Journal of The Institute of Electronics and Information Engineers Vol.63, NO.3, March 2026 https://doi.org/10.5573/ieie.2026.63.3.63
ISSN 2287-5026(Print) / ISSN 2288-159X(Online)

논문 2026-63-3-7

메타버스 환경에서 인구통계학적 속성정보를 반영한 인간 행동 인식 모델의 성능 향상 연구

(Enhancing Human Action Recognition with Demographic Attribute Information in Metaverse Environments)

한 석 호*, 김 종 성*

(Seok-Ho Han and Jong-Sung Kim©)

요 약

본 연구는 메타버스 환경에서 성별 및 연령대와 같은 인구통계학적 속성정보를 반영하여 인간 행동 인식(HAR) 모델이 인식 성능을 향상시키는 방법을 제안한다. 이를 위해 성별과 연령대에 따라 사용자를 그룹화하고, XR 장비를 착용한 상태에서 골프 스윙과 볼링 투구 동작 데이터를 수집하여 세부 동작 단계로 라벨링하였다. 수집한 데이터를 기반으로 LSTM, GRU, 1D-CNN, Transformer 모델을 학습시켜 인구통계학적 속성정보 반영 여부에 따른 모델 성능을 비교하였다. 실험 결과 인구통계학적 속성정보를 반영한 모든 모델에서 골프 스윙 동작은 평균 약 3%, 볼링 투구 동작은 평균 약 1.5%의 성능 향상을 확인하였다. 특히 가장 성능이 우수했던 Transformer 모델은 각 동작에서 3.8%, 1.8%의 성능 향상을 보였다. 또한 성별 및 연령대별 세부 분석에서도 모든 그룹에서 성능이 개선되었으며, t-SNE 시각화 결과에서도 인구통계학적 속성정보를 반영했을 때 특징이 더 명확하게 분리되는 것을 확인하였다. 이러한 결과를 통해 인구통계학적 속성정보가 HAR 모델의 인식 성능 향상과 사용자 그룹 간 성능 개선에 영향을 주는 것을 확인할 수 있었으며, 이는 향후 포용적이고 사용자 친화적인 메타버스 환경 구현을 위한 행동 인식 기술 개발에 의미 있는 기여할 수 있을 것으로 기대된다.

Abstract

This study proposes a method to enhance the recognition performance of Human Action Recognition (HAR) models by incorporating Demographic Attribute Information such as gender and age group in a metaverse environment. To achieve this, participants were grouped by gender and age, and action data of golf swings and bowling throws were collected using full-body XR devices. Each action sequence was labeled into detailed sub-actions for fine-grained analysis. Based on the collected data, LSTM, GRU, 1D-CNN and Transformer models were trained and compared according to whether Demographic Attribute Information was included in the input. Experimental results show that the models incorporating Demographic Attribute Information consistently outperformed those without it, achieving an average improvement of approximately 3.0% for golf swings and 1.5% for bowling throws. In particular, the Transformer model demonstrated the highest performance, showing improvements of 3.8% and 1.8% for each action, respectively. Furthermore, group-wise evaluation revealed performance improvements across all gender and age groups, while t-SNE visualization demonstrated clearer separability among feature clusters when Demographic Attribute Information was utilized. These findings confirm that integrating Demographic Attribute Information can improve both the overall recognition performance and the balance across user groups in HAR models, contributing to the development of more inclusive and user-friendly metaverse environments.

Keywords : Human action recognition, Demographic attribute information, Transformer, LSTM, GRU, 1-D CNN

*비회원, 한국전자통신연구원 콘텐츠연구본부(Creative Contents Research Division, Electronics and Telecommunications Research Institute)
© Corresponding Author(E-mail : js.kim@etri.re.kr)

※ 본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2023-2024년도 문화기술 연구개발사업으로 수행되었음(과제명:디지털 문화 포용성 지원 메타버스 구축을 위한 유니버설 XR 플랫폼 기술개발, 과제번호:RS-2023-00270006, 기여율:80%; 신체 장애인 온라인 활동 접근성 향상을 위한 배리어프리 체험형 XR 콘텐츠 기술개발, 과제번호: RS-2024-00396700, 기여율:20%).

Received ; December 10, 2025 Revised ; December 24, 2025 Accepted ; December 30, 2025

Ⅰ. 서 론

메타버스(Metaverse)는 가상(Meta)과 세계(Universe)의 합성어로 현실과 같은 가상 공간에서 아바타를 통해 소통하고 게임, 교육, 관광 등 다양한 콘텐츠를 체험할 수 있는 환경을 의미한다[1]. 이러한 메타버스 환경에서는 일반적으로 XR 헤드셋, 컨트롤러, 트래커 등 XR 기기를 착용하여 콘텐츠에 참여할 수 있다.

초기 XR 기술은 고가의 XR 장비, 무겁고 불편한 착용감, 낮은 해상도 및 제한된 콘텐츠 등으로 인해 대중적으로 활용되기에는 한계가 있었다. 그러나 최근에는 상대적으로 저렴해진 XR 장비, 무선 연결 기능, 고해상도 디스플레이, 경량화된 하드웨어와 함께 다양한 산업 및 소비자용 콘텐츠가 개발되면서 XR 기술의 접근성과 편의성이 향상되었다. 이러한 발전으로 인해 많은 사용자가 메타버스 콘텐츠에 쉽게 참여할 수 있게 되었으며, 메타버스 산업은 빠르게 성장하고 있다[2].

메타버스 환경에서는 XR 장비를 통해 손 흔들기, 고개 끄덕임 등 사용자의 신체 동작을 인식하여 사용자들 간 원활한 의사소통과 상호작용을 할 수 있다. 이러한 기술에는 주로 인간 행동 인식(Human Action Recognition, HAR) 기술이 활용된다. HAR은 착용 중인 기기에서 획득한 센서 데이터나 카메라 영상을 기반으로 사용자의 움직임과 동작을 자동으로 식별하는 기술이다. 주로 사람의 동작이 필요한 스포츠, 예술, 재활치료 등 다양한 응용 분야에서 활용되고 있다[3, 4].

그러나 대부분 기존 HAR 연구는 성별이나 연령대와 같은 사용자의 인구통계학적 속성정보를 고려하지 않고 있다[5, 6]. 이는 메타버스에 참여 중인 특정 사용자 그룹에서 행동 인식 정확도 저하로 이어질 수 있다. 예시로, 일반적으로 청소년이나 청년 사용자 그룹은 신체 가동 범위가 넓고 동작 속도가 빠르지만, 노년층은 제한된 가동 범위와 느린 동작 속도를 보인다. 성별 역시 신체적 차이로 인해 동작 패턴에 영향을 끼칠 수 있다. 이러한 차이를 고려하지 않는다면 동일 동작을 수행하더라도 특정 계층에서는 행동 인식이 저하될 수 있다. 이는 특정 사용자 그룹의 원활한 메타버스 참여를 제한하는 요인으로 작용할 수 있으며, 전 연령층이 차별 없이 콘텐츠를 활용하고 즐기는 데 제약이 발생할 수 있다.

따라서 본 연구에서는 이러한 한계를 보완하기 위해 성별과 연령대 두 가지 인구통계학적 속성정보를 HAR 모델에 통합하는 방법을 제안한다. 이를 위해 성별과 연령대에 따라 참가자를 모집하여 사용자 그룹을 구분하고 골프 스윙 및 볼링 투구 동작 데이터를 수집하였다. 이후 LSTM, GRU, 1D-CNN, Transformer 네 가지 모델에 인구통계학적 속성정보를 입력으로 추가하여 학습을 진행하였다. 이를 통해 인구통계학적 속성정보의 반영 여부에 따른 모델 간 성능 차이를 비교하고, 사용자 그룹별 성능 평가를 통해 인구통계학적 속성정보가 모델 성능에 미치는 영향을 분석하였다.

Ⅱ. 관련 연구

2.1. 인간 행동 인식 연구

인간 행동 인식 연구는 시계열 데이터를 기반으로 한 비전 기반 접근법과 센서 기반 접근법으로 구분된다.

비전 기반 접근법은 스마트폰이나 CCTV 등 일상에서 쉽게 접할 수 있는 RGB 카메라 영상을 활용한다[7]. RGB 영상은 사람의 외형, 색상, 주변 환경 등 다양한 시각 정보를 포함하고 있기에 사람이 동작을 보고 이해하는 것과 유사하게 행동을 직관적으로 예측할 수 있다는 장점이 있다. 하지만 조명 변화, 배경의 복잡성, 촬영 시점의 변화 등 외부 환경 변화에 민감하여 모델이 실제 동작 패턴보다 부수적인 요인에 과적합 될 수 있다[8].

이러한 한계를 보완하기 위해 RGB-D 카메라가 활용되기도 한다. RGB-D 카메라는 RGB 픽셀 정보에 깊이 정보를 추가하여 객체와 카메라 사이의 거리를 측정하고 3차원 공간 구조를 파악할 수 있기에 조명이나 배경 변화 등 외부 환경 변화의 영향을 덜 받아 동작 자체에 집중할 수 있다는 장점이 있다. 다만, 측정 가능 거리가 제한적이고 RGB 카메라만큼 보급률이 높지 않다는 한계가 있다[9].

센서 기반 접근법은 IMU(관성 측정 장치), 가속도계, 자이로스코프 센서를 활용하여 사용자의 움직임 시계열 정보를 활용한다[10]. 이 방식은 조명이나 배경 같은 외부 환경 변화에 영향을 받지 않기 때문에 안정적으로 동작을 측정할 수 있다는 장점이 있다. 그러나 같은 동작이라도 사용자별 움직임 특성이나 센서 착용 위치와 방향에 따라 데이터 분포가 달라지는 도메인 이동(domain shift) 한계가 있다[11].

최근에는 인간 행동 인식 연구는 카메라와 센서를 통해 획득한 원시 데이터를 스켈레톤 데이터로 변환하여 행동을 예측하는 방식으로 확장되고 있다. 스켈레톤 데이터는 사람의 주요 신체 관절의 2D 또는 3D 좌표를 추출하고 이를 뼈대 형태로 연결한 데이터 구조로 외부 환경 요인에서 발생하는 불필요한 정보를 제거하고 사

람의 움직임에 집중할 수 있다[12].

스켈레톤 데이터는 RGB 및 RGB-D 카메라에서 자세 추정(Pose Estimation) 기술을 통해 스켈레톤 데이터를 획득할 수 있다. 특히 RGB-D 카메라는 깊이 정보를 활용하여 더 정확하고 안정적인 3D 스켈레톤 좌표를 얻을 수 있다[13]. 한편 센서 데이터는 직접적으로 관절 좌표를 제공하지 않지만, 여러 IMU 센서를 신체 부위에 착용하여 각 센서의 움직임 데이터를 기반으로 스켈레톤과 유사한 관절의 움직임을 추정하는 방식으로 사용한다[14].

이처럼 스켈레톤 데이터는 다양한 장치로부터 획득할 수 있으며, 사람의 구조적 움직임을 효과적으로 표현할 수 있기에 시계열 기반 모델이나 그래프 신경망과 같은 딥러닝 모델의 입력으로 널리 사용되고 있다. 하지만 스켈레톤 데이터 추출이 어려운 환경이나 센서 원시 신호 자체에 담긴 미세한 특징이 중요한 경우 여전히 전통적인 시계열 기반 접근법이 유효하게 활용된다.

2.2. 딥러닝 기반 인간 행동 인식 연구

딥러닝 기반 행동 인식 연구는 시계열 데이터의 시간적 특징을 학습하는 방향으로 발전하였다. 대표적으로 RNN 계열의 LSTM과 GRU 모델은 시간에 따른 동작 패턴의 연속성과 맥락을 학습하는 데 효과적이다. 하지만 시퀀스가 길어지면 이전 정보를 잊어버리는 기울기 소실(Vanishing Gradient) 문제가 발생할 수 있고 데이터를 순차적으로 처리하기에 연산 효율성에 한계가 있다[15].

이와 함께 활용되던 1D-CNN 모델은 합성곱 연산을 적용하여 지역적 시간 패턴을 효율적으로 추출할 수 있으며, 병렬 연산이 가능하고 실시간성이 우수하다는 장점이 있다[16]. 하지만 고정된 커널 크기로 인해 장기적인 시계열 의존성이나 신체의 구조적 관계를 충분히 반영하기 어렵다는 한계가 존재한다.

최근에는 이러한 시계열 데이터 처리의 한계를 보완하기 위해 자연어 처리 분야에서 사용되는 Transformer 구조가 인간 행동 인식 분야에도 활용되고 있다[17]. Transformer는 Self-Attention 메커니즘을 기반으로 시퀀스 내 모든 시점의 관계를 병렬적으로 계산한다. 이를 통해 기존 RNN 계열 모델이 가진 장기 의존성 문제를 해결할 수 있다는 장점이 있다. 이러한 장점을 기반으로 Transformer 구조를 활용하여 신체 관절의 시공간 정보를 통합적으로 학습하고 복잡한 신체 동작 패턴을 인식하는 연구들이 활발히 진행되고 있다[18].

이처럼 딥러닝 기반 행동 인식 행동 모델은 초기의 단순한 시계열 패턴 학습에서 신체의 공간적 구조와 복잡한 시공간적 맥락을 통합적으로 학습하는 방향으로 발전하고 있다. 이에 본 연구에서는 앞서 언급한 LSTM, GRU, 1D-CNN, Transformer 모델에 인구통계학적 속성정보 여부에 따른 각 모델의 성능 향상을 비교하고 검증하고자 한다.

Ⅲ. 본 론

3.1. 데이터 전처리

본 연구에서 활용된 모든 동작 데이터는 그림 1과 같은 환경에서 XR 헤드셋과 양손 컨트롤러, 5개의 바디 트래커(허리, 양손목, 양발목)를 포함한 총 8개의 XR 장비를 활용하여 데이터를 획득하였다. 각 프레임은 8개의 장비에서 머리, 양손, 허리, 양손목, 양발목 등 주요 관절 지점의 3차원 위치 좌표와 4차원 회전 정보를 포함하며, 각 장비당 7차원 원시 데이터로 구성된다.

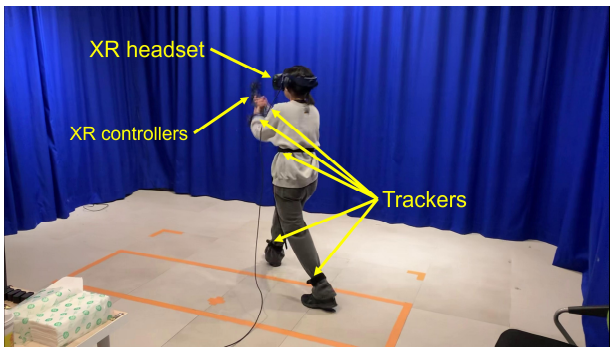

그림 1. XR 장비를 활용한 동작 데이터 획득 환경
Fig. 1. Configuration of action data acquisition using XR devices.

8개의 장비에서 획득한 7차원 원시 데이터를 그대로 모델의 입력으로 사용하기에는 차원 수가 증가하고 연산 복잡도가 증가하기 때문에, 각 장비의 원시 데이터를 위치 및 회전 정보를 대표하는 2차원으로 축소하였다. 3차원 위치 좌표는 벡터 크기를 계산하여 관절이 원점으로부터 떨어진 거리를 나타내는 스칼라 값으로 변환하였으며, 4차원 회전 정보는 벡터 성분의 크기를 계산하여 회전 강도를 나타내는 스칼라 값으로 변환하였다. 이렇게 변환된 2차원 특징에 대해 1차 미분과 2차 미분을 통해 속도와 가속도 정보를 추가로 계산하여 모델이 위치 및 회전 정보와 함께 동작의 변화까지 함께 학습할 수 있도록 데이터를 전처리하였다.

전처리 된 데이터는 하나의 프레임에서 8개의 장비

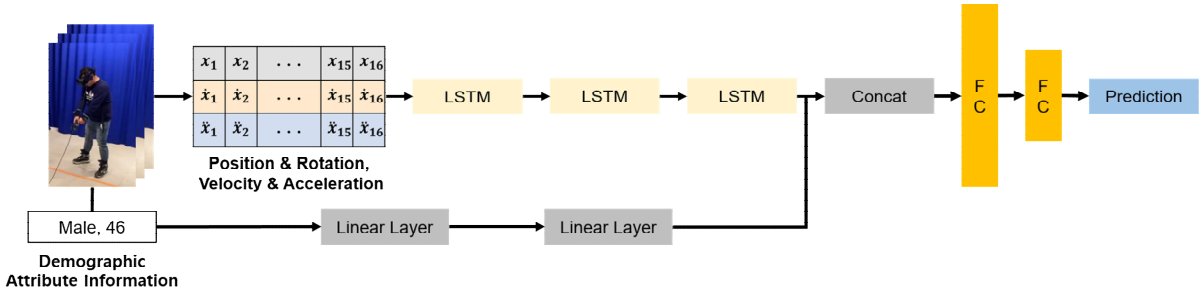

그림 2. 제안하는 LSTM 모델 구조
Fig. 2. Proposed LSTM model architecture.

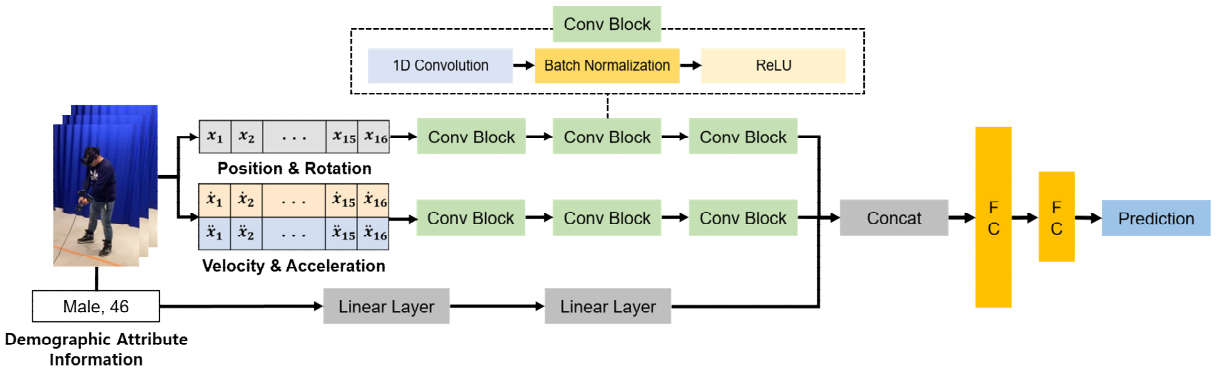

그림 3. 제안하는 1D-CNN 모델 구조
Fig. 3. Proposed 1D-CNN model architecture.

에서 얻은 축소된 2차원 특징(16차원), 1차 미분 특징(16차원), 2차 미분 특징(16차원) 총 48개의 특징으로 구성된다. 이후 일정 길이의 프레임을 묶는 윈도윙(windowing)을 통해 입력 시퀀스를 구성하였으며, 각 윈도우는 3개의 프레임으로 설정하였다. 이때 각 윈도우의 중앙 프레임 라벨을 정답 라벨로 지정하여 시간적 맥락을 반영하였다.

3.2. 모델 설계

본 연구에서는 사용자 인구통계학적 속성정보가 동작 인식 성능에 미치는 영향을 검증하기 위해, HAR 분야에서 널리 활용되는 LSTM, GRU, 1D-CNN, Transformer 모델을 비교 대상으로 선정하였다.

LSTM 모델 구조는 그림 2와 같다. 모델은 전처리된 위치 및 회전, 속도와 가속도 시계열 데이터를 입력으로 받아 3개의 LSTM 셀을 통해 시간적 의존성을 학습한다. 이때 사용자의 인구통계학적 속성정보(성별, 연령대)는 원-핫 인코딩(One-Hot Encoding) 방식으로 변환되어 성별 2차원과 연령대 4차원 총 6차원 벡터로 구성된다. 이 벡터는 2개의 Linear Layer를 순차적으로 통과하여 64차원 임베딩 벡터로 변환된다. 이후 임베딩된 인구통계학적 속성정보 벡터는 3개의 LSTM 셀의 최종 출력과 결합된다. 이렇게 결합된 특징은 Fully Connected(FC) Layer를 통과하여 최종 동작을 예측한다. GRU 모델은 LSTM 모델과 동일한 구조로, LSTM 셀 대신 GRU 셀로 대체하여 구성하였다.

1D-CNN 모델 구조는 그림 3과 같다. 전처리 된 시계열 데이터를 입력받는 방식은 동일하지만, 위치 및 회전 정보와 속도와 가속도 정보를 2개의 독립적인 스트림으로 분리하여 처리한다. 이때 각 스트림은 1D Convolution, Batch Normalization, ReLU로 구성된 3개의 Convolution Block을 거쳐 특징을 추출한다. 인구통계학적 속성정보는 앞서 설명한 LSTM 모델과 동일하게 임베딩 과정을 거친 후, 두 스트림의 출력 특징과 결합되어 최종적으로 FC Layer를 통과하여 최종 동작을 예측한다.

Transformer 모델 구조는 그림 4와 같다. 전처리 된 시계열 데이터는 순서 정보를 위해 Positional Encoding이 더해진 후 Encoder에 입력된다. Encoder는 Multi-Head Attention과 Feed-Forward Network (FFN)로 구성되며, 이를 통해 데이터 내의 전역적인 상관관계를 학습한다. 이후 Encoder의 출력은 하나의 특

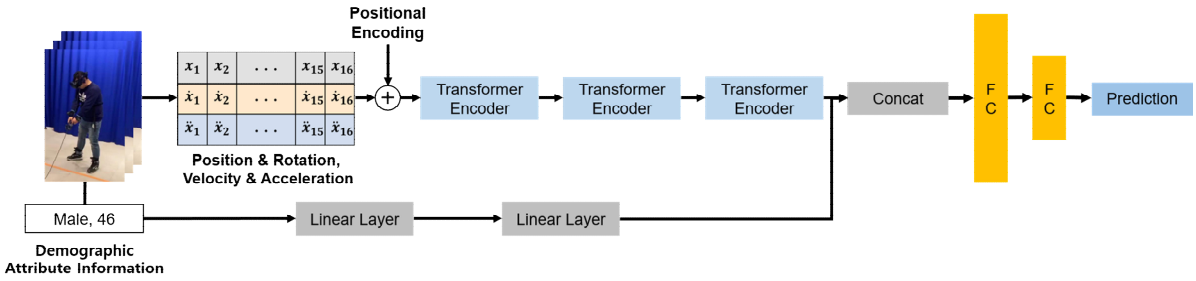

그림 4. 제안하는 Transformer 모델 구조
Fig. 4. Proposed Transformer model architecture.

징 벡터로 압축되고, 사용자의 인구통계학적 속성정보 또한 앞선 모델들과 동일하게 임베딩 된 후 압축된 시계열 특징과 결합되어 최종 FC Layer를 통해 최종 동작을 예측한다.

본 연구에서는 제안된 네 가지 모델에 대해 인구통계학적 속성정보의 결합 유무에 따른 성능 차이를 검증하기 위해 각 모델을 인구통계학적 속성정보를 포함한 버전(Demog)과 포함하지 않은 버전(No-Demog)으로 나눠 실험을 진행하였다. 이때 인구통계학적 속성정보를 포함하지 않은 모델은 인구통계학적 속성정보 임베딩을 위한 Linear Layer를 제거하고 시계열 데이터만을 사용하여 동작을 예측하도록 구성하였다. 이를 통해 성별과 연령대와 같은 사용자 인구통계학적 속성정보가 HAR 모델의 동작 인식 정확도 향상에 미치는 영향을 분석하였다.

Ⅳ. 실 험

4.1. 데이터 수집

본 연구는 성별 및 연령대에 따른 모델의 성능 차이를 평가하기 위해 총 8개의 사용자 그룹에서 데이터를 수집하였다. 사용자 그룹은 2개의 성별(남성, 여성), 4개의 연령대(청소년, 청년, 중년, 노년)로 구성하였으며 각 그룹당 10명씩 총 80명의 참가자를 모집하였다.

모든 참가자는 XR 헤드셋, 컨트롤러, 트래커를 착용한 상태에서 골프 스윙 동작과 볼링 투구 동작을 각각 60회씩 반복하여 수행하였다. 이를 통해 골프 스윙 데이터는 그룹별 600개씩 총 4,800개의 데이터를 수집하였으며, 볼링 투구 데이터는 노년 그룹의 한 명을 제외한 7개 그룹에서 600개씩 총 4,740개의 데이터를 수집하였다.

수집한 데이터는 그림 5와 같이 7개의 세부 동작 단계로 라벨링을 진행하였다. 골프 스윙 동작은 어드레스

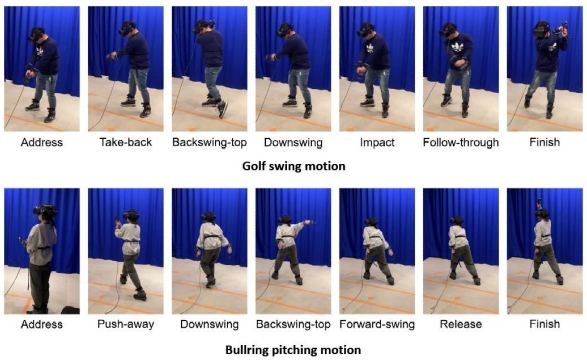

그림 5. 골프 스윙 및 볼링 투구 동작 세부 라벨링
Fig. 5. Detailed labeling of golf swing and bowling throwing actions.

(Address), 테이크백(Takeback), 백스윙 탑(Backswing top), 다운스윙(Downswing), 임팩트(Impact), 팔로우 스루(Follow through), 피니시(Finish)로 구성하였으며, 볼링 투구 동작은 어드레스(Address), 푸시 어웨이(Push away), 다운스윙(Down swing), 백스윙 탑(Backswing top), 포워드 스윙(Forward swing), 릴리스(Release), 팔로우 스루(Follow through)로 구성하였다.

4.2. 모델 학습

모델 학습은 LSTM, GRU, 1D-CNN, Transformer 네 가지 모델에 대해 골프 스윙 동작 데이터와 볼링 투구 동작 데이터를 각각 독립적으로 학습하였다. 데이터는 학습, 검증, 테스트 데이터를 6:2:2 비율로 분할하였으며, 성별과 연령대 비율이 모든 데이터에 동일하게 유지되도록 구성하였다. 골프 스윙 동작 데이터는 학습 2,880개, 검증 960개, 테스트 960개로 구성하였으며, 볼링 투구 데이터는 학습 2,820개, 검증 960개, 테스트 960개로 분할하였다. 모델 학습은 NVIDIA GeForce RTX 4080 GPU 환경에서 배치 크기(Batch Size)는 128, 학습률(Learning Rate) 0.001, 에포크(Epoch) 30

표 1. 골프 스윙 동작 성능 평가 결과
Table 1. Performance evaluation of golf swing action recognition.

	Model	Acc	Prec	Rec	F1-score
Demog	LSTM	0.920	0.894	0.902	0.897
	GRU	0.920	0.902	0.900	0.900
	1D-CNN	0.931	0.783	0.774	0.778
	Transformer	**0.935**	**0.911**	**0.922**	**0.915**
No-Demog	LSTM	0.887	0.853	0.860	0.854
	GRU	0.908	0.886	0.884	0.883
	1D-CNN	0.920	0.759	0.753	0.756
	Transformer	0.908	0.887	0.870	0.877

으로 설정하고, 조기 종료(Early stopping)를 10으로 설정하여 과적합을 방지하였다.

4.3. 실험 결과

표 1과 표 2는 각각 골프 스윙 동작과 볼링 투구 동작을 학습시킨 LSTM, GRU, 1D-CNN, Transformer 모델의 테스트 데이터에 대한 성능 평가 결과를 나타낸다. 여기에서 Demog는 모델에 사용자의 인구통계학적 속성정보를 반영한 모델, No-Demog는 모델에 사용자의 인구통계학적 속성정보를 반영하지 않은 모델을 의미한다.

골프 스윙 동작의 경우 LSTM 모델은 F1-score가 85.4%에서 89.7%로 약 4.3%의 성능 향상되었으며, GRU 모델은 88.3%에서 90%로 약 1.7%, 1D-CNN 모델은 75.6%에서 77.8%로 약 2.2%, Transformer 모델은 87.7%에서 91.5%로 약 3.8% 성능 향상을 보인 것을 확인할 수 있었다. 결과적으로 네 가지 모델 중 Transformer 모델이 인구통계학적 속성정보를 반영했을 때 가장 높은 F1-score 성능을 보였으며, LSTM 모델이 가장 큰 성능 향상을 보인 것을 확인하였다.

볼링 투구 동작에서도 F1-score를 기준으로 인구통계학적 속성정보를 반영한 모델이 전반적으로 더 우수한 성능을 보였다. LSTM 모델은 88.1%에서 89.2%로 약 1.1%, GRU 모델은 87.8%에서 89%로 약 1.2%, 1D-CNN 모델은 73.5%에서 75.4%로 약 1.9%, Transformer 모델은 87.8%에서 89.6%로 약 1.8% 성능 향상을 보인 것을 확인할 수 있었다. 결과적으로 네 가지 모델 중 Transformer 모델이 인구통계학적 속성정보를 반영했을 때 가장 높은 F1-score 성능을 보였으며, 1D-CNN 모델이 가장 큰 성능 향상을 보인 것을 확인하였다.

표 2. 볼링 투구 동작 성능 평가 결과
Table 2. Performance evaluation of bowling throwing action recognition.

	Model	Acc	Prec	Rec	F1-score
Demog	LSTM	0.898	0.890	0.894	0.892
	GRU	0.897	0.888	0.892	0.890
	1D-CNN	0.887	0.762	0.747	0.754
	Transformer	**0.901**	**0.893**	**0.900**	**0.896**
No-Demog	LSTM	0.890	0.885	0.878	0.881
	GRU	0.887	0.882	0.877	0.878
	1D-CNN	0.864	0.744	0.729	0.735
	Transformer	0.890	0.890	0.871	0.878

앞선 실험에서 모델 전체 수준에서 성능 향상을 확인하였다. 이어서 성별과 연령대에 따른 세부적인 성능 차이를 분석하기 위해 추가 실험을 진행하였다.

표 3은 골프 스윙 동작에서 가장 높은 성능을 보인 Transformer 모델을 기반으로 한 그룹별 성능 평가 결과를 나타낸다. 그룹은 남성, 여성, 청소년, 청년, 중년, 노년 6개 그룹으로 구성하였다.

골프 스윙 동작의 그룹별 성능 평가 결과를 살펴보면, 인구통계학적 속성정보를 반영한 모델에서 중년 그룹의 F1-score가 95.5%로 가장 높은 성능을 보인 것을 확인하였다. 이는 중년 그룹의 동작이 다른 그룹에 비해 일관성이 높고 패턴이 명확해서 모델이 시계열 특

표 3. 그룹별 골프 스윙 동작 성능 평가 결과
Table 3. Performance evaluation by group for golf swing action recognition.

	Group	Acc	Prec	Rec	F1-score
Demog	Male	0.923	0.893	0.918	0.903
	Female	0.947	0.927	0.928	0.927
	Teenager	0.949	0.929	0.930	0.929
	Youth	0.910	0.888	0.886	0.886
	Middle-aged	0.970	0.966	0.948	0.955
	Senior	0.907	0.872	0.920	0.886
	Total	**0.935**	**0.911**	**0.922**	**0.915**
No-Demog	Male	0.898	0.869	0.872	0.870
	Female	0.917	0.904	0.869	0.883
	Teenager	0.909	0.900	0.863	0.876
	Youth	0.874	0.846	0.816	0.827
	Middle-aged	0.951	0.940	0.913	0.925
	Senior	0.892	0.862	0.883	0.870
	Total	0.908	0.887	0.870	0.877

성을 안정적으로 학습했기 때문으로 해석된다. 반면 노년 그룹의 경우 F1-score가 88.6%로 가장 낮은 성능을 보인 것을 확인하였다. 이는 가동 범위의 제한과 느린 동작 속도 등의 제약 사항이 모델 학습에 영향을 준 것으로 해석된다.

인구통계학적 속성정보를 반영한 모델은 모든 그룹에서 반영하지 않았을 때보다 일관된 성능 향상을 보였으며, 남성은 약 3.3%, 여성은 약 4.4%, 청소년은 약 5.3%, 청년은 약 5.9%, 중년은 약 3.0%, 노년은 약 1.6%의 성능 향상을 확인하였다. 특히 청년 그룹에서 약 5.9%로 가장 높은 성능 향상을 보였으며, 다른 그룹에서도 일관된 성능 향상을 확인하였다. 이는 성별과 연령대 정보를 반영하는 것이 그룹 간의 동작 패턴 차이를 학습하여 모델이 일반화 성능을 개선한 것으로 해석된다.

표 4. 그룹별 볼링 투구 동작 성능 평가 결과
Table 4. Performance evaluation by group for bowling throwing action recognition.

	Group	Acc	Prec	Rec	F1-score
Demog	Male	0.877	0.872	0.883	0.874
	Female	0.925	0.919	0.918	0.918
	Teenager	0.910	0.900	0.892	0.895
	Youth	0.903	0.888	0.886	0.885
	Middle-aged	0.971	0.965	0.967	0.965
	Senior	0.837	0.843	0.857	0.835
	Total	**0.901**	**0.893**	**0.900**	**0.896**
No-Demog	Male	0.870	0.871	0.847	0.856
	Female	0.909	0.909	0.894	0.899
	Teenager	0.867	0.867	0.820	0.834
	Youth	0.875	0.871	0.839	0.844
	Middle-aged	0.958	0.955	0.952	0.953
	Senior	0.974	0.863	0.859	0.860
	Total	0.890	0.890	0.871	0.878

이어서 볼링 투구 동작에서도 표 4와 같이 성능이 가장 높았던 Transformer 모델을 기반으로 그룹별 성능 평가를 진행하였다. 실험 결과 중년 그룹의 F1-score가 96.5%로 가장 높은 성능을 보인 것을 확인할 수 있었으며, 노년 그룹은 F1-score가 83.5%로 가장 낮은 성능을 보인 것을 확인하였다. 세부적으로 살펴보면 노년 그룹을 제외한 모든 그룹에서 인구통계학적 속성정보를 반영한 모델이 일관된 성능 향상을 보였으며, 남성은 약 1.8%, 여성은 약 1.9%, 청소년은 약 6.1%, 청년은 약 4.1%, 중년은 약 1.2%의 성능 향상을 확인하였다. 반면 노년 그룹은 약 2.5%의 성능 하락을 확인하였다. 이는 앞서 데이터 수집 과정에서 노년 그룹의 일부 데이터가 제외됨에 따라 다른 그룹에 비해 학습 데이터 수가 부족하여 클래스 불균형이 발생했기 때문으로 분석된다. 또한 볼링 투구 동작에서는 노년층의 신체적 가동 범위 편차로 인해 모델이 일관된 패턴을 학습하는 데 한계가 있던 것으로 해석된다.

볼링 투구 동작의 노년 그룹이 비록 성능 저하가 있었지만 골프 스윙 동작에서는 모든 그룹에서 일관된 성능 향상을 보였으며, 볼링 투구 동작 또한 전체 모델의 F1-score는 87.8%에서 89.6%로 약 1.8% 향상되었다. 이를 통해 두 실험 모두 대다수의 그룹에서 인구통계학적 속성정보를 반영한 모델이 긍정적인 영향을 끼친 것으로 확인되었다.

이처럼 골프와 볼링 두 가지 동작에 대한 전체 모델 성과 대부분의 사용자 그룹별 세부 평가 결과에서 인구통계학적 속성정보를 반영한 Demog 모델이 반영하지 않은 No-Demog 모델에 비해 일관된 성능 향상을 확인할 수 있었다. 이는 성별과 연령대 같은 속성정보가 HAR 모델이 사용자 그룹별 특화된 동작 패턴을 학습하는 데 실질적으로 기여하는 요소임을 시사한다. 비록 볼링 투구 동작의 노년 그룹에서는 예외적인 성능 하락이 관찰되었으나, 이는 데이터 불균형 및 해당 연령대의 신체적 가동 범위 편차라는 특수성에 기인한 것으로 판단된다. 이러한 성능 향상의 원인과 모델 내부의 특징 분포 변화를 확인하기 위해 t-SNE 시각화 분석을 진행하였다.

4.4. t-SNE 기반 시각화 분석

사용자의 인구통계학적 속성정보 포함 여부에 따라 모델이 학습한 내부 특징 표현 차이를 분석하기 위해 골프 스윙 동작에서 가장 높은 성능을 보인 Transformer 모델을 기반으로 t-SNE(t-Distributed Stochastic Neighbor Embedding) 시각화 분석을 진행하였다. t-SNE는 고차원 특징 공간의 데이터를 저차원으로 축소하여 데이터 간 분포 관계를 시각적으로 표현하는 차원 축소 기법이다.

이를 위해 인구통계학적 속성정보를 반영한 모델과 인구통계학적 속성정보를 반영하지 않은 모델의 FC layer 직전 내부 특징 벡터를 대상으로 PCA를 통해 30차원으로 축소한 후, t-SNE를 적용하여 2차원으로 시

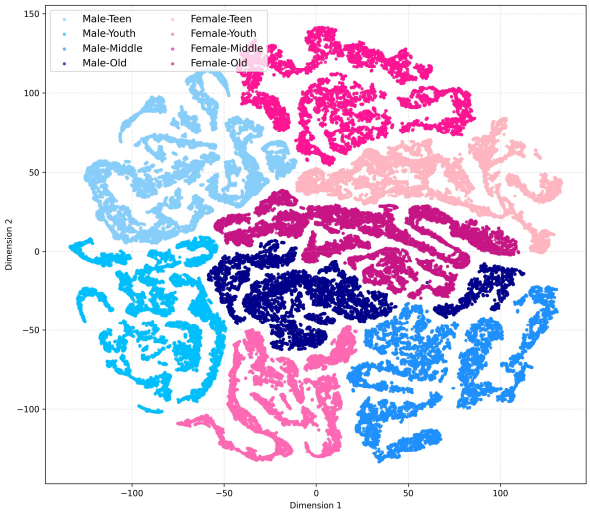

그림 6. 인구통계학적 속성정보를 반영한 모델의 성별 및 연령대별 특징 공간 시각화

Fig. 6. Feature space visualization of the Demog model by gender and age group.

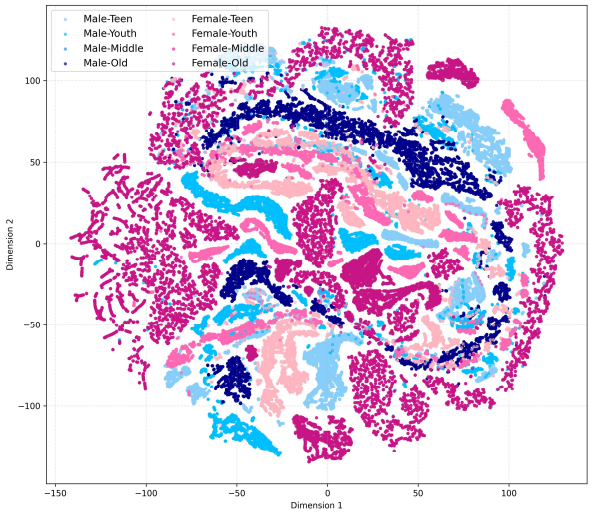

그림 7. 인구통계학적 속성정보를 반영하지 않은 모델의 성별 및 연령대별 특징 공간 시각화

Fig. 7. Feature space visualization of the No-Demog model by gender and age group.

각화하였다. 그림 6은 인구통계학적 속성정보를 반영한 경우의 시각화 결과이며, 그림 7은 인구통계학적 속성정보를 반영하지 않았을 때 시각화 결과를 나타낸다.

시각화 결과를 살펴보면 인구통계학적 속성정보를 반영하지 않았을 경우 각 그룹 간 경계가 불명확하고 서로 다른 그룹의 데이터가 혼재되어 분포하는 모습을 보였다. 반면 인구통계학적 속성정보를 반영한 경우 각 그룹 간 경계가 비교적 명확하게 분리되며, 각 그룹 내 데이터의 응집도 또한 크게 향상된 것을 확인하였다.

이러한 특징 공간의 구조적 변화는 위치, 회전, 속도 및 가속도 데이터가 인구통계학적 속성정보와 결합되어 모델이 각 그룹별 동작 패턴의 미세한 차이를 변별력 있게 학습한 것으로 해석된다. 결과적으로 네 가지 서로 다른 딥러닝 모델에서 공통으로 나타난 일관된 성능 향상과 t-SNE 시각화 분석을 통해 확인된 특징 분포의 차이는 본 연구에서 제안한 방식이 HAR 모델의 인식 성능 향상에 유효한 영향을 미치고 있음을 확인할 수 있었다.

V. 결 론

본 연구에서는 XR 기반 메타버스 환경에서 성별 및 연령대 두 가지 인구통계학적 속성정보를 활용하여 HAR 모델의 성능 향상 가능성을 평가하였다. 이를 위해 LSTM, GRU, 1D-CNN, Transformer 네 가지 모델을 기반으로 골프 스윙 및 볼링 투구 동작 데이터를 학습하고 인구통계학적 속성정보 반영 여부에 따른 모델 간 성능을 비교하였다.

실험 결과 네 가지 모델에서 인구통계학적 속성정보를 반영했을 때 성능이 일관되게 향상되는 것을 확인하였다. 골프 스윙 동작의 경우 F1-score 기준 LSTM 모델은 약 4.3%, GRU 모델은 약 1.7%, 1D-CNN 모델은 약 2.2%, Transformer 모델은 약 3.8% 성능이 향상되었다. 볼링 투구 동작에서도 LSTM 모델은 1.1%, GRU 모델은 1.2%, 1D-CNN 모델은 1.9%, Transformer 모델은 1.8%의 성능 향상을 확인할 수 있었다. 이중 Transformer 모델이 가장 높은 동작 인식 성능을 달성하였다.

가장 높은 동작 인식 성능을 보인 Transformer 모델을 기반으로 성별 및 연령대 그룹별 성능 평가 결과 F1-score 기준 인구통계학적 속성정보를 포함했을 때 중년 그룹은 골프 스윙 동작에서 95.5%, 볼링 투구 동작에서 96.5% 성능을 기록하며 높은 성능을 보였다. 하지만 노년 그룹은 각각 88.6%, 83.5%로 상대적으로 낮은 성능을 기록하였다. 또한 볼링 투구 동작에서 노년 그룹은 인구통계학적 속성정보를 포함했을 때가 포함하지 않았을 때보다 성능이 하락하였는데 이는 데이터 불균형과 신체적 제약에 따른 동작의 가변성이 성능 향상을 저해한 원인으로 분석된다. 그러나 이를 제외한 나머지 그룹에서는 인구통계학적 속성정보를 반영했을 때 성능이 일관되게 향상되었으며, 이는 성별과 연령대 같은 속성정보가 모델의 인식 성능 개선에 실질적으로

기여하는 요소임을 의미한다.

t-SNE 시각화 분석에서도 인구통계학적 속성정보를 반영한 Transformer 모델의 내부 특징 공간이 성별 및 연령대별로 명확히 구분되는 것을 확인하였다. 이러한 특징 공간의 구조적 변화는 위치, 회전, 속도 및 가속도 데이터가 인구통계학적 속성정보와 결합되어 모델이 각 그룹별 동작 패턴의 미세한 차이를 효과적으로 학습하는 데 도움을 준 것으로 해석된다.

이를 통해 본 연구에서는 사용자 인구통계학적 속성정보를 모델에 통합함으로써 사용자 간 행동 인식 성능 차이를 완화하고, 보다 포용적이고 사용자 친화적인 메타버스 환경 구현에 기여할 수 있음을 확인하였다. 향후 연구에서는 키, 체중 등 추가적인 인구통계학적 속성정보를 반영하고 다양한 동작 유형으로 일반화 실험을 수행할 예정이다.

REFERENCES

[1] K. Plupattanakit, P. Suntichaikul, P. Taveekitworachai, R. Thawonmas, J. White, K. Sookhanaphibarn, and W. Choensawat, “LLMs in Eduverse: LLM-integrated English educational game in metaverse,” in Proc. IEEE 13th Global Conf. on Consumer Electronics, pp. 257–258, Nara, Japan, Oct. 2024.

[2] S. Agrawal, R. Singh, A. Singh, and S. Kapoor, “A study of entrepreneurial opportunities in metaverse: Navigating the virtual frontier,” in Big Data in Finance: Transforming the Financial Landscape, vol. 1, pp. 659–669, 2025.

[3] S. Deng, J. Chen, D. Teng, C. Yang, D. Chen, T. Jia, and H. Wang, “LHAR: Lightweight human activity recognition on knowledge distillation,” IEEE J. Biomed. Health Inform., vol. 27, no. 12, pp. 6192–6203, Dec. 2023.

[4] J. H. Park et al., “Implementation of metaverse-based interactive art exhibition platform utilizing artificial intelligence,” in Proc. IEIE Conf., pp. 4756–4759, 2025.

[5] Y. Yin, L. Xie, Z. Jiang, F. Xiao, J. Cao, and S. Lu, “A systematic review of human activity recognition based on mobile devices: Overview, progress and trends,” IEEE Commun. Surveys Tuts., vol. 26, no. 2, pp. 890–929, Apr. 2024.

[6] Y. H. Jeon, B. H. Heo, D. U. Jo, J. I. Lim, and J. Y. Choi, “Real-time action recognition algorithm based on pose estimation,” in Proc. IEIE Conf., pp. 659–661, 2018.

[7] M. H. Arshad, M. Bilal, and A. Gani, “Human activity recognition: Review, taxonomy and open challenges,” Sensors, vol. 22, no. 17, p. 6463, Sept. 2022.

[8] S. P. Sahoo, S. Ari, K. Mahapatra, and S. P. Mohanty, “HAR-Depth: A novel framework for human action recognition using sequential learning and depth estimated history images,” IEEE Trans. Emerg. Top. Comput. Intell., vol. 5, no. 5, pp. 813–825, Oct. 2021.

[9] M. B. Shaikh and D. Chai, “RGB-D data-based action recognition: A review,” Sensors, vol. 21, no. 12, p. 4246, June 2021.

[10] M. Straczkiewicz, P. James, and J.-P. Onnela, “A systematic review of smartphone-based human activity recognition methods for health research,” NPJ Digit. Med., vol. 4, no. 1, p. 148, Oct. 2021.

[11] J. Pan, L. Zhang, Z. Hu, X. Zhang, and X. Cai, “Generalizing sensor-based human activity recognition with domain specific ensemble learning,” IEEE Sensors J., early access, 2025.

[12] J.-L. Chung, L.-Y. Ong, and M.-C. Leow, “A systematic literature review of optimization methods in skeleton-based human action recognition,” IEEE Access, early access, 2025.

[13] V.-H. Hoang, J. W. Lee, M. J. Piran, and C.-S. Park, “Advances in skeleton-based fall detection in RGB videos: From handcrafted to deep learning approaches,” IEEE Access, vol. 11, pp. 92322–92352, Aug. 2023.

[14] P. Zolfaghari, V. F. Rey, L. Ray, H. Kim, S. Suh, and P. Lukowicz, “Sensor data augmentation from skeleton pose sequences for improving human activity recognition,” in Proc. 2024 IEEE Int. Conf. on Activity and Behavior Computing, pp. 1–8, Bangkok, Thailand, Aug. 2024.

[15] N. Bansal, A. Bansal, and M. Gupta, “Analyzing the variability of RNN hyperparameters and architectures for HAR with wearable sensor data,” in Proc. 2024 IEEE 3rd Int. Conf. on Power Electronics and IoT Applications in Renewable Energy and Its Control, pp. 232–237, Mathura, India,

Feb. 2024.
[16] J. Zhao, M. Shao, Y. Wang, and R. Xu, “Real-time recognition of in-place body actions and head gestures using only a head-mounted display,” in Proc. 2023 IEEE Conf. on Virtual Reality and 3D User Interfaces, pp. 105–114, Shanghai, China, Mar. 2023.
[17] Q. Chen, D. Lu, W. Feng, C. Ye, L. Qiu, J. Pan, and J. Li, “HARFormer: A masked self-supervised Transformer-base model for human activity recognition with predicting somatosensory tokens,” IEEE Trans. Instrum. Meas., early access, 2025.
[18] T. Zhaksylyk, K. Taniyev, and N. A. Tu, “Spatial-temporal Transformer with contrastive learning for skeleton-based human action recognition,” in Proc. 2024 Int. Conf. on Advanced Technologies for Communications, pp. 637–642, Hanoi, Vietnam, Oct. 2024.

저 자 소 개

한 석 호(비회원)
2023년 2월 원광대학교 디지털콘텐츠공학과 학사
2026년 2월 충남대학교 컴퓨터공학과 석사
2024년 9월~현재 한국전자통신연구원 콘텐츠연구본부 위촉연구원
<주관심분야 : 컴퓨터 비전, 인공지능, 행동인식, 이상탐지>

김 종 성(비회원)
2000년 2월 고려대학교 전파공학 학사
2002년 2월 포항공과대학교 전자전기공학 석사
2008년 2월 포항공과대학교 전자전기공학 박사
2008년 3월~현재 한국전자통신연구원 콘텐츠연구본부 책임연구원
<주관심분야 : 컴퓨터비전, 영상처리, 혼합현실, 메타버스, 인공지능, 콘텐츠>