Learning Graspability of Unknown Objects via Intrinsic Motivation

Ercin Temel[1], Beata J. Grzyb[2], and Sanem Sariel[1]

[1] Artificial Intelligence and Robotics Laboratory
Computer Engineering Department,
Istanbul Technical University, Istanbul, Turkey
{ercintemel,sariel}@itu.edu.tr
[2] Centre for Robotics and Neural Systems,
Plymouth University, Plymouth, United Kingdom
beata.grzyb@plymouth.ac.uk

Abstract. Interacting with unknown objects, and learning and producing effective grasping procedures in particular, are challenging problems for robots. This paper proposes an intrinsically motivated reinforcement learning mechanism for learning to grasp uknown objects. The mechanism uses frustration to determine when grasping of an object is not possible. The critical threshold of frustration is dynamically regulated by impulsiveness of the robot. Here, the artificial emotions regulate the learning rate according to the current task and performance of the robot. The proposed mechanism is tested in a real world scenario where the robot, using the grasp pairs generated in simulation, has to learn which objects are graspable. The results shows that the robot equipped with frustration and impulsiveness learns faster than the robot with standard action selection strategies providing some evidence that the use of artificial emotions can improve the learning time.

Keywords: Reinforcement Learning, Intrinsic motivation, Grasping unknown objects, Frustration, Impulsiveness, Visual scene representation, Vision-based grasping

1 Introduction

Robots need effective grasp procedures to interact with and manipulate unknown objects. In unstructured environments, challenges arise mainly due to uncertainties in sensing and control, and lack of prior knowledge and model of objects. Effective learning methods are essential to deal with these challenges. One classic approach here is to use reinforcement learning (RL) where an agent actively interacts with an environment and learns from the consequences of its actions, rather than from being explicitly taught. An agent selects its actions on basis of its past experiences (exploitation) and also by new choices (exploration). The goal of an agent is to maximize the global reward, therefore the agent needs to rely on actions that led to high rewards in the past. However, if the agent is

too greedy and neglects exploration, it might never find the optimal strategy for the task. Hence, to find the best ways to perform an action they need to find a balance between exploitation of current knowledge and exploration to discover new knowledge that might lead to better performance in the future.

We propose a competence-based approach to reinforcement learning where exploration and exploitation is balanced while learning to grasp novel objects. In our approach, the dynamics of balancing between exploration and exploitation is tightly related to the level of frustration. The failures in obtaining a new goal may significantly increase the robot's level of frustration, and push it into searching new solutions in order to achieve its goal. However, a prolonged state of frustration, when no solution can been found, will lead to a state of learned helplessness, and the goal will be marked as unachievable at the current state (i.e., object not graspable). Simply speaking, an optimal level of frustration favours more explorative behaviour, whereas low or high level of frustration favours more exploitative behaviour. Additionally, we dynamically change the robot's impulsiveness that influences how fast the robot gets frustrated, and indirectly how much time it devotes to learning a particular task.

To demonstrate the advantages of our approach, we compare it with three other action selection methods: ε-greedy algorithm, softmax function with constant temperature parameter, softmax function with variable temperature depending on agent's overall frustration level. The results shows that the robot equipped with frustration and impulsiveness learns faster than the robot with standard action selection strategies providing some evidence that the use of artificial emotions can improve the learning time.

The rest of the paper is organized as follows. We first present related work in the area. Then, we give the details of the learning system including visual processing of objects, the RL framework and the proposed action selection strategies. In the next section, we present the experimental results and then conclude the paper.

2 Related Work

Our main focus is on learning graspability of objects. Previously, analytical methods are proposed for grasping objects [3], [8], [4]. These methods use contact point locations on objects and the gripper, and then find the friction coefficients by tactile sensors to compute force [15]. With these data, grasp stability values or promising grasp positions can be determined. Another approach for grasping is learning by exploration. In a recent work [6], grasp successes are associated with 3D object models which can lead algorithms to memorize object grasp coordination. According to their work, grasping unknown objects is a challenging problem and it varies in accordance with system complexity. This complexity depends on the chosen sensors, prior knowledge about environment and scene configuration. In [11], 2D contours are used for approximating the center of mass of objects for grasping.

In our work, we use reinforcement learning (RL) framework for learning and incorporate competence-based intrinsic motivation for guidance in search. The complexity of reinforcement learning is high in terms of the number of state-action pairs and the computations needed to determine utility values [14]. Approximate policy iteration methods can be used to alleviate this problem based on sampling [7]. Imitation learning before reinforcement learning [12] is one of the methods for decreasing the complexity in RL [5]. Furthermore, it is also used for robots learn crucial parameters in movement to accomplish the task.

In our work, we use a competence-based approach for intrinsic motivation for balancing exploration in RL. Frustration level of the robot is taken into account. We further extend this approach by adopting an adaptive frustration level depending on a task. Intrinsic motivation is investigated in earlier works. Lenat [13] propose a system considering "interestingness" and Schmidhuber introduce curiosity concept for reinforcement learning [19]. Uchibe and Doya [22] also consider intrinsic motivation as learning objective. Different from curiosity and reward functions, Wong [24] point out that ideal level of frustration is beneficial for exploration and faster learning. In addition, Baranes and Oudeyer [1] propose competence-based intrinsic motivation for learning. In our work, main difference is that impulsiveness [20] is adapted into the frustration rate in order to change the learning rate dynamically based on a task in real world environment for robots.

3 Learning to Grasp Unknown Objects

We propose an intrinsically motivated reinforcement learning system for robots to learn graspability of unknown objects. The system includes two main phases for determination of grasp points on objects and experimentation of them in the real world (Fig. 1). The first phase includes the required methods to determine candidate grasp point pairs in simulation. Note that a robot arm with a two-fingered end effector is selected as the target platform. For this reason, grasp points are determined as point pairs. In the second phase of the system, the grasp points determined in the first phase are experimented in the real world through reinforcement learning. The following subsections explain the details of these processes.

3.1 Visual Representation of Objects

In our system, objects are detected in the scene by using an ASUS Xtion Pro Live RGB-D camera mounted on a linear platform for interpreting the scene for tabletop manipulation scenarios by a robotic arm. We use a scene interpretation system that can both recognize known objects and detect unknown objects in the scene [9]. For unknown object detection, Organized Point Cloud Segmentation with Connected Components algorithm [21] from PCL [16] is used. This algorithm finds and marks connected pixels coming from the RGB-D camera

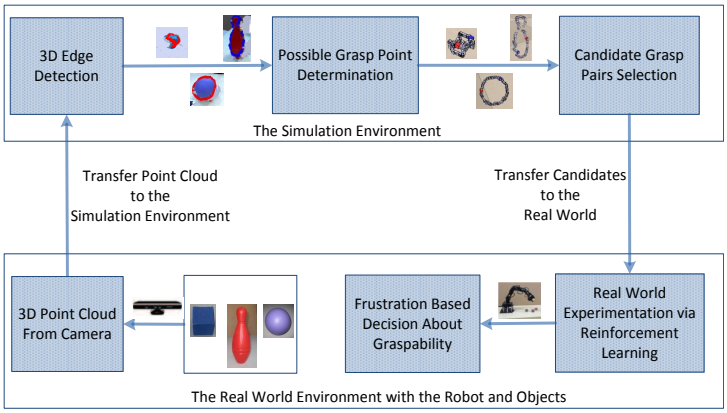


Fig. 1. Overview of the intrinsically motivated reinforcement learning system.

and finds the outlier 3D edges by RANdom SAmple Consensus (RANSAC) algorithm [17]. Hence, the object's center of mass and its edges are detected to be used by the grasp point detection algorithm that finds candidate grasp point pairs for a two-fingered robotic hand.

3.2 Detection of Candidate Grasp Points in the Simulator

Objects are represented by their center of masses (μ) and 3D edges (H). Then candidate grasp point pairs ($\rho = [p_1, p_2]$) are determined as in Algorithm 1. In the algorithm, initially the reference points are determined. The center of mass, the upside and the bottom side center points are chosen as references. Based on these points, cross section points coplanar with the reference points and parallel to the table surface are determined. In the next step, the algorithm detects the closest point to the reference points on the same planar and draw a line crossing with reference points and closest to it. The second step is determining the opposite point to the closest one on the same line. This procedure continues until all points are tested. The algorithm produces the candidate grasp pairs (two grasp points with x,y,z values) and orientation of each pair according to (0,0) point in 2D (x,y) plane. These grasp points are tested in the simulator for finding out only the feasible ones.

In Fig. 2, the edges and sample grasp points for six different objects along with the number of grasp points are presented.

Algorithm 1 Grasp Point Detection (μ, H)

Input: Object Center Of Mass μ, Edge Point Cloud H
Output: Grasp Pairs P
Detect $maxZ$, $minZ$ and C as reference point ref.
for each reference point **do**
  $cPoints$ = `findPointsOnTheSamePlane()`
  $mPoint$ = `findClosestPointToReferencePoint(`$cPoints$`)`
  $slope$ =`findSlope(`$mPoint$`,`ref`)`
  for each $p \in cPoints$ **do**
    $Pslope$ =`findSlope(`$mPoint$`,`p`)`
    if `onTheSameLine(`$Pslope$`,`$slope$) **then**
      $P \leftarrow \{ p, mPoint \}$
    end if
  end for
end for

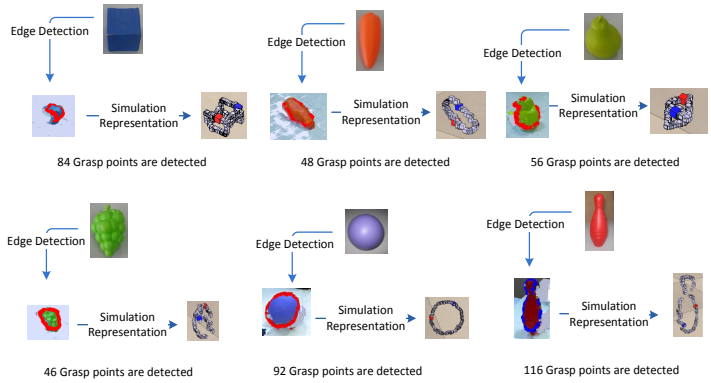


Fig. 2. Candidate grasp points on unknown objects are determined through a sequence of processes. Samples points for six objects are illustrated. The first step is 3D edge detection from 3D point cloud data. The second step is determination of candidate grasp point pairs for which samples are marked with red points on the 3D edges extracted from point clouds of objects. The number of feasible grasp points for each object is presented.

3.3 Learning When to Give Up Grasping

In the system, the output of the simulation environment is fed to the robotic arm to apply real-world experimentation. Intrinsic motivation with frustration level and new proposed impulsiveness method are evaluated to increase the learning process speed for the robot in order to give up quickly for the objects that are not graspable.

The main task of the robot is to learn which objects are graspable. We use a Reinforcement Learning (RL) framework with Q-learning [23] algorithm and softmax action selection [2] strategy. The state space here are all grasp point pairs generated during the simulation phase. A general state S is defined as:

$$S = [\mu, \rho, \phi, \omega, \boldsymbol{O_v}] \tag{1}$$

where, μ is the center of mass of the object, ρ is the selected set of two grasp points $\rho = [p_1, p_2]$, ϕ is the grasp orientation, ω is the approach direction of the gripper and $\boldsymbol{O_v}$ is the 3D translation vector for object during grasp trial. A collision between the robotic arm and the object may occur when there is a trajectory error that results in a non-zero vector.

Actions can be represented as follows,

$$A = [||\boldsymbol{R_v}||, \omega] \tag{2}$$

where, $||\boldsymbol{R_v}||$ is the slide amount on the x axis and ω represents the approach vector to the object of interest.

In our framework, the robot receives the reward value of 10 (R_{max}) when the grasp is successful and 0.1 (R_{min}) when the grasp is unsuccessful [10]. The Q-values are updated according to Eq. 3.

$$Q'(s,a) = Q(s,a) + \alpha * [R + (\gamma * maxQ(s',a)) - Q(s,a)] \tag{3}$$

where, $Q'(s,a)$ the next Q-value for state action pair (s,a), $Q'(s,a)$ is the current Q-value, α is the learning rate, R is the immediate reward after performing an action a in state s, γ is the discount factor, $maxQ(s',a)$ is the maximum estimate of optimal future value.

We investigate four action selection strategies. The first (and the simplest) one is the ε-greedy action selection method (M_1). This method most of the time selects the action with the highest estimated action value, but once in a while (with a small probability ε), selects an action at random, uniformly, independently of the action-value estimates.

The second one is the SoftMax Action selection (Eq. 4) method (M_2)with constant temperature value [2]:

$$P(a)_t = \frac{e^{Q_t(a)/\tau}}{\sum_{b=1}^{n} e^{Q_t(b)/\tau}} \tag{4}$$

where, $P(a)_t$ is the probability of selecting an action a at the time step t, $Q_t(a)$ is the value function for an action a, and τ is the positive parameter called the temperature that controls the stochasticity of a decision. A high value of the temperature will cause the actions to be almost equiprobable and a low value will cause a greater difference in selection probability for actions that differ in their value estimates.

The third strategy (M_3) also uses the Softmax action selection rule. In this approach, however, the τ parameter is flexible and changes dynamically in relation to the robot's level of frustration and sense of control [10]. An optimal

level of frustration favours more explorative behavior, whereas low or high level of frustration leads to a more exploitative behavior. For the purpose of our simulations, frustration was represented as a simple leaky integrator:

$$\frac{df}{dt} = -L * f + A_0 \tag{5}$$

where, f is the current level of frustration, A_0 is the outcome of the action ($success$ or $failure$) and L is the fixed rate of the 'Leak'.

In Eq. 5 the 'leak' rate (L) was fixed and kept at value 1 for all simulations [10]. Higher values of L cause the frustration rate to increase slower compared to smaller values of L. That means that the robot with a high value of L spends more time on exploration and possibly learns faster. Hence, we propose the forth method (M_4) that builds on this method and changes the value of L dynamically using an expected utilization motivation formula [20]:

$$L = \frac{expectancy * value}{Z + \Gamma(T - t)} \tag{6}$$

where *expectancy* represents the probability of getting the highest estimated action value (as in the greedy action selection method), *value* refers to the expected action reward (here $value = R_{max}$), Z is a constant derived from when rewards are immediate, Γ indicates agent's sensitivity to delay (impulsiveness) and $(T - t)$ refers to the delay of the reward in terms of "time reward" minus "time now".

The impulsiveness is main focus of ours to develop interaction with frustration rate competence based motivation. According to triad "Frustration - Impulse - Temper", a person who has high impulsiveness is considered as "short tempered" and it means quickly get frustrated so that changes on frustration level for learning behavior. Our proposal with that, different values on impulsiveness directly affect rate of leak, L, on frustration formula so frustration rate of agent also will be dependent on impulsiveness.

The robot apart from learning how to grasp an object, also needs to learn whether the target object is graspable or not. The learning of a selected grasp pair ρ and action a finishes when overall frustration level becomes equal or greater than a certain threshold value. This value is determined based on a tolerance formula:

$$Tolerance = e^{-||\boldsymbol{O_v}|| * \varphi} \tag{7}$$

where, $||\boldsymbol{O_v}||$ denotes the translation of the object on the table because of the collision with the end effector and φ the number of trials from the beginning of learning.

Additionally, the online learning process may also end when the following criterium has been met:

$$FrustrationLimit = e^{1/\sqrt{n}} \tag{8}$$

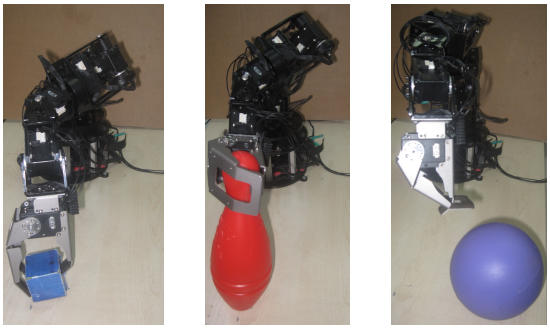
where, n refers to the number of grasp pairs.

3.4 Impulsiveness and Learning Rate

The main focus of the presented work is investigating an effect that impulsiveness has on frustration level and on learning. The learning rate and the speed of decision making is an important issue in human-robot interaction [18]. For example, when a robot plays a quick game with a human, it has to learn quickly. However, when the robot is alone, it can spend relatively more time on exploring different states. By changing the impulsiveness, the robot may dynamically control its level of frustration and therefore the time devoted for learning a particular task. Hence, the robot could behave differently in different environments and for different tasks.

4 Experimental Results

As mentioned before, the candidate grasping points are first determined in simulation, and then transferred to a robotic arm for real-word experimentation. V-REP simulator is used as the simulator and the Cyton-Veta Robotic 7-DOF robot arm by Robai (shown in Fig 3) is used as the experimental platform. The reachability of the arm is about 45 cm. Also in the experiments, we used three objects of different size and shape (i.e., a small cubic plastic block, a plastic bowling pin and a spherical plastic ball). We compare the performance of four



(a) Success of lengthwise grasp on the block.

(b) Success of transverse grasp on the pin.

(c) Failure of lengthwise grasp on the ball.

Fig. 3. Illustrative examples for grasping three different objects. (a)A cubic block which is relatively easy to grasp (b) a plastic bowling pin which can be grasped from top but not for all prasp points (c) a plastic ball which cannot be grasped as it is solid and too large.

different action selection methods discussed in the previous section. A high value

of impulsiveness results in a faster increase in a frustration level (in other words, in a "short-tempered" agent). For comparison reasons, we use here two different values of impulsiveness: a low value of 0.01 and a high value of 100. The results of our experiments support our proposed hypothesis. An agent with low impulsiveness spends more time on exploration, testing more grasp pair possibilities than an agent with a higher value of impulsiveness. For demonstration purposes, we chose three different objects that vary in their graspability properties: a cube that is relatively easy to grasp, a plastic bowling pin that is easily graspable but it is liable of toppling down, and finally, a ball that is not graspable at all. We compare the decision and learning rate of the robot that uses our proposed strategy (M_4) with the one based only on frustration (M_3). Fig. 4 shows robot's level of frustration for each learning epoch while the robot was learning how to grasp the block. The 84 possible grasp pairs generated in simulation were used in a real world scenario. Since the robot can easily grasp the cube, the frustration level is kept low and the learning process terminates before it reaches its limit value, 1.115 (i.e., according to Eq. 8.). In case of the pin (see Fig. 5), the simu-

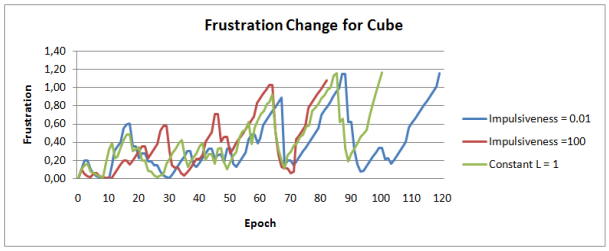


Fig. 4. Frustration Rate Changes For Block Grasping with Methods M_3 and M_4.

lation generated 116 possible grasp pair candidates that were subsequently used by the robotic arm. Since the pin is quite light, the arm pulls it down for some grasp pairs. When the pin fells down, the frustration threshold is decreased for the related grasp pairs according to the Eq. 7. Hence, the robot learns that these grasp pairs should be eliminated from the set and immediately proceeds to test another grasp pair. While for some grasp pairs grasping of the pin was possible, the robot was not able to grasp the ball for any of grasp pairs. The ball was made of a hard plastic material and quite light, so every robot's attempt to grasp it resulted in a ball rolling over on the scene Fig. 3(c). After each trial, the robot's tolerance for frustration decreased rapidly resulting in that the robot switches to another grasp pair. With each failure, the overall frustration level was raising and quickly exceeded the tolerance threshold (that at the same time was being decreased). Although 92 grasp pairs were transferred to the real world scenario, only after a few steps the robot learned that the object is not graspable. Fig.

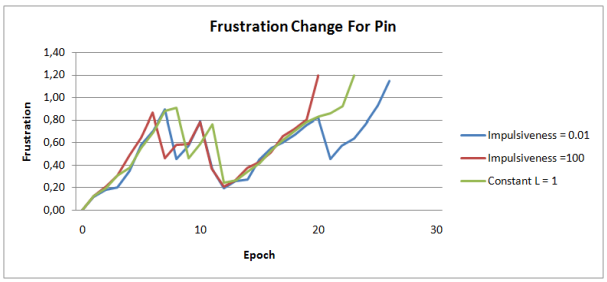


Fig. 5. Frustration Rate Changes For Pin Grasping with Methods M_3 and M_4.

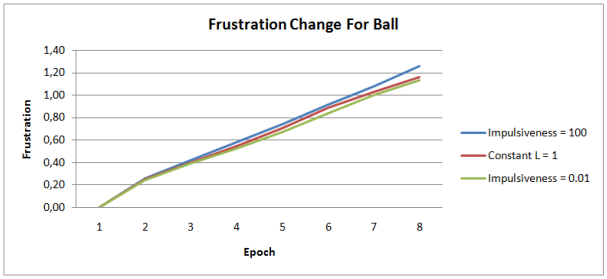


Fig. 6. Frustration Rate Changes For Ball Grasping with Methods M_3 and M_4.

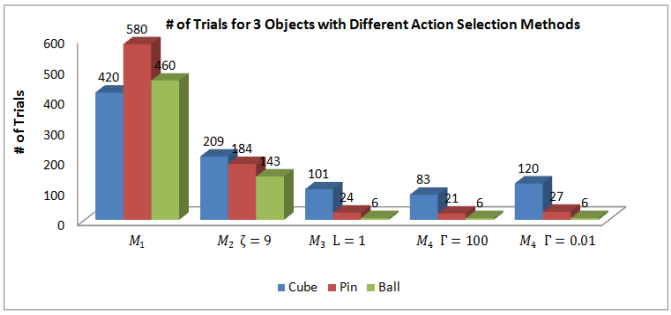


Fig. 7. Trial Count for Three Selected Object and Action Selection Method.

7 shows the comparison of the results for all four strategies of action selection. The frustration-based action selection methods require a lower number of trials

to learn the graspability of the objects compared to the standard softmax action selection with fixed temperature parameter and ε-greedy action selection. The agent with higher value of impulsiveness performs slightly better than the agent with low value.

5 Conclusion

We have presented our intrinsically motivated reinforcement learning system for learning graspability of novel objects. Intrinsic motivation is provided by frustration-based action selection methods during learning, and tolerance values are determined based on impulsiveness of the robot. Our claim is that impulsiveness can be adjusted based on the task that the robot is executing. We have analyzed this mechanism on a robotic arm to learn graspability of different-shaped objects. Our results reveal that the intrinsic motivation helps the robot learn faster. Furthermore, the decision on graspability is made earlier by taking impulsiveness into account. Our future work includes extending the experiment set and investigating impulsiveness parameters in detail for different domains with varying time constraints.

Acknowledgment

This research is funded by a grant from the Scientific and Technological Research Council of Turkey (TUBITAK), Grant No. 111E-286. TUBITAK's support is gratefully acknowledged. We thank Burak Topal for his contribution for robotic arm movement and also thank Mehmet Biberci for his effort on vision algorithms.

References

1. Baranes, A., Oudeyer, P.Y.: Maturationally-constrained competence-based intrinsically motivated learning. In: Development and Learning (ICDL), 2010 IEEE 9th International Conference on. pp. 197–203. IEEE (2010)
2. Barto, A.G.: Reinforcement learning: An introduction. MIT press (1998)
3. Bicchi, A.: On the closure properties of robotic grasping. The International Journal of Robotics Research 14(4), 319–334 (1995)
4. Buss, M., Hashimoto, H., Moore, J.B.: Dextrous hand grasping force optimization. Robotics and Automation, IEEE Transactions on 12(3), 406–418 (1996)
5. Chebotar, Y., Kroemer, O., Peters, J.: Learning robot tactile sensing for object manipulation
6. Detry, R., Baseski, E., Popovic, M., Touati, Y., Kruger, N., Kroemer, O., Peters, J., Piater, J.: Learning object-specific grasp affordance densities. In: Development and Learning, 2009. ICDL 2009. IEEE 8th International Conference on. pp. 1–7. IEEE (2009)
7. Dimitrakakis, C., Lagoudakis, M.G.: Rollout sampling approximate policy iteration. Machine Learning 72(3), 157–171 (2008)
8. Ding, D., Liu, Y.H., Wang, S.: The synthesis of 3-d form-closure grasps. Robotica 18(01), 51–58 (2000)

9. Ersen, M., Ozturk, M.D., Biberci, M., Sariel, S., Yalcin, H.: Scene interpretation for lifelong robot learning. In: The 9th International Workshop on Cognitive Robotics (CogRob 2014) held in conjunction with ECAI-2014. Prague, Czech Republic (2014)
10. Grzyb, B., Boedecker, J., Asada, M., Del Pobil, A.P., Smith, L.B.: Between frustration and elation: Sense of control regulates the lntrinsic motivation for motor learning. In: Lifelong Learning (2011)
11. Huebner, K., Ruthotto, S., Kragic, D.: Minimum volume bounding box decomposition for shape approximation in robot grasping. In: Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on. pp. 1628–1633. IEEE (2008)
12. Kober, J., Peters, J.: Learning motor primitives for robotics. In: Robotics and Automation, 2009. ICRA'09. IEEE International Conference on. pp. 2112–2118. IEEE (2009)
13. Lenat, D.B.: Am: An artificial intelligence approach to discovery in mathematics as heuristic search. Tech. rep., DTIC Document (1976)
14. Peters: Machine learning of motor skills for robotics (2007)
15. Platt, R.: Learning grasp strategies composed of contact relative motions. In: Humanoid Robots, 2007 7th IEEE-RAS International Conference on. pp. 49–56. IEEE (2007)
16. Rusu, R.B., Cousins, S.: 3D is here: Point Cloud Library (PCL). In: IEEE International Conference on Robotics and Automation (ICRA). Shanghai, China (May 9-13 2011)
17. Rusu, R.B., Cousins, S.: 3d is here: Point cloud library (pcl). In: Robotics and Automation (ICRA), 2011 IEEE International Conference on. pp. 1–4. IEEE (2011)
18. Sauser, E.L., Billard, A.G.: Biologically inspired multimodal integration: Interferences in a human-robot interaction game. In: Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on. pp. 5619–5624. IEEE (2006)
19. iirgen Schmidhuber, J.: A possibility for lmplementing curiosity and boredom in model-building neural controllers (1991)
20. Steel, P., König, C.J.: Integrating theories of motivation. Academy of Management Review 31(4), 889–913 (2006)
21. Trevor, A., Gedikli, S., Rusu, R., Christensen, H.: Efficient organized point cloud segmentation with connected components. Semantic Perception Mapping and Exploration (SPME) (2013)
22. Uchibe, E., Doya, K.: Finding intrinsic rewards by embodied evolution and constrained reinforcement learning. Neural Networks 21(10), 1447–1455 (2008)
23. Watkins, C.J., Dayan, P.: Q-learning. Machine learning 8(3-4), 279–292 (1992)
24. Wong, P.T.: Frustration, exploration, and learning. Canadian Psychological Review/Psychologie canadienne 20(3), 133 (1979)