Human-in-the-loop Reinforcement Learning

Huanghuang Liang, Lu Yang, Hong Cheng, Wenzhe Tu, Mengjie Xu
Center for Robotics, University of Electronic Science and Technology of China.

***Abstract*—This paper focuses on presenting a human-in-the-loop reinforcement learning theory framework and foreseeing its application to driving decision making. Currently, the technologies in human-vehicle collaborative driving face great challenges, and do not consider the Human-in-the-loop learning framework and Driving Decision-Maker optimization under the complex road conditions. The main content of this paper aimed at presenting a study framework as follows: (1) the basic theory and model of the hybrid reinforcement learning; (2) hybrid reinforcement learning algorithm for human drivers; (3)hybrid reinforcement learning algorithm for autopilot; (4) Driving decision-maker verification platform. This paper aims at setting up the human-machine hybrid reinforcement learning theory framework and foreseeing its solutions to two kinds of typical difficulties about human-machine collaborative Driving Decision-Maker, which provides the basic theory and key technologies for the future of intelligent driving. The paper serves as a potential guideline for the study of human-in-the-loop reinforcement learning.**

***Index Terms*—Human-in-the-loop Reinforcement Learning; Driving Decision-Maker; Human-Driving**

I. Introduction

Human-machine hybrid intelligence is one of the core research directions in the new generation of artificial intelligence. First, from the perspective of human to machine learninghuman can enhance the speed of machine learning, for example, in the process of searching, human can play the guiding role in the early search strategy of machine learning; In addition, its difficult for machine to optimize some objective functions that are subjective and difficult to be digitized. But people can rely on their own intuition and experience to give machine the optimization guide in a relatively fast and efficient way. Secondly, from the perspective of machine learning to human: because of its high computing power, real-time computation abilities, machine can predict some information which is easily neglected by human; additionally, machine can calculate and make decisions about things that are impossibly judged by human, therefore, they can help make some auxiliary decisions for human. In order to achieve human-machine integration, it is meaningful to launch research on the human-machine hybrid intelligence.

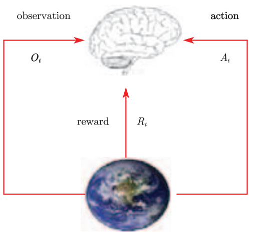
On the other hand, as one of the key technologies in artificial intelligence, reinforcement learning has become one of the hottest topics in the field of machine learning and artificial intelligence in recent years. At present, the field of artificial intelligence has entered the third wave stage. As the representative of artificial intelligence, interpretive intelligence and general intelligence break the narrow artificial intelligence in the traditional sense. The research uses the interpretation and correction interface to further enhance the learning model, and operate artificial intelligence systems (communication and interaction) in a highly transparent way. Reinforcement learning focuses on communication and interaction with the environment, and has attracted the attention of researchers in other disciplines such as operations research, control theory, robotics and so on. With the development of mathematical theory, reinforcement learning has been getting better and better results at the algorithm level. Moreover, the integration of cognitive science and the deep neural network is implemented to achieve end-to-end learning from perception to decision. Reinforcement learning has the potential to break through the key technology of artificial intelligence. Reinforcement learning as a general learning algorithm that can pass through perception and perceive decision control, it will be applied to a variety of areas in real life [3], [9], [23], [24], [30], [36].

Hybrid reinforcement learning is designed to combine human-machine hybrid intelligence and reinforcement learning, so as to achieve true human-machine communication, and it is still an emerging area of research in recent years, some scholars have gradually expanded their research. Abel et al. proposed a reinforcement learning model for Human-in-the-loop, independent of the agent, in 2016, with the goal of reinforcement learning for a tagged useful feature [23]. Through simulation training, a preliminary assessment can be made in some simple areas and verify the validity of the protocol; Modares et al. considered human factors, studied a new design method of human-machine interaction control, and used reinforcement learning to solve the optimization problem of linear quadratic regulator [30]. So far, the research on hybrid reinforcement learning is mainly focused on some relative static problems, which is not fully considered and organically integrated human intelligence and machine intelligence. Moreover, the current hybrid reinforcement learning scenarios are also relatively simple, mainly focused on some basic mathematical problems (such as quadratic linear programming) or some toy problems. The purpose of this paper is to propose a set of 'Human-in-the-loop' reinforcement learning theory framework, and verify it in a typical and practical application environment-human-machine cooperative driving decision maker.

Generally, it is of great academic and social significance to carry out the basic theory research of human-machine hybrid reinforcement learning and its application in driving decision-maker. With the development of artificial intelligence and electronic information technology, it is of great significance and prospect to apply human-machine hybrid reinforcement learning theory to human-vehicle cooperative driving.

The Corresponding Author is Lu Yang (yanglu1995uestc@aliyun.com).

978-1-5386-3524-7/17/$31.00 ©2017 IEEE



At each step t the agent:
Executes action A_t
Receives observation O_t
Receives scalar reward R_t
The environment:
Receives action A_t
Emits observation O_t
Emits scalar reward R_t
t increments at env. step

Fig. 1. Basic framework of reinforcement learning.

II. Literature Review

A. *Human-machine Hybrid Intelligence*

Human-machine hybrid intelligence is an important direction of artificial intelligence. Some international research teams have launched a series of studies about human-machine hybrid intelligence: Marcel Walch suggests that when the intelligent vehicle's autopilot system reaches its capability boundary, the system has to respond (such as an emergency brake) or hand over control to the driver. This shows that it is necessary to design a system with auxiliary switching function in the human-machine cooperative driving [37]; In December 2016, the Stanford University research team in published a paper about human-machine cooperative driving [29]. It pointed out that in the foreseeable future, there will be a transition between full driver driving and fully autonomous driving between modes of switching. That is the embodiment of human-machine Hybrid Intelligence. It is found that there is a switching adaptation period in the process of driving switch, which provides a reference for the design and actual manufacture of intelligent vehicles. In addition, the literature in [17] proposes that 'Human-in-the-loop' robot system has the potential to handle complex tasks. In the unstructured environment, combined with the cognitive ability of the human operator and the independent tools and behavior, implement a system of remote 'Human-in-the-loop' grab behavior. Furthermore, the study of 'Human-in-the-loop' is also reflected in the study of Demonstration Learning [1], [6], [14]. It is about how the robot to complete specific tasks through the guidance of human. Inspired by the above mentioned literature, in order to better realize the intelligent vehicle driving decision, it is necessary to study the human-machine hybrid Intelligence theory and fully consider the 'Human-in-the-loop' learning framework in the human-machine cooperative driving task.

B. *Reinforcement Learning*

In recent years, reinforcement learning has gradually become one of the focuses in the field of machine learning and artificial intelligence. Reinforcement learning (RL) as shown in Fig. 1. addresses the problem of a decision-maker faced with a sequential decision problem and using evaluative feedback as a performance measure [32].Its technology has been widely studied and applied in the fields of artificial intelligence, machine learning and automatic control [12], [20], [31], [33], [39], [41]. With the deepening of reinforcement learning algorithm and theory, especially the study of reinforcement learning has made breakthrough progress, reinforcement learning methods have been widely used in the fields of robotics, intelligent vehicles [21], [40], [43]–[45]. According to the type of learning system and the interaction environment, reinforcement learning can be divided into two categories: non-associative reinforcement learning and associative reinforcement learning [40]. Non-associative reinforcement learning system returns only from the environment without distinguishing the state of the environment; the structure of associative reinforcement learning is similar to the feedback control system, at the same time the state feedback information of the environment to get return. Whether the return of the study has a delay can be divided into types: Immediate return associative reinforcement learning and Sequential decision reinforcement learning. According to the stationarity and optimization index of MDP behavior selection strategy, reinforcement learning algorithm can be divided into two types: discounted reward and average reward.

1) Discounted reward reinforcement learning algorithm: TD learning algorithm: The temporal difference learning method plays an important role in early reinforcement learning and artificial intelligence, and has achieved some successful applications, but has not established a unified formal system and theoretical foundation. Sutton [32] et al proposed the formal description of temporal difference learning for the first time, and proved the convergence of the algorithm under certain conditions, thus laying a theoretical foundation for the temporal difference learning.

Q-learning algorithm: In order to optimize the control problem of the temporal difference learning, Watkins [38]proposed a table-based Q-learning algorithm for solving the MDP optimal value function and the optimal value strategy. Peng [25] et al proposed the $Q(\lambda)$ algorithm, which combines the Q-learning algorithm and the Eligibility Traces of the TD learning algorithm to further improve the convergence speed of the algorithm. In order to further improve the learning efficiency of reinforcement learning algorithm, Jonsson [15] et al proposed a Dyna-Q learning algorithm with online model estimation based on the idea of model identification in adaptive control. This method evaluates the MDP model online during the learning process, although it can significantly improve the efficiency, but must be at the expense of huge calculation and storage.

Sarsa learning algorithm: Singh [31] et al proposed an On-policy of the Q-learning algorithm, called the Sarsa learning algorithm. In the Q-learning algorithm, the behavior selection strategy and the iteration of the value function are independent of each other. While the Sarsa learning algorithm implements the iteration of the behavior value function in strict TD learning form, that is, the behavior selection strategy is consistent with the iteration of the value function. The Sarsa learning algorithm has been proven to be superior to the Q-learning algorithm in some applications of learning control problems.

Actor-Critic learning algorithm: The above three learning algorithms have a common feature that they only estimate the value function of MDP, and the behavior selection strategy is completely determined by the valuation of the value function. In order to estimate the value function and strategy at the same time, Barto [4] et al. Proposed the Actor-Critic learning algorithm, Critic uses TD learning algorithm to estimate the value function, Actor uses a strategy gradient estimation method for gradient descent learning. Barto et al. Only consider the case of discrete space, and the paper further studies the Actor-Critic learning algorithm for solving the optimal strategy of continuous behavior MDP.

2) Average reward reinforcement learning algorithm: Because in some practical problems, he average reward is more suitable to describe the optimization goal. average reward reinforcement learning algorithm has been paid more and more attention. This type of learning algorithm mainly includes the following three branches:

Time domain difference learning algorithm based on average reward: Preux et al. applied the dynamic programming theory and method to solve the problem of MDP strategy evaluation of average reward in the temporal difference learning, and proposed a temporal difference learning algorithm based on average reward [27]. In this algorithm, by introducing the concept of Relative Value Function in dynamic programming, the estimation of the value function of stationary strategy MDP is realized when the MDP model is unknown.

R-learning algorithm: In document [2], the R-learning algorithm is proposed. The algorithm achieves the iterative process of generalized strategy by iterating the relative function and greedy behavior selection strategy. Simulation studies in this paper show that R-learning algorithms perform better than Q-learning algorithms in some cases.

H-learning algorithm: In document [8], the H-learning algorithm is proposed, which can be regarded as the R-learning algorithm based on online model estimation. Because the discounted reward reinforcement learning algorithm has similar performance to the average reward reinforcement learning algorithm when the discount factor tends to 1, and the discounted reward reinforcement learning algorithm is much easier than the average reward reinforcement learning algorithm in theoretical analysis.

C. Driving Decision-maker

Through a variety of sensors installed on vehicles and roads to master the vehicle, roads and the surrounding vehicle status and other information for the driver to provide advice or warning signals and under certain conditions can control the vehicle. From the perspective of the international intelligent vehicle and intelligent transportation conference in recent years, the research direction is divided into two major categories: the direct response to sensor data and the environment modeling. The first type of driving decision maker mainly based on real-time sensor data for some simple and effective control, such as radar obstacle avoidance system . The second type of driving decision maker makes an environmental modeling of the input sensor information, to achieve a certain awareness on the basis of the vehicle for some control, such as based on the camera data lane line detection system [26] and vision-based navigation system [22], [34]. Human-machine cooperative driving in China is still a relatively new matter, but has made some progress through the detection of the driver's situation and the perception of the outside world, it can determine when the vehicle driver's right to convert. So it is not only necessary to have an auxiliary driving, but also need to do a certain degree of unmanned driving. Auxiliary driving system is required when the driver is driving; unmanned system is required to operate when the driver is out of the driving [13], [16], [18], [32]. It focuses on the applications human-vehicle interaction learning, as depicted in Fig. 2.

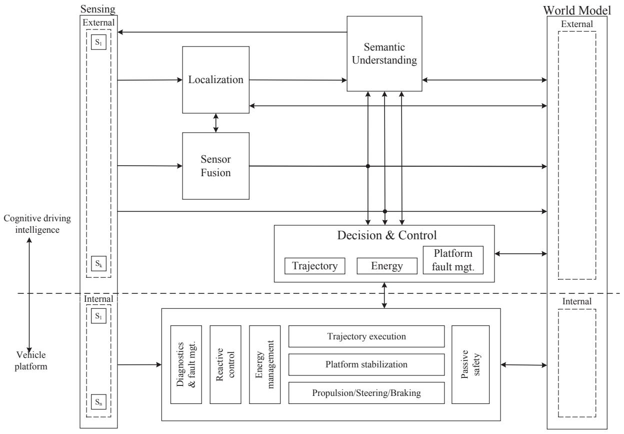


Fig. 2. Human-vehicle interaction learning.

The decision-maker [16], [34], [42] is based on the prediction of the behavior of other vehicles and makes decisions accordingly. This decision maker must be accepted by passengers (comfortable, reliable, agile, etc.), and also be accepted by other traffic participants (for example, cannot cause panic, ambiguity, strange and other associations) the detailed framework is shown in Fig. 3.

In recent years, scholars in China have made a lot of progress in the research of intelligent vehicle assistant driving decision maker. In the scientific research units, there are some representative researchers such as Zheng Nanning academician, Professor Wang Feiyue, Professor Li Deyi and so on [18], [26], [35]. Jilin University is one of the earliest intelligent vehicle research units in China. From the late 80s of last century Professor Wang Rongben led the intelligent vehicle research group began research on Autonomous Navigation of Intelligent Vehicle. Their team has in-depth study on Environmental perception, navigation technology and so on [32]. Intelligent vehicle research group led by Professor He Hangen of National University of Defense Technology is also an early intelligent vehicle research team in China [42]. Professor Gu Weikang, director of the Institute of intelligent and communication systems, Zhejiang University, has undertaken the 'military ground intelligent robot' project of the national defense science and engineering commission. In addition,

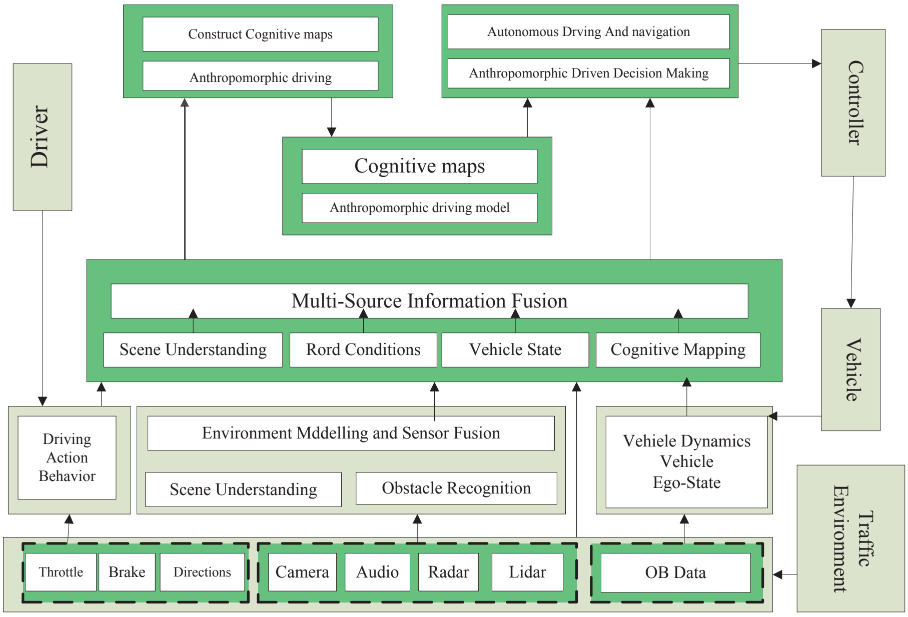


Fig. 3. The framework of decision maker.

Shanghai Jiao Tong University, Xi'an Jiao Tong University, Beijing Institute of Technology, Wuhan University, Hunan University, Hefei Institute of Physical Science.CAS, Academy of military transportation in recent years to carry out intelligent vehicle driving assistant decision, and achieved certain results [19], [34]. In addition, Professor Mao Qirong, Professor Chu Duanfeng, Professor Wu Chaozhong and Professor Hu Zhaizheng have also obtained a series of results in the aspects of driver detection and vehicle control algorithms. [10], [28]. The above research results have promoted the development of intelligent vehicles in China, and promoted the research of human-machine hybrid intelligent driving. In the aspect of intelligent driving technology, the foreign country started earlier, and has proved the feasibility of the technology and carried out road tests. Typical research on behalf of the United States Carnegie Mellon University NavLab-5 and Boss smart car [11], [26], Google's Google Driverless Car, the University of Parma, the ARGO vehicle [5], [35], the German Federal Defense Force University VaMP Intelligent driving system [7] and so on.

Although the research on the application of intelligent vehicle has made great progress, most of the existing research often uses only the environmental perception information outside the vehicle, and has some limitations. Referring to the questions and opinions put forward in Professor Chen Hong's paper, the driver's driving data should be taken into consideration, which has potential application value in the human-machine cooperative driving. In order to solve the current research problem, this paper is expected to be based on the study of human-machine hybrid intelligence theory. This is also the innovation of this paper. The advanced reinforcement learning method will be applied to the human-machine cooperative driving in an intelligent vehicle to solve the problem of human-machine cooperative driving. Considering the pilot's driving situation, we can realize the 'Human-in-the-loop' driving, so as to realize the human-machine cooperative information interaction, and then make the best driving decision.

III. Research Contents

The research is divided into four parts: (1) Man-machine hybrid reinforcement learning theory. (2) Man-machine hybrid reinforcement learning applied to human-vehicle oriented driving decision-making devices. (3) Man-machine hybrid reinforcement learning applied to automatic driving oriented driving decision-making devices. (4) Driving decision making verification platform. The logical relations is shown in Fig. 4.

Content (1) studies human-machine hybrid reinforcement learning theory, including the form of man-machine hybrid Markov decision processes model, establishment of the new 'Human-in-loop' optimal value function estimation methods, and the study of the guide of the exploration strategy of the behaviors like Softmax or ϵ-greedy; based on the content (1), content (2) and content (3) make content (1) concreted, we will do further research in two aspects: human-vehicle oriented driving decision-making devices and auto-driving oriented driving decision-making devices. content (2) focus on hybrid reinforcement learning theory on the condition of 'Human-in-loop', and the application of the hybrid reinforcement learning theory in human-vehicle oriented driving decision-making devices, Content (3) is focused on the study of auto-driving, And how to materialized hybrid reinforcement learning theory (including the define of elements in Markov decision chain, re-estimation of Value function and the value estimation of anthropogenic influence of the behavior strategies).All these will converting theory into practical driving conditions. Content (4) provides a verification platform, and the platform is

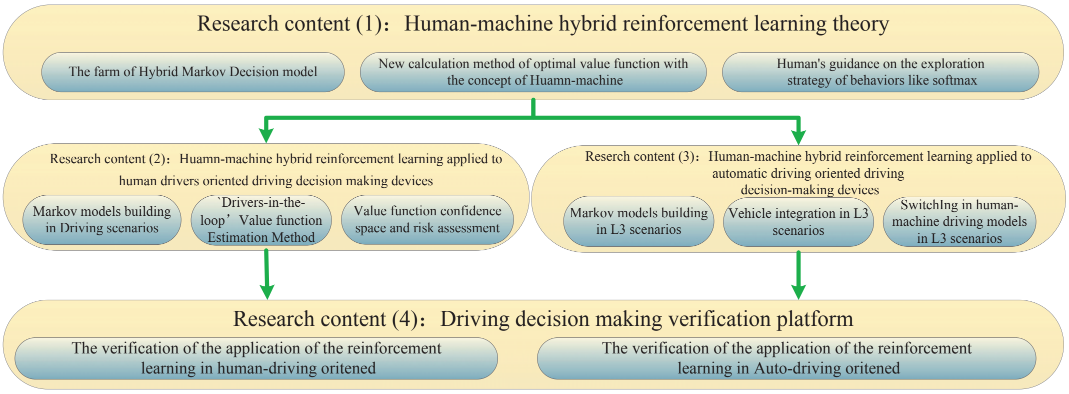

Research content (1)：Human-machine hybrid reinforcement learning theory
The farm of Hybrid Markov Decision model
New calculation method of optimal value function with the concept of Huamn-machine
Human's guidance on the exploration strategy of behaviors like softmax
Research content (2)：Huamn-machine hybrid reinforcement learning applied to human drivers oriented driving decision making devices
Markov models building in Driving scenarios
`Drivers-in-the-loop' Value function Estimation Method
Value function confidence space and risk assessment
Reserch content (3)：Human-machine hybrid reinforcement learning applied to automatic driving oriented driving decision-making devices
Markov models building in L3 scenarios
Vehicle integration in L3 scenarios
SwitchIng in human-machine driving models in L3 scenarios
Research content (4)：Driving decision making verification platform
The verification of the application of the reinforcement learning in human-driving oritened
The verification of the application of the reinforcement learning in Auto-driving oritened

Fig. 4. The relationship between the contents of this study.

used to verify the algorithm in content (2) and content (3).

A. Human-machine Hybrid Reinforcement Learning Theory

Content: Hybrid reinforcement learning fundamental research is aimed at establishing a reinforcement learning mechanism in the condition of ‘Human-in-the-loop’. Includes (1) established a new hybrid Markov decision processes (MDP) model. From an empirical point of view, (2) proposed a iterative method of Parameterized value function, The method dynamically configured the effect of experience of human and machine to value function, improve the convergence speed of hybrid reinforcement learning, and to explore the best configuration parameter of man and machine, in addition it also proposed a Convergent boundary of man-machine hybrid reinforcement theory; (3) fully using human intuitive judgment about their surroundings and inherent experience on weighted combination of selection strategies such as ϵ-greedy and Softmax, and dynamic adjustment of the weights.

Study objectives: (1) Establishment of basic theory of human-machine hybrid reinforcement learning, redefining the Human-in-loop Markov decision process (2) hybrid intelligent theory to improve the convergence speed of reinforcement learning, and calculate the maximum theoretical speed of convergence boundary values on the condition of digitalized human experience, explore the iteration formula of the dynamic parameter value function, as well as the convergence conditions; (3) Quantitative human curiosity and intuition of the surrounding environment to establish intuition value function and curiosity quantization function to find out the optimum balance of value Q with them, and optimize the balance between exploitation and exploration, and quantify the risk with local optimum and the earning with the improve of the convergence speed.

Key scientific questions: (1) How to define and quantify the State, how to measure /estimate the status of human, and the establishment of three important elements of Markov decision processes (status, returns, state transition probability) and dynamic relationship with human state. This subject intends to human efficacy theory to define and measure the state value of human, then use the Dynamic Bayesian network function to from a non-analytic mathematical model to describe the relationship between the state of human beings and the elements of Markov decision process, meanwhile we will try to find a conjugate posterior probability function distribution expression based on Prior probability distribution, even if we cannot found conjugate posterior probability function, we still have a reserve solution: Using Markov Chain Monte Carlo (MCMC) to express the conjugate posterior probability function. (2) How to optimize the configuration between the value function with the machine and experience of human, how to calculate the biggest value of convergence with the concern of the optimum configuration of hybrid reinforcement learning. This subject is based on Lagrangian theory of convex optimization, bold assumptions, and proved the convex properties of configuration functions, used convex programming duality theory to prove the prove that salient characteristics of the dual function of the constraints, so as to ensure optimum and accessibility (the Uniformity cannot be guaranteed), and use the master-slave dual method (Primal-Dual) to find the optimal weight. Using elipsoid methods to prove optimal point available in polynomial time, and at the same time detect the optimal value of the upper and lower boundary according to the dual theory of boundaries.

B. Human-in-the-loop Reinforcement Learning Applied in Human Drivers Oriented Driving Decision-maker

Content: Human-machine hybrid reinforcement learning was used in human-driving decision-making devices in order to study how to realize human-machine hybrid reinforcement learning theory under human driving scenarios, and study its particularities. Contents include: (1) Such as lane keeping, automatic parking and braking assist, backup and car safety driving assist-specific issues such as lane Assistant / Scenes, monitoring the real-time status, establishing corresponding to the Human-in-the-loop model of Markov decision processes. (2) Based on the driver’s driving habits and the monitoring status of the driver, exploring the method of the dynamic parameter value function estimation. Adjusting value function estimation functions on the condition of unbalanced in the sample space (there are only trace amounts of negative samples). (3) Studying value calculation method of function space of confidence, giving some confidence level and risk evaluation of function, and propose to switched strategies reference to the risk evaluation function and the corresponding the extent of trust.

Study objectives: (1) Establishing the reinforcement learning training framework under the scenario of 'Human-in-the-loop'. According to different scenarios (such as lane keeping, automatic parking) Re-specific parameters are defined (human status as one of the most important parameter) Markov decision process elements (status, returns, and state transition probability); (2) Proposed 'individual (personalized)' value function estimation methods ,according to mass data and real-time monitoring to driver status; (3) Under the condition of very unbalanced sample space (very few negative samples), an accurate value function estimation method is proposed, where the false negative rate is reduced in the case of minimizing the false error rate (4) Establishing a reliable high-valued function confidence space method.

Key scientific questions: (1) How to integrate large volumes of different driver data and real-time status monitoring data of the same driver, and to build a 'personalized' value function estimation methods. This paper intends to begin using dynamic Bayesian network for data fusion, the output was defined as probability distribution of the true state of the driver , Conducting stochastic programming according to the probability distribution to achieve optimal value function estimation and (2) How to design an accurate value function estimation method under very few negative samples? This paper proposes to adopt (a) Adaboost machine learning methods small sample space sampling frequency to do the adjustment of their adaptation, thereby reducing the 'fake' error rate; (b) at the same time, programming method of using iterative weight (Iterated Reweight Programming), continuously improve 'fake' wrong weight, thereby enabling the machine to reduce 'false' error rate, (3) And how to calculate confidence intervals for the value function. This paper intends to continue to use Bayesian Networks, based on mass drivers the prior probability of the data given the model, the posterior probability is based on numerical computation of statistical models, and function space of confidence given predictive value.

C. Human-machine Hybrid Reinforcement Learning Applied in Autopilot Oriented Driving Decision-maker

Content: Human-machine hybrid reinforcement learning is used in autopilot driving decision-maker aims to study how to make a car can quickly and accurately to achieve the human driving level, switching and pre-judge when vehicle is required. Contents include: (1) Based on the basic theory of hybrid learning on how to use human experience and intuition to help improve the convergence of learning speed and fast parametric search criteria (ϵ-greedy and Softmax); (2) Research on the effective integration of human experience and the machine's experience function, it would make the car have a certain level of intelligence and avoid some unnecessary mistakes trying at the beginning of reinforcement learning and training. (3) To study how to define the experience of the car and decide whether and when to implement person driving mode switching. For extremely complex driving situations or 'no solution' (must be illegal) driving, Smart cars need to choose the right handed human driving, the third study mainly study the necessity and switch the timing of switching.

Study objectives: (1) Using experience and intuition of human to improve the process of vehicle learning, and detecting the optimal human-vehicle experience balancing strategy, giving the fastest boundary value of theoretical learning. (2) Defining and quantifying silly driving mistakes, minimizing errors in reinforcement learning process; (3) Define and quantify the vehicle safe driving confidence, switching driving mode when the confidence in safe driving value is lower than the threshold.

Key scientific questions: (1) How to define the mathematical problem of fusing human-vehicle experience values on the condition of auto-driving oriented scene, and to give a precise boundary analysis. Autonomous driving is a dynamic model, it would change as the environment and the changes of the vehicle itself, and its hard to use Classic to model a system identification method. This paper attempts to explore the model in the loop method, and to use the application of model predictive control to solve the problem. Dynamic system identification method is applied first online digital drive models using model predictive control method based on this model, and the model was built in a very short time, and then using model predictive control method. In the EU project SEAM4US, the problem of similar complexity has been successfully solved by me, so there is confidence in the model ring and model predictive control method and its application to autopilot scenarios. (2) How to determine when a person driving mode to switch. This issue will be solved by follow measures: (a) Build vehicles safe driving confidence value function by using machine learning techniques (such as support vector machines) and sparse smooth convex optimization theory to optimize the safety driving confidence value function and control the complexity of the model. (b) At the same time, Papers published in accordance with the latest in robotics [29], parameter switch of time. (c) Using the value function iteration method of reinforcement learning, building confidence value to switch from safe driving time parameter mapping.

D. Driving Decision Verification Platform

Content: The safety and reliability of Smart driving were achieved by a mass of functional and performance tests. Considering of the possibility of normal and abnormal condition, so decision-making platform for intelligent drive system is validation of safer and more effective in economy alternative choice. Operating vehicles in collaborative simulation platform for driving, can keep the driver's perception of monitoring and environmental scenarios; using multi-modal data fusion, we can also use judgment on the interactive driving rights

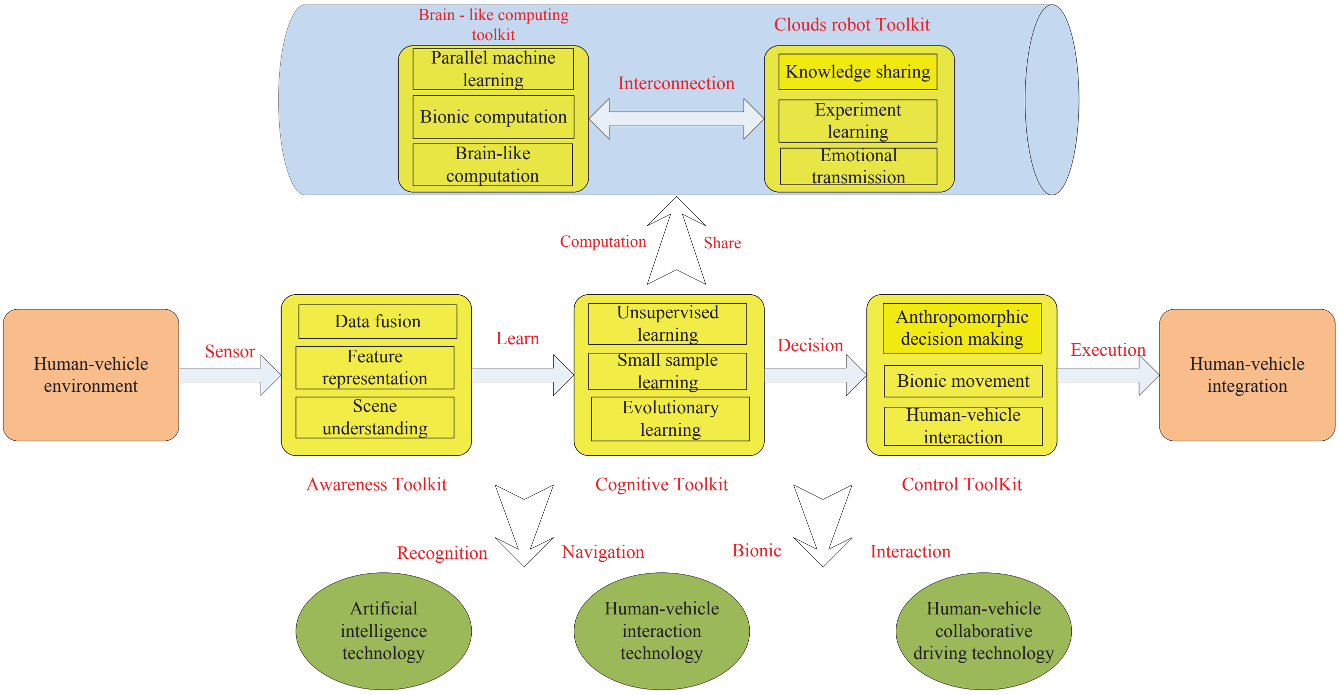


Fig. 5. Collaboration of human-vehicle intelligent driving system based on AI.

between human-vehicle, and online real time validation and an algorithm to choose the most suitable model.

Study objectives: Through the operation in human-vehicle collaboration simulation platform, following progress can be expected: First using the platform to gets driver situation and outside scene of perception data accurately. These data are processed by using multi-modal fusion techniques. All these made theoretical research and experiment can be done simultaneously; comparing the hints of platform and real-time situation to verify the model and algorithm of human-vehicle collaboration driving; using the platform interface displayed real-time image to validate the stability of platform on the condition of 'Human-in-the-loop'; it is necessary to upgrade the platforms at the same time of verification, to make vehicle hybrid drive can be simulated completely on the platform. It focuses on the applications of artificial intelligence, machine learning and automatic control to vehicles, as depicted in Fig. 5.

IV. Conclusion and Future Works

Compared with the domestic and foreign related research, in terms of academic thought, research program and technical route. The characteristics and innovations of this paper are the integration of human-machine interaction problem into the 'Human-in-the-loop' under the loop intelligent driving framework. Using human-machine hybrid reinforcement learning basic theories. Integration of man and the integration of experience of human and machine, which make decisions in the human-driving-dominated (such as L2) and the autopilot-dominated (such as L3) application fields more efficient.

The main characteristics and innovations are shown in:

(1) At present, most of the research on smart vehicles has been moving towards unmanned vehicles. In the market only auxiliary driving equipment has been applied without taking the problem of human-vehicle interaction into account. This paper will achieve an intelligent balance by sensing the exterior environment and monitoring the driver's posture, and timely do the transfer of human-vehicle driving rights, which can be intelligent and humanized. It can do human-oriented, serve the public better, and promote the development of intelligent driving.

(2) By human-machine hybrid reinforcement learning basic theory, a new hybrid Markov decision process model can be established, and by using human experience value parameterization function and using human intuition for strategic choice, which can achieve intelligent fusion of human-vehicle. The most important thing in the human and vehicle interaction is to be able to determine when the person should be driving and when the machine should be driving, so that this paper makes theoretical innovation in the human and vehicle collaboration control to meet the above requirements in order to make driving experience more comfortable.

(3) Human-machine hybrid reinforcement learning theory applied to two applications: human driving dominated application and the autopilot dominated application. It can be used in lane maintenance, automatic parking and other specific safe driving scene. Fusing experience value of human and machine, researching calculation method of function belief space can enhance the training effect of primal learning.

References

[1] S. R. Ahmadzadeh, R. Kaushik, and S. Chernova, "Trajectory learning from demonstration with canal surfaces: A parameter-free approach," in *Ieee-Ras International Conference on Humanoid Robots*, 2017.

[2] N. Aissani, B. Beldjilali, and D. Trentesaux, "Dynamic scheduling of maintenance tasks in the petroleum industry: A reinforcement approach," *Engineering Applications of Artificial Intelligence*, vol. 22, no. 7, p. 10891103, 2009.

[3] S. Arai and K. Sycara, "Multi-agent reinforcement learning for planning and conflict resolution in a dynamic domain," in *International Conference on Autonomous Agents*, 2000, pp. 104–105.

[4] A. G. Barto, R. S. Sutton, and C. W. Anderson, "Neuronlike adaptive elements that can solve difficult learning control problems," *IEEE transactions on systems, man, and cybernetics*, no. 5, pp. 834–846, 1983.

[5] M. Biehl and P. Riegler, "On-line learning with a perceptron," vol. 28, no. 7, pp. 525–530, 1994.

[6] A. Billard, S. Calinon, R. Dillmann, and S. Schaal, "Robot programming by demonstration," *Handbook of Robotics*, vol. chapter 59, no. 4, pp. 1371–1394, 2008.

[7] A. Broggi, M. Bertozzi, A. Fascioli, and G. Conte, "Automatic vehicle guidance: The experience of the argo vehicle," *IN PROCS. SPIE'98 - AEROSENSE CONF*, pp. 218–229, 1999.

[8] A. Broggi, M. Bertozzi, and A. Fascioli, "The 2000 km test of the argo vision-based autonomous vehicle," *IEEE Intelligent Systems*, vol. 14, no. 1, pp. 55–64, 1999.

[9] L. BuOniu and S, "A comprehensive survey of multiagent reinforcement learning."

[10] Y. Chen and J. Wang, "Adaptive vehicle speed control with input injections for longitudinal motion independent road frictional condition estimation," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 3, pp. 839–848, 2011.

[11] D. Chu, X. Y. Lu, C. Wu, Z. Hu, and M. Zhong, "Smooth sliding mode control for vehicle rollover prevention using active antiroll suspension," *Mathematical Problems in Engineering*, vol. 2015, pp. 1–8, 2015.

[12] T. Degris, P. M. Pilarski, and R. S. Sutton, "Model-free reinforcement learning with continuous action in practice," in *American Control Conference*, 2012, pp. 2177–2182.

[13] X. Gang, W. Kang, F. Wang, F. Zhu, Y. Lv, X. Dong, J. Riekki, and S. Pirttikangas, "Continuous travel time prediction for transit signal priority based on a deep network," in *IEEE International Conference on Intelligent Transportation Systems*, 2015, pp. 523–528.

[14] B. Huang, M. Li, R. L. D. Souza, J. J. Bryson, and A. Billard, "A modular approach to learning manipulation strategies from human demonstration," *Autonomous Robots*, vol. 40, no. 5, pp. 903–927, 2016.

[15] A. Jonsson and A. Barto, "Causal graph based decomposition of factored mdps." *Journal of Machine Learning Research*, vol. 7, no. 3, pp. 2259–2301, 2006.

[16] J. Koutnk, G. Cuccu, J. Schmidhuber, and F. Gomez, "Evolving large-scale neural networks for vision-based reinforcement learning," *Machine Learning*, pp. 541–548, 2013.

[17] A. Leeper, K. Hsiao, M. Ciocarlie, L. Takayama, and D. Gossow, "Strategies for human-in-the-loop robotic grasping," in *ACM/IEEE International Conference on Human-Robot Interaction*, 2012, pp. 1–8.

[18] Q. Li, N. Zheng, and H. Cheng, "Springrobot: a prototype autonomous vehicle and its algorithms for lane detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 5, no. 4, pp. 300–308, 2004.

[19] S. Li, K. Li, R. Rajamani, and J. Wang, "Model predictive multi-objective vehicular adaptive cruise control," *IEEE Transactions on Control Systems Technology*, vol. 19, no. 3, pp. 556–566, 2011.

[20] D. Liu and Q. Wei, "Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 3, pp. 621–634, 2014.

[21] H. R. Maei, C. Szepesvri, S. Bhatnagar, D. Precup, D. Silver, and R. S. Sutton, "Convergent temporal-difference learning with arbitrary smooth function approximation," in *Advances in Neural Information Processing Systems 22: Conference on Neural Information Processing Systems 2009. Proceedings of A Meeting Held 7-10 December 2009, Vancouver, British Columbia, Canada*, 2009, pp. 1204–1212.

[22] M. Maurer, R. Behringer, S. Furst, F. Thomanek, Dickmanns, and E. D., "A compact vision system for road vehicle guidance," in *International Conference on Pattern Recognition*, 1996, p. 313.

[23] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[24] Y. K. Na and S. Y. Oh, *Hybrid Control for Autonomous Mobile Robot Navigation Using Neural Network Based Behavior Modules and Environment Classification*. Kluwer Academic Publishers, 2003.

[25] J. Peng and R. J. Williams, "Incremental multi-step q-learning," *Machine Learning Proceedings*, vol. 22, no. 1-3, pp. 226–232, 1994.

[26] D. Pomerleau and T. Jochem, *Rapidly Adapting Machine Vision for Automated Vehicle Steering*. IEEE Educational Activities Department, 1996.

[27] P. Preux, S. Girgin, and M. Loth, "Feature discovery in approximate dynamic programming," in *Adaptive Dynamic Programming and Reinforcement Learning, 2009. ADPRL'09. IEEE Symposium on*. IEEE, 2009, pp. 109–116.

[28] Q. Rao, X. Qu, Q. Mao, and Y. Zhan, "Multi-pose facial expression recognition based on surf boosting," in *International Conference on Affective Computing and Intelligent Interaction*, 2015, pp. 630–635.

[29] H. E. B. Russell, L. K. Harbott, I. Nisky, S. Pan, A. M. Okamura, and J. C. Gerdes, "Motor learning affects car-to-driver handover in automated vehicles," vol. 1, no. 1, p. eaah5682, 2016.

[30] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, d. D. G. Van, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, and M. Lanctot, "Mastering the game of go with deep neural networks and tree search." *Nature*, vol. 529, no. 7587, p. 484, 2016.

[31] S. Singh, T. Jaakkola, M. L. Littman, and C. Szepesvri, "Convergence results for single-step on-policy reinforcement-learning algorithms," *Machine Learning*, vol. 38, no. 3, pp. 287–308, 2000.

[32] R. Sutton and A. Barto, "Reinforcement learning: An introduction, adaptive computation and machine learning series," 1998.

[33] R. S. Sutton, D. Mcallester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," *Advances in Neural Information Processing Systems*, vol. 12, pp. 1057–1063, 1999.

[34] E. Ur and E. Ahin, "Traversability: A case study for learning and perceiving affordances in robots," *Adaptive Behavior*, vol. 18, no. 18, pp. 258–284, 2010.

[35] C. Urmson, J. Anhalt, D. Bagnell, C. Baker, R. Bittner, M. N. Clark, J. Dolan, D. Duggins, T. Galatali, and C. Geyer, "Autonomous driving in urban environments: Boss and the urban challenge," *Journal of Field Robotics*, vol. 25, no. 8, p. 425466, 2008.

[36] P. Vrancx, K. Verbeeck, and A. Nowé, "Decentralized learning in markov games," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 38, no. 4, pp. 976–981, 2008.

[37] M. Walch, K. Lange, M. Baumann, and M. Weber, "Autonomous driving: investigating the feasibility of car-driver handover assistance," in *International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 2015, pp. 11–18.

[38] C. J. C. H. Watkins, "Learning from delayed rewards," Ph.D. dissertation, King's College, Cambridge, 1989.

[39] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[40] J. Wu, X. Xu, J. Wang, and H.-G. He, "Recent advances of reinforcement learning in multi-robot systems: a survey," *Control and Decision*, vol. 26, no. 11, pp. 1601–1610, 2011.

[41] H. Zhang, D. Liu, Y. Luo, and D. Wang, *Adaptive Dynamic Programming for Control: Algorithms and Stability*. Springer Publishing Company, Incorporated, 2013.

[42] J. Zhang and K. Cho, "Query-efficient imitation learning for end-to-end autonomous driving," 2016.

[43] D. Zhao, Z. Xia, and D. Wang, "Model-free optimal control for affine nonlinear systems with convergence analysis," *IEEE Transactions on Automation Science and Engineering*, vol. 12, no. 4, pp. 1461–1468, 2015.

[44] D. Zhao and Y. Zhu, "Meca near-optimal online reinforcement learning algorithm for continuous deterministic systems," *IEEE transactions on neural networks and learning systems*, vol. 26, no. 2, pp. 346–356, 2015.

[45] Y. Zhu, D. Zhao, and X. Li, "Using reinforcement learning techniques to solve continuous-time non-linear optimal tracking problem without system dynamics," *IET Control Theory & Applications*, vol. 10, no. 12, pp. 1339–1347, 2016.