(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0176509
(43) 공개일자 2024년12월24일

(51) 국제특허분류(Int. Cl.)
G06V 20/52 (2022.01) *G06N 3/045* (2023.01)
G06N 3/0464 (2023.01) *G06V 10/44* (2022.01)
G06V 10/80 (2022.01) *G06V 10/82* (2022.01)
G06V 20/40 (2022.01)
(52) CPC특허분류
G06V 20/52 (2023.08)
G06N 3/045 (2023.01)
(21) 출원번호 10-2023-0077204
(22) 출원일자 2023년06월16일
심사청구일자 2023년06월16일

(71) 출원인
세종대학교산학협력단
서울특별시 광진구 능동로 209 (군자동, 세종대학교)
(72) 발명자
백성욱
서울특별시 강남구 영동대로 22, 813동 1303호(일원동, 디에이치 자이 개포)
이미영
서울특별시 강남구 언주로85길 13, 102호 (역삼동, 경남아파트)
(뒷면에 계속)
(74) 대리인
특허법인위더피플

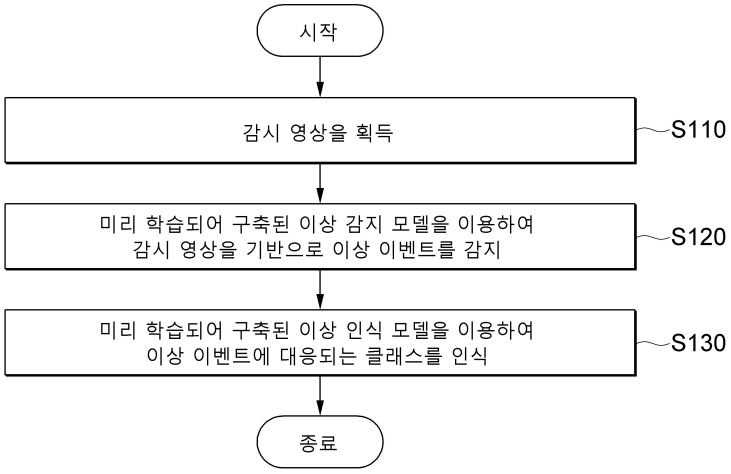
전체 청구항 수 : 총 15 항

(54) 발명의 명칭 **비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 방법, 이를 수행하는 장치 및 컴퓨터 프로그램**

(57) 요 약

본 발명의 바람직한 실시예에 따른 비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 방법, 이를 수행하는 장치 및 컴퓨터 프로그램은, 비전 트랜스포머 어텐션(vision transformer attention)과 멀티-리저보이어 에코 스테이트 네트워크(multi-reservoir echo state network)를 이용하여 감시 영상에서 이상을 감지하고 인식함으로써, 감시 영상에서 이상을 인식하는 정확도를 향상시킬 수 있다.

대 표 도 - 도2



(52) CPC특허분류
G06N 3/0464 (2023.01)
G06V 10/454 (2023.08)
G06V 10/80 (2023.08)
G06V 10/82 (2022.01)
G06V 20/44 (2022.01)

(72) 발명자
와심 울라
서울특별시 광진구 능동로 209 세종대학교 대양AI 센터 411호
사미 울라 칸
서울특별시 광진구 능동로 209 세종대학교 대양AI 센터 411호
김민제
서울특별시 광진구 천호대로110길 83-5, 101호(능동, 에코빌)
이수민
경기도 안산시 단원구 선부로 166, 120동 509호(선부동, 공작한양아파트)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711188556
과제번호	2023R1A2C1005788
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	개인기초연구(과기정통부)
연구과제명	커넥티드 비전 환경에서 범죄/사고 예방을 위한 이상행동 상황인식 기술
기 여 율	1/1
과제수행기관명	세종대학교
연구기간	2023.03.01 ~ 2024.02.29

명 세 서

청구범위

청구항 1

감시 영상을 획득하는 단계;

미리 학습되어 구축된 이상 감지 모델을 이용하여 상기 감시 영상을 기반으로 이상 이벤트를 감지하는 단계; 및

미리 학습되어 구축된 이상 인식 모델을 이용하여 상기 이상 이벤트에 대응되는 클래스를 인식하는 단계;

를 포함하는 비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 방법.

청구항 2

제1항에서,

상기 이상 이벤트 감지 단계는,

원-클래스(one-class) 심층 신경망을 포함하는 상기 이상 감지 모델을 이용하여 상기 감시 영상에서 상기 이상 이벤트를 감지하는 것으로 이루어지는,

비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 방법.

청구항 3

제2항에서,

상기 이상 감지 모델은,

기하학적 중앙값(geometric median)을 기반으로 하는 필터 가지치기(filter pruning)를 이용하여 크기가 압축된 모델인,

비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 방법.

청구항 4

제2항에서,

상기 클래스 인식 단계는,

비전 트랜스포머 어텐션(vision transformer attention)과 멀티-리저보이어 에코 스테이트 네트워크(multi-reservoir echo state network)를 포함하는 상기 이상 인식 모델을 이용하여 상기 이상 이벤트에 대응되는 클래스를 인식하는 것으로 이루어지는,

비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 방법.

청구항 5

제4항에서,

상기 클래스 인식 단계는,

상기 감시 영상에서 상기 비전 트랜스포머 어텐션을 통해 비전 트랜스포머 특징을 획득하고, 상기 비전 트랜스포머 특징을 기반으로 병목 어텐션(bottleneck attention) 메커니즘을 이용하여 정제된 특징을 획득하며, 상기 정제된 특징을 기반으로 상기 멀티-리저보이어 에코 스테이트 네트워크를 통해 상기 이상 이벤트에 대응되는 클래스를 인식하는 것으로 이루어지는,

비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 방법.

청구항 6

제5항에서,

상기 이상 인식 모델은,

멀티 헤드 셀프 어텐션 블록(multi-head self-attention block)과 완전 연결된 피드 포워드 고밀도 블록(fully connected feed-forward dense block)을 포함하고 상기 감시 영상에서 상기 비전 트랜스포머 특징을 획득하는 트랜스포머 인코더(transformer encoder) 모듈, 상기 비전 트랜스포머 특징을 기반으로 상기 정제된 특징을 획득하는 병목 어텐션 모듈(bottleneck attention module, BAM), 및 상기 정제된 특징을 기반으로 상기 이상 이벤트에 대응되는 클래스를 인식하는 상기 멀티-리저보이어 에코 스테이트 네트워크를 포함하는,

비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 방법.

청구항 7

제6항에서,

상기 멀티 헤드 셀프 어텐션 블록은,

선형(linear) 레이어, 셀프-어텐션(self-attention) 레이어, 연결(concatenation) 레이어, 및 최종 선형(final linear) 레이어를 포함하는,

비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 방법.

청구항 8

제1항 내지 제7항 중 어느 한 항에 기재된 비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 방법을 컴퓨터에서 실행시키기 위하여 컴퓨터 판독 가능한 저장 매체에 저장된 컴퓨터 프로그램.

청구항 9

비전 트랜스포머 어텐션(vision transformer attention)과 멀티-리저보이어 에코 스테이트 네트워크(multi-reservoir echo state network)를 이용하여 감시 영상에서 이상을 감지하고 인식하기 위한 하나 이상의 프로그램을 저장하는 메모리; 및

상기 메모리에 저장된 상기 하나 이상의 프로그램에 따라 상기 감시 영상에서 이상을 감지하고 인식하기 위한 동작을 수행하는 하나 이상의 프로세서;

를 포함하며,

상기 프로세서는,

감시 영상을 획득하고,

미리 학습되어 구축된 이상 감지 모델을 이용하여 상기 감시 영상을 기반으로 이상 이벤트를 감지하며,

미리 학습되어 구축된 이상 인식 모델을 이용하여 상기 이상 이벤트에 대응되는 클래스를 인식하는,

비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 장치.

청구항 10

제9항에서,

상기 프로세서는,

원-클래스(one-class) 심층 신경망을 포함하는 상기 이상 감지 모델을 이용하여 상기 감시 영상에서 상기 이상 이벤트를 감지하는,

비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 장치.

청구항 11

제10항에서,

상기 이상 감지 모델은,

기하학적 중앙값(geometric median)을 기반으로 하는 필터 가지치기(filter pruning)를 이용하여 크기가 압축된 모델인,

비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 장치.

청구항 12

제10항에서,

상기 프로세서는,

상기 비전 트랜스포머 어텐션과 상기 멀티-리저보이어 에코 스테이트 네트워크를 포함하는 상기 이상 인식 모델을 이용하여 상기 이상 이벤트에 대응되는 클래스를 인식하는,

비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 장치.

청구항 13

제12항에서,

상기 프로세서는,

상기 감시 영상에서 상기 비전 트랜스포머 어텐션을 통해 비전 트랜스포머 특징을 획득하고, 상기 비전 트랜스포머 특징을 기반으로 병목 어텐션(bottleneck attention) 메커니즘을 이용하여 정제된 특징을 획득하며, 상기 정제된 특징을 기반으로 상기 멀티-리저보이어 에코 스테이트 네트워크를 통해 상기 이상 이벤트에 대응되는 클래스를 인식하는,

비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 장치.

청구항 14

제13항에서,

상기 이상 인식 모델은,

멀티 헤드 셀프 어텐션 블록(multi-head self-attention block)과 완전 연결된 피드 포워드 고밀도 블록(fully connected feed-forward dense block)을 포함하고 상기 감시 영상에서 상기 비전 트랜스포머 특징을 획득하는 트랜스포머 인코더(transformer encoder) 모듈, 상기 비전 트랜스포머 특징을 기반으로 상기 정제된 특징을 획득하는 병목 어텐션 모듈(bottleneck attention module, BAM), 및 상기 정제된 특징을 기반으로 상기 이상 이벤트에 대응되는 클래스를 인식하는 상기 멀티-리저보이어 에코 스테이트 네트워크를 포함하는,

비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 장치.

청구항 15

제14항에서,

상기 멀티 헤드 셀프 어텐션 블록은,

선형(linear) 레이어, 셀프-어텐션(self-attention) 레이어, 연결(concatenation) 레이어, 및 최종 선형(final linear) 레이어를 포함하는,

비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 방법, 이를 수행하는 장치 및 컴퓨터 프로그램에 관한 것으로서, 더욱 상세하게는 감시 영상에서 이상을 감지하고

인식하는, 방법, 장치 및 컴퓨터 프로그램에 관한 것이다.

배 경 기 술

[0002] 통계에 따르면 2022년에는 전 세계에 450억 대 이상의 감시 카메라가 설치될 것으로 예상된다. 비디오 감시는 특히 법 집행, 교통 및 환경 모니터링 목적에 필수적이다. 이 기술은 교통 관리, 범죄 예방 및 수사 분야에서 필수적인 요소가 되고 있다. 영상에서 비정상적인 활동을 정확하고 즉각적으로 감지하고 인식하는 것은 광범위한 감시 애플리케이션으로 인해 머신 러닝 및 컴퓨터 비전 분야에서 오랫동안 해결해야 할 과제였다. 기존의 보안 감시 시스템은 실시간으로 생성되는 대량의 스트리밍 비디오를 따라잡을 수 없었다. 사람이 24시간 내내 이러한 감시 카메라를 수동으로 모니터링하는 것은 어렵고 번거로운 작업으로 오류가 발생할 수 있다. 그러나 비디오 콘텐츠 분석은 다양한 보안 감시 애플리케이션에 사용될 수 있으며, 그 중 하나는 비정상적인 이벤트를 감지하는 것이다. 정상 이벤트에 비해 비정상적인 행동은 거의 발생하지 않기 때문에 정상 세그먼트와 비정상 세그먼트 간의 데이터 불균형으로 인해 다른 유형의 비디오 분석에 비해 이상 행동 인식 및 감지가 더 어렵다. 따라서 현재 사용 가능한 감시 인프라를 적절히 활용하지 못하고 있으며, 감시 이벤트를 수동으로 분석하는 것은 비용이 많이 들고 인적 한계로 인해 오산이 발생하기 쉽기 때문에 지능형 컴퓨터 비전 알고리즘에 의존하는 자동 비디오 이상 감지 시스템이 절실히 요구된다.

발명의 내용

해결하려는 과제

[0003] 본 발명이 이루고자 하는 목적은, 비전 트랜스포머 어텐션(vision transformer attention)과 멀티-리저보이어 에코 스테이트 네트워크(multi-reservoir echo state network)를 이용하여 감시 영상에서 이상을 감지하고 인식하는, 비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 방법, 이를 수행하는 장치 및 컴퓨터 프로그램을 제공하는 데 있다.

[0004] 본 발명의 명시되지 않은 또 다른 목적들은 하기의 상세한 설명 및 그 효과로부터 용이하게 추론할 수 있는 범위 내에서 추가적으로 고려될 수 있다.

과제의 해결 수단

[0005] 상기의 기술적 과제를 달성하기 위한 본 발명의 바람직한 실시예에 따른 비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 방법은, 감시 영상을 획득하는 단계; 미리 학습되어 구축된 이상 감지 모델을 이용하여 상기 감시 영상을 기반으로 이상 이벤트를 감지하는 단계; 및 미리 학습되어 구축된 이상 인식 모델을 이용하여 상기 이상 이벤트에 대응되는 클래스를 인식하는 단계;를 포함한다.

[0006] 여기서, 상기 이상 이벤트 감지 단계는, 원-클래스(one-class) 심층 신경망을 포함하는 상기 이상 감지 모델을 이용하여 상기 감시 영상에서 상기 이상 이벤트를 감지하는 것으로 이루어질 수 있다.

[0007] 여기서, 상기 이상 감지 모델은, 기하학적 중앙값(geometric median)을 기반으로 하는 필터 가지치기(filter pruning)를 이용하여 크기가 압축된 모델일 수 있다.

[0008] 여기서, 상기 클래스 인식 단계는, 비전 트랜스포머 어텐션(vision transformer attention)과 멀티-리저보이어 에코 스테이트 네트워크(multi-reservoir echo state network)를 포함하는 상기 이상 인식 모델을 이용하여 상기 이상 이벤트에 대응되는 클래스를 인식하는 것으로 이루어질 수 있다.

[0009] 여기서, 상기 클래스 인식 단계는, 상기 감시 영상에서 상기 비전 트랜스포머 어텐션을 통해 비전 트랜스포머 특징을 획득하고, 상기 비전 트랜스포머 특징을 기반으로 병목 어텐션(bottleneck attention) 메커니즘을 이용하여 정제된 특징을 획득하며, 상기 정제된 특징을 기반으로 상기 멀티-리저보이어 에코 스테이트 네트워크를 통해 상기 이상 이벤트에 대응되는 클래스를 인식하는 것으로 이루어질 수 있다.

[0010] 여기서, 상기 이상 인식 모델은, 멀티 헤드 셀프 어텐션 블록(multi-head self-attention block)과 완전 연결된 피드 포워드 고밀도 블록(fully connected feed-forward dense block)을 포함하고 상기 감시 영상에서 상기 비전 트랜스포머 특징을 획득하는 트랜스포머 인코더(transformer encoder) 모듈, 상기 비전 트랜스포머 특징을 기반으로 상기 정제된 특징을 획득하는 병목 어텐션 모듈(bottleneck attention module, BAM), 및 상기 정제된 특징을 기반으로 상기 이상 이벤트에 대응되는 클래스를 인식하는 상기 멀티-리저보이어 에코 스테이트 네트워크를 포함할 수 있다.

[0011] 여기서, 상기 멀티 헤드 셀프 어텐션 블록은, 선형(linear) 레이어, 셀프-어텐션(self-attention) 레이어, 연결(concatenation) 레이어, 및 최종 선형(final linear) 레이어를 포함할 수 있다.

[0013] 상기의 기술적 과제를 달성하기 위한 본 발명의 바람직한 실시예에 따른 컴퓨터 프로그램은 컴퓨터 판독 가능한 저장 매체에 저장되어 상기한 비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 방법 중 어느 하나를 컴퓨터에서 실행시킨다.

[0015] 상기의 기술적 과제를 달성하기 위한 본 발명의 바람직한 실시예에 따른 비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 장치는, 비전 트랜스포머 어텐션(vision transformer attention)과 멀티-리저보이어 에코 스테이트 네트워크(multi-reservoir echo state network)를 이용하여 감시 영상에서 이상을 감지하고 인식하기 위한 하나 이상의 프로그램을 저장하는 메모리; 및 상기 메모리에 저장된 상기 하나 이상의 프로그램에 따라 상기 감시 영상에서 이상을 감지하고 인식하기 위한 동작을 수행하는 하나 이상의 프로세서;를 포함하며, 상기 프로세서는, 감시 영상을 획득하고, 미리 학습되어 구축된 이상 감지 모델을 이용하여 상기 감시 영상을 기반으로 이상 이벤트를 감지하며, 미리 학습되어 구축된 이상 인식 모델을 이용하여 상기 이상 이벤트에 대응되는 클래스를 인식한다.

[0016] 여기서, 상기 프로세서는, 원-클래스(one-class) 심층 신경망을 포함하는 상기 이상 감지 모델을 이용하여 상기 감시 영상에서 상기 이상 이벤트를 감지할 수 있다.

[0017] 여기서, 상기 이상 감지 모델은, 기하학적 중앙값(geometric median)을 기반으로 하는 필터 가지치기(filter pruning)를 이용하여 크기가 압축된 모델일 수 있다.

[0018] 여기서, 상기 프로세서는, 상기 비전 트랜스포머 어텐션과 상기 멀티-리저보이어 에코 스테이트 네트워크를 포함하는 상기 이상 인식 모델을 이용하여 상기 이상 이벤트에 대응되는 클래스를 인식할 수 있다.

[0019] 여기서, 상기 프로세서는, 상기 감시 영상에서 상기 비전 트랜스포머 어텐션을 통해 비전 트랜스포머 특징을 획득하고, 상기 비전 트랜스포머 특징을 기반으로 병목 어텐션(bottleneck attention) 메커니즘을 이용하여 정제된 특징을 획득하며, 상기 정제된 특징을 기반으로 상기 멀티-리저보이어 에코 스테이트 네트워크를 통해 상기 이상 이벤트에 대응되는 클래스를 인식할 수 있다.

[0020] 여기서, 상기 이상 인식 모델은, 멀티 헤드 셀프 어텐션 블록(multi-head self-attention block)과 완전 연결된 피드 포워드 고밀도 블록(fully connected feed-forward dense block)을 포함하고 상기 감시 영상에서 상기 비전 트랜스포머 특징을 획득하는 트랜스포머 인코더(transformer encoder) 모듈, 상기 비전 트랜스포머 특징을 기반으로 상기 정제된 특징을 획득하는 병목 어텐션 모듈(bottleneck attention module, BAM), 및 상기 정제된 특징을 기반으로 상기 이상 이벤트에 대응되는 클래스를 인식하는 상기 멀티-리저보이어 에코 스테이트 네트워크를 포함할 수 있다.

[0021] 여기서, 상기 멀티 헤드 셀프 어텐션 블록은, 선형(linear) 레이어, 셀프-어텐션(self-attention) 레이어, 연결(concatenation) 레이어, 및 최종 선형(final linear) 레이어를 포함할 수 있다.

발명의 효과

[0022] 본 발명의 바람직한 실시예에 따른 비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 방법, 이를 수행하는 장치 및 컴퓨터 프로그램에 의하면, 비전 트랜스포머 어텐션(vision transformer attention)과 멀티-리저보이어 에코 스테이트 네트워크(multi-reservoir echo state network)를 이용하여 감시 영상에서 이상을 감지하고 인식함으로써, 감시 영상에서 이상을 인식하는 정확도를 향상시킬 수 있다.

[0023] 본 발명의 효과들은 이상에서 언급한 효과들로 제한되지 않으며, 언급되지 않은 또 다른 효과들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0024] 도 1은 본 발명의 바람직한 실시예에 따른 비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크

를 이용한 이상 인식 장치를 설명하기 위한 블록도이다.

도 2는 본 발명의 바람직한 실시예에 따른 비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 방법을 설명하기 위한 흐름도이다.

도 3은 본 발명의 바람직한 실시예에 따른 이상 인식 동작을 설명하기 위한 도면이다.

도 4는 본 발명의 바람직한 실시예에 따른 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프레임워크의 구조를 설명하기 위한 도면이다.

도 5는 도 4에 도시한 멀티-리저보이어 에코 스테이트 네트워크의 구조를 설명하기 위한 도면이다.

도 6은 도 4에 도시한 병목 어텐션 모듈의 구조를 설명하기 위한 도면이다.

도 7은 본 발명의 바람직한 실시예에 따른 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프레임워크와 최첨단(SOTA) 기법을 비교한 도면이다.

도 8은 본 발명의 바람직한 실시예에 따른 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프레임워크의 혼동 행렬을 나타내는 도면으로, 도 8의 (a)는 UCF-Crime 데이터 세트를 나타내고, 도 8의 (b)는 LAD-2000 데이터 세트를 나타낸다.

도 9는 본 발명의 바람직한 실시예에 따른 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프레임워크의 클래스별 성능을 나타내는 도면으로, 도 9의 (a)는 UCF-Crime 데이터 세트에서 2스트림 네트워크와 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프레임워크의 클래스별 정확도 성능 비교를 나타내고, 도 9의 (b)는 LAD-2000 데이터 세트에서 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프레임워크의 클래스별 정확도를 나타낸다.

도 10은 본 발명의 바람직한 실시예에 따른 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프레임워크와 다른 최신 기법을 ROC 파라미터를 기준으로 UCF-Crime 데이터 세트에서 비교한 결과를 나타내는 도면이다.

도 11은 본 발명의 바람직한 실시예에 따른 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프레임워크가 UCF-Crime 및 LAD-2000 데이터 세트에서 이상을 인식한 예시를 나타내는 도면으로, UCF-Crime 데이터 세트의 (a)는 방화를 나타내고, UCF-Crime 데이터 세트의 (b)는 폭발을 나타내며, UCF-Crime 데이터 세트의 (c)는 교통 사고를 나타내고, UCF-Crime 데이터 세트의 (d)는 폭행을 나타내며, LAD-2000 데이터 세트의 (a)는 떨어짐을 나타내고, LAD-2000 데이터 세트의 (b)는 넘어짐을 나타내며, LAD-2000 데이터 세트의 (c)는 폭력을 나타내고, LAD-2000 데이터 세트의 (d)는 절도를 나타낸다.

발명을 실시하기 위한 구체적인 내용

[0025] 이하, 첨부된 도면을 참조하여 본 발명의 실시예를 상세히 설명한다. 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나, 본 발명은 이하에서 게시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 발명의 게시가 완전하도록 하고, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다. 명세서 전체에 걸쳐 동일 참조 부호는 동일 구성 요소를 지칭한다.

[0026] 다른 정의가 없다면, 본 명세서에서 사용되는 모든 용어(기술 및 과학적 용어를 포함)는 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에게 공통적으로 이해될 수 있는 의미로 사용될 수 있을 것이다. 또한, 일반적으로 사용되는 사전에 정의되어 있는 용어들은 명백하게 특별히 정의되어 있지 않는 한 이상적으로 또는 과도하게 해석되지 않는다.

[0027] 본 명세서에서 "제1", "제2" 등의 용어는 하나의 구성 요소를 다른 구성 요소로부터 구별하기 위한 것으로, 이들 용어들에 의해 권리범위가 한정되어서는 아니 된다. 예컨대, 제1 구성 요소는 제2 구성 요소로 명명될 수 있고, 유사하게 제2 구성 요소도 제1 구성 요소로 명명될 수 있다.

[0028] 본 명세서에서 각 단계들에 있어 식별부호(예컨대, a, b, c 등)는 설명의 편의를 위하여 사용되는 것으로 식별부호는 각 단계들의 순서를 설명하는 것이 아니며, 각 단계들은 문맥상 명백하게 특정 순서를 기재하지 않는 이상 명기된 순서와 다르게 일어날 수 있다. 즉, 각 단계들은 명기된 순서와 동일하게 일어날 수도 있고 실질적으로 동시에 수행될 수도 있으며 반대의 순서대로 수행될 수도 있다.

[0029] 본 명세서에서, "가진다", "가질 수 있다", "포함한다" 또는 "포함할 수 있다" 등의 표현은 해당 특징(예컨대,

수치, 기능, 동작, 또는 부품 등의 구성 요소)의 존재를 가리키며, 추가적인 특징의 존재를 배제하지 않는다.

[0032] 이하에서 첨부한 도면을 참조하여 본 발명에 따른 비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 방법, 이를 수행하는 장치 및 컴퓨터 프로그램의 바람직한 실시예에 대해 상세하게 설명한다.

[0034] 먼저, 도 1을 참조하여 본 발명의 바람직한 실시예에 따른 비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 장치에 대하여 설명한다.

[0035] 도 1은 본 발명의 바람직한 실시예에 따른 비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 장치를 설명하기 위한 블록도이다.

[0036] 도 1을 참조하면, 본 발명의 바람직한 실시예에 따른 비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 장치(이하 '이상 인식 장치'라 한다)(100)는 비전 트랜스포머 어텐션(vision transformer attention)과 멀티-리저보이어 에코 스테이트 네트워크(multi-reservoir echo state network)를 이용하여 감시 영상에서 이상을 감지하고 인식할 수 있다. 이에 따라, 본 발명은 종래 기술보다 감시 영상에서 이상을 인식하는 정확도를 향상시킬 수 있다.

[0037] 즉, 이상 이벤트 인지는 인명과 재산 피해를 줄이기 위해 즉각적인 대응이 필요하지만, 기존의 자동화된 시스템은 영상의 시간적 영역에 대한 고려와 공간 정보의 중요한 역할을 무시한 채 제한적인 성능을 보이고 있다. 또한, 현재의 감시 시스템은 이상 이벤트를 감지할 수는 있지만, 이상 이벤트 감지와 해석을 동시에 수행할 수 있는 완전 자동 감시 기법이 없기 때문에 이상 이벤트의 특성을 인식하고 적절한 대응책을 선택하기 위해서는 사람의 개입이 필요하다. 따라서 본 발명은 스마트시티 감시 영상에서 이상 행위를 감지하고 해석할 수 있는 비전 트랜스포머 이상 인식(Vision Transformer Anomaly Recognition, ViT-ARN)이라는 프레임워크를 제안한다. 이 프레임워크는 두 단계로 구성되며, 첫 번째 단계에서는 감시 환경에서 이상을 감지하기 위해 맞춤형 경량 원-클래스(one-class) 심층 신경망을 개발하고, 두 번째 단계에서는 감지된 이상을 해당 클래스로 추가 분류한다. 이상 감지 모델의 크기는 리소스가 제한된 장치에 쉽게 적용할 수 있도록 기하학적 중앙값(geometric median)에 기반한 필터 가지치기(filter pruning) 전략을 사용하여 압축된다. 이상 분류는 비전 트랜스포머 특징을 기반으로 이루어지며, 그 다음에는 병목 어텐션(bottleneck attention) 메커니즘을 통해 표현력을 향상시켜 정제된 특징을 획득한다. 정제된 특징은 기물 파손 및 도로 사고와 같은 실제 이상 현상을 자세히 분석하기 위해 멀티-리저보이어 에코 스테이트 네트워크로 전달된다. 두 데이터 세트의 총 858개와 1,600개의 비디오가 제안된 모델을 학습하는 데 사용되었으며, 290개와 400개의 테스트 비디오로 구성된 LD-2000 및 UCF-Crime 데이터 세트에 대한 광범위한 실험을 통해 프레임워크가 다른 최신 접근 방식을 능가하는 10.14%와 3%의 정확도로 이상을 더 효과적으로 인식할 수 있음을 알 수 있다.

[0039] 이를 위해, 이상 인식 장치(100)는 하나 이상의 프로세서(110), 컴퓨터 판독 가능한 저장 매체(130) 및 통신 버스(150)를 포함할 수 있다.

[0040] 프로세서(110)는 이상 인식 장치(100)가 동작하도록 제어할 수 있다. 예컨대, 프로세서(110)는 컴퓨터 판독 가능한 저장 매체(130)에 저장된 하나 이상의 프로그램(131)을 실행할 수 있다. 하나 이상의 프로그램(131)은 하나 이상의 컴퓨터 실행 가능 명령어를 포함할 수 있으며, 컴퓨터 실행 가능 명령어는 프로세서(110)에 의해 실행되는 경우 이상 인식 장치(100)로 하여금 비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용하여 감시 영상에서 이상을 감지하고 인식하기 위한 동작을 수행하도록 구성될 수 있다.

[0041] 컴퓨터 판독 가능한 저장 매체(130)는 비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용하여 감시 영상에서 이상을 감지하고 인식하기 위한 컴퓨터 실행 가능 명령어 내지 프로그램 코드, 프로그램 데이터 및/또는 다른 적합한 형태의 정보를 저장하도록 구성된다. 컴퓨터 판독 가능한 저장 매체(130)에 저장된 프로그램(131)은 프로세서(110)에 의해 실행 가능한 명령어의 집합을 포함한다. 일 실시예에서, 컴퓨터 판독 가능한 저장 매체(130)는 메모리(랜덤 액세스 메모리와 같은 휘발성 메모리, 비휘발성 메모리, 또는 이들의 적절한 조합), 하나 이상의 자기 디스크 저장 디바이스들, 광학 디스크 저장 디바이스들, 플래시 메모리 디바이

스들, 그 밖에 이상 인식 장치(100)에 의해 액세스되고 원하는 정보를 저장할 수 있는 다른 형태의 저장 매체, 또는 이들의 적합한 조합일 수 있다.

[0042] 통신 버스(150)는 프로세서(110), 컴퓨터 판독 가능한 저장 매체(130)를 포함하여 이상 인식 장치(100)의 다른 다양한 컴포넌트들을 상호 연결한다.

[0043] 이상 인식 장치(100)는 또한 하나 이상의 입출력 장치를 위한 인터페이스를 제공하는 하나 이상의 입출력 인터페이스(170) 및 하나 이상의 통신 인터페이스(190)를 포함할 수 있다. 입출력 인터페이스(170) 및 통신 인터페이스(190)는 통신 버스(150)에 연결된다. 입출력 장치(도시하지 않음)는 입출력 인터페이스(170)를 통해 이상 인식 장치(100)의 다른 컴포넌트들에 연결될 수 있다.

[0046] 그러면, 도 2 및 도 3을 참조하여 본 발명의 바람직한 실시예에 따른 비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 방법에 대하여 설명한다.

[0047] 도 2는 본 발명의 바람직한 실시예에 따른 비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 이용한 이상 인식 방법을 설명하기 위한 흐름도이고, 도 3은 본 발명의 바람직한 실시예에 따른 이상 인식 동작을 설명하기 위한 도면이다.

[0048] 도 2를 참조하면, 이상 인식 장치(100)의 프로세서(110)는 감시 영상을 획득할 수 있다(S110).

[0050] 그런 다음, 프로세서(110)는 미리 학습되어 구축된 이상 감지 모델을 이용하여 감시 영상을 기반으로 이상 이벤트를 감지할 수 있다(S120).

[0051] 즉, 프로세서(110)는 도 3에 도시된 바와 같이, 원-클래스 심층 신경망을 포함하는 이상 감지 모델을 이용하여 감시 영상에서 이상 이벤트를 감지할 수 있다.

[0052] 여기서, 이상 감지 모델은 기하학적 중앙값을 기반으로 하는 필터 가지치기를 이용하여 크기가 압축된 모델일 수 있다.

[0054] 그런 다음, 프로세서(110)는 미리 학습되어 구축된 이상 인식 모델을 이용하여 이상 이벤트에 대응되는 클래스를 인식할 수 있다(S130).

[0055] 여기서, 이상 인식 모델은 비전 트랜스포머 어텐션과 멀티-리저보이어 에코 스테이트 네트워크를 포함할 수 있다.

[0056] 즉, 프로세서(110)는 감시 영상에서 비전 트랜스포머 어텐션을 통해 비전 트랜스포머 특징을 획득하고, 비전 트랜스포머 특징을 기반으로 병목 어텐션(bottleneck attention) 메커니즘을 이용하여 정제된 특징을 획득하며, 정제된 특징을 기반으로 멀티-리저보이어 에코 스테이트 네트워크를 통해 이상 이벤트에 대응되는 클래스를 인식할 수 있다.

[0057] 이를 위해, 이상 인식 모델은 도 3에 도시된 바와 같이, 트랜스포머 인코더(transformer encoder) 모듈, 병목 어텐션 모듈(bottleneck attention module, BAM), 및 멀티-리저보이어 에코 스테이트 네트워크를 포함할 수 있다. 트랜스포머 인코더 모듈은 멀티 헤드 셀프 어텐션 블록(multi-head self-attention block)과 완전 연결된 피드 포워드 고밀도 블록(fully connected feed-forward dense block)을 포함하고, 감시 영상에서 비전 트랜스포머 특징을 획득할 수 있다. 여기서, 멀티 헤드 셀프 어텐션 블록은 선형(linear) 레이어, 셀프-어텐션(self-attention) 레이어, 연결(concatenation) 레이어, 및 최종 선형(final linear) 레이어를 포함할 수 있다. 병목 어텐션 모듈(BAM)은 비전 트랜스포머 특징을 기반으로 정제된 특징을 획득할 수 있다. 멀티-리저보이어 에코 스테이트 네트워크는 정제된 특징을 기반으로 이상 이벤트에 대응되는 클래스를 인식할 수 있다.

[0060] 그러면, 도 4 내지 도 11을 참조하여 본 발명의 바람직한 실시예에 따른 비전 트랜스포머 어텐션과 멀티-리저보

이어 에코 스테이트 네트워크를 이용한 이상 인식 동작에 대하여 보다 자세하게 설명한다.

[0061] 1. 개요

[0062] 현재 비정상적인 이벤트에 대한 주석이 불충분하고 해상도가 낮은 비디오, 두 개 이상의 클래스 간에 큰 클래스 내/클래스 간 차이가 있기 때문에 비정상적인 이벤트를 자동으로 인식하는 것은 어려운 작업이다. 여러 연구자들이 확장된 칼만 필터 하이브리드 서포트 벡터 머신(expanded Kalman filter hybrid support vector machine), 스택형 자동 인코더(stacked autoencoder), 기울기 히스토그램(histogram of gradients, HOG) 등 다양한 접근 방식을 사용해 교통사고를 조사해 이상을 감지했다. 그러나 이러한 알고리즘은 다른 비정상적인 행동을 분류하는 데는 제한적인 성능을 보였기 때문에 실제 감시를 위한 이러한 기법의 성능은 제한적이다. 또한, 여러 연구에서 시공간 자동 인코더가 포함된 컨볼루션 네트워크를 사용하여 감시 비디오에서 정상 이벤트와 비정상 이벤트만 분류함으로써 이상 행위 감지를 이진 분류 문제로 처리했다. 그러나 이러한 방법은 강도, 싸움, 폭발, 학대, 폭행, 기물 파손과 같은 다양한 유형의 이상 현상을 구분하는 데 적합한 솔루션을 제공하지 못한다.

[0063] 빅데이터 세트에는 모델이 다양한 정상 및 비정상 패턴을 학습할 수 있는 시간적 주석이 부족하고 학대, 절도, 충격, 절도, 강도와 같은 비정상 행동의 시간적 범위가 넓기 때문에 비정상 감지 및 인식은 복잡한 작업이다. 종래의 대부분의 기술은 이상 행위 인식보다는 이상 행위 감지에 초점을 맞추고 있으며, 그 결과 이상 행위 인식 분야에서는 상당한 진전이 없었다. 대부분의 연구자들은 이상 인식을 별도의 과제로 간주하고 있으며, 비디오 수준의 지도와 시각적 인식 작업 기반 백본을 사용하는 행동 인식 기법을 통해 모델을 비디오에 대해 학습시킨다. TCNN 및 C3D와 같은 백본 모델은 기본 집계 알고리즘을 사용하여 비디오 세그먼트 특징을 추출하고 트리밍되지 않은 전체 비디오를 정상 및 비정상 샘플로 분류하기 때문에 이러한 방법론은 최적이 아니다. 그 결과, 비정상을 분류하는 것은 여전히 어려운 과정이다.

[0064] 앞서 언급한 최신 접근법의 한계에 착안하여, 본 발명의 목표는 다양하고 까다로운 이상 데이터 세트의 시공간적 특징을 학습하고 그 유형을 정확하게 인식하는 것이다. 본 발명에 따른 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프레임워크에서는 비전 트랜스포머를 사용하여 비디오 프레임의 특징을 추출하고 멀티-리저보이어 에코 스테이트 네트워크(MrESN)를 통해 시간 정보를 학습하여 감시 영상에서 이상을 인식할 수 있도록 한다. 본 발명에 따른 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프레임워크는 기존 기술과 경쟁할 수 있을 만큼 효율적이고 강력하며, 다중 클래스 분류를 위한 다양한 대규모의 까다로운 이상 데이터 세트에서 최첨단 성능을 발휘한다. 본 발명의 주요 기여는 다음과 같이 요약할 수 있다.

[0065] - 본 발명은 완전 자동화된 감시가 가능한 모델을 개발하기 위해 약한 지도 패러다임에 기반한 이상 감지 및 인식을 위한 새로운 프레임워크인 비전 트랜스포머 이상 인식 네트워크(ViT-ARN)를 제안한다. 제안된 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프레임워크는 리소스가 제한된 디바이스에 효과적으로 적용하기 위해 압축된 단일 클래스 분류 모델을 이상 감지에 사용하고, 이상 인식에는 순차적 학습 모델과 어텐션을 갖춘 비전 트랜스포머를 사용하여 이상 활동을 효과적으로 분류할 수 있다.

[0066] - 기하학적 중앙값을 통한 필터 가지치기(filter pruning via geometric median, FPGM)를 사용하여 원본 이상 행위 감지 모델의 크기를 압축하고 줄인다. 이 기술은 모델의 크기를 약 3배로 줄여 비용 효율적이며 리소스가 제한된 지능형 보안 감시 시스템에 쉽게 적용할 수 있다.

[0067] - 비정상적인 활동을 인식하기 위해 비전 트랜스포머를 특징 추출기로 사용하고, 차별적인 비디오 특징을 표현하기 위해 멀티-리저보이어 에코 스테이트 네트워크(MrESN)를 사용한다. 우리가 아는 한, 복잡한 이상 활동의 동적 특성을 효율적으로 포착하기 위해 어텐션 메커니즘을 갖춘 멀티-리저보이어를 비디오 이상 활동 인식에 사용한 접근 방식은 이것이 처음이다.

[0068] - 두 개의 대규모 벤치마크 이상 행동 데이터 세트에서 제안한 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프레임워크를 검증하고, 기존 모델과 비교하여 최첨단 이상 행동 인식 성능을 달성했다.

[0070] 2. 관련 연구

[0071] 실제 감시 영상에서 이상을 감지하고 인식하는 것은 어려운 작업이지만, 실제 감시, 교통 및 보안 시스템과 범죄 현장 조사에 다양하게 적용되고 있어 활발히 연구되고 있는 분야이다. 이상 영상 감지 기법은 기본적으로 전통적인 머신 러닝과 딥 러닝 기법으로 나눌 수 있지만, 머신 러닝 기법은 성능이 매우 제한적이기 때문에 최

근 문헌에서 많이 활용되지 않고 있다. 따라서 본 발명에서는 머신 러닝 기법에 대한 설명은 생략하고 대신 최신 딥 러닝 모델에 초점을 맞춘다. 이상 감지를 위한 딥러닝 모델은 크게 (i)하이브리드 모델과 (ii)3D CNN 모델로 분류할 수 있다. 다음 섹션에서는 이러한 기법에 대해 자세히 설명한다.

[0072] 2.1. 하이브리드 이상 감지 기법

[0073] 기존의 이상 감지 접근 방식은 주로 정상 이벤트만을 학습에 포함하는 비지도 학습 기법에 집중해 왔다. 특징 추출 기법과 마찬가지로 기존 접근 방식은 수작업 특징 기반 접근 방식과 심층 특징 기반 접근 방식의 두 가지로 나눌 수 있다. 전자의 경우 수작업으로 만든 특징은 이상을 감지하기에 부적절하기 때문에 복잡한 영상 감시 장면을 표현하는 데 한계가 있다. 그러나 딥 러닝 모델에 기반한 몇 가지 이상 감지 접근 방식은 효과적인 성능을 보여주었다. 정상 데이터를 둘러싸고 있는 중간 수준의 하이퍼플레인을 계산하기 위해 원-클래스 SVM 접근법과 비교하여 일련의 원-클래스 딥 접근법이 소개되었다. 기초 연구 논문에서 저자들은 원-클래스 SVM 모델을 학습하고 사용하여 이상 점수를 얻기 위한 스택형 자동 인코더 메커니즘을 제시했다. Hasan 등은 시각 프레임의 재구성 오류로 렌더링되는 비정상 이벤트를 분류하는 컨볼루션 기반 자동 인코더를 채택했다. 또 다른 방법은 서포트 벡터 데이터에 대한 설명을 기반으로 모양과 동작 단서를 결합하여 시각 프레임에서 비정상적인 행동과 객체를 구분했다. Luo 등은 시간적으로 일관된 예비 코딩 접근법을 도입하여 유사한 인접 프레임의 재구성 계수 특징 간의 유사성을 보장하고 시간적으로 일관된 예비 코딩 접근법을 스택형 RNN과 매핑하여 이상을 감지했다. Gong 등은 비디오 프레임에서 비정상적인 동작을 감지하기 위해 메모리 증강 자동 인코더 메커니즘을 제시했다. Zhang 등은 지역 민감 해싱 필터 기법을 채택하여 비정상적인 활동을 메싱하고 분류했다.

[0074] 최근 듀얼 스트림 기반 네트워크는 이상 감지에 사용하기 위해 많은 관심을 받고 있다. 예를 들어, Ullah 등은 감시 이상 감지를 위한 듀얼 스트림 지능형 시스템을 제안했다. 첫 번째 스트림에서는 순차적 모델에서 전달된 2D 정제된 최적 특징을 사용했고, 두 번째 스트림에서는 효율적인 이상 감지를 위해 3D 외관 특징을 사용했다. 또 다른 기반 라인 접근법에서는 이상 감지를 위한 이중 특징 기반 모델을 제안했다. 이 연구에서는 모션에는 3D 인코더를 사용하고 입력 비디오의 예측 프레임을 재구성하는 데는 2D 디코더를 사용했다. Huang 등은 셀프-어텐션 판별자와 예측자로 구성된 이상 감지를 위한 자기 지도형 엔드-투-엔드 모델을 도입했다. 예측자는 시간적 종속성을 포착하고 입력 비디오 프레임의 예측 품질을 향상시키는 데 사용되며, 판별자는 회전 및 참/거짓 판별을 수행했다. Chongke 등은 노이즈 제거 자동 인코더와 이상 감지를 위한 샤플리 부가 설명 기법을 결합한 설명 가능한 딥러닝 모델을 제안했다. 이러한 기법의 주요 한계는 훈련 데이터 세트에 불규칙성이 빈번하게 발생하여 이상이나 이상값뿐만 아니라 오해의 소지가 있는 특징 표현을 생성하기 때문이다. 데이터 재구성의 목적 함수는 이상 감지가 아니라 모델의 크기를 줄이거나 데이터를 압축하는 데 있다. 따라서 이러한 표현은 기본 규칙성을 일반적인 방식으로 표현하기 때문에 불규칙성을 감지하는 데 최적화되어 있지 않다.

[0075] 2.2. 3D CNN 기반 이상 감지

[0076] 최근의 또 다른 이상 감지 연구 그룹에서는 시각 프레임의 시퀀스를 표현하기 위해 시공간 패턴을 표현하기 위해 3D CNN을 채택했다. Tran 등은 대규모 행동 인식 데이터 세트에 심층 3D CNN을 적용하여 가치 있는 결과를 얻었다. Varol 등은 동작 인식을 위한 장기적인 시간적 특징을 학습하고 정확도와 성능 측면에서 최첨단 2D CNN 모델을 개선하기 위해 3D CNN 아키텍처를 채택했다. 또 다른 연구에서 저자들은 3D CNN의 고유한 잔존 연결을 확장하여 동작 인식 데이터 세트에 대한 경쟁력 있는 성능을 보고했다. 그러나 이상 감지에 이러한 3D CNN 아키텍처를 사용할 때의 문제점은 높은 수준의 특징을 기반으로 하는 경우가 많기 때문에 이상 감지에 직접 적용하기 어렵고 지도 방식으로 학습해야 한다는 점이다. Lee 등은 비디오 기반 이상 감지 및 인식을 위한 시공간적 적대적 학습 아키텍처를 제안했는데, 여기에는 3D CNN 기반 판별기가 있는 LSTM 기반 생성기가 포함되어 있다. Khan 등은 시공간적 적대적 학습을 기반으로 한 이전 연구를 3D CNN을 사용한 단일 클래스 분류기로 확장했다. 이러한 유형의 네트워크는 주로 입력 비디오 프레임을 재구성하는 데 사용되며, 시공간 컨볼루션 네트워크는 원본 비디오 프레임에서 장면을 구별하는 데 사용된다. 이러한 접근 방식을 사용한 시퀀스 데이터 분석은 이상 감지에만 제한적으로 적용되며, 이러한 방법의 성능은 감시 데이터에서 이상을 인식하는 데는 효율성이 떨어진다.

[0077] 3. 본 발명에 따른 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프레임워크

[0078] 도 4는 본 발명의 바람직한 실시예에 따른 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프레임워크의 구조를 설명하기 위한 도면이다.

[0079] 본 발명에서는 이상 감지 및 인식을 위한 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프레임워크를

제안한다. 비전 트랜스포머 이상 인식 네트워크(ViT-ARN)의 전체 흐름은 도 4와 같으며, (i)원-클래스 심층 신경망을 사용하여 수행하는 이상 감지, (ii)시공간 모델을 사용하여 감지된 이상의 유형을 인식하는 두 가지 주요 단계로 구성된다. 제안된 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프레임워크의 전체 아키텍처는 다음 섹션에서 자세히 설명한다.

[0080] 3.1. 원-클래스 이상 감지 모델

[0081] 원-클래스 분류(one-class classification, OCC) 문제는 최근 컴퓨터 비전의 다양한 분야에서 상당한 관심을 받고 있다. 원-클래스 분류(OCC)는 한 클래스의 인스턴스만 분류해야 하는 특정 유형의 다중 및 이진 분류 작업이다. 원-클래스 분류(OCC)의 주요 목표는 훈련 인스턴스와 잠재적 이상값을 구별할 수 있도록 훈련 인스턴스의 특성을 파악하는 것이다. 다중 또는 이진 분류와 달리 원-클래스 분류(OCC)는 일반적으로 관심 클래스인 한 클래스의 인스턴스만 검사하고 분석한다. 긍정적인 클래스의 레이블이 지정된 인스턴스는 원-클래스 분류(OCC) 문제 시나리오에서 사용할 수 없거나 기존 머신 러닝 알고리즘을 훈련하기에 충분한 수로 제공되지 않는다. 많은 실제 감시 데이터 세트는 부분적으로 레이블이 지정되거나 약하게 지도되며, 결과적으로 레이블이 지정되지 않은 샘플의 수가 레이블이 지정된 샘플의 수에 비해 많을 수 있으며, 이는 일반적으로 데이터 세트의 크기가 크기 때문에 기존 분류 기법의 학습 시간을 증가시킨다. 현재 이러한 분류 문제를 해결하기 위한 한 가지 해결책은 관찰되지 않은 이벤트를 정상 또는 비정상(이상) 이벤트로 인식하는 원-클래스 분류(OCC) 기법을 사용하는 것이다. 원-클래스 분류(OCC)는 분류 패러다임이 표준에서 벗어난 이벤트를 감지하는 데 사용되는 이상 감지의 한 형태로 간주될 수도 있는데, 이를 단원 분류 또는 클래스 모델링이라고 한다.

[0082] 본 발명은 감시 환경에서 이상 이벤트를 감지하기 위해 원-클래스 분류(OCC) 모델에 기반한 경량 심층 신경망을 제안한다. 일반적인 원-클래스 분류(OCC) 구조를 따르는 저희 모델은 도 4의 이상 감지 모델 다이어그램과 같이 특징 추출기와 분류 네트워크로 구성된다. 이미지 특징 추출을 위한 네트워크는 입력 대상 클래스를 특징 벡터 공간으로 인코딩하는 데 사용된다. 추출된 특징 외에도 의사 음성(pseudo-negative) 클래스 데이터를 사용하여 정보를 보강한다. 이렇게 추출된 특징은 특징 공간에서 영 중심 가우시안(zero-centred Gaussian)으로 생성된 유사 음성 클래스 데이터에 추가되며, 추출된 특징에서 파생된다. 완전 연결 신경망은 완전히 연결된 신경망에 공급되는 추가된 특징에 의해 성능이 결정되는 분류 모델을 정의한다. 분류 네트워크의 특징 표현에는 해당 표현에 대한 신뢰도에 따라 신뢰도 점수가 할당된다. 분류 네트워크의 결과로 출력은 1 또는 0이 된다. 이 데이터 샘플 분류 체계에서 점수가 1인 데이터 샘플은 목표 클래스에 속하고, 점수가 0인 데이터 샘플은 네거티브 클래스에 속한다. 전체 네트워크의 엔드-투-엔드 학습은 각 노드의 이진 교차 엔트로피 손실을 분석하여 수행된다. 모델에 입력되는 데이터는 198×198 RGB 프레임의 형태이다. 제안된 원-클래스 분류(OCC) 모델은 VGG[K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556,* 2014.] 네트워크에서 영감을 얻었으며, 동일한 컨볼루션 필터(3×3), 보폭(stride) 2를 가지는 최대 풀링 레이어(2×2), 패딩이 없다. 제안된 모델에는 4개의 컨볼루션 블록과 완전히 연결된 256개의 뉴런과 분류를 위한 소프트맥스(Softmax) 함수가 포함된 최종 블록 등 총 5개의 블록이 있다. 최종 예측 단계에서는 비디오 스트림에서 비정상적인 활동이 감지되는지 여부를 판단하기 위해 소프트맥스 출력을 사용하며, 이 경우 비디오 프레임은 추가 분석 및 비정상 유형 식별을 위해 이상 인식 모델에 전달된다.

[0083] 3.2. 이상 인식 모델

[0084] 도 5는 도 4에 도시한 멀티-리저보이어 에코 스테이트 네트워크의 구조를 설명하기 위한 도면이다.

[0085] 도 6은 도 4에 도시한 병목 어텐션 모듈의 구조를 설명하기 위한 도면이다.

[0086] 이 섹션에서는 이상 인식 모델에 대한 전체 개요를 설명한다. 이상 인식을 포함한 다양한 유형의 지능형 감시 비디오 모델의 첫 번째 단계는 원시 입력 비디오에서 특징을 추출하는 것이다. 비전 트랜스포머를 사용하여 트리밍되지 않은 감시 영상에서 특징을 추출한다. 비전 트랜스포머는 최근 이미지 인식, 동작 이해, 객체 감지, 이미지 생성, 비디오 이해 등 많은 까다로운 작업에서 인상적인 성능을 보여줬다. 이 트랜스포머는 인코더-디코더 설계를 채택하고 있으며 반복 네트워크를 사용하지 않고도 순차 입력을 병렬로 처리할 수 있다. 시퀀스 세그먼트 간의 장거리 상관관계를 포착하기 위해 개발된 셀프-어텐션 메커니즘은 트랜스포머 모델의 성공에 크게 기여했다. 표준 트랜스포머의 사용을 이미지 분류로 확장하기 위해 비전 트랜스포머가 제시된다. 주요 목적은 특정 아키텍처를 통합하지 않고 텍스트 양식 이상으로 일반화하는 것이다. 비전 트랜스포머는 트랜스포머 인코더 모듈을 사용하여 이미지 패치 시퀀스를 시맨틱 레이블로 분류한다. 비전 트랜스포머의 어텐션 메커니즘은 표준 합성곱 신경망(CNN) 설계와 달리 로컬 수신 필드가 있는 필터를 사용하는 경우가 많으며, 이를 통해 다

양한 이미지 패치를 커버하고 전체 이미지에 걸쳐 정보를 통합할 수 있다.

[0087] 비전 트랜스포머의 전체 아키텍처는 도 4에 나와 있다. 주로 인코더(encoder), 임베딩 레이어(embedding layer), 최종 헤드 분류기(final head classifier)로 구성된다. 첫 번째 단계에서는 입력 프레임 $\mathcal{X}$를 훈련 세트에서 겹치지 않는 고정 크기 패치의 시퀀스로 세분화한다. 트랜스포머는 각 패치를 별도의 토큰으로 간주한다. 주어진 입력 크기 $\mathcal{H} \times \mathrm{W} \times \mathrm{C}$(여기서 $\mathcal{H}$는 입력 프레임의 높이, W는 너비, C는 채널 수)의 입력 프레임 $\mathcal{X}$의 경우, 프레임에서 $\mathrm{C} \times \mathcal{P} \times \mathcal{P}$ 차원의 패치를 추출한다. 이렇게 하면 길이 $n = \mathcal{H} \times \mathrm{W} / \mathcal{P}^2$의 패치 시퀀스 $(x_1, x_2, .., x_{\mathrm{N}})$가 생성된다. 패치 크기 $\mathcal{P}$는 일반적으로 32×32 또는 16×16으로 선택되며, 패치 크기가 작을수록 시퀀스 길이가 길어지고 그 반대도 마찬가지이다.

[0088] 3.2.1. 선형 임베딩 레이어(Linear Embedding Layer)

[0089] 인코더에 일련의 패치를 적용하기 전에 학습된 임베딩 행렬 $\mathcal{E}$를 사용하여 모델 차원 벡터 $\mathcal{D}$에 선형적으로 투영된다. 그런 다음 임베딩된 이미지가 분류 작업을 수행하는 데 필요한 학습된 분류 토큰 v_{class}와 결합된다. 트랜스포머는 임베딩된 이미지 패치를 배열에 대한 정보 없이 단순히 일련의 패치로만 인식한다. 따라서 패치의 공간적 배열을 원본 이미지와 동일하게 유지하기 위해 위치 정보 $\mathcal{E}_{pos}$를 인코딩하여 패치 표현에 첨부한다. 그 결과 통합된 패치 시퀀스는 [수학식 1]에서 토큰 z_0으로 표시된다.

[0090] 3.2.2 트랜스포머 인코더 모듈(Transformer Encoder Module)

[0091] 도 4의 이상 인식 모델에 표시된 대로 트랜스포머 인코더 모듈에서 일련의 임베디드 패치 z_0을 승인한다. 이 모듈은 두 개의 주요 하위 구성 요소인 멀티 헤드 셀프 어텐션 블록(multi-head self-attention block, MHSA)과 완전 연결된 피드 포워드 고밀도 블록(fully connected feed-forward dense block, MLP)으로 구성된 동일한 레이어로 구성된다. 최종 분류 결과가 나오기 전에 두 개의 고밀도 레이어 사이에 GeLU 활성화 기능이 적용된다. 잔여 스킵(residual skip) 연결은 비전 트랜스포머 인코더 모듈의 두 하위 구성 요소 각각을 보존하고 그 뒤에 정규화 레이어(normalisation layer) $\mathcal{NL}$가 이어진다.

[0092] [수학식 1]

[0093] $$z_{\mathcal{L}}^{'} = MHSA(\mathcal{NL}(z_{\mathcal{L}-1})) + z_{\mathcal{L}-1}, \mathcal{L} = 1.....\ell$$

[0094] [수학식 2]

[0095] $$z_{\mathcal{L}} = MLP(\mathcal{NL}(z_{\mathcal{L}}^{'})) + z_{\mathcal{L}}^{'}, \ \mathcal{L} = 1.....\ell$$

[0096] 인코더의 마지막 레이어에서 시퀀스 $z_{\mathcal{L}}^{0}$의 초기 부분을 가져와서 헤드 분류기에 전달하여 레이블을 예측한다.

[0097] [수학식 3]

[0098] $$y = \mathcal{NL}(z_{\mathcal{L}}^{0})$$

[0099] 인코더의 멀티 헤드 셀프 어텐션 블록(MHSA)은 트랜스포머의 핵심 요소이며, 이 블록의 주요 역할은 나머지 시퀀스와 비교하여 단일 패치 임베딩의 상대적 중요성을 설정하는 것이다. 멀티 헤드 셀프 어텐션 블록(MHSA)은

선형 레이어(linear layer), 셀프 어텐션(self-attention), 연결(concatenation), 최종 선형 레이어(final linear layer)의 네 가지 주요 레이어로 구성된다. 멀티 헤드 셀프 어텐션 블록(MHSA)은 각 헤드에 대한 다양한 투영 행렬을 통해 단일 헤드의 셀프 어텐션을 확장한다. 특히, 주어진 토큰 x_1은 쿼리($\mathcal{Q}$), 키($\mathcal{K}$), 그리고 투영 행렬을 적용한 값($\mathcal{V}$)에 투영된 초기 토큰으로, 즉 $\mathcal{Q},\mathcal{K},\mathcal{V} = x_1\mathrm{W}_Q, x_1Q_\mathcal{K}, x_1Q_\mathcal{V}$으로, 여기서 $\mathrm{W}_{Q/\mathcal{K}/\mathcal{V}} \in \mathcal{R}^{\mathcal{D}\times\mathcal{D}}$는 각각 값, 쿼리, 키에 대한 투영 행렬을 나타낸다. 다음으로 셀프 어텐션은 다음의 [수학식 4]와 같이 수행된다.

[0100] [수학식 4]

[0101] $$SA(\mathcal{Q},\mathcal{K},\mathcal{V}) = softmax\left(\mathcal{Q}\mathcal{K}^{\mathcal{T}}/\sqrt{\mathcal{D}}\right)\mathcal{V}$$

[0102] 이 단일 헤드 셀프 어텐션 블록을 $\hbar$회 반복하여 멀티 헤드 셀프 어텐션 블록(MHSA)을 생성하며, 여기서 $\hbar$는 헤드 수를 나타내고 헤드의 출력 특징을 채널 차원을 따라 통합하여 멀티 헤드 셀프 어텐션 블록(MHSA)의 출력을 계산한다.

[0103] 3.2.3. 멀티-리저보이어 에코 스테이트 네트워크

[0104] Jaeger와 Haas는 비선형 시스템을 학습하고 불안정한 시계열을 식별하기 위해 처음에 에코 스테이트 네트워크(echo state network, ESN)를 제안했다. 기본 개념은 RNN의 넓은 리저보이어를 활용하는 것이다. 에코 스테이트 네트워크(ESN) 방법에서는 무작위로 반복되는 연결과 -1에서 1까지 다양한 수치값을 가진 보통보다 큰 뉴런층이 사용되며, 이 값은 훈련 내내 수정되지 않는다. 에코 스테이트 네트워크(ESN)는 다음과 같이 설정할 수 있다. 입력 계층 뉴런은 환경으로부터 데이터를 수신하여 리저보이어로 전송하며, 주요 목표는 주어진 데이터를 고차원 공간으로 확장하는 것이다. 입력 정보의 강도는 이 새로운 공간에서 더 명확하게 나타나고 선형 분할이 더 쉬워지지만 입력의 비선형 변화는 그 상태로 유지된다(메모리 효과).

[0105] 다음으로, 리저보이어 출력의 선형 조합을 사용하여 원하는 출력 신호를 얻는다. 최적의 가중치 행렬을 식별하기 위해 다중 선형 회귀를 통해 단일 선형 조합을 개발한다. 일반적으로 복잡계는 수많은 변수로 구성되며, 단일 ESN-리저보이어 모델을 사용하여 복잡한 시계열 데이터를 예측할 경우 활동/공간 재구성 차원에 대해 단 하나의 값만 결정할 수 있어 복잡계의 동적 특성을 효과적으로 보여줄 수 없다. 이는 예측 모델의 자유도를 어느 정도 제한하고 예측 정확도에 부정적인 영향을 미친다. 본 발명에서는 이러한 문제를 해결하기 위해 멀티-리저보이어 에코 스테이트 네트워크(MrESN)를 제안한다. 복잡계 시스템의 각 변수는 복잡계의 동적 고유성을 나타내기 위해 각 리저보이어별로 적절하게 매핑된다. 이 아키텍처의 가장 큰 장점은 기존 에코 스테이트 네트워크(ESN)에 비해 넓은 스펙트럼과 다양한 리저보이어 역학으로 교사 신호(teacher's signal)를 탐색, 캡슐화 및 재현하는 데 기여하여 유연성을 높이고 계산 비용을 절감할 수 있다는 점이다. 제안된 멀티-리저보이어 에코 스테이트 네트워크(MrESN)의 아키텍처는 도 5에 나와 있다.

[0106] 그러면 (두 개의 리저보이어를 예로 들어) 멀티-리저보이어 에코 스테이트 네트워크(MrESN)는 다음의 [수학식 5] ~ [수학식 7]과 같이 표현할 수 있다.

[0107] [수학식 5]

[0108] $$\mathcal{X}_1(\mathcal{K}+1) = tanh(\mathcal{W}_\mathcal{X}\cdot\mathcal{X}_1(\mathcal{K}) + \mathcal{W}_{In}\cdot\mathrm{u}_1(\mathcal{K}))$$

[0109] [수학식 6]

[0110] $$\mathcal{X}_2(\mathcal{K}+1) = tanh(\mathcal{W}_\mathcal{X}\cdot\mathcal{X}_2(\mathcal{K}) + \mathcal{W}_{In}\cdot\mathrm{u}_2(\mathcal{K}))$$

[0111] [수학식 7]

[0112] $$\mathcal{Y}(\mathcal{K}+1) = tanh(\mathcal{W}_1{}^{\mathcal{T}}\mathcal{X}_1(\mathcal{K}) + \mathcal{W}_2{}^{\mathcal{T}}\mathcal{X}_2(\mathcal{K}) + \in$$

[0113] 여기서 입력 뉴런의 가중치 행렬은 $\mathcal{W}_{In}$ 이고, 내부 연결 행렬은 $\mathcal{W}_x$로 표시되며, 여기서 $\mathcal{u}_1(\mathcal{K})$ 및 $\mathcal{u}_2(\mathcal{K})$는 네트워크에 대한 입력이고, $\mathcal{X}_1(\mathcal{K})$ 및 $\mathcal{X}_2(\mathcal{K})$는 시간 단계 $\mathcal{K}$의 내부 상태 행렬이다. [수학식 5]와 같이 네트워크를 계산하는 데 선형 함수가 사용되며, 여기서 $\mathcal{W}_1$과 $\mathcal{W}_2$는 출력 유닛에 대한 링크이고, $\in$는 노이즈 신호를 나타낸다. 에코 스테이트 네트워크(ESN)의 주요 특징 중 하나는 학습/훈련 과정에서 출력 유닛 $\mathcal{W}_1$과 $\mathcal{W}_2$에 대한 연결 가중치만 수정된다는 것이다. 히든 레이어(hidden layer)와 다른 가중치 행렬에 대한 토폴로지는 변경되지 않는다.

[0114] 본 발명에서는 정확한 이상 인식을 위해 병목 어텐션 모듈(BAM)이 있는 멀티-리저보이어 에코 스테이트 네트워크(MrESN)에서 시간 "t"에 청크(chunk) 형태로 비전 트랜스포머(Vision Transformer, ViT)의 특징을 전달한다. 병목 어텐션 모듈(BAM)은 채널 및 공간 어텐션 축을 모두 활용할 수 있다는 점, 단순하고 가벼운 설계 등 다른 어텐션 메커니즘에 비해 몇 가지 장점이 있다. 이 시스템의 또 다른 장점은 모듈을 쉽게 배치할 수 있고, 효과적인 역전파를 제공하며, 글로벌 컨텍스트 모델링을 효율적으로 처리할 수 있다는 점이다. 병목 어텐션 모듈(BAM)은 도 6과 같이 채널 및 공간 어텐션 분기를 결합하여 관심도 3D 맵을 생성한다. 효율적인 그라디언트 흐름을 위해 요소별 계산을 사용하고, 합산 후 시그모이드 함수를 적용하여 정규화된 형태의 최종 3D 어텐션 맵을 얻는다. 정규화 후 이 3D 어텐션 맵은 요소 단위로 멀티-리저보이어 에코 스테이트 네트워크(MrESN)의 입력 특징에 곱해진 다음 원본 특징과 통합되어 정제된 특징을 제공한다. 첫 번째 프레임 패치 시퀀스의 1×1024 특징의 초기 청크는 멀티-리저보이어 에코 스테이트 네트워크(MrESN)에 입력으로 전달되고 시간 "t"에서 병목 어텐션 모듈(BAM)을 통해 정제된다. 이 출력 기법을 통해 제안된 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프레임워크는 최첨단 정확도를 달성할 수 있다. 프레임 출력은 시퀀스의 이전 프레임과 다음 프레임을 모두 기반으로 계산되며, 각 레이어는 멀티-리저보이어 에코 스테이트 네트워크(MrESN)를 통해 정보를 처리한다. 시퀀스의 연속된 프레임에서 어텐션 메커니즘으로 추출한 딥 피처를 통해 프레임에 이상이 있는지 또는 정상적인 활동이 있는지 판단할 수 있으며, 이를 멀티-리저보이어 에코 스테이트 네트워크(MrESN)에서 직접 전달하고 소프트맥스(Softmax) 함수를 최종 예측 레이어로 사용한다.

[0116] 4. 실험 결과

[0117] 이 섹션에서는 여러 가지 까다로운 이상 인식 데이터 세트에 대해 본 발명에 따른 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프레임워크를 테스트한다.

[0118] 4.1. 데이터 세트

[0119] 본 발명은 두 가지 까다로운 이상 인식 데이터 세트인 LD-2000 및 UCF-Crime에서 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프레임워크를 테스트했다.

[0120] 4.2. 결과

[0121] 도 7은 본 발명의 바람직한 실시예에 따른 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프레임워크와 최첨단(SOTA) 기법을 비교한 도면이고, 도 8은 본 발명의 바람직한 실시예에 따른 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프레임워크의 혼동 행렬을 나타내는 도면으로, 도 8의 (a)는 UCF-Crime 데이터 세트를 나타내고, 도 8의 (b)는 LAD-2000 데이터 세트를 나타내며, 도 9는 본 발명의 바람직한 실시예에 따른 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프레임워크의 클래스별 성능을 나타내는 도면으로, 도 9의 (a)는 UCF-Crime 데이터 세트에서 2스트림 네트워크와 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프레임워크의 클래스별 정확도 성능 비교를 나타내고, 도 9의 (b)는 LAD-2000 데이터 세트에서 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프

레임워크의 클래스별 정확도를 나타내고, 도 10은 본 발명의 바람직한 실시예에 따른 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프레임워크와 다른 최신 기법을 ROC 파라미터를 기준으로 UCF-Crime 데이터 세트에서 비교한 결과를 나타내는 도면이며, 도 11은 본 발명의 바람직한 실시예에 따른 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프레임워크가 UCF-Crime 및 LAD-2000 데이터 세트에서 이상을 인식한 예시를 나타내는 도면으로, UCF-Crime 데이터 세트의 (a)는 방화를 나타내고, UCF-Crime 데이터 세트의 (b)는 폭발을 나타내며, UCF-Crime 데이터 세트의 (c)는 교통 사고를 나타내고, UCF-Crime 데이터 세트의 (d)는 폭행을 나타내며, LAD-2000 데이터 세트의 (a)는 떨어짐을 나타내고, LAD-2000 데이터 세트의 (b)는 넘어짐을 나타내며, LAD-2000 데이터 세트의 (c)는 폭력을 나타내고, LAD-2000 데이터 세트의 (d)는 절도를 나타낸다.

[0122] 이 섹션에서는 다양한 대규모 이상 감지 데이터 세트에 대해 제안한 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프레임워크의 실험적 평가에 대해 자세히 설명한다. 제안한 경량 원-클래스 이상 감지 프레임워크는 처리 시간을 단축하면서 유망한 결과를 얻었다. 제안된 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프레임워크의 혼동 행렬, 클래스별 정확도, 이상 감지 데이터 세트에 대한 ROC 결과는 각각 도 8, 도 9, 도 10에 표시되어 있다. 제안된 비전 트랜스포머 이상 인식 네트워크(ViT-ARN) 프레임워크의 시각적 결과에 대한 정성적 분석은 테스트 비디오의 다양한 이상 이벤트 클래스에 대해 도 11에 제시되어 있다. 비디오의 일부 시퀀스는 폭행으로 정확하게 예측되었지만, 모델의 일부 비디오 시퀀스 예측은 혼동되었다. 빨간색 상자는 다른 클래스와 혼동된 모델의 잘못된 예측을 나타낸다. UCF-Crime 데이터 세트에서 예측이 잘못되는 이유는 여러 가지가 있는데, 가장 일반적인 이유는 비디오 세그먼트 간의 클래스 내 유사성이다. UCF-Crime 데이터 세트에서 싸움 또는 폭행으로 분류된 동영상을 육안으로 검사한 결과, 이러한 결과를 뒷받침하는 증거를 발견했다. 이 데이터 세트의 일부 폭행 동영상은 폭행 중 싸움 장면을 보여주는 반면, 다른 싸움 장면에서는 동일한 이벤트 시나리오가 발견된다. 예를 들어, 동영상에서 5초가 지나면 폭행을 당한 사람이 자신을 공격하는 사람과 싸우기 시작한다. 또 다른 싸움 영상에서도 똑같은 상황의 예가 있는데, 영상 초반에 한 사람이 남성에게 다가와 싸울 의사가 있는 것처럼 보이지만 몇 초 후 싸움이 시작되는 것을 볼 수 있다. 폭행 클래스에는 싸움 클래스와 동일한 기본 패턴을 가진 약 20~25개의 동영상이 있다. 마찬가지로 강도, 절도, 절도 클래스 내에서도 모델 성능에 영향을 미친다. 이러한 기술의 성능은 도 7에 도시된 표에 요약되어 있다.

[0125] 본 실시예들에 따른 동작은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능한 저장 매체에 기록될 수 있다. 컴퓨터 판독 가능한 저장 매체는 실행을 위해 프로세서에 명령어를 제공하는데 참여한 임의의 매체를 나타낸다. 컴퓨터 판독 가능한 저장 매체는 프로그램 명령, 데이터 파일, 데이터 구조 또는 이들의 조합을 포함할 수 있다. 예컨대, 자기 매체, 광기록 매체, 메모리 등이 있을 수 있다. 컴퓨터 프로그램은 네트워크로 연결된 컴퓨터 시스템 상에 분산되어 분산 방식으로 컴퓨터가 읽을 수 있는 코드가 저장되고 실행될 수도 있다. 본 실시예를 구현하기 위한 기능적인(Functional) 프로그램, 코드, 및 코드 세그먼트들은 본 실시예가 속하는 기술 분야의 프로그래머들에 의해 용이하게 추론될 수 있을 것이다.

[0126] 본 실시예들은 본 실시예의 기술 사상을 설명하기 위한 것이고, 이러한 실시예에 의하여 본 실시예의 기술 사상의 범위가 한정되는 것은 아니다. 본 실시예의 보호 범위는 아래의 청구범위에 의하여 해석되어야 하며, 그와 동등한 범위 내에 있는 모든 기술 사상은 본 실시예의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

부호의 설명

[0127] 100 : 이상 인식 장치,

110 : 프로세서,

130 : 컴퓨터 판독 가능한 저장 매체,

131 : 프로그램,

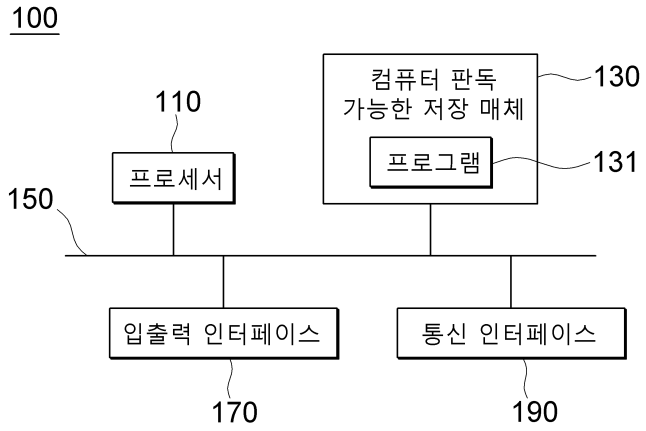
150 : 통신 버스,

170 : 입출력 인터페이스,

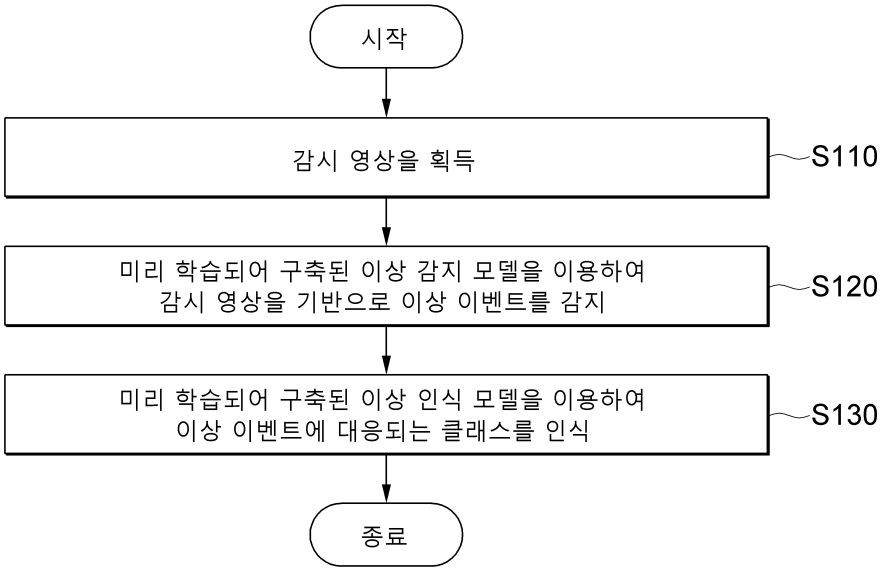
190 : 통신 인터페이스

도면

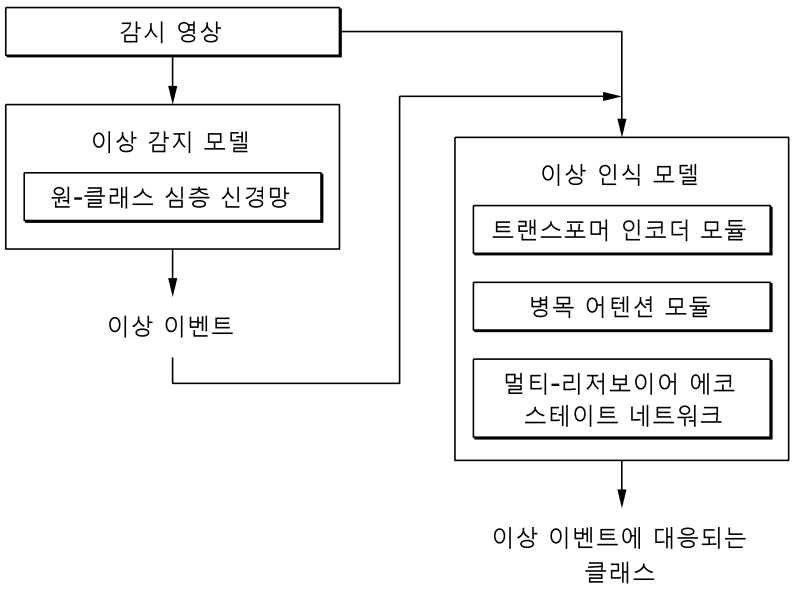
도면1


100
110
프로세서
150
컴퓨터 판독
가능한 저장 매체
130
프로그램
131
입출력 인터페이스
통신 인터페이스
170
190

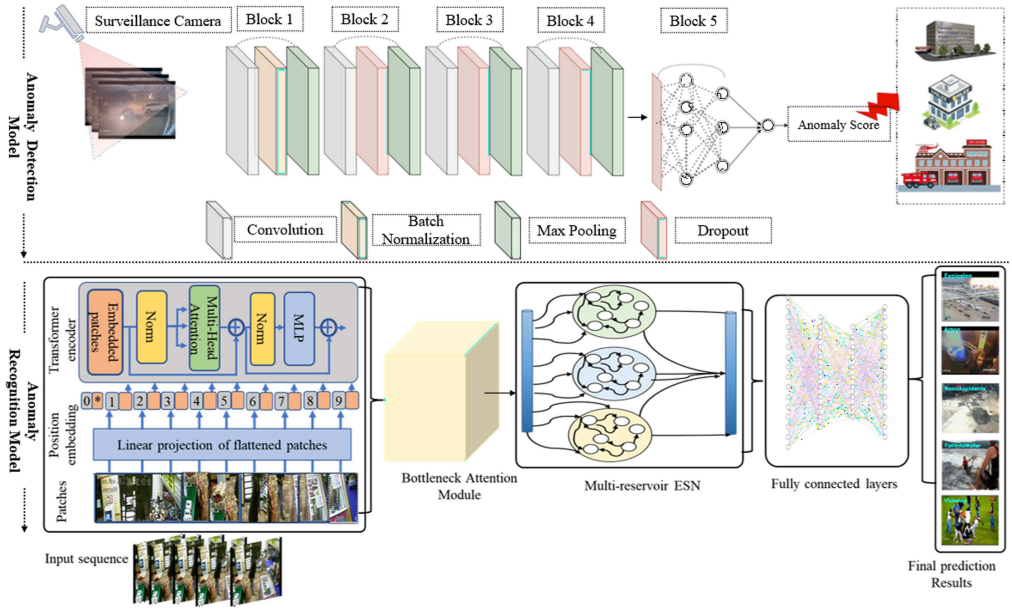
도면2


시작
감시 영상을 획득
S110
미리 학습되어 구축된 이상 감지 모델을 이용하여
감시 영상을 기반으로 이상 이벤트를 감지
S120
미리 학습되어 구축된 이상 인식 모델을 이용하여
이상 이벤트에 대응되는 클래스를 인식
S130
종료

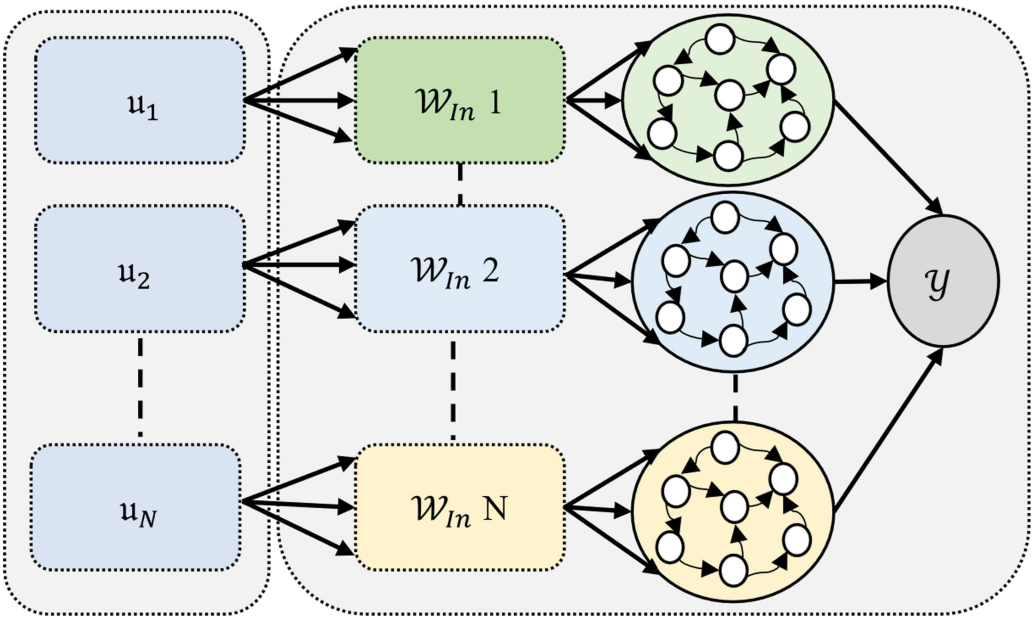
도면3


감시 영상
이상 감지 모델
원-클래스 심층 신경망
이상 이벤트
이상 인식 모델
트랜스포머 인코더 모듈
병목 어텐션 모듈
멀티-리저보이어 에코 스테이트 네트워크
이상 이벤트에 대응되는 클래스

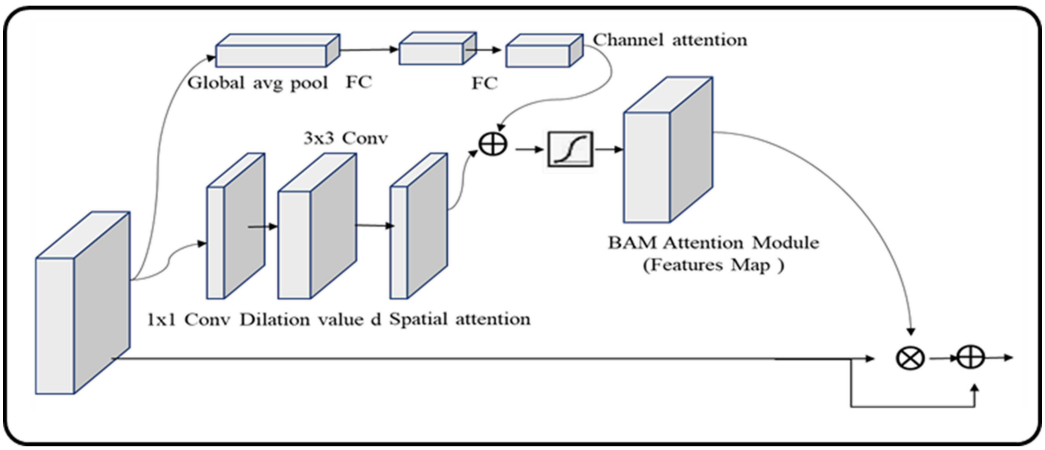
도면4


Anomaly Detection Model
Surveillance Camera
Block 1
Block 2
Block 3
Block 4
Block 5
Anomaly Score
Convolution
Batch Normalization
Max Pooling
Dropout
Anomaly Recognition Model
Transformer encoder
Embedded patches
Norm
Multi-Head Attention
Norm
MLP
Position embedding
Linear projection of flattened patches
Patches
Input sequence
Bottleneck Attention Module
Multi-reservoir ESN
Fully connected layers
Final prediction Results

도면5



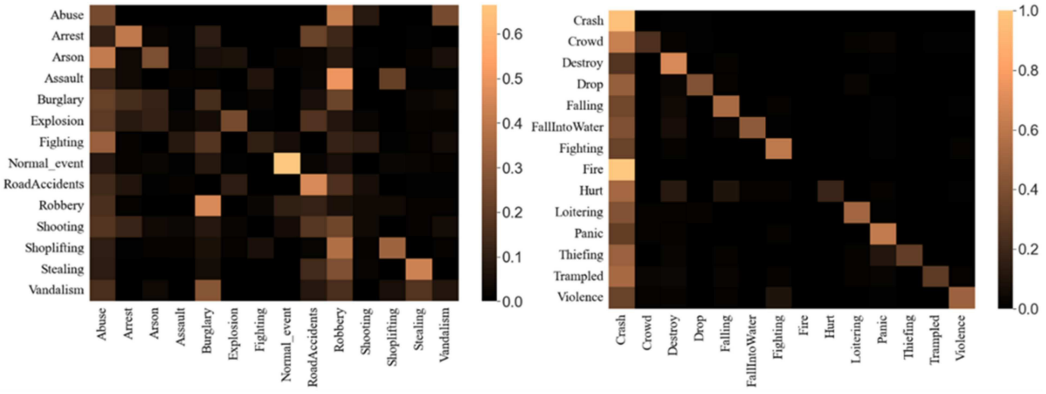
도면6



도면7

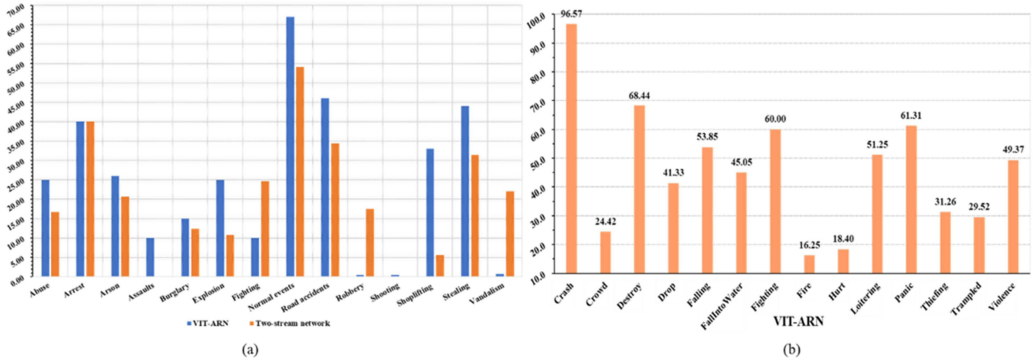
Model	Spatio-temporal features	Attention module	Temporal features	SOTA
3D Siamese [35]	-	-	✓	-
3DCovNet [38]	-	-	✓	-
C3D-KNN [9, 15]	-	-	✓	-
DACM [39]	-	-	-	-
C3D [15]	-	-	✓	-
Two-stream network [34]	✓	-	-	-
Proposed ViT-ARN framework	✓	✓	-	✓

도면8


Abuse
Arrest
Arson
Assault
Burglary
Explosion
Fighting
Normal_event
RoadAccidents
Robbery
Shooting
Shoplifting
Stealing
Vandalism
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Crash
Crowd
Destroy
Drop
Falling
FallIntoWater
Fighting
Fire
Hurt
Loitering
Panic
Thiefing
Trampled
Violence
1.0
0.8
0.6
0.4
0.2
0.0

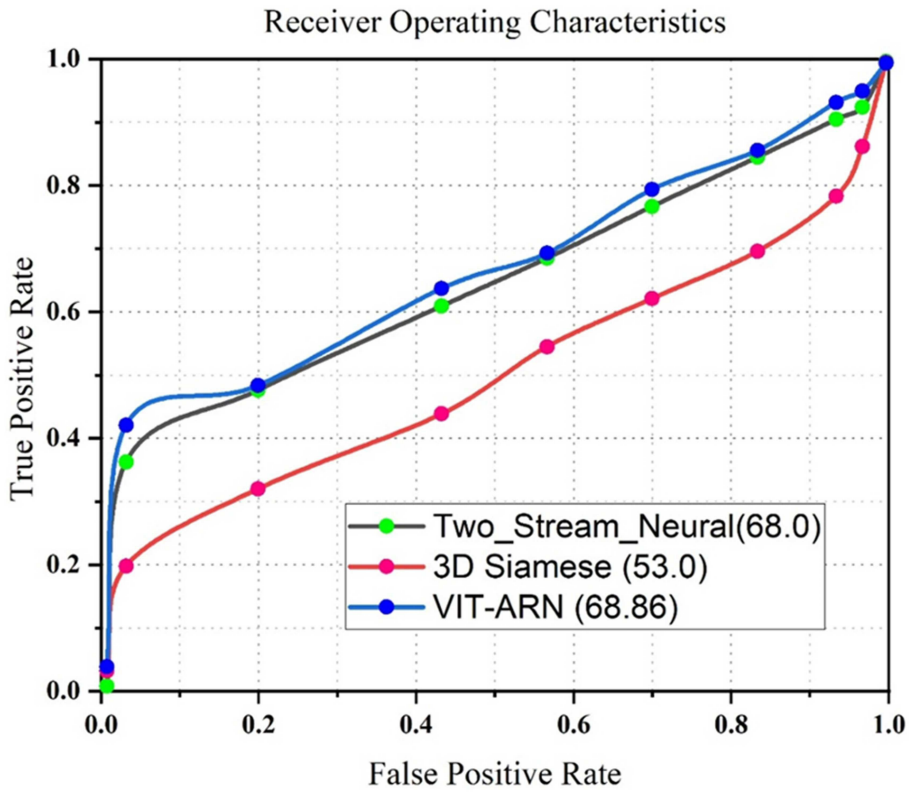
(a) (b)

도면9

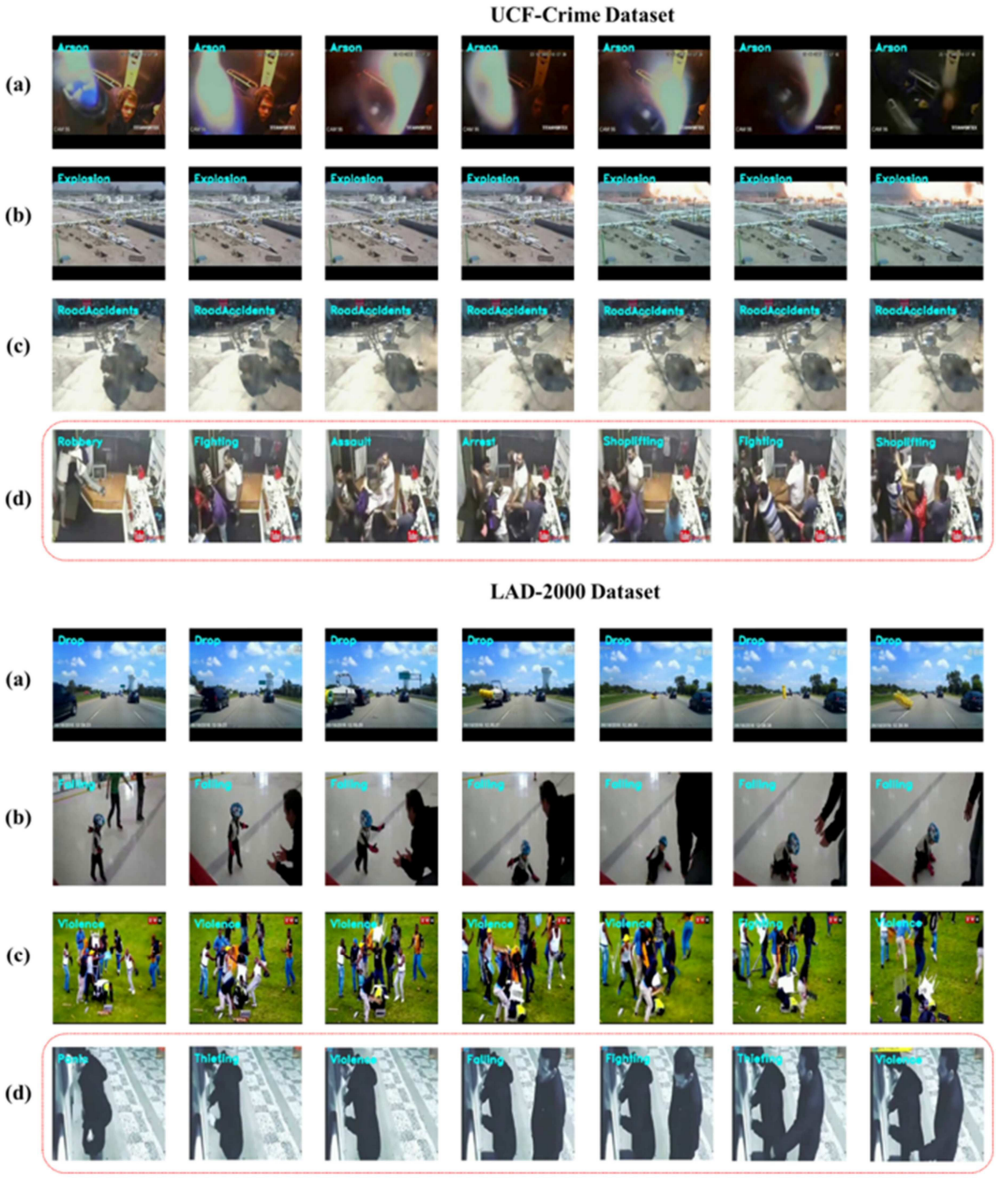

VIT-ARN
Two-stream network
96.57
24.42
68.44
41.33
53.85
45.05
60.00
16.25
18.40
51.25
61.31
31.26
29.52
49.37
VIT-ARN

(a) (b)

도면10


Receiver Operating Characteristics
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Two_Stream_Neural(68.0)
3D Siamese (53.0)
VIT-ARN (68.86)

도면11


UCF-Crime Dataset
(a)
Arson
(b)
Explosion
(c)
RoadAccidents
(d)
Robbery
Fighting
Assault
Arrest
Shoplifting
Fighting
Shoplifting
LAD-2000 Dataset
(a)
Drop
(b)
Falling
(c)
Violence
(d)
Theft
Theft
Violence
Falling
Fighting
Theft
Violence