نشریه مهندسی مکانیک ایران، دوره ۲۶، شماره ۳، شماره پیاپی۷۶، پاییز ۱۴۰۳، صفحه ۴۴-۲۹
انجمن مهندسان مکانیک ایران
مقاله علمی پژوهشی

DOI: https://doi.org/10.30506/ijmep.2024.558221.1893

کنترل راه رفتن ربات انسان نمای با پنجه فعال به کمک یادگیری تقویتی

عارف توانگر[۱]
کارشناسی ارشد

مجید ساده‌دل[۲]
استادیار

کاربرد کنترل کننده های مبتنی بر هوش مصنوعی در رباتیک نتایج درخشانی را به دنبال داشته است. از بین روش های مبتنی بر هوش مصنوعی، روش های یادگیری تقویتی بیشترین سهم استفاده را به خود اختصاص داده اند. با وجود مزایای وجود پنجه فعال، ربات های انسان‌نمای بسیار پیشرفته بدون پنجه فعال به کمک یادگیری تقویتی کنترل شده‌ا ند. در این پژوهش دو الگوریتم DDPG و TD3 بر روی ربات انسان نمای با پنجه به کار گرفته شده و با یکدیگر مقایسه شده اند. کارایی چهارچوب طراحی شده در کنترل ربات انسان نمای با پنجه، به کمک شبیه سازی سنجیده شده و ربات انسان نما توانسته با سرعت ۰/۹ متر بر ثانیه مسیر صاف را طی نماید.

واژه‌های راهنما: ربات انسان نما، مفصل پنجه فعال، یادگیری راه رفتن، یادگیری تقویتی عمیق

۱- مقدمه

ربات‌های انسان‌نما از اهمیت ویژه‌ای در حوزه ربات‌های متحرک برخوردارند. ربات‌های دوپا اغلب مزیت‌هایی را نسبت به ربات‌های چندپایی دارند. ربات‌های دوپا انطباق‌پذیری قابل‌توجهی دارند که آن‌ها را قادر می‌سازد به راحتی موانعی مانند گذرگاه‌های باریک و پله‌ها را پشت سر بگذارند. این انطباق‌پذیری به ویژه زمانی اهمیت پیدا می‌کند که ربات قرار است وظایف انسان-محور را به عهده بگیرد. پیشرفت‌ها در حوزه‌ی ربات‌های انسان‌نما می‌تواند نوید ساخت پاهای مصنوعی با قابلیت راه رفتن طبیعی برای افراد معلول، تسهیل کمک به انسان در انجام وظایف و همچنین جایگزینی انسان در مواردی که خطرناک است را بدهد. اضافه کردن پنجه به ربات انسان‌نما پیچیدگی مسئله را دوچندان می‌کند. اضافه کردن پنجه درجات آزادی سیستم را افزایش می‌دهد و می‌تواند تعداد فاز و سوییچ‌های بین آن‌ها را نیز تغییر دهد. اضافه کردن مفصل پنجه به سادگی اضافه کردن درجات آزادی در لینک‌های بالاتر از خود نیست زیرا مستقیماً پایداری ربات را با مشکل روبرو می‌کند.

[۱] کارشناسی ارشد، دانشکده مهندسی مکانیک، دانشگاه تربیت مدرس، تهران، جمهوری اسلامی ایران،
aref.tavangar@modares.ac.ir
[۲] نویسنده مسئول، استادیار، دانشکده مهندسی مکانیک، دانشگاه تربیت مدرس، تهران، جمهوری اسلامی ایران،
majid.sadedel@modares.ac.ir

تاریخ دریافت: ۱۴۰۱/۰۴/۰۳، تاریخ بازنگری: ۱۴۰۲/۰۸/۲۵، تاریخ پذیرش: ۱۴۰۳/۰۴/۱۳

با این‌حال در صورتی که به خوبی کنترل شود می‌تواند به نرمی راه رفتن ربات کمک کند و امکان مانورهای بیشتری را برای ربات فراهم آورد.

مهم‌ترین حرکت در ربات‌های انسان‌نما، قابلیت راه رفتن ربـات شبیه به انسان است. به همین دلیل کنترل حرکت ربات انسان‌نما به منظور راه رفتن از اهمیت ویژه‌ای برخوردار است. کنترل راه رفتن ربات‌های انسان‌نما به دلیل ماهیت غیرخطی، چندفازی بودن و تعداد درجات آزادی بالا بسیار پیچیده است. ربات انسان‌نما در حین راه رفتن یک سیستم چندورودی-چندخروجی است که الزاماً تعداد درجات آزادی آن با تعداد عملگرها برابر نیست. مهم‌ترین مسئله در کنترل حرکت ربات‌های انسان‌نما پایداری است. از آنجاییکه ربات انسان‌نما می‌تواند عدم قطعیت‌های متعددی در پارامترهای خود، مشاهده و کنترل داشته باشد، روش‌های مستقل‌ازمدل گزینه‌های خوبی برای استفاده هستند.

یادگیری تقویتی در کنترل به عنوان یک روش مستقل از مدل محسوب می‌شود. یادگیری تقویتی زیرمجموعه‌ای از روش‌های یادگیری ماشین است. الگوریتم‌های یادگیری به سه دسته اصلی یادگیری با ناظر[1]، یادگیری تقویتی[2] و یادگیری بدون ناظر[3] دسته‌بندی می‌شوند. یکی از تفاوت‌های اساسی این سه الگوریتم در میزان اطلاعات و اطمینان از صحت اطلاعات است [۱]. یادگیری نظارت‌نشده نیازی به اطلاعات درست و غلط ندارد و برای خوشه‌بندی کاربرد دارد. یادگیری با ناظر کاملاً اطلاعات تاییدشده‌ای را در دست دارد و می‌تواند خوب یا بد بودن هر عملی را بسنجد. این در حالی است که تنها مؤلفه‌ای که عامل یادگیری تقویتی در اختیار دارد پاداش است.

Russel [۲] با استفاده از یادگیری تقویتی ربات انسان‌نما را برای حرکت روی سطح شیب‌دار و رو به بالا کنترل کرد. علاوه بر آن نشان داد که ربات پس از یادگیری می‌تواند از عامل تربیت‌شده خود حتی در سطح شیب‌دار با شیب‌های متغیر نیز بهره ببرد. Wu و همکاران [۳] الگوریتم Q-learning را به نحوی به کار گرفتند که ربات بدون هیچ‌گونه اطلاعی درباره مدل دینامیکی راه برود. کنترل تعادل به این صورت است که نقطه گشتاور صفر[4] را به کمک یادگیری به نواحی پایدار منتقل می‌کند. روش پیشنهادشده می‌تواند بر روی هر راه‌رونده دوپا و روی هر صفحه‌ای اعم از شیب‌دار و صاف به کار گرفته شود. به علاوه از یادگیری تقویتی برای کنترل مکان ربات نیز استفاده شده است. نشان داده شده است که اگر ربات به اندازه کافی تحت آموزش قرار بگیرد می‌تواند صفحات شیب‌دار را سریع‌تر طی نماید. با استفاده از الگوریتم‌های Q-Learning عمیق می‌توان مسئله‌های در فضای پیوسته و دارای ابعاد بالا را حل کرد. Lilicrap و همکاران [۴] با معرفی الگوریتم گرادیان سیاست قطعی عمیق[5] را مدلی رل طراحی کردند که برای فضاهای عمل پیوسته نیز کاربرد دارد. Fujimoto و همکاران [۵] الگوریتم گرادیان سیاست قطعی عمیق با تاخیر[6] که به نوعی ارتقاء یافته‌ی الگوریتم گرادیان سیاست قطعی عمیق است را معرفی کردند. این عامل مستقل از مدل، برخط و سیاست-خاموش[7] است. از جمله تفاوت‌های

[1] Supervised Learning (SL)
[2] Reinforcement Learning (RL)
[3] Unsupervised Learning (UL)
[4] Zero Moment Point (ZMP)
[5] Deep Deterministic Policy Gradient (DDPG)
[6] Twin-Delayed Deep Deterministic Policy Gradient (TD3)
[7] Off-Policy

مهم این الگوریتم با گرادیان سیاست عمیق در استفاده‌ی از دو تابع ارزش-Q، به‌روزرسانی دیر به دیر و اضافه کردن نویز به کنش است. مورد آخر برای دور شدن از انتخاب کنشی که ارزش-Qی آن بیشترین باشد و فایده آن جست‌وجوی بیشتر است. Garcia و همکاران[۶] به کمک الگوریتم یادگیری تقویتی عمیق امن[1] الگوریتم کنترلی طراحی کردند که به کمک آن ربات قادر است در محیط‌های ناشناخته در انجام وظیفه‌ی راه رفتن بهتر عمل کند. Melo و همکاران[۷] یک روش حل مسئله را بر مبنای یادگیری تقویتی عمیق و بهینه‌سازی سیاست تقریبی[2] را معرفی کردند که بر اساس آن ربات انسان‌نما می‌توانست بدون هیچ‌گونه داده‌ی پیشینی از محیط بتواند در آن راه برود.

ساده‌دل و همکاران[۸-۱۰] پارامترهای راه رفتن ربات انسان‌نمای با پنجه فعال و غیرفعال را بر اساس توابع هدف مختلف بهینه‌سازی کردند و اثر آن‌ها بر پایداری و مصرف انرژی را بررسی کردند. Weiler و Dorer[۱۱] نشان دادند ربات با پنجه فعال در انجام وظایفی مانند شوت زدن از ربات بدون پنجه بهتر عمل می‌کند. ربات در یک آزمون گل‌زنی به‌طور متوسط می‌تواند نرخ شوت‌زنی ۰/۶۵۶ را به دست آورد که از نرخ شوت‌زنی ربات بدون پنجه که برابر است با ۰/۳۸۵ بهتر است. Fischer و Dorer[۱۲] به منظور انجام وظیفه‌ی راه رفتن توسط ربات انسان‌نمای با پنجه از الگوریتم کنترلی مبتنی بر الگوریتم ژنتیک استفاده کردند و توانستند ربات نائو[3] را با سرعت متوسط ۱/۳ متر بر ثانیه در مسیر راست به حرکت درآورند. Duburcq و همکاران[۱۳] از یادگیری تقویتی عمیق به منظور افزایش پایداری حرکات ربات انسان‌نما در مقابل اغتشاشات بهره بردند. ربات به کمک یک الگوریتم مبتنی بر شکل دادن پاداش ربات انسان‌نما در محیط واقعی توانست در مقابل تحرکات خارجی مقاومت نشان دهد و به وضعیت پایدار بازگردد. kim و همکاران[۱۴] برای وظیفه‌ی بازیابی مسیر پس از هل دادن ربات انسان‌نما با استفاده از یادگیری تقویتی مسیری برای لگن، زانو و گام‌ها طراحی کردند. تمامی مفاصل در طراحی کنترلی روش کیم و همکاران فعال در نظر گرفته شده است. bijing و همکاران[۱۵] برای طراحی ربات همیار انسان در فرآیند راه رفتن از ترکیبی از یادگیری تقویتی برای مفاصل فعال پایین‌تنه و ترکیبی از مفاصل غیرفعال استفاده کردند.

در تمامی پژوهش‌های ذکر شده ربات‌های پیچیده با درجات آزادی بالا مورد آزمایش قرار گرفته‌اند. اما راه رفتن ربات انسان‌نما به کمک یادگیری تقویتی در حالتی که ربات دارای پنجه‌ی فعال باشد بررسی نشده است. در این پژوهش این مورد تحت آزمایش قرار گرفته است.

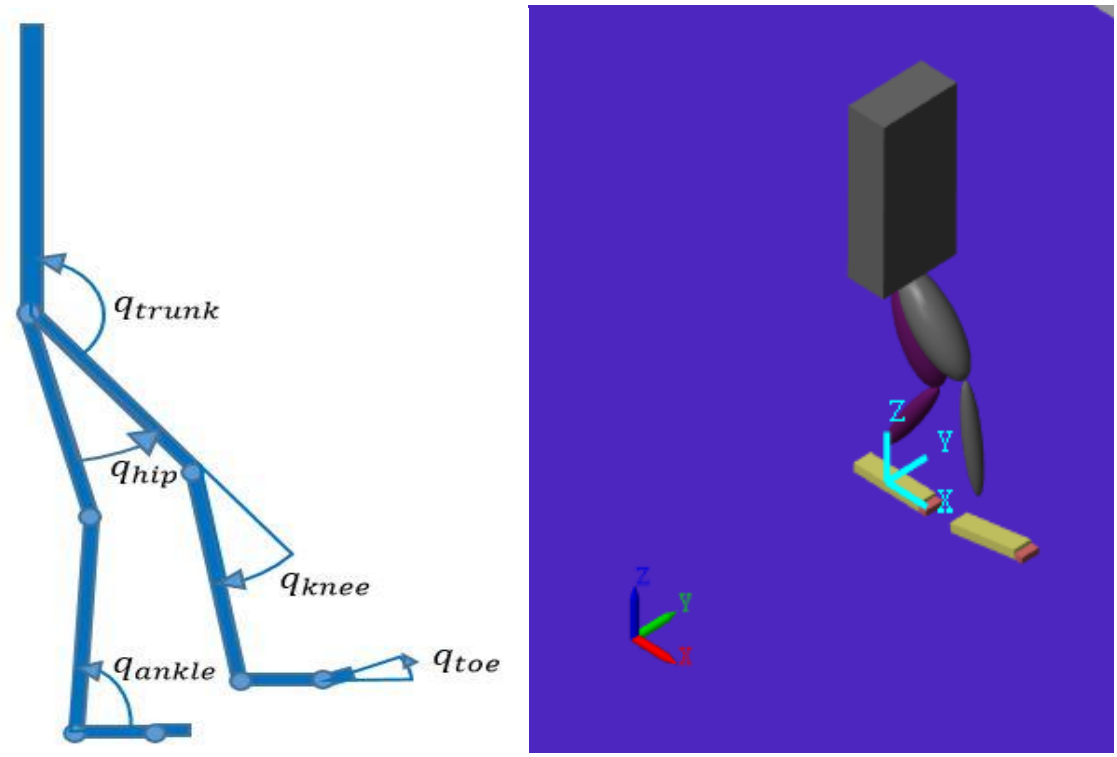
۲- مدل ربات

ساده‌دل و همکاران [۸] یک ربات انسان‌نمای دو بعدی با پنجه‌ی فعال را شبیه‌سازی و کنترل کردند. مدل کنترل‌شده یک ربات انسان‌نمای دوبعدی با هشت درجه آزادی (دو عدد در پنجه، دو عدد در مچ پا، دو عدد در زانو و دو درجه آزادی در کمر) بود. بر اساس مدل ذکرشده می‌توان یک قاب مرجع متحرک نسبت به قاب اینرسی تعریف کرد. جابه‌جایی و دوران ربات بر اساس جابه‌جایی قاب ربات تعریف می‌شود. مرکز قاب ربات در مفصل کمر قرار دارد. زمانی که تمام بازوها (به جز پا و پنجه) عمود بر زمین قرار دارند محورهای قاب ربات

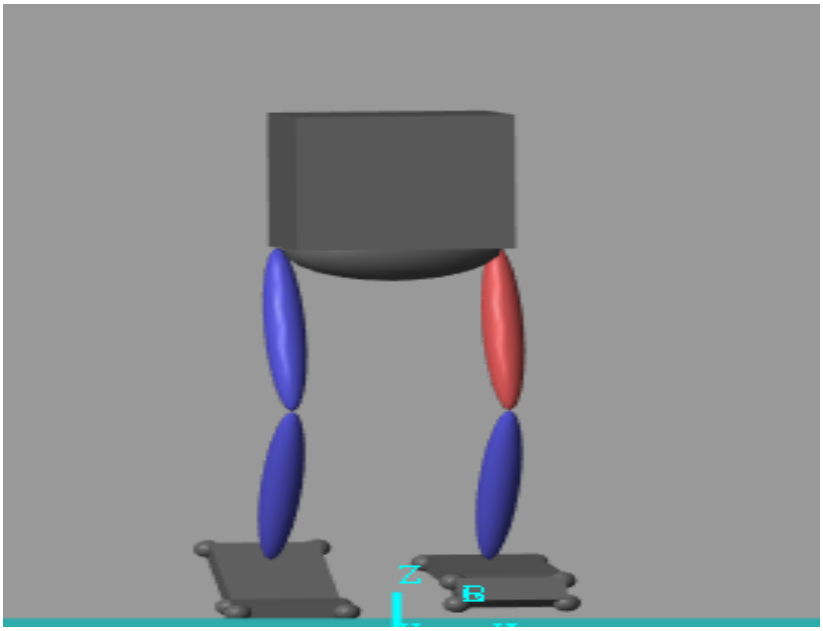
[1] Safe Reinforcement Learning
[2] Proximal Policy Optimization (PPO)
[3] Nao

موازی محورهای قاب اینرسی است. مدل ربات انسان‌نمای دوبعدی و راستای محورهای قاب اینرسی در شکل (۱) نشان داده شده است.

در این پژوهش از یک مدل اصلاح‌شده از مدل ساده‌دل و همکاران استفاده می‌شود که تنها تفاوت آن در ایجاد یک فاصله بین نقطه‌ی اتصال لینک‌های ران چپ و راست به لگن است. این مدل در شکل (۲) نشان داده شده است.



شکل ۱– مدل ربات انسان نما



شکل ۲– مدل اصلاح‌شده

به کمک مدل اصلاح‌شده می‌توان راه رفتن در فضای سه‌بعدی را نیز شبیه‌سازی کرد.
طول لینک‌های مدل، جرم لینک‌های مدل و ممان اینرسی لینک‌های مدل به ترتیب در جدول (۱) تا (۳) نشان داده شده‌اند.

جدول ۱- طول لینک‌های مدل

نام پارامتر	توضیحات	طول (cm)
L_{tr}	طول بالاتنه	8.5
L_{th}	طول ران	9.5
L_{sh}	طول ساق	9.5
L_{an}	فاصله‌ی مچ تا کف پا	1
L_{ab}	فاصله‌ی کف پا تا پاشنه	1
L_{af}	فاصله‌ی پاشنه پا تا پنجه	5
L_{ft}	طول پنجه	2

جدول ۲- جرم لینک‌های مدل

نام پارامتر	توضیحات	جرم (g)
M_{tr}	جرم بالاتنه	150
M_{th}	جرم ران	20
M_{sh}	جرم ساق	20
M_f	جرم کف پا	20
M_t	جرم پنجه	60

جدول ۳- ممان اینرسی لینک‌های مدل

نام پارامتر	توضیحات	ممان اینرسی $(N.m^2 \times 10^{-6})$
J_{tr}	ممان اینرسی بالاتنه	11.5
J_{th}	ممان اینرسی ران	7.5
J_{sh}	ممان اینرسی ساق	7.5
J_f	ممان اینرسی کف پا	4.5
J_t	ممان اینرسی پنجه	1.6

به منظور مدلسازی تماس پا با زمین مدل‌های مختلفی معرفی شده است. مدل کردن تماس کف پا با زمین پیچیده است. احسانی‌سرشت و مقدم، الزاماتی را برای مدل‌سازی تماس پا با زمین که در راه رفتن، پریدن و دویدن معتبر باشند، بیان کرده‌اند [۱۶]. مانند اینکه نیروی عکس‌العمل عمودی همواره مثبت باشد، برخورد سریع رخ دهد، شامل برخورد پلاستیک باشد. بر همین اساس از مدل شورجه و McPhee که مدل‌سازی نیروهای سطح را با استفاده از مدل‌های فرا-حجمی[1] انجام می‌دهد، استفاده می‌شود. معادلات (۱) و (۲) به ترتیب برای نیروی نرمال و نیروی مماسی نتیجه گرفته می‌شود [۱۷].

$$f_n = \left(k_h V^{\mathcal{H}} + a_h V_{V_{cn}}\right)\hat{n} \quad (۱)$$

$$f_f = -\mu(v_{ct}) f_n, \qquad \mu(v_{ct}) = \mu_f arctan\left(\frac{v_{ct}}{v_s}\right) \quad (۲)$$

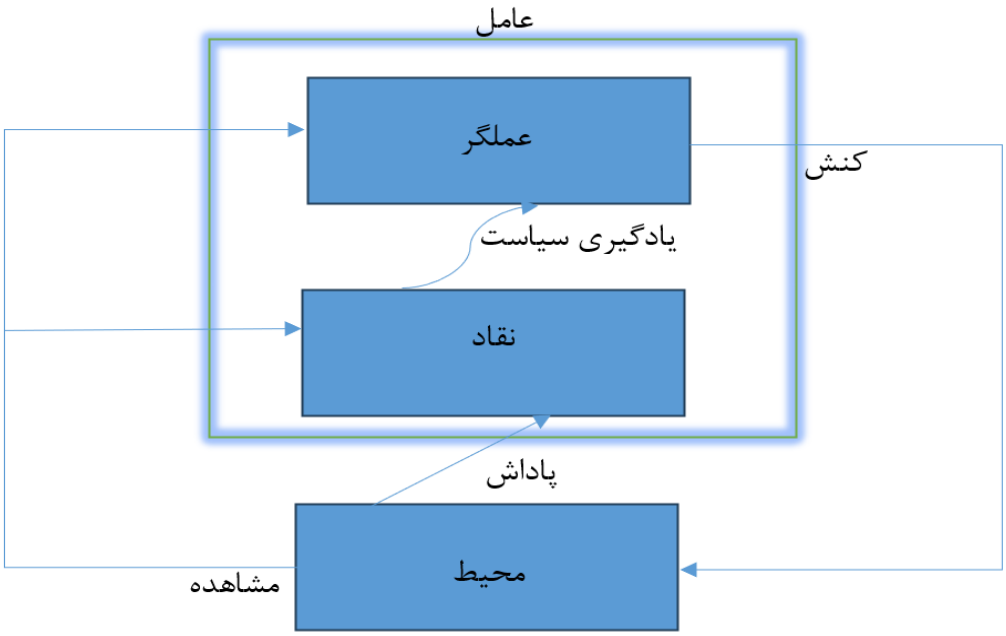
که k_h شبه‌نرمی حجمی غیرخطی[2] نامیده می‌شود. $\mathcal{H}$ به نرمی حجمی و مشخصه‌های هندسی بستگی دارد. a_h نرمی زمین ضرب در دمپینگ زمین است. V حجم فرورفتگی در زمین، v_{cn} سرعت نرمال مرکز جرم فرورفتگی در زمین و v_{ct} سرعت مرکز جرم فرورفتگی در زمین است. v_s ضریب شکل و μ_f ضریب مجانبی اصطکاک است. در نهایت شورجه و مک‌آفی نشان دادند که استفاده از مدل فراحجمی نسبت به برخورد نقطه‌ای[3] و حجمی خطی[4] به ترتیب ۷۵ و ۶۲ درصد تطابق بیشتری با نتایج عملی دارد.

۳- کنترل به کمک یادگیری تقویتی عمیق

عموماً روش‌های کنترل ربات انسان‌نما در دو دسته‌ی با مدل و بدون مدل دسته‌بندی می‌شود. روش‌های با مدل با افزایش پیچیدگی مدل، کارایی کمتری از خود بروز می‌دهند. راه رفتن ربات را باید با استفاده از چند دینامیک مختلف و سوییچ بین آن‌ها توصیف کرد. سوییچ کردن بین حالت‌های دینامیکی در حالت حضور و عدم حضور پنجه بسیار متفاوت است. به همین دلیل اضافه شدن پنجه از همه‌ی مفصل‌های دیگر در پیچیدگی توصیف دینامیک ربات تأثیرگذارتر است. در حالتی که ربات بدون پنجه باشد تغییر بین دینامیک‌های تک‌تکیه‌گاهی و دوتکیه‌گاهی در یک لحظه اتفاق می‌افتد. در حالی که با اضافه شدن پنجه، حالت ایستایی هر پا به دو قسمت تقسیم شده که در یک قسمت آن تنها پنجه‌ی پا با زمین در تماس است. این موجب می‌شود که ربات مدت‌زمان کمتری در حالت ناپایداری باشد و گام‌های بلندتر و سریعتری بردارد. از آن‌جاکه پنجه‌ی مساحت بسیار کمتری از کف پا دارد، جدایش و فرود را راحت‌تر می‌کند. تأثیر مثبت اضافه کردن پنجه به پای ربات همراه با افزایش پیچیدگی مدل‌سازی و کنترل است. یادگیری تقویتی یک روش بدون مدل است که پیچیدگی مدل مشکلات کمتری را برای الگوریتم (به نسبت روش‌های مدل-مبنا) وجود می‌آورد.

---

[1] Hyper-Volumetric
[2] Nonlinear Volumetric Pseudo-stiffness
[3] Point Contact
[4] Linear Volumetric



شکل ۳- الگوریتم

المان‌های اصلی یک مسئله‌ی یادگیری تقویتی عبارتند از حالت، سیاست، پاداش و کنش. به منظور تشخیص این المان‌ها برای مسئله و شروع پروسه‌ی یادگیری بایستی به ترتیب سیاست، کنش، حالت‌ها (از بین مشاهدات)، تابع پاداش و در نهایت شرایط اتمام دوره تعیین شود. ارتباط این المان‌ها در شکل (۳) نمایش داده شده است.

۳-۱- سیاست

سیاست بایستی پیش از مراحل دیگر تعیین شود زیرا انتخاب الگوریتم حل سایر المان‌ها را نیز تحت تأثیر قرار می‌دهد. در این پژوهش از دو الگوریتم کاربردی گرادیان سیاست قطعی عمیق و گرادیان سیاست قطعی عمیق با تاخیر جفتی به منظور حل راه رفتن ربات انسان‌نمای با پنجه استفاده شده است.

۳-۲- کنش‌ها

در مورد تعداد کنش عملاً بحثی وجود ندارد. به این علت که هشت مفصل فعال وجود دارد که بایستی کنترل شود.. اما برای نوع ورودی عملگر حالت‌های متفاوتی را می‌توان در نظر گرفت. یک حالت می‌تواند این باشد که زوایای مفاصل از عامل به دست بیاید و مفاصل کنترل زاویه شوند. در این حالت یک حلقه کنترلی بین خروجی عامل و موتورهای مفاصل قرار می‌گیرد. حالت دیگر این است که گشتاور مفاصل از خروجی عامل به دست بیاید. در این حالت نیاز به حلقه دومی نیست اما کنترل ربات به ورودی گشتاور حساس‌تر است و شبکه‌ها باید مدت زمان بیشتری را برای یادگیری صرف کنند. از بین دو روش معرفی شده روشی انتخاب می‌شود که خروجی‌ها در آن به صورت گشتاور انتخاب می‌شوند. دلیل این انتخاب این است که کل بار راه بردن یک ربات انسان‌نما بر دوش عامل یادگیری تقویتی تحمیل شود و هیچ نیازی به مدل ربات نباشد. خروجی‌ها نرمالیزه‌شده و بین -۱ تا ۱ است و برای خروجی هر گشتاور ضریبی در نظر گرفته می‌شود.

۳-۳- حالت

انتخاب حالت را از دو جنبه می‌توان بررسی کرد. اولاً اینکه در مسئله شرط مارکو[1] برقرار است. بنابراین اطلاعات هر گام برای گام بعدی کافی است. جنبه‌ی دوم انتخاب حالت‌ها، انتخاب با دید کنترلی است. ربات انسان‌نما از نظر دینامیکی یک سیستم هشت درجه آزادی است که در فضای دوبعدی به علاوه سه درجه آزادی و در فضای سه‌بعدی به علاوه‌ی شش درجه آزادی می‌شود. حالت‌ها باید به گونه‌ای انتخاب شوند که به کمک آن بتوان حالت مرحله بعد را پیش‌بینی کرد (فرض فضای مشاهده‌پذیر). برای پیش‌بینی حالت بعد بایستی هم حالت فعلی در دسترس باشد و هم ورودی فعلی. بنابراین حالت‌ها به شرح زیر خواهد بود:

- موقعیت مرکز مختصات محلی ربات در مختصات اینرسی
- زوایای دوران مختصات ربات نسبت به مختصات اینرسی
- زاویه‌ی مفاصل ربات
- سرعت زاویه‌ی مفاصل ربات
- کنش گام قبلی

۳-٤- شرط اتمام دوره

طبیعتاً با زمین خوردن بایستی دوره خاتمه پیدا کند. بنابراین شروط زیر هدف را ارضا می‌کند:

- سه زاویه مختصات محلی ربات نسبت به مختصات اینرسی از حد از پیش‌تعیین‌شده بگذرند.
- ارتفاع ربات از مقدار معینی کمتر شود.

علاوه بر زمین خوردن ربات، خروج از حرکت در راستای مستقیم نیز بایستی تعریف شود. بنابراین انحراف بیش از مقدار معین از خط مستقیم نیز جزو شروط اتمام دوره در نظر گرفته می‌شود.

۳-۵- تابع پاداش

تابع پاداش حاصل جمع مولفه‌هایی است که عامل را در اصلاح راه رفتن هدایت می‌کند. تابع پاداش از مولفه‌های زیر می‌تواند تشکیل شود:

پاداش مقدار سرعت در راستای افقی

R_1، ترم اول تابع پاداش، در رابطه (۳) نشان داده شده است.

$$R_1 = \dot{X} \qquad (۳)$$

X مولفه اول مختصات مرکز قاب ربات در قاب اینرسی و نشان‌دهنده حرکت در راستای افقی است. مقدار مثبت محور X به عنوان جهت مطلوب حرکت انتخاب می‌شود. $\dot{X}$ سرعت رو به جلو است. در حقیقت سرعت

[1] Markov condion

در راستای افقی هدفی است که راه رفتن بنا بوده آن را برآورده کند. بنابراین سرعت رو به جلو باید به صورت پاداش برای عامل در نظر گرفته شود.

جریمه برای انحراف جانبی

R_2، ترم دوم تابع پاداش، در رابطه (۴) نشان داده شده است.

$$R_2 = Y^2 \qquad (۴)$$

Y مقدار مؤلفه دوم مختصات مرکز قاب ربات در قاب اینرسی و نشان‌دهنده انحراف ربات از مسیر اصلی در راستای جانبی است.

جریمه ارتفاع نامطلوب ربات

R_3، ترم سوم تابع پاداش، در رابطه (۵) نشان داده شده است.

$$R_3 = f(Z, Z_i, Z_a) \qquad (۵)$$

Z مؤلفه سوم مختصات مرکز قاب ربات در قاب اینرسی و نشان‌دهنده فاصله از زمین است. Z_i میانگین درنظرگرفته‌شده برای جابه‌جایی مولفه Z مرکز قاب ربات است. Z_a شعاع قابل قبول برای فاصله مولفه Z مرکز قاب ربات از میانگین است. f تابع ناحیه مرده به مرکز Z_i و شعاع Z_a است. کاهش بیش از حد ارتفاع ربات به منزله نزدیک شدن به ناپایداری و افزایش بیش از حد آن نیز به منزله پرش است. به همین دلیل یک بازه برای ارتفاع ربات در نظر گرفته می‌شود.

پاداش مدت زمان دوام حرکت

R_4، ترم چهارم پاداش، در رابطه (۶) نشان داده شده است.

$$R_4 = \frac{t_s}{T_f} \qquad (۶)$$

t_s گام زمانی و T_f مدت زمان کل شبیه‌سازی است. از آنجاییکه زمین خوردن ربات یا کاهش از یک ارتفاع از قبل‌تعیین‌شده منجر به پایان دوره تمرین می‌شود مدت زمان باقی ماندن ربات در شبیه‌سازی یک ترم مطلوب در نظر گرفته می‌شود.

جریمه اندازه گشتاورهای وارد به مفاصل

R_5، ترم پنجم پاداش، در رابطه (۷) نشان داده شده است.

$$R_5 = \sum T_j\,, \qquad j = 1, \dots, n \qquad (۷)$$

T_j گشتاور هر مفصل و n تعداد مفاصل فعال ربات است. اگر برای گشتاور ربات مجازاتی در نظر گرفته نشود عامل آزادانه سعی می‌کند از گشتاوری که پاداش بیشتری فراهم می‌اورد، استفاده کند. این امر موجب می‌شود که ربات پاداش‌های کوتاه‌مدت را ترجیح داده و در هر گام از حداکثر گشتاور استفاده کند.

جریمه شتاب روبه‌جلوی ربات

R_6، ترم ششم پاداش، در رابطه (۸) نشان داده شده است.

$$R_6 = \ddot{X} \tag{۸}$$

که در آن $\ddot{X}$ شتاب رو به جلو ربات است. انسان در هنگام راه رفتن روی یخ سعی می‌کند شتاب رو به جلوی خود را کاهش دهد تا از ناپایداری جلوگیری کند. در نظر گرفتن این ترم به عنوان جریمه منجر به راه رفتن با شتاب کمتر می‌شود و می‌تواند یک ترم برای تابع پاداش باشد.

۴- شبیه‌سازی و نتایج

شبیه‌سازی راه رفتن ربات انسان‌نما به کمک نرم‌افزار سیمولینک انجام شده است. راه رفتن ربات انسان‌نمای با پنجه با هر دو عامل $DDPG$ و $TD3$ به ازای هر سه تابع پاداش معرفی‌شده در روابط (۹)، (۱۰) و (۱۱) شبیه‌سازی شده است. رابطه (۱۰) تابع پاداش در پژوهش Heess و همکاران [۱۸] معرفی شده است.

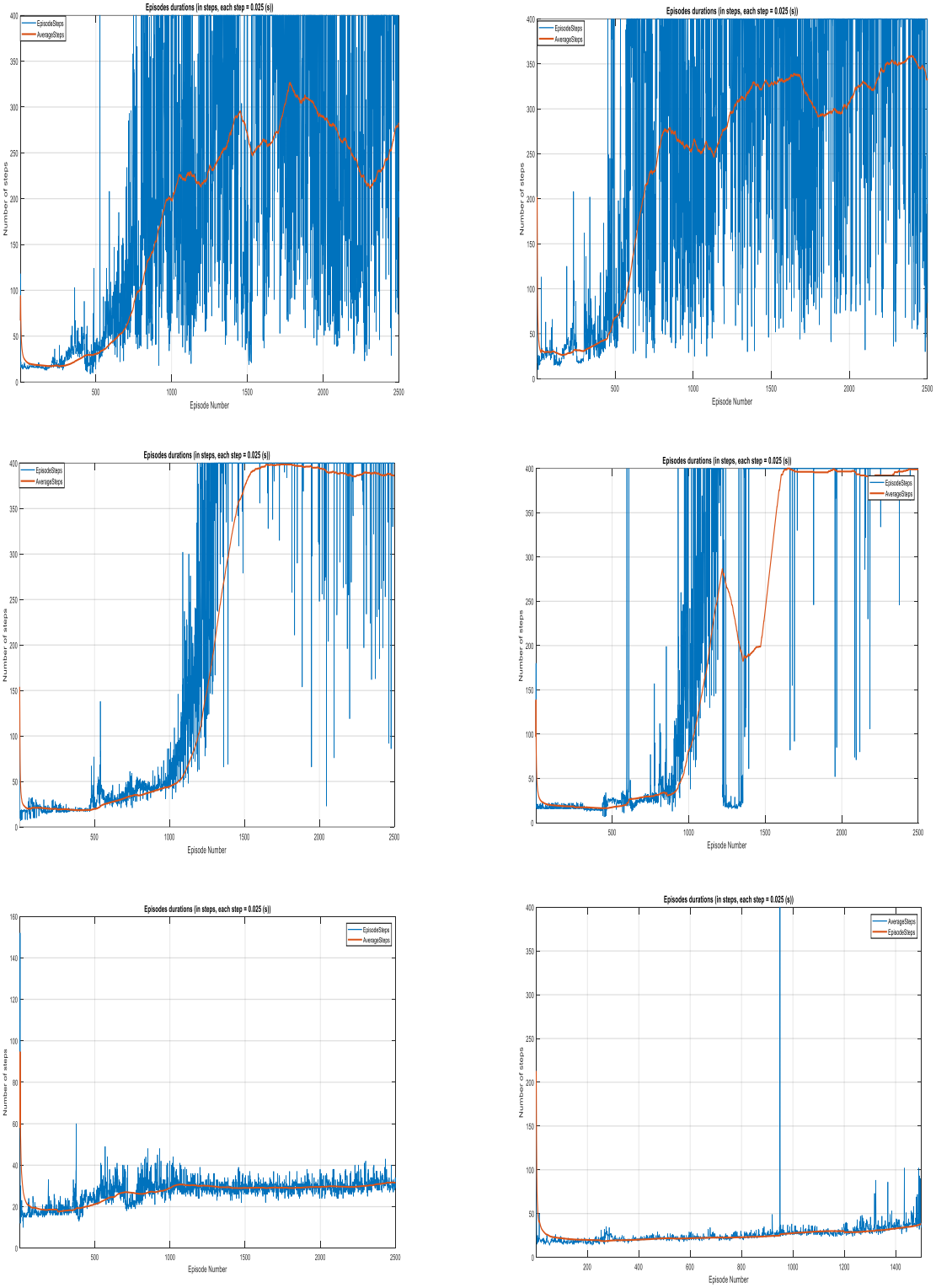
$$\bar{R}_1 = \sum_{i=1}^{4} R_i \tag{۹}$$

$$\bar{R}_2 = \sum_{i=1}^{5} R_i \tag{۱۰}$$

$$\bar{R}_3 = \sum_{i=1}^{6} R_i \tag{۱۱}$$

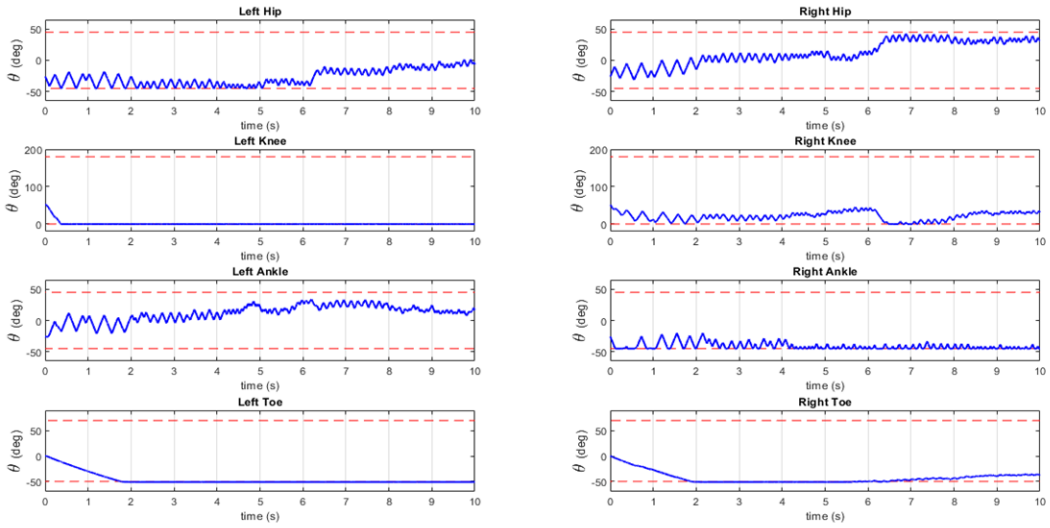
نمودارهای تعداد گام طی‌شده در هر دوره نشان‌دهنده مزایای روش‌های مختلف یادگیری بر یکدیگر است و در شکل (۴) نشان داده شده‌اند. زاویه و گشتاور ورودی مفاصل ربات انسان‌نما زمانی که تابع پاداش رابطه (۸) و عامل TD3 انتخاب می‌شود، به ترتیب در شکل (۵) و شکل (۶) نشان داده شده‌اند. تصاویر لحظه‌ای گام برداشتن ربات در این حالت در شکل (۶) نمایش داده شده است که در نهایت با سرعت ۰/۹ متر بر ثانیه مسیر ۲۵ متری را طی می‌کند. مورد قابل توجه در نتایج، استفاده‌ی ربات از مفصل پنجه است. با توجه به نمودار زاویه‌ی مفاصل که در شکل (۴) نشان داده شده است، مفصل پنجه نقش فعالی در راه رفتن ایفا کرده است. همانطور که در شکل (۷) نشان داده شده است، مفصل پنجه در لحظاتی تنها لینک در تماس مستقیم با زمین بوده و پایداری و راه رفتن ربات به کنترل درست این مفصل بستگی داشته است. بنابراین مفصل پنجه نه تنها

نقش مهمی در راه رفتن ایفا کرده است؛ بلکه در لحظاتی تنها عامل کنترل پایداری بوده است. علاوه بر این می‌توان مشاهده کرد که استفاده از پنجه راه رفتن ربات را خیلی به انسان شبیه‌تر کرده است و اثر پنجه در طبیعی بودن گام برداری ربات کاملاً مشهود است.

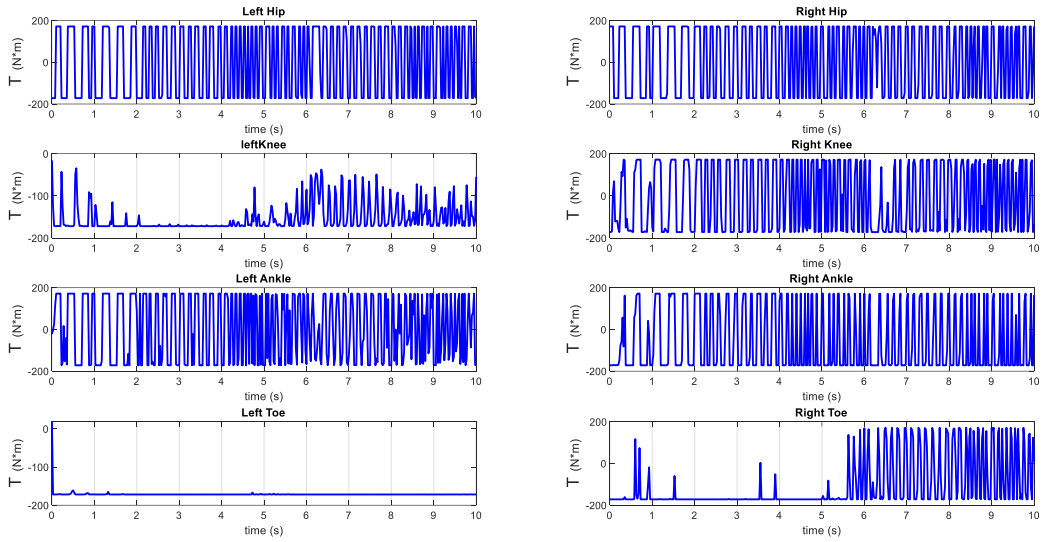


DDPG TD3

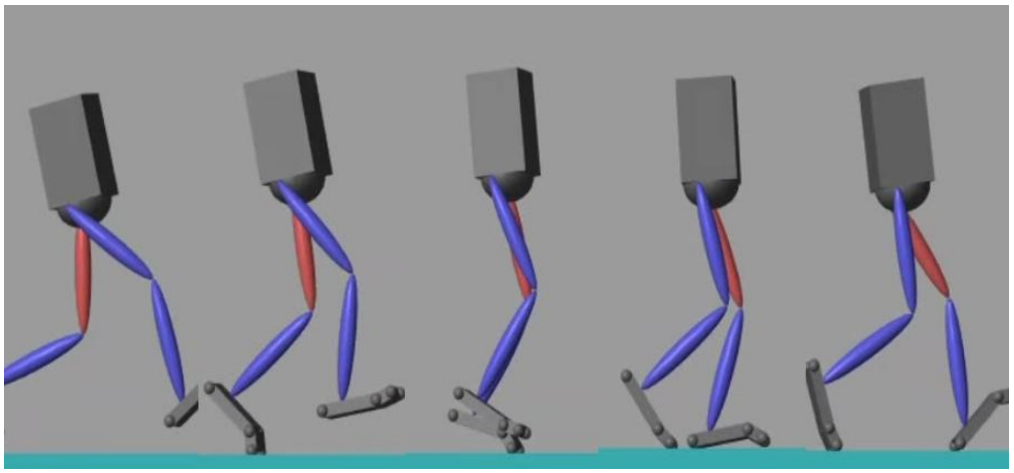
شکل ۴– تعداد گام طی‌شده در هر دوره از یادگیری (به ترتیب از بالا $\bar{R}_1$، $\bar{R}_2$ و $\bar{R}_3$)



شکل ۵– مقادیر زاویه مفاصل ربات انسان‌نما



شکل ۶– مقادیر گشتاور مفاصل ربات انسان‌نما



شکل ۷– تصاویر پی در پی گام برداشتن ربات

۵- نتیجه‌گیری و پیشنهادات

در این پژوهش ربات انسان‌نمای با پنجه فعال به کمک یادگیری تقویتی کنترل شد. با توجه به پاداشی که برای ربات در نظر گرفته شده بود علاوه بر اینکه قادر به انجام حرکت پایدار بود از حداقل گشتاور برای این حرکت نیز استفاده می‌کند. در مقایسه الگوریتم‌های DDPG و TD3 در آزمونی که از یک تابع پاداش استفاده شده است (بدین معنی که قابلیت مقایسه وجود دارد) الگوریتم TD3 به طور تقریبی بیش از ۱/۵ برابر DDPG پاداش بیشتر دریافت می‌کند. الگوریتم DDPG و تابع پاداش دارای مولفه‌ی هزینه‌ی گشتاور از الگوریتم TD3 و تابع هزینه‌ی بدون مولفه‌ی گشتاور بهتر عمل می‌کند. می‌توان گفت در این مثال انتخاب تابع پاداش بهتر به انتخاب الگوریتم یادگیری بهتر اولویت دارد. یک نتیجه اساسی این می‌تواند باشد که هرچه مولفه‌های پایداری کمتری به ربات تحمیل شود الزاماً منجر به حرکات بد و نامطلوب نمی‌شود. اتفاقاً این امکان وجود دارد که ربات متناسب با وضعیت راه بهتری را انتخاب کند که جزء گزینه‌های اصلی طراح نبوده است. تعمیم این نتیجه‌گیری در [۱۸] مفصل بررسی شده است و با نتیجه‌ی این پژوهش همخوانی دارد.

کنترل راه رفتن ربات با پنجه پیچیده‌تر است. بنابراین در بسیاری از موارد از آن استفاده نمی‌شود. تاکنون پژوهشی به طور مشخص استفاده از یادگیری تقویتی برای ربات با پنجه (و تاثیر آن در کاهش پیچیدگی کنترل پنجه و تسهیل استفاده‌ی از پنجه) را مورد بررسی قرار نداده است. اما در این پژوهش نشان داده شد که می‌توان از پنجه در بهبود گام‌برداری ربات استفاده کرد در عین اینکه درگیر پیچیدگی‌هایی که در کنترل کلاسیک ربات با پنجه به وجود می‌آید نشد که ناشی از استفاده‌ی از روش یادگیری تقویتی است.

بعضی از شرایط خاص زمین می‌تواند بیشتر کاربردهای پنجه را بروز دهد. به طور مثال پیش‌بینی می‌شود ربات با پنجه در حرکت بر روی سطح شیب‌دار در مقایسه با ربات انسان‌نمای بدون پنجه کارآمدتر عمل کند. به طور دقیق‌تر پیشنهاد می‌گردد از ربات با پنجه برای یادگیری ورود از زمین صاف به زمین شیب‌دار بررسی‌های مفصل‌تری انجام پذیرد.

مراجع

[1] R. S. Sutton, and A. G. Barto, "*Reinforcement Learning: An Introduction*", MIT Press, 2018, https://mitpress.mit.edu/9780262039246/reinforcement-learning/.

[2] L. T. Russell, "Applied Optimal Control for Dynamically Stable Legged Locomotion," PhD Thesis, Massachusetts Institute of Technology, 2004, https://dspace.mit.edu/handle/1721.1/28742.

[3] Y. Wu, D. Yao, X. Xiao, and Z. Guo, "Intelligent Controller for Passivity-based Biped Robot using Deep Q Network," *Journal of Intelligent & Fuzzy Systems*, Vol. 36, No. 1, pp. 731-745, 2019, doi: https://doi.org/10.3233/JIFS-172180.

[4] T. P. Lillicap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous Control with Deep Reinforcement Learning," *arXiv Preprint arXiv:1509.02971*, 2015, doi: https://doi.org/10.48550/arXiv.1509.02971.

[5] S. Fujimoto, H. Hoof, and D. Meger, "Addressing Function Approximation Error in Actor-critic Methods," *arXiv Preprint arXiv. 1802.09477*, 2018, doi: https://doi.org/10.48550/arXiv.1802.09477.

[6] J. García, and D. Shafie, "Teaching a Humanoid Robot to Walk Faster through Safe Reinforcement Learning," *Engineering Applications of Artificial Intelligence*, Vol. 88, p. 103360, 2020, doi: https://doi.org/10.1016/j.engappai.2019.103360.

[7] L. C. Melo, D. C. Melo, and M. R. Maximo, "Learning Humanoid Robot Running Motions with Symmetry Incentive through Proximal Policy Optimization," *Journal of Intelligent & Robotic Systems*, Vol. 102, No. 3, pp. 1-15, 2021, doi: https://doi.org/10.1007/s10846-021-01355-9.

[8] M. Sadedel, A. Yousefi Koma, and F. Iranmanesh, "Heel-off and Toe-off Motions Optimization for A2d Humanoid Robot Equipped with Active Toe Joints," *Modares Mechanical Engineering*, Vol. 16, No. 3, pp. 87-97, 2016, http://mme.modares.ac.ir/article-15-8715-en.html

[9] M. Sadedel, A. Yousefi-Koma, M. Khadiv, and F. Iranmanesh, "Heel-strike and Toe-off Motions Optimization for Humanoid Robots Equipped with Active Toe Joints," *Robotica*, Vol. 36, No. 6, pp. 925-944, 2018, doi: https://doi.org/10.1017/S0263574718000140.

[10] M. Sadedel, A. Yousefi-Koma, M. Khadiv, and M. Mahdavian, "Adding Low-cost Passive Toe Joints to the Feet Structure of SURENA III Humanoid Robot," *Robotica*, Vol. 35, No. 11, pp. 2099-2121, 2017, doi: https://doi.org/10.1017/S026357471600059X.

[11] M. Spitznagel, D. Weiler, and K. Dorer, “Deep Reinforcement Multi-directional Kick-learning of a Simulated Robot with Toes,” in 2021 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC), 28-29 April, Santa Maria da Feira, Portugal, pp. 104–110, Apr. 2021, doi: 10.1109/ICARSC52212.2021.9429811.

[12] J. Fischer, and K. Dorer, "Learning a Walk Behavior Utilizing Toes from Scratch," Robocup.info, Jul. 22, 2019, Available: https://archive.robocup.info/Soccer/Simulation/3D/FCPs/RoboCup/2019/magmaOffenburg_SS3D_RC2019_FCP.pdf. [Accessed: Nov. 01, 2019].

[13] A. Duburcq, F. Schramm, G. Boéris, N. Bredeche, and Y. Chevaleyre, "Reactive Stepping for Humanoid Robots using Reinforcement Learning: Application to Standing Push Recovery on the Exoskeleton Atalante," *arXiv Preprint arXiv: 2203.01148*, 2022, doi: https://doi.org/10.48550/arXiv.2203.01148.

[14] H. Kim, D. Seo, and D. Kim, "Push Recovery Control for Humanoid Robot using Reinforcement Learning," in *2019 Third IEEE International Conference on Robotic Computing (IRC)*, 2019: IEEE, 25-27 February, Naples, Italy, pp. 488-492, doi: https://doi.org/10.1109/IRC.2019.00102.

[15] G. Bingjing, H. Jianhai, L. Xiangpan, and Y. Lin, "Human–robot Interactive Control Based on Reinforcement Learning for Gait Rehabilitation Training Robot," *International Journal of Advanced Robotic Systems*, Vol. 16, No. 2, p. 1729881419839584, 2019, doi: https://doi.org/10.1177/1729881419839584.

[16] A. Ehsaniseresht, and M. M. Moghaddam, "A New Ground Contact Model for the Simulation of Bipeds' Walking, Running and Jumping," in *2015 3rd RSI International Conference on Robotics and Mechatronics (ICROM)*, 2015: IEEE, 07-09 October, Tehran, Iran, pp. 535-538, doi: https://doi.org/10.1109/ICRoM.2015.7367840.

[17] M. S. Shourijeh, and J. McPhee, "Foot-gound Contact Modeling within Human Gait Simulations: from Kelvin-Voigt to Hyper-volumetric Models," *Multibody System Dynamics*, Vol. 35, No. 4, pp. 393-407, 2015, doi: https://doi.org/10.1007/s11044-015-9467-6.

[18] N. Heess, D. TB, S. Sriram, J. Lemmon, J. Merel, G. Wayne, Y. Tassa, T. Erez, Z. Wang, S. M. Ali Eslami, M. Riedmiller, and D. Silver, "Emergence of Locomotion Behaviours in Rich Environments," *arXiv Preprint arXiv: 1707.02286*, 2017, doi: https://doi.org/10.48550/arXiv.1707.02286.

فهرست نمادهای انگلیسی

X، Y و Z	مؤلفه‌های مختصات مرکز قاب ربات با توجه به قاب اینرسی
R_i	مؤلفه‌های کاندید تشکیل تابع پاداش
$\bar{R}_i$	تابع پاداش iام
t_s	گام زمانی
T_f	زمان پایان دوره
T_j	گشتاور مفصل jام
k_h	شبه‌نرمی حجمی غیرخطی
V	حجم فرورفتگی در زمین
a_h	حاصلضرب نرمی و دمپینگ زمین

نمادهای یونانی

v_{ct}	سرعت مرکز جرم فرورفتگی جسم در زمین
v_s	ضریب شکل
μ_f	ضریب مجانبی اصطکاک
$\mathcal{H}$	ضریب وابسته به نرمی حجمی و مشخصه‌های هندسی

Gait Control of Humanoid Robot with Toe Joints Based on Reinforcement Learning

Aref Tavangar
M.Sc., Department of Mechanical Engineering, Tarbiat Modares University, Tehran, Iran
aref.tavangar@modares.ut.ac.ir

*Corresponding author: **Majid Sadedel**
Assistant Professor, Department of Mechanical Engineering, Tarbiat Modares University, Tehran, Iran
majid.sadedel@modares.ut.ac.ir

Abstract

Controlling a humanoid robot is a complicated task because it deals with a high degree of freedom, a non-holonomic and underactuated system. Many model-based control strategies have been implied on humanoid robots. Over time model-free and AI-based strategies have taken place. Among AI strategies, Reinforcement Learning has the largest share. Many complex systems have successfully controlled to perform complicated tasks such as jumping and running. Toe joints is almost missing in all of these systems and does not have the application it performs in humans. Toed robots can outperform, so implementing Reinforcement Learning algorithms on a humanoid with an active toe joint has been studied. Two algorithms, DDPG, and TD3 were applied and compared. A customized RL framework was designed to teach a humanoid to walk. Simulations showed that the task of controlling a humanoid to walk was accomplished. Learned robot was able to gait on a flat surface at the average speed of 0.9 m/s.

Keywords: Humanoid Robot, Active toe joint, Learn to walk, Deep reinforcement learning