doi: 10.7690/bgzdh.2025.04.019

基于无建图的强化学习人工势场法编队

丁 磊[1,2]，骆云志[1]，洪华杰[2]，黄 杰[2]，樊 鹏[1]，赵 伟[3]，陈斯灏[1]
（1. 中国兵器装备集团自动化研究所有限公司系统总体部，四川 绵阳 621000；
2. 国防科技大学智能科学学院，长沙 410073；3. 陆装驻广元地区军代室，四川 广元 628000）

摘要：针对同步定位与建图（simultaneous localization and mapping，SLAM）技术对计算资源的高需求、有限环境适应性、累积误差问题、系统复杂度高、成本昂贵、大场景处理能力受限以及缺乏有效的回环检测机制的缺点，提出一种结合人工势场法和深度强化学习的方法。利用图论模拟人工势场在机器人间的相互作用以及机器人与目的地之间的势场力，并采用孪生延迟深度确定性策略梯度（twin delayed deep deterministic policy gradient，TD3）算法来优化机器人对障碍物信息的感知和处理。仿真试验结果表明：该方法使机器人能够在未知环境中快速、准确地进行定位、移动，同时维持队形的稳定性和一致性。

关键词：人工势场法；强化学习；双延时确定策略梯度；图论
中图分类号：TP242.6 **文献标志码：**A

Artificial Potential Field Formation Method Based on Reinforcement Learning without Graph Construction

Ding Lei[1,2], Luo Yunzhi[1], Hong Huajie[2], Huang Jie[2], Fan Peng[1], Zhao Wei[3], Chen Sihao[1]
(1. *Department of System General*, *Automation Research Institute Co*., *Ltd. of China South Industries Group Corporation*, *Mianyang* 621000, *China*; 2. *College of Intelligence Science and Technology*, *National University of Defense Technology*, *Changsha* 410073, *China*;
3. *Military Representative Office of Army Equipment Department in Guangyuan District*, *Guangyuan* 628000, *China*)

Abstract: For simultaneous localization and mapping (SLAM) technology has the disadvantages of high demand for computing resources, limited environmental adaptability, cumulative error problem, high system complexity, high cost, limited large scene processing capacity and lack of effective loop detection mechanism, so a method combining artificial potential field method and deep reinforcement learning is proposed. The graph theory is used to simulate the interaction between robots and the potential force between robots and the destination, and the twin delayed deep deterministic policy gradient algorithm is used to optimize the robot's perception and processing of obstacle information. The simulation results show that the method can make the robot locate and move quickly and accurately in the unknown environment, while maintaining the stability and consistency of the formation.

Keywords: artificial potential field method; reinforcement learning; two-delay deterministic policy gradient; graph theory

0 引言

随着科技的不断发展，大量机器人被使用在日常工作中。单个机器人由于受载荷能力、电池、算力、观察实验范围的限制，导致在执行一些复杂任务的时候力不从心。机器人编队算法将多个机器人编队在一起执行一个复杂任务，相比单个机器人在观测、计算、任务处理上有很大的优势。

为了实现机器人编队，研究人员提出了许多编队算法，包括基于行为法[1]、虚拟结构法[2]，领航者跟随者法[3]以及人工势场法[4]。

传统人工势场法以及之后的改进方法容易在通道狭窄时不能通过，极度依赖同步定位与建图（SLAM）[5]等可以获知环境信息的技术，这导致它对计算资源的高需求、有限的环境适应性、累积误差、系统复杂度高、成本昂贵、大场景处理能力受限以及缺乏有效的回环检测机制的缺点。这些方法在处理动态障碍物和陌生环境时不能做出及时反应。在多机器人编队场景下，传统方法也存在编队时间长、稳定性差，容易陷入局部死区和在障碍物前振动徘徊等问题。

针对这些挑战，笔者提出了一种改进的编队策略。具体而言，利用图论方式对各个机器人之间的势场力建模，同时对单个机器人的障碍物识别进行马尔可夫决策过程建模，使其能够实时处理环境信息，有效解决了机器人编队时间过长、编队稳定性差、易陷入死区以及对动态障碍物避障效果不佳的

收稿日期：2024-08-10；修回日期：2024-09-10
第一作者：丁 磊（1993—），男，内蒙古人，硕士。

问题，从而显著提升了多机器人编队的效率和稳定性。

1　方法设计

1.1　基于图论的编队势场构建

1) 建立编队势场。

系统中的每个机器人可以通过里程计感知其他群体成员相对于参考坐标系的位置。在无人车编队系统中，内部控制回路用于跟踪参考路径。笔者假设每个机器人是 n 维空间中的一个系统，具有以下动力学特征：

$$P_i=U_i \qquad i=1, 2, 3, 4\cdots。 \tag{1}$$

式中：P_i 为机器人现在的位置；U_i 为机器人现在所受的势场力。

假设系统一共有 N 个机器人，对于每个机器人所收到的编队势场力为 u_i^f，δ_{ij} 表示第 i 个机器人和第 j 个机器人之间的期望位置差。此外 δ_{ij} 必须满足 $\delta_{ji}=-\delta_{ij}$，$\delta_{ii}=0$。那么编队问题有如下定义：

假定在 $t>0$ 时，每个机器人的位置 p 有：

$$\lim_{t\to\infty} p_j(t)-p_i(t)=\delta_{ji}\ ;\ 1\leqslant i, j\leqslant N, i\neq j\ 。 \tag{2}$$

那么每个编队的势场力可以通过机器人的相对位置控制器来计算：

$$u_i^f=k_p\sum\nolimits_{j=1}^{N}(p_j-p_i-\delta_{ji})\ 。 \tag{3}$$

式中 $k_p>0$ 为编队控制增益。式(3)的策略也可以被认为是一个共识问题[6]。

上述公式也可被写作：

$$k_p\sum\nolimits_{j=1}^{N}(p_j-p_i-\delta_{ji})=k_p\sum\nolimits_{j=1}^{N}(p_j-p_i)+b_i\ 。 \tag{4}$$

式中 $b_i=-k_p\sum\nolimits_{j=1}^{N}\delta_{ji}$ 为输入的偏置。如文献[7]所指出的，沿各轴的动力学是解耦的，这表明 1 维分析对 n 维情况是有效的。假设 n=1 结果适用于一般有限 $n\geqslant 1$。为了利用拉普拉斯矩阵的特性，以一种更精简的矩阵形式重新表达了具有控制信号的群体系统动态。

$$\dot{p}=-k_p\boldsymbol{L}p+b\ 。 \tag{5}$$

式中：$p=[p_1, p_2,\cdots, p_N]^{\mathrm{T}}$，$b=[p_1, p_2,\cdots, p_N]^{\mathrm{T}}$ 分别小车位置信息和偏置参数；$\boldsymbol{L}\in R^{N\times N}$ 为群体相互作用拓扑图的拉普拉斯矩阵。这个拉普拉斯矩阵定义为 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{A}$，其中 $\boldsymbol{D}\in R^{N\times N}$ 是一个对角矩阵，$\boldsymbol{D}$ 的第 i 个对角线元素是第 i 个元素的度，其余元素为 0，$\boldsymbol{A}$ 是表示各个机器人间距离的邻接矩阵。又由于控制矩阵 $\boldsymbol{u}_i^f$ 是一个完全连通图，故 $\boldsymbol{L}$ 的元素定义为：

$$l_{ij}=\begin{cases}-1 & i\neq j\\ N-1 & i=j\end{cases}\ 。 \tag{6}$$

设向量 $\boldsymbol{b}$ 是 $\boldsymbol{L}$ 的特征向量，其对应特征值 N。对于矩阵 $\boldsymbol{L}$，其特征值具有特定的性质：矩阵 $\boldsymbol{L}$ 仅拥有一个零特征值，其余所有特征值均为相同的正数，这表明该矩阵是正半定的，因此有：

$$\boldsymbol{Lb}=N\boldsymbol{b}\ 。 \tag{7}$$

接下来定义该编队群系统的类李雅普诺夫函数为：

$$V(p)=\frac{1}{2}(p-\frac{1}{k_pN}b)^{\mathrm{T}}(p-\frac{1}{k_pN}b)\ 。 \tag{8}$$

对 V 沿 p 求导，得到：

$$\begin{aligned}\dot{V}(p)&=(p-\boldsymbol{b}/k_pN)^{\mathrm{T}}(-k_p\boldsymbol{L}p+b)=\\ &-k_p(p-\boldsymbol{b}/k_pN)^{\mathrm{T}}\boldsymbol{L}(p-\boldsymbol{b}/k_pN)\leqslant 0\ 。\end{aligned} \tag{9}$$

定义集合 $\Omega_e=\{p|V(p)=0\}$，当 $t\to\infty$，各机器人位置集合 p 收敛到 Ω_e 的最大不变性子集。

设 $p=\boldsymbol{b}/k_pN+\alpha 1_v$，$p\in\Omega_e$。其中，$1_v$ 是一个由 1 组成的 N 维向量，α 是一个实值常数。这里可以知道，Ω_e 的所有元素都是对应于 $p=b/k_pN$ 的常数平移，因此可知，Ω_e 具有不变性，即当 Ω_e 趋向稳定时小车呈编队队形。

小车在 t 时刻的每个小车的势场力：

$$\boldsymbol{L}_p=\boldsymbol{Lb}/k_pN+\alpha\boldsymbol{L}1_v=\boldsymbol{Lb}/k_pN=\boldsymbol{b}/k_p。 \tag{10}$$

笔者采用这个式子近似获得机器人的编队势场函数。

2) 建立引力势场。

$$U_{\mathrm{att_{goal}}}(x)=K(X_{\mathrm{goal}}-X)^2/2\ 。 \tag{11}$$

式中：$U_{\mathrm{att_{goal}}}(x)$ 代表小车此刻在 x 位置上与目标点产生的引力；K 为引力系数；$X_{\mathrm{goal}}-X$ 为小车与目标点的距离。机器人小车与目标点间的距离和引力势场值的平方成正比关系，小车与目标点越近，势场力越小，反之越大。

3) 建立斥力势场。

在人工势场法中，斥力场主要包括机器人间的斥力和机器人与障碍物的斥力。由于机器人的位置信息容易通过里程计获取，在处理机器人间的碰撞时候，数据计算难度小，速度快。机器人间的人工势场斥力场函数为：

$$U_{\mathrm{rep}}(x)=\eta(1/\rho_{\mathrm{object}}-1/\rho_{\mathrm{o}})^2/2。 \tag{12}$$

式中：η 为斥力势场函数系数；ρ_{o} 为斥力发生的最

小距离；ρ_{object} 为机器人之间的距离。当 $\rho_{object}-\rho_0$ 大于 0，不计算这 2 个机器人之间的斥力，定义斥力为 0；当 $\rho_{object}-\rho_0$ 小于 0，通过单数函数拟合距离和斥力的变换关系。其中：

$$\rho_{object}=\sqrt{(X_{object}-X)^2+(Y_{object}-Y)^2}\text{。} \tag{13}$$

在该方法中，斥力场的大小与 2 个机器人之间的距离成反比关系：即 2 个机器人之间的距离越近，斥力场的强度就越大；2 个机器人之间的距离越远，斥力场的强度就越小。当 2 个机器人之间的距离超过某个特定的阈值时，斥力场的影响将不再存在。

1.2　基于 TD3 算法的障碍物势场构建

因小车障碍物的斥力不易直接从环境获取，故采用强化学习方法，通过小车与环境交互训练策略网络获取这一数据。小车的障碍物斥力如下：

$$U_{TD3}(x)=\text{network}_{车,环境}\text{。} \tag{14}$$

通过输入小车的激光雷达数据，前进方向和小车与目的地夹角，计算小车在此刻的斥力。

本文中的强化学习采用 TD3 算法[8]，该算法采用确定性策略，如图 1 所示，Actor 网络接收状态信息作为输入，并产生相应的动作作为输出。具体而言，Actor 网络包含：主网络和目标网络 2 个网络结构。这 2 个网络结构在训练过程中并行更新，但目标网络的参数更新速度较慢，其参数通过软更新与 Actor 网络同步，目的是为了稳定训练过程。

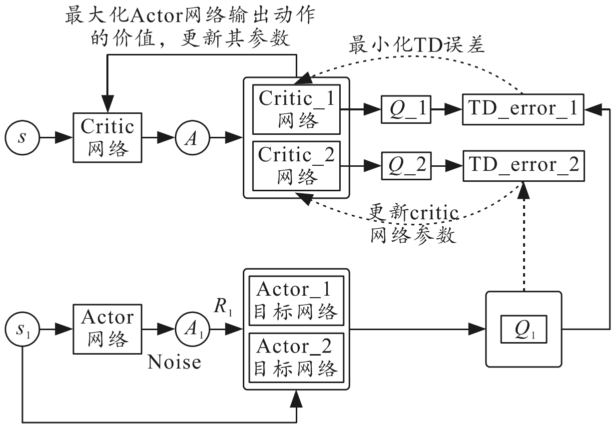

图 1　TD3 算法系统结构

Critic 网络部分较复杂，包含 2 对网络：Critic_1 网络及其对应的目标网络 Critic-target_1 和 Critic_2 网络及其对应的目标网络 Critic-target_2。这 2 对网络共同评估在特定环境下执行动作的价值，即 Q 值。Critic_1 和 Critic_2 网络分别独立地学习状态-动作对的价值函数，而它们的目标网络则用于计算目标 Q 值，以减少过估计问题。这种双 Critic 结构有助于提高算法的性能和稳定性，TD3 每部分的设计细节具体如下：

1) 状态空间。构建多智能体系统的初始状态空间包括：获取无人车感知系统接收到的外部环境的观测数据，所述观测数据包括前向激光雷达数据、无人车里程计信息以及目标点位置。

根据所述观测数据计算当前无人车与目的地的距离、当前无人车方向角和无人车与目的地的连线的夹角，当前小车的线速度、当前小车的角速度；将无人车正前方提取的 180° 数据每 9° 划分为一个区间；选择该区间上最大值作为一个 20 维度的数据，得到初始状态空间[distance, theta, action[0], action[1], laserdat[0], …, laserdat[19]]。其中，Distance 为无人车与目标点的距离，Theta 为无人车朝向与目标点连线的夹角；action[0]、action[1]分别为无人车的线速度和角速度；laserdat[0], …, laserdat[19]为雷达数据。

2) 动作空间。由于本文中使用的是 TD3 算法，产生的动作是一个连续动作。强化学习的动作空间是一个含有机器人在 X 轴方向上的速度和机器人此刻的方向角的动作空间，这 2 个变量决定了机器人的前进方向和前进速度。由于机器人需要保证在移动时保持编队队形、防止机器人之间相互碰撞、躲避障碍物且机器人移动速度不能太慢，所以要求机器人保证在满足上述条件下尽量快速移动。本文中设定了机器人最快移动速度，防止机器人由于移动过快导致通信不能及时获取强化学习所需信息。

Critic 网络计算某个动作的价值函数：

$$\nabla_{\phi}J(\phi)=N^{-1}\sum\nabla_{a}Q_{\theta_1}(s,a)|_{a=\pi_{\phi}(s)}\nabla_{\phi}\pi_{\phi}(s)\text{。} \tag{15}$$

式中：$\nabla_{\phi}J(\phi)$ 为机器人在某时刻下采取某种动作下的价值函数；$Q_{\theta_1}(s,a)|_{a=\pi_{\phi}(s)}$ 为执行某次动作后获取的价值；$\pi_{\phi}(s)$ 为机器人在状态 s 下采取的某个动作。这个函数表示在训练经验池中随机选取一定批次的数据，计算这些数据在执行这些动作的价值函数。机器人的动作和价值曲线如图 2 所示。

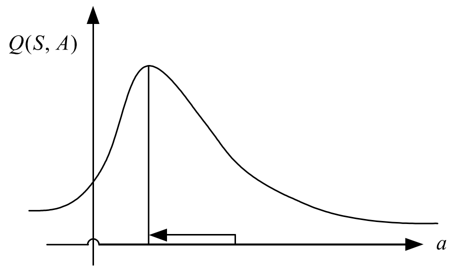

图 2　TD3 网络动作价值训练

根据 TD3 网络动作价值训练曲线可知：动作网络可以输出的数据应保证尽可能覆盖所有动作，同时要求在最佳动作周围多取值、多训练，尽量拟合最佳输出动作，所以笔者为 v_x 和 ω 设计一个扰动，这个扰动在一开始保证动作选取变化范围足够广，训练一段时间后，扰动变小，动作变化范围逐渐缩小在动作策略函数输出的附近。

3) 奖励函数。

笔者为机器人在强化学习障碍物势场设计了奖励函数：

$$r(s_t, a_t) = \begin{cases} r_g & D_t < \eta \\ r_c & \text{collision} \\ v - |\omega| & \text{others} \end{cases} 。\tag{16}$$

式中：r_g 为+100；r_c 为−100。r_g 为机器人在到达目标点后给一个+100 的奖励；r_c 为在机器人碰到障碍物时，给出一个−100 的奖励；$v-|\omega|$为机器人在移动时，当机器人朝着目标点移动的奖励，速度越大奖励越大，该设计可以保证机器人尽可能跑快。

2　仿真实验

2.1　实验设置

采用深度强化学习(deep reinforcement learning，DRL)来实现机器人的本地导航任务。实验是在一台配备有 4 个 NVIDIA GeForce GTX 4090 显卡、64 GB RAM 以及 Intel Core i7-6800K CPU 的计算机上进行的。笔者使用 TD3 算法在 Gazebo 仿真环境中训练机器人。每个回合在机器人发现达到目标、检测到碰撞或执行了 500 个动作步骤后结束。在实验中测试了最大线速度为 0.5、1.0、1.5 和 2.0 m/s 的运动轨迹。为了稳定学习过程，采用延迟奖励机制，即在最近 10 个连续步骤内更新奖励，同时设置了 2 个回合的参数更新延迟。训练环境是一个 10 m×10 m 的模拟空间。为了促进策略的泛化和探索，笔者在传感器读数和动作值中引入了高斯噪声。此外，为了增加环境的多样性，在每个回合开始时随机改变盒子形障碍物的位置。

2.2　实验结果分析

为了验证笔者所提方法的有效性，选用胡铮等[9]提出的引入角度定义障碍物人工势场法作为对比，其法引入了碰撞距离筛选障碍物、引入角度定义障碍物的影响范围和引入距离因子解决目标不可达。2 种方法在不同的环境设置中进行了多次实验，实验中记录的数据包括运动速度、重编队时间，以及该方法成功重编队的比例。

第 1 个实验是测试 2 种算法在陌生环境下的运动轨迹。图 3 和 4 分别表示 2 种方法的小车运动路径，实验说明基于双延时确定策略人工势场法可以不依赖事先建图，利用激光雷达实时躲避陌生环境下的障碍物。

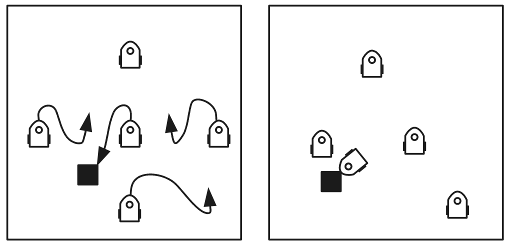

图 3　引入角度定义障碍物人工势场法机器人路径

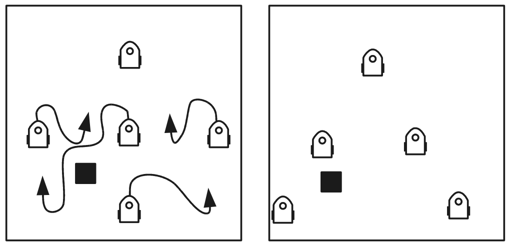

图 4　基于双延时确定策略人工势场法机器人路径

第 2 个实验是测试 2 种算法在尽快提高小车速度后，机器人编队的成功率。图 5 展示了实验次数为 50 的情况下，不同速度的小车用 2 种编队方式的成功率。

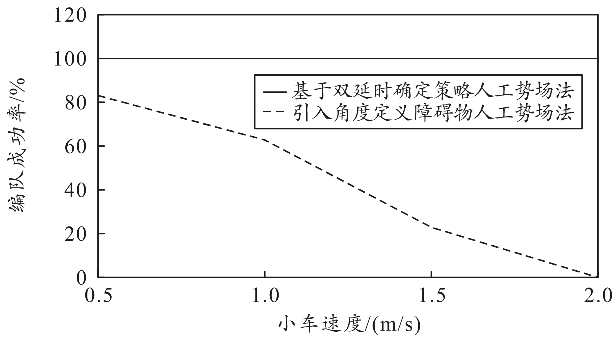

图 5　不同速度编队成功率对比

从图 5 中可以看到：改进人工势场在速度为 2 m/s 时依然可以正常编队，而传统人工势场法，在速度过大的情况下，小车出现轨迹过大，振荡严重，无法正常完成编队。

此外，相比于引入角度定义障碍物人工势场法本算法可以躲避移动的障碍物、不需要提前准备点云图，在短暂中断其中一个机器人的通信后仍可完成编队。

仿真结果表明：笔者提出的基于强化学习编队

算法能够在较小的时间内完成编队，并具有可行性和较优的编队能力，相比与传统人工势场法有着更高的编队效率和成功率，对于一些动态障碍物也可完成避障。

3 结论

笔者提出一种基于强化学习的算法，旨在替代传统人工势场法中的障碍物斥力场。该算法通过图论构建编队引力场，实现了对小车位置的动态调整。具体而言，每个小车利用其搭载的激光雷达感知前方视野内的障碍物，并借助里程计将自身位置信息传递给其他小车。通过图论方法，能够优化每个小车的位置，以维持编队结构。在 Gazebo 仿真环境中进行的测试结果表明：该策略显著提高了编队的成功率，并实现了更高效的重编队过程。此外，该算法还使得小车能够以更高的速度进行编队运动，从而在效率和速度上相较于传统人工势场法展现出显著优势。这些发现为未来在复杂环境中实现高效、可靠的多智能体编队提供了新的视角和方法。

参考文献：

[1] NACER H, MENDIL B. Behavior-based Autonomous Navigation and Formation Control of Mobile Robots in Unknown Cluttered Dynamic Environments with Dynamic Target Tracking[J]. International Journal of Automation & Computing, 2021, 18(5): 766-786.

[2] YAN X, JIANG D P, MIAO R L, et al. Formation control and obstacle avoidance algorithm of a multi-usv system based on virtual structure and artificial potential field[J]. Journal of Marine Science and Engineering, 2021, 9(2): 1-17.

[3] GÓMEZ N, PENA N, RINCON S, et al. Leader-follower Behavior in Multi-agent Systems for Search and Rescue Based on PSO Approach[J]. SoutheastCon, 2022: 413-420.

[4] HAO G Q, LYU Q, HUANG Z, et al. UAV Path Planning Based on Improved Artificial Potential Field Method[J]. Aerospace, 2023, 10(6): 562.

[5] CARLOS C, ELVIRA R, GOMEZ R J J, et al. An accurate open-source library for visual, visual–inertial, and multimap SLAM[J]. IEEE, 2020, 37: 1874-1890.

[6] OLFATI-SABER R, FAX I A, MURRAY R M. Consensus and cooperation in networked multi-agent systems[J]. Proceedings of the IEEE, 2007, 95(1): 215-233.

[7] FAX J A, MURRAY R M. Information flow and cooperative control of vehicle formations[J]. IEEE transactions on automatic control, 2004, 49(9): 1465-1476.

[8] FUJIMOTO S, HOOF H, MEGER D. Addressing function approximation error in actor-critic methods[J]. International conference on machine learning, 2018: 1587-1596.

[9] 胡铮，徐斌．改进人工势场法的轨迹规划[J]．电光与控制, 2023, 30(3): 38-41, 53.

****

（上接第 63 页）

[8] 吴丹杨．小型倾转多旋翼无人机飞控系统研究与设计实现[D]．成都：电子科技大学, 2020.

[9] 杨姗姗，王彪．基于 FlightGear 的三维可视化飞行控制仿真实验平台的设计[J]．实验室研究与探索，2017, 36(7): 113-117.

[10] 董斐，刘剑超，林亚军，等．飞行仿真系统虚拟仪表关键技术[J]．兵工自动化, 2021, 40(11): 23-26.

[11] 漆卫微．EAD200 飞机的总体方案设计与性能分析[D]．南京：南京航空航天大学, 2015.

[12] 安子聪，院老虎，陈旭．常规布局无人机动力学解算和三维视景仿真[J]．沈阳航空航天大学学报, 2020, 37(4): 31-39.

[13] 关永亮．复合材料无人机结构和飞行动力学关键技术研究[D]．长春：中国科学院长春光学精密机械与物理研究所, 2017.

[14] 陈琦．基于 FlightGear 的低空风切变飞行模拟研究[D]．天津：中国民航大学, 2014.

[15] 王振川．超临界压力流体湍流换热实验与数值模拟研究[D]．北京：清华大学, 2018.

[16] 王岳，汪磊，郭世广．基于 FlightGear/Matlab 的运输类飞机飞行仿真实验设计[J]．实验技术与管理，2019, 36(7): 129-133.