임베디드 환경을 위한 MediaPipe 기반 행동 인식

김태현, 김명섭, 박세호*

한국전자기술연구원

taehyeon.kim@keti.re.kr, myeongseopkim@keti.re.kr, *sehopark@keti.re.kr

MediaPipe-Based Action Recognition For Embedded System

Taehyeon Kim, Myeongseop Kim, Seho Park*

Korea Electronics Technology Institute.

요 약

본 논문은 임베디드 환경에서 구동이 가능한 인공지능 기반 행동 인식 기술을 제안한다. 최근, 임베디드 시스템과 인공지능 기술의 결합을 통해 다양한 산업에서 높은 수준의 인공지능 서비스 제공이 가능해졌으며 많은 사용자가 높은 수준의 신경망 기반 인공지능 기술의 혜택을 누리고 있다. 본 논문에서 제안하는 기술은 임베디드 환경에서 최적화된 MediaPipe 프레임워크를 활용하여 자세 추정 기반의 행동 인식 기술에 대해서 다룬다. 추가로, 제안 기술의 일반화 성능 평가를 위해 총 3가지 임베디드 환경에서 실험을 수행하였으며, 각 환경에서 높은 수준의 성능을 달성하는 것을 확인하였다.

Ⅰ. 서론

최근 신경망을 활용한 인공지능 기술의 발전과 더불어 임베디드 환경에서의 신경망을 구동하기 위한 기술 개발에도 지속적인 박차를 가하고 있다. 이를 통해 현재의 사용자들은 서비스용 로봇, 자율주행 자동차, 지능형 CCTV 등 임베디드 기술과 인공지능 기술의 결합으로 탄생한 다양한 산업 혜택을 누리고 있으며, 산학 및 국가 단위의 천문학적인 투자는 해당 분야의 기술이 큰 잠재력을 보유하고 귀납적으로 증명하고 있다 [1].

이러한 기술 연구 및 개발 경향과 발맞추어, 본 논문은 임베디드 환경에서 구동할 수 있는 자세 추정 기반의 행동 인식 기술을 제안한다. 본 기술의 차별점은 임베디드 시스템을 지향하고 있으므로, Google Research 팀에서 제공하는 MediaPipe [2] 프레임워크를 활용하는 것이며, 이를 통해 대표적인 임베디드 환경인 안드로이드 환경에서의 동작을 목적한다.

Ⅱ. 본론

자세 추정 기술이란, 2차원 데이터 형태의 영상을 신경망이 입력받아 사용자가 사전에 정의한 핵심 위치의 영상 내 좌표정보를 추정하는 것을 의미한다. MediaPipe의 자세 추정 기술은 영상 내 존재하는 사람의 총 33개의 핵심 좌표를 추출하며 이는 다음과 같다. {1. nose, 2. left eye inner, 3. left eye, 4. left eye outer, 5. right eye inner, 6. right eye, 7. right eye outer, 8. left ear, 9. right ear, 10. mouth left, 11. mouth right, 12. left shoulder, 13. right shoulder, 14. left elbow, 15. right elbow, 16. left wrist, 17. right wrist, 18. left pinky, 19. right pinky, 20. left index, 21. right index, 22. left thumb, 23. right thumb, 24. left hip, 25. right hip, 26. left knee, 27. right knee, 28. left ankle, 29. right ankle, 30. left heel, 31. right heel, 32. left foot index, 33. right foot index.}

위와 같은 다양한 인간의 핵심 좌표를 추정하는 과정은 영상 내 인간이 취하고 있는 자세를 정확하게 예측할 수 있도록 한다. 또한, 위의 다양한 핵심 좌표를 추정하기 위한 신경망 학습은 신경망 스스로가 인간의 신체적 특징을 이해할 수 있도록 유도할 뿐만 아니라, 핵심 좌표 간의 상관관계 또한 동시에 학습할 수 있도록 한다.

MediaPipe를 이용한 영상 내 인간의 핵심 좌표 추출 이후, 행동 인식을 수행하기 위해 Stanford 40 Actions 데이터세트를 활용한다 [3]. Stanford 40 Actions 데이터세트는 총 9,532장의 영상 데이터로 구성되어 있으며 40개의 행동이 정의되어있는 데이터세트다. 행동별 영상의 개수는 대략 180-300 영상으로 구성된다. 제안하는 기법은 행동 인식을 수행하기 위해, Stanford 40 Action 데이터에 MediaPipe를 적용하여 데이터별 핵심 좌표를 추정하였으며, 추후 임베디드 환경에서 인식하는 핵심 좌표의 위치 변량이 Stanford 40 Action 데이터의 변화량과 일치하는 경향성을 보이는지를 판단하여 행동 인식을 수행한다.

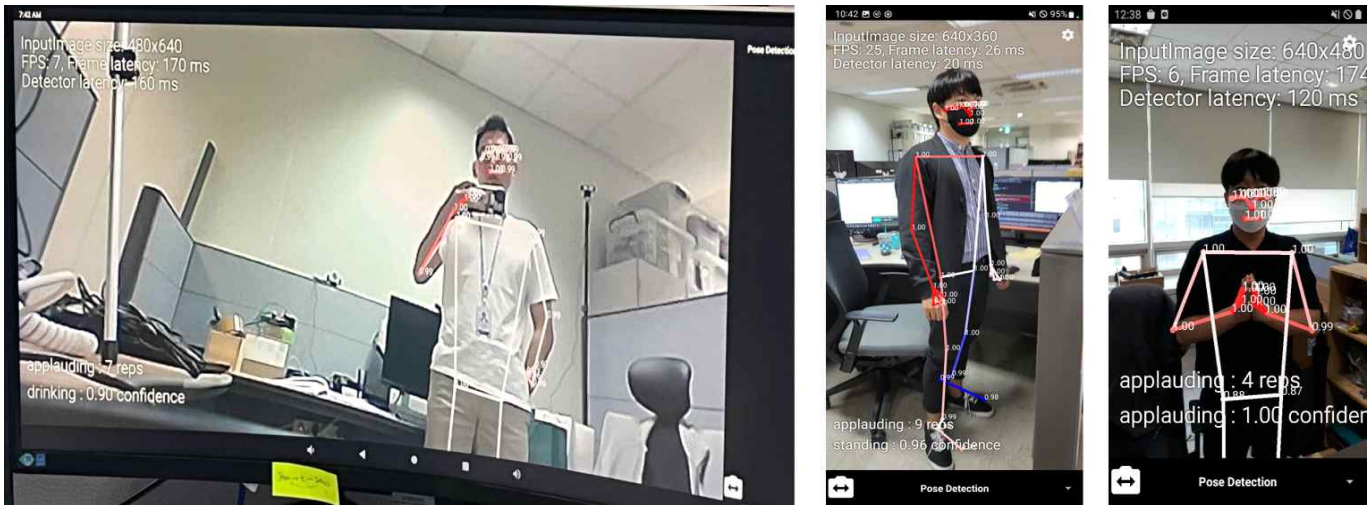

그림 1 임베디드 환경별 제안 기법의 행동 인식 결과 영상 **(좌측)** Qber Board, "Drinking"인식, 7FPS, **(중앙)** Galaxy Note 20, "Standing"인식, 25FPS, **(우측)** Galaxy J5, "Applauding", 6FPS.

표 1 임베디드 환경별 행동 인식 처리 시간

임베디드 환경	Application Processor	행동 인식 처리 시간
Qber Board	Rockship RK3399	5-9 frames/sec
SAMSUNG Galaxy Note 20	Snapdragon 855 SDM855	24-30 frames/sec
SAMSUNG Galaxy J5	Exynos 7870	6-9 frames/sec

비록, Recurrent Neural Network (RNN) 과 같은 기법을 활용하여 입력 동영상의 시간별 핵심 좌표의 변화량을 인식하는 기법으로 행동 인식 기술을 제안할 수 있지만, 이는 별도의 추가적인 계산 복잡도를 요구하는 기법이므로 임베디드 환경에 적합하지 않다.

그러므로, 제안 기법은 임베디드 환경에서 MediaPipe의 결과만을 바탕으로 행동 인식이 가능한 딕셔너리 기반의 행동 인식을 제안한다. 이는 별도의 신경망을 활용하는 기법 대비 메모리 복잡도와 계산 복잡도 모두 긍정적인 효과를 가질 수 있다.

제안 기술의 일반화 성능을 평가하기 위해, 다양한 임베디드 환경에서의 구현 가능성을 평가하였고, 각 임베디드 실험 환경에서 제안 기법의 행동 인식 실험 결과는 표 1에서 확인할 수 있다. 그림 1은 제안하는 행동 인식 기술의 실제 실험 영상을 나타낸다. 해당 그림에서 볼 수 있듯이, 모든 임베디드 환경에서 제안 기법이 정상적으로 동작하는 것이 확인할 수 있으며, 표 1과 그림 1을 통해 제안하는 행동 인식 기법의 일반화 성능의 증명이 가능하다.

Ⅲ. 결론

본 논문에서는 임베디드 환경에 적합한 신경망 기반의 인공지능 기반 인간 행동 인식 기술을 제안한다. 제안된 기법은 MediaPipe의 인간의 핵심 좌표 추출 기술을 적극적으로 활용하고 있으며, 행동 인식의 수행을 위해서는 공인 데이터세트인 Stanford 40 Action을 활용함으로써 임베디드 환경임에도 불구하고 높은 행동 인식 성능의 달성이 가능하다. 제안하는 기술의 차별성은 행동 인식을 수행하는 과정에서 별도의 신경망을 따로 구성하는 것이 아닌 딕셔너리 기법을 활용함으로써 기존 행동 인식 기법 대비 높은 메모리 및 계산 효율성을 달성할 수 있었으며, 이와 동시에 높은 성능의 행동 인식 결과도 제공할 수 있다. 본 기술은 임베디드 기술과 신경망 기술이 결합한 다양한 신산업 및 서비스에 적극적으로 활용할 수 있으므로 인공지능 주요 산업의 매우 중요한 중추 기술로 간주할 수 있다.

ACKNOWLEDGMENT

This work was supported by the IT R&D program of MOTIE/KEIT (20009543, Development of Smart Home Manager for providing Edge AI-based Emotional Services).

참 고 문 헌

[1] Sun, Zehua, et al. "Human action recognition from various data modalities: A review." IEEE transactions on pattern analysis and machine intelligence (2022).

[2] Lugaresi, Camillo, et al. "Mediapipe: A framework for building perception pipelines." arXiv preprint arXiv:1906.08172 (2019).

[3] B. Yao, X. Jiang, A. Khosla, A.L. Lin, L.J. Guibas, and L. Fei-Fei. Human Action Recognition by Learning Bases of Action Attributes and Parts. Internation Conference on Computer Vision (ICCV), Barcelona, Spain. November 6-13, 2011.