다중 이미지를 이용한 광원 추정*

조창호[0], 하인우, 윤성의
한국과학기술원 전산학부
timegate@kaist.ac.kr, iw.ha@samsung.com sungeui@kaist.edu

Light estimation using multiple images

Changho Jo[0], Inwoo Ha, Sung-eui Yoon
School of computing, KAIST

요약

광원 추정(Light estimation) 문제는 가상 현실 어플리케이션에서의 현실적인 표현을 위한 주요한 문제이며, 제한된 시야로부터 가상의 물체를 렌더링하기 위해 반드시 필요한 과정이다. 본 논문에서는 기존까지의 단일 이미지 기반으로 접근해왔던 광원 추정 문제를, 다중 이미지를 활용하여 더 많은 정보로부터 풀어내는 것을 제안한다.

1. 서론

최근 들어, 많은 어플리케이션들은 가상의 물체들을 현실로 가져와 새로운 경험들을 제공하고 있다. 하지만 여전히 가상 물체와 실제 물체 사이의 격차를 제거하기 위한 많은 문제들이 남아있으며, 역 렌더링(Inverse rendering) 문제[1]에 그 기반을 둔다. 즉, 가상의 물체를 현실 환경에 놓기 위해서는 주어진 장면으로부터 그 환경의 기하, 반사 성질, 그리고 빛의 3가지 요소를 분리해서 찾아내야 하는 어려움이 있게 된다. 더욱이, 모바일 어플리케이션이라면 오브젝트를 둘러싼 모든 방향에서 오는 빛을 전체 파노라마의 6%밖에 되지 않는 모바일 기기의 화면에서 예측해야만 한다.

그러나 실제로는 여러 장의 관련된 이미지들을 활용할 수 있음에도 불구하고, [2]와 [3]과 같은 기존 연구들은 모두 단일 이미지 기반이었다. 따라서 우리는 이 다중 이미지들을 활용해 더 좋은 결과를 내고자 하였고, 이를 위해 매터포트3D 데이터셋을 가지고 다중 이미지들을 잘 활용할 수 있는 ConvLSTM 인코더-디코더 구조의 네트워크를 활용하여 문제를 풀었다.

2. 광원 추정이란 무엇인가

2.1. 광원 추정 문제와 기존 해결 방법

광원 추정 문제란 주어진 장면으로부터 HDR 파노라마와 같은, 가상의 물체를 렌더링하기에 충분한 결과를 만드는 과정이다. [2]는 처음으로 기하, 물질 속성, 빛 등 어떠한 것에도 큰 가정을 두지 않고 광원을 추정해냈으며, 가상 물체가 놓일 특정 위치를 중심으로 일정 크기로 잘라낸 이미지를 인풋으로 하고, 그들이 직접 만든 광원 분류기로 찾아낸 광원 마스크와 위에서 정한 위치로 구형 워핑을 한 파노라마를 아웃풋으로 하여 하나의 심층망으로 구현하였다. 그러나 이 연구는 세부적인 광원들까지 맞추지는 못한다는 한계를 가진다.

다음으로 [3]는, 어려운 역 렌더링 문제에 기반한 광원 추정 문제를 하나의 블랙박스 네트워크로 풀기보다, 서브 태스크들로 나누어 풀 때 더 좋은 결과를 낸다는 것을 보여주었다. 광원 추정 문제를 기하 추정 문제, 기하를 이용한 특정 위치로의 워핑, 파노라마 장면 완성, 완성된 장면의 HDR로의 전환, 이 4가지로 분리해서 세부적인 광원들까지 더 잘 찾아내는 결과를 보여주었다. 하지만 다중 이미지를 활용할 수 있는 경우에도 단일 이미지 기반으로만 광원을 추정한다는 한계를 가진다.

3. 다중 이미지를 이용한 광원 추정

3.1. 매터포트3D 데이터셋

매터포트3D 데이터셋은 다양한 시각에서 촬영한 194,400개의 HDR RGB-D 이미지들을 제공한다. 이 이미지들은 10,800개의 파노라마로 구성되어있고 90개 빌딩의 실내 모습을 보여주며, 데이터셋은 또한 실내 환경에 관련된 특징들을 학습하는 데 사용될 수 있도록 세그멘테이션 정보와 메쉬 정보도 가지고 있다.
더 자세하게는 하나의 파노라마를 모두 덮을 수 있는 18장의 이미지를 제공하는데, 우리는 그림1과 같이 18장의 이미지 중 광원의 직접적인 정보를 찾기 힘든 수직적으로 가운데에 있는 6장의 이미지들만을 활용하여 문제를 풀었다. 본 논문의 실험에서는 LDR 파노라마만을 사용하였다.

3.2. ConvLSTM

ConvLSTM은 비디오 예측 모델에서의 유명한 네트워크 블록들 중 하나이다. 기존의 Fully-connected LSTM 네트워크보다 모델의 복잡성을 줄이면서 연속적인 이미지들의 공간적인 정보를 더 잘 저장하며, 계속해서 실험적으로도 좋은 결과를 보여주었기에 계속해서 사용돼 왔다. 우리도 연속적인 여러 이미지들의 시공간적 정보

---

* 포스터 발표논문
* 본 연구는 과학기술정보통신부/정보통신기획평가원의 지원(IITP-2015-0-00199) 및 과학기술정보통신부/한국연구재단-차세대정보·컴퓨팅기술개발사업의지원(No. NRF-2017M3C4A7066317)을 받아 수행된 연구임.

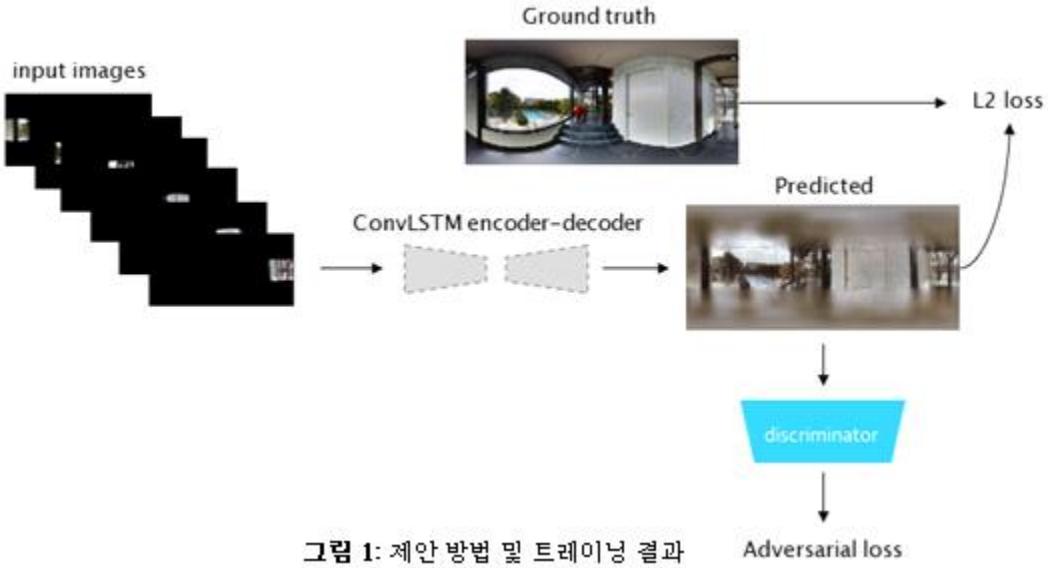

그림 1: 제안 방법 및 트레이닝 결과

를 완전히 활용하기 위해 ConvLSTM을 사용하였다.

3.3 제안 방법

우리는 [4]에서 사용한 ConvLSTM 인코더-디코더 구조를 RGB 이미지의 3채널을 받아들일 수 있도록 변형하여 사용하였다. 트레이닝 과정에서는 매터포트3D의 각각의 파노라마를 네트워크의 아웃풋으로, 앞서 언급한 6장의 이미지를 네트워크의 인풋으로 하였고, 테스트 과정에서는 보여주지 않았던 위치에서의 6장의 이미지만을 주고 전체 파노라마를 출력하도록 하였다. 여기에서의 인풋으로 들어가는 6장의 이미지들은 각각 파노라마의 일부분과 공간적으로 같을 수 있도록, 카메라가 바라보는 방향에 맞게 파노라마처럼 구형 왜곡을 적용해주고, 공간적으로 대응되는 곳에 위치시켰다.

3.3. 실험 결과

트레이닝은 9,810개의 파노라마로 20 epoch 동안 진행하였으며, 손실 함수는 [3]과 같이 L2 loss와 adversarial loss를 사용하였다. c개의 채널과 k개의 필터를 가진 Convolution-Relu를 C(c, k)로 표기할 때, adversarial loss를 위한 discriminator는 C(3, 16)-C(16, 32)-C(32, 64)-C(64, 128)-C(128, 256)-C(256, 512)와 같이 구성하였다.

트레이닝 결과에서는 인풋에서 제공된 수직적으로의 가운데 부분뿐만 아니라 파노라마의 위와 아래 부분도 괜찮은 결과를 내었지만, 테스트 결과에서는 위와 아래 부분으로 중간 부분의 이미지에서의 컨벌루젼이 충분히 되지 않는 결과를 얻었다.

그림 2: 테스트 결과

4. 결론

이 논문은 다중 이미지를 활용해 광원 추정 문제를 더 정확히 풀어낼 수 있음을 제안하였다. 실제 어플리케이션에는 하나의 이미지로부터 광원을 추정해야 하는 기존 연구들과는 달리 다중 이미지를 활용해 광원을 추정하기 위한 더 많은 정보들을 얻을 수 있기 때문이다. 비록 기존 비디오 예측 모델에서 주로 사용하던 ConvLSTM 블록을 변형하여 사용한 이 실험은 좋은 테스트 결과를 내지 못했지만, 이러한 광원 추정 문제에 맞는 연속적인 이미지들을 더 잘 활용할 수 있는 네트워크를 도입하고, 매터포트3D 데이터셋에서 제공하는 공간적인 정보, 시맨틱한 정보까지 모두 잘 이용한다면 더 좋은 결과를 낼 수 있을 것이다.

참고문헌

[1] R. Ramamoorthi and P. Hanrahan. A signal-processing framework for inverse rendering. In Proceedings of the 28th annual conference on Computer graphics and interactive techniques, pages 117–128. ACM, 2001.

[2] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. 2017. Learning to predict indoor illumination from a single image. ACM Trans. Graph. 36, 6, Article 176 (November 2017), 14 pages.

[3] S. Song and T. Funkhouser. Neural illumination: Lighting prediction for indoor environments. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6918–6926, 2019

[4] WANG, Rui; PIZER, Stephen M.; FRAHM, Jan-Michael. Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019. p. 5555-5564.