추종로봇을 위한 RGBD 센서 기반의 End-to-End 학습

이준구, 오지용, 이혜민, 이송, 남승우
한국전자통신연구원

{leejg01679, jiyongoh, leehaemin90, li.song, swnam}@etri.re.kr

End-to-End Learning Based on RGBD Sensor for Robotic Following

Joon-Goo Lee, Jiyong Oh, Hea-Min Lee, Song Li, Seung Woo Nam
Electronics and Telecommunications Research Institute (ETRI)

요 약

본 논문은 추종로봇의 인지를 위하여 RGBD 센서 기반의 End-to-End 학습을 제안한다. 입력 영상은 RGB 영상에 정렬된 depth 영상을 병합하여 4 채널로 구성하며, End-to-End 학습에 사용될 딥러닝 네트워크는 CNN 을 기반으로 하는 특징 추출기와 추종 중심 및 거리를 예측하기 위한 분류 모델로 구성된다. 실험에서 RGB 만 사용하는 방법보다 depth 정보를 추가로 사용하는 방법이 거리 예측에 대하여 약 20% 이상 높은 성능을 보였다.

Ⅰ. 서 론

사용자 추종은 대표적인 인간-로봇 상호작용으로 농업, 수산업과 같은 전통적인 1 차 산업에서부터 물류센터 및 배송, 제조공장, 병원 등 다양한 산업에서 응용될 수 있다. 영상 기반 추적 기술은 사용자 추종 로봇의 필수 요소 기술 중 하나로 다양한 딥러닝 기술과 접목되어 과거에 비해 크게 성능이 향상되고 있다. 최근 Zhang 은 Siameses 네트워크 기반의 추적기에 적합한 backbone 네트워크를 제안하였으며, deformable convolution 을 적용하여 anchor-free 한 네트워크를 제안하기도 하였다[1-2]. 일반적으로 사용자 추종은 추적 기술로 추정된 대상의 위치를 기반으로 로봇의 주행을 제어하는 방식으로 구현된다. 한편, [3]에서는 자율주행 자동차의 조향각 조절을 위해 차선 검출 및 곡선 피팅과 같은 복잡한 절차를 대신하여 입력 영상에서 조향각을 직접 추정하는 End-to-End 방식을 제안하였으며, 해당 연구를 시작으로 End-to-End 접근법에 대한 다양한 후속 연구가 이루어져 로봇 추종 연구에 영향을 미치고 있다. 특히, John 은 시간적으로 연속한 일련의 RGB 이미지를 다양한 방법으로 구성하여 End-to-End 기반의 네트워크에 입력으로 사용하였으며, 추종 대상의 좌, 우, 중앙의 방향과 추종 대상과의 거리를 유지하기 위한 가속의 여부를 분류함으로써 로봇 추종을 성공하였다[4]. 본 논문에서는 [4]와 같이 사용자 추종을 위한 End-to-End 방식에 대한 연구를 제안한다. 하지만 [4]와 달리 RGB 영상 대신 최근 다양하게 활용되고 있는 RGBD 센서를 활용하여 RGB 영상과 함께 depth 정보를 추가로 활용한다.

Ⅱ. 본론

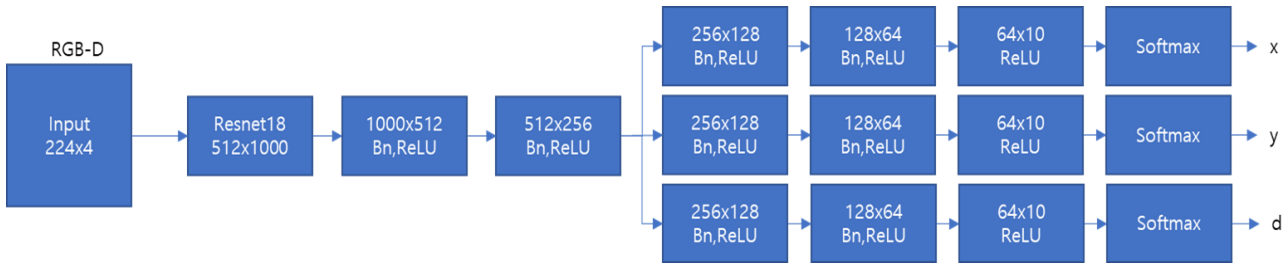
본 논문에서는 RGBD 센서를 입력으로 사용하여 추종 대상의 위치 및 거리를 출력하는 End-to-End 학습 방법을 제안한다.

먼저 입력 영상은 RGB 영상에 정렬된 depth 영상을 병합하여 4 채널로 구성하며, 각 영상의 레이블은 추종 대상의 중심 x, y 좌표와 중심의 거리인 d 로 설정한다. 여기서 x, y, d 각각은 분류 문제로 해결하기 위해 각각 균등한 10 개의 구간으로 샘플링 하여 $X=\{x_0, x_1, \cdots, x_9\}$, $Y=\{y_0, y_1, \cdots, y_9\}$, $D=\{d_0, d_1, \cdots, d_9\}$와 같이 표현한다.

End-to-End 학습에 사용할 네트워크는 그림 1 과 같이 특징 추출을 위한 CNN 과 추종 대상의 중심 좌표 및 거리를 예측하기 위한 분류 모델로 구성된다. CNN 에 사용된 모델은 우수한 경량 네트워크로 잘 알려진 Resnet18 을 기반으로 하였으며[5], 4 채널의 영상을 입력 받기 위해 3 채널로 설정 되어있는 입력 부분을 수정하였다. 분류에 사용된 모델은 2 개의 linear layer 를 연결한 후 ReLU 함수로 활성화하였으며, 추출된 특징을 공유하여 x, y, d 각각을 예측하기 위해 출력이 3 개인 multi-task learning 으로 구성하였다.

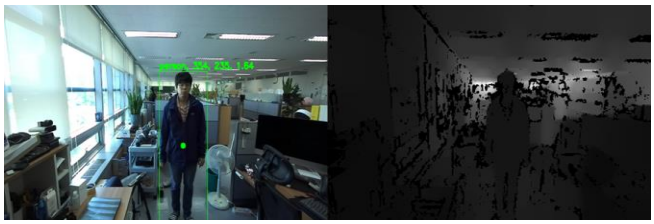
학습 파라미터로 분류기의 weight 변수는 Xavier 초기화를 사용하였으며, optimizer 는 Adam 을 사용하고, 손실 함수는 cross entropy loss 를 사용하였다. 학습률은 0.0001 부터 0.000001 까지 cosine annealing scheduler 를 사용하였으며, batch size 는 16, epoch 는 1000 을 사용하였다.

실험을 위해 그림 2 와 같이 RGBD 영상과 영상 내에 존재하는 추종 대상의 중심점 x, y 좌표 및 거리 d 로 구성된 데이터셋을 구축하였다. 영상 수집에 사용된 센서는 스테레오 기반의 RGBD 카메라인 ZED2 를 사용하였으며, depth 영상은 32 비트 부동소수점 영상을 8 비트의 부호 없는 영상으로 정규화 하여 사용하였다. 대상의 중심은 YOLO 검출기를 사용하고, 대상의 거리는 해당 중심의 depth 영상으로부터 추출하였다. 학습과



<그림 1> End-to-End 구성도

테스트 영상은 각각 다른 사람으로 촬영하였으며, 학습을 위한 영상은 약 1600 장, 테스트를 위한 영상은 약 600 장을 사용하였다.



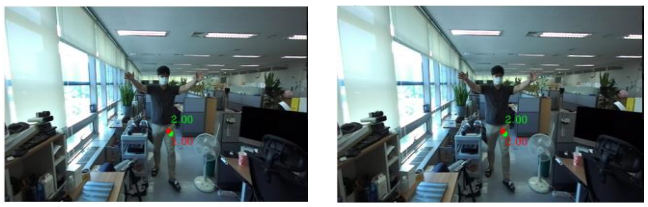
<그림 2> 학습 영상

3 채널 RGB 영상과 4 채널 RGBD 영상의 추종 성능 평가를 위해 표 1 과 같이 두 가지의 학습된 모델을 이용하여 예측 정확도를 비교하였다. 추종 대상의 좌표 중 y 에 대한 예측 결과는 유사한 정확도를 보였으나, x 에 대한 예측 결과는 RGBD 영상이 약 4% 높게 나타났다. 특히 추종 대상의 거리인 d 의 경우 RGBD 영상이 RGB 영상만 이용한 결과 보다 약 20% 이상 높은 성능을 보였다.

<표 1> RGB 기반과 RGBD 기반의 성능 비교

	X_acc	Y_acc	D_acc
RGB	95.67%	99.99%	66.50%
RGBD	99.50%	99.83%	86.33%

그림 3 은 x, y 좌표가 중앙에 위치하며, 2m 의 거리를 정답으로 하는 이미지에 대하여 RGB 와 RGBD 각각의 예측 결과를 표시한 것이다. 그림 3(a)는 RGB 영상을 기반으로 하여 예측한 것으로 x, y 좌표는 정확하게 예측하였으나 거리는 1m 로 다르게 예측하였다. 그림 3(b)는 RGBD 영상을 기반으로 하여 예측한 것으로 x, y 좌표를 정확하게 예측하였으며, 거리도 2m 로 정확하게 예측하였다. 두 영상 모두 추종 대상의 위치는 정확하게 예측하고 있으나 추종 대상의 거리는 RGBD 영상을 이용한 경우가 더 정확하게 예측하였다.



(a) RGB 예측 (b) RGBD 예측

<그림 3> RGB 및 RGBD 예측 결과

Ⅲ. 결론

본 논문에서는 추종 로봇의 추종 대상 인지를 위한 End-to-End 학습 방법에 대해 초기 연구를 수행하였다. RGBD 4 채널의 영상을 입력으로 하여 ResNet18 을 특징 추출기로 사용하였으며, x, y 좌표와 거리 d 를 예측하기 위해 3 개의 multi-task 로 네트워크를 구축하였다. 실험에서 RGB 이미지만 사용하는 방법보다 RGBD 를 사용하는 방법이 거리 예측에서 약 20% 이상 높은 성능을 보였다.

추후 샘플링으로 인해 손실된 값을 보완하기 위해 regression 을 사용한다면 보다 정확한 예측이 가능할 것으로 보인다.

ACKNOWLEDGMENT

본 연구는 한국전자통신연구원 주요사업의 일환으로 수행되었음. [21ZD1130, 지능제어기반 스마트기계 및 로봇 기술 개발]

참 고 문 헌

[1] Z. Zhang, et al., “Deeper and wider Siameses networks for real-time visual tracking,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4591-4600, 2019.

[2] Z. Zhang, et al., " Ocean: Object-aware Anchor-free Tracking,” Proceedings of 16th European Conference on Computer Vision, pp. 771-787, 2020.

[3] M. Bojarski et al., “End to End Learning for Self-Driving Cars,” arXiv:1604.07316.

[4] J. M. Pierre, et al., "End-to-End Deep Learning for Robotic Following,” ICMSCE 2018, pp. 77-85, Feb. 2018.

[5] He, K., et al., "Deep residual learning for image recognition," Proceedings of the IEEE converence on computer vision and pattern recognition, pp. 770-778, 2016.