인공지능 기반 영상 콘텐츠 생성 기술 동향

Artificial Intelligence-Based Video Content Generation

손정우 (J.-W. Son, jwson@etri.re.kr) 스마트미디어연구그룹 선임연구원
한민호 (M.-H. Han, mhhan@etri.re.kr) 스마트미디어연구그룹 책임연구원
김선중 (S.-J. Kim, kimsj@etri.re.kr) 스마트미디어연구그룹 책임연구원/PL

ABSTRACT

This study introduces artificial intelligence (AI) techniques for video generation. For an effective illustration, techniques for video generation are classified as either semi-automatic or automatic. First, we discuss some recent achievements in semi-automatic video generation, and explain which types of AI techniques can be applied to produce films and improve film quality. Additionally, we provide an example of video content that has been generated by using AI techniques. Then, two automatic video-generation techniques are introduced with technical details. As there is currently no feasible automatic video-generation technique that can generate commercial videos, in this study, we explain their technical details, and suggest the future direction for researchers. Finally, we discuss several considerations for more practical automatic video-generation techniques.

KEYWORDS 영상 생성, 기계학습, 심화 학습, GANs, AI, Video Contents Generation, Video Prediction, Generative Adversarial Networks

Ⅰ. 서론

영화, 방송 드라마, 다큐멘터리 등 영상 콘텐츠는 다양한 예술 장르가 복합적으로 적용되는 종합 예술 장르일 뿐 아니라 거대 엔터테인먼트 산업을 형성하는 핵심 매체이다. 인공지능 기술 관점에서 영상 콘텐츠는 시각과 청각 정보를 담은 순차 데이터로 다룰 수 있다. 즉 이미지, 음성, 텍스트 등 여타의 단일 모달리티(Modality) 데이터와 달리 복합 모달리티를 가지는 데이터이며, 이들 복합 모달리티가 프레임을 통해서 순차적으로 표출된다. 영상 콘텐츠가 담고 있는 시대상, 역사적 의미 등 시각적으로 관측하기 어려운 정보까지 고려하면 일반적인 데이터와는 극명하게 구분된다. 본 논문에서

* DOI: 10.22648/ETRI.2019.J.340304
* 본 연구는 한국전자통신연구원 연구운영비지원사업의 일환으로 수행되었음[19ZH1300, 오픈 시나리오 기반 프로그래머블 인터랙티브 미디어 창작 서비스 플랫폼 개발].

는 이와 같은 영상 콘텐츠의 자동 생성을 위한 인공지능 기술의 동향을 소개하고자 한다.

딥러닝(Deep learning)의 제안 이후, 인공지능 기술 기반의 콘텐츠 생성 기술은 크게 발전하고 있다. 특히, Generative Adversarial Networks(GANs)[1]의 제안은 데이터의 일반적인 구조적 특성뿐만 아니라 세부적인 정보를 재현할 수 있는 학습 방법을 제공함으로써 흡사 인간에 의해 창작된 듯한 데이터의 생성을 가능하게 하였다. 임의의 사진을 생성하거나[2], 주어진 사진을 바탕으로 그림을 생성하는 등[3]의 예는 이제 쉽게 찾아볼 수 있다. 뿐만 아니라 인간의 영역으로 간주하였던 문학에서도 에세이[4], 시[5] 등 인공지능 기술이 적용되어 시장에 유통되는 작품까지 접할 수 있게 되었다.

콘텐츠 창작을 위한 인공지능 기술은 일견 산업적 가치가 크지 않아 보인다. 예컨대, 인공지능에 의해 창작된 시나 에세이 혹은 그림 등은 예술적으로 가치를 측정하기 어려우며 자연스러운 결과물을 얻기까지 많은 기술적 장애를 넘어야 한다. 즉, 기술 개발에 투입되는 비용 대비 그 결과물의 직접적 활용은 제한된다. 그렇다면 왜 많은 연구자 혹은 기업이 콘텐츠 생성을 위한 인공지능 기술을 개발하고 있는가를 따져봐야 한다. 콘텐츠 생성 기술의 목표는 인간과 동일한 창작 능력의 획득에 있다. 이를 위해서 기계는 기존 콘텐츠를 분석하여 표현하고, 표현된 형상을 기반으로 새로운 데이터를 생성할 수 있어야 한다. 인간 또한 창작을 위해 기존 분야의 지식을 습득하고 이해한 지식을 바탕으로 자신만의 창작물을 생성한다. 즉 콘텐츠 창작은 데이터의 표현, 이해, 생성의 전 과정을 인공지능 기술이 인간과 유사하게 수행하는 데 그 목적이 있다. 이를 통해 최종 산출물뿐만 아니라 다양한 분야에서 활용 가능한 중간 산출 기술들을 얻을 수 있다.

사진을 기반으로 그림을 생성하는 Google의 Deep dream[6]을 생각해 보면 기계는 기존의 작품을 바탕으로 데이터의 형상을 분석하고, 특징을 추출한다. 이를 토대로 사진의 정보를 유지하면서 특정 화풍을 삽입한 새로운 그림을 그릴 수 있다. 해당 기술의 개발을 통해 얻어지는 이점이 화풍을 따라 하는 이미지 생성에 그친다면 그 기술적 가치는 낮아진다. 하지만 이미지 생성을 위해 이미지 데이터를 분석하고, 잠재된 요소들을 대체하는 기술은 다양한 분야에 적용 가능하기에 다양한 생성 기술이 연구되고 있다.

영상 콘텐츠 생성을 위한 인공지능 기술 또한 기존 콘텐츠 생성 기술과 동일한 과정을 거쳐야 한다. 다만, 현재까지 널리 알려진 기술들이 단일 모달리티를 고려하는 데 비해 복수의 모달리티를 동시에 생성해야 한다는 차이점이 있다. 시각, 청각 등 각각의 모달리티는 서로 간의 연관성을 가지고 생성되어야 하기 때문에 단일 모달리티에 비해 기술 개발의 어려움이 크다. 특히, 이들 모달리티의 관계가 의미에 따라 변화하거나 유지되어야 하며, 이때의 의미가 데이터에 직접적으로 나타나지 않는 경우까지 고려하면 기술 개발의 난이도는 더 높아진다. 이로 인해 사진, 시, 소설 등의 콘텐츠 생성 기술과 비교할 때 영상 콘텐츠 생성 기술은 학계를 중심으로 연구가 이제 시작되는 시점이라 볼 수 있다.

본 논문에서는 현재까지 발표된 영상 콘텐츠의 자동 생성 기술에 대해 논하고자 한다. 영상 콘텐츠 자동 생성 기술은 기저에 객체 및 행위 인식, 이미지 생성, 음성 및 음악 생성, 컨텍스트 분석 등 다양한 기술들을 내포하고 있다. 따라서 이들 기술까지 논할 경우 영상 및 음성 처리, 자연어 처리 등 다양한 기반 기술을 함께 다루어야 한다. 본 동향에서는 공간의 제약을 감안하여 영상 콘텐츠를 산출물로 생성하는 기술에 대해 한정하여 소개

하고자 한다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 영상 콘텐츠 생성 기술을 크게 반자동 및 자동 기술로 구분하여 정의한다. Ⅲ장과 Ⅳ장에서는 각 분류에 해당하는 기술을 소개한다. Ⅴ장에서는 각 기술의 장단점을 분석하고, 이후 기술 개발 방향을 제시하며 결론을 맺는다.

Ⅱ. 영상 콘텐츠 생성 기술 분류

본 논문에서는 영상 콘텐츠 생성 기술을 크게 반자동기술과 자동 기술로 나누어 설명한다. 반자동 기술은 영상 콘텐츠의 생성 과정에서 결과물의 질적 향상을 위해 수동 정보 삽입 혹은 수정이 요구되는 기술을 의미하며, 자동 기술의 경우 초기 입력을 기준으로 최종 결과물인 영상 콘텐츠 생성까지 인공지능 모델에 의해 전 과정이 수행되는 기술을 의미한다.

인공지능 기술을 기반으로 생성된 결과물로 대내외에 알려지는 대부분의 영상 콘텐츠는 현재까지 반자동 기술을 기반으로 구축되었다. 이는 초기 개발 단계에 있는 영상 콘텐츠 생성 기술의 특성상, 산출물의 질적 수준을 담보할 수 없기 때문이다. 따라서 현재까지 제안된 기술은 자동 영상 콘텐츠 생성 기술의 성능이 크게 떨어지는 상황이나 미래에 충분한 기술 개발이 진행된다면 그 활용성은 자동 영상 콘텐츠 생성 기술이 더 크다고 볼 수 있다. 표 1은 현재까지의 반자동/자동 영상 콘텐츠 생성 기술들이 가지는 상대적인 장단점을 보여주고 있다. 표에서와 같이 현재 제안된 자동 영상 콘텐츠 생성 기술은 영상 콘텐츠의 생성 과정에서의 비용 절감 효과를 제외하고는 반자동 기술을 넘어서기는 어렵다. 하지만 반자동 기술의 적용을 통해 파악되는 영상 콘텐츠 생성의 세부 과정 및 요구사항들이 자동 영상 콘텐츠 생성 기술에 적용되는 시점이 멀지 않았다고 판단된다.

표 1 영상 콘텐츠 생성 기술 분류별 장단점

기준	반자동 영상	자동 영상 생성
생성 비용	▲	○
생성 시간	×	○
의도 반영	▲	×
질적 우수성	○	×
후보정 용이성	○	×

○: 우수, ▲: 보통, ×: 나쁨

따라서 본 논문에서는 반자동 영상 콘텐츠 생성 기술을 먼저 소개함으로써 인공지능에 의한 콘텐츠 생성의 결과물을 공유하고, 현재까지의 자동 영상 콘텐츠 생성 기술을 살펴봄으로써 향후 개발 방향을 논하고자 한다. 반자동 기술에서는 영화 생성을 위한 '벤자민' 모델[7]과 한국전자통신연구원의 인터랙티브 미디어 창작 플랫폼[8,9]을 소개한다. 이후, 자동 영상 콘텐츠 생성 기술의 예로 video prediction 기술은 MCNet[10]과 MoCoGan[11]을 설명하고, 이들 기술의 결과물이 영상 콘텐츠로써 가치를 가지기 위해 요구되는 기술들을 나열하고자 한다.

Ⅲ. 반자동 영상 콘텐츠 생성 기술

반자동 영상 콘텐츠 생성 기술은 영상 콘텐츠의 제작과정에서 일부 자동화 혹은 비용 절감을 위해 활용되는 인공지능 기술 혹은 시스템이다. 2016년 IBM Watson은 SF 영화 〈Morgan〉의 예고편을 인공지능을 활용하여 제작한 바 있으며, 윔블던 경기의 하이라이트 영상을 생성한 바 있다[12]. 반자동 영상 콘텐츠 생성 기술은 영상 콘텐츠의 질적 수준은 떨어뜨리지 않으면서 제작 시 영상 표현, 장소 및

시간의 제약 등을 해소하는 데 그 목적이 있다.

1. 벤자민(Benjamin)

최근 영화감독 오스카 샤프(Oscar Sharp)와 인공지능 연구자인 로스 굿윈(Ross Goodwin)은 영화 시나리오 창작을 위한 인공지능 모델 '벤자민(Benjamin)'[7]을 기반으로 〈Sunspring〉이라는 영화를 일반에 공개한 바 있다. 공개된 영화는 9분짜리의 SF 단편 영화로, 3명의 배우가 출현하며, 이야기의 전체 구성 및 제목까지 벤자민에 의해 작성되었다.

좀 더 상세히 설명하면, 벤자민은 Long and Short Term Memory(LSTM)를 기반으로 설계된 Recurrent Neural Networks(RNN) 모델로써 문장을 생성하도록 학습되었다. 학습 데이터는 스탠리 큐브릭(Stanley Kubrick)의 〈Space Odyssey〉를 비롯하여 〈Brazil〉, 〈Mad Max〉, 〈The Matrix〉, 〈Star Wars〉 등의 영화 스크립트와 30,000곡 이상의 팝송 가사를 이용하여 구축하였다. 스토리 시퀀스의 일관성을 LSTM-RNN 기반의 모델을 통해 유지함으로써 벤자민은 연속성을 가지는 문장들을 생성할 수 있었고, 이를 토대로 영화를 제작할 수 있었다.

비록 벤자민은 스크립트 작성에 한정하여 인공지능을 활용하고 있으나, 영화 〈Sunspring〉 제작에서는 Deepfake[13]와 같은 얼굴 교체 기술이 활용되었다. 즉, 영화의 시나리오 창작부터 촬영 및 후편집 과정의 요소에서 인공지능 기술이 활용된 것을 알 수 있다. 이를 근거로 〈Sunspring〉을 인공지능이 만든 최초의 상업 영화라 평한다.

벤자민은 여전히 자동으로 스크립트를 작성하고 있다. 2019년 3월에도 매일 한 편씩 생성된 스크립트를 일반에 공개하고 있다. 이와 같은 성공에도 불구하고 최초의 인공지능 기반 영화 제작은 그 과정에서 다양한 문제점을 보여주었으며, 이는 향후 영상 콘텐츠 생성을 위한 인공지능 모델 개발 시 고려되어야 할 사항으로 정리되었다.

벤자민은 일반적인 문장 생성 모델과 달리 영화의 스크립트를 생성하도록 설계되었다. 그럼에도 불구하고 생성된 문장은 너무 일반적인 문장이며, 생성한 장면은 특색이 없는 경우가 많았다. 〈Sunspring〉 제작에서는 이들 내용을 각본가가 손을 봐서 활용하였다. 뿐만 아니라 영화 스크립트는 현재 화면에 대한 구성, 분위기, 인물의 감정 등 상세한 설명이 요구되나 벤자민은 이를 놓치는 경우가 많았다. 기술적으로 LSTM-RNN 기반의 모델은 컨텍스트의 연속성을 고려하여 전체 문장의 구조적 특성을 학습할 뿐 문장의 세세한 특징을 학습하여 인간과 유사한 형태를 생성하기는 어렵다. 이러한 모델의 특성이 결과로 나타난 것으로 파악된다.

다행스럽게도 문장 생성에서의 문제점은 최근 두각을 드러내는 GANs 기반의 모델[1]을 통해 해소되고 있는 상황이다. GANs는 adversarial loss의 최소화를 통해 전체 데이터의 구조적 특성뿐만 아니라 각 인스턴스의 세밀한 정보를 재현하도록 학습된다. 뉴스, 소설, 시 등 다양한 장르에 활용되고 있으며, 이러한 모델들이 영화의 스크립트 생성에 충분히 활용될 것으로 기대된다.

이후 로스 굿윈과 오스카 샤프 감독은 벤자민의 스크립트를 얼굴 교체, 음성 합성 등의 기술을 적용하여 영화 〈Zone Out〉을 제작한 바 있다[14]. 이는 〈Sunspring〉의 제작 과정에서 제작진이 역설한 음성의 생성, 얼굴 생성 등 더 다양한 인공지능 모델의 적용을 실험해 본 작품이라 볼 수 있다. 이들 기술을 통해 제작 비용 및 기간을 크게 단축할 수 있을 것으로 기대된다. 다만, 영상 콘텐츠의 모든 요소를 일일이 생성하는 것이 현재 기술을 활용할 경우 높은 모델의 복잡도와 연산 시간으로 인해 비용 측면에서의 이점이 크다고 보긴 힘들다. 이는

차후 기술 개발을 통해 실현될 것으로 판단된다. 이에 대한 해법 중 하나로 한국전자통신연구원은 인터랙티브 미디어 창작 플랫폼을 제안한 바 있다.

2. 인터랙티브 미디어 창작 플랫폼

한국전자통신연구원의 인터랙티브 미디어 창작 플랫폼(이하 ETRI 창작 플랫폼)은 사용자의 의도에 맞는 새로운 인터랙티브 미디어를 쉽게 창작할 수 있도록 기능을 제공한다. ETRI 창작 플랫폼의 사용 대상은 일반인을 가정하고 있다. 일반 사용자의 경우, 영상의 스토리가 있더라도 스토리에 맞는 고품질의 영상을 촬영 및 편집하기는 어렵다. ETRI 창작 플랫폼에서는 기존 영화 및 방송 등 영상을 분석하여 의미 단위로 분할하고, 분할된 영상에 대한 인공지능 기반의 태깅 기능을 제공한다. 사용자는 자신의 스토리를 질의로 하여 분할된 영상을 검색 및 선택함으로써 기존에 촬영된 고품질의 영상을 기반으로 새로운 스토리를 구성할 수 있다.

스토리의 구성은 다수의 장면을 조합하여 하나의 영상 콘텐츠를 생성하도록 하고 있다. 이때, 추가로 스토리의 분기에 따른 사용자 인터랙션을 삽입할 수 있는 기능을 제공함으로써 최근 이슈가 되고 있는 Netflix의 벤더 스내치와 같은 인터랙티브 영화를 쉽게 제작할 수 있다.

각 기능에 대해 상세히 살펴보면, 먼저 Content Provider(CP)로부터 얻어진 영화, 방송 드라마 등의 영상은 영상의 최소 단위인 샷(shot)으로 분할된다. 분할된 샷의 영상/음성/텍스트 정보를 기반으로 입력 영상을 다수의 장면으로 분할한다. 각 장면이 가지는 자막/각본 등의 정보를 기반으로 장면의 내용을 설명할 수 있는 벡터를 생성하여 각 장면에 태깅한다. 이를 위해 LSTM 기반의 RNN 모델을 활용하고 있다. 뿐만 아니라 누가, 언제, 어디서, 무엇을, 어떻게에 해당하는 온톨로지를 사전에 구축하여 시스템 관리자에 의한 수동 태깅을 함께 지원하고 있다.

분할 및 태깅된 장면은 검색의 대상으로 활용된다. 그림 1은 창작 플랫폼의 검색 화면을 보여준다. 사용자는 자신의 스토리 중 하나의 장면을 구성할 것으로 판단되는 일부를 질의(query)로 입력하고, 이에 매칭되는 장면의 리스트를 창작 플랫폼이 제공한다. 적합한 장면을 선택하면, 각 장면에 대한 편집 및 인터랙션 삽입을 위한 간단한 도구들이 제공된다.

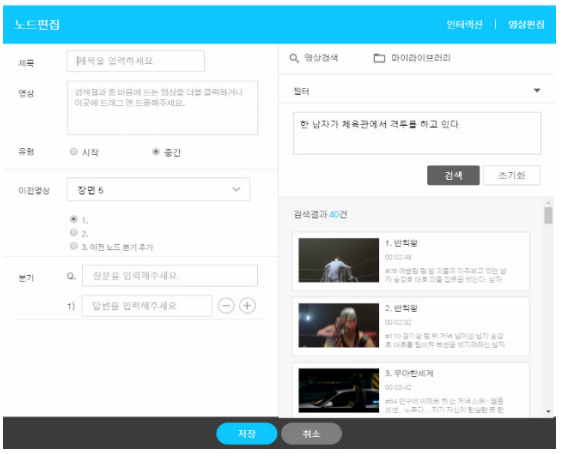

그림 1 창작 플랫폼의 장면 검색 화면

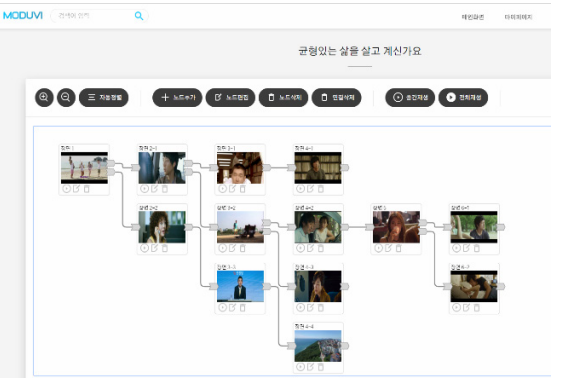

그림 2 창작 플랫폼의 스토리 구성 화면

하나의 영상 콘텐츠는 다수의 장면으로 구성된다. ETRI 창작 플랫폼에서는 각 장면과 장면들이 선형적으로 구성되는 기존의 영상 콘텐츠에 더해 다양한 분기가 가능한 인터랙티브 콘텐츠를 쉽게 생성할 수 있는 도구를 제공하고 있다. 그림 2는 ETRI 창작 플랫폼의 스토리 구성 화면을 보여준다. 그림에서 알 수 있듯이 스토리 검색을 통해 선택된 장면들은 이전, 이후 장면들과 연결을 맺을 수 있으며, 이때 1:1 관계뿐만 아니라 1:*N* 관계를 지원함으로써 스토리의 분기를 사용자가 선택할 수 있다.

지난 2018년 ETRI는 기술의 활용성 검증을 위해 100여 편의 상업 영화와 부산을 주제로 촬영된 영상을 기반으로 데이터베이스를 구축하여, 이를 이용한 인터랙티브 미디어 창작 공모전을 개최하였다[15]. 10여 개의 대학생 팀이 참가하였으며, 다양한 아이디어를 가미한 영상을 창작함으로써 ETRI 창작 플랫폼이 일반인의 영상 창작에 도움이 될 수 있음을 보였다.

벤자민과 ETRI 창작 플랫폼은 영상 콘텐츠의 창작 과정에서 인공지능 기술들이 효과적으로 활용될 수 있음을 잘 보여주었다. 하지만 여전히 사람에 의한 촬영 및 편집 작업이 요구되고 있으며, 인공지능에 의한 영상 창작과는 거리가 있다. 예컨대, 벤자민의 경우 생성 결과물에 대한 후보정 및 편집, 각 장면의 촬영에는 인공지능 기술이 적용하지 못했으며, ETRI 창작 플랫폼의 경우에도 기존 영상을 활용할 뿐 영상 자체에 대한 생성은 지원하지 못하고 있다. 따라서 창작자의 의도를 반영하는 새로운 영상 콘텐츠를 생성한다고 보기는 힘들다.

Ⅳ. 자동 영상 콘텐츠 생성 기술

반자동 영상 콘텐츠 생성 기술의 한계점은 자동 영상 콘텐츠 생성 기술을 통해 해소될 수 있다. 자동 영상 콘텐츠 생성 기술은 사용자의 의도에 부합하는 영상을 '생성'하는 데 목표를 두고 있다. 즉 벤자민이 문장을 생성하여 각본을 작성하듯이, 임의의 사용자 의도에 대해 적합한 프레임의 집합을 자동 생성한다. 자동 영상 콘텐츠 생성 기술의 개발인 이제 시작하는 단계로 고품질의 영상 콘텐츠를 생성하는 데는 한계가 있다. 하지만, 기술의 개발을 통해 얻어지는 제작 환경의 변화, 시장의 확장성 등을 고려하면 그 가치가 크다.

본 논문에서는 MIT CSAIL 연구팀이 발표한 이미지 기반의 영상 생성 기술[16]이나 PredNet[17] 등 현재 발표된 주요 자동 영상 콘텐츠 생성 기술 중 MCNet과 MoCoGan을 소개하고, 영상 콘텐츠 생성의 측면에서 연구 방향성을 논하고자 한다.

1. MCNet

Villegas 등 Michigan University, Adobe Research, POSTECH, Beihang University, 그리고 Google brain 팀의 연구자들이 제안한 MCNet[10]은 임의 수의 프레임을 관측한 값을 토대로 이후 프레임을 생성하는 모델이다. 즉 영상의 0.5초를 관측하고, 이후 0.5초의 가상 영상을 생성한다. 해당 문제는 video prediction이라 명명되어 컴퓨터 비젼 분야에서 연구되기 시작했다.

MCNet에서는 영상의 프레임을 콘텐츠와 모션으로 분리하여 처리한다. 임의 시점의 프레임을 Convolution Neural Network(CNN)를 이용하여 벡터화한다. 생성된 벡터는 영상 프레임의 형상을 담게 된다. 모션의 경우, 현재 프레임과 이전 프레임의 차를 표현하는 residual 프레임을 CNN과 LSTM 기반의 RNN을 이용하여 벡터화하여 얻는다. 얻어진 콘텐츠와 모션 벡터는 CNN 기반의 decoder를 통해

프레임으로 변환된다.

모델의 학습은 생성된 프레임과 실제 프레임 간의 차이와 adversarial loss를 최소화하도록 이루어진다. MCNet은 두 가지 구조로 제안되었다. 설명한 MCNet의 특징은 기본 모델을 통해 구현되며, 생성된 프레임의 질적 향상을 위해서 skip connection을 활용하여 확장된 모델 또한 함께 제안되었다. MCNet은 UCF101 데이터와 KTH 데이터를 활용하여 성능 검증을 하였으며, Structural Similarity Index Measure(SSIM) 기준 10프레임 후의 생성 결과물이 0.7~0.8 사이를 보였다. 흥미로운 점은 생성 프레임의 일부 구간에서 기존 프레임을 변형하지 않는 경우 SSIM이 모델을 통해 생성한 경우에 비해 높다는 것이다. 이는 매우 짧은 시간을 표현하는 프레임의 특성상 매 프레임 모션과 콘텐츠의 변화를 반영하는 것이 효과적이지 못함을 의미한다. 뿐만 아니라 잘못된 영상 생성 모델의 설계는 변화하지 않는 프레임 생성을 통한 성능 확보를 학습 과정에서 시도할 가능성 있음을 의미한다. 따라서 영상 생성 기술 개발에서 반드시 염두에 두어야 할 부분이다.

모델에서는 영상을 구성하는 두 가지 정보로 콘텐츠와 모션을 제안하고 있다. 각각의 프레임에 대한 형상 정보와 프레임의 변화를 의미하는 모션 정보를 효과적으로 추출하고, 이를 활용하여 이후 프레임을 예측함으로써 영상을 생성하고 있다. 이러한 접근은 MoCoGan에서도 찾아볼 수 있다.

2. MoCoGan

Snap research의 Tulyakov가 NVIDIA 연구팀과 함께 제안한 MoCoGAN[11]은 MCNet과 동일하게 영상의 프레임을 콘텐츠와 모션 정보로 분할한다. 이때 크게 달라지는 부분이 영상의 콘텐츠는 변화하지 않음을 가정한다. 즉, 영상을 구성하는 프레임들은 동일한 콘텐츠에 대해 모션을 기반으로 변형한 결과물임을 가정하는 것이다. 얼굴의 표정 변화를 촬영한 영상을 상상하면 쉽게 이해된다. 영상의 콘텐츠는 얼굴의 형상으로 정의할 수 있으며, 특정 표정을 담는 것은 모션 정보를 기반으로 생성된 결과물이다.

MoCoGan은 먼저 CNN을 통해 공통된 콘텐츠 벡터를 추출한다. 임의의 노이즈로부터 생성된 모션 벡터는 콘텐츠 벡터와 결합되어 프레임을 생성하게 된다. 각각의 프레임들을 순차적으로 취합하여 영상을 생성한다. 모델은 크게 두 가지 측면에서 학습된다. 먼저 각 생성된 프레임에 대한 개별 adversarial loss를 최소화하도록 하며, 동시에 프레임 집합인 생성된 영상에 대한 adversarial loss를 최소화하도록 모델을 학습한다.

MoCoGan의 구조를 살펴보면, 생성기 GI의 입력으로 콘텐츠 ZC와 모션 ZM이 입력되며, 이때 ZC는 모든 프레임에 대해 고정된다. 이와 같은 구조는 MCNet과 차별되는 결과물 생성이 가능하다. 예컨대 모션 생성에 클래스 정보를 삽입하여 조건부 생성이 가능하다. 논문에서는 콘텐츠를 고정하고 모션 생성 시 감정, 특정 동작을 삽입한 생성 결과물을 제시한 바 있다.

흔히, 영상 콘텐츠라 함은 다양한 요소를 내재하고 있다. 이는 제작 과정에서도 쉽게 드러난다. 영화를 구성하는 하나의 장면을 생각해 보자. 해당 장면의 배경과 구성에 필요한 소품들이 준비되고, 배우가 연기를 한다. 장면에는 분위기, 감정에 적합한 조명이 있으며, 이를 잘 표현하도록 카메라 구도를 설정하여 촬영한다. 이들 요소는 장면이 표현하는 스토리를 중심으로 유기적으로 결합되어 있다. 안타깝게도 본 논문에서 소개되었던 것과 같이 현재까지의 영상 콘텐츠 자동 생성 기술은 이러

한 복합적인 요소를 고려하지 못하고 있다. 데이터로서의 영상은 각 픽셀 값들의 집합인 프레임이 순차적으로 묶여 있을 뿐이다. 물리적인 형상과 움직임은 분석하여 추출되고 있으나, 그 이면의 내용을 다루지는 못하고 있다. 이로 인해 생성되는 결과물 또한 간단한 동작을 담고 있는 영상을 넘기 힘들다. 따라서 현재까지의 영상 자동 생성 기술이 영상 콘텐츠 자동 생성 기술이라 칭하기는 무리가 있다. 다만, 영상의 측면에서 기계에 의한 프레임 생성이 가능함으로 보이는 수준에 있으며, 이를 통해 영상 콘텐츠 자동 생성의 단초를 제공하고 있음을 알 수 있다.

V. 결론

본 논문에서는 영상 콘텐츠 자동 생성을 위한 인공지능 기술들을 소개하였다. 먼저 발표된 반자동 영상 콘텐츠 자동 생성 기술을 소개함으로써 영상 콘텐츠 제작 환경에서의 인공지능이 대체 가능한 역할들과 그 형상을 설명했다. 이후, 자동 영상 콘텐츠 생성 기술을 통해 제작 시 요구되는 인간의 노력을 줄이기 위한 방향을 제시하였다. 본 논문에서 소개한 자동 영상 콘텐츠 생성 기술은 video prediction 문제를 풀기 위한 방법들로 큰 범주에서 영상 콘텐츠 자동 생성을 video prediction으로 보고 설명하였다.

세부적으로 영상 콘텐츠 자동 생성과 video prediction의 차이를 보면, 영상 콘텐츠는 제작자의 의도, 생각이 반영되어야 한다. 임의의 영상 생성은 영상 콘텐츠 측면에서 큰 가치를 가지지 않기 때문이다. 다음으로 영상에 직접적으로 표출되지 않고 잠재된 정보들이 반영되어야 한다. Video prediction에서 다뤄지는 영상 데이터와 달리 영상 콘텐츠는 복합적인 사고의 산물이다. 영상 제작 시 고려되는 다양한 요소를 추출하고 적용하는 것은 생성된 영상이 콘텐츠로서 가치를 가질 수 있는 최소한의 조건이다. 부차적으로 기존 제작 환경에서 적용되는 전통적인 촬영, 미술, 조명 기법들에 대한 기술적 해석은 인공지능에 의한 창작 결과물이 좀 더 인간과 같은 형상을 가지도록 해줄 것으로 생각된다.

이러한 차이를 고려하면 인공지능 기반의 자동 영상 창작 기술을 위해서는 다양한 형상을 가지는 창작자의 의도 및 생각을 해석할 수 있는 기술이 추가적으로 요구된다. 시놉시스 형태의 텍스트, 사진 등을 토대로 스토리를 생성하는 벤자민이나 사용자 시나리오 질의와 영상의 매칭은 이러한 기술의 시작점에 있다고 볼 수 있다. 이에 대한 고도화 및 추가적인 기술 개발이 요구된다. 생성된 영상의 콘텐츠로서의 질적 수준을 보장하기 위해 인공지능 모델은 영상의 제작 방법, 과정과 다양한 전통적 기법들을 반영할 수 있어야 한다. 이를 위해서는 기술적으로 다양한 정보들이 반영된 영상 생성 모델이 개발되어야 하며, 개발 과정에서 영상 제작자, 작가, 감독 등 제작에 참여하는 다양한 분야의 인력이 참여하여야 할 것이다.

이외에도 영상의 품질 향상을 위한 video restoration 및 편집 도구 등 부수적으로 다양한 기술들이 요구된다. 따라서 영상 콘텐츠 자동 생성 기술의 완성은 매우 먼 미래의 일일 것이다. 현재 미디어 시장에서는 Netflix, Amazon 등 기술적 우위를 바탕으로 시장을 선점하는 거대 글로벌 기업들을 손 쉽게 찾아 볼 수 있다. 이는 현재의 낮은 기술적 수준과 기술 완성까지의 긴 시간이 영상 콘텐츠 자동 생성 기술의 개발을 미루는 이유가 되어서는 안 됨을 의미한다.

용어해설

RNN 동일한 모델을 순차적으로 적용하여 시퀀스 데이터를 처리하는 딥러닝 모델 구조

LSTM RNN 모델에서 시퀀스의 길이에 따라 정보 손실이 되는 현상을 방지하기 위해 최근 시퀀스에서 나타난 정보와 멀리 떨어진 과거에 나타난 정보를 분리하여 관리 및 적용하는 구조

CNN Convolution 연산을 통해 부분적으로 반복하여 나타나는 구조적 패턴을 검출하고, 이를 결합하여 개념적인 패턴을 추출하는 딥러닝 모델

GAN 생성자와 구분자 모델을 두어 상호 간 경쟁을 통해 전체 모델의 성능을 높이도록 설계된 딥러닝 모델

Adversarial Loss GAN 모델에서 생성자가 생성한 데이터를 구분자가 구분할 확률과 구분자가 모든 데이터를 구분할 확률을 결합한 손실 함수

샷 하나의 카메라 워크로 생성된 연속된 프레임들

장면 동일한 장소, 시간, 스토리를 가지는 연속된 샷들

약어 정리

CNN	Convolutional Neural Network
CP	Content Provider
GANs	Generative Adversarial Networks
LSTM	Long and Short Term Memory
RNN	Recurrent Neural Network
SSIM	Structural Similarity Index Measure

참고문헌

[1] I. J. Goodfellow et al., "Generative Adversarial Nets," in *Adv. Neural Inform. Process. Syst.*, Montreal, Canada, 2014, pp. 2672-2680.

[2] A. Odena, C. Olah, J. Shlens, "Conditional Image Synthesis with Auxiliary Classifier GANs," in *Proc. Int. Conf. Mach. Learning*, Sydney Australia, 2017, pp. 2642-2651.

[3] Z. Yi, H. Zhang, P. Tan, M. Gong, "DualGAN: Unsupervised Dual Learning for Image-to-Image Translation," in *Proc. IEEE Int. Conf. Comput. Vision*, Venice, Italy, Oct. 2017, pp. 2868-2876.

[4] rossgoodwin, "WordCar," GitHub, available: https://github.com/rossgoodwin/wordcar

[5] B. Liu, J. Fu, M. Kato, M. Yoshikawa, "Beyond Narrative Description: Generating Poetry from Images by Multi-Adversarial Training," in *Proc. ACM Int. Conf. Multimedia*, Seoul, Rep. of Korea, Oct. 2018, pp. 783-791.

[6] A. Mordvintsev, C. Olah, M. Tyka, "DeepDream - a Code Example for Visualizing Neural Networks," Google Research, 2015.

[7] R. Goodwin, O. Sharp, "Benjamin," http://benjamin-ai.tumblr.com.

[8] A. Lee, C. Kwak, J. Son, S. Kim, "SVIAS: Scene-segmented Video Information Annotation System," in *Proc. ACM Int. Conf. Multimedia*, Seoul, Rep. of Korea, Oct. 2018, pp. 1278-1269.

[9] C. Kwak, M. Han, S. Kim, G. Hahm, "Interactive Story Maker: Tagged Video Retrieval System for Video Re-creation Service," in *Proc. ACM Int. Conf. Multimedia*, Seoul, Rep. of Korea, Oct. 2018, pp. 1270-1271.

[10] R. Villegas, J. Yang, S. Hong, X. Lin, H. Lee, "Decomposing Motion and Content for Natural Video Sequence Prediction," in *Proc. Int. Conf. Learning Representations*, Toulon, France, Apr. 24-26, 2017.

[11] S. Tulyakov, M. Liu, X. Yang, J. Kautz, "MoCoGAN: Decomposing Motion and Content for Video Generation," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Salt Lake City, UT, USA, June 18-23, pp. 1526-1535.

[12] IBM THINK Blog, "IBM Research Takes Watson to Holloywood with the First Cognitive Movie Trailer," Aug. 31, 2016.

[13] J. Brandon, "Terrifying High-Tech Porn: Creepy 'Deepfake' Videos are on the Rise," Fox News, Feb. 20, 2018.

[14] S. Machkovech, "This Wild, AI-Generated Film is the Next Step in 'Whole-Movie Puppetry'," Arstechnica, June 12, 2018.

[15] 조원진, "부산시와 ETRI, 인터랙티브 영상 공모전 개최," 서울경제, June 11, 2018.

[16] A. Conner-Simons, R. Gordon, "Creating Videos of the Future," MIT News, Nov. 28, 2018.

[17] B. Lotter, G. Kreiman, D. Cox, "Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning," arXiv:1605.08104, 2016.