내용 기반의 정렬을 통한 HDR 동영상 생성 방법

정혜수, 조남익
서울대학교
reneeish@snu.ac.kr, nicho@snu.ac.kr

HDR Video Reconstruction via Content-based Alignment Network

Haesoo Chung, Nam Ik Cho
Seoul National University

요 약

최근 인터넷을 통한 동영상 제공 서비스가 확대됨에 따라 높은 품질의 온라인 컨텐츠에 대한 수요가 급증하고 있다. 그런데 넓은 동적 범위를 표현할 수 있는 High Dynamic Range (HDR) 컨텐츠의 공급은 수요를 따라가지 못하고 있는 실정이다. 본 논문에서는 밝기가 다른 프레임들로 구성된 Low Dynamic Range (LDR) 동영상을 이용해 HDR 영상을 생성하는 방법을 제안한다. 우선, 프레임들 간에 움직임이 존재하기 때문에 정렬 과정을 통해 이웃 프레임들을 중심 프레임에 맞추어 정렬한다. 이때 내용 (content) 기반으로 정렬을 해 정확도를 높이고, 원래 크기의 입력을 그대로 이용하는 모듈을 함께 사용하여 세부 정보도 잘 살려준다. 그리고 나서 잘 정렬된 다중 프레임들을 합쳐서 하나의 HDR 프레임을 생성한다. 실험을 통해 기존 방법들에 비해 우수한 성능을 보임을 확인하였다.

1. 서론

최근 들어 온라인 동영상 제공 서비스의 확대와 함께 고품질의 동영상 컨텐츠에 대한 수요가 증가하고 있다. 디스플레이 기술의 발전으로 높은 화질과 동적 범위 (dynamic range)를 표현할 수 있는 기기는 많아졌지만, 그에 상응하는 High Dynamic Range (HDR) 컨텐츠들은 현저히 부족한 상황이다. HDR 동영상 생산을 위해 직접 특수 장비를 이용하여 취득하는 방법도 있지만, 촬영 장비의 가격이 비싸고 구하기 어렵기 때문에 Low Dynamic Range (LDR) 동영상을 HDR 동영상으로 변환하는 기술에 대한 연구가 필요하다.

정지 영상에 대한 HDR 이미징 방법들이 활발히 연구되어 온 것에 비하여 [1, 2, 3, 4, 5, 6], HDR 동영상 생성 방법에 대한 연구는 상대적으로 더디게 진행되었다. HDR 동영상을 만들어 내기 위해서는 기본적으로 노출 값을 변화시켜 가면서 촬영한 LDR 동영상을 이용해 다양한 영역에 대한 충분한 정보를 취득할 수 있게 한다. 그런데 입력으로 여러 장의 프레임을 사용할 때, 프레임 간의 밝기가 다를 뿐만 아니라 물체나 카메라의 움직임도 존재하기 때문에 이를 고려할 필요가 있다. 기존 방법들은 프레임 간의 움직임을 보정하기 위하여 optical flow나 patch 기반의 방법들을 이용하였는데, 보정이 정확하지 않고 시간이 많이 소요된다는 단점이 있었다 [7, 8, 9].

본 논문에서는 밝기가 다른 여러 장의 프레임들을 효과적으로 이용하기 위하여, 이웃 프레임들을 중심 프레임의 움직임에 맞추어 정렬한 뒤 정렬된 피쳐 (feature)들을 하나로 합쳐 HDR 프레임을 만들어 낸다. 이때 그림 1에서 보이는 것과 같이 HDR 프레임 한 장을 만들기 위해 중심 프레임 한 장과 이웃 프레임 네 장을 입력으로 사용하고, 정렬 네트워크를 이용해 이웃 프레임의 움직임을 정렬해 준다. 정렬 네트워크는 크게 두 가지 모듈로 구성되는데, 첫번째 모듈은 다운샘플링된 입력 프레임들을 이용해 내용 기반의 정렬을 수행하는 Low Resolution (LR) 모듈이고, 두번째 모듈은 두 장의 입력 프레임을 그대로 이용하여 LR 모듈에서 미처 생성하지 못한 세부 정보들을 만들어 내는 High Resolution (HR) 모듈이다. LR 모듈에서는 RGB 색상 값이 아닌 내용 (content)에 기반한 key와 query를 추출하여 key-query 매칭을 진행함으로써 신뢰도 높은 정렬을 할 수 있다.

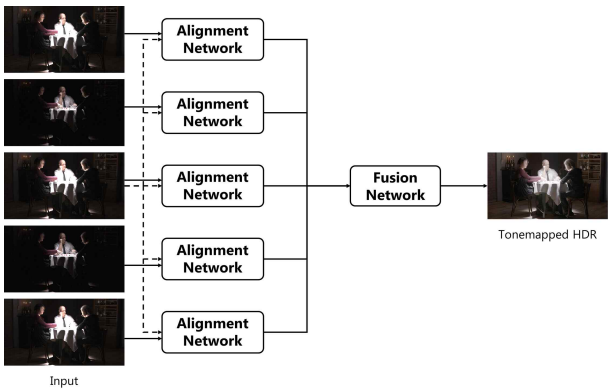


그림 1. 전체 구조

본 논문의 구성은 다음과 같다. 2절에서 내용 기반의 정렬 방법을 이용한 HDR 동영상 생성 방법에 대해 설명하고, 3 절에서는 제안 방법을 이용한 실험 결과를 제시한다. 마지막으로 4절에서 결론을 지으며 마무리한다.

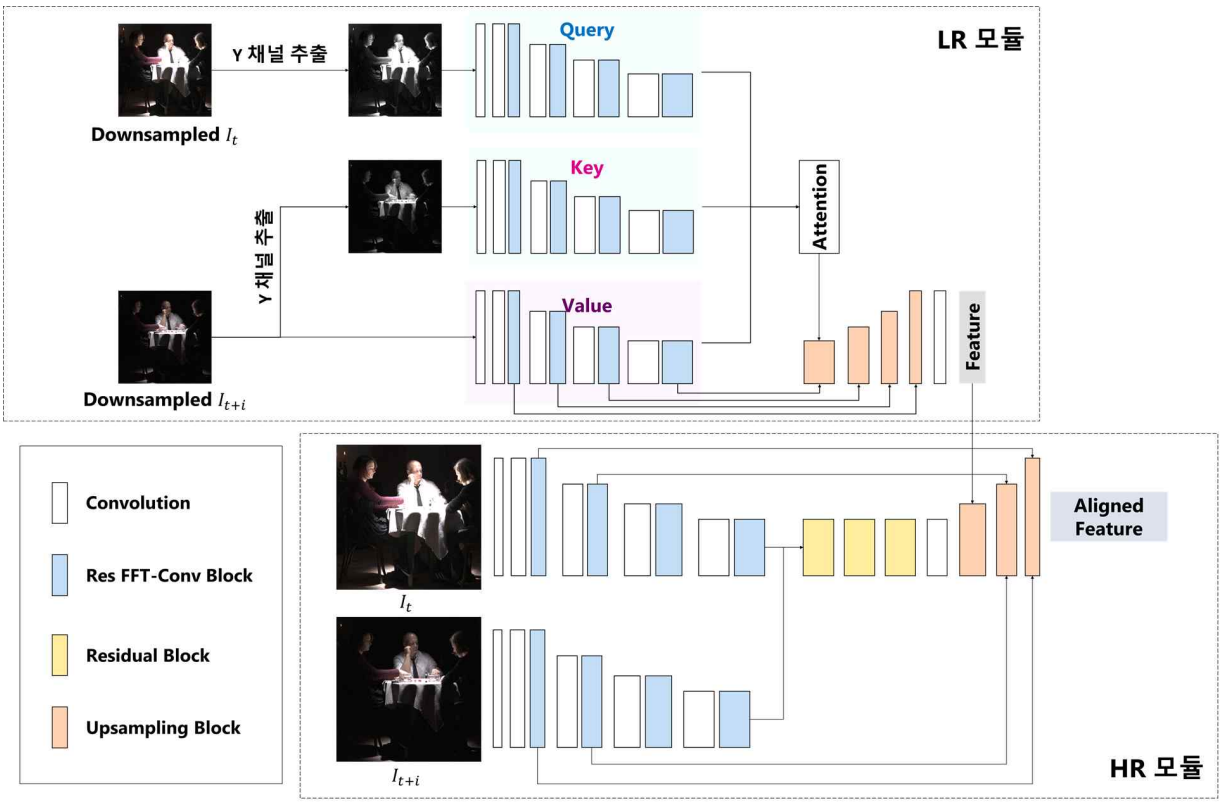


그림 2. 정렬 네트워크의 구조

2. 제안 방법

제안하는 방법은 연속된 다섯 장의 LDR 프레임 $\{I_{t-2}, I_{t-1}, I_t, I_{t+1}, I_{t+2}\}$을 이용하여 HDR 프레임 H_t을 생성하는 것을 목표로 한다. 다섯 장의 입력 프레임들은 교대로 다른 노출 값을 가진다. 그림 1과 같이 각각의 입력 프레임들은 중심 프레임과 함께 정렬 네트워크에 들어간다. 정렬 네트워크는 내용에 기반하여 이웃 프레임을 중심 프레임의 움직임에 맞추어 정렬한다. 합성 네트워크는 잘 정렬된 피쳐들의 정보를 합쳐 최종 HDR 프레임을 생성한다.

정렬 네트워크는 그림 2와 같이 두 가지 모듈로 구성된다. LR 모듈은 다운샘플링된 중심 프레임 I_t와 이웃 프레임 I_{t+i} 간의 내용 기반 매칭을 통한 명시적인 정렬을 담당하고, HR 모듈은 인코더-디코더 구조를 이용해 부족한 고주파 세부 정보를 살려준다.

LR 모듈부터 자세히 살펴보면, 먼저 LR 모듈의 입력으로 원래의 입력 영상을 $1/4$ 크기로 다운샘플링하여 넣어주어서 메모리 사용량을 크게 줄여 준다. 모듈 내에서 정렬을 할 때 중심 프레임과 이웃 프레임 간에 유사한 부분을 찾아야 하는데, 단순한 색상 정보를 이용한 매칭이 아닌 내용 기반의 매칭을 하기 위해 RGB 영상을 YCbCr로 변환한 뒤 색상 정보는 제외한 Y 채널 정보만 이용하여 key, query를 추출한다. Key와 query는 각 인코더의 마지막 피쳐에서 임베딩 (embedding)을 거친 피쳐이며, 두 인코더는 가중치를 공유한다. 내용 기반의 매칭을 통해 얻은 attention 값에 따라 이웃 프레임 로부터 얻은 value 피쳐를 적절히 가져오면 피쳐 수준에서 정렬이 수행된다. 이렇게 얻어진 피쳐는 디코더를 통과하여 정렬 네트워크의 입력 영상 크기의 피쳐로 복원된다.

HR 모듈은 입력 영상의 크기를 줄이지 않고 그대로 이용해서, LR 모듈에서 만들어 내기 어려운 세부 정보들을 만들어낸다. HR 모듈은 중심 프레임과 이웃 프레임으로부터 각각 딥 피쳐를 추출한 뒤 residual block [10]을 통해 적절히 합쳐 준다. 피쳐 추출을 위한 인코더는 간단한 합성곱 레이어와 주파수 도메인에서 연산을 수행하는 Res FFT-Conv Residual Block [11]으로 이루어져 있다. 이후 디코딩 과정에서 Upsampling block에서 LR 모듈의 출력을 받아 합쳐 주면서 복원을 진행한다. 해당 block에서는 HR 모듈의 피쳐와 LR 모듈의 피쳐의 차원을 맞춰준 뒤, 두 피쳐를 더해주고 residual block을 통과시켜준다.

정렬 과정이 끝나면 각 정렬 네트워크의 출력 피쳐는 합성 네트워크를 통과하여 하나의 HDR 영상으로 만들어 진다. 합성 네트워크는 합성곱 레이어와 Res FFT-Conv Residual Block으로 구성된다. 정렬된 피쳐들을 채널 축을 따라 이어 붙여준 뒤, 합성곱 레이어와 다섯 개의 Res FFT-Conv Residual Block을 통과시킨다. 마지막으로 3 채널로 줄여주고

시그모이드 함수를 통과시키면 최종 HDR 프레임이 만들어진다.

네트워크 학습을 위한 손실 함수로는 L_1 손실 함수와 VGG 손실 함수, 주파수 손실 함수를 사용하는데, 톤맵핑 (tonemapping) 한 HDR 영상에 대해 계산한다. 주파수 손실 함수는 각 HDR 프레임에 fast Fourier transform (FFT)을 적용한 뒤 주파수 영역에서 L_1 거리를 측정하는 함수이다. 각 손실 함수는 $1:0.001:0.1$의 가중치로 사용한다. 톤맵핑 함수로는 다음과 같이 정의되는 $\mu-law$를 사용한다.

$$T(H)=\frac{\log(1+\mu H)}{\log(1+\mu)} \quad (1)$$

μ값은 5000으로 설정한다.

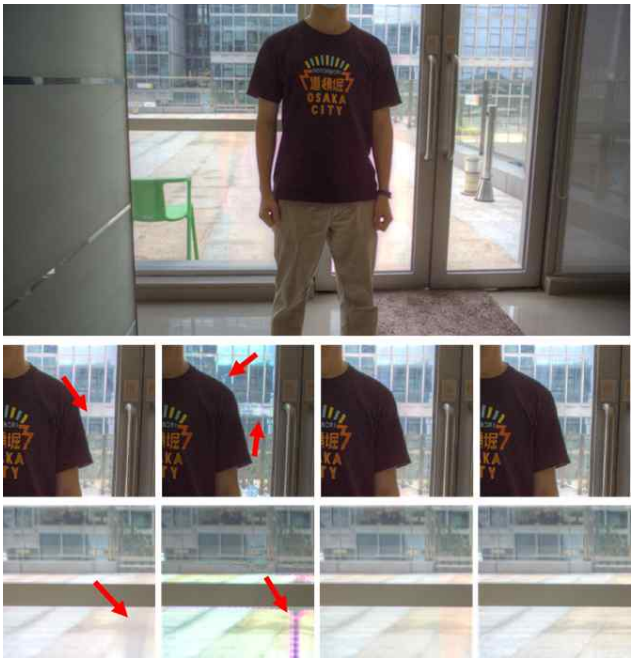
3. 실험 결과 및 분석

네트워크 학습 데이터로는 Vimeo-90K dataset [12]을 사용하였고, 테스트 데이터로는 실제 데이터인 DeepHDRVideo 데이터셋 [9]을 이용하였다. 학습 데이터와 테스트 데이터 모두 그림 3과 같이 노출 값이 번갈아 가면서 바뀌게 구성되어 있다. 비교를 위해 다중 노출 영상을 이용한 HDR 이미징 방법 [6]과 HDR 동영상 생성 방법 [9]를 같은 데이터로 학습하였다. 먼저 정량적 결과 비교를 위해 HDR 영상을 톤맵핑한 후 측정한 PSNR (PSNR-T)과 SSIM (SSIM-T)을 평가 지표로 사용하였다. 표 1에서 볼 수 있듯이 제안하는 방법이 다른 방법들에 비해 높은 성능을 보인다. 또한 결과 영상을 그림 4에 나타내었다. 상단에 위치한 그림이 제안 방법을 이용하여 얻은 결과 영상이고, 하단에 위치한 그림들이 다른 방법들과 비교한 결과를 확대하여 나타낸 것이다. 제안하는 방법이 프레임 간의 움직임으로 인한 왜곡을 최소화하고 깨끗한 HDR 프레임을 만들어 내는 것을 확인할 수 있다.

그림 3. 데이터 예시

표 1. 정량적 결과

	PSNR-T	SSIM-T
Yan [6]	43.71	0.9682
Chen [9]	35.95	0.9640
제안 방법	**44.67**	**0.9709**



Yan [6] Chen [9] Ours GT

그림 4. 정성적 결과 비교

4. 결론

본 논문에서는 내용 기반의 정렬을 수행하는 모듈과 세부 정보를 잘 만들어 내는 모듈로 구성된 정렬 네트워크와 간단한 합성 네트워크를 이용한 HDR 동영상 생성 구조를 제안하였다. 실험을 통해 제안하는 방법이 세부 정보들이 잘 표현되는 HDR 동영상을 만들어 낼 수 있음을 확인하였다.

감사의 글

이 논문은 2022년도 BK21 FOUR 정보기술 미래인재 교육연구단 및 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (2021R1A2C2007220).

참고 문헌

[1] Thorsten Grosch et al. Fast and robust high dynamic range image generation with camera and object movement. Vision, Modeling and Visualization, RWTH Aachen, 277284, 2006.

[2] Shanmuganathan Raman and Subhasis Chaudhuri. Reconstruction of high contrast images for dynamic

scenes. The Visual Computer, 27(12):1099-1114, 2011.

[3] Yong Seok Heo, Kyoung Mu Lee, Sang Uk Lee, Youngsu Moon, and Joonhyuk Cha. Ghost-free high dynamic range imaging. In Asian Conference on Computer Vision, pages 486-500. Springer, 2010.

[4] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. ACM Trans. Graph., 36(4):144-1, 2017.

[5] Pradeep Sen, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B Goldman, and Eli Shechtman. Robust patch-based hdr reconstruction of dynamic scenes. ACM Trans. Graph., 31(6):203-1, 2012.

[6] Qingsen Yan, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Attention-guided network for ghost-free high dynamic range imaging. In CVPR, 2019.

[7] Nima Khademi Kalantari, Eli Shechtman, Connelly Barnes, Soheil Darabi, Dan B Goldman, and Pradeep Sen. Patch-based high dynamic range video. TOG, 2013.

[8] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep HDR video from sequences with alternating exposures. In Computer Graphics Forum, 2019.

[9] Chen, Guanying, et al. "HDR video reconstruction: A coarse-to-fine network and a real-world benchmark dataset." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016.

[11] X. Mao, Y. Liu, W. Shen, Q. Li, and Y. Wang, “Deep residual fourier transformation for single image deblurring,” arXiv preprint arXiv:2111.11745, 2021.

[12] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. “Video enhancement with task-oriented flow,” International Journal of Computer Vision, 2019.