SISTEMA DE ORIENTAÇÃO PARA ROBÔS MÓVEIS UTILIZANDO APRENDIZAGEM POR REFORÇO.

Rafael Nunes de ALMEIDA PRADO(1); José Henrique d'SOUZA(2)
(1) Centro Federal de Educação Tecnológica do RN – CEFET – RN Departamento Acadêmico de Tecnologia Industrial – DATIN Núcleo de Desenvolvimento em Mecatrônica – NUDEM
Av. Senador Salgado Filho, 1559, Tirol, fone/fax: (84)3208-8195 Natal-RN, CEP 59015-000
e-mail: rnaprado@yahoo.com.br
(2) Centro Federal de Educação Tecnológica do RN – CEFET – RN, e-mail: ricky@cefetrn.br

RESUMO

O presente trabalho apresenta o desenvolvimento de um sistema de orientação inteligente empregado em robôs autônomos móveis. O projeto foi baseado no uso de técnicas de aprendizagem por reforço (AR) para orientar-se em ambientes e situações desconhecidas. A técnica de AR baseia na distribuição de recompensas ao robô e estas podem ser positivas e/ou negativas (decisão certa implica em reforço positivo e decisão errada em reforço negativo). O sistema consiste em uma modelagem cinemática restritiva (restrições não-holomônicas) e de comportamento dinâmico do robô através de um modelo SISO iterativo de controle. Os resultados preliminares atestam a eficácia da estratégia de controle para auxiliar o robô a aprender com seus erros. O agente implementado se mostrará capaz de aprender comportamentos coerentes com algumas expectativas de desempenho a partir dos resultados de suas ações. A modelagem em ambiente virtual (simuladores) está sendo realizada para confrontação dos resultados com os anteriormente obtidos e a partir destes poder iniciar a construção de um protótipo em escala real para a técnica possa ser testada em situação real de trabalho.

Palavras-chave: Aprendizagem por reforço, sistemas inteligentes, robô autônomo, simulação.

1. INTRODUÇÃO

A inteligência artificial vem fascinando o mundo científico há muitas décadas em diversos campos da ciência e da ficção. Dentre os estudos envolvidos, estão os sistemas autônomos, providos de parcial, ou total capacidade de realizar tarefas complexas e exatas, podendo obter um aprimoramento ao longo do tempo. Este setor vem sendo intensamente explorado, devido ao aumento do grau de complexidade dos processos realizados hoje, à exigência de velocidade em processamento, risco das operações, e a necessidade de sistemas cada vez mais independentes, adaptáveis e inteligentes.

Neste contexto, o desenvolvimento de robôs móveis tem lugar de destaque. Isto se justifica pela grande quantidade de atividades abrangidas pela aplicabilidade destes robôs. As tarefas feitas por eles variam em diversos graus de sofisticação e robustez. Hoje se podem realizar procedimentos que há alguns anos atrás não eram possíveis.

Tais robôs devem aceitar ordens para a execução de tarefas com um elevado grau de dificuldade e as cumprirão sem a intervenção humana. Ao colocar o robô como substituto do homem, deve-se dotá-lo de capacidade para tomada de decisões, a fim de que ele trabalhe conjuntamente com as demais máquinas. Esta premissa tem sido uma das principais motivações para a pesquisa de veículos móveis autônomos ou AGVs, foco deste trabalho. Vemos que é possível fabricar robôs que servem quase para qualquer operação, seja de limpeza, operações cirúrgicas a distancia, cortar a grama, fazer o chá, etc. Existem micromáquinas (col ocados num relógio de pulso, por exemplo) capazes de obter dados de temperatura, pressão, pulso, etc. robôs móveis poderão monitorar a poluição no meio ambiente com maior eficácia (NEHMZOW, 2000).

A habilidade de decisão depende da inteligência, que depende do aprendizado. A inteligência não pode existir sem a capacidade de aprender ou de adquirir novos conhecimentos. Nos sistemas robóticos não é diferente, e a necessidade de aplicação da inteligência artificial em robótica não é nova. Ao tornar o robô mais complexo, adicionando a ele novos sensores e atuadores, estamos também acrescentando mais complexidade ao trabalho de programação de ações do robô, além de dificultar a tarefa de calibração dos sensores e coordenação dos movimentos. O desenvolvimento de algoritmos de aprendizado permite que o robô calibre seu comportamento e desempenhe a sua tarefa de forma mais confiável e adaptável.

Este trabalho tem por objetivo analisar o desempenho de um sistema para orientação baseada na aprendizagem por reforço (AR). A aprendizagem por reforço (SUTTON & BARTO, 1998), consiste, basicamente, em fazer um agente escolher suas ações se baseando apenas na interação com o ambiente. Diferentemente da aprendizagem supervisionada, na qual existe um professor que diz ao agente qual deveria ter sido a ação correta para cada estado, na aprendizagem por reforço, existe apenas um crítico, que indica o caminho correto, mas não diz exatamente a resposta correta. Esse tipo de aprendizagem é inspirado na aprendizagem infantil humana. Uma criança costuma realizar ações aleatórias, e, de acordo com as respostas de seus pais (elogio ou reclamação), ela aprende quais destas ações são boas e quais são ruins (RIBEIRO, 1999).

2. FUNDAMENTAÇÃO TEÓRICA

Nesta seção será apresentada uma revisão bibliográfica sobre os conceitos em robótica móvel, sistemas autônomos e sistemas inteligentes.

2.1. Modelagem de robôs móveis com acionamento diferencial

Acionamento diferencial (ou direção diferencial) é o mecanismo de direção mais simples, consiste de duas rodas em um eixo comum, em que cada roda é controlada independentemente. Utiliza uma roda adicional (caster) para balanço, e é sensível a velocidade relativa das duas rodas (pequeno erro resulta em diferentes trajetórias, não apenas velocidade).

Será mostrado o modelo cinemático do robô usando acionamento diferencial, que representa as características de movimento e as restrições destes, o modelo cinemático também é um modelo dinâmico, pois o estado do robô, definido por este modelo, varia com as excitações de entrada e depende do estado no instante anterior, porém este não inclui as forças dinâmicas que atuam sobre o robô, daí a separação entre modelos cinemático e dinâmico.

2.1.1. Não-holonomia

Os robôs móveis com rodas e os com pernas, juntamente com os satélites, pertencem a uma classe de sistema mecânico denominada de sistemas não-holonômicos, que se caracterizam por ter restrições cinemáticas. Os algoritmos de controle e planejamento de movimento de tais sistemas requerem, portanto, uma classe diferente de procedimentos que aqueles empregados em manipuladores mecânicos estacionários, (sistema holonômico).

No caso de robôs móveis com acionamento diferencial a restrição é imposta pela impossibilidade do robô se movimentar em todas as direções, devido ao sistema não possuir atuadores que permitam tais movimentos, bem como pela condição de não-deslize (considera-se que não há derrapagem). Um robô com um sistema de locomoção por rodas e com a mesma configuração de um carro, é um clássico exemplo de robô móvel não-holonômico. Uma restrição não-holonômica impede que o robô execute movimentos normais à superfície do corpo de suas rodas, quando não há deslizamento(ALSINA, 2002).

2.1.2. Modelo Cinemático

A configuração do robô é representada por sua posição no espaço cartesiano (x e y – posição do centro do robô em relação um referencial fixo no espaço de trabalho), e pela sua orientação θ (ângulo entre o vetor de orientação do robô e o eixo x do referencial fixo no espaço de trabalho). A Figura 1 mostra a representação do robô em questão (ALSINA, 2002).

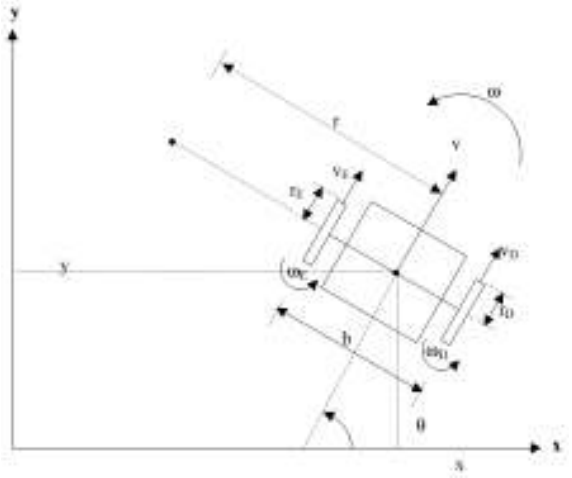

Figura 1 - Representação esquemática das variáveis cinemáticas do robô.

$(x, y) =$ Posição do referencial fixo no robô em relação ao referencial fixo no espaço de trabalho.

$\theta =$ Ângulo de orientação do robô em relação ao referencial fixo no espaço de trabalho.

$b =$ Comprimento do eixo.

$r =$ Raio de giro do robô.

$r_d\,(r_e) =$ Raio da roda direita (esquerda)

$\omega =$ Velocidade angular do robô.

$\omega_d\,(\omega_e) =$ Velocidade angular da roda direita (esquerda).

$v =$ Velocidade linear do robô

$v_d\,(v_e) =$ Velocidade linear da borda da roda direita (esquerda).

As relações entre as velocidades lineares e angulares são:

$$v = \omega\, r \tag{1}$$

$$v_d = \omega_d r_d \tag{2a}$$

$$v_e = \omega_e r_e \tag{2b}$$

Para deslocamentos incrementais em um intervalo de tempo dt como mostra a Figura 2 abaixo:

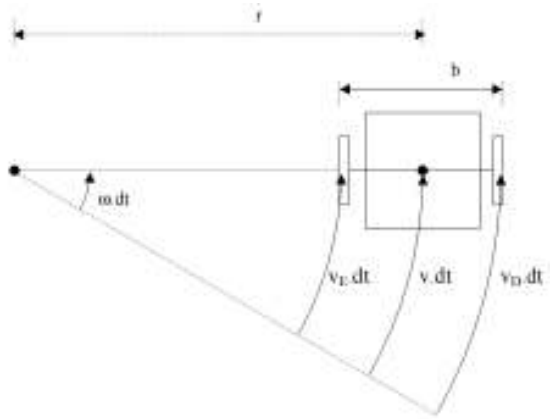

Figura 2 - movimento infinitesimal em um intervalo *dt.*

$$v_d dt = \omega(r + \frac{b}{2})dt \tag{3a}$$

$$v_e dt = \omega(r - \frac{b}{2})dt \tag{3b}$$

A partir da equação (3) e utilizando as expressões (1) e (2) temos:

$$v = \omega_d \frac{r_d}{2} + \omega_e \frac{r_e}{2} \tag{4a}$$

$$\omega = \omega_d \frac{r_d}{b} - \omega_e \frac{r_e}{b} \tag{4b}$$

Que podem ser representadas na forma matricial como:

$$\mathrm{V} = {}^{v}\mathrm{T}_{\omega}.\mathrm{W} \longrightarrow \mathrm{W} = ({}^{v}\mathrm{T}_{\omega})^{-1}.\mathrm{V} \longrightarrow \mathrm{W} = {}^{\omega}\mathrm{T}_{v}.\mathrm{V} \quad \text{onde} \quad {}^{\omega}\mathrm{T}_{v} = ({}^{v}\mathrm{T}_{\omega})^{-1} \tag{5}$$

onde,

$$\mathrm{V} = \begin{bmatrix} v \\ \omega \end{bmatrix} \tag{6a}$$

$${}^{v}\mathrm{T}_{\omega} = \begin{bmatrix} \frac{r_d}{2} & \frac{r_e}{2} \\ \frac{r_d}{b} & -\frac{r_e}{b} \end{bmatrix} \tag{6b}$$

$$\mathrm{W} = \begin{bmatrix} \omega_d \\ \omega_e \end{bmatrix} \tag{6c}$$

O vetor V representa as velocidades em referencial de eixos principais e W é o vetor de velocidades em espaço de atuadores.

A partir dessas relações, encontram-se as velocidades das rodas para que o robô possa mover-se com raio de giro r.

$$\frac{\omega_e}{\omega_d} = \frac{(r - b/2)r_d}{(r + b/2)r_e} \tag{7}$$

3. METODOLOGIA

Trataremos rapidamente de um embasamento, que propiciará a obtenção da lógica na qual foi baseado o algoritmo do sistema, através do entendimento dos processos de decisão Markovianos, da aprendizagem por reforço, da técnica de aprendizagem aplicada no problema (Q-Learning) e o planejamento do comportamento.

3.1. Decisões de Markov

Um processo de Markov é uma seqüência de estados, com propriedade de que qualquer predição de valor de estado futuro dependera apenas do estado e ação atuais e não da seqüência de estados passados. Um ambiente satisfaz a propriedade de Markov se o seu estado resume o passado de forma compacta sem perder a habilidade de prever o futuro, ou seja, pode-se dizer qual será o próximo estado e a próxima recompensa do estado e ações atuais (SUTTON & BARTO, 1998).

Um processo de aprendizagem por reforço que satisfaz a propriedade de Markov é chamado de processo decisório de Markov (MDP – Markov Decision Process). Se o espaço de estados e ações for finito, então ele é chamado de processo decisório de Markov finito, base para teoria de aprendizagem por reforço.

3.2. Aprendizagem por reforço

Formalmente, no problema de aprendizagem por reforço temos um agente, que atua em um ambiente. O agente percebe um conjunto discreto 'S' de estados, e pode realizar um conjunto discreto 'A' de ações. A cada instante de tempo *t*, o agente pode detectar seu estado atual "*s"* e, de acordo com esse estado, escolher uma ação "a" ser executada, que o levará para um outro estado *s*'. Para cada par estado/ação, (*s,a*), há um sinal de reforço, $r(s,a) \rightarrow \Re$, que é dado ao agente quando ele executa a ação *a* no estado *s*. O relacionamento do agente com ambiente é ilustrado na Figura 3.

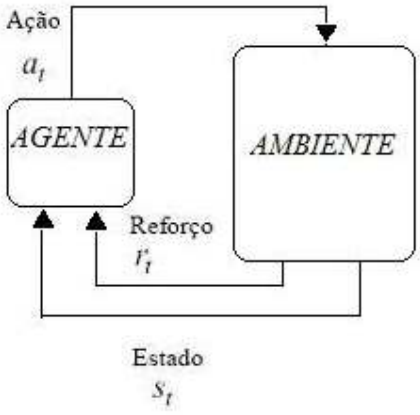

Figura 3 – Relacionamento na aprendizagem.

O sinal de reforço é a base do aprendizado do agente. O reforço deve indicar o objetivo a ser alcançado. Por exemplo, em um jogo de damas o reforço pode ser dado ao agente apenas ao final do jogo, sendo positivo quando o agente ganhar ou negativo quando perde ou empata. Com isso, o reforço está mostrando ao agente que seu objetivo é ganhar o jogo, e não perder ou empatar (JÚNIOR, 2006).

O problema de aprendizagem por reforço consiste em escolher uma política de ações que maximize o total de recompensas recebidas pelo agente. Uma política de ações corresponde a uma função Π(s)→*a*, que diz, para cada estado, qual deve ser a ação realizada pelo agente. Um agente pode seguir várias políticas de ações, mas o objetivo da aprendizagem é calcular a política que maximize a soma das recompensas futuras, isto é, o total de recompensas recebidas após a adoção dessa política. (ANDRADE, 2004).

3.2.1. Q-learning

Uma das técnicas mais utilizados em problemas de aprendizagem por reforço é o algoritmo Q-Learning (WATKINS, 1989). Tal algoritmo é baseado nos conceitos do método de diferenças temporais, utiliza os princípios de acúmulo de reforço e ganho mostrado na seção anterior e tem sua convergência para valores excelentes de $Q(Q^*(s,a))$ independente da política que está sendo utilizada. A expressão de atualização do valor de Q do algoritmo Q-Learning é a seguinte:

$$Q(s,a) = Q(s,a) + \alpha.[r_{t+1} + \gamma \max_{a'} Q(s',a') - Q(s,a)] \quad (8)$$

onde r_{t+1} é o retorno associado à transição do estado s para o estado s', α é a taxa de aprendizagem γ é o fator de desconto, com $0 \leq \gamma \leq 1$. A função de valor do estado atual ($Q(s,a)$) é atualizada a partir do seu valor atual, do reforço imediato (r_{t+1}) e da diferença entre a máxima função de valor no estado seguinte $(\max_{a'} Q(s',a'))$ e o valor da função de valor do estado atual.

Na equação do Q-Learning, uma questão importante é a analise do cálculo do termo $(\max_{a'} Q(s',a'))$. No caso geral, esse cálculo pode ser visualizado através do diagrama mostrado na Figura 4 abaixo.

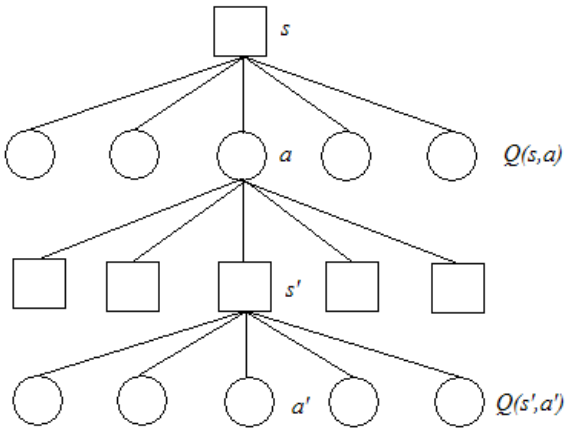

Figura 4 - Diagrama da equação Q-Learning.

Na Figura 4, os estados são os quadriláteros. As ações são os círculos e ação a está nomeada. Quando passados s e a, tem-se o estado s'. O valor do termo $(\max_{a'} Q(s',a'))$ é então escolhido entre os valores de $Q(s',a')$ de todas as ações possíveis de serem tomadas a partir de *s*. Observe-se que não existe então a necessidade de se conhecer qual a ação deverá ser tomada em *s*, mas sim quais todas as possíveis ações permitidas para os estados possíveis, filtrando todas as possibilidades qualificando-as (YANG & GU, 2004).

Uma característica do Q-Learning é que a função de valor Q^* aprendida aproxima-se diretamente da função de valor Q desejado como satisfatório, sem depender da política que está sendo utilizada. Este fato simplifica a construção do algoritmo. A política ainda mantém algum efeito ao determinar qual dos pares estado-ação deve-se visitar e atualizar. A convergência exige que todos os pares estado-ação sejam visitados. Logo, a política Π a ser utilizada para a determinação do Q^* pode ser uma política robusta e arriscada para tentar uma convergência mais rápida. O algoritmo Q-Learning tradicional é apresentado abaixo:

Inicializar uma ação de forma arbitrária, gerando Q(*s,a*) qualquer ;

Repetir (para cada novo estado)

Inicializar s;

Repetir (para cada nova ação)

Escolher *a* para *s* usando uma política *p* (definida no projeto);

Tomar a ação *a*, receber e analisar *r, s'* ;

Obter Q(*s*,*a*);

Atualizar *s*;

Até *s* ser o estado final desejado;

Fim

O Q-Learning foi o primeiro método de aprendizagem por reforço a possuir provas de convergência [3]. É uma das mais utilizadas por ser uma técnica muito simples que calcula diretamente as ações sem avaliações intermediárias e sem uso de modelo. Em (WATKINS, 1989) mostra-se uma avaliação que se cada estado-ação for visitado um número infinito de vezes e com um valor de a adequado, a função de valor Q^* irá convergir com probabilidade 1 para Q. A convergência do algoritmo Q-Learning não depende da política de exploração utilizada. O agente está livre para explorar suas ações a qualquer momento. Não existem requisitos para a execução de ações estimadas como as melhores, mas a busca de ações que maximizem o retorno é necessária durante o aprendizado (BAKKER, ZHUMATIY & GRUENER, 2006).

3.3. Planejamento do comportamento

O comportamento do robô foi baseado nas teorias anteriormente comentadas com o intuito de alcançar um ponto de luz em um ambiente qualquer. Partindo do pressuposto, que na aprendizagem por reforço não é preciso saber qual a seqüência de passos necessária para alcançar algum objetivo, a elaboração de um planejamento de comportamento (ações) torna-se extremamente fácil. A dificuldade está em definir-se como definir os estados a serem considerados, determinar as ações para cada estado, calcular um reforço para cada par de ação-estado e implementar essas situações para o modelo usado.

No modelo utilizado, serão considerados seis sensores de luminosidade distribuídos nos cento e oitenta graus frontais do robô. No caso em que se deseja achar um ponto mais luminoso possível em um ambiente, foi criado um modelo com valores e pesos para a resposta de cada sensor. É calculada uma média ponderada dos valores existentes em sensores ativados pela sensibilidade a luz, onde cada lado do robô, direito ou esquerdo, terão valores positivos de um lado e negativo de outro, de modo a indicar a orientação que está a luz e o quanto ela está deslocada do eixo principal (melhor trajetória para se encontrar um ponto mais luminoso rapidamente). Desse modo resta apenas definir velocidade de cruzeiro, velocidades das curvas e o tamanho de cada passo (ação) e castigo, além de verificar seus estados e erros continuamente e ininterruptamente.

4. ALGORITMO

4.1. Algoritmo

Basicamente o algoritmo desenvolvido para este sistema em questão, é muito semelhante ao algoritmo da técnica Q-Learning, porém com algumas alterações e adaptações, para comportar o modelo utilizado no trabalho e as anomalias de comportamento observadas em simulação. Ele também está de forma mais esquemática e menos procedural, para que se tenha um entendimento geral e rápido.

Início do programa

Laço sem fim para cada passo (os robôs autônomos não morrem ou param de tentar realizar suas tarefas)

Leitura de estado (lêem-se todos os sensores).

Calcula-se a média ponderada dos sensores.

Escolhe-se uma ação aleatoriamente para esse estado (alteração na velocidade do robô).

Calcula-se a punição através da equação do algoritmo Q-Learning.

Verifica-se se o número de ações com reforço positivo são maiores que as com reforço ruim, se for maior o número de reforços bons, comece a escolher apenas ações com melhor reforço, senão, escolha aleatoriamente as ações para os estados.

Final do laço quando o robô obtiver condição ótima se não lhe for determinado outro objetivo (geralmente escolhe-se outros objetivos e o robô dificilmente parará, sua característica de sempre trabalhar e nunca parar).

Final do programa.

5. CONSIDERAÇÕES FINAIS

O trabalho proposto foi fundamentado em diversos artigos, apostilas e outros trabalhos científicos com o intuito de elaborar um sistema de orientação para robôs móveis utilizando aprendizagem por reforço, onde foi elaborado algoritmos de planejamento de trajetória, aprendizagem por reforço (Q-Learning) e cálculos dos reforços. Observou-se que o problema da aprendizagem por reforço é de simples aplicação e entendimento, notando-se também a sua economia computacional para poder-se, em uma etapa posterior deste trabalho, simular e embarcar em uma plataforma móvel real desenvolvida em laboratório. O trabalho pretende propor uma maior aplicação de técnicas de inteligência, principalmente de aprendizagem por reforço, e disseminar com segurança as propostas de sistemas cada vez mais especialistas, mais adaptativos e com uma segurança que dê margem para aplicações reais em ambientes industriais com total segurança. Ao tentar aplicar o sistema proposto, propõe-se fazer a implementação da AR em outras partes do sistema, além do planejamento de ações, como nos erros de controle dos atuadores, cálculo de ruídos em espaço de atuadores e até uma mudança geral na estratégia das ações, considerando grupos de ações a serem selecionados na AR juntamente com cada ação dentro desses grupos a ser selecionado para análise com AR.

REFERÊNCIAS

ALSINA, P. J. Sistemas Robóticos Autonomos. Publicação interna UFRN, DCA, 2002.

NEHMZOW, U. Mobile Robotics: A Practical Introduction. Springer, Verlag, 2000.

SUTTON, R. & BARTO, A. Reinforcement Learning: an introduction. MIT Press 1998.

WATKINS, C. J. C. H., Learning from Delayed Rewards, Phd thesis, University of Cambridge, 1989.

ANDRADE, G. D. Aprendizagem por Reforço e Adaptação ao Usuário em Jogos Eletrônicos, Recife, 2004

YANG, E. & GU, D. Multiagent Reinforcement Learning for Multi-Robot Systems: A Survey, Department of Computer Science, University of Essex Wivenhoe Park, Colchester, Essex, United Kingdom, 2004, CO4 3SQ.

JÚNIOR, L. A. C. & BIANCHI, R. A. C. Aprendizado por Reforço Acelerado por Heurística para um Sistema Multi-Agentes, 2006.

BAKKER, B. & ZHUMATIY, V. & GRUENER, G. Quasi-Online Reinforcement Learning for Robots Informatics Institute, University of Amsterdam, the Netherlands, 2006.

RIBEIRO, C. H. C. Aprendizado por reforço. V Escola de Redes Neurais: Conselho Nacional de Redes Neurais, ITA, 1999, pp. c028-c072.