PROCEEDINGS A

rspa.royalsocietypublishing.org

Research

Article submitted to journal

Subject Areas:
biomechanics, mathematical modeling, artificial intelligence

Keywords:
Locomotion, Modeling, Reinforcement Learning

Author for correspondence:
Marcello Calisti
e-mail: m.calisti@santannapisa.it

THE ROYAL SOCIETY
PUBLISHING

Learning to Stop: A Unifying Principle for Legged Locomotion in Varying Environments

T. G. Thuruthel[1], G. Picardi[2,3], F. Iida[1], C. Laschi[2,3] and M. Calisti[2,3]

[1]Bio-Inspired Robotics Laboratory,Department of Engineering, University of Cambridge, UK
[2]The BioRobotics Institute, Scuola Superiore Sant'Anna, Pisa, Italy
[3]Department of Excellence in Robotics & AI, Scuola Superiore Sant'Anna, Pisa, Italy

Evolutionary studies have unequivocally proven the transition of living organisms from water to land. Consequently, it can be deduced that locomotion strategies must have evolved from one environment to the other. However the mechanism by which this transition happened and its implications on bio-mechanical studies and robotics research have not been explored in detail. This paper presents a unifying control strategy for locomotion in varying environments based on the principle of 'learning to stop'. Using a common reinforcement learning framework, deep deterministic policy gradient (DDPG), we show that our proposed learning strategy facilitates a fast and safe methodology for transferring learned controllers from the facile water environment to the harsh land environment. Our results not only propose a plausible mechanism for safe and quick transition of locomotion strategies from a water to land environment, but also provide a novel alternative for safer and faster training of robots.

1. Introduction

Fundamental models of dynamic gaits (e.g. running, trotting, galloping, etc.) provide a high level picture of biomechanics and control of legged locomotion. Models such as the Spring Loaded Inverted Pendulum (SLIP, [1]) elucidate the fundamental relationships among legs' compliance, speed, control, and shred a new light on the tight relationships among self-stabilizing mechanics and

feedback control [2, 3, 4, 5]. Beside its simplicity, the model successfully describe the behaviour of higher-order systems and it is employed as a reference for the control of single and multi-legged robotic devices [6, 7].

Such generality suggested the relevance of SLIP model also to locomotion in different media, i.e. in low gravity environments and in underwater environments. However, significant changes in the dynamics of the latter are difficult to be explained by the SLIP model alone and, indeed, animals employ a slightly different gait in the underwater environment, which is referred to as *punting* [8]. An extension of the SLIP model, called Underwater SLIP (USLIP, [9]), which takes into account the non-conservative nature of the system, the buoyancy, drag, and added mass effects of the punting gait, captures the dynamic of such locomotion [10], and it is successfully employed as a reference for the locomotion of underwater robots [11, 12].

The similarity between the two models envision the possibility that an overall and unifying control strategy could be employed by animals and humans to move in both environments. In particular, it is demonstrated by evolutionary algorithms that locomotion learnt in water can be beneficial for land gaits [13], and similarly that smooth locomotion changes can be elicited by moving robots from land to water [14, 15]. A similar conceptual hypothesis is proposed in this study, while we are looking for a control strategy who could be learnt and transferred from one environment to the other.

With this respect, reinforcement learning (RL) algorithms have been fairly successful in developing locomotion controllers. This is because of the complexity in modelling locomotion dynamics which makes learning-based approaches more suitable over analytical methods and the relative ease in defining the RL objective function. In fact, the locomotion problem has become a popular benchmark task in the reinforcement learning field. One of the earliest works on using RL for locomotion was done by Tedrake et. al. [16]. The key aspect of their work was the design of a simple passive dynamic bipedal walker which greatly simplified the learning problem. RL for single legged locomotion, a problem we are investigating in this work, was addressed using a model-free RL method called policy improvement using path integrals [17]. Their work showed how RL algorithms can be effectively used for exploiting the complex dynamics arising from compliant joints. Another interesting work showed how complex environment conditions can lead to the emergence of rich behaviours without explicit reward shaping [18]. Recent works have looked into deep Reinforcement learning algorithms for learning locomotion skills on physical prototypes [19, 20].

A significant challenge in using RL algorithms for legged locomotion is the one of running real-world trials. Accurate simulation models are difficult to develop, especially with complex environmental interactions. Recent techniques like domain randomization can be way around this problem, but it requires a large number of simulation trials and appropriate parameter settings [21]. Another solution, which this work will be adopting, is the concept of *shaping* [22, 23]. *Shaping* is the idea of smoothly changing the physics of the problem to accelerate the learning process. The underlying principle is that by learning to solve problems in a simpler environment will facilitate the learning of a similar problem in a more complex version of the environment. Randlov proved that for a finite Markovian Decision Process with a limited reward signal, it is guaranteed that if a series of tasks converges to the original one, then the optimal value function converges to the original one as well [23]. In [24], a temporary device to reduce gravity helped the learning of single-leg hopping, showcasing the potential of shaping in real world applications. In this paper we are proposing the transition from water to land, a landmark of earth colonization, as a naturally emerging *shaping* mechanism. We validate this hypothesis through simulations of fundamental legged models. For this, we propose the strategy of 'learning-to-stop' as a general locomotion plan that smoothly unifies a learned locomotion controller in both air and water. Using a deep reinforcement learning framework we show that the transfer of such learned controllers from water to air is faster and more efficient, as we would expect from an evolutionary viewpoint.

The goals of this work are: to investigate the relationships between SLIP and USLIP, by finding a possible unifying strategy for the control of legged locomotion in multiple media; to establish the role of the environment on the learning progress of legged locomotion; and eventually to propose a learning procedure which have evolutionary basis to explain animal locomotion and that could be employed in the training of effective legged devices, e.g. prosthetic devices or legged robots.

2. Materials and Methods

(a) Background on previous fundamental models

The steady state motion of the centre of mass (CoM) of animals performing various dynamic legged gaits can be described, abstracting the complexity of the animals' physical bodies [25], by a simple system consisting of a point-mass vaulting around a springy leg, the so-called Spring Loaded Inverted Pendulum (SLIP, [1]). Extensive analysis of the SLIP model revealed the relations between physical body properties (i.e. mass, leg length and leg stiffness) and locomotion features and stability [26] and stimulated research in the fields of biomechanics (e.g. [27, 28]) and control (e.g. [29, 30, 31]). In robotics, SLIP gave great impulse to the development of hopping and running machines, which achieved unprecedented degrees of speed and agility through the integration of compliant elements carefully dimensioned to meet the criteria dictated by the model (e.g. [32, 33, 34]).

In SLIP it is assumed that the resistance of air is negligible, thus the model is purely conservative and does not require the presence of an actuator to inject energy into the system. While this assumption is perfectly sensible in the case of terrestrial legged locomotion, the same cannot be said underwater, where the drag imposed on the body by a dense fluid introduces significant dissipation. Biological studies revealed fundamentals changes in the gait employed by crabs while running in water [8, 35] that simply cannot be obtained through SLIP solutions. At the same time, there has been a growing interest towards underwater legged vehicles (ULR) [36, 37, 38, 39, 11, 40, 12], which can extend the capabilities of traditional underwater robots thanks to their improved interaction with the seabed. In order to extend the benefits of an approach based on the SLIP model also to the underwater environment, an extension of SLIP which accounts for the contribution of water, namely Underwater SLIP (USLIP), has been introduced [9] and employed as a reference model in the design and control of underwater legged machines [11, 40, 12]. The SLIP and USLIP schematics are reported in Figure 1.

The equations of USLIP (Eq. 2.1) are reported below:

$$\begin{aligned} \ddot{x} &= -\frac{X}{m+M}\dot{x}|\dot{x}| + \frac{k(x-x_t)}{m+M}\left(\frac{r_0+r-l}{l}\right) \\ \ddot{y} &= -\frac{Y}{m+M}\dot{y}|\dot{y}| + \frac{ky}{m+M}\left(\frac{r_0+r-l}{l}\right) - \frac{(\rho_w V - m)g}{m+M} \end{aligned} \tag{2.1}$$

where $l = \sqrt{(x-x_t)^2 + y^2}$ is the length of the leg, $X = \frac{1}{2}c_x A_x \rho$ and $Y = \frac{1}{2}c_y A_y \rho$ are respectively the horizontal and vertical drag constants, $r(t) = r_s t$ is the linear leg elongation law and all variables and parameters are defined in Tab. 1. Typically, the system is started in *swimming phase*, (foot not in contact with the ground and mathematically modelled by setting the leg stiffness $k = 0$) with initial conditions $[x_0, \dot{x}_0, y_0, \dot{y}_0]$ and touch down angle α. In this phase the agent follows a ballistic trajectory and dissipates its energy to drag until the *touch-down condition* ($y = l \sin\alpha$) is met. Here the *punting phase* begins ($k \neq 0$), and the agent, following a minimal control law, extends its leg with constant velocity r_s to actively compress the leg spring and gain elastic energy. A new swimming phase begins when the forces on the spring balance out as geometrically expressed by the *lift-off condition* $l = l_0 + r$. For adequate choices of

the control parameters α and r_s the system converges to period trajectories which correspond to stable locomotion patterns.

Up to now USLIP has only been used to model agents moving in water on planet Earth, thus setting $g = 9.81m/s^2$ and $\rho = 1000kg/m^3$. However other environments can be modelled by the same equations by tuning the environment dependent parameters of Tab. 1. For example, running on land on Earth can be obtained with $g = 9.81m/s^2$ and $\rho = 0kg/m^3$ and running on the Moon with $g = 1.62m/s^2$ and $\rho = 0kg/m^3$. For these environments with negligible air density, the conservative SLIP model emerges by simply setting the leg extension speed $r_s = 0$.

(b) Learning Methodology

The common strategy for learning to locomote involves framing the problem as a reinforcement learning problem with the objective to reach a specific velocity or maximize the locomotion speed while satisfying other user constraints like avoiding obstacles, reducing energy expenditure, etc. However, such learned policies will be highly specific to the particular objective and hence typically not generalizable and robust. Recent findings show that robustness can be achieved by learning in a rich environment [19] or using domain randomization techniques [21]. Yet, none of these techniques provide a methodology to develop a learned policy that can maintain stable locomotion while being able to control other state variables. In this work, we derive a strategy inspired from RL approached used in autonomous flight, i.e learning to hover, or in our case, hop-in-place/learning-to-stop [41]. The idea is to learn a controller that brings the system to a zero horizontal velocity hop from the current state, in as few steps as possible. A closely related concept of capture point, where the strategy of placing the foot for stopping as been investigated for push recovery [42]. Hof showed that planning locomotion strategies based on the capture

Table 1. All variables and parameters of the presented model. Parameters are subdivided into control parameters, which can be modified online; design-dependent parameters, which relates to the intrinsic physical properties of the agent; and environment-dependent parameters, which model the interaction between the agent and the surrounding environment.

Variables	
x	CoM horizontal position
$\dot{x}$	CoM horizontal velocity
y	CoM vertical position
$\dot{y}$	CoM vertical velocity
x_t	horizontal foot position
Control parameters	
r_s	leg elongation speed
Δr	maximum elongation
α	leg angle at touch down
Design dependent parameters	
m	dry mass
r_0	rest length of leg
V_r	volume of the agent
k	spring stiffness
A_x	horizontal projection area
A_y	vertical projection area
c_x	horizontal drag coefficient
c_y	vertical drag coefficient
Environment dependent parameters	
g	gravity constant
ρ	density of fluid
M	added mass

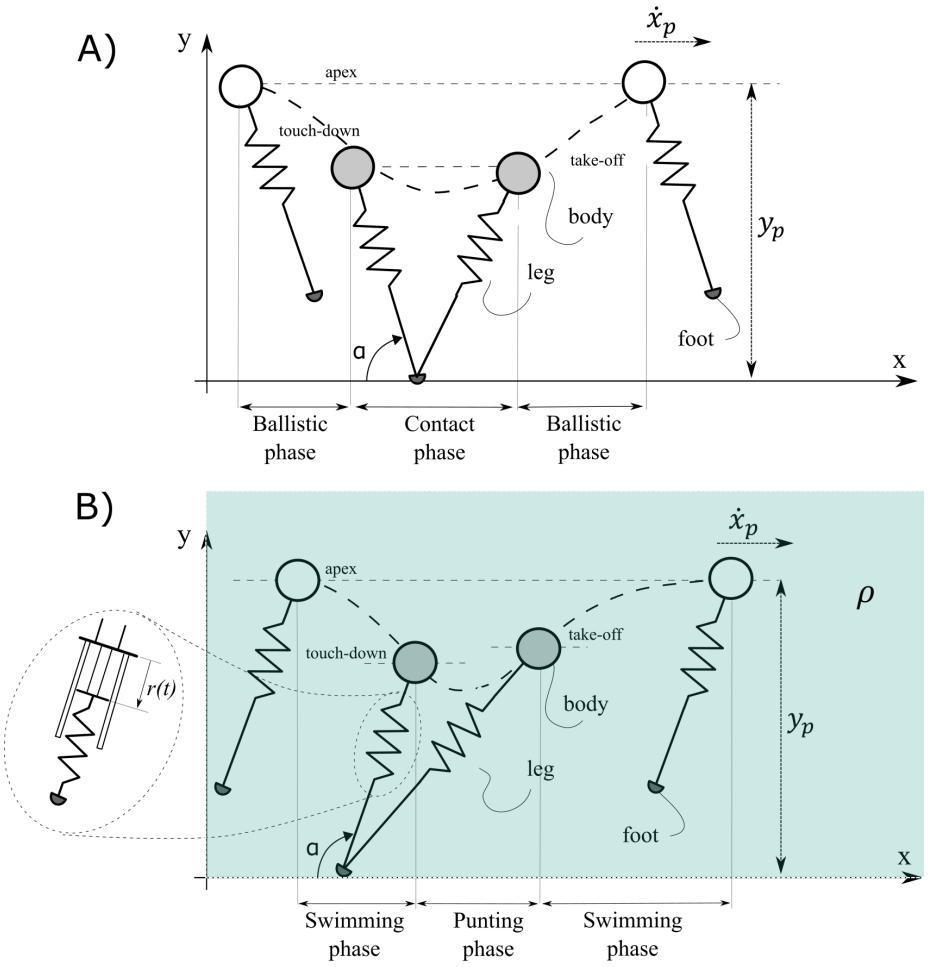

Figure 1. The SLIP (A) and USLIP (B) models with the state and control variables. The state of the system is full defined by the density of the environment ρ and the horizontal velocity $\dot{x}_p$ and height of the model y_p at the peak of each hop. The two control parameters are the angle of contact α and the elongation speed of the spring after contact $r(t)$.

point locations, in contrast to the typical ZMP formulations, allowed for stable walking in a simpler form [43]. Further studies showed the Capture Point methodology to be an effective and simple tool for the design of robust trajectory generators and feedback controllers for bipedal walking robots [44].

Once a controller for hopping-in-place is learned, tracking a set velocity can be done by offsetting the observed state of the model by the desired velocity to 'trick' the controller to move in the desired velocity. As the zero horizontal velocity is symmetric with respect to the design of the model, motion in either directions can be achieved without additional components. The next section presents the procedure for learning the hopping-in-place controller using a common reinforcement learning algorithm and the control architecture of the horizontal velocity tracker.

(i) Deep Deterministic Policy Gradient

As mentioned before, our objective is to develop locomotion controllers capable of locomotion in different environmental conditions and investigate any underlying properties among the controllers. For developing the controller we use a simple 2D SLIP model that is placed in a variable medium environment (See Figure 1). The state of the system is full defined by the density of the environment (ρ) and the horizontal velocity ($\dot{x}_p$) and height of the model (y_p) at the apex of each hop. The two control parameters are the angle of contact α and the elongation speed r_s of the spring after contact. Note that the density of the environment can be directly estimated by the vertical acceleration at the peak, but for simplicity, we assume this information is directly available to the controller.

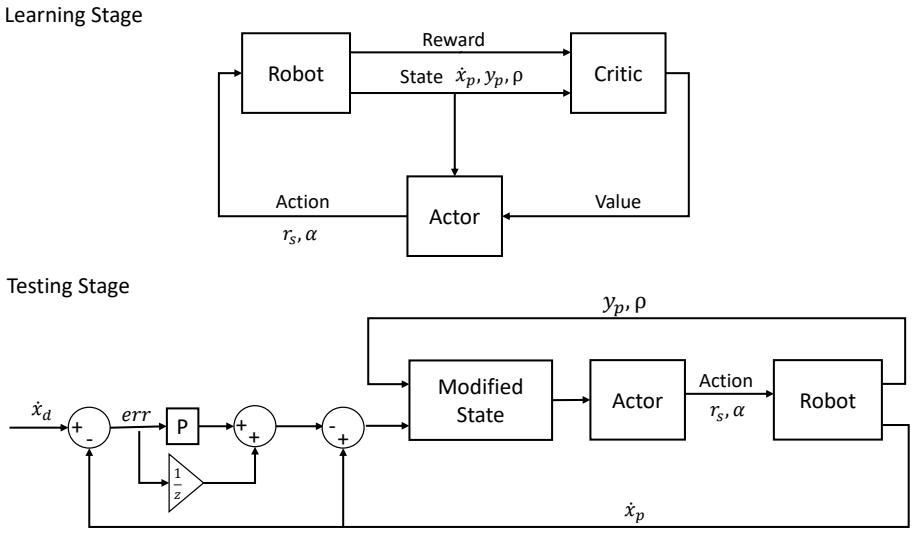

Figure 2. Above: The DDPG learning process. Below: The cascaded control architecture developed using the learned agent. Errors in velocity tracking are compensated by the proportional controller.

The deep deterministic policy gradient (DDPG) algorithm is a model-free, online, off-policy reinforcement learning method [45]. A DDPG agent is an actor-critic reinforcement learning agent that concurrently learns a value function and a policy. It uses off-policy data and the Bellman equation to learn the value function, and uses the value function to learn the policy (See Figure 2). The DDPG agent obtains the observation S ($\dot{x}_p, y_p, \rho$) and reward from the model after each step and updates the actor and critic using a mini-batch of experiences randomly sampled from the experience buffer. Each episode can have a maximum of n (5, in our case) steps and the model is reinitialized at a random state after the end of each episode. The reward obtained after each step is defined as:

$$StepReward, R = log(1 + 1/|\dot{x}_p|), if y_p > threshold, else \quad R = 0 \tag{2.2}$$

As described in the previous section, the objective of the step reward is to reduce the horizontal velocity at the apex. A discount factor of 0.5 is multiplied to the reward value after each step to obtain the episode reward. This is done to favor faster convergence to the zero velocity state. The threshold for detecting failure is based on the height of the model at peak with respect to the ground. The action A (α, r_s) is chosen by the policy using a stochastic noise model at each training

step. The variance of the noise is gradually reduced by a constant factor during training to favour exploitation over exploration.

At the start of training, the DDPG algorithm creates the critic $Q(S, A)$ the target critic $Q'(S, A)$ with the same random parameters. The critic, in our case, is a multi-layered neural network with ReLu activation layers. The output of the critic is the expected value for the given observation and action. The expected value is simply the sum of the current reward and the discounted future reward. Similarly, the actor $\mu(S)$ and the target actor $\mu'(S)$ is initialized with the same random parameters. The actor is also a multi-layered neural network with ReLu activation layers. At each step the actor executes the action A, observe the reward R and next observation S'. The experience (S, A, R, S') is then stored in the experience buffer. For updating the parameters of the actor and critic, random mini-batches of experience from the replay buffer are used. A small replay buffer will cause the algorithm to overfit and instability in learning, while a large replay buffer will slow down the learning process. The target networks are time-delayed copies of their original networks that slowly track the learned networks. Using these target value networks greatly improve stability in learning [45]. We use the MATLAB reinforcement learning toolbox for initializing and training the networks. The source code of the algorithm and the trained agents can be downloaded from: https://github.com/tomraven1/DDPG-hop

3. Results

This section presents the results of training an agent in various environments and how the strategy of transferring agents learned in a specific environment to another fares. Further analysis is done on the performance of the controller and the possibility of a universal controller for locomotion in water and air.

(a) Training Results

To test the learning algorithm on different mediums, two controllers are learned, separately in land and water. In order to introduce an understanding of medium properties on the controller strategy, the medium density is randomly initialized along with the other states for each episode. For the water environment, the medium density is defined as: $1 - |randn|/10$, (where 1 is the normalized water density), while for the land environment, the medium density is randomly initialized to: $0 + |rand|/10$. Here, $randn$ is a standard normally distributed random number. The same network architecture and agent/critic parameters are used for both scenarios for comparison purposes.

The training progress, indicated by the average reward of the agent in both the mediums, are shown in Figure 3. As observed in other works, learning to hop from-scratch in a dense environment is much easier [9]. This is because of the low inertial effects in a dense medium combined with the stabilizing effect of buoyancy. When we transfer the pre-trained actor and critic from the water environment to the air environment, without the experience replay, we see that the agent converges to high reward much faster (Figure 3). As the RL algorithm reaches a higher reward quickly in the water environment, even with the combined number of episodes, the transferred agent can obtain higher rewards than an agent trained directly on Land. This implies that the actor and/or critic learned in water and land must have significant similarities amongst themselves that is reflected in their parameter space too. Note that the variance and decay for exploration is also reinitialized for this network. So, the faster convergence is not because of a lower exploration parameter. As there is no additional cost in changing the medium (algorithmically), this is in fact a good strategy to guide your training, especially in cases where high rewards are sparse. Moreover, looking at the number of failures during training (See Figure 4), it is evident that transferred agent has lower number of falls (steps <5). This is important for this case, as falling in land is more perilous than falling underwater. Hence, such a transfer strategy can be useful for safe training for real robots too.

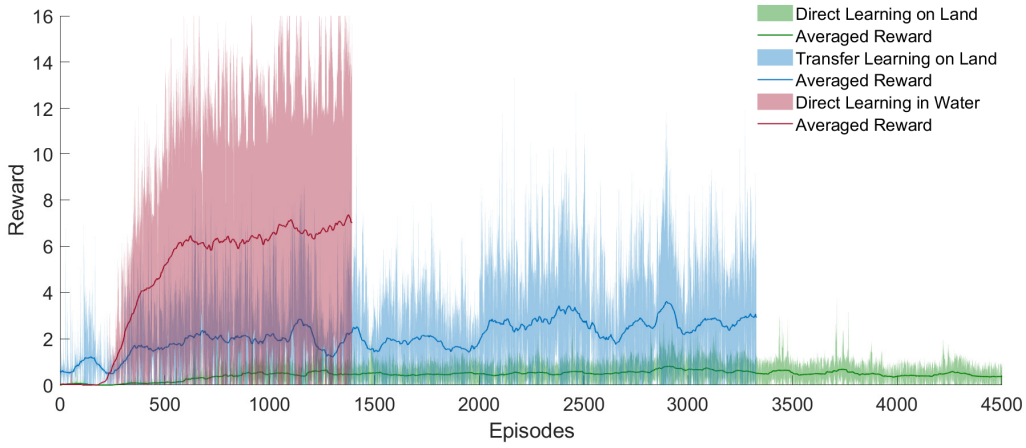

Figure 3. Agent reward during the training phase in different mediums. Transfer learning on land is done by retraining the actor and critic obtained in water.

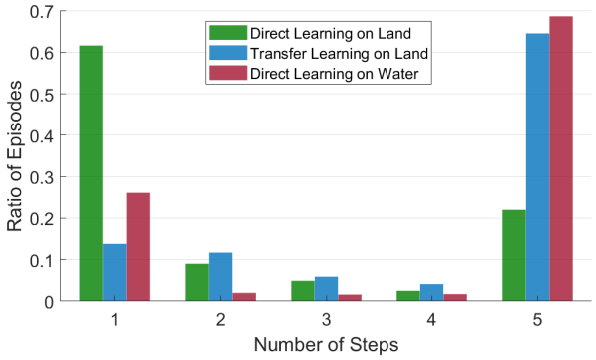

Figure 4. Number of steps during training for the first 1000 episodes shown as the ratio of the total episodes (A ratio of 1 implies that all the episodes resulted in the same number of steps). For our training, the maximum number of steps that the system can take for each episodes has been set at 5. Lower number of steps, most likely indicate a lower episode reward too.

The reason that learning to locomote in water is easier can be understood by looking at the reward landscape for stopping with respect to the control inputs for varying environmental mediums (Figure 5). In the water environment, high reward regions are dense and continuous with respect to the control inputs. Hence, gradient based approaches, like the one we adopt here, are well suited. On the other hand, for a pure air environment, the reward regions are sparse and discrete, making it difficult to be found. More importantly, when transitioning from water to land smoothly, we also observe a smooth variation in the reward landscape too, validating our hypothesis. This allows the agent transferred from water to air to quickly adapt. Notice that the reward function is almost independent from the elongation speed for the single step case, but with multiple steps, the elongation speed makes an effect to the reward obtained by the agent. This is because a non-optimal elongation speed can drive the next state of the system to regions outside the bounds of the problem, making the learning process unstable. For example, in the conservative air environment, a high elongation speed will eventually bring the height of the model to regions beyond the learning samples, causing instability in the learning process. In the water medium, a low elongation speed will eventually drive the model to a low height, causing

the system to fall. Hence, to ensure optimal convergence of the elongation law, it is necessary to increase the maximum number of steps allowed for each episodes to a high value.

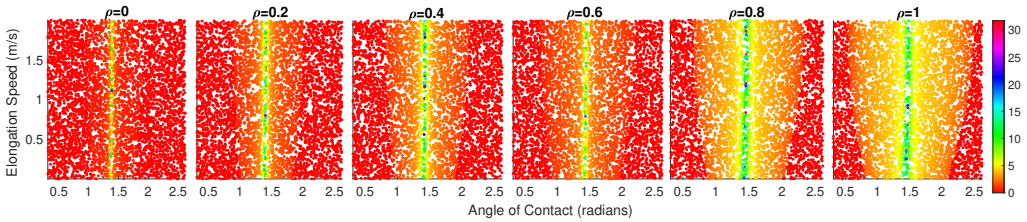

Figure 5. Reward landscape with respect to the control inputs for a fixed random initial state for different environmental mediums.

This smooth transition in the reward landscape is observed for all initial conditions. This is, however, only the case for the objective of minimizing the horizontal peak velocity. For other reward formulations, the transition across different medium is not as smooth, although certain level of information can still be transferred from one medium to the other. One such example is shown in Figure 6 for an objective to reach a horizontal velocity of 1m/s. More flexible objectives like maximizing the horizontal velocity would have higher variations because of the diverse limits on the locomotion speeds in different mediums.

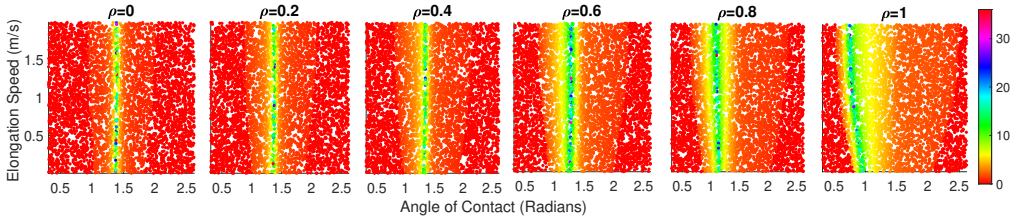

Figure 6. Reward landscape with respect to the control inputs for a fixed random initial state for different environmental mediums with a modified reward of moving at a fixed velocity of 1m/s.

(b) Controller Analysis

This section presents the results of the learned controller. After the actor is trained, the agent can be directly used as the zero-velocity tracking controller. By offsetting the observed horizontal velocity of the model by the desired velocity and PID term of the error, we obtain a cascaded control architecture for controlling the locomotion speed while maintaining stability (Figure 2). For simplicity, we only use a proportional component to stabilize the horizontal velocity. The base performance of the actor (zero desired velocity) is shown in Figure 7. For all these analyses, we use an environment condition of pure water and air ($\rho = 1 \; and \; 0$). The results of the controller show that the agent performs very well in water; reaching a stable limit cycle (i.e. hopping in place) in few steps, with the horizontal velocity reaching almost zero. This is as expected, based on the training results. The agent trained directly on land, manages to take few steps, but is not stable enough to sustain locomotion. This explains the relatively low rewards obtained during training. The transferred agent is able to maintain stable locomotion and reaches a bi-periodic limit cycle (i.e. hopping around zero speed). The final velocities are near zero, but not as accurate as the case of the agent in water. This could be because of the lack of energy dissipating components in land, which makes it much harder to reach the low energy state of zero-velocity hopping. Looking at

the control actions prescribed by the actor, we can see that the agent in water uses the maximum elongation speed for locomotion while the transferred agent uses the minimum elongation speed, essentially adapting to a passive SLIP system.

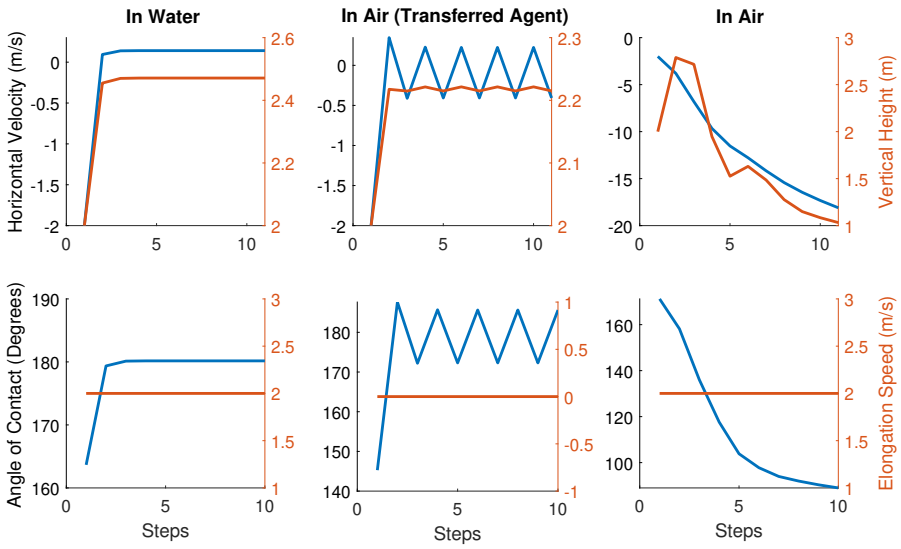

Figure 7. Base performance of the learned actor in different medium. The agent is trained to minimize the horizontal velocity at apex at each step. The state/action values at the apex are only shown here.

The performance of the cascaded control architecture for the agent trained in water and transferred to land is shown in Figure 8 and 9, respectively. As the agent learned directly on land was not stable enough for longer periods of locomotion, we ignore that case here. For both cases, a desired velocity of 1m/s is prescribed and initialized from the same initial conditions. The tracked velocity for different P values are shown in the results.

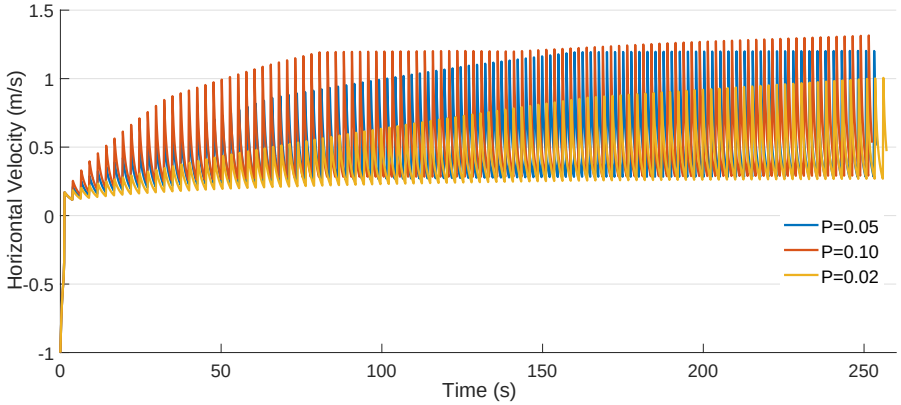

Figure 8. Performance of the cascaded controller in water for different P values for a desired velocity ($\dot{x}_d$) of 1m/s (See Figure 2 for reference). The whole trajectory over time is shown here unlike Figure 7.

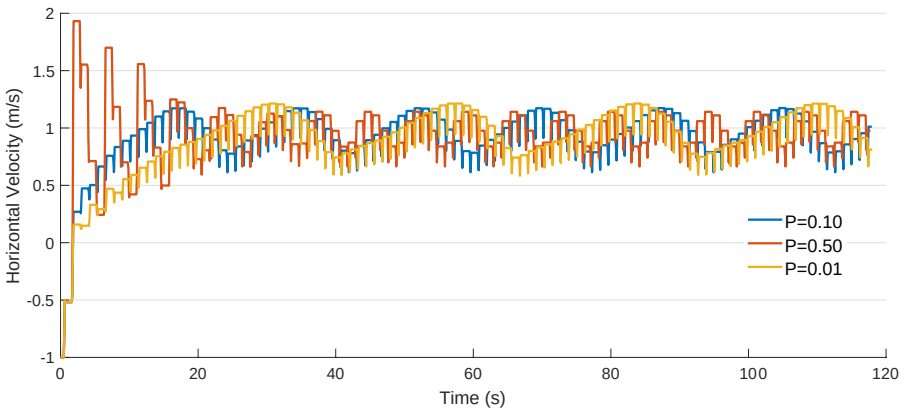

Figure 9. Performance of the cascaded controller in air for different P values for a desired velocity ($\dot{x}_d$) of 1m/s (See Figure 2 for reference). The whole trajectory over time is shown here unlike Figure 7.

(c) Gait Analysis

In this section we analyze the characteristics of an universal controller that is adapted to hop in both land and water. For this, we transfer the agent learned in water to an environment that alternates between pure water and air. The training is performed again with the same parameters as the previous sections. The maximum number of steps per episode is however increased to 10. The gaits observed for the single controller in both air and water is shown in Figure 10. Here, we are showing the gait after convergence for a prescribed horizontal velocity of 1 m/s. Note that in air, the controller settles to the conservative SLIP model, characterized by symmetric touch-down and lift-off angles and in water, the controller settles to the USLIP model with an acute touch-down and lift-off angles. As our model ignores many of the dissipative terms found in reality like friction and damping, the vertical displacements observed are higher than what is observed in nature.

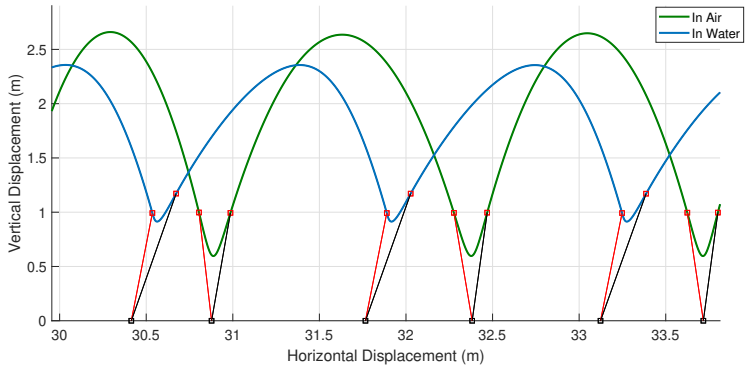

Figure 10. Gaits emergent from the universal controller in different mediums.

4. Discussion

The seminal SLIP model was developed about 30 years ago, and since then several extensions were proposed, among the latter the USLIP model for underwater legged locomotion. Our results shred a light on the tight relationship between them, and on a possible unique fundamental model which represents legged locomotion on different media. As shown in Figure 10, the learnt strategy is coherently employed in diverse media by adapting the angle of contact and elongation speed, the two key control parameters that differ between SLIP and USLIP, for the air and the water scenarios. This differentiate qualitatively the behaviors and it shows an energy-conservative behavior for the air case, where energy is transferred to the elastic components of the leg during the deceleration of the CoM and further released as kinetic energy during the acceleration, and a energy-dissipated behavior for the water case, where energy is dissipated by drag forces and it is injected by means of pushing actions. The model in our analysis being the same, the transition between the two diverse employments is dictated by the environment itself, through the reading of the density of the medium.

This emergent change in behaviour appears very similar to what is observed in a robotic salamander when the environment dictated the swap between walking and swimming gaits [14], and could explain the effectiveness of animals (including humans) to swiftly transition from land to low gravity or underwater environments [46, 8]: the fundamental model is indeed the same, and by sensing the environment the locomotion is directly adapted. It is worth to mention that, in our model, the change in density of the medium affects directly buoyancy and drag forces: both contributions could be plausibly detected by extero-perceptive or proprio-perceptive biological systems, such as the vestibular system.

It has to be noted, however, that prolonged exposition to low-gravity environment decreases locomotion capabilities during low-gravity to high-gravity transitions, e.g. on astronauts returning to the Earth, or going from space flights to Moon [47]. These impairments are related to significant changes in skeletal, muscular and sensing systems, which alter the sensory-motor coordination [48]: nevertheless, the hypothesis of a unique locomotion model still holds while such defects could be linked to the impossibility of complying to the model. Eventually, in the very first Moon locomotion, astronauts mainly reported falls related to the interaction with the ground or structures (rocks, ladders, etc.), rather than on actual locomotion impairments [48], while the natural evidences of underwater pedestrian locomotion positively confirm the existence of a transferable skill.

Moreover we claim that the abstract concept of balancing the centre of mass could be a reference strategy for legged locomotion in different media, and that legs' action is led by this balancing goal (i.e. learning to hop in place, or equivalently to reach $x_v = 0$). We report the learning progress of this strategy on different media (Figures 3), and the performance in air is much better for the transferred agent (Figures 7, middle column). With the help of the proposed cascaded controller, the model can reach a desired apex horizontal speed too, as shown in Figure 8. The cascaded controller in water modifies the system to act like a highly damped spring system with small steady state errors. Similarly, the controller transferred on land perform well in reaching the average target speed, with slight oscillations around the desired velocity (Figure 9). The oscillations can be further reduced by adding a derivative terms to the error values and the steady state error can be reduced with an integral term.

Beside no explicit references, to the authors knowledge, are present in the biological literature, several evidences support our "learning to stop" approach. Termination of gaits it is considered to involve prediction of center of mass position and speed [49], and how to place the foot accordingly; infants locomotion development involves the learning of braking actions [50]; facilitating actions in real environment promotes an early skill acquisition [51]; and most bouts of young infants are made of a few steps only (one to three, before stopping without an apparent reason) [52]. These observations could promote a view of "learning to stop", rather than explicitly "learning to move". This is supported also by the reward landscape for different objectives: learning to stop (Figure 5) appears much smoother and continuous than learning to move at

a constant velocity, Figure 6. This enables training of gradient based learning algorithms much more stable, robust and faster. Due to the symmetric nature of the reward with respect to the initial conditions, the strategy is also well suited for developing the cascaded architectures for building higher level controllers. Practically, such an approach will reduce the risk of damage as the controller tries to reduce the momentum of the system during the training process.

We do not presume that our hypothesis can explain all the facets of this complex research topic, however it appears as a simple and, to the author knowledge, plausible unifying objective to be taken into account during development. It is general enough to be resilient to change in body or environments (as happens continuously with children), it can be extended, with additional sensory input, to cluttered or complex surfaces (slippery grounds, gravel, grass, etc.), and it does not impose a hierarchical structure of skills developments, e.g. crawling as essential step for walking (which does not happen in all infants) [51].

By comparing the learning on the two different tested environments, we reported a clear asymmetry in the quality of the learned behaviour. Learning in water appears to be easy when compared to learning in air, and moreover the learnt control strategy could be positively transferred to the air environment (Figure 3). Reward rapidly grew and settle on a high value for the training in water with respect to air, which support the facilitating physical change proposed in [23]. Due to the low inertial conditions and stabilizing effects of buoyancy in water, the reward landscape is much wider (Figure 5), hence enabling faster and higher reward accumulation during training. With a massive number of training episodes ending at the first step for direct learning in air Figure 4, the agent has reduced number of chances to learn the correct control parameters: on the other hand, an agent transferred from the water environment allows an higher number of steps per training episode along with some prior knowledge of the control parameters, thus promoting the possibility of faster learning with fewer falls.

The last point we would like to discuss is related to the effectiveness of the transferred learning, i.e. water to land. The transferred agent performs surprisingly well both as overall reward (compared to air-only learning, Figure 3) and as number of steps for training event (Figure 4). The reward landscape explains the high performances both in water and how it would facilitate the transferred agent (Figure 5). With respect to previous works which explored the evolution of artificial creatures, it was already found how moving (swimming) in water was eventually beneficial for moving (crawling) on land [13]. Our results point in the same direction, pushing forward the same idea and promoting the concept that even legged agents can benefit from an initial learning phase in water. The use of fins as limbs for walking, jumping or crawling is reported in several fishes [53, 54, 55], and our results may suggest that effective land locomotion with limbs *must* be initiated in water environment first, then transferred to shore, and eventually to land. The underwater environment was not only a convenient environment where legged animals evolved, but it was the pivotal factor which enabled the development of limbed locomotion. The implication is that legged locomotion could not be produced directly on land, since it was too difficult and ineffective.

The presented results has also important implications on robotic learning approaches: the intrinsic self-stabilizing properties of the USLIP model allows the robotic hardware to experience less harmful impacts; diminishes the learning time and increases the learning episodes (pushes) for session. Eventually, the facility of learning in water and the possibility to transfer the learning on land may promote the adoption of on-line learning with the actual hardware, and to diminish the issues related to the reality gap between learning in simulation and in real environment.

5. Conclusion

Fundamental simplified models are powerful tools for studying and analysing complex dynamical systems. The SLIP and USLIP models have been fairly successful in explaining the observed gait patterns in nature with highly simplified mathematical models. In this work, we use these models along with the state of the art reinforcement learning algorithms to study the transition of locomotion behavior from water to air. Our findings show that transitioning from

water to land is not just an evolutionary manifestation, but has significant advantages from a controller development perspective. We show that transferring knowledge from controllers and models learned in water is much more efficient and safer than developing them from scratch in air. This *shaping* strategy is tested in simulation using a reinforcement learning algorithm called Deep Deterministic Policy Gradient. In order to unify and generalize the control strategy we also propose the concept of 'learning to stop' as a novel locomotion objective that enables us to smoothly transition from one medium to the other while allowing us freedom to develop higher level controllers with a cascaded architecture. The underlying rationale behind the transfer learning principle and the performance of the controller are quantitatively analyzed in this paper. This study not only provides corroborative evidences to the unification of locomotion models and behavior in various environments but also proposes the idea of training robots in simpler environments as a efficient safer methodology for acquiring new skills. An extension of the work would be to investigate the ideal training conditions and scenarios for better transfer of skills our environment to extraterrestrial environments.

Ethics. Insert ethics text here.

Data Accessibility. All the source code are available here: https://github.com/tomraven1/DDPG-hop

Authors' Contributions. Insert author contribute text here.

Competing Interests. Insert competing text here.

Funding. Insert funding text here.

Acknowledgements. Insert acknowledgment text here.

Disclaimer. Insert disclaimer text here.

References

[1] Reinhard Blickhan. "The spring-mass model for running and hopping". In: *Journal of Biomechanics* 22.11-12 (1989), pp. 1217–1227. DOI: https://doi.org/10.1016/0021-9290(89)90224-8.

[2] Reinhard Blickhan et al. "Intelligence by mechanics". In: *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 365.1850 (2007), pp. 199–220.

[3] Daniel P Ferris, Micky Louie, and Claire T Farley. "Running in the real world: adjusting leg stiffness for different surfaces". In: *Proc. R. Soc. B* 265 (1998), pp. 989–994.

[4] Andre Seyfarth et al. "A movement criterion for running". In: *Journal of Biomechanics* 35 (2002), pp. 649–655.

[5] Hartmut Geyer, André Seyfarth, and Reinhard Blickhan. "Compliant leg behaviour explains basic dynamics of walking and running". In: *Proc. R. Soc. B* 273.1603 (2006), pp. 2861–2867.

[6] Marc H Raibert. "Hopping in legged systems—modeling and simulation for the two-dimensional one-legged case". In: *IEEE Transactions on Systems, Man, and Cybernetics* 3 (1984), pp. 451–463.

[7] Mark H. Raibert. *Legged robots that balance*. MIT Press, 1986.

[8] MARLENE M Martinez, RJ Full, and MA Koehl. "Underwater punting by an intertidal crab: a novel gait revealed by the kinematics of pedestrian locomotion in air versus water". In: *Journal of Experimental Biology* 201.18 (1998), pp. 2609–2623.

[9] Marcello Calisti and Cecilia Laschi. "Morphological and control criteria for self-stable underwater hopping". In: *Bioinspiration & biomimetics* 13.1 (2017), p. 016001.

[10] Mrudul Chellapurath et al. "Locomotory behaviour of the intertidal marble crab (Pachygrapsus marmoratus) supports the underwater spring-loaded inverted pendulum as a fundamental model for punting in animals". In: *Bioinspiration & Biomimetics* 15.5 (July 2020), p. 055004. DOI: 10.1088/1748-3190/ab968c. URL: https://doi.org/10.1088%2F1748-3190%2Fab968c.

[11] Giacomo Picardi, Cecilia Laschi, and Marcello Calisti. "Model-based open loop control of a multigait legged underwater robot". In: *Mechatronics* 55 (2018), pp. 162–170.

[12] G Picardi et al. "Bioinspired underwater legged robot for seabed exploration with low environmental disturbance". In: *Science Robotics* 5.42 (2020).

[13] Francesco Corucci et al. "Evolving soft locomotion in aquatic and terrestrial environments: effects of material properties and environmental transitions". In: *Soft robotics* 5.4 (2018), pp. 475–495.

[14] A Crespi and A.˜J. Ijspert. "Amphibot II: an amphibious snake robot that crawls and swims using a central pattern generator". In: *9th International Conference on Climbing and Walking Robots*. Brussles, Belgium, 2006, pp. 19–27.

[15] Auke Jan Ijspeert. "Biorobotics: Using robots to emulate and investigate agile locomotion". In: 346.6206 (2014), pp. 196–203. DOI: 10.1126/science.1254486.

[16] Russ Tedrake, Teresa Weirui Zhang, and H Sebastian Seung. "Learning to walk in 20 minutes". In: *Proceedings of the Fourteenth Yale Workshop on Adaptive and Learning Systems*. Vol. 95585. Beijing. 2005, pp. 1939–1412.

[17] Péter Fankhauser et al. "Reinforcement learning of single legged locomotion". In: *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2013, pp. 188–193.

[18] Nicolas Heess et al. "Emergence of locomotion behaviours in rich environments". In: *arXiv preprint arXiv:1707.02286* (2017).

[19] Tuomas Haarnoja et al. "Learning to walk via deep reinforcement learning". In: *arXiv preprint arXiv:1812.11103* (2018).

[20] Zhaoming Xie et al. "Feedback control for cassie with deep reinforcement learning". In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2018, pp. 1241–1246.

[21] Josh Tobin et al. "Domain randomization for transferring deep neural networks from simulation to the real world". In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2017, pp. 23–30.

[22] Vijaykumar Gullapalli and Andrew G Barto. "Shaping as a method for accelerating reinforcement learning". In: *Proceedings of the 1992 IEEE international symposium on intelligent control*. IEEE. 1992, pp. 554–559.

[23] J RANDLOV. "Shaping in reinforcement learning by changing the physics of the problem". In: *Proceedings of the seventeenth international conference on machine learning, 2000*. Morgan Kaufmann. 2000.

[24] Steve Heim et al. "Shaping in practice: training wheels to learn fast hopping directly in hardware". In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2018, pp. 1–6.

[25] Reinhard Blickhan and Robert J Full. "Similarity in multilegged locomotion: Bouncing like a monopode". In: *Journal of Comparative Physiology A* 173.5 (1993), pp. 509–517. DOI: https://doi.org/10.1007/BF00197760.

[26] Andre Seyfarth et al. "A movement criterion for running". In: *Journal of biomechanics* 35.5 (2002), pp. 649–655.

[27] Robert J Full and Daniel E Koditschek. "Templates and anchors: neuromechanical hypotheses of legged locomotion on land". In: *Journal of experimental biology* 202.23 (1999), pp. 3325–3332.

[28] Richard Altendorfer, Daniel E Koditschek, and Philip Holmes. "Stability analysis of legged locomotion models by symmetry-factored return maps". In: *The International Journal of Robotics Research* 23.10-11 (2004), pp. 979–999.

[29] Ioannis Poulakakis and Jessy W Grizzle. "The spring loaded inverted pendulum as the hybrid zero dynamics of an asymmetric hopper". In: *IEEE Transactions on Automatic Control* 54.8 (2009), pp. 1779–1793.

[30] Jessica K. Hodgins and Mark H. Raibert. "Adjusting Step length for Rough Terrain Locomotion". In: *IEEE Transactions of Robotics and Automation* 7.3 (1991), pp. 289–298.

[31] Dominic Lakatos, Werner Friedl, and Alin Albu-Schäffer. "Eigenmodes of nonlinear dynamics: Definition, existence, and embodiment into legged robots with elastic elements". In: *IEEE Robotics and Automation Letters* 2.2 (2017), pp. 1062–1069.

[32] R. Altendorfer et al. "RHex: A biologically inspired hexapod runner". In: *Autonomous Robots* 11.3 (2001), pp. 207–213. ISSN: 09295593. DOI: 10.1023/A:1012426720699.

[33] Marc Raibert. *BigDog, the rough-terrain quadruped robot*. Vol. 17. 1 PART 1. IFAC, 2008, pp. 10822–10825. ISBN: 9783902661005. DOI: 10.3182/20080706-5-KR-1001.4278. URL: http://dx.doi.org/10.3182/20080706-5-KR-1001.01833.

[34] Sangok Seok et al. "Design principles for energy-efficient legged locomotion and implementation on the MIT cheetah robot". In: *Ieee/asme transactions on mechatronics* 20.3 (2014), pp. 1117–1129.

[35] Mrudul Chellapurath et al. "Locomotory behaviour of the intertidal marble crab (Pachygrapsus marmoratus) supports the underwater spring loaded inverted pendulum as fundamental model for punting in animals". In: *Bioinspiration & Biomimetics* (2020).

[36] Helen Greiner et al. "Autonomous legged underwater vehicles for near land warfare". In: *Proceedings of Symposium on Autonomous Underwater Vehicle Technology*. IEEE. 1996, pp. 41–48.

[37] Jun'ichi Akizono et al. "Seabottom roughness measurement by aquatic walking robot". In: *Oceans' 97. MTS/IEEE Conference Proceedings*. Vol. 2. IEEE. 1997, pp. 1395–1398.

[38] Joseph Ayers and Jan Witting. "Biomimetic approaches to the control of underwater walking machines". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 365.1850 (2007), pp. 273–295.

[39] Jung-Yup Kim and Bong-Huan Jun. "Design of six-legged walking robot, Little Crabster for underwater walking and operation". In: *Advanced Robotics* 28.2 (2014), pp. 77–89.

[40] Giacomo Picardi et al. "Morphologically induced stability on an underwater legged robot with a deformable body". In: *The International Journal of Robotics Research* (2019), p. 0278364919840426.

[41] H Jin Kim et al. "Autonomous helicopter flight via reinforcement learning". In: *Advances in neural information processing systems*. 2004, pp. 799–806.

[42] Jerry Pratt et al. "Capture point: A step toward humanoid push recovery". In: *2006 6th IEEE-RAS international conference on humanoid robots*. IEEE. 2006, pp. 200–207.

[43] At L Hof. "The 'extrapolated center of mass' concept suggests a simple control of balance in walking". In: *Human movement science* 27.1 (2008), pp. 112–125.

[44] Johannes Englsberger et al. "Bipedal walking control based on capture point dynamics". In: *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2011, pp. 4420–4427.

[45] Timothy P Lillicrap et al. "Continuous control with deep reinforcement learning". In: *arXiv preprint arXiv:1509.02971* (2015).

[46] Brittany L Coughlin and Frank E Fish. "Hippopotamus Underwater Locomotion: Reduced-Gravity Movements for a Massive Mammal". In: *Journal of Mammalogy* 90.3 (2009), pp. 675–679. ISSN: 0022-2372. DOI: 10.1644/08-MAMM-A-279R.1. URL: http://jmammal.oxfordjournals.org/content/90/3/675.

[47] Francesco Lacquaniti et al. "Human locomotion in Hypogravity: from basic research to clinical applications". In: *Frontiers in physiology* 8 (2017), p. 893.

[48] Jacob J Bloomberg et al. "Risk of impaired control of spacecraft/associated systems and decreased mobility due to vestibular/sensorimotor alterations associated with space flight". In: (2015).

[49] David A Winter. "Human balance and posture control during standing and walking". In: *Gait & posture* 3.4 (1995), pp. 193–214.

[50] Simone V Gill, Karen E Adolph, and Beatrix Vereijken. "Change in action: How infants learn to walk down slopes". In: *Developmental science* 12.6 (2009), pp. 888–902.

[51] Karen E Adolph, Justine E Hoch, and Whitney G Cole. "Development (of walking): 15 suggestions". In: *Trends in cognitive sciences* 22.8 (2018), pp. 699–711.

[52] Whitney G Cole, Scott R Robinson, and Karen E Adolph. "Bouts of steps: The organization of infant exploration". In: *Developmental psychobiology* 58.3 (2016), pp. 341–354.

[53] Heather M King et al. “Behavioral evidence for the evolution of walking and bounding before terrestriality in sarcopterygian fishes”. In: *Proceedings of the National Academy of Sciences* 108.52 (2011), pp. 21146–21151.

[54] Sandy M Kawano and Richard W Blob. “Propulsive forces of mudskipper fins and salamander limbs during terrestrial locomotion: implications for the invasion of land”. In: *Integrative and comparative biology* 53.2 (2013), pp. 283–294.

[55] CM Pace and Alice C Gibb. “Mudskipper pectoral fin kinematics in aquatic and terrestrial environments”. In: *Journal of Experimental Biology* 212.14 (2009), pp. 2279–2286.