(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2021년12월01일
(11) 등록번호 10-2334388
(24) 등록일자 2021년11월29일

(51) 국제특허분류(Int. Cl.)
G06K 9/00 (2006.01) *G06F 40/216* (2020.01)
G06K 9/46 (2006.01) *G06N 3/08* (2006.01)
(52) CPC특허분류
G06K 9/00711 (2013.01)
G06F 40/216 (2020.01)
(21) 출원번호 10-2019-0168077
(22) 출원일자 2019년12월16일
심사청구일자 2019년12월16일
(65) 공개번호 10-2021-0076659
(43) 공개일자 2021년06월24일
(56) 선행기술조사문헌
WO2017150211 A1*
JP2011118777 A
KR1020180125885 A
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
변혜란
서울특별시 서대문구 연세로 50, 제4공학관 810호(신촌동, 연세대학교)
이제욱
서울특별시 서대문구 연세로 50, 제4공학관 810호(신촌동, 연세대학교)
김호성
서울특별시 서대문구 연세로 50, 제4공학관 810호(신촌동, 연세대학교)
(74) 대리인
특허법인우인

전체 청구항 수 : 총 10 항 심사관 : 황승희

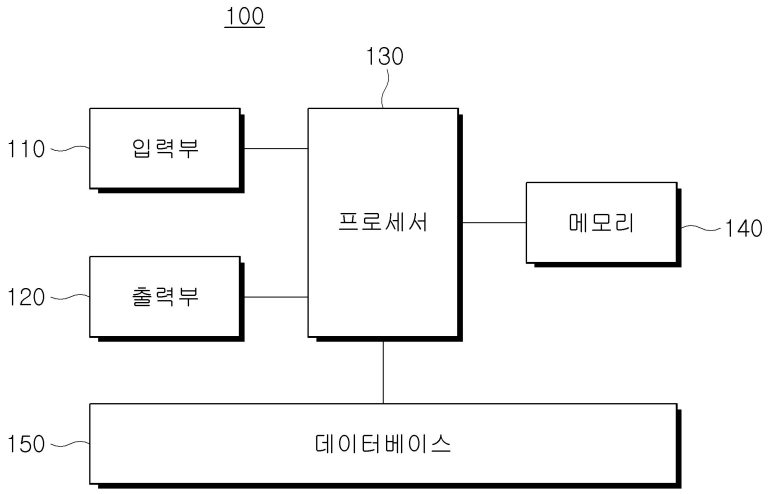
(54) 발명의 명칭 **순차적 특징 데이터 이용한 행동 인식 방법 및 그를 위한 장치**

(57) 요 약

순차적 특징 데이터 이용한 행동 인식 방법 및 그를 위한 장치를 개시한다.

본 발명의 실시예에 따른 하나 이상의 프로세서 및 상기 프로세서에 의해 실행되는 하나 이상의 프로그램을 저장하는 메모리를 포함하는 컴퓨팅 디바이스에 의해 수행되는 행동 인식 학습 방법은 자연어 벡터를 획득하는 자연어 획득 단계; 자연어 벡터를 입력으로 적어도 하나의 특징값을 포함하는 자연어 특징 데이터를 생성하는 자연어 처리 단계; 상기 자연어 특징 데이터를 기반으로 소스 영상의 소스 특징 데이터와 분류를 위한 대상 특징 데이터를 생성하는 생성 처리 단계; 및 상기 소스 특징 데이터와 상기 자연어 특징 데이터 및 상기 대상 특징 데이터 중 적어도 하나를 기반으로 시퀀스(Sequence) 및 세그먼트(Segment) 각각에 대한 분류를 처리하여 객체의 행동 인식이 수행되도록 하는 감별 처리 단계를 수행할 수 있다.

대 표 도 - 도2



(52) CPC특허분류
G06K 9/00335 (2013.01)
G06K 9/46 (2013.01)
G06N 3/08 (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711134177
과제번호 2019R1A2C2003760
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 중견연구자지원사업
연구과제명 특성 정보 자동 생성을 통한 처음 보는 복합카테고리의 이미지와 비디오 생성 및 인식을 위한 제로샷 학습 기술 연구
기 여 율 1/1
과제수행기관명 연세대학교 산학협력단
연구기간 2021.03.01 ~ 2022.02.28

명 세 서

청구범위

청구항 1

하나 이상의 프로세서 및 상기 프로세서에 의해 실행되는 하나 이상의 프로그램을 저장하는 메모리를 포함하는 컴퓨팅 디바이스에 의해 수행되는 행동 인식 학습 방법에 있어서, 상기 컴퓨팅 디바이스는,

입력부를 통해 외부 장치로부터 자연어 벡터를 획득하는 자연어 획득 단계;

상기 자연어 벡터에 순차적 정보를 부여하여 생성된 적어도 하나의 특징값을 포함하는 자연어 특징 데이터를 생성하는 자연어 처리 단계;

소스 영상의 소스 특징 데이터와의 분류를 위하여 상기 자연어 특징 데이터와 기 생성된 랜덤 변수를 기반으로 페이크(Fake) 영상에 대한 대상 특징 데이터를 생성하는 생성 처리 단계; 및

상기 대상 특징 데이터와 상기 소스 특징 데이터를 이용하여 데이터 간 시퀀스(Sequence)에 대한 분류를 처리하고, 상기 자연어 특징 데이터 및 상기 대상 특징 데이터를 결합한 대상 결합 데이터와 상기 소스 특징 데이터를 이용하여 데이터 간 세그먼트(Segment)에 대한 분류를 처리하여 객체의 행동 인식이 수행되도록 하는 감별 처리 단계를 수행하되,

상기 감별 처리 단계는, 순차적 정보가 포함된 복수의 소스 특징값을 결합한 상기 소스 특징 데이터와 순차적 정보가 포함된 복수의 대상 특징값을 결합한 상기 대상 특징 데이터를 비교하여 상기 대상 특징 데이터의 진위 여부를 학습한 제1 학습 결과를 출력하는 제1 감별 처리 단계; 및 상기 소스 특징 데이터의 세그먼트 단위와 상기 대상 결합 데이터의 세그먼트 단위를 비교하여 상기 대상 결합 데이터의 진위 여부를 학습한 제2 학습 결과를 출력하는 제2 감별 처리 단계를 포함하는 것을 특징으로 하는 행동 인식 학습 방법.

청구항 2

제1항에 있어서,

상기 자연어 처리 단계는,

상기 자연어 벡터에 순차적 정보를 부여하여 복수 개의 벡터로 확장하고, 상기 복수 개의 벡터 각각에 대응되는 상기 적어도 하나의 특징값 각각을 생성하는 것을 특징으로 하는 행동 인식 학습 방법.

청구항 3

제2항에 있어서,

상기 자연어 처리 단계는,

상기 자연어 벡터의 평균, 표준 편차 및 노이즈 중 적어도 하나를 이용하여 상기 자연어 벡터를 정규 분포 상에서 분포를 갖는 상기 적어도 하나의 특징값 각각을 생성하는 것을 특징으로 하는 행동 인식 학습 방법.

청구항 4

제2항에 있어서,

상기 자연어 처리 단계는,

재귀 신경망을 기반으로 상기 자연어 벡터를 상기 복수 개의 벡터로 확장하며, 상기 순차적 정보를 포함하는 상기 복수 개의 벡터 각각은 이전 시점에 생성된 벡터에 근거하여 생성되는 것을 특징으로 하는 행동 인식 학습 방법.

청구항 5

삭제

청구항 6

제1항에 있어서,

상기 생성 처리 단계는,

컨볼루션 뉴럴 네트워크(CNN) 학습을 통해 상기 대상 특징 데이터를 생성하는 것을 특징으로 하는 행동 인식 학습 방법.

청구항 7

제1항에 있어서,

상기 생성 처리 단계는,

상기 자연어 특징 데이터와 동일한 개수의 세그먼트 단위로 상기 대상 특징 데이터를 생성하는 것을 특징으로 하는 행동 인식 학습 방법.

청구항 8

삭제

청구항 9

삭제

청구항 10

제1항에 있어서,

상기 제1 감별 처리 단계는,

상기 제1 학습 결과에 근거하여 상기 대상 특징 데이터를 생성하는 단계로 피드백 정보를 전달하며, 상기 소스 특징 데이터와 상기 대상 특징 데이터를 비교하여 상기 대상 특징 데이터가 참 신호에 해당할 때까지 반복하여 상기 대상 특징 데이터의 진위 여부를 학습하는 것을 특징으로 하는 행동 인식 학습 방법.

청구항 11

삭제

청구항 12

제1항에 있어서,

상기 제2 감별 처리 단계는,

상기 제2 학습 결과에 근거하여 상기 대상 특징 데이터를 생성하는 단계로 피드백 정보를 전달하며, 상기 소스 특징 데이터와 상기 대상 결합 데이터를 비교하여 상기 대상 결합 데이터가 참 신호에 해당할 때까지 반복하여 상기 대상 결합 데이터의 진위 여부를 학습하는 것을 특징으로 하는 행동 인식 학습 방법.

청구항 13

제1항에 있어서,

상기 제2 감별 처리 단계는,

상기 소스 특징 데이터의 세그먼트 단위의 데이터와 상기 자연어 특징 데이터의 특징값과 상기 대상 특징 데이터의 특징값을 결합한 세그먼트 단위의 상기 대상 결합 데이터를 이용하여 세그먼트에 대한 분류를 처리하는 것을 특징으로 하는 행동 인식 학습 방법.

청구항 14

본적 없는 행동을 인식하는 장치로서,

하나 이상의 프로세서; 및

상기 프로세서에 의해 실행되는 하나 이상의 프로그램을 저장하는 메모리를 포함하며, 상기 프로그램들은 하나 이상의 프로세서에 의해 실행될 때, 상기 하나 이상의 프로세서들에서,

입력부를 통해 외부 장치로부터 자연어 벡터를 획득하는 자연어 획득 단계;

상기 자연어 벡터에 순차적 정보를 부여하여 생성된 적어도 하나의 특징값을 포함하는 자연어 특징 데이터를 생성하는 자연어 처리 단계;

소스 영상의 소스 특징 데이터와의 분류를 위하여 상기 자연어 특징 데이터와 기 생성된 랜덤 변수를 기반으로 페이크(Fake) 영상에 대한 대상 특징 데이터를 생성하는 생성 처리 단계; 및

상기 대상 특징 데이터와 상기 소스 특징 데이터를 이용하여 데이터 간 시퀀스(Sequence)에 대한 분류를 처리하고, 상기 자연어 특징 데이터 및 상기 대상 특징 데이터를 결합한 대상 결합 데이터와 상기 소스 특징 데이터를 이용하여 데이터 간 세그먼트(Segment)에 대한 분류를 처리하여 객체의 행동 인식이 수행되도록 하는 감별 처리 단계를 포함하는 동작들을 수행하되,

상기 감별 처리 단계는, 순차적 정보가 포함된 복수의 소스 특징값을 결합한 상기 소스 특징 데이터와 순차적 정보가 포함된 복수의 대상 특징값을 결합한 상기 대상 특징 데이터를 비교하여 상기 대상 특징 데이터의 진위 여부를 학습한 제1 학습 결과를 출력하는 제1 감별 처리 단계; 및 상기 소스 특징 데이터의 세그먼트 단위와 상기 대상 결합 데이터의 세그먼트 단위를 비교하여 상기 대상 결합 데이터의 진위 여부를 학습한 제2 학습 결과를 출력하는 제2 감별 처리 단계를 포함하는 것을 특징으로 하는 행동 인식 장치.

청구항 15

삭제

발명의 설명

기 술 분 야

[0001] 본 발명은 순차적 시퀀스 데이터를 이용하여 영상 내 행동을 인식하는 방법 및 그를 위한 장치에 관한 것이다.

배 경 기 술

[0002] 이 부분에 기술된 내용은 단순히 본 발명의 실시예에 대한 배경 정보를 제공할 뿐 종래기술을 구성하는 것은 아니다.

[0003] 종래의 관련 연구는 동영상 데이터의 경우 동영상 하나에서 여러 개의 특징 벡터가 추출되더라도, 이미지와 유사한 방식으로 처리하기 위하여 추출된 여러 벡터에 대한 평균 벡터를 사용하여 행동을 인식한다.

[0004] 다시 말해, 종래의 제로샷 행동인식(Zero-shot Action Recognition)기술은 시계열 정보가 담겨져 있는 순차적 데이터를 사용함에도 불구하고, 제로샷 이미지 분류 연구와 유사한 방법을 적용하기 위해 심층 신경망을 통해 추출된 순차적 특징 벡터를 평균 낸 특징 벡터로 변환하여 행동 인식에 사용하였다. 하지만, 이러한 방식은 순차적 특징 벡터에 포함된 시계열을 무시함으로써 중간 과정이 비슷한 행동이 존재하는 경우 잘못된 판단 결과를 도출하게 된다. 예를 들어, 도 1에 도시된 바와 같이, 달리기 행동과 점프 행동에 대한 동영상에 대한 행동 인식을 수행하는 경우, 달리기 행동과 점프 행동 각각에 대한 영상에서 추출된 특징 벡터(10)의 시계열을 무시하는 하는 경우, 달리기 행동과 점프 행동을 동일한 행동으로 오인하게 되는 문제가 발생한다. 즉, 도 1에 도시된 바와 같이, 시계열의 흐름을 잃어버림에 따라 특징 데이터(10)를 정확하게 구분하여 생성하지 못하고, 중간 단계가 비슷한 행동으로 잘못 구분하게 될 수 있다.

발명의 내용

해결하려는 과제

[0005] 본 발명은 자연어 벡터를 기반으로 하는 처음 보는 영상에 대한 행동 특징 데이터를 생성하여 학습을 수행함으로써, 실제 영상을 통해 학습하지 않은 행동을 인식할 수 있는 순차적 특징 데이터 이용한 행동 인식 방법 및

그를 위한 장치를 제공하는 데 주된 목적이 있다.

과제의 해결 수단

[0006] 본 발명의 일 측면에 의하면, 상기 목적을 달성하기 위한 하나 이상의 프로세서 및 상기 프로세서에 의해 실행되는 하나 이상의 프로그램을 저장하는 메모리를 포함하는 컴퓨팅 디바이스에 의해 수행되는 행동 인식 학습 방법은, 자연어 벡터를 획득하는 자연어 획득 단계; 자연어 벡터를 입력으로 적어도 하나의 특징값을 포함하는 자연어 특징 데이터를 생성하는 자연어 처리 단계; 상기 자연어 특징 데이터를 기반으로 소스 영상의 소스 특징 데이터와 분류를 위한 대상 특징 데이터를 생성하는 생성 처리 단계; 및 상기 소스 특징 데이터와 상기 자연어 특징 데이터 및 상기 대상 특징 데이터 중 적어도 하나를 기반으로 시퀀스(Sequence) 및 세그먼트(Segment) 각각에 대한 분류를 처리하여 객체의 행동 인식이 수행되도록 하는 감별 처리 단계를 수행할 수 있다.

[0007] 또한, 본 발명의 다른 측면에 의하면, 상기 목적을 달성하기 위한 행동 인식 장치는, 하나 이상의 프로세서; 및 상기 프로세서에 의해 실행되는 하나 이상의 프로그램을 저장하는 메모리를 포함하며, 상기 프로그램들은 하나 이상의 프로세서에 의해 실행될 때, 상기 하나 이상의 프로세서들에서, 자연어 벡터를 획득하는 자연어 획득 단계; 자연어 벡터를 입력으로 적어도 하나의 특징값을 포함하는 자연어 특징 데이터를 생성하는 자연어 처리 단계; 상기 자연어 특징 데이터를 기반으로 소스 영상의 소스 특징 데이터와 분류를 위한 대상 특징 데이터를 생성하는 생성 처리 단계; 및 상기 소스 특징 데이터와 상기 자연어 특징 데이터 및 상기 대상 특징 데이터 중 적어도 하나를 기반으로 시퀀스(Sequence) 및 세그먼트(Segment) 각각에 대한 분류를 처리하여 객체의 행동 인식이 수행되도록 하는 감별 처리 단계를 포함하는 동작들을 수행할 수 있다.

[0008] 또한, 본 발명의 다른 측면에 의하면, 상기 목적을 달성하기 위한 하나 이상의 프로세서 및 상기 프로세서에 의해 실행되는 하나 이상의 프로그램을 저장하는 메모리를 포함하는 컴퓨팅 디바이스에 의해 수행되는 행동 인식 방법은, 본적 없는 소스 영상을 입력 받고, 상기 소스 영상 데이터의 소스 특징 데이터에 시퀀스 특징 데이터를 학습한 제1 학습 결과와 세그먼트 특징 데이터를 학습한 제2 학습 결과를 적용하여 행동을 판단하고, 판단된 행동 인식 결과를 출력할 수 있다.

발명의 효과

[0009] 이상에서 설명한 바와 같이, 본 발명은 자연어 벡터를 기반으로 행동에 대한 순차적 데이터를 생성하여 행동 인식을 수행할 수 있는 효과가 있다.

[0010] 또한, 본 발명은 자연어 벡터를 기반으로 행동 특징 데이터를 생성함으로써, 학습 시 볼 수 없었던 처음 보는 행동(새로운 행동)을 인식할 수 있어 행동 인식 성능을 향상 시킬 수 있는 효과가 있다.

도면의 간단한 설명

[0011] 도 1은 종래 기술의 문제점 및 본 발명의 개략적인 동작 특성을 설명하기 위한 도면이다.

도 2는 본 발명의 실시예에 따른 행동 인식 장치를 개략적으로 나타낸 블록 구성도이다.

도 3은 본 발명의 실시예에 따른 프로세서의 학습을 위한 동작 구성을 개략적으로 나타낸 블록 구성도이다.

도 4는 본 발명의 실시예에 따른 행동 인식을 위한 학습 방법을 설명하기 위한 순서도이다.

도 5는 본 발명의 실시예에 따른 프로세서의 행동 인식을 위한 동작 구성을 개략적으로 나타낸 블록 구성도이다.

도 6은 본 발명의 실시예에 따른 행동 인식 방법을 설명하기 위한 순서도이다.

도 7은 본 발명의 실시예에 따른 행동 인식 장치의 학습 동작을 설명하기 위한 예시도이다.

도 8은 본 발명의 실시예에 따른 입력 영상을 처리하여 특징 데이터를 생성하는 동작을 설명하기 위한 예시도이다.

도 9는 본 발명의 실시예에 따른 자연어 벡터를 처리하여 특징 데이터를 생성하는 동작을 설명하기 위한 예시도이다.

도 10은 본 발명의 실시예에 따른 인코더의 동작 구성을 나타낸 도면이다.

도 11은 본 발명의 실시예에 따른 감별자의 동작 구성을 나타낸 도면이다.

발명을 실시하기 위한 구체적인 내용

[0012] 이하, 본 발명의 바람직한 실시예를 첨부된 도면들을 참조하여 상세히 설명한다. 본 발명을 설명함에 있어, 관련된 공지 구성 또는 기능에 대한 구체적인 설명이 본 발명의 요지를 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명은 생략한다. 또한, 이하에서 본 발명의 바람직한 실시예를 설명할 것이나, 본 발명의 기술적 사상은 이에 한정하거나 제한되지 않고 당업자에 의해 변형되어 다양하게 실시될 수 있음은 물론이다. 이하에서는 도면들을 참조하여 본 발명에서 제안하는 순차적 특징 데이터 이용한 행동 인식 방법 및 그를 위한 장치에 대해 자세하게 설명하기로 한다.

[0013] 도 1에 도시된 바와 같이, 본 발명은 원본 동영상의 시계열 정보를 잃지 않기 위해 특징 벡터의 평균이 아닌 순차적 특징 데이터(20)를 생성하고, 이를 통해 처음 보는 행동을 인식하는 성능을 개선하기 위한 장치 및 방법을 제안한다.

[0014] 도 2는 본 발명의 실시예에 따른 행동 인식 장치를 개략적으로 나타낸 블록 구성도이다.

[0015] 본 실시예에 따른 행동 인식 장치(100)는 입력부(110), 출력부(120), 프로세서(130), 메모리(140) 및 데이터 베이스(150)를 포함한다. 도 2의 행동 인식 장치(100)는 일 실시예에 따른 것으로서, 도 2에 도시된 모든 블록이 필수 구성요소는 아니며, 다른 실시예에서 행동 인식 장치(100)에 포함된 일부 블록이 추가, 변경 또는 삭제될 수 있다. 한편, 행동 인식 장치(100)는 컴퓨팅 디바이스로 구현될 수 있고, 행동 인식 장치(100)에 포함된 각 구성요소들은 각각 별도의 소프트웨어 장치로 구현되거나, 소프트웨어가 결합된 별도의 하드웨어 장치로 구현될 수 있다.

[0016] 행동 인식 장치(100)는 자연어 벡터를 입력 받고, 자연어 벡터에 순차적 정보를 부여하여 생성된 자연어 특징 데이터를 입력으로 생성자를 통해 대상 특징 데이터를 생성하고, 생성자와 연동하는 적어도 2 개의 감별자를 통해 소스 영상(원본 동영상)의 소스 특징 데이터, 자연어 특징 데이터, 대상 특징 데이터 등을 분류 처리하여 처음 보는 영상에서 행동을 인식하는 동작을 수행한다.

[0017] 입력부(110)는 행동 인식 장치(100)에서의 행동 인식 동작을 수행하기 위한 신호 또는 데이터를 입력하거나 획득하는 수단을 의미한다. 입력부(110)는 프로세서(130)와 연동하여 다양한 형태의 신호 또는 데이터를 입력하거나, 외부 장치와의 연동을 통해 신호 또는 데이터를 획득하여 프로세서(130)로 전달할 수도 있다. 여기서, 입력부(110)는 소스 영상(원본 동영상), 자연어 벡터, 랜덤 변수 등을 입력하기 위한 모듈로 구현될 수 있으나 반드시 이에 한정되는 것은 아니다.

[0018] 출력부(120)는 프로세서(130)와 연동하여 특징 데이터 기반의 시퀀스(Sequence) 학습 결과, 특징 데이터 기반의 세그먼트(Segment) 학습 결과, 행동 인식 결과 등 다양한 정보를 출력할 수 있다. 출력부(120)는 행동 인식 장치(100)에 구비된 디스플레이(미도시)를 통해 다양한 정보를 출력할 수 있으나 반드시 이에 한정되는 것은 아니며, 다양한 형태의 방식으로 출력을 수행할 수 있다.

[0019] 프로세서(130)는 메모리(140)에 포함된 적어도 하나의 명령어 또는 프로그램을 실행시키는 기능을 수행한다.

[0020] 본 실시예에 따른 프로세서(130)는 입력부(110) 또는 데이터 베이스(150)로부터 획득한 자연어 벡터 및 소스 영상을 기반으로 기계학습을 수행하고, 기계학습 결과를 기반으로 기 학습되지 않은 처음 보는 영상에 대한 행동을 인식하는 동작을 수행한다.

[0021] 프로세서(130)는 소스 영상을 입력 받고, 소스 영상을 기반으로 전처리를 수행하여 소스 특징 데이터를 생성한다. 또한, 프로세서(130)는 자연어 벡터를 입력 받고, 자연어 벡터에 순차적 정보를 부여하여 자연어 특징 데이터를 생성하고, 자연어 특징 데이터를 입력으로 대상 특징 데이터를 생성한다.

[0022] 또한, 프로세서(130)는 소스 특징 데이터와 자연어 특징 데이터 및 대상 특징 데이터 중 적어도 하나를 기반으로 시퀀스(Sequence) 및 세그먼트(Segment) 각각에 대한 분류를 처리하여 객체의 행동 인식이 수행되도록 한다. 여기서, 프로세서(130)는 영상의 순차적 특징을 고려하여 행동 인식을 수행하기 위하여, 소스 특징 데이터와 대상 특징 데이터를 이용하여 시퀀스에 대한 분류를 처리하여 제1 학습 결과를 생성한다. 또한, 프로세서(130)는 영상의 기 설정된 단위의 특징을 고려하여 행동 인식을 수행하기 위하여, 소스 특징 데이터와 자연어 특징 데이터 및 대상 특징 데이터를 결합한 대상 결합 데이터를 이용하여 세그먼트에 대한 분류를 처리하여 제2 학습 결과를 생성한다. 프로세서(130)는 시퀀스 및 세그먼트 각각에 대한 분류를 처리하여 생성된 제1 학습 결과 및 제

2 학습 결과를 기반으로 학습 시 입력된 적이 없는 처음 보는 영상의 행동 인식을 수행한다.

[0023] 본 실시예에 따른 프로세서(130)의 자세한 동작은 도 3 내지 6에서 설명하도록 한다.

[0024] 메모리(140)는 프로세서(130)에 의해 실행 가능한 적어도 하나의 명령어 또는 프로그램을 포함한다. 메모리(140)는 소스 특징 데이터를 생성하는 동작, 자연어 특징 데이터를 생성하는 동작, 대상 특징 데이터를 생성하는 동작, 대상 결합 데이터를 생성하는 동작, 시퀀스에 대한 분류를 처리하는 동작, 세그먼트에 대한 분류를 처리하는 동작 등을 위한 명령어 또는 프로그램을 포함할 수 있다. 또한, 메모리(140)는 학습 결과를 적용하는 동작, 행동 인식을 수행하는 동작 등을 위한 명령어 또는 프로그램을 포함할 수 있다.

[0025] 데이터베이스(150)는 데이터베이스 관리 프로그램(DBMS)을 이용하여 컴퓨터 시스템의 저장공간(하드디스크 또는 메모리)에 구현된 일반적인 데이터구조를 의미하는 것으로, 데이터의 검색(추출), 삭제, 편집, 추가 등을 자유롭게 행할 수 있는 데이터 저장형태를 뜻하는 것으로, 오라클(Oracle), 인포믹스(Infomix), 사이베이스(Sybase), DB2와 같은 관계형 데이타베이스 관리 시스템(RDBMS)이나, 겜스톤(Gemston), 오리온(Orion), O2 등과 같은 객체 지향 데이타베이스 관리 시스템(OODBMS) 및 엑셀론(Excelon), 타미노(Tamino), 세카이주(Sekaiju) 등의 XML 전용 데이터베이스(XML Native Database)를 이용하여 본 발명의 일 실시예의 목적에 맞게 구현될 수 있고, 자신의 기능을 달성하기 위하여 적당한 필드(Field) 또는 엘리먼트들을 가지고 있다.

[0026] 본 실시예에 따른 데이터베이스(150)는 행동 인식과 관련된 데이터를 저장하고, 기 저장된 행동 인식과 관련된 데이터를 제공할 수 있다.

[0027] 데이터베이스(150)에 저장된 데이터는 소스 영상, 특징 데이터(예: 소스 특징 데이터, 자연어 특징 데이터, 대상 특징 데이터, 대상 결합 데이터 등), 학습 결과(예: 제1 학습 결과, 제2 학습 결과, 행동 인식 학습 결과 등), 행동 인식 결과 등에 대한 데이터일 수 있다. 데이터베이스(140)는 행동 인식 장치(100) 내에 구현되는 것으로 기재하고 있으나 반드시 이에 한정되는 것은 아니며, 별도의 데이터 저장장치로 구현될 수도 있다.

[0028] 도 3은 본 발명의 실시예에 따른 프로세서의 학습을 위한 동작 구성을 개략적으로 나타낸 블록 구성도이다.

[0029] 본 실시예에 따른 행동 인식 장치(100)에 포함된 프로세서(130)는 기계 학습을 기반으로 처음보는 영상 내에서 행동을 인식하는 동작을 수행한다. 여기서, 기계 학습은 생성적 적대 신경망(GAN: Generative Adversarial Network)을 이용한 학습인 것이 바람직하나 반드시 이에 한정되는 것은 아니다.

[0030] 행동 인식 장치(100)에 포함된 프로세서(130)는 소스 영상을 입력 받고, 소스 영상을 기반으로 전처리를 수행하여 소스 특징 데이터를 생성하는 모델, 자연어 벡터를 입력 받고, 자연어 벡터에 순차적 정보를 부여하여 자연어 특징 데이터를 생성하고, 자연어 특징 데이터를 입력으로 대상 특징 데이터를 생성하는 모델, 소스 특징 데이터와 자연어 특징 데이터 및 대상 특징 데이터 중 적어도 하나를 기반으로 시퀀스(Sequence) 및 세그먼트(Segment) 각각에 대한 분류를 처리하는 모델 등을 기반으로 본적 없는 행동을 인식하는 동작이 수행되도록 하며, 행동 인식을 수행하는 모든 기기에 탑재되거나, 행동 인식을 수행하는 소프트웨어와 연동할 수 있다.

[0031] 본 실시예에 따른 프로세서(130)는 영상 획득부(310), 전처리부(320), 영상 특징값 처리부(322), 자연어 벡터 획득부(330), 인코더(340), 제1 특징값 처리부(342), 생성자(350), 제2 특징값 처리부(352) 및 감별자(360)를 포함할 수 있다. 도 3의 프로세서(130)는 일 실시예에 따른 것으로서, 도 3에 도시된 모든 블록이 필수 구성요소는 아니며, 다른 실시예에서 프로세서(130)에 포함된 일부 블록이 추가, 변경 또는 삭제될 수 있다. 한편, 프로세서(130)에 포함된 각 구성요소들은 각각 별도의 소프트웨어 장치로 구현되거나, 소프트웨어가 결합된 별도의 하드웨어 장치로 구현될 수 있다.

[0032] 영상 획득부(310)는 소스 영상을 획득하는 동작을 수행한다. 여기서, 소스 영상은 소스 비디오의 비디오 클립을 의미하며, 비디오 클립은 복수의 영상 세그먼트로 구성될 수 있다. 여기서, 영상 세그먼트는 복수의 움직임 벡터 영상 프레임을 포함한다. 움직임 벡터 영상 프레임 사이에는 차분 영상이 추가로 포함될 수 있으며, 차분 영상은 인접한 두 개의 움직임 벡터 영상 프레임의 차이를 통해 생성된 영상을 의미한다.

[0033] 전처리부(320)는 소스 영상을 입력으로 소스 영상에 대한 소스 특징 데이터를 생성한다. 전처리부(320)에서 생성된 소스 특징 데이터는 복수의 세그먼트 단위 별 특징값를 포함한다.

[0034] 전처리부(320)는 소스 영상에 대해 컨볼루션 뉴럴 네트워크(CNN: Convolutional Neural Networks) 학습을 위한 전처리(Pre-traning)를 수행하여 소스 특징 데이터를 생성할 수 있다. 여기서, 전처리(Pre-traning)에 대한 기술은 일반적으로 알려진 기술이므로 자세한 설명은 생략하도록 한다.

[0035] 영상 특징값 처리부(322)는 전처리부(320)에서 출력된 소스 특징 데이터를 감별자(360)로 전달하는 동작을 수행한다. 영상 특징값 처리부(322)는 소스 특징 데이터를 제1 감별자(372) 및 제2 감별자(374) 각각으로 전달한다.

[0036] 한편, 영상 특징값 처리부(322)는 전처리부(320)에서 소스 특징 데이터를 감별자(360)로 직접 전달하는 경우 생략되거나, 전처리부(320)에 포함된 형태로 구현될 수 있다.

[0037] 자연어 벡터 획득부(330)는 기 설정된 조건에 대응되는 자연어 벡터를 획득한다. 여기서, 자연어 벡터는 시계열적인 정보를 포함하지 않고, 소정의 행동에 대하여 자연어 기반으로 생성된 벡터를 의미한다.

[0038] 인코더(340)는 자연어 벡터를 입력으로 적어도 하나의 특징값을 포함하는 자연어 특징 데이터를 생성하는 동작을 수행한다.

[0039] 인코더(340)는 자연어 벡터에 순차적 정보를 부여하여 복수 개의 벡터로 확장하고, 복수 개의 벡터 각각에 대응되는 적어도 하나의 특징값 각각을 생성한다.

[0040] 인코더(340)는 자연어 벡터의 평균, 표준 편차 및 노이즈 등 중 적어도 하나를 이용하여 자연어 벡터를 정규 분포 상에서 분포를 갖는 적어도 하나의 특징값 각각을 생성한다.

[0041] 인코더(340)는 재귀 신경망(RNN: Recurrent Neural Network)을 기반으로 자연어 벡터를 복수 개의 벡터로 확장하며, 순차적 정보를 포함하는 복수 개의 벡터 각각은 이전 시점에 생성된 벡터에 근거하여 생성될 수 있다.

[0042] 제1 특징값 처리부(342)는 인코더(340)에서 출력된 자연어 특징 데이터를 생성자(350)로 전달하는 동작을 수행한다. 제1 특징값 처리부(342)는 자연어 특징 데이터에 랜덤 변수(잠재 잡음에 대한 랜덤 변수)를 추가로 결합시켜 생성자(350)로 전달할 수 있다.

[0043] 한편, 제1 특징값 처리부(342)는 인코더(340)에서 자연어 특징 데이터를 생성자(350)로 직접 전달하는 경우 생략되거나, 인코더(340)에 포함된 형태로 구현될 수 있다.

[0044] 생성자(350)는 자연어 특징 데이터를 기반으로 소스 영상의 소스 특징 데이터와 분류를 위한 대상 특징 데이터를 생성하는 동작을 수행한다.

[0045] 생성자(350)는 자연어 특징 데이터와 기 생성된 랜덤 변수를 기반으로 페이크(Fake) 영상에 대한 대상 특징 데이터를 생성한다. 여기서, 생성자(350)는 컨볼루션 뉴럴 네트워크(CNN: Convolutional Neural Networks) 학습을 통해 상기 대상 특성 데이터를 생성하는 것이 바람직하나 반드시 이에 한정되는 것은 아니다.

[0046] 생성자(350)는 적어도 하나의 특징값을 포함하는 대상 특징 데이터를 생성한다. 여기서, 생성자(350)는 자연어 특징 데이터와 동일한 개수의 세그먼트 단위로 대상 특징 데이터를 생성한다. 여기서, 세그먼트 단위는 대상 특징 데이터에 포함된 각각의 특징값으로 구분될 수 있다.

[0047] 제2 특징값 처리부(352)는 생성자(350)에서 출력된 대상 특징 데이터를 감별자(360)로 전달하는 동작을 수행한다. 제2 특징값 처리부(352)는 대상 특징 데이터를 제1 감별자(372) 및 제2 감별자(374) 각각으로 전달한다.

[0048] 한편, 제2 특징값 처리부(352)는 생성자(350)에서 대상 특징 데이터를 감별자(360)로 직접 전달하는 경우 생략되거나, 생성자(350)에 포함된 형태로 구현될 수 있다.

[0049] 감별자(360)는 소스 특징 데이터와 자연어 특징 데이터, 대상 특징 데이터 등 중 적어도 하나를 기반으로 시퀀스(Sequence) 및 세그먼트(Segment) 각각에 대한 분류를 처리하여 객체의 행동 인식이 수행되도록 한다. 본 실시예에 따른 감별자(360)는 제1 감별자(372) 및 제2 감별자(374)를 포함한다.

[0050] 제1 감별자(372)는 대상 특징 데이터와 소스 특징 데이터를 이용하여 시퀀스(Sequence)에 대한 분류를 처리하는 동작을 수행한다. 제1 감별자(372)는 대상 특징 데이터와 소스 특징 데이터를 입력 받고, 대상 특징 데이터의 진위 여부를 판별할 수 있다.

[0051] 구체적으로, 제1 감별자(372)는 순차적 정보가 포함된 복수의 소스 특징값을 결합한 소스 특징 데이터와 순차적 정보가 포함된 복수의 대상 특징값을 결합한 대상 특징 데이터를 비교하여 대상 특징 데이터의 진위 여부를 학습한 제1 학습 결과를 출력한다.

[0052] 제1 감별자(372)는 제1 학습 결과에 근거하여 대상 특징 데이터를 생성하는 생성자(350)로 피드백 정보를 전달하며, 소스 특징 데이터와 대상 특징 데이터를 비교하여 대상 특징 데이터가 참 신호에 해당할 때까지 반복하여 대상 특징 데이터의 진위 여부를 학습할 수 있다. 여기서, 제1 감별자(372)는 생성자(350)와 연동하여 대상 특

징 데이터가 참 신호에 해당하도록 분류하기 위하여 생성적 적대 신경망(GAN: Generative Adversarial Network)을 기반으로 학습을 수행하는 것이 바람직하나 반드시 이에 한정되는 것은 아니다._

[0053] 제2 감별자(374)는 자연어 특징 데이터 및 대상 특징 데이터를 결합한 대상 결합 데이터와 소스 특징 데이터를 이용하여 세그먼트(Segment)에 대한 분류를 처리하는 동작을 수행한다. 제2 감별자(374)는 대상 결합 데이터와 소스 특징 데이터를 입력 받고, 대상 결합 데이터의 진위 여부를 판별할 수 있다.

[0054] 구체적으로, 제2 감별자(374)는 소스 특징 데이터의 세그먼트 단위와 대상 결합 데이터의 세그먼트 단위를 비교하여 대상 결합 데이터의 진위 여부를 학습한 제2 학습 결과를 출력한다. 여기서, 제2 감별자(374)는 소스 특징 데이터의 세그먼트 단위의 데이터와 자연어 특징 데이터의 특징값과 대상 특징 데이터의 특징값을 결합한 세그먼트 단위의 대상 결합 데이터를 비교하여 세그먼트에 대한 분류를 처리할 수 있다.

[0055] 제2 감별자(374)는 제2 학습 결과에 근거하여 대상 특징 데이터를 생성하는 생성자(350)로 피드백 정보를 전달하며, 소스 특징 데이터와 대상 결합 데이터를 비교하여 대상 결합 데이터가 참 신호에 해당할 때까지 반복하여 대상 결합 데이터의 진위 여부를 학습할 수 있다. 여기서, 제2 감별자(374)는 생성자(350)와 연동하여 대상 결합 데이터가 참 신호에 해당하도록 분류하기 위하여 생성적 적대 신경망(GAN: Generative Adversarial Network)을 기반으로 학습을 수행하는 것이 바람직하나 반드시 이에 한정되는 것은 아니다.

[0056] 도 4는 본 발명의 실시예에 따른 행동 인식을 위한 학습 방법을 설명하기 위한 순서도이다.

[0057] 행동 인식 장치(100)는 소스 영상의 입력 여부를 확인한다(S410).

[0058] 단계 S410에서 소스 영상이 입력된 경우, 행동 인식 장치(100)는 소스 영상을 획득한다(S420). 행동 인식 장치(100)는 소스 영상을 전처리하여 복수의 영상 특징값을 생성하고, 복수의 영상 특징값을 포함하는 소스 특징 데이터를 생성한다(S430).

[0059] 한편, 단계 S410에서 소스 영상이 입력되지 않고 자연어 벡터가 입력된 경우, 행동 인식 장치(100)는 자연어 벡터를 획득한다(S440).

[0060] 행동 인식 장치(100)는 자연어 벡터를 입력으로 적어도 하나의 특징값(제1 특징값)을 포함하는 자연어 특징 데이터를 생성한다(S450).

[0061] 또한, 행동 인식 장치(100)는 자연어 특징 데이터에 포함된 특징값(제1 특징값)을 입력으로 소스 영상의 소스 특징 데이터와 분류를 위한 적어도 하나의 특징값(제2 특징값)을 포함하는 대상 특징 데이터를 생성한다(S460).

[0062] 행동 인식 장치(100)는 대상 특징 데이터와 소스 특징 데이터를 이용하여 시퀀스(Sequence)에 대한 분류를 처리(제1 감별 처리)를 통해 제1 학습 결과를 생성한다(S470). 구체적으로, 행동 인식 장치(100)는 순차적 정보가 포함된 복수의 소스 특징값을 결합한 소스 특징 데이터와 순차적 정보가 포함된 복수의 대상 특징값을 결합한 대상 특징 데이터를 비교하여 대상 특징 데이터의 진위 여부를 학습한 제1 학습 결과를 출력한다.

[0063] 또한, 행동 인식 장치(100)는 자연어 특징 데이터 및 대상 특징 데이터를 결합한 대상 결합 데이터와 소스 특징 데이터를 이용하여 세그먼트(Segment)에 대한 분류를 처리(제2 감별 처리)를 통해 제2 학습 결과를 생성한다(S480). 구체적으로, 행동 인식 장치(100)는

[0064] 소스 특징 데이터의 세그먼트 단위의 데이터와 자연어 특징 데이터의 특징값과 대상 특징 데이터의 특징값을 결합한 세그먼트 단위의 대상 결합 데이터를 비교하여 대상 결합 데이터의 진위 여부를 학습한 제2 학습 결과를 출력한다.

[0065] 도 4에서는 각 단계를 순차적으로 실행하는 것으로 기재하고 있으나, 반드시 이에 한정되는 것은 아니다. 다시 말해, 도 4에 기재된 단계를 변경하여 실행하거나 하나 이상의 단계를 병렬적으로 실행하는 것으로 적용 가능할 것이므로, 도 4는 시계열적인 순서로 한정되는 것은 아니다.

[0066] 도 4에 기재된 본 실시예에 따른 행동 인식 학습 방법은 애플리케이션(또는 프로그램)으로 구현되고 단말장치(또는 컴퓨터)로 읽을 수 있는 기록매체에 기록될 수 있다. 본 실시예에 따른 행동 인식 학습 방법을 구현하기 위한 애플리케이션(또는 프로그램)이 기록되고 단말장치(또는 컴퓨터)가 읽을 수 있는 기록매체는 컴퓨팅 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록장치 또는 매체를 포함한다.

[0067] 도 5는 본 발명의 실시예에 따른 프로세서의 행동 인식을 위한 동작 구성을 개략적으로 나타낸 블록 구성도이다.

[0068] 본 실시예에 따른 행동 인식 장치(100)에 포함된 프로세서(130)는 입력 영상 획득부(510), 신경망 처리부(520), 학습 결과 적용부(530), 영상 판단부(540) 및 결과 출력부(550)를 포함한다. 도 5의 프로세서(130)는 일 실시예에 따른 것으로서, 도 5에 도시된 모든 블록이 필수 구성요소는 아니며, 다른 실시예에서 프로세서(130)에 포함된 일부 블록이 추가, 변경 또는 삭제될 수 있다. 한편, 프로세서(130)에 포함된 각 구성요소들은 각각 별도의 소프트웨어 장치로 구현되거나, 소프트웨어가 결합된 별도의 하드웨어 장치로 구현될 수 있다.

[0069] 입력 영상 획득부(510)는 행동 인식을 위한 본적 없는 소스 영상을 획득한다. 여기서, 본적 없는 소스 영상은 행동 인식을 위한 학습 시 입력되지 않은 영상을 의미한다.

[0070] 신경망 처리부(520)는 획득된 소스 영상을 입력으로 소스 특징 데이터를 생성한다. 신경망 처리부(520)는 컨볼루션 뉴럴 네트워크(CNN: Convolutional Neural Networks) 학습을 기반으로 전처리를 수행하여 소스 특징 데이터를 생성할 수 있다. 여기서, 소스 특징 데이터는 복수의 영상 특징값을 포함할 수 있다.

[0071] 학습 결과 적용부(530)는 소스 영상 데이터의 소스 특징 데이터에 시퀀스 특징 데이터를 학습한 제1 학습 결과와 세그먼트 특징 데이터를 학습한 제2 학습 결과를 적용하며, 영상 판단부(540)는 적용된 학습 결과를 기반으로 소스 영상의 행동을 인식한다.

[0072] 결과 출력부(550)는 인식된 행동을 기반으로 행동 인식 결과를 출력한다.

[0073] 도 6은 본 발명의 실시예에 따른 행동 인식 방법을 설명하기 위한 순서도이다.

[0074] 행동 인식 장치(100)는 행동 인식을 위한 본적 없는 소스 영상을 획득한다(S610). 여기서, 본적 없는 소스 영상은 행동 인식을 위한 학습 시 입력되지 않은 영상을 의미한다.

[0075] 행동 인식 장치(100)는 획득된 소스 영상을 입력으로 신경망 학습 기반의 전처리를 수행하여 영상 특징값을 추출하여 소스 특징 데이터를 생성한다(S620). 행동 인식 장치(100)는 컨볼루션 뉴럴 네트워크(CNN: Convolutional Neural Networks) 학습을 기반으로 전처리를 수행하여 소스 특징 데이터를 생성할 수 있다.

[0076] 행동 인식 장치(100)는 기 학습된 학습 결과를 적용하여 특징값 비교한다(S630). 구체적으로, 행동 인식 장치(100)는 소스 영상 데이터의 소스 특징 데이터에 시퀀스 특징 데이터를 학습한 제1 학습 결과와 세그먼트 특징 데이터를 학습한 제2 학습 결과를 적용하며 특징값을 비교한다.

[0077] 행동 인식 장치(100)는 적용된 학습 결과를 기반으로 소스 영상(입력 영상)의 행동을 판단하고(S640), 인식된 행동을 기반으로 행동 인식 결과를 출력한다(S650).

[0078] 도 6에서는 각 단계를 순차적으로 실행하는 것으로 기재하고 있으나, 반드시 이에 한정되는 것은 아니다. 다시 말해, 도 6에 기재된 단계를 변경하여 실행하거나 하나 이상의 단계를 병렬적으로 실행하는 것으로 적용 가능할 것이므로, 도 6은 시계열적인 순서로 한정되는 것은 아니다.

[0079] 도 6에 기재된 본 실시예에 따른 행동 인식 방법은 애플리케이션(또는 프로그램)으로 구현되고 단말장치(또는 컴퓨터)로 읽을 수 있는 기록매체에 기록될 수 있다. 본 실시예에 따른 행동 인식 방법을 구현하기 위한 애플리케이션(또는 프로그램)이 기록되고 단말장치(또는 컴퓨터)가 읽을 수 있는 기록매체는 컴퓨팅 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록장치 또는 매체를 포함한다.

[0080] 도 7은 본 발명의 실시예에 따른 행동 인식 장치의 학습 동작을 설명하기 위한 예시도이다.

[0081] 비디오 데이터를 학습 과정에 사용할 수 없거나 학습을 위한 비디오 데이터가 없는 경우, 종래의 행동 인식 장치는 비디오에서 추출된 특징들을 평균화하여 처음 보는 이미지의 분류(ZSIC: Zero-shot Image Classification)를 통해 상속된 방법을 통해 행동을 인식하는 동작을 수행한다. 그러나, 이러한 종래의 행동 인식 방법은 비디오의 시계열적인 순차적 정보를 무시하여 행동을 인식하게 됨에 따라 비디오에 포함된 전체의 행동에 대한 인식 오류가 발생할 수 있다.

[0082] 이러한 종래의 문제점을 해결하기 위해, 본 실시예에 따른 행동 인식 장치(100)는 순차적 정보를 고려한 시퀀스 생성적 모델을 통해 단일 샘플이 아니라 본 적 없는 클래스에 대한 일련의 동작을 합성할 수 있도록 하고, 처음 보는 행동에 대한 인식을 완전한 지도 학습 방식으로 전환한다.

[0083] 본 실시예에 따른 행동 인식 장치(100)는 처음 보는 행동을 인식하기 위한 시퀀스를 생성하기 위해 속성 인코더(340), 생성자(350), 감별자(360) 등으로 구성될 수 있다. 구체적으로, 속성 인코더(340)는 시퀀스 생성을 위하여 자연어 벡터를 복수의 벡터로 변환하여 순차적 정보를 부여할 수 있다. 또한, 행동 인식 장치(100)의 시퀀스

생성적 모델은 생성된 시퀀스를 합성시, 행동의 세그먼트 뿐만 아니라, 시퀀스 감별자를 통해 실제 분포로 전체 행동의 시퀀스를 함께 샘플링한다. 여기서, 행동 인식 장치(100)는 순차적 특징 데이터 기반의 생성적 적대 신경망(SFGAN: Sequence Feature Generative Adversarial Network)으로 구현될수 있다.

[0084] 행동 인식 장치(100)는 행동의 특징 시퀀스를 생성하는 생성적 모델을 포함하며, 이러한 모델은 단일 조건에서 시퀀스를 생성하고, 생성된 시퀀스는 현실적이어야 한다.

[0085] 행동 인식 장치(100)는 우리는 시간적인 정보를 포함하는 시맨틱 임베딩 공간을 탐색하고 시퀀스 큐를 조건에서 전개하기 위해 재귀 신경망에 기반한 속성 엔코더(340)를 포함한다. 또한, 행동 인식 장치(100)는 행동의 순서를 무시하는 생성자에 대한 패널티를 적용하기 위한 시퀀스 감별자를 포함한다. 또한, 행동 인식 장치(100)는 제로샷 학습(ZSL: Zero-Shot Learning)의 시간 조건에 상응하는 처음 보는 행동의 특징을 생성할 수 있다.

[0086] 본 발명에 따른 행동 인식 장치(100)는 기존의 준 지도 학습(Semi-Supervised Learning)을 완전한 지도 학습(Fully-Supervised Learning)으로 전환하기 위해 본적 없는 행동을 생성한다.

[0087] 행동 인식 장치(100)는 평균화된 특징을 생성하는 종래의 제로샷 행동 인식 방식과는 달리, 행동 특징에 대한 시퀀스를 생성하는 시퀀스 기반의 생성적 적대 신경망(GAN: Generative Adversarial Network) 모델을 적용하며, 순차적 데이터를 처리하기 위한 속성 인코더(340), 생성자(350), 감별자(360) 등을 포함한다.

[0088] 본 실시예에서 본적 있는 클레스(Class)에 대한 데이터 세트는 Ds로 정의될 수 있고, 본적 있는 데이터 세트 Ds는 $\{(x_v, x_f, y, c(y))|x_v \in X_v^s, x_f \in X_f^s, y \in Y^s, c(y) \in C\}$ 와 같이 표현될 수 있다. 여기서, x_v는 X_v^s의 RGB 시각적 특징이고, x_f는 X_f^s의 광학 흐름 특징이고, y는 Y^s의 클레스 라벨(Class label)을 나타내며, c(y)는 클레스의 의미를 의미론적으로 나타낸 클레스 y의 자연어 임베딩을 의미한다.

[0089] 이와 유사하게, 본 실시예에서 본적 없는 클레스에 대한 데이터 세트는 Du로 정의될 수 있고, Du는 Y^s와 분리되어 있다. 본 실시예에서 본적 없는 데이터 세트 Du는 $\{(x_v, x_f, y, c(y))|x_v \in X_v^u, x_f \in X_f^u, y \in Y^u, c(y) \in C\}$ 와 같이 표현될 수 있다.

[0090] 본 실시예에 따른 행동 인식 장치(100)에서는 처음 보는 행동을 인식(ZSAR: Zero-shot Action Recognition)을 위한 제약을 기반으로, 본적 있는 데이터 세트(Ds)와 본적 없는 데이터 세트(Du) 두 개의 데이터 세트 사이의 포함 관계는 $Y = Y^s \cup Y^u$ 및 $Y^s \cap Y^u = \emptyset$ 를 만족하도록 설정된다.

[0091] 행동의 시퀀스는 N의 특징 벡터 길이로 표현될 수 있고, 여기서 N은 시퀀스의 시간적 길이를 의미한다. 본적 있는 데이터는 행동 인식을 위한 학습 단계에서 접근할 수 있으나, 본적 없는 데이터의 RGB 특징 및 흐름 특징은 테스트 단계에서만 접근할 수 있다.

[0092] 이하, 본 실시예에 따른 행동 인식 장치(100)에서 사용되는 처음 보는 행동의 인식을 위한 생성적 적대 학습 동작(GAN for Zero-shot Action Recognition)을 설명하도록 한다.

[0093] 행동 인식 장치(100)에 적용되는 생성적 적대 신경망(GAN: Generative Adversarial Network)은 생성자(Generator, 350)와 감별자(discriminator, 360) 사이의 최소 극대화 알고리즘(Minimax Algorithm)을 통해 실제 분포에서 샘플을 생성하는 것을 목표로 한다. 여기서, 생성자(350)는 가짜 샘플을 생성하여 감별자(360)을 속이려 하는 동작을 수행하고, 반면 감별자(360)는 실제 샘플을 가짜 샘플과 구별하려 하는 동작을 수행한다.

[0094] 또한, 본 실시예에 따른 행동 인식 장치(100)는 생성적 적대 신경망의 학습 안정성을 위해 그라디언트 패널티(gradient penalty)가 있는 목적 함수로 Wasserstein 거리를 조정한다. 행동 인식 장치(100)에서 본적 없는 클래스에서 샘플을 생성하기 위하여 생성 모델은 조건부 WGAN(Wasserstein GAN)을 기반으로 생성한다.

[0095] 행동 인식 장치(100)에서 사용되는 목적 함수는 수학식 1과 같이 정의될 수 있다.

수학식 1

[0096]
$$L_{cWGAN} = \underset{x\sim P_r}{E}[D(x,c)] - \underset{\tilde{x}\sim P_g}{E}[D(\hat{x},c)] + \gamma E[(\|\nabla D(\hat{x},c)\|_2 - 1)^2],$$

[0097] 여기서 P_r과 P_g는 실제 분포와 생성된 분포를 의미하고, $\tilde{x}$ 는 생성자(350)의 출력을 의미하고, $\hat{x}$ 는 x와 $\tilde{x}$ 의 보간을 의미하며, 마지막 항은 페널티를 주어 그라디언트의 폭발(Gradient Exploding)하는 것을 방지하는 정규화항이며, γ 는 항의 매개 변수를 의미한다.

[0098] 이하, 본 실시예에 따른 행동 인식 장치(100)에서 본적 없는 행동 시퀀스를 생성하는 동작(Generating Unseen Action Sequence)을 설명하도록 한다.

[0099] 행동 인식을 위한 비디오를 생성하는 것은 단일 프레임을 생성하는 것보다 어려운 동작이다. 비디오는 시간 축과 함께 더 복잡하므로, 동작 시퀀스를 완료하기 위하여 생성된 세그먼트가 조립될 때 각 세그먼트 사이의 간격은 자연스럽게 연결되어야 한다.

[0100] 따라서, 본 실시예에 따른 행동 인식 장치(100)에서는 2 개의 조건을 기반으로 본적 없는 클레스의 비디오 특징 시퀀스를 생성한다. 첫 번째 조건은 단일 조건에서 시퀀스를 생성하는 것이고, 두 번째 조건은 시퀀스의 충실도를 보장하기 위하여 복수의 특징을 결합하여 시퀀스를 생성하는 것이다. 여기서, 단일 조건은 하나의 자연어 벡터를 의미하는 것이 바람직하나 반드시 이에 한정되는 것은 아니다.

[0101] 행동 인식 장치(100)는 본적 없는 비디오의 특징 시퀀스의 생성을 위해 단일 조건만 제공되는 경우, 생성자(350)에서는 두 가지 방법을 이용하여 시퀀스를 합성할 수 있다. 여기서, 두 가지 방법은 단일 조건에서 전체 기능을 간단히 생성하는 일대다 매핑과 주어진 조건을 생성하기 전에 예상 길이의 복수의 조건으로 확장하는 일대일 매핑일 수 있다. 여기서, 실제 비디오 생성하는 동작을 참고하면, 일대다 매핑을 수행할 경우, 부족한 조건과 네트워크 용량으로 인해 본적 없는 비디오의 특징 시퀀스를 생성하기는 어렵다. 따라서, 본 발명의 행동 인식 장치(100)에서는 재귀 신경망(RNN: Recurrent Neural Network)을 통해 시간 정보를 단일 조건에서 전개할 수 있는 속성 엔코더(340)를 포함한다. 즉, 속성 인코더(340)를 통해 시간 정보가 포함된 시맨틱 임베드 공간을 탐색한다.

[0102] 다음으로, 행동 인식 장치(100)는 생성된 본적 없는 비디오의 특징 시퀀스의 충실도를 보장해야 한다. 행동 인식 장치(100)에서 단일 조건은 복수 개로 확장되고, 확장된 조건에서 세그먼트가 생성되고, 생성된 세크먼트를 수집하여 행동 시퀀스가 생성된다. 생성된 행동 시퀀스 전체의 흐름은 실제 행동 시퀀스만큼 자연스럽게 연결되어야만 한다.

[0103] 따라서, 생성 모델은 시각적 공간에서 세그먼트와 시퀀스의 분포를 동시에 탐색해야만 한다. 이를 위해, 본 발명의 행동 인식 장치(100)의 감별자(360)는 실제 행동 시퀀스를 가짜 시퀀스와 구별하기 위한 시퀀스 감별자(372)를 포함한다.

[0104] 이하, 본 실시예에 따른 행동 인식 장치(100)에서 적용된 행동 특징 기반의 생성적 적대 신경망(Action Feature Generative Adversarial Networks)의 동작을 설명하도록 한다.

[0105] 도 7에서는 본 발명에 따른 행동 인식 장치(100)에 적용된 행동 시퀀스 특징 기반의 생성적 적대 신경망(SFGAN: Sequence Feature Generative Adversarial Networks)의 세부 구조를 나타낸다. 도 7을 참조하면, 행동 인식 장치(100)는 인코더(340), 생성자(350), 감별자(372, 374) 등으로 구성된 행동 시퀀스 특징 기반의 생성적 적대 신경망을 통해 본 적 없는 행동을 인식하기 위한 학습을 수행한다.

[0106] 이하, 본 실시예에 따른 행동 인식 장치(100)에 포함된 속성 인코더(340)에 대해 설명하도록 한다.

[0107] 인코더(340)는 입력된 단일 조건을 인코딩하여 출력값 c(y)을 출력한다. 여기서, 인코더(340)는 입력된 단일 조건의 시간 스트림을 풀기 위하여 재귀 신경망을 사용한다. 예를 들어, 인코더(340)는 자연어 벡터를 단일 조건으로 입력 받고, 자연어 벡터를 인코딩하여 자연어 특징 데이터를 출력할 수 있다.

[0108] 또한, 인코더(340)는 GRU(Gated Recurrent Unit) 셀로 구성될 수 있으며, 인코더(340)의 GRU 동작은 수학식 2와 같이 정의될 수 있다.

수학식 2

[0109] $$\hat{c}^{k+1} = GRU(\hat{c}^{k})$$

[0110] 여기서 $\hat{c}^{0}$ = c(y)이고 k 는 0 < k < N, k ∈ N을 만족시킵니다.

[0111] 인코더(340)는 단일 조건에서 다수의 조건으로 푸는 동작에 의해 잠재 공간에서 불연속이 발생한다. 따라서, 인코더(340)는 컨디셔닝 증강 기술(Conditioning Augmentation)을 추가로 사용한다.

[0112] 다수의 조건으로 확장된 각각의 조건은 가우스 분포 $\mathcal{N}(\mu(a^{i}), \Sigma(a^{i}))$에서 표본으로 다시 매개 변수화되며, 여기서 μ는 평균을 의미하고, Σ는 공분산 행렬을 의미한다.

[0113] 인코더(340)는 시맨틱 공간을 과도하게 조정하고 매끄러움을 강화하는 것을 방지하기 위하여 KL-divergence(Kullback-Leibler divergence)를 정규화 용어로 사용한다.

[0114] 따라서, 도 10에 도시된 바와 같이, 인코더(340)는 $\hat{c}^{0\sim N-1}$에서 매개 변수화된 조건 $a^{0\sim N-1}$는 생성자(350)으로 전달되어 생성자(350)의 입력 조건의 역할을 한다.

[0115] 또한, 생성자(350)에서 본적 없는 특징을 생성하기 위해서 조건 $a^{0\sim N-1}$는 행동 사이의 관계정보를 포함해야 한다. 이에, 본 발명의 인코더(340)는 삼중항 손실함수를 사용하며, 삼중항 손실 함수는 GRU에 의해 처리된 조건을 원래 조건과 유사하게 처리하고 다른 행동의 조건과는 다르게 처리한다. 인코더(340)에 삼중항 손실 함수에서 사용되는 목적 함수 및 정규화 용어는 수학식 3 및 4와 같이 정의될 수 있다.

수학식 3

[0116] $$D_{KL}(\mathcal{N}(\mu(\hat{c}^{i}), \sigma(\hat{c}^{i}))||\mathcal{N}(0, I))$$

수학식 4

[0117] $$L_{tri} = max(0, m + d^{+}(c(y), \hat{c}^{i}) - d^{-}(c(y), \hat{c}^{j}))$$

[0118] 여기서, d^{+}는 파지티브(positive) 쌍의 거리를 의미하고, d^{-}는 네거티브(negative) 쌍의 거리를 의미하며, c(y), $\hat{c}^{i}$, $\hat{c}^{j}$ 각각은 앵커(anchor), 파지티브 샘플 및 네거티브 샘플이다. m 은 삼중항 손실의 마진이며, 코사인 유사성을 삼중항 손실 거리 측정법으로 사용한다. $\hat{c}^{i}$는 동일한 클립의 피처에서 샘플링되고 네거티브는 다른 동작의 클립에서 샘플링된다.

[0119] 이하, 본 실시예에 따른 행동 인식 장치(100)에 포함된 생성자(350)에 대해 설명하도록 한다.

[0120] 본 실시예에 따른 행동 인식 장치(100)는 완전한 지도 학습 방식을 통해 행동을 인식하며, 이러한 방식은 광학적 흐름의 특징을 사용하는 것이 바람직하다.

[0121] 행동 인식 장치(100)는 본적 없는 행동인식을 위하여 생성자(350)를 포함하며, 생성자(350)는 RGB 특징과 흐름

특징이 결합된 결합 특징을 생성한다.

[0122] 생성자(350)는 매개 변수화된 조건 a^t와 잠재 잡음 벡터 z를 입력으로 RGB 특징과 흐름 특징이 결합된 결합 특징을 생성한다.

[0123] 흐름 특징은 원래의 RGB 비전에서 추출됨에 따라, 생성자(350)는 RGB 특징과 흐름 특징 간의 관계를 모델링하기 위해 풀리 커넥티드 레이어(fully connected layer)로 구성된다. 생성자(350)의 동작은 수학식 5와 같이 정의될 수 있다.

수학식 5

[0124] $$\hat{x} = G(a^n, z)$$

[0125] 여기서 z는 잠재 잡음에 대한 랜덤 변수이고, n은 n 번째 임베디드 매개 변수화된 조건을 의미한다.

[0126] 이하, 본 실시예에 따른 행동 인식 장치(100)에 포함된 감별자(360)에 대해 설명하도록 한다.

[0127] 행동 인식 장치(100)에 포함된 감별자(360)는 생성자(350)에서 생성된 특징의 분포와 실제 분포의 차이를 판별하여 생성자(350)에 피드백을 제공한다.

[0128] 본 실시예에 따른 감별자(360)는 도 11에 도시된 바와 같이, 세그먼트에 대한 판별을 위한 세그먼트 감별자(372)와 시퀀스에 대한 판별을 위한 시퀀스 감별자(374)로 구성될 수 있다.

[0129] 세그먼트 감별자(372) 및 시퀀스 감별자(374) 각각은 실제 특징과 실제 시퀀스를 가짜와 구별하기 위한 복수의 풀리 커넥티드 레이어(fully connected layer)로 구성될 수 있다.

[0130] 세그먼트 감별자(372)는 특징과 조건을 동시에 처리하고, 시퀀스 감별자(374)는 특징만을 처리한다.

[0131] 본 실시예에 따른 행동 인식 장치(100)는 본적 없는 행동 시퀀스를 생성하는 것이기 때문에 훈련 중 과도한 컨디셔닝으로 인해 클래스에 편견이 생길 수 있다. 따라서, 행동 인식 장치(100)는 시퀀스 감별자(374)가 포함된 감별자(360)로 구성되어야 한다.

[0132] 이하, 본 실시예에 따른 행동 인식 장치(100)에서 사용되는 목적 함수(Objective function)에 대해 설명하도록 한다.

[0133] 본 실시예에 따른 행동 인식 장치(100)의 모델 학습을 위한 목적 함수는 조건부 Wasserstain GAN을 기반으로 한다. 그러한, 행동 인식 장치(100)에 포함된 시퀀스 감별자(374)는 본적 없는 행동 시퀀스를 생성하기 위해 무조건적으로 설계되었으며, 시퀀스 감별자(374)에서는 일반적인 Wasserstain 거리를 사용한다. 또한, 시퀀스 감별자(374)에 대한 기울기 패널티(gradient penalty)는 수학식 6과 같이 정의된다.

수학식 6

[0134] $$R_{uncond} := \gamma E[(\|\nabla D(\hat{x})\|_2 - 1)^2]$$
$$R_{cond} := \gamma E[(\|\nabla D(\hat{x}, a)\|_2 - 1)^2]$$

[0135] 여기서 R_{uncond}은 시퀀스 감별자(374)에 대한 무조건부 정규화를 의미하고, R_{cond}은 세그먼트 감별자(372)에 대한 조건부 정규화를 의미한다. 따라서, 생성 모델에 대한 손실 함수는 수학식 7과 같이 정의될 수 있다.

수학식 7

$$L_{WGAN} = E[D_{seg}(\hat{x}^n, a^n)] - E[D_{seg}(x^n, a^n)] + E[D_{seg}(\hat{X}_{seq})] - E[D_{seg}(X_{seq})] + R_{uncond} + R_{seq},$$

[0136]

[0137] 여기서, 0 ≤ n <N 이며, a^n은 인코딩된 조건(특징 데이터)를 나타내고, x는 실제 특징 데이터의 샘플이다. 또한, $\hat{x}$ 은 생성자(350)에서 생성된 대상 특징 데이터를 나타내며, $\hat{X}_{seq} = \{\hat{x}^0, \ldots, \hat{x}^{N-1}\}$ 이고, $X_{seq} = \{x^0, \ldots, x^{N-1}\}$ 이다. R_{uncond} 및 R_{cond}는 각각 Dseq 및 Dseg에 대한 정규화 용어를 의미한다.

[0138] 결과적으로, 행동 인식 장치(100)에서 사용되는 매개 변수가 있는 엔드-투-엔드 모델의 전체 목적 함수는 수학식 8로 정의될 수 있다.

수학식 8

$$minimize\ \alpha L_{WGAN} + \beta L_{tri} + \gamma D_{KL}$$

[0139]

[0140] 이하, 본 실시예에 따른 행동 인식 장치(100)에서 본적 없는 행동을 인식하는 동작에 대해 설명하도록 한다.

[0141] 행동 인식 장치(100)는 본적 있는 데이터 세트(Ds)를 생성적 적대 신경망을 통해 학습한 후, 본적 없는 클래스의 조건으로부터 본적 없는 행동 특징 $\hat{X}^u$ 을 생성한다.

[0142] 행동 인식 장치(100)는 처음보는 행동을 인식하기 위한 문제를 행동 인식을 위한 완전한 지도 학습 방식으로 처리하고, 평가시에는 다중 계층 퍼셉트론 분류기(Multi-Layer Perceptron classier)를 사용한다. 여기서, 분류기는 음의 로그 우도 손실을 최소화하여 최적화되며, 수학식 9와 같이 정의될 수 있다.

수학식 9

$$\min_{\theta} - \sum_{(x,y)\in F} logP(y|x;\theta)$$

[0143]

[0144] 여기서, Θ는 분류기에서 풀리 커넥티드 레이어(fully connected layer)의 가중치이고, F는 GZSL 또는 ZSL일 때 $\mathcal{D}_u \cup \mathcal{D}_s$ 또는 Du를 의미한다. 또한, 분류를 위한 예측 함수는 수학 식 10과 같이 정의될 수 있다.

수학식 10

$$f(x) = \arg\max_{y} P(y|x)$$

[0145]

[0146] 여기서 softmax 함수는 $P(y|x) = \frac{\exp(x)}{\sum \exp(x)}$ 이며, GZSL에서 $y \in Y^s \cup Y^u$, ZSL에서 $y \in Y^u$를 의미한다.

[0147] 도 8은 본 발명의 실시예에 따른 입력 영상을 처리하여 특징 데이터를 생성하는 동작을 설명하기 위한 예시도이다.

[0148] 도 8을 참조하면, 행동 인식 장치(100)는 소스 영상을 획득한다. 여기서, 소스 영상은 비디오 클립(810)을 의미하며, 비디오 클립(810)은 5 개의 영상 세그먼트(811, 812, 813, 814, 815)로 구성될 수 있다. 여기서, 비디오 클립(810)은 농구, 야구, 출구 등에 대한 행동을 포함하는 클립일 수 있다. 영상 세그먼트(811, 812, 813, 814, 815) 각각은 32 개의 움직임 벡터 영상 프레임을 포함할 수 있다. 움직임 벡터 영상 프레임(820) 사이에는 차분 영상(821)이 추가로 포함될 수 있으며, 차분 영상(821)은 인접한 두 개의 움직임 벡터 영상 프레임(820)의 차이를 통해 생성된 영상을 의미한다.

[0149] 도 8을 참조하면, 행동 인식 장치(100)는 컨볼루션 뉴럴 네트워크(CNN: Convolutional Neural Networks) 학습을 위한 전처리(Pre-traning)를 수행하여 소스 특징 데이터(X)를 생성할 수 있다. 여기서, 소스 특징 데이터는 복수의 세그먼트 단위 별 특징값(831, 832, 833, 834, 835)를 포함하며, 각각의 특징값은 1×1024의 크기를 갖는 행렬 특징값일 수 있다.

[0150] 도 9는 본 발명의 실시예에 따른 자연어 벡터를 처리하여 특징 데이터를 생성하는 동작을 설명하기 위한 예시도이다.

[0151] 인코더(340)는 자연어 벡터(910)를 입력으로 적어도 하나의 특징값을 포함하는 자연어 특징 데이터를 생성하는 동작을 수행한다. 여기서, 자연어 벡터는 시계열적인 정보를 포함하지 않고, 소정의 행동에 대하여 자연어 기반으로 생성된 벡터를 의미한다.

[0152] 인코더(340)는 자연어 벡터(910)에 순차적 정보를 부여하여 복수 개의 벡터로 확장하고, 복수 개의 벡터 각각에 대응되는 적어도 하나의 특징값(921, 922, 923, 924, 925) 각각을 생성한다.

[0153] 인코더(340)는 자연어 벡터의 평균, 표준 편차 및 노이즈 등 중 적어도 하나를 이용하여 자연어 벡터를 정규 분포 상에서 분포를 갖는 적어도 하나의 특징값(921, 922, 923, 924, 925) 각각을 생성한다.

[0154] 제1 특징값 처리부(342)는 인코더(340)에서 출력된 적어도 하나의 특징값(921, 922, 923, 924, 925)을 포함하는 자연어 특징 데이터를 생성자(350)로 전달하는 동작을 수행한다. 제1 특징값 처리부(342)는 자연어 특징 데이터에 랜덤 변수(잠재 잡음에 대한 랜덤 변수)를 추가로 결합시켜 생성자(350)로 전달할 수 있다.

[0155] 또한, 제1 특징값 처리부(342)는 인코더(340)에서 출력된 적어도 하나의 특징값(921, 922, 923, 924, 925)을 포함하는 자연어 특징 데이터를 제2 감별자(374)로 전송한다.

[0156] 한편, 제1 특징값 처리부(342)는 인코더(340)에서 자연어 특징 데이터를 생성자(350)로 직접 전달하는 경우 생략되거나, 인코더(340)에 포함된 형태로 구현될 수 있다.

[0157] 생성자(350)는 자연어 특징 데이터를 기반으로 소스 영상의 소스 특징 데이터와 분류를 위한 대상 특징 데이터를 생성하는 동작을 수행한다.

[0158] 생성자(350)는 자연어 특징 데이터와 기 생성된 랜덤 변수를 기반으로 페이크(Fake) 영상에 대한 대상 특징 데이터를 생성한다. 여기서, 생성자(350)는 컨볼루션 뉴럴 네트워크(CNN: Convolutional Neural Networks) 학습을 통해 상기 대상 특성 데이터를 생성하는 것이 바람직하나 반드시 이에 한정되는 것은 아니다.

[0159] 생성자(350)는 적어도 하나의 특징값(931, 932, 933, 934, 935)을 포함하는 대상 특징 데이터를 생성한다. 여기서, 생성자(350)는 자연어 특징 데이터와 동일한 개수의 세그먼트 단위로 대상 특징 데이터를 생성한다. 여기서, 세그먼트 단위는 대상 특징 데이터에 포함된 각각의 특징값으로 구분될 수 있다.

[0160] 제2 특징값 처리부(352)는 생성자(350)에서 출력된 대상 특징 데이터를 감별자(360)로 전달하는 동작을 수행한다. 제2 특징값 처리부(352)는 대상 특징 데이터를 제1 감별자(372) 및 제2 감별자(374) 각각으로 전달한다. 한편, 제2 특징값 처리부(352)는 생성자(350)에서 대상 특징 데이터를 감별자(360)로 직접 전달하는 경우 생략되거나, 생성자(350)에 포함된 형태로 구현될 수 있다.

[0161] 도 10은 본 발명의 실시예에 따른 인코더의 동작 구성을 나타낸 도면이다.

[0162] 인코더(340)는 자연어 벡터에 순차적 정보를 부여하여 복수 개의 벡터로 확장을 수행한다. 여기서, 단일 조건의 자연어 벡터는 LSTM(Long short-term memory), GRU(Gated recurrent unit) 등의 방식을 이용하여 확장될 수 있다.

[0163] 또한, 인코더(340)는 복수 개의 벡터 각각에 대응되는 적어도 하나의 특징값 각각을 생성한다. 인코더(340)는 자연어 벡터의 평균(μ), 표준 편차(σ) 및 노이즈(ε) 등 중 적어도 하나를 이용하여 자연어 벡터를 정규 분포 상에서 분포를 갖는 적어도 하나의 특징값(a^i) 각각을 생성한다.

[0164] 도 11은 본 발명의 실시예에 따른 감별자의 동작 구성을 나타낸 도면이다.

[0165] 감별자(360)는 소스 특징 데이터와 자연어 특징 데이터, 대상 특징 데이터 등 중 적어도 하나를 기반으로 시퀀스(Sequence) 및 세그먼트(Segment) 각각에 대한 분류를 처리하여 객체의 행동 인식이 수행되도록 한다. 본 실시예에 따른 감별자(360)는 제1 감별자(372) 및 제2 감별자(374)를 포함한다.

[0166] 제1 감별자(372)는 대상 특징 데이터와 소스 특징 데이터를 이용하여 시퀀스(Sequence)에 대한 분류를 처리하는 동작을 수행한다. 제1 감별자(372)는 대상 특징 데이터와 소스 특징 데이터를 입력 받고, 대상 특징 데이터의 진위 여부를 판별할 수 있다. 구체적으로, 제1 감별자(372)는 순차적 정보가 포함된 복수의 소스 특징값을 결합(Concatenation)한 소스 특징 데이터와 순차적 정보가 포함된 복수의 대상 특징값을 결합한 대상 특징 데이터를 비교하여 대상 특징 데이터의 진위 여부를 학습한 제1 학습 결과를 출력한다. 여기서, 제1 학습 결과는 [0, 1] 사이의 값으로 표현될 수 있다. 제1 감별자(372)에서 대상 특징 데이터의 진위 여부의 판단 결과, 0 값에 가까울수록 거짓(Fake) 신호로 분류된 것이고 1 값에 가까울수록 참(Real) 신호로 분류된 것이다.

[0167] 제2 감별자(374)는 자연어 특징 데이터 및 대상 특징 데이터를 결합한 대상 결합 데이터와 소스 특징 데이터를 이용하여 세그먼트(Segment)에 대한 분류를 처리하는 동작을 수행한다. 제2 감별자(374)는 대상 결합 데이터와 소스 특징 데이터를 입력 받고, 대상 결합 데이터의 진위 여부를 판별할 수 있다.

[0168] 구체적으로, 제2 감별자(374)는 소스 특징 데이터의 세그먼트 단위와 대상 결합 데이터의 세그먼트 단위를 비교하여 대상 결합 데이터의 진위 여부를 학습한 제2 학습 결과를 출력한다. 여기서, 제2 학습 결과는 [0, 1] 사이의 값으로 표현될 수 있다. 제2 감별자(374)에서 대상 결합 데이터의 진위 여부의 판단 결과, 0 값에 가까울수록 거짓(Fake) 신호로 분류된 것이고 1 값에 가까울수록 참(Real) 신호로 분류된 것이다.

[0169] 제2 감별자(374)는 소스 특징 데이터의 세그먼트 단위의 데이터와 자연어 특징 데이터의 특징값과 대상 특징 데이터의 특징값을 결합((Concatenation))한 세그먼트 단위의 대상 결합 데이터를 비교하여 세그먼트에 대한 분류를 처리할 수 있다.

[0170] 이상의 설명은 본 발명의 실시예의 기술 사상을 예시적으로 설명한 것에 불과한 것으로서, 본 발명의 실시예가 속하는 기술 분야에서 통상의 지식을 가진 자라면 본 발명의 실시예의 본질적인 특성에서 벗어나지 않는 범위에서 다양한 수정 및 변형이 가능할 것이다. 따라서, 본 발명의 실시예들은 본 발명의 실시예의 기술 사상을 한정하기 위한 것이 아니라 설명하기 위한 것이고, 이러한 실시예에 의하여 본 발명의 실시예의 기술 사상의 범위가 한정되는 것은 아니다. 본 발명의 실시예의 보호 범위는 아래의 청구범위에 의하여 해석되어야 하며, 그와 동등한 범위 내에 있는 모든 기술 사상은 본 발명의 실시예의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

부호의 설명

[0171] 100: 행동 인식 장치

110: 입력부 120: 출력부

130: 프로세서 140: 메모리

150: 데이터 베이스

310: 영상 획득부 320: 전처리부

322: 영상 특징값 처리부 330: 자연어 벡터 획득부

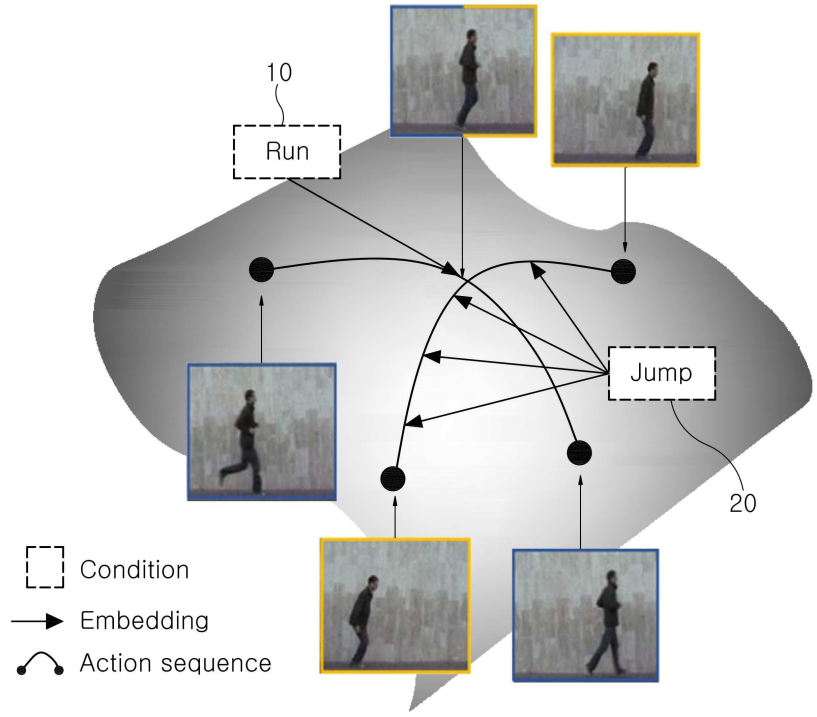
340: 인코더 342: 제1 특징값 처리부

350: 생성자 352: 제2 특징값 처리부

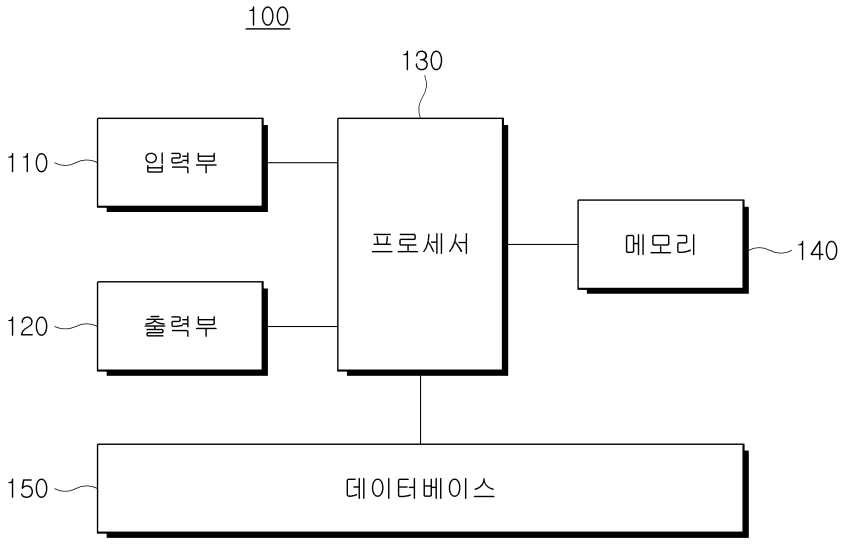
360: 감별자

도면

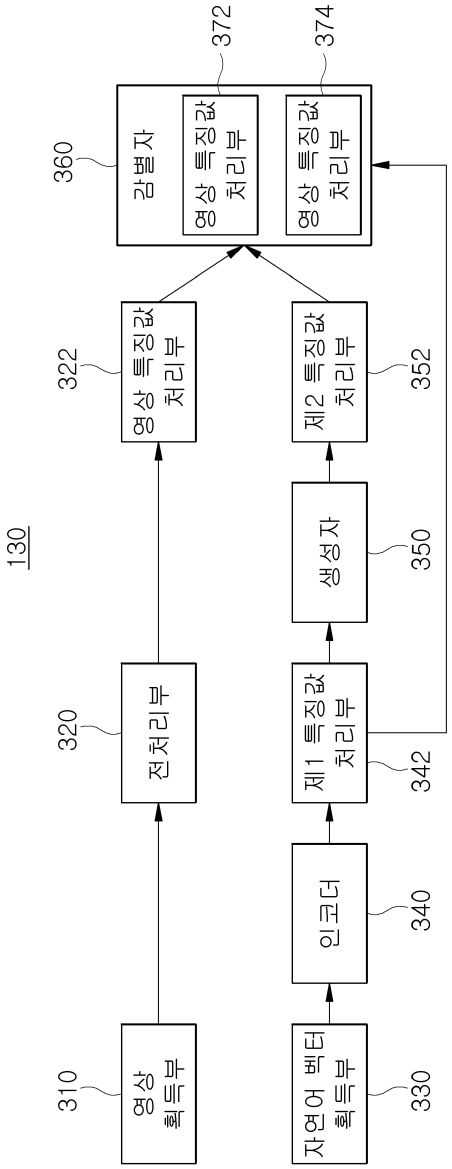
도면1


10
Run
Jump
20
Condition
Embedding
Action sequence

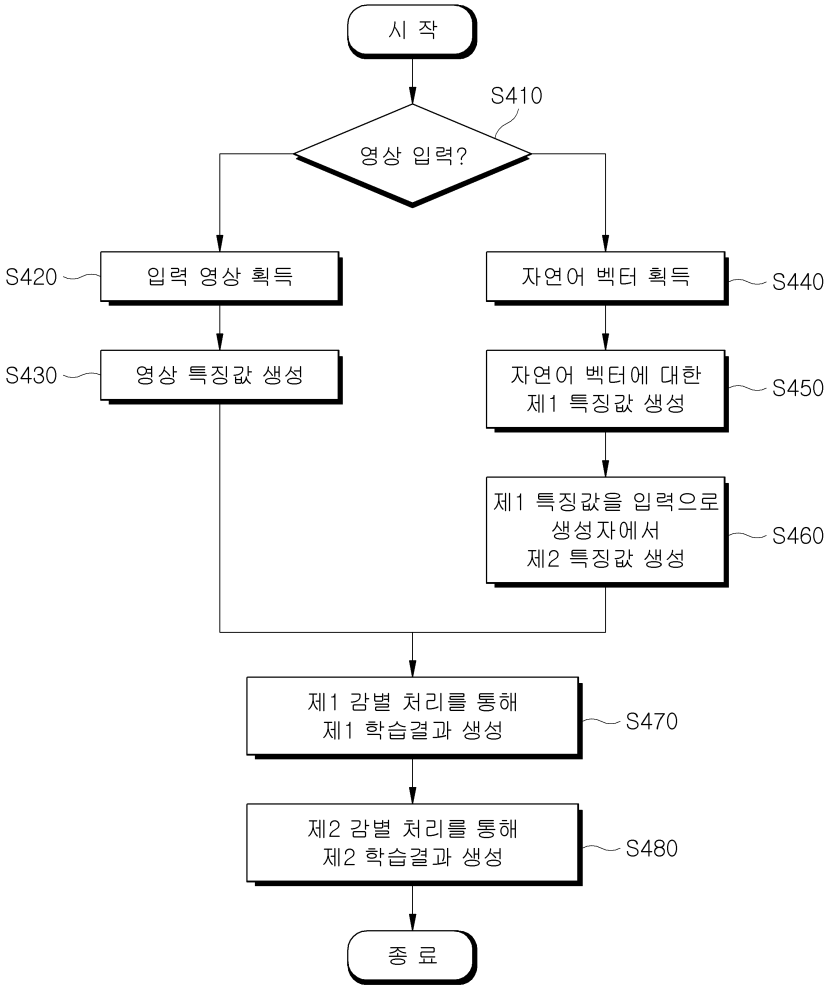
도면2


100
130
110
입력부
120
출력부
프로세서
메모리
140
150
데이터베이스

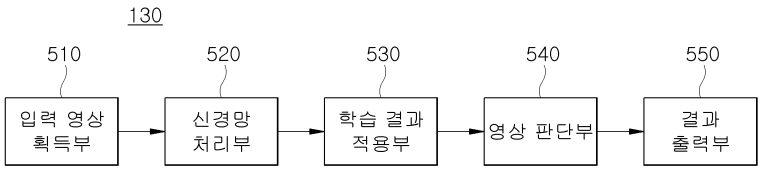
도면3


130
310
영상 획득부
320
전처리부
322
영상 특징값 처리부
330
자연어 벡터 획득부
340
인코더
342
제1 특징값 처리부
350
생성자
352
제2 특징값 처리부
360
감별자
372
영상 특징값 처리부
374
영상 특징값 처리부

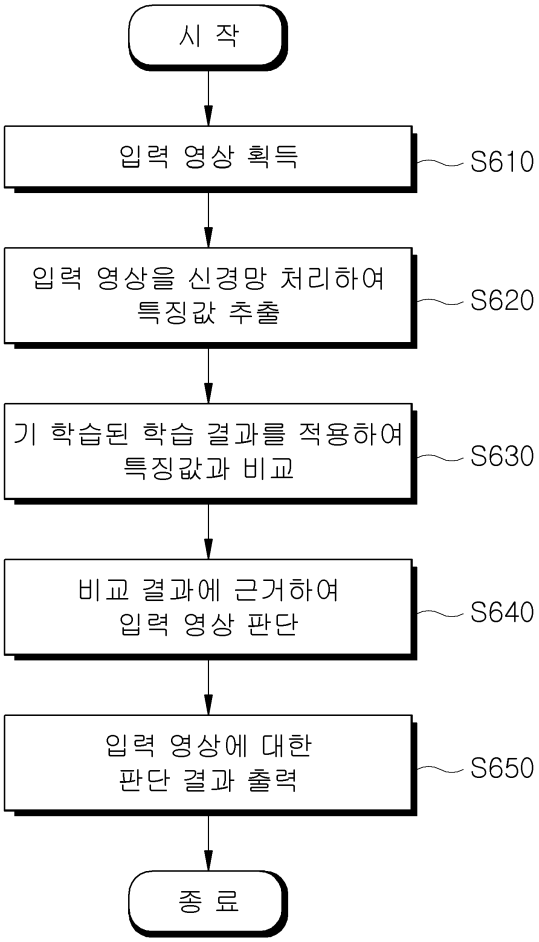
도면4


시 작
S410
영상 입력?
S420
입력 영상 획득
S430
영상 특징값 생성
자연어 벡터 획득
S440
자연어 벡터에 대한
제1 특징값 생성
S450
제1 특징값을 입력으로
생성자에서
제2 특징값 생성
S460
제1 감별 처리를 통해
제1 학습결과 생성
S470
제2 감별 처리를 통해
제2 학습결과 생성
S480
종 료

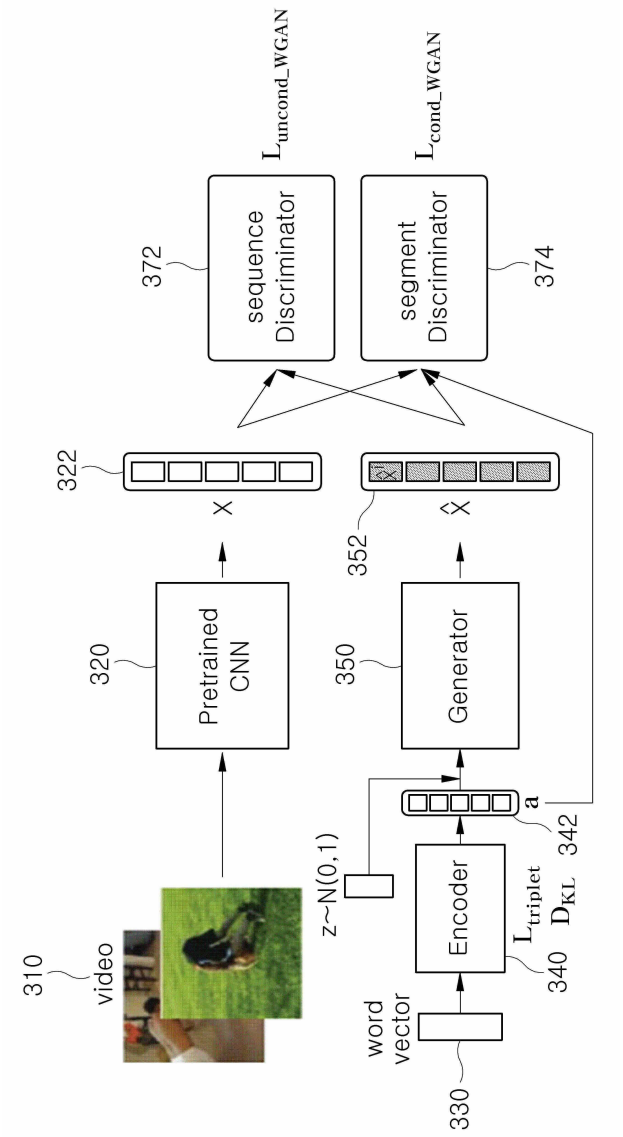
도면5


130
510
입력 영상
획득부
520
신경망
처리부
530
학습 결과
적용부
540
영상 판단부
550
결과
출력부

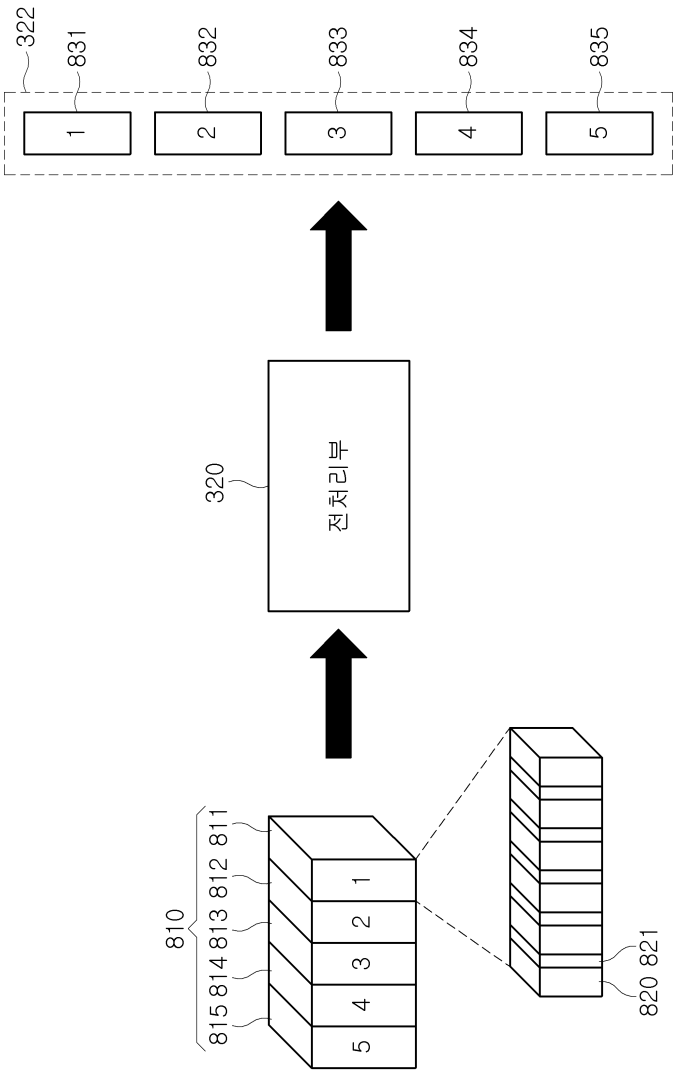
도면6


시 작
입력 영상 획득
S610
입력 영상을 신경망 처리하여 특징값 추출
S620
기 학습된 학습 결과를 적용하여 특징값과 비교
S630
비교 결과에 근거하여 입력 영상 판단
S640
입력 영상에 대한 판단 결과 출력
S650
종 료

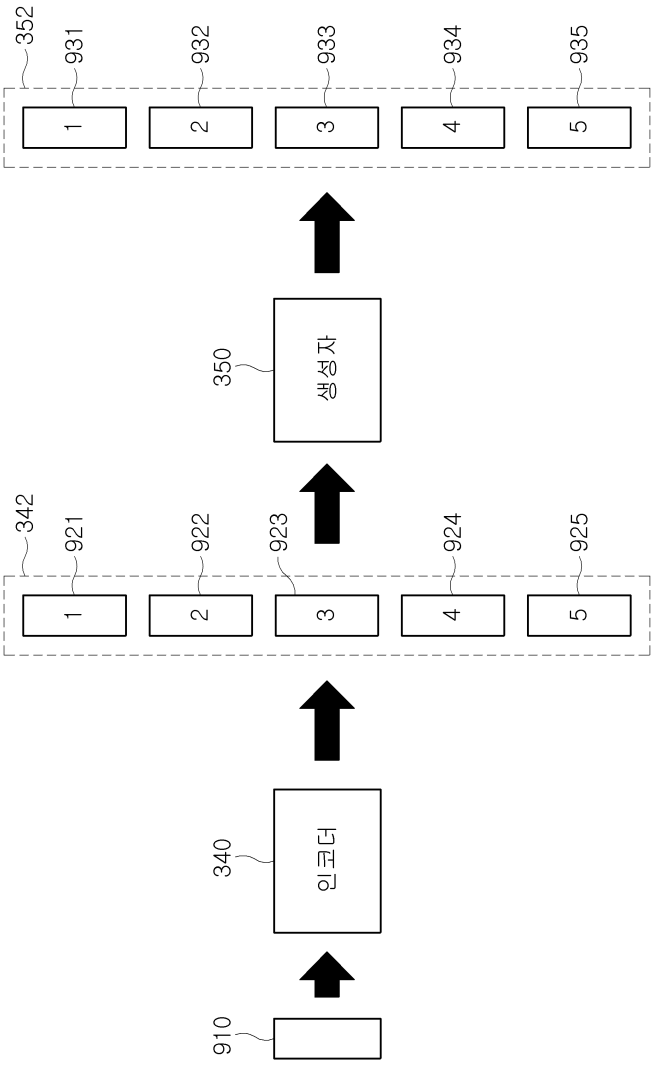
도면7


L_{uncond_WGAN}
L_{cond_WGAN}
372
sequence Discriminator
segment Discriminator
374
322
352
x
$\hat{x}$
320
Pretrained CNN
350
Generator
a
342
z~N(0,1)
Encoder
$L_{triplet}$ D_{KL}
340
310
video
word vector
330

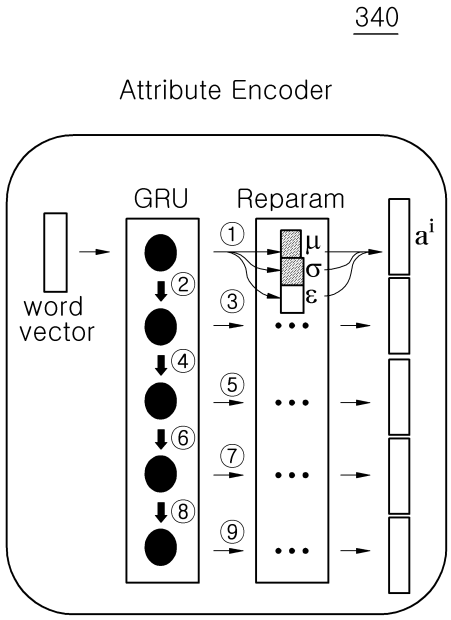
도면8


322
831
832
833
834
835
1
2
3
4
5
320
전처리부
810
811
812
813
814
815
1
2
3
4
5
820
821

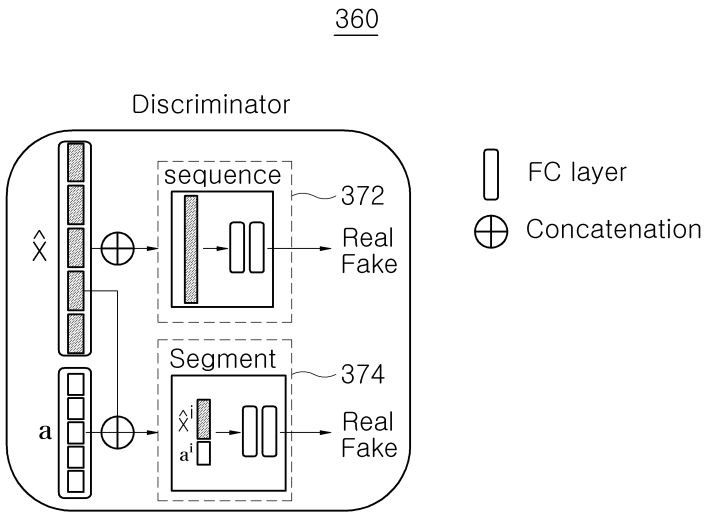
도면9


352
931
932
933
934
935
1
2
3
4
5
350
생성자
342
921
922
923
924
925
1
2
3
4
5
340
인코더
910

도면10


340
Attribute Encoder
GRU
Reparam
word vector
μ
σ
ε
a^i
①
②
③
④
⑤
⑥
⑦
⑧
⑨

도면11


360
Discriminator
sequence
372
Real
Fake
Segment
374
Real
Fake
a
FC layer
Concatenation