Journal of Korea Robotics Society (2017) 12(3):297-305
https://doi.org/10.7746/jkros.2017.12.3.297 ISSN: 1975-6291 / eISSN: 2287-3961

실시간 장애물 회피 자동 조작을 위한 차량 동역학 기반의 강화학습 전략

Reinforcement Learning Strategy for Automatic Control of Real-time Obstacle Avoidance based on Vehicle Dynamics

강 동 훈[1], 봉 재 환[2], 박 주 영[3], 박 신 석[†]

Dong-Hoon Kang[1], Jae Hwan Bong[2], Jooyoung Park[3], Shinsuk Park[†]

Abstract As the development of autonomous vehicles becomes realistic, many automobile manufacturers and components producers aim to develop 'completely autonomous driving'. ADAS (Advanced Driver Assistance Systems) which has been applied in automobile recently, supports the driver in controlling lane maintenance, speed and direction in a single lane based on limited road environment. Although technologies of obstacles avoidance on the obstacle environment have been developed, they concentrates on simple obstacle avoidances, not considering the control of the actual vehicle in the real situation which makes drivers feel unsafe from the sudden change of the wheel and the speed of the vehicle. In order to develop the 'completely autonomous driving' automobile which perceives the surrounding environment by itself and operates, ability of the vehicle should be enhanced in a way human driver does. In this sense, this paper intends to establish a strategy with which autonomous vehicles behave human-friendly based on vehicle dynamics through the reinforcement learning that is based on Q-learning, a type of machine learning. The obstacle avoidance reinforcement learning proceeded in 5 simulations. The reward rule has been set in the experiment so that the car can learn by itself with recurring events, allowing the experiment to have the similar environment to the one when humans drive. Driving Simulator has been used to verify results of the reinforcement learning. The ultimate goal of this study is to enable autonomous vehicles avoid obstacles in a human-friendly way when obstacles appear in their sight, using controlling methods that have previously been learned in various conditions through the reinforcement learning.

Keywords Reinforcement Learning, Obstacle Avoidance

Received : Mar. 7. 2017; Revised : May. 12. 2017; Accepted : May. 15. 2017
※This work was financially supported by the Agency for Defense Development (ADD) under the contract UD140073ID and the Technology Innovation Program (No. 2017-10069072) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea).
The work of Dong-Hoon Kang was partially supported by Hyundai Motor Company.
†Corresponding author: Mechanical Engineering, Korea University, Anam-dong 5-ga, Seongbuk-gu, Seoul, Korea (drsspark@korea.ac.kr)
[1]Automotive Convergence Engineering, Korea University (dongguss@korea.ac.kr)
[2]Mechanical Engineering, Korea University (delitian@korea.ac.kr)
[3]The Department of Control and Instrumentation Engineering, Korea University (parkj@korea.ac.kr)

1. 서 론

과거의 자동화 기술은 사람의 반복 작업을 대신 하는 단순 자동화 기술 위주였으나, 현재의 자동화 기술 개발 방향은 실생활에 밀접한 부분으로 그 영역을 넓히고 있다. 자동차의 자율 주행 기술 개발도 이러한 맥락에서 하나의 중요한 자동화 기술 분야로 대두되고 있다. 자동

차의 자율 주행 기술 개발은 주행 중 운전자의 페달 및 핸들 조작과 같은 단순 작업에서 운전자를 자유롭게 해 줄 수 있을 뿐만 아니라 운전자의 부주의로 인한 실수를 줄이고 도로 환경에 따른 최적화된 차량제어를 통해 사고를 미연에 방지할 수 있다. 자율주행에 대한 가이드라인을 구축하기 위해 미국 도로교통 안전국(National Highway Traffic Safety Administration, NHTSA)에서는 총 5가지 단계로 자동차 시스템을 분류하였다[1]. 0~4단계 중 3단계 이상부터 자율주행 자동차로 인지되며 4단계는 운전자의 차량 통제 없이 모든 제어를 차량이 스스로 담당하는 단계로 자동차 및 부품업체에서는 4단계를 목표로 연구를 진행하고 있다. 현재 자동차에 적용되고 있는 운전자 보조 시스템(Advanced Driver Assistance Systems, ADAS)의 경우, NHTSA 기준 2,3 단계의 기술로써 단일 차선에서 차량의 속도 및 방향을 제어하여 노면 위 장애물 상황 및 차선 변경과 같은 다양한 도로 환경에 적용이 힘들다는 한계점이 있다. 노면에 장애물이 놓인 상황에서 운전자의 장애물 회피 조작을 도와주기 위한 Evasive Steering이란 기술이 연구 되었지만 급격한 방향 전환으로 인해 운전자에게 불안감을 야기하는 문제점이 있다. 또한 장애물 환경에서 해당 장애물을 회피하기 위한 방법으로 Fuzzy Logic Control[2], Potential Field Method[3], Genetic Algorithm[4]의 방법이 사용되었지만, 경로 생성에만 초점이 맞춰져 있으며 실제 장애물 상황에서 사람이 차량을 조작하는 요인들을 고려하지 않고 수식을 통한 계산 결과로만 장애물 회피 경로를 생성하기 때문에 사람이 탑승하는 자동차에 적용함에 있어 사용자의 기능 거부감 및 불안감을 야기할 수 있다.

NHTSA 기준으로 4단계에 해당하는 완전 자율 주행 자동차를 구현하기 위해서는 기존 기술들의 한계점을 보완하여 다양한 장애물 환경과 도로 환경에 대해서도 운전자가 거부감이나 불안감을 느끼지 않도록 최적화된 차량제어를 수행할 수 있어야 한다. 본 논문에서는 기계학습 중 하나인 강화학습에 차량 동역학 요소를 반영하여 장애물 등의 다양한 도로 환경에서도 실제 사람이 운전하는 것과 유사한 사용자 친화적인 차량 조작을 구현하였다. 이를 통해 장애물 환경에서도 실제 운전자가 차량을 운전하는 것과 같은 차량 거동을 보이는 자율주행 제어 전략을 제시하였다.

2장에서는 기존의 장애물 회피 방법에 대해 소개한 후, 기계학습 방법 중 하나인 강화학습에 대한 설명과 Q-Learning에 대한 적용 사항을 언급한다. 3장에서는 장애물 상황에 대한 강화학습 결과와 Driving Simulator 결과를 설명하고 마지막으로 결론에서 이를 비교하여 강화학습 적용 가능성을 기술하였다.

2. 자율주행 제어 전략

먼저, 이 장에서는 장애물 회피와 관련된 기존의 회피 경로 생성 방법의 한계점을 언급한 후, 이를 보완할 방법으로 강화학습 법의 일종인 Q-learning에 대해서 언급한다.

2.1 기존 경로 생성 알고리즘

장애물 회피와 관련하여 Fuzzy Logic Control, Potential Field Method, Genetic Algorithm과 같은 알고리즘이 대표적으로 사용된다.

먼저 Fuzzy Logic Control의 경우, 1964년 Zedeh 교수가 제안한 이론으로 불분명한 상태, 모호한 상태를 이진 논리에서 벗어나 다치성으로 표현하는 논리 개념이다[2,5]. Fuzzy Algorithm의 경우 각 대상이 해당 집합(fuzzy set)에 속하는 정도를 소속함수로 나타내고 그 소속함수에 대응되는 대상과 함께 표기를 한다.

Potential Field Method의 경우, 주행 제어에 사용되는 대표적인 알고리즘으로 장애물로부터 이동 대상을 밀어내는 가상의 힘을 생성하고 목표 지점으로 당기는 힘을 생성하여 최종적으로 해당 목표 지점에 도달하는 주행 알고리즘이다. 위 알고리즘의 경우, 간단한 구조의 이론이라는 장점이 있지만 장애물을 회피하는 경로 생성만을 고려하고, 속도 제어에 대한 고려는 없다는 한계점이 있다[6].

Genetic Algorithm의 경우, 적자 생존과 같은 자연 생태계의 진화 현상과 유전학에 근거한 계산 모델이다. 해를 나타내는 개체들로 개체군을 형성한 후, 그 개체군에 대하여 각 개체의 적합도에 따라 유전 연산자를 적용한다[4]. 이 후 다음 세대의 개체군을 생성하는 과정을 반복함으로써 전체적으로 우수한 해들로 진화시키는 방법이다.

위 3가지 방법들은 모두 경로 생성에만 집중되어 있는 방안이며 다양한 장애물 상황에서 적용하기 힘들다는 문제점이 있다. 또한 경로 생성에 초점을 두었기 때문에 차량의 동역학적 특성을 고려하지 않았다는 문제점이 있으며 실제 운전자가 각 도로 및 장애물 상황에 대해서 운전시 중요하게 여기는 요소들을 고려하지 않은 차량 제어로 인해 실 적용에 있어서 사용자에게 거부감을 준다는 문제점이 있다. 위 문제점을 해결하기 위해서 다양한 환경에서 시스템이 스스로 학습을 통해 행동을 결정하는 강화학습을 적용하고자 하였다.

2.2 강화학습

본 논문에서는 기존의 장애물 회피 기술들의 한계점을 극복하고자 기계학습 중 하나인 강화학습을 자율 주행에 적용하였다. 강화학습은 환경 정보를 정확히 알기 힘든 로봇 제어에서 많이 사용되고 있는 방법이다[7]. 강화학습은 환경과 에이전트 사이의 상호 정보 교환을 통해 에이전트의 각 상태에 대해서 미지의 환경에 대한 행동을 결정한다.

Fig. 1은 강화학습의 개념도이다. 강화학습에서는 학습대상을 에이전트라 하며, 에이전트 외부에 위치한 환경과 정보를 상호 교환한다[8]. 위 과정에서 에이전트는 해당 시간(t)에 해당하는 행동(a_t)을 선택하고 행동의 결과로써 환경으로부터 상태 정보(s_t)를 얻게 된다. 에이전트는 행동의 결과로 보상(r_t)을 받게 된다. 보상(r_t)은 선택한 행동이 좋은 행동인지 혹은 아닌지에 대해서 스칼라 값으로 표현된다. 에이전트는 행동선택을 조정함으로써 보상 값을 최대화 할 수 있도록 노력한다.

시간 t 이후의 보상 값들의 집합을 다음과 같이 $r_t, r_{t+1}, r_{t+2}, \cdots$로 표현한다. 강화학습에서 학습의 목표는 보상 값들의 집합에 속하는 원소들의 누적 합이 최대가 되도록 하는 것이다. 이를 리턴(R_t)이라 하며 일반적으로 할인율이 적용된 보상 값들의 합의 형태로 아래와 같이 표현된다.

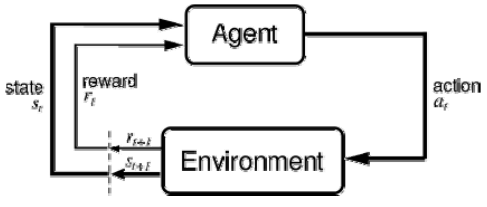

Fig. 1. The agent-environment interaction in reinforcement learning

$$R_t = r_{t+1} + \gamma^1 \cdot r_{t+2} + \gamma^2 \cdot r_{t+3} + \cdots + \gamma^T \cdot r_{t+T+1} = \sum_{k=o}^{T} \gamma^k \cdot r_{t+k+1} \quad (1)$$

여기서 γ는 0과 1 사이의 값을 가지며, 미래의 보상이 현재에 얼마나 가치가 있는지를 표현하는 할인율을 의미한다. 에이전트가 수립된 전략(π)에 의해서 행동을 선택하면, 보상들의 기대 합을 나타내는 가치 함수(value function, $V^{\pi}(s)$)는 상태정보 s_t에 대해서 식 (2)와 같이 표현된다.

$$V^{\pi}(s) = E_{\pi}\{R_t | s_t = s\} = E_{\pi}\left\{\sum_{k=0}^{T} \gamma^k \cdot \sum r_{t+k+1} | s_t = s\right\} \quad (2)$$

강화학습에서는 반복 시행을 통해 가치함수를 갱신하면서 최대값으로 수렴할 때까지 진행한다.

2.3 Q-Learning

Q-learning은 강화학습의 일종으로 보상 값 정보를 활용하여 최적 제어 전략을 구축하기 위해 사용되는 방법이다[9]. Q-learning은 적용성 및 이식성이 우수하며 빠른 연산속도를 갖고 있어 다양한 장애물 환경에서의 차량의 회피 조작 학습 기법으로 선택하였다. 앞서 강화학습 구조에서 설정한 목표를 에이전트가 학습 과정을 통해 달성하기 위해서는 모든 상태 s에 해당하는 $V^{\pi}(s)$를 최대화 해주는 전략 π를 에이전트가 학습하는 과정이 필요하다. 최적 전략을 π^*로 기호화 하면 식 (5)와 같이 표현할 수 있다.

$$\pi^* \equiv \underset{\pi}{arg\max} V^{\pi}(s), \ \forall s \quad (3)$$

최적 전략의 가치 함수인 $V^{\pi^*}(S)$를 $V^*(s)$로 간단히 표현한다고 했을 때, $V^*(s)$는 에이전트가 해당 상태(s)

에서 획득한 할인율이 적용된 누적 보상 값들의 최대값을 제공한다. 할인율이 적용된 누적 보상 값들의 최대값을 표현하기 위해서 다음과 같이 $Q(s,a)$을 설정한다. 여기서 Q 값은 상태에 대한 행동이 실행될 때 받게 되는 보상 값을 의미하며 아래와 같이 표현된다.

$$Q(s,a) \equiv r(s,a) + \gamma V^*(\delta(s,a)) \tag{4}$$

식 (4)에서 $Q(s,a)$는 상태에 대한 최적 행동을 선택하기 위한 최대 값을 의미하기 때문에 다음과 같은 관계식을 유도 할 수 있다.

$$\pi^*(s) = \underset{a}{arg\max} Q(s,a) \tag{5}$$

식 (5)를 통해 에이전트가 V^*을 아는 것 대신에 Q 값을 통해 학습할 수 있다는 것을 알 수 있다. 즉, 현재 상태에 대해서 에이전트의 각 가능 행동과 $Q(s,a)$을 최대로 만들어주는 행동을 선택하면 된다는 것을 의미한다. 이것이 Q-learning의 핵심이며 다음과 같은 규칙을 따른다. 여기서 Q는 학습자가 추정한 것이 되거나 실제 Q 함수의 가정이 된다. Q는 상태와 행동의 쌍으로 표현되는 표 형태로 구축된다. 에이전트는 현재 상태를 주기적으로 관찰하면서 행동을 선택하고 실행하면서 보상인 $r = r(s,a)$와 새로운 상태인 s'를 얻게 된다. 그 다음 해당 순서인 테이블의 $\overline{Q}(s,a)$값을 업데이트 한다. 위 과정을 실행할 때 식 (6)과 같은 규칙을 따른다.

$$Q(s,a) \leftarrow r(s,a) + \underset{a'}{\gamma \max} Q(s',a') \tag{6}$$

3. 장애물 환경에 대한 강화학습

3.1 장애물 환경 및 가정사항

본 논문에서는 다음과 같은 가정을 설정하였다. 차량이 Fig. 2의 개념도에 묘사된 것과 같은 차량 센서부의 정보를 바탕으로 장애물을 인지했다고 가정하였다. 이를 통해 장애물의 종류와 위치를 차량이 알고 있으며 차량의 경우, 정밀한 GPS를 가지고 있어 도로 위에 정확

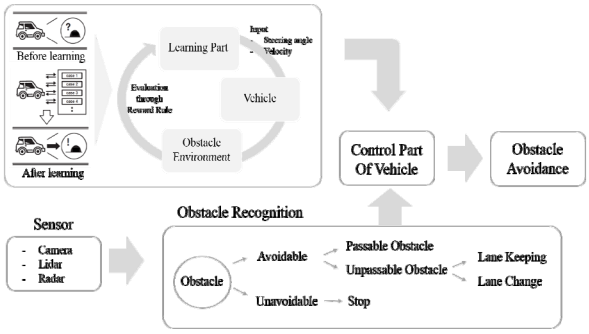

Fig. 2. Conceptual diagram

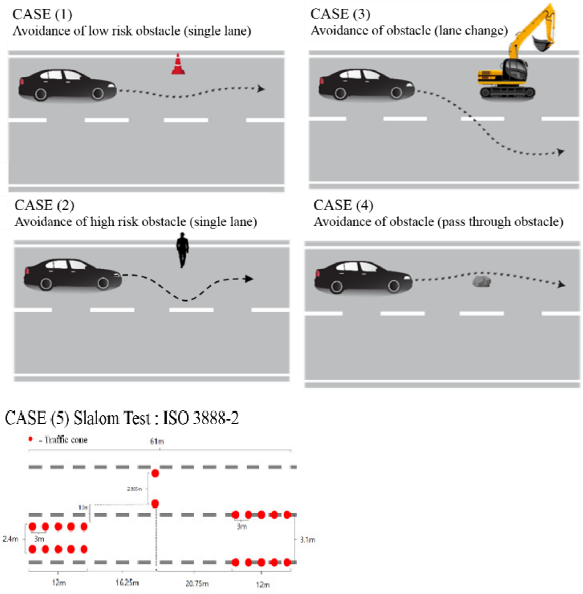

Fig. 3. Case of obstacles on the road environment

한 위치를 파악할 수 있다고 가정하였다. 위 정보를 바탕으로 도로 환경에서 차량과 장애물을 위치시키는 것으로 환경 설정을 하였다. 노면 위 장애물 상황의 경우, 장애물 상황에 대한 규정이 정해져 있지 않아 장애물 상황에 대해 가정한 네 가지 도로 환경과 ISO 3888-2의 Slalom test[10] 상황을 구축하여 실험을 진행하였다. 구축한 다섯 가지 장애물 상황은 Fig. 3과 같다.

3.2 CarSim-Simulink 연동

Fig. 3과 3.1절에서 묘사한 장애물 환경을 자동차 시뮬레이션인 CarSim (Mechanical Simulation Corperation, USA)에서 도로 환경으로 설정하였다. 차량 모델로는

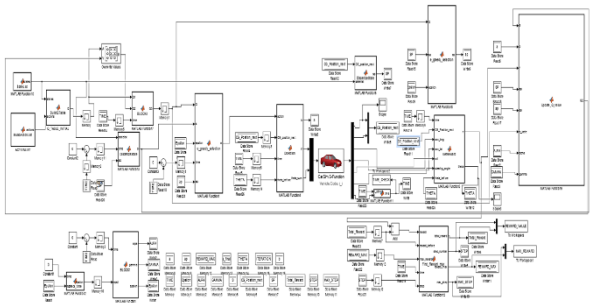

Fig. 4. CarSim-Simulink linkage

Carsim에서 제공하고 있는 15자유도 모델인 C-Class Hatchback을 사용하였다(Spring mass-6DOF, each Suspension-2DOF, each Wheel-1DOF, Steering-1DOF). 이러한 설정을 통해 15자유도 다물체 동역학 모델을 Carsim에서 자동 생성해주고, 이를 수치해석적인 방법으로 풀이해준다.

강화학습과 CarSim의 차량 모델 구성환경을 연동하기 위해서 Simulink (Mathworks, USA) 를 사용하였다. CarSim과 Simulink를 연동하기 위해 Simulink에 CarSim 라이브러리 경로를 추가한 후, 강화학습 알고리즘을 Simulink의 Function box에 구현하여 CarSim에 Fig. 4와 같이 연동하였다. 이때 Carsim 모델의 입력 값으로 차량 속도와 조향각을 사용하였고, 출력 값으로 차량 무게중심의 좌표, heading값, 횡/종 가속도, 차량의 roll, 위험지역 도달 여부를 산출하여 학습을 진행하였다.

Fig. 3의 장애물 환경에 따라 강화학습의 학습 및 적용 과정을 크게 두 부분으로 나누어 3.2.1절과 3.2.2절에서 설명하였다.

3.2.1 Q-Table 설정 및 차량 조작 입력 값 선택

이 절에서는 Q-learning을 기반으로 한 강화학습에 필요한 환경 구축과 차량의 조작 입력 값을 선택하는 방법에 대해서 설명한다. 차량의 무게중심 좌표, 장애물과의 거리, 차량의 heading angle, 위험지역 접근 여부와 같은 차량 상태와 입력 값인 차량 속도 및 조향각의 조합으로 인덱스를 구축하였다. 이를 바탕으로 Q-Table을 만들었다. 초기 Q-Table은 각 인자를 0으로 초기화 하고 강화학습을 진행하면서 각 인자의 값을 업데이트하였다. 조작 입력 값을 선택하기 위해서는 ε-greedy selection[11] 기법을 적용하였다. 위 기법을 장애물 회피 조작에 적용하여 Q-Table에서 Q-value를 파악하여 그 순간 차량의 상태에 적합한 조작 입력 값을 선택하여 CarSim의 차량 모델에 적용하였다.

3.2.2 출력 값에 대한 보상평가 및 Q-value 업데이트

CarSim의 차량모델에 입력된 조향각과 속도에 대한 출력으로 차량의 무게중심의 좌표, 횡/종 가속도, 롤, 차량 바퀴 각도, 시간 성분을 전달받는다. 이렇게 받은 정보를 토대로 구축한 보상 정책과 비교하여 행동 평가를 진행하고 이를 반복적으로 진행함으로써 사람이 해당 장애물 상황에서 차량을 조작하는 것과 유사한 결과를 구축할 수 있도록 알고리즘을 구축하였다. 이때 사람의 거동과 유사한지 아닌지를 판단하기 위해 사람은 장애물을 회피할 때는 차량의 속도가 감소하도록 조작하고 장애물을 회피한 후에는 차량의 속도가 증가하도록 조작한다는 가설을 세웠다. 이러한 가설과 선택한 조작 입력 값을 차량에 적용했을 때의 차량의 행동을 평가하기 위해 총 6쌍의 보상 기준을 Table 1과 같이 설정하였다. 한 쌍의 보상 기준은 가중치(Weight)와 보상 요소

Table 1. Reward index for reward evaluation

Weights		Reward Components	
W_{LL}	10	R_{LL}	$(-\frac{1}{D_{LL}})$
W_{RL}	1	R_{RL}	$(-\frac{1}{D_{RL}})$
W_{Obs}	20	R_{Obs}	$(-\frac{1}{D_{Obs}-R_{CS}})$
W_{Vel}	15	R_{Vel}	-T
W_{Roll}	20	R_{Roll}	$(-\frac{1}{CA-\theta_R})$
W_{TC}	20	R_{TC}	$(-\frac{1}{1g-A})$

- D_{LL} : DistancetoLeftLane
- D_{RL}: DistancetoRightLane
- D_{Obs}: DistancetoObstacle
- R_{CS}: RadiusofCriticalSection
- T: Obstacle Avoidance Time
- CA: Critical Angle
- θ_R: Roll Angle
- A: Acceleration of the tire

(Reward Component)로 구성된다. 주어진 장애물 상황에 대한 차량의 행동을 평가하기 위한 보상 값은 식 (7)과 같이 총 6쌍의 가중치와 보상 요소의 곱을 합하여 계산된다.

$$\begin{aligned}\text{Reward} = & W_{LL} \cdot R_{LL} + W_{RL} \cdot R_{RL} + W_{Obs} \cdot R_{Obs} \\ & + W_{Vel} \cdot R_{Vel} + W_{Roll} \cdot R_{Roll} + W_{TC} \cdot R_{TC}\end{aligned} \tag{7}$$

식 (7)에서 보상 값 계산을 위해 사용된 가중치와 보상 요소의 상세 내용은 아래 Table 1과 같다. 총 6쌍의 가중치와 보상 요소는 다음의 6가지 차량 행동을 식 (7)의 보상 값에 반영하기 위해 설정되었다: 왼쪽 차선에 접근($W_{LL} \cdot R_{LL}$), 오른쪽 차선에 접근($W_{RL} \cdot R_{RL}$), 장애물 영역에 접근($W_{Obs} \cdot R_{Obs}$), 장애물 상황에서 속도($W_{Vel} \cdot R_{Vel}$), Roll 임계각에 접근($W_{Roll} \cdot R_{Roll}$), Traction circle 경계에 접근($W_{TC} \cdot R_{TC}$).

Table 1의 보상 요소(Reward Components)를 계산하기 위해 사용된 변수들은 Fig. 5와 같이 측정되었으며, 이때 T는 장애물 회피에 걸린 시간을 측정하였고 Roll 임계각(Critical Angle)은 차량 모델의 무게중심의 수직 위치와 윤거 폭을 사용하여 계산하였다. Traction circle[12]

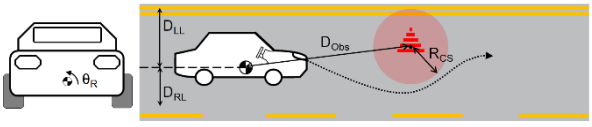

Fig. 5. Components of reward evaluation: Roll Angle (θ_R), Distance to Left Lane (D_{LL}), Distance to Right Lane (D_{RL}), Distance to Obstacle (D_{Obs}), Radius of Critical Section (R_{CS})

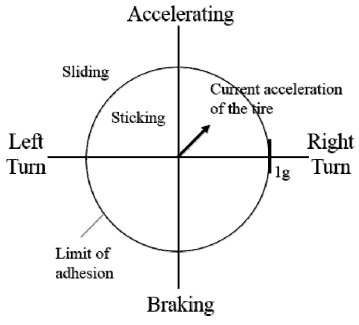

Fig. 6. Traction circle

은 Fig. 6과 같은 타이어의 접촉 면적을 중심으로 타이어의 한계 그립을 나타낸 원 그래프에서 접촉 한계선에 접근할수록 작은 Reward를 부여하였다.

보상 결과를 바탕으로 Q-Value를 업데이트하고 해당 결과를 학습 이전에 구축한 Q-Table에 적용한다. 위 과정을 계속적으로 반복하여 해당 장애물 환경에 부합하는 차량 조작법을 구축하였다

3.3 강화학습 결과

CarSim과 Simulink를 연동하여 구축한 네 가지 실험 환경과 ISO 3888-2 규격의 Slalom test 환경을 구축하여 총 다섯 가지 도로 상황에 대한 장애물 회피 강화학습을 적용하였다.

먼저 Case(1)과 (2)는 단일 차선 내에서 장애물을 회피하는 경우로, (1)의 경우 위험도가 낮은 장애물이며 (2)의 경우 위험도가 큰 장애물로써 장애물 영역의 위험도를 다르게 설정하여 강화학습을 진행하였다.

다음으로 큰 장애물이 도로에 있어 차선을 변경하여 회피하는 경우인 Case(3)에 대해서 강화학습을 진행하였다. 차선을 변경하여 회피한 후 다시 원래 차선으로 돌아오도록 실험을 진행하였다.

Case(4)의 경우는 노면 위 차량이 통과 가능한 작은 장애물 및 요철이 있는 경우로 장애물 회피에 치중을 둔 보수적인 회피 전략을 사용할 경우 사용자에게 과도한 응답으로 보여줄 수 있기 때문에 통과하여 회피하도록 강화학습을 진행하였다.

마지막으로 Case(5)의 경우, ISO 3888-2의 Slalom test 규격에 맞춰 교통 콘을 배치한 후 교통 콘을 부딪치지 않고 설정 코스를 통과하는 실험을 진행하였다. 각 장애물 상황 별 강화학습을 적용한 결과는 Fig. 7에서 Fig. 11에 점선으로 표시하였다. Fig. 7에서 Fig. 11의 그래프에서 y축 데이터는 강화학습결과와 Driving Simulator 결과를 비교하기 위해 식 (8)을 이용해 Normalize 하였다.

$$y_N = y / Max(|y|) \tag{8}$$

y_N: Normalized Data

$Max(|y|)$: 절대값을 취한 Data의 최대값

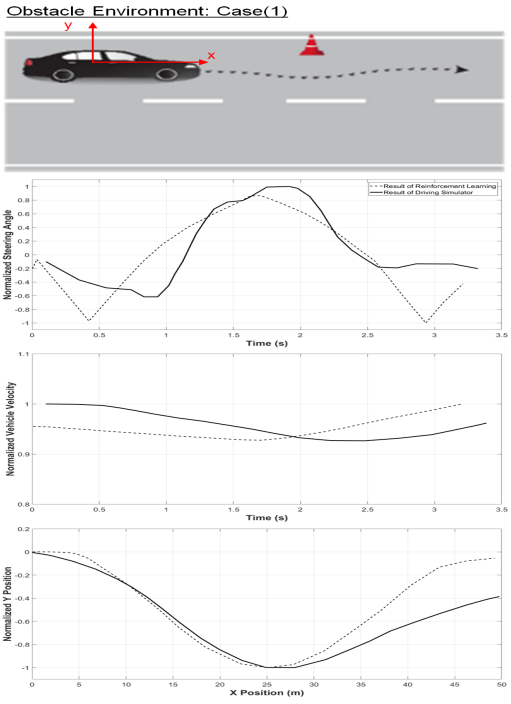

Fig. 7. Comparative on the result of reinforcement learning & driving simulator based on Case (1): Avoidance result of low risk obstacle (single lane)

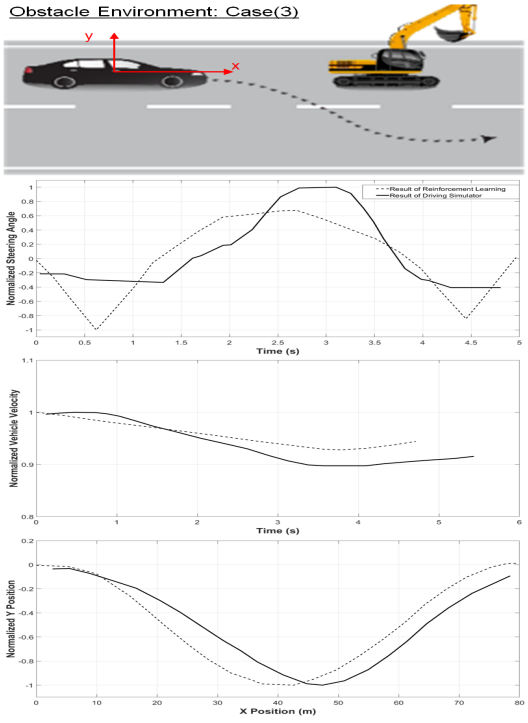

Fig. 9. Comparative on the result of reinforcement learning & driving simulator based on Case (3): Avoidance result of obstacle (lane change)

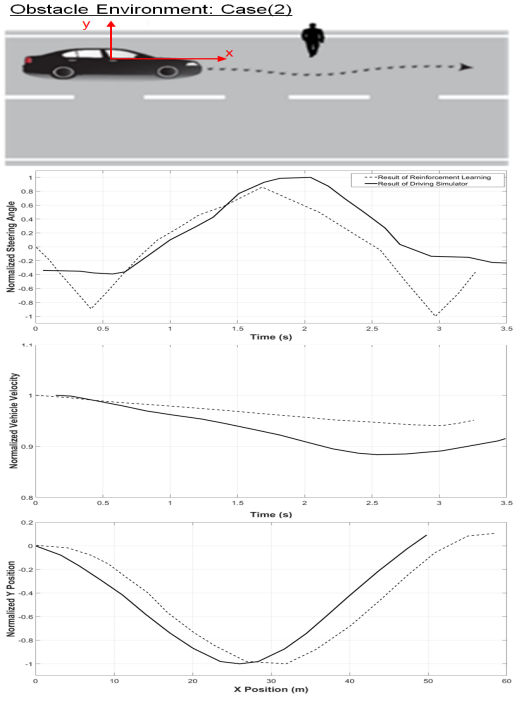

Fig. 8. Comparative on the result of reinforcement learning & driving simulator based on Case (2): Avoidance result of high risk obstacle (single lane)

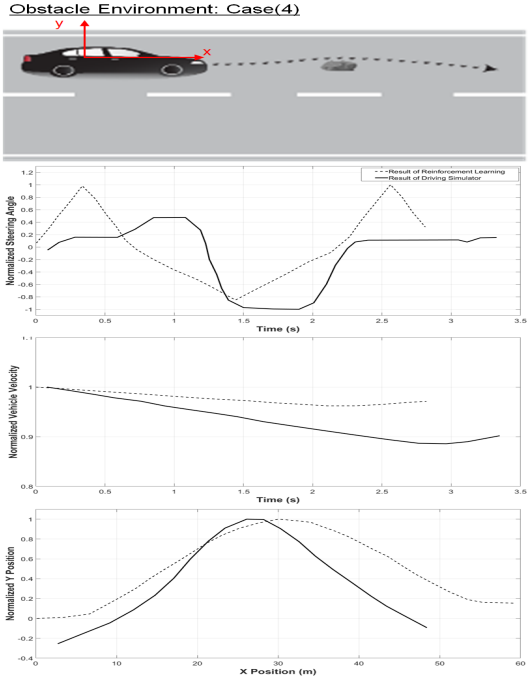

Fig. 10. Comparative on the result of reinforcement learning & driving simulator based on Case (4): Avoidance result of obstacle (pass through obstacle)

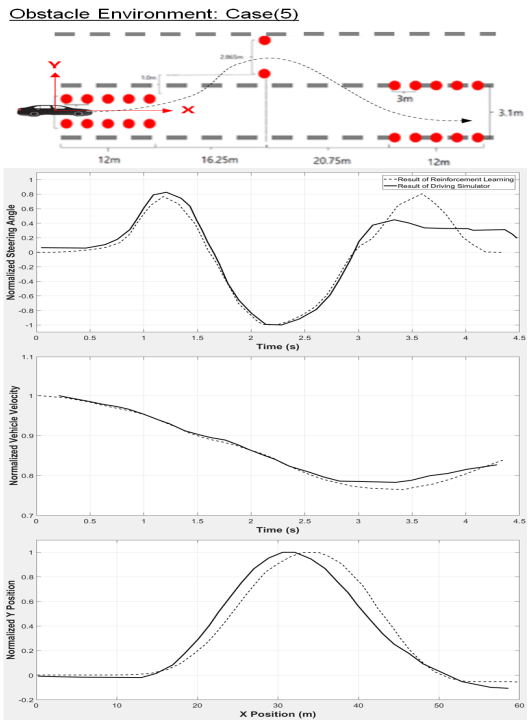

Fig. 11. Comparative on the result of reinforcement learning & driving simulator based on Case (5): Avoidance result of slalom test (ISO 3888-2)

Table 2. Specification of driving simulator

Hardware Specification	
Vehicle model	NF Sonata
Monitor	42"LED 3ch + 7"Toucj Screen
Power	220V, 2KW
Size	2000(L)X2400(W)X1500(H)
Weight	350kg

강화학습으로 얻은 결과를 실제 사람이 동일한 장애물 상황 속에서 차량을 조작할 때의 데이터와 비교하여 검증하였다. 이를 위해서 Driving Simulator를 사용하여 앞서 설정한 총 다섯 가지의 도로 환경을 구축하여 운전자 데이터를 추출하였다. 사용한 장비의 사양은 Table 2와 같다.

Driving Simulator로 추출한 운전자 데이터와 강화학습 결과를 비교한 결과는 각 장애물 상황 별로 Fig. 7에서 Fig. 11에 실선으로 표시하였다.

Fig. 7에서 Fig. 11까지를 통해 비교해보면 장애물 회피 시, 차량 이동 궤적 및 속도의 변화가 강화학습 결과와 실제 운전자의 결과가 유사함을 알 수 있었다. 다만 Steering angle에 있어 강화학습의 거동과 실제 운전자와의 데이터에 있어 차이가 있었다. Driving Simulator 내에서 사람이 차량을 조작할 때 실험 속도에 도달하는 과정에서 핸들 오차가 발생하는 부분과 CarSim에서 제공하고 있는 차량 재원과 Driving Simulator의 차량 특성을 맞춰주는 작업이 병행된다면 강화학습의 거동과 실제 운전자와의 데이터 사이에 좀 더 유사한 결과를 얻을 수 있을 것으로 예상된다.

4. 결 론

본 논문에서는 장애물 상황에 대해서 자율 주행 자동차가 사람과 유사한 거동을 보일 수 있도록 차량을 제어하는 방법을 제안하였다. 이를 구현하기 위해 Q-learning을 기반으로 한 강화학습을 적용하였다. 총 다섯 가지 장애물 상황에 대해서 강화학습을 적용하였으며 이때 학습 모델로는 Full car 모델을 기반으로 하는 CarSim을 사용하여 장애물 회피 학습을 진행하였다. 보상 정책을 구축하고 차량의 거동을 평가 할 때 사람이 주어진 장애물 환경에서 어떤 점을 중요시 여기는지에 대한 항목을 가정하여 설정하였으며 학습 결과를 Driving Simulator를 사용해 추출한 운전자 데이터와 비교를 해보았다. 학습결과와 운전 데이터를 비교한 결과 강화학습 결과와 유사한 차량 거동을 보임을 알 수 있었다. 이를 통해 강화학습을 적용한 장애물 회피 전략을 사전에 구축하고 차량에 적용함으로써 실시간으로 장애물 상황에 상응하는 차량 조작법을 사용할 수 있고 사용자 친화적인 장애물 회피 구현의 가능성을 확인하였다. 본 논문에서는 장애물 상황 및 위치 정보를 차량이 사전에 인지했다는 것을 가정하였기 때문에 차량이 장애물을 어떻게 인식할 것인가에 대한 보완사항이 남아있다. 이미 차량에 사용되고 있는 차량 센서(Radar, LIDAR, 카메라 등)을 사용하여 장애물 인식 문제를 보완한다면 강화학습을 적용한 장애물 회피 전략을 실제 환경에서 사용할 수 있을 것으로 기대한다.

References

[1] NHTSA (National Highway Traffic Safety Administration), *Federal Automated Vehicles Policy*, [Online], https://one.nhtsa.gov/nhtsa/av/av-policy.html, Accessed: December 12, 2016.

[2] L.A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338-353, Jun, 1965.

[3] R.B. Tilove, "Local obstacle avoidance for mobile robots based on the method of artificial potentials," in *IEEE International Conference on Robotics and Automation*, Ohio, USA, pp. 566-571, 1990.

[4] A.E. Eiben, P-E. Raue, and Zs. Ruttkay, "Genetic algorithms with multi-parent recombination," in *International Conference on Evolutionary Computation the Third Conference on Parallel Problem Solving from Nature*, Jerusalem, Israel, pp. 78-87, 1994.

[5] R. Malhotra, A. Sarkar, "Development of a fuzzy logic based mobile robot for dynamic obstacle avoidance and goal acquisition in an unstructured environment," in *IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, California, USA, pp. 1198-1203, 2005.

[6] L. Jong-Yeon, J. Hah-Min, and K Dong-Hun, "Amorphous obstacle avoidance based on APF methods for local path planning," *Journal of Korean Institute of Intelligent Systems*, vol. 1, no. 1, pp. 19-24, Feb, 2011.

[7] R.S. Sutton and A.G. Barto, "Introduction" in *Reinforcement learning : An Introduction*, MIT Press, 2012, ch.1, sec. 1.1, pp. 18-38s

[8] G.J. Tesauro, "Temporal difference learning and TDGammon," *Communications of the ACM*, vol. 38, no. 3, pp. 58-68, Mar, 1995.

[9] S. Lu, X. Liu, and S. Dai, "Incremental multistep Q-learning for adaptive traffic signal control based on delay minimization strategy," in *2008 7th World Congress on Intelligent Control and Automation*, Chongqing, China, pp. 2854-2858, 2008.

[10] ISO (International Organization for Standardization), *ISO 38888-2:2011*, [Online], http://www.iso.org/iso/catalogue_detail.htm?csnumber=57253, Accessed : September 16, 2016.

[11] C. Watkins, "Learning from delayed rewards," Ph.D. Thesis, King's College, Cambridge, England,1989.

[12] B. Beckman's, "Part 7: The traction budget," *The Physics of Racing*, [Online], http://phors.locost7.info/contents.htm, Accessed : July 03, 2016.

강 동 훈

2015 성균관대학교 기계공학부(공학사)
2017 고려대학교 자동차융합학과(공학석사)
현재 현대자동차 연구원

관심분야: 강화 학습, Q-Learning

봉 재 환

2012 고려대학교 기계공학부(공학사)
2014 고려대학교 기계공학부(공학석사)
현재 고려대학교 기계공학부(박사과정)

관심분야: 증강현실, 착용형 로봇, 기계학습

박 주 영

1983 서울대학교 전기공학과(공학사)
1985 KAIST 핵공학과(공학석사)
1992 University of Texas at Austin 전기및컴퓨터공학과(공학박사)
1993~현재 고려대 제어계측공학과 교수

관심분야: 기계학습, 제어이론

박 신 석

1989 서울대학교 기계설계학과(공학사)
1991 서울대학교 기계설계학과(공학석사)
1999 MIT, 기계공학과(공학박사)
1999~2000 Nissan 자동차 방문연구원
2000~2002 Harvard 대학교 Postdoctor
2002~2004 Keio 대학교 방문 교수
2010~2011 Harvard 의대 방문 교수
2004~현재 고려대 기계공학부 교수

관심분야: 로봇제어, 인간-기계 인터페이스