단안 카메라 깊이 추정기를 이용한 미지 물체의 자세 추정

송성호, 김인철
경기대학교 AI컴퓨터공학부
ssh10032@kyonggi.ac.kr, kic@kyonggi.ac.kr

Unseen Object Pose Estimation using a Monocular Depth Estimator

Sung-Ho Song, Incheol Kim
School of Artificial Intelligence & Computer Science, Kyonggi University

요 약

3차원 물체의 탐지와 자세 추정은 실내외 환경에서 장면 이해, 로봇의 물체 조작 작업, 자율 주행, 증강 현실 등과 같은 다양한 응용 분야들에서 공통적으로 요구되는 매우 중요한 시각 인식 기술이다. 깊이 지도를 요구하는 기존 연구들과는 달리, 본 논문에서는 RGB 컬러 영상만을 이용해 미지의 물체들, 즉 3차원 CAD 모델을 가지고 있지 않은 새로운 물체들을 탐지해내고, 이들의 자세를 추정해낼 수 있는 새로운 신경망 모델을 제안한다. 제안 모델에서는 최근 빠른 속도로 발전하고 있는 깊이 추정 기술을 이용함으로써, 깊이 측정 센서 없이도 물체 자세 추정에 필요한 깊이 지도를 컬러 영상에서 구해낼 수 있다. 본 논문에서는 벤치마크 데이터 집합을 이용한 실험을 통해, 제안 모델의 유용성을 평가한다.

1. 서론

3차원 물체의 탐지(object detection)와 자세 추정(pose estimation)은 실내외 환경에서 3차원 장면 이해(scene understanding), 로봇의 물체 조작 작업(robotic manipulation tasks), 자율 주행(autonomous driving), 증강 현실(augmented reality) 등과 같은 다양한 응용 분야들에서 공통적으로 요구되는 매우 중요한 시각 인식 기술이다. 특히 이 중에서 물체 6D 자세 추정은 3차원 공간에서 카메라를 중심으로 물체의 3축 회전(rotation)과 3축 변환(translation)으로 표현되는 해당 물체의 실시간 자세를 알아내는 작업이다. 이것은 3차원 공간에 놓인 물체의 위치를 단순히 직육면체 경계 상자로 알아내려는 3차원 물체 탐지와는 달리, 더 높은 정밀도를 요구하는 작업이다.

물체 6D 자세 추정에 관한 과거 연구들은 대부분 대상 물체의 정확한 3차원 CAD 모델을 이용하는 소위 개체-수준 자세 추정(instance-level pose estimation) 방식을 채택하였다. 최근에 와서는 이러한 개체-수준의 자세 추정기들은 매우 높은 수준의 자세 정확도를 얻는 데 성공하였으나, 인식 대상 물체마다 모두 3차원 CAD 모델이 확보되어야만 자세 추정이 가능하다는 한계는 뛰어넘지 못하고 있다 [1]. 반면에, 최근 들어서는 이러한 개체-수준의 자세 추정기들의 한계성을 극복하기 위해, 인식 대상 물체가 속한 범주(category)나 해당 범주에 속한 다른 개체들의 3차원 표현은 알 수 있으나 해당 물체의 3차원 CAD 모델은 가지고 있지 않다고 가정하는 소위 미지의 물체(unseen object)에 관한 범주-수준의 자세 추정(category-level pose estimation)에 관한 연구가 활발하다 [2-4].

기존 연구들에서 제시된 미지 물체에 대한 대표적인 범주-수준의 자세 추정 방식들로는 (a) 3차원 재건과 랜더링(3D reconstruction and rendering) 방식과 (b) 범주별 3차원 NOCS 표현(3D NOCS representation for each category)을 이용하는 방식 등이 있다. 3차원 재건과 랜더링 방식[2]은 다수의 다른 물체 데이터들로 학습된 신경망을 이용해, 인식 대상 물체에 관한 소량의 참조 영상(reference image)들로부터 해당 물체의 잠재적 3차원 표현을 재건한 뒤, 이를 임의 관점에서 랜더링함으로써 깊이 지도(depth 지도)를 예측한다. 그리고 이렇게 예측된 깊이 지도와 대상 물체를 포함한 입력 깊이 지도 간의 매칭을 통해 대상 물체의 3차원 자세를 추정한다. 반면에, 범주별 3차원 NOCS(Normalized Object Coordinate Space) 표현을 이용하는 방식[3, 4]은 인식 대상 물체에 관한 별도의 참조 영상 없이 입력 RGB 영상과 깊이 지도로부터 각각 해당 물체가 속한 범주의 표준 NOCS 표현과 해당 물체의 포인트 클라우드(point cloud)를 생성한 뒤, 이 2개의 3차원 표현을

* 본 연구는 정보통신기획평가원의 재원으로 정보통신방송 기술개발사업의 지원을 받아 수행한 연구 과제(클라우드에 연결된 개별 로봇 및 로봇그룹의 작업 계획 기술 개발, 2020-0-00096)입니다.

서로 매칭함으로써 해당 물체의 자세를 추정한다.

하지만, 앞서 언급한 기존의 두 가지 미지 물체 자세 추정 방식 모두 인식 대상 물체를 포함한 깊이 지도를 입력으로 요구한다. 따라서 물체 자세 추정을 위해서는 일반 RGB 카메라 외에 추가로 적외선(IR)이나 라이다(Lidar) 등 깊이 측정 센서가 필요하다. 한편 최근에는 RGB 단안 카메라 영상으로부터 깊이 지도를 추정해내는 단안 카메라 깊이 추정(monocular depth estimation) 기술이 급속히 발전하여 비교적 높은 성능을 보여주고 있다 [5, 6].

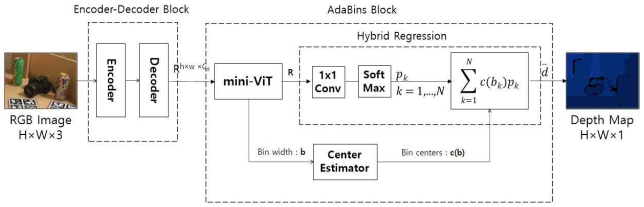
본 논문에서는 깊이 지도를 추가 입력으로 요구하는 기존 연구 모델들과는 달리, RGB 컬러 영상만을 이용해 미지 물체들의 자세를 추정해낼 수 있는 새로운 범주-수준 자세 추정 신경망 모델을 제안한다. 제안 모델에서는 단안 카메라 깊이 추정기를 이용하여 깊이 측정 센서 없이도 물체 자세 추정에 필요한 깊이 지도를 RGB 컬러 영상에서 구해낼 수 있다. 본 논문에서는 벤치마크 데이터 집합 NYU-Depth-V2[6]와 REAL[3]을 이용한 정량 및 정성 평가 실험을 통해, 제안 모델의 유용성을 입증한다.

2. 깊이 추정

학습 기반의 단안 카메라 깊이 추정 방식들은 크게 감독 학습 기법(supervised method)과 비감독 학습 기법(unsupervised method)들로 나눌 수 있다. 감독 학습 기법들은 깊이 추정기의 학습을 위해 입력 RGB 영상과 함께 조밀한 정답 깊이 지도(ground truth depth 지도)가 제공되어야 한다. 반면에, 비감독 기법은 단안 카메라의 연속적인 영상을 기반으로 에피폴라 기하학을 적용함으로써 깊이 지도를 예측한다. 이러한 비감독 기법은 정답 깊이 지도를 수집하지 않아도 된다는 장점이 있지만, 아직은 감독 학습 기법에 비해 안정적인 성능을 보장할 수 없다. 따라서 본 논문에서 제안하는 미지 물체 자세 추정 모델에서는 대표적인 감독 학습 기반의 단안 카메라 깊이 추정기인 AdaBins[6]를 이용한다.

AdaBins 깊이 추정기는 (그림 1)과 같이 표준 인코더-디코더 합성곱 신경망(encoder-decoder convolutional neural network) 구조에, 전역적 정보 처리를 위해 트랜스포머(Transformer) 기반 신경망 블록을 추가하였다. 트랜스포머 기반 신경망 블록인 AdaBins 블록은 깊이 범위(depth range)를 다수의 구간들(bins)로 나누며, 각 구간의 중심 값은 입력 영상에 맞게 적응적으로 추정된다. 따라서 AdaBins 깊이 추정기는 불충분한 전역 정보 처리로 인해 깊이 추정의 품질이 낮았던 기존의 깊이 추정기들의 단점을 극복하고 높은 깊이 추정 성능을 보여주고 있다.

(그림 1)과 같이 AdaBins 깊이 추정기의 인코더-디코더 블록은 RGB 입력 영상으로부터 최종 깊이 지도가 아닌 텐서 $x_d \in R^{h \times w \times C_d}$를 출력한다.



(그림 1) 깊이 추정기 구성

한편, Adabins 블록은 mini-ViT라는 트랜스포머(Transformer) 서브 블록과 구간 중심 추정기(Center Estimator) 서브 블록, 복합 회귀(Hybrid Regression) 서브 블록들로 구성된다. 첫 번째 mini-ViT 트랜스포머 서브 블록은 입력 영상에 대한 깊이 간격을 나누는 방법을 정의하는 구간 넓이 벡터 b(bin-width vector)와 픽셀 수준의 깊이 계산에 유용한 정보를 포함하는 크기 h×w×C의 범위-집중(Range-Attention) 지도 R을 각각 출력한다.

두 번째 구간 중심 추정기 서브 블록은 (식 2)와 같이 N개의 깊이 구간별로 구간 중심값(depth-bin-center) c(b)를 계산한다. (식 1)에서 b는 구간 넓이 벡터를 나타낸다.

$$c(b_i) = d_{\min} + (d_{\max} - d_{\min})(b_i/2 + \sum_{j=1}^{i-1} b_j) \qquad \text{(식 1)}$$

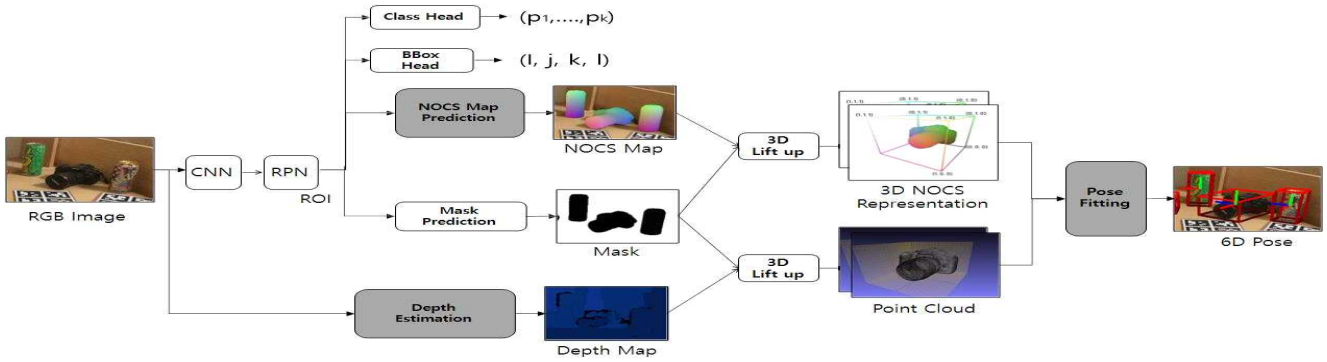
세 번째 복합 회귀 서브 블록에서는 (식 2)와 같이 범위-집중(Range-Attention) 지도 R에서 구해진 각 구간별 점수인 p_k와 각 깊이 구간 중심값 c(b)의 선형 조합(linear combination)으로 각 픽셀의 최종 깊이 값 $\hat{d}$를 추정한다.

$$\hat{d} = \sum_{k=1}^{N} c(b_k) p_k \qquad \text{(식 2)}$$

3. 미지 물체 자세 추정 모델

3.1 모델 개요

단일 RGB 영상 기반 미지 물체의 6D 자세 추정을 위해 본 논문에서 제안하는 모델의 구성은 (그림 2)와 같다. 제안 모델 역시 범주별 3차원 NOCS 표현을 이용하는 방식에 기초하지만, 기존 모델들과는 달리 별도의 깊이 추정 모듈을 채용해 깊이 지도를 자체적으로 생성한다. 제안 모델은 크게 NOCS 지도 예측(NOCS Map Prediction) 헤드가 추가되어 새롭게 확장된 Mask-RCNN 신경망 모듈과 깊이 추정(Depth Estimation) 모듈, 3차원 확장(3D Lift-Up) 모듈, 자세 추론(Pose Fitting) 모듈들로 구성된다. 확장된 Mask-RCNN 신경망 모듈은 합성 곱 신경망(Convolutional Neural Network, CNN)을 통해 RGB



(그림 2) 제안 모델의 구성

입력 영상으로부터 시각적 특징 지도(visual feature map)를 추출한 뒤, 관심 영역 제안 망(Region Proposal Network, RPN)을 통해 구해진 각 관심 영역별로 물체의 종류(class), 물체의 경계 상자(bounding box, bbox), 물체의 마스크(mask), 물체의 NOCS 지도(NOCS map) 등을 예측한다. 한편, 깊이 추정 모듈은 RGB 입력 영상으로부터 그것에 대응하는 깊이 지도를 예측해낸다. 3차원 확장 모듈은 물체 마스크를 기초로, (1) NOCS 지도와 결합하여 해당 물체의 3차원 NOCS 표현을 구하기도 하고, (2) 깊이 지도와 결합하여 해당 물체의 3차원 포인트 클라우드(point cloud)를 얻기도 한다. 마지막으로 자세 추론 모듈에서는 각 물체의 3차원 NOCS 표현과 포인트 클라우드를 서로 매칭함으로써, 해당 물체의 6D 자세와 크기를 예측한다.

3.2 경계상자와 마스크 예측

제안 모델에서 입력 영상의 관심 영역(ROI)별로 물체의 종류와 경계 상자, 그리고 마스크를 예측하는 부분은 영상 기반 물체 개체 분할(image instance segmentation)을 목적으로 개발된 본래의 Mask R-CNN 신경망과 큰 차이가 없다. 물체의 경계 상자 예측을 위한 좌표 회귀(regression)에는 기울기 폭주 현상을 방지하기 위하여, (식 3)과 같이 소프트 L1 손실 함수(L_{bbox})가 사용된다. 물체 종류를 판별하기 위한 분류(classification)에는 교차 엔트로피 손실 함수(L_{class})를 사용한다. 또, 물체 마스크는 픽셀 단위로 분류(pixel-wise classification)가 이루어지기 때문에 (식 3)과 같이 교차 엔트로피 손실 함수(L_{mask})가 사용된다. (식 3)에서 y는 정답 값을, p는 예측치를, k는 관심 영역 안에 있는 픽셀의 수를 각각 나타낸다.

$$L_{bbox} = \begin{cases} 0.5*(p-y)^2, \text{if } (p-y)^2 < 1 \\ \mid p-yvert-0.5, else \end{cases},$$

$$L_{class} = \sum_{c=1}^{k} y_c \log(p_c), \quad \text{(식 3)}$$

$$L_{mask} = \sum_{c=1}^{k} y_c \log(p_c)$$

3.3 NOCS 지도 예측

3차원 NOCS 표현은 동일 범주에 속한 다양한 물체들을 하나의 정규화된 3차원 좌표공간에 통합해 나타낸 것으로서, 해당 범주를 나타내는 표준화된 3차원 표현으로 해석할 수 있다. 반면에 NOCS 지도는 물체의 3차원 NOCS 표현을 카메라의 관점에서 투영해서 얻는 2차원 지도를 의미한다. 제안 모델의 확장된 Mask-RCNN 신경망 모듈은 RGB 입력 영상으로부터 물체의 종류, 경계 상자, 마스크 외에, 물체의 NOCS 지도도 예측한다. NOCS 지도 예측을 위한 회귀(regression)에는 (식 4)와 같은 소프트 L1 손실 함수 $L(y,y^*)$를 사용한다.

$$L(y,y^*) = \frac{1}{n}\begin{cases} 5(y-y^*)^2, & \mid y-y^* \mid \le 0.1, \\ \mid y-y^* \mid -0.05, & \mid y-y^* \mid > 0.1 \end{cases} \quad \text{(식 4)}$$
$$\forall y \in N, y^* \in N_p,$$

(식 4)에서 y는 정답 NOCS 지도 정답 픽셀 값을, y*은 예측된 픽셀 값을, n은 관심 영역(ROI) 내부의 마스크 픽셀 수를 각각 나타낸다.

3.4 6D 자세 추정

RGB 영상으로터 예측된 NOCS 지도는 3차원 확장 모듈에 의해 물체 마스크와 결합되어, 해당 물체의 3차원 NOCS 표현 P_n을 구하는데 이용된다. 또한 깊이 추정 모듈에 의해 예측된 깊이 지도 역시 3차원 확장 모듈에 의해 물체 마스크와 결합됨으로써, 해당 물체의 3차원 포인트 클라우드 P_m를 얻는데 이용된다. 예측이 완료된 후에는 마스크 영상을 이용하여 NOCS 지도의 물체 영역만을 잘라낸 후 컬러 코딩된 3차원 좌표를 복원함으로써 NOCS 표현 P_n을 구성한다. 자세 추론 모듈에서는 이렇게 구해진 물체의 3차원 NOCS 표현 P_n을 포인트 클라우드 P_m과 정렬(align)을 통해, 해당 물체의 크기(scale)와 회전(rotation), 변환(translation) 값을 추정한다. 이 강체 정렬 추정 문제 해결을 위해서는 Umeyama 알고리즘을 이용하고, 이상치 제거를 이해서는 RANSAC 알고리즘을 사용한다.

4. 구현 및 실험

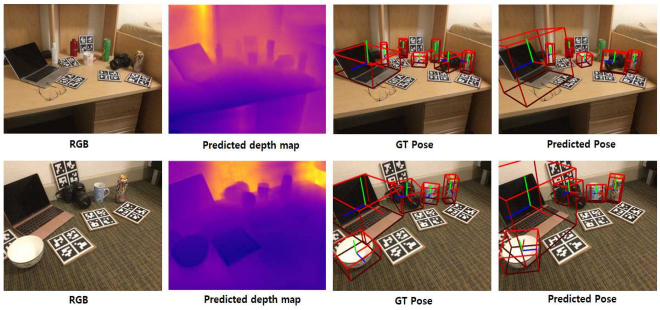
본 논문에서는 제안 모델의 깊이 추정 성능 평가를 위해 NYU-Depth-V2[6] 데이터 집합을 사용하였고, 자세 추정 성능 평가를 위해서는 REAL[3] 데이터 집합을 사용하였다. Keras로 구현된 제안 모델은 GeForce RTX 3090 GPU가 탑재된 하드웨어와 Ubuntu 18.04.6 LTS 플랫폼에서 학습과 평가를 수행하였다. 첫 번째 실험은 AdaBins 기반의 깊이 추정 모듈을 채용한 제안 모델의 깊이 추정 성능을 분석하기 실험이다. 이 실험에서는 제안 모델의 깊이 추정기를 별도의 전역적 정보 처리 기능이 포함되지 않은 인코더-디코더 기반의 깊이 추정기들인 Dense Depth와 BTS들과 성능을 비교하였다. 성능 평가 지표로는 절대 상대 오차(Absolute Relative Error, Abs Rel), 제곱근 상대 오차(Square Relative Error, Sq Rel), 평균 제곱근 오차(Root Mean Square Error, RMSE), RMSE log, 척도 불변 로그(Scale Invariant Log) 와 픽셀에 대한 깊이 참값 d, 추정된 깊이 $\hat{d}$에 대하여 $\max(\frac{d_i}{\hat{d}_i}, \frac{\hat{d}_i}{d_i})$의 임계값을 각각 1.25, 1.25^2, 1.25^3을 적용한 정확도 지표 $\delta_1, \delta_2, \delta_3$ 등을 이용하였다.

Model	δ1	δ2	δ3	AbsRel	SqRel	RMSE	RMSElog	SILog	Log_10
DenseDepth	0.840	0.972	0.993	0.132	0.099	0.558	0.179	17.711	0.055
BTS	0.880	0.979	0.994	0.113	0.067	0.394	0.143	11.602	0.048
AdaBins	0.903	0.984	0.997	0.103	0.056	0.367	0.136	10.892	0.044

(표 1) 깊이 추정 성능 비교

(표 1)의 실험 결과에서 보듯이, 제안 모델의 AdaBins 깊이 추정기는 제곱근 상대 오차(Square Relative Error:Sq Rel)에서 DenseDepth보다 43.4%, BTS보다 16.4% 더 향상된 성능을 보였고, 척도 불변 로그(SILog)에서도 각각 38.5%, 6.1% 더 높은 성능을 보였다. 이와 같은 실험 결과를 통해, 전역적 정보 처리 능력을 포함한 제안 모델의 AdaBins 깊이 추정기의 우수성을 확인할 수 있었다.

두 번째 실험은 제안 모델의 자세 추정 성능을 정성적으로 분석하기 위한 실험이다. 이 실험에서는 REAL 데이터 집합의 사례들을 이용하였다. (그림 3)는 제안 모델을 이용해 물체들의 6D 자세 추정을 수행한 결과들을 나타낸다. (그림 3)의 사례들에서 보듯이, 제안 모델은 실제 깊이 지도 대신 예측된 깊이 지도를 이용하는 불리한 조건에도 불구하고, 비교적 정답에 근접한 자세 추정 결과들을 보여준다.



(그림 3) 물체의 6D 자세 추정 결과들

하지만 한편으로는 (그림 3)의 두 사례에서 보듯이, 요철이 심한, 즉 깊이 값의 변동이 심한 물체들에 대한 크기(scale) 및 변환(translation) 예측, 그리고 대칭 물체(symmetric object)에 대한 회전(rotation) 예측에는 성능 개선의 여지 남아있음을 동시에 확인할 수 있었다.

5. 결론

본 논문에서는 깊이 지도를 추가 입력으로 요구하는 기존 연구 모델들과는 달리, RGB 컬러 영상만을 이용해 미지 물체들의 자세를 추정해낼 수 있는 새로운 범주-수준 자세 추정 신경망 모델을 제안하였다. 제안 모델에서는 단안 카메라 깊이 추정기를 이용하여 물체 자세 추정에 필요한 깊이 지도를 RGB 컬러 영상에서 구해낼 수 있다. 본 논문에서는 벤치마크 데이터 집합을 이용한 실험을 통해, 제안 모델의 유용성을 입증하였다.

참고문헌

[1] Zaixing He, Wuxi Feng, Xinyue Zhao, et al, “6D Pose Estimation of Objects: Recent Technologies and Challenges,” *Applied Science, 2021, 11, 228.*

[2] K. Park, A. Mousavian, Y. Xiang, et al, “LatentFusion: End-to-End Differentiable Reconstruction and Rendering for Unseen Object Pose Estimation.” *Proc of CVPR-2019,* 2019..

[3] H. Wang, S. Sridhar, J. Huang, et al, “Normalized Object Coordinate Space for Category-Level 6D Object Pose and Size Estimation,” *Proc of CVPR-2019,* 2019.

[4] J. Wang, K. Chen, and Q. Dou, “Category-Level 6D Object Pose Estimation via Cascaded Relation and Recurrent Reconstruction Networks,” *Proc of IROS-2021,* 2021.

[5] C. Lee, D. Shim, H. Kim, “Deep Learning Based Monocular Depth Estimation: Survey,”*JPNT,* Vol. 10, No. 4, 2021.

[6] Shriq Farroq Bhat, Ibraheem Alhashim, Peter Wonka "AdaBins: Depth Estimation Using Adaptive Bins.”*Proc of CVPR-2021,* 2021.