Attentional View Pooling 을 이용한 조립 부품 이미지 기반 3 차원 물체 검색

이은지, 강이삭, 김민우, 박선지, *조남익
서울대학교 전기정보공학부, 뉴미디어통신공동연구소
jane0119@snu.ac.kr, isaackang@snu.ac.kr, mwk0614@ispl.snu.ac.kr,
seonjipark@ispl.snu.ac.kr, *nicho@snu.ac.kr

Assembly Part Image-based 3D Shape Retrieval using Attentional View Pooling

Eun Ji Lee, Isaac Kang, Min Woo Kim, Seon Ji Park and *Nam Ik Cho
Department of ECE, INMC, Seoul National University

요 약

조립 부품 이미지에 해당하는 3D CAD 모델 매칭 기술은 최근 로봇 조립 기술의 발전으로 필요성이 대두되고 있다. 이미지 기반 3 차원 모델 매칭 연구는 진행되어 왔지만 가구 부품 이미지와는 특성이 다른 RGB[5] 이미지나 스케치 이미지를 다루는[1] 접근들이었다. 딥러닝을 사용하는 스케치 이미지 기반 3 차원 물체 검색 연구에서는 대부분 3 차원 이미지를 다각도에서 렌더링한 view 이미지들에서 feature 를 추출하고 pooling 하여 하나의 feature 를 출력한다. 그러나 기존의 view pooling 방식은 단순한 평균 방식으로, 부품 이미지에 따른 view 를 반영하기에는 한계가 있었다. 따라서 본 논문에서는 조립 부품 이미지 기반 3 차원 물체 검색을 위해 query 부품 이미지에 따라 다른 view 이미지에 집중할 수 있는 방식의 attentional view pooling 을 제안한다. 또한 조립 부품 데이터의 특성 상 class 당 CAD 모델이 하나인 상황이므로 학습 데이터가 터무니없이 부족하여 이를 해결하기 위한 학습 데이터 증강 방법을 제안한다. 실험은 의자 부품 11 가지에 대해 진행하였고 이를 통해 제안하는 방식의 성능을 입증하였다.

1. 서론

최근 로봇의 가구 조립 작업에 대한 연구가 늘어나고 있다. 몇몇 로봇 가구 조립 연구에서는[2] 조립 설명서를 참고하지 않고 가구 조립을 진행한다. 그러나 로봇이 단독적으로 조립을 진행하지 않고 인간의 가구 조립을 도와주거나, 혹은 로봇의 조립 과정에 인간의 개입이 가능하도록 하는 것 또한 미래지향적인 관점에서 중요하다. 인간친화적인 로봇의 가구 조립 기술에 있어 인간이 이해할 수 있는 조립 설명서를 입력으로 하는 로봇의 가구 조립 기술은 인간의 사고방식과 로봇의 작업 처리능력을 조화롭게 연결하는 기본적인 기술이다.

이를 위해 조립 설명서 이미지를 인식하는 기술이 필요하며 조립 설명서의 부품과 CAD 모델간의 매칭 작업은 필수적인 요소이다. 이는 조립 부품 이미지를 query 로 하여 3 차원 CAD 모델 데이터베이스에서 매칭되는 모델을 검색하는 task 이다.

딥러닝 기반의 feature 추출방식을 이용한 2 차원 이미지와 3 차원 CAD 모델 간의 매칭 연구는 좋은 성능을 거두었다. 그러나 대부분의 연구는 RGB 이미지[5], 또는 흑백 이미지인 스케치 이미지를 다루었다[1]. 스케치 이미지 기반의 연구에서는 비행기, 의자, 자전거 등의 물체들을 다른 클래스로 구별하는 반면, 조립 부품의 경우 의자 다리, 등판, 엉덩이판 등의 비슷한 생김새의 부품들을 각각 다른 클래스로 구별한다. 또한 이 때문에 클래스 하나에는 2개 이상의 CAD 모델을 할당할 수 없어 클래스 당 CAD 모델을 여러 개 구할 수 있는 스케치 기반 검색 기술과는 달리 학습 데이터가 터무니없이 부족한 상황이다. 이 때문에 스케치 이미지 기반의 모델 검색 방법을 조립 부품 데이터에 그대로 적용하기는 쉽지 않다.

조립 부품 기반의 연구를 위해 본 논문에서는 attentional view pooling 방법을 제안한다. 기존의 스케치 이미지 기반

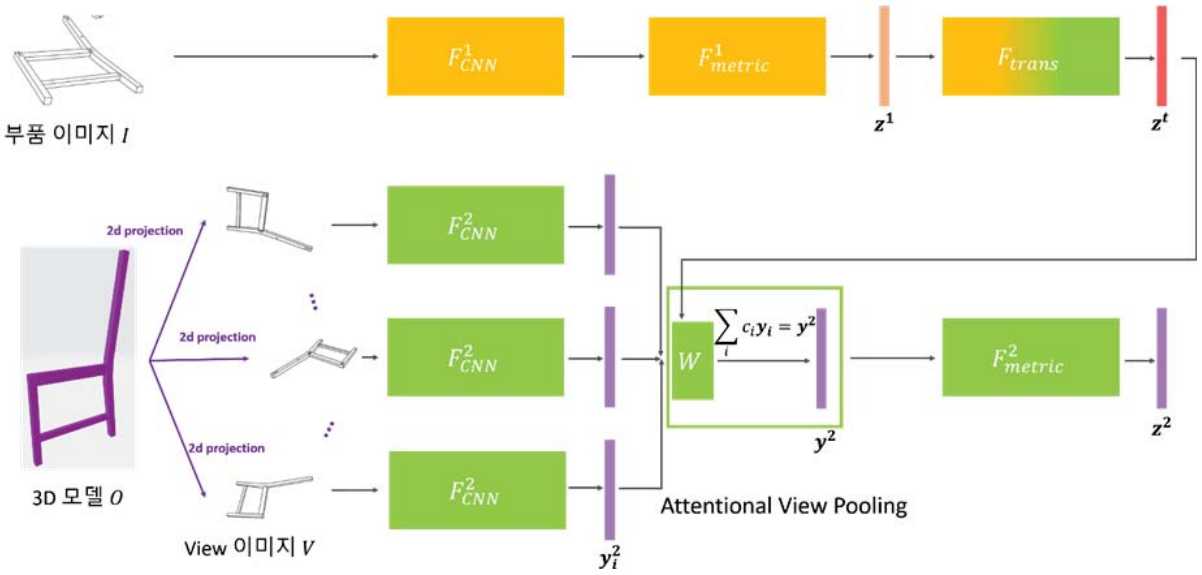

그림 1. 네트워크 구조

전체 네트워크는 부품 이미지의 feature 를 추출하는 모듈(F^1_{CNN}, F^1_{metric}), 3D 모델의 feature 를 추출하는 모듈(F^2_{CNN}, F^2_{metric}), 그리고 부품 이미지 feature 를 3D 모델 도메인으로 변환하기 위해 사용하는 모듈 F_{trans}로 이루어져있다. View pooling 에 사용되는 네트워크 W는 F^2_{CNN}의 출력인 $\mathbf{y}_i^2$들과 $\mathbf{z}^t$를 입력으로 받아 score c_i를 출력한다.

연구에서는[1] 3 차원 CAD 모델에서 다각도로 렌더링한 view 이미지들을 하나의 feature 로 합치기 위해 view pooling 기법을 사용하는데, 이 때 단순한 average pooling 을 사용하였다. 이는 query 스케치 이미지마다 주요하게 다루어야 할 각도가 다름에도 그것을 제한적으로 고려하게 되는 문제가 있다. 그리하여 본 논문에서는 각 view 마다의 가중치(attention)를 추가로 학습하여 평균을 구하는 방식의 attentional view pooling 방법을 제안한다. 더불어 학습 시 한가지의 view 조합이 아닌 랜덤한 12 개의 view 이미지 조합을 사용하여 데이터 증강 효과를 얻었다. 실험은 나무로 만든 IKEA 의자 부품 11 개로 구성된 데이터를 구축하여 진행하였다.

2. 조립 부품 이미지 기반 3 차원 물체 검색 방법

부품 이미지 기반 3 차원 물체 검색은 query 로 주어진 부품 이미지에서 추출한 feature 와 3 차원 물체에서 추출한 feature 들 간의 거리를 비교하여 가까이 위치한 feature 의 3 차원 물체를 반환하는 방식으로 진행한다.

2.1 네트워크 구조

그림 1 에 나와있듯이, 전체 네트워크는 세가지 부분(색으로 구분)으로 나뉘어진다. 먼저 부품 이미지 feature 추출단계에서는 부품 이미지 I 를 입력으로 받아 F^1_{CNN}, F^1_{metric} 을 차례로 지나 feature $\mathbf{z}^1$ 을 출력한다. 다음으로는 3D 모델 O 에서 feature $\mathbf{z}^2$ 를 추출한다. 이 때 3D 모델 데이터를 12 가지 각도에서 렌더링한 이미지들 $\{V_i\}_{i=1}^{12}$ 을 입력으로 받아 F^2_{CNN} 를 지나 중간 feature $\mathbf{y}_i^2$ 를 출력한다. 이 때 12 개의 feature $\mathbf{y}_i^2$ 를 하나의 feature 로 만들어주기 위해 view pooling 과정을 거친다. view pooling 을 거친 feature 를 F^2_{metric}에 통과시켜 3D 모델에 대한 feature $\mathbf{z}^2$를 추출한다. 세번째로, 부품 이미지 feature $\mathbf{z}^1$을 3D 모델 feature 의 도메인으로 adaptation 하기 위해 F_{trans}를 거쳐 feature $\mathbf{z}^t$를 추출하여 $\mathbf{z}^t$와 $\mathbf{z}^2$의 거리를 계산한다.

2.1.1 Attentional View Pooling

본 논문에서는 3D 모델에서 feature 를 추출할 때 다각도에서 렌더링한 이미지 조합을 이용하여 feature 를 추출하는데, 이는 성능이 좋은 기존 스케치 이미지 기반 검색 연구의 방법을 따랐다[1].

3D 모델 하나에 대해 여러 개의 View 이미지를 입력으로 삼아 하나의 feature 를 추출하기 위해서는 pooling 과정이 필요하다. 기존의 접근에서는[1] 각각 feature 들을 단순 평균하는 방식인 average view pooling 방식을 제안하였다. 그러나 이 방식에서는 모든 View 를 같은 중요도로 바라보기 때문에 query 이미지의 자세를 고려한 View feature 를 추출하기 어려워 각 View 이미지의 특성을 반영하기 어렵다.

이를 극복하기 위해 본 논문에서는 Attentional view pooling 을 제안한다. 이는 query 부품 이미지에 따라 어떤 각도의 View 이미지가 중요하게 반영될 지를 예측하여 그 값을 가중치로 두어 가중 평균으로 pooling 하는 방식이다. F^2_{CNN} 를 통과한 feature $\mathbf{y}_i^2$ 들의 가중치를 구하기 위해 행렬 W 를

추가하였다. 가중치는 $\mathbf{z}$ 와 $W \cdot \mathbf{y}_i^2$ 의 내적에 softmax 를 씌운 값으로 정의하였으며, 이는 일반적인 attention mechanism 의 사용하는 방식을 참고하였다[3]. F_{CNN}^2 를 통과한 feature $\mathbf{y}_i^2$ 들을 pooling 하여 $\mathbf{y}^2$ 로 출력하기 위한 과정을 수식으로 나타내면 다음과 같다.

$$c_i = \sigma(\mathbf{z}^T W \mathbf{y}_i) \quad (1)$$

$$\mathbf{y}^2 = \sum_{i=1}^{12} c_i \cdot \mathbf{y}_i \quad (2)$$

$\mathbf{y}_i$ 는 i 번째 View 이미지의 중간 feature, W는 score 를 위해 학습해야 할 행렬, $\mathbf{z}^t$ 는 이미지에서 추출한 최종적인 feature, c_i 는 $\mathbf{y}_i$ 의 중요도, $\mathbf{y}^2$ 는 pooling 된 3D 모델의 중간 feature 를 나타낸다.

2.2 학습 방법

네트워크의 학습은 크게 3 가지로 나뉜다. Feature learning 을 위한 cross-entropy loss 를 이용한 학습, domain adaptation 을 위한 class mean discrepancy loss 를 이용한 학습, 그리고 domain adaptation 을 위한 generative adversarial networks 기반 학습이다.

먼저 feature learning 을 위한 cross-entropy loss 를 $\mathbf{z}^1$, $\mathbf{z}^2$, $\mathbf{z}^t$ 를 출력하는 네트워크(각 F_{CNN}^1 과 F_{metric}^1, F_{CNN}^2 와 F_{metric}^2, 그리고 F_{trans})의 학습에 사용하였다. 학습을 위해 추가적인 분류기(각 C^1, C^2, 그리고 C^t)를 설계하였다. 각 손실 함수를 수식으로 정리하면 다음과 같다.

$$L_{1\_classification} = -\sum_j \widehat{C^1(\mathbf{z}^1)}_{(j)} \log C^1(\mathbf{z}^1)_{(j)} \quad (3)$$

$$L_{2\_classification} = -\sum_j \widehat{C^2(\mathbf{z}^2)}_{(j)} \log C^2(\mathbf{z}^2)_{(j)} \quad (4)$$

$$L_{t\_classification} = -\sum_j \widehat{C^t(\mathbf{z}^t)}_{(j)} \log C^t(\mathbf{z}^t)_{(j)} \quad (5)$$

$\widehat{C^1(\mathbf{z}^1)}_{(j)}$, $\widehat{C^2(\mathbf{z}^2)}_{(j)}$ 와 $\widehat{C^t(\mathbf{z}^t)}_{(j)}$ 는 각각 $\mathbf{z}^1$, $\mathbf{z}^2$, $\mathbf{z}^t$ 의 classification 결과의 추정값을 one-hot vector 로 나타내었을 때의 j 번째 값을, $C^1(\mathbf{z}^1)_{(j)}$, $C^2(\mathbf{z}^2)_{(j)}$, $C^t(\mathbf{z}^t)_{(j)}$ 는 참값을 의미한다.

다음으로 이미지에서 추출한 feature $\mathbf{z}^1$ 을 $\mathbf{z}^2$ 의 도메인으로 adaptation 하기 위해 class mean discrepancy loss 를 F_{trans} 와 F_{CNN}^2, F_{metric}^2 의 학습에 사용하였다. 이는 각 class 마다 이미지 도메인의 변형된 feature 들의 평균과 3 차원 물체 도메인의 feature 들의 평균 간의 거리를 줄이는 것을 목표로 한다.

$$L_{cmd} = \sum_{class} \left\| E_{\mathbf{z}^t \sim p^t(\mathbf{z}^t|class)}[\mathbf{z}^t] - E_{\mathbf{z}^2 \sim p(\mathbf{z}^2|class)}[\mathbf{z}^2] \right\|_2 \quad (6)$$

실제 mini-batch 단위로 학습할 때는 각 class 에 해당하는 feature 들의 평균을 구할 때 해당 batch 에 있는 feature 들의 평균을 구하여 근사한다.

마지막으로 $\mathbf{z}^t$의 학습을 위해 GAN framework 를 도입하였다. F_{trans} 를 생성기로 삼고 추가적인 분류기(D)를 설계하여 real feature 를 $\mathbf{z}^2$ 로, fake feature 를 $\mathbf{z}^t$ 로 삼아 LSGAN loss[5]를 사용하여 학습하였다.

$$L_{gan\_G} = \frac{1}{2} E[(\mathbf{z}^t - 1)^2] \quad (7)$$

$$L_{gan\_D} = \frac{1}{2} E[(\mathbf{z}^2 - 1)^2] + \frac{1}{2} E[(\mathbf{z}^t - 0)^2] \quad (8)$$

각 모델 별 loss function 을 정리하면 다음과 같다.

$$L_1 = L_{1\_classification} \quad (9)$$

$$L_2 = L_{2\_classification} + 0.1\, L_{cmd} \quad (10)$$

$$L_t = L_{t\_classification} + L_{cmd} + L_{gan\_G} \quad (11)$$

$$L_D = L_{gan\_D} \quad (12)$$

네트워크를 학습할 때는 3 번의 pre-training 을 거친 후 최종적으로 전체 모델을 학습하였다. 3 번의 pre-training 은 다음과 같다. 먼저 F_{CNN}^1, F_{metric}^1 을 L_1 을 사용하여 학습하고 F_{CNN}^2, F_{metric}^2 를 $L_{2\_classification}$ 을 사용하여 학습한다. 이 때 attentional view pooling 을 위한 matrix W 를 학습할 수 없으므로 average view pooling 을 사용한다, 다음으로 F_{trans}, D를 각각 L_t, L_D 를 사용하여 학습한다. 마지막 최종 학습 시에는 F_{CNN}^1과 F_{metric}^1, F_{CNN}^2과 F_{metric}^2, D, 그리고 F_{trans}를 차례로 L_1, L_2, L_D, L_t를 최소화하는 방향으로 학습을 진행한다.

2.3 학습 View 이미지 데이터 증강

부품 데이터의 특성 상 한 가지 클래스에는 3D CAD 모델이 하나만 존재한다. 이는 샘플 개수를 늘릴 수 있는 부품 이미지와도 데이터 불균형을 이룰 뿐 아니라 데이터의 개수가 너무 적어 네트워크 학습을 어렵게 한다.

이를 해결하기 위해 본 논문에서는 12 가지였던 실험 단계에서의 카메라 세팅을 48 가지로 늘려, 전체 이미지 중 12 개의 각도를 랜덤하게 조합하여 학습 데이터로 사용하였다. 이를 통해 클래스 당 실제 CAD 모델은 하나지만 여러 개의 3D 모델을 학습에 사용하는 효과를 얻었다.

3. 실험 결과 및 분석

실험 데이터는 11 가지의 부품으로, IKEA 나무 의자인 IVAR 와 STEFAN 의자의 기본 부품 및 중간 산출물로 구성되어 있다. 부품 이미지는 학습 셋으로 660 장(부품 당 60 장), 실험 셋으로 110 장(부품 당 10 장)을 구축하였다. CAD 모델은 학습 셋과 실험 셋 모두 동일하게 총 11 개로 부품 당 1 개이다. 부품

이미지는 blender 로 렌더링한 이미지이다. 실제 IKEA 조립 설명서에서 추출한 이미지를 사용할 경우 부품 별로 한 장의 샘플 밖에 얻을 수 없어 렌더링한 이미지를 실험에 사용하였다.

표 1. 부품 이미지 기반 3 차원 모델 검색 실험 결과

	NN
Average view-pooling	92.73%
Attentional view-pooling	93.64%

실험은 두 가지(average, attention) view pooling 방식에 대해 실험하였고 성능은 nearest neighbor 로 측정하였다. 표 1 에 NN 방식의 정확도를 나타내었다. Attentional view pooling 방식의 네트워크가 기존의 Average view pooling 방식의 네트워크보다 성능이 높음을 보였다.

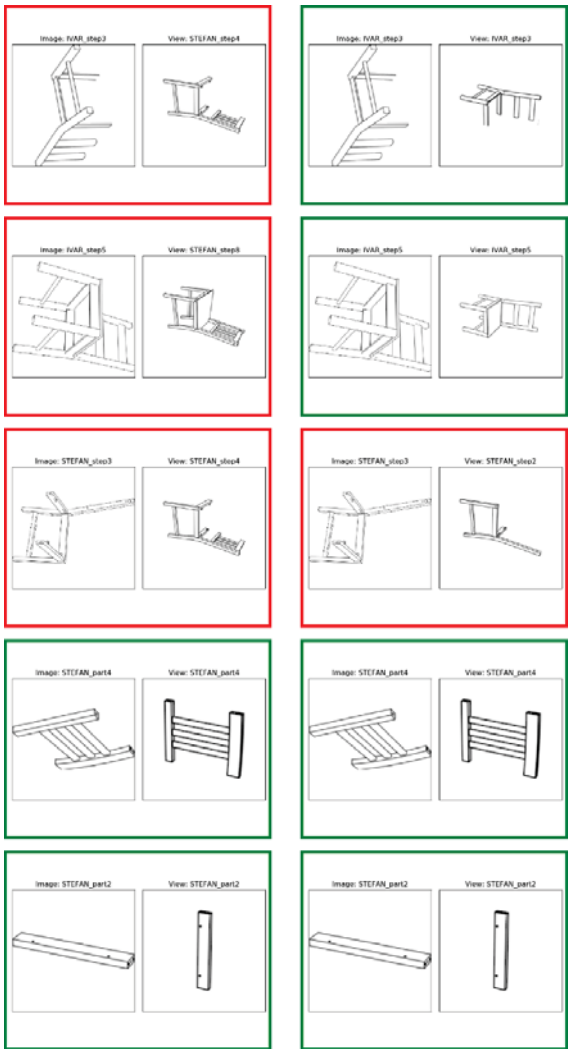

그림 2. 부품 이미지와 CAD 모델 매칭 실험 예시
average view-pooling(왼), attention view-pooling(오른)

또한 그림 2 에 각 방식의 결과 예시를 나타내었다. 왼쪽은 average view pooling 방식의 결과이며 오른쪽은 attentional view pooling 방식의 결과이다. 윗줄의 두 샘플(IVAR_step3, IVAR_step5)에 대해 attentional view pooling 방식에서 좋은 성능을 거두었다. 그러나 세번째 샘플(STEFAN_step3)에 대해서는 두 방식 모두 맞추지 못하였지만 사람의 눈으로 보기에 오른쪽 결과가 query 부품과 상대적으로 비슷해 보이는 점에서 성능 향상이 존재한다고 볼 수 있다. 그러나 아직 occlusion 이 심한 이미지나 STEFAN_step3 과 STEFAN_step2 처럼 비슷한 모양의 샘플에 대해서 성능을 개선할 필요가 있다.

4. 결론

본 연구에서는 조립 부품 이미지 기반의 3 차원 물체 검색을 위하여 부품 이미지에 따라 attention 을 다르게 주는 새로운 방식의 view pooling 기법을 제안하였으며, 학습 데이터 증강을 통해 네트워크의 성능을 높였다. 실험을 통해 제안하는 방식의 효과를 입증하였다. 추후 occlusion 이 심한 부품 이미지 데이터에 대한 연구도 진행할 계획이다.

감사의 글

이 논문은 2020 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. NI190004, 사람을 위한 조립 설명서를 이해하고 조립작업 계획을 생성하는 AI 기술 개발)에 의하여 지원되었음.

참고문헌

[1] Jiaxin Chen and Yi Fang, “Deep Cross-modality Adaptation via Semantics Preserving Adversarial Learning for Sketch-based 3D Shape Retrieval”, 15th European Conference on Computer Vision (ECCV), 2018.

[2] Jiaxin Chen and Yi Fang, “Deep Cross-modality Adaptation via Semantics Preserving Adversarial Learning for Sketch-based 3D Shape Retrieval”, 15th European Conference on Computer Vision (ECCV), 2018.

[3] Francisco Suarez-Ruiz *et al*., “Can robots assemble an IKEA chair?”, Science Robotics, vol 3, 2018.

[4] Minh-Thang Luong *et al*., “Effective Approaches to Attention-based Neural Machine Translation”, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2015.

[5] Manolis Savva *et el.*, “SHREC’16 Track Large-Scale 3D Shape Retrieval from ShapeNet Core55”, Eurographics Workshop on 3D Object Retrieval, 2016.

[6] Xudong Mao *et el*., “Least Squares Generative Adversarial Networks”, IEEE international Conference on Computer Vision (ICCV), 2017.