(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2020년11월05일
(11) 등록번호 10-2174656
(24) 등록일자 2020년10월30일

(51) 국제특허분류(Int. Cl.)
G06K 9/00 (2006.01) *G06N 3/08* (2006.01)
(52) CPC특허분류
G06K 9/00711 (2013.01)
G06N 3/08 (2013.01)
(21) 출원번호 10-2019-0034501
(22) 출원일자 2019년03월26일
심사청구일자 2019년03월26일
(65) 공개번호 10-2020-0119386
(43) 공개일자 2020년10월20일
(56) 선행기술조사문헌
KR101731461 B1
KR1020150065370 A
KR1020170070298 A

(73) 특허권자
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
손광훈
서울특별시 서대문구 연세로 50, 제3공학관 C129호(신촌동, 연세대학교)
이지영
서울특별시 서대문구 연세로 50, 제3공학관 C129호(신촌동, 연세대학교)
(74) 대리인
민영준

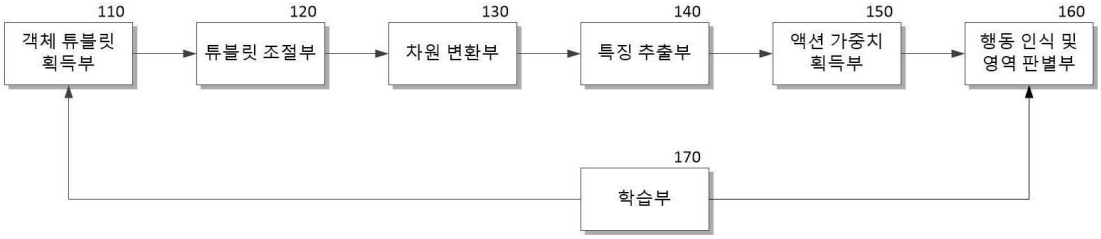
전체 청구항 수 : 총 12 항 심사관 : 황승희

(54) 발명의 명칭 **비디오 액션 인식 및 액션 영역 탐지 장치 및 방법**

(57) 요 약

본 발명은 액션 레이블만이 주석된 학습용 비디오를 이용하여 학습되어 학습용 비디오를 획득하기 위한 시간적 비용적 부담을 경감하고, 비디오에 포함된 객체의 액션을 인식하여 액션 영역을 정확하게 추출하여 액션 로컬라이제이션을 수행할 수 있는 비디오 액션 인식 및 액션 영역 탐지 장치 및 방법을 제공할 수 있다.

대 표 도



이 발명을 지원한 국가연구개발사업

과제고유번호	NRF-2018M3E3A1057289
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	복합인지기술개발사업
연구과제명	[Ezbaro] (2세부)이종 CCTV 영상에서의 딥러닝 기반 실종자 초동 신원확인 및 추적 시스템 (1단계)(1/2)
기 여 율	1/1
과제수행기관명	연세대학교 산학협력단
연구기간	2018.10.23 ~ 2019.04.22

명 세 서

청구범위

청구항 1

미리 학습된 제1 패턴 추정 방식에 따라 비디오의 다수 프레임 각각에서 기지정된 객체가 포함된 영역인 경계 박스를 탐색하고, 다수의 프레임에서 대응하는 경계 박스를 연결하여 객체 튜블릿을 생성하는 객체 튜블릿 획득부;

액션 레이블이 주석된 액션 학습용 비디오를 이용하는 약지도 학습 방식으로 제2 패턴 추정 방식이 미리 학습되어, 상기 객체 튜블릿의 다수의 경계 박스의 크기를 조절하여 튜블릿을 획득하는 튜블릿 조절부;

상기 튜블릿의 크기가 조절된 다수의 경계 박스를 시간 평균 풀링하여 튜블릿 이미지로 변환하고, 미리 학습된 제3 패턴 추정 방식에 따라 상기 튜블릿 이미지의 특징을 추출하여 특징맵을 생성하는 특징맵 획득부;

상기 특징맵에서 액션 가중치를 획득하여 대응하는 특징맵에 가중하여 가중 특징맵을 획득하는 액션 가중치 획득부; 및

상기 가중 특징맵이 기지정된 다수의 액션 클래스 각각에 대응하는 수준을 나타내는 액션 클래스 스코어를 계산하고, 상기 액션 클래스 스코어에 따라 튜블릿에 대응하는 액션을 선택하고, 상기 튜블릿에 포함된 상기 크기가 조절된 경계 박스의 위치 정보를 출력하는 액션 인식 및 영역 판별부; 를 포함하는 비디오 액션 인식 및 액션 영역 탐지 장치.

청구항 2

제1항에 있어서, 상기 튜블릿 조절부는

상기 제2 패턴 추정 방식에 따라 객체 튜블릿의 다수의 경계 박스(B_t^n) 각각의 폭(u_t^n)에 대한 조절 폭(∇u_t^n)과 높이(v_t^n)에 대한 조절 높이(∇v_t^n)를 획득하고, 획득된 조절 폭(∇u_t^n)과 조절 높이(∇v_t^n)로부터 수학식

$$\begin{bmatrix} \bar{u}_t^n \\ \bar{v}_t^n \end{bmatrix} = \begin{bmatrix} \nabla u_t^n & 0 \\ 0 & \nabla v_t^n \end{bmatrix} \begin{bmatrix} u_t^n \\ v_t^n \end{bmatrix}$$

에 따라 상기 크기가 조절된 경계 박스($\bar{B}_t^n$)를 획득하는 비디오 액션 인식 및 액션 영역 탐지 장치.

청구항 3

제1항에 있어서, 상기 액션 인식 및 영역 판별부는

인공 신경망을 포함하여 구성되고 비디오에 포함된 N(N은 자연수)개 튜블릿 중 n번째 튜블릿(P_n)에 대한 상기 액션 클래스 스코어($\lambda^n(c)$)를

수학식

$$\lambda^n(c) = \sum_d w_T(c,d)\alpha^n y^n(d)$$
$$= \alpha^n \sum_d w_T(c,d) y^n(d)$$

(여기서 α^n은 의 액션 가중치이고, y^n 은 특징맵이며, $w_T(c, d)$는 지정된 액션 클래스($c \in \{1, \ldots, C\}$)를 식별하기 위한 액션 클래스 분류자에 대응하는 d번째 요소로서 인공 신경망의 연산 레이어의 가중치를 나타낸다.)

에 따라 획득하는 비디오 액션 인식 및 액션 영역 탐지 장치.

청구항 4

제1항에 있어서, 상기 액션 인식 및 영역 판별부는

상기 액션 클래스 스코어 중 기지정된 기준 액션 클래스 스코어 이상인 액션 클래스 스코어를 선택하고, 선택된 액션 클래스 스코어에 대응하는 액션 클래스를 객체의 액션으로 출력하고, 선택된 액션 클래스 스코어에 대응하는 튜블릿의 크기가 조절된 경계 박스의 위치 정보를 출력하는 비디오 액션 인식 및 액션 영역 탐지 장치.

청구항 5

제4항에 있어서, 상기 액션 인식 및 영역 판별부는

동일한 튜블릿에 대해 기준 액션 클래스 스코어 이상인 액션 클래스 스코어가 다수개인 경우, 기지정된 설정에 따라 액션 클래스 스코어가 가장 높은 하나의 액션 클래스를 출력하거나, 기준 액션 클래스 스코어 이상으로 나타난 다수의 액션 클래스를 함께 출력하는 비디오 액션 인식 및 액션 영역 탐지 장치.

청구항 6

제1항에 있어서, 상기 비디오 액션 인식 및 액션 영역 탐지 장치는

액션 레이블만이 주석된 액션 학습용 비디오를 기반으로 상기 튜블릿 조절부, 상기 특징맵 획득부, 상기 액션 가중치 획득부 및 액션 인식 및 영역 판별부를 약지도 학습시키기 위한 학습부; 를 더 포함하고,

상기 학습부는

상기 액션 학습용 비디오에 응답하여, 액션 가중치 획득부(150)에서 모든 액션 튜블릿(P_n)에 대해 출력되는 가중 특징맵을 가산하여 비디오 특징맵을 획득하고, 비디오 특징맵으로부터 비디오 액션 클래스 스코어를 획득하며, 비디오 액션 클래스 스코어와 액션 학습용 비디오의 액션 레이블과의 차이를 액션 손실로 획득하여 역전파하여 약지도 학습을 수행하는 비디오 액션 인식 및 액션 영역 탐지 장치.

청구항 7

제6항에 있어서, 상기 객체 튜블릿 획득부는

상기 제1 패턴 추정 방식에 따라 비디오의 다수 프레임 각각에서 기지정된 객체가 포함된 영역인 경계 박스를 탐색하고, 각 경계 박스에 검출해야 하는 객체가 존재할 확률을 나타내는 객체 스코어를 함께 획득하고, 획득된 객체 스코어가 기지정된 기준 객체 스코어 이상인 경계 박스를 이용하여 객체 튜블릿을 생성하고,

상기 학습부는

객체 레이블만이 주석된 객체 학습용 비디오가 인가되어 상기 객체 튜블릿 획득부에서 획득된 상기 객체 스코어와 객체 학습용 비디오에 주석된 객체 레이블 사이의 차이를 객체 손실로 획득하여 역전파함으로써, 상기 객체 튜블릿 획득부를 약지도 학습시키는 비디오 액션 인식 및 액션 영역 탐지 장치.

청구항 8

미리 학습된 제1 패턴 추정 방식에 따라 비디오의 다수 프레임 각각에서 기지정된 객체가 포함된 영역인 경계 박스를 탐색하고, 다수의 프레임에서 대응하는 경계 박스를 연결하여 객체 튜블릿을 생성하는 단계;

액션 레이블이 주석된 액션 학습용 비디오를 이용하는 약지도 학습 방식으로 학습된 제2 패턴 추정 방식에 따라 상기 객체 튜블릿의 다수의 경계 박스의 크기를 조절하여 튜블릿을 획득하는 단계;

상기 튜블릿의 크기가 조절된 경계 박스를 시간 평균 풀링하여 튜블릿 이미지로 변환하고, 미리 학습된 제3 패턴 추정 방식에 따라 상기 튜블릿 이미지의 특징을 추출하여 특징맵을 생성하는 단계;

상기 특징맵에서 액션 가중치를 획득하여 대응하는 특징맵에 가중하여 가중 특징맵을 획득하는 단계; 및

상기 가중 특징맵이 기지정된 다수의 액션 클래스 각각에 대응하는 수준을 나타내는 액션 클래스 스코어를 계산하고, 상기 액션 클래스 스코어에 따라 튜블릿에 대응하는 액션을 선택하고, 상기 튜블릿에 포함된 상기 크기가

조절된 경계 박스의 위치 정보를 출력하는 단계; 를 포함하는 비디오 액션 인식 및 액션 영역 탐지 방법.

청구항 9

제8항에 있어서, 상기 위치 정보를 출력하는 단계는

인공 신경망을 포함하여 구성되고 비디오에 포함된 N(N은 자연수)개 튜블릿 중 n번째 튜블릿(P_n)에 대한 상기 액션 클래스 스코어($\lambda^n(c)$)를

수학식

$$\lambda^n(c) = \sum_d w_T(c,d)\,\alpha^n y^n(d)$$
$$= \alpha^n \sum_d w_T(c,d) y^n(d)$$

(여기서 α^n은 의 액션 가중치이고, y^n 은 특징맵이며, $w_T(c, d)$는 지정된 액션 클래스($c \in \{1, \ldots, C\}$)를 식별하기 위한 액션 클래스 분류자에 대응하는 d번째 요소로서 인공 신경망의 연산 레이어의 가중치를 나타낸다.)

에 따라 획득하는 비디오 액션 인식 및 액션 영역 탐지 방법.

청구항 10

제8항에 있어서, 상기 위치 정보를 출력하는 단계는

상기 액션 클래스 스코어 중 기지정된 기준 액션 클래스 스코어 이상인 액션 클래스 스코어를 선택하는 단계;

선택된 액션 클래스 스코어에 대응하는 액션 클래스를 객체의 액션으로 출력하는 단계; 및

선택된 액션 클래스 스코어에 대응하는 튜블릿의 크기가 조절된 경계 박스의 위치 정보를 출력하는 단계; 를 포함하는 비디오 액션 인식 및 액션 영역 탐지 방법.

청구항 11

제8항에 있어서, 상기 비디오 액션 인식 및 액션 영역 탐지 방법은

액션 레이블만이 주석된 액션 학습용 비디오를 기반으로 약지도 학습시키는 단계; 를 더 포함하고,

상기 약지도 학습시키는 단계는

상기 액션 학습용 비디오에 응답하여, 획득되는 모든 액션 튜블릿(P_n)에 대해 출력되는 가중 특징맵($\alpha^n y^n$)을 가산하여 비디오 특징맵(y^*)을 획득하는 단계;

상기 비디오 특징맵(y^*)으로부터 비디오 액션 클래스 스코어($\lambda(c)$)를 획득하는 단계; 및

비디오 액션 클래스 스코어($\lambda(c)$)와 액션 학습용 비디오의 액션 레이블과의 차이를 액션 손실로 획득하여 역전파하는 단계; 를 포함하는 비디오 액션 인식 및 액션 영역 탐지 방법.

청구항 12

제8항에 있어서, 상기 객체 튜블릿을 생성하는 단계는

상기 제1 패턴 추정 방식에 따라 비디오의 다수 프레임 각각에서 기지정된 객체가 포함된 영역인 경계 박스를 탐색하는 단계;

각 경계 박스에 검출해야 하는 객체가 존재할 확률을 나타내는 객체 스코어를 함께 획득하는 단계: 및

획득된 객체 스코어가 기지정된 기준 객체 스코어 이상인 경계 박스를 이용하여 객체 튜블릿을 생성하는 단계; 를 포함하고,

상기 약지도 학습시키는 단계는

객체 레이블만이 주석된 객체 학습용 비디오가 인가되어 상기 객체 튜블릿 획득부에서 획득된 상기 객체 스코어와 객체 학습용 비디오에 주석된 객체 레이블 사이의 차이를 객체 손실로 획득하여 역전파함으로써, 상기 객체 튜블릿 획득부를 약지도 학습시키는 단계; 를 더 포함하는 비디오 액션 인식 및 액션 영역 탐지 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 비디오 동작 인식 및 동작 구간 탐지 장치 및 방법에 관한 것으로, 비디오에서 객체의 동작을 인식하고 동작 영역을 추출할 수 있는 동작 인식 및 동작 구간 탐지 장치 및 방법에 관한 것이다.

배 경 기 술

[0002] 비디오에 포함된 객체의 액션을 인식하고, 액션 영역을 추출하는 것은 비디오 감시, 비디오 요약 및 비디오 캡션과 같은 다양한 비디오 이용 분야에서 필수적이다. 비디오에서 객체를 탐지하는 다양한 기술이 공개되었으며, 이로부터 객체의 액션을 인식하는 기법 또한 큰 발전을 이루어 왔으나, 액션의 위치를 정확하게 추출하는 것은 액션의 다양성과 복잡한 배경 등을 포함한 다양한 이유로 인해 성능의 제약이 있어왔다.

[0003] 이에 최근에는 딥 러닝(Deep learning) 기법으로 학습된 인공 신경망(artificial neural network)을 이용하여 비디오에서 객체의 액션 영역을 추출하는 액션 로컬라이제이션을 수행하기 위한 다양한 연구가 진행되었다. 딥 러닝 기법을 이용함에 의해 비디오에 대한 액션 로컬라이제이션 작업의 성능이 크게 향상되었다.

[0004] 기존의 딥러닝 기법에서 인공 신경망은 완전 지도(fully supervised) 학습 방식으로 학습되었다. 따라서 학습시에 학습용 비디오 내의 객체의 액션 경계에 대한 검증 자료 레이블(ground truth label)이 완전하게 주석(full annotation)될 것이 요구되었다.

[0005] 그러나 비디오에서 각 액션 각각에 대한 경계를 수작업으로 주석 처리하는 것은 시간적으로나 비용적으로 매우 비효율적이다. 뿐만 아니라, 각 액션의 경계는 작업자에 따라 주관적으로 판단될 수 있어, 인공 신경망을 부정확하게 학습시킬 수 있다는 문제가 있다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 한국 등록 특허 제10-1900237호 (2018.09.13 등록)

발명의 내용

해결하려는 과제

[0007] 획득이 용이한 간단한 액션 레이블만이 주석된 학습용 비디오를 이용하는 약지도 학습(weakly-supervised learning) 방식을 기반으로 학습시킬 수 있는 비디오 액션 인식 및 액션 영역 탐지 장치 및 방법을 제공하는데 있다.

[0008] 본 발명의 다른 목적은 약지도 학습으로 학습되어 비디오에 대한 액션 로컬라이제이션을 수행할 수 있는 비디오 액션 인식 및 액션 영역 탐지 장치 및 방법을 제공하는데 있다.

[0009] 본 발명의 또 다른 목적은 비디오로부터 객체의 액션 영역을 정확하게 추출할 수 있는 비디오 액션 인식 및 액션 영역 탐지 장치 및 방법을 제공하는데 있다.

과제의 해결 수단

[0010] 상기 목적을 달성하기 위한 본 발명의 일 실시예에 따른 비디오 액션 인식 및 액션 영역 탐지 장치는 미리 학습된 패턴 추정 방식에 따라 비디오의 다수 프레임 각각에서 기지정된 객체가 포함된 영역인 경계 박스를 탐색하고, 다수의 프레임에서 대응하는 경계 박스를 연결하여 객체 튜블릿을 생성하는 객체 튜블릿 획득부; 액션 레이

블이 주석된 액션 학습용 비디오를 이용하는 약지도 학습 방식으로 패턴 추정 방식이 미리 학습되어, 상기 객체 튜블릿의 다수의 경계 박스의 크기를 조절하여 튜블릿을 획득하는 튜블릿 조절부; 상기 튜블릿의 다수의 최적 경계 박스를 시간 평균 풀링하여 튜블릿 이미지로 변환하고, 미리 학습된 패턴 추정 방식에 따라 상기 튜블릿 이미지의 특징을 추출하여 특징맵을 생성하는 특징맵 획득부; 상기 특징맵에서 액션 가중치를 획득하여 대응하는 특징맵에 가중하여 가중 특징맵을 획득하는 액션 가중치 획득부; 및 상기 가중 특징맵이 기지정된 다수의 액션 클래스 각각에 대응하는 수준을 나타내는 액션 클래스 스코어를 계산하고, 상기 액션 클래스 스코어에 따라 튜블릿에 대응하는 액션을 선택하고, 튜블릿에 포함된 최적 경계 박스의 위치 정보를 출력하는 액션 인식 및 영역 판별부; 를 포함한다.

[0011] 상기 액션 인식 및 영역 판별부는 인공 신경망을 포함하여 구성되고 비디오에 포함된 N(N은 자연수)개 튜블릿 중 n번째 튜블릿(P_n)에 대한 상기 액션 클래스 스코어($\lambda^n(c)$)를 수학식

$$\lambda^n(c) = \sum_d w_T(c,d)\alpha^n y^n(d)$$
$$= \alpha^n \sum_d w_T(c,d) y^n(d)$$

[0012]

[0013] (여기서 α^n은 의 액션 가중치이고, y^n 은 특징맵이며, $w_T(c, d)$는 지정된 액션 클래스($c \in \{1, \ldots, C\}$)를 식별하기 위한 액션 클래스 분류자에 대응하는 d번째 요소로서 인공 신경망의 연산 레이어의 가중치를 나타낸다.)에 따라 획득할 수 있다.

[0014] 상기 액션 인식 및 영역 판별부는 상기 액션 클래스 스코어 중 기지정된 기준 액션 클래스 스코어 이상인 액션 클래스 스코어를 선택하고, 선택된 액션 클래스 스코어에 대응하는 액션 클래스를 객체의 액션으로 출력하고, 선택된 액션 클래스 스코어에 대응하는 튜블릿의 최적 경계 박스의 위치 정보를 출력할 수 있다.

[0015] 상기 액션 인식 및 영역 판별부는 동일한 튜블릿에 대해 기준 액션 클래스 스코어 이상인 액션 클래스 스코어가 다수개인 경우, 기지정된 설정에 따라 액션 클래스 스코어가 가장 높은 하나의 액션 클래스를 출력하거나, 기준 액션 클래스 스코어 이상으로 나타난 다수의 액션 클래스를 함께 출력할 수 있다.

[0016] 상기 비디오 액션 인식 및 액션 영역 탐지 장치는 액션 레이블만이 주석된 액션 학습용 비디오를 기반으로 상기 튜블릿 조절부, 상기 특징맵 획득부, 상기 액션 가중치 획득부 및 액션 인식 및 영역 판별부를 약지도 학습시키기 위한 학습부; 를 더 포함하고, 상기 학습부는 상기 액션 학습용 비디오에 응답하여, 액션 가중치 획득부(150)에서 모든 액션 튜블릿(P_n)에 대해 출력되는 가중 특징맵을 가산하여 비디오 특징맵을 획득하고, 비디오 특징맵으로부터 비디오 액션 클래스 스코어를 획득하며, 비디오 액션 클래스 스코어와 액션 학습용 비디오의 액션 레이블과의 차이를 액션 손실로 획득하여 역전파하여 약지도 학습을 수행할 수 있다.

[0017] 상기 객체 튜블릿 획득부는 미리 학습된 패턴 추정 방식에 따라 비디오의 다수 프레임 각각에서 기지정된 객체가 포함된 영역인 경계 박스를 탐색하고, 각 경계 박스에 검출해야 하는 객체가 존재할 확률을 나타내는 객체 스코어를 함께 획득하고, 획득된 객체 스코어가 기지정된 기준 객체 스코어 이상인 경계 박스를 이용하여 객체 튜블릿을 생성하고, 상기 학습부는 객체 레이블만이 주석된 객체 학습용 비디오가 인가되어 상기 객체 튜블릿 획득부에서 획득된 상기 객체 스코어와 객체 학습용 비디오에 주석된 객체 레이블 사이의 차이를 객체 손실로 획득하여 역전파함으로써, 상기 객체 튜블릿 획득부를 약지도 학습시킬 수 있다.

[0018] 상기 목적을 달성하기 위한 본 발명의 다른 실시예에 따른 비디오 액션 인식 및 액션 영역 탐지 장치 및 방법은 미리 학습된 패턴 추정 방식에 따라 비디오의 다수 프레임 각각에서 기지정된 객체가 포함된 영역인 경계 박스를 탐색하고, 다수의 프레임에서 대응하는 경계 박스를 연결하여 객체 튜블릿을 생성하는 단계;

[0019] 액션 레이블이 주석된 액션 학습용 비디오를 이용하는 약지도 학습 방식으로 학습된 패턴 추정 방식에 따라 상기 객체 튜블릿의 다수의 경계 박스의 크기를 조절하여 튜블릿을 획득하는 단계; 상기 튜블릿의 다수의 최적 경계 박스를 시간 평균 풀링하여 튜블릿 이미지로 변환하고, 미리 학습된 패턴 추정 방식에 따라 상기 튜블릿 이미지의 특징을 추출하여 특징맵을 생성하는 단계; 상기 특징맵에서 액션 가중치를 획득하여 대응하는 특징맵에 가중하여 가중 특징맵을 획득하는 단계; 및 상기 가중 특징맵이 기지정된 다수의 액션 클래스 각각에 대응하는 수준을 나타내는 액션 클래스 스코어를 계산하고, 상기 액션 클래스 스코어에 따라 튜블릿에 대응하는 액션을

선택하고, 튜블릿에 포함된 최적 경계 박스의 위치 정보를 출력하는 단계; 를 포함한다.

발명의 효과

[0020] 따라서, 본 발명의 실시예에 따른 비디오 액션 인식 및 액션 영역 탐지 장치 및 방법은 액션 레이블만이 주석된 학습용 비디오를 이용하여 학습되어 학습용 비디오를 획득하기 위한 시간적 비용적 부담을 경감할 수 있다. 또한 비디오에 포함된 객체의 액션을 인식하고, 액션 영역을 정확하게 추출하여 액션 로컬라이제이션을 수행할 수 있다.

도면의 간단한 설명

[0021] 도 1은 본 발명의 일 실시예에 따른 비디오 액션 인식 및 액션 영역 탐지 장치의 개략적 구조를 나타낸다.

도 2는 도 1의 특징맵 획득부의 상세 구성을 나타낸다.

도 3은 약지도 학습을 위한 액션 학습용 비디오의 일예를 나타낸다.

도 4는 도 1의 튜블릿 조절부에서 크기가 조절된 경계 박스의 일예를 나타낸다.

도 5 및 도 6은 본 실시예에 따른 비디오 액션 인식 및 액션 영역 탐지 장치에서 액션 로컬라이제이션이 수행된 결과의 일예를 나타낸다.

도 7은 본 발병의 일 실시예에 따른 비디오 액션 인식 및 액션 영역 탐지 방법을 나타낸다.

발명을 실시하기 위한 구체적인 내용

[0022] 본 발명과 본 발명의 동작상의 이점 및 본 발명의 실시에 의하여 달성되는 목적을 충분히 이해하기 위해서는 본 발명의 바람직한 실시예를 예시하는 첨부 도면 및 첨부 도면에 기재된 내용을 참조하여야만 한다.

[0023] 이하, 첨부한 도면을 참조하여 본 발명의 바람직한 실시예를 설명함으로써, 본 발명을 상세히 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.

[0024] 명세서 전체에서, 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 의미한다. 또한, 명세서에 기재된 "...부", "...기", "모듈", "블록" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어나 소프트웨어 또는 하드웨어 및 소프트웨어의 결합으로 구현될 수 있다.

[0025] 도 1은 본 발명의 일 실시예에 따른 비디오 액션 인식 및 액션 영역 탐지 장치의 개략적 구조를 나타내고, 도 2는 도 1의 객체 튜블릿 획득부의 상세 구성을 나타낸다.

[0026] 도 1을 참조하면, 본 실시예에 따른 비디오 액션 인식 및 액션 영역 탐지 장치는 객체 튜블릿 획득부(110), 튜블릿 조절부(120), 차원 변환부(130), 특징 추출부(140), 액션 가중치 획득부(150), 액션 인식 및 영역 판별부(160) 및 학습부(170)를 포함한다.

[0027] 객체 튜블릿 획득부(110)는 액션 로컬라이제이션이 수행되어야 하는 비디오를 획득하고, 획득된 비디오의 다수의 프레임에서 객체가 포함된 영역을 검출하여 객체 튜블릿(tubelet)을 생성한다. 객체 튜블릿 획득부(110)는 약지도 학습(weakly-supervised) 방식에 따라 패턴 추정 방식이 미리 학습된 인공 신경망을 포함하여, 비디오의 다수의 프레임에서 기지정된 객체를 검출하고, 연속되는 다수의 프레임에서 객체가 검출된 영역을 연결함으로써 객체 튜블릿을 생성한다.

[0028] 도 2를 참조하면, 객체 튜블릿 획득부(110)는 비디오 제공부(111), 프레임 그룹화부(112), 객체 검출부(113) 및 튜블릿 생성부(114)를 포함할 수 있다.

[0029] 비디오 제공부(111)는 액션 로컬라이제이션이 수행되어야 하는 다수 프레임으로 구성된 비디오를 획득한다. 여기서 액션 로컬라이제이션은 비디오 내에 포함된 적어도 하나의 객체의 액션 영역을 구분하는 것으로서, 객체의 액션을 인지하고, 각 프레임 내에서 인지된 객체의 액션이 포함된 영역을 추출하는 것을 의미한다.

[0030] 프레임 그룹화부(112)는 연속되는 다수의 프레임이 포함된 비디오에서 기지정된 개수(여기서는 일예로 T개(T는

자연수)) 단위로 프레임(f_t, t = {1, ..., T})을 그룹화하여 구분한다. 일반적으로 비디오에는 매우 많은 프레임이 포함되고, 비디오에 포함된 모든 프레임에서 객체가 동일하게 유지되는 경우는 거의 없다. 또한 스트리밍과 같이 비디오의 모든 프레임을 일괄적으로 획득할 수 없는 경우도 발생한다. 따라서 프레임 그룹화부(112)는 비디오에 포함된 객체를 용이하게 검출할 수 있도록, 비디오를 기지정된 개수의 프레임 단위로 그룹화한다. 여기서 그룹화되는 프레임의 개수는 다양하게 설정될 수 있으나, 일예로 8개의 프레임 단위로 그룹화될 수 있다.

[0031] 객체 검출부(113)는 패턴 추정 방식이 미리 학습된 인공 신경망을 포함하여, 학습된 패턴 추정 방식에 따라 그룹화된 다수의 프레임에 포함된 객체를 탐지하고, 다수의 프레임에서 탐지된 객체가 나타나는 객체 영역에 대한 경계 박스(B = {B_1, ..., B_T})를 검출한다.

[0032] 여기서 객체 검출부(113)는 검출할 객체가 지정되어 미리 학습될 수 있다. 본 실시예에서는 비디오 액션 인식 및 액션 영역 탐지 장치가 사람의 액션을 인식하고 액션 영역을 검출하는 것으로 가정하며, 이에 객체 검출부(113)는 그룹화된 다수의 프레임에서 사람이 포함된 영역을 검출한다.

[0033] 객체 검출부(113)는 그룹화된 다수의 프레임(f_t)에서 객체가 나타나는 객체 영역을 사각의 경계 박스(bounding box)(B)로 검출하고, 검출된 경계 박스(B)의 좌표를 출력할 수 있다.

[0034] 여기서 객체 검출부(113)는 일예로 컨볼루션 신경망(Convolutional Neural Networks)으로 구현될 수 있으며, 학습의 편의성을 위해 약지도 학습 방식에 따라 학습될 수 있다.

[0035] 객체 검출부(113)가 완전 지도 학습 방식으로 학습되는 경우, 객체 검출 성능이 우수하지만, 객체의 경계, 즉 경계 박스(B)에 대한 검증 자료 레이블이 완전하게 주석된 대량의 학습용 비디오를 필요로 한다. 그리고 검증 자료 레이블이 주석된 학습용 비디오는 기본적으로 수작업으로 획득되므로, 학습용 비디오를 획득하는 것이 용이하지 않다.

[0036] 그에 비해 본 실시예에서 객체 검출부(113)는 단순한 객체 레이블만이 주석된 객체 학습용 비디오를 기반으로 약지도 학습된다.

[0037] 단순 객체 레이블만이 주석된 객체 학습용 비디오는 객체 영역에 대한 별도의 주석 없이 비디오 전체에 대해 객체 레이블만이 제공되는 비디오를 의미한다. 일예로, 본 실시예에서 객체 학습용 비디오에는 사람, 개, 고양이, 염소 등과 같이 단순히 객체의 레이블만이 주석으로 제공되며, 객체가 나타나는 객체 영역에 대해서는 별도의 주석이 제공되지 않는다.

[0038] 간단한 객체 레이블만이 제공되는 비디오를 이용한 약지도 학습은 객체의 영역 경계를 수작업으로 주석 처리할 필요가 없으므로, 대량의 객체 학습용 비디오를 저비용으로 빠르고 용이하게 제작할 수 있다.

[0039] 여기서 객체 검출부(113)는 각 경계 박스(B)에 검출해야 하는 객체가 존재할 확률을 나타내는 객체 스코어(h^*)를 함께 획득하고, 획득된 객체 스코어(h^*)가 기지정된 기준 객체 스코어 이상인 경우에만 정상 경계 박스(B)로 판별하여 출력할 수도 있다. 이는 객체 검출부(113)의 객체 검출 신뢰도를 향상시키기 위해서이다.

[0040] 객체 검출부(113)는 미리 학습된 패턴 추정 방식에 따라 다수의 프레임(f_t) 중 n번째 프레임(f_n)에서 검출되는 경계 박스(B_n)의 객체 특징($x_n \in \mathbb{R}^{D\times 1}$)을 추출하고, 추출된 객체 특징($x_n$)로부터 수학식 1에 따라 프레임별 객체 스코어(h^n)를 획득할 수 있다.

수학식 1

[0041]
$$h^n = \sum_d w_h(d) x^n(d)$$

[0042] (여기서 $w_h(d)$는 지정된 객체를 식별하기 위한 객체 분류자(w_h)에 대응하는 d번째 요소(element)로서 인공 신경망의 연산 레이어(예를 들면 컨볼루션 레이어)의 가중치를 나타낸다.)

[0043] 객체 검출부(113)는 다수의 프레임(f_t) 각각에 대해 획득되는 프레임별 객체 스코어(h^n)에 대해 평균값 풀링(average pooling)과 시그모이드(sigmoid) 함수를 적용하여, 수학식 2에 따라 다수의 프레임(f_t)의 경계 박스(B)에서 객체가 존재할 확률인 객체 스코어(h^*)를 획득할 수 있다.

수학식 2

[0044] $$h^* = \sum_n \text{sigmoid}(h^n)$$

[0045] 한편 객체 검출부(113)는 다수의 프레임에 다수의 객체가 포함된 경우, 다수 객체의 다양한 조합에 따른 영역을 검출할 수도 있다. 예를 들면, 비디오에 다수의 객체가 나타나며, 다수의 객체는 서로 이격되어 나타나거나 인접 또는 일부 영역에서 중첩되어 나타날 수도 있다. 이에 객체 검출부(113)는 서로 이격된 객체는 각각 구분된 객체 영역으로 검출하고, 인접하거나 일부 영역이 중첩된 객체는 각 객체별로 구분된 객체 영역으로 검출할 뿐만 아니라, 인접 또는 중첩된 객체가 함께 포함된 객체 영역 또한 검출할 수 있다. 여기서 객체 검출부(113)는 일예로 객체가 인접 또는 중첩 배치되어 각 객체에 대한 객체 영역의 적어도 일부가 중첩되는 경우에 객체가 함께 포함된 객체 영역을 추가로 검출하도록 구성될 수 있다.

[0046] 객체 검출부(113)는 T개의 프레임(f_1, ..., f_T) 각각에 대응하는 경계 박스(bounding box)($B = \{B_1, ..., B_T\}$)를 검출하며, 각 프레임(f_t)에 다수의 객체 영역이 탐지되는 경우, 각 객체 영역에 대응하는 개수(N)의 경계 영역(B_t^n, 여기서 $\{n = 1, ..., N\}$)를 검출할 수 있다.

[0047] 객체 검출부(113)에 의해 객체 영역이 검출되면, 튜블릿 생성부(114)는 다수의 프레임에서 동일 객체에 대해 검출된 객체 영역을 연결하여 객체 튜블릿을 생성한다. 즉 튜블릿은 그룹화된 다수 프레임에서 동일한 객체가 포함된 영역에 대한 경계 박스(B)들의 집합으로 획득될 수 있다.

[0048] 튜블릿 생성부(114)는 두개의 연속되는 프레임(f_{t-1}, f_t)에서 경계 박스(B_t^m, B_{t-1}^n)(여기서 $m, n \in \{1, ..., N\}$)가 획득되면, 경계 박스(B_t^m)와 경계 박스(B_{t-1}^n) 사이의 링크 스코어(E_{link})를 수학식 3에 따라 획득한다.

수학식 3

[0049] $$E_{\text{link}}(B_t^n, B_{t-1}^m) = h(B_t^n) + h(B_{t-1}^m) + \beta_1 E_{\text{feat}}(B_t^n, B_{t-1}^m) + \beta_2 E_{\text{IoU}}(B_t^n, B_{t-1}^m)$$

[0050] (여기서 $h(B_t^n)$는 경계 박스에 지정된 객체가 포함될 확률을 나타내는 객체 스코어이고, $E_{feat}(B_t^n, B_{t-1}^n)$는 L_2-norm 함수에 의해 경계 박스(B_t^n)와 경계 박스(B_{t-1}^n)의 정규화된 특징 사이의 유사성을 나타내고, $E_{IoU}(B_t^m, B_{t-1}^n)$는 경계 박스($B_t^m$)와 경계 박스($B_{t-1}^n$) 사이의 중첩 스코어로서 Union of IoU (Intersection of Union)를 측정한 결과를 나타내며, β_1, β_2는 각각 특징 유사도와 중첩 스코어에 대한 가중치를 제어하는 매개 변수이다.)

[0051] 링크 스코어(E_{link})는 연속되는 두 개의 프레임(f_{t-1}, f_t)에서 객체의 특징이 유사하고, 객체가 나타나는 영역이 중첩될수록 큰 값을 가져 강력하게 연결된다.

[0052] 그리고 프레임(f_t)에서 n 번째 객체에 대한 튜블릿을 생성하기 위해, 경로 수학식 4에 따른 인덱스($\pi_t(n)$)를 갖는 연결 경로를 구성하여, 수학식 5로 표현되는 객체 튜블릿(0^n)을 생성한다.

수학식 4

[0053]
$$\pi_1(n) = n,$$
$$\pi_t(n) = \underset{l}{\operatorname{argmax}} E_{\text{link}}(B_t^l, B_{t-1}^{\pi_t(n)})$$

[0054] (여기서, l ∈ {1, ..., N}이고, t ∈ {2, ..., T}이다.)

수학식 5

[0055]
$$\mathbf{O}^n = [B_1^{\pi_1(n)}, ..., B_T^{\pi_T(n)}]$$

[0056] 다만 객체 튜블릿을 획득하는 다양한 방식이 기존에 공개되어 있으므로, 경우에 따라서 객체 튜블릿 획득부(110)는 기존의 방식으로 미리 학습되어 객체 튜블릿(O^n)을 생성 할 수도 있다.

[0057] 튜블릿 조절부(120) 또한 패턴 추정 방식이 미리 학습된 인공 신경망을 포함하여 객체 튜블릿 획득부(110)에서 획득된 객체 튜블릿(O^n)을 인가받고, 인가된 객체 튜블릿(O^n) 각각에서 경계 박스(B)들의 크기를 조절한다. 즉 객체 튜블릿(O^n) 각각의 크기를 조절하여 튜블릿(P_n)을 획득한다.

[0058] 상기한 바와 같이, 객체 튜블릿 획득부(110)가 약지도 학습되는 경우, 학습용 비디오를 매우 용이하게 획득할 수 있으나, 경계 박스(B)의 검출 성능은 완전 지도 학습 방식보다 낮아질 수 있다. 즉 경계 박스(B)가 객체가 나타나는 객체 영역에 정확하게 대응하지 않고, 불필요한 영역을 포함하여 추출될 수 있다. 이에 튜블릿 조절부(120)는 경계 박스(B)가 정확하게 객체 영역만을 지정하도록 객체 튜블릿의 경계 박스(B)에서 이러한 불필요한 영역을 제거하도록 한다.

[0059] 튜블릿 조절부(120)는 t번째 프레임(f_t)의 n번째 경계 박스(B_t^n)의 중심 위치를 기준으로 대해 폭(u_t^n)과 높이(v_t^n)에 대한 오프셋을 줄여 경계 박스(B)의 크기를 조절한다.

[0060] 구체적으로 튜블릿 조절부(120)는 미리 학습된 패턴 추정 방식에 따라 경계 박스(B_t^n)의 폭(u_t^n)에 대한 조절 폭(∇u_t^n)과 높이(v_t^n)에 대한 조절 높이(∇v_t^n)를 획득하고, 획득된 조절 폭(∇u_t^n)과 조절 높이(∇v_t^n)에 따라 수학식 6과 같이 크기가 조절된 최적 경계 박스($\bar{B}_t^n$)를 획득한다.

수학식 6

[0061]
$$\begin{bmatrix} \bar{u}_t^n \\ \bar{v}_t^n \end{bmatrix} = \begin{bmatrix} \nabla u_t^n & 0 \\ 0 & \nabla v_t^n \end{bmatrix} \begin{bmatrix} u_t^n \\ v_t^n \end{bmatrix}$$

[0062] 튜블릿 조절부(120)는 다수의 컨볼루션 레이어와 적어도 하나의 활성화 함수 레이어(Activation function layer)(여기서는 일예로 ReLU)로 구성된 컨볼루션 신경망을 포함하여, 조절 폭(∇u_t^n)과 조절 높이(∇v_t^n)를 획득할 수 있다.

[0063] 도 3은 약지도 학습을 위한 액션 학습용 비디오의 일예를 나타내고, 도 4은 도 1의 튜블릿 조절부에서 크기가 조절된 경계 박스의 일예를 나타낸다.

[0064] 튜블릿 조절부(120)는 객체 튜블릿 획득부(110)가 객체 학습용 비디오에 의해 약지도 학습된 이후, 약지도 학습된 객체 튜블릿 획득부(110)가 액션 학습용 비디오에서 획득한 객체 튜블릿을 인가받아 추가적으로 약지도 학습될 수 있다. 여기서 액션 학습용 비디오는 단순히 액션 레이블이 주석된 비디오로서, 일예로 도 3에 도시된 바와 같이, 다이빙, 골프, 아이스 댄싱, 펜싱 등의 액션 레이블이 주석된 단일 액션이 포함된 비디오일 수 있다.

[0065] 도 4에서는 연속되는 다수의 프레임에서 객체 튜블릿 획득부(110)가 검출한 경계 박스(B_t^n)와 튜블릿 조절부(120)에서 조절된 최적 경계 박스($\bar{B}_t^n$)를 나타내고 있다. 도 4에 도시된 바와 같이, 경계 박스(B_t^n)는 객체가 나타나는 영역에 대해 상대적으로 큰 영역으로 검출되어 여백이 포함되는 반면, 최적 경계 박스($\bar{B}_t^n$)는 객체의 영역에 매우 타이트하게 설정되었음을 알 수 있다.

[0066] 차원 변환부(130)는 튜블릿(P_n) 각각의 다수의 최적 경계 박스($\bar{B}_t^n$)들에 대해 시간축을 기준으로 시간 평균 풀링(time average pooling)을 수행하여, 다수의 최적 경계 박스($\bar{B}_t^n$)를 포함하는 3차원의 튜블릿(P_n) 각각을 2차원의 튜블릿 이미지로 변환한다.

[0067] 특징 추출부(140)는 튜블릿 이미지를 인가받고, 미리 학습된 패턴 추정 방식에 따라 튜블릿 이미지의 특징을 추출하여 특징맵($y^n \in \mathbb{R}^D$)을 획득한다.

[0068] 액션 가중치 획득부(150)는 미리 학습된 패턴 추정 방식에 따라 특징 추출부(140)에서 획득된 특징맵에서 액션 가중치(α^n)를 획득하고, 획득된 액션 가중치를 대응하는 특징맵(y^n)에 적용하여 가중 특징맵($\alpha^n y^n$)을 획득한다.

[0069] 여기서 액션 가중치(α^n)는 튜블릿 이미지에서 객체의 액션 수준, 즉 움직임을 나타내는 가중치이다. 액션 가중치 획득부(150)가 액션 가중치(α^n)를 획득하여 특징맵(y^n)에 가중하는 것은, 비록 객체 튜블릿 획득부(110)가 객체를 탐지하여 객체 튜블릿(0^n)을 획득하더라도, 객체 튜블릿(0^n)의 객체에 움직임이 없다면 객체의 액션 영역을 검출하는 액션 로컬라이제이션에서는 무의미하기 때문이다.

[0070] 차원 변환부(130)와 특징 추출부(140)는 특징맵 획득부로 통합될 수 있다.

[0071] 액션 인식 및 영역 판별부(160)는 가중 특징맵($\alpha^n y^n$)을 인가받고, 미리 학습된 패턴 추정 방식에 따라 기지정된 다수의 액션 클래스 중 적어도 하나의 액션 클래스로 분류한다. 액션 인식 및 영역 판별부(160) 또한 인공 신경망으로 구현될 수 있다.

[0072] 액션 인식 및 영역 판별부(160)는 우선 튜블릿(P_n) 각각에 대응하는 가중 특징맵($\alpha^n y^n$)이 기지정된 다수의 액션 클래스 각각에 대응하는 수준을 나타내는 액션 클래스 스코어($\lambda^n(c) = \{\lambda^n(1), \ldots, \lambda^n(C)\}$)를 수학식 7에 따라 획득한다.

수학식 7

$$\lambda^n(c) = \sum_d w_T(c,d)\,\alpha^n y^n(d) = \alpha^n \sum_d w_T(c,d) y^n(d)$$

[0073]

[0074] (여기서 $w_T(c, d)$는 지정된 액션 클래스($c \in \{1, \ldots, C\}$)를 식별하기 위한 액션 클래스 분류자($w_T \in \mathbb{R}^{D \times C}$)에 대응하는 d번째 요소로서 인공 신경망의 연산 레이어(예를 들면 컨볼루션 레이어)의 가중치를 나타낸다.)

[0075] 수학식 7에서 $\sum_{d} w_T(c,d)y^n(d)$ 는 n번째 튜블릿(P_n)의 클래스(c)에 대한 연관성을 나타내는 분류 스코어로서 수학식 8와 같이 표현될 수 있다.

수학식 8

[0076] $$s^n(c) = \sum_{d} w_T(c,d)y^n(d)$$

[0077] (여기서 $s^n = [s^n(1), \ldots, s^n(C)]^T \in \mathbb{R}^{C\times 1}$ 이다.)

[0078] 수학식 8에 의해 수학식 7는 수학식 9으로 표현될 수 있다.

수학식 9

[0079] $$\lambda^n(c) = \alpha^n s^n(c)$$

[0080] 즉 n번째 튜블릿(P_n)의 클래스(c)에 대한 액션 클래스 스코어($\lambda^n(c)$)는 수학식 9과 같이, 액션 가중치(α^n)와 분류 스코어(S^n)로 획득된다.

[0081] 액션 인식 및 영역 판별부(160)는 튜블릿(P_n) 각각의 다수의 액션 클래스(c)에 대한 액션 클래스 스코어($\lambda^n(c)$)가 획득되면, 기지정된 기준 액션 클래스 스코어 이상인 액션 클래스 스코어($\lambda^n(c)$)를 선택하고, 선택된 액션 클래스 스코어($\lambda^n(c)$)에 대응하는 튜블릿(P_n)을 액션 튜블릿으로 추출한다. 그리고 추출된 액션 튜블릿의 최적 경계 박스($\bar{B}_t^n$)와 액션 클래스(c)를 획득하여 액션 로컬라이제이션의 결과로 출력한다. 즉 액션의 종류와 함께 비디오에서 액션이 나타난 객체 영역을 출력한다.

[0082] 이때 하나의 튜블릿(P_n)이 다수의 액션 클래스(c)에 대해서 액션 클래스 스코어($\lambda^n(c)$)가 기준 액션 클래스 스코어 이상으로 나타날 수 있다. 즉 하나의 튜블릿(P_n)이 다수의 액션 클래스에 대응하는 경우가 발생할 수 있다. 이 경우, 액션 인식 및 영역 판별부(160)는 기지정된 설정에 따라 액션 클래스 스코어($\lambda^n(c)$)가 가장 높은 하나의 액션 클래스(c)를 출력하거나, 기준 액션 클래스 스코어 이상으로 나타난 다수의 액션 클래스(c) 모두를 출력할 수 있다.

[0083] 학습부(170)는 액션 인식 및 액션 영역 탐지 장치를 약지도 학습시키기 위한 구성으로 학습 수행 시에만 추가되고, 학습된 이후에는 생략될 수 있다.

[0084] 학습부(170)는 객체 학습용 비디오를 이용하여 객체 튜블릿 획득부(110)를 우선 약지도 학습시키고, 이후, 약지도 학습된 객체 튜블릿 획득부(110)와 액션 학습용 비디오를 이용하여 튜블릿 조절부(120), 차원 변환부(130), 특징 추출부(140), 액션 가중치 획득부(150) 및 액션 인식 및 영역 판별부(160)를 약지도 학습시킬 수 있다.

[0085] 학습부(170)는 객체 학습용 비디오가 객체 튜블릿 획득부(110)에 인가되어 획득되는 경계 박스(B)에 검출해야 하는 객체가 존재할 확률을 나타내는 객체 스코어(h^*)를 전달받는다. 그리고 객체 스코어(h^*)와 객체 학습용 비디오에 주석된 객체 레이블 사이의 차이를 객체 손실로 획득하여 객체 튜블릿 획득부(110)로 역전파하여 객체 튜블릿 획득부(110)를 약지도 학습시킨다. 이때, 학습부(170)는 객체 손실을 일예로 표준 다중 레이블 교차 엔트로피 손실과 같은 공지된 함수에 적용하여 획득할 수 있다.

[0086] 객체 튜블릿 획득부(110)가 약지도 학습되면, 학습부(170)는 액션 학습용 비디오를 객체 튜블릿 획득부(110)에

인가하고, 액션 가중치 획득부(150)에서 모든 액션 튜블릿(P_n)에 대해 출력되는 가중 특징맵($\alpha^n y^n$)을 수학식 10과 같이 모두 더하여 비디오 레벨에서 액션 튜블릿에 대한 특징을 나타내는 비디오 특징맵(y^*)을 획득한다.

수학식 10

$$y^* = \sum_n \alpha^n y^n$$

[0087]

[0088] 그리고 학습부(170)는 비디오 특징맵(y^*)으로부터 비디오 액션 클래스 스코어($\lambda(c)$)를 수학식 7와 유사하게 수학식 11에 따라 획득한다.

수학식 11

$$\begin{aligned}\lambda(c) &= \sum_d w_T(c,d) y^*(d), \\ &= \sum_d w_T(c,d) \sum_n \alpha^n y^n(d) \\ &= \sum_n \alpha^n \sum_d w_T(c,d) y^n(d) \\ &= \sum_n \alpha^n s^n(c)\end{aligned}$$

[0089]

[0090] 비디오 액션 클래스 스코어($\lambda(c)$)가 획득되면, 학습부(170)는 비디오 액션 클래스 스코어($\lambda(c)$)와 액션 학습용 비디오의 액션 레이블과의 차이를 액션 손실로 획득하여 역전파함으로써, 튜블릿 조절부(120), 차원 변환부(130), 특징 추출부(140), 액션 가중치 획득부(150) 및 액션 인식 및 영역 판별부(160)를 약지도 학습시킬 수 있다. 여기서 학습부(170)는 액션 손실을 일예로 표준 다중 레이블 교차 엔트로피 손실과 같은 공지된 함수에 적용하여 획득할 수 있다.

[0091] 도 5 및 도 6은 본 실시예에 따른 비디오 액션 인식 및 액션 영역 탐지 장치에서 액션 로컬라이제이션이 수행된 결과의 일예를 나타낸다.

[0092] 도 5에서 (a)와 (b)는 각각 농구와 아이스 댄싱에 대해 액션 로컬라이제이션을 수행한 결과를 나타내고, 도6 에서 (a) 내지 (d)는 각각 다이빙, 축구, 농구 및 사이클에 대해 액션 로컬라이제이션을 수행한 결과를 나타낸다. 그리고 도 5 및 도 6에서는 본 실시예에 따른 비디오 액션 인식 및 액션 영역 탐지 장치의 성능을 비교하기 위해 기존에 수작업 등으로 수행된 검증 자료 레이블(ground truth label)을 함께 표시하였다.

[0093] 도 5 및 도 6에 도시된 바와 같이, 본 실시예에 따른 비디오 액션 인식 및 액션 영역 탐지 장치는 약지도 학습 방식으로 학습이 수행됨에도 객체의 액션이 발생된 영역을 정확하게 추출할 수 있음을 확인할 수 있다.

[0094] 도 7은 본 발명의 일 실시예에 따른 비디오 액션 인식 및 액션 영역 탐지 방법을 나타낸다.

[0095] 도 1 내지 도 6을 참조하여, 도 7의 비디오 액션 인식 및 액션 영역 탐지 방법을 설명하면, 우선 학습부(170)는 객체 레이블이 주석된 객체 학습용 비디오를 이용하여 객체 튜블릿 획득부(110)를 약지도 학습시킨다(S12). 학습부(170)는 객체 튜블릿 획득부(110)가 객체 학습용 비디오에 응답하여 출력하는 객체 스코어(h^*)와 객체 레이블 사이의 차이를 객체 손실로 획득하여 객체 튜블릿 획득부(110)로 역전파함으로써, 객체 튜블릿 획득부(110)를 학습시킬 수 있다.

[0096] 이후 학습부(170)는 학습된 객체 튜블릿 획득부(110)와 객체의 액션 레이블이 주석된 액션 학습용 비디오를 이용하여 튜블릿 조절부(120), 차원 변환부(130), 특징 추출부(140), 액션 가중치 획득부(150) 및 액션 인식 및

영역 판별부(160)를 약지도 학습시킨다(S12).

[0097] 학습부(170)는 액션 가중치 획득부(150)에서 모든 액션 튜블릿(P_n)에 대해 출력되는 가중 특징맵($\alpha^n y^n$)으로부터 수학식 10에 따라 비디오 특징맵(y^*)을 획득하고, 비디오 특징맵(y^*)으로부터 비디오 액션 클래스 스코어($\lambda(c)$)를 획득한다. 그리고 획득된 비디오 액션 클래스 스코어($\lambda(c)$)와 액션 학습용 비디오의 액션 레이블과의 차이를 액션 손실로 획득하여 역전파하여 약지도 학습을 수행할 수 있다.

[0098] 학습이 수행된 이후, 비디오 액션 인식 및 액션 영역 탐지 장치는 액션 로컬라이제이션이 수행되어야 하는 비디오를 인가받고, 패턴 추정 방식이 약지도 학습된 객체 튜블릿 획득부(110)는 비디오에서 기지정된 객체가 포함된 영역인 경계 박스(B)를 검출하여 객체 튜블릿(O^n)을 획득한다(S21). 이때 비디오에 포함된 객체의 수에 따라 획득되는 객체 튜블릿(O^n)의 개수는 가변될 수 있다.

[0099] 튜블릿 조절부(120)는 획득된 객체 튜블릿(O^n)의 경계 박스(B_t^n) 각각에 대해 약지도 학습된 패턴 추정 방식에 따라 수학식 6과 같이 객체 튜블릿(O^n)의 경계 박스(B)의 크기를 조절하여, 최적 경계 박스($\bar{B}_t^n$)를 갖는 튜블릿(P_n)을 획득한다.

[0100] 튜블릿(P_n)이 획득되면, 차원 변환부(130)가 튜블릿(P_n) 각각의 다수의 최적 경계 박스($\bar{B}_t^n$)들에 대해 시간축을 기준으로 시간 평균 풀링을 수행하여, 튜블릿 이미지로 변환한다 (S23).

[0101] 그리고 미리 학습된 패턴 추정 방식에 따라 특징 추출부(140)가 튜블릿 이미지의 특징을 추출하여 특징맵(y^n)을 획득하고, 액션 가중치 획득부(150)가 특징맵에서 액션 가중치(α^n)를 획득하여 대응하는 특징맵(y^n)에 적용함으로써 가중 특징맵($\alpha^n y^n$)을 획득한다(S24).

[0102] 액션 인식 및 영역 판별부(160)는 튜블릿(P_n) 각각에 대응하는 가중 특징맵($\alpha^n y^n$)이 기지정된 다수의 액션 클래스 각각에 대응하는 수준을 나타내는 액션 클래스 스코어($\lambda^n(c) = \{\lambda^n(1), \ldots, \lambda^n(C)\}$)를 수학식 9와 같이 획득한다(S25).

[0103] 그리고 획득된 액션 클래스 스코어($\lambda^n(c)$) 중 기지정된 기준 액션 클래스 스코어 이상인 액션 클래스 스코어($\lambda^n(c)$)를 선택하고, 선택된 액션 클래스 스코어($\lambda^n(c)$)에 대응하는 튜블릿(P_n)을 액션 튜블릿으로 추출한다. 이와 함께 추출된 액션 튜블릿의 최적 경계 박스($\bar{B}_t^n$)와 액션 클래스(c)를 획득하여 출력한다(S26).

[0104] 본 발명에 따른 방법은 컴퓨터에서 실행 시키기 위한 매체에 저장된 컴퓨터 프로그램으로 구현될 수 있다. 여기서 컴퓨터 판독가능 매체는 컴퓨터에 의해 액세스 될 수 있는 임의의 가용 매체일 수 있고, 또한 컴퓨터 저장 매체를 모두 포함할 수 있다. 컴퓨터 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터와 같은 정보의 저장을 위한 임의의 방법 또는 기술로 구현된 휘발성 및 비휘발성, 분리형 및 비분리형 매체를 모두 포함하며, ROM(판독 전용 메모리), RAM(랜덤 액세스 메모리), CD(컴팩트 디스크)-ROM, DVD(디지털 비디오 디스크)-ROM, 자기 테이프, 플로피 디스크, 광데이터 저장장치 등을 포함할 수 있다.

[0105] 본 발명은 도면에 도시된 실시예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다.

[0106] 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 청구범위의 기술적 사상에 의해 정해져야 할 것이다.

부호의 설명

[0107] 110: 객체 튜블릿 획득부 120: 튜블릿 회귀부

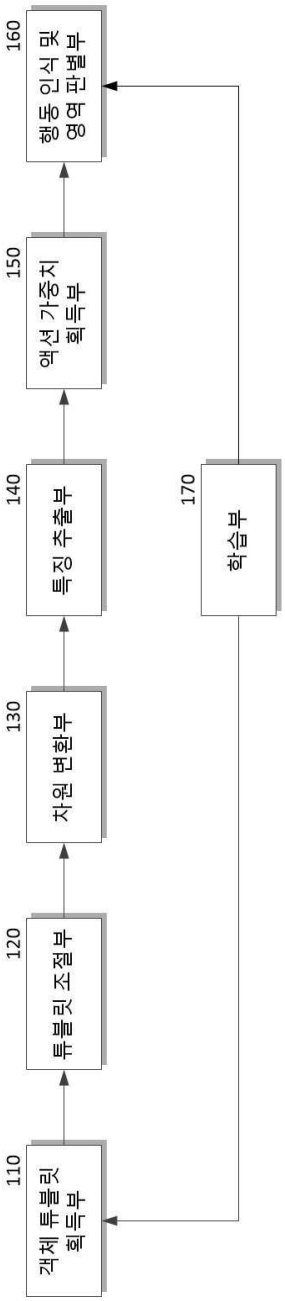
130: 차원 변환부 140: 특징 추출부

150: 액션 가중치 획득부 160: 액션 인식 및 영역 판별부

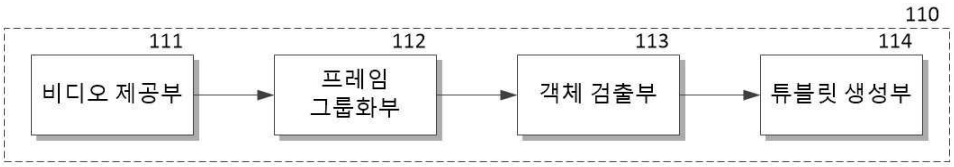
170: 학습부

도면

도면1

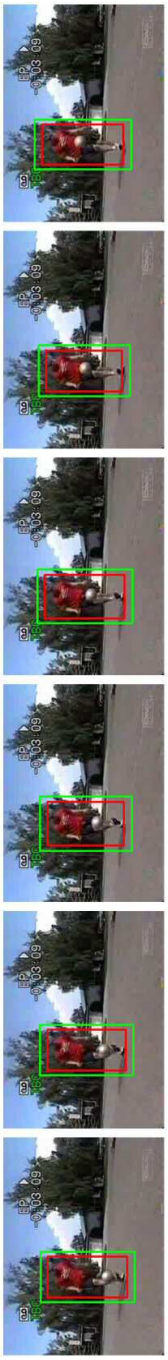


도면2

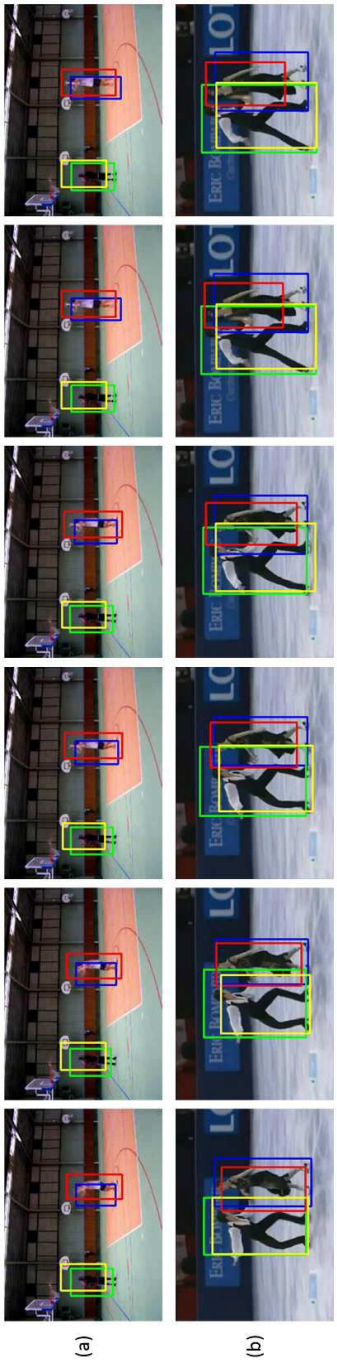


도면3

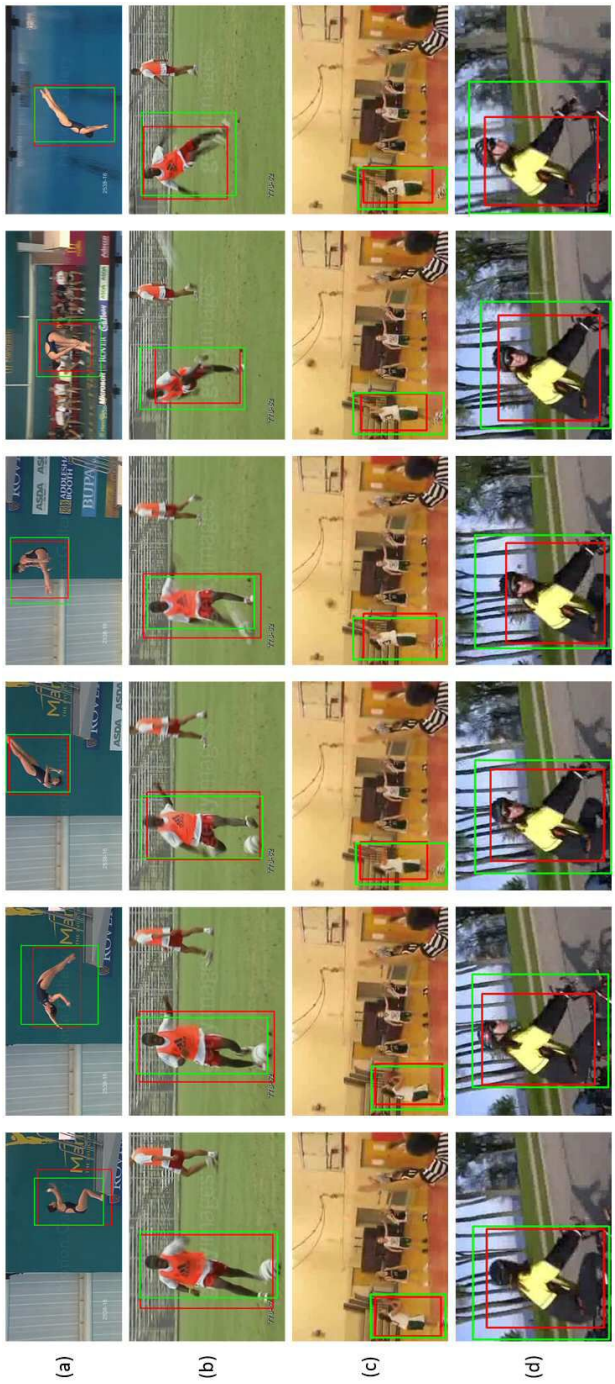
도면4



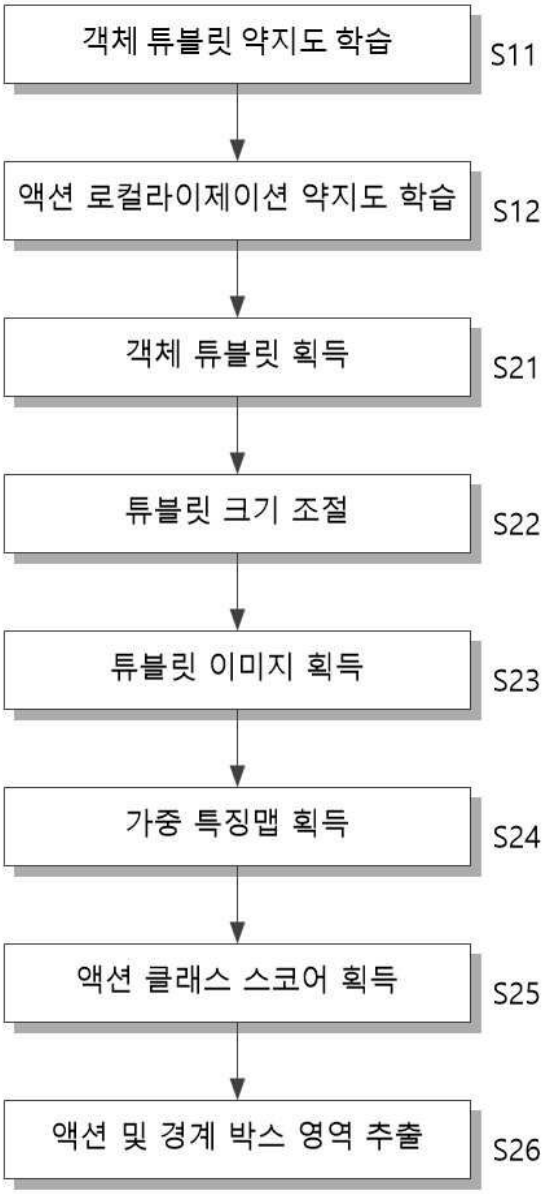
도면5



도면6



도면7


객체 튜블릿 약지도 학습
S11
액션 로컬라이제이션 약지도 학습
S12
객체 튜블릿 획득
S21
튜블릿 크기 조절
S22
튜블릿 이미지 획득
S23
가중 특징맵 획득
S24
액션 클래스 스코어 획득
S25
액션 및 경계 박스 영역 추출
S26