CNN 을 이용한 전방위 비디오 합성 시점의 화질 개선 알고리즘

박현수, 강제원
이화여자대학교 엘텍공과대학 전자전기공학과
gustn824@ewha.ac.kr, jewonk@ewha.ac.kr

CNN-based Denoising Algorithm for Synthesized Views in 6 Degree-of-Freedom Videos

Hyeonsu Park, Je-Won Kang
Department of Electronic and Electrical Engineering, Ewha Womans University

요 약

본 논문은 최근 MPEG-I 에서 논의되고 있는 전방위 6 자유도 영상의 가상시점 합성의 기존 공개 소프트웨어의 문제점 해결방안을 제안한다. 참조시점을 사용하여 합성된 가상시점의 영상을 대상으로 묶음 조정 (bundle adjustment) 개념의 딥 러닝을 적용하여 영상 간 시공간적 품질 차이를 낮춘다. 실험에 따르면 중간시점 영상 합성 후 같은 시간적 특성을 같은 묶음을 MF-CNN (Multi-Frame Convolutional Neural Networks)에 적용함으로써 단순 VVS2.0 의 합성 결과 대비 평균 공간적으로 0.34dB, 시간적으로 0.81dB 의 성능 향상을 제공하였다.

1. 서론

최근 급속한 디지털 영상 처리 기술의 발전과 더불어, 사용자에게 보다 실감적인 경험을 제공하는 몰입형 미디어에 대한 수요가 높아졌다. 이에 따라 국제 표준화 단체인 MPEG (Moving Picture Expert Group)에서는 몰입형 미디어를 위한 표준을 제정하기 위해 2016 년 10 월 116 차 청두 회의에서 MPEG-I 라는 새로운 표준화를 시작하였다. 특히 시청자에게 높은 몰입감과 현장감을 제공하기 위해 머리 회전운동의 3 자유도 (degree of freedom)를 넘어 움직임 시차에 따른 자유도를 증가한 6 자유도 (6 degree-of-freedom)의 지원이 MPEG-I 의 최종 목표이다.

실제로 6 자유도를 지원하는 콘텐츠를 제공하기 위해서는 시점 변화에 따른 시점 영상을 모두 제공해야 한다. 하지만 모든 시점에서의 영상을 직접 취득하는 것은 사실상 불가능하다. 이를 극복하기 위해 제한된 입력 영상을 사용하여 원하는 위치에서의 영상을 합성하는 중간시점, 즉 가상시점에서의 영상 합성에 대한 연구가 진행되고 있다 [1]. MPEG 에서는 중간시점 합성에 대한 연구의 일환으로 VSRS (View Synthesis Reference Software)를 공개 소프트웨어를 개발해왔다. [1] 현재는 콘텐츠 별로 기존 VSRS 의 좁은 기준선과 낮은 주관적 성능 문제점을 개선한 RVS (Reference View Synthesizer)와 VVS (Versatile View Synthesizer)가 공개 소프트웨어로 개발되고 있다 [2-3].

해당 공개 소프트웨어들은 매 회의마다 기술 기고를 바탕으로 발전되고 있지만, 중간시점 합성의 결과들로 여전히 낮은 품질의 영상이 생성되는 문제가 있다. 이는 합성된 중간시점은 근본적으로 깊이 영상 기반의 투사 (projection) 방식으로 영상을 합성하기 때문이다. 특히 시간적 품질 차이는 주로 깊이 영상 (depth map)의 정확도에, 공간적 품질 차이는 주로 입력 영상의 위치 변이에 따른 가려짐 현상으로 기인한다.

본 논문에서는 위 문제점을 해결하여 합성된 중간시점 영상에 묶음 조정(bundle adjustment)개념의 딥러닝을 적용하여 영상 간 시공간적 품질 차이를 낮추는 방법을 제안한다.

2. 중간시점에서의 영상 합성

앞서 소개된 중간시점에서의 영상 합성의 공개 소프트웨어들은 입력 시점의 색상영상 (texture map)과 깊이영상 (depth map) 및 카메라 매개변수를 필요로 하며, 크게 3 단계의 영상합성 과정을 따른다. 첫 번째는 3 차원 워핑 (warping) 과정이다. 3 차원 워핑이란 깊이 영상과 카메라 매개변수를 이용하여 색상영상을 3 차원 공간상에 역투영 (back-projection) 시킨 후 이를 다시 목표 시점에 투영하는 투사 방식의 일종이다. 이 때, 시점 이동으로 인해 참조시점에서 존재하지 않았던 영역은 홀(hole)로 나타난다. 홀은 대게 두번째 과정인 영상 통합 과정을 통해 해결된다. 영상 통합 과정이란 참조시점으로부터 중간시점에 워핑된 두 영상을 하나로 합치는 과정으로 다른 시점의 홀 영역을 서로 채워주는 역할을 한다. 마지막 세 번째 과정은 투사 과정 중 생긴 결함이나 영상 통합 과정에서 처리되지 못한 홀 영역에 대한 필터링을 통한 화질 개선 과정이다 [4].

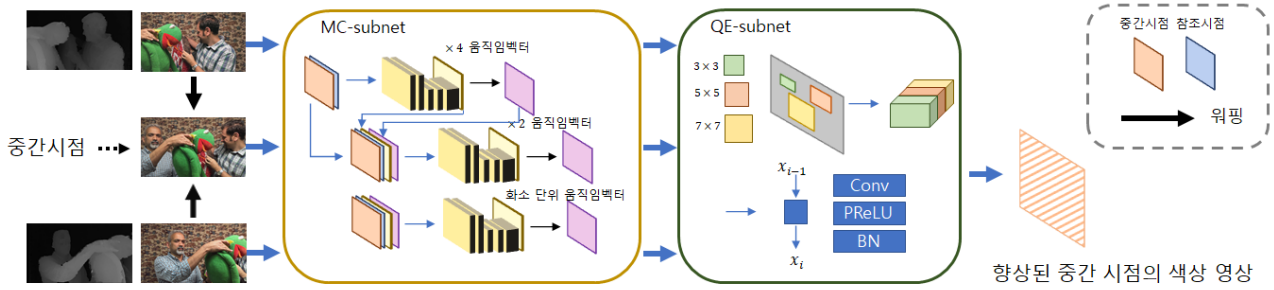

그림 1 제안 알고리듬의 도식도

그 중 본 논문에서 사용한 VVS 는 기존 VSRS 대비 주관적인 품질 향상을 우선적으로 고려하여 설계된 알고리듬이다. 이는 MPEG-I 영상 그룹의 124 차 마카오 미팅에서 6 자유도 영상을 위해 채택되었다. 특히, VVS 는 입력 영상들이 가상 시점 영상에서 얼마나 떨어져 있는지 워핑 거리 를 척도로 구분하여, 투사한 결과를 통합할 때 사용한다. 또한, 홀 뿐만 아니라 주관적 화질 향상을 위하여 영상 내 객체의 경계를 주위로 필터링이 적용된다. 그러나 기존의 중간시점합성에 관하여 시공간적으로 균일한 화질을 제공하지 못하는 한계가 있다.

3. 제안 방법

그림 1 은 제안 알고리듬의 도식도이다. 제한된 입력 영상이 존재할 때, 해당 시점의 색상영상과 깊이 영상을 이용하여 중간 시점의 영상을 생성 후, MF-CNN (Multil-Frame Convolutional Neural Networks)에 묶음으로 넣어 전체적으로 향상된 프레임들을 얻는다 [5]. 이때 같은 묶음들은 서로 다른 위치의 입력 영상으로부터 얻어지지만 시간적으로 같은 특성을 갖는다. 참조시점을 사용하여 위치적으로 중간에 위치한 가상시점을 VVS2.0 으로 합성한 뒤 시간이 동일한 프레임들을 묶음으로 MF-CNN 에 적용하면, 참조시점들을 중간에 위치한 가상시점 합성 과정만 아니라 잡음제거에도 활용할 수 있는 이점이 있다.

그림 1 에서 보듯이 먼저 참조시점들과 중간시점 간의 공간적 움직임을 보상하기 위해 움직임 벡터를 다중 스케일로 측정한 뒤 참조시점을 중간시점에 워핑하여 보완된 참조시점들과 중간시점을 다중 스케일로 특징을 추출한다. 이어서 참조시점들과 중간시점 간의 차이를 보완하는 차분 학습을 진행하여 향상된 중간시점의 색상 영상을 구한다. 이러한 네트워크의 구조는 시간적 특성이 동일한 다른 시점에서의 영상들에 적용 가능할 뿐 아니라, 한 시점에서 얻어진 서로 다른 시간 상에 위치한 영상들 사이에도 적용이 가능하다. 프레임 간 깊이 영상의 정확도 차이를 MF-CNN 에 묶음으로 적용함으로써 시간적 품질 저하도 역시 향상 가능하다.

A. 중간시점 합성

두 개의 입력영상에 대해 위치적으로 중간에 위치해 있는 중간시점의 영상을 합성하기 위해서 6 자유도 영상의 공개 소프트웨어인 VVS2.0 (Versatile View Synthesis)을 사용한다. VVS2.0 의 구조는 그림 2 와 같다. 이 구조는 다른 알고리듬에 비해 주관적 성능이 우수하다. 특히 세 번째와 아홉 번째에서의 경계처리 방식에서 우수한 주관적 성능 향상이 달성된다.

VVS2.0 은 2 장에서 설명한 영상합성 과정의 3 단계의 거치기 전에 그림 2 의 첫 번째 구성요소에서 입력영상이 중간시점의 영상으로부터 얼마나 떨어져 있는지 확인하기 위해 워핑 거리를 계산하고 이에 따라 정렬하여 입력이 어느 순서로 들어오는지에 따라 결과가 변하지 않도록 한다. 더불어 세 번째 구성요소를 보면 입력영상의 깊이 영상을 가장 결함이 많이 나는 경계 위주로 안전한 정보와 그렇지 않은 정보를 구분한다. 이 이후에 영상합성 과정을 거친다. 또한 합성이 끝난 뒤 아홉 번째 구성요소에서 홀과 더불어 경계 위주의 필터링을 거친다.

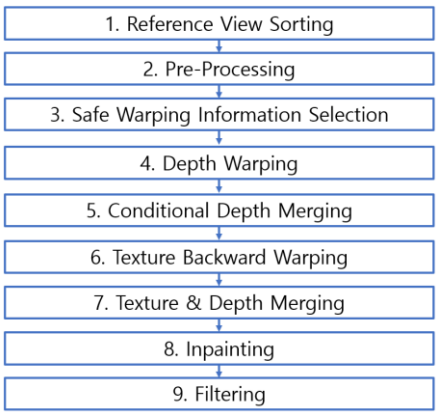

그림 2 VVS (Versatile View Synthesis) 구조

B. MF-CNN

MF-CNN 의 구조는 그림 1 에서는 보는 바와 같이 두 개의 서브넷으로 구성된다. 첫 번째 서브넷은 MC-subnet (Motion Compensation subnet) 으로 인접 시점 간의 움직임 벡터를 다중 스케일로 예측하여 시간적 움직임에 따른 움직임 벡터를 스케일 별로 획득한다. 획득된 움직임 벡터를 기준으로 참조시점을 중간시점에 워핑하여 시점 간의 움직임을 보상한다. 이 때 다중스케일로 움직임을 측정하는 것은 지역적인

움직임과 전역적인 움직임을 모두 보상하기 위함이다. MC-subnet 을 거쳐 워핑된 참조시점들은 중간시점과 두 번째 서브넷을 거치게 된다. 두 번째 서브넷인 QE-subnet (Quality Enhancement subnet) 에서는 입력의 특징을 3*3, 5*5, 7*7 의 다중 스케일로 추출해내어 이후 서로 간의 차이를 보완하기 위해 차분 학습을 하는 구조를 가진다 [5].

4. 실험 결과

본 논문에서 제안하는 가상시점 합성 영상 변경의 성능을 평가하기 위하여 참조 소프트웨어인 VVS2.0.1 에서 공통 실험 조건 CTC (Common Test Condition)을 참고하였다 [6]. 또한 MF-CNN 모델은 별도의 재학습 없이 기존 사전학습 (Pre-trained) 모델을 사용하였다. 또한, 실험을 위해 서로 다른 열 다섯 개의 시점들이 3.675cm 간격으로 수평적으로 배열된 IntelKermit 시퀀스와 스테레오 카메라 다섯 개가 호를 그리며 20cm 간격으로 떨어져 있는 PoznanFencing 시퀀스를 사용하였다. [7-8]

실험은 두 가지 방식으로 진행하였다. 첫 째로는 동일한 시간적 특성을 같는 서로 다른 시점의 영상들을 묶음으로 MF-CNN 적용하여 공간적 품질 저하를 개선하였다. 둘 째로는 동일한 공간적 특성을 갖는 한 시점의 연속된 영상들에 대해 MF-CNN 적용하여 시간적 품질 저하를 개선하였다.

첫 번째 실험 환경은 그림 3 와 같다. IntelKermit 의 경우 열 다섯 개의 시점 중 여덟 개의 시점을 참조시점으로 사용하였고, PoznanFencing 의 경우 열 개 중 다섯 개를 참조시점으로 사용하였고, 가장 자리에 있는 마지막 시점의 영상은 사용하지 않았다.

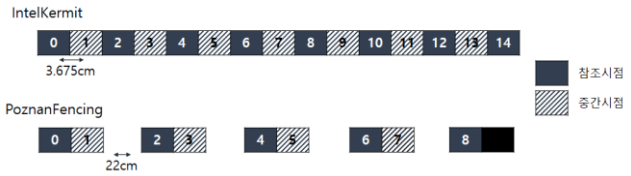

그림 3 첫 번째 실험 환경

두 번째 실험 환경을 그림 4 와 같다. 하나의 시퀀스에 대해 전체 프레임 중 주기적으로 8 프레임마다 하나의 참조시점이 존재하고, 그 사이의 일곱개의 중간시점이 존재한다.

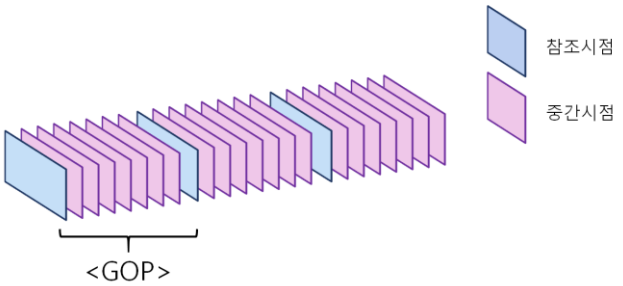

그림 4 두 번째 실험 환경

첫 번째 실험 결과 기존 중간시점 영상 합성 방식 대비 딥러닝 기반 시간적 특성이 동일한 묶음 조정 방법은 평균적으로 약 0.34dB 의 성능 향상을 제공하였다. 그림 5 와 6 은 첫 번째 실험의 결과로 IntelKermit 시퀀스와 PoznanFencing 시퀀스의 VVS 대비 제안 알고리즘의 향상된 PSNR 을 확인할 수 있다. 가로축은 그림 3 의 해당하는 중간시점의 번호이며, 세로축은 증가한 PSNR 이다. IntelKermit 과 PoznanFencing 시퀀스는 MPEG 의 깊이영상 예측 참조 소프트웨어인 DERS (Depth Estimation Reference Software)로 깊이 정보를 취득하였는데, CG 로 생성된 영상들과 달리 자연영상인 관계로 그림 7 과 같이 깊이 정보의 정확도가 상대적으로 정확하지 않다 [1]. 제안 알고리즘은 깊이 정보의 부정확성에 의한 공간적인 방향으로의 합성 잡음을 효과적으로 개선하는 것을 관찰할 수 있다.

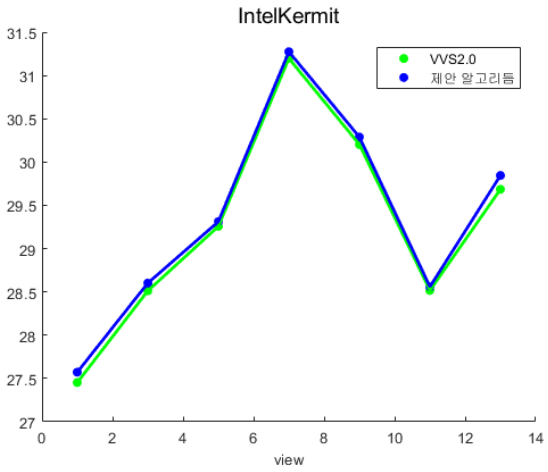

그림 5 IntelKermit 의 실험결과

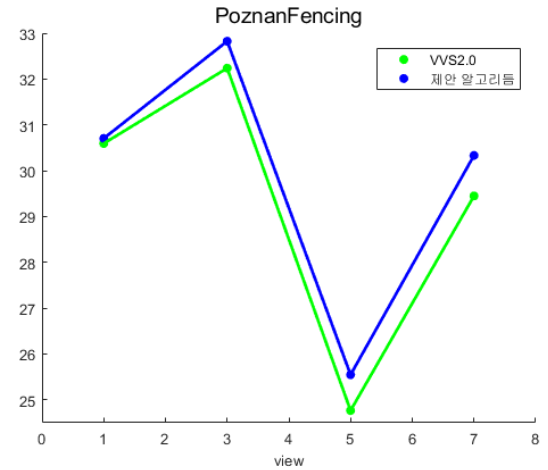

그림 6 PoznanFencing 의 실험결과

<색상 영상> <깊이 영상>

그림 7 Poznan Fecing 시퀀스의 색상영상(좌) 깊이영상(우)

두 번째 실험 결과 기존 중간시점 영상 합성 방식 대비 딥러닝 기반 공간적 특성이 동일한 묶음 조정방법은 평균적으로 0.81dB 의 성능 향상을 제공하였다. IntelKermit 시퀀스의 300 프레임, PoznanFencing 시퀀스의 250 프레임 중 주기적으로 8 장마다 프레임들이 MF-CNN 적용되어 화질이 향상됨을 PSNR 로 확인할 수 있다. 그림 8 과 9 은 각 시퀀스에 대한 두 번째 실험의 전체 결과 중 한 주기에 대한 PSNR 향상을 확인할 수 있다.

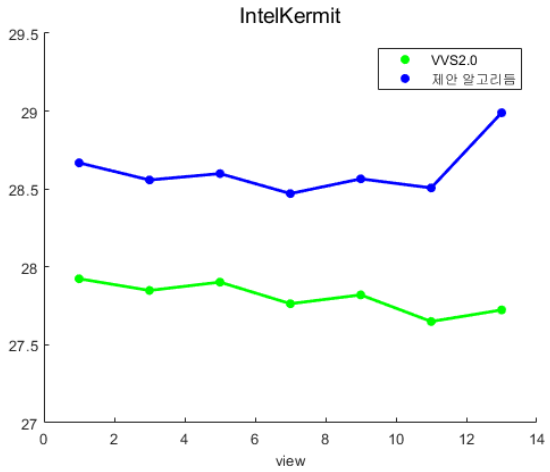

그림 8 IntelKermit 의 실험 결과

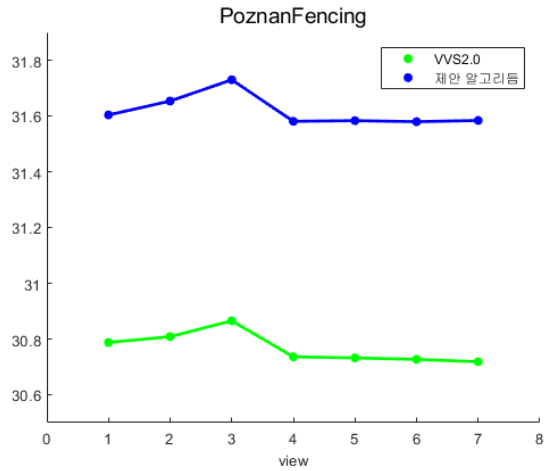

그림 9 PoznanFencing 의 실험 결과

5. 결론

본 논문에서는 참조시점을 사용하여 합성된 중간시점의 영상을 대상으로 묶음 조정 개념의 딥러닝을 적용하여 영상 간 공간적 품질 차이를 낮추는 방법을 제안하였다. 제안 방법의 실험 결과, 기존 방법보다 공간적으로 약 0.34dB, 시간적으로 0.81dB 개선하였다. 중간시점과 참조시점을 묶음으로 조정하는 데에 딥 러닝 기법을 적용한다는 점에 의의를 가진다. 향 후 연구로는 MF-CNN 보다 우수한 딥 러닝 알고리즘을 개발하여 화질을 개선하는 것이다.

감사의 글

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2018-0-00765, Development of Compression and Transmission Technologies for Ultra High Quality Immersive Videos Supporting 6DoF.)

참고문헌

[1] ISO/IEC JTC1/SC29/WG11, *FTV software user guidelines, m36590*, June. 2015.

[2] ISO/IEC JTC1/SC29/WG11, *Reference View Synthesizer (RVS) manual, N18068*, October. 2018.

[3] ISO/IEC JTC1/SC29/WG11, *Versatile View Synthesizer 2.0 (VVS 2.0) manual, w18172*, October. 2018.

[4] 호요성 외 1 인, “3 차원 비디오의 이해와 분석,” 진샘미디어, 2011.

[5] Yang, R., Xu, M., Wang, Z., &Li, T. (2018). Multi-frame quality enhancement for compressed video. In IEEE conference on computer vision and pattern recognition (pp 6664-6673).

[6] ISO/IEC JTC1/SC29/WG11, *Common Test Conditions for Immersive Video, w18443*, March. 2019.

[7] ISO/IEC JTC1/SC29/WG11, *Kermit test sequence for Windowed 6DoF Activities, m43748,* July. 2018.

[8] ISO/IEC JTC1/SC29/WG11, Multiview test video sequences for free navigation exploration oabtained using pairs of cameras, *m38247,* may. 2016.