끝단 학습을 통한 영상 내 사람 영역 검출 및 소거

김은서*, 구윤회*, 정윤영*, 백승렬*
울산과학기술원*

End-to-End Image-based Human Detection and Inpainting

Eunseo Kim*, Yunhoe Ku*, Yunyoung Jeong*, Seungryul Baek*
Ulsan National Institute of Science and Technology*

Abstract - CCTV, 웹캠 등의 보급으로 일상생활에서 수집되는 이미지 및 비디오 데이터의 양은 많아지는데 반해, 이에 노출되는 개인의 프라이버시는 잘 지켜지지 않고 있다. 본 논문에서는 딥러닝 기술을 활용해 이미지에서 사람을 검출하고 지우고 이를 자연스럽게 소거하는 방법을 제시하였다. 제안된 방법론은 입력부터 출력까지 미분가능하도록 구현되었으며, GPU를 활용하여 20 fps 이상의 속도가 나옴을 확인하였다.

1. 서 론

정보화 시대에서 수집되는 이미지 및 비디오 데이터의 양은 빠른 속도로 늘어나고 있다. 하지만 이 데이터에 노출되는 개개인의 프라이버시는 제대로 지켜지지 않고 있으며, 이는 사회적으로 중요한 문제 중 하나로 대두되고 있다. 우리는 이미지에서 사람이 포함된 영역을 지우고 이를 채워주는 모델을 제시해 개인의 프라이버시가 보장되게 하는 기술을 개발하고자 한다. 제안된 모델은 딥러닝 기술을 활용하였으며, 모델을 크게 (1) 사람 검출 네트워크와 (2) 사람 소거 네트워크로 나누었다.

사람을 검출하는 Instance segmentation은 이미지에 등장하는 사람 instance들을 식별하고, 각 instance에 해당하는 영역에 mask를 검출하는 과정을 뜻한다. 이미지 내에서 사람 객체를 mask 형태로 식별하는 것을 첫번째 네트워크의 목표로 잡았다.

두번째 네트워크인 사람 소거 네트워크는 이미지에서 첫 번째 네트워크에서 검출된 mask에 해당되는 픽셀들을 지우고, 지워진 영역을 주변 배경과 잘 동화되는 픽셀값으로 채우는 과정이다. 없는 정보를 생성해 내는 image generation 과정이기 때문에, 주변 배경이 많아야 이를 문맥 정보로 활용하여 이미지를 잘 생성해 낼 수 있다.

사람 검출과 소거 네트워크는 각각 YOLACT [1]와 CR-Fill [2] 논문에서 활용된 네트워크 구조를 활용하였다.

2. 본 론

2.1 관련 연구

2.1.1 Instance segmentation

Tian. et al [3]이 제시한 논문은 instance segmentation과 가장 비슷한 분야인 semantic segmentation에서 잘 사용되고 있는 fully convolutional network가 instance segmentation에 잘 적용되지 못함에 의문을 제기한다. 이는 기존의 region of interest (ROI) based method들이 instance segmentation에 중요한 요소 중 location information을 명시적으로 인코딩하지 못 했기 때문이며, 이를 해결하기 위해 instance-sensitive convolution filter를 사용하여 instance를 잘 반영하여 인식하도록 해서 이러한 문제를 완화하였다.

BlendMask [4]는 기존에 널리 사용되었던 FCOS (Fully Convolutional One-Stage Object Detection) 모델을 이용하는데, 1x1 convolution을 이용하여 box간의 attention mask를 학습하게 하고 FPN(Feature Pyramid Network)의 input인 feature를 사용하여 bottom module은 객체의 bases, 즉 객체의 boundary와 같은 핵심적인 정보를 학습하도록 한 후 이를 attention mask와 선형적으로 결합시켜 attention map을 학습하도록 한다.

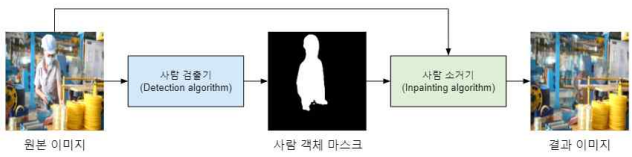
2.1.2 Image inpainting

DeepFillv2 [5]은 이전의 모델들에서 성공적인 결과를 보여준 partial convolution layer의 channel들이 동일한 mask를 공유하고 있어서 complex한 mask를 표현하기 어려움과 동시에 user-guided input을 사용할 수 없음을 해결하고자 노력했다. 저자들은 이를 각 채널 및 각 공간 위치에 대한 특성을 학습할 수 있는 gated convolution을 제안함과 동시에 SN-PatchGAN을 도입하여 기존의 모델들이 가지고 있었던 문제들을 해결할 수 있었다.

Hi-Fill [6]은 generator에 coarse-to-fine manner를 사용하여 high-resolution image에도 image inpainting이 잘 되도록 모델을 구성했다. Generator는 image를 down-sampling하여 low-resolution으로 만든 image에 mask를 적용하여 coarse network에서 inpainting result를 만든다. 이후 결과물을 다시 up-sampling하여 masking하기 전의 image에 inpainting area를 덮어 씌우고 이를 refine network에 집어넣어 inpainting area의 quality를 향상시킴과 동시에 convolution을 거친 feature를 이용하여 attention map을 형성한다. 이는 high-resolution image에서 계산한 contextual residual과 함께 attention transfer module로 들어가 각 scale에 따른 feature aggregation을 구성하게 되고 최종적으로 해당 정보는 generator에서 나온 결과물과 합쳐져 high-resolution inpainting result를 형성할 수 있게 된다.

2.2. 사람 검출 및 소거 모델

우리는 딥러닝 기반의 사람 검출 기술과 소거 기술을 이용하여 실제 환경에서의 데이터를 기반으로 다양한 크기와 자세의 사람을 검출해내고 소거하는 기술을 제안한다. 우리 모델의 전체 개요는 그림 1과 같다.

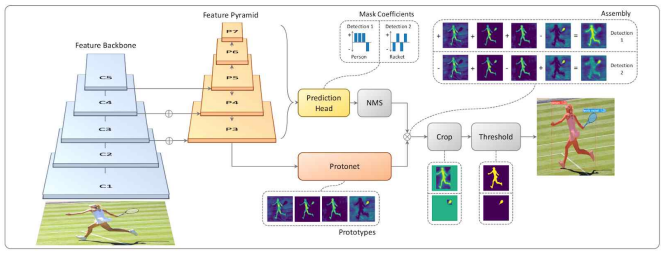


<그림 1> 모델 개요

모델에 사람이 포함된 이미지가 입력으로 들어가게 되면 사람 검출기와 사람 소거기를 거쳐 사람이 제거된 이미지를 얻을 수 있다. 첫째, 사람 검출기에서는 사람이 있는 특정 영역을 파악하여 객체의 마스크 이미지를 획득한다. 그 후, 이전 단계에서 획득한 마스크와 원본 이미지를 사용하여 사람 소거기에서 사람 영역을 제거하고 해당 영역의 픽셀값을 자연스럽게 생성한 결과 이미지를 생성한다.

2.2.1. 사람 검출기

우리의 사람 검출기 알고리즘은 실시간으로 객체 분할을 수행하는 fully-convolutional 모델인 YOLACT [1]를 기반으로 하였다. YOLACT [1]의 전체 구조는 그림 2과 같으며, 명시적인 특징 위치 추정 단계 (explicit feature localization step) 없이 기존의 단일 단계 객체 검출 모델에 마스크 분기를 추가하는 방식을 사용하고 있다.



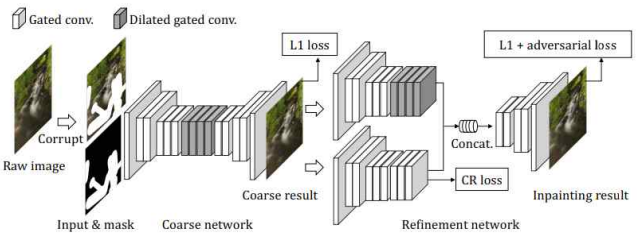
<그림 2> YOLACT의 구조

해당 모델에서는 객체 분할 작업을 최종적인 마스크를 형성하기 위한 두 가지의 간단하고 병렬적인 과정으로 나누었다. 첫번째 분기는 FCN (Fully convolutional networks) [7] 인 Protonet을 사용하여 어떤 하나의 객체에 의존하지 않는 이미지 크기의 프로토타입 마스크 (prototype masks) 세트를 생성한다. 두번째는 객체 탐지 분기에 추가적인 head를 추가 하여 각 앵커에 대한 마스크 계수 (mask coefficients) 의 벡터를 예측한다. 최종적으로, NMS (Non-Maximum Suppression) 에서 살아남은 각 객체에 대해 앞에 언급된 두 분기의 결과물을 선형적으로 결합하여 해당 객체에 대한 마스크를 구성한다.

우리의 사람 검출기 모델은 이러한 구조를 통해 사람 클래스에 대한 검출과 객체 분할을 수행하여 원본 이미지에서 사람 영역이 흰색, 배경 영역이 검정색으로 이루어진 마스크 이미지를 획득한다.

2.2.2. 사람 소거기

사람 소거기에 사용된 알고리즘은 생성기 및 구별기를 포함하는 적대적 생성 신경망 방식의 이미지 인페인팅 (Image inpainting) 모델인 CR-Fill [2]을 기반으로 하였다. 해당 모델의 생성기 구조는 그림 3과 같다.



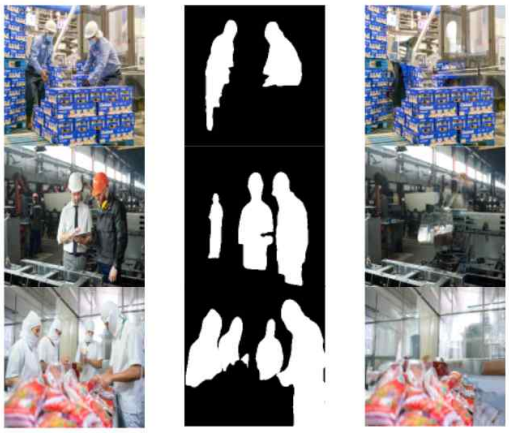
<그림 3> CR-Fill의 생성기 구조

생성기는 두 가지 네트워크로 구성되어 있는데 합성곱 인코더-디코더 형태인 거친 네트워크 (Coarse network) 와 개선 네트워크 (Refinement network) 이다. 거친 네트워크는 원본 이미지에서 손상된 픽셀이 0으로 설정된 이미지와 손상 영역을 나타내는 마스크를 입력으로 받는다. 개선 네트워크에서는 거친 네트워크에서 생성된 거친 결과물 (Coarse result) 의 세부적인 부분을 보완하여 최종적인 인페인팅 결과를 출력한다. 생성기 네트워크는 DeepFillv2 [5]와 유사한 구조이지만, CA 층 (Contextual Attention layer) 이 제거되었고 대신 CR 손실함수 (Contextual Reconstruction Loss)가 적용되었다.

우리의 사람 소거기 모델은 사람 영역을 소거하기 위해 생성기의 입력으로 사람 검출기 모델에서 추출된 마스크와 사람이 포함된 원본 이미지를 사용한다. 모델의 생성기는 사람 영역이 손실된 부분을 채워넣어서 원본 이미지에서 사람이 제거된 고해상도의 이미지를 결과로 출력한다.

2.3. 결과 및 평가

YOLACT 모델이 학습된 COCO dataset를 활용하여 제안된 사람 검출 및 소거 네트워크가 학습되었으며, 그림 4는 학습된 모델을 사람 검출 및 소거 기술을 실제 환경에서의 이미지에 적용해본 결과이다. 공장과 같은 복잡한 배경에서도 높은 품질의 사람 객체 검출 마스크와 소거 결과 이미지를 얻을 수 있었다. 또한 다양한 크기와 자세의 사람이 포함된 이미지에서 사람을 정확히 검출하였고, 마스킹된 영역이 주변 환경과 조화롭게 인페인팅 된 것을 확인할 수 있었다. GPU 활용하여 20 fps 이상의 속도를 확보하였다.



(a) 원본 이미지 (b) 사람 객체 마스크 (c) 결과 이미지

<그림 4> 사람 검출 및 소거 모델의 결과

3. 결 론

사람 검출 및 소거 네트워크를 결합하여 영상내 사람을 소거하는 기술을 개발함으로써 개인의 프라이버시를 보장하는 기술을 개발하였다. CCTV, 웹캠 등 다양한 영상에서 개인의 프라이버시를 보장할 수 있는 응용분야에 활용될 것으로 기대된다.

감사의 글

본 연구는 울산과학기술원의 CCTV 내 현장 작업자의 고성능, 실시간 검출 및 소거 기술 연구 과제 (2.210894.01)의 지원을 받아 수행되었습니다.

[참 고 문 헌]

[1] Bolya, Daniel, et al., "YOLACT: Real-Time Instance Segmentation", In ICCV, pp. 9157-9166, 2019

[2] Zeng, Yu, et al, "CR-Fill: Generative Image Inpainting With Auxiliary Contextual Reconstruction", In ICCV, pp. 14164-14173, 2021

[3] Tian, Z., Shen, C., & Chen, H., "Conditional convolutions for instance segmentation", In ECCV, pp. 282-298, 2020

[4] Chen, Hao, et al., "BlendMask: Top-Down Meets Bottom-Up for Instance Segmentation", In CVPR, pp. 8573-8581, 2020

[5] Yu, Jiahui, et al., "Free-form image inpainting with gated convolution", In ICCV, pp. 4471-4480, 2019

[6] YI, Zili, et al., "Contextual residual aggregation for ultra high-resolution image inpainting", In CVPR, pp. 7508-7517, 2020

[7] LIN, Tsung-Yi, et al., "Feature pyramid networks for object detection", In CVPR, pp. 2117-2125, 2017