Научная статья
УДК 517.977.5
URL: https://trudymai.ru/published.php?ID=187462
EDN: https://www.elibrary.ru/GYVJQI

ПРЕДОТВРАЩЕНИЕ ВОЗДУШНЫХ СТОЛКНОВЕНИЙ С ИСПОЛЬЗОВАНИЕМ РЕКУРРЕНТНОГО ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ С УЧЕТОМ ЗАДЕРЖКИ РЕАКЦИИ ПИЛОТА

Л. Цзочэн ✉

Северо-Западный политехнический университет, г. Сиань, Китай
✉ liuzuocheng@mail.nwpu.edu.cn

Цитирование: Цзочэн Л. Предотвращение воздушных столкновений с использованием рекуррентного обучения с подкреплением с учетом задержки реакции пилота // Труды МАИ. 2026. № 146. URL: https://trudymai.ru/published.php?ID=187462

Аннотация. Системы предотвращения воздушных столкновений имеют критическое значение для обеспечения безопасности полетов, особенно в условиях роста воздушного движения. В то время как традиционные системы, такие как система предупреждения столкновения самолетов в воздухе (TCAS,) предлагают решения, основанные на процессах принятия решений Маркова (MDP), эти модели не учитывают важные реальные факторы, такие как задержки реакции пилота. В данной работе мы формулируем задачу предотвращения воздушных столкновений как частично наблюдаемый марковский процесс принятия решений (POMDP), чтобы решить проблемы, вызванные задержками реакции пилота. Для решения полученной задачи POMDP мы применяем алгоритм Long Short-Term Memory Soft Actor-Critic дискретный (LSTM SAC-d), который расширяет фреймворк Soft Actor-Critic дискретный (SAC-d) за счет включения временных зависимостей. Наш модель учитывает задержку реакции пилота в 3 секунды, что отражает реальные ограничения. Мы сравниваем производительность LSTM SAC-d с марковским SAC-d и демонстрируем, что LSTM SAC-d значительно превосходит последний по эффективности предотвращения столкновений и общей стабильности решений. Экспериментальные результаты

показывают, что LSTM SAC-d значительно улучшает работу системы за счет лучшего учета задержек реакции пилота и оптимизации рекомендаций в реальном времени.

Ключевые слова: обучение с подкреплением, учет реакции пилота, предотвращение воздушных столкновений, модель столкновения воздушных судов, динамическая модель самолета

COLLISION AVOIDANCE IN AIR TRAFFIC USING RECURRENT REINFORCEMENT LEARNING ACCOUNTING FOR PILOT REACTION DELAY

L. Zuocheng ✉
Northwestern Polytechnical University, Xi'An, P. R. China
✉ liuzuocheng@mail.nwpu.edu.cn

Citation: Zuocheng L. Collision avoidance in air traffic using recurrent reinforcement learning accounting for pilot reaction delay // Trudy MAI. 2026. No. 146. (In Russ.). URL: https://trudymai.ru/published.php?ID=187462

Abstract. Air collision avoidance systems are critical for flight safety, especially in growing air traffic. While traditional systems such as Traffic Collision Avoidance System (TCAS) offer solutions based on Markov Decision Processes (MDPs), these models do not account for important real-world factors such as pilot reaction delays. In this work, we formulate the air collision avoidance problem as a partially observed Markov Decision Process (POMDP) to solve problems caused by pilot response delays. To solve the resulting POMDP problem, we use the Long Short-Term Memory Soft Actor-Critic discrete (LSTM SAC-d) algorithm, which extends the Soft Actor-Critic discrete (SAC-d) framework by including time dependencies. Our model takes into account a pilot reaction delay of 3 seconds, which reflects real limitations. We compare the performance of LSTM SAC-d with Markov's SAC-d and demonstrate that LSTM SAC-d significantly outperforms the latter in collision avoidance efficiency and overall solution stability. Experimental results show that the SAC-d LSTM significantly improves system performance by better accounting for pilot response delays and optimizing recommendations in real time.

Keywords: reinforcement learning, pilot response influence, air collision avoidance, aircraft collision model, dynamic aircraft model

Введение

Системы предотвращения воздушных столкновений имеют критическое значение для обеспечения безопасности полетов, особенно в условиях роста воздушного движения. Хотя традиционные системы, такие как система предотвращения столкновений на воздушных трассах (TCAS), значительно улучшили предотвращение столкновений в воздухе, они часто предполагают немедленную реакцию пилота на рекомендации по предотвращению столкновения1. На практике реакция пилота часто задерживается из-за таких факторов, как когнитивная нагрузка, стресс и внешние отвлекающие факторы. Задержка в 3-5 секунд может значительно увеличить риск столкновения, как это было доказано несколькими громкими происшествиями, где человеческая ошибка сыграла ключевую роль2. Учитывая важность своевременных решений при предотвращении столкновений, крайне важно разработать системы, которые могут учитывать и адаптироваться к этим задержкам. Однако традиционные системы с трудом предсказывают и минимизируют влияние задержанных человеческих реакций.

Искусственный интеллект (ИИ), такой как обучение с подкреплением (RL), модельно-предсказательный контроль (MPC), поиск Монте-Карло по дереву решений (MCTS) и динамическое программирование (DP), продемонстрировал перспективность в оптимизации принятия решений в сложных динамичных средах, таких как предотвращение столкновений34. Хотя подходы на основе модели RL превосходят в контролируемых условиях, они сталкиваются с трудностями в неопределенных, динамичных ситуациях5. Модели без модели, такие как глубокие Q-сети (DQN) и алгоритмы градиента политики, более адаптируемы, но страдают от медленной сходимости и высоких вычислительных затрат5. MPC и DP предлагают оптимизацию в реальном времени, но сталкиваются с трудностями из-за больших пространств состояний и сложной динамики67.

Несмотря на свои преимущества, эти методы имеют ограничения при обработке частичной наблюдаемости и задержанных откликов, характерных для предотвращения столкновений, что делает их менее подходящими для реальных приложений в авиации. С другой стороны, частично наблюдаемые марковские

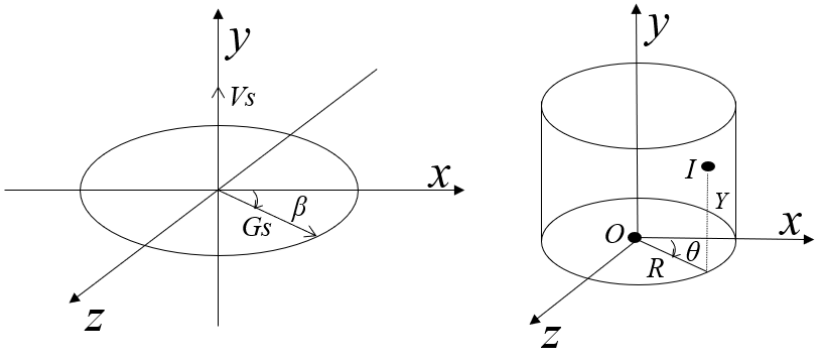
процессы принятия решений (POMDP) предлагают надежную структуру для моделирования таких сред, учитывая частичные наблюдения состояния и задержанные действия, что делает их хорошо подходящими для задач предотвращения столкновений 8.

В данной работе предлагается новый подход на основе POMDP для решения неопределенности, возникающей из-за задержек реакции пилота и частичной наблюдаемости в сценариях предотвращения столкновений. Расширив алгоритм SAC-d с использованием сетей LSTM, данная структура захватывает временные зависимости, что позволяет учитывать задержки реакции пилота и улучшать принятие решений в реальном времени.

Модель предотвращения столкновений между частично наблюдаемыми воздушными судами

Модель столкновения воздушных судов

В данном исследовании мы описываем трехмерное столкновение двух воздушных судов с использованием параметрической модели, которая представляет их динамику движения и пространственную геометрию. Эта модель служит основой для определения пространства состояний и действий, что облегчает разработку стратегий предотвращения столкновений на основе обучения с подкреплением.



а б

Рисунок 1. а - представление скорости воздушного судна;
б - относительное положение нарушителя (*I*) относительно собственного судна (*O*) на момент минимального расстояния (CPA)

Как показано на рисунке 1а скорость каждого воздушного судна характеризуется тремя ключевыми параметрами: наземной скоростью *Gs*, углом направления β и вертикальной скоростью *Vs*. Для заданного воздушного судна (нарушителя) его компоненты скорости в декартовых координатах выражаются следующим образом:

$$\begin{bmatrix} Vx_i \\ Vy_i \\ Vz_i \end{bmatrix} = \begin{bmatrix} Gs_i * \cos(\beta_i) \\ Vs_i \\ Gs_i * \sin(\beta_i) \end{bmatrix} \quad (1)$$

где Vx_i, Vy_i и Vz_i представляют собой компоненты скорости в направлениях *x*-, *y*- и *z*- соответственно.

При моделировании относительного движения между двумя воздушными судами предполагается, что начальное состояние собственного воздушного судна задано его скоростью $[Gs_o\ ,\ \beta_o\ ,\ Vs_o]^T$ и положением $[x_o\ ,\ y_o\ ,\ z_o]^T$. Положение нарушителя в момент времени *T* относительно собственного воздушного судна можно выразить следующим образом:

$$\begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix} = \begin{bmatrix} x_o \\ y_o \\ z_o \end{bmatrix} + \begin{bmatrix} Gs_o * \cos(\beta_o) \\ Vs_o \\ Gs_o * \sin(\beta_o) \end{bmatrix} * T + \begin{bmatrix} R * \cos(\theta) \\ Y \\ R * \sin(\theta) \end{bmatrix} - \begin{bmatrix} Vx_i \\ Vy_i \\ Vz_i \end{bmatrix} * T \quad (2)$$

где *R*, θ и *Y* определяют геометрию относительного смещения между двумя воздушными судами, когда нарушитель достигает ближайшей точки сближения (CPA) собственного воздушного судна, как показано на рисунке 1б, а *T* обозначает время до CPA.

Модель столкновения фокусируется на относительных состояниях между собственным воздушным судном и нарушителем, чтобы упростить описание проблемы, сохраняя при этом основные динамические характеристики для принятия решений по предотвращению столкновений. Закрепив начальное положение $[x_o\ ,\ y_o\ ,\ z_o]^T$ и угол направления β_o собственного воздушного судна,

относительная кинематика нарушителя может быть полностью описана компактным набором параметров: {Gs_o, Vs_o, T, R, θ, Y, Gs_i, β_i, Vs_i}.

Модель столкновения предоставляет компактное представление 3D-относительных динамик между двумя воздушными судами, захватывая ключевые временные и пространственные особенности, такие как время до конфликта и относственная геометрия в точке CPA. Несмотря на то, что модель описывает полное 3D-столкновение, это исследование фокусируется на вертикальном предотвращении столкновений, чтобы сохранить согласованность с рекомендациями TCAS, которые в основном касаются вертикального измерения 9. Этот переход упрощает задачу, сохраняя важную временную и вертикальную информацию, позволяя фреймворку обучения с подкреплением разрабатывать эффективные стратегии разрешения вертикальных конфликтов.

Модель динамики воздушного судна

Динамическая модель, описанная уравнением (3), предполагает шаг времени в одну секунду, что соответствует системе предотвращения столкновений, работающей с частотой обновления 1 Гц. Для увеличения сложности среды симуляции ускорение нарушителя ограничено в пределах [$-a_{\text{int}}$, a_{int}]. Более того, на каждом временном шаге нарушитель корректирует свою траекторию в сторону потенциального состояния столкновения на основе информации о скорости собственного воздушного судна, с учетом его допустимых ограничений по скорости и ускорению. Эта настройка вводит более сложные и реалистичные динамические характеристики для фреймворка обучения с подкреплением.

$$\begin{bmatrix} h \\ \dot{h}_{own} \\ \dot{h}_{\text{int}} \\ \tau \\ a_{prev} \end{bmatrix} \leftarrow \begin{bmatrix} h + \dot{h}_{\text{int}} + 0.5\ddot{h}_{\text{int}} - \dot{h}_{own} - 0.5\ddot{h}_{own} \\ \dot{h}_{own} + \ddot{h}_{own} \\ \dot{h}_{\text{int}} + \ddot{h}_{\text{int}} \\ \tau - 1 \\ a'_{prev} \end{bmatrix} \tag{3}$$

Модель принятия решений по предотвращению столкновений воздушных судов с задержкой реакции пилота

1. Моделирование задержки реакции пилота и структура принятия решений

При предотвращении столкновений пилоты проявляют фиксированную задержку реакции, обычно около 3 секунд, из-за когнитивных и операционных ограничений. Эта задержка вводит временное несоответствие между рекомендациями, выданными системой, и их выполнением, что значительно влияет на производительность в рамках традиционных моделей марковских процессов принятия решений (MDP).

MDP предполагает полностью наблюдаемую среду с немедленным выполнением действий, определенную как (S,A,P,R,γ). Однако в сценариях с задержкой отклика состояние изменяется в течение периода задержки, что приводит к субоптимальности решений агента.

Для решения этой проблемы мы применяем POMDP, который учитывает задержанную обратную связь и частичную наблюдаемость. POMDP определяется как (S,A,O,P,R,Z,γ), где O представляет собой множество возможных наблюдений, доступных агенту, а Z обозначает функцию наблюдения, которая отображает истинное состояние и действие на распределение вероятностей по наблюдениям. Агент наблюдает за задержанными состояниями из-за времени реакции пилота, поддерживает состояние убеждения для оценки истинного состояния системы и принимает решения, предсказывая влияние задержки на будущие состояния.

Интеграция фиксированной задержки реакции пилота (Δt=3 s) отражена в проектировании функции вознаграждения, которая оценивает рекомендации на основе их выполнения в момент времени $t+\Delta t$. Это гарантирует, что агент изучает стратегии, которые учитывают влияние задержки, что ведет к улучшению принятия решений и эффективности предотвращения столкновений. Результаты симуляции демонстрируют, что внедрение задержек в рамках POMDP значительно повышает безопасность и производительность системы по сравнению с методами стандартных MDP.

2. Пространство состояний

Пространство состояний для задачи предотвращения столкновений состоит из пяти переменных, как изложено в таблице 1 и показано на рисунке 2. Первые три переменные представляют собой относительные положения и вертикальные скорости собственного воздушного судна и нарушителя. Четвертая переменная, τ, захватывает горизонтальную геометрию, указывая время до того момента, когда горизонтальное разделение между двумя воздушными судами станет меньше 500 футов, что называется временем до конфликта. Наконец, включение предыдущей рекомендации в пространство состояний позволяет системе наказывать развороты или усиленные рекомендации, обеспечивая сохранение марковского свойства.

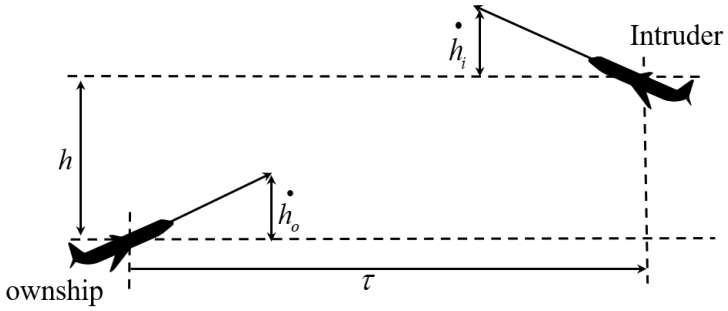


Рисунок 2 - Визуальное представление переменных состояния

Таблица 1

Переменные пространства состояний

Переменная	**Описание**	**Значения**	**Единицы измерения**
h	Относительная высота нарушителя	[-2500, 2500]	ft
$\dot{h}_o$	Вертикальная скорость собственного самолета	[-70, 70]	ft/s
$\dot{h}_i$	Вертикальная скорость нарушителя	[-50, 50]	ft/s
τ	Время до потери горизонтального разделения	[0, 40]	s
a_{prev}	Предыдущий совет	см. Таблица 2	-

3. Пространство действий

Пространство действий состоит из семи рекомендаций, которые система предотвращения столкновений может выдать во время полета, как изложено в таблице 2. Все рекомендации, кроме COC (Clear of Conflict), вызывают сигнал тревоги и направляют воздушное судно в определенный диапазон вертикальных скоростей с соответствующим ускорением. Рекомендация COC указывает, что немедленной угрозы столкновения со стороны нарушителя нет.

Таблица 3 описывает условия для выдачи каждой рекомендации на основе текущей рекомендации. Например, рекомендация COC может быть выдана в любой момент времени. Однако начальные рекомендации, такие как DES1500 и CL1500, могут быть выданы только при наличии текущей рекомендации COC. SDES1500 может следовать за такими рекомендациями, как CL1500, SCL1500 и SCL2500, действуя как разворот, или следовать за SDES2500, служащим ослаблением. Важно отметить, что SDES1500 не может напрямую следовать за COC или заменять DES1500, поскольку эти рекомендации имеют схожую природу.

Таблица 2

Набор советов

Действие	**Описание**	**Ускорение**
COC	Нет конфликта	0
DES1500	Спуск<=-25ft/s	-g/3
CL1500	Взлет>=25ft/s	g/3
SDES1500	Спуск<=-25ft/s	-g/2.5
SCL1500	Взлет>=25ft/s	g/2.5
SDES2500	Спуск<=-42ft/s	-g/2.5
SCL2500	Взлет>=42ft/s	g/2.5

Таблица 3

Доступность советов

Действие	**Доступно от**
COC	В любое время
DES1500	COC
CL1500	COC
SDES1500	CL1500, SCL1500, SCL2500, SDE2500
SCL1500	DES1500, SDE1500, SDES2500, SCL1500
SDES2500	DES1500, SDES1500
SCL2500	CL1500, SCL1500

4. Формирование вознаграждения

Для обеспечения безопасности и эффективности предотвращения столкновений воздушных судов функция вознаграждения разработана с целью сбалансировать предотвращение столкновений, минимизацию предупреждений и управление высотой. Функция вознаграждения разделена на два основных компонента: штраф за финальную высоту и управление предупреждениями и высотой.

а) Штраф за финальную высоту

Этот компонент оценивает относительную высоту между собственным воздушным судном и нарушителем в критические моменты, в частности, когда Время до Потери Горизонтального Разделения (TLOS) достигает нуля. Он штрафует небезопасные действия, такие как столкновения или значительные отклонения от безопасного диапазона высот, чтобы усилить эффективные стратегии предотвращения столкновений. Математически вознаграждение выражается следующим образом:

$$R_1 = R_{1,1} + R_{1,2} + R_{1,3} \tag{4}$$

где: штраф за столкновение

$$R_{1,1} = -\omega_{NMAC} 1\left\{|h_{rel}| \le h_{col}\right\} \tag{5}$$

штрафует сценарии, когда относительная высота h_{rel} между двумя воздушными судами падает ниже порога столкновения h_{col};

штраф за отклонение от высоты:

$$R_{1,2} = \omega_{NMAC}\left(\frac{|h_{rel}| - h_{col}}{h_{rel\_s} - h_{col}}\right) 1\left\{h_{col} \le |h_{rel}| < h_{rel\_s}\right\} \tag{6}$$

штрафует за умеренные отклонения относительной высоты от желаемого диапазона h_{rel_s};

штраф за чрезмерное отклонение высоты:

$$R_{1,3} = \max\left(-2\omega_{leav}\left(|h_{rel}| - h_{rel\_s}\right), -\gamma\right) 1\left\{|h_{rel}| \ge h_{rel\_s}\right\} \tag{7}$$

наложение штрафов за чрезмерные отклонения, где γ обозначает максимальный предел штрафа.

б) Управление предупреждениями и высотой

Этот компонент учитывает ограничения высоты собственного воздушного судна, выдачу предупреждений и эффективность разрешения конфликтов. Вознаграждение определяется следующим образом:

$$R_2 = R_{2,1} + R_{2,2} + R_{2,3} \tag{8}$$

где: штраф за пределы высоты:

$$R_{2,1} = -\omega_{leav} 1\left\{\left|h_{own}\right| > h_{upper}\right\} \tag{9}$$

штрафует за превышение высоты собственного воздушного судна выше заранее установленного верхнего предела h_{upper};

штраф за предупреждение:

$$R_{2,2} = -\left(\omega_{alert} + \omega_{reversal} + \omega_{strength} + \omega_{crossing}\right)\exp\left(t - tau\right) \tag{10}$$

штрафует за выдачу ненужных предупреждений (ω_{alert}), отмену рекомендаций ($\omega_{reversal}$), ненужное повышение степени серьезности предупреждений ($\omega_{strength}$) и ненужное пересечение высот воздушным судном ($\omega_{crossing}$). Эти штрафы увеличиваются экспоненциально по мере приближения времени до конфликта (tau);

награда за устранение конфликта:

$$R_{2,3} = \omega_{COC} \tag{11}$$

предоставляет небольшое положительное вознаграждение, когда конфликт успешно устранен.

Избежание столкновений воздушных судов с использованием рекуррентного обучения с подкреплением

Марковский SAC-дискретное для предотвращения столкновений

Алгоритм SAC-d, основанный на структуре процесса принятия решений Маркова (MDP), является популярным методом обучения с подкреплением для дискретных пространств действий. В MDP предполагается, что текущее состояние s_t содержит всю необходимую информацию для оптимального принятия решений,

что позволяет агенту принимать решения, основываясь только на текущем состоянии. Алгоритм SAC-d оптимизирует стохастическую политику $\pi(a|s)$, максимизируя комбинацию ожидаемого кумулятивного вознаграждения и энтропии политики, тем самым способствуя как эффективному исследованию, так и надежному принятию решений.

В модели предотвращения столкновений воздушных судов пространство состояний S включает критически важные параметры, такие как относительная высота, вертикальная скорость и время до конфликта между собственным судном и нарушителем. Пространство действий A состоит из дискретных рекомендаций, представляющих вертикальные маневры, включая «Подъем», «Спуск» и «Сохранить высоту». Функция вознаграждения $R(s_t,a_t)$ разработана для штрафования близких столкновений в воздухе (NMAC), при этом вознаграждаются безопасные разделения и эффективные разрешения конфликтов. Алгоритм итеративно оптимизирует мягкую Q-функцию $Q(s_t,a_t)$, которая обновляется с использованием уравнения Беллмана:

$$Q(s_t, a_t) = r_t + \gamma \mathrm{E}_{s_{t+1} \sim p(\cdot|s_t,a_t)} \left[\min_{i=1,2} Q_i(s_{t+1}, a') - \alpha \log \pi(a' | s_{t+1}) \right] \quad (12)$$

где r_t представляет вознаграждение в момент времениt, γ — коэффициент дисконтирования, а α — параметр температуры, который регулирует компромисс между вознаграждением и энтропией.

Хотя SAC-d демонстрирует высокую эффективность в принятии решений в полностью наблюдаемых средах, его зависимость от марковского предположения ограничивает его эффективность в реальных сценариях предотвращения столкновений. Задержанные реакции пилота, шум датчиков и частично наблюдаемые состояния вносят неопределенности, которые нарушают марковское свойство. В результате SAC-d не учитывает исторические зависимости, которые важны для точного принятия решений в таких условиях. Это ограничение служит мотивом для расширения SAC-d, которое явно учитывает временные зависимости и решает проблему частичной наблюдаемости.

LSTM SAC-d: Временное расширение SAC-дискретного

Для устранения ограничений Markovian SAC-d мы используем структуру LSTM SAC-d, которая интегрирует сети LSTM в SAC-d для захвата временных зависимостей и кодирования исторической информации. Модуль LSTM позволяет агенту принимать решения на основе последовательностей прошлых наблюдений, действий и вознаграждений, эффективно решая проблемы, возникающие в немарковских средах, где текущее состояние недостаточно для оптимального принятия решений 10.

В LSTM SAC-d входное состояние s_t расширяется для учета исторических зависимостей через скрытое состояние LSTM h_t, которое вычисляется как:

$$h_t = \mathrm{LSTM}\left(h_{t-1}, \left[s_t, a_{t-1}, r_{t-1}\right]\right) \tag{13}$$

где a_{t-1} и r_{t-1} — это действие и вознаграждение на предыдущем временном шаге. Сетевой политику генерирует стохастическую политику $\pi(a_t|h_t)$, которая прогнозирует вероятности действий на основе выхода LSTM. Аналогично, мягкие функции Q, $Q_1(h_t,a_t)$ и $Q_2(h_t,a_t)$, условно зависят от h_t , позволяя критику оценивать действия в контексте временной информации. Уравнение Беллмана для функции Q модифицируется следующим образом:

$$Q(h_t, a_t) = r_t + \gamma \mathrm{E}_{h_{t+1} \sim p(\cdot|h_t, a_t)}\left[\min_{i=1,2} Q_i(h_{t+1}, a') - \alpha \log \pi(a'|h_{t+1})\right] \tag{14}$$

Этот временной расширение позволяет агенту обрабатывать задержанные реакции, такие как фиксированная задержка реакции пилота в 3 секунды, смоделированная в этом исследовании. Используя последовательную информацию, закодированную в LSTM, LSTM SAC-d предсказывает последствия задержанных рекомендаций и адаптирует свою политику соответственно. Включение временной осведомленности не только улучшает стабильность принятия решений, но и повышает эффективность предотвращения столкновений. Экспериментальные результаты показывают, что LSTM SAC-d достигает значительно более высоких показателей успеха предотвращения столкновений, реже возникают NMAC и обеспечивается лучшая стабильность рекомендаций по сравнению с Марковским SAC-d. Эти результаты подчеркивают

важность временного моделирования в обучении с подкреплением для реальных авиационных приложений.

Экспериментальная симуляция и анализ результатов

Условия симуляции

Для оценки эффективности стратегий избегания столкновений в сценариях с фиксированными задержками реакции пилота мы провели эксперименты, сравнив LSTM SAC-d и Марковский SAC-d. Оба алгоритма были протестированы в специально разработанной среде, предназначенной для моделирования дискретных вертикальных маневров для предотвращения столкновений самолетов. В этой среде учитывается задержка реакции пилота в 3 секунды, что делает систему частично наблюдаемой и добавляет временные зависимости, создавая дополнительные трудности для традиционных марковских подходов.

LSTM SAC-d расширяет модель SAC-d, интегрируя сеть LSTM для моделирования исторических зависимостей, что позволяет предсказывать задержанные реакции и адаптировать политику в соответствии с ними.

В отличие от этого, Марковский SAC-d использует традиционную архитектуру на базе MLP, предполагая полностью наблюдаемые состояния и немедленную реакцию.

Оба метода обучались на протяжении 30,000 итераций, с оценкой их эффективности в 20 независимых задачах. Размеры буфера воспроизведения и размеры батчей были настроены в зависимости от требований каждого алгоритма, при этом LSTM SAC-d использовал меньшие батчи для учета последовательного моделирования. Параметры алгоритма были установлены следующим образом (таблица 4).

Таблица 4

Параметры симуляции

Параметр	**LSTM SAC-d**	**Markovian SAC-d**
Скорость обучения	0.0003	0.0003
Коэффициент дисконтирования (γ)	0.900	0.975
Размер буфера повторов	4×10^4	4×10^4
Размер пакета	16	512
Число итераций обучения	30000	30,000
Архитектура политики	LSTM	MLP

Оценка производительности алгоритмов

Для оценки эффективности алгоритмов в решении задачи избегания столкновений использовались следующие метрики:

- **Средний возврат (Average Return)**: Среднее вознаграждение, полученное за последние 100 эпизодов, отражающее общую производительность и стабильность алгоритма.
- **Частота столкновений (Collision Rate)**: Средняя частота столкновений за последние 100 эпизодов, показывающая количество случаев "почти воздушного столкновения" (NMAC).
- **Процент успешного избегания столкновений (Success Avoidance Rate)**: Доля эпизодов за последние 100, в которых относительная высота между самолетами была поддержана в пределах безопасного диапазона от 900 до 1500 футов.
- **Среднее количество предупреждений (Average Alert Count)**: Среднее количество выданных предупреждений за эпизод за последние 100 эпизодов, оценивающее частоту выдачи предупреждений.
- **Среднее количество усилений предупреждений (Average Advisory Strengthening Count)**: Среднее количество предупреждений, которые были усилены за эпизод в последние 100 эпизодов, отражающее корректировки политики в условиях повышенного риска.
- **Среднее количество отмен предупреждений (Average Advisory Reversal Count)**: Среднее количество отмен предупреждений за эпизод за последние 100 эпизодов, указывающее на стабильность выданных предупреждений.
- **Среднее количество пересечений высот (Average Altitude Crossing Count)**: Среднее количество случаев пересечения высоты между двумя самолетами за эпизод в последние 100 эпизодов, показывающее потенциальные сценарии конфликта.

Экспериментальные результаты и анализ

В этом разделе представлен всесторонний анализ экспериментальных результатов, сравнивающих производительность LSTM SAC-d и Markovian SAC-d по различным метрикам. Результаты, показанные на нескольких графиках и

обобщенные в Таблице 5, демонстрируют превосходные возможности LSTM SAC-d в решении задач частично наблюдаемой среды избегания столкновений в воздухе с фиксированными задержками реакции пилота.

Таблица 5

Производительность двух алгоритмов в задачах избегания столкновений

Алгоритм	Частота столкновений	Процент успешного избегания столкновений	Количество оповещений	Количество укреплений оповещений	Количество изменений оповещений	Количество пересечений высоты
LSTM SAC-d	0	0.915	13.251	7.56	0.344	0.062
Markovian SAC-d	0.366	0.278	1.583	5.409	2.062	0.205

Кумулятивный возврат, достигнутый LSTM SAC-d (рисунок 3), показывает стабильный рост на протяжении всего процесса обучения, стабилизируясь на значительно более высоком уровне по сравнению с Markovian SAC-d. Эта тенденция указывает на то, что LSTM SAC-d последовательно учится и улучшает свою политику, используя временную информацию, захваченную с помощью структуры LSTM, для учета задержек реакции пилота и частичной наблюдаемости среды. В отличие от этого, Markovian SAC-d не достигает аналогичных результатов, достигая плато на значительно более низком уровне возврата из-за своей ограниченной способности обрабатывать историческую информацию, что препятствует созданию оптимальных рекомендаций в динамических сценариях.

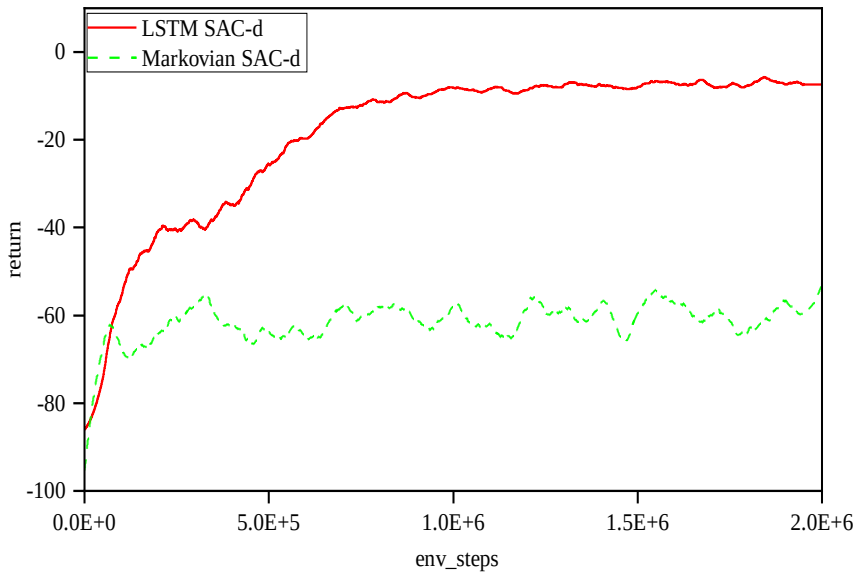


Рисунок 3 - Производительность тренировки двух алгоритмов в задачах избегания столкновений с задержкой реакции пилота

Рисунок 4(b) дополнительно подтверждает эффективность LSTM SAC-d в избегании столкновений, поскольку алгоритм быстро снижает частоту столкновений до значений, близких к нулю, в процессе обучения. Этот результат отражает способность алгоритма предсказывать потенциальные конфликты и предоставлять своевременные и эффективные решения. В свою очередь, Markovian SAC-d демонстрирует постоянно более высокую частоту столкновений, что указывает на его неспособность адекватно снижать риски в условиях частичной наблюдаемости. Левый подграфик Рисунка 4(a) показывает, что LSTM SAC-d позволяет двум самолетам достичь разумного интервала относительных высот в значительно большем числе эпизодов, что подчеркивает его устойчивые способности к принятию решений.

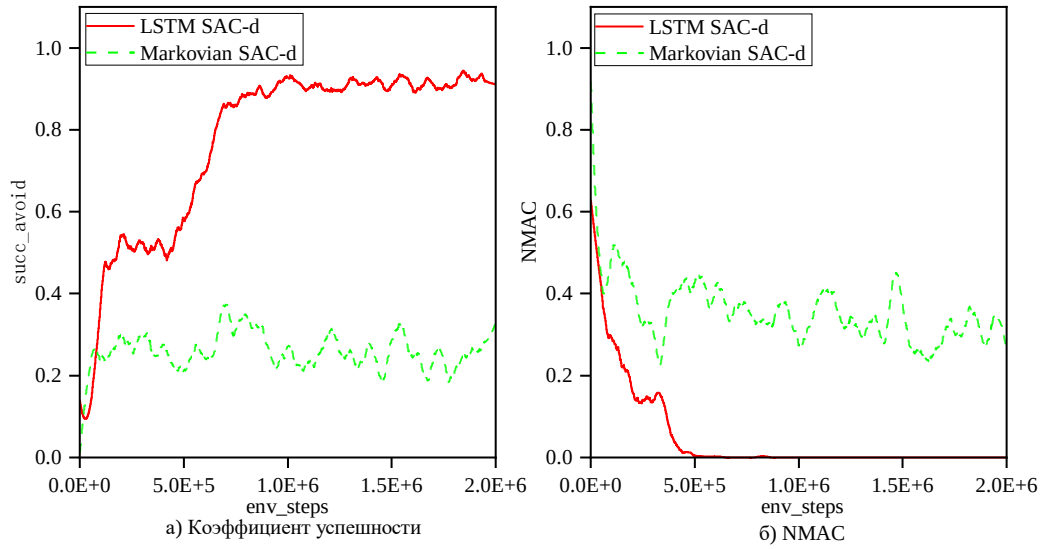


Рисунок 4 - Сравнение частоты столкновений и успеха избегания столкновений между двумя алгоритмами

Метрики оповещений и усиления оповещений на Рисунке 5 дают дополнительные сведения о стратегиях принятия решений двух алгоритмов. LSTM SAC-d демонстрирует более высокое среднее количество оповещений на эпизод, как показано на Рисунке 5(a), что указывает на его проактивный подход в выявлении потенциальных рисков и своевременном выдаче рекомендаций.

Однако Рисунок 5(b) показывает, что среднее количество усилений оповещений со временем снижается, сходясь к значению, которое лишь немного превышает результат Markovian SAC-d.

Это указывает на то, что, хотя LSTM SAC-d генерирует больше оповещений, его рекомендации становятся более точными и стабильными по мере прохождения обучения, требуя меньшего количества корректировок.

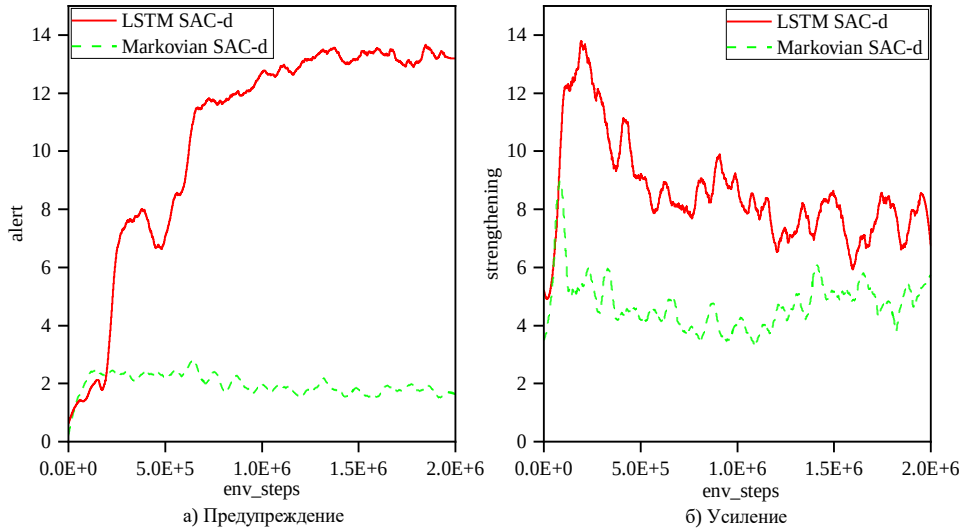


Рисунок 5 - Сравнение частоты выдачи предупреждений и усилений advisories между двумя алгоритмами

Метрики реверсий и пересечений на Рисунке 6 дают более глубокое понимание поведения алгоритмов при направлении воздушных судов к безопасным разрешениям. Рисунок 6(a) показывает, что LSTM SAC-d значительно снижает количество реверсий на эпизод по сравнению с Markovian SAC-d, что демонстрирует его способность предоставлять более решительные и эффективные рекомендации. В то время как Рисунок 6(b) показывает, что уровень пересечений для LSTM SAC-d стабильно ниже, чем у Markovian SAC-d, что свидетельствует о лучшем управлении разделением и меньшем количестве наложений по высоте.

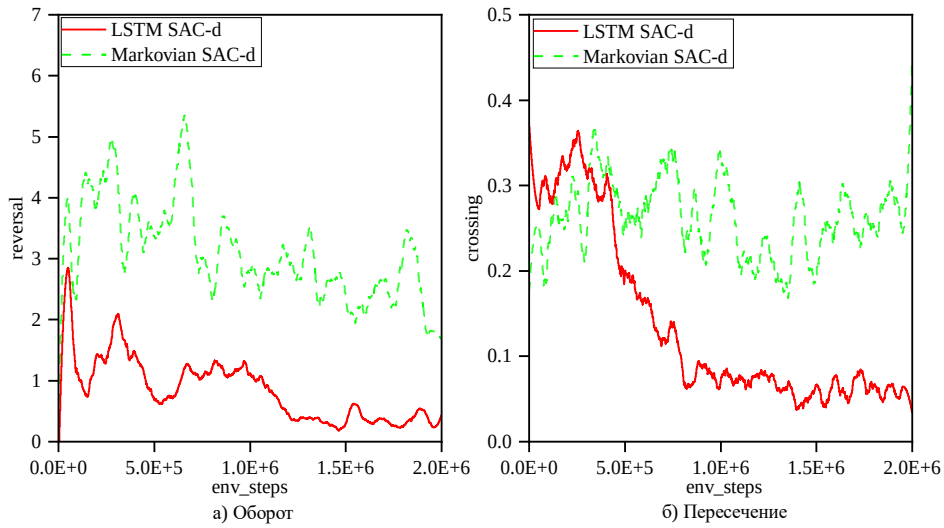


Рисунок 6 - Сравнение числа отмен и пересечений высоты между двумя алгоритмами

В заключение, результаты экспериментов по всем меткам последовательно демонстрируют превосходство LSTM SAC-d над Markovian SAC-d. Эффективно справляясь с частичной наблюдаемостью в среде предотвращения столкновений в воздухе и учитывая задержку реакции пилота, LSTM SAC-d достигает более высокого уровня безопасности, лучшей стабильности принятия решений и более эффективных стратегий разрешения конфликтов. Эти выводы подтверждают целесообразность применения LSTM SAC-d как надежного решения для реальных систем предотвращения столкновений в воздухе.

Заключение

Проведенное исследование демонстрирует эффективность применения модели предотвращения столкновений в воздухе как POMDP и решение этой задачи с использованием LSTM SAC-d. Включив 3-секундную задержку реакции пилота, модель показывает более высокую эффективность предотвращения столкновений по сравнению с Markovian SAC-d, достигая больших вознаграждений, меньшего числа пересечений по высоте, уменьшения числа реверсий, при этом поддерживая проактивное предупреждение и адаптивное принятие решений. Эти результаты подчеркивают устойчивость LSTM SAC-d в динамичных, частично наблюдаемых средах, предлагая многообещающий подход для будущих систем.

Конфликт интересов

Автор заявляет об отсутствии конфликта интересов.

Conflict of interest

The author declares no conflict of interest.

Список источников/ References

1. Holland J E, Kochenderfer M J, Olson W A. Optimizing the next generation collision avoidance system for safe, suitable, and acceptable operational performance[J]. Air Traffic Control Quarterly, 2013, 21(3): 275-297.

2. Londner E H, Moss R J. Bayesian network model of pilot response to collision avoidance system resolution advisories[J]. Journal of Air Transportation, 2018, 26(4): 171-182.

3. Panoutsakopoulos C, Yuksek B, Inalhan G, et al. Towards safe deep reinforcement learning for autonomous airborne collision avoidance systems[C]//AIAA SCITECH 2022 Forum. 2022: 2102.

4. Li S, Egorov M, Kochenderfer M. Optimizing collision avoidance in dense airspace using deep reinforcement learning[J]. arXiv preprint arXiv:1912.10146, 2019.

5. Rizk H, Chaibet A, Kribèche A. Model-based control and model-free control techniques for autonomous vehicles: A technical survey[J]. Applied Sciences, 2023, 13(11): 6700.

6. Lindqvist B, Mansouri S S, Agha-mohammadi A, et al. Nonlinear MPC for collision avoidance and control of UAVs with dynamic obstacles[J]. IEEE robotics and automation letters, 2020, 5(4): 6001-6008.

7. Kochenderfer M J, Chryssanthacopoulos J P. Robust airborne collision avoidance through dynamic programming[J]. Massachusetts Institute of Technology, Lincoln Laboratory, Project Report ATC-371, 2011, 130.

8. Brechtel S, Gindele T, Dillmann R. Probabilistic decision-making under uncertainty for autonomous driving using continuous POMDPs[C]//17th international IEEE conference on intelligent transportation systems (ITSC). IEEE, 2014: 392-399.

9. Kochenderfer M J, Chryssanthacopoulos J P. Robust airborne collision avoidance through dynamic programming[J]. Massachusetts Institute of Technology, Lincoln Laboratory, Project Report ATC-371, 2011, 130.

10. Ni T, Eysenbach B, Salakhutdinov R. Recurrent model-free rl can be a strong baseline for many pomdps[J]. arXiv preprint arXiv:2110.05038, 2021.

Информация об авторах

Лю Цзочэн, Северо-Западный политехнический университет, г. Сиань, Китай; e-mail: liuzuocheng@mail.nwpu.edu.cn

Information about the authors

Liu Zuocheng, Northwestern Polytechnical University, Xi'An, P. R. China**;** e-mail: liuzuocheng@mail.nwpu.edu.cn

Получено 29 сентября 2025 ● Принято к публикации 17 февраля 2026 ● Опубликовано 27 февраля 2026
Received 29 September 2025 ● Accepted 17 February 2026 ● Published 27 February 2026