홈트레이닝을 위한 관절 특징점 검출 및 행동 유사도 측정

*강도희, 송병철
인하대학교 전자공학과
asrs777@gmail.com, bcsong@inha.ac.kr

Joint keypoints detection and behavioral similarity measurement for home training

*Dohee Kang, Byung Cheol Song
Dept. Electronic Engineering, Inha University

요약

언택트 문화가 활성화되면서 다양한 업체에서 홈트레이닝 어플리케이션이 출시되고 있다. 많은 어플리케이션이 관절 특징점 검출 기능을 제공하여 사용자에게 편리함을 제공하지만, 자체 컨텐츠만 사용가능하다는 점에서 한계를 갖는다. 본 작품에서는 딥러닝 기반의 관절 특징점 검출기 및 특징 추출기를 결합하여 실시간 자세 유사도 측정기를 구현하였다. 목표영상 및 사용자의 관절 위치를 파악함과 동시에 관절 위치 정보에 대한 특징을 추출하여 자세 유사도를 실시간으로 점수화해 사용자에게 제공한다.

1. 작품의 제작 동기

지난 1월부터 현재까지 이어지고 있는 COVID-19 바이러스 확산으로 인해 재택근무, 온라인수업, 홈트레이닝 등 다양한 언텍트 문화가 활성화되고 있다. 이 중에서도 홈트레이닝의 경우 카카오VX의 '스마트 홈트' 등 관절 특징점 검출을 통해 사용자의 자세를 추정 및 교정해주는 어플리케이션들이 출시되고 있다. 이러한 제품들은 자체적으로 제공하는 영상에 대해서만 서비스를 사용할 수 있다는 점에서 컨텐츠 다양성의 한계점을 갖는다.

이에 본 작품에서는 관절 특징점 검출기와 함께 특징 추출기를 결합해 실시간 자세 유사도 측정기를 구현하였다. 사용자가 따라하기 원하는 목표영상과 함께 자신의 모습을 실시간으로 촬영하여 시스템에 입력하면 시스템 내부에서 두 영상의 관절 위치를 파악함과 동시에 그에 따른 특징을 추출하여 두 영상의 자세에 대한 유사도를 점수화하여 출력한다.

2. 작품의 설계 및 구현

관절 특징점 추정기의 경우 주어진 영상 내에서 대략적인 사람의 위치를 찾아낸 후 그 안에서 관절의 위치를 파악하는 Top-down 방식과 우선 관절 위치를 파악한 후에 관절 특징점간의 관계를 분석하여 자세를 추정하는 Bottom-up 방식 두 가지로 분류할 수 있다. Top-down 방식의 경우 사람 영역을 먼저 정한 후에 해당 영역 내에서 관절 특징점을 추정하기 때문에 그 정확도가 Bottom-up 방식보다 현저히 높으며, 한 명의 사람의 경우 더 빠르게 추정이 가능하다는 장점이 있지만, 영상 내 여러 명의 사람이 등장할 경우 각 사람에 대한 영역을 먼저 파악해야 하기 때문에 속도저하가 심해진다는 단점을 가진다. 실제 Top-down 방식을 사용하는 Alphapose [1]와 Bottom-up 방식을 사용하는 Openpose [2]를 비교한 결과 영상 내 1명의 사람이 등장할 경우 각각 기본 설정에서 15.5fps/5.8fps의 속도를 보였으며 관절 특징점 파악에 대한 정확도 또한 Alphapose가 높은 것을 확인할 수 있었다. 본 제품의 경우 홈트레이닝에 초점을 두고 있어 여러 사람에 대한 빠른 특징점 추정보다는 한 명 또는 두명의 사람에 대한 정확한 추정이 중요하다고 판단하여 top-down 방식의 Alphapose를 사용할 모듈로 선정하였다.

Method	AP @0.5:0.95	AP @0.5	AP @0.75	AP medium	AP large
OpenPose (CMU-Pose)	61.8	84.9	67.5	57.1	68.2
Detectron (Mask R-CNN)	67.0	88.0	73.1	62.2	75.6
AlphaPose	**73.3**	**89.2**	**79.1**	**69.0**	**78.6**

그림 1. AlphaPose와 OpenPose의 정확도 비교

이에 해당 시스템 Alphapose를 활용해 자세 추정 및 관절 특징점 정보를 저장하고, 사용자 영상 내 출현 인물에 대한 자세 추적 기능을 제공하는 관절 특징점 검출부와 관절 특징점 위치를 파악한 목표영상과 실시간으로 입력되는 사용자영상의 관절 특징점 정보를 입력으로 받아 자세 유사도를 점수화하여 출력하는 AutoEncoder부로 이루어진다.

관절 특징점 검출부는 AlphaPose v0.3.0 version 과 함께 72.0AP의 정확도를 갖는 ResNet50 기반 FastPose Model 를 사용하였다[3]. 이를 통해 입력 영상에서 사람을 검출하여 17개의 관절 특징점쌍을 도출한다. 이를 통해 640*480 웹캠 영상을 기준으로 사람 1명이 등장하는 영상에선 최대 약 20fps의 속도, 2명의 경우 18fps의 결과를 보였다. 2명의 사용자에 대한 영상의 경우 각 사용자의 관절 위치를 실시간으로 추적해야 각각의 유사도 측정이 가능한데, 이를 위해 Alphapose에서 제공하는 pose tracking 기능을 사용할 경우 각 18fps, 14fps의 결과를 보였고, 이를 통해 속도저하가 매우 크다는 것을 확인할 수 있었다. 이에 본 시스템에서는 pose tracking 기능을 사용하는 것이 아닌 영상 속 사용자가 서로 교차되지 않는다는 가정 하에 특정 좌표값을 기준으로 두 사람을 분리하여 관절 특징점을 검출 및 추적한다.

AutoEncoder부는 자세 유사도 측정을 위해 관절 특징점 검출부로부터 출력되는 목표 영상과 실시간 사용자 영상의 각각 17쌍의 관절점 정보를 사용한다. 본 시스템에서 구현된 특징 추출기 AutoEncoder는 각각의 영상 내에서 17개의 특징점 간 유클리디안 거리 17C2에 해당하는 136개의 vector를 학습된 특징 추출기의 입력으로 넣으면 해당 자세

의 관절간의 특징이 담긴 Latent vector를 얻을 수 있다. 각 영상에 대한 Latent vector에 대해 cosine similarity를 비교해 그 값을 통해 두 자세의 유사도를 측정하여 이를 점수화할 수 있다. 시스템의 전반적인 흐름은 그림2와 같다.

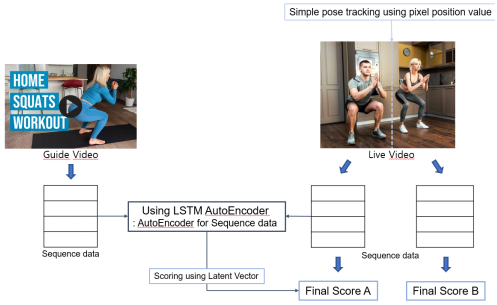

그림 2. 시스템 흐름도

추가적으로, 사용자는 목표 영상을 실시간으로 시청하면서 따라하는 것이기 때문에 목표영상과 매 순간 같은 동작을 할 수 없으며, 동작에 따라 따라하는 데에 걸리는 시간 또한 차이가 날 수 밖에 없다. 이에 사용자와 목표영상간의 속도 차이를 무시하기 위해서 Dynamic time Warping 알고리즘을 사용한다 [4]. 해당 알고리즘의 경우 길이가 다른 두 시계열 데이터의 유사도를 계산하기 위해 사용되며, 비교할 데이터가 목표 데이터의 파장과 최대한 비슷해지도록 비교할 데이터를 인위적으로 휘게 만든다. 이를 통해서 두 영상에 대해 똑같은 속도로 비교하는 것이 아닌 가장 비슷한 동작에 대해서 유사도를 비교하여 사용자가 목표 영상을 잘 모방하고 있는가에 대해 정확한 판단이 가능하다

3. 작품의 예상 구현 결과

두 영상의 자세 유사도를 비교하기 위해서는 관절 위치 정확도를 측정하는 PCK(Percentage of Correct Keypoints) 또는 PDJ (Percentage of Detected Joints) 또한 활용이 가능하다. 그러나 이러한 방법들의 경우 영상 내 사람의 위치 및 크기를 동일하게 맞추기 위해 사람 정규화 과정이 필요하며, 이를 실시간 상황 내에서 프레임 단위로 처리하기에는 한계가 있다. 본 작품에서 제안하는 AutoEncoder의 경우 사람의 크기 등 위치에 대해 관계없이 자세의 특징을 파악할 수 있도록 학습되었기 때문에 빠르게 자세 비교가 가능하며, 이를 통해서 실시간으로 사용자와 목표 영상의 자세 유사도를 파악할 수 있다.

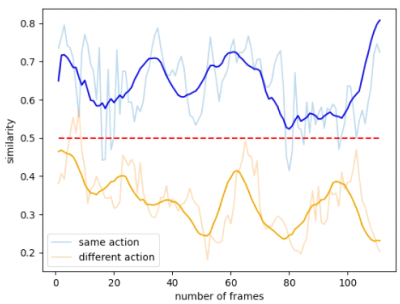

그림 3. AutoEncoder 출력

AutoEncoder의 실제 동작 모습은 그림3과 같다. 실제 4초 분량의 같은 동작을 하면서 사람의 위치 및 크기가 다른 두 사람의 영상을 입력하였을 때, 평균적으로 약 0.65의 cosine similarity를 보였으며, 다른 동작을 하는 영상의 경우 약 0,33의 결과를 보였다. 이를 통해서 출력되는 유사도가 어느 정도의 신뢰도를 갖는다는 것을 확인하였다.

현재 본 작품은 구현 단계에 있으며, Alphapose를 통한 관절 특징점 추출 및 유사한 영상과 자세 차이가 큰 영상에 대한 Latent vector 유사도의 차이가 있음을 확인하였다. 아직 구현이 완료된 상태가 아니기 때문에 정확도 측면에서 부족함이 있지만, 가장 적절한 AutoEncoder 구조와 Hyper parameter를 찾아 시스템에 적용한다면 자세 유사도를 더욱 정확하게 구할 수 있을 것이다.

4. 작품의 기대효과

COVID-19 바이러스 확산이 장기간 지속되면서 언택트 문화는 사회적 거리두기를 위한 일시적인 대체 문화가 아닌 일상으로 자리잡고 있다. 이와 함께 장소 및 시간에 대한 제한이 없는 홈트레이닝 시장의 발전 또한 가속화되고 있는 상황이다. 이에 유튜브 등의 매체를 통해 다양한 운동 컨텐츠를 쉽게 접근할 수 있다. 하지만 컨텐츠에서 제공되는 영상 및 음성만으로는 사용자가 자세를 완벽하게 따라하기에는 무리가 있다. 자세추정 기능을 제공하는 애플리케이션들도 출시되고 있지만 컨텐츠에 제한이 있고 이는 홈트레이닝의 가장 큰 장점인 컨텐츠의 다양성을 제대로 활용하지 못한다는 것을 의미한다.

본 애플리케이션은 사용자가 원하는 영상을 시스템에 입력하면 즉각적으로 관절 keypoint 정보를 추출 및 저장하고, 이를 사용자의 정보와 비교한다. 따라서 사용자가 원하는 컨텐츠를 제한 없이 자유롭게 사용이 가능하며, 실시간으로 사용자와 목표영상을 비교하고 이를 점수화하여 제공하기 때문에 영상 및 음성으로는 부족했던 컨텐츠에 대한 이해도를 향상시킬 것으로 기대할 수 있다.

더 나아가 단순히 매체에서 얻을 수 있는 영상을 모방하는 것을 넘어 현재 사회적 거리두기로 인해 크게 피해를 입고 있는 피트니스 업체 및 체육 수업 등에도 적용하여 더 다양한 분야에서 활용할 것으로 기대할 수 있다.

5. 사사

이 논문은 2020년도 현장맞춤형 이공계 인재양성 지원사업의 재원으로 한국연구재단 지원을 받아 수행된 연구임.

6. 참고문헌

[1] https://github.com/MVIG-SJTU/AlphaPose

[2] https://github.com/CMU-Perceptual-Computing-La/openpose

[3] Zhang, J., Zhu, Z., Zou, W., Li, P., Li, Y., Su, H., & Huang, G. (2019). Fastpose: Towards real-time pose estimation and tracking via scale-normalized multi-task networks. arXiv preprint arXiv:1908.05593.

[4] Müller, M. (2007). Dynamic time warping. Information retrieval for music and motion, 69-84.