CONTROLE DE MOVIMENTAÇÃO DE ROBÔ HUMANÓIDE COMANDADO POR VOZ E COM PROCESSAMENTO BASEADO EM REDE NEURAL ARTIFICIAL

Nícolas Carnizello Accarini [1]; Wânderson de Oliveira Assis [2]; Alessandra Dutra Coelho [2]

[1] Aluno de Iniciação Científica da Escola de Engenharia Mauá (EEM-CEUN-IMT);
[2] Professor(a) da Escola de Engenharia Mauá (EEM-CEUN-IMT).

Resumo. *O controle de movimentação de robôs pode ser efetuado remotamente ou de forma autônoma. Neste trabalho de pesquisa pretende-se desenvolver um sistema de controle remoto para robô humanoide comandado por voz. O sistema de reconhecimento de voz consiste em microfone, sistema de aquisição de dados e computador por meio do qual o usuário poderá enviar comandos para o robô. Um algoritmo de processamento de sinal baseado em banco de filtros e rede neural artificial foi desenvolvido em computador para fazer a identificação de um conjunto de seis comandos. O resultado do processamento de sinal será utilizado para definir a movimentação de um robô humanoide conforme a identificação do comando de voz.*

Introdução

Comandos remotos para o controle de movimentação de veículos são utilizados em uma grande variedade de aplicações. Podemos citar, na área de engenharia, o controle de robôs e braços robóticos (Rowe et al., 2009), (Yamamoto et al., 2006) e na medicina, em cirurgias controladas remotamente (Palep, 2009). Dentre estes, pode-se citar os comandos por meio do reconhecimento de voz, que também podem ser utilizados em robôs militares ou no controle automático de cadeiras de rodas (Simpson e Levine, 2002), (Al-Rousan e Assaleh, 2009).

As técnicas mais utilizadas para o reconhecimento de voz são as redes neurais artificiais, o Modelo Oculto de Markov (*HMM – Hidden Markov Model*), o modelo híbrido e por áudio-visual. Projetos que envolvem o reconhecimento de um pequeno número de comandos feitos por palavras isoladas e com vocabulário pequeno, tipicamente apresentam bom desempenho se realizados com redes neurais artificiais, com a vantagem de permitir a simplicidade na realização, o que resulta em sistemas com processamento rápido, inclusive com possibilidade de implementação em aplicações embarcadas (Cardoso et al., 2010).

Neste trabalho propõe-se o desenvolvimento de um sistema de reconhecimento de voz, implementável computacionalmente para permitir o controle de movimentação de um robô ou veículo móvel. Utiliza-se um sistema constituído de microfone, algoritmo para processamento de sinal baseado em análise de espectro e filtragem de componentes de frequência, rede neural artificial e sistema de aquisição de dados por meio do qual o usuário poderá enviar comandos para o robô. O algoritmo para processamento de sinal e identificação foi desenvolvido em software LabVIEW®. A rede neural utilizada é uma rede *MLP* (*Multilayer Perceptron*) treinada em MATLAB®.

Destaca-se como principal objetivo obter um sistema de reconhecimento de voz que possa ser utilizado em robôs móveis – priorizando o tipo humanóide – que não possuam alto poder de processamento, pois a rede neural artificial possui apenas 20 entradas.

A estrutura de um sistema clássico de reconhecimento de voz pode ser dada pela Figura 1, consistindo de 5 etapas fundamentais (Cardoso et al., 2010).

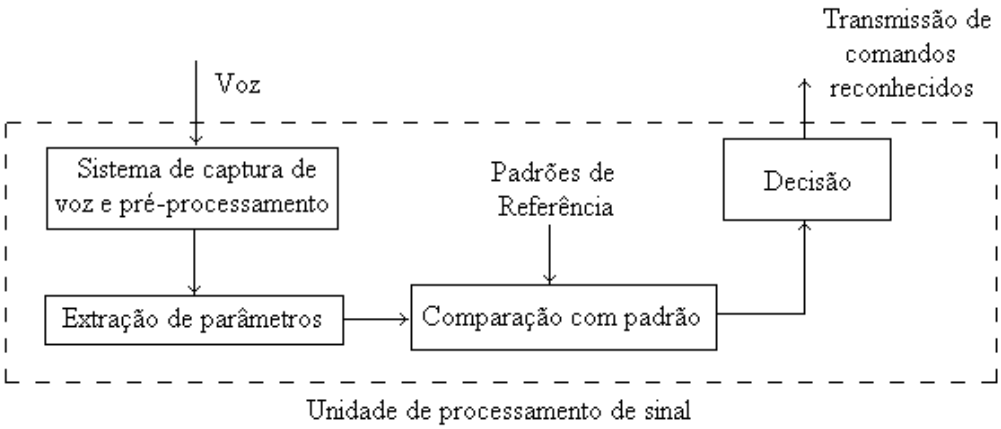

Figura 1 – Sistema de Reconhecimento de Comandos de Voz

Na etapa inicial realiza-se a captura do comando de voz e o pré-processamento deste sinal com o objetivo de minimizar ruídos na aquisição.

Na etapa seguinte faz-se a extração de parâmetros. Há várias formas de representar a informação da voz. Cardoso et. al cita três modelos de representação: o LPC (*Linear Predictive Coding*), o modo CEPSTRAL e o banco de filtros. O LPC é um modelo de codificação por predição linear que fornece coeficientes lineares que predizem o comportamento futuro do sinal. O modelo Cepstral (Sanchez, 2008) é definido como a transformada inversa de Fourier do logaritmo do espectro do sinal. O banco de filtros consiste na utilização de filtros passa-faixa, próximos um do outro, cobrindo a faixa de frequências perceptíveis pelo ouvido.

Após extrair as características do sinal, é formado um conjunto de padrões de referência durante a fase de treinamento que são utilizados como base de conhecimento sobre as palavras ou sons "ensinados". Durante a fase de reconhecimento realiza-se a comparação com os parâmetros de cada padrão de referência e a partir desta comparação toma-se uma decisão, escolhendo-se o padrão de referência que mais se aproximou do novo padrão. Nesta etapa as técnicas mais utilizadas para o reconhecimento de voz são as redes neurais artificiais (Al-Rousan e Assaleh, 2009,) (Cardoso et al., 2010), (Rhe et al., 2000), o modelo oculto de Markov (Huang e Lee, 1991), (Kutjic et al., 2007), o modelo híbrido (Trentin e Gori, 2003) e por análise áudio-visual.

A utilização de redes neurais nesta etapa é reconhecidamente a melhor solução para lidar com padrões estáticos, permitindo que uma palavra inteira possa ser reconhecida de forma simples e direta.

Material e Métodos

O sistema de reconhecimento de voz elaborado neste projeto possui diversas etapas, vistas a seguir, e cujo diagrama de blocos está representado na Figura 2.

Na etapa de pré-processamento, utilizou-se um microfone estéreo de 2 canais, de 16 bits e 48000 Hz (*DVD Quality*) acoplado a um sistema de aquisição de dados pela placa controladora de áudio de alta definição (Intel 82801B ICH9), ambos integrados à um computador portátil (*notebook*). Um aplicativo distribuído pelo próprio fabricante da placa controladora de áudio permite realizar um pré-tratamento do sinal, pela captura dos dois canais e filtragem de ruídos em frequência muito elevadas (acima de 40000 Hz). O sistema ainda permite alterar a sensibilidade do microfone, ajustando a amplitude do sinal capturado para evitar a saturação do mesmo.

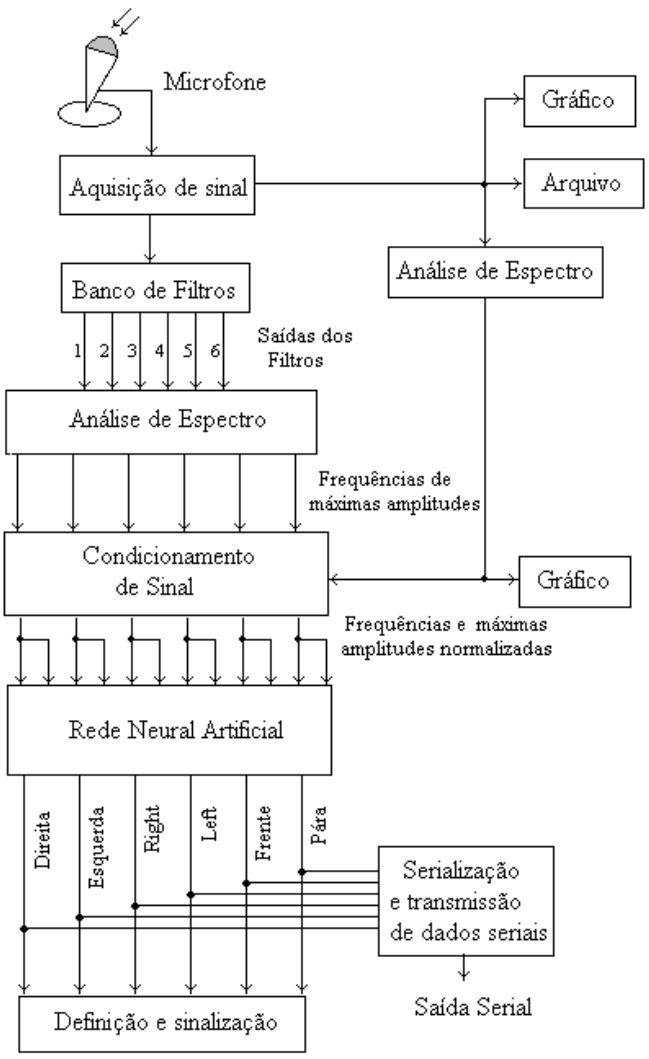

Figura 2 – Diagrama de Blocos do Sistema de Reconhecimento de Voz

Na etapa de extração de parâmetros, desenvolveu-se um algoritmo em LabVIEW® por meio do qual faz-se a aquisição de sinais através do microfone integrado, em uma frequência de amostragem de 40 kHz. Posteriormente realiza-se a análise de espectro e, em seguida, o resultado é dividido em faixas de frequências definidas, para então selecionar a maior amplitude dentro de cada uma destas e utilizá-las como entrada na rede.

A etapa de reconhecimento de padrões, foi realizada por meio de uma rede neural, previamente treinada em MATLAB®, que retorna 1 de 6 possíveis saídas, de acordo com a entrada da rede. Como critério de escolha da saída ativada, convencionou-se que apenas será considerada ativa a entrada que possuir valor superior a 0,4 e, caso duas ou mais saídas o tenham, a saída com maior valor será considerada ativa.

Aquisição de Sinais

Na aquisição de sinais considerou-se que o usuário deve enviar a mensagem de voz imediatamente após o acionamento de um botão de comando. Após este comando, o software em LabVIEW® realiza a aquisição de dados por meio de um microfone conectado na placa de som do computador. Para isto utiliza-se a função *Acquire Sound* disponível na tela de ferramentas de processamento de sinal onde se configura a aquisição em um canal (mono) durante 2 segundos (visualização em gráfico de 1,5 segundos) e utilizando uma frequência de amostragem de 40 kHz. A escolha do tempo de amostragem deve satisfazer ao teorema da

amostragem, ou seja, $\Delta T < 1 / (2*f_{máx})$ onde $f_{máx}$ é a máxima frequência que será considerada na etapa de filtragem.

Antes da definição das faixas de frequência para o banco de filtros realizou-se uma análise dos espectros de frequência produzidos por um usuário e utilizando seis comandos de voz: “DIREITA”, “ESQUERDA”, “RIGHT”, “LEFT”, “FRENTE” e “PÁRA”. Por meio da identificação destes comandos podemos definir os movimentos desejados para o robô que receberá estes comandos, respectivamente, movimenta-se para a direita, movimenta-se para a esquerda, gire em sentido horário, gire em sentido anti-horário, ande para frente e pare.

Filtragem e Análise de Espectro

Os espectros de frequências são obtidos utilizando a função *Spectral Measurements* que realiza a análise espectral do sinal com base na transformada rápida de Fourier (*FFT – Fast Fourier Transform*). A *FFT* é essencialmente um algoritmo eficiente para o cálculo numérico da transformada discreta de Fourier (*DFT - Discrete Fourier Transform*). A *DFT* (1) é muito usada no estudo do espectro de sinais sendo tipicamente determinada numericamente utilizando processadores digitais de sinais (*DSP’s*) ou computadores. Considerando-se *N* amostras do sinal no domínio do tempo, denotadas por *f(k), k = 0, 1, 2, ..., N-1*, a *DFT* é dada por um conjunto de *N* amostras do sinal no domínio da frequência dadas por *F(n), n = 0, 1, 2, ..., N-1* e definidas por:

$$F(n) = \frac{1}{N} \sum_{k=0}^{N-1} f(k) e^{-2\pi nk / N} \qquad (1)$$

Para selecionar corretamente as faixas de frequência utilizadas, realizou-se a análise dos espectros de sinais obtidos para todos os comandos possíveis ditos por quatro usuários com timbres diferentes, observou-se que há uma considerável variação no posicionamento das componentes de frequência para cada um deles. Contudo, notou-se que todos os comandos produzem raias concentradas na faixa entre 60 Hz e 6,4 kHz. Como outra constatação, observou-se que para cada comando, as raias presentes estão concentradas em determinadas faixas de frequências.

Com o objetivo de simplificar o algoritmo e reduzir o número de entradas para a etapa de reconhecimento de padrões, optou-se por uma solução simplificada onde foram delimitadas vinte faixas de frequências. Desta forma, a rede neural utilizada na etapa de reconhecimento poderá utilizar um número de entradas reduzido. Diante disto, o algoritmo desenvolvido realiza a análise de espectro e faz a separação do resultado nas seguintes faixas de frequência (com uma variação de 320 Hz entre cada uma): Faixa 1: 60 Hz até 320 Hz; Faixa 2: 320 Hz até 640; ...; Faixa 20: 6080 Hz até 6400 Hz.

Redes Neurais Artificiais

O neurônio pode ser entendido como um dispositivo que tem muitas entradas e somente uma saída conforme o modelo apresentado na Figura 3 que é amplamente aceito pela comunidade científica (Haykin, 1998):

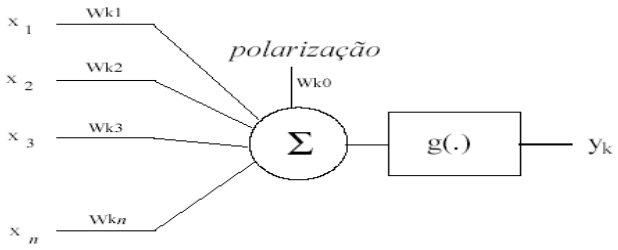

Figura 3 – Modelo Não-Linear de Um Neurônio

onde x_i é a excitação de entrada na sinapse i; y_k é a resposta (ou saída) do neurônio k;w_{ki} é o peso sináptico da entrada i do neurônio k; w_{k0} é o peso sináptico da entrada de polarização do neurônio k *e* $g(.)$ é a função de ativação do neurônio. Esse modelo consiste basicamente em um *perceptron* com função de ativação semilinear (Haykin, 1998). O *perceptron* modela um neurônio processando uma soma ponderada de suas entradas e submetendo o resultado a uma camada de processamento de limiares g(.) (Figura 3).

Os principais tipos de função de ativação são: reta, degrau, linear, sigmoide e tangente hiperbólica.

Nas redes neurais que incluem neurônios com pesos fixos, os pesos de alguns (ou de todos eles) são fixados em um dado valor, ou fixados ao peso (variável) de outro neurônio. Esse tipo de rede é tipicamente aplicado a soluções de problemas particulares e tem aplicação tradicionalmente ligada à invariância espacial.

Nas redes com treinamento não supervisionado não existe a apresentação de mapeamentos entrada-saída à rede; caberá exclusivamente a ela a tarefa de realizar a classificação, com base na informação de número de classes e topologia da rede. É o caso das redes auto-organizadas de Kohonen (Haykin, 1998).

Nas redes com treinamento supervisionado, tipicamente, uma sequência de padrões de entrada associados a padrões de saída é apresentada à rede. Esta utiliza as comparações entre a sua classificação para o padrão de entrada e a classificação correta dos exemplos para recalibrar seus pesos. Enquadram-se nesse contexto a maioria das redes utilizadas, como o *perceptron* multicamada (*MLP*).

A rede neural utilizada neste trabalho é uma rede MLP (*Multilayer Perceptron*). O treinamento da rede foi realizado em MATLAB® utilizando comandos previamente gravados de apenas um usuário, pela possibilidade de obter uma grande quantidade de amostras. A qualidade da rede neural está diretamente relacionada ao número de amostras conseguidas.

Resultados e Discussão

O fator mais importante para o correto funcionamento do projeto como um todo está na escolha de uma rede que apresente os melhores resultados de reconhecimento possíveis. Foram realizadas algumas possíveis combinações de número de neurônios em cada camada e também número de camadas, todas apresentadas na Tabela 1.

As funções de ativação sigmóide utilizadas no treinamento das redes estão presentes na Tabela 1, outras das funções disponíveis foram utilizadas para verificar qual seria a melhor para o corrente propósito, porém como os resultados foram muito inferiores aos aqui apresentados, alguns inclusive não possuíram convergência, estes dados não foram incluídos neste relatório.

Os seguintes parâmetros de configuração foram utilizados em todas as redes neurais criadas e apresentadas: *show* = 25, *showWindow* = *true*, *showCommandLine* = *false*, *epochs* = 5000, *time* = *Infinity*, *goal* = 0, *max_fail* = 200, *mem_reduc* = 1, *min_grad* = $1*10^{-10}$, *mu* = 0.001, *mu_dec* = 0.1, *mu_inc* = 10 e *mu_max* = 10000000000.

Na rede neural construída foram coletados 300 sinais de áudio sendo 50 sinais relativos a cada um dos seis comandos definidos. Dentre estes, 40 amostras foram utilizadas para o treinamento da rede neural e as 10 restantes para a verificação do resultado obtido para cada comando, de forma a fazer a validação e verificar a eficiência da rede desenvolvida. O total utilizado para o treinamento foi de 240 vetores de entradas, porém o MATLAB® utiliza parte desses dados para uma verificação interna e não os utiliza no treinamento em si, portanto para contornar este problema criou-se um novo vetor com quatro repetições do vetor original, totalizando 960 dados.

As redes elaboradas possuem 20 entradas, portanto 20 neurônios na camada inicial, 6 neurônios na camada de saída, e algumas variações do número de neurônios da(s) camada(s) do centro, cujos resultados estão apresentados na Tabela 1.

Tabela 1 – Resultados das Redes Neurais

Rede	**Camada 1** (neurônios / função)	**Camada 2** (neurônios / função)	**Acerto com dados do treinamento** (%)	**Acerto com novos dados** (%)
1	20 / TANSIG	20 / TANSIG	91,67%	43,33%
2	20 / TANSIG	20 / TANSIG	98,75%	88,33%
3	20 / TANSIG	20 / TANSIG	19,17%	13,33%
4	20 / TANSIG	-	97,08%	68,33%
5	20 / TANSIG	-	98,75%	60,00%
6	20 / TANSIG	-	99,58%	71,67%
7	30 / TANSIG	30 / TANSIG	99,17%	63,33%
8	30 / TANSIG	30 / TANSIG	98,75%	70,00%
9	30 / TANSIG	30 / TANSIG	97,08%	70,00%

Para todas as redes acima, a função de ativação sigmóide utilizada para a saída foi o *PURELIN* (linear). A opção *TANSIG* se refere à tangente hiperbólica utilizada no treinamento. Percebe-se uma grande variação entre redes com características idênticas, isso se deve ao fato de que a matriz de pesos é iniciada com valores quase aleatórios – estes devem seguir certas restrições – e, dependendo destes valores, a resposta pode ou não convergir. A diferença entre as redes idênticas de número 2 e 3 (Tabela 1) pode ser utilizada para ilustrar esse problema.

As redes escolhidas foram as de número 2 (88.33% de acerto para novas entradas) e 6 (71.67%), sendo que a escolha final depende do sistema computacional em que será inserido, pois uma delas possui duas camadas centrais de 20 neurônios cada e a outra apenas uma camada central, com 20 neurônios. A exigência computacional aumenta consideravelmente com o número de camadas e neurônios.

Conclusão

Os resultados obtidos foram comparados com os melhores de outras propostas, conforme pesquisa realizada.

Foram considerados na análise comparativa os seguintes trabalhos:

- Rhe et al. (2000), que utilizou banco de filtros, rede neural MLP com 256 entradas e 39 neurônios ocultos, dígitos em inglês e 16 usuários;
- Trentin e Gori (2003), que utilizou processamento híbrido com HMM e rede neural artificial, dígitos em italiano e 40 usuários;

Os resultados comparativos estão apresentados na Tabela 2.

Tabela 2 – Resultados Comparativos para Reconhecimento de Voz em Aplicações com Múltiplos Usuários

Origem	Índice de Acerto
Rhe et al. (2000)	96 ,00%
Trentin e Gori (2003)	94,65 %

Embora os resultados obtidos (88,33%) estejam abaixo daqueles obtidos por outros trabalhos ainda assim a proposta aqui apresentada neste trabalho é relevante, pois:

- utiliza-se uma rede neural limitada, com apenas 20 entradas e um número reduzido de comandos de voz no reconhecimento;
- é necessário avaliar melhor os resultados obtidos utilizando um número maior de comandos na análise dos resultados.

Adicionalmente, a principal contribuição deste trabalho é o desenvolvimento de um sistema de identificação em tempo real que permite controlar automaticamente a movimentação de robôs humanoides utilizando uma interface computacional flexível e de fácil utilização desenvolvida em LabVIEW®.

Adicionalmente deve-se salientar que o sistema permite criar arquivos executáveis autônomos de forma que o executável resultante pode rodar em qualquer sistema operacional, mesmo sem a necessidade de instalar todo o programa. Isto significa que o aplicativo pode rodar em qualquer processador com capacidade de memória e velocidade de processamento compatível, o que permite realizar futuramente a aplicação "embarcada" no robô humanoide.

Agradecimentos

Sinceros agradecimentos ao programa PIBIC / CNPq / CEUN-IMT pela bolsa de iniciação científica que possibilitou o desenvolvimento deste projeto.

Referências Bibliográficas

Al-Rousan, M.; Assaleh, K. (2009) A Wavelet and Neural Network Based Voice System for a Smart Wheelchair Control. *Journal of the Franklin Institute*.

Cardoso, S. A.; Castanho, J. E. C.; Franchin, M. N.; Fontes, I. R. (2010). *Sesame: Sistema de Reconhecimento de Comandos de Voz Utilizando PDS e RNA*. XVIII Congresso Brasileiro de Automática, 12 a 16 de setembro, 2010, Bonito, MS, 1316 – 1323.

Haykin, S. (1998) *Neural Networks: A Comprehensive Foundation*, 2nd edition, Prentice-Hall, 842 p.

House, B.; Malkin, J.; Bilmes, J. (2009). *The VoiceBot: a Voice Controlled Robot Arm. Proceedings of the 27th International Conference on Human Factors in Computing Systems*. ACM Digital Library, Boston, MA, USA, April, 4-9.

Huang X.; Lee K. (1991) On Speaker-independent, speaker-dependent, and speaker-adaptative speech recognition. *Proceedings of ICASSP 91 – International Conference on Acoustics, Speech, and Signal Processing*. Toronto, 877-880.

Kutjic, B.; Janos, S.; Tibor, S. (2007). *Móbile Robot Controlled By Voice*. SiSY 2007 – 5th International Symposium on Intelligent Systems and Informatics.

Palep, J. H. (2009). Robotic Assisted Minimally Invasive Surgery. *Journal of Minimal Access Surgery*. V.5, I.1, p. 1-7.

Rhee, M. K.; Young-ik K.; Geon, H. L. (2000) A Noise-Robust Front-End Based on Tree-Structured Filter-Bank for Speech Recognition. IJCNN 2000 – *Proceedings of the IEEE – INNS – ENNS International Joint Conference on Neural Networks*. p.81-86 vol. 6.

Sanchez, F. L. (2008) *Análise Cepstral Baseada em Diferentes Famílias de Transformada Wavelet*. Dissertação de Mestrado. Universidade de São Paulo, USP.

Simpson, R. C.; Levine, S. P. (2002). Voice Control of a Powered Wheelchair. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. October, v. 10, I. 2, p.122-125.

Trentin E.; Gori, M. (2003) Robust Combination of Neural Networks and Hidden Markov Models for Speech Recognition. *IEEE Transactions on Neural Networks*. 14(6): 1519-1531.

Yamamoto, S.; Nakadal, K.; Nakano, M.; Tsujino, H.; Valin, J. M.; Komatani, K.; Ogata, T.; Okuno, H. G. (2006). Real-Time Robot Audition System That Recognizes Simultaneous Speech in The Real World. *IEEE / RSJ International Conference on Intelligent Robots and Systems*, Beijing, october, p. 5333-5338.