U-Net 구조를 이용한 이미지에서의 보행자 분할

김승택*, 이효종*,°
*전북대학교 컴퓨터공학부
*kis7279@naver.com, **hlee@chonbuk.ac.kr, ◦교신저자

Pedestrian Segmentation Using U-Net

Seung Taek Kim*, Hyo Jong Lee*
*Division of Computer Science and Engineering, Chonbuk National University
Chonbuk National University

요 약

자율주행 자동차에서의 보행자 인식 및 사람의 행동 인식과 같은 분야 등에 대한 연구들이 활발하게 진행되고 그에 기반을 둔 기술들이 많이 개발되고 있다. 그리고 대부분의 연구에서는 사람에 대한 경계 박스를 검출한다. 영상에서 사람의 유무 혹은 위치를 판단하는 문제에서는 경계 박스만을 검출하는 것이 효율적일 수 있으나 경계 박스는 행동 인식과 같은 분야에 사용하기에는 많은 정보의 손실이 발생할 수 있다. 본 논문에서는 U-NET 구조의 딥러닝 모델을 사용해 경계 박스로 인한 정보 손실을 줄일 수 있는 보행자 분할 방법을 제안한다. 모델의 학습을 위해 2017 COCO 데이터셋의 사람 카테고리를 사용하였으며 Penn-Fudan 보행자 데이터셋을 이용하여 제안 방법을 테스트하였으며 기존의 방법들과 비교하여 의미 있는 결과를 얻었다.

1. 서론

최근 몇 년간 자율주행 및 행동인식과 같은 분야에 대한 관심이 높아짐에 따라 이러한 기술의 기반이 되는 많은 기술들이 큰 발전을 이룩하였다. 특히 보행자 검출 연구는 자율주행 자동차, 인공지능 CCTV, 장면 자막 생성, 행동인식 등 많은 분야에서 연구 되고 있으며 큰 성공을 거두었다. 그러나 기존의 보행자 인식을 위한 대부분의 연구들은 보행자의 경계 박스를 검출하는데 그치고 있다[1, 2, 3]. 경계 박스 검출 방법은 객체의 유무를 판단하는 문제에 있어서 효율적 일수 있으나 많은 노이즈를 포함하고 있거나 반대로 필요한 정보의 손실을 야기할 수 있다. 이러한 경계박스의 문제를 해결하기 위하여 최근에는 CNN(Convolutional Neural Networks)를 이용한 픽셀 단위의 객체 분할 연구가 진행되고 있다[4, 5, 6].

그림 1. FCN의 보행자 검출 결과

FCN(Fully Convolutional Networks)[7]은 위의 연구들에서 픽셀단위의 예측을 위해 CNN의 하단에서 쓰이며 객체 분할에서 정확성을 향상 시킨다[7]. 그러나 FCN은 그림 1.과 같이 국부적인 밀집된 예측을 야기하며, 객체의 전체 구조와 일치하지 않는 문제를 나타내는 경우도 있다.

최근에는 이러한 같은 동일 객체의 픽셀간 불일치를 완화하기 위하여 CRF(Conditional Random Field)를 이용한 연구가 진행되었다[8, 9, 10, 11, 12]. CRF를 이용하여 잘못된 픽셀 예측을 개선할 수 있지만 CRF는 FCN의 결과를 이용하는 후처리 기법이므로 FCN에 의한 예측 결과에 크게 의존한다. 그러므로 CRF를 이용하여 회복할 수 없을 만큼 큰 잘못된 예측이 FCN에서 발생하면 여전히 나쁜 결과를 보인다. 또한 CRF는 지역적인 정보만을 이용하여 객체를 분할하므로 사람의 다리와 같은 국부적인 부분의 모호성을 피하기 어렵다[8, 9, 10, 11, 12].

우리는 이러한 FCN이 야기하는 문제를 해결하기 위하여 FCL(Fully Connected Layer)이 없는 U-Net[13] 구조를 이용한 모델을 제안한다. 제안된 방법은 이미지로부터 고수준의 전역적인 특징을 추출하여 FCN-CRF에서 발생하는 국부적인 부분의 모호성을 피할 수 있다.

우리는 보행자의 특징을 추출하기 위하여 2017 COCO 데이터셋[14]의 사람 카테고리의 일부 이미지 7000장을 학습(Training)과 검증(Validation)에 사용하였다. 또한 모델의 성능을 테스트하기 위하여 Penn-Fudan 데이터셋[15]을 이용하였다.

2. 제안 방법

2.1. 모델 구조

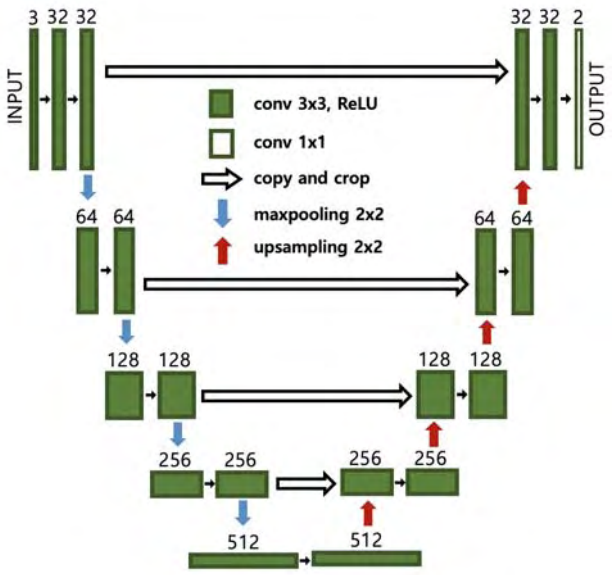

그림 2. 픽셀 단위의 보행자 분할을 위한 제안된 U-Net 구조.

픽셀 단위의 보행자 분할을 위한 제안된 모델의 구조는 그림 2.와 같다. 본 모델은 기존의 U-Net과 같이 다운샘플링을 위한 부분과 업샘플링 부분으로 구성되었으며 서로 다른 3개의 컨볼루션 레이어에서 특징맵을 연결해 공유한다. 본 모델은 256x256x3 크기의 컬러 이미지를 입력으로 받도록 설계되어 있다. 각 컨폴루션 레이어는 활성함수로 ReLU를 사용하며, 정규화를 위하여 He 정규화[16]를 사용한다. He 정규화는 0 이하의 값을 제거하는 ReLU로 인하여 분산이 줄어드는 것을 방지한다.

2.2. 데이터 및 학습

학습 및 검증을 위해 2017 COCO 데이터셋의 사람 카테고리의 이미지 7000장을 사용하였다. 또한 픽셀 단위의 분할을 위해 모델의 입력으로 마스크 이미지를 픽셀 단위로 레이블링하여 이미지와 함께 주어준다. 배치 사이즈(Batch-Size)는 32, 모델 최적화 기법으로는 Adam 이용하여 0.0001의 속도로 학습 시켰으며 Binary Cross Entropy를 손실함수로 사용하였다. 학습을 위해 NVIDIA GTX 1080ti 그래픽카드 2개를 이용하여 전체 학습 데이터셋에 대한 학습을 150번 반복(150 epoch)하였다.

또한 보행자 분할에 대한 성능을 테스트하기 위하여 Penn-Fudan 보행자 데이터셋을 이용하였다. Penn-Fudan 데이터셋은 170장의 보행자 이미지-마스크 쌍을 갖는다.[15] 성능 비교를 위하여 기존의 방법인 FCN, CRF, CHOPPs[17], MMBM1[18] 및 MMBM2[18]과 비교하였으며 기존의 방법에 비해 본 연구에서 제안한 방법이 좋은 성능을 보임을 입증하였다.

3. 결과

<표 1> Penn-Fudan 데이터셋 테스트 결과

Method	IoU
CRF	68.35
CHOPPs	71.33
MMBM1 (case4)	76.92
MMBM2	77.30
MMBM1 (case4) + GraphCut	77.97
MMBM2 (case4) + GraphCut	79.42
Proposed Method	**80.91**

위의 <표 1>은 Penn-Fudan 데이터셋의 기존의 방법들과 제안된 방법에 대한 테스트 결과이다. 성능 지표로는 IoU(Intersection over Union)를 사용하였으며 본 논문에서 제안된 방법은 80.91% IoU를 얻었다. 이는 최신의 연구 결과인 'MMBM2 + GraphCut'[18]와 비교하여 약 +1.49% IoU의 향상된 결과를 보임으로서 본 논문에서 제안된 방법의 성능이 보행자를 픽셀 수준에서 분할해 검출함에 있어 유의미한 결과를 보여줌을 입증한다.

그림 3. 제안 방법의 Penn-Fudan 데이터셋에 대한 예측 결과. (a)입력 이미지, (b)Ground-Truth, (c)예측결과

그림 3. 제안 방법의 Penn-Fudan 데이터셋에 대한 결과를 보면 보행자를 올바르게 분할하는 것을 확인할 수 있다. 그림 3.의 (a)는 입력 데이터, (b)는 Ground-Truth 이며 (c)는 입력데이터에 대한 예측 결과이다. 그림 1. FCN과 단리 보행자의 전역 적인 특징을 잘 찾아내고 있고 다리와 같은 부분에서 국부적인 부분에 대한 모호함을 잘 피하고 있음을 확인할 수 있다.

4. 결론

본 논문에서는 U-Net 구조를 이용한 픽셀 수준의 보행자 분할 방법을 제안하였다. 기존의 방법들은 이미지에서 보행자에 대한 전역적인 특징을 잘 추출 하지 못하며 다

리와 같은 부분의 모호함을 잘 찾아내지 못하지만 제안된 방법을 통해 그러한 기존의 방법에 대한 문제를 해결했음을 확인할 수 있었다. AP지표를 이용하여 기존 최신의 방법인 'MMBM2 + GraphCut'[18]와 비교하여 약 +1.49% 향상된 정확도를 확인할 수 있었다.

후속 연구에서는 세밀하게 보행자를 분할하고 군집 보행자에 대한 분할이 잘 이루어질 수 있도록 하는 방법에 대해 연구할 수 있으며, 또한 의미론적 분할(Semantic Segmentation)이 아닌 객체에 대한 인스턴스 분할(Instance Segmentation) 방법에 대해 고찰해볼 필요가 있다. 이러한 연구들은 장면 인식 혹은 사람의 행동 인식에 대한 연구의 초석이 될 것이라고 생각된다.

사 사

이 논문은 2016년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No.: 2016R1D1A3B03931911)

참고문헌

[1] Dollar, Piotr, et al. "Pedestrian detection: An evaluation of the state of the art." IEEE transactions on pattern analysis and machine intelligence 34.4 (2012): 743-761.

[2] Dollár, Piotr, et al. "Pedestrian detection: A benchmark." (2009): 304-311.

[3] Dollar, Piotr, et al. "Pedestrian detection: An evaluation of the state of the art." IEEE transactions on pattern analysis and machine intelligence 34.4 (2012): 743-761.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, pages 1097-1105, 2012.

[5] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. arXiv preprint:1409.4842, 2014.

[7] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. 2015.

[8] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr. Conditional random fields as recurrent neural networks. In CVPR, 2015.

[9] A. G. Schwing and R. Urtasun. Fully connected deep structured networks. arXiv preprint:1503.02351, 2015.

[10] G. Lin, C. Shen, I. Reid, and A. van den Hengel. Deeply learning the messages in message passing inference. In NIPS, 2015.

[11] S. Ross, D. Munoz, M. Hebert, and J. A. Bagnell. Learning message-passing inference machines for structured prediction. In CVPR, pages 2737-2744, 2011.

[12] G. Lin, C. Shen, I. Reid, et al. Efficient piecewise training of deep structured models for semantic segmentation. arXiv preprint :1504.01013, 2015.

[13] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.

[14] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." European conference on computer vision. Springer, Cham, 2014.

[15] Object Detection Combining Recognition and Segmentation. Liming Wang, Jianbo Shi, Gang Song, I-fan Shen. To Appear in ACCV 2007.

[16] He, Kaiming, et al. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification." Proceedings of the IEEE international conference on computer vision. 2015.

[17] Y. Li, D. Tarlow, and R. Zemel. Exploring compositional high order pattern potentials for structured output learning. In CVPR, pages 49-56, 2013.

[18] J. Yang, S. Safar, and M.-H. Yang. Max-margin boltzmann machines for object segmentation. In CVPR, pages 320-327, 2014.