基于逆向强化学习的舰载机甲板调度优化方案生成方法*

李耀宇,朱一凡,杨 峰,贾 全
(国防科技大学 信息系统与管理学院,湖南 长沙 410073)

摘 要:针对计算机辅助指挥调度舰载机甲板作业的决策过程无法脱离人参与这一特点,引入基于逆向学习的强化学习方法,将指挥员或专家的演示作为学习对象,通过分析舰载机的甲板活动,建立舰载机甲板调度的马尔可夫决策模型(MDP)框架;经线性近似,采用逆向学习方法计算得到回报函数,从而能够通过强化学习方法得到智能优化策略,生成舰载机甲板调度方案。经仿真实验验证,本文所提方法能够较好地学习专家演示,结果符合调度方案优化需求,为形成辅助决策提供了基础。

关键词:逆向强化学习;强化学习;舰载机甲板调度;优化方案生成
中图分类号:TP391.9 **文献标志码**:A **文章编号**:1001-2486(2013)04-0171-05

Inverse reinforcement learning based optimal schedule generation approach for carrier aircraft on flight deck

LI Yaoyu, ZHU Yifan, YANG Feng, JIA Quan
(College of Information and System and Management, National University of Defense Technology, Changsha 410073, China)

Abstract: Traditional aircraft scheduling on carrier flight deck relies heavily on human commander decisions. To improve the computer aided decision making, an inverse reinforcement learning method was proposed. Learning from the commander or expert's demonstration, a Markov decision process (MDP) based aircraft scheduling model by analyzing the aircraft operations on deck was proposed. Then, the optimal policy and schedule were generated by using the linear approximating and inverse reinforcement learning method. Simulation results show that our method can learn expert's demonstration well, satisfy the requirement of scheduling optimization, and facilitate the computer aided decision making.

Key words: inverse reinforcement learning; reinforcement learning; aircraft scheduling on flight deck; optimal schedule generation

舰载机甲板调度作为影响航母战斗力的重要因素,一直是各军事强国争先研究的热点问题。航母甲板环境复杂,围绕舰载机的各项操作必须制定严格的行动计划,才能确保其在紧急情况下的出动能力和人员安全[1]。多年来,舰载机的调度方案一直由指挥人员的经验制定,但由于设备故障、操作失误或战情变化等多种不确定性因素的存在,使得人工制定调度方案变得十分繁琐而且影响效率。智能规划系统的出现为指挥人员提供了舰载机甲板调度辅助规划决策的手段。

国内学者在舰载机调度方面的研究仍处于理论探索阶段,其中,司维超[2]应用PSO算法研究戴高乐航母的舰载机甲板布列出动问题并建立了布列调度模型,但假设和模型比较简单,只考虑了近似距离和静态出动问题;魏昌全[3]针对舰载机两种出动方式,应用工业生产的车间调度方法研究了航空保障调度问题并给出了两种数学调度模型;马登武等[4]用遗传算法研究了舰载机的弹药调度问题,为舰载机的保障问题提供了优化方法。此外,冯强[5]针对舰载机调度中的不确定性问题,采用多主体技术对舰载机动态调度问题进行了研究。

最早用计算机辅助舰载机调度的是Giardina[6]和Johnson等[7],1974年,他们设计了最早的航母甲板操作控制系统(CADOCS),但没有得到充分的应用。2002年,Timothy[8]论述了智能数字化甲板调度系统代替人工规划操作的必然趋势,提出了基于智能Agent调度系统的需求分析。Jeffrey等[9]为减少舰载机作业过程中的失误和伤亡,设计了一种甲板持续监控系统,能够提供危险预警和舰载机甲板路径规划能力。2009年,MIT的Ryan等[10]始为美海军自动项目研究所开发一款名为航空母舰甲板行动过程规划者(Aircraft Carrier Deck Course of Action Planner, DCAP)的系统,旨在为航母甲板指控人员提供舰载机的作业流程和调度辅助方案。为应对人工制

* **收稿日期**:2012-10-25
基金项目:国家自然科学基金资助项目(71031007)
作者简介:李耀宇(1984—),男,黑龙江牡丹江人,博士研究生,E-mail:Garett_1984@hotmail.com;
朱一凡(通信作者),男,教授,博士,博士生导师,E-mail:nudtzyf@hotmail.com

定调度方案效率低且动态显示能力不足等问题，美国海军已开发了航空数据管理与控制系统（Aviation Data Management And Control System，ADMACS）[11] 和舰船综合信息系统（Integrated Shipboard Information System，ISIS）[12]，并首次尝试在舰船上建立整体数据库和分布式数据共享的系统，用来管理舰载机的飞行与甲板布列调运。

综合上述研究可以看出，智能辅助规划已成为舰载机甲板调度研究的发展趋势，但具体的作战仍然离不开人的参与，为使计算机生成的调度方案具备真正的辅助决策能力，必须将指挥人员的经验加入方案制定过程中[13]。近几年兴起的基于逆向学习的强化学习方法为我们研究这种需要人机交互的规划问题提供了解决途径[14]，它针对传统强化学习里 MDP 决策过程中的回报函数不易确定这一问题[15]，通过观察专家的演示，将专家制定的策略作为最优对象，计算优化回报函数，最后通过学徒学习方法生成优化调度策略，而调度策略对应的状态变化即为舰载机甲板调度方案。

根据舰载机甲板调度辅助决策问题与逆向强化学习方法都需要“专家”经验作为优化输入这一共同点，本文通过分析舰载机的甲板活动过程，以“尼米兹”级航母为研究对象，建立基于 MDP（Markov Decision Process）的舰载机状态转移模型，采用逆向强化学习方法从“专家”，即模拟指挥人员的调度演示中确定优化回报函数，进而由强化学习方法生成优化调度方案，为形成人在回路的辅助决策做准备。最后通过仿真实验验证方法的有效性和适用性。

1 基于 MDP 的舰载机甲板调度决策描述框架

1.1 舰载机甲板活动分析

舰载机主要的甲板调运活动如图 1 所示，作战任务下达后，需要出动的舰载机以拖运或滑行的方式至维修站进行加油和装载弹药的操作，然后拖动至指定的弹射器，等待弹射器空闲后进行飞行前准备，飞机弹射起飞并执行任务返航后，进入降落队列等待塔台指令，指挥人员根据待降舰载机的优先级安排着舰顺序，一旦成功降落，立即引导其滑行或拖运至停机位，或者到维修站进行补给准备再次出动，如果拦阻降落失败，立即重新起飞再次进入降落队列等待着舰。

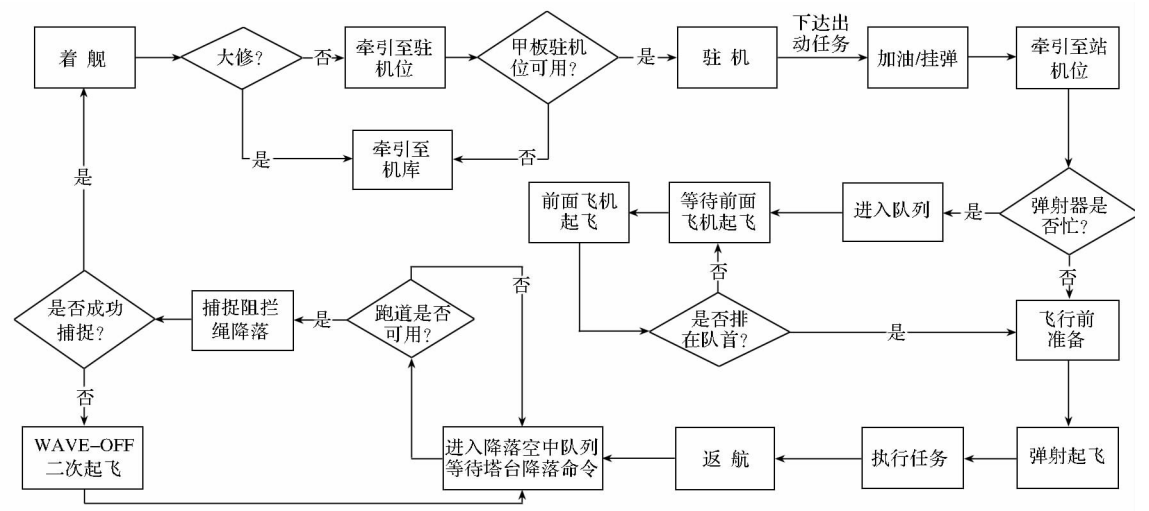

图 1 舰载机甲板操作调度过程

Fig. 1 Aircraft operation process on carrier flight deck

1.2 基于 MDP 的舰载机状态转移模型

MDP 决策过程可以用数组表示：(S, A, T, γ, R)，其中，S 代表状态变量集合；A 为行动集合；$T = P_{sa}$ 为状态转移概率（P_{sa} 代表状态 s 上采取行动 a 的状态转移分布）；$\gamma \in [0,1)$ 为一个折扣因子；$R: S \to \mathbf{R}$，或 $R(s)$ 为强化学习的回报函数。

策略可用 $\pi: S \to A$ 的映射关系表示，其任意状态点 s_1 的值函数可表示为：

$$V^{\pi}(s_1) = \mathrm{E}[R(s_1) + \gamma R(s_2) + \gamma^2 R(s_2) + \cdots | \pi]$$

其中，期望值由状态序列 $(s_1, s_2, \cdots)$ 的分布决定。对 $s_1 \to a_1$ 每一步的值函数我们可以用 Q 函数表示，即：

$$Q^{\pi}(s,a) = R(s) + \gamma E_{s_t \sim P_{sa}(\cdot)}[V^{\pi}(s_t)]$$

其中 $s_t \sim P_{sa}(\cdot)$ 表示 s_t 根据概率 $P_{sa}(\cdot)$ 转移的期望值。这样便可得到最优值函数：$V^*(s) = \sup_{\pi} V^{\pi}(s)$，以及最优 Q 函数：$Q^*(s,a) = \sup_{\pi} Q^{\pi}(s,a)$，而最优策略 π 则可表示为：$\pi(s) \in \arg\max_{a \in A}$

$Q^{\pi}(s,a)$。

用 MDP 决策过程描述舰载机甲板调度问题有如下特点:

• 调度过程中的许多不确定因素方便由 MDP 状态转移模型描述。

• 根据策略 $\pi: S \rightarrow A$ 的映射关系,可得到策略与调度方案的一一对应关系。

• 逆向强化学习的目的是得到 MDP 回报函数,以便强化学习生成优化策略。

舰载机甲板调度的 MDP 模型中状态变量主要包括设备完好性、位置、优先级和燃油余量。完好性是指舰载机、弹射器或降落跑道是否可用,即 YES 和 NO 两个布尔状态;位置变量包括停机位、四个弹射器、空中执行任务、降落队列和降落跑道;优先级用于规定舰载机的起飞和降落顺序;燃油余量可划分为 3 个级别,如果舰载机尚未降落而燃油余量变为 0,则意味着其已坠毁。由此可以得出 MDP 模型的状态空间数量为:

$$S_{num} = (2Y/N)^5 \times (8loc \cdot 3level \cdot 2pri)^n$$

其中 Y/N 指设备是否正常,n 为飞机的数量,loc 指位置,$level$ 为燃油级,pri 指优先级。MDP 的动作集合由舰载机的甲板作业过程决定,如飞机在降落队列中,其下一动作即为等待命令降落,而下一状态则根据该动作的转移概率决定,即成功降落,则状态转移为驻机,降落失败则重新起飞进入降落队列。作为 MDP 决策过程的关键,回报函数 R 直接对应着调度策略,所以,一旦确定了回报函数,我们便可以用强化学习方式生成优化调度策略。

2　基于逆向学习的调度方案生成

2.1　回报函数的线性近似描述方法

由 S_{num} 可以看出,如果飞机数量较多,舰载机甲板调度 MDP 的状态集合会非常庞大,用列表方式描述每一个 $s \rightarrow a$ 的回报函数不现实,为此,我们采用特征属性的线性近似方式描述回报函数集合:

$$R(s) = \omega^{\mathrm{T}} \phi(s) \tag{1}$$

其中 $\phi(s)$ 为特征属性向量,舰载机调度 MDP 的特征属性选取应能充分反映飞行甲板调度情况的状态变化,为人机交互生成优化调度策略做准备。根据对舰载机甲板作业过程的分析,本文选取如下 7 种特征属性作为研究对象:

• 每个位置变量上的飞机数量(1 个)

• 各甲板位置上,三种燃油余量对应的飞机数量(3 个)

• 空中执行任务中,三种燃油余量对应的飞机数量(3 个)

回报函数中,ω 为权重向量,它既可以根据特征属性由人工制定,也可根据逆向学习方法从专家演示中学习得到。

2.2　基于逆向学习的优化回报函数生成

根据 MDP 过程的原理和近似回报函数的描述,一个策略 π 的值可以描述为:

$$\begin{aligned} \mathrm{E}_{s_0 \sim P_{sa}(\cdot)}[V\pi(s_0)] &= \mathrm{E}\Big[\sum_{t=0}^{\infty} \gamma^t R(s_t) \mid \pi\Big] \\ &= \mathrm{E}\Big[\sum_{t=0}^{\infty} \gamma^t \omega^{\mathrm{T}} \varphi(s_t) \mid \pi\Big] = \omega^{\mathrm{T}} \mathrm{E}\Big[\sum_{t=0}^{\infty} \gamma^t \varphi(s_t) \mid \pi\Big] \end{aligned} \tag{2}$$

与式(1) 对比,我们可以定义特征属性向量的值 $\mu(\pi)$ 为:

$$\mu(\pi) = E\Big[\sum_{t=0}^{\infty} \gamma^t \phi(s_t) \mid \pi\Big] \in \mathbf{R} \tag{3}$$

为从专家演示中得到回报函数,先假设“专家”的特征属性期望值 $\mu_E = \mu(\pi_E)$,并给定专家演示的状态序列轨迹:$\{s_0^{(i)}, s_1^{(i)}, \cdots\}_{i=1}^{m}$,则可得到 μ_E 的经验估计:

$$\hat{\mu}_E = \frac{1}{m} \sum_{i=1}^{m} \sum_{t=0}^{\infty} \gamma^t \phi(s_t^{(i)}) \tag{4}$$

至此,根据专家期望值 μ_E 便能生成新的优化策略 $\pi^* \sim \mu(\pi^*)$,迭代算法如下:

1. 任意选择一个策略 $\pi^{(0)}$,通过计算或蒙特卡罗近似得到期望值 $\pi^{(0)} = \mu(\pi^{(0)})$,令 $i = 1$。

2. 增加一个新的线性优化问题:

$$\begin{aligned} &\min_{\lambda,\mu} \|\mu_E - \mu\|_2 \\ &\sum_{j=0}^{i-1} \lambda_j \mu^{(j)} = \mu \\ &\text{s.t. } \lambda \geqslant 0 \\ &\sum_{j=0}^{i-1} \lambda_j = 1 \end{aligned}$$

设 $t^{(i)} = \|\mu_E - \mu\|_2, \omega^{(i)} = \dfrac{\mu_E - \mu}{\|\mu_E - \mu\|_2}$。

3. 如果 $t^{(i)} \leqslant \varepsilon$,终止计算。

4. 运用策略 $\pi^{(i)}$ 的回报函数:$R = (\omega^{(i)})^{\mathrm{T}} \phi(s)$,通过强化计算方法生成 $\mu^{(i)} = \mu(\pi^{(i)})$。

5. 令 $i = i + 1$,返回步骤 2。

3　仿真实验与分析

为验证逆向强化学习生成调度方案方法的有效性。我们在 JAVA Script 环境下设计了一款基于离散事件仿真的舰载机甲板调度仿真评估系统,该系统以舰载机作为主要仿真实体,将弹射器、降

落跑道、加油站、弹药装载、升降机、驻机位以及拖车作为主要资源，以命令驱动方式调度舰载机进行各种甲板操作。系统输出包括可视化和任务信息界面，其中甲板环境与网格重合，可提供路径规划和碰撞检测功能；控制台按仿真时间顺序输出任务信息，其系统的输出界面如图2所示。

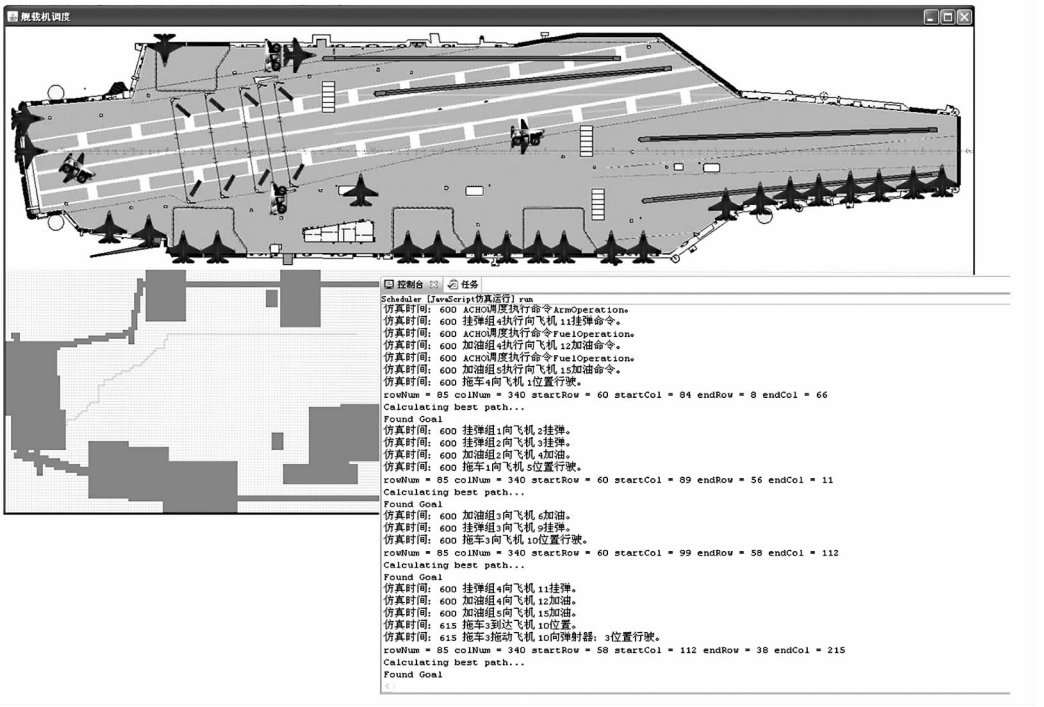

图2 舰载机甲板调度仿真评估系统输出界面

Fig. 2 Interface of aircraft scheduling simulation and evaluation system output

实验想定以两组舰载机出动保障作为例子(第一组5架，第二组10架)。甲板调度从驻机位开始，到所有飞机降落在跑道并成功停驻为仿真结束，优化目标为总调度时间和坠毁概率。两组实验首先由作者作为指挥人员制定命令执行仿真，并将所有系统状态以日志的方式保存下来作为“专家演示”，供逆向强化学习生成“逆向调度策略”，再用它作为命令输入执行仿真系统。仿真系统设定的运行参数如表1所示。

表1 仿真评估系统运行参数表

Tab. 1 Running parameter of simulation system

序号	活动	设施(数量)	操作时间(s)
1	飞机拖运	拖车(4)	Path/v
2	弹药装载	弹药组(5)	60
3	加注燃油	燃油组(5)	45
4	飞行前检查	保障组(4)	30
5	起飞	弹射器(4)	10
6	空中作战	舰载机(n)	5000
7	降落	跑道(1)	30

经仿真实验，两组结果如表2所示。其中，$\mu(E)$为指挥员调度系统得到的策略期望值，$\mu(\pi^*)$为逆向强化学习生成调度策略得到的期望值。

表2 实验结果对比分析表

Tab. 2 Contrast analyze table of experiment result

	第一组(5架)		第二组(10架)	
	$\mu(E)$	$\mu(\pi^*)$	$\mu(E)$	$\mu(\pi^*)$
总调度时间(s)	6245	6307	6958	7083
坠毁概率(%)	0.02	0.03	0.08	0.12

逆向强化学习以专家演示作为学习对象，目的是使生成的调度策略接近指挥人员的调度水平。从表2中的实验结果可以看出，第一组实验由于飞机数量较少，优化结果较第二组更接近$\mu(E)$，这是因为随着实体、状态数量的增加，导致计算量变大，逼近程度有所弱化，符合实际需求。

4 结论

本文将专家演示作为最优学习对象，采用逆向强化学习方法计算回报函数，达到生成优化策略、辅助指挥员调度舰载机甲板作业的目的。提出的方法经仿真验证，具备较好的拟合程度和优化精度，为进一步结合优化算法，开展人在回路的舰载机甲板调度优化研究奠定了基础。

参考文献(References)

[1] 孙诗南. 现代航空母舰[M]. 上海科学普及出版社,1998.
SUN Shinan. Modern aircraft carrier[M]. Shanghai: Shanghai science popularization Press, 1998. (in Chinese)

[2] 司维超,等. 基于 PSO 算法的舰载机舰面布放调度方法研究[J]. 航空学报,2012, 33(1): 1000 - 6893.
SI Weichao. Research on deck-disposed scheduling method of carrier plane based on PSO [J]. ACTA Aeronautica et Astronautica Sinica. 2012, 33 (1): 1000 - 6893. (in Chinese)

[3] 魏昌全,等. 基于出动方式的舰载机航空保障调度模型[J]. 海军航空工程学院学报,2012, 27(1): 111 - 114.
WEI Changquan, et al. Research on the aircraft support scheduling model of carrier-based aircraft based on launch mode[J]. Journal of Naval Aeronautical and As-tronautical University,2012, 27(1): 111 - 114. (in Chinese)

[4] 马登武,等. 基于改进遗传算法的舰载机弹药调度[J]. 计算机工程与应用, 2012,48(8):246 - 248.
MA Dengwu, et al. Modified genetic algorithms for scheduling scheme of carrier-based aircraft ammunition [J]. Computer Engineering and Applications, 2012, 48(8): 246 - 248. (in Chinese)

[5] 冯强,等. 基于 MAS 的舰载机动态调度模型[J]. 航空学报,2009, 30(11): 2119 - 2125.
FENG Qiang, et al. A MAS based model for dynamic scheduling of carrier aircraft [J]. ACTA Aeronautica et Astronautica Sinica, 2009, 30 (11): 2119 - 2125. (in Chinese)

[6] Giardina T J. An interactive graphics approach to the flight deck handling problem[R]. Master's thesis. Monterey: Naval Postgraduate School, 1974.

[7] Johnson A K, Kriston P. A simulation of a computer graphics-aided aircraft handling system [D]. Monterey: Naval Postgraduate School, 1975.

[8] Timothy. Requirements for digitized aircraft spotting (Ouija) board for use on U. S. Navy Aircraft Carriers[D]. Monterey: Naval Postgraduate School, 2002.

[9] Johnston J S. A feasibility study of a persistent monitoring system for the flight deck of U. S. Navy aircraft carriers[D]. Ohio: Department of the Air Force Air University, 2009.

[10] Ryana. Designing an interactive local and global decision support system for aircraft carrier deck scheduling[C]. AIAA Infotech@ Aerospace St. Louis, 2011.

[11] Aircraft Platform Interface Laboratory. Aviation Data Management and Control System[EB/OL]. (2012 - 03 - 15). www. lakehurst. navy. mil.

[12] Jeffrey. A persistent monitoring system to reduce navy aircraft carrier flight deck mishaps [C]. AIAA Guidance, Navigation, and Control Conference, Chicago, Illinois, 10 - 13 August 2009.

[13] Robert K. Outer-loop control Factors for carrier aircraft[R]. Contract Summary Report, Department of the Navy, 1990.

[14] Abbeel , Ng A Y. Apprenticeship learning via inverse reinforcement learning[C]. ICML, 2004.

[15] Abbeel P. Apprenticeship learning and reinforcement learning with application to robotic control[D]. Stanford: Stanford University, 2008.

(上接第 133 页)

[6] 钟慧婷,廖俊必,吴瑞. 一种有效消除超声测量拖尾的新方法[J]. 仪器仪表学报, 2007, 28(6): 1075 - 1079.
ZHONG Huiting, LIAO Junbi, WU Rui. New method of eliminating ultrasonic tailing efficiently[J]. Chinese Journal of Science Instrument, 2007, 28 (6): 1075 - 1079. (in Chinese)

[7] Sabatini A M. Correlation receivers using Laguerre filter banks for modelling narrowband ultrasonic echoes and estimating their time - of - flights [J]. IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control, 1997, 44 (6): 1253 - 1263.

[8] Silva T O E. On the determination of the optimal pole position of Laguerre filters [J]. IEEE Transactions on Signal Processing, 1995, 43(9): 2079 - 2087.

[9] 王跃科,陈建云,张传胜,等. 测量原理[M]. 北京:清华大学出版社,2012.
WANG Yueke, Chen Jianyun, ZHANG Chuansheng, et al. Principle of Measurement[M]. Beijing: Tsinghua University Press, 2012. (in Chinese)

[10] Ohta Y. Realization of input-output maps using generalized orthonormal basis functions[J]. Systems & Control Letters, 2004, 54(6): 521 - 528.

[11] Lee Y W. Statistical theory of communication[M]. New York: Wiley, 1960.

[12] Clowes G. Choice of the time-scaling factor for linear system approximations using orthonormal Laguerre functions [J]. IEEE Transactions on Automatic Control, 1965. 10(4): 487 - 489.

[13] Barkat M. Signal detection and estimation (Second Edition) [M]. Boston: Artech House, 2005.

[14] Ianniello J. Comparison of the Ziv-Zakai lower bound on multipath time delay estimation with autocorrelator performance [C]. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1984.

[15] Weiss A J, Weinstein E. Fundamental limitations in passive time delay estimation——Part I: Narrow-Band Systems[J]. IEEE Transactions on Acoustic, Speech and Signal Processing, 1983, 31(2): 472 - 486.

[16] Yao Z J, Meng Q H, Zeng M. Improvement in the accuracy of estimation the time-of-flight in an ultrasonic ranging system using multiple square-root unscented Kalman filters [J]. Review of Scientific Instruments,2010, 81(1): 1004901 - 1 - 7.

[17] Marioli D, Narduzzi C, Offelli C, et al. Digital time-of-flight measurement for ultrasonic sensors[J]. IEEE Transactions on Instrumentation and Measurement, 1992, 41(1): 93 - 97.