단일 단계 검출 방법을 위한 이미지 합성기반 학습 데이터 증강에 관한 연구

이선경*,**, 정치윤*, 문경덕*, 김채규**
*한국전자통신연구원 인공지능연구소 휴먼증강연구실
**부경대학교 IT융합응용공학부
ssunkyung00@gmail.com, iamready@etri.re.kr, kdmoon@etri.re.kr, kyu0707@pknu.ac.kr

A Study on Synthesizing Training Data for One-stage Object Detector

Seon-Gyeong Lee*,**, Chi Yoon Jeong*, KyeongDeok Moon*, Chae-Kyu Kim**
*Human Enhancement & Assistive Technology Research Section, Artificial Intelligence Research Lab, ETRI
**Dept. of IT Convergence & Application Engineering, Bukyong University

요 약

딥러닝 기반의 영상 분석 방법들은 많은 양의 학습 데이터가 필요하며, 학습 데이터 구축에는 많은 시간과 노력이 소요된다. 특히 객체 검출 분야의 경우 영상 내 객체의 위치, 크기, 범주 등의 정보가 모두 필요하여 학습 데이터 구축에 더 많은 어려움이 있으며, 이를 해결하기 위해 최근 이미지 합성기반 데이터 증강에 관한 연구가 활발히 진행되고 있다. 이미지 합성기반 데이터 증강 방법은 배경 영상에 객체를 합성할 때 객체와 배경 영상이 접한 영역에서 아티팩트(Artifact)가 발생하며, 이는 객체 검출 모델이 아티팩트를 객체의 특징으로 모델링하여 검출 성능이 저하되는 원인이 된다. 이러한 문제를 해결하기 위하여 본 논문에서는 양방향 필터 기반의 이미지 합성 방법을 제안하고, 단일 단계 검출의 대표적인 방법인 RetinaNet을 이용하여 이미지 합성기반 데이터 증강 방법의 성능을 분석하였다. 공개 데이터셋에 대한 실험 결과 본 논문에서 사용한 단일 검출 방법 및 데이터 증강 기법을 사용하면 더 적은 양의 증강 데이터로 기존 방법과 동일한 성능을 보여주는 것을 확인하였다.

1. 서론[1)]

디지털 기기들의 발전으로 각양각색의 개인 데이터들이 대량으로 증가하고 있으며, 이를 분석하기 위하여 딥러닝 기반 영상 분석에 관한 연구가 활발히 진행되고 있다. 딥러닝 기반의 영상 분석 방법들은 많은 양의 학습 데이터가 필요하며, 학습 데이터 구축에는 많은 시간과 노력이 소요된다. 특히 영상에 존재하는 객체를 검출하는 객체 검출 분야의 경우 영상 내 객체의 위치, 크기, 범주 등의 정보가 모두 필요하여 학습 데이터 구축에 더 많은 시간과 노력이 필요하다.

이러한 어려움을 해결하기 위해 객체 검출 분야에서 이미지 합성기반의 데이터 증강에 관한 연구가 진행되고 있다[1-3]. 이미지 합성기반의 데이터 증강 기법은 배경 영상에 원하는 객체를 특정한 위치에 배치하고 합성하여 학습 데이터를 증강시켜 객체 검출 방법의 성능을 향상시키는 기법이다. 기존 연구에서는 배경 영상의 레이아웃을 분석하여 실제와 유사한 영상이 생성되도록 객체를 배치하는 방법이 제안되었다[2]. 이렇게 생성된 합성 영상을 학습데이터로 활용하면 SSD(Single shot multibox detector)와 Faster R-CNN 등의 객체 검출 방법의 성능이 향상된다는 것을 보여주었다. 그러나 배경 영상의 레이아웃을 분석하는 방법은 새로운 형태의 배경 영상에 적용하기 어렵고, 배경 영상에 객체를 합성할 때 객체와 배경 영상이 접한 영역에서 아티팩트(Artifact)가 발생하는 문제가 있다. 또한, 객체 검출 알고리즘이 아티팩트를 객체의 특징으로 모델링할 수 있기 때문에 객체 검출 성능이 저하될 수 있다. 이러한 문제를 개선하려는 노력으로, 동일한 객체를 배경 영상의 같은 위치에 가우시안 필터(Gaussian filter), 포아송 블렌딩(Poisson blending) 등의 다양한 블렌딩 방법을 적용하여 합성하고 이렇게 생성된 다양한 영상을 학습 데이터로 활용하는 방법이 제안되었다[3]. 이러한 방법은 배경과 객체가 합성된 영역에 다양한 블렌딩 방법이 적용된 학습 데이터를 사용함으로써 아티팩트 효과를 줄이고 객체 검출 방법의 성

* 본 연구는 한국전자통신연구원(ETRI) 연구운영비지원 사업의 일환으로 수행되었음[20ZS1200, 인간의 감각·지각 능력을 증강하는 다중 감각 융합 기술 개발 사업]

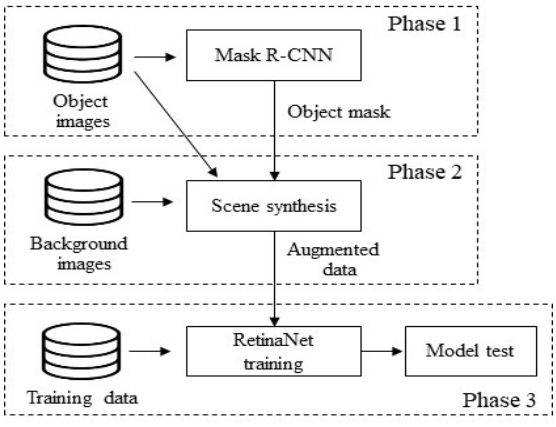
능 향상 시킬 수 있다. 하지만, 이 방법의 경우 블렌딩 방법마다 학습 데이터가 생성되므로 전체 학습 데이터의 양이 증가하여 객체 검출 모델의 학습에 많은 시간이 소요되는 문제점이 있다.

본 논문에서는 아티팩트 효과를 줄이면서도 학습에 필요한 데이터의 양을 줄 일 수 있는 이미지 합성기반 데이터 증강 방법을 제안하였다. 먼저, 본 논문에서는 배경 영상에 객체 영상을 합성할 때 발생하는 아티팩트 효과를 줄이기 위하여 양방향 필터(Bilateral filter) 기반의 이미지 합성 데이터 생성 방법을 제안하였다. 또한, 단일 단계 검출의 대표적 방법인 RetinaNet[4]을 사용하여 이미지 합성기반 데이터 증강 방법의 성능을 분석하였다. 본 논문에서 제안한 방법을 공개 데이터셋을 사용하여 실험한 결과 기존 방법보다 더 적은 양의 증강 데이터를 사용하여 기존 방법과 동일한 성능을 가지는 것을 확인하였다.

본 논문은 2장에서 이미지 합성기반 객체 검출 성능 향상 방법을 제안하여 검증하고 3장에서는 실험 결과를 종합하였으며 4장에서 결론을 서술하였다.

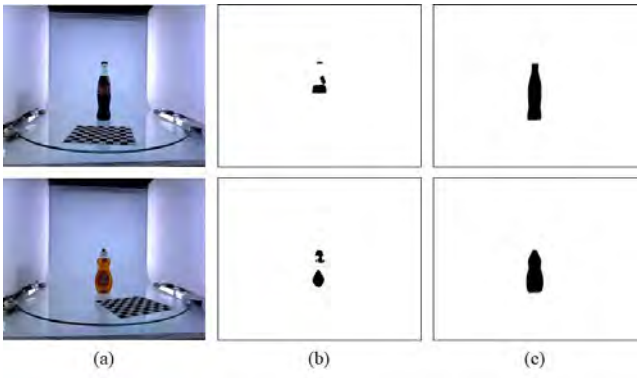
2. 이미지 합성기반 객체 검출 성능 향상 방법

본 논문에서 제안하는 이미지 합성기반 객체 검출 성능 향상 방법은 그림 1과 같이 3단계로 구성된다. 첫 번째 단계에서는 배경 영상에 합성할 객체 영상을 생성하는 단계로써, 객체가 포함된 영상에서 해당 객체의 영역으로만 구성된 마스크 영상을 생성한다. 두 번째 단계에서는 배경 영상에 랜덤하게 객체의 위치를 설정한 후 해당 위치에 객체 영상을 합성함으로써 객체의 범주 및 위치에 대한 학습용 태깅 데이터를 생성한다. 마지막 단계에서는 실제 데이터 및 합성 데이터를 사용하여 객체 검출 방법 및 이미지 합성 방법의 성능을 분석한다.



(그림 1) 이미지 합성기반 객체 검출 성능 향상 방법 흐름도

첫 번째 단계인 객체 영역 추출 단계에서는 Big Berkeley Instance Recognition Dataset (BigBIRD) 데이터셋을 사용하였다. BigBIRD 데이터셋은 125개의 객체를 카메라의 각도 및 위치를 달리하여 촬영한 영상으로 구성되어 있으며, 본 논문에서는 GMU-Kitchen 데이터셋[6]에 존재하는 11개의 객체 데이터를 사용하였다. BigBIRD 데이터셋은 객체 영상마다 객체 마스크 영상이 존재하지만, 투명한 재질로 이뤄진 객체의 경우 마스크 영상이 부정확하게 검출되는 문제점이 존재한다. 따라서 객체 마스크를 정확하게 추출하기 위하여 Mask R-CNN[7] 알고리즘을 사용하였다. Mask R-CNN은 Faster R-CNN이 검출한 객체 영역에서 각 픽셀이 객체에 해당되는지를 판단하는 Fully Convolutional Network을 결합한 알고리즘이다. 그림 2는 Mask R-CNN을 사용하여 객체 영역을 검출한 결과를 나타낸다. 유리병과 같이 투명한 재질의 물체는 그림 2의 (b)와 같이 원본 마스크 영상이 부정확하게 나타나지만, Mask R-CNN을 활용하면 그림 2의 (c)와 같이 개체 영역을 정확하게 검출함을 확인할 수 있다.

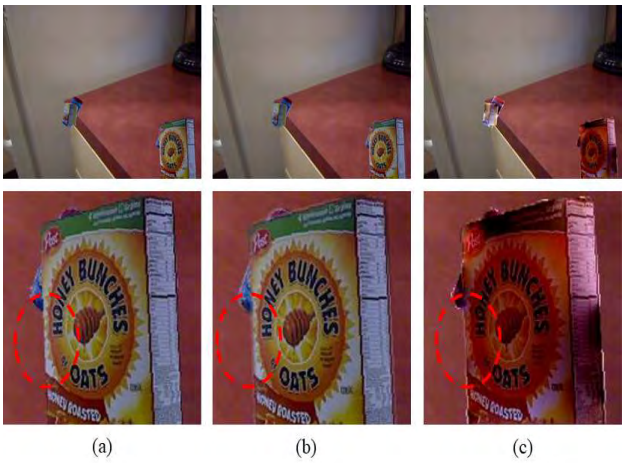


(그림 2) 객체 영역 추출 결과 (a) 입력 영상 (b) 원본 마스크 영상 (c) Mask R-CNN 결과 영상

영상 합성 단계에서는 배경 영상과 객체 영상을 합성하여 객체 검출 모델의 학습 데이터를 생성한다. 영상 합성을 위한 배경 영상은 UW Scenes 데이터셋[8]을 사용하였다. 배경 영상에 객체 영상을 합성하게 되면 객체 영상의 경계에 아티팩트가 생성되며, 이는 객체 검출 방법의 성능을 저하하는 원인이 된다. 따라서 기존 연구[3]에서는 아티팩트로 인한 성능저하를 해결하기 위하여 배경 영상의 같은 위치에 동일 객체를 배치하고 다양한 블렌딩 방법을 적용한 다수의 영상을 생성하여 객체 검출 알고리즘

을 학습시키는 방법을 사용하였다. 다양한 블렌딩 방법을 적용하면 배경 영상과 객체 영역이 만나는 영역의 정보가 비일관성을 갖게 되어 객체 검출 알고리즘 학습 과정에서 아티팩트로 인한 성능저하를 해결할 수 있지만, 학습 데이터의 양이 증가하게 되어 네트워크 모델 학습에 많은 시간이 소요되는 단점이 있다.

본 논문에서는 다양한 블렌딩 방법을 적용하지 않고 양방향 필터 기반의 영상 합성 방법을 제안한다. 기존 연구에서 사용한 가우시안 필터의 경우 경계의 모든 영역을 평활화(Smoothing)하여 객체의 윤곽선도 평활화되는 문제점이 존재하였다. 객체 검출 알고리즘에서 윤곽선 정보는 중요한 특징 중 하나이므로 윤곽선이 평활화되는 경우 객체 검출 알고리즘의 성능이 저하될 수 있다. 양방향 필터[9]는 에지 성분을 보존하면서 다른 영역을 평활화 할 수 있기 때문에 객체 검출 알고리즘의 성능을 향상시킬 수 있을 것으로 기대된다.



(그림 3) 객체와 배경 영상 합성 결과 (a) 양방향 필터 (b) 가우시안 필터 (c) 포아송 블렌딩

그림 3은 다양한 필터를 사용하여 객체를 배경 영상에 합성한 결과를 나타내며, 상단은 합성한 영상의 전체 영역을 나타내고 하단은 객체와 배경의 경계 영역을 나타낸다. 그림 3 하단의 객체 영역을 확대한 영상에서 알 수 있듯이 가우시안 필터를 사용하면 객체의 경계선 영역이 평활화되는 반면, 양방향 필터의 경우 경계선을 유지함을 확인할 수 있다. 그림 3의 (c)는 포아송 블렌딩를 사용한 경우를 나타내며, 경계와 주변 영역이 부드럽게 연결되지 않고 색상의 변화가 발생하는 단점이 있다. 기존 연구[3]에서도 포아송 블렌딩을 사용하는 경우 색상 변화가 발생하는 단점이 있다고 지적하였다.

객체 검출 단계에서 기존 연구들은 Faster R-CNN 과 같은 두 단계 검출 방법을 주로 사용하였다. 두 단계로 수행되는 검출 방법의 경우, 높은 정확도를 가지지만 검출 속도가 느리기 때문에 실제 활용에는 어려움이 있다. SSD로 대표되는 단일 단계 검출 방법은 두 단계 검출 방법과 비교하면 정확도가 낮지만 검출 속도가 빠른 장점이 있다. 단일 단계 검출 방법의 낮은 검출 정확도는 학습 과정에서 클래스 간 불균형으로 인하여 발생하며, 최근 클래스의 불균형 문제를 해결하면서 단일 검출 방법의 성능을 향상시킨 RetinaNet이 발표되었다. RetinaNet은 Focal loss를 사용하여 학습 과정에서 분류하기 쉬운 예제들의 학습 기여도를 낮춤으로써 클래스의 불균형 문제를 해결하였으며, 두 단계 검출 방법보다 높은 검출 성능과 처리 속도를 보여주었다[4]. 따라서, 본 논문에서는 RetinaNet을 사용하여 이미지 합성기반 데이터 증강 방법의 성능을 분석하였다.

3. 실험 결과

이미지 합성기반 데이터 증강 방법의 성능 분석에는 GMU-Kitchen 데이터셋을 사용하였다. GMU-Kitchen 데이터셋은 9개의 영상으로 구성되어 있으며, 각 영상은 11개의 객체 정보를 포함한다. GMU-Kitchen 데이터셋은 3-겹 교차 검증으로 구분되어 있으며, 본 실험에서도 데이터셋에서 정의한 3-겹 교차 검증을 수행하여 성능을 측정하였다.

데이터 증강을 위한 객체 데이터는 BigBIRD 데이터셋 중 GMU-Kitchen 데이터와 중복되는 11개 객체 데이터를 사용하였다. BigBIRD 데이터셋에서 한 객체는 5개의 카메라 위치와 120 개의 다른 각도 촬영된 총 600장의 영상으로 구성되어 있으며, 본 논문에서는 객체별로 2개의 카메라 위치에서 촬영된 240장 영상을 랜덤하게 선택하여 합성에 사용하였다.

영상 합성 단계에서는 1.358장의 배경 영상이 사용되었으며, 성능 비교를 위한 영상 합성 방법은 기존 연구[3]에서 사용한 가우시안 필터, 포아송 블렌딩, 박스 필터와 본 논문에서 제안한 양방향 필터를 사용하였다. 영상 합성을 위해 사용된 방법의 파라미터는 기존 연구에서 사용한 값을 사용하였으며, 본 논문에서 제안한 양방향 필터의 경우 거리는 5, 색상과 공간에 대한 표준편차 값은 25를 사용하였다. 배경 영상에 최소 1개에서 최대 4개의 객체가 랜덤하게 선택되어 배치되며, 동일 영상에 대해서 각기 다른 필터를 적용한 합성 영상이 생성된다.

<표 1> 이미지 합성기반 객체 검출 성능 평가 결과

	coca cola	coffee mate	honey bunches	hunt's sauce	mahatma rice	nature v1	nature v2	palmolive orange	pop secret	pringles bbq	red bull	mAP
No augmentation	82.9	93.7	90.5	85.8	90.5	97.2	86.5	88.9	89.7	89.4	71.6	87.9
No blending	84.7	94.3	91.7	87.5	88.9	96.7	86.6	88.1	90.8	90.1	61.2	88.1
Box blurring	84.5	95.5	92.1	87.5	89.1	96.8	87.3	87.5	88.3	90.6	69.3	88.0
Gaussian blurring	83.9	93.8	82.2	88.0	91.0	96.8	87.7	88.8	90.3	89.9	69.7	88.5
Poisson	84.1	95.0	91.2	87.6	86.5	97.1	87.1	88.5	91.3	89.7	71.0	88.1
Ours	83.9	95.1	91.6	88.7	88.6	97.3	88.6	88.7	91.8	91.2	70.8	88.7
Georgios's [2]	82.6	92.9	91.4	85.5	81.9	95.5	88.6	78.5	93.6	90.2	54.1	85.0
Dwibedi's [3]	88.5	95.5	94.1	88.1	90.3	97.2	91.8	80.1	94.0	92.2	65.4	88.8

객체 검출을 위한 RetinaNet 모델은 백본 네트워크로 ResNet-50을 사용하였으며, 에포크(epochs)는 30으로 설정하고, 반복(iteration) 횟수는 객체 데이터의 수로 설정하였다. 학습 비율은 0.00001을 설정하였으며 네트워크 모델은 Keras 라이브러리를 사용하여 구현하였다. 객체 검출 방법의 성능 측정은 IoU (Intersection over Union) 가 0.5일 때의 mAP (mean Average Precision)을 사용하였다.

표 1은 이미지 합성기반 객체 검출 성능평가 결과를 나타낸다. 실험 결과를 살펴보면 이미지 합성기반 증강 데이터를 사용하여 RetinaNet을 학습하면 증강 데이터를 사용하지 않은 경우에 비하여 성능이 증가하게 되며, 이미지 합성 방법 중 본 논문에서 제안한 양방향 필터를 사용한 합성 방법의 성능 향상이 가장 크게 나타났다. 제안 방법은 배경 영상의 레이아웃을 분석하여 데이터를 증강하는 기존 방법[2]보다 높은 성능을 나타내며, 다양한 블렌딩 방법을 적용한 기존 연구[3]와 동일한 성능을 보여주었다. 기존 연구[3]는 블렌딩을 사용하지 않은 합성 영상과 가우시안 필터 및 포아송 블렌딩을 적용한 합성 영상을 모두 학습 데이터로 사용하므로 본 논문에서 제안된 방법보다 3배나 많은 증강 데이터를 사용하게 된다. 따라서 본 논문에서 제안한 방법은 기존 방법보다 더 적은 양의 증강 데이터로 동일하거나 더 높은 성능을 보여주는 것을 표1의 결과에서 확인할 수 있다.

4. 결론

본 논문에서는 양방향 필터 기반의 이미지 합성 데이터 생성 방법을 제안하고, 단일 단계 검출의 대표적인 방법인 RetinaNet에서의 이미지 합성기반 데이터 증강 방법의 성능을 분석하였다. 제안 방법에서는 Mask R-CNN을 사용하여 객체 영역의 마스크를 추출한 후, 객체와 배경 영상을 합성하는 과정에서 객체의 경계선을 보존하면서도 주변 영역을 평활화하여 영상 합성의 품질을 높일 수 있는 양방향 필터를 적용하였다. 실제 데이터 및 합성 데이터를 사용하여 RetinaNet의 객체 검출 성능을 분석하였다. 공개 데이터셋을 사용한 실험 결과를 통해 본 논문에서 제안한 양방향 필터 기반의 합성 방법이 다른 블렌딩 방법보다 객체 검출 방법의 성능을 더 향상시키는 것을 확인하였다. 또한, 본 논문에서 사용한 단일 검출 방법 및 이미지 데이터 증강 기법을 사용하면 기존 방법보다 더 적은 양의 증강 데이터로 동일한 성능을 보여주는 것을 실험 결과로 확인하였다.

참고문헌

[1] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," *Proc. CVPR*, 2016.

[2] G. Georgakis, A. Mousavian, A. C. Berg, and J. Kosecka, "Synthesizing training data for object detection in indoor scenes," *arXiv preprint arXiv:1702.07836*, 2017.

[3] D. Dwibedi, I. Misra, and M. Hebert, "Cut, paste and learn: Surprisingly easy synthesis for instance detection," *Proc. ICCV*, 2017.

[4] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *Proc. ICCV*, 2017.

[5] A. Singh, J. Sha, K. Narayan, T. Achim, and P. Abbeel, "BigBIRD: A large-scale 3D database of object instances," *Proc. ICRA*, 2014.

[6] G. Georgakis, Md. Reza, A. Mousavian, P. Le, and J. Kosecka, "Multiview. RGB-D Dataset for Object Instance Detection," *arXiv preprint arXiv*

:1609.07826, 2016.
[7] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," *Proc. ICCV*, 2018.
[8] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view rgb-d object dataset," *Proc. ICRA*, 2011.
[9] C. Tomasi and R. Maduchi, "Bilateral filtering for gray and color images," *Proc. ICCV*, 1998.