멀티모달 퓨전을 이용한 실시간 1인칭 손동작 인식

김보경[1], 정희용[2], 신춘성[3]

[1,2]전남대학교 인공지능학부
[3]전남대학교 문화전문대학원
bokyung08@jnu.ac.kr, h.jeong@jnu.ac.kr, cshin@jnu.ac.kr

Real-Time First-Person Hand Gesture Recognition using Multimodal Fusion

Bo-Kyung Kim[1], Hie-Yong Jeong[2], Choon-Sung Shin[3]

[1,2]Dept. of Artificial Intelligence, Chonnam National University
[3]Graduate School of Culture, Chonnam National University

요 약

인간-컴퓨터 상호작용(HCI), 증강현실(AR), 가상현실(VR) 기술이 발전함에 따라, 실시간 손동작 인식 기술의 중요성이 커지고 있다. 특히 1인칭 시점의 손동작 인식 기술은 사용자 친화적인 모델을 구축한다는 의미가 있지만, 영상의 각도 등으로 인해 self-occlusion 현상이 빈번하게 발생하기에 기술적 어려움이 존재한다. 본 논문에서는 이러한 문제를 해결하기 위해 어텐션 메커니즘에 기반한 멀티모달 퓨전 모델을 제안한다. 모델은 두 가지 입력 형태를 갖는다. 첫 번째 스트림은 비디오 프레임으로부터 3D-CNN 모델을 통해 시각적·공간적 특징을 추출하고, 두 번째 스트림은 MediaPipe를 통해 얻은 손의 관절 좌표로부터 움직임을 학습한다. 추출된 두 특징 벡터는 어텐션 메커니즘을 통해 동적으로 Fusion 되어 최종 특징을 생성한다. 본 모델을 통해 자체 구축한 1인칭 손동작 데이터셋을 이용해 실험한 결과, 높은 인식 정확도를 보였으며 실시간 환경에서도 높은 인식률을 보여주었다. 본 연구는 1인칭 시점에서 기기와의 상호작용을 위한 효과적인 손동작 인식 방법론을 제시하였다는 의의를 갖는다.

1. 서론

최근 가상현실(VR) 기기의 상용화로 가상환경과의 상호작용이 보편화 되었다. 초기 제어 방식은 물리적 컨트롤러를 통해 가상 환경과 상호작용하는 방식이 대부분이었으나, 이는 사용자에게 움직임의 불편함을 초래한다는 문제점이 있다. 이러한 한계를 극복하고자 Apple Vision Pro와 같은 별도의 컨트롤러 없이 제어 가능한 기술에 대한 요구가 증가하고 있다. 이 인터페이스를 구현하기 위한 핵심 기술은 사용자의 의도를 실시간으로 파악하는 1인칭 시점 손동작 인식 기술이 포함된다. 본 논문에서는 어텐션 메커니즘 기반 멀티모달 네트워크를 이용하여 실시간으로 동작하는 1인칭 시점 손동작 인식 기술 개발을 목표로 한다.

2. 관련 연구

1) 3D-CNN 기반 동작 인식

동작 인식 분야의 초기 연구는 주로 2D-CNN 기술을 사용하여 비디오 데이터의 시간축을 고려하지 않은 이미지 인식에 사용되었다. 이 방법은 시간적 연속성을 학습하게 어렵다는 한계가 있어 시간적 특징과 공간적 특징을 함께 추출할 수 있는 3D-CNN 방식으로 발전하였다.[1]

2) Gating Mechanism

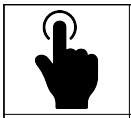
멀티모달 학습에서 각기 다른 모달리티의 데이터를 융합하는 것은 성능 향상을 위한 핵심적인 요소이다. 기존의 융합 방식은 단순히 데이터를 결합하는 방식으로 인해 모달리티별 특성을 반영하지 못하거나 중요한 데이터에 대한 가중치 없이 학습을 진행하는 경우가 많았다. 이러한 문제를 해결하기 위한 방법에는 게이팅 기반 멀티모달 융합 기법이다. 게이팅 메커니즘은 각 모달리티에서 추출된 특징 벡터에 동적인 가중치를 할당하여, 데이터의 중요한 특징을 반영할 수 있도록 한다. 이러한 모델은 단일 모달 접근 방식보다 높은 성능을 보여주고 있다.

3. 1인칭 시점 실시간 손동작 인식 모델 구축

1) 데이터셋 구축

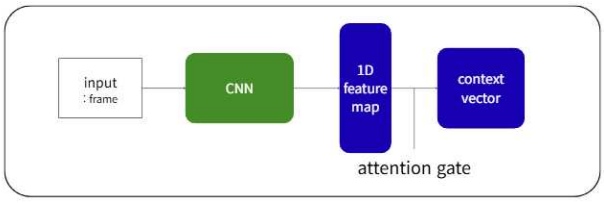
본 연구에서는 Motion Array 플랫폼에서 Touch Screen Hand Gesture 키워드를 통해 수집한 영상과 직접 촬영한 영상을 활용하여 자체 데이터셋을 구축하였다. 수집한 원본 영상에서는 라벨이 제공되지 않았기 때문에, 동작 클래스별로 분류 후 수작업 라벨링을 진행하였다. 각 영상은 5초 내외의 길이로 구성되며, 연속적인 손동작 수행 과정을 포함하고 있다.

<표 1> 데이터셋 클래스의 종류

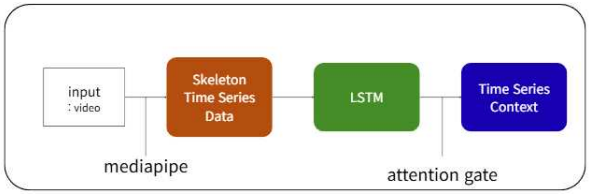
				
tap	double tap	pinch	zoom	long press
				
swipe right	swipe left	swipe up	swipe down	rotate

2) 모델 정의

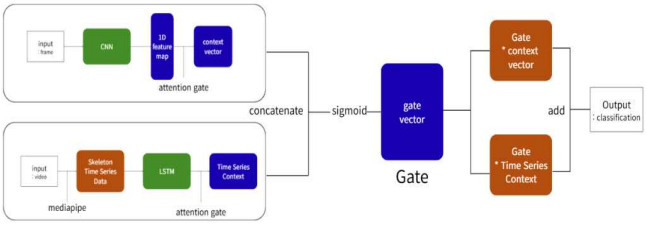
본 연구에서는 1인칭 시점의 실시간 손동작을 위해 이중 스트림 모델을 설계하였다. 제안된 모델은 손동작으로부터 추출된 시각적 특징과 skeleton 좌표로부터 추출된 시계열 특징을 동시에 입력 받아 처리한다. 첫 번째 스트림은 영상으로부터 공간적·시각적 특징을 추출하고, 어텐션 메커니즘을 통해 중요한 맥락이 담긴 프레임을 학습한다. 두 번째 스트림은 MediaPipe 라이브러리를 이용해 비디오 데이터를 skeleton 좌푯값으로 변환해 시계열 특징을 추출 후, 어텐션 메커니즘을 적용하여 중요한 시계열 특징에 집중하여 학습한다.



(그림 1) 첫 번째 스트림 아키텍처



(그림 2) 두 번째 스트림 아키텍처



(그림 3) 최종 아키텍처

4. 실험 결과

앞서 정의한 모델의 아키텍처를 적용한 실험을 통해 단일 모델에 비해 나은 성능을 보임을 확인할 수 있었다. 동일 환경에서 3D-CNN-LSTM 단일 모델을 적용한 경우에는 0.85의 정확도를 확인할 수 있었지만, 본 논문에서 사용된 모델을 적용하였을 때 정확도가 약 0.97로 크게 오른 점을 확인할 수 있었다.

<표 2> 모델 성능 비교

Model	Accuracy
3D-CNN-LSTM	0.85
ST-GCN	0.46
Multimodal Model	0.97

5. 결론

본 연구에서는 1인칭 시점 손동작 인식을 위한 이중 스트림 멀티모달 모델을 제안하였다. 실험 결과, 제안된 모델은 단일 모달 기반 모델에 비해 우수한 성능을 보였으며, 실시간 환경에서도 안정적인 손동작 인식이 가능함을 확인하였다. 이는 멀티모달 융합방식이 1인칭 시점 손동작 인식에서 효과적인 접근법임을 시사한다.

사 사

본 연구는 차세대통신혁신융합대학사업단 실험실 연계 소그룹 운영 지원을 받아서 수행하였다.

참고문헌

[1] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, Manohar Paluri “Learning Spatiotemporal Features with 3D Convolutional Networks” Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4489-4497.