가상 스튜디오 시스템을 위한 생성모델 기반 영상 합성 방법

정진우, 조인휘

한양대학교

jinwoo0427@hanyang.ac.kr, iwjoe@hanyang.ac.kr

Image Synthesis Method based Generation Model for Virtual Studios System

Zheng Zhen Yu, Joe In Whee*

Hanyang Univ.

요 약

코로나 19 로 시작된 팬더믹으로 인해 비접촉식 시스템에 대한 관심이 폭발적으로 늘어나게 되었다. 그 중에서 컴퓨터비전 기술을 이용하여 사진을 합성해 낼 수 있는 가상 스튜디오에 대한 연구가 각광을 받고 있다. 기존의 가상스튜디오 시스템에서는 원하는 포즈의 사진을 촬영한 후 배경과 합성하는 방식을 사용하고 있다. 하지만 이는 사용자가 특별한 포즈의 영상을 제공해야 할 뿐만 아니라 원하는 씬 과의 자연스러운 합성을 위해 심도, 크기 등 다양한 요소를 고려해야 하며, 이는 실제 응용의 발전을 저해하고 있다. 본 논문에서는 이미지 생성모델을 이용하여 한 장의 인물 영상으로부터 필요한 포즈의 이미지를 자동으로 생성하고 씬을 재구성하는 자동화된 영상 합성 방법을 제안한다.

Ⅰ. 서 론

SNS 의 발전으로 현재 사진은 생활 속의 중요한 구성 부분 중 하나로 인식되고 있다. 결혼, 여행 및 생일 등과 같은 중요한 행사 뿐만 아니라 일상생활 곳곳에서 사진을 찍고 공유하며 기록을 남긴다. 따라서 마음에 드는 사진 한 장을 얻기 위해서 스튜디오를 찾는 사람이 많아지고 있다. 하지만 팬데믹 시대에 접어 들 면서 스튜디오를 일일이 찾아다니며 발품을 파는 전통적인 방식은 필연적으로 접촉을 강요받게 되어 사회적인 거부감을 야기하게 된다. 또한 다 양한 스타일로 꾸며진 현실 세트장 부지와 사진작가, 사진보조 및 디자이너에 의류와 장신구 등 소모되는 비용 또한 만만치 않다. 하지만 영상합성 기술을 이용하는 소프트웨어 기반의 방식은 불필요 한 접촉을 차단할 뿐만 아니라 비용 또한 획기적으로 낮출 수 있어 청년 세대의 많은 관심을 받고 있다.

시중의 영상합성 프로그램들은 일반적으로 적절한 포즈로 잘 세팅된 인물과 배경을 합성하는 방식이 주를 이루고 있다. 하지만 중요한 사진 만큼, 한 두 번의 시도로 원하는 결과를 얻어 내기는 어렵다. 따라서 전통적인 현실 속 스튜디오에서는 많은 시간과 체력을 들여 의상과 포즈를 테스트 해가며 사진을 촬영해야 한다. 이는 사용자가 적절한 사진을 제공해야 하는 현 시중의 스튜디오 프로그램들 또한 크게 다를 바가 없이, 반복적으로 다양한 포즈의 사진을 촬영하고 합성결과를 테스트를 해야 한다.이런 불편함을 해결하고자, 본 논문에서는 사용자가 원하는 시나리오의 사진 한 장과 사용자 본인의 사진 한 장만 제공하면, 자동으로 유저의 사진으로부터 시나리오에 적합한 포즈의 영상을 생성하고 합성하여 새로운 영상을 합성해 내는, 자동화된 가상 스튜디오 시스템을 제안한다.

Ⅱ. 가상 스튜디오 시스템

본 논문에서 제안하는 시스템의 전반적인 구조는 그림 1에서 보여주는 바와 같다. 우선 원하는 시나리오가 포함된 스크립트 영상(Script Image) 한 장과 본인의 사진 한 장을 선택하여 시스템의 입력으로 한다. 여기서 입력된 본인의 사진을 참조영상(Reference Image) 이라고 칭한다. 입력된 스크립트 영상은 영상분할(Image Segmentation) 방법을 이용해 배경과 자신의 이미지로 교체를 해야 할 대상객체(Target Object)를 분리해 낸다. 본 논문에서는 빠른 실행속도를 자랑하는 YOLACT [1] 모델을 이용하여 영상분할을 진행하였다. 계산 양을 좀 더 줄이기 위해 우선 입력된 영상으로부터 사람검출 알고리즘을 이용하여 사람을 검출 한 후 검출된 관심영역에 대해서만 화소단위의 영상분할을 진행한다. 이때 대상객체는 참조영상과 다른 포즈를 취하고 있으며 교체를 진행하기 위해서는 우선 대상객체의 포즈 정보를 알아야 한다. 따라서 본 논문에서는 AlphaPose [2] 알고리즘을 이용하여 대상객체의 포즈를 획득하였다. AlphaPose는 먼저 사람을 검출하고 다음 포즈를 추정하는 two-step 포즈추정 전략을 사용하여 정확도를 향상하였는데, 본 논문에서는 전 단계에서 검출된 사람의 검출영역을 AlphaPose의 첫 단계에 그대로 사용하였다. 그림 2 에서 영상분할 알고리즘의 결과를 보여준다.

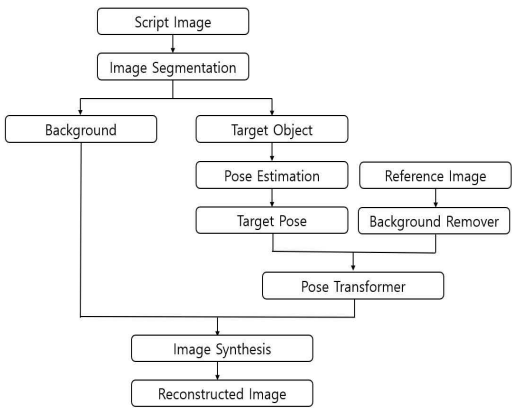


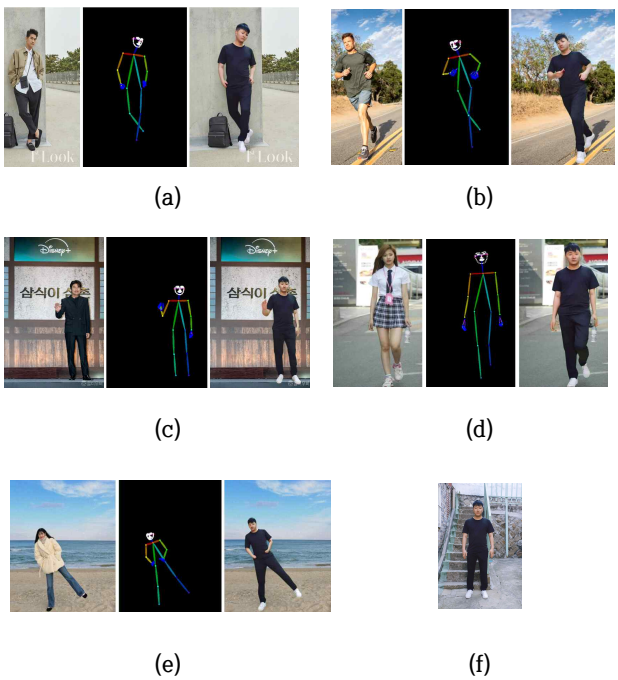
그림 1. 시스템 흐름도

(1) (2) (3) (4)

그림 2.

다음 참조영상에서 배경을 제거한 뒤 포즈전이(Pose Transfer) 알고리즘을 이용하여 대상객체의 포즈를 참조영상에 전이한다. 이때 본인의 사진을 목표영상과 합성하기 위해서는 참조영상의 신원정보(identity)와 외관정보(appearance)의 일관성이 유지되는 것이 중요하다. 본 논문에서는 Diffusion model [3]에 기반 한 Image-to-video 생성 모델인 AnimateAnyone[4]을 이용하여 새로운 포즈의 참조영상을 생성하였다. Image-to-video를 위한 모델들은 PIDM[6] 과 같은 Image-to-Image 모델들에 비해 참조영상에 대한 일관성을 유지하면서 포즈가 부드럽게 변화하는 영상을 생성하는 것이 주요 목표이기에 보다 안정적으로 신원정보와 외관정보를 유지하면서 새로운 포즈영상을 생성할 수 있다. 본 논문에서 제안하는 시스템에 적용한 AnimateAnyone 모델에서는 Clip encoder 나 Cross-attention을 사용하지 않고 Denoising Unet과 유사한 구조를 가지는 ReferenceNet을 이용하여 참조영상에 대한 일관성을 보장하였다.

이로써 얻어진 새로운 포즈의 참조영상은 영상합성 기술을 이용하여 스크립트영상으로부터 분리된 배경과 합성하여 새로운 영상을 얻게 된다. 영상을 합성할 때 부자연스러운 경계가 발생하는 것을 방지하기 위해 이미지 블렌딩(Image Blending) 방법을 사용한다. 본 논문에서는 화소와 윤곽선사이 거리를 나타내는 거리지도(Distancemap) 을 생성하고 거리에 따른 가중치를 부가하는 선형블랜딩 기법을 사용하여 부자연스러운 경계를 제거하였다.



(a) (b) (c) (d) (e) (f)

그림 3. 실험결과

Ⅲ. 실험결과

본 논문에서 제안하는 시스템이 실제 목표가 되는 스크립트 영상 속의 대상객체와 대상객체의 포즈를 검출, 분리 해 내고, 그 결과를 이용하여 참조영상 속 인물을 대상객체와 같은 포즈의 영상으로 생성한 후 합성하는 결과를 보여주기 위해 인터넷에서 무작위로 다운받은 사진을 이용하여 실험을 진행 하였다. 그 실험결과를 그림 3에서 잘 보여주고 있다. 여기서 (f)는 참조영상이며, (a)~(e) 에서 좌측 영상은 타겟 영상이다. 그리고 중간영상에서는 타겟 영상으로부터 추정된 포즈를 보여주고 있다. 우측 영상은 입력받은 참조영상을 타겟 영상에서와 같은 포즈로 변화시키고 합성을 진행 한 결과를 보여준다. 이로서 본 논문에서는 제안하는 가상스튜디오 시스템이 효과적으로 참조영상을 원하는 포즈의 이미지로 변화시켜 새로운 스튜디오 사진을 생성 할 수 있다는 것을 실험을 통해 확인하였다.

Ⅳ. 결론

본 논문에서는 본인이 원하는 스크립트 영상과 자신의 영상 만 입력하면, 스크립트 영상을 본인이 참여한 영상으로 바꿔주는 가상 스튜디오 시스템을 제안하였다. 입력된 스크립트 영상은 우선 영상분할을 통해 배경과 목표객체 영상으로 분리가 된다. 다음 포즈추정 알고리즘을 통해 객체영상의 포즈를 추출 한 후 생성모델을 이용하여 레퍼런스 영상을 목표 포즈영상으로 바꾸어 준다. 마지막으로 생성된 새로운 포즈영상을 분리된 배경영상과 다시 합성하여 원하는 영상을 얻을 수 있다. 본 논문에서는 실험을 통해 생성된 영상을 보여 주었다. 이런 가상 스튜디오 시스템은 비 접촉식으로 작동할 뿐만 아니라 원하는 영상을 생성하기 위한 비용을 감소시킬 수 있어 후 팬데믹 시대의 광범위한 응용을 기대한다. 다만 좀 더 고품질의 영상을 생성하기 위해서는 생성모델을 통해서 생성된 영상이 교체를 하게 될 목표객체의 조명 상황도 반영해야 하지만 이는 추후의 연구에서 계속하여 보완해 나가려고 한다.

참 고 문 헌

[1] Bolya, Daniel, et al. "Yolact: Real-time instance segmentation." Proceedings of the IEEE/ CVF international conference on computer vision. 2019.

[2] Fang, Hao-Shu, et al. "Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time." IEEE Transactions on Pattern Analysis and Machine Intelligence 45.6 (2022): 7157-7173.

[3] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[4] Hu, Li. "Animate anyone: Consistent and controllable image-to-video synthesis for character animation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

[5] Jiang, Peiyuan, et al. "A Review of Yolo algorithm developments." Procedia computer science 199 (2022): 1066-1073.

[6] Shibasaki, et al. "PIDM: Personality-Aware Interaction Diffusion Model for Gesture Generation." International Conference on Artificial Neural Networks. Cham: Springer Nature Switzerland, 2024.