졸업논문
소개

생성형 AI를 사용한 3차원 체적 모델 합성에 관한 연구

박병서 / 광운대학교 Intelligent Computing Lab.

최근 생성형 AI 분야의 연구에서 자연어로 작성된 텍스트 프롬프트를 이용해 다양한 고품질의 영상 합성을 가능하게 하는 Text-to-Image 모델이 발표되어 영상 합성 분야의 비약적인 발전을 이루고 있다. 최근, 생성형 AI들은 다양한 분야에서 다양한 용도로 널리 활용되고 있다. 이와 같은 생성형 AI에서 확산 모델(Diffusion Model)은 영상 합성과 초고해상도 분야에서 굉장히 좋은 성능을 보였다. 해당 분야에서 구축된 모델의 학습을 위해서는 많은 컴퓨팅 리소스가 필요하다. DDPM(Denoising Diffusion Probabilistic Model) 등의 LDM(Latent Diffusion Model)을 지원하는 샘플러들은 초기 노이즈 제거 단계에서 적은 양의 데이터를 샘플링하여 컴퓨팅 파워 문제를 해결하고자 시도하였다. 또한 이들은 모델의 성능은 동등하지만 계산적으로 더 적합한 잠재 공간을 찾아서 고해상도 영상 합성을 위한 확산 모델을 학습시켰다. 파인 튜닝(Fine-tuning)은 미리 학습된 모델(Pretrained model)을 기반으로 추가적으로 새로운 데이터 셋을 학습하여 모델의 파라미터를 미세 조정하는 것을 의미하며, 처음부터 모델을 학습할 필요가 없기 때문에 기존 모델에 이어서 빠르게 학습시킬 수 있어 더욱 효율적이다. LDM에서는 파인 튜닝이 가능한 영역은 U-Net과 텍스트 인코더 영역이며, 드림부스(Dreambooth)는 두 가지 영역 모두에서 파인 튜닝을 지원한다. Ben Mildenhall 등은 NeRF(Neural Radiance Fields)라는 새로운 3차원 체적 모델의 표현 방법을 제안하였다. 5차원 데이터를 입력받아 객체의 새로운 시점에서의 모습을 알아내는 방법을 통해, 딥러닝 알고리즘을 기반으로 입력 데이터로부터 밝기와 밀도를 추출하는 함수를 만든다.

최근에 RGB와 깊이 센서가 결합된 RGB-D 센서가 보편화되어 다양한 분야에서 널리 사용되고 있다. RGB-D 센서를 사용하면 촬영되는 객체의 형태를 비교적 정확하고 빠르게 확인할 수 있다. RGB-D 센서는 SLAM(Simultaneous Localization and Mapping)과 내비게이션, 물체추적, 물체인식과 위치식별, 자세추정 그리고 3차원 체적 모델 합성 등과 같은 다양한 분야를 급속히 발전시켜 왔다. RGB-D 센서에서 컬러 정보는 RGB 카메라를 사용하여 획득된다. 반면에 깊이 영상은 레이저 거리 측정 스캐너, 그리고 구조광 방식의 센서와 같은 다양한 형태의 센서를 통하여 획득된다. RGB-D 센서를 사용하여 정확하고 신뢰성 있는 3차원 volume을

reconstruction하기 위해서는 각 카메라의 내부 파라미터 계산과 두 센서 사이의 외부 파라미터 계산이 필요하다. 카메라가 생산되는 시점에서 카메라에 포함된 메모리에 룩업테이블의 형태로 캘리브레이션에 필요한 파라미터가 미리 제공되는 형태도 있다. 다수의 RGB-D 센서를 사용하여 촬영을 수행하는 실시간 스캐닝과 3차원 체적 영상 촬영 같은 응용 분야에서 다수의 카메라들 사이의 위치 및 자세 추정은 매우 중요한 문제이다. 카메라의 정확한 내·외부 파라미터를 획득하기 위해서 다양한 연구들이 수행되어 왔다.

따라서 본 논문은 다시점 RGB-D 카메라를 이용하여 reconstruction된 3D point cloud의 품질을 개선하기 위해 LDM을 이용하는 방법을 시도하고자 한다. AI 모델은 무수히 많고, 이와 같이 많은 AI들 중에서 특정 문제를 해결하는데 어떤 AI가 가장 적합한 솔루션인지를 이론적으로 결정하기는 매우 어렵다. 우리의 접근 방식은 최근에 널리 이용되는 AI 모델 중의 하나인 LDM을 선택한 후에, 이것이 우리의 문제를 해결하는데 유용한지를 확인하는 것이다.

본 논문에서는 정렬되지 않은 다수 개의 구조광 방식의 RGB-D 센서가 설치된 촬영 시스템을 위한 외부 캘리브레이션 기법 및 생성형 AI를 사용한 3D 체적 모델의 합성 방법을 제안한다. 우리의 방식은 크게 세 가지 단계의 알고리즘들로 구성된다. 첫 번째는 다시점 카메라의 반복적 캘리브레이션 방법을 도입하여 캘리브레이션 과정에서 발생하는 파라미터의 오차를 최소화하는 과정이다. 두 번째는 깊이 정보의 필터링을 통해 구조광 카메라가 갖는 거리에 따른 깊이 영상의 왜곡과 잡음을 효율적으로 완화시킴으로써 캘리브레이션 결과와 합성된 포인트 클라우드의 품질을 향상시키는 과정이다. 이 두 가지 과정은 이전 연구에서 제안된 것들이다. 세 번째는 이 시스템에 적합한 AI 학습 모델과 학습 방법을 이용하여 카메라들 사이의 시점의 연속성을 증가시키고 표면 정보를 정확히 합성하는 것이다.

본 논문은 다수의 RGB-D 카메라들을 효율적으로 캘리브레이션한 후에, 이 카메라들로부터 입력된 데이터들을 3D point cloud로 3D reconstruction을 수행하는 알고리즘을 사용하였다. 이 기법은 강력한 캘리브레이션 및 reconstruction 솔루션을 제공하고 있지만, RGB-D 카메라가 필연적으로 발생시키는 depth의 오차로 인해 발생하는 point cloud의 outlier를 제거하고 mesh를 생성하는데 한계점을 갖고 있었다. 따라서 본 논문에서는 최근 급격히 발달하고 있는 딥러닝 기술을 이용하여 이와 같은 한계를 극복하기 위한 해법을 제시해 보고자 하였다.

본 논문에서 사용하는 3차원 포인트 클라우드를 생성하는 과정은 크게 Depth Processing, Iterative Calibration, Registration의 세 과정으로 나뉜다. 1) Depth Processing: 저가의 구조광 센서로부터 센싱된 깊이 영상은 많은 잡음 성분을 가지고 있기 때문에, 깊이 영상에 포함된 잡음에 적합한 처리를 수행하여 깊이 영상의 오차를 제거한다. 각 카메라별로 잡음이 제거된 깊이 영상은 RGB 영상을 기준으로 캘리브레이션을 수행한다. 2) Iterative Calibration: 이 과정은 크게 두 단계로 구성되는데 위와 아래에 배치된 카메라들 간에 공유된 특징점을 찾고 이 정보를 이용하여 외부 파라미터를 구한다. 다음으로 네 개의 시점들 사이의 특징점 위치를 통해 카메라 위치를 예측하는 외부 파라미터를 계산한다. 이 과정은 오차가 수렴될 때까지 반복적으로 수행된다. 3) Registration: Iterative Calibration 과정을 통해 획득된 외부 파라미터를 사용하여 모든 포인트 클라우드를 통합하고, 포인트 클라우드에 대한 개선 과정을 수행하여 포인트 클라우드의 품질을 향상한다. 이 과정을 통해 포인트 클라우드를 획득하고, 다음으로 메시 시퀀스를 형성한다.

모든 시퀀스 구성이 완료된 뒤 메시 품질 향상을 위해 OpenGL 등 3차원 데이터를 표시하는 그래픽스 파이프라인의 카메라 시점에서 3차원 체적 모델의 6방향

(정면, 좌측면, 후면, 우측면, 상, 하)의 깊이 영상을 획득하여 LDM의 입력으로 사용한다. LDM의 인퍼런스는 각 RGB-D 센서의 깊이 영상을 사용해 U-Net을 학습하고 이 가중치와 실제 피사체의 형태를 컨디션화할 수 있도록 학습한 Control-Net 가중치를 함께 사용한다. 보다 효율적인 학습을 위해 고유한 문자(Unique Identifier)에 할당하여 고품질 깊이 영상의 특징을 전이(Style Transfer)하도록 구성하였다. 3차원 체적 모델로부터 획득된 각 깊이 영상은 VAE를 통해 전체 입력 해상도의 1/8사이즈의 잠재 벡터로 전환된다. 깊이 영상으로부터 전환된 잠재 벡터는 Control-Net을 거쳐 컨디션화 된다. 각 U-Net 입력은 샘플러(DDPM, DDIM)를 거쳐 순 전파(Forward Propagation) 단계에서 점진적으로 노이즈가 추가되고, 역 전파(Backward Propagation) 단계에서 각각의 타임스탭마다 반복적으로 잠재 벡터에 포함된 노이즈의 양을 추정하고 복원하는 과정을 거치게 된다. 샘플러를 통과한 잠재 벡터는 VAE의 디코더를 통해 다시 영상 도메인으로 전환되고 이때 품질이 향상된 각 시점별 깊이 영상을 획득하게 된다. 각 시점에서 획득된 개선된 깊이 영상으로부터 각 픽셀별 깊이 값만큼 피사체를 향해 레이를 생성하고 각 레이의 길이 평균을 통해 해당하는 메시 표면의 정점 위치를 조정한다.

제안한 방법을 통해 3차원 차루코 보드의 보정 오류가 약 0.00926 mm로 줄어들 수 있었다. 물리적 거리를 고려하면 제안한 방법을 사용한 보정을 통해 카메라 간의 위치를 거의 정확하게 찾아낼 수 있음을 의미한다. 모아이상의 경우 정합 결과의 정확도를 측정하였으며, 본 실험에서는 오차 평균과 표준 편차가 각각 약 8 mm, 3.9 mm 이하로 감소할 수 있음을 확인하였다. 이때 양방향 필터링과 포인트 클라우드 정제 알고리즘을 통해 정합 결과를 향상시킬 수 있었다. 이 후 LDM을 이용한 메시 표면의 개선을 진행하였다. 그래픽스 파이프라인으로부터 획득된 깊이 영상을 컨디션으로 전환해 인퍼런스 과정에 적용한 결과 실험 데이터에서 오차 평균은 54.8% 표준 편차는 65.9% 향상된 결과를 확인하였다. 이후 3차원 체적 데이터 처리를 위한 대응을 사용한 메시 표면의 비강체 변형에 대한 동적 재구성 알고리즘을 적용하였다. 제안된 알고리즘은 900개의 프레임으로 구성된 3차원 체적 모델을 사용하여 검증되었다. 본 논문에서 제안한 동적 합성 알고리즘을 적용한 결과 체적 모델이 변형된 이후 타겟 프레임과 오차 평균은 0.23 mm, 표준 편차는 0.13 mm로 높은 정확도를 보였다. 또한, 제안한 방법은 기존 연구와 비교하여 최대 98.88%, 최소 20.39%의 향상된 정확도를 보였다.

본 논문의 분산 카메라 시스템과 제안하는 알고리즘은 비교적 단순하고 적은 수의 카메라를 사용하지만, 높은 품질의 3차원 체적 모델을 합성하여 비디오 게임이나 영화 제작 같은 응용 분야에서 높은 활용도를 보일 수 있도록 하는 것을 목표로 구성되었다. 실험을 통해 제안하는 방식이 3차원 체적 모델의 구조를 단순화하고 프레임별 높은 일관성을 확보하며 좋은 품질의 표면을 구성할 수 있음을 확인하였다.

후속 연구로는 LDM의 U-Net이 반복적인 샘플링을 통해 깊이 정보를 재구성하는 구간의 병목을 줄여 실시간(초당 30프레임) 랜더링이 가능하게 하는 것을 목표로 연구가 진행될 예정이다.

졸업논문
소개

박 병 서

- 2019년 2월 : 광운대학교 경영학 학사
- 2024년 2월 : 광운대학교 전자재료공학과 박사 (석박통합과정)