

硕 士 研 究 生 读 书 报 告



题目 基于轨迹优化和深度强化学习的

篮球运球技能学习分析报告

作者姓名 林炳如

作者学号 21851196

指导教师 李启雷

学科专业 软件工程

所在学院  软件学院

提交日期 2018 年 12 月

Analysis on *Learning Basketball Dribbling Skills Using Optimization and Deep Reinforcement Learning*

A Dissertation Submitted to

Zhejiang University

in partial fulfillment of the requirements for

the degree of

Master of Engineering

Major Subject: Software Engineering

Advisor: Qilei Li

By

Bingru Lin

Zhejiang University, P.R. China

2018

摘要

本文为发表在2018年Siggraph会议上题为*Learning Basketball Dribbling Skills Using Optimization and Deep Reinforcement Learning*【1】的读书报告，从研究背景，研究难点，贡献点，相关工作，实现方法以及实验等方面介绍论文。

在篮球动画制作中，运球的灵活性和精准度是大家所希望优化的目标，动画越逼真效果越好。打篮球的特点是运动员的身体需要表现出高灵活性，因为往往需要在极短时间内对球施力以改变球向，同时预判下一次控球点并马上调整身体朝向和动作，如此循环反复。传统的基于物理学的仿真方法很难满足上述要求。该论文提出了一种通过将动画人物动作分解为身体控制和手部控制，利用轨迹优化和深度强化学习来学习运篮球的动作捕捉数据的模型，达到了较好的效果。

**关键词**：篮球运球，控制分解，轨迹优化，深度强化学习

Abstract

This is a report on a paper published in the 2018 Siggraph entitled Learning Basketball Dribbling Skills Using Optimization and Deep Reinforcement Learning. The report will introduce research background, research difficulties, contribution points, related work, implementation methods and experiments of this paper.

In basketball animation production, the agility and precision of dribbling is the goal that everyone wants to optimize. The more realistic the animation, the better. The characteristic of playing basketball is that the athlete needs to be highly agile, because it is often necessary to apply force to the ball in a very short time to change the direction of the ball. At the same time, the athlete needs to predict the next position of applying force and immediately adjust the orientation and movement of the body, repeatedly. Traditional physics-based simulation methods are difficult to meet the above requirements. This paper proposes a model that learns from the motion capture data of basketball dribbling by decomposing the animation character motion into body control and hand control, and uses trajectory optimization and deep reinforcement learning to learn the model. The algorithm achieves better results than traditional ways.

**Keywords：**basketball dribbling, control decomposition, trajectory optimization, deep reinforcement learning

1总体概述

本文提出了一个基于控制分解、轨迹优化和深度强化学习的篮球运球动画仿真算法并实现了一个流水线系统，在实验中具有较好的效果。

系统将篮球控制分解为身体控制和手部控制两个模块。系统首先学习身体控制模块，然后学习和训练手部控制模块，从而实现对各种篮球动作的高水平控制。其中，对于身体控制模块，论文采用刘等人【2】提出的基于采样的学习方法；对于手部控制模块，论文采用基于轨迹优化和深度强化学习的学习方法。

为了使仿真球员的动作连贯，需要考虑如何衔接两个不同的篮球动作，论文提出通过学习控制图的方法来学习过渡性动作。

**2研究难点**

对篮球运球进行动画仿真的难点在于：

* 使用运动捕捉数据来动画仿真篮球运动的难点在于在运动操作（例如混合和变形）期间球的运动和运动员的运动之间的精确耦合可能容易被破坏。
* 设计这样的控制器的难点在于，球和地面的物理相互作用的细节，执行所需任务所需的灵活性和优雅性，以及运动控制和快速移动篮球的操纵之间的耦合对控制器提出了很高的要求。

当前在篮球动画领域最前沿的技术是NBA 2K系列以及NBA Live系列，它们的缺点是动画中的篮球有时可能会往一个不符合物理规律的轨迹移动，有时则会让玩家感觉球粘在了球员的手上。

**3 贡献点**

该论文的贡献点如下：

* 为了实现对球的稳健性控制，开发了基于轨迹优化和深度强化学习的高效流水线模型，并学习了非线性臂控制策略。
* 提出了一种能同时学习运球技能和技能间过渡的技术。
* 该模型能够为各种篮球运球技术学习强大的控制器，例如腿下运球和交叉运球。
* 使模拟玩家能够在这些技能之间进行转换并响应用户的交互。
* 是第一个在篮球控制领域的实时物理控制模型
* 成功采用了深层强化学习，以便有效学习非线性手臂控制政策，从而在高度动态的篮球技巧中实现对球的强力控制。

**4 相关工作**

过去的几十年来，运动控制一直是基于物理学方法的角色动画的核心话题。具体的运动过程的例子有走路【3】，跑步【4】，保持平衡【5】，体操【6】，杂技表演【7】等等。不同的动作有不同的特点，于是科学家们提出了各种各样的方案和策略。当前最为先进的方法是基于采样的方法：它能从动作捕捉数据学习出适用于各种运动动作的身体动作控制器。然而，论文的作者发现这种基于采样的算法并不能很好地用于学习篮球动作的控制，原因就在于：为了很好地控制篮球和身体，角色需要在极短时间内且在球离开手前精准地控制球并且预判下一次的碰球点在哪，而基于采样的方法没有前向预判的策略，所以很难做到对球的持续性控制。

与运动控制相比，控制能够与运动物体相互作用的仿真人形角色的研究比较少。球类移动，自行车特技，滑板等等均属于该范畴，同样地，人们也提出了一些相应的解决办法。其中，篮球运球也属于该领域，它涵盖了一些在分析过程中需要重视的特点，包括篮球的移动速度快，用于单次控制球的时间短以及角色身体运动和篮球移动的耦合匹配等等。这些特点使得篮球运球的动画仿真是一个棘手的问题。

另一方面，在计算机动画、机器人和生物力学领域，综合手部的抓取和移动物体的动作也有一段比较长的研究历史。主流的方法是采用基于物理学的模型来生成具体的，完全遵循物理规律的手部动作，例如通过PD servos生成真实的抓取动作【8】，通过获取接触力的数据可以将手指与物理的交互抽象转换成其它类型的问题【9】，通过对球动态位置进行细致分析和建模来指导物理机器人与球的交互【10】等等。值得注意的是，这些方法通常关注如何控制手指，手或者手臂，而忽略了对角色的身体的分析考虑【11】。作为对上述问题的改进，论文通过定位角色的手指、手、手臂以及整个身体和手臂的耦合来实现对快速移动物体的控制交互。

接着，在基于物理学的角色仿真中，轨迹优化常常被用来生成人体和虚拟物体的运动轨迹【12】。在论文的控制系统中，对于每一种篮球控制任务（例如原地运球，胯下运球，跑步运球等等），都用轨迹优化来学习一种对应的线性控制策略。接着线性控制器会作为深度强化学习网络的输入，学习出一个更复杂的非线性控制策略。

其实，在角色动画领域，相当一部分的先前工作已经表明强化学习方法能够被成功地应用在交互型动作合成和基于物理学的控制等方向【13】。具体的先前工作的例子有离散化动作空间加深度Q-learning算法【14】，引导型策略搜索算法【15】，actor-critic算法【16】等等。论文中的模型采用了深度确定性策略梯度方法【17】（deep deterministic policy gradient method）来学习篮球控球过程中手部的控制策略。具体来讲，模型首先对每一种运球动作学习一种控制策略，然后训练学习不同类型的运球动作之间的过渡动作。

**5 系统概述**

**5.1 概要描述**

* 系统输入

动作捕捉片段

* 系统输出

强大的允许模拟篮球运动员执行各种篮球技能的控制器模型

* 模型组成

控制器由两部分组成，分别是用于稳定控制篮球的手部控制器和控制角色的关节、移动和平衡的身体控制器，如图1所示。

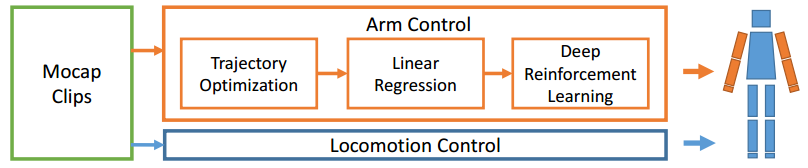


图 1

* 学习策略

1. 首先使用刘等人提出的基于采样的方法来学习身体控制器。在这个阶段，算法不考虑篮球的运动，算法尽可能使角色模仿动作捕捉数据的动作。这种做法的理由在于与人类运动员相比，篮球比较轻，而且球的轨迹变化对手臂的调整很小。
2. 手部控制器的学习从重构篮球的运动轨迹开始。这个过程并不需要真实的篮球轨迹数据。学习算法会求解一个优化问题，用于为每个能够生成最佳匹配动作捕捉片段的篮球轨迹的控制片段计算其手部控制策略。
3. 在实践中发现，2）中训练得到的手臂控制策略可以成功控制简单的篮球技能，如双手携带球，但不能用于控制更复杂的技能，如运球; 因此，学习一种非线性的手部控制策略，该策略用人工神经网络建模，以用于每种篮球技能。

* 控制图

控制图有一系列篮球动作以及动作之间的过渡动作组成。为了保证能学习到有效的过渡动作，控制图是以增量形式学习的。具体地，当有一组新的篮球动作需要学习时，系统会把该新动作连接的已学动作、过渡动作和该新动作作为一个整体进行学习训练。

**5.2 仿真框架**

论文将角色篮球运动员模型化为一个关节型骨骼系统，这当中包括一个浮动的根关节和使用PD-servos的内部关节，如图2所示。

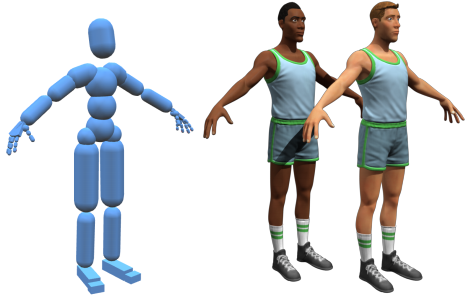


图 2

在每一个时间片，身体控制部件和手臂控制部件计算包含每个内部关节的目标角度的目标姿势。然后通过使用PD-servos跟踪该目标姿势来计算关节扭矩。

一般来讲，对手部的建模是通过对手指的建模来完成的，每个手都有5个自由度并且被设定为独立控制，如图3所示。

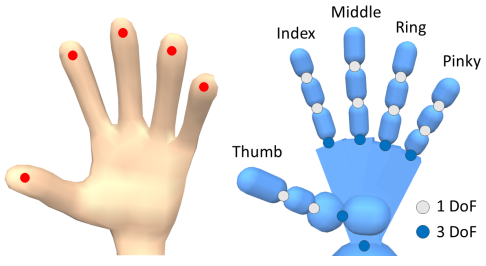


图 3

该论文提出了一种对手部的改进建模方法。因为手指的移动本来是包含冗余的自由度的，为了简化问题，系统会计算每只手的目标姿势，作为手展开姿势和握拳姿势之间的插值，如图4所示。

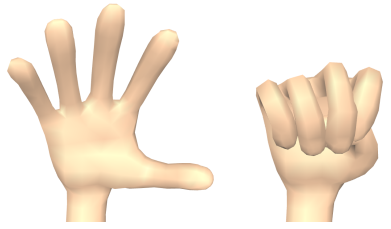


图 4

另外，为了实现协调的手指运动，我们的系统使用绑定拇指，食指和小指的三个控制信号控制每只手; 然后通过它们间接计算中指和无名指的位置。

**5.3 控制序列**

接下来主要介绍控制器的构成。如图5所示，身体控制器和手部控制器都分别由一个反馈策略和一个被分片的开环控制轨迹组成。

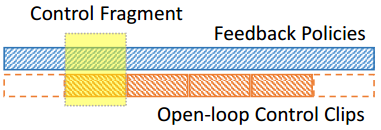


图 5

具体地，一个控制器由一系列控制分片组成，记为{}，每一个分片，它由开环控制片和相应的控制策略组成。

控制策略的定义为状态空间中的状态到决策空间中的决策的映射。状态由仿真系统的当前状态和控制分片的序号组成。

在运行时，我们的系统在每个控制分片的开始处计算反馈策略，并根据仿真状态来计算身体控制器和手臂控制器的补偿偏移。

**5.4 基于采样的动作控制方法回顾**

该论文采用Liu等人提出的基于采样的运动控制方法来学习篮球控制系统中的身体控制器。

当以参考运动作为输入时，基于采样的方法首先将动作捕捉数据进行分片并通过这些分片逐步学习线性控制策略，策略被定义为：



其中是增益矩阵，是对应的仿射项。因此，学习该策略就被转化为学习和的问题。

Liu等人构建了一个长开环控制轨迹，它可以被PD-servos跟踪到并多次重现目标运动过程。同时，轨迹会被分片，每一分片都将被视为状态决策元组并作为线性回归模型的输入数据。

基于采样的算法的核心在于长开环控制轨迹的构建，LIU等人发明了被称为Guided SAMCON的算法来实现上述思想。Guided SAMCON算法通过生成样本决策动作、在应用决策动作时运行仿真以及记录产生最接近参考的运动的样本来为输入序列中的每个控制分片构建开环控制。

**6 身体控制器的学习过程**

该论文的系统遵循与LIU等人相同的学习过程来学习篮球技术的身体控制器。该论文以0.05s为分片间隔。

一旦身体控制器被学习完成后，该模块就会被固定，然后系统开始学习手部控制器。

**7 手部控制器的学习过程**

正如前文所述，Guided SAMCON算法直适用于简单的篮球动作，无法生成允许玩家成功运球的开环手臂控制器。

为了解决这个问题，论文中的系统采用轨迹优化来计算适用于运球任务的开环手臂控制策略。开环控制器以肩部、肘部为目标，手腕也是根据输入的动作捕捉数据来计算的。另外，开环控制分片以这些目标初始化。因为在学习身体控制器的过程中不考虑手指运动，所以中的手指关节的目标角度初始化为0。

然而，因为在上述过程中没有考虑篮球的运动轨迹，所以控制分片序列{}并不能精准地控制篮球运球。

**7.1 轨迹优化**

论文使用轨迹优化来找到一组校正偏移量以应用于初始化的开环手臂控制分片序列，以便使角色可以成功地运球。

具体地，给定动作捕捉切片数据作为输入，论文的做法是首先指定数据中的哪些帧球与角色的手应该接触。这个指定将会作为优化过程的软约束，不必非常精确。然后优化的目标是不断调节手部控制器使得在这些指定帧中球与角色的手距离最小。

其中，一些帧会被指定为篮球动作的检查点，并且将被视为硬约束，篮球和玩家的手指之间的接触将由优化器强制实行。论文选择球在运球动作循环中达到最高点时的帧作为检查点。

模型所需要优化的变量是纠正偏移集合{，}。它们是分别应用到控制分片中的左手和右手的偏移量。其中每个有10个自由度，由3个肩部自由度、1个肘部自由度、3个手腕自由度和3个手指自由度组成。

为了解决学习复杂动作时需要大量参数的问题，论文中的系统将篮球动作技能划分为检查点处的一系列优化窗口，然后每次优化两个连续的窗口。优化问题的目标函数可以被定义为：角色手指与篮球的平均距离与过度修正的惩罚项以及引导训练向可行解进行的辅助成本项。

即使使用滑动窗口方案，优化问题仍然难以解决，因为优化过程可能会容易地在较差的局部最小值处停止，在这个数值点上角色在运球之后就无法再次接球了。 当篮球运动的仿真过程不满足某些要求时，辅助成本项是一个很大的损失值，使目标函数取值变大，当优化过程收敛时，辅助成本项变为零，使目标函数取值变小。为了解决这个优化问题，论文采用了协方差矩阵适应性进化策略（the Covariance Matrix Adaption Evolution Strategy）。

另外，在程序初始化时，初始状态设为输入数据的值，同时篮球的位置设置为使得它跟手掌和手指都接触。

**7.2 线性控制策略的学习过程**

手臂控制策略以状态向量作为输入并计算出一个包含应用于左右手的纠正偏移量的决策向量。

如前文所述，状态向量由仿真过程的当前状态，控制序列分片的序号组成，其中是一个165自由度的向量，代表着角色和篮球的状态。所有这些数量都在一个身体局部参考坐标系中计算，该坐标系与角色的根关节水平移动，并且一个轴指向角色的面向方向。

论文使用线性回归来学习逐步线性控制策略；此线性策略将被用来初始化深度强化学习过程。

在学习过程中，当控制分片已经被学习次并且正被从待优化序列中移除时，文中通过如下公式更新相应的：



**7.3 深度强化学习**

论文将控制问题建模为马尔可夫决策过程（MDP），并使用深度强化学习训练非线性手臂控制器，以实现对篮球动作的鲁棒控制。强化学习的目标是找到最优的策略，使得任何起始状态的预期回报最大化。 该目标可以用目标表示。

其中，深度确定性策略梯度（The deep deterministic policy gradient）算法[Lillicrap et al2015]是一种基于确定性策略梯度的actor-critic方法[Silver et al2014]：策略称为actor; 动作值函数，称为critic。动作值函数计算在状态采用动作后期望的回报值。

论文的系统利用两个中等大小的神经网络来逼近actor和critic函数。 critic网络有两个完全连接的隐藏层，分别包含400和600个ReLU单元，如图6所示。critic函数的动作分量直到第二隐藏层才被包括在网络中。

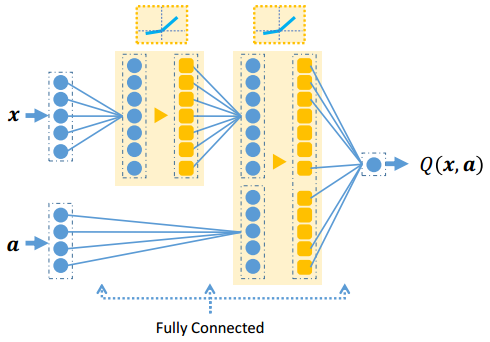


图 6

actor网络具有一个完全连接的隐藏层，其包含500个ReLU单元，如图7所示。

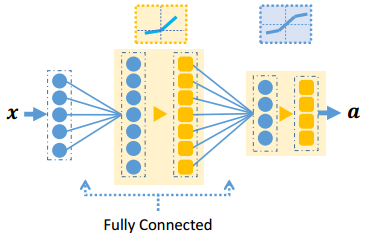


图 7

总的来讲，强化学习网络的设置如下：

* 目标损失函数定义为网络输出与实际值的L2范数和正则化项的和
* 学习算法采用Q-learning
* 参数更新策略是批量随机梯度下降，每批次含有32个数据样本
* 参数更新算法采用Adam算法
* 采用移动平均法来解决不收敛问题
* 在学习过程中，每个动作都会被叠加服从正态分布的随机噪声
* 当critic值增加时，critic网络的学习率会被减少。原因在于该方法能独立于收敛速度，能很好地泛化到不同难度的动作

**8 控制图的学习过程**

控制图是一种图数据结构，其中的结点是控制分片。控制图允许角色在不同篮球动作之间切换。

控制图可以通过指定动作之间的过渡来进行创建。过渡动作仅发生在不同帧间角色状态相似的时候。

控制图的学习过程为：

1. 使用上面讨论的方法优化用于循环型的动作技能的开环控制分片，并学习逐步线性策略；
2. 非循环型动作技能与循环型动作技能相结合，将其作为一个整体进行优化; 论文的系统在训练组合动作时仍然优化循环技能，但是相应的学习到的循环型动作技能的线性反馈策略不会更新；由于线性策略被用作指导，循环型动作技能的再训练会加速收敛，从而降低了组合型的计算成本；
3. 在深层强化学习阶段，论文的系统首先使用DDPG算法来训练循环型动作技能的非线性控制策略，然后使用组合型技能来训练非循环型技能

**9 实验**

论文通过让角色学习一套篮球动作技巧来测试论文提出的方法。

**9.1 实验配置**

角色的设置是1.8m高，体重76kg，包括101个自由度。篮球的半径的大小为11.93cm，重为623.7g。

系统基于c++实现。仿真时间间隔为0.01s，程序运行在配备Intel Core i5-6500 @ 3.2 GHz CPU的机器上。

**9.2 循环型动作技能**

论文通过对一组循环型动作技能来评估所提方法的能力。

对于每个目标动作，算法优化运动序列，在这个过程中角色重复执行动作技能100轮。CMA-ES的输入样本是独立的，这允许算法通过在不同工作线程中计算每个样本来并行优化过程。然后，系统使用线性回归从这些优化的运动序列中学习逐步线性臂控制策略。然后，系统使用图8中算法所描述的DDPG算法学习技能的非线性手臂控制策略。

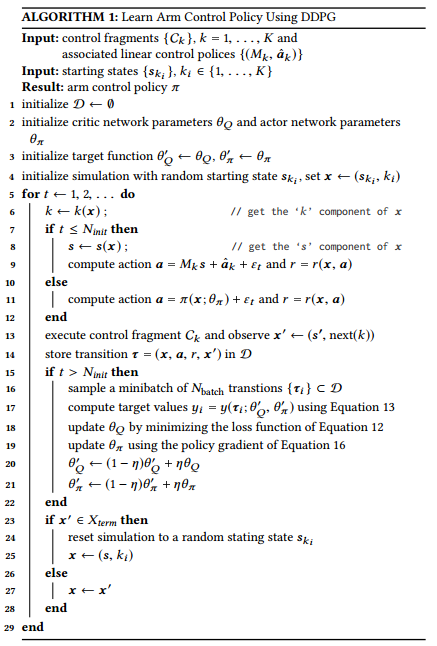


图 8

在学习过程中，论文的系统监视最后10000个过渡动作元组中的结束状态的数量，这提供了对控制策略稳健性的度量。

值得注意的是，经过训练出来的角色能够在承受一定程度外力的情况下保持篮球动作。

**9.3 控制图**

在具体实验中，算法构造两组控制图，如图9所示。每组控制图包含一个循环型动作技能和两个非循环型动作技能，这使得角色能执行两套集成的篮球技巧。

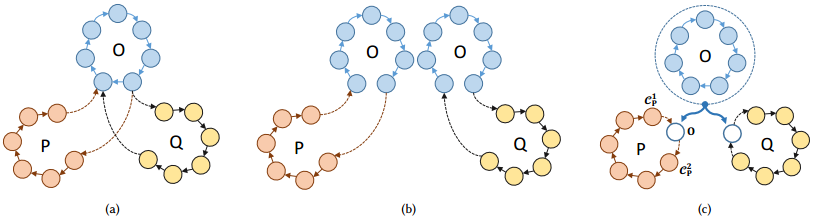


图 9

**9.4 结果**

图10表明实验中对5中运球动作的参数测量比较。图11表明有无线性策略作为初始化数据情况下模型的学习情况。

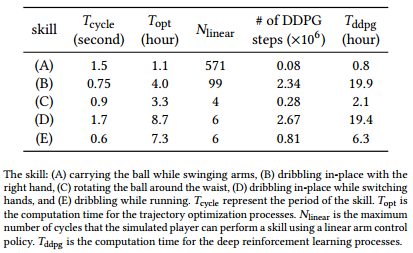


图 10

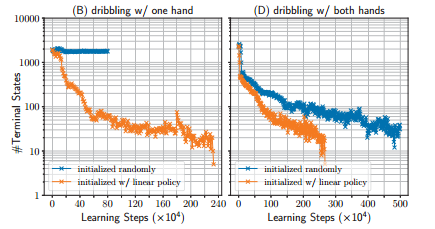


图 11

**10 总结**

这是第一次使用基于物理的方法实时合成各种运球技巧的尝试。

虽然论文的框架是专为篮球技术而设计的，但作者相信它可以扩展到其他角色和被操纵对象之间的相互作用不会显著影响角色的平衡的动作，例如杂耍。

在未来的工作中，作者有兴趣调查其他运动，如足球，足球的特点是平衡控制与运动机动紧密结合。

论文选择CMA-ES作为轨迹优化方法，是因为它易于实现，并且独立于物理引擎的选择。该方法可以扩展到使用更有效的轨迹优化方法。

监督学习可以替代文中学习控制策略的方法; 在学习复杂的非线性控制策略时，监督学习通常需要大量数据来防止过拟合，这是采用监督学习方式应该注意的地方。

直接优化控制策略是使用强化学习的另一种选择; 当学习具有大量参数的复杂控制策略时，优化问题可能难以解决并且易于收敛到较差的局部最小值。

因为学习篮球控制器的身体控制器不允许在运行时改变分片步长，这个限制使论文的框架无法学习控制器以获得需要精确控制步骤的技能，例如上篮和扣篮。

**11参考文献**

[1] Liu L, Hodgins J. Learning basketball dribbling skills using trajectory optimization and deep reinforcement learning[J]. ACM Transactions on Graphics (TOG), 2018, 37(4): 142．

[2] Libin Liu, Michiel van de Panne, and KangKang Yin. 2016. Guided Learning of Control Graphs for Physics-Based Characters. ACM Trans. Graph. 35, 3, Article 29 (May 2016), 14 pages.

[3] Stelian Coros, Philippe Beaudoin, and Michiel van de Panne. 2010. Generalized Biped Walking Control. ACM Trans. Graph. 29, 4, Article 130 (July 2010), 9 pages.

[4] Kai Ding, Libin Liu, Michiel van de Panne, and KangKang Yin. 2015. Learning Reducedorder Feedback Policies for Motion Skills. In Proceedings of the 14th ACM SIGGRAPH/ Eurographics Symposium on Computer Animation (SCA ’15). ACM, 83–92.

[5] Jie Tan, Yuting Gu, C. Karen Liu, and Greg Turk. 2014. Learning Bicycle Stunts. ACM Trans. Graph. 33, 4, Article 50 (July 2014), 12 pages.

[6] Jie Tan, Yuting Gu, C. Karen Liu, and Greg Turk. 2014. Learning Bicycle Stunts. ACM Trans. Graph. 33, 4, Article 50 (July 2014), 12 pages.

[7] Mazen Al Borno, Martin de Lasa, and Aaron Hertzmann. 2013. Trajectory Optimization for Full-Body Movements with Complex Contacts. IEEE Trans. Visual. Comput. Graph. 19, 8 (Aug 2013), 1405–1414.

[8] Nancy S. Pollard and Victor Brian Zordan. 2005. Physically Based Grasping Control from Example. In Proceedings of the 2005 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA ’05). 311–318.

[9] Paul G. Kry and Dinesh K. Pai. 2006. Interaction Capture and Synthesis. ACM Trans. Graph. 25, 3 (July 2006), 872–880.

[10] Sheldon Andrews and Paul G. Kry. 2013. Goal directed multi-finger manipulation: Control policies and analysis. Comput. Graph. 37, 7 (2013), 830 – 839.

[11] Yunfei Bai, Kristin Siu, and C. Karen Liu. 2012. Synthesis of Concurrent Object Manipulation Tasks. ACM Trans. Graph. 31, 6, Article 156 (Nov. 2012), 9 pages.

[12] Mazen Al Borno, Martin de Lasa, and Aaron Hertzmann. 2013. Trajectory Optimization for Full-Body Movements with Complex Contacts. IEEE Trans. Visual. Comput. Graph. 19, 8 (Aug 2013), 1405–1414.

[13] Yongjoon Lee, Kevin Wampler, Gilbert Bernstein, Jovan Popović, and Zoran Popović. 2010b. Motion Fields for Interactive Character Locomotion. ACM Trans. Graph. 29, 6, Article 138 (Dec. 2010), 8 pages.

[14] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015b. Human-level control through deep reinforcement learning. Nature 518, 7540 (26 Feb 2015), 529–533. Letter.

[15] Sergey Levine and Vladlen Koltun. 2013. Guided Policy Search. In Proceedings of the 30th International Conference on Machine Learning, Vol. 28(3). 1–9.

[16] Xue Bin Peng, Glen Berseth, Kangkang Yin, and Michiel Van De Panne. 2017. DeepLoco: Dynamic Locomotion Skills Using Hierarchical Deep Reinforcement Learning. ACM Trans. Graph. 36, 4, Article 41 (July 2017), 13 pages.

[17] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. CoRR abs/1509.02971 (2015).