CycleGAN을 이용한 비지도학습 기반 인체 검출

Unsupervised Human Segmentation with Cycle Consistent Adversarial Networks

반송하[O1,2] 김병희[2]

sb5449@nyu.edu bhkim@surromind.ai

[1] New York University Shanghai

[2] 써로마인드로보틱스

연구 동기

문제 : 모션 데이터 부족

필요한 모션데이터 구축 시
- 키넥트 등 장비 필요
- RGB이외의 정보 필요
- 지도학습으로 신체단위 인식 필요

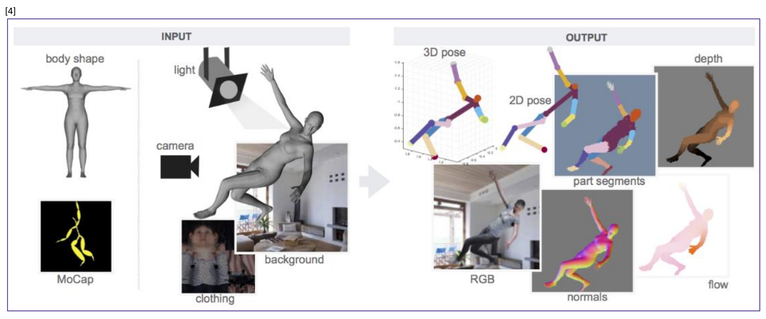
수요 : 부가적인 장비나 RGB값 이외의 정보가 존재하지 않을 때 이미지로부터 자세추정 혹은 영상으로부터의 동작인식

본 연구는 아직 자세 추정 단계에는 미치지 못했지만, CycleGAN을 이용한 인체 검출 실험을 통해 얻은 결과로 자세추정으로의 적용 가능성을 뒷받침하고자 한다.

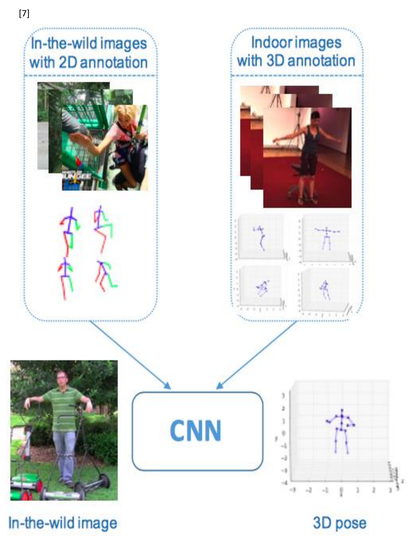
관련 연구

현존하는 연구들은 RGB 이외 depth 정보를 이용하거나 지도학습을 통해 신체 인식 후 자세정보 추출

[4]



[7]



모션데이터 부족 문제를 극복하기 위해 SURREAL DATASET이라 불리는 새로운 인위적인 데이터를 만든 연구가 있다[4]. 이는 동작인식 데이터에 대한 접근성을 크게 높였다는 점에서 의의가 크다. 하지만 현재는 공개된 모델을 통해 동작인식 데이터에서 인조 영상을 생성하는 단방향만 가능한 상태이다.

이외에 CNN을 통해 자세추정을 한 연구들이 여럿 있으나 모두 지도학습 또는 준지도학습에 기반한 것으로, labeling된 데이터 없이는 학습을 진행할 수 없다는 한계점이 있다. 무작위 데이터를 바탕으로 완전한 비지도학습을 통해 자세추정을 할 수 있다면 동작인식 데이터의 범위를 훨씬 넓힐 수 있을 뿐만 아니라 컴퓨터 비전분야의 다른 영역에서도 다양하게 적용될 수 있을 것이다.

인체 검출 실험

신경망 구조 / 데이터

$$\mathcal{L}_{\text{GAN}}(G_{AB}, D_B, A, B) = \mathbb{E}_{b\sim Pdata(b)}[\log D_B(b)] + \mathbb{E}_{a\sim Pdata(a)}[\log(1 - D_B(G_{AB}(a)))]$$

($\mathbb{E}_{b\sim Pdata(b)}[(D_B(b)-1)^2]$, $\mathbb{E}_{a\sim Pdata(a)}[(D_A(G(a))-1)^2]$)

$$\mathcal{L}_{\text{cyc}}(G_{AB}, G_{BA}) = \mathbb{E}_{a\sim Pdata(a)}[\|G_{BA}(G_{AB}(a)) - a\|_1] + \mathbb{E}_{b\sim Pdata(b)}[\|G_{AB}(G_{BA}(b)) - b\|_1]$$

$$\mathcal{L}(G_{AB}, G_{BA}, D_A, D_B) = \mathcal{L}_{\text{GAN}}(G_{AB}, D_B, A, B) + \mathcal{L}_{\text{GAN}}(G_{BA}, D_A, B, A) + \lambda\mathcal{L}_{\text{cyc}}(G_{AB}, G_{BA})$$

$$G_{AB}, G_{BA} = \arg\min_{G_{AB},G_{BA}} \max_{D_A,D_B} \mathcal{L}(G_{AB}, G_{BA}, D_A, D_B)$$

모델		[6]에 명시된 모델	본 실험 모델
생성모델 (Generator)	Encoder	Conv 2개	Conv 3개
	Transformation	Residual blocks 9개	Residual blocks 9개
	Decoder	Conv 2개	Conv 3개
판별모델 (Discriminator)		70 x 70 PatchGANs	32 x 32 PatchGANs

데이터 A

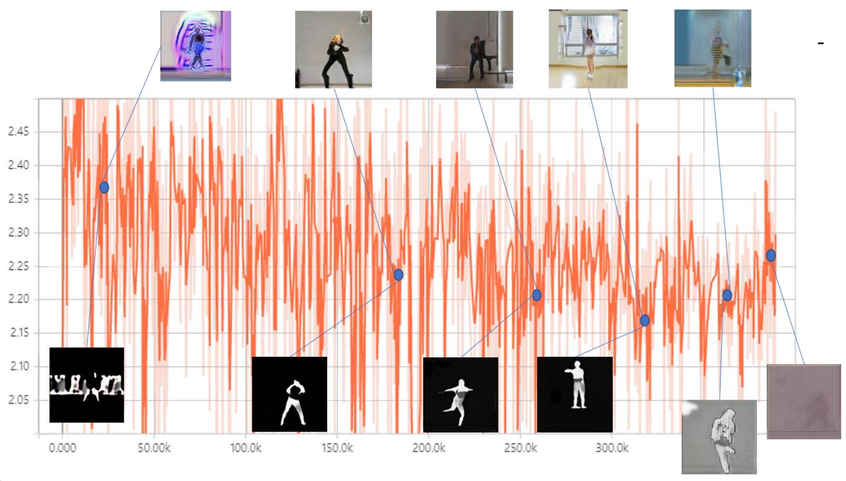
유튜브 댄스 영상 15개에서 10초간격으로 2,291개 프레임 추출.

데이터 B

SURREAL DATASET에서 segmentation 정보만 5,535개 추출하여 흑백으로 이미지화

학습 과정

- 학습용데이터 : 시험용데이터 = 9 : 1
- 반복학습 : 200회



- 학습이 진행될수록 손실(loss)이 감소하는 추세였으나, 총 손실이 학습 후반부에 갑자기 증가
- 반복학습 횟수에 따른 학습 상황
  - 50회 : 인체뿐만 아니라 배경에 속하는 물건들을 함께 분할
  - 100회 : 인체 윤곽을 보다 정확하게 인식하지만 신체단위를 구분하지는 못함.
  - 150회 : 제법 정확하게 인체 검출, 신체 의미단위 인식. (하체부위 어둡게 표시)
  - 170회 이상 : 모드붕괴(mode collapse). 이미지의 형태가 완전히 파괴되거나 심한 노이즈 발생.

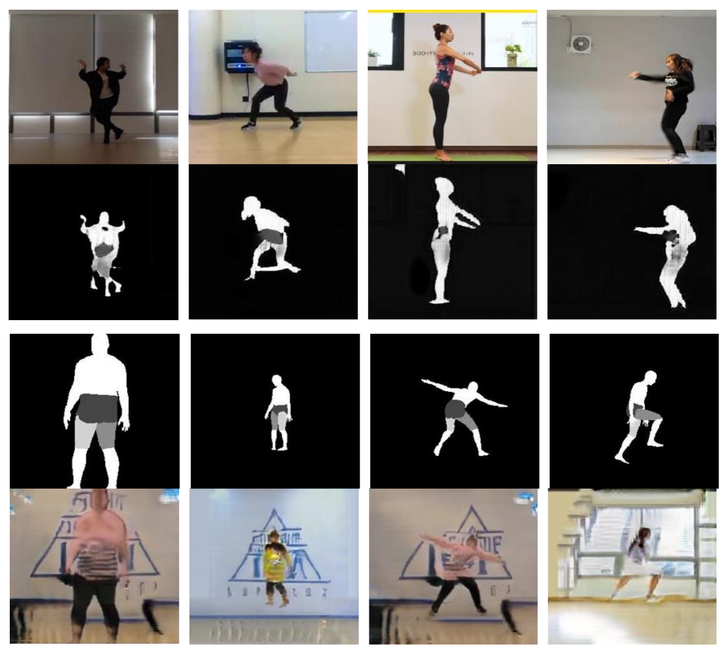
결과 분석

적정 학습 횟수로 보여지는 160회에서 생성된 결과를 확인했을 때,

A → B'
- 비교적 정확한 인체 검출.
- 하체부위는 어둡게 잘 표시됨.
  ⇒ 신체 의미단위 구분 가능

B → A'
- 얼굴이 세밀하지 못하고 배경이 무너짐.
- 인체가 바닥으로부터 떠다니는 현상 발생.
- 머리, 팔, 몸통, 다리 등 신체단위 구분하여 새로운 이미지 적절하게 생성 가능.

∴ 생성 모델과 판별 모델이 적대적으로, 그리고 주기 순환적으로 학습하면서 단순히 인체라는 의미특징을 학습할 뿐만 아니라 학습에 필요한 신체의 의미단위를 구분해 낼 수 있음이 확인됨.



Real Data A
실제 이미지

Fake Data B'
위 이미지에 대한 인체검출 마스크

Real Data B
실제 인체검출 마스크

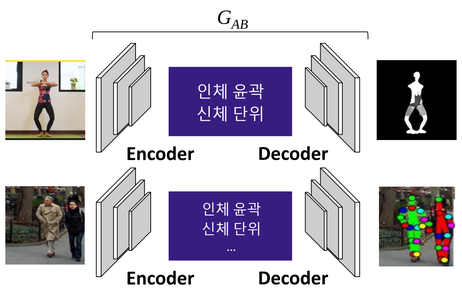
Fake Data A'
인체검출 마스크 기반 생성된 이미지

한계점

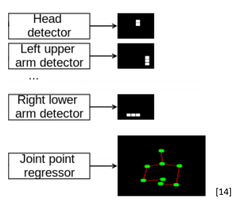
- 사람 1명에 대해 자세 추정 하는 것을 염두에 둔 관계로 데이터가 매우 제한적.
- 신경망 자체가 매우 복잡하고 깊기 때문에 계산비용이 높고 다른 지도학습 기반 이미지 segmentation 알고리즘에 비해 비효율적.
- 다른 연구들은 보다 다양한 종류의 데이터에서 객체를 검출하는 작업을 했다는 점에서, 객체 종류를 인체로 한정시킨 본 실험과는 결과 비교 어려움.

자세추정 적용 가능성

연구 [2]에서 인체 bounding box 및 신체 부위 검출을 바탕으로 자세추정에 성공한 것을 미루어 볼 때, 비지도학습을 통한 인체 검출은 비지도학습 기반 자세추정에 큰 기반이 되는 실험 결과라 여겨진다.



- 제한된 정보의 인체검출 마스크에서 RGB 사람 이미지 생성.
  ⇒ 생성모델 G_{AB}, G_{BA} 내의 인코딩, 디코딩 하는 과정에서 추출된 잠재적 특징이 인체의 모양과 부위별 특징을 잘 함축.
- 데이터B에 인체 검출 정보 대신 관절 위치 정보를 입력
  ⇒ 생성모델이 학습한 잠재적 특징이 신체 단위 위치정보를 포함한다면 자세 추정 가능할 것.
- 실제 현존하는 정확한 자세추정 방법 중 하나가 joint point regression [14]



다만, 관절 정보는 인체 검출 마스크에 비해서도 훨씬 더 제한된 정보만을 포함하고 있기 때문에, G_{BA}의 학습뿐만 아니라 전체적인 생성모델에서 의미 있는 잠재적 특징을 추출하는 것이 어려울 수 있다. 이 경우 G_{AB}에서 B′ 이외에 추가적인 정보를 함께 생성하여 함께 학습을 하되, 판별 모델에는 B′만 입력해주는 방법을 해결방안으로 생각해 볼 수 있다.

결 론

본 논문에서는 CycleGAN이라는 강력한 생성적 적대신경망을 통해 비지도학습을 기반으로 이미지로부터 인체를 검출한 실험 과정과 결과에 대해 분석하였고, 같은 신경망을 통해 비지도학습 기반 자세추정의 가능성을 제안하였다. 인체 검출을 통해 신경망이 RGB 정보만으로 비지도학습으로 신체의 의미단위를 구분할 수 있음을 확인하였고, 같은 원리로 자세 추정을 수행하는 것이 향후 연구과제이다.

참고문헌

[2] F. Xia, P. Wang, X. Chen, and A. Yuille. Joint multi-person pose estimation and semantic part segmentation. *axXiv preprint arXiv:1708.03383*, 2017.

[4] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *CVPR*, 2017.

[6] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *ICCV*, 2017.

[7] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei. Towards 3D human pose estimation in the wild: a weakly-supervised approach. *arXiv preprint arXiv:1704.02447v2*, 2017.

[14] S. Li, Z.-Q. Liu, and A. B. Chan. Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network. *International Journal of Computer Vision*, 113(1):19-36, 2015.