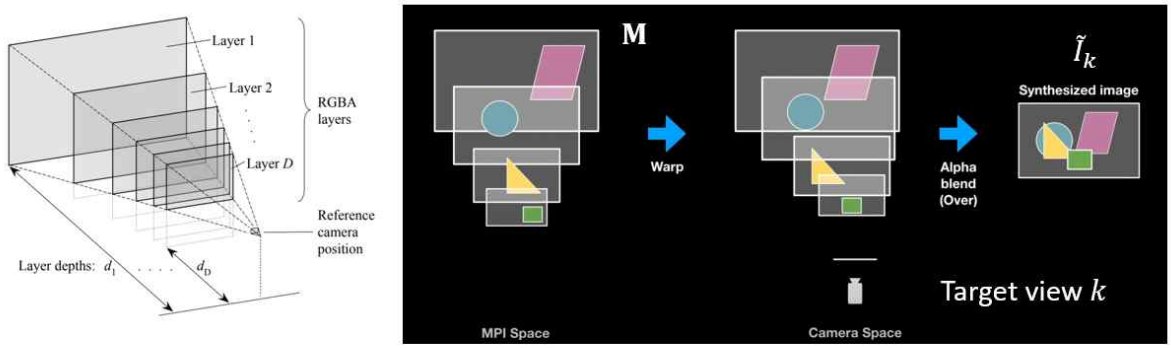
(그림 6-38) 디스크 기반 LF 렌더링 기술 및 기존 방식과의 비교

6.2.3.2 딥러닝 기반 다중 평면 영상 학습 및 가상시점 재현

최근에는 딥러닝 기술을 적용하여 장면 정보를 학습하고, LF 영상 재현에 적합한 형태로 표현하고자 하는 연구도 활발히 진행되고 있다. 이러한 시도들은 장면 내 존재하는 객체의 분포를 기반으로 공간 내 위치 및 방향에 따른 각 광선 별 색상 값을 예측하는 방식들이 주를 이룬다. 이러한 접근 방법 내에도 다양한 세부 기법들이 존재하는데, 최근 활발히 진행되고 있는 다중 평면 영상(Multi-Plane Image, MPI) 기반 재현 기법을 중심으로 살펴본다.

■ 다중 평면 영상 기반 재현 기술

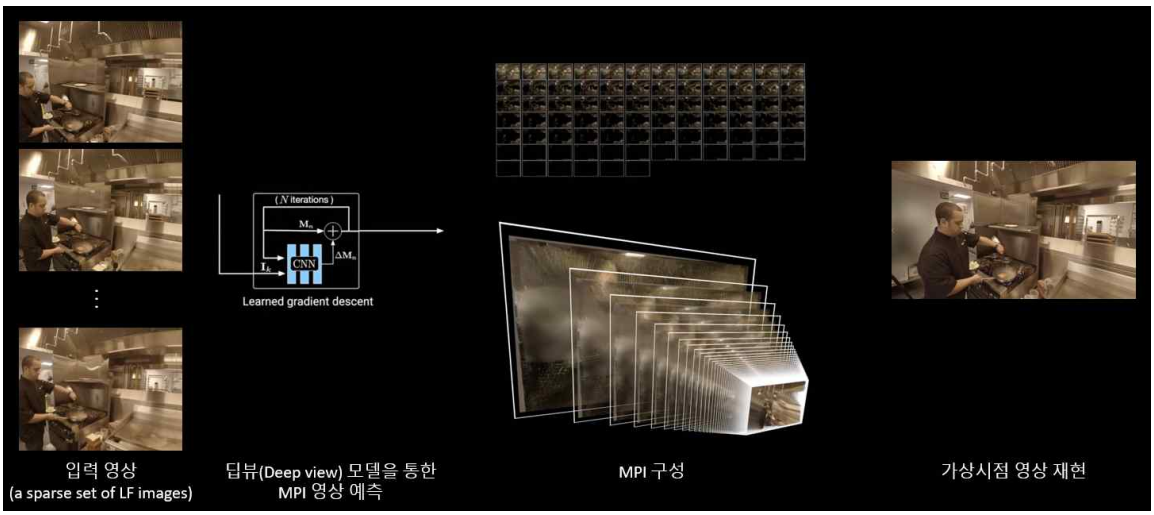
Google 사에서는 2019년 딥러닝을 활용한 MPI 기반 재현 기술을 발표하였다. MPI란 LF 영상을 효과적으로 표현하기 위해 전체 장면 공간을 유한한 수의 계층형 평면 영상에 투영시켜 나타낸 것을 의미한다. (그림 6-39)는 MPI의 개념 및 MPI 기반 가상시점 영상 합성 과정을 나타낸다. 계층 형태의 각 평면 영상은 장면으로부터 얻어지는 색상 값과 불투명도 값으로 이루어진 RGBA 값을 가진다. 가상시점의 위치가 결정되면 해당 가상시점 영상좌표계로 MPI 영상들을 모두 워핑 및 투영시킨 후, 투영된 영상들을 각각의 불투명도를 가중치로 이용하여 혼합하는 알파 블렌드(alpha blend)함으로써 합성 영상을 생성한다.



MPI의 개념 MPI 기반 가상시점 영상 합성 과정

(그림 6-39) MPI의 개념 및 MPI기반 가상시점 영상 합성 과정

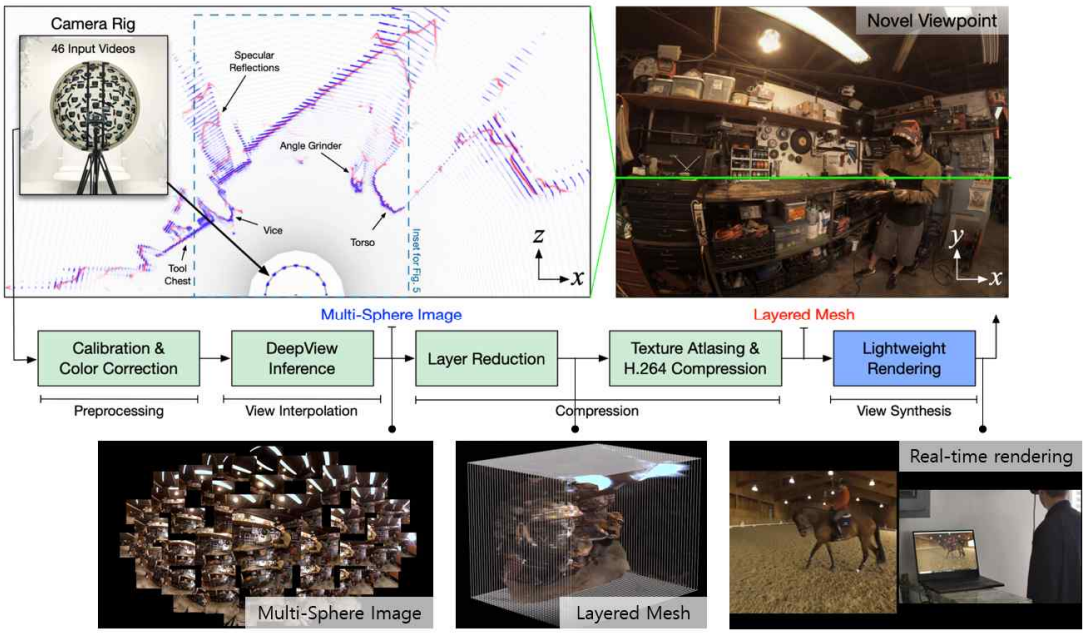
입력 영상으로부터 MPI를 효과적으로 예측하기 위해, MPI를 구성하는 각 계층 영상들 간의 기울기 차분 값을 재귀적으로 학습하는 합성곱 신경망을 설계하였다. (그림 6-40)은 MPI 기반 재현 기술의 개념도를 나타낸다. 구성된 MPI를 통해 단방향의 제한된 시점 공간 내에서 6DoF를 지원할 수 있다.



(그림 6-40) Google의 MPI기반 재현 기술

■ 다중 구면 영상 기반 재현 기술

상기 MPI 기술을 전방위로 확장하기 위한 시도로서, Google 사에서는 2020년 MSI(Multi-Sphere Image) 기술 및 이를 렌더링 측면에서 최적화하기 위한 계층화된 메쉬 표현(layered mesh representation) 기술을 발표하였다. (그림 6-41)은 제안된 이머시브 라이트필드 기술의 전체 파이프라인을 나타낸다.



(그림 6-41) Google의 이머시브 라이트필드 기술 전체 파이프라인

제안된 파노라믹 라이트필드 획득 구조체는 반구 형태의 리그에 부착된 46개 카메라를 통해 입력 영상을 획득한다. 이 영상들을 이용하여 딥뷰 모듈에서는 MSI 영상들을 예측한다. MSI 영상은 MPI와 유사하게 전체 장면 공간을 유한한 수의 계층 영상으로 표현하되, 피라미드 형태의 평면 영상이 아닌 반구 형태의 곡면 영상들로 투영시킨 것이다. 그런데 이때 전체 화각 범위에 대해 MSI를 매우 촘촘하게 구성하면 처리할 데이터가 많아지고 반대로 성기게 구성하면 장면에 대한 정밀한 표현이 어려운 문제가 있다. 이를 위해 상대적으로 듬성한(sparse) MSI를 추론한 후 분할하여 일련의 레이어 그룹을 만들고, 깊이 정보를 이용하여 결합 및 정제한 뒤 계층간에서 메쉬화(meshify)함으로써 MSI를 렌더링 측면에서 최적화하였다. 이를 통해 데이터를 효율적으로 처리하면서 상대적으로 고품질 및 실시간 재현이 가능하며, 직경 80cm, 화각 220° 범위의 시점 공간 내에서 6 자유도의 라이트필드 비디오 실시간 재현을 제공한다.

6.3 볼류메트릭 콘텐츠 기술

볼류메트릭(Volumetric) 콘텐츠 생성 기술은 미리 정의된 공간 내에 있는 객체나 사람을 볼륨 공간 형태로 복원하여 획득한 3D 또는 4D(외형+움직임) 콘텐츠를 지칭하는 것으로, 이와 같이 획득한 콘텐츠는 전방향에서 자유롭게 바라볼 수 있다는 특징이 있다. 특히, 사람을 실제와 같이 복원하여 전송하는 기술은 다가오고 있는 메타버스 시대의 핵심 기술 중 하나라고 할 수 있으며 본 절에서는 볼류메트릭 콘텐츠를 획득하기 위한 획득 기술과 휴먼 복원 기술 동향에 살펴보고자 한다.

6.3.1 볼류메트릭 영상 획득 기술

볼류메트릭 콘텐츠 획득을 위해서는 적게는 수 십대에서 많게는 백대 이상의 다중 카메라를 사용하며 동기화된 영상을 실시간으로 획득하고 저장하는 기술의 개발이 필요하다. 딥 러닝 기술이 널리 사용되기 전에는 주로 스튜디오 환경에서 이와 같은 다중 시점의 영상을 이용하여 복원을 수행하였기 때문에 주로 폐색영역이나 텍스처가 부족한 영역에서의 복원 품질이 주요한 문제였다. 이에 관련 연구들은 카메라의 수를 늘려 폐색 영역을 최소화하거나 대응점 탐색을 높여 복원된 모델의 품질을 높이는 방향으로 진행이 되었다. 기술 개발의 초창기라 할 수 있는 1990년대와 2000년대 초에는 크로마키 환경에서 전배경 분리를 통해 얻어진 분할 지도를 이용한 복원 방식이 주를 이루었으며, 그 후 카메라와 컴퓨팅 자원의 발전과 더불어 컴퓨터 비전 알고리즘도 함께 발전하여 복원 성능이 크게 향상되었다. 2010년대에 이르러 관련 기술이 솔루션 형태로 판매되기 시작하였으며 대표적으로 미국 Microsoft 사의 혼합 현실 캡처 스튜디오, 뉴질랜드의 8i 스튜디오, 미국 Intel 사의 Intel Studio, Canon 사의 볼류메트릭 비디오 스튜디오 등이 있으며 각 솔루션의 특징을 요약하면 다음과 같다.

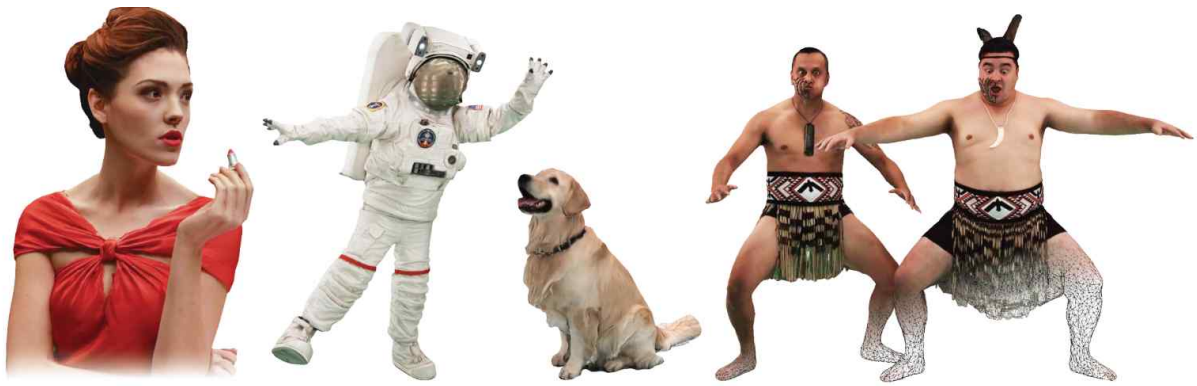
<표 6-1> 볼류메트릭 콘텐츠 제작 솔루션 비교

	Mixed Reality Capture Studio	8i Studio	Intel Studio	Canon Studio
복원 대상	1-4인	1-4인	50명 이상 (300평 규모)	최대 10인 (8m X 8m)
카메라 대수	106	20 - 60	100+	100+
해상도	2K+	2-4K	8K	4K
특징	IR 구조광 기반 깊이 추정 및 컬러 영상 융합	컬러 영상 기반	뮤지컬 및 영화 제작 목적	스포츠, 공연 복원 목적

■ Microsoft (Mixed Reality Capture Studio)

Microsoft 사의 Mixed Reality Capture Studio는 106대의 산업용 카메라로 구성되며, 획득된 영상은 깊이 추정, 전배경 분할, 깊이 융합, 텍스처 생성 과정 등을 거쳐 최종적으로 3D 모델링이 이루어진다. 특이할 만한 점은 다른 기업의 방식과 다르게 (그림 6-42)에 보이는 것과 같이 적외선 기반의 구조광(structured light)을 사용하여 깊이 정보를 추정한다는 것이며, 해당 기술은 Kinect 센서 초기 모델과 유사하게 랜덤 패턴을 투사하는 접근 방식을 갖는다. 적외선 기반의 구조광을 사용하면 인체나 의상 표면이 갖는 모호함을 줄여주어 복원의 품질을 높일 수 있다는 장점이 있으며, 컬러 정보, 실루엣 정보를 동시에 사용하여 각각의 정보가 갖는 한계점을 극복하고자 하였다. 또한 연속되는 메쉬 간 연결성을 추정하는 기술과 얼굴 부분에 대한 품질 향상을 위한 텍스처 맵 생성 기술을 제안하였으며, 복원된 결과는 (그림 6-43)에서 확인할 수 있다. 해당 솔루션은 국내 통신사에서 도입한 사례가 있으며, 이를 활용하여 다양한 광고/공연 등 응용 서비스를 위한 콘텐츠를 획득하고 있다.

(그림 6-42) Microsoft사의 Mixed Reality Capture Studio(좌)와 카메라 구성도(우)



(그림 6-43) Mixed Reality Capture Studio 콘텐츠 획득 결과

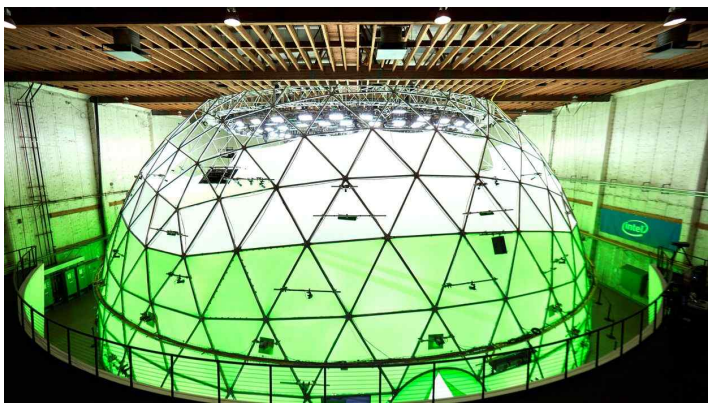
■ 8i Studio

8i Studio의 특징은 컬러 영상만을 이용하여 볼류메트릭 복원을 수행한다는 것이며, 세밀한 복원을 위해 공간에 대한 사전 지식을 바탕으로 다양한 초점 거리를 갖는 카메라를 배치한다는 특징이 있다. 예를 들어 얼굴 영역을 촬영하는 카메라는 초점 거리가 큰 카메라를 사용하여 (예: 25mm) 얼굴 영역을 크게 촬영하여 복원하고, 외형 변화가 몸통의 경우 초점 거리가 작은 카메라를 사용하여 보다 넓은 영역을 복원한 뒤, 최종적으로 다양한 품질의 깊이 영상을 융합하는 접근 방법을 갖는다. 카메라는 응용 서비스에 따라 구성이나 해상도 변경이 가능하며, 적게는 24대에서 많게는 60여대의 카메라에 대한 복원 기능을 제공한다. 우리나라에 가장 먼저 도입된 볼류메트릭 솔루션이라고 할 수 있으며, 대표적으로 청하의 CF 제작에 활용되어 큰 이목을 끌었다.

(그림 6-44) 국내 스튜디오 솔루션 도입 사례: LGU+(좌측)와 SKT(우측)의 Jump Studio

■ Intel Studio

Intel Studio는 기존의 스튜디오와는 달리 대규모 공연, 자유 시점 영화 콘텐츠 제작을 위해 제작되었으며 이를 위해 8K급 카메라 100대 이상을 사용하였다. 최대 50명에 이르는 사람을 동시에 복원할 수 있다는 특징이 있으며, 해당 시스템을 통해 획득되는 데이터의 크기가 10초당 1테라바이트에 이르기 때문에 방대한 양의 데이터를 효과적으로 처리하고 복원을 수행하는 것이 Intel Studio의 핵심 기술이라 할 수 있다. 또한 기존 시스템이 사람 복원에 초점이 맞춰진 것과 달리 (그림 6-46)과 같이 여러 명의 사람과 주변 환경까지 동시에 복원할 수 있다는 특징이 있다.



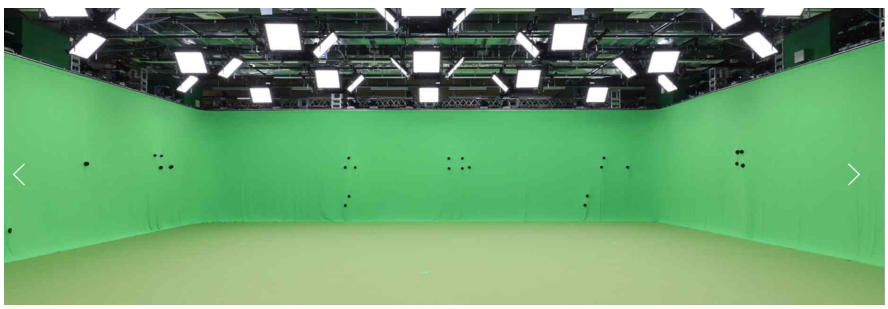
(그림 6-45) Intel Studio

해당 기술은 뮤지컬 그리스의 3차원 복원을 비롯하여 다양한 헐리우드 영화 제작에 사용되었으며, (그림 6-46)은 자유 시점에서 바라본 영화의 한 장면을 보여주고 있다. 하지만 스튜디오를 유지하기 위해 엄청난 유지비용과 인력이 필요했기 때문에, Intel Studio는 2020년 10월에 서비스를 종료하였으며 아직 대형 공간에 대한 복원을 위해서는 풀어야 할 숙제가 남아있다는 것을 시사하였다.

(그림 6-46) Intel Studio 콘텐츠 획득 결과

■ Canon: Volumetric Video Studio

Canon 사의 Volumetric Video Studio는 10명 이내의 사용자로 구성된 공연이나 스포츠 경기에 대한 볼류메트릭 콘텐츠를 생성하기 위해 설계되었다. 약 8mx8m 공간에 대한 복원을 지원하며, 100대 이상의 4K 급 카메라를 벽면에 설치하였으며 최대 60fps 속도로 영상을 획득하고 복원하는 것이 가능하다는 특징이 있다.



(그림 6-47) Canon Volumetric Video Studio

(그림 6-48) Canon Volumetric Video Studio 콘텐츠 획득 결과

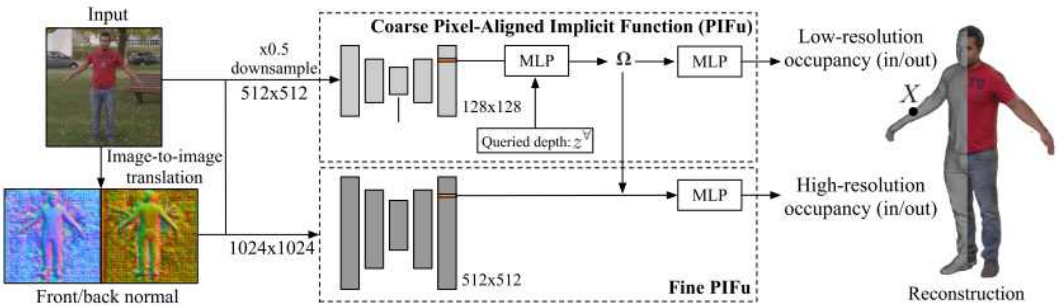
6.3.2 볼류메트릭 콘텐츠 제작을 위한 휴먼 복원 기술

볼류메트릭 콘텐츠 기술의 목적이 주로 사람에 대한 외형이나 움직임 등을 복원하는데 있기 때문에 가장 핵심이 되는 기술 또한 휴먼 복원 및 모델링 기술이라 할 수 있다. 앞서 설명한 Microsoft나 8i Studio의 경우 스테레오 정합 기술, 전배경 분할 기술, 깊이 영상 융합 기술 등 주로 전통적으로 사용되던 기술을 활용해 솔루션을 개발하였으며, 최근 들어 각각의 요소 기술이 딥 러닝 기술로 대체가 되고 있는 실정이다. 한편으로는 딥 러닝 기술의 발전으로 인해 다수의 카메라를 이용하지 않고 휴먼 모델링을 수행하는 연구 결과를 Facebook이나 Google과 같은 글로벌 IT 기업에서 보여주고 있으며, 앞으로의 기술은 다수의 카메라 보다는 한 대의 카메라나 뎁스 센서를 이용하는 방향으로 발전될 것으로 예상이 된다. 본 절에서는 단안 영상 기반의 휴먼 복원 기술과 단일 뎁스 센서 기반의 휴먼 복원 기술 동향에 대해 살펴본다.

깊이 영상을 사용하지 않고 한 장의 영상만으로 전신 복원을 수행하는 기술은 기술의 난이도가 높지만 대신 휴대폰이나 웹캠 등으로 간편하게 아바타를 만들 수 있다는 장점이 있어 기술이 갖는 파급력 측면에 있어서는 가장 중요한 기술 중 하나라고 할 수 있다. 관련된 대표적 연구로는 BodyNet, PIFu, PIFuHD, Moulding Human 등이 있으며 기술의 접근 방법과 특징은 다음과 같다.

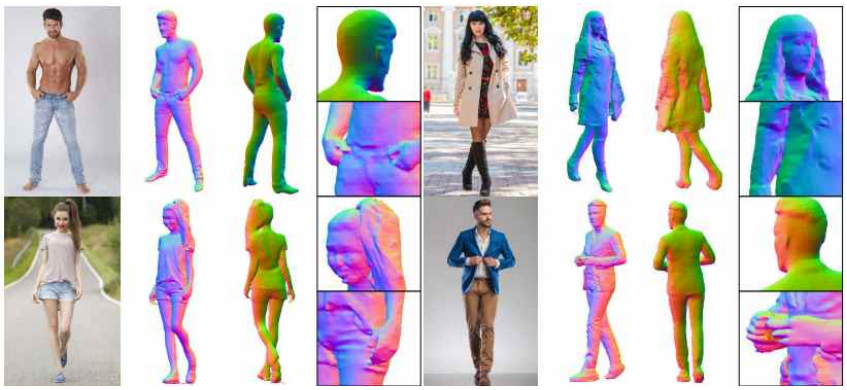
■ 딥 러닝 기반의 암시적 볼륨(implicit volume) 예측을 통한 휴먼 복원 기술

BodyNet에서는 종단(end-to-end)간 학습이 가능한 휴먼 모델링 기술을 제안하였으며 입력 영상으로부터 2D/3D 자세 정보와 전경 정보를 추정 한 뒤 이를 동시에 활용하여 볼륨 공간(implicit volume)을 예측하는 방법을 제안하였다. 하지만 종단 간 학습을 통해 볼륨 공간을 예측하기 위해서는 학습 시 많은 양의 메모리를 필요로 하여 높은 해상도의 복원이 어렵다. 높은 해상도의 모델 복원을 위해 PIFu와 PIFuHD에서는 볼륨 공간 전체를 예측하지 않고 개별 복셀(볼륨 공간을 구성하는 기본 단위)에 대해 점유(occupancy) 여부를 예측하는 접근 방법을 제안하였다. 다시 말해 학습되는 네트워크는 영상으로부터 추출된 특징 벡터를 입력으로 받아 특정 복셀이 객체의 내/외부에 존재하는지 여부를 예측하는데, 복셀 단위로 예측이 이루어지기 때문에 네트워크 구조가 단순하고 적은 양의 데이터만을 이용하여 학습하는 것이 가능하다는 장점이 있다.



(그림 6-49) PIFuHD: 단일 고해상도 영상에 대한 휴먼 복원 기술

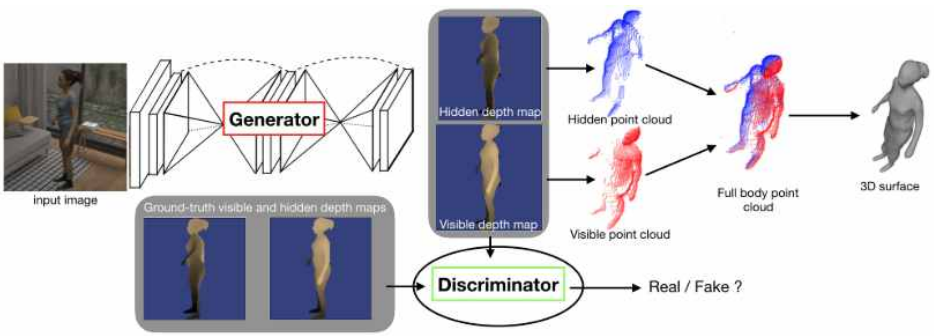
결과적으로 PIFuHD의 경우 단일 영상을 사용하여 (그림 6-50)과 같은 수준의 복원이 가능함을 보였고, 최근 수행되고 있는 많은 연구들이 본인들의 연구 결과를 PIFuHD와 비교하거나 발전시키는 형태로 이루어지고 있다. 예를 들어, PIFu, PIFuHD의 경우 복셀 단위로 예측을 하다보니 전체 형상(global shape)에 대한 제약사항이 없어 형상이 알 수 없는 형태로 일그러지는 경우가 있는데, Geo-PIFu의 경우 형상 정보를 동시에 이용하여 이와 같은 문제점를 해결하였으며, StereoPIFu의 경우 단안 영상을 양안으로 확장하여 안정성과 정확성을 한 단계 더 향상시키고자 하였다.



(그림 6-50) PIFuHD: 단일 입력 영상에 대한 휴먼 복원 결과

■ 깊이 영상 예측을 통한 휴먼 복원 기술

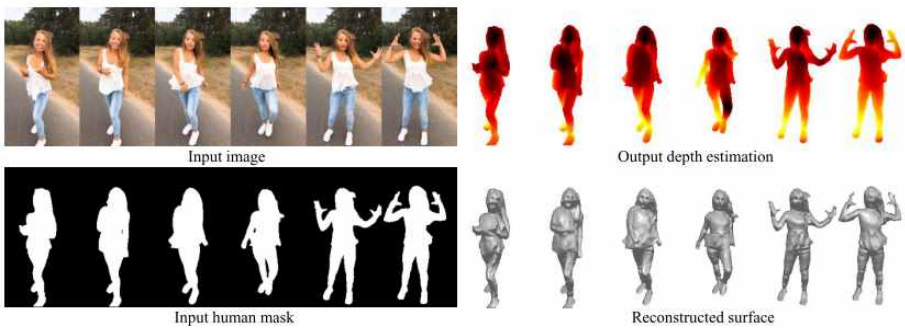
영상에서 볼륨 공간을 예측하는 것은 가장 널리 사용되는 접근 방법이기는 하지만 볼륨 공간 자체가 영상과 해상도가 동일한 3차원 공간이기 때문에 예측하는데 걸리는 시간이 오래 걸린다는 단점이 있고, 종단 간 학습 시 높은 메모리 사용량 때문에 고해상도 복원이 어렵다는 한계점이 있다. 이를 극복할 수 있는 접근 방법은 볼륨 공간 대신 여러 시점에 대한 깊이 영상을 생성하고 융합하는 방식이 있다. Moulding Humans에서는 입력 영상에서 정면과 후면에 대한 깊이 영상을 예측하고 이를 융합하여 3차원 메쉬를 구성하는 적대적 신경망(adversarial neural network) 기반의 접근 방법을 제안하였다. 볼륨 공간을 예측하는 것과 대비했을 때 정확도는 높지 않지만 종단 간 학습을 통해 효과적으로 모델이 복원 가능함을 입증하였다.



(그림 6-51) Moulding Humans: 단안 영상 기반의 3차원 모델 생성 과정

■ 소셜 미디어 동영상 기반의 깊이 영상 예측 기술

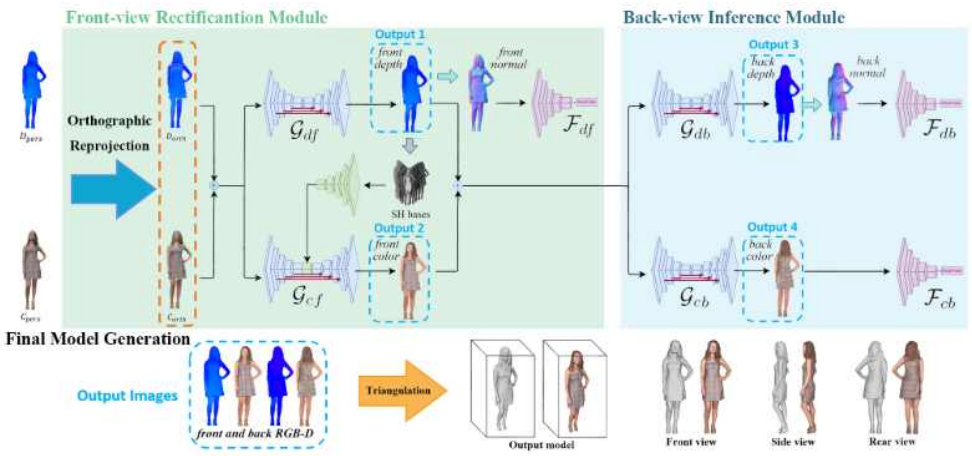
최근에는 사람에 대한 깊이 영상을 추정하는데 있어서 실측 모델이나 깊이 정보 없이 동영상 정보만을 사용하여 깊이 영상 시퀀스를 예측하는 방법이 미국 미네소타 대학에서 제안되었으며, 새롭게 데이터 셋을 제작하거나 획득하지 않고 기존 소셜 미디어 데이터에서 (그림 6-52)처럼 3D 모델링이 가능함을 보여주었다.



(그림 6-52) 소셜 미디어 동영상 기반의 깊이 영상 예측 기술

■ 깊이 센서 및 템플릿 모델 기반 휴먼 복원 기술

한편으로는 단일 영상 또는 비디오 대신 깊이 센서를 사용해서 전신에 대한 복원을 수행하고자 하는 시도가 자주 이루어지고 있다. 관련 연구는 칭화대의 Yebin Liu 교수 연구진에 의해 활발하게 진행되고 있으며, 대표적으로는 최근 발표된 NormalGAN이나 DeepHuman 논문이 있다. NormalGAN의 경우 정면에서 촬영된 깊이 영상으로부터 후면 깊이 영상을 예측하는 접근 방법을 제안하였으며 보다 좋은 결과를 생성하기 위해 입력 영상의 투영 성분을 제거하는 방법과 음영을 제거하는 방법, 법선 지도에 대한 식별자(discriminator)를 이용하는 방법을 제안하였다. DeepHuman의 경우, 템플릿 모델을 이용하여 전체적인 3차원 형상을 추정하고 이후 영상 정보를 활용하여 세부적인 디테일을 복원하는 접근 방법을 갖는다.



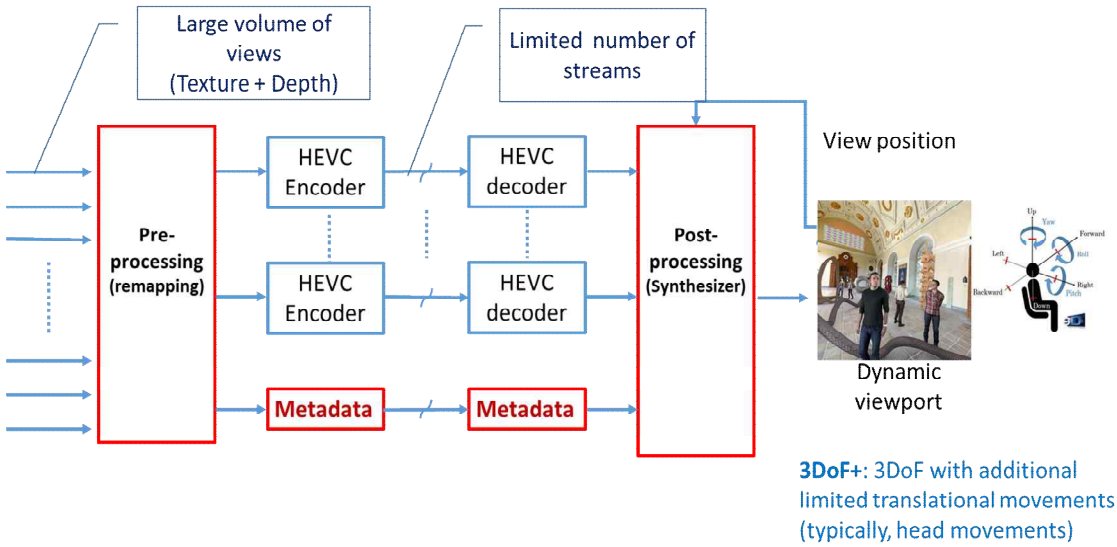
(그림 6-53) NormalGAN: RGB-D 영상을 입력으로 받아 음영이 제거된 영상과 후면에 대한 깊이 영상과 컬러 영상을 예측하는 기술

7 몰입형 비디오 표준화 동향

7.1 MPEG-I Visual 표준화

비디오, 오디오와 같은 멀티미디어 데이터의 압축 기술을 표준화하는 단체로 널리 알려진 MPEG(Moving Picture Experts Group)에서는 '16년 10월부터 VR(Virtual Reality), AR (Augmented Reality), 다시점 비디오(Multiview Video) 등 몰입형 미디어(Immersive Media)의 대중화를 끌어내기 위해 MPEG-I(Immersive) 프로젝트를 진행 중이다. 본 프로젝트의 목표는 전 세계 IT 기업과 연구기관의 요구사항 및 기술을 반영하여 최대 6 자유도를 지원하는 몰입형 미디어의 서비스에 사용될 수 있는 비디오, 오디오 및 시스템 표준 기술을 개발하는 것이다.

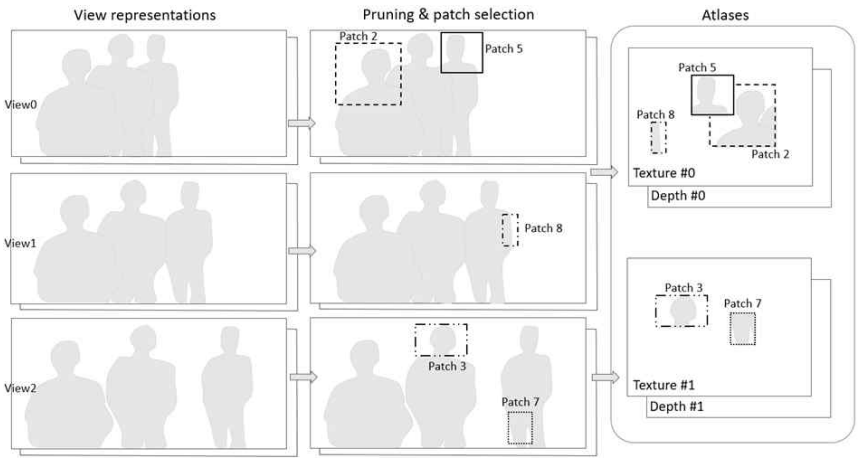
MPEG 산하의 조직 중 하나인 MPEG Video Coding에서는 몰입형 미디어의 핵심 요소인 운동시차(motion parallax)를 지원하는 방안으로 깊이 정보를 활용한 가상시점 합성 (virtual view synthesis) 기술을 고려했으며, (그림 7-1)과 같은 방향성을 기반으로 관련 표준화 작업을 '19년 3월부터 "MPEG Immersive Video(MIV)"라는 이름으로 진행해왔다. MIV 표준화의 목적은 HEVC(High Efficiency Video Coding) 등과 같은 기존의 2D 동영상 압축 기술을 그대로 사용하면서 여러 쌍의 텍스쳐(texture) 및 깊이맵(depthmap)으로 구성된 다시점 영상을 데이터양 측면에서 최소화하면서 재현 품질은 최대화하기 위한 전처리, 후처리 그리고 메타데이터 기술이다. 본절에서는 '21년 4월에 표준의 마무리 단계인 FDIS(Final Draft International Standard)에 이른 MIV 1.0 표준의 기술에 대해 상세히 알아보고, '22년 6월부터 시작될 것으로 예상하는 MIV 버전 2.0의 진행 방향성에 관해 설명한다.



(그림 7-1) MIV 표준화 개념도

7.1.1 테스트 모델

MPEG 내에서 테스트 모델(Test Model)이란 어떤 주제에 대해 표준화를 진행할 때 개발 중인 표준 기술의 유효성을 검증하기 위해 구현한 소프트웨어를 의미한다. MIV 표준화에서도 “Test Model for Immersive Video(TMIV)”란 이름으로 테스트 모델을 개발하고 있으며, 표준화가 본격적으로 시작된 ’19년 3월에 버전 1을 발간한 이후 ‘21년 10월 현재까지 채택한 다양한 기술을 반영하여 버전 11 상태이다. TMIV 소프트웨어를 구성하는 여러 기술적 요소에 관해 설명하기 전 먼저 MIV 표준 기술이 어떤 방식으로 다시점 영상의 방대한 데이터양을 줄이는지에 대해 간략히 살펴보고자 한다. 일반적으로 다시점 영상은 조밀한 간격으로 배치된 복수의 카메라를 통해 획득하므로, (그림 7-2) (좌)처럼 시점 영상 간에는 상당한 양의 중복 데이터가 존재할 수밖에 없다. 이러한 특성을 고려하여 (그림 7-2) (중간)과 같이 “프루닝(pruning)”이라는 과정을 통해 중복 데이터를 제거하고, 각 시점에서만 보이는 장면 정보를 “패치(patch)”라는 조각 영상 형태로 추출한다. 그리고 (그림 7-2) 우측처럼 여러 시점에서 추출한 다수의 패치를 “아틀라스(atlas)”라는 영상으로 결합(이하 패킹(packing))한다. 이렇게 다수의 시점 영상을 훨씬 적은 수의 아틀라스로 재구성함으로써 데이터양을 획기적으로 줄임과 동시에 이전부터 미디어 서비스에 널리 사용해온 기존의 동영상 코덱(HEVC, VVC 등) 및 전송(DASH[2]) 기술과도 호환이 되도록 한 것이다.



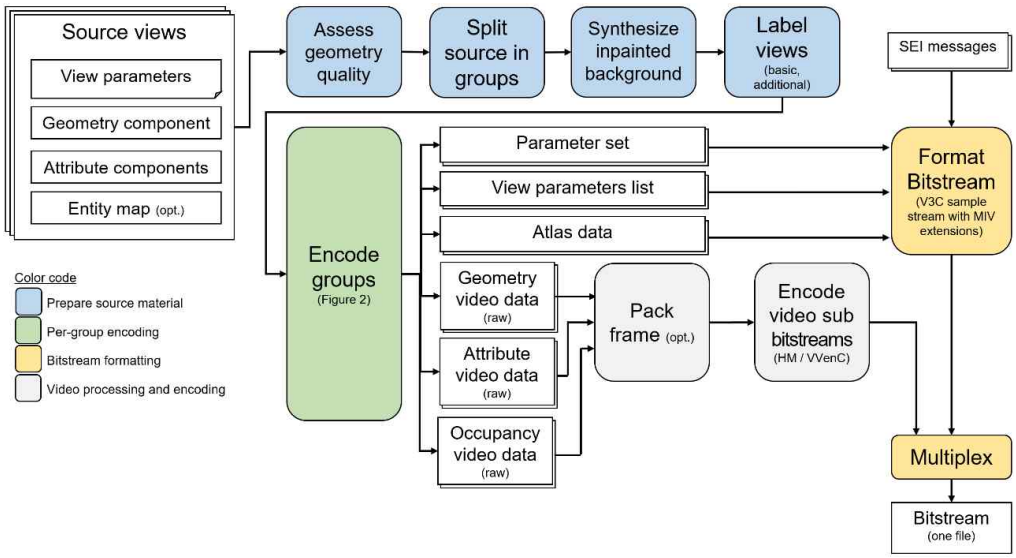
(그림 7-2) 패치 및 아틀라스 개념을 활용한 MIV 표준 개념

TMIV 테스트 모델은 크게 “TMIV 부호화기(TMIV Encoder)”와 “TMIV 복호화기(TMIV Decoder)”로 구성되어 있다. 미디어 서비스 워크플로우에 비유하면 TMIV 부호화기는 시청자에게 송신할 데이터를 준비하는 단계이며, 다시점 영상을 가공하여 아틀라스와 이와 관련된 메타데이터를 생성한다. 핸드폰, TV 등 재현 단말에 해당하는 TMIV 복호화기는 아틀라스와 메타데이터를 수신하여 시청자가 보는 영역(이하 뷰포트(viewport))에 대응하는 영상을 합성하여 재현한다.

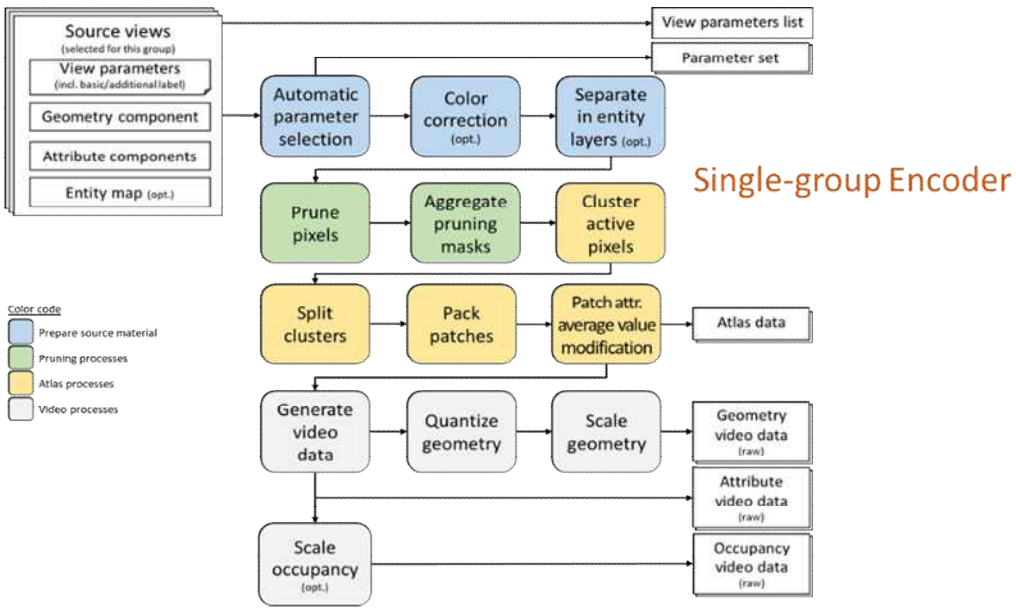
2) DASH: Dynamic Adaptive Streaming over HTTP

7.1.1.1 TMIV 부호화기

TMIV 부호화기의 구성도는 (그럼 7-3)과 같으며, 모듈 중 하나인 “Encode groups” 모듈은 (그림 7-4)와 같이 더 세분될 수 있다. 본 장에서는 TMIV 부호화기가 입력값으로 다시점 영상과 시점 위치, 각도 등의 카메라 변수 정보를 받아 최종 출력 결과인 아틀라스 영상과 메타데이터로 구성된 하나의 “MIV 비트스트림(Bitstream)”을 생성하는 과정에 관해 설명한다.



(그림 7-3) TMIV 부호화기 구성도



(그림 7-4) “Encoder group” 모듈 구성도

■ Assess Geometry Quality

가장 먼저 “Assess geometry quality” 과정을 통해 콘텐츠의 깊이맵이 정확한지 또는 부정확한지를 자동으로 평가한다. 본 과정의 목적은 깊이맵의 품질에 맞춰 TMIV 부호화기

및 TMIV 복호화기의 일부 처리 과정을 가변적으로 수행하기 위함이다. 첫 번째 프레임의 정보를 기준으로 각 시점의 깊이맵을 다른 시점의 위치로 투영(projection)하고, 두 시점의 정보가 겹치는 영역에서 같은 위치에 대응하는 픽셀의 깊이 값 차이의 합이 특정 문턱값 이상이 되는 경우가 발생하는 시점이 한 개라도 존재하면 해당 콘텐츠는 저품질로 판정된다. 평가에 사용되는 문턱값은 사용자가 임의로 설정할 수 있으며, 기본값의 경우, 실사 테스트 콘텐츠는 저품질로 CG 테스트 콘텐츠는 고품질로 판정하도록 설정되어 있다.

■ Split Source in Groups

“Split source in groups” 과정은 여러 개의 입력 시점을 위치의 근접성 정도를 기반으로 여러 개의 그룹으로 겹치지 않게 나눈다. 본 과정의 목적은 (그림 7-3)과 같은 아틀라스 생성 과정(“Single-group Encoder”)을 그룹별로 독립적으로 수행함으로써 아틀라스 간의 공간적 의존성을 없애어 특정 영역에 대응하는 영상을 재현할 때 모든 아틀라스가 아닌 필요한 일부만 선택적으로 가져오는 “공간적 임의 접근(spatial random access)” 서비스를 지원하기 위함이다.

■ Label Views

“Label views” 과정은 입력 시점을 기본시점(basic view)과 부가시점(additional view) 두 종류로 겹치지 않게 나눈다. 본 과정의 목적은 아틀라스를 구성할 때 (그림 7-5)와 같이 정보를 원상태 그대로 유지할 기본시점과 프루닝 과정을 통해 중복 정보를 제거한 이후 여러 개의 패치로 나눌 부가시점을 찾는 것이다. 기본시점이 필요한 이유는 일정 수준의 재현 품질을 보장하기 위함이다.

(그림 7-5) 아틀라스 내 기본시점 및 부가시점 형태

첫 번째 단계는 기본시점의 개수를 결정하는 것이다. “Automatic parameter selection” 과정의 아틀라스 크기 계산 방법을 빌려 아틀라스의 너비 및 높이를 계산하고, “아틀라스 총 샘플(또는 픽셀) 수”를 의미하는 두 값의 곱에 아틀라스 개수를 곱하여 “프레임 당 총 샘플 수”를 구한다. 그리고 이 값과 기본시점에 의한 픽셀 수의 비율이 특정 문턱값을 초과하지 않는 범위 내에서 기본시점의 개수를 결정한다. 두 번째 단계 기본시점의 개수만큼 기본시점을 선택하는 것이다. 이를 위해 시점 간 거리 정보를 바탕으로 두 종류의 비용함수(cost function)를 정의한다. 기본시점이 두 개 이상일 때 사용하는 “반발 비용(repulsion cost)” 함수는, 가장 멀리 떨어진 시점의 조합이 기본시점이 되게 설계돼있다. 기본시점이 한 개일 때 사용하는 “당김 비용(attraction cost)” 함수는 다른 시점과 거리의 합이 최소화되는 시점이 기본시점이 된다. 즉, 가장 중앙에 있는 시점을 선택한다. 다시점 콘텐츠 획득에 사용되는 카메라의 구조는 대부분 대칭 구조이며, 이로 인해 비용 함수를 최소화하는 여러 쌍의 기본시점 조합이 나올 수 있다. 이때는 임의로 한 쌍을 선택하는 것이 아닌 리그 구조상 가장 중앙에 있는 시점을 기본시점으로 먼저 고정하고 이 시점을 기준으로 비용 함수가 최소화되는 시점의 조합을 나머지 기본시점으로 선택한다.

■ Automatic Parameter Selection

“Automatic parameter selection” 과정은 아틀라스의 크기를 자동으로 계산한다. 우선 아틀라스 너비를 시점 영상 너비와 같게 설정하고, 아틀라스 너비 및 높이, 콘텐츠 프레임율(frame rate), 최대 복호화기 개수를 곱한 값이 “초당 최대 샘플 수(maximum luma sample rate)” 값을 초과하지 않는 범위에서 아틀라스 높이를 결정한다. 콘텐츠 프레임율은 이미 주어진 정보이며, 최대 복호화기 개수는 단말의 현실적인 사양을 고려하여 총 네 개로, 초당 최대 샘플 수 값은 미디어 서비스에 많이 사용되는 HEVC 코덱 Main 10 프로파일(profile) 사양이 지원하는 1,069,547,520로 설정했다. 나중에 깊이맵 아틀라스는 데이터양 최소화를 위해 “Geometry downscaling” 과정에서 공간적으로 축소될 수 있는데, TMIV 복호화에서는 렌더링 작업 수행 전 먼저 본 과정을 통해 계산된 아틀라스의 크기로 복원된다.

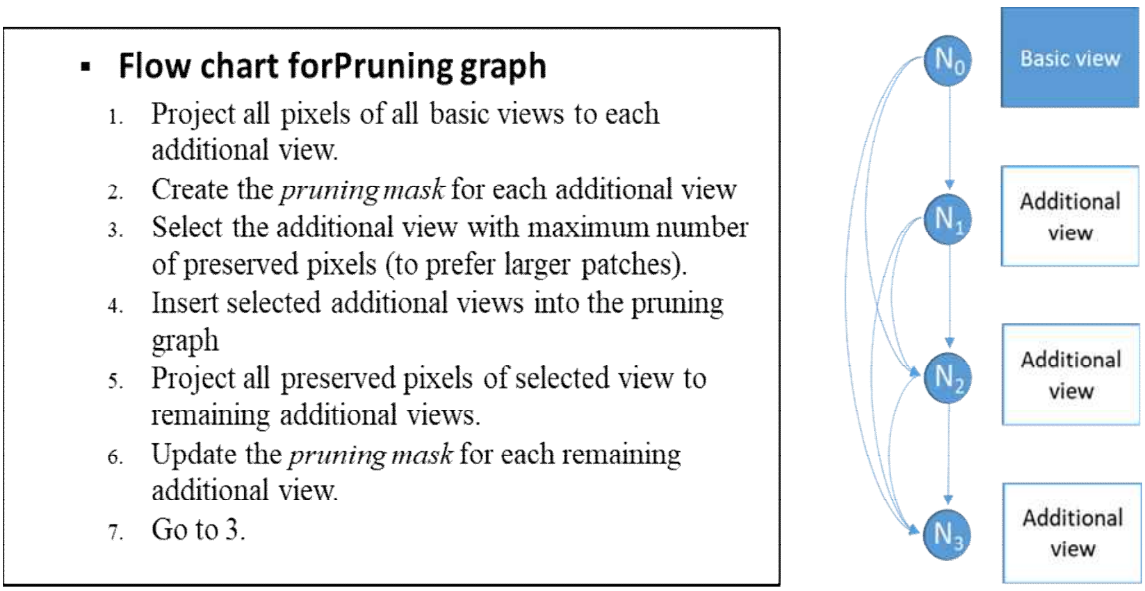
■ Color Correction

“Color correction” 과정은 각 시점의 색상 값을 조절하여 시점 간의 색상 일관성(color consistency)을 올리는 작업을 수행한다. 본 과정의 목적은 아틀라스를 구성하는 텍스쳐 정보를 유사하게 만들어 압축 효율성을 높이기 위함이다. 카메라 리그 중심점에서 가장 근접한 시점을 기준 시점으로 설정하고, 시점별로 색상 정보의 평균값이 기준 시점과 일치하도록 만들기 위해 기준 시점과의 평균값의 차이를 모든 화소에 더하거나 뺀다. 이후 TMIV 복호화기는 이 평균치 조정값을 받아 색상을 원 상태로 복원한다.

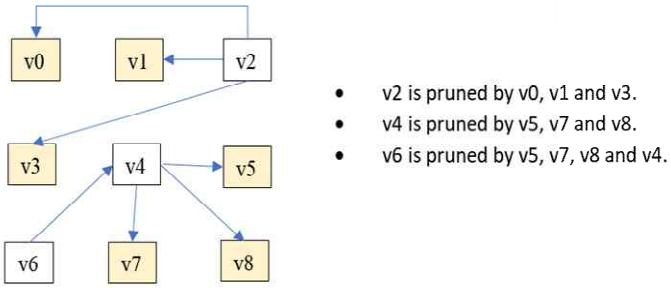
■ Prune Pixels

TMIV 부호화기를 구성하는 핵심 과정 중 하나인 “Prune pixels”은 각 부가시점의 정보를 기본시점 그리고 다른 부가시점과 비교하여 중복성 유무를 바탕으로 제거돼도(이하 프루닝)

괜찮은 영역을 판별하는 작업을 수행한다. 프루닝 작업의 수행을 위해 (그림 7-6) (우)와 같은 계층 구조의 "프루닝 그래프(pruning graph)"를 구성한다. 첫 프레임 정보만 활용하여 구성하며, 그 구조는 시간상에서 변하지 않는다. 그래프에서 점(node)은 시점 번호를, 선은 시작점에 대응하는 시점의 정보(텍스쳐 및 깊이맵)를 끝점의 시점으로 투영하는 과정을 의미한다. 그래프 설계 과정에 대한 설명은 (그림 7-6) (좌)와 같은데, 먼저 기본 시점을 가장 상위 계층에 할당하고, 기본시점의 정보를 각 부가 시점으로 투영한다. 각 부가시점에 대해 화소 단위로 기본시점과 중복된 정보인지의 유무가 0(중복)과 1(미중복)로 표기된 "프루닝 마스크(pruning mask)"라는 이진법 영상을 생성한다. 이때 중복 정보 인지에 대한 판단은 같은 위치의 화소에 대해 깊이 값 및 Y 채널 색상 차이를 특정 문턱값과 비교하여 결정한다. 그리고 미중복 화소 개수가 가장 많은 부가시점을 두 번째 계층에 할당하는데, 이는 크기가 큰 패치를 많이 만들어 아틀라스의 공간적 복잡도를 최소화함으로써 압축 효율을 올리기 위함이다. 뒤이어 기본시점과 두 번째 계층의 부가 시점의 정보를 이전 과정처럼 아직 계층 할당이 안 된 부가 시점으로 투영하여 프루닝 마스크를 만들고, 미중복 픽셀 수가 가장 많은 부가 시점을 세 번째 계층에 할당한다. 모든 부가 시점에 대한 계층 할당이 끝날 때까지 이러한 과정을 반복하여 최종적인 프루닝 그래프를 만든다. 하지만, 위 절차를 단 한 개의 프루닝 그래프로 수행할 경우 계층 구조가 너무 길어져 프루닝 작업에 상당한 시간이 소요된다. 이를 위해 여러 개의 시점을 (그림 7-7)과 같이 여러 무리로 중복되지 않게 모아 "프루닝 클러스터(pruning cluster)"를 구성하고, 클러스터별로 독립적으로 프루닝 작업을 수행한다. 클러스터를 구성할 때는 각 클러스터가 비슷한 개수의 기본시점을 가지게 하고, 서로 인접한 기본시점을 같은 클러스터로 배정한다. 부가 시점의 경우 장면 정보 측면에서 중첩 정도가 가장 큰 기본 시점이 있는 클러스터에 배치한다.

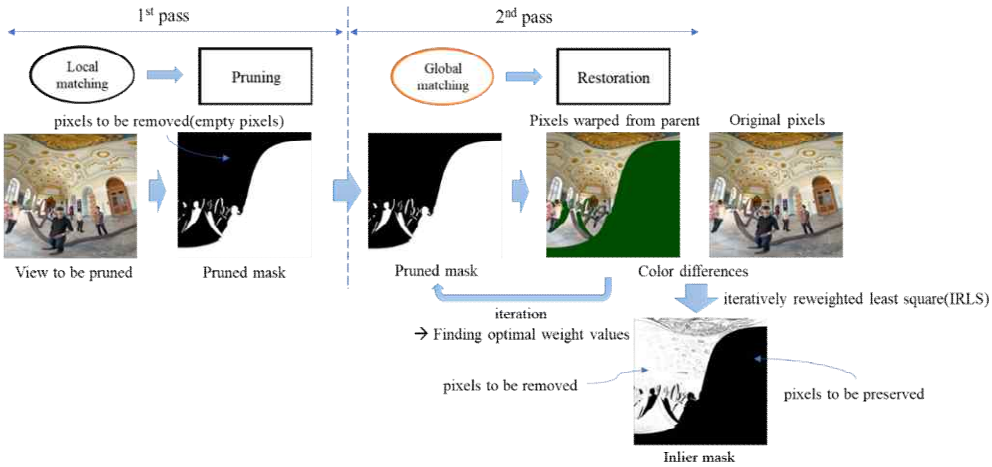


(그림 7-6) 프루닝 절차(좌) 및 프루닝 그래프 개념도(우)



(그림 7-7) 프루닝 클러스터 예시

위 과정을 통해 일차적으로 중복 및 미중복 정보를 판별한 이후, (그림 7-8)과 같은 "2차 프루닝(Second pass pruning)" 과정을 통해 프루닝 마스크를 수정한다. "최소 제곱법(least square method)"을 통해 시점 영상 간의 전역적 색상 차이를 예측하는 하나의 "적합식(fitting function)"을 계산하고, 특정 화소의 색상 값이 이 식을 통해 예측한 색상 값에서 일정 문턱 값만큼 벗어나면 해당 화소의 표식을 중복에서 미중복으로 변경한다. 본 과정의 목적은 시점 간에 깊이 값은 유사하나, 외부 조명 등에 의해 텍스쳐 정보가 다른 영역을 보존하기 위함이다.



(그림 7-8) 2차 프루닝 개념

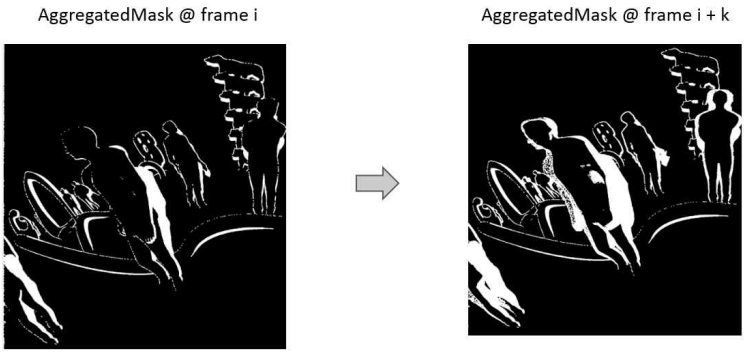
위 과정으로 생성된 프루닝 마스크를 보면 미중복 화소로 구성된 영역의 형태가 불규칙하거나 중간에 작은 구멍이 있는데, 이를 마지막으로 3x3 크기의 "침식(erosion)" 및 "확장(dilation)" 필터를 통해 다듬어 최종의 프루닝 마스크를 생성한다.

■ Aggregate Pruning Mask

"Aggregate pruning mask" 과정은 부가시점별로 (식 7-1)과 같은 "OR" 연산자를 통해 프레임 단위의 마스크 영상을 "인트라 주기(intraperiod)"라는 기간 단위로 누적하여 하나의 "누적 마스크(aggregated mask)"를 생성한다. (그림 7-9)는 본 과정을 시각화한 것인데, 예를 들어 사람 경계 부분의 흰 영역이 움직임으로 인해 점점 두꺼워짐을 알 수 있다. 인트라 주기는 TMIV 사용자가 설정하는 값으로, 현재는 HEVC 코덱의 기본 설정값인 32를

사용한다. 이렇게 마스크를 누적하는 이유는 패치의 모양을 특정 기간만큼 같게 유지하여 패치와 관련된 메타데이터를 프레임 단위가 아닌 인트라 주기 별로 정의하여 전송 데이터양을 최소화함과 동시에 아틀라스 내에서 패치의 위치를 시간 축으로 같게 유지하여 프레임 간에 일관성을 주어 아틀라스의 압축률을 높이기 위함이다.

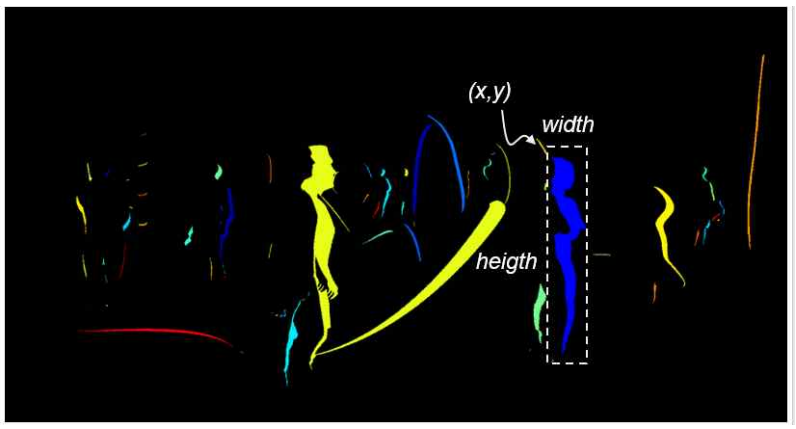
$$\begin{aligned} &\text{aggregatedMask}[i]@\text{current\_frame} \\ &\quad = \max(\text{Mask}[i]@\text{current\_frame},\ \text{aggregatedMask}[i]@\text{previous\_frame}) \end{aligned} \qquad \text{(식 7-1)}$$



(그림 7-9) 프루닝 마스크 영상 누적 과정 예시

■ Cluster Active Pixels

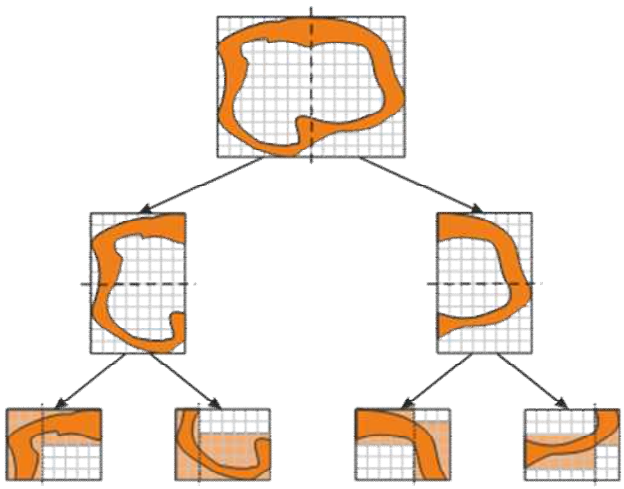
“Cluster active pixel” 과정은 부가시점별로 “Aggregate pruning mask” 과정을 통해 생성된 누적 마스크에서 “클러스터”를 생성한다. 클러스터는 여덟 방향(좌우, 상하, 대각선)으로 이웃한 값이 1인 유효 화소를 연결하여 (그림 7-10)과 같이 만든 화소의 집합을 의미한다(그림 상에서 다른 색상은 다른 클러스터를 의미). 그리고 임의 형태를 가진 클러스터를 감싸는 직사각형을 정의하여 유효 화소와 값이 0인 무효 화소로 구성된 패치의 기본 구조를 결정한다. 직사각형의 너비와 높이는 16 또는 32의 배수가 되도록 하는데, 이는 HEVC 코덱 기술의 원리를 고려하여 압축 과정에서 발생하는 정보 손실을 최소화하기 위함이다.



(그림 7-10) 클러스터 예시

■ Split Clusters

카메라 리그 구조에서 외각에 위치한 시점의 경우 (그림 7-11) 상단과 같이 유효 픽셀의 분포가 알파벳 C 또는 L 형태인 패치가 생성된다. 이러한 패치는 유효 화소가 점유한 공간에 비해 패치 크기가 너무 커져 아틀라스의 공간 낭비를 초래한다. 이에 TMIV 부호화기는 (그림 7-11) 하단과 같이 C 또는 L 형태의 유효 픽셀로 구성된 패치를 반복적으로 더 작은 패치로 나누는 기술을 지원한다. 패치 분할은 높이와 너비 중 짧은 방향으로 수행하고, 정중앙을 기준으로 정확히 같은 넓이로 이 등분한다. 패치 분할 여부는 분할 전 패치의 넓이와 분할 후 두 패치의 넓이의 합의 차이를 문턱값과 비교하여 결정한다.



(그림 7-11) C 또는 L 형태의 유효 픽셀로 구성된 패치 분할 예시

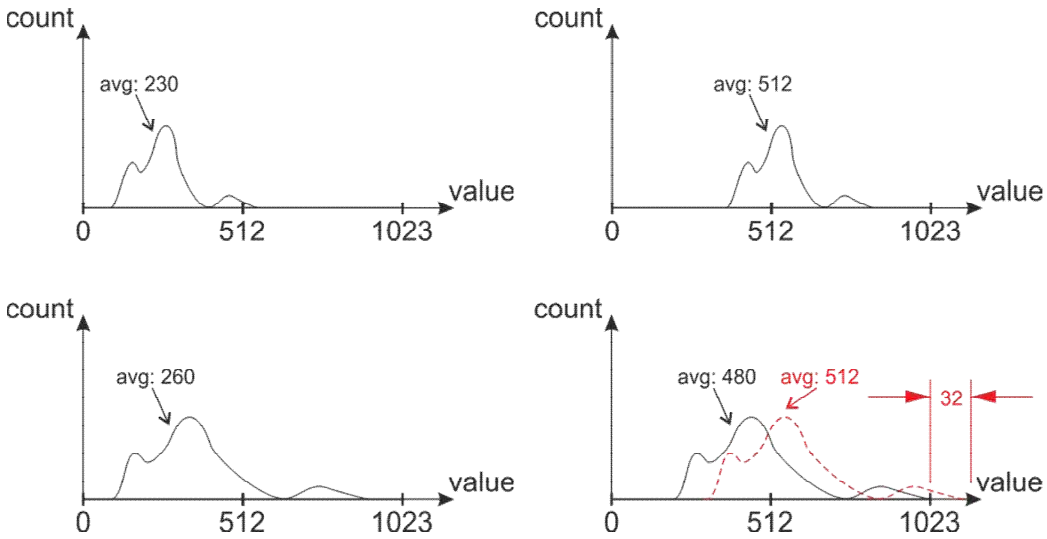
■ Pack Patches

“Pack patches” 과정은 여러 과정을 통해 생성된 다수의 패치를 아틀라스로 패킹하기 위해 각 패치를 아틀라스의 어떤 위치에 90도 회전하여 배치할 것인가를 결정한다. 패치 패킹에는 “MaxRect”라는 알고리즘을 사용하는데, 이는 패치를 크기가 큰 순서대로 아틀라스의 왼쪽 위부터 집어넣고, 패치를 이미 배치한 공간 중에서 무효 화소 영역에 들어갈 수 있으면 그 영역에 집어넣고, 없으면 패치가 아예 없는 공간에 넣는 방식이다. 이를 고려했을 때 프루닝 과정 없이 원상태 그대로인 기본시점은 항상 아틀라스의 가장 위쪽에 위치함을 알 수 있다. 아틀라스 내에서 패치의 위치는 직사각형의 왼쪽 위 모서리를 기준으로 정의하며, 위치 또한 코덱 기술의 원리를 고려하여 16 또는 32의 배수로 설정한다. 패킹 과정 이후 TMIV 부호화기는 아틀라스 구조와 관련된 정보를 “어떤 패치, 어떤 시점의 어떤 위치에서 어떤 크기로 추출하여 아틀라스의 어떤 위치에 90도 회전 후 배치했다”의 형식으로 저장한다. 참고로 텍스쳐 및 깊이 정보에 대한 아틀라스 구조는 동일하다.

■ Patch Attribute Average Value Modification

텍스쳐 아틀라스를 보면 같은 패치 내 유효 영역과 무효 영역의 경계 또는 다른 패치의 유효 영역 간 경계에서 색상 값이 갑작스레 변하는 경우가 발생하며, 이는 텍스쳐 아틀라스의 압축 효율을 저하하는 요인으로 작용한다. 이를 위해 패치 단위에서 (그림 7-12)

(상)와 같이 유효 영역의 색상 평균값이 무효 영역에 할당한 중앙값(10비트 영상은 512)을 기준으로 한쪽으로 편향되면, 그 평균값이 중앙값이 되도록 유효 영역의 모든 화소값에 평균값과 중앙값의 차이를 더하거나 빼서 재조정한다. (그림 7-13) (하)와 같이 일부 패치에서는 보정 후 유효 픽셀의 값이 색상 최댓값(10비트 영상의 최댓값은 1023)을 넘어가는 경우가 발생하는데, 이를 보정에 사용하는 값을 자동으로 재조정한다. 이러한 색상 보정 작업은 Y, Cb, Cr 색상 채널 별로 수행하며, 평균치 조정을 위해 사용한 값은 메타데이터로 정의하여 TMIV 복호화기로 보내어 텍스쳐 아틀라스를 원래의 색상으로 복원할 때 사용한다.



(그림 7-12) 패치 단위 텍스쳐 정보 보정 전(좌) 및 후(우) 색상 분포도 형태

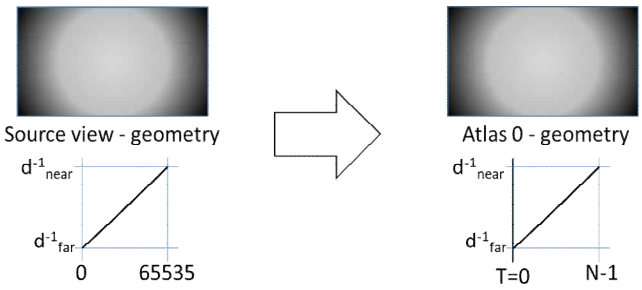
■ Generate Video Data

이전 과정은 아틀라스 크기, 패치 모양, 아틀라스 내 패치 위치 등을 구해서 아틀라스의 설계도를 제작하는 과정이었다면, “Generate video data” 과정은 이 설계도를 바탕으로 실제로 텍스쳐 및 깊이 정보 각각에 해당하는 아틀라스 영상을 생성한다. 먼저 아틀라스 개수만큼 빈 버퍼(buffer) 영상을 만들고, “Pack patches” 과정에서 계산한 패킹 구조에 맞춰 패치 인덱스 순서대로 패치를 특정 시점에서 추출하여 아틀라스의 특정 위치에 배치한다.

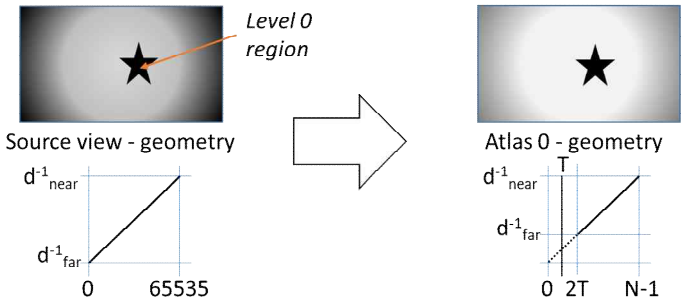
■ Quantize Geometry

TMIV 복호화기는 아틀라스의 각 화소가 렌더링 작업에 실제로 필요한 유효 화소인지 또는 불필요한 무효 화소인지를 판단하기 위해 화소 별 “점유(occupancy)” 정보가 필요하다. TMIV 부호화기는 점유 정보를 깊이맵 아틀라스 안에 삽입(embed)하는 방식과 0(무효 화소)과 1(유효 화소) 값으로 구성된 “점유지도(occupancy map)”라는 별도의 이진법 영상을 구성하는 방식을 지원한다. TMIV 소프트웨어는 첫 번째 방식을 기본 방식으로 사용하고 있는데, 이때 TMIV 부호화기는 특정 문턱 값을 메타데이터로 정의하고 TMIV 복호화기는 이를 받아 (그림 7-15) (우)와 같이 깊이 값이 문턱 값 이상이면 유효 화소 이하이면 무효

화소로 판정한다. 기본시점의 경우 유무효 화소 개념이 없기 때문에 문턱 값은 항상 0이다. 더불어 실사 콘텐츠는 CG 콘텐츠와 비교했을 때 깊이맵이 부정확하므로 깊이 값의 비트 수를 줄여도 렌더링 품질 열화가 적게 발생한다. 이러한 특성을 고려하여 깊이 값의 비트 수를 낮춰 데이터양을 줄이고 있다.



(그림 7-14) 기본시점의 깊이 정보 부호화 방법



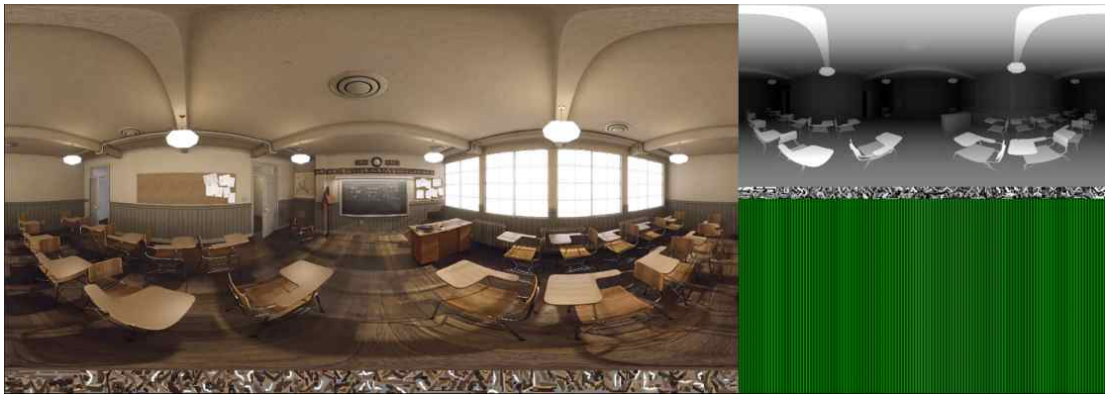
(그림 7-15) 부가시점의 깊이 정보 부호화 방법

■ Geometry Downscaling

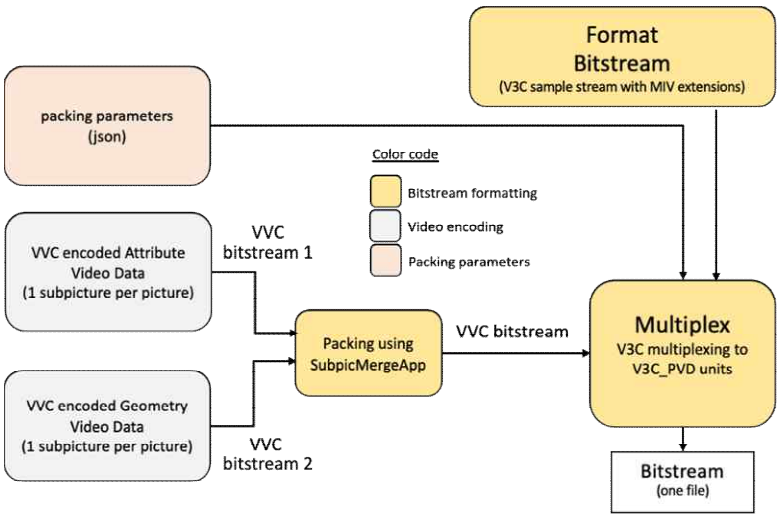
“Geometry downscaling” 과정은 데이터양 최소화를 위해 깊이맵 아틀라스의 크기를 일정 비율만큼 축소한다. 깊이맵 영상은 텍스쳐 영상에 비해 낮은 주파수 성분으로 구성돼 있어, 원래의 크기로 복원했을 때 상대적으로 정보 손실이 덜 발생한다. TMIV 부호화기는 축소 작업에 2x2 크기의 “최대 풀링(max pooling)” 필터를 사용하여 렌더링 품질에 중요한 전경 객체의 정보를 최대한 보존하도록 했다. 복원 과정에 대해서는 7.1.1.2장의 “Geometry scaling” 부분에서 설명한다.

■ Frame Packing

“Frame packing” 과정은 (그림 7-16)과 같이 여러 개의 텍스쳐 및 깊이맵 아틀라스를 아틀라스 단위로 결합하여 더 큰 크기의 아틀라스로 만든다. 본 과정은 아틀라스를 각자 따로 복호화할 경우 발생하는 프레임 간 비동기 문제를 해결하여 정확한 렌더링 작업을 가능케 하고, 복호화에 필요한 복호화기 개수를 절반으로 줄이는 것이다. TMIV 부호화기는 (그림 7-16)과 같이 각 아틀라스를 VVC로 부호화하여 생성한 여러 개의 비트스트림(bitstream)을 “Sub-picture merging software”라는 별도의 소프트웨어를 통해 하나의 비트스트림으로 결합하는 방식으로 프레임 패킹을 수행한다.



(그림 7-16) 프레임 패킹 된 아틀라스 예시



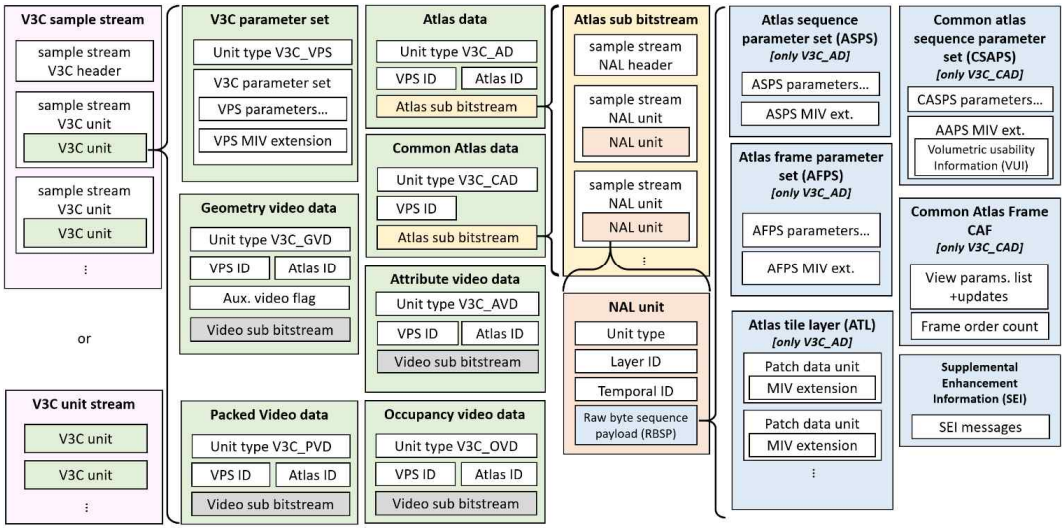
(그림 7-17) 프레임 패킹 과정

■ Encode Video Sub-bitstreams

아틀라스를 비디오 코덱으로 압축하여 하위 비트스트림(sub-bitstream)을 생성하는 과정이다. TMIV 부호화기는 현재 HEVC와 VVC 기술만 사용하고 있으나, 아틀라스가 2D 비디오 형태인 만큼 두 기술 이외에 다른 코덱 기술 또한 적용할 수 있다.

■ Multiplex

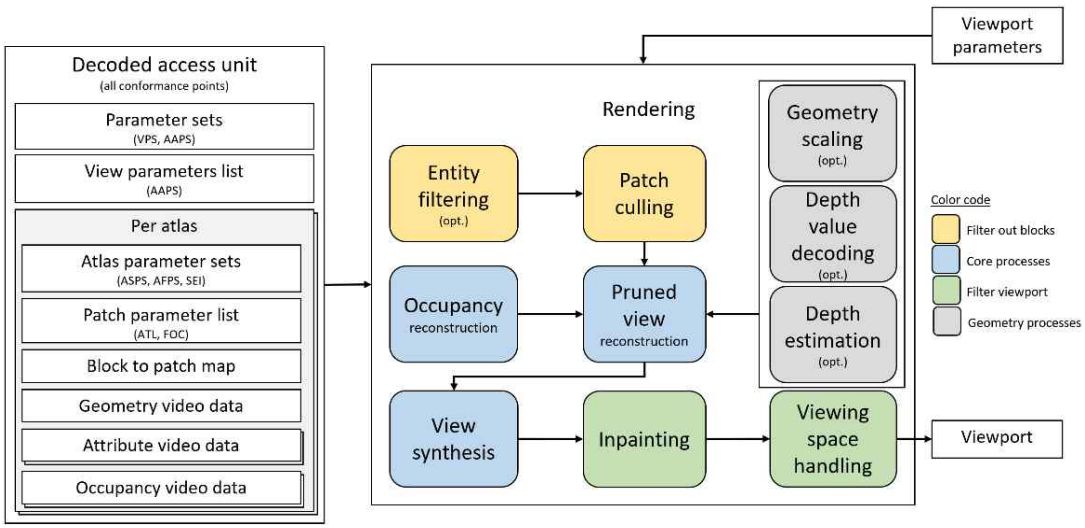
TMIV 부호화기를 구성하는 마지막 과정인 “Multiplex”는 아틀라스 관련 하위 비트스트림과 메타데이터를 (그림 7-18)과 같은 “V3C(Visual Volumetric Video-based Coding)” 표준의 구조로 배치하여 최종의 MIV 비트스트림을 생성한다. V3C 표준은 아틀라스와 패치 개념을 사용하여 포인트 클라우드(point cloud) 객체를 압축하는 V-PCC 표준과 MIV 표준의 기술을 결합한 것이다.



(그림 7-18) MIV 비트스트림 구조

7.1.1.2 TMIV 복호화기

TMIV 복호화기는 TMIV 부호화기에서 출력된 아틀라스 영상과 메타데이터를 받아 시청자의 뷰포트에 대응하는 영상을 재현한다. 본 절에서는 (그림 7-19)을 바탕으로 TMIV 복호화기의 구조에 관해 설명한다.



(그림 7-19) TMIV 복호화기 구조

■ Patching Culling

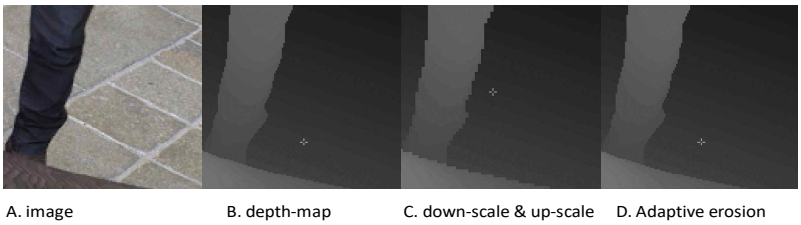
“Patch Culling” 과정은 뷰포트의 위치 및 각도 정보를 반영하여 재현에 불필요한 패치 정보를 제거한다. 물론 본 과정의 목적은 계산량을 줄여 렌더링 시간을 단축하기 위함이다. 패치의 네 모서리를 최소 및 최대 깊이 값을 이용하여 뷰포트로 투영한 이후 중첩 영역 여부에 따라 컬링 여부를 결정한다. 컬링된 패치의 픽셀은 아틀라스에서 무효 픽셀로 표시되고 렌더링 시 무시된다.

■ Occupancy Reconstruction

깊이맵 아틀라스에 임베디드 되던지, 별도로 전송되던지 아니면 보내지지 않은 경우에도 점유지도(occupancy map)는 최종적으로 아틀라스와 동일한 해상도로 복원되어야 한다. 점유지도가 기하정보에 임베디드 되었을 때, 점유지도는 기하정보를 업 스케일링한 후 기하값을 depthOccMapThreshold와 비교하여 결정된다. depthOccMapThreshold보다 작을 경우 점유지도는 0으로 세팅되고 클 경우 1로 세팅된다. 점유지도가 별도로 전송될 경우 (explicity occupancy map) 아틀라스 해상도로 복원하기 위해 nearest neighbor interpolation 알고리즘이 적용된다. 점유지도가 시그널링되지 않을 경우에는 점유지도는 아틀라스 해상도와 동일하게 1로 채운다.

■ Geometry Scaling

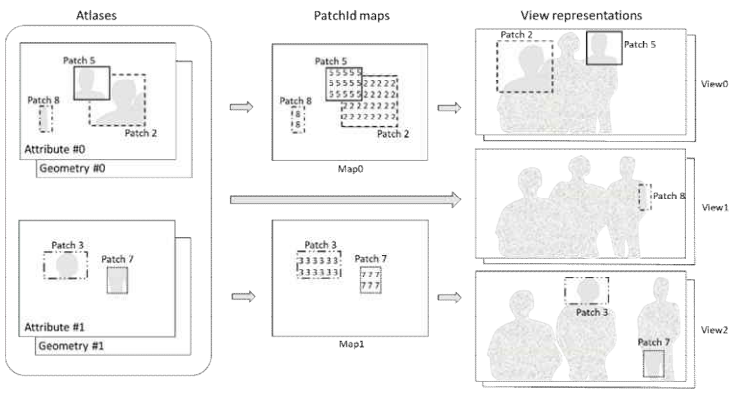
“Geometry scaling” 모듈은 크기가 절반으로 축소된 깊이맵 아틀라스를 원래의 크기로 복원한다. 가장 먼저 “Nearest Neighbor” 필터를 통해 업스케일링을 하고, 색상 정보 기반 침식 필터(color adaptive erosion filtering)를 통해 전경 객체의 에지가 약간 팽창할 수 있도록 하여 렌더링 시 화질열화를 최소화한다.



(그림 7-20) 기하정보 업스케일링 과정

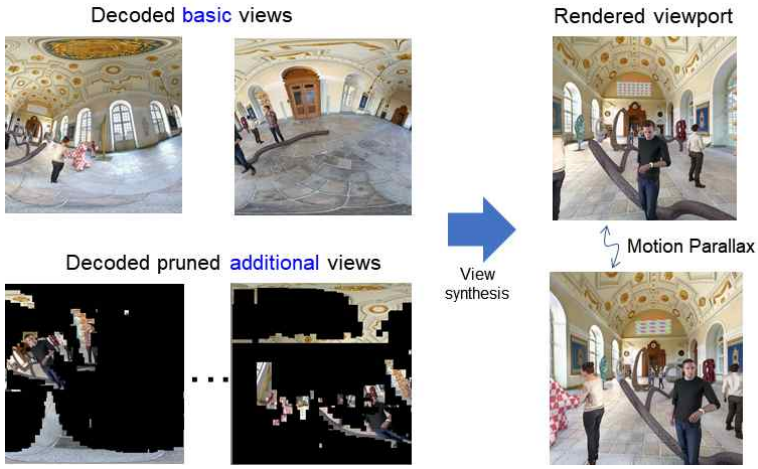
■ Pruned View Reconstruction

TMIV 부호화 단계에서 패치를 패킹할 때 patch-in-patch로 중첩되기 때문에, 아틀라스에서 pruned view를 복원하려면 중첩 순서를 표현할 수 있는 block-to-patch map에 대한 생성이 필요하다. (그림 7-21)과 같이 아틀라스에서 컬링 안되는 모든 패치는 원본 시점으로의 대응되는 위치로 복사된다.



(그림 7-21) Pruned view 복원 과정

(그림 7-22)는 복원된 기본시점 및 부가시점 영상들을 이용하여, 시청자의 시청위치 및 각도에 따라 가상시점을 합성하고 이를 통해 운동시차가 지원되는 뷰포트 영상을 렌더링 하는 개념을 설명하고 있다.

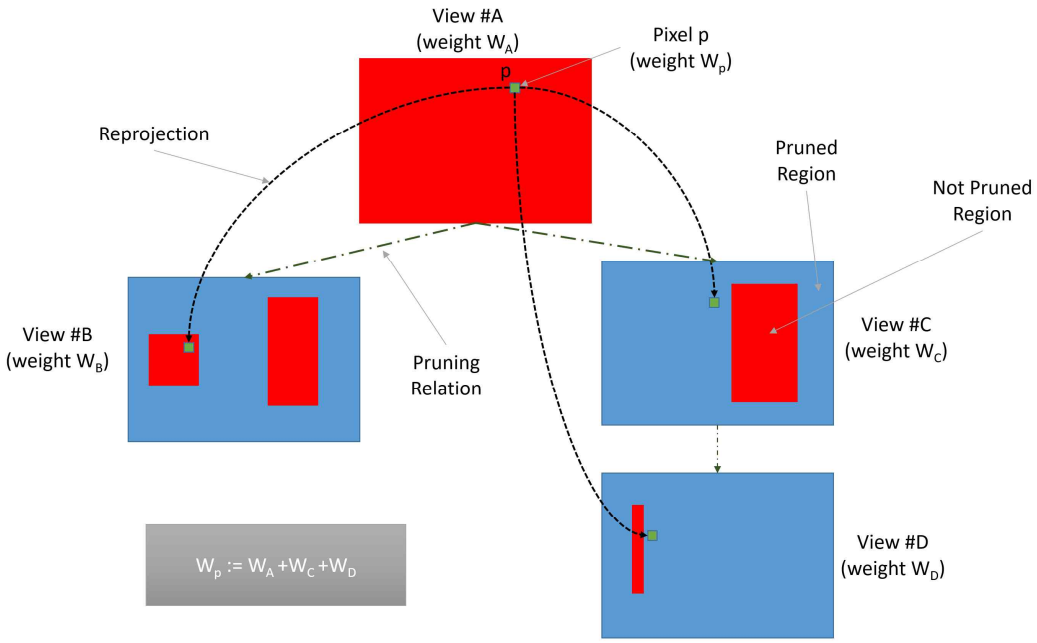


(그림 7-22) 복원된 기본/부가 시점영상들을 이용한 뷰포트 영상 렌더링 개념

■ View Synthesis and Inpainting

첫번째 Visibility map은 소스뷰로부터 기하맵을 unprojecting/reprojecting하여 목표 뷰포트에 대한 기하맵을 생성한다. 여기서 mesh 방식 대신 splat기반의 rasterization을 사용하여 연산량은 많지만 품질을 향상시킬 수 있었다. 소스뷰로 reprojection된 다수의 기하값은 voting process 및 median filtering 등을 통해 단일 기하맵으로 생성된다. 두 번째 “쉐이딩(shading)”은 뷰포트 영상의 색상 값을 계산하는 단계로, visibility map을 기반으로 시점의 화소를 backward-warping하고 블랜딩한다. 시점 영상의 경계를 추출하여 ghost를 제거하는 알고리즘도 포함되어 있다. 전통적으로 가상시점 합성 시 블랜딩은 주변 시점 영상에서 공간적으로 대응하는 픽셀들을 weighted 합을 하는 과정이다. 하지만 인코더에서 pruned된 픽셀에 대해서는 렌더링 시 blending weight값을 알 수가 없었는데, 세번째 특징으로서 Weighting 복원단계는 시그널링된 프루닝 그래프를 통해 blending weight값을 복원할 수 있다. 이를 위해 우선 Visibility map과 shading 단계는 전통적인 view weighting 개념을 따르며 시점 위치와 뷰포트 간의 거리 정보를 이용하여 초기 weight값을 계산한다. 다음으로 Pruned view를 처리할 때 다음과 같이 view weight 정보를 복원한다. 프루닝되지 않은 픽셀의 weight는 시점합성 단계에서 업데이트 되는데 이는 (그림 7-22)와 같이 메타데이터로 전송되는 프루닝 그래프의 내림순서로 대응하는 시점영상에서 프루닝된 픽셀을 고려함으로써 구해진다. 즉, 프루닝되지 않은 픽셀인 Non-pruned 픽셀의 weight를 보정하기 위해 다음과 같은 절차가 적용된다. wP와 wN을 위와 같이 시점 간 거리 정보에 의해 결정되는 초기 weight라고 하면, 블랜딩 작업 시 최종 weight는 다음과 같이 결정한다.

- 만일 픽셀 p가 자식노드 시점영상의 pruned 픽셀에 재투영된다면, 그 픽셀이 가지는 weight wP 는 자식노드의 weight wO와 함께 누적된다. (wP := wP + wO) 누적 계산은 손자노드까지 반복해서 진행된다.
- 만일 픽셀 p가 자식노드의 시점 영상으로 재투영되지 않는다면, 이전까지의 법칙은 손자노드 시점까지 반복해서 확장된다.
- 만일 픽셀 p가 어느 한 자식노드 시점영상의 unpruned 픽셀로 재투영된다면, weight는 변하지 않고 프루닝 그래프 상에서 weight 탐색은 중단된다.



(그림 7-23) 프루닝 그래프를 통한 Weight 복원 과정

7.1.2 CTC(Common Test Condition)

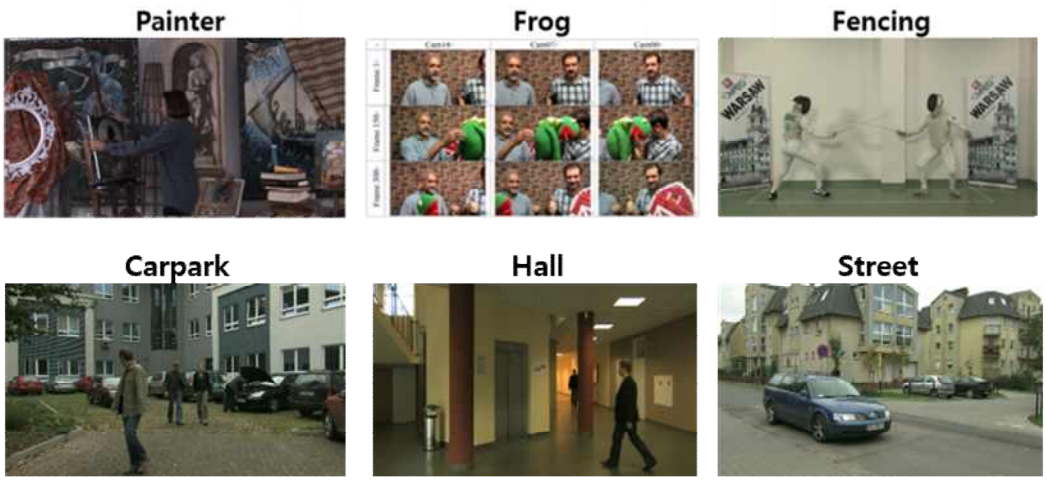
CTC란 “Common Test Condition”의 줄임말로 어떤 주제에 대해 표준화를 진행할 때 여러 기업에서 제안한 다양한 기술의 성능을 공정히 평가하여 표준으로의 반영 여부를 결정하기 위해 설립한 실험 수행절차와 평가 기준을 의미한다. 본 절에서는 MIV 표준에 신규 기술을 제안할 경우 성능 입증에 필요한 결과 생성을 위해 준수해야 하는 실험 조건과 그 결과를 어떤 과정을 통해 평가하는지에 대해 설명한다.

7.1.2.1 테스트 콘텐츠 정보

제안 기술의 평가에 테스트 콘텐츠는 필수적이며, 현재 MIV 표준화에서는 총 열세 종의 다시점 형태의 콘텐츠를 사용하고 있다. 각 콘텐츠는 CfTM 문서에 기재된 요구사항에 따라 시점별 텍스쳐 영상, 시점별 깊이맵 영상, 카메라 내/외부 변수 정보가 나열된 JSON(JavaScript Object Notation) 파일로 구성되어 있다. (그림 7-24)와 같은 CG(Computer Graphics) 콘텐츠는 복수의 가상 카메라를 통해 획득한 콘텐츠이며, 매우 정밀한 3차원 공간 정보로 구성된 실측 깊이맵(GTD: Ground Truth Depthmap)을 제공한다. 반면 (그림 7-25)와 같은 실사 콘텐츠는 복수의 물리적 카메라로 구성된 카메라

배열로 획득한 실사 영상이며, 깊이맵은 이전부터 MPEG-I Visual 그룹에서 깊이맵 추출에 사용해온 DERS(Depth Estimation Reference Software) 또는 IVDE(Immersive Video Depth Estimation) 소프트웨어를 통해 추출하였다. 물론 CG 콘텐츠와는 달리 컴퓨터 비전 알고리즘을 통해 텍스쳐 영상과 카메라 변수 정보에 기반한 깊이 정보를 추정하기 때문에, 어느 정도의 오차가 불가피하게 존재한다. CTC 문서 내에는 프레임률, 영상 포맷, 해상도 등 각 콘텐츠에 대한 정보가 상세히 기술되어 있으며, 이를 요약하면 <표 7-1> 및 <표 7-2>와 같다.

(그림 7-24) CG 콘텐츠 형태



(그림 7-25) 실사 콘텐츠 형태

<표 7-1> CG 콘텐츠 정보 요약

	Classroom Video	Museum	Hijack	Kitchen	Chess	Fan	Group
문서 번호	m42944	m42349	m42349	m43318	m50787	m54732	m54731
ID	A	B	C	J	N	O	R
프레임 수	120	300	300	90	300	97	99
프레임률	30	30	30	30	30	30	30
시점수	15	24	10	25	10	15	21
텍스쳐 포맷	YUV 4:2:0 10bit	YUV 4:2:0 10bit	YUV 4:2:0 10bit	YUV 4:2:0 10bit	YUV 4:2:0 10bit	YUV 4:2:0 10bit	YUV 4:2:0 10bit
깊이맵 포맷	YUV 4:2:0 16bit	YUV 4:2:0 16bit	YUV 4:2:0 16bit	YUV 4:2:0 10bit	YUV 4:2:0 16bit	YUV 4:2:0 16bit	YUV 4:2:0 16bit
해상도 (가로x세로)	4096×2048	2048×2048	4096×4096	1920×1080	2048x2048	2048x2048	2048x2048
화각 (가로x세로)	360°×180°	180°×180°	180°×180°	53°×31°	180°×180°	50°×38°	75°×48°
투영 형식	ERP	Half-ERP	Half-ERP	Rectilinear	Half-ERP	Rectilinear	Rectilinear

<표 7-2> 실사 콘텐츠 정보 요약

	Painter	Frog	Fencing	Carpark	Hall	Street
문서 번호	m47445	m47445	m38247	m51598	m51598	m51598
ID	NC - D	NC - E	NC - L	NC - P	NC - T	NC - U
프레임 수	300	300	300	250	250	250
프레임율	30	30	25	25	25	25
시점수	16	15	10	9	9	9
텍스쳐 포맷	YUV 4:2:0 10bit	YUV 4:2:0 10bit	YUV 4:2:0 10bit	YUV 4:2:0 10bit	YUV 4:2:0 10bit	YUV 4:2:0 10bit
깊이맵 포맷	YUV 4:2:0 16bit	YUV 4:2:0 16bit	YUV 4:2:0 16bit	YUV 4:2:0 16bit	YUV 4:2:0 16bit	YUV 4:2:0 16bit
해상도 (가로x세로)	2048×1088	1920×1080	1920×1080	1920×1088	1920×1088	1920×1088
화각 (가로x세로)	46°×25°	63.6°×38.5°	63°×48°	63°×48°	63°×48°	63°×48°
투영 형식	Rectilinear	Rectilinear	Rectilinear	Rectilinear	Rectilinear	Rectilinear

7.1.2.2 앵커 정보

제안 기술의 성능이 객관적 수치 및 시청 품질 측면에서 어느 정도 우수한지를 판단하기 위해서는 비교 대상이 필요하며, 이를 MPEG에서는 앵커(anchor)라 부른다. 앵커는 바로 이전 회의 때까지 개발된 표준 기술을 바탕으로 생성되며, 현재 MIV 표준화에서는 다음의 절차에 따라 앵커를 생성하고 있다.

- 단계 1: 테스트 콘텐츠를 TMIV 부호화기에 입력값으로 넣어 아틀라스 영상 데이터와 본 데이터와 연관된 메타데이터를 생성한다. 앵커 생성에는 테스트 콘텐츠의 모든 프레임을 사용하지 않고 일부 17 및 97프레임을 활용하여 “17fr” 및 “97fr” 두 종류의 앵커를 만든다. “17fr” 앵커의 목적은 개발 중인 기술의 성능을 신속히 확인할 수 있게 하는 등 개발의 편의성을 제공하기 위함이며, 제안 기술의 반영 여부 결정에는 “97fr” 앵커와의 비교 결과를 사용하고 있다.
- 단계 2: 영상 데이터를 다섯 개의 다른 QP(Quantization Parameter) 값을 적용하여 VVenC(Versatile Video Encoder)을 통해 부호화한다. 메타데이터는 정보를 원상태로 유지하기 위해 무손실 압축 기법을 적용한다. 다섯 개의 QP 값은 테스트 콘텐츠마다 서로 상이한데, 이는 각 QP 별로 설정한 최대 데이터량을 맞추기 위함이다.
- 단계 3: 각 QP 별로 압축된 영상 데이터와 메타데이터의 데이터량을 초당 킬로비트(kbps) 단위로 계산한다.
- 단계 4: 각 QP 별로 압축된 영상 데이터를 VVdeC(Versatile Video Decoder)를 통해 복호화한다.
- 단계 5: 각 QP 별로 복호화된 영상 데이터 및 메타데이터를 TMIV 복호화기에 입력값으로 넣어 원본 영상의 위치에 대응하는 텍스처 영상을 합성한다. 그리고 합성 영상과 원본 영상을 WS-PSNR(Weighted to Spherically uniform PSNR)과 IV-PSNR(Immersive PSNR) 두 종류의 평가 지표 측면에서 비교한다. WS-PSNR은 전방위 영상의 대표 투영 형식인 ERP(Equi-Rectangular Projection) 형식에서 발생하는 왜곡 특성을 고려하여, PSNR 계산 시 왜곡이 크게 발생하는 극점 부분은 적게 반영하고 적게 발생하는 적도 부분은 많이 반영하는 화질 평가 방법을 의미한다. IV-PSNR은 영상 합성 과정에서 발생하는 픽셀 이동(Pixel Shift) 현상을 고려하여 영상을 픽셀 단위 대신 블록 단위로 비교하는 방법이다.
- 단계 6: 각 QP 별로 복호화된 영상 데이터 및 메타데이터를 TMIV 복호화기에 입력값으로 넣어 (그림 7-26)과 같은 형태로, “포즈트레이스(posetrace)” 엑셀 파일 내에 나열된 시점의 위치(X, Y, Z) 및 방향(Yaw, Pitch, Roll)에 대응하는 영상을 합성한다. 이는 제안 기술의 시청 품질을 평가할 때 사용된다.

X	Y	Z	Yaw	Pitch	Roll
0	0	0	1,85E-05	-2,99E-06	3,20E-05
-2,40E-05	-0,00186	-6,27E-05	0,031693	-0,00152	-0,03308
-4,80E-05	-0,00371	-0,00013	0,063368	-0,00304	-0,0662
-7,20E-05	-0,00557	-0,00019	0,095042	-0,00455	-0,09931
-0,00012	-0,00727	-0,00028	0,121472	-0,01075	-0,12784
-0,00017	-0,00897	-0,00037	0,147901	-0,01695	-0,15636
-0,00022	-0,01068	-0,00046	0,17433	-0,02315	-0,18489
-0,00032	-0,01221	-0,0005	0,191171	-0,02616	-0,20674
-0,00043	-0,01375	-0,00055	0,208012	-0,02917	-0,2286

(그림 7-26) 포즈트레이스 파일 예시

7.1.2.3 제안 기술 객관적 품질 평가 과정

제안 기술에 대한 객관적 품질 평가는 다음의 절차를 통해 진행하고 있다.

- 단계 1: 앵커 생성에 사용한 동일한 프레임에 대응하는 테스트 콘텐츠를 제안 기술이 적용된 TMIV 부호화기에 입력값으로 넣어 아틀라스 형식의 영상 데이터 및 메타데이터를 생성한다.
- 단계 2: (단계 1) 과정에서 생성된 아틀라스 형식의 영상 데이터를 앵커 생성에 적용된 동일한 다섯 개의 QP 값을 적용하여 VVC 부호화기를 통해 압축한다.
- 단계 3: 각 QP 지점에 대해 압축된 영상 데이터 및 메타데이터의 데이터 양을 각각 초당 kbps 단위로 계산한다.
- 단계 4: 각 QP 지점에 대해 압축된 영상 데이터를 VVC 복호화기를 통해 복원한다.
- 단계 5: (단계 4) 과정에서 복원된 영상 데이터와 메타데이터를 제안 기술이 구현된 TMIV 복호화기에 입력 값으로 넣어 원본 영상이 획득된 모든 위치에 대응하는 영상을 합성하고, 각 영상을 동일한 위치에 대응하는 원본 영상과 WS-PSNR 및 두 종류의 객관적 품질 평가 지표 측면에서 비교한다.
- 단계 6: 각 테스트 콘텐츠 별로 앵커 및 제안 기술을 통해 생성된 결과 (QP 지점 별로 kbps 단위로 계산된 영상 데이터 및 메타데이터의 총 데이터 양 그리고 WSPSNR, IVPSNR 수치)를 엑셀 양식에 기입한다.
- 단계 7: 값 기입을 완료하면 엑셀 양식이 자동으로 앵커와 제안 기술 간의 “BD-Rate(Bjontegaard Rate)”을 계산하고 모든 콘텐츠에 대한 결과를 (그림 7-27)과 같은 형태로 출력한다. BD-Rate은 어떤 신규 기술 A에 의한 결과가 어떤 기준 기술 B에 비해 총 데이터량은 얼마나 감소 또는 증가하면서 상대적으로 WS-PSNR 및 IV-PSNR 수치가 얼마나 감소 또는 증가한 지를 고려하여 신규 기술의 성능이 얼마나 좋은지를 백분율 단위로 표현하는 기준을 의미한다. 녹색은 제안 기술이 앵커에 비해 성능이 좋음을 빨간색은 그 반대를 의미한다.

Mandatory content - Proposal vs. Low/High-bitrate Anchors

Sequence		High-BR BD rate Y-PSNR	Low-BR BD rate Y-PSNR	Max delta Y-PSNR	High-BR BD rate VMAF	Low-BR BD rate VMAF	High-BR BD rate IV-PSNR	Low-BR BD rate IV-PSNR
ClassroomVideo	SA	-0.2%	3.1%	1.91	-1.6%	2.9%	2.6%	4.0%
Museum	SB	0.3%	1.8%	16.66	0.3%	1.9%	1.2%	2.2%
Hijack	SC	-2.4%	0.8%	9.64	1.3%	2.3%	-2.8%	0.6%
Chess	SN	-23.2%	-7.9%	15.83	-21.6%	-4.5%	-5.7%	1.0%
Kitchen	SJ	-19.7%	-5.9%	16.12	-19.1%	-2.8%	-10.6%	-0.8%
Painter	SD	0.3%	0.6%	8.16	0.6%	0.7%	0.5%	0.7%
Frog	SE	9.8%	15.6%	5.36	14.9%	18.5%	15.0%	18.8%
Carpark	SP	-10.9%	-3.2%	7.16	-12.5%	-3.9%	-2.6%	0.8%
MIV		**-5.8%**	**0.6%**	**10.11**	**-4.7%**	**1.9%**	**-0.3%**	**3.4%**

(그림 7-27) 앵커 및 제안 기술 BD-Rate 비교 결과 예시

7.1.2.4 제안 기술 주관적 품질 평가 과정

대면 회의가 가능했던 '20년 1월까지 제안 기술에 대한 주관적 품질 평가는 다음 절차를 통해 진행되어왔다.

- 단계 1: 신규 기술을 제안하는 기관에서는 모든 테스트 콘텐츠, 다섯 개 QP, 세 종류의 포즈트레이스에 대응하는 (그림 7-28)과 같은 형태의 좌우대칭형 동영상을 생성하여 MPEG 회의에 가져온다. 본 동영상의 좌측 및 우측에는 각각 앵커와 제안 기술로 생성한 포즈트레이스 영상이 위치한다.

(그림 7-28) 주관적 화질 평가를 위한 좌우대칭형 동영상 예시

- 단계 2: 주관적 품질 평가를 진행하는 일시에 대형 4K 모니터가 설치된 공간에 모인다.
- 단계 3: 임의 평가자가 시청을 원하는 동영상을 테스트 콘텐츠, QP, 포즈트레이스 번호 형식으로 얘기하면 피평가자 측에서는 해당 동영상을 반복적으로 재생한다.
- 단계 4: 참여자들이 어떤 부분이 개선 또는 열화되었는지 의견을 얘기하면 평가 진행자가 그 내용을 기록한다.
- 단계 5: 평가 의견을 종합하여 해당 기술의 반영 여부를 결정한다.

비대면 회의로 전환된 '20년 4월부터 주관적 품질 평가는 다음 절차를 통해 진행되고 있는데, 본 절차가 과연 비대면 회의에 최적화된 방식인지에 대해서는 계속 논의하고 있다.

- 단계 1: 대면 방식과 마찬가지로 신규 기술을 제안하는 기관에서는 모든 테스트 콘텐츠, 다섯 개 QP, 세 종류의 포즈트레이스에 대응하는 좌우대칭형 동영상을 만든다. 그리고 이를 MPEG 콘텐츠 서버에 올린 이후 메일을 통해 평가자들에게 공지한다.
- 단계 2: 주관적 품질 평가를 진행하는 일시 전까지 평가자들은 각자 컴퓨터에 모든 동영상을 다운받는다.
- 단계 3: 임의 평가자가 시청을 원하는 동영상을 테스트 콘텐츠, QP, 포즈트레이스 번호 형식으로 얘기하면 각자 PC에서 해당 동영상을 시청한다.
- 단계 4: 평가자는 5점 만점에 기준한 점수 그리고 평가 의견이 적힌 메시지를 평가 진행자에게 보낸다.
- 단계 5: 평가 진행자는 받은 의견과 점수를 정리하여 평가자/피평가자에게 공유하고 이를 바탕으로 제안 기술의 반영 여부를 결정한다.

7.1.3 MIV 버전 2.0 표준화 방향 및 현황

’21년 4월 MIV 1.0 표준이 마무리 단계인 FDIS 단계에 이름에 따라 ‘21년 7월부터 버전 2.0 표준화 추진을 위한 사전 작업을 진행하고 있다. 버전 2.0 표준화는 완전히 새로운 표준 기술을 개발하는 것이 아닌, 아틀라스 및 패치 개념을 사용하는 버전 1.0의 전체적인 구조를 그대로 유지하면서 이를 개선하는 방향으로 진행될 것으로 보인다. 준비 작업의 첫 번째로 표준화 방향성 설립을 목적으로 표준 기술이 만족해야 할 요구사항과 지원해야 할 서비스 시나리오(이하 유즈케이스)를 정의하고 있다. 두 번째로는 정의한 요구사항 및 유즈케이스의 특성이 반영된 테스트 콘텐츠를 모집하고 있다. 세 번째로는 “Exploration of Experiment on Future MPEG Immersive Video (EE on Future MIV)”라는 이름으로 MIV 1.0 표준의 성능 개선을 위한 실험 아이템을 제안받아 이에 대한 실험을 그룹 차원에서 수행한 이후 결과를 공유 및 분석하는 절차를 진행하고 있다. 현재까지 정의된 여러 요구사항 및 유즈케이스 중 중요한 몇 가지를 요약하면 다음과 같다.

- MIV 1.0 표준 기술은 시청 위치 및 방향에 따라서 텍스쳐 정보가 변하는 유리, 거울 등의 “비랑베르(non-lambertian)” 객체를 사실적으로 재현하지 못하는 한계가 있었다. 버전 2.0 표준에서는 이러한 단점을 보완할 수 있는 기술의 필요성이 제시되었다.
- 6DoF 콘텐츠의 재현에는 일반적으로 고사양의 단말이 필요하다. 하지만, 저사양의 단말을 고려하여 시점을 단순히 스위칭(switching)하면서 시청할 수 있게 하되, 좀 더 몰입감 있게 보고 싶은 프레임이 있으면 이를 다운로드하여 6DoF를 체험하는 시나리오가 포함되어 있다.
- 메타버스(metaverse)와 같이 특성이 서로 다른 CG 및 실사 객체가 하나의 장면으로 재현돼야 하는 시나리오가 포함되어 있다.

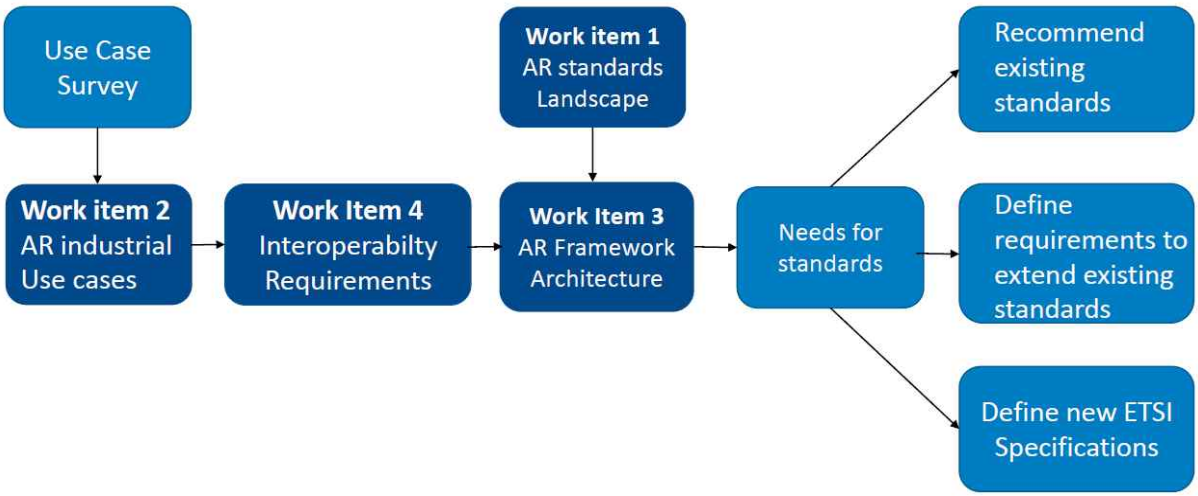
위의 요구사항 및 유즈케이스를 고려하여 버전 2.0 표준화 진행에 필요한 테스트 콘텐츠의 모집을 위해 “Call for MPEG-I Visual Test Materials”라는 문서를 발간했으며, 이에 대응하여 비램버트 객체가 포함된 콘텐츠, 촬영 반경이 확장된 콘텐츠 등이 제안 및 채택되었다. EE on Future MIV에서는 실사 콘텐츠의 깊이맵 추출에 사용되고 있는 소프트웨어인 IVDE(Immersive Video Depth Estimator)의 성능 개선, 비랑베르 객체 처리를 위한 프루닝 기법 개선, 주관적 화질 평가 과정 개선 등을 주제로 여러 실험이 진행되고 있다.

7.2 ETSI ISG ARF 표준화

ETSI는 2017년 12월부터 전체적으로 다양한 AR 응용 서비스를 위하여 요구되는 산업체 요구사항, 서비스 시나리오, 공통적으로 적용될 수 있는 프레임워크 및 요소기술 개발에 표준화 목적으로 ISG(Industry Specification Group) ARF(Augmented Reality Framework)를 발족하였다. 개발되는 AR 프레임워크 구조는 서로 다른 AR 구성 요소 간의 안정적인 상호 운용성(Interoperability)을 지원함과 동시에 AR 응용 서비스의 성공적인 시장 진입을 위한 핵심기술로 다양한 AR 응용 서비스 아래 관련 요소간의 프레임워크 구조 및 기술에 대한 표준화가 진행 중이다. 개발되는 AR 프레임워크의 기본적인 가용성은 다음과 같다.

- 소규모 플레이어 등 다양한 신규 진입자를 위한 다양한 솔루션 및 공통의 프레임 웍 구조 제공
- 공통 플랫폼기반 다양한 AR 응용서비스 개발 및 모듈식 기술 호환성 제공

ISG ARF는 네 가지의 워크아이템으로 토대로 한 표준화를 논의 중으로, 'AR standards Landscape (Work item 1)'는 AR에 대한 관련 표준화 기구 동향 및 AR 프레임워크 및 서비스에 적용될 수 있는 기존 표준을 분석, 'AR industrial Use cases(Work item 2)'는 다양한 산업적 서비스 시나리오 관련 요구사항 정의, 'AR Framework Architecture(Work item 3)'는 AR 서비스를 위하여 공통적으로 요구되는 요소기술을 포함하는 프레임워크 정의, 'AR Interoperability Requirements(Work item 4)'는 상기 Work item 2에서 도출된 시나리오에 따른 구체적인 AR 요소기술/기능, 시스템 및 서비스에 대한 상호운용성 요구사항을 정의한다. (그림 7-29)는 전체적인 ISG ARF 워크아이템 구조를 나타낸다.



(그림 7-29) ISG ARF 워크아이템

7.2.1 워크아이템 기반 표준화 진행 현황

7.2.1.1 AR standards landscape (Work item 1)

ISG ARF는 AR과 연관된 기존 표준을 분석하고 관련 핵심기술이 무엇인지 인지하고 이를 토대로 다양한 AR 응용서비스 개발 및 기 표준기술과의 상호 운용성을 높이기 위하여 2019년 4월 'ETSI GR ARF 001, ARF; AR standards landscape' 문서를 발간하였다. <표 7-3>은 AR 응용서비스를 제공하기 위한 관련된 기존 표준을 나타낸다.

<표 7-3> AR Related Standards

ISO/ IEC JTC1	ARAF-Augmented Reality Application Format	ISO IEC 23000-13	https://www.iso.org/standard/69465.html
	MAR-RM Mixed and Augmented Reality Reference Model	ISO/IEC 18039	https://www.iso.org/standard/30824.html
OGC	ARML Augmented Reality Markup Language		https://www.opengeospatial.org/standards/arml
W3C	WebXR		https://www.w3.org/blog/tags/webxr/

■ ARAF(Augmented Reality Application Format)

ARAF는 MAR(Mixed and Augmented Reality) 경험을 제공하기 위한 것으로 종래 관련 MPEG 표준(MPEG-4 Part 1, MPEG-4 Part 16, MPEG-V)과 결합된 MPEG-4 Part 11(Scene Description and Application Engine) 표준을 토대로 설계되었다. 실사 환경하에서 2D 및 3D 멀티미디어, 다이내믹, 인터랙티브, 실사/CG 콘텐츠의 소비를 가능하게 하는 요소기술을 토대로 증강현실 응용 프로그램의 생성을 용이하게 하기 위한 저장 및 전송 포맷으로 사용될 수 있다. <표 7-4>는 AR 서비스를 위하여 사용될 수 있는 종래의 MPEG 표준을 나타낸다.

<표 7-4> AR 서비스에 적용될 수 있는 MPEG 표준

특징	표준
Image(video) & audio capturing	MPEG-V
Capture real camera position and orientation	MPEG-V & CDVS
Detection and tracking of visual objects	CDVS
Transmission of media assets	MPEG-4 Part 1, 2, 3, 10, 11, 16, 25
Image & video rendering as a background	MPEG-4 Systems

상기 MPEG 기술을 사용하여 일반적인 AR 브라우저 구성이 가능하며, 특히 ARAF 브라우저는 IP 환경하에서 사용자가 원하는 비디오 획득, 이미지와 물체 추적, 사용자 이동에 따른 카메라 위치 추적, 실사 환경에서 2D/3D 콘텐츠 다운로드 및 재현 등 요소기술을 토대로 사용자는 스마트폰 등 다양한 디바이스를 통해 최적의 AR 경험을 제공한다. <표 7-5>는 다양한 AR 서비스를 위하여, ARAF 표준에서 지원하는 구성 요소 및 범위를 나타낸 것으로 AR 프레임워크를 설계에 있어 요구되는 구성 요소의 참조 정보로 사용될 수 있다.

<표 7-5> ARAF 표준에서 지원하는 구성 요소 및 지원범위

구성 요소	지원 범위
Elementary media	Audio
	Image and video
	Texture
	Graphics
User interactivity	Time sensors
	Touch sensors
	Media sensors
	Access to sensors and actuators of physical devices
Scene graph related information	Organization of scene elements
	Navigation of scene elements
	Layouting
	Visual identification and tracking, local and remote
	Remote visual registration and composition
	Audio identification and synchronization, local and remote
User localization	GPS
Dynamic and animated scene	Interpolated and valuators
	Scripting
	Sensors
Communication and compression	Media control
	Map support

■ ARML(Augmented Reality Markup Language) 2.0

ARML은 AR 장면을 설명하고 상호 작용하기 위한 데이터 표준으로 OGC(Open Geospatial Consortium)를 통하여 개발되었다. ARML은 장면에서 가상 개체의 위치와 모양을 설명하는 XML 문법과 가상 객체의 모델 속성에 대한 동적 액세스 및 이벤트 처리를 허용하는 ECMAScript 바인딩으로 구성되며 현재 버전 2.0 표준이 발간되었다. ARML은 시각적 증강현실에 중점을 두고 있으며 오디오 또는 햅틱기반의 AR은 제외하였다.
ARML 객체 모델은 다음의 개념으로 구성된다.

- Features : 증강되는 현실 세계의 물리적 객체 표현
- Visual Assets : 증강 장면에서의 가상 객체 형태 표현
- Anchors : 물리적 객체와 가상 객체 사이의 공간적 관계 표현

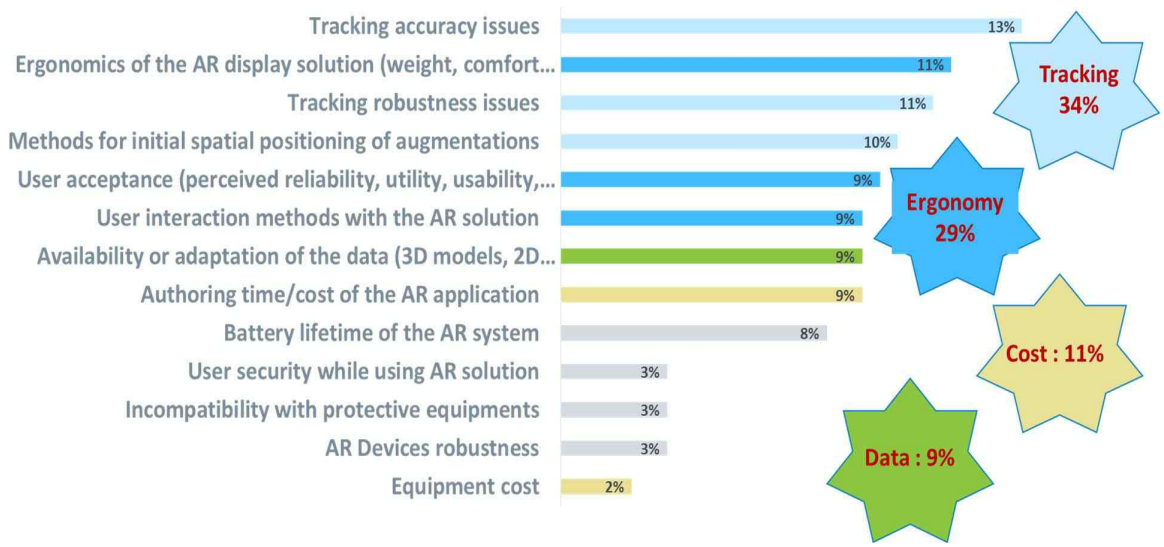
■ W3C(World Wide Web Consortium) WebXR

WebXR은 웹에서 가상 환경을 표시하거나 그래픽 이미지를 실제 환경에 오버레이하여 표현하기 위해 설계된 하드웨어에 3D 장면 렌더링을 지원하는 데 사용되는 표준이다. WebXR Device API는 가상환경을 구현하는 출력 디바이스의 선택을 관리하고, 적절한

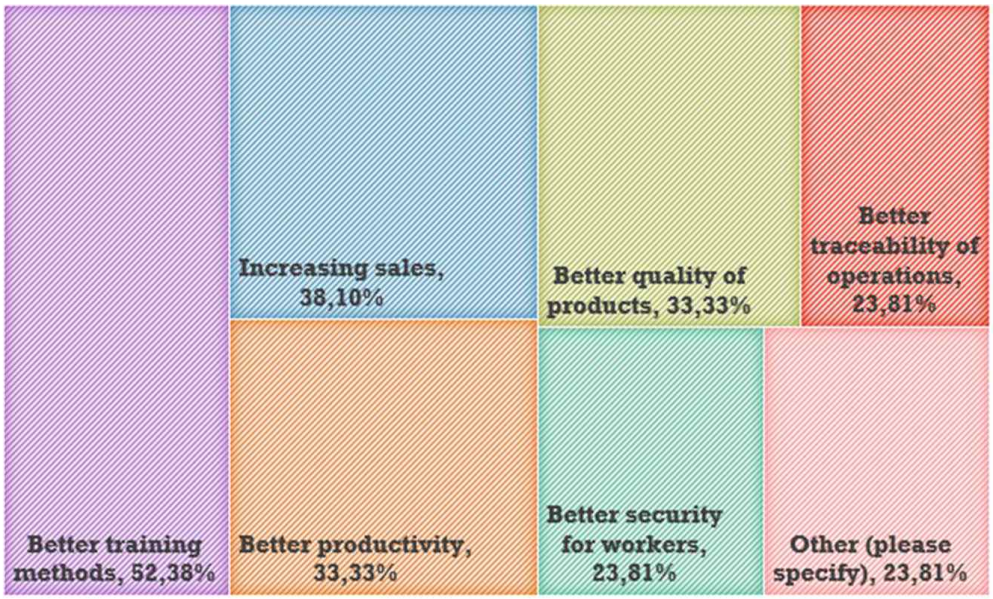
프레임 속도로 장치에 3D 장면 렌더링, 출력을 2D 디스플레이로 미러링, 입력 컨트롤의 움직임을 나타내는 벡터 생성과 같은 주요 기능을 제공한다. 가장 기본적인 수준에서 장면은 각 눈의 위치를 계산하고 그 위치에서 장면을 렌더링 하여 사용자의 각 눈의 관점에서 렌더링하기 위해 장면에 적용할 원근에 관한 내용을 계산하여 3D로 표현된다. WebXR Device API(Application Programming Interface)는 W3C 그룹에서 2018년에 처음 공개하였으며, 최근까지도 활발하게 개발 중인 API이다. 2020년 6월에는 WebXR 앵커 모듈과 WebXR 증강현실 모듈의 최신 버전이 공개되었다. 이전에는 WebVR(가상현실용)과 WebAR(증강현실용)은 API가 별도로 개발되었으나, WebXR Device API는 웹에서 가상현실 및 증강현실을 모두 지원하고, WebXR 증강현실 모듈을 통해 증강현실에 대한 다양한 기능을 지원한다. WebXR Device API는 웹 환경에서 확장 현실을 서비스하기 위해 개발된 API로, 출력 장치의 선택을 관리하며 선택된 장치에 적절한 프레임 속도로 3D 장면을 렌더링한다. WebXR 호환 장치에는 동작 및 방향 추적 기능이 있는 완전 몰입형 3D 헤드셋, 프레임을 통과하는 실제 장면 위에 그래픽을 오버레이하는 안경, 카메라로 환경을 캡처하고 컴퓨터로 해당 장면을 확대하여 현실을 강화하는 스마트폰이 포함된다.

7.2.1.2 Industrial use cases for AR applications and services(Work item 2)

ISG ARF는 다양한 AR 응용서비스에 공통적으로 사용될 수 있는 AR 프레임워크 개발을 위하여 관련된 주요 서비스 시나리오를 분석하고, 이를 토대로 AR 서비스를 위하여 요구되는 구성요소, 인터페이스 및 상호 운용성을 보장하기 위한 기술 요구사항을 도출하기 위하여 2019년 7월 'ETSI GR ARF 002, ARF; Industrial use cases for AR applications and services' 문서를 발간하였다. 다양한 서비스 시나리오 도출을 위하여 ISG ARF는 회원사를 토대로 AR 서비스를 위하여 요구되는 주요 고려 사항 및 산업에 적용될 수 있는 주요 서비스에 대하여 설문 조사를 시행하였다. (그림 7-30) 및 (그림 7-31)은 AR 서비스 시 고려되는 주요 사항들 및 AR 서비스에 따라 산업적으로 기대되는 이점을 각각 나타내며, <표 7-6>은 도출된 주요 AR 서비스 시나리오를 나타낸다.

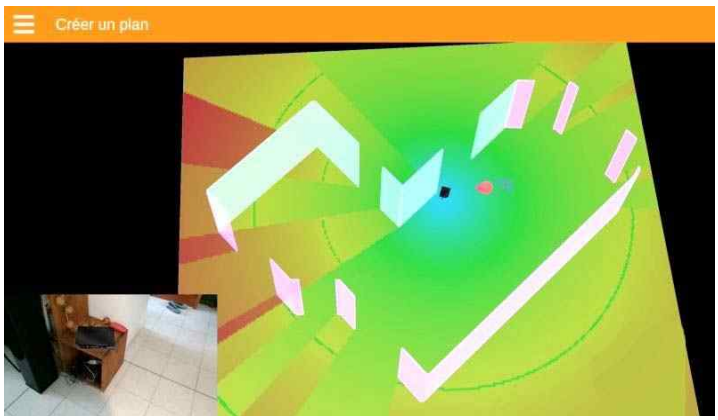


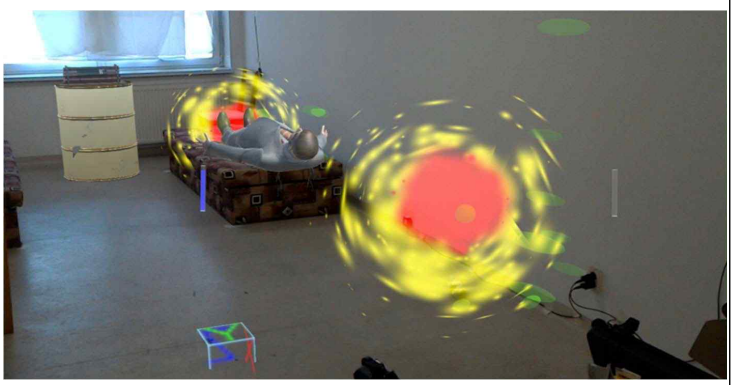
(그림 7-30) AR 서비스를 위한 주요 사항



(그림 7-31) AR 서비스를 토대로 기대되는 이점(benefits)

<표 7-6> 주요 AR 서비스 시나리오

구분	특징
Use Case 1: Wireless network and IoT installation	· 무선 네트워크는 산업체 및 가정에서 다양한 장치 연동(IoT)하기 위한 백본 망으로 사용 · 이러한 장치들은 무선 네트워크 범위내에서 안전하게 식별되고 통합되어 관리되어야 하며, 현실 세계에서의 위치 정보를 스마트하게 상황별로 정보(액세스 및 해당 장치에 대한 인터페이스) 제공 필요 · AR는 다양한 장치의 설치 및 무선 네트워크 범위 최적화를 할 수 있으며 풍부한 커넥티드(connected) 서비스 및 홈 자동화 서비스 구축을 쉽게 하는데 적용할 수 있음  (그림 7-32) AR기반 무선 네트워크 커버리지 예측 예

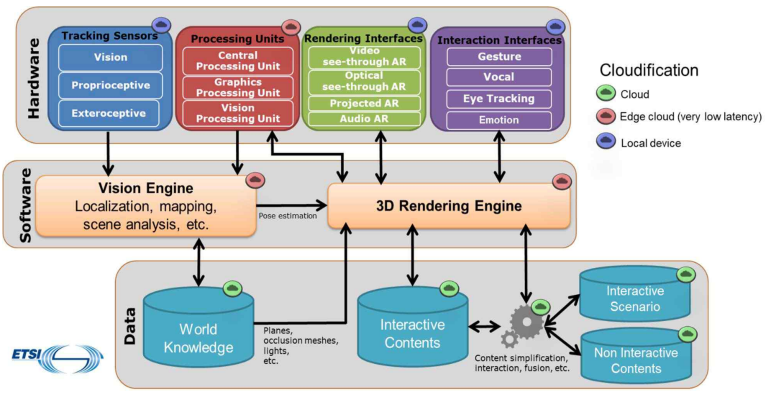
Use Case 2: Service Remote Support	· 경쟁적인 작업 환경에서 비용과 시간은 시장에서의 비즈니스와 직결되고 있으며, 초 연결 네트워크를 통하여 솔루션의 기술적 복잡성과 필요한 데이터 양이 폭발적으로 증가되고 있는 상황임 · 디지털화 된 세계에서 특정 수준의 서비스 품질과 경험 및 기술 요구사항은 보장함과 동시에 효율적으로 유지 관리 및 초 연결 네트워크기반 연결된 작업 흐름에서 가용성을 극대화하기 위한 실용적 솔루션 제공 필요 · AR은 초연결 네트워크 환경하에서 다양한 장치의 원격 제어 및 시뮬레이션 등 효율적인 리모트(remote) 작업 환경 및 유지 관리 플랫폼 구축에 적용할 수 있음  (그림 7-33) AR기반 원격 장치 제어 예
Use Case 3: Training	· AR 서비스 중 가장 많은 이점을 제공할 수 있는 서비스 시나리오로 실제 조건에 가장 가까운 훈련 수준을 제공함으로써 준비성 향상 및 실제 자원을 효율적으로 사용할 수 있음 · AR은 관련된 운영자의 안전을 보장하면서 보다 안전한 모의 실험 환경을 제공함으로써 다양한 산업체 분야에서 증강 교육, 학습, 실험에 직접적으로 적용할 수 있음  (그림 7-34) AR기반 모의 훈련 예

Use Case 4: Manufacturing	· 전 세계 주요 제조 및 교육 프로세스 내 증강현실, 클라우드 컴퓨팅, 인공지능, 사물 인터넷 등의 신기술이 등장 · 제조 공정 내에서, 제품의 개념 단계부터 재료 공급 및 워크스테이션 밸런싱에 이르기까지 생산 공장에서 가장 중요한 작업은 제품의 실제 조립 과정이며, 이를 개선하기 위한 노력이 이루어지고 있음 · AR은 Industry 4.0을 위한 10가지 전략 기술 중 하나로, 제조 공정에서 조립 시간 및 교육 환경을 개선하고 작업 오류를 최소화함으로써 생산 성과 지표를 직접적으로 증대시킬 시킬 수 있음 

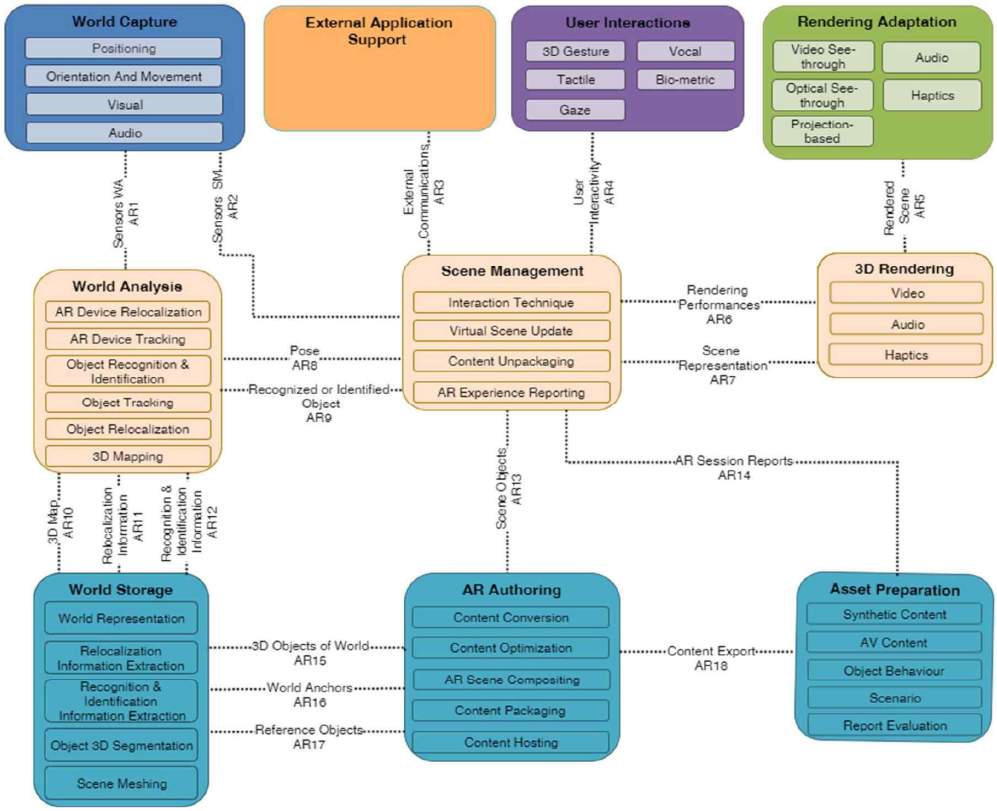
(그림 7-35) AR기반 제조 공정 예

7.2.1.3 AR framework architecture(Work item 3)

ISG ARF는 상기 제정된 'ETSI GR ARF 001' 및 'ETSI GR ARF 002'를 토대로 AR 서비스를 위하여 공통적으로 사용될 수 있는 AR 구성 요소(components), 시스템 및 서비스에 대한 AR 프레임워크를 정의하고, 각 구성 요소 간 상호 관계 및 기능에 대하여 2020년 3월 'ETSI GR ARF 003, ARF; AR framework architecture' 문서를 발간하였다. (그림 7-36)은 크게 3가지 레이어로 구성되는 전체적인 AR 프레임워크 구조를 나타내며, (그림 7-37)은 AR 프레임워크 기반의 AR 시스템 구현을 위한 각 요소 간의 기능 및 상관관계를 나타낸다.



(그림 7-36) AR 프레임워크 구조



(그림 7-37) AR 요소 기능 및 상관관계

■ Hardware Layer

- Tracking Sensors: 실제 물리적 공간에 가상 객체를 올리기 위하여 실시간으로 공간의 위치 및 방향을 추적하기 위한 센서. 대부분 AR 시스템(ex. 스마트 폰 또는 AR 안경)은 하나 이상의 비전 센서(RGB 카메라, GPS, 깊이센서 등)가 내장
- Processing Units: 영상기반 컴퓨터 비전, 기계 학습 기반 추론, 3D 렌더링 등 공간 인지 및 분석을 위하여 요구되는 공간컴퓨팅 처리 구조(AR 시스템에 내장 또는 원격으로 지원)
- Rendering Interface: 렌더링 SW에 의하여 생성된 AR 객체를 각 AR 시스템의 고유한 특성에 맞게 변환하여 재현하기 위한 인터페이스
- Interaction Interface: 제스처, 위치 추적 등 다양한 사용자 상호작용 정보를 추출하고 분석하기 위한 인터페이스

■ Software Layer

- Vision Engine: AR 객체를 현실 공간과 융합을 목적으로 현실 세계의 3D 표현 재구성 및 분석(localization), 현실 세계 인지 및 추적, 현실 공간에 상대적인 AR 객체 위치 및 방향 정보 등 렌더링을 위하여 필요한 모든 정보를 분석 및 생성하기 위한 엔진
- 3D Rendering Engine: 사용자 상호작용, AR 객체 행동 표현 등에 따라 현실 세계를 기반으로 AR 객체를 실시간으로 업데이트 및 각 AR 시스템에 맞는 형태의 변환/재현하기 위한 엔진

■ Data Layer

- World Knowledge: 상기 Vision Engine 또는 외부에서 생성된 것으로 3D 표현 재구성 및 분석, 렌더링을 위한 부가 데이터 등 전반적인 AR 재현하기 위한 정보
- Interactive Contents: 사용자 상호작용에 따라 실시간으로 동작하기 위하여 생성된 가상 콘텐츠(3D 모델, 동적 애니메이션 콘텐츠 등)

■ World Capture

AR 시스템 또는 실제 물체의 위치 및 방향을 인지하고 분석에 관련된 비디오, 오디오, 위치정보 등 요구되는 정보 전달을 목적으로 현실 세계에 AR 객체의 정확한 등록을 위한 현실 세계 정보 수집 기능을 수행한다.

■ World Analysis

입력된 다양한 형태의 정보로부터 현실 세계에 존재하는 요소 식별 및 움직임 추정, 3D 공간 재구성, 객체 추적 등 현실 세계에 안정적인 AR 객체 배치 및 재현을 위한 현실 세계 분석 기능을 수행한다.

■ World Storage

현실 세계 분석을 토대로, 지속해서 변하는 현실 세계 정보를 업데이트하는 데 필요한 정보(3D 객체 인식 및 식별, 객체 추출(3D segmentation), AR 객체 저작, 위치 변화에 따른 3D 공간 데이터 등) 전달 기능을 수행한다.

■ Asset Preparation

AR 공간 구성 및 다양한 AR 객체 속성을 표현하기 위한 멀티미디어 콘텐츠 전달 기능을 수행한다.

■ AR Authoring

AR 장면 편집 및 구성, AR 객체 상호 관계 정의 및 애니메이션 등 다양한 AR 장면 구성 및 특정 포맷으로의 패키징(ex. ARAF 등) 기능을 수행한다.

■ User Interactions

제스처 인식, 음성 인식, 시선 추적 등 다양한 사용자 상호작용 디바이스를 통하여 생성된 정보 전달 기능을 수행한다.

■ Scene Management

AR 프레임워크 구조 기능 중 핵심 기능으로, 다른 모든 기능과 연동됨과 동시에 사용자 상호작용 또는 요구에 따라 관련된 AR 장면 구성 및 표현, 실시간 업데이트 등 장면 유지 기능을 수행한다.

■ 3D Rendering
전체 또는 부분적 AR 장면 복원 및 재현 기능을 수행한다.

■ Rendering Adaptation
3D Rendering에 의해 생성된 영상, 오디오 및 진동 등을 재현하기 위한 물리적 렌더링 장치를 의미한다.

7.2.1.4 Interoperability Requirements for AR components, systems and services (Work item 4)

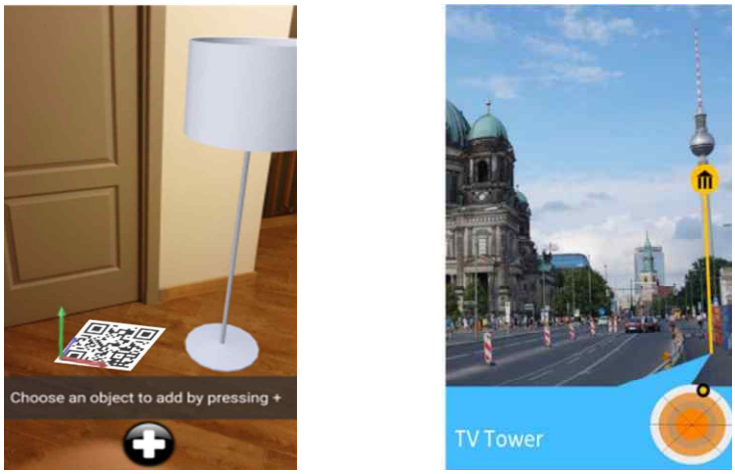
상기 'ETSI GR ARF 003'은 AR 구성요소, 시스템 및 서비스를 위한 상호운용성 프레임워크 구조를 정의함과 동시에 AR 솔루션에 필요한 주요 구성 요소와 인터페이스를 기술하고 있다. 이는 다양한 AR 시스템 간의 상호운용성을 지원하기 위한 첫 번째 단계이며, 두번째로 주요 기능 및 인터페이스에 대한 구체적인 상호운용성 요구사항을 정의한다. 이에, ISG ARF Work item 4를 아래와 같이 3가지의 파트로 분류하여 표준화를 진행 중에 있으며, 2021년 8월 'ETSI GS ARF 004-2, ARF; Interoperability Requirements for AR components, systems and services Part 2: World Storage and AR Authoring functions' 표준을 발간하였다.

- Part 1: Overview
- Part 2: World storage and AR authoring functions
- Part 3: Sensors for world capture

'ETSI SG ARF 004-2' 표준은 (그림 7-37) 내 AR authoring과 World Storage 사이의 'World Anchors'(AR16) 및 'Reference Objects'(AR17)에 대한 상호운용성 요구사항을 정의한다.

■ World Anchors
World Anchor는 현실세계에 포함되는 하나 이상의 요소와 관련하여 고정된 위치를 나타낸 것으로, 각 AR asset(현실 공간 내 위치되는 3D 모델, 이미지, 비디오, 텍스트, 음원 등 모든 종류의 콘텐츠)은 공간상 표현하기 위하여 기준이 되는 좌표계 참조 시스템을 가진다.

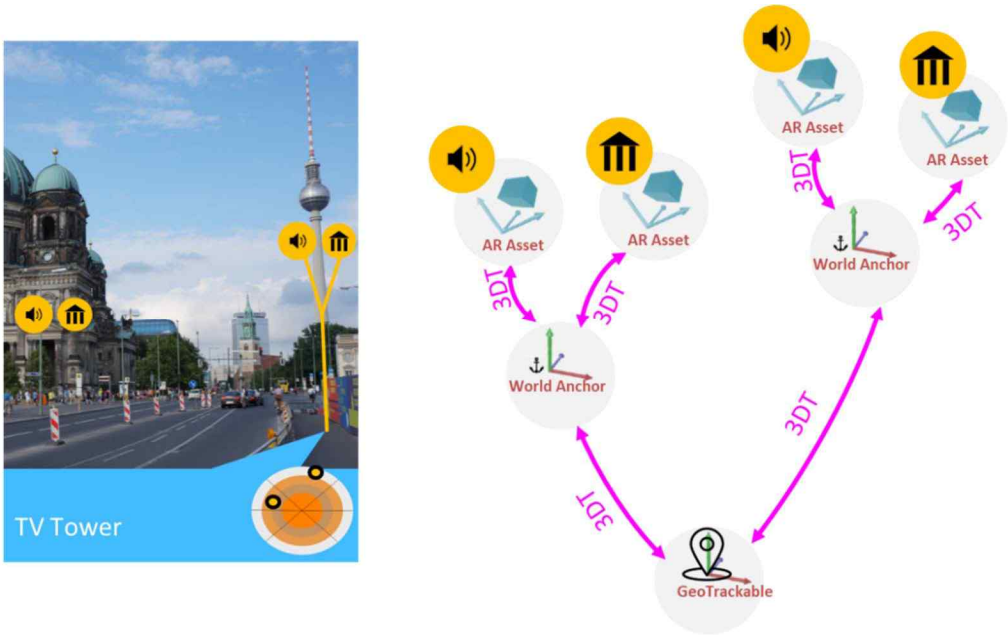
- World Anchor는 실제 세계의 좌표계 참조 시스템을 정의해야 한다
- World Anchor는 좌표계 참조 시스템 내에서 0개 이상의 AR asset을 가져야 한다
- World Anchor는 고유의 식별자를 가져야 한다



(그림 7-38) World Anchor 예(좌:Mark기반, 우:GNSS[3]기반)

■ AR Authoring

- AR Authoring은 AR 장면에서 World Anchor 포즈 추정과 관련된 특정 Trackable[4] 세트를 정의해야 한다
- AR Authoring은 World Anchor 사이의 논리적 관계(World Graph)를 정의하는 구조를 제공해야 한다
- AR Authoring은 World Anchor 및 AR Asset 사이의 상대적인 위치와 방향을 유지해야 한다



(그림 7-39) World Anchor 사이 및 AR Asset 간의 논리적 관계 예[5]

3) GNSS: Global Navigation Satelite System
4) Trackable: 추출할 수 있는 실제 세계의 요소
5) 3DT: 3D Transform은 Trackable, World Anchor 및 AR Assets 간의 상대적인 위치와 방향 정의

■ World Storage

- World Storage는 World Anchor의 고유의 식별자를 제공해야 한다
- World Storage는 Trackable의 고유의 식별자를 제공해야 한다
- World Storage는 Trackable과 World Anchor의 상대적인 위치를 제공 및 유지해야 한다
- World Storage는 World Graph를 지속적으로 체크해야 한다
- World Storage는 World Anchor를 토대로 AR 장치의 포즈를 추정하는 기능을 제공해야 한다

7.2.2 ISG ARF 표준화 전망

ETSI는 2017년 12월부터 서로 다른 AR 구성 요소 간의 안정적인 상호 운용성(Interoperability)을 지원함과 동시에 AR 응용 서비스의 성공적인 시장 진입을 위한 핵심 기술로 다양한 AR 응용 서비스 아래 관련 요소 간의 프레임워크 구조 및 기술에 대한 표준화를 진행하고 있다. 현재 ISG ARF는 Work item 4 내에서 세부적인 기능에 대한 상호운용성 요구사항에 대한 표준화를 수행하고 있으며, SC29/WG7(MPEG 3D Graphics Coding) 표준화 단체와 정보공유(2021년 5월, SC29/WG7은 MPEG-I 비디오 기반 포인트 클라우드 압축(V-PCC), MPEG-I Geometry 기반 포인트 클라우드 압축(G-PCC), 미디어 사물 인터넷(IoMT) 및 MPEG-V 표준화 현황 및 정보 관련 Liaison 송부)를 토대로 AR 프레임워크에 대한 시장 활성화를 모색하고 있다.

부 록 Ⅰ-1

(본 부록은 기술보고서를 보충하기 위한 내용으로 기술보고서의 일부는 아님)

지식재산권 확약서 정보

해당 사항 없음

부 록 Ⅰ-2

(본 부록은 기술보고서를 보충하기 위한 내용으로 기술보고서의 일부는 아님)

시험인증 관련 사항

해당 사항 없음

부 록 Ⅰ-3

(본 부록은 기술보고서를 보충하기 위한 내용으로 기술보고서의 일부는 아님)

본 기술보고서의 연계(family) 표준

해당 사항 없음

부 록 Ⅰ-4

(본 부록은 기술보고서를 보충하기 위한 내용으로 기술보고서의 일부는 아님)

참고 문헌

[5.1. 몰입형 미디어 서비스 발전 방향]

[1] 과학기술정책연구원 Future Horizon 제49호, “미래연구 포커스, Metaverse, 가상과 현실의 경계를 넘어”, 2021.

[2] 정보통신기획평가원, ICT Spot Issue 2021-11호, “디지털 전환의 핵심, ‘메타버스’ 르네상스”, 2021.8.

[3] KIET 산업경제, “다가오는 메타버스 시대, 차세대 콘텐츠산업의 방향과 시사점”, 2021.05.

[5.2. 메타버스 서비스 동향]

[1] 소프트웨어정책연구소 ISSUE REPORT IS-115, “로그인(Log In) 메타버스: 인간x공간x시간의 혁명”, 2021.

[2] https://www.naverz-corp.com/

[3] https://www.roblox.com/

[4] https://www.minecraft.net/

[5] https://www.epicgames.com/fortnite/

[6] https://www.oculus.com/facebook-horizon/

[7] https://www.facebook.com/

[8] https://www.nintendo.co.kr/software/switch/acbaa/

[9] https://hopin.com/

[10] https://www.teooh.com/

[11] https://www.dob.world/

[12] http://sidus-x.com/

[13] https://live.lge.co.kr/virtual_influencer_reah/

[6.1. AR/VR/XR 기술]

[1] 한국인터넷진흥원 2021 KISA REPORT Volume 02, “메타버스를 위한 소프트웨어 플랫폼‘, 2021.

[2] https://developers.google.com/ar

[3] https://developer.apple.com/kr/augmented-reality/arkit/

[4] https://unity.com/kr/unity/features/arfoundation

[5] https://aws.amazon.com/ko/sumerian/

[6] https://nianticlabs.com/blog/nrwp-update-110619/

[7] https://sparkar.facebook.com/ar-studio/

[8] https://www.unrealengine.com/ko/metahuman-creator

[9] https://spatial.io/

[10] https://www.oculus.com/quest-2/?locale=ko_KR
[11] https://www.microsoft.com/en-us/hololens
[12] https://www.kat-vr.com/
[13] https://www.cybershoes.com/us/
[14] https://ektovr.com/

[6.2. LF 영상 기술]

[1] http://lightfield-forum.com/
[2] https://www.insta360.com/
[3] https://www.z-cam.com/
[4] https://www.kandaovr.com/
[5] R. S. Ryan, et al., “A system for acquiring, processing, and rendering panoramic light field stills for virtual reality,” ACM Transactions on Graphics, Vol. 37, No. 6, Article No. 197, 2018.11.
[6] https://www.blog.google/products/google-ar-vr/experimenting-light-fields/
[7] https://augmentedperception.github.io/lowcost-panoramic-LFV/
[8] M. Broxton, et al., “A low cost multi-camera array for panoramic light field video capture,” SIGGRAPH Asia 2019 Posters, Article No. 25, 2019.11.
[9] M. Broxton, et al., “Immersive light field video with a layered mesh representation,” ACM Transactions on Graphics, Vol. 39, No. 4, Article No. 86, 2020.7.
[10] https://www.facebook.com/Facebook360
[11] https://facebook360.fb.com
[12] A. P. Pozo, et al., “An integrated 6DoF video camera and system design,” ACM Transactions on Graphics, Vol. 38, No. 6, Article No. 216, 2019.11.
[13] 정원식, “Light Field 미디어 기술 개발 및 표준화 동향”, 방송과미디어 학회지, 2018. 1.
[14] https://www.youtube.com/watch?v=cAg0A9gld5c&feature=youtu.be&ab_channel=IDEAImmersiveAlliance.
[15] 정준영, “플렌옵틱 영상 서비스를 위한 적응적 저지연 전송 기술”, 방송과미디어 학회지, 2021.1.
[16] Wijnants, Maarten, et al. "Standards-compliant HTTP adaptive streaming of static light fields." Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology, pp. 1-12, 2018. 11.
[18] Broxton, Michael, et al. "Immersive light field video with a layered mesh representation." ACM Transactions on Graphics, vol. 39. No. 4, 2020. 7.
[19] Kara, Peter A., et al. “Evaluation of the concept of dynamic adaptive streaming of light field video.” IEEE Transactions on Broadcasting vol. 64. No. 2, pp. 407-421, 2018. 6.

[20] Cserkaszky, Aron, et al. "Real-time light-field 3D telepresence." 2018 7th European Workshop on Visual Information Processing, pp. 1-5, 2018. 11.
[21] R. S Overbeck et al., “A system for acquiring, processing, and rendering panoramic light field stills for virtual reality,” ACM Trans. Graph. 2018.
[22] J. Flynn et al., “Deepview: View synthesis with learned gradient descent,” Proc. Conference on Computer Vision and Pattern Recognition, 2019.
[23] M. Broxton et al., “Immersive light field video with a layered mesh representation,” ACM Trans. Graph. 2020.

[6.3. Volumetric 콘텐츠 기술 동향]

[1] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. Kirk, and S. Sullivan, High-Quality Streamable Free-Viewpoint Video, in SIGGRAPH, 2015.
[2] https://8i.com/
[3] https://newsroom.intel.com/press-kits/intel-studios/
[4] https://global.canon/en/vvs/
[5] SKT 점프 스튜디오, 혼합현실 기술로 한층 더 진화하는 K팝 열풍, https://www.jumpstudio.co.kr/news-detail?newsSeq=11, 2021.
[6] LG유플러스, 세계 최초 4K 화질 3D AR 콘텐츠 제작, http://www.uplus.co.kr, 2019.
[7] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid, BodyNet: Volumetric Inference of 3D Body Shapes, in ECCV, 2018.
[8] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and Hao Li. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In Proceedings of the International Conference on Computer Vision (ICCV), 2019.
[9] S. Saito, T. Simon, J. Saragih, and H. Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
[10] V. Gabeur, J.-S. Franco, X. Martin, C. Schmid, and G. Rogez, Moulding Humans: Non-parametric 3D Human Shape Estimation from a Single Image", in ICCV, 2019.
[11] T. He, J. Collomosse, H. Jin, and S. Soatto, Geo-PIFu: Geometry and Pixel Aligned Implicit Functions for Single-view Human Reconstruction, in NIPS, 2020.
[12] Y. Hong, J. Zhang, B. Jiang, Y. Guo, L. Liu, and H. Bao, StereoPIFu: Depth Aware Clothed Human Digitization via Stereo Vision. In Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[13] Y. Jafarian and H. S. Park, Learning High Fidelity Depths of Dressed Humans by Watching Social Media Dance Videos, in CVPR, 2021.
[14] Y. Lizhen, Z. Xiaochen, Y. Tao, W. Songtao, and L. Yebin, Z, NormalGAN: Learning Detailed 3D Human from a Single RGB-D Image, in ECCV, 2020.
[15] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu, DeepHuman: 3D Human Reconstruction from a Single Image, in ICCV, 2019.

[7.1. MPEG-I Visual 표준화 동향]

[1] ISO/IEC JTC1/SC29/WG4 135, “Test Model 10 for MPEG Immersive Video”, July, 2021.
[2] Text of ISO/IEC FDIS 23090-12 MPEG Immersive Video”, July, 2021.
[3] ISO/IEC JTC1/SC29/WG4 135, “Common Test Conditions for MPEG Immersive Video”, July, 2021.

[7.2. ETSI ISG ARF 표준화 동향]

[1] https://www.etsi.org/committee/1420-arf
[2] ETSI GR ARF 001 V1.1.1 ARF; AR standards landscape, April. 2019.
[3] ETSI GR ARF 002 V1.1.1 ARF; Industrial use cases for AR applications and services, July. 2019.
[4] ETSI GR ARF 003 V1.1.1 ARF; AR framework architecture, March. 2020.
[5] ETSI GR ARF 004-2 V1.1.1 ARF; Interoperability Requirements for AR components, systems and services Part 2: World Storage and AR Authoring functions, Aug. 2021.
[6] ISO/IEC JTC1/SC29/WG7 123, “Liaison statement to ETSI ARF on MPEG-I Video-based Point Cloud Compression, Geometry-based Point Cloud Compression, Internet of Media Things and MPEG-V”, April, 2021.

※ 상기 기재된 참고 문헌의 발간일이 기재된 경우, 해당 표준(문서)의 해당 버전에 대해서만 유효하며, 연도를 표시하지 않은 경우에는 해당 표준(권고)의 최신 버전을 따름

부 록 Ⅰ-5

(본 부록은 기술보고서를 보충하기 위한 내용으로 기술보고서의 일부는 아님)

영문기술보고서 해설서

해당 사항 없음

부 록 Ⅰ-6

(본 부록은 기술보고서를 보충하기 위한 내용으로 기술보고서의 일부는 아님)

기술보고서의 이력

판수	채택일	기술보고서번호	내용	담당 위원회
제1판	2021.11.29	제정 FBMF-TR-009	몰입형 비디오 기술 및 표준화 동향	미래미디어 분과위원회
오류정정				
오류정정				
제2판				