(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0111050
(43) 공개일자 2024년07월16일

(51) 국제특허분류(Int. Cl.)
G06V 40/20 (2022.01) *G06N 3/045* (2023.01)
G06N 3/0464 (2023.01) *G06N 3/048* (2023.01)
G06V 10/764 (2022.01) *G06V 10/82* (2022.01)
(52) CPC특허분류
G06V 40/23 (2022.01)
G06N 3/045 (2023.01)
(21) 출원번호 10-2023-0002662
(22) 출원일자 2023년01월09일
심사청구일자 2023년01월09일

(71) 출원인
계명대학교 산학협력단
대구광역시 달서구 달구벌대로 1095, 계명대학교 산학협력관 201호(신당동)
(72) 발명자
고병철
대구광역시 수성구 범어천로 107 범어 STX KAN, 101동 1002호
김선빈
대구광역시 달서구 비슬로 2724
(뒷면에 계속)
(74) 대리인
김건우

전체 청구항 수 : 총 16 항

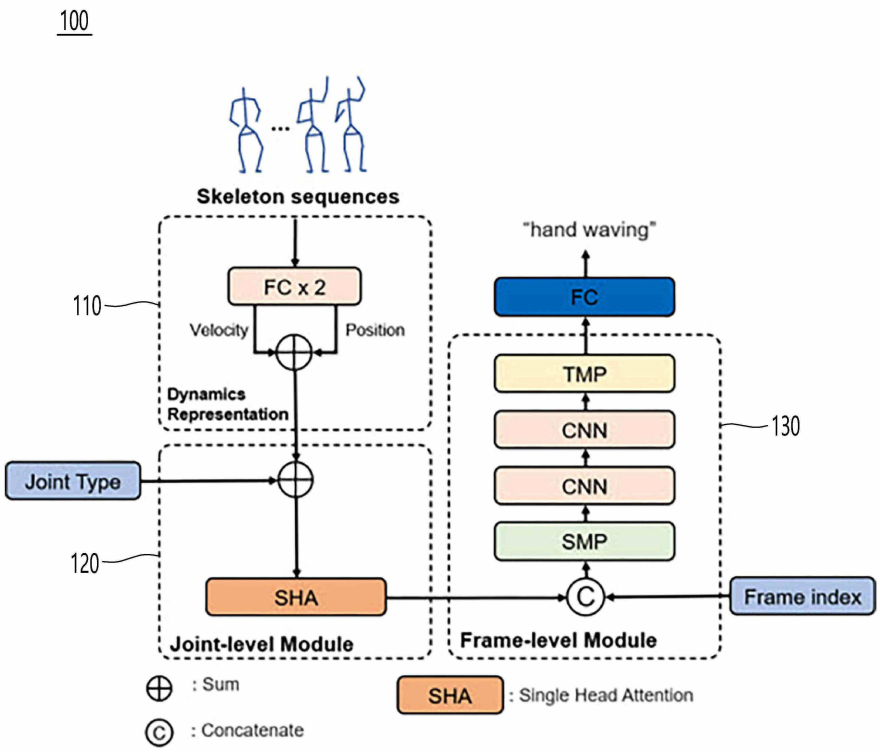
(54) 발명의 명칭 **싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템 및 방법**

(57) 요 약

본 발명은 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템에 관한 것으로서, 보다 구체적으로는 골격 시퀀스를 입력받아 의미-유도 신경망(Semantic-Guided Neural Networks, SGN)을 이용해 사람 행동을 인식하는 사람 행동 인식 시스템으로서, 상기 골격 시퀀스에서 추출된 입력 특징에 싱글 헤드 어텐션(Single

(뒷면에 계속)

대 표 도 - 도2



Head Attention, SHA)을 적용해 동일한 프레임에서 관절의 상관관계를 나타내는 조인트-레벨 출력 특징을 생성하는 조인트-레벨 모듈을 포함하는 것을 그 구성상의 특징으로 한다.

또한, 본 발명은 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 방법에 관한 것으로서, 보다 구체적으로는 골격 시퀀스를 입력받아 의미-유도 신경망을 이용해 사람 행동을 인식하는, 컴퓨터에서 수행되는 사람 행동 인식 방법으로서, 상기 골격 시퀀스에서 추출된 입력 특징에 싱글 헤드 어텐션(SHA)을 적용해 동일한 프레임에서 관절의 상관관계를 나타내는 조인트-레벨 출력 특징을 생성하는 조인트-레벨 모듈을 이용해, 상기 골격 시퀀스로부터 최종 행동 분류를 출력하는 것을 그 구성상의 특징으로 한다.

본 발명에서 제안하고 있는 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템 및 방법에 따르면, 골격 시퀀스를 입력받아 의미-유도 신경망(SGN)을 이용해 사람 행동을 인식하되, 골격 시퀀스에서 추출된 입력 특징에 싱글 헤드 어텐션(SHA)을 적용해 동일한 프레임에서 관절의 상관관계를 나타내는 조인트-레벨 출력 특징을 생성하는 조인트-레벨 모듈을 포함함으로써, 골격 관절의 공간적 특징 추출을 위해 GCN 대신 단일 SHA를 적용해 매개변수의 개수 및 연산 수를 대폭 줄이면서도 높은 정확도를 달성할 수 있다.

(52) CPC특허분류
G06N 3/0464 (2023.01)
G06N 3/048 (2023.01)
G06V 10/764 (2023.08)
G06V 10/82 (2022.01)
G06T 2207/20044 (2013.01)
G06T 2207/20084 (2013.01)
G06T 2207/30221 (2013.01)

(72) 발명자
정찬혁
대구광역시 달서구 월곡로 320

김병일
대구광역시 달서구 서당로3길 22

이 발명을 지원한 국가연구개발사업

과제고유번호	1345354368
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	2022.06.01 ~ 2023.02.28

명 세 서

청구범위

청구항 1

골격 시퀀스를 입력받아 의미-유도 신경망(Semantic-Guided Neural Networks, SGN)을 이용해 사람 행동을 인식하는 사람 행동 인식 시스템(100)으로서,

상기 골격 시퀀스에서 추출된 입력 특징에 싱글 헤드 어텐션(Single Head Attention, SHA)을 적용해 동일한 프레임에서 관절의 상관관계를 나타내는 조인트-레벨 출력 특징을 생성하는 조인트-레벨 모듈(120)을 포함하는 것을 특징으로 하는, 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템(100).

청구항 2

제1항에 있어서,

상기 골격 시퀀스로부터 관절의 위치 및 속도 정보를 융합해 입력 특징을 생성하는 동적 표현 모듈(110); 및

상기 조인트-레벨 모듈(120)의 조인트-레벨 출력 특징과 프레임 인덱스를 연결해 프레임-레벨 입력을 구성하고 프레임 간의 상관관계를 이용해 최종 행동 분류를 출력하는 프레임-레벨 모듈(130)을 더 포함하는 것을 특징으로 하는, 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템(100).

청구항 3

제2항에 있어서, 상기 조인트-레벨 모듈(120)은,

상기 동적 표현 모듈(110)에서 생성한 입력 특징에 관절 유형을 결합한 입력 벡터에 싱글 헤드 어텐션(SHA)을 적용해 상기 조인트-레벨 출력 특징을 생성하는 것을 특징으로 하는, 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템(100).

청구항 4

제2항에 있어서, 상기 조인트-레벨 모듈(120)은,

상기 입력 벡터를 서로 다른 FC(Fully Connected layer)를 통해 쿼리(Q) 및 키(K)로 임베딩한 다음, 상기 쿼리(Q)와 키(K)를 이용해 어텐션을 생성하고, 생성된 어텐션을 값(V)과 결합해 최종 어텐션을 생성하며, 상기 최종 어텐션을 고려하는 상기 조인트-레벨 출력 특징을 생성하는 것을 특징으로 하는, 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템(100).

청구항 5

제4항에 있어서, 상기 조인트-레벨 모듈(120)은,

4개의 단층 FC, 활성화 함수 및 잔여 네트워크(residual network)를 포함해 구성되는 것을 특징으로 하는, 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템(100).

청구항 6

제5항에 있어서, 상기 활성화 함수는,

Leaky ReLU인 것을 특징으로 하는, 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템(100).

청구항 7

제2항에 있어서, 상기 프레임-레벨 모듈(130)은,

두 개의 CNN(Convolutional Neural Network)을 포함하는 것을 특징으로 하는, 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템(100).

청구항 8

제7항에 있어서, 상기 프레임-레벨 모듈(130)은,

상기 두 개의 CNN 이전에 공간 최대 풀링(Spatial Max-Pooling layer, SMP)을 포함하고, 상기 두 개의 CNN 이후에 시간 최대 풀링(Temporal Max-Pooling layer, TMP) 및 FC를 더 포함하는 것을 특징으로 하는, 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템(100).

청구항 9

골격 시퀀스를 입력받아 의미-유도 신경망을 이용해 사람 행동을 인식하는, 컴퓨터에서 수행되는 사람 행동 인식 방법으로서,

상기 골격 시퀀스에서 추출된 입력 특징에 싱글 헤드 어텐션(SHA)을 적용해 동일한 프레임에서 관절의 상관관계를 나타내는 조인트-레벨 출력 특징을 생성하는 조인트-레벨 모듈(120)을 이용해, 상기 골격 시퀀스로부터 최종 행동 분류를 출력하는 것을 특징으로 하는, 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 방법.

청구항 10

제9항에 있어서,

(1) 동적 표현 모듈(110)에서, 상기 골격 시퀀스로부터 관절의 위치 및 속도 정보를 융합해 입력 특징을 생성하는 단계; 및

(2) 상기 조인트-레벨 모듈(120)에서, 상기 단계 (1)의 입력 특징에 싱글 헤드 어텐션(SHA)을 적용해 조인트-레벨 출력 특징을 생성하는 단계; 및

(3) 프레임-레벨 모듈(130)에서, 상기 조인트-레벨 출력 특징과 프레임 인덱스를 연결해 프레임-레벨 입력을 구성하고 프레임 간의 상관관계를 이용해 상기 최종 행동 분류를 출력하는 단계를 포함하는 것을 특징으로 하는, 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 방법.

청구항 11

제10항에 있어서, 상기 단계 (2)에서는,

상기 조인트-레벨 모듈(120)은, 상기 동적 표현 모듈(110)에서 생성한 입력 특징에 관절 유형을 결합한 입력 벡터에 싱글 헤드 어텐션(SHA)을 적용해 조인트-레벨 출력 특징을 생성하는 것을 특징으로 하는, 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 방법.

청구항 12

제10항에 있어서, 상기 단계 (2)에서는,

상기 조인트-레벨 모듈(120)은, 상기 입력 벡터를 서로 다른 FC를 통해 쿼리(Q) 및 키(K)로 임베딩한 다음, 상기 쿼리(Q)와 키(K)를 이용해 어텐션을 생성하고, 생성된 어텐션을 값(V)과 결합해 최종 어텐션을 생성하며, 상기 최종 어텐션을 고려하는 조인트-레벨 출력 특징을 생성하는 것을 특징으로 하는, 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 방법.

청구항 13

제12항에 있어서, 상기 조인트-레벨 모듈(120)은,

4개의 단층 FC, 활성화 함수 및 잔여 네트워크를 포함해 구성되는 것을 특징으로 하는, 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 방법.

청구항 14

제13항에 있어서, 상기 활성화 함수는,

Leaky ReLU인 것을 특징으로 하는, 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식방법.

청구항 15

제10항에 있어서, 상기 단계 (3)에서는,

두 개의 CNN을 이용해 상기 최종 행동 분류를 출력하는 것을 특징으로 하는, 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 방법.

청구항 16

제15항에 있어서, 상기 단계 (3)에서는,

상기 조인트-레벨 출력 특징과 프레임 인덱스를 연결해 프레임-레벨 입력을 구성하고, 구성된 프레임-레벨 입력으로부터 공간 최대 풀링(SMP), 상기 두 개의 CNN, 시간 최대 풀링(TMP) 및 FC를 통해 상기 최종 행동 분류를 출력하는 것을 특징으로 하는, 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 사람 행동 인식 시스템 및 방법에 관한 것으로서, 보다 구체적으로는 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템 및 방법에 관한 것이다.

배 경 기 술

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

[0004] 컴퓨터 비전 및 기계 학습 분야에서 인간 행동 인식은 인간의 의도를 인식하고 이러한 의도에 맞는 적절한 서비스를 제공하는 것을 목표로 한다. 따라서 행동 인식은 인간-기계 인터페이스(HMI), 감시 시스템, 게임, 증강현실, 가상 현실 등 다양한 분야에 적용되고 있다.

[0006] 인간의 행동은 크게 독립적인 행동과 상호작용으로 나눌 수 있다. 독립적인 행동은 서로 다른 신체 부위가 지속적으로 움직이는 것을 의미하며, 상호 작용은 두 개 이상의 물체가 행동을 제공하고 받는 것을 의미한다. 독

립적인 행동은 행동과 활동으로 더 나눌 수 있다. 행동(action)은 20-30 프레임의 비교적 짧은 기간 동안 발생하는 동작을 말한다. 행동의 예로는 '달리기', '걷기', '줍기' 등이 있다. 활동(activity)은 비교적 오랜 시간 동안 지속적으로 발생하고 다양한 행동이 동시에 결합되는 행동의 집합이다. 대표적인 활동으로는 '자동차 문 열고 운전하기', '주머니에서 스마트폰 꺼내 전화 걸기' 등이 있다. 대조적으로, 상호작용은 '악수', '싸움', '포옹'과 같이 인간과 사물(인간 또는 사물) 사이에 주고받는 행동을 말한다.

[0008] 본 발명은 짧은 시간 내에 발생하는 실시간 인간 행동에 관한 것이다. 따라서 이하에서는 활동 인식보다 행동 인식에 초점을 맞추어 설명한다.

[0010] 행동 인식은 카메라를 이용한 센서 및 시각 기반 방법을 포함한다. 센서 기반 방법은 센서를 구매하는 추가 비용과 센서 착용에 대한 거부감을 유발한다. 따라서, 비전 기반 행동 인식이 주로 연구되어 왔다. 비전 기반 행동 인식은 비디오 프레임의 RGB만을 이용하는 방식, 비디오의 인간 골격(포즈)만을 이용하는 방식, 비디오 프레임과 골격을 함께 이용하는 방식으로 크게 나눌 수 있다. 비디오 프레임의 RGB 기반 동작 인식의 장점으로, 시각적 정보만 사용되기 때문에 모델 구조가 간단하다. 그러나 배경 소음(예: 카메라 각도, 차체 크기 및 복잡한 배경) 또는 조도의 변화에 민감하게 반응하기 때문에 성능이 저하될 수 있다. 반대로, 골격 기반 행동 인식은 골격 구조를 특징으로 전환함으로써 신체의 구조적 변화를 수반한다. 이 접근 방식은 다양한 실제 소음의 영향을 덜 받기 때문에 높은 성능을 달성할 수 있다. 그러나 골격 데이터를 추출하기 위해 별도의 모델이 필요하다는 단점이 있다. RGB 비디오 프레임과 골격을 함께 사용하면 RGBe 정보를 사용하여 동일한 포즈의 모호한 행동(예: '테니스'와 '배드민턴')을 구별할 수 있다. 그러나 두 세트의 교차 모달 정보를 결합하는 것은 어렵고, 각 인식 모듈에 개별 특징을 적용할 경우 실시간 행동 인식에 한계가 있다. 따라서 실시간으로 동작을 인식할 수 있는 빠른 학습 속도를 가진 행동 인식 알고리즘이 필요하다.

[0012] 도 1은 골격 시퀀스를 이용한 사람 행동 인식을 설명하기 위해 도시한 도면이다. 골격 기반 행동 인식에서는 스켈레톤 시퀀스에서 시공간 특징을 추출하여 그래프 컨볼루션 네트워크(graph convolutional network, GCN)와 결합하는 방법이 널리 사용되고 있다. 그 중에서도 의미-유도 신경망(Semantics-Guided Neural Networks, SGN)은 GCN을 사용하여 공간적 및 시간적 특징을 계층적으로 학습하는 빠르고 유망한 행동 인식 알고리듬이다(Zhang, P.; Lan, C.; Zeng, W.; Xing, J.; Xue, J.; Zheng, N. Semantics-Guided Neural Networks for Efficient Skeleton-Based Human Action Recognition. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), virtual, 14-19 June 2020; pp. 1112-1121.). 그러나 SGN은 구조적 특성으로 인해 로컬 특징 학습보다는 글로벌 특징 학습에 중점을 두기 때문에 인접 노드 간의 의존성이 중요한 동작 인식에 한계가 있다. 또한, SGN은 경량 모델이기는 하지만, 실시간 동작 인식을 위해 사용하기에는 상당한 계산 복잡도를 가지고 있다는 점에서 여전히 문제가 있다.

[0014] 전술한 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

[0015] 본 발명은 기존에 제안된 방법들의 상기와 같은 문제점들을 해결하기 위해 제안된 것으로서, 골격 시퀀스를 입력받아 의미-유도 신경망(SGN)을 이용해 사람 행동을 인식하되, 골격 시퀀스에서 추출된 입력 특징에 싱글 헤드 어텐션(Single Head Attention, SHA)을 적용해 동일한 프레임에서 관절의 상관관계를 나타내는 조인트-레벨 출력 특징을 생성하는 조인트-레벨 모듈을 포함함으로써, 골격 관절의 공간적 특징 추출을 위해 GCN 대신 단일 SHA를 적용해 매개변수의 개수 및 연산 수를 대폭 줄이면서도 높은 정확도를 달성할 수 있는, 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템 및 방법을 제공하는 것을 그 목적으로 한다.

[0017] 다만, 본 발명이 이루고자 하는 기술적 과제는 상기한 바와 같은 기술적 과제로 한정되지 않으며, 또 다른 기술

적 과제들이 존재할 수 있고, 명시적으로 언급하지 않더라도 과제의 해결수단이나 실시 형태로부터 파악될 수 있는 목적이나 효과도 이에 포함됨은 물론이다.

과제의 해결 수단

[0018] 상기한 목적을 달성하기 위한 본 발명의 특징에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템은,

[0019] 골격 시퀀스를 입력받아 의미-유도 신경망(SGN)을 이용해 사람 행동을 인식하는 사람 행동 인식 시스템으로서,

[0020] 상기 골격 시퀀스에서 추출된 입력 특징에 싱글 헤드 어텐션(SHA)을 적용해 동일한 프레임에서 관절의 상관관계를 나타내는 조인트-레벨 출력 특징을 생성하는 조인트-레벨 모듈을 포함하는 것을 그 구성상의 특징으로 한다.

[0022] 바람직하게는,

[0023] 상기 골격 시퀀스로부터 관절의 위치 및 속도 정보를 융합해 입력 특징을 생성하는 동적 표현 모듈; 및

[0024] 상기 조인트-레벨 모듈의 조인트-레벨 출력 특징과 프레임 인덱스를 연결해 프레임-레벨 입력을 구성하고 프레임 간의 상관관계를 이용해 최종 행동 분류를 출력하는 프레임-레벨 모듈을 더 포함할 수 있다.

[0026] 더욱 바람직하게는, 상기 조인트-레벨 모듈은,

[0027] 상기 동적 표현 모듈에서 생성한 입력 특징에 관절 유형을 결합한 입력 벡터에 싱글 헤드 어텐션(SHA)을 적용해 상기 조인트-레벨 출력 특징을 생성할 수 있다.

[0029] 더욱 바람직하게는, 상기 조인트-레벨 모듈은,

[0030] 상기 입력 벡터를 서로 다른 FC(Fully Connected layer)를 통해 쿼리(Q) 및 키(K)로 임베딩한 다음, 상기 쿼리(Q)와 키(K)를 이용해 어텐션을 생성하고, 생성된 어텐션을 값(V)과 결합해 최종 어텐션을 생성하며, 상기 최종 어텐션을 고려하는 상기 조인트-레벨 출력 특징을 생성할 수 있다.

[0032] 더더욱 바람직하게는, 상기 조인트-레벨 모듈은,

[0033] 4개의 단층 FC, 활성화 함수 및 잔여 네트워크(residual network)를 포함해 구성될 수 있다.

[0035] 더욱더 바람직하게는, 상기 활성화 함수는, Leaky ReLU일 수 있다.

[0037] 더욱 바람직하게는, 상기 프레임-레벨 모듈은,

[0038] 두 개의 CNN(Convolutional Neural Network)을 포함할 수 있다.

[0040] 더더욱 바람직하게는, 상기 프레임-레벨 모듈은,

[0041] 상기 두 개의 CNN 이전에 공간 최대 풀링(Spatial Max-Pooling layer, SMP)을 포함하고, 상기 두 개의 CNN 이후에 시간 최대 풀링(Temporal Max-Pooling layer, TMP) 및 FC를 더 포함할 수 있다.

[0043] 상기한 목적을 달성하기 위한 본 발명의 특징에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 방법은,

[0044] 골격 시퀀스를 입력받아 의미-유도 신경망을 이용해 사람 행동을 인식하는, 컴퓨터에서 수행되는 사람 행동 인

식 방법으로서,

[0045] 상기 골격 시퀀스에서 추출된 입력 특징에 싱글 헤드 어텐션(SHA)을 적용해 동일한 프레임에서 관절의 상관관계를 나타내는 조인트-레벨 출력 특징을 생성하는 조인트-레벨 모듈을 이용해, 상기 골격 시퀀스로부터 최종 행동 분류를 출력하는 것을 그 구성상의 특징으로 한다.

[0047] 바람직하게는,

[0048] (1) 동적 표현 모듈에서, 상기 골격 시퀀스로부터 관절의 위치 및 속도 정보를 융합해 입력 특징을 생성하는 단계; 및

[0049] (2) 상기 조인트-레벨 모듈에서, 상기 단계 (1)의 입력 특징에 싱글 헤드 어텐션(SHA)을 적용해 조인트-레벨 출력 특징을 생성하는 단계; 및

[0050] (3) 프레임-레벨 모듈에서, 상기 조인트-레벨 출력 특징과 프레임 인덱스를 연결해 프레임-레벨 입력을 구성하고 프레임 간의 상관관계를 이용해 상기 최종 행동 분류를 출력하는 단계를 포함할 수 있다.

[0052] 더욱 바람직하게는, 상기 단계 (2)에서는,

[0053] 상기 조인트-레벨 모듈은, 상기 동적 표현 모듈에서 생성한 입력 특징에 관절 유형을 결합한 입력 벡터에 싱글 헤드 어텐션(SHA)을 적용해 조인트-레벨 출력 특징을 생성할 수 있다.

[0055] 더욱 바람직하게는, 상기 단계 (2)에서는,

[0056] 상기 조인트-레벨 모듈은, 상기 입력 벡터를 서로 다른 FC를 통해 쿼리(Q) 및 키(K)로 임베딩한 다음, 상기 쿼리(Q)와 키(K)를 이용해 어텐션을 생성하고, 생성된 어텐션을 값(V)과 결합해 최종 어텐션을 생성하며, 상기 최종 어텐션을 고려하는 조인트-레벨 출력 특징을 생성할 수 있다.

[0058] 더더욱 바람직하게는, 상기 조인트-레벨 모듈은,

[0059] 4개의 단층 FC, 활성화 함수 및 잔여 네트워크를 포함해 구성될 수 있다.

[0061] 더욱더 바람직하게는, 상기 활성화 함수는, Leaky ReLU일 수 있다.

[0063] 더욱 바람직하게는, 상기 단계 (3)에서는,

[0064] 두 개의 CNN을 이용해 상기 최종 행동 분류를 출력할 수 있다.

[0066] 더더욱 바람직하게는, 상기 단계 (3)에서는,

[0067] 상기 조인트-레벨 출력 특징과 프레임 인덱스를 연결해 프레임-레벨 입력을 구성하고, 구성된 프레임-레벨 입력으로부터 공간 최대 풀링(SMP), 상기 두 개의 CNN, 시간 최대 풀링(TMP) 및 FC를 통해 상기 최종 행동 분류를 출력할 수 있다.

발명의 효과

[0068] 본 발명에서 제안하고 있는 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템 및 방법에 따르면, 골격 시퀀스를 입력받아 의미-유도 신경망(SGN)을 이용해 사람 행동을 인식하되, 골격 시퀀스에서 추출된 입력 특징에 싱글 헤드 어텐션(SHA)을 적용해 동일한 프레임에서 관절의 상관관계를 나타내는 조인트-레벨 출력 특징을 생성하는 조인트-레벨 모듈을 포함함으로써, 골격 관절의 공간적 특징 추출을 위해 GCN 대신 단일 SHA를 적용해 매개변수의 개수 및 연산 수를 대폭 줄이면서도 높은 정확도를 달성할 수 있다.

[0070] 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

[0071] 도 1은 골격 시퀀스를 이용한 사람 행동 인식을 설명하기 위해 도시한 도면.

도 2는 본 발명의 일실시예에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템의 구조를 나타낸 도면.

도 3은 본 발명의 일실시예에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템의 구성을 도시한 도면.

도 4는 GCN 기반의 의미-유도 신경망(SGN)에서, 조인트 레벨 모듈의 구성을 도시한 도면.

도 5는 본 발명의 일실시예에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템에서, 조인트-레벨 모듈의 구성을 도시한 도면.

도 6은 본 발명의 일실시예에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 방법의 흐름을 도시한 도면.

도 7은 NTU60 데이터셋에서 본 발명의 일실시예에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템 및 방법과 다른 최신 기술의 정확도, 매개변수 개수, 계산 복잡도를 나타낸 도면.

도 8은 NTU120 데이터셋에서 본 발명의 일실시예에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템 및 방법과 다른 최신 기술의 정확도, 매개변수 개수, 계산 복잡도를 나타낸 도면.

도 9는 NW-UCLA 데이터셋의 세부 행동에 대한 의미-유도 신경망(기본 SGN)의 혼동 행렬(confusion matrix)을 나타낸 도면.

도 10은 NW-UCLA 데이터셋에서 본 발명의 일실시예에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템 및 방법(SGN-SHA)의 혼동 행렬을 나타낸 도면.

발명을 실시하기 위한 구체적인 내용

[0072] 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.

[0074] 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 이해되어야 한다.

[0076] 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.

[0078] 또한, 본 발명의 각 실시예에 포함된 각 구성, 과정, 공정 또는 방법 등은 기술적으로 상호간 모순되지 않는 범위 내에서 공유될 수 있다.

[0080] 본 발명은 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템 및 방법에 관한 것으로서, 본 발명의 특징에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템 및 방법은, 메모리 및 프로세서를 포함한 하드웨어에서 기록되는 소프트웨어로 구성될 수 있다. 예를 들어, 본 발명의 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 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.

[0082] 도 2는 본 발명의 일실시예에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템(100)의 구조를 나타낸 도면이고, 도 3은 본 발명의 일실시예에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템(100)의 구성을 도시한 도면이다. 도 2 및 도 3에 도시된 바와 같이, 본 발명의 일실시예에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템(100)은, 골격(skeleton) 시퀀스를 입력받아 의미-유도 신경망(SGN)을 이용해 사람 행동을 인식하는 사람 행동 인식 시스템(100)으로서, 골격 시퀀스에서 추출된 입력 특징에 싱글 헤드 어텐션(SHA)을 적용해 동일한 프레임에서 관절의 상관관계를 나타내는 조인트-레벨 출력 특징을 생성하는 조인트-레벨 모듈(120)을 포함하여 구성될 수 있으며, 골격 시퀀스로부터 관절의 위치 및 속도 정보를 융합해 입력 특징을 생성하는 동적 표현 모듈(110); 및 조인트-레벨 모듈(120)의 조인트-레벨 출력 특징과 프레임 인덱스를 연결해 프레임-레벨 입력을 구성하고 프레임 간의 상관관계를 이용해 최종 행동 분류를 출력하는 프레임-레벨 모듈(130)을 더 포함하여 구성될 수 있다.

[0084] 즉, 본 발명의 일실시예에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템(100)은, 실시간 행동인식이 가능하도록 SGN을 행동인식을 위한 기본 구조로 사용하되, 기존의 SGN이 사용하는 그래프 구조를 SHA 구조로 대체하였다(SGN-SHA). 이러한 방식으로 본 발명의 SGN-SHA는 동작 인식 성능을 향상시키고 매개변수의 수와 계산 복잡도를 획기적으로 줄일 수 있다.

[0086] 본 발명의 일실시예에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템(100) 및 방법은 다음과 같은 특징이 있다. 1) 골격 관절의 공간적 특징 추출을 위한 조인트-레벨 모듈(120)을 GCN 대신 단일 SHA로 대체함으로써 매개변수의 개수를 절반으로 줄이고, 기본 SGN에 비해 연산 수를 2.6배 감소시킬 수 있다. 2) GCN과 달리 관절의 공간적 특성을 임베딩하기 위해 쿼리(query, Q), 키(key, K) 및 값(value, V)으로 구성된 SHA는 관절 간의 특성을 강조할 수 있다. 3) 제안된 조인트-레벨 모듈(120)은 입력 특징의 손실을 줄이고 잔여 네트워크를 SHA 출력과 결합하여 임베딩 특징을 강조할 수 있다. 4) 다양한 실험을 통해 제안된 SGN-SHA 모델이 다른 모델에 비해 정확도를 낮추지 않고 매개변수 수와 계산 복잡도를 크게 줄일 수 있음을 증명했다.

[0088] 이하에서는, 본 발명의 일실시예에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템(100) 및 방법에 대해 상세히 설명하기 전에, 관련된 연구 중 골격 시퀀스를 이용한 행동 인식 접근법의 세부적인 내용을 설명하도록 한다.

[0090] Yan 외.는 비디오 프레임에서 추출된 골격 시퀀스에서 공간 및 시간 패턴을 자동으로 학습하여 행동 인식의 성능을 향상시킬 수 있는 공간-시간 그래프 컨볼루션 네트워크(ST-GCN)를 제안했다(Yan, S.; Xiong, Y.; Lin, D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In Proceedings of the 2018 Association for the Advancement of Artificial Intelligence (AAAI), New Orleans, LA, USA, 2-7 February 2018; pp. 7444-7452.). ST-GCN은 시공간 골격 기능에 대해 더 큰 표현력과 일반화 능력을 갖추고 있다. Takkar와 Narayan은 골격 그래프를 4개의 하위 그래프로 나누고 부분 기반 GCN을 사용한 행동 인식 모델을 제안했다(Thakkar, K.; Narayanan, P.J. Part-based Graph Convolutional Network for Action Recognition. In Proceedings of the 29th British Machine Vision Conference (BMVC), Newcastle, UK, 3-6 September 2018; pp. 1-13.). 그들은 또한 이 모델이 전체 골격 그래프를 사용하는 모델보다 인식 성능이 향상

되었음을 보여주었다. Li 외.는 보다 풍부한 관절 관계를 표현하기 위해 ST-GCN에 행동-링크 및 구조-링크 구조를 추가하는 행동 구조 GCN을 제안했다(Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Actional-Structural Graph Convolutional Networks for Skeleton-based Action Recognition. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16-20 June 2019; pp. 3595-3603.). 행동 링크는 인코더-디코더 구조로 구성되며, 이를 통해 근처에서 연결된 관절 사이의 관계와 더 먼 거리의 관절 사이의 관계를 찾을 수 있다. 그러나 ST-GCN 기반 방법은 정확도에 비해 많은 수의 연산이 필요하다는 단점이 있다.

[0092] 이 문제를 해결하기 위해 Cheng 외.는 Shift Graph Convolutional Network를 제안했다. 이 방법은 무거운 일반 그래프 컨볼루션 대신 시공간 이동 그래프 연산과 가벼운 point-wise 컨볼루션으로 구성되어 행동 인식 성능이 향상된다(Cheng, K.; Zhang, Y.; He, X.; Chen, W.; Cheng, J.; Lu, H. Skeleton-Based Action Recognition with Shift Graph Convolutional Network. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), virtual, 14-19 June 2020; pp. 183-192.). Zhang 외.는 효율적인 SGN을 제안했다(Zhang, P.; Lan, C.; Zeng, W.; Xing, J.; Xue, J.; Zheng, N. Semantics-Guided Neural Networks for Efficient Skeleton-Based Human Action Recognition. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), virtual, 14-19 June 2020; pp. 1112-1121.). 이 방법에서는 특징 표현 능력을 향상시키기 위해 높은 수준의 관절 의미론(예: 관절 유형) 또는 프레임 인덱스(프레임 수)가 사용되었다. Cai 외.는 JOLO-GCN이라는 이중 스트림 GCN에서 인간 포즈 골격과 경량 관절 중심 정보를 공동으로 사용하기 위한 행동 인식 프레임워크를 제안했다(Cai, J.; Jiang, N.; Han, X.; Jia, K.; Lu, J. JOLO-GCN: Mining Joint-Centered Light-Weight Information for Skeleton-Based Action Recognition. In Proceedings of the 2021 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), virtual, 5-9 January 2021; pp. 2735-2744.). 단일 골격 기반 방법과 비교하여, 이 하이브리드 접근 방식은 적은 계산 및 메모리 오버헤드를 유지하면서 성능을 효과적으로 향상시키는 것으로 입증되었다. Mazzia 외 연구진은 컨볼루션, 반복 및 주의(attention) 레이어를 결합한 더 단순하지만 자기-주의적 아키텍처인 행동 변환기(transformer)를 제안했다(Mazzia, V.; Angarano, S.; Salvetti, F.; Angelini, F.; Chiaberge, M. Action Transformer: A Self-Attention Model for Short-Time Pose-Based Human Action Recognition. *Pattern Recognit.* 2022, 124, 1-11. doi: 10.1016/j.patcog.2021.108487.). 이 방법은 작은 타임 윈도우 내에서 2D 포즈 표현과 낮은 지연 시간 솔루션을 사용하여 실시간 포즈 기반 행동 인식을 가능하게 하여 행동 인식에 필요한 작업 수를 줄인다.

[0094] 이와 같은 GCN(Graph Convolutional Network, 그래프 컨볼루션 네트워크) 기반 행동 인식은 골격 기반 행동 인식에서 우수한 성능을 보여주지만, 시공간 비디오 시퀀스가 지나치게 길어지면 다른 접근 방식과 마찬가지로 계산 복잡도와 매개변수 수를 증가시키는 단점이 있다.

[0096] 그밖에, RGB와 골격 시퀀스를 함께 이용한 행동 인식 접근법도 있다. 이러한 멀티모달 특징 융합 방법은 비디오 프레임의 골격 시퀀스와 RGB 특징을 융합함으로써 행동 인식 성능을 향상시킬 수 있다. 그러나 이 접근법은 비디오 프레임의 RGB에서 특징을 추출하는 데 사용되는 네트워크와는 별도로, 골격을 추출하기 위한 별도의 네트워크를 필요로 한다. 또한, 두 가지 이기종 특징을 결합하고 행동 인식을 위한 다운스트림 네트워크를 필요로 한다. 따라서 높은 네트워크 복잡성을 위해 효율성을 희생하는 한계가 있다.

[0098] 반면에, 도 2 및 도 3에 도시된 바와 같은 본 발명의 일실시예에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템(100)은, 의미-유도 신경망(SGN)과 싱글 헤드 어텐션(SHA)의 조합을 기반으로 하는 가볍고 빠른 행동 인식 알고리듬으로, 계산 복잡도가 낮고 매개변수가 적은 효율적인 행동 인식 모델이다. 보다 구체적으로, 기본 SGN에서 입력 골격 시퀀스는 조인트 레벨(위치, 3D 좌표, 속도)과 프레임 레벨(관절 유형 및 프레임 인덱스) 모듈로 구분되어 행동 인식에 사용된다. 두 개의 무거운 모듈로 구성되어 있기 때문에 계산 복잡도가 높고 많은 모델 매개변수가 필요하다. 본 발명에서는, 기본 SGN의 조인트 레벨 모듈을 더 적은 작업이 필요한 SHA로 대체함으로써 정확도 수준을 유지하면서 실시간 행동 인식을 할 수 있다. 이하에서는, 도 2

및 도 3을 참조하여 본 발명의 일실시예에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템(100)의 각 구성에 대해 상세히 설명하도록 한다.

[0100] 동적 표현 모듈(110)은, 골격 시퀀스로부터 관절의 위치 및 속도 정보를 융합해 입력 특징을 생성할 수 있다. 먼저, 동적 표현 모듈(110)에서는, 골격 시퀀스에 대해 모든 관절(joint) 세트를 $S = \{X_t^k | t = 1,2,...,T; k = 1,2,...,J\}$ 로 표시할 수 있다. 여기서 X_t^k는 시간 t에서 유형 k의 관절을 나타내고 관절은 3D 위치(x,y,z)로 표시된다. 또한, T는 골격 시퀀스의 프레임 수를 나타내고, J는 단일 프레임에서 인체의 관절 총수를 나타낸다.

[0102] 도 2에 도시된 바와 같이, 동적 표현 모듈(110)에서 관절 위치(position) 및 속도(velocity) 정보는 FC(Full Connected layer)를 사용하여 각각 별도로 변환될 수 있다. 관절 속도는 T 프레임의 관절 X_t^k와 T-1 프레임의 관절 X_{t-1}^k 사이의 거리 차이를 이용할 수 있다. 그 다음, 관절 위치와 속도를 연결해 T × J × C_1의 차원의 특징으로 표현해, 입력 특징을 생성할 수 있다. 여기서 C_1은 관절 표현의 차원이다. 동적 표현 모듈(110)의 출력인 입력 특징은 이하에서 상세히 설명할 조인트-레벨 모듈(120)에 전달되며, 관절 유형(예: 머리 및 발)과 결합해 SHA에 입력되어 동일한 프레임에서 관절의 상관관계를 모델링할 수 있다.

[0104] 조인트-레벨 모듈(120)은, 골격 시퀀스에서 추출된 입력 특징에 싱글 헤드 어텐션(SHA)을 적용해 동일한 프레임에서 관절의 상관관계를 나타내는 조인트-레벨 출력 특징을 생성할 수 있다. 보다 구체적으로, 조인트-레벨 모듈(120)은, 동적 표현 모듈(110)에서 생성한 입력 특징에 관절 유형을 결합한 입력 벡터에 싱글 헤드 어텐션(SHA)을 적용해 조인트-레벨 출력 특징을 생성할 수 있다.

[0106] 기본 SGN은 다른 최첨단(Sate of The Art, SoTA) 방법에 비해 상대적으로 낮은 연산 수로 우수한 행동 인식 정확도를 보여주지만, 조인트 레벨 모듈에서의 GCN의 구조적 한계로 인해 불필요한 작업 및 매개변수를 요구하는 문제가 있다.

[0108] 도 4는 GCN 기반의 의미-유도 신경망(SGN)에서, 조인트 레벨 모듈의 구성을 도시한 도면이다. 도 4에 도시된 바와 같이, GCN을 기반으로 하는 기본 SGN의 조인트 레벨 모듈은, 2개의 FC와 3개의 연속된 GCN으로 구성되므로, 기본 SGN을 이용하면 DR(Dynamics Representation) 모듈의 임베딩 특징은 조인트 레벨 모듈에서 그래프 형태로 변환되어야 한다. 따라서 그래프 변환 과정에서 각 관절의 노드 특징이 잘못 일치할 수 있다. 또한, 각 프레임에 대해 인접 행렬이 유지되어야 하고 노드 특징이 인접 행렬과 함께 3개의 연속된 GCN 계층에 적용되어야 하므로 메모리의 양과 작업 수가 증가한다. 즉, 개별 GCN에 대한 학습 매개변수와 인접 행렬이 필요하며, GCN 작업을 위해서는 많은 수의 연산이 필요하다. 따라서 본 발명에서는 SGN의 연산 횟수와 매개변수를 크게 줄이면서 정확도를 유지하기 위해 SHA를 기반으로 하는 새로운 조인트-레벨 모듈(120)을 제안한다.

[0110] 도 5는 본 발명의 일실시예에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템(100)에서, 조인트-레벨 모듈(120)의 구성을 도시한 도면이다. 도 5에 도시된 바와 같이, 본 발명의 일실시예에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템(100)의 조인트-레벨 모듈(120)은, 트랜스포머의 주의 메커니즘에서 영감을 받은 것으로, 골격 시퀀스에서 추출된 입력 특징에 싱글 헤드 어텐션(SHA)을 적용함으로써, 골격의 전역적 특징뿐만 아니라 인접한 관절 사이의 의존성을 고려하는 국소적 특징도 반영하는 새로운 특징을 구성할 수 있다. 즉, 도 4에 도시된 바와 같은 기본 SGN은 DR의 출력과 관절 유형 벡터가 연결되어 3개의 연속된 GCN을 통해 조인트 레벨 특징을 생성하는 반면에, 도 5에 도시된 바와 같은 본 발명의 조인트-레벨 모듈(120)은 4개의 단층 FC와 잔여 네트워크를 통해 중요한 관절을 강조하는 조인트-레벨 특징을 생성할 수 있다.

[0112] 조인트-레벨 모듈(120)은, 입력 벡터를 서로 다른 FC를 통해 쿼리(Q) 및 키(K)로 임베딩한 다음, 쿼리(Q)와 키(K)를 이용해 어텐션을 생성하고, 생성된 어텐션을 값(V)과 결합해 최종 어텐션을 생성하며, 최종 어텐션을 고려하는 조인트-레벨 출력 특징을 생성할 수 있다. 보다 구체적으로, 도 5에 도시된 바와 같이, 조인트-레벨 모듈(120)은, 4개의 단층 FC, 활성화 함수 및 잔여 네트워크를 포함해 구성될 수 있으며, 이때 활성화 함수는, Leaky ReLU일 수 있다.

[0114] 트랜스포머의 핵심 아이디어인 어텐션 헤드는 쿼리(Q), 키(K), 값(V)의 세 가지 행렬을 사용하여 입력 토큰(특징) 사이의 어텐션을 계산한다. 트랜스포머는 헤드가 입력과 다른 주의(어텐션)를 학습하고 이를 결합하여 강력한 주의 표현을 얻을 수 있도록 멀티 헤드 어텐션(Multi-Head Attention, MHA)을 사용한다. 그러나, 멀티 헤드 중 일부만이 트랜스포머의 주의 성능에 영향을 미치고, 일부 헤드는 불필요한 부분에 주의를 기울인다는 것이 입증되었다(Michel, P.; Levy, O.; Neubig, G. Are Sixteen Heads Really Better than One?. In Proceedings of the 2019 Annual Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 2019, 8-14 December 2019; pp. 1-11.). 따라서 본 발명에서는, 기본 SGN 모델의 복잡한 조인트 레벨 모듈 대신, 트랜스포머의 MHA를 SHA로 대체해 조인트-레벨 모듈(120)을 구성하였다. MHA 대신 SHA만 사용하더라도 중요한 관절을 강조하는 공간적 특징을 만들 수 있고, 기본 SGN보다 적은 수의 매개변수와 낮은 계산 복잡도를 확보할 수 있다. 이하에서는, 도 5를 참조하여 본 발명의 조인트-레벨 모듈(120)의 조인트-레벨 출력 특징 생성 과정을 상세히 설명하도록 한다.

[0116] 먼저, 동적 표현 모듈(110)에서 출력된 입력 특징과 관절 유형을 합산한 입력 벡터 $I \in \mathbb{R}^{T\times J\times C_1}$ 은 별도의 FC를 통해 다음 수학식 1 및 수학식 2와 같이 각각 Q와 K로 임베딩될 수 있다.

수학식 1

[0118] $$Q = FC^{q}(I) \quad \in \mathbb{R}^{T\times J\times C_2}$$

수학식 2

[0120] $$K = FC^{k}(I) \quad \in \mathbb{R}^{T\times J\times C_2}$$

[0122] 어텐션은 Q와 전치 K의 스케일드 닷 연산(scaled dot product)을 사용하여 계산될 수 있으며, Softmax 연산 결과를 다시 V와 연산하여 최종 어텐션을 생성할 수 있다.

수학식 3

[0124] $$\text{Attention} \quad (Q,K) = \text{softmax}\left(\frac{QK^{T}}{\sqrt{C'}}\right)V \in \mathbb{R}^{T\times J\times J}$$

[0126] 도 5에 도시된 바와 같이, 조인트-레벨 모듈(120)에 적용되는 SHA에서는, 순수 트랜스포머와 달리 입력 벡터도

$V \in \mathbb{R}^{T \times J \times C_1}$로 대체되며 별도의 FC를 거치지 않고 그대로의 어텐션을 곱한다. 최종 어텐션은 FC에 입력되고, 최종 어텐션을 고려하는 새로운 특징을 만들기 위해 T × J × C_3의 차원으로 변환될 수 있다.

[0128] FC 출력은 단일 레이어로 구성된 잔여 네트워크의 같은 크기의 출력과 합계(sum) 연산을 수행할 수 있다. 레이어 간의 잔여 연결은 입력 특징의 손실을 줄이고 임베딩된 특징을 강조하기 위해 사용될 수 있다. 최종 T × J × C_3 차원의 조인트-레벨 특징 표현은 Leaky ReLU, FC 및 잔여 네트워크로 순차로 합계 연산을 적용하여 얻을 수 있다. 조인트-레벨 모듈(120)의 최종 출력(조인트-레벨 출력 특징)은 프레임 간의 상관관계를 이용하기 위해 프레임-레벨 모듈(130)에 공급된다. 이와 같은 SHA를 이용한 조인트-레벨 모듈(120)을 통해, Q와 K 행렬에 대한 FC가 중요한 관절을 강조하도록 훈련된 경우, SHA만을 사용하여 중요한 관절의 어텐션을 얻을 수 있다.

[0130] 프레임-레벨 모듈(130)은, 조인트-레벨 모듈(120)의 조인트-레벨 출력 특징과 프레임 인덱스를 연결해 프레임-레벨 입력을 구성하고 프레임 간의 상관관계를 이용해 최종 행동 분류를 출력할 수 있다. 프레임-레벨 모듈(130)은, 두 개의 CNN을 포함할 수 있다. 또한, 프레임-레벨 모듈(130)은, 두 개의 CNN 이전에 공간 최대 풀링(SMP)을 포함하고, 두 개의 CNN 이후에 시간 최대 풀링(TMP) 및 FC를 더 포함할 수 있다.

[0132] 보다 구체적으로, 도 2에 도시된 바와 같이, 프레임 인덱스(Frame index)는 조인트-레벨 모듈(120)의 출력 특징 T × J × C_3와 연결되어 공간 차원을 압축하는 공간 최대 풀링(SMP)에 입력될 수 있다. 그런 다음 두 개의 CNN을 통과한 후 시간 차원을 압축하는 시간 최대 풀링(TMP)에 입력될 수 있다. TMP는 모든 프레임의 정보를 집계하고 C_3 차원의 시퀀스 레벨 특징 표현(C_{out})을 얻을 수 있다. C_{out} 특징 벡터는 softmax를 사용하여 마지막 FC를 통해 최종 분류를 출력할 수 있다.

[0134] 한편, 본 발명의 일실시예에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 방법은, 골격 시퀀스를 입력받아 의미-유도 신경망을 이용해 사람 행동을 인식하는, 컴퓨터에서 수행되는 사람 행동 인식 방법으로서, 골격 시퀀스에서 추출된 입력 특징에 싱글 헤드 어텐션(SHA)을 적용해 동일한 프레임에서 관절의 상관관계를 나타내는 조인트-레벨 출력 특징을 생성하는 조인트-레벨 모듈(120)을 이용해, 골격 시퀀스로부터 최종 행동 분류를 출력할 수 있다.

[0136] 도 6은 본 발명의 일실시예에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 방법의 흐름을 도시한 도면이다. 도 6에 도시된 바와 같이, 본 발명의 일실시예에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 방법은, 입력 특징에 싱글 헤드 어텐션(SHA)을 적용해 조인트-레벨 출력 특징을 생성하는 단계(S120)를 포함하여 구성될 수 있으며, 골격 시퀀스로부터 관절의 위치 및 속도 정보를 융합해 입력 특징을 생성하는 단계(S110) 및 조인트-레벨 출력 특징과 프레임 인덱스를 연결해 프레임-레벨 입력을 구성하고 최종 행동 분류를 출력하는 단계(S130)를 더 포함하여 구성될 수 있다. 보다 구체적으로, 단계 S120에서는, 조인트-레벨 모듈(120)은, 동적 표현 모듈(110)에서 생성한 입력 특징에 관절 유형을 결합한 입력 벡터에 싱글 헤드 어텐션(SHA)을 적용해 조인트-레벨 출력 특징을 생성할 수 있다.

[0138] 각각의 단계들과 관련된 상세한 내용들은, 앞서 본 발명의 일실시예에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템(100)과 관련하여 충분히 설명되었으므로, 상세한 설명은 생략하기로 한다.

[0140] 실험

[0142] 본 발명의 일실시예에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템(100) 및 방법

의 성능을 검증하기 위해 다른 최첨단(SoTA) 방법과 비교 실험을 수행하였다.

[0144] 데이터 세트

[0146] 본 발명의 일실시예에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템(100) 및 방법이 다양한 행동에 대해 매우 강력한 성능을 보인다는 것을 증명하기 위해 다양한 환경에서 얻은 골격 데이터 세트의 정확도를 측정했다.

[0148] (1) NTU-RGB+D 60 (NTU60)

[0149] NTU-RGB+D 60 데이터 세트는 인간 행동 인식을 위한 깊이 기반(depth-based) 및 RGB 기반(RGB-based) 행동 인식 벤치마크로, 40개의 개별 피사체에서 수집된 56,000개 이상의 비디오 샘플과 400만 개 이상의 프레임을 포함한다. 깊이 맵, 3D 골격 관절 위치, RGB 프레임, 적외선 시퀀스 등 4가지 행동 정보를 제공한다. 이 데이터 세트에는 40개의 일상적인 행동(daily action), 11개의 상호 행동(mutual action) 및 9개의 건강 관련 행동(health-related action)을 포함하여 60개의 서로 다른 행동 클래스(action classes)가 포함되어 있다. 두 가지 평가 프로토콜, 즉 교차-주제(cross-subject, XSub) 및 교차-뷰(cross-view, XView) 프로토콜을 사용한다.

[0151] (2) NTU-RGB+D 120 (NTU120)

[0152] NTU RGB+D 120은 NTU 60의 확장으로, 106개의 개별 주제로부터 수집되었으며 114,000개 이상의 비디오 샘플과 800만 개 이상의 프레임을 포함한다. 이 데이터 세트에는 NTU60 이상으로 확장된 120개의 다양한 행동 클래스가 포함되어 있다. 두 가지 평가 프로토콜, 즉 XSub 프로토콜과 교차 설정(cross-setup, XSet) 프로토콜을 사용한다.

[0154] (3) NW-UCLA DB

[0155] NW-UCLA 데이터 세트는 3개의 키넥트(Kinect) 카메라를 사용하여 동시에 캡처된 RGB, 깊이 및 인간 골격 데이터를 포함한다. 골격에는 20개의 관절과 19개의 뼈가 연결되어 있다. 이 데이터 세트에는 ‘한 손으로 줍기’, ‘두 손으로 줍기’, ‘쓰레기 버리기’, ‘걸어 다니기’, ‘앉기’, ‘일어나기’ 등을 포함하는 10명의 10개의 행동 카테고리가 포함되어 있다.

[0157] 구현 세부 정보

[0159] 모든 실험은 파이토치(Pytorch) 프레임워크가 있는 RTX 2080Ti GPU에서 수행되었다. 훈련은 최대 120개 에포크(epoch) 동안 모든 데이터 세트에 동일하게 적용되었다. 데이터 세트의 배치 크기도 64로 설정되었다. 모든 FC는 단일 레이어로 구성되었으며 배치 정규화, Leaky ReLU 및 Adam 옵티마이저가 0.0005의 중량 감소(weight decay)로 사용되었다. 초기 학습률(learning rate)은 NTU 데이터 세트의 경우 0.001, UCLA 데이터 세트의 경우 0.01였고, 학습률 스케줄러는 파이토치에서 제공한 LambdaLA(lr_lambda = lambda, epoch = 0.93 ** epoch)를 사용하였다. 분류 훈련을 위해서는 교차 엔트로피 손실(cross-entropy loss)을 사용하였다.

[0161] NTU60의 성능 비교

[0163] 행동 인식에 널리 적용되는 NTU60 데이터 세트를 사용하여 SHA를 SGN과 결합한 모델(본 발명에서 제안하는 SGN-SHA)이 행동 인식 성능을 유지하면서 경량화될 수 있음을 증명한다. 실험을 통해 NTU60의 두 가지 평가 프로토콜, 즉 XSub와 XView를 사용하여 성능을 비교하였다.

[0165] 비교 실험에 사용된 SoTA 방법은 다음과 같다: 1) ST-GCN, 2) 그래프 컨볼루션(MS-G3D), 3) 분리된 공간-시간 주의 네트워크(DSTA-Net), 4) 이동 그래프 컨볼루션 네트워크(Shift-GCN), 5) 인간 골격에 대한 표현을 학습한 info-GCN, 6) 극도로 증강된 골격 시퀀스로부터의 대조 학습에 기반한 3S-aimCLR, 7) 기본 SGN, 및 8) 본 발명에서 제안된 SGN-SHA.

[0167] 도 7은 NTU60 데이터셋에서 본 발명의 일실시예에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템(100) 및 방법과 다른 최신 기술의 정확도(accuracy), 매개변수 개수(Params (M: Million)), 계산 복잡도(Giga FLOating Point Operation Per Second(GFLOPs))를 나타낸 도면이다. 여기서, 정확도는 NTU60의 두 가지 평가 프로토콜인 XSub와 XView을 비교했다. 도 7에서 알 수 있듯이, Info-GCN은 NTU60의 XSub 및 XView에 대해 가장 높은 정확도를 달성했다. 그러나 Info-GCN은 SGN-SHA보다 4.8배 더 많은 매개변수와 29.5배 더 많은 계산 복잡도를 필요로 하므로, SGN-SHA가 실시간 처리 측면에서 더 실현 가능하다는 것을 알 수 있다. Info-GCN을 제외한 SGN-SHA와 다른 방법 간의 정확도 차이는 크지 않지만 매개변수의 수와 계산 복잡도 측면에서 SGN-SHA가 상당히 우수하다. 특히, 기본 SGN과 비교하여 SGN-SHA의 정확도는 XSub에서 1.5%, XView에서 1.9%로 SGN보다 약간 낮지만 매개변수 개수 측면에서는 약 2.15%, 계산 복잡도 측면에서는 약 2.58%로 SGN을 능가한다. 도 7과 같이 정확도 면에서 SoTA 방법과 거의 차이가 없었다는 점을 고려할 때, 제안된 SGN-SHA의 계산 복잡도가 크게 감소한 것은 저가형 장치(low-end device)에서 실시간 행동 인식을 가능하게 하므로 중요한 의미를 가진다.

[0169] NTU120의 성능 비교

[0171] NTU120 데이터 세트를 사용하여 실험을 수행하였으며, 이 데이터 세트는 XSub와 XSet의 두 가지 평가 프로토콜을 포함한다.

[0173] 도 8은 NTU120 데이터셋에서 본 발명의 일실시예에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템(100) 및 방법과 다른 최신 기술의 정확도, 매개변수 개수, 계산 복잡도를 나타낸 도면이다. 도 8에 나타난 바와 같이, SGN-SHA는 가장 적은 수의 매개변수와 상당히 낮은 계산 복잡도를 보여주었다. Info-GCN은 복잡한 네트워크를 사용했기 때문에 클래스 수가 60개에서 120개로 늘어도 여전히 가장 높은 인식률을 보였다. 제안된 SGN-SHA는 Info-GCN보다 낮은 정확도를 달성했지만, SoTA 방식인 3s-AimCLR과 비교하면 XSub에서는 1.6% 낮지만 XSet에서는 6.2% 높다. 특히 기본 SGN과 비교하면 XSub이 0.7%로 다소 낮은 반면 XSet는 5.6%로 높은 편이었다. 도 8의 결과를 통해 클래스 수가 증가하더라도 SGN-SHA의 성능은 크게 감소하지 않으며, 여전히 매개변수가 적고 계산 복잡도가 낮아 실시간 처리에 적합하다는 것을 확인할 수 있다.

[0175] NW-UCLA 데이터 세트의 성능 비교

[0177] NW-UCLA 데이터 세트는 10개의 행동 클래스로 구성되며, 전체 행동 정확도(action accuracy)는 기본 SGN의 경우 92.5%, SGN-SHA의 경우 92.6%로 제안된 방법이 0.1% 높다. 본 실험에서는 기본 SGN과 제안된 SGN-SHA 사이의 각 세부 행동 클래스의 정확도를 혼동 행렬을 이용하여 비교하였다.

[0179] 도 9는 NW-UCLA 데이터셋의 세부 행동에 대한 의미-유도 신경망(기본 SGN)의 혼동 행렬(confusion matrix)을 나타낸 도면이고, 도 10은 NW-UCLA 데이터셋에서 본 발명의 일실시예에 따른 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템(100) 및 방법(SGN-SHA)의 혼동 행렬을 나타낸 도면이다. 도 9 및 도 10에 도시된 바와 같이, 두 방법 모두 세부 행동 클래스에 대해 유사한 정확도를 보였지만, SGN은 75%, SGN-SHA는 83%로 '한 손으로 줍기' 클래스에서 가장 낮은 정확도를 달성했다. SGN의 경우 '두 손으로 줍기', '쓰레기

버리기', '쓰레기 버리기' 등 3개 클래스에서 100% 정확도를 보였다. 반면, SGN-SHA는 '두 손으로 줍기', '일어나기', '착용(donning)'의 3가지 클래스에서 100% 정확도를 보여 두 방법의 결과가 약간 달랐다. 그러나 전반적으로 두 방법 사이에 각 행동 클래스의 정확도에 큰 차이가 없었기 때문에, SGN-SHA는 매개변수가 더 적고 계산 복잡도가 더 낮음에도 불구하고 기본 SGN과 유사한 성능을 가지고 있음을 알 수 있다.

[0181] 전술한 바와 같이, 본 발명에서 제안하고 있는 싱글 헤드와 경량 의미-유도 신경망을 이용한 사람 행동 인식 시스템(100) 및 방법에 따르면, 골격 시퀀스를 입력받아 의미-유도 신경망(SGN)을 이용해 사람 행동을 인식하되, 골격 시퀀스에서 추출된 입력 특징에 싱글 헤드 어텐션(SHA)을 적용해 동일한 프레임에서 관절의 상관관계를 나타내는 조인트-레벨 출력 특징을 생성하는 조인트-레벨 모듈(120)을 포함함으로써, 골격 관절의 공간적 특징 추출을 위해 GCN 대신 단일 SHA를 적용해 매개변수의 개수 및 연산 수를 대폭 줄이면서도 높은 정확도를 달성할 수 있다.

[0183] 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.

[0185] 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.

[0187] 본 발명은 특정 실시예와 관련하여 설명되었지만, 그것들의 구성 요소 또는 동작의 일부 또는 전부는 범용 하드웨어 아키텍처를 갖는 컴퓨터 시스템을 사용하여 구현될 수 있다.

[0189] 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.

[0191] 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

[0192] 100: 본 발명의 특징에 따른 사람 행동 인식 시스템

110: 동적 표현 모듈

120: 조인트-레벨 모듈

130: 프레임-레벨 모듈

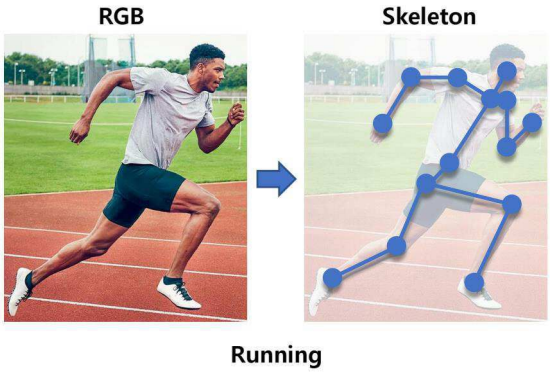
S110: 골격 시퀀스로부터 관절의 위치 및 속도 정보를 융합해 입력 특징을 생성하는 단계

S120: 입력 특징에 싱글 헤드 어텐션(SHA)을 적용해 조인트-레벨 출력 특징을 생성하는 단계

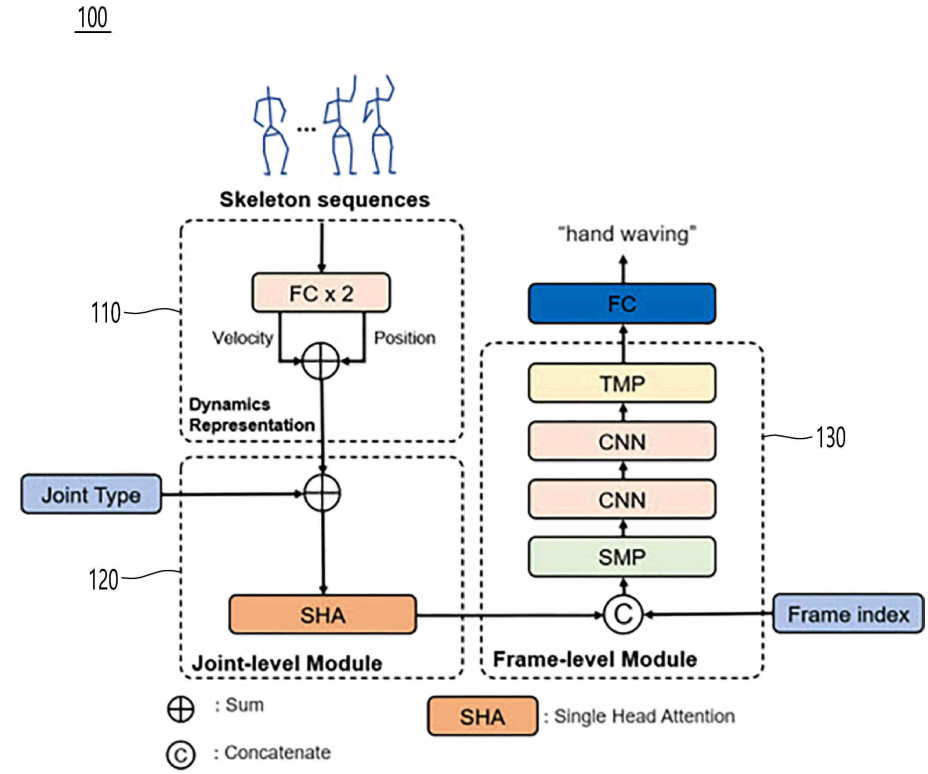
S130: 조인트-레벨 출력 특징과 프레임 인덱스를 연결해 프레임-레벨 입력을 구성하고 최종 행동 분류를 출력하는 단계

도면

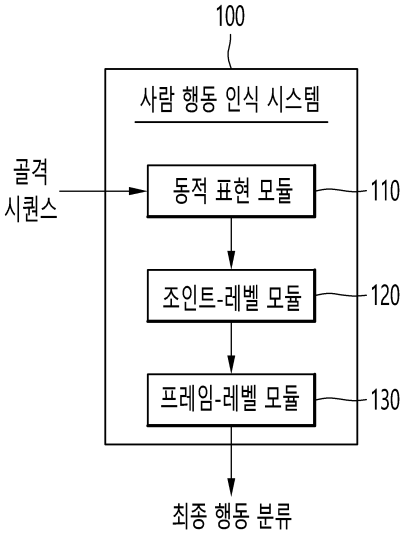
도면1



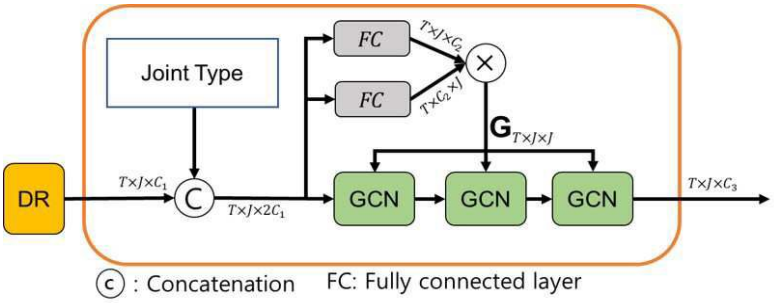
도면2



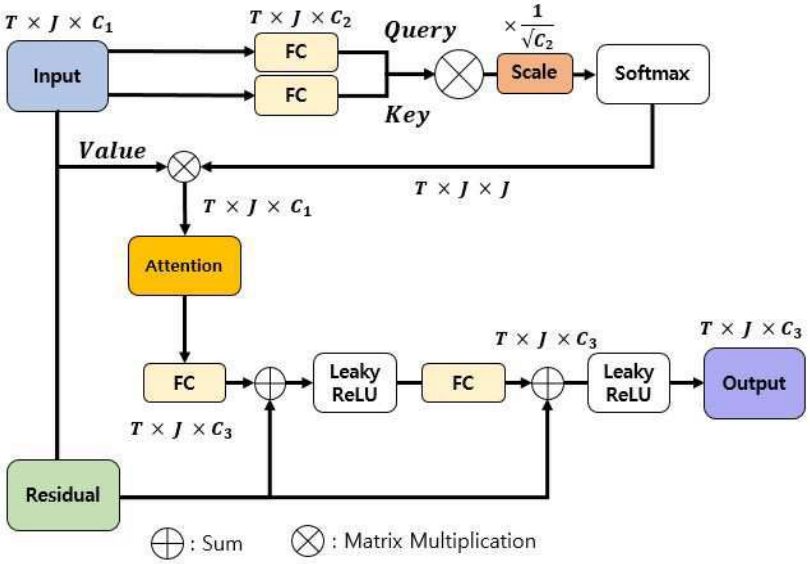
도면3


100
사람 행동 인식 시스템
골격 시퀀스
동적 표현 모듈
110
조인트-레벨 모듈
120
프레임-레벨 모듈
130
최종 행동 분류

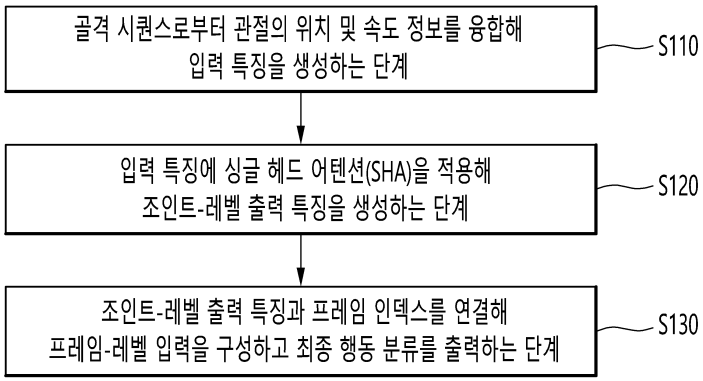
도면4


Joint Type
FC
FC
DR
GCN
GCN
GCN
C : Concatenation
FC: Fully connected layer

도면5


Input
FC
FC
Query
Key
Scale
Softmax
Value
Attention
FC
Leaky ReLU
FC
Leaky ReLU
Output
Residual
: Sum
: Matrix Multiplication

도면6



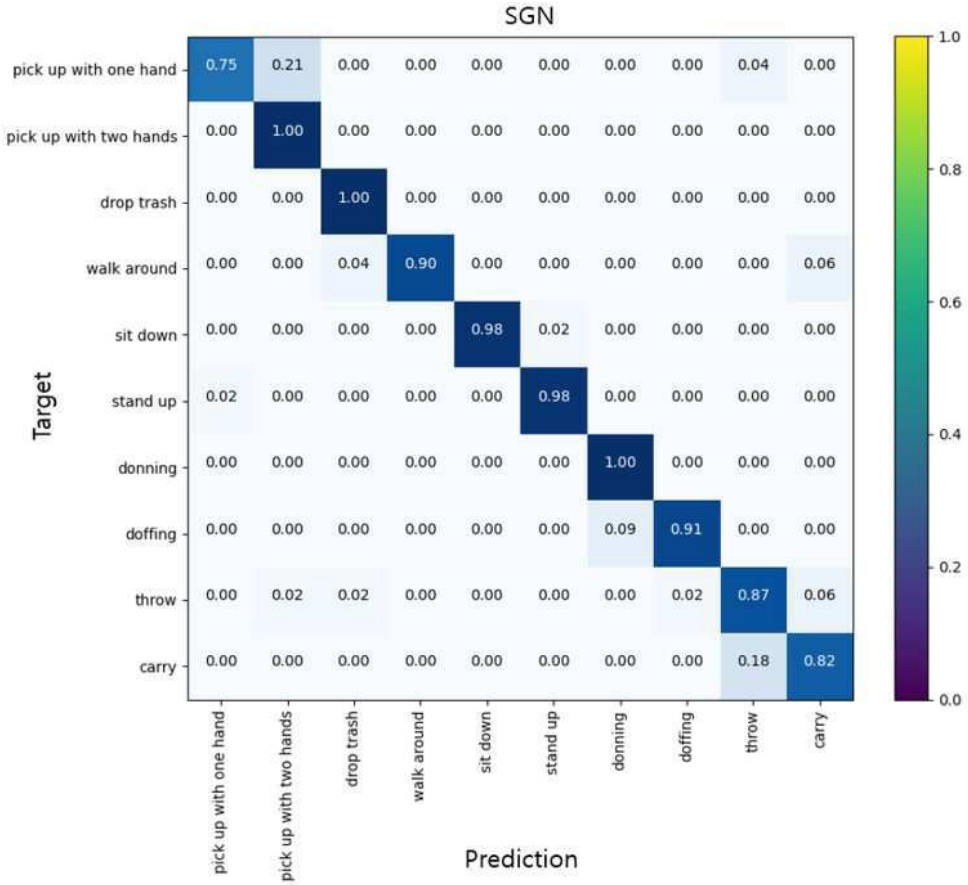
도면7

Method	NTU60		Lightweight	
	XSub (%)	XView (%)	Params (M)	GFLOPs
ST-GCN [4]	81.5	88.3	3.10	16.3
MS-G3D [26]	91.5	96.2	3.50	48.8
DSTA-Net [27]	91.5	96.4	3.40	55.6
Shift-GCN [12]	90.7	96.5	0.76	10.0
Info-GCN [25]	93.0	97.1	1.56	1.83
3s-AimCLR [28]	86.9	92.8	2.51	1.86
SGN [1]	89.0	94.5	0.69	0.16
SGN-SHA	87.5	92.6	0.32	0.062

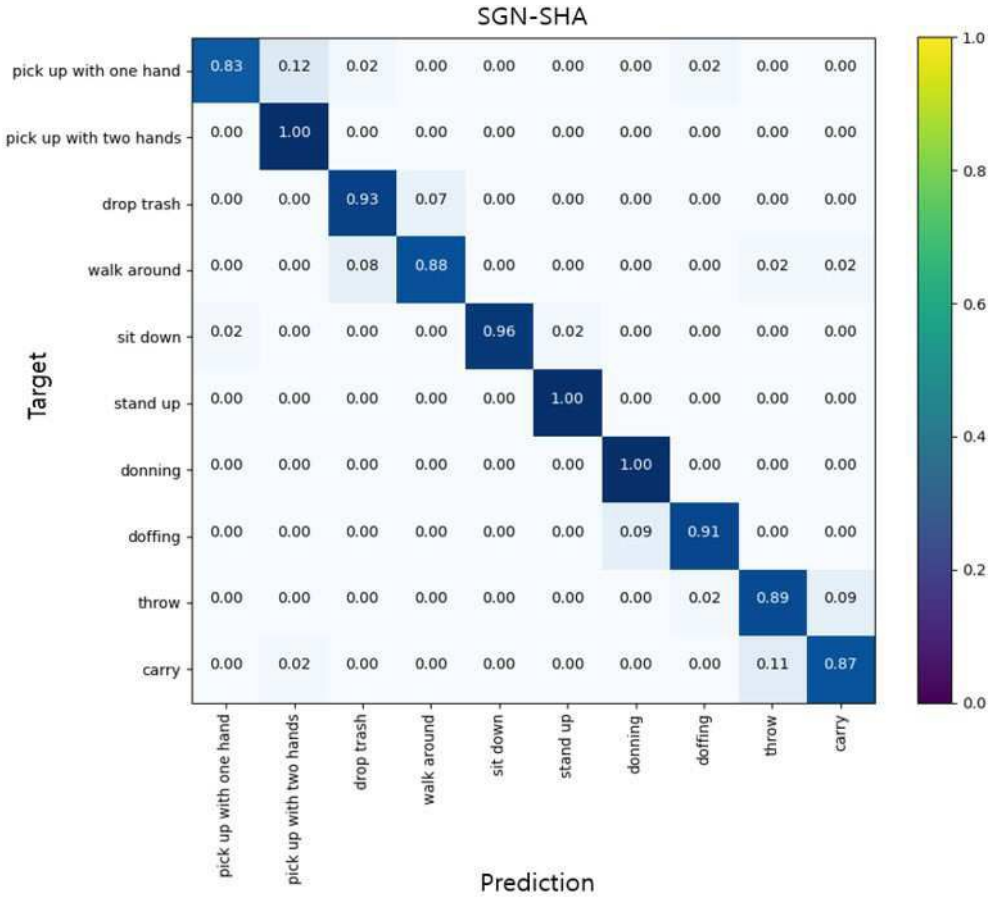
도면8

Method	NTU120		Lightweight	
	XSub (%)	XSet (%)	Params (M)	GFLOPs
MS-G3D [4]	86.9	88.4	3.50	48.8
DSTA-Net [27]	86.6	89.0	3.40	55.6
Shift-GCN [12]	85.9	87.6	0.76	10.0
Info-GCN [25]	89.8	91.2	1.56	1.83
3s-AimCLR[28]	80.1	80.9	2.51	1.86
SGN [1]	79.2	81.5	0.69	0.16
SGN-SHA	78.5	87.1	0.32	0.062

도면9


SGN
pick up with one hand
pick up with two hands
drop trash
walk around
sit down
stand up
donning
doffing
throw
carry
Target
Prediction
0.75 0.21 0.00 0.00 0.00 0.00 0.00 0.00 0.04 0.00
0.00 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.00 0.00 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.00 0.00 0.04 0.90 0.00 0.00 0.00 0.00 0.00 0.06
0.00 0.00 0.00 0.00 0.98 0.02 0.00 0.00 0.00 0.00
0.02 0.00 0.00 0.00 0.00 0.98 0.00 0.00 0.00 0.00
0.00 0.00 0.00 0.00 0.00 0.00 1.00 0.00 0.00 0.00
0.00 0.00 0.00 0.00 0.00 0.00 0.09 0.91 0.00 0.00
0.00 0.02 0.02 0.00 0.00 0.00 0.00 0.02 0.87 0.06
0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.18 0.82
1.0
0.8
0.6
0.4
0.2
0.0

도면10


SGN-SHA
pick up with one hand
pick up with two hands
drop trash
walk around
sit down
stand up
donning
doffing
throw
carry
Target
Prediction
0.83 0.12 0.02 0.00 0.00 0.00 0.00 0.02 0.00 0.00
0.00 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.00 0.00 0.93 0.07 0.00 0.00 0.00 0.00 0.00 0.00
0.00 0.00 0.08 0.88 0.00 0.00 0.00 0.00 0.02 0.02
0.02 0.00 0.00 0.00 0.96 0.02 0.00 0.00 0.00 0.00
0.00 0.00 0.00 0.00 0.00 1.00 0.00 0.00 0.00 0.00
0.00 0.00 0.00 0.00 0.00 0.00 1.00 0.00 0.00 0.00
0.00 0.00 0.00 0.00 0.00 0.00 0.09 0.91 0.00 0.00
0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.02 0.89 0.09
0.00 0.02 0.00 0.00 0.00 0.00 0.00 0.00 0.11 0.87
1.0
0.8
0.6
0.4
0.2
0.0