Journal of Applied And Basic Machine Intelligence Research (ABMIR)

پژوهش‌های نظری و کاربردی هوش ماشینی

سال سوم، شماره اول، بهار و تابستان ۱٤۰٤، صفحات ۱۵ - ۱

ارائه یک روش جدید ناوبری خودگردان کپسول درون‌بین مبتنی بر یادگیری تقویتی عمیق با بهینه‌سازی سیاست مجاور در یک محیط مجازی

حمیدرضا قهرمانی [۱]، وحید جوهری مجد [۲*]

[۱] دانشجوی دکترای دانشکده مهندسی برق و کامپیوتر، دانشگاه تربیت مدرس، تهران، ایران

[۲] دانشیار دانشکده مهندسی برق و کامپیوتر، دانشگاه تربیت مدرس، تهران، ایران

مقاله پژوهشی

تاریخ دریافت:
۱٤۰٤/۲/۷
تاریخ پذیرش:
۱٤۰٤/۳/۲۹

کلیدواژه‌ها:
ناوبری خودگردان، کپسول درون‌بین، یادگیری تقویتی عمیق، بهینه‌سازی سیاست مجاور، محیط مجازی

نویسنده‌مسئول:
majd@modares.ac.ir

چکیده

با توجه به نگرانی‌های عمومی از عوارض درون‌بینی سنتی، تحقیقات در زمینه استفاده از کپسول‌های درون‌بین به عنوان روشی کمتر تهاجمی مورد توجه قرار گرفته است. اما حرکت غیرفعال کپسول باعث عدم دسترسی به زوایای مدنظر پزشک می‌شود. برای رفع این محدودیت، یک رویکرد نوین ناوبری خودگردان مبتنی بر یادگیری تقویتی عمیق با استفاده از الگوریتم بهینه‌سازی سیاست مجاور ارائه شده است تا فرآیند موقعیت‌یابی، مسیریابی و کنترل حرکت کپسول را به‌صورت اتوماتیک انجام دهد. در این روش با ادغام داده‌های چندوجهی حسگرها، نقطه هدف در طول زمان تخمین زده می‌شود. از آنجا که آموزش اولیه الگوریتم نیاز به داده‌های فراوانی دارد یک محیط مجازی نزدیک به واقعیت برای آموزش عامل هوشمند شامل مدل عملگر با ساختاری از سیم‌پیچ‌های مغناطیسی، کپسولی مجهز به آهنربای دوقطبی، دوربین، حسگر اینرسی، و مدل سه‌بعدی روده بزرگ، فراهم شده است. هدف اصلی در این پژوهش، کاهش مداخله‌های عملیاتی اپراتور جهت تمرکز بیشتر بر جنبه‌های بالینی و پزشکی درون‌بینی است. روش ارائه‌شده با اَبَرمتغیرهای مختلفی آموزش داده شده و نتایج آن بر اساس شاخص‌های حرکت و هم‌جهتی به سمت هدف و آنتروپی مقایسه شده است. ارزیابی نتایج نشان می‌دهد که با تنظیم بهینه اندازه بافر و مقدار دسته، الگوریتم عملکرد مناسبی در ردیابی و پایداری دارد.

doi : 10.22034/ABMIR.2025.23028.1122

E-ISSN: 2821-2037

۱- مقدمه

روده‌بینی[1] به عنوان موثرترین روش برای تشخیص سرطان روده به کار گرفته می‌شود. مطالعات گسترده‌ای نشان داده‌اند که کولونوسکوپی غربالگری به‌طور قابل توجهی بروز و مرگ و میر ناشی از سرطان روده را کاهش می‌دهد [۱]. اما عمل تشخیصی کولونوسکوپی عوارض متعددی از جمله پارگی روده را به همراه دارد. همچنین بسیاری از افراد واجد شرایط غربالگری به دلیل عوامل متعددی نظیر ترس، احساس ناراحتی و درد ناشی از تجربه کولونوسکوپی، از انجام آن صرف‌نظر می‌کنند [۲]. درون‌بینی[2] با کپسول بی‌سیم به عنوان یک راه حل جایگزین مناسب در مقابل روش‌های سنتی درون‌بینی، یک روش کم‌تهاجمی و بدون درد است که عمل تشخیص با بلعیدن یک کپسول کوچک مجهز به دوربین انجام می‌شود و بیماران را قادر می‌سازد تا معاینات گوارشی را به‌طور مکرر و راحت‌تر انجام دهند، در نتیجه تشخیص زودهنگام را افزایش می‌دهد [۳] و [٤].

محدودیت اصلی سیستم کپسول درون‌بین غیرفعال، عدم امکان مانور پذیری توسط اپراتور برای بررسی زوایای خاص بوده که این موضوع منجر به بررسی ناقص روده و در نتیجه کاهش دقت تشخیصی می‌شود. برای غلبه بر این چالش، محققان سیستم‌های محرک مغناطیسی خارجی را پیشنهاد داده‌اند که امکان کنترل دقیق حرکت کپسول درون‌بین[3] (CE) را فراهم می‌آورد. در این سیستم‌ها، یک آرایه‌ای از سلونوئیدها به‌صورت مدور در اطراف بدن فرد قرار می‌گیرد و با ایجاد میدان مغناطیسی، از طریق برهمکنش با آهنربای دائمی موجود در CE، امکان کنترل و هدایت حرکت آن فراهم می‌شود. مقاله [۵] با معرفی یک سیستم عملگر مغناطیسی جدید با ۸ سیم‌پیچ توانسته است CE را در ۵ درجه آزادی کنترل کند به‌صورتی که فضای کاری بین سیم‌پیچ‌ها بیشینه شود. مقاله [٦] با آرایشی از ۸ سیم‌پیچ الکترومغناطیس موقعیت CE را به‌وسیله روش کنترل حلقه باز و جهت‌گیری کپسول را به‌وسیله کنترل‌کننده تناسبی مد لغزشی مبتنی بر بینایی کنترل کرده است. با این حال، بسیاری از سیستم‌های موجود هنوز به یک اپراتور متخصص نیاز دارند تا حرکت CE را به‌صورت دستی کنترل کند که مقیاس‌پذیری و کارایی این سیستم‌ها را محدود می‌کند. پژوهشگران در مقاله [۷] موفق به ارائه روش‌های ناوبری شده‌اند که آندوسکوپ با سیم را به‌صورت خودگردان در روده با کمترین مداخله انسانی حرکت می‌دهد. این رویکرد به افزایش سطح استقلال CE کمک می‌کند و منجر به کاهش زمان و بار کاری فرآیند عمل تشخیصی درون‌بینی می‌شود.

برای پیاده‌سازی روش‌های ناوبری خودگردان[4]، در گام اول نیاز به پایش مداوم مسیر پیمایش شده، مکان‌یابی محلی و تولید مسیر آینده است و در گام بعدی سیستم کنترلی باید CE را در مسیر مطلوب هدایت کند. با این حال، محلی سازی موقعیت کپسول درون بدن به دلیل ترکیب بافتی متنوع بدن، فضای محدود درون CE برای بهره‌گیری از حسگرهای متنوع و تداخل میدان‌های مغناطیسی چالش برانگیز است [٤]. از طرفی مسئله کنترل حرکت CE در محیط منعطف روده با وجود عدم قطعیت بالا و همچنین استفاده بهینه از ترکیب داده‌های چندوجهی حسگرها به عنوان بازخورد، چالش جدی در سیستم‌های کنترلی است.

برای غلبه بر این مشکلات، تحقیقات جدید بر پایه پیشرفت‌های هوش مصنوعی، به‌ویژه یادگیری تقویتی (DRL)[5]، به‌منظور پیاده‌سازی استراتژی‌های کنترل تطبیقی که از محیط یاد می‌گیرند و حرکت را در زمان واقعی بهینه می‌کنند، انجام شده است. در مقاله [۸] از روش بهینه‌سازی سیاست مجاور (PPO)[6] مبتنی بر DRL برای کنترل خودگردان یک دستگاه آندوسکوپ نرم استفاده شده است که با تکیه بر داده‌های تصویری دوربین و بهره‌گیری از رفتار انطباقی، امکان هدایت مؤثر این دستگاه را در بخش‌های پیچیده روده بزرگ فراهم می‌کند. محققان در مقاله [۹]، کاری معطوف به کنترل CE بر پایه یادگیری تقویتی عمیق جهت پایش معده ارائه کرده‌اند که در یک محیط مجازی، شبیه‌سازی شده و مورد آزمایش قرار گرفته است. در این پژوهش از روش قدرتمند DRL استفاده شده است که توانسته با افزایش بهره‌وری در مسیرپیمایی، پوشش

[1] Colonoscopy
[2] Endoscopy
[3] Capsule Endoscope
[4] Autonomous Navigation
[5] Deep Reinforcement Learning
[6] Proximal Policy Optimization

گسترده‌ای از سطح داخلی معده را در بازه زمانی محدود به دست آورد.

برخلاف روش‌های سنتی کنترل غیرخطی که عمدتاً متکی بر مدل‌سازی صریح سیستم و استخراج روابط دقیق دینامیکی هستند، رویکردهای مبتنی بر DRL به‌طور موثری قادرند از داده‌های ورودی با ابعاد بالا بهره‌برداری کرده و سیاست‌های کنترلی بهینه‌ای را استخراج کنند. این ویژگی امکان تعمیم‌پذیری بالاتر این روش‌ها را در مواجهه با عدم قطعیت ناشی از تغییرات آناتومیکی و تفاوت‌های فردی بین بیماران را فراهم می‌کند.

بااین‌حال، علی‌رغم پتانسیل بالای DRL در بهینه‌سازی فرآیند کنترل درون‌بین کپسولی، چالش‌های متعددی در ناوبری خودگردان آن، از جمله یافتن مسیر مطلوب و هدایت به سمت آن باقی می‌ماند. همچنین توسعه یک محیط برای آموزش عامل هوشمند، پیاده‌سازی ساختارهای کنترلی کارآمد و آزمایش الگوریتم‌های مختلف، از الزامات اساسی این حوزه به شمار می‌رود.

در این مقاله، سه نوآوری اساسی مطرح شده است:

۱- بر اساس مطالعات انجام شده، این اولین پژوهش در زمینه کنترل CE با عملگرهای سیم‌پیچ ایستا به‌وسیله DRL داخل روده بزرگ بر پایه ادغام داده‌های چندوجهی[1] حسگر اینرسی، دوربین و حسگر جریان است.

۲- این پژوهش یک روش جدید ناوبری خودگردان CE با بهره‌گیری از الگوریتم ارائه‌شده تخمین نقطه هدف سه‌بعدی در سیستم کنترل حلقه بسته یادگیری تقویتی با الگوریتم بهینه‌سازی سیاست مجاور معرفی می‌کند که هدف آن کاهش بار کاری اپراتور در فرآیند درون‌بینی است.

۳- همچنین در این پژوهش یک محیط مجازی نزدیک به واقعیت با پیاده‌سازی مدل‌های عملگر، عامل و محیط فراهم شده است و بر مبنای آن عملکرد چندین الگوریتم DRL ارزیابی و مقایسه شده است که می‌تواند زمینه توسعه رویکردهای کارآمدتر برای آموزش عامل هوشمند را ایجاد کند.

۲- مفاهیم اولیه

۲-۱ محیط مجازی آموزش

از آنجایی که الگوریتم‌های DRL برای یادگیری اقدامات پیچیده به تجربه زیادی نیاز دارند، بنابراین آموزش مستقیم آن‌ها در سناریوهای دنیای واقعی غیرعملی است. علاوه‌براین، در مراحل اولیه یادگیری، رفتار CE تقریباً غیرقابل پیش‌بینی است و می‌تواند به‌طور بالقوه به دستگاه گوارش آسیب برساند. برای پرداختن به این مسائل، الگوریتم‌های DRL معمولاً در محیط‌های مجازی آموزش داده می‌شوند که نه‌تنها فرآیند یادگیری روش‌های گوناگون کنترل درون‌بین کپسولی را تسریع می‌کند، بلکه هزینه‌های مواد را نیز به حداقل می‌رساند [۱۰]. همچنین بهره‌برداری از محیط شبیه‌سازی در جهت توانمندسازی اپراتور باعث کاهش زمان و مهارت لازم برای یادگیری درون‌بینی می‌شود.

در این پژوهش، محیط مجازی با استفاده از نرم‌افزار Unity3D طراحی شده است که رفتارهای عامل هوشمند و مدل‌سازی تعاملات بین عامل و محیط موردنظر را نزدیک به واقعیت شبیه‌سازی می‌نماید [۱۱]. این محیط از موتور فیزیک پیشرفته Nvidia PhysX بهره می‌برد که علاوه بر شبیه‌سازی دقیق پدیده های فیزیکی، قابلیت نونمایی[2] بصری با کیفیت بالا را نیز فراهم می‌آورد [۱۲]. برای آموزش عامل هوشمند در این محیط، از جعبه ابزار ML-Agents استفاده شده است که یک کتابخانه منبع باز برای پیاده‌سازی الگوریتم‌های یادگیری تقویتی در یونیتی است.

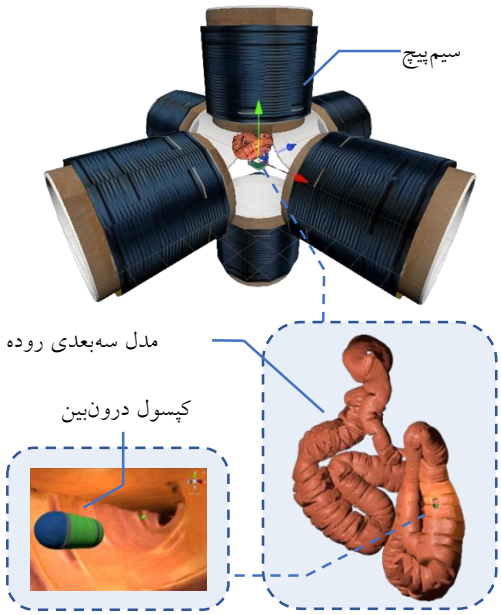
جزئیات عامل و محیط مجازی در شکل (۱) نشان داده شده است. عامل موردنظر، کپسولی است که درون روده بزرگ قرار دارد و مجهز به یک دوقطبی مغناطیسی و دوربین تک‌چشمی است. این کپسول تحت تأثیر میدان مغناطیسی حاصل از شش سیم‌پیچ خارجی قرار می‌گیرد که با اعمال جریان الکتریکی، میدان‌های مغناطیسی لازم برای حرکت CE را تولید می‌نمایند. معادلات میدان مغناطیسی برای هر سیم‌پیچ بر اساس مدل سیم‌پیچ با طول محدود پیاده‌سازی شده‌اند که با دقت بسیار بالا امکان محاسبه دقیق

[1] Multimodal

[2] Rendering

نیروهای وارد بر آهنربای CE را در هر نقطه از فضای درون روده فراهم می‌کند.

مدل مجازی تولیدشده قابلیت پیاده‌سازی انواع ساختارهای عملگر، عامل و محیط و همچنین اجرای مدل‌های آموزشی متنوع را داراست و به عنوان یک پلتفرم عمومی، زمینه را برای کاربردهای پیشرفته در پژوهش‌های حوزه CE فراهم می‌نماید.



شکل (۱): شمایی از محیط مجازی

۲-۲ یادگیری تقویتی عمیق

الگوریتم DRL قابلیت تصمیم‌گیری بهینه برای کنترل حرکت CE را در محیط پیچیده و پویای روده فراهم می‌سازد. این روش به CE این امکان را می‌دهد که با تعامل مداوم با محیط روده و دریافت بازخورد پاداش، استراتژی‌های بهینه‌ای برای حرکت و هدایت را یاد بگیرد. این مسئله کنترل با چارچوب فرآیندهای تصمیم‌گیری مارکوف $M = (S, A, \mathrm{T}, R, \gamma)$ صورت‌بندی می‌شود، که در آن S مجموعه حالات است، A مجموعه اعمال، T تابع انتقال است: $\mathrm{T}: S \times A \to S$، R تابع پاداش است: $R: S \times A \to \mathbb{R}$، و $\gamma \in (0,1]$ ضریب تخفیف است. توجه داشته باشید که $A(s) \subseteq A$ نشان‌دهنده مجموعه اعمال موجود در حالت s است. در هر مرحله زمانی t که عامل درحالت $s_t \in S$ است، یک اقدام $a_t \in A$ انجام می‌دهد،یک پاداش اسکالر $r_t \in R$ دریافت می‌کند و به یک وضعیت جدید s_{t+1} می‌رسد. تعداد معینی از مراحل زمانی محدود که در آن کاری انجام می‌شود، یک اپیزود را تشکیل می‌دهند. اپیزودها با موفقیت (تکمیل کار) یا شکست (در غیر این صورت)، با تعریفی از جانب کاربر به پایان می‌رسند. مسئله یادگیری تقویتی، یافتن اقدام مناسب عامل در محیط، به‌وسیله یک تابع تصمیم‌گیری است. این تابع که به آن سیاست $\pi: S \to A$ می‌گویند و می‌تواند قطعی یا تصادفی باشد، یک نگاشت از حالت‌ها به مجموعه‌ای از اعمال است که تعیین می‌کند عامل در هر حالت S باید چه عملی انجام دهد. در الگوریتم DRL، این سیاست معمولاً توسط یک شبکه عصبی عمیق با ساختار پیشرو مدل‌سازی می‌شود که پارامترهای آن طی فرایند یادگیری در طول اپیزودها به‌روزرسانی می‌گردند. هدف یک الگوریتم DRL یافتن سیاست بهینه π^* است که تابع پاداش تجمعی مورد انتظار را در یک مسیر ξ به حداکثر برساند. این تابع را می‌توان به‌صورت زیر بیان کرد:

$$\max_{\pi_\theta \in \Pi} J_r(\pi_\theta) := \mathbb{E}_{r_i, s_i, a_i \sim \pi_\theta}[\sum_{i=0}^{\infty} \gamma^i r_i] \qquad (۱)$$

که در آن π_θ ،سیاست با پارامترهای θ است.

الگوریتم‌های DRL به‌طور کلی به دو دسته ارزش‌محور و سیاست‌محور تقسیم می‌شوند. در مسئله کنترل CE با فضای حالت و عملگر پیوسته و با ابعاد بالا، روش بهینه‌سازی سیاست مجاور به عنوان یک روش سیاست محور به دلیل پایداری، کارایی بالا و توانایی یادگیری مؤثر در فضای بزرگ با بار محاسباتی کم، گزینه‌ای مناسب محسوب می‌شود.

۲-۳ بهینه‌سازی سیاست مجاور

الگوریتم بهینه‌سازی سیاست مجاور (PPO) معرفی شده توسط [۱۳]، یکی از روش‌های پیشرفته مبتنی‌بر گرادیان سیاست است که چالش تغییرات بزرگ در به‌روز رسانی پارامترهای سیاست که منجر به ناپایداری می‌شود را با محدود کردن دامنه به‌روزرسانی‌ها به‌منظور جلوگیری از افت عملکرد ناگهانی، حل می‌کند. ایده اصلی، بهینه‌سازی سیاست پارامتری $\pi_\theta(a \mid s)$با استفاده از یک تابع هدف جانشین قطع‌شده است:

$$L^{CLIP}(\theta) = \mathbb{E}_t[\min(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t)] \quad (۲)$$

که در آن $r_t(\theta) = \frac{\pi_\theta(a_r|s_r)}{\pi_{\theta_{old}}(a_t|s_t)}$ نسبت احتمال بین سیاست جدید و قدیم، A_t تخمین مزیت در گام زمانی t و ϵ یک پارامتر کنترل (معمولاً $0.1 \leq \epsilon \leq 0.2$) است که محدوده قطع را تعیین می‌کند. اگر $r_t(\theta) < 1$، عامل به سمت بهره‌گیری تمایل دارد، چون احتمال اجرای آن عمل در سیاست جدید افزایش یافته است و اگر $r_t(\theta) > 1$، عامل به سمت کاوش حرکت می‌کند، چون سیاست جدید تمایل کمتری به اقدام قبلی دارد. درنهایت، تابع زیان کلی به‌منظور بهینه‌سازی عملکرد، در عین حفظ پایداری الگوریتم، با ترکیبی از سه جزء زیر طراحی شده است:

$$L_t^{CLIP+VF+S}(\theta) = \mathbb{E}_t[L^{CLIP}(\theta) - c_1 L^{VF}(\theta) + c_2 S[\pi_\theta](s_t)] \quad (۳)$$

که در آن $L^{VF}(\theta)$ زیان ارزش و $S[\pi_\theta](s_t)$ آنتروپی[1] برای تشویق به اکتشاف و جلوگیری از همگرایی زودرس به سیاست‌های قطعی طراحی شده است. جمله آنتروپی به عنوان یک تنظیم‌کننده عمل می‌کند که با افزایش تنوع اقدامات، امکان کشف راهکارهای بهینه را فراهم می‌آورد. این ترکیب هوشمندانه از اجزای مختلف، PPO را به الگوریتمی کارآمد، پایدار و قابل اعتماد تبدیل کرده است که می‌تواند در طیف وسیعی از مسائل یادگیری تقویتی مورد استفاده قرار گیرد.

۳- ناوبری خودگردان

کنترل حرکت CE در روده بزرگ با چالش عدم وجود مسیرهای از پیش تعریف شده مواجه است، زیرا شکل و حرکات بافتی روده به‌طور پویا تغییر می‌کنند و تعیین مسیر مطلوب برای حرکت کپسول را در عمل غیرممکن می‌سازد. برای رفع این مسئله، سیستم ناوبری خودگردان به‌گونه‌ای طراحی شده است که با شناسایی پویای هدف بتواند مسیر حرکت مداوم CE را هم‌راستا با لومن روده تعیین کند. این فرآیند از طریق اعمال سیگنال‌های کنترلی به سیستم مغناطیسی عملگرها صورت می‌پذیرد که باعث می‌شود کپسول با کاهش وابستگی به نظارت مستقیم اپراتور، به‌طور خودکار در جهت عمیق‌ترین نقاط روده حرکت کند. برای دستیابی به این هدف، از تکنیک تخمین نقطه هدف مبتنی بر پردازش تصویر استفاده شده است که با یک الگوریتم DRL ترکیب شده است. این روش نیاز به استخراج ویژگی‌های خاص از چین‌های هاسترال روده بزرگ را حذف می‌کند و عملکرد تخمین نقطه هدف را در شرایط مختلف روده بهبود می‌بخشد. مزیت این روش، تخمین نقطه هدف بدون نیاز به استخراج ویژگی خاص برای چین‌های هاسترال روده بزرگ و عملکرد مناسب در شرایط متغیر محیطی است.

۳-۱ تخمین نقطه هدف

برای حرکت CE در روده به‌صورت مستقل، شناسایی و تخمین دقیق نقطه هدف در مختصات سه‌بعدی از اهمیت بالایی برخوردار است. زیرا انتخاب بهینه این نقطه به‌طوری‌که کپسول بیشترین فاصله را از موقعیت فعلی خود طی کند منجر به آن می‌شود که CE کل مسیر روده را پوشش دهد. از آنجا که در زاویه دید دوربین کپسول، نقاط عمیق‌تر روده به دلیل فاصله بیشتر از منابع نور LED تعبیه‌شده مجاور دوربین کپسول، شدت روشنایی کمتری دریافت کرده و به‌صورت تاریک‌تری در تصویر ثبت می‌شوند، الگوریتم تخمین نقطه هدف از این ویژگی فیزیکی بهره می‌برد تا با تحلیل پیکسل‌های تصویر دوبعدی دوربین، تاریک‌ترین ناحیه شناسایی شده و به عنوان کاندیدای اصلی هدف در دو بعد معرفی می‌گردد. سپس متناظر سه‌بعدی آن محاسبه می‌شود. برای شناسایی تاریک‌ترین نقطه ابتدا، مقدار درخشندگی هر پیکسل محاسبه شده (که با مقدار بیشینه درخشندگی یعنی ۲۵۵ نرمال شده است) و با استفاده از رابطه زیر میزان تاریکی آن به دست می‌آید.

$$\text{میزان تاریکی} = ۱ - \text{مقدار درخشندگی} \quad (٤)$$

پیکسل‌هایی که مقدار تاریکی آن‌ها از آستانه ۰/۷۵ فراتر رفته باشد، زیرمجموعه تاریک از تصویر را به عنوان کاندید برای هدف تشکیل می‌دهند. با میانگین وزن‌دار از نقاط این زیرمجموعه (با وزن‌دهی نمایی به پیکسل‌های تاریک‌تر)، تاریک‌ترین نقطه به عنوان مرکز هدف دوبعدی به دست می‌آید. سپس، معیارهای آماری کلیدی

[1] Entropy

مانند نسبت تاریکی تصویر[1]، شدت تاریکی تصویر[2] و پراکندگی مکانی[3] زیرمجموعه تاریک محاسبه‌شده که برای ارزیابی اعتبار و کیفیت هدف شناسایی‌شده در فرآیند تخمین نقطه هدف به کار می‌روند. همچنین برای مشخص کردن هر حالت شاخصی به‌نام اطلاعات تعریف شده که سطوح بالاتر اطلاعات، بیانگر غنای داده‌ای بیشتر تصویر و جزئیات کافی برای تحلیل دقیق‌تر موقعیت نقطه هدف است. شاخص اطلاعات به عنوان بازخوردی از رفتار CE، نقش مهمی در تشخیص عملکرد اقدامات عامل در آن گام آموزشی دارد.

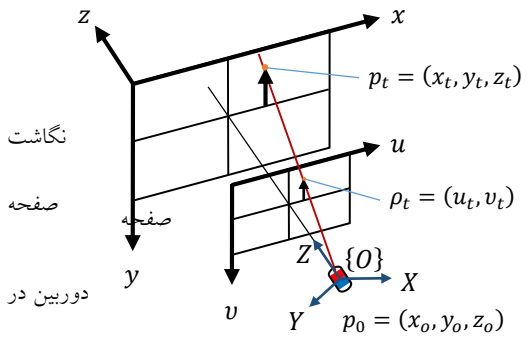
نسبت تاریکی تصویر، با استفاده از نسبت تعداد پیکسل‌های موجود در زیرمجموعه تاریک به کل تعداد پیکسل‌های تصویر به دست می‌آید و بیانگر گستردگی ناحیه تاریک در تصویر است. شدت تاریکی تصویر نیز به‌صورت مقدار مطلق تاریکی تاریک‌ترین پیکسل تعریف می‌شود که نماینده میزان کلی تاریکی تصویر است. معیار سوم، واریانس مکانی پیکسل‌های زیرمجموعه تاریک است که نشان‌دهنده میزان تمرکز یا پراکندگی ناحیه تاریک در تصویر است. حال برمبنای این معیارها، یک درخت تصمیم‌گیری، میزان اعتبار نقطه عمیق و فاصله نقطه هدف با مکان فعلی CE در فضای سه‌بعدی (که عمق هدف نام‌گذاری می‌شود)، را مشخص می‌کند. در شرایطی که شدت تاریکی بسیار بالا، نسبت تاریکی مناسب و پراکندگی مکانی محدود باشد، بیشترین اعتبار به نقطه هدف اختصاص می‌یابد و عمق هدف نیز مقداری بیشتر نسبت به سایر حالات در نظر گرفته می‌شود. با کاهش شدت تاریکی، هم میزان اعتبار و هم عمق هدف کاهش می‌یابد. همچنین اگر شدت تاریکی بسیار کم باشد نشان‌دهنده روبرویی با دیوار روده است که برای پرهیز از این شرایط باید عامل محیط اطراف خود را برای اصلاح جهت کاوش کند. درنهایت، پس از تعیین مرکز هدف و عمق هدف در هر حالت از درخت تصمیم‌گیری، با استفاده از مدل دوربین سوراخ‌سوزنی[4] نگاشت نقطه هدف در مختصات سه‌بعدی محاسبه می‌شود.

شکل (۲): موقعیت‌های مختلف کپسول درون‌بین

هر بخش شکل (۲) بیانگر موقعیت سه‌بعدی کپسول است و تصویر پایین زاویه نگاه دو بعدی دوربین CE را نشان می‌دهد. در هر کدام از دو تصویر موقعیت نقطه هدف تخمین زده شده به‌صورت گوی یا علامت + رنگی نشان داده شده است. در تصویر (الف) معیارهای شدت تاریکی بالاتر از ۰/۸۵، نسبت تاریکی بالاتر از ۵ و پراکندگی مکانی زیرمجموعه تاریک در بازه ۳۰ و ۵۰۰ تعریف شده‌اند. در این حالت شاخص اطلاعات سطح ۳ در نظر گرفته می‌شود. همان‌طور که در تصویر بالا مشاهده می‌شود نقطه تخمین زده شده (گوی سبز) با فاصله نسبتاً زیادی با CE قرار دارد زیرا در الگوریتم تخمین نقطه هدف، به علت دید کافی عمق نقطه هدف مقدار نسبتاً زیادی انتخاب شده است. در تصویر پایین نیز

[1] Dark Ratio
[2] Darkness Intensity
[3] Spatial Variance
[4] pin-hole

به‌وضوح مشخص است که به لطف اطلاعات کافی تصویر، مکان نقطه هدف (علامت + سبز) به‌درستی تشخیص داده شده است و دقت قابل قبولی دارد. در تصویر (ب) معیارهای شدت تاریکی بالاتر از ۰٫۸، نسبت تاریکی بالاتر از ۰/٦ و پراکندگی مکانی زیرمجموعه تاریک در بازه ۲۰ و ۵۰۰ تعریف شده‌اند. اگرچه در این حالت اطلاعات کافی وجود دارد، اما کیفیت بالایی برای تخمین نقطه هدف ندارد لذا عمق در نظر گرفته شده برای نقطه هدف کاهش می‌یابد و همچنین شاخص اطلاعات سطح ۲ در نظر گرفته می‌شود. در تصویر (ج) معیارهای شدت تاریکی پایین‌تر از ۰٫۸ تعریف شده‌اند. در این حالت میزان تاریکی تصویر بسیار اندک است و در بیشتر مواقع کپسول به سمت دیواره روده قرار گرفته است. شاخص اطلاعات در این حالت سطح ۱ در نظر گرفته می‌شود. در تصویر (د) نیز حالتی است که زیرمجموعه تاریک تهی باشد. به عبارتی هیچ پیکسلی از تصویر با تاریکی مطلوب شناسایی نشده است و به‌دلیل عدم وجود اطلاعات کافی جهت تصمیم‌گیری شاخص اطلاعات سطح صفر در نظر گرفته می‌شود.



شکل (۳): مدل دوربین سوراخ‌سوزنی

۳-۲ مدل دوربین سوراخ‌سوزنی

برای تخمین موقعیت سه‌بعدی هدف (x, y, z) از داده‌های دوبعدی تصویری (u, v)، از مدل دوربین سوراخ‌سوزنی استفاده شده است که نگاشت هندسی بین صفحه دوبعدی تصویر دوربین و فضای سه‌بعدی واقعی را تعریف می‌کند. این مدل بر اساس اصل فیزیکی عبور نور از یک سوراخ کوچک (مرکز دید) به صفحه تصویر عمل می‌کند و فرض می‌کند که نقاط سه‌بعدی از طریق این سوراخ به نقاط دوبعدی روی سنسور دوربین نگاشته می‌شوند. تبدیل مختصات دوبعدی هدف $\rho_t = (u_t, v_t)$ به مختصات سه‌بعدی $p_t = (x_t, y_t, z_t)$ با استفاده از ماتریس ذاتی دوربین، که بر اساس فاصله کانونی (فاصله بین مختصات کپسول تا مرکز صفحه دوربین) و مرکز تصویر محاسبه می‌شود، انجام می‌گیرد. همان‌طور که در شکل (۳) نشان داده شده است، سیستم مختصات دوربین با محور X در جهت راست، Y در جهت عمودی، و Z در جهت رو به جلو نام‌گذاری شده است. اندازه طول و عرض صفحه دوربین برابر تعداد پیکسل‌های دوربین است و نقطه ρ_t مختصات هدف در فضای دوبعدی صفحه دوربین است که به اندازه u_t و v_t پیکسل از مرکز صفحه فاصله دارد. اگرچه این مدل به دلیل سادگی و کارایی در سیستم‌های ادراک بینایی پراستفاده است، اما شرایط را ایده‌آل فرض می‌کند و از تغییر شکل‌های عدسی (مانند اعوجاج شعاعی یا مماسی) چشم‌پوشی می‌کند که ممکن است دقت را در پیاده‌سازی واقعی محدود کند و برای کاربردهای عملی باید از مدل کامل آن بهره جست [۱٤].

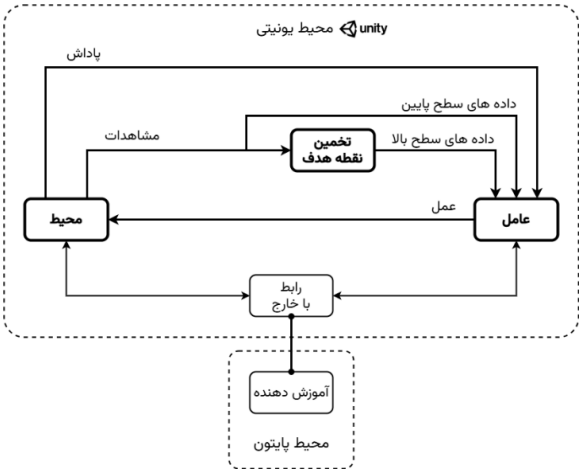
۳-۳ کنترل کپسول درون‌بین

با روش کنترل هوشمند توسعه‌یافته می‌توان بر پیچیدگی کنترل کپسول توسط میدان مغناطیسی ساکن با رفتار غیرخطی غلبه کرد. لذا در این پژوهش برای رسیدگی به چالش ناوبری خودگردان CE و کنترل آن جهت حرکت در مسیر هدف تخمینی، با بهره‌برداری از ترکیب حسگرهای جریان، اینرسی و بینائی از روش DRL استفاده شده است که روابط اجزای آن در شکل (٤) نشان داده شده است.

۳-٤ فضای عمل

حرکت CE از طریق تعامل میدان مغناطیسی خارجی و آهنربای دائمی تعبیه‌شده در داخل کپسول محقق می‌شود. میدان مغناطیسی خارجی در هر مکان p از فضا، توسط آرایه‌ای از محرک‌های الکترومغناطیسی تولید می‌شود که به‌صورت ریاضی با برهم‌نهی میدان‌ها با رابطه $\vec{\mathbf{B}}(p) = \sum_{i=1}^{n_e} \vec{B}_i(p)$، مدل‌سازی می‌شود که در آن n_e تعداد الکترومغناطیس‌ها و $\vec{B}_i(p)$ سهم میدان مغناطیسی تولیدشده توسط الکترومغناطیس i-ام است. همان‌طور که در شکل (۱) نشان داده شده است، سیستم از شش الکترومغناطیس تشکیل شده است که چهار الکترومغناطیس به‌صورت دو به دو در مقابل هم در یک صفحه و با زاویه ۹۰ درجه نسبت به یکدیگر قرار

گرفته‌اند و دو الکترومغناطیس باقیمانده نیز عمود بر این صفحه و در طرفین صفحه به‌صورت روبروی یکدیگر جایگذاری شده‌اند. این آرایش از الکترومغناطیس‌ها، یک فضای کاری را که توسط یک کره محیطی پوشش داده می‌شود را شکل می‌دهد. این الکترومغناطیس‌ها با استفاده از پیکربندی‌های سولنوئیدی مکسول[1] و هلمهولتز[2] طراحی شده‌اند که به‌ترتیب مؤلفه‌های نیرو و گشتاور لازم برای حرکت انتقالی و دورانی CE را تأمین می‌کنند.



شکل (٤): چیدمان اجزای سیستم کنترل یادگیری تقویتی عمیق

میدان مغناطیسی هر سولنوئید با جریان عبوری از آن متناسب است که این امر کنترل دقیق حرکت کپسول را از طریق تنظیم جریان ممکن می‌سازد. برای مدل‌سازی دینامیک دوقطبی مغناطیسی کپسول با سولنوئیدها، از یک کتابخانه محاسباتی استفاده شده است که مدل‌های تحلیلی تعمیم‌یافته برای معادلات میدان‌های مغناطیسی ناشی از سیم‌پیچ‌های با طول محدود را پیاده‌سازی کرده است [۱۵]. یک عامل DRL با تعیین عمل‌های پیوسته a_i، جریان ورودی به هر سولنوئید را کنترل می‌کند. برای محاسبه این جریان ورودی، روی هر یک از خروجی‌های تابع سیاست یک اشباع‌کننده سخت اعمال می‌شود تا مقادیر آن در بازه $[-1,1]$ به دست آید. سپس، این مقادیر با ضرب در حداکثر جریان مجاز κ مقیاس‌دهی شده و به جریان نهایی ورودی به سیستم الکترومغناطیسی به‌صورت $I_i = \kappa \cdot a_i$ محاسبه می‌شوند که یک فضای عمل ٦ بعدی را تشکیل می‌دهند.

۳-۵ فضای مشاهدات

فضای مشاهدات سیستم از ترکیب داده‌های حسگر اندازه‌گیری مقدار جریان، اندازه‌گیری اینرسی شش محوره (IMU)[3] و دوربین تک‌چشمی تعبیه‌شده در CE تشکیل شده است که به‌طور هم‌زمان اطلاعات سینماتیکی و محیطی را جمع‌آوری می‌کنند. به‌منظور کاهش هزینه‌های محاسباتی و استخراج ویژگی‌های کلیدی برای تخمین نقطه هدف، تصاویر خام دوربین با تغییر ابعاد به اندازه ٦٤×٦٤ پیکسل پیش‌پردازش شده‌اند. بردار مشاهده نهایی که ابعاد ۲۱ بعدی دارد و برای پایداری آموزش و تعمیم بهتر با حداکثر مقادیر ممکن نرمال‌سازی شده‌اند، از دو دسته داده سطح پایین و سطح بالا ساخته می‌شود.

داده‌های سطح پایین شامل شتاب خطی و سرعت زاویه‌ای CE، هر کدام در سه محور مختصاتی با استفاده از حسگر اینرسی، جریان سیم‌پیچ‌های الکترومغناطیسی است. از سوی دیگر، داده‌های سطح بالا از پردازش این ورودی‌های خام به دست می‌آیند که شامل جهت‌گیری کپسول (محاسبه‌شده از طریق انتگرال‌گیری کواترنیون سرعت زاویه‌ای در چهار بعد)، سرعت خطی کپسول (محاسبه‌شده از طریق انتگرال‌گیری از شتاب خطی در سه بعد)، میزان اطمینان از نقطه هدف (محاسبه‌شده در الگوریتم تخمین نقطه هدف)، بردار نرمال‌شده سه‌بعدی جهت هدف تخمینی نسبت به موقعیت فعلی کپسول، میزان هم‌جهتی سرعت خطی کپسول و همچنین جهت‌گیری سر کپسول با راستای بردار هدف (با اعمال تبدیل نمایی برای تقویت حساسیت در نزدیکی آستانه‌های بحرانی) هستند. این ترکیب داده‌ها به عامل هوش مصنوعی اجازه می‌دهد تا هم وضعیت لحظه‌ای خود و هم روابط پویا با محیط را به‌طور دقیق درک کند.

۳-٦ تابع پاداش

تابع پاداش که بر اساس داده‌های حسگر اینرسی، دوربین و نقطه هدف مورد نظر طراحی شده است، عامل را برای انجام اقدامات بهینه تشویق می‌کند و از سه جزء کلیدی زیر تشکیل شده است:

[1] Maxwell
[2] Helmholtz
[3] Inertial Measurement Unit

۳-۷ حرکت و هم‌جهتی به سمت نقطه هدف

برای فعال کردن ردیابی مؤثر هدف، پاداش حرکت به سمت هدف (MTT)[1] طراحی شده است تا جهت حرکت عامل را به سمت نقطه هدف هدایت کند. همچنین این استراتژی از یک پاداش تکمیلی استفاده می‌کند که به عامل به‌منظور هم راستا شدن به سمت هدف پاداش می‌دهد و جهت‌گیری به‌سوی هدف (FTT)[2] نام‌گذاری می‌شود. به عبارتی این پاداش علاوه بر حرکت دادن عامل به سمت نقطه هدف، انگیزه‌ای برای صفر شدن زاویه حمله آن برای ردیابی دقیق‌تر نیز فراهم می‌کند. پاداش‌ها به‌صورت ریاضی به‌این‌ترتیب بیان می‌شوند:

$$R_{MTT} = \overline{v_{cap}} \cdot \overline{d_{ctt}} \quad (۵)$$

$$R_{FTT} = \overline{d_{for}} \cdot \overline{d_{ctt}}. \quad (۶)$$

که مقدار اسکالر R_{MTT} پاداش حرکت به سمت هدف، بردار $\overline{v_{cap}}$ سرعت خطی عامل و بردار $\overline{d_{ctt}}$، بردار نرمال شده جهت موقعیت عامل به سمت موقعیت نقطه هدف را نشان می‌دهد. زاویه بین این دو بردار (θ_1) نماینده میزان هم‌جهتی حرکت CE به سمت نقطه هدف است که این پاداش مقدار این زاویه را به حداقل می‌رساند. همچنین مقدار اسکالر R_{FTT} نشان‌دهنده پاداش جهت‌گیری به‌سوی هدف است، $\overline{d_{for}}$ نیز بردار جهت رو به جلو نرمال شده CE است و زاویه بین این دو بردار (θ_2) نماینده زاویه حمله است. هر دو عبارت R_{MTT} و R_{FTT} به عنوان حاصل ضرب داخلی بردارها، متناسب با $\cos(\theta)$ که در حالتی‌که θ به صفر نزدیک می‌شود، عملکرد ضعیفی دارد زیرا حساسیت $\cos(\theta)$ به تغییرات θ نزدیک به صفر کاهش می‌یابد.

برای رفع این محدودیت، یک تابع تبدیل برای افزایش حساسیت اعمال می‌شود:

$$T(x) = a_T * \left(1 - \exp\left(b_T * (x - c_T)\right)\right) \quad (۷)$$

که در آن مقدار ضرایب a_T، b_T و c_T به‌ترتیب ۱.۰۲۱۳-، ۲.۲۷۵۴ و ۰.۶۹۹۹ است. درنهایت پاداش ترکیبی برای حرکت به سمت هدف با زاویه حمله صفر (R_{MFTT}) به‌صورت زیر تعریف می‌شود:

$$R_{MFTT} = 0.5 * T(R_{MTT}) + 0.5 * T(R_{FTT}) \quad (۸)$$

که این رابطه تضمین می‌کند پاداش همیشه در محدوده $[-1,1]$ باقی می‌ماند.

۳-۸ جلوگیری از برخورد

برای تشویق رفتار ایمن CE، تابع پاداش، هرگونه تماس با روده بزرگ را جریمه می‌کند و بر اهمیت اجتناب از آسیب فیزیکی، به‌ویژه در کاربردهای دنیای واقعی تاکید می‌کند. این تابع پاداش که به‌صورت R_{col} مشخص می‌شود، به‌صورت زیر تعریف می‌شود:

$$R_{col} = \begin{cases} -1 & \text{اگر برخورد صورت گیرد،} \\ -0.5 & \text{اگر کپسول پس از برخورد ثابت بماند،} \\ 0 & \text{در غیر این صورت.} \end{cases} \quad (۹)$$

عبارت R_{col} یک تابع پاداش منفی است که در لحظه برخورد یک جریمه فوری را اعمال می‌کند و اگر CE نتواند پس از آن حرکت اصلاحی را نشان دهد، جریمه ثانویه را تحمل می‌کند. این طراحی تشویق می‌کند که عامل پس از برخورد، فعالانه مانور دهد تا از تله‌های احتمالی فرار کند و ناوبری بدون برخورد حاصل شود.

۳-۹ پاداش کل

تابع پاداش کلی با ترکیب خطی هر دو جزء، به‌صورت زیر محاسبه می‌شود:

$$R_{Total} = R_{MFTT} + R_{col} \quad (۱۰)$$

برخلاف روش‌های پیشین مانند [۹] و [۱۵] که با محدودیت‌های پیاده‌سازی در عمل مواجه بودند، تابع پاداش پیشنهادی در این پژوهش کاملاً کاربردی و قابل پیاده‌سازی در شرایط واقعی است. در مطالعات مذکور، تابع پاداش بر اساس نرخ پوشش سطح معده توسط CE طراحی شده و سعی در هدایت عامل برای پوشش کامل محیط معده داشت. اما این رویکرد در عمل با دو چالش اساسی روبروست: اول آن‌که داده‌های مربوط به موقعیت رئوس شبکه‌بندی روده در محیط واقعی در دسترس نیست، و دوم آن‌که سنجش میزان پوشش این شبکه‌بندی نیز مشکل‌آفرین است. این محدودیت‌ها باعث می‌شود روش‌های قبلی برای کاربردهای عملی مناسب نباشند.

[1] Moving Toward Target

[2] Facing Toward Target

در هر اپیزود[1]، اگر پاداش انباشته‌شده عامل به مقدار قابل توجهی منفی شود، با افزودن یک پاداش منفی نسبتاً بزرگ برای جریمه کردن رفتارهای ناکارآمد، سیگنال پایان اپیزود فعال می‌شود. علاوه‌براین، هنگامی که اختلاف بین کل گام‌های اجرا شده و گام‌های با R_{MFTT} مثبت از یک آستانه از پیش تعیین‌شده نیز فراتر رود، با افزودن جریمه‌ای به تابع پاداش متناسب با این اختلاف، سیگنال خاتمه اپیزود صادر می‌گردد. همچنین اگر تعداد کل گام‌ها در یک اپیزود از یک حدی بیشتر شود، آن اپیزود با یک پاداش مثبت نسبتاً بزرگ برای تشویق عملکرد به پایان می‌رسد. پس از هر اپیزود، هم عامل و هم محیط به شرایط اولیه بازنشانی می‌شوند تا اطمینان حاصل شود که اپیزودهای بعدی با شرایط تصادفی جدید شروع می‌شوند.

٤- آزمایش و نتایج

در این پژوهش، آموزش عامل در محیط مجازی با روش یادگیری تقویتی PPO انجام شده است. در هر اپیزود، فرآیند آموزشی با حالت‌های اولیه با استفاده از توزیع تصادفی یکنواخت در اطراف ایستگاه شروع CE، در یک فضای کاری محدود پراکنده شده‌اند انجام می‌شود که منجر به کاهش سوگیری شروع از یک حالت اولیه می‌شود. این آموزش‌ها بر روی پردازنده Intel Core(TM) i7-10510U با ١٦ گیگابایت حافظه رم و پردازنده گرافیکی Nvidia GeForce MX250 انجام شده است و تقریباً هر آموزش ۷ ساعت به‌طول انجامید.

برای سنجش عملکرد روش‌های کنترلی، برخی پژوهش‌ها [١٦] با در نظر گرفتن مسیر فرضی از پیش تعیین‌شده معیار ردیابی آن را اندازه‌گیری و گزارش می‌کنند. اما این مسیر بهینه از پیش تعیین‌شده در عمل وجود ندارد لذا در این پژوهش، الگوریتم ناوبری خودگردان مسیر مطلوب را به‌صورت خودکار در طول زمان معین کرده و با کنترل CE سعی در ردیابی آن دارد. به‌منظور ارزیابی و تحلیل عملکرد الگوریتم پیشنهادی، شاخص‌های کلیدی حاصل از اجرای آن در طول زمان، با استفاده از پیکربندی مبتنی بر اَبَرمتغیرهای تعیین‌شده در جدول (۱)، در پنج نسخه مختلف از الگوریتم PPO مورد بررسی قرار گرفته‌اند که از جمله مهم‌ترین آن‌ها اندازه دسته[2] و اندازه بافر[3] می‌باشند. این پارامترها نقش تعیین‌کننده‌ای در پایداری و کارایی فرآیند یادگیری دارند.

جدول (۱): مقدار اَبَرمتغیرها

Trainer	PPO1	PPO2		PPO3	PPO4	PPO5
Batch Size	1024	512		256	128	256
Buffer Size	5120	5120		5120	2560	2560
Learning Rate	0.0003					
Hidden Units	256					
Beta	۰/۰۰۵					
Epsilon	۰/۲					
Lambd	۰/۹۵					

اندازه دسته: انتخاب اندازه دسته تأثیر قابل توجهی بر عملکرد مدل و سرعت همگرایی دارد. آزمایش‌های انجام‌شده با اندازه دسته‌های ١٢٨، ٢٥٦، ٥١٢ و ١٠٢٤ نشان داد که اندازه دسته ١٢٨ سریع‌تر همگرا می‌شود، زیرا دسته‌های کوچک‌تر به‌روزرسانی‌های گرادیان سریع‌تری را ارائه می‌دهند، هرچند که ممکن است دقت کافی را نداشته باشند. این یافته با نتایج مطالعه مولر و همکاران آن [١٧] هماهنگ است که بیان می‌دارد یک اندازه دسته بهینه برای کمینه کردن زمان آموزش وجود دارد. به عبارتی دیگر از یک سو کاهش اندازه دسته به پایین‌تر از این حد بهینه، به‌دلیل گستردگی فضای سیاست، زمان آموزش را افزایش می‌دهد، از سوی دیگر، افزایش آن نیز انعطاف‌پذیری عامل را محدود کرده و لذا منجر به طولانی شدن فرآیند آموزش می‌شود.

اندازه بافر: اندازه بافر مشخص‌کننده مقدار داده‌های تجربه‌شده (حالت-عمل-پاداش) است که پیش از به‌روزرسانی سیاست از همه عامل‌های موازی جمع‌آوری می‌شوند. در آموزش با مدل PPO، تأثیر استفاده از اندازه بافرهای ۵/۱۲۰ و ۲/۵٦۰ بررسی شده است. اندازه بافر بزرگ منجر به افزایش فاصله بین به‌روزرسانی‌های

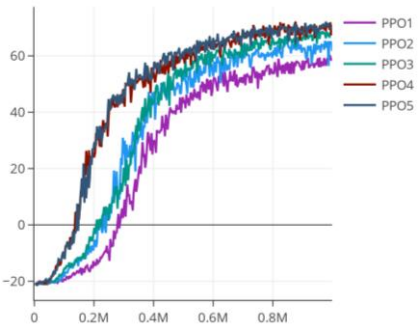
[1] Episode
[2] Batch Size
[3] Buffer Size

سیاست شد که سرعت آموزش را بسیار کاهش می‌دهد. به‌هرحال، اندازه بافر کوچک نیز همگرایی را مختل می‌کند.

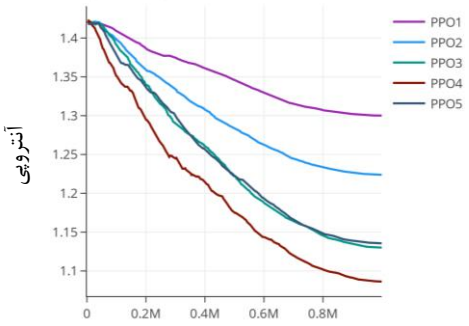
فرآیند یادگیری عامل در محیط در یک میلیون گام انجام شده است و تغییرات دو شاخص کلیدی پاداش کل و آنتروپی در طول زمان در دو شکل (۵) و (٦) رسم شده است.

همان‌طور که مشاهده می‌شود همه الگوریتم‌ها در شروع یادگیری، پاداش کل منفی کسب می‌کنند ولی در انتها به مقدار ثابتی از پاداش مثبت همگرا می‌شوند که این نشانگر همگرایی الگوریتم PPO است. مطابق نمودار شکل (۵)، الگوریتم تا گام ۲۰۰ هزار در فاز کاوش و سپس تا گام ۸۰۰ هزار در فاز همگرایی و سرانجام تا انتهای فرآیند آموزش (گام ۱ میلیون) در فاز بهره‌گیری عمل می‌کند. اما در مقایسه نسخه‌های مختلف آن، دو الگوریتم PPO4 و PPO5 سرعت همگرایی بیشتری دارند به‌طوری‌که در ۳۰۰ هزار گام اول به ۸۰ درصد از مقدار نهایی پاداش کل همگرا شده (یعنی نزدیک مقدار ۵۳) می‌رسند.

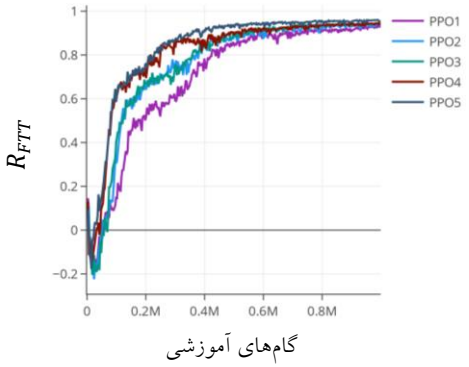
به‌علاوه مقدار نهایی همگرا شده تابع پاداش کل به دست آمده برای الگوریتم‌های PPO4 و PPO5 نیز در مقایسه با نتایج الگوریتم‌های دیگر بیشتر است. این برتری نسبی برای دو الگوریتم PPO4 و PPO5 در شاخص‌های دیگر نیز به وضوح قابل مشاهده است. در دو شکل (۷) و (۸) تغییرات دو معیار R_{FTT} و R_{MTT} که به‌ترتیب بیان‌گر پاداش حرکت به سمت هدف و پاداش جهت‌گیری به‌سوی هدف هستند و تشکیل‌دهنده پاداش کل می‌باشند، در طول زمان آموزش نشان داده شده است که هر دو به مقدار مطلوب ۱ همگرا شده‌اند.



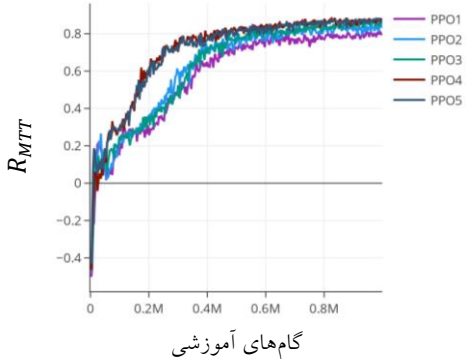
شکل (۵): پاداش جمع‌شونده کل



شکل (٦): آنتروپی



شکل (۷): پاداش جهت‌گیری به‌سوی هدف



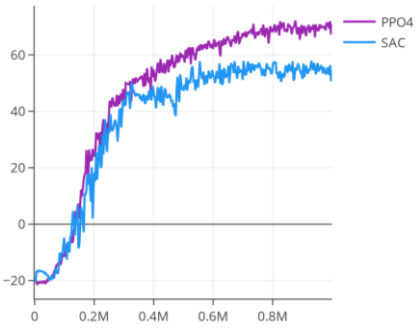
شکل (۸): پاداش حرکت به سمت هدف

همان‌طور که در قسمت پاداش توضیح داده شده است، مقدار نزدیک به ۱ برای پاداش حرکت به سمت هدف به معنی هم‌راستایی بردار سرعت حرکت به سمت هدف و برای پاداش جهت‌گیری به‌سوی هدف به معنی صفر بودن زاویه حمله (θ_2) است.

در این راستا، استفاده از ترکیب خطی پاداش‌های R_{MTT} و R_{FTT} با ضرایب وزنی مساوی (۰/۵) در پاداش کل، به تعادل مناسبی بین سرعت حرکت و دقت جهت‌گیری منجر شده است. به عبارتی، در

حالی‌که R_{MTT} حرکت کلی به سمت هدف را تشویق می‌کند، R_{FTT} با تمرکز بر کاهش خطای زاویه حمله، به عامل کمک می‌کند تا با دقت بیشتری محور خود را برای ردیابی هدف تنظیم کند. در مقایسه این دو معیار، دو الگوریتم PPO4 و PPO5 هم سرعت همگرایی بیشتر هم مقدار نهایی بالاتری نسبت به الگوریتم‌های دیگر کسب کرده‌اند. همان‌طور که از مقایسه شاخص کلیدی پاداش مشخص است دو الگوریتم PPO4 و PPO5 نسبت به نسخه‌های دیگر عملکرد بهتری دارند. با این حال، این دو الگوریتم در مقایسه بر اساس شاخص کلیدی دوم یعنی آنتروپی اختلاف معناداری دارند.

همان‌طور که در شکل (٦) نشان داده شده است نسخه PPO4 در طول زمان یادگیری به سطح پایین‌تری از معیار آنتروپی دست پیدا کرده است. در یادگیری تقویتی، کاهش آنتروپی بیانگر کاهش تصادفی بودن سیاست و گرایش به بهره‌برداری[1] از مدل یادگرفته شده است. آنتروپی پایین به‌تنهایی ضامن موفقیت نیست، این کاهش تنها زمانی مطلوب است که با عملکرد بهتر (پاداش بالاتر) همراه باشد، زیرا نشان‌دهنده همگرایی به سیاستی قطعی و مؤثر است. بنابراین، آنتروپی همواره باید همراه با ارزیابی عملکرد کلی سیستم مورد تحلیل قرار گیرد. لذا الگوریتم PPO4 توانسته است در عین بروز عملکرد مناسب در یادگیری تابع پاداش، آنتروپی پایین‌تری نیز داشته باشد.



شکل (۹): پاداش جمع‌شونده کل (مقایسه PPO4 و SAC)

برای مشخص شدن اعتبار عملکرد الگوریتم PPO4 در مقایسه با سایر روش‌های یادگیری تقویتی عمیق، مسئله ناوبری خودگردان کپسول آندوسکوپی با روش SAC[2] نیز پیاده‌سازی شده و مورد مقایسه قرار گرفته است. الگوریتم SAC به دلیل بهره‌گیری از اصل بیشینه‌سازی آنتروپی قدرت بالایی در اکتشاف دارد به‌طوری‌که از هر تعامل با محیط بیشتر یاد می‌گیرد. به همین دلیل نتایج شکل (۹) شده نشان می‌دهد، الگوریتم SAC توانسته سرعت خوبی در همگرایی داشته باشد. با این حال، الگوریتم PPO4 نتیجه بهتری را رقم زده و مشخص است که سرعت بیشتری در همگرایی دارد. بعلاوه الگوریتم PPO4 به مقدار پاداش کل بیشتری نسبت به الگوریتم SAC دست پیدا کرده است.

۵- نتیجه‌گیری

در این مطالعه، یک چارچوب جدید برای کنترل و ناوبری خودگردان یک کپسول درون‌بین در روده بزرگ ارائه شد که با بهره‌گیری از روش یادگیری تقویتی عمیق، امکان تولید و ردیابی مستقل مسیرهای مطلوب توسط کپسول درون‌بین با کنترل جریان عملگرهای آرایه‌ای الکترومغناطیسی، فراهم گردید. در این چارچوب، داده‌های حاصل از انواع حسگر کپسول درون‌بین (شامل دوربین، حسگر اینرسی و حسگر جریان الکتریکی) به‌وسیله روش ارائه‌شده با هم ادغام شدند که مجموعه مشاهدات عامل را در دو لایه داده‌های سطح پایین شامل داده‌های بدون پردازش حسگرها و داده‌های سطح بالا شامل داده‌های پردازش شده و همچنین الگوریتم تخمین نقطه هدف سه‌بعدی تشکیل می‌دادند. علاوه‌براین، برای پیاده‌سازی روش یادگیری تقویتی عمیق، یک محیط مجازی نزدیک به واقعیت فراهم شد که امکان آموزش مؤثر عامل در شرایط متنوع و شبیه‌سازی‌های دقیق سه‌بعدی را میسر ساخت. کپسول درون‌بین به‌وسیله روش ناوبری خودگران ارائه‌شده، با ترکیب الگوریتم تخمین نقطه هدف در فضای سه‌بعدی و الگوریتم لبه تکنولوژی بهینه‌سازی مجاور به‌وسیله مهندسی تابع پاداش، مسیر خود را یافت و با موفقیت آن را دنبال کرد. تابع پاداش طراحی‌شده به‌صورت هم‌زمان جهت حرکت کپسول درون‌بین و زاویه حمله آن را در طول مسیر تنظیم کرد و همچنین به‌منظور جلوگیری از آسیب دیواره به روده هرگونه برخورد را نیز جریمه کرد. برای ارزیابی

[1] Exploitation

[2] Soft Actor-Critic

موفقیت ناوبری خودگران دو شاخص کلیدی تابع پاداش شامل عناصر سازنده آن و آنتروپی مورد سنجش قرار گرفتند که از بین اَبرمتغیرهای مختلف مورد آزمایش، اندازه دسته ۱۲۸ و اندازه بافر ۲۵۶۰ که متناظر با نسخه PPO4 بود بهترین نتیجه را به همراه داشت. در ارزیابی نتایج، مشخص شد که این الگوریتم در کنار ارائه بهترین عملکرد نسبت به سایر نسخه‌ها، قطعیت در اجرای عمل بهینه در هر حالت را نیز افزایش داد. این پژوهش با ارائه ناوبری خودگردان کپسول درون‌بین فرصت‌های جدیدی را برای پزشکان غیر متخصص جهت انجام کولونوسکوپی با اطمینان و دقت بیشتر را به وجود می‌آورد. هدف اصلی این رویکرد کاهش پیچیدگی فرآیند طراحی، ساخت و به‌کارگیری کپسول درون‌بین توسط اپراتور است که با کاهش بار کاری عمل درون‌بینی، امکان تمرکز بیشتر بر تصمیم‌گیری‌های بالینی و جنبه‌های درمانی را فراهم آورد.

جهت‌های تحقیقاتی آینده شامل اصلاح تابع پاداش به‌منظور سایر اهداف، مانند ذخیره‌سازی انرژی مصرفی توسط سیستم الکترومغناطیسی عملگر، پرداختن به چالش ایمنی حرکت کپسول درون‌بین در حین آموزش و اجرای الگوریتم ناوبری خودگردان در واقعیت و همچنین اعتبارسنجی سیستم کنترل کپسول درون‌بین از طریق آزمایش‌های بالینی است. با پرداختن به این چالش‌ها، این فناوری پتانسیل ایجاد انقلابی در روش‌های درون‌بینی را دارد که درنهایت منجر به افزایش میزان عمل‌های تشخیصی درون‌بینی در جامعه برای پیشگیری از سرطان می‌شود.

References

[1] Bretthauer, M., et al., Effect of colonoscopy screening on risks of colorectal cancer and related death. New England Journal of Medicine, 2022. 387(17): p. 1547-1556.

[2] Scaglioni, G., et al., Facing the emotional barriers to colorectal cancer screening. The roles of reappraisal and situation selection. International journal of behavioral medicine, 2024: p. 1-10.

[3] Cao, Q., et al., Robotic wireless capsule endoscopy: recent advances and upcoming technologies. Nature Communications, 2024. 15(1): p.٤٥٩٧ .

[4] Ali, M.A., et al., Recent Advancements in Localization Technologies for Wireless Capsule Endoscopy: A Technical Review. Sensors (Basel, Switzerland), 2025. 25(1): p. 253.

[5] Hoang, M.C., et al., Independent electromagnetic field control for practical approach to actively locomotive wireless capsule endoscope. IEEE TranSACtions on Systems, Man, and Cybernetics: Systems, 2019. 51(5): p. 3040-3052.

[6] Zhang, H., et al., Towards Automatic Stomach Screening Using a Wireless Magnetically Actuated Capsule Endoscope. IEEE TranSACtions on Medical Robotics and Bionics, 2024.

[7] Martin, J.W., et al., Enabling the future of colonoscopy with intelligent and autonomous magnetic manipulation. Nature machine intelligence, 2020. 2(10): p. 595-606.

[8] Pore, A ,.et al. Colonoscopy navigation using end-to-end deep visuomotor control: A user study. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2022.

[9] Zhang, Y., et al. Deep reinforcement learning-based control for stomach coverage scanning of wireless capsule endoscopy. IEEE International Conference on Robotics and Biomimetics (ROBIO). 2022.

[10] Iriondo, A., et al., Pick and place operations in logistics using a mobile manipulator controlled with deep reinforcement learning. Applied Sciences, 2019. 9(2): p. 348.

[11] Juliani, A., et al., Unity: A general platform for intelligent agents. arXiv preprint arXiv:1809.02627, 2018.

[12] Cheliotis, K., ABMU: an agent-based modelling framework for Unity3D. SoftwareX, 2021. 15: p. 100771.

[13] Schulman, J., et al., Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017.

[14] Tao, X., et al., A Fast and Robust Camera-IMU Online Calibration Method For Localization System. arXiv preprint arXiv:2305.08247, 2023.

[15] İncetan, K., et al., VR-Caps: a virtual environment for capsule endoscopy. Medical image analysis, 2021. 70: p. 101990.

[16] Xu, Y., et al., Trajectory following strategies for wireless capsule endoscopy under reciprocally

rotating magnetic actuation in a tubular environment. arXiv preprint arXiv:2108.11620, 2021.

[17] Müller, A., F. Grumbach, and M. Sabatelli. Smaller Batches, Bigger Gains? Investigating the Impact of Batch Sizes on Reinforcement Learning Based Real-World Production Scheduling. IEEE 29th International Conference on Emerging Technologies and Factory Automation (ETFA). 2024.

Journal of Applied And Basic Machine Intelligence Research (ABMIR)

Applied and Basic Machine Intelligence Research

Vol. 3, No. 1, Spring & Summer 2025, pp. 1 - 15

A novel autonomous navigation of capsule endoscopy based on deep reinforcement learning with proximal policy optimization in a virtual environment

Hamidreza Ghahremani [1], Vahid Johari Majd [2*]

[1]Ph.D. student, School electrical and computer Engineering, Tarbiat Modares University, Tehran, Iran

[2]Associate professor, School of electrical and computer Engineering, Tarbiat Modares University, Tehran, Iran

Article Information

Original Research Paper

Received:
2025 April 27

Accepted:
2025 June 19

Keywords:
Autonomous navigation, capsule endoscopy, deep reinforcement learning, neighborhood policy optimization, virtual environment

Corresponding Author*:
majd@modares.ac.ir

Abstract

Given widespread concerns about the complications associated with traditional endoscopy, research into the use of less invasive endoscopic capsules has gained significant attention. However, the passive movement of these capsules often prevents access to specific angles and areas of interest to the clinician. To overcome this limitation, we propose a novel autonomous navigation approach based on deep reinforcement learning, utilizing a proximal policy optimization (PPO) algorithm. This method automates the capsule's positioning, pathfinding, and motion control. Our approach integrates multi-modal sensor data to estimate the target point over time. Recognizing that initial algorithm training requires substantial data, we developed a near-realistic virtual environment. This environment facilitates the training of an intelligent agent and includes an actuator model with a magnetic coil structure, a capsule equipped with a dipole magnet, a camera, an inertial sensor, and a 3D model of the large intestine. The primary objective of this research is to reduce operator intervention, allowing clinicians to focus more on the clinical and medical aspects of endoscopy. The proposed method was trained with various hyperparameters, and its performance was evaluated based on metrics such as movement and alignment toward the target, as well as entropy. The evaluation results demonstrate that by optimally adjusting the buffer size and batch size, the algorithm achieves effective tracking and stability.

: 10.22034/ABMIR.2025.23028.1122

E-ISSN: 2821-2037