(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0133077
(43) 공개일자 2023년09월19일

(51) 국제특허분류(Int. Cl.)
G05D 1/02 (2020.01) *G05D 1/00* (2006.01)
G06N 20/00 (2019.01) *G06V 40/20* (2022.01)
(52) CPC특허분류
G05D 1/0221 (2013.01)
G05D 1/0005 (2013.01)
(21) 출원번호 10-2022-0030165
(22) 출원일자 2022년03월10일
심사청구일자 2022년03월10일

(71) 출원인
한국기술교육대학교 산학협력단
충청남도 천안시 동남구 병천면 충절로 1600 (한국기술교육대학교내)
(72) 발명자
한연희
대전광역시 유성구 가정로 306-6, 주공타운하우스 11동 104호
최호빈
충청남도 천안시 동남구 병천면 충절로 1636, 명품에덴빌 206호
(뒷면에 계속)
(74) 대리인
김견수

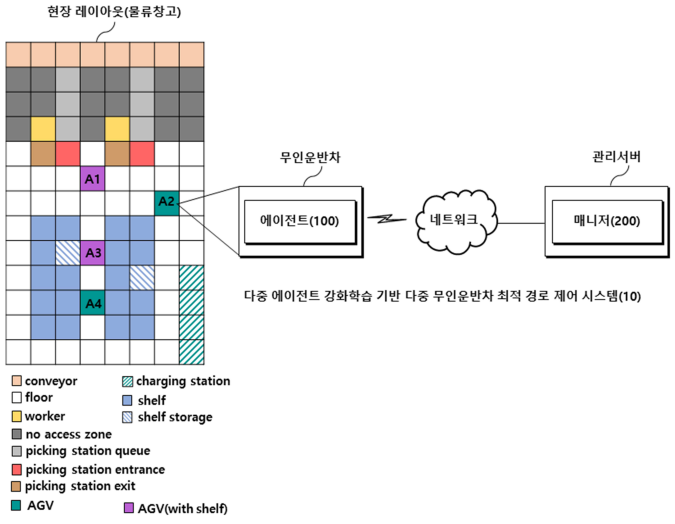
전체 청구항 수 : 총 12 항

(54) 발명의 명칭 **다중 에이전트 강화학습 기반 다중 무인운반차 최적 경로 제어 시스템**

(57) 요 약

본 발명은 다중 에이전트 강화학습 기반 다중 무인운반차 최적 경로 제어 시스템에 관한 것으로, 복수의 무인운반차 각각을 제어하기 위한 에이전트에서, 사전에 설정한 시나리오에 따라 목적지까지 이동하기 위한 무인운반차의 행동을 추출하기 위한 강화학습을 수행할 때, 개별보상을 늘려 누적 전역보상을 최대화하고 손실(loss)을 최소화하는 방향으로 상기 강화학습을 수행함으로써 복수의 무인운반차의 충돌이나 고립을 방지하고 목적지까지 최적의 경로로 이동할 수 있도록 하는 시스템에 관한 것이다.

대 표 도 - 도1



(52) CPC특허분류
G06N 20/00 (2021.08)
G06V 40/20 (2022.01)

(72) 발명자
김주봉
충청북도 청주시 흥덕구 원터로 44, 402호
임현교
대구광역시 중구 태평로 65, 35동 96호(태평동, 삼부아파트3단지)
지창훈
충청북도 청주시 흥덕구 1순환로577번길 36, 401호

이 발명을 지원한 국가연구개발사업

과제고유번호	1345315470
과제번호	2018R1A6A1A03025526
부처명	교육부
과제관리(전문)기관명	한국연구재단
연구사업명	이공분야 대학중점 연구소 지원사업
연구과제명	상호작용 가상현실 기반 몰입형 교육 훈련 플랫폼
기 여 율	1/1
과제수행기관명	한국기술교육대학교 산학협력단
연구기간	2018.06.01 ~ 2027.02.28

명 세 서

청구범위

청구항 1

무인운반차에 대한 현재의 관측데이터를 에이전트 강화학습모델에 입력하여 행동가치가 높은 행동을 추출하는 에이전트; 및

각 상기 에이전트별로 상기 추출한 행동에 따른 개별보상을 제공하는 매니저;를 포함하며,

상기 에이전트 강화학습모델에 상기 개별보상과 상기 행동가치를 이용하여 계산한 로컬손실을 적용하여 강화학습하는 것을 특징으로 하는 다중 에이전트 강화학습 기반 다중 무인운반차 최적 경로 제어 시스템.

청구항 2

청구항 1에 있어서,

상기 에이전트는,

상기 현재의 관측데이터와 이전에 추출한 행동을 상기 에이전트 강화학습모델에 입력하여 상기 행동가치가 높은 행동을 추출하는 행동 추출부;

상기 현재의 관측데이터를 포함하는 과거의 관측데이터 및 상기 추출한 행동을 포함하는 상기 행동가치를 상기 매니저로 제공하는 행동가치 제공부; 및

상기 매니저로부터 수신한 상기 개별보상에 따라 상기 행동가치를 조정하여 상기 로컬손실을 계산하는 로컬손실 계산부;를 더 포함하며,

상기 로컬손실은, 상기 개별보상이 높을수록 상기 행동가치를 높게 조정함으로써 최소화되며, 상기 에이전트 강화학습모델에서 역전파되어 상기 현재의 관측데이터에 따라 상기 행동가치가 높은 행동을 추출하도록 상기 강화학습에 이용되는 것을 특징으로 하는 다중 에이전트 강화학습 기반 다중 무인운반차 최적 경로 제어 시스템.

청구항 3

청구항 1에 있어서,

상기 관측데이터는, 상기 무인운반차가 적용된 현장을 복수의 셀로 구성한 레이아웃에서, 각 상기 무인운반차를 중심으로 사전에 설정한 범위에 대한 제1 채널 이미지 및 제2 채널 이미지를 포함하여 구성되며,

상기 제1 채널 이미지는, 상기 범위 내의 각 상기 셀에 대한 상기 무인운반차의 이동 가능여부를 포함하며,

상기 제2 채널 이미지는, 상기 범위 내의 각 상기 셀에 대한 상기 무인운반차의 목적지까지 남은 거리정보를 포함하는 것을 특징으로 하는 다중 에이전트 강화학습 기반 다중 무인운반차 최적 경로 제어 시스템.

청구항 4

청구항 3에 있어서,

상기 매니저는,

각 상기 에이전트로부터 제공받은 각 행동가치와 상기 레이아웃에서 각 셀에 대한 에이전트의 현재 누적 방문횟수를 포함하는 전역 상태데이터를 매니저 강화학습모델에 입력하여 모든 에이전트의 행동에 따른 공동행동가치를 추출하는 공동행동가치 추출부; 및

상기 개별보상을 모두 합산한 전역보상 및 상기 공동행동가치에 따라 전역손실을 계산하여 각 상기 에이전트로 제공하는 전역손실 계산부;를 더 포함하며,

상기 에이전트는, 상기 로컬손실과 상기 전역손실을 합산하여 최종 손실을 계산하고, 상기 에이전트 강화학습모델에 상기 최종 손실을 역전파하여 상기 강화학습을 수행하는 것을 더 포함하는 것을 특징으로 하는 다중 에이

전트 강화학습 기반 다중 무인운반차 최적 경로 제어 시스템.

청구항 5

청구항 1에 있어서,

상기 에이전트는,

사전에 부여된 식별자에 따라 순차적으로 상기 행동을 추출하며,

상기 행동을 순차적으로 추출하기 위해, 상기 식별자 순으로 상기 추출한 행동에 따른 상기 무인운반차의 제어 결과를 현장에 대한 레이아웃에 마스킹하여 마스킹 데이터를 생성하고, 상기 매니저를 통해 상기 마스킹 데이터를 다음 순서의 식별자를 가지는 에이전트로 순차적으로 공유하도록 함으로써 무인운반차 상호간에 충돌이 발생하지 않도록 하는 행동을 추출하도록 하는 것을 더 포함하는 것을 특징으로 하는 다중 에이전트 강화학습 기반 다중 무인운반차 최적 경로 제어 시스템.

청구항 6

청구항 1에 있어서,

상기 에이전트는,

상기 강화학습을 수행한 결과에 따라 상기 무인운반차에 대한 실제 관측데이터를 상기 에이전트 강화학습모델에 입력하여 행동가치가 높은 행동을 추출하여 다른 에이전트와 독립적으로 상기 무인운반차의 목적지까지 최적 경로로 제어하는 무인운반차 제어부;를 더 포함하는 것을 특징으로 하는 다중 에이전트 강화학습 기반 다중 무인운반차 최적 경로 제어 시스템.

청구항 7

무인운반차의 에이전트에서, 무인운반차에 대한 현재의 관측데이터를 에이전트 강화학습모델에 입력하여 행동가치가 높은 행동을 추출하는 단계; 및

매니저에서, 각 상기 에이전트별로 상기 추출한 행동에 따른 개별보상을 제공하는 단계;를 포함하며,

상기 에이전트 강화학습모델에 상기 개별보상과 상기 행동가치를 이용하여 계산한 로컬손실을 적용하여 강화학습하는 것을 특징으로 하는 다중 에이전트 강화학습 기반 다중 무인운반차 최적 경로 제어 방법.

청구항 8

청구항 1에 있어서,

상기 방법은,

상기 에이전트에서, 상기 현재의 관측데이터와 이전에 추출한 행동을 상기 에이전트 강화학습모델에 입력하여 상기 행동가치가 높은 행동을 추출하는 행동 추출 단계;

상기 에이전트에서, 상기 현재의 관측데이터를 포함하는 과거의 관측데이터 및 상기 추출한 행동을 포함하는 상기 행동가치를 상기 매니저로 제공하는 행동가치 제공 단계; 및

상기 에이전트에서, 상기 매니저로부터 수신한 상기 개별보상에 따라 상기 행동가치를 조정하여 상기 로컬손실을 계산하는 로컬손실 계산 단계;를 더 포함하며,

상기 로컬손실은, 상기 개별보상이 높을수록 상기 행동가치를 높게 조정함으로써 최소화되며, 상기 에이전트 강화학습모델에서 역전파되어 상기 현재의 관측데이터에 따라 상기 행동가치가 높은 행동을 추출하도록 상기 강화학습에 이용되는 것을 특징으로 하는 다중 에이전트 강화학습 기반 다중 무인운반차 최적 경로 제어 방법.

청구항 9

청구항 7에 있어서,

상기 관측데이터는, 상기 무인운반차가 적용된 현장을 복수의 셀로 구성한 레이아웃에서, 각 상기 무인운반차를 중심으로 사전에 설정한 범위에 대한 제1 채널 이미지 및 제2 채널 이미지를 포함하여 구성되며,

상기 제1 채널 이미지는, 상기 범위 내의 각 상기 셀에 대한 상기 무인운반차의 이동 가능여부를 포함하며,

상기 제2 채널 이미지는, 상기 범위 내의 각 상기 셀에 대한 상기 무인운반차의 목적지까지 남은 거리정보를 포함하는 것을 특징으로 하는 다중 에이전트 강화학습 기반 다중 무인운반차 최적 경로 제어 방법.

청구항 10

청구항 9에 있어서,

상기 방법은,

상기 매니저에서, 각 상기 에이전트로부터 제공받은 각 행동가치와 상기 레이아웃의 각 셀에 대한 에이전트의 현재 누적 방문횟수를 포함하는 전역 상태데이터를 매니저 강화학습모델에 입력하여 모든 에이전트의 행동에 따른 공동행동가치를 추출하는 공동행동가치 추출 단계; 및

상기 매니저에서, 상기 개별보상을 모두 합산한 전역보상 및 상기 공동행동가치에 따라 전역손실을 계산하여 각 상기 에이전트로 제공하는 전역손실 계산 단계;를 더 포함하며,

상기 에이전트는, 상기 로컬손실과 상기 전역손실을 합산하여 최종 손실을 계산하고, 상기 에이전트 강화학습모델에 상기 최종 손실을 역전파하여 상기 강화학습을 수행하는 것을 더 포함하는 것을 특징으로 하는 다중 에이전트 강화학습 기반 다중 무인운반차 최적 경로 제어 방법.

청구항 11

청구항 7에 있어서,

상기 에이전트는,

사전에 부여된 식별자에 따라 순차적으로 상기 행동을 추출하며,

상기 행동을 순차적으로 추출하기 위해, 상기 식별자 순으로 상기 추출한 행동에 따른 상기 무인운반차의 제어결과를 현장에 대한 레이아웃에 마스킹하여 마스킹 데이터를 생성하고, 상기 매니저를 통해 상기 마스킹 데이터를 다음 순서의 식별자를 가지는 에이전트로 순차적으로 공유하도록 함으로써 무인운반차 상호간에 충돌이 발생하지 않도록 하는 행동을 추출하도록 하는 것을 더 포함하는 것을 특징으로 하는 다중 에이전트 강화학습 기반 다중 무인운반차 최적 경로 제어 방법.

청구항 12

청구항 7에 있어서,

상기 방법은,

상기 에이전트에서, 상기 강화학습을 수행한 결과에 따라 상기 무인운반차에 대한 실제 관측데이터를 상기 에이전트 강화학습모델에 입력하여 행동가치가 높은 행동을 추출하여 다른 에이전트와 독립적으로 상기 무인운반차의 목적지까지 최적 경로로 제어하는 무인운반차 제어 단계;를 더 포함하는 것을 특징으로 하는 다중 에이전트 강화학습 기반 다중 무인운반차 최적 경로 제어 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 다중 에이전트 강화학습 기반 다중 무인운반차 최적 경로 제어 시스템에 관한 것으로, 더욱 상세하게는 복수의 무인운반차 각각을 제어하기 위한 에이전트에서, 사전에 설정한 시나리오에 따라 목적지까지 이동하기 위한 무인운반차의 행동을 추출하기 위한 강화학습을 수행할 때, 개별보상을 늘려 누적 전역보상을 최대화하고 손실(loss)을 최소화하는 방향으로 상기 강화학습을 수행함으로써 복수의 무인운반차의 충돌이나 고립을 방지하고 목적지까지 최적의 경로로 이동할 수 있도록 하는 시스템에 관한 것이다.

배 경 기 술

[0002] 최근 산업기술의 급격한 발전으로 전자상거래가 급속하게 발전하고 있다. 전자상거래의 발전은 물류시장에 많은 변화를 가져왔으며 많은 기업들이 고객의 요구에 부응하기 위한 주문처리서비스를 도입하고 있다.

[0003] 전자상거래와 같은 온라인 유통산업에서 주문처리서비스는 고객의 주문에 따라 물류창고에서 제품을 피킹(picking), 포장(packing) 및 배송하기 위한 일련의 프로세스를 말한다.

[0004] 물류창고 내에서는 수많은 물품을 실시간으로 이송해야 하므로 자동으로 목적지까지 물품을 이송하기 위한 복수의 무인운반차가 필수적으로 활용되고 있다.

[0005] 무인운반차는 물류창고의 운영에 필수 구성요소로 자리 잡고 있으며, 효율적인 재고 관리 및 생산성 향상을 위해서 자동으로 제어하는 것이 필요하다.

[0006] 이를 위해 다중 에이전트 강화학습(multi agent reinforcement learning)을 통해 각 에이전트에서 무인운반차를 충돌 없이 최적경로를 따라 이동하도록 자동으로 제어하고자 하는 방법이 연구개발 되고 있다.

[0007] 종래의 다중 에이전트 강화학습은 에이전트 수(즉, 무인운반차의 수)가 증가함에 따라 행동공간(action space)이 기하급수적으로 증가하는 치명적인 문제점이 있고, 비정상성(non stationarity)이 더욱 심화되는 문제점을 내포하고 있다.

[0008] 이러한 문제점을 해결하기 위해 최근에는 단일 에이전트 학습에 이용되는 Q-학습(Q-learning)을 확장한 IQL(independent Q-learning) 알고리즘과 VDN(value deposition networks) 알고리즘이 다중 에이전트 강화학습에 적용되고 있다.

[0009] 그러나 IQL(Independent Q-Learning) 알고리즘은 복수의 에이전트 각각이 Q-학습을 통해 자신이 부분적으로 관측한 상황을 기반으로 자신의 행동을 선택할 수 있는 확률이 높은 Q값(행동가치)을 다른 에이전트에 대한 정보 없이 추출하도록 독립적으로 강화학습하는 방식이다.

[0010] 이러한 IQL 알고리즘은 다른 에이전트들(즉, 무인운반차)을 환경의 일부로 간주하기 때문에 다른 무인운반차의 독립적인 이동에 따른 모든 에이전트의 공동행동(joint action)에 대한 협력을 반영하지 못해 안정적인 강화학습이 어려우며, 무인운반차간의 충돌이나 측정 위치에서의 유휴상태(고립)가 발생할 수 있는 문제점이 있다.

[0011] 또한 VDN(Value Decomposition Networks) 알고리즘은 모든 에이전트의 Q값을 더하여 계산된 전체 Q값을 통해 다중 에이전트의 공동행동가치(전체 Q값)를 추정하여, 공동행동가치에 따라 계산한 손실(loss)을 이용하여 각 에이전트에 대한 강화학습을 수행하도록 하는 것이다.

[0012] 이러한 VDN 알고리즘은 각 에이전트들에 대해서 독립적인 행동가치로 추상화하도록 하는 것이 가능지만, 단순 가산성 제약으로 인한 문제가 발생될 수 있으며, 전체 공동행동가치에 따른 손실을 이용하여 개별 에이전트에서 강화학습을 수행하는 구조이기 때문에 개별 에이전트의 강화학습에 대한 손실을 반영하지 못해 실제 정확한 제어가 어려운 문제점이 있다.

[0013] 이러한 문제점을 해결하기 위해 최근에는 Qmix 알고리즘을 이용하여 각 에이전트에 대한 강화학습을 수행하는 방법에 대한 연구개발이 진행되고 있다.

[0014] Qmix 알고리즘은 VDN 알고리즘의 단순 합산 방식이 아닌 MLP(multi layer perceptron)를 포함하는 혼합 네트워크(mixing network)를 사용하여 복잡한 비선형 방법으로 전체 공동행동가치를 추정함으로써 VDN의 가산성 제약을 완화시킨다.

[0015] 또한 Qmix 알고리즘은 혼합 네트워크에서 추가적인 상태정보를 이용할 수 있기 때문에 더 풍부한 공동 행동에 대한 가치를 추정할 수 있는 장점이 있다.

[0016] 그러나 Qmix 알고리즘 또한 VDN 알고리즘과 같이 전체 공동행동에 따른 손실을 계산하여 개별 에이전트에 대한 강화학습을 수행하도록 하는 구조이기 때문에 개별 에이전트의 강화학습에 대한 손실을 반영하지 못해 실제 정확한 제어가 어려운 문제점이 있다.

[0017] 이에 따라 본 발명에서는 복수의 에이전트(즉, 복수의 무인운반차)간의 협업이 필요한 작업(물품 이송)을 해결하기 위한 다중 에이전트 환경을 고려하여 각 에이전트의 충돌이나 고립 없이 목적지까지 최적의 경로로 정확하게 제어할 수 있도록 하는 강화학습 방법을 제안하고자 한다.

[0018] 또한 본 발명은 각 에이전트에서 자체 관찰한 관측데이터를 에이전트 인공지능 강화학습모델에 입력하여 행동가치(Q값)가 높은 행동(action)을 추출하고, 각 에이전트에서 추출한 행동에 대한 행동가치에 따라 모든 에이전트에 대한 공동행동가치를 추정하고, 상기 각 행동에 따른 전역보상과 개별보상에 따라 상기 추정한 공동행동가치를 이용하여 전역소실과 로컬손실을 계산하여 상기 각 손실을 이용하여 각 에이전트의 인공지능 강화학습모델을

강화학습함으로써 실제 무인운반차를 제어할 때 자신의 관측데이터만으로도 다른 무인운반차와의 충돌 없이 목적지까지 최적의 루트로 이동하도록 제어할 수 있도록 하는 방안을 제안하고자 한다.

[0019] 또한 본 발명은 각 에이전트에서 강화학습을 수행할 때, 개별보상을 늘려 전역보상을 최대화함과 동시에 로컬손실을 최소화하는 방향으로 강화학습을 수행함으로써 정확하게 무인운반차를 제어할 수 있도록 하는 방안을 제안하고자 한다.

[0020] 다음으로 본 발명의 기술분야에 존재하는 선행기술에 대하여 간단하게 설명하고, 이어서 본 발명이 상기 선행기술에 비해서 차별적으로 이루고자 하는 기술적 사항에 대해서 기술하고자 한다.

[0021] 먼저 한국공개특허 제2019-0076628호(2019.07.02.)는 보상 제어기를 이용한 강화학습 방법 및 이를 위한 장치에 관한 것으로, 주어진 환경에서, 현재의 상태를 인식하여 선택 가능한 행동 중 보상과 관련된 값을 최대화하고, 각 행동에 대한 보상을 제공하여 강화학습을 수행하도록 하는 보상 제어기를 이용한 강화학습 방법 및 이를 위한 장치에 관한 것이다. 즉, 한국공개특허 제2019-0076628호는 하나의 에이전트에 대한 강화학습을 수행하는 정도만 기재되어 있다.

[0022] 그러나 본 발명은 복수의 에이전트가 협업하는 환경에서 복수의 에이전트에 대한 에이전트 강화학습모델을 강화학습하여 복수의 무인운반차를 최적 경로로 제어하기 위한 것으로, 한국공개특허 제2019-0076628호와 본 발명은 그 목적에 있어서 현저한 차이점이 있다.

[0023] 더욱이 한국공개특허 제2019-0076628호는 단순 보상을 최대화하는 방향으로 강화학습하는 것일 뿐, 본 발명에서 제안하고 있는 전역손실과 로컬손실을 이용하여 강화학습하는 기술적 특징에 대해서는 전혀 기재하고 있지 않으며 시사 혹은 그 어떠한 암시도 없다.

[0024] 또한 한국공개특허 제2020-0062887호(2020.06.04.)는 강화학습에 기반하여 시스템의 제어 동작에 대한 품질을 보장하기 위한 장치 및 방법에 관한 것으로, 에이전트에서, 자신의 상태정보에 기초하여 Q 네트워크를 통해 특정 행동을 계산한 후, 환경 시스템으로 제공하여 상기 행동에 따른 리워드를 수신하고, 상기 리워드에 기초하여 상기 Q 네트워크를 업데이트(강화학습)하는 것을 특징으로 한다. 즉, 한국공개특허 제2020-0062887호는 단순히 행동에 따른 리워드에 따라 강화학습을 수행하는 것이다.

[0025] 반면에 본 발명은 전역손실과 로컬손실을 이용하여 각 에이전트에 대한 에이전트 강화학습모델을 강화학습하여 각 에이전트에서 해당하는 무인운반차를 최적의 경로로 목적지에 도착할 수 있도록 제어하는 것으로, 한국공개특허 제2020-0062887호는 이러한 본 발명의 기술적 특징을 전혀 기재하고 있지 않으며 시사 혹은 그 어떠한 암시도 없다.

발명의 내용

해결하려는 과제

[0026] 본 발명은 상기와 같은 문제점을 해결하기 위해 창안된 것으로서, 복수의 무인운반차에 대한 각 에이전트에서 각 무인운반차의 관측데이터를 토대로 행동을 각각 추출하기 위한 에이전트 강화학습모델을 통해 복수의 무인운반차간의 충돌, 고립 또는 이들의 조합을 방지하여 최적 경로로 목적지까지 제어할 수 있도록 하는 다중 에이전트 강화학습 기반 다중 무인운반차 최적 경로 제어 시스템을 제공하는 것을 그 목적으로 한다.

[0027] 또한 본 발명은 사전에 설정한 시나리오에 따라 각 에이전트에서 무인운반차의 관측데이터를 에이전트 강화학습모델에 입력하여 과거 관측데이터에 따른 행동에 대한 행동가치가 높은 행동을 각각 추출하고, 상기 추출한 행동에 대한 행동가치를 관리서버로 제공하여, 상기 관리서버에서 상기 각 행동가치와 전역 상태데이터를 매니저 강화학습모델에 입력하여 전체 에이전트의 행동에 따른 공동행동가치를 추출하는 시스템을 제공하는 것을 그 목적으로 한다.

[0028] 또한 본 발명은 관리서버에서 개별 에이전트의 각 행동에 따라 개별 에이전트에 대한 개별보상을 산출하여 각 에이전트로 제공하는 시스템을 제공하는 것을 그 목적으로 한다.

[0029] 또한 본 발명은 관리서버에서 개별보상을 합산한 전역보상과 공동행동가치를 추출한 결과에 따라 전역손실을 계산하여 상기 공동행동가치를 추출하기 위한 매니저 강화학습모델을 강화학습하는 시스템을 제공하는 것을 그 목적으로 한다.

[0030] 또한 본 발명은 각 에이전트에서 개별보상에 따른 로컬손실을 각각 계산하여 에이전트 강화학습모델을 업데이트

하여 강화학습을 수행하는 시스템을 제공하는 것을 그 목적으로 한다.

[0031] 또한 본 발명은 각 에이전트에서, 로컬손실을 계산할 때, 특정 타임스탭에서 관측한 관측데이터에 따라 추출한 행동에 대한 행동가치를 개별보상에 따라 조정하여 로컬손실이 최소화되도록 하는 시스템을 제공하는 것을 그 목적으로 한다.

[0032] 또한 본 발명은 개별보상이 높을수록 행동가치를 높게 조정하여 로컬손실을 최소화하고, 로컬손실을 에이전트 강화학습모델에 역전파시켜 강화학습하여 특정 관측데이터에서 개별보상이 높으면서 동시에 행동가치가 높은 특정 행동을 추출할 수 있도록 하는 시스템을 제공하는 것을 그 목적으로 한다.

[0033] 또한 본 발명은 강화학습을 통해 실제 무인운반차가 실제 환경에 적용되었을 때, 다른 무인운반차와의 협력 없이 각 에이전트에서 관측한 관측데이터만으로도 최적의 행동을 추출하여 최적의 경로로 무인운반차를 제어할 수 있도록 하는 시스템을 제공하는 것을 그 목적으로 한다.

과제의 해결 수단

[0034] 본 발명의 일 실시예에 따른 다중 에이전트 강화학습 기반 다중 무인운반차 최적 경로 제어 시스템은, 복수의 무인운반차에 대한 관측데이터를 각각의 복수의 에이전트 강화학습모델에 입력하여 행동가치가 높은 행동을 추출하는 복수의 에이전트, 및 상기 복수의 에이전트에 각각에 대해서 상기 추출한 행동에 따른 개별보상을 제공하는 매니저를 포함하며, 각 상기 에이전트 강화학습모델에 상기 개별보상과 상기 행동가치를 이용하여 계산한 로컬손실을 적용하여 강화학습하는 것을 포함하는 것을 특징으로 한다.

[0035] 또한 상기 복수의 에이전트 각각은, 상기 관측데이터와 이전에 추출한 행동을 각 상기 에이전트 강화학습모델에 입력하여 상기 행동가치가 높은 행동을 추출하는 행동 추출부, 상기 관측데이터를 포함하는 과거의 관측데이터 및 상기 추출한 행동을 포함하는 상기 행동가치를 상기 매니저로 제공하는 행동가치 제공부 및 상기 매니저로부터 수신한 상기 개별보상에 따라 상기 행동가치를 조정하여 상기 로컬손실을 계산하는 로컬손실 계산부를 더 포함하며, 상기 로컬손실은, 상기 개별보상이 높을수록 상기 행동가치를 높게 조정함으로써 최소화되며, 각 상기 에이전트 강화학습모델에서 역전파되어 상기 관측데이터에 따라 상기 행동가치가 높은 행동을 추출하도록 상기 강화학습에 이용되는 것을 특징으로 한다.

[0036] 또한 상기 관측데이터는, 상기 복수의 무인운반차가 각각 적용된 현장을 복수의 셀로 구성한 레이아웃에서, 각 상기 복수의 무인운반차를 중심으로 사전에 설정한 범위에 대한 제1 채널 이미지 및 제2 채널 이미지를 포함하여 구성되며, 상기 제1 채널 이미지는, 상기 범위 내의 각 상기 셀에 대한 각 상기 복수의 무인운반차의 이동가능 여부정보를 포함하며, 상기 제2 채널 이미지는, 상기 범위 내의 각 상기 셀에 대한 각 상기 복수의 무인운반차의 목적지까지 남은 거리정보를 포함하는 것을 특징으로 한다.

[0037] 또한 상기 매니저는, 상기 복수의 에이전트 각각으로부터 제공받은 각 행동가치와 상기 레이아웃에서 각 셀에 대한 상기 복수의 에이전트 각각의 누적 방문횟수를 포함하는 전역 상태데이터를 매니저 강화학습모델에 입력하여 상기 복수의 에이전트 모두의 행동에 따른 공동행동가치를 추출하는 공동행동가치 추출부 및 상기 복수의 에이전트 각각의 상기 개별보상을 모두 합산한 전역보상 및 상기 공동행동가치에 따라 전역손실을 계산하여 상기 복수의 에이전트로 제공하는 전역손실 계산부를 더 포함하며, 상기 복수의 에이전트 각각은, 상기 로컬손실과 상기 전역손실을 합산하여 최종 손실을 계산하고, 상기 복수의 에이전트 강화학습모델 각각에 상기 최종 손실을 역전파하여 상기 강화학습을 수행하는 것을 더 포함하는 것을 특징으로 한다.

[0038] 또한 상기 복수의 에이전트 각각은, 사전에 부여된 식별자에 따라 순차적으로 상기 행동을 추출하며, 상기 행동을 순차적으로 추출하기 위해, 상기 식별자 순으로 상기 추출한 행동에 따른 상기 복수의 무인운반차 각각의 제어결과를 현장에 대한 레이아웃에 마스킹한 마스킹 데이터를 생성하고, 상기 매니저를 통해 상기 마스킹 데이터를 다음 순서의 식별자를 가지는 각 상기 복수의 에이전트로 순차적으로 공유하도록 함으로써 상기 복수의 무인운반차 상호간에 충돌이 발생하지 않는 행동을 추출하도록 하는 것을 더 포함하는 것을 특징으로 한다.

[0039] 또한 상기 복수의 에이전트 각각은, 상기 강화학습을 수행한 결과에 따라 상기 복수의 무인운반차 각각에 대한 관측데이터를 상기 복수의 에이전트 강화학습모델 각각에 입력하여 행동가치가 높은 행동을 추출하여 다른 에이전트와 독립적으로 각 상기 복수의 무인운반차의 목적지까지 최적 경로로 제어하는 무인운반차 제어부를 더 포함하는 것을 특징으로 한다.

[0040] 아울러 본 발명의 일 실시예에 따른 다중 에이전트 강화학습 기반 다중 무인운반차 최적 경로 제어 방법은, 복수의 무인운반차 각각에 대한 각 복수의 에이전트에서, 상기 복수의 무인운반차 각각에 대한 관측데이터를 복수

의 에이전트 강화학습모델에 각각 입력하여 행동가치가 높은 행동을 추출하는 단계 및 매니저에서, 상기 복수의 에이전트에 각각에 대해서 상기 추출한 행동에 따른 개별보상을 제공하는 단계를 포함하는 것을 특징으로 한다.

[0041] 또한 상기 방법은, 상기 복수의 에이전트 각각에서, 상기 관측데이터와 이전에 추출한 행동을 각 상기 복수의 에이전트 강화학습모델에 입력하여 상기 행동가치가 높은 행동을 추출하는 행동 추출 단계, 상기 복수의 에이전트 각각에서, 상기 관측데이터를 포함하는 과거의 관측데이터 및 상기 추출한 행동을 포함하는 상기 행동가치를 상기 매니저로 제공하는 행동가치 제공 단계 및 상기 복수의 에이전트 각각에서, 상기 매니저로부터 수신한 상기 개별보상에 따라 상기 행동가치를 조정하여 상기 로컬손실을 계산하는 로컬손실 계산 단계를 더 포함하는 것을 특징으로 한다.

[0042] 또한 상기 방법은, 상기 매니저에서, 상기 복수의 에이전트 각각으로부터 제공받은 각 행동가치와 상기 레이아웃에서 각 셀에 대한 상기 복수의 에이전트 각각의 누적 방문횟수를 포함하는 전역 상태데이터를 매니저 강화학습모델에 입력하여 상기 복수의 에이전트 모두의 행동에 따른 공동행동가치를 추출하는 공동행동가치 추출 단계 및 상기 매니저에서, 상기 복수의 에이전트 각각의 상기 개별보상을 모두 합산한 전역보상 및 상기 공동행동가치에 따라 전역손실을 계산하여 상기 복수의 에이전트 각각으로 제공하는 전역손실 계산 단계를 더 포함하는 것을 특징으로 한다.

[0043] 또한 상기 방법은, 상기 복수의 에이전트 각각에서, 상기 강화학습을 수행한 결과에 따라 상기 복수의 무인운반차 각각에 대한 관측데이터를 상기 복수의 에이전트 강화학습모델 각각에 입력하여 행동가치가 높은 행동을 추출하여 다른 에이전트와 독립적으로 각 무인운반차의 목적지까지 최적 경로로 제어하는 무인운반차 제어 단계를 더 포함하는 것을 특징으로 한다.

발명의 효과

[0044] 이상에서와 같이 본 발명의 다중 에이전트 강화학습 기반 다중 무인운반차 최적 경로 제어 시스템은 복수의 무인운반차가 적용된 환경에서, 무인운반차를 제어하기 위한 강화학습을 수행할 때, 서로 다른 무인운반차에 대한 제어결과를 반영하여 타임스탭별로 개별보상을 늘려 결과적으로 누적 전역보상을 최대화하고, 상기 개별보상에 따른 손실을 최소화하는 방향으로 상기 강화학습을 수행하여 상기 강화학습모델을 점진적으로 고도화함으로써 상기 무인운반차의 충돌, 고립 또는 이들이 조합이 발생하지 않고 최적의 경로로 목적지까지 이동할 수 있도록 하는 효과가 있다.

도면의 간단한 설명

[0045] 도 1은 본 발명의 일 실시예에 따른 다중 에이전트 강화학습 기반 다중 무인운반차 최적 경로 제어 시스템을 설명하기 위해 나타낸 개념도이다.

도 2는 본 발명의 시나리오에 따른 에이전트의 상태의 천이를 나타낸 도면이다.

도 3은 본 발명의 일 실시예에 따른 관측데이터의 제1 채널 이미지를 설명하기 위해 나타낸 도면이다.

도 4는 본 발명의 일 실시예에 따른 관측데이터의 제2 채널 이미지를 설명하기 위해 나타낸 도면이다.

도 5는 본 발명의 일 실시예에 따른 마스킹 데이터를 생성하는 과정을 설명하기 위해 나타낸 도면이다.

도 6은 본 발명의 일 실시예에 따른 주문처리서비스를 제공하기 위한 무인운반차 경로 제어 작업을 수행하는 과정을 나타낸 도면이다.

도 7은 본 발명의 일 실시예에 따른 무인운반차에 대한 에이전트의 구성을 나타낸 블록도이다.

도 8은 본 발명의 일 실시예에 따른 매니저의 구성을 나타낸 블록도이다.

도 9는 본 발명의 일 실시예에 따른 다중 에이전트 강화학습 기반 다중 무인운반차를 최적의 경로로 제어하는 절차를 나타낸 흐름도이다.

도 10은 본 발명의 일 실시예에 따른 스몰 레이아웃과 라지 레이아웃을 도시한 도면이다.

도 11은 본 발명의 일 실시예에 따른 종래 기술과의 성능비교 결과를 나타낸 도면이다.

도 12는 본 발명의 또 다른 일 실시예에 따른 종래 기술과의 성능비교 결과를 나타낸 도면이다.

발명을 실시하기 위한 구체적인 내용

[0046] 이하, 첨부한 도면을 참조하여 본 발명의 다중 에이전트 강화학습 기반 다중 무인운반차 최적 경로 제어 시스템에 대한 바람직한 실시예를 상세히 설명한다. 각 도면에 제시된 동일한 참조부호는 동일한 부재를 나타낸다. 또한 본 발명의 실시예들에 대해서 특정한 구조적 내지 기능적 설명들은 단지 본 발명에 따른 실시예를 설명하기 위한 목적으로 예시된 것으로, 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는 것이 바람직하다.

[0047] 도 1은 본 발명의 일 실시예에 따른 다중 에이전트 강화학습 기반 다중 무인운반차 최적 경로 제어 시스템을 설명하기 위해 나타낸 개념도이다.

[0048] 도 1에 도시한 바와 같이, 본 발명의 일 실시예에 따른 다중 에이전트 강화학습 기반 다중 무인운반차 최적 경로 제어 시스템(10)(이하, 제어 시스템이라 칭함)은 복수의 무인운반차에 각각 구비되는 에이전트(100) 및 관리 서버에 구비되는 매니저(200)를 포함하여 구성된다.

[0049] 즉, 에이전트(100)는 무인운반차량에 구비되는 것으로, 이하에서는 무인운반차량과 에이전트(100)를 동일하게 취급하여 설명하도록 하며, 필요에 따라 구분하여 설명하도록 한다.

[0050] 복수의 무인운반차는 물류창고 등의 주문처리센터와 같은 다양한 환경에 적용되어 복수의 물품을 정해진 목적지까지 운반하는 자동화된 차량을 의미한다.

[0051] 물류창고는 물품을 다른 스테이션으로 이동하기 위한 컨베이어(conveyor), 컨베이어로 물품을 이동하기 위한 피킹스테이션 큐(picking station queue), 상기 피킹스테이션의 입구(entrance) 및 출구(exit), 무인운반차량이 이송해온 물품을 피킹스테이션 큐에 올려놓기 위한 작업자(worker), 플로어(floor), 벽 등과 같이 무인운반차가 이동하지 못하는 이동불가 영역(no access zone), 무인운반차량의 충전을 위한 충전 스테이션(charging station), 물품이 올려진 선반(shelf), 빈 선반을 보관하기 위한 선반 보관소(shelf storage), 복수의 무인운반차(AGV) 등과 같이 다양한 객체를 포함하여 구성된다.

[0052] 또한 에이전트(100)는 현장의 객체를 포함하는 현장에 대해 생성한 레이아웃을 미리 저장하고 있다.

[0053] 에이전트(100)는 타임스탭별로 행동가치가 높은 무인운반차의 행동을 추출하기 위한 에이전트 강화학습모델을 구비하고 있으며, 매니저(200)는 각 에이전트에서 추출한 행동에 따른 전역보상과 개별보상을 생성하여 개별보상을 각 에이전트(100)로 제공하며, 모든 에이전트(100)의 행동에 따른 공동행동가치를 추출하기 위한 매니저 강화학습모델을 구비하고 있다.

[0054] 제어 시스템(10)은 각 무인운반차량을 목적지로 최적 경로로 제어하기 위한 에이전트 강화학습모델에 대한 강화학습을 수행한다.

[0055] 이를 위해 에이전트(100)는 사전에 설정한 시나리오에 따라 타임스탭별로 자신을 중심으로 사전에 설정한 범위에 대한 관측데이터를 생성한다.

[0056] 각 강화학습모델은 과거의 특정 데이터에서 특정 결과에 대한 행동가치(Q값)를 강화학습하는 네트워크로 구성된다.

[0057] 또한 에이전트(100)는 현재의 관측데이터를 에이전트 강화학습모델에 입력하여, 상기 에이전트 강화학습모델을 통해 상기 현재의 관측데이터와 동일한 과거의 관측데이터와 과거의 관측데이터에서 추출한 행동 중 행동가치(Q값)가 높은 행동을 추출한다. 즉, 에이전트(100)는 행동가치 중 확률이 높은 행동가치를 선택함으로써 현재의 관측데이터를 토대로 다음의 행동을 추출하는 것이다.

[0058] 한편, 관측데이터는 현장에 대한 복수의 셀로 구성한 레이아웃을 토대로 각 상기 무인운반차(즉, 에이전트)를 중심으로 사전에 설정한 범위에 대한 제1 채널 이미지 및 제2 채널 이미지를 포함하여 구성된다.

[0059] 또한 관측데이터는 다른 에이전트(100)로부터 행동에 따라 제어한 결과를 현장의 레이아웃에 마스킹한 마스킹 데이터에 따라 생성되는 것으로, 도 5를 참조하여 상세히 설명하도록 한다.

[0060] 제1 채널 이미지는 상기 범위 내의 각 상기 실에 대해서, 상기 무인운반차의 이동 가능여부를 포함하며, 제2 채

널 이미지는 상기 범위 내에서 각 상기 셀에 대한 상기 무인운반차의 목적지까지 남은 거리정보(맨해튼 거리)를 포함하여 구성된다.

[0061] 또한 에이전트는 현재 관측데이터를 포함하는 과거 관측데이터, 추출한 행동을 포함하는 행동가치를 관리서버의 매니저(200)로 전송한다.

[0062] 한편 관측데이터는 도 3 및 도 4를 참조하여 설명하도록 한다.

[0063] 강화학습을 수행하는 과정에서, 각 에이전트(100)에서 행동을 추출하는 것은, 사전에 각 에이전트(100)에 부여된 식별자(ID) 순으로 순차적으로 수행된다.

[0064] 예를 들어, 5개의 무인운반차가 환경에 적용되어 있고, 각 무인운반차에 대한 에이전트(100)에 #1 내지 #5의 식별자가 부여되어 있다고 가정한다면, #1의 에이전트(100)부터 #5의 에이전트(100) 순으로 순차적으로 행동을 추출한다.

[0065] 이때, 행동은 에이전트(100)가 바라보는 방향을 기준으로, 정지(stop), 앞쪽으로 이동(move forward), 오른쪽으로 이동(move right), 왼쪽으로 이동(move left) 및 뒤로 이동(move back)을 포함하여 총 5개로 구성된다.

[0066] 또한 행동을 순차적으로 수행하는 것은 각 에이전트(100)에서 추출한 행동에 의해 무인운반차의 충돌을 방지하기 위한 것으로, 각 에이전트(100)의 행동 추출 순서에 따라 순차적으로 공유되는 마스킹 데이터에 따라 관측데이터를 생성하여 행동을 추출한다.

[0067] 이를 위해 각 에이전트(100)는 현장의 레이아웃에 행동에 따른 제어결과(이동결과)를 마스킹한 마스킹 데이터를 매니저(200)를 통해 다음의 에이전트(100)로 순차 공유하도록 한다.

[0068] 마지막 에이전트(100)는 이전 에이전트(100)에서 마스킹한 결과를 포함하여 자신이 마스킹한 마스킹 데이터를 매니저(200)를 통해 최초 에이전트(100)로 공유하여 각 에이전트(100)(즉, 무인운반차)의 제어결과에 따라 관측데이터를 타임스탭별로 생성할 수 있도록 한다.

[0069] 즉, 각 에이전트(100)는 마스킹 데이터를 통해 상호 협력하여 충돌이 발생되지 않도록 행동을 추출할 수 있도록 하는 것이다.

[0070] 또한 매니저(200)는 각 에이전트(100)에서 수신한 행동가치의 행동에 따른 개별보상(individual reward)을 산출하여 각 에이전트(100)로 제공한다.

[0071] 이때, 매니저(200)는 각 에이전트(100)가 행동을 수행한 후, 목적지까지 남은 거리(맨해튼 거리)가 줄어들거나, 무인운반차량이 목적지에 도착하면 양의 보상(+1)을 부여하고, 그 외에는 음의 보상(-1)을 부여한 후, 각 보상을 총 에이전트의 수로 나눔으로써 개별보상을 산출한다.

[0072] 한편, 에이전트(100)가 다른 에이전트(100)와의 충돌을 회피하기 위해 정지 행동을 수행하는 경우, 음의 보상에 해당하는 개별보상을 받는 것이 가혹하다고 생각할 수 있지만 이러한 상황을 방지하기 위해 해당 관측데이터에 따라 정지라는 행동을 추출하였을 때의 행동가치(Q값)가 업데이트 될 것이다.

[0073] 또한 매니저(200)는 개별보상을 합산하여 전역보상을 산출한다.

[0074] 또한 각 에이전트(100)로 제공되는 개별보상의 절대값은 (보상)/ 총 에이전트의 수이므로 전역보상은 -1 내지 +1의 범위를 가진다.

[0075] 또한 매니저(200)는 각 에이전트(100)로부터 수신한 행동가치 및 매니저 강화학습모델과 각 에이전트(100)의 행동을 수행한 결과에 따른 전체 레이아웃에서 각 셀에 대한 누적 방문수를 포함하는 전역 상태데이터를 매니저 강화학습모델에 입력하여 공동행동가치를 추출하고, 상기 공동행동가치에 따라 전역손실을 계산하여 각 에이전트(100)로 제공할 수 있다.

[0076] 에이전트(100)는 매니저(200)로부터 수신한 개별보상에 따라 로컬손실을 계산하여, 상기 계산한 로컬손실을 에이전트 강화학습모델에 역전파시켜 업데이트함으로써 강화학습을 수행한다.

[0077] 한편, 에이전트(100)는 전역손실이 수신되면, 로컬손실과 전역손실을 합산한 최종 손실을 계산하고, 상기 계산한 최종 손실을 에이전트 강화학습모델에 역전파시켜 업데이트함으로써 강화학습을 수행한다.

[0078] 한편, 강화학습을 수행하는 과정과 각 손실을 계산하는 방법은 도 5를 참조하여 상세히 설명하도록 한다.

[0079] 또한 강화학습은 사전에 설정한 복수의 시나리오를 복수회 반복하여 수행되며, 에이전트(100)는 강화학습을 수

행한 결과에 따라 무인운반차를 목적지까지 최적 경로로 제어하는 기능을 수행한다.

[0080] 이때, 에이전트(100)는 실제 무인운반차를 제어하는 경우, 다른 에이전트(100)와의 마스킹 데이터 공유와 같은 별도의 협력가정을 수행하지 않는다.

[0081] 즉, 에이전트(100)는 자신이 관측한 관측데이터만을 강화학습을 수행한 에이전트 강화학습모델에 입력하여 현재의 관측데이터에 따라 행동가치가 높은 행동을 추출하여 다른 에이전트(100)와 독립적으로 해당 무인운반차량을 제어하게 된다.

[0082] 도 2는 본 발명의 시나리오에 따른 에이전트의 상태를 나타낸 도면이다.

[0083] 도 2에 도시한 바와 같이, 본 발명의 일 실시예에 따른 강화학습은 사전에 설정한 복수의 시나리오를 복수회 수행함으로써 수행됨은 상술한 바와 같다.

[0084] 시나리오는 (1)강화학습에 대한 동작이 수행되면 모든 무인운반차는 충전소에서 물품에 대한 선반의 할당을 대기, (2)각 피킹스테이션의 작업자는 필요한 선반을 임의의 무인운반차에 할당, (3)선반이 할당된 무인운반차는 해당 선반이 위치하는 장소로 이동하여 해당 선반을 적재(들어올림), (4)선반을 적재한 무인운반차는 피킹스테이션 입구로 이동, (5)피킹스테이션에서 작업이 완료된 무인운반차량은 임의의 선반 보관소로 이동, (6) 선반 보관소에 이동한 무인운반차량은 선반을 내려놓고 (2)부터 반복하는 총 6개의 단계로 구성될 수 있다.

[0085] 즉, 도 2에 나타낸 것과 같이 무인운반차(즉, 에이전트)는 충전소에서의 대기, 선반으로 이동, 피킹스테이션으로의 이동 및 선반 보관소로의 이동을 포함하여 총 4가지의 상태 중 하나 일 수 있다.

[0086] 예를 들어 도 1에 도시한 것과 같이 에이전트 A2와 A4는 선반을 적재하지 않은 에이전트(100)이고, A1은 물품을 포함하는 선반을 적재한 에이전트(100)이며, A4는 물품을 포함하지 않은 에이전트(100)라고 가정하면, 에이전트(100) A2와 A4는 작업자가 할당한 선반의 위치로 이동하는 상태에 있을 것이다.

[0087] 또한 에이전트(100) A1은 선반을 할당한 작업자가 위치하는 피킹스테이션으로 이동하는 상태에 있을 것이며, A4는 물품을 피킹스테이션에 내려놓았으므로 빈 선반을 내려놓기 위한 선반 보관소로 이동하는 상태에 있을 것이다.

[0088] 도 3은 본 발명의 일 실시예에 따른 관측데이터의 제1 채널 이미지를 설명하기 위해 나타낸 도면이다.

[0089] 한편 강화학습을 수행하는 도중에서 각 에이전트(100)에서 생성하는 관측데이터는 다른 에이전트(100)로부터 공유받은 마스킹 데이터를 기반으로 생성된다.

[0090] 마스킹 데이터는 도 5를 참조하여 보다 상세히 설명하도록 한다.

[0091] 도 3에 도시한 바와 같이, 본 발명의 일 실시예에 따른 에이전트(100)는 마스킹 데이터를 이용하여 해당 에이전트(100)를 중심으로 사전에 설정한 사전에 설정한 범위(예: 5 x 5 혹은 9 x 9) 내의 관측데이터를 생성한다.

[0092] 관측데이터는 에이전트(100)를 중심으로 하는 사전에 설정함 범위에 대한 2개의 채널 이미지로 정규화된 2차원 정보로 구성된다.

[0093] 관측데이터의 제1 채널 이미지는 에이전트(100)의 현재 위치를 포함하여 범위내의 셀에 대한 이동 가능 여부를 나타내는 2차원 정보를 포함하여 구성된다.

[0094] 에이전트(100)는 제1 채널 이미지를 생성할 때, 이동 가능한 셀은 1.0으로 표시하며, 이동 불가능한 셀은 0.0으로 표시하고, 에이전트(100) 자신은 0.5로 표시한다.

[0095] 이동 가능한 셀은 플로어(floor)와 에이전트(100)의 목적지(Tn 중 어느 하나)를 의미하며, 이동 불가능한 셀은 이동불가 영역(no access zone), 다른 에이전트(100), 선반을 적재한 상태에서의 다른 선반이 위치하는 셀, 상기 목적지가 아닌 셀을 의미한다. 목적지는 피킹스테이션, 선반 보관소, 선반이 위치하는 장소 중 어느 하나이다.

[0096] 예를 들어, 선반을 적재하지 않고 목적지가 특정 선반이 위치하는 장소(T2)라고 가정한다면, 다른 에이전트(100)가 위치하지 않는 셀(플로어), 목적지(T2)에 이동 가능함을 나타내는 상수 1.0을 표시하고, 그 외에는 이동 불가능함을 나타내는 상수 0.0을 표시한다.

[0097] 이때, 도 3에 도시한 제1 채널 데이터의 경우, 목적지(T2)가 관측되지 않으므로 목적지(T2)에 대해서는 나타나 있지 않다.

[0098] 이러한 제1 채널 이미지는 각 에이전트(100)에 따라 상이하게 생성됨은 당연하다.

[0099] 도 4는 본 발명의 일 실시예에 따른 관측데이터의 제2 채널 이미지를 설명하기 위해 나타낸 도면이다.

[0100] 도 4에 도시한 바와 같이, 본 발명의 일 실시예에 따른 관측데이터의 제2 채널 이미지는 에이전트(100)의 현재 위치를 포함하여 각 셀에 대해서, 상기 에이전트(100)의 목적지(target)까지 남은 거리정보(맨해튼 거리)를 나타내는 2차원 정보를 포함하여 구성된다.

[0101] 에이전트(100)는 사전에 설정한 범위 내의 각 셀과 목적지(target)까지 남은 거리정보를 계산하여 각 셀별로 계산한 거리정보를 표시함으로써 제2 채널 이미지를 생성한다.

[0102] 이때, 제2 채널 이미지는 표시한 거리정보를 정규화하여 각 거리정보가 0 내지 1의 사이 값으로 가지도록 생성될 수 있다.

[0103] 한편, 도 4에 나타낸 것과 같이 에이전트(100) A2에서 생성한 관측데이터의 제2 채널 이미지는 해당 범위 내에 목적지(T2)가 나타나 있지 않다.

[0104] 도 5는 본 발명의 일 실시예에 따른 마스킹 데이터를 생성하는 과정을 설명하기 위해 나타낸 도면이다.

[0105] 도 5에 도시한 바와 같이, 본 발명의 일 실시예에 따른 에이전트(100)는 생성한 관측데이터를 에이전트 강화학습모델에 입력하여 다음으로 수행할 행동을 추출할 때, 에이전트(100)에 부여한 식별자 순으로 수행됨은 상술한 바와 같다.

[0106] 관측데이터는 공유한 마스킹 데이터에 따라 생성된다.

[0107] 즉, 관측데이터의 제1 채널 이미지는 마스킹 데이터에 따른 이동 가능한 셀과 이동 불가능한 셀에 대한 상수를 표시함으로써 생성된다.

[0108] 도 4를 참조하여 마스킹 데이터를 설명할 때, 3개의 에이전트를 통해 순차적으로 공유하는 것을 설명하도록 한다.

[0109] 먼저 행동을 추출할 에이전트(100)(A1)는 이전에 공유받은 마스킹 데이터(즉, 이전 타임스탭에서 마지막으로 행동을 추출한 에이전트로부터 공유받은 마스킹 데이터)에 따라 자신을 중심으로 관측데이터를 생성한다.

[0110] 이때, 모든 에이전트(100)(A1 내지 A3)의 방향이 도 5에서 위쪽 방향으로 가정하면 에이전트(100)(A1)에서 생성한 관측데이터의 제1 채널 이미지에는 좌측의 경우 이동 불가 영역(no access zone)이 존재하므로 좌측으로 이동했을 때 해당하는 셀에는 이동 불가를 나타내는 상수 0이 표시될 것이고, 선반을 적재하지 않은 에이전트(100)(A1)는 전진(앞쪽으로 이동), 후진(뒤로 이동), 정지 및 우측 이동의 경우에는 이동 가능하므로 해당하는 셀에는 이동 가능을 나타내는 상수 1이 표시될 것이다.

[0111] 또한 에이전트(100)(A1)는 관측데이터를 에이전트 강화학습모델에 입력하여 행동가치가 높은 행동을 추출한다. 이때, 추출한 행동이 전진이면 이전에 공유받은 마스킹 데이터에 무인운반차를 제어한 결과를 마스킹함으로써 자신의 마스킹 데이터를 생성한다.

[0112] 상기 마스킹 데이터를 관리서버(200)의 매니저를 통해 다음의 에이전트(100)(A2)로 공유된다.

[0113] 또한 선반을 적재한 에이전트(100)(A2)는 에이전트(100)(A1)에서 공유한 마스킹 데이터를 토대로 자신을 중심으로 하는 관측데이터를 생성한다.

[0114] 이때, 에이전트(100)(A2)에서 생성한 관측데이터의 제1 채널 이미지에서 에이전트(100)(A2)의 좌측에는 이미 에이전트(100)(A1)가 이동되어 있고, 이미 선반을 적재하고 있으므로 좌측으로 이동했을 때 해당하는 셀과 후진했을 때 해당하는 셀은 이동 불가는 나타내는 상수 0이 표시되어 있을 것이고, 나머지 정지, 전진, 우측으로 이동했을 때 해당하는 셀에는 이동 가능을 나타내는 상수 1이 표시되어 있을 것이다.

[0115] 또한 에이전트(100)(A2)는 관측데이터를 에이전트 강화학습모델에 입력하여 행동가치가 높은 행동을 추출한다. 이때, 추출한 행동이 우측으로 이동이면, 에이전트(100)(A1)로부터 공유받은 마스킹 데이터(즉, 전체 레이아웃에 제어한 결과를 표시한 데이터)에 무인운반차를 제어한 결과를 추가 마스킹하여 자신의 마스킹 데이터를 생성하여 매니저(200)를 통해 다음의 에이전트(100)(A3)로 공유하도록 한다.

[0116] 선반을 적재하지 않은 에이전트(100)(A3)는 에이전트(100)(A2)에서 공유한 마스킹 데이터를 토대로 자신을 중심으로 하는 관측데이터를 생성한다.

[0117] 이때, 에이전트(100)(A3)에서 생성한 관측데이터의 제1 채널 이미지에서, 에이전트(100)(A3)의 좌측에는 이미 에이전트(100)(A2)가 이동되어 있고, 우측에 또 다른 에이전트(100) 혹은 이동 불가 영역 등의 존재로 인해 이동이 불가하면 우측 및 좌측으로 이동했을 때 해당하는 셀에는 이동 불가를 나타내는 상수 0이 표시되어 있을 것이고, 정지, 전진 및 후진했을 때 해당하는 셀에는 이동 가능을 나타내는 상수 1이 표시되어 있을 것이다.

[0118] 또한 에이전트(100)(A3)는 관측데이터를 에이전트 강화학습모델에 입력하여 행동가치가 높은 행동을 추출한다. 이때, 행동이 후진이면 에이전트(100)(A2)로부터 공유받은 마스킹 데이터에 무인운반차를 제어한 결과를 추가 마스킹하여 마스킹 데이터를 생성하고, 매니저(200)를 통해 다음의 에이전트(100)로 공유하도록 한다.

[0119] 이러한 공유는 마지막 순서에 해당하는 에이전트(100)까지 순차적으로 수행되며, 마지막 에이전트(100)에서 생성한 마스킹 데이터는 제일 처음에 해당하는 에이전트(100)(A1)로 공유되어, 다음의 타임스탭에서 상기 마스킹 데이터에 따라 관측데이터를 생성할 수 있도록 한다.

[0120] 이렇게 마스킹 데이터를 순차적으로 공유하는 것은 각 에이전트(100)에서 추출한 행동에 따라 무인운반차를 제어하였을 때 무인운반차간의 충돌을 회피하기 위해 상호 협력하는 것을 의미한다.

[0121] 도 6은 본 발명의 일 실시예에 따른 강화학습을 수행하는 과정을 나타낸 도면이다.

[0122] 도 6에 도시한 바와 같이, 본 발명의 일 실시예에 따른 주문처리서비스를 제공하기 위한 무인 운반 경로 제어 작업은 <A, S, O, Z, U, P, Rg, R1, γ>로 표시되는 분산형 부분 관찰 가능 마르코프 결정 프로세스(Dec-POMPD, decentralized partially observable Markov decision process)로 공식화될 수 있다.

[0123] 여기서 A는 에이전트(100)의 집합을 의미한다. 즉, 각 에이전트(100)는 a∈ {1,...,n}≡A로 정의된다.

[0124] S는 전역 상태 공간을 의미하며, 전역 상태데이터 s는 st ∈ S 관계에 있다.

[0125] 또한 Z는 각 에이전트의 관측공간을 나타내며, O는 관측함수로써 O: S x A -> Z로 정의된다.

[0126] U는 각 에이전트의 행동공간(action space)을 나타내며 타임스탭별로 추출되는 행동(u_t^a)은 u_t^a ∈ U의 관계에 있다.

[0127] P는 상태천이함수를 타나내며 P(st+1|st, ut):S x U x S ->[0,1]로 정의된다. 결과적으로 상태천이함수에 따라 천이되며, 상기 천이는 전역보상(rt = Rg(st, ut))과 개별보상(r_t^a = R1(st, u_t^a))으로 완료된다.

[0128] Rg는 전역보상을 산출하기 위한 전역보상 함수를 의미하며 R1은 개별보상 함수를 나타내며 Rg는 Rg: S x U -> R로 정의되며, R1은 R1: S x U -> R로 정의된다.

[0129] γ는 감가팩터(discount factor)를 의미하는 것으로 γ ∈ (0.0, 1.0]의 범위를 가진다.

[0130] 또한 개별 에이전트(100)는 타임스탭에 따라 관측데이터(o_t^n)를 생성하여 에이전트 강화학습모델에 관측데이터와 이전의 타임스탭에서 추출한 행동(u_{t-1}^n)를 입력하여 현재의 관측데이터와 현재의 관측데이터를 포함하여 해당 관측데이터와 동일한 과거의 관측데이터에 따른 행동가치가 높은 행동을 추출한다. 여기서 t는 타임스탭을 의미하며 n은 각 에이전트(100)를 나타낸다(예: 에이전트의 식별자).

[0131] 이때, 에이전트 강화학습모델은 5가지 행동에 대한 행동가치를 출력하며, 상기 행동을 추출하는 것은 행동가치가 높은 특정 행동을 선택함으로써 수행된다.

[0132] 이전에 추출한 행동을 관측데이터와 함께 에이전트 강화학습모델에 입력하는 것은 상기 이전에 추출한 행동에 따른 무인운반차량의 방향을 인식하기 위함이다.

[0133] 예를 들어, 특정 위치에서 무인운반차량을 우측 방향으로 제어한 경우, 무인운반차량은 우측방향으로 바라보고 있으며, 이러한 방향은 이전에 추출한 행동에 따라 알 수 있다.

[0134] 또한 에이전트 강화학습모델은 에이전트 네트워크로 구성되며, 상기 에이전트 네트워크는 2차원의 관측데이터를 1차원 특징으로 변환하기 위한 CNN(convolutional neural network)과 CNN의 출력과 이전에 추출한 행동을 입력으로 하여 현재의 관측데이터와 동일한 상태의 과거의 관측데이터에 따른 행동가치별 각 행동을 출력하는 멀티레이어 퍼셉트론(MLP, multi layer perceptron)이 포함된 순환 신경망(GRU, Gated Recurrent Units)을 포함하여 구성된다.

[0135] 또한 에이전트(100)는 조건(condition)으로써, 확률적 정책(stochastic policy) πa(ua|τa): T x U -> [0, 1]을 사용하여 고유한 관측데이터 및 행동 기록 τa ∈ T ≡ (Z x U)*를 생성한다.

[0136] 여기서 (Z x U)*는 가능한 모든 관찰데이터 및 행동 이력세트를 의미한다.

[0137] 각 에이전트의 정책 πa로 구성된 공동정책 π는 다음의 [수학식 1]과 같이 공식화되는 공동행동 가치 함수를 가실 수 있다.

[0138] [수학식 1]

[0139] $$Q^{\pi}(s_t, \boldsymbol{u}_t) = \mathbb{E}_{s_{t+1:\infty}, \boldsymbol{u}_{t+1:\infty}}[\sum_{i=0}^{\infty} \gamma^i \boldsymbol{r}_{t+i} | s_t, \boldsymbol{u}_t]$$

[0140] 또한 매니저(200)는 각 에이전트별 행동가치에 포함된 행동에 따른 개별보상을 산출하여 각 에이전트(100)로 제공하며, 개별보상을 모두 합산한 전역보상을 산출한다. 개별보상을 산출하는 것은 도 1을 참조하여 설명하였으므로 여기서는 생략한다.

[0141] 매니저(200)는 매니저 강화학습모델에 각 에이전트(100)로부터 수신한 행동가치(Q값)와 각 에이전트(100)의 누적 방문수를 포함하는 전역 상태데이터(s)를 입력하여 모든 에이전트(100)의 전체 관측데이터의 이력(τ)과 행동(u)에 대한 공동행동가치를 추출한다.

[0142] 매니저 강화학습모델은 단순 합산 방식이 아닌 복잡한 비선형 방법으로 전체 공동행동가치를 추정하여 출력하도록 구성된다.

[0143] 또한 매니저 강화학습모델은 각 에이전트(100)의 행동가치를 혼합하여 공동행동가치를 추정하는 것으로 혼합 네트워크를 포함하여 구성되며, 상기 혼합 네트워크는 각 에이전트(100)의 행동가치를 입력으로 각 에이전트(100)의 행동에 따른 공동행동가치를 출력하는 멀티 레이어 퍼셉트론으로 구성된다.

[0144] 또한 매니저(200)는 추출한 공동행동가치와 목표 사이의 평균 제곱 오차를 사용하여 다음의 [수학식 2]에 따라 전역손실(Lglobal)을 계산한다.

[0145] [수학식 2]

[0146] $$L_{global} = \sum_{i=1}^{b} \left[\left(y_i^{tot} - Q_i^{tot}(\boldsymbol{\tau}, \boldsymbol{u}, s;\ \theta^{tot}) \right)^2 \right]$$

[0147] 여기서, Q^{tot}는 공동행동가치를 의미하며, Θ^{tot}는 매니저 강화학습모델의 가중치(weight)와 편향(bias)들을 의미한다.

[0148] 또한 [수학식 2]의 y^{tot}는 다음의 [수학식 3]에 따라 계산된다.

[0149] [수학식 3]

[0150] $$y^{tot} = r + \gamma \max_{\boldsymbol{u}'} Q^{tot}(\boldsymbol{\tau}', \boldsymbol{u}', s';\ \bar{\theta}^{tot})$$

[0151] 여기서 r은 전역보상을 의미하며, γ는 감가팩터를 의미한다. τ', u', s'은 최대 행동가치를 가지는 관측데이터, 행동 및 전역 상태데이터를 의미한다.

[0152] 상기 전역손실은 매니저 강화학습모델에서 역전파되어 상기 매니저 강화학습모델을 강화학습하는데 이용된다.

[0153] 이때, 전역손실은 Θ^{tot}기준으로 편미분되어 매니저 강화학습모델에서 역전파된다.

[0154] 또한 매니저(200)는 산출한 전역손실을 각 에이전트(100)로 제공할 수 있다.

[0155] 또한 각 에이전트(100)는 매니저(200)로부터 제공받은 개별보상을 이용하여 다음의 [수학식 4]에 따라 로컬손실을 계산한다.

[0156] [수학식 4]

[0157] $$L_{local}^{a} = \sum_{i=1}^{b} \left[\left(y_i^a - Q_i^a(\tau^a, u^a; \theta^a) \right)^2 \right]$$

[0158] 여기서, a는 각 에이전트(100)를 의미하며, θ^a는 각 에이전트(100)의 에이전트 강화학습모델의 가중치와 편향들을 의미한다.

[0159] 또한 y^a는 다음의 [수학식 5]에 따라 계산된다.

[0160] [수학식 5]

[0161] $$y^a = r^a + \gamma \max_{u_a'} Q^a(\tau_a', u_a'; \bar{\theta}^a)$$

[0162] 여기서, r^a는 개별보상을 의미하며, τ', u'은 가장 높은 행동가치를 가지는 관측데이터와 행동을 각각 의미한다.

[0163] 상기 [수학식 4] 및 [수학식 5]와 같이 개별보상이 주어지면 해당 관측데이터에서의 행동가치를 조정하여 로컬손실이 최소화한다.

[0164] 즉, 개별보상이 높으면 해당 행동에 대한 행동가치를 높게 설정하고, 개별보상이 낮으면 해당 행동에 대한 행동가치를 낮게 설정하여 로컬손실을 최소화하고, 상기 로컬손실을 에이전트 강화학습모델에 역전파하여 상기 에이전트 강화학습모델을 업데이트함으로써 강화학습을 수행한다.

[0165] 이때, 로컬손실은 θ^a를 기준으로 편미분되어 에이전트 강화학습모델에 역전파된다.

[0166] 이를 통해 관측데이터에 따라 개별보상이 높으면서 행동가치가 높은 행동을 추출하는 것이 가능해진다.

[0167] 결국 에이전트(100)는 전역보상으로 집계되는 자신의 개별보상을 늘리고 궁극적으로 감가팩터에 따라 할인된 누적 전역보상을 최대화하도록 강화학습을 수행하는 것이다.

[0168] 로컬손실을 통한 역전파는 개별 에이전트(100)의 동작에 대한 피드백을 명확하게 할 수 있다. 한편 전역손실은 에이전트 네트워크에 대한 영향을 줄 수 있다.

[0169] 따라서 에이전트(100)는 매니저(200)로부터 전역손실이 수신되면, 다음의 [수학식 6]에 따라 최종 손실을 타임스탭별로 계산하여 상기 최종 손실을 에이전트 네트워크에 역전파하여 업데이트할 수 있다.

[0170] [수학식 6]

[0171] $$\mathcal{L} = \mathcal{L}_{global} + \sum_{a=1}^{n} \mathcal{L}_{local}^{a}$$

[0172] 상기에서 설명한 것과 같이, 매니저(200)는 혼합 네트워크로 구성되는 매니저 강화학습모델을 이용하여 모든 에이전트(100)의 행동가치(Q값)를 총 가치인 공동행동가치로 추정하여 각 에이전트 네트워크를 업데이트하는데 사용되도록 하는 것이다.

[0173] 또한 매니저 강화학습모델은 환경과 직접 상호작용하지 않고, 각 에이전트(100)의 상태와 행동가치만을 필요로 한다.

[0174] 또한 매니저 강화학습모델은 강화학습이 완료된 후, 실제 무인운반차량을 제어하는 데에는 사용되지 않는다.

[0175] 즉, 각 에이전트(100)는 강화학습이 완료된 후, 강화학습한 결과에 따라 자체 생성한 관측데이터를 에이전트 강화학습모델에 입력하여 행동을 추출하며, 다른 에이전트(100)와 협력하지 않고 독립적으로 상기 추출한 행동에 따라 무인운반차를 목적지에 따라 최적 경로로 제어하게 된다.

[0176] 도 7은 본 발명의 일 실시예에 따른 무인운반차에 대한 에이전트의 구성을 나타낸 블록도이다.

[0177] 도 7에 도시한 바와 같이, 본 발명의 일 실시예에 따른 에이전트(100)는 관측데이터 생성부(110), 행동 추출부(120), 마스킹 데이터 처리부(130), 행동가치 전송부(140), 보상 수신부(150), 로컬손실 계산부(160), 에이전트

강화학습모델 강화학습부(170) 및 무인운반차 제어부(180)를 포함하여 구성된다.

[0178] 관측데이터 생성부(110)는 강화학습을 수행할 때, 이전에 공유한 마스킹 데이터에 따라 에이전트(100)를 중심으로 사전에 설정한 범위 내의 관측데이터를 생성하는 기능을 수행한다.

[0179] 관측데이터 생성부(110)는 제1 채널 이미지 생성부(111) 및 제2 채널 이미지 생성부(112)를 포함하여 구성된다.

[0180] 제1 채널 이미지 생성부(111)는 상기 사전에 설정한 범위 내의 각 셀에 대한 이동 가능여부를 표시한 제1 채널 이미지를 생성한다.

[0181] 또한 제2 채널 이미지 생성부(112)는 목적지까지 남은 거리정보(맨해튼 거리)를 상기 사전에 설정한 범위 내의 각 셀에 표시한 제2 채널 이미지를 생성한다.

[0182] 즉, 관측데이터는 제1 및 제2 채널 이미지를 포함하여 구성되며, 상기 제1 및 제2 채널 이미지를 생성하는 것은 도 3 및 도 4를 각각 참조하여 설명하였으므로 여기서는 생략하도록 한다.

[0183] 한편, 관측데이터 생성부(110)는 강화학습이 종료된 경우, 마스킹 데이터 필요 없이 독립적으로 관측데이터를 생성하게 된다.

[0184] 또한 행동 추출부(120)는 강화학습을 수행할 때, 현재 타임스텝에서 생성한 관측데이터와 이전의 타임스탭에서 추출한 행동을 에이전트 강화학습모델에 입력하여 행동가치가 제일 높은 행동을 추출하는 기능을 수행한다.

[0185] 즉, 행동 추출부(120)는 에이전트 강화학습모델의 출력에서 제일 높은 행동가치를 선택함으로써 행동을 추출하게 된다.

[0186] 한편, 행동 추출부(120)는 강화학습이 종료되면 독립적으로 생성한 관측데이터와 이전에 추출한 행동을 에이전트 강화학습모델에 입력하여 행동가치가 제일 높은 행동을 추출한다.

[0187] 마스킹 데이터 처리부(130)는 강화학습을 수행할 때, 무인운반차 제어부(180)를 통해 추출한 행동에 따라 무인운반차를 제어한 결과에 대한 마스킹 데이터를 생성하는 기능을 수행하고, 상기 생성한 마스킹 데이터를 관리서버의 매니저(200)를 통해 다음의 에이전트(100)로 공유하도록 한다.

[0188] 특정 에이전트의 상기 마스킹 데이터는 이전 에이전트(100)로부터 공유받은 마스킹 데이터에 자신의 제어결과를 마스킹함으로써 생성됨은 상술한 바와 같다.

[0189] 한편 마스킹 데이터를 생성하고 공유하는 과정은 도 5를 참조하여 설명하였으므로 더 이상상의 상세한 설명은 생략하도록 한다.

[0190] 행동가치 전송부(140)는 강화학습을 수행할 때, 상기 행동 추출부(120)의 선택에 따라 현재의 관측데이터를 포함한 과거의 관측데이터, 상기 추출한 행동을 포함하는 행동가치를 매니저(200)로 전송하는 기능을 수행한다.

[0191] 보상 수신부(150)는 매니저(200)로부터 해당 에이전트(100)에 대해 제공한 개별보상을 수신하는 기능을 수행한다.

[0192] 로컬손실 계산부(160)는 보상 수신부(150)를 통해 수신한 개별보상과 행동을 추출할 때 선택한 행동가치를 이용하여 로컬손실을 계산하는 기능을 수행한다.

[0193] 상기 로컬손실 계산부(160)는 개별보상에 따라 상기 행동가치를 조정하여 로컬손실이 최소화되도록 한다.

[0194] 한편 로컬손실을 계산하는 것은 도 6을 참조하여 상세히 설명하였으므로 더 이상의 상세한 설명은 생략한다.

[0195] 에이전트 강화학습모델 강화학습부(170)는 로컬손실을 에이전트 강화학습모델에서 역전파하여 에이전트 강화학습모델을 업데이트함으로써 상기 에이전트 강화학습모델에 대한 강화학습을 수행하는 기능을 수행한다. 이를 통해 관측데이터에 따라 개별보상이 높은 행동가치에 해당하는 행동을 추출할 수 있도록 함으로써 강화학습이 종료된 후, 타 에이전트(100)와의 협력 없이 독립적으로 생성한 관측데이터만으로도 최적의 행동을 추출하도록 하여 목적지에 따라 최적 경로로 무인운반차를 제어할 수 있도록 한다.

[0196] 또한 무인운반차 제어부(180)는 행동 추출부(120)에서 추출한 행동에 따라 해당 무인운반차를 제어하는 기능을 수행한다.

[0197] 도 8은 본 발명의 일 실시예에 따른 매니저의 구성을 나타낸 블록도이다.

[0198] 도 8에 도시한 바와 같이, 본 발명의 일 실시예에 따른 매니저(200)는 마스킹 데이터 공유부(210), 행동가치 수

신부(220), 보상 처리부(230), 공동행동가치 추출부(240), 전역손실 계산부(250) 및 매니저 강화학습모델 강화학습부(260)를 포함하여 구성된다.

[0199] 마스킹 데이터 공유부(210)는 각 매니저(100)로부터 수신한 마스킹 데이터를 다음의 에이전트(100)로 전송하여 공유하는 기능을 수행한다.

[0200] 이때, 매니저(200)는 강화학습을 수행하는 경우, 에이전트(100)에 부여된 식별자 순으로 에이전트(100)에 행동을 순차적으로 추출하도록 명령하고 순차적으로 마스킹 데이터를 각각 수신하여 다음으로 행동을 추출할 에이전트(100)로 마스킹 데이터를 공유하도록 할 수 있다.

[0201] 행동가치 수신부(220)는 각 에이전트(100)로부터 행동가치를 각각 수신하는 기능을 수행한다.

[0202] 공동행동가치 추출부(240)는 매니저 강화학습모델에 각 에이전트(100)로부터 수신한 행동가치와 각 에이전트(100)에 대한 전역 상태데이터를 입력하여 모든 행동에 대한 공동행동가치를 추출하는 기능을 수행한다.

[0203] 여기서 전역 상태데이터는 현장의 레이아웃(예: 1 x 20 x 20)으로 구성되며, 시나리오에 따라 레이아웃을 구성하는 각 객체에 대한 정규화된 누적 방문수를 포함하여 구성된다.

[0204] 또한 전역손실 계산부(250)는 전역손실을 계산하는 기능을 수행한다.

[0205] 상기 전역손실은 공동행동가치를 이용하여 계산된다. 한편 공동행동가치를 추출하는 것과 전역손실을 계산하는 것은 도 6을 참조하여 설명하였으므로 여기서는 생략하도록 한다.

[0206] 매니저 강화학습모델 강화학습부(260)는 전역손실을 매니저 강화학습모델에 역전파시켜 상기 매니저 강화학습모델을 업데이트함으로써 상기 매니저 강화학습모델에 대한 강화학습을 수행한다.

[0207] 한편, 전체적인 강화학습이 종료되면, 매니저(200)의 역할은 중단되며, 개별 에이전트(100)는 다른 에이전트(100)간의 협력 없이 현재의 관측데이터에 따라 행동을 추출하여 목적지까지 최적 경로로 무인운반차를 독립적으로 제어한다.

[0208] 도 9는 본 발명의 일 실시예에 따른 다중 에이전트 강화학습 기반 다중 무인운반차를 최적의 경로로 제어하는 절차를 나타낸 흐름도이다.

[0209] 도 9에 도시한 바와 같이, 본 발명의 일 실시예에 따른 다중 에이전트 강화학습 기반 다중 무인운반차를 최적의 경로로 제어하는 절차는 우선, 다중 에이전트 강화학습 기반 다중 무인운반차 최적 경로 시스템(10)은 사전에 설정한 시나리오를 복수회 반복하여 무인운반차를 최적 경로로 제어하기 위한 강화학습 단계를 수행한다.

[0210] 상기 강화학습 단계는 우선, 각 에이전트(100)에서 관측데이터를 생성하는 관측데이터 생성 단계를 수행한다(S105).

[0211] 관측데이터는 제1 채널 이미지와 제2 채널 이미지를 포함하여 구성됨은 상술한 바와 같다.

[0212] 다음으로 에이전트(100)에서, 관측데이터와 이전에 추출한 행동을 에이전트 강화학습모델에 입력하여 행동가치가 높은 행동을 추출하는 행동 추출 단계를 수행한다(S110).

[0213] 한편, 관측데이터를 생성하고, 행동을 추출하는 것은 에이전트(100)별로 사전에 부여한 식별자 순으로 수행되며, 상기 관측데이터는 이전의 에이전트(100)로부터 공유받은 마스킹 데이터를 참조하여 생성됨은 상술한 바와 같다.

[0214] 따라서 행동을 추출한 에이전트(100)는 이후의 에이전트(100)로 상기 추출한 행동에 따라 무인운반차를 제어한 결과를 현장의 레이아웃에 마스킹한 마스킹 데이터를 관리서버의 매니저(200)를 통해 공유한다.

[0215] 이러한 마스킹 데이터는 이후의 각 에이전트(100)에서 무인운반차를 제어한 결과가 추가되어 이후의 에이전트(100)로 순차 공유된다.

[0216] 다음으로 에이전트(100)에서, 추출한 행동에 따라 현재의 관측데이터를 포함하여 과거의 관측데이터, 상기 추출한 행동을 포함하는 행동가치를 매니저(200)로 전송하는 행동가치 전송 단계를 수행한다(S115).

[0217] 다음으로 매니저(200)에서, 각 에이전트(100)별로 수신한 행동가치의 행동에 따라 개별보상 및 전역보상을 산출하는 보상 산출 단계를 수행한다(S120).

[0218] 개별보상은 상기 행동에 따라 무인운반차를 제어한 결과가 목적지에 가까워지거나 목적지에 도착하면 양(+1)의

보상 및 그 외에 부여할 음(-1)의 보상을 책정하고 상기 책정한 보상을 에이전트(100)의 총 수로 나눔으로써 산출되며, 전역보상은 상기 개별보상을 총합함으로써 산출됨은 상술한 바와 같다.

[0219] 다음으로 매니저(200)에서, 각 행동가치와 전역 상태데이터를 입력하여 공동행동가치를 추정하는 공동행동가치 추정 단계를 수행한다(S125).

[0220] 전역 상태데이터는 각 시나리오에 따라 현장의 레이아웃에서 각 객체에 대한 정규화된 누적 방문수를 포함하여 구성됨은 상술한 바와 같다.

[0221] 다음으로 매니저(200)에서, 공동행동가치와 전역보상에 따라 계산한 전역손실, 개별보상을 각 에이전트(100)로 전송하는 단계를 수행한다(S130).

[0222] 다음으로 매니저(200)는 개별보상을 이용하여 로컬손실을 계산하고, 상기 로컬손실을 에이전트 강화학습모델에 역전파시켜 에이전트 강화학습모델에 대한 강화학습을 수행하는 에이전트 강화학습모델 강화학습 단계를 수행한다(S135).

[0223] 상기 로컬손실은 개별보상에 따라 최소화되는 방향으로 계산됨은 상술한 바와 같다.

[0224] 다음으로 전체적인 강화학습이 종료되면(S140), 에이전트(100)에서 강화학습을 수행한 결과에 따른 에이전트 강화학습모델을 이용하여 무인운반차를 독립적으로 제어하는 무인운반차 제어 단계를 수행한다(S145).

[0225] 이하에서는 가장 기본적인 다중 에이전트 강화학습 알고리즘으로 간주되는 IQL 알고리즘과 성능을 비교한 결과를 설명하도록 한다. 본 출원인은 상기 비교를 통해 본 발명의 효과 및 타당성을 검증하였다.

[0226] 또한 본 발명은 분산 실행 프레임워크인 Qmix를 기반으로 설계하였다.

[0227] 또한 비교를 위한 실험은 스몰(small) 레이아웃과 라지(large) 레이아웃으로 수행되었다.

[0228] 스몰 레이아웃은 12 x 16 크기로 8개의 무인운반차(에이전트)와 3개의 피킹스테이션을 포함하여 구성하였으며, 라지 레이아웃은 20 x 20 크기로 14개의 무인운반차(에이전트)와 5개의 피킹스테이션을 포함하여 구성하였다.

[0229] 스몰 레이아웃과 라지 레이아웃은 도 10에 각각 표시하였다.

[0230] 자세한 평가를 위해 에피소드 보상, 평균 경로 길이 및 피킹스테이션에 도착한 선반 수를 포함한 세 가지 측정 항목(메트릭)을 채택하였다.

[0231] 강화학습 중, 각 메트릭은 모든 에이전트(100)가 탐욕스러운 행동을 취한 5개의 에피소드에 대한 평균값으로 측정하였다.

[0232] 두 레이아웃의 모든 메트릭에 대해 95% 신뢰구간으로 5개의 에피소드에 따라 실행한 평균을 표시하였다.

[0233] 강화학습은 로컬손실 혹은 전연손실과 로컬손실을 합산한 최종손실에 의한 에이전트 강화학습모델의 업데이트를 의미한다.

[0234] 도 11은 본 발명의 일 실시예에 따른 종래 기술과의 성능비교 결과를 나타낸 도면이다.

[0235] 도 11의 (a)는 본 발명의 일 실시예에 따른 스몰 레이아웃에서 성능비교 결과를 나타낸 도면이며, 도 11의(b)는 본 발명의 일 실시예에 따른 라지 레이아웃에서 성능비교 결과를 나타낸 도면이다.

[0236] 도 11의 (a) 및 도 11의 (b)에 도시한 것과 같이, 에피소드 보상과 피킹스테이션에 도착한 선반 수에 대한 그래프를 볼 수 있듯이 본 발명이 훨씬 더 빠르고 일관되게 강화학습하는 것을 알 수 있다.

[0237] 더 구체적으로 설명하면, 본 발명은 보다 더 나은 값(에피소드 보상, 피킹스테이션에 도착한 선반 수)에서 안정적으로 수렴을 달성하는 반면에 IQL은 더 나쁜 값에서도 불안정한 경향이 있음을 알 수 있다.

[0238] 이러한 대비는 스몰 레이아웃보다 라지 레이아웃에서 더 분명하게 나타남을 알 수 있다.

[0239] IQL는 에이전트 간의 정보교환이 없기 때문에 에이전트의 수가 증가할수록 IQL의 성능은 급격하게 저하된다.

[0240] 반면에 본 발명 에이전트 수가 증가하더라도 에이전트 간의 효율적인 정보교환(즉, 마스킹 데이터 공유)을 통한 협력으로 강력한 성능을 발휘함을 알 수 있다.

[0241] 이러한 결과는 평균 경로 길이에 대한 그래프에서 확인할 수 있다.

[0242] 즉, 평균 경로 길이에 대한 그래프에서 알 수 있듯이 본 발명은 강화학습이 진행됨에 따라 안정적으로 짧은 경

로로 목적지까지 무인운반차를 제어함을 알 수 있다. 반면에 IQL은 수렴하지 않고 매우 불안정한 형태를 나타냄을 알 수 있다.

[0243] 특히 IQL의 경우, 평균 경로 길이는 라지 레이아웃에서 더욱더 길어짐을 알 수 있다.

[0244] 결과적으로 레이아웃의 구조가 복잡할수록(즉, 무인운반차가 많을수록) 에이전트간의 협업이 더 중요한 것임을 알 수 있다.

[0245] 도 12는 본 발명의 또 다른 일 실시예에 따른 성능비교 결과를 나타낸 도면이다.

[0246] 도 12는 에이전트(100)의 실제 행동패턴을 분석하기 위해 무인운반차의 이동에 따른 누적 방문 횟수를 히트맵(heat map)으로 시각화한 것이다.

[0247] 또한 500번의 타임스탭 동안의 에피소드에 따라 본 발명과 IQL에 따른 무인운반차의 누적 방문 횟수를 라지 레이아웃에 시각화하였다.

[0248] 도 12에 도시한 바와 같이, IQL의 경우, 특정 무인운반차량이 임의적으로 부적절한 위치에서 유휴상태(고립)에 빠지는 사실을 알 수 있다.

[0249] 즉, IQL에 따른 히트맵의 경우, 주요 경로가 아닌 히트맵에 표시된 매우 밝은 특정 위치에서 유휴상태에 빠지는 경우가 발생하였다. 이는 에이전트(100)간의 협력부재가 치명적인 문제로 발전하는 것으로 해석할 수 있다.

[0250] 대조적으로 본 발명은 모든 무인운반차가 주요 경로인 피킹스테이션과 선반 사이를 이동함을 알 수 있다.

[0251] 이러한 결과를 봤을 때, 본 발명의 강화학습한 결과에 따라 에이전트(100)가 전체 레이아웃에서 주요 경로로 이동함을 알 수 있으며, 혼잡, 충돌, 고립 등이 발생하지 않음을 알 수 있다.

[0252] 따라서 본 발명은 모든 무인운반차의 상태 순환이 원활하게 이루어지며 현장의 생산성을 높일 수 있음을 확인할 수 있다.

[0253] 이상에서 설명한 바와 같이, 본 발명의 다중 에이전트 강화학습 기반 다중 무인운반차 최적 경로 제어 시스템은 복수의 에이전트 간에 협력과, 로컬손실을 통해 무인운반차를 제어하기 위한 에이전트 강화학습모델에 대한 강화학습을 수행함으로써 각 무인운반차의 목적지에 따라 최적의 경로로 제어할 수 있는 효과가 있다.

[0254] 상기에서는 본 발명에 따른 바람직한 실시예를 위주로 상술하였으나, 본 발명의 기술적 사상은 이에 한정되는 것은 아니며 본 발명의 각 구성요소는 동일한 목적 및 효과의 달성을 위하여 본 발명의 기술적 범위 내에서 변경 또는 수정될 수 있을 것이다.

[0255] 또한, 이상에서는 본 발명의 바람직한 실시예에 대하여 도시하고 설명하였지만, 본 발명은 상술한 특정의 실시예에 한정되지 아니하며, 청구범위에서 청구하는 본 발명의 요지를 벗어남이 없이 당해 발명이 속하는 기술분야에서 통상의 지식을 가진 자에 의해 다양한 변형 실시가 가능한 것은 물론이고, 이러한 변형 실시들은 본 발명의 기술적 사상이나 전망으로부터 개별적으로 이해되어서는 안 될 것이다.

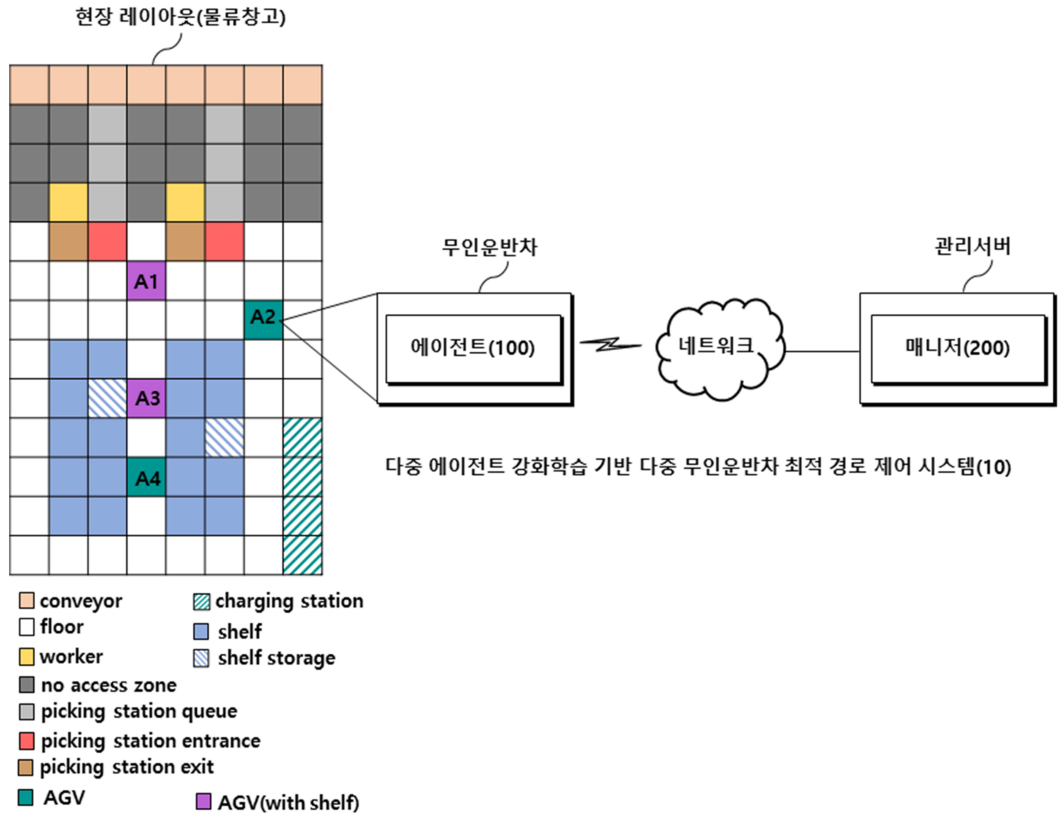
부호의 설명

[0256] 10: 다중 에이전트 강화학습 기반 다중 무인운반차 최적 경로 제어 시스템

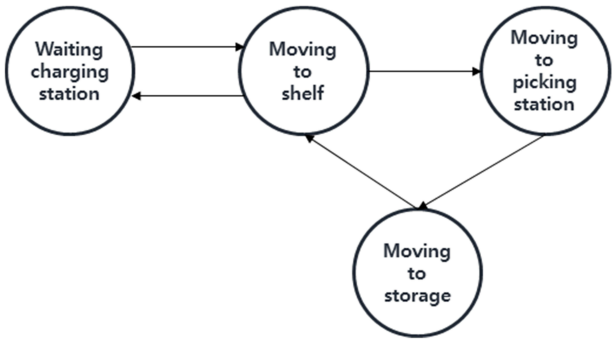
100: 에이전트	110: 관측데이터 생성부
111: 제1 채널 이미지 생성부	112: 제2 채널 이미지 생성부
120: 행동 추출부	130: 마스킹 데이터 처리부
140: 행동가치 전송부	150: 보상 수신부
160: 로컬손실 계산부	170: 에이전트 강화학습모델 강화학습부
180: 무인운반차 제어부	200: 매니저
210: 마스킹 데이터 공유부	220: 행동가치 수신부
230: 보상 처리부	240: 공동행동가치 추출부
250: 전역손실 계산부	260: 매니저 강화학습모델 강화학습부

도면

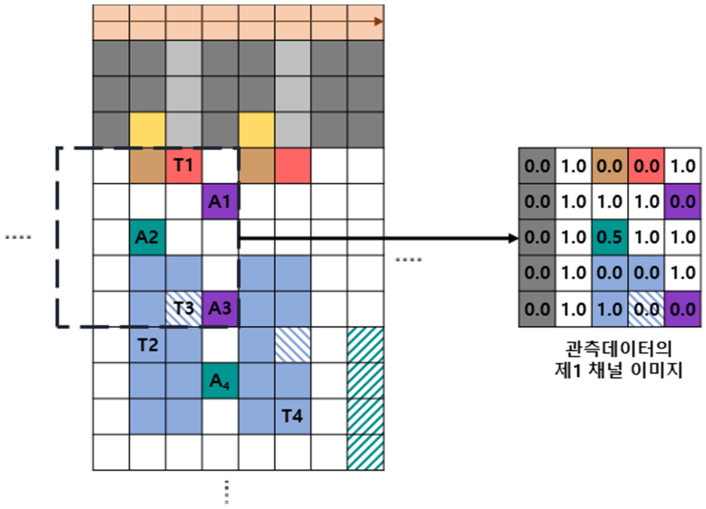
도면1


현장 레이아웃(물류창고)
A1
A2
A3
A4
무인운반차
에이전트(100)
네트워크
관리서버
매니저(200)
다중 에이전트 강화학습 기반 다중 무인운반차 최적 경로 제어 시스템(10)
conveyor
floor
worker
no access zone
picking station queue
picking station entrance
picking station exit
AGV
charging station
shelf
shelf storage
AGV(with shelf)

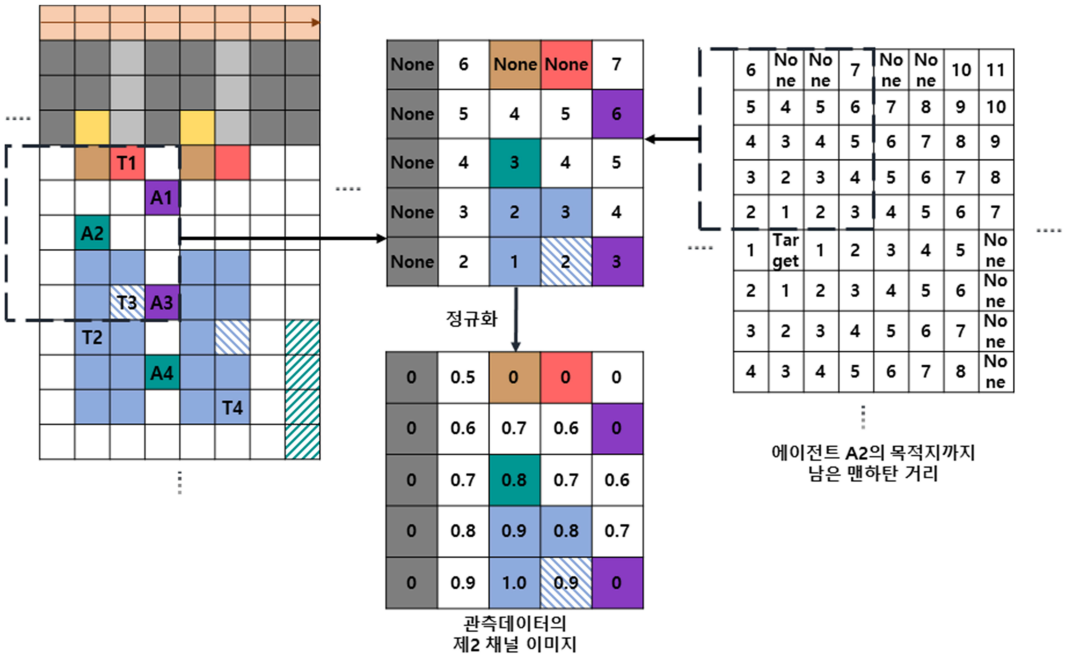
도면2


Waiting charging station
Moving to shelf
Moving to picking station
Moving to storage

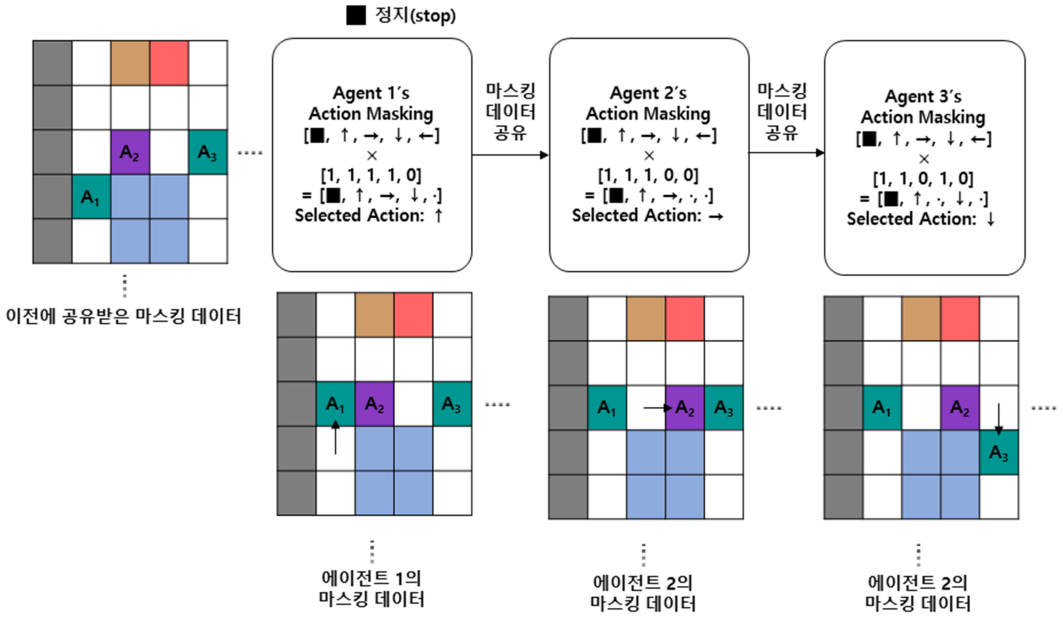
도면3


T1
A1
A2
T3
A3
T2
A4
T4
0.0 1.0 0.0 0.0 1.0
0.0 1.0 1.0 1.0 0.0
0.0 1.0 0.5 1.0 1.0
0.0 1.0 0.0 0.0 1.0
0.0 1.0 1.0 0.0 0.0
관측데이터의
제1 채널 이미지

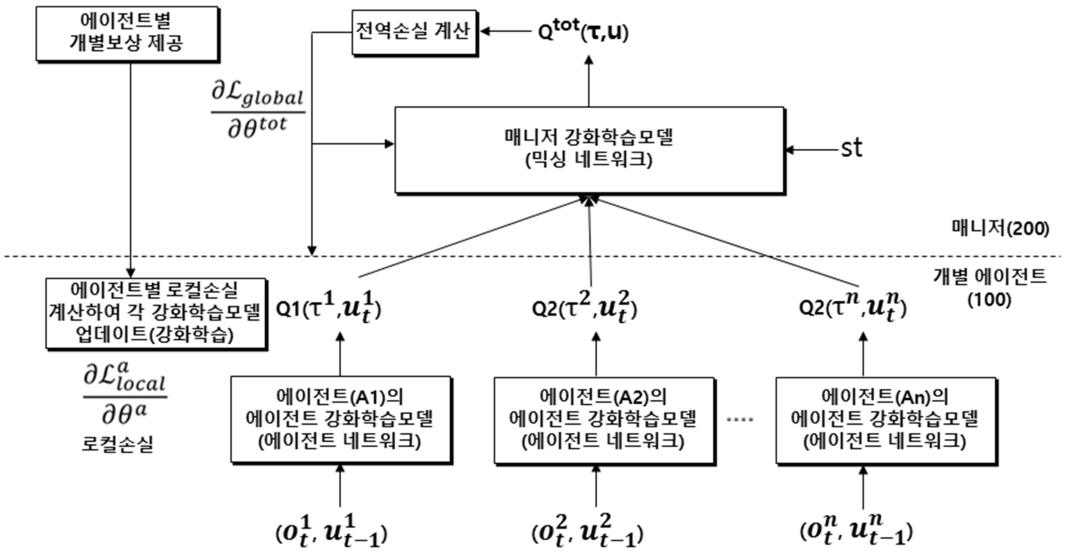
도면4


T1
A1
A2
T3
A3
T2
A4
T4
None 6 None None 7
None 5 4 5 6
None 4 3 4 5
None 3 2 3 4
None 2 1 2 3
정규화
0 0.5 0 0 0
0 0.6 0.7 0.6 0
0 0.7 0.8 0.7 0.6
0 0.8 0.9 0.8 0.7
0 0.9 1.0 0.9 0
관측데이터의
제2 채널 이미지
6 None None 7 None None 10 11
5 4 5 6 7 8 9 10
4 3 4 5 6 7 8 9
3 2 3 4 5 6 7 8
2 1 2 3 4 5 6 7
1 Target 1 2 3 4 5 None
2 1 2 3 4 5 6 None
3 2 3 4 5 6 7 None
4 3 4 5 6 7 8 None
에이전트 A2의 목적지까지
남은 맨하탄 거리

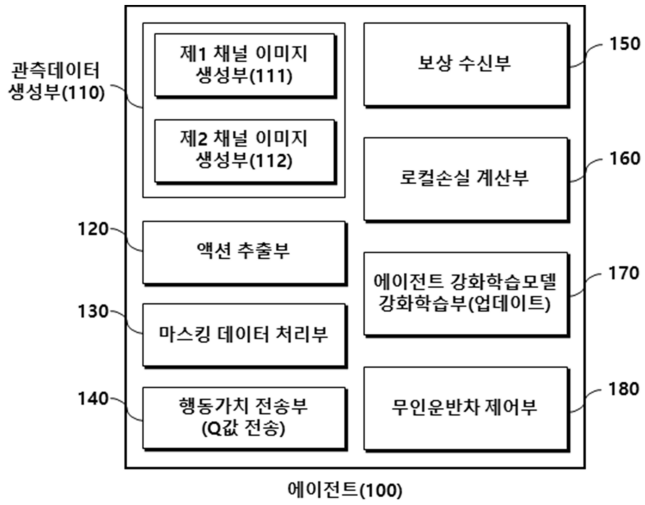
도면5


■ 정지(stop)
Agent 1's Action Masking
[■, ↑, →, ↓, ←]
×
[1, 1, 1, 1, 0]
= [■, ↑, →, ↓, ·]
Selected Action: ↑
마스킹 데이터 공유
Agent 2's Action Masking
[■, ↑, →, ↓, ←]
×
[1, 1, 1, 0, 0]
= [■, ↑, →, ·, ·]
Selected Action: →
마스킹 데이터 공유
Agent 3's Action Masking
[■, ↑, →, ↓, ←]
×
[1, 1, 0, 1, 0]
= [■, ↑, ·, ↓, ·]
Selected Action: ↓
이전에 공유받은 마스킹 데이터
에이전트 1의 마스킹 데이터
에이전트 2의 마스킹 데이터
에이전트 2의 마스킹 데이터

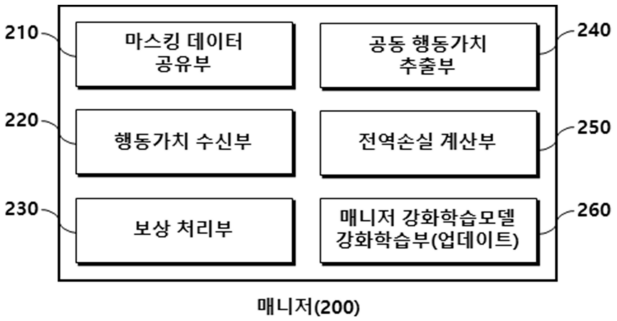
도면6


에이전트별 개별보상 제공
전역손실 계산
$Q^{tot}(\tau,u)$
$\frac{\partial \mathcal{L}_{global}}{\partial \theta^{tot}}$
매니저 강화학습모델 (믹싱 네트워크)
st
매니저(200)
개별 에이전트 (100)
에이전트별 로컬손실 계산하여 각 강화학습모델 업데이트(강화학습)
Q1(τ^1,u_t^1)
Q2(τ^2,u_t^2)
Q2(τ^n,u_t^n)
$\frac{\partial \mathcal{L}_{local}^{a}}{\partial \theta^{a}}$
로컬손실
에이전트(A1)의 에이전트 강화학습모델 (에이전트 네트워크)
에이전트(A2)의 에이전트 강화학습모델 (에이전트 네트워크)
에이전트(An)의 에이전트 강화학습모델 (에이전트 네트워크)
(o_t^1, u_{t-1}^1)
(o_t^2, u_{t-1}^2)
(o_t^n, u_{t-1}^n)

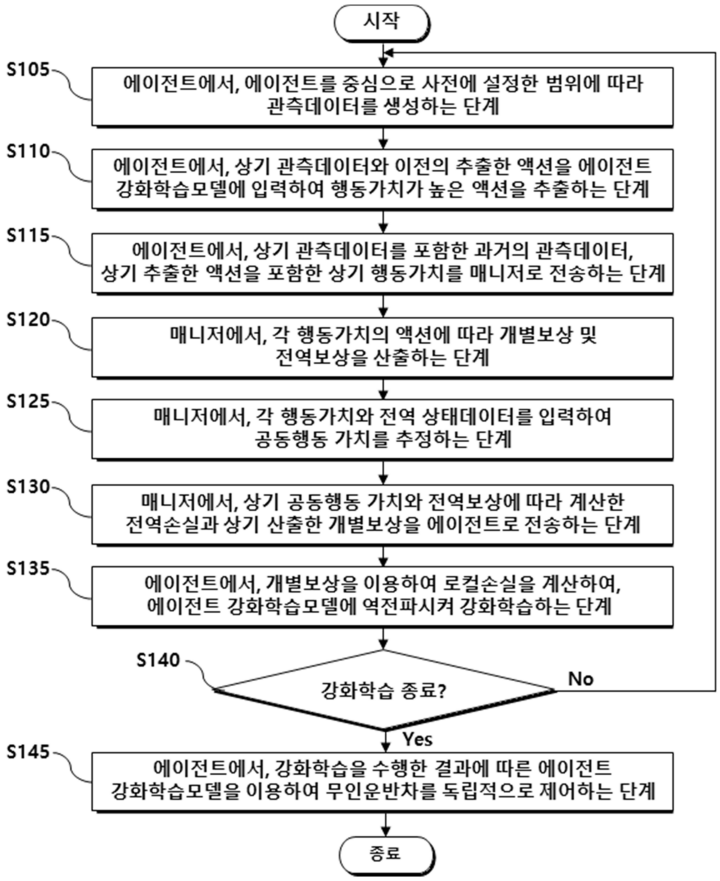
도면7


관측데이터
생성부(110)
제1 채널 이미지
생성부(111)
제2 채널 이미지
생성부(112)
120
액션 추출부
130
마스킹 데이터 처리부
140
행동가치 전송부
(Q값 전송)
보상 수신부
150
로컬손실 계산부
160
에이전트 강화학습모델
강화학습부(업데이트)
170
무인운반차 제어부
180
에이전트(100)

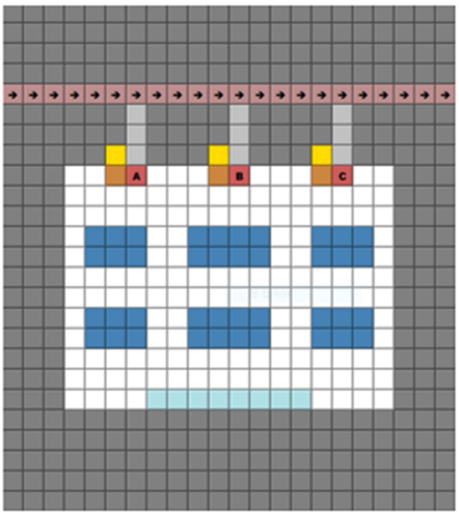
도면8


210
마스킹 데이터
공유부
220
행동가치 수신부
230
보상 처리부
공동 행동가치
추출부
240
전역손실 계산부
250
매니저 강화학습모델
강화학습부(업데이트)
260
매니저(200)

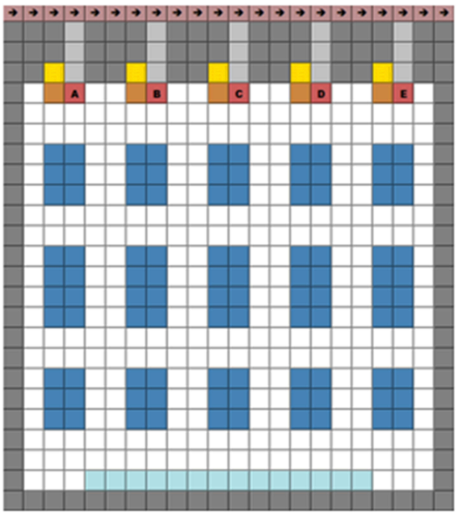
도면9



도면10

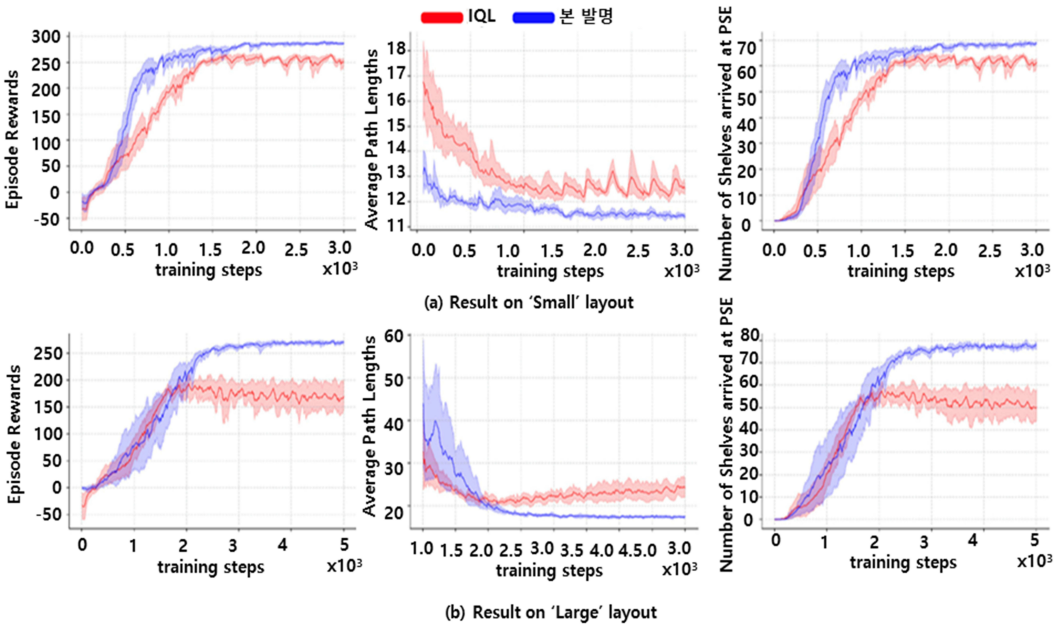


스몰(small) 레이아웃

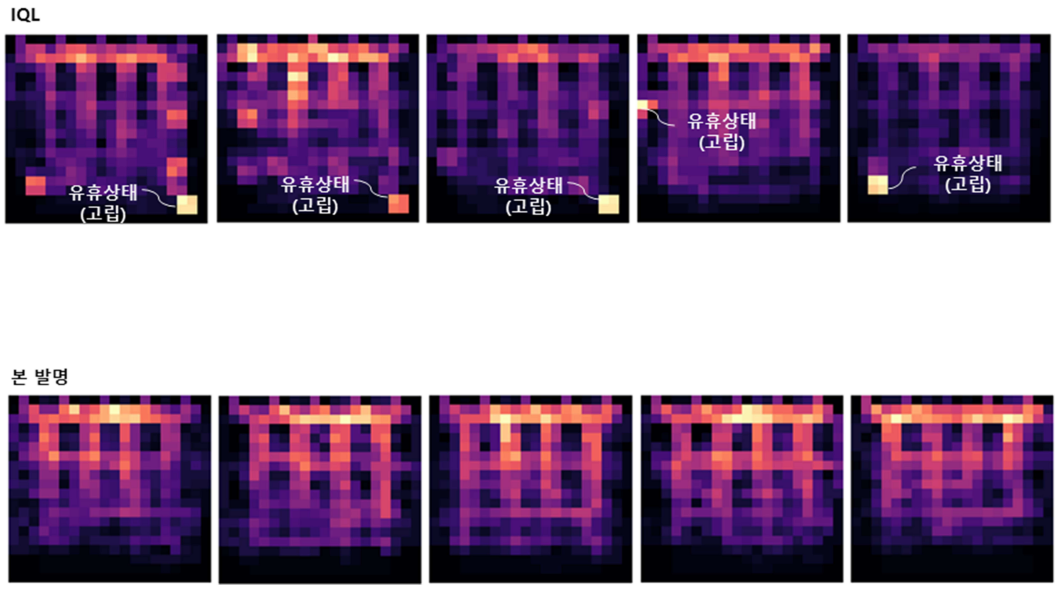


라지(large) 레이아웃

도면11


IQL
본 발명
Episode Rewards
Average Path Lengths
Number of Shelves arrived at PSE
training steps
x10³
(a) Result on 'Small' layout
(b) Result on 'Large' layout

도면12


IQL
유휴상태
(고립)
본 발명