딥러닝 기반의 Human Part Segmentation 기술을 활용한 체형 변화 추적 서비스

이나경 ˚, 최윤종, 박상훈, 손종수†
CJ 올리브네트웍스 AI Core 연구소

Body Progress Tracking Service using Deep Learning-based Human Part Segmentation

Na-Kyung Lee ˚, Yun-Jong Choi, Sang-Hoon Park, Jong-Soo Sohn†
CJ Olivenetworks AI Core Research
E-mail : nk.lee3@cj.net, yunjong.choi@cj.net, sanghoon.park6@cj.net, jongsoo.sohn@cj.net

요 약

최근 COVID-19의 확산으로 체육시설 방문이 어려워짐에 따라 홈트레이닝 시장이 크게 성장하고 있다. 홈트레이닝의 경우, 대개 스스로 운동의 진행을 점검해야 하므로 이를 도와줄 수 있는 서비스가 필요하다. 특히 운동하는 사람이 신체 변화를 이미지로 기록하면서 운동의 효과를 체감하는 '눈바디' 루틴은 매우 중요한데, 이를 체계적으로 기록하고 분석하기는 쉽지 않다. 이에, 본 연구는 사용자에게 AI 기반의 촬영 가이드를 제공하고 딥러닝 기반의 Human Part Segmentation 기술을 활용하여 체형을 비교·분석한 결과를 지원하는 서비스를 개발함으로써, 기존의 '눈바디' 루틴이 가진 문제를 개선하였다. 본 연구는 Image Segmentation 기술로 체형의 변화를 수치적으로 표현했다는 점에서 의미가 있으며, 향후 홈트레이닝 분야에서 활용될 것이라 기대한다.

1. 서론

COVID-19의 영향으로 외부 활동이 제한되면서 집에서 운동하는 인구가 늘고 있다. 흔히 집에서 하는 운동을 홈트레이닝(At-home Workout)이라고 하며, 이는 체육관에서 트레이너에게 받는 트레이닝과 구별된다. 홈트레이닝 환경에는 일반적으로 몸 상태를 체크해줄 수 있는 트레이너가 없기 때문에, 운동하는 사람 스스로 운동의 효과를 점검할 필요가 있다. 운동의 효과를 확인하는 방법의 하나로, 주기적으로 신체의 사진을 찍고 변화를 체크하는 일명 '눈바디'라는 루틴이 대표적이다. 하지만 기존의 '눈바디' 루틴은 사진 촬영 작업과 비교 작업을 모두 인간이 직접 수행한다는 점에서, 1) 촬영의 일관성을 유지하기 어렵다는 문제, 2) 이로 인해 비교 시 편향이 생기는 문제, 그리고 3) 비교하는 작업의 번거로움이 루틴의 지속을 저해한다는 문제가 있다.

본 연구에서는 이러한 문제를 개선할 수 있도록,

촬영 가이드 제공 및 픽셀 단위의 체형 비교·분석을 포함하는 딥러닝 기반의 체형 변화 추적 서비스를 제안한다. 이는 Image Segmentation을 기반으로 구현되었으며, 특히 신체를 부위별로 분할하는 Human Part Segmentation 기술을 활용하여 체형의 변화를 정량적으로 표현하고 분석한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문과 관련된 연구를 서술하며, 3장에서는 딥러닝 기반의 체형 변화 추적 서비스를 제안한다. 4장에서는 실험결과를 제시하고, 5장에서는 본 논문의 결론을 정리한다.

2. 관련 연구

2.1. Human Shape Estimation (체형 예측)

최근의 Human Shape Estimation은 인간의 모습이 담긴 2차원 싱글 이미지로부터 3차원의 체형을 예측[3, 4]하는 문제를 주로 다룬다. H. Zhu et al.[3]은 인간의 Mesh 이미지를 추론한 후, 관절, 실루엣, 음영 정보를 추가로 학습하여 형태를 보정한다. S. Saito et al.[4]은 내포된 Geometry 정보를 학습하여 생략된 부분을 예측해낸다. 이러한 3차원 예측은 공통적으로 깊이 정보를 정확하게 복구할 수 없기 때문에 오차가 크다는 단점이 있고, 빠르게 추론해야 하는 환경에 적합하지 않다. 이에 본 논문에서는 불확실한 깊이 정보 추론은 고려하지 않았고, 추론 시간이 빠른 2차원 면적계산으로 체형을 근사하는 Human Part Segmentation 방법을 사용한다.

2.2. Human Part Segmentation (신체부위 이미지 분할)

Human Part Segmentation은 이미지에서 픽셀 단위로 머리, 얼굴, 팔, 다리 등의 신체 부위와 옷(opt.)을 의미론적으로 분할하는 것을 말한다. Lu Yang et al.[6]은 Mask R-CNN[5]의 FPN 구조와 RoIAlign 컨셉을 따르며, RoI의 크기를 키워 정확도를 높이는 데에 기여했다. 또한, Pseudo-labeling[2, 7]에 대한 연구가 이루어졌고, Kevin Lin et al.[9]은 Skeleton 정보를 활용하여 학습 데이터의 양을 증가시켰다. SCHP[2]는 CE2P[12]의 Encoder-Decoder 구조의 CNN 아키텍처를 따르며, Model Aggregation과 Label Refinement를 반복하며 GT의 노이즈를 감소시킨다. 이 모델은 Single Human 데이터셋에 대해서 성능이 뛰어남이 증명되었다. (LIP[8] Validation Set에 대하여 mIoU: 59.36)

2.3. Instance Segmentation (개체 이미지 분할)

Instance Segmentation은 이미지에서 픽셀 단위로 클래스를 예측하고 각 객체를 개별적으로 구별한다. Yolact[1]는 이미지 크기의 프로토타입 마스크의 집합을 생성하고, 이를 객체마다 계산된 마스크 상수와 선형 연산하여 하나의 최종 분할 결과를 얻는다. Mask-RCNN[5] 계열의 2단계 모델과는 다르게 주요 작업을 병렬적으로 수행하기 때문에 실시간 추론에 적합한 모델이다.

3. 딥러닝 기반의 체형 변화 추적 서비스

제안하는 서비스는 촬영 가이드 제시 파트와 체형 비교·분석 파트로 나뉜다. 그림 1은 전체 서비스 개요를 보여준다.

3.1. SCHP 모델 학습에 사용된 데이터셋

SCHP 모델 학습에 다음의 세 가지 데이터셋 LIP[8], ATR[10], Pascal-Person-Part[11]가 각각 사용되었으며, 데이터셋별로 학습 가능한 신체부위와 클래스 개수가 각각 다르다. 상세 내용은 표 1에 기재되어 있다.

데이터셋	#데이터	#클래스	클래스 종류
LIP	50,462	19 + 1(배경)	모자, 머리카락, 장갑, 선글라스, 상의, 드레스, 코트, 양말, 바지, 점프수트, 스카프, 치마, 얼굴, 왼팔, 오른팔, 왼다리, 오른다리, 왼쪽신발, 오른쪽신발
ATR	17,000 +	17 + 1(배경)	모자, 머리카락, 선글라스, 상의, 치마, 바지, 드레스, 벨트, 왼쪽신발, 오른쪽신발, 얼굴, 왼다리, 오른다리, 왼팔, 오른팔, 가방, 스카프
Pascal-Person-Part	3,533	6 + 1(배경)	머리, 몸통, 팔뚝, 팔, 허벅지, 종아리

표 1 SCHP 모델 학습에 사용된 데이터셋 정보

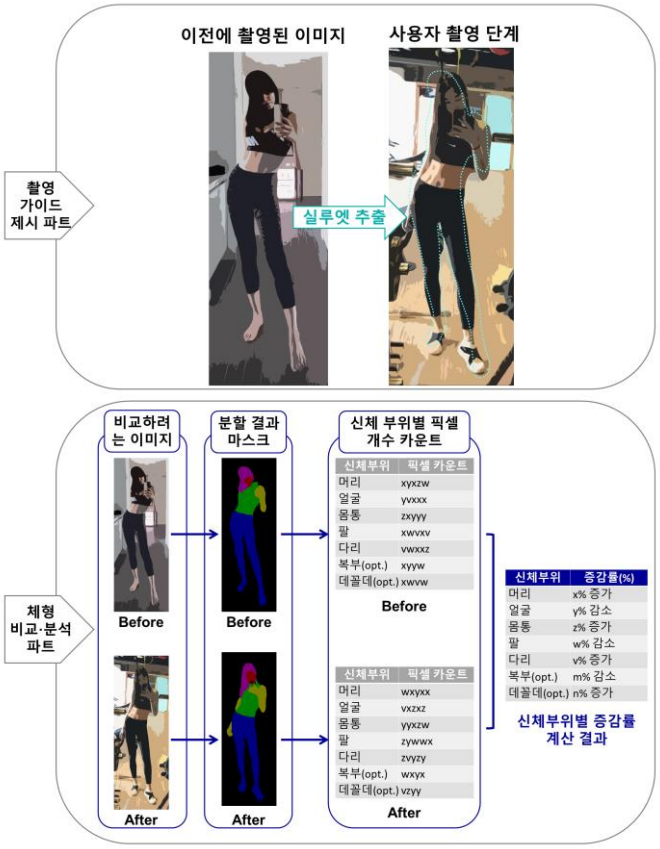

그림 1 딥러닝 기반의 체형 변화 추적 서비스 개요도

3.2. 촬영 가이드 제시 파트

촬영 가이드 제시 파트에서는 그림 1의 상단처럼 이전에 사용자가 촬영한 사진의 실루엣을 추출한 후, 이를 다음 사진 촬영을 위한 가이드 이미지로 사용자 촬영 단계에 사용자에게 제시한다. 실루엣 추출은 Resnet101-FPN을 백본으로 하는 Yolact 모델을 사용하여 'Person' 객체로 분류되는 픽셀들의 윤곽선을 추출하는 방식이며, 가장 큰 실루엣 하나만 추출한다. 이 방법은 비슷한 포즈의 일관된 사진 촬영이 가능하게 하여 사진을 비교할 때 발생하는 오차를 줄여준다. 또한, 실시간 추론이 가능하기 때문에 촬영되는 사진의 객체와 이전 실루엣이 겹치는 부분이 일정한 임계값 이상인지를 실시간으로 확인할 수 있다.

3.3. 체형 비교·분석 파트

체형 비교·분석 파트에서는 비교하고자 하는 두 개의 이미지에 대하여 각각 SCHP 모델을 활용하여 신체부위 이미지 분할을 진행한 후, 신체부위별 증감률(%)을 계산하여 체형 변화를 수치화한 결과를 제공한다. 상세 구현 방법은 아래와 같다.

먼저 데이터셋 LIP, ATR, Pascal-Person-Part를 각각 SCHP로 학습한 후, 세 가지 모델이 예측하는 픽셀별 클래스를 휴리스틱한 방법에 따라 N개의 클래스로 재편한다. (본 서비스의 구현에서는 LIP 모델과 ATR모델에서 N=5(머리, 얼굴, 팔, 몸통, 다리), Pascal모델은 머리와 얼굴을 합쳐서 N=4)

재편된 클래스에 대해서 세 가지 모델의 예측 결과에 Majority Voting을 적용하는 앙상블 접근(Majority가 없을 경우 우선순위는 Pascal모델)을 통해 최종 모델의 성능을 향상시킨다.

추가적으로 LIP모델이 '배경'으로 예측하거나 Pascal모델이 '얼굴'로 예측하는데, ATR모델은 '몸통'으로 예측하는 경우가 목 아래에서 가슴윗부분과 복부를 특정한다는 점에서 착안하여 기존에 없던 신체 부위 클래스인 '데꼴데'와 '복부'를 구분할 수 있다. 상반신의 가장 두꺼운 가로를 기준선으로 설정하여 윗부분은 '데꼴데', 아랫부분은 '복부'로 산술 계산하였다. (Optional)

(x, y)는 2차원 이미지 좌표이고 $f(x, y) = 1$일 때, 각 신체부위 c에 대하여 픽셀의 개수를 구하는 함수 $pc(c)$는 다음과 같이 표현할 수 있다.

$$pc(c) = \sum_{(x,y)\in c} f(x, y) = \sum_{(x,y)\in c} 1$$

여기서 각 신체부위 c별 증감률(%)은 다음과 같이 계산할 수 있다. ($pc_A(c)$, $pc_B(c)$는 각각 이후(After)이미지, 이전(Before)이미지에서의 신체부위 c의 픽셀 개수)

$$\frac{pc_A(c) - pc_B(c)}{pc_B(c)} \times 100 \ (\%)$$

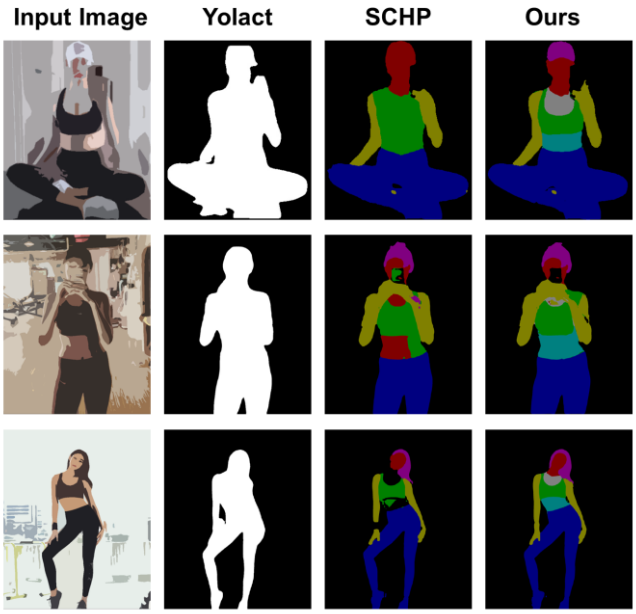

그림 2 Yolact, SCHP, 제안하는 방법에 대한 분할 결과

Method	mAcc	Bkg (Acc)	Face (Acc)	Torso (Acc)	Arms (Acc)	Legs (Acc)	Hair (Acc)	mIoU
SCHP-LIP	77.73	94.56	**78.40**	27.68	74.78	64.62	78.01	69.67
SCHP-ATR	86.40	**98.13**	22.84	50.88	**82.39**	**93.49**	74.62	70.39
SCHP-Pascal	92.73	97.51	75.43	77.25	78.28	89.09	-	83.51
Ours	**93.04**	98.04	70.95	**83.12**	80.28	91.67	**78.13**	**83.70**

표 2 자체 제작한 '눈바디' 테스트셋 실험 결과

4. 실험 및 결과

그림 2는 예시 이미지 3개를 Yolact모델, SCHP 모델(위에서부터 순서대로 각각 Pascal, ATR, LIP을 학습한 모델), 제안하는 체형 분석 방법에 통과시켜 얻은 분할 결과이다. Yolact모델로 추론된 결과는 실루엣 제공이라는 목표를 달성할 수 있는 수준인 것을 알 수 있다. SCHP모델로 추론된 결과보다 제안하는 체형 분석 방법에서 더 정확하게 신체가 분류된 것으로 보아 앙상블 접근과 클래스 세분화가 유의미했음을 알 수 있다.

웹 상의 '눈바디' 이미지를 크롤링한 후, 자체적으로 레이블링하여 '눈바디' 테스트셋(데이터 10개)을 만들었다. 표 2는 이 테스트셋을 이용하여, SCHP 모델 세 가지와 본 논문에서 제안하는 방법을, 전체 평균 픽셀 정확도(%), 재편된 신체 부위 클래스에 대한 평균 픽셀 정확도(%), 전체 평균 IoU (Intersection over Union) 값을 계산하여 평가한 결과이다. 본 논문에서 제안하는 방법이 전체 평균 픽셀 정확도(93.04%), 몸통과 머리 분할(83.12%, 78.13%), 평균 IoU값(83.70)에서 다른 모델들보다 좋은 성능을 보여준다.

5. 결론

본 논문에서는 Image Segmentation 기술을 활용한 딥러닝 기반의 체형 추적 서비스를 제안한다. 촬영 가이드 제시를 통해서 촬영의 일관성을 유도할 수 있고, 이로 인해 발생하는 비교 시의 편향도 줄일 수 있다. 또한, 촬영된 이미지를 비교하는 과정을 딥러닝 기반으로 자동화하고, 사용자에게 신체부위별 증감률을 수치화하여 제시함으로써 '눈바디' 루틴 지속의 원동력을 제공한다. 다만, 현재는 이미지를 비교할 때 포즈에 따라서 면적의 편차가 발생하기 때문에, 향후 지속적인 연구를 통해 이를 해결하고자 한다.

마지막으로 본 연구는 시간에 따라 변화하는 신체를 추적할 수 있기 때문에 앞으로 홈트레이닝 분야의 다양한 애플리케이션에서 활용될 수 있을 것으로 기대한다.

[참고문헌]

[1] Daniel Bolya et al. "YOLACT: Real-time Instance Segmentation". Em: ICCV. 2019.
[2] Peike Li et al. "Self-Correction for Human Parsing". IEEE TPAMI. 2020.
[3] Hao Zhu et al. "Detailed human shape estimation from a single image by hierarchical mesh deformation". Em: CVPR. 2019.
[4] Shunsuke Saito et al. "PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization". CVPR. 2020.
[5] Kaiming He et al. "Mask R-CNN". Em: ICCV. 2017.
[6] Lu Yang et al. "Parsing R-CNN for Instance-Level Human Analysis". Em: CVPR. 2019.
[7] Tao Li et al. "Self-Learning with Rectification Strategy for Human Parsing". Em: CoRRabs/ 2004.08055 (2020).
[8] Xiaodan Liang et al. "Look into Person: Joint Body Parsing Pose Estimation Network and A New Benchmark". Em: IEEE TPAMI. 2018.
[9] Kevin Lin et al. "Cross-Domain Complementary Learning Using Pose for Multi-Person Part Segmentation". Em: IEEE Transactions on Circuits and Systems for Video Technology PP. 2020.
[10] Xiaodan Liang et al., "Deep Human Parsing with Active Template Regression". Em: IEEE TPAMI. 2015.
[11] Xianjie Chen et al. "Detect What You Can: Detecting and Representing Objects Using Holistic Models and Body Parts". Em: CVPR. 2014.
[12] Tao Ruan et al. "Devil in the Details: Towards Accurate Single and Multiple Human Parsing". Em: AAAI. 2019.