생성 AI 기술을 활용한 스토리텔링 영상 컨텐츠 제작*

소병욱, 윤선민[0], 김도희, 나경민, 최명걸
가톨릭대학교 미디어기술콘텐츠학과
{thquddnr123, tjsals914, dhdh2040, nkm3266, mgchoi}@catholic.ac.kr

Generating Story Telling Video Contents Using Generative AIs

Byunguk So, Seonmin Yoon [0], Dohee kim, Kyungmin Na, Myung Geol Choi
Dept. of Media Technology and Media Contents, The Catholic University of Korea

요약

본 연구에서는 창의적인 스토리텔링 영상 컨텐츠를 제작하기 위해 인공지능 기술을 활용하는 방법을 소개한다. 간단한 주제어 입력을 시작으로 시나리오를 생성하고, 생성된 시나리오에 기초하여 스토리에 부합하는 이미지를 생성한다. 이미지의 깊이 정보를 추정하여 입체감 효과 등의 특수효과를 적용하고, 이를 3D 애니메이션으로 렌더링한다. 각 단계에서 사용된 인공지능 기술과 사용 사례를 설명한다. 이를 통해 인간의 창의성은 최소한으로 발휘하여도 흥미롭고 창의적인 스토리텔링 콘텐츠를 제작할 수 있음을 검증한다. 또한 각 단계에서 아직 인공지능만으로 해결되지 못하고 사람의 개입이 필요한 부분을 확인하고, 향후 연구의 방향을 제시한다.

1. 서론

최근 발전하고 있는 AI 기술은 인간의 창작활동을 보조하는 역할을 넘어서 창작활동의 주체가 되어가고 있다. 특히, 최근 이슈가 되고 있는 ChatGPT 비롯한 대화형 생성 AI는 사용자가 입력한 텍스트에 기초하여 새로운 창작물을 생성할 수 있기 때문에 스토리텔링 콘텐츠 제작에 적합하다. 본 연구에서는 간단한 주제어를 입력하는 것에서부터 시작하여 ChatGPT를 포함한 여러 가지 인공지능 기술을 단계적으로 활용하여 수준 높은 스토리텔링 영상 콘텐츠를 제작한 사례를 소개한다. 이 과정에서 사람의 창의성은 최소한으로 발휘하는 것을 목표로 하였다. Table 1은 본 연구에서 사용된 기술을 단계별로 정리한 것이다. 회색으로 표시된 두 부분에서만 사람의 노력이 필요하였고, 나머지 부분은 인공지능 기술만으로 해결할 수 있었다. 각 단계에서 사용된 인공지능 기술과 사용 사례를 설명하고 한계점을 분석한다.

* 구두발표논문
* 본 논문은 학부생 발표 논문임
* 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2021R1F1A1048002)

Table 1 Technologies used in each stage and invested time.

Production Stage	Technology Used	Time Investment
Scenario Generation	ChatGPT3.5	0.17 hours
Image Prompt Command Generation	ChatGPT3.5	0.5 hours
Prompt Command Selection	Human effort	4 hours
Image Generation	Stable Diffusion v2.1, Sci-Fi-Diffusion v1.0 Model	
Image Depth Information Generation	Boosting monocular Depth	0.5 hours
Stereoscopic Special Effects	Blender 3.4, Human effort	10 hours
Narration Script Generation	ChatGPT4	1 hours
Narration Voice Generation	TypeCast	

2. 시나리오 생성

시나리오 생성에는 ChatGPT를 사용하였다. 시나리오 작성에서는 주제어 결정에만 사람의 창의력이 필요하다. 본 연구에서는 "*Polynesians and Space Exploration*"을 주제어로 결정하였다. 다음 Table 2와 같은 명령어(Prompt)를 입력하였다. ChatGPT는 이 요청에 기초하여 450 단어로 작성된 영문 시나리오를 생성하였다. 본 연구에서는 생성된 시나리오에 더이상 사람이 개입하지 않고, 이를 기초로 다음 단계를 진행하였다. ChatGPT에 다시 생성된 시나리오를 몇 문장으로 요약해달라 요청하여 얻은 결과는 Table 2와 요약과 같다.

Table 2 Prompt command used to generate our scenario and the summary of the scenario.

Prompt	Write a scenario about Polynesians and space exploration.
Summary	In 2045, Polynesian scientists launched a spacecraft, Te Mana O Te Moana, to explore space, inspired by their ancestors' star navigation skills. During their journey, they celebrated their culture and overcame challenges with ancestral wisdom. Upon exploring a planet, they made discoveries that inspired humanity upon their return to Earth.

3. 이미지 생성

생성된 시나리오에 기초하여 스토리에 부합하는 이미지를 생성하기 위해 Stable diffusion v2.1의 Sci-Fi-Diffusion v1.0 모델을 사용하였다. 이 모델은 SF영화와 유사한 이미지를 생성하도록 학습된 모델이다. 그림 생성 요청을 위한 프롬프트 명령어를 생성하기 위해 ChatGPT에 각 문단 별로 키워드 리스트를 작성해줄 것을 요청하였다. 자동 생성된 키워드를 그대로 사용하여 이미지를 생성하는 경우 기대에 미치지 못하는 결과가 생성되는 경우를 자주 발견하였다. 생성 모델 학습 데이터에 없었던 키워드에 대한 문제인 것으로 추정된다. 예를 들어 본 연구의 실험에서 사용된 생성 모델은 프롬프트에 "Polynesian"를 포함시키는 경우 그에 대한 내용을 그림으로 표현하지 못하고 문자로 된 제목이나 설명으로 표현하려는 경향이 있다. Figure 1은 키워드에 *Polynesian*을 포함시켰을 때와 아닐 때의 결과를 비교한 것이다. 이와 같이 문제를 피하기 위해 일부 생성된 프롬프트 키워드를 선택하여 제거하였다. 이 때 사람의 판단과 노력이 투입되었다. 총 7장의 1820 x 1024 크기 이미지를 생성하였다.

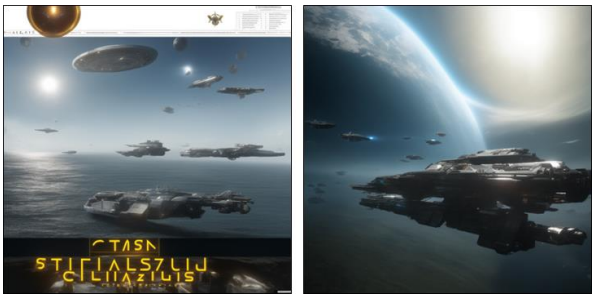

Figure 1 Generated images with (left) and without (right) *Polynesian* keyword.

4. 입체감 효과

이미지에 입체감 특수 효과를 적용하기 위해 먼저 생성된 이미지를 Boosting Monocular Depth (BMD) [1] 모델에 적용시켜 픽셀 단위의 깊이 값을 추정한다. BMD 모델은 고해상도 이미지에 대한 깊이 추정에 유리하도록 설계된 인공신경망으로 고해상도 콘텐츠 제작에 적합하다. 깊이 정보를 이용하여 해당 이미지에 대한 3차원 메시 모델을 생성하였다. Figure 2는 생성된 이미지와 추정된 깊이 정보를 바탕으로 3차원으로 복원된 결과이다. 다음 사람의 판단으로 장면마다 카메라 모션 설정, 라이팅 조건 변경 등을 통해 입체감을 높였다.

Figure 2 Generated 2D image (left) and its 3D mesh reconstructed (right) by using estimated depth map.

5. 오디오 생성

내레이션 생성을 위해 다시 한번 ChatGPT를 사용하였다. 최초 생성된 시나리오를 입력으로 하여 각 장면에 대한 내레이션 스크립트를 생성하도록 요청하였다. 별도의 한국어 콘텐츠를 제작하기 위해 생성된 스크립트를 DeepL 서비스를 사용하여 한국어로 번역하였다. 다음 생성된 스크립트를 읽는 오디오 데이터를 생성하기 위해 TypeCast 서비스를 사용하였다. 배경 음악의 경우 다양한 생성 모델을 사용하여 보았으나 주어진 스크립트의 내용에 부합하는 음악을 생성하는 것은 어려움이 있었다. 따라서 적합한 음원을 구매하여 사용하였다.

6. 결론

제작 과정을 통해 인공지능 기술로 창작된 스토리텔링 콘텐츠가 인간의 창의력으로 창작된 콘텐츠와 비교하여 충분히 흥미롭고 창의적일 수 있음을 검증하였다. Table 1의 세번째 열은 각 단계에서 투입된 대략의 시간을 보여준다. 3D 렌더링에 걸린 시간과 최종 영상 합성에 들어간 시간은 제외하였다. 사람의 노력이 투입되는 부분에서 대부분의 작업 시간이 소요되었음을 알 수 있다. 향후 유사한 작업에 대해 전체 작업 시간과 사람의 노력을 줄이기 위해서는 이미지 생성 모델의 특성에 맞는 키워드 생성 방법, 또는 부적합한 이미지 제거를 위해 생성된 이미지를 자동 평가하는 방법의 개발이 필요하다. 또한, 카메라 모션 설정 등 단순 작업에 대해서도 인공지능의 도움을 받을 수 있는 기술이 개발되면 작업의 효율성을 크게 높일 수 있을 것으로 기대된다.

참고문헌

[1] Miangoleh, S.M.H., Dille, S., Mai, L., Paris, S. and Aksoy, Y., Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging, *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.9685-9694.