(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2023년05월17일
(11) 등록번호 10-2533041
(24) 등록일자 2023년05월11일

(51) 국제특허분류(Int. Cl.)
G06V 20/52 (2022.01) *G06N 3/04* (2023.01)
G06N 3/08 (2023.01) *G06V 10/74* (2022.01)
G06V 10/774 (2022.01) *G06V 10/82* (2022.01)
(52) CPC특허분류
G06V 20/52 (2023.01)
G06N 3/045 (2023.01)
(21) 출원번호 10-2022-0130516
(22) 출원일자 2022년10월12일
심사청구일자 2022년10월12일
(56) 선행기술조사문헌
US20210158048 A1*
Kaiming He 외 5명, "Masked Autoencoders Are Scalable Vision Learners", arXiv:2111.06377, pp.1-14 (2021.11.11.) 1부.*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
국방과학연구소
대전광역시 유성구 북유성대로488번길 160 (수남동)
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
이주연
서울 성북구 안암로 145 고려대학교 우정정보관 408호
남우정
서울 성북구 안암로 145 고려대학교 우정정보관 408호
이성환
서울 성북구 안암로 145 고려대학교 우정정보관 408호
(74) 대리인
특허법인 광장리앤고

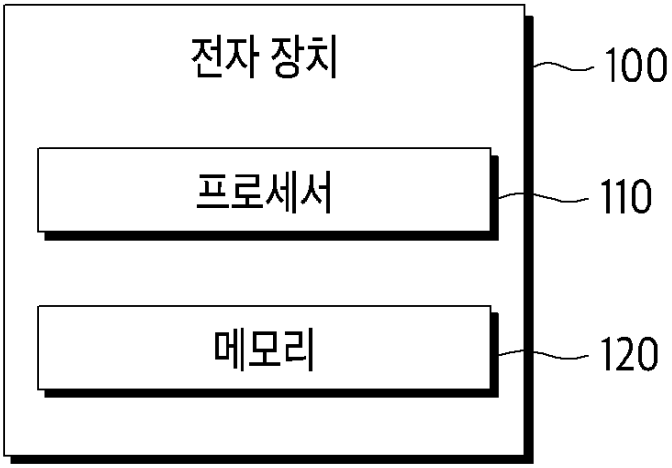
전체 청구항 수 : 총 9 항 심사관 : 정수진

(54) 발명의 명칭 **비디오 이상치 탐지를 위한 전자 장치 및 그의 동작 방법**

(57) 요 약

본 개시의 비디오 이상치 탐지를 위한 전자 장치의 동작 방법은, 비디오의 연속된 프레임들 중에서 적어도 하나의 프레임에 대해 마스킹을 수행하여 적어도 하나의 마스킹되지 않은 프레임에 관한 제1정보 및 적어도 하나의 마스킹된 프레임에 관한 제2정보를 확인하는 단계; 제1정보에 대해 인코딩을 수행하여 인코딩된 제1정보를 확인하는 단계; 제1모델을 이용하여 인코딩된 제1정보, 및 제2정보에 대응되는 프레임들에 관한 제1예측 정보를 확인하고, 제2모델을 이용하여 제1예측 정보에 대응되는 프레임들에 후행하는 후행 프레임에 관한 제2예측 정보를 확인하고, 제3모델을 이용하여 인코딩된 제1정보에 대응되는 후행 프레임에 관한 제3예측 정보를 확인하는 단계; 및 제1예측 정보 중 적어도 하나의 마스킹된 프레임에 관한 예측 정보, 제2예측 정보, 및 제3예측 정보를 기초로 비디오의 이상치를 판단하는 단계를 포함한다.

대 표 도 - 도1



(52) CPC특허분류
G06N 3/08 (2023.01)
G06V 10/74 (2022.01)
G06V 10/774 (2023.01)
G06V 10/82 (2022.01)

공지예외적용 : 있음

명 세 서

청구범위

청구항 1

비디오 이상치 탐지를 위한 전자 장치의 동작 방법으로서,

비디오의 연속된 프레임들 중에서 적어도 하나의 프레임에 대해 마스킹을 수행하여 적어도 하나의 마스킹되지 않은 프레임에 관한 제1정보 및 적어도 하나의 마스킹된 프레임에 관한 제2정보를 확인하는 단계;

상기 제1정보에 대해 인코딩을 수행하여 인코딩된 제1정보를 확인하는 단계;

제1모델을 이용하여 상기 인코딩된 제1정보 및 상기 제2정보에 대응되는 상기 프레임들에 관한 제1예측 정보를 확인하고, 제2모델을 이용하여 상기 제1예측 정보에 대응되는 상기 프레임들에 후행하는 후행 프레임에 관한 제2예측 정보를 확인하고, 제3모델을 이용하여 상기 인코딩된 제1정보에 대응되는 상기 후행 프레임에 관한 제3예측 정보를 확인하는 단계; 및

상기 제1예측 정보 중 적어도 하나의 마스킹된 프레임에 관한 예측 정보, 상기 제2예측 정보, 및 상기 제3예측 정보를 기초로 상기 비디오의 이상치를 판단하는 단계를 포함하되,

상기 제1모델은,

상기 인코딩된 제1정보 및 상기 제2정보를 입력 정보로서 이용하여 적어도 하나의 마스킹된 프레임에 관한 정보 및 적어도 하나의 마스킹되지 않은 프레임에 관한 정보를 복원하는 비전 트랜스포머 디코더를 포함하고,

상기 제2모델은,

선형 함수를 통해 상기 제1예측 정보로부터 상기 제2예측 정보를 생성하는 선형 함수 레이어를 포함하고,

상기 제3모델은,

상기 인코딩된 제1정보를 입력 정보로서 이용하여 적어도 하나의 마스킹되지 않은 프레임에 관한 정보를 복원하는 비전 트랜스포머 디코더 및 선형 함수를 통해 상기 복원된 적어도 하나의 마스킹되지 않은 프레임에 관한 정보로부터 상기 제3예측 정보를 생성하는 선형 함수 레이어를 포함하는, 동작 방법.

청구항 2

제1항에 있어서,

상기 인코딩된 제1정보를 확인하는 단계는,

비전 트랜스포머(Vision Transformer) 인코더를 이용하여 상기 제1정보에 대해 인코딩을 수행하는 단계를 포함하는, 동작 방법.

청구항 3

삭제

청구항 4

비디오 이상치 탐지를 위한 전자 장치의 동작 방법으로서,

비디오의 연속된 프레임들 중에서 적어도 하나의 프레임에 대해 마스킹을 수행하여 적어도 하나의 마스킹되지 않은 프레임에 관한 제1정보 및 적어도 하나의 마스킹된 프레임에 관한 제2정보를 확인하는 단계;

상기 제1정보에 대해 인코딩을 수행하여 인코딩된 제1정보를 확인하는 단계;

제1모델을 이용하여 상기 인코딩된 제1정보 및 상기 제2정보에 대응되는 상기 프레임들에 관한 제1예측 정보를 확인하고, 제2모델을 이용하여 상기 제1예측 정보에 대응되는 상기 프레임들에 후행하는 후행 프레임에 관한 제2예측 정보를 확인하고, 제3모델을 이용하여 상기 인코딩된 제1정보에 대응되는 상기 후행 프레임에 관한 제3예

측 정보를 확인하는 단계; 및

상기 제1예측 정보 중 적어도 하나의 마스킹된 프레임에 관한 예측 정보, 상기 제2예측 정보, 및 상기 제3예측 정보를 기초로 상기 비디오의 이상치를 판단하는 단계를 포함하되,

상기 판단하는 단계는,

상기 적어도 하나의 마스킹된 프레임에 관한 예측 정보와 상기 적어도 하나의 마스킹된 프레임에 관한 실측 정보간의 제1유클리드 거리(Euclidean Distance), 상기 제2예측 정보와 상기 후행 프레임에 관한 실측 정보간의 제2유클리드 거리, 및 상기 제3예측 정보와 상기 후행 프레임에 관한 실측 정보간의 제3유클리드 거리를 계산하는 단계; 및

제1가중치가 부여된 제1유클리드 거리, 제2가중치가 부여된 제2유클리드 거리, 및 제3가중치가 부여된 제3유클리드 거리로부터 비디오 이상치 탐지를 위한 비디오 이상치 점수를 결정하고, 상기 비디오 이상치 점수를 소정의 임계 값과 비교하여 상기 비디오의 이상치를 판단하는 단계를 포함하는, 동작 방법.

청구항 5

비디오 이상치 탐지를 위한 전자 장치의 동작 방법으로서,

비디오의 연속된 프레임들 중에서 적어도 하나의 프레임에 대해 마스킹을 수행하여 적어도 하나의 마스킹되지 않은 프레임에 관한 제1정보 및 적어도 하나의 마스킹된 프레임에 관한 제2정보를 확인하는 단계;

상기 제1정보에 대해 인코딩을 수행하여 인코딩된 제1정보를 확인하는 단계;

제1모델을 이용하여 상기 인코딩된 제1정보 및 상기 제2정보에 대응되는 상기 프레임들에 관한 제1예측 정보를 확인하고, 제2모델을 이용하여 상기 제1예측 정보에 대응되는 상기 프레임들에 후행하는 후행 프레임에 관한 제2예측 정보를 확인하고, 제3모델을 이용하여 상기 인코딩된 제1정보에 대응되는 상기 후행 프레임에 관한 제3예측 정보를 확인하는 단계; 및

상기 제1예측 정보 중 적어도 하나의 마스킹된 프레임에 관한 예측 정보, 상기 제2예측 정보, 및 상기 제3예측 정보를 기초로 상기 비디오의 이상치를 판단하는 단계를 포함하되,

상기 비디오의 연속된 프레임들 중 마지막 프레임 내 객체의 동작 패턴에 관한 정보를 확인하는 단계; 및

제4모델을 이용하여 상기 동작 패턴에 관한 정보에 대응되는 상기 객체의 동작 패턴을 재구성한 재구성 정보를 확인하는 단계를 더 포함하고,

상기 판단하는 단계는,

상기 제1예측 정보 중 적어도 하나의 마스킹된 프레임에 관한 예측 정보, 상기 제2예측 정보, 상기 제3예측 정보, 및 상기 재구성 정보를 기초로 상기 비디오의 이상치를 판단하는 단계를 포함하는, 동작 방법.

청구항 6

제5항에 있어서,

상기 제4모델은,

상기 동작 패턴에 관한 정보를 입력 정보로서 이용하여 상기 객체의 동작 패턴을 재구성한 재구성 정보를 생성하는 컨볼루션 오토인코더(Convolutional Autoencoder)를 포함하는, 동작 방법.

청구항 7

제5항에 있어서,

상기 판단하는 단계는,

상기 적어도 하나의 마스킹된 프레임에 관한 예측 정보와 상기 적어도 하나의 마스킹된 프레임에 관한 실측 정보간의 유클리드 거리(Euclidean Distance), 상기 제2예측 정보와 상기 후행 프레임에 관한 실측 정보간의 유클리드 거리, 및 상기 제3예측 정보와 상기 후행 프레임에 관한 실측 정보간의 유클리드 거리의 합인 예측 유클리드 거리를 계산하고, 상기 재구성 정보와 상기 동작 패턴에 관한 정보간의 유클리드 거리인 재구성 유클리드 거

리를 계산하는 단계; 및

제1가중치가 부여된 예측 유클리드 거리 및 제2가중치가 부여된 재구성 유클리드 거리로부터 비디오 이상치 탐지를 위한 비디오 이상치 점수를 결정하고, 상기 비디오 이상치 점수를 소정의 임계 값과 비교하여 상기 비디오의 이상치를 판단하는 단계를 포함하는, 동작 방법.

청구항 8

제1항에 있어서,

통신 디바이스를 통해 비디오 이상치 판단 결과를 외부 장치로 전송하거나, 디스플레이를 통해 상기 비디오 이상치 판단 결과를 출력하는 단계를 더 포함하는, 동작 방법.

청구항 9

비디오 이상치 탐지를 위한 전자 장치로서,

적어도 하나의 프로그램이 저장된 메모리; 및

상기 적어도 하나의 프로그램을 실행함으로써, 비디오의 연속된 프레임들 중에서 적어도 하나의 프레임에 대해 마스킹을 수행하여 적어도 하나의 마스킹되지 않은 프레임에 관한 제1정보 및 적어도 하나의 마스킹된 프레임에 관한 제2정보를 확인하고,

상기 제1정보에 대해 인코딩을 수행하여 인코딩된 제1정보를 확인하고,

제1모델을 이용하여 상기 인코딩된 제1정보, 및 상기 제2정보에 대응되는 상기 프레임들에 관한 제1예측 정보를 확인하고, 제2모델을 이용하여 상기 제1예측 정보에 대응되는 상기 프레임들에 후행하는 후행 프레임에 관한 제2예측 정보를 확인하고, 제3모델을 이용하여 상기 인코딩된 제1정보에 대응되는 상기 후행 프레임에 관한 제3예측 정보를 확인하고, 및

상기 제1예측 정보 중 적어도 하나의 마스킹된 프레임에 관한 예측 정보, 상기 제2예측 정보, 및 상기 제3예측 정보를 기초로 상기 비디오의 이상치를 판단하는 프로세서를 포함하되,

상기 제1모델은,

상기 인코딩된 제1정보 및 상기 제2정보를 입력 정보로서 이용하여 적어도 하나의 마스킹된 프레임에 관한 정보 및 적어도 하나의 마스킹되지 않은 프레임에 관한 정보를 복원하는 비전 트랜스포머 디코더를 포함하고,

상기 제2모델은,

선형 함수를 통해 상기 제1예측 정보로부터 상기 제2예측 정보를 생성하는 선형 함수 레이어를 포함하고,

상기 제3모델은,

상기 인코딩된 제1정보를 입력 정보로서 이용하여 적어도 하나의 마스킹되지 않은 프레임에 관한 정보를 복원하는 비전 트랜스포머 디코더 및 선형 함수를 통해 상기 복원된 적어도 하나의 마스킹되지 않은 프레임에 관한 정보로부터 상기 제3예측 정보를 생성하는 선형 함수 레이어를 포함하는, 전자 장치.

청구항 10

비디오 이상치 탐지를 위한 전자 장치의 동작 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 비일시적 기록매체로서,

상기 동작 방법은,

비디오의 연속된 프레임들 중에서 적어도 하나의 프레임에 대해 마스킹을 수행하여 적어도 하나의 마스킹되지 않은 프레임에 관한 제1정보 및 적어도 하나의 마스킹된 프레임에 관한 제2정보를 확인하는 단계;

상기 제1정보에 대해 인코딩을 수행하여 인코딩된 제1정보를 확인하는 단계;

제1모델을 이용하여 상기 인코딩된 제1정보, 및 상기 제2정보에 대응되는 상기 프레임들에 관한 제1예측 정보를 확인하고, 제2모델을 이용하여 상기 제1예측 정보에 대응되는 상기 프레임들에 후행하는 후행 프레임에 관한 제2예측 정보를 확인하고, 제3모델을 이용하여 상기 인코딩된 제1정보에 대응되는 상기 후행 프레임에 관한 제3예

측 정보를 확인하는 단계; 및

상기 제1예측 정보 중 적어도 하나의 마스킹된 프레임에 관한 예측 정보, 상기 제2예측 정보, 및 상기 제3예측 정보를 기초로 상기 비디오의 이상치를 판단하는 단계를 포함하되,

상기 제1모델은,

상기 인코딩된 제1정보 및 상기 제2정보를 입력 정보로서 이용하여 적어도 하나의 마스킹된 프레임에 관한 정보 및 적어도 하나의 마스킹되지 않은 프레임에 관한 정보를 복원하는 비전 트랜스포머 디코더를 포함하고,

상기 제2모델은,

선형 함수를 통해 상기 제1예측 정보로부터 상기 제2예측 정보를 생성하는 선형 함수 레이어를 포함하고,

상기 제3모델은,

상기 인코딩된 제1정보를 입력 정보로서 이용하여 적어도 하나의 마스킹되지 않은 프레임에 관한 정보를 복원하는 비전 트랜스포머 디코더 및 선형 함수를 통해 상기 복원된 적어도 하나의 마스킹되지 않은 프레임에 관한 정보로부터 상기 제3예측 정보를 생성하는 선형 함수 레이어를 포함하는, 비일시적 기록매체.

발명의 설명

기 술 분 야

[0001] 본 개시는 비디오 이상치 탐지를 위한 전자 장치 및 그의 동작 방법에 관한 것이다.

배 경 기 술

[0002] 비디오 이상치 탐지(Video Anomaly Detection) 기술은 현재 줄어드는 인력 자원을 고려할 때 미래 전장을 대비하기 위한 주요 기술 중 하나이다. 미래 전장에서 군 경계 지역에 인력의 적소 배치는 전장의 승패를 가르는 중요한 요소이고, 비디오 이상치 탐지 기술을 통해 최소한의 인력 배치만으로 적의 공격수단 또는 적군의 이상 행동을 정확하게 탐지하는 능력이 필수적으로 요구된다. 종래의 딥러닝 기반의 비디오 이상치 탐지 모델은 정상 샘플만을 사용하여 모델을 학습시키므로, 모델에 비정상 샘플이 입력될 때에도 모델의 출력이 정상 샘플과 큰 차이가 없어 이상치 탐지가 어려워지는 소위 '생성 문제'가 발생할 수 있다. 또한, 최근 비디오 이상치 탐지 분야에서 적용되는 어텐션 기반의 트랜스포머 모델인 비전 트랜스포머는 딥러닝 방식과 달리 이미지의 지역성을 학습할 수 없어 모델의 성능이 데이터 풍부도에 큰 의존성을 보이는 소위 '유도 편향 문제'가 발생할 수 있다. 이에 '생성 문제'와 '유도 편향 문제'에 덜 민감하면서 높은 이상치 탐지 성능을 갖는 비디오 이상치 탐지 기술이 요구된다.

발명의 내용

해결하려는 과제

[0003] 개시된 실시예들은 비디오 이상치 탐지를 위한 전자 장치 및 그의 동작 방법을 개시하고자 한다. 본 실시예가 이루고자 하는 기술적 과제는 상기된 바와 같은 기술적 과제들로 한정되지 않으며, 이하의 실시예들로부터 또 다른 기술적 과제들이 유추될 수 있다.

과제의 해결 수단

[0004] 일 실시예에 따른 비디오 이상치 탐지를 위한 전자 장치의 동작 방법으로서, 비디오의 연속된 프레임들 중에서 적어도 하나의 프레임에 대해 마스킹을 수행하여 적어도 하나의 마스킹되지 않은 프레임에 관한 제1정보 및 적어도 하나의 마스킹된 프레임에 관한 제2정보를 확인하는 단계; 제1정보에 대해 인코딩을 수행하여 인코딩된 제1정보를 확인하는 단계; 제1모델을 이용하여 인코딩된 제1정보, 및 제2정보에 대응되는 프레임들에 관한 제1예측 정보를 확인하고, 제2모델을 이용하여 제1예측 정보에 대응되는 프레임들에 후행하는 후행 프레임에 관한 제2예측 정보를 확인하고, 제3모델을 이용하여 인코딩된 제1정보에 대응되는 후행 프레임에 관한 제3예측 정보를 확인하는 단계; 및 제1예측 정보 중 적어도 하나의 마스킹된 프레임에 관한 예측 정보, 제2예측 정보, 및 제3예측 정보를 기초로 비디오의 이상치를 판단하는 단계를 포함하는 방법이 제공될 수 있다.

[0005] 일 실시예에 따른 비디오 이상치 탐지를 위한 전자 장치의 동작 방법으로서, 인코딩된 제1정보를 확인하는 단계

는, 비전 트랜스포머(Vision Transformer) 인코더를 이용하여 제1정보에 대해 인코딩을 수행하는 단계를 더 포함할 수 있다.

[0006] 일 실시예에 따른 비디오 이상치 탐지를 위한 전자 장치의 동작 방법으로서, 제1모델은 인코딩된 제1정보 및 제2정보를 입력 정보로서 이용하여 적어도 하나의 마스킹된 프레임에 관한 정보 및 적어도 하나의 마스킹되지 않은 프레임에 관한 정보를 복원하는 비전 트랜스포머 디코더를 포함하고, 제2모델은, 선형 함수를 통해 제1예측 정보로부터 제2예측 정보를 생성하는 선형 함수 레이어를 포함하고, 제3모델은, 인코딩된 제1정보를 입력 정보로서 이용하여 적어도 하나의 마스킹되지 않은 프레임에 관한 정보를 복원하는 비전 트랜스포머 디코더 및 선형 함수를 통해 복원된 적어도 하나의 마스킹되지 않은 프레임에 관한 정보로부터 제3예측 정보를 생성하는 선형 함수 레이어를 포함할 수 있다.

[0007] 일 실시예에 따른 비디오 이상치 탐지를 위한 전자 장치의 동작 방법으로서, 판단하는 단계는, 적어도 하나의 마스킹된 프레임에 관한 예측 정보와 적어도 하나의 마스킹된 프레임에 관한 실측 정보간의 제1유클리드 거리(Euclidean Distance), 제2예측 정보와 후행 프레임에 관한 실측 정보간의 제2유클리드 거리, 및 제3예측 정보와 후행 프레임에 관한 실측 정보간의 제3유클리드 거리를 계산하는 단계; 및 제1가중치가 부여된 제1유클리드 거리, 제2가중치가 부여된 제2유클리드 거리, 및 제3가중치가 부여된 제3유클리드 거리로부터 비디오 이상치 탐지를 위한 비디오 이상치 점수를 결정하고, 상기 비디오 이상치 점수를 소정의 임계 값과 비교하여 상기 비디오의 이상치를 판단하는 단계를 포함할 수 있다.

[0008] 일 실시예에 따른 비디오 이상치 탐지를 위한 전자 장치의 동작 방법으로서, 비디오의 연속된 프레임들 중 마지막 프레임 내 객체의 동작 패턴에 관한 정보를 확인하는 단계; 및 제4모델을 이용하여 동작 패턴에 관한 정보에 대응되는 객체의 동작 패턴을 재구성한 재구성 정보를 확인하는 단계를 더 포함하고, 판단하는 단계는, 제1예측 정보 중 적어도 하나의 마스킹된 프레임에 관한 예측 정보, 제2예측 정보, 제3예측 정보, 및 재구성 정보를 기초로 비디오의 이상치를 판단하는 단계를 포함할 수 있다.

[0009] 일 실시예에 따른 비디오 이상치 탐지를 위한 전자 장치의 동작 방법으로서, 제4모델은, 동작 패턴에 관한 정보를 입력 정보로서 이용하여 객체의 동작 패턴을 재구성한 재구성 정보를 생성하는 컨볼루션 오토인코더(Convolutional Autoencoder)를 포함할 수 있다.

[0010] 일 실시예에 따른 비디오 이상치 탐지를 위한 전자 장치의 동작 방법으로서, 판단하는 단계는, 적어도 하나의 마스킹된 프레임에 관한 예측 정보와 적어도 하나의 마스킹된 프레임에 관한 실측 정보간의 유클리드 거리(Euclidean Distance), 제2예측 정보와 비디오의 연속된 프레임들에 후행하는 후행 프레임에 관한 실측 정보간의 유클리드 거리, 및 제3예측 정보와 비디오의 연속된 프레임들에 후행하는 후행 프레임에 관한 실측 정보간의 유클리드 거리의 합인 예측 유클리드 거리를 계산하고, 재구성 정보와 동작 패턴에 관한 정보간의 유클리드 거리인 재구성 유클리드 거리를 계산하는 단계; 및 제1가중치가 부여된 예측 유클리드 거리 및 제2가중치가 부여된 재구성 유클리드 거리로부터 비디오 이상치 탐지를 위한 비디오 이상치 점수를 결정하고, 비디오 이상치 점수를 소정의 임계 값과 비교하여 비디오의 이상치를 판단하는 단계를 포함할 수 있다.

[0011] 일 실시예에 따른 비디오 이상치 탐지를 위한 전자 장치의 동작 방법으로서, 통신 디바이스를 통해 비디오 이상치 판단 결과를 외부 장치로 전송하거나, 디스플레이를 통해 비디오 이상치 판단 결과를 출력하는 단계를 포함할 수 있다.

[0012] 일 실시예에 따른 비디오 이상치 탐지를 위한 전자 장치로서, 적어도 하나의 프로그램이 저장된 메모리; 및 적어도 하나의 프로그램을 실행함으로써, 비디오의 연속된 프레임들 중에서 적어도 하나의 프레임에 대해 마스킹을 수행하여 적어도 하나의 마스킹되지 않은 프레임에 관한 제1정보 및 적어도 하나의 마스킹된 프레임에 관한 제2정보를 확인하고, 제1정보에 대해 인코딩을 수행하여 인코딩된 제1정보를 확인하고, 제1모델을 이용하여 인코딩된 제1정보, 및 제2정보에 대응되는 프레임에 관한 제1예측 정보를 확인하고, 제2모델을 이용하여 프레임들에 후행하는 후행 프레임에 관한 제2예측 정보를 확인하고, 제3모델을 이용하여 인코딩된 제1정보에 대응되는 후행 프레임에 관한 제3예측 정보를 확인하고, 및 제1예측 정보 중 적어도 하나의 마스킹된 프레임에 관한 예측 정보, 제2예측 정보, 및 제3예측 정보를 기초로 비디오의 이상치를 판단하는 프로세서를 포함하는 전자 장치가 제공될 수 있다.

[0013] 일 실시예에 따른 비디오 이상치 탐지를 위한 전자 장치의 동작 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 비일시적 기록매체로서, 방법은, 비디오의 연속된 프레임들 중에서 적어도 하나의 프레임에 대해 마스킹을 수행하여 적어도 하나의 마스킹되지 않은 프레임에 관한 제1정보 및 적어도 하나의

마스킹된 프레임에 관한 제2정보를 확인하는 단계; 제1정보에 대해 인코딩을 수행하여 인코딩된 제1정보를 확인하는 단계; 제1모델을 이용하여 인코딩된 제1정보, 및 제2정보에 대응되는 프레임들에 관한 제1예측 정보를 확인하고, 제2모델을 이용하여 제1예측 정보에 대응되는 프레임들에 후행하는 후행 프레임에 관한 제2예측 정보를 확인하고, 제3모델을 이용하여 인코딩된 제1정보에 대응되는 후행 프레임에 관한 제3예측 정보를 확인하는 단계; 및 제1예측 정보 중 적어도 하나의 마스킹된 프레임에 관한 예측 정보, 제2예측 정보, 및 제3예측 정보를 기초로 비디오의 이상치를 판단하는 단계를 포함하는, 비일시적 기록매체가 제공될 수 있다.

발명의 효과

[0014] 본 개시에 따르면 전자 장치는 마스킹되지 않은 프레임에 관한 제1정보 및 마스킹된 프레임에 관한 제2정보를 기초로 각각의 모델을 이용하여 각각의 예측 정보를 확인하고, 확인된 각각의 예측 정보를 기초로 비디오의 이상치를 판단할 수 있다. 전자 장치는 마스킹되지 않은 프레임에 관한 제1정보와 마스킹된 프레임에 관한 제2정보를 동시에 고려하므로 '생성 문제'를 회피할 수 있다. 또한, 전자 장치는 하나의 프레임을 하나의 패치로 이용하므로 '유도 편향 문제'를 회피할 수 있다. 전자 장치는 프레임에 관한 예측 정보(즉, 비디오의 시간적인 맥락을 파악하기 위한 정보)와 객체의 동작 패턴에 관한 정보(즉, 비디오의 공간적인 맥락을 파악하기 위한 정보)를 통해 비디오의 시공간적 맥락을 종합적으로 고려하여 비디오 이상치를 탐지하므로 보다 정확하게 비디오 이상치를 탐지할 수 있다. 국방 분야에서 전자 장치는 향상된 비디오 이상치 탐지 성능을 제공할 수 있다. 예를 들어, 전자 장치는 감시·정찰 임무를 수행할 때 영상에서 적군의 이상 행동(예를 들어, 적기 출현, 적 군수 물자의 비정상적 이동 등)을 보다 정확하게 탐지할 수 있다.

도면의 간단한 설명

[0015] 도 1은 본 개시에 따른 전자 장치를 나타낸다.

도 2는 일 실시예에 따른 전자 장치의 개략도를 나타낸다.

도 3은 일 실시예에 따른 예측 모듈을 학습시키는 방법을 나타낸다.

도 4는 일 실시예에 따른 재구성 모듈을 학습시키는 방법을 나타낸다.

도 5는 일 실시예에 따른 전자 장치와 종래 기술들 간의 이상치 탐지 비교 결과를 나타낸다.

도 6은 일 실시예에 따른 전자 장치와 종래 기술 간의 이상치 탐지 비교 결과들을 나타낸다.

도 7은 일 실시예에 따른 전자 장치의 비디오 이상치 점수 그래프를 나타낸다.

도 8은 다른 실시예에 따른 전자 장치의 비디오 이상치 점수 그래프를 나타낸다.

도 9는 일 실시예에 따른 전자 장치의 동작 방법을 나타낸다.

발명을 실시하기 위한 구체적인 내용

[0016] 이하, 본 발명의 실시예를 첨부된 도면을 참조하여 상세하게 설명한다.

[0017] 실시예들에서 사용되는 용어는 본 개시에서의 기능을 고려하면서 가능한 현재 널리 사용되는 일반적인 용어들을 선택하였으나, 이는 당 분야에 종사하는 기술자의 의도 또는 판례, 새로운 기술의 출현 등에 따라 달라질 수 있다. 또한, 특정한 경우는 출원인이 임의로 선정한 용어도 있으며, 이 경우 해당되는 설명 부분에서 상세히 그 의미를 기재할 것이다. 따라서 본 개시에서 사용되는 용어는 단순한 용어의 명칭이 아닌, 그 용어가 가지는 의미와 본 개시의 전반에 걸친 내용을 토대로 정의되어야 한다.

[0018] 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 있음을 의미한다. 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.

[0019] 명세서 전체에서 기재된 “a, b, 및 c 중 적어도 하나” 의 표현은, ‘a 단독’, ‘b 단독’, ‘c 단독’, ‘a 및 b’, ‘a 및 c’, ‘b 및 c’, 또는 ‘a,b,c 모두’ 를 포괄할 수 있다.

[0020] 아래에서는 첨부한 도면을 참고하여 본 개시의 실시예에 대하여 본 개시가 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다. 그러나 본 개시는 여러 가지 상이한 형태로 구현될 수

있으며 여기에서 설명하는 실시예에 한정되지 않는다.

[0021] 종래의 딥러닝 기반 비디오 이상치 탐지 기술은 '생성 문제'를 내포하고, 트랜스포머 기반 비디오 이상치 탐지 기술은 '유도 편향 문제'를 내포한다. '생성 문제'를 해결하기 위해 메모리 모듈 기반 오토인코더(Autoencoder)를 이용하고자 하는 시도가 있었으나 오토인코더의 높은 재구성 성능으로 인해'생성 문제'를 완전히 해결할 수는 없었다. 마찬가지로, 종래의 트랜스포머(Transformer)는 트랜스포머를 학습시키기 위해서 방대한 양의 데이터가 요구되어, 비교적 데이터가 적은 비디오 이상치 탐지 분야에 적합하지 않았다.

[0022] 비전 트랜스포머(Vision Transformer)는 자연어 처리(Natural Language Processing)에서 사용되는 트랜스포머의 구조를 따르고, 순차적인 정보를 학습할 수 있다는 장점이 있다. 비전 트랜스포머는 하나의 이미지를 여러 패치들로 분할하고, 위치 정보와 더불어 패치들 각각을 선형 임베딩한다. 선형 임베딩된 패치들은 셀프-어텐션(self-attention) 모듈을 포함하는 트랜스포머 인코더로 입력된다. 셀프-어텐션 모듈은 현재 패치와 주어진 모든 패치 사이의 관계에 관한 정보(즉, 어텐션)를 획득할 수 있으므로, 현재 패치에 주어진 모든 패치의 정보를 고려하여 인코딩할 수 있다. 트랜스포머 인코더를 통해 인코딩된 패치는 분류 작업을 수행하는데 사용된다. 본 개시는 비전 트랜스포머와 메모리 모듈에 기반하지 않은 오토인코더를 이용하여'생성 문제'와 '유도 편향 문제'를 회피하고 비디오의 시공간적 맥락을 종합적으로 고려하여 향상된 이상치 탐지 성능을 갖는 비디오 이상치 탐지 방법을 제안하고자 한다.

[0024] 도 1은 본 개시에 따른 전자 장치를 나타낸다.

[0025] 전자 장치(100)는 프로세서(110) 및 메모리(120)를 포함한다. 도 1에 도시된 전자 장치(100)에는 본 실시예들과 관련된 구성요소들만이 도시되어 있다. 따라서, 전자 장치(100)에는 도 1에 도시된 구성요소들 외에 다른 범용적인 구성요소들이 더 포함될 수 있음은 당해 기술분야의 통상의 기술자에게 자명하다.

[0026] 전자 장치(100)는 비디오의 이상치를 탐지할 수 있다. 구체적으로, 전자 장치(100)는 비디오의 프레임들에 관한 예측 정보를 기초로 비디오의 이상치를 판단할 수 있다. 프레임에 관한 예측 정보는 비디오의 프레임들을 모델에 입력하여 출력으로 복원한 프레임을 의미할 수 있다. 예를 들어, 제1프레임에 관한 예측 정보는 비디오의 프레임들을 모델에 입력하여 출력으로 복원한 제1프레임을 의미할 수 있다.

[0027] 프로세서(110)는 전자 장치(100)의 전반적인 기능들을 제어하는 역할을 한다. 예를 들어, 프로세서(110)는 전자 장치(100) 내의 메모리(120)에 저장된 프로그램들을 실행함으로써, 전자 장치(100)를 전반적으로 제어한다. 프로세서(110)는 전자 장치(100) 내에 구비된 CPU(central processing unit), GPU(graphics processing unit), AP(application processor) 등으로 구현될 수 있으나, 이에 제한되지 않는다.

[0028] 메모리(120)는 전자 장치(100) 내에서 처리되는 각종 데이터들을 저장하는 하드웨어로서, 메모리(120)는 전자 장치(100)에서 처리된 데이터들 및 처리될 데이터들을 저장할 수 있다. 또한, 메모리(120)는 전자 장치(100)에 의해 구동될 애플리케이션들, 드라이버들 등을 저장할 수 있다. 메모리(120)는 DRAM(dynamic random access memory), SRAM(static random access memory) 등과 같은 RAM(random access memory), ROM(read-only memory), EEPROM(electrically erasable programmable read-only memory), CD-ROM, 블루레이 또는 다른 광학 디스크 스토리지, HDD(hard disk drive), SSD(solid state drive), 또는 플래시 메모리를 포함할 수 있다.

[0029] 일 실시예에 따라, 프로세서(110)는 비디오의 연속된 프레임들로부터 적어도 하나의 프레임에 관한 정보를 확인할 수 있다. 구체적으로, 프로세서(110)는 비디오의 연속된 프레임들 중에서 적어도 하나의 프레임에 대해 마스킹을 수행하여 적어도 하나의 마스킹되지 않은 프레임에 관한 제1정보 및 적어도 하나의 마스킹된 프레임에 관한 제2정보를 확인할 수 있다. 여기서 적어도 하나의 프레임에 관한 정보는 적어도 하나의 프레임의 토큰을 포함할 수 있다. 프레임은 패치 임베딩 프로세스를 통해 벡터 형태로서 토큰화될 수 있다. 즉, 토큰은 프레임으로부터 생성된 벡터를 의미할 수 있다. 예를 들어, 적어도 하나의 마스킹되지 않은 프레임에 관한 제1정보는 적어도 하나의 마스킹되지 않은 프레임의 토큰을 포함할 수 있고, 적어도 하나의 마스킹된 프레임에 관한 제2정보는 적어도 하나의 마스킹된 프레임의 토큰을 포함할 수 있다.

[0030] 일 실시예에 따라, 프로세서(110)는 프레임 내 객체의 동작 패턴에 관한 정보를 확인할 수 있다. 구체적으로, 프로세서(110)는 비디오의 연속된 프레임들 중 마지막 프레임 내 객체의 동작 패턴에 관한 정보를 확인할 수 있다. 여기서 객체의 동작 패턴에 관한 정보는 이전 프레임(t)과 다음 프레임(t+1) 간에 객체를 이루는 픽셀이 이동한 속력과 거리에 관한 정보를 포함할 수 있다. 예를 들어, 동작 패턴에 관한 정보는 제3프레임과 제4프레임 간에 픽셀이 이동한 속력과 거리의 벡터를 포함할 수 있다. 일 실시예에 따라, 프로세서(110)는 비디오의 연속된 프레임으로부터 프레임 내 객체의 동작 패턴에 관한 정보를 계산할 수 있다. 예를 들어, 프로세서(110)는

Lucas-Kanade 알고리즘, Horn-Schunck 알고리즘, 또는 Bruhn 알고리즘을 통해 프레임 내 객체의 동작 패턴에 관한 정보를 계산할 수 있다.

[0031] 일 실시예에 따라, 프로세서(110)는 비전 트랜스포머 인코더를 이용하여 제1정보에 대해 인코딩을 수행할 수 있다. 비전 트랜스포머 인코더는 제1정보를 입력 정보로서 이용하여 인코딩된 제1정보를 생성할 수 있고, 프로세서(110)는 비전 트랜스포머 인코더를 통해 생성된 인코딩된 제1정보를 확인할 수 있다. 프로세서(110)는 비전 트랜스포머 인코더의 입력 정보로서 마스킹되지 않은 프레임에 관한 정보인 제1정보만을 사용하므로, 제1정보와 제2정보 모두를 비전 트랜스포머 인코더의 입력 정보로 사용하는 경우보다 연산량이 감소하는 효과를 가질 수 있다.

[0032] 일 실시예에 따라, 프로세서(110)는 모델을 이용하여 모델의 입력 정보에 대응되는 프레임에 관한 예측 정보를 확인할 수 있다. 예를 들어, 비디오가 4개의 프레임(I_{t_1}, I_{t_2}, I_{t_3}, 및 I_{t_4})을 포함하는 경우, 비디오의 4개의 프레임들에 관한 제1예측 정보는 제1 모델을 이용하여 확인된 $\hat{I}_{t_1}'$, $\hat{I}_{t_2}'$, $\hat{I}_{t_3}'$, 및 $\hat{I}_{t_4}'$ 프레임을 의미하고, 후행 프레임(예를 들어, 제5프레임)에 관한 제2예측 정보는 제2모델을 이용하여 확인된 $\hat{I}_{t_5}$ 프레임을 의미하고, 후행 프레임에 관한 제3예측 정보는 제3모델을 이용하여 확인된 $\hat{I}_{t_5}$ 프레임을 의미할 수 있다. 예측 정보를 확인하는 과정은 이하 도 2에서 상세히 설명될 것이다.

[0034] 일 실시예에 따라, 프로세서(110)는 예측 정보를 기초로 비디오의 이상치를 탐지할 수 있다. 구체적으로 프로세서(110)는 제1예측 정보 중 적어도 하나의 마스킹된 프레임에 관한 예측 정보, 제2예측 정보, 및 제3예측 정보를 기초로 비디오의 이상치를 판단하거나, 프로세서(110)는 제1예측 정보 중 적어도 하나의 마스킹된 프레임에 관한 예측 정보, 제2예측 정보, 제3예측 정보, 및 재구성 정보를 기초로 비디오의 이상치를 판단할 수 있다. 일 실시예에 따라, 프로세서(110)는 예측 정보를 기초로 비디오 이상치 점수를 결정하고, 비디오 이상치 점수를 소정의 임계 값과 비교하여 비디오의 이상치를 판단할 수 있다. 다른 실시예에 따라, 프로세서(110)는 예측 정보 및 재구성 정보를 기초로 비디오 이상치 점수를 결정하고, 비디오 이상치 점수를 소정의 임계 값과 비교하여 비디오의 이상치를 판단할 수 있다. 비디오 이상치 판단 과정은 이하 도 2에서 상세히 설명될 것이다.

[0035] 일 실시예에 따라, 프로세서(110)는 통신 디바이스를 통해 비디오 이상치 판단 결과를 외부 장치로 전송하거나, 디스플레이를 통해 비디오 이상치 판단 결과를 출력할 수 있다.

[0037] 도 2는 일 실시예에 따른 전자 장치(200)의 개략도를 나타낸다.

[0038] 도 2의 전자 장치(200)는 도 1에 따른 전자 장치(100)의 각 구성요소를 포함할 수 있다. 도 2를 참조하면 전자 장치(200)는 예측 모듈(210)과 재구성 모듈(220)을 포함할 수 있다. 예측 모듈(210)은 훈련용 비디오로부터 객체 위주로 추출된 프레임들로 이루어진 정상 샘플들(즉, 이상 데이터로 판단되지 않은 샘플들)로 훈련될 수 있다. 또한, 재구성 모듈(220)도 객체 위주로 추출된 연속적인 프레임들로 이루어진 정상 샘플들로 훈련될 수 있다. 일 실시예에 따라, 예측 모듈(210)은 한 개의 비전 트랜스포머 인코더와 두 개의 비전 트랜스포머 디코더(제1디코더 및 제2디코더)를 포함할 수 있고, 재구성 모듈(220)은 한 개의 컨볼루션 오토인코더를 포함할 수 있다.

[0039] 일 실시예에 따라, 전자 장치(200)는 적어도 하나의 마스킹되지 않은 프레임에 관한 제1정보 및 적어도 하나의 마스킹된 프레임에 관한 제2정보를 확인할 수 있다. 구체적으로, 전자 장치(200)는 비디오의 연속된 프레임들 중에서 적어도 하나의 프레임에 대해 마스킹을 수행하여 제1정보 및 제2정보를 확인할 수 있다. 도 2를 참조하면, I_{t_1}~I_{t_4} 각각은 비디오의 연속된 4개의 프레임(순서대로 제1프레임, 제2프레임, 제3프레임, 및 제4프레임)을 의미한다. 전자 장치(200)는 I_{t_1}~I_{t_4} 프레임에 대해 기 설정된 마스킹 비율에 따라 마스킹을 수행할 수 있다. 예를 들어 전자 장치(200)는 I_{t_1}과 I_{t_4}에 기 설정된 마스킹 비율에 따라 마스킹을 수행할 수 있다. 도 2에서는 I_{t_1} 과 I_{t_4}프레임에 대해서만 마스킹을 수행한 것으로 가정한다. 전자 장치(200)는 각각의 마스킹된 프레임과 마스킹 되지 않은 프레임을 플래튼(Flatten) 레이어와 선형(Linear) 임베딩 레이어를 이용하여 적어도 하나의 마스킹되지 않은 토큰과 적어도 하나의 마스킹된 토큰을 생성할 수 있다. 플래튼 레이어는 프레임의 사이즈를 축소하고, 선형 임베딩 레이어는 프레임의 특징을 추출할 수 있다. 예를 들어, 전자 장치(200)는

3x32x32 사이즈를 갖는 마스킹된 프레임(즉, 마스킹된 I_{t_1} 및 I_{t_4})을 플래튼 레이어를 거쳐 1x1024 사이즈로 축소하고 선형 임베딩 레이어를 거쳐 1x1000 사이즈를 갖는 마스킹되지 않은 토큰을 생성할 수 있다. 다른 예시로서, 전자 장치(200)는 3x32x32 사이즈를 갖는 마스킹되지 않은 프레임(즉, I_{t_2} 및 I_{t_3})을 플래튼 레이어 및 선형 임베딩 레이어를 거쳐 1x1000 사이즈를 갖는 마스킹된 토큰을 생성할 수 있다. 전자 장치(200)는 적어도 하나의 마스킹되지 않은 토큰을 제1정보로서 확인하고, 적어도 하나의 마스킹된 토큰을 제2정보로서 확인할 수 있다. 도 2에 기재된 프레임과 토큰의 사이즈는 일 예시일 뿐 이에 제한되지 않는다.

[0040] 일 실시예에 따라, 예측 모듈(210)은 제1정보에 대해 인코딩을 수행하여 인코딩된 제1정보를 확인할 수 있다. 다시 말해, 예측 모듈(210)은 제1토큰에 대해 인코딩을 수행하여 인코딩된 제1토큰을 확인할 수 있다. 예측 모듈(210)은 비전 트랜스포머 인코더를 이용하여 제1토큰에 대해 인코딩을 수행할 수 있다. 예측 모듈(210)은 비전 트랜스포머 인코더로부터 생성된 인코딩된 제1토큰을 확인할 수 있다. 인코딩된 제1토큰은 아래의 수학식1로 표현할 수 있다.

수학식 1

[0041] $$LN(E(q_{i\in M^c})) \mapsto R, r_i \in R \cup \{q_j | j \in M\}$$

[0042] $$q_i \in \mathbb{R}^C$$

[0043] 수학식1에서 LN은 선형 임베딩 레이어에서 수행되는 선형 임베딩을 의미하고, E는 비전 트랜스포머 인코더에서 수행되는 인코딩을 의미하고, S는 모든 토큰들의 인덱스를 의미하고, M은 마스킹된 토큰들의 인덱스를 의미하고, M^c는 마스킹되지 않은 토큰들의 인덱스를 의미하고, C는 토큰의 사이즈를 의미하고, q_i는 C사이즈를 갖는 마스킹되지 않은 토큰을 의미하고, q_j는 C사이즈를 갖는 마스킹된 토큰을 의미하고, i와j는 토큰의 인덱스를 의미하고, R은 인코딩된 제1정보의 집합을 의미하고, r_i는 인코딩된 제1정보(즉, 인코딩된 마스킹되지 않은 토큰)와 제2정보를 모두 포함하는 것을 의미한다. 수학식 1을 참조하면, 제1정보(즉, 마스킹되지 않은 토큰인 q_i)는 인코더에서 인코딩(E) 후 선형 임베딩 레이어에서 선형 임베딩(LN)되어 인코딩된 제1정보(R)로 변환된다. 수학식 1을 참조하면, 제1정보(마스킹되지 않은 토큰인 q_i)만이 인코딩 함수(E)를 통해 인코딩된다는 것을 알 수 있다. 전자 장치(200)는 제1정보와 제2정보 모두가 아닌 제1토큰에 대해서만 인코딩을 수행하므로 연산량이 감소되는 효과를 가질 수 있다. 수학식 1에서 r_i는 제1예측 정보를 생성하는 비전 트랜스포머 디코더(제1디코더)의 입력 정보인 인코딩된 제1정보 및 제2정보를 정의하기 위해 사용된다.

[0044] 비디오는 프레임들 간의 변화들의 묶음이고, 따라서 프레임들의 변화는 비디오를 이해하기 위한 비결일 수 있다. 이러한 직관을 반영하기 위해 본 개시에서는 비전 트랜스포머 모델을 사용한 다중 프레임 예측을 제안한다. 예측 모듈(210)은 다중 프레임(예를 들어, 현재 프레임인 제1프레임 내지 제4프레임과 미래 프레임(또는, 후행 프레임)인 제5프레임)을 예측할 수 있다. 도 2를 참조하면, 예측 모듈(210)은 입력 정보인 인코딩된 제1정보 및 제2정보로부터 제1예측 정보, 제2예측 정보, 및 제3예측 정보를 생성할 수 있다. 여기서 제1예측 정보는 제1모델(도 2의 상단 비전 트랜스포머 디코더)을 이용하여 인코딩된 제1정보, 및 제2정보에 대응되는 프레임들을 예측한 프레임이고, 제2예측 정보는 제2모델(도 2의 상단 선형 함수 레이어)을 이용하여 제1예측 정보에 대응되는 프레임들에 후행하는 후행 프레임을 예측한 프레임이고, 제3예측 정보는 제3모델(도 2의 하단 비전 트랜스포머 디코더 및 하단 선형 함수 레이어)을 이용하여 생성된 인코딩된 제1정보에 대응되는 후행 프레임을 예측한 프레임일 수 있다. 예를 들어, 도 2를 참조하면 비디오의 연속된 프레임들은 I_{t_1}~I_{t_4}(즉, 제1프레임, 제2프레임, 제3프레임, 및 제4프레임)이고, 인코딩된 제1정보, 및 제2정보에 대응되는 프레임은 마스킹되지 않은 프

레임(즉, I_{t_2} 및 I_{t_3})과 마스킹된 프레임(즉, 마스킹된 I_{t_1} 및 I_{t_4})이고, I_{t_1}~I_{t_4}의 후행 프레임은 I_{t_5}(즉, 제5프레임)이므로 제1예측 정보는 제1모델을 이용하여 마스킹되지 않은 I_{t_2} 및 I_{t_3} 프레임과 마스킹된 I_{t_1}및 I_{t_4} 프레임을 예측한 프레임을 의미하고, 제2예측 정보는 제2모델을 이용하여 I_{t_5}을 예측한 프레임을 의미하고, 제3예측 정보는 제3모델을 이용하여 I_{t_5}을 예측한 프레임을 의미할 수 있다.

[0045] 일 실시예에 따라, 예측 모듈(210)은 제1모델을 이용하여 인코딩된 제1정보, 및 제2정보에 대응되는 프레임에 관한 제1예측 정보를 확인할 수 있다. 일 실시예에 따라, 제1모델은 인코딩된 제1정보 및 제2정보를 입력 정보로서 이용하여 적어도 하나의 마스킹된 프레임에 관한 정보를 복원하는 비전 트랜스포머 디코더(제1디코더)를 포함할 수 있다. 예측 모듈(210)은 선형 임베딩 레이어의 입력 정보로서 제1디코더의 출력을 사용하여 제1예측 정보를 생성할 수 있다. 제1예측 정보는 아래 수학식 2로 표현할 수 있다.

수학식 2

[0046] $$LN(D^{w}(r_{i\in S})) \mapsto \hat{I}'_{t_{i\in M}}$$

[0047] 수학식 2를 참조하면, LN은 선형 임베딩 레이어에서 수행되는 선형 임베딩을 의미하고, D^{w}는 비전 트랜스포머 디코더(제1디코더)에서 수행되는 디코딩을 의미하고, $r_{i\in S}$는 인코딩된 제1정보(즉, 인코딩된 마스킹되지 않은 토큰)와 제2정보(마스킹된 토큰)를 포함하는 모든 정보를 의미하고, S는 모든 토큰들의 인덱스를 의미하고, $\hat{I}'_{t_i}$은 I_{t_i} 프레임을 제1모델을 이용하여 예측한 프레임을 의미할 수 있다. 인코딩된 제1정보와 제2정보 모두($r_{i\in S}$)는 제1디코더에서 디코딩(D^{w})된 후 선형 임베딩 레이어에서 임베딩 (LN) 되어 인코딩된 제1정보, 및 제2정보에 대응되는 프레임에 관한 제1예측 정보($\hat{I}'_{t_{i\in S}}$)로 변환된다. 제1예측 정보($\hat{I}'_{t_{i\in S}}$)는 마스킹된 프레임에 관한 예측 정보($\hat{I}'_{t_{i\in M}}$)를 포함하므로, 인코딩된 제1정보와 제2정보 모두($r_{i\in S}$)는 제1디코더에서 디코딩(D^{w})된 후 선형 임베딩 레이어에서 임베딩(LN) 되어 마스킹된 프레임에 관한 예측 정보($\hat{I}'_{t_{i\in M}}$)로 변환된다. 예측 모듈(210)은 마스킹된 토큰과 마스킹되지 않은 토큰 내 정보들을 모두 고려하기 위해, 제2정보와 마스킹된 프레임의 일반화를 위한 학습 변수로 구성된 인코딩된 제1정보 모두(r_i)를 제1디코더의 입력 정보로서 사용할 수 있다. 예측 모듈(210)에서 제1디코더는 마스킹된 프레임과 마스킹되지 않은 프레임을 모두 복원할 수 있다. 예측 모듈(210)에서 제1디코더는 마스킹되지 않은 프레임뿐만 아니라 마스킹된 프레임도 복원하므로, 제1예측 정보에 마스킹된 프레임에 관한 예측 정보인 $\hat{I}'_{t_{i\in M}}$가 포함되는 것을 알 수 있다. 다시 말해, 예측 모듈(210)은 제1디코더를 이용하여 마스킹된 프레임(즉, 도 2에서 I_{t_1} 및 I_{t_4})을 예측한 프레임(즉, $\hat{I}'_{t_1}$ 및 $\hat{I}'_{t_4}$)를 생성할 수 있다. 도 2를 참조하면, 마스킹된 프레임 예측(Masked Frame Prediciton)은 $\hat{I}'_{t_{i\in M}}$ 프레임(즉, $\hat{I}'_{t_1}$ 및 $\hat{I}'_{t_4}$)을 의미하고, 마스킹되지 않은 프레임 예측은 $\hat{I}'_{t_2}$ 및 $\hat{I}'_{t_3}$을 의미한다. 제1예측 정보는 마스킹된 프레임 예측과 마스킹되지 않은 프레임 예측을 포함할 수 있다. 마스킹된 프레임 예측은 이하에서 비디오 이상치를 판단할 때 사용되고, 예측 프레임과 실측 프레임의 유클리드 거리가 최소화되도록 모델을 훈련시킬 때 사용될 수 있다.

[0048] 일 실시예에 따라, 예측 모듈(210)은 제2모델을 이용하여 비디오의 프레임들에 후행하는 후행 프레임에 관한 제2예측 정보를 확인할 수 있다. 일 실시예에 따라, 제2모델은 선형 함수를 통해 제1예측 정보로부터 제2예측 정

보를 생성하는 선형 함수 레이어를 포함할 수 있다. 제2예측 정보는 아래 수학식 3로 표현할 수 있다.

수학식 3

[0049] $$\phi^{w}(\hat{I}'_{t_{i\in S}}) \rightarrow \hat{I}^{w}_{t_5}$$

[0050] 수학식 3을 참조하면, ϕ^{w}는 제2예측 정보를 생성하기 위한 선형 함수를 의미하고, $\hat{I}'_{t_{i\in S}}$는 제1예측 정보(즉, 마스킹된 프레임 예측 및 마스킹되지 않은 프레임 예측)를 의미하고, $\hat{I}^{w}_{t_5}$은 제2모델을 이용하여 제5프레임(제1예측 정보에 대응되는 프레임들(도 2에서 제1프레임, 제2프레임, 제3프레임, 및 제4프레임)에 후행하는 후행 프레임, 또는 미래 프레임)을 예측한 프레임을 의미할 수 있다. 수학식 3을 참조하면, 제1예측 정보($\hat{I}'_{t_{i\in S}}$)는 선형 함수(ϕ^{w})를 통해 제2예측 정보($\hat{I}^{w}_{t_5}$)로 변환된다. 도 2를 참조하면, 선형 함수(ϕ^{w})를 통한 제2예측 정보($\hat{I}^{w}_{t_5}$) 생성 과정은 선형 함수 레이어에서 수행될 수 있다. 제2예측 정보($\hat{I}^{w}_{t_5}$)는 마스킹된 프레임 예측($\hat{I}'_{t_1}$ 및 $\hat{I}'_{t_4}$) 및 마스킹되지 않은 프레임 예측($\hat{I}'_{t_2}$ 및 $\hat{I}'_{t_3}$) 모두에 기초하여 생성되므로 전체 미래 예측(Whole Future Prediction)이라고 지칭될 수 있다.

[0051] 일 실시예에 따라, 예측 모듈(210)은 제3모델을 이용하여 인코딩된 제1정보에 대응되는 후행 프레임에 관한 제3예측 정보를 확인할 수 있다. 일 실시예에 따라, 제3모델은 인코딩된 제1정보를 입력 정보로서 이용하여 적어도 하나의 마스킹되지 않은 프레임에 관한 정보를 복원하는 비전 트랜스포머 디코더(제2디코더) 및 선형 함수를 통해 복원된 적어도 하나의 마스킹되지 않은 프레임에 관한 정보로부터 제3예측 정보를 생성하는 선형 함수 레이어를 포함할 수 있다. 제3예측 정보는 아래 수학식 4로 표현될 수 있다.

수학식 4

[0052] $$\phi^{p}(D^{p}(r_{i\in M^{c}})) \rightarrow \hat{I}^{p}_{t_5}$$

[0053] 수학식 4를 참조하면, ϕ^{p}는 제3예측 정보를 생성하기 위한 선형 함수를 의미하고, D^{p}는 비전 트랜스포머 디코더(제2디코더)에서 수행되는 디코딩을 의미하고, M^{c}는 마스킹되지 않은 토큰들의 인덱스를 의미하고, $r_{i\in M^{c}}$는 인코딩된 제1정보를 의미하고, $\hat{I}^{p}_{t_5}$는 제3모델을 이용하여 제5프레임(후행 프레임 또는 미래 프레임)을 예측한 프레임을 의미할 수 있다. 수학식 4를 참조하면, 인코딩된 제1정보($r_{i\in M^{c}}$)는 제2디코더에서 디코딩 (D^{p})된 후 선형 함수(ϕ^{p})를 통해 제3예측 정보($\hat{I}^{p}_{t_5}$)로 변환된다. 도 2를 참조하면, 선형 함수(ϕ^{p})를 통한 제3예측 정보 생성 과정은 선형 함수 레이어에서 수행될 수 있다. 제2디코더는 미래 프레임을 예측하기 위해 마스킹되지 않은 정보만을 입력 정보로서 이용하므로 제3예측 정보($\hat{I}^{p}_{t_5}$)는 부분 미래 예측(Partial Future Prediction)이라고 지칭될 수 있다.

[0054] 광학적 흐름(Optical Flow)은 주로 비디오로부터 동작 정보를 추출하기 위해 사용된다. 광학적 흐름은 훈련용 샘플에서 객체의 픽셀이 이동한 속력과 거리에 관한 정보를 포함하므로 비디오 이상치 탐지에 유용하게 사용될 수 있다. 광학적 흐름은 객체의 동작 정보를 포함하므로, 단순히 광학적 흐름에 대해 재구성 작업을 수행함으로써 객체의 동작의 정규성을 학습할 수 있다. 따라서, 본 개시는 재구성 모듈(220)을 사용함으로써 정상 훈련용 샘플에서 학습되지 않은 객체의 이상 동작을 효과적으로 탐지할 수 있다. 도 2를 참조하면, 재구성 모듈(220)은 컨볼루션 오토인코더를 포함할 수 있다. 일 실시예에 따라, 전자 장치(200)는 비디오의 연속된 프레임들 중 마지막 프레임 내 객체의 동작 패턴에 관한 정보를 확인하고, 제4모델을 이용하여 동작 패턴에 관한 정보에 대응

되는 객체의 동작 패턴을 재구성한 재구성 정보를 확인할 수 있다. 일 실시예에 따라, 전자 장치(200)는 비디오의 연속된 프레임으로부터 프레임 내 객체의 동작 패턴에 관한 정보(또는, 광학적 흐름)를 계산할 수 있다. 예를 들어, 전자 장치(200)는 Lucas-Kanade 알고리즘, Horn-Schunck 알고리즘, 또는 Bruhn 알고리즘을 통해 프레임 내 객체의 동작 패턴에 관한 정보를 계산할 수 있다. 전자 장치(200)는 컨볼루션 오토인코더의 입력 정보로서 비디오의 연속된 프레임들 중 마지막 프레임의 동작 패턴에 관한 정보(O_t)를 이용하여 객체의 동작 패턴을 재구성한 재구성 정보($\hat{O}_t$)를 생성할 수 있다. 예를 들어, 도 2를 참조하면, 전자 장치(200)는 컨볼루션 오토인코더의 입력 정보로서 비디오의 마지막 제4프레임(I_{t_4})의 동작 패턴에 관한 정보(O_4)를 이용하여 제4프레임의 재구성 정보($\hat{O}_4$)를 확인할 수 있다.

[0055] 일 실시예에 따라, 전자 장치(200)는 제1예측 정보 중 적어도 하나의 마스킹된 프레임에 관한 예측 정보, 제2예측 정보, 및 제3예측 정보를 기초로 비디오의 이상치를 판단할 수 있다. 다른 실시예에 따라, 전자 장치(200)는 제1예측 정보 중 적어도 하나의 마스킹된 프레임에 관한 예측 정보, 제2예측 정보, 제3예측 정보 및 재구성 정보를 기초로 비디오의 이상치를 판단할 수 있다. 일 실시예에 따라, 전자 장치(200)는 적어도 하나의 마스킹된 프레임에 관한 예측 정보와 적어도 하나의 마스킹된 프레임에 관한 실측 정보간의 유클리드 거리, 제2예측 정보와 비디오의 연속된 프레임들에 후행하는 후행 프레임에 관한 실측 정보간의 유클리드 거리, 및 제3예측 정보와 비디오의 연속된 프레임들에 후행하는 후행 프레임에 관한 실측 정보간의 유클리드 거리의 합인 예측 유클리드 거리를 계산하고, 재구성 정보와 동작 패턴에 관한 정보간의 유클리드 거리인 재구성 유클리드 거리를 계산할 수 있다. 프레임에 관한 실측 정보란 해당 프레임에 관한 실측 값(Ground Truth)을 의미할 수 있다. 예를 들어, 후행 프레임에 관한 실측 정보란 후행 프레임에 관한 실측 값을 의미할 수 있다. 전자 장치(200)는 예측 유클리드 거리 및 재구성 유클리드 거리에 기초하여 비디오의 이상치를 탐지할 수 있다. 예측 유클리드 거리 및 재구성 유클리드 거리는 아래 수학식 5로 계산될 수 있다.

수학식 5

$$\mathcal{L}_{pred} = \|\hat{I}^{w}_{t_5} - I_{t_5}\|_2^2 + \|\hat{I}^{p}_{t_5} - I_{t_5}\|_2^2 + \sum_{m \in M} \|\hat{I'}_{t_m} - I'_{t_m}\|_2^2$$

$$\mathcal{L}_{recon} = \|\hat{O}_t - O_t\|_2^2$$

[0056]

[0057] 수학식 5를 참조하면, 전술한 바와 같이 $\hat{I}^{w}_{t_5}$는 전체 미래 예측(즉, 제2예측 정보)를 의미하고, I_{t_5}는 비디오의 미래 프레임(도 2에서 제5 프레임)의 실측 값을 의미하고, $\hat{I}^{p}_{t_5}$는 부분 미래 예측(즉, 제3예측 정보)을 의미하고, $\hat{I'}_{t_m}$은 마스킹된 프레임 예측(즉, 제1예측 정보 중 마스킹된 프레임에 관한 예측 정보)을 의미하고, I'_{t_m}은 마스킹된 프레임 예측의 실측 값을 의미하고, $\hat{O}_t$는 t번째 프레임의 재구성 정보를 의미하고, O_t는 t번째 프레임의 광학적 흐름(도 2에서 컨볼루션 오토인코더의 입력 정보인 비디오의 연속된 프레임들 중 마지막 프레임 내 객체의 동작 패턴에 관한 정보)을 의미할 수 있다. 수학식 5를 참조하면, $\mathcal{L}_{pred}$는 각각의 예측(전체 미래 예측, 부분 미래 예측, 및 마스킹된 프레임 예측)과 각각의 예측에 대응하는 실측 프레임간의 유클리드 거리(즉, L2 distance)의 합인 예측 유클리드 거리를 의미하고, $\mathcal{L}_{recon}$는 t번째 프레임의 재구성 정보와 t번째 프레임의 광학적 흐름간의 유클리드 거리인 재구성 유클리드 거리를 의미한다. 수학식 5에서는 $\mathcal{L}_{recon}$을 계산할 때 t번째 프레임을 기준으로 계산하는 것으로 기재되었으나, 본원의 이상치 탐지 성능을 측정할 때에는 비디오의 현재 프레임인 제4프레임(즉, t=4)을 기준으로 $\mathcal{L}_{recon}$을 계산할 수 있다. 예측 유클리드 거리($\mathcal{L}_{pred}$)와 재구성 유클리드 거리($\mathcal{L}_{recon}$)는 이하 비디오 이상치 점수를 계산할 때 사용될 수 있다.

[0059] 일 실시예에 따라 전자 장치(200)는 마스킹된 프레임 예측과 마스킹된 프레임 예측의 실측 값 간의 제1유클리드 거리, 전체 미래 예측과 전체 미래 예측의 실측 값 간의 제2유클리드 거리, 및 부분 미래 예측과 부분 미래 예측의 실측 값 간의 제3유클리드 거리를 계산할 수 있다. 일 실시예에 따라, 전자 장치(200)는 제1가중치가 부여된 제1유클리드 거리, 제2가중치가 부여된 제2유클리드 거리, 및 제3가중치가 부여된 제3유클리드 거리로부터 비디오 이상치 탐지를 위한 비디오 이상치 점수를 결정할 수 있다. 제1가중치, 제2가중치, 및 제3가중치 각각은 기 설정될 수 있다. 일 실시예에 따라 전자 장치(200)는 제1가중치가 부여된 예측 유클리드 거리 및 제2가중치가 부여된 재구성 유클리드 거리로부터 비디오 이상치 점수를 결정할 수 있다. 비디오 이상치 점수는 아래 수학식 6로 계산될 수 있다.

수학식 6

[0060] $$\mathcal{S} = \lambda_a \mathcal{L}_{pred} + \lambda_o \mathcal{L}_{recon}$$

[0061] 수학식 6을 참조하면, $\mathcal{S}$는 비디오 이상치 점수를 의미하고, λ_a는 제1가중치를 의미하고, $\mathcal{L}_{pred}$은 예측 유클리드 거리를 의미하고, λ_o는 제2가중치를 의미하고, $\mathcal{L}_{recon}$은 재구성 유클리드 거리를 의미한다. 비디오 이상치 점수($\mathcal{S}$)는 제1가중치가 부여된 예측 유클리드 거리($\lambda_a \mathcal{L}_{pred}$)와 제2가중치가 부여된 재구성 유클리드 거리($\lambda_o \mathcal{L}_{recon}$)의 합으로 계산될 수 있다. 제1가중치 및 제2가중치 각각은 기 설정될 수 있다. 예를 들어, (λ_a, λ_o)는 (0.05,0.94), 또는 (2.0,1.0)으로 설정될 수 있다. 일 실시예에 따라, 전자 장치(200)는 계산된 비디오 이상치 점수($\mathcal{S}$)를 소정의 임계 값과 비교하여 비디오의 이상치를 판단할 수 있다. 예를 들어, 전자 장치(200)는 비디오 이상치 점수($\mathcal{S}$)가 소정의 임계 값 이상인 경우 제4프레임을 '이상'이라고 판단할 수 있고, 비디오 이상치 점수($\mathcal{S}$)가 소정의 임계 값 미만인 경우 제4프레임을 '정상'이라고 판단할 수 있다.

[0063] 도 3은 일 실시예에 따른 예측 모듈(210)을 학습시키는 방법을 나타낸다.

[0064] S310 단계에서 전자 장치(200)는 비디오로부터 입력받은 입력 프레임으로부터 물체(또는 객체) 위주로 프레임을 분할할 수 있다. 도 2를 참조하면, 분할된 프레임 각각은 순서대로 I_{t_1}, I_{t_2}, I_{t_3}, 및 I_{t_4}일 수 있다.

[0065] S320 단계에서 전자 장치(200)는 기 설정된 마스킹 비율에 따라 특정 프레임들을 임의로 마스킹할 수 있다. 도 2를 참조하면, 전자 장치(200)는 분할된 각각의 프레임 I_{t_1}, I_{t_2}, I_{t_3}, 및 I_{t_4}중에서 I_{t_1}과 I_{t_4}를 마스킹할 수 있다. 마스킹된 프레임과 마스킹되지 않은 프레임 각각은 플래튼 레이어와 선형 임베딩 레이어를 거쳐 마스킹된 토큰과 마스킹되지 않은 토큰 각각으로 변환될 수 있다. 마스킹되지 않은 토큰은 인코더를 통해 인코딩 될 수 있다.

[0066] S330 단계에서 예측 모듈(210)은 마스킹된 프레임 예측, 전체 미래 프레임 예측, 및 부분 미래 프레임 예측을 수행할 수 있다. 예측 모듈(210)은 마스킹된 토큰과 마스킹되지 않은 토큰을 기초로 마스킹된 프레임 예측(제1예측 정보 중 마스킹된 프레임에 관한 예측 정보)을 확인할 수 있고, 마스킹된 프레임 예측을 기초로 전체 미래 예측(제2예측 정보)을 확인할 수 있고, 마스킹되지 않은 토큰을 기초로 부분 미래 예측(제3예측 정보)을 확인할 수 있다.

[0067] S340 단계에서 예측 모듈(210)은 각각의 예측된 프레임과 실측 프레임과의 차이를 최소화하도록 학습할 수 있다. 예측 모듈(210)은 각각의 예측된 프레임과 실측 프레임과의 차이의 합인 예측 유클리드 거리를 계산하고, 예측 유클리드 거리를 최소화하도록 인코더 및 디코더(제1디코더 및 제2디코더)를 학습할 수 있다. 일 실시예에 따라, 예측 모듈(210)은 제1예측 정보 중에서 마스킹된 프레임에 관한 예측 정보($\hat{I}'_{t_{i\in M}}$)만을 사용하여 예측 유클리드 거리를 최소화하도록 인코더 및 디코더를 학습할 수 있다.

[0069] 도 4는 일 실시예에 따른 재구성 모듈(220)을 학습시키는 방법을 나타낸다.

[0070] S410 단계에서, 재구성 모듈(220)은 광학적 흐름 프레임들을 객체(또는, 물체) 위주로 분할할 수 있다. 일 실시예에 따라, 재구성 모듈(220)은 비디오의 프레임들을 객체 위주로 분할하고 연속된 두 프레임간의 광학적 흐름(또는 객체의 동작 패턴에 관한 정보)을 계산할 수 있다. 광학적 흐름은 이전 프레임과 다음 프레임간에 픽셀이 이동한 속력과 거리에 관한 정보를 포함할 수 있다. 예를 들어, 제4프레임의 광학적 흐름은 제3프레임과 제4프레임간에 픽셀이 이동한 속력과 거리의 벡터가 표현된 이미지를 포함할 수 있다.

[0071] S420 단계에서, 재구성 모듈(220)은 프레임 내 객체의 광학적 흐름을 재구성하도록 오토인코더 모델을 학습시킬 수 있다. 일 실시예에 따라, 재구성 모듈(220)은 오토인코더 모델의 입력 정보인 광학적 흐름과 오토인코더 모델의 출력 정보인 재구성 정보간의 유클리드 거리인 재구성 유클리드 거리(L_{recon})를 최소화하도록 오토인코더를 학습시킬 수 있다.

[0073] 도 5는 일 실시예에 따른 전자 장치와 종래 기술을 통해 이상치 탐지 결과들을 나타낸다.

[0074] 도 5를 참조하면, 510의 1열은 비정상으로 규정한 객체(달리고 있는 사람)의 프레임을 나타내고, 520의 1열은 정상으로 규정한 객체(걷고 있는 사람)의 프레임을 나타낸다. 종래 기술과 본원을 통해 이상치를 탐지하여 시각화한 결과들은 색상이 밝을수록 이상치 점수가 높은 것을 의미한다. 종래 기술의 이상치 점수는 본원과 마찬가지로 유클리드 거리를 통해 계산되는 것으로 가정한다. 510의 2열은 비정상으로 규정한 객체의 프레임을 MemAE(Memory-augumented Autoencoder)를 통해 이상치를 탐지하여 시각화한 결과이고, 520의 2열은 정상으로 규정한 객체의 프레임을 MemAE를 통해 이상치를 탐지하여 시각화한 결과이다. 510의 3열은 비정상으로 규정한 객체의 프레임을 ML-MemAE-SC(Multi-Level Memory Modules in an Autoencoder with Skip Connections)를 통해 이상치를 탐지하여 시각화한 결과이고, 520의 3열은 정상으로 규정한 객체의 프레임을 ML-MemAE-SC를 통해 이상치를 탐지하여 시각화한 결과이다. 510의 4열은 비정상으로 규정한 객체의 프레임을 본원을 통해 이상치를 탐지하여 시각화한 결과이고, 520의 4열은 정상으로 규정한 객체의 프레임을 본원을 통해 이상치를 탐지하여 시각화한 결과이다. 510의 2열과 520의 2열을 참조하면, MemAE는 비정상으로 규정한 객체의 프레임의 이상치를 탐지한 결과와 정상으로 규정한 객체의 프레임의 이상치를 탐지한 결과의 차이가 크지 않은 것을 볼 수 있다. 따라서, MemAE를 사용할 경우 이상치 판단을 위한 임계 값 설정이 난해한 문제점이 있어 이상치 탐지 결과가 부정확할 수 있다. 510의 3열과 520의 3열을 참조하면, ML-MemAE-SC의 경우 MemAE에 대비하면 정상으로 규정된 객체의 프레임의 이상치 탐지 결과와 비정상으로 규정된 객체의 프레임의 이상치 탐지 결과가 조금 더 구분된다. 다만, 본원을 통해 이상치를 탐지한 결과인 510의 4열과 520의 4열의 차이가 510의 3열과 520의 3열의 차이보다 뚜렷하게 구별되는 것을 볼 수 있다. 따라서, 메모리 모듈을 사용하지 않는 본원이 메모리 모듈을 사용하는 MemAE 및 ML-MemAE-SC 대비 우월한 이상치 탐지 성능을 제공함을 알 수 있다. 아래 표1은 MemAE, ML-MemAE-SC, 및 본원의 예측 모듈(210)을 통한 AUROC(Area Under the Receiver Operating Characteristic Curve) 성능(%) 비교 결과를 나타낸다.

표 1

[0075]

Method	MemAE	ML-MemAE-SC	MAE	Ours
AUROC	70.2	81.9	81.6	**86.2**

[0076] 표 1을 참조하면, MemAE는 70.2%의 낮은 이상치 탐지 성능을 제공하는 것을 알 수 있다. ML-MemAE-SC와 MAE(Masked Autoencoder)는 각각 81.9%와 81.6%로 MemAE 대비 나은 성능을 보여주지만, 본원의 86.2%에 비하면 5% 이상의 낮은 성능을 제공하는 것을 알 수 있다. 따라서, 종래 기술인 MemAE, ML-MemAE-SC, 및 MAE 대비 본원은 소위 '생성 문제'를 회피할 수 있어 비디오의 시간적인 맥락에서 향상된 이상치 탐지 성능을 제공한다는 것을 알 수 있다.

[0077] 아래 표2는 MemAE, ML-MemAE-SC, 및 본원의 재구성 모듈(220)을 통한 AUROC 성능(%) 비교 결과를 나타낸다.

표 2

Method	MemAE	ML-MemAE-SC	CAE
Flow	77.9	86.8	79.8
Ovarall	89.3	91.7	**92.1**

[0078]

[0079] 표 2를 참조하면, Flow 행은 본원의 재구성 모듈(220)만을 사용한 경우의 비디오 이상치 탐지 성능을 나타내고, Overall 행은 본원의 예측 모듈(210)과 재구성 모듈(220)을 모두 사용한 경우의 비디오 이상치 탐지 성능을 나타낸다. 즉, Overall 행에서 MemAE 열은 본원의 재구성 모듈(220)을 MemAE로 구성하면서 예측 모듈(210)과 재구성 모듈(220)을 모두 사용한 경우의 이상치 탐지 성능을 나타내고, ML-MemAE-SC 열은 본원의 재구성 모듈(220)을 ML-MemAE-SC로 구성하면서 예측 모듈(210)과 재구성 모듈(220)을 모두 사용한 경우의 이상치 탐지 성능을 나타내고, CAE(Convolutional Autoencoder) 열은 본원의 재구성 모듈(220)을 CAE로 구성하면서 예측 모듈(210)과 재구성 모듈(220)을 모두 사용한 경우의 이상치 탐지 성능을 나타낸다.

[0080] 표 2의 Flow 행을 참조하면 본원의 재구성 모듈(220)만을 사용하여 비디오 이상치를 탐지하는 경우, 본원의 재구성 모듈(220)을 MemAE로 구성하면 이상치 탐지 성능이 77.9%이고, ML-MemAE-SC로 구성하면 이상치 탐지 성능이 86.8%이고, CAE로 구성하면 이상치 탐지 성능이 79.8%임을 알 수 있다. 표 2의 Overall 행을 참조하면 본원의 예측 모듈(210)을 공통적으로 사용하는 경우에, 본원의 재구성 모듈(220)을 MemAE로 구성하면 이상치 탐지 성능이 89,3%이고, ML-MemAE-SC로 구성하면 이상치 탐지 성능이 91.7%이고, CAE로 구성하면 이상치 탐지 성능이 92.1%임을 알 수 있다.

[0081] 결과적으로, 재구성 모듈(220)만을 사용한 경우의 성능에 비해 예측 모듈(210)과 재구성 모듈(220)을 사용한 경우의 성능이 우월하고, 예측 모듈(210)과 재구성 모듈(220)을 모두 사용하여 비디오 이상치를 탐지하는 경우에 재구성 모듈(220)을 CAE로 구성하는 것이 가장 높은 비디오 이상치 탐지 성능을 제공할 수 있다는 것을 확인할 수 있다. 따라서, 표 2를 통해 본원은 예측 모듈(210)과 재구성 모듈(220)을 동시에 사용하면서 재구성 모듈(220)이 CAE를 포함할 수 있으므로 종래 기술 대비 최고 92.1%에 달하는 이상치 탐지 성능을 제공할 수 있다는 것을 확인할 수 있다.

[0083] 아래 표3은 본원의 예측 모듈(210)을 통해 예측된 마스킹된 프레임 예측, 전체 미래 예측, 및 부분 미래 예측 각각을 조합한 다중 프레임 예측의 AUROC 성능 비교 결과를 나타낸다.

표 3

masked	whole	partial	result
✗	✗	✗	**70.5**
✓	✗	✗	**81.6**
✓	✓	✗	**84.9**
✓	✓	✓	**86.2**

[0084]

[0085] 표 3은 본원의 재구성 모듈(220)을 제외하고 비디오 이상치 탐지 성능을 평가한 결과이다. 표 3을 참조하면, 마스킹된 프레임 예측, 전체 미래 예측, 및 부분 미래 예측을 모두 사용하지 않은 경우(일반적인 비전 트랜스포머의 경우) 70.5%의 성능을 갖는 것을 볼 수 있다. 마스킹된 프레임 예측만을 사용하는 경우 본원의 예측 모듈(210)은 81.6%의 성능을 갖는 것을 볼 수 있다. 마스킹된 프레임 예측만을 사용하더라도 비디오의 시간적 맥락을 반영할 수 있고, '생성 문제'를 회피할 수 있어 마스킹된 프레임 예측을 사용하지 않는 경우 대비 최대 11.6%의 성능이 향상되는 것을 알 수 있다. 마스킹된 프레임 예측과 전체 미래 예측을 사용하는 경우 본원의 예측 모듈(210)은 84.9%의 성능을 갖는 것을 볼 수 있다. 전체 미래 예측을 사용하는 경우 마스킹된 프레임 예측

만을 사용하는 경우보다 예측된 프레임이 정상 상태(normality)를 더 반영할 수 있고, 전체 미래 예측은 이전 프레임들로만 예측되어 '생성 문제'에 보다 덜 민감해지므로 최대 3.3% 성능이 향상되는 것을 알 수 있다. 마스킹된 프레임 예측, 전체 미래 예측, 및 부분 미래 예측을 모두 사용하는 경우 본원의 예측 모듈(210)은 86.2%의 성능을 갖는 것을 볼 수 있다. 부분 미래 예측은 마스킹된 프레임들로만 예측되어 프레임의 숨겨진 특징을 더 반영할 수 있으므로 부분 미래 예측을 사용하지 않는 경우보다 최대 1.3%의 성능이 향상되는 것을 알 수 있다.

[0087] 도 6은 일 실시예에 따른 전자 장치와 종래 기술 간의 이상치 탐지 비교 결과들을 나타낸다.

[0088] 도 6에서 이상치를 탐지하여 시각화한 결과들은 색상이 밝을수록 이상치 점수가 높은 것을 의미한다. 610의 제1행, 제2행, 및 제3행의 제1열은 각각 비정상으로 규정한 객체(달리고 있는 사람)의 프레임을 나타낸다. 620의 제1행, 제2행, 및 제3행의 제1열은 각각 정상으로 규정한 객체(걷고 있는 사람)의 프레임을 나타낸다. 610의 제1행, 제2행, 및 제3행의 제2열은 각각 비정상으로 규정한 객체의 프레임을 ML-Mem-SC를 통해 이상치를 탐지하여 시각화한 결과를 나타낸다. 620의 제1행, 제2행, 제3행의 제2열은 각각 정상으로 규정한 객체의 프레임을 ML-Mem-SC를 통해 이상치를 탐지하여 시각화한 결과를 나타낸다. 610의 제1행, 제2행, 및 제3행의 제3열은 각각 비정상으로 규정한 객체의 프레임을 본원을 통해 이상치를 탐지하여 시각화한 결과를 나타낸다. 620의 제1행, 제2행, 및 제3행의 제3열은 각각 정상으로 규정한 객체의 프레임을 본원을 통해 이상치를 탐지하여 시각화한 결과를 나타낸다. 도 6을 참조하면 610의 제1행, 제2행, 및 제3행의 제2열과 620의 제1행, 제2행, 및 제3행의 2열의 차이보다 610의 제1행, 제2행, 및 제3행의 제3열과 620의 제1행, 제2행, 및 제3행의 제3열의 차이가 확연히 구분되는 것을 볼 수 있다.

[0089] 610의 제4행의 제1열은 비정상으로 규정한 객체(가방을 던지는 사람)의 프레임을 나타내고, 620의 제4행의 제1열은 정상으로 규정한 객체(걷고 있는 사람)의 프레임을 나타낸다. 610의 제5행의 제1열은 비정상으로 규정한 객체(쓰레기를 버리는 사람)의 프레임을 나타내고, 620의 제5행의 제1열은 정상으로 규정한 객체(서 있는 사람)의 프레임을 나타낸다. 제1행 내지 제3행과 마찬가지로, 610의 제4행 제2열과 620의 제4행 제2열의 차이보다 610의 제4행 제3열과 620의 제4행 제3열의 차이가 뚜렷하고, 610의 제5행 제2열과 620의 제5행 제2열의 차이보다 610의 제5행 제3열과 620의 제5행 제3열의 차이가 뚜렷한 것을 볼 수 있다. 따라서, 본원이 ML-Mem-SC 대비 향상된 이상치 탐지 성능을 제공하는 것을 볼 수 있다.

[0091] 도 7은 USCD Pedestrian 2 데이터 세트를 이용해 본원의 전자 장치를 훈련하고, 특정 비디오의 비디오 이상치 점수를 계산하여 그래프로 나타낸 것이다. 도 7을 참조하면, 710은 자전거를 타는 사람 및 스케이트 보드를 타는 사람을 비정상 객체로 규정하고 비디오 이상치 탐지를 수행한 결과이다. 710의 그래프의 x축은 프레임 순서를 나타내고, y축은 비디오 이상치 점수를 나타낸다. 710의 노란색 박스의 양 끝은 비정상으로 규정한 객체의 프레임의 시점과 종점을 나타낸다. 710에서 노란색 박스의 시점을 기점으로 y축의 비디오 이상치 점수가 급증하고, 노란색 박스의 종점에 도달할수록 y축의 비디오 이상치 점수가 감소하는 것을 볼 수 있다. 따라서, 본원의 전자 장치가 USCD Pedestrian 2 데이터 세트로 훈련된 경우에도 높은 비디오 이상치 탐지 성능을 제공하는 것을 알 수 있다.

[0093] 도 8은 CUHK Avenue 데이터 세트를 이용해 본원의 전자 장치를 훈련하고, 비디오 이상치 점수를 계산하여 특정 비디오의 비디오 이상치 점수를 계산하여 그래프로 나타낸 것이다. 도 8을 참조하면, 810은 자전거를 타는 사람을 비정상 객체로 규정하고 비디오 이상치 탐지를 수행한 결과이다. 810의 그래프의 x축은 프레임 순서를 나타내고, y축은 비디오 이상치 점수를 나타낸다. 810의 노란색 박스의 양 끝은 비정상으로 규정한 객체의 프레임의 시점과 종점을 나타낸다. 810에서 노란색 박스의 시점을 기점으로 y축의 비디오 이상치 점수가 급증하고, 노란색 박스의 종점에 도달할수록 y축의 비디오 이상치 점수가 감소하는 것을 볼 수 있다. 따라서, 본원의 전자 장치가 CUHK Avenue 데이터 세트로 훈련된 경우에도 높은 비디오 이상치 탐지 성능을 제공하는 것을 알 수 있다.

[0095] 도 9는 일 실시예에 따른 전자 장치의 동작 방법을 나타낸다.

[0096] 단계 S910에서 전자 장치는 비디오의 연속된 프레임들 중에서 적어도 하나의 프레임에 대해 마스킹을 수행하여 적어도 하나의 마스킹되지 않은 프레임에 관한 제1정보 및 적어도 하나의 마스킹된 프레임에 관한 제2정보를 확인할 수 있다. 일 실시예에 따라 전자 장치는 비디오의 연속된 프레임들 중 마지막 프레임 내 객체의 동작 패턴에 관한 정보를 확인할 수 있다.

[0097] 단계 S920에서 전자 장치는 제1정보에 대해 인코딩을 수행하여 인코딩된 제1정보를 확인할 수 있다. 일 실시예에 따라 전자 장치는 비전 트랜스포머 인코더를 이용하여 제1정보에 대해 인코딩을 수행할 수 있다.

[0098] 단계 S930에서 전자 장치는 제1모델을 이용하여 인코딩된 제1정보, 및 제2정보에 대응되는 프레임들에 관한 제1

예측 정보를 확인하고, 제2모델을 이용하여 제1예측 정보에 대응되는 프레임들에 후행하는 후행 프레임에 관한 제2예측 정보를 확인하고, 제3모델을 이용하여 인코딩된 제1정보에 대응되는 후행 프레임에 관한 제3예측 정보를 확인할 수 있다. 일 실시예에 따라, 제1모델은 인코딩된 제1정보 및 제2정보를 입력 정보로서 이용하여 적어도 하나의 마스킹된 프레임에 관한 정보 및 적어도 하나의 마스킹되지 않은 프레임에 관한 정보를 복원하는 비전 트랜스포머 디코더를 포함할 수 있다. 일 실시예에 따라, 제2모델은 선형 함수를 통해 제1예측 정보로부터 제2예측 정보를 생성하는 선형 레이어를 포함할 수 있다. 일 실시예에 따라, 제3모델은 인코딩된 제1정보를 입력 정보로서 이용하여 적어도 하나의 마스킹되지 않은 프레임에 관한 정보를 복원하는 비전 트랜스포머 디코더 및 선형 함수를 통해 복원된 적어도 하나의 마스킹되지 않은 프레임에 관한 정보로부터 제3예측 정보를 생성하는 선형 함수 레이어를 포함할 수 있다. 일 실시예에 따라 전자 장치는 제4모델을 이용하여 동작 패턴에 관한 정보에 대응되는 객체의 동작 패턴을 재구성한 재구성 정보를 확인할 수 있다. 일 실시예에 따라 제4모델은 동작 패턴에 관한 정보를 입력 정보로서 이용하여 객체의 동작 패턴을 재구성한 재구성 정보를 생성하는 컨볼루션 오토인코더를 포함할 수 있다.

[0099] 단계 S940에서 전자 장치는 제1예측 정보 중 적어도 하나의 마스킹된 프레임에 관한 예측 정보, 제2예측 정보, 및 제3예측 정보를 기초로 비디오의 이상치를 판단할 수 있다. 일 실시예에 따라 전자 장치는 제1예측 정보 중 적어도 하나의 마스킹된 프레임에 관한 예측 정보, 제2예측 정보, 및 제3예측 정보를 기초로 비디오의 이상치를 판단할 수 있다. 다른 실시예에 따라 전자 장치는 제1예측 정보 중 적어도 하나의 마스킹된 프레임에 관한 예측 정보, 제2예측 정보, 제3예측 정보, 및 재구성 정보를 기초로 비디오의 이상치를 판단할 수 있다.

[0100] 일 실시예에 따라 전자 장치는 통신 디바이스를 통해 비디오 이상치 판단 결과를 외부 장치로 전송하거나, 디스플레이를 통해 비디오 이상치 판단 결과를 출력할 수 있다.

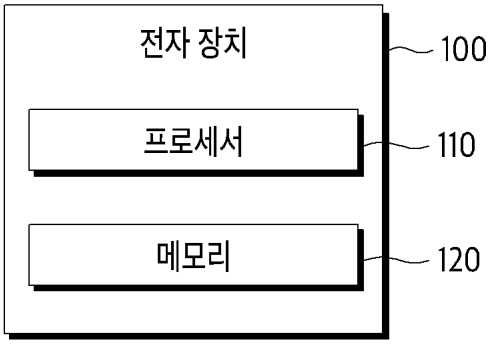
[0101] 전술한 실시예들에 따른 전자 장치는, 프로세서, 프로그램 데이터를 저장하고 실행하는 메모리, 디스크 드라이브와 같은 영구 저장부(permanent storage), 외부 장치와 통신하는 통신 포트, 터치 패널, 키(key), 버튼 등과 같은 사용자 인터페이스 장치 등을 포함할 수 있다. 소프트웨어 모듈 또는 알고리즘으로 구현되는 방법들은 상기 프로세서상에서 실행 가능한 컴퓨터가 읽을 수 있는 코드들 또는 프로그램 명령들로서 컴퓨터가 읽을 수 있는 기록 매체 상에 저장될 수 있다. 여기서 컴퓨터가 읽을 수 있는 기록 매체로 마그네틱 저장 매체(예컨대, ROM(read-only memory), RAM(random-Access memory), 플로피 디스크, 하드 디스크 등) 및 광학적 판독 매체(예컨대, 시디롬(CD-ROM), 디브이디(DVD: Digital Versatile Disc)) 등이 있다. 컴퓨터가 읽을 수 있는 기록 매체는 네트워크로 연결된 컴퓨터 시스템들에 분산되어, 분산 방식으로 컴퓨터가 판독 가능한 코드가 저장되고 실행될 수 있다. 매체는 컴퓨터에 의해 판독가능하며, 메모리에 저장되고, 프로세서에서 실행될 수 있다.

[0102] 본 실시예는 기능적인 블록 구성들 및 다양한 처리 단계들로 나타내어질 수 있다. 이러한 기능 블록들은 특정 기능들을 실행하는 다양한 개수의 하드웨어 또는/및 소프트웨어 구성들로 구현될 수 있다. 예를 들어, 실시예는 하나 이상의 마이크로프로세서들의 제어 또는 다른 제어 장치들에 의해서 다양한 기능들을 실행할 수 있는, 메모리, 프로세싱, 로직(logic), 룩 업 테이블(look-up table) 등과 같은 직접 회로 구성들을 채용할 수 있다. 구성 요소들이 소프트웨어 프로그래밍 또는 소프트웨어 요소들로 실행될 수 있는 것과 유사하게, 본 실시 예는 데이터 구조, 프로세스들, 루틴들 또는 다른 프로그래밍 구성들의 조합으로 구현되는 다양한 알고리즘을 포함하여, C, C++, 자바(Java), 어셈블러(assembler) 등과 같은 프로그래밍 또는 스크립팅 언어로 구현될 수 있다. 기능적인 측면들은 하나 이상의 프로세서들에서 실행되는 알고리즘으로 구현될 수 있다. 또한, 본 실시 예는 전자적인 환경 설정, 신호 처리, 및/또는 데이터 처리 등을 위하여 종래 기술을 채용할 수 있다. “매커니즘”, “요소”, “수단”, “구성”과 같은 용어는 넓게 사용될 수 있으며, 기계적이고 물리적인 구성들로서 한정되는 것은 아니다. 상기 용어는 프로세서 등과 연계하여 소프트웨어의 일련의 처리들(routines)의 의미를 포함할 수 있다.

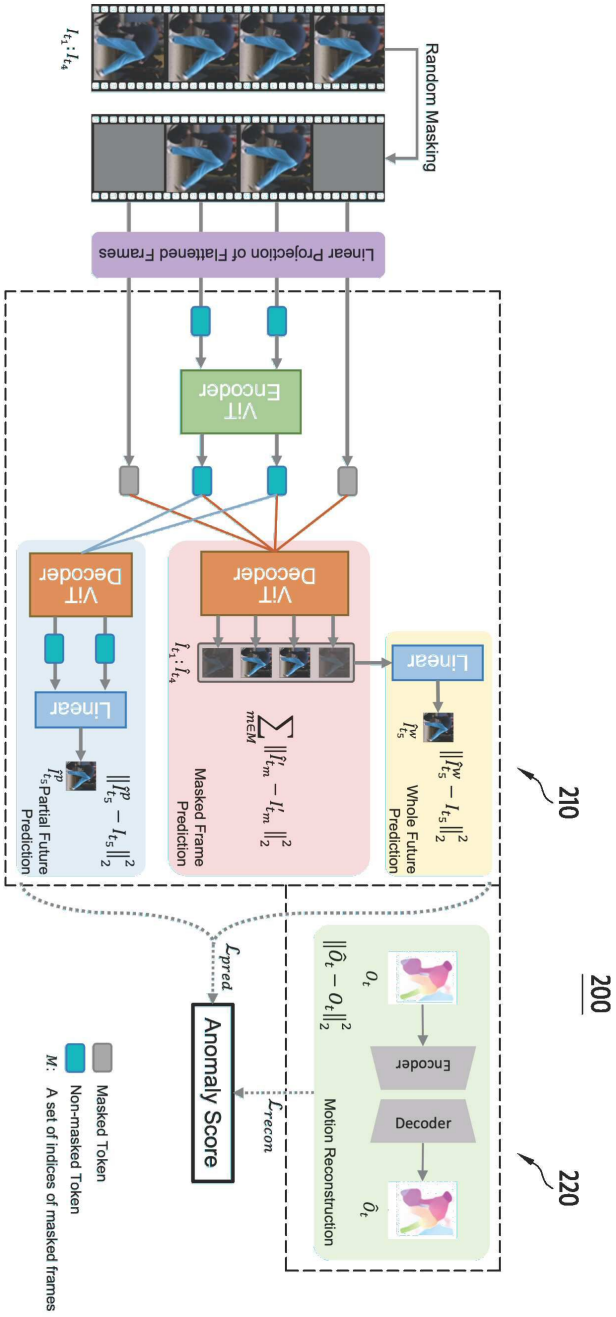
[0103] 전술한 실시예들은 일 예시일 뿐 후술하는 청구항들의 범위 내에서 다른 실시예들이 구현될 수 있다.

도면

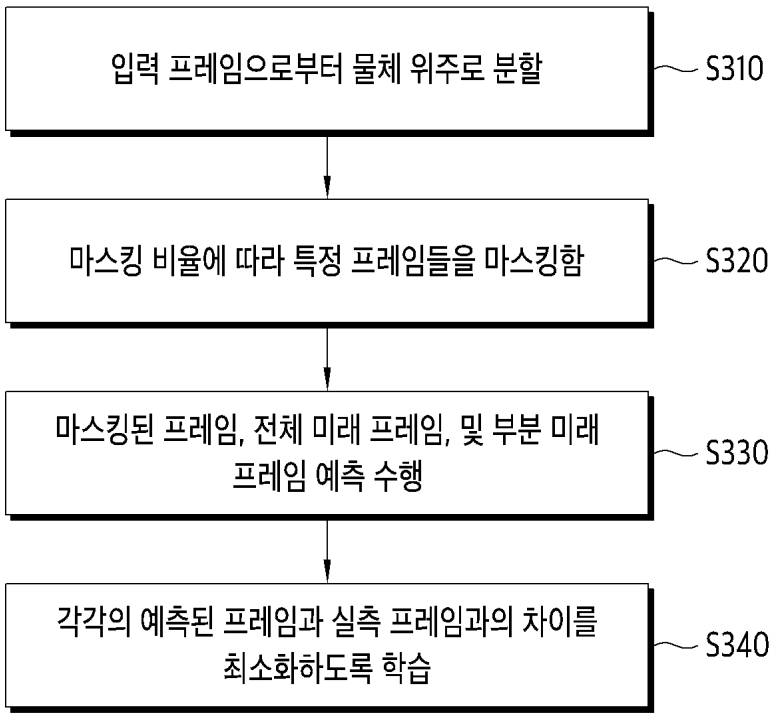
도면1



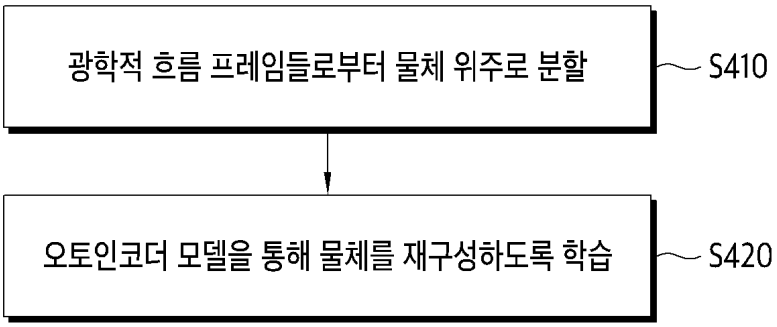
도면2


Random Masking
Linear Projection of Flattened Frames
ViT Encoder
ViT Decoder
Linear
Masked Frame Prediction
ViT Decoder
Linear
Whole Future Prediction
Linear
Partial Future Prediction
210
200
220
Encoder
Decoder
Motion Reconstruction
Anomaly Score
Masked Token
Non-masked Token
A set of indices of masked frames

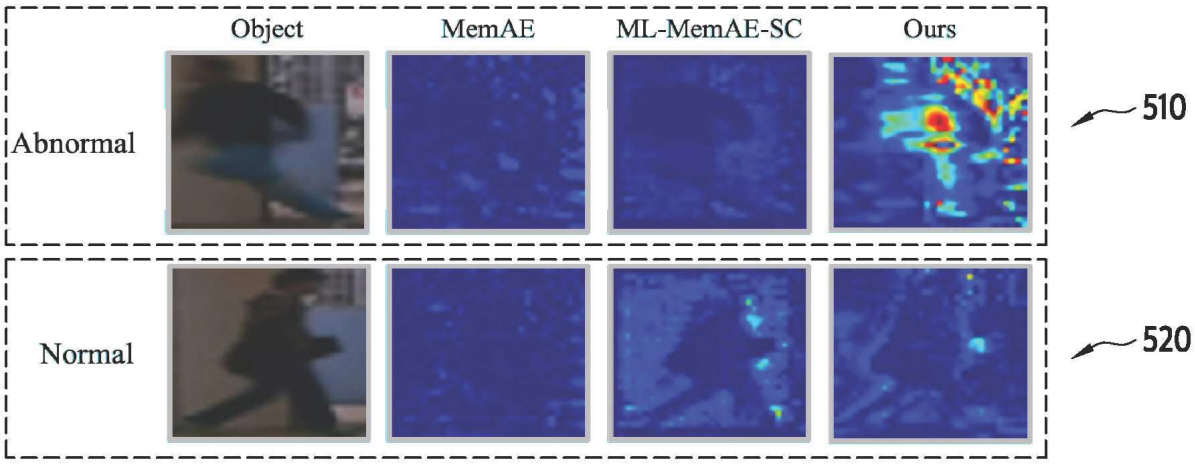
도면3


입력 프레임으로부터 물체 위주로 분할
S310
마스킹 비율에 따라 특정 프레임들을 마스킹함
S320
마스킹된 프레임, 전체 미래 프레임, 및 부분 미래 프레임 예측 수행
S330
각각의 예측된 프레임과 실측 프레임과의 차이를 최소화하도록 학습
S340

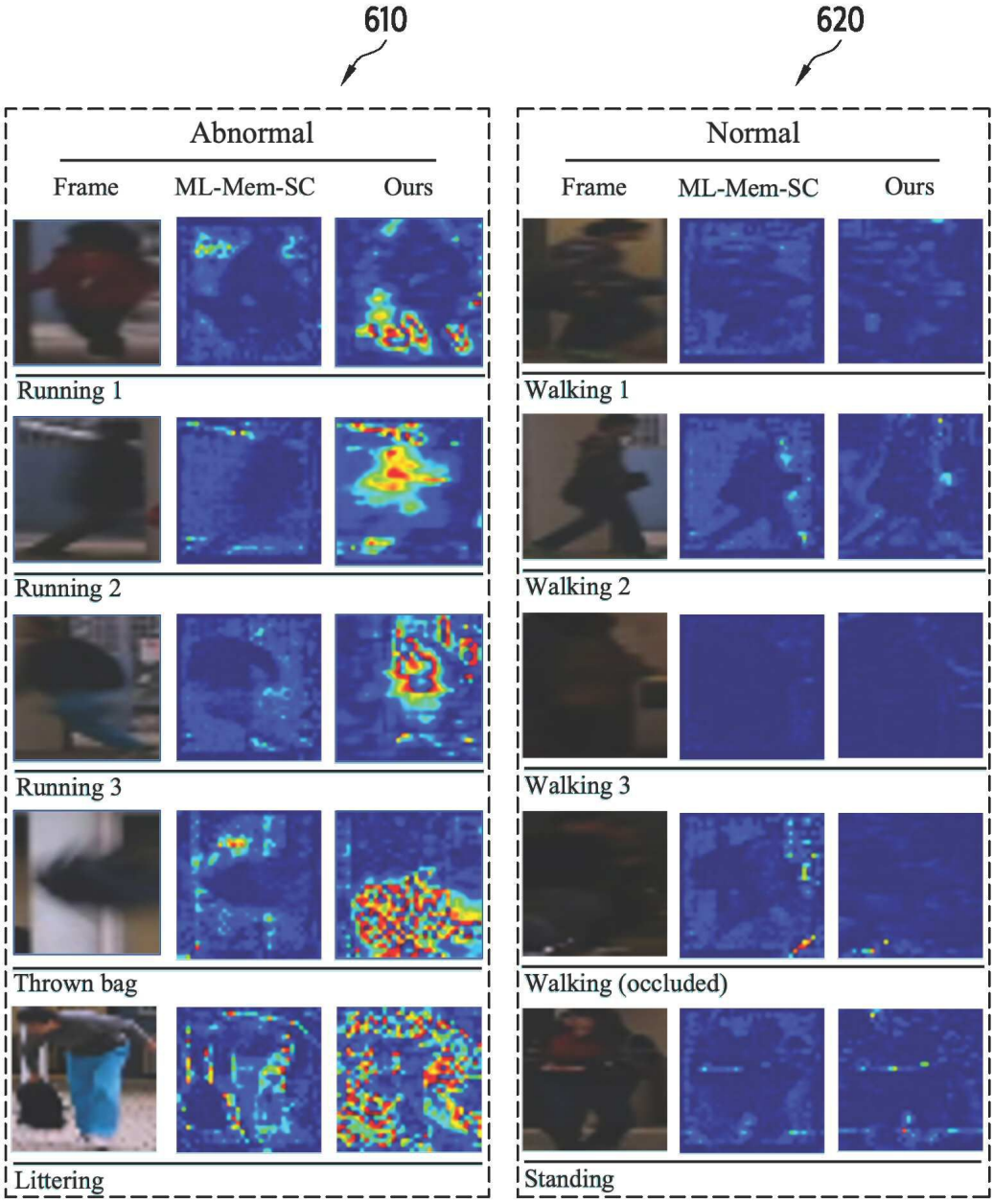
도면4


광학적 흐름 프레임들로부터 물체 위주로 분할
S410
오토인코더 모델을 통해 물체를 재구성하도록 학습
S420

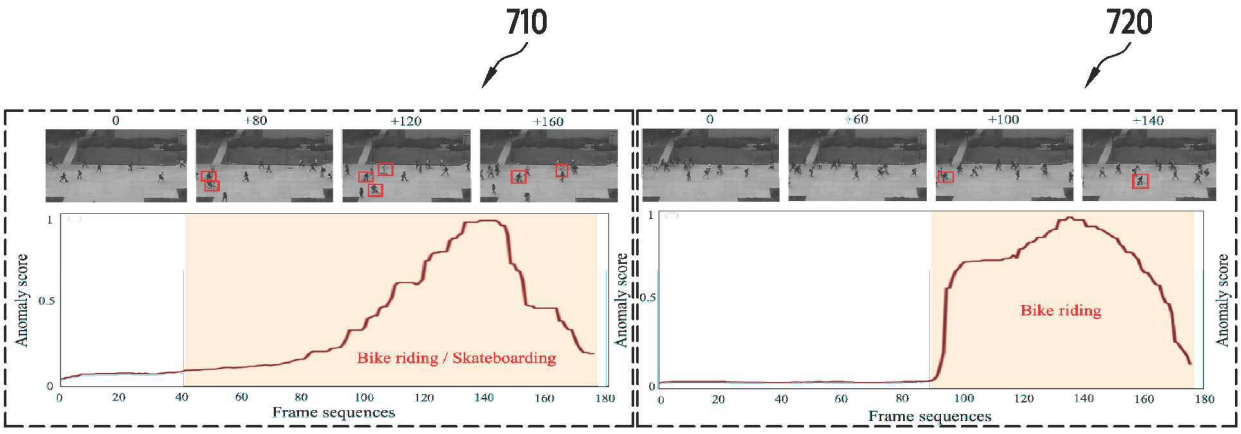
도면5


Object
MemAE
ML-MemAE-SC
Ours
Abnormal
Normal
510
520

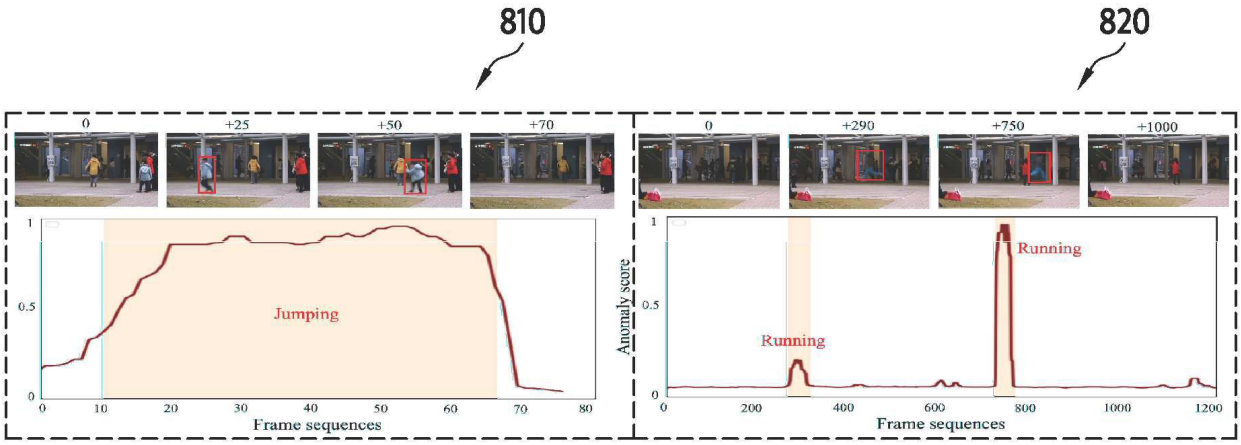
도면6


610
620
Abnormal
Frame
ML-Mem-SC
Ours
Running 1
Running 2
Running 3
Thrown bag
Littering
Normal
Frame
ML-Mem-SC
Ours
Walking 1
Walking 2
Walking 3
Walking (occluded)
Standing

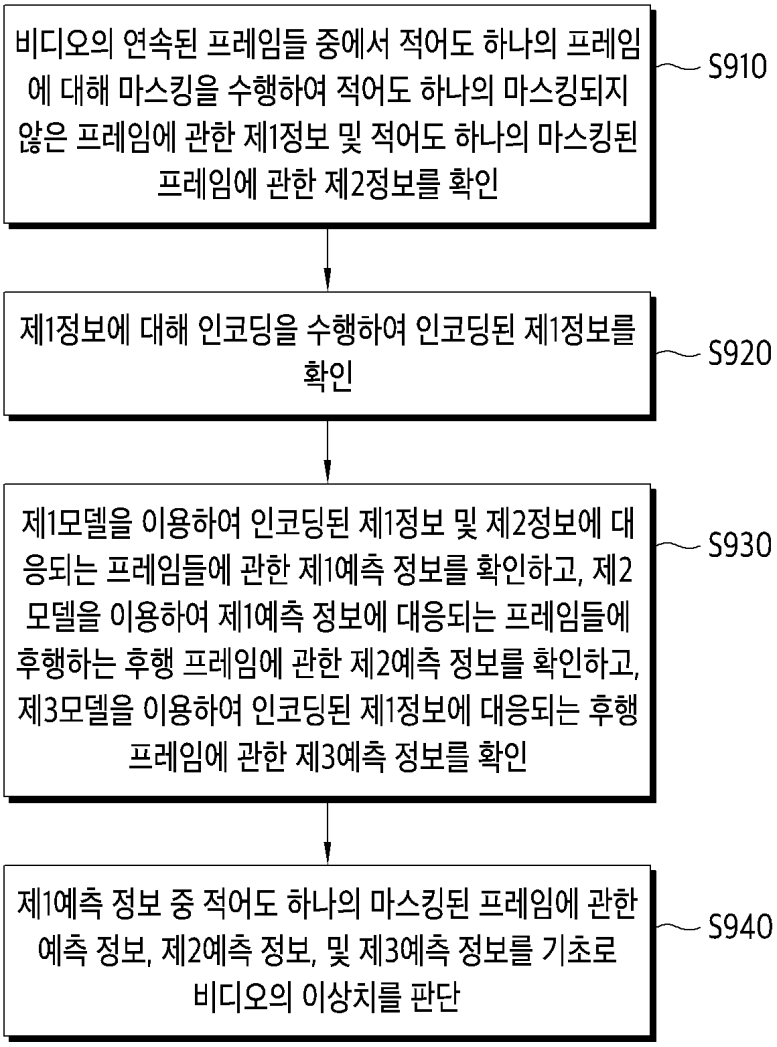
도면7


710
720
Anomaly score
Frame sequences
Bike riding / Skateboarding
Bike riding

도면8


810
820
0
+25
+50
+70
0
+290
+750
+1000
Jumping
Running
Running
Anomaly score
Frame sequences
Frame sequences

도면9


비디오의 연속된 프레임들 중에서 적어도 하나의 프레임에 대해 마스킹을 수행하여 적어도 하나의 마스킹되지 않은 프레임에 관한 제1정보 및 적어도 하나의 마스킹된 프레임에 관한 제2정보를 확인
S910
제1정보에 대해 인코딩을 수행하여 인코딩된 제1정보를 확인
S920
제1모델을 이용하여 인코딩된 제1정보 및 제2정보에 대응되는 프레임들에 관한 제1예측 정보를 확인하고, 제2모델을 이용하여 제1예측 정보에 대응되는 프레임들에 후행하는 후행 프레임에 관한 제2예측 정보를 확인하고, 제3모델을 이용하여 인코딩된 제1정보에 대응되는 후행 프레임에 관한 제3예측 정보를 확인
S930
제1예측 정보 중 적어도 하나의 마스킹된 프레임에 관한 예측 정보, 제2예측 정보, 및 제3예측 정보를 기초로 비디오의 이상치를 판단
S940