Engajamento social e estimativa de orientação 3D de pessoas

Marcos Vinicius Cruz (CTI) mvcruz@cti.gov.br

Resumo

A área de Interação Humano Robô visa sempre criar robôs que são capazes de entender e interagir com humanos ao mesmo tempo que realizam tarefas mais complexas e de forma autônoma. Com isso, é muito importante entender o comportamento das pessoas antes de iniciar uma interação. Neste artigo é introduzido o conceito de engajamento social e mecanismos para identificar orientação de múltiplas pessoas. Com experimentos em ambiente controlado, foram atingidos os resultados esperados.

Palavras-chave: HRI, Engajamento Social, HOE, Orientação.

1. Introdução

O projeto Fapesp ROSANA visa o desenvolvimento de robôs socialmente interativos atuando em ambientes públicos. Entre as contribuições científicas e tecnológicas previstas pelo projeto encontram-se:

a. A inclusão de capacidades sociais em um robô móvel, permitindo que o mesmo seja utilizado como recepcionista no CTI – Campinas e constitua referência para eventuais produtos relacionados a robôs sociais;

b. A implementação de uma arquitetura de software robótica reconhecida na literatura apta a gerenciar um robô.

As Atividades relacionadas à contribuição "b" foram realizadas e apresentadas na Jornada PCI 2021. Neste ano estudou-se soluções para alcançar resultados relacionados à contribuição "a", neste sentido serão apresentados métodos e técnicas utilizadas em engajamento social para melhorar as capacidades sociais dos robôs móveis antes de começar uma conversação e o novo componente do sistema capaz de obter a orientação de pessoas e identificar grupos sociais através de uma imagem RGB.

Segundo Leite (2016), engajamento pode ser definido como uma métrica que sistemas robóticos usam para avaliar a qualidade da interação com humanos e se os mesmos estão interessados em manter a interação, consequentemente tendo a capacidade de se adaptar a diferentes cenários ou possuir alguma característica (física ou virtual) que motive a pessoa a se aproximar para uma interação. Existem, ainda, as definições de engajamento de Inoue (2018) e Sidner (2003) que definem como um estado de interação e como processo de início, manutenção e término de interação, respectivamente. Mostrando que não existe um senso comum entre os pesquisadores sobre engajamento social em Interação Humano-Robô (HRI).

A ideia apresentada neste plano de trabalho é de avaliar se um usuário possui a intenção de interagir com o robô utilizando o conceito de engajamento como um estado antes do processo

de interação e criar um modelo para definir um plano de ação para motivar o usuário a se aproximar.

2. Engajamento social

O contato inicial entre humanos e robôs é de extrema importância tanto para um engajamento de curto prazo quanto para uma interação de longo prazo. Sua importância pode ser extraída de estudos antropológicos que relatam que os primeiros encontros determinam a direção dos relacionamentos e se as pessoas desejam continuar em contato. Humanos espontaneamente começam a formar impressões e julgamentos sobre o outro, e essas impressões podem durar um tempo significativo após interações (SUNNAFRANK, 2004).

Embora esses efeitos se refiram à impressão na interação entre seres humanos, os estudos indicam que pessoas tendem a avaliar e julgar robôs e agentes virtuais da mesma forma que fazem com outros humanos. Assim, por exemplo, as pessoas podem fazer julgamentos negativos se um robô se comportar de maneira inadequada nos primeiros momentos de interação, o que afetará a confiança no robô a longo prazo (REEVES, 1996).

Zhang et.al (2021) propõe engajamento como intenção da pessoa em iniciar uma interação com uma inteligência artificial (AI), focando no período antes de começar uma interação. Deste modo, as AIs avaliam continuamente se as pessoas têm vontade de iniciar uma conversa e, se sim, elas podem cumprimentar ativamente os participantes em potencial. Desta forma, as AIs dão impressão de serem inteligentes, e o HRI torna-se mais natural e socialmente amigável.

Baseado na análise de comportamento feito por Troung et.al em (2017), durante o processo de interação, o robô precisa enviar sinais sociais para a pessoa de forma mais ativa e esta interação deve ser aprimorada durante este processo, chamada de “interação progressiva”. Conforme a pessoa se aproxima ou afasta do robô, ela possui expectativas de que o robô realizará alguma ação. O autor divide esse processo em três estágios:

- Estágio distante: o objetivo dos robôs nesta fase é chamar a atenção dos usuários;
- Estágio intermediário: robôs devem expressar ainda mais a intenção de interagir, e tornar os usuários claramente cientes;
- Estágio próximo: robôs precisam começar um diálogo.

3. Localização e orientação 3D com múltiplas pessoas

Detecção de pessoas em imagens e estimar o seu esqueleto é um problema amplamente estudado. Os métodos mais avançados baseiam-se em Redes Neurais Convolucionais e podem ser agrupados em métodos top-down e bottom-up. Como mostra em Xiao (2018) as abordagens de top-down consistem em detectar cada instância na imagem primeiro e depois estimar as articulações do corpo dentro do limites da caixa de delimitação inferida. Abordagens bottom-up estimam separadamente cada juntas do corpo através de arquitecturas convolucionais e depois combinam para obter uma pose humana completa.

Mais recentemente Kreiss (2019) propôs um método adaptado para cenários de condução autónoma que funcionam bem em cenas de baixa resolução, com muita gente e ocluídas. Relacionado com o nosso trabalho base, que mostra a eficácia de informação latente contida

em estímulos de articulações 2D. Conseguem resultados através da simples previsão de juntas 3D em dados 2D através de uma rede neural totalmente conectada. Embora estes trabalhos estimem posições relativas em 3D das juntas do corpo, não fornecem qualquer informação sobre a localização real em uma cena 3D (KU, 2019).

A estimativa da orientação corporal (HOE) fornece pistas visuais cruciais em muitas aplicações, incluindo a robótica e a condução autônoma. É particularmente desejável quando a estimativa da pose 3D é difícil de inferir devido à má resolução da imagem, oclusão, ou partes indistinguíveis do corpo (WU, 2020).

HOE justifica ser abordada como um problema autônomo por três razões. Primeiro, a pose 3D pode ser difícil de inferir devido à má resolução da imagem, oclusões presentes em imagens de ambientes não controlados. Em segundo lugar, sob certos cenários, a orientação do corpo já é suficiente para ser utilizado como ponto de partida para tarefas de previsão ou planejamento estratégicos. Em terceiro lugar, o custo computacional de um modelo de orientação corporal é muito mais reduzido em comparação com um modelo 3D , isso o torna mais atrativo para a implementação no dispositivo. Além disso, a estimativa da orientação corporal e a pose 3D pode ser complementar na abordagem do mundo real em desafios que envolvem a compreensão dos comportamentos humanos (WU, 2020).

4. Metodologia

Apresenta-se a seguir a metodologia utilizada nos estudos relacionados a HRI que serão aplicados no projeto FAPESP.

4.1 Engajamento Social

Levando em consideração restrições sociais e proxêmicas, as informações dos sensores do robô, como posição e distância, combinadas influenciam em como o robô realiza o engajamento definindo expressão facial e a sentença mais adequada para iniciar o processo de engajamento. Um conjunto de estados (Figura 1) de comportamentos para conseguir sincronizar as expressões e fala do avatar quando o comportamento de um dela muda, dependendo do seu estado.

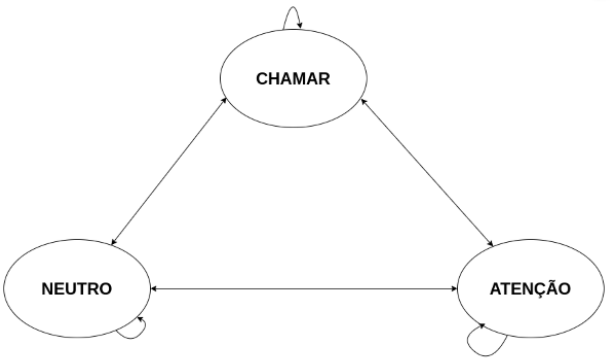

Figura 1 – Padrão dos estados comportamentais do avatar

Para o controle do robô durante o processo de engajamento, utiliza-se um modelo oculto de Markov (HMM); as zonas de interação são um processo observável independente que manifesta um processo estocástico latente capaz de realizar as tomadas de decisão de ação do

robô. Assim, de acordo com as observações das zonas que o usuário se encontra, realiza-se uma cadeia de ações que definem o estado atual. Para criar o modelo foi necessário realizar um treinamento com uma base de dados simulada (Figura 2).

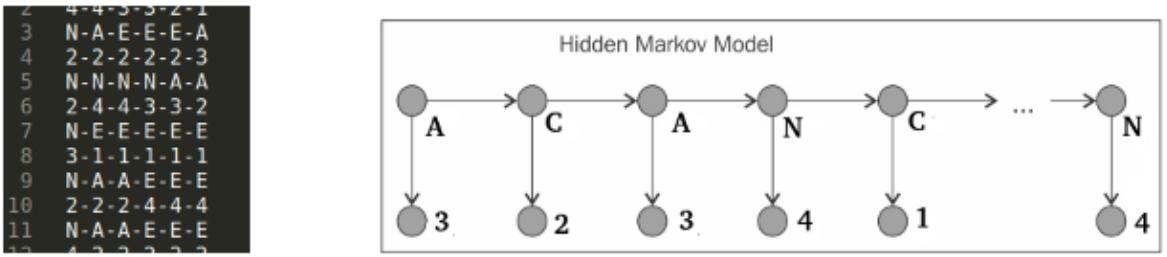

Figura 2 – A esquerda base de dados simulada; a direita modelo HMM treinado

4.2 Estimativa de HOE

Anteriormente a orientação era obtida através de um treinamento realizado numa máquina especializada para processamento usando uma ferramenta chamada OpenPose, que calcula as juntas do esqueleto da pessoa e retorna em forma de vetor essas coordenadas. Com as coordenadas das juntas é possível construir a postura da pessoa e, utilizando uma estratégia de calcular o ângulo entre as coordenadas presentes no vetor, é possível estimar a orientação do corpo humano. Entretanto, a solução não se enquadra no atual projeto, pois o sistema precisa de maior precisão na estimativa do ângulo e um tempo de resposta menor. Para decidir a melhor forma de obter as orientações realizou-se uma revisão bibliográfica sobre o estado da arte em HOE, tendo como prioridades facilidade de implementação, precisão e com um desempenho aceitável para rodar em nuvem.

Durante os estudos o MEBOW (Monocular Estimation of Body Orientation In the Wild) foi escolhido como potencial modelo para estimar orientação. O modelo tem um desempenho significativamente superior ao se comparar com soluções de última geração para a estimativa da posição humana utilizando câmera monocular, onde a formação utiliza apenas informações 3D e 2D. Após algumas tentativas de adaptar o modelo ao sistema como um componente de arquitetura distribuída não foi possível realizar a detecção em tempo real pois o modelo foi feito para rodar offline processando um grande volume de dados simultaneamente (WU, 2020).

Após novas pesquisas, a melhor escolha foi uma biblioteca criada por Bertoni (2019) chamada monoloco, que possui vários conjuntos de modelos para trabalhar com identificação de pedestres e identificação de esqueleto, além de conseguir detectar o ângulo de pessoas e identificar grupos sociais. Após algumas adaptações no código-fonte e implementação de paralelismo no código foi possível implementá-lo como um módulo do sistema para detecção.

5. Experimentos e discussão de resultados

Apresenta-se dois experimentos realizados, um associado ao engajamento social e outro associado à estimativa de orientação. Os experimentos foram realizados em ambiente controlado indoor com pessoas reais com membros da divisão DISCF. Com o intuito de testar a eficácia dos sistemas em diversas situações. Os testes de engajamento foram realizados com o agente virtual ANA – Robô recepcionista e utilizando uma câmera estéreo que provê

informações de distância. Os testes de orientação e detecção de grupos foram feitos em uma máquina com suporte à GPU de forma a otimizar o processamento diminuindo o tempo de resposta.

5.1 Engajamento Social

Os resultados dos testes do engajamento social foram satisfatórios, como mostra na figura 3: no item (a) o avatar está repousando enquanto não há presença de usuário; no (b) avatar fica triste quando é ignorada; em (c) Avatar segue pessoas com os olhos para chamar atenção; (d) Início do diálogo após o engajamento social. Como o engajamento é uma etapa antes do início do diálogo não se usa o microfone. Em alguns momentos existia delay nas respostas da recepcionista devido à latência na rede.

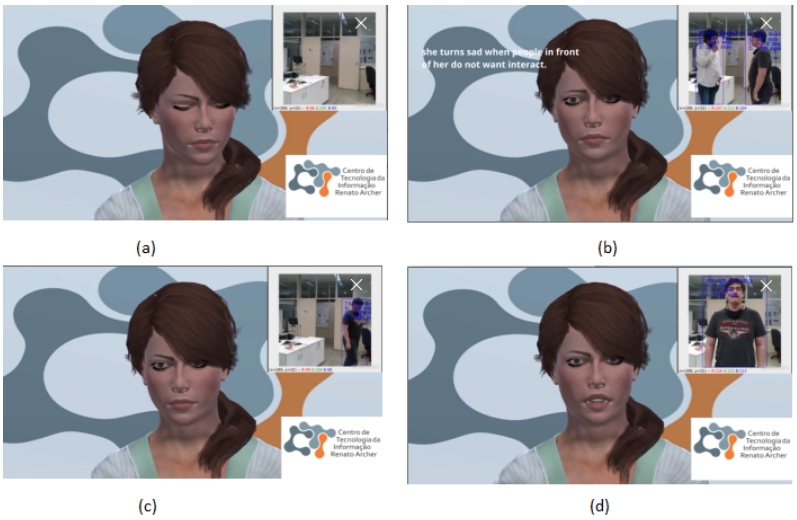

Figura 3 – Estudos de casos do engajamento social

5.2 Estimativa de orientação

Em relação ao experimento de estimativa da orientação tivemos uma acurácia em torno de 94%, porém em relação à localização real da pessoa os resultados foram abaixo da expectativa, pois o algoritmo não consegue identificar a distância real da pessoa em imagens de pequena escala e sem o esqueleto inteiro e, consequentemente afeta diretamente a detecção de grupos sociais.

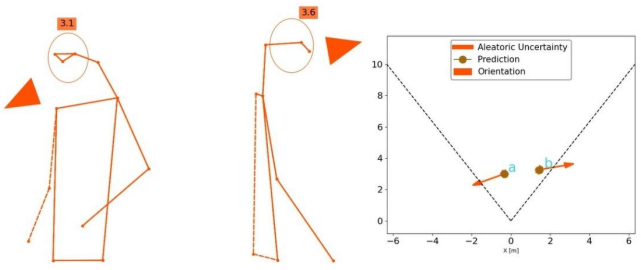

Figura 4 – Estimativa da orientação da pessoa

6. Conclusão

Com os avanços alcançados esse ano podemos dizer que os trabalhos são promissores e os planos para as próximas são aprimoramento do engajamento que, devido a limitação do

avatar, não é possível ser tão expressivo. Para tal, será utilizado o robô humanoide Pepper, já adquirido, que possui alto nível de expressão..

Para o projeto Fapesp é necessário identificar a posição relativa da pessoa e transformá-la para o mundo real. Sendo assim, serão realizados maiores estudos de localização 3D para conseguir melhores resultados partindo do princípio que a estimativa de orientação funciona.

Referências

BERTONI, L.; KREISS, S.; ALAHI, A.. *Monoloco: Monocular 3d pedestrian localization and uncertainty estimation*. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019. p. 6861-6871.

INOUE, K. et al. *Engagement recognition by a latent character model based on multimodal listener behaviors in spoken dialogue.* APSIPA Transactions on Signal and Information Processing, v. 7, 2018.

KREISS, S.; BERTONI, L.; ALAHI, A. *Pifpaf: Composite fields for human pose estimation.* In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019. p. 11977-11986.

KU, J.; PON, A. D.; WASLANDER, S. L. *Monocular 3d object detection leveraging accurate proposals and shape reconstruction.* In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019. p. 11867-11876.

LEITE, I. et al. *Autonomous disengagement classification and repair in multiparty child-robot interaction.* In: 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). IEEE, 2016. p. 525-532.

REEVES, B.; NASS, C. *The media equation: How people treat computers, television, and new media like real people.* Cambridge, UK, v. 10, p. 236605, 1996.

SIDNER, C. L.; LEE, C.; LESH, N. *Engagement when looking: behaviors for robots when collaborating with people.* In: Diabruck: Proceedings of the 7th workshop on the Semantic and Pragmatics of Dialogue. University of Saarland, 2003. p. 123-130.

SUNNAFRANK, M.; RAMIREZ JR, A. *At first sight: Persistent relational effects of get-acquainted conversations.* Journal of Social and Personal Relationships, v. 21, n. 3, p. 361-379, 2004.

TRUONG, X.; NGO, T. *"To approach humans?": A unified framework for approaching pose prediction and socially aware robot navigation.* IEEE Transactions on Cognitive and Developmental Systems, v.10, n. 3, p. 557-572, 2017.

WU, C., et al. *MEBOW: Monocular Estimation of Body Orientation in the Wild.* In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020. p. 3451-3461.

XIAO, B.; WU, H.; WEI, Yi. *Simple baselines for human pose estimation and tracking.* In: Proceedings of the European conference on computer vision (ECCV). 2018. p. 466-481.

ZHANG, Z.; ZHENG, J.; THALMANN, N. M. *Engagement Intention Estimation in Multiparty Human-Robot Interaction.* In: 2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN). IEEE, 2021. p. 117-122.