프로그래밍 방식의 객체 기반 영상 콘텐츠 제작 기술 동향

Trends in Programmable Object-Based Content Production Technologies

이재영 (J.-Y. Lee, jaeyl@etri.re.kr) 미디어방송연구실 책임연구원
김태원 (T.O. Kim, kimm@etri.re.kr) 디지털홀로그래피연구실 선임연구원
추현곤 (H.-G. Choo, hyongonchoo@etri.re.kr) 실감미디어연구실 책임연구원
이한규 (H.K. Lee, hkl@etri.re.kr) 미디어지능화연구실 책임연구원
석왕헌 (W.H. Seok, whseok@etri.re.kr) 지능화정책연구실 선임연구원
강정원 (J.W. Kang, jungwon@etri.re.kr) 미디어부호화연구실 책임연구원
허남호 (N.H. Hur, namho@etri.re.kr) 미디어방송연구실 책임연구원
김흥묵 (H.M. Kim, hmkim@etri.re.kr) 미디어연구본부 책임연구원/본부장

ABSTRACT

With the rapid growth in media service platforms providing broadcast programs or content services, content production has become more important and competitive. As a strategy to meet the diverse needs of global consumers for a variety of content and to retain them as long-term repeat customers, global over-the-top service providers are increasing not only the number of content productions but also their production efficiency. Moreover, a considerable amount of scene composition in the flow of content production work appears to be combined with rendering technology from a game engine and converted to object-based computer programming, thereby enhancing the creativity, diversity, quality, and efficiency of content production. This study examines the latest technology trends in content production such as virtual studio technology, which has emerged as the center of content production, the use cases in various fields of artificial intelligence, and the metadata standards for content search or scene composition. This study also examines the possibility of using object-based computer programming as one of the future candidate technologies for content production.

KEYWORDS 가상 스튜디오, 객체, 렌더링, 메타데이터, 영상 기반 모델링, 인공지능, 장면 구성

* DOI: https://doi.org/10.22648/ETRI.2022.J.370408
* 본 기술동향 보고서는 미디어연구본부 내 "POBM기획-TF" 활동의 최초 결과물임.
* 이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임[2017-0-00081, 초고품질 UHD (UHQ) 전송 기술 개발].

Ⅰ. 서론

오늘날 콘텐츠를 전달하는 플랫폼의 다양화로 콘텐츠 자체의 중요성이 점점 더 부각되고 있으며, 콘텐츠는 플랫폼 선택의 지배적인 요인 중 하나로 작용하고 있다. 그간 짜임새 있는 이야기를 기반으로 흥행한 콘텐츠들은 소수 전문가에 의해 기획되고 장기간 현지 촬영과 후반작업을 거쳐 제작되어왔으며, 대작의 경우 엄청난 규모의 예산, 인력, 장비, 시간을 투입하였다. 최근 스마트폰 카메라와 액션 캠, 자동 편집기술 등 촬영 기기와 편집 소프트웨어의 보급, 유튜브와 같은 동영상 공유 플랫폼에서의 수많은 개인 창작자의 성공 사례가 나오면서 '미디어 민주화'가 언급되는 등 콘텐츠 창작과 소비의 영역에서 큰 변화가 일고 있다.

일례로 스마트폰 카메라로 이벤트 현장에서 HDR 비디오를 실시간으로 촬영해 원격 저장소에 업로드하면 인공지능 기반의 편집 툴이 특정 인물(객체), 시간, 위치와 같은 제어신호 입력에 따라 자동으로 편집한 후 여러 플랫폼에 바로 공유할 수 있는 시대가 되었다[1]. 아직은 편집 툴의 기능이나 품질이 충분하지 않을 수 있으나 연구개발이 지속된다면 기능과 품질이 크게 향상될 것으로 보인다. 이는 일반인으로 하여금 콘텐츠 제작을 보다 수월하게 하고, 품질도 개선함으로써 개인 창작자의 콘텐츠 제작 수준을 전문가 수준으로 끌어올리게 될 것이다. 또한 카메라로 촬영한 실시간 비디오뿐만 아니라 저장매체나 인터넷에서 필요한 장면이나 객체를 고속으로 검색하여 추출한 후 새로운 장면을 재구성하는 데 편리하게(컴퓨터 프로그래밍을 하는 것처럼 어떤 규칙에 따라 손쉽게) 활용할 수 있다면 편집 툴에 대한 개인 창작자의 수용도가 한층 높아질 것이다[2]. 이와 더불어 메타버스와 같은 가상세계로까지 개인 창작물의 수익 실현이 확장되면 개인 창작자를 더 넓은 창작의 세계로 유입시켜, 창작–수익–창작의 선순환 고리를 만들어내는 촉매제로 작용할 수도 있다.

또 다른 변화 중 눈에 띄는 것은 방송사 중심의 콘텐츠 제작이 점점 원격 스튜디오나 가상 스튜디오 중심으로 옮겨가고 있다는 점이다[3]. 방송사 스튜디오와 원격 스튜디오가 5세대 이동통신망이나 초고속망으로 연결되면서 실시간 콘텐츠 제작에 대한 실증도 이뤄지고 있으며, 그 가능성을 점점 높이고 있다[4]. 가상 스튜디오는 LED Wall, 게임 엔진을 탑재한 고속 렌더러, 카메라, 조명, 무대로 기본 구성이 되는데 통합제어시스템은 이들을 정밀하게 동기화한다[5]. 또한 촬영한 비디오에 대한 러프 컷(Rough Cut)은 장면의 전개 상황을 바로 확인하고 조정할 수 있는 수준에 이르렀다. 이로 인해 가상 스튜디오 콘텐츠 제작에서는 배경이나 객체를 빠르게 변경해 장면을 재구성할 수 있게 됨으로써 콘텐츠 제작 시간과 비용을 현저하게 줄일 수 있게 되었으며, 최근 콘텐츠 제작 방식의 변화를 주도하고 있다.

이처럼 콘텐츠 제작 기술이 날로 고도화되면서 향후 표준화된 콘텐츠 제작 기법을 학습한 인공지능 에이전트까지 등장한다면 입력신호와 제어 프로그램에 따라 콘텐츠를 자동으로 대량 생산하는 시대가 도래할지도 모르겠다[4,6]. 조선시대 신숙주가 쓴 《보한재집(保閑齋集)》의 〈화기(畫記)〉에 "무릇 그림이란 반드시 천지의 조화를 자세히 살피고 음양의 운행을 파악하여 만물의 성정(性情)과 사리의 변화를 가슴속에 새긴 연후에 붓을 잡고 화폭에 임하면 신명(神冥)과 만나게 되어, 산을 그리고자 하면 산이 보이고 물을 그리고자 하면 물이 보이며 무엇이든 붓으로 그대로 나타내니 가상(假像)에서 참모습이 나타나게 된다. 이것이 화가의 법이다."라는 글이 실려 있다고 한다[7]. 수백 년이 흘러 콘텐츠 창작의 영역에서도 민주화의 바람이 부는 지금 신숙

주의 글을 다음과 같이 각색할 수도 있겠다 싶다. “무릇 ‘콘텐츠’란 반드시 창작의 본질을 이해하고 제작 방법을 가슴속에 새긴 연후에 제작 툴을 이용해 컴퓨터 화면에 임하면 신명과 만나게 되어, ‘산’이라는 객체는 인터넷에서 검색하고, ‘물’이라는 객체는 로컬서버에서 호출해 ‘자연’이라는 어떤 장면을 맘껏 구성하니 가상에서 참모습이 나타나게 된다. 이것이 미래 콘텐츠 ‘창작자’의 법이다.”라고 말이다.

본고에서는 가상 스튜디오에서의 객체 기반, 컴퓨터 프로그래밍 방식의 콘텐츠 제작 방식으로 진화해가는 최신 콘텐츠 제작 기술의 동향을 살펴보고 향후 어떤 방향으로 나아갈지 간단하게 전망해 보고자 한다.

II. 최신 콘텐츠 제작 기술 동향

1. 가상 스튜디오 제작 기술의 프로그래밍화

앞서 서론에서도 언급한 바와 같이 콘텐츠 제작 기술은 급속하게 발전하고 있으며, 특히 최근 OTT 시장의 활성화, VFX, 인공지능(AI), 하드웨어 기기 등과 같은 관련 기술의 발전에 따라 가상 스튜디오 제작 기술의 활용 및 그 중요성은 날로 확대되고 있다.

해외에서는 이미 오래전부터 콘텐츠 제작 후반 작업에 CG를 활용하여 왔다. 이후 가상 스튜디오 제작 기술을 도입하기 시작하면서 실시간으로 가상 환경을 제작 현장에 활용할 수 있게 됨에 따라, 콘텐츠의 품질뿐만 아니라 제작 비용과 공간의 효율성도 향상시킬 수 있게 되었다. 영화 〈아바타(2009)〉는 가상 스튜디오 제작 기술을 본격적으로 활용한 최초의 사례로 꼽히고 있으며, 제작의 많은 부분을 그린 스크린 혹은 블루 스크린 기반의 VFX 장면으로 제작하였다[8,9]. 이후 영화 〈레디 플레이어 원(2018)〉, 〈웰컴 투 마웬(2018)〉 등에서는 실시간 렌더링이 가능한 게임 엔진을 도입하기 시작하였으며, 이를 활용하여 실제 세계와 가상 세계의 동시 촬영이 가능해짐에 따라 실시간 가상 스튜디오 제작의 효율성을 한 단계 끌어올려 주었다. 특히 가상 스튜디오 제작 전문 기업인 The Third Floor는 사전 시각화(Pre-Visualization) 기술을 도입하여, 실제 제작 현장과 연동되는 가상 공간을 사전 제작하고, 가상 스튜디오 제작을 실시간으로 수행하는 기술을 제공하고 있다[8]. 이 밖에도 글로벌 기업인 Netflix, Amazon, Bron Studio, Weta Digital 등은 최근 글로벌 OTT 시장의 폭발적인 성장에 맞추어 각자의 기술력을 활용한 가상 스튜디오 제작 환경을 구축하고 있다[10].

국내에서도 글로벌 제작 현장에서 적용되던 가상 스튜디오 제작 및 VFX 기술을 2020년부터 도입하기 시작하였으며, 국내 콘텐츠의 글로벌 시장 진출에 힘입어 투자가 본격적으로 확대되고 있다. 위지윅스튜디오는 국내 최초의 가상 스튜디오인 XON을 2020년에 설립하였으며, 2021년 3월 신규 XR 스테이지를 추가로 오픈하였다. CJ ENM은 삼성전자, 에픽게임즈와 파트너십을 맺고 경기도 파주에 국내 최대 규모의 가상 스튜디오를 설립 중에 있으며, 500평 규모의 공간에 LED Wall과 VFX 장비 등을 갖추어 다양한 콘텐츠 제작에 활용할 예정이다. 이 밖에도 덱스터, 자이언트 스텝이 실감 콘텐츠 제작을 위한 가상 스튜디오 구축 및 투자를 발표한 바 있다[11,12].

가상 스튜디오 제작을 위해 최근 활용되고 있는 주요 시스템과 기술은 다음과 같다[13].

- **모션 캡처 카메라 시스템**: 가상 카메라와 제어 시스템으로 구성되며, 스튜디오 내부 곳곳에 설치된 센서들을 활용하여 가상 카메라가 찍는 장면을 실시간으로 추적하는 기능을 갖는다.

- **디지털 액터 기술**: 실제 배우와 같은 수준의 캐릭터 및 장면을 연출할 수 있는 CG 기술로서 모션 캡처 카메라 시스템을 통해 획득한 실제 배우의 움직임을 CG 캐릭터로 변환하여 활용한다.
- **실시간 렌더링 기술**: 본래 게임 엔진에서 발전된 기술로써 CG로 연출된 3차원(3D) 캐릭터 및 장면을 화면을 통해 실시간으로 구현하는 기술이다. 현재의 가상 스튜디오 제작 방식에서는 모션 캡처 카메라와 디지털 액터 기술을 통해 재생된 장면이 화면상에 어떻게 반영되는지 실시간으로 확인이 가능하여, 후반작업에 소요되는 시간과 비용을 획기적으로 단축할 수 있다.
- **실시간 LED Wall 기술**: CG 기반으로 제작된 배경 이미지를 LED Wall에 투영시켜 카메라에서 얻은 장면과 LED Wall의 배경을 실시간으로 합성하여 재생할 수 있다. 기존의 그린이나 블루 스크린을 이용하는 방식보다 실시간성을 높여주어 제작 결과물의 불확실성을 줄여줄 수 있으며, 또한 보다 사실감을 갖는 배경 제작이 가능하다.

2. 객체 기반 콘텐츠 생성 기술 동향

가상 스튜디오 제작의 모션 캡처, 디지털 액터, 실시간 렌더링 기술은 CG/CV 기술을 복합적으로 활용한 고도의 사실적 콘텐츠 생성 기술에 속한다. 최근의 가상 스튜디오 제작 기술로 만들어진 영화 장면들은 전문가의 눈에도 실제 세계와 가상 세계를 구분하기 어려운 수준에 도달한 상황이다. 향후 메타버스와의 연계를 고려한 가상 스튜디오 기술의 활용 가능성은 무궁무진하다고 볼 수 있다. 이 기술을 가능케 하는 여러 핵심기술 중 실제 세계의 3D 정보 획득을 위한 CV 기술과 획득된 3D 정보에 기반하여 사실적인 렌더링을 수행하는 CG 기술이 복합된 IBMR 기술에 대해 알아볼 필요가 있다. 최근에 새로운 양상으로 진화된 DL 기술과 결합된 IBMR 기술은 직전까지 보여준 기술의 한계를 극복하고 보다 사실적이고 자연스러운 콘텐츠 생성에 활용할 수 있는 가능성을 보여주고 있는데, 이에 대한 기술 동향을 간단히 소개하고자 한다.

DL 적용 이전의 IBMR 기술은 기존 CG 모델에 기반한 사실적 렌더링 기법의 한계를 극복하기 위해 실사 RGBD 영상 또는 비디오로부터 물체의 3D 정보, 표면 반사(Reflectance) 정보, 장면(Scene) 조명(Illumination) 정보, 카메라 기하 정보 등을 추출한 후에 이를 이용하여 사실적 영상 렌더링을 하는 방식이었다. 이로 인해 보다 사실적 CG 렌더링을 위한 방대한 양의 컴퓨팅 파워 없이도 사실적 영상 실시간 렌더링이 가능하게 되었다. 그러나, 획득 또는 추정된 깊이 정보의 정확도에 따라 3D 모델의 품질이 결정되고, 새로운 시점의 경우 카메라 기하 정보와 깊이 정보의 정확도에 쉽게 영향을 받는 근본적인 문제점이 있었다[14-16]. IBMR 방식 이외에도 3D 모델 복원 없이 라이트 필드(Light Field)[17] 정보를 활용해 새로운 시점의 영상을 합성하는 방식도 하나의 주요 연구 흐름으로 떠올랐으나, 정확한 깊이 정보가 없어 합성 영상의 품질이 떨어지는 문제가 있었다.

다음은 DL 기법을 적용해 합성 영상의 품질을 개선하는 방식에 대해 소개하고자 한다. 먼저 SPS는 의미론적(Semantic) 영상 학습을 통해 구성된 DL 네트워크에 사용자가 구성하고자 의미론적 구성(Semantic Layout)을 입력으로 주면 그에 해당하는 새로운 영상을 합성하는 기술로 네트워크 내부적으로 CN을 채용하여 입력 영상을 자동으로 분할(Segmentation)시켜 학습한 후 새로운 영상 합성을 가능하게 한다[18]. 영상으로부터 분할 정보와 더불어 시간의

흐름에 따른 장면의 특성 정보(Cloudy/Clear/Rainy/Day/Night 등)를 함께 학습하고, 의미론적 구성과 더불어 제어 정보를 입력으로 받아 사용자가 원하는 장면의 특성에 부합하는 영상을 합성하는 방법도 개발되었다[19]. 또한 기존 방식의 낮은 영상 해상도 문제는 적층 네트워크(Cascaded Network)를 사용하여 개선할 수 있고[20], 비디오 합성으로도 확장 가능하다[21].

적은 수의 observation 영상을 사용하여 복원된 3D 모델을 기반으로 하여, 참고문헌[14]와 같은 방식으로 새로운 시점의 영상을 합성할 때 폐색(Occlusion) 영역에 구멍(Hole)과 같은 결함(Artifacts)이 나타나게 되는데, 이러한 문제점을 최소한의 3차원 표현에 기반한 DL을 사용해 해결하는 다양한 방식이 개발되고 있다. 미분가능 표면(Differentiable Surface)을 이용한 최소한의 3차원 정보 표현에 기반한 DL 렌더링[22], 3D voxel 및 texture 표현에 기반한 DL 렌더링[23], Point cloud 표현에 기반한 DL 렌더링[24], 복수 plane(or Layer) 표현에 기반한 DL 렌더링[25], 암시된 함수(Implicit Function) 표현 기반 DL 렌더링[26] 기법 등이다.

자유시점 비디오 합성의 경우 기존의 정확한 RGBD 정보에 기반한 실시간 렌더링[23,24]을 하는 기법들을 적용하더라도 최종 합성된 영상 안에 고주파 정보 표현이 누락되는 문제가 발생하는데, DL에 기반하여 다시점 비디오로부터 자동으로 고품질 모델을 구성하고 이를 바탕으로 자유시점의 비디오를 합성할 수 있고, 더불어 새로운 동작 애니메이션 합성도 가능하게 되었다[27].

재조명(Relighting)의 경우 기존의 영상 획득 장비에 기반한 물체의 reflectance 복원 방식[28]은 기본적으로 많은 수의 영상 획득을 필요로 하고, 가능한 설치 공간 한계로 인해 획득하고자 하는 물체의 역학(Dynamics)도 함께 제한되며, 옥외 공간의 물체에 대한 reflectance 복원의 어려움이 발생한다. 이러한 문제점을 해결하기 위해 DL 기법을 적용하여 특수한 영상 획득 장비 없이도 CG 기반하에서 적은 수의 영상 렌더링을 통해서도 고품질의 영상 재조명이 가능하게 되었다[29].

얼굴 재현(Face Reenactment)의 경우 기존의 입력 영상으로부터 복원된 3D 모델 편집을 통한 새로운 합성 방식[30]은 기본적으로 복원된 3D 모델의 정확도와 얼굴 재연을 위한 객체 추적의 정확도에 따라 최종 합성되는 영상의 품질이 결정되는데, 기존의 방식을 통해 합성되는 영상 합성 결과를 조건적인(Conditional) GAN을 통해 정제함으로써 합성 영상의 화질을 향상시킬 수 있게 되었다[31].

3. AI 기반 콘텐츠 제작 기술 동향

앞서 소개한 바와 같이 DL을 적용해 합성 영상의 품질 개선뿐만 아니라 AI는 이미 콘텐츠 제작, 유통, 이용자의 미디어 이용 패턴 분석에 이르기까지 광범위하게 활용되고 있어 간단하게 소개하고자 한다.

AI 기반 콘텐츠 제작 기술은 콘텐츠의 제작과정에서 일부 자동화 혹은 비용 절감을 위해 활용되는 인공지능 기술 혹은 인공지능이 적용된 시스템으로 볼 수 있다. 2016년 IBM Watson은 SF 영화 〈Morgan〉의 예고편 제작에 인공지능을 활용한 바 있으며, 윔블던 경기의 하이라이트 영상 생성에 이용하였다. 영화감독 Oscar Sharp와 인공지능 연구자인 Ross Goodwin은 인공지능 모델 '벤자민(Benjamin)'을 이용하여 만든 대본을 기반으로 제작된 9분 정도 길이의 〈Sunspring〉이라는 영화를 공개하였다. 또한 국내외 여러 기업에서 스포츠 영상의 하이라이트, 영화/드라마의 요약 등에 인공지능이 활용되고 있다.

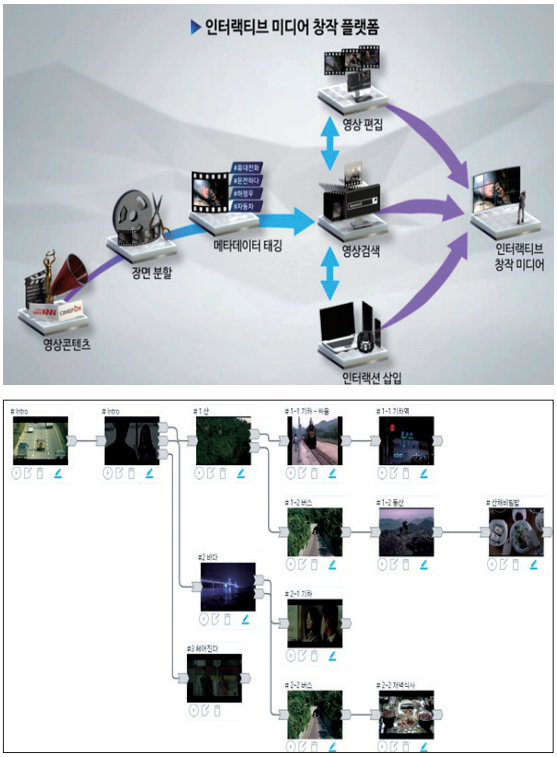

그림 1 ETRI 인터랙티브 미디어창작 플랫폼

Geniussports 연구팀은 CVPR 2020 학술대회의 CVSports 워크숍에서 농구 경기 비디오를 자동 편집하는 기술을 소개하였으며[32], 스탠포드 대학에서는 2020년 많은 수의 테니스 경기 비디오를 분석해 선수, 라켓, 테니스공의 움직임을 학습하고 이를 토대로 실제와 유사한 테니스 경기 비디오를 생성하는 Vid2Player 모델을 발표하였다[33]. 한국전자통신연구원에서는 사용자의 의도에 맞는 새로운 인터랙티브 미디어를 쉽게 창작할 수 있도록 시나리오에 기반한 인터랙티브 미디어 창작 플랫폼을 개발하였다(그림 1)[34].

AI 기반 텍스트-비디오 편집기술은 사용자가 텍스트를 입력하면 이와 상응하는 영상을 수집 및 가공하여 하나의 영상 또는 비디오로 만들어주는 기술이다. 2019년 스탠포드 대학과 막스플랑크 정보과학연구소, 프린스턴 대학, Adobe Research에서 AI 기술을 이용하여 텍스트만 바꿔주면 음성과 영상이 동시에 수정되는 기술을 발표하였으며[35], KAIST 등 국내 대학에서도 TiVGAN과 같은 텍스트에 따라 영상 및 비디오를 생성해 주는 GAN 기반의 네트워크 기술을 발표하였다[36].

Synthesia, Synths Video, Rephrase.ai, Lumen5, Designs.ai, Invideo 등 많은 업체에서 Text-to-Video AI 편집기를 이용하여 텍스트 입력을 통한 비디오 생성 기능을 제공하고 있다. 대부분 편집기의 경우, 주어진 텍스트 또는 키워드를 기반으로 기존의 비디오 라이브러리에서 영상을 추천하고 이를 편집하여 새로운 비디오를 생성하는 기능을 제공하고 있다.

국내외 방송사 및 제작사의 방송프로그램 제작에서도 인공지능이 활용되고 있다. 고인이 된 가수나 배우의 목소리와 영상 생성이 좋은 예이다. 2020년 Mnet, SBS, MBC 등에서는 인공지능 기반 얼굴 및 음성 생성기술을 이용하여 고인이 된 가수의 생전 모습과 생전 목소리를 구현, 이를 통해 새로운 공연을 시연하였다.

한편 인공지능 기술로 제작한 가상의 인물이 광고, 뮤직비디오 등에 속속 등장하고 있다. 2019년 브러드사에서는 '릴미퀼라'라는 가상의 인물을 만들고 패션모델 및 뮤직미디어 등을 통해 2019년 140억 원 규모의 수익을 달성하였고, 국내에서도 '로지'라는 가상의 인물이 TV 광고 모델로 나섰으며, LG전자의 가상 인물 '김래아'는 2021년 국제전자제품박람회(CES)에서 연사로 출연하였다. 또한 2022년 대통령선거에서는 가상으로 제작된 후보가 선거운동에 활용되기도 하였다.

인공지능은 특정 영상 또는 장면을 이용하여 연속된 비디오 장면을 생성하는 데도 이용되고 있다. Michigan University, Adobe Research, POSTECH, Beihang University, 그리고 Google Brain에서는 적

은 수의 프레임을 관측하여 이후 프레임을 생성하는 모델인 MCNet을 발표하였으며, Snap research와 NVIDIA는 CVPR 2018에서 영상의 프레임을 객체와 모션 정보로 분할하여 새로운 영상을 생성하는 MoCoGAN 기술을 발표하였다[37].

4. 장면 구성 메타데이터 표준화 동향

메타데이터는 객체, 장면, 영상, 비디오와 같은 데이터의 활용을 위한 데이터(Data for Data)를 의미한다. 메타데이터는 콘텐츠의 제작, 콘텐츠의 검색 및 브라우징(Search and Browsing), 송수신 및 렌더링 등 콘텐츠 생성 및 소비 전 과정상의 다양한 활용목적에 다양한 규격이 존재한다. 여기서는 공간미디어를 구성하고 렌더링 및 상호작용을 위하여 필요한 장면 기술(Scene Description) 메타데이터 기술규격의 동향에 대하여 MPEG 표준규격을 중심으로 소개한다.

장면 기술 메타데이터는 다양한 2차원 또는 3차원 콘텐츠 객체들로 구성되는 장면의 구성을 기술하기 위한 메타데이터로써, 복수의 영상 및 오디오 객체들로 구성되는 장면을 생성하고, 이를 렌더링하여 표출하기 위하여 사용된다. 장면 기술 메타데이터는 통상적으로 그래프 또는 트리 형식으로 표현되며, 장면을 구성하는 객체들의 기하학적 계층구조를 기술한다. 장면 기술 메타데이터의 일례로 VRML은 XML 신택스 기반으로 웹에서 사용하기 위한 장면구성 메타데이터이다. VRML은 이후 2010년에 JSON 신택스 기반의 X3D 규격으로 정의되었다.

MPEG은 객체 기반 미디어의 송수신 및 부복호화를 위한 MPEG-4 규격을 위하여 객체로 구성되는 장면을 기술하기 위하여 BIFS를 정의하였다. BIFS는 VRML의 구조를 기반으로 MPEG의 객체 기반 미디어 규격을 위한 데이터 스트리밍, 장면의 업데이트, 메타데이터의 압축 등의 규격요소를 추가로 정의하였다[38]. 한편 OpenGL은 게임과 같은 3차원 그래픽 렌더링을 위한 규격으로 정의되었으며, 이후 glTF 규격으로 발전하였다. glTF는 장면 기술 포맷으로 3차원 콘텐츠 관련 서비스를 위한 상호호환성을 보장하기 위한 공통규약으로써의 활용을 목적으로 한다[39].

MPEG은 AR/VR 등 새롭게 대두되는 몰입형 미디어 환경을 위한 표준화 MPEG-I에서 장면 기술 메타데이터 규격을 glTF를 기반으로 하여 정의하였다[40]. MPEG-I 장면 기술 메타데이터는 glTF의 장면 기술 구조 기반으로 확장된 요소를 정의하여 구성되며, 다음과 같이 추가 기능을 지원한다.

- 비디오 및 동적 메쉬 및 포인트 클라우드
- 오디오 노드
- 장면 구성 요소들에 대한 상호작용
- 실시간 미디어 및 AR 지원

현재 MPEG-I의 장면 기술 규격화 작업은 국제표준(ISO/IEC 23090-14 IS) 제정을 목표로 진행 중이며, phase 2 추가 기술요소를 탐색 중이다.

Ⅲ. 결론

지금까지 가상 스튜디오 기술의 부상, DL 적용으로 합성 영상의 고품질화, 콘텐츠 제작에서 소비 분야에 걸쳐 폭넓은 인공지능의 접목, 콘텐츠 검색, 장면 구성 등에 필요한 메타데이터 표준화에 걸쳐 최신 콘텐츠 제작과 활용에 관한 기술 동향을 간단하게 소개하였다. 이러한 콘텐츠 제작 기술 동향에서 파악할 수 있는 것 중 하나는 게임 엔진의 실시간 렌더링 및 LED Wall 투영과 같이 콘텐츠를 이루는 객체 기반 장면 구성의 프로그래밍화가 가속화되고

있다는 점이다. 콘텐츠는 시간 축으로 단순하게 보면 여러 장면으로 구성되는데 하나의 장면은 여러 객체의 관계로 구성된다. 만약 장면의 기본 단위가 되는 중요한 객체를 표준화된 포맷으로 DB화하여 공개할 수 있다면 향후 콘텐츠를 제작할 때 매우 유용할 것이다. 현재 콘텐츠 제작에서 사용하고 있는 하드웨어 기기나 소프트웨어 기술이 극도로 고도화되고, 또한 공통으로 활용할 수 있는 표준화된 객체 DB까지 완전하게 제공된다면 아마도 먼 미래에는 콘텐츠 제작의 작업 흐름이 마치 컴퓨터 프로그래밍을 하는 것처럼 바뀌면서 다양한 콘텐츠를 컴퓨터가 대량 생산하는 시대를 맞이할지도 모르겠다. 물론 이 과정에서 객체 데이터 공개 및 가공에 따른 프라이버시 보호 문제나 저작권 문제도 함께 짚어 나가야 할 것이다.

용어해설

콘텐츠 제작 「콘텐츠산업 진흥법」 제2조에서 "콘텐츠 제작"이란 창작·기획·개발·생산 등을 통하여 콘텐츠를 만드는 것을 말하며, 이를 전자적인 형태로 변환하거나 처리하는 것을 포함하는 것으로 정의됨

방송프로그램 「방송법」 제2조에서 "방송프로그램"을 방송편성의 단위가 되는 방송내용물로 정의. 여기서 "방송편성"은 방송되는 사항의 종류·내용·분량·시각·배열을 정하는 것이고, "방송내용물"은 정지 또는 이동하는 사물의 순간적 영상과 이에 따르는 음성·음향 등으로 이루어짐

장면 동일한 장소, 시간, 스토리를 가지는 연속된 영상 프레임

약어 정리

3D	Three Dimensional
AI	Artificial Intelligence
AR	Augmented Reality
BIFS	Binary Format for Scenes
CG	Computer Graphics
CN	Convolutional Networks
CV	Computer Vision
DB	DataBase
DL	Deep Learning
GAN	Generative Adversarial Networks
glTF	Graphics Library Transmission Format
HDR	High Dynamic Range
IBMR	Image Based Modeling and Rendering
LED	Light Emitting Diode
MPEG	Moving Picture Expert Group
OpenGL	Open Graphics Library
OTT	Over-The-Top
POBM	Programmable Object-Based Media
RGBD	Red, Green, Blue, Depth
SPS	Semantic Photo Synthesis
TF	Task Force
TiVGAN	Text-to-Image-to-Video GAN
VFX	Visual Effects
VR	Virtual Reality
VRML	Virtual Reality Modeling Language
X3D	eXtensible 3D
XML	eXtended Markup Language
XR	eXtended Reality

참고문헌

[1] J. Svensson, "Watch out, wedding videographers, AI is coming for you," IEEE Spectrum, Nov. 2021.

[2] Ofcom, "Object-based media report," Sept. 2021.

[3] 유건식, "OTT와 위드코로나 시대, 방송 제작 현장의 변화," 방송트렌드&인사이트, 제28권 제3호, 2021.

[4] Nevion, "5G VIRTUOSA project introduction," IRT, 2021.

[5] A. Pennington, "Virtual production can be real for everybody-Here's how," TVTECH, June 2020.

[6] R. Krishna et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," arXiv preprint, CoRR, 2016, arXiv: 1602.07332v1.

[7] 유홍준, "명작순례: 유홍준의 미를 보는 눈 2," 눌와, 2013.

[8] 유미, "가상 제작의 개념과 해외 제작 사례 분석," 애니메이션연구, 제17권 제1호, 2020, pp. 98-113.

[9] C.K. Ellie, "Graphic masters: Creating real time VFX with virtual production," Genero, https://genero.com/insights/graphics-masters-creating-real-time-vfx-with-virtual-production

[10] 김미라, "포스트 코로나, 영상 콘텐츠 제작 기술," 영상기술연구, 제1권 제35호, 2021, pp. 27-44.

[11] 이남수, 이한결, "VFX는 달리는 말이다," 키움증권 리서치센터, 2021. 4. 6.

[12] 삼성전자 뉴스룸, "삼성전자, CJ ENM과 버추얼 스튜디오 구축 파트너십 체결," 2021. 7. 26.

[13] 이동호, "버추얼 프로덕션 기술을 활용한 제작 기술 동향 연구," 영상기술연구, 제1권 제31호, 2019, pp. 61-78.

[14] P. Debevec, Y. Yizhou, and G. Borshukov, "Efficient view-dependent image-based rendering with projective texture-mapping," in Eurographics Workshop on Rendering Techniques, Springer, Vienna, Austria, 1998, pp. 105-116.

[15] M. Dou et al., "Fusion4d: Real-time performance capture of challenging scenes." ACM Trans. Graph., vol. 35, no. 4, 2016, pp. 1-13.

[16] K. Guo et al., "The relightables: Volumetric performance capture of humans with realistic relighting," ACM Trans. Graph., vol. 38, no. 6, 2019, pp. 1-19.

[17] M. Levoy and P. Hanrahan, "Light field rendering," in Proc. 23rd Annu. Conf. Comput. Graph. Interact. Techniques (SIGGRAPH), (New Orleans, LA, USA), Aug. 1996.

[18] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), (Boston, MA, USA), 2015, pp. 3431-3440.

[19] L. Karacan et al., "Learning to generate images of outdoor scenes from attributes and semantic layouts," arXiv preprint, CoRR, 2016, arXiv: 1612.00215.

[20] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), (Venice, Italy), Oct. 2017, pp. 1511-1520.

[21] T.C. Wang et al., "Video-to-video synthesis," arXiv preprint, CoRR, 2018, arXiv: 1808.06601.

[22] S.M.A. Eslami et al., "Neural scene representation and rendering," Science, vol. 360, no. 6394, 2018, pp. 1204-1210.

[23] J.Y. Zhu et al., "Visual object networks: Image generation with disentangled 3d representation," arXiv preprint, CoRR, 2018, arXiv: 1812.02725.

[24] M. Meshry et al., "Neural rerendering in the wild," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), (Long Beach, CA, USA), June 2019, pp. 6878-6887.

[25] Z. Xu et al., "Deep view synthesis from sparse photometric images," ACM Trans. Graph., vol. 38, no. 4, 2019, pp. 1-13.

[26] S. Saito et al., "Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), (Seoul, Republic of Korea), Oct. 2019, pp. 2304-2314.

[27] S. Lombardi et al., "Neural volumes: Learning dynamic renderable volumes from images," arXiv preprint, CoRR, 2019, arXiv: 1906.07751.

[28] P. Debevec et al., "Acquiring the reflectance field of a human face," in Proc. 27th Annu. Conf. Comput. Graph. Interact. Techniques (SIGGRAPH), (New Orleans, LA, USA), July 2000, pp. 145-156.

[29] A. Meka et al., "Deep reflectance fields: High-quality facial reflectance field inference from color gradient illumination," ACM Trans. Graph., vol. 38, no. 4, 2019, pp. 1-12.

[30] J. Thies et al., "Face2face: Real-time face capture and reenactment of rgb videos," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), (Las Vegas, NV, USA), June 2016, pp. 2387-2395.

[31] H. Kim et al., "Deep video portraits," ACM Trans. Graph., vol. 37, no. 4, 2018, pp. 1-14.

[32] J. Quiroga et al., "As seen on TV: Automatic basketball video production using Gaussian-based actionness and game states recognition," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), (Seattle, WA, USA), June 2020, pp. 3911-3920.

[33] H. Zhang et al., "Vid2player: Controllable video sprites that behave and appear like professional tennis players," ACM Trans. Graph., vol. 40, no. 3, 2021, pp. 1-16.

[34] 손정우, 한민호, 김선중, "인공지능 기반 영상 콘텐츠 생성 기술 동향," 전자통신동향분석, 제34권 제3호, 2019, pp. 34-42.

[35] O. Fried et al., "Text-based editing of talking-head video," ACM Trans. Graph., vol. 38, no. 4, 2019.

[36] D. Kim, D. Joo, and J. Kim, "TiVGAN: Text to image to video generation with step-by-step evolutionary generator," IEEE Access, vol. 8, 2020, pp. 153113-153122.

[37] S. Tulyakov et al., "MoCoGAN: Decomposing motion and content for video generation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), (Salt Lake City, UT, USA), June 2018, pp. 1526-1535.

[38] ISO/IEC 14496-11, Coding of audio-visual objects, Part 11: Scene description and Application engine(BIFS, XMT, MPEG-J).

[39] GL Transmission Format(glTF) version 2.0, 2017.

[40] Imed Bouazizi, MPEG-I Scene Description Overview, mpeg-sg.org, 2021.

[41] 조용성 외, "미디어와 AI 기술: 미디어 지능화," 전자통신동향분석, 제35권 제5호, 2020, pp. 92-101.

부록

A. 사용 사례

A.1 Compression domain에서의 3D 미디어 검색 및 브라우징

〈내용〉

- 공간미디어 제작자는 원하는 3차원 공간상의 내용객체를 배치하고 편집하는 환경에서 원하는 3차원 내용객체를 리소스 DB에서 검색할 수 있음
- 이때 검색의 속도를 향상하기 위해서는 리소스 DB 내의 압축된 3차원 내용객체에 대하여 별도의 디코딩 과정 없이 검색 및 브라우징이 가능하여야 함

〈기술고려사항〉

- 부호화 또는 압축된 상태의 3차원 내용객체 집합에 대한 검색
- 유사한 내용객체들의 클러스터링을 제공하거나, 내용객체 간의 유사도를 측정하기 위한 수단
- 쿼리 인터페이스 형태

A.2 2D Cartoon 기반 3D 입체 콘텐츠 제작

〈내용〉

- 2D cartoon을 입력받아 실사 수준의 3D 입체 콘텐츠를 자동으로 생성하여 소비할 수 있도록 함

〈기술고려사항〉

- 2D cartoon을 실사 수준의 3D 입체 영상으로 변환하는 기술
- 변환된 입체영상의 이웃장면 자동 생성 기술
- 장면에 적합한 배경 음악 자동 생성 기술

A.3 Description 기반 3D 입체 콘텐츠 제작

〈내용〉

- 텍스트, 2D 이미지/동영상, 혹은 객체와 배경을 포함한 장면 서술자에 기반한 시나리오를 입력받아 3D 입체 콘텐츠를 자동으로 생성할 수 있도록 함

〈기술고려사항〉

- 서술자 혹은 2D 영상 기반 3D 입체 콘텐츠 생성 및 렌더링 기술
- 서술자 기반 장면, 시나리오 표현 기술

A.4 소비자 중심의 미디어 소비

〈내용〉

- 단말 환경에 따른 최적의 콘텐츠 소비: 소비자가 사용하는 장치의 성능(2D, 3D, HMD)에 적합한 콘텐츠 소비를 할 수 있도록 함
- 소비자가 원하는 시점에서의 콘텐츠 소비: 관찰자 또는 1인칭 주인공 등 소비자가 원하는 시점에서의 콘텐츠 소비를 할 수 있도록 함
- 소비자와의 상호작용에 따른 콘텐츠 소비: 드라마의 배우를 변경/선택하거나 줄거리를 선택하여 소비할 수 있도록 함

〈기술고려사항〉

- 콘텐츠 변환 기술(3D to 2D, 2D to 3D 등)
- 특정 시점 기반 렌더링 기술
- 실사 수준의 object 기반 콘텐츠 생성 기술

A.5 기 제작된 콘텐츠 재구성

〈내용〉

- 사전 제작된 영상 콘텐츠(영화, 드라마 등)를 서비

스하는 과정에서 외부 환경변화(배우 스캔들)로 인한 피해사례 방지
- 학교폭력 등 범죄나 사회적 물의를 일으킨 배우만 따로 떼어내고 다른 배우가 새롭게 연기한 부분만 덮어쓰는 형태로 조작 가능한 미디어 기능 제공
- 제작비가 모두 매몰비용(이미 지출해서 회수할 수 없는 비용)이 되는 것을 방지 가능

〈기술고려사항〉
- 장면, 객체 식별 및 대체 기술

B. 분야별 AI 활용 사례[41]

〈부호화〉
- MPEG에서는 AV 부호화를 위한 NN 기술에 대한 Ad-hoc그룹 생성 및 표준화에 대한 논의 중
- JPEG에서는 인공지능 기반 이미지 부호화를 위해 JPEG-AI에 대한 표준화를 진행 중
- CVPR 2020 등 주요학술대회에서 기존 부호화 코딩 대비 10% 이상 개선된 이미지 압축 기술을 발표

〈메타데이터〉
- AI 비전 및 음성인식 기술을 통해 영상 내 일부 객체 및 음악 등을 자동 검출(SKT, 보이저X 등)
- AI 기술을 활용하여 동영상의 메타데이터를 자동 추출하는 서비스를 제공(IBM, Google, Leankr 등)

〈화질개선〉
- 캐논메디컬은 초고해상도 및 선명한 화질의 진단 영상을 제공하는 AiCE(Advanced Intelligent Clear-IQ Engine) 엔진 개발(2019. 7.)
- 삼성전자는 입력 화질과 관계없이 8K 수준의 화질로 업스케일링해 주는 AI 퀀텀프로세서가 탑재된 TV 출시(2020. 1.)
- LG전자는 원본 영상의 화질을 분석해 품질을 향상시키는 인공지능 프로세서 알파9 3세대 칩이 내장된 OLED 8K TV 출시(2020. 1.)

〈콘텐츠 크리에이터〉
- 영국 BBC는 인공지능 시스템인 'Ed'를 이용하여 다수의 고정된 카메라를 설치하고 샷 추출과 컷 편집 등 분절된 가상의 샷들을 상황에 맞는 자동 편집 기술 개발
- Synthesia, Synths Video, Rephrase.ai, Lumen5, Designs.ai 등에서는 인공지능기반 Text-AI 편집기를 이용하여 텍스트 입력을 통한 비디오 생성 기능 제공

〈콘텐츠 전송〉
- KAIST 등 학계를 중심으로 기계 학습을 오류정정, MIMO 채널 모델링, 저전력 통신 등 다양한 통신 분야로 적용 시도
- 버지니아 공대, AT&T에서는 2015년부터 딥러닝 기반의 통신시스템 설계 관련 연구 수행
- ITU-T SG13 FG-ML5G에서 기계학습 기반으로 5G를 포함한 차세대 네트워크에 관한 표준화 논의 중