단안 비디오로부터의 5D 라이트필드 비디오 합성 프레임워크

배규호, Andre Ivan, 박인규
인하대학교 정보통신공학과
kyuho1104@gmail.com, andreivan13@gmail.com, pik@inha.ac.kr

Deep Learning Framework for 5D Light Field Synthesis from Single Video

Kyuho Bae, Andre Ivan, and In Kyu Park
Department of Information and Communication Engineering, Inha University

요약

본 논문에서는 기존의 연구를 극복하여 단일 영상이 아닌 단안 비디오로부터 5D 라이트필드 영상을 합성하는 딥러닝 프레임워크를 제안한다. 현재 일반적으로 사용 가능한 Lytro Illum 카메라 등은 초당 3프레임의 비디오만을 취득할 수 있기 때문에 학습용 데이터로 사용하기에 어려움이 있다. 이러한 문제점을 해결하기 위해 본 논문에서는 가상 환경 데이터를 구성하며 이를 위해 UnrealCV를 활용하여 사실적 그래픽 렌더링에 의한 데이터를 취득하고 이를 학습에 사용한다. 제안하는 딥러닝 프레임워크는 두 개의 입력 단안 비디오에서 5×5의 각 SAI(sub-aperture image)를 갖는 라이트필드 비디오를 합성한다. 제안하는 네트워크는 luminance 영상으로 변환된 입력 영상으로부터 appearance flow를 추측하는 플로우 추측 네트워크(flow estimation network), appearance flow로부터 얻어진 두 개의 라이트필드 비디오 프레임 간의 optical flow를 추측하는 광학 플로우 추측 네트워크(optical flow estimation network)로 구성되어있다.

1. 서론

라이트필드 영상은 다양한 방향에서의 빛의 정보를 취득함으로써 한 장의 영상만으로 깊이 영상 추측(image depth estimation), 영상 재초점(refocusing), 시점 이동(view-point change) 등의 다양한 영상처리가 가능하다는 장점이 있다.

일반적으로 라이트필드 영상은 플렌옵틱(plenoptic)카메라 혹은 카메라 배열(camera arrays)을 사용하여 취득한다. 하지만 일반 사용자가 사용 가능하던 *Lytro*사의 카메라는 더 이상 지원이 되지 않고 유일하게 남아있는 플렌옵틱 카메라는 *Raytrix* [1]사의 카메라가 있지만 이는 주로 산업 현장에서 사용함을 목적으로 만들어지므로 일반 사용자가 사용하기엔 어려움이 있다. 이러한 한계점을 극복하기 위해 일반 영상으로부터 라이트필드 영상을 합성하는 다양한 기법이 소개되었다 [2, 3, 4]. 하지만 해당 기법들은 모두 비디오가 아닌 일반 영상으로부터 라이트필드 영상을 합성하는 기법으로 라이트필드 비디오 합성에는 어려움이 있다.

기존의 기법 중 [2]의 경우 한 장의 라이트필드 영상을 합성하기 위해 특정 시점의 입력 영상 4장을 요구하며 이는 일반 사용자 입장에서 취득하기 어렵다는 단점이 있다. [3]의 경우 한 장의 라이트필드 영상을 합성하기 위해 한 장의 입력 영상만을 사용하여 깊이 영상 기반 렌더링 기법을 사용해 라이트필드 영상을 합성한다. 하지만 해당 기법의 경우 합성된 깊이 영상의 품질에 크게 의존하며, 깊이 영상을 사용하여 라이트필드 영상을 합성하기 때문에 non-lambertian 효과를 복원하는데 어려움을 겪고 폐색(occulusion) 영역을 복원하는데 어려움이 있다. [4]의 경우 [3]의 한계점을 극복하기 위해 깊이 영상 대신 appearance flow를 사용하여 라이트필드 영상을 합성하였다. 하지만 해당 기법의 경우 비디오가 아닌 정적인 물체에 대한 단일 라이트필드 영상을 합성하였기 때문에 단안 비디오로부터 라이트필드 비디오를 합성하는데는 적합하지 않다.

본 논문에서는 이상의 한계점을 극복하여 단일 영상이 아닌 단안 비디오로부터 5D 라이트필드 비디오(x, y, u, v, t)를 합성하는 딥러닝 프레임워크를 제안한다. 본 논문의 구성은 다음과 같다. 먼저, 딥러닝 네트워크를 학습하기 위해 사용한 가상 환경 데이터를 제시한다. 둘째, 입력 단안 비디오 프레임에서 라이트필드 비디오를 합성하는 end-to-end 딥러닝 네트워크를 제시한다. 셋째, 제안하는 기법의 결과를 확인한다. 마지막으로, 본 논문에 대한 결론을 맺는다.

2. 가상 환경 데이터 구성

컴퓨터 비전 분야에서 다양하게 활용되고 있는 딥러닝 기반 영상처리 기법들은 각각의 기법에서 제안하는 네트워크를 학습하기 위해 다량의 영상 데이터를 요구한다. 하지만 특정한 목적에 맞는 영상 데이터를 취득하기엔 어려움이 많으며 특히 다량의 데이터가 요구되는 경우에는 더 많은 어려움이 따른다. 본 논문에서는 이러한 한계점을 극복하기 위해 여러 연구에서 그 효용성을 보인 사실적 그래픽 렌더링에 의한 가상 환경 데이터를 활용한다. 가상 환경 데이터를 구성하기 위해 *Unreal Engine*을 기반으로 하는 *UnrealCV* [5]를 활용한다. 해당 기법을 활용하여 9×9의 angulr domain(u, v)을 갖는 라이트필드 데이터를 취득하였으며 학습에 사용된 가상 환경은 실재 도시와 유사한

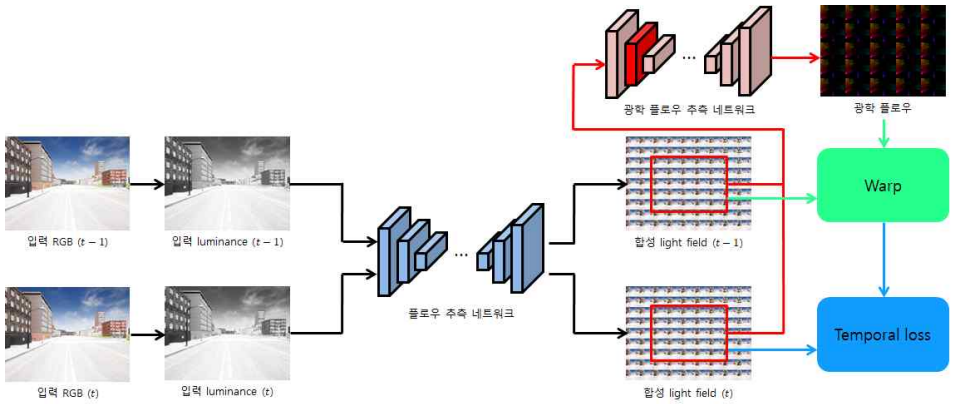

그림 1. 제안하는 5D 라이트필드 비디오 합성 프레임워크.

외관을 가진 2개의 환경을 사용하였다. 2개의 환경은 [5]에서 제공하는 가상 환경과 본 논문에서 직접 구성한 환경으로 구성되어있다. 각 환경에서 구성된 도로를 따라 카메라를 이동하며 총 1,818장의 9×9의 SAI로 구성된 라이트필드비디오 데이터를 취득하였다.

3. 딥러닝 네트워크

본 논문에서 제안하는 입력 단안 비디오로부터 5D 라이트필드 비디오를 합성하는 딥러닝 프레임워크를 **그림 1**에 나타내었다. 전체 프레임워크는 입력 비디오로부터 9×9의 각 SAI에 대응되는 appearance flow를 추측하는 플로우 추측 네트워크, 그리고 추측된 플로우를 각 입력 단안 비디오에 워핑하여 얻은 라이트필드 비디오간의 광학 플로우를 추측하는 광학 플로우 추측 네트워크로 구성되어 있다. 각 단계들을 아래와 같이 수식으로 표현할 수 있다.

$$L_a(t) = A(L(x,y,0,0,t), L_{af}(x,y,u,v,t)) \quad (1)$$

$$F_{(t-1)\to t} = O(L_a(t-1), L_a(t)), \quad (2)$$

수식 (1)에서 $L_{af}(x,y,u,v)$은 플로우 추측 네트워크로부터 얻은 9×9 의 각 SAI에 대응되는 appearance flow를 나타내고 $L(x,y,0,0,t)$는 입력 luminance 영상을 나타낸다. A는 bilinear sampler module [6]을 나타내며 이를 통해 시간 t에서 워핑된 라이트필드 영상인 $L_a(t)$를 얻을 수 있다. O는 시간 $t-1$과 t간의 광학 플로우 $F_{(t-1)\to t}$를 추측하는 광학 플로우 추측 네트워크를 나타낸다.

Appearance flow를 추측하는 네트워크인 플로우 추측 네트워크는 [4]에서 사용한 프레임워크를 바탕으로 네트워크 구조를 해당 기법에서 사용한 네트워크 구조 대신 encoder-decoder 구조를 사용하여 구현하였다.

두 개의 9×9 라이트필드 영상으로부터 9×9의 각 SAI에 대응되는 광학 플로우를 추측하는 과정은 막대한 양의 GPU 메모리가 필요하다. 따라서 본 논문에서는 하드웨어적 제한으로 인하여 9×9의 각 SAI중 가운데 5×5영역만을 선택하여 해당 영역의 SAI에 대응되는 광학 플로우만을 추측한다. 광학 플로우를 추측하는 네트워크인 광학 플로우 추측 네트워크는 플로우 추측 네트워크와 유사한 encoder-decoder 구조를 바탕으로 구성되어 있으며 (1)을 통해 얻은 시간 $t-1$과 시간 t에서의 라이트필드 프레임인 $L_a(t-1)$과 $L_a(t)$를 입력으로 시간 $t-1$과 시간 t 사이의 광학 플로우를 추측한다. 광학 플로우 추측 네트워크는 두 프레임간의 상관관계를 추측하기 위해 각 프레임의 특징을 추출한 뒤 이를 결합하고 3D convolution을 사용하여 두 특징간의 상관관계를 사용하여 광학 플로우를 추측하게 된다.

추측된 광학 플로우와 $L_a(t-1)$중 5×5영역을 워핑하여 얻은 영상과 $L_a(t)$간의 시간적 일관성을 부여하기 위해 다음과 같은 손실함수를 정의한다.

$$\ell_{temporal} = \sum_t w \left| L_a(t) - L_{warp}^{t-1\to t} \right| \quad (3)$$

$$w = \exp(-\lambda |L(x,y,0,0,t-1) - L(x,y,0,0,t)|), \quad (3a)$$

수식 (3a)의 w는 [7]에서 영감을 받아 두 입력 영상간의 차이에 따라 시간적 일관성을 다르게 부여하기 위한 가중함수이다. 즉, 시간 $t-1$에서의 프레임과 시간 t에서의 프레임간의 차이가 클 경우 시간적 일관성을 적게 부여하고 두 프레임간의 차이가 적은 경우 상대적으로 더 큰 시간적 일관성을 부여한다. 수식 (3)의 $L_{warp}^{t-1\to t}$는 (2)에서 얻은 광학 플로우를 사용하여 시간 $t-1$의 프레임을 시간 t로 워핑하여 얻은 라이트필드 프레임을 나타낸다.

4. 실험 결과

본 논문에서 제안한 기법의 실험 결과를 **그림 2**와 **그림 3**에 각각 나타내었다. **그림 2**는 가상 환경 데이터 중 학습에 사용되지 않은 데이터에 대하여 라이트필드 비디오를 합성한 결과를 나타낸다. **그림 3**은 가상 환경 데이터가 아닌 다양한 컴퓨터 비전 분야에서 활용되고 있는 실재 환경에 대한 비디오 데이터인 KITTI 데이터셋에 대한 실험 결과를 나타낸다. **그림 3**에서 나타내듯이 가상 환경 데이터를 사용해 학습한 네트워크를 사용하여 실재 환경의 데이터에 대해서도 라이트필드 비디오를 합성해 낼 수 있음을 보였다.

5. 결론

본 논문에서 제안한 기법은 가상 환경 데이터를 활용해 딥러닝 네트워크를 학습하고 appearance flow를 사용해 라이트필드를 합성하며 광학 플로우를 사용해 시간적 일관성을 부여한다. 실험 결과를 통해 가상 환경 데이터 뿐만 아니라 실재 환경의 데이터에 대해서도 라이트필드 비디오를 합성해 낼 수 있음을 보였다.

그림 2. 가상 환경 데이터에 대한 실험 결과.

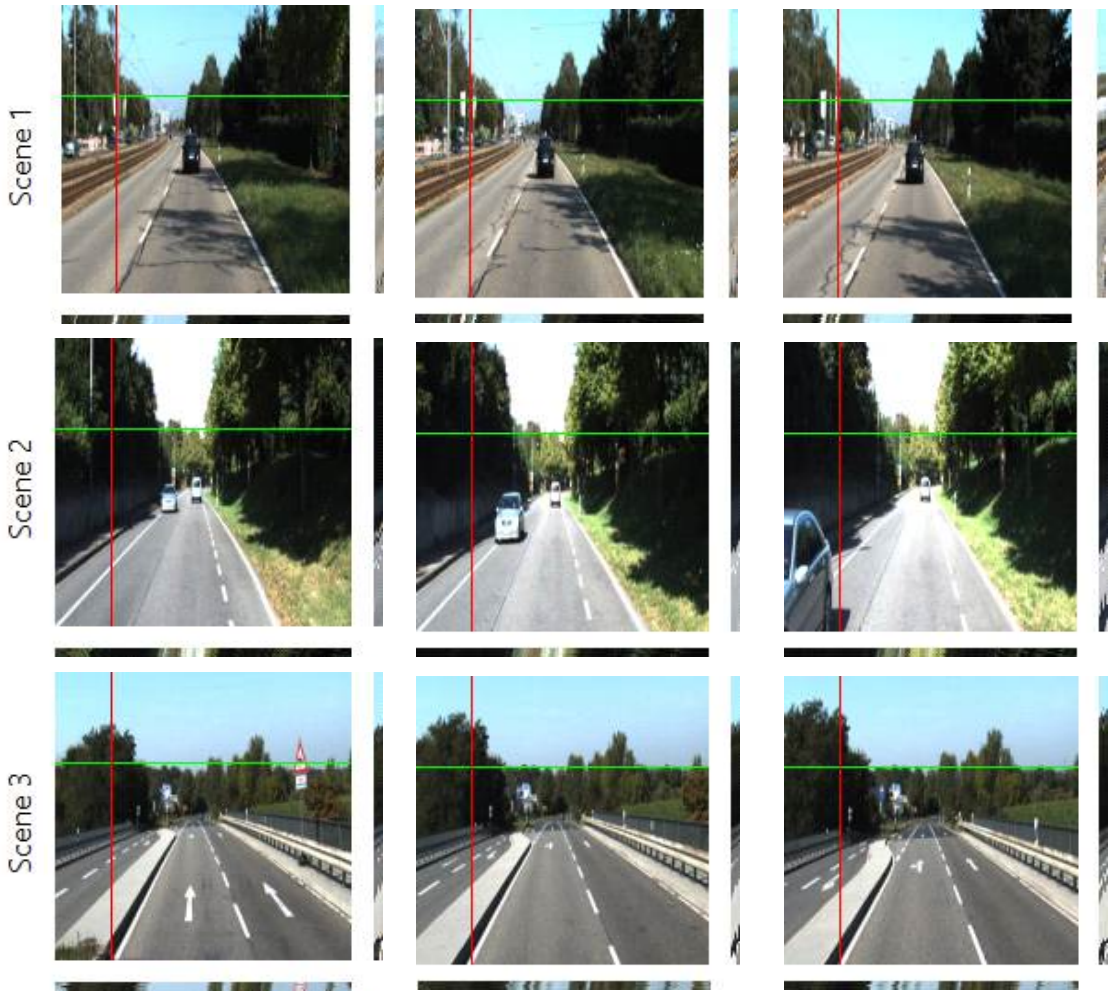

그림 3. 실재 환경 데이터에 대한 실험 결과.

6. 감사의 글

본 논문은 삼성전자 미래기술육성센터의 지원을 받아 수행한 연구결과임 (과제번호 SRFC-IT1702-06)

참고문헌(Reference)

[1] Rayrtix 3D light field camera, https://raytrix.de/products/.

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

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

[4] I. Andre, Williem, and I. K. Park, "Synthesizing a 4D spatio-angular consistent light field from a single image," *arXiv preprint arXiv:1903.12364*, 2019.

[5] W. Qiu, and Y. Alan, "UnrealCV: Connecting computer vision to unreal engine," In *Proc. of ECCV*, pages 909-916, 2017.

[6] M. Jaderberg, K. Simonyan, A. Zisserman, et al., "Spatial transformer networks," In *Proc. of NIPS*, pages 2017-2025, 2015.

[7] N. Bonnel, et al., "Blind video temporal consistency," *ACM TOG*, 34(6):196, 2015.