A Continuous-Time Markov Decision Process-Based Method With Application in a Pursuit-Evasion Example

Shengde Jia, Xiangke Wang, *Member, IEEE*, and Lincheng Shen, *Member, IEEE*

***Abstract*—This paper presents a novel method—continuous-time Markov decision process (CTMDP)—to address the uncertainties in pursuit-evasion problem. The primary difference between the CTMDP and the Markov decision process (MDP) is that the former takes into account the influence of the transition time between the states. The policy iteration method-based potential performance for solving the CTMDP and its convergence are also presented. The results obtained by MDP-based method demonstrate that it is a special case of CTMDP-based method involving the identity transition rate matrix. To compare the methods, a well-known pursuit-evasion problem, involving two identical cars, is solved as a benchmark. The CTMDP-based method can provide a discretization solution that is close to the analytical solution obtained by the differential game method. Besides, it shows strong robustness against changes in the transition probability, as compared with the traditional MDP-based method. To the best of our knowledge, this is the first attempt to validate the influence of the transition time between the states in such a pursuit-evasion scenario, or in a similar application, solved by an MDP-related model. The CTMDP-based method offers a new approach to solving the pursuit-evasion problem and can be extended to similar optimization applications.**

***Index Terms*—Continuous-time Markov decision process (CTMDP), dynamic programming (DP), policy iteration, potential performance, pursuit-evasion.**

I. Introduction

PURSUIT-evasion is a family of problems in control theory and computer science, in which one group attempts to track down the members of another group in a given environment. An important reason of studying techniques for solving pursuit-evasion problems is that automated pursuit systems perform better than human pursuers [1]. This problem is difficult and is usually considered a dynamic, stochastic, continuous-space, and continuous- or discrete-time discrete-space game [2].

Manuscript received May 23, 2015; accepted July 31, 2015. Date of publication October 9, 2015; date of current version August 16, 2016. This work was supported in part by the National Natural Science Foundation of China under Grant 61403406 and in part by the Research Project of National University of Defense Technology. This paper was recommended by Associate Editor W.-K. V. Chan. *(Corresponding author: Xiangke Wang.)*

The authors are with the College of Mechatronic Engineering and Automation, National University of Defense Technology, Changsha 410073, China (e-mail: jia.shde@gmail.com; xkwang@nudt.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMC.2015.2478875

There are two main types of techniques for addressing such a problem. According to differential game theory, the pursuit-evasion problem is essentially an optimal control problem. When capture can occur, the Hamiltonian condition must be satisfied by the saddle-point controls of the players [3]. This method is always used to address problems with continuous time and continuous space, but it is impossible to obtain a solution when the equations become complex. Another approach could be to consider the problem as one in which the agent of the pursuer is set against nature. In this approach, the pursuer must maximize nature's utility, although the environment is unknown. To handle these uncertainties, a probabilistic framework, such as the Markov decision process (MDP), can be used in modeling the evolution of the system [4]. The transition probabilities between states could be used to describe the uncertainties in the form of a probability. For example, the dynamic programming (DP) technique has been used to solve an MDP [5], [6]. Two major solution methods, involving construction of a value function to be optimized and the space of states, are value iteration and policy iteration. DP is the most popular technique to address sequential decision problems. However, the principal problem facing DP is the curse of dimensionality: the number of parameters increases exponentially with the size of the compact state representation. This problem appears to be even more complex in a dynamic system, such as pursuit-evasion. To overcome this problem, DP-based techniques must be extended. Thus, we extend our research from the MDP to the continuous-time Markov decision process (CTMDP) because the latter is expected to overcome the curse of dimensionality by temporal abstraction.

This paper is based on those previous works wherein CTMDP was used to model the dynamic environment [7]. Here, we introduce a novel mathematical framework to derive our algorithms. The "game of two identical cars" on a plane will be used to demonstrate the presented method. In the game, both the pursuer P and the evader E have positive minimum-turn radii and constant speeds. The game terminates when E's separation from P becomes less than a specified capture radius. The controls are the normalized turn rates of P and E. The purpose of this paper is to apply a novel method, namely, the CTMDP-based method, to such a problem and to present an approximate technique for a fairly reliable technique for pursuit-evasion situations that are imperfectly observed. Unlike in the differential game theory, only the decision or control of the pursuer is considered here. The game

of two identical cars problem is used to demonstrate that the CTMDP-based algorithm is better than the MDP-based one, in terms of usability and robustness.

A. Motivation

As mentioned previously, the MDP probability framework and the temporal abstraction technique are expected to overcome the two challenges complicating the benchmark game: 1) the uncertainties and 2) the curse of dimensionality. In this section, we will briefly explain the CTMDP was chosen from the perspectives of these two issues.

Regarding the uncertainties in the environment, MDP provides an efficient probability framework. The MDP, which is also referred to as a controlled Markov chain [8], describes a problem in which a single agent must choose an action at every node of the chain to maximize some reward-based optimization criterion. The basic idea used in modeling the environment is the Markovian property, which states that the occurrence of the current state depends only on the previous state. Consider a Markov chain $X^+ = \{X_0, X_1, \ldots, X_n\}$ as a sequence of random variables. Thus, the property can be described as $P(X_{l+1} = j|X_0 = i_0, \ldots, X_l = i_l) = P(X_{l+1}|X_l = i_l)$, where the values of X are selected in the state space S. In the real world, this property can be explained as a result of the agent considering only the information of the previous and the current states when selecting its current action. Under the MDP framework, the uncertainties could be formalized based on the transition probabilities and the Markovian property.

In this paper, our attempt is to handle the curse of dimensionality by adding the temporal influence. Most of the previous works on MDPs focused on the decision processes without considering the effects caused by transition time, which commences at a state and continues until the next state arrives [5], [9]. Consequently, conventional MDP methods cannot take advantage of the simplicity and efficiency that are sometimes available at higher levels of temporal abstraction [10]. For example, when modeling a robotic path-planning problem within an MDP framework, the state is defined as its position, and the transition from state i to state j as action a. In conventional MDP, regardless of the length of the transition time, the solution will be the same because the transition probabilities do not change. In fact, it is more reasonable to assume that the transitions whose transition times are closer to the expected transition times have a greater probability of occurring. In summary, the method containing temporal abstraction might perform better with the same state representation or it might require fewer states to exhibit equal performance. A detailed explanation of this method is given in Section III.

The CTMDP is expected to provide an appropriate model that contains the temporal influence. The principal difference between MDP and CTMDP is that the former corresponds to a Markov chain $X^\dagger$, and the latter to a Markov stochastic process X_t (for simplicity, we denote $X(t)$ as X_t in this paper). Then, the Markovian property can be described as $P(X_{s+t} = j|X_u; u \leq s) = P(X_{s+t} = j|X_s)$, and for any time variables t, s maintains $0 < t, s < \infty$. Among the Markov theories, the Markov process X_t is proposed to be described in two parts [11]: 1) the embedded Markov chain $X^\dagger$ and 2) the expected sojourn time at each state. As a result, the CTMDP considers not only the Markovian property but also the time property at each transition.

B. Literature Review

A large body of literature has addressed the related Homicidal Chauffeur and Two Cars games since the seminal work of Isaacs [3] in 1965. In the early stages, the focus is on how to obtain new types of singular lines for problems in the plane. Cockayne [12] showed that the pursuer satisfies the capture condition when the players are modeled as Dubins vehicles. Merz [13], [14] provided a systematic description of the solution structure for the Homicidal Chauffeur game depending on the given parameters. Later on, the game was extended to a 3-D game. Shinar and Gutman [15] solved a 3-D linearized kinematic model with bounded controls as a zero-sum game with perfect information. Miloh [16] showed that in the 3-D version of the Homicidal Chauffeur game, it would be optimal for both players to steer the system into a common plane and remain there until termination. Miloh *et al.* [17] generalized the 2-D capturability criteria for the Homicidal Chauffeur and Two Cars games for a genuine 3-D pursuit-evasion encounter. Pang [18] studied a pursuit-evasion problem in which both vehicles had limited acceleration and turning ability and relied on sensors to determine the location of their opponent. These studies are treated as perfect information games in degree or kind, which imposed restrictions on their detailed analysis, e.g., a linear game; they fixed the velocity ratios or the players, or divided the constant radii of curvature piecewise.

The description of the evader's behavior allows for incorporating probabilities into opponent locations, intents, and/or sensor observations [19]. The probabilistic pursuit-evasion game framework was first considered for solving a game with multiple pursuers and one randomly moving evader [19], [20] extended the method to a game involving multiple pursuers and multiple randomly moving evaders. The MDP is now widely accepted as a preferred framework for modeling control and planning problems with probabilities [21]–[23]. In [4], the map-learning and pursuit process are integrated into a partial-information Markov game, and DP is used to find the Stackelberg equilibrium. It is very natural and suitable to use CTMDP for a pursuit-evasion problem because the combat result will be sensitive to time, which means that the results might depend on the duration of transition times. The CTMDP provides an extended model of the usual discrete-time MDP [8], [24], [25].

Treating the MDPs as a discrete-event dynamic system, Cao [11] proposed a potential performance-based approach to find the optimal solutions to the discounted and average reward problems. This approach provides a new perspective to optimization problems and establishes a relationship among MDPs, perturbation analysis, and reinforcement learning. The concept of potential performance, which is discussed thoroughly in [11], forms a major thread throughout this paper.

C. Original Contributions

The original contributions of this paper are as follows.

1) By addressing the applications of dynamic processes modeled in MDP, we focus on the fact that the transition time between every two states is variable. The well-known pursuit-evasion problem called the game of two identical cars, is selected as an example. To characterize the system dynamics, CTMDP is introduced to model this game.
2) Based on the concept of potential performance, we present a policy iteration algorithm to solve the CTMDP. We also prove the convergence of the presented algorithm and describe some of its theoretical properties. The MDP-based method is also demonstrated to be a special case of the CTMDP-based method in which the transition rates matrix is the identity matrix I, indicating that the CTMDP has a stronger modeling capacity.
3) It is demonstrated that the transition rates (i.e., the reciprocals of the transition time) are significant parameters in modeling the dynamic processes via the MDP. With a few assumptions, the transition probabilities and rates are reasonably constructed. The solution obtained by the CTMDP-based method is closer to the analytical solution obtained by the differential game method, than the one obtained by MDP-based method. Through simulations, the results of two different methods are compared at a set of slices with different relative headings.
4) The CTMDP-based method is also shown to be more robust than the MDP-based one in solving inaccurately estimated transition probabilities. Given that the transition probabilities are inaccurate or even worse, the CTMDP-based method can still produce a solution that approximates the analytical one, whereas the MDP-based method cannot.

D. Organization

The rest of this paper is organized as follows. The dynamics of the game of two identical cars and the solutions obtained by the differential game method are presented in Section II. Section III gives a detailed description of the CTMDP-based method, and proposes a definition for the CTMDP. Besides, it presents the main contribution of this paper, i.e., the policy iteration algorithm. The application of the presented method to the game of two identical cars is described in Section IV. The presented method is demonstrated by using some numerical examples in Section V. Finally, the conclusion is given in Section VI.

II. Problem Description

In this section, the equations of motion and the terminal conditions of the game of two identical cars are introduced, and then, the solutions to the game of kind and the game of degree are given directly from [14].

A. Equations of Motion and Terminal Conditions

Consider the example of Merz's [14] game of two identical cars in which the speeds and maximum turn rates of both players are kept equal. Fig. 1 shows the notations used in the game, among which x and y are the relative coordinates and the y-axis is the velocity of the pursuer v_P; $\theta \in [-\pi, \pi)$ is the relative heading, which is positive if the angle from P to E is obtained clockwise; σ_1 and σ_2 are the control inputs of P and E, respectively, that maintain $|\sigma_i| \leq 1$; and $\beta > 0$ is the ratio of the capture radius to the turn radius. The line of sight (LOS) is represented by the line between the players, and the positive direction extends from the blue to the red. The polar coordinate $\phi \in [-\pi, \pi)$ is the angle measured counterclockwise from the positive y-axis to the LOS. AA $\in [0, \pi]$ represents the aspect angle, and ATA $\in [0, \pi]$ the antenna train angle, which is equal to $|\phi|$.

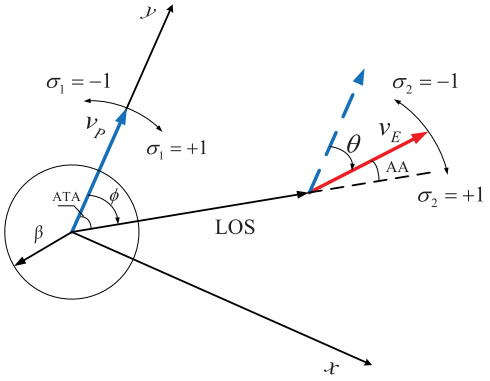


Fig. 1. Notations and coordinates.

Using these notations, the equations of relative motion of the game are found to be

$$\begin{aligned} \dot{x} &= -\sigma_1 y + \sin\theta \\ \dot{y} &= -1 + \sigma_1 x + \cos\theta \\ \dot{\theta} &= -\sigma_1 + \sigma_2. \end{aligned} \tag{1}$$

Merz [14] claims that the terminal conditions for all trajectories, when capture occurs, can be expressed as the vector

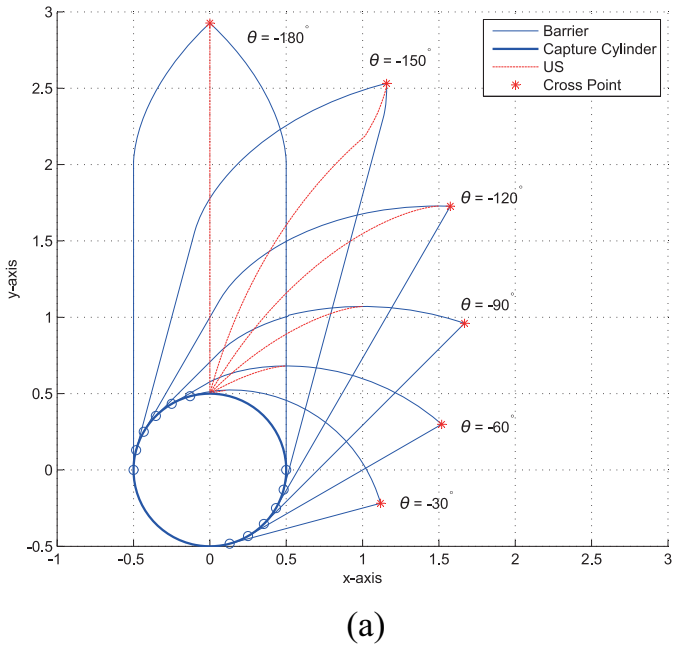
$$\mathbf{x}_0 = [\beta\sin\phi_0, \beta\cos\phi_0, \theta_0]$$

in which ϕ_0 must make the radial velocity nonpositive, that is

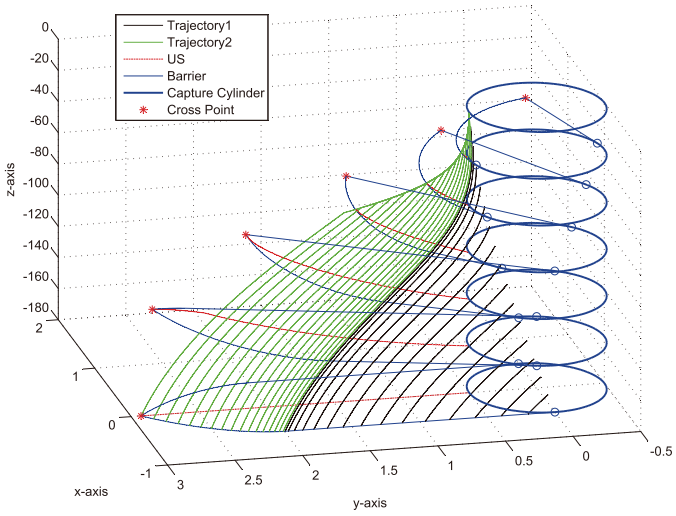
$$\dot{r}_0 = \cos(\theta_0 - \phi_0) - \cos\phi_0 \leq 0.$$

B. Results Obtained by the Differential Game Method

Merz [14] presented the results of the two cars game according to a group of cross sections of barriers, and Mitchell [26] fleshed out these results. The capture cylinder is enclosed by the barriers, and the capture region fills the region between the cylinder and the barriers. The player P will win when P is inside the region and lose when outside the region. On the barriers, none of the players can win even though both have optimal strategies. Sections are shown at $\theta = -180°$, $-150°$, $-120°$, $-90°$, $-60°$, $-30°$, and $0°$. Fig. 2(a), a reproduction of [14, Fig. 7], illustrates the game with the pursuer at the origin. The $\theta = 0°$ section is difficult to see because it is precisely the capture circle. The end points of the barriers at each section are marked by a circle. The crossover point for

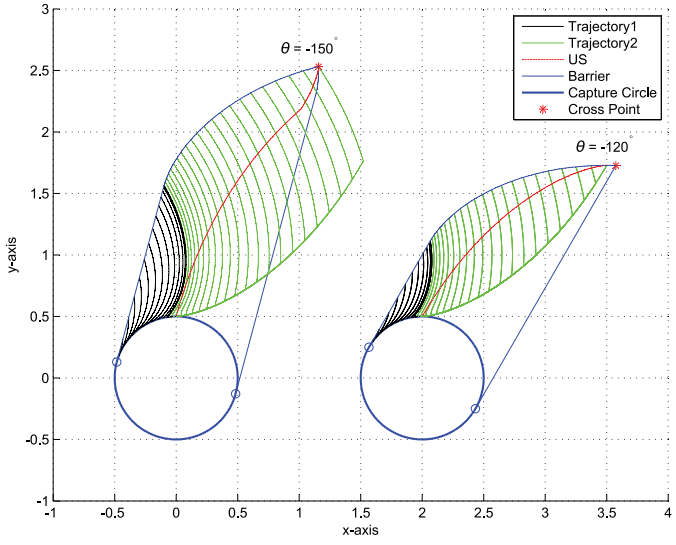


(a)

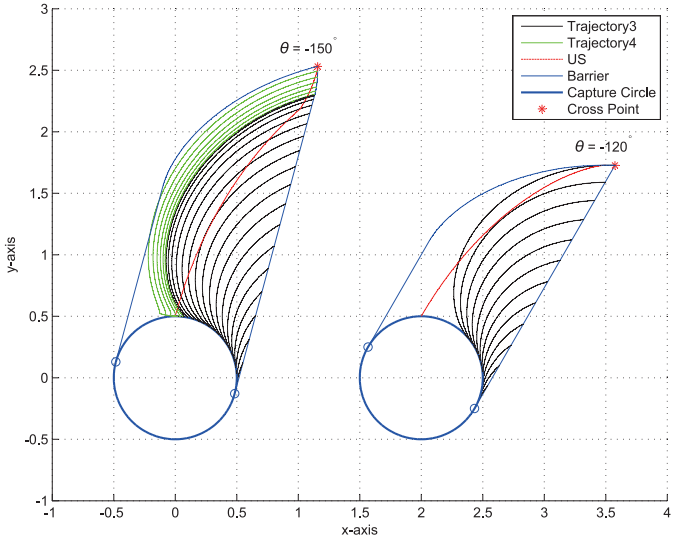


(b)

Fig. 2. Cross sections of the barriers. (a) $x-y$ view. (b) 3-D view.



(a)



(b)

Fig. 3. Trajectories from the barriers at $\theta = -150°$ and $-120°$. The shapes of $\theta = -120°$ have been moved toward the right by two units on x-axis. Trajectories from (a) left and (b) right barriers.

each section is marked with a star. To provide a better view, Fig. 2(b) is presented in three dimensions: 1) x-axis; 2) y-axis; and 3) θ-axis. A larger cluster of trajectories that start from the left barrier of the slice at $\theta = -180°$ are included in Fig. 2(b) than those in Fig. 2(a). The lines that start from the point $x = 0, y = \beta$, and extend to the left barriers are the intersection lines of the universal surface (US) and the capture region.

In the capture circle, the evader has already been captured. However, outside the capture region, i.e., beyond the barrier, P will not capture E, if E always selects the right choice. Thus, the capture region has been delineated.

Remark 1: The analytical results obtained by the differential game method are given in Fig. 2. The pursuer's controls are known to be a function of $\mathbf{x} = (x, y, \theta)$. A group of cross sections, depending on $\theta = -180°, -150°, -120°, -90°, -60°$, and $-30°$ are used to depict the solution. The US will intersect with all the sections. Then, each cross section will be divided into two subregions. In the subregion surrounded by the US, the left barrier and the capture circle, namely, the left subregion, $\sigma_1 = -1$ is the optimal strategy for P. Additionally, $\sigma_1 = 1$ corresponds to the right subregion, which is surrounded by the US, the right barrier and the capture circle.

The trajectories that start from the barriers will help us in understanding the dynamic processes of the game. To maximize the clarity of the illustrations, the trajectories from the left and the right barriers are shown separately. These trajectories are 3-D [see Fig. 2(b)].

Remark 2: Consider the projection of the trajectories on the $x-y$ plane with $\theta = -180°, -150°, -120°, -90°, -60°, -30°$, and $0°$. The trends of the trajectories' directions can be illustrated by two components: 1) "left to right ($L \to R$)" and

2) "up to down ($U \to D$)." Then, the trends of "$L \to R$" and "$U \to D$" will decrease when θ increases.

To explain this behavior, the cases at $\theta = -150°$ and $-120°$ will be used as examples. Denote the trajectories that start from the left side of the barrier as "*L*-path" and those starting from the other side as "*R*-path." The *L*-path trajectories at $\theta = -150°$ and $-120°$ are shown in Fig. 3(a), and the corresponding *R*-path trajectories in Fig. 3(b). First, the *L*-path trajectories at $\theta = -120°$ coincide with the underpart of the *L*-path trajectories at $\theta = -150°$, but are closer to the vertical angle compared to the upper part at $\theta = -150°$. In other words, the $L \to R$ will tend to decrease. Second, the right barrier at $\theta = -120°$ is closer to the horizontal line, which means that the trajectories starting from it will walk a longer distance on the *x*-axis. Thus, the right-to-left trend will increase, i.e., the $L \to R$ will tend to decrease.

III. CTMDP-Based Method

The given pursuit-evasion case, as discussed above, can be modeled as a CTMDP. In this section, the theoretical background of the CTMDP and the policy iteration algorithm will be presented.

A. Preliminaries

The definition and some properties of the CTMDP will be described first, followed by the key concept, namely, the potential performance function [11].

The CTMDP problem is a broader version of the MDP problem. It is the same as the MDP in the discrete space of states, but differs in the continuous domain of time. A typical description of the CTMDP problem is that it is a five-tuple

$$\left\{S, S_0, (A(i), i \in S), q^{\mu}(j|i), r^{\mu}(i)\right\} \tag{2}$$

where the state space S represents a finite set of fully observable states of the system, S_0 the initial state that belongs to S, and $A(i)$ a family of measurable subsets of actions applicable in $i \in S$. The policy μ consists of all the actions that have been selected in every state, i.e., $\mu = [a_1, \ldots, a_i, \ldots, a_{|S|}]$ and $\mu(i) = a_i \in A(i)$, where $|S|$ represents the total number of the states in S. $Q^{\mu} = [q^{\mu}(j|i)]$ is the infinitesimal generator of the transition probabilities that maintains the following equation:

$$P^{\mu}(t) = \exp\left(tQ^{\mu}\right). \tag{3}$$

The reward function $r^{\mu} = [r^{\mu}(i)]$ describes the immediate rewards when selecting actions $\mu(i) = a_i$. Then, the expected average reward function $\bar{J}^{\mu}$ under a policy μ, which must be maximized by finding an optimal policy $\mu = \mu^*$, is defined as the sum of the future reward over an infinite horizon

$$\bar{J}^{\mu}(i) \coloneqq \lim_{T\to\infty} \frac{1}{T} E\left\{\int_0^T r^{\mu}(X_t)dt | X_0 = i\right\}. \tag{4}$$

The CTMDP theory shows that $\bar{J}^{\mu}(i)$ converges to a value denoted as η^{μ} for any initial state shown by the following formula:

$$\eta^{\mu} = \bar{J}^{\mu}(i) = \pi^{\mu} r, w.p.1 \tag{5}$$

where π^{μ} is the steady-state probability using the policy. Moreover, η^{μ} allows

$$\eta = \lim_{l\to\infty}, w.p.1, \eta_l = \frac{1}{T_l}\int_0^{T_l} r(X_t)dt. \tag{6}$$

In this paper, we have introduced the following equations regarding the steady-state probability and the infinitesimal generator:

$$\pi^{\mu} e = 1 \tag{7}$$

$$\pi^{\mu} Q^{\mu} = 0 \tag{8}$$

$$Q^{\mu} e = 0. \tag{9}$$

The Markov theory shows that a Markov process can be described in two separate parts: 1) an embedded Markov chain and 2) the sojourn time of each state. More specifically, the transition probabilities of the embedded Markov chain P satisfy the equation

$$Q^{\mu} = \Lambda^{\mu}\left(P^{\mu} - I\right) \tag{10}$$

in which $\Lambda^{\mu} = \text{diag}[\lambda_1^{\mu}$ and $\lambda_2^{\mu}, \ldots, \lambda_n^{\mu}]$ and $\lambda_i^{\mu} > 0$ indicate the transition rate at state i. The time length of the stay in state i denoted as τ_i (sojourn time), satisfies the probability distribution

$$p\left(\tau_i^{\mu} \le t\right) = 1 - e^{-\lambda^{\mu}(i)t}. \tag{11}$$

For the transition probability matrix, the theory of Perron–Frobeniu [27] shows some characteristics of its eigenvalues.

Remark 3: By incorporating the temporal information, CTMDP can facilitate reducing the number of states compared to those in MDP.

Recall (6), which is used in calculating the expected average reward μ of the CTMDP. If all the sojourn times are assumed to be equal, then, (6) will become $\eta = \lim_{L\to\infty}(1/L)\sum_{l=0}^{L-1} r(X_l)$, which is the same as the one for the MDP. Hence, MDP is a special case of CTMDP that has a fixed transition time. Consider a fragment of an MDP chain $\{X_1, X_2, X_3, X_4, X_5, X_6\}$ with a fixed transition time Δ_t. Assume that these states can be assigned to three groups—$\{X_{1,2,3}\}$, $\{X_{4,5}\}$, and $\{X_6\}$—and that the states in the same group have the same immediate reward. This assumption regarding the reward is easy to sustain. A more intense example of this occurs frequently in the reinforcement learning problem, in which only the state that can result in failure will receive a -1 reward, whereas the other states will receive only a null reward [28]. Based on this division, we can obtain an equivalent CTMDP fragment with only three states, whose sojourn times are $\{3\Delta_t, 2\Delta_t, \Delta_t\}$. From the above analysis, we find that CTMDP can reduce the number of states and bring lesser influence on the expected average reward.

Obviously, too much reduction will render the problem insolvable. Consider the extreme case of the pendulum problem in [28], wherein the state space consists of only one state, and the policy for the agent is "10 *N*" or "-10 *N*," regardless of the policy chosen. It will lead the system to failure. Thus, reducing the number of states is a tradeoff between the speed

of algorithm and the performance of solution. Complicated as it is, this problem will be studied in our future work.

Lemma 1: Let P be the transition probability matrix of an irreducible Markov chain. Then, the number 1 is a simple eigenvalue of P. For any other eigenvalue α of P, we have $|\alpha| \leq 1$. If P is aperiodic, then $|\alpha| < 1$ for all the other eigenvalues of P.

The CTMDP would thus satisfy the Poisson equation

$$Q^{\mu} g^{\mu} = -r^{\mu} + \eta^{\mu} e \tag{12}$$

in which $g^{\mu}(i)$ is the potential performance function of state i. Moreover, for a CTMDP, the potential performance function has the formula

$$g^{\mu}(i) = \lim_{T \to \infty} E\left\{ \int_0^T \left\{ r^{\mu}(X_t) - \eta^{\mu} \right\} dt | X_0 = i \right\}. \tag{13}$$

B. Policy Iteration Algorithm

Two basic algorithms can be used to obtain the optimal solution for the traditional MDP problem: the value iteration algorithm and the policy iteration algorithm. In this paper, we consider only the policy iteration algorithm; therefore, a policy algorithm for the CTMDP problem will be presented in this section.

First, the formula describing the reward difference between two policies, namely, μ_1 and μ_2, is given. With the property (7) and (8), we have

$$\begin{aligned} \eta^{\mu_1} - \eta^{\mu_2} &= \pi^{\mu_1} r^{\mu_1} - \pi^{\mu_1} e \eta^{\mu_2} \\ &= \pi^{\mu_1} \left[r^{\mu_1} - \left(r^{\mu_2} + Q^{\mu_2} g^{\mu_2} \right) \right] \\ &= \pi^{\mu_1} \left[\left(r^{\mu_1} + Q^{\mu_1} g^{\mu_2} \right) - \left(r^{\mu_2} + Q^{\mu_2} g^{\mu_2} \right) \right]. \end{aligned} \tag{14}$$

Definition 1: For vectors with the same dimension, it is denoted as $u \preceq v$ if at least one entry sustains $u(i) < v(i)$, and the remaining are equal. $u \succeq v$ can be defined in the same way.

Lemma 2: Assume that μ_1 and μ_2 are two policies of the CTMDP problem. If $r^{\mu_1} + Q^{\mu_1} g^{\mu_2} \succeq r^{\mu_2} + Q^{\mu_2} g^{\mu_2}$, then $\eta^{\mu_1} > \eta^{\mu_2}$. Moreover, if $r^{\mu_1} + Q^{\mu_1} g^{\mu_2} \preceq r^{\mu_2} + Q^{\mu_2} g^{\mu_2}$, then $\eta^{\mu_1} < \eta^{\mu_2}$.

Proof: According to the Definition 1, (14) and the fact that the steady-state probability of an ergodic Markov process is greater than zero, the lemma can be easily obtained. ■

Lemma 3: Consider a CTMDP problem. Policy μ^* is called an optimal policy, if and only if for any $\mu \in D$

$$r^{\mu^*} + Q^{\mu^*} g^{\mu^*} \succeq r^{\mu} + Q^{\mu} g^{\mu^*}. \tag{15}$$

Proof: We first prove the necessary condition. If μ^* is assumed as the optimal policy, then we need to prove that (15) holds. Assume that there exists a policy μ' that does not make (15) hold, i.e., at least, there exists a state i maintaining that

$$r^{\mu^*}(i) + \sum_{j=1}^{S} q^{\mu^*}(j|i) g^{\mu^*}(j) < r^{\mu'}(i) + \sum_{j=1}^{S} q^{\mu'}(j|i) g^{\mu^*}(j).$$

Construct a new policy $\hat{\mu}$ in this way. $\hat{\mu}(j) = \mu^*(j)$, while $j \neq i$; $\hat{\mu}(j) = \mu'(j)$, while $j = i$. Thus, we can obtain $r^{\mu^*} + Q^{\mu^*} g^{\mu^*} \succeq r^{\mu} + Q^{\mu} g^{\mu^*}$. By applying Lemma 2, we can see that $\eta^{\hat{\mu}} > \eta^{\mu^*}$, which contradicts the previous assumption.

Next, we will present proof of the sufficient condition. Suppose (15) holds. By applying Lemma 2, we can see that $\eta^{\mu^*} > \eta^{\mu}$ for any $\mu \in D$. Hence, μ^* is the optimal policy. ■

Second, using the previous description as a theoretical foundation, the CTMDP-based policy iteration algorithm is presented below.

Algorithm 1 CTMDP-Based Policy Iteration Algorithm

Require: Q^{μ}, r^{μ}

1: Select an initial policy μ_0, and set $k = 0$, $g^{\mu_0} = 0$.
2: Obtain the potential performance function g^{μ_k}.
  Obtain π^{μ} by formula (7, 8);
  Obtain η^{μ} by formula (5);
  Obtain g^{μ_k} by the Poisson equation:
$$Q^{\mu} g^{\mu} = -r^{\mu} + \eta^{\mu} e. \tag{16}$$
3: Improve the policy by: $\mu_{k+1} = \underset{\mu \in D}{\operatorname{argmax}} \{ r^{\mu} + Q^{\mu} g^{\mu_k} \}$
4: **if** $\mu_{k+1} = \mu_k$ **then**
5: Set $\mu^* = \mu_{k+1}$
6: Stop.
7: **else**
8: Set $k = k + 1$;
9: Jump to Step 2.
10: **end if**

Corollary 1: The result of the Algorithm 1 is the same when the transition rate matrix Λ^{μ} is multiplied by a constant $c > 0$.

Proof: Suppose that the transition rate matrix is replaced by $(\Lambda')^{\mu} = c\Lambda^{\mu}$. The reward r^{μ} is thus the same. From (7), (8), and (10), we can obtain

$$\pi^{\mu} Q^{\mu} = \pi^{\mu} \Lambda^{\mu} (P^{\mu} - I) = 0 \tag{17}$$

$$\left(\pi'\right)^{\mu} \left(Q'\right)^{\mu} = \left(\pi'\right)^{\mu} c \Lambda^{\mu} \left(P^{\mu} - I\right) = 0. \tag{18}$$

First, we consider only the irreducible Markov chain. P^{μ} is a random and non-negative matrix, according to Lemma 1. Then, we can see that the eigenvalues of P^{μ} are $1, \alpha_2, \ldots, \alpha_S$, where $|\alpha_i| < 1$ for $i = 2, \ldots, S$. It is not difficult to see that matrix $P - I$ has only a zero eigenvalue; thus, the rank of $P - I$ is $S - 1$. Additionally, $c > 0$, $\pi e = 1$, and Λ is invertible. Then, $\pi^{\mu} = (\pi')^{\mu}$. Referring to (5), if r does not change, we can see that $(\eta')^{\mu} = \eta^{\mu}$. Furthermore, from (16), we obtain $r^{\mu} + (Q')^{\mu} g^{\mu_k} = r^{\mu} + Q^{\mu} g^{\mu_k}$. Hence, the policy improved by step 3 in Algorithm 1 also does not change. The corollary is thus proved. ■

Proposition 1: For a discrete CTMDP, Algorithm 1 will converge to the optimal solution in finite steps.

Proof: Rewrite the equation in step 3 as

$$\mu_{k+1}(i) = \underset{a \in A(i)}{\operatorname{argmax}} \left\{ r(i,a) + \sum_{j \in S} q(j|i,a) g^{\mu_k}(i) \right\}.$$

For $\mu_{k+1} \neq \mu_k$, it is easy to see that the ith entry of $r^{\mu_{k+1}} + Q^{\mu_{k+1}} g^{\mu_k}$ is larger than the ith entry of $r^{\mu_k} + Q^{\mu_k} g^{\mu_k}$, i.e., $r^{\mu_{k+1}} + Q^{\mu_{k+1}} g^{\mu_k} \succeq r^{\mu_k} + Q^{\mu_k} g^{\mu_k}$ has been satisfied.

Algorithm 2 MDP-Based Policy Iteration Algorithm

Require: P^{μ}, r^{μ}

1: Select an initial policy μ_0, and set $k = 0$, $g^{\mu 0} = 0$.
2: Obtain the potential performance function g^{μ_k}.
Obtain π^{μ} by the balance equation $\pi P = \pi$ and $\pi e = 1$;
Obtain η^{μ} by $\eta = \pi r$;
Obtain g^{μ_k} by the Poisson equation:

$$(I - P^{\mu})g^{\mu} + \eta^{\mu} e = r^{\mu}. \tag{19}$$

3: Improve the policy by: $\mu_{k+1} = \underset{\mu \in D}{argmax}\{r^{\mu} + P^{\mu} g^{\mu_k}\}$
4: **if** $\mu_{k+1} = \mu_k$ **then**
5: Set $\mu^* = \mu_{k+1}$
6: Stop.
7: **else**
8: Set $k = k + 1$;
9: Jump to Step 2.
10: **end if**

According to Lemma 3, we then obtain $\eta^{\mu_{k+1}} > \eta^{\mu_k}$. Because the number of decisions μ is finite, the iteration will terminate at the optimal value of η^{μ^*} in finite steps. ■

C. Compare With MDP-Based Policy Iteration Algorithm

The MDP-based policy iteration algorithm, which used the concept of potential performance as Algorithm 2, is presented [11]. The meanings of the notations are the same as those used for the CTMDP-based algorithm.

Proposition 2: Consider a CTMDP problem and its embedded MDP problem. If the matrix of transition rates satisfies $\Lambda = I$, then the results of the CTMDP-based policy iteration algorithm and the MDP-based policy iteration algorithm will be equal.

Proof: According to (10) and because $\Lambda = I$, we have $P = I + Q$. Replace P^{μ} with $I + Q^{\mu}$ in the balance equation and Poisson equation at step 2. Then, we can obtain $\pi Q^{\mu} = 0$ and $Q^{\mu} g^{\mu} = -r^{\mu} + \eta^{\mu} e$, which are the same as in those Algorithm 1. Next, after performing the same replacement, step 3 in Algorithm 2 becomes $\mu_{k+1} = \underset{\mu \in D}{\text{argmax}}\{r^{\mu} + Q^{\mu} g^{\mu_k} + g^{\mu_k}\}$, that will return μ_{k+1}, which is the same as that produced by step 3 in Algorithm 1. Thus, the proposition is proved. ■

IV. Applying CTMDP-Based Method to the Game of Two Identical Cars

In this section, the space of states, actions, rewards, transition probability matrix, and transition rate matrix, which are necessary for a CTMDP, are constructed.

A. Space of States, Actions, and Rewards

Consider the problem in the relative coordinates shown in Fig. 1. The states vector in the game of two identical cars is continuous. To adopt the CTMDP, discretization is needed. If too much of the capacity for expressing the solution is not lost and the number of states does not become too high, a grid, which overlaps all the cross sections described in Section II-B, can be used to define the space of states.

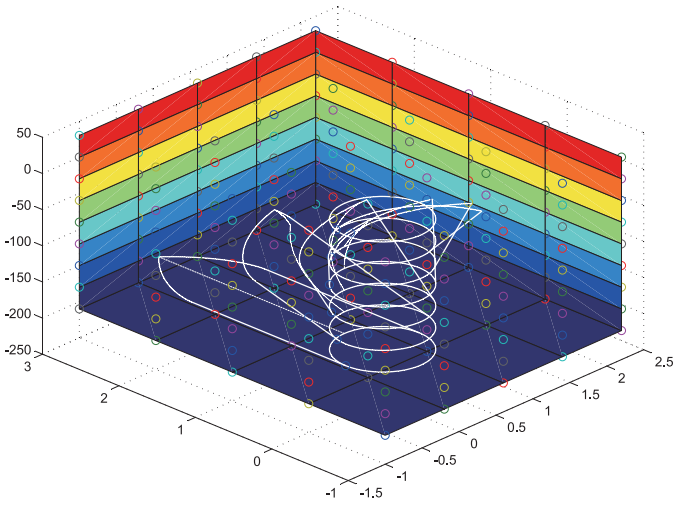


Fig. 4. Discrete space of states in CTMDP.

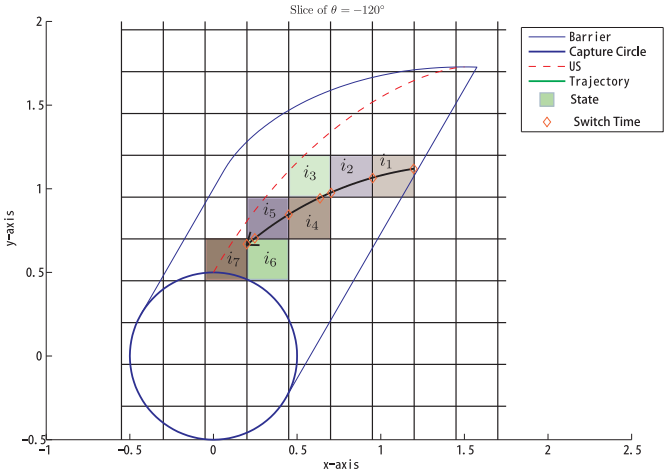


Fig. 5. Example of the trajectory on the $\theta = -120°$ slice.

Use the sequence $x_0 = -\infty, x_1, \ldots, x_{m-1}, x_m = +\infty$ to divide the x-axis into m intervals as $I_l^x = [x_{l-1}, x_l), l = 1, \ldots, m$, and thus, the intervals on the y-axis can be obtained as $I_l^y = [y_{l-1}, y_l), l = 1, \ldots, n$, where $y_0 = -\infty$ and $y_n = +\infty$. θ-axis can be divided into intervals $I_l^{\theta} = [\theta_{l-1}, \theta_l)$, $l = 1, \ldots, p$, where $\theta_0 = -210°$ and $\theta_p = 0°$. Assuming that state i of CTMDP is related to intervals $I_{\iota_1}^x$, $I_{\iota_2}^y$, and $I_{\iota_3}^{\theta}$, these state vectors $\mathbf{x} = (x, y, \theta)$ ensure that $x \in I_{\iota_1}^x$, $y \in I_{\iota_2}^y$, and $\theta \in I_{\iota_3}^{\theta}$ will correspond to state i. For a short notation, we denote state i as $i = (\iota_1, \iota_2, \iota_3)$. Except for the first and last intervals in x, y, and θ axes, the remaining intervals will be considered as Δ_x, Δ_y, and Δ_{θ}, respectively. Hence, the continuous space is separated by a grid (see Fig. 4).

Player P can perform three actions, namely, "turn left," "go straight," and "turn right," which are denoted as "L," "S," and "R," corresponding to the controls $\sigma_1 = -1$, $\sigma_1 = 0$, and $\sigma_1 = +1$, respectively.

Remark 4: This remark will use an example to show that the transition times are not fixed. Given the initial state $\mathbf{x_0} = (1.2, 1.12, -120°)$ and controls $\sigma_1 = \sigma_2 = +1$, the trajectory in Fig. 5 is generated according to (1).

Under the definition of the space of states mentioned above, those states through which the trajectory has passed are denoted as $i_1, i_2, \ldots, i_7$. The switch times between every two successive states are marked by diamonds and calculated to be $[0, 0.1275, 0.2585, 0.2915, 0.4000, 0.5265, 0.5540]$. Thus, the stay times in states from i_1 to i_6 are $[0.1275, 0.1310, 0.0330, 0.1085, 0.1265, 0.0275]$. The transition time starts when the current state ends and lasts until the next state arrives, which is the same as the stay time. This example thus explains why the transition times are not fixed in this paper.

Defining a reward function, an expert heuristic function that is used to reasonably capture the relative merits of every possible state in the adversarial game, is helpful for generating a better solution [6]. The reward in state i is denoted as $r(i)$, which is the product of the angle-related part and the distance-related part

$$r(i) = \left(1 - \frac{\text{ATA}}{180}\right)\exp\left(-\frac{d}{\delta_d}\right) \tag{20}$$

where d satisfies that

$$d = \sqrt{x^2 + y^2}, \text{ATA} = \cos^{-1}\frac{y}{d} \text{ (rad)}.$$

It can be seen that the states with lower values of ATA and d will have a higher reward, which is reasonable for the capture process.

B. Construction of Matrix Q

In this section, the transition probability matrix P^μ and the transition rate matrix Λ^μ are constructed to match with reality. Then, matrix Q can be obtained by (10).

Suppose that state $i = (\imath_1, \imath_2, \imath_3)$. Then, states $j = (\imath_1 + \Delta_1, \imath_2 + \Delta_2, \imath_3 + \Delta_3)$ are the adjoining states in which $\Delta_1, \Delta_2, \Delta_3 \in \{-1, 0, 1\}$. Because the state vector changes continuously, as mentioned in Section IV-A, all the states in the gridded space cannot be transferred to other states through any nonadjoining states. Hence, we can reasonably assume that the transition probability between two nonadjoining states is zero.

Consider the transition probabilities from state i to the adjoining state j. To simplify the construction, the changes in x, y, and θ from i to j, denoted as Δ_1, Δ_2, and Δ_3, are assumed to be conditionally independent depending on i'θ. Then, the transition probability can be written as

$$\begin{aligned} p(j|i,a) &:= p((\jmath_1 - \imath_1, \jmath_2 - \imath_2, \jmath_3 - \imath_3)|\imath_3, a) \\ &= p(\jmath_1 - \imath_1|\imath_3, \jmath_2 - \imath_2|\imath_3, \jmath_3 - \imath_3|\imath_3, a) \\ &= p(\Delta_1|\imath_3, a)p(\Delta_2|\imath_2, a)p(\Delta_3|\imath_3, a). \end{aligned} \tag{21}$$

Intuitively, the values of Δ_1 and Δ_2 have no obvious causations relative to the selected action a because there are no limits on the values of x and y after action a. Hence, we can assume that $p(\Delta_1|\imath_3) = p(\Delta_1|\imath_3, a)$ and $p(\Delta_2|\imath_3) = p(\Delta_2|\imath_3, a)$. Furthermore, the change in angle θ depends heavily on the actions. The effect of $\imath_3$ will be ignored. Thus, (21) can be written as

$$p(j|i,a) = p(\Delta_1|\imath_3)p(\Delta_2|\imath_2)p(\Delta_3|a). \tag{22}$$

Next, the elements in (22) will be calculated. Define the matrices P_{Δ_1} and P_{Δ_2}, where the rows represent $\imath_3$ (corresponding to $\theta = -210°, -180°, \ldots, 0°$) and the columns Δ_1 and Δ_2. The left-side columns, from left to right, show that $\Delta_1 = -1, 0, +1$, respectively. The matrix P_{Δ_3} is defined for probability distribution of $p(\Delta_3|a)$, where the rows represent a and the columns Δ_3.

Recall the sojourn time τ_i^μ at state i in (11). The expected value of τ_i^μ is equal to $1/\lambda^\mu(i)$. Based on this, matrix Λ^μ will be given in the next section. Corollary 1 shows that multiplying the transition rates by a constant $c > 0$ will not change the optimal policy. Therefore, we can limit the value of the transition rate to $\lambda^\mu(i) \in (0, 1]$.

V. Numerical Results

To verify the effectiveness of the proposed CTMDP-based method in solving the pursuit-evasion problem, some numerical experiments are conducted. First, a numerical simulation of the MDP-based method will be presented. For comparison, an example of the CTMDP-based method with different prior probabilities produced by the situation is also described.

A. Parameters Setting and Simulation Results

For the game of two identical cars with $\beta = 0.5$ and $\delta_d = 0.25$, the scales of x, y, and θ axes are first set as $\Delta_x = \Delta_y = 0.25$ and $\Delta_\theta = 30°$, and the bounds as $x_0 = -0.55$, $x_m = 1.75$, $y_0 = -0.5$, $y_n = 3$, and $\theta_0 = -210°$, $\theta_p = 0°$ (to reduce the influence of the end points, we consider the bounds of θ more than $[-180°, -30°]$). Approximately, 1200 states are generated.

The matrices P_{Δ_1} and P_{Δ_2} are selected as

$$P_{\Delta_1} = \frac{1}{18}\begin{bmatrix} 5 & 6 & 7 \\ 6 & 6 & 6 \\ 7 & 6 & 5 \\ 8 & 6 & 4 \\ 9 & 6 & 3 \\ 10 & 6 & 2 \\ 11 & 6 & 1 \\ 12 & 6 & 0 \end{bmatrix} \quad P_{\Delta_2} = \frac{1}{18}\begin{bmatrix} 11 & 6 & 1 \\ 12 & 6 & 0 \\ 11 & 6 & 1 \\ 10 & 6 & 2 \\ 9 & 6 & 3 \\ 8 & 6 & 4 \\ 7 & 6 & 5 \\ 6 & 6 & 6 \end{bmatrix}. \tag{23}$$

θ of the first row is $-210°$ and is symmetrical to the case at $150°$ (the results of the game of two identical cars show that the cases at $\theta < -180°$ are symmetrical to the ones at $\theta > -180°$). Except for the first row, the values in the left column of P_{Δ_1} increase with the value of θ from $-180°$ to $0°$. Thus, as θ increases, the probability that the x-coordinate of state i transfers from $\imath$ to $\imath - 1$ also increases. Because of the symmetry, the right column is inverse. This agrees with Remark 2. Similarly, P_{Δ_2} also agree with Remark 2.

Third, the matrix P_{Δ_3} is

$$P_{\Delta_2} = \frac{1}{6}\begin{bmatrix} 3 & 2 & 1 \\ 2 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix}. \tag{24}$$

It is difficult to determine the opponent's actions, and therefore, as a compromise, we assume that the opponent does not move. According to Fig. 1, the pursuer's action L will

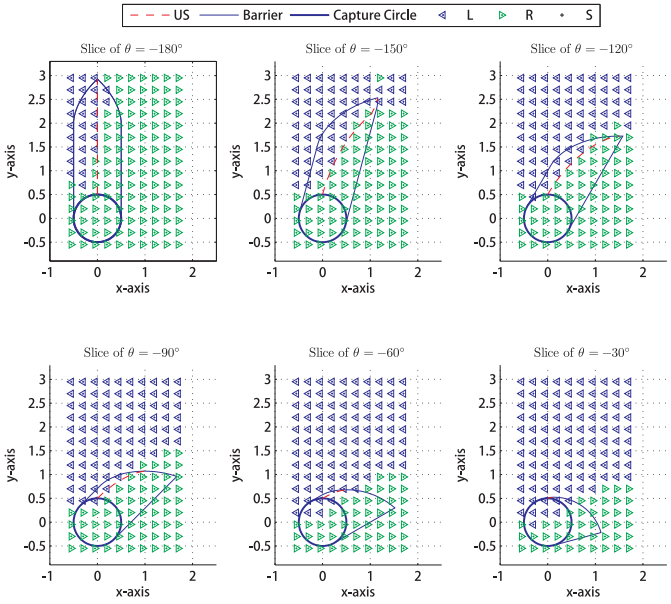


Fig. 6. Policy obtained by the MDP-based method.

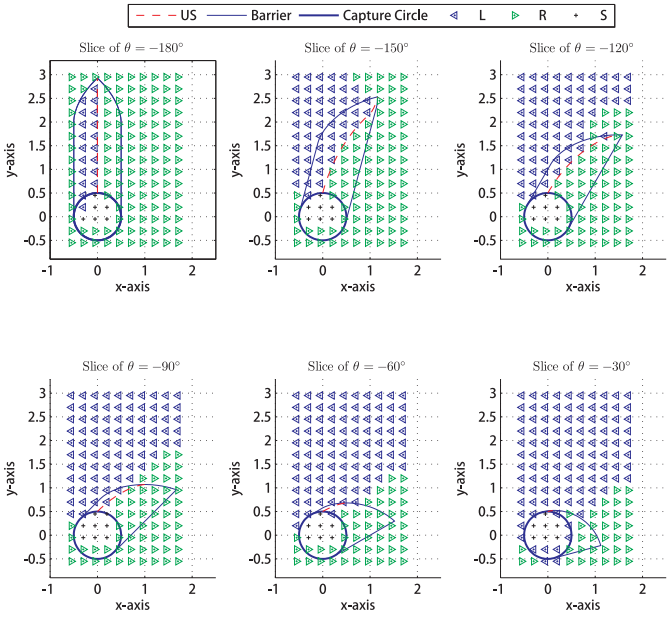


Fig. 7. Policy obtained by the CTMDP-based method.

cause a decrease in θ. Then, we can set the first row of P_{Δ_3} as $3/6, 2/6, 1/6$, which indicates that the probabilities satisfy "$\Delta_3 = -1$" > "$\Delta_3 = 0$" > "$\Delta_3 = +1$." The second row represents a uniform distribution of θ with action S. In addition, the pursuer's action R will cause an increase in θ. Thus, the third row is the inverse of the first row because of symmetry.

The transition rate matrix Λ^{μ} will be set by assuming that if action a is optimal for state i, then the sojourn time τ_i^a is the minimum for all allowable actions. Intuitively, the optimal policy will drive the game to the end state as soon as possible. Thus, the transition rate will be maximized. In this paper, if action L is optimal, then the transition rates are $\lambda^a(i) = 1, 0.1, 0.1$ with respect to $a = L, S, R$, respectively; if action R is optimal, then $\lambda^a(i) = 0.1, 0.1, 1$ for $a = L, S, R$; if the action S is optimal, then $\lambda^a(i) = 0.1, 1, 0.1$ for $a = L, S, R$.

The following are some figures illustrating the results of the game of two identical cars based on the above-mentioned parameter settings, in which the pursuer's actions L, R, and S are shown, respectively, as left-facing triangles, right-facing triangles, and crosses. The analytical solution obtained by the differential game method is also included in the figures. As mentioned in Remark 1, P will choose an action L in the left subregion and R in the right subregion.

Fig. 6 shows the results produced by the MDP-based method. Although P tries to make a correct decision and follows the analytical solution in most states, many states still exist in which the action of P does not match with the analytical solution.

Fig. 7 shows the results obtained by the CTMDP-based method. It can be seen that, in the left and right subregions, the selected actions of P match with the analytical solution.

Fig. 8 shows the values of the average rewards and different elements, with the previous policy, at each iteration of the MDP-based and CTMDP-based methods. After six iterations, both the algorithms converge in their results.

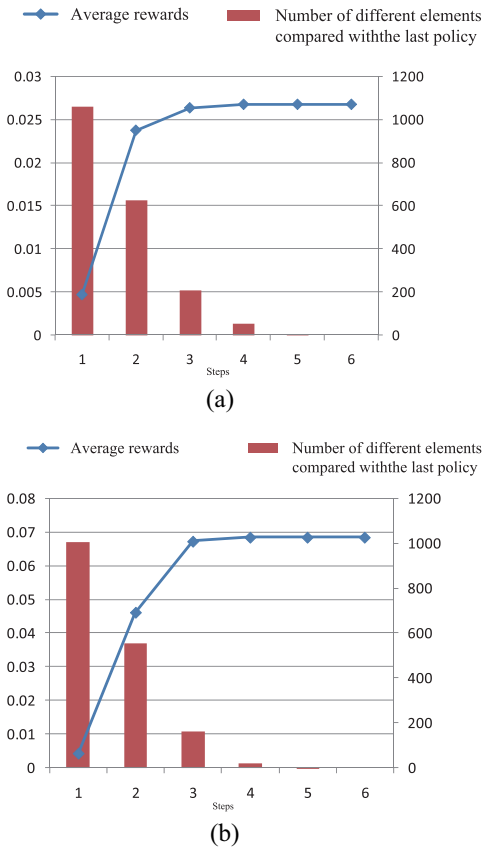


Fig. 8. Performances of the algorithms. (a) MDP-based case. (b) CTMDP-based case.

Given a drastic situation, the probability matrix P_{Δ_2} can be reversed by rows to give P'_{Δ_2}. As a result, the pursuer has poor knowledge of the game. Figs. 9 and 10 show the results produced by the MDP-based and CTMDP-based methods, respectively. The results of MDP-based method are wholly

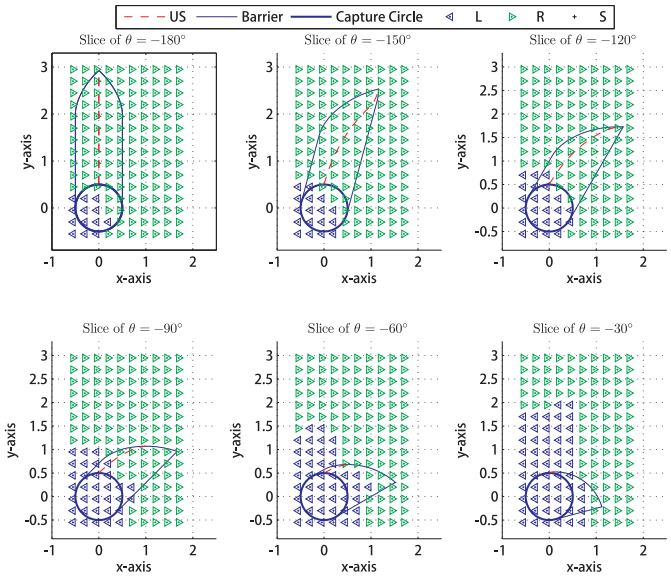


Fig. 9. Policy obtained by the MDP-based method under an inaccurate estimation of transition probabilities.

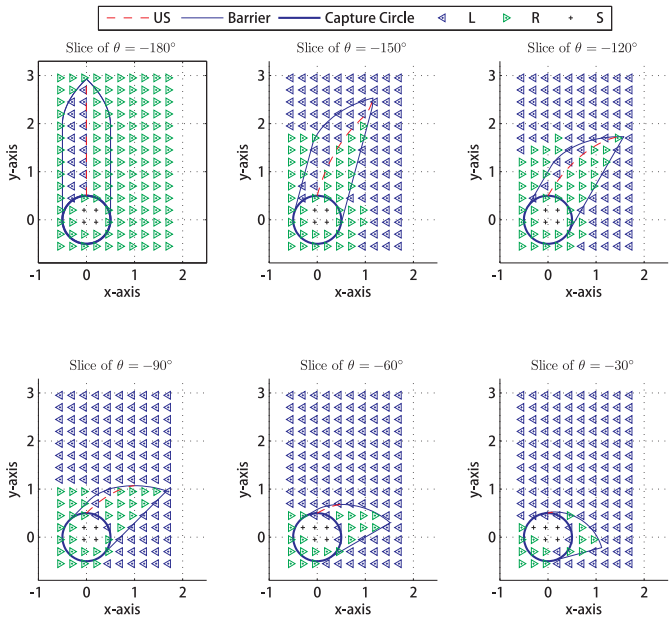


Fig. 10. Policy obtained by the CTMDP-based method under an inaccurate estimation of transition probabilities.

inappropriate for the game, whereas those of CTMDP-based method are appropriate

$$P'_{\Delta_2} = \frac{1}{18}\begin{bmatrix} 6 & 6 & 6 \\ 7 & 6 & 5 \\ 8 & 6 & 4 \\ 9 & 6 & 3 \\ 10 & 6 & 2 \\ 11 & 6 & 1 \\ 12 & 6 & 0 \\ 13 & 6 & 1 \end{bmatrix}. \tag{25}$$

B. Discussion

The MDP-based and CTMDP-based methods approach the game of two identical cars from different points of view. In differential games, each player has perfect knowledge of the opponent's information, and as a result, the optimal policy for both players will be obtained by solving the differential equations. In the MDP-based or CTMDP-based method, the behaviors of the opponents are depicted as probabilities, because they are not known exactly. In other words, in such a situation, "we have no idea about the opponent's choice at the current time, but we can know his habits"; this knowledge will be closer to reality than perfect knowledge. In this paper, the habits of the opponent are interpreted as transition probabilities and rates, which are constructed directly from some experimental results. For a more general situation, such model information can be trained by a large number of sampling trajectories; this process is known as reinforcement learning. Compared with the differential game method, the MDP-based and CTMDP-based methods produce approximate solutions; besides, they do not require model information.

Additionally, the CTMDP-based method can provide a better approximation of the results by the differential method than the MDP-based method can. The simulation results in Fig. 7 are very close to the analytical solutions in these states after the application of discretization approach. However, the results in Fig. 6 include some states that do not match with the analytical results. In addition, the inclusion of a larger number of states will lead to a better approximated solution, but it requires more time. Fortunately, many techniques exist to reduce the number of states, such as state aggregation and approximate DP [5], [29].

Finally, the simulations also show that the CTMDP-based method is more robust than the MDP-based method in coping with inaccurate estimations of the transition probabilities. The reversing of the matrix P_{Δ_2} is used to simulate dramatic deviations of the probabilities. Figs. 9 and 10 show that the CTMDP-based method remains functional, while the MDP-based method ceases to be functional. This observation supports the main theme of this paper: not only the transition probabilities but also the transition rates are the key factors in dynamic processes.

VI. Conclusion

This paper presents a novel framework, the CTMDP-based method, in which the game of two identical cars is modeled as a CTMDP. Qualitative analysis of the analytical solutions is used to construct the transition probability matrices and transition rate matrices. The main contribution of this paper is the policy iteration algorithm for CTMDP, based on the concept of potential performance. In the numerical examples, the policies produced by the CTMDP-based method are better than those produced by the MDP-based methods; besides, they remain robust against probability changes.

The transition probabilities and rates in this paper are constructed in an intuitive way, although, a more practical method is needed to adaptively estimate such information. Our future work will focus on using the reinforcement learning technique to replace the direct construction method used here.

A potential difficulty in achieving this goal will be how to quickly evaluate the potential function based on sampling trajectories. Consequently, the method for developing a policy in the CTMDP should be extended to more applications.

REFERENCES

[1] P. W. Koken, H. J. J. Jonker, and C. J. Erkelens, "A model of the human smooth pursuit system based on an unsupervised adaptive controller," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 26, no. 2, pp. 275–280, Mar. 1996.

[2] P. Kachroo, S. A. Shedied, and H. Vanlandingham, "Pursuit evasion: The herding noncooperative dynamic game—The stochastic model," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 32, no. 1, pp. 37–42, Feb. 2002.

[3] R. Isaacs, *Differential Games: A Mathematical Theory with Application to Warfare and Pursuit, Control and Optimization*. New York, NY, USA: Wiley, 1965.

[4] J. P. Hespanha, M. Prandini, and S. Sastry, "Probabilistic pursuit-evasion games: A one-step Nash approach," in *Proc. 39th IEEE Conf. Decis. Control*, vol. 3. Sydney, NSW, Australia, Dec. 2000, pp. 2272–2277.

[5] D. P. Bertsekas and J. N. Tsitsiklis, "Neuro-dynamic programming," in *Anthropological Field Studies*. Belmont, MA, USA: Athena Scientific, 1996.

[6] J. S. McGrew, J. P. How, B. Williams, and N. Roy, "Air combat strategy using approximate dynamic programming," *AIAA J. Guid. Control Dyn.*, vol. 33, no. 5, pp. 1641–1654, 2010.

[7] S. Jia, X. Wang, X. Ji, and H. Zhu, "A continuous-time Markov decision process based method on pursuit-evasion problem," in *Proc. 19th IFAC World Congr.*, vol. 19. Cape Town, South Africa, Aug. 2014, pp. 620–625.

[8] X. P. Guo, *Continuous-Time Markov Decision Processes*. Berlin, Germany: Springer, 2009.

[9] R. A. Howard, *Dynamic Programming and Markov Processes*. Cambridge, MA, USA: MIT Press, 1960.

[10] R. S. Sutton, D. Precup, and S. Singh, "Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning," *Artif. Intell.*, vol. 112, nos. 1–2, pp. 181–211, Aug. 1999.

[11] X. R. Cao, *Stochastic Learning and Optimization: A Sensitivity-Based Approach*. New York, NY, USA: Springer, 2008.

[12] E. Cockayne, "Plane pursuit with curvature constraints," *SIAM J. Appl. Math.*, vol. 15, no. 6, pp. 1511–1516, 1967.

[13] A. W. Merz, "The homicidal chauffeur: A differential game," Ph.D. dissertation, Dept. Aeronaut. Astronaut., Stanford Univ., Stanford, CA, USA, 1971.

[14] A. W. Merz, "The game of two identical cars," *J. Optim. Theory Appl.*, vol. 9, no. 5, pp. 324–343, May 1972.

[15] J. Shinar and S. Gutman, "Three-dimensional optimal pursuit and evasion with bounded controls," *IEEE Trans. Autom. Control*, vol. 25, no. 3, pp. 492–496, Jun. 1980.

[16] T. Miloh, "A note on three-dimensional pursuit-evasion game with bounded curvature," *IEEE Trans. Autom. Control*, vol. 27, no. 3, pp. 739–741, Jun. 1982.

[17] T. Miloh, M. Pachter, and A. Segal, "The effect of a finite roll rate on the miss-distance of a bank-to-turn missile," *Comput. Math. Appl.*, vol. 26, no. 6, pp. 43–54, Sep. 1993.

[18] J. E. Pang, "Pursuit evasion with acceleration, sensing limitation, and electronic counter measures," M.S. thesis, Holcombe Dept. Elect. Comput. Eng., Clemson Univ., Clemson, SC, USA, 2007.

[19] R. Vidal, "Probabilistic pursuit-evasion games: Theory, implementation and experimental evaluation," *IEEE Trans. Robot. Autom.*, vol. 18, no. 5, pp. 662–669, Oct. 2002.

[20] J. P. Hespanha, J. K. Hyoun, and S. Sastry, "Multiple-agent probabilistic pursuit-evasion games," in *Proc. 38th IEEE Conf. Decis. Control*, vol. 3. Phoenix, AZ, USA, Dec. 1999, pp. 2432–2437.

[21] S. M. LaValle, *Planning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2006.

[22] S. J. Russell, P. Norvig, and E. Davis, *Artificial Intelligence: A Modern Approach* (Prentice Hall Series in Artificial Intelligence). Upper Saddle River, NJ, USA: Prentice Hall, 2010.

[23] L. F. Bertuccelli, A. Wu, and J. P. How, "Robust adaptive Markov decision processes: Planning with model uncertainty," *IEEE Control Syst. Mag.*, vol. 32, no. 5, pp. 96–109, Oct. 2012.

[24] S. J. Bradtke and M. O. Duff, "Reinforcement learning methods for continuous-time Markov decision problems," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 1994, pp. 393–400.

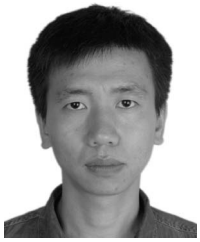
[25] S. Mahadevan, N. Marchalleck, T. K. Das, and A. Gosavi, "Self-improving factory simulation using continuous-time average-reward reinforcement learning," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 1997, pp. 202–210.

[26] I. Mitchell, "Games of two identical vehicles," Dept. Aeronaut. Astronaut., Stanford Univ., Stanford, CA, USA, Tech. Rep. SUDAAR#740, 2001.

[27] E. Çinlar, *Introduction to Stochastic Processes*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1975.

[28] A. Barto, R. Sutton, and C. Anderson, "Neuronlike adaptive elements that can solve difficult learning control problems," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-13, no. 5, pp. 834–846, Sep./Oct. 1983.

[29] D. Bertsekas, *Dynamic Programming and Optimal Control*. Belmont, MA, USA: Athena Scientific, 2001.



Shengde Jia received the B.S. degree in automatic control from the Harbin Institute of Technology, Harbin, China, in 2008, and the M.S. degree from the National University of Defense Technology, Changsha, China, in 2011, where he is currently pursuing the Ph.D. degree with the College of Mechatronic Engineering and Automation.

His current research interest include machine learning.

Xiangke Wang (M'12) received the B.S., M.S., and Ph.D. degrees in automatic control from the National University of Defense Technology, Changsha, China, in 2004, 2006, and 2012, respectively.

He is currently an Associate Professor with the College of Mechatronic Engineering and Automation, National University of Defense Technology. His current research interests include coordination and control of multiple unmanned aerial vehicles, cooperative decision, and nonlinear control.

Lincheng Shen (M'10) received the B.E., M.S., and Ph.D. degrees in automatic control from the National University of Defense Technology (NUDT), Changsha, China, in 1986, 1989, and 1994, respectively.

In 1989, he joined the Department of Automatic Control, NUDT, where he is currently a Full Professor and serves as the Dean of the College of Mechatronic Engineering and Automation. He has published over 100 technical papers in refereed international journals and academic conferences proceedings. His current research interests include mission planning, autonomous and cooperative control for unmanned systems, biomimetic robotics, and intelligent control.

Mr. Shen has initiated and organized several workshops and symposia, including the International Workshop on Bionic Engineering in 2012 and the Chinese Automation Congress in 2013. He has been serving as an Editorial Board Member of the *Journal of Bionic Engineering* since 2007.