International Journal of Humanoid Robotics (2014). 11(2), 1450012
DOI: http://dx.doi.org/10.1142/S0219843614500121

Full-Body Postural Control of a Humanoid Robot with Both Imitation Learning and Skill Innovation

Miguel González-Fierro* and Carlos Balaguer†

Robotics Lab, Universidad Carlos III of Madrid,
Avda. de la Universidad 30, Leganés, Madrid 28912, Spain
**mgpalaci@ing.uc3m.es*
†balaguer@ing.uc3m.es

Nicola Swann

School of Life Sciences,
Kingston University, Penrhyn Road,
Kingston Uppon Thames, KT1 2EE, UK
nicola.swann@kingston.ac.uk

Thrishantha Nanayakkara

Center for Robotics Research,
King's College London, Strand St., London WC2R 2LS, UK
thrish.antha@kcl.ac.uk

In this paper, we present a novel methodology to obtain imitative and innovative postural movements in a humanoid based on human demonstrations in a different kinematic scale. We collected motion data from a group of human participants standing up from a chair. Modeling the human as an actuated 3-link kinematic chain, and by defining a multi-objective reward function of zero moment point and joint torques to represent the stability and effort, we computed reward profiles for each demonstration. Since individual reward profiles show vari-ability across demonstrating trials, the underlying state transition probabilities were modeled using a Markov chain. Based on the argument that the reward profiles of the robot should show the same temporal structure of those of the human, we used differential evolution to compute a trajectory that fits all humanoid constraints and minimizes the difference between the robot reward profile and the predicted profile if the robot imitates the human. Therefore, robotic imitation involves developing a policy that results in a temporal reward structure, matching that of a group of human demonstrators across an array of demonstrations. Skill innovation was achieved by optimizing a signed reward error after imitation was achieved. Experimental results using the humanoid HOAP-3 are shown.

Keywords: Learning from demonstration; skill innovation; postural control; humanoid robot.

1. Introduction

Consider a child learning motor skills based on demonstrations performed by his parent. In this case, the problem of relating demonstrations performed by the parent to the child's own kinematic scale, weight and height, known as the correspondence problem, would be one of the complex challenges that should be solved first. The correspondence problem is one of the crucial problems of imitation and can be stated as the mapping of action sequences between the demonstrator and the imitator.[1,2] This problem can be solved by mapping movements made in a different kinematic scale to a common domain, such as a set of optimality criteria. From that perspective, the child could find a solution which fits his own muscular strength, size, reachable space and kinematic characteristics which somehow matches the level of optimality of demonstrations performed by the parent. Moreover, if comparisons are made in an optimality domain, the child could even innovate solutions that can be more relevant to his kinematic structure, but closely follow the optimal solution demonstrated by the parent. This comparison is best done in a common reward landscape, specified by a set of reward functions rather than in the kinematic domain or in the muscle effort domain, since similar behavioral goals should give similar trajectories in a common reward landscape subject to a set of constraints.

This paper presents an advancement in how a humanoid robot can learn to imitate and innovate motor skills from demonstrations of human teachers of larger kinematic structures and different actuator constraints. We present experimental results for the task of standing up from a chair to a stable upright posture, where the robot has to transit from one stable posture to another via a set of unstable states.

1.1. *Foundations of skill innovation in humans*

A wide range of work has been done in the area of *observational learning* from a psychological point of view. Thompson suggested that children learn not only by imitating, but also by understanding how the process works, what is known as *emulation learning*.[3] For example, to understand that a doorknob twist will open a door, will help to learn how to leave a room. Even, the high predisposition of children to learn from observation suggested a more appropriate name for the human species: *homo imitans*, which means "man who imitates".[4] There are also many experiments conducted with apes[5] or more recently[6] that support the argument that learning based on demonstrations can happen in intrinsic domains to do with the context of the kinematic domain.

It has been demonstrated that even animals can outperform the optimality of the demonstrated behavior in certain contexts. In a task of pushing a lever to obtain a food reward, rats finally associated the amount of food to the rate of pushing the lever, which was not demonstrated at the beginning.[7] Similar observations have been made in other experiments with birds[8] and apes.[9] Even Piaget, the father of the constructivist theory of knowing, hypothesized that the likelihood of matching a response may depend on the expected outcome for the observer.[10]

The phenomenon known as *goal emulation* shows that the observer can reproduce the result of a behavior with a method slightly different from that of the demonstrator.[11,12] This is similar to what Mitchell calls *fourth-level imitation*, where an individual tends to reproduce a model understanding the consequences of that behavior, and performs a different behavior maintaining what they called *intentionality*.[13]

Recent work in emulation show that apes are more suitable to emulate while children show more tendency to *over-imitate*, in the sense that children make an attempt to improve the optimality of the learnt skills. In that sense, skill innovation is an essential part of the human behavior.[14]

1.2. *Learning from demonstration and skill innovation in robots*

A humanoid robot sharing tools and space in a human society can benefit from a sound framework of learning based on demonstrations, that can vary across trials.[15] Despite the challenges to solve the correspondence problem,[1,2,16–18] there has been a growing interest in this area due to several advantages such as the simplification of communicating a complex behavior through a demonstration,[19] the absence of the need to have complex mathematical models of the dynamical system to learn an optimal behavior, and the fact that it does not require an expert teacher to perform the demonstrations, which simplifies the information gathering process.[20]

Similar to the biological world, robotic imitation is achieved under a set of schemes, as stated by Schaal[1]:

- Determine what to imitate, inferring the goal of the demonstrator.
- Establish a metric for imitation.
- Map between dissimilar bodies.
- Compute the control commands to perform the imitation behavior.

Several studies have been conducted in the area of learning from demonstration (LfD) using Gaussian mixture models (GMM), that encode a set of trajectories, and Gaussian mixture regression (GMR), to obtain a generalized version of these trajectories to perform a robot movement.[21]

The problem of skill transfer and whole body motion transfer has been an interesting area of research in recent years. Some studies addressed the problem of manipulating the angular momentum of the center of mass (COM),[22] using graphs and Markov chains,[23] defining a spatio-temporal models based on movement primitives[24] or encoding and organizing learned skills.[25]

In the process of learning, as psychological and biological studies support for animals and humans, robots should be able to innovate new behavioral solutions, that fit their constraints, to behave more efficiently.[14] In this regard, reinforcement learning (RL) is a good framework to innovate behaviors, since we can construct a reward landscape such that some search mechanism could explore for better actions in the neighborhood of demonstrations.[18,26,27]

Mixing imitation learning with RL produces a set of benefits, as claimed by Barrios.[28] It diminishes the computational time of convergence, since the search space is reduced. The innovation is based on actions that the robot has observed, so it is easier to improve this behavior. Furthermore, RL algorithms do not need to have the states and actions defined *a priori*. Some works combine LfD with RL, to teach the humanoid robot a constrained task of placing a cylinder in a box[29] or to teach a robot how to hit a baseball.[18]

1.3. *Overview of the proposed method*

In this paper, we address the problem of imitation and innovation learning in a task of standing up from a chair. This particular posture control task takes the human body from one stable posture (seated) to the other (standing) through a series of unstable postures. In this task, a robot would benefit from demonstrations to avoid excessive activation of joints to reach the second stable posture. We captured data

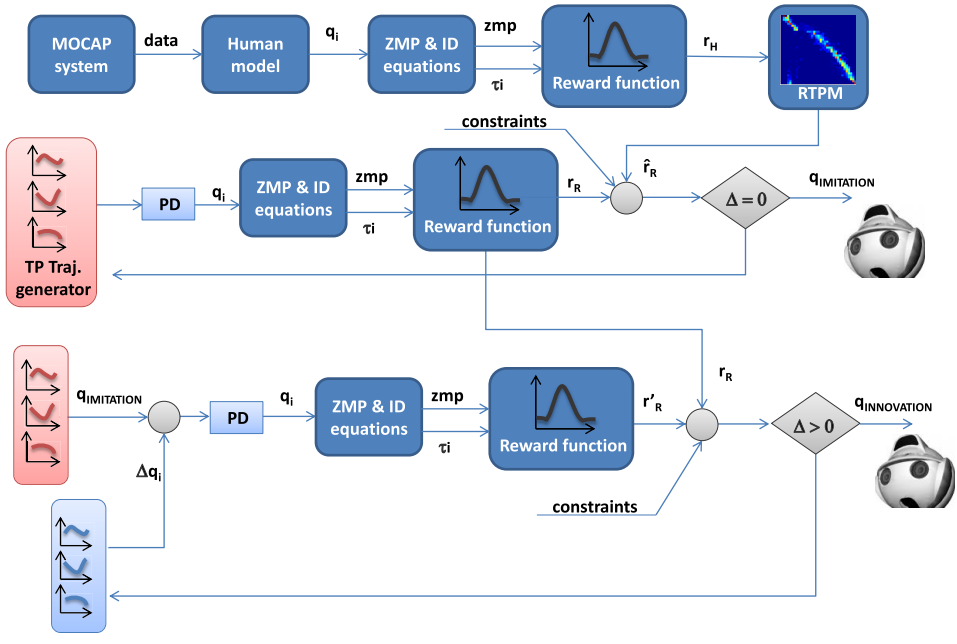


Fig. 1. Overview of the algorithm. We collected data from a MOCAP system and model the human as an actuated 3-link kinematic chain, where q_i are the joint positions. After computing the ZMP and joint torque τ_i, we define a reward function for the human r_H. This is done for all 160 demonstrations of standing up from a chair. Then, the RTPM is obtained. Using DE, we generate a new triple pendulum articular trajectory and obtain the reward profile r_R. This profile is compared with the predicted reward profile if the robot behaves like a human $\hat{r}_R$. The optimization process ends when the difference is small producing the imitation solution. Furthermore, we perturb the imitation solution Δq_i and compute a new reward profile r'_R that is compared with the imitation reward (r_R). The optimization process ends when the imitation reward is higher than the innovation reward, producing the innovation solution.

from eight human participants performing 20 consecutive demonstrations each, using a Qualisys motion capturing system. Modeling the humans as an actuated 3-link kinematic chain, we computed a reward function of stability and effort which we used as a common basis of comparison with the humanoid robot. Using Markov chains theory, we summarized the performance of all human participants in a transition probability matrix of the scalar reward that defines the optimality of the behavior. Using differential evolution (DE),[30] we optimized the robot joint trajectory to minimize the difference between the predicted reward and the real reward achieved by human demonstrators, subject to the constraints of zero moment point (ZMP) limits, maximum and minimum torque and joint limits. Once this first stage of imitative learning is accomplished, we proceed to explore new solutions with better rewards in the neighborhood of human-like movement subject to constraints, which we call skill innovation. Figure 1 shows the summary of the algorithm discussed in this paper.

The rest of the paper is organized as follows. First, in Sec. 2, an overview of Markov decision process is presented. Afterwards, in Sec. 3, we discuss the imitation process and how it is applied to the humanoid robot. In addition, in Sec. 4, we explain the skill innovation process. In Sec. 5, we define the mathematical tools that will be used and the representation of the demonstrated behavior. Finally, in Sec. 6 the experimental results are presented and in Sec. 7 the whole method is discussed and compared with other related approaches.

2. Extracting a Stochastic Template from Human Demonstrations

This section presents the process of extracting a stochastic model of the human behavior, that will be transferred to the robot. This behavior is presented in the form of a probability reward transition probability matrix (RTPM) domain, that can be used to compute an optimal robotic behavior learned based on human demonstrations.

Robot LfD, also called robot programming by demonstration or imitation learning, is a powerful method to reduce the space of solutions and accelerate the learning process. LfD is a natural way to interact with the robot and does not require an expert teacher. Furthermore, in contrast to slow RL or trial-and-error learning methods, it can easily find a good solution from the observed demonstrations (local optima).

We address two main challenges when using human demonstrations to train a humanoid robot. First, the robot is of a much smaller kinematic structure compared to the human demonstrators, while being limited by different actuator constraints, causing a correspondence problem.[1] Second, we noticed that the demonstrations performed by a group of eight human participants were variable across trials of a given participant and across the average behaviors of individuals.

We proposed to solve the correspondence problem by finding a common constrained reward domain for the behavior of both the humans and the robot. Once a set of reward functions have been identified, we approach to solve the second

problem by taking a stochastic approach to model the reward transition probability distribution for all demonstrated trials by all participants. More specifically, we construct a Markov chain using the discretized reward profiles for each demonstrated trial.

The advantage of a Markov chain to model human demonstrations in this manner is that we can use a particular reward value, obtained by the robot at any given time to predict the future reward it would obtain if it follows a policy similar to humans in a qualitative sense. Errors of such predictions can be used as feedback signals to update the policy of the robot. Another advantage of the Markov chain is that we can find a single state transition probability distribution, or a transition matrix that summarizes the dominant intrinsic structure of demonstrations performed by many individuals subject to variability.

2.1. *Markov chains and transition matrix*

A Markov chain is a random process that can define the behavior of a dynamical system under the Markov property. This property assumes that a state has all the required information to make a decision about the future.

A first order Markov chain is defined as a series of random variables or states $s_1, \ldots, s_N$ where

$$P(s_{N+1}|s_1, \ldots, s_N) = P(s_{N+1}|s_N). \tag{1}$$

If we use a Markov chain to explain a behavior, we could predict the next state in a sequence. The distribution of the prediction will depend only on the previous state and will be independent of all early states. In other words, the defining characteristic of a Markov chain is that its future trajectories depend on its present and its past only through the current value.[31]

For nth and $(n+1)$th trials, if a state s_N has an outcome j (i.e., $s_N = j$) and $s_{N+1} = k$, the transition probability associated with both trials is $P(s_{N+1} = k|s_N = j) = p_{jk}$.

We can specify a Markov chain given the initial probability distribution $P(s_1)$ and the conditional probabilities in the form of a transition probability matrix or *Markov transition matrix* T, where T is a stochastic matrix, i.e., it satisfies that every row is a probability distribution and it is a square matrix with non-negative elements.

$$p_{jk} \geq 0, \quad \sum_j p_{jk} = 1 \quad \text{for all } j. \tag{2}$$

This matrix may be written in the form

$$T = \begin{pmatrix} p_{11} & p_{12} & p_{13} & \cdots \\ p_{21} & p_{22} & p_{23} & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix}. \tag{3}$$

Algorithm 1. Transition Probability Matrix Computation

Create a 1D array R of size 1 by N from the minimum to maximum possible value in the trajectory
Create an N by N null matrix T.
for all $trial = 1 : t$ **do**
  for all $SamplingStep = 1 : s$ **do**
    Calculate the trajectory and find its bin in the 1D array R.
    Store the bin number in the array $B(SamplingStep)$
    if $SamplingStep > 1$ **then**
      $T(B(SamplingStep - 1), B(SamplingStep)) += 1$
    end if
  end for
end for
Normalize each raw of T so that the sum of each raw adds up to 1.

The transition probabilities and the transition matrix are defined for a unit-step transition, however, we can define a m-step transition in the future. The m-step transition probability is defined by

$$P(s_{N+M} = k | s_N = j) = p_{jk}^M, \tag{4}$$

and the m-step transition matrix is denoted by T^M.

Thus, the probability of a state with m-steps in the future can be denoted by

$$P(s_{N+M}) = P(s_N) T^M. \tag{5}$$

In Algorithm 1, the computation of the transition matrix is presented.

3. Imitation of the Human Behavior

In this section, we explain the process of how the robot performs a meta-stable standing up postural movement based on human demonstrations.

Since a human and a humanoid robot are morphologically similar, the optimality criteria guiding the behavioral policies should be comparable. In this paper, we use this premise to make a humanoid robot learn from an array of demonstrations performed by a group of eight human participants, in the task of standing up from a seated posture to an upright posture. In the case of the robot, both postures were manually selected. The initial posture was achieved by manually placing the robot on a small chair. The height of the chair was selected to make sure that the robot does not exceed the maximum torques and the seated posture is such that the ZMP is a little outside the sole of the feet. The final posture was a stable upright stand-up position. These postures can be changed as long as the initial posture do not surpass the maximum torque and the final posture is stable.

Postural control can be defined as controlling the body's position in space for the purpose of stability and orientation for the robot to move from one static posture to another.[32] In that sense, we can define a set of static posture primitives, such as to be sitting down, to be standing up, to be lying down, etc. We also can define a set of dynamic postures as the transitions between static postures. An example of these can be the actions of standing up, sitting down, walking, running, jumping, etc.

3.1. *Behavior prediction through reward transition probability matrix*

Using the transition matrix, we can predict the probability of the future reward of the human.

$$P_{\text{human}}(k+n) = P_{\text{human}}(k)\, T_{\text{human}}^{n}, \tag{6}$$

where $P_{\text{human}}(k)$ is the probability in step k, $P_{\text{human}}(k+n)$ is the probability in N steps in the future and T_{human} is the transition matrix.

If the robot is going to behave like the human, we can suppose that their transition matrices are the same:

$$T_{\text{human}} = T_{\text{robot}} = T. \tag{7}$$

So if we know the reward probability in step k, we can predict the future probability in step $k+n$.

$$P_{\text{robot}}(k+n) = P_{\text{robot}}(k)\, T_{\text{robot}}^{n} = P_{\text{robot}}(k)\, T^{n}. \tag{8}$$

We defined the fitness function as the difference between the predicted reward of the robot if it behaves like a human and the actual reward, under some constraints. The predicted reward is obtained as the expected value of the probability in (6), and it is given by

$$\hat{r}_R(k) = E[P_{\text{robot}}(k)\, T^{n}]. \tag{9}$$

Furthermore, we added as a constraint the ZMP limits, torque limits and joint limits.

We defined the fitness function as

$$\min f = \sum_{k=1}^{N-1} (\hat{r}_R(k+1) - r_R(k+1))^2, \tag{10}$$

$$\theta_{\min} \leq \theta \leq \theta_{\max}, \tag{11}$$

where θ represents ZMP, torque or joint position.

In Algorithm 2, an outline of the imitation process is presented. This algorithm can be easily implemented minimizing the fitness function (10) using DE.

Algorithm 2. Imitation Learning

Require: Identification of human and robot model
Require: K_p, K_d
Compute ZMP_H for all trials of every human
Compute τ_H for all trials of every human
Compute r_H for all trials of every human
Calculate the RTPM T
while $F \neq 0$ **do**
  New q_{m1}, q_{m2}, q_{m3} using Differential Evolution
  Compute piecewise polynomial
  Simulate system using K_p, K_d
  Compute ZMP($\mathbf{q}, \dot{\mathbf{q}}, \ddot{\mathbf{q}}$)
  Compute ID($\mathbf{q}, \dot{\mathbf{q}}, \ddot{\mathbf{q}}$)
  Compute the reward $\hat{r}_R$
  Compute the estimated reward r_R using (9)
  Compute fitness function F using (10)
end while

4. Innovation of the Human Behavior

Similar to how a child would try to improve learned behaviors based on demonstrations, often known as *over-imitation*, a robot could use heuristic search algorithms to explore the reward landscape for better behavioral policies in the neighborhood if the policies acquired are based on demonstrations.[14] In this section, we will discuss how we achieved this, based on the demonstrations of the eight human participants.

4.1. *Innovative solution to stand up process*

We can compute a new desired joint trajectory as a perturbation of the imitation trajectory $\mathbf{q}_{\text{innovation}} = \mathbf{q}_{\text{imitation}} + \Delta\mathbf{q}$ and maximize the difference between the innovation reward and imitation reward, while fitting the constraints.

The new fitness function maximizes the positive difference between the innovation reward r'_R and imitation reward r_R. It is given by

$$\min f = \sum_{k=1}^{N} e^{-\mu(r'_R(k) - r_R(k))}, \tag{12}$$

$$\theta_{\min} \leq \theta \leq \theta_{\max}, \tag{13}$$

where θ represents ZMP, torque or joint position.

Algorithm 3 presents the innovation behavior.

Algorithm 3. Skill Innovation

Require: Identification of human and robot model
Require: K_p, K_d
Require: $\mathbf{q}_{imitation}$
Require: Compute the imitation reward r_R
while $f \neq 0$ **do**
  New $\Delta q_{m1}, \Delta q_{m2}, \Delta q_{m3}$ using DE
  Compute piecewise polynomial
  Simulate system using K_p, K_d
  Compute ZMP($\mathbf{q}$,$\dot{\mathbf{q}}$,$\ddot{\mathbf{q}}$)
  Compute ID($\mathbf{q}$,$\dot{\mathbf{q}}$,$\ddot{\mathbf{q}}$)
  Compute the reward r'_R
  Compute fitness function f using (12)
end while

4.2. *Imitative and innovative learning*

Imitation and innovation process during learning is a complex process that can be formulated together using a simple variation of (10) and (12).

$$\min f = (1-\rho)\sum_{k=1}^{N-1}(\hat{r}_R(k+1) - r_R(k+1))^2 + \rho\sum_{k=1}^{N}\frac{1}{r_R(k)}, \tag{14}$$

$$\theta_{\min} \leq \theta \leq \theta_{\max}, \tag{15}$$

where θ represents ZMP, torque or joint position.

The first term of (14) represents the imitation part, the second part represents the innovation, tuned by the term $\rho \in (\ 0,\ 1)$. The third term corresponds to the ZMP, torque and joint limits constraints. If we are looking for more innovation, we just need to adjust the term ρ.

5. Model and Behavior Representation in the Reward Domain

In this section, we develop the mathematical tools that allow to transfer the behavior from the human to the robot. We modeled both human and robot as an actuated 3-link kinematic chain. Next, we obtain the ZMP and torque of every link trajectory to compute a common basis of comparison, the reward domain.

5.1. *Kinematic model*

We approximated both the humans and the robot using an actuated 3-link kinematic chain in the sagittal plane, to represent the third scheme of robotic imitation, the mapping between dissimilar bodies,[1] since the human standing up movement occurs

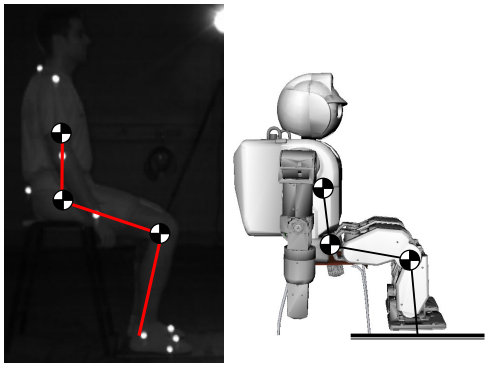


Fig. 2. Snapshot of the high frequency camera of the MOCAP system with a subject seated on a chair and markers on his body. Actuated 3-link kinematic chain is overlaid (left). A simulation of the humanoid HOAP-3 seated on a chair and the 3-link kinematic chain (right).

in that plane without relative movement of legs. It should be noted that this does not account for the role of the swing of hands during standing up.

Figure 2 (left) shows a snapshot of the high frequency camera of the MOCAP system, where a human is seated on a chair with all the markers on his body. An actuated 3-link kinematic chain is overlaid. In Fig. 2 (right), we show the position of the 3-link actuated kinematic chain over the humanoid robot. The COM of each link in the kinematic chain is located at its tip. The first joint of the kinematic chain corresponds to the ankle joint in both human and humanoid, the second joint of the kinematic chain corresponds to the knee and the third one corresponds to the hip.

It is clear that a 3-link kinematic chain does not completely represent the behavior of a humanoid robot in every situation, however the model has some advantages that make it suitable for the task of standing up from a chair. We chose a 3-link kinematic chain with continuous boundary conditions (starting from static torques needed to keep balance soon after lift-off from the chair) to represent a human and robot standing up due to the model suitability for this task. It is a simple model with easy to solve equations, with low computational cost, which is an advantage to use an optimization process like in our case. The task involves relatively low velocity and acceleration profiles. Motor tasks like standing up in healthy adults is a sub-conscious process with minimum cognitive inhibition. Furthermore, the movement is symmetric (legs do not move relative to each other). Therefore, a kinematic chain in the sagittal plane is suitable to represent the movement. In a previous work, we discussed the advantages and limitations of using a reduced model and how the performance can be improved using a robust control technique like Fractional Calculus.[33] It should be noted that, using a more complete model like in the work of Dr. Sentis[34] would reduce real-time feedback control effort, it would not affect the application or results of our method. A complete model would require a more complex formulation with higher computational cost and room for errors due to wrong estimation of an increased number of parameters. But most important of all, it does

not improve the method we propose in any way. It is clear that a 3-link kinematic chain does not represent completely the behavior of a humanoid robot in every situation, however the model has some advantages that make it suitable for the task of standing up form a chair.

Our motion data shows that there is no slip between the foot and the ground in the human demonstrations, therefore we assume that the friction coefficient was high enough for the reaction force vector to stay within the friction cone with no slip. Furthermore, for simplicity we did not model the contact with the chair when the human is seated.

5.1.1. *Human kinematic model and simplifications*

To calculate the masses of the actuated 3-link kinematic chain for the human, we took into account the total weight of the subject and an estimate of the mean distribution of human body parts presented by NASA.[35] The mass of the first link is composed by the mass from the feet to the knee, the mass of the second link is composed by the mass from the knee to the hip and the mass of the third link is composed by the mass from the rest of the body.

The length of the links is estimated using the distance between the markers (see Fig. 10). For the first link, the length is the distance between ankle and knee, for the second one, the distance between knee and hip and for the third one, the distance between the hip and the middle of the chest.

5.1.2. *Robot kinematics identification*

To identify the actuated 3-link kinematic chain parameters of the Fujitsu HOAP-3 humanoid robot, i.e., the length and mass of every link, we used DE[30] and data of the robot sensors. This method is based on the work of Tang.[36]

We manually planned several stand up trajectories for the robot and obtained the ZMP measurement of the force sensors in the feet. Later, we used the ZMP multi-body equation (16) to obtain the theoretical ZMP trajectory in the saggital plane.

$$x_{\mathrm{ZMP}} = \frac{\sum m_i x_i(\ddot{z}_i + g) - \sum m_i \ddot{x}_i z_i - \sum I_{iy}\alpha_{iy}}{\sum m_i(\ddot{z}_i + g)}, \tag{16}$$

where x_{ZMP} is the ZMP in the sagittal plane, m_i is the link mass, x_i and z_i are the positions of the link tip, $\ddot{x}_i$ and $\ddot{z}_i$ are the accelerations, I_{iy} is the inertia, α_{iy} is the angular velocity and g is the gravity acceleration.

To identify the system, we optimized the kinematic chain parameters minimizing the difference between the theoretical ZMP and the real ZMP. The results are shown in Table 1.

The trajectories to identify the system were planned performing a stand up movement starting from seated (Fig. 3). At first, the robot seems unstable because its ZMP is slightly outside the limits. Actually it is not, the robot is seated on a small chair. Since we do not take into account the contacts with this chair, both theoretical

Table 1. Parameter identification of the 3-link kinematic chain for the robot.

	Mass (g)	Length (m)
Link 1	505	0.167
Link 2	500	0.260
Link 3	3900	0.264

and real ZMP are outside the limits. However, the robot at this moment has three contacts, the chair and both feet. When the movement starts, the robot rapidly reaches stability.

5.2. *Equations of motion*

To derive the equation of motion for the 3-link kinematic chain (see Fig. 3), we used the Lagrange theory. m_i is the link mass, l_i is the link length, q_i is the joint position, and τ_i is the joint torque.

$$\frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}_i}\right) - \frac{\partial \mathcal{L}}{\partial q_i} = \tau_i, \tag{17}$$

where $\dot{q}_i$ is the joint velocity and the Lagrangian $\mathcal{L}$ is the difference between kinetic $\mathcal{T}$ and potential energy $\mathcal{V}$.

$$\mathcal{L} = \mathcal{T} - \mathcal{V}. \tag{18}$$

Let us define the potential energy

$$\mathcal{V} = m_1 g z_1 + m_2 g z_2 + m_3 g z_3. \tag{19}$$

Let us define the kinetic energy

$$\mathcal{T} = \frac{1}{2} m_1 v_1^2 + \frac{1}{2} m_2 v_2^2 + \frac{1}{2} m_3 v_3^2, \tag{20}$$

where v_1, v_2 and v_3 are the speed of the centers of mass of the 3-link kinematic chain, $v_i^2 = \dot{x}_i^2 + \dot{z}_i^2$. Substituting (19) and (20) into (18), we obtain the equation of motion

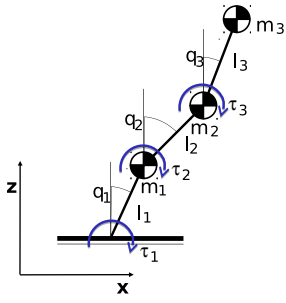


Fig. 3. An actuated 3-link kinematic chain in the sagittal plane.

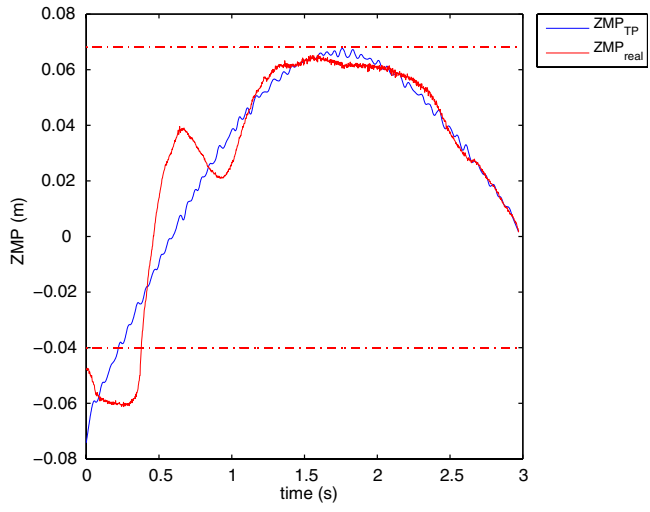


Fig. 4. An example of a theoretical versus real ZMP trajectory used in the parameter identification. The limits of the ZMP are showed in dotted red.

of the 3-link kinematic chain, whose compact form is stated as follows:

$$\tau = \mathbf{H}(\mathbf{q})\ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{q}} + \mathbf{G}(\mathbf{q}), \tag{21}$$

where $\mathbf{H} \in \mathbb{R}^{3\times3}$ is the inertia matrix, $\mathbf{C} \in \mathbb{R}^{3\times3}$ is the matrix of centrifugal and Coriolis forces, $\mathbf{G} \in \mathbb{R}^{3\times1}$ is the gravity matrix, τ is the vector of joint torques, $\mathbf{q}, \dot{\mathbf{q}}$ and $\ddot{\mathbf{q}}$ are the vectors for joint position, velocity and acceleration.

5.3. *State space representation of the 3-link kinematic chain*

The 3-link kinematic chain can be expressed as a dynamical system in the standard form

$$\dot{\mathbf{x}} = A\mathbf{x} + B\mathbf{u}, \tag{22}$$

$$\mathbf{y} = C\mathbf{x}, \tag{23}$$

where $\mathbf{x}$ is the state vector, $\mathbf{u}$ is the control vector and $\mathbf{y}$ is the output vector.

To obtain the representation of the triple pendulum system, let us define the following state variables: $\mathbf{x} = [q_1, \dot{q}_1, q_2, \dot{q}_2, q_3, \dot{q}_3]^T$.

Taking this into account, and reordering Eq. (21), the matrices A, B and C can be obtained knowing that

$$\dot{\mathbf{x}}_1 = \mathbf{x}_2, \quad \dot{\mathbf{x}}_3 = \mathbf{x}_4, \quad \dot{\mathbf{x}}_5 = \mathbf{x}_6, \tag{24}$$

$$\begin{pmatrix} \dot{\mathbf{x}}_2 \\ \dot{\mathbf{x}}_4 \\ \dot{\mathbf{x}}_6 \end{pmatrix} = \hat{f}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6), \tag{25}$$

where $\hat{f}$ contains nonlinear terms of the state variables.

To get rid of the nonlinear terms, we linearized over the point of maximum acceleration, $\mathbf{x}_{i0}$ and $\mathbf{u}_{i0}$, using a Taylor expansion.

$$\dot{\tilde{\mathbf{x}}} = \mathbf{A}\tilde{\mathbf{x}} + \mathbf{B}\tilde{\mathbf{u}}, \tag{26}$$

where

$$A = \left.\frac{\partial f}{\partial \mathbf{x}}\right|_{\substack{\mathbf{x}=\mathbf{x}_0 \\ \mathbf{u}=\mathbf{u}_0}}; \quad B = \left.\frac{\partial f}{\partial \mathbf{u}}\right|_{\substack{\mathbf{x}=\mathbf{x}_0 \\ \mathbf{u}=\mathbf{u}_0}} \tag{27}$$

and $\tilde{\mathbf{x}}_\mathbf{i} = \mathbf{x}_\mathbf{i} - \mathbf{x}_{\mathbf{i0}}$, $\tilde{\mathbf{u}}_\mathbf{i} = \mathbf{u}_\mathbf{i} - \mathbf{u}_{\mathbf{i0}}$.

5.4. *Trajectory generation*

To perform the imitation, we defined a desired joint trajectory for the robot, computed as a cubic spline with one via point with an initial, middle and final point. The initial and final points correspond to the static postures of being seated and being standing up and are known. The middle point is optimized using DE to obtain the imitative and innovative behavior, using (10) and (12), respectively.

It is important to highlight that there is no dynamic control. We use the data obtained via the MOCAP system to compute a stochastic model of the human behavior. The model is transferred to the humanoid robot by computing an optimal trajectory, which imitates the human fitting all humanoid constraints.

The trajectory optimization is computed offline at the moment, because the robot has to learn the average behavior demonstrated by a group of humans. However, DE can use individual demonstrations in a pool of references in an online learning framework.

5.5. *Definition of reward profile*

Through human demonstrations, the humanoid robot learns how to imitate the human performance, taking into account the robot constraints. Furthermore, it is able to improve the imitation, obtaining a better solution than the one demonstrated by the human.

For this purpose, we defined a reward function of stability and effort for all human participants, which are modeled as 3-link kinematic chains. To check the stability, we used the ZMP and to check the effort, the torques of the three joints.

To compute the inverse dynamics, we used (21). To compute the ZMP for the 3-link kinematic chain, we used the equation of multibody ZMP in the sagittal plane (16).

In Fig. 5, the ZMP profile for all 20 demonstrations of one human participant is plotted, whose weight is 68.3 kg and height is 179.6 cm. As it can be observed, at the beginning, the ZMP is outside the limits because we do not model the contact with the chair. The ZMP limits are obtained by measuring the feet size of all subjects with the data provided in the MOCAP system. We took as the stable zone the mean of the

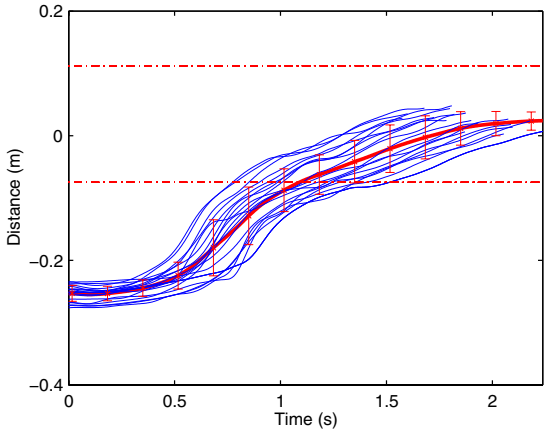


Fig. 5. Actuated 3-link kinematic chain's ZMP profile for the 20 demonstrations of one of the subjects standing up. The mean and standard deviation are in red. The ZMP limits are in dotted red.

feet measurements in every demonstration. As it can be seen, not all trajectories have the same duration, since not all demonstrations are equal. To solve this problem, we took the slowest movement as the basis and stretch the other trajectories as if the human is still.

Figure 6 shows the joint torques for the same participant. Since we cannot measure the maximum torque that the muscles support, for simplicity, we took the maximum torque of the 20 demonstrations as the maximum value. This value will be used in the definition of the reward function.

5.6. *Selection of the reward function*

We used two different functions to evaluate the human behavior in the reward space: a polynomial and a Gaussian-like function. Every function is used to obtain the ZMP reward profile and the torque reward profile. θ_j represents the ZMP minimum, medium and maximum in the case of the ZMP reward function and similarly with the torque reward function. The torque reward function is the normalized mean of the 3-link kinematic chain's joint torques.

The polynomial reward function has order four and is centered in θ_{med}. Values outside the limits have zero reward. It is given by

$$f_1(x) = ax^4 + bx^3 + cx^2 + dx + e, \tag{28}$$

where x can be ZMP or joint torque trajectory and the coefficients $a, b, c, d, e,$ are obtained solving

$$f(\theta_{\min}) = 0; \quad f(\theta_{\max}) = 0; \quad f(\theta_{\text{med}}) = 1; \quad f(\theta_{\text{med}}/2) = 0.8; \quad f(3\theta_{\text{med}}/4) = 0.8. \tag{29}$$

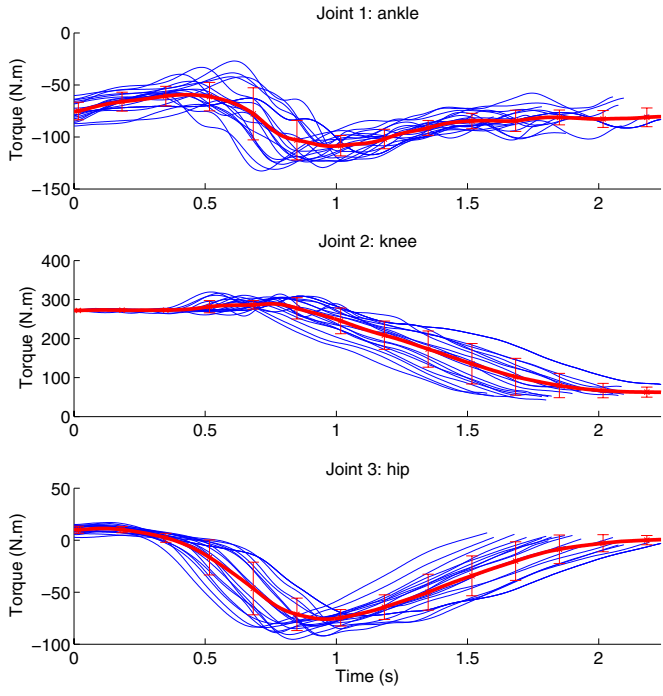


Fig. 6. Actuated 3-link kinematic chain's torques of the 20 demonstrations of one of the subjects standing up. The first joint of the 3-link kinematic chain corresponds to the human ankle, the second one to the knee and the third one to the hip. The mean and standard deviation are in red.

The Gaussian-like function follows the next equation:

$$f_2(x) = \exp \frac{-36(x - \theta_{\text{med}})^2}{2(\theta_{\max} - \theta_{\min})^2}. \tag{30}$$

These functions allow the mapping from the ZMP or torque space to the reward space (Fig. 7).

The total reward profile for the human is the sum of stability and effort functions

$$r_H(t) = \frac{w_{\text{ZMP}}(t) r_{\text{ZMP}}(t) + w_\tau(t) r_\tau(t)}{2}, \tag{31}$$

where w_{ZMP} and w_τ are weights of ZMP and torque respectively, that can vary from 0 to 1, r_{ZMP} is the reward of the zmp and r_τ is the reward of the torque, which is the sum of the reward of every joint torque divided by three.

Figure 8 shows the mean reward of the 20 demonstrations of every participant standing up (in blue). The mean of all 160 demonstrations and the standard deviation are plotted in red. For this profile, we chose the following weights:

$$w_{\text{zmp}}(t) = at^3 + bt^2 + c, \tag{32}$$

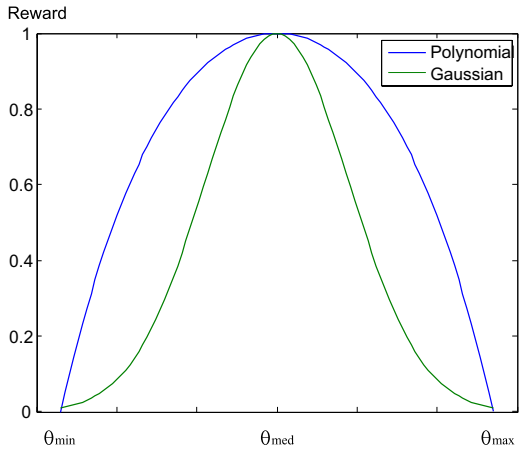


Fig. 7. Two reward functions. The blue one is a polynomial function of 4th degree, the green one is a Gaussian-like function whose maximum is one. The parameter θ represent the ZMP or the torque of the actuated 3-link kinematic chain.

$$a = \frac{-2(\phi_1 - \phi_0)}{T^3}; \quad b = \frac{3(\phi_1 - \phi_0)}{T^2}; \quad c = \phi_0, \tag{33}$$

where $\phi_1 = 1$, $\phi_0 = 0$ and $T = t_{\max}$

$$w_\tau(t) = 1. \tag{34}$$

Equations (32) and (33) represent a third order polynomial, that starts in 0 and finishes in 1, which means that at the beginning of the stand up motion, we do not care if the ZMP is outside the limits, but we care about the torques (34).

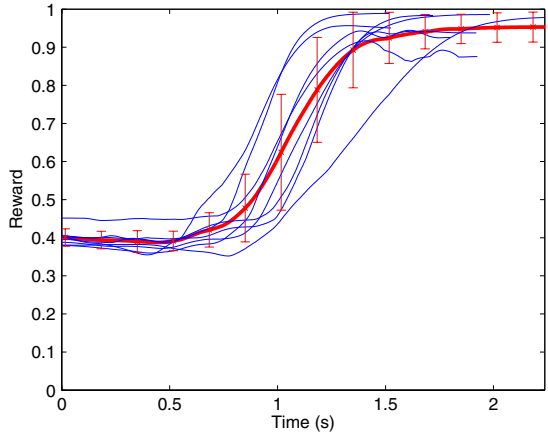


Fig. 8. The mean reward of the 20 demonstrations of each human participant is in blue. The mean and standard deviation of all demonstrations of eight human participants are in red.

Analyzing Fig. 8, the predominantly subconscious operation of motor programs to execute standing up show some stereotypical pattern across all subjects irrespective of their variability in terms of weight and height. A detailed discussion of this can be found in Appendix A.

5.7. *Generalization and discussion of the reward profile*

The shape of the reward functions are selected in accordance to the task. The final posture or goal posture of standing up has the feature of being stable and of minimum effort, as we showed in a previous work.[37] In that case, the ZMP is almost in the middle of the feet and the torque is minimum. Therefore, the reward function is selected to be an attractor to these conditions. That is why the middle point of the reward function is the mean ZMP in the case of stability function and zero in the case of effort function (Fig. 7).

The selection of a suitable reward function has been discussed extensively. Sometimes, it is the observer who will manually set the reward value[2]; it can be defined as a mathematical function that maps from states and actions to rewards[17,38,39,29] or the reward function can be learned from the demonstration set, what is called inverse RL.[40]

We used stability as a criterion because this task involves moving from a statically stable posture (seated) to an unstable fixed point (upright posture) through a process of dynamic stabilizing. This can be achieved by humanoid robots with regulator type feedback control, that leads to high peak torques at the start of the movement. In our experiments, we show that humans do not use such a regulator type feedback control. In contrast, humans use an optimum strategy in terms of effort minimization. Therefore, it is more meaningful to combine the intention to maintain stability while minimizing effort in a learning based on demonstration framework.

Our method needs to define a reward function for each task. In a complete different task, as for example opening a door, the reward function will have to account for a complete different shape. This has all the sense since the goal is completely different. In the case of opening the door, the goal is grabbing the knob successfully and pulling the door until it is open, then the reward function has to be selected to take that into account. The goal and reward function are completely different in the case of standing up from a chair where it is important to maintain stability and minimize the effort.

6. Experimental Results

We used the humanoid HOAP-3 to show the robustness of our method and present the experimental results.

6.1. *Experimental setup*

We collected data from eight human participants of age between 20 to 40 years, weights between 60 and 99 kg and heights between 1.68 and 1.88 m. For this task, we

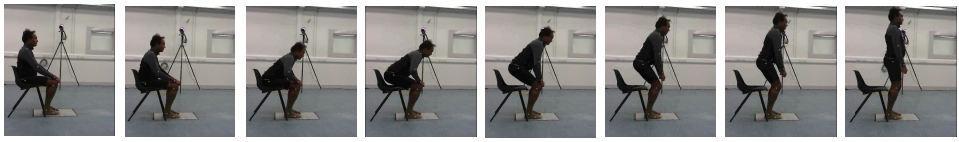


Fig. 9. Snapshots of one of the human participants standing up from a chair in the MOCAP environment.

recruited healthy adult participants with no known history of motor dysfunction. The experimental protocol was approved by the ethics committee on using human participants in experiments of Kingston University of London. The height and weight of all human participants are presented in Appendix A.

Every participant performed 20 consecutive demonstrations of standing up from a chair. There were no special training for the participants, since it is a simple task, only a few instructions like do not stand up very fast or put your feet straight. A 6-camera Oqus motion capturing system made by Qualisys, Sweden, collected position data of 21 markers attached to the subject's body at 240 Hz sampling rate.

In Fig. 9, the experimental procedure is shown. The participant is seated on a chair and performs the movement of standing up.

The markers were distributed as follows: first and fifth metatarsi, lateral malleolus (ankle), lateral epicondyle of the femur (knee), greater trochanter (hip), anterior superior iliac spine (ASIS), posterior superior iliac spine (PSIS), seventh cervical vertebra (top of spine), acromion process (shoulder), lateral epicondyle of the humerus (elbow) and lateral styloid process (wrist). All markers are bilateral, they were located on both sides of the body, except the seventh cervical vertebra. In Fig. 10, the position of the markers is shown.

6.2. *Extraction of human behavior*

After computing the reward function for every trial of every human participant, and assuming that the reward fits the Markov property, we computed the RTPM.[41] This matrix summarizes, in one singular metric, the behavior of several human participants doing the action of standing up from a chair. Its computation is presented in Algorithm 1.

This matrix changes with the reward function we select. Therefore, we computed several RTPM depending on the reward function selected, if it is the polynomial or the Gaussian, and depending of the weights selected in Eq. (31).

Figure 11(a) represents the RTPM using the polynomial function and the weights (32) and (34). Figure 11(b) represents the RTPM using the Gaussian-like function (30) and $w_{\text{zmp}} = w_{\tau} = 1$. These matrices, Fig. 11, represent the behavior of the human standing up taking into consideration the stability and the torques, and of course, it strongly depends on the selection of the reward function.

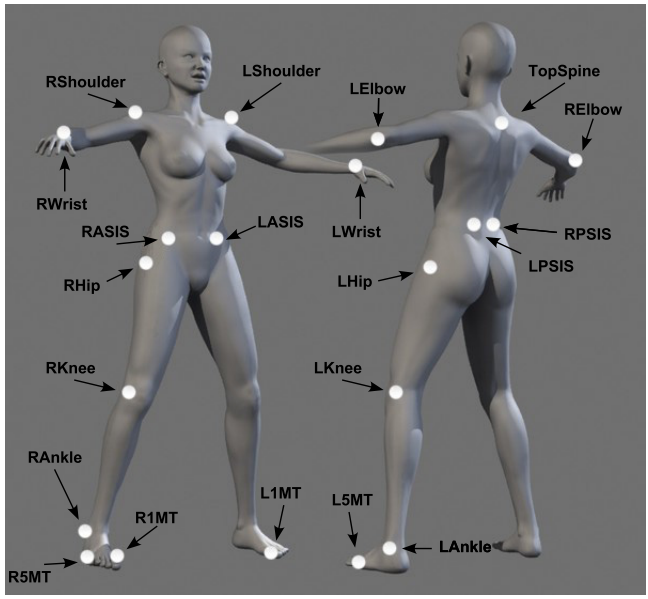
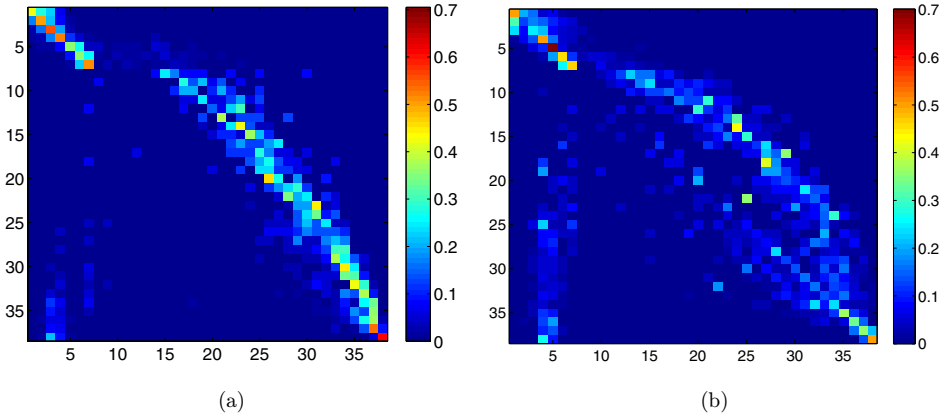


Fig. 10. Distribution of the 21 markers in a human body. R stands for right and L stands for left. 1 and 5MT stands for first and fifth metatarsi, respectively.

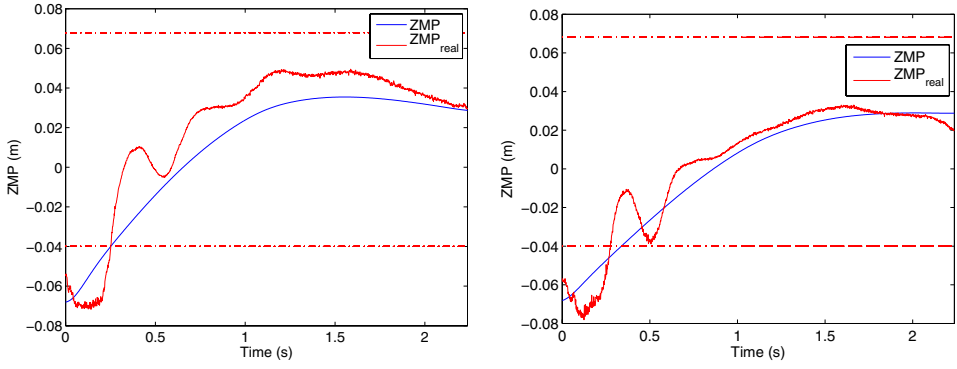


(a) (b)

Fig. 11. (a) Normalized RTPM for all human participants using the polynomial function (28). (b) Normalized RTPM for all human participants using the Gaussian-like function (30).

6.3. *Stand up experiments*

Figures 12(a) and 12(b) shows the theoretical ZMP, calculated using (16) and the ZMP measured from the robot sensors for both imitation and innovation. This measurement is the mean of the ZMP trajectory of both feet. As it can be seen, initially the ZMP is outside the stability region. This happens because at that time

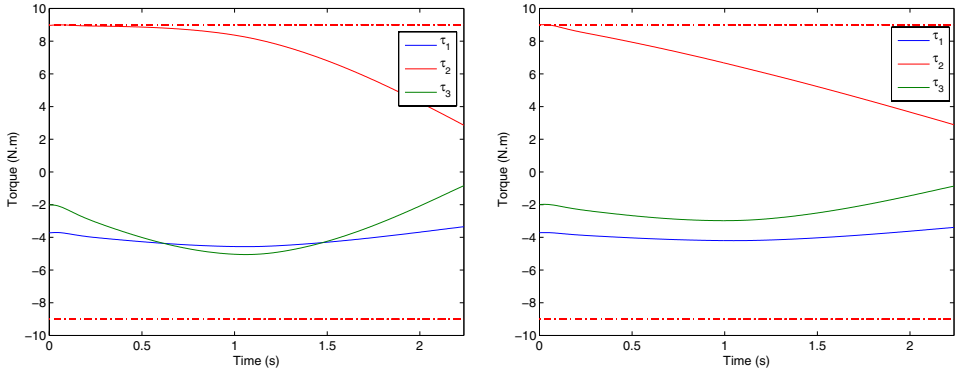


(a) (b)

Fig. 12. Computed ZMP of the actuated 3-link kinematic chain approximation and that for the real robot for the imitation behavior (a) and for the innovation behavior (b).

the robot is slightly leaned on the chair. The ZMP in the innovation motion goes straighter to the middle, which is translated in a higher reward. The imitation ZMP profile also stays near the middle value, but not as much as the innovation profile. The explanation is simple, in the case of the imitation, the solution minimizes the difference between the predicted reward if the robot behaves like a human and the actual reward, instead, in the case of the innovation, the solution maximizes the reward, always fitting the constraints. Furthermore, as it can be noted in Fig. 5, the ZMP of the human and those of the robot is not the same, which is obvious as their sizes and kinematic structure are different.

Figures 13(a) and 13(b) plots the 3-link kinematic chain torques. As it can be seen, they are between the limits. It is remarkable that the second joint has the higher value, because it supports the heaviest part of the robot. Again, if we analyze



(a) (b)

Fig. 13. 3-link kinematic chain torques in the imitation trajectory (a) and in the skill innovation trajectory (b).

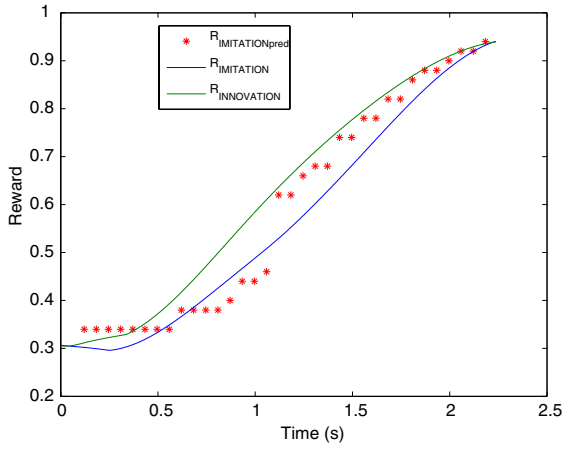


Fig. 14. Reward profiles for imitation and innovation robot behaviors. The imitation reward in blue, the predicted reward imitating a human in red and the innovation reward in green.

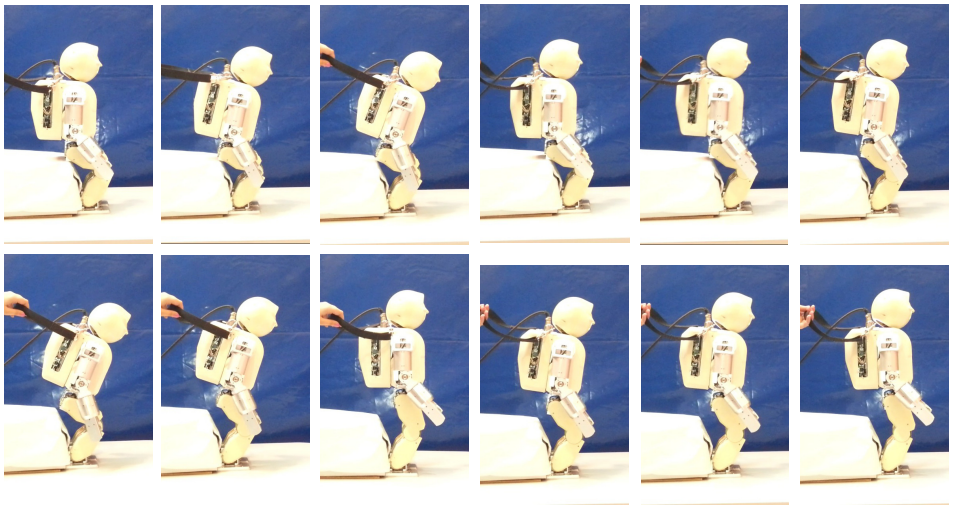
the imitation and innovation torque profiles, we observe that in the imitation movement, the knee joint stays near the limit almost until the second one. However, in the innovation movement, the reward is higher, and the torque decrease faster to a comfortable posture.

In the movement of standing up to an upright posture, the torque limits play an important role. They define the initial posture. It is the same when a human stands up. If the torque that our legs have to create is too much high, we help ourselves with our hands, finding another contact or a different stand up strategy. Therefore, our method as we presented it, can cope with postural movements starting from a safe and feasible initial posture.

In Fig. 14, the computed reward profiles for imitation and innovation behavior are plotted. Blue line represents the imitation reward, and red dots represent the predicted reward if the robot behaves like a human and the green line represents the innovation reward. Comparing the robot reward with the human participants reward in Fig. 8, we observe that they are very similar, since the predicted robot reward is related to human reward. However, it is not the same, because the variability of the different demonstrator performances is encoded in the RTPM, which is the key element to transfer a behavior.

The results presented in this section were obtained using the polynomial function that maps from ZMP and torque space to reward space. The results using the Gaussian-like function (30) were not showed here, for reasons of space and that they would be very similar to the results of the polynomial function.

In the imitation approach (see snapshots in Fig. 15(a)), the kinematic chain lean forward producing a movement very similar to that of the human demonstrations. In that case, the optimizer minimize the difference between the actual reward and the predicted reward if the robot behaves like a human.



(a) (b)

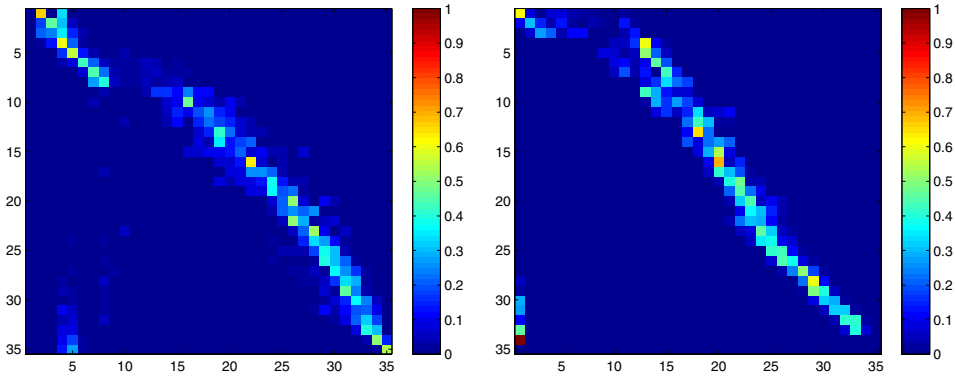
Fig. 15. (a) Snapshots of the robot standing up in the imitation process. (b) Snapshots of the robot standing up in the innovation process. In the imitation process, the robot lean forward too much, trying to follow the strategy of the human. However, in the innovation process the robot stands up more straightly, since it is maximizing its reward. This behavior is logical because the feet size in the case of the robot is larger in relation with the feet size of the human. Therefore, the robot does not really have to lean forward so much.

In the innovation approach (see snapshots in Fig. 15(b)), we obtained a new reward which is greater than the imitation reward, fitting all the constraints. In that case, the movement of the 3-link kinematic chain is straighter, and it is more adequate to the robot structure. The ratio between the sole of human's foot and human's height is around 0.14. The ratio for the robot is 0.18. Then, the robot's feet are greater in relation with its height than the human's feet. As the robot has a wider surface, its ZMP is wider, in relation with its height. Therefore, the robot does not need to lean forward so much as when it is imitating the human performance, instead, it can go straighter, obtaining a better reward for the movement.

6.4. *Hypothesis testing and generality*

As it can be seen in Fig. 15, the robot is not seated with the second joint at 90° as a human would do. Due to the torque limitation in the robot motors, it is impossible to generate a trajectory with that initial posture, robot servos could be damaged. The initial posture is selected to obey the maximum torque limits. This phenomenon is equal in humans. It is common to use our hands to help us to stand up if our legs cannot generate enough torque. Furthermore, we use our hands to generate an easier and safer movement.

Furthermore, the surface of contact between the robot and the chair is smaller than in the case of a human seated. Again, this is due to the robot morphology which is different to the human's in structure, not only in height and weight.



(a) (b)

Fig. 16. (a) Normalized transition matrix of the reward for the human using the polynomial function (28). (b) Normalized transition matrix of the reward for the robot using the polynomial function (28) and the experimental solutions.

To prove the generality of our method, we generate up to 35 experimental trajectories of the robots standing up behavior. This trajectories have different initial and final postures. Our method allows to robustly transit from a seated posture to a stand up posture. The initial posture is selected to not surpass the maximum torque and the final posture is stable.

We used all these trajectories to compute the RTPM for the robot as shown in Fig. 16. This matrix represents the real behavior of the humanoid when it imitates the repertoire of human demonstrations. The human and the robot are morphologically similar though the exact scales are different. Therefore, we hypothesized that their stand up strategies should be similar. In order to test this, we can compare the human RTPM (Fig. 16(a)) with the robot RTPM (Fig. 16(b)) as similar strategies should result in similar probability transitions in the reward space (see (7)).

To compare the matrices, we compute the mean square error, e, given by

$$e = \sqrt{\left(T_{\text{human}}(i,j) - T_{\text{robot}}(i,j)\right)^2} = 0.0395, \tag{35}$$

and then obtain the average probability error of a cell, P_e, given by

$$P_e = eP(s = s_i) = 0.1128\%, \tag{36}$$

where $P(s = s_i)$ is the probability of staying in state s_i, which in the case of this RTPM is $1/35$. For a more detailed discussion of the human demonstration consistency, please refer to Appendix A.

7. Discussion

In this paper, we presented an original method to obtain imitative and innovative postural behaviors in a humanoid robot through human demonstrations.

We collected data from a group of eight human participants standing up from a chair. We modeled both human and humanoid using an actuated 3-link kinematic chain approximation and computed a reward profile in terms of ZMP and inverse dynamics. We used 20 demonstrations each from eight participants to obtain the Markov probability transition matrix of the compound reward for the human demonstrations.

Provided the humanoid robot should follow the same optimality criteria and profile as the human if it were to imitate the human in a qualitative sense, we can use the Markov chain obtained for human demonstrations to predict the future humanoid reward starting from any state of the humanoid robot. We then optimized a joint trajectory to obtain imitation, where the robot reward is equal to the predicted human-like reward along the whole posture control period. Having achieved imitation, we proceeded to achieve robotic skill innovation where the average reward profile of the humanoid is higher than that of the average human demonstrations.

The approach discussed in this paper emphasizes the fact that intelligent behavior of an embodied agent is in the eyes of the observer.[42] Therefore, different observers can use different criteria to compare two embodied agents trying to achieve a given goal. Here, we propose that the observer can compare a behavior enacted by two different embodiments in a common reward space. This paper considers the case where one multimodal reward function is used throughout the standing up behavior. However, it should be noted that there can exist state dependent heterogeneous reward functions in more complex cases. An example is to consider acceleration and joint torque optimization at the start and ZMP variability minimization in the neighborhood of the standing posture. Well established techniques of GMM can be a good technique to model such reward landscapes.

The developed algorithm produces a dynamic posture (standing up), which is the transition between the static posture of being seated and the static posture of being standing up. Both initial and final static postures are calculated in advance.

The main features of our method are discussed here:

- Our method allows to transfer a stand up behavior from a human teacher to a robot learner, even if there is a wide mismatch in their kinematic structures.
- The robot learns to perform smooth and stable standing up movements based on human demonstrations.
- The robot does not simply imitate the human movement, rather learns an optimal behavior subject to a set of internal constraints.
- It takes into account the ZMP, torque, and joint limits of the robot, so the trajectory is always executable.
- We defined a multi-objective reward profile of ZMP and joint torques and encoded the demonstrating trials of the human in a RTPM.
- Based on neuroscientific theories that suggest that human skill transfer is achieved by imitating the goal of the action, we suppose that the reward transition probabilities of the robot show the same structure as that in the human demonstrations.

Thus, we computed a constrained policy that minimizes the predicted error in the reward profile.[11]

- After the imitation is achieved so that the robots RTPM is statistically significantly equal to that of the human demonstrations, we moved on to find a new policy that improves the robot reward profile leading to skill innovation.

7.1. *Key contributions*

The key contributions of this paper are summarized as follows.

We have presented a new skill transfer method of stand up behavior from human demonstrators to humanoid robots, that involves comparing temporal transition in a common multi-objective reward landscape. The main advantage is that we could accommodate the behavior even if the human and the robot have a mismatch in their kinematic structures, weights and heights. The transfer is obtained using a Markov transition matrix that summarizes the state transition probabilities in the reward space. This generic method can be extended to other movements like sitting down, crouching or grasping an object subject to a set of robot constraints.

We achieve imitation learning by finding a policy that minimizes the error between the predicted robot reward profile, if it behaves like a human, and the actual reward profile. The consequence is a trajectory that fits stability, torque and joint limit constraints while producing a stand up movement that imitates the human behavior.

Finally, we refine the robot behavior by maximizing the positive difference between the new reward and the imitation reward, producing skill innovation that is translated in a more suitable behavior of the humanoid robot.

7.2. *Comparison with related work*

We perform the same task as Mistry but in a completely different way.[43] In their work, a full-size humanoid robot stand up from a chair using different strategies, imitating a young and an elder person. Their approach is based on mimicking, they adapt the human trajectories to the robot structure. By contrast, our approach is more general. We are able to transfer the stand up behavior to a robot much smaller than the human with a kinematic structure, weight and height completely different.

Our work is somewhat similar to Billard that used HMM to recognize and generate motion patterns.[44] They address the question of what to imitate and how to imitate. First, they encode the demonstration in a HMM that are treated with Bayesian information criterion (BIC) to optimize the number of model states. They defined a metric in the form of a cost function to evaluate the robot's performance. Finally, they optimized the reproduction of the task in another context. The key differences between Billard and our work are that they used kinesthetic information, instead of transferring the behavior from several humans to a small robot. Moreover, the behavior is more complex in our case, since it has the problem of handling stability. However, the clearest difference is the selection of the cost function. They

used a cost function that takes into account the joint trajectories, which is not a generic method to transfer the skill to a different robot. On the contrary, we construct our reward function taking into account the stability and the effort, so that the robot and the human have different joint trajectories in their successful standing up behaviors.

An interesting approach that have synergies with our work is the concept of goal oriented behavior understanding.[1,17,45] This field studies the recognition and posterior imitation of other agent's behavior. Aksoy present a method of understanding a manipulation behavior using graphs.[46] Similar to us, they define a transition matrix of semantic events that allows to understand a behavior and reproduce it under different conditions. Takahashi presents a multi-agent behavior imitation procedure based on RL.[38] Their method can be divided into two phases. First, they recognize an observed behavior through the estimation of the state and action reward, encoding it as a state value function. Afterwards, the imitator develops a similar behavior optimizing a reward function which is a weighted combination of the imitator reward and the teacher reward. This work is similar to ours in the sense that they used a reward profile as a basis of behavior comparison.

Argall presents a combination of LfD and teacher advice is used to improve the policy in the continuous space.[47] Similarly, Bentivegna makes a humanoid robot to learn from observation a set of task that using a library of manually predefined primitives.[48] The performance of the robot is improved through practice based on observations of the teacher's outcomes.

A number of other approaches use a framework based on ask for help to speed up and enhance learning. Here, an agent request advice for other similar agents which are combined with information of the environment.[49,50]

Our work is partially inspired by a framework to perform imitation by solving the correspondence problem.[2,51,52] They define three metrics for imitation: *end-point level*, where it is only considered the overall goal, *trajectory level*, where the imitator considers a set of subgoals that are sequentially reached, and finally *path level*, where the imitator tries to replicate the teacher's trajectory as closely as possible. Trajectory level and path level metric is similar to program level and action level of Byrne.[16] The method for imitation we present in this paper is based on the trajectory level of Alissandrakis.[2] Instead of using a set of sub-states as the metric of the imitation, we used the reward of the state, which is our basis of evaluation. Furthermore, our method not only imitates but innovates new behaviors, which are evaluated producing an improvement of the demonstrator behavior.

Another solution to the problem of what to imitate is presented by Billard.[17] Similar to our approach, they establish a metric to evaluate the performance of the imitation process, paying attention to the manipulation task of writing the letters A, B, C and D. This metric is divided into three levels of imitation of Alissandrakis and a mimic metric which reproduces the exact trajectory of the robot. They optimized the robot control signal to minimize this four metrics which are expressed as costs functions.[2]

A recent approach addresses how to obtain a model of the locomotion behavior that can be transferred form a human demonstrator to a robot, which is called *inverse optimal control*.[53] The authors select an objective function which is a linear combination of position, velocity and other features of the movement as the metric. Parameters for linear combination are obtained thought optimization. This model can be transferred to the robot to produce a similar behavior. The differences with our approach are: First, the selection of the metric to optimize, that in our case is a combination of a reward function of stability and effort, which is more intuitive, and second, they used the model of the human to compute the humanoid locomotion, which produce a similar movement, whereas in our case, we use two 3-link kinematic chain models of different dimensions which perform drastically different trajectories to render in similar optimality of standing up behaviors.

In our previous works, a policy improvement method is used over a large number of machine operators to improve their expertise and enhance their skills in a global way.[41,39] We demonstrated that individuals innovate better skills while mixing their behavior with that of an elite individual, producing new elite members with better skills. Furthermore, this paper is an improvement of another previous work, where we compared the behavior of a human and a robot in a common reward space for the same task, standing up from a chair.[37]

7.3. *Future directions*

Several open lines will be addressed in the future. The first one is to try our method in a more complex model of human and robot. The second one is, instead of defining a reward profile, learn the optimal reward profile of the human through, for example, inverse RL.[40]

Finally, we will extend our method to more complex behaviors like opening a door and walk to leave a room. In that case, the reward function will be completely different as the goal is completely different.

Acknowledgments

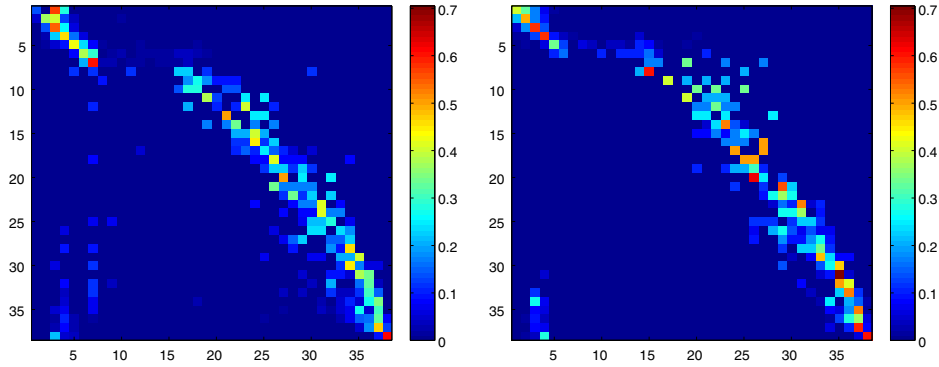
The research leading to these results has received funding from the ARCADIA project DPI2010-21047-C02-01, funded by CICYT project grant on behalf of Spanish Ministry of Economy and Competitiveness and from the RoboCity2030-II-CM project (S2009/DPI-1559), funded by Programas de Actividades I+D en la Comunidad de Madrid and co-funded by Structural Funds of the EU.

Appendix A. Consistency of Human Demonstrations

Our participants varied in physical characteristics in terms of their weight, height and limb kinematics (see Table 2). We wanted to test the consistency of the demonstrations and therefore the reward profile among different groups of humans. For that purpose, we divided the demonstrators in two groups. Group number 1 is

Table 2. Weight and height of the human participants.

	Weight (Kg)	Height (m)
Participant 1	68.3	1.79
Participant 2	60.9	1.68
Participant 3	78.1	1.82
Participant 4	75.2	1.75
Participant 5	83.4	1.84
Participant 6	99.0	1.88
Participant 7	81.0	1.85
Participant 8	67.8	1.71



(a) (b)

Fig. 17. (a) Normalized RTPM for the first group of human participants using the polynomial function (28). (b) Normalized RTPM for the second group of human participants using the polynomial function (28).

composed of people with greater height and weight, they are the participants number 3, 5, 6 and 7. The group number 2 is composed of people with lower height and low weight, they are the participants number 1, 2, 4 and 8.

We computed the RTPM for both groups obtaining Fig. 17. To see the difference between them, we compute the mean square error (35) which is $e = 0.0264$ and the average probability error of a cell (36) which is $P_e = 0.0694\%$.

Observing the results, we can conclude that there is no significant difference in the reward profiles between the two groups. Therefore, there seems to be a reward profile that is independent of the human body size and can define the behavior of standing up. Furthermore, this is equivalent to say that all humans, no matter what the size and weight share the same strategy to accomplish a task, which can be defined as a reward profile and transmitted to a robot. Our results are in accordance to those saying that what should be imitated is the goal of the action, not just the movements.[3,11,12,14]

References

1. S. Schaal, A. Ijspeert and A. Billard, Computational approaches to motor learning by imitation, *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **358**(1431) (2003) 537–547.
2. A. Alissandrakis, C. L. Nehaniv and K. Dautenhahn, Imitation with Alice: Learning to imitate corresponding actions across dissimilar embodiments, *IEEE Trans. Syst. Man Cybern. A, Syst. Humans* **32**(4) (2002) 482–496.
3. D. E. Thompson and J. Russell, The ghost condition: Imitation versus emulation in young children's observational learning, *Develop. Psychol.* **40**(5) (2004) 882.
4. A. N. Meltzoff, The human infant as homo imitans. *Social Learning: Psychological and Biological Perspectives* (Psychology Press, 1988), pp. 319–341.
5. A. Whiten, V. Horner, C. A. Litchfield and S. Marshall-Pescini, How do apes ape? *Learn. Behav.* **32**(1) (2004) 36–52.
6. L. M. Hopper, S. P. Lambeth, S. J. Schapiro and A. Whiten, Observational learning in chimpanzees and children studied through ghost conditions, *Proc. R. Soc. B, Biol. Sci.* **275**(1636) (2008) 835–840.
7. C. M. Heyes, E. Jaldow and G. R. Dawson, Imitation in rats: Conditions of occurrence in a bidirectional control procedure, *Learn. Motivation* **25**(3) (1994) 276–287.
8. N. H. Nguyen, E. D. Klein and T. R. Zentall, Imitation of a two-action sequence by pigeons, *Psychon. Bull. Rev.* **12**(3) (2005) 514–518.
9. A. Whiten, D. M. Custance, J. C. Gomez, P. Teixidor and K. A. Bard, Imitative learning of artificial fruit processing in children (homo sapiens) and chimpanzees (pan troglodytes), *J. Comparative Psychol.* **110**(1) (1996) 3–14.
10. J. Piaget, *Play, Dreams and Imitation*, Vol. 24 (Norton, New York, 1962).
11. G. Metta, G. Sandini, L. Natale, L. Craighero and L. Fadiga, Understanding mirror neurons: A bio-robotic approach, *Interact. Stud.* **7**(2) (2006) 197–232.
12. L. Craighero, G. Metta, G. Sandini and L. Fadiga, The mirror-neurons system: Data and models, *Progr. Brain Res.* **164** (2007) 39–59.
13. R. W. Mitchell, A comparative-developmental approach to understanding imitation, *Perspect. Ethol.* **7** (1987) 183–215.
14. A. Whiten, N. McGuigan, S. Marshall-Pescini and L. M. Hopper, Emulation, imitation, over-imitation and the scope of culture for child and chimpanzee, *Philos. Trans. R. Soc. B, Biol. Sci.* **364**(1528) (2009) 2417–2428.
15. B. D. Argall, S. Chernova, M. Veloso and B. Browning, A survey of robot learning from demonstration, *Robot. Auton. Syst.* **57**(5) (2009) 469–483.
16. R. W. Byrne, Imitation as behaviour parsing, *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **358**(1431) (2003) 529–536.
17. A. Billard, Y. Epars, S. Calinon, S. Schaal and G. Cheng, Discovering optimal imitation strategies, *Robot. Auton. Syst.* **47**(2) (2004) 69–77.
18. J. Peters and S. Schaal, Reinforcement learning of motor skills with policy gradients, *Neural Netw.* **21**(4) (2008) 682–697.
19. M. Jeannerod, *The Neural and Behavioural Organization of Goal-Directed Movements* (Clarendon Press/Oxford University Press, 1988).
20. S. Schaal, Is imitation learning the route to humanoid robots? *Trends Cogn. Sci.* **3**(6) (1999) 233–242.
21. S. M. Khansari-Zadeh and A. Billard. Learning stable nonlinear dynamical systems with Gaussian mixture models, *IEEE Trans. Robot.* **27**(5) (2011) 943–957.
22. N. Naksuk, C. S. G. Lee and S. Rietdyk, Whole-body human-to-humanoid motion transfer, *5th IEEE-RAS Int. Conf. Humanoid Robots* (2005), pp. 104–109.

23. D. Kulić, W. Takano and Y. Nakamura, Incremental learning, clustering and hierarchy formation of whole body motion patterns using adaptive hidden Markov chains, *Int. J. Robot. Res.* **27**(7) (2008) 761–784.
24. W. Ilg, G. H. Bakir, J. Mezger and M. A. Giese, On the representation, learning and transfer of spatio-temporal movement characteristics, *Int. J. Human. Robot.* **1**(04) (2004) 613–636.
25. H. I. Lin and C. S. G. Lee, Self-organizing skill synthesis, *IEEE/RSJ Int. Conf. Intelligent Robots and Systems* (2008), pp. 828–833.
26. R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, Vol. 1 (Cambridge University Press, 1998).
27. J. Peters and S. Schaal, Learning to control in operational space, *Int. J. Robot. Res.* **27**(2) (2008) 197–212.
28. D. Barrios-Aranibar, L. M. G. Gonçalves and P. J. Alsina, Learning by experience and by imitation in multi-robot systems, *International Journal of Advanced Robotic Systems (Org.), Livro: Frontiers in Evolutionary Robotics* (Viena, Aleksandar Lazinica, 2008).
29. F. Guenter, M. Hersch, S. Calinon and A. Billard, Reinforcement learning for imitating constrained reaching movements, *Adv. Robot.* **21**(13) (2007) 1521–1544.
30. R. Storn and K. Price, Differential evolution — A simple and efficient heuristic for global optimization over continuous spaces, *J. Global Opt.* **11**(4) (1997) 341–359.
31. J. Medhi, *Stochastic Processes*, 3rd edn. (New Age International, 2010).
32. C. Shumway, Motor control: Theory and practical applications, *Recherche* **67**(02) (2000).
33. M. Gonzlez-Fierro, C. A. Monje and C. Balaguer, Robust control of a reduced humanoid robot model using genetic algorithms and fractional calculus, *Mathematical Methods in Engineering Int. Conf.* (2013), pp. 183–194.
34. L. Sentis and O. Khatib, Synthesis of whole-body behaviors through hierarchical control of behavioral primitives, *Int. J. Human. Robot.* **2**(4) (2005) 505–518.
35. NASA, Man-systems integration standards, Technical report, National Aeronautics and Space Administration (1995).
36. H. Tang, S. Xue and C. Fan, Differential evolution strategy for structural system identification, *Comput. Struct.* **86**(21–22) (2008) 2004–2012.
37. M. Gonzlez-Fierro, C. Balaguer, N. Swann and T. Nanayakkara, A humanoid robot standing up through learning from demonstration using a multimodal reward function, *IEEE-RAS Int. Conf. Humanoid Robots* (2013).
38. Y. Takahashi, Y. Tamura, M. Asada and M. Negrello, Emulation and behavior understanding through shared values, *Robot. Auton. Syst.* **58**(7) (2010) 855–865.
39. T. Nanayakkara, C. Piyathilaka, A. Subasingha and M. Jamshidi, Development of advanced motor skills in a group of humans through an elitist visual feedback mechanism, *IEEE Int. Conf. Systems of Systems Engineering* (San Antonio, 2007).
40. P. Abbeel and A. Y. Ng, Apprenticeship learning via inverse reinforcement learning, in *Proc. Twenty-First Int. Conf. Machine Learning*, p. 1 (ACM, 2004).
41. T. Nanayakkara, F. Sahin and M. Jamshidi, *Intelligent Control Systems with an Introduction to System of Systems Engineering* (CRC Press, 2009).
42. R. A. Brooks, Intelligence without representation, *Artif. Intell.* **47**(1) (1991) 139–159.
43. M. Mistry, A. Murai, K. Yamane and J. Hodgins, Sit-to-stand task on a humanoid robot from human demonstration, *10th IEEE-RAS Int. Conf. Humanoid Robots (Humanoids).* (2010), pp. 218–223.
44. A. G. Billard, S. Calinon and F. Guenter, Discriminative and adaptive imitation in uni-manual and bi-manual tasks, *Robot. Auton. Syst.* **54**(5) (2006) 370–384.

45. L. Jamone, L. Natale, F. Nori, G. Metta and G. Sandini, Autonomous online learning of reaching behavior in a humanoid robot, *Int. J. Human. Robot.* **9**(03) (2012).
46. E. E. Aksoy, A. Abramov, J. Dörr, K. Ning, B. Dellen and F. Wörgötter, Learning the semantics of object–action relations by observation, *Int. J. Robot. Res.* **30**(10) (2011) 1229–1249.
47. B. D. Argall, B. Browning and M. Veloso, Learning robot motion control with demonstration and advice-operators, *IEEE/RSJ Int. Conf. Intelligent Robots and Systems, IEEE/RSJ International Conference* (2008), pp. 399–404.
48. D. C. Bentivegna, C. G. Atkeson, A. Ude and G. Cheng, Learning to act from observation and practice, *Int. J. Human. Robot.* **1**(04) (2004) 585–611.
49. E. Oliveira and L. Nunes, Learning by exchanging advice, *Stud. Fuzziness Soft Comput.* **162** (2004) 279–314.
50. A. Alissandrakis, C. L. Nehaniv and K. Dautenhahn, Towards robot cultures? Learning to imitate in a robotic arm test-bed with dissimilarly embodied agents, *Interact. Stud.* **5**(1) (2004) 3–44.
51. A. Alissandrakis, C. L. Nehaniv and K. Dautenhahn, Correspondence mapping induced state and action metrics for robotic imitation, *IEEE Trans. Syst. Man Cybern. B, Cybern* **37**(2) (2007) 299–307.
52. A. Alissandrakis, D. S. Syrdal and Y. Miyake, Helping robots imitate, *New Frontiers in Human–Robot Interact* (Amsterdam, Netherlands, 2011), p. 9.
53. K. Mombaur, A. Truong and J. P. Laumond, From human to humanoid locomotion: An inverse optimal control approach, *Auton. Robots* **28**(3) (2010) 369–383.