(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0088179
(43) 공개일자 2025년06월17일

(51) 국제특허분류(Int. Cl.)
G06V 20/40 (2022.01) *G06N 3/08* (2023.01)
G06V 10/74 (2022.01) *G06V 10/82* (2022.01)
(52) CPC특허분류
G06V 20/46 (2022.01)
G06N 3/08 (2023.01)
(21) 출원번호 10-2023-0177999
(22) 출원일자 2023년12월08일
심사청구일자 2023년12월08일

(71) 출원인
세종대학교산학협력단
서울특별시 광진구 능동로 209 (군자동, 세종대학교)
(72) 발명자
최유경
서울특별시 광진구 능동로 209, 세종대학교 대양AI센터 601호(군자동)
조원
서울특별시 광진구 광나루로14길 26, 503호(화양동, 아젠시아)
(뒷면에 계속)
(74) 대리인
송인호, 최관락

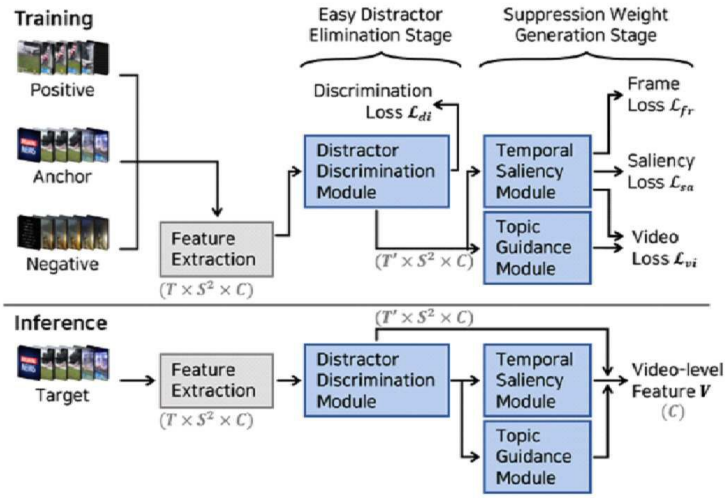
전체 청구항 수 : 총 13 항

(54) 발명의 명칭 **방해자 프레임 억제를 적용한 비디오 특징 추출 방법 및 그 장치**

(57) 요 약

방해자 프레임 억제를 적용한 비디오 특징 추출 방법 및 그 장치가 개시된다. 방해자 프레임 억제를 적용한 비디오 특징 추출 방법은 (a) 입력 비디오의 각 프레임에 대한 프레임 수준 특징 벡터를 추출하는 단계; (b) 상기 프레임 수준 특징 벡터를 학습된 방해자 구별 모델에 적용하여 방해자 프레임을 선별하여 제거하는 단계; (c) 상기 방해자 프레임이 제거된 프레임 수준 특징 벡터들을 학습된 시간적 중요도 모델에 적용하여 시간적 중요도 가중치를 계산하고, 상기 시간적 중요도 가중치를 이용하여 상기 프레임 수준 특징 벡터들에 대한 방해자 프레임 억제 가중치를 생성하는 단계; 및 (d) 상기 방해자 프레임 억제 가중치와 프레임 수준 특징 벡터들을 집계하여 비디오 수준 특징 벡터를 생성하는 단계를 포함한다.

대 표 도 - 도3



(52) CPC특허분류
G06V 10/761 (2023.08)
G06V 10/82 (2022.01)

(72) 발명자
임근택
서울특별시 광진구 아차산로21길 86 (화양동)
이광진
소서울특별시 광진구 천호대로137길 9, A동 502호
(구의동, 한울타워)
김현우
서울특별시 송파구 올림픽로43길 34, 105동 1001호(풍납동, 잠실 올림픽공원 아이파크)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711195810
과제번호	00231894
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	ICTR&D혁신바우처지원사업(예타)
연구과제명	비대면 콘텐츠 제작을 위한 AI 기반의 비디오 검색 시스템 개발
기 여 율	50/100
과제수행기관명	피아스페이스(주)
연구기간	2023.04.01 ~ 2023.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193731
과제번호	2022-0-00331-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	스마트엣지디바이스기술개발(R&D)
연구과제명	멘탈헬스케어를 위한 감성인지/생성 기반의 교감형 AI 엣지 디바이스 기술 개발
기 여 율	25/100
과제수행기관명	한국과학기술연구원(KIST)
연구기간	2023.01.01 ~ 2023.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711198096
과제번호	00254529
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성
연구과제명	메타버스 융합대학원(세종대학교)
기 여 율	25/100
과제수행기관명	세종대학교산학협력단
연구기간	2023.07.01 ~ 2023.12.31

명 세 서

청구범위

청구항 1

(a) 입력 비디오의 각 프레임에 대한 프레임 수준 특징 벡터를 추출하는 단계;

(b) 상기 프레임 수준 특징 벡터를 학습된 방해자 구별 모델에 적용하여 방해자 프레임을 선별하여 제거하는 단계;

(c) 상기 방해자 프레임이 제거된 프레임 수준 특징 벡터들을 학습된 시간적 중요도 모델에 적용하여 시간적 중요도 가중치를 계산하고, 상기 시간적 중요도 가중치를 이용하여 상기 프레임 수준 특징 벡터들에 대한 방해자 프레임 억제 가중치를 생성하는 단계; 및

(d) 상기 방해자 프레임 억제 가중치와 프레임 수준 특징 벡터들을 집계하여 비디오 수준 특징 벡터를 생성하는 단계를 포함하는 방해자 프레임 억제를 적용한 비디오 특징 추출 방법.

청구항 2

제1 항에 있어서,

상기 (b) 단계는,

상기 프레임 수준 특징 벡터의 크기가 제1 임계치 미만인 프레임을 방해자 프레임으로 선별하여 제거하는 것을 특징으로 하는 방해자 프레임 억제를 적용한 비디오 특징 추출 방법.

청구항 3

제1 항에 있어서,

상기 (c) 단계는,

상기 방해자 프레임이 제거된 프레임 수준 특징 벡터들을 학습된 시간적 중요도 모델에 적용하여 자기 유사도 맵을 생성하고, 상기 자기 유사도 맵을 인공 신경망에 적용한 후 대각 성분만 추출하여 시간적 중요도 가중치를 생성하는 것을 특징으로 하는 방해자 프레임 억제를 적용한 비디오 특징 추출 방법.

청구항 4

제1 항에 있어서,

상기 (a) 단계 이전에, 상기 방해자 구별 모델이 학습되되,

상기 방해자 구별 모델의 학습 과정에서, 상기 프레임 수준 특징 벡터의 크기를 각각 산출한 후 임계치보다 낮은 크기값을 가지는 프레임들을 방해자 프레임으로 선별하여 방해자 프레임 임시 라벨을 부여하고, 상기 방해자 프레임 임시 라벨을 이용하여 상기 입력 비디오(X)에 대해 일정 비율만큼 방해자 프레임 수준 특징 벡터를 추가하는 것을 특징으로 하는 방해자 프레임 억제를 적용한 비디오 특징 추출 방법.

청구항 5

제4 항에 있어서,

상기 학습 과정에서 상기 입력 비디오(X)는 앵커, 포지티브 및 네거티브 샘플로 입력되는 것을 특징으로 하는

방해자 프레임 억제를 적용한 비디오 특징 추출 방법.

청구항 6

제4 항에 있어서,

상기 (a) 단계 이전에, 시간적 중요도 모델의 학습 과정이 선행되되,

상기 학습 과정에서 상기 시간적 중요도 모델은 상기 앵커와 상기 포지티브 쌍의 프레임 수준 특징 벡터들에 대한 제1 프레임 수준 유사도 맵을 산출하고, 포지티브 프레임 수준 특징 벡터의 시간 차원에 대해 맥스 풀링을 통해 제1 시간적 중요도 가중치를 생성하고, 상기 제1 시간적 중요도 가중치를 이용하여 앵커 샘플들에 대한 제1 시간적 중요도 임시 라벨을 부여하도록 학습되고,

상기 앵커와 상기 네거티브 쌍의 프레임 수준 특징 벡터들에 대한 제2 프레임 수준 유사도 맵을 산출하고, 네거티브 프레임 수준 특징 벡터의 시간 차원에 대해 맥스 풀링을 통해 제2 시간적 중요도 가중치를 생성하고, 상기 제2 시간적 중요도 가중치를 이용하여 앵커 샘플들에 대한 제2 시간적 중요도 임시 라벨을 부여하도록 학습되며,

상기 앵커 프레임 수준 특징 벡터들에 대한 자기-유사도 맵을 산출하고 병목 레이어와 트랜스포머 인코더를 적용한 후 대각 성분만 추출하여 최종 시간적 중요도 가중치를 산출하도록 학습되는 것을 특징으로 하는 방해자 프레임 억제를 적용한 비디오 특징 추출 방법.

청구항 7

제1 항 내지 제6 항 중 어느 하나의 항에 따른 방법을 수행하기 위한 프로그램 코드를 기록한 컴퓨터로 판독 가능한 기록 매체.

청구항 8

입력 비디오의 각 프레임에 대한 프레임 수준 특징 벡터를 추출하는 특징 추출부;

상기 프레임 수준 특징 벡터를 학습된 방해자 구별 모델에 적용하여 방해자 프레임을 선별하여 제거부;

상기 방해자 프레임이 제거된 프레임 수준 특징 벡터들을 학습된 시간적 중요도 모델에 적용하여 시간적 중요도 가중치를 계산하고, 상기 시간적 중요도 가중치를 이용하여 상기 프레임 수준 특징 벡터들에 대한 방해자 프레임 억제 가중치를 생성하는 억제 가중치 생성부; 및

상기 방해자 프레임 억제 가중치와 프레임 수준 특징 벡터들을 집계하여 비디오 수준 특징 벡터를 생성하는 비디오 수준 특징 생성부를 포함하는 비디오 특징 추출 장치.

청구항 9

제8 항에 있어서,

상기 제거부는,

상기 프레임 수준 특징 벡터의 크기가 제1 임계치 미만인 프레임을 방해자 프레임으로 선별하여 제거하는 것을 특징으로 하는 비디오 특징 추출 장치.

청구항 10

제8 항에 있어서,

상기 억제 가중치 생성부는,

상기 방해자 프레임이 제거된 프레임 수준 특징 벡터들을 학습된 시간적 중요도 모델에 적용하여 자기 유사도 맵을 생성하고, 상기 자기 유사도 맵을 인공 신경망에 적용한 후 대각 성분만 추출하여 시간적 중요도 가중치를 생성하는 것을 특징으로 하는 비디오 특징 추출 장치.

청구항 11

제8 항에 있어서,

상기 제거부는,

상기 프레임 수준 특징 벡터의 크기를 각각 산출한 후 임계치보다 낮은 크기값을 가지는 프레임들을 방해자 프레임으로 선별하여 방해자 프레임 임시 라벨을 부여하고, 상기 방해자 프레임 임시 라벨을 이용하여 상기 입력 비디오(X)에 대해 일정 비율만큼 방해자 프레임 수준 특징 벡터를 추가하여 상기 방해자 구별 모델을 사전 학습시키는 것을 특징으로 하는 비디오 특징 추출 장치.

청구항 12

제11 항에 있어서,

상기 학습 과정에서 상기 입력 비디오(X)는 앵커, 포지티브 및 네거티브 샘플로 입력되는 것을 특징으로 하는 비디오 특징 추출 장치.

청구항 13

제12 항에 있어서,

상기 억제 가중치 생성부는, 시간적 중요도 모델을 사전 학습시키되,

상기 학습 과정에서 상기 시간적 중요도 모델은 상기 앵커와 상기 포지티브 쌍의 프레임 수준 특징 벡터들에 대한 제1 프레임 수준 유사도 맵을 산출하고, 포지티브 프레임 수준 특징 벡터의 시간 차원에 대해 맥스 풀링을 통해 제1 시간적 중요도 가중치를 생성하고, 상기 제1 시간적 중요도 가중치를 이용하여 앵커 샘플들에 대한 제1 시간적 중요도 임시 라벨을 부여하도록 학습되고,

상기 앵커와 상기 네거티브 쌍의 프레임 수준 특징 벡터들에 대한 제2 프레임 수준 유사도 맵을 산출하고, 네거티브 프레임 수준 특징 벡터의 시간 차원에 대해 맥스 풀링을 통해 제2 시간적 중요도 가중치를 생성하고, 상기 제2 시간적 중요도 가중치를 이용하여 앵커 샘플들에 대한 제2 시간적 중요도 임시 라벨을 부여하도록 학습되며,

상기 앵커 프레임 수준 특징 벡터들에 대한 자기-유사도 맵을 산출하고 병목 레이어와 트랜스포머 인코더를 적용한 후 대각 성분만 추출하여 최종 시간적 중요도 가중치를 산출하도록 학습되는 것을 특징으로 하는 비디오 특징 추출 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 방해자 프레임 억제를 적용한 비디오 특징 추출 방법 및 그 장치에 관한 것이다.

배 경 기 술

[0003] 비디오 입력 기반 유사 영상 검색(Content-based Video Retrieval)은 정보 인출 문제의 일종으로, 특정한 비디오간 유사도 척도에 근거하여 입력된 비디오와의 유사도가 높은 비디오들을 데이터베이스에서 인출하는 것을 목

표로 한다. 비디오 간 유사도는 영상의 시각적 특징을 요약하는 특징 벡터의 유사도를 산출함으로써 측정되며, 우수한 비디오 검색 성능을 얻기 위해서는 표현력이 높으면서도 효율적인 비교가 가능하도록 특징 벡터를 설계하는 것이 바람직하다.

[0004] 비디오에 대한 특징 추출 연구의 흐름은 크게 두 가지로 볼 수 있는데, 한 가지 갈래는 비디오의 특징을 개별 프레임들에 대한 이미지 특징 벡터의 모음으로 표현하는 접근 방식을 취하고 있으며 다른 한 갈래는 비디오 전체를 단일한 특징 벡터로 표현하는 접근 방식을 취하고 있다. 이하에서 전자는 '프레임 수준 특징 벡터' 방식으로, 후자는 '비디오 수준 특징 벡터' 방식으로 지칭하도록 한다. 두 방법의 차이는 두 비디오의 유사도를 비교하는 방법에서 명확하게 드러나는데, 프레임 수준 특징 벡터 방식은 두 비디오의 프레임들에 대해 각각 유사도를 비교한 뒤에 산출된 유사도를 종합하여 최종적인 영상 유사도 점수를 산출 하는데 반해, 비디오 수준 특징 벡터 방식은 별도의 점수 종합 과정이 없이 두 비디오의 유사도 점수를 바로 계산할 수 있다.

[0005] 프레임 수준 특징 벡터 방식은 비디오 간 유사도를 산출하는 과정에서 비디오 수준 특징 벡터 방식보다 더 많은 정보를 활용하기 때문에 높은 검색 정확도를 보이지만 연산량과 처리 속도에 한계가 있다. 이에 반해 비디오 수준 특징 벡터 방식은 검색 속도 및 메모리 측면에서 이득이 있지만 프레임 수준 방식에 비해 낮은 검색 정확도를 보여준다. 정보 검색 (information retrieval) 문제는 검색 속도 및 메모리 측면도 중요하기 때문에 비디오 수준 특징 벡터의 방식에서 검색 정확도를 높이는 방향으로 개선이 바람직하다.

[0006] 하지만 비디오 수준 특징 벡터 방식은 비디오를 구성하는 다수의 프레임 정보를 하나의 단일한 특징 벡터로 표현하는 과정에서 발생하는 정보의 손실과 비디오 인식에 방해가 되는 distractor 프레임들의 영향으로 좋은 구별력을 가지는 특징 벡터를 얻는데 어려움을 겪고 있다.

[0007] 도 1은 앞서 설명한 distractor 프레임의 예시를 보여준다. 도 1에서 빨간 박스로 표시된 프레임들은 비디오의 주제와 관련성을 가지지 못하는 정보로써 비디오 수준 특징 벡터를 집적하는 과정에서 특징 벡터의 표현력을 저해시키는 원인으로 작용한다.

발명의 내용

해결하려는 과제

[0009] 본 발명은 방해자 프레임 억제를 적용한 비디오 특징 추출 방법 및 그 장치를 제공하기 위한 것이다.

[0010] 또한, 본 발명은 시각적 중정보가 부족한 프레임들을 제거하고, 중요도가 높은 특징 벡터들을 비디오 수준 특징 벡터로 집적하여 프레임의 시각적/의미론적 정보에 기반하여 비디오 수준 서술자의 표현력을 개선할 수 있는 방해자 프레임 억제를 적용한 비디오 특징 추출 방법 및 그 장치를 제공하기 위한 것이다.

과제의 해결 수단

[0012] 본 발명의 일 측면에 따르면, 방해자 프레임 억제를 적용한 비디오 특징 추출 방법이 제공된다.

[0013] 본 발명의 일 실시예에 따르면, (a) 입력 비디오의 각 프레임에 대한 프레임 수준 특징 벡터를 추출하는 단계; (b) 상기 프레임 수준 특징 벡터를 학습된 방해자 구별 모델에 적용하여 방해자 프레임을 선별하여 제거하는 단계; (c) 상기 방해자 프레임이 제거된 프레임 수준 특징 벡터들을 학습된 시간적 중요도 모델에 적용하여 시간적 중요도 가중치를 계산하고, 상기 시간적 중요도 가중치를 이용하여 상기 프레임 수준 특징 벡터들에 대한 방해자 프레임 억제 가중치를 생성하는 단계; 및 (d) 상기 방해자 프레임 억제 가중치와 프레임 수준 특징 벡터들을 집계하여 비디오 수준 특징 벡터를 생성하는 단계를 포함하는 방해자 프레임 억제를 적용한 비디오 특징 추출 방법이 제공될 수 있다.

[0014] 상기 (b) 단계는, 상기 프레임 수준 특징 벡터의 크기가 제1 임계치 미만인 프레임을 방해자 프레임으로 선별하여 제거할 수 있다.

[0015] 상기 (c) 단계는, 상기 방해자 프레임이 제거된 프레임 수준 특징 벡터들을 학습된 시간적 중요도 모델에 적용하여 자기 유사도 맵을 생성하고, 상기 자기 유사도 맵을 인공 신경망에 적용한 후 대각 성분만 추출하여 시간적 중요도 가중치를 생성할 수 있다.

[0016] 상기 (a) 단계 이전에, 상기 방해자 구별 모델이 학습되되, 상기 방해자 구별 모델의 학습 과정에서, 상기 프레임 수준 특징 벡터의 크기를 각각 산출한 후 임계치보다 낮은 크기값을 가지는 프레임들을 방해자 프레임으로 선별하여 방해자 프레임 임시 라벨을 부여하고, 상기 방해자 프레임 임시 라벨을 이용하여 상기 입력 비디오(X)에 대해 일정 비율만큼 방해자 프레임 수준 특징 벡터를 추가할 수 있다.

[0017] 상기 학습 과정에서 상기 입력 비디오(X)는 앵커, 포지티브 및 네거티브 샘플로 입력될 수 있다.

[0018] 상기 (a) 단계 이전에, 시간적 중요도 모델의 학습 과정이 선행되되, 상기 학습 과정에서 상기 시간적 중요도 모델은 상기 앵커와 상기 포지티브 쌍의 프레임 수준 특징 벡터들에 대한 제1 프레임 수준 유사도 맵을 산출하고, 포지티브 프레임 수준 특징 벡터의 시간 차원에 대해 맥스 풀링을 통해 제1 시간적 중요도 가중치를 생성하고, 상기 제1 시간적 중요도 가중치를 이용하여 앵커 샘플들에 대한 제1 시간적 중요도 임시 라벨을 부여하도록 학습되고, 상기 앵커와 상기 네거티브 쌍의 프레임 수준 특징 벡터들에 대한 제2 프레임 수준 유사도 맵을 산출하고, 네거티브 프레임 수준 특징 벡터의 시간 차원에 대해 맥스 풀링을 통해 제2 시간적 중요도 가중치를 생성하고, 상기 제2 시간적 중요도 가중치를 이용하여 앵커 샘플들에 대한 제2 시간적 중요도 임시 라벨을 부여하도록 학습되며, 상기 앵커 프레임 수준 특징 벡터들에 대한 자기-유사도 맵을 산출하고 병목 레이어와 트랜스포머 인코더를 적용한 후 대각 성분만 추출하여 최종 시간적 중요도 가중치를 산출하도록 학습될 수 있다.

[0020] 본 발명의 다른 측면에 따르면, 방해자 프레임 억제를 적용한 비디오 특징 추출 장치가 제공된다.

[0021] 본 발명의 일 실시예에 따르면, 입력 비디오의 각 프레임에 대한 프레임 수준 특징 벡터를 추출하는 특징 추출부; 상기 프레임 수준 특징 벡터를 학습된 방해자 구별 모델에 적용하여 방해자 프레임을 선별하여 제거부; 상기 방해자 프레임이 제거된 프레임 수준 특징 벡터들을 학습된 시간적 중요도 모델에 적용하여 시간적 중요도 가중치를 계산하고, 상기 시간적 중요도 가중치를 이용하여 상기 프레임 수준 특징 벡터들에 대한 방해자 프레임 억제 가중치를 생성하는 억제 가중치 생성부; 및 상기 방해자 프레임 억제 가중치와 프레임 수준 특징 벡터들을 집계하여 비디오 수준 특징 벡터를 생성하는 비디오 수준 특징 생성부를 포함하는 비디오 특징 추출 장치가 제공될 수 있다.

[0022] 상기 제거부는, 상기 프레임 수준 특징 벡터의 크기가 제1 임계치 미만인 프레임을 방해자 프레임으로 선별하여 제거할 수 있다.,

[0023] 상기 억제 가중치 생성부는, 상기 방해자 프레임이 제거된 프레임 수준 특징 벡터들을 학습된 시간적 중요도 모델에 적용하여 자기 유사도 맵을 생성하고, 상기 자기 유사도 맵을 인공 신경망에 적용한 후 대각 성분만 추출하여 시간적 중요도 가중치를 생성할 수 있다.

[0024] 상기 제거부는, 상기 프레임 수준 특징 벡터의 크기를 각각 산출한 후 임계치보다 낮은 크기값을 가지는 프레임들을 방해자 프레임으로 선별하여 방해자 프레임 임시 라벨을 부여하고, 상기 방해자 프레임 임시 라벨을 이용하여 상기 입력 비디오(X)에 대해 일정 비율만큼 방해자 프레임 수준 특징 벡터를 추가하여 상기 방해자 구별 모델을 사전 학습시킬 수 있다.

[0025] 상기 억제 가중치 생성부는, 시간적 중요도 모델을 사전 학습시키되, 상기 학습 과정에서 상기 시간적 중요도 모델은 상기 앵커와 상기 포지티브 쌍의 프레임 수준 특징 벡터들에 대한 제1 프레임 수준 유사도 맵을 산출하고, 포지티브 프레임 수준 특징 벡터의 시간 차원에 대해 맥스 풀링을 통해 제1 시간적 중요도 가중치를 생성하고, 상기 제1 시간적 중요도 가중치를 이용하여 앵커 샘플들에 대한 제1 시간적 중요도 임시 라벨을 부여하도록 학습되고, 상기 앵커와 상기 네거티브 쌍의 프레임 수준 특징 벡터들에 대한 제2 프레임 수준 유사도 맵을 산출하고, 네거티브 프레임 수준 특징 벡터의 시간 차원에 대해 맥스 풀링을 통해 제2 시간적 중요도 가중치를 생성하고, 상기 제2 시간적 중요도 가중치를 이용하여 앵커 샘플들에 대한 제2 시간적 중요도 임시 라벨을 부여하도록 학습되며, 상기 앵커 프레임 수준 특징 벡터들에 대한 자기-유사도 맵을 산출하고 병목 레이어와 트랜스포머 인코더를 적용한 후 대각 성분만 추출하여 최종 시간적 중요도 가중치를 산출하도록 학습될 수 있다.

발명의 효과

[0027] 본 발명의 일 실시예에 따른 방해자 프레임 억제를 적용한 비디오 특징 추출 방법 및 그 장치를 제공함으로써, 시각적 중정보가 부족한 프레임들을 제거하고, 중요도가 높은 특징 벡터들을 비디오 수준 특징 벡터로 집적하여

프레임의 시각적/의미론적 정보에 기반하여 비디오 수준 서술자의 표현력을 개선할 수 있다.

도면의 간단한 설명

[0029] 도 1은 방해자 프레임의 예시를 도시한 도면.

도 2는 본 발명의 일 실시예에 따른 비디오 특징 추출 장치의 내부 구성을 개략적으로 도시한 도면.

도 3은 본 발명의 일 실시예에 따른 전체 프레임워크를 도시한 도면.

도 4는 본 발명의 일 실시예에 따른 방해자 구별 모델의 상세 구조를 도시한 도면.

도 5는 본 발명의 일 실시예에 따른 시간적 중요도 모델의 상세 구조를 도시한 도면.

도 6은 본 발명의 일 실시예에 따른 주제 지도 모델의 상세 구조를 도시한 도면.

도 7은 종래 기술들과의 비디오 검색 정확도-속도를 비교한 그래프.

도 8은 본 발명의 일 실시예에 따른 검색 정확도 개선 효과를 나타낸 도면.

도 9는 본 발명의 일 실시예에 따른 비디오 특징 추출 장치의 학습 과정을 나타낸 순서도.

도 10은 본 발명의 일 실시예에 따른 학습 과정 파이프 라인을 도시한 도면.

도 11은 본 발명의 일 실시예에 따른 방해자 프레임 억제를 적용한 비디오 특징 추출 방법을 나타낸 순서도.

발명을 실시하기 위한 구체적인 내용

[0030] 본 명세서에서 사용되는 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "구성된다" 또는 "포함한다" 등의 용어는 명세서상에 기재된 여러 구성 요소들, 또는 여러 단계들을 반드시 모두 포함하는 것으로 해석되지 않아야 하며, 그 중 일부 구성 요소들 또는 일부 단계들은 포함되지 않을 수도 있고, 또는 추가적인 구성 요소 또는 단계들을 더 포함할 수 있는 것으로 해석되어야 한다. 또한, 명세서에 기재된 "...부", "모듈" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어 또는 소프트웨어로 구현되거나 하드웨어와 소프트웨어의 결합으로 구현될 수 있다.

[0031] 이하, 첨부된 도면들을 참조하여 본 발명의 실시예를 상세히 설명한다.

[0033] 도 2는 본 발명의 일 실시예에 따른 비디오 특징 추출 장치의 내부 구성을 개략적으로 도시한 도면이고, 도 3은 본 발명의 일 실시예에 따른 전체 프레임워크를 도시한 도면이고, 도 4는 본 발명의 일 실시예에 따른 방해자 구별 모델의 상세 구조를 도시한 도면이며, 도 5는 본 발명의 일 실시예에 따른 시간적 중요도 모델의 상세 구조를 도시한 도면이고, 도 6은 본 발명의 일 실시예에 따른 주제 지도 모델의 상세 구조를 도시한 도면이다.

[0034] 도 2를 참조하면, 본 발명의 일 실시예에 따른 비디오 특징 추출 장치(200)는 특징 추출부(210), 제거부(220), 억제 가중치 생성부(230), 비디오 수준 특징 생성부(240), 메모리(250) 및 프로세서(260)를 포함하여 구성된다.

[0035] 특징 추출부(210)는 비디오의 각 프레임에서 프레임 수준 특징 벡터를 추출하기 위한 수단이다.

[0036] 본 발명의 일 실시예에 따르면 비디오의 프레임들에 대한 시각적 특징 정보를 요약하는 특징 벡터가 추출될 수 있다. 특징 벡터는 영상내 국소 특징점 서술자(SIFT)를 집적하거나 인공신경망을 활용하여 산출될 수 있다.

[0037] 프레임 수준 특징 벡터는 복수의 프레임을 참조하여 산출될 수 있다. 예를 들어, f_t를 산출하기 위해서는 t번째 프레임 전후 N개의 프레임이 참조될 수 있다. 본 발명의 일 실시예에서는 이해와 설명의 편의를 도모하기 위해 프레임 수준 특징 벡터가 단일 프레임 xt로부터 산출되는 L_n-iMAC 특징 벡터인 것을 가정하기로 한다.

[0038] 각 프레임은 백본 네트워크 Φ 에 입력되고, 백본 네트워크 Φ 의 중간 컨볼루션 레이어들이 생성하는 중간

특징맵($\mathcal{M}^{(k)} \in \mathbb{R}^{S^2 \times C^{(k)}} (k=1,\cdots,K)$)들로부터 R-MAC(Regional Maximum Activation of Convolution)이 추출되어 L_n-iMAC 특징 벡터가 구성될 수 있다.

[0039] 백본 네트워크 Φ 의 K개 중간 컨볼루션 레이어의 출력에 대해 미리 정해진 (SxS)개 구획별로 regional maximum activation pooling을 수행한 후 풀링된 MAC 특징 벡터를 채널 축으로 결합(concatenation)함으로써 프레임 수준 특징 벡터($x \in \mathbb{R}^{S^2 \times C} (C=\sum_{k=1}^{K} C^{(k)})$)가 생성될 수 있다.

[0040] 각 프레임 수준 특징 벡터들에 PCA 화이트닝(whitening)을 적용하면 L_n-iMAC 특징($X \in \mathbb{R}^{T \times S^2 \times C}$)이 획득될 수 있다. PCA 화이트닝을 적용할 때 채널 축 차원을 다른 크기로 줄일 수 있으나, 편의상 프레임 수준 특징의 차원을 C라고 가정하기로 한다. 따라서, 프레임 수준 특징 벡터의 차원은 S x S x C가 된다. T개의 프레임에 대한 특징 벡터 모음을 F로 표기하기로 하며, 그 차원은 (TxSxSxC)로 나타낼 수 있다.

[0041] 도 3은 본 발명의 일 실시예에 따른 전체 프레임워크가 도시되어 있다.

[0042] 학습 과정에서는 3개의 비디오(포지티브 비디오, 앵커 비디오, 네거티브 비디오)가 입력되어 각 프레임에 대한 프레임 수준 특징 벡터가 추출될 수 있다.

[0043] 추론 과정에서는 도 3에 도시된 바와 같이, 타겟 비디오만 특징 추출부(210)로 입력되어 프레임 수준 특징 벡터가 추출될 수 있다.

[0044] 제거부(220)는 프레임 수준 특징 벡터를 이용하여 방해자 프레임을 제거하기 위한 수단이다.

[0045] 본 발명의 일 실시예에 따르면 방해자 프레임은 픽셀을 비롯한 저수준 특징(에지, 코너 등)이 적은 프레임으로 정의하기로 한다.

[0046] 제거부(220)는 방해자 구별 모델을 이용하여 시각 정보가 적은 프레임들을 방해자 프레임으로 판단하여 제거할 수 있다. 이를 위해 제거부(220)는 프레임 수준 특징 벡터를 방해자 구별 모델에 입력한 후 방해자 구별 모델이 이를 구별하도록 최적화할 수 있다. 방해자 구별 모델의 상세 구조는 도 4에 도시되어 있다.

[0047] 방해자 구별 모델이 방해자 프레임을 잘 구분할 수 있도록 학습시키기 위해, 프레임 수준 특징 벡터의 크기를 사용하여 임시 라벨을 생성할 수 있다. 적은 양의 저수준 정보를 가진 프레임들은 백본 네트워크의 컨볼루션 레이어에서 더 적은 값들을 활성화시키고 결과적으로 L_n-iMAC의 크기가 작아지게 된다.

[0048] 따라서, 학습에 앞서 프레임 수준 특징 벡터의 크기가 임계값 이하인 프레임들을 방해자 프레임으로 선별하여 임시 라벨을 부여하여 방해자 세트를 구성할 수 있다.

[0049] 학습 과정에서 임시 라벨을 이용하여 입력 비디오(X)에 대해 임의의 비율(예를 들어, 20% 내지 50%)만큼 방해자 프레임이 추가될 수 있다. 동시에 방해자 프레임이 주입된 지점은 0으로 설정되고, 방해자 프레임이 주입되지 않은 지점은 1로 설정되어 임시 라벨($Y_{di} = \{y_{di}^{(t)}\}_{t=1}^{T'}$)이 정의될 수 있다. 또한, 방해자 프레임이 추가됨에 따라 입력 비디오(X)에 대한 프레임 수준 특징 벡터들의 차원은 시간적 차원이 늘어나($T'xSxSxC$)가 된다. 주입된 특징 벡터는 방해자 프레임 억제 가중치(W_{di})를 생성하기 위해 멀티 레이어를 통해 투영될 수 있다. 각 프레임에 대한 특징 벡터만 이용하여 방해자 프레임을 식별하는 기준이 결정되기 때문에 멀티 레이어는 FC 레이어로만 구성되어 프레임 간 상호 작용 없이 각 프레임을 독립적으로 처리할 수 있다.

[0050] 결과적으로, 방해자 구별 모델은 방해자 프레임 억제 가중치(W_{di})와 임시 라벨 사이(Y_{di})의 이진 교차 엔트로피 손실로 계산된 구별 손실($\mathcal{L}_{di}$)을 통해 최적화될 수 있다.

[0051] 방해자 구별 모델의 목적은 프레임 수준 특징 벡터를 다음 단계로 전달하고 방해자 프레임 억제 가중치를 위한 임계값 지정을 통해 방해자 프레임에 대한 특징 벡터를 제거하는 것이다. 이 경우, 학습 과정에서는 임계값 연산이 미분 불가능하므로 출력은 방해자 프레임 억제 가중치(W_{di})와 입력 특징 X의 아다마르 곱에서 파생되며, 추론 단계에서는 출력이 방해자 임계값(λ_{di})을 기반으로 한 임계값 연산에서 추론될 수 있다.

[0052] 억제 가중치 생성부(230)는 방해자 프레임이 제거된 프레임 수준 특징 벡터들을 이용하여 중요도 가중치를 계산하고, 중요도 가중치를 이용하여 상기 프레임 수준 특징 벡터들에 대한 방해자 프레임 억제 가중치를 생성한다.

[0053] 이러한 억제 가중치 생성부(230)는 시간적 중요도 모델과 주제 지도 모델을 포함할 수 있다. 시간적 중요도 모델의 상세 구조는 도 5에 도시되어 있다.

[0054] 시간적 중요도 모델은 프레임 수준 유사도로부터 얻어지는 중요한 정보(saliency information)를 통해 각 프레임이 갖는 중요성을 평가하고, 주제 지도 모델은 비디오의 주제와 각 프레임이 관련된 정보를 측정할 수 있다.

[0055] 학습 과정에서 시간적 중요도 모델로 트리플렛 형태의 입력에 따라 앵커, 포지티브, 네거티브 샘플로 구성된 3개의 프레임 수준 특징 벡터들이 입력될 수 있다. 시간적 중요도 모델은 앵커와 포지티브 쌍간의 TD(tensor dot) 연산을 수행하여 $\mathbb{R}^{T' \times S^2 \times S^2 \times T'}$ 크기의 출력을 얻을 수 있다.

[0056] 이해와 설명의 편의를 도모하기 위해 TD 연산에 대해 간략하게 설명하기로 한다. TD 연산은 주어진 두 텐서($\mathcal{A}, \mathcal{B} \in \mathbb{R}^{T \times S^2 \times C}$) 사이의 특정 축에 대한 합으로 정의될 수 있다. 여기서, T, S, C는 각각 시간, 공간, 및 채널 축을 나타낸다. 본 발명의 일 실시예에서는 채널 축으로의 TD 연산을 수행하기 때문에, TD 연산은 $\mathbb{R}^{T' \times S^2 \times S^2 \times T'}$ 크기의 출력을 갖는다.

[0057] 이어, 시간적 중요도 모델은 TD 연산 수행 결과에 CS 연산을 수행하여 $\mathbb{R}^{T' \times T'}$ 크기를 가지는 프레임 수준 유사도 맵을 도출할 수 있다. 이해와 설명의 편의를 도모하기 위해 CS 연산에 대해 간략하게 설명하기로 한다.

[0058] 포지티브 비디오에 대한 프레임 수준 특징 벡터들과 앵커 비디오에 대한 프레임 수준 특징 벡터를 각각 행과 열에 위치시킨 후 최대값만 샘플링하여 1D 벡터를 도출하고, 이를 다시 평균하여 프레임 수준 유사도 맵이 도출될 수 있다.

[0059] CS 연산은 수학식 1과 같이 도출될 수 있다.

수학식 1

[0060]
$$CS(\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N}\max_{j\in[1,M]}\mathcal{D}^{(i,j)}$$

[0061] 여기서, $\mathcal{D} \in \mathbb{R}^{N \times M}$ 는 행렬을 나타낸다. CS 연산은 수학식 1과 같이 주어진 행렬에 대해 열 축의 최

대값의 평균으로 정의될 수 있다. 행렬(D)의 크기가 $\mathbb{R}^{T\times S^2\times S^2\times T}$ 인 경우 출력의 크기는 $\mathbb{R}^{T\times T}$ 이 된다. 만일 행렬의 크기가 $\mathbb{R}^{T\times T}$ 라면, 출력의 크기는 $\mathbb{R}^{1}$ 이 된다.

[0062] 앵커와 네거티브 샘플에 대해서도 동일하게 프레임 수준 유사도 맵이 도출될 수 있다. 여기서, 프레임 수준 유사도 맵의 각 행은 앵커의 프레임 인덱스에 해당할 수 있다.

[0063] 두 개의 프레임 수준 유사도 맵은 복수의 컨볼루션 레이어를 통과하여 $\mathcal{D}_p, \mathcal{D}_n \in \mathbb{R}^{T''\times T''}$ 크기로 조정될 수 있다. 여기서, $\mathcal{D}_p$ 는 앵커와 포지티브 쌍과의 프레임 수준 유사도 맵을 나타내고, $\mathcal{D}_n$ 는 앵커와 네거티브 쌍과의 프레임 수준 유사도 맵을 나타내며, T'' 는 축소된 유사도 맵의 크기를 나타낸다.

[0064] 이후, 두개의 프레임 수준 유사도 맵들의 CS 연산 후 계산된 삼중항 마진 손실(triplet margin loss)는 정규화 손실과 합쳐저 사용되어 조정된 프레임 수준 유사도 맵이 발산되는 것을 방지할 수 있다.

[0065] 프레임 수준 유사도 맵의 조정 과정에서 포지티브 쌍의 프레임 수준 유사도 맵 상의 중요도 정보(saliency information)를 가지는 영역들이 더욱 강조되며, 중요도 임시 라벨은 이를 이용하여 생성될 수 있다.

[0066] 이를 수학식으로 나타내면, 수학식 2와 같다.

수학식 2

$$\rho_i = \max_{j\in[1,T'']} \mathcal{D}_p^{(i,j)},$$

$$\rho = \left[\rho_1,\ \rho_2,\ \cdots,\ \rho_i,\ \cdots,\ \rho_{T''}\right]^{\mathbf{T}},$$

$$Y_{sa} = H\left(\rho - \frac{1}{T''}\sum_{i=1}^{T''}\rho_i\right) \in \mathbb{R}^{T''},$$

[0067]

[0068] 여기서, $\mathbf{T}$ 는 전치 연산(transpose operation)을 나타내고, H는 단위 계단 함수(Heaviside step function)를 나타내고, ρ 는 포지티브 비디오에 대한 앵커 비디오의 각 프레임 중 가장 높은 유사도를 나타낸다. ρ 의 평균은 양수 점수(positive score), 즉 포지티브 비디오쌍 간의 비디오 수준 유사도이다.

[0069] 포지티브 비디오의 시간 차원에 대해 맥스 풀링(max-pooling)을 통해 시간적 중요도 가중치를 정의할 수 있다. 시간적 중요도 가중치가 일정 임계치(ρ)보다 높은 값을 가지는 프레임을 시간적 중요도 프레임으로 정의하고, 이를 앵커 비디오에 대한 시간적 중요도 임시 라벨(saliency label)로 정의할 수 있다.

[0070] 시간적 중요도 라벨(saliency label)은 ρ 값이 평균보다 크면 1면 ρ_i 는 1로 설정되고, 그렇지 않으면 0으로 설정될 수 있다.

[0071] 시간적 중요도 라벨 생성 과정을 거친 후 시간적 중요도 모델은 앵커 비디오의 프레임 수준 특징 벡터들에 대해

TD 및 CS 연산을 수행한 후 자기 유사도 맵(self-similarity map)을 생성하고, 대각 성분만 추출하는 대각선 샘플링을 통해 시간적 중요도 가중치(W_{sa})를 생성할 수 있다.

[0072] 자기 유사도 맵은 앵커 비디오의 프레임 수준 특징 벡터들에 대해 TD 연산과 CS 연산을 적용한 후 병목 레이어와 트랜스포머 인코더를 통과한 후 시그모이드 레이어를 통과함으로써 시간적 중요도 가중치($W_{sa} = \{w_{sa}^{(t)}\}_{t=1}^{T'}$) 생성에 이용될 수 있다.

[0073] 여기서, 시간적 중요도 손실($\mathcal{L}_{sa}$)은 시간적 중요도 가중치(W_{sa})와 시간적 중요도 라벨(Y_{sa}) 사이의 이진 교차 엔트로피 손실로 계산될 수 있다. 여기서, 최근접 인터폴레이션이 출력 길이(T')와 일치하도록 레이블에 적용될 수 있다.

[0074] 학습 과정에서 시간적 중요도 모델은 시간적 중요도 손실($\mathcal{L}_{sa}$)과 프레임 손실($\mathcal{L}_{fr}$)로 학습될 수 있다.

[0075] 학습이 완료된 이후 추론 과정에서는 타겟 비디오에 대한 자기 유사성 맵을 통해 생성된 시간적 중요도 가중치만 생성할 수 있다.

[0076] 주제 지도 모델의 상세 구조는 도 6에 도시된 바와 같다.

[0077] 비디오 특성상 영상내에는 주제와 무관한 프레임들(예를 들어, 뉴스 앵커의 설명, 편집 효과 등)이 포함될 수 있다. 따라서, 주제 지도 모델은 영상 전체에 대해 시공간 글로벌 에버리지 풀링(ST-GAP)을 통해 영상의 주요 성분만 담은 의사 주제($G \in \mathbb{R}^C$)를 추출할 수 있다.

[0078] 이어, 주제 지도 모델은 영상의 시간 축을 남기고 공간 축에 대해 글로벌 에버리지 풀링(S-GAP)을 수행하여 $T' \times C$ 크기의 행렬을 얻고, 각 프레임과 G 사이의 코사인 유사도를 구하여 초기 상태($I \in \mathbb{R}^{T'}$)를 생성할 수 있다.

[0079] 이는 T'개의 각 프레임이 영상의 주된 내용(주제)와 얼만큼 유사한지를 대략적으로 나타낼 수 있다.

[0080] 초기 상태를 신경망에 입력하여 정제함으로써 더욱 상세하게 각 프레임이 주제와 얼마나 유관한지를 나타내는 주제 지도 가중치($W_{gu} = \{w_{gu}^{(t)}\}_{t=1}^{T'}$)를 생성할 수 있다.

[0081] 즉, 주제 지도 모델은 1 x 3 크기의 컨볼루션 레이어를 통해 인접한 프레임들을 살펴보고, 최종적으로 1x1 컨볼루션 레이어를 통해 채널 수를 축소할 수 있다. 여러 계층을 쌓음으로써 수용 필드가 점차 커지게 되는데, 각기 다른 수용 필드를 갖는 특징 맵들을 결합하고, 1 x 1 커널을 한번 더 통과하여 출력 가중치를 생성할 수 있다.

[0082] 주제 지도 모델은 활성화 함수로 Tempered 시그모이드를 사용하여 노이즈로부터 강건하게 가중치를 학습할 수 있다.

[0083] 다시 정리하면, 억제 가중치 생성부(230)는 시간적 중요도 모델에서 출력되는 시간적 중요도 가중치를 이용하여 프레임 수준 특징 벡터들에 대한 방해자 프레임 억제 가중치를 생성할 수 있다.

[0084] 비디오 수준 특징 생성부(240)는 방해자 프레임 억제 가중치를 이용하여 비디오 수준 특징 벡터를 생성한다.

[0085] 즉, 학습 단계에서는 프레임 수준 특징 벡터들은 방해자 프레임 억제 가중치와의 아다마르곱(Hadamard product, $X \odot W_{di}$)을 통해 비디오 수준 특징 벡터($V \in \mathbb{R}^C$)로 누적될 수 있다.

[0086] 본 발명의 일 실시예에 따르면, 전체 손실 함수는 수학식 3과 같이 계산될 수 있다.

수학식 3

$$\mathcal{L} = \mathcal{L}_{vi} + \mathcal{L}_{fr} + \mathcal{L}_{sa} + \alpha\mathcal{L}_{di}$$

[0087]

[0088] 여기서, α는 방해자 구별 모델의 학습 속도를 조절하는 매개변수이며, 해당 모델이 다른 모델에 비해 빠르게 학습되기 때문에 도입되었다.

[0089] 학습이 완료되면, 비디오 수준 특징 생성부(240)는 방해자 프레임 억제 가중치의 값이 임계치(예를 들어, 0.5) 이상인 프레임 수준 특징 벡터만 사용하여 입력 비디오(X)에 대해 방해자 프레임을 억제하여 비디오 수준 특징 벡터를 생성할 수 있다.

[0090] 메모리(250)는 본 발명의 일 실시예에 따른 방해자 프레임 억제를 적용한 비디오 특징 추출 방법을 수행하기 위한 명령어를 저장한다.

[0091] 프로세서(260)는 본 발명의 일 실시예에 따른 비디오 특징 추출 장치(200)의 내부 구성 요소들(예를 들어, 특징 추출부(210), 제거부(220), 억제 가중치 생성부(230), 비디오 수준 특징 생성부(240), 메모리(250) 등)을 제어하기 위한 수단이다.

[0092] 도 7은 종래 기술들과의 비디오 검색 정확도-속도를 비교한 그래프이다. 도 7의 x축은 검색 반응 속도를 나타내고, y축은 검색 정확도를 나타낸다. 도 7에서 보여지는 바와 같이, 본 발명의 일 실시예에 따른 방법이 비디오 수준 특징 벡터 방식 중 가장 높은 검색 정확도를 보여주며 일부 프레임 수준 특징 벡터 방식과도 비슷한 검색 정확도를 보이는 것을 알 수 있다.

[0093] 도 8은 본 발명의 일 실시예에 따른 검색 정확도 개선 효과를 나타낸 도면이다. 도 8의 (a)를 살펴보면 보면 FIVR5K 비디오 검색 데이터 상황에서의 기초 실험 구조에 대한 정확도 성능을 확인할 수 있다. 도 8의 (b)에는 기초 실험 구조에서 추가적으로 방해자 구별 모델(DDM)을 활용한 경우 모든 검색 상황에서 개선된 검색 정확도를 확인할 수 있다. 또한, 도 8의 (c)에서 보여지는 바와 같이, 기초 실험 구조에서 추가적으로 시간적 중요도 모델(TSM)을 활용한 경우 동일하게 모든 검색 상황에서 개선된 검색 정확도를 확인할 수 있다. 도 8의 (f)를 살펴보면 기초 실험 구조에서 방해자 구별 모델(DDM)과 시간적 중요도 모델(TSM)을 함께 활용했을 경우 각각 사용한 경우보다 더 좋은 검색 정확도 개선을 확인할 수 있다.

[0095] 도 9는 본 발명의 일 실시예에 따른 비디오 특징 추출 장치의 학습 과정을 나타낸 순서도이고, 도 10은 본 발명의 일 실시예에 따른 학습 과정 파이프 라인을 도시한 도면이다.

[0096] 단계 910에서 비디오 특징 추출 장치(200)는 포지티브, 앵커 및 네거티브 샘플(프레임)을 학습 데이터 셋으로 입력받는다.

[0097] 단계 915에서 비디오 특징 추출 장치(200)는 포지티브 프레임, 네거티브 프레임 및 앵커 프레임에 대한 프레임 수준 특징 벡터를 각각 추출한다.

[0098] 이어, 단계 920에서 비디오 특징 추출 장치(200)는 프레임 수준 특징 벡터를 방해자 구별 모델에 입력하여 방해자 프레임을 선별하여 임시 라벨을 부여한다.

[0099] 예를 들어, 비디오 특징 추출 장치(200)는 프레임 수준 특징 벡터의 크기를 각각 산출한 후 임계치보다 낮은 크기값을 가지는 프레임을 방해자 프레임으로 선별하여 방해자 프레임 임시 라벨을 부여할 수 있다.

[0100] 이어, 단계 925에서 비디오 특징 추출 장치(200)는 방해자 프레임 임시 라벨을 이용하여 입력 비디오(X)에 대해 일정 비율만큼 방해자 프레임 수준 특징 벡터를 추가한다. 이에 따라 입력 비디오(X)에 대한 프레임 수준 특징 벡터들의 차원이 $T' \times S \times S \times C'$로 늘어나게 된다.

[0101] 단계 930에서 비디오 특징 추출 장치(200)는 방해자 프레임 수준 특징 벡터가 추가된 특징 벡터들을 시간적 중요도 모델에 적용하여 시간적 중요도 가중치를 산출한다.

[0102] 즉, 시간적 중요도 모델은 앵커, 포지티브, 네거티브 샘플로 구성된 트리플렛 입력이 들어오면, 앵커와 포지티브 쌍에 대해 유사도 맵을 산출하고, 포지티브 샘플의 시간 차원에 대해 맥스 풀링을 통해 시간적 중요도 가중치를 생성한다.

[0103] 이어, 비디오 특징 추출 장치(200)는 시간적 중요도 가중치가 임계치 이상인 프레임들을 시간적 중요도 프레임으로 정의하고, 앵커 비디오에 대한 시간적 중요도 임시 라벨로 정의할 수 있다.

[0104] 이어, 비디오 특징 추출 장치(200)는 앵커 비디오에 대해 자기 유사도 맵을 생성한 후 인공 신경망과 대각 성분만 추출하는 대각선 샘플링을 통해 시간적 중요도 가중치를 산출할 수 있다.

[0105] 도 9에서는 앵커와 포지티브 쌍에 대한 시간적 중요도 가중치를 산출하는 것만 기술하고 있으나, 도 10에 도시된 바와 같이, 앵커와 네거티브 쌍에 대해서도 시간적 중요도 가중치가 산출될 수 있다.

[0106] 단계 935에서 비디오 특징 추출 장치(200)는 방해자 프레임 수준 특징 벡터가 추가된 특징 벡터들을 주제 지도 모델에 적용하여 주제 지도 가중치를 산출한다.

[0107] 여기서, 단계 935는 생략될 수도 있다.

[0108] 단계 940에서 비디오 특징 추출 장치(200)는 시간적 중요도 가중치를 이용하여 상기 프레임 수준 특징 벡터들에 대한 방해자 프레임 억제 가중치를 생성한다.

[0109] 단계 945에서 비디오 특징 추출 장치(200)는 방해자 프레임 억제 가중치를 이용하여 입력 비디오(X)에 대한 방해자 프레임을 억제하는 비디오 수준 특징 벡터를 생성하도록 학습한다.

[0111] 도 11은 본 발명의 일 실시예에 따른 방해자 프레임 억제를 적용한 비디오 특징 추출 방법을 나타낸 순서도이다.

[0112] 단계 1110에서 비디오 특징 추출 장치(200)는 타겟 프레임을 입력받는다.

[0113] 단계 1115에서 비디오 특징 추출 장치(200)는 타겟 프레임에 대한 프레임 수준 특징 벡터를 생성한다.

[0114] 단계 1120에서 비디오 특징 추출 장치(200)는 프레임 수준 특징 벡터들을 학습된 방해자 구별 모델에 적용하여 방해자 프레임을 제거한다.

[0115] 예를 들어, 비디오 특징 추출 장치(200)는 프레임 수준 특징 벡터들의 크기가 임계치 이상인 프레임을 방해자 프레임을 선별하여 제거할 수 있다.

[0116] 단계 1125에서 비디오 특징 추출 장치(200)는 방해자 프레임이 제거된 프레임 수준 특징 벡터들을 시간적 중요도 모델에 적용하여 자기 유사도 맵을 생성한 후 인공 신경망과 대각 성분만 추출하는 대각선 샘플링을 통해 시간적 중요도 가중치를 산출한다.

[0117] 단계 1130에서 비디오 특징 추출 장치(200)는 시간적 중요도 가중치를 이용하여 프레임 수준 특징 벡터들에 대한 방해자 프레임 억제 가중치를 생성한다.

[0118] 단계 1135에서 비디오 특징 추출 장치(200)는 방해자 프레임 억제 가중치와 프레임 수준 특징 벡터들을 집계하여 비디오 수준 특징 벡터를 생성한다.

[0120] 본 발명의 실시 예에 따른 장치 및 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 컴퓨터 판독 가능 매체에 기록되는 프로그램 명령은

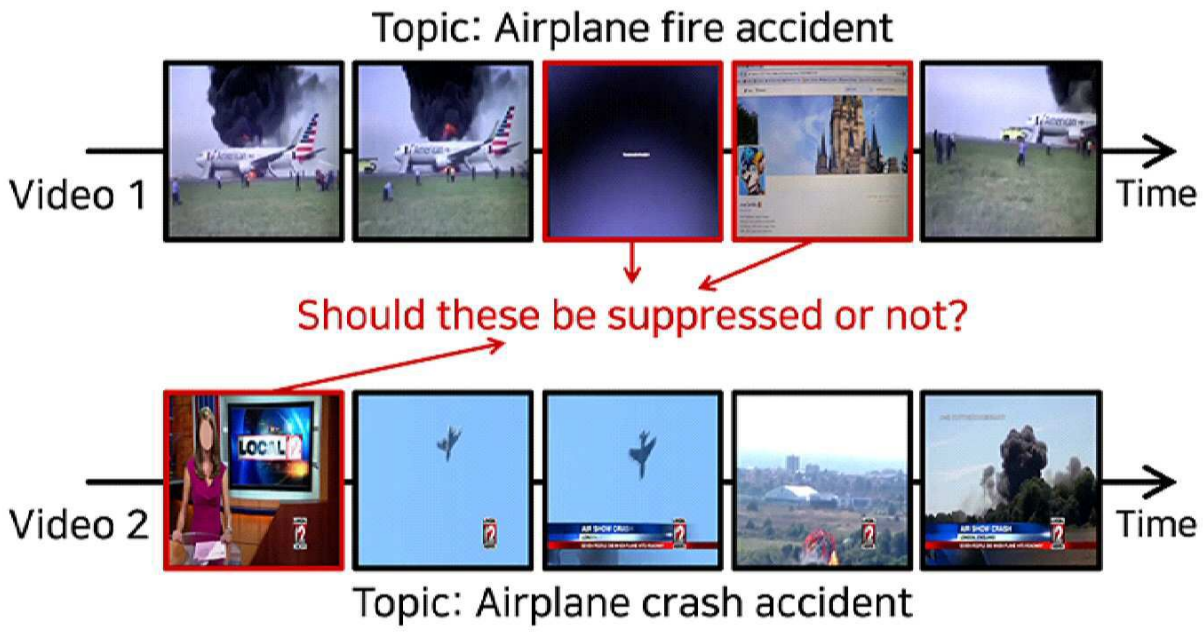
본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 분야 통상의 기술자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media) 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

[0121] 상술한 하드웨어 장치는 본 발명의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

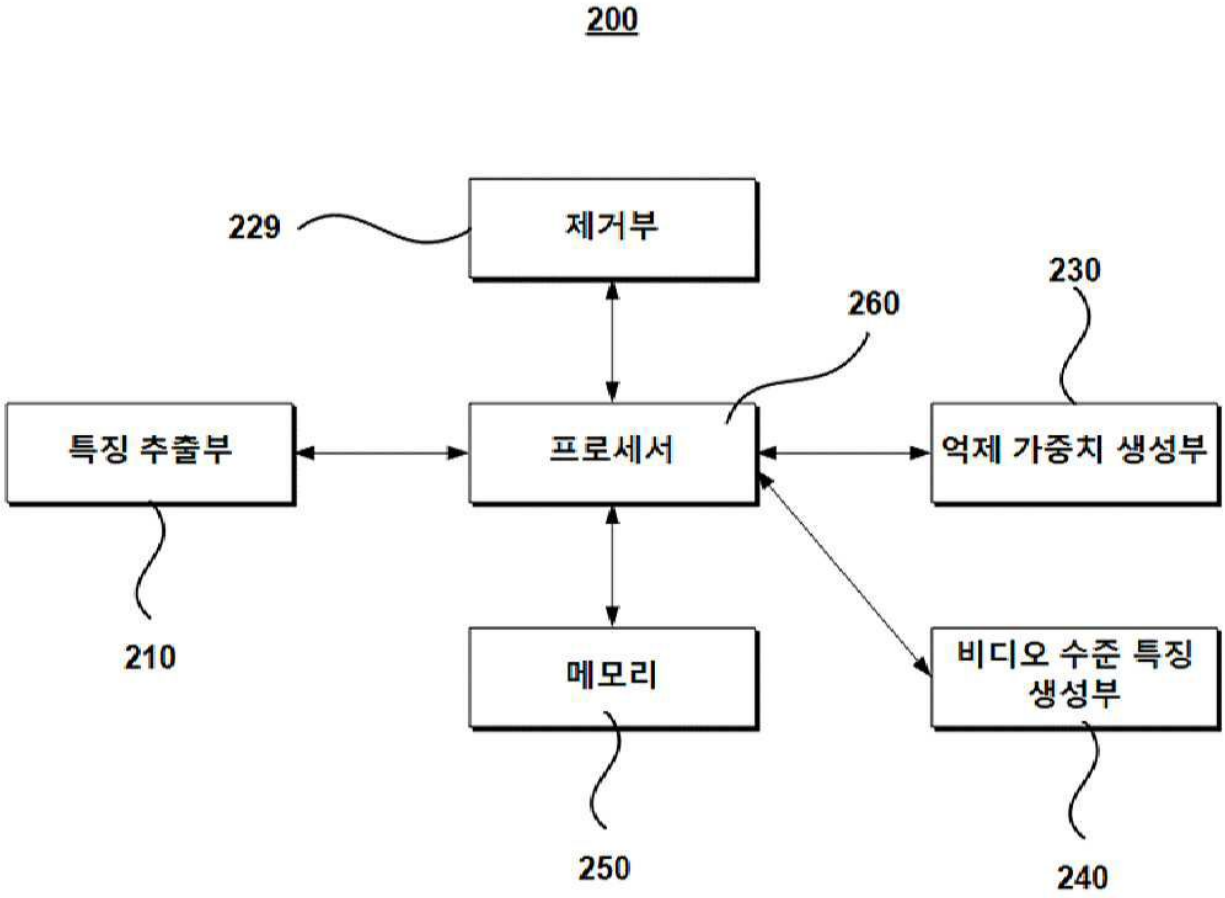
[0122] 이제까지 본 발명에 대하여 그 실시 예들을 중심으로 살펴보았다. 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명이 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 변형된 형태로 구현될 수 있음을 이해할 수 있을 것이다. 그러므로 개시된 실시 예들은 한정적인 관점이 아니라 설명적인 관점에서 고려되어야 한다. 본 발명의 범위는 전술한 설명이 아니라 특허청구범위에 나타나 있으며, 그와 동등한 범위 내에 있는 모든 차이점은 본 발명에 포함된 것으로 해석되어야 할 것이다.

도면

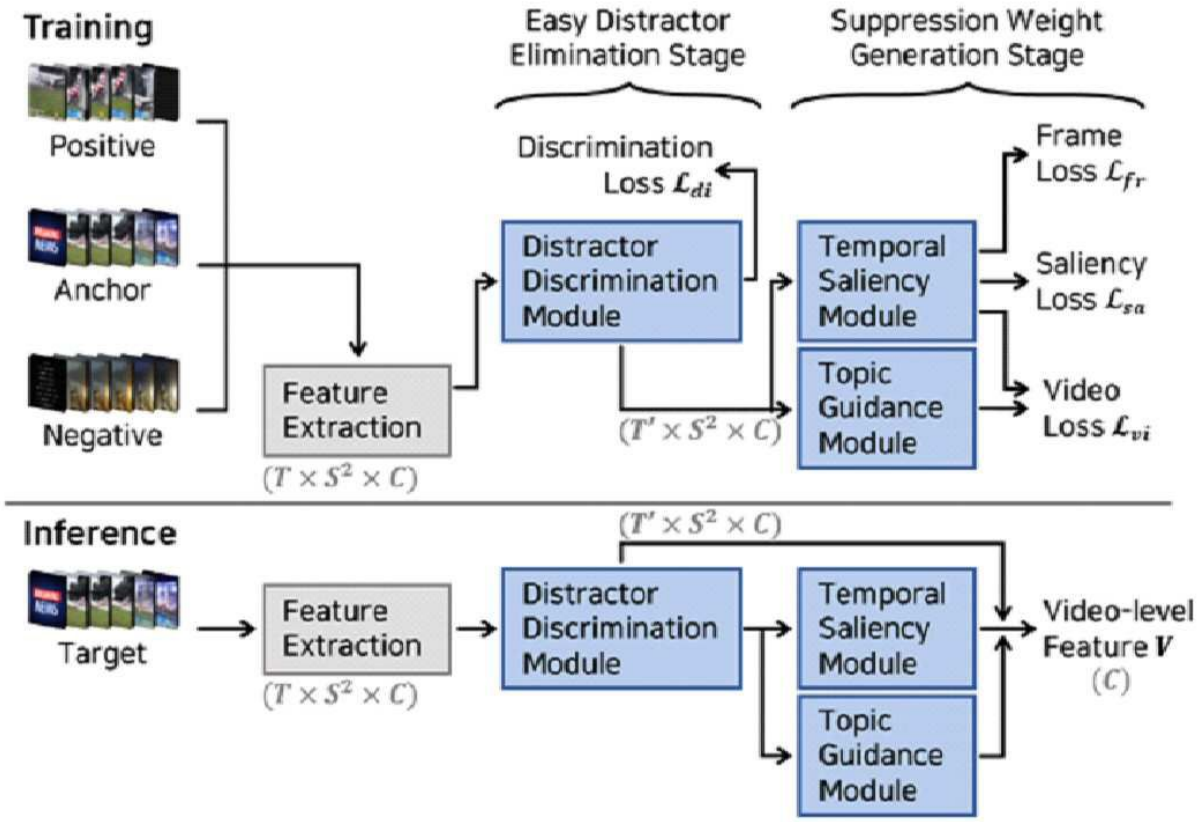
도면1



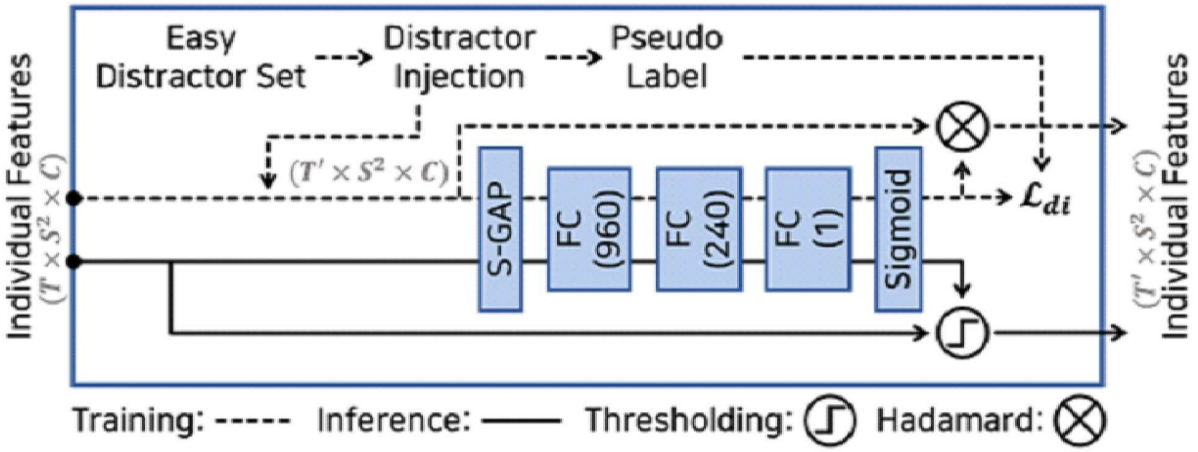
도면2


200
229
제거부
230
260
특징 추출부
프로세서
억제 가중치 생성부
210
메모리
비디오 수준 특징 생성부
250
240

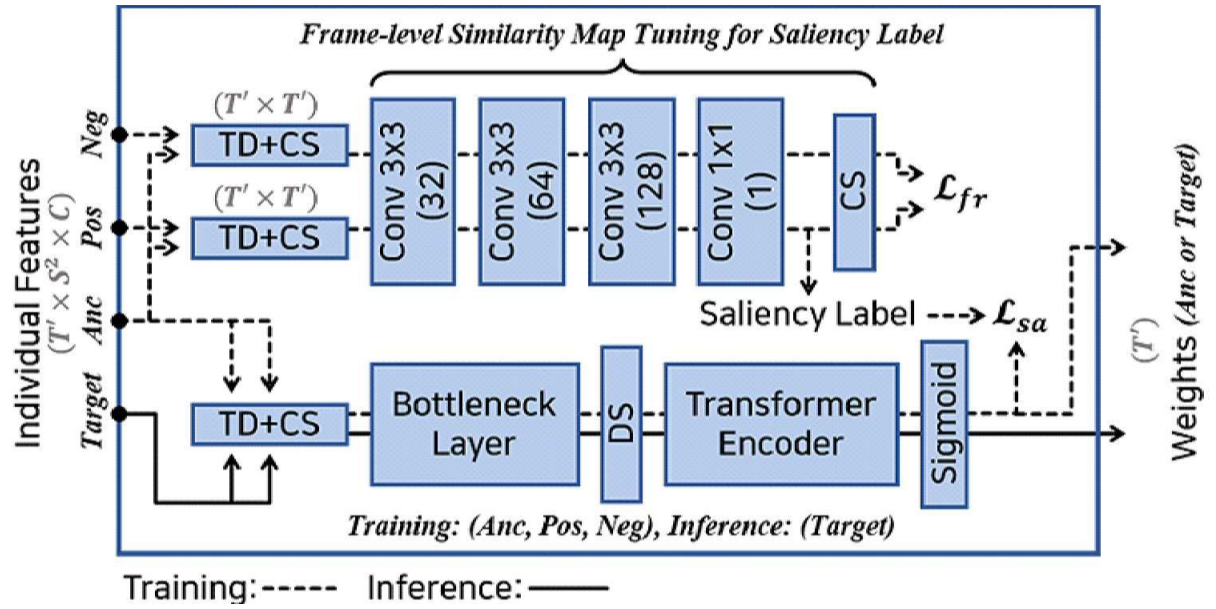
도면3


Training
Positive
Anchor
Negative
Feature Extraction
$(T \times S^2 \times C)$
Easy Distractor Elimination Stage
Suppression Weight Generation Stage
Discrimination Loss $\mathcal{L}_{di}$
Distractor Discrimination Module
$(T' \times S^2 \times C)$
Temporal Saliency Module
Topic Guidance Module
Frame Loss $\mathcal{L}_{fr}$
Saliency Loss $\mathcal{L}_{sa}$
Video Loss $\mathcal{L}_{vi}$
Inference
Target
Feature Extraction
$(T \times S^2 \times C)$
$(T' \times S^2 \times C)$
Distractor Discrimination Module
Temporal Saliency Module
Topic Guidance Module
Video-level Feature V
(C)

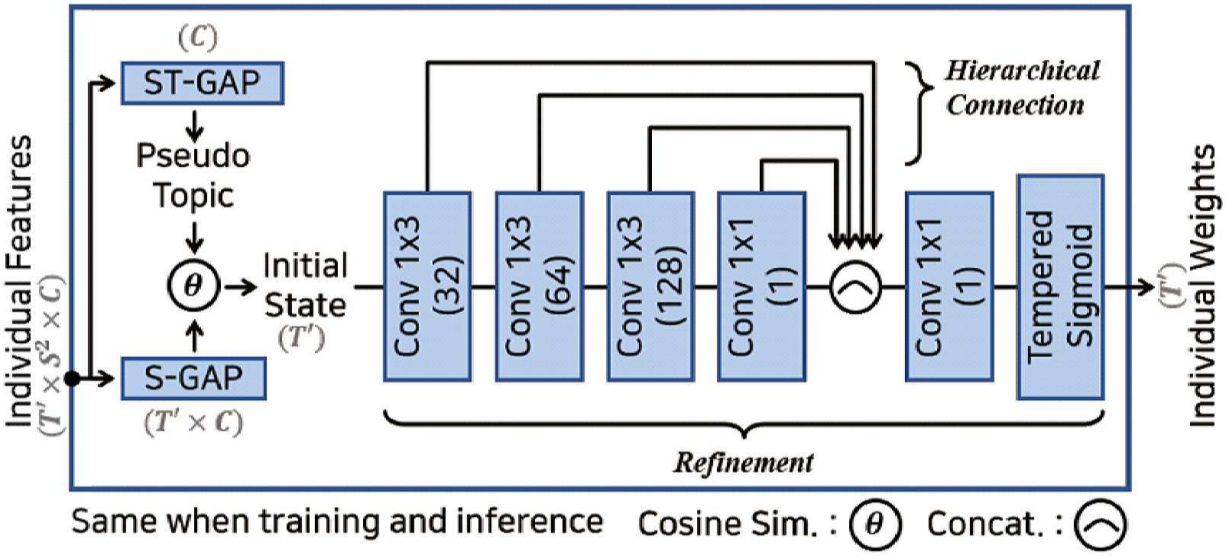
도면4


Easy Distractor Set
Distractor Injection
Pseudo Label
Individual Features
$(T \times S^2 \times C)$
$(T' \times S^2 \times C)$
S-GAP
FC (960)
FC (240)
FC (1)
Sigmoid
$\mathcal{L}_{di}$
$(T' \times S^2 \times C)$
Individual Features
Training: ----- Inference: —— Thresholding: Hadamard:

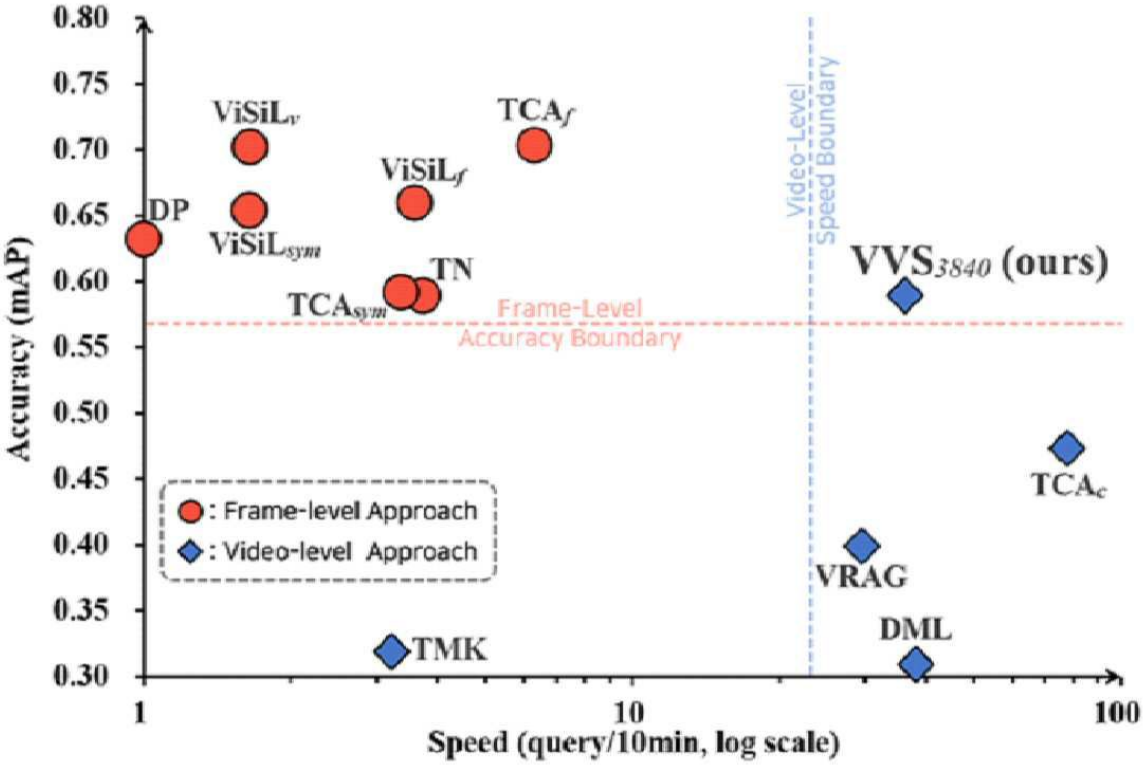
도면5


Frame-level Similarity Map Tuning for Saliency Label
Individual Features
$(T' \times S^2 \times C)$
Neg
Pos
Anc
Target
$(T' \times T')$
TD+CS
$(T' \times T')$
TD+CS
Conv 3x3 (32)
Conv 3x3 (64)
Conv 3x3 (128)
Conv 1x1 (1)
CS
$\mathcal{L}_{fr}$
Saliency Label
$\mathcal{L}_{sa}$
TD+CS
Bottleneck Layer
DS
Transformer Encoder
Sigmoid
(T')
Weights (Anc or Target)
Training: (Anc, Pos, Neg), Inference: (Target)
Training:----- Inference:——

도면6


(C)
ST-GAP
Pseudo Topic
θ
Initial State
(T')
S-GAP
$(T' \times C)$
Individual Features
$(T' \times S^2 \times C)$
Conv 1x3 (32)
Conv 1x3 (64)
Conv 1x3 (128)
Conv 1x1 (1)
Hierarchical Connection
Conv 1x1 (1)
Tempered Sigmoid
(T')
Individual Weights
Refinement
Same when training and inference
Cosine Sim. : θ
Concat. :

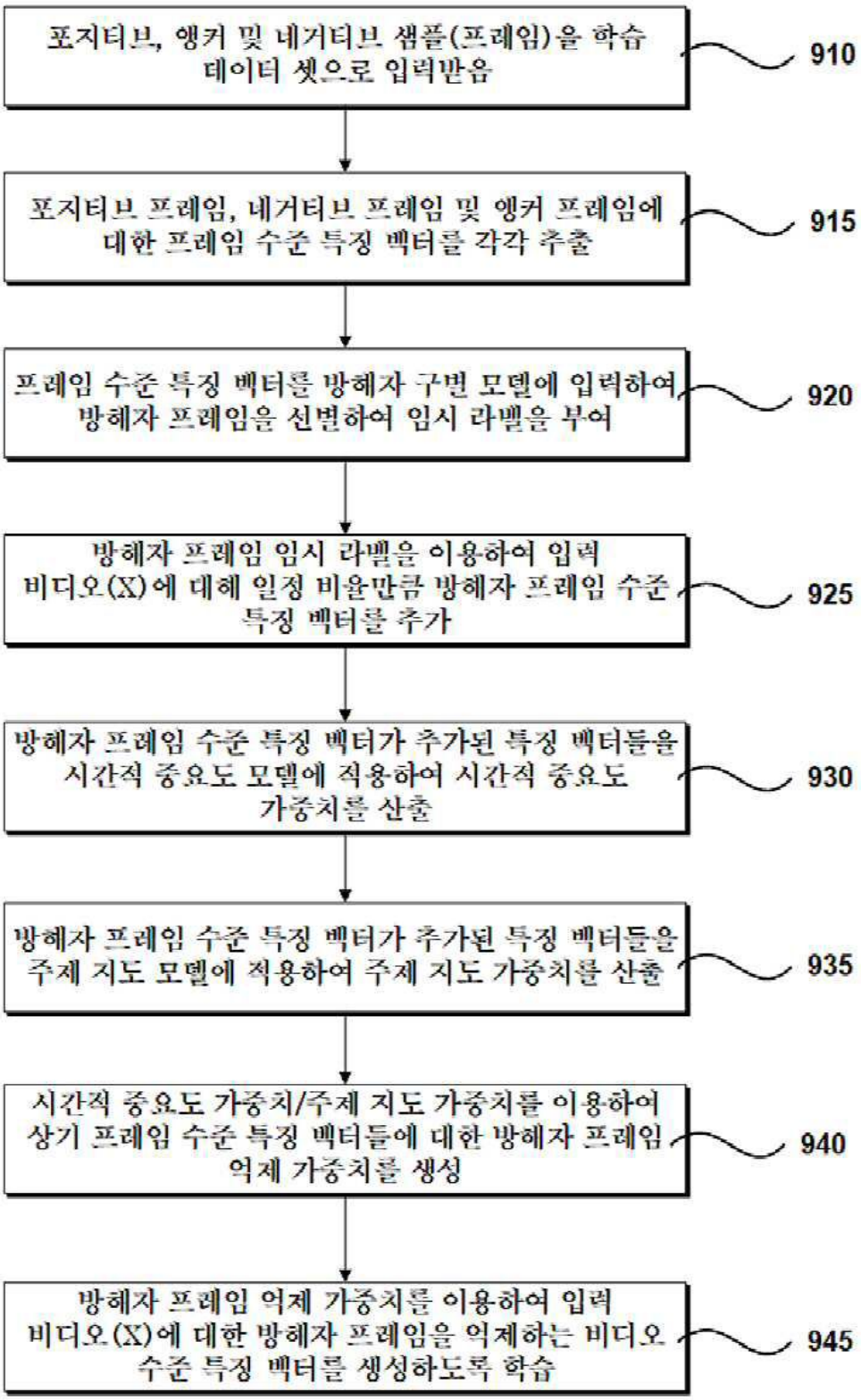
도면7



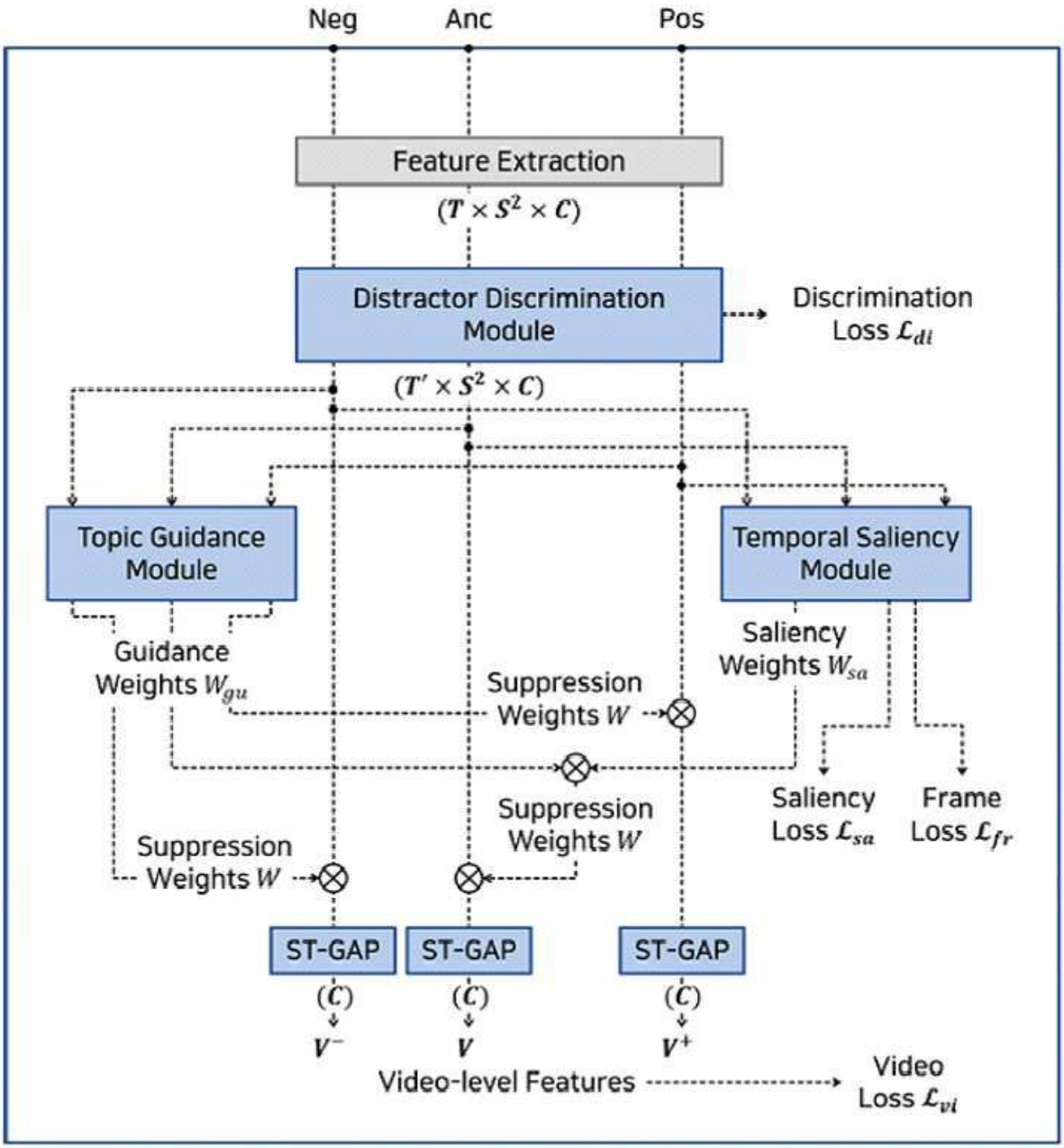
도면8

	Elim.	*Gen.*		FIVR-5K		
	DDM	TSM	TGM	DSVR	CSVR	ISVR
(a)				0.692	0.700	0.651
(b)	✓			0.715	0.725	0.672
(c)		✓		0.702	0.710	0.661
(d)			✓	0.716	0.724	0.677
(e)		✓	✓	0.719	0.726	0.680
(f)	✓	✓		0.724	0.732	0.683
(g)	✓		✓	0.738	0.746	0.698
(h)	✓	✓	✓	**0.744**	**0.752**	**0.705**

도면9

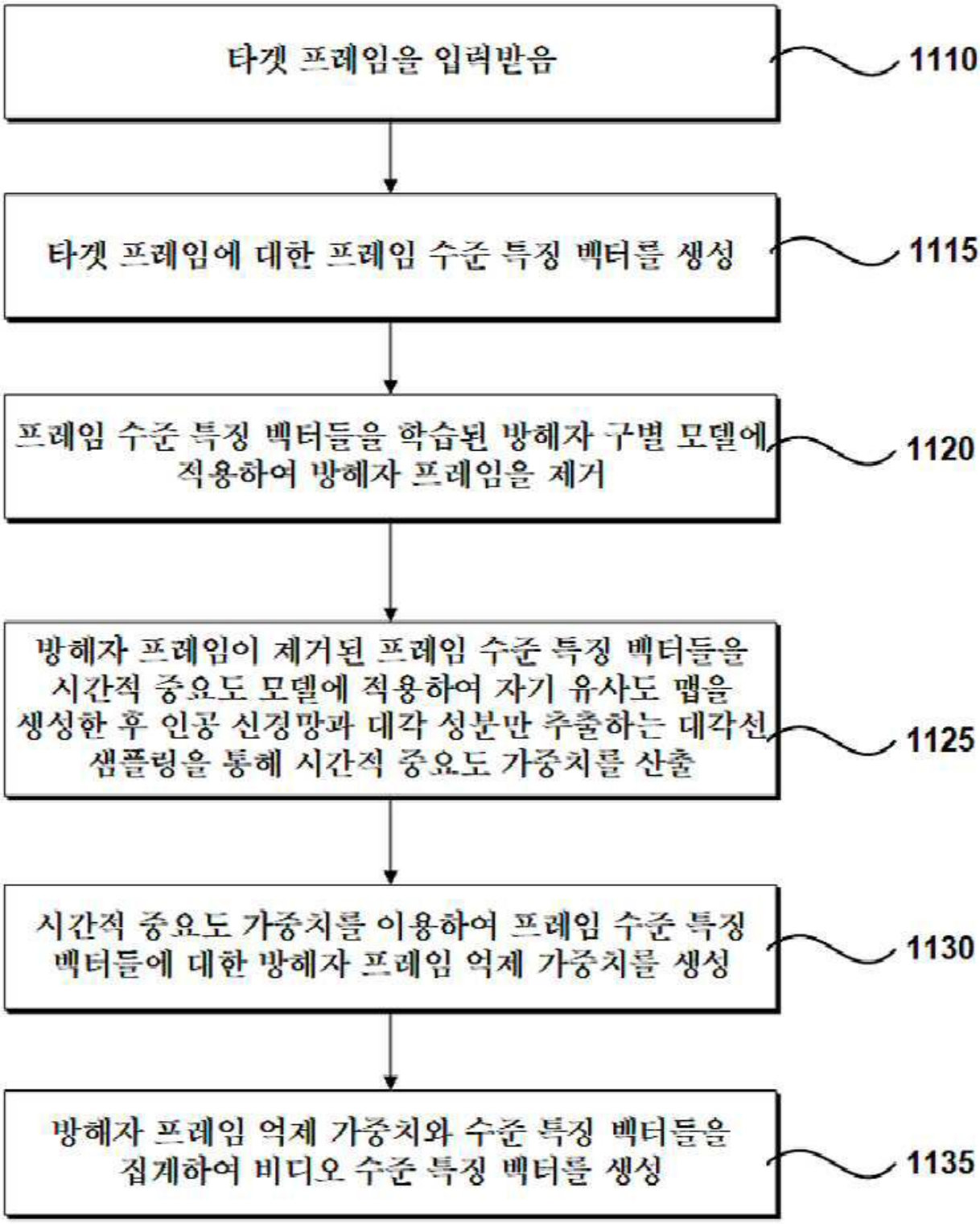

포지티브, 앵커 및 네거티브 샘플(프레임)을 학습 데이터 셋으로 입력받음
910
포지티브 프레임, 네거티브 프레임 및 앵커 프레임에 대한 프레임 수준 특징 벡터를 각각 추출
915
프레임 수준 특징 벡터를 방해자 구별 모델에 입력하여 방해자 프레임을 선별하여 임시 라벨을 부여
920
방해자 프레임 임시 라벨을 이용하여 입력 비디오(X)에 대해 일정 비율만큼 방해자 프레임 수준 특징 벡터를 추가
925
방해자 프레임 수준 특징 벡터가 추가된 특징 벡터들을 시간적 중요도 모델에 적용하여 시간적 중요도 가중치를 산출
930
방해자 프레임 수준 특징 벡터가 추가된 특징 벡터들을 주제 지도 모델에 적용하여 주제 지도 가중치를 산출
935
시간적 중요도 가중치/주제 지도 가중치를 이용하여 상기 프레임 수준 특징 벡터들에 대한 방해자 프레임 억제 가중치를 생성
940
방해자 프레임 억제 가중치를 이용하여 입력 비디오(X)에 대한 방해자 프레임을 억제하는 비디오 수준 특징 벡터를 생성하도록 학습
945

도면10



(a) Training

도면11


타겟 프레임을 입력받음
1110
타겟 프레임에 대한 프레임 수준 특징 벡터를 생성
1115
프레임 수준 특징 벡터들을 학습된 방해자 구별 모델에 적용하여 방해자 프레임을 제거
1120
방해자 프레임이 제거된 프레임 수준 특징 벡터들을 시간적 중요도 모델에 적용하여 자기 유사도 맵을 생성한 후 인공 신경망과 대각 성분만 추출하는 대각선 샘플링을 통해 시간적 중요도 가중치를 산출
1125
시간적 중요도 가중치를 이용하여 프레임 수준 특징 벡터들에 대한 방해자 프레임 억제 가중치를 생성
1130
방해자 프레임 억제 가중치와 수준 특징 벡터들을 집계하여 비디오 수준 특징 벡터를 생성
1135