몰입형 비디오 부호화를 위한 신경망 기반 아틀라스 후처리 필터링

임성균, 이건우, 김정우, 윤용욱, 김재곤
한국항공대학교
{sglim, yuyoon}@kau.kr, jgkim@kau.ac.kr

Neural Network-Based Post Filtering of Atlas for Immersive Video Coding

Sung-Gyun Lim, Kun-Woo Lee, Jeong-Woo Kim, Yong-Uk Yoon, and Jae-Gon Kim
Korea Aerospace University

요 약

MIV(MPEG Immersive Video) 표준은 제한된 3D 공간의 다양한 위치의 뷰(view)들을 효율적으로 압축하여 사용자에게 임의의 위치 및 방향에 대한 6 자유도(6DoF)의 몰입감을 제공한다. MIV 의 참조 소프트웨어인 TMIV(Test Model for Immersive Video)에서는 몰입감을 제공하기 위한 여러 시점의 입력 뷰들 간의 중복 영역을 제거하고 남은 영역들을 패치(patch)로 만들어 패킹(packing)한 아틀라스(atlas)를 생성하고 이를 압축 전송한다. 아틀라스 영상은 일반적인 영상 달리 많은 불연속성을 포함하고 있으며 이는 부호화 효율을 크게 저하시키다 본 논문에서는 아틀라스 영상의 부호화 손실을 줄이기 위한 신경망 기반의 후처리 필터링 기법을 제시한다. 제안기법은 기존의 TMIV 와 비교하여 아틀라스의 복원 화질 향상을 보여준다.

1. 서론

최근 사용자에게 높은 자유도의 시각적 경험을 제공하는 몰입형(immersive) 비디오가 주목받고 있다. 몰입형 비디오는 사용자의 움직임에 따른 시차를 지원하기 위해 여러 위치에서 획득한 입력 시점 뷰(view)들을 필요로 한다. MPEG 비디오 그룹(JTC 1/SC 29/WG 4)은 몰입형 비디오 부호화를 위한 MIV(MPEG Immersive Video) 표준의 버전 1 개발을 완료했으며[1]-[3], 버전 2 표준화를 진행중이다[4]. MIV 의 참조 소프트웨어 TMIV(Test Model for Immersive Video)에서는 다시점 입력 뷰들 간의 중복되는 영역을 제거하고 남은 영역들을 각각 패치(patch)로 생성하여 최대한 조밀하게 아틀라스에 패킹(packing)한 아틀라스(atlas) 영상을 압축 및 전송한다. 아틀라스 영상은 일반적인 영상과는 다르게 서로 다른 텍스처(texture)를 가진 패치들이 인접해 있기 때문에 공간적 상관성이 크게 떨어지고 패치 경계에 불연속성이 존재하여 부호화 효율이 좋지 않다는 문제가 있다. 본 논문에서는 TMIV 로 생성된 아틀라스 영상의 부호화 손실을 줄이기 위한 신경망(neural network) 기반의 후처리 필터링(post filtering) 기법을 제시한다.

2. TMIV

TMIV 는 몰입형 비디오를 구성하는 다수의 뷰간의 중복성을 줄이는 전처리를 통하여 압축할 화소률을 최대한 줄인다. 그림 1 은 TMIV 부호화기의 전체 구조이다. 그림 1 과 같이 입력된 서로 다른 시점의 비디오들을 몇 개의 그룹으로 나누고 각 그룹 내에서 기본 뷰와 추가 뷰으로 분류한다. 그룹별로 기본 뷰와의 중복성이 있는 추가 뷰의 영역들을 제거하는 프루닝(pruning) 과정이 수행되고 남은 영역들은 직사각형의 패치(patch)로 생성된다. 하나의 패치에는 전송되어야 할 잔여 화소들을 포함하는 유효(valid) 영역과 프루닝에 의해 제거된 부분인 무효(invalid) 영역이 모두 포함된다.

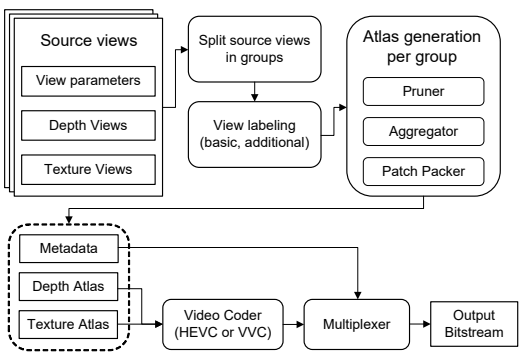

그림 1. TMIV 부호화기의 전체 구조도[5]

추가 뷰에서 생성된 패치들은 기본 뷰와 함께 소수의 아틀라스에 유효 영역이 겹치지 않도록 하되 최대한 조밀하게 패킹된다. 이를 위해 패치를 회전하여 패킹할 수도 있으며, 각

패치의 패킹 위치 및 회전 정보들은 메타데이터(matadata)로 아틀라스 부호화 비트스트림(bitstream)과 함께 전송되어 TMIV 복호화기에서 아틀라스에 패킹된 패치들이 추가 뷰들의 원래 위치로 복원될 수 있도록 한다.

3. 아틀라스 후처리 필터링

직사각형 패치의 유효 여부는 16 x 16 또는 32 x 32 크기의 블록 단위로 결정되며 패치내에는 화소 단위로 유효 영역과 비유효 영역이 모두 포함된다. 패치내의 유효 화소에는 입력 뷰의 해당 위치로부터 가져온 화소값으로 비유효 화소에는 중간값이나 0 값으로 채워진다. 따라서, 아틀라스 영상은 서로 다른 패치 사이 또는 유효 영역과 비유효 영역 사이의 많은 불연속 경계면을 포함하게 된다. 이로 인하여 아틀라스의 부호화 효율이 크게 떨어지게 된다 이 문제를 해결하기 위해서 그림 2 의 구조와 같은 구조로 원본 아틀라스와의 차이가 적도록 복원된 아틀라스 화질을 향상시키는 신경망 기반의 후처리 필터링을 제안한다.

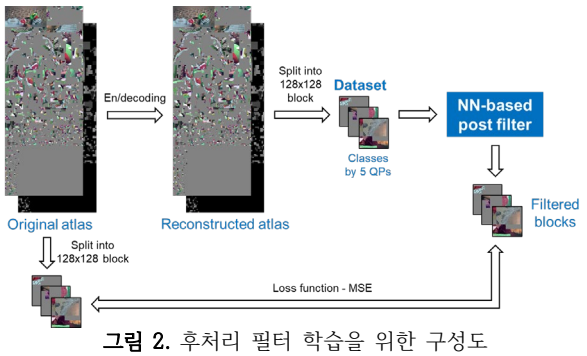

그림 2. 후처리 필터 학습을 위한 구성도

제안한 후처리 필터링 신경망 모델은 그림 3 과 같이 12 개의 합성곱(convolution) 은닉층으로 구성된 CNN(Convolutional Neural Network) 구조를 사용한다[6]. 원본 아틀라스 영상과 복원된 아틀라스 영상을 각각 128 x 128 크기로 나눈 블록쌍이 모델에 입력되며, 두 블록의의 MSE(Mean Square Error)가 최소가 되도록 모델을 학습한다. 입력 데이터로 사용하는 아틀라스의 부호화 오류 정도에 따라 학습되는 모델이 달라지기 때문에 부호화에 사용하는 5 개의 QP(Quantization Parameter)에 따라 각각 모델을 별도로 학습한다. 또한, 아틀라스에서 공간적적 상관성이 높은 기본 뷰가 패킹된 영역과 많은 불연속 경계면을 포함하고 있는 패치들이 패킹된 영역의 영상은 특성이 다르므로, 우선 패치 영역만 필터링 하기 위해서 패치 영역만 포함한 데이터들로 학습을 진행하였다.

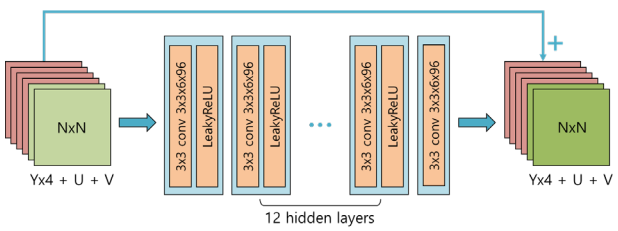

그림 3. 후처리 필터의 모델 구조[6]

4. 실험결과

MIV 공통 테스트 조건(Common Test Condition, CTC)[7]에 규정된 선택적 시퀀스(optional sequence)로 학습 데이터 셋을 구성하고 테스트에는 CTC 의 필수 시퀀스(mandatory sequence)를 사용했다. 학습을 위해서 아틀라스를 블록 단위로 분할하여 많은 양의 데이터셋을 구성했지만 사용된 시퀀스 수는 한정적이다. 따라서, 충분히 다양한 특성을 반영하여 학습하는 것이 어려울 수 있고, 과적합(overfitting) 문제가 발생하여 학습이 원활하게 수행되지 않을 수 있다. 이를 해결하기 위해서 학습 모델에 드롭아웃(drop-out) 기법을 적용하여 모델을 학습하였고, 드롭아웃 비율(drop-out rate)는 0.4 로 설정했다. 또한, 학습에 RMSProp 최적화기를 사용하고 학습률(learning rate)은 0.001, 배치(batch) 크기는 64 로 설정하고 50 에폭(epoch) 학습했다. 그림 4 는 제안하는 아틀라스 후처리 필터 모델의 학습 손실 곡선을 나타낸 것으로 에폭이 증가함에 따라 학습 손실과 검증 손실이 약 670 ~ 690 의 값으로 수렴하는 것을 보여준다.

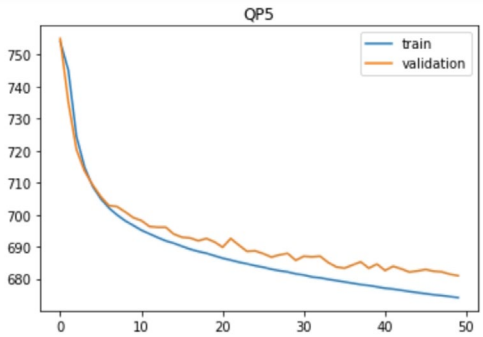

그림 4. 제안 후처리 필터 모델의 학습 손실 곡선

표 1 은 가장 가장 낮은 비트율 조건(QP5 in CTC)에서 부호화된 아틀라스에 제안하는 후처리 필터를 적용했을 때, PSNR 로 계산한 객관적 화질이 개선된 정도를 보여주는 표이다. TMIV11.0 으로 생성한 아틀라스를 사용했으며, 아틀라스 압축 및 복원에는 MIV CTC 에 따라 VVenC[8]와 VVdeC[9]를 사용하였다. 제안하는 후처리 필터를 적용함으로써 휘도성분 PSNR 에서 대하여 평균 1.77dB 의 화질 개선을 얻었다.

표 1. 제안 후처리 필터링의 아틀라스 화질 비교(QP5)

Sequences (resolution)	Without filtering (dB)	With filtering (dB)	Δ-PSNR (dB)
SA (4096x4096)	32.52	32.91	0.39
SB (2048x2048)	29.35	29.50	0.15
SO (1920x1080)	29.77	32.28	2.50
SJ (1920x1080)	36.40	39.67	3.27
SD (2048x1088)	30.96	31.16	0.21
SE (1920x1080)	28.14	30.38	2.23
SP (1920x1080)	35.86	39.39	3.53
SN (2048x2048)	33.77	34.03	0.27
SR (1920x1080)	29.61	33.03	3.41
Average	**31.82**	**33.59**	**1.77**

그림 5 는 제안하는 후처리 필터를 적용했을 때, 주관적으로 아틀라스의 화질이 개선된 것을 보여준다. 후처리 필터를 적용하지 않은 복원된 아틀라스는 패치 경계 부분에 부호화 오류로 인한 평활화(smoothing)를 확인할 수 있는데, 제안하는 후처리 필터를 적용함으로써 원본 영상과 비슷하게 경계가 선명해진 것을 볼 수 있다.

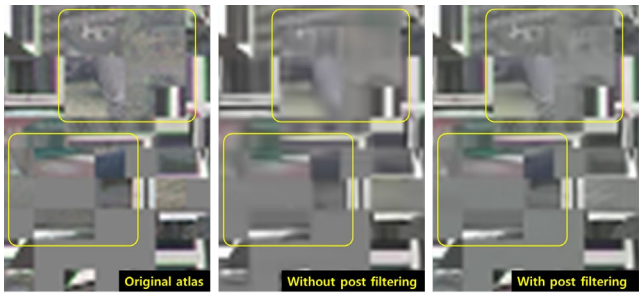

그림 5. 제안 후처리 필터를 적용한 아틀라스의 예
(왼쪽: 원본, 가운데: 필터링 미적용, 오른쪽: 필터링 적용)

또한, 제안 필터링 기법을 적용했을 때, 종단간 부호화 성능을 평가하기 위해서 입력 뷰 비디오와 동일한 시점으로 합성된 렌더링 비디오를 비교하여 WS-PSNR[10]과 IV-PSNR[11]로 화질을 측정했다. 표 3 은 BD-rate 측면에서 제안기법을 적용한 부호화 성능을 결과이다. 제안하는 후처리 필터를 적용함으로써 휘도성분 WS-PSNR 에서 대하여 평균 0.1%의 이득을 보여준다. 아틀라스의 기본 뷰 영역에도 후처리 필터링을 추가로 적용하면 보다 의미 있는 부호화 이득이 있을 것으로 예상된다.

표 3. 제안기법을 적용한 몰입형 비디오 부호화 성능(Anchor: TMIV11.0)

Sequences (resolution)	WS-PSNR (Y)	IV-PSNR
SA (4096x4096)	0.3%	0.5%
SB (2048x2048)	0.1%	0.0%
SO (1920x1080)	-0.2%	0.0%
SJ (1920x1080)	-0.3%	1.5%
SD (2048x1088)	-0.4%	0.8%
SE (1920x1080)	0.5%	2.5%
SP (1920x1080)	-0.1%	0.4%
SN (2048x2048)	-2.4%	1.7%
SR (1920x1080)	1.3%	0.9%
Average	**-0.1%**	**0.9%**

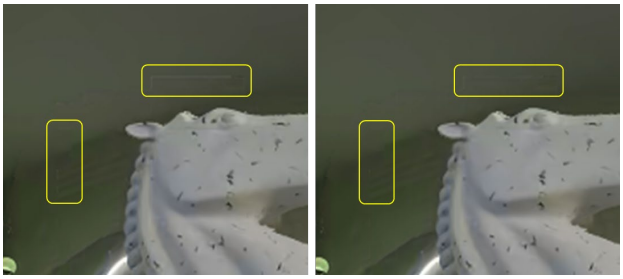

그림 6. 'SN(Chess)' 시퀀스에서 렌더링 된 시점 영상의 예
(좌: 필터링 미적용, 우: 필터링 적용)

그림 6 은 주관적 화질을 비교한 것으로 필터링을 적용하지 않은 아틀라스로 렌더링한 뷰에서 보이는 시각적 아티팩트(visual artifact)가 제안기법을 적용한 결과에서는 제거된 것을 확인할 수 있다. 이를 통해 제안기법은 주관적 화질 측면에서도 눈에 띄는 개선을 보이는 것을 알 수 있다.

5. 결론

본 논문에서는 몰입형 비디오 부호화 과정에서 아틀라스에 존재하는 많은 불연속성으로 인한 화질 저하를 개선하고, 이를 통해 렌더링 뷰 화질 향상을 위한 신경망 기반의 후처리 필터링 기법을 제시한다. 제안하는 후처리 필터링은 아틀라스 영상의 부호화 손실을 줄임으로써 아틀라스 복원 영상의 화질이 개선됨을 확인하였다. 제안기법을 적용한 종단간 부호화 성능에서는 다소 미미하지만 객관적인 부호화 이득과 주관적인 화질 개선을 얻을 수 있음을 확인하였다. 추후 아틀라스의 기본 뷰 영역에도 후처리 필터링을 추가로 적용하면 보다 의미 있는 부호화 이득이 있을 것으로 예상된다.

Acknowledgement

이 논문은 2022 년도 정부(과학기술정보통신부)의 재원으로정보통신기획평가원의 지원(No. 2017-0-00207, 이머시브 미디어 전문연구실)과 한국연구재단의 지원(No. 2020-R1A6A3A13073358)을 받아 수행된 연구임.

References

[1] "MPEG-I Use Cases for Omnidirectional 6DoF, Windowed 6DoF, and 6DoF," ISO/IEC JTC1/SC29/WG11, N16768, Apr. 2017.

[2] M. Wien, J. M. Boyce, T. Stockhammer, and W.-H. Peng, "Standardization Status of Immersive Video Coding," IEEE J. Emerg. Select. Topics Circuits Syst., vol. 9, no. 1, pp. 5-17, Mar. 2019.

[3] "Text of ISO/IEC FDIS 23090-12 MPEG Immersive Video," ISO/IEC JTC1/SC29/WG04, N00111, Jul. 2021.

[4] "Preliminary WD1 of ISO/IEC 23090-12 MPEG Immersive video Ed. 2," ISO/IEC JTC1/SC29/WG04, N00176, Jan. 2022.

[5] B. Salahieh, J. Jung, A. Dziembowski (Eds.), "Test Model 11 for Immersive Video," ISO/IEC JTC1/SC29/WG04, N00142, Oct. 2021.

[6] H. Wang, M. Karczewicz, J. Chen, and A. Kotra, "AHG11: Neural Network-based In-Loop Filter," ISO/IEC JTC1/SC29/WG05, JVET-T0079, Oct. 2020.

[7] J. Jung, B. Kroon, "Common Test Conditions for MPEG Immersive Video," ISO/IEC JTC1/SC29/WG04, N00169, Jan. 2022.

[8] VVenC software, [Online]. Available at: https://github.com/fraunhoferhhi/vvenc/tree/v0.2.0.0

[9] VVdeC software, [Online]. Available at: https://github.com/fraunhoferhhi/vvdec/tree/v1.0.1

[10] "WS-PSNR Software Manual," ISO/IEC JTC1/SC29/WG11, N18069, Oct. 2019.

[11] A. Dziembowski, "Software manual of IV-PSNR for Immersive Video," ISO/IEC JTC1/SC29/WG04, N00013, Oct. 2020.