X-ray영상 기반 단안식 깊이 예측 기법

김낙우, 이현용, 이준기, 황유민, 이병탁
한국전자통신연구원
{nwkim,hyunyonglee,jungi,yumin,bytelee}@etri.re.kr

X-ray image-based monocular depth prediction technique

Nac-Woo Kim, Hyun-Yong Lee, Jun-Gi Lee, Yu Min Hwang, Byung-Tak Lee
Electronics and Telecommunications Research Institute

요 약

본 논문은 X-ray영상 기반으로 단안식 예측 기법을 통해 지도식으로 깊이를 추정하는 방법에 관한 것이다. 임의의 설비 혹은 장치 내부의 특정 객체에 대한 위치를 추정하고, 연속된 촬영을 통해 객체 위치를 추적하기 위해서는 X-ray를 통한 사물 투사 영상이 요구되고, X-ray 영상을 기반으로 한 특정 객체의 X, Y, Z 좌표 인식이 필요하다. 영상 내 객체의 X, Y 좌표 인식은 YOLO(You Only Look Once) 등 다양한 객체 추적 기법이 활성화 되어 있으나, Z좌표를 인식을 위한 X-ray기반 단안식 깊이 예측 기법은 여전히 연구가 미흡한 상황이다. 본 논문에서는, 전이학습을 위한 backbone 모델로 RGB 영상으로 사전학습된 DenseNet 모델을 사용하고, X-ray영상과 같이 단일 채널을 갖는 gray영상 기반으로 모델에 대한 재학습을 진행하였다. 이 때, U-net구조에 DenseNet모델을 인코더부로 활용하여, 깊이추정 영상을 출력하도록 모델을 구성하였다. 시험을 통해 X-ray 영상에 대해서도 객체에 대한 깊이 예측이 충분히 가능한 것을 확인하였다.

Ⅰ. 서 론

2차원 영상기반 3차원 깊이 추정기법은 가상현실(VR), 증강현실(AR), 자율주행자동차, 이동형 로봇 등 많은 응용분야에서 핵심 기술로 각광받고 있다. 이전에는 단안식 혹은 양안식 카메라 기반의 다중뷰 epipolor geometry를 통한 깊이 정보 재구성에 관한 연구가 많이 이루어져왔다. Fundamental 행렬과 essential 행렬로부터 카메라 고유정보를 구하고, 삼각측량법을 통해 원래의 3D 공간좌표를 구하는 방식이다. 이러한 방식은 최근 딥러닝 기술이 보편화됨에 따라, 심층학습 모델을 기반으로 한 깊이 예측 기법으로 전환되고 있다. Fu는 깊이를 disparity space에 따라 양자화하여 단안식 기법으로 예측하는 방법을 제안하면서, atrous convolution을 통해 멀티 스케일에 대응하고, 거리측정을 oridinal regression 문제로 해결하는 기법을 제시한 바 있다[1]. Godard는 양안식 영상을 통해 좌측 영상으로 우측 영상에 대한 disparity를 예측하여 우측 영상을 얻고, 우측 영상으로 좌측 영상에 대한 disparity를 예측하여 좌측 영상을 얻는 monodepth 기법을 제안하였다[2]. 영상재구성 기법을 통해 좌우 영상의 disparity를 예측하는 아이디어를 적용한 것이다. 이 논문에서는 카메라 포즈를 안다고 가정하였는데, 카메라 포즈 또한 심층학습 모델을 통해 추정하는 기법을 새롭게 제안하며 monodepth2를 소개한 바도 있다[3]. Monodepth와 monodepth2는 학습 시에는 양안식 영상을 이용하고, 시험 시에 단안 영상을 통해 깊이영상을 추정하는 방법으로, 학습 시에 양안영상이 필요하다는 단점이 있다. 본 논문에서는 뉴욕대학교의 NYU depth dataset v2[4]를 이용하여, 단안식 영상으로 심층학습모델의 학습과 시험을 진행할 수 있도록 구성하면서, X-ray영상에 대한 깊이 추정이 가능하도록 하였다. NYU depth dataset v2는 kinect를 이용한 RGB+D 실내 비디오 시퀀스 데이터로, 3개의 사이트로부터 얻은 464개의 씬 구성이며, 1,449개의 조밀한 RGB+D 영상쌍 및 407,024개의 unlabeled 프레임으로 제공된다.

Ⅱ. 단안식 깊이 예측 모델 구조

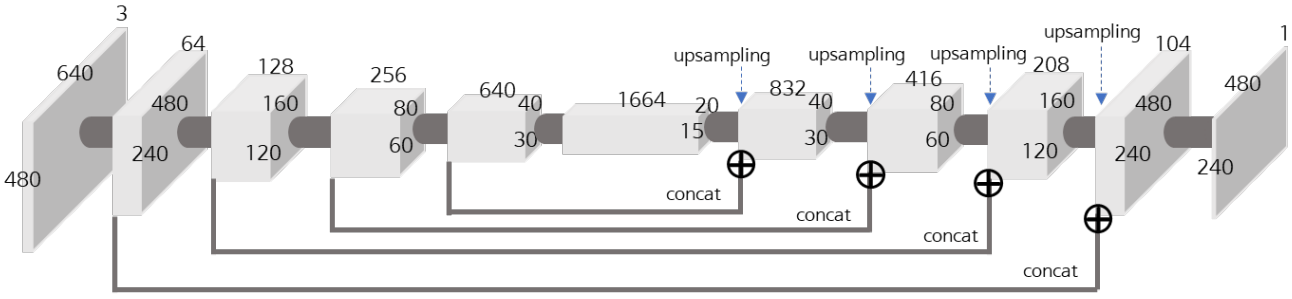
본 연구에서는 단안식 깊이 추정을 위해 U-net 모델 구조를 이용하였다. U-net 구조의 인코더 부분은 사전학습된 DenseNet-169 모델을 사용하였고, 전이학습을 통해 비교적 적은량의 깊이 label로도 높은 수준의 성능을 나타날 수 있었다. DenseNet 기반의 인코더부는 1,664개의 채널과 함께 15x20의 특징맵을 출력하고, 이후 디코더부는 upsampling을 통해 U-net 구조를 완성한다.

그림 1과 같이 U-net구조를 위한 skip-connection은 4개가 연결되어있고, 이를 위해 총 4번의 upsampling이 이루어진다. 디코더의 upsampling 이후의 layer구성은 인코더부의 skip-connection을 위한 concatenate layer, 채널 감소를 위한 convolution layer와 batch normalization 레이어 쌍을 중첩하여 사용하였다. 디코더부의 최종레이어는 한 개의 채널을 갖는 convolution 레이어를 sigmoid함수로 activation하여 사용한다. 이를 통해 입력영상에 대한 깊이 추정 영상을 출력할 수 있다. 손실함수는 l1 거리함수와 edge 거리함수, SSIM(structural similarity) 구조함수를 사용한다. L1거리함수는 깊이 정답영상과 깊이 예측영상 간 픽셀 간 차분, edge 거리함수는 깊이 정답영상을 미분한 값과, 깊이 예측영상을 미분한 값과의 차분, SSIM 구조함수는 깊이 영상에서의 구조적 품질 평가를 위한 손실함수이다. 또한, learning rate decay를 위해 step기반 decay함수를 사용하였다. Epoch당 1/max_epoch만큼 base learning rate 값을 감소시킨다. Base learning rate는 0.0001이다.

Optimizer함수는 AdamW를 이용하였다. AdamW함수는 기존의 Adam 함수의 단점인 낮은 일반화 성능을 개선한 것으로, Adam 함수가 SGD(Stochastic gradient decent)에 비해 일반화가 떨어지는 이유로 L2 정규화 및 weight decay기법에 문제가 있다는 것을 밝히면서, 이를 개선한 기법이다.

[표 1] 단안식 깊이 예측 시험 결과

	l1	log_l1	abs_rel	sq_rel	rmse	log_rmse	δ<1.25	δ<1.25^2	δ<1.25^3
mean	0.109	0.142	0.360	0.061	0.145	0.195	0.482	0.751	0.879
std	0.022	0.027	0.102	0.029	0.030	0.035	0.083	0.078	0.056
min	0.070	0.095	0.232	0.025	0.090	0.137	0.295	0.512	0.677
25%	0.097	0.123	0.265	0.038	0.128	0.171	0.431	0.702	0.840
50%	0.110	0.142	0.341	0.059	0.143	0.193	0.481	0.755	0.887
75%	0.119	0.159	0.403	0.074	0.157	0.217	0.535	0.803	0.920
max	0.178	0.226	0.607	0.144	0.237	0.287	0.661	0.869	0.955



[그림 1] X-ray영상 기반 단안식 깊이 예측 모델 구조.

Ⅲ. 실험 결과

깊이 추정 실험에서의 metric은 다음의 지표를 일반적으로 많이 사용한다. 평균 제곱근 오차(Root Mean Squre Error; RMSE), 로그-RMSE, 절대 상대 오차(Absolute Relative Error; Abs_rel), 제곱근 상대 오차(Square Relative Error; Sq_rel), L1 거리, 로그-L1 거리 등은 낮을수록 성능이 뛰어난 것이고, 정답 레이블과 예측값과의 비율을 통해 정확도를 나타내는 δ는 높을수록 좋은 성능을 나타낸다.

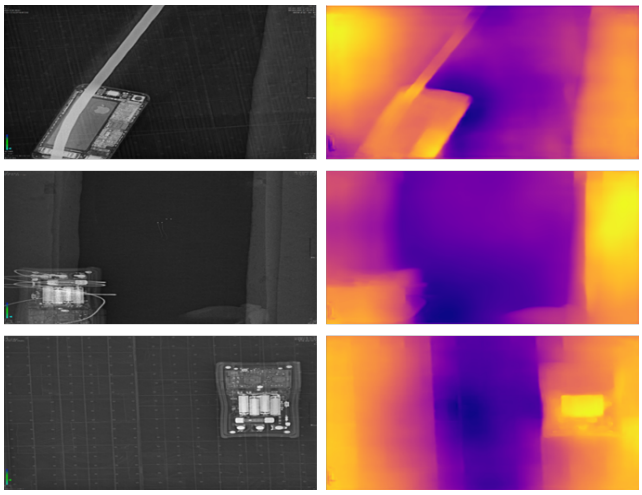
사전학습된 DenseNet 파라미터가 RGB영상에서 학습된 것임을 알기 때문에, 실험 성능을 높이기 위해, NYU depth 데이터세트를 gray영상으로 전환하고 단안식 깊이 예측 모델 구조에 맞춰 전체 모델을 재학습한다. 원하는 태스크인 X-ray영상 기반의 깊이 추정을 위해서, X-ray 영상은 gray영상과 마찬가지로 1채널 영상이기 때문에, gray영상으로 모델을 재학습하는 것은 성능향상에 큰 영향을 끼친다.

표 1은 NYU depth 데이터세트 중 40x16(batch) 시험 데이터를 기반으로 측정한 성능값이다. 40번의 배치시험을 통해 δ<1.25^3 성능 평균값은 87.9% 수준임을 확인하였고, 최대 성능값은 95.5%였다. RMSE값은 평균 0.145이고, 최소값은 0.09 수준으로 굉장히 낮은 수치값을 보였다. L1 거리 값은 평균 0.109, 최대 0.178이었으며, 절대 상대 오차 평균값은 0.36, 최대 값은 0.607로 나타났다.

그림 2는 NYU depth 데이터세트를 통해 학습된 단안식 깊이 예측 모델 구조에 X-ray영상을 입력하고 깊이 예측을 실험한 결과이다. 그림에서 보는 바와 같이, X-ray영상으로 학습하지 않았음에도, X-ray영상 입력에 대해 적절한 깊이 예측 결과를 보이고 있다.

Ⅳ. 결론

본 논문에서는 U-net구조의 단안식 깊이 예측 모델을 구성하고, 사전학습된 DenseNet-169 모델을 U-net구조의 인코더부에 적용함으로써, 적은 량의 깊이 영상 데이터를 활용하여 X-ray영상에 대한 깊이 영상을 획득하는 방법을 제안하였다. X-ray영상에 대한 깊이 레이블을 획득할 수 없기 때문에, 일반적인 RGB영상+Depth영상의 데이터세트를 활용하여, X-ray영상의 깊이 정보를 획득하는 새로운 방법을 소개하였다. 향후 보다 다양한 데이터세트를 활용하여 성능시험을 진행할 예정이다.



[그림 2] X-ray영상 기반 단안식 깊이 예측 결과

ACKNOWLEDGMENT

이 논문은 2022년도 정부(산업통상자원부)의 재원으로 한국에너지기술평가원(KETEP)의 '신재생에너지핵심기술개발사업'으로 지원을 받아 수행한 연구 과제입니다. (No. 20223030020070)

참 고 문 헌

[1] Fu, Huan et al. "Deep Ordinal Regression Network for Monocular Depth Estimation." 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018): 2002-2011.

[2] Godard, Clément et al. "Unsupervised Monocular Depth Estimation with Left-Right Consistency." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017): 6602-6611.

[3] Godard, Clément et al. "Digging Into Self-Supervised Monocular Depth Estimation." 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019): 3827-3837.

[4] Silberman, Nathan et al. "Indoor Segmentation and Support Inference from RGBD Images." ECCV (2012).