



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0102395
(43) 공개일자 2022년07월20일

(51) 국제특허분류(Int. Cl.)
G08G 1/081 (2006.01) G08G 1/01 (2006.01)
G08G 1/0967 (2006.01)
(52) CPC특허분류
G08G 1/081 (2013.01)
G08G 1/0145 (2013.01)
(21) 출원번호 10-2021-0004701
(22) 출원일자 2021년01월13일
심사청구일자 2021년01월13일

(71) 출원인
부경대학교 산학협력단
부산광역시 남구 용소로 45, 부경대학교내 (대연동)
에스에이엠(주)
부산광역시 남구 용소로 45, 환경해양대학 606호 (대연동)
(72) 발명자
배상훈
부산광역시 남구 용소로 45 부경대학교대연캠퍼스
(74) 대리인
오위환, 나성곤, 정기택

전체 청구항 수 : 총 15 항

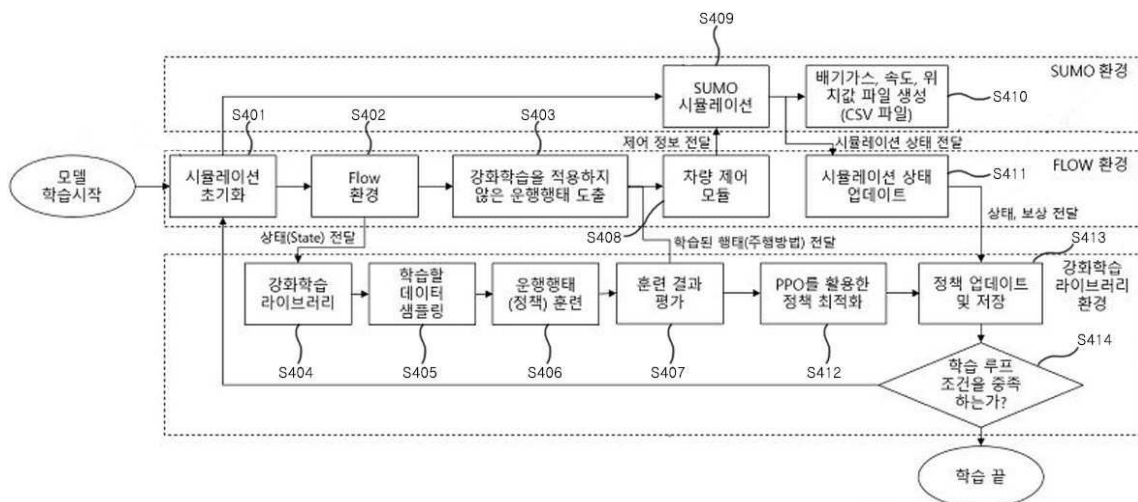
(54) 발명의 명칭 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치 및 방법

(57) 요약

본 발명은 군집된 자율주행차량과 인간운전자 차량이 혼재되어 있는 혼합 교통류 상황에서 자율주행차량 군집주행 학습으로 비신호 교차로 통행을 개선하고 안전성을 확보할 수 있도록 한 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치 및 방법에 관한 것으로, SUMO(Simulation of Urban

(뒷면에 계속)

대표도 - 도4



MObility)를 활용하여 시뮬레이션 환경을 구축하고, 자율 주행 차량의 속도, 위치, 센서에서 얻어진 데이터를 FLOW 적용부로 전달하는 SUMO 시뮬레이션 실행부; SUMO와 연동할 수 있는 강화학습 플랫폼 FLOW 환경에서 시뮬레이션 환경을 구축하고, 강화학습을 적용하지 않은 운전 행태 도출, 차량 제어 및 시뮬레이션 상태 업데이트를 하여 상태 및 보상 정보를 강화학습 라이브러리 환경 구축부로 전달하는 FLOW 적용부; SUMO와 연동할 수 있는 강화학습 플랫폼 FLOW를 활용하여 멀티 에이전트 심층강화학습(Multi agent Deep Reinforcement Learning)으로 통행 제어를 최적화하는 강화학습 라이브러리 환경 구축부;를 포함하고, SUMO 시뮬레이션 실행부는 군집 차량이 비신호 교차로에 접근하여 네 가지 다른 방향을 따라 직진 주행하는 상황에서, 1% ~ 100%의 AV 보급률에 대한 결과를 10% 단위로 제시하고, 비신호 교차로에서 모든 차량에 대해 차선 변경과 좌회전을 무시한다.

(52) CPC특허분류

G08G 1/096725 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	2020184
과제번호	2020-184
부처명	중소벤처기업부
과제관리(전문)기관명	중소기업기술정보진흥원
연구사업명	2020년 예비창업패키지[특화분야]
연구과제명	고속도로 주요 결렬지점 고정밀 도로지도 제작 및 자율주행 안전제고 기술
기 여 율	1/1
과제수행기관명	한국도로공사
연구기간	2020.06.01 ~ 2021.01.31

명세서

청구범위

청구항 1

SUMO(Simulation of Urban MObility)를 활용하여 시뮬레이션 환경을 구축하고, 자율 주행 차량의 속도, 위치, 센서에서 얻어진 데이터를 FLOW 적용부로 전달하는 SUMO 시뮬레이션 실행부;

SUMO와 연동할 수 있는 강화학습 플랫폼 FLOW 환경에서 시뮬레이션 환경을 구축하고, 강화학습을 적용하지 않은 운전 행태 도출, 차량 제어 및 시뮬레이션 상태 업데이트를 하여 상태 및 보상 정보를 강화학습 라이브러리 환경 구축부로 전달하는 FLOW 적용부;

SUMO와 연동할 수 있는 강화학습 플랫폼 FLOW를 활용하여 멀티 에이전트 심층강화학습(Multi agent Deep Reinforcement Learning)으로 통행 제어를 최적화하는 강화학습 라이브러리 환경 구축부;를 포함하는 것을 특징으로 하는 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치.

청구항 2

제 1 항에 있어서, SUMO 시뮬레이션 실행부는,

SUMO(Simulation of Urban MObility)를 활용하여 시뮬레이션 환경을 구축하여 SUMO 시뮬레이션을 수행하는 SUMO 시뮬레이션부와,

배기가스, 속도 및 위치값 파일을 생성하여 자율 주행 차량의 속도, 위치, 센서에서 얻어진 데이터를 FLOW 적용부로 전달하는 결과 파일생성부를 포함하는 것을 특징으로 하는 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치.

청구항 3

제 1 항에 있어서, SUMO 시뮬레이션 실행부는 군집 차량이 비신호 교차로에 접근하여 네 가지 다른 방향을 따라 직진 주행하는 상황에서,

1% ~ 100%의 AV 보급률에 대한 결과를 10% 단위로 제시하고, 비신호 교차로에서 모든 차량에 대해 차선 변경과 좌회전을 무시하는 것을 특징으로 하는 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치.

청구항 4

제 1 항에 있어서, FLOW 적용부는,

인간운전자 정의, 심층 강화학습 입력값 설정 및 차량의 속도, 가속도, 출발점을 포함하는 시뮬레이션 환경 설정을 하는 시뮬레이션 초기화부와,

FLOW 환경 구축을 하여 상태(state)를 강화학습 라이브러리로 전달하는 FLOW 환경 구축부와,

강화학습을 적용하지 않은 운전 행태 도출을 하는 운행행태 도출부와,

차량 제어를 하고 제어 정보를 SUMO 시뮬레이션 실행부로 전달하는 차량 제어 모듈과,

SUMO 시뮬레이션 실행부로부터 시뮬레이션 상태를 받아 시뮬레이션 상태 업데이트를 하여 상태 및 보상 정보를 강화학습 라이브러리 환경 구축부로 전달하는 업데이트부를 포함하는 것을 특징으로 하는 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치.

청구항 5

제 1 항에 있어서, 강화학습 라이브러리 환경 구축부는,

FLOW 적용부로부터 상태(state)를 전달받는 강화학습 라이브러리와,

학습할 데이터를 샘플링하는 데이터 샘플링부와,

운전 행태(정책) 훈련을 하는 정책 훈련부와,

훈련 결과를 평가하고 학습된 행태를 FLOW 적용부로 전달하는 훈련 결과 평가부와,

자율주행차량이 관찰할 수 있는 범위 내의 정보를 통하여 학습하는 방법으로 강화학습과 마르코프 의사결정 모델 사용(Partial Observability MDP, POMDP)을 적용하여 행동에 대한 강화학습의 보상을 최대화할 수 있도록 하는 정책 최적화부를 포함하는 것을 특징으로 하는 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치.

청구항 6

제 5 항에 있어서, 강화학습 라이브러리 환경 구축부는,

FLOW 적용부로부터 자율 주행 차량의 상태를 받아 정책 업데이트 및 저장을 하는 정책 업데이트 저장부와,

업데이트된 정책이 학습 루프 조건을 만족하는지 판단하는 학습 루프조건 판단부를 더 포함하는 것을 특징으로 하는 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치.

청구항 7

제 5 항에 있어서, 강화학습 라이브러리 환경 구축부는,

정책 최적화(Policy Optimization)를 위하여 동작 값이나 상태 값 함수가 아닌 경사 강하 알고리즘을 사용하여 매개 변수화된 정책 함수의 추정기를 계산하는 정책 경사 방법(Policy gradient methods)을 적용하여,

비선형 근사 및 부분 관측으로 인해 추정 함수에 발생하는 수렴 문제를 피하도록 하는 것을 특징으로 하는 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치.

청구항 8

제 7 항에 있어서, 강화학습 라이브러리 환경 구축부는,

비신호화된 교차로의 시뮬레이션에서 제어 정책을 직접 최적화하기 위해 MLP(multilayer perceptron)정책을 적용하고,

정책 행동($\log \pi_{\theta}$)의 확률에 대한 기대치와 시간 스텝 $t(\hat{A}_t)$ 에서의 어드밴티지 함수(advantage function)의 추정치에 기초하는 정책 경사법은,

$$\hat{g} = \hat{E}_t[\nabla_{\theta} \log \pi_{\theta}(a_t|s_t)\hat{A}_t] \quad \text{으로 정의하고,}$$

여기서, $\hat{E}_t[\cdot]$ 는 유한한 표본 배치에 대한 기대 연산자, π_{θ} 는 확률적 정책, $\hat{A}_t$ 는 디스카운트된 보상 합계와 기준 추정치로 정의되며, a_t 와 s_t 는 시간 스텝 t 의 행동과 상태인 것을 특징으로 하는 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치.

청구항 9

제 5 항에 있어서, 강화학습 라이브러리 환경 구축부는,

훈련 과정 중 성능 저하를 방지하기 위하여 대리 손실 함수를 채택하여 정책 업데이트를 생성하는 PPO(Proximal policy optimization)를 적용하고,

$$\text{대리 객체}(J^{CPI}) \text{는} \quad J^{CPI}(\theta) = \hat{E}_t\left[\frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}\hat{A}_t\right] = \hat{E}_t[r_t(\theta)\hat{A}_t] \quad \text{으로 정의되고,}$$

$\pi_{\theta_{old}}$ 는 업데이트 전 정책 매개 변수, π_{θ} 는 업데이트 후 정책 매개 변수, $r_t(\theta)$ 는 확률비인 것을 특징으로 하

는 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치.

청구항 10

제 9 항에 있어서, 연속 행동의 경우 PPO의 정책 출력은 각 행동에 대한 가우스 분포의 매개 변수이고, 적응형 KL 패널티를 가진 PPO는 미니 배치(minibatch) 확률적 경사 하강(SGD)을 사용하여 KL 패널티 목표를 최적화하는 데 사용되고,

$$\begin{aligned} & \underset{\theta}{\text{maximize}} \hat{E}_t[r_t(\theta)\hat{A}_t] - \beta \hat{E}_t[KL[\pi_{\theta_{old}}(a_t|s_t), \pi_{\theta}(a_t|s_t)]] \\ & \text{Subject to } \hat{E}_t[KL[\pi_{\theta_{old}}(a_t|s_t), \pi_{\theta}(a_t|s_t)]] \leq \delta \end{aligned}$$

여기서, β 는 매 정책 업데이트 후 업데이트되는 가중 조절 계수(weight control coefficient)인 것을 특징으로 하는 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치.

청구항 11

제 10 항에 있어서, 현재 KL 차이가 목표 KL 편차보다 클 경우 증가되고, 현재 KL 발산이 목표 KL 발산보다 작으면 감소되고,

PPO 알고리즘에서 먼저 현재 정책이 환경과 상호 작용하여 에피소드 시퀀스를 생성하고, 어드밴티지 함수(advantage function)는 상태 값에 대한 기준 추정치를 사용하여 추정되어 모든 경험을 수집하고 정책 네트워크를 통해 경사 하강 알고리즘을 실행하는 것을 특징으로 하는 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치.

청구항 12

SUMO(Simulation of Urban Mobility)를 활용하여 시뮬레이션 환경을 구축하고, 자율 주행 차량의 속도, 위치, 센서에서 얻어진 데이터를 FLOW 적용부로 전달하는 SUMO 시뮬레이션 실행 단계;

SUMO와 연동할 수 있는 강화학습 플랫폼 FLOW 환경에서 시뮬레이션 환경을 구축하고, 강화학습을 적용하지 않은 운전 행태 도출, 차량 제어 및 시뮬레이션 상태 업데이트를 하여 상태 및 보상 정보를 강화학습 라이브러리 환경 구축부로 전달하는 FLOW 적용 단계;

SUMO와 연동할 수 있는 강화학습 플랫폼 FLOW를 활용하여 멀티 에이전트 심층강화학습(Multi agent Deep Reinforcement Learning)으로 통행 제어를 최적화하는 강화학습 라이브러리 환경 구축 단계;를 포함하는 것을 특징으로 하는 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 방법.

청구항 13

제 12 항에 있어서, SUMO 시뮬레이션 실행 단계는,

SUMO(Simulation of Urban Mobility)를 활용하여 시뮬레이션 환경을 구축하여 SUMO 시뮬레이션을 수행하는 SUMO 시뮬레이션 단계와,

배기가스, 속도 및 위치값 파일을 생성하여 자율 주행 차량의 속도, 위치, 센서에서 얻어진 데이터를 FLOW 적용부로 전달하는 결과 파일생성 단계를 포함하는 것을 특징으로 하는 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 방법.

청구항 14

제 12 항에 있어서, FLOW 적용 단계는,

인간운전자 정의, 심층 강화학습 입력값 설정 및 차량의 속도, 가속도, 출발점 등 시뮬레이션 환경 설정을 하는 시뮬레이션 초기화 단계와,

FLOW 환경 구축을 하여 상태(state)를 강화학습 라이브러리로 전달하는 FLOW 환경 구축 단계와,

강화학습을 적용하지 않은 운전 행태 도출을 하는 운행행태 도출 단계와,

차량 제어를 하고 제어 정보를 SUMO 시뮬레이션 실행부로 전달하는 차량 제어 단계와,

SUMO 시뮬레이션 실행부로부터 시뮬레이션 상태를 받아 시뮬레이션 상태 업데이트를 하여 상태 및 보상 정보를 강화학습 라이브러리 환경 구축부로 전달하는 업데이트 단계를 포함하는 것을 특징으로 하는 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 방법.

청구항 15

제 12 항에 있어서, 강화학습 라이브러리 환경 구축 단계는,

강화학습 라이브러리가 FLOW 적용부로부터 상태(state)를 전달받는 단계와,

학습할 데이터를 샘플링하는 데이터 샘플링 단계와,

운전 행태(정책) 훈련을 하는 정책 훈련 단계와,

훈련 결과를 평가하고 학습된 행태를 FLOW 적용부로 전달하는 훈련 결과 평가 단계와,

자율주행차량이 관찰할 수 있는 범위 내의 정보를 통하여 학습하는 방법으로 강화학습과 마르코프 의사결정 모델 사용(Partial Observability MDP, POMDP)을 적용하여 행동에 대한 강화학습의 보상을 최대화할 수 있도록 하는 정책 최적화 단계와,

FLOW 적용부로부터 자율 주행 차량의 상태를 받아 정책 업데이트 및 저장을 하는 정책 업데이트 저장 단계와,

업데이트된 정책이 학습 루프 조건을 만족하는지 판단하는 학습 루프조건 판단 단계를 포함하는 것을 특징으로 하는 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 방법.

발명의 설명

기술 분야

[0001] 본 발명은 다수의 자율주행 차량 운행 제어에 관한 것으로, 구체적으로 군집된 자율주행차량과 인간운전자 차량이 혼재되어 있는 혼잡 교통류 상황에서 자율주행차량 군집주행 학습으로 비신호 교차로 통행을 개선하고 안전성을 확보할 수 있도록 한 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치 및 방법에 관한 것이다.

배경 기술

[0002] 자율 주행 차량(Autonomous Vehicle)은 카메라 또는 전방물체 감지센서를 이용하여 차선을 인식하고 자동 조향을 행하는 기술이 탑재된 차량이다. 자율 주행 차량은 카메라의 이미지 프로세싱 또는 전방물체 감지센싱을 기반으로 차선 폭, 차선상의 차량의 횡방향 위치, 양측 차선까지의 거리 및 차선의 형태, 도로의 곡률 반경이 측정되며, 이와 같이 얻어진 차량의 위치와 도로의 정보를 사용하여 차량의 주행 궤적을 추정하고, 추정된 주행 궤적을 따라 차선을 변경한다.

[0003] 자율 주행 차량(Autonomous Vehicle)은 차량 전방에 장착된 카메라 또는 전방물체 감지센서에서 검출되는 선행 차량의 위치 및 거리를 통하여 차량의 쓰로틀밸브, 브레이크 및 변속기를 자동 제어하여 적절한 가속을 수행함으로써, 선행차량과 적정거리를 유지하도록 할 수도 있다.

[0004] 그러나 이와 같은 자율 주행 차량(Autonomous Vehicle)이 교차로를 통과하는 경우에는 신호등의 교통신호에 따라 정차 후 출발시 선행 차량의 움직임을 감지한 다음 출발하므로 차량들 간의 출발이 지체되어 교차로에서 정체가 발생할 수 있다.

[0005] 특히, 자율주행 차량과 같이 센서로부터 입력되는 정보를 이용하여 주행 환경을 파악하는 경우 비신호 교차로에서의 주행은 일반적인 도로에서의 주행보다 훨씬 어려운 과제가 된다.

[0006] 한편, 무선 통신 기술의 발전으로 인하여 IoT 관련 연구가 활발히 진행되고 있으며, 그와 같이 주목 받고 있는 것이 IoV(Internet of Vehicles)이다. 차량 사이의 통신을 위해 각 차량이 노드 역할을 수행하는 무선 네트워크

인 Vehicular Ad-hoc Network (VANET)은 Mobile Ad-hoc Network (MANET)의 한 형태이다.

- [0007] Simulation of Urban MObility(SUMO)는 도로 상에서의 교통 네트워크를 시뮬레이션 할 수 있도록 디자인되어 있는 오픈 소스이다.
- [0008] SUMO를 이용하여 도로 위에서 차량 간의 움직임을 파악함으로써 교통의 흐름을 예측할 수 있다.
- [0009] 이와 같은 기술들을 통하여 자율주행 차량이 주행 환경을 파악하여 비신호 교차로에서의 효율적인 주행을 위한 연구들이 이루어지고 있으나, 혼합 교통류 상황(자율주행차량과 인간운전자의 혼재)에서 자율주행차량 군집주행에 따른 비신호 교차로 통행에서는 아직도 해결하여야 하는 과제가 많다.
- [0010] 따라서, 자율주행차량 군집주행에 따른 비신호 교차로 통행 개선 및 안전성 확보를 위한 새로운 기술의 개발이 요구되고 있다.

선행기술문헌

특허문헌

- [0011] (특허문헌 0001) 대한민국 공개특허 제10-2020-0071406호
- (특허문헌 0002) 대한민국 공개특허 제10-2020-0058613호
- (특허문헌 0003) 대한민국 공개특허 제10-2018-0065196호

발명의 내용

해결하려는 과제

- [0012] 본 발명은 종래 기술의 자율주행 차량 운행 제어 기술의 문제점을 해결하기 위한 것으로, 군집된 자율주행차량과 인간운전자 차량이 혼재되어 있는 혼합 교통류 상황에서 자율주행차량 군집주행 학습으로 비신호 교차로 통행을 개선하고 안전성을 확보할 수 있도록 한 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치 및 방법을 제공하는데 그 목적이 있다.
- [0013] 본 발명은 실제 상황과 같이 자율주행차량이 관찰할 수 있는 범위 내의 정보를 통하여 학습하는 방법으로 강화학습과 마르코프 의사결정 모델 사용(Partial Observability MDP, POMDP)을 적용하여 행동에 대한 강화학습의 보상을 최대화할 수 있도록 한 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치 및 방법을 제공하는데 그 목적이 있다.
- [0014] 본 발명은 군집된 자율주행차량과 인간운전자 차량이 혼재되어 있는 혼합 교통류 상황에서 자율주행차량 운행 행태를 학습하는 방법으로 인공신경망에 학습을 최적화하기 위한 알고리즘인 PPO 적용으로 통행 제어를 최적화할 수 있도록 한 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치 및 방법을 제공하는데 그 목적이 있다.
- [0015] 본 발명은 SUMO(Simulation of Urban MObility)를 활용하여 실험환경을 구축하고 ACC(Adaptive Cruise Control) 시스템으로 인간운전자 정의를 하여, SUMO와 연동할 수 있는 강화학습 플랫폼 FLOW를 활용하여 멀티 에이전트 심층강화학습(Multi agent Deep Reinforcement Learning)으로 통행 제어를 최적화할 수 있도록 한 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치 및 방법을 제공하는데 그 목적이 있다.
- [0016] 본 발명은 강화학습 파라미터 조정 및 자율주행차량 점유율별 운행 최적화 및 검증으로 비신호 교차로에서 완전 인간운전자환경에 비해 완전 자율주행차량 환경에서 평균 통행 속도를 향상시킬 수 있도록 한 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치 및 방법을 제공하는데 그 목적이 있다.
- [0017] 본 발명은 부분관찰 마르코프 의사결정과정(POMDP)에 따라 시뮬레이션 환경 내의 자율주행차량의 행태를 결정하며 평균속도를 보상으로 학습하고, 멀티 에이전트 심층강화학습을 하기 위해 PPO(Proximal Policy Optimization) 알고리즘을 적용하여 행동 결정을 최적화할 수 있도록 한 자율주행 차량 군집 운행을 위한 비신

호 교차로에서의 강화학습기반 통행 개선을 위한 장치 및 방법을 제공하는데 그 목적이 있다.

[0018] 본 발명은 시뮬레이션 환경에서 실제 자율주행 환경을 모사하기 위해 학습과 행동 결정의 근거를 시뮬레이션의 모든 환경이 아닌 자율주행차량 센서를 통하여 얻어진 데이터(부분만 관찰)를 기반으로 하여 행동을 결정하고 행동에 대해 강화학습의 보상을 최대화할 수 있도록 한 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치 및 방법을 제공하는데 그 목적이 있다.

[0019] 본 발명의 다른 목적들은 이상에서 언급한 목적으로 제한되지 않으며, 언급되지 않은 또 다른 목적들은 아래의 기재로부터 당업자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0020] 상기와 같은 목적을 달성하기 위한 본 발명에 따른 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치는 SUMO(Simulation of Urban Mobility)를 활용하여 시뮬레이션 환경을 구축하고, 자율 주행 차량의 속도, 위치, 센서에서 얻어진 데이터를 FLOW 적용부로 전달하는 SUMO 시뮬레이션 실행부; SUMO와 연동할 수 있는 강화학습 플랫폼 FLOW 환경에서 시뮬레이션 환경을 구축하고, 강화학습을 적용하지 않은 운전 행태 도출, 차량 제어 및 시뮬레이션 상태 업데이트를 하여 상태 및 보상 정보를 강화학습 라이브러리 환경 구축부로 전달하는 FLOW 적용부; SUMO와 연동할 수 있는 강화학습 플랫폼 FLOW를 활용하여 멀티 에이전트 심층강화학습(Multi agent Deep Reinforcement Learning)으로 통행 제어를 최적화하는 강화학습 라이브러리 환경 구축부;를 포함하는 것을 특징으로 한다.

[0021] 다른 목적을 달성하기 위한 본 발명에 따른 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 방법은 SUMO(Simulation of Urban Mobility)를 활용하여 시뮬레이션 환경을 구축하고, 자율주행 차량의 속도, 위치, 센서에서 얻어진 데이터를 FLOW 적용부로 전달하는 SUMO 시뮬레이션 실행 단계; SUMO와 연동할 수 있는 강화학습 플랫폼 FLOW 환경에서 시뮬레이션 환경을 구축하고, 강화학습을 적용하지 않은 운전 행태 도출, 차량 제어 및 시뮬레이션 상태 업데이트를 하여 상태 및 보상 정보를 강화학습 라이브러리 환경 구축부로 전달하는 FLOW 적용 단계; SUMO와 연동할 수 있는 강화학습 플랫폼 FLOW를 활용하여 멀티 에이전트 심층강화학습(Multi agent Deep Reinforcement Learning)으로 통행 제어를 최적화하는 강화학습 라이브러리 환경 구축 단계;를 포함하는 것을 특징으로 한다.

발명의 효과

[0022] 이상에서 설명한 바와 같은 본 발명에 따른 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치 및 방법은 다음과 같은 효과가 있다.

[0023] 첫째, 군집된 자율주행차량과 인간운전자 차량이 혼재되어 있는 혼합 교통류 상황에서 자율주행차량 군집주행 학습으로 비신호 교차로 통행을 개선하고 안전성을 확보할 수 있도록 한다.

[0024] 둘째, 실제 상황과 같이 자율주행차량이 관찰할 수 있는 범위 내의 정보를 통하여 학습하는 방법으로 강화학습과 마르코프 의사결정 모델 사용(Partial Observability MDP, POMDP)을 적용하여 행동에 대한 강화학습의 보상을 최대화할 수 있도록 한다.

[0025] 셋째, 군집된 자율주행차량과 인간운전자 차량이 혼재되어 있는 혼합 교통류 상황에서 자율주행차량 운행 행태를 학습하는 방법으로 인공지능망에 학습을 최적화하기 위한 알고리즘인 PPO 적용으로 통행 제어를 최적화할 수 있다.

[0026] 넷째, SUMO(Simulation of Urban Mobility)를 활용하여 실험환경을 구축하고 ACC(Adaptive Cruise Control) 시스템으로 인간운전자 정의를 하여, SUMO와 연동할 수 있는 강화학습 플랫폼 FLOW를 활용하여 멀티 에이전트 심층강화학습(Multi agent Deep Reinforcement Learning)으로 통행 제어를 최적화할 수 있도록 한다.

[0027] 다섯째, 강화학습 파라미터 조정 및 자율주행차량 점유율별 운행 최적화 및 검증으로 비신호 교차로에서 완전 인간운전자환경에 비해 완전 자율주행차량 환경에서 평균 통행 속도를 향상시킬 수 있도록 한다.

[0028] 여섯째, 부분관찰 마르코프 의사결정과정(POMDP)에 따라 시뮬레이션 환경 내의 자율주행차량의 행태를 결정하며 평균속도를 보상으로 학습하고, 멀티 에이전트 심층강화학습을 하기 위해 PPO(Proximal Policy Optimization) 알고리즘을 적용하여 행동 결정을 최적화할 수 있도록 한다.

[0029] 일곱째, 시뮬레이션 환경에서 실제 자율주행 환경을 모사하기 위해 학습과 행동 결정의 근거를 시뮬레이션의 모

든 환경이 아닌 자율주행차량 센서를 통하여 얻어진 데이터(부분만 관찰)를 기반으로 하여 행동을 결정하고 행동에 대해 강화학습의 보상을 최대화할 수 있도록 한다.

도면의 간단한 설명

[0030]

도 1은 본 발명에 따른 비신호 교차로에서의 심층 강화 학습 아키텍처를 나타낸 구성도

도 2는 적응형 KL 페널티 알고리즘을 사용한 PPO 알고리즘

도 3은 본 발명에 따른 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치 구성도

도 4는 본 발명에 따른 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 방법을 나타낸 동작 흐름도

도 5는 비신호 교차로에서의 일반적인 SUMO 시뮬레이터 구성도

도 6은 일반적인 관측 영역(Observation Space)의 일 예를 나타낸 구성도

도 7은 비신호 교차로에서의 선도 자율 주행 차량 실험 특성을 나타낸 구성도

도 8은 비신호화된 교차로에서의 실험 비교 구성도

도 9는 AV 점유율을 기반으로 한 200회 이상의 평균 보상 곡선 그래프

도 10은 비신호화된 교차로에서 AV 점유율을 통한 시공간 역학 특성 그래프

도 11은 SUMO 시뮬레이션 환경에서 평균속도, 평균 지체시간, 평균연료 소모량, 평균 배기가스 값 도출 특성 그래프

발명을 실시하기 위한 구체적인 내용

[0031]

이하, 본 발명에 따른 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치 및 방법의 바람직한 실시 예에 관하여 상세히 설명하면 다음과 같다.

[0032]

본 발명에 따른 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치 및 방법의 특징 및 이점들은 이하에서의 각 실시 예에 대한 상세한 설명을 통해 명백해질 것이다.

[0033]

도 1은 본 발명에 따른 비신호 교차로에서의 심층 강화 학습 아키텍처를 나타낸 구성도이고, 도 2는 적응형 KL 페널티 알고리즘을 사용한 PPO 알고리즘이다.

[0034]

본 발명에 따른 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치 및 방법은 군집된 자율주행차량과 인간운전자 차량이 혼재되어 있는 혼합 교통류 상황에서 자율주행차량 군집주행 학습으로 비신호 교차로 통행을 개선하고 안전성을 확보할 수 있도록 한 것이다.

[0035]

이를 위하여, 본 발명은 실제 상황과 같이 자율주행차량이 관찰할 수 있는 범위 내의 정보를 통하여 학습하는 방법으로 강화학습과 마르코프 의사결정 모델 사용(Partial Observability MDP, POMDP)을 적용하여 행동에 대한 강화학습의 보상을 최대화할 수 있도록 하는 구성을 포함할 수 있다.

[0036]

본 발명은 군집된 자율주행차량과 인간운전자 차량이 혼재되어 있는 혼합 교통류 상황에서 자율주행차량 운행 행태를 학습하는 방법으로 인공신경망에 학습을 최적화하기 위한 알고리즘인 PPO 적용으로 통행 제어를 최적화할 수 있도록 하는 구성을 포함할 수 있다.

[0037]

본 발명은 SUMO(Simulation of Urban MObility)를 활용하여 실험환경을 구축하고 ACC(Adaptive Cruise Control) 시스템으로 인간운전자 정의를 하여, SUMO와 연동할 수 있는 강화학습 플랫폼 FLOW를 활용하여 멀티 에이전트 심층강화학습(Multi agent Deep Reinforcement Learning)으로 통행 제어를 최적화할 수 있도록 하는 구성을 포함할 수 있다.

[0038]

본 발명은 강화학습 파라미터 조정 및 자율주행차량 점유율별 운행 최적화 및 검증 구성, 부분관찰 마르코프 의사결정과정(POMDP)에 따라 시뮬레이션 환경 내의 자율주행차량의 행태를 결정하며 평균속도를 보상으로 학습하는 구성, 멀티 에이전트 심층강화학습을 하기 위해 PPO(Proximal Policy Optimization) 알고리즘을 적용하여 행동 결정을 최적화하는 구성을 포함할 수 있다.

- [0039] 강화 학습(RL)은 기계 학습의 하위 영역이며 에이전트가 환경과 상호 작용하고 누적 보상을 극대화하는 조치를 학습하는 것이다.
- [0040] RL 알고리즘의 전형적인 형태는 마르코프 결정 과정(MDP)으로, 전체 관측치 집합이 주어진 적절한 동작을 결정하는 데 사용되는 강력한 프레임워크이다.
- [0041] MDP는 튜플 $(S, A, P, R, \rho_0, \gamma, T)$ 이며, 여기서 S 와 A 는 각각 참가자의 상태와 행동이다. $P(S', S, a)$ 는 전이 확률을 정의하며, $R(a, S)$ 은 선택된 작용에 따라 보상을 정의하며, ρ_0 은 초기 상태 분포를 정의하며, γ 는 0에서 1까지의 할인 계수(discount factor)를 정의하며, T 는 시간 범위를 나타낸다.
- [0042] 그러나 자동차 차량은 부정확성, 의도 및 센서 노이즈를 포함하는 불확실한 환경에서 기동한다. 이 문제를 해결하기 위해 관측치 집합을 정의하는 O 와 관측 함수인 Z 라는 두 가지 요소를 더 사용하는 부분 관측 가능한 MDP(POMDP)가 제안되었다.
- [0043] RL의 객관적 학습 에이전트는 정책 π 를 최적화하여 몇 가지 타임 스텝에 걸쳐 예상 누적 할인 보상을 극대화한다.
- [0044] 심층 신경 네트워크(DNN)는 여러 개의 숨겨진 표현 계층으로 인해 형상 추출을 자동으로 수행할 수 있는 기능을 가지고 있다. 연속 제어기의 경우, 인공 신경 네트워크(ANN)는 복잡한 기능을 나타내기 위해 여러 개의 숨겨진 레이어를 사용하는 일반적으로 사용되는 방법이다.
- [0045] 이 작업에서는 MLP를 적용하여 입력 세트(상태 및 관찰)에서 출력 세트(정책)를 생성한다. 또한, DNN의 성능을 향상시키기 위해 경사 하강 최적화 방법에 기초한 PPO를 적용한다.
- [0046] MLP와 RL을 융합하는 제안된 심층 RL 프레임워크는 비신호화된 교차점에서 AV의 효과를 고려하도록 설계되었다.
- [0047] 첫째, SUMO 시뮬레이터는 하나의 시뮬레이션 단계를 실행한다.
- [0048] 둘째, Flow 프레임워크는 SUMO 시뮬레이터의 상태에 대한 정보를 RL 라이브러리에 보낸다. 그런 다음, RL 라이브러리(RLlib)는 MLP를 통해 SUMO 시뮬레이터의 상태에 따라 적절한 조치를 계산한다. MLP 정책은 트래픽 데이터를 기반으로 RL 알고리즘에 대한 누적 보상을 최대화하기 위해 적용된다.
- [0049] 마지막으로 시뮬레이션은 RL 프로세스를 재설정하고 반복한다.
- [0050] 도 1은 비신호화 교차로에서 심층 강화 학습 아키텍처를 나타낸 것이다.
- [0051] 중요한 것은, '정책'은 환경에서의 인식과 행동 사이의 의사소통의 청사진을 가리킨다. 즉, 정책은 트래픽 시뮬레이션의 컨트롤러와 유사하다.
- [0052] 이 작업에서 컨트롤러는 여러 개의 숨겨진 계층이 있는 MLP 정책이다.
- [0053] 컨트롤러의 매개변수는 MLP 정책을 사용하여 반복적으로 업데이트되어 SUMO 시뮬레이터에서 샘플링된 트래픽 데이터를 기반으로 누적 보상을 최대화한다.
- [0054] 에이전트의 주요 목표는 다음과 같이 확률적 정책을 최적화하는 방법을 학습하는 것이다.

수학식 1

$$\theta^* := \operatorname{argmax}_{\theta} \eta(\pi_{\theta})$$

[0055]

[0056] 여기서, $\eta(\pi_{\theta})$ 는 할인 계수(γ^i)와 보상(r)에 의해 계산되는 예상 누적 할인 보상이다.

수학식 2

$$\eta(\pi_0) = \sum_{i=0}^T \gamma_i r_i$$

종방향 역학 모델(Longitudinal Dynamic Models)을 설명하면 다음과 같다.

기본적인 차량 역학은 차량 자체와 전방 차량의 관찰에 기초하여 수동 작동 차량의 세로 방향 역학을 설명하는 차량 추종 모델에 의해 정의될 수 있다.

표준 차량 추종 모델은 다음과 같다.

수학식 3

$$a_i = f(h_i, \dot{h}_i, v_i)$$

여기서, a_i 는 차량 i의 가속도이고, $f()$ 는 비선형 함수이며, v_i , $\dot{h}_i$ 및 h_i 는 각각 차량 i의 속도, 상대 속도 및 방향이다.

본 발명에서는 운전자 행동을 묘사할 수 있는 능력으로 인해 인간 구동 차량의 세로 방향 제어를 위해 ACC 시스템의 일종인 IDM을 적용한다.

IDM은 일반적으로 사용되는 자동차 추종 모델이다.

IDM의 가속도 명령에서 비신호화된 교차로 환경에서의 차량 속도와 선두 차량의 식별(ID) 및 선두 차량의 진행(headway of the leading vehicle)은 "get" 방법으로 얻을 수 있도록 설정할 수 있다.

차량의 가속도는 다음과 같이 계산한다.

수학식 4

$$a_{IDM} = a \left[1 - \left(\frac{v}{v_0} \right)^\delta - \left(\frac{s^*(v, \Delta v)}{s} \right)^2 \right]$$

여기서, a_{IDM} 은 차량의 가속이고, v_0 는 원하는 속도이며, δ 는 가속도지수, s 는 차량의 앞길(앞차와의 거리)이며, $s^*(v, \Delta v)$ 는 원하는 방향을 나타내며, 다음과 같이 표현된다.

수학식 5

$$s^*(v, \Delta v) = s_0 + \max \left(0, vT + \frac{v\Delta v}{2\sqrt{ab}} \right)$$

여기서, s_0 는 최소 갭을, T 는 a시간 갭을, Δv 는 선두 차량과 비교한 속도 차이(현재 속도 - 선두 속도), a 는 가속 구간, b 는 편안한 감속을 나타낸다.

[0071] 도시 교통에 대한 IDM 컨트롤러의 대표적인 매개변수는 표 1에서와 같다.

표 1

Parameters	Value
Desired speed (m/s)	15
Time gap (s)	1.0
Minimum gap (m)	2.0
Acceleration exponent	4.0
Acceleration (m/s ²)	1.0
Comfortable acceleration (m/s ²)	1.5

[0072]

[0073] 정책 최적화(Policy Optimization)를 설명하면 다음과 같다.

[0074] 정책 경사 방법(Policy gradient methods)은 동작 값이나 상태 값 함수가 아닌 경사 강하 알고리즘을 사용하여 매개 변수화된 정책 함수의 추정기를 계산하려고 한다.

[0075] 따라서 비선형 근사 및 부분 관측으로 인해 추정 함수에 발생하는 수렴 문제를 피한다.

[0076] 본 발명은 비신호화된 교차로의 시뮬레이션에서 제어 정책을 직접 최적화하기 위해 MLP 정책을 적용한다. 정책 행동($\log \pi_{\theta}$)의 확률에 대한 기대치와 시간 스텝 t 에서의 어드밴티지 함수(advantage function)의 추정치에 기초하는 정책 경사법은 다음과 같이 표현된다.

수학식 6

$$\hat{g} = \hat{E}_t[\nabla_{\theta} \log \pi_{\theta}(a_t|s_t)\hat{A}_t]$$

[0077]

[0078] 여기서, $\hat{E}_t[\cdot]$ 는 유한한 표본 배치에 대한 기대 연산자이며, π_{θ} 는 확률적 정책을 나타내며, $\hat{A}_t$ 는 디스카운트된 보상 합계와 기준 추정치로 정의되며, a_t 와 s_t 는 시간 스텝 t 의 행동과 상태를 각각 나타낸다.

[0079] 술만(Schulman) 등에 의해 제안된 PPO는 RLlib 라이브러리에서 제공하는 간단한 TRPO이다.

[0080] 즉, PPO의 목표는 TRPO와 동일하며, TRPO는 신뢰 지역 제약 조건을 사용하여 새 정책이 이전 정책에서 너무 멀리 있지 않도록 정책을 업데이트하도록 강제한다.

[0081] PPO에는 적응형 쿨백-라이블러(adaptive Kullback-Leibler;KL) 페널티와 클리핑 목표(clipped objective)의 두 가지 유형이 있다.

[0082] PPO는 대리 손실 함수를 채택하여 정책 업데이트를 생성한다. 이 프로세스는 훈련 과정 중 성능 저하를 방지한다.

[0083] 대리 객체(J^{CPI})는 다음과 같이 설명된다.

수학식 7

$$J^{CPI}(\theta) = \hat{E}_t\left[\frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}\hat{A}_t\right] = \hat{E}_t[r_t(\theta)\hat{A}_t]$$

[0084]

[0085] 여기서, $\pi_{\theta_{old}}$ 는 업데이트 전 정책 매개 변수, π_{θ} 는 업데이트 후 정책 매개 변수, $r_t(\theta)$ 는 확률비를 나타낸다.

[0086] 연속 행동의 경우 PPO의 정책 출력은 각 행동에 대한 가우스 분포의 매개 변수이다.

[0087] 그런 다음 정책은 이러한 분포를 기반으로 연속 출력을 생성한다.

[0088] 본 발명에서 적응형 KL 패널티를 가진 PPO는 다음과 같이 미니 배치(minibatch) 확률적 경사 하강(SGD)을 사용하여 KL 패널티 목표를 최적화하는 데 사용된다.

수학식 8

$$\text{maximize}_{\theta} \hat{E}_t[r_t(\theta)\hat{A}_t] - \beta \hat{E}_t[KL[\pi_{\theta_{old}}(a_t|s_t), \pi_{\theta}(a_t|s_t)]]$$

수학식 9

$$\text{Subject to } \hat{E}_t[KL[\pi_{\theta_{old}}(a_t|s_t), \pi_{\theta}(a_t|s_t)]] \leq \delta$$

[0091] 여기서, β 는 매 정책 업데이트 후 업데이트되는 가중 조절 계수(weight control coefficient)이다.

[0092] 현재 KL 차이가 목표 KL 편차보다 클 경우 증가되고, 현재 KL 발산이 목표 KL 발산보다 작으면 감소한다.

[0093] PPO 알고리즘에서는 먼저 현재 정책이 환경과 상호 작용하여 에피소드 시퀀스를 생성한다. 다음으로, 어드밴티지 함수(advantage function)는 상태 값에 대한 기준 추정치를 사용하여 추정된다.

[0094] 마지막으로, 모든 경험을 수집하고 정책 네트워크를 통해 경사 하강 알고리즘을 실행한다. 적응형 KL 패널티 알고리즘의 전체 PPO는 도 2의 알고리즘 1의 유사 코드로 표시된다.

[0095] 도 3은 본 발명에 따른 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치 구성도이다.

[0096] 본 발명에 따른 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치는 SUMO(Simulation of Urban Mobility)를 활용하여 시뮬레이션 환경을 구축하고, 자율 주행 차량의 속도, 위치, 센서에서 얻어진 데이터를 FLOW 적용부(200)로 전달하는 SUMO 시뮬레이션 실행부(100)와, SUMO와 연동할 수 있는 강화학습 플랫폼 FLOW 환경에서 시뮬레이션 환경을 구축하고, 강화학습을 적용하지 않은 운전 행태 도출, 차량 제어 및 시뮬레이션 상태 업데이트를 하여 상태 및 보상 정보를 강화학습 라이브러리 환경 구축부(300)로 전달하는 FLOW 적용부(200)와, SUMO와 연동할 수 있는 강화학습 플랫폼 FLOW를 활용하여 멀티 에이전트 심층강화학습(Multi agent Deep Reinforcement Learning)으로 통행 제어를 최적화하는 강화학습 라이브러리 환경 구축부(300)를 포함한다.

[0097] 여기서, SUMO 시뮬레이션 실행부(100)는 SUMO(Simulation of Urban Mobility)를 활용하여 시뮬레이션 환경을 구축하여 SUMO 시뮬레이션을 수행하는 SUMO 시뮬레이션부(10)와, 배기가스, 속도 및 위치값 파일을 생성하여 자율 주행 차량의 속도, 위치, 센서에서 얻어진 데이터를 FLOW 적용부(200)로 전달하는 결과 파일생성부(11)를 포함한다.

[0098] 그리고 FLOW 적용부(200)는 인간운전자 정의, 심층 강화학습 입력값 설정 및 차량의 속도, 가속도, 출발점 등 시뮬레이션 환경 설정을 하는 시뮬레이션 초기화부(20)와, FLOW 환경 구축을 하여 상태(state)를 강화학습 라이브러리로 전달하는 FLOW 환경 구축부(21)와, 강화학습을 적용하지 않은 운전 행태 도출을 하는 운행행태 도출부(22)와, 차량 제어를 하고 제어 정보를 SUMO 시뮬레이션 실행부(100)로 전달하는 차량 제어 모듈(23)과, SUMO 시뮬레이션 실행부(100)로부터 시뮬레이션 상태를 받아 시뮬레이션 상태 업데이트를 하여 상태 및 보상 정보를 강화학습 라이브러리 환경 구축부(300)로 전달하는 업데이트부(24)를 포함한다.

- [0099] 그리고 강화학습 라이브러리 환경 구축부(300)는 FLOW 적용부(200)로부터 상태(state)를 전달받는 강화학습 라이브러리(31)와, 학습할 데이터를 샘플링하는 데이터 샘플링부(32)와, 운전 행태(정책) 훈련을 하는 정책 훈련부(33)와, 훈련 결과를 평가하고 학습된 행태(주행방법)를 FLOW 적용부(200)로 전달하는 훈련 결과 평가부(34)와, 자율주행차량이 관찰할 수 있는 범위 내의 정보를 통하여 학습하는 방법으로 강화학습과 마르코프 의사결정 모델 사용(Partial Observability MDP, POMDP)을 적용하여 행동에 대한 강화학습의 보상을 최대화할 수 있도록 하는 정책 최적화부(35)와, FLOW 적용부(200)로부터 자율 주행 차량의 상태를 받아 정책 업데이트 및 저장을 하는 정책 업데이트 저장부(36)와, 업데이트된 정책이 학습 루프 조건을 만족하는지 판단하는 학습 루프조건 판단부(37)를 포함한다.
- [0100] 도 4는 본 발명에 따른 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 방법을 나타낸 동작 흐름도이다.
- [0101] 본 발명에 따른 자율주행 차량 군집 운행을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 방법은 SUMO(Simulation of Urban Mobility)를 활용하여 시뮬레이션 환경을 구축하고, 자율 주행 차량의 속도, 위치, 센서에서 얻어진 데이터를 FLOW 적용부(200)로 전달하는 SUMO 시뮬레이션 실행 단계와, SUMO와 연동할 수 있는 강화학습 플랫폼 FLOW 환경에서 시뮬레이션 환경을 구축하고, 강화학습을 적용하지 않은 운전 행태 도출, 차량 제어 및 시뮬레이션 상태 업데이트를 하여 상태 및 보상 정보를 강화학습 라이브러리 환경 구축부(300)로 전달하는 FLOW 적용 단계와, SUMO와 연동할 수 있는 강화학습 플랫폼 FLOW를 활용하여 멀티 에이전트 심층강화학습(Multi agent Deep Reinforcement Learning)으로 통행 제어를 최적화하는 강화학습 라이브러리 환경 구축 단계를 포함한다.
- [0102] 여기서, SUMO 시뮬레이션 실행 단계는 SUMO(Simulation of Urban Mobility)를 활용하여 시뮬레이션 환경을 구축하여 SUMO 시뮬레이션을 수행하는 SUMO 시뮬레이션 단계(S409)와, 배기가스, 속도 및 위치값 파일을 생성하여 자율 주행 차량의 속도, 위치, 센서에서 얻어진 데이터를 FLOW 적용부(200)로 전달하는 결과 파일생성 단계(S410)를 포함한다.
- [0103] 그리고 FLOW 적용 단계는 인간운전자 정의, 심층 강화학습 입력값 설정 및 차량의 속도, 가속도, 출발점 등 시뮬레이션 환경 설정을 하는 시뮬레이션 초기화 단계(S401)와, FLOW 환경 구축을 하여 상태(state)를 강화학습 라이브러리로 전달하는 FLOW 환경 구축 단계(S402)와, 강화학습을 적용하지 않은 운전 행태 도출을 하는 운행행태 도출 단계(S403)와, 차량 제어를 하고 제어 정보를 SUMO 시뮬레이션 실행부(100)로 전달하는 차량 제어 단계(S408)와, SUMO 시뮬레이션 실행부(100)로부터 시뮬레이션 상태를 받아 시뮬레이션 상태 업데이트를 하여 상태 및 보상 정보를 강화학습 라이브러리 환경 구축부(300)로 전달하는 업데이트 단계(S411)를 포함한다.
- [0104] 그리고 강화학습 라이브러리 환경 구축 단계는 강화학습 라이브러리(31)가 FLOW 적용부(200)로부터 상태(state)를 전달받는 단계(S404)와, 학습할 데이터를 샘플링하는 데이터 샘플링 단계(S405)와, 운전 행태(정책) 훈련을 하는 정책 훈련 단계(S406)와, 훈련 결과를 평가하고 학습된 행태(주행방법)를 FLOW 적용부(200)로 전달하는 훈련 결과 평가 단계(S407)와, 자율주행차량이 관찰할 수 있는 범위 내의 정보를 통하여 학습하는 방법으로 강화학습과 마르코프 의사결정 모델 사용(Partial Observability MDP, POMDP)을 사용하여 행동에 대한 강화학습의 보상을 최대화할 수 있도록 하는 정책 최적화 단계(S412)와, FLOW 적용부(200)로부터 자율 주행 차량의 상태를 받아 정책 업데이트 및 저장을 하는 정책 업데이트 저장 단계(S413)와, 업데이트된 정책이 학습 루프 조건을 만족하는지 판단하는 학습 루프조건 판단 단계(S414)를 포함한다.
- [0105] 도 5는 비신호 교차로에서의 일반적인 SUMO 시뮬레이터 구성도이다.
- [0106] 독일 항공우주센터의 교통 시스템 연구소가 개발한 SUMO는 오픈소스 마이크로소프트 교통 시뮬레이터이다. SUMO는 신호등, 차량, 보행자 및 대중 교통과 함께 도시 규모의 교통 네트워크를 시뮬레이션할 수 있다. 또한 TraCI는 SUMO 시뮬레이터에 심층 RL을 적용하기 위해 SUMO를 Python에 연결할 수 있도록 한다.
- [0107] 비신호화된 교차로에서 일반적인 SUMO 시뮬레이터는 도 5에서와 같다.
- [0108] UC Berkeley에서 개발한 Flow는 심층 RL 알고리즘과 맞춤형 도로 네트워크 간의 인터페이스를 제공한다. 또한 Flow는 훈련 정책을 분석하고 검증할 수 있다.
- [0109] Flow의 장점은 심층 RL을 통해 자율 주행 차량의 제어기를 개선하기 위해 다양한 도로망을 쉽게 구현할 수 있는 능력을 포함한다. Flow에서 사용자 지정 환경은 다양한 시나리오에 대한 초기화된 시뮬레이션, 관찰 공간, 상태 공간, 작업 공간, 제어기 및 보상 기능을 포함한 주요 부분 집합 클래스를 생성하는 데 사용될 수 있다.

- [0110] 초기화된 시뮬레이션은 시작 에피소드에 대한 시뮬레이션 환경의 초기 설정을 나타낸다.
- [0111] 본 발명에서는 IDM 규칙과 심층 RL 프레임워크의 매개 변수뿐만 아니라 위치, 속도, 가속, 출발점, 궤적 및 차량 수를 설정한다.
- [0112] 특히, 모든 차량의 궤적은 특정 노드(네트워크의 포인트 위치), 특정 에지(노드를 함께 연결) 및 특정 경로(에지 차량이 통과하는 시퀀스)를 포함하여 SUMO 시뮬레이터에 의해 초기 시뮬레이션 프로세스에서 설정된다.
- [0113] 다음으로, 인간 운전 차량의 가속은 SUMO 시뮬레이터에 의해 제어되고 AV의 가속은 RLlib 라이브러리에 의해 제어된다.
- [0114] 도 6은 일반적인 관측 영역(Observation Space)의 일 예를 나타낸 구성도이다.
- [0115] 관측 공간은 AV 속도(자기 차량 속도), AV 위치(자기 차량 위치) 및 해당 선행 및 AV의 속도 및 범퍼 투 범퍼 헤드웨이와 같은 관측 가능한 형상의 수와 유형을 나타낸다.
- [0116] 관찰 가능한 출력이 상태 공간으로 공급되어 적절한 정책을 예측한다.
- [0117] 그리고 상태 공간(state space)은 AV의 위치 및 속도뿐만 아니라 이전 및 이후의 AV를 포함하여 관찰 공간을 기반으로 하는 자율 에이전트 및 주변 차량의 벡터를 나타낸다.
- [0118] 환경 내의 기능은 get_state 방법을 사용하여 추출되어 정책에 공급된다.
- [0119] 첫째, 비신호화된 교차로에서 모든 차량의 ID를 얻는다. 그런 다음 모든 차량의 위치와 속도를 파악하여 상태 공간을 생성한다.
- [0120] 중요한 것은 현재 위치가 미리 지정된 시작 지점을 기반으로 한다는 것이다.
- [0121] 상태 공간은 다음과 같이 정의된다.

수학식 10

$$S = \begin{pmatrix} x_0 \\ v_0 \\ v_l \\ d_l \\ v_f \\ d_f \end{pmatrix}$$

- [0122]
- [0123] 여기서, S는 특정 차량의 상태이고, x_0 은 AV의 해당 좌표이고, v_0 , v_l 및 v_f 는 각각 AV, 이전 AV 및 다음 AV의 해당 속도이고,
- [0124] d_l 와 d_f 는 각각 이전 AV와 다음 AV의 범퍼-대-범퍼 헤드웨이이다.
- [0125] 행동 공간(Action Space)은 OpenAI gym에서 제공하는 트래픽 환경에서 자율 에이전트의 행동을 나타낸다.
- [0126] 자동화 차량의 표준 행동은 가속이고, 행동 공간에서 행동의 범위는 최대 감속부터 최대 가속까지이다.
- [0127] apply_RL_ 행동 함수는 SUMO 시뮬레이터에서 특정 명령을 실제 행동으로 변환하기 위해 적용된다.
- [0128] 첫째, 비신호화된 교차로에서 모든 AV를 식별한다. 그런 다음 행동 명령은 기본 환경 방법을 사용하여 가속으로 변환된다.
- [0129] 컨트롤러는 사람이 운전하는 차량과 AV를 포함하여 행위자들의 행동을 통제한다. 공유 제어를 사용하여 단일 컨트롤러를 여러 행위자에 적용할 수 있다. 본 발명에서는 인간이 운전하는 차량은 플로우 프레임워크에 의해 제어되고, 자동화 차량은 RLlib 라이브러리에 의해 제어된다.
- [0130] 보상 함수(Reward Function)를 설명하면 다음과 같다.

- [0131] 트래픽 정체를 줄이기 위해서는 지연 시간, 대기열 길이를 줄임으로써 네트워크의 평균 속도를 최적화해야 한다. 따라서, 평균 속도는 현실에서 심층 RL 정책을 훈련하는 유망한 측정 기준이 된다.
- [0132] 보상 함수는 자율 에이전트가 정책을 최적화하는 방법을 정의한다.
- [0133] 본 발명에서 RL 에이전트의 목표는 비신호화된 교차로에서 차량 간 충돌을 억제하는 동시에 높은 평균 속도를 얻는다.
- [0134] 본 발명에서, L2 규범은 목표 속도(비신호 교차로에서 모든 차량의 원하는 속도)에 기초하여 비신호 교차로에서 주어진 차량 속도에 주어진 양의 거리를 추정하는 데 사용된다.
- [0135] 특히, 비신호화된 교차로에서 모든 차량의 현재 속도를 구한 다음 평균 속도를 보상으로 돌려주는 Get-speed 방법을 적용한다.
- [0136] 보상 함수는 수학적 식 11에서와 같이 표현된다.

수학적 식 11

[0137]
$$r_t := \max\left(\|v_{des} \cdot \mathbb{I}^k\|_2 - \|v_{des} - v\|_2, 0\right) / \|v_{des} \cdot \mathbb{I}^k\|_2$$

- [0138] 여기서, v_{des} 는 임의의 원하는 속도를 나타내고 $v \in R^k$ 는 비신호화된 교차로에서 모든 차량의 속도를 나타낸다.
- [0139] 본 발명에 따른 자율주행 차량 군집 운영을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치 및 방법을 이용한 시뮬레이션 환경 설정 및 결과를 설명하면 다음과 같다.
- [0140] 도 7은 비신호 교차로에서의 선도 자율 주행 차량 실험 특성을 나타낸 구성도이고, 도 8은 비신호화된 교차로에서의 실험 비교 구성도이다.
- [0141] 표 2는 비신호 교차로 시뮬레이션 설정값의 일 예를 나타낸 것이다.

표 2

시뮬레이션 설정 내용	설정값
학습 횟수	200
학습 1회의 학습 시간	6000
보상 조정값(PPO 조정값, Gamma)	0.99
보상 조정값(PPO 조정값, Lamda)	0.95
심층 신경망의 형태	256*256*256
정책 조정값 Kullback-Leibler (KL)	0.01
최적화 방법 SGD 횟수	10
자율주행차량의 비율	0%, 10%, 20% ... 100%

- [0142]
- [0143] 시뮬레이션 시나리오는 다음과 같다.
- [0144] 본 발명에서 비신호 교차로를 횡단한 차량은 SUMO 시뮬레이터가 제공하는 선로선택 규칙(right-of-way rule)을

따랐다. 선로설비 규칙의 목적은 교통 규칙을 시행하고 교통 충돌을 방지하는 것이다.

- [0145] 또한 모든 차량의 위치를 관찰하고 POMDP에서 MDP으로 환경을 전환했다. 중요한 것은, 자율 에이전트는 RLlib 라이브러리를 사용하여 물아웃에 대한 특정 보상을 최적화하는 방법을 학습한다. 시뮬레이션은 RL 에이전트를 사용하여 인간 운전 주행과 혼합 자율 주행에서 전체 주행 흐름을 나타낸다.
- [0146] RL 에이전트는 업데이트된 상태를 수신하고 0.1초의 시간 단계에서 새 상태를 가져오고, 인간 운전 차량의 경우 가속 동작은 IDM 모델에 의해 제어된다. 또한, 연속 라우팅은 네트워크 내에서 차량을 유지하기 위해 적용된다.
- [0147] 0.1초의 시간 스텝, 3.2m의 차선 폭, 각 방향으로 2차선, 420m의 차선 길이, 최대 가속도 3m/s^2 , 최소 가속도 -3m/s^2 , 최대 속도 12m/s, 600의 시야, 훈련 과정에 대한 200회의 반복으로 시뮬레이션 실험을 수행했다.
- [0148] 각 방향으로 시간당 1000대의 차량이 유입되도록 설정하고, 비신호 교차로의 범위는 200m에서 220m 사이였다.
- [0149] 현장에서 다양한 시나리오를 시뮬레이션해야 하는데, 본 발명에서는 비신호화된 교차로에서 선도적인 자율 주행 차량의 효과로 초점을 제한했다.
- [0150] 군집 차량은 비신호 교차로에 접근하여 네 가지 다른 방향을 따라 직진 주행한다. 또한 1% ~ 100%의 AV 보급률에 대한 결과를 10% 단위로 제시하고, 비신호 교차로에서 모든 차량에 대해 차선 변경과 좌회전을 무시한다.
- [0151] 도 7의 (a)는 10% ~ 90% 범위의 자율 주행(AV) 점유율을 가진 혼합 교통 상황에서의 비신호화된 교차로에서 선도 자율 주행 차량 실험 환경이고, (b)는 100% 자율 주행(AV) 점유율의 실험 환경이다.
- [0152] 선도적인 자율 주행 차량 실험의 우수성을 입증하기 위해 선도적인 자율 주행 차량 실험을 선도적인 인간 주도 차량 실험과 모든 인간 주도 차량 실험을 포함한 다른 실험과 비교했다. 도 8은 비신호화된 교차로에서 실험의 비교를 보여준다.
- [0153] 도 9는 AV 점유율을 기반으로 한 200회 이상의 평균 보상 곡선 그래프이다.
- [0154] 훈련 정책의 성능(Training Policy's Performance)은 다음과 같다.
- [0155] AV 점유율을 통한 RL 훈련 성과는 학습 성과를 평가하기 위해 사용되었다. 도 9는 AV 점유율을 기반으로 한 200회 이상의 평균 보상 곡선을 나타낸 것이다.
- [0156] 모든 상황에서 곡선이 평평해졌다는 것은 교육 정책이 거의 융합되었음을 나타낸다. 또한, 비신호 교차로의 AV 점유율이 50% AV 점유율을 제외하고 증가함에 따라 평균 보상이 증가했다. 완전 자율 주행은 다른 AV 점유율을 능가했으며, 가장 높은 평균 보상과 상당한 곡선 평탄화를 초래했다. 특히, 전체 자율 주행은 10% AV 점유율에 비해 6.8배 향상되었다.
- [0157] 따라서 전체 자율 주행은 모든 상황에서 다른 AV 점유율을 능가했고, 비신호화된 교차로에서 선도적인 자율 주행 차량 실험의 효과는 AV 점유율이 증가함에 따라 더욱 분명해졌다.
- [0158] 도 10은 비신호화된 교차로에서 AV 점유율을 통한 시공간 역학 특성 그래프이다.
- [0159] 선도적 자율 주행 차량이 부드러운 주행 속도에 미치는 영향은 다음과 같다.
- [0160] 도 10에서 점(point)은 속도에 따라 색상으로 구분되고, 맨 위에 가까운 점은 원활한 교통을 나타낸다. 이와는 대조적으로, 바닥에 가까운 지점은 혼잡한 교통량을 나타낸다.
- [0161] 낮은 AV 점유율의 경우, 사람이 운전하는 차량 거동의 정지 및 이동 파동으로 인해 교란이 발생하여 비신호화된 교차로 영역(200m에서 220m 범위)의 속도가 감소했다. 도 10에서와 같이, AV 점유율이 낮은 비신호화된 교차로에서 거의 모든 지점이 바닥에 근접해 있다.
- [0162] 이는 인간이 운전하는 차량이 비신호화된 교차로 구역에 동시에 접근하고 선로설비 규칙에 따라 속도를 늦추기 때문이다. 높은 AV 점유율에서 포인트는 상단에 가깝고, AV는 더 짧은 시간 내에 느려지며, 따라서 비신호화된 교차로에서 정지 및 이동 파동이 점점 더 적어진다.
- [0163] 전체 자율 주행은 모든 AV 점유율 중 가장 높은 부드러운 주행 속도를 달성했다. 따라서, 교통 체증이 부분적으로 해소되었고, AV 점유율이 증가함에 따라 교통 흐름이 원활해졌다.
- [0164] 도 11은 SUMO 시뮬레이션 환경에서 평균속도, 평균 지체시간, 평균연료 소모량, 평균 배기가스 값 도출 특성 그

래프이다.

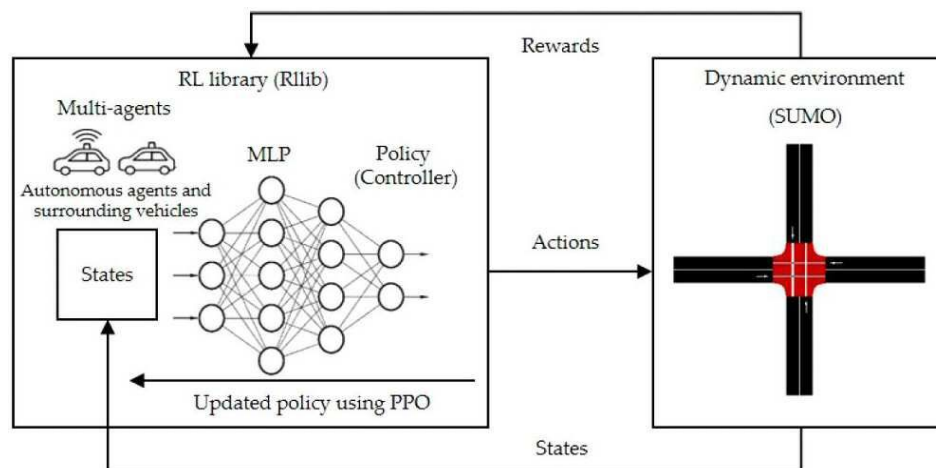
- [0165] 도 11은 평균 속도, 지연 시간, 연료 소비량 및 AV 점유율에 따른 배출량 측면에서 MOE 평가를 나타낸 것으로, MOE 평가 결과는 AV 점유율이 증가함에 따라 시뮬레이션이 더욱 효과적이었음을 나타낸다.
- [0166] 이동성과 관련하여, 평균 속도는 AV 점유율이 증가함에 따라 점차적으로 증가하였고 지연 시간은 점차 감소하였다.
- [0167] 도11의 (a)(b)에서와 같이, 완전 자율 주행은 10% AV 점유율에 비해 평균 속도가 1.19배, 지연 시간은 1.76배 향상되었다. 에너지 효율, 연료 소비 및 배출량은 AV 보급률이 증가함에 따라 약간 감소했다.
- [0168] 도 11의 (c)(d)에서와 같이, 완전 자율 주행은 10% AV 점유율에 비해 연료 소비량이 1.05배, 배기 가스 배출량이 1.22배 향상되었다.
- [0169] 따라서, 선도적인 자율 주행 차량은 AV 점유율이 증가할 때 이동성과 에너지 효율 측면에서 더 효과적인 것을 확인할 수 있다.
- [0170] 이상에서 설명한 본 발명에 따른 자율주행 차량 군집 운영을 위한 비신호 교차로에서의 강화학습기반 통행 개선을 위한 장치 및 방법은 군집된 자율주행차량과 인간운전자 차량이 혼재되어 있는 혼합 교통류 상황에서 자율주행차량 군집주행 학습으로 비신호 교차로 통행을 개선하고 안전성을 확보할 수 있도록 한 것이다.
- [0171] 본 발명은 실제 상황과 같이 자율주행차량이 관찰할 수 있는 범위 내의 정보를 통하여 학습하는 방법으로 강화학습과 마르코프 의사결정 모델 사용(Partial Observability MDP, POMDP)을 적용하여 행동에 대한 강화학습의 보상을 최대화할 수 있도록 한 것이다.
- [0172] 이상에서의 설명에서와 같이 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 변형된 형태로 본 발명이 구현되어 있음을 이해할 수 있을 것이다.
- [0173] 그러므로 명시된 실시 예들은 한정적인 관점이 아니라 설명적인 관점에서 고려되어야 하고, 본 발명의 범위는 전술한 설명이 아니라 특허청구 범위에 나타나 있으며, 그와 동등한 범위 내에 있는 모든 차이점은 본 발명에 포함된 것으로 해석되어야 할 것이다.

부호의 설명

- [0174] 100. SUMO 시뮬레이션 실행부
200. FLOW 적용부
300. 강화학습 라이브러리 환경 구축부

도면

도면1



도면2

Algorithm 1 PPO with an Adaptive KL Penalty Algorithm

- 1: Initial policy parameters θ_0 , weight control β_0 , target KL-divergence δ_{tag}
- 2: For $k = 0, 1, 2 \dots$ do
- 3: Gather set of trajectories on policy $\pi_k = \pi(\theta_k)$
- 4: Optimize the KL penalized using minibatch SGD

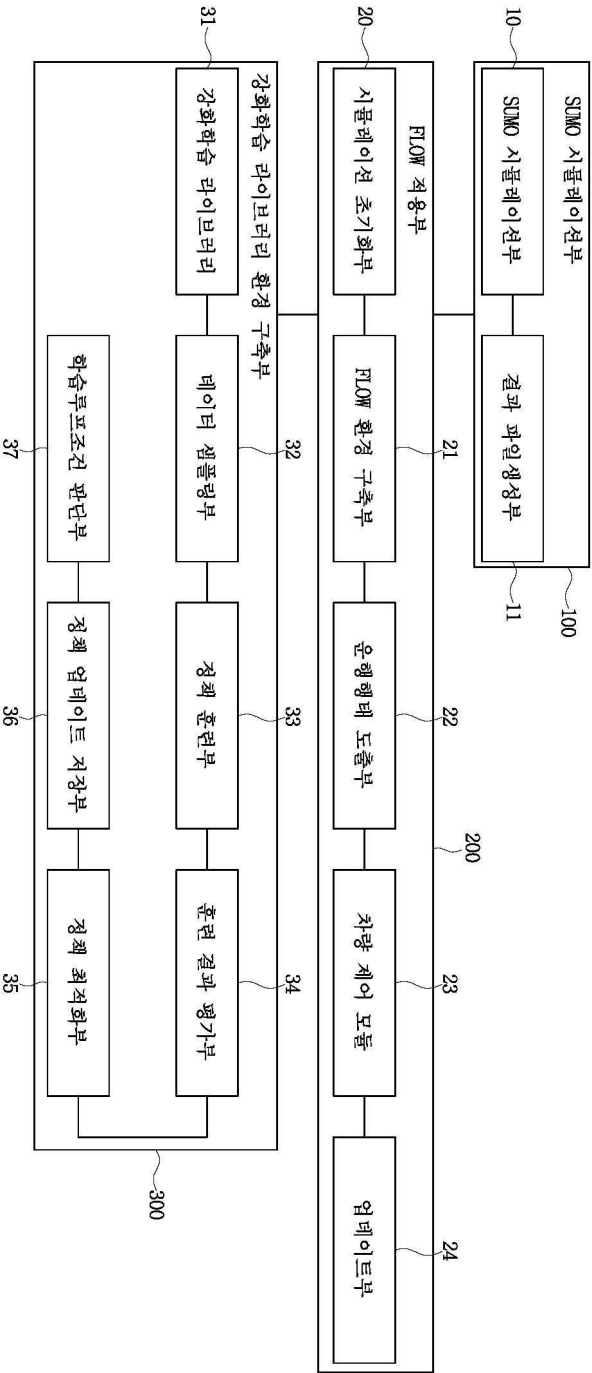
$$J^{KL PEN}(\theta) = \hat{E}_t[r_t(\theta)\hat{A}_t] - \beta \hat{E}_t[KL[\pi_{\theta_{old}}(a_t|s_t), \pi_{\theta}(a_t|s_t)]]$$

- 5: Compute KL divergence between the new and old policy

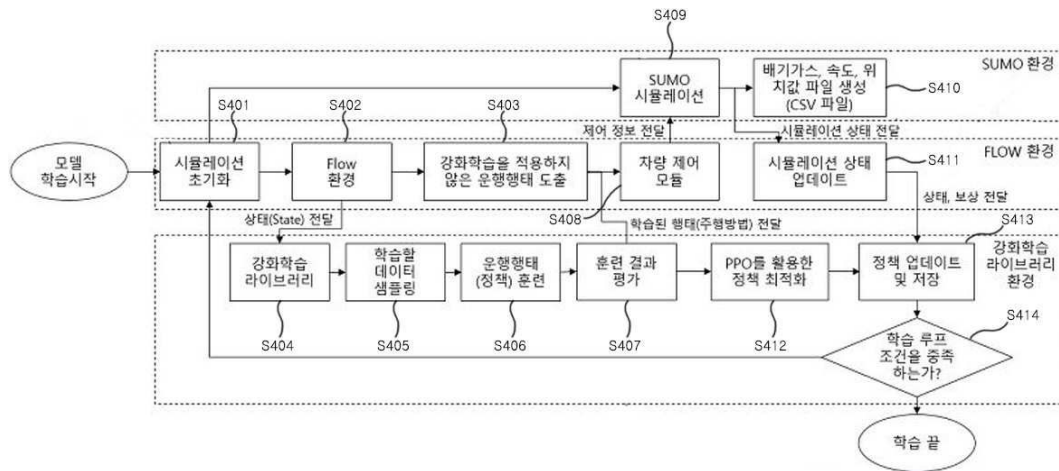
$$\delta = \hat{E}_t[KL[\pi_{\theta_{old}}(a_t|s_t), \pi_{\theta}(a_t|s_t)]]$$

- 6: If $\delta > 1.5 \delta_{tag}$ then
 - $\beta_{k+1} = 2\beta_k$
 - 7: Else if $\delta < \delta_{tag}/1.5$ then
 - $\beta_{k+1} = \beta_k/2$
 - 8: Else
 - pass
 - 9: End if
 - 10: End for
-

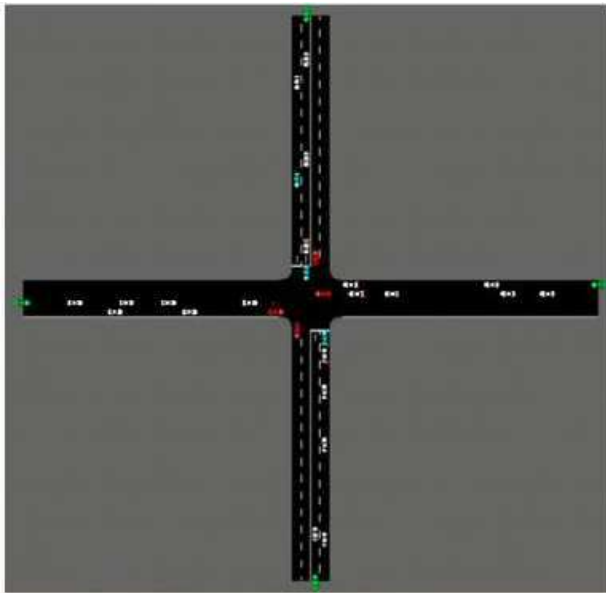
도면3



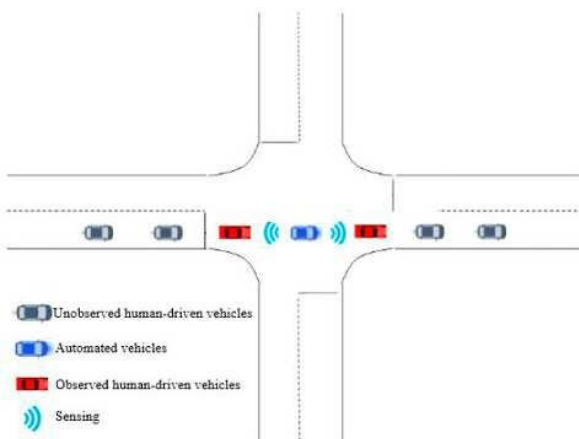
도면4



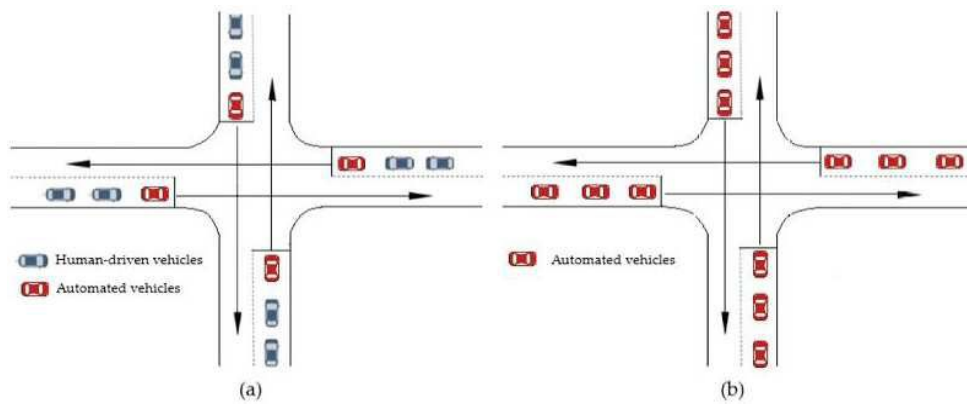
도면5



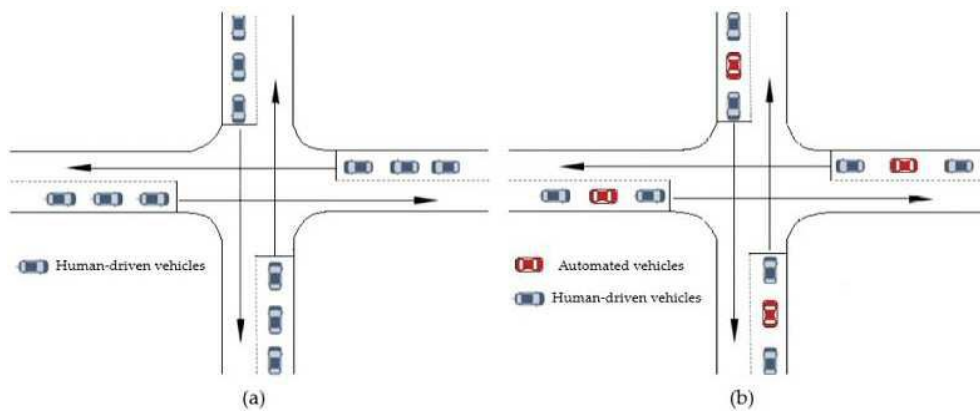
도면6



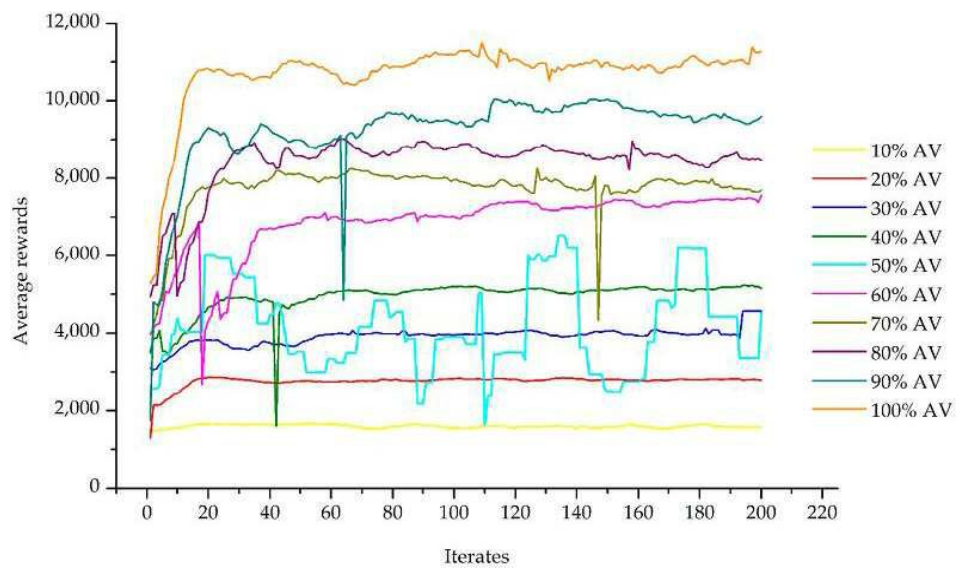
도면7



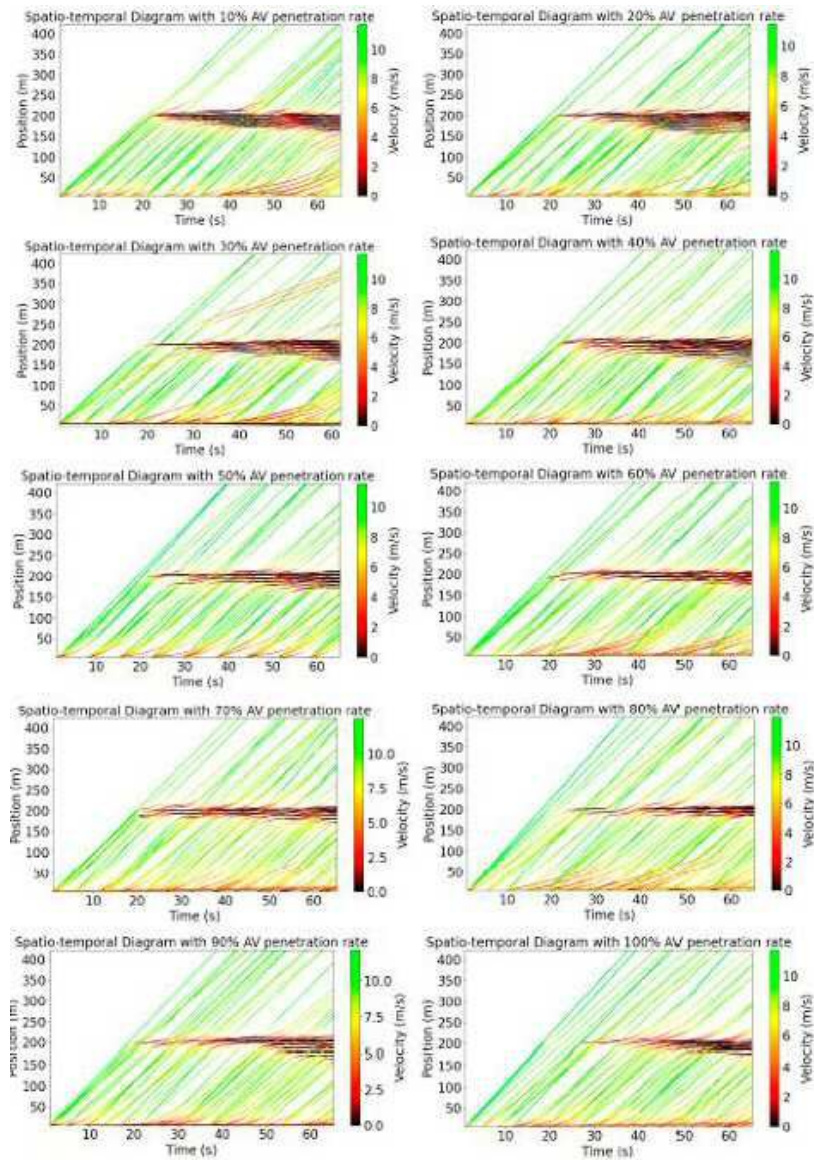
도면8



도면9



도면10



도면11

