세그먼테이션과 스타일 변환을 활용한 영상 재구성 시스템

방연준 이의진 박주형 *강병근

서울과학기술대학교

yjbang0529@seoultech.ac.kr, yeejinlee@seoultech.ac.kr, sks3389@seoultech.ac.kr,
*byeongkeun.kang@seoultech.ac.kr

Image Recomposition System Using Segmentation and Style-transfer

Bang, Yeonjun Lee, Yeejin Park, Juhyeong Kang, Byeongkeun

Seoul National University of Science and Technology

요약

기존 영상 콘텐츠에 새로운 물체를 삽입하는 등의 영상 재구성 기술은 새로운 게임, 가상현실, 증강현실 콘텐츠를 생성하거나 인공신경망 학습을 위한 데이터 증대를 위해 사용될 수 있다. 하지만, 기존 기술은 컴퓨터 그래픽스, 사람에 의한 수동적인 영상 편집에 의존하고 있어 금전적/시간적 비용이 높다. 이에 본 연구에서는 인공지능 신경망을 활용하여 낮은 비용으로 영상을 재구성하는 기술을 소개하고자 한다. 제안하는 방법은 기존 콘텐츠와 삽입하고자 하는 객체를 포함하는 영상이 주어졌을 때, 객체 세그먼테이션 네트워크를 활용하여 입력 영상에서 객체를 분리하고, 스타일 변환 네트워크를 활용하여 입력 영상을 스타일 변환한 후, 사용자 입력과 두 네트워크의 결과를 활용하여 기존 콘텐츠에 새로운 객체를 삽입하는 것이다. 실험에서는 기존 콘텐츠는 온라인 영상을 활용하였으며 삽입 객체를 포함한 영상은 ImageNet 영상 분류 데이터 세트를 활용하였다. 실험을 통해 제안한 방법을 활용하면 기존 콘텐츠와 잘 어우러지게끔 객체를 삽입할 수 있음을 보인다.

1. 서론

기존에 촬영했거나 새롭게 제작한 콘텐츠의 일부 영역에 새로운 객체를 자연스럽게 추가하는 기술은 많은 경우에 활용될 수 있다. 예로써, 촬영한 사진에 해당 사진 촬영 시 참석하지 못했던 사람을 추가한 사진을 생성할 수 있고, 기존 영상에 새로운 객체를 자연스럽게 추가한 영상을 생성한 후, 해당 영상을 인공신경망 학습에 활용하는 데이터 증대 기술로도 사용할 수 있다. 또한, 게임 콘텐츠, VR/AR 콘텐츠에서도 지속적으로 새로운 물체를 생성/삽입하기 위하여, 실세계에 가상의 물체를 자연스럽게 시각화하기 위하여 활용할 수 있다.

하지만, 기존 기술의 경우, 컴퓨터 그래픽스를 활용하여 새로운 객체를 생성한 후, 조도, 카메라 각도, 날씨 등을 고려하여 추가로 편집하거나, 혹은 해당 객체를 추가로 촬영한 후, 기존 콘텐츠의 환경을 고려하여 추가로 편집하여 영상을 재구성해야 하며 이러한 기술들은 금전적, 시간적 비용이 높다는 문제가 있다. 이에 본 연구에서는 인공지능 신경망을 활용하여 낮은 비용으로 자연스럽게 콘텐츠를 재구성하거나 물체를 삽입하는 기술을 소개하고자 한다.

본 논문에서 제안하는 방법은 기존 콘텐츠와 삽입하고자 하는 객체를 포함한 입력 영상이 주어졌을 때, 인공지능 신경망 기반 세그먼테이션 네트워크를 활용하여 주어진 입력 영상에서 삽입하고자 하는 객체의 마스크를 추정하고, GAN(Generative Adversarial Networks) 기반 스타일 변환 네트워크를 활용하여 입력 영상의 스타일을 기존 콘텐츠와 유사하게 변환한다. 그리고, 객체 마스크(instance mask)와 스타

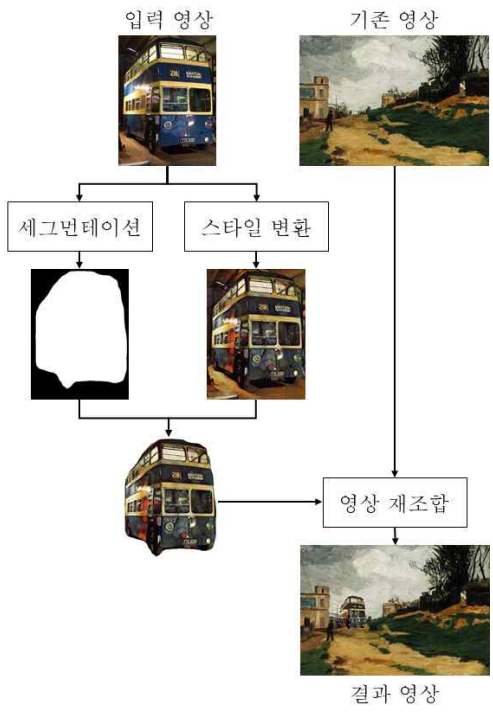

그림 1. 시스템 구성도

일 변환된 입력 영상을 활용하여 스타일 변환된 객체를 추출한 후, 사용자로부터 해당 객체를 배치하고자 하는 위치를 입력받아 위치시킨다. 제안하는 시스템의 구성은 그림 1에 시각화되어있다.

2. 관련 연구

2.1. 영상 재구성

영상 재구성 기술에 대한 연구는 지속적으로 있었다. Zhang 등은 배경 영상이 주어졌을 때 해당 영상에 잘 어울리는 물체 영상을 삽입하는 방법에 대해서 제안했다 [1]. 그러나 해당 논문에서는 이미 세그먼테이션된 객체 영상을 이용하고, 재구성 시의 스타일 변화는 다루지 않았다. MetaOD는 탐지 모델의 성능 평가를 위해 영상 재구성하는 방법을 발표했다 [2]. MetaOD는 입력 영상을 세그먼테이션한 후, 영상을 재구성하는 방법을 다루지만 스타일 변환은 수행하지 않았다. Tarko 등은 3D 영상의 영상 재구성 방법을 제안했다 [3]. 그러나 영상 재구성 시 이미 세그먼테이션된 영상을 활용하고 각각의 톤을 맞추는 과정은 있지만, 추가적인 스타일 변환 과정을 적용하지 않았다. 본 논문은 세그먼테이션 모듈과 스타일 변환 모듈을 모두 채택하여 영상 재구성을 진행한다.

2.2. 세그먼테이션

세그먼테이션이란 영상을 여러 개의 픽셀 집합으로 나누는 과정을 말한다. 세그먼테이션 기술 중 각 픽셀은 사물 종류별로 분류하는 시맨틱(semantic) 세그먼테이션과 각 픽셀을 사물 종류별 그리고 객체별로 분류하는 객체(instance) 세그먼테이션이 본 연구에서 세그먼테이션을 통해 얻고자 하는 결과물과 가장 연관이 있다. 두 세그먼테이션 기술 분야에서는 최근 인공신경망 기반 세그먼테이션 알고리즘들이 주로 가장 높은 정확도를 달성했다. 이에, 본 장에서는 인공신경망 기반 시맨틱 세그먼테이션 알고리즘과 인공신경망 기반 객체 세그먼테이션에 대해 간략히 소개하고자 한다.

초기 인공신경망 기반 시맨틱 세그먼테이션 알고리즘 중 하나인 FCN [4]은 기존 분류(classification) 모델의 구조에서 마지막 완전 연결 층(fully connected layer)을 컨볼루션 층(convolution layer)으로 대체하고 업 샘플링(upsampling)을 통해 세그먼테이션을 수행하였다. 또한, 스킵 커넥션(skip connection)을 통해 일부 앞단의 컨볼루션 층과 뒤쪽 컨볼루션 층을 연결함으로써 위치 정보를 보다 더 보존하여 세그먼테이션 결과를 개선하였다. U-Net [5]은 대칭적인 구조의 인코더-디코더 구조를 이용하였고 이를 통해 상대적으로 더 정교한 위치정보를 얻을 수 있다고 한다.

객체 세그먼테이션 알고리즘 중에서 YOLACT [6]는 기존 알고리즘의 객체별 마스크 추정 단계를 대신하여 프로토타입 마스크를 생성과 객체별 프로토타입 마스크 계수를 추정함으로써 실시간 객체 세그먼테이션을 가능하게 했다. Mask R-CNN [7] 모델은 YOLACT [6] 모델보다 속도는 느리지만, 성능이 더 좋다는 장점이 있고 본 논문에서는 동작 시간보다는 성능이 더 중요하므로 Mask R-CNN [7] 모델을 채택했다.

2.3. 스타일 변환

스타일 변환이란 주어진 영상을 다른 스타일로 변화시키는 기술을 말한다. 예를 들면 풍경 사진을 고흐 풍의 풍경 그림으로 바꾸거나 낮에 촬영한 사진을 밤에 촬영한 사진으로 바꾸는 기술을 말한다. 스타일 변환은 여러 가지 방법으로 수행 가능한데, 그중 하나는 입력 영상과 스타일 영상, 두 영상의 특징 맵을 이용하여 새롭게 만들 영상의 특징 맵을 일치시키는 방법이 있다. 이 방법의 경우 영상은 두 개만 있으면 스타일 변화가 가능하지만, 입력 영상이 변화될 때마다 매번 학습을 다시 해야 한다는 단점이 있다. Image Style Transfer Using Convolutional Neural Networks [8]에서는 해당 방법을 이용하여 스타일 변화하는 방법을 소개한다. 그러나 해당 방법은 텍스쳐 합성을 할 때 불안정하다는 단점이 있다.

또 다른 방법은 스타일 변화 네트워크를 학습시키는 방법이다. 네트워크 자체를 학습시키다 보니 입력 영상과 스타일 영상이 여러 장 필요하고 학습 시에도 시간이 많이 소요되지만, 한번 학습을 하면 서로 다른 입력 영상에 대하여 스타일 변화가 가능하다는 장점이 있다. Pix2Pix [9]는 이 방법을 이용하여 네트워크를 학습한다. 그러나 Pix2Pix [9]의 경우 기존 GAN 구조에 입력 영상과 스타일 영상의 대응되는 좌푯값의 차이를 추가적인 손실 함수로 이용하여 학습을 진행하므로 입력 영상과 스타일 영상이 짝을 이뤄야 한다는 단점이 있다. 이러한 단점을 해결하기 위해 Cycle GAN [10]이 제안되었다. Cycle GAN [10]의 경우 입력 영상과 스타일 영상이 짝을 이루지 않기 때문에 추가적인 손실 함수로 입력 영상을 GAN 구조를 통과시켰을 때 스타일 변화된 영상이, 그리고 스타일 변화된 영상을 다시 GAN 구조를 통과시켰을 때 기존의 입력 영상이 출력되게끔 하는 순환 일관성 손실 함수(cycle consistency loss)를 도입한다. 본 논문에서는 입력 영상과 스타일 영상을 매칭시키기 어렵기 때문에 Cycle GAN [10]을 사용하였다.

3. 방법

본 논문에서 제안하는 방법은 그림 1에서 확인할 수 있듯이 세 개의 모듈(세그먼테이션, 스타일 변환, 영상 재구성)로 구성되어 있다. 먼저, 온라인에 존재하는 영상들 중 사용자가 기존 콘텐츠와 재구성하고자 하는 입력 영상을 선택하고 해당 영상을 세그먼테이션 모듈과 스타일 변환 모듈에 입력한다. 세그먼테이션 모듈에서는 미리 학습된 객체 세그먼테이션 네트워크를 적용하여 입력 영상에 존재하는 객체의 마스크를 획득한다. 스타일 변환 모듈에서는 입력 영상을 기존 콘텐츠의 분위기와 일치시키기 위해 GAN을 활용하여 스타일 변화를 수행한다. 마지막으로 영상 재구성 모듈에서는 기존 콘텐츠, 스타일 변환된 입력 영상, 객체 마스크를 활용하여 스타일 변환된 입력 영상에서 객체 영역을 분리하고 사용자가 지정한 기존 콘텐츠 위치에 해당 객체를 삽입하여 재구성된 결과 영상을 획득한다.

3.1. 세그먼테이션 모듈

세그먼테이션 모듈에서는 입력 영상을 객체 또는 배경으로 픽셀 단위로 분할하고 이를 통해 객체 영역에 해당하는 마스크를 획득하기 위해 Mask R-CNN [7] 모델을 활용한다. 해당 모델은 물체 탐지를 수행하는 Faster R-CNN [11] 모델에 기반을 두고 있어 해당 물체 탐지 모델과 유사한 방식으로 객체의 경계 박스(bounding box)와 사물 종류(object class)를 추정하는 동시에 추가된 마스크 브랜치를 활용하여 객체별로

픽셀 단위 객체 마스크를 추정한다.

구체적으로는 주어진 입력 영상을 해당 영상의 비율을 유지하면서 짧은 변의 길이를 800픽셀 혹은 긴 변의 길이를 1,333픽셀의 사이즈로 변환한 후, 위치 정보 손실을 최소화하기 위하여 FPN(Feature Pyramid Network) [12]을 통과 시켜 특징 맵을 추출한다. 추출한 특징 맵들을 RPN(Region Proposal Network)를 통과 시켜 여러 관심 영역 중 최적의 관심 영역을 추정하고 이 관심 영역과 특징 맵을 이용하여 객체의 경계 박스 위치와 사물의 종류를 추정한다. 이후 해당 객체에 해당하는 픽셀들을 추정하여 객체 마스크를 구할 수 있다.

3.2. 스타일 변환 모듈

스타일 변환 모듈은 입력 영상을 기존 콘텐츠와 유사한 스타일로 변환시켜주는 역할을 한다. 본 논문에서는 이를 위해 Cycle GAN [7]모델을 사용하였다. Cycle GAN [7] 모델의 경우 기존 GAN 구조를 응용하여 입력 영상을 기존 콘텐츠와 유사한 스타일로 변화시켜 새로운 영상을 생성하는 네트워크(F), 생성된 영상을 입력 영상으로 기존 입력 영상과 유사한 영상을 생성하는 네트워크(G)를 동시에 학습한다. 그러나 이때 기존의 GAN과 같은 목적함수를 이용하여 학습하면 기존의 콘텐츠를 유지하지 않으므로 새롭게 생성된 영상이 다시 기존의 영상으로 변환될 수 있게 만들어주는 순환 일관성 손실 함수(cycle consistency loss)를 추가하였다. 다시 말해, 기존 입력 영상(X)을 네트워크(F)를 이용하여 새로운 영상(Y)을 만들고 이때 새롭게 만들어진 영상을 네트워크(G)를 이용하여 새로운 영상(Z)을 만들었을 때 기존 영상(X)과 새로운 영상(Z)이 같은 영상이어야 한다. 이때 기존 영상(X)과 새로운 영상(Z)의 차이를 순환 일관성 손실 함수로 정의하고 기존 함수에 추가하여 학습을 진행한다. 이 방법으로 네트워크들을 학습시키면 입력 영상에 대하여 객체는 유지한 상태로 다른 스타일로 변화시킬 수 있다. 따라서 이 모듈을 통해 입력 영상을 기존 콘텐츠와 비슷한 스타일로 변환할 수 있다.

4. 결과

본 논문에서 제안한 방법으로 콘텐츠를 재구성한 영상은 그림 2와 다. 입력 영상은 ImageNet 데이터 세트의 콘텐츠를 이용하였고, 기존 콘텐츠는 온라인에 있는 데이터를 활용하였다.

또한, 모듈별 결과는 그림 3, 그림 4에서 확인할 수 있다. 먼저 그림 3은 입력 영상을 세그먼테이션 모듈을 통과 시켜 획득한 영상이다. 가운데 열은 입력 영상의 객체의 마스크를 흰색으로, 다른 영역을 검은색으로 표현하였다. 오른쪽 열은 해당 마스크를 이용하여 객체를 픽셀 단위로 배경과 분리한 결과이다.

그림 4는 스타일 변환 모듈을 이용하여 원본 영상을 스타일 변환시킨 결과이다. GAN을 활용한 스타일 변환 결과를 보면 다음과 같이 원하는 콘텐츠에 맞게 스타일 변화를 할 수 있음을 확인할 수 있다. 첫번째 열은 원본 영상, 두 번째 열은 원본 영상을 모네의 그림 스타일로 변화시킨 영상, 세 번째 열은 원본 영상을 어스름이 질 때 촬영한 영상처럼 스타일 변화시킨 결과, 네 번째 열은 원본 영상을 해 뜰 녘에 촬영한 영상처럼 스타일 변화시킨 결과이다. 각각 스타일에 맞게 원본 영상을 스타일 변화시키는 것을 알 수 있다.

그림 2. 제안한 기술을 적용한 재구성 결과 영상

그림 3. 세그먼테이션 모듈을 이용한 세그먼테이션 결과 영상

그림 4. 스타일 변환 모듈을 이용한 스타일 변화 영상

5. 결론

본 논문에서는 인공지능 신경망을 이용한 영상 재구성 방법을 제안하였다. 이 방법을 통해 기존 방법과는 다르게 영상을 재구성할 때 추가적인 촬영 또는 편집이 요구되지 않으므로 금전적/시간적으로 큰 비용을 절감할 수 있다. 또한 추가적인 편집 없이 기존 콘텐츠와 재구성한 객체가 자연스럽게 어우러짐을 확인할 수 있다. 추후에는 객체를 영상에 위치시키는 과정 또한 인공지능 신경망을 이용하여 자동화한다면 사용자의 비용 부담을 더욱 줄일 수 있으리라 생각한다.

Acknowledgement

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2020-0-00994, 이용환경을 반영하는 자율적 VR·AR 콘텐츠 생성 기술개발).

참고문헌

[1] SH. Zhang, ZP. Zhou, B. Liu, X. Dong, D. Liang, P. Hall and S.-M. Hu, "What and Where: A Context-based Recommendation System for Object Insertion," Comp. Visual Media 6, 79-93, 2020. https://doi.org/10.1007/s41095-020- 0158-8.

[2] S. Wang and Z. Su, "Metamorphic Object Insertion for Testing Object Detection Systems," 2020 35th IEEE/ACM International Conference on Automated Software Engineering (ASE), 2020, pp. 1053-1065.

[3] Joanna Tarko, James Tompkin, and Christian Richardt. 2019. Real-time Virtual Object Insertion for Moving 360° Videos. In The 17th International Conference on Virtual-Reality Continuum and its Applications in Industry (VRCAI '19). Association for Computing Machinery, New York, NY, USA, Article 14, 1-9. DOI:https://doi.org/10.1145/3359997.3365708.

[4] J. Long, E. Shelhamer and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431-3440, doi: 10.1109/CVPR.2015.7298965.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," In: Navab N., Hornegger J., Wells W., Frangi A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science, vol 9351. Springer, Cham. https://doi.org/10.1007/ 978-3-319-24574-4_28.

[6] D. Bolya, C. Zhou, F. Xiao and Y. J. Lee, "YOLACT: Real-Time Instance Segmentation," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9156-9165, doi: 10.1109/ICCV.2019.00925.

[7] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980-2988, doi: 10.1109/ICCV.2017.322.

[8] L. A. Gatys, A. S. Ecker and M. Bethge, "Image Style Transfer Using Convolutional Neural Networks," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2414-2423, doi: 10.1109/CVPR.2016.265.

[9] P. Isola, J. Zhu, T. Zhou and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5967-5976, doi: 10.1109/CVPR.2017.632.

[10] J. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2242-2251, doi: 10.1109/ICCV.2017.244.

[11] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017, doi: 10.1109/TPAMI.2016.2577031.

[12] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature Pyramid Networks for Object Detection," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 936-944, doi: 10.1109/CVPR.2017.106.