다차원영상기술표준화포럼표준(국문표준)

XDFK_01.0044/R0 제정일 : 2023년 11월 20일

국문 표준명	딥러닝 기반 3차원 재구성 모델의 가중치 공유를 사용한 6자유도 가상현실 영상합성
영문 표준명	6DoF Video Synthesis Using Parameter Sharing of Deep Learning-based 3D Reconstruction Models
표준초안 작성자	최재열(성균관대학교), 류은석(성균관대학교)

XD FORUM
다차원영상기술표준화포럼

서 문

1 표준의 목적

이 표준의 목적은 딥러닝 기반 3차원 재구성 기술을 응용하여 다수의 시점으로 구성된 몰입형 영상으로부터 가상현실 (virtual reality, VR) 을 위한 6자유도 사용자 정의 시점의 가상현실 영상을 생성하는 시스템을 제공하는 데 있다. 이 때 프레임간 인공신경망 가중치 공유 기법을 적용하여 학습 시간을 단축하고 프레임간 일관성이 향상되도록 하는 것, 인공신경망 내 일부 계층을 동결하여 불필요한 연산량을 감소시키는 것 또한 이 표준의 목적이다.

2 주요 내용 요약

이 표준은 다수의 시점으로 구성된 몰입형 영상으로부터 3차원 동적 장면 재구성을 위해 신경 복사 필드 (neural radiance fields, NeRF) 딥러닝 모델을 사용한다. 해당 모델을 사용한다면 깊이 정보의 입력 없이 색상 정보만으로 장면을 재구성할 수 있으며, 볼륨 렌더링 기법을 통해 임의의 시점에서의 이미지를 렌더링할 수 있다. 프레임별로 개별적 모델을 사용하되, 인공신경망 내 가중치 공유 기법 및 가중치 동결 기법을 적용하여 동일 학습 시간 대비 렌더링 이미지의 품질 향상 효과를 갖도록 하였다.

3 인용 표준과의 비교

이 표준은 MPEG 비디오 시스템 표준 및 특정 신경 복사 필드 모델과 관련성이 없는 범용의 표준임.

목 차

딥러닝 기반 3차원 재구성 모델의 가중치 공유를 사용한 6자유도 가상현실 영상합성
(6DoF Video Synthesis Using Parameter Sharing of Deep Learning-based 3D Reconstruction Models)

1 적용 범위

본 표준의 적용 범위는 딥러닝 모델을 사용한 몰입형 영상의 표현 및 입출력이며, 이는 파일 포맷을 통한 정보전달 규격을 포함할 수 있다. 또한 본 표준은 실감형 미디어 서비스의 종단 간 (end-to-end) 시스템에 적용될 수 있다.

2 인용 표준

해당 사항 없음.

3 용어 정의

3.1 Neural Radiance Fields
3 차원 공간의 위치 정보와 바라보는 방향 정보를 입력으로 하고, 해당 점의 색상과 밀도를 출력으로 하는 딥러닝 모델. 이미지 집합으로부터 3 차원 재구성을 실시하는데 사용됨.

3.2 Structure from Motion
동일한 객체를 중첩되도록 촬영한 다시점 이미지들로부터 카메라 매개변수를 복원하는 알고리즘. 특성 추출, 특징 매칭. 재구성 단계로 이루어짐.

3.3 Signal-to-noise Ratio
최대 신호에서 잡음 비율. 생성 혹은 손실 압축된 영상의 화질에 대하여 손실 정보를 평가하는데 사용됨.

3.4 Structural Similarity Index
두 이미지 사이의 유사도를 휘도, 밝기의 대조, 픽셀의 구조적 차이를 이용하여 평가하는 방법. 일반적으로 RGB, YUV 각 요소에 가중치를 두어 사용됨.

3.5 Learned Perceptual Image Patch Similarity

VGG 네트워크를 통해 두 이미지의 추출된 특성 사이의 비교를 통해 유사도를 측정하는 방법.

3.6 전이 학습 (Transfer Learning)

특정 작업을 위한 기계학습 모델의 훈련을 위해 관련된 작업에 대해 훈련된 모델의 일부를 가져와 재사용하는 기법.

3.6 미세 조정 (Fine Tuning)

사전에 학습된 기계학습 모델을 새로운 문제에 적용하기 위해 일부 가중치를 조정하는 학습 과정.

4 약어

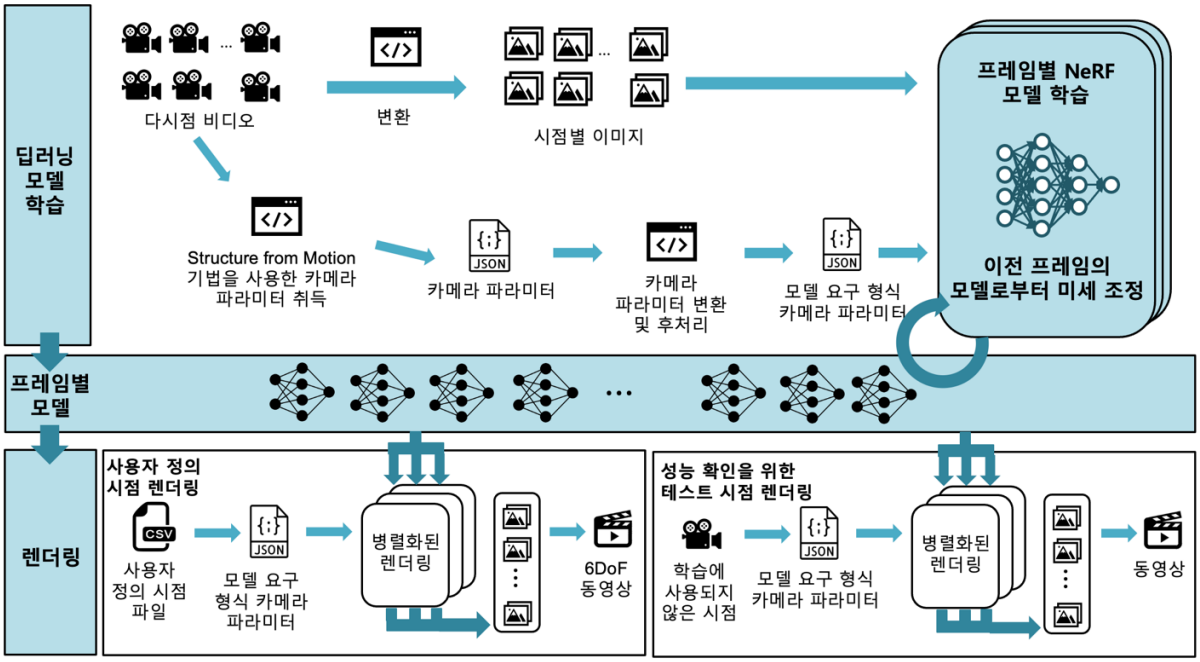
6DoF	6 Degrees of Freedom
FOV	Field of View
HMD	Head Mounted Display
LPIPS	Learned Perceptual Image Patch Similarity
MIV	MPEG Immersive Video
MLP	Multi-Layer Perceptron
NeRF	Neural Radiance Fields
RVS	Reference View Synthesizer
SfM	Structure from Motion
SSIM	Structural Similarity Index
VWS	View Weighting Synthesizer

5 딥러닝 기반 3차원 재구성 모델을 사용한 6자유도 가상현실 영상합성

5.1 다중 인공신경망 모델을 사용한 영상처리 프레임워크

NeRF를 사용한다면 인공신경망을 통해 3차원 장면을 복원할 수 있다. NeRF의 학습은 다시점의 2차원 이미지와 이에 해당하는 카메라 매개변수를 요구한다. 여러 시점을 나타내는 2차원 이미지로부터 특정 픽셀의 집합을 batch 단위로 선택하여 해당 픽셀로부터 장면의 방향으로 출발하는 광선 (ray) 를 구성한다. Ray상에서 점들을 샘플링하여 MLP에 대입하게 되면 MLP는 각 점들의 색상 및 밀도 정보를 출력하며, 샘플링된 모든 점들의 정보를 종합하여 한 ray에 대한 색상 예측값을 계산할 수 있다. 이 기법을 볼륨 렌더링 (volume rendering) 이라고 한다. 실제값과 예측값 간 오차를 계산하여 MLP에 역전파 (back-propagation) 를 수행하여 모델을 학습시킬 수 있다. 이와 같은 볼륨 렌더링 기반의 접근법은 깊이 영상 (depth map) 을 요구하지 않는다는 점에서 기존 중간 시점 합성 방식에 비해 편의성을 제공하며 시점 간 거리 값이 클 경우 정확성을 제공한다.
여러 프레임에 대해 NeRF 모델을 학습한다면 3차원 재구성을 통해 동적 장면을 표현할 수 있으며 이를 활용한다면 전방위 6자유도 동영상을 출력할 수 있다. 본 표준의 프레임워크는 시간 동기화가 이루어진 다수의 카메라를 사용해 취득한 몰입형 영상 및 사용자 시점 정보를 입력으로 하며, 해당 사용자 시점에 대한 6자유도 동영상을 출력으로 한다. (그림 5-1)은 본 표준의 프레임워크에 대한 구조도를 보여준다. 우선 각 영상으로부터 각 시점의 카메라 매개변수를 추정하는 과정을 거치는데, 이 때 특징점으로부터 카메라 간 관계를 추정하는 SfM 알고리즘이 적용된다. 카메라 위치와 방향을 나타내는 외부 매개변수 및 카메라 고유의 초점 거리 및 FOV 등을 영상으로부터 추정하는 과정이다. 하지만 상용 SfM 소프트웨어에 따라 카메라 매개변수의 좌표계 및 포맷이 상이하게 된다. 또한 학습 및 렌더링에 사용되는 NeRF 모델에 따라서도 상이한 카메라 좌표계를 요구한다. 따라서 SfM 소프트웨어의 출력 카메라 매개변수를 특정 NeRF 모델에서 요구하는 카메라 좌표계에 맞추어 변환하는 소프트웨어가 요구된다. 이 변환 과정을 거친 후 특정 NeRF 모델로부터 학습이 일어난다. 상기 과정은 특정 NeRF 모델에 종속되지 않고 통상의 딥러닝 기반 정적 3차원 재구성 모델에 전역적으로 적용될 수 있다.
학습 결과 프레임별 개별 모델이 산출되고, 특정 데이터셋을 통해 한 번 학습된 모델은 지속적으로 디바이스에 보관되며 사용자 정의 시점이 주어지는 대로 렌더링 될 수 있다. 모델은 별도의 파일 (예: pytorch 의 pt , pth 또는 tensorflow의 .tf 등) 로 저장 및 전송될 수 있다. 렌더링 과정은 디바이스의 GPU개수에 비례하여 병렬화 될 수 있으며 개별 프레임에서 임의의 시점에 따라 렌더링된 결과물 이미지가 병합되어 6자유도 동영상이 생성된다. 본 표준에서는 두 가지 렌더링 모드를 지원한다. 첫 번째는 앞서 언급한 사용자 시점 렌더링 모드이다. HMD 를 착용한 사용자의 머리 회전, 사용자의 위치 이동을 나타내는 정보 (x, y, z, yaw, pitch, roll) 를 각 프레임별로 입력 받고, 저장된 모델을 이용하여 각 프레임의 영상을 렌더링한다. 두 번째 모드는 성능 측정을 위해 실제값 (ground truth)와 비교 가능한 학습에 사용되지 않은 테스트셋 (test set) 렌더링 방법이다.

상기 방법에서는 원본 동영상과의 PSNR, SSIM, LPIPS 등의 지표를 측정함으로써 실험에 사용한 모델에 대한 성능을 측정할 수 있다. 이 경우에는 가변형 사용자 시점 정보 대신 고정된 카메라 위치와 내부 매개변수를 요구한다.

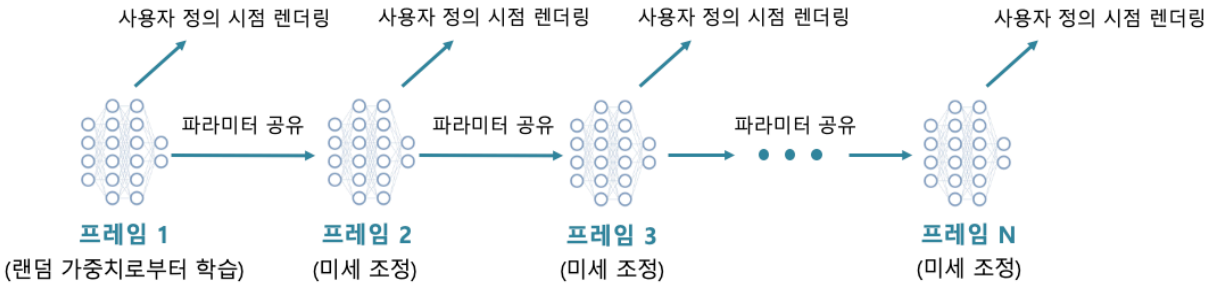


(그림 5-1) 인공신경망 모델을 사용한 6자유도 영상처리 프레임워크 구조도

5.2 가중치 공유 기법을 적용한 렌더링 품질 향상

본 절은 프레임을 시간 순으로 나타냈을 때 인접 프레임간 유사성을 고려한 가중치 공유 기법을 소개한다. (그림 5-2)는 해당 기법의 개념도를 나타낸다. 첫 프레임은 아무런 참조를 하지 않고 무작위로 초기화된 인공신경망의 가중치로부터 학습을 시작한다. 무작위로 초기화된 가중치로부터 최적화를 진행해야 하기에 상대적으로 높은 학습 횟수 설정이 권장된다. 이후 프레임 (t > 0) 부터는 프레임 t에 대해 프레임 t-1을 참조한다. 즉, t번째 프레임을 학습시킬 때 t-1번째 프레임에서 학습된 모델의 가중치로부터 미세 조정 (fine-tuning) 한다. 인접 프레임간 데이터셋이 유사하다는 것을 고려했을 때, 모델의 최적화를 위해 진행해야 할 학습량이 현저히 감소한다. 즉, 프레임 단위로 모델을 구성 시, 동일 횟수의 학습을 진행했을 때 생성된 6자유도 동영상의 품질이 증가한다.

각 NeRF 모델에서는 학습 후 가중치 (모델) 를 저장하는 함수가 구현되어 있다. 본 표준에서는 사용하는 모델에 따라 해당 함수를 호출할 수 있는 모듈을 포함한다. 프레임워크에서는 저장된 모델의 가중치를 다음 프레임을 위한 새로운 데이터셋 (전처리 된 상태)에서 이어서 학습하도록 처리한다. 이를 통해 결과적으로 이전 프레임의 모델과, 그것을 통해 생성된 이후 프레임의 모델을 동시에 저장 가능하다.



(그림 5-2) 프레임간 가중치 공유 기법 개념도

5.3 가중치 동결 기법을 적용한 학습 시간 단축

본 표준에서는 NeRF 모델에서 미세 조정을 수행할 때 전반부 계층을 동결하고 후반부 계층에 대해서만 학습을 하는 기법을 사용한다. 전반부 계층의 가중치를 동결함으로써 얻는 효과는, 손실 (loss) 계산 후 역전파 (back-propagation) 수행 시 가중치에 대한 손실의 변화율 (gradient) 계산 횟수가 감소하여 학습 시간의 단축이다.

<표 5-1>은 가중치 동결을 위한 기준 모델 기준의 구현 사항 내 구문 의미론을 나타낸다. 통상의 모델에서 채택된 pytorch 라이브러리에 범용적으로 모듈 단위의 적용이 가능하다. 가중치 동결 계층의 가중치를 나타내는 텐서 (tensor)에 대한 미분값 추적 옵션을 해제함으로써 해당 계층에 대한 업데이트를 제한할 수 있다. 모델에서 업데이트를 활성화 할 후반부 계층을 제외한 모든 계층을 동결하는 과정을 거친다. 결과적으로 가중치 공유 및 가중치 동결 기법을 통해 프레임 단위 개별 모델 학습 시 시간 대비 품질을 향상시킬 수 있다.

<표 5-1> 가중치 동결시 모델 훈련 설정 사항에서 사용되는 정보에 대한 구문 의미론

구문	의미론
requires_grad	해당 매개변수의 변화율을 추적할지 여부를 의미함. True/False로 표현됨. True 지정 시 해당 매개변수의 변화율을 계산함
model.parameters()	모델에 존재하는 전체 가중치를 반환하는 파이썬 제너레이터 (generator) 자료형. 반복문 조건식에서 사용됨
exclude_layers	학습에서 제외할 동결 계층 (layers to be freezed). 정수 인덱스의 리스트로 표현됨

다음 <표 5-2>은 인공신경망 가중치 동결 사항의 구현을 위한 기준 코드를 나타낸다. 해당 사항은 통상의 모델에서 모듈 단위로 확장하여 적용이 가능하다.

<표 5-2> 인공신경망 내 사용자 지정 계층에 대한 가중치 동결 기준 코드

기준 코드	설명

for i , layer in enumerate(layers):
if i in exclude_layers:
backup[j++] = deepcopy(layer)
	변화율 추정 활성화 계층의 가중치를 저장

for param in model.parameters(): param.requires_grad = False	가중치 동결 작업 실시

for i , layer in enumerate(layers):
if i in exclude_layers:
layer = backup[j++]
	활성화 계층의 매개변수를 원 상태로 복원

부 록 Ⅰ

(본 부록은 표준을 보충하기 위한 내용으로 표준의 일부는 아님)

필요성

Ⅰ.1 본 표준의 필요성

6자유도 가상현실 영상은 머리장착형 영상장치를 착용한 사용자의 머리를 돌리는 행위 뿐 아니라 움직임까지 반영하여 렌더링할 수 있는 동영상을 의미한다. 다수 카메라로 동시에 촬영된 영상을 통해 이를 구현할 수 있는데, 촬영에 사용되는 카메라 수에는 현실적인 제한이 있으므로 중간 시점의 영상을 합성하는 작업을 필요로 한다. 제한된 카메라 개수의 다시점 비디오로부터 시점과 시점 사이에 끊임 없는 영상을 출력하기 위해 다양한 중간 시점 합성 기법이 사용되고 있다. 예를 들어 reference view synthesizer (RVS)와 view weighting synthesizer (VWS) 기술이 있는데, 6자유도 가상현실 영상 압축 표준인 MIV 에서는 렌더링 과정에서 이를 선택적으로 사용할 수 있다. 하지만 이와 같은 보간 기반의 중간 시점 합성 기법은 사용자 시점에서 부자연스럽고 시점 간 거리가 멀면 합성 품질이 급격히 저하되므로 몰입형 가상 현실 영상 처리에는 3차원 재구성 기반의 접근법이 요구된다. 자세히 설명하자면, 사전에 제공된 깊이 정보를 사용하여 두 이미지 사이의 중간 이미지를 합성하는 것이 아닌, 공간 자체를 복원하여 이로부터 렌더링을 하는 기법을 요구하는 것이다.

정적 3차원 재구성을 실시하는 인공신경망 모델을 사용할 때, 매 프레임별 개별적으로 학습을 진행하는 것은 인접 프레임 간 유사성을 활용하지 않는 것이기 때문에 비효율적이다. 또한 딥러닝을 통한 3차원 재구성의 특성상 매번 가중치를 무작위로 초기화 한 상태에서 해당 프레임 데이터에 맞추어 학습을 진행한다면 연속적으로 동영상을 재생했을 때 주관적 품질 평가에서 일관성 있는 결과가 나오지 않을 가능성이 크다. 따라서 본 표준에서는 프레임 간 미세 조정을 통해 매개변수를 공유함으로써 데이터의 일관성 문제를 해결하고 학습에 소요되는 시간 절감을 위한 기술을 정의한다.

부 록 II-1

(본 부록은 표준을 보충하기 위한 내용으로 표준의 일부는 아님)

지식재산권 확약서 정보

II-1.1 지식재산권 확약서(1)

- 해당 사항 없음

부 록 II-2

(본 부록은 표준을 보충하기 위한 내용으로 표준의 일부는 아님)

시험인증 관련 사항

해당 사항 없음.

부 록 II-3

(본 부록은 표준을 보충하기 위한 내용으로 표준의 일부는 아님)

본 표준의 연계(family) 표준

해당 사항 없음.

부 록 II-4

(본 부록은 표준을 보충하기 위한 내용으로 표준의 일부는 아님)

참고 문헌

해당 사항 없음.

부 록 II-5

(본 부록은 표준을 보충하기 위한 내용으로 표준의 일부는 아님)

영문표준 해설서

해당 사항 없음.

부 록 II-6

(본 부록은 표준을 보충하기 위한 내용으로 표준의 일부는 아님)

표준의 이력

판수	채택일	표준번호	내용	담당 위원회
제1판	2023.11.20	제정 XDFK.01.0044/R0	-	운영위원회