단안비디오로부터 광폭 베이스라인을 갖는 라이트필드 합성기법

백형선, 박인규
인하대학교 정보통신공학과
baekhys8801@gmail.com, pik@inha.ac.kr

Wide-baseline LightField Synthesis from monocular video

Hyungsun Baek, In Kyu Park
Department of Information and Communication Engineering, Inha University

요 약

본 논문에서는 단안비디오 입력으로부터 각 SAI(sub-aperture image)간의 넓은 기준선을 갖는 라이트필드 합성기법을 제안한다. 기존의 라이트필드 영상은 취득의 어려움에 의해 규모가 작고 특정 물체위주로 구성되어 있어 컴퓨터 비전 및 그래픽스 분야의 최신 딥러닝 기법들을 라이트필드 분야에 적용하기 어렵다는 문제를 갖고 있다. 이러한 문제점들을 해결하기 위해 사실적 렌더링 기반의 가상환경상에서 실제환경과 유사함을 갖는 데이터를 취득하였다. 생성한 데이터셋을 이용하여 기존의 새로운 시점을 생성하는 기법 중 하나인 다중 평면 영상(Multi Plane Image) 기반 합성기법을 통해 라이트필드 영상을 합성한다. 제안하는 네트워크는 단안비디오의 연속된 두개의 프레임으로부터 MPI 추정하는 네트워크와 입력영상의 깊이 정보를 추정하는 네트워크로 구성되어 있다.

1. 서론

라이트필드 영상은 여러 방향의 빛의 정보를 갖고 있다는 특성에 의해 일반적으로 사용되는 2D 카메라 영상에서는 거의 불가능한 촬영 후 영상 재 초점, 깊이 정보 추정, 시점 변환이 단일 라이트필드 영상에서 가능하다는 장점을 갖고 있다.

하지만 라이트필드는 플렌옵틱 카메라 또는 카메라배열을 통해 획득가능하기에 일반 사용자들이 쉽게 취득하기가 어렵다는 한계점이 존재한다. 이러한 문제를 해결하기 위해 특정 장면당 여러 장의 영상 혹은 하나의 영상으로부터 딥러닝 네트워크를 통해 라이트필드를 합성하는 연구들[1, 2]이 소개되어왔다. 해당 기법들은 플렌옵틱 카메라로 취득한 데이터를 사용하므로 결과로 생성된 라이트필드의 SAI 들의 기준선들이 짧게 생성이 된다는 한계를 갖고 있다.

본 논문에서는 가상환경상의 특성을 활용하여 간단히 SAI 간의 기준선이 넓은 라이트필드 데이터들을 취득하고, 입력 단안 비디오로부터 [4]에 의해 제안된 MPI 장면 표현방식과 깊이 추정 네트워크를 결합한 광시야각을 갖는 라이트필드 합성기법을 제안한다.

2. 가상환경 데이터셋 생성

컴퓨터 비전분야 및 그래픽스 분야에서의 뛰어난 성능을 보여주고 있는 딥러닝 기반 방식들을 사용하기 위해서는 대량의 영상데이터를 필요로 하는 경우가 많다. 하지만 라이트필드 분야의 경우에는 데이터 취득의 어려움으로 인해 데이터셋이 특정 물체 위주 또는 소량으로 구성이 되는 경우가 대부분이다.

본 논문에서는 이러한 문제점을 해결하기 위해 가상환경을 활용한 라이트필드 비디오 데이터셋 생성하였다. 실제 환경과 가상 환경의 차이를 최소화 하기위해 사실적 렌더링 기반의 비디오 게임인 Grand Theft Auto V(GTA-V)를 사용하여 실제와 유사한 데이터를 취득한다. 데이터 취득과정은 다음과 같이 요약할 수 있다. GTA-V 비디오 게임으로부터 영상 및 카메라 파라미터를 얻기 위해 Script Hook V library[3]를 사용하였으며, 해당 라이브러리를 통해 카메라 배열을 생성하고 각 카메라 별로 렌더링을 수행하여 영상을 취득하였다. 렌더링 기반인 게임의 특성상 신호 처리 지연에 의해 발생하는 카메라간 취득시간 불일치 문제는 정적인 물체만을 렌더링하고 동적인 물체는 제거함으로써 해결하였다.

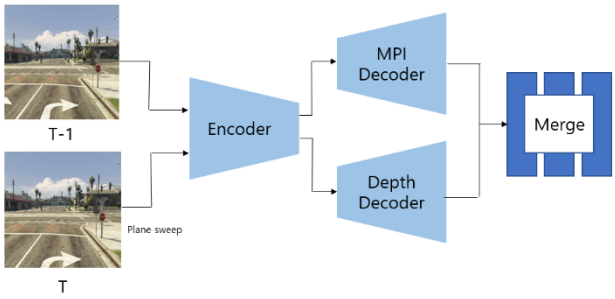

그림 1. 제안하는 라이트필드 합성 네트워크

3. 제안하는 네트워크

본 논문에서 제안하는 입력 단안 비디오로부터 광시야각을 갖는 라이트필드 합성 네트워크를 그림 1 에 나타내었다. 전체 네트워크는 입력으로 들어온 연속된 프레임들로부터 MPI 를 추정하는 네트워크, 깊이 정보를 추정하는 네트워크, 두가지 네트워크들의 결과를 결합하는 네트워크로 총 3 개의 네트워크로 구성하였다. 각각의 네트워크는 Encoder-Decoder 구조를 통해 학습을 진행하였으며, MPI 를 추정하는 네트워크는 [4]에서 사용한 프레임워크를 기반으로 구현하여 사용하였다. MPI 과 Depth 를 추정하는 네트워크는 공유된 Encoder 를 가지는 구조로 각각의 목적에 부합하는 Decoder 를 학습하였다. 최종적으로 MPI 를 이용하여 합성한 영상과 Depth 기반 워핑으로 생성한 영상을 입력으로 하여 최종 결과를 추정한다.

세부적인 구현사항으로 입력영상의 해상도는 224x224 를 사용하였으며 9x9 의 각해상도를 갖는 라이트필드를 합성한다. 또한 라이트필드의 기준선에 따른 학습과정을 용이하게 하고 결과영상의 품질향상을 위하여 기준선이 짧은 라이트필드에서 점차 넓어지는 라이트필드 데이터셋을 이용하여 순차적으로 학습을 진행한다.

4. 실험 결과

본 논문에서 제안하는 네트워크의 실험 결과를 그림 2 에 나타내었다. 평가에 사용된 데이터는 학습에 사용한 데이터처럼 가상환경에서 취득한 데이터를 사용하였으며, 학습에는 사용되지 않은 영상을 나타낸다. 그림 2 에서 가상환경의 4 개의 장면에 대한 결과를 나타내며 각각의 영상들은 입력영상을 보여주고 있다. 영상의 우측과 하단의 작은 이미지는 수직, 수평방향의 EPI 를 의미한다. 결과적으로 제안하는 네트워크를 이용하여 단안비디오로부터 넓은 기준선을 갖는 라이트필드를 합성함을 보였다.

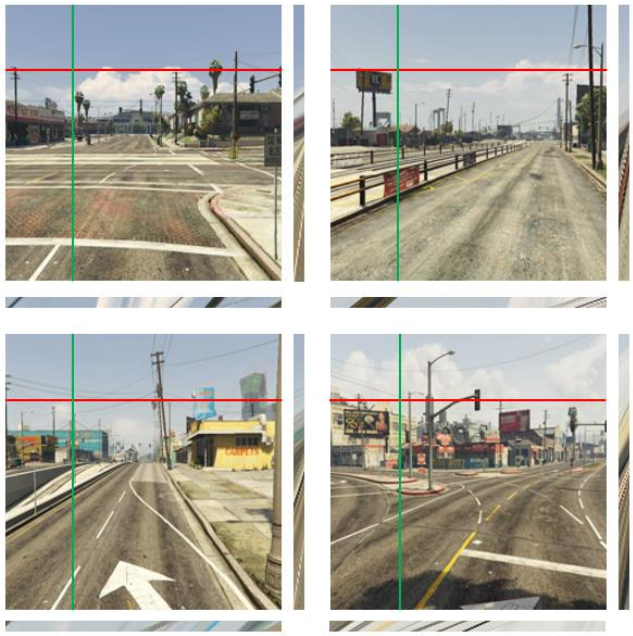

그림 2. 가상환경 데이터에 대한 실험 결과

5. 결론

본 논문에서 라이트필드 데이터 취득이 어렵다는 문제를 해결하기 위하여 가상환경 기반의 데이터셋 생성하였다. 취득한 단안 입력비디오로부터 다중 평면 영상과 깊이영상을 추정하고 이를 통해 라이트필드의 각 SAI 들을 합성할 수 있음을 보였다.

감사의 글

이 논문은 삼성전자 미래기술육성센터의 지원을 받아 수행된 연구임(SRFC-IT1702-54). 이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (2020-0-01389, 인공지능융합연구 센터지원(인하대학교)).

참고문헌

[1] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," ACM TOG, 35(6):193, 2016.

[2] P. P. Srinivasan, et al., "Learning to synthesize a 4D RGBD light field from a single image," In Proc. of IEEE ICCV, 2017.

[3] A. Blade. Script Hook V, Software available at http://www. devc. com/gtav/scripthookv, 2017.

[4] T. Zhou, R. Tucker, J. Flynn, G. Fyffe and N. Snavely, "Stereo magnification: Learning view synthesis using multiplane images." ACM SIGGRAPH, 2018.