Transferindo movimentos humanos de vídeos para robôs com Aprendizado por Reforço Profundo

Nayari Marrie Lessa[1,3], **Esther Luna Colombini**[2], **Alexandre da Silva Simões** [1]

[1]Universidade Estadual Paulista (Unesp)
Instituto de Ciência e Tecnologia (ICT) – Sorocaba, SP – Brasil

[2]Universidade Estadual de Campinas (Unicamp)
Instituto de Computação (IC) – Campinas, SP – Brasil

[3]Deutsches Forschungszentrum fuer Kuenstliche Intelligenz (DFKI)
Centro de Inovação em Robótica – Bremen – Alemanha

{nayari.lessa,alexandre.simões}@unesp.br, esther@ic.unicamp.br

Resumo. *Treinar robôs para aprender políticas complexas tem se mostrado um desafio monumental. Nesse contexto, o Aprendizado por Imitação (IL) tem como foco extrair políticas de referência de um especialista e transferi-las para robôs com a máxima fidelidade possível, geralmente através do Aprendizado por Reforço Profundo (DRL). Este trabalho apresenta um novo processo de imitação para robôs bípedes, composto por três fases distintas: i) extração de poses de especialistas humanos a partir de vídeos; ii) geração de trajetórias de referência de movimento para o robô; e iii) treinamento do robô utilizando DRL para adaptar os movimentos considerando a anatomia e dinâmica específicas do robô. Nos experimentos conduzidos em um ambiente simulado, um robô humanoide foi capaz de chutar uma bola a uma distância de 1 metro, utilizando como referência vídeos de movimentos similares realizados por humanos e extraídos do YouTube.*

Palavras-chave: *Aprendizado por Imitação. Robôs humanoides. Aprendizado por Reforço Profundo. Estimativa de postura humana.*

Abstract. *Training robots to learn complex policies has proven to be a monumental challenge. In this context, Imitation Learning (IL) focuses on extracting reference policies from experts and transferring them to robots with the highest possible fidelity, usually through Deep Reinforcement Learning (DRL). This work presents a novel imitation process for bipedal robots, consisting of three distinct phases: i) extraction of human expert poses from videos; ii) generation of reference motion trajectories for the robot; and iii) training the robot using DRL to adapt the movements, taking into account the specific anatomy and dynamics of the robot. In the experiments conducted in a simulated environment, a humanoid robot was able to kick a ball at a distance of 1 meter, using videos of similar human movements as reference, extracted from YouTube.*

Keywords: *Imitation Learning. Humanoid Robots. Deep Reinforcement Learning. Human posture estimation.*

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Eng. Elétrica (PGEE) da Universidade Estadual Paulista (Unesp) em **02/06/2022**. Orientador: **Prof. Dr. Alexandre da Silva Simões**. Coorientadora: **Profa. Dr.a Esther Luna Colombini**. Documento completo em: https://repositorio.unesp.br/handle/11449/235583. Artigo submetido ao Concurso de Teses e Dissertações em Robótica (CTDR-2023).

1. Introdução

Seres humanos possuem características e habilidades próprias construídas ao longo de milhares de anos, capazes de se locomover em ambientes diversos e complexos. A reprodução de tais habilidades em robôs é um desafio que persiste há décadas.

Dentro das estratégias de controle de robôs, as abordagens baseadas em **aprendizagem** se tornaram uma alternativa promissora. Em particular, o **Aprendizado por Reforço** (*Reinforcement Learning* – RL) tem possibilitado que o agente aprenda a executar tarefas por tentativa e erro interagindo com o ambiente. No entanto, mesmo com a combinação das vantagens dessa abordagem com as do Aprendizado Profundo no **Aprendizado por Reforço Profundo** (*Deep Reinforcement Learning - DRL*), ensinar políticas sofisticadas a robôs tem se mostrado altamente desafiador. Como destacado por [Dong et al. 2020], dentre os desafios dessa abordagem estão: *i)* dificuldade de seleção de amostras adequadas; *ii)* falta de estabilidade do treinamento; *iii)* dificuldade de exploração; *iV)* dificuldade de transferência do ambiente simulado para o mundo real. Uma alternativa em destaque nesse cenário consiste no **Aprendizado por Imitação** (*Imitation Learning* - IL), o qual viabiliza a extração de políticas de especialistas e sua subsequente transferência para robôs.

Com o intuito de contribuir para as investigações nessa área, este estudo propõe uma nova abordagem baseada no Aprendizado por Imitação para a locomoção de robôs bípedes. O processo de imitação proposto é composto de três fases: *i)* extração de poses de especialistas humanos a partir de vídeos; *ii)* geração de trajetórias de referência de movimento para o robô; e *iii)* treinamento do robô por meio de DRL para adaptar ou aprimorar os movimentos, considerando as diferenças entre o esquema corporal e a dinâmica do robô em relação ao especialista humano. A tarefa selecionada para a imitação foi o chute de uma bola de futebol. Os experimentos foram conduzidos com a robô humanoide Marta, desenvolvida pelos grupos de trabalho GASI (Unesp) e LaRoCS (Unicamp) [Chenatti et al. 2018]. Nos experimentos realizados, a robô foi capaz de chutar a bola, posicionada a 1m de distância, com base em vídeos de humanos realizando movimentos semelhantes obtidos do YouTube.

2. Contribuições

Como principais contribuições do presente trabalho, destacamos:

- O desenvolvimento de um novo conjunto de dados (*SoccerKicks*), que disponibiliza uma coleção de vídeos de chutes de bola parada para fornecer movimentos de referência para robôs humanoides [Lessa et al. 2021a];
- O desenvolvimento de metodologias para seleção de vídeos, estimativa de pose humana e obtenção de movimentos de referência nesse conjunto de dados;
- A criação de uma nova metodologia para treinar robôs humanoides usando vídeos do conjunto de dados, com base no Aprendizado por Reforço Profundo (DRL);

- O desenvolvimento de uma abordagem capaz de lidar com o fato de que robôs humanoides podem ter uma forma corporal significativamente diferente dos humanos nos vídeos de referência;
- A melhoria do algoritmo de DRL, adaptando-o para utilizar coordenadas de posição em vez de orientações;
- O desenvolvimento de novas funções de recompensa para o DRL;
- O desenvolvimento de um framework capaz de comportar o treinamento por reforço de agentes robóticos a partir de vídeos do youtube;
- A proposição e comparação de diferentes funções de recompensa e a investigação de técnicas que possam estimular posturas mais naturais e realistas a partir do aprendizado.

Parte das contribuições já foi publicada no trabalho [Lessa et al. 2021a], apresentado e publicado no IEEE SMC (Qualis A2).

3. Referencial Teórico

3.1. Estimativa de Pose Humana

A Estimativa de Pose Humana (EPH) é o processo de localização das articulações em conjuntos vetoriais bidimensionais ou tridimensionais que compõem a pose do copo humano em imagens ou vídeos. Para a EPH 2D, o *AlphaPose* [Xiu et al. 2018] e o *OpenPose* [Cao et al. 2019] são amplamente utilizados, fornecendo estimativas em tempo real para múltiplas pessoas a partir de uma única imagem. Dentre os sistemas para o EPH 3D, *Human Mesh and Motion Recovery* (HMMR) [Kanazawa et al. 2019] fornece uma predição da pose 3D, na geração das informações geométricas (malha) do corpo de um modelo humano e na geração de informações de câmera a partir da qual a imagem da pessoa foi obtida na estimativa 2D. Vários fatores influenciam a qualidade da EPH como movimento da câmera, ponto de vista, rotação, iluminação e oclusão das partes do corpo.

3.2. Aprendizado por Reforço Profundo

No Aprendizado por Reforço (RL), o objetivo é aprender uma política π que habilite o agente a maximizar o retorno esperado para uma dada tarefa. Em cada instante de tempo *t*, o agente interage com o seu ambiente descrito por uma política $\pi_\theta(a|s)$, parametrizada por θ, que modela uma distribuição de probabilidade condicional $p_{\pi\theta}(a|s)$, dado a observação do estado no instante de tempo s_t e da ação consequente a_t. O agente executa a ação, transiciona para um novo estado s_{t+1} e recebe uma recompensa escalar r_t, que reflete uma transição desejada. Este processo gera uma trajetória $\tau = \{(s_0, a_0, r_0), (s_1, a_1, r_1), \ldots\}$, descrito como um Processo Decisório de Markov. Portanto, o objetivo é aprender parâmetros ótimos que maximizam o retorno esperado $J_{\pi\theta} = E_{p_{\pi\theta}(\tau)}[\sum_{t=0}^{T} \gamma^t r_t]$, onde $p_{\pi\theta}(\tau)$ é a distribuição de probabilidade sobre a trajetória induzida pela política π_θ, em um horizonte de T passos e um fator de desconto de $\gamma \in [0, 1]$.

A combinação do RL com as Redes Neurais Profundas resultou no Aprendizado por Reforço Profundo (*Deep Reinforcement Learning* - DRL), que caracteriza-se pelas múltiplas observações dimensionais no espaço de estados, permitindo ao agente tomar melhores decisões [Arulkumaran et al. 2017]. Dentre os algoritmos de DRL, destacamos aqui o *Soft Actor Critic* (SAC) [Haarnoja et al. 2018], um método Off-policy onde o objetivo do ator é maximizar a recompensa esperada e a entropia $\mathcal{H}$ em cada estado visitado.

O coeficiente não-negativo α regula a estocasticidade do sistema, onde um valor alto de α indica mais exploração e um valor baixo sugere menos exploração.

3.3. Aprendizado por Imitação

O Aprendizado por Imitação (Imitation Learning - IL) consiste em utilizar demonstrações de especialistas para guiar o agente a aprender uma política ótima que imite o especialista. Uma abordagem simples para IL é usar trajetórias geradas pelo processo de HPE. Desta forma, dados os movimento de referência, o objetivo é encontrar uma política π^* que minimize a divergência entre a distribuição das características induzidas pelo aprendiz (pose do robô simulado) p_π e as características induzidas pela política de referência (movimento de referência) $\widehat{p}_\pi$ em cada quadro do vídeo.

No entanto, a imitação das ações de um especialista em condições ambientais, morfológicas e dinâmicas diferentes do aprendiz apresenta vários desafios [Cheng et al. 2021]. Dois problemas críticos relacionam-se à percepção, ou seja, como o sistema observa os estados: *i)* as diferenças corporais entre o especialista e o aprendiz, e *ii)* as discrepâncias nos pontos de vista da câmera que gravou as imagens em ambientes não controlados. Além disso, existem problemas decorrentes do controle, ou seja, da abordagem utilizada para aprender a política de imitação. Algumas abordagens são **baseadas em modelos** e envolvem modelagem direta ou inversa da dinâmica do ambiente. Por outro lado, as abordagens **livres de modelo** fazem uso de algoritmos de aprendizado que empregam diferentes estratégias para aprender a política a ser imitada. Essas abordagens serão discutidas em detalhes na próxima seção.

4. Trabalhos Correlatos

Nos últimos anos, os métodos de aprendizado têm demonstrado resultados promissores no controle de robôs humanoides. O RL [Benbrahim and Franklin 1997], e o DRL *actor-critic* [Kim et al. 2017] foram utilizados para controlar robôs bípedes. Para o controle da caminhada da robô humanoide Marta em ambiente simulado, diversas abordagens foram investigadas, incluindo o uso do Algoritmo Genético para selecionar padrões de movimento gerados por Série de Fourier Truncada [Tejada Begazo 2020], a aplicação de DRL com algoritmo SAC [Chenatti et al. 2018] [Tomazela 2019] e a abordagem combinada SAC-PPO [Soares et al. 2020].

Outra linha de pesquisa explorou o **aprendizado por imitação** de políticas de referência e sua transferência eficiente para outros sistemas. Abordagens mais recentes propuseram o uso de algoritmos baseados em em redes adversariais para permitir que caracteres imitem movimentos humanos a partir de vídeos [Peng et al. 2021]. Similarmente, foram aplicados no controle de robôs humanoides [Hudson et al. 2021]. Uma abordagem relevante nesse contexto é o treinamento da política de imitação por DRL utilizando o algoritmo de *Proximal Policy Otimization* (PPO) [Peng et al. 2018]. Entre os trabalhos avaliados, [Peng et al. 2018] propôs uma metodologia completa para utilizar vídeos do YouTube, diferentemente de estudos anteriores que utilizaram vídeos gravados em ambientes controlados com sistemas de captura.

5. Abordagem Proposta

Este trabalho propõe uma abordagem que combina a extração de movimentos humanos a partir de vídeos (política de referência) e a imitação desses movimentos em um robô

humanoide. A imitação consiste no aprendizado de uma nova política adaptada, ou seja, na adaptação da política do especialista à estrutura e dinâmica do robô.

A abordagem proposta segue três fases distintas:

1. **Estimativa de Pose Humana:** consiste na extração de um conjunto de poses humanas 3D a partir de vídeos do YouTube;
2. **Movimento de Referência:** compreende na aplicação de um conjunto de transformações para extrair poses humanas 3D ajustadas ao formato e à dinâmica do corpo do robô, levando em consideração que essas poses podem não ser idênticas às dos especialistas.
3. **Aprendizado por Imitação a partir de Observações:** Nesta etapa, DRL é utilizado para aprender a política de imitação dos movimentos de referência adaptados para o corpo e dinâmica do robô.

A figura 1 apresenta uma representação detalhada desses processos.

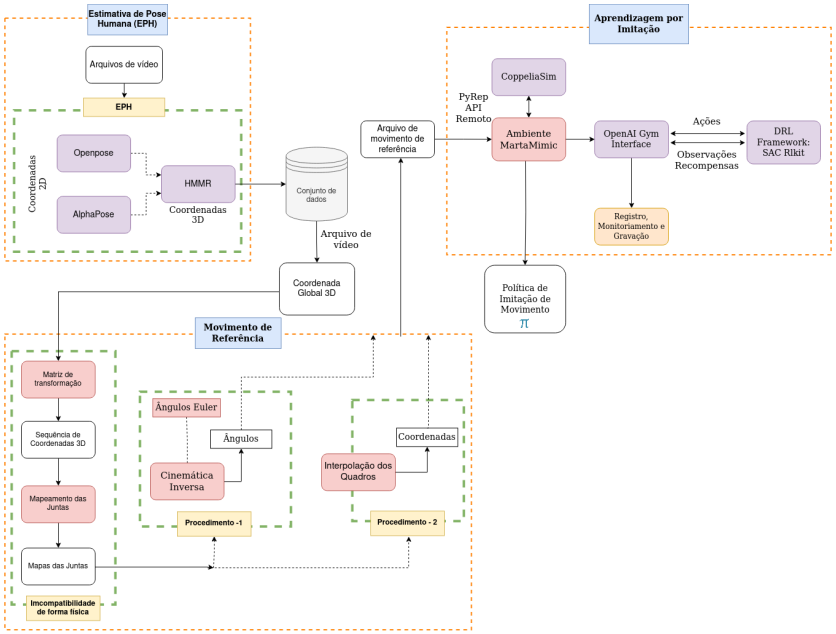


Figura 1. Diagrama esquemático detalhado das três fases do processo de imitação.

5.1. Estimativa de Pose Humana

Como parte deste trabalho, os autores desenvolveram um conjunto de dados chamado *SoccerKicks*, conforme publicado em [Lessa et al. 2021b]. Os vídeos foram selecionados da plataforma do YouTube seguindo critérios específicos: *i)* Tema: os vídeos escolhidos são jogadores de futebol executando chutes em bolas paradas (pênaltis e faltas); *ii)* Ações completas: os vídeos selecionados contêm a sequência completa do movimento realizado pelo jogador; *iii)* Ponto de vista: os vídeos selecionados não possuem movimentos abruptos, como balanços, rotações e outros; *iv)* Tamanho e escala: o jogador está completamente enquadrado em todos os quadros do vídeo; *v)* Ausência de oclusão: os vídeos selecionados não possuem obstruções de jogadores ou partes do corpo; *vi)* Presença mínima de pessoas em cena: os vídeos selecionados têm poucas ou quase nenhuma pessoa além

do jogador; *vii)* Qualidade: os vídeos foram escolhidos considerando iluminação, ruídos e outros fatores que possam afetar a qualidade da imagem.

Além disso, o conjunto de dados inclui a EPHs gerada a partir de cada vídeo. Para a geração das EPHs, foram utilizados dois modelos de estimativa 2D, o *OpenPose* com 25 *keypoints* e o *Alphapose* com 26 *keypoints*, e o HMMR para estimativa 3D a partir de ambos os modelos. Mais detalhes sobre o conjunto de dados estão disponíveis em [Lessa et al. 2021b] e podem ser encontrados em: [1].

5.2. Movimento de Referência

Aplicamos as matrizes de transformação (rotação, escala e translação) obtidas anteriormente para converter as coordenadas 3D das juntas para o sistema de referência local. Para tratar da diferença de estrutura corporal entre o ser humano de referência e o robô, foram investigados dois métodos: *i)* Utilização da cinemática inversa nas coordenadas 3D para obter um conjunto de rotações que mapeiam cada junta do corpo humano para o corpo do robô. *ii)* Direcionamento direto das coordenadas escaladas para a altura do robô ($\sim$1m).

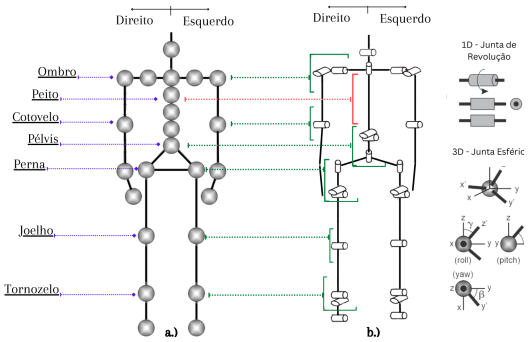


Figura 2. Adaptação corporal entre humano e robô. a) representação do esqueleto de referência (video) destacando as juntas alvo; b) Mapeamento entre as juntas da robô Marta e as juntas da referência.

5.3. Transferindo Movimentos Humanos como um Problema de Aprendizagem por Reforço

Após adaptar as sequências de poses $\widehat{q}_t$, avançamos para a etapa de transferência dos movimentos humanos para o robô. Nosso objetivo é aprender uma política π que permita ao robô imitar com máximo realismo a habilidade demonstrada no ambiente de simulação. Utilizamos o movimento de referência da fase anterior como objetivo de imitação e treinamos a política com algoritmo SAC.

Espaço de observação e ações. O espaço de observação (s) inclui a propriocepção do corpo da Marta, como as coordenadas cartesianas 3D para os membros (cintura-pélvica, tórax e tornozelos), velocidades linear (tórax) e angular (juntas) e as orientação em ângulos euler (juntas e tórax). No contexto do Procedimento -2, são adicionadas as coordenadas cartesianas 3D de todas as juntas. Para complementar o espaço de observação,

[1]Conjunto de dados SoccerKicks, página GitHub: https://github.com/larocs/SoccerKicks.

são adicionadas as ações anteriores, informações sensoriais (sensores de força, acelerômetros, centro de massa e inércia do robô), e informações relevantes à tarefa suplementar de chutar a bola (distância mínima entre a bola e o pé de chute, estado de colisão, posição do dedo do pé, força do chute e dados de posição, orientação e velocidades linear e angular da bola). A Figura 3 exemplifica a tarefa adicional. Cada ação ($a_i \in A$) determina as orientações das juntas-alvo, permitindo o controle de suas posições.

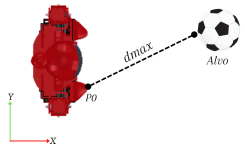


Figura 3. Ilustração da tarefa adicional. A bola é o alvo em d_{max}, (P_0) é a distância entre o pé de chute para a bola alvo.

Função de Recompensa. A função de recompensa incentiva a robô Marta a seguir a sequência de poses de referência $\hat{q}$ em cada instante de tempo, enquanto almeja o objetivo adicional de chutar a bola. A abordagem adotada segue uma metodologia semelhante à utilizada em [Peng et al. 2018] no treinamento das políticas no Método Procedimento -1. A função de recompensa é uma combinação ponderada de funções que descrevem a tarefa de imitação e a tarefa adicional, composta por: Recompensa da pélvis (R_R), Recompensa da Pose (R_P), Recompensa da Velocidade da Pose (R_V), Recompensa do tornozelo (R_{EE}), Recompensa da velocidade do Centro de Massa (R_{CoM}) e Recompensa da tarefa adicional (R_G). A função de recompensa total é calculada conforme Equação 1.

$$R_t = 0.2R_R + 0.5R_P + 0.05R_V + 0.15R_{EE} + 0.1R_{CoM} + 0.3R_G. \quad (1)$$

Na função de recompensa da pélvis ($R_R = exp[-5(\sum_j \|R_{error}\|)]$) é computado o erro ponderado das medidas de rotação, posição, velocidade linear e angular entre a junta da cintura do robô (junta Yaw) e a cintura-pélvica da referência humana.

A recompensa da Pose ($R_P = exp[-2(\sum_j \|P_{error}\|)$ tem o objetivo de incentivar o robô a seguir a sequência de orientações das juntas da referência. O cálculo da R_P envolve a computação do erro das rotações 1D para cada instante de tempo. A recompensa relativa à velocidade angular da pose ($R_V = exp[-0.1(\sum_j \|Vel_{error}\|)]$) é calculada a partir do erro entre as velocidades das poses alvo do robô e sua função exponencial.

A estabilidade do robô é promovida pela recompensa da velocidade do Centro de Massa ($R_{CoM} = exp[-10(\sum_j |CoM_{error}|)]$), que incentiva o robô a acompanhar a velocidade do Centro de Massa de referência. Além disso, a recompensa relativa à posição do tornozelo ($R_{EE} = exp[-40(\sum_j |E_{error}|)]$) serve como um estímulo adicional para o robô seguir as coordenadas de posição de referência.

Por fim, a recompensa da tarefa adicional ($R_G = exp[-4(\sum_j |G_{error}|)]$) é calculada com base no erro entre a distância do pé de chute e a posição inicial da bola. Quando o robô alcança a bola, a recompensa recebida a partir desse momento até o final do episódio é igual a um (1).

No Procedimento -2, a recompensa da pose passa a incentivar o robô a seguir a sequência de coordenadas cartesianas 3D, enquanto que a recompensa da velocidade da

Pose é determinada pela velocidade linear para cada uma das juntas alvo da referência. Similarmente, a recompensa da pélvis é determinada pelas coordenadas de posição e velocidades linear da cintura-pélvica. As funções de recompensa do tornozelo, velocidade do Centro de Massa e tarefa adicional permanecem inalteradas.

6. Procedimento Experimental

A robô Marta, na sua versão simulada neste estudo, apresenta uma altura aproximada de 1 metro e um peso de 8 quilogramas. Ela é composta por 25 juntas de revolução, sendo três delas localizadas na cintura permitindo movimentos esféricos. Destaca-se que ela possui características distintas, tais como pés com área de contato reduzida e um grau de liberdade adicional nos dedos dos pés. A simulação foi realizada no CoppeliaSim (com a Biblioteca de física Bullet), com codificação em *Python* e *PyRep*. A interface *OpenAI Gym* foi usada para abstrair o ambiente para o algoritmo SAC, executado no *framework Rlkit*. Cada episódio iniciou com a robô em pé, em um ambiente sem obstáculos, sofrendo uma perturbação inicial aleatória. Os episódios foram encerrados se a robô perdesse estabilidade ou após 20 segundos.

A estrutura de rede utilizada para o treinamento das políticas consistiu em duas camadas ocultas de 512 neurônios cada, com ativação *Relu*, seguidas por uma camada de saída linear com 1 unidade. Os hiperparâmetros utilizados no algoritmo SAC foram: taxa de aprendizagem (α) de 0.0003, tamanho de lote de 256, tamanho de *buffer* de 50000, fator de desconto (γ) de 0.99 e coeficiente de atualização (τ) de 0.005.

Os experimentos foram conduzidos para avaliar o desempenho do sistema na tarefa de chute de bola, utilizando o vídeo *6freekick* (1.24s) do SoccerKicks como referência. Em cada experimento, diferentes políticas e funções de recompensa foram avaliadas. A bola foi posicionada a aproximadamente 0.3m do pé direito da robô. Para corrigir a orientação do jogador em relação a bola, uma rotação de correção foi aplicada aos quadros do vídeo de referência. Adicionalmente, A posição da pélvis do jogador foi normalizada para o intervalo de 0 a 0.5m (primeiro e último quadro). Esse processo é ilustrado na Figura 4.

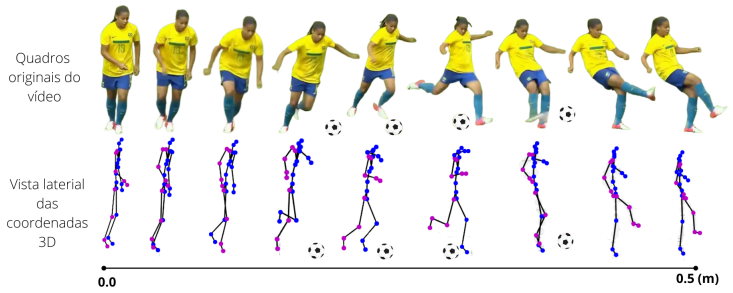


Figura 4. Imagens do vídeo selecionado do conjunto de dados SoccerKicks e suas coordenadas 3D correspondentes geradas.

Foram conduzidos quatro experimentos:

1. **EXP-01:** imitação de ângulos usando torque real. Neste experimento, o objetivo foi reproduzir os ângulos das articulações do especialista. Os motores no ambiente de simulação foram configurados com os torques máximos informados pelos fabricantes. A cinemática inversa foi utilizada;

2. **EXP-02:** imitação de ângulos usando torque ilimitado. Esse experimento foi idêntico ao anterior, porém utilizando motores sem limite máximo de torque;
3. **EXP-03:** imitação de posição com torque ilimitado e recompensa por pose. Neste experimento, o objetivo foi reproduzir a posição espacial relativa das articulações. O sistema foi recompensado pela diferença entre as posições das juntas (excluindo as juntas do peito, pélvis, pescoço e cabeça) na referência e a posição atual;
4. **EXP-04:** imitação de posição com torque ilimitado e recompensa de todas as juntas. Este experimento foi uma replicação do experimento anterior, excluindo apenas as articulações do pescoço e cabeça da recompensa.

7. Resultados e Discussões

Ao incentivar as políticas a imitar os movimentos ricos e complexos dos seres humanos a partir da observação de vídeos do conjunto de dados SoccerKicks, os resultados obtidos confirmam a viabilidade da transferência desses movimentos para robôs por meio da imitação por observação. A Figura 5 mostra registros das execuções das políticas aprendidas, evidenciando que o robô adquiriu posturas semelhantes às da referência, adaptando-se às suas próprias condições físicas e dinâmicas, mesmo quando a tarefa adicional de chutar a bola não foi alcançada (EXP-01, EXP-02).

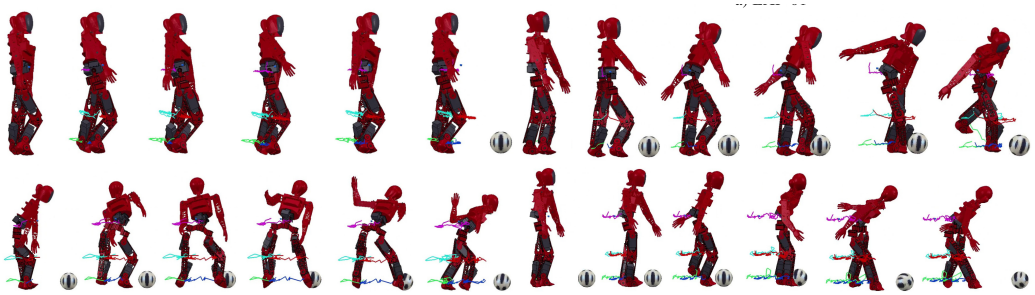


Figura 5. Fotografias do robô executando as políticas aprendidas para os quatro experimentos (EXP-01 a EXP-04). Com as abordagens adotadas, o robô foi capaz de obter posturas similares a da referência enquanto se atentava a chutar a bola, satisfatoriamente alcançados nos EXP-03 e EXP-04.

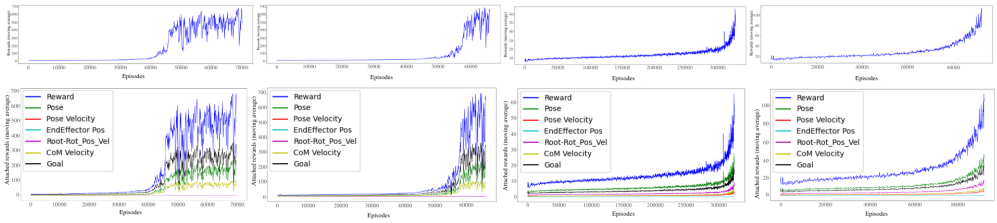
A Figura 6 apresenta as curvas de aprendizado para as diferentes configurações adotadas nos experimentos. A performance das políticas treinadas é resumida na Tabela 1. A avaliação das políticas foi realizada em 50 execuções, e os resultados referentes à estabilidade do tronco, à distância do pé de chute para a bola e à distância percorrida pela bola são apresentados nos gráficos das Figuras 7 (EXP-01, EXP-02, EXP-03 e EXP-04), respectivamente.

As curvas de aprendizado mostram a média total de recompensas recebidas a cada 100 episódios para cada experimento. As Figuras a, b, c e d (Superior) representam as curvas individuais do total médio recebido durante o treinamento. Por outro lado, as Figuras a, b, c e d (Inferior) exibem as curvas de contribuição para cada tipo de reforço, conforme descrito na Seção 5.3.

O método Procedimento -1 utilizado no treinamento das políticas π_{kick1} e π_{kick2}, apresentou resultados inferiores aos esperado, visto que este método tenta encaixar o movimento humano no corpo da Marta não levando em conta as diferenças estruturais e

Tabela 1. Sumário dos resultados obtidos das políticas aprendidas para realizar a transferência de movimentos humanos para o robô com a Aprendizagem por Reforço Profundo.

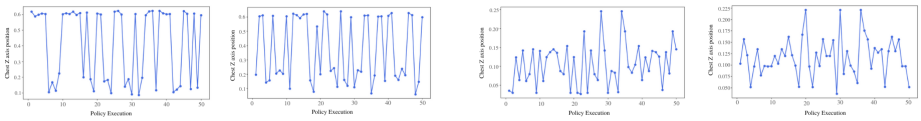
Experimentos	Ciclo de tempo (s)	Escala de tempo de simulação (s)	Episódios (10^3)	Recompensa máxima (10^3)	Média da Recompensa Máxima (10^1)	Em pé, após o treinamento (%)	Distância média viajada pela bola em (m)
π_{kick1}	1.24	0.04	70.3	1.04	67.7	58%	0
π_{kick2}	1.24	0.04	66.16	1.02	70	43%	0
π_{kick3}	1.24	0.1	328.08	1.23	6.5	4%	0.57
π_{kick4}	1.24	0.05	92.60	0.29	11.3	6%	1.02



a) EXP-01 b) EXP-02 c) EXP-03 d) EXP-04

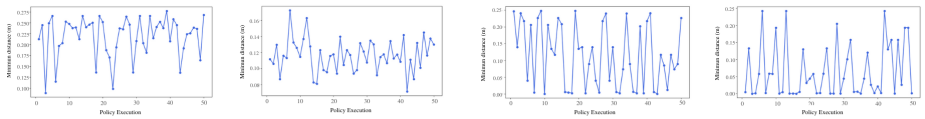
Figura 6. Curvas de aprendizado da média total de recompensas recebidas a cada 100 episódios foram plotadas para cada experimento. As Figuras a,b,c e d (Superior) representam as curvas individuais do total médio recebido ao longo treinamento. Já as Figuras a, b, c, e d (Inferior) são as curvas de contribuição para cada tipo de reforço (Descrito na Seção 5.3).

1. Posição final do peito no eixo-z.



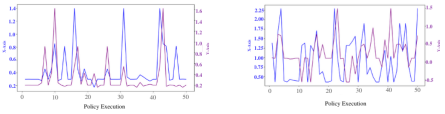
a) EXP-01 b) EXP-02 c) EXP-03 d) EXP-04

2. Mínima distância entre o pé direito da perna de chute para a posição da bola.



e) EXP-01 f) EXP-02 g) EXP-03 h) EXP-04

3. Distância percorrida pela bola após o chute.



i) EXP-03 j) EXP-04

Figura 7. Gráficos para a avaliação das políticas treinadas para os experimentos EXP-01, EXP-02, EXP-03 E EXP-04.

dinâmicas em relação a referência. Como resultado, o robô se limitou a reproduzir os movimentos que lhe trouxessem maior estabilidade levando a uma menor exploração da trajetória. Por outro lado, ao utilizar o segundo método, Procedimento -2, para treinar as políticas π_{kick3} e π_{kick4}, obtivemos resultados promissores. A Marta teve maior li-

berdade para movimentar suas juntas de forma a alcançar as coordenadas reescaladas da referência. Conclui-se, portanto, que a estratégia de coordenadas reescaladas foi vantajosa para lidar melhor com as diferenças corporais, especialmente em relação à altura, entre a Marta e a referência humana. Além disso, nas políticas π_{kick3} e π_{kick4}, a Marta foi capaz de realizar com sucesso a tarefa adicional de chutar a bola, percorrendo em média uma distância de 0,57m e 1,02m, respectivamente. Assim, os resultados demonstram que, ao utilizar a Aprendizagem por Imitação a partir da Observação de informações extraídas do vídeo, a Marta conseguiu reproduzir movimentos humanos semelhantes aos da referência e executar a tarefa adicional de chutar a bola, mesmo que não seguindo a trajetória da referência em todos os treinamentos.

8. Conclusões e Trabalhos Futuros

Apresentamos um framework baseado em Aprendizagem por Reforço Profundo e investigamos metodologias para transferir movimentos humanos para o robô Marta usando vídeos monoculares. Nossos resultados demonstraram uma reprodução satisfatória, em certa medida, dos movimentos humanos no robô Marta, assim como a realização de uma tarefa adicional nos últimos dois treinamentos. As contribuições deste trabalho são significativas e trazem perspectivas promissoras para a adaptação a novas tarefas. Embora os resultados obtidos sejam auspiciosos, identificamos limitações que precisam ser abordadas em futuras pesquisas. Primeiramente, a Marta foi incapaz de reproduzir completamente o ciclo de movimento de referência devido à duração limitada do vídeo em comparação com o tempo prolongado do episódio. Uma possível solução seria a introdução de quadros interpolados para ampliar o ciclo de tempo do vídeo e ajustar a duração do episódio de acordo com o vídeo. Além disso, o método utilizado para lidar com as diferenças corporais entre humanos e robôs requer estudos adicionais para abranger a diversidade morfológica dos robôs. No desenvolvimento deste trabalho, a engenharia da recompensa de imitação foi baseada em métricas de similaridade de estado definidas manualmente, assim como os pesos na função de reforço exigiram ajustes manuais. É importante ressaltar que nossas políticas foram projetadas para um ambiente personalizado para o robô Marta, o que demandaria adaptações para outros modelos robóticos. Como perspectivas futuras, propomos o refinamento das políticas existentes para corrigir as limitações atuais e a incorporação de uma maior diversidade de habilidades e tarefas. Isso permitiria que o robô executasse atividades desafiadoras e interações complexas com o ambiente, utilizando vídeos de humanos como referência. Essas contribuições pioneiras e o caráter inovador deste trabalho colocam-nos como fortes candidatos ao prêmio de melhor dissertação de mestrado em robótica.

Referências

Arulkumaran, K., Deisenroth, M. P., Brundage, M., and Bharath, A. A. (2017). Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6):26–38.

Benbrahim, H. and Franklin, J. A. (1997). Biped dynamic walking using reinforcement learning. *Robotics and Autonomous Systems*, 22(3-4):283–302.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2019). Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186.

Chenatti, S. F., Previato, G., Tomazela, R., Kopp, V. G., Begazo, M. F. T., Salaro, L. G., Rohmer, E., Colombini, E. L., and da Silva Simoes, A. (2018). Larocs+ unesp team description paper for the ieee humanoid racing 2018. *Latin American Robotics Competition - IEEE Humanoid Racing*.

Cheng, Z., Liu, L., Liu, A., Sun, H., Fang, M., and Tao, D. (2021). On the guaranteed almost equivalence between imitation learning from observation and demonstration. *IEEE Transactions on Neural Networks and Learning Systems*.

Dong, H., Ding, Z., and Zhang, S. (2020). *Deep Reinforcement Learning: Fundamentals, Research and Applications*. Springer Nature.

Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., et al. (2018). Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*.

Hudson, E., Warnell, G., Torabi, F., and Stone, P. (2021). Skeletal feature compensation for imitation learning with embodiment mismatch. *arXiv preprint arXiv:2104.07810*.

Kanazawa, A., Zhang, J. Y., Felsen, P., and Malik, J. (2019). Learning 3d human dynamics from video. In *Computer Vision and Pattern Recognition (CVPR)*.

Kim, D., Lee, J., and Sentis, L. (2017). Robust dynamic locomotion via reinforcement learning and novel whole body controller. *arXiv preprint arXiv:1708.02205*.

Lessa, N. M., Colombini, E. L., and Da Silva Simões, A. (2021a). Soccerkicks: a dataset of 3d dead ball kicks reference movements for humanoid robots. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3472–3478.

Lessa, N. M., Colombini, E. L., and Simões, A. D. S. (2021b). Soccerkicks: a dataset of 3d dead ball kicks reference movements for humanoid robots. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3472–3478. IEEE.

Peng, X. B., Kanazawa, A., Malik, J., Abbeel, P., and Levine, S. (2018). Sfv: Reinforcement learning of physical skills from videos. *ACM Transactions on Graphics (TOG)*, 37(6):1–14.

Peng, X. B., Ma, Z., Abbeel, P., Levine, S., and Kanazawa, A. (2021). Amp: Adversarial motion priors for stylized physics-based character control. *ACM Trans. Graph.*, 40(4).

Soares, Y. C. P. et al. (2020). *Deep reinforcement learning for bipedal locomotion: Aprendizado por reforço profundo para locomoção bípede*. PhD thesis, Universidade Estadual de Campinas, Instituto de Computação.

Tejada Begazo, M. F. (2020). A learning-based model-free controller for decoupled humanoid robot walking.

Tomazela, R. M. (2019). A combined model-based planning and model-free reinforcement learning approach for biped locomotion: Uma abordagem combinada de planejamento baseado em modelo e aprendizado por reforço para locomoção bípede. Master's thesis, Universidade Estadual de Campinas, Instituto de Computação, Campinas,SP.

Xiu, Y., Li, J., Wang, H., Fang, Y., and Lu, C. (2018). Pose Flow: Efficient online pose tracking. In *BMVC*.