프레쳇 거리 손실함수를 이용한 RGBD 파노라마 영상 생성

김수지, 박인규
인하대학교 정보통신공학과
soojie@inha.edu, pik@inha.ac.kr

RGBD Panoramic Image Generation Using Frechet Distance Loss Function

Soo Jie Kim, In Kyu Park
Department of Information and Communication Engineering, Inha University

요 약

RGBD 영상은 다양한 3 차원 비전 연구에서 유용하게 사용되며 고품질 RGBD 영상을 취득하기 위한 많은 연구들이 수행되었다. 기존의 영상 생성 연구들은 주로 좁은 FoV(Field of View) 영상을 사용하여서 전체 장면 중 상당 부분이 소실된 영상에 대한 정보를 생성한다. 본 논문에서는 기존의 좁은 FoV 영상으로부터 360 도 전방향 RGBD 영상을 생성하는 기법을 제안한다. 오버랩 되지 않는 4 장의 소수 영상으로부터 전체 파노라마 영상에 대해서 상대적인 FoV 를 추정하고, 360 도 RGBD 영상을 동시에 생성하는 적대적 생성 신경망 기반의 영상 생성 네트워크이다. 360 도 영상의 특징을 반영하도록 설계하여서 개선된 성능을 보인다.

1. 서론

RGBD 영상은 3 차원 객체 인식, 장면 재구성 등 다양한 3 차원 비전 연구에 활용된다. 이에 따른 고품질 RGBD 영상 생성 연구가 진행되어 왔으나 대부분 좁은 FoV 영상을 활용하며, 도로영상과 같은 실외 데이터셋에 기반한다 [1,2]. 또한 360 도 영상을 생성하는 연구의 경우에 파노라마 영상 특성을 고려한 네트워크를 설계하여서 생성된 파노라마 영상의 왜곡을 줄이고자 하는 연구가 수행되었다. 하지만 대부분 전체 파노라마 영상을 입력으로 사용하거나, 센서 기반의 360 도 RGBD 영상을 생성하는 연구들이 수행되었다.

본 논문에서는 4 장의 좁은 FoV 영상에 대해서 360 도 RGBD 영상을 동시에 생성하는 적대적 생성 신경망 기반의 네트워크를 제안하여서 전체 파노라마에 대해서 상대적인 FoV 를 추정하고 이로부터 얻어진 부분 Equirectangular 영상에 대해서 전체 파노라마 영상을 생성하도록 하였다. 파노라마 생성 단계에서 두 모달리티의 특징을 공유하도록 네트워크를 구성하며, 360 도 영상의 특징을 반영한 목적함수를 적용하여서 정량적, 정성적으로 우수한 성능을 보인다.

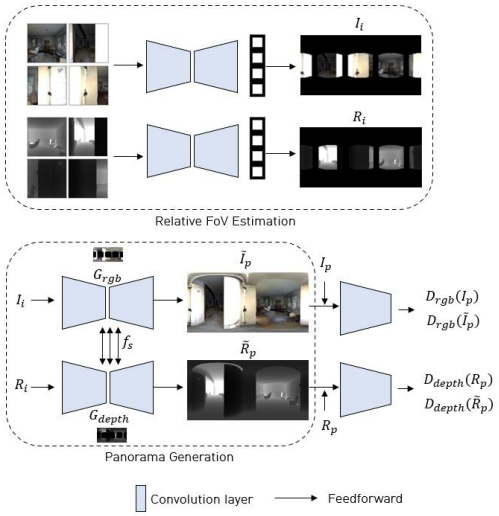

그림 1. 제안하는 방법의 전체적인 흐름도

2. 파노라마 생성

먼저 4 개의 영상으로부터 전체 파노라마에 대해서 상대적인 FoV 를 추정하고[3] Equirectangular 영상으로 변환하여서 파노라마 생성 단계의 입력으로 사용한다.

파노라마 생성 네트워크는 U-Net 기반의 적대적 생성 신경망 네트워크이다. 훈련시에 적대적 손실함수로 LSGAN 을 사용하며 생성된 영상과 참값 영상 사이의 픽셀 손실 함수는 L1 손실함수를 사용한다. 또한 사실적인 영상 생성을 위해서 지각 손실 목적 함수로 사전 훈련된 VGG 네트워크를 사용한다.

360 도 영상 특징을 반영한 특징을 추출하기 위해서 제안하는 모델을 훈련하고 사전 훈련된 모델을 이용하여서 RGBD 특징이 공유된 모듈에서 잠재공간의 마지막 레이어의 특징 f_{latent} 를 추출한다. 그런 다음 [4]에 따라서 longitudinal invariant 특징 f'_{latent} 와 latitudinal equivariant 특징 f''_{latent} 을 추출하고 Frechet Distance [5]를 측정하여서 손실함수로 사용하였다. Frechet Distance 손실함수는 훈련 중 0 번째 iteration 에서 측정된 값을 나누어서 정규화하여서 사용하였으며 500 개의 데이터셋 분포에 대해서 2000 Iteration 마다 한 번씩 값을 계산하여서 손실함수로 반영하였다.

각 RGBD 네트워크에서 생성된 영상에 대하여 입력 영상의 참값 영역을 제외한 나머지 부분에 해당하는 이진 마스크를 적용한 영상의 특징을 공유하며 가장 큰 클래스에 해당하는 마스크를 생성된 영상에 적용하고 각 RGBD 네트워크에서 공유하도록 하였다. 각 RGBD 네트워크의 레이어는 픽셀 합계를 수행하며 마지막 레이어 f_s는 각 네트워크의 마지막 블록에 채널 연결을 수행하여서 각 디코더에 전달된다. 제안하는 네트워크의 전체적인 프레임워크는 **그림 1** 에서 도식화하였다.

3. 실험 결과

네트워크를 훈련 환경은 ADAM 옵티마이저를 사용하며, Learning rate 는 $\alpha = 0.0002$, $\beta_1 = 0.5$, $\beta_2 = 0.99$, 배치 사이즈는 2 로 설정하고, NVIDIA RTX A6000 GPU 를 사용하여 훈련하였다.

총 21,600 개의 실내 데이터셋 Matterport3D[6], Stanford3D[7], SunCG[8]에서 깊이 영상을 기준으로 손상된 영역이 일정 값 이상에 해당하는 훈련에 부적절한 데이터셋 3,446 개를 제거하는 과정을 거치며, 총 18,154 개의 데이터셋에 대하여 훈련 데이터셋(80%), 테스트 데이터셋(20%)으로 나누어서 적용한다. Matterport3D[6], Stanford3D[7], SunCG[8] 데이터셋에 대해서 큐브맵 형식으로 변환하여 4 개의 면을 훈련에 사용할 수 있도록 데이터셋을 구축하였으며, 세 개의 데이터셋을 각각 따로 훈련하였다. 정량평가에서 생성 모델의 유사도를 위한 PSNR, SSIM 을 측정하여서 Frechet distance 손실함수를 사용하지 않은 네트워크와 비교하였다.

표 1 은 생성된 RGBD 영상의 PSNR 과 SSIM 평균을 나타낸다. 제안하는 Frechet distance 손실함수를 적용하지 않은 네트워크의 결과를 비교하였으며, 제안하는 네트워크에서 우수한 결과를 확인할 수 있다. **그림 2** 는 제안하는 모델과 제안하는 모델에 대한 정성적인 결과이며, FoV 60 도에 대한 입력 영상의 결과이다.

표 1. 생성된 파노라마 영상의 PSNR, SSIM 평균

Method		Dataset	PSNR	SSIM
w/o Frechet distance	RGB	Matterport3D	17.6667	0.6705
		Stanford3D	18.2139	0.6748
		SunCG	**21.7096**	**0.8072**
	Depth	Matterport3D	26.1095	0.9294
		Stanford3D	25.4359	0.9185
		SunCG	**28.4385**	0.9418
w/ Frechet distance	RGB	Matterport3D	**18.1325**	**0.6850**
		Stanford3D	**18.4782**	**0.6859**
		SunCG	21.0893	0.8051
	Depth	Matterport3D	**26.1831**	**0.9301**
		Stanford3D	**25.6019**	**0.9187**
		SunCG	28.3960	**0.9435**

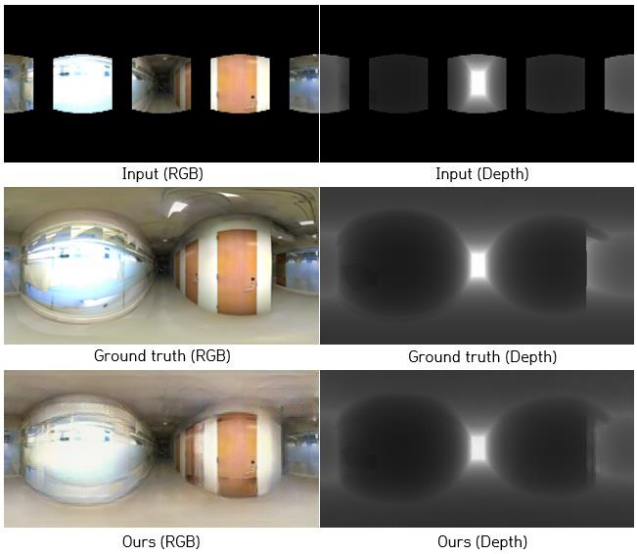

그림 2. 파노라마 생성 네트워크에 대한 정성적 평가 결과

5. 결론

본 논문에서는 소수의 영상으로부터 RGBD 영상을 동시에

생성하는 적대적 생성 신경망 기반 네트워크를 제안하였다. 두 모달리티의 특징을 공유한 생성 모델과 360 도 영상의 특징이 반영된 Frechet Distance 손실함수를 적용하여서 정량적, 정성적 결과를 통해서 개선된 성능을 보인다. 복잡한 장면을 재구성하는 분야와 3D 실내 환경 복원에서 사용될 수 있다.

감사의 글

이 논문은 2022 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행한 연구임(2020-0-01389, 인공지능융합연구센터지원(인하대학교)). 이 논문은 삼성전자 미래기술육성센터의 지원을 받아 수행한 연구임(과제번호 SRFC-IT1702-54).

참고문헌

[1] Y. Wang, W. L. Chao, D.Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, “Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving,” Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Aug. 2019.

[2] F. E. Wang, Y. H. Yeh, M. Sun, W. C. Chiu, and Y. H. Tsai, “LED2-Net: Monocular 360 Layout Estimation via Differentiable Depth Rendering,” Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Apr. 2021.

[3] J. S. Sumantri and I. K. Park, “360 Panorama synthesis from a sparse set of images with unknown field of view,” Proc. IEEE Trans. on Computational Imaging, vol. 6, pp. 1179-1193, Jul. 2020.

[4] C. O. W.J. Cho and K. Yoon, “Rgbd panorama synthesis using normal field-of-view cameras and mobile depth sensors in arbitrary configuration,” Proc. The 33rd Workshop on Image Processing and Image Understanding, p1-11, 2021.

[5] D. C. Dowson and B. V. Landau, “The frechet distance between ´ multivariate normal distributions,” in Journal of Multivariate Analysis, 12:450-455, 1982.

[6] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, “Matterport3D: Learning from RGB-D data in indoor environments,” Proc. International Conference on 3D Vision, Sep. 2017.

[7] I. Armeni, S. Sax, A. R. Zamir, and S. Savarese, “Joint 2D-3Dsemantic data for indoor scene understanding,” arXiv preprint arXiv:1702.01105, Apr. 2017.

[8] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, “Semantic scene completion from a single depth image,” Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun. 2017.