

(11) 공개번호 10-2024-0152157
(43) 공개일자 2024년10월21일

(51) 국제특허분류(Int. Cl.)
G06V 40/20 (2022.01) *G06V 10/34* (2022.01)
G06V 10/70 (2022.01) *G06V 10/80* (2022.01)
G06V 20/40 (2022.01) *G06V 20/70* (2022.01)

(52) CPC특허분류
G06V 40/20 (2022.01)
G06V 10/34 (2023.08)

(21) 출원번호 10-2023-0048469

(22) 출원일자 2023년04월12일
 심사청구일자 2023년04월12일

(71) 출원인
 계명대학교 산학협력단
 대구광역시 달서구 달구벌대로 1095, 계명대학교
 산학협력관 201호(신당동)

(72) 발명자
 고병철
 대구광역시 수성구 범어천로 107
 안다솜
 대구광역시 북구 침산로21길 39

(74) 대리인
 김건우

전체 청구항 수 : 총 16 항

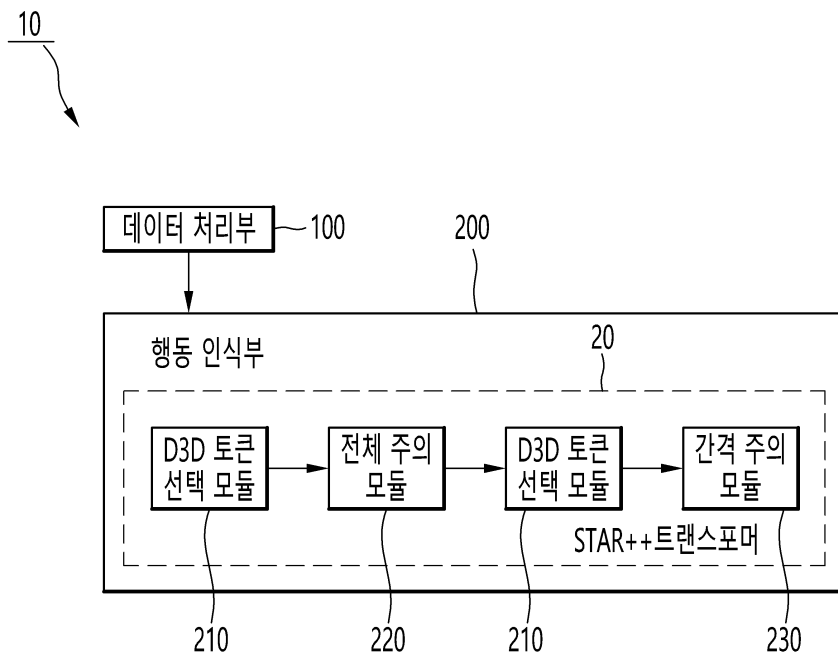
(54) 발명의 명칭 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템

(57) 요약

본 발명은 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템에 관한 것으로서, 보다 구체적으로는 입력 비디오 및 골격(skeleton) 데이터를 사용해 인간 행동을 인식하는 행동 인식 시스템으로서, 상기 입력 비디오와 골격 데이터로부터 입력 토큰을 구성하는 데이터 처리부; 및 상기 입력 토큰을 개선된 시공간 교차

(뒷면에 계속)

대표도 - 도3



주의(improved Spatio-TemproAl cRoss attention, STAR++) 트랜스포머에 입력해 최종 행동 레이블을 추정하는 행동 인식부를 포함하여 구성되며, 상기 STAR++ 트랜스포머는, 주의 메커니즘을 수행하되, 상기 입력 토큰을 시간 차원을 기준으로 미리 설정된 간격 항(interval term)에 따라 나누어 두 토큰 그룹을 생성하고, 두 토큰 그룹 간의 교차 주의 메커니즘을 수행해 행동 변화를 학습하는 간격 주의(Interval Attention, IAttn) 모듈을 포함하는 복수의 인코더 레이어로 구성되는 것을 그 구성상의 특징으로 한다.

또한, 본 발명은 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 방법에 관한 것으로서, 보다 구체적으로는 컴퓨터에서 각 단계가 수행되는, 입력 비디오 및 골격(skeleton) 데이터를 사용해 인간 행동을 인식하는 행동 인식 방법으로서, (1) 상기 입력 비디오와 골격 데이터로부터 입력 토큰을 구성하는 단계; 및 (2) 상기 입력 토큰을 개선된 시공간 교차 주의(improved Spatio-TemproAl cRoss attention, STAR++) 트랜스포머에 입력해 최종 행동 레이블을 추정하는 단계를 포함하여 구성되며, 상기 STAR++ 트랜스포머는, 주의 메커니즘을 수행하되, 상기 입력 토큰을 시간 차원을 기준으로 미리 설정된 간격 항(interval term)에 따라 나누어 두 토큰 그룹을 생성하고, 두 토큰 그룹 간의 교차 주의 메커니즘을 수행해 행동 변화를 학습하는 간격 주의(Interval Attention, IAttn) 모듈을 포함하는 복수의 인코더 레이어로 구성되는 것을 그 구성상의 특징으로 한다.

본 발명에서 제안하고 있는 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템 및 방법에 따르면, 비디오 프레임과 사람의 골격 특징을 함께 학습해 행동 인식을 수행하되, 다중 표현 학습을 통해 영상 및 골격 기반 데이터를 유연하게 결합하고 다중 클래스 토큰을 적용해 이중 도메인의 데이터를 효과적으로 학습할 수 있으며, 개선된 시공간 교차 주의가 적용된 트랜스포머(STAR++ 트랜스포머)를 통해 시공간 특징을 함께 학습할 수 있어서 행동 인식 성능을 높일 수 있다.

또한, 본 발명에서 제안하고 있는 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템 및 방법에 따르면, 시간 차원을 기준으로 미리 설정된 간격 항(interval term)에 따라 나누어 두 토큰 그룹을 생성하고, 두 토큰 그룹 간의 교차 주의 메커니즘을 수행해 행동 변화를 학습하는 간격 주의 모듈을 포함하는 복수의 인코더 레이어로 STAR++ 트랜스포머를 구성하여, 깊이에 따라 시간 축의 크고 작은 특징 사이의 차이를 점진적으로 학습할 수 있고, 인코더-디코더 구조가 아닌 단일 인코더 형태로 통합하여 네트워크를 단순화해 계산 비용을 줄일 수 있다.

(52) CPC특허분류

G06V 10/70 (2023.08)

G06V 10/803 (2023.08)

G06V 20/40 (2023.08)

G06V 20/46 (2022.01)

G06V 20/70 (2023.08)

이 발명을 지원한 국가연구개발사업

과제고유번호	1345363225
과제번호	2022R1I1A3058128
부처명	교육부
과제관리(전문)기관명	한국연구재단
연구사업명	지역대학우수과학자지원사업
연구과제명	자기-지도학습 기반의 경량 시공간 그래프 트랜스포머 연구
기 여 율	1/1
과제수행기관명	계명대학교
연구기간	2023.03.01 ~ 2024.02.29

명세서

청구범위

청구항 1

입력 비디오 및 골격(skeleton) 데이터를 사용해 인간 행동을 인식하는 행동 인식 시스템으로서,
 상기 입력 비디오와 골격 데이터로부터 입력 토큰을 구성하는 데이터 처리부(100); 및
 상기 입력 토큰을 개선된 시공간 교차 주의(improved Spatio-Temporal cRoss attention, STAR++) 트랜스포머에
 입력해 최종 행동 레이블을 추정하는 행동 인식부(200)를 포함하여 구성되며,
 상기 STAR++ 트랜스포머(20)는,
 주의 메커니즘을 수행하되, 상기 입력 토큰을 시간 차원을 기준으로 미리 설정된 간격 항(interval term)에 따라 나누어 두 토큰 그룹을 생성하고, 두 토큰 그룹 간의 교차 주의 메커니즘을 수행해 행동 변화를 학습하는 간격 주의(Interval Attention, IAttn) 모듈을 포함하는 복수의 인코더 레이어로 구성되는 것을 특징으로 하는, 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템(10).

청구항 2

제1항에 있어서, 상기 간격 항은,
 상기 인코더 레이어마다 미리 설정되며, 레이어가 깊어짐에 따라 값이 커지도록 설정된 것을 특징으로 하는, 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템(10).

청구항 3

제1항에 있어서, 상기 데이터 처리부(100)는,
 상기 입력 비디오 프레임에서 전역(global) 특징 및 로컬(local) 특징을 추출하고, 상기 전역 특징을 토큰화한 전역 그리드 토큰과 상기 로컬 특징과 골격 데이터를 결합해 토큰화한 관절 맵 토큰을 결합하되, 다중 클래스 토큰을 포함해 결합하여 입력 토큰을 구성하는 것을 특징으로 하는, 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템(10).

청구항 4

제3항에 있어서, 상기 데이터 처리부(100)는,
 상기 골격 데이터의 관절 좌표를 사용해 생성한 관절 열 지도(joint heatmap)를 상기 로컬 특징과 결합해 토큰화하여 상기 관절 맵 토큰을 생성하는 것을 특징으로 하는, 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템(10).

청구항 5

제3항에 있어서, 상기 데이터 처리부(100)는,
 상기 전역 그리드 토큰과 관절 맵 토큰을 결합하되, 상기 다중 클래스 토큰을 전체 토큰의 양 끝과 상기 전역 그리드 토큰과 관절 맵 토큰의 사이에 연결해 상기 입력 토큰을 구성하는 것을 특징으로 하는, 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템(10).

청구항 6

제1항에 있어서, 상기 인코더 레이어는,

상기 입력 토큰을 구성하는 전체 토큰에 대해 전체 시공간 주의(Full Spatio-temporal Attention, FAttn) 메커니즘을 수행하는 전체 주의 모듈(220)과; 상기 전체 주의 모듈(220)의 출력을 시간 차원을 기준으로 미리 설정된 간격 항에 따라 나누어 두 토큰 그룹을 생성하고, 두 토큰 그룹 간의 교차 주의 메커니즘을 수행하는 상기 간격 주의 모듈(230)을 포함하여 구성되는 것을 특징으로 하는, 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템(10).

청구항 7

제6항에 있어서, 상기 STAR++ 트랜스포머(20)는,

입력된 토큰에서 상기 전역 그리드 토큰을 추출해 오프셋과 기준점을 결합하여 최종 좌표를 생성하며, 상기 최종 좌표로부터 키(key) 및 값(value)을 생성하는 D3D 토큰 선택 모듈(210)을 더 포함하는 것을 특징으로 하는, 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템(10).

청구항 8

제7항에 있어서, 상기 인코더 레이어는,

- (a) 상기 D3D 토큰 선택 모듈(210)이 상기 입력 토큰에서 키 및 값을 생성하는 단계;
- (b) 상기 전체 주의 모듈(220)이 상기 D3D 토큰 선택 모듈(210)에서 생성된 키 및 값을 이용해 전체 시공간 주의 메커니즘을 수행하는 단계;
- (c) 상기 전체 주의 모듈(220)의 출력을 시간 차원을 기준으로 미리 설정된 간격 항에 따라 나누어 두 토큰 그룹을 생성하는 단계;
- (d) 상기 D3D 토큰 선택 모듈(210)이 생성된 두 토큰 그룹에서 키 및 값을 생성하는 단계; 및
- (e) 상기 간격 주의 모듈(230)이, 상기 단계 (d)에서 생성한 키 및 값을 이용해, 상기 단계 (c)에서 생성한 두 토큰 그룹 간의 교차 주의 메커니즘을 수행하는 단계를 수행하는 것을 특징으로 하는, 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템(10).

청구항 9

컴퓨터에서 각 단계가 수행되는, 입력 비디오 및 골격(skeleton) 데이터를 사용해 인간 행동을 인식하는 행동 인식 방법으로서,

- (1) 상기 입력 비디오와 골격 데이터로부터 입력 토큰을 구성하는 단계; 및
- (2) 상기 입력 토큰을 개선된 시공간 교차 주의(improved Spatio-TemproAl cRoss attention, STAR++) 트랜스포머에 입력해 최종 행동 레이블을 추정하는 단계를 포함하여 구성되며,

상기 STAR++ 트랜스포머(20)는,

주의 메커니즘을 수행하되, 상기 입력 토큰을 시간 차원을 기준으로 미리 설정된 간격 항(interval term)에 따라 나누어 두 토큰 그룹을 생성하고, 두 토큰 그룹 간의 교차 주의 메커니즘을 수행해 행동 변화를 학습하는 간격 주의(Interval Attention, IAttn) 모듈을 포함하는 복수의 인코더 레이어로 구성되는 것을 특징으로 하는, 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 방법.

청구항 10

제9항에 있어서, 상기 간격 항은,

상기 인코더 레이어마다 미리 설정되되, 레이어가 깊어짐에 따라 값이 커지도록 설정된 것을 특징으로 하는, 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 방법.

청구항 11

제9항에 있어서, 상기 단계 (1)에서는,

상기 입력 비디오 프레임에서 전역(global) 특징 및 로컬(local) 특징을 추출하고, 상기 전역 특징을 토큰화한 전역 그리드 토큰과 상기 로컬 특징과 골격 데이터를 결합해 토큰화한 관절 맵 토큰을 결합하되, 다중 클래스 토큰을 포함해 결합하여 입력 토큰을 구성하는 것을 특징으로 하는, 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 방법.

청구항 12

제11항에 있어서, 상기 단계 (1)에서는,

상기 골격 데이터의 관절 좌표를 사용해 생성한 관절 열 지도(joint heatmap)를 상기 로컬 특징과 결합해 토큰화하여 상기 관절 맵 토큰을 생성하는 것을 특징으로 하는, 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 방법.

청구항 13

제11항에 있어서, 상기 단계 (1)에서는,

상기 전역 그리드 토큰과 관절 맵 토큰을 결합하되, 상기 다중 클래스 토큰을 전체 토큰의 양 끝과 상기 전역 그리드 토큰과 관절 맵 토큰의 사이에 연결해 상기 입력 토큰을 구성하는 것을 특징으로 하는, 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 방법.

청구항 14

제9항에 있어서, 상기 인코더 레이어는,

상기 입력 토큰을 구성하는 전체 토큰에 대해 전체 시공간 주의(Full Spatio-temporal Attention, FAttn) 메커니즘을 수행하는 전체 주의 모듈(220)과; 상기 전체 주의 모듈(220)의 출력을 시간 차원을 기준으로 미리 설정된 간격 항에 따라 나누어 두 토큰 그룹을 생성하고, 두 토큰 그룹 간의 교차 주의 메커니즘을 수행하는 상기 간격 주의 모듈(230)을 포함하여 구성되는 것을 특징으로 하는, 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 방법.

청구항 15

제14항에 있어서, 상기 STAR++ 트랜스포머(20)는,

입력된 토큰에서 상기 전역 그리드 토큰을 추출해 오프셋과 기준점을 결합하여 최종 좌표를 생성하며, 상기 최종 좌표로부터 키(key) 및 값(value)을 생성하는 D3D 토큰 선택 모듈(210)을 더 포함하는 것을 특징으로 하는, 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 방법.

청구항 16

제15항에 있어서, 상기 인코더 레이어는,

- (a) 상기 D3D 토큰 선택 모듈(210)이 상기 입력 토큰에서 키 및 값을 생성하는 단계;
- (b) 상기 전체 주의 모듈(220)이 상기 D3D 토큰 선택 모듈(210)에서 생성된 키 및 값을 이용해 전체 시공간 주의 메커니즘을 수행하는 단계;
- (c) 상기 전체 주의 모듈(220)의 출력을 시간 차원을 기준으로 미리 설정된 간격 항에 따라 나누어 두 토큰 그룹을 생성하는 단계;
- (d) 상기 D3D 토큰 선택 모듈(210)이 생성된 두 토큰 그룹에서 키 및 값을 생성하는 단계; 및
- (e) 상기 간격 주의 모듈(230)이, 상기 단계 (d)에서 생성한 키 및 값을 이용해, 상기 단계 (c)에서 생성한 두 토큰 그룹 간의 교차 주의 메커니즘을 수행하는 단계를 수행하는 것을 특징으로 하는, 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 방법.

발명의 설명

기술 분야

[0001] 본 발명은 인간 행동 인식 시스템 및 방법에 관한 것으로서, 보다 구체적으로는 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템 및 방법에 관한 것이다.

배경 기술

[0002] 이 부분에 기술된 내용은 단순히 본 발명의 일실시예에 대한 배경 정보를 제공할 뿐 종래기술을 구성하는 것은 아니다.

[0004] 최근 컴퓨터 비전과 관련된 딥러닝 연구에서 다양한 데이터 형식의 조합을 모델링하는 멀티-모달(multi-modal) 학습이 주목받고 있다. 컴퓨터 비전과 관련된 많은 응용 분야 중 비디오 동작 인식은 멀티-모달 접근법을 사용하여 성능을 향상시킬 수 있는 예이다. 기존 비디오 동작 인식 방법은 일반적으로 비디오 시퀀스 또는 골격 특징과 같은 특징을 독립적으로 사용하지만, 최근에는 오디오, 광학 흐름, 모션 및 텍스처도 사용하는 멀티-모달 접근 방식이 연구되었다. 멀티-모달 학습을 사용하는 비디오 동작 인식의 기존 방법으로는 RGB 프레임과 광학 흐름을 모델링하기 위한 2가지 경향이 있다. 그러나 이 방법은 멀티-모달 데이터를 별도로 사용하여 결과를 도출한 다음 결합하기 때문에 동일한 정보에 대한 예측 결과를 표현하는 데 시간이 오래 걸리고 모호하다.

[0006] 시계열 처리를 위한 비전 트랜스포머(Vision Transformer, ViT) 기반 모델은 다중 헤드 자기-주의(self-attention) 메커니즘을 통해 우수한 성능을 보여주었지만, 상당한 계산 비용이 필요하다. 이러한 계산 문제를 해결하기 위해 최근 효율적인 트랜스포머 기반 비디오 인식 알고리즘이 연구되었다. 또한, 변형 가능한 토큰 선택 방법 또는 토큰을 점진적으로 줄이거나 삭제하는 방법을 적용하는 트랜스포머 기반 모델이 증가하고 있다. 이러한 모델은 변형 가능한 토큰 선택 및 토큰 중단 기술(token halting techniques)을 사용하여 토큰의 크기를 줄여 연산 비용의 대부분을 차지하는 자기-주의 연산을 효과적으로 줄이면서 성능을 유지하거나 개선한다.

[0008] 한편, 본 발명과 관련된 선행기술로서, 공개특허 제10-2022-0078893호(발명의 명칭: 영상 속 사람의 행동 인식 장치 및 방법, 공개일자: 2022년 06월 13일) 등이 개시된 바 있다.

[0010] 진술한 배경 기술은 발명자가 본 발명의 도출을 위해 보유하고 있었거나, 본 발명의 도출 과정에서 습득한 기술 정보로서, 반드시 본 발명의 출원 전에 일반 공중에게 공개된 공지 기술이라 할 수는 없다.

발명의 내용

해결하려는 과제

[0011] 본 발명은 기존에 제안된 방법들의 상기와 같은 문제점들을 해결하기 위해 제안된 것으로서, 비디오 프레임과

사람의 골격 특징을 함께 학습해 행동 인식을 수행하되, 다중 표현 학습을 통해 영상 및 골격 기반 데이터를 유연하게 결합하고 다중 클래스 토큰(Multi class token)을 적용해 이중 도메인의 데이터를 효과적으로 학습할 수 있으며, 개선된 시공간 교차 주의가 적용된 트랜스포머(STAR++ 트랜스포머)를 통해 시공간 특징을 함께 학습할 수 있어서 행동 인식 성능을 높일 수 있는, 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템 및 방법을 제공하는 것을 그 목적으로 한다.

[0013] 또한, 본 발명은, 시간 차원을 기준으로 미리 설정된 간격 항(interval term)에 따라 나누어 두 토큰 그룹을 생성하고, 두 토큰 그룹 간의 교차 주의 메커니즘을 수행해 행동 변화를 학습하는 간격 주의 모듈을 포함하는 복수의 인코더 레이어로 STAR++ 트랜스포머를 구성하여, 깊이에 따라 시간 축의 크고 작은 특징 사이의 차이를 점진적으로 학습할 수 있고, 인코더-디코더 구조가 아닌 단일 인코더 형태로 통합하여 네트워크를 단순화해 계산 비용을 줄일 수 있는, 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템 및 방법을 제공하는 것을 또 다른 목적으로 한다.

[0015] 다만, 본 발명이 이루고자 하는 기술적 과제는 상기한 바와 같은 기술적 과제로 한정되지 않으며, 또 다른 기술적 과제들이 존재할 수 있고, 명시적으로 언급하지 않더라도 과제의 해결수단이나 실시 형태로부터 파악될 수 있는 목적이나 효과도 이에 포함됨은 물론이다.

과제의 해결 수단

[0016] 상기한 목적을 달성하기 위한 본 발명의 특징에 따른 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템은,

[0017] 입력 비디오 및 골격(skeleton) 데이터를 사용해 인간 행동을 인식하는 행동 인식 시스템으로서,

[0018] 상기 입력 비디오와 골격 데이터로부터 입력 토큰을 구성하는 데이터 처리부; 및

[0019] 상기 입력 토큰을 개선된 시공간 교차 주의(improved Spatio-Temporal cross attention, STAR++) 트랜스포머에 입력해 최종 행동 레이블을 추정하는 행동 인식부를 포함하여 구성되며,

[0020] 상기 STAR++ 트랜스포머는,

[0021] 주의 메커니즘을 수행하되, 상기 입력 토큰을 시간 차원을 기준으로 미리 설정된 간격 항(interval term)에 따라 나누어 두 토큰 그룹을 생성하고, 두 토큰 그룹 간의 교차 주의 메커니즘을 수행해 행동 변화를 학습하는 간격 주의(Interval Attention, IAttn) 모듈을 포함하는 복수의 인코더 레이어로 구성되는 것을 그 구성상의 특징으로 한다.

[0023] 바람직하게는, 상기 간격 항은,

[0024] 상기 인코더 레이어마다 미리 설정되되, 레이어가 깊어짐에 따라 값이 커지도록 설정될 수 있다.

[0026] 바람직하게는, 상기 데이터 처리부는,

[0027] 상기 입력 비디오 프레임에서 전역(global) 특징 및 로컬(local) 특징을 추출하고, 상기 전역 특징을 토큰화한 전역 그리드 토큰과 상기 로컬 특징과 골격 데이터를 결합해 토큰화한 관절 맵 토큰을 결합하되, 다중 클래스 토큰을 포함해 결합하여 입력 토큰을 구성할 수 있다.

[0029] 더욱 바람직하게는, 상기 데이터 처리부는,

[0030] 상기 골격 데이터의 관절 좌표를 사용해 생성한 관절 열 지도(joint heatmap)를 상기 로컬 특징과 결합해 토큰화하여 상기 관절 맵 토큰을 생성할 수 있다.

- [0032] 더욱 바람직하게는, 상기 데이터 처리부는,
- [0033] 상기 전역 그리드 토큰과 관절 맵 토큰을 결합하되, 상기 다중 클래스 토큰을 전체 토큰의 양 끝과 상기 전역 그리드 토큰과 관절 맵 토큰의 사이에 연결해 상기 입력 토큰을 구성할 수 있다.
- [0035] 바람직하게는, 상기 인코더 레이어는,
- [0036] 상기 입력 토큰을 구성하는 전체 토큰에 대해 전체 시공간 주의(Full Spatio-temporal Attention, FAttn) 메커니즘을 수행하는 전체 주의 모듈과; 상기 전체 주의 모듈의 출력을 시간 차원을 기준으로 미리 설정된 간격 항에 따라 나누어 두 토큰 그룹을 생성하고, 두 토큰 그룹 간의 교차 주의 메커니즘을 수행하는 상기 간격 주의 모듈을 포함하여 구성될 수 있다.
- [0038] 더욱 바람직하게는, 상기 STAR++ 트랜스포머는,
- [0039] 입력된 토큰에서 상기 전역 그리드 토큰을 추출해 오프셋과 기준점을 결합하여 최종 좌표를 생성하며, 상기 최종 좌표로부터 키(key) 및 값(value)을 생성하는 D3D(Deformable 3D) 토큰 선택 모듈을 더 포함할 수 있다.
- [0041] 더더욱 바람직하게는, 상기 인코더 레이어는,
- [0042] (a) 상기 D3D 토큰 선택 모듈이 상기 입력 토큰에서 키 및 값을 생성하는 단계;
- [0043] (b) 상기 전체 주의 모듈이 상기 D3D 토큰 선택 모듈에서 생성된 키 및 값을 이용해 전체 시공간 주의 메커니즘을 수행하는 단계;
- [0044] (c) 상기 전체 주의 모듈의 출력을 시간 차원을 기준으로 미리 설정된 간격 항에 따라 나누어 두 토큰 그룹을 생성하는 단계;
- [0045] (d) 상기 D3D 토큰 선택 모듈이 생성된 두 토큰 그룹에서 키 및 값을 생성하는 단계; 및
- [0046] (e) 상기 간격 주의 모듈이, 상기 단계 (d)에서 생성한 키 및 값을 이용해, 상기 단계 (c)에서 생성한 두 토큰 그룹 간의 교차 주의 메커니즘을 수행하는 단계를 수행할 수 있다.
- [0048] 상기한 목적을 달성하기 위한 본 발명의 특징에 따른 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 방법은,
- [0049] 컴퓨터에서 각 단계가 수행되는, 입력 비디오 및 골격(skeleton) 데이터를 사용해 인간 행동을 인식하는 행동 인식 방법으로서,
- [0050] (1) 상기 입력 비디오와 골격 데이터로부터 입력 토큰을 구성하는 단계; 및
- [0051] (2) 상기 입력 토큰을 개선된 시공간 교차 주의(improved Spatio-Temporal cRoss attention, STAR++) 트랜스포머에 입력해 최종 행동 레이블을 추정하는 단계를 포함하여 구성되며,
- [0052] 상기 STAR++ 트랜스포머는,
- [0053] 주의 메커니즘을 수행하되, 상기 입력 토큰을 시간 차원을 기준으로 미리 설정된 간격 항(interval term)에 따라 나누어 두 토큰 그룹을 생성하고, 두 토큰 그룹 간의 교차 주의 메커니즘을 수행해 행동 변화를 학습하는 간격 주의(Interval Attention, IAttn) 모듈을 포함하는 복수의 인코더 레이어로 구성되는 것을 그 구성상의 특징으로 한다.
- [0055] 바람직하게는, 상기 간격 항은,
- [0056] 상기 인코더 레이어마다 미리 설정되되, 레이어가 깊어짐에 따라 값이 커지도록 설정될 수 있다.

- [0058] 바람직하게는, 상기 단계 (1)에서는,
- [0059] 상기 입력 비디오 프레임에서 전역(global) 특징 및 로컬(local) 특징을 추출하고, 상기 전역 특징을 토큰화한 전역 그리드 토큰과 상기 로컬 특징과 골격 데이터를 결합해 토큰화한 관절 맵 토큰을 결합하되, 다중 클래스 토큰을 포함해 결합하여 입력 토큰을 구성할 수 있다.
- [0061] 더욱 바람직하게는, 상기 단계 (1)에서는,
- [0062] 상기 골격 데이터의 관절 좌표를 사용해 생성한 관절 열 지도(joint heatmap)를 상기 로컬 특징과 결합해 토큰화하여 상기 관절 맵 토큰을 생성할 수 있다.
- [0064] 더욱 바람직하게는, 상기 단계 (1)에서는,
- [0065] 상기 전역 그리드 토큰과 관절 맵 토큰을 결합하되, 상기 다중 클래스 토큰을 전체 토큰의 양 끝과 상기 전역 그리드 토큰과 관절 맵 토큰의 사이에 연결해 상기 입력 토큰을 구성할 수 있다.
- [0067] 바람직하게는, 상기 인코더 레이어는,
- [0068] 상기 입력 토큰을 구성하는 전체 토큰에 대해 전체 시공간 주의(Full Spatio-temporal Attention, FAttn) 메커니즘을 수행하는 전체 주의 모듈과; 상기 전체 주의 모듈의 출력을 시간 차원을 기준으로 미리 설정된 간격 항에 따라 나누어 두 토큰 그룹을 생성하고, 두 토큰 그룹 간의 교차 주의 메커니즘을 수행하는 상기 간격 주의 모듈을 포함하여 구성될 수 있다.
- [0070] 더욱 바람직하게는, 상기 STAR++ 트랜스포머는,
- [0071] 입력된 토큰에서 상기 전역 그리드 토큰을 추출해 오프셋과 기준점을 결합하여 최종 좌표를 생성하며, 상기 최종 좌표로부터 키(key) 및 값(value)을 생성하는 D3D 토큰 선택 모듈을 더 포함할 수 있다.
- [0073] 더더욱 바람직하게는, 상기 인코더 레이어는,
- [0074] (a) 상기 D3D 토큰 선택 모듈이 상기 입력 토큰에서 키 및 값을 생성하는 단계;
- [0075] (b) 상기 전체 주의 모듈이 상기 D3D 토큰 선택 모듈에서 생성된 키 및 값을 이용해 전체 시공간 주의 메커니즘을 수행하는 단계;
- [0076] (c) 상기 전체 주의 모듈의 출력을 시간 차원을 기준으로 미리 설정된 간격 항에 따라 나누어 두 토큰 그룹을 생성하는 단계;
- [0077] (d) 상기 D3D 토큰 선택 모듈이 생성된 두 토큰 그룹에서 키 및 값을 생성하는 단계; 및
- [0078] (e) 상기 간격 주의 모듈이, 상기 단계 (d)에서 생성한 키 및 값을 이용해, 상기 단계 (c)에서 생성한 두 토큰 그룹 간의 교차 주의 메커니즘을 수행하는 단계를 수행할 수 있다.

발명의 효과

- [0079] 본 발명에서 제안하고 있는 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템 및 방법에 따르면, 비디오 프레임과 사람의 골격 특징을 함께 학습해 행동 인식을 수행하되, 다중 표현 학습을 통해 영상 및 골격 기반 데이터를 유연하게 결합하고 다중 클래스 토큰을 적용해 이중 도메인의 데이터를 효과적으로 학습할 수 있으며, 개선된 시공간 교차 주의가 적용된 트랜스포머(STAR++ 트랜스포머)를 통해 시공간 특징을 함께 학습할 수 있어서 행동 인식 성능을 높일 수 있다.
- [0081] 또한, 본 발명에서 제안하고 있는 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템 및 방

법에 따르면, 시간 차원을 기준으로 미리 설정된 간격 항(interval term)에 따라 나누어 두 토큰 그룹을 생성하고, 두 토큰 그룹 간의 교차 주의 메커니즘을 수행해 행동 변화를 학습하는 간격 주의 모듈을 포함하는 복수의 인코더 레이어로 STAR++ 트랜스포머를 구성하여, 깊이에 따라 시간 축의 크고 작은 특징 사이의 차이를 점진적으로 학습할 수 있고, 인코더-디코더 구조가 아닌 단일 인코더 형태로 통합하여 네트워크를 단순화해 계산 비용을 줄일 수 있다.

[0083] 더불어, 본 발명의 다양하면서도 유익한 장점과 효과는 상술한 내용에 한정되지 않으며, 본 발명의 구체적인 실시 형태를 설명하는 과정에서 보다 쉽게 이해될 수 있을 것이다.

도면의 간단한 설명

[0084] 도 1은 기존의 STAR-Transformer의 시공간 교차 주의를 나타낸 도면.

도 2는 본 발명의 일실시예에 따른 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템에서 제안하는 STAR++ 트랜스포머의 시공간 교차 주의를 나타낸 도면.

도 3은 본 발명의 일실시예에 따른 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템의 구성을 도시한 도면.

도 4는 본 발명의 일실시예에 따른 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템의 전체 구조를 설명하기 위해 도시한 도면.

도 5는 본 발명의 일실시예에 따른 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템에서, STAR++ 트랜스포머의 인코더 레이어 상세 구조를 도시한 도면.

도 6은 본 발명의 일실시예에 따른 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템에서, D3D 토큰 선택 모듈의 D3D 토큰 선택 구조를 도시한 도면.

도 7은 본 발명의 일실시예에 따른 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템에서, 간격 주의(Interval Attention, IAttn)를 설명하기 위해 도시한 도면.

도 8은 본 발명의 일실시예에 따른 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템에서, 간격 주의 모듈에서 수행되는 IAttn 프로세스의 의사 코드(pseudo code)를 도시한 도면.

도 9는 본 발명의 일실시예에 따른 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 방법의 흐름을 도시한 도면.

도 10은 본 발명의 일실시예에 따른 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템 및 방법을 이용한 NTU-RGB+D 60 및 Penn 행동 데이터 세트의 시각화 결과를 도시한 도면.

도 11은 본 발명의 일실시예에 따른 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템 및 방법에서, 골격 정보를 나타내는 관절 맵 토큰에 대한 주의 점수를 시각화하여 도시한 도면.

발명을 실시하기 위한 구체적인 내용

[0085] 아래에서는 첨부한 도면을 참조하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 본 발명의 실시예를 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며, 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

[0087] 명세서 전체에서, 어떤 부분이 다른 부분과 "연결"되어 있다고 할 때, 이는 "직접적으로 연결"되어 있는 경우뿐 아니라, 그 중간에 다른 소자를 사이에 두고 "간접적으로 연결"되어 있는 경우도 포함한다. 또한 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미하며, 하나 또는 그 이상의 다른 특징이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해

되어야 한다.

- [0089] 이하의 실시예는 본 발명의 이해를 돕기 위한 상세한 설명이며, 본 발명의 권리 범위를 제한하는 것이 아니다. 따라서 본 발명과 동일한 기능을 수행하는 동일 범위의 발명 역시 본 발명의 권리 범위에 속할 것이다.
- [0091] 또한, 본 발명의 각 실시예에 포함된 각 구성, 과정, 공정 또는 방법 등은 기술적으로 상호간 모순되지 않는 범위 내에서 공유될 수 있다.
- [0093] 또한, 본 발명에 있어서 단말, 장치 또는 디바이스가 수행하는 것으로 기술된 동작이나 기능 중 일부는 해당 단말, 장치 또는 디바이스와 연결된 서버에서 대신 수행될 수 있다. 마찬가지로, 서버가 수행하는 것으로 기술된 동작이나 기능 중 일부도 해당 서버와 연결된 단말, 장치 또는 디바이스에서 수행될 수도 있다.
- [0095] 특히, 본 발명의 각 실시예에 따른 시스템을 실행시키기 위한 수단으로는 애플리케이션(Application), 또는 웹 서버일 수 있으며, 이 애플리케이션, 또는 웹 서버를 기록한 기록매체를 읽을 수 있는 수단인 단말로는, 일반적인 데스크톱이나 노트북 등의 일반 PC뿐만 아니라, 스마트 폰, 태블릿 PC, 등의 모바일 단말기를 포함할 수 있다.
- [0097] 이하, 첨부된 도면을 참고하여 본 발명의 실시예들을 상세히 설명하도록 한다.
- [0099] 본 발명은 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템 및 방법에 관한 것으로서, 본 발명의 특징에 따른 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템 및 방법은, 메모리 및 프로세서를 포함한 하드웨어에서 기록되는 소프트웨어로 구성될 수 있다. 예를 들어, 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템 및 방법은, 개인용 컴퓨터, 노트북 컴퓨터, 서버 컴퓨터, PDA, 스마트폰, 태블릿 PC 등에 저장 및 구현될 수 있다. 이하에서는 설명의 편의를 위해, 각 단계를 수행하는 주체는 생략될 수 있다.
- [0101] 도 1은 기존의 STAR-Transformer의 시공간 교차 주의를 나타낸 도면이고, 도 2는 본 발명의 일실시예에 따른 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템에서 제안하는 개선된 시공간 교차 주의(improved Spatio-Temporal Cross attention, STAR++) 트랜스포머의 시공간 교차 주의를 나타낸 도면이다.
- [0103] 특히, 멀티-모달 동작 인식 모델에 사용되는 STAR-Transformer (D. Ahn, S. Kim, H. Hong and B. C. Ko, “STAR-Transformer: A Spatio-temporal Cross Attention Transformer for Human Action Recognition,” in Proc. WACV, Waikoloa, HI, USA, 2023, pp. 3330-3339.)는 다중 클래스 토큰이 있는 멀티-모달 학습 메커니즘을 사용하는 기존 접근 방식인 2-스트림 방법과 달리, 여러 특징을 하나의 입력 벡터로 유연하게 결합하고 계산 수를 효율적으로 줄이고 동작의 시간적 차이를 학습할 수 있는 시공간 교차 주의 메커니즘을 제안한다. 그러나 STAR-Transformer의 시공간 교차 주의를 포함한 인코더-디코더 구조는 여전히 비효율적이며 개선이 필요하다. 따라서 STAR-Transformer의 효율성을 더욱 높이고 성능을 향상시키기 위해 변형 가능한 3D(Deformable 3D, D3D) 토큰 선택을 포함하여 개선한 STAR++ 구조를 제안한다.
- [0105] 도 1에 도시된 바와 같이, STAR-Transformer는 전체 입력 시퀀스를 지그재그 패턴으로 분할하는 지그재그 인코더(zigzag encoder)와 이를 다시 하프 프레임으로 분할하는 이진 디코더(binary decoder)로 구성된다. 반면에, 도 2에 도시된 바와 같이 STAR++ 트랜스포머는 간격 항(interval term)을 사용하여 레이어 깊이당 디커플링 프레임의 수를 조정하는 하나의 간격 주의(interval attention) 인코더로 구성될 수 있다. 여기서, 간격 항은 키

(key) 및 값(value)으로 필요한 토큰만 선택하여 주의를 생성할 수 있다.

[0107] 보다 구체적으로, 제안된 STAR++은 기본 STAR-Transformer 방법과 유사하게 모든 토큰을 시간 축을 중심으로 하는 두 개의 토큰 세트로 분리하여 시공간 교차 주의 메커니즘을 적용한다. 그러나 도 1에 도시된 바와 같이 STAR-Transformer는 인코더와 디코더 구조를 가지므로 레이어가 깊어질수록 더 많은 매개변수와 연산이 필요하다. 그러나 도 2에 도시된 바와 같이 STAR++는 간격 항을 사용하여 레이어 깊이당 디커플링 프레임의 수를 조정하는 인코더 하나로 구성되며, 필요한 토큰만 키와 값으로 선택하여 주의를 적용하므로 연산 수를 크게 줄일 수 있다. 또한 STAR++는 시간 축을 따라 로컬 특징 차이에서 전역 특징 차이를 점진적으로 학습할 수 있다. 이는 시공간 교차 주의로 토큰을 만드는 새로운 간격 주의(Interval Attention, IAttn)를 통해 가능하다.

[0109] 본 발명의 기여는 다음과 같다:

- [0110] - 기본 STAR-Transformer의 교차-모달 학습 메커니즘을 개선하여 멀티-모달 학습을 유연하게 결합할 수 있는 새로운 STAR++ 동작 인식 알고리즘을 제안한다.
- [0111] - 제안된 STAR++에서는 인코더 구조만 사용하여, 기본 STAR-트랜스포머의 인코더-디코더 구조를 개선한다.
- [0112] - 시공간 교차 주의를 기반으로 한 IAttn을 제안하여, 로컬 및 전역 특징을 점진적으로 학습하여 인식 성능을 향상시킨다.
- [0113] - STAR++는 D3D 토큰 선택을 제안하여 학습 결과에 영향을 미치는 토큰의 좌표를 효율적으로 학습하는 동적 토큰 선택을 가능하게 한다.

[0115] 이하에서는, 먼저 기본 모델인 STAR-Transformer에 대해 설명한다. 토큰화를 위해 ResNet Mixed Convolution 18 (MC18) (D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in Proc. CVPR, Salt Lake City, UT, USA, 2018, pp. 6450-6459.)을 사용하여 입력 비디오 프레임에서 전역 및 로컬 특징을 추출할 수 있다. 전역 특징은 비디오 프레임의 전역 정보를 포함하기 때문에 RGB 특징을 나타내는 전역 그리드 토큰 $\mathbf{T}_g \in \mathbb{R}^{T \times P \times C}$ 로 생성된다. 여기서, T는 비디오 프레임의 수를 나타내고, P는 전역 특징을 중복되지 않는 크기로 분할하여 토큰화한 전역 그리드 토큰의 수를 나타내며, C는 차원을 나타낸다. 로컬 특징에는 로컬 정보가 포함되어 있으므로 골격(skeleton) 좌표를 사용하여 생성된 관절 열 지도(heatmap)와 결합되고 토큰화되어 $\mathbf{T}_j \in \mathbb{R}^{T \times N \times C}$ 가 생성된다. 여기서, N은 관절의 수이다. 관절 맵(joint map) 토큰을 만드는 공식은 다음 수학적 식 1과 같다.

수학적 식 1

$$[0117] \quad \mathbf{T}_{j,n} = \bigoplus_{c'=1}^{C'} (\sum_t^T \sum_{h'}^{H'} \sum_{w'}^{W'} \mathbf{F}_{c'}^t(h', w') \times \mathbf{H}_n^t(h', w'))$$

[0119] 여기서, $\mathbf{F} \in \mathbb{R}^{T \times h' \times w' \times c'}$ 는 로컬 특징맵을 나타내며, $\mathbf{H} \in \mathbb{R}^{T \times N \times h' \times w' \times c'}$ 는 골격 좌표로 만들어진 관절 열 지도이다.

[0121] 토큰화된 두 개의 멀티-모달 토큰은 다중 클래스 토큰과 결합되어 STAR-Transformer에 대한 입력으로 사용된다. 다중 클래스 토큰은 전체 토큰의 양 끝과 멀티-모달 토큰 사이에 연결되어, 위치 임베딩(pos)과 함께 전역 그리드 토큰($\mathbf{T}_g$)과 관절 맵 토큰($\mathbf{T}_j$)을 균등하게 학습할 수 있다.

수학식 2

$$\mathbf{Z} = (\mathbb{C}_g \oplus \mathbb{T}_g) \oplus (\mathbb{C}_j \oplus (\mathbb{T}_j + \text{pos})) \oplus \mathbb{C}_{tot}$$

여기서, $\mathbb{C}_g$ 는 $\mathbb{T}_g$ 를 학습하기 위한 클래스 토큰이고, $\mathbb{C}_j$ 는 $\mathbb{T}_j$ 를 학습하기 위한 클래스 토큰이며, $\mathbb{C}_{tot}$ 은 모든 멀티-모달 토큰을 학습하기 위한 클래스 토큰이다. 최종 토큰 $\mathbf{Z}$ 는 STAR-Transformer의 입력으로 사용된다.

시공간 특징을 학습하기 위해 STAR-Transformer는 전체 시공간 주의(full spatio-temporal attention, FAttn), 시공간 교차 주의(spatio-temporal cross-attention, STCA), 지그재그 주의(zigzag attention, ZAttn) 및 이진 주의(binary attention, BAttn)를 사용할 수 있다. FAttn은 시공간 차원의 모든 토큰을 사용하여 모든 토큰과 자기-주의 사이의 관계를 학습할 수 있다. 그러나 FAttn만으로는 동작 인식에 필요한 시간 차이를 학습하기에 충분하지 않다. 따라서 STAR-Transformer에서는, ZAttn과 BAttn을 비디오 시퀀스의 미세한 차이와 거친 차이를 순차적으로 학습하는 데 사용한다.

ZAttn과 BAttn은 먼저 모든 토큰을 시간 축 중심의 프레임으로 분리하는 과정을 거친다. 도 1에 도시된 바와 같이, ZAttn은 시간 축 T 를 중심으로 한 프레임을 건너뛰어 지그재그 모양으로 분리한다. BAttn은 시간 축 T 를 중심으로 프레임을 반으로 분할하여 이전 프레임에 해당하는 토큰과 다음 프레임에 해당하는 토큰을 디커플링하며, 이진 형태의 토큰 디커플링을 보여준다. 이때 ZAttn과 BAttn은 모든 토큰이 분리되는 순서에 따라 각각 $Q, K, V \in \mathbb{R}^{S \times (T/2)}$ 와 $Q', K', V' \in \mathbb{R}^{S \times (T/2)}$ 의 두 개의 토큰 세트를 생성할 수 있다. 이후, 이 두 토큰 세트 O 와 O' 는 교차-주의(cross-attention)의 출력을 연결하여 최종 STCA 결과를 생성하는 데 사용된다. 이에 대한 공식은 다음 수학식 3 내지 5와 같이 표현할 수 있다.

수학식 3

$$O = \sum_t^{T/2} \sum_s^S \text{Softmax} \left\{ \frac{Q_{s,t} \cdot K_{s,t}}{\sqrt{d_h}} \right\} V_{s,t}$$

수학식 4

$$O' = \sum_t^{T/2} \sum_s^S \text{Softmax} \left\{ \frac{Q'_{s,t} \cdot K'_{s,t}}{\sqrt{d_h}} \right\} V'_{s,t}$$

수학식 5

$$\text{STCA}(\text{Decoupling}(\mathbf{Z})) = O \oplus O'$$

- [0137] 이러한 STCA를 사용하여 시공간 교차 주의 메커니즘을 수행함으로써, STAR-Transformer는 큰 비디오 프레임과 작은 비디오 프레임 간의 특징 차이를 학습할 수 있다.
- [0139] 도 3은 본 발명의 일실시예에 따른 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템(10)의 구성을 도시한 도면이고, 도 4는 본 발명의 일실시예에 따른 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템(10)의 전체 구조를 설명하기 위해 도시한 도면이다. 도 3 및 도 4에 도시된 바와 같이, 본 발명의 일실시예에 따른 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템(10)은, 입력 비디오 및 골격(skeleton) 데이터를 사용해 인간 행동을 인식하는 행동 인식 시스템으로서, 데이터 처리부(100) 및 행동 인식부(200)를 포함하여 구성될 수 있다. 여기서, STAR++ 트랜스포머(20)는, 주의 메커니즘을 수행하되, 입력 토큰을 시간 차원을 기준으로 미리 설정된 간격 항(interval term)에 따라 나누어 두 토큰 그룹을 생성하고, 두 토큰 그룹 간의 교차 주의 메커니즘을 수행해 행동 변화를 학습하는 간격 주의(Interval Attention, IAttn) 모듈을 포함하는 복수의 인코더 레이어로 구성될 수 있다.
- [0141] 도 4의 (a)에 도시된 멀티-모달 특징을 사용한 토큰화 과정은 데이터 처리부(100)에서 수행되는데, STAR-Transformer의 토큰화 과정과 동일하다. 도 4의 (b)에 도시된 바와 같이, STAR++ 트랜스포머(20)는, D3D 토큰 선택 및 시공간 교차 주의 간격(interval spatio-temporal cross attention)을 기반으로 한 IAttn 과정을 포함할 수 있다.
- [0143] 데이터 처리부(100)는, 입력 비디오와 골격 데이터로부터 입력 토큰을 구성할 수 있다. 보다 구체적으로, 데이터 처리부(100)는, 입력 비디오 프레임에서 전역(global) 특징 및 로컬(local) 특징을 추출하고, 전역 특징을 토큰화한 전역 그리드 토큰과 로컬 특징과 골격 데이터를 결합해 토큰화한 관절 맵 토큰을 결합하되, 다중 클래스 토큰을 포함해 결합하여 입력 토큰을 구성할 수 있다. 이때, 데이터 처리부(100)는, 골격 데이터의 관절 좌표를 사용해 생성한 관절 열 지도를 로컬 특징과 결합해 토큰화하여 관절 맵 토큰을 생성할 수 있다. 그 다음, 데이터 처리부(100)는, 전역 그리드 토큰과 관절 맵 토큰을 결합하되, 다중 클래스 토큰을 전체 토큰의 양 끝과 전역 그리드 토큰과 관절 맵 토큰의 사이에 연결해 입력 토큰을 구성할 수 있다.
- [0145] 즉, STAR++ 트랜스포머(20)에서는, 전술한 STAR-Transformer와 동일하게 입력 토큰을 구성할 수 있다. 보다 구체적으로, 비디오 프레임과 골격을 입력으로 수신하고 MC18을 통해 비디오 프레임에서 전역 및 로컬 특징을 추출할 수 있다. 전역 특징은 RGB를 나타내는 전역 그리드 토큰이며, 로컬 특징은 골격을 나타내는 관절 맵 토큰으로 토큰화될 수 있다. 이후 두 개의 멀티-모달 토큰을 멀티 클래스 토큰과 결합하여 유연하게 처리하고 STAR++에 대한 입력으로 사용할 수 있다.
- [0147] 행동 인식부(200)는, 입력 토큰을 STAR++ 트랜스포머(20)에 입력해 최종 행동 레이블을 추정할 수 있다. 보다 구체적으로, 행동 인식부(200)에서는 STAR++가 D3D 토큰 선택 및 IAttn을 사용하여 시간 경과에 따른 비디오의 특징 차이를 학습할 수 있다. 마지막으로, 학습된 다중 클래스 토큰을 평균화한 후, 간단한 다층 퍼셉트론(MLP)을 사용하여 최종 비디오 인식에 대한 확률 값을 도출할 수 있다. 도 2 및 도 4에 도시된 바와 같이, STAR++ 트랜스포머(20)는 주의 메커니즘을 수행하되, 입력 토큰을 시간 차원을 기준으로 미리 설정된 간격 항(interval term)에 따라 나누어 두 토큰 그룹을 생성하고, 두 토큰 그룹 간의 교차 주의 메커니즘을 수행해 행동 변화를 학습하는 간격 주의(Interval Attention, IAttn) 모듈을 포함하는 복수의 인코더 레이어로 구성될 수 있다. 여기서, 간격 항은, 인코더 레이어마다 미리 설정되되, 레이어가 깊어짐에 따라 값이 커지도록 설정될 수 있다.
- [0149] 도 5는 본 발명의 일실시예에 따른 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템(10)에서, STAR++ 트랜스포머(20)의 인코더 레이어 상세 구조를 도시한 도면이다. 도 4 및 도 5에 도시된 바와 같이, 본 발명의 일실시예에 따른 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템(10)에서, STAR++ 트랜스포머(20)는 STAR-Transformer가 인코더-디코더 구조로 된 것과는 달리 인코더 레이어로만 구성되며, L층

의 인코더 레이어로 구성될 수 있다. 인코더 레이어는, 입력 토큰을 구성하는 전체 토큰에 대해 전체 시공간 주의(Full Spatio-temporal Attention, FAttn) 메커니즘을 수행하는 전체 주의 모듈(220)과; 전체 주의 모듈(220)의 출력을 시간 차원을 기준으로 미리 설정된 간격 항에 따라 나누어 두 토큰 그룹을 생성하고, 두 토큰 그룹 간의 교차 주의 메커니즘을 수행하는 간격 주의 모듈(230)을 포함하여 구성될 수 있으며, D3D 토큰 선택 모듈(210)을 더 포함하여 구성될 수 있다. 보다 구체적으로, 각 인코더 레이어는 D3D 토큰 선택 모듈(210), 전체 주의 모듈(220), D3D 토큰 선택 모듈(210), 간격 주의 모듈(230)의 순서로 연산될 수 있다.

[0151] 이하에서는, 도 3 내지 도 5를 참조하여 본 발명의 일실시예에 따른 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템(10)에서, STAR++ 트랜스포머(20)의 세부적인 구성에 대해 상세히 설명하도록 한다.

[0153] D3D 토큰 선택 모듈(210)은, 변형 가능한 3D 토큰(D3D token) 선택을 처리할 수 있다. 즉, D3D 토큰 선택 모듈(210)은, 입력된 토큰에서 전역 그리드 토큰을 추출해 오프셋과 기준점을 결합하여 최종 좌표를 생성하며, 최종 좌표로부터 키 및 값을 생성할 수 있다. D3D 토큰 선택은 동적 토큰 선택 방법으로, 3D RGB의 시공간 특징을 변형하여 캡처할 수 있다.

[0155] 트랜스포머의 자기-주의 메커니즘은 각 이미지의 공간적 관계성을 판단하는 데 특화되어 있어 영상 분류나 객체 탐지에서 좋은 성능을 보여주었지만 상당한 계산 비용이 필요하다. 따라서 변형 가능한 토큰 선택 방법 또는 토큰을 점진적으로 줄이거나 삭제하는 방법을 적용하는 트랜스포머 기반 모델이 증가하고 있다. 이러한 모델은 변형 가능한 토큰 선택 및 토큰 중단 기술을 사용하여 토큰의 크기를 줄여 연산 비용의 대부분을 차지하는 자기-주의 연산을 효과적으로 줄이면서 성능을 유지하거나 개선한다.

[0157] D3D 토큰 선택 모듈(210)은, 오프셋을 사용하여 학습 결과에 영향을 미치는 토큰의 위치를 학습하고 토큰 선택을 적용하는 기술(Z. Xia, X. Pan, S. Song, L. E. Li and G. Huang, "Vision transformer with deformable attention," in Proc. CVPR, New Orleans, LA, USA, 2022, pp. 4794-4803. 및 M. Jaderberg, K. Simonyan and A. Zisserman, "Spatial transformer networks," in Proc. NeurIPS, Montreal, QC, Canada, 2015, pp. 1-9.)의 접근 방식에서 영감을 받은 새로운 동적 토큰 선택 방법을 적용할 수 있다. 이러한 D3D 토큰 선택 모듈(210)은, 토큰을 동적으로 선택하기 때문에 결과에 큰 영향을 미치는 토큰을 선택하고 학습하여 성능을 향상시킬 수 있다. 또한 토큰 크기를 줄일 수 있으므로 운영 비용을 효과적으로 줄일 수 있다.

[0159] 도 6은 본 발명의 일실시예에 따른 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템(10)에서, D3D 토큰 선택 모듈(210)의 D3D 토큰 선택 구조를 도시한 도면이다. 도 6에 도시된 바와 같이, 본 발명의 일실시예에 따른 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템(10)의 D3D 토큰 선택 모듈(210)은, 토큰 세트 X에서 전역 그리드 토큰 X_g 만을 추출하고 오프셋과 기준점을 결합하며 토큰 선택 프로세스를 통해 사용 가능한 키와 값 토큰을 생성할 수 있다. 이하에서는, 도 6을 참조하여 D3D 토큰 선택을 상세히 설명하도록 한다.

[0161] 토큰 세트 X는 다음과 같이 구성될 수 있다.

수학식 6

$$X = [X_{g_{ck}} || X_g || X_{j_{ck}} || X_j || X_{tot_{ck}}]$$

[0163]

[0165] 여기서, $X_{g_{cls}}$ 는 전역 그리드에 대한 클래스 토큰, X_g 는 전역 그리드 토큰, $X_{j_{cls}}$ 는 관절 맵에 대한 클래스 토큰, X_j 는 관절 맵 토큰, $X_{tot_{cls}}$ 는 다중 클래스 토큰이다. 여기서, $||$ 은 연결을 나타낸다.

[0167] 도 6에 도시된 바와 같이, 최종 키(K) 및 값(V)을 생성하기 위해, 토큰 세트 X에서 RGB 이미지에 해당하는 전역 그리드 토큰 X_g 만 추출하여 D3D 토큰 선택에 공급할 수 있다. 그런 다음 3D 컨볼루션을 사용하여 좌표를 학습하기 위한 오프셋 θ_{offset} 함수를 생성하고 기준점 p를 추가하여 최종 좌표를 생성할 수 있다.

수학식 7

$$[0169] \quad \tilde{X}_g = \phi(X_g; p + \Delta p), \Delta p = \theta_{offset}(X_g)$$

[0171] 여기서, ϕ 는 기준점과 오프셋을 사용하여 토큰의 좌표값을 학습하고 샘플링하는 함수이다. 다음으로, $\tilde{X}_g$ 에 학습 가능한 키 및 값 가중치인 W_k 및 W_v 를 곱하여 시간적(temporal) 키 $\bar{K}$ 및 값 $\bar{V}$ 를 생성할 수 있다.

수학식 8

$$[0173] \quad \bar{K} = \tilde{X}_g W_k, \bar{V} = \tilde{X}_g W_v$$

[0175] $\bar{K}$ 와 $\bar{V}$ 의 좌표는 실수이기 때문에, 3D 토큰의 좌표는 2선형 보간법을 이용한 정수 과정을 통해 계산될 수 있다.

[0177] 도 6에 도시된 바와 같이, 시간적 $\bar{K}$ 와 $\bar{V}$ 는 $X_{g_{cls}}$, $X_{j_{cls}}$, X_j , $X_{tot_{cls}}$ 와 연결되어 최종 K와 V가 생성될 수 있다.

수학식 9

$$[0179] \quad K = [X_{g_{cls}} || \bar{K} || X_{j_{cls}} || X_j || X_{tot_{cls}}]$$

수학식 10

$$[0181] \quad V = [X_{g_{cls}} || \bar{V} || X_{j_{cls}} || X_j || X_{tot_{cls}}]$$

- [0183] D3D 토큰 선택은 도 5와 같이 FAttn 및 IAttn 이전에 각각 수행될 수 있다. 첫 번째는, D3D 토큰 선택을 통해 생성된 K와 V를 이용하여 FAttn을 적용하고, 계층 정규화를 통해 토큰 디커플링을 수행할 수 있다. 그 다음에 두 번째 D3D 토큰 선택이 적용되고, IAttn을 적용한 후 레이어 정규화 및 MLP를 통해 출력 벡터가 생성될 수 있다.
- [0185] 전체 주의 모듈(220)은, 입력 토큰을 구성하는 전체 토큰에 대해 전체 시공간 주의 메커니즘을 수행할 수 있다. 시공간 데이터를 학습하는 기존 트랜스포머 모델은 일반적으로 시공간 차원 T의 모든 토큰들에 대해 주의 메커니즘을 수행하는 전체 시공간 주의를 사용한다. 시간 차원인 T와 공간차원인 S의 모든 토큰들에 대해 주의 메커니즘을 적용하므로, 복잡도는 $O(T^2S^2)$ 로 매우 커진다.
- [0187] 간격 주의 모듈(230)은, 입력 토큰을 시간 차원을 기준으로 미리 설정된 간격 항에 따라 나누어 두 토큰 그룹을 생성하고, 두 토큰 그룹 간의 교차 주의 메커니즘을 수행해 행동 변화를 학습할 수 있다.
- [0189] 종래의 STAR-Transformer는 우수한 동작 인식 성능을 보여주지만, 인코더와 디코더로 구분되어 두 배의 레이어를 사용하기 때문에 레이어 깊이에 따른 특징 학습에 최적화되어 있지 않다. 본 발명에서는 STAR-Transformer의 ZAttn과 BAttn을 사용하는 대신 레이어 깊이에 따라 시간 축의 크고 작은 특징 사이의 차이를 점진적으로 학습할 수 있는 IAttn을 제안한다. STAR-Transformer의 계산 복잡성을 줄이기 위해 IAttn은 인코더-디코더 구조를 단일 인코더 형태로 통합할 수 있다. 따라서 $(L \times 2)$ 레이어를 사용하는 STAR-Transformer와 달리 STAR++는 L 레이어만 사용할 수 있으므로 네트워크가 단순해질 수 있다.
- [0191] 본 발명에서 제안한 STAR++의 층상구조에서는 도 5에 도시된 바와 같이 모든 토큰에 대한 관계를 FAttn을 통해 학습하고(Full Attention) 정규화(Norm) 후 IAttn 토큰 디커플링(Token Decoupling)을 적용할 수 있다. 여기서, 간격 항은, 인코더 레이어마다 미리 설정되며, 레이어가 깊어짐에 따라 값이 커지도록 설정될 수 있다. 즉, 레이어의 깊이에 따라 디커플링 간격 항의 값은 점진적으로 증가할 수 있다. 예를 들어, 3개 레이어($L=3$)에 대해 $I=\{1,4,8\}$ 또는 $I=\{1,2,4\}$ 의 간격 항을 가진 구조를 사용할 수 있다.
- [0193] 도 7은 본 발명의 일실시예에 따른 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템(10)에서, 간격 주의(Interval Attention, IAttn)를 설명하기 위해 도시한 도면이다. 도 7에는 간격 항 $I=\{1,2,4\}$ 를 사용한 예를 도시하였다.
- [0195] 1번째 레이어에서 간격 항은 1이기 때문에, STAR-Transformer의 ZAttn과 동일하게 한 프레임 간격으로 토큰 세트가 분리된다(녹색 토큰). 디커플링 토큰은 총 2개의 Q, K 및 V 토큰 세트에서 생성되며, K 및 V 토큰 세트는 D3D 토큰 선택(파란색 토큰)을 통해 IAttn에서 사용할 K 및 V를 생성할 수 있다. Q, K 및 V를 사용하여 교차 주의를 적용한 후 두 출력이 연결되어 $(1+1)$ 계층의 입력값으로 사용될 수 있다.
- [0197] $(1+1)$ 번째 레이어에서 토큰은 간격 항이 2가 되고 두 개의 프레임 간격을 갖도록 토큰 세트가 분리될 수 있다. D3D 토큰 선택과 IAttn은 1 레이어와 동일한 방식으로 적용될 수 있다. L번째 계층에서는 간격 항이 4로 설정되고 동일한 D3D 토큰 선택 및 IAttn이 적용될 수 있다.
- [0199] 도 8은 본 발명의 일실시예에 따른 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템(10)에서, 간격 주의 모듈(230)에서 수행되는 IAttn 프로세스의 의사 코드(pseudo code)를 도시한 도면이다.
- [0201] 도 3 내지 도 5에 도시된 바와 같이, 기본 STAR-Transformer와 달리 STAR++는 인코더-디코더 구조가 아닌 L 인

코더 레이어로만 구성된다. 기본 STAR-Transformer의 인코더-디코더 구조는 L 레이어로 구성되어 있지만 ZAttn 및 BAttn을 포함한 교차 주의를 적용하기 때문에 실제로 $(L \times 2)$ 레이어를 사용하게 된다. 그러나 STAR++은 인코더 레이어와 L 레이어만 사용하도록 개선되었기 때문에 기본 모델보다 계산 복잡도 측면에서 더 효율적이다. 또한, 레이어가 깊어짐에 따라 전역 특징의 공급에 초점을 맞추어 점진적으로 학습할 수 있다. 즉, STAR++은 레이어가 깊어짐에 따라 전역 특징에 초점을 맞추는 트랜스포머의 특성을 고려하여 학습한다고 할 수 있다.

[0203] 도 5에 도시된 바와 같이, STAR++ 트랜스포머(20)의 인코더 레이어는, D3D 토큰 선택 모듈(210)이 입력 토큰에서 키 및 값을 생성하는 단계(S10), 전체 주의 모듈(220)이 D3D 토큰 선택 모듈(210)에서 생성된 키 및 값을 이용해 전체 시공간 주의 메커니즘을 수행하는 단계(S20), 전체 주의 모듈(220)의 출력을 시간 차원을 기준으로 미리 설정된 간격 항에 따라 나누어 두 토큰 그룹을 생성하는 단계(S30), D3D 토큰 선택 모듈(210)이 생성된 두 토큰 그룹에서 키 및 값을 생성하는 단계(S40) 및 간격 주의 모듈(230)이, 단계 S40에서 생성한 키 및 값을 이용해, 단계 S30에서 생성한 두 토큰 그룹 간의 교차 주의 메커니즘을 수행하는 단계(S50)를 수행할 수 있다.

[0205] 도 5에 도시된 STAR++ 인코더 레이어의 구조는 다음 수학적 식 11 내지 13을 통해 표현할 수 있다.

수학적 식 11

[0207] $\bar{Z}_l = \text{LN}\{\text{FAttn}(\text{D3D}(\text{Z}_{l-1}))\} + \text{Z}_{l-1}, l \in \{1, 2, \dots, L\}$

수학적 식 12

[0209] $\hat{Z}_l = \text{IAttn}(\text{D3D}(\text{Decoupling}(\bar{Z}_l, \text{I}_l)))$

수학적 식 13

[0211] $\text{Z}_l = \text{LN}\{\text{MLP}(\bar{Z}_l)\} + \hat{Z}_l$

[0213] 여기서, l은 트랜스포머 레이어의 수이고, LN은 레이어 정규화이다. 마지막으로 행동 인식부(200)는, 공식을 기반으로 학습된 다중 클래스 토큰을 평균화하고 간단한 MLP를 사용하여 행동 레이블을 예측할 수 있다.

[0215] 도 9는 본 발명의 일실시예에 따른 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 방법의 흐름을 도시한 도면이다. 도 9에 도시된 바와 같이, 본 발명의 일실시예에 따른 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 방법은, 컴퓨터에서 각 단계가 수행되는 영상 및 골격(skeleton) 데이터를 사용해 인간 행동을 인식하는 행동 인식 방법으로서, 입력 비디오와 골격 데이터로부터 입력 토큰을 구성하는 단계(S100) 및 입력 토큰을 STAR++ 트랜스포머(20)에 입력해 최종 행동 레이블을 추정하는 단계(S200)를 포함하여 구현될 수 있다.

[0217] 각각의 단계들과 관련된 상세한 내용들은, 앞서 본 발명의 일실시예에 따른 본 발명의 일실시예에 따른 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템(10)과 관련하여 충분히 설명되었으므로, 상세한

설명은 생략하기로 한다.

[0219] 실험

[0220] 제안된 STAR++에 대한 교육 데이터 세트 및 하이퍼 파라미터를 포함한 세부 정보를 설명할 수 있다. 또한 해당 벤치마크 데이터 세트에 대해 최첨단(SoTA) 모델을 사용하여 성능 비교 및 정량적 평가를 수행했다. 제안된 STAR++의 D3D 토큰 선택 및 IAttn을 포함하여 각 모듈의 효과를 입증하기 위해 절제 연구가 수행되었다.

[0222] A. 데이터 세트

[0223] 행동 인식 벤치마크 데이터 세트인, Penn-Action 데이터 세트(Penn action dataset), NTU-RGB+D 60 및 NTU-RGB+D 120을 사용하여 광범위한 실험을 수행했다.

[0225] Penn-Action 데이터 세트에는 baseball_pitch, clean_and_jerk, pull_ups 등 15개 스포츠 동작의 비디오 시퀀스 2,326개와 각 비디오 시퀀스에 대한 인간 골격 주석 13개가 포함되어 있다.

[0227] NTU-RGB+D 60(이하, NTU60)은 일상, 건강 관련 및 상호작용 행동을 위한 60개의 행동 클래스와 56,880개의 비디오 샘플을 포함하는 대규모 데이터 세트이다. 데이터 세트는 각 샘플에 대한 RGB 비디오 시퀀스, 깊이 맵 시퀀스, 3D 골격 데이터 및 적외선(IR) 비디오 클립을 제공한다. RGB 비디오의 픽셀 해상도는 1920×1080 이고, 깊이 맵 시퀀스와 IR 비디오의 픽셀 해상도는 모두 512×512 이며, 3D 골격 데이터에는 각 프레임에 대한 25개의 관절 좌표가 포함되어 있다. NTU60을 사용한 평가는 교차 피사체(XSub)와 교차 뷰(XView)의 두 가지 방법을 따랐다. (1) XSub의 경우 40명의 피실험자의 행동 중 20명의 데이터가 학습에 사용되었고, 나머지 20명의 피실험자의 데이터가 평가에 사용되었다. (2) XView의 경우 3개의 카메라 뷰 중에서, 2개는 학습에 사용되었고 1개는 평가에 적용되었다.

[0229] NTU-RGB+D 120(이하, NTU120)은 NTU60의 확장으로 60개의 동작 클래스와 57,600개의 비디오 샘플을 추가하고 총 120개의 액션 클래스와 114,480개의 비디오 샘플을 포함한다. NTU120을 이용한 평가는 XSub과 교차설정(Xset) 방식의 두 가지 방법을 사용하였으며, (1) XSub의 경우 106명의 피실험자의 행동 중 53명의 피실험자에 대한 데이터를 학습에 사용하였으며, 나머지 53명의 피실험자에 대한 데이터를 평가에 사용하였다. (2) Xset의 경우, 모든 샘플은 교육용 짝수 집합 ID와 테스트용 홀수 집합 ID를 가지고 있었다. 즉, 16개의 설정은 교육에 사용되었고, 나머지 16개의 설정은 평가에 사용되었다.

[0231] B. 구현 세부 정보

[0233] 제안된 STAR++ 트랜스포머(20)는 Kinetics-400 (W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman and A. Zisserman, "The kinetics human action video dataset," 2017, arXiv preprint arXiv:1705.06950.)을 백본으로 사전 훈련된 PyTorch 및 ResNet MC18을 사용하여 구현되었다. STAR++ 트랜스포머(20) 학습에 총 16개 프레임이 사용되었으며, 배치 크기는 8, 학습 에포크 100, 확률적 기울기 하강 최적화기(a stochastic gradient descent optimizer), 학습 속도는 $2e-4$, 모멘텀은 0.9이다. 실험은 4개의 NVIDIA Tesla V100 GPU로 구성된 환경에서 수행되었다.

[0235] C. 최첨단 방법과의 비교

[0237] Penn-Action과 동작 인식 벤치마크 데이터 세트인 NTU-RGB+D 60 및 120을 사용하여 SoTA 방법과 STAR++의 성능을 비교했다.

[0239] NTU-RGB+D 60 및 120의 결과

[0240] 표 1은 NTU-RGB+D 60 및 120 데이터 세트를 사용한 SoTA 모델의 성능 비교 결과를 나타낸 것이다. SoTA 방법으로는, PoseMap (M. Liu and J. Yuan, "Recognizing human actions as the evolution of pose estimation maps," in Proc. CVPR, Salt Lake City, UT, USA, 2018, pp. 1159-1168.), MMTM (V. Reza, H. Joze, A. Shaban, M. L. Iuzzolino and K. Koishida, "Mmtm: Multimodal transfer module for cnn fusion," in Proc. CVPR, Virtual, 2020, pp. 13289-13299.), VPN (S. Das, S. Sharma, R. Dai, F. Bremond and M. Thonnat, "Vpn: Learning video-pose embedding for activities of daily living," in Proc. ECCV, Virtual, 2020, pp. 72-90.), DualHead-Net (T. Chen, D. Zhou, J. Wang, S. Wang, Y. Guan, X. He and E. Ding, "Learning multi-granular spatio-temporal graph network for skeleton-based action recognition," in Proc. ACM MM., Chengdu, China, 2021, pp. 4334-4342.), Skeletal GNN (A. Zeng, X. Sun, L. Yang, N. Zhao, M. Liu and Q. Xu, "Learning skeletal graph neural networks for hard 3d pose estimation," in Proc. ICCV, Virtual, 2021, pp. 11436-11445.), CTR-GCN (Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in Proc. ICCV, Virtual, 2021, pp. 13359-13368.), InfoGCN (H. G. Chi, M. H. Ha, S. Chi, S. W. Lee, Q. Huang and K. Ramani, "Infogcn: Representation learning for human skeleton-based action recognition," in Proc. CVPR, New Orleans, LA, USA, 2022, pp. 20186- 20196.), 3s-AimCLR (T. Guo, H. Liu, Z. Chen, M. Liu, T. Wang and R. Ding, "Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition," in Proc. AAAI, Washington, DC, USA, 2022, pp. 762-770.), PoseC3D (H. Duan, Y. Zhao, K. Chen, D. Lin and B. Dai, "Revisiting skeleton-based action recognition," in Proc. CVPR, New Orleans, LA, USA, 2022, pp. 2969-2978.), KA-AGTN (Y. Liu, H. Zhang, D. Xu and K. He, "Graph transformer network with temporal kernel attention for skeleton-based action recognition," Knowledge-Based Syst., vol. 240, pp. 108146, Jan. 2022, doi: 10.1016/j.knosys.2022.108146.), STAR (STAR Transformer)와 비교하였다.

표 1

Method	Pre-training	Feature			NTU-RGB+D 60		NTU-RGB+D 120	
		RGB	Est. Skeleton	Anno. Skeleton	XSub	XView	XSub	XSet
PoseMap [52]	×	✓	✓		91.7	95.2	-	-
MMTM [31]	×	✓		✓	91.9	-	-	-
VPN [29]	×	✓		✓	95.5	98.0	86.3	87.8
DualHead-Net [53]	×			✓	92.0	96.6	88.2	89.3
Skeletal GNN [54]	×			✓	91.6	96.7	87.5	89.2
CTR-GCN [26]	×			✓	92.4	96.8	88.9	90.6
InfoGCN [27]	×			✓	93.0	97.1	89.8	91.2
3s-AimCLR [55]	✓			✓	86.9	92.8	80.1	80.9
PoseC3D [30]	✓	✓	✓		97.0	99.6	95.3	96.4
KA-AGTN [50]	×			✓	90.4	96.1	86.1	88.0
STAR [13]	×	✓		✓	92.0	96.5	90.3	92.7
STAR++	×	✓		✓	92.7	97.6	90.9	92.9

[0242]

[0244] 표 1에서 확인할 수 있는 바와 같이, PoseC3D는 NTU-RGB+D 60 및 120에서 가장 우수한 성능을 나타냈지만, 사전 훈련을 사용하기 때문에 PoseC3D를 제외한 실험 결과를 분석했으며, 추정된 포즈는 주석이 달린 포즈가 아닌 모델에 최적화되었다. 따라서 우리는 주로 유사한 조건에서 테스트된 SoTA 모델을 비교할 수 있다.

[0246] NTU-RGB+D 60 XSub의 경우 STAR++ 트랜스포머(20)는 VPN 및 InfoGCN 모델보다 약간 낮은 성능을 보였다. 이는 STAR++ 트랜스포머(20)는 16프레임과 2D 골격 데이터만 사용하는 반면, VPN은 64프레임과 3D 골격 데이터를 학

습에 사용하기 때문이다. 따라서 STAR++ 트랜스포머(20)보다 풍부한 특징을 학습할 수 있으며 성능이 약간 더 높은 것으로 판단되었다. 그러나 보다 복잡한 NTU-RGB+D 120 데이터 세트의 경우 STAR++ 트랜스포머(20)은 VPN에 비해 XSub의 경우 4.6%, XSet의 경우 5.1%의 정확도를 향상시켰다. 골격만 사용한 InfoGCN의 경우 Xsub에만 적용된 STAR++보다 0.3% 높은 정확도를 보였으나 카메라 뷰 차이가 있는 XView의 경우 STAR++ 트랜스포머(20)보다 0.5% 낮은 정확도를 보였다. 그러나 NTU-RGB+D 120으로, STAR++는 InfoGCN에 비해 Xsub의 경우 1.1%, Xset의 경우 1.7%의 정확도 향상을 보였다.

[0248] NTU-RGB+D 60, NTU-RGB+D 120 XSub과 XSet의 결과, STAR++ 트랜스포머(20)가 학습이 부족함에도 불구하고 PoseC3D를 제외하고는 SoTA 방법 중 가장 높은 점수를 보였다. 결과로부터 제안된 STAR++ 알고리즘이 동작 인식을 위한 매우 경쟁력 있는 알고리즘임을 확인할 수 있다. STAR++ 트랜스포머(20)와 마찬가지로 트랜스포머 기반 골격 정보를 토큰화하여 동작 인식에 사용하는 KA-AGTN은 STAR++보다 XSub의 정확도가 4.8%, XSet의 정확도가 4.9% 낮은 것으로 나타났다. 또한 기본 네트워크인 STAR-Transformer에 비해 NTU-RGB+D 60 및 120에 걸쳐 STAR++ 트랜스포머(20)의 정확도가 고르게 증가했다. 이러한 결과는 STAR++ 트랜스포머(20)의 토큰화 방법이 각 다중 모드 특징을 더 유연하게 결합할 수 있음을 보여준다.

[0250] Penn-Action 데이터 세트의 결과

[0251] 표 2는 Penn-Action 데이터 세트를 사용하여 SoTA 모델과 제안된 STAR++ 트랜스포머(20) 간의 성능 비교 결과를 나타낸 것이다. SoTA 방법으로는, 3D Deep (C. Cao, Y. Zhang, C. Zhang and H. Lu, "Body joint guided 3-d deep convolutional descriptors for action recognition," IEEE Trans. on Cybernetics, vol. 48, no. 3, pp.1095-1108, Mar. 2017, doi: 10.1109/TCYB.2017.2756840.) PoseMap, Multi CNN (D. C. Luvizon, D. Picard and H. Tabia, "2d/3d pose estimation and action recognition using multitask deep learning," in Proc. CVPR, Salt Lake City, UT, USA, 2018, pp. 5137-5146.), HDM-BG (R. Zhao, W. Xu, H. Su and Q. Ji, "Bayesian hierarchical dynamic model for human action recognition," in Proc. CVPR, Long Beach, CA, USA, 2019, pp 7733-7742.), Pr-VIPE (J. Sun, J. Zhao, L.-C. Chen, F. Schroff, H. Adam and T. Liu, "Viewinvariant probabilistic embedding for human pose," in Proc. ECCV, Virtual, 2020, pp 53-70.), UNIK (D. Yang, Y. Wang, A. Dantcheva, L. Garattoni, G. Francesca and F. Bremond, "Unik: A unified framework for real-world skeleton-based action recognition," in Proc. BMVC, Virtual, 2021, pp. 1-13.) 및 STAR Transformer와 비교하였다.

표 2

Method	Pre-train	Feature		Acc. (%)
		RGB	Anno. Skeleton	
3D Deep [56]	×	✓	✓	98.1
PoseMap [52]	×	✓	✓	98.2
Multi CNN [57]	×	✓	✓	98.6
HDM-BG [58]	×		✓	93.4
Pr-VIPE [51]	✓		✓	97.5
UNIK [28]	✓		✓	97.9
STAR [13]	×	✓	✓	98.7
STAR++	×	✓	✓	99.3

[0253]

[0255] GNN을 기반으로 골격 정보를 사용하는 단일 모델 접근법인 Pr-VIPE와 UNIK는 사전 훈련된 모델로 각각 97.5%와 97.9%의 정확도를 보였다. 그러나 이러한 방법은 사전 훈련 없이 멀티-모달 접근 방식을 사용하는 Muti CNN, STAR-Transformer 및 STAR++ 트랜스포머(20) 모델에 비해 상대적으로 낮은 정확도를 보여주었다. 실험 결과로부터, 우리는 단일-모달 접근법에 대해 멀티-모달 방법을 사용했을 때 성능이 더 낫다는 것을 알 수 있다. 제안된 STAR++ 트랜스포머(20) 모델의 정확도는 99.3%로 기본 STAR-Transformer 모델보다 0.6% 높고 Penn-Action

데이터 세트에서 SoTA 성능을 달성했다. 이러한 결과는 STAR-Transformer의 확장 버전인 STAR++ 트랜스포머(20)에 포함된 각 모듈이 효율적으로 작동함을 보여준다.

[0257] D. 모델 분석 및 절제 연구(ablation study)

[0259] 제안된 STAR++ 알고리즘의 성능을 실험적으로 검증하였다. 모든 실험은 Penn-Action 데이터 세트를 사용하여 수행되었다.

[0261] IAttn 및 D3D의 성능 평가

[0262] STAR++의 핵심 모듈인 간격 주의 모듈(230)의 교차 주의 메커니즘(IAttn)과 D3D 토큰 선택 모듈(210)의 토큰 선택(D3D)이 모델 학습에 어떤 영향을 미치는지 테스트했다. 표 3은 각 모듈을 함께 사용할 때와 개별적으로 사용할 때의 성능 비교 결과를 보여준다.

표 3

IAttn	D3D	Accuracy (%)
×	×	98.7
×	✓	98.9
✓	×	99.1
✓	✓	99.3

[0264]

[0266] STAR++ 트랜스포머(20)는 STAR-Transformer를 기반으로 하는 구조이기 때문에 두 모듈을 사용하지 않을 때 기본 STAR-Transformer를 사용하는 것과 동일할 수 있다. 표 3에서 보는 바와 같이 D3D만을 사용하였을 때는 어느 모듈도 사용하지 않은 경우에 비해 정확도가 0.2% 증가하였다. 마찬가지로, IAttn을 단독으로 사용했을 때, 두 모듈 모두 적용되지 않았을 때와 비교하여 정확도가 0.4% 증가했다. 마지막으로, 모든 모듈을 함께 사용했을 때 최고 성능(99.3%)을 달성했다. 결과는 각 STAR++ 트랜스포머(20)의 모듈이 상호 보완적인 특성을 가지고 있으며 조화롭게 결합할 때 성능을 더욱 향상시킬 수 있음을 보여준다.

[0268] 단일 및 다중(멀티) 모달 접근 방식 간의 성능 비교

[0269] 교차-모달 메커니즘이 단일 및 멀티-모달 토큰에 어떤 영향을 미치는지 확인하기 위해 다중 클래스 토큰을 사용할 때와 사용하지 않을 때의 성능을 추가로 비교했다. 표 4에서 보는 바와 같이, 단일-모달 특징 중 다중 클래스 토큰 사용 여부와 관계없이 RGB만 사용했을 때 스켈레톤만 적용했을 때보다 성능이 우수함을 확인할 수 있다. 또한 다중 클래스 토큰을 사용할 때는 사용하지 않을 때보다 성능이 더 높았다. 다중 클래스 토큰을 사용하는 경우 멀티-모달 성능이 단일-모달 토큰보다 0.8%~1.1% 더 높았다. 이 결과로부터, 다중 클래스 토큰이 모델의 성능을 향상시킬 뿐만 아니라 멀티-모달 토큰을 유연하게 결합하여 성능을 향상시키는 데 도움이 된다는 것을 알 수 있다.

표 4

Multi Class Token	RGB	Skeleton	Accuracy (%)
✗	✓		97.7
		✓	97.5
	✓	✓	97.2
✓	✓		98.5
		✓	98.2
	✓	✓	99.3

[0271]

[0273] 간격 항에 따른 성능 비교

[0274] STAR++ 트랜스포머(20)의 IAttn에 사용된 간격 항이 학습에 미치는 영향을 확인하기 위해 추가 실험을 수행했다. 표 5와 같이 STAR++ 트랜스포머(20)에서 $I=\{1,4,8\}$ 의 간격 항을 사용하였을 때 가장 우수한 성능을 나타냄을 알 수 있었으며, 트랜스포머는 층(레이어)이 깊어질수록 국소 및 전역 특징으로부터 학습하는 데 집중하기 때문에 간격 항이 층의 수에 따라 증가할 때 STAR++ 트랜스포머(20)가 가장 높은 성능을 나타냄을 알 수 있었다. STAR++ 트랜스포머(20)는 이러한 특성을 본질적으로 이어받았기 때문에 간격 항이 점차 증가하고 층이 깊어질수록 전역 특성도 배울 수 있었다.

표 5

Interval term			Accuracy (%)
layer-1	layer-2	layer-3	
8	4	1	98.8
4	2	1	98.7
8	4	2	98.4
2	4	8	98.8
1	2	4	99.0
1	4	8	99.3

[0276]

[0278] 따라서 간격 항이 점차 증가하는 구조는 간격 항이 점차 감소하는 구조보다 더 나은 성능을 나타내는 것으로 나타났다. 실험 결과에 따라 제안된 STAR++ 트랜스포머(20) 모델은 수행된 모든 실험에서 간격 항 $\{1, 4, 8\}$ 을 사용했다.

[0280] 복잡성의 비교

[0281] 본 발명에서는 효율적인 토큰 선택 접근법으로 D3D 토큰 선택 방법을 제안했다. 제안된 D3D 토큰 선택의 효율성을 입증하기 위해 기본 STAR-Transformer와 STAR++ 트랜스포머(20)의 매개 변수 수와 연산량을 비교했다. 표 6에서 볼 수 있듯이 STAR++ 트랜스포머(20)에서 사용되는 매개변수의 수는 거의 2배 증가했지만 GFLOP 연산의 수는 약 1.06배 감소했다. 모델 크기가 증가했지만 제안된 STAR++ 트랜스포머(20)는 GFLOP 작업 수를 줄이고 성능을 향상시키는 측면에서 STAR-Transformer보다 더 효율적인 것으로 입증되었다.

표 6

Method	Param(M)	GFLOPs
STAR	16.8	25.6
STAR++	32.6	24.2

[0283]

[0285]

시각화

[0286]

행동 인식을 위해 제안된 IAttn 및 D3D 토큰 선택 접근 방식에 의해 입력 프레임의 시공간 특징 중 어느 부분에 초점이 맞춰져 있는지 확인하기 위해 주의 점수를 시각화했다.

[0288]

도 10은 본 발명의 일실시예에 따른 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템(10) 및 방법을 이용한 NTU-RGB+D 60 및 Penn 행동 데이터 세트의 시각화 결과를 도시한 도면이다. 즉, 학습에 사용되는 프레임의 시각화된 주의 지도를 나타낸 것이다. 도 10에 도시된 바와 같이 같이, 1인의 경우 동작이 크게 변화하는 위치에서 관심(주의)이 높았고, 여러 사람의 경우 동작이 큰 사람을 중심으로 관심이 높았다. 주의 점수는 악수 동작에서 손의 위치, 안경을 벗었을 때 손의 움직임의 궤적과 얼굴의 위치에 집중되었다. 또 테니스 서브와 폴업의 경우 크게 움직인 사람을 중심으로 주의 점수가 높았다. 이와 같이 RGB를 나타내는 전역 그리드 토큰의 관심은 동작 강도와 객체에 집중됨을 알 수 있다.

[0290]

도 11은 본 발명의 일실시예에 따른 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템(10) 및 방법에서, 골격 정보를 나타내는 관절 맵 토큰에 대한 주의 점수를 시각화하여 도시한 도면이다. 각 관절에 대한 주의 점수가 높은 4개의 프레임을 추출하여 시각화했다. 도면에서 주의 점수가 클수록 빨간색 원이 크고 주의 점수가 작을수록 원이 작다. 도 11에 도시된 바와 같이, 각 관절의 작용에 큰 차이가 있을 때 관절의 위치가 크게 변화하여 주의 점수가 높게 나타난다. 특히 시간이 지날수록 관절의 위치가 크게 바뀌거나 관절의 위치가 가까워졌을 때 주의 점수가 높게 나왔다. 예를 들어, 첫 번째 줄에 표시된 공 던지기 동작의 경우 던지기 전에 몸을 웅크렸을 때, 머리 높이가 낮아졌을 때, 공을 던진 후 발이나 손의 위치가 크게 변화했을 때 주의 점수가 높다. 또 쪼그려 앉을 때 웅크린 자세에서 기립 자세로 크게 이동하는 어깨와 무릎에 대해서도 주의 점수가 높았다. 4열 배드민턴 동작에서는 선수가 점프하는 동안 셔틀콕을 맞았기 때문에 스켈레톤 전체의 위치가 바뀌었고, 3번째 프레임에서는 전체 관절 점수가 증가했다. 이러한 결과로부터 제안된 STAR++ 트랜스포머(20)가 액션 비디오의 시공간 특성을 정확하게 식별하면서 학습한다는 것을 확인할 수 있다.

[0292]

전술한 바와 같이, 본 발명에서 제안하고 있는 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템(10) 및 방법에 따르면, 비디오 프레임과 사람의 골격 특징을 함께 학습해 행동 인식을 수행하되, 다중 표현 학습을 통해 영상 및 골격 기반 데이터를 유연하게 결합하고 다중 클래스 토큰을 적용해 이중 도메인의 데이터를 효과적으로 학습할 수 있으며, 개선된 시공간 교차 주의가 적용된 트랜스포머(STAR++ 트랜스포머(20))를 통해 시공간 특징을 함께 학습할 수 있어서 행동 인식 성능을 높일 수 있다.

[0294]

또한, 본 발명에서 제안하고 있는 개선된 시공간 교차 주의 트랜스포머를 이용한 인간 행동 인식 시스템(10) 및 방법에 따르면, 시간 차원을 기준으로 미리 설정된 간격 항(interval term)에 따라 나누어 두 토큰 그룹을 생성하고, 두 토큰 그룹 간의 교차 주의 메커니즘을 수행해 행동 변화를 학습하는 간격 주의 모듈(230)을 포함하는 복수의 인코더 레이어로 STAR++ 트랜스포머(20)를 구성하여, 깊이에 따라 시간 축의 크고 작은 특징 사이의 차이를 점진적으로 학습할 수 있고, 인코더-디코더 구조가 아닌 단일 인코더 형태로 통합하여 네트워크를 단순화해 계산 비용을 줄일 수 있다.

[0296]

한편, 본 발명은 다양한 통신 단말기로 구현되는 동작을 수행하기 위한 프로그램 명령을 포함하는 컴퓨터에서

판독 가능한 매체를 포함할 수 있다. 예를 들어, 컴퓨터에서 판독 가능한 매체는, 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media) 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치를 포함할 수 있다.

[0298] 이와 같은 컴퓨터에서 판독 가능한 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 이때, 컴퓨터에서 판독 가능한 매체에 기록되는 프로그램 명령은 본 발명을 구현하기 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 예를 들어, 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해 실행될 수 있는 고급 언어 코드를 포함할 수 있다.

[0300] 전술한 본 발명의 설명은 예시를 위한 것이며, 본 발명이 속하는 기술분야의 통상의 지식을 가진 자는 본 발명의 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다. 예를 들어, 단일형으로 설명된 각 구성요소는 분산되어 실시될 수도 있으며, 마찬가지로 분산된 것으로 설명된 구성 요소들도 결합된 형태로 실시될 수 있다.

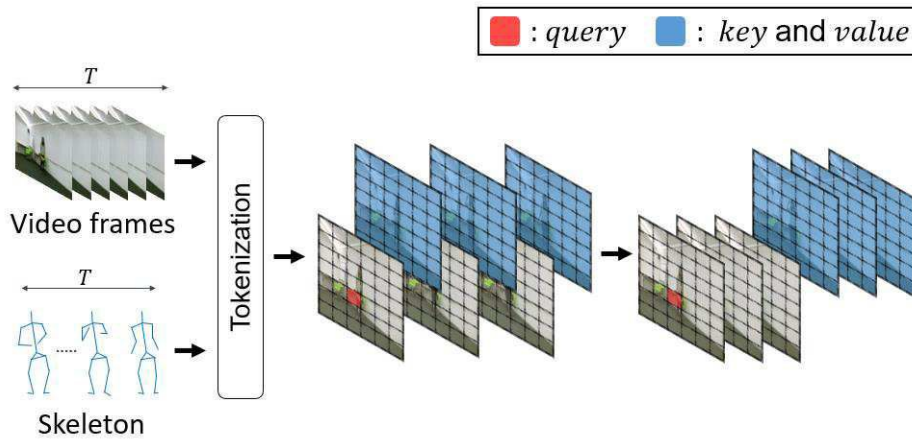
[0302] 본 발명의 범위는 상기 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

부호의 설명

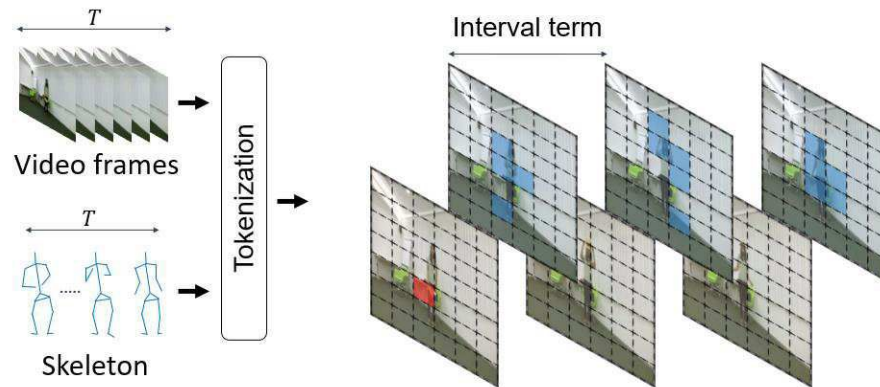
[0303] 10: 인간 행동 인식 시스템
 20: STAR++ 트랜스포머
 100: 데이터 처리부
 200: 행동 인식부
 210: D3D 토큰 선택 모듈
 220: 전체 주의 모듈
 230: 간격 주의 모듈
 S10: D3D 토큰 선택 모듈이 입력 토큰에서 키 및 값을 생성하는 단계
 S20: 전체 주의 모듈이 전체 시공간 주의 메커니즘을 수행하는 단계
 S30: 전체 주의 모듈의 출력을 시간 차원을 기준으로 미리 설정된 간격 항에 따라 나누어 두 토큰 그룹을 생성하는 단계
 S40: D3D 토큰 선택 모듈이 생성된 두 토큰 그룹에서 키 및 값을 생성하는 단계
 S50: 간격 주의 모듈이 두 토큰 그룹 간의 교차 주의 메커니즘을 수행하는 단계
 S100: 입력 비디오와 골격 데이터로부터 입력 토큰을 구성하는 단계
 S200: 입력 토큰을 STAR++ 트랜스포머에 입력해 최종 행동 레이블을 추정하는 단계

도면

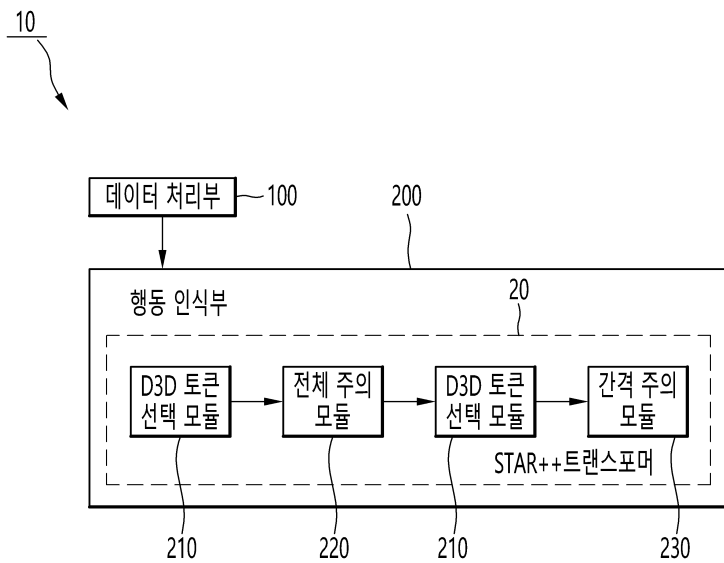
도면1



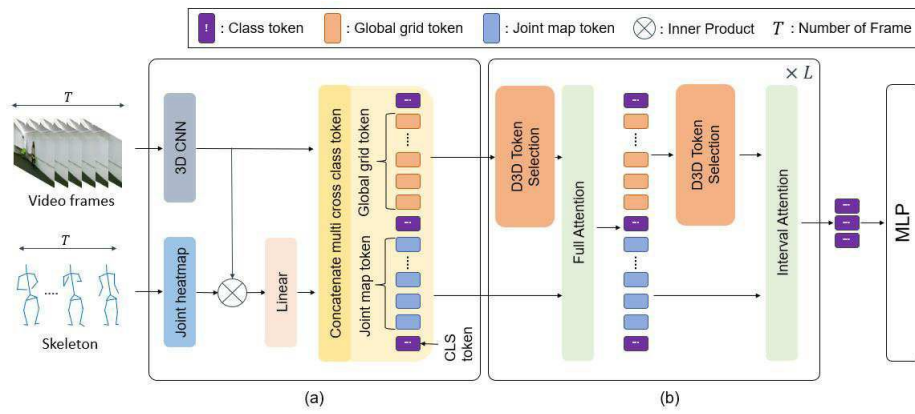
도면2



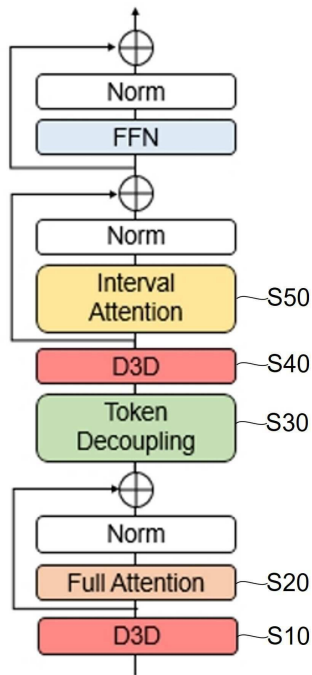
도면3



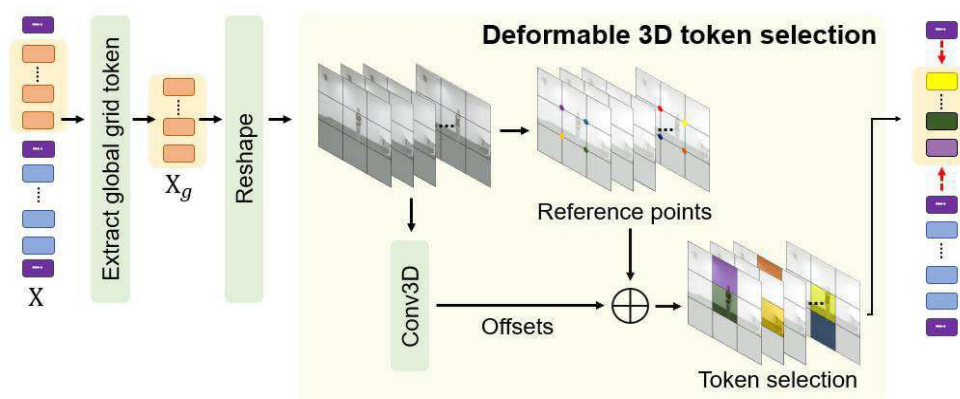
도면4



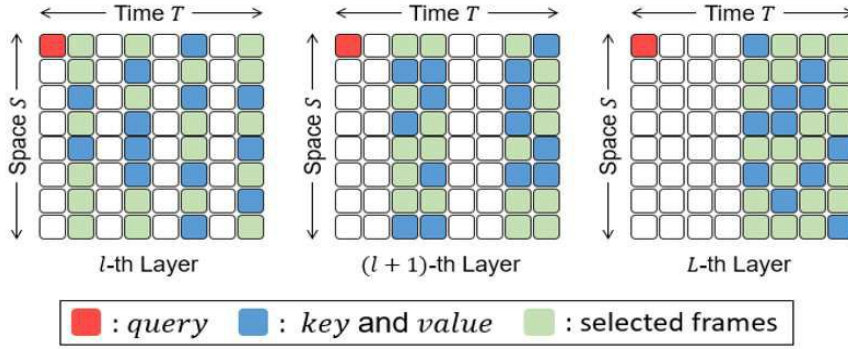
도면5



도면6



도면7



도면8

Algorithm 1 Pseudo code for the process of IAttn

```

# z: input tokens
# i: interval term of current layer
# t: number of frames
# p: token projection

s1 = [] # frame index for first token set
s2 = [] # frame index for second token set
# frame index including interval term
index = [range(0, t, i)]

# frame index selection for decoupling
for y in range(0, (t/i), 2):
    s1 += [range(index[y], index[y+1], 1)]
    s2 += [range(index[y+1], index[y+2], 1)]

o1 = Cross Attn(z, s1, s2)
o2 = Cross Attn(z, s2, s1)

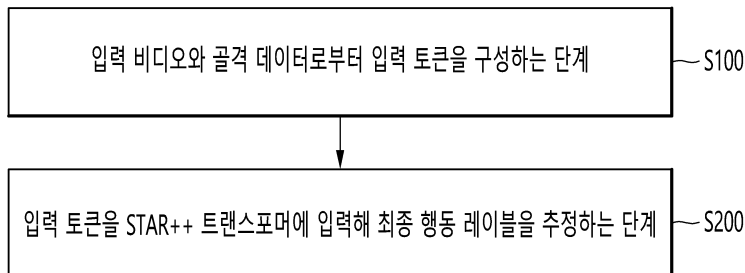
o = concat[o1, o2] # final output

def Cross Attn(z, iq, ikv):
    q = p(z[:, iq]) # extract token for query
    skv = z[:, ikv] # extract token for key and value
    k, v = D3D(skv) # deformable 3D token selection

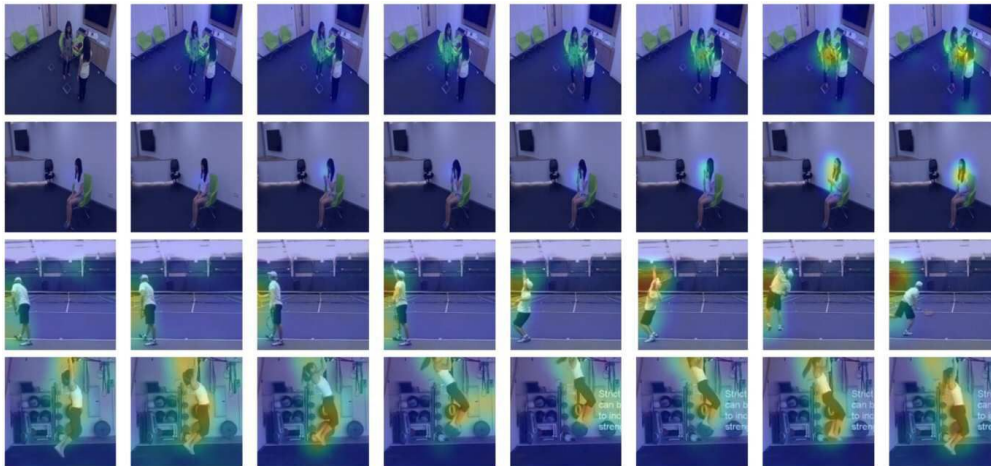
    o = z + MHA(q, k, v) # multi-head cross
                        # attention
    o = o + FFN(o) # output
    return o

```

도면9



도면10



도면11

