《基于众包和仿真平台的人机合作学习系统》

项目研究报告

项目研究摘要

应用众包的数据库建立模式，研究了在双人合作的 2D 游戏中，计算机能否在搜索数据库而不是复杂编程的环境下实现与人类玩家的配合。数据库是通过两名人类玩家，事先多次在游戏中分别扮演计算机和人类相互配合完成游戏，在此过程中记录两名玩家的控制方法和游戏状态建立而成。实验表明，通过增加数据库中的游戏次数，能明显的提高计算机与人类配合完成游戏的效果。当数据库中的样本达到一定数量后，人机合作完成游戏的效果甚至超过了两名人类玩家的配合。实验结果表明，计算机可以在简单的场景下通过数据库的众包式建立而非复杂编程，完成与人类的合作。

关键词：人机合作，数据库，众包，2D 游戏

第一章 立项背景

我们常说的机器人主要可以分为两种， 即工业机器人和服务机器人。在与人们生活息息相关的制造业中，工业机器人已经扮演了不可或缺的角工业机器人的诞生与应用将人类从重复，低附加值的工作中解放出来，使人类可以将更多的时间与精力投入到创造性的工作中。而服务机器人例如软银公司的 pepper 机器人，也越来越多地出现在人们的日常生活中并承担了一部分工作。但是机器人仍然难以胜任一些复杂的，需要配合的工作，这也就意味着在将来人类需要与机器人合作去完成任务，而机器人也需要学会如何安全，高效率地与人类进行合作。

第一节 机器学习技术

机器学习领域在近些年的蓬勃发展为人机合作技术的发展提供了充足的动力。深度学习的发展，使得机器人精确感知世界成为了可能。DeepMind 公司中负责 AlphaGo 项目的研究院 David Silver 提出了“AI = RL + DL”的看法。即人工智能技术的关键在于结合深度学习的表示能力和强化学习的推理能力。毫无疑问，以深度神经网络为代表的强大的深度学习技术为机器人提供了精确感知周遭世界，获取有用信息的能力，例如物体识别分割，特征提取等等。在人与机器人合作执行任务的过程中，深度学习技术的应用可以使机器人快速且精确的感知并理解人类的行为和决策，同时能够精确地提取周遭环境中有用的特征信息。例如在机器人与人类合作收拾房间地过程中，机器人可以识别出来人类正在处理的物体，识别并记录房间内物体位置的变化，进而可以选择去配合人类整理正确的物体。

在拥有强大感知能力的基础上，机器人需要的是推理和决策的能力。机器人在执行任务的过程中需要做出许多个相互关联的决策，并需要根据实际情况在庞大的决策空间中选择当前情况下最优的决策。实际上机器人在与人类合作执行任务的过程可以看成是一条马尔可夫决策链（Markov Chain），即相互影响的一系列条件概率的组合，在每个状态下，下一状态的概率分布仅由当前状态确定，这一过程与人机合作完成任务的过程十分相像。在人类执行了一次动作之后机器人需要理解并记录人类做出的动作，进而做出在当前状态条件下的最优决策来配合人类继续执行任务。而机器人学会如何做出正确的选择来配合人类完成任务的学习过程可以看作为一个强化学习的过程。在机器人学习的过程中，可以人为地设置奖惩机制与评估策略，机器人

在做出正确的抉择时可以得到奖励，反之则受到惩罚；评估策略则会对最终任务的完成情况进行评估并依此调整机器人在合作过程中采取的策略。通过人与机器人不断地进行交互合作，不断地尝试通过不同的方法完成任务并记录这一过程中人类与机器人的决策链和奖惩结果，便可以完成这一类似强化学习地训练过程。

但是机器学习技术在机器人领域的应用仍然存在诸多难点。首先，训练过程十分缓慢。深度学习与强化学习的训练过程依赖于庞大的数据资源和算力资源，而实际情况下我们难以对一些任务进行数学建模并设计合适地奖惩机制，这样可能导致最终的训练结果无法收敛。其次，实验环境的缺乏与限制，要想得到具有广泛适用性的机器人，就必须在复杂多样的环境条件下对机器人进行训练，而这往往是难以实现的。再其次，受限于机器人硬件和控制系统，人类并不能在现实环境中与机器人高效且安全地进行互动合作来获取训练过程中机器人所需要的庞大的训练数据。

第二节 虚拟仿真平台

真实环境中的机器人实验十分不方便并且成本非常高，使得将机器人实验从现实迁移到仿真环境中，然后再将训练的结果从虚拟环境应用到真实环境中成为了一种几乎必然的选择。那么如何设计一个与真实环境十分接近的虚拟仿真环境，并将机器人和人以及两者交互合作的过程迁移到虚拟环境中也就成为了实现人机合作学习的一个关键点。从人机合作的定义不难看出，这一合作过程非常类似于我们在游戏中与虚拟人物进行沟通交互的过程。我们的目标是训练机器人使其懂得如何与人类合作完成任务，那么机器学习的对象其实是对于不同的任务，人和人是如何进行合作的。在现实环境中受限于人力和物力资源，难以获取到数量足够并且复杂度足够的训练数据。我们不难想到，虚拟游戏其实是人与人之间进行互动和合作非常频繁的一个平台。如果我们将各种复杂的任务和奖惩机制制作成一个虚拟游戏，然后让不同的玩家通过合作来完成任务得到奖励，那么来自世界各地的玩家与玩家之间合作完成任务的过程便可以提供足够庞大，复杂度足够高的训练数据。时至今日，诸如 Unity，Unreal 等强大的游戏引擎完全可以完成虚拟环境的搭建，这些引擎中提供的接口也可以使我们能够很方便地对游戏过程进行记录。并且，我们可以很简单地在虚拟环境中设计复杂多样的任务，例如不同的周遭环境，不同的任务要求等等。通过在复杂多样的虚拟环境中训练学习，可以大大提升训练得到的智能体的泛化能力，使其可以快速适应实际应用中的各

种场景。

而在训练过程中使用 ROS（Robot Operating System）中的 Gazebo 软件包可以很方便地对机器人以及机器人的硬件设施例如传感器，机械臂进行虚拟仿真。并可以通过 Unity 等虚拟引擎的接口将仿真的机器人接入虚拟环境中，让其在虚拟环境中尝试完成任务。训练数据便是之前设计的游戏中收集到的庞大的合作过程信息，让机器人在虚拟环境中尝试配合人类采取的决策。

除此之外，日益强大的 VR 技术可以为实验者带来沉浸式的用户体验，使得实验者可以在虚拟环境中如同真实环境那样与机器人进行合作，但却更加地安全和高效。

虚拟环境的应用会大大加速机器人与人合作学习的过程。但是也面临着问题，那就是训练数据的迁移和应用。我们获取到的只有虚拟环境中的实验数据和结果。那么应该如何将这些数据应用到实际的机器人上，也是近段时间机器人领域的研究重点。

第三节 网络众包

网络众包（Crowdsourcing）对许多人来说并不是一个熟悉的名词，它表示将一个大型任务分配给众多成员分别完成。如果说虚拟环境解决了人机交互学习中实验环境的限制，那么网络众包和大数据则可以为这一训练过程注入充足的燃料：数据资源。虚拟游戏有着传统训练环境不可比拟的优势：网络优势。通过类似 Steam，Origin，微信等游戏发布平台和社交平台，我们可以很轻松地将我们制作的包含数据收集功能的虚拟游戏发布到云端。然后成千上万的用户可以下载我们的游戏并进行游玩，在他们游玩的同时，他们在游戏内与其他玩家合作解决各种各样任务时生成的数据和决策链便会上传至云端被我们收集。正如之前提到的，机器学习需要庞大的数据资源与计算资源，通过这样网络众包便可以分撒获取所需要的庞大数据资源。

第四节 相关研究

2018 世界机器人大会上来自日本东北大学的机器人系教授 Kazuhiro Kosuge 在论坛上进行了题为“人机协作的最新演进方向与前景展望”的演讲。演讲中指出近些年关于人机协作的发展十分迅速，前景十分开阔。外骨骼系统是人机协作的方式之一，已经逐渐应用在医疗康复领域，残障人士辅助，以及

军事领域。人类在外骨骼机器人的协助下可以搬运数百千克的物体，获得强大的运动能力。而另外一种人机协作的模式是共同工作机器人，这类机器人在日常条件下的应用更加广发。例如波士顿动力公司开发的机械狗，已经可以自主地配合人类搬运货物。但是实际上大多数的机器人都是基于具体的场景的，为了能更好更高效更安全地使用这些机器人，提高机器人对于多样的场景的适应能力是十分重要的。

第二章 研究的目的和意义

第一节 研究的目的和内容

近些年来，随着技术的不断发展，通过数学建模编码，机器人能实现的任务越来越复杂，但是在某些特定的场合中，数学模型是很难建立的，在这种情形下，单纯的机器程序就很难应对。这种场合在不同的人机合作中格外多，比如在某些依赖于人类习惯的人和机器人的合作项目中，因为每个人的习惯都不太一样，使得机器人很难以单一死板的建模方式完成人机合作任务。这种情况下，如果利用机器学习就能很好地解决这个问题。

机器学习是继专家系统之后人工智能应用又一重要研究领域，也是人工智能和神经计算的核心研究课题之一。在一项人机合作的任务中，我们可以先用人类的角色替代机器人，来进行人与人之间的合作，并在合作中设置几个较为关键的变量，将这些变量记录下来。通过反复大量的实验，将这些人与人之间的合作数据建立数据库，然后将其中的一个人替换为机器人，利用数据库进行机器学习，这样机器人就会学习人类的习惯，以实验中另一个人的习惯为基础进行反馈，来完成人机合作的任务。

在我们的研究中，我们已经建立了一种通过人机交互数据产生机器学习经验的机器学习系统，并且通过建立知识库使得机器人在与人的合作过程中对人类的动作做出反应，进而配合人类的行为完成任务。并且通过实验证明，训练的次数越多，数据库中记录的动作越多，机器人与人类合作完成任务的效果就越好，无论是从完成任务的时间还是从完成任务的效果上来看，都是与训练次数成正比的。

除此之外，因为直接从现实场景之中获取数据较为困难，我们利用 unity 制作了一个游戏系统，通过在游戏中对人机合作任务进行模拟，从游戏后台中获取任务的参数。这样获取人机合作任务的数据就会方便很多，同时游戏的形式也更有意思。我们先由两个人来进行游戏，一个人作为主人物，另一个人作为机器人，反复进行游戏并收集游戏中的数据，利用数据进行机器学习，再进行人与机器人

的合作练习。

第二节 研究的意义

此项研究的意义，在于使复杂场景的人机合作任务变得简单起来，同时让人机合作过程中机器人的反应能够更贴合合作者的习惯，而不是死板地识别和下命令。数据库的存在从最简单的层面上解决了机器人运行的复杂的代码逻辑的问题，使得人机合作变得更加直接明白。

除此之外，在未来的应用场合中，家庭服务机器人是要商业化的，面向每一个普通的家庭，每一个家庭中的家庭习惯都不一样，这样的机器学习方式使得每一个家庭都可以根据自己的习惯定制自己独有的机器人。

第三章 研究的方法过程以及问题

整个项目的研究可以分为两个部分分开介绍，分别是研究的方法过程以及在研究过程中遇到的一些问题。

第一节 研究的方法过程

课题的研究过程大概可以分为四个阶段，分别是准备阶段、编写阶段、实验阶段以及最后的总结阶段。

3.1.1 准备阶段

准备这阶段的主要任务是查阅此次研究所需要的相关资料论文等，学习相关的专业知识，具体到我们的课题中，主要有熟悉 unity 的应用以及简易数据库的编写等。除了学习知识外，为课题中的人机合作设想一个合适而又具有代表性的模型也是十分重要的，并且在这一方面我们经过了较多次的讨论，反复删改确认，是第一部分中最重要的一块。

3.1.2 编写阶段

这一阶段的主要任务是完成游戏的编写和数据库的建立，即项目中主要的动手部分。

3.1.3 实验阶段

这一阶段的主要任务是收集实验的相关数据，从而建立数据库。在游戏程序编写好之后，按我们的实验设想需要收集两个人类合作者一起完成游戏时的数据，从而利用这个数据对机器人进行训练。我们将游戏制作好后，召集了几个志愿者，反复游玩游戏，然后收集数据。

3.1.4 总结阶段

最后一个阶段的主要任务是总结实验结果，验证数据的可行性以及整个实验过程中可能出现的问题，并对实验做出调整，反复进行实验及数据收集的过程，得出最后的结论。

第二节 研究过程中的一些问题

在整个实验设计的过程中，怎样选择合适的场景进行实验是一个很麻烦的事情，因为具体的场景需要符合实验的目的并且要有代表性。最开始的时候我们选择餐具摆放这样的一个场景，同时为了游戏更具有趣味性，打算制作 VR 游戏，借用 Oculus 的 VR 头显来进行实验，但因为一些技术原因以及实验的便携性问题，最后我们放弃了这个方案，采取了用键盘控制的游戏形式，线上直接把游戏发给对方然后进行数据收集。

受到马里奥派对这个游戏的启发，我们重新设计了游戏，将其设计为两个人一起搬运箱子的游戏。这样游戏数据比起之前更容易收集，并且在合作方面也更有代表性。

第四章 基于众包式数据库的学习模式

第一节 人机合作方式探究

人机合作的实现需要在一定的场景下才具有实际意义。虽然现在实体化的人机合作在生活中的例子还不是很多，很大原因是因为机器人技术的尚未成熟，以及基于人机合作安全性的考虑，但是现实生活中有很多人类合作的例子可以借鉴。

4.1.1 人人合作

人类作为一个独立的个体，无论是生活还是工作，或多或少都避不开与他人合作这一场景。合作的模式也是千变万化。我想可以把合作分为两种：强合作，和弱合作。

强合作，我将它定义为：在某一项任务中，通过人与人之间的合作能较大幅度的提升任务完成效果，或者必需由两人或多人才能完成的合作模式。在生活中有很多这样的例子。比如，一场手术的完成不可能只靠一名主刀医生的努力，还需要主刀医生的助手在旁边协助，器械护士在旁边传递手术器械，巡回护士时刻观察生命体征来向主刀医生反馈等。这些合作人员的存在，原因主要有一下几种：1、这些合作者负责的工作是并行的，而不是串行的，也就是说这些工作可能在同一个时间点发生。而主刀医生就算能力再强也不可能同时完成这些工作。这个时候就需要其他人的协助，帮他分担这些工作。2、手术这样一个场景具有很高昂的失误代价，和时间成本。也就是说无论在哪个环节出现差错，后果都是不能接受的。随着时间的流逝，患者的生存概率降低和器械的损耗都代表着，手术必需要有很高的效率和成功率。手术这样一个场景很能代表一类合作场景，即具有很高昂的失败代价的任务场景，由此产生的合作关系便是将此类任务的完成度提升至单人工作不可能触及的程度。再比如：多人体育运动竞技足球、篮球、排球等，需要同一队成员相互配合——无论是在肢体动作上还是语言沟通，甚至是眼神交流，来共同实现一样的目的，即得分。虽然这些运动在比赛规则，运动形式上都不太一样，但是队员们之间的合作模式都是有很多相似的地方。体育竞技这样的场景虽然不像手术具有那么高昂的失败代价，但是通过合作能将任务的成功率大大提升。

弱合作，我将它定义为：在某项任务中，合作的加入能加速任务的实现进程，或者降低单个参与者劳动成本的合作模式。弱合作相比于强合作，在现实生活中更为常见。比如，大扫除这一任务，并不是非要多人一起做才能完成，也并不存在所谓的失败代价。然而合作的加入能更快的将这一任务实现，并且降低单个参与者在大扫除过程中付出的劳动力。有时候，弱合作者的需求并不只是简简单单的完成当前任务，而是伴随着相关的情感需求。比如，在合作的过程中，合作者之间的交流能带给合作者愉快的体验，增进合作者之间的关系等等，不过这个方面已经超出了本研究涉及的领域，暂且不予讨论。

合作所面临的另一个问题是，多人的合作能否达成一加一大于二的合作效果。答案是不一定，有些时候合作得到的是一加一小于二的效果。这取决于多人之间的合作方式，也是本文重点关注的地方之一。只有了解到人与人合作的方式，才能将其应用在人机合作上面。我们拿篮球这项体育竞技举例：我们先抛开个人技术不谈，来看看合作在取得胜利中扮演了什么角色。在篮球比赛中，合作有这样

几项实例：

1、传球：将球权交给队员，往往发生在队员处于比较有利的位置，或者自己处于困境的时候。当然这只是传球发生的比较简单的两种场景。我们来看参与者是如何判断传球的时机的：首先传球者需要对每一位队员所处的情况加以分析；其次传球者需要对自己的情况加以分析。被传球者一次好的传球从长远的角度来看，是增加了这一轮进攻得分的可能性的。

2、跑位：通过跑动得到被传球的时机，或者将防守球员调出内场。跑位者需要得到全场所有人的信息，此处的信息包括所有人的位置，所有人下一秒的意图等。

3、战术配合：以一套固定的流程规定出每一个参与者在合作中的任务。需要每一位参与者提前得知自己的任务内容及发生时间。合作的成功由每一个战术配合的参与者单人任务的完成度决定。

4、语言沟通，手势交流：这一项合作实例可以发生在以上所有实例之前。即用语言或者肢体动作，传达合作的信息。

由以上分析可以总结得到，合作者的决策需要建立在对任务环境的整体感知——包括对于环境中所有物体信息的获取，以及对于合作者下一步动作的猜测。一次好的决策应当是合作双方对于预判结果的高度重合。拿传球举例就是说，当传球者想传球的时候，被传球者刚好想接球。这时，一个成功的合作才算达成。

4.1.2 人机合作

人机合作若要达到人人合作那样的效果，计算机在充当合作参与者做出合作决策的时候就要就需要参考人类合作者的合作方式。由上一节的分析我们知道，在人人合作的过程中，每一个合作动作的决策都需要两方面的信息：1、当前时刻任务场景的信息；2、合作者动作的预判信息。这里我们设计算机在第i时刻做出的决策为S_i，则有:

$$S_i = f(\Omega_i, \omega(p_i, p_{i-1}), q_i) \tag{1}$$

其中Ω_i代表第i时刻任务的场景信息，p_{i-1}代表合作者上一时刻的动作，q_i代表计算机当前时刻的动作，ω代表了从第$i-1$和i时刻到第$i+1$时刻动作之间的映射关系。

当前，在机器学习邻域，我们可以通过复杂的算法来实现对S_i的计算，比如我们可以用强化学习的思路来对计算机当前时刻的决策进行推演。但是，有时候，复杂的编程不一定适合合作任务的场景。比如，我们想要计算机适应个别合作者特别的合作习惯的时候，编程的方法就会很难实现。其次如果合作任务

的场景十分复杂，则编程的复杂度也会随之升高。

如果使用数据库记录示教样本的模式，或许能成为编程的替代方式之一。具体来说，首先，通过人人合作，将合作的信息全部记录到数据库中，形成示教样本，再让计算机通过搜索数据库的形式，推演出S_i。

第五章 实验及分析

第一节 实验设计

5.1.1 游戏环境设计

游戏为二人合作避障小游戏，两名玩家操作两个人物（分别为机器人智能体和人）共同完成游戏。人物分别由 WASD 以及上下左右键控制，完成向不同方向的加速以及减速，整个模型仿真真实的物理环境，同时具有各种障碍物，均具有碰撞体积，整体环境如图 5.1 所示。

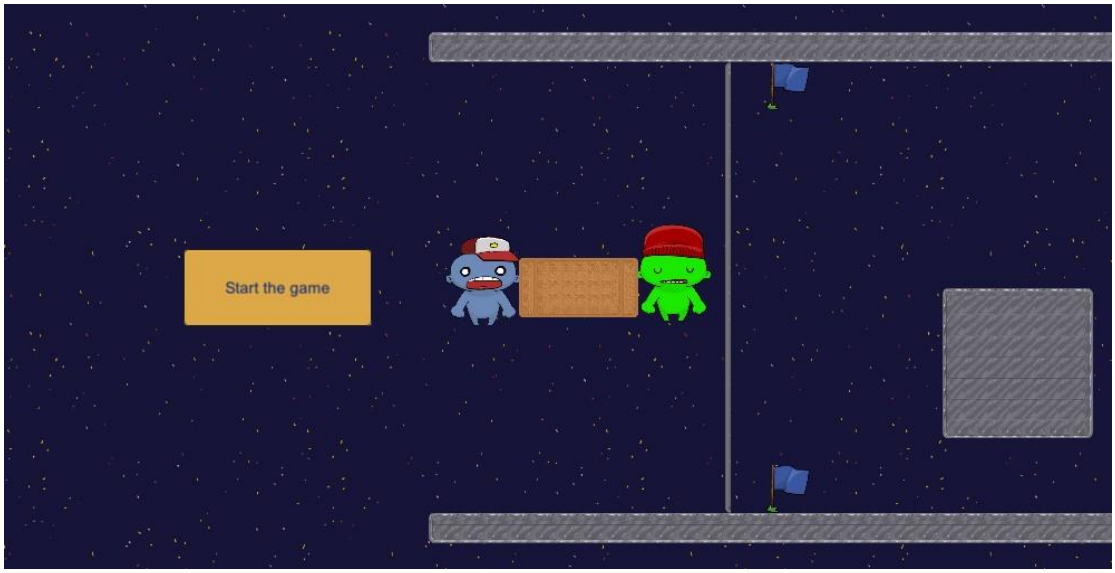

图 5.1 游戏效果图

5.1.2 游戏奖惩设计

为了实现对数据的分析处理，游戏奖惩设置必须明确。不同的目标可以具有不同的奖惩策略。比如，若最终目标为实现最快达到终点，则这一幕每一步的得

分均给时间的倒数，最终到达目标点后给 0 的奖励，在此奖惩机制下，智能体完成任务的首要目标是快速性，一幕终止时间越短，此幕内每个动作的得分越高；若最终目标为实现最少碰撞，则对碰撞动作进行大幅度惩罚，这样每幕结束后，具有碰撞的幕动作的得分就会远少于没有碰撞的幕动作。在本实验中，目标权衡二者，即在快速到达终点的同时降低碰撞概率。因此整体奖惩机制如下：

（1） 对于一幕中没有碰撞的动作，本幕结束后得分增加整幕时间倒数 1/T；

（2） 对于一幕中具有碰撞的动作，本幕结束后得分增加整幕时间倒数的 1/k，其中 k 为碰撞次数，即增加 1/kT。

以上奖惩措施对于碰撞的惩罚力度适中，只进行(k-1)/kT 的惩罚，若需要大幅度惩罚碰撞，可以在此基础上增加系数 α，完成超额惩罚。

5.1.3 数据集获取设计

为了验证数据集对人机合作效果的影响，本实验收集不同大小的数据集，并为智能体训练。其中分别从 25 幕、50 幕、75 幕、100 幕游戏中共获取四组数据集。对于人机合作，如果在不同幕数的数据下具有显著不同的效果，则证明此方法适用于对智能体的训练。比如，若从 25 幕数据到 100 幕数据训练出的智能体与人合作时的速度逐渐变快，且碰撞次数逐渐降低，则说明智能体随着训练数据的增多而接近准确模型，能够更好地完成人机交互合作。

第二节 实验分析

针对本实验总共收集了 4 组数据，分别从 25 幕、50 幕、75 幕、100 幕游戏中获取。在智能体上使用每组数据进行游戏 50 次并取每幕平均耗时以及每幕平均碰撞次数。得到如下表格：

表 1 不同数据下每幕平均耗时及每幕平均碰撞次数对比

	每幕平均耗时(秒)	每幕平均碰撞(次)
25 幕数据	73.1667	31
50 幕数据	56.2687	23

75 幕数据	43.4670	20
100 幕数据	39.4333	14

从表中可以看出，随着数据增加，每幕平均耗时和每幕平均碰撞次数均有显著降低，说明对于状态-动作二元组的得分机制能够有效提高智能体完成任务的能力。通过游戏得到的数据越多，就能获得越多的状态-动作二元组，且其得分越接近准确得分。

在获取原始数据的过程中，我们计算玩家通过游戏的平均每幕耗时为 71.1479 秒，平均每幕碰撞为 25 次。根据表 1，在 25 幕数据的实验中，每幕平均耗时比原始数据的获取耗时长，说明仅获得 25 幕数据并不能很好的对重要的状态-动作二元组进行得分优化；同时由于实验中智能体使用 ε 贪心算法选择动作，在数据不完全的情况下，很多状态-动作二元组的得分为 0，此时随机选取动作的概率大大增加，而我们动作空间有 9 种选择，因此很难做到顺利完成任务，导致耗时和碰撞次数都比原始数据获取时增加。

在 50 幕数据的实验中，每幕平均耗时和平均碰撞次数都均有改善，其中耗时大大降低。说明在 50 幕数据中，状态-动作二元组的得分已经能够满足重要状态的使用，此时智能体可以通过 ε 贪心算法选择较优动作，来加快完成任务。但此时每幕平均碰撞次数仅仅与原始数据获取时的平均碰撞次数相差不大，由以下两点原因造成：

（1） 在此得分机制下，50 幕数据还不足以对碰撞进行有效惩罚，导致可能称为碰撞前提的动作得分较大，在 ε 贪心算法下选择此动作的概率较大。

（2） 由于合作过程中操作员不一定选择到最优动作，因此在合作过程中，智能体选择最有动作的前提下，操作员无法有效配合，最终导致碰撞。

75 幕数据和 100 幕数据时，每幕平均耗时和每幕平均碰撞次数均有显著改善。耗时达到了原始数据获取耗时的一半且碰撞次数也几乎降低至原始数据获取时碰撞次数的一半。此时由于数据足够多，因此我们已经覆盖了充分多的状态-动作二元组。由于我们采取的瓦片编码只有简单的一个覆盖，因此其泛化能力只能覆盖到单个瓦片内的连续小数值。在大于 75 幕数据时，从起点开始的重要路

径及其边缘泛化到的路径均得到了有效地分数赋值，因此在利用 ε 贪心算法选择动作时几乎可以选取到最优动作，再配合操作员的操作，可以达到灵活地合作，能够较快经过拐点以及动态障碍物，最终使耗时大幅度降低。对于碰撞的惩罚在多幕数据下的优势也体现了出来，由于我们采用碰撞来弱化得分，因此在数据较少的情况下，弱化得分不明显，累积得分相差不大，在一定误差下对碰撞惩罚力度不够；而多幕数据下累积的得分弱化就使较差动作与较好动作的得分形成了较大差异。比如在某拐点平均会碰撞 5 次，获取数据时平均一幕需要 71.1479 秒，根据得分机制，最开始的每幕数据能够对此状态-动作二元组带来 0.14 的得分，由于碰撞了 5 次，因此弱化得分后，最终得分增加量为 0.028，与 0.14 差了一个数量级。平均碰撞 5 次意味着总会有小于 5 次的碰撞出现，碰撞次数越少，意味着得分增加量越大，因此在数据较少的前提下，某些较差状态-动作二元组由于统计误差，并未总是得到弱化得分，因此与较好状态-动作二元组得分差异不大；而数据增多后，随着统计误差的消除，较差状态-动作二元组弱化得分次数趋于增加，便与较好状态-动作二元组产生较大差异，最终使动作选择更加准确。

5.2.1 瓦片编码

瓦片编码是一种用于多维连续空间的粗编码，具有灵活并且计算高效的特点。在状态空间中利用一系列划分来编码，每一个划分称为一个覆盖，每个覆盖中的元素被称为一个瓦片。不同覆盖需要具有一定比例的偏移，使其相互交叠。不同状态会激活不同覆盖下的一个瓦片，我们使用被激活的瓦片作为该状态的某个特征，因此总共使用 k 个覆盖，得到瓦片编码的激活维度就是 k。如图 4.1 所示，左边阴影部分为状态空间，右边为利用四个覆盖进行瓦片编码，每个覆盖中有 4×4 个瓦片，其中白点为在状态空间中待表示的点，四个加粗的瓦片代表该点激活的特征。由于使用了 4×4×4 总共 64 个瓦片，因此最终得到的特征向量也有 64 维，其中被激活的维度为 4。

瓦片编码具有较强的泛化能力，如图 4.1 所示，如果更新白点处的状态，4 个加粗的瓦片均会被激活，意味着 4 个瓦片所包含的区域内的所有状态均会被更新。对应不同的位置具有不同的泛化模式，因此整个区域内的状态都能被快速且有效地更新。本文使用的瓦片编码仅有一个覆盖，因此所有的泛化能力均给到单个瓦片之间的关系。

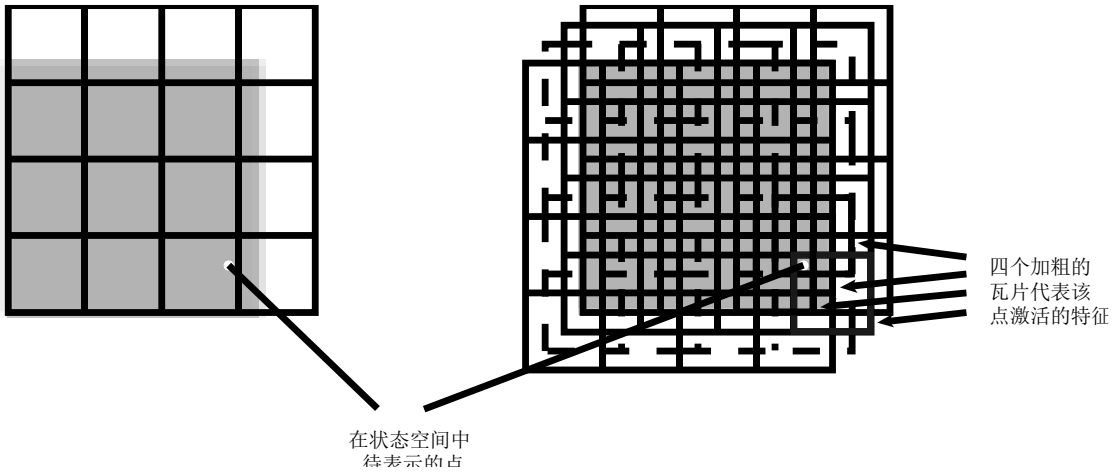

图 5.2 瓦片编码示意图

5.2.2 ε 贪心算法

在智能体对未知环境交互的过程中，总会出现试探与寻优的权衡问题：当智能体过于依赖已有模型进行寻优，就会有部分状态无法取得；相反，当智能体过于依赖试探，最终得到的模型可能并非最优模型。ε 贪心算法即为解决试探-寻优权衡问题的基本算法，具体算法如下：

（1） 以$1-\varepsilon$的概率进行贪心搜索，即在当前模型搜索最优值；

（2） 以ε的概率在动作空间随机选取。

经过计算可以发现，假设当前最优得分为动作 a，则选取到动作 a 的概率为$1-\varepsilon+\varepsilon/N$，其中，N 为动作空间大小。而选取到其他动作的概率为ε/N，由此看出，在ε贪心算法中，模型整体倾向于选择最优动作，但同时保留一定的概率对其他动作进行探索。一般的，取ε值为 0.1 或 0.01。

第六章 总结

从实验中可以得知，当训练次数达到一定数量后，计算机便会从建立的数据库中检索出对人机交互有效的数据，即获得人机交互中的实际指令。因此将 crowdsourcing 应用到实验平台将极大的丰富数据库的有效信息，使得计算机可以更好的与人类完成相应的任务。丰富的数据库也减小无效的训练经验在完成任务期间对计算机获取指令的影响。

第七章 不足与展望

虽然我们的方法可以在 2D 的实验平台上得到验证，但是离最终在三维环境下的实现还很遥远。原因主要在于，三维环境下能够感知的信息剧增，无法通过简单的数据库收集与检索来得到较好的人机交互数据来源。下一步我们将在此工作的基础上继续开发与研究，将理论部分扩展到三维空间。

第八章 参考文献

[1] Benson, S., “Learning Action Models for Reactive Autonomous Agents.” Ph.D. dissertation, Stanford University Computer Science Department.

[2] BioWare. Neverwinter Nights. Atari.

[3] Blizzard Entertainment. World of Warcraft. Vivendi Universal Games Inc.

[4] “Game Designers Test the Limits of Artificial Intelligence.” The Boston Globe, June 17, 2007.

[5] Bruner, J., “Early Social Interaction and Language Acquisition.” In H.R. Schaffer (Ed.),Studies in Mother-Infant Interaction. New York: Academic.

[6] Clark, H. Using Language. Cambridge University Press.

[7] Fikes, R.E. and Nilsson, N.J. (1971). STRIPS: A new approach to the application of theorem proving to problem solving. Artificial Intelligence, 2(3–4), 189–208.

[8] Gil, Y. “Acquiring Domain Knowledge for Planning by Experimentation.” Ph.D. dissertation, School of Computer Science, Carnegie-Mellon University.

[9] Gorin, A., Riccardi, G., and Wright, J., “How may I help you?” Speech Communication, Volume 23. Elsevier Science.

[10] Gorniak, P. and Roy, D., “Probabilistic Grounding of Situated Speech using Plan Recognition and Reference Resolution.” Seventh International Conference on Multimodal Interfaces (ICMI 2005).

[11] Housz, T. I., “The Elephant’s Memory: An Interactive Visual Language.” www.khm.de/~timot/PageElephant.html.

[12] Jurafsky, D. and Martin, J. Speech and Language Processing. Prentice Hall.

[13] Mateas, M., and Stern, A., “Procedural Authorship: A Case Study of the Interactive Drama Façade.” Digital Arts and Culture (DAC 2005).