UV-map 기반의 신경망 학습을 이용한 조립 설명서에서의 부품의 자세 추정

*강이삭 **조남익

서울대학교

*isaackang@ispl.snu.ac.kr **nicho@snu.ac.kr

UV Mapping Based Pose Estimation of Furniture Parts in Assembly Manuals

*Kang, Isaac **Cho, Nam Ik

Seoul National University

요약

최근에는 증강현실, 로봇공학 등의 분야에서 객체의 위치 검출 이외에도, 객체의 자세에 대한 추정이 요구되고 있다. 객체의 자세 정보가 포함된 데이터셋은 위치 정보만 포함된 데이터셋에 비하여 상대적으로 매우 적기 때문에 인공 신경망 구조를 활용하기 어려운 측면이 있으나, 최근에 들어서는 기계학습 기반의 자세 추정 알고리즘들이 여럿 등장하고 있다. 본 논문에서는 이 가운데 Dense 6d Pose Object detector (DPOD) [11]의 구조를 기반으로 하여 가구의 조립 설명서에 그려진 가구 부품들의 자세를 추정하고자 한다. DPOD [11]는 입력으로 RGB 영상을 받으며, 해당 영상에서 자세를 추정하고자 하는 객체의 영역에 해당하는 픽셀들을 추정하고, 객체의 영역에 해당되는 각 픽셀에서 해당 객체의 3D 모델의 UV map 값을 추정한다. 이렇게 픽셀 개수만큼의 2D – 3D 대응이 생성된 이후에는, RANSAC과 PnP 알고리즘을 통해 RGB 영상에서의 객체와 객체의 3D 모델 간의 변환 관계 행렬이 구해지게 된다. 본 논문에서는 사전에 정해진 24개의 자세 후보들을 기반으로 가구 부품의 3D 모델을 2D에 투영한 RGB 영상들로 인공 신경망을 학습하였으며, 평가 시에는 실제 조립 설명서에서의 가구 부품의 자세를 추정하였다. 실험 결과 IKEA의 Stefan 의자 조립 설명서에 대하여 100%의 ADD score를 얻었으며, 추정 자세가 자세 후보군 중 정답 자세에 가장 근접한 경우를 정답으로 평가했을 때 100%의 정답률을 얻었다. 제안하는 신경망을 사용하였을 때, 가구 조립 설명서에서 가구 부품의 위치를 찾는 객체 검출기(object detection network)와, 각 개체의 종류를 구분하는 객체 리트리벌 네트워크(retrieval network)를 함께 사용하여 최종적으로 가구 부품의 자세를 추정할 수 있다.

1. 서론

객체 검출은 컴퓨터 비전 분야에서 중요한 비중으로 다루어져 왔으며, 지난 몇 년간 인공신경망의 발전과 함께 Faster-RCNN [1], YOLO [2], SSD [3]등의 등장으로 성능 향상이 이루어져왔다. 이러한 검출기들은 모두 이미지에서 검출을 목표로 하고 있는 객체에 잘 들어맞는 사각형을 그리고자 한다. 하지만 최근에는 증강현실, 로봇공학 등의 분야에서 객체의 경계 사각형 이외에도, 객체의 3차원 공간상에서의 회전까지 고려한 자세 추정에 대한 요구가 생기고 있다.

객체의 자세를 추정하는 과제에서는 객체와 관찰자 사이의 상대적 위치관계에 따라서 객체의 모습이 크게 변하는 점, 다른 객체가 관심 객체를 가림으로 인해 객체의 일부분만을 통해 자세 추정을 해야 하는 점, 객체의 주변에 복잡한 환경이 존재하는 점 등의 어려움이 있다. 만약 깊이 정보까지 주어지는 RGB-D 영상이 있는 경우 과제가 조금 더 수월해지지만, 깊이 정보와 함께 객체의 자세 정보까지 주어지는 데이터셋은 일반적으로 얻기 힘들다. 따라서 본 논문에서는 일반적인 카메라에서 얻을 수 있고 깊이 정보가 주어지지 않은 RGB 영상만으로 영상 내 객체의 자세를 추정하는 신경망 구조를 다루고자 한다. RGB 영상 기반 객체 자세 추정 과제에서는 깊이 정보가 없는 RGB 영상과, 해당 영상 내 관심 객체의 3D 모델을 가지고 있는 상태에서, 2D-3D 관계를 올바르게 표현하는 전이행렬(transition matrix)과 회전행렬(rotation matrix), 혹은 이를 포함하는 투영행렬(projection matrix)을 구하고자 한다.0

일반적으로 RGB 영상을 기반으로 물체의 자세를 추정하는 연구들은 실제 환경에서 얻어진 입체적인 물체의 자세를 추정하는 것을 목표로 한다. 반면, 본 논문에서는 가구 조립 설명서를 RGB 영상 입력으로 하고, 설명서에 그려진 부품들의 자세를 추정하고자 한다. 가구 조립 설명서에서는 부품의 자세가 실제 환경의 임의의 물체들에 비하여 제한되어 있어 신경망이 탐색해야할 범위가 줄어든다는 이점이 있지만, 한편으로 부품 그림이 흰 배경에 검은색 선들로만 이루어져 있다는 점에서 실체 물체와 달리 물체 표면의 질감과 색상을 기반으로 학습할 수 없다는 단점이 있다. 본 논문에서는 ground truth 자세 정보가 주어진 실제 환경에서의 RGB 영상을 학습셋과 평가셋으로 분리하여 학습을 진행하는 일

반적인 경우와 다르게, 가구 부품의 3D 모델을 2D로 렌더링 한 합성셋으로 학습을 진행하고 실제 가구 조립설명서에 대해서 성능평가를 진행하였다.

2. 관련 연구

초기의 자세 추정 알고리즘들은 SIFT [4]와 같이 신경망을 쓰지 않는 방법에 의존하였으나, 이는 객체 주변에 복잡한 배경이 있거나 조명 상황이 변하는 등의 객체의 환경에 변화에 대해 취약하였다. 최근에는 일반적인 상황에서의 안정적인 성능을 위하여 신경망 구조를 사용하는 방식들이 등장하기 시작했다. 초기에는 PoseNet [5]과 PoseCNN [6]과 같이 RGB 영상에서 직접 회전행렬을 회귀적인 방법으로 추정하는 방법들이 제안되었다. 그러나 이 방식에는 영상의 깊이 정보가 주어지지 않은 상태에서 신경망의 매개 변수가 너무 넓은 범위를 탐색해야 한다는 한계점이 있었다.

이후에는 영상에서 직접 회전행렬을 구하는 대신에, 영상에서 객체의 특징점들을 먼저 구하고, 2D 영상과 3D 모델의 특징점들의 대응 관계를 통하여 PnP (Perspective-n-Point) 알고리즘으로 전이행렬과 회전행렬을 추정하는 방식이 제안되었다. RGB 영상에서 직접 객체의 자세를 추정하는 것에 비하여 객체의 자세 변화에 따른 특징점들의 이동을 추정하는 과제가 신경망 입장에서 상대적으로 쉽기 때문에 성능 향상이 이루어질 수 있었다.

이와 같은 두 단계 방식은 객체에서 어떤 특징점들을 찾는가에 따라서 더 분류할 수 있다. BB8 [7]과 YOLO6D [8]에서는 객체의 3차원에서의 경계 직육면체의 꼭짓점들을 특징점으로 삼는다. 하지만 경계 직육면체의 꼭짓점들은 객체에서 다소 거리가 있는 곳에 위치하기 때문에, 특징점이 RGB 영상의 경계를 넘어서 위치하는 경우가 발생하는 문제점이 있다. 또한 객체의 모양을 기반으로 객체에서 어느 정도 떨어진 곳의 꼭짓점들의 위치를 추정하는 것은 다소 포괄적인 과제이기 때문에 객체의 일부가 가려진 경우에는 성능 하락의 폭이 큰 문제점이 있었다.

이에 대한 대응방안으로 PVNet [9]에서는 FPS (Farthest Point Sampling) 알고리즘을 통하여 객체의 표면상에서 특징점들을 정의할 수 있도록 하였다. 또한, 특징점들의 위치를 객체 전체를 보고 회귀적으로 추정하는 것이 아니라, 영상의 모든 픽셀들이 특정 특징점의 해당 픽셀에서 보았을 때의 방향을 추정하도록 하여, 특징점들의 위치 추정에 대한 안정성을 높였다. 객체의 모든 픽셀이 개별적인 추정을 하도록 하여 안정성을 높이는 PVNet [9]과 비슷한 개념으로, DPOD [11]에서는 객체의 모든 픽셀이 3D 모델의 UV map 값을 추정하도록 하여, 모든 픽셀이 2D-3D 대응 관계를 추정하는 특징점과 같은 역할을 하게 함으로써 자세 추정의 두 번째 단계인 RANSAC (Random Sampling Consensus)기반의 PnP 알고리즘의 성능을 높였다.

3. 실험 방법

3.1. UV map

본 논문에서는 DPOD [11]를 기본 구조로 하였다. DPOD [11]에서는 객체의 각 픽셀이 3D 모델의 UV map을 추정하도록 하는데, UV map이란 2차원의 이미지를 3차원 모델에 투영할 때의 근사 값에 해당된다. 이를 반대로 해석하면, 3개의 변수로 표현되어야 하는 3D 모델을 2개의 변수로 근사하는 효과가 있다. DPOD [11]에서는 2D 영상에서 객체의 각 픽셀이 3D 모델에서 대응되는 픽셀의 위치를 추정하는데 필요한 3개의 변수를 직접 추정하는 대신에, UV map의 2개의 변수를 추정하도록 하여 신경망의 매개변수의 탐색 범위를 제한하는 한편, 2D-3D 대응 관계 추정의 개수가 늘어나는 효과를 얻었다.

본 논문에서 사용된 3D 모델의 UV map은 다음과 같이 유도된다. 물체 위의 임의의 점 P에서 물체의 중점을 향하는 단위 벡터를 $\hat{d}=(d_x,d_y,d_z)$라고 하면, 다음 두 수식을 통해 물체 위의 모든 점을 $u,v\in[0,1]$인 UV공간으로 매핑 할 수 있다.

$$u = 0.5 + \frac{\arctan 2(d_x, d_z)}{2\pi}$$
$$v = 0.5 - \frac{\arcsin(d_y)}{\pi} \quad (1)$$

위의 수식을 그대로 사용했을 때, 3D 모델의 y축 방향 길이가 짧은 경우에는 3D 모델 상에서 위치가 다르지만 같은 u,v값으로 매핑 되는 점들의 개수가 많아져 나중에 PnP 알고리즘으로 자세를 추정하기 어려워진다. 따라서 3D 모델의 x, y, z축 방향 길이를 구한 후, 가장 긴 축 방향 성분이 <수식 1>의 두 번째 식에 사용될 수 있도록 하였다. 또한 본 논문에서는 DPOD [11]의 방식을 따라, u와 v값의 추정을 회귀 문제(regression)가 아닌 분류 문제(classification)으로 바꾸기 위하여 위의 수식에 255를 곱하여 각 픽셀이 0에서 255의 정수 값을 추정하도록 한다. 이러한 ground truth $UV-map$은 3D 모델 하나 당 사전에 정의된 24개의 자세 후보에 대하여 각각 구해진다.

3.2. 신경망 구조

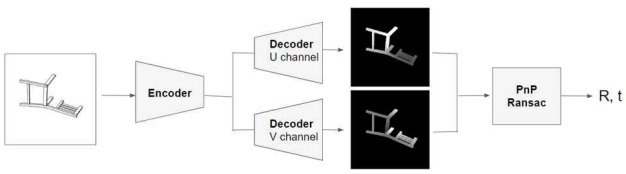

그림 1 신경망 구조

본 논문의 관심사는 가구 부품의 자세 추정에 있기 때문에, 조립 설명서에서 특정 객체의 경계 사각형은 객체 검출기(object detector)를 통하여 이미 구해졌다고 가정한다. 또한 객체 리트리벌 네트워크(image retrieval network)를 통해 조립설명서에서 객체의 경계 사각형을 잘라냄으로 인해 얻는 RGB 영상의 객체가 현재 가지고 있는 3D 모델 가운데 어느 모델과 대응되는지도 사전에 알고 있다고 가정한다. 전체 신경망의 구조는 DPOD [11]와 동일하게 U-net [10] 기반의 구조이다. 처음에 $H\times W\times C$의 RGB 영상을 받아서 인코더(Encoder)를 통과시키며, 본 논문에서는 $224\times224\times3$의 입력을 사용하였다. 이후에 두 갈래의 출력으로 나뉘는데, 첫 번째로 U 채널 디코더(Decoder)에서는 $H\times W\times D$ 크기의 텐서가 출력된다. 여기서 $D=256$이며

(i,j,k) 위치에서의 값은 (i,j)위치 픽셀에 대응되는 $U-map$의 값이 k가 될 확률이다. 두 번째 D 채널 디코더(Decoder)는 앞선 출력단과 동일하지만, 다만 $U-map$ 대신에 $V-map$에 대한 확률 값이 출력된다. 이후에 각 채널별 max 값만 남기는 방식으로 $U-map$과 $V-map$을 얻은 뒤, 이를 합쳐서 전체 $UV-map$을 구할 수 있다. 이렇게 구한 추정 $UV-map$은 24개의 ground truth $UV-map$과 각각 비교되어, 동일한 (i,j) 위치에서 동일한 (u,v)값을 가지는 픽셀의 개수가 최대인 ground truth $UV-map$이 선택된다. 해당 $UV-map$은 다음 RANSAC 기반의 PnP 알고리즘으로 넘어가고, 결과적으로 입력 RGB 영상 내의 객체와, 현재 가지고 있는 객체의 3D 모델간의 변환 관계를 나타내는 회전행렬 R과 전이행렬 t를 추정하게 된다. 전체 네트워크의 Loss는 각 디코더(Decoder)에서의 Cross-entropy loss인 L_u와 L_v의 합이며 다음과 같이 표현할 수 있다.

$$L = L_u + L_v \tag{2}$$

신경망 학습 시에는 사전에 정의된 24개의 자세 후보로 3D 모델을 2D로 투영한 RGB 영상을 입력으로 삼는다. 이와 같이 사전에 자세 후보를 미를 정해 놓은 이유는, 조립 설명서에서 객체의 자세의 가짓수에 한계가 있기 때문이다. 이렇게 자세 후보를 미리 정해놓음으로 인해서 3D 모델이 투영된 2D 영상으로 이루어진 학습셋의 크기가 무한히 커지지 않고, 신경망의 파라미터의 탐색 범위도 한정되어 학습이 용이한 장점이 있다. 실제 조립 설명서에서는 객체의 경계 사각형을 잘라내었을 때 사각형 내에 다른 객체의 일부분이 함께 담기는 경우가 많은데, 따라서 학습을 위한 2D 영상을 제작할 때 일정한 확률로 배경에 조립 설명서의 임의의 부분을 잘라내서 첨부하였다. 실험 결과를 얻을 때에는 학습에 사용된 2D 영상을 사용하지 않고, 대신 실제 조립설명서에서 특정 객체를 경계 사각형으로 잘라낸 2D 영상을 입력으로 하였다.

3.2. ADD score

일반적으로 자세 추정 알고리즘은 ADD score를 평가 지수로 사용한다. Ground truth 회전행렬 R과 전이행렬 t, 그리고 예측된 회전행렬 $\hat{R}$과 $\hat{t}$가 주어졌을 때, 다음과 같이 3D 모델을 이루는 점들의 실제 위치와 예측된 위치 사이의 거리의 평균을 구할 수 있다.

$$m = avg_{x \in M} \| (Rx+t) - (\hat{R}x+\hat{t}) \|_2 \tag{3}$$

여기서 M은 특정 3D 모델을 이루는 점들의 집합이다. 이 때 m의 값이 3D 모델의 직경의 0.1배 값보다 작은 경우 정답으로 처리한다. 이러한 방식으로 전체 평가셋에 대하여 점수를 평균 내어 ADD score를 구할 수 있다.

3.3. 학습셋 생성

본 논문에서는 IKEA사의 Stefan 조립도와, 해당 조립도에 나타나는 가구 부품의 CAD 모델을 사용하여 학습과 평가가 진행되었다. 이때 조립설명서에서 각 객체의 경계 사각형을 알고 있으며, 또한 각 객체에 대응되는 3D 모델 (CAD 모델)이 어느 것인지 미리 알고 있다는 가정을 하며, 본 논문에서 제안된 신경망에서는 해당 객체의 자세만 추정된다.

학습셋을 만들기 위하여 사전에 정의된 24개의 자세 후보들로 3D 모델을 2D로 투영시킨 영상을 얻는다. 실험 환경에서 조립설명서 내에서 객체의 경계 사각형이 주어졌다고 가정하고, 이를 잘라낸 결과를 신경망의 입력으로 사용하기 때문에 전이행렬 t는 항상 고정된 값이다. 따라서 학습 셋을 구성하기 위한 자세 후보들 또한 t는 고정시키고 회전행렬 R만 변화시키며 생성하였다. 또한 실제 조립 설명서에서는 객체들이 서로 겹쳐져 있는 경우가 있기 때문에, 학습셋 구성 시에도 배경으로 조립 설명서의 임의의 부분을 잘라내서 첨부하였다. 학습셋 생성에는 stefan의 부품에 해당되는 총 11개의 CAD 모델이 사용되었다.

한편 부품에 대칭성이 있는 경우에는 서로 다른 자세 후보에서 3D 모델을 투영하였음에도 불구하고, 동일한 투영 영상이 얻어지는 경우가 생긴다. 이를 학습에 사용하였을 경우에 동일한 투영 영상 입력에 대하여 신경망이 다른 $UV-map$을 추정해야하기 때문에 학습이 불안정해진다. 따라서 부품의 대칭성으로 인해 두 가지 자세 후보에서 완전히 동일한 투영 영상을 얻는 경우에는 둘 중 하나의 자세 후보는 자세 후보군에서 제외시켰다. 해당되는 부품은 <그림 2>에서 '부품 6', '부품 7', '부품 8'이며, 각각 24개 대신에 12개, 18개, 18개의 자세 후보가 학습과 평가 시에 사용되었다.

신경망 평가 시에 사용된 RGB 입력은 조립설명서에서 각 가구 부품의 경계 사각형 부분을 잘라내어 만들어졌으며, 신경망 학습 시에 사용된 영상들은 평가에 사용되지 않는다.

4. 실험 결과

Stefan의 CAD 모델에서 생성된 RGB 영상들로 학습을 진행하고, 실제 Stefan 조립설명서에 대하여 평가를 진행한 결과는 <그림 2>와 같으며, 100%의 ADD score와 추정 자세가 자세 후보군 중 정답 자세에 가장 근접한 경우를 정답으로 평가했을 때 100%의 정답률을 얻었다.

본 실험의 한계점으로는 우선 조립설명서에서 각 객체의 위치를 알며, 해당 객체에 대응되는 3D 모델이 무엇인지 알고 있다는 실험의 가정이 있으며, 각각을 위한 객체 검출기(object detector)와 객체 리트리벌 네트워크(retrieval network)가 별도로 필요하다. 두 번째 한계점으로 3D 모델을 2D에 투영한 영상과, 실제 조립도에서 가구 부품 그림의 선 두께 등의 특성이 다르기 때문에, 학습을 오래 진행할수록 신경망의 성능이 떨어지는 문제점이 있다. 이는 특성이 다른 학습셋과 평가셋에 대하여 신경망의 성능을 보존시키는 도메인 어댑테이션(Domain Adaptation)을 적용하여 개선해야할 사항이다.

5. 결론

본 논문에서는 가구 조립설명서에 그려져 있는 가구 부품들의 자세를 추정하기 위한 알고리즘이 제안되었다. 이를 위하여 일반적인 객체의 자세 추정을 위한 신경망 구조인 DPOD [11]의 구조를 기본으로 한 신경망 구조를 제안하였다. IKEA사의 Stefan 가구에 대하여 CAD 모델로 학습셋을 생성하고, 조립 설명서에 대하여 평가를 수행한 결과 높은 정확성으로 조립 설명서에 그려져 있는 가구 부품들의 자세를 추정할 수 있었다.

감사의 글

이 논문은 2020년도 BK21 플러스 창의정보기술 인재양성사업단에 의하여 지원되었음.

참고문헌

[1] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems. 2015.

[2] Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[3] Liu, Wei, et al. "Ssd: Single shot multibox detector." European conference on computer vision. Springer, Cham, 2016.

[4] Lowe, David G. "Object recognition from local scale-invariant features." Proceedings of the seventh IEEE international conference on computer vision. Vol. 2. Ieee, 1999.

[5] Kendall, Alex, Matthew Grimes, and Roberto Cipolla. "Posenet: A convolutional network for real-time 6-dof camera relocalization." Proceedings of the IEEE international conference on computer vision. 2015.

[6] Xiang, Yu, et al. "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes." arXiv preprint arXiv:1711.00199 (2017).

[7] Rad, Mahdi, and Vincent Lepetit. "Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth." Proceedings of the IEEE International Conference on Computer Vision. 2017.

[8] Tekin, Bugra, Sudipta N. Sinha, and Pascal Fua. "Real-time seamless single shot 6d object pose prediction." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[9] Peng, Sida, et al. "Pvnet: Pixel-wise voting network for 6dof pose estimation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

[10] 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.

[11] Zakharov, Sergey, Ivan Shugurov, and Slobodan Ilic. "Dpod: Dense 6d pose object detector in rgb images." Proceedings of the IEEE international conference on computer vision. 2019.

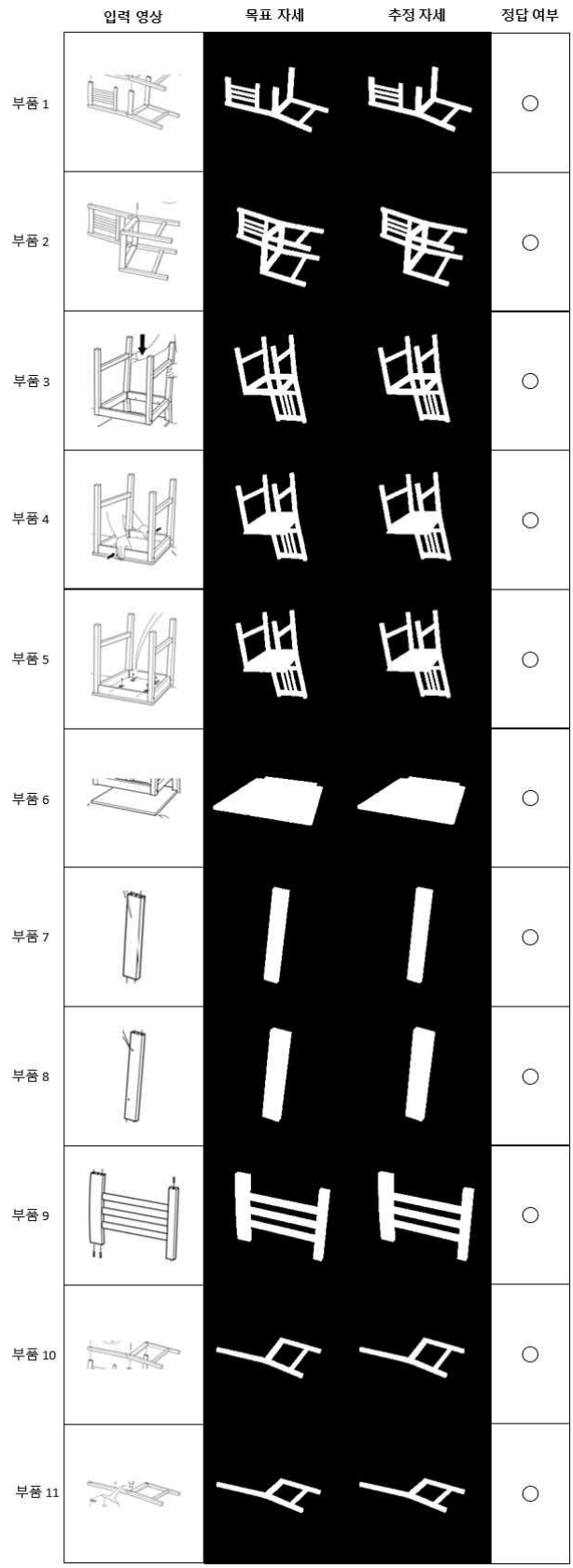

그림 2 IKEA사의 Stefan 가구조립도에 대한 실험 결과. 차례대로 입력 영상, 목표로 하고 있는 출력 자세, 신경망의 추정 자세이다. 24개의 자세 후보군에서 추정 자세가 정답 자세에 가장 가까운 경우 정답으로, 그 외의 경우 오답으로 처리하였다.