OTT 기반의 모션 검출 및 객체인식 모델 연구

김유림[1], 이승재[1], 최휘녕[1]
전북대학교 IT정보공학과
e-mail : yoo9379@gmail.com, leesengjie@naver.com, soonso00@naver.com

A Study on OTT-based Motion Detection and Object Recognition Models

Yoo-Rim Kim, Seung-Jae Lee, Hwi-Nyeong Choi
Dept. of IT Information Engineering, Jeonbuk National University

요 약

최근 OTT 서비스의 이용 증가로 인해 많은 사람들이 일방향 미디어에 장시간 노출되고 있다[1]. 이는 영유아의 발달 과정에 부정적인 영향을 미치게 된다. 따라서 본 논문은 영유아의 건강한 미디어 매체의 접근을 위해 영유아가 화면 속 동작을 따라 움직이며 학습에 능동적으로 참여할 수 있는 상호작용 학습 콘텐츠를 개발하는 것을 목적으로 한다. OTT 셋톱박스에 내장된 카메라를 활용해 사용자 모션 인식을 진행하여, 영유아가 콘텐츠에서 요구하는 행동을 올바르게 수행하는지를 판단하는 기능을 구현하였다. 사용자 모션 인식은 MediaPipe Holistic을 활용하여 실시간으로 움직임을 감지하고, 학습에 사용되는 교구재 인식을 위한 객체 인식은 Custom CNN 모델을 통해 그의 사용 여부를 판단할 수 있게 제안했으며, 본 논문에서 객체 인식 모델은 85.71%의 정확도를 구현하였다.

1. 서 론

코로나(COVID-19) 사태 이후 OTT(Over-The-Top) 서비스에 대한 수요가 크게 급증해 콘텐츠 소비 패턴이 재편되었다. 연구에 따르면 봉쇄 기간 동안 더 많은 사람들이 집에 머무르며 엔터테인먼트를 위해 스트리밍으로 전환하면서 OTT 플랫폼 사용과 기존 TV 시청률이 모두 증가했다[1].

하지만 뇌 발달이 중요한 시기인 영유아기에 장시간 동안 일방향 미디어 매체에 노출되는 것은 그들의 발달 과정에 부정적인 영향을 미칠 수 있다[2]. 2015년 옥스퍼드 대학의 '텔레비전 시청이 뇌구조에 미치는 영향: 횡단면 및 종단면 분석' 논문에 따르면 TV 시청이 전두엽을 포함한 특정 뇌 영역의 회백질 및 백질 구조에 변화를 일으켜 언어 지능(IQ) 저하에 영향을 미칠 수 있다고 보고되었다[3]. 전두엽은 결정을 내리고, 계획을 세우고, 사회적 상호작용을 처리하고, 감정을 조절하는 등의 중요한 기능을 담당한다.

그러나 미디어 매체의 인기와 사용량이 증가함에 따라 현대인의 일상에서 TV는 필수불가결한 도구로 기능한다. TV와 같은 일방향 미디어의 한계를 극복하고 전두엽 발달을 촉진하기 위해, 사용자가 능동적으로 반응하고 참여할 수 있는 상호작용 가능한 뉴미디어의 필요성이 더욱 강조되고 있다.

따라서 본 논문은 게임 콘솔인 Nintendo Wii와 같이 사용자와 상호작용할 수 있는 학습 콘텐츠를 개발하기 위한 목적으로 연구를 진행한다. 이러한 상호작용 학습 콘텐츠는 사용자가 화면 속 동작을 따라 하며 자신의 신체를 움직이고, 이를 통해 능동적으로 학습에 참여하도록 설계된다. 모션 인식 기술을 활용해 사용자의 동작을 실시간으로 추적하고 분석함으로써, 상호작용을 유도하는 방식이다.

MediaPipe Holistic을 활용해 사용자의 움직임을 효과적으로 인식하고, 사용자의 몸에 Bounding Box를 처리해 범위 내에서 특정 동작의 수행 여부를 판단한다. 더 나아가 Custom CNN을 통해 객체 분류 기술을 구현해 학습에 사용되는 교구재를 인식할 수 있도록 하였다.

본 논문의 구성은 다음과 같다. 2장에서는 사용자 모션 인식과 특정 행동 수행 여부 판단, 그리고 특정 교구재 객체 인식에 대해 나누어 설명한다. 마지막으로 3장에서 결론과 향후 연구방향으로 맺는다.

2. 관련 연구

2.1 사용자 모션 인식 및 특정 행동 수행 여부 판단

MediaPipe 솔루션은 구글에서 제공하는 인공지능(AI) 및 머신러닝(ML) 기법을 빠르게 적용할 수 있는 라이브러리 서비스이다. 그 중 MediaPipe-Holistic은 기존 Mediapipe의 Face-Mesh, Pose, Multi-hand를 합쳐놓은 모듈로 구성되어 있다. Face-Mesh의 경우 468개의 3D 얼굴 랜드마크를 실시간으로 추정하는 얼굴 형상 솔루션이다. Pose는 고충실도 신체 자세 추적을 위한 ML 솔루션으로, BlazePose 연구를 사용하여 RGB 비디오 프레임에서 전신의 33개의 3D 랜드마크를 추론한다. 마지막으로 Hands는 충실도가 높은 손 및 손가락 추적 솔루션으로, ML을 사용하여 단일 프레임에서 손의 3D 랜드마크 21개를 추론한다.

본 논문은 MediaPipe Holistic을 사용하여 사용자 모션 인식을 구현하였으며, 랜드마크 시각화를 위해 MediaPipe의 drawing_utils를 사용하고, 그림 1의 좌와 같이 각각의 랜드마크를 선으로 이어 화면에 나타내었다.

* 제 1저자: 학생 회원 김유림(전북대학교 IT정보공학과), 공동 저자: 이승재(전북대학교 IT정보공학과), 최휘녕(전북대학교 IT정보공학과)
** "본 연구는 2024년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업의 지원을 받아 수행되었음"(2022-0-01067)

특정 행동의 정답 데이터를 이미지 형태로 입력한 후 그림 1의 좌와 같이 바운딩 박스를 둘러싼다. 정답 데이터의 좌표값을 바운딩 박스 기준으로 정규화 시켜 이후 비디오의 랜드마크 값과 비교할 수 있도록 한다. 정답 이미지와 실시간 모션이 인식되는 프레임 속 사람에게 각각 바운딩 박스를 생성하고, 그 랜드마크를 바운딩 박스를 기준으로 정규화시킨다. 그리고 두 개의 3D 랜드마크 좌표값 집합을 비교해 일정 임계값 내에서 일치하는 점들의 개수를 집계했다. 이때 임계값이란, 두 랜드마크의 좌표가 일정 거리를 벗어나지 않은 범위에서 일치하는 것으로 간주하는 기준값으로, 이 과정에서 표 1과 같이 임계값은 0.05로 설정하였다. 이는 유아의 특성상 오랫동안 동일한 행동을 유지하기 어렵다는 점을 고려해, 정확한 좌표로부터 일정 수준의 오차 값을 허용한 상태로 앞서 추출했던 랜드마크 값을 활용해 특정 행동 수행 여부 판단을 진행하였다.

(표 1). 임계값 수식 관한 함수

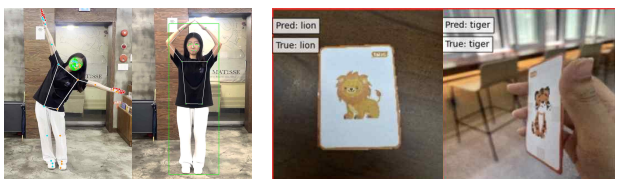
```
def compare_landmarks(landmarks1, landmarks2, threshold=0.05):
    matching_points = 0
    for lm1, lm2 in zip(landmarks1, landmarks2):
        if abs(lm1['x'] - lm2['x']) <= threshold and abs(lm1['y'] - lm2['y']) <= threshold and abs(lm1['z'] - lm2['z']) <= threshold:
            matching_points += 1
    return matching_points
```

2.2 특정 교구재 객체 인식

특정 교구재를 인식하기 위한 맞춤형 CNN(Convolutional Neural Network) 모델을 설계하고, 이를 학습시키기 위해 제작한 교구재 데이터셋을 사용했다. 주 교구재로는 그림1에 우와 같이 사자와 호랑이 그림이 그려진 카드를 활용했으며, 두 객체를 정확히 분류하는 모델을 구현했다. CNN 모델은 두 개의 Convolution layer와 Pooling layer, 그리고 두 개의 Fully connected layer를 기반으로 설계했다.

데이터셋은 그림 1의 우와 같이 사자와 호랑이 카드를 각각 클래스 0과 1로 라벨링하여, 약 70%의 데이터를 Train data set으로, 나머지 30%를 Validation data와 Test data로 각각 나누었다. 데이터셋이 비교적 작다고 판단하여, Resizing한 이미지 파일을 Blur, 밝기 조정, 이미지 축소, 좌우 반전을 적용시켜 기존 데이터를 7배로 증강시켰다. 이외에도 훈련 중 Overfitting을 방지하기 위해 Dropout[4]과 Early Stopping 기법을 적용하였다.

이러한 기법들을 적용한 결과, CNN 모델은 Test set에서 85.71%의 정확도를 가졌으며, 이는 교구재 객체 인식 모델로서 실용적인 성능을 보인다고 판단된다.



(그림 1). 모션 인식 및 교구재 객체 인식

3. 결론

본 논문에서는 미디어가 일상화된 현대사회에서 영유아의 건강한 미디어 매체의 접근을 유도하고 능동적인 참여를 요구하는 방법을 연구하였다. 영유아 대상으로 체육 및 요가 등의 분야에서 각 개인의 특성과 요구에 맞춘 차별화된 서비스를 제공하고 인지, 운동, 사회성 발달 등의 발달 과업을 지원하며 이를 통해 일방향 미디어 문제를 효과적으로 해결할 수 있기를 기대한다. 결론적으로 본 논문에서는 상호작용할 수 있는 미디어 개발을 위해 MediaPipe Holistic을 활용하여 영유아 모션 인식을 구현하고, 실시간으로 영유아의 동작을 추적하여 특정 행동 수행 여부를 판단할 수 있는 시스템을 개발하였다. 또한, Custom CNN 모델을 통해 교구재 인식을 수행하였으며, 도출된 객체 인식 모델의 정확도는 85.71% 정도의 결과를 보였으며, 향후 연구에서 학습 데이터셋을 더 확보하여 정확도를 높이는 연구를 진행할 예정이다.

참 고 문 헌

[1] Rajani, M., & Rajani, S., "A study on adoption of OTT platforms and its growth in the post-COVID era", Webology, Vol.18 No.6, 2021. pp.417.
DOI: http://www.webology.org

[2] Heffler, K. F., Sienko, D. M., Subedi, K., McCann, K. A., & Bennett, D. S., "Association of Early-Life Social and Digital Media Experiences With Development of Autism Spectrum Disorder-Like Symptoms," JAMA Pediatrics, Vol.174 No.7, 2020, pp.690-696.
DOI: https://doi.org/10.1001/jamapediatrics.2020.0230

[3] Takeuchi, H., Taki, Y., Hashizume, H., Asano, K., Asano, M., Sassa, Y., ... & Kawashima, R., "The impact of television viewing on brain structures: Cross-sectional and longitudinal analyses", Cerebral Cortex, Vol.25 No.5, 2015. pp.1188.
DOI: https://doi.org/10.1093/cercor/bht315

[4] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R., "Dropout: A simple way to prevent neural networks from overfitting", Journal of Machine Learning Research, Vol.15 No.56, 2014. pp.193-1931.