(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0059879
(43) 공개일자 2023년05월04일

(51) 국제특허분류(Int. Cl.)
G06F 18/00 (2023.01) *G06N 3/04* (2023.01)
G06N 3/08 (2023.01)
(52) CPC특허분류
G06N 3/04 (2023.01)
G06N 3/08 (2023.01)
(21) 출원번호 10-2021-0142851
(22) 출원일자 2021년10월25일
심사청구일자 2021년10월25일

(71) 출원인
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
김선주
서울특별시 양천구 목동서로 70, 228동 201호 (목동, 목동신시가지아파트2단지)
강효림
서울특별시 서대문구 신촌로 11길 49, 201호
(뒷면에 계속)
(74) 대리인
정부연

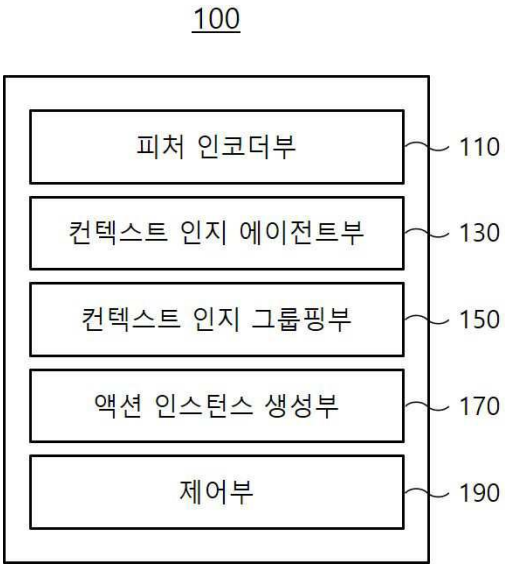
전체 청구항 수 : 총 16 항

(54) 발명의 명칭 **모방 학습을 이용한 실시간 비디오 동작 검출 장치 및 방법**

(57) 요 약

본 발명은 실시간 비디오 동작 검출 장치 및 방법에 관한 것으로, 상기 장치는 비디오 프레임의 피처를 인코딩하는 피처 인코더부; 이전 시간의 이산 출력을 입력받는 결정 큐(Qd)의 출력, 상기 인코딩된 피처를 입력받는 제1 OAD (Online Action Detection) 모델의 출력과 연결되는 액션 큐(Qa)의 출력을 입력받아, 현재 시간의 이산 출력을 출력하는 컨텍스트 인지 에이전트부; 상기 이전 시간의 이산 출력 및 현재 시간의 이산 출력을 기초로 해당 비디오 프레임들에 대하여 컨텍스트 인지 그룹핑을 수행하는 컨텍스트 인지 그룹핑부; 및 상기 해당 비디오 프레임들에 관한 액션을 검출하여 액션 인스턴스를 생성하는 액션 인스턴스 생성부;를 포함한다.

대 표 도 - 도1



(72) 발명자

김경민

서울특별시 영등포구 신길로 77래미안에스티움 107동 702호

고유민

서울특별시 서초구 잠원로 221 1동 205호

이 발명을 지원한 국가연구개발사업

과제고유번호	1711126082
과제번호	2020-0-01361-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성(R&D)
연구과제명	인공지능대학원지원(연세대학교)
기 여 율	1/1
과제수행기관명	연세대학교 산학협력단
연구기간	2021.01.01 ~ 2021.12.31

명 세 서

청구범위

청구항 1

비디오 프레임의 피처를 인코딩 하는 피처 인코더부;

이전 시간의 이산 출력을 입력받는 결정 큐(Qd)의 출력, 상기 인코딩된 피처를 입력받는 제1 OAD (Online Action Detection) 모델의 출력과 연결되는 액션 큐(Qa)의 출력을 입력받아, 현재 시간의 이산 출력을 출력하는 컨텍스트 인지 에이전트부;

상기 이전 시간의 이산 출력 및 현재 시간의 이산 출력을 기초로 해당 비디오 프레임들에 대하여 컨텍스트 인지 그룹핑을 수행하는 컨텍스트 인지 그룹핑부; 및

상기 해당 비디오 프레임들에 관한 액션을 검출하여 액션 인스턴스를 생성하는 액션 인스턴스 생성부;를 포함하는 실시간 비디오 동작 검출 장치.

청구항 2

제1항에 있어서, 상기 피처 인코더부는

기 설정된 프레임 간격마다 연속된 프레임들을 인코딩하여 프레임 피처를 생성하고 프레임 피처들에 관한 피처 시퀀스를 생성하는 것을 특징으로 하는 실시간 비디오 동작 검출 장치.

청구항 3

제1항에 있어서, 상기 컨텍스트 인지 에이전트부는

상기 제1 OAD 모델을 통해 상기 인코딩된 피처에 대한 액션 스코어를 상기 출력으로 결정하고, 상기 액션 스코어의 값은 상기 인코딩된 피처의 액션 정도에 해당하는 것을 특징으로 하는 실시간 비디오 동작 검출 장치.

청구항 4

제3항에 있어서, 상기 컨텍스트 인지 에이전트부는

상기 제1 OAD 모델을 LSTM (Long Short-Term Memory) 네트워크와 복수의 FC (Fully Connected) 레이어들로 구성된 바이너리 OAD 모델로 구현하는 것을 특징으로 하는 실시간 비디오 동작 검출 장치.

청구항 5

제1항에 있어서, 상기 컨텍스트 인지 에이전트부는

상기 결정 큐의 출력, 상기 액션 큐의 출력 및 상기 인코딩된 피처를 입력받는 제2 OAD 모델의 출력을 입력받아, 상기 현재 시간의 이산 출력을 출력하는 것을 특징으로 하는 실시간 비디오 동작 검출 장치.

청구항 6

제5항에 있어서, 상기 컨텍스트 인지 에이전트부는

상기 제2 OAD 모델을 통해 상기 인코딩된 피처에 대한 클래스 확률을 상기 출력으로 결정하고, 상기 클래스 확률의 값은 상기 인코딩된 피처의 액션 시점에 해당할 가능성에 해당하는 것을 특징으로 하는 실시간 비디오 동

작 검출 장치.

청구항 7

제6항에 있어서, 상기 컨텍스트 인지 에이전트부는

상기 제2 OAD 모델을 LSTM (Long Short-Term Memory) 네트워크와 복수의 FC (Fully Connected) 레이어들로 구성된 멀티-클래스 OAD 모델로 구현하는 것을 특징으로 하는 실시간 비디오 동작 검출 장치.

청구항 8

제1항에 있어서, 상기 컨텍스트 인지 그룹핑부는

상기 이산 출력의 전환 시점들을 검출하고, 상기 전환 시점들 사이에 있는 비디오 프레임들에 관해 상기 컨텍스트 인지 그룹핑을 수행하는 것을 특징으로 하는 실시간 비디오 동작 검출 장치.

청구항 9

제1항에 있어서, 상기 액션 인스턴스 생성부는

상기 액션 인스턴스를 상기 액션에 관한 텍스트 기반의 동작설명으로 생성하는 것을 특징으로 하는 실시간 비디오 동작 검출 장치.

청구항 10

비디오 프레임의 피처를 인코딩 하는 단계;

이전 시간의 이산 출력을 입력받는 결정 큐(Qd)의 출력, 상기 인코딩된 피처를 입력받는 제1 OAD (Online Action Detection) 모델의 출력과 연결되는 액션 큐(Qa)의 출력을 입력받아, 현재 시간의 이산 출력을 출력하는 단계;

상기 이전 시간의 이산 출력 및 현재 시간의 이산 출력을 기초로 해당 비디오 프레임들에 대하여 컨텍스트 인지 그룹핑을 수행하는 단계; 및

상기 해당 비디오 프레임들에 관한 액션을 검출하여 액션 인스턴스를 생성하는 단계를 포함하는 온라인 비디오의 컨텍스트 인지 그룹핑 단계;를 포함하는 실시간 비디오 동작 검출 방법.

청구항 11

제10항에 있어서, 상기 피처를 인코딩 하는 단계는

기 설정된 프레임 간격마다 연속된 프레임들을 인코딩하여 프레임 피처를 생성하고 프레임 피처들에 관한 피처 시퀀스를 생성하는 단계를 포함하는 것을 특징으로 하는 실시간 비디오 동작 검출 방법.

청구항 12

제10항에 있어서, 상기 현재 시간의 이산 출력을 출력하는 단계는

상기 제1 OAD 모델을 LSTM (Long Short-Term Memory) 네트워크와 복수의 FC (Fully Connected) 레이어들로 구성된 바이너리 OAD 모델로 구현하는 단계를 포함하는 것을 특징으로 하는 실시간 비디오 동작 검출 방법.

청구항 13

제10항에 있어서, 상기 현재 시간의 이산 출력을 출력하는 단계는

상기 결정 큐의 출력, 상기 액션 큐의 출력 및 상기 인코딩된 피처를 입력받는 제2 OAD 모델의 출력을 입력받아, 상기 현재 시간의 이산 출력을 출력하는 단계를 포함하는 것을 특징으로 하는 실시간 비디오 동작 검출 방법.

청구항 14

제13항에 있어서, 상기 현재 시간의 이산 출력을 출력하는 단계는

상기 제2 OAD 모델을 LSTM (Long Short-Term Memory) 네트워크와 복수의 FC (Fully Connected) 레이어들로 구성된 멀티-클래스 OAD 모델로 구현하는 단계를 포함하는 것을 특징으로 하는 실시간 비디오 동작 검출 방법.

청구항 15

제10항에 있어서, 상기 컨텍스트 인지 그룹핑을 수행하는 단계는

상기 이산 출력의 전환 시점들을 검출하고, 상기 전환 시점들 사이에 있는 비디오 프레임들에 관해 상기 컨텍스트 인지 그룹핑을 수행하는 단계를 포함하는 것을 특징으로 하는 실시간 비디오 동작 검출 방법.

청구항 16

제10항에 있어서, 상기 액션 인스턴스를 생성하는 단계는

상기 액션 인스턴스를 상기 액션에 관한 텍스트 기반의 동작설명으로 생성하는 단계를 포함하는 것을 특징으로 하는 실시간 비디오 동작 검출 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 실시간 비디오 동작 검출 기술에 관한 것으로, 보다 상세하게는 비디오 스트리밍 상황에서 액션의 시작과 끝, 그리고 액션 인스턴스의 클래스를 생성하여 제공하는 실시간 비디오 동작 검출 장치 및 방법에 관한 것이다.

배 경 기 술

[0003] 비디오 플랫폼의 발전에 따라 비디오 이해 작업들(video understanding tasks)은 컴퓨터 비전 연구 분야에서 상당한 관심을 끌고 있다. 많은 비디오 이해 작업들 중 무편집 비디오(untrimmed video)에서 액션 인스턴스(action instance)를 추출하는 작업인 시간적 행동 국지화(TAL, Temporal Action Localization)는 가장 인기 있는 주제 중 하나일 수 있다. TAL에서는 많은 작업들이 수행되어 왔으며, 이는 비디오 이해에 있어 액션 인스턴스의 중요성을 나타낼 수 있다.

[0004] 그러나, 실시간 및 온라인 접근이 필요한 비디오 스트리밍 서비스가 점점 더 많이 제공되고 있음에도 불구하고 스트리밍 비디오에서 액션 인스턴스를 감지하는 작업은 많은 주목을 받지 못하고 있다. 이와 달리, 객체 감지(object detection) 및 추적(tracking), 비디오 객체 분할(video object segmentation) 및 비디오 인스턴스 분할(video instance segmentation) 등과 같은 다른 비디오 이해 작업들에서 많은 온라인 알고리즘들이 소개되고 있다.

[0005] 인기 있는 스포츠 웹사이트에서는 스포츠 경기의 진행 상황을 실시간으로 보여주는 실시간 경기 중계 시스템

(live play-by-play system)이라는 기능을 제공하고 있다. AI 기반 실시간 경기 중계 시스템을 개발하기 위해서는 알고리즘이 온라인 방식으로 발생하는 액션의 시작, 종료 시점 및 클래스 정보를 모두 감지할 필요가 있다. 이전의 시간적 행동 국지화 방법들은 오프라인 방식으로 동작하고, 이에 따라 전체 비디오 시퀀스를 볼 수 있어야 하기 때문에 사용되기 어려울 수 있다. 도 3은 스포츠 생중계를 위한 실황 시스템에서 온라인 시간적 행동 국지화의 적용 사례를 도시하고 있다. 시간적 행동 국지화의 온라인 버전의 또 다른 중요한 사용은 로봇 공학 영역에서 찾을 수 있다. 로봇이 실시간으로 인간과 상호작용하기 위해서는 반응 방법을 결정하기 전에 전체 액션 인스턴스에 대한 정보가 필요할 수 있다.

[0006] 무편집 스트리밍 비디오(untrimmed streaming video)에서 액션 인스턴스를 즉석에서 생성하는 작업은 Online Temporal Action Localization 또는 On-TAL에 해당할 수 있다. 이름에서 알 수 있듯이 On-TAL의 최종 출력은 시작 및 종료 타임스탬프가 있는 액션 인스턴스인 offline TAL과 동일할 수 있다. 그러나, On-TAL 설정에서는 작업이 종료되는 즉시 액션 인스턴스를 생성해야 하므로 다음과 같이 offline TAL과 구별하는 몇 가지 문제가 있다.

[0007] 첫 번째로, 미래 프레임(future frame)에 접근하지 않고, 모델은 액션이 종료되는 즉시 액션 인스턴스를 반환해야 하기 때문에 현재 프레임에 액션이 포함되어 있는지 여부를 결정할 필요가 있다. 이때, 해당 결정은 모든 프레임에 대해 발생할 수 있다. 두 번째로, 액션 인스턴스는 신속하게 생성되고 시간을 되돌릴 수 없으므로 이전 결과에 대한 수정이 엄격히 금지되어 (Soft-)Non Maximum Suppression(NMS) 등과 같은 일반적인 후처리 방법(post-processing method)을 사용할 수 없다.

[0008] 대부분의 이전 TAL 방법들은 이러한 제약으로 인해 전체 비디오를 보고 NMS 기술을 사용하여 중복 결과를 제거해야 하는 점에서 온라인 설정에 대한 이전의 TAL 접근들을 단순 확장할 수 없다.

[0009] 온라인 행동 탐지(OAD, Online Action Detection)는 스트리밍 비디오에서 프레임별 레이블을 추출하는 것을 목적으로 하는 온라인 비디오 처리 작업에 해당할 수 있다. OAD는 프레임별 레이블링을 제공한다는 점에서 On-TAL은 OAD를 중간 절차에 사용하여 설계될 수 있다. 즉, 중간 절차는 액션 프레임을 구별하고 그룹화하는 바이너리 OAD 모델을 학습하는 과정을 포함할 수 있다. 그러나, 이 접근 방식에는 행동 단편화(action fragmentation) 및 행동 틱(action tick)(도 4 참조)이라는 무시하기 힘든 제한 사항이 존재할 수 있다. 이러한 문제는 모델이 현재 프레임에 대한 올바른 결정을 내리는데 필수적임에도 불구하고 과거의 결정들을 반영하지 못함으로써 결과적으로 결정 컨텍스트(decision context)를 인식하지 못하기 때문에 발생할 수 있다.

선행기술문헌

특허문헌

[0011] (특허문헌 0001) 한국공개특허 제10-2018-0054453호 (2018.05.24)

발명의 내용

해결하려는 과제

[0012] 본 발명의 일 실시예는 비디오 스트리밍 상황에서 액션의 시작과 끝, 그리고 액션 인스턴스의 클래스를 생성하여 제공하는 실시간 비디오 동작 검출 장치 및 방법을 제공하고자 한다.

[0013] 본 발명의 일 실시예는 온라인 시간적 행동 국지화(On-TAL, Online Temporal Action Localization)에 대한 해결책으로서 Q 모방 학습(Q Imitation Learning, QIL) 프레임워크를 통한 컨텍스트 인지 행동성 그룹핑(Context-Aware Actionness Grouping, CAG)을 수행할 수 있는 실시간 비디오 동작 검출 장치 및 방법을 제공하고자 한다.

과제의 해결 수단

[0015] 실시예들 중에서, 실시간 비디오 동작 검출 장치는 비디오 프레임의 피처를 인코딩 하는 피처 인코더부; 이전

시간의 이산 출력을 입력받는 결정 큐(Qd)의 출력, 상기 인코딩된 피처를 입력받는 제1 OAD (Online Action Detection) 모델의 출력과 연결되는 액션 큐(Qa)의 출력을 입력받아, 현재 시간의 이산 출력을 출력하는 컨텍스트 인지 에이전트부; 상기 이전 시간의 이산 출력 및 현재 시간의 이산 출력을 기초로 해당 비디오 프레임들에 대하여 컨텍스트 인지 그룹핑을 수행하는 컨텍스트 인지 그룹핑부; 및 상기 해당 비디오 프레임들에 관한 액션을 검출하여 액션 인스턴스를 생성하는 액션 인스턴스 생성부;를 포함한다.

[0016] 상기 피처 인코더부는 기 설정된 프레임 간격마다 연속된 프레임들을 인코딩하여 프레임 피처를 생성하고 프레임 피처들에 관한 피처 시퀀스를 생성할 수 있다.

[0017] 상기 컨텍스트 인지 에이전트부는 상기 제1 OAD 모델을 통해 상기 인코딩된 피처에 대한 액션 스코어를 상기 출력으로 결정할 수 있다. 이때, 상기 액션 스코어의 값은 상기 인코딩된 피처의 액션 정도에 해당할 수 있다.

[0018] 상기 컨텍스트 인지 에이전트부는 상기 제1 OAD 모델을 LSTM (Long Short-Term Memory) 네트워크와 복수의 FC (Fully Connected) 레이어들로 구성된 바이너리 OAD 모델로 구현할 수 있다.

[0019] 상기 컨텍스트 인지 에이전트부는 상기 결정 큐의 출력, 상기 액션 큐의 출력 및 상기 인코딩된 피처를 입력받는 제2 OAD 모델의 출력을 입력받아, 상기 현재 시간의 이산 출력을 출력할 수 있다.

[0020] 상기 컨텍스트 인지 에이전트부는 상기 제2 OAD 모델을 통해 상기 인코딩된 피처에 대한 클래스 확률을 상기 출력으로 결정할 수 있다. 이때, 상기 클래스 확률의 값은 상기 인코딩된 피처의 액션 시점에 해당할 가능성에 해당할 수 있다.

[0021] 상기 컨텍스트 인지 에이전트부는 상기 제2 OAD 모델을 LSTM (Long Short-Term Memory) 네트워크와 복수의 FC (Fully Connected) 레이어들로 구성된 멀티-클래스 OAD 모델로 구현할 수 있다.

[0022] 상기 컨텍스트 인지 그룹핑부는 상기 이산 출력의 전환 시점들을 검출하고, 상기 전환 시점들 사이에 있는 비디오 프레임들에 관해 상기 컨텍스트 인지 그룹핑을 수행할 수 있다.

[0023] 상기 액션 인스턴스 생성부는 상기 액션 인스턴스를 상기 액션에 관한 텍스트 기반의 동작설명으로 생성할 수 있다.

[0024] 실시예들 중에서, 실시간 비디오 동작 검출 방법은 비디오 프레임의 피처를 인코딩 하는 단계; 이전 시간의 이산 출력을 입력받는 결정 큐(Qd)의 출력, 상기 인코딩된 피처를 입력받는 제1 OAD (Online Action Detection) 모델의 출력을 입력받아, 현재 시간의 이산 출력을 출력하는 단계; 상기 이전 시간의 이산 출력 및 현재 시간의 이산 출력을 기초로 해당 비디오 프레임들에 대하여 컨텍스트 인지 그룹핑을 수행하는 단계; 및 상기 해당 비디오 프레임들에 관한 액션을 검출하여 액션 인스턴스를 생성하는 단계를 포함하는 온라인 비디오의 컨텍스트 인지 그룹핑 단계;를 포함한다.

[0025] 상기 피처를 인코딩 하는 단계는 기 설정된 프레임 간격마다 연속된 프레임들을 인코딩하여 프레임 피처를 생성하고 프레임 피처들에 관한 피처 시퀀스를 생성하는 단계를 포함할 수 있다.

[0026] 상기 현재 시간의 이산 출력을 출력하는 단계는 상기 제1 OAD 모델을 LSTM (Long Short-Term Memory) 네트워크와 복수의 FC (Fully Connected) 레이어들로 구성된 바이너리 OAD 모델로 구현하는 단계를 포함할 수 있다.

[0027] 상기 현재 시간의 이산 출력을 출력하는 단계는 상기 결정 큐의 출력, 상기 액션 큐의 출력 및 상기 인코딩된 피처를 입력받는 제2 OAD 모델의 출력을 입력받아, 상기 현재 시간의 이산 출력을 출력하는 단계를 포함할 수 있다.

[0028] 상기 현재 시간의 이산 출력을 출력하는 단계는 상기 제2 OAD 모델을 LSTM (Long Short-Term Memory) 네트워크와 복수의 FC (Fully Connected) 레이어들로 구성된 멀티-클래스 OAD 모델로 구현하는 단계를 포함할 수 있다.

[0029] 상기 컨텍스트 인지 그룹핑을 수행하는 단계는 상기 이산 출력의 전환 시점들을 검출하고, 상기 전환 시점들 사이에 있는 비디오 프레임들에 관해 상기 컨텍스트 인지 그룹핑을 수행하는 단계를 포함할 수 있다.

[0030] 상기 액션 인스턴스를 생성하는 단계는 상기 액션 인스턴스를 상기 액션에 관한 텍스트 기반의 동작설명으로 생성하는 단계를 포함할 수 있다.

발명의 효과

[0032] 개시된 기술은 다음의 효과를 가질 수 있다. 다만, 특정 실시예가 다음의 효과를 전부 포함하여야 한다거나 다음의 효과만을 포함하여야 한다는 의미는 아니므로, 개시된 기술의 권리범위는 이에 의하여 제한되는 것으로 이해되어서는 아니 될 것이다.

[0033] 본 발명에 따른 실시간 비디오 동작 검출 장치 및 방법은 비디오 스트리밍 상황에서 액션의 시작과 끝, 그리고 액션 인스턴스의 클래스를 생성하여 제공할 수 있다.

[0034] 본 발명에 따른 실시간 비디오 동작 검출 장치 및 방법은 온라인 시간적 행동 국지화(On-TAL, Online Temporal Action Localization)에 대한 해결책으로서 Q 모방 학습(Q Imitation Learning, QIL) 프레임워크를 통한 컨텍스트 인지 행동성 그룹핑(Context-Aware Actionness Grouping, CAG)을 수행할 수 있다.

도면의 간단한 설명

[0036] 도 1은 본 발명에 따른 실시간 비디오 동작 검출 장치의 기능적 구성을 설명하는 도면이다.

도 2는 본 발명에 따른 실시간 비디오 동작 검출 방법을 설명하는 순서도이다.

도 3은 비디오에 적용된 play-by-play 시스템을 설명하는 도면이다.

도 4는 온라인 OAD의 단순 확장의 제약을 설명하는 도면이다.

도 5는 본 발명에 따른 컨텍스트 인지 행동성 그룹핑 과정을 설명하는 도면이다.

도 6은 본 발명에 따른 컨텍스트 인지 행동성 그룹핑 과정의 추론 단계를 설명하는 도면이다.

도 7은 프레임 단위의 OAD 과정의 일 실시예를 설명하는 도면이다.

도 8은 전문가 데이터베이스의 구축 과정을 설명하는 도면이다.

도 9는 본 발명에 따른 MDP 과정을 설명하는 도면이다.

도 10 내지 15는 본 발명에 관한 실험 결과를 설명하는 도면이다.

도 16은 본 발명에 따른 실시간 비디오 동작 검출 장치의 시스템 구성을 설명하는 도면이다.

도 17은 본 발명에 따른 실시간 비디오 동작 검출 시스템을 설명하는 도면이다.

발명을 실시하기 위한 구체적인 내용

[0037] 본 발명에 관한 설명은 구조적 내지 기능적 설명을 위한 실시예에 불과하므로, 본 발명의 권리범위는 본문에 설명된 실시예에 의하여 제한되는 것으로 해석되어서는 아니 된다. 즉, 실시예는 다양한 변경이 가능하고 여러 가지 형태를 가질 수 있으므로 본 발명의 권리범위는 기술적 사상을 실현할 수 있는 균등물들을 포함하는 것으로 이해되어야 한다. 또한, 본 발명에서 제시된 목적 또는 효과는 특정 실시예가 이를 전부 포함하여야 한다거나 그러한 효과만을 포함하여야 한다는 의미는 아니므로, 본 발명의 권리범위는 이에 의하여 제한되는 것으로 이해되어서는 아니 될 것이다.

[0038] 한편, 본 출원에서 서술되는 용어의 의미는 다음과 같이 이해되어야 할 것이다.

[0039] "제1", "제2" 등의 용어는 하나의 구성요소를 다른 구성요소로부터 구별하기 위한 것으로, 이들 용어들에 의해 권리범위가 한정되어서는 아니 된다. 예를 들어, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다.

[0040] 어떤 구성요소가 다른 구성요소에 "연결되어"있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결될 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어"있다고 언급된 때에는 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다. 한편, 구성요소들 간의 관계를 설명하는 다른 표현들, 즉 "~사이에"와 "바로 ~사이에" 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.

[0041] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한 복수의 표현을 포함하는 것으로 이해되어야 하고, "포함하다"또는 "가지다" 등의 용어는 실시된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것이며, 하나 또는 그 이상의 다른 특징이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이

들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0042] 각 단계들에 있어 식별부호(예를 들어, a, b, c 등)는 설명의 편의를 위하여 사용되는 것으로 식별부호는 각 단계들의 순서를 설명하는 것이 아니며, 각 단계들은 문맥상 명백하게 특정 순서를 기재하지 않는 이상 명기된 순서와 다르게 일어날 수 있다. 즉, 각 단계들은 명기된 순서와 동일하게 일어날 수도 있고 실질적으로 동시에 수행될 수도 있으며 반대의 순서대로 수행될 수도 있다.

[0043] 본 발명은 컴퓨터가 읽을 수 있는 기록매체에 컴퓨터가 읽을 수 있는 코드로서 구현될 수 있고, 컴퓨터가 읽을 수 있는 기록 매체는 컴퓨터 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록 장치를 포함한다. 컴퓨터가 읽을 수 있는 기록 매체의 예로는 ROM, RAM, CD-ROM, 자기 테이프, 플로피 디스크, 광 데이터 저장 장치 등이 있다. 또한, 컴퓨터가 읽을 수 있는 기록 매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어, 분산 방식으로 컴퓨터가 읽을 수 있는 코드가 저장되고 실행될 수 있다.

[0044] 여기서 사용되는 모든 용어들은 다르게 정의되지 않는 한, 본 발명이 속하는 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한 이상적이거나 과도하게 형식적인 의미를 지니는 것으로 해석될 수 없다.

[0046] 도 1은 본 발명에 따른 실시간 비디오 동작 검출 장치의 기능적 구성을 설명하는 도면이다.

[0047] 도 1을 참조하면, 실시간 비디오 동작 검출 장치(100)는 피처 인코더부(110), 컨텍스트 인지 에이전트부(130), 컨텍스트 인지 그룹핑부(150), 액션 인스턴스 생성부(170) 및 제어부(190)를 포함할 수 있다.

[0048] 피처 인코더부(110)는 비디오 프레임의 피처를 인코딩 할 수 있다. 여기에서, 비디오 프레임은 소정의 시간동안 촬영된 영상, 즉 비디오에 해당할 수 있다. 특히, 비디오 프레임은 편집되지 않은 무편집 비디오(untrimmed video)의 연속된 프레임들의 집합에 해당할 수 있다.

[0049] 일 실시예에서, 피처 인코더부(110)는 기 설정된 프레임 간격마다 연속된 프레임들을 인코딩하여 프레임 피처를 생성하고 프레임 피처들에 관한 피처 시퀀스를 생성할 수 있다. 피처 인코더부(110)는 비디오 프레임을 수신하여 프레임 순서에 따라 소정의 프레임 간격마다 연속된 프레임들로 분할할 수 있다. 이때, 프레임 간격은 사전에 설정될 수 있으며, 예를 들어, 프레임 간격이 5인 경우 연속된 5 프레임들로 분할될 수 있다. 피처 인코더부(110)는 연속된 프레임들에 대해 인코딩한 결과로서 피처들을 순차적으로 생성할 수 있다. 이때, 순차적으로 생성된 피처들의 집합은 피처 시퀀스(feature sequence)에 해당할 수 있다.

[0050] 컨텍스트 인지 에이전트부(130)는 이전 시간의 이산 출력을 입력받는 결정 큐(Qd)의 출력, 인코딩된 피처를 입력받는 제1 OAD (Online Action Detection) 모델의 출력과 연결되는 액션 큐(Qa)의 출력을 입력받아, 현재 시간의 이산 출력을 출력할 수 있다. 즉, 컨텍스트 인지 에이전트부(130)는 기존의 OAD 프레임워크를 확장하여 도출되는 새로운 프레임 워크를 포함할 수 있다. 구체적으로, 컨텍스트 인지 에이전트부(130)는 1개의 OAD 모델의 출력과 이전 단계의 출력을 기초로 현재 단계의 출력을 생성하는 동작을 수행할 수 있다.

[0051] 일 실시예에서, 컨텍스트 인지 에이전트부(130)는 이전 시간의 이산 출력을 입력받는 결정 큐(Qd)의 출력, 인코딩된 피처를 입력받는 제1 OAD (Online Action Detection) 모델의 출력과 연결되는 액션 큐(Qa)의 출력 및 인코딩된 피처를 입력받는 제2 OAD 모델의 출력을 입력받아, 현재 시간의 이산 출력을 출력할 수 있다. 즉, 컨텍스트 인지 에이전트부(130)는 서로 다른 2개의 OAD 모델들의 출력과 이전 단계의 출력을 기초로 현재 단계의 출력을 생성하는 동작을 수행할 수 있다.

[0052] 특히, 컨텍스트 인지 에이전트부(130)는 이전 시간의 이산 출력을 결정 큐(Qd)를 통해 수신하여 입력을 사용할 수 있고, 제1 OAD 모델의 출력을 액션 큐(Qa)를 통해 수신하여 입력으로 사용할 수 있다. 또한, 컨텍스트 인지 에이전트부(130)는 제2 OAD 모델의 출력으로서 현재 프레임의 클래스 확률(Pt)를 수신하여 입력으로 사용할 수 있다. 예를 들어, 3개의 연속된 프레임들에 대한 이산 출력 d를 생성하는 경우 컨텍스트 인지 에이전트부(130)는 각 프레임의 클래스 확률 [p_{t1}, p_{t2}, p_{t3}]를 입력으로 수신할 수 있다.

[0053] 일 실시예에서, 컨텍스트 인지 에이전트부(130)는 제1 OAD 모델을 통해 인코딩된 피처에 대한 액션 스코어를 출력으로 결정할 수 있다. 제1 OAD 모델은 시점 t에서 피처 f를 입력으로 수신할 수 있고, 해당 피처 f에 대응되는 액션 스코어를 출력으로 생성할 수 있다. 여기에서, 액션 스코어는 행동성 점수(actionness score)에 해당할

수 있다. 또한, 액션 스코어의 값은 인코딩된 피처의 액션 정도에 해당할 수 있다. 즉, 행동성 점수는 해당 피처에 액션이 포함된 정도에 따라 1에 가까운 값을 가질 수 있으며, 해당 피처에 액션이 포함되지 않은 경우 0에 해당할 수 있다. 제1 OAD 모델이 생성한 출력은 액션 큐(Qa)에 입력되어 저장될 수 있으며, 액션 큐를 통해 컨텍스트 인지 에이전트에 입력될 수 있다.

[0054] 일 실시예에서, 컨텍스트 인지 에이전트부(130)는 제1 OAD 모델을 LSTM (Long Short-Term Memory) 네트워크와 복수의 FC (Fully Connected) 레이어들로 구성된 바이너리 OAD 모델로 구현할 수 있다. 예를 들어, 컨텍스트 인지 에이전트부(130)는 단일 레이어(single layer) LSTM 네트워크와 2개의 FC 레이어들로 구성된 바이너리 OAD 모델을 구축하여 제1 OAD 모델로 사용할 수 있다.

[0055] 일 실시예에서, 컨텍스트 인지 에이전트부(130)는 제2 OAD 모델을 통해 인코딩된 피처에 대한 클래스 확률을 출력으로 결정할 수 있다. 제2 OAD 모델은 시점 t에서 피처 f를 입력으로 수신할 수 있고, 해당 피처 f에 대응되는 클래스 확률(class probability)를 출력으로 생성할 수 있다. 이때, 클래스 확률의 값은 인코딩된 피처의 액션 시작점에 해당할 가능성의 정도에 해당할 수 있다. 제2 OAD 모델이 생성한 출력은 프레임 간격 단위로 저장되어 컨텍스트 인지 에이전트에 입력될 수 있다.

[0056] 일 실시예에서, 컨텍스트 인지 에이전트부(130)는 제2 OAD 모델을 LSTM (Long Short-Term Memory) 네트워크와 복수의 FC (Fully Connected) 레이어들로 구성된 멀티-클래스 OAD 모델로 구현할 수 있다. 예를 들어, 컨텍스트 인지 에이전트부(130)는 단일 레이어(single layer) LSTM 네트워크와 2개의 FC 레이어들로 구성된 멀티 클래스 OAD 모델을 구축하여 제2 OAD 모델로 사용할 수 있다.

[0057] 컨텍스트 인지 그룹핑부(150)는 이전 시간의 이산 출력 및 현재 시간의 이산 출력을 기초로 해당 비디오 프레임들에 대하여 컨텍스트 인지 그룹핑을 수행할 수 있다. 여기에서, 컨텍스트 인지 그룹핑은 컨텍스트 인지 행동성 그룹핑(Context-Aware Actionness Grouping, CAG)에 해당할 수 있다. 컨텍스트 인지 그룹핑부(150)는 컨텍스트 인지 에이전트부(130)에 의해 출력되는 피처별 이산 출력의 결과들을 기초로 컨텍스트 인지 그룹핑을 수행할 수 있다. 예를 들어, 컨텍스트 인지 그룹핑부(150)는 컨텍스트 인지 에이전트부(130)에 의해 출력되는 값들 중에서 0이 아닌 값들과 연관된 피처들과 연관된 비디오 프레임들을 묶어 그룹화할 수 있다.

[0058] 일 실시예에서, 컨텍스트 인지 그룹핑부(150)는 이산 출력의 전환 시점들을 검출하고, 전환 시점들 사이에 있는 비디오 프레임들에 관해 컨텍스트 인지 그룹핑을 수행할 수 있다. 여기에서, 전환 시점은 컨텍스트 인지 에이전트부(130)에 의한 출력이 0에서 0이 아닌 값으로 변하거나 또는 0이 아닌 값에서 0으로 변하는 시점에 해당할 수 있다. 컨텍스트 인지 그룹핑부(150)는 피처 시퀀스에 대한 컨텍스트 인지 에이전트부(130)의 출력을 기초로 전환 시점을 검출할 수 있으며, 전환 시점들 사이의 비디오 프레임들을 기초로 컨텍스트 인지 그룹핑을 수행할 수 있다. 예를 들어, 컨텍스트 인지 그룹핑부(150)는 컨텍스트 인지 에이전트부(130)에 의한 출력이 0이 아닌 값들과 연관된 비디오 프레임들을 하나로 묶어 그룹화 할 수 있다.

[0059] 액션 인스턴스 생성부(170)는 해당 비디오 프레임들에 관한 액션을 검출하여 액션 인스턴스를 생성할 수 있다. 액션 인스턴스 생성부(170)는 컨텍스트 인지 그룹핑의 결과로 생성된 비디오 프레임들의 집합에 대해 액션 인스턴스를 생성할 수 있다. 즉, 해당 비디오 프레임들의 집합에는 적어도 하나의 행동들이 포함될 수 있으며, 액션 인스턴스는 해당 적어도 하나의 행동들을 기초로 생성될 수 있다. 예를 들어, 액션 인스턴스는 전환 시점에 관한 정보와 행동의 클래스 정보를 포함하여 생성될 수 있다.

[0060] 일 실시예에서, 액션 인스턴스 생성부(170)는 액션 인스턴스를 액션에 관한 텍스트 기반의 동작설명으로 생성할 수 있다. 액션 인스턴스는 전환 시점 정보와 검출된 행동의 클래스 정보를 포함할 수 있으며, 행동의 클래스 정보는 텍스트 기반의 동작설명을 포함할 수 있다. 예를 들어, 검출된 행동이 테스트 관련 행동으로 분류되는 경우, 클래스 정보는 'tennis swing'을 표현될 수 있다. 액션 인스턴스는 전환 시점과 함께 행동의 클래스 정보로서 'tennis swing'를 포함하여 생성될 수 있다.

[0061] 제어부(190)는 실시간 비디오 동작 검출 장치(100)의 전체적인 동작을 제어하고, 피처 인코더부(110), 컨텍스트 인지 에이전트부(130), 컨텍스트 인지 그룹핑부(150) 및 액션 인스턴스 생성부(170) 간의 제어 흐름 또는 데이터 흐름을 관리할 수 있다.

[0063] 도 2는 본 발명에 따른 실시간 비디오 동작 검출 방법을 설명하는 순서도이다.

[0064] 도 2를 참조하면, 실시간 비디오 동작 검출 장치(100)는 피처 인코더부(110)를 통해 비디오 프레임의 피처를 인

코딩 할 수 있다(단계 S210). 실시간 비디오 동작 검출 장치(100)는 컨텍스트 인지 에이전트부(130)를 통해 이전 시간의 이산 출력을 입력받는 결정 큐(Qd)의 출력, 인코딩된 피처를 입력받는 제1 OAD (Online Action Detection) 모델의 출력과 연결되는 액션 큐(Qa)의 출력 및 인코딩된 피처를 입력받는 제2 OAD 모델의 출력을 입력받아, 현재 시간의 이산 출력을 출력할 수 있다(단계 S230).

[0065] 또한, 실시간 비디오 동작 검출 장치(100)는 컨텍스트 인지 그룹핑부(150)를 통해 이전 시간의 이산 출력 및 현재 시간의 이산 출력을 기초로 해당 비디오 프레임들에 대하여 컨텍스트 인지 그룹핑을 수행할 수 있다(단계 S250). 실시간 비디오 동작 검출 장치(100)는 액션 인스턴스 생성부(170)를 통해 해당 비디오 프레임들에 관한 액션을 검출하여 액션 인스턴스를 생성할 수 있다(단계 S270).

[0067] 이하, 도 3 내지 9를 참조하여 본 발명에 따른 실시간 비디오 동작 검출 방법에 대해 보다 자세히 설명한다.

[0068] 시간적 행동 국지화(TAL, Temporal Action Localization)는 이미지 영역(image domain)에서 객체 탐지(object detection)와 유사할 수 있으며, 강화 학습(reinforcement learning)을 활용한 방법을 포함하여 다양한 방법들이 연구되어 왔다.

[0069] 본 발명에 따른 실시간 비디오 동작 검출 방법은 시간적 행동성 그룹핑(TAG, Temporal Actionness Grouping)과 연관될 수 있다. 여기에서, 행동성(Actionness)은 'Objectness'에 대응되는 개념으로서 행동 가능성에 대한 기준 신뢰도(confidence threshold)에 해당할 수 있다. 그러나, 본 발명의 경우 행동성 그룹핑(actionness grouping)은 미래 프레임에 접근하지 않고 수행되어야 하며, 이와 달리 TAG는 전체 비디오를 활용하고 행동 단편화(action fragmentation)을 피하기 위한 원시 제안(primitive proposal)들의 그룹핑이나 중복 제거를 위한 NMS 등과 같이 생성된 제안들의 후처리(post-processing)를 활용할 수 있다.

[0070] 본 발명에 관한 On-TAL 작업과 유사한 목표를 공유하는 온라인 비디오 이해 작업들이 존재할 수 있다. 예를 들어, 온라인 행동 탐지(OAD, Online Action Detection) 및 온라인 행동 개시 탐지(ODAS, Online Detection of Action Start)가 이에 해당할 수 있다.

[0071] OAD(Online Action Detection)는 프레임별 클래스 레이블을 추출하는 작업에 해당할 수 있다. OAD의 성능을 향상시키기 위해 다양한 방법들이 제안되었으며, 여기에는 초기 행동 탐지를 가능하게 하기 위해 강화 학습을 활용한 방법이 포함될 수 있다. OAD는 각 프레임에 대해 조밀 클래스 점수(dense class score)를 제공하기 때문에 추가 처리를 위한 좋은 중간 표현(intermediate representation)으로 간주될 수 있지만, 실세계 문제(real-world problem)에 직접 적용하기에는 소정의 한계가 존재할 수 있다.

[0072] 반면에, ODAS(Online Detection of Action Start)는 가능한 빨리 행동의 시작을 탐지하는 것을 목표로 하며, 희소 시작점 예측(sparse start point prediction)을 생성함으로써 작업을 보다 실용적으로 만들 수 있다.

[0073] 전반적으로, 이전의 온라인 비디오 처리 작업들은 주로 프레임 수준 정보(frame-level information)에 초점을 두고 있다. 이와 반대로, On-TAL은 실제 컴퓨터 비전 문제에 직접 배포할 수 있는 보다 풍부한 의미를 갖는 액션 인스턴스를 제공할 수 있다. 또한, 액션 인스턴스가 액션 시작점(action start point)을 요소로 포함하므로 On-TAL을 처리할 수 있는 모델은 ODAS 문제를 자동으로 해결할 수 있으며, 이는 On-TAL이 ODAS보다 상위 수준의 작업에 해당함을 의미할 수 있다.

[0075] 무편집 비디오(untrimmed video)는 $V = \{x_\tau\}_{\tau=1}^{T}$ 이고, m개의 액션 인스턴스들은 $\Psi = \{\psi_m\}_{m=1}^{M} = \{(s_m, e_m, c_m)\}_{m=1}^{M}$ 으로 가정할 수 있다. 여기에서, x_τ는 τ번째 프레임이고, s_m, e_m 및 c_m은 각각 m번째 액션 인스턴스 ψ_m의 시작 프레임 인덱스(start frame index), 끝 프레임 인덱스(end frame index) 및 클래스 레이블(class label)이다. 또한, x_τ는 온라인 조건 하에서 순차적으로 제공될 수 있다.

[0076] 최근의 시간적 제안 생성 방법(temporal proposal generation method)들에 따라, 연속된 k개의 프레임들은 시각적 특징(visual feature) f로 변환될 수 있고 행동 제안(action proposal)들의 온라인 생성을 포함하여 다음 모든 단계들은 해당 특징 시퀀스에 대해 실행될 수 있다. 해당 제안 생성을 위한 모델은 행동 종료를 감지하는

즉시 적절한 행동 제안을 출력하여야 한다. 즉, 모델은 모든 특징들에 대해 행동 제안을 생성할지 여부를 결정할 수 있다. 특징 시퀀스의 단위(granularity)는 On-TAL에서 중요할 수 있다. 즉, k가 작을수록 더 정밀한 결정을 함으로써 더 정확한 결과가 도출될 수 있는 반면, k가 클수록 이와 반대로 동작하여 더 부정확한 결과가 도출될 수 있다.

[0077] 본 발명에 따른 On-TAL의 최종 목표는 NMS와 같은 후처리 없이 온라인 생성된(online-generated) 각 액션 인스턴스 ψ를 집계하여 Ψ를 복원하는 것일 수 있다.

[0078] On-TAL은 액션 인스턴스 간에 중복이 없다고 가정하면 OAD 프레임워크를 확장하여 해결할 수 있다. 즉, 연속적인 k개의 프레임들은 인코더 E에 의해 피처 f로 변환되어 원시 프레임 시퀀스(raw frame sequence) $\{x_\tau\}_{\tau=1}^{T}$대신 특징 시퀀스(feature sequence) $\{f_t\}_{t=1}^{[T/k]}$ 이 될 수 있다. 각 시간 단계 t에서 OAD 모델 M은 피처 f_t를 입력으로 사용하고 행동성 점수(actionness score) $\alpha_t(0 < \alpha_t < 1)$를 출력할 수 있다. 여기에서, 특징에 액션이 포함되어 있으면 α_t는 1에 가깝고 그렇지 않으면 0일 수 있다. 단순히 온라인 방식으로 $\{\alpha_t|\alpha_t > threshold\}$를 그룹핑하면 액션 인스턴스들이 생성될 수 있다.

[0080] 단순한(naive) OAD 확장은 OAD 모델 M에 대해 상태 유지 RNN 아키텍처(stateful RNN architecture)를 사용하여 프레임 컨텍스트(frame context)를 처리할 수 있는 반면 여전히 결정 컨텍스트(decision context)를 반영할 수 없다. 이에 따라, 도 6의 알고리즘 1에서 설명하는 컨텍스트 인지 행동성 그룹핑(CAG, Context-Aware Actionness Grouping)은 On-TAL을 해결하기 위한 새로운 프레임워크(framework)에 해당할 수 있다.

[0081] 도 5를 참조하면, 본 발명에 따른 CAG에서는 새로운 구성에 해당하는 컨텍스트 인지 에이전트(context-aware agent) γ가 적용될 수 있다. 컨텍스트 인지 에이전트는 두 개의 대기열들(즉, 길이가 모두 n인 행동성 대기열 q_a 및 결정 대기열 q_d)과 현재 프레임의 클래스 확률(class probability) p_t를 입력으로 사용하고 이산 출력 $d \in \{0,1\}$를 반환할 수 있다. 해당 모델은 액션 인스턴스를 생성하기 위해 OAD 확장(OAD extension)에서와 같이 $\{\alpha_t|\alpha_t > threshold\}$를 그룹핑하는 대신 $\{d_t|d_t = 1\}$을 집계(aggregate)할 수 있다.

[0082] CAG에서, 이전 단계에서의 결정은 현재 프레임에서의 결정을 위해 고려될 수 있으며, 이에 따라 γ에 관한 학습이 복잡해질 수 있다. 이와 같은 반복을 무시하면서, 표준 지도 학습 방법을 사용하여 현재 상태(q_a, q_d, p)를 γ의 출력 d에 직접 매핑할 수 있으며, 해당 방법은 행동 복제(BC, Behavioral Cloning)에 해당할 수 있다. 그러나, BC 학습된(BC-trained) 에이전트는 BC가 전환 역학(transition dynamics)을 완전히 무시하기 때문에 스스로 좋은 상태로의 진행을 계획할 수 없다. 또한, 합성 오차(compounding error)로 인해 해당 에이전트는 분포 외 상태(out-of-distribution)가 발생하는 경우 적절한 행동을 결정할 수 없다. 이러한 단점으로 인해 BC 학습된 에이전트가 본 발명에 따른 방법이 좋은 성능을 달성하는 것을 방해할 수 있다.

[0083] 상기의 반복(recurrency)을 모델링하기 위해, 해당 문제는 마르코프 결정 과정(MDP, Markov Decision Process)으로 공식화될 수 있다. 시간 단계(timestep) t에서, 상태 s_t는 $\{q_a^t; q_d^t; pt\}$와 같이 표현될 수 있다. 여기에서, q_a^t = [α_{t-n},...,α_t]이고 q_d^t = [d_{t-n-1},...,d_{t-1}]이다. 결정 d_t는 이산 결정 공간(discrete decision space) {0,1}에 존재할 수 있다. 여기에서 d_t = 1은 $x_{k(t-1):kt}$가 액션 프레임들을 나타낼 수 있고, 그렇지 않은 경우에는 배경(background) 프레임들을 나타낼 수 있다. MDP의 전환 역학(transition dynamics)은 시간 단계 t+1에서 에이전트가 알 수 없는 α_{t+1} 및 p_{t+1}에 직면하기 때문에 확률적일 수 있다. 즉, 모델은 미래에 직면하게 될 상태에 대해 오직 부분적으로만 제어할 수 있다.

[0084] 강화 학습(RL, Reinforcement Learning)은 경로의 즉각적인 보상(immediate reward)이 아닌 누적 보상

(cumulative reward)을 최대화하는 것이 목표이기 때문에 모델이 MDP 설정 하에서 계획되도록 할 수 있다. 즉, RL은 모델이 MDP의 전환 역학을 인식하고 예측할 수 있도록 하여 BC의 주요 한계(limitation)를 해결할 수 있다.

[0085] 그럼에도 불구하고, MDP에 대한 공정한 보상이 무엇인지는 여전히 문제로 존재할 수 있다. 도 7과 같이, 더 나은 프레임 단위 OAD 성능이 더 나은 행동 제안 성능을 보장하지 않을 수 있다.

[0086] 이 문제를 해결하는 한 가지 방법은 정교한 보상 함수(sophisticated reward function)를 적용하는 것일 수 있다. 예를 들어, 연속적인 올바른 결정에 대해 보상하는 경우 행동이 단편화(fragmentation)되는 것을 방지할 수 있다. 그러나, 해당 보상 함수는 긴 액션 세그먼트(long action segment)들에 과도한 보상(excessive reward)을 부여할 수 있고, 액션 틱(action tick)들에 대해서는 아무 것도 부여하지 않을 수 있다. 따라서, RL을 CAG 학습에 적용하기 위해서는 적절한 보상 함수에 대한 철저한 검색(exhaustive search)이 필수적일 수 있다.

[0087] 모방 학습(IL, Imitation Learning)에서 작업의 목표는 주어진 전문가 경로(expert trajectory)들에 의해 정의될 수 있다. 즉, 어떠한 보상 신호(reward signal)도 사용될 수 없으며, 모델은 주어진 전문가 전환(상태, 행동, 다음 상태)들만 사용하여 좋은 정책(good policy)이 무엇인지를 결정하여야 한다. 모델의 학습 절차는 MDP의 전환 역학을 통합해야 하기 때문에, 많은 IL 방법들은 주어진 전문가 전환들로부터 보상 함수 R(state,action)을 복원할 수 있다.

[0088] SQIL은 완전히 다른 접근 방식을 제공할 수 있다. 기본적인 직관은 매우 간단할 수 있다. 궁극적인 목표는 전문가를 모방하는 것이므로 전문가 전환(expert transition)은 좋은 전환(good transition)으로 간주되어야 한다. 즉, 전문가 전환의 보상은 1로 설정되고 에이전트의 전환 보상은 0으로 설정되어야 한다. 이러한 전환들은 리플레이 버퍼(replay buffer)에 저장될 수 있으며, soft-Q learning이 실행될 때 사용될 수 있다. SQIL은 적대적 학습(adversarial traning)을 활용하는 다른 방법보다 더 안정적인 경향을 나타낼 수 있다. 또한, SQIL은 보상 함수 R(state, action)을 추정하기 위해 별도의 네트워크를 사용하지 않고 Q 함수를 직접 근사하므로 보상 함수 근사(reward function approximation)를 포함하는 다른 방법보다 더 파라미터 효율적(parameter efficient)일 수 있다.

[0089] 모방 학습은 사전 정의된 보상 방법을 요구하지 않으면서도 MDP 구조를 충분히 활용하기 때문에, BC와 RL에 관한 상기의 문제들을 완화할 수 있다. 따라서, 컨텍스트 인지 행동성 그룹핑(CAG)을 위한 학습 방법에는 모방 학습, 특히 SQIL이 적용될 수 있다.

[0090] 그럼에도 불구하고, SQIL을 CAG에 직접 적용하는 경우 만족스럽지 못한 결과가 도출될 수 있다. 즉, CAG는 하나의 잘못된 결정을 기초로 액션 틱 또는 단편화를 야기할 수 있으며, 이는 soft-Q 설정의 최대 엔트로피 가정(maximum entropy assumption)과의 불일치(mismatch)가 원인일 수 있다.

[0091] 따라서, 널리 사용되는 DQN 방법과 유사한 SQIL의 hard-Q 변형(variant)이 적용될 수 있다. 원래의 DQN과의 유일한 차이점은 Q-함수를 근사화하기 위해 일정한 보상 +0.1을 갖는 주어진 전문가 전환들(expert transitions)과 일정한 보상 -0.1을 갖는 에이전트 전환들(agent transitions)을 사용하는 것일 수 있다. 구체적으로, Q 네트워크 $Q_\theta(s, d)$는 다음의 수학식 1과 같이 갱신될 수 있다.

[0092] [수학식 1]

[0093] $$\theta \leftarrow \theta - \eta \nabla_\theta (\delta^2(D_{expert}, +0.1) + \delta^2(D_{agent}, -0.1)),$$

[0094] $$where\ \delta^2(D,r) \triangleq$$

[0095] $$\frac{1}{|D|} \sum_{(s_t, d_t, s_{t+1}) \in D} (Q_\theta(s_t, d_t) - (r + \gamma \max_{d_{t+1}} Q_\theta(s_{t+1}, d_{t+1})))^2$$

[0097] 여기에서, D는 아래 첨자(subscript)에서 가져온 전환들로 구성된 미니 배치(minibatch)이고 γ와 η는 각각 할인율(discount factor)와 학습률(learning rate)이다. 이하, 해당 방법은 SQIL에서 'soft' 접두사(prefix)를 뺀 것과 같이, Q-Imitation Learning(QIL)이라 한다.

[0098] 전체 학습 절차는 2단계로 이루어질 수 있으며, 이는 OAD 모델 M1, M2 및 컨텍스트 인식 에이전트 γ가 개별로 학습됨을 의미할 수 있다. 학습의 첫 번째 단계에서는, 단순한 단일 레이어 LSTM 네트워크(one layer LSTM network)와 2개의 추가적인 FC 레이어들로 구성된 이진 OAD 모델 M_1과 다중 클래스 OAD 모델 M_2를 교차 엔트로피 손실(cross entropy loss)을 이용하여 학습시킬 수 있다. 그 후, 학습된 OAD 모델을 사용하여 모든 학습 비디오들에 대한 α(actionness) 및 p_t(class probability) 시퀀스들이 산출될 수 있다. 산출된 시퀀스들로부터 모방 학습(imitation learning)을 위한 전문가 데이터베이스(expert database)가 도 8과 같이 구축될 수 있다.

[0099] 학습의 두 번째 단계에서는, LeakyReLU 활성화 함수를 갖는 두 개의 FC 레이어들로 구성된 컨텍스트 인식 에이전트 γ를 학습시키기 위해 MDP 환경(도 9 참조)에서 QIL(상기 수학식 1)이 실행될 수 있다. 제안 분류(proposal classification)를 위해, 독립적인 TSM 분류기(classifier) C가 사용될 수 있고, 분류기로부터의 클래스 확률은 mAP를 계산할 때 신뢰 점수(confident score)로 간주될 수 있다.

[0101] 이하, 도 10 내지 15를 참조하여 본 발명에 따른 실시간 비디오 동작 검출 방법에 관한 실험 결과를 구체적으로 설명한다.

[0102] 여기에서는, 두 가지 표준 데이터셋, 즉 THUMOS14 및 Activitynet1.3을 통해 본 발명에 따른 방법을 검증한다. THUMOS14는 20개의 액션 클래스와 413개의 무편집 비디오(untrimmed video)들을 포함할 수 있으며, 무편집 비디오들은 200개의 학습용 비디오(training video)들과 213개의 테스트용 비디오(test video)들로 분할될 수 있다. 데이터셋에는 비디오당 15개 이상의 액션 인스턴스들이 포함될 수 있다. Activitynet1.3은 19,994개의 무편집 비디오들로 구성될 수 있으며, 무편집 비디오들은 2:1:1의 비율로 학습용, 검증용 및 테스트용 셋으로 분류될 수 있다. THUMOS14와 달리, Activitynet1.3의 비디오들은 비디오당 1.5개의 액션 인스턴스들이 포함될 수 있으며, 이는 스트리밍 설정에서 여러 액션 인스턴스들을 감지하는 것이 주요 목표인 점을 고려하면 해당 데이터셋이 가장 적합한 데이터셋이 아님을 의미할 수 있다.

[0103] 프레임 특징(frame feature) f을 위해, 6개의 연속된 프레임들(k = 6)은 Kinetics에서 학습된 2-스트림 TSN(two-stream TSN)에 입력될 수 있으며 그 출력이 사용될 수 있다. Activitynet1.3에 대해, 각 비디오의 특징 시퀀스는 선형 보간법(linear interpolation)을 통해 길이 100으로 재조정될 수 있다.

[0104] 그러나, THUMOS14에서 ODAS 성능을 평가할 때에는 다른 ODAS 작업들과의 공정한 비교를 보장하기 위해 특징들을 사전학습한 Activitynet이 사용될 수 있다.

[0105] TAL 메트릭(metric)의 경우, 본 발명에 따른 모델과 다른 오프라인 TAL 모델 간의 공정한 비교를 위해 평균 정밀도 평균(mAP, mean Average Precision)이 사용될 수 있다. 여기에서는, THUMOS14의 경우 {0.3, 0.4, 0.5, 0.6, 0.7} 집합, Acticitynet1.3의 경우 {0.5, 0.75, 0.95} 집합에서 다중 tIOU들을 갖는 mAP들이 사용될 수 있다.

[0106] ODAS 성능을 평가하기 위해, 포인트 레벨 평균 정밀도(p-AP, point-level average precision)를 측정할 수 있고, 모든 액션 클래스들에 대한 p-AP를 평균하여 p-mAP를 계산할 수 있다. 각 액션 클래스에 대해, 감지된 시작점들은 신뢰도 점수(confidence score)에 따라 내림차순(descending order)으로 정렬될 수 있고, 그에 따라 AP가 측정될 수 있다. 각 액션 시작점은 해당 액션 클래스가 정확하고 정답 포인트(ground truth point)로부터의 시간적 거리(temporal distance)가 오프셋(offset)보다 작은 경우에만 진실 양성(true positive)으로 카운트될 수 있다. 또한, 동일한 정답 포인트(ground truth point)에 대해 중복 예측(duplicate prediction)은 허용되지 않을 수 있다.

[0107] 본 발명에 따른 모델은 OAD 출력을 중간 표현(intermediate representation)으로 사용하고 프레임 단위(frame-wise) 레이블을 제공하지 않기 때문에 다른 OAD 모델과의 비교가 적절하지 않을 수 있다. 컨텍스트 인지 에이전트(context-aware agent)의 목표는 OAD 출력을 후처리(post-process)하고 유효한 액션 인스턴스들을 생성하는 것일 수 있다.

[0108] 한편, 본 발명에 따른 접근 방식의 효율성(effectiveness)을 입증하기 위해 본 발명에 따른 방법은 다양한 베이스라인(baseline)들과 비교될 수 있다.

[0109] a) OAD-Grouping: 모델은 바이너리 OAD 모델의 결과를 그룹핑하여 액션 인스턴스들을 생성할 수 있다. 일정한

기준값(constant threshold)을 가진 모델과 동일할 수 있다.

[0110] b) OAD-Grouping w/ Hindsight Threshold: 모델은 각 클래스에 대해 서로 다른 기준값을 할당하여 가장 높은 클래스 AP를 생성할 수 있다. 해당 모델은 테스트셋(test set)을 사용하고 각 클래스마다 가장 성능이 좋은 기준값을 찾기 위해 ([0.3:0.05:0.7])에서 그리드 검색(grid search)을 수행하기 때문에 'hindsight threshold model'로 표현될 수 있다. 해당 설정에서 mAP는 각 클래스의 가장 높은 AP를 평균하여 산출될 수 있다. 해당 모델은 mAP를 계산할 때 클래스 별로 서로 다른 최고 성능의 기준값을 사용하기 때문에 정교한 기준값 조정(sophisticated threshold tuning)의 상한(upper bound)으로 간주될 수 있다.

[0111] c) OAD-Grouping w/ Temporal Smoothing: 바이너리 OAD 모델의 출력에 임시 스무딩 필터(temporal smoothing filter)를 추가적으로 적용할 수 있다. 구체적으로, k 크기의 평균 필터(average filter)가 적용될 수 있다. 이에 따라, $k \in \{3, 5, 7, 9, 11, 13\}$에 관한 다양한 크기가 적용될 수 있으며, 결과적으로 k = 5가 가장 성능이 좋은 필터 크기임이 도출될 수 있다. 이러한 방식을 통해 액션 틱 및 단편화(action tick and fragmentation)를 어느 정도 완화할 수 있다.

[0112] d) CAG-BC: 모델은 표준 지도 학습 방법(행동 복제라고도 함)을 사용하여 상태(q_a; q_d; p)에서 결정 d로의 직접 매핑을 학습할 수 있다. CAG-BC는 CAG-QIL과 동일한 입력 및 출력 형식을 사용하지만 매핑 학습을 위한 알고리즘은 서로 상이할 수 있다.

[0113] e) CAG-RL: 모델은 도 7과 같이 수작업 보상 체계(프레임 단위 OAD보상)로 훈련된 에이전트에 해당할 수 있다. 또한, 모델은 CAG-QIL과 마찬가지로 DQN 알고리즘을 사용하여 강화 학습을 수행할 수 있다.

[0114] 도 10을 참조하면, THUMOS14 데이터셋에서 본 발명에 따른 방법인 CAG-QIL의 효과가 명확히 도시되어 있다. 추가 기법(additional trick)을 갖는 OAD 그룹핑 모델들은 실세계의 On-TAL 문제에 직접 적용될 수 없다. 즉, 사후적 기준값(hindsight threshold)을 갖는 모델들은 테스트 셋을 사용하여 최상의 클래스별 기준값을 선택해야 하고, 시간적 스무딩(temporal smoothing)을 사용하는 모델들은 온라인으로 실행할 수 없기 때문일 수 있다. 단순하지 않은 OAD 그룹핑 모델들과 큰 차이를 보이는 것은 CAG-QIL이 On-TAL을 처리하는 탁월한 능력을 가지고 있음을 나타낼 수 있다. 또한, 도 11의 결과는 CAG-QIL이 Activitynet1.3 등의 다른 데이터 셋에서도 잘 수행됨을 나타낼 수 있다.

[0115] On-TAL에 대한 첫 번째 작업으로서, 동일한 조건 하에서 비교 가능한 방법들이 존재하지 않을 수 있다. 따라서, 도 12에서 출력이 모두 동일한 최근의 offline TAL 방법들을 이용하여 본 발명에 따른 방법과의 비교를 수행할 수 있다. 본 발명에 따른 방법과 최근의 offline 시간적 행동 국지화(TAL) 방법들 사이에는 여전히 성능 차이가 존재하지만, 강력한 제약에도 불구하고 본 발명에 따른 방법의 성능은 최근의 단일 스테이지(one-stage) 오프라인 접근 방식과 유사할 수 있다. 도 12에서, 오직 [39] 방법과 본 발명에 따른 방법(CAG-QIL)만이 NMS와 같은 후처리(post processing)를 포함하지 않을 수 있다.

[0116] 모델의 견고성(robustness)을 검증하기 위해, THUMOS14에서 각 모델에 대해 3개의 추가 실험들을 수행할 수 있다. 각 시도(each trial)에서, OAD 모델 M1, M2에 대해 서로 다른 가중치가 사용될 수 있다. OAD 모델에 다른 가중치를 사용하면 비디오들의 미리 계산된 α 및 p 시퀀스가 달라질 수 있으며, CAG 학습에 대해 상이한 데이터가 생성될 수 있다. 모든 시도는 베이스라인(OAD-그룹핑 모델) 성능이 상이하기 때문에 알고리즘과 베이스라인의 mAP 점수 간의 차이를 측정함으로써 베이스라인보다 알고리즘의 개선을 도출할 수 있다(도 13 참조). 해당 결과는 CAG-QIL이 베이스라인 뿐만 아니라 강화 학습이 적용된 CAG보다 일관되게 더 높은 성능을 보이고 있음을 나타낼 수 있고, 이와 달리 CAG-BC는 지속적으로 더 낮은 성능을 보이고 있음을 나타낼 수 있다.

[0117] 도 14를 참조하면, CAG-QIL의 효과가 정성적으로 도시되어 있다. OAD 그룹핑(OAD-Grouping) 및 CAG-QIL은 동일한 OAD 모델 M1을 공유하기 때문에 입력으로서 동일한 α 시퀀스를 포함할 수 있다. CAG-QIL은 α 시퀀스가 합리적으로 계산되는 경우 결정 컨텍스트(decision context)를 활용하여 액션 단편화 문제(action fragmentation issue)를 성공적으로 해결할 수 있다.

[0118] 전체적인 실험 결과에 따르면, CAG-RL이 TAL 성능을 거의 향상시키지 않는다는 사실이 도출될 수 있다. 즉, 프레임 단위(frame-wise) OAD 보상은 도 7에서 예상한 것처럼 액션 단편화에 페널티를 부여하지 않을 수 있다. 해당 내용은 CAG-RL이 THUMOS14 테스트 셋에서 OAD 그룹핑 베이스라인(4841)보다 제안 번호(5641)를 감소시키지 않는다는 사실로서 증명될 수 있다.

[0119] 경량화된 OAD 모델들(light-weighted OAD models)과 컨텍스트 인지 에이전트(context-aware agent) 덕분에, 본

발명에 따른 방법은 계산 비용이 낮을 수 있으며 온라인으로 실행될 수 있다. 또한, 주요 병목 현상(bottleneck)은 6개의 중첩된 프레임들에 대해 126ms가 소요되는 광학 흐름 계산(OAD에 사용됨)에서 발생할 수 있다. 실험에서 전체 파이프라인은 29.4 fps로 실행될 수 있다.

[0120] 본 발명에 따른 모델은 액션 인스턴스 생성의 부산물(by-product)로 액션 시작점(action start point)을 신속하게 산출할 수 있기 때문에, 행동이 발생하는 즉시 액션 시작점의 발생과 클래스를 감지하는 것을 목표로 하는 액션 개시의 온라인 탐지(Online Detection of Action Start, ODAS) 작업에 대한 메트릭으로 시작점 생성 성능을 평가할 수 있다. 그러나, 본 발명에 따른 분류 절차는 액션 종료 시점으로 연기되고 전체 액션 인스턴스를 사용하기 때문에, 직접 비교는 본 발명에 따른 모델에 불공정한 이익을 제공할 수 있다. 이를 방지하고 공정한 비교를 보장하기 위해, 독립적인 분류기 C의 출력을 사용하는 대신 다중 클래스 OAD 모델 M_2의 출력 p_t(여기에서, t는 감지된 액션 시작 시점에 해당함)를 사용할 수 있다.

[0121] 도 15를 참조하면, 현재 ODAS의 최신 모델과 본 발명에 따른 모델(CAG-QIL)을 비교한 결과가 도시되어 있다. 본 발명에 따른 모델이 ODAS 작업을 해결하도록 특별히 학습되지 않았음에도 불구하고 오프셋마다 StartNet[13]보다 성능이 더 높을 수 있다. 즉, 본 발명에 따른 모델은 정확한 시작점을 감지할 수 있다.

[0122] 결과적으로, 본 발명에 따른 실시간 비디오 동작 검출 방법은 새롭게 정의된 온라인 시간적 행동 국지화(On-TAL, Online Temporal Action Localization)에 대한 해결책으로서 Q 모방 학습(Q Imitation Learning, QIL) 프레임워크를 통한 컨텍스트 인지 행동성 그룹핑(Context-Aware Actionness Grouping, CAG), 즉 CAG-QIL을 포함할 수 있다.

[0124] 도 16은 본 발명에 따른 실시간 비디오 동작 검출 장치의 시스템 구성을 설명하는 도면이다.

[0125] 도 16을 참조하면, 실시간 비디오 동작 검출 장치(100)는 프로세서(1610), 메모리(1630), 사용자 입출력부(1650) 및 네트워크 입출력부(1670)를 포함할 수 있다.

[0126] 프로세서(1610)는 본 발명의 실시예에 따른 실시간 비디오 동작 검출 프로시저를 실행할 수 있고, 이러한 과정에서 읽혀지거나 작성되는 메모리(1630)를 관리할 수 있으며, 메모리(1630)에 있는 휘발성 메모리와 비휘발성 메모리 간의 동기화 시간을 스케줄 할 수 있다. 프로세서(1610)는 실시간 비디오 동작 검출 장치(100)의 동작 전반을 제어할 수 있고, 메모리(1630), 사용자 입출력부(1650) 및 네트워크 입출력부(1670)와 전기적으로 연결되어 이들 간의 데이터 흐름을 제어할 수 있다. 프로세서(1610)는 실시간 비디오 동작 검출 장치(100)의 CPU(Central Processing Unit)로 구현될 수 있다.

[0127] 메모리(1630)는 SSD(Solid State Disk) 또는 HDD(Hard Disk Drive)와 같은 비휘발성 메모리로 구현되어 실시간 비디오 동작 검출 장치(100)에 필요한 데이터 전반을 저장하는데 사용되는 보조기억장치를 포함할 수 있고, RAM(Random Access Memory)과 같은 휘발성 메모리로 구현된 주기억장치를 포함할 수 있다. 또한, 메모리(1630)는 전기적으로 연결된 프로세서(1610)에 의해 실행됨으로써 본 발명에 따른 실시간 비디오 동작 검출 방법을 실행하는 명령들의 집합을 저장할 수 있다.

[0128] 사용자 입출력부(1650)은 사용자 입력을 수신하기 위한 환경 및 사용자에게 특정 정보를 출력하기 위한 환경을 포함하고, 예를 들어, 터치 패드, 터치 스크린, 화상 키보드 또는 포인팅 장치와 같은 어댑터를 포함하는 입력장치 및 모니터 또는 터치 스크린과 같은 어댑터를 포함하는 출력장치를 포함할 수 있다. 일 실시예에서, 사용자 입출력부(1650)은 원격 접속을 통해 접속되는 컴퓨팅 장치에 해당할 수 있고, 그러한 경우, 실시간 비디오 동작 검출 장치(100)는 독립적인 서버로서 수행될 수 있다.

[0129] 네트워크 입출력부(1670)은 네트워크를 통해 사용자 단말(1710)과 연결되기 위한 통신 환경을 제공하고, 예를 들어, LAN(Local Area Network), MAN(Metropolitan Area Network), WAN(Wide Area Network) 및 VAN(Value Added Network) 등의 통신을 위한 어댑터를 포함할 수 있다. 또한, 네트워크 입출력부(1670)는 데이터의 무선 전송을 위해 WiFi, 블루투스 등의 근거리 통신 기능이나 4G 이상의 무선 통신 기능을 제공하도록 구현될 수 있다.

[0131] 도 17은 본 발명에 따른 실시간 비디오 동작 검출 시스템을 설명하는 도면이다.

[0132] 실시간 비디오 동작 검출 시스템(1700)은 사용자 단말(1710), 실시간 비디오 동작 검출 장치(100) 및 데이터베

이스(1730)를 포함할 수 있다.

[0133] 사용자 단말(1710)은 사용자에 의해 운용되는 단말 장치에 해당할 수 있다. 본 발명의 실시예에서 사용자는 하나 이상의 사용자로 이해될 수 있으며, 복수의 사용자들은 하나 이상의 사용자 그룹으로 구분될 수 있다. 또한, 사용자 단말(1710)은 실시간 비디오 동작 검출 시스템(1700)을 구성하는 하나의 장치로서 실시간 비디오 동작 검출 장치(100)와 연동하여 동작하는 컴퓨팅 장치에 해당할 수 있다. 예를 들어, 사용자 단말(1710)은 실시간 비디오 동작 검출 장치(100)와 연결되어 동작 가능한 스마트폰, 노트북 또는 컴퓨터로 구현될 수 있으며, 반드시 이에 한정되지 않고, 태블릿 PC 등 포함하여 다양한 디바이스로도 구현될 수 있다. 또한, 사용자 단말(1710)은 실시간 비디오 동작 검출 장치(100)와 연동하기 위한 전용 프로그램 또는 어플리케이션(또는 앱, app)을 설치하여 실행할 수 있다.

[0134] 실시간 비디오 동작 검출 장치(100)는 본 발명에 실시간 비디오 동작 검출 방법을 수행하는 컴퓨터 또는 프로그램에 해당하는 서버로 구현될 수 있다. 또한, 실시간 비디오 동작 검출 장치(100)는 사용자 단말(1710)과 유선 네트워크 또는 블루투스, WiFi, LTE 등과 같은 무선 네트워크로 연결될 수 있고, 네트워크를 통해 사용자 단말(1710)과 데이터를 송·수신할 수 있다.

[0135] 또한, 실시간 비디오 동작 검출 장치(100)는 관련 동작을 수행하기 위하여 독립된 외부 시스템(도 1에 미도시함)과 연결되어 동작하도록 구현될 수 있다. 예를 들어, 실시간 비디오 동작 검출 장치(100)는 포털 시스템, SNS 시스템, 클라우드 시스템 등과 연동하여 다양한 서비스를 제공하도록 구현될 수 있다.

[0136] 데이터베이스(1730)는 실시간 비디오 동작 검출 장치(100)의 동작 과정에서 필요한 다양한 정보들을 저장하는 저장장치에 해당할 수 있다. 예를 들어, 데이터베이스(1730)는 비디오에 관한 정보를 저장할 수 있고, 학습 데이터와 모델에 관한 정보를 저장할 수 있으며, 반드시 이에 한정되지 않고, 실시간 비디오 동작 검출 장치(100)가 본 발명에 따른 실시간 비디오 동작 검출 방법을 수행하는 과정에서 다양한 형태로 수집 또는 가공된 정보들을 저장할 수 있다.

[0137] 또한, 도 17에서, 데이터베이스(1730)는 실시간 비디오 동작 검출 장치(100)와 독립적인 장치로서 도시되어 있으나, 반드시 이에 한정되지 않고, 논리적인 저장장치로서 실시간 비디오 동작 검출 장치(100)에 포함되어 구현될 수 있음은 물론이다.

[0139] 상기에서는 본 발명의 바람직한 실시예를 참조하여 설명하였지만, 해당 기술 분야의 숙련된 당업자는 하기의 특허 청구의 범위에 기재된 본 발명의 사상 및 영역으로부터 벗어나지 않는 범위 내에서 본 발명을 다양하게 수정 및 변경시킬 수 있음을 이해할 수 있을 것이다.

부호의 설명

[0141] 100: 실시간 비디오 동작 검출 장치

110: 피처 인코더부 130: 컨텍스트 인지 에이전트부

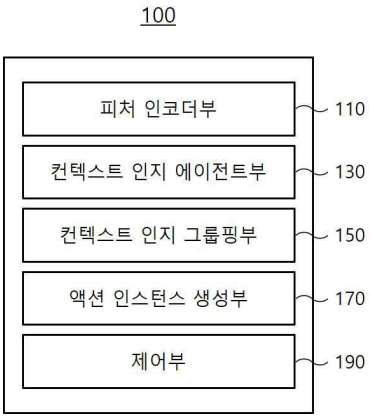
150: 컨텍스트 인지 그룹핑부 170: 액션 인스턴스 생성부

190: 제어부

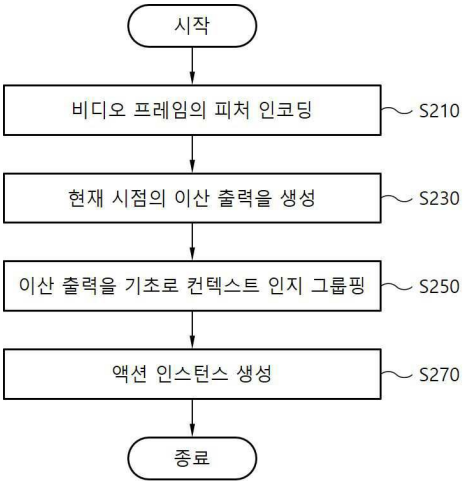
1700: 실시간 비디오 동작 검출 시스템

도면

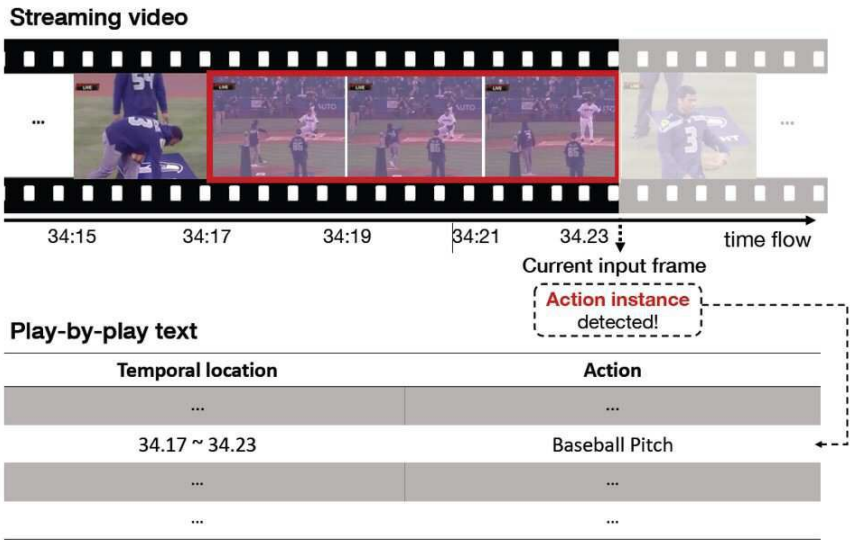
도면1



도면2



도면3



도면4

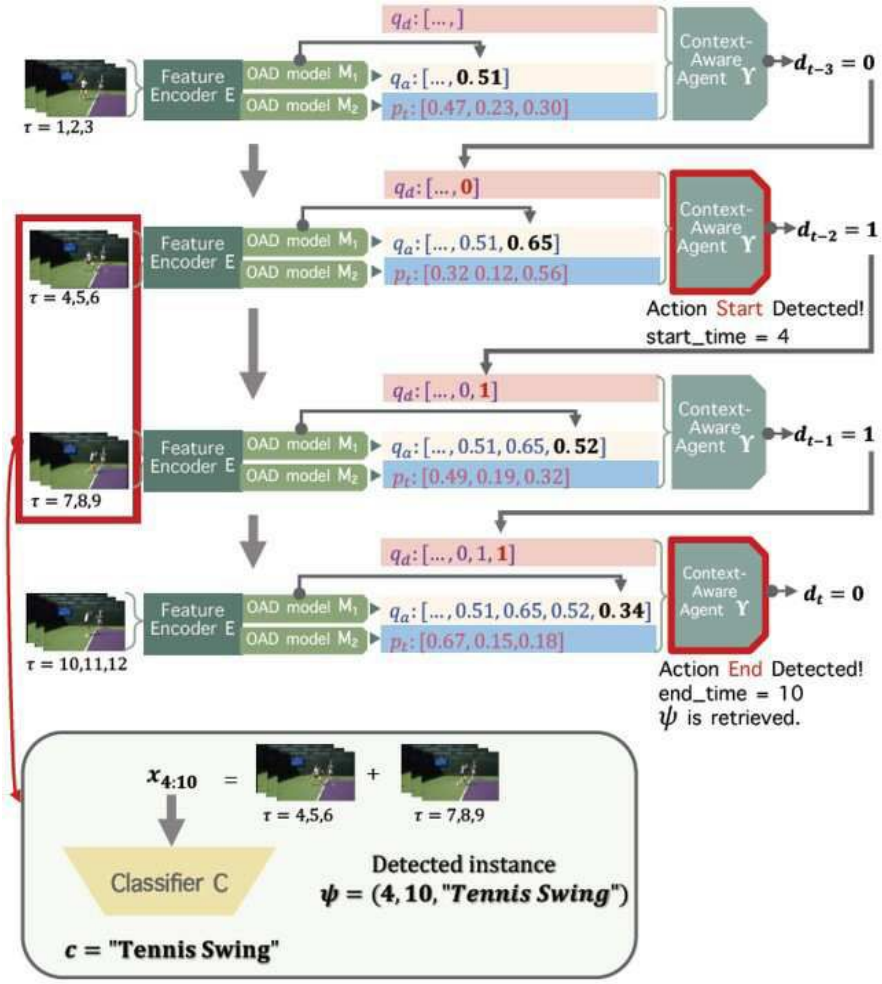
Ground Truth	0	0	0	0	0	0	0	0	0	0
α Sequence	0.34	0.41	0.53	0.45	0.41	0.33	0.44	0.51	0.43	0.36
OAD-Grouping	0	0	1	0	0	0	0	1	0	0

(a)

Ground Truth	0	0	1	1	1	1	1	1	1	1
α Sequence	0.45	0.51	0.55	0.62	0.52	0.49	0.55	0.57	0.60	0.56
OAD-Grouping	0	0	1	1	1	0	1	1	1	1

(b)

도면5



도면6

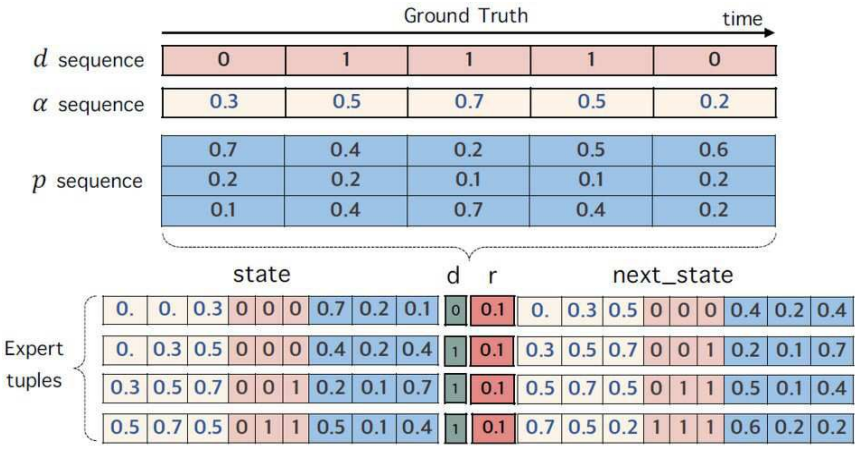
Algorithm 1: Context-Aware Actionness Grouping (CAG) at Inference Stage

Component:
Feature Encoder E
OAD model for actionness output M_1,
OAD model for class probability output M_2,
Context-Aware Agent Υ,
Classifier C,
Input: Video Stream $\{x_\tau\}_{\tau=1}^{T}$
Output: Action instance set Ψ

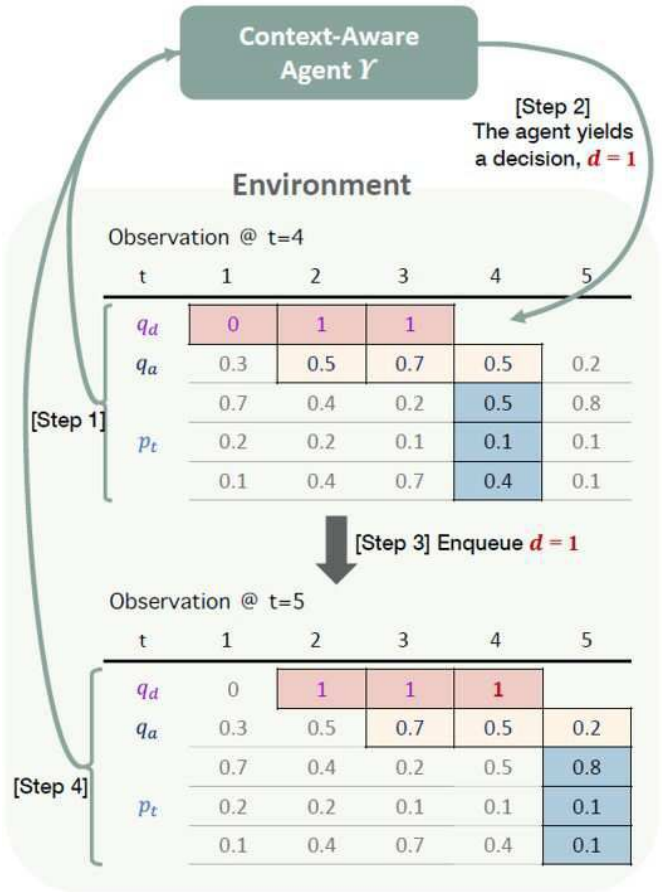
$\Psi \leftarrow \phi$
$d_{prev} \leftarrow 0$
$q_d.initialize()$
$q_a.initialize()$
for $t \leftarrow 1$ **to** $\lceil T/k \rceil$ **do**
  $f_t \leftarrow E(x_{k(t-1):kt})$
  $\alpha_t \leftarrow M_1(f_t)$
  $p_t \leftarrow M_2(f_t)$
  $q_a.dequeue()$
  $q_a.enqueue(\alpha_t)$
  $d \leftarrow \Upsilon(q_a, q_d, p_t)$
  if $d_{prev} = 0$ *and* $d = 1$ **then**
    $s \leftarrow k(t-1)+1$
  else if $d_{prev} = 1$ *and* $d = 0$ **then**
    $e \leftarrow k(t-1)+1$
    $c \leftarrow C(x_{s:e})$
    $\psi \leftarrow (s, e, c)$
    $\Psi \leftarrow \Psi \cup \psi$
  $q_d.dequeue()$
  $q_d.enqueue(d)$
  $d_{prev} \leftarrow d$
end

도면7

도면8


Ground Truth
time
d sequence
α sequence
p sequence
state
d
r
next_state
Expert tuples

도면9


Context-Aware Agent Y
[Step 2] The agent yields a decision, d = 1
Environment
Observation @ t=4
[Step 1]
[Step 3] Enqueue d = 1
Observation @ t=5
[Step 4]

도면10

Method	0.3	0.4	0.5	0.6	0.7
OAD-Grouping	33.3	28.0	22.0	16.8	10.4
w/ Hindsight Threshold	35.9	30.3	24.5	18.6	12.1
w/ Temporal Smoothing	38.3	32.4	24.9	18.2	10.7
CAG-BC	6.4	4.8	3.3	2.2	1.5
CAG-RL	32.8	27.0	22.2	16.8	10.8
CAG-QIL w/o p_t	43.0	34.9	27.2	19.6	12.4
CAG-QIL	**44.7**	**37.6**	**29.8**	**21.9**	**14.5**

도면11

Method	0.5	0.75	0.9
OAD-Grouping	28.1	15.7	3.3
CAG-BC	9.5	7.4	3.4
CAG-RL	21.4	11.3	2.2
CAG-QIL	**30.5**	**18.5**	**4.1**

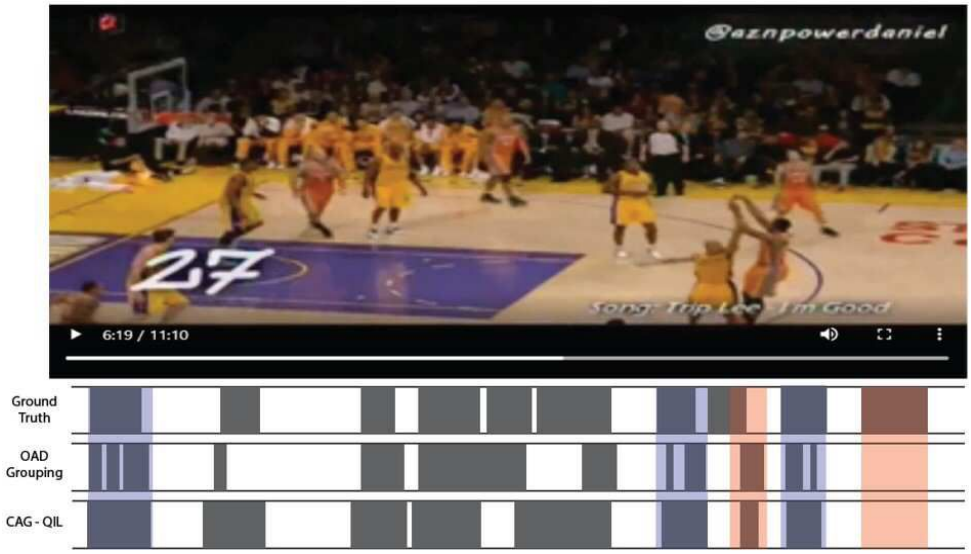
도면12

		Method	0.3	0.4	0.5	0.6	0.7
Offline	Two-Stage	S-CNN [30]	36.3	28.7	19.0	10.3	5.3
		CDC [28]	40.1	29.4	23.3	13.1	7.9
		SST [5]	41.2	31.5	20.0	10.9	4.7
		SSN [42]	51.9	41.0	29.8	-	-
		BSN [20]	53.5	45.0	36.9	28.4	20.0
		TAL-Net [7]	53.2	48.5	42.8	33.8	20.8
		G-TAD [37]	54.5	47.6	40.2	30.8	23.4
		G-TAD+P-GCN [37]	66.4	60.4	51.6	37.6	22.9
	One-Stage	End-to-End learning [39]	36.0	26.4	17.1	-	-
		SMS [40]	36.5	27.8	17.8	-	-
		SSAD [19]	43.0	35.0	24.6	-	-
		SS-TAD [4]	45.7	-	29.2	-	9.6
		GTAN [21]	57.8	47.2	38.8	-	-
Online		**CAG-QIL**	44.7	37.6	29.8	21.9	14.5

도면13

Method	0.3		0.4		0.5		0.6		0.7	
	Avg	SD	Avg	SD	Avg	SD	Avg	SD	Avg	SD
CAG-RL	0.4	1.5	-0.1	1.3	-0.1	0.8	0.0	0.8	0.0	0.8
CAG-BC	-25.2	5.4	-21.6	3.8	-17.2	2.7	-13.1	1.9	-8.6	1.0
CAG-QIL	12.6	1.1	9.7	0.6	7.5	0.7	5.2	1.2	3.5	1.5

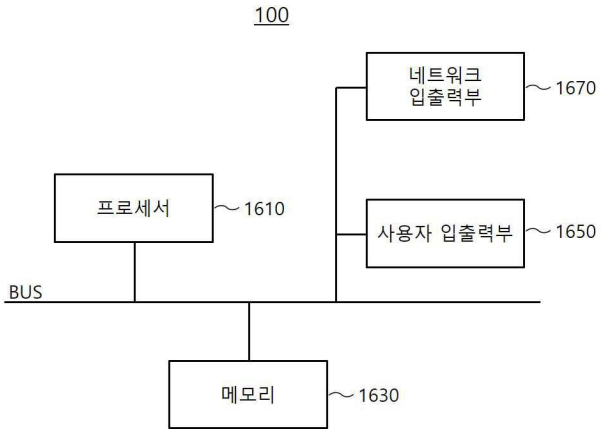
도면14



도면15

Offsets (second)	1	2	3	4	5	6	7	8	9	10
Shou et al. [29]	3.1	4.3	4.7	5.4	5.8	6.1	6.5	7.2	7.6	8.2
ClsNet-only [13]	13.9	21.6	25.8	28.9	31.1	32.5	33.5	34.3	34.8	35.2
StartNet-CE [13]	17.4	25.4	29.8	33.0	34.6	36.3	37.2	37.7	38.6	38.8
StartNet-PG [13]	19.5	27.2	30.8	33.9	36.5	37.5	38.3	38.8	39.5	39.8
CAG-QIL (Ours)	**20.3**	**31.2**	**37.2**	**41.4**	**44.2**	**46.0**	**47.3**	**48.1**	**48.9**	**49.8**

도면16



도면17