控制与决策
Control and Decision

基于深度/单目融合视觉及强化学习的机器人定位棋局与行棋策略

吴启宇, 谢非, 黄磊, 刘宗熙, 赵静, 刘锡祥

引用本文:
吴启宇,谢非,黄磊,刘宗熙,赵静,刘锡祥. 基于深度/单目融合视觉及强化学习的机器人定位棋局与行棋策略[J]. 控制与决策, 2022, 37(12): 3278−3288.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2021.0756

基于深度/单目融合视觉及强化学习的机器人定位棋局与行棋策略

吴启宇[1], 谢 非[1†], 黄 磊[2], 刘宗熙[1], 赵 静[3], 刘锡祥[4]

(1. 南京师范大学 电气与自动化工程学院，南京 210023；2. 南京林业大学 机械电子工程学院，南京 210037；
3. 南京邮电大学 自动化学院、人工智能学院，南京 210003；4. 东南大学 仪器工程与科学学院，南京 210018)

摘 要: 中国象棋对弈机器人系统实现的关键包括棋局识别定位和自主行棋策略. 首先,针对棋局识别与定位问题,提出一种基于单目相机与深度相机视觉融合的棋局识别定位方法. 该方法利用立体棋子三维特征获取棋子位置,与二维图像识别结果融合计算定位,以提高棋子的识别定位精度. 其次,针对行棋策略问题,提出一种基于深度神经网络与蒙特卡洛树搜索的决策方法. 该方法利用具有终局特征判断的蒙特卡洛树进行搜索,使用优化的随机行棋策略指导模拟行棋,训练具有多尺度及残差结构的策略价值网络模型. 最后,通过自对弈获取训练数据,通过智能体对抗验证、更新模型参数. 实验表明,相较于单目视觉识别,所提出方法具有更高的精确度和稳定性,识别率达到97 %;相较于基准剪枝搜索算法,所提出方法对弈时最多赢得82 %的对局,且所需运算时间缩短41 %.

关键词: 中国象棋；行棋策略；目标检测；深度图像；蒙特卡洛树搜索；强化学习

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.0756

引用格式: 吴启宇, 谢非, 黄磊, 等. 基于深度/单目融合视觉及强化学习的机器人定位棋局与行棋策略 [J]. 控制与决策, 2022, 37(12): 3278-3288.

Chess positioning and playing strategy of robot based on integrated depth/mono vision and reinforcement learning

WU Qi-yu[1], *XIE Fei*[1†], *HUANG Lei*[2], *LIU Zong-xi*[1], *ZHAO Jing*[3], *LIU Xi-xiang*[4]

(1. School of Electrical and Automation Engineering，Nanjing Normal University，Nanjing 210023，China；2. School of Mechanical and Electronic Engineering, Nanjing Forestry University，Nanjing 210037，China；3. College of Automation & College of Artificial Intelligence，Nanjing University of Posts and Telecommunications，Nanjing 210003，China；4. College of Instrument Science and Engineering，Southeast University，Nanjing 210018，China)

Abstract: The key to the realization of the Chinese chess system lies in the board recognition and chess strategy. Firstly, for the problem of chessboard recognition, a method based on the fusion of mono vision and depth vision is proposed. This method designs a chess piece grid recognition network, uses the three-dimensional characteristics of chess pieces to convert the depth image into a chessboard grid, and integrates the chess piece coordinates with the chessboard grid to effectively improve the recognition accuracy of the chessboard. Secondly, aiming at the problem of the chess strategy, a method based on the deep neural network and Monte-Carlo tree search is proposed. This method uses the improved random search strategy with end-game feature judgment to guide the simulation of chess, which trains a policy and value network with residual structure. Finally, the training data is obtained through self-playing, and the parameters are updated and verified through the agent confrontation. Experiments show that compared with mono-only visual recognition, this method has higher accuracy and stability, and the recognition rate reaches 97 %. Compared with the pruning search algorithm baseline, this method wins 82 % of the games, and the computing time is reduced by 41 %.

Keywords: chinese chess；strategy；object detection；depth image；Monte-Carlo tree search；reinforcement learning

收稿日期: 2019-04-29；**录用日期:** 2021-08-26.

基金项目: 国家自然科学基金项目(41974033)；江苏省科技成果转化项目(BA2020004)；江苏省省级工业和信息产业转型升级专项资金项目(JITC-2000AX0676-71)；南京市优势产业关键技术突破招标项目(2018003).

责任编委: 张文安.

†通讯作者. E-mail: xiefei@njnu.edu.cn.

0　引　言

中国象棋对弈机器人系统是面向服务型机器人的一个典型复杂应用,其实现的关键在于棋盘识别定位和自主行棋策略,具有较高的研究价值. 目前,在目标检测领域,除了传统的基于模板匹配的算法[1],已经出现了许多基于卷积神经网络[2]的算法,比如YOLO[3-5]、Faster-RCNN[6]等算法,在多类别目标、多尺度小目标、三维物体检测等方面已有较多研究和应用[7-10]. 但是,其基于二维图像的目标检测方法在中国象棋这种高密度、大数量、小目标、立体化目标的检测任务中,效果不够理想. 同时,国内外机构在棋类自主决策问题方面已有不少的研究,许多基于深度强化学习的智能体,如AlphaGo Zero[11-12]、绝艺等,在围棋对弈领域取得较大研究进展. 此后,基于蒙特卡洛树搜索(Monte Carlo tree search, MCTS)与深度神经网络的决策算法在各种棋类中均有研究应用[13-15]. 在中国象棋行棋策略方面,有许多基于minimax、alpha-beta剪枝搜索的算法[16-17],然而目前少有高水平的中国象棋对弈智能体. 传统的基于剪枝搜索或数据库算法的智能体,棋力受限于数据库体积和边缘计算资源,往往不具备高水平的对弈能力和直接识别棋盘的能力. 一个功能完善可靠的中国象棋机器人系统,需要有基于机器视觉的高精度中国象棋棋局识别能力,以及一个能够进行高水平决策的智能体. 因此,本文研究一种具有自主识别、行棋决策能力的中国象棋对弈机器人系统. 该系统能够利用深度相机和彩色单目相机自主获取棋子种类和位置信息,进行高精确度棋局识别.采用基于深度强化学习的行棋决策算法,给出优化的行棋策略,进行高水平中国象棋行棋决策,形成完整的端到端系统.

本文针对中国象棋定位与行棋策略问题,提出一种基于深度视觉融合定位的中国象棋棋子定位方法和一种基于深度强化学习的行棋策略. 本文方法具有以下优点:

1) 针对单目视觉定位算法难以识别平面特征不显著,但立体特征突出的物体的缺点,提出一种基于深度视觉图像与单目视觉图像的融合计算定位方法. 针对棋子等具备较强立体特征的物体具有较好的定位效果,在保证识别准确率的情况下,大幅减小模型体积,加快识别速度,适用于嵌入式平台部署,具有更高的实用性.同时,此方法具有较强泛用性,可以应用于多数平面特征不明显,但具有明显立体特征与侧面信息的工业物体的识别与定位.

2) 提出一种改进的中国象棋行棋策略,在保证实时性的同时,棋力优于剪枝搜索算法. 使用多特征复合棋盘作为神经网络输入,减少数据量; 使用改进的具有多尺度和残差结构的神经网络,增加网络深度,减少训练时间; 使用改进的具有终局特征判断的蒙特卡洛树搜索算法,能够提升强化学习的泛化能力. 针对中国象棋动作空间大,但高价值动作少的特点,使用改进的价值优化的随机行棋策略,高效探索中国象棋动作空间,提升智能体决策能力.

1　中国象棋定位与行棋策略整体框架

1.1　系统整体框架

基于深度/单目融合视觉及深度强化学习的象棋定位与行棋策略主要由两部分组成,其整体框架如图1所示.

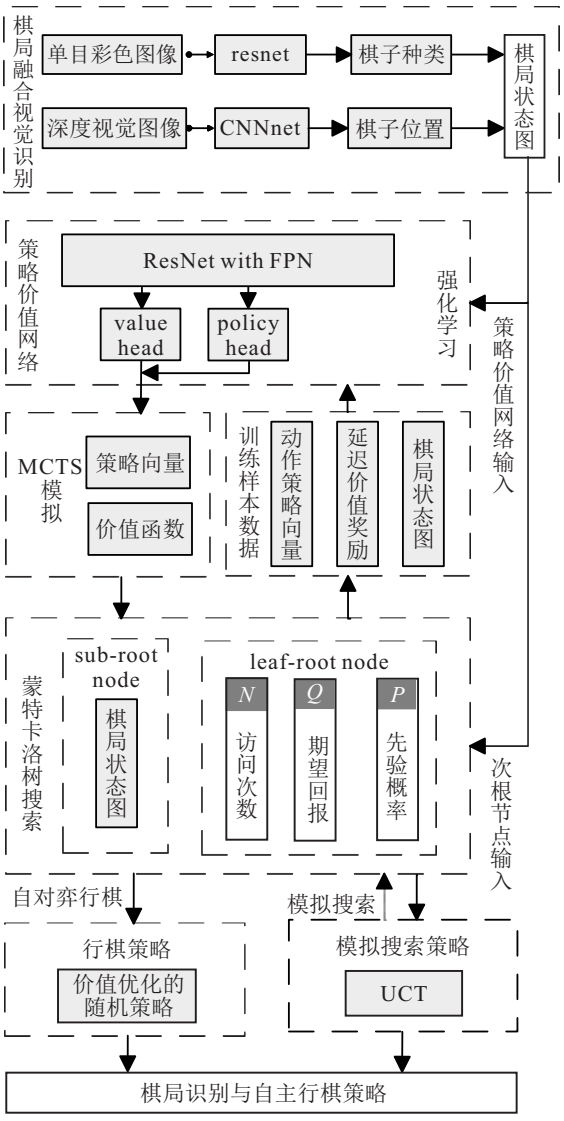


图 1　总体框架

1) 棋局识别部分. 通过彩色摄像头采集棋盘彩色图片,输出棋子目标检测信息. 通过深度相机采集棋盘深度数据,输出棋子定位信息. 两者利用隶属度融合计算,合成棋局信息,经过棋局模型编码之后得

到棋局状态图.

2) 对弈策略部分. 对弈策略结构包含一个深度神经网络和一个蒙特卡洛树. 蒙特卡洛树的子节点记录所有棋局状态图,叶节点记录状态动作对的相关参数. 深度神经网络的输入为棋局状态图,输出为价值函数和策略向量两部分.

1.2 系统整体流程

整体系统分为行棋部分和训练部分. 其流程如图2和图3所示.

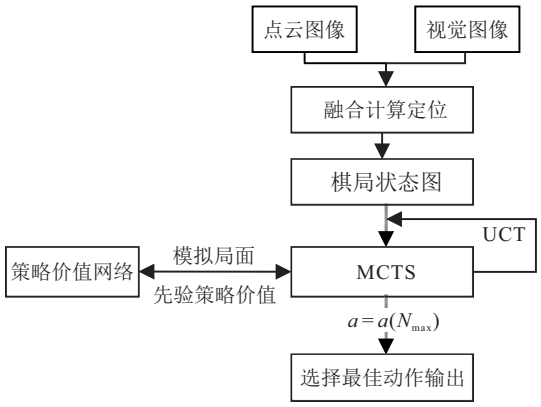


图 2 行棋部分

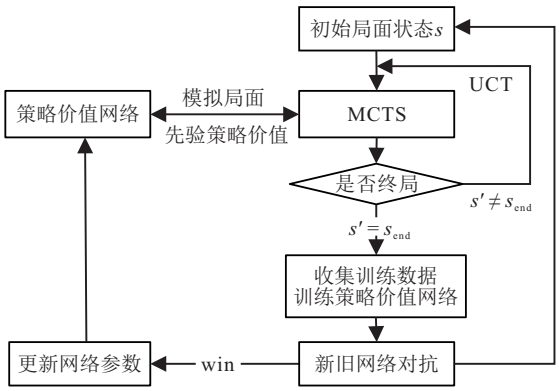


图 3 训练部分

1.2.1 行棋部分

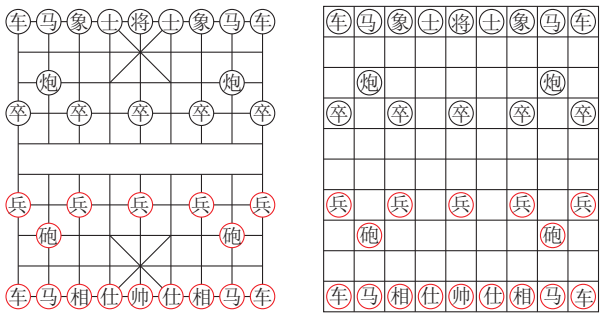
行棋是棋局识别与决策的过程. 如图2所示,首先采集深度视觉图像与单目彩色图像,采用融合定位算法,并编码得到棋局状态图;其次,进行若干次蒙特卡洛树搜索和模拟循环,使用训练完成的策略价值网络输出先验策略及价值作为参考;最后,将随机策略退化为最优动作策略,选择最佳行棋动作并输出.

1.2.2 训练部分

决策智能体训练的过程是自对弈的过程. 如图3所示,首先,从初始棋局状态开始进行若干次蒙特卡洛树搜索与模拟的循环,其中策略价值网络使用随机权值初始化,输出先验策略及价值作为参考;其次,使用改进的随机策略选择行棋动作,执行并进入下一棋局状态,直到进入终局状态;然后,收集MCTS过程中的训练数据,训练策略价值网络;最后,进行新旧网络对抗并更新网络参数. 如此循环迭代直至棋力达到设定目标要求.

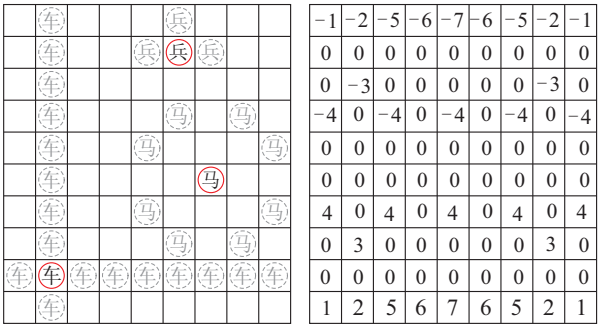
1.3 基于棋子映射的中国象棋多特征复合棋盘建模

中国象棋的状态空间为9×10的网格,其映射如图4所示. 中国象棋共32枚、7种棋子,一次只能移动一枚棋子. 忽略其他棋子的影响,一方16枚棋子最多有115种走法. 例如:“车”在不考虑其他棋子阻挡的情况下,可以移动至另外17个格点,即“车”共有17种可选动作,以此类推. 故中国象棋的最大动作空间为1×115的向量,图5左侧为部分棋子的动作空间示意. 由于其他棋子的阻碍,实际动作空间小于115.



(a) chessboard (b) board gridding

图 4 中国象棋状态空间



(a) action space (b) board state map

图 5 中国象棋动作空间和棋盘状态

中国象棋双方一共14种棋子,包括格点内没有棋子的状态,使用4 bit整数编码这15种状态. 将一盘9×10的中国象棋棋盘,变换为一张9×10的4 bit向量图像,称为棋局状态图. 其中,行棋方棋子按车、马、炮、兵、象、士、将的顺序,分别编码为1至7. 对方棋子编码为-1至-7,空格编码为0. 图4所示初始局面经过编码后得到图5右侧所示的棋局状态图.

1.4 蒙特卡洛树结构

一颗蒙特卡洛树具有两种存储节点,其存储结构如图6所示. 以当前棋局状态作为根节点(root node),每一个可能的棋局状态作为一个子根节点(sub-root node),每一个棋局状态可能采取的动作(action)作为一个叶节点(leaf node).

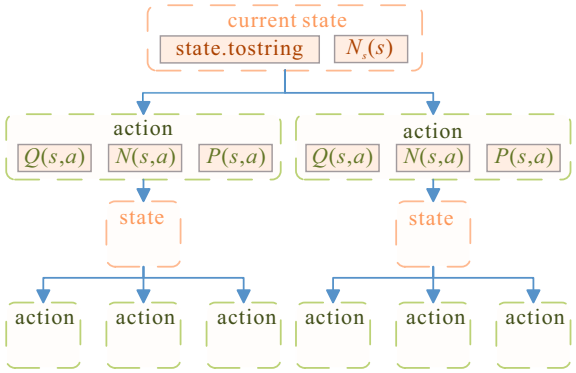


图 6　蒙特卡洛树存储结构

子节点中保存棋局状态图和被访问次数 N_s; 叶节点中保存状态动作对 (s,a) 的3个参数,访问次数 $N(s,a)$、行动价值 $Q(s,a)$ 和先验概率 $P(s,a)$. 其中: $N(s,a)$ 表示在状态 s 采取动作 a 的次数, $Q(s,a)$ 表示在状态 s 采取动作 a 的期望回报, $P(s,a)$ 表示在当前状态 s 可能采取动作 a 的先验概率.

2　中国象棋定位与行棋策略

2.1　基于深度视觉与单目视觉融合的象棋棋局识别

针对以棋子为典型的具有立体特征的物体,本文提出一种基于单目相机与深度相机融合的视觉识别方法. 单目视觉识别采用 Yolov5s 网络,体积小、运算速度快. 输入彩色图像,输出棋子种类 c、棋子中心点位置 $p_c(x_c,y_c)$ 以及棋盘位置框 p_b. 深度视觉图像网络输入为 16 bit 一维深度图. 经过若干层卷积、下采样和全连接操作,输出为 9×10 的棋盘网格图 b_{TOF}, 每一网格为 g_t 及各棋子中心点位置. 之后编码得到棋局状态图 s. 其总体结构如图7所示.

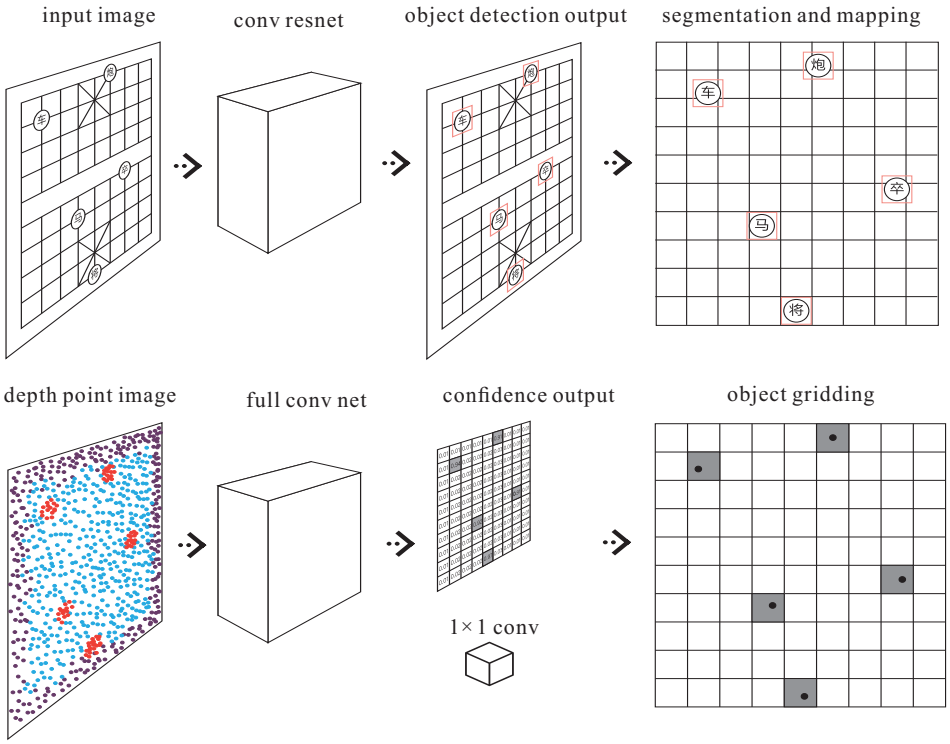


图 7　深度视觉图像与彩色图像识别网络

2.2　基于隶属度融合计算的棋子定位

不同于一般的目标检测任务,棋局定位与识别不仅需要完成目标检测与分类,还需精确定位棋子. 利用深度视觉图像获得棋子真实位置,与视觉图像识别结果进行融合定位计算,得到棋子的高精度确切定位信息,其计算定位过程如图8所示.

首先,将视觉网络得到的棋盘位置框 p_b 透视变换成为俯视图,裁切框外部分,同时变换棋子坐标 p'_c,得到俯视棋盘 p_{bs}; 然后,将俯视棋盘 p_{bs} 分割为9×10的俯视网格图 g,其中心点坐标加上棋子偏移坐标后,记为 (x_{gk},y_{gk}); 最后,计算每一网格 g_t 与棋盘中每

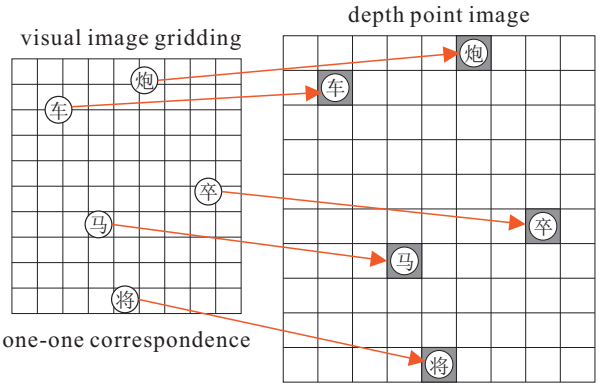


图 8　融合计算定位示意图

一棋子p_c的隶属度$m(g_k,p_c)$. 计算公式如下:

$$m(g_k,p_c)=\lg\sqrt{(x_{gk}-x_c)^2+(y_{gk}-y_c)^2}. \quad (1)$$

这些网格g_k与深度视觉图像网络输出的棋盘网格图$b_{\rm TOF}$中的网格g_t是一一对应的关系,同一位置的隶属度相等,即$m(g_k,p_c)=m(g_t,p_c)$. 在棋盘网格图$b_{\rm TOF}$中,从左上角开始,依次搜索每一网格g_t. 对存在棋子的网格g_{tc},选择隶属度最高的棋子填入网格中,即$g_t(p_c)=m_{\max}(g_{tc},p_c)$.

由于嵌入式边缘平台算力限制,轻量化神经网络往往参数较少,难以处理中国象棋棋局这一密排列、大数量的目标检测任务,容易出现误识别. 同时,由于网络深度较浅,输出目标检测结果往往置信度不高. 但深度视觉图像数据量小,网络结构简单易于训练,能够直接确定棋子的位置和数量. 利用隶属度、IoU匹配、序列置信度等算法,可筛除误识别目标框.

2.3 改进的具有多尺度及残差结构的策略价值网络

最优行动策略选择与价值预测是强化学习决策的关键,使用深度神经网络进行策略价值预测,适用于行棋策略等具有高维状态和动作的决策问题[18]. 采取更深层的网络以及多尺度的输入,能够带来更好的网络性能,缩短所需训练时间与训练样本数量[19]. 为此,本文提出一种具有多尺度特征输入及残差结构的神经网络,其主要特征如图9所示.

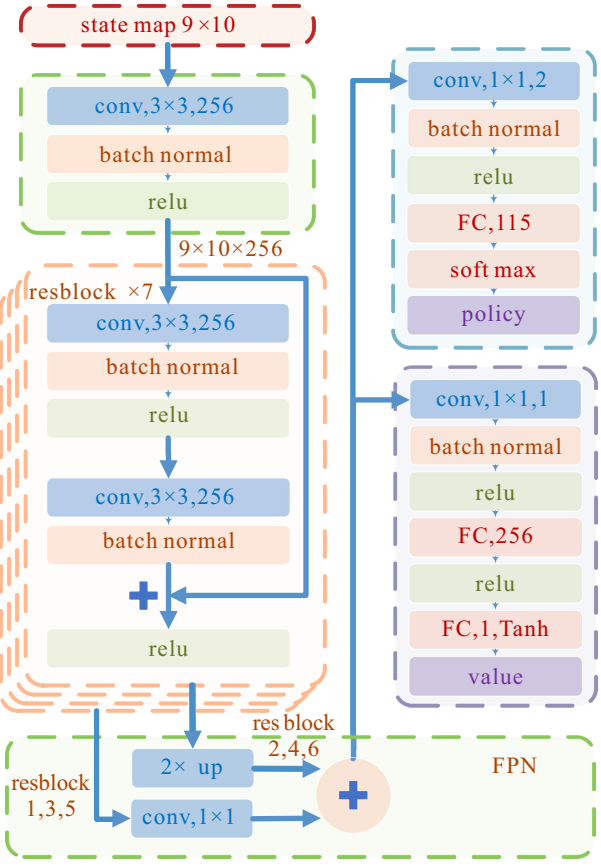


图 9 策略价值网络结构

1) 策略价值网络的输入为一个9×10的4 bit棋局状态图s. 策略价值网络将行棋方的棋局状态作为输入,从中提取价值与策略特征.

2) 策略价值网络的输出分为两部分: 第1部分为价值输出v_θ,表示神经网络对当前棋局状态胜负的预测; 第2部分为策略输出p_θ,为1×115的归一化概率向量,表示采取各可能动作的概率. 具体如下所示:

$$\begin{cases} p_\theta=f_\theta(s)=\overbrace{[p_1,p_2,p_3,\ldots]}^{1\times115,\,\rm FP64},\\ v_\theta=f_\theta(s)\in[-1,1]. \end{cases} \quad (2)$$

3) 策略价值网络由1个初始卷积层、7个残差卷积模块、FPN和2层全连接组成,其策略输出和价值输出具有共享的权重参数.

2.4 改进的具有终局特征判断的蒙特卡洛树搜索

蒙特卡洛树搜索对动作空间进行采样,利用强化学习方法输出的策略价值参考进行动作选择和价值评估,并提供策略价值网络的训练样本,二者相辅相成. 蒙特卡洛树搜索循环过程,是若干次对当前棋局状态的模拟走子动作过程. 一次蒙特卡洛树搜索的过程,以当前棋局状态$s_{\rm cur}$作为模拟搜索的起点,以棋局终局的状态或者不在蒙特卡洛树中的状态作为模拟搜索的终点.

这一过程共包括3个阶段: 探索阶段、扩展阶段和回溯阶段. 每次行棋前进行若干次循环,扩充蒙特卡洛树,并输出一个策略向量,为行棋采用的随机策略提供参考. 利用终局特征判断能够得到更加准确的策略价值,从而为神经网络提供更加精确的策略价值样本[20]. 本文提出一种改进的具有终局特征判断的MCTS算法,能够解决中国象棋特有的“困毙”特殊终局状态问题,同时提升智能体的决策性能. 如图10所示,下文分别说明MCTS的3个阶段.

2.4.1 MCTS探索阶段

探索阶段是模拟行棋的过程. 从当前状态$s_{\rm cur}$开始,从所有合法的动作空间中选择一个最佳动作a并模拟行棋,进入新的状态s'.

选择最佳动作a时,为平衡探索的深度与准确度,本文采用UCT算法[21-22]计算动作空间a_s中所有动作的价值. UCT的计算公式如下:

$$U(s,a)=Q(s,a)+c_{\rm uct}P(s,a)\frac{\sqrt{\sum\limits_b N(s,b)}}{1+N(s,a)}, \quad (3)$$

其中: $U(s,a)$是在状态s下采取动作a的综合价值; 对于动作a, $Q(s,a)$是期望回报, $P(s,a)$是先验概率, $N(s,a)$是累计模拟次数, $\sum\limits_b N(s,b)$是状态s的动作

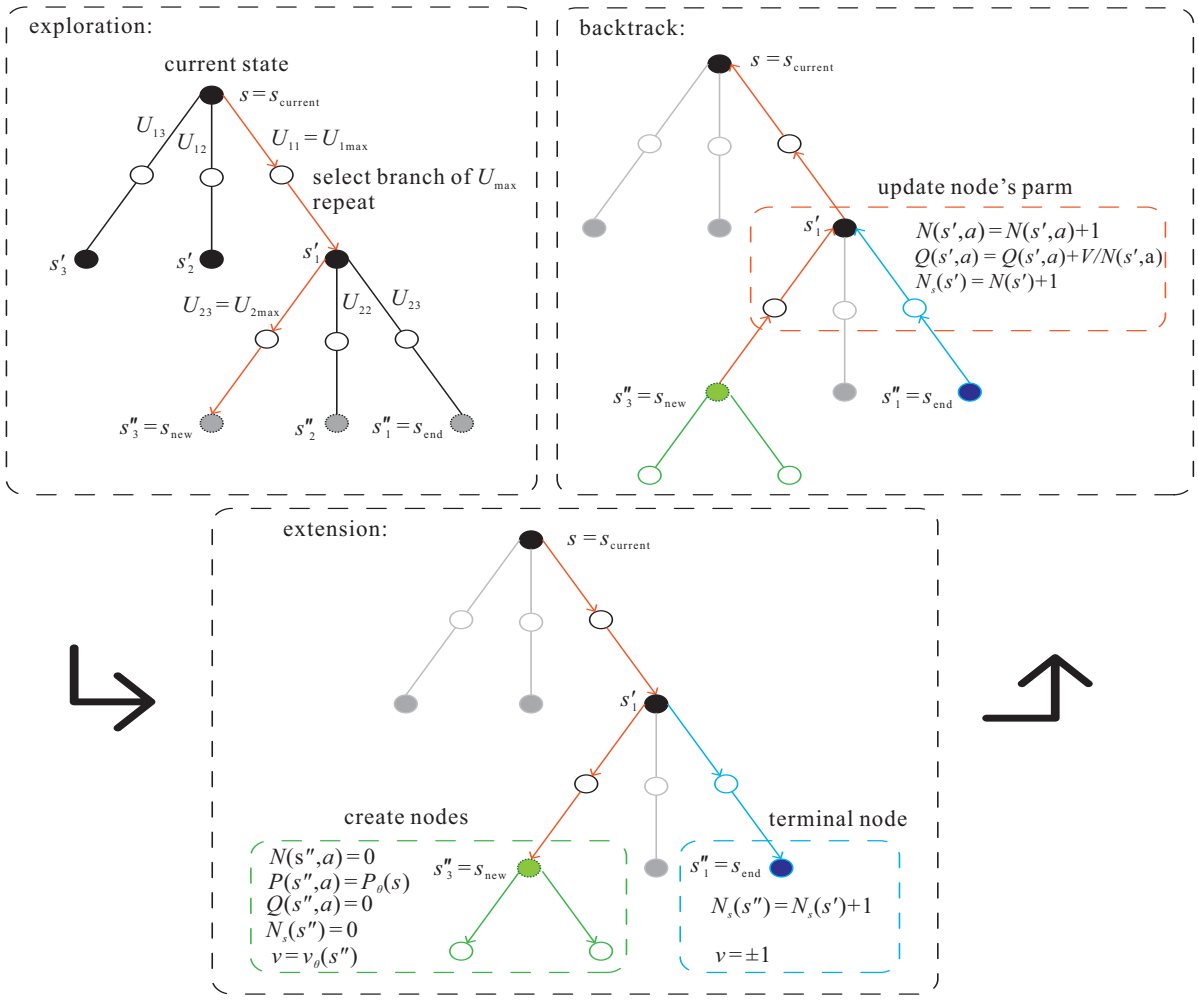


图 10　蒙特卡洛树搜索 3 个阶段

空间 a_s 中除去动作 a 的所有动作累计模拟次数之和; c_{uct} 是超参数,其值越大,越倾向于探索模拟次数较少的动作. 式 (3) 中第 1 项是行动价值,第 2 项是频数价值,模拟次数低的动作具有更高的频数价值. UCT 方法相比随机-贪婪策略 (ϵ-greedy) 方法,前者能够通过控制超参数 c_{uct} 调节探索与利用的平衡,选择频数价值与先验概率综合得分最高的动作. 在有限的搜索次数和计算资源下,此方法能够增强MCTS对于中国象棋动作空间的探索能力,优化搜索速度.

2.4.2　MCTS扩展阶段

在探索阶段重复选择动作并模拟行棋后进入扩展阶段,此时存在两种情况. 若遇到终局状态 s_{end},则利用下式得到终局状态的延迟奖励 v_{end}:

$$v = v_{\text{end}} = \begin{cases} 1, \text{ won;} \\ -1, \text{ otherwise.} \end{cases} \tag{4}$$

本文提出一种针对中国象棋的终局状态判断. 因为中国象棋存在“困毙”这一特殊终局状态,所以不使用行棋方的“将”棋子作为判断终局状态的依据,而对每一 s_{new} 做终局特征判断操作. 首先,若行棋方处于“被将军”状态,则遍历合法动作空间模拟进行“应将”动作,若均不能合法“应将”,则判断此状态为终局状态 s_{end}. 终局特征判断使终局状态提前,有利于决策智能体更好地学习“防御”动作和更深层次的“进攻”动作.

若状态 s_{new} 不为终局状态且不在蒙特卡洛树中,则扩展蒙特卡洛树. 新建 1 个子节点,存储棋局状态图 s_{new} 的字符串编码,频次为 $N_s(s_{\text{new}})$. 同时新建若干个叶节点,存储频次 $N(s_{\text{new}}, a)$、先验概率 $P(s_{\text{new}}, a)$ 和行动价值 $Q(s_{\text{new}}, a)$,新建叶节点的数量与状态 s_{new} 的合法动作空间 $a_{(s_{\text{new}})}$ 大小对应相等. 然后将 s_{new} 输入策略价值网络,得到策略输出 $p = p_\theta(s_{\text{new}})$ 和价值输出 $v = v_\theta(s_{\text{new}})$. 然后按照下式对新建子节点和叶节点中的参数进行初始化:

$$\begin{cases} N(s_{\text{new}}, a) = 1, \\ P(s_{\text{new}}, a) = p_\theta(s_{\text{new}}), \\ Q(s_{\text{new}}, a) = 0, \\ N_s(s_{\text{new}}) = 1, \end{cases} \tag{5}$$

同时得到棋局状态 s_{cur} 对应的延迟奖励 v.

2.4.3 MCTS回溯阶段

完成蒙特卡洛树扩充,并在得到延迟奖励v之后进入回溯阶段. 其任务是更新蒙特卡洛树中相关子节点和叶节点的参数值. 从新建状态$s_{\rm new}$或终局状态$s_{\rm end}$(即探索阶段的终点),沿选择的动作a回溯至当前起始状态$s_{\rm cur}$,更新这一路径中所有的子节点s_p和叶节点(s_p,a_p)的参数. 更新方法如下所示:

$$\begin{cases} N(s_p,a_p)=N(s_p,a_p)+1,\\ Q(s_p,a_p)=Q(s_p,a_p)+\dfrac{v}{N(s_p,a_p)},\\ N_s(s_p)=N_s(s_p)+1. \end{cases} \tag{6}$$

其中:v为扩展阶段得出的延迟奖励,s_p为回溯过去经历的状态,a_p为回溯过去采取的动作.

在若干次搜索中,总是选择$U(s,a)$值最大的动作并模拟. 每次搜索循环结束之后都会对终局或者新局面的奖励进行回溯,如果之前选择的动作分支能够得到正向的延迟奖励,则鼓励继续探索这一动作分支. 反之,若一个$U(s,a)$值较大的动作分支在终局状态中得到负向的延迟奖励,则这个动作的$U(s,a)$值下降,MTCS自动倾向于探索其他的分支.

2.5 改进的价值优化的随机行棋策略

行棋时需要一种策略,从动作空间a_s中选择一个最佳行棋动作. 中国象棋的动作空间大,但高价值动作少,动作价值相差较大. 因此,需要在训练初期鼓励探索,同时在后期着重对高价值动作深入探索.

为满足这一要求,本文提出一种改进的价值优化的随机策略,其计算公式如下:

$$\pi(a)=\frac{N(s,a)^{\frac{1}{\tau}}}{\sum\limits_b N(s,b)^{\frac{1}{\tau}}}\varepsilon N(s,a)\lg N(s,a). \tag{7}$$

其中:$\pi(a)$是动作a的先验概率;$N(s,a)$是状态动作对(s,a)被模拟的次数;$\sum\limits_b[N(s,b)]$是动作空间s_a中,动作a以外的其他动作被模拟次数之和;τ是温度参数. 一个动作被模拟的次数越多,表明其通过UCT计算出的价值越高,被随机策略选中的概率越高. τ值越大,不同动作之间价值差越小,有利于探索更多的动作. 可以通过控制温度参数τ的大小,控制行棋策略中探索的程度. ϵ为特定方向控制参数,其值应随着棋局的深入不断提高,强化对中国象棋中少数高价值动作的探索.

2.6 策略价值网络训练与损失函数

2.6.1 训练样本数据获取

神经网络训练所需的样本数据由自对弈产生与记录. 训练数据$T(n)$是一组形如$[s,p_{a(s)},v(s,p_a)]$的三元组列表. 其中:s为某一特定局面状态;$p_{a(s)}$为特定状态s下,采取动作a_s的策略向量;$v(s,p_a)$表示在状态s下采取动作策略p_a的最终胜负. 首先从初始棋局状态s_0开始,进行k次模拟;然后进行真实行棋,利用随机策略从当前棋局状态s_t的合法动作空间中选择一个动作,得到状态s_{t+1},并取反得到相对于对方的棋局状态图s_{t+1}^{-1}.称这一过程为一个自对弈行棋循环,交替行棋直至$s_t=s_{\rm end}$,循环若干轮收集足够的训练样本数据,送入策略价值网络训练.

2.6.2 损失函数

策略价值网络包含两个分支输出头:policy head输出p_θ和value head输出v_θ. 因此,设计策略价值网络的损失函数

$$L_\theta=\sum_t(v_\theta(s_t)-z_t)^2-\pi_t\log(p_\theta(s_t)). \tag{8}$$

其中:损失函数的前项为价值输出v_θ的均方误差,后项为策略向量输出p_θ的交叉熵损失.

2.6.3 策略价值网络参数的更新与验证

在旧网络参数$f(\theta)$的基础上,继续训练得到新网络参数$f(\theta')$. 通过智能体对抗来验证$f(\theta')$的棋力. 若新智能体赢得超过52%的对局,则接受$f(\theta')$. 不断循环进行训练与对弈,直至棋力达到预定目标.

3 实验与分析

为验证中国象棋识别与决策的效果,设计以下实验内容. 验证基于单目/深度相机融合定位计算方法的有效性和优势,验证改进的行棋策略的效果,对比本文算法与不同基准算法的决策能力.

3.1 实验平台介绍与数据增强方法分析

本文的实验平台包括一台服务器,使用Ubuntu 18. 04,搭载NVIDIA Titan RTX GPU,一台笔记本计算机,无GPU. 使用Intel RealSense D435i传感器,包含一个深度相机和一个彩色相机. 中国象棋自主对弈机器人实验平台如图11所示.

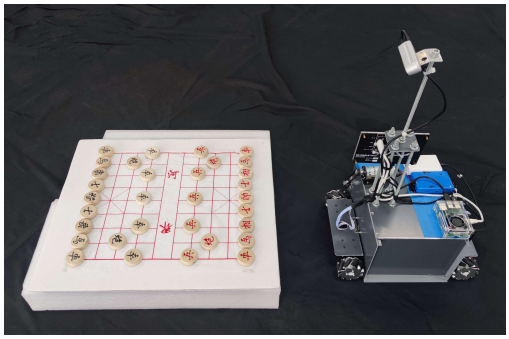


图11 中国象棋自主对弈机器人实验平台

共采集2 800张原始棋局图片,保留280张作

为验证集. 包含棋子数量从 32 到 2 枚, 每类超过 50 张. 共标定 15 类目标, 包括 14 种棋子和棋盘. 采集 4 000 张深度图像数据集, 并标定棋子网格位置. 图 12 和图 13 是数据集标定的部分内容.

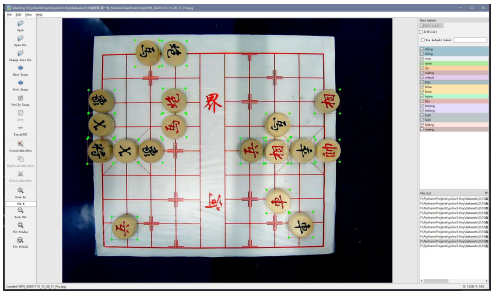
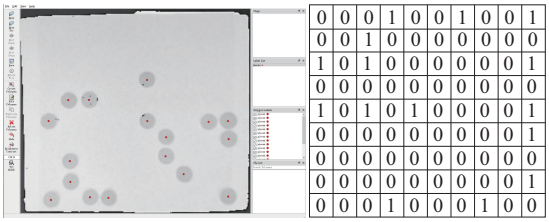


图 12　单目图像数据集标注



0	0	0	1	0	0	1	0	0	1
0	0	1	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0
1	0	1	0	1	0	0	0	0	1
0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1
0	0	0	1	0	0	0	1	0	0

图 13　深度图像数据集标注

利用中国象棋棋盘的对称性与网格状分布特性, 采用方向旋转、随机网格遮盖等数据增强方法对原始训练集进行扩充. 将棋盘旋转 90°、180° 或 270°, 同时使用白色色块对棋子进行随机遮盖, 获得棋子数量较少的棋局图像, 成倍扩充数据集. 2 800 张原始图像经过数据增强处理后, 扩充为约 25 000 张图像的训练集. 数据增强方法如图 14 所示.

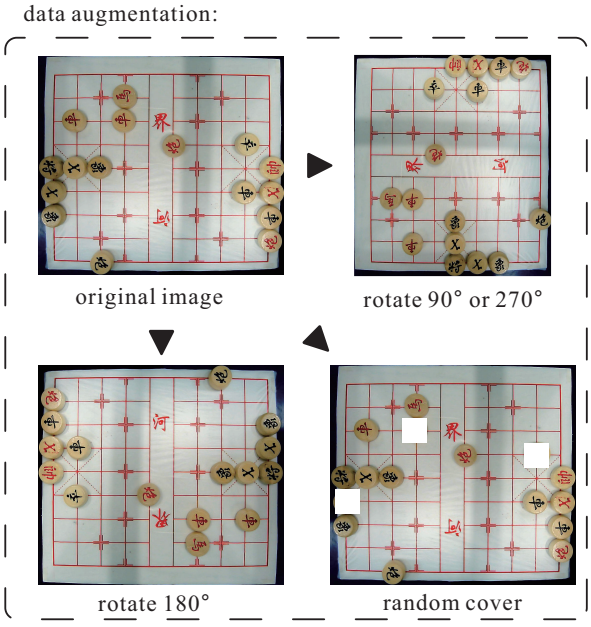


图 14　旋转与随机遮盖数据增强示意图

3.2　棋局识别与定位效果分析

为验证本文算法针对棋局识别与定位问题的有效性, 以较多棋子数量的一盘棋局为例, 图 15 左侧为深度相机拍摄图像与定位结果, 右侧为单目视觉棋盘识别效果. 实验结果表明, 此方法能够实现大量中国象棋棋子的准确定位与识别.

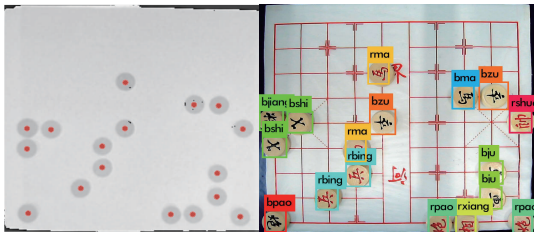


图 15　深度视觉图像与彩色图像融合计算定位示例

使用深度相机获取准确位置信息, 可以修正二维图像的识别结果, 达到更高的识别定位准确率. 为验证本文提出的深度相机融合计算定位方法对于小型网络的准确率提升效果, 使用 Yolov5s 轻量化模型进行二维图像识别, 对棋局进行目标识别与修正实验.

如图 16 所示, 左下方棋子置信度仅为 0.61, 易被置信度阈值摒除, 形成漏识别. 使用深度图像确定棋子的确切位置后, 该棋子虽然置信度偏低, 但对应网格的隶属度较高, 融合计算后能够被正确识别. 如图 17 所示, 左上方棋子出现置信度相近的重框, 但误识别目标框置信度略高. 若使用常用的 nms 方法, 易遗漏正确目标, 造成棋子定位偏离真实位置. 使用深度图像获取棋子真实位置之后, 即可保留隶属度更高的正确目标框. 如图 18 所示, 左上方棋子位于图像边缘,

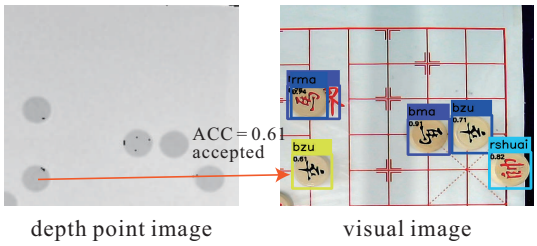


图 16　正确识别低置信度目标

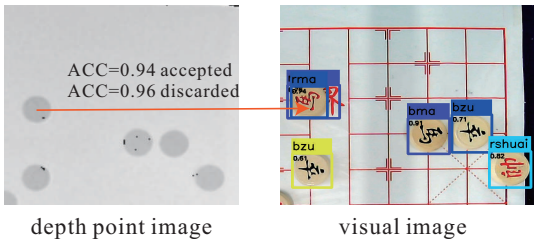


图 17　修正重复识别目标框

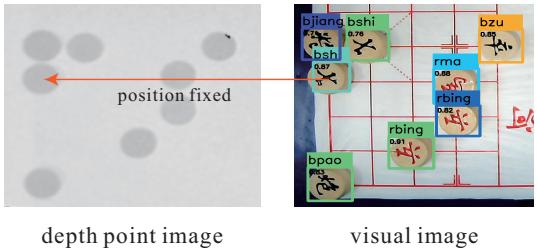


图 18　棋子目标框位置修正

受畸变影响,目标框位置不够准确,易造成棋盘映射错误. 利用深度图像获取精确位置信息,融合计算后可以修正目标位置.

为了评估本文所提出深度图像融合计算方法在棋局识别中的优势,使用本文算法与仅使用模板匹配方法、Yolov5、Faster-RCNN、NanoDet等方法对比,测试数据为280张验证集,对比平均置信度、类别准确度、定位准确率、用时(实时性)、运算量等. 由表1可以看出,模板匹配由于不具有旋转不变性,在随机摆放的测试集中表现不理想. 而仅使用二维平面图像对棋盘进行定位时,虽然在中大型网络如Yolov5x、Faster-RCNN中,取得了不错的平均置信度成绩,但是由于其定位框不够准确,在实际完成棋盘识别定位任务时,棋子定位准确度相较目标检测准确率有所降低,达不到理想的棋盘生成效果. 使用深度图像网络搭配轻量化神经网络(如Yolov5s),采用本文融合定位算法识别棋局时,其定位准确度超过大型网络Yolov5x,同时运算时间大幅缩短. 配合深度图像进行融合定位时,平均置信度有所降低,这是因为深度图像定位算法减少了重框和较高置信度的误识别现象,从而提升了总体的定位准确率.

表1 本文算法与其他算法识别结果性能对比

对比模型	平均置信度	类别准确率/%	定位准确率/%	用时(ms)	FLOPS
模板匹配	—	66	66	10	—
Yolov5s	0.78	95.5	92	80	17B
NanoDet	0.65	92.3	88	56	1.2B
Yolov5m	0.88	97	95.5	150	21.4B
Yolov5x	0.93	97.9	97.1	420	87.7B
Faster-RCNN-Resnet50	0.92	97	96	1120	3.8B
深度图像网络+Yolov5s	0.75	95.5	98	170	20B
深度图像网络+NanoDet	0.63	92.3	96.3	136	4.2B

实验表明,本文算法搭配轻量化Yolov5s模型时,棋子定位准确度达到98%,略优于Yolov5x. 其网络模型较小,运算速度快,能够在不依赖GPU算力的边缘计算平台上,达到或超过大型网络的中国象棋棋局识别效果.

3.3 策略价值网络训练与对弈分析

本文根据回合数动态设置随机策略的探索参数. 在自对弈前10手,设置温度参数τ为1. 在第10手至30手,设置温度参数τ从1逐步降低为0.2,在行棋前中期使得各个动作分支概率分布更加平均,各动作分支之间的差异变小,以此在对弈的初期鼓励探索的广度. 在30手之后,则将此参数逐渐减小至0或无穷小,以此在对弈的后期充分利用前期取得的先验价值,增加探索的深度. 随着棋局的深入,不断提升定向探索参数ϵ,在行棋后期提升对高价值动作的探索程度.

3.3.1 策略价值网络训练分析

本文使用1 700轮迭代训练神经网络,每轮迭代使用100盘对弈产生样本训练数据,每次行棋进行550次MTCS模拟,训练时间约19 h. 最终神经网络的损失函数值持续下降,稳定收敛.

中国象棋开局具有如下特点:可选动作多,但高价值动作少. 因此在行棋的前两步,即蒙特卡洛树的前两层内,使用固定的8个动作空间,而不去探索整个动作空间,以此加快自对弈数据采集速度.

为验证本文所提出行棋策略的可行性与有效性,将本文算法与基于策略树剪枝搜索的基准算法作比较,对比平均步长耗时、胜率、平均回合数. 使用本文算法分别进行300、550和800次MCTS模拟,与搜索深度为2、3和4的α-β剪枝搜索算法及深度为2、3的minimax算法分别进行100盘对弈. 由表2可以看出,本文算法在300次MCTS模拟时,智能体棋力已经优于深度为2的剪枝搜索算法. 本文算法单步进行800次模拟时,相比其他算法,在基准胜率方面取得一定的优势,最高赢得82%的对局,同时运算时间缩短41%.

表2 算法对比基准胜率

模型	基准算法	平均步长耗时/s	基准耗时/s	基准胜率/%	平均回合
本文-800	α-β-depth 2	2.6	4.4	82	34
本文-800	α-β-depth 3	2.6	6.3	75	45
本文-800	α-β-depth 4	2.6	13.2	72	49
本文-550	α-β-depth 3	2.1	6.3	62	46
本文-300	α-β-depth 2	1.2	4.4	55	57
本文-800	minimax-depth 2	2.6	5.9	80	36
本文-800	minimax-depth 3	2.6	18.9	78	43

同时,为验证本文基于深度强化学习的行棋策略对于全局价值判断与决策的效果,采用本文算法-800与几种其他算法对比平均领先价值.

由表3可以看出,在面对数据库的算法时,由于中国象棋开局棋谱基本在数据库中全面收录,开局往往不利,但进入中盘后,数据库无法全面收录,本文算法具有较大领先. 在对比α-β博弈树算法时,面对棋力更弱的深度2算法,其平均子力价值领先不如棋力更高的其他算法. 经分析可以得出,基于局面价值的博弈树算法在搜索深度较浅时一般倾向于选择得子,即局部最优. 而本文算法虽然子力优势不大,但能够在中局时选择全局更优的行棋策略,往往在局面优势不大的情况下取得棋局的胜利.

表 3　平均局面领先价值对比

对弈基准算法	前 15 手	前 30 手	终局
基于数据库	−1.3	2.6	3
α-β-depth 2	0.87	−0.13	0.45
α-β-depth 4	0.19	1.12	1.45
minimax-depth 2	0.90	0.26	1.87

为验证本文所提出具有终局特征判断与价值优化随机策略的蒙特卡洛树搜索效果,将不同方法训练出的网络模型进行对比. 其相较于深度为 2 的 α-β 剪枝搜索算法的基准胜率如图 19 所示. 可以看出,本文改进的具有多特征输入、终局判断及特定方向优化的随机行棋策略,能够有效减少训练所需轮次与时间,在更快的时间内取得神经网络的收敛,最终收敛胜率超过原基本算法.

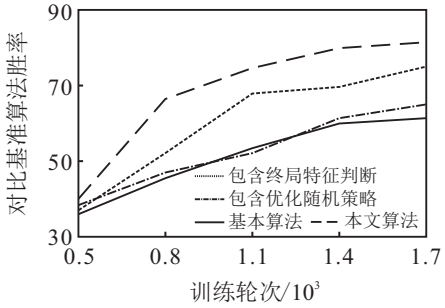


图 19　本文改进算法的基准胜率对比

3.3.2　智能体行棋策略分析

为分析本文所提出自主行棋策略效果,进行残局行棋分析,其中一盘典型残局的部分走子策略如图 20 所示,智能体能够采取正确的行棋策略应对中国象棋特有的“困毙”棋局.

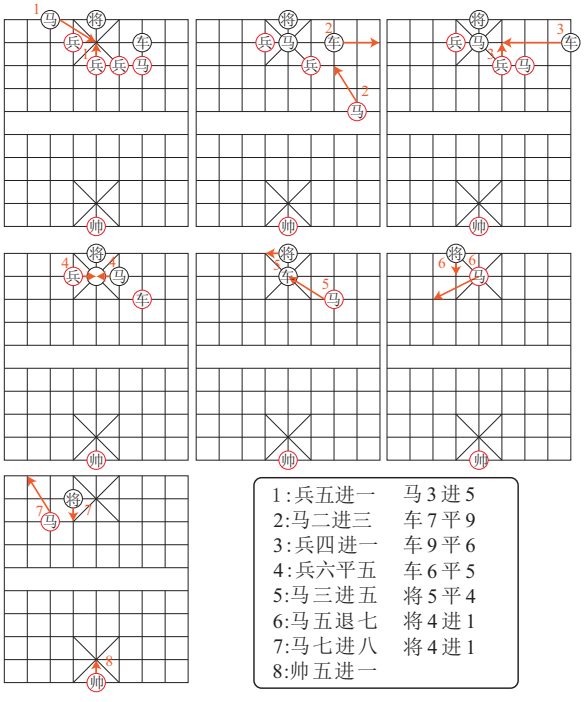


图 20　残局行棋策略

为验证本文算法在选取全局最优动作上的优越性,设计残局破译对比实验. 中国象棋往往高价值动作较少,且存在连贯性和路径依赖性,突出体现在残局破解过程中,往往采用“弃子争先”的策略以取得全局最优. 这是基于策略与价值预测的行棋策略擅长的地方,从中国象棋云库中随机抽取 100 盘残局,选取不同搜索深度的基于博弈树的算法进行对比,记录破解成功率、对局用时、消耗回合数和局面领先价值.

由表 4 可以看出,本文算法用时较短,且取得较高的破解成功率,能够较好地判断棋局状态价值,选择最优动作. 其中,α-β-depth4 棋力与本文算法接近,而 α-β-depth2 用时与本文算法接近. 可以看出,基于博弈树剪枝的算法随着搜索深度的增加,其破解成功率显著上升,原因是残局中棋子数量普遍较少,且总回合数不多. 同时,破解成功率较高的算法,其平均子力领先往往为负数,意味着此算法能够避免局部最优,而选取全局最优的动作.

表 4　残局破解性能对比

算法	破解成功率 /%	对局平均用时 /s	平均消耗回合数	平均局面领先价值
本文-800	92	18	10.2	−1.1
α-β-depth 2	77	17	12.6	0.6
α-β-depth 3	89	29	10.3	−0.5
α-β-depth 4	94	44	9.2	−1.6
minimax-depth 2	77	46	12.6	0.3

4　结　论

为实现中国象棋的自主定位识别与人机对弈系统,本文提出了基于深度相机融合定位的中国象棋棋子定位方法和一种基于深度强化学习的行棋策略. 利用彩色单目与深度相机相融合识别棋局信息,深度相机能够更好地处理棋子等立体物体的识别. 利用改进的蒙特卡洛树搜索与具有多尺度及残差结构的深度神经网络相结合进行棋局对弈,参考神经网络输出进行蒙特卡洛树搜索,极大地提高了搜索效率. 实验结果表明,对于目标密度大、数量多、更加立体的棋局识别,本文方法具有更好的识别效果. 相比于传统的中国象棋博弈树剪枝搜索算法,本文方法所需的运算资源更少、对弈水平更高. 后续工作将增强定位识别算法的适应性与应用性,并围绕实际需求,针对特定工业行业物体等立体特征明显的物体定位与识别作出进一步完善,着重于将定位与决策系统应用于嵌入式设备,提升系统的使用能力和应用水平.

参考文献(References)

[1] 郭晓峰, 王耀南, 周显恩, 等. 中国象棋机器人棋子定位与识别方法[J]. 智能系统学报, 2018, 13(4): 517-523.
(Guo X F, Wang Y N, Zhou X N, et al. Chess-piece localization and recognition method for Chinese chess robot[J]. CAAI Transactions on Intelligent Systems, 2018, 13(4): 517-523.)

[2] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[3] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal speed and accuracy of object detection[EB/OL]. 2020, arXiv: 2004.10934.

[4] Wang C P, Wang H Q, Yu F J, et al. A high-precision fast smoky vehicle detection method based on improved Yolov5 network[C]. 2021 IEEE International Conference on Artificial Intelligence and Industrial Design. Guangzhou, 2021: 255-259.

[5] Feng Z, Guo L, Huang D, et al. Electrical insulator defects detection method based on YOLOv5[C]. The 10th Data Driven Control and Learning Systems Conference(DDCLS). IEEE, 2021: 979-984.

[6] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[7] 赵静, 王弦, 王奔, 等. 基于神经网络的多类别目标识别[J]. 控制与决策, 2020, 35(8): 2037-2041.
(Zhao J, Wang X, Wang B, et al. Multi-category target recognition based on neural network[J]. Control and Decision, 2020, 35(8): 2037-2041.)

[8] 王殿伟, 杨旭, 韩鹏飞, 等. 复杂背景下全景视频运动小目标检测算法[J]. 控制与决策, 2021, 36(1): 249-256.
(Wang D W, Yang X, Han P F, et al. Panoramic video motion small target detection algorithm in complex background[J]. Control and Decision, 2021, 36(1): 249-256.)

[9] 刘芳, 吴志威, 杨安喆, 等. 基于多尺度特征融合的自适应无人机目标检测[J]. 光学学报, 2020, 40(10): 133-142.
(Liu F, Wu Z W, Yang A Z, et al. Multi-scale feature fusion based adaptive object detection for UAV[J]. Acta Optica Sinica, 2020, 40(10): 133-142.)

[10] 王刚, 王沛. 基于深度学习的三维目标检测方法研究[J]. 计算机应用与软件, 2020, 37(12): 164-168.
(Wang G, Wang P. 3d object detection method based on deep learning[J]. Computer Applications and Software, 2020, 37(12): 164-168.)

[11] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge[J]. Nature, 2017, 550(7676): 354-359.

[12] 唐川, 陶业荣, 麻曰亮. AlphaZero原理与启示[J]. 航空兵器, 2020, 27(3): 27-36.
(Tang C, Tao Y R, Ma Y L. Principle and enlightenment of AlphaZero[J]. Aero Weaponry, 2020, 27(3): 27-36.)

[13] Hsueh C H, Wu I C, Chen J C, et al. AlphaZero for a non-deterministic game[C]. 2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI). Taichung, 2018: 116-121.

[14] Li M Y, Huang W Z. Research and implementation of Chinese chess game algorithm based on reinforcement learning[C]. The 5th International Conference on Control, Robotics and Cybernetics (CRC). Wuhan, 2020: 81-86.

[15] Xiao C J, Zhu T, Lin C, et al. Applying determinized MCTS in Chinese military chess[C]. The 26th Chinese Control and Decision Conference (2014 CCDC). Changsha, 2014: 3941-3946.

[16] Takada K, Iizuka H, Yamamoto M. Reinforcement learning to create value and policy functions using minimax tree search in hex[J]. IEEE Transactions on Games, 2020, 12(1): 63-73.

[17] Zhao Z C, Wu S Z, Liang J, et al. The game method of checkers based on alpha-beta search strategy with iterative deepening[C]. The 26th Chinese Control and Decision Conference (2014 CCDC). Changsha, 2014: 3371-3374.

[18] Naeem M, Rizvi S T H, Coronato A. A gentle introduction to reinforcement learning and its application in different fields[J]. IEEE Access, 2020, 8: 209320-209344.

[19] Cazenave T. Residual networks for computer go[J]. IEEE Transactions on Games, 2018, 10(1): 107-110.

[20] Nakayashiki T, Kaneko T. Learning of evaluation functions via self-play enhanced by checkmate search[C]. 2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI). Taichung, 2018: 126-131.

[21] Xiao F, Liu Z Q. Modification of UCT algorithm with quiescent search in computer GO[C]. 2010 International Conference on Technologies and Applications of Artificial Intelligence. Hsinchu, 2010: 481-484.

[22] Sironi C F, Winands M H M.Comparing randomization strate gies for search-control parameters in Monte-Carlo tree search[C]. 2019 IEEE Conference on Games. IEEE, 2019: 1-8.

作者简介

吴启宇(2000−), 男, 硕士生, 从事机器学习、目标检测的研究, E-mail: wqy11888@126.com;

谢非(1983−), 男, 副教授, 博士, 从事机器学习、视觉识别等研究, E-mail: xiefei@njnu.edu.cn;

黄磊(1975−), 男, 副教授, 博士, 从事嵌入式系统设计、SLAM等研究, E-mail: huanglei@njfu.edu.cn;

刘宗熙(2000−), 男, 硕士生, 从事深度学习、机械臂控制的研究, E-mail: 1754627980@qq.com;

赵静(1983−), 女, 副教授, 博士, 从事非线性系统故障诊断、自适应控制等研究, E-mail: zhaojing@njupt.edu.cn;

刘锡祥(1976−), 男, 教授, 博士生导师, 从事机器学习、自动控制等研究, E-mail: scliuseu@163.com.

(责任编辑：孙艺红)