DETR 기반 객체탐지를 사용한 댄스 자세교정 방법

우상철, 지수미, 성연식*
동국대학교 멀티미디어공학과
woo.si@dgu.ac.kr, sumi@dongguk.edu, sung@dongguk.edu

Dance Posture Correction Method using DETR-based Object Detection

Sangchul Woo, Sumi Ji. Yunsick Sung*
Dept. of Multimedia Engineering, Dongguk University-Seoul, South Korea

요 약

전 세계적으로 코로나 바이러스가 확산되면서 언택트 시대가 되었다. 언택트 시대에서는 대부분의 대면활동이 비대면으로 전환되고 있다. 전 세계적으로 열광중인 케이팝 댄스의 대중화를 위해 우리는 비대면으로 댄스 학습이 가능한 DETR 기반 객체탐지를 사용한 댄스 자세교정 연구를 제안한다. 본 논문에서 제안한 댄스 자세교정은 객체탐지에 DETR을 적용한 방식이다. DETR은 기존 객체탐지 모델에서 앵커박스, 바운딩박스 중복처리를 제거하는 NMS같은 휴리스틱한 방법을 사용하지 않고 트랜스포머를 통해 자동으로 학습하도록 만든 모델이다. DETR로 객체탐지를 한 후 강사와 사용자의 동작유사성을 샴 뉴럴 네트워크를 통해 계산한다.

1. 서론

케이팝이 전 세계적으로 열풍을 일으키면서 케이팝을 배우기 위한 수요가 점점 증가하고 있다. 하지만 코로나 바이러스가 전 세계적으로 확산되면서 언택트 시대라는 새로운 국면으로 접어들었다. 언택트 시대에서는 우리가 상상도 못 했던 무관중 경기, 비대면 교육 등 다양한 삶의 변화가 일어나고 있다. 모든 활동이 언택트로 전환되면서 케이팝 중에서도 특히 댄스를 배우고 싶어하는 수요를 받아들이기 어려워졌다. 지금까지 우리가 생각했던 댄스 교육은 강사와 학생이 같이 마주 보면서 댄스를 따라하고, 자세를 교정하는 방식이다.

본 논문은 비대면 환경에서 댄스를 효과적이고 더 쉽게 학습하기 위해 DETR 기반 객체탐지를 사용한 댄스 자세교정 연구를 제안한다. 사용자는 영상을 통해 강사의 댄스를 보고 동작을 따라한다. 사용자가 따라한 동작을 카메라로 인식하여 동일시간에 대해서 강사와 사용자의 동작을 탐지하고 해당하는 동작과 강사의 동작을 비교하여 어떤 동작을 취해야 하는지 동작 유사성을 계산하고 알려준다.

사용자와 강사의 객체를 탐지할 때, DETR[1]을 적용한 객체탐지 모델을 사용한다. DETR은 기존의 휴리스틱한 작업으로 해야했던 앵커박스와 바운딩박스 중복처리 제거 등을 트랜스포머를 통해 자동 학습하게 하여 End-to-End 학습이 가능하게 만든 새로운 모델이다. 강사와 사용자의 동작 유사성 비교는 합성곱 신경망으로 이루어진 샴 뉴럴 네트워크[2]를 사용한다. 합성곱 신경망을 사용해 특징벡터를 추출하고 특징벡터간 유사성을 계산하는데, 이는 더 정확한 동작 비교가 가능하게 해준다. 사용자는 실시간으로 동작을 피드백 받으면서 댄스를 학습할 수 있다. 이는 높은 하드웨어 성능과 발전된 딥러닝 모델을 통해 모바일 환경에서도 실시간으로 동작추정이 가능하게 되었다.

전 세계적 인기를 끌고있는 케이팝이 언택트라는 난관에 막히지 않고, DETR 기반 객체탐지를 사용한 댄스 자세교정 방법을 통해 활발히 전파하는데 기여할 수 있을 것이라 예상한다.

2. 제안방법

그림 1은 우리가 제안한 방법에서 DETR[1]과 샴 뉴럴 네트워크[2]의 동작과정을 보여준다. 우리가 제안한 연구에서는 2가지 과정에 따라서 댄스 자세교정이 이루어진다. 1)DETR기반 객체탐지 기술을 활용해

*교신저자 : 성연식(sung@dongguk.edu)

서 영상에서 사용자 객체를 탐지한다. 2)탐지된 사용자 객체와 강사 객체의 동작 유사성을 샴 뉴럴 네트워크를 통해 동작 유사성을 비교한다.

(그림 1) DETR과 샴 뉴럴 네트워크 동작과정

2.1 객체탐지

객체 탐지 방식으로 동일시간에 대해서 강사의 동작과 사용자의 동작을 검출한다. 본 논문에서 사용할 모델은 DETR 구조를 사용하여 기존에 사용하던 객체탐지 구조를 단순화시켰다. 기존모델은 오래전부터 휴리스틱한 방법에 기반을 두고 있어 End-to-End로 학습하기에는 어려움이 있었지만 DETR 구조를 사용함으로써 이를 해결했다. DETR 구조는 자연어 처리 모델에서 사용하던 트랜스포머를 객체탐지에 활용하였다. DETR은 합성곱 신경망으로 특징을 추출하는 부분과 트랜스포머를 사용한 인코더-디코더 구조로 구성된다 DETR은 모든 객체를 한번에 예측하며 앵커박스같은 전처리를 하지 않고 간단한 파이프라인을 가진다. 또한 멀티 객체가 아닌 단일 객체만 탐지하면 되기 때문에 사용자 동작 인식이 더 쉬울 것으로 기대한다. 합성곱 신경망 모델로는 ResNet을 사용하였지만, 성능과 시스템 응답성을 고려해서 ResNet-152나 모바일넷으로 대체가능하다. DETR구조에서는 예측한 객체와 정답으로 예측해야하는 객체 사이에 이분 매칭[3]을 통해 손실함수를 계산한다.

2.2 동작 유사성 비교

사용자와 강사의 동작을 샴 뉴럴 네트워크[1]를 통해 같은 동작인지 다른 동작인지 인식한다. 우리는 이를 이미지 유사도 문제 관점에서 해결하고자 한다.

단순 픽셀값 비교를 통해서 이미지 유사도를 측정하기는 어렵다. 같은 그림이지만 그림의 중심점이 다를 수 있고, 그림이 완전히 일치하지 않는 이상 성능을 기대하기 어렵다. 하지만 샴 뉴럴 네트워크는 벡터영역에서 이미지 유사도를 비교하기 때문에 유사성 비교가 가능하다.

샴 네트워크는 두 개의 이미지를 입력으로 받아 같은 클래스인지 다른 클래스인지만 학습한다. 같은 클래스일 경우 합성곱 신경망 모델을 통해 추출된 특징벡터 거리를 더 가깝게 만들고, 다른 클래스일 경우 특징벡터 거리를 더 멀게 만든다. 학습이 완료되면 특징벡터간 거리를 통해 같은 클래스인지 다른 클래스인지 구분한다. 샴 뉴럴 네트워크에서는 일반적인 정규화방법 및 이진 크로스엔트로피 손실함수가 사용된다[4]. 또한 학습에 사용하지 않은 동작에 대해서도 구분이 가능하기 때문에 우리가 제안한 연구에서도 높은 성능을 발휘할 수 있다.

3. 결론

위에서 사용자의 자세교정을 수행하기 위한 전체 과정을 기술하였다. 제안한 방식은 사람을 탐지한 후 동작을 비교하는 2단계로 신경망이 구성되어 있다. 이후에는 DETR을 통해 사람을 탐지하는게 아닌, 사용자의 동작을 바로 탐지하는 신경망을 1단계로 줄여 모델을 단순화하는 연구를 수행할 것이다.

사사표기

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 글로벌핵심인재양성지원사업의 연구결과로 수행되었음" (2020-0-01576)

참고문헌

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, Sergey Zagoruyko, "End-to-End Object Detection with Transformers," European Conference on Computer Vision (ECCV), 2020.

[2] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, "Siamese neural networks for one-shot image recognitio" deep learning workshop International Conference on Machine Learning(ICML), Lille France, 2015.

[3] Birnbaum Benjamin, Claire Mathieu. "On-line bipartite matching made simple," Special Interest Group on Algorithms & Computation Theory (SIGACT), Vol.39, No.1, pp.80-87, 2008.

[4] Elad Hoffer, Ron Banner, Itay Golan, Daniel Soudry, "Norm matters: efficient and accurate normalization schemes in deep networks," Neural Information Processing Systems (NIPS), Montreal Canada, 2018.