CycleGAN을 이용한 비지도학습 기반 인체 검출

반송하[O1,2] 김병희[2]
[1]뉴욕대학교 상하이
[2]써로마인드로보틱스
sb5449@nyu.edu, bhkim@surromind.ai

Unsupervised Human Segmentation with Cycle Consistent Adversarial Networks

Songha Ban [O1,2] Byoung-Hee Kim[2]
[1]New York University Shanghai
[2]Surromind Robotics

요 약

사람의 동작을 인식하고 자세를 추정하기 위한 연구가 많이 진행되었지만 대부분 RGB정보 이외에 깊이(depth) 정보를 추가로 수집하거나 ground truth가 있는 상태에서 지도학습을 통해 연구가 진행되었다. 본 연구는 이미지에 상응하는 모션 데이터가 존재하지 않을 때 비지도학습을 통해 자세 추정을 하는 방법을 모색하고자 CycleGAN이라는 생성적 적대신경망을 이용하였고, 1차적으로 비지도학습에 기반하여 이미지로부터 인체를 검출하는데 성공하였다. 이 과정에서 신경망이 신체의 의미단위를 구분하는 것이 가능하다는 점을 발견했으며 이는 비지도학습을 통한 영상으로부터의 자세 추정에 적용 가능함을 보인다.

1. 서 론

인체가 표현할 수 있는 동작은 무한하고 이를 나타내는 모션 데이터는 다양한 분야에 쓰일 수 있다. 하지만 현존하는 모션 데이터는 매우 한정적이고 고유한 정보를 얻기 위해서는 키넥트(Kinect) 등의 기타 장비를 사용해야하는 등 원하는 분야의 데이터를 수집하기가 쉽지 않은 실정이다. 예를 들어 음악에 맞는 춤을 생성하기 위한 모델을 학습시켜야 하는 경우, 음악과 함께 춤추는 모션 데이터를 찾기 어렵기 때문에 키넥트를 이용하여 댄서가 춤추는 모습을 녹화해야 하는 번거로움이 있다[1]. 부가적인 장비나 별도의 정보 없이 영상의 RGB데이터로만 자세추정(pose estimation)이 가능하다면 인체 동작과 관련된 다른 연구를 진행하는 데에 매우 유용할 것이다.

컴퓨터비전 분야에서는 단순 이미지나 영상에서 자세추정을 하는 연구[2,3,4]들이 다양하게 진행된 바 있다. 하지만 대부분은 지도학습을 통해 신체를 부위별로 인식한 후 자세 정보를 추출하는 방식으로, 지도학습이나 준지도학습에 속한다. 데이터 부족 문제를 해결하기 위해 synthetic한 데이터를 만드는 연구 또한 진행되었으나 RGB데이터 이외에 추가 정보를 많이 들여 데이터를 생성한 것을 확인할 수 있다. 본 연구는 RGB값 이외의 정보가 존재하지 않을 때 이미지로부터 자세추정 정보를 얻고자 하는 수요에서 비롯되었고, CycleGAN(Cycle Consistent Adversarial Networks)[6]을 통해 이를 해결할 수 있다고 가정하였다.

본 연구는 아직 자세 추정 단계에는 미치지 못했지만, CycleGAN을 이용한 인체 검출 실험을 통해 얻은 결과로 자세 추정으로의 적용 가능성을 뒷받침하고자 한다. 실험 결과를 통해 해당 신경망이 인체 검출을 잘 해낼 뿐만 아니라 학습에 필요한 신체 단위 또한 구분한다는 점을 알 수 있다. 같은 원리로 신경망이 신체 단위를 잘 함축 시킬 수 있도록 신경망을 디자인 한다면, 비지도학습 기반 관절 인식과 자세 추정 또한 충분히 가능할 것이다.

2. 관련연구

모션데이터 부족 문제를 극복하기 위해 SURREAL DATASET이라 불리는 새로운 인위적인 데이터를 만든 연구가 있다[4]. 이들은 여러 프레임의 연속으로 된 3D 동작인식(motion capture) 데이터를 기반으로 실제 사람이 움직이는 영상을 생성하여 공개하였다. 이는 동작인식 데이터에 대한 접근성을 크게 높였다는 점에서 의의가 크다. 하지만 현재는 공개된 모델을 통해 동작인식 데이터에서 인조 영상을 생성하는 단방향만 가능한 상태이다.

이외에도 합성곱 신경망(Convolutional Neural Networks)을 통해 단순 이미지에서 지도학습 기반으로 자세추정을 한 연구[3], 지도학습 기반 인체 부위별 검출을 통해 2명 이상의 사람에 대한 자세추정을 한 연구[2,7], 무작위한 실외 배경에서의 자세추정을 바탕으로 3D 동작인식 데이터를 생성하는 연구[8] 등이 존재하고, 자세추정을 정확하게 하기 위해 여러가지 방법들이 고민되고 발전되어 왔음을 알 수 있다. 하지만 이들은 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.

3. 적대신경망 활용 인체 검출

3.1. 신경망 구조

본 실험을 위해 사용된 학습 모델은 CycleGAN [6]이다. A를 바탕으로 B'를 생성하는 $G_{AB}(A)$, 그리고 B'가 진짜 B인지 혹은 생성된 B'인지를 구별하기 위한 $D_B(B')$가 있다고 할 때, G_{AB}는 B'가 D_B에서 진짜 B로 인식될 수 있도록, D_B는 실제 B만을 진짜로 판별하고 생성된 B'는 가짜로 판별할 수 있도록, 적대 손실(adversarial loss)를 최소화 해야 한다. B'를 생성하는 과정에서 A의 특징 또한 유지할 수 있도록 B'에서 다시 본래 입력 이미지인 A를 생성하는 $G_{BA}(B') = A'$를 학습시키기 위해 A'와 A의 차이를 나타내는 순환일관성

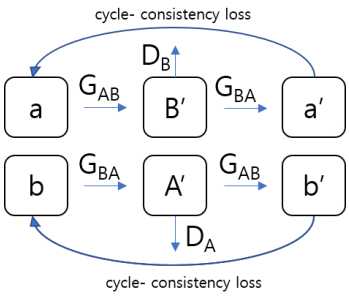

그림 1. CycleGAN 구조

손실(cycle-consistency loss) 또한 최소화 해야 한다. 각 손실은 다음과 같이 표현된다.

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

$$\mathcal{L}_{cyc}(G_{AB}, G_{BA}) = E_{a\sim Pdata(a)}[\|G_{BA}(G_{AB}(a)) - a\|_1] + E_{b\sim Pdata(b)}[\|G_{AB}(G_{BA}(b)) - b\|_1]. \quad (2)$$

이 과정을 거쳤을 때 부족제한(under-constrained) 문제가 있기 때문에, 변수를 공유하는 같은 생성, 판별 모델을 역방향으로 적용하여 같이 학습시키게 된다. 이러한 방식으로 CycleGAN[6]은 쌍으로 이루어지지 않은 이미지들 속에서 한 이미지의 의미특징(semantic feature)은 유지하면서 다른 이미지의 부가적인 특징을 추출하여 스타일이 전환된 새로운 이미지 생성이 가능하다. 본 실험 또한 사람 이미지에 대응되는 인체 검출 마스크가 존재하지 않기 때문에 쌍으로 이루어지지 않은 데이터이고, 스타일 전환 하듯이 이미지로부터의 인체 검출이 가능하다고 전제하였다.

신경망 구조는 표 1에서 볼 수 있듯이 [6]의 구조에서 미세하게 변형시켜 구현하였다. 생성모델은 [6]에서 명시된 바와 같이 instance normalization[9]을 사용하였고, 안정적인 학습을 위해 적대손실은 negative log likelihood를 least square loss로 대체하였다. 판별모델은 PatchGAN[10,11]을 이용하여 만들어진 32x32 크기의 출력물에서 샘플이 진짜인지 가짜인지를 판별하게 된다.

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

표 1. 연구[6]과 본 실험에서 사용된 신경망 구조

3.2. 데이터

실험에 사용된 데이터는 크게 두 종류로 나뉜다. 신경망에서 A에 해당하는 실제 사람 사진과, B에 해당하는 인체 검출 마스크이다. 데이터 A는 유튜브에 올라와 있는 아이돌의 댄스 영상, 일반인들의 댄스 영상 등 15가지 동영상에서 프레임을 추출하였다. 데이터는 배경은 실내이고 사람이 1명인 경우로 제한하였으며, 10초간격으로 총 2,291장의 프레임을 추출하여 동작이 겹치지 않을 수 있게 하였다. 데이터 B로는 synthetic하게 생성된 SURREAL DATESET[4]에서 본 연구에 필요한 segmentation 정보만 추출하여 총 5,535개를 이미지화 하였다. B 데이터에는 인체 이외에 배경은 검정색으로 표시되며, 그림 5에서 볼 수 있듯이 하얗게 표시된 인체중에서도 하체의 특정 부위는 어두운 색으로 표시된 것을 확인할 수 있다. A, B 이미지 데이터 모두 가로, 세로 256픽셀로 크기를 조정하였고 이외에는 별도의 전처리과정 없이 이미지 그대로 RGB정보만을 갖고 학습을 진행하였다.

3.3. 학습과정

학습용 데이터와 시험용 데이터를 9:1의 비율로 나누어 학습을 진행하였고, 총 200회 반복학습을 시켰다. 약 50회 인체뿐만 아니라 배경에 속하는 창문이나 기타 물건을 함께 분할하는 경우가 생겼다. 반복학습을 100회 이상 진행했을 때, 인체 윤곽은 보다 정확해지고 인체를 보다 정확하게 검출하는 듯하지만, 어둡게 표시되어야 할 하체 대신 상체나 기타 부위가 강조되기도 하였다. 150회 이상 학습이 진행되었을 때에는 제법 정확하게 인체를 검출할 뿐만 아니라 신체 부위까지 적절하게 표시하는 모습을 보였다. 하지만 학습이 170회가 넘어가자, 이미지의 형태가 완전히 파괴되어 선명하게 인체를 그려내는데 실패하고 노이즈가 많이 추가되는 현상이 나타났다.

학습 초반부터 완료시까지의 손실(loss) 변화를 분석했을 때, 모든 손실이 학습이 진행될수록 감소하는 추세를 보였으나, 총 손실이 학습 후반부에 갑자기 증가하는 것이 확인되었다. 이 때 생성된 결과를 확인했을 때 이미지의 형태를 유지하지 못한 것으로 보아 모드붕괴(mode collapse)[12]가 일어난 것으로 확인된다. 다음 그래프는 학습 진행에 따른 손실의 변화와 주요 지점별 생성된 샘플이다.

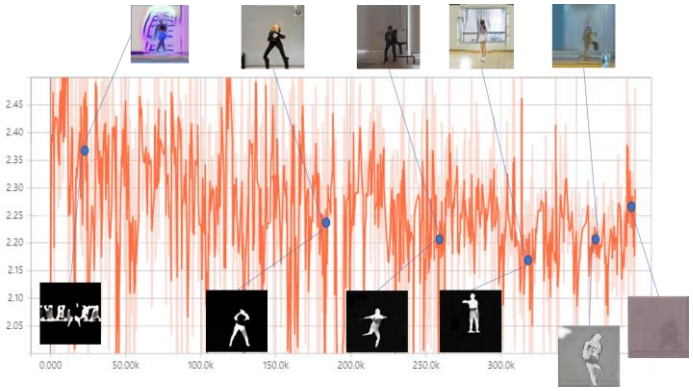

그림 2. 학습 진행에 따른 손실 변화 및 생성 샘플

3.4. 결과 분석 및 한계점

모드붕괴가 일어나기 전, 적정 학습 횟수로 보여지는 160회에서 생성된 결과를 확인했을 때 인체 검출이 비교적 정확하게 이루어진 것을 확인하였으며, 하체부위가 어둡게 잘 표시된 점 또한 의미 있는 결과로 보여진다. 인체 검출 마스크에서 실제와 같은 이미지를 생성하는 역방향 결과를 확인했을 때에도, 인체 검출 마스크에 표시된 동작, 크기에 맞게 입력 데이터 A 같은 이미지가 생성되었다. 그림 6의 결과를 보면 얼굴이나 인체가 세밀하고 부드럽지 못하고, 배경이 무너지거나 인체가 바닥으로부터 떨어져 떠다니는 현상이 발생한다. 하지만 기본적으로 머리, 팔, 몸통, 다리 등 신체 단위가 잘 구분되어 새로운 이미지가 적절하게 생성된 것을 확인할 수 있다. 이를 통해 양방향 생성 모델과 판별 모델이 적대적으로, 그리고 주기 순환적으로 학습하면서 단순히 인체라는 의미특징을 학습할 뿐만 아니라 학습에 필요한 신체의 의미단위를 구분해 낼 수 있음이 확인되었다.

본 실험은 사람 1명에 대해 자세 추정 하는 것을 염두에 두고 진행된 관계로 데이터가 매우 제한적이다. 또한 segmentation 관점에서만 봤을 때에는 사용된 신경망 자체가 매우 복잡하고 깊기 때문에 계산비용이 높고 다른 지도학습 기반 이미지 segmentation 알고리즘에 비해 효율적이지 못하다는 한계점이 있다. 또한 [13]와 같은 다른 연구들은 보다 다양한 종류의 데이터에서 객체를 검출하는 작업을 했다는 점에서, 객체 종류를 인체로 한정시킨 본 실험과는 결과 비교를 하기 어려운 점이 있다. 하지만 비지도학습 기반으로 인체 검출을 이뤄냈다는 것과 신체가 의미단위별로 잘 학습되었다는 점에 큰 의의가 있으며, 검출 결과의 질만 보았을 때에도 준지도학습 기반으로 이미지 분할을 수행한 [13] 결과의 질에 뒤떨어지지 않음을 확인할 수 있다.

그림 3. 실제 데이터 A: 신경망에 input으로 쓰인 실제 사람 이미지

그림 4. 생성된 데이터 B': 반복학습 160회 이후 그림 2의 이미지로부터 생성된 인체 검출 마스크

4. 자세추정으로의 적용 가능성

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

본 연구에서는 인체 검출을 통해 비지도학습을 통해서도 신경망이 신체단위를 구분할 수 있음을 입증했다. 인체 검출 마스크는 색, 배경, 사람에 관한 세부정보 등이 부재한 상태로 RGB이미지에 비해 제한된 정보만을 갖고 있지만, adversarial loss와 cycle-consistency loss 최소화를 통해 그 이상의 정보를 포함하는 실제 이미지를 충분히 생성할 수 있다. 이는 생성모델 G_{AB}와 G_{BA}에서 인코딩, 디코딩 하는 과정에서 추출된 잠재적 특징이 인체의 모양과 부위별 특징을 잘 함축하고 있다고 말할 수 있다. 같은 방식으로, 인체 검출 마스크 대신 관절의 위치를 표시하는 정보를 데이터 B로 입력했을 때, 생성모델이 학습한 잠재적 특징이 이미지 A에서의 인체 위치와 신체 부위별 정보를 포함하고 있기만 한다면 관절 인식을 통한 자세 추정이 가능해질 것이다. 실제로 현존하는 자세추정 모델 중 가장 정확한 방법 중 하나가 지도학습 기반으로 관절인식을 먼저 수행한 후 joint point regression[14]을 통해 자세추정 정보를 추출하는 것이다. 따라서 비지도학습 기반 관절인식이 가능할 때 이를 기반으로 높은 질의 자세추정 정보를 추출할 수 있을 것으로 예상된다.

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 문제점에 대해서는 G_{AB}에서 B' 이외에 추가적인 정보를 함께 생성하여 함께 학습을 하되, 판별 모델에는 B'만 입력해주는 방법을 해결방안으로 생각해 볼 수 있다. 현재 비지도학습 기반 관절인식에 대해서는 추가정보 생성 없이 관절 위치 정보만 갖고 실험 진행중에 있으며, 위와 같은 문제가 발생할 경우 B'를 생성할 때 추가정보를 생성하고 함께 학습을 진행하여 결과를 지켜볼 예정이다.

5. 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 비지도 학습으로 신체의 의미단위를 구분할 수 있음을 확인하였고, 같은 원리로 자세 추정을 수행하는 것이 향후 연구 과제이다.

그림 5. 실제 데이터 B: 신경망에 또다른 input으로 쓰인 인체 검출 이미지

그림 6. 생성된 데이터 A': 반복학습 160회 이후 그림 4의 인체 검출 마스크로부터 생성된 이미지

감사의 글

이 논문은 2017년도 산업통상자원부의 재원으로 IITP의 지원(10069126)과 과학기술정보통신부의 재원으로 KEIT의 지원(2017-0-00162)을 받아 수행된 연구임.

참고문헌

[1] O. Alemi, J. Francoise, and P. Pasquier. GrooveNet: Real-time music-driven dance movement generation using artificial neural networks. In Workshop on *Machine Learning for Creativity*, 2017.

[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.

[3] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional Pose Machines. In *CVPR*, pages 4724-4732, 2016.

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

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

[6] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *CVPR*, pages 7291-7299, 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.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770-778, 2016.

[9] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[10] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[11] C. Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. *ECCV*, 2016.

[12] I. Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *axXiv preprint arXiv:1701.00160v4*, 2017.

[13] E. Haller, and M. Leordeanu. Unsupervised object segmentation in video by efficient selection of highly probable positive features. *ICCV*, pages 5095-5093, 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.