(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2022년01월27일
(11) 등록번호 10-2357000
(24) 등록일자 2022년01월25일

(51) 국제특허분류(Int. Cl.)
G06K 9/00 (2022.01) *G06V 10/46* (2022.01)
(52) CPC특허분류
G06V 40/20 (2022.01)
G06N 3/0454 (2013.01)
(21) 출원번호 10-2020-0029743
(22) 출원일자 2020년03월10일
심사청구일자 2020년03월10일
(65) 공개번호 10-2021-0114257
(43) 공개일자 2021년09월23일
(56) 선행기술조사문헌
KR101930940 B1*
KR1020190120489 A*
US20150023590 A1*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
김은태
서울특별시 서대문구 연세로 50, 제3공학관 C607호(신촌동, 연세대학교)
성홍제
서울특별시 서대문구 연세로 50, 제3공학관 C607호(신촌동, 연세대학교)
현준혁
서울특별시 서대문구 연세로 50, 제3공학관 C607호(신촌동, 연세대학교)
(74) 대리인
특허법인우인

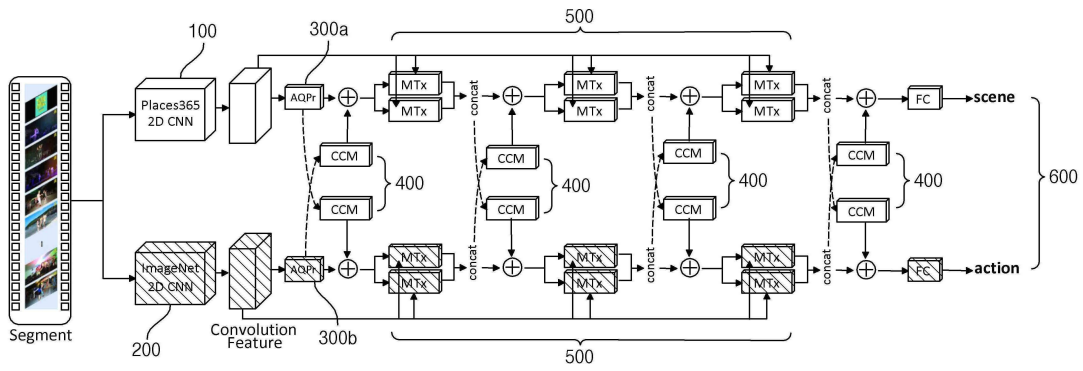
전체 청구항 수 : 총 12 항 심사관 : 노용완

(54) 발명의 명칭 **인공 신경망 기반의 비정제 동영상에서의 행동 인식 방법 및 장치**

(57) 요 약

본 발명에 따르면, 프로세서가 분석 대상 영상을 입력 받고, 시간 영역을 기준으로 기 설정된 구간별로 상기 분석 대상 영상에서 일부의 프레임 영상을 선택하고, 상기 선택된 프레임 영상에서 장소와 행동을 인식하여 인식한 장소와 행동에 따른 특징값을 상기 선택된 프레임 영상에 라벨링하여 비정제 동영상에서 클립 단위 장소와 행동 정보로 학습한 인공신경망으로 장소와 행동이 어느 프레임의 어느 공간 영역에서 나타나고 있는지 찾는 인공 신경망 기반의 비정제 동영상에서의 행동 인식 방법 및 장치가 개시된다.

대 표 도



(52) CPC특허분류
G06N 3/08 (2013.01)
G06V 10/469 (2022.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 NRF- 2017M3C4A7069370
부처명 미래창조과학부
과제관리(전문)기관명 한국연구재단
연구사업명 차세대정보컴퓨팅기술개발사업
연구과제명 딥러닝 기반 의미론적 상황 이해 원천기술 연구
기 여 율 1/1
과제수행기관명 연세대학교
연구기간 2017.09.01 ~ 2020.12.31
공지예외적용 : 있음

명 세 서

청구범위

청구항 1

프로세서가, 분석 대상 영상을 입력 받고, 시간 영역을 기준으로 기 설정된 구간별로 상기 분석 대상 영상에서 일부의 프레임 영상을 선택하는 단계;

상기 선택된 프레임 영상에서 장소와 행동을 인식하는 단계; 및

인식한 장소와 행동에 따른 특징값을 상기 선택된 프레임 영상에 라벨링하는 단계;를 포함하며,

상기 선택된 프레임 영상에서 장소와 행동을 인식하는 단계는,

제1 합성곱 신경망을 이용하여 상기 선택된 프레임 영상의 장소 인식을 위한 제1 특징 텐서를 추출하는 단계;

제2 합성곱 신경망을 이용하여 상기 선택된 프레임 영상의 객체 인식을 위한 제2 특징 텐서를 추출하는 단계;

CCM(class conversion matrix) 연산부를 이용하여 상기 장소와 상기 행동의 특징 정보를 공유하되 기울기를 전파하지 않고,

MTx(Multitask Transformer unit) 연산부를 이용하여 제1 그룹에 속하는 MTx는 상기 장소에 대해 집중 학습하고, 제2 그룹에 속하는 MTx는 상기 행동에 대해 집중 학습하는 것을 특징으로 하는 행동 인식 방법.

청구항 2

삭제

청구항 3

제1항에 있어서,

상기 제2 합성곱 신경망은,

인식하고자 하는 행동 정보와 유사한 객체 인식 데이터셋에 학습된 것을 특징으로 하는 행동 인식 방법.

청구항 4

제1항에 있어서,

상기 선택된 프레임 영상에서 장소와 행동을 인식하는 단계는,

어텐션 함수(Attention Function)를 이용한 연산을 수행하여 시공간 영역의 상기 제1 특징 텐서를 기 설정된 크기의 제1 특징 벡터로 추출하는 단계; 및

어텐션 함수(Attention Function)를 이용한 연산을 수행하여 시공간 영역의 상기 제2 특징 텐서를 기 설정된 크기의 제2 특징 벡터로 추출하는 단계;를 더 포함하는 것을 특징으로 하는 행동 인식 방법.

청구항 5

제4항에 있어서,

상기 선택된 프레임 영상에서 장소와 행동을 인식하는 단계는,

상기 제1 특징 벡터 또는 상기 제2 특징 벡터의 차원 변환을 통해 연산이 가능하도록 변환하는 클래스 변환 연산을 수행하는 단계; 및

상기 제1 특징 벡터와 클래스 변환 연산을 수행한 상기 제2 특징 벡터를 합하여 시공간 영역에 해당하는 특징을 추출하기 위한 멀티태스크 트랜스포머 유닛을 이용한 트랜스포머 연산을 수행하여 결합 특징 벡터를 추출하는 단계;를 더 포함하는 것을 특징으로 하는 행동 인식 방법.

청구항 6

제5항에 있어서,

상기 멀티태스크 트랜스포머 유닛을 이용하여 트랜스포머 연산을 수행하는 단계는,

쿼리 입력부가 상기 제1 특징 벡터를 입력받고, 상기 입력된 제1 특징 벡터와 미리 결정된 연결 가중치에 따른 풀리 커넥티트 특징값을 생성하는 단계;

상기 선택된 프레임 영상을 컨볼루션 변환을 통해 컨볼루션 특징값을 생성하는 단계;

상기 풀리 커넥티트 특징값과 상기 컨볼루션 특징값의 행렬 곱 연산을 수행하는 단계; 및

상기 풀리 커넥티트 특징값과 상기 행렬 곱 연산을 수행한 상기 컨볼루션 특징값을 합하여 정규화를 수행하는 단계;를 포함하는 것을 특징으로 하는 행동 인식 방법.

청구항 7

제5항에 있어서,

상기 트랜스포머 연산은, 제1 트랜스포머 연산 내지 제3 트랜스포머 연산을 포함하며,

상기 선택된 프레임 영상에서 장소와 행동을 인식하는 단계는,

상기 제3 트랜스포머 연산을 수행하여 추출한 상기 결합 특징 벡터를 풀리 커넥티드 레이어(fully-connected layer)를 이용하여 장소와 행동을 분류하는 단계;를 더 포함하는 것을 특징으로 하는 행동 인식 방법.

청구항 8

외부로부터 인식하고자 하는 행동 정보가 포함된 분석 대상 영상을 획득하는 영상 획득부;

하나 이상의 인스트럭션을 저장하는 메모리; 및

상기 메모리에 저장된 하나 이상의 인스트럭션을 실행하는 프로세서;를 포함하고,

상기 프로세서는, 인공신경망을 기반으로 상기 분석 대상 영상으로부터 장소와 행동을 인식하며,

상기 프로세서는, 상기 분석 대상 영상에서 시간 영역을 기준으로 기 설정된 구간별로 상기 분석 대상 영상에서 일부의 프레임 영상을 선택하는 단계;

상기 선택된 프레임 영상에서 장소와 행동을 인식하는 단계; 및

인식한 장소와 행동에 따른 특징값을 상기 선택된 프레임 영상에 라벨링하는 단계;를 수행하며,

상기 프로세서는, 제1 합성곱 신경망을 이용하여 상기 선택된 프레임 영상의 장소 인식을 위한 제1 특징 텐서를 추출하는 단계; 및

제2 합성곱 신경망을 이용하여 상기 선택된 프레임 영상의 객체 인식을 위한 제2 특징 텐서를 추출하는 단계;를 수행하며,

상기 프로세서는, CCM(class conversion matrix) 연산부를 이용하여 상기 장소와 상기 행동의 특징 정보를 공유하되 기울기를 전파하지 않고,

MTx(Multitask Transformer unit) 연산부를 이용하여 제1 그룹에 속하는 MTx는 상기 장소에 대해 집중 학습하고, 제2 그룹에 속하는 MTx는 상기 행동에 대해 집중 학습하는 것을 특징으로 하는 행동 인식 장치.

청구항 9

삭제

청구항 10

삭제

청구항 11

제8항에 있어서,

상기 제2 합성곱 신경망은,

상기 인식하고자 하는 행동 정보와 유사한 객체 인식 데이터셋에 학습된 것을 특징으로 하는 행동 인식 장치.

청구항 12

제8항에 있어서,

상기 프로세서는, 어텐션 함수(Attention Function)를 이용한 연산을 수행하여 시공간 영역의 상기 제1 특징 텐서를 기 설정된 크기의 제1 특징 벡터로 추출하는 단계; 및

어텐션 함수(Attention Function)를 이용한 연산을 수행하여 시공간 영역의 상기 제2 특징 텐서를 기 설정된 크기의 제2 특징 벡터로 추출하는 단계;를 수행하는 것을 특징으로 하는 행동 인식 장치.

청구항 13

제12항에 있어서,

상기 프로세서는, 상기 제1 특징 벡터 또는 상기 제2 특징 벡터의 차원 변환을 통해 연산이 가능하도록 변환하는 클래스 변환 연산을 수행하는 단계; 및

상기 제1 특징 벡터와 클래스 변환 연산을 수행한 상기 제2 특징 벡터를 합하여 시공간 영역에 해당하는 특징을 추출하기 위한 멀티태스크 트랜스포머 유닛을 이용한 트랜스포머 연산을 수행하여 결합 특징 벡터를 추출하는 단계;를 수행하는 것을 특징으로 하는 행동 인식 장치.

청구항 14

제13항에 있어서,

상기 프로세서는, 상기 제1 특징 벡터를 입력받고, 상기 입력된 제1 특징 벡터와 미리 결정된 연결 가중치에 따른 풀리 커넥티트 특징값을 생성하는 단계;

상기 선택된 프레임 영상을 컨볼루션 변환을 통해 컨볼루션 특징값을 생성하는 단계;

상기 풀리 커넥티트 특징값과 상기 컨볼루션 특징값의 행렬 곱 연산을 수행하는 단계; 및

상기 풀리 커넥티트 특징값과 상기 행렬 곱 연산을 수행한 상기 컨볼루션 특징값을 합하여 정규화를 수행하는 단계;를 수행하는 것을 특징으로 하는 행동 인식 장치.

청구항 15

제13항에 있어서,

상기 트랜스포머 연산은, 제1 트랜스포머 연산 내지 제3 트랜스포머 연산을 포함하며,

상기 프로세서는, 상기 제3 트랜스포머 연산을 수행하여 추출한 상기 결합 특징 벡터를 풀리 커넥티드 레이어(fully-connected layer)를 이용하여 장소와 행동을 분류하는 단계;를 수행하는 것을 특징으로 하는 행동 인식 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 행동 인식 방법 및 장치에 관한 것으로, 특히 인공 신경망을 기반으로 한 비정제 동영상에서의 행동 인식 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 최근 딥 러닝이 컴퓨터 비전 분야에서 다양한 문제를 해결하는데 많이 사용되고 있다. 하지만 딥 러닝 기반 시스템을 사용하기 위해선 학습에 필요한 많은 양의 데이터를 필요로 한다. 또한, 실제 상황에 사용하기에 적합한 시스템을 구축하기 위해선 실제 상황과 유사한 데이터에 학습을 해야 하며, 이는 비정제 동영상이 매우 적합하

고, 장소와 행동 인식은 실제 상황에 적용하기 위해 가장 기본적으로 수행되어야 할 기술이다. 하지만, 학습에 사용될 데이터들은 정확한 정보를 제공해야 하기 때문에 사람이 직접 만들어야 하며, 비정제 동영상은 필요 없는 프레임을 많이 포함하고 있기 때문에 비정제 동영상 데이터를 구축하기 위해 장소와 행동이 나타나는 프레임을 찾는 것은 매우 많은 시간을 필요로 한다. 게다가 각 프레임 마다 장소와 행동이 일어나는 구역을 나타내는 것 또한 매우 많은 노동력을 필요로 한다. 그에 반해, 클립 단위 (약 5초 간격)으로 해당 클립이 어떠한 장소와 행동으로 이루어져 있는지 라벨링 하는 것은 비교적 적은 노동력을 필요로 한다.

발명의 내용

해결하려는 과제

[0003] 본 발명은 인공 신경망 기반의 비정제 동영상에서의 행동 인식 방법 및 장치로 프로세서가 분석 대상 영상을 입력 받고, 시간 영역을 기준으로 기 설정된 구간별로 상기 분석 대상 영상에서 일부의 프레임 영상을 선택하고, 상기 선택된 프레임 영상에서 장소와 행동을 인식하여 인식한 장소와 행동에 따른 특징값을 상기 선택된 프레임 영상에 라벨링하여 비정제 동영상에서 클립 단위 장소와 행동 정보로 학습한 인공신경망으로 장소와 행동이 어느 프레임의 어느 공간 영역에서 나타나고 있는지 찾는데 그 목적이 있다.

[0004] 또한, 장소와 행동의 영역을 함께 도출해내는 단일 인공신경망을 사용하며, 서로 다른 2가지 이상의 태스크(task)를 해결하는 멀티태스킹(multitasking) 방법을 이용하여, 서로 연관성이 높은 장소와 행동의 영역을 함께 도출하는데 또 다른 목적이 있다.

[0005] 본 발명의 명시되지 않은 또 다른 목적들은 하기의 상세한 설명 및 그 효과로부터 용이하게 추론할 수 있는 범위 내에서 추가적으로 고려될 수 있다.

과제의 해결 수단

[0006] 상기 과제를 해결하기 위해, 본 발명의 일 실시예에 따른 행동 인식 방법은, 프로세서가, 분석 대상 영상을 입력 받고, 시간 영역을 기준으로 기 설정된 구간별로 상기 분석 대상 영상에서 일부의 프레임 영상을 선택하는 단계, 상기 선택된 프레임 영상에서 장소와 행동을 인식하는 단계 및 인식한 장소와 행동에 따른 특징값을 상기 선택된 프레임 영상에 라벨링하는 단계를 포함한다.

[0007] 여기서, 상기 선택된 프레임 영상에서 장소와 행동을 인식하는 단계는, 제1 합성곱 신경망을 이용하여 상기 선택된 프레임 영상의 장소 인식을 위한 제1 특징 텐서를 추출하는 단계 및 제2 합성곱 신경망을 이용하여 상기 선택된 프레임 영상의 객체 인식을 위한 제2 특징 텐서를 추출하는 단계를 포함한다.

[0008] 여기서, 상기 제2 합성곱 신경망은, 상기 인식하고자 하는 행동 정보와 유사한 객체 인식 데이터셋에 학습된 것이다.

[0009] 여기서, 상기 선택된 프레임 영상에서 장소와 행동을 인식하는 단계는, 어텐션 함수(Attention Function)를 이용한 연산을 수행하여 시공간 영역의 상기 제1 특징 텐서를 기 설정된 크기의 제1 특징 벡터로 추출하는 단계 및 어텐션 함수(Attention Function)를 이용한 연산을 수행하여 시공간 영역의 상기 제2 특징 텐서를 기 설정된 크기의 제2 특징 벡터로 추출하는 단계를 더 포함한다.

[0010] 여기서, 상기 선택된 프레임 영상에서 장소와 행동을 인식하는 단계는, 상기 제1 특징 벡터 또는 상기 제2 특징 벡터의 차원 변환을 통해 연산이 가능하도록 변환하는 클래스 변환 연산을 수행하는 단계 및 상기 제1 특징 벡터와 클래스 변환 연산을 수행한 상기 제2 특징 벡터를 합하여 시공간 영역에 해당하는 특징을 추출하기 위한 멀티태스크 트랜스포머 유닛을 이용한 트랜스포머 연산을 수행하여 결합 특징 벡터를 추출하는 단계를 더 포함한다.

[0011] 여기서, 상기 멀티태스크 트랜스포머 유닛을 이용하여 트랜스포머 연산을 수행하는 단계는, 쿼리 입력부가 상기 제1 특징 벡터를 입력받고, 상기 입력된 제1 특징 벡터와 미리 결정된 연결 가중치에 따른 풀리 커넥티트 특징값을 생성하는 단계, 상기 선택된 프레임 영상을 컨볼루션 변환을 통해 컨볼루션 특징값을 생성하는 단계, 상기 풀리 커넥티트 특징값과 상기 컨볼루션 특징값의 행렬 곱 연산을 수행하는 단계 및 상기 풀리 커넥티트 특징값과 상기 행렬 곱 연산을 수행한 상기 컨볼루션 특징값을 합하여 정규화를 수행하는 단계를 포함한다.

[0012] 여기서, 상기 트랜스포머 연산은, 제1 트랜스포머 연산 내지 제3 트랜스포머 연산을 포함하며, 상기 선택된 프레임 영상에서 장소와 행동을 인식하는 단계는, 상기 제3 트랜스포머 연산을 수행하여 추출한 상기 결합 특징

벡터를 풀리 커넥티드 레이어(fully-connected layer)를 이용하여 장소와 행동을 분류하는 단계를 더 포함한다.

[0013] 본 발명의 일 실시예에 따른 행동 인식 장치는, 외부로부터 인식하고자 하는 행동 정보가 포함된 분석 대상 영상을 획득하는 영상 획득부, 하나 이상의 인스트럭션을 저장하는 메모리 및 상기 메모리에 저장된 하나 이상의 인스트럭션을 실행하는 프로세서를 포함하고, 상기 프로세서는, 인공신경망을 기반으로 상기 분석 대상 영상으로부터 장소와 행동을 인식한다.

[0014] 여기서, 상기 프로세서는, 상기 분석 대상 영상에서 시간 영역을 기준으로 기 설정된 구간별로 상기 분석 대상 영상에서 일부의 프레임 영상을 선택하는 단계, 상기 선택된 프레임 영상에서 장소와 행동을 인식하는 단계 및 인식한 장소와 행동에 따른 특징값을 상기 선택된 프레임 영상에 라벨링하는 단계를 수행한다.

[0015] 여기서, 상기 프로세서는, 제1 합성곱 신경망을 이용하여 상기 선택된 프레임 영상의 장소 인식을 위한 제1 특징 텐서를 추출하는 단계 및 제2 합성곱 신경망을 이용하여 상기 선택된 프레임 영상의 객체 인식을 위한 제2 특징 텐서를 추출하는 단계를 수행한다.

[0016] 여기서, 상기 제2 합성곱 신경망은, 상기 인식하고자 하는 행동 정보와 유사한 객체 인식 데이터셋에 학습된 것이다.

[0017] 여기서, 상기 프로세서는, 어텐션 함수(Attention Function)를 이용한 연산을 수행하여 시공간 영역의 상기 제1 특징 텐서를 기 설정된 크기의 제1 특징 벡터로 추출하는 단계 및 어텐션 함수(Attention Function)를 이용한 연산을 수행하여 시공간 영역의 상기 제2 특징 텐서를 기 설정된 크기의 제2 특징 벡터로 추출하는 단계를 수행한다.

[0018] 여기서, 상기 프로세서는, 상기 제1 특징 벡터 또는 상기 제2 특징 벡터의 차원 변환을 통해 연산이 가능하도록 변환하는 클래스 변환 연산을 수행하는 단계 및 상기 제1 특징 벡터와 클래스 변환 연산을 수행한 상기 제2 특징 벡터를 합하여 시공간 영역에 해당하는 특징을 추출하기 위한 멀티태스크 트랜스포머 유닛을 이용한 트랜스포머 연산을 수행하여 결합 특징 벡터를 추출하는 단계를 수행한다.

[0019] 여기서, 상기 프로세서는, 상기 제1 특징 벡터를 입력받고, 상기 입력된 제1 특징 벡터와 미리 결정된 연결 가중치에 따른 풀리 커넥티트 특징값을 생성하는 단계, 상기 선택된 프레임 영상을 컨볼루션 변환을 통해 컨볼루션 특징값을 생성하는 단계, 상기 풀리 커넥티트 특징값과 상기 컨볼루션 특징값의 행렬 곱 연산을 수행하는 단계 및 상기 풀리 커넥티트 특징값과 상기 행렬 곱 연산을 수행한 상기 컨볼루션 특징값을 합하여 정규화를 수행하는 단계를 수행한다.

[0020] 여기서, 상기 트랜스포머 연산은, 제1 트랜스포머 연산 내지 제3 트랜스포머 연산을 포함하며, 상기 프로세서는, 상기 제3 트랜스포머 연산을 수행하여 추출한 상기 결합 특징 벡터를 풀리 커넥티드 레이어(fully-connected layer)를 이용하여 장소와 행동을 분류하는 단계를 수행한다.

발명의 효과

[0021] 이상에서 설명한 바와 같이 본 발명의 실시예들에 의하면, 프로세서가 분석 대상 영상을 입력 받고, 시간 영역을 기준으로 기 설정된 구간별로 상기 분석 대상 영상에서 일부의 프레임 영상을 선택하고, 상기 선택된 프레임 영상에서 장소와 행동을 인식하여 인식한 장소와 행동에 따른 특징값을 상기 선택된 프레임 영상에 라벨링하여 비정제 동영상에서 클립 단위 장소와 행동 정보로 학습한 인공신경망으로 장소와 행동이 어느 프레임의 어느 공간 영역에서 나타나고 있는지 찾을 수 있다.

[0022] 또한, 장소와 행동의 영역을 함께 도출해내는 단일 인공신경망을 사용하며, 서로 다른 2가지 이상의 태스크(task)를 해결하는 멀티태스킹(multitasking) 방법을 이용하여, 서로 연관성이 높은 장소와 행동의 영역을 함께 도출할 수 있다.

[0023] 여기에서 명시적으로 언급되지 않은 효과라 하더라도, 본 발명의 기술적 특징에 의해 기대되는 이하의 명세서에서 기재된 효과 및 그 잠정적인 효과는 본 발명의 명세서에 기재된 것과 같이 취급된다.

도면의 간단한 설명

[0024] 도 1은 본 발명의 일 실시예에 따른 행동 인식 장치의 블록도이다.

도 2는 본 발명의 일 실시예에 따른 행동 인식 장치의 프로세서를 설명하기 위한 도면이다.

도 3은 본 발명의 일 실시예에 따른 행동 인식 장치 및 방법의 인공 신경망 구조를 나타낸 것이다.

도 4는 본 발명의 일 실시예에 따른 행동 인식 장치 및 방법의 라벨링을 예로 들어 나타낸 것이다.

도 5는 본 발명의 일 실시예에 따른 행동 인식 장치 및 방법의 MTx 연산 구조를 나타낸 것이다.

도 6은 본 발명의 일 실시예에 따른 행동 인식 장치 및 방법의 AQPr 연산 구조를 나타낸 것이다.

도 7은 본 발명의 일 실시예에 따른 행동 인식 장치 및 방법을 이용한 실험 결과를 나타낸 것이다.

도 8 내지 도 10은 본 발명의 일 실시예에 따른 행동 인식 방법을 이용한 나타낸 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0025] 이하, 본 발명에 관련된 인공 신경망 기반의 비정제 동영상에서의 행동 인식 방법 및 장치에 대하여 도면을 참조하여 보다 상세하게 설명한다. 그러나, 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며, 설명하는 실시예에 한정되는 것이 아니다. 그리고, 본 발명을 명확하게 설명하기 위하여 설명과 관계없는 부분은 생략되며, 도면의 동일한 참조부호는 동일한 부재임을 나타낸다.

[0026] 이하의 설명에서 사용되는 구성요소에 대한 접미사 "모듈" 및 "부"는 명세서 작성의 용이함만이 고려되어 부여되거나 혼용되는 것으로서, 그 자체로 서로 구별되는 의미 또는 역할을 갖는 것은 아니다.

[0027] 본 발명은 인공 신경망 기반의 비정제 동영상에서의 행동 인식 방법 및 장치에 관한 것이다.

[0028] 도 1은 본 발명의 일 실시예에 따른 행동 인식 장치의 블록도이다.

[0029] 도 1을 참조하면, 본 발명의 일 실시예에 따른 행동 인식 장치(1)는 프로세서(10), 영상 획득부(20), 메모리(30), I/O 인터페이스(40)를 포함한다.

[0030] 본 발명의 일 실시예에 따른 행동 인식 장치(1)는 비정제 동영상에서 행동을 인식하기 위한 장치로서, 딥 러닝 구조 중 메모리 네트워크의 한 종류인 Multitask Transformer Network에서 사용되는 QKV (Query, Key, Value) 컨셉을 활용하여 비정제 동영상에서 클립 단위 장소와 행동 정보로 학습한 인공신경망으로 장소와 행동이 어느 프레임의 어느 공간 영역에서 나타나고 있는지 찾는 알고리즘을 이용한다.

[0031] 본 발명의 일 실시예에 따른 행동 인식 장치(1)는 비정제 동영상에서 클립 단위로 라벨링 되어있는 장소와 행동 정보로 학습된 인공 신경망을 활용하여 장소와 행동이 각각 나타나고 있는 구체적인 시공간 영역을 찾는 것을 목적으로 한다.

[0032] 프로세서(10)는 인공신경망을 기반으로 상기 분석 대상 영상으로부터 장소와 행동을 인식한다.

[0033] 본 발명의 일 실시예에 따른 행동 인식 장치(1)의 프로세서(10)는 적은 정보량을 가진 데이터로 학습을 한 후, 더 자세하고 구체적인 정보를 생성해내도록 학습하는 방법을 약지도학습(Weakly supervised learning)을 사용한다. 딥 러닝에서 사용되는 약지도학습의 대표적인 예로 CAM (Class Activation Map)이 있다. 이 방법은, image-level로 라벨링 되어있는 데이터로 인공 신경망을 학습한 후 pixel-level로 결과를 도출해내는 방법이다. 학습할 때 인공신경망 구조는, 컨벌루션 필터를 통해 추출된 특징(feature) 텐서(tensor)를 분류(classifying)하는 마지막 레이어(layer)의 구성을 GAP (Global Average Pooling)과 FC (Fully Connected) Layer로 구성한다. 그러면, 특징 텐서가 GAP를 통과하여 공간(spatial)영역으로 pooling되어 특징 벡터(vector)가 되며, FC layer가 이 특징 벡터를 분류하게 되기 때문에 이미지-레벨(image-level)으로 학습이 된다. 픽셀-레벨(Pixel-level)로 결과를 도출할 때는, 인공 신경망 구조에서 GAP를 제거한 구조로 진행이 된다. 이 경우, 컨벌루션 필터를 통해 추출된 특징 텐서가 그대로 FC layer를 거치게 되므로, 특징 텐서의 공간 영역별로 존재하는 특징 벡터 각각이 FC layer를 통해 분류가 되기 때문에 공간 영역마다(pixel-level) 분류를 진행할 수 있게 된다.

[0034] 또한, 공간 영역으로 정보를 확장하던 방법을 시간영역으로 변경한 T-CAM (Temporal Class Activation Map)방법도 존재한다. 본 방법은, 이미지-레벨(image-level)로 학습하던 인공신경망 구조에서 공간영역으로 풀링(pooling)하던 GAP를, 비디오-레벨(video-level)로 학습이 되도록 시간(temporal) 영역으로 풀링(pooling)이 되도록 GAP를 구성을 한 후, CAM과 같은 방식으로 결과를 도출할 때 GAP를 제거하여 시간마다 분류가 되도록 하는 구조이다.

[0035] 이외에도 CAM은 딥 러닝을 활용한 약지도학습 방식에 대표적으로 많이 사용되고 있다. 본 발명의 일 실시예에

따른 행동 인식 장치(1)의 프로세서(10)는, CAM과는 다른 성격을 많이 띄고 있는 약지도학습법으로 비디오-레벨(video-level) 정보를 시공간(spatio-temporal) 영역으로 구체화(또는 지역화, localization) 하는 방법을 사용한다. 본 발명의 일 실시예에 따른 행동 인식 장치(1)의 프로세서(10)는 인공신경망 구조를 멀티태스크 트랜스포머 네트워크(Multitask Transformer Network)를 기반으로 진행한다. 이 멀티태스크 트랜스포머 네트워크(Multitask Transformer Network)는 액션 트랜스포머 네트워크(Action Transformer Network)가 비디오에서 행동(action)만을 분류하던 인공신경망 구조를 장소(scene)과 행동이 함께 분류가 되도록 확장한 인공신경망 구조이다.

[0036] 액션 트랜스포머 네트워크(Action Transformer Network)는 트랜스포머 네트워크(Transformer Network)가 자연어 처리 (언어 번역)에 사용되던 것을 비디오에 적용이 가능하도록 인공 신경망 구조를 바꾼 것이다. 이 트랜스포머 네트워크(Transformer Network)는 메모리 네트워크(Memory Network)의 구조 중 QKV (Query, Key, Value) 컨셉을 활용하여 고성능의 자연어 처리가 가능하도록 구성한 인공신경망 구조이다.

[0037] CAM의 경우 인공신경망 구조가 Convolution Filter - GAP - FC 의 구조로, 마지막 GAP-FC의 구조가 필수적으로 사용되어야 한다. 따라서 인공신경망 구조 설계에 제한이 많으며, 이는 학습할 image(video)-level 분류에서부터 높은 성능을 기대하기 어렵다. 낮은 성능의 image(video)-level 분류는 결국 최종적으로 해내어야 할 시공간 영역의 지역화 성능 역시 낮을 수밖에 없다. 하지만 QKV 컨셉의 경우 인공신경망 구조의 어디에도 사용될 수 있으며, 이러한 높은 설계에서의 자유도는 높은 성능을 도출해내며, 최근 많이 사용되고 있다. 또한, CAM은 학습할 때와 결과를 도출해낼 때 인공신경망 구조가 다르기 때문에 (학습할 때 존재하였던 GAP를 결과 도출 시엔 없앰), 좋은 성능을 기대하기 어렵다. QKV 컨셉을 사용한 약지도학습법의 경우엔, 학습할 때와 결과를 도출해낼 때 같은 구조의 인공신경망을 사용하기 때문에 잘 학습된 인공신경망의 경우 좋은 성능을 기대할 수 있다.

[0038] 영상 획득부(20)는 외부로부터 인식하고자 하는 행동 정보가 포함된 분석 대상 영상을 획득한다. 별도의 입력부를 통해 분석 대상 영상을 획득하면, 프로세서를 통해 분석 대상 영상에서의 장소와 행동을 인식하게 된다.

[0039] 메모리(30)는 프로세서(10)의 처리 및 제어를 위한 프로그램들(하나 이상의 인스트럭션들)을 저장할 수 있다.

[0040] I/O 인터페이스(40)는 시스템 또는 장비를 연결 할 수 있는 연결매체를 장착할 수 있는 장치로서 본 발명에서는 영상 획득부와 프로세서를 연결한다.

[0041] 도 2는 본 발명의 일 실시예에 따른 행동 인식 장치의 프로세서를 설명하기 위한 도면이다.

[0042] 도 2를 참조하면, 본 발명의 일 실시예에 따른 행동 인식 장치(1)의 프로세서(10)는 장소 인식 특징 추출부(100), 객체 인식 특징 추출부(200), AQPr 연산부(300), CCM 연산부(400), MTx 연산부(500), 특징값 분류부(600)를 포함한다.

[0043] 본 발명의 일 실시예에 따른 행동 인식 장치(1)의 프로세서(10)에서 사용되는 인공신경망 구조는 하기 도 3에서 상세히 설명한다.

[0044] 프로세서(10)는 기능에 따라 복수 개의 모듈들로 구분될 수도 있고, 하나의 프로세서에서 기능들을 수행할 수도 있다.

[0045] 프로세서(10)는 인공신경망을 기반으로 상기 분석 대상 영상으로부터 장소와 행동을 인식한다. 구체적으로, 상기 분석 대상 영상에서 기 설정된 구간별로 선택된 프레임을 선택하고, 선택된 프레임에서 장소와 행동을 인식하며, 인식한 장소와 행동에 따른 특징값을 상기 선택된 프레임 별로 라벨링한다.

[0046] 장소 인식 특징 추출부(100)는 제1 합성곱 신경망을 이용하여 상기 선택된 프레임의 장소 인식을 위한 제1 특징 텐서를 추출한다.

[0047] 객체 인식 특징 추출부(200)는 제2 합성곱 신경망을 이용하여 상기 선택된 프레임의 객체 인식을 위한 제2 특징 텐서를 추출한다.

[0048] 여기서, 상기 제2 합성곱 신경망은, 상기 인식하고자 하는 행동 정보와 유사한 객체 인식 데이터셋에 학습된 것이다.

[0049] AQPr 연산부(300)는 제1 AQPr 연산부(300a)와 제2 AQPr 연산부(300b)를 포함한다. 제1 AQPr 연산부(300a)는 어텐션 함수(Attention Function)를 이용한 연산을 수행하여 시공간 영역의 상기 제1 특징 텐서를 기 설정된 크기의 제1 특징 벡터로 추출한다.

[0050] 제2 AQPr 연산부(300b)는 어텐션 함수(Attention Function)를 이용한 연산을 수행하여 시공간 영역의 상기 제2 특징 텐서를 기 설정된 크기의 제2 특징 벡터로 추출한다.

[0051] CCM 연산부(400)는 상기 제1 특징 벡터 또는 상기 제2 특징 벡터의 차원 변환을 통해 연산이 가능하도록 클래스 변환 연산을 수행한다.

[0052] MTx 연산부(500)는 상기 제1 특징 벡터와 클래스 변환 연산을 수행한 상기 제2 특징 벡터를 합하여 시공간 영역에 해당하는 특징을 추출하기 위한 멀티태스크 트랜스포머 유닛을 이용한 트랜스포머 연산을 수행하여 결합 특징 벡터를 추출한다.

[0053] MTx 연산부(500)는 제1 MTx 연산부 내지 제3 MTx 연산부를 포함하여 제1 트랜스포머 연산 내지 제3 트랜스포머 연산을 수행하고, 특징값 분류부(600)는 상기 제3 트랜스포머 연산을 수행하여 추출한 상기 결합 특징 벡터를 풀리 커넥티드 레이어(fully-connected layer)를 이용하여 장소와 행동을 분류한다.

[0054] 도 3은 본 발명의 일 실시예에 따른 행동 인식 장치 및 방법의 인공 신경망 구조를 나타낸 것이다.

[0055] 본 발명의 일 실시예에 따른 행동 인식 장치 및 방법은 비 정제 동영상에서 약 지도 학습법으로 장소와 행동이 나타나는 시공간 영역을 찾는 방법을 제안한다.

[0056] 본 발명의 일 실시예에 따른 행동 인식 장치(1)의 프로세서(10)에서 사용되는 인공신경망 구조인 멀티태스크 트랜스포머 네트워크(Multitask Transformer Network)는 QKV 컨셉을 사용하고 있으며, 이 QKV컨셉을 활용하여 학습할 땐 video-level의 장소 및 행동 정보로 학습하지만, 결과는 장소와 행동이 시공간 영역 중 어디에 해당하는지를 도출해낸다.

[0057] 또한, 본 발명은 장소와 행동의 영역을 함께 도출해내는 단일 인공신경망을 사용한다. 이렇게 서로 다른 2가지 이상의 태스크(task)를 해결하는 방법을 멀티태스킹(multitasking)이라 하는데, 각각의 task인 장소와 행동은 서로 연관성이 높기 때문에, 장소와 행동 각각 독립적인 인공신경망을 사용하는 것 보다 더 좋은 성능을 기대할 수 있다.

[0058] 본 발명의 일 실시예에 따른 행동 인식 장치(1)의 프로세서(10)에서 사용되는 멀티태스크 트랜스포머 네트워크(Multitask Transformer Network)는 도 3에 나타난 바와 같이, 비 정제 동영상의 일정 구간인 segment를 입력으로 받고, 해당 segment가 어떤 장소와 행동인지를 출력으로 내놓는다. 입력인 segment의 경우, 비 정제 동영상 전체를 한번에 입력으로 사용할 경우, 딥러닝 가속화를 위한 병렬연산에 사용될 그래픽카드의 메모리의 한계 때문에 5초 동안의 32프레임, 즉 6.4fps 의 간격으로 32개의 RGB 이미지가 입력으로 들어가게 된다. 따라서, 도 3의 Multitask Transformer Network는 해당 5초짜리 segment가 어떤 장소와 행동을 나타내는지를 출력으로 내며, 전체 비 정제 동영상에서는 5초 간격으로 장소와 행동이 무엇인지를 나타내준다.

[0059] Multitask Transformer Network의 학습은 딥 러닝에서 일반적으로 사용하는 Stochastic Gradient Descent (SGD) 방식을 따라 진행되며, 실험적으로 본 발명의 성능을 보이기 위해 학습에 사용된 데이터 셋은 CoVieW 2019 dataset이 사용되는 것이 바람직하다. CoVieW 2019 dataset은 비 정제 동영상에서 5초 간격으로 장소(scene), 행동(action), 그리고 해당 5초짜리 segment가 그 동영상에서 얼마나 중요한지에 대한 점수(importance score) 이렇게 3개가 라벨링 되어있는 데이터셋이다. CoVieW 2019 데이터셋에서 하나의 비 정제 동영상 에서의 라벨링 예시는 하기 도 4와 같다.

[0060] 도 3의 인공신경망 구조를 자세히 살펴보면 다음과 같다. 먼저, 장소 인식 특징 추출부(100)와 객체 인식 특징 추출부(200)에서 Places365와 ImageNet dataset에 각각 학습된 Places365 2D CNN, ImageNet 2D CNN을 사용하여 segment의 프레임의 특징을 추출해낸다. 이 때 사용하는 2D CNN은 ResNet18을 사용하며, 이는 (높이×너비×RGB 채널)으로 **H×W×3** 크기의 RGB (3채널) 이미지를 입력으로 받고 출력으로 **H/32×W/32×512** 크기의 특징 텐서를 출력으로 내놓는 인공신경망 (2D CNN) 이다. 본 발명에서는 segment (시간길이×높이×너비×RGB채널)으로 **32×224×224×3**을 입력으로 넣어 **32×7×7×512**의 크기의 특징을 각각의 2D CNN에서 추출해낸다.

[0061] 여기서 Places365 2D CNN는 장소 인식 데이터셋인 Places365에 학습되어 있으므로, 장소 관련 특징을 추출해내고, ImageNet 2D CNN은 객체 인식 데이터셋인 ImageNet에 학습 되어있으므로 객체 관련 특징을 추출해낸다. ImageNet 2D CNN의 경우, 우리는 최종적으로 행동을 인식할 것인데 행동 인식 데이터셋에 학습된 2D CNN이 아닌 객체 인식 데이터셋에 학습한 이유는, 2D CNN은 2-Dimensional CNN으로 입력으로 동영상이 아닌 이미지를 받기 때문에 이미지로 구성된 데이터셋을 사용해야 하는데, 이미지로 구성된 행동 인식 데이터셋이 존재하지 않아, 행동 인식과 가장 유사한 객체 인식 데이터셋에 학습된 2D CNN을 사용하였다. 그리고, 2D CNN을 미리 다른 데이

터셋에 학습하여 사용한 이유는, CoVieW 2019 데이터셋이 총 1500개의 비 정제 동영상을 제공하여 딥 러닝 구조를 학습하기엔 비교적 적은 양의 데이터를 제공하여 과적합 (overfitting)문제가 발생하기 때문에 비교적 많은 양의 데이터를 제공하는 Places365 (약 800만 장 이미지), ImageNet (약 120만 장 이미지)에 2D CNN을 미리 학습하여 일반적인 (general한) 특징을 추출해낼 수 있게 하였다.

[0062] 그 후, 도 3에 나타난 바와 같이 2D CNN으로 추출된 특징(Convolution feature)은 제1 AQPr 연산부(300a)와 제2 AQPr 연산부(300b)를 구현하는 AQPr (Attentional Query Processor)에 입력으로 들어가고, MTx 연산부(500)를 구현하는 MTx (Multitask Transformer units) 들의 입력으로도 사용된다. 2D CNN으로 추출된 **32×7×7×512**의 크기의 특징을 $X_{t,h,w} \in \mathbb{R}^{512}$ 으로 나타낼 때, AQPr에서의 수학식 1과 같이 진행된다.

수학식 1

[0063]
$$M_{t,h,w} = \sigma\left(\text{InstanceNorm}\left(W^{Attention} X_{t,h,w}\right)\right)$$

$$Y^{attention} = \sum_{t,h,w} \frac{M_{t,h,w} X_{t,h,w}}{\sum_{t,h,w} M_{t,h,w}}$$

[0064] 여기서, $Y^{attention} \in \mathbb{R}^{512}$ 이 AQPr의 출력이고, $\sigma(z) = \frac{1}{1+e^{-z}}$ 는 sigmoid function이고, $W^{Attention}$ 는 학습되는 weight (trainable parameter)이고 InstanceNorm은 Instance Normalization으로 이 역시 학습되는 weight를 포함한 정규화 (normalization) 방식 중 하나이다. 따라서, AQPr을 거치면 **32X7X7X512**의 크기의 시공간 영역의 특징 텐서가 **512** 크기의 특징 벡터로 추출되며, 이 특징 벡터와 도 3의 CCM 연산부(400)를 구현하는 CCM (class conversion matrix) 에서 추출된 특징 벡터가 더해져서 MTx의 query 입력으로 사용된다.

[0065] 도 3의 CCM의 경우, 입력을 $X^{CCM} \in \mathbb{R}^n$, 출력을 $Y^{CCM} \in \mathbb{R}^n$, **n**은 특징 벡터의 채널 (도 3에서 AQPr의 출력과 더해지는 부분은 **n=512** 이며, 나머지 "MTx->concat->"과 더해지는 부분은 **n=256**이다.) 이라 하였을 때, CCM에서의 연산은 수학식 2와 같이 진행된다.

수학식 2

[0066]
$$Y^{CCM} = \text{ReLU}\left(W^{CCM} X^{CCM} + b^{CCM}\right)$$

[0067] 여기서 $\text{ReLU}(z) = \max(0, z)$ 는 ReLU function이고, W^{CCM}, b^{CCM} 는 학습되는 weight이다.

[0068] 그리고 MTx의 연산은 하기 도 5와 같이 이루어 진다.

[0069] 도 3에서 실선은 "학습할 때 기울기가 전파되는 곳"을 나타내며, 회색 점선은 "학습할 때 기울기가 전파되지 않는 곳"을 나타낸다. 즉, 장소와 행동의 특징 (정보) 공유가 일어나는 부분 인 CCM에서 기울기를 전파해주지 않으므로, MTx 는 각각 본인이 맡고있는 문제(task)인 장소 또는 행동에 대해서만 학습을 하게 된다. 즉, 도 3에서 윗줄 민무늬로 표시된 MTx들은 장소에 대한 전문성을 띄게되고, 아랫줄 빗금 표시된 MTx들은 행동에 대한 전문성을 띄게 된다.

[0070] 도 3의 Multitask Transformer Network에 대해 부연 설명을 하면, 앞에서 잠깐 언급 했듯이, CCM에서 장소와 행동에 대한 정보 공유가 일어나며, AQPr은 시공간 영역의 특징 텐서를 하나의 특징 벡터로 합쳐주는 (feature aggregation) 역할을 하며, MTx는 하기 도 5에서의 연산과 같이, memory (2D CNN으로 추출된 특징 벡터) 중 보고자 하는 영역을 query 특징과 matrix multiplication (시공간 영역으로 각각 inner product 연산으로 진행됨) 연산을 통해 찾아내고, 그 시공간 영역에 해당하는 특징을 추출해내기 위해 한번 더 matrix multiplication 연산을 통해 추출해낸다. 즉, 첫번째 matrix multiplication으로는 보고자 하는 시공간 영역을 찾아내는

것이고, 두번째 matrix multiplication은 그 시공간 영역에 해당하는 특징을 추출해내는 역할을 한다. 그리고 뒤의 Layer Norm, FC 등 은 일반적인 MLP (multi-layer perceptron)과 같은 역할로, feature를 고차원으로 embedding 하는 역할을 한다.

[0071] 첫번째 Matrix Multiplication에서 장소/행동 인식과 관련된 찾고자 하는 정보가 담긴 특징 (Query, 1x512, Query Embedding인 FC를 거치면 1x128)와 가장 잘 매칭이 되는 시공간영역을 (Memory 32x7x7x512, Key Embedding을 거치면, 32x7x7x128) 에서 찾고자 Matrix Multiplication을 통해 32x7x7x1의 특징을 추출해내어 시공간 영역으로 장소/행동이 존재할 확률을 얻어내고 두번째 Matrix Multiplication에서는 시공간영역 32x7x7에 해당하는 장소/행동 특징을 추출해내기 위해 위에서 추출된 32x7x7x1 과 Value Embedding을 통해 추출된 32x7x7x128과 Matrix Multiplication을 통해 1x128 크기의 특징을 추출해낸다.

[0072] 즉, Query(1x512, FC를 거치면 1x128)는 이전에 추출된 (MTx의 입력으로 들어가는) 장소/행동에 대한 특징이며, Key Embedding(32x7x7x128)에서 추출되는 특징은 Query와 매칭을 하기위해 추출하는 시공간 특징이며, Value Embedding(32x7x7x128)에서 추출되는 특징은 행동/장소에 대한 시공간 영역마다의 특징이다.

[0073] 도 4는 본 발명의 일 실시예에 따른 행동 인식 장치 및 방법의 라벨링을 예로 들어 나타낸 것이다.

[0074] Multitask Transformer Network의 학습은 딥 러닝에서 일반적으로 사용하는 Stochastic Gradient Descent (SGD) 방식을 따라 진행되며, 실험적으로 본 발명의 성능을 보이기 위해 학습에 사용된 데이터 셋은 CoVieW 2019 dataset이 사용된다. CoVieW 2019 dataset은 비 정제 동영상에서 5초 간격으로 장소(scene), 행동(action), 그리고 해당 5초짜리 segment가 그 동영상에서 얼마나 중요한지에 대한 점수 (importance score) 이렇게 3개가 라벨링 되어있는 데이터셋이다. CoVieW 2019 데이터셋에서 하나의 비 정제 동영상 에서의 라벨링 예시는 도 4와 같다.

[0075] 도 5는 본 발명의 일 실시예에 따른 행동 인식 장치 및 방법의 MTx 연산 구조를 나타낸 것이다.

[0076] MTx 연산부(500)는 상기 제1 특징 벡터와 클래스 변환 연산을 수행한 상기 제2 특징 벡터를 합하여 멀티태스크 트랜스포머 유닛을 이용한 트랜스포머 연산을 수행하여 결합 특징 벡터를 추출한다.

[0077] MTx 연산부(500)의 멀티태스크 트랜스포머 유닛은, 쿼리 입력부(510), 메모리 입력부(520), 적어도 하나의 풀리 커넥티드 레이어(fully-connected layer)(530)를 포함한다.

[0078] MTx 연산부(500)는, 쿼리 입력부(510)로 상기 제1 특징 벡터를 입력받고, 상기 입력된 제1 특징 벡터와 미리 결정된 연결 가중치에 따른 풀리 커넥티트 특징값을 생성한다.

[0079] 또한, 메모리 입력부(520)로 입력된 선택된 프레임 영상을 컨볼루션 변환을 통해 컨볼루션 특징값을 생성한다.

[0080] 풀리 커넥티트 특징값과 상기 컨볼루션 특징값의 행렬 곱 연산을 수행하고, 상기 풀리 커넥티트 특징값과 상기 행렬 곱 연산을 수행한 상기 컨볼루션 특징값을 합하여 정규화를 수행한다.

[0081] 도 5에서 쿼리 입력부(510)를 구현하는 쿼리(Query)는 도 3의 MTx 블록의 왼쪽에서 들어오는 입력이며, Memory는 MTx 블록의 위 또는 아래에서 들어오는 입력이다. FC는 fully connected layer를 나타내며, 1x1x1 Conv는 **1x1x1** Convolution 을 나타내며, Layer Norm은 Layer Normalization 으로, 앞에서 언급한 InstanceNorm과 비슷하게, 학습되는 weight를 포함한 정규화 (normalization) 방식 중 하나이다. Softmax는 softmax function을 나타낸다. Dropout은 과적합 (overfitting) 문제를 완화하기 위해 사용되었다.

[0082] $\otimes$ 는 matrix multiplication 연산자를, $\oplus$ 는 element-wise sum (일반적인 덧셈) 연산자를 나타낸다. 그리고 빨간색 화살표의 경우, 특별한 연산이 따로 있는 것이 아니며, 추후 설명할 본 발명의 약 지도 학습법에서 사용할 특징 텐서가 빨간색 화살표에서 추출된 특징 텐서를 사용할 것임을 나타낸다. 도 5의 MTx 에서 최종적으로 출력되는 맨 오른쪽 화살표에선, $\mathbb{R}^{128}$ 의 크기를 갖는 특징 벡터가 추출이 된다.

[0083] 이후, 상기 도 3에 나타난 바와 같이, 2개의 MTx에서 추출된 특징이 "concat"을 거치게 되는데, 이는 concatenation의 줄임말로, 두 개의 $\mathbb{R}^{128}$ 크기 벡터가 단순히 연결되어 $\mathbb{R}^{256}$ 의 크기가 됨을 나타낸다.

[0084] 그리고 특징값 분류부(600)는 상기 도 3에서 2개씩 쌓인 MTx를 총 3번 걸쳐 추출된 특징 벡터가 최종적으로 하나의 FC (fully connected layer)의 입력으로 들어가게 되고, 이를 통해 장소(scene), 행동(action)을 분류해

낸다.

[0085] 상기 도 3의 Multitask Transformer Network는 5초짜리 segment (동영상 클립)의 장소와 행동을 분류하는 인공신경망으로, 학습할 때 동영상 5초마다의 해당 장소와 행동의 정보만 필요로 한다. 그리고 이렇게 학습된 인공신경은 하기 도 5의 화살표 부분에서 어느 시공간 영역을 보아야 하는지 사람이 정보를 주지 않더라도, 장소와 행동이 무엇인지 잘 찾아낼 수 있도록 학습이 된다. 따라서 이 빨간색 화살표에서 추출되는 특징 텐서는 장소와 행동이 나타나는 시공간 영역을 나타낼 것이며, 이를 위한 학습 정보를 주지 않았으므로, 이 정보로 장소와 행동이 나타내는 영역을 표현한다면 이는 약 지도 학습법이 된다. (비교적 쉬운 정보인 5초마다의 행동, 장소 정보만으로 학습 후 장소, 행동이 나타나는 시공간 영역을 표현하므로)

[0086] 정확한 표현을 위해, 화살표에서 추출되는 특징 텐서를 $Y \in \mathbb{R}^{THW\times 1}$, query가 FC를 거쳐 추출된 Query Embedding 특징 벡터를 $Q \in \mathbb{R}^{128\times 1}$, memory가 1x1x1 Conv를 거쳐 추출된 Key Embedding 특징 텐서를 $K \in \mathbb{R}^{THW\times 128}$ 으로 표현할 때 수학식 3과 같이 계산된다 (T는 time, H는 height, W는 width으로 처음에 설명한 T=32, H=7, W=7이 사용된다)

수학식 3

[0087] $$Y = \mathrm{softmax}(KQ)$$

[0088] 여기서 softmax는 softmax function이며, 특징 텐서 Y의 전체 합이 1이 되도록 한다. 따라서, 시공간 영역 $\mathbb{R}^{THW}$ 으로 각각 상수값(scalar) 를 하나씩 갖게 되며, 이는 시공간 영역의 장소 또는 행동이 나타나는 정도를 띄게 된다.

[0089] 하지만 이 **T×H×W=32×7×7** 는 2D CNN에서 추출된 크기로, 원본 이미지의 높이(H)와 너비(W)의 비해 1/32 배 된 크기이다. 따라서 이 $Y \in \mathbb{R}^{THW\times 1}$ 를 bilinear interpolation 방식으로 높이와 너비를 32배 늘려준 후, 원본 이미지 크기로 맞춰주고 visualization을 위해 값이 낮은 곳은 파란색, 값이 높은 곳은 빨간색으로 표현하면 하기 도 7과 같은 결과를 얻을 수 있다.

[0090] 마지막으로, 본 발명에서 하고자 하는 MTx에서 Softmax결과로 출력되는 32(시간)×7(높이)×7(너비) 마다의 중요도를 원본 크기인 32(시간)×224(높이)×224(너비) 의 크기로 늘리는 방법은 bilinear interpolation 방식을 사용한다.

[0091] CNN이 높너비를 1/2 크기로 줄이는 pooling을 5번 사용하기 때문에 input image의 높너비 보다 1/32배 크기를 갖는 특징을 추출하게 된다. 따라서 본 발명에서 사용하는 특징인 32×7×7마다의 중요도는 시간방향으로는 압축되지 않고, 공간 방향으로만 1/32배 압축된 크기이다. 시간방향으로는 크기 변화가 없기 때문에 하나의 frame인 224×224 입장에서만 보았을때, 7×7 특징의 각 한칸마다는 224×224의 image에서 32×32의 크기를 담당하게 된다. 이를 보기 좋게 visualization 하기 위해서 7×7 특징을 224×224으로 키우게 되는데, interpolation 방식은 무엇을 쓰던지 상관 없이 7×7 크기의 특징을 224×224으로만 키우기만 하면 됩니다. 저희는 관련 약지도 학습법인 CAM에서 사용한 방식인 bilinear ineterpolation 방식을 채택한다.

[0092] 도 6은 본 발명의 일 실시예에 따른 행동 인식 장치 및 방법의 AQPr 연산 구조를 나타낸 것이다.

[0093] 상기 도 3에서 나와있듯이 2D CNN으로 추출된 특징 (Convolution feature)은 AQPr (Attentional Query Processor)에 입력으로 들어가고, MTx (Multitask Transformer units) 들의 입력으로도 사용된다. 2D CNN으로 추출된 **32×7×7×512**의 크기의 특징을 $X_{t,h,w} \in \mathbb{R}^{512}$ 으로 나타낼 때, AQPr에서의 상기 수학식 1과 같이 진행된다.

[0094] AQPr을 거치면 **32×7×7×512**의 크기의 시공간 영역의 특징 텐서가 **512** 크기의 특징 벡터로 추출되며, 이 특징 벡터와 상기 도 3의 CCM (class conversion matrix) 에서 추출된 특징 벡터가 더해져서 MTx의 query 입력으로

사용된다.

[0095] 도 7은 본 발명의 일 실시예에 따른 행동 인식 장치 및 방법을 이용한 실험 결과를 나타낸 것이다.

[0096] 도 7은 2가지 segment (video clip)에 대한 실험 결과를 나타낸다. Input video는 원본 이미지를 나타내며, scene은 장소의 시공간 영역, action은 행동의 시공간 영역을 나타낸다. 2가지 segment 중 위의 실험 결과에선 장소가 대부분의 영역에 퍼져서 골고루 나타나며, 행동은 사람에게만 영역이 집중된 것을 볼 수 있다. 이는 원래 해당 segment가 나타내는 concert라는 장소와 performance라는 행동을 잘 나타내는 것을 확인할 수 있다. 아래 segment 결과에 대해서는, 장소와 행동이 나타나는 영역이 일부 특정 시간에서만 집중되어 있으며, 나타내는 영역도 특정 구역에 집중되어 있는 것을 확인할 수 있다. 이는, 특정 프레임의 특정 영역에서 장소 또는 행동이 눈에 띄게 잘 나타나는 경우에, 그 특정 프레임에서만 시공간 영역이 집중되는 결과가 나타나는 것이다. 따라서, 행동인 cycling은 자전거 타는 사람에게 영역이 잘 나타났으며, 장소인 park의 경우 실내 공간에서는 시공간 영역이 나타나지 않고, 야외에만 영역이 잘 표현된 것을 확인할 수 있다.

[0097] 도 8 내지 도 10은 본 발명의 일 실시예에 따른 행동 인식 방법을 이용한 나타낸 흐름도이다.

[0098] 도 8을 참조하면, 본 발명의 일 실시예에 따른 행동 인식 방법은 프로세서가, 분석 대상 영상을 입력 받고, 시간 영역을 기준으로 기 설정된 구간별로 상기 분석 대상 영상에서 일부의 프레임 영상을 선택하는 단계(S100), 상기 선택된 프레임 영상에서 장소와 행동을 인식하는 단계(S200) 및 인식한 장소와 행동에 따른 특징값을 상기 선택된 프레임 영상에 라벨링하는 단계(S300)를 포함한다.

[0099] 도 9를 참조하면, 선택된 프레임에서 장소와 행동을 인식하는 단계(S200)는, 단계 S210에서 제1 합성곱 신경망을 이용하여 상기 선택된 프레임의 장소 인식을 위한 제1 특징 텐서를 추출한다.

[0100] 단계 S220에서 제2 합성곱 신경망을 이용하여 상기 선택된 프레임의 객체 인식을 위한 제2 특징 텐서를 추출한다.

[0101] 여기서, 제2 합성곱 신경망은 상기 인식하고자 하는 행동 정보와 유사한 객체 인식 데이터셋에 학습된 것이다.

[0102] 단계 S230에서 어텐션 함수(Attention Function)를 이용한 연산을 수행하여 시공간 영역의 상기 제1 특징 텐서를 기 설정된 크기의 제1 특징 벡터로 추출한다.

[0103] 단계 S240에서 어텐션 함수(Attention Function)를 이용한 연산을 수행하여 시공간 영역의 상기 제2 특징 텐서를 기 설정된 크기의 제2 특징 벡터로 추출한다.

[0104] 단계 S250에서 상기 제1 특징 벡터 또는 상기 제2 특징 벡터의 차원 변환을 통해 연산이 가능하도록 변환하는 클래스 변환 연산을 수행한다.

[0105] 단계 S260에서 상기 제1 특징 벡터와 클래스 변환 연산을 수행한 상기 제2 특징 벡터를 합하여 시공간 영역에 해당하는 특징을 추출하기 위한 멀티태스크 트랜스포머 유닛을 이용한 트랜스포머 연산을 수행하여 결합 특징 벡터를 추출한다.

[0106] 단계 S260에서 상기 트랜스포머 연산은, 제1 트랜스포머 연산 내지 제3 트랜스포머 연산을 포함하며,

[0107] 단계 S270에서 상기 제3 트랜스포머 연산을 수행하여 추출한 상기 결합 특징 벡터를 풀리 커넥티드 레이어(fully-connected layer)를 이용하여 장소와 행동을 분류한다.

[0108] 도 10을 참조하면, 멀티태스크 트랜스포머 유닛을 이용하여 트랜스포머 연산을 수행하는 단계(S260)는, 단계 S261에서 제1 특징 벡터를 입력받고, 상기 입력된 제1 특징 벡터와 미리 결정된 연결 가중치에 따른 풀리 커넥티트 특징값을 생성한다.

[0109] 단계 S262에서 상기 선택된 프레임 영상을 컨볼루션 변환을 통해 컨볼루션 특징값을 생성한다.

[0110] 단계 S263에서 풀리 커넥티트 특징값과 상기 컨볼루션 특징값의 행렬 곱 연산을 수행한다.

[0111] 단계 S264에서 풀리 커넥티트 특징값과 상기 행렬 곱 연산을 수행한 상기 컨볼루션 특징값을 합하여 정규화를 수행한다.

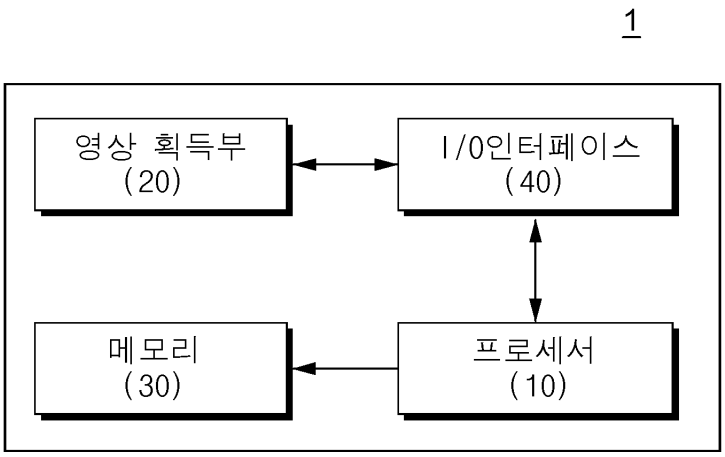
[0112] 본 발명의 일 실시예에 따른 행동 인식 장치 및 방법은 사람이 비 정제 동영상에서 장소와 행동이 일어나는 시공간 영역을 모두 찾기엔 많은 노동력을 필요로 하지만, 비디오 단위로(약 5초 단위) 장소와 행동을 라벨링 하는 것은 비교적 훨씬 쉬운 일이다. 본 발명에서는 QKV컨셉을 활용하여 비정제 동영상에서의 장소와 행동이 나타

나고 있는 시공간 영역을 비디오 단위의 장소와 행동 라벨링을 통해 찾는 약지도학습 방법을 제안하고 있다. 본 발명은 장소와 행동의 시공간 영역을 도출해내는 단일 인공신경망을 사용함으로써, 장소와 행동의 연관성을 통한 강인한 인공 신경망이 학습이 가능하여 장소 또는 행동의 결과만 도출해내는 인공신경망에 비해 좋은 성능을 기대할 수 있다.

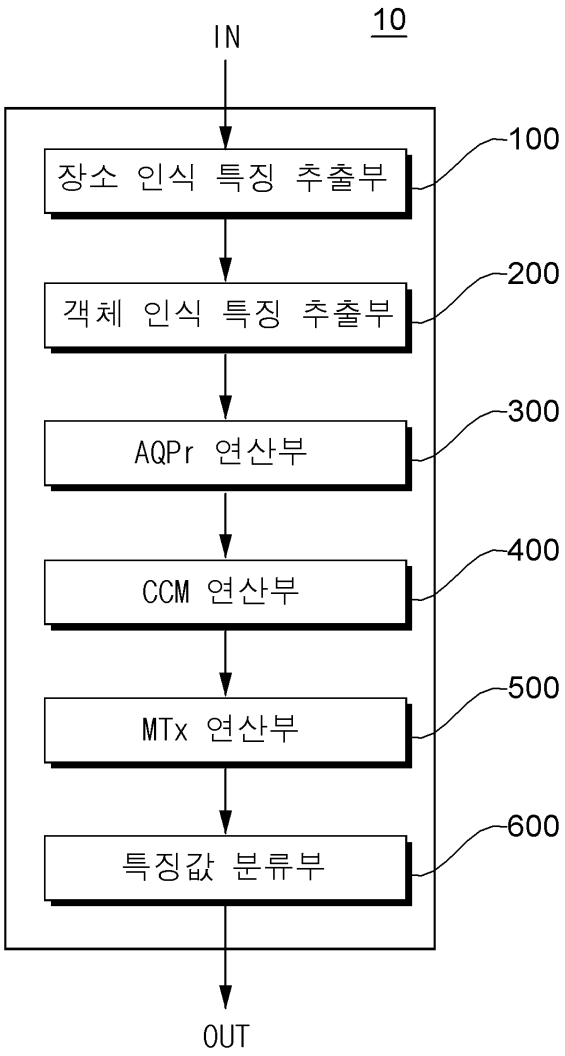
[0113] 이상의 설명은 본 발명의 일 실시예에 불과할 뿐, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명의 본질적 특성에서 벗어나지 않는 범위에서 변형된 형태로 구현할 수 있을 것이다. 따라서 본 발명의 범위는 전술한 실시예에 한정되지 않고 특허 청구 범위에 기재된 내용과 동등한 범위 내에 있는 다양한 실시 형태가 포함되도록 해석되어야 할 것이다.

도면

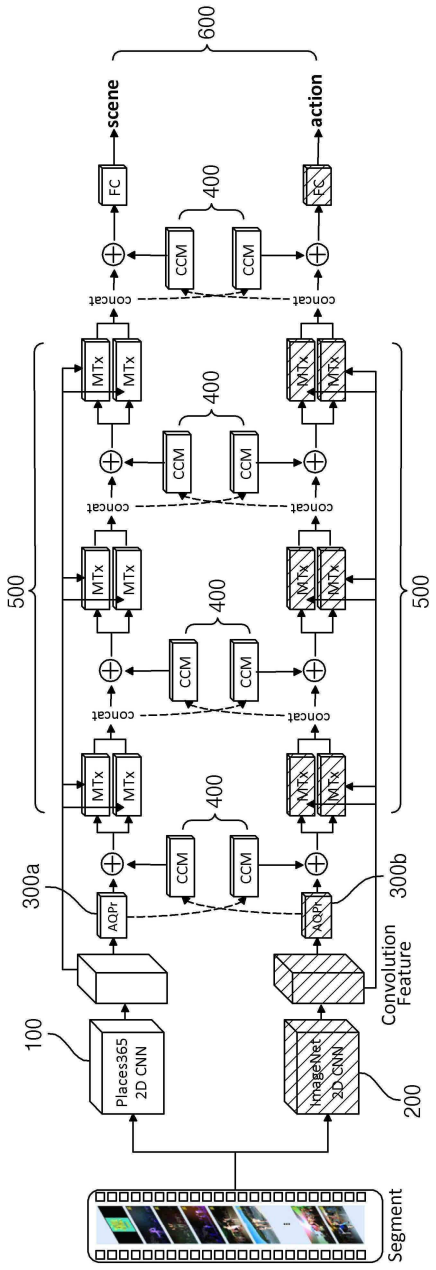
도면1



도면2


10
IN
장소 인식 특징 추출부
100
객체 인식 특징 추출부
200
AQPr 연산부
300
CCM 연산부
400
MTx 연산부
500
특징값 분류부
600
OUT

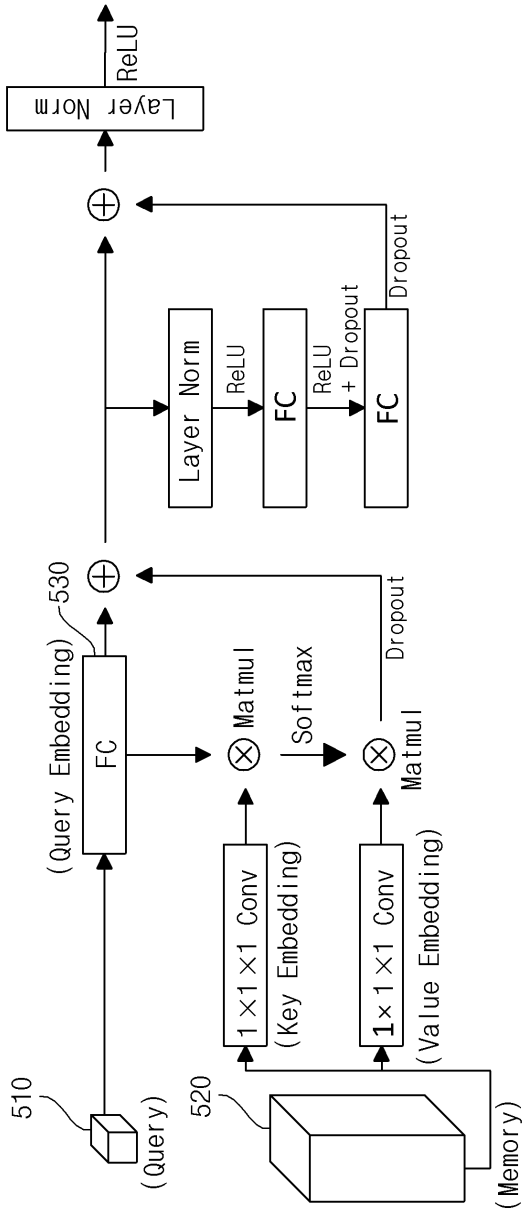
도면3


600
scene
action
FC
400
CCM
concat
MTx
500
300a
300b
AQP
Convolution Feature
100
Places365 2D CNN
ImageNet 2D CNN
200
Segment

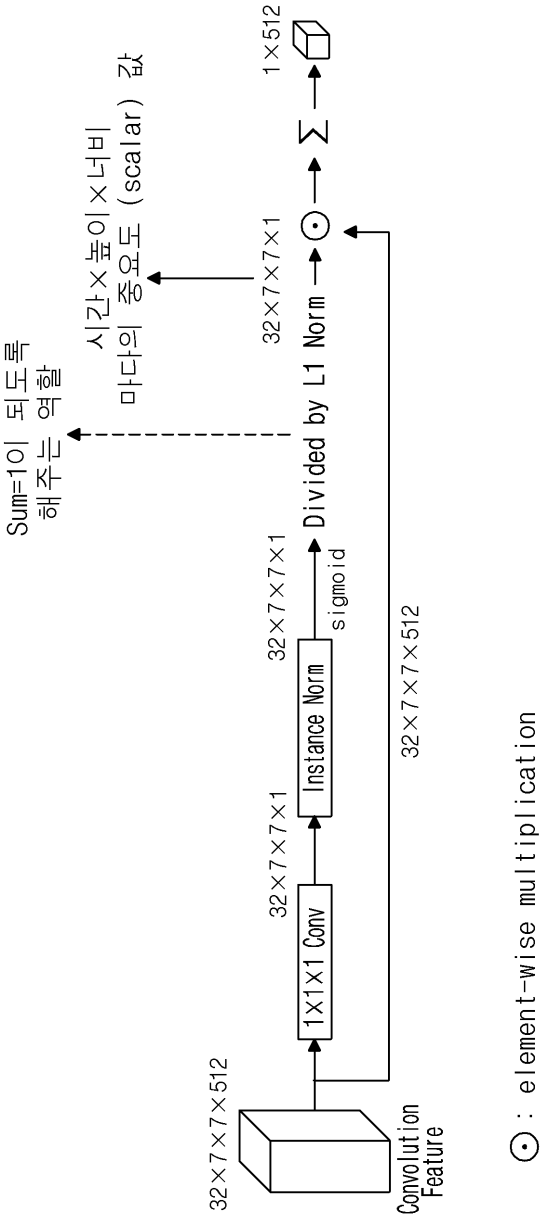
도면4


Time(sec)
0
5
10
15
20
25
30
Scene
Action
Importance Score
beach
no Person
0.8
concert
dance
1.4
concert
dance
1.4
pub
dance
1.2
concert
playing instrument
1.6
outdoor
etc
0.4

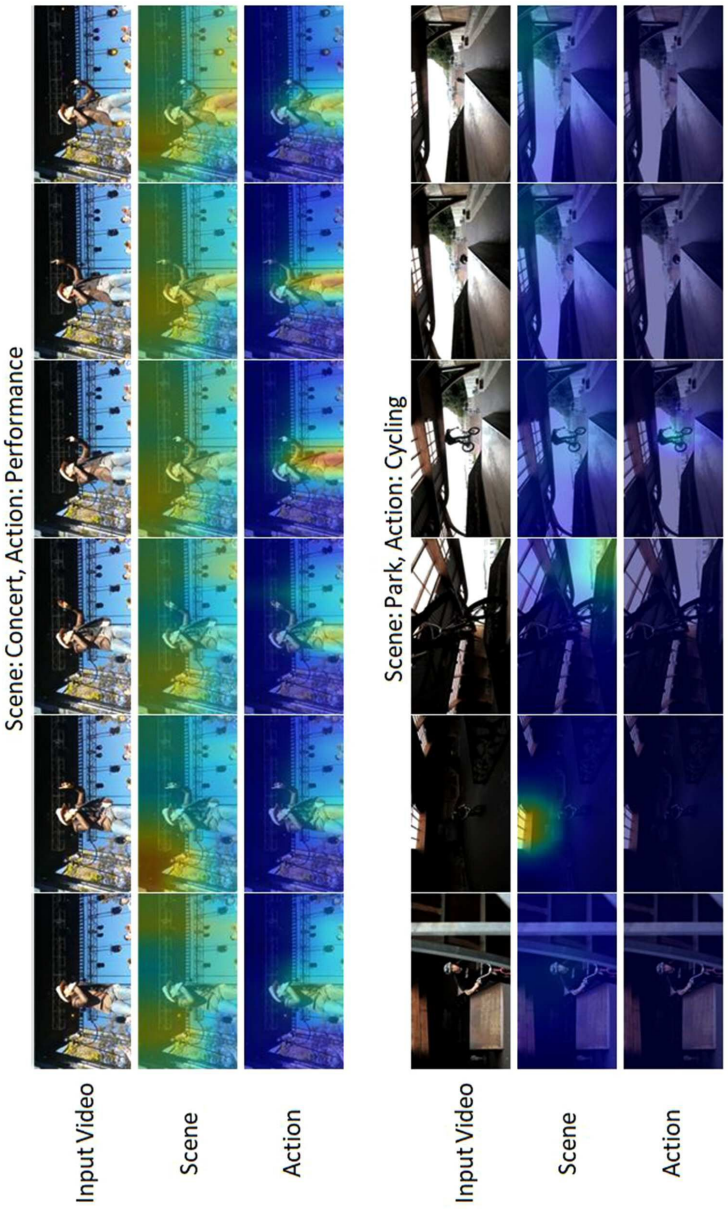
도면5


ReLU
Layer Norm
Dropout
FC
ReLU + Dropout
Layer Norm
ReLU
FC
530
(Query Embedding)
FC
Matmul
Softmax
Dropout
Matmul
1×1×1 Conv
(Key Embedding)
1×1×1 Conv
(Value Embedding)
510
(Query)
520
(Memory)

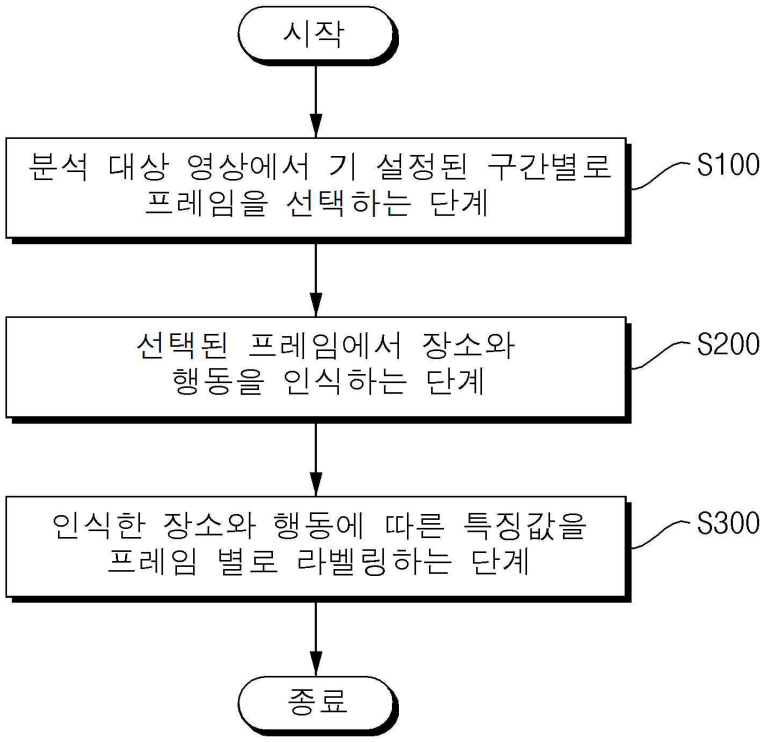
도면6


Sum=1이 되도록
해주는 역할
시간×높이×너비
마다의 중요도 (scalar) 값
32×7×7×1
1×512
Divided by L1 Norm
32×7×7×1
sigmoid
Instance Norm
32×7×7×1
1X1X1 Conv
32×7×7×512
32×7×7×512
Convolution
Feature
⊙ : element-wise multiplication

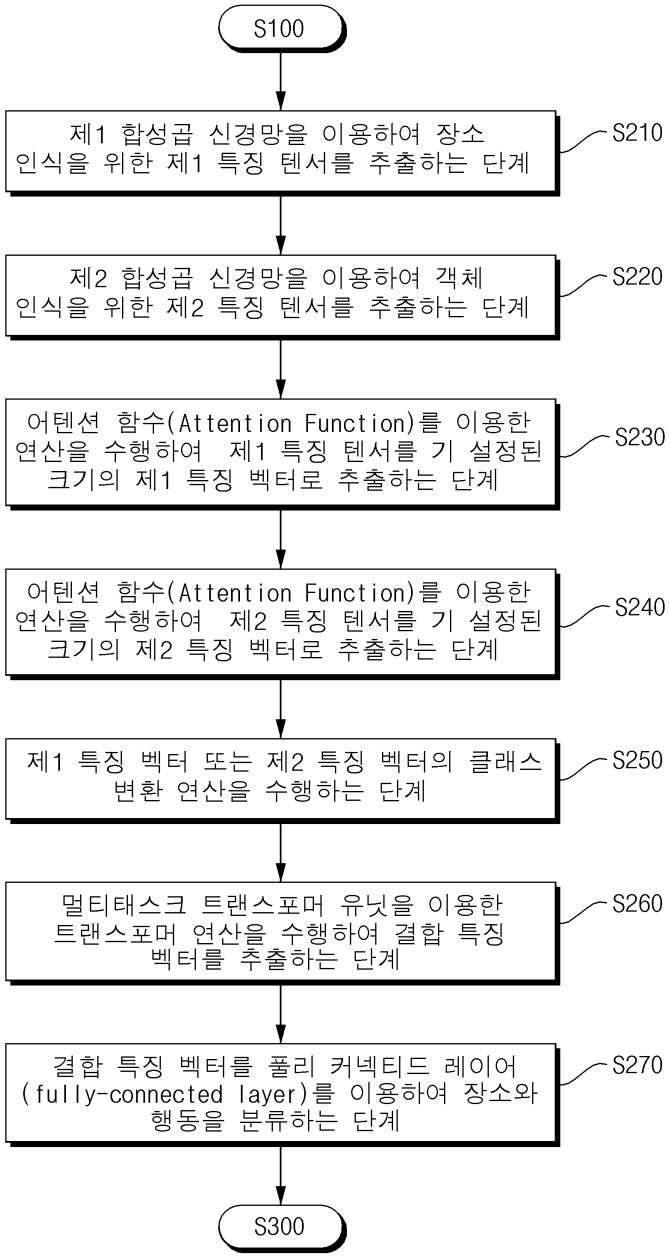
도면7


Scene: Concert, Action: Performance
Input Video
Scene
Action
Scene: Park, Action: Cycling
Input Video
Scene
Action

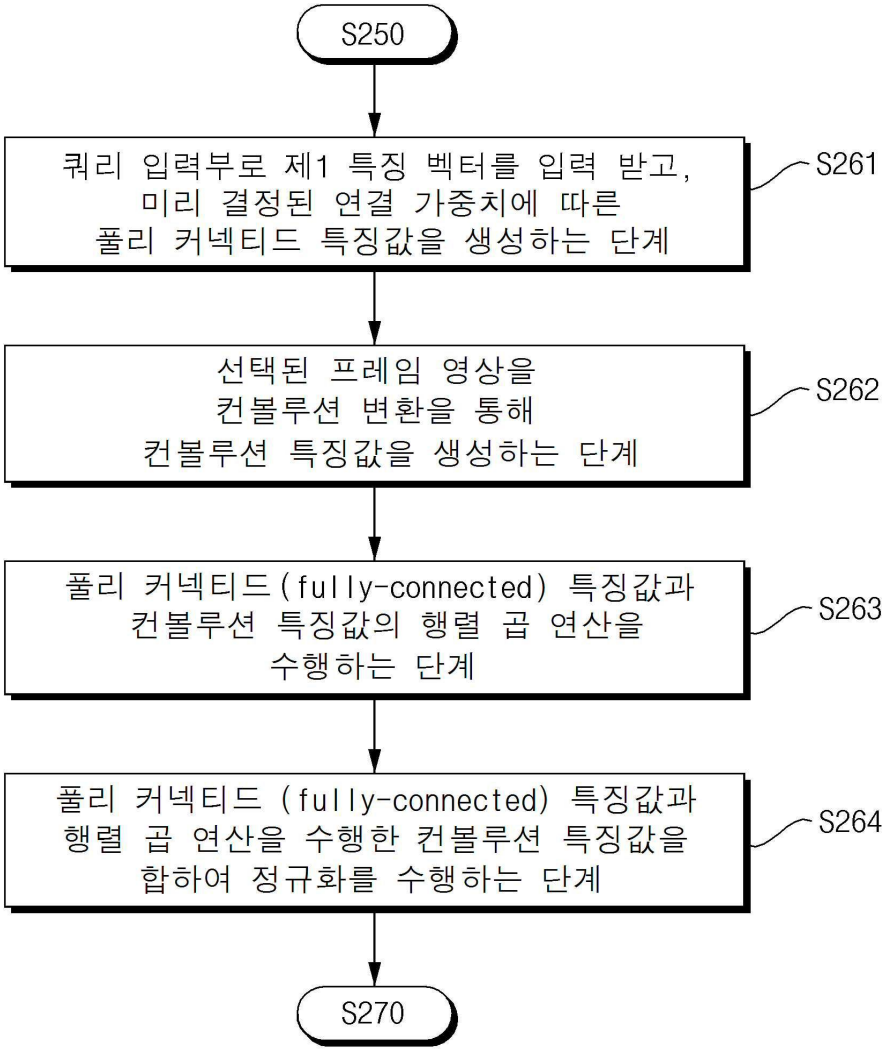
도면8


시작
분석 대상 영상에서 기 설정된 구간별로 프레임을 선택하는 단계
S100
선택된 프레임에서 장소와 행동을 인식하는 단계
S200
인식한 장소와 행동에 따른 특징값을 프레임 별로 라벨링하는 단계
S300
종료

도면9


S100
제1 합성곱 신경망을 이용하여 장소 인식을 위한 제1 특징 텐서를 추출하는 단계
S210
제2 합성곱 신경망을 이용하여 객체 인식을 위한 제2 특징 텐서를 추출하는 단계
S220
어텐션 함수(Attention Function)를 이용한 연산을 수행하여 제1 특징 텐서를 기 설정된 크기의 제1 특징 벡터로 추출하는 단계
S230
어텐션 함수(Attention Function)를 이용한 연산을 수행하여 제2 특징 텐서를 기 설정된 크기의 제2 특징 벡터로 추출하는 단계
S240
제1 특징 벡터 또는 제2 특징 벡터의 클래스 변환 연산을 수행하는 단계
S250
멀티태스크 트랜스포머 유닛을 이용한 트랜스포머 연산을 수행하여 결합 특징 벡터를 추출하는 단계
S260
결합 특징 벡터를 풀리 커넥티드 레이어(fully-connected layer)를 이용하여 장소와 행동을 분류하는 단계
S270
S300

도면10


S250
쿼리 입력부로 제1 특징 벡터를 입력 받고, 미리 결정된 연결 가중치에 따른 풀리 커넥티드 특징값을 생성하는 단계
S261
선택된 프레임 영상을 컨볼루션 변환을 통해 컨볼루션 특징값을 생성하는 단계
S262
풀리 커넥티드(fully-connected) 특징값과 컨볼루션 특징값의 행렬 곱 연산을 수행하는 단계
S263
풀리 커넥티드 (fully-connected) 특징값과 행렬 곱 연산을 수행한 컨볼루션 특징값을 합하여 정규화를 수행하는 단계
S264
S270