US 20250263075A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2025/0263075 A1**

KWON et al. (43) **Pub. Date: Aug. 21, 2025**

(54) **METHOD AND APPARATUS FOR DETERMINING BEHAVIOR BASED ON DEEP REINFORCEMENT LEARNING FOR AUTONOMOUS VEHICLE MERGING STRATEGY**

(71) Applicant: **FOUNDATION OF SOONGSIL UNIVERSITY-INDUSTRY COOPERATION**, Seoul (KR)

(72) Inventors: **Min Hae KWON**, Seoul (KR); **Jae Hwi LEE**, Seoul (KR)

(73) Assignee: **FOUNDATION OF SOONGSIL UNIVERSITY-INDUSTRY COOPERATION**, Seoul (KR)

(21) Appl. No.: **18/766,357**

(22) Filed: **Jul. 8, 2024**

(30) **Foreign Application Priority Data**

Feb. 16, 2024 (KR) 10-2024-0022573

Publication Classification

(51) **Int. Cl.**

B60W 30/18	(2012.01)
B60W 40/04	(2006.01)
B60W 60/00	(2020.01)

(52) **U.S. Cl.**

CPC ***B60W 30/18163*** (2013.01); ***B60W 40/04*** (2013.01); ***B60W 60/001*** (2020.02); *B60W 2556/45* (2020.02)

(57) **ABSTRACT**

A deep reinforcement learning-based vehicle action decision apparatus for merging strategy of an autonomous vehicle in an on-ramp merging zone is disclosed. The deep reinforcement learning-based vehicle action decision apparatus comprises an information observation unit for collecting observation information from a sensing module or roadside unit (RSU) of an autonomous vehicle; a policy execution unit for deciding on a current action, including acceleration control and lane change of the autonomous vehicle, based on the current observation information and policy; and a reward determination unit for determining a reward according to the current observation information, the current action, and the next observation information according to the current action, wherein reward in the reward determination unit is determined through a reward term related to speed, lane change, safety distance compliance, and an accident of the autonomous vehicle and a merge reward term related to merge of an autonomous vehicle in the on-ramp merging zone.

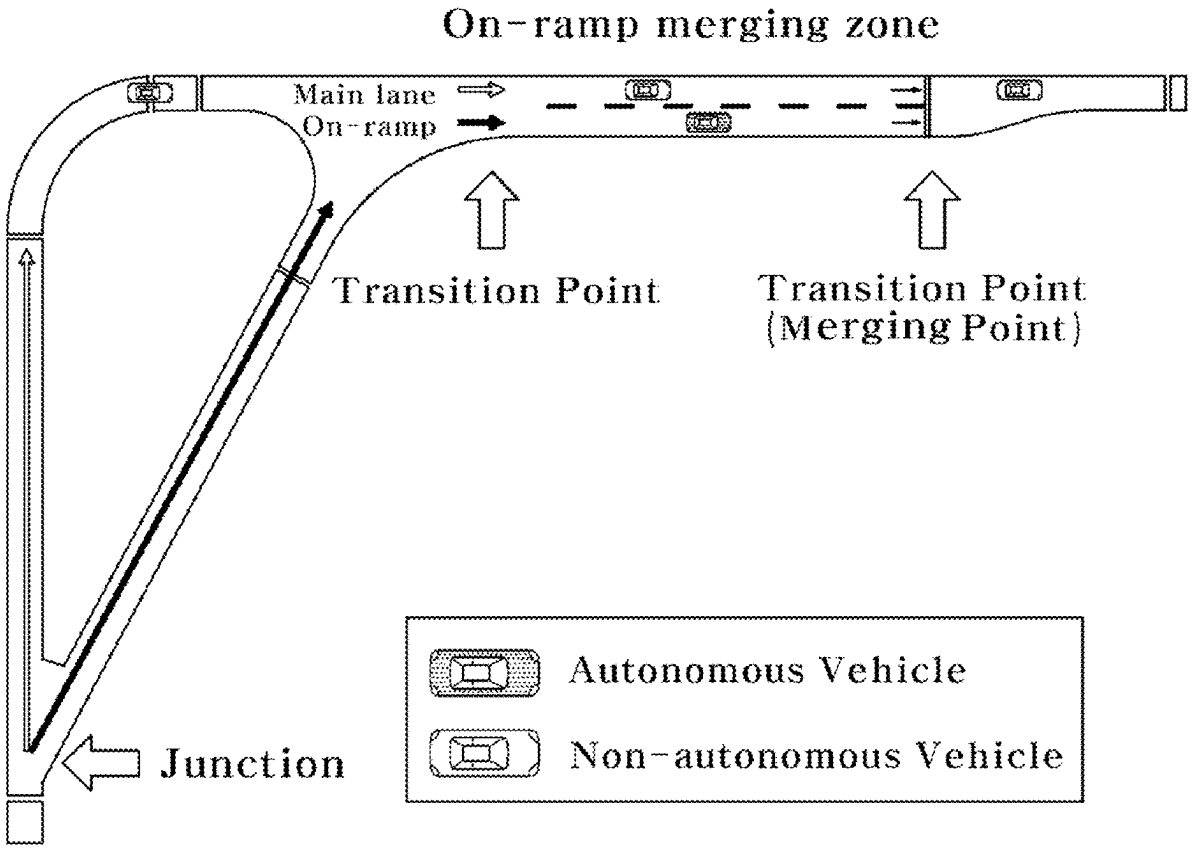


FIG. 1

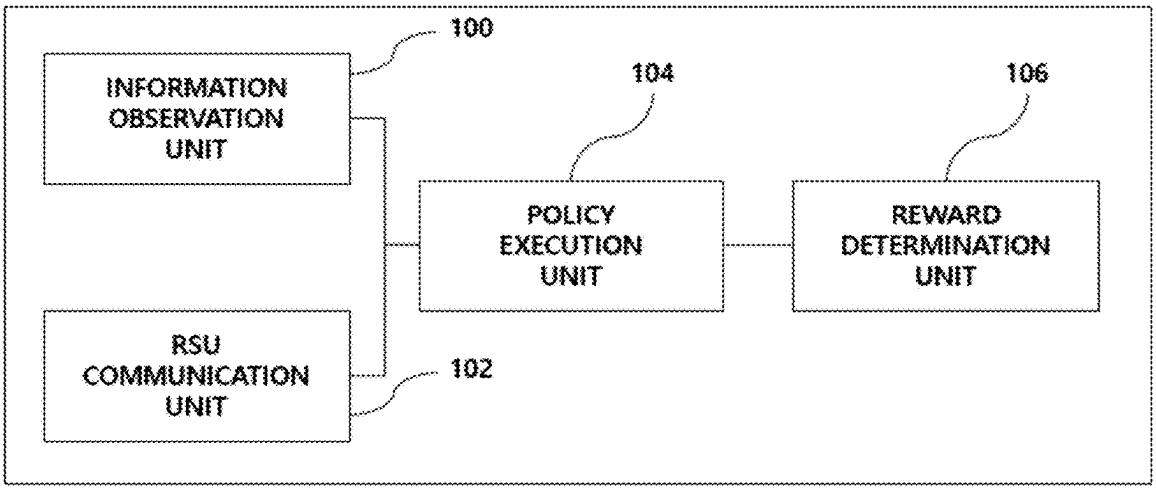

100
INFORMATION OBSERVATION UNIT
104
POLICY EXECUTION UNIT
106
REWARD DETERMINATION UNIT
RSU COMMUNICATION UNIT
102

FIG. 2

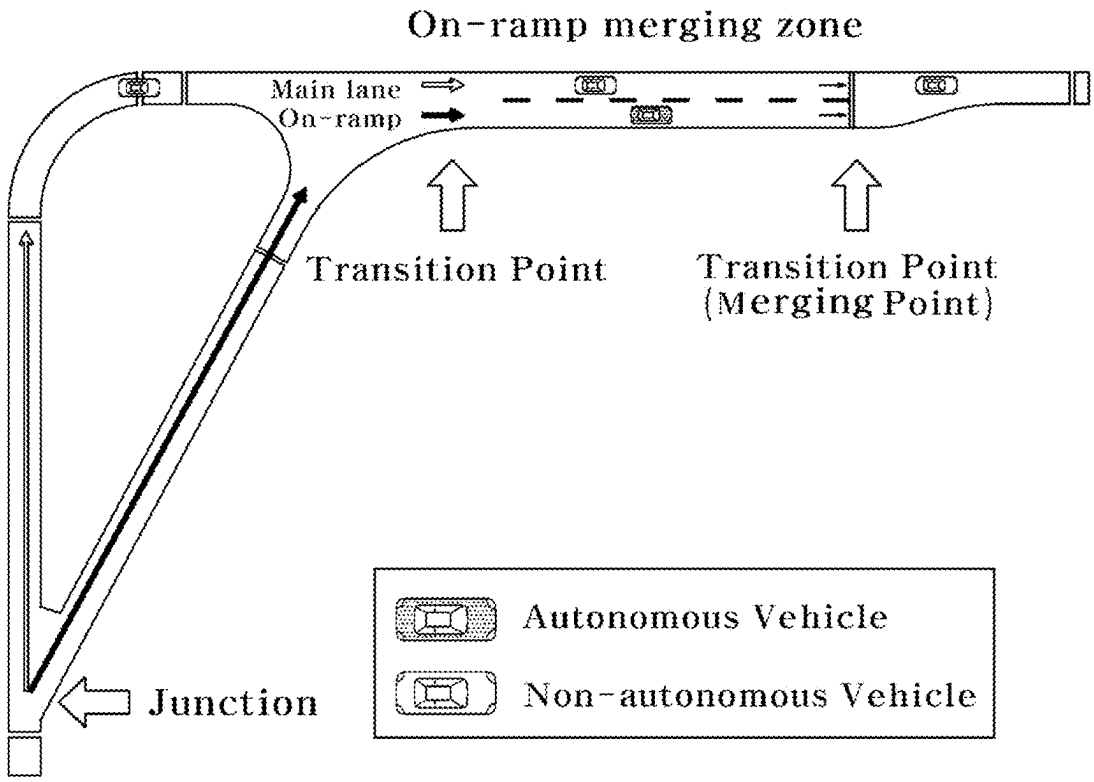

On-ramp merging zone
Main lane
On-ramp
Transition Point
Transition Point
(Merging Point)
Junction
Autonomous Vehicle
Non-autonomous Vehicle

FIG. 3

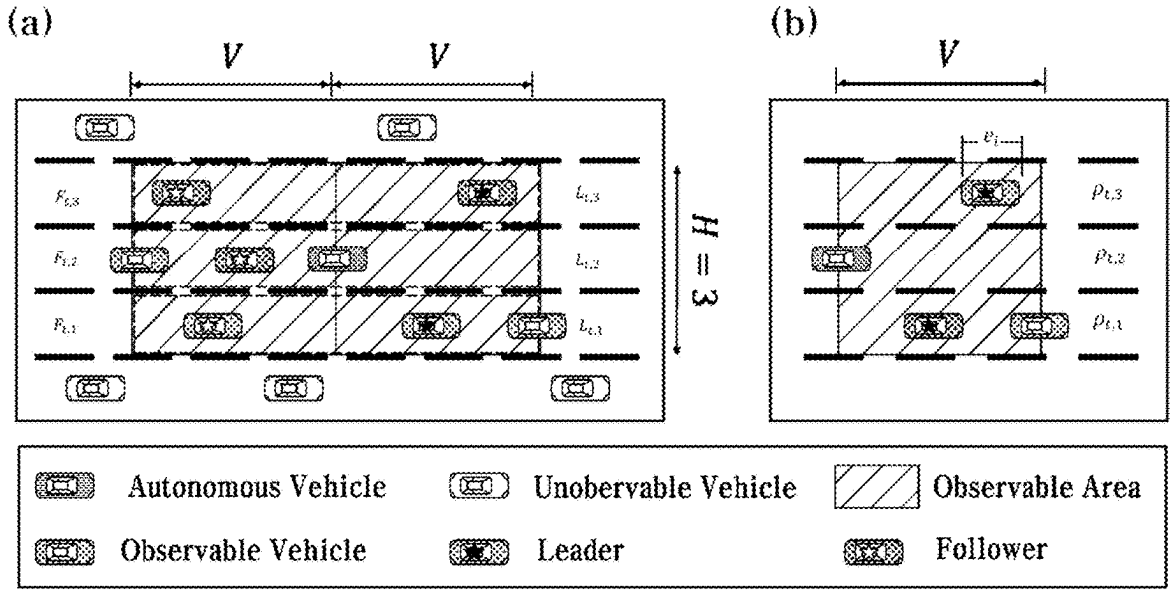
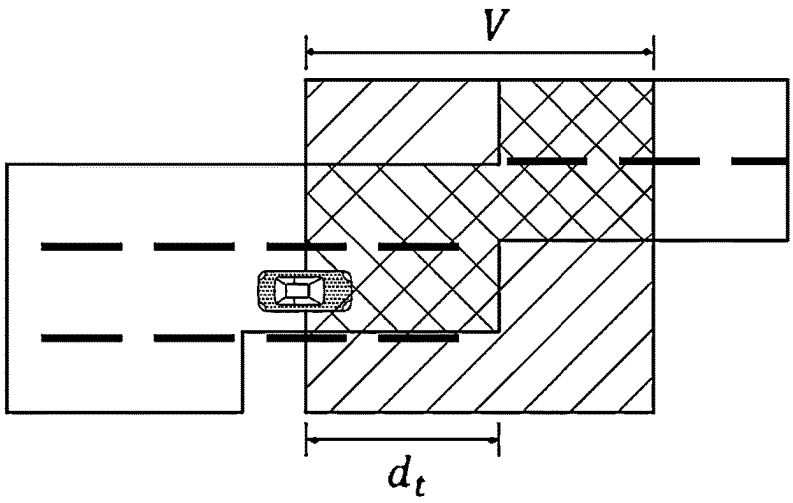

(a)
V
V
H = 3
(b)
V
Autonomous Vehicle
Unobervable Vehicle
Observable Area
Observable Vehicle
Leader
Follower

FIG. 4



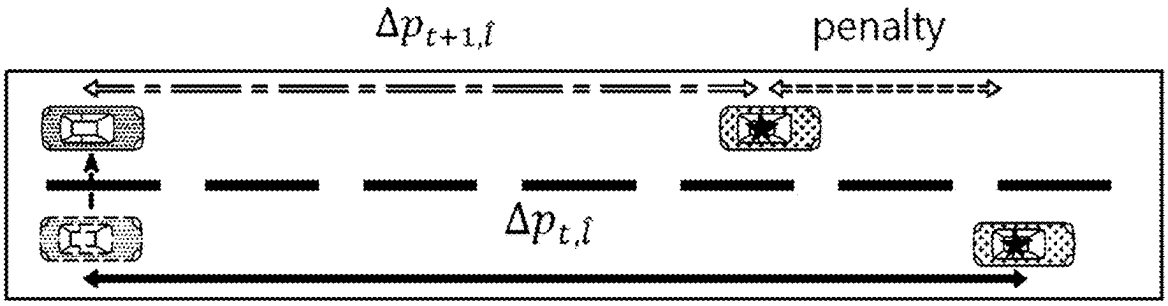
$$\zeta_{t,H} = -(d_t - V)$$

$$\zeta_{t,H-1} = V$$

$$\zeta_{t,2} = d_t - V$$

$$\zeta_{t,1} = -V$$

FIG. 5



$$|a_{t,lc}| = 1, \Delta p_{t,l} > \Delta p_{t+1,l}$$

FIG. 6

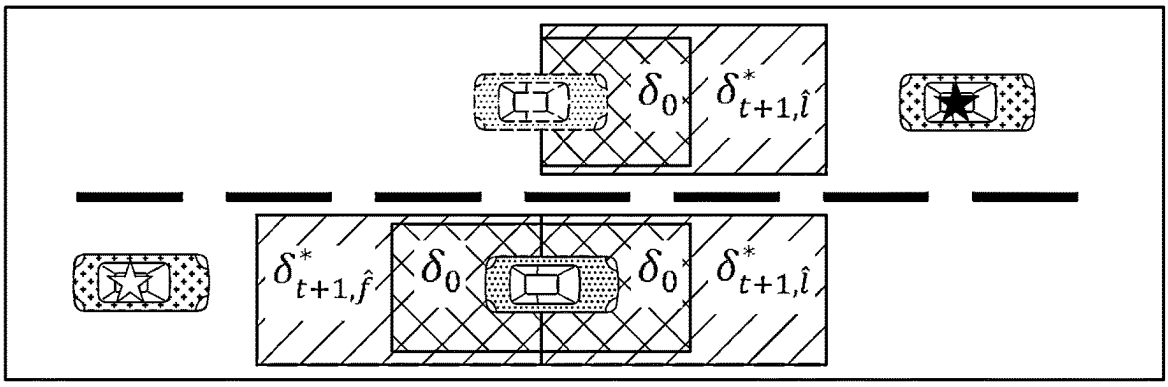

δ_0
$\delta^*_{t+1,\hat{l}}$
$\delta^*_{t+1,\hat{f}}$
δ_0
δ_0
$\delta^*_{t+1,\hat{l}}$

FIG. 7

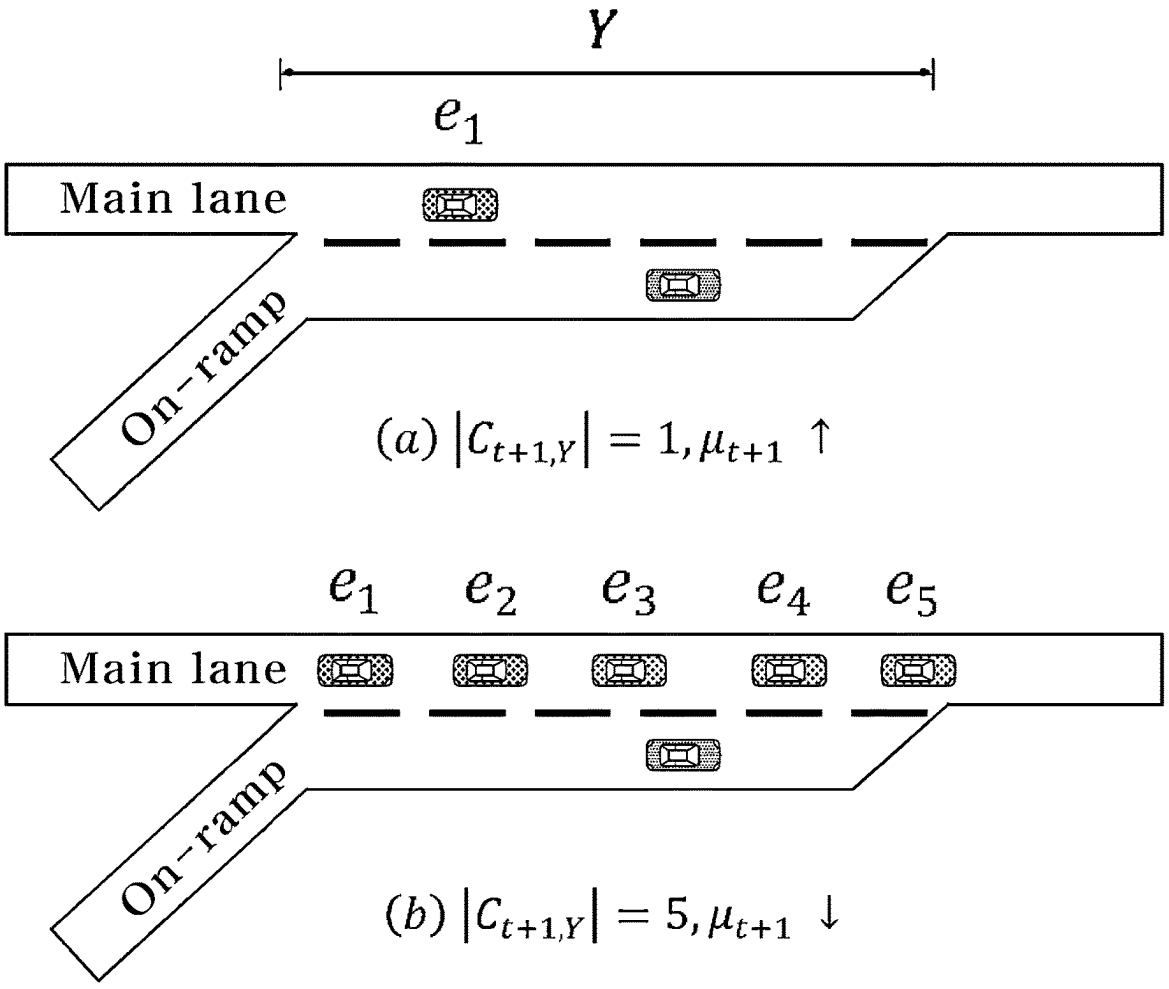

Y
e_1
Main lane
On-ramp
$(a)\ |C_{t+1,Y}| = 1, \mu_{t+1}\ \uparrow$
e_1 e_2 e_3 e_4 e_5
Main lane
On-ramp
$(b)\ |C_{t+1,Y}| = 5, \mu_{t+1}\ \downarrow$

FIG. 8

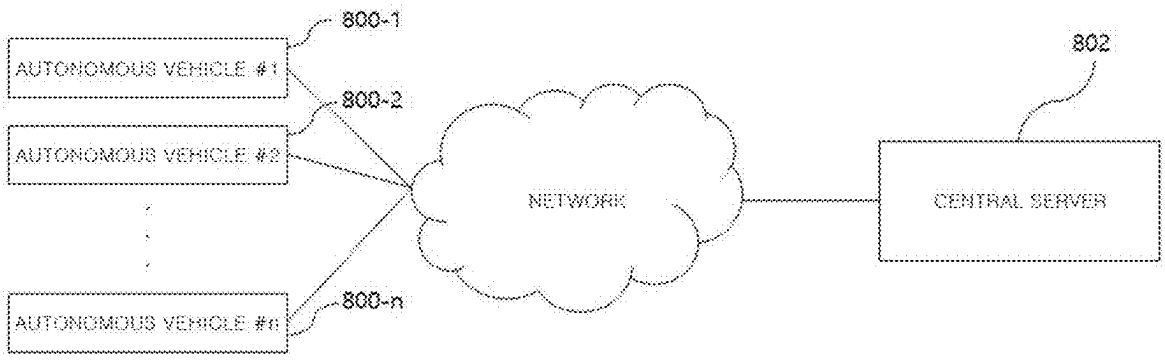

800-1
AUTONOMOUS VEHICLE #1
800-2
AUTONOMOUS VEHICLE #2
AUTONOMOUS VEHICLE #n
800-n
NETWORK
802
CENTRAL SERVER

FIG. 9

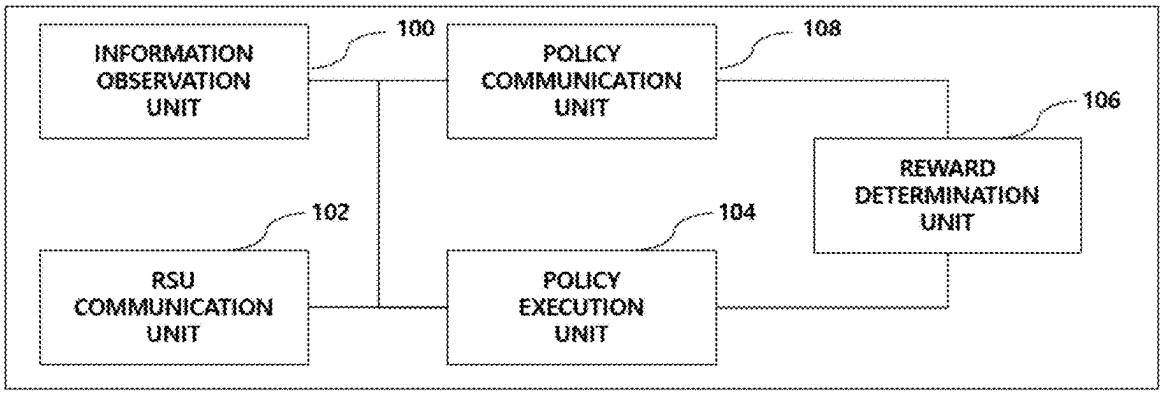

INFORMATION OBSERVATION UNIT
100
POLICY COMMUNICATION UNIT
108
REWARD DETERMINATION UNIT
106
RSU COMMUNICATION UNIT
102
POLICY EXECUTION UNIT
104

FIG. 10

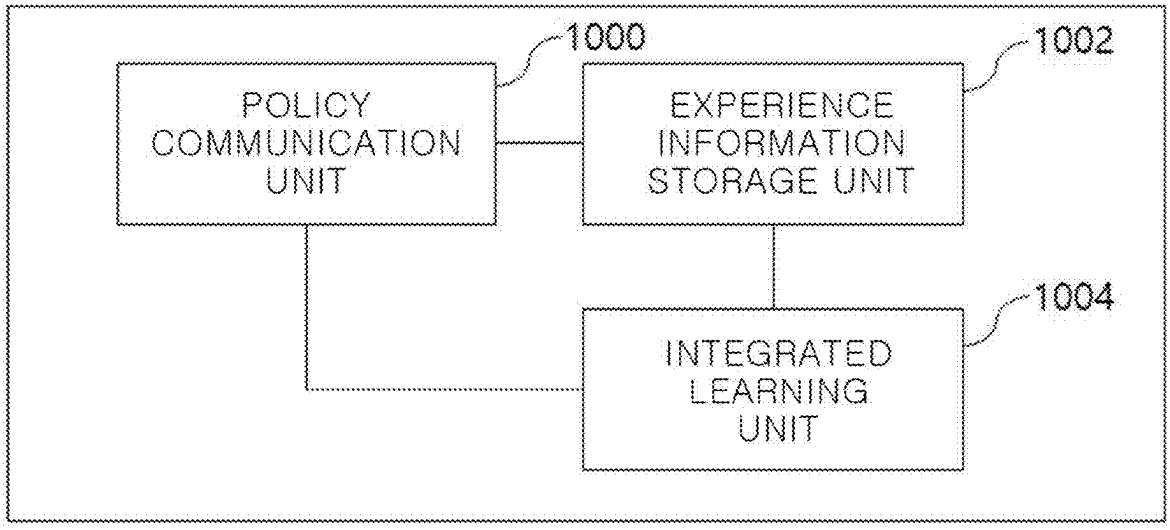

1000
1002
POLICY COMMUNICATION UNIT
EXPERIENCE INFORMATION STORAGE UNIT
1004
INTEGRATED LEARNING UNIT

FIG. 11

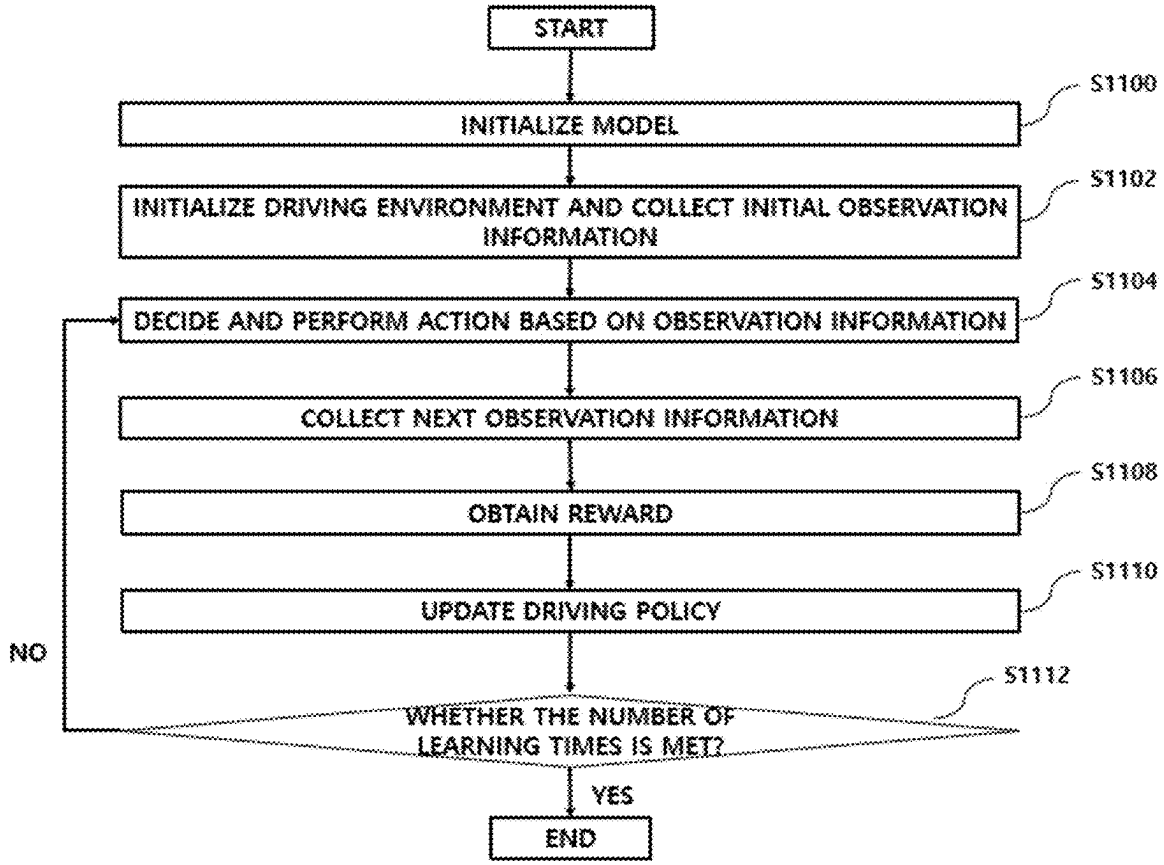
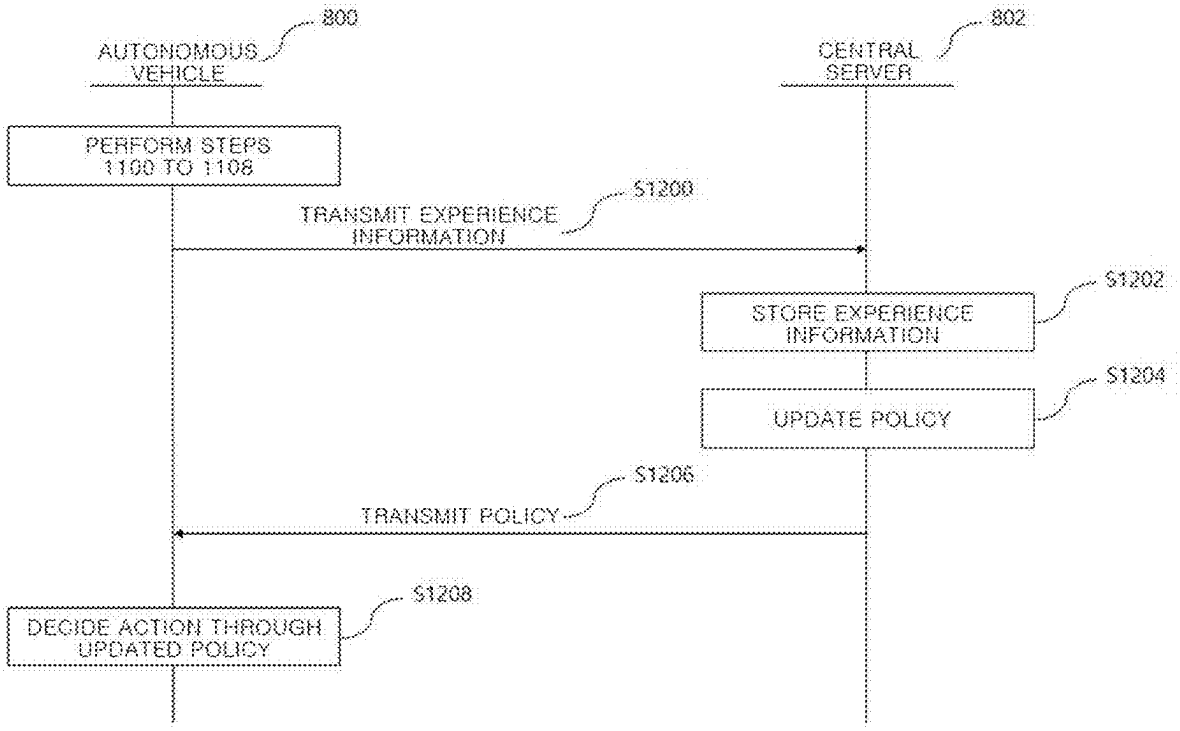

START
INITIALIZE MODEL
S1100
INITIALIZE DRIVING ENVIRONMENT AND COLLECT INITIAL OBSERVATION INFORMATION
S1102
DECIDE AND PERFORM ACTION BASED ON OBSERVATION INFORMATION
S1104
COLLECT NEXT OBSERVATION INFORMATION
S1106
OBTAIN REWARD
S1108
UPDATE DRIVING POLICY
S1110
WHETHER THE NUMBER OF LEARNING TIMES IS MET?
S1112
NO
YES
END

FIG. 12


800
AUTONOMOUS VEHICLE
802
CENTRAL SERVER
PERFORM STEPS 1100 TO 1108
S1200
TRANSMIT EXPERIENCE INFORMATION
S1202
STORE EXPERIENCE INFORMATION
S1204
UPDATE POLICY
S1206
TRANSMIT POLICY
S1208
DECIDE ACTION THROUGH UPDATED POLICY

METHOD AND APPARATUS FOR DETERMINING BEHAVIOR BASED ON DEEP REINFORCEMENT LEARNING FOR AUTONOMOUS VEHICLE MERGING STRATEGY

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of Korean Application No. 10-2024-0022573 filed on Feb. 16, 2024, in the Korean Intellectual Property Office. All disclosures of the document named above are incorporated herein by reference.

TECHNICAL FIELD

[0002] The present invention relates to a vehicle action decision method and apparatus based on deep reinforcement learning for merging strategies of autonomous vehicles.

BACKGROUND ART

[0003] With recent rapid developments in the field of artificial intelligence, research on autonomous driving technology is being conducted in various fields.

[0004] Conditional autonomous driving technology has recently been commercialized, and it is expected that a level **4** or higher fully autonomous driving system that does not require driver intervention will be introduced soon.

[0005] However, there are still limitations in technological completeness in the development of autonomous driving technology for highway ramp sections.

[0006] Because the autonomous driving environment in the ramp section is uncertain and dynamic, it requires flexible and precise decision-making ability compared to other road structures.

[0007] On roads that require merging, many risk factors are difficult to predict, such as sudden vehicle stops and lane changes.

RELATED ART REFERENCES

[0008] Korean Patent Application Publication No. 10-2022-0102694

[0009] Korean Patent Application Publication No. 10-2023-0127946

DISCLOSURE

Technical Problem

[0010] In order to solve the problems of the prior art described above, the present invention proposes a deep reinforcement learning-based vehicle action decision method and apparatus for the merging strategy of autonomous vehicles that allows autonomous vehicles to successfully pass the on-ramp merging zone where vehicle collisions frequently occur due to sudden stops and lane changes.

Technical Solution

[0011] In order to achieve the above-described object, according to one embodiment of the present invention, a deep reinforcement learning-based vehicle action decision apparatus for merging strategy of an autonomous vehicle in an on-ramp merging zone comprises an information observation unit for collecting observation information from a sensing module or roadside unit (RSU) of an autonomous vehicle; a policy execution unit for making a decision regarding a current action, including acceleration control and lane change of the autonomous vehicle, based on the current observation information and policy; and a reward determination unit for determining a reward according to the current observation information, the current action, and the next observation information according to the current action, wherein reward in the reward determination unit is determined through a reward term related to speed, lane change, safety distance compliance, and an accident of the autonomous vehicle and a merge reward term related to merge of an autonomous vehicle in the on-ramp merging zone

[0012] The observation information may comprise at least one of the absolute speed of the autonomous vehicle, the relative speed between the closest leader vehicle in front and the closest follower vehicle in rear among vehicles in each lane that the autonomous vehicle can observe, the relative distance between the leader vehicle and the follower vehicle, vehicle density for each lane within front observation range of the autonomous vehicle, and presence or absence of a lane within front observation range of the autonomous vehicle.

[0013] The merge reward term may be determined using the remaining drivable distance from an on-ramp where the autonomous vehicle is located to a merging point and traffic density.

[0014] The closer the autonomous vehicle merges to the merging point, the reward determination unit may give a higher penalty as it is considered a delayed merge.

[0015] The reward determination unit may alleviate the penalty using a traffic density weight according to the traffic density of the main lane of the on-ramp merging zone.

[0016] The traffic density weight may be determined using at least one of the length of the on-ramp merging zone, the number of vehicles driving in the main lane of the on-ramp merging zone, the length of a vehicle belonging to a vehicle set in the main lane of the on-ramp merging zone, and minimum safety distance.

[0017] The merge reward term may be defined by the following equation,

$$R_{t,5} = \begin{cases} \mu_{t+1} \times \zeta_{t+1,\hat{h}}, & \zeta_{t+1,\hat{h}} < 0 \\ 0, & \zeta_{t+1,\hat{h}} \geq 0 \end{cases} \quad \text{[Equation]}$$

[0018] wherein $\zeta_{t+1,\hat{h}} \in \zeta_{t=1}$ is the remaining drivable distance from an on-ramp where the autonomous vehicle is located to a merging point and μ_{t+1} is traffic density weight.

[0019] The traffic density weight may be defined by the following equation,

$$\mu_{t+1} = 1 - \left(\frac{\sum_{i=1}^{|C_{t+1,Y}|} (e_i + \delta_0)}{Y} \right) \quad \text{[Equation]}$$

[0020] wherein Y is the length of the on-ramp merging zone, $|C_{t+1,Y}|$ is the number of vehicles driving in a main lane of the on-ramp merging zone, e_i is the length of the ith vehicle in a set of vehicles $|C_{t+1,Y}|$ in a main lane of the on-ramp merging zone, and δ_0 is the minimum safety distance.

[0021] The learning of the policy may be performed by the autonomous vehicle itself or by a central server connected to multiple autonomous vehicles through a network.

[0022] When the learning of the policy is performed by the autonomous vehicle itself, the apparatus may further comprise a policy learning unit for updating the policy according to the collected current observation information, the current action, the next observation information, and experience information including the reward, and determining whether the number of learning times is met to learn the policy.

[0023] When the learning of the policy is performed by the central server, the apparatus may further comprise a policy communication unit for transmitting experience information including the collected current observation information, the current action, the next observation information, and the reward to the central server, and receiving a policy that the central server has learned using experience information received from the multiple autonomous vehicles.

[0024] According to another embodiment of the present invention, a deep reinforcement learning-based vehicle action decision apparatus for merging strategy of an autonomous vehicle in an on-ramp merging zone comprises a processor; and a memory connected to the processor, wherein the memory stores program instruction, when executed by the processor, configured to perform operations comprises collecting observation information from a sensing module or roadside unit (RSU) of an autonomous vehicle; making a decision regarding a current action, including acceleration control and lane change of the autonomous vehicle, based on the current observation information and policy; and determining a reward according to the current observation information, the current action, and the next observation information according to the current action, wherein reward is determined through a reward term related to speed, lane change, safety distance compliance, and an accident of the autonomous vehicle and a merge reward term related to merge of an autonomous vehicle in the on-ramp merging zone.

[0025] According to another embodiment of the present invention, a deep reinforcement learning-based vehicle action decision method for merging strategy of an autonomous vehicle in an on-ramp merging zone comprises collecting observation information from a sensing module or roadside unit (RSU) of an autonomous vehicle; making a decision regarding a current action, including acceleration control and lane change of the autonomous vehicle, based on the current observation information and policy; and determining a reward according to the current observation information, the current action, and the next observation information according to the current action, wherein the reward is determined through a reward term related to speed, lane change, safety distance compliance, and an accident of the autonomous vehicle and a merge reward term related to merge of an autonomous vehicle in the on-ramp merging zone.

Advantageous Effects

[0026] According to the present invention, there is an advantage that autonomous vehicles can smoothly merge in an on-ramp merging zone where vehicle collisions frequently occur due to sudden stops and lane changes.

DESCRIPTION OF DRAWINGS

[0027] These and/or other aspects will become apparent and more readily appreciated from the following description of the embodiments, taken in conjunction with the accompanying drawings in which:

[0028] FIG. **1** is a diagram illustrating the configuration of a deep reinforcement learning-based vehicle action decision apparatus for a merging strategy of autonomous vehicles according to the present embodiment;

[0029] FIG. **2** is a diagram showing the structure of an on-ramp merging zone;

[0030] FIG. **3** is a diagram showing the observable area and traffic density according to the present embodiment;

[0031] FIG. **4** is a diagram illustrating various types of the existence of a lane;

[0032] FIG. **5** is a diagram showing a case where a penalty is given after changing lanes;

[0033] FIG. **6** is a diagram for explaining driving action for maintaining a safe distance between vehicles;

[0034] FIG. **7** is a diagram for explaining traffic density weights according to the present embodiment;

[0035] FIG. **8** is a diagram showing the configuration of a system for autonomous vehicles in a multi-autonomous driving environment according to the present embodiment;

[0036] FIG. **9** is a diagram showing the configuration of an autonomous vehicle in a multi-autonomous driving environment;

[0037] FIG. **10** is a diagram showing the configuration of a central server in a multi-autonomous driving environment;

[0038] FIG. **11** is a diagram illustrating a policy learning process in a single autonomous vehicle environment according to the present embodiment; and

[0039] FIG. **12** is a flowchart showing the policy learning process in a multi-autonomous driving environment according to the present embodiment.

DETAILED DESCRIPTION OF EMBODIMENTS

[0040] Since the present invention can make various changes and have various embodiments, specific embodiments will be illustrated in the drawings and described in the detailed description. However, this is not intended to limit the present invention to specific embodiments, and should be understood to include all changes, equivalents, and substitutes included in the spirit and technical scope of the present invention.

[0041] The terms used herein are only used to describe specific embodiments and are not intended to limit the invention. Singular expressions include plural expressions unless the context clearly dictates otherwise. In this specification, terms such as "comprise" or "have" are intended to designate the presence of features, numbers, steps, operations, components, parts, or combinations thereof described in the specification, but it should be understood that this does not exclude in advance the presence or addition of one or more other features, numbers, steps, operations, components, parts, or combinations thereof.

[0042] In addition, the components of the embodiments described with reference to each drawing are not limited to the corresponding embodiments, and may be implemented to be included in other embodiments within the scope of maintaining the technical spirit of the present invention, and

even if separate description is omitted, a plurality of embodiments may be re-implemented as a single integrated embodiment.

[0043] In addition, when describing with reference to the accompanying drawings, identical or related reference numerals will be given to identical or related elements regardless of the reference numerals, and overlapping descriptions thereof will be omitted. In describing the present invention, if it is determined that a detailed description of related known technologies may unnecessarily obscure the gist of the present invention, the detailed description will be omitted.

[0044] This embodiment proposes a system for learning and performing a driving policy (hereinafter referred to as policy) for a successful merging strategy in an on-ramp merging zone, and can be applied regardless of single and multiple autonomous driving environments.

[0045] Here, policy can be defined as a policy model or policy network that makes decisions (action decisions) for driving autonomous vehicles.

[0046] At this time, policy learning is performed through deep reinforcement learning, and includes the design of a Markov Decision Process (MDP) for this purpose.

[0047] In this embodiment, the decision-making of the autonomous vehicle may be actions related to acceleration control and lane change in the on-ramp merging zone.

[0048] FIG. **1** is a diagram illustrating the configuration of a deep reinforcement learning-based vehicle action decision apparatus for a merging strategy of autonomous vehicles according to the present embodiment.

[0049] As shown in FIG. **1**, the apparatus according to the present embodiment may comprise an information observation unit **100**, a roadside unit (RSU) communication unit **102**, a policy execution unit **104**, and a reward determination unit **106**.

[0050] The configuration in FIG. **1** may be configured inside an autonomous vehicle, but is not necessarily limited thereto.

[0051] The information observation unit **100** collects observation information from a roadside unit (RSU) using the sensing module of the autonomous vehicle or the roadside unit communication unit **102**.

[0052] Observation information may be information including at least part of state information about the surrounding environment, and if all information about the surrounding environment can be collected, observation information may be used in the same sense as state information.

[0053] The roadside unit communication unit **102** allows the autonomous vehicle to obtain information about the target vehicle through communication with the roadside unit when there is a vehicle that cannot be sensed through its own sensing module. At this time, information exchange between autonomous vehicle and roadside unit is done through V2I (Vehicle to Infrastructure). Additionally, the roadside unit may communicate with adjacent roadside units to obtain or transmit information about the target vehicle. Communication between roadside units is done through I2I (Infrastructure to Infrastructure).

[0054] In addition, policy learning for decision-making by the policy execution unit **104** can be performed by the autonomous vehicle itself (single autonomous vehicle environment) or by a central server connected to multiple autonomous vehicles through a network.

[0055] The policy execution unit **104** makes decisions regarding current actions, including acceleration control and lane changes of the autonomous vehicle, based on current observation information and policy.

[0056] The reward determination unit **106** determines the reward based on current observation information, current action, and next observation information according to the current action.

[0057] A reward is determined through reward terms related to the autonomous vehicle's speed, lane change, compliance with safety distance, and accidents, and merge reward terms related to the merge of autonomous vehicles in the on-ramp merging zone, which will be described in detail below.

[0058] Vehicle action decisions according to this embodiment are performed based on deep reinforcement learning, and the Markov Decision Process will be described in detail below.

[0059] In MDP, there is an assumption that an agent is fully observable of all state information in the environment.

[0060] However, in realistic environments such as autonomous driving, perfect observation of all state information is limited, so in this embodiment, a reinforcement learning problem is defined through POMDP (Partially Observable MDP), which makes decisions based on partial state information.

[0061] POMDP is defined as a tuple $\langle S, A, T, O, R, \Omega, \gamma \rangle$, where $s_t \in S$ means the state information of the road environment (state), $a_t \in A$ means the driving action of the autonomous driving agent, $T(s_{t+1}|s_t,a_t)$ means the state transition probability, and $o_t \in O$ means the observable information by the autonomous driving agent at a specific time point t in state s_t. In addition, $\Omega(o_t|s_t)$ means the observation probability, $R(s_t,a_t,s_{t+1})$ means the reward function, and $\gamma \in [0,1)$ means the discount factor over time.

[0062] In the present embodiment, to learn the merging strategy policy in the on-ramp merging zone, a road structure including a transition point where the number of lanes on the road increases/decreases is considered, as shown in FIG. **2**.

[0063] As shown in FIG. **2**, the on-ramp merging zone is defined as the main lane, the on-ramp where the autonomous vehicle to merge is located, and the merging point where the main lane and the on-ramp merge, and the merging point can also be defined as a transition point.

[0064] Specifically, an environment in which a total of N vehicles $C=\{c_1, c_2, \ldots, c_N\}$ drive on a road with M transition points $\mathcal{M}=\{1,2, \ldots, M\}$ is considered.

[0065] In the relevant road environment, merging from the on-ramp to the main lane is essential, and an irregular road environment is created due to a decrease in traffic capacity due to a reduction in lanes.

[0066] In the present embodiment, an autonomous vehicle can safely and efficiently perform merging strategy policy learning in this environment.

[0067] In the present embodiment, an environment is considered in which an autonomous vehicle branches off at a junction and enters the on-ramp, and a non-autonomous vehicle enters the main lane without branching.

[0068] The set of vehicles $C=C_{NAV}\cup C_{AV}$ on the road comprises N−1 non-autonomous vehicles $C_{NAV}=\{c_i|i\neq N\}$ and 1 autonomous vehicle

$$C_{AV}=\{c_i \mid i=N\}.$$

[0069] Here, road driving state information s_t is defined as follows.

$$s_t=[v_t^T,p_t^T,k_t^T,d_t^T]^T$$

[0070] Here, $v_t=[v_{t,1},v_{t,2},\ldots,v_{t,N}]^T$ represents the absolute speed of all vehicles on the road and $p_t=[p_{t,1},p_{t,2},\ldots,p_{t,N}]^T$ represents the absolute positions of all vehicles.

[0071] In addition, $[k_{t,1},k_{t,2},\ldots,k_{t,N}]^T$ denotes the lane number of the road on which each vehicle is located, and $d_t=[d_{t,1},d_{t,2},\ldots,d_{t,N}]^T$ denotes the distance to the nearest transition point for each vehicle.

[0072] In the present embodiment, a POMDP-based reinforcement learning problem is considered for agent decision-making through partial observation information.

[0073] FIG. **3** is a diagram showing the observable area and traffic density according to this embodiment.

[0074] Referring to FIG. **3***a*, the autonomous vehicle can observe a total of H lanes, including the located lane, and a distance of V forward and behind the absolute position of the agent, which is defined as an observable area.

[0075] Vehicles within the observable area are defined as observable vehicles and are denoted as a set $C_{t,obs}$.

[0076] At this time, $C_{t,obs}$ can be divided into a front vehicle set L_t and a rear vehicle set F_t, and is defined as follows.

$$C_{t,obs}=L_t*\bigcup F_t,$$

where

$$L_t=\bigcup_{h=1}^{H}L_{t,h},$$

$$F_t=\bigcup_{h=1}^{H}F_{t,h}$$

[0077] Here, $L_{t,h}\subset L_t$ and $F_{t,h}\subset F_t$ mean the front and rear vehicle sets within the observable area for each lane h.

[0078] At this time, in the set of vehicles observed for each front and rear lane, the front vehicle closest to the autonomous vehicle is defined as the leader vehicle leader $l_{t,h}\in L_{t,h}$, and the rear vehicle closest to the autonomous vehicle is defined as the follower vehicle follower $f_{t,h}\in F_{t,h}$.

[0079] The observation information $o_t\in O$ of an autonomous vehicle at time t is defined as follows.

$$o_t=[v_{t,N},\Delta v_t^T,\Delta p_t^T,\rho_t^T,\zeta_t^T]^T$$

[0080] $v_{t,N}$ is the absolute speed of the autonomous vehicle.

[0081] $[\Delta v_{t,l_1},\Delta v_{t,l_2},\ldots,\Delta v_{t,l_H},\Delta v_{t,f_1},\Delta v_{t,f_1},\ldots,\Delta v_{t,f_H}]^T$ is the relative speed between the leader vehicle and the follower vehicle in each lane that can be observed by the autonomous vehicle.

[0082] $[\Delta p_{t,l_1},\Delta p_{t,l_2},\ldots,\Delta p_{t,l_H},\Delta p_{t,f_1},\Delta p_{t,f_1},\ldots,\Delta p_{t,f_H}]^T$ means the relative distance between the leader vehicle and the follower vehicle in each lane.

[0083] $\rho_t=[\rho_{t,1},\rho_{t,2},\ldots,\rho_{t,H}]^T$ represents the vehicle density for each lane within the front observation range of the autonomous vehicle.

[0084] Referring to FIG. **3***b*, the vehicle density $\rho_{t,h}$ in a specific lane h means the ratio of vehicles in that lane compared to the front observation range V.

$$\rho_{t,h}=\frac{\sum_{i=1}^{|L_{t,h}|}(\delta_0+e_i)}{Y}$$

[0085] Here, vehicle density is defined by the number of vehicles $|L_{t,h}|$ observed in a specific lane h, the length of the ith vehicle in that lane, and the minimum safe distance between vehicles δ_0.

[0086] $\zeta_t=[\zeta_{t,1},\zeta_{t,2},\ldots,\zeta_{t,H}]^T$ indicates the presence or absence of a lane within the observation range ahead of the autonomous vehicle.

[0087] FIG. **4** is a diagram illustrating various types of existence of a lane.

[0088] Referring to FIG. **4**, the existence of a lane is defined by the observable range V and the remaining distance d_t to the transition point.

[0089] If a lane within the front observation range of an autonomous vehicle exists and then is cut off, it is defined as d_t-V, and if it expands, it is defined as $-(d_t-V)$.

[0090] In addition, if the lane is already connected based on the autonomous vehicle, it is defined as V, and if the lane does not exist, it is defined as −V.

[0091] The action of the autonomous vehicle at time t is as follows.

$$a_t=\{a_{t,acc},a_{t,lc}\}$$

[0092] Here, $a_{t,acc}$ means acceleration control action, and $a_{t,lc}$ means lane change action.

[0093] Acceleration control action $a_{t,acc}\in[a_{min},a_{max}]$ has values within a continuous range of minimum acceleration a_{min} and maximum acceleration a_{max}.

[0094] Lane change action $a_{t,lc}\in\{-1,0,1\}$ has discrete values, and each value represents the lane change direction of the autonomous vehicle. Specifically, −1 indicates a lane change to the right, 1 indicates a lane change to the left, and 0 indicates lane-keeping action.

[0095] The reward r_t at time t of an autonomous vehicle is defined in the form of a function $r_t=R(s_t,a_t,s_{t+1})$ for the current state (current observation information) s_t, current action a_t, and next state s_{t+1} (next observation information according to the current action), and comprises a linear combination of each reward term and penalty term.

$$R(s_t,a_t,s_{t+1})=\sum_{i=1}^{6}\eta_i R_{t,i}$$

[0096] Here, $R_{t,i\in\{1,\ldots,6\}}$ means the reward term or penalty term, and $\eta_{i\in\{1,\ldots,6\}}$ means the coefficient for each term.

[0097] The first term $R_{t,1}$ is a speed reward term regarding the speed $v_{t+1,N}$ of the autonomous vehicle at time t+1 due

to the acceleration control action $a_{t,acc}$ of the autonomous vehicle at time t. The agent learns the action that does not exceed the speed limit v_{limit} while driving close to the target speed v*, and is defined as follows.

$$R_T,1=\begin{cases}\frac{v_{t+1,N}}{v^*}, & v_{t+1,N}\le v^*\\ \frac{v_{limit}-v_{t+1,N}}{v_{limit}-v^*}, & v_{t+1,N}>v^*\end{cases}$$

[0098] In the first reward term, the autonomous vehicle obtains the maximum positive reward when driving close to the target speed, and when it exceeds the target speed, the positive reward decreases linearly. Additionally, a penalty is imposed if driving exceeds the speed limit.

[0099] $R_{t,2}$ is a penalty term for meaningless lane changes, and is activated only when the autonomous vehicle performs a lane change action and is defined as follows.

$$R_{t,2}=|a_{t,lc}|\min\left(0,\Delta p_{t+1,\hat{l}}-\Delta p_{t,\hat{l}}\right)$$

[0100] Here, $\hat{l}$ refers to the leader vehicle in the same lane as the autonomous vehicle, $\Delta p_{t,\hat{l}}$ refers to the relative distance between the autonomous vehicle and the leader vehicle in the same lane at time t, and $\Delta p_{t+1,\hat{l}}$ refers to the relative distance between the autonomous vehicle and the leader vehicle in the same lane at the next time point (t+1).

[0101] At this time, if the relative distance after the lane change action is reduced compared to the relative distance before the lane change action ($\Delta p_{t+1,\hat{l}}>\Delta p_{t,\hat{l}}$), it is considered a meaningless lane change and a penalty is given.

[0102] FIG. **5** is a diagram showing a case where a penalty is given after changing lanes.

[0103] Autonomous vehicles maintain a safe distance between vehicles and learn safe driving action through $R_{t,3}$ and $R_{t,4}$.

[0104] $R_{t,3}$ guides the autonomous vehicle to drive without violating the safety distance $\delta_{t+1,\hat{l}}^*$ from the leader vehicle in the same lane, $R_{t,4}$ and guides the autonomous vehicle to change lanes without violating the safety distance $\delta_{t+1,\hat{f}}^*$ from the follower vehicle in the same lane.

[0105] FIG. **6** is a diagram for explaining driving action for maintaining a safe distance between vehicles.

[0106] $R_{t,3}$ and $R_{t,4}$ are defined as follows.

$$R_{t,3}=\min\left[0,1-\left(\frac{\delta_{t+1,\hat{l}}^*}{\Delta p_{t+1,\hat{l}}}\right)^2\right]$$

$$R_{t,4}=|a_{t,lc}|\min\left[0,1-\left(\frac{\delta_{t+1,\hat{f}}^*}{\Delta p_{t+1,\hat{f}}}\right)^2\right)$$

[0107] Here, the safety distance $\delta_{t+1,\hat{l}}^*$ and $\delta_{t+1,\hat{f}}^*$ are defined as follows.

$$\delta_{t+1,\hat{l}}^*=\delta_0+\max\left[0,v_{t+1,\hat{l}}\left(t^*+\frac{\left(v_{t+1,\hat{l}}-v_{t+1,N}\right)}{2a_{max}\cdot a_{min}}\right)\right]$$

$$\delta_{t+1,\hat{f}}^*=\delta_0+\max\left[0,v_{t+1,\hat{f}}\left(t^*+\frac{\left(v_{t+1,\hat{f}}-v_{t+1,N}\right)}{2a_{max}\cdot a_{min}}\right)\right]$$

[0108] δ_0 means the minimum safety distance, means the minimum time to prevent an accident, $v_{t+1,\hat{l}}$ and $v_{t+1,\hat{f}}$ mean the absolute speed of the leader and follower vehicles in the same lane at time t+1, respectively.

[0109] $R_{t,5}$ is a merge reward term. According to this embodiment, the autonomous vehicle $R_{t,5}$ weakens the delayed merging action through $R_{t,5}$, which is defined as follows.

$$R_{t,5}=\begin{cases}\mu_{t+1}\times\zeta_{t+1,\hat{h}}, & \zeta_{t+1,\hat{h}}<0\\ 0, & \zeta_{t+1,\hat{h}}\ge 0\end{cases}$$

[0110] Here, $\zeta_{t+1,\hat{h}}\in\zeta_{t+1}$ means the remaining drivable distance from the lane where the autonomous vehicle is located to the merging point.

[0111] In other words, the closer the autonomous vehicle merges to the merging point, the higher the penalty is given as it is considered a delayed merge.

[0112] μ_{t+1} is a traffic density weight, which alleviates the penalty due to $R_{t,5}$ when changing lanes is difficult due to high traffic density in the lane, and is defined as follows.

$$\mu_{t+1}=1-\left(\frac{\sum_{i=1}^{|C_{t+1,Y}|}(e_i+\delta_0)}{Y}\right)$$

[0113] Here, Y means the length of the on-ramp merging zone, $|C_{t+1,Y}|$ means the number of vehicles driving in the main lane of the on-ramp merging zone, and e_i means the length of the ith vehicle in the vehicle set $|C_{t+1,Y}|$ in the main lane of the on-ramp merging zone.

[0114] FIG. **7** is a diagram for explaining traffic density weights according to this embodiment.

[0115] Referring to FIG. **7**, the traffic density weight has a value that is inversely proportional to the number of vehicles in the main lane in the on-ramp merging zone, and alleviates the degree of penalty when changing lanes is difficult due to traffic congestion in the main lane.

[0116] The last term $R_{t,6}$ is a penalty term related to an accident, which imposes a penalty if a vehicle accident occurs and is defined as follows.

$$R_{t,6}=\begin{cases}-1, & \text{Accident}\\ 0, & \text{Otherwise}\end{cases}$$

[0117] Learning of the policy for decision-making by the policy execution unit **104** according to this embodiment can be performed by the autonomous vehicle itself (single

autonomous vehicle environment) or by a central server connected to multiple autonomous vehicles through a network.

[0118] FIG. **8** is a diagram showing the configuration of a system for an autonomous vehicle in a multi-autonomous driving environment according to this embodiment, FIG. **9** is a diagram showing the configuration of an autonomous vehicle in a multi-autonomous driving environment, and FIG. **10** is a diagram showing the configuration of the central server in a multi-autonomous driving environment.

[0119] As shown in FIG. **8**, a plurality of autonomous driving apparatuses **800** are connected to the central server **802** through a network.

[0120] Here, the network may include wired or wireless Internet and mobile communication networks.

[0121] As shown in FIG. **8**, the autonomous driving apparatus **800** according to this embodiment may comprise the information observation unit **100**, the roadside unit communication unit **102**, the policy execution unit **104**, and the reward determination unit **106** of FIG. **1**. In addition, it may comprise a policy communication unit **108**.

[0122] In addition, the central server **802** is connected to multiple autonomous driving apparatuses **800** through a network and may comprise a policy communication unit **1000**, an experience information storage unit **1002**, and an integrated learning unit **1004**.

[0123] As described above, in a multi-autonomous driving environment, the central server **802** is connected to a plurality of autonomous vehicles **800** through a network, and at this time, policy learning is performed in the integrated central server **802**.

[0124] Specifically, the individual autonomous vehicle **800** transmits experience information to the central server **802** through the policy communication unit **108**. Here, the experience information of the individual autonomous vehicle **800** may include current observation information, current action, next observation information, and reward.

[0125] The central server **802** integrates individual experience information received through the policy communication unit **1000** in the experience information storage unit **1002**, and the integrated experience information can be used as base data for later policy learning.

[0126] The integrated learning unit **1004** of the central server **802** updates the decision-making policy for each autonomous vehicle using experience information from multiple autonomous vehicles based on deep reinforcement learning.

[0127] This is not limited to a specific reinforcement learning algorithm and can be comprehensively applied to most algorithms based on deep reinforcement learning methodology. Learning of the policy is repeated a predefined number of times, and the policy updated during the learning process is transmitted to each autonomous vehicle **800** through the policy communication unit **1000**.

[0128] After receiving the updated policy, the autonomous vehicle **800** makes actual driving decisions at the policy execution unit **104** by inputting observation information obtained through the information observation unit **100** and the roadside unit communication unit **102**.

[0129] At this time, interaction with the central server and policy communication unit is not considered in the decision-making stage of the autonomous vehicle. Since the learned policy according to this embodiment considers the adaptive target speed in the learning stage, flexible decision-making is possible even in irregular road congestion.

[0130] Meanwhile, in a single-driving vehicle environment, the policy communication unit **108** as shown in FIG. **9** is not included, and policy learning can be performed independently by providing a policy learning unit.

[0131] FIG. **11** is a diagram illustrating a policy learning process in a single autonomous vehicle environment according to this embodiment.

[0132] FIG. **11** shows the process of learning a policy by an autonomous vehicle itself. Referring to FIG. **11**, the apparatus according to this embodiment initializes the model (step **1100**), initializes the driving environment, and collects initial observation information (step **1102**).

[0133] Here, the observation information may be information collected from a sensing module of an autonomous vehicle or a roadside unit.

[0134] Afterwards, an action is decided based on the collected observation information (step **1104**).

[0135] The action decision according to this embodiment includes the acceleration control and the lane change direction determination of the autonomous vehicle in the on-ramp merging zone.

[0136] The next observation information is changed by the action decided in step **1104**, and the next observation information is collected accordingly (step **1106**).

[0137] A reward is determined based on the current observation information in steps **1104** and **1106**, current action according to the current observation information, and next observation information (step **1108**), and the driving policy for decision-making is updated according to the experience information including the determined reward (step **1110**).

[0138] The apparatus according to this embodiment determines whether the number of learning times is met (step **1112**) and ends learning.

[0139] FIG. **12** is a flowchart showing the policy learning process in a multi-autonomous driving environment according to this embodiment.

[0140] FIG. **12** is a diagram showing the process, in which each autonomous vehicle collects observation information using an initial policy, determines actions and rewards, and then updates the policy at each autonomous vehicle and a central server connected through a network.

[0141] Referring to FIG. **12**, after performing steps **1100** to **1108** in FIG. **11**, the experience information is transmitted to the central server **802** through the policy communication unit **108** (step **1200**).

[0142] Here, the experience information may include current observation information from each autonomous vehicle, current actions according to the current observation information, next observation information after the current action, and reward using these.

[0143] The central server **802** stores experience information (step **1202**), samples some of it, and updates the policy for driving the autonomous vehicle (step **1204**).

[0144] The update of the policy can be performed repeatedly until the preset number of learning times is met.

[0145] The central server **802** transmits the updated policy to the autonomous vehicles (step **1206**), and each autonomous vehicle performs a decision-making process through the updated policy (step **1208**).

[0146] Table 1 shows simulation results for the decision-making process according to this embodiment and the conventional decision-making process.

TABLE 1

Environment		Average speed(km/h) On-ramp AV	Main lane NAV
RL-based	PPO	47.964 ± 3.064	44.813 ± 0.118
	DDPG	47.146 ± 1.617	44.772 ± 0.235
	TD3	46.626 ± 1.074	44.872 ± 0.078
Control-theoretic		43.684 ± 0.641	43.792 ± 0.23

[0147] In Table 1, RL-based represents an autonomous vehicle environment that performs learned policies, and Control-theoretic represents a control theory-based autonomous vehicle environment that does not use learned policies.

[0148] In RL-based, PPO stands for Proximal Policy Optimization, DDPG stands for Deep Deterministic Policy Gradient, and TD3 stands for Twin Delayed DDPG.

[0149] Additionally, On-ramp AV refers to autonomous vehicles in the on-ramp, and Main lane NAV refers to non-autonomous vehicles in the main lane. Referring to Table 1, it can be seen that the RL-based environment according to this embodiment provides an autonomous driving system that minimizes disruption of traffic flow in the main lane because the average speed of non-autonomous vehicles in the main lane is higher compared to the control-theoretic environment.

[0150] The aforementioned vehicle action decision method based on deep reinforcement learning for the merging strategy of autonomous vehicles in the on-ramp merging zone may be also implemented in the form of a recording medium containing instructions executable by a computer, such as an application or program module executed by a computer. Computer-readable media can be any available media that can be accessed by a computer and includes both volatile and non-volatile media, removable and non-removable media. Additionally, computer-readable media may include computer storage media. Computer storage media includes both volatile and non-volatile, removable and non-removable media implemented in any method or technology for storage of information such as computer-readable instructions, data structures, program modules, or other data.

[0151] The above-described embodiments of the present invention have been disclosed for illustrative purposes, and those skilled in the art will be able to make various modifications, changes, and additions within the spirit and scope of the present invention, and such modifications, changes, and additions should be regarded as falling within the scope of the patent claims below.

1. A deep reinforcement learning-based vehicle action decision apparatus for merging strategy of an autonomous vehicle in an on-ramp merging zone comprising:

an information observation unit for collecting observation information from a sensing module or roadside unit (RSU) of an autonomous vehicle;

a policy execution unit for making a decision regarding a current action, including acceleration control and lane change of the autonomous vehicle, based on the current observation information and policy; and

a reward determination unit for determining a reward according to the current observation information, the current action, and the next observation information according to the current action,

wherein reward in the reward determination unit is determined through a reward term related to speed, lane change, safety distance compliance, and an accident of the autonomous vehicle and a merge reward term related to merge of an autonomous vehicle in the on-ramp merging zone.

2. The apparatus of claim **1**, wherein the observation information comprises at least one of an absolute speed of the autonomous vehicle, a relative speed between the closest leader vehicle in front and the closest follower vehicle in rear among vehicles in each lane that the autonomous vehicle can observe, a relative distance between the leader vehicle and the follower vehicle, a vehicle density for each lane within front observation range of the autonomous vehicle, and presence or absence of a lane within front observation range of the autonomous vehicle.

3. The apparatus of claim **1**, wherein the merge reward term is determined using a remaining drivable distance from an on-ramp where the autonomous vehicle is located to a merging point and traffic density.

4. The apparatus of claim **3**, wherein the closer the autonomous vehicle merges to the merging point, the reward determination unit gives the higher penalty as it is considered a delayed merge.

5. The apparatus of claim **4**, wherein the reward determination unit alleviates the penalty using a traffic density weight according to a traffic density of a main lane of the on-ramp merging zone.

6. The apparatus of claim **5**, wherein the traffic density weight is determined using at least one of a length of the on-ramp merging zone, the number of vehicles driving in the main lane of the on-ramp merging zone, a length of a vehicle belonging to a vehicle set in the main lane of the on-ramp merging zone, and a minimum safety distance.

7. The apparatus of claim **1**, wherein the merge reward term is

defined by the following equation,

$$R_{t,5} = \begin{cases} \mu_{t+1} \times \zeta_{t+1,\hat{h}}, & \zeta_{t+1,\hat{h}} < 0 \\ 0, & \zeta_{t+1,\hat{h}} \geq 0 \end{cases} \quad \text{[Equation]}$$

wherein $\zeta_{t+1,\hat{h}} \in \zeta_{t=1}$ is the remaining drivable distance from an on-ramp where the autonomous vehicle is located to a merging point, and μ_{t+1} is a traffic density weight.

8. The apparatus of claim **7**, wherein the traffic density weight is defined by the following equation,

$$\mu_{t+1} = 1 - \left(\frac{\sum_{i=1}^{|C_{t+1,Y}|} (e_i + \delta_0)}{Y} \right) \quad \text{[Equation]}$$

wherein Y is a length of the on-ramp merging zone, $|C_{t+1,Y}|$ is the number of vehicles driving in a main lane of the on-ramp merging zone, e_i is a length of the ith vehicle in a set of vehicles $|C_{t+1,Y}|$ in a main lane of the on-ramp merging zone, and δ_0 is a minimum safety distance.

9. The apparatus of claim **1**, wherein the learning of the policy is performed by the autonomous vehicle itself or by a central server connected to multiple autonomous vehicles through a network.

10. The apparatus of claim **9**, when the learning of the policy is performed by the autonomous vehicle itself, further comprises,

a policy learning unit for updating the policy according to the collected observation information, the current action, the next observation information, and experience information including the reward, and determining whether the number of learning times is met to learn the policy.

11. The apparatus of claim **9**, when the learning of the policy is performed by the central server, further comprises,

a policy communication unit for transmitting experience information including the collected current observation information, the current action, the next observation information, and the reward to the central server, and receiving a policy that the central server has learned using experience information received from the multiple autonomous vehicles.

12. A deep reinforcement learning-based vehicle action decision apparatus for merging strategy of an autonomous vehicle in an on-ramp merging zone comprising:

a processor; and

a memory connected to the processor,

wherein the memory stores program instruction, when executed by the processor, configured to perform operations comprising,

collecting observation information from a sensing module or roadside unit (RSU) of an autonomous vehicle;

making a decision regarding a current action, including acceleration control and lane change of the autonomous vehicle, based on the current observation information and policy; and

determining a reward according to the current observation information, the current action, and the next observation information according to the current action,

wherein a reward is determined through a reward term related to speed, lane change, safety distance compliance, and an accident of the autonomous vehicle and a merge reward term related to a merge of an autonomous vehicle in the on-ramp merging zone.

13. A deep reinforcement learning-based vehicle action decision method for merging strategy of an autonomous vehicle in an on-ramp merging zone comprising:

collecting observation information from a sensing module or roadside unit (RSU) of an autonomous vehicle;

making a decision regarding a current action, including acceleration control and lane change of the autonomous vehicle, based on the current observation information and policy; and

determining a reward according to the current observation information, the current action, and the next observation information according to the current action,

wherein reward is determined through a reward term related to speed, lane change, safety distance compliance, and an accident of the autonomous vehicle and a merge reward term related to a merge of an autonomous vehicle in the on-ramp merging zone.

14. The method of claim **13**, wherein the merge reward term is determined according to at least one of the remaining drivable a distance from an on-ramp where the autonomous vehicle is located to a merging point and traffic density.

15. The method of claim **14**, wherein the determining the reward comprises,

the closer the autonomous vehicle merges to the merging point, giving the higher penalty as it is considered a delayed merge.

* * * * *