NeRF 배경 장면 제작을 위한 모바일 카메라 궤적 분석 및 성능 향상 실험

강준구[1], 유경민[1], 손보성[1], 박노갑[2], 박용현[2], 김현지[2], 홍현기*

중앙대학교[1,*], SK 텔레콤[2]

{engineerjkk[1], ykmcau[1], sonbosung[1], honghk*}@cau.ac.kr, {tony.nokap.park[2], calm.ardent[2], hyeonji[2]}@sk.com

Analysis of Mobile Camera Trajectories and Performance Enhancement Experiments for NeRF Background Scene Production

Junekoo Kang[1], Kyeongmin Yu[1], Bosung Sonn[1], Nokap Park[2], Yonghyun Park[2], Hyeonji Kim[2], Hyunki Hong*

Chungang Univ[1,*], SK Telecom[2]

요 약

본 논문은 실사 및 컴퓨터 그래픽 요소를 합성하기 위해 사용되는 Neural Radiance Fields(NeRF) 기술을 분석한다. NeRF는 2D 이미지를 학습하여 어떤 각도에서든 사실적인 장면을 생성하는 인공지능 기법이다. 이를 통해 배경의 3D 가상 공간을 구현하는 과정에서 소요되는 비용과 시간을 크게 줄일 수 있다. 본 논문에서는 NeRF를 통해 생성한 3D 배경 장면의 성능 향상을 위해 궤적 등의 조건에 따라 생성된 결과 영상의 화질 등을 분석하고, NeRF의 적절한 입력 영상 촬영 방법 등을 제시한다.

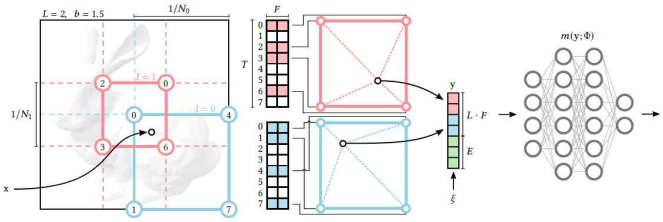
Ⅰ. 서 론

합성 콘텐츠를 제작[1]하는 과정에서는 컴퓨터 그래픽스 기술과 실제로 촬영된 영상 등이 결합된다. 이것은 3차원 모델링[2] 및 렌더링으로 생성되는데, 이 과정에서 많은 비용과 시간이 소요된다. 또한 3D 스캐너를 활용한 3D 모델링 방식이 활용될 수 있으나, 공간에 제약이 있다. 최근 등장한 NeRF[3]는 이러한 문제를 해결해 3D 모델링에 비약적인 발전을 가져왔다. NeRF는 미리 촬영한 2D 이미지를 COLMAP[4]을 통해 카메라 파라미터를 추출한 다음, 이미지와 카메라 파라미터를 함께 학습하여 어떤 각도에서든 사실적인 뷰를 새롭게 생성하는 인공지능 기법이다. 하지만 입력 영상에 따라 최종 생성된 결과 영상의 성능에 많은 영향을 준다. 본 논문에서는 3D 배경 장면 제작을 위한 NeRF 입력 영상 가이드를 실험 및 분석을 통해 제시한다. 추가적인 실험 결과들은 다음 프로젝트 웹사이트에서 확인할 수 있다. https://sonbosung.github.io/NeRFMaverick/

Ⅱ. 본론

Ⅱ-1. Instant-NGP[5]

본 논문에서 실험에 사용한 NeRF[3] 모델은 Instant-NGP[5] 이다. Instant-NGP는 기존 NeRF의 Positional encoding을 Mulri-resolution hash encoding으로 대체한 방식이다.



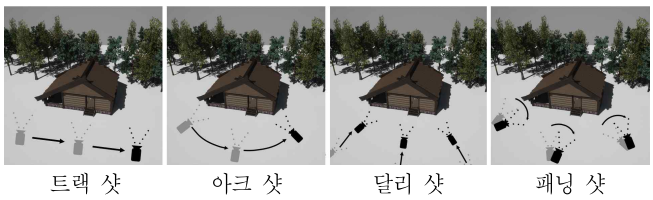
[그림 1] Instant-NGP[5]의 Multi-resolution hash encoding

카메라 Ray상의 샘플 포인트가 Multi-level 복셀에 입력되면서 꼭짓점의 거리에 따라 다른 가중치로 삼선형 보간 연산을 수행함으로써 인코딩해 더 많은 정보를 포함하고 빠른 속도로 학습이 가능하다.

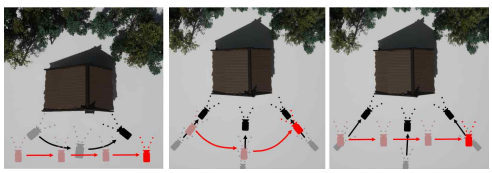
Ⅱ-2. NeRF 배경 장면 제작을 위한 가이드 라인.

NeRF를 이용해 가상 배경 장면을 생성하는 경우, 다음 다섯 가지를 고려해야 한다. 첫째, 빠른 속도로 촬영 시 모션 블러가 생겨 3D 재구성의 품질을 크게 저하시킬 수 있으므로, 모바일 기기를 천천히 움직인다. 둘째, 정지된 상태에서 카메라를 회전하기보다는 카메라를 이동하면서 촬영해야 한다. 셋째, 거울 및 투명한 객체, 특징점을 찾기 어려운 흰 벽 등, 장면 내 물체 표면의 재질에 영향을 받는다. 넷째, 어두운 조명환경을 피하고 텍스쳐를 온전히 식별할 수 있는 조명 조건에서 촬영한다. 다섯째, 나무나 사람과 같은 동적인 물체는 NeRF 최종 품질을 저하시키므로, 최대한 정적인 환경을 촬영한다.

Ⅱ-3. NeRF 배경 장면 제작 카메라 궤적 실험



트랙 샷 아크 샷 달리 샷 패닝 샷



트랙+아크 샷 아크+달리 샷 트랙+달리 샷

[그림 2] 실험 카메라 궤적

NeRF를 이용하여 가상 배경 장면을 생성하는 경우, [그림 2]는 모바일 기기를 통한 입력 영상의 일반적인 촬영 방법이다. 이러한 카메라 궤적에 의해 촬영된 입력 영상을 이용하여 3D Point를 2D 이미지에 재투영하여 오차를 구하는 R/E(Reprojection error) 값과 생성된 NeRF 영상을 원본 영상과 비교하여 정량적 값(PSNR, SSIM, LPIPS, BRISQUE)으로 평가한다.

평가		서울 잠실 트레비 분수 (근경)					서울 올림픽공원 엄지손가락 (근경+원경)				
		PSNR↑	SSIM↑	LPIPS↓	BRISQUE↓	R/E↓	PSNR↑	SSIM↑	LPIPS↓	BRISQUE↓	R/E↓
가	정지 영상	22.22	0.70	0.24	43.81	0.78	19.51	0.64	0.43	38.17	0.97
	2 fps	21.21	0.70	0.25	37.71	0.82	23.34	0.80	0.18	43.31	0.77
	5 fps	22.08	0.74	0.27	50.57	1.02	23.60	0.80	0.17	43.77	0.87
	10 fps	21.74	0.73	0.28	50.53	1.13	24.71	0.82	0.16	43.94	0.97
나	아크 샷	22.08	0.74	0.27	50.57	1.02	23.60	0.80	0.17	43.77	0.87
	트랙 샷	18.79	0.61	0.43	53.05	1.02	18.27	0.77	0.40	49.37	0.93
	패닝 샷	-	-	-	-	-	-	-	-	-	-
	달리 샷	18.69	0.62	0.24	42.35	1.06	19.57	0.75	0.43	40.74	0.99
	트랙+아크 샷	15.71	0.59	0.45	33.46	1.01	30.34	0.92	0.18	43.13	0.88
	아크+달리 샷	17.91	0.68	0.33	52.31	1.02	20.81	0.79	0.32	43.96	0.95
	트랙+달리 샷	23.66	0.78	0.22	44.07	0.97	19.07	0.78	0.43	48.06	0.99

[표 1] 각 데이터셋 별 실험 결과. 가 : 정지 영상과 동영상 비교, 나 : 카메라 궤적별 비교, 다 : 모바일 기기별 비교.

Ⅱ-4. 데이터셋


트레비 분수


엄지손가락

[그림 3] 촬영 데이터셋

실험 영상의 촬영 장소로써 서울 잠실에 위치한 트레비 분수는 실내 환경으로 근경이다. 정지 영상 40장과 동영상 2, 5, 10 fps(frame per second)로 촬영되었으며, 각각 40, 100, 200장이다. 서울 올림픽공원에 위치한 엄지손가락은 가까이 객체가 있는 실외 환경으로써 근경과 원경을 모두 포함하고 있다. 정지 영상 80장과 동영상 2, 5, 10 fps로 각각 80, 200, 400장으로 구성되어 있다. 모두 동일한 해상도(1920x1080)이며 , 동영상과 달리 정지 영상은 GPS와 카메라 초점 거리가 추가로 제공된다. 모든 데이터셋은 아이폰(iPhone 13 Pro Max)으로 촬영되었다.

Ⅱ-5. 실험

[표 1]의 (가) 실험은 정지영상과 동영상의 성능 비교 실험이다. 공정한 실험을 위해 정지 영상과 2 fps에서는 이미지 수가 같다. (나) 실험은 카메라 궤적에 따른 성능 평가이다. 아이폰, 5 fps로 동일한 조건에서 실험을 진행했다. 빨강 볼드는 가장 좋은 성능을 보인 결과이며, 검정 볼드는 두 번째로 좋은 성능을 보인 결과이다.

Ⅱ-6. 실험 결과

(가) 실험에서 정지영상은 가까운 객체에 포커싱되고 배경에는 아웃 포커싱 기능이 있다. 실내에서 촬영한 트레비 분수는 근경으로만 구성되어 있기 때문에 정지 영상에서 아웃 포커싱 없이 전체적으로 선명하므로 일관성 있는 특징점 추출과 매칭으로 더 정확한 카메라 포즈 추정이 가능하여 R/E가 가장 낮다. 따라서 정확한 위치에서 NeRF를 학습했기 때문에 높은 성능을 보인다. 또한 동영상의 경우 근경이기 때문에 카메라 프레임마다 급격한 변화가 있어 모션 블러가 생겨 선명도가 낮다. 이는 결국 R/E와 결과 영상 성능에 불리한 영향을 미치게 된다. 하지만 실외에서 촬영한 엄지손가락은 정지 영상에서 객체는 포커싱되어 선명하나, 배경에서 아웃 포커싱이 적용되어 선명하지 않다. 이와 반대로 상대적으로 일관성이 있는 해상도로 촬영한 동영상에서는 전체적으로 높은 성능을 보인다. 동영상 2 fps에서는 카메라 베이스라인이 10 fps보다 크기 때문에 삼각법 과정에서 오차범위가 더 적다. 이러한 이유로 R/E에러가 더 낮으나, 그럼에도 불구하고 NeRF 렌더링 시에는 더 많은 데이터셋을 통해 다양한 뷰를 학습하는 것이 중요하기 때문에 10 fps에서 가장 좋은 성능을 보인다.

(나) 카메라 궤적별 비교 실험이다. 트레비 분수는 트랙과 달리 샷을 조합한 궤적에서 가장 높은 성능을 보인다. 트랙의 경우 이미지가 담고 있는 객체의 균형이 높으며, 근경이기 때문에 달리 샷에서도 확실한 프레임 변화가 있다. 이러한 이유로 R/E가 낮고 렌더링 된 결과 영상의 성능이 높다. 하지만 엄지손가락 데이터셋은 트랙의 경우 모든 프레임에서 객체가 담기는 균형이 맞지 않으므로, 많은 이미지에서 객체 대비 배경을 많이 포함하게 되어 낮은 성능을 보인다. 또한 달리 샷은 원경에서 프레임 간의 변화가 적기 때문에 낮은 성능을 보인다. 하지만 원경에서는 모션 블러의 영향이 적다는 장점이 있다. 따라서 가장 좋은 성능을 보인 궤적은 모든 프레임의 중심에 객체가 있어 다양한 뷰에서 학습할 수 있는 아크 샷이 포함된 궤적에서 가장 높은 성능이 보임을 알 수 있다. 마지막으로 모든 데이터셋에서 고정된 위치에서 회전하는 패닝 샷은 카메라 궤적을 해석하지 못해 COLMAP[4]를 실패하였다.

Ⅲ. 결론

본 논문에서는 일반 사용자가 자신의 모바일 기기를 통해 3D 가상 배경 장면을 제작할 때, 필요한 NeRF 입력 영상 매뉴얼 및 가이드를 제공한다. 이를 통해 NeRF에 대한 지식이 부족하여도, 본 논문의 가이드를 이용해 입력 영상을 구성하고, 더욱 나은 NeRF 렌더링 결과를 얻을 수 있다.

ACKNOWLEDGMENT

이 논문은 2023년도 SK텔레콤의 재원으로 진행된 SKT AI Fellowship 과정에서 수행된 공동연구결과임 (02.AI기반 고화질 3D 변환 기술 연구)

참 고 문 헌

[1] 김미라. “포스트 코로나, 영상 콘텐츠 제작 기술: 가상 제작 (Virtual Production) 시스템,” *영상기술연구,* 1(35), pp. 27-44. 2021

[2] 오민정, 서용덕. “언리얼 엔진(Unreal Engine) 기반의 2D 이미지를 이용한 가상공간 제작 연구,” *한국방송미디어공학회 학술발표대회 논문집,* pp 175-176. 2022

[3] Ben M, Pratul P. S. Matthew T., Jonathan T. B., Ravi R and Ren N."NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis," *ECCV* 2020.

[4] Johannes L. S.,Jan M. F. "Structurefrom-motion revisited," *In Proceedings of the IEEE conference on computer vision and pattern recognition,* pp 4104-4113, 2016.

[5] Thomas M., Alex E, Christoph S. and Alexander K. "Instant Neural Graphics Primitives with a Multiresolution Hash Encoding," *ACM Trans. Graph.* 2022.