強化学習により獲得したファジィ目標に基づく四輪車の移動制御

Four-wheeled Vehicle Control based on Fuzzy Target acquired by Reinforcement Learning

筑波大学理工学研究科 ○松原 智也　　筑波大学機能工学系 安信 誠二

○ Tomoya Matsubara and Seiji Yasunobu,　　University of Tsukuba

Abstract: The target setting knowledge acquired by trial and error in reinforcement learning is including the degree of success for the achievement of the task. In this paper, we focus attention on this point and handled the target setting knowledge (fuzzy set) acquired by reinforcement learning. This target is "fuzzy target". And it applied to the four-wheeled vehicle control, and the effectiveness of it was confirmed by the computer simulation.

1　はじめに

人間は，知識や経験を獲得し，これらを柔軟に利用することによって，環境が変化した場合でも，適切に行動を決定している．我々は，コンピュータを用いて，このような，知識や経験の獲得とその利用に基づいた，柔軟な知的制御の実現を目指している．

コンピュータを用いて知識や経験を獲得する，機械学習の一つとして「強化学習」[1]がある．この手法は，人間や動物が試行錯誤的に成功を学習する仕組みを模倣する学習法である．ある行動を選択したことによって，成功した場合には報酬が，失敗した場合には罰が与えられる．これを繰り返した結果，タスクを達成するために必要な知識が得られる．そして，過去に試みた行動の中で，報酬を得るために効果的な知識を利用する．我々はこれまで，行動として制御目標を決定する「目標設定知識」を強化学習を用いて獲得することを提案し，四輪車の知的駐車制御を実現した [2][3]．

一般的に，強化学習で獲得した知識は学習した環境に特化されており，汎用性が無い．そのため，過去に学習した環境に対し，新たに制約が付加されたことによって環境が変化した場合には，その環境に適した知識を再び獲得する必要があった．しかし，最適性を重視せず，タスクの達成だけを考えれば，過去の獲得した知識が，新たに制約が付加された環境下でも，部分的に利用できる．

本論文では，強化学習で獲得した知識や経験が，タスク達成に関する，成功の度合いを含む点に着目し，目標設定知識をファジィ集合 [4]で取り扱う．この，目標設定知識のファジィ集合が「ファジィ目標」[5]である．この目標に基づいて，制約の無い環境で過去に獲得した知識を，障害物により新たに制約が付加された環境下で柔軟に利用し，四輪自動車の移動制御を行う．この，ファジィ目標に基づいた知的制御システムの有効性を，実際の車両を想定したコンピュータシミュレーションによって評価する．

2　ファジィ目標の概要

本論文では，強化学習を用いて，行動として制御目標を決定する「目標設定知識」を獲得する．目標設定知識とは，制御対象の状態 c_n と，この状態のときに制御器に指示する制御目標（行動）a_n の，状態と行動のペアと，それに対する評価値を記述したものである．ここでの行動 a_n とは，状態 c_n における制御指令ではなく，制御器に指示する制御目標である．

この，強化学習で得られる目標設定知識には，学習の過程で，タスクが成功した場合には報酬が，失敗した場合には罰が与えられる．そのため，ある状態 c_n において，行動 a_n がタスク達成のために効果的であるほど，状態 c_n と行動 a_n のペアは，高い価値を持つ．逆に，状態 c_n において，行動 a_n がタスク達成のために効果的で無いならば，状態 c_n と行動 a_n のペアは，低い価値を持つこととなる．このように，強化学習で得られる目標設定知識は，タスク達成に関する，成功の度合いを含んでいる．そこで，本研究ではこの点に着目し，学習の過程で得られた報酬をメンバーシップ値として，目標設定知識をファジィ集合で取り扱うことで,「ファジィ目標」を獲得する．

「ファジィ目標」は，度合いを持った複数の目標値を，ファジィ集合で一つにまとめたものである (図 1)．学習の

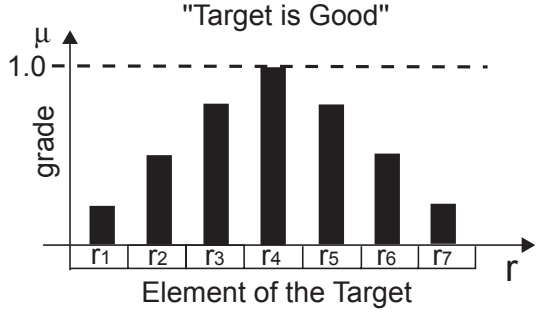

Fig.1 The Fuzzy Target

際に獲得した各目標要素に対する報酬を，ファジィ目標 “Target is Good” のメンバーシップ値とする．

制御を行う際にファジィ目標を用いる利点として，「目標の代替案」を含んでいる点が挙げられる．一般的に，強化学習で獲得した知識は，学習した環境に特化されたものであり，汎用性が無い．そのため，過去に学習した環境に対し，新たに制約が付加されることによって環境が変化した場合には，その環境に適した知識を再び獲得する必要があった．

しかし，図 1 に示すように，ファジィ目標は目標値となる複数の要素を含んでいるので，最適性を重視せず，タスクの達成だけを考えれば，新たに制約が加わった場合でも，適用可能な代替案を用いることで，適切に制御目標を設定し，柔軟に対応することが可能である．

本論文では，四輪車の移動制御において，障害物の無い，十分に広い環境下で獲得したファジィ目標を，障害物を配置し，制約が付加されることによって環境が変化した状態で利用する．そして，適用可能な代替案を用いて，環境の変化に柔軟に対応する．

3　ファジィ目標に基づく四輪車の移動制御

四輪自動車は，２入力 (ハンドルと速度の操作) で，３出力 (位置 (x, y) と車体の向き) を制御する，ノンホロノミックな系として知られている．そのため，現在状態から最終目標へ車両を移動させるためには，自車および周囲の状況に応じて，適切に途中経由地点 (サブゴール) を設定し，移動制御を行う必要がある．

本章では，四輪車の特性について詳しく触れた後，強化学習によるファジィ目標の獲得法と，ファジィ目標に基づいた知的制御システムについて述べる．

3.1　四輪車の特性

前輪操舵の四輪自動車において，十分速度が遅いとき，タイヤの滑りや遠心力の発生を無視することができる．その際，四輪自動車は，図 2 に示すような幾何学的な動きをする．左右前輪の角度 (ϕ_L,ϕ_R) は，四輪全てが旋回中心と直角に旋回半径 R で動くように，*Ackerman-Jeantaud* の操舵機構により構成されている．

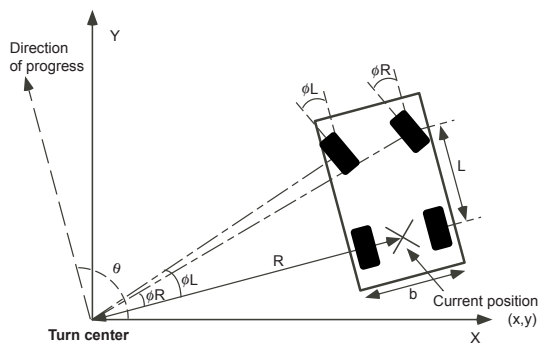

Fig.2 Kinematics constraint in nonholonomic vehicle

この自動車の動きは，次に示す拘束条件式によって記述できる:

$$\frac{dx}{dt} = v\cos\phi\cos\theta,$$
$$\frac{dy}{dt} = v\cos\phi\sin\theta,$$
$$\frac{d\theta}{dt} = \frac{v}{L}\sin\phi$$

ここで，x,y は自動車の位置，θ は x 軸と自動車の進行方向のなす角，v は車の速度，ϕ は旋回半径 R での動きを指令するハンドルの切れ角，L は車のホイールベースの長さである．

四輪自動車は上記のような特性を持つので，現在状態 (x_0, y_0, θ_0) から，任意の状態 (x, y, θ) への移動は不可能である．例えば，車両を真横に移動させるためには，切返し地点を設定する必要がある．このような理由で，現在状態から最終目標 (x_G, y_G, θ_G) へ車両を移動させるためには，自車両および周囲の状況に応じた，適切なサブゴールを設定し，これに基づいて移動制御を行う必要がある．

3.2　ファジィ目標の獲得

サブゴールを適切に設定する際に必要な「目標設定知識」を，強化学習によって獲得する．本研究では，強化学習を，制御指令の獲得のために使用するのではなく，サブゴールを設定する, 目標設定知識を獲得するために使用する．

PSP(ProfitSharingPlan)-learning [6]は，強化学習の手法の一つとして提案されている．

この学習法は，初期状態から最終目標に至る，一つのエピソードにおいて，何段階かの行動の後の成功報酬を各段階にさかのぼって分配する．PSP-learning は離散的な状態遷移に対して適用されている場合が多いが，ここでは，連続値の状態の制御である車の運転に適用し，その報酬に個々の目標達成度をファジィ評価によって行う，fPSP(fuzzyProfitSharingPlan)-learning [3]を用いる．

fPSP-learning の概要を図 3 に示す．

fPSP-learning によって得られる知識は，“IF <condition:c_n>THEN<target:a_n>” なる形式の，状態と行動のペアと，その評価値からなる規則 $S(c_n, a_n)$ であり，これをまとめたものを S-table とする．

学習時の状態 p_n（連続値）の地点 c_n（離散値）において，行動 a_n を ϵ-グリーディ手法により選択する．ϵ-グリーディ手法とは，学習の際に未知の経験の探査と知識利用のバランスを取るための簡単な方法で，全実行時間に対してある小さな割合でランダムな選択を行うものである．本研究では，10 回の試行のうち 1 回を，ルーレットによりラ

ンダムに行動選択する．

行動 a_n を選択した結果，状態 p_{n+1} に遷移する．その際の目標達成度 $\mu(p_{n+1}-a_n)$ を図 4 に示すようなメンバーシップ関数を用いて，

$$\mu(p_{n+1}-a_n)=\mu_d x(\Delta x_{n+1})\wedge\mu_d y(\Delta y_{n+1})\wedge\mu_d\theta(\Delta\theta_{n+1})$$

のようにファジィ評価によって求める．

そして，状態 p_{n+1} の地点 c_{n+1} において，同様に次の行動 a_{n+1} を選択する．

何段階かのステップの後，最終目標に到達した際，獲得した報酬 R を各ステップに対して分配し，S-table 値を更新する．その際，報酬 R を獲得するために使用したステップ数を N，割引率を γ，学習率を α とすると，n 番目のステップにおける S-table 値 $S(c_n,a_n)$ は，

$$S(c_n,a_n)=(1-\alpha)S(c_n,a_n)+\alpha\mu(p_{n+1}-a_n)R*\gamma^{(N-n)}$$

のように更新される．

実行した結果に得られる，エピソードで用いた行動 a_n 全体に対する報酬 R は，最終目標に達することができた場合，報酬は制限時間 (250 秒) と所要時間の差で与える．例えば，最終的な目標に 80 秒で達したならば，報酬 R は 170 となる．制限時間内に到達できない時や，障害物などで動きが取れない状況に陥った場合，罰を -100 とし，そのエピソードで用いた全ての行動 a_n に対する評価を下げる．

学習後，得られた評価値をメンバーシップ値としたファジィ集合を，ファジィ目標とする．

なお，学習時には，位置と車体方向で与えられる目標 (x_T,y_T,θ_T) に達するように，車両を移動するため，図 5 のようなカスケードファジイ制御器[2]を用いた．目標方向が y 軸となるよう座標系を変換し，目標 X_T との x 軸方向の偏差 X_e から，現時点の目標角度 θ_T をファジィ推論し，さらに車体角度の誤差 θ_e からハンドルの操作角度 ϕ をファジィ推論により求めている．

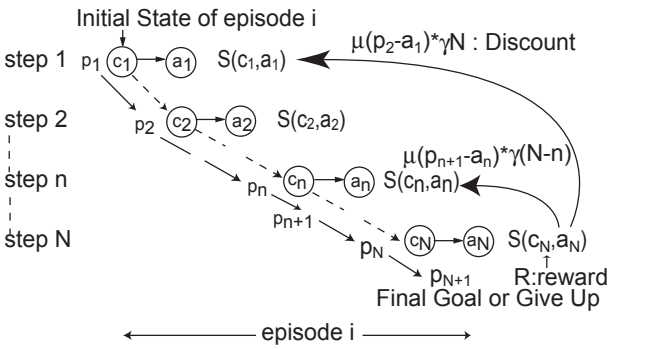

Fig.3 PSP-learning deitributes reward of penalty to the previous fired rules

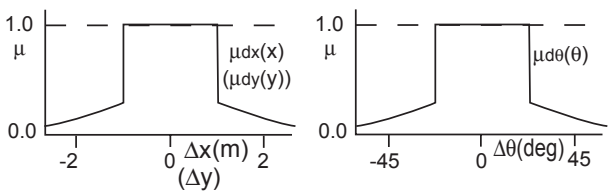

Fig.4 Fuzzy set of position and angle error

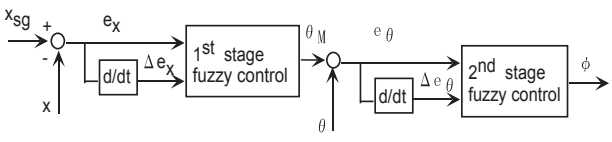

Fig.5 Cascade fuzzy controller

3.3 制御システムの概要

獲得したファジィ目標に基づいて四輪車の移動制御を行う，階層型知的制御システムの概要を図 6 に示す．

このシステムは 3 つのブロックに分かれている．状況監視部では，最終目標およびサブゴールへの到達と，障害物への接触を判断する．ファジィ目標設定部では，車両の現在位置におけるファジィ目標を，強化学習によって得られた S-table から取得する．自動運転部では，ファジィ目標を利用して適切な制御指令を算出し，障害物に接触しないよう車両を制御する．

3.3.1 状況監視部

状況監視部は，障害物への接触，最終目標およびサブゴールへの到達について評価し，新しい目標の必要性を判断する．何らかの理由で現在のサブゴールへの到達が困難となった場合は，現在のサブゴールをリセットし，ファジィ目標設定部に対し目標設定指令を出力する．

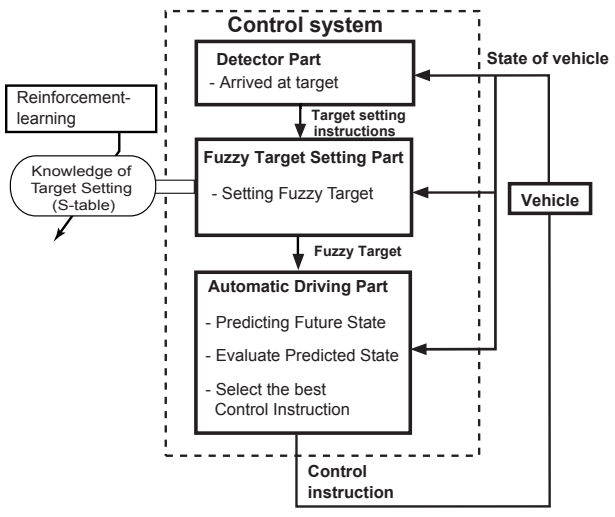

Fig.6 Outline of the control system

3.3.2　ファジィ目標設定部

目標設定指令を受け，ファジィ目標設定部は，強化学習によって得られた S-table から，ファジィ目標を設定する．

ここで重要なことは，強化学習によって得られた S-table は過去に学習した環境に特化されている，という点である．そのため，自動運転部において，適切にファジィ目標を利用する必要がある．

3.3.3　自動運転部

自動運転部における制御指令の算出過程を図 7 に示す．ファジィ目標中の個々の要素を，サブゴールとして仮定し，カスケードファジィ制御器によって操作指令候補を算出する．算出された操作指令候補を用いて車両を仮に動かし，車両の将来状態を予見する．予見した将来状態と，サブゴールの評価値 (メンバーシップ値)，サブゴールとの距離偏差，角度偏差，および障害物との距離を，総合的にファジィ評価し，操作指令候補の評価値を求める．これを，ファジィ目標の全ての要素に対して行い，最も評価値の高かった操作指令候補を選択し，これを制御指令として実車両に与える．

4　コンピュータシミュレーションによる評価

構築した知的制御システムの有効性を，実際の車両を想定したコンピュータシミュレーションにより検証する．

4.1　シミュレーションの内容

任意の位置 (x, y, θ) から最終目標 $(20m, 20m, 0.25\pi)$ まで，四輪車の移動制御を行う．今回は，四輪車を，位置 $(0m, 0m, 0.25\pi)$ および $(6m, 16m, 0.0\pi)$ から，最終目標まで移動させる．

シミュレーションに用いた四輪車の諸元は，車幅：1.7m，ホイールベース：2.6m，最小旋回半径：6.0m であり，移動速度は前後進ともに 0.4m/s である．

学習を行った環境を図 8 に示す．障害物の無い広い環境を離散化し，2m 毎の格子点を配置した．各格子点を中心とし，車両の向き θ を，0，0.25π，0.5π，0.75π，π，1.25π，1.5π，1.75π，の 8 方位に区分しラベル化している．図 8 に示す環境では，121 個の格子点が置かれ，$121\times8 = 968$ 個のラベルがある．そのため，状態 $c_n = (x_T, y_T, \theta_T)$ は 968 通り，行動選択のための目標 $a_n = (x_T, y_T, \theta_T)$ は 968 分割されている．位置 $(0m, 0m, 0.25\pi)$ は c_1，$(6m, 16m, 0.0\pi)$ は c_{328} というようにラベル化した．

上記の環境下で，fPSP-learning により，ファジィ目標を獲得した．獲得したファジィ目標の一例として，車両の位置 $c_1 = (0m, 0m, 0.25\pi)$ における，ファジィ目標 $\widetilde{a_1}$ の要素のうち，メンバーシップ値が 0.2 以上のものを次に示す．また，ファジィ目標 $\widetilde{a_1}$ の各要素の位置を図 9 に示す．

Target position	Membersip value
(4m, 2m, 0.25π)	0.711
(4m, 10m, 0.5π)	0.669
(6m, 2m, 0.25π)	0.528
(6m, 6m, 0.0π)	0.702
(6m, 10m, 0.0π)	0.695
(6m, 12m, 0.5π)	0.664
(8m, 12m, 0.25π)	0.701
(10m, 0m, 0.0π)	0.608
(10m, 4m, 0.25π)	0.690
(12m, 16m, 0.0π)	0.467
(14m, 6m, 0.25π)	0.673
(14m, 16m, 0.25π)	0.706
(16m, 2m, 0.25π)	0.609
(16m, 4m, 0.25π)	0.634
(16m, 10m, 0.5π)	0.623
(16m, 14m, 0.5π)	0.619
(18m, 14m, 0.0π)	0.213
(20m, 20m, 0.25π)	0.880

このファジィ目標 $\widetilde{a_1}$ の中の，要素 $(20m, 20m, 0.25\pi)$，

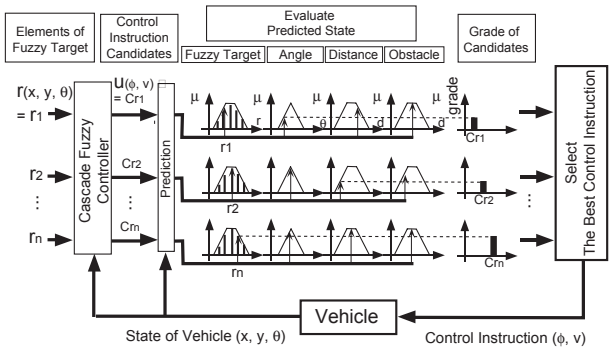

Fig.7 Detail of the Automatic Driving Part

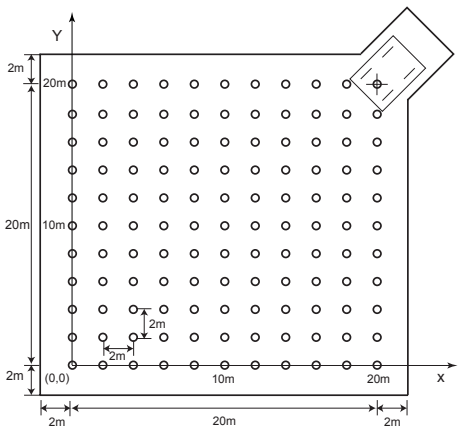

Fig.8 Demension of the field

すなわち最終目標が持つメンバーシップ値は 0.880 で最も高い．これは，学習を行った，障害物の無い広い環境下において，状態 c_1 の時にこの要素を利用し，制御指令を算出することが，短時間で正確にタスクを達成のために最も有効であることを示している．しかし，障害物を配置したことによって環境が変化した場合においても，この要素を選択することが常に有効であるとは限らない．

車両が位置している，状態 c_n におけるファジィ目標 $\widetilde{a_n}$ に基づいて制御指令を決定し，最終目標 $(20m, 20m, 0.25\pi)$ まで移動させる．シミュレーションを行う初期位置は，(A) $c_1 = (0m, 0m, 0.25\pi)$，および (B) $c_{328} = (6m, 16m, 0.0\pi)$ である。(A) および (B) において，(i) 障害物が無い場合，(ii) 新たに障害物を配置し制約を付加した場合 の２通りのシミュレーションを行う．その際，環境の変化（障害物の有無）に関わらず，同一のファジィ目標を用いる．

ファジィ目標 $\widetilde{a_{328}}$ のうち，(B) で使用する，メンバーシップ値が 0.1 以上のものを次に示す．

Target position	Membersip value
(2m, 0m, 0.5π)	0.128
(8m, 16m, 0.0π)	0.486
(12m, 18m, 0.0π)	0.739
(20m, 20m, 0.25π)	0.999

4.2　シミュレーション結果

(A) 初期位置 $c_1 = (0m, 0m, 0.25\pi)$ の場合（図 10）

図 10 (i) のように，障害物が無い場合，最終目標まで直進し、到達することができた．このときの経過時間は 70 秒であった。

(ii) のように障害物を初期位置の前方に配置し，可動領域を制約した場合，直接最終目標に向うことができない．

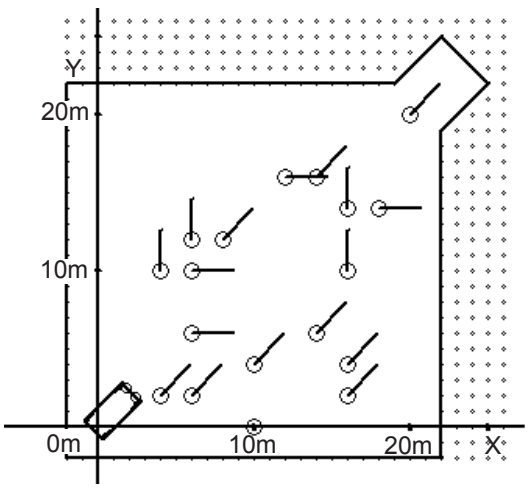

Fig.9 Fuzzy target at the vehicle state (0m,0m,0.25π)

この場合，ファジィ目標中の利用可能な代替案を用いて，ファジィ目標中の要素 $(4m, 10m, 0.5\pi)$ をサブゴールとし，これに向かうよう制御指令を算出した．この地点に到達した後，今度は，最終目標 $(20m, 20m, 0.25\pi)$ へ向かうよう，制御指令を算出した．この結果，障害物を迂回するように移動し，最終目標に到達することができた．最終目標到達までの経過時間は 82 秒であった．

(B) 初期位置 $c_{328} = (6m, 16m, 0.0\pi)$ の場合（図 11）

図 11 (i) のように，障害物が無い場合，最終目標を目指し，到達することができた．最終目標到達までの経過時間は 37 秒であった．

(ii) のように障害物を初期位置の前方に配置した場合，前方に移動することができない．また、四輪車のノンホロノミックな特性により，真横に平行移動して障害物を回避することもできない．そのため，ファジィ目標中の利用

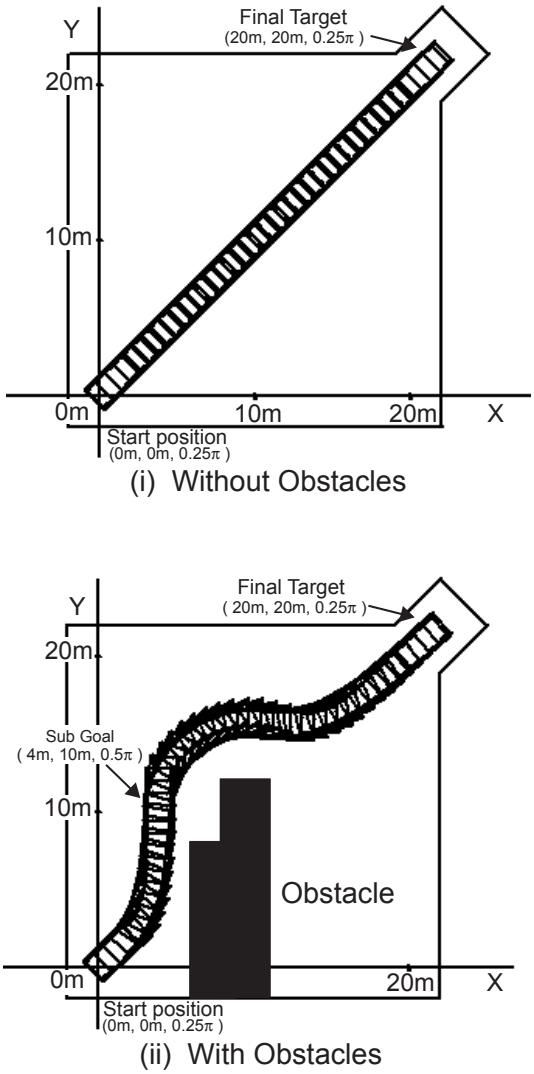

Fig.10 Case (A) : The trajectory from (0m,0m,0.25π) to (20m,20m,0.25π)

可能な代替案を用いて，適切にサブゴールを設定し、切返しを行うことによって，障害物を回避している．まず，ファジィ目標 $\widetilde{a_{328}}$ の中から，要素 $(2m, 0m, 0.5\pi)$ をサブゴールとして選択し，後退で向かうよう制御指令を算出した．次に，サブゴール $(2m, 0m, 0.5\pi)$ に向けて十分に後退したところ，障害物を避け，最終目標 $(20m, 20m, 0.25\pi)$ へ直接移動可能な状態となった．そして，目標を最終目標 $(20m, 20m, 0.25\pi)$ に切り替え，前進することにより，最終目標に到達することができた．最終目標到達までの経過時間は 113 秒であった．

(A) および (B) のように，過去に獲得したファジィ目標に基づいて車両の移動制御を行うことで，障害物の配置によって変化した環境に適応し，最終目標に到達することができた．これにより，構築した，ファジィ目標に基づいた移動制御システムの有効性を，シミュレーションによって確認した．

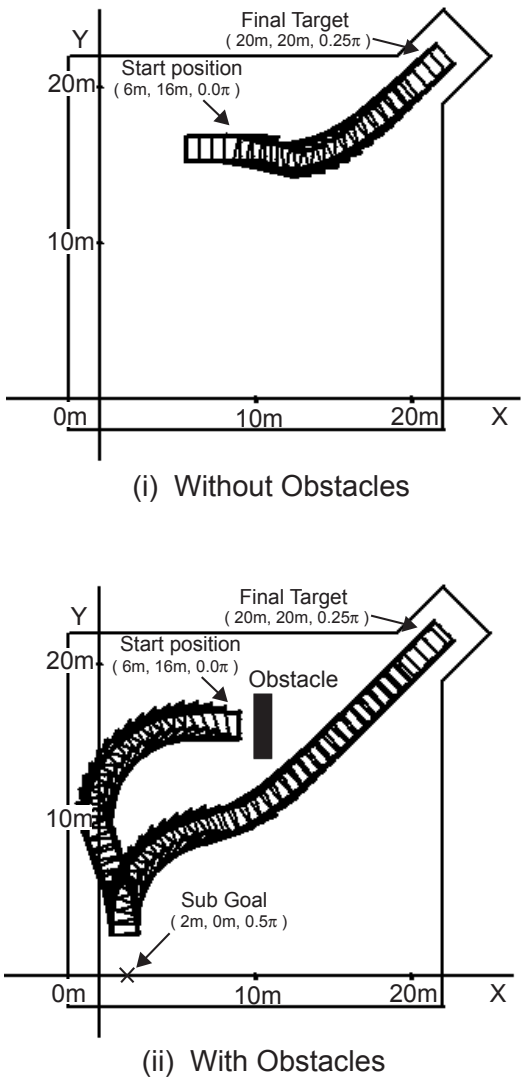

Fig.11 Case (B) : The trajectory from (6m,16m,0.0π) to (20m,20m,0.25π)

5 おわりに

本論文では，強化学習で試行錯誤的に得られる目標設定知識が，タスク達成に関する，成功の度合いを含む点に着目し，目標設定知識をファジィ集合で取り扱った「ファジィ目標」に基づいて，四輪車の移動制御システムを構築した．このシステムは，獲得したファジィ目標によって，新たに制約を付加した環境下においても，環境の変化に柔軟に対応でき，タスクを達成することが可能であった．この，強化学習で獲得した，ファジィ目標に基づいた知的制御システムの有効性を，実際の車両を想定したコンピュータシミュレーションによって確認した．

参考文献

[1] Richard S. Sutton and Andrew G. Barto: REINFORCEMENT LEARNING: An Introduction, A Bradford Book (1998).

[2] Y. Suryana and S. Yasunobu: Hierarchical Intelligent Control by PSP-learning and Cascade Fuzzy Control for Parking Vehicle in Narrow Area, The Transactions of The Institute of Electrical Engineers of Japan, Vol. 122-C, No. 2, 315/316 (2002).

[3] S. Yasunobu: Fuzzy-Reinforcment-learning Design of Intelligent Controller for Nonholonomic Vehicle, Proc. of The 12^{th} FAN Intelligent Systems Symposium, 105/108 (2002).

[4] L. A. Zadeh: Fuzzy sets, Information and Control, Vol. 8, 338/353 (1965).

[5] S. Yasunobu and T. Matsubara: Fuzzy Target Acquired by Reinforcement Learning for Parking Control, Proc. of SICE Annual Conference 2003 in Fukui (SICE2003), 1303/1308 (2003).

[6] T. Horiuchi, A. Fujino, O. Katai and T. Sawaragi: Q-PSP learning: An Exploration-Oriented Q learning and Its Applications, The Society of Instument and Control Enggineers, Vol. 35, No. 5, 645/653 (1999).