보행 보조 로봇의 환경 인지를 위한 의미론적 영상 분할 기법에 관한 준비 연구

이서영*, 박지성**, 김강건**
*건국대학교 응용통계학과, **한국과학기술연구원 AI·로봇연구소
cococindy98@gmail.com, jisungpark@kist.re.kr, danny@kist.re.kr

A Preliminary Study on Semantic Segmentation Techniques for Environment Recognition of Walking Assistant Robot

SeoYoung Lee*, JiSung Park**, KangGeon Kim**
*Dept. of Applied Statistics, Konkuk University
**Artificial Intelligence and Robotics Institute, Korea Institute of Science and Technology

요 약

보행 환경 인지 기술은 보행 보조 로봇의 지능화를 위한 핵심 기술 중 하나다. 이 논문은 국내 보행 환경에 대한 보행 보조 로봇의 인지 지능을 고도화하는 방법으로 심층 학습 기반의 의미론적 영상 분할 기법을 고려한다. 이 논문은 국내 보행 환경에 대한 기존 영상 분할 기법의 성능을 비교 분석하고, 국내 보행 환경에 적합한 영상 분할 기술의 개발 방향과 인지 센서의 구성 및 배치에 대해 논한다.

1. 서론

보행 보조 로봇은 고령자를 포함한 보행 장애인의 보행 안전을 위한 효과적인 대안으로 여겨지고 있으며 고령화 인구의 지속적인 증가로 인해 그 중요성이 커지고 있다. 이러한 이유로 보행 보조 로봇은 지금까지 많은 기술적 진보를 이루어 왔으며 최근 인공지능 기술을 이용해 사용자에게 여러 보행 기능과 편의를 제공하는 보행 보조 로봇의 지능에 대한 연구적 관심도 함께 증가하고 있다.

이 논문은 보행 보조 로봇의 지능화를 위한 여러 요소 기술 중에서 보행 환경 내 객체를 구분하는 보행 환경 인지 기술을 고려한다. 이 기술은 여러 방법으로 구체화할 수 있으며 이 중에서 심층 학습 기법을 이용해 영상 내 관심 객체의 영역을 찾는 의미론적 영상 분할 기법이 대표적이다.

기존 영상 분할 기법은 도전적인 영상 분할 문제에서 높은 성능에 도달했고, 보행 장애인의 보행 보조를 위한 응용에서도 높은 유용성을 보였다. 그러나 지금까지 국내 보행 환경에 적합한 보행 환경 인지와 영상 분할 기법에 대한 논의와 연구가 다소 부족했다. 그리고 이것은 국내 보행 보조 로봇의 실제적 활용 측면에서 중요하다.

이 논문은 보행 보조 로봇의 국내 보행 환경 인지 문제에 의미론적 영상 분할 기법을 효과적으로 적용하기 위해 국내 보행 환경에 적합한 영상 분할 기법의 개발 방향과 환경 인지를 위한 센서의 구성 및 배치 전략을 설명한다. 이를 위해, 이 연구는 국내 보행 환경의 특징을 나타내는 영상 자료를 획득하고 이 자료에 대한 기존 영상 분할 기법들의 동작 특성과 성능을 비교 및 분석한다.

2. 관련 연구

지금까지 영상 분할 기법은 도전적인 영상 분할 문제에서 높은 성능을 보여주었다. FCN[1]은 1차원 합성곱을 이용해 망 전체를 합성곱 망으로 구성하고, 직접 연결(skip connections)을 이용해 객체 분할 성능을 개선했다. UNet++[2]은 고밀도 직접 연결(dense skip connections)과 심층 감독(deep supervision)을 이용하여 자료와 객체 크기의 변화에 일정한 성능과 빠른 추론 속도를 달성했다. PSPNet[3]은 피라미드 풀링(Pyramid pooling)을 이용해 전역 참조 기능을 강화했고, DeepLabV3+[4]는 인코더-디코더(encoder-decoder)와 어트로우스 합성곱 (atrous convolution)을 이용해 크고 작은 객체에 대한 분할 성능을 높였다. 이 외에도 영상 분할의 속도를 높인 ERFNet[5]이 있다.

영상 분할 기법을 이용한 응용은 보행 안내 시스템을 위한 보행로 탐지에 ERFNet과 PSPNet을 결합

(그림 1) 학습을 위한 영상 자료의 수집: 자료 수집 장소의 위성 사진과 이동 경로 (왼쪽), 수집된 영상 자료 (오른쪽)

해 영상 분할 기법, ERF-PSPNet[6]을 제안한 연구가 있으며, 시각 장애인의 보행을 보조하기 위해 DenseNet[7]과 FPN[8]을 결합한 영상 분할 기법을 제안한 연구가 있었다. 앞선 연구에서 보행 안내 및 보조를 위한 시스템은 시스템 경량화 및 시스템 자원을 고려해 RGB 카메라 또는 깊이 카메라를 주로 탑재했다. 센서의 배치는 인지 대상과 목적에 따라 머리, 가슴, 골반, 무릎 등 모든 신체 부위를 활용했으며 원거리 장애물 탐지와 보행로 구분을 위해서 머리 또는 가슴 부위를, 근거리 지형 구분을 위해서 골반 또는 무릎 부위를 이용했다.

3. 국내 보행 환경과 영상 분할 기법

이 장은 국내 보행 환경에 대한 기존 영상 분할 기법의 성능을 실험을 통해 확인한다. 이 실험은 FPN, UNet++, PSPNet, DeepLabV3+의 성능을 여러 백본망(Backbone network)을 적용해 비교한다.

3.1 자료 수집 및 영상 분할 망의 학습

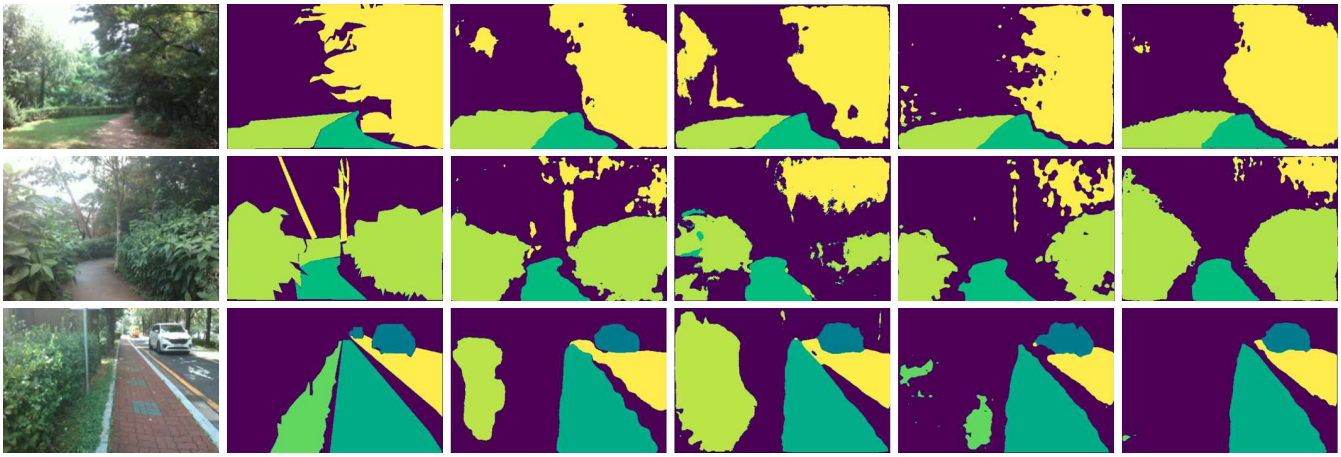
국내 보행 환경에 대한 기존 영상 분할 기법의 성능을 확인하기 위해 국내 보행 환경의 특징을 나타내는 도심 주변의 산에서 RGB 영상을 수집했다(그림 1). 여기서 사용한 카메라는 인텔의 L515 카메라이며 카메라를 가슴에 배치해 지면으로부터 약 1.2 미터 높이에 위치하고, 보행 방향을 향하도록 했다. 영상은 VGA 해상도로 수집했으며 자료 수집을 위한 이동 경로는 포장도로와 등산로 영상을 포함하도록 정했다.

이 연구는 수집한 자료를 이용해 학습이 완료된 망을 전이 학습을 통해 재학습한 뒤 기존 기법의 성능을 확인한다. FPN은 MobileNetV2[9]와 EfficientNet [10] 백본망을 이용해 디코더-인코더 피라미드 채널 등을 달리하여 재학습했다. UNet++은 DenseNet[11], MobileNetV2, EfficientNet를 백본망으로 구성해 재학습했다. 그리고 PSPNet은 ResNet[12]을 백본망으로 채널 수를 달리하여 재학습했다. DeepLabV3+는 ResNet과 EfficientNet 백본망을 이용해 재학습했다. 재학습 과정에서 전체 수집 영상 총 130장 중 76%는 훈련용, 14%는 검증용, 그리고 10%는 시험용으로 사용했다. 분류할 객체의 수는 13 개로 보행 환경을 구성하는 건물, 보행로, 나무, 잔디, 차도, 계단, 보행자, 배경 등을 포함한다.

3.2 영상 분할 결과

그림 2는 각 기법을 이용한 객체 분할 결과의 일부를 나타낸다. 이 결과에서 모든 기법은 보행로 식별에 큰 문제가 없는 수준에서 보행로 및 차도 객체를 분류했고, 객체 경계면에서 정밀함에 차이가 있었으나 크게 유의할만한 차이는 아니었다. 모든 기법이 나무와 관목 객체에 대해 낮은 객체 분할 성능을 보였는데, 이것은 최초 학습을 위해서 사용한 자료나 이 연구의 재학습에 사용한 자료가 해당 객체를 충분히 나타내지 못한 영향일 수 있다. 그리고 해당 객체의 형상이 다소 복잡하고, 다양한 위치 관계를 가지는 것도 하나의 이유일 수 있다. 그리고 모든 기법은 보행자, 차량, 가로등과 같은 장애물에 대해서 인지 정보로 활용할 수 있는 수준의 객체 분할 성능을 보였다.

표 1은 수집 자료에 대한 각 기법의 성능을 mIoU(mean intersection of union)와 추론시간을 이용해 나타낸다. 컴퓨터는 Intel Core i9-9900KF CPU 3.60GHz*16, 개발 환경은 Ubuntu 18.04.5 LTS, 그래픽카드는 NVIDIA GeForce RTX 2080을 사용하였다. 여기서 DeepLabV3+와 FPN을 이용한 영상 분할 성능은 모든 백본망에 대해서 다른 기법에 비해 높은 성능을 보여주었다. 하지만 모든 기법이 도로와 장애물을 제외한 나머지 객체에서 낮은 분할 성능을 가지기 때문에 전체적인 성능은 기존



(그림 2) 보행 영상에 대한 영상 분할 결과: 왼쪽부터 입력 RGB 영상, 영상 분할 정답 레이블, FPN, UNet++, PSPNet, DeepLabV3+을 이용한 영상 분할 추론 결과

<표 1> 영상 분할 기법 성능의 정량적 비교: DPC(decoder-pyramid channel), DSC(decoder-segmentation channel), DMP(decoder-merge policy), DC(decoder channel), PD(psp-dropout), POC(psp-dropout channel), EOS(encoder-output stride), DOS(decoder-output stride), BS(batch size)

영상 분할 기법	백본 망	망 구성	mIoU	추론시간(초)
FPN	MobileNetV2	256-DPC, 128-DSC, add-DMP, 4-BS	0.55	0.06
	EfficientNet-b0	512-DPC, 128-DSC, add-DMP, 8-BS	0.56	0.06
	EfficientNet-b0	512-DPC, 128-DSC, cat-DMP, 4-BS	0.58	0.05
UNet++	DenseNet121	(512,256,128,64,32)-DC, 4-BS	0.52	0.12
	MobileNetV2	(512,256,128,64,32)-DC, 4-BS	0.50	0.07
	EfficientNet-b0	(512,256,128,64,32)-DC, 4-BS	0.53	0.08
PSPNet	ResNet101	0.2-PD, 512-POC, 8-BS	0.47	0.12
	ResNet101	0.2-PD, 256-POC, 4-BS	0.47	0.08
	ResNet101	0.5-PD, 512-POC, 8-BS	0.47	0.12
DeepLabV3+	ResNet101	16-EOS, (6,12,18)-DOS, 4-BS	0.53	0.12
	ResNet101	16-EOS, (6,12,18)-DOS, 8-BS	0.55	0.14
	EfficientNet-b0	16-EOS, (6,12,18)-DOS, 4-BS	0.55	0.06

알고리즘이 보고하는 성능에 비해 다소 낮은 편이다. 이는 아직 재학습에 사용된 국대 보행 환경 데이터 셋이 충분하지 못한 이유일 가능성이 크며, 추후 국내 보행 환경에 맞는 데이터 셋이 충분히 모였을 때의 결과를 비교해볼 필요가 있다.

모든 기법의 추론 속도는 대체로 백본망에 지배적인 영향을 받았다. 모든 기법은 백본망으로 MobileNetV2 또는 EfficientNet을 이용할 경우 빠른 추론 속도를 보인 반면에 DenseNet과 ResNet을 이용한 경우 약 2배에서 3배까지 추론 시간이 증가했다. 보행 보조 로봇의 여러 응용을 감안한다면 전자와 같이 가벼운 네트워크의 사용이 필요하다.

4. 국내 보행 환경에 대한 영상 분할 연구

국내 보행 환경은 도심과 도심 주변의 산과 강에 보행로가 발달해 있고, 인구밀도가 높아 환경 인지 문제에 다소 도전적인 부분이 있다. 국내 보행 환경의 인지 문제에서 기존 영상 분할 기법의 전체적인 성능을 높이는 것도 중요하지만, 이것보다 우선으로 고려해야 하는 몇 가지 문제가 있다. 구체적으로, 국내 보행 환경이 차도와 보행로의 경계가 불분명한 경우가 많아 안전한 보행 공간을 찾아내는 것이 중요하다. 따라서 보행 환경 인지에서 차도로부터 안전한 보행로를 구분하는 것이 필요하다. 한편, 인구밀도가 높은 국내 보행 환경은 장애물이 높은 밀도로 존재하고, 장애물의 위치가 급변하기 때문에 보행 환경 인지는 장애물을 빠르고 정확하게 구분해야 한다. 보통 객체 인지 문제에서 객체 분류의 정확도와 속도는 상호 배치하기 때문에 모든 객체에 대한 분류 성능을 높이기보다는 안전한 보행에 필요한 객체를 선별하고, 해당 객체에 대한 분류 성능을 우선적으로 높이는 전략이 필요하다.

보행 환경 인지를 위한 센서의 구성은 국내 보행 환경이 보행자를 포함한 장애물이 주변에 고밀도로

존재하는 경우가 많기 때문에 높은 신뢰도로 장애물 구분을 위한 고려가 필요하다. 환경 인지를 위한 대표적인 센서인 RGB 카메라와 더불어 깊이 카메라는 근거리 장애물 탐지에 매우 효과적이기 때문에 적극적으로 고려할 필요가 있다. 한편, 국내 보행 환경이 산지의 영향으로 험지와 경사로가 많고, 이러한 환경에서 영상 정보만으로 공간 해석은 도전적일 수 있다. 관성측정장치의 사용은 이러한 문제를 해결할 수 있는 효과적이고 경제적인 시스템으로 적극적으로 활용해야 한다. 센서 배치의 경우, 2장 관련 연구에서 언급한 것처럼 국내 보행 환경에서도 동일하게 원거리 장애물 탐지와 보행로 구분을 위해서는 머리 또는 가슴 부위를, 근거리 지형 구분을 위해서는 골반 또는 무릎 부위를 이용할 것을 제안한다. 또한, 국내 보행 환경을 반영한 환경 센싱을 위해 추가적인 데이터 셋의 확보가 필수적이라 판단한다.

학습 기반 기법의 성능은 학습 자료에 지배적인 영향을 받는다. 앞서 언급한 여러 주장도 국내 보행 환경의 특징을 나타내는 자료를 확보하는 것을 전제한다. 국내 보행 환경에 대한 대규모 학습 자료를 확보하는 것과 학습 자료가 날씨와 계절에 따른 국내 보행 환경의 다양한 모습을 포함하는 것이 필요하다.

5. 결론

이 연구는 국내 보행 환경에 대한 인지 지능을 고도화하기 위한 기초 연구로서 영상 분할 기법을 설명한다. 이 연구는 국내 보행 환경의 특징을 나타내는 영상 자료에 대한 기존 영상 분할 기법들의 동작 특성과 성능을 비교 및 분석하고, 국내 보행 환경에 대한 영상 분할 기법의 개발 방향과 인지 센서의 구성 및 배치 전략을 설명한다. 이 연구 결과는 향후 국내 보행 환경에 적합한 보행 보조 지능에 관한 연구에서 기초 자료로 활용될 것이다.

참고문헌

[1] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[2] Zhou, Zongwei, et al. "Unet++: A nested u-net architecture for medical image segmentation." Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer, Cham, 2018. 3-11.

[3] Zhao, Hengshuang, et al. "Pyramid scene parsing network." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[4] Chen, Liang-Chieh, et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation." Proceedings of the European conference on computer vision (ECCV). 2018.

[5] Romera, Eduardo, et al. "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation." IEEE Transactions on Intelligent Transportation Systems 19.1 (2017): 263-272.

[6] Yang, Kailun, et al. "Unifying terrain awareness through real-time semantic segmentation." 2018 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2018.

[7] Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[8] Lin, Tsung-Yi, et al. "Feature pyramid networks for object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[9] Sandler, Mark, et al. "Mobilenetv2: Inverted residuals and linear bottlenecks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[10] Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." International Conference on Machine Learning. PMLR, 2019.

[11] Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[12] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.