(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0128418
(43) 공개일자 2024년08월26일

(51) 국제특허분류(Int. Cl.)
G06V 10/74 (2022.01) *G06N 3/08* (2023.01)
G06V 10/40 (2022.01) *G06V 10/774* (2022.01)
G06V 10/88 (2022.01) *G06V 40/20* (2022.01)
(52) CPC특허분류
G06N 3/088 (2023.01)
(21) 출원번호 10-2023-0021570
(22) 출원일자 2023년02월17일
심사청구일자 2023년02월17일

(71) 출원인
한국과학기술원
대전광역시 유성구 대학로 291(구성동)
울산과학기술원
울산광역시 울주군 언양읍 유니스트길 50
(72) 발명자
성영철
대전광역시 유성구 대학로 291 (구성동, 한국과학기술원)
최성호
대전광역시 유성구 대학로 291 (구성동, 한국과학기술원)
(뒷면에 계속)
(74) 대리인
양성보

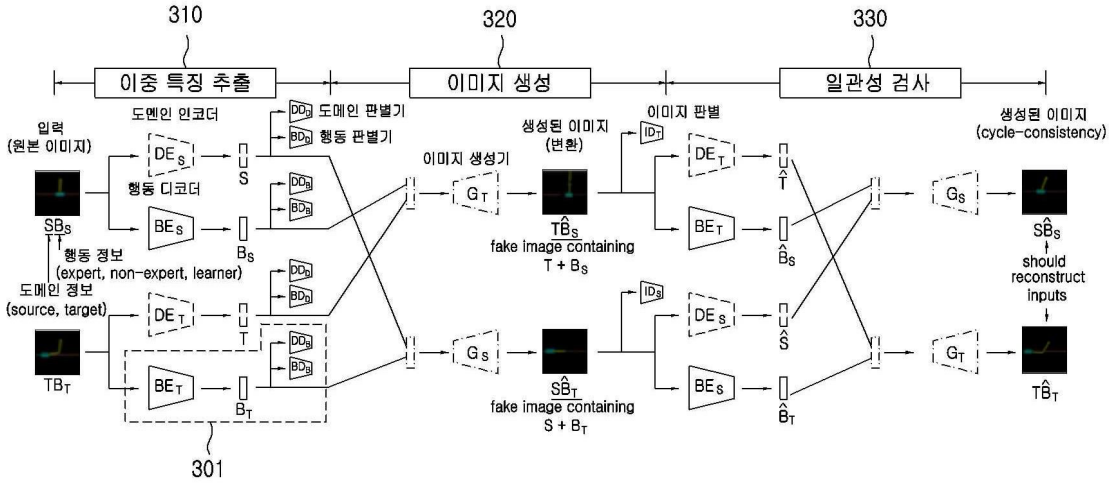
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 **시각적 관찰데이터를 활용한 도메인 적응형 모방학습 방법 및 시스템**

(57) 요 약

시각적 관찰데이터를 활용한 도메인 적응형 모방학습 방법 및 시스템이 개시된다. 일 실시예에 따른 모방학습 시스템에 의해 수행되는 도메인 적응형 모방학습 방법은, 원본 이미지의 소스 도메인과 타겟 도메인 각각에 채택된 인코더로부터 이중 특징(dual feature)을 추출하는 단계; 적대적 학습을 기반으로 훈련된 이미지 생성기를 이용하여 상기 추출된 이중 특성을 결합함에 따라 이미지를 재구성하는 단계; 및 상기 재구성된 이미지의 일관성 검사를 통해 상기 재구성된 이미지가 상기 원본 이미지와 동일한 지 여부를 판단하는 단계를 포함할 수 있다.

대 표 도



(72) 발명자

김우준

대전광역시 유성구 대학로 291 (구성동, 한국과학기술원)

채종성

대전광역시 유성구 대학로 291 (구성동, 한국과학기술원)

정휘영

대전광역시 유성구 대학로 291 (구성동, 한국과학기술원)

한승열

울산광역시 울주군 언양읍 유니스트길 50

이 발명을 지원한 국가연구개발사업

과제고유번호	1711160557
과제번호	2022-0-00124-001
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신.방송 기술개발사업
연구과제명	(N01220828)(통합EZ)스스로 학습역량을 인지하고 활용하여 적정한 결과를 제공하는 인공지능 기술 개발(2022년도)
기 여 율	90/100
과제수행기관명	한국전자통신연구원
연구기간	2022.04.01 ~ 2022.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711152473
과제번호	2020-0-01336-003
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성
연구과제명	인공지능대학원지원(울산과학기술원)
기 여 율	5/100
과제수행기관명	울산과학기술원
연구기간	2022.01.01 ~ 2022.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711174108
과제번호	00156361
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성(R&D)
연구과제명	지역지능화혁신인재양성(울산과학기술원)
기 여 율	5/100
과제수행기관명	울산과학기술원
연구기간	2022.07.01 ~ 2022.12.31

명 세 서

청구범위

청구항 1

모방학습 시스템에 의해 수행되는 도메인 적응형 모방학습 방법에 있어서,

원본 이미지의 소스 도메인과 타겟 도메인 각각에 채택된 인코더로부터 이중 특징(dual feature)을 추출하는 단계;

적대적 학습을 기반으로 훈련된 이미지 생성기를 이용하여 상기 추출된 이중 특성을 결합함에 따라 이미지를 재구성하는 단계; 및

상기 재구성된 이미지의 일관성 검사를 통해 상기 재구성된 이미지가 상기 원본 이미지와 동일한 지 여부를 판단하는 단계

를 포함하는 적응형 모방학습 방법.

청구항 2

제1항에 있어서,

상기 이중 특징을 추출하는 단계는,

상기 소스 도메인과 타겟 도메인의 각각에 행동 인코더와 도메인 인코더를 포함하는 복수 개의 인코더를 채택하고, 상기 채택된 복수 개의 인코더를 통해 상기 소스 도메인과 타겟 도메인의 각각에 대한 도메인 특징 및 행동 특징을 추출하는 단계

를 포함하고,

상기 행동 인코더는, 도메인 정보를 제외한 행동 정보를 포함하는 행동 특징을 추출하고,

상기 도메인 인코더는, 행동 정보를 제외한 도메인 정보를 포함하는 도메인 특징을 추출하는

것을 특징으로 하는 도메인 적응형 모방학습 방법.

청구항 3

제1항에 있어서,

상기 이미지를 재구성하는 단계는,

상기 타겟 도메인의 도메인 특징과 상기 소스 도메인의 행동 특징을 결합한 후 상기 타겟 도메인의 제1 이미지 생성기(G_T)를 이용하여 제1 이미지($\widehat{TB_S}$)를 생성하고, 상기 소스 도메인의 도메인 특징과 상기 타겟 도메인의 행동 특징을 결합한 후 상기 소스 도메인의 제2 이미지 생성기(G_S)를 사용하여 제2 이미지($\widehat{SB_T}$)를 생성하는 단계

를 포함하는 도메인 적응형 모방학습 방법.

청구항 4

제3항에 있어서,

상기 제1 이미지 생성기 및 제2 이미지 생성기는, 각각의 이미지 판별기를 통해 적대적 학습으로 훈련되는 것을 특징으로 하는 도메인 적응형 모방학습 방법.

청구항 5

제1항에 있어서,

상기 이미지를 재구성하는 단계는,

상기 이중 특징의 추출을 위해 훈련된 복수 개의 인코더를 이용하여 상기 생성된 제1 이미지 및 상기 생성된 제2 이미지의 각각에 대한 도메인 특징 및 행동 특징을 추출하는 단계

를 포함하는 도메인 적응형 모방학습 방법.

청구항 6

제5항에 있어서,

상기 이미지를 재구성하는 단계는,

제1 이미지 생성기 및 제2 이미지 생성기를 통해 상기 생성된 1 이미지 및 상기 생성된 제2 이미지의 각각으로부터 추출된 도메인 특징 및 행동 특징을 이용한 이미지 변환을 통해 재구성된 이미지를 획득하는 단계

를 포함하는 도메인 적응형 모방학습 방법.

청구항 7

제1항에 있어서,

상기 판단하는 단계는,

상기 재구성된 이미지의 일관성 검사와 상기 추출된 이중 특징의 일관성 검사를 수행하는 단계

를 포함하는 도메인 적응형 모방학습 방법.

청구항 8

제1항에 있어서,

상기 도메인 적응형 모방학습 방법은,

보상 생성을 위한 보상 생성용 판별기와 학습자 정책이 적대적 방식을 통해 공동 학습을 수행하되, 상기 보상 생성용 판별기에서 행동 인코더의 행동 특징 정보를 이용하여 학습자와 숙련자의 행동을 구별하도록 학습하고, 상기 학습자 정책에서 상기 행동 특징 정보가 숙련자처럼 보이도록 관찰데이터를 생성하는 방법을 학습하는 것을 특징으로 하는 도메인 적응형 모방학습 방법.

청구항 9

모방학습 시스템에 의해 수행되는 도메인 적응형 모방학습 방법을 실행시키기 위해 컴퓨터 판독 가능한 저장매체에 저장된 컴퓨터 프로그램에 있어서,

상기 도메인 적응형 모방학습 방법은,

원본 이미지의 소스 도메인과 타겟 도메인 각각에 채택된 인코더로부터 이중 특징(dual feature)을 추출하는 단계;

적대적 학습을 기반으로 훈련된 이미지 생성기를 이용하여 상기 추출된 이중 특성을 결합함에 따라 이미지를 재구성하는 단계; 및

상기 재구성된 이미지의 일관성 검사를 통해 상기 재구성된 이미지가 상기 원본 이미지와 동일한 지 여부를 판단하는 단계

를 포함하는 컴퓨터 판독 가능한 저장매체에 저장된 컴퓨터 프로그램.

청구항 10

모방학습 시스템에 있어서,

원본 이미지의 소스 도메인과 타겟 도메인 각각에 채택된 인코더로부터 이중 특징(dual feature)을 추출하는 이중 특징 추출부;

적대적 학습을 기반으로 훈련된 이미지 생성기를 이용하여 상기 추출된 이중 특성을 결합함에 따라 이미지를 재

구성하는 이미지 생성부; 및

상기 재구성된 이미지의 일관성 검사를 통해 상기 재구성된 이미지가 상기 원본 이미지와 동일한 지 여부를 판단하는 일관성 검사부

를 포함하는 모방학습 시스템.

발명의 설명

기 술 분 야

[0001] 아래의 설명은 모방학습 기술에 관한 것이다.

배 경 기 술

[0003] 모방학습은 에이전트가 숙련자가 특정 행동을 수행하는 시연 데이터를 모방함으로써 행동을 학습하는 프레임워크로, 기존 강화학습에서 보상함수를 설계해야 하는 것의 어려움을 해결할 수 있다. 기존의 모방학습의 경우 숙련자와 학습자가 같은 도메인에 있어서 학습자가 숙련자의 시연 데이터를 그대로 모방할 수 있어야 하는데, 이것은 현실성이 떨어지는 가정이다. 도메인 적응형 모방학습은 학습자가 숙련자의 시연을 직접 따를 수 없기 때문에 어렵기 때문에 모방학습의 도메인 이동을 극복하기 위해 관찰데이터에서 도메인별 정보를 제거하여 시각적 관찰에서 도메인 독립 특징 벡터를 추출하는 모델을 구성하여 문제를 해결하려고 했다. 그러나, 도메인 적응형 모방학습 방법은 도메인 차이가 크지 않은 모방학습 태스크에서 잘 작동하지만, 필요한 행동 특징을 잘 얻지 못하고 도메인 격차가 큰 어려운 모방학습 태스크에 대해 만족스럽게 수행되지 않는다.

발명의 내용

해결하려는 과제

[0005] 행동 특성을 더욱 효과적으로 추출하는 학습 모델을 제시하고, 기존 알고리즘에서 제시하는 기본적인 학습구조보다 더 큰 도메인 차이에서도 타겟 도메인에서 에이전트의 정책을 학습하는 방법 및 시스템을 제공할 수 있다.

과제의 해결 수단

[0007] 모방학습 시스템에 의해 수행되는 도메인 적응형 모방학습 방법은, 원본 이미지의 소스 도메인과 타겟 도메인 각각에 채택된 인코더로부터 이중 특징(dual feature)을 추출하는 단계; 적대적 학습을 기반으로 훈련된 이미지 생성기를 이용하여 상기 추출된 이중 특성을 결합함에 따라 이미지를 재구성하는 단계; 및 상기 재구성된 이미지의 일관성 검사를 통해 상기 재구성된 이미지가 상기 원본 이미지와 동일한 지 여부를 판단하는 단계를 포함할 수 있다.

[0008] 상기 이중 특징을 추출하는 단계는, 상기 소스 도메인과 타겟 도메인의 각각에 행동 인코더와 도메인 인코더를 포함하는 복수 개의 인코더를 채택하고, 상기 채택된 복수 개의 인코더를 통해 상기 소스 도메인과 타겟 도메인의 각각에 대한 도메인 특징 및 행동 특징을 추출하는 단계를 포함하고, 상기 행동 인코더는, 도메인 정보를 제외한 행동 정보를 포함하는 행동 특징을 추출하고, 상기 도메인 인코더는, 행동 정보를 제외한 도메인 정보를 포함하는 도메인 특징을 추출할 수 있다.

[0009] 상기 이미지를 재구성하는 단계는, 상기 타겟 도메인의 도메인 특징과 상기 소스 도메인의 행동 특징을 결합한 후 상기 타겟 도메인의 제1 이미지 생성기(G_T)를 이용하여 제1 이미지($\widehat{TB_S}$)를 생성하고, 상기 소스 도메인의 도메인 특징과 상기 타겟 도메인의 행동 특징을 결합한 후 상기 소스 도메인의 제2 이미지 생성기(G_S)를 사용하여 제2 이미지($\widehat{SB_T}$)를 생성하는 단계를 포함할 수 있다.

[0010] 상기 제1 이미지 생성기 및 제2 이미지 생성기는, 각각의 이미지 판별기를 통해 적대적 학습으로 훈련되는 것일

수 있다.

[0011] 상기 이미지를 재구성하는 단계는, 상기 이중 특징의 추출을 위해 훈련된 복수 개의 인코더를 이용하여 상기 생성된 제1 이미지 및 상기 생성된 제2 이미지의 각각에 대한 도메인 특징 및 행동 특징을 추출하는 단계를 포함할 수 있다.

[0012] 상기 이미지를 재구성하는 단계는, 제1 이미지 생성기 및 제2 이미지 생성기를 통해 상기 생성된 1 이미지 및 상기 생성된 제2 이미지의 각각으로부터 추출된 도메인 특징 및 행동 특징을 이용한 이미지 변환을 통해 재구성된 이미지를 획득하는 단계를 포함할 수 있다.

[0013] 상기 판단하는 단계는, 상기 재구성된 이미지의 일관성 검사와 상기 추출된 이중 특징의 일관성 검사를 수행하는 단계를 포함할 수 있다.

[0014] 상기 도메인 적응형 모방학습 방법은, 보상 생성을 위한 보상 생성용 판별기와 학습자 정책이 적대적 방식을 통해 공동 학습을 수행하되, 상기 보상 생성용 판별기에서 행동 인코더의 행동 특징 정보를 이용하여 학습자와 숙련자의 행동을 구별하도록 학습하고, 상기 학습자 정책에서 상기 행동 특징 정보가 숙련자처럼 보이도록 관찰데이터를 생성하는 방법을 학습할 수 있다.

[0015] 모방학습 시스템에 의해 수행되는 도메인 적응형 모방학습 방법을 실행시키기 위해 컴퓨터 판독 가능한 저장매체에 저장된 컴퓨터 프로그램에 있어서, 상기 도메인 적응형 모방학습 방법은, 원본 이미지의 소스 도메인과 타겟 도메인 각각에 채택된 인코더로부터 이중 특징(dual feature)을 추출하는 단계; 적대적 학습을 기반으로 훈련된 이미지 생성기를 이용하여 상기 추출된 이중 특성을 결합함에 따라 이미지를 재구성하는 단계; 및 상기 재구성된 이미지의 일관성 검사를 통해 상기 재구성된 이미지가 상기 원본 이미지와 동일한 지 여부를 판단하는 단계를 포함할 수 있다.

[0016] 모방학습 시스템은, 원본 이미지의 소스 도메인과 타겟 도메인 각각에 채택된 인코더로부터 이중 특징(dual feature)을 추출하는 이중 특징 추출부; 적대적 학습을 기반으로 훈련된 이미지 생성기를 이용하여 상기 추출된 이중 특성을 결합함에 따라 이미지를 재구성하는 이미지 생성부; 및 상기 재구성된 이미지의 일관성 검사를 통해 상기 재구성된 이미지가 상기 원본 이미지와 동일한 지 여부를 판단하는 일관성 검사부를 포함할 수 있다.

발명의 효과

[0018] 숙련자의 시연 데이터가 학습자와는 다른 도메인에서 생성되어 학습자가 숙련자의 시연 데이터를 그대로 모방할 수 없는 상황에서도 시연 데이터로부터 학습에 필요한 행동 정보를 적절하게 추출하여 타겟 도메인에 있는 학습자의 정책을 학습할 수 있다.

도면의 간단한 설명

[0020] 도 1은 일 실시예에 있어서, 모방학습 시스템의 구성을 설명하기 위한 블록도이다.

도 2는 일 실시예에 있어서, 도메인 적응형 모방학습 방법을 설명하기 위한 흐름도이다.

도 3은 일 실시예에 있어서, 도메인 적응형 모방학습 동작을 설명하기 위한 도면이다.

도 4는 일 실시예에 있어서, 모방학습(IL) 태스크에 대한 예제 이미지 쌍을 나타낸 예이다.

도 5는 일 실시예에 있어서, 150 Epoch 시간에 해당하는 RE2-3 태스크의 샘플링된 관찰에서 추정된 예상 보상을 나타낸 예이다.

도 6은 일 실시예에 있어서, PointUMaze 및 AntUMaze에 대한 원래 관찰과 PointUMaze-ant 태스크에 대한 관찰을 설명하기 위한 예이다.

발명을 실시하기 위한 구체적인 내용

[0021] 이하, 실시예를 첨부한 도면을 참조하여 상세히 설명한다.

[0023] 실시예에서는 타겟 도메인의 에이전트가 소스 도메인의 숙련자 시연을 관찰하여 태스크 수행을 학습하는 시각적 관찰을 통한 도메인 적응형 모방학습에 대하여 설명하기로 한다.

[0024] 도 1은 일 실시예에 있어서, 모방학습 시스템의 구성을 설명하기 위한 블록도이고, 도 2는 일 실시예에 있어서, 도메인 적응형 모방학습 방법을 설명하기 위한 흐름도이다.

[0025] 모방학습 시스템(100)의 프로세서는 이중 특징 추출부(110), 이미지 생성부(120) 및 일관성 검사부(130)를 포함할 수 있다. 이러한 프로세서의 구성요소들은 모방학습 시스템에 저장된 프로그램 코드가 제공하는 제어 명령에 따라 프로세서에 의해 수행되는 서로 다른 기능들(different functions)의 표현들일 수 있다. 프로세서 및 프로세서의 구성요소들은 도 2의 도메인 적응형 모방학습 방법이 포함하는 단계들(210 내지 230)을 수행하도록 모방학습 시스템을 제어할 수 있다. 이때, 프로세서 및 프로세서의 구성요소들은 메모리가 포함하는 운영체제의 코드와 적어도 하나의 프로그램의 코드에 따른 명령(instruction)을 실행하도록 구현될 수 있다.

[0026] 프로세서는 도메인 적응형 모방학습 방법을 위한 프로그램의 파일에 저장된 프로그램 코드를 메모리에 로딩할 수 있다. 예를 들면, 모방학습 시스템에서 프로그램이 실행되면, 프로세서는 운영체제의 제어에 따라 프로그램의 파일로부터 프로그램 코드를 메모리에 로딩하도록 모방학습 시스템을 제어할 수 있다. 이때, 프로세서는 이중 특징 추출부(110), 이미지 생성부(120) 및 일관성 검사부(130) 각각은 메모리에 로딩된 프로그램 코드 중 대응하는 부분의 명령을 실행하여 이후 단계들(210 내지 230)을 실행하기 위한 프로세서의 서로 다른 기능적 표현들일 수 있다.

[0027] 단계(210)에서 이중 특징 추출부(110)는 원본 이미지의 소스 도메인과 타겟 도메인 각각에 채택된 인코더로부터 이중 특징(dual feature)을 추출할 수 있다. 이중 특징 추출부(110)는 소스 도메인과 타겟 도메인의 각각에 행동 인코더와 도메인 인코더를 포함하는 복수 개의 인코더를 채택하고, 채택된 복수 개의 인코더를 통해 소스 도메인과 타겟 도메인의 각각에 대한 도메인 특징 및 행동 특징을 추출할 수 있다.

[0028] 단계(220)에서 이미지 생성부(120)는 적대적 학습을 기반으로 훈련된 이미지 생성기를 이용하여 추출된 이중 특성을 결합함에 따라 이미지를 재구성할 수 있다. 이미지 생성부(120)는 타겟 도메인의 도메인 특징과 소스 도메인의 행동 특징을 결합한 후 타겟 도메인의 제1 이미지 생성기(G_T)를 이용하여 제1 이미지($\widehat{TB_S}$)를 생성하고, 소스 도메인의 도메인 특징과 타겟 도메인의 행동 특징을 결합한 후 소스 도메인의 제2 이미지 생성기(G_S)를 사용하여 제2 이미지($\widehat{SB_T}$)를 생성할 수 있다. 이미지 생성부(120)는 이중 특징의 추출을 위해 훈련된 복수 개의 인코더를 이용하여 생성된 제1 이미지 및 생성된 제2 이미지의 각각에 대한 도메인 특징 및 행동 특징을 추출할 수 있다. 이미지 생성부(120)는 제1 이미지 생성기 및 제2 이미지 생성기를 통해 생성된 1 이미지 및 생성된 제2 이미지의 각각으로부터 추출된 도메인 특징 및 행동 특징을 이용한 이미지 변환을 통해 재구성된 이미지를 획득할 수 있다.

[0029] 단계(230)에서 일관성 검사부(130)는 재구성된 이미지의 일관성 검사를 통해 재구성된 이미지가 원본 이미지와 동일한 지 여부를 판단할 수 있다. 일관성 검사부(130)는 재구성된 이미지의 일관성 검사와 추출된 이중 특징의 일관성 검사를 수행할 수 있다.

[0030] 도 3은 일 실시예에 있어서, 도메인 적응형 모방학습 동작을 설명하기 위한 도면이다.

[0031] 마르코프 의사결정 프로세스(MDP)는 튜플 $(\mathcal{S}, \mathcal{A}, P, R, \gamma, \rho_0)$ 로 설명되며, 여기서 S는 상태 공간, A는 행동 공간, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^+$ 는 상태 전이 확률, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ 은 보상 함수, $\gamma \in (0,1)$은 카운트 계수, ρ_0는 초기 상태 분포이다. 마르코프 의사결정 프로세스(MDP)가 주어졌을 때, 확률적 정책 $\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^+$는 상태 $s \in \mathcal{S}$가 주어졌을 때 행동에 대한 조건부 확률이다. 강화학습의 목표는 $J(\pi) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)]$ 을 최대화하는 최적의 정책 π^*를 찾는 것이다. 여기서, $s_0 \sim \rho_0, a_t \sim \pi(\cdot|s_t), s_{t+1} \sim P(\cdot|s_t, a_t)$ 이다. 모방학습에서는 실제 R 대신 숙련자의 시연이 제공된다. 단일 도메인을 가진 모방학습의 목표는 숙련자 정책 π_E에 의해 생성된 시연을 모방하는 것이다.

[0032] 도메인 적응형 모방학습을 정의하기로 한다. 두 개의 도메인, 즉 마르코프 의사결정 프로세스(MDP)로 모델링된

소스 도메인(첨자 S 사용)과 타겟 도메인(첨자 T 사용)은 각각 $\mathcal{M}_S = (\mathcal{S}_S, \mathcal{A}_S, P_S, R_S, \gamma_S, \rho_{0,S})$ 와 $\mathcal{M}_T = (\mathcal{S}_T, \mathcal{A}_T, P_T, R_T, \gamma_T, \rho_{0,T})$ 로 표시될 수 있다. 숙련자 정책 $\pi_E : \mathcal{S}_S \times \mathcal{A}_S \rightarrow \mathbb{R}^+$ 은 소스 도메인에 있고 학습자 정책 $\pi_\theta : \mathcal{S}_T \times \mathcal{A}_T \rightarrow \mathbb{R}^+$ 은 타겟 도메인에 있다. 또한, 소스 도메인 $\pi_{SN} : \mathcal{S}_S \times \mathcal{A}_S \rightarrow \mathbb{R}^+$ 와 타겟 도메인 $\pi_{TN} : \mathcal{S}_T \times \mathcal{A}_T \rightarrow \mathbb{R}^+$ 에 하나의 비숙련자 정책이 있다고 가정하기로 한다. 모방학습 시스템은 특징 추출 및 이미지 재구성 프로세스를 지원하는 추가 데이터를 생성할 수 있다. 소스 도메인 π_{SN} 및 타겟 도메인 π_{TN} 은 여러 가지 방법으로 구성될 수 있지만, 균일하게 임의의 행동을 수행하는 정책으로 선택한다. 모방학습 시스템은 모방학습에 대해 이미지 관측치가 제공되고 제공된 이미지를 O_{SE}, O_{SN}, O_{TN}이라 가정하며, 각각 π_E, π_{SN}, π_{TN} 이 생성한 관측치 셋을 n_{demo} 크기로 나타낸다. 학습자 정책 π_θ 에 의해 생성된 관찰 셋은 O_{TL}로 표시될 수 있다. 각 관찰 o_t는 4개의 연속된 타임스텝에 해당하는 4개의 이미지 $o_t = (o_t^0, o_t^{-1}, o_t^{-2}, o_t^{-3})$ 로 구성될 수 있다. O_{SE}, O_{SN}, O_{TN}, O_{TL}를 사용하여 타겟 도메인에서 주어진 태스크를 잘 수행할 수 있도록 학습자 정책 π_θ를 훈련하는 것이 목표이다.

[0033] 도 3을 참고하면, 모방학습 시스템은 이미지 재구성 및 일관성 검사를 기반으로 획득된 특징이 정보 손실 없이 필요한 정보를 잘 포함하고 있는지 확인할 수 있는 추가 구조를 추가하여 간단한 적대적 학습을 강화할 수 있다.

[0034] 모방학습 시스템(100)은 이중 특징 추출(310)을 통해 원본 이미지(SB_S, TB_R)로부터 특징 정보를 추출할 수 있다. 모방학습 시스템(100)은 소스 도메인과 타겟 도메인의 각각에 두 개의 인코더를 채택할 수 있다. 모방학습 시스템은 소스 도메인의 도메인 인코더 DE_S와 행동 인코더 BE_S를 포함할 수 있고, 타겟 도메인의 도메인 인코더 DE_T와 행동 인코더 BE_T를 포함할 수 있다. 여기서, 소스 도메인 및 타겟 도메인의 각각의 도메인 인코더는 행동 정보를 제외한 도메인 정보만을 포함하는 도메인 특징(소스 도메인 특징/타겟 도메인 특징)을 추출하려고 하는 반면, 소스 도메인 및 타겟 도메인 각각의 행동 인코더는 도메인 정보를 제외한 행동 정보(숙련자 또는 비숙련자)만을 포함하는 행동 특징(소스 행동 특징/타겟 행동 특징)을 추출하려고 한다. 이러한 인코딩 각각은, 하나의 인코더와 두 개의 수반되는 판별기(도메인 판별기 및 행동 판별기)로 구성된 적대적 학습 블록을 기반으로 한다. 모방학습 시스템은 복수 개(예를 들면, 4개)의 인코더를 통해 소스 도메인 인코더의 특징 정보(소스 도메인 특징S와 소스 행동 특징 B_S), 타겟 도메인 인코더의 특징 정보(타겟 도메인 특징 T와 타겟 행동 특징 B_T)를 포함하는 이중 특징을 추출할 있다(1단계 특징 추출).

[0035] 모방학습 시스템은 추출된 각각의 특징 정보를 사용하여 이미지를 생성(320)할 수 있다. 모방학습 시스템은 추출된 각각의 특징 정보(특징 벡터)를 사용하여 이미지 변환(translation)을 적용할 수 있다. 모방학습 시스템(100)은 타겟 도메인 특징 T와 소스 행동 특징 B_S를 결합하고 이미지 생성기 G_T를 사용하여 새로운 이미지 $\widehat{TB_S}$를 생성하고, 소스 도메인 특징 S와 타겟 행동 특징 B_T를 결합하고 이미지 생성기 G_S를 사용하여 새로운 이미지 $\widehat{SB_T}$를 생성할 수 있다. 이때, 각 이미지 생성기(G_T 또는 G_S)는 도 3의 중간 부분에 표시된 것처럼 해당 이미지 판별기(ID_T 또는 ID_S)의 도움을 받아 적대적 학습을 기반으로 훈련될 수 있다.

[0036] 모방학습 시스템(100)은 생성된 새로운 이미지 $\widehat{TB_S}$ 및 새로운 이미지 $\widehat{SB_T}$에서 도 3의 오른쪽에 표시된 것처럼 1단계 특징 추출에 사용된 동일한 훈련된 인코더 DE_T, BE_T, DE_S, BE_S를 사용하여 특징 추출(2단계 특징 추출)을 다시 적용할 수 있다. 모방학습 시스템(100)은 2단계 특징 추출을 통해 추출된 복수 개의 특징 정보 $(\widehat{T}, \widehat{B}_S, \widehat{S}, \widehat{B}_T)$를 이용하여 도 3의 가장 오른쪽에 나타낸 것과 같이 1단계 특징 추출 후에 사용하였던 동일한 이미지 생성기 G_S 및 G_T를 활용하여 이미지 변환을 다시 적용함으로써 원본 이미지와 동일한 또는 유사한 이미지 $\widehat{SB_S}$ 및 $\widehat{TB_T}$를 재구성할 수 있다.

[0037] 모방학습 시스템(100)은 인코더가 정보 손실 없이 해당 특징을 적절하게 추출한 경우 재구성된 이미지 $\widehat{SB_S}$ 및 $\widehat{TB_T}$ 와 원본 이미지 SB_S 및 TB_T와 동일한지 일관성 검사(330)를 수행할 수 있다. 이러한 주기 일관성을 특징 추출 인코더에 대한 추가 기준 중 하나로 활용할 수 있다. 또한, 모방학습 시스템(100)은 이미지 재구성을 기반으로 하는 아키텍처를 통해 추출된 특징 정보에 대한 다른 일관성을 확인할 수 있다. 첫째, 모방학습 시스템(100)은 1단계 특징 추출에서 추출된 소스 도메인 인코더 DE_S와 타겟 행동 인코더 BE_S의 소스 도메인 특징 S와 타겟 행동 특징 B_S을 결합하고 특징 조합 (S, B_S)를 이미지 생성기 G_S에 입력함에 따라 생성된 소스-도메인 이미지 $\widehat{SB_S}$가 원본 이미지(관찰 이미지)인 원본 이미지 SB_S와 동일한지 판단할 수 있다. 이와 같이, 이미지 생성기 G_T가 있는 타겟 도메인의 특징 조합 (T, B_T)에도 동일한 일관성이 적용될 수 있다. 이러한, 일관성을 이미지 재구성 일관성이라고 한다. 둘째, 모방학습 시스템(100)은 생성된 소스-도메인 이미지 $\widehat{SB_S}$를 소스 도메인 인코더 DE_S와 소스 행동 인코더 BE_S에 입력하여 도메인 특징 $\tilde{S}$와 행동 특징 $\tilde{B}_S$를 획득하고, 획득된 도메인 특징 $\tilde{S}$와 행동 특징 $\tilde{B}_S$이 1단계 특징 추출에서 추출된 소스 도메인 특징 S 및 소스 행동 특징 B_S와 동일한 지 여부를 판단할 수 있다. 동일한 방식으로 타겟 도메인에도 적용될 수 있다. 이러한 일관성을 특징 재구성 일관성이라고 한다. 특징 재구성 일관성은 특징 인코더와 이미지 생성기를 통해 특징 간 일관성을 검사하는 반면, 이미지 재구성 일관성은 특징 인코더와 이미지 생성기를 통해 이미지 간 일관성을 검사한다.

[0038] 이에 따라, 도메인 및 행동 판별기를 사용한 기존의 단순 적대적 학습에 더하여 특징 추출을 위한 추가 지침을 통해 제안된 체계는 이전의 단순한 접근 방식에 비해 훨씬 더 나은 특징을 추출할 수 있다. 전체 구조가 학습되면 타겟 도메인(도 3의 점선 블록 301 내부의 블록)에서 행동 인코더 BE_T를 사용하여 타겟 도메인에서 학습자를 훈련시킨다. 타겟 도메인에서 향상된 행동 특징 추출을 통해 타겟 도메인에서 학습자의 태스크 성능을 향상시킬 수 있다.

[0039] 앞서 언급한, 일관성 검사 외에도, 성능을 향상시키기 위해 특징 벡터 정규화 및 특징 벡터 유사성의 추가 기술이 적용될 수 있다. 특징 벡터 정규화는 특징 벡터의 L2-norm을 유지하고 특징 벡터의 값의 폭발적 증가를 방지한다. 특징 벡터 유사성은 동일한 도메인 관찰의 도메인 특징 벡터가 유사한 값을 가져야 하고 동일한 행동 관찰의 행동 특징 벡터가 유사한 값을 갖도록 모델을 안내하는 또 다른 일관성이다. 실시예에 따르면, 이중 특징 추출 및 이미지 재구성을 기반으로 한 특징 추출을 개선할 수 있다.

[0040] 모방학습 시스템(100)은 보상 생성 및 학습자 정책 업데이트를 수행할 수 있다. 학습자 정책 π_θ를 업데이트하는 방법에 대해 설명하기로 한다. 타겟 도메인에서 행동 판별기 BD_T를 행동 인코더 BE_T와 함께 사용하여 정책 업데이트에 대한 보상을 생성할 수 있다. 그러나 보상 생성의 경우 행동 인코더 BE_T를 사용하지만 행동 판별기 BD_T는 사용하지 않는다. 행동 판별기 BD_T는 숙련자의 행동 E와 비숙련자의 행동 N을 구별하도록 훈련되었지만, 학습자 정책 π_θ가 숙련자를 모방하여 적절한 보상을 받을 수 있도록 숙련자의 행동 E와 학습자의 행동 L을 구별하는 판별기가 필요하다. 따라서, 모방학습 시스템은 보상 생성을 위해 보상 생성용 판별기 D를 사용할 수 있다. 보상 생성용 판별기 D와 학습자 정책 π_θ은 적대적인 방식으로 공동 학습을 수행할 수 있다. 보상 생성용 판별기 D는 행동 인코더 BE_T에서 행동 특징 벡터를 입력으로 가져와 입력의 행동 레이블(E 또는 L)을 예측하는 방법을 학습할 수 있다. 반면, 학습자 정책 π_θ은 해당 행동 특징 벡터가 숙련자처럼 보이도록 관찰 O_{TL}을 생성하는 방법을 학습할 수 있다. 학습자는 SAC를 사용하여 정책을 업데이트하고 관찰자에 대한 추정 보상은 다음과 같이 정의될 수 있다. 비특허문헌 1(Haarnoja, T.; Zhou, A.; Hartikainen, K.; Tucker, G.; Ha, S.; Tan, J.; Kumar, V.; Zhu, H.; Gupta, A.; Abbeel, P.; et al. 2018. Soft actor-critic algorithms and applications.)에 SAC에 대한 설명이 개시된 바 있다.

[0041] $$\hat{r}(o_t) = \log(D(BE_T(o_t))) - \log(1 - D(BE_T(o_t))), \quad (1)$$

[0042] 실시예에 따르면, 시뮬레이션 환경에서 타겟 도메인과 소스 도메인에 대해서 시각적인 효과의 차이, 로봇을 비추는 카메라의 각도의 차이, 에이전트의 구조의 차이 등의 차이를 가지는 다양한 모방학습 태스크에서 기존 알

고리즘보다 타겟 도메인에서 학습하는 정책의 성능을 향상시킬 수 있다. 또한 기존 강화학습의 희소 보상 환경에서는 정책을 학습할 수 없는 상황에서 실시예에서 제안된 방법으로 타겟 도메인에서 정책 학습에 성공할 수 있다. 이러한 실험 및 실험 결과를 설명하기 위하여, 8개의 모방학습(IL) 태스크에 대해 실시예에서 제안된 적응형 모방학습 방법이 평가될 수 있다. 각 IL 태스크는 소스 도메인과 타겟 도메인으로 구성되며, 각 도메인은 MuJoCo 시뮬레이터의 기본 RL 태스크를 포함한다. 먼저 기본 강화학습(RL) 태스크를 설명하기로 한다.

[0043] IP(inverted pendulum, 역진자): 폴이 넘어지는 것을 막기 위해 에이전트가 카트를 움직인다. 에피소드 길이는 1000이다. 폴이 떨어져도 에피소드는 중간에 종료되지 않는다. 각 에피소드는 태스크를 해결하는 데 유용한 관찰을 덜 제공하기 때문에 이는 태스크를 훨씬 더 어렵게 만든다.

[0044] IDP(역 이중 진자): 카트에 두 개의 폴이 있는 IP 태스크의 더 어려운 버전이다.

[0045] RE2(Reacher-2): 에이전트는 한쪽 끝이 고정된 두 링크 암을 이동하여 암의 이동 가능한 끝점이 목표 지점에 도달하도록 한다. 에피소드 길이는 50 이다. 목표 지점의 극좌표(r, ϕ)가 $r \in \{0.15, 0.2\}$이고 $\phi \in \{0, \Pi/4, \Pi/2, 3\Pi/4, \cdots, 7\Pi/4\}$일 수 있는 목표 지점 후보를 16점으로 좁혔다.

[0046] RE3(reacher-3): 에이전트가 3링크 암 로봇을 이동하는 RE2 태스크의 더 어려운 버전이다.

[0047] 이 네 가지 기본 태스크를 기반으로 8개의 IL 태스크를 고려한다. 소스 도메인, 타겟 도메인, IL 태스크가 표 1에 요약되어 있다.

[0048] 표 1:

[0049]

IL task	Source	Target	Domain difference
IP-to-color	IP	IP	Pole and cart colors
IDP-to-one	IDP	IP	2 poles → 1 pole
IDP-to-color	IDP	IDP	Pole and cart colors
IP-to-two	IP	IDP	1 pole → 2 poles
RE2-to-tilted	RE2	RE2	Different camera angle
RE3-to-two	RE3	RE2	3 links → 2 links
RE3-to-tilted	RE3	RE3	Different camera angle
RE2-to-three	RE2	RE3	2 links → 3 links

[0050] 도 4를 참고하면, IL 태스크에 대한 예제 이미지 쌍을 나타낸 것이다. 소스 도메인의 이미지와 타겟 도메인의 이미지를 포함하는 예제 이미지 쌍에 대한 시각적 관찰 사례를 나타낸 도면이다. IL 태스크는 각각 픽셀 색 상 교체, 카메라 각도 변경 또는 심지어 다른 MDP를 포함한 다른 유형의 도메인 차이를 제공하기 때문에 어렵다.

[0051] 실시예에 제안된 방법의 성능과 종래의 기술(TPIL과 DisentanGAIL)이 실험 설정에 설명된 8개의 IL 태스크를 이용하여 비교될 수 있다. 모든 IL 태스크에서 고정된 하이퍼 파라미터 셋과 손실 계수가 사용될 수 있다. 실시예에서 제안된 방법이 모든 IL 태스크에서 다른 태스크보다 큰 마진으로 우수한 성능을 보인다는 것을 확인할 수 있었다. 도 5를 참고하면, 150 Epoch 시간에 해당하는 RE2-3 태스크의 샘플링된 관찰에서 추정된 예상 보상의 인스턴스를 나타낸 것으로, 4회 연속 이미지가 하나의 관찰을 구성하며, 예상 보상은 네 번째 이미지에 표시된다. 도 5의 왼쪽 열은 비숙련자의 행동이고 오른쪽 열은 숙련자의 행동이다. 리치 암의 한쪽은 이미지의 중앙에 고정된다. 왼쪽 열에서는 이동 가능한 암 엔드가 4개의 시간 연속 이미지를 통해 목표 위치에서 멀어지는 반면, 이동 가능한 암 엔드는 오른쪽 열에서 목표 위치를 향해 이동한다. 이러한 경우 기준 알고리즘은 비숙련자 행동에 더 높은 보상을 할당하는 반면, 제안된 방법은 숙련자 행동에 더 높은 보상을 올바르게 할당할 수 있다.

[0052] 표 1에 있는 8개의 IL 태스크는 난이도가 다른 다양한 IL 태스크에 대해 제안된 방법을 평가하는 것으로 간주된다. 실제로 이러한 태스크의 경우 학습자 정책은 타겟 도메인에서 직접 숙련자 시연을 통해 IL 또는 타겟 도메인 환경과 상호 작용하는 RL에 의해 타겟 도메인에서 직접 학습될 수 있다. 도메인 적응형 IL의 진정한 힘은 알려진 RL 알고리즘에 의해 타겟 도메인에서 그러한 직접적인 정책 학습이 매우 어렵거나 불가능할 때 나타난다. 여기서, AntUMaze라고 알려진 사례를 고려하기로 한다. 실제 타겟 태스크는 다음과 같다. MuJoCo 시뮬레이터에 정의된 4족 보행 ant 로봇은 도 6의 상단 행 우측에 나타난 바와 같이 Ushaped maze를 통해 빨간색 목표 위치로 이동해야 한다. 현재의 최첨단 RL 알고리즘으로는 AntUMaze 문제를 직접 해결하기가 매우 어렵다는 것을 확인할 수 있다. 복잡한 로봇 조작과 경로 찾기가 동시에 수반되는 문제이기 때문이다. 어려운 타겟 태스크를 해결하기 위해, PointUMaze라는 소스 도메인에서 더 간단한 태스크를 고려하기로 한다.

PointUMaze에서, 로봇은 도 6의 상단 행(지점의 막대는 지점의 방향을 나타낸다)의 좌측에 도시된 바와 같이 지점 위치 및 지점 방향으로 정의된 지점으로 단순화되는 반면, 원래 ant 로봇의 상태는 훨씬 복잡하다. 이 point 로봇이 Ushape 미로를 통해 목표 위치에 도달하는 더 쉬운 태스크는 종래의 방법(예를 들면, SAC)를 통해 학습할 수 있다. point 로봇 소스 도메인의 숙련자 정책이 획득되면 이미지 시퀀스로 숙련자 시연을 생성하지만, 시연 이미지의 경우 도 6의 하단 행 왼쪽에 표시된 것처럼 point 로봇을 동심 원으로 마스킹한다. ant 로봇 에이전트는 타겟 도메인에서 UMaze 문제를 학습하려고 한다. 마스크 숙련 자(point 로봇) 시연에 대해 설명하기로 한다. ant 로봇은 또한 미로에서 자신의 이미지(즉, 타겟 도메인 이미지)를 촬영한다. 여기서 다시 도 6의 하단 행 오른쪽에 표시된 것처럼 ant 로봇을 동심 원으로 마스킹할 수 있다. 이 마스킹은 ant 로봇의 역학이 아닌 관찰에 대한 수정이다. 따라서, IL을 사용하여 ant 로봇은 자체 역학과 두 집합의 마스킹된 이미지로 UMaze 문제를 해결할 수 있다. 실시예에서 제안된 방법은 숙련자의 지식을 point 로봇에서 ant 로봇으로, 더 쉬운 태스크에서 더 어려운 태스크로 전달하고, 타겟 도메인에서 에이전트가 어려운 태스크를 성공적으로 수행하도록 훈련시킬 수 있다.

[0053] 이상에서 설명된 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0054] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치에 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

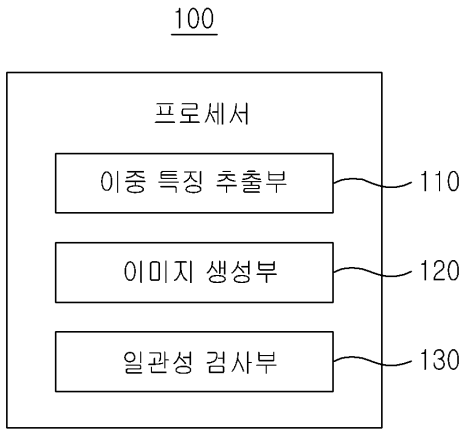
[0055] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

[0056] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

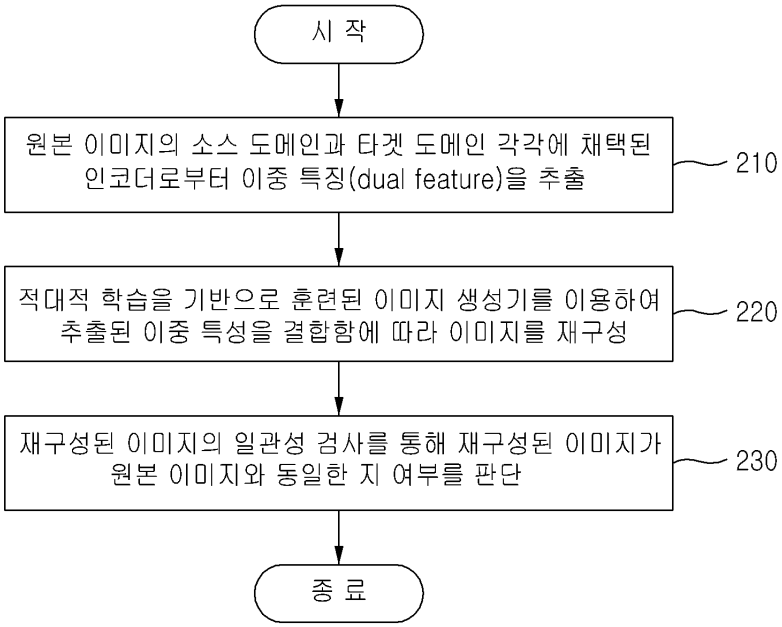
[0057] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

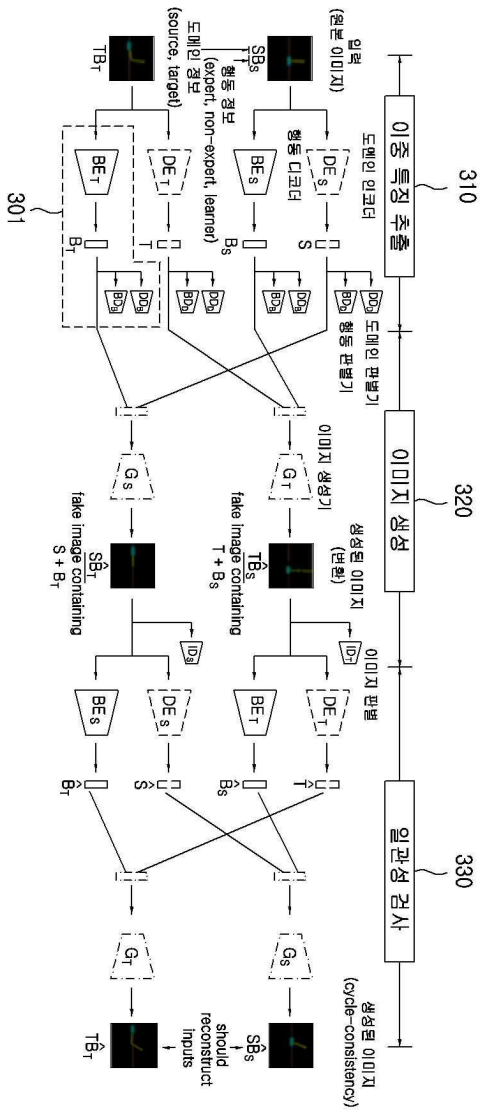
도면1


100
프로세서
이중 특징 추출부
110
이미지 생성부
120
일관성 검사부
130

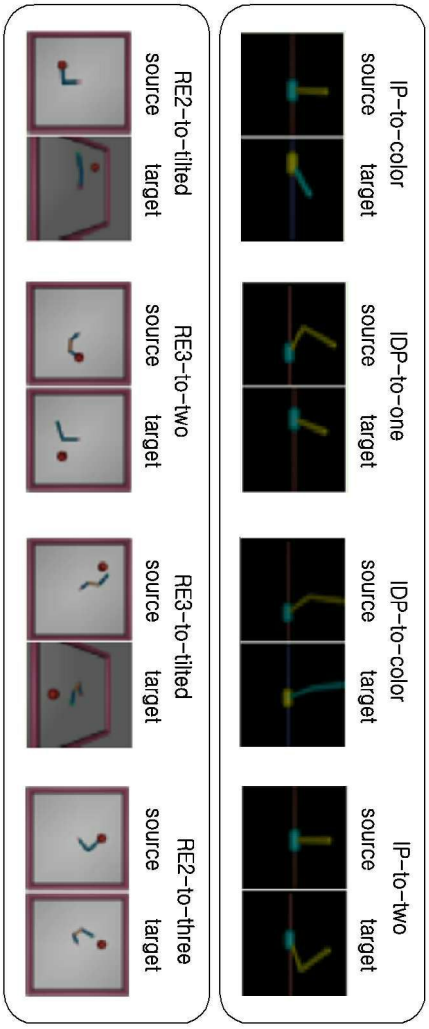
도면2


시 작
원본 이미지의 소스 도메인과 타겟 도메인 각각에 채택된 인코더로부터 이중 특징(dual feature)을 추출
210
적대적 학습을 기반으로 훈련된 이미지 생성기를 이용하여 추출된 이중 특성을 결합함에 따라 이미지를 재구성
220
재구성된 이미지의 일관성 검사를 통해 재구성된 이미지가 원본 이미지와 동일한 지 여부를 판단
230
종 료

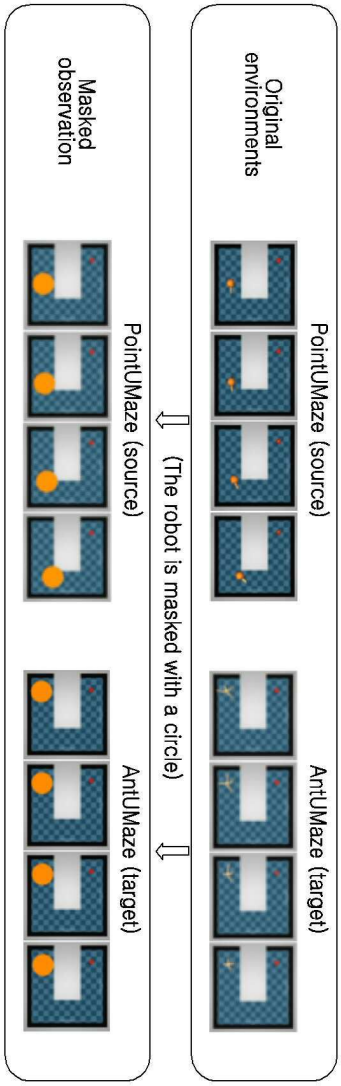
도면3


310
이중 특징 추출
320
이미지 생성
330
일관성 검사
301
입력 (현실 이미지)
도메인 정보 (source, target)
행동 정보 (expert, non-expert, learner)
도메인 인코더
행동 디코더
도메인 판별기
행동 판별기
이미지 생성기
생성된 이미지 (변환)
이미지 판별
생성된 이미지 (cycle-consistency)
fake image containing S + B_T
fake image containing T + B_S
should reconstruct inputs

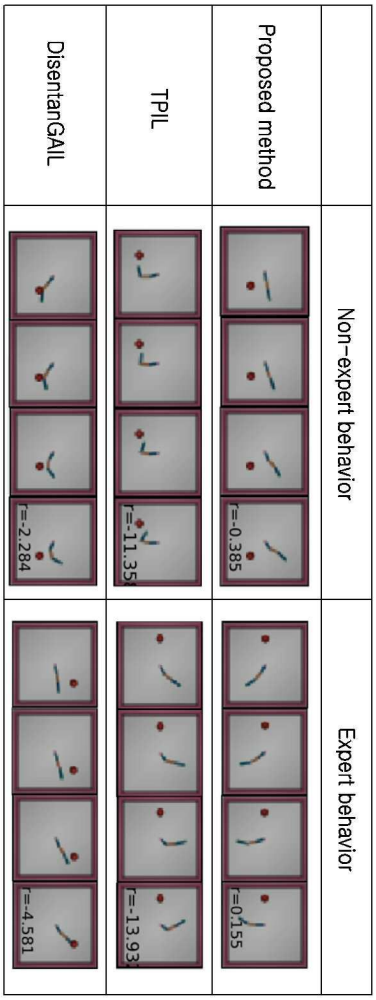
도면4


IP-to-color
source
target
IDP-to-one
source
target
IDP-to-color
source
target
IP-to-two
source
target
RE2-to-tilted
source
target
RE3-to-two
source
target
RE3-to-tilted
source
target
RE2-to-three
source
target

도면5


Original environments
PointUMaze (source)
AntUMaze (target)
(The robot is masked with a circle)
Masked observation
PointUMaze (source)
AntUMaze (target)

도면6


Proposed method
TPIL
DisentanGAIL
Non-expert behavior
Expert behavior
r=-0.385
r=0.155
r=-2.284
r=-4.581