Phase Transfer Sequence Learning for Humanoids' Whole Body Motor Control in Nonstationary Environments

Toshihiko Shimizu, Ryo Saegusa, *Member, IEEE,* Shuhei Ikemoto, Hiroshi Ishiguro, Giorgio Metta

***Abstract*—This paper proposes an invariant feature named "phase transfer sequence" to encapsulate knowledge of humanoids' whole body movements in nonstationary environments. The phase transfer sequence represents turning points in sequential signals: it encodes dynamical aspects of the signals with absorbing timing and amplitude factors. The proposed feature is applied for description of sensory-motor knowledge of a robot while being instructed actions by a human. The robot, then, exploits the knowledge to guide learning of motor control in demonstration of the actions without humans' support. We evaluated the effect of the phase transfer sequence for guiding reinforcement learning of a sit-up motion and walk motion in simulations and with an actual humanoid robot. The experimental results show that in the both motor tasks the proposed feature described the actions as less environment dependent than conventional methods do, and enhanced convergence speed of action learning in autonomous demonstrations.**

***Index Terms*—change detection, feature extraction, imitation learning, nonstationary environment, human-robot interaction**

I. Introduction

SOCIAL motor interaction helps developments of infants' motor skills towards self-sustained walk [1] [2]. In the process of the development, however, a question remains: how do infants extract skills from experiences of walks supported by adults and finally acquire skills to self-sustained walk? Even though from outside these motion patterns look similar, internal motor controls are physically different because interaction with adults derive external forces and spatial constrain in locomotion.

In robotics, imitation learning is known as a solution for acquiring motor skills from interaction with a teacher. Imitation learning has been broadly applied to humanoid robots in the learning of whole body movements such as tennis swing [3], bipedal walk [4], whole body dance [5], and dynamical roll and rise motions [6]. However, environmental changes make it difficult to reproduce the motions in these methods, because motor information is directly represented by raw sensory sequences (e.g., joint angles or velocities) which can be strongly influenced by the surrounding environment. Moreover, in the conventional methods, movements are measured by external sensors or environmentally-embedded sensors such as optical motion capture systems. This kind of sensors give reliable measurements for localization, but use of these environment-dependent sensors is against our concerns in which we focus on motor developments of agents that act in a "nonstationary" environment.

In order to relax the dependency on timing for the representation of sequential information, dynamic time warping [7] is often used in literature. However, differences of the signal amplitude still influence the representation. The more robust representation of the motor information is required to reuse actions learned in the different environment.

In contrast to previous studies in imitation learning, we focus on motion phase transition. A movement, composed

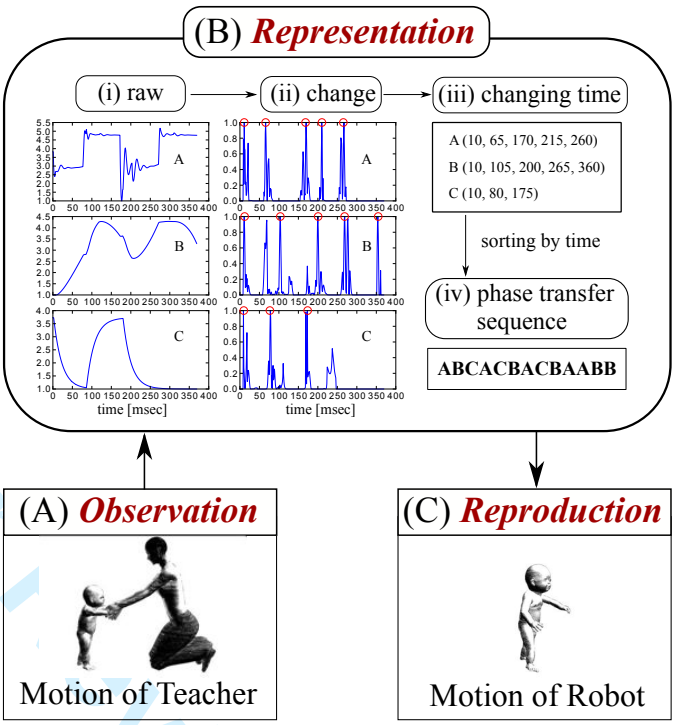


Fig. 1. The process consists of three phases; Observation, Representation, and Reproduction. The concept of the phase transfer sequence is shown in (B) Representation. For the sensor, the raw data (i) is transformed into a change scores (ii). Then the peaks of each change scores are collected (iii), and the phase transfer sequence (iv) is obtained by sorting the sensor labels by time.

T. Shimizu and H. Ishiguro are with Department of Systems Innovation, Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama Toyonaka Osaka 560-8531 Japan. (e-mail: shimizu.toshihiko@is.sys.es.osaka-u.ac.jp; ishiguro@sys.es.osaka-u.ac.jp).

S. Ikemoto is with Department of Multimedia Engineering, Graduate School of Information Science and Technology, Osaka University, E6-411 2-1 Yamada-oka Suita Osaka Japan. (e-mail: ikemoto@ist.osaka-u.ac.jp).

Ryo Saegusa and G. Metta are with Department of Robotics, Brain and Cognitive Sciences, Italian Institute of Technology, Via Morego, 30 16163 Genova, Italy. (e-mail: ryos@ieee.org; giorgio.metta@iit.it).

Manuscript received May 7, 2012; revised ***, 2012.

by a series of actions, involves some critical motion phase transitions in the sequence, and the satisfaction of the motor phase transitions in the correct order is important to achieve the movement [8]. For example, a walk motion consists of the following sequential phases: one leg supports the body, the other leg moves to the front, and this leg supports the body. This motion phase transition seems to be common in a variety of walking conditions, such as self-sustained walk, supported walk, walk up and down on slopes or stairs. The motion phase decision trees proposed in [9] stand on this idea: the system creates decision trees of the motion phases from successful and failure trials of stand-up actions on a flat floor, and reproduces imitated motions on a slope. The technique we propose is closely related to this work. However, in our case the representation of the information is not based on raw data sequences, which limit the adaptability for environmental changes.

In order to extract latent relations from multiple different sensory signals, Singular Spectrum Transformation (SST [10]) was introduced in the data mining field. The SST evaluates the variation of a temporal sequence at the center of a sampling window. By feeding temporal sensory-motor signals to the SST at every time step, the presences of changes are sequentially detected. This technique is prospectively applicable to encode the motion phase transitions in robot's multi-modal sensor signals.

We propose a novel motor representation "phase transfer sequence" which represents the motion phase transitions in multiple sensory-motor signals. We then apply the phase transfer sequence to a robotic imitation learning system, and verify the robustness of the representation in different environments as well as its contribution to the learning convergence speed in simulations and experiments with a real humanoid robot.

This paper is organized as follows. In Section II, a novel representation of robot movements, "phase transfer sequence" is derived. An imitation learning system based on this representation is described in Section III. Experimental results of learning in nonstationary environments are compared among the proposed system and some conventional methods in Section IV. In Section V, related works are reviewed and a comparison between the proposed method and existing literature is provided. Finally, Section VI concludes this paper.

II. Phase Transfer Sequence

Figure I shows an overview of the proposed system. When an action is executed, multiple raw sensor signals are fed to the system (a). Change scores of the sequences are computed using SST (b), and the peaks of change scores are collected (c). The peaks are labeled with a symbol corresponding to the sensor identifier, and the representation is generated concatenating the symbols in the temporal order (d). Note that we assume that all the sensor signals are derived from internal sensors equipped with the robot platform.

Here, we employ the phase transfer sequence to measure the difference between two temporal sequences. It is compared to other distance measures; Euclidean distance and time warping [7] as shown in Fig.2. All of them measure

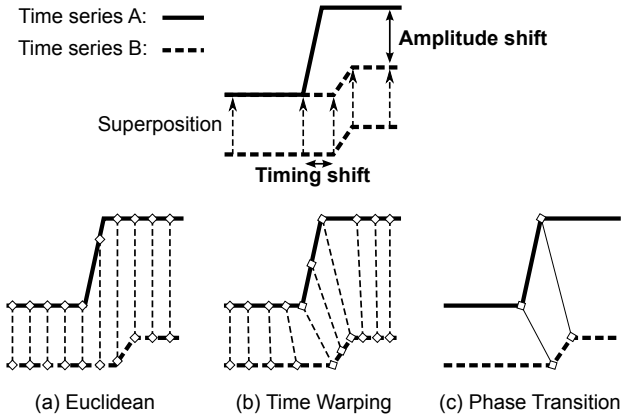


Fig. 2. Conceptual illustrations of three distance measures between two time series. The shifts in the timing and amplitude are compared by superposing the time series. Euclidean distance is affected by the both shifts and time warping is only affected by the amplitude shift. Phase transfer sequence are not affected by the both shifts, because both two time series have the two phase transitions in common.

the distance between two temporal sequences as shown in Fig.2. The Euclidean distance is an intuitive and popular measure, but it is sensitive to environmental changes, since it is affected both by temporal shifts and variation in the amplitudes. The time warping approach is more robust against temporal shifts, however it is still influenced by the amplitude values. In contrast, the phase transfer sequence only represents phase transitions, which are invariant to shifts in timing and amplitude. As shown in the figure, the phase transfer sequence encodes the two sequences as the same pattern with two phase transitions. The timing and amplitude are environment dependent, while phase transfer sequence extracts transitions of motion phases from the sequences: therefore, the motion representation is more universal over different environments.

The proposed symbolic representation serves as a dimensional reduction technique for multiple sensory sequences. Furthermore, its symbolic nature allows the application of traditional string computation algorithms, such as the longest common subsequence (LCS) [11].

III. Phase Transfer Sequence Learning

We apply the phase transfer sequence to robotic imitation learning. As shown in Fig.I, the proposed process consists of three phases; observation, representation and reproduction.

1) Observation phase: A robot observes temporal sequences of multiple internal sensors, while being instructed by the teacher on the task execution, as shown in Fig.I(A). The robot is not provided with any other knowledge on the task motion.

2) Representation phase: In the representation phase, phase transfer sequences are extracted. We use Robust Singular Spectrum Transformation (RSST) [12] to detect the phase transitions of sensor sequences. The RSST is an improved version of the SST that decreases the number of parameters. We describe the computation procedure of the RSST in Algorithm 1. After calculating the changing score of each sensor $x_s^p(t)$, we get the changing times t^p by taking the local maximums of $x_s^p(t)$, where p is a sensor in the sensor

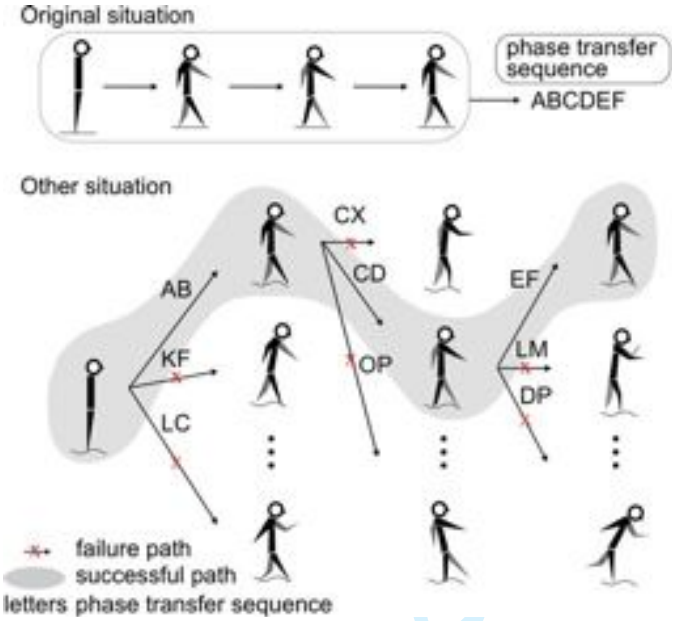


Fig. 3. The schematic view of motion learning.

set $P = \{p_1, ..., p_{N_p}\}$ and N_p is the number of the sensors. Finally, we get the phase transfer sequence O by sorting the the labels of each sensor by temporal order. Then we select the reference trial $reference$ to generate the reference phase transfer sequence O_r which is used for the reference point of the distance calculation.

3) Reproduction phase: Finally, in the reproduction phase, phase transfer sequences are used as a guide for motor exploration in a novel environment, as shown in Fig.I(C). In this phase, the system autonomously choose actions and learns how the taught motion should be adapted for execution in a different environment.

In order to learn the motions, we use reinforcement learning (RL). In the RL framework, an agent performs actions until achieving a goal state as shown in Fig.3. The agent is given rewards from the environment when it reaches some valuable goal states. If the reward is given only in the final goal state, all trials with intermediate failures are wasted, as shown in Fig.3. As a result, a huge number of trials are necessary to achieve learning convergence. Effective rewards for achievement of sub-goal should be designed in order to speed up the learning.

We, thus, apply the phase transfer sequence for guiding the motor exploration of the learning in the novel environment. An agent obtains a reward at intermediate states by measuring similarity between the current phase transfer sequence O_t and the reference phase transfer sequence O_r learned previously in the observation phase. The similarity is measured by the length of the Longest Common Subsequence (LCS) [11]. By this measure, the longer the common subsequence of O_t and O_r is, the more similar the motion phase transitions in these environments are.

We employed Q learning [13] for motion learning. The action-value function was defined as follows:

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha(r(s,a,\acute{s}) + \lambda \max_{\acute{a} \in A} Q(\acute{s},\acute{a})), \quad (1)$$

where a and s are the current state and the action, respectively, and A and S is the set of actions and states. $\acute{a}$ and $\acute{s}$ are the next state and the next action, respectively. The strategy employed for the action selection is ϵ-greedy. In this paper, $episode$ denotes a trial which consists of the state transitions by the selected actions, and run consists of multiple episodes. The following parameters are constant in all the experiments; $\alpha = 0.25$, $\lambda = 0.9$, $\epsilon = 0.5$.

We defined the reward function as follows:

$$r(s,a,\acute{s}) = \begin{cases} 1 + f(\acute{s}) & (\acute{s} = s_g \text{ and } t_j = t_e), \\ f(\acute{s}) & (\acute{s} \in S), \\ -1 & (\acute{s} \notin S \text{ or } fail), \end{cases} \quad (2)$$

$$f(\acute{s}) = \begin{cases} 0 & (s = s_s), \\ g(\acute{s}) - f(s) & (\text{else}), \end{cases} \quad (3)$$

$$g(s) = \begin{cases} 0 & (\text{without sub-goal reward}), \\ g^E(s) & (\text{with Euclidean}), \\ g^T(s) & (\text{with TimeWarping}), \\ g^P(s) & (\text{with Phase Transfer Sequence}) \end{cases} \quad (4)$$

where s_s and s_g is the start and goal state, t_e is the end time of the experiment, $fail$ denotes a task failure condition, $f(\acute{s})$ is the sub-goal reward obtained by performing a at s, $g(s)$ is the sub-goal reward obtained when the agent reaches s, $g^E(s)$, $g^T(s)$, $g^P(s)$ are the sub-goal rewards described later in this section. In order to evaluate the sub-goal reward obtained from s to $\acute{s}$ with a, (3) subtracts the previous sub-goal reward $f(s)$ from the current sub-goal reward $g(\acute{s})$.

Three types of the sub-goal rewards; Euclidean distance, time warping and phase transfer sequence are compared in our experiments. We assume a set of sensors $P = \{p_1, ..., p_{N_p}\}$, where N_p denotes the number of the sensors to be used. The sensory information from time 0 to time t is defined as follows: $X(t) = \{\vec{x}_1(t), ..., \vec{x}_{N_p}(t)\}$, where $\vec{x}_i(t) = \{x_i(t_1), ..., x_i(t)\}$ is the sequence of the values of p_i over time.

The reward based on Euclidean distance is computed as follows:

$$g^E(s) = \frac{1}{1 + c^E \sum_{i=1}^{N_p} |\vec{x}_i^c(t_s) - \vec{x}_i^r(t_s)|}, \quad (5)$$

where t_s is the current time, $\vec{x}_i^c(t_s)$ and $\vec{x}_i^r(t_s)$ are the sensor sequences of the current and the $reference$ trial, respectively, and c^E represents a constant to scale the sum of Euclidean distance. As shown in Fig.2(a), this function compares the sensor sequences of the current trial to those of $reference$ trial from the start time until the current time t_s.

The reward based on time warping is computed as follows:

$$g^T(s) = \frac{1}{1 + c^T \sum_{i=1}^{N_p} TW(\vec{x}_i^c(t_s), \vec{x}_i^r(t_e^r))}, \quad (6)$$

where function TW represents the time warping distance between $\vec{x}_i^c(t_s)$ and $\vec{x}_i^r(t_e^r)$, and t_e^r is the end time of the $reference$ trial, because two sequences with different temporal length can be evaluated by time warping. Thus this function compares the sensor sequences of the current trial from the start time to the current time t_s and those of the $reference$ trial from the start time to t_e^r, as shown in Fig.2(b).

The reward based on phase transfer sequence is defined as follows:

$$g^P(s) = \frac{|LCS(O_r, O_c(t_s))|}{|O_r|}, \quad (7)$$

where $LCS(O_r, O_c(t_s))$ is the LCS of O_r and $O_c(t_s)$. $O_c(t_s)$ represents a phase transfer sequence of $X(t_s)$ of the current trial. $|O|$ is the length of the phase transfer sequence. If the two sequences are similar, these sub-goal rewards are close to 1 in all indices. Therefore, (3) provides a greater, positive reward to motions similar to $reference$.

IV. Experiment

The proposed method was evaluated with a real humanoid robot and its corresponding simulator. The robot platform is iCub [14], which is a child-scale full-body humanoid robot (about 104cm tall) with 53 degrees of freedom (DOF) [15]. The robot platform is controlled with CPU clusters via YARP [16]. In the experiment, we used force-torque sensors mounted in the four limbs [17], an inertial sensor mounted in the head, joint encoders and corresponding motors for all the DOF. The simulated robot [18] is corresponding to the real robot platform.

We selected sit-up and walk movements as motor tasks for the robot. First, we instructed the robot on the task motions with human assistance (supported condition), and then executed the reinforcement learning without this assistance (unsupported condition). The first objective of the experiments is to investigate how the learning convergence in the unsupported condition is improved by the knowledge extracted from trials in the supported condition. The supported and unsupported motions were remarkably different due to the difference from contact conditions. The supported motion, therefore, cannot be directly used to achieve the unsupported motions. The second objective is to compare the robustness of the representation of sensory sequences when Euclidean distance, time warping and the proposed phase transfer sequence are employed. Table.I shows the configurations of the phase transfer sequence in the sit-up and walk experiment.

In the sit-up experiment, teachers (experimenters) lifted the robot to teach the sit-up motions. In order to ease the teacher to lift the robot, the pitch joints of both shoulders and hips of the robot were freed, and the pitch joint of the torso was controlled by a torque-based control. The other joints were kept at their home position by a position-based control. The teacher lifted the robot by holding both lower arms of the robot. As the initial posture, the robot was laid down on the ground as shown in Fig.4 left.

In the walk experiment, teachers (experimenters) grabbed the arms of the robot and taught it a walking motion. In order to achieve a safe interaction between humans and the robot, we implemented a reactive leg motion control as follows: When the teacher pulls up one arm of the robot, the robot detects the stimuli and steps forward with the opposite side leg as shown in Fig.5. This movement is similar to a postural reaction in infants to recover the stability. The initial posture for the learning phase is shown in Fig.5 left. The configurations of the joint angles for the postural reactions are shown in TableII. The hip yaw joints of both legs were externally rotated for gaining the stability and propulsion [19]. Both arms were extended forward to interact with a teacher.

TABLE I
THE CONFIGURATIONS FOR GENERATING THE PHASE TRANSFER SEQUENCE.

The configurations for sit-up experiment.				
part name	sensor name	RSST (real/simulation)		label
		w	n	
left arm	force x	8 / 8	8 / 8	0
right arm	force x	6 / 6	6 / 6	1
left leg	force x	8 / 8	8 / 8	2
right leg	force x	8 / 8	8 / 6	3
left leg	hip pitch encoder	10 / 10	4 / 10	4
right leg	hip pitch encoder	10 / 10	10 / 10	5
left arm	shoulder pitch encoder	10 / 10	4 / 10	6
right arm	shoulder pitch encoder	10 / 10	10 / 10	7
torso	torso pitch encoder	10 / 10	10 / 10	8
head	inertial pitch	8 / 8	4 / 4	a
head	inertial gyro y	16 / 16	4 / 4	b
The configurations for walk experiment.				
part name	sensor name	RSST (real/simulation)		label
		w	n	
torso	torso roll encoder	10 / 8	10 / 6	0
right leg	hip pitch encoder	20 / 8	10 / 6	1
left leg	hip pitch encoder	16 / 8	16 / 6	2
right leg	hip roll encoder	16 / 8	16 / 6	3
left leg	hip roll encoder	14 / 8	14 / 6	4
left leg	force z	8 / 8	8 / 8	5
right leg	force z	8 / 8	8 / 8	6

TABLE II
THE JOINT ANGLE CONFIGURATION FOR THE WALK EXPERIMENT.

joint name	reactions			reproduction
	home	right	left	home
torso roll	8.0	-8.0	8.0	8.0
hip pitch (left leg)	0.0	-2.4	8.0	0.0
hip pitch (right leg)	0.0	8.0	-2.4	0.0
hip roll (both leg)	0.0	8.0	8.0	0.0
hip yaw (both leg)	48.0			48.0
shoulder pitch (both arm)	-30.0			0.0
shoulder roll (both arm)	30.0			30.0
elbow pitch (both arm)	45.0			0.0
other joints	0.0			

A. Experiment in Simulation

We implemented an interactive GUI for the robot simulator [18]. A movement of a pointing device is translated into an external force to be applied to the robot arms. The sensor information was sampled from $0.1[sec]$ before the interaction and to $0.1[sec]$ after the interaction with a sampling interval of

Fig. 4. The screenshots of the teaching experiment of the sit-up motion. The left figure shows the initial state. The figure in the center shows an intermediate state of lifting up. The right figure corresponds to the final state of the sit-up.

Fig. 5. The initial posture and the implemented reactions. The left figure shows the initial posture, the center figure is the reaction to a stimuli on the left arm and the right figure is the reaction to a stimuli on the right arm.

$0.005[sec]$. A session of interaction in the sit-up experiment is segmented when the robot achieves the sitting posture. In the walk experiment, it is segmented when the robot achieves a certain number of successful walk steps. We conducted 10 trials to teach supported motions and sampled a reference phase transfer sequence O_r as a reference motion by choosing the best and smoothest trial.

1) Learning of unsupported sit-up: The initial, intermediate and final states are defined in Fig.4 in the left, center and right, respectively. Each state is described with the Euler angles of the head, the joint angles of the shoulder and hip pitch encoders of the limbs. The data of each state were recorded at $\{0.0, 0.6, 1.2\}[sec]$. The actions $A = a_i(i = 0, 1, 2)$ were defined in Table III. The a_2 means $no\text{-}action$ which keeps the current posture. The agent selected the action $a(t_j)$ at time $t_j = \{0.0, 0.6\}[sec](j = 0, 1)$. The goal state s_g is the sit-up state and the end time is $t_e = 1.2[sec]$. The $episode$ was regarded successful, if the robot reached s_g at $1.2[sec]$, or else it is regarded as failed. If 3 $episodes$ are consecutively successful, a run is completed. If the number of $episode$ exceeds 40, the run is regarded failed and stopped. We conducted 10 $runs$ for each reward function and compared the learning speed.

Figure 6 and Table IV show profiles of the average reward and the average number of episodes for learning convergence. Fig.7 shows screenshots of a learned sit-up motion in the unsupported condition. The average of the convergence speed was improved by the use of sub-goal rewards based on phase transfer sequence and time warping, but not on Euclidean distance. In the case of the phase transfer sequence, 8 $runs$ converged within 14 $episodes$. In the case of the phase transfer sequence and Euclidean Index, however, at least one run of the learning did not converged and the learning was trapped in a local solution by the wrong guide of the sub-goal rewards. On average, the convergence speed was improved by the phase transfer sequence, time warping, and Euclidean distance in this order.

We conducted experiments to investigate the robustness of the proposed motion feature. We evaluated the distances from the unsupported and the supported motions to the $reference$ motion which is a supported motion with an "ideal" pattern. The three trials of learned sit-up and 9 trials of the taught sit-up except the "ideal" pattern from 10 trials were used for this comparison. Fig.8 reports plots of the unsupported and the supported motions with respect to phase transfer sequence and the other measures. As we see in the figures, Euclidean and time warping indices generate separable clusters (along

TABLE III
THE ACTION DEFINITIONS.

joint name	walk action			sit-up action	
	0	1	2	0	1
torso roll	-8.0	8.0	-8.0	0.0	
torso pitch	0.0			0.0	8.0
hip pitch (left leg)	-2.4	19.8	-2.4	0.0	88
hip pitch (right leg)	11.0	-2.4	25.6	0.0	88
hip roll (both leg)	8.0			0.0	0.0
hip yaw (both leg)	48.0			0.0	0.0
shoulder pitch (both arm)	0.0			0.0	-88
shoulder roll (both arm)	30.0			30.0	
elbow pitch (both arm)	0.0			45.0	
other joints	0.0				

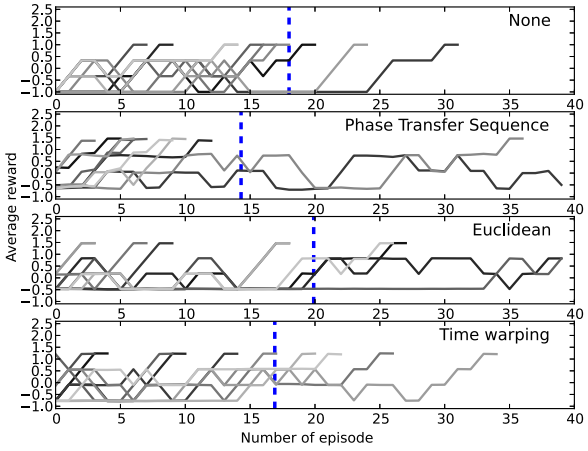


Fig. 6. The average reward acquired during 10 $runs$ with each sub-goal reward function. The top figure shows the result without the sub-goal reward, and the second, third and forth figure show the results obtained with the phase transfer sequence, the Euclidean, time warping respectively. The vertical dot-line shows the average number of $episodes$ required for the learning convergence. The profiles were smoothed by a low-pass filter with a 3 $episodes$ sampling window.

Fig. 7. Screenshots of a learned sit-up motion in the unsupported condition.

TABLE IV
THE LEARNING RESULTS OF THE SIT-UP EXPERIMENTS.

reward	average speed	successful run	under 14 $episodes$
None	18.0	10	2
LCS	14.3	9	8
Euclidean	19.9	8	4
Time warping	16.9	10	4

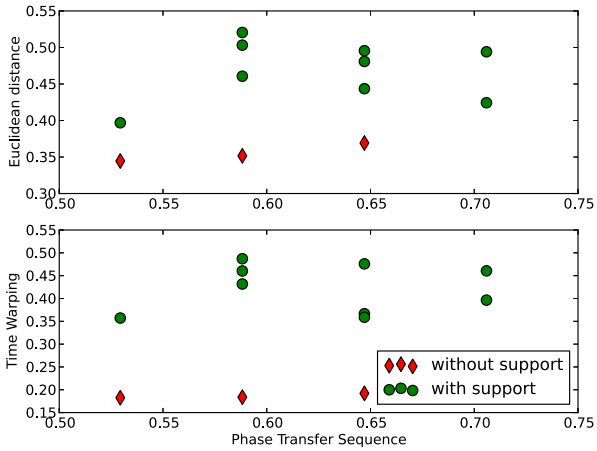


Fig. 8. The distributions of the supported and unsupported sit-up movements. The figures show the results with the phase transfer sequence versus the Euclidean distance (the upper figure) and Time Warping (the lower figure).

the vertical axis) corresponding to the supported motions and unsupported motions. The phase transfer sequence, however, gives inseparable distribution of support motions and unsupported motions (along the horizontal axis). This suggests the robustness of the phase transfer sequence in describing motions, and the knowledge of the reference trial (supported motion) was effectively reused for the reinforcement learning in the unsupported condition.

2) Learning of unsupported walk: In the learning of unsupported walk, the state is defined as the absolute position of the robot head, as shown in Fig.9. We divided the workspace of the robot with grids along the x, y, and z axis of the world coordinates, and assigned indices to the grid sub-spaces. In particular, each axis was divided into $(d_x, d_y, d_z) = (4, 3, 2)$ partitions, for a total number of the subspaces of 24. The range of each axis is $xlim = [-0.1, 0.5]$, $ylim = [-0.25, 0.25]$, $zlim = [0.5, 1.0][m]$, respectively. As shown in Fig.9(left), we defined the set of states as $S = \{s_0, s_1, s_2, ..., s_{23}\}$ with notations, $s_0 = \{0, 0, 0\}$, $s_1 = \{0, 0, 1\}$, ...,$s_{23} = \{3, 0, 1\}$. The goal state was $s_g = s_{19}$. The actions were defined $A = a_i (i = 0, 1, 2)$ as shown in Table III. The agent selected the action $a(t_j)$ at time $t_j[sec] = \{0.0, 0.4, 0.8, 1.2\} (j = 0, 1, 2, 3)$. The initial posture of the robot at t_0 was set at the home position, reported in Table II with the label home.

The $fail$ condition, mentioned in the rewarding rule of Eq. (2), is defined as a trial in which the Euler angle of the robot head is out of the range $\{-20°, 20°\}$. Given this definition, when the agent reaches the goal state, the reward rule gives a positive reward to the agent. On the contrary, when the agent exits the area (out of the states) or $fail$, the reward rule gives a negative reward to the agent. Each trial lasts $1.6[sec]$ or stops earlier when the robot reaches the goal state s_g. If the robot fails in the middle of the movement execution, the trial is also stopped at the time. One run of experiments is organized as follows: if the $episodes$ are successful for 3 consecutive times, the run is completed, while if the number of $episodes$ exceed 1000, the run is regarded as failed. We performed 10 $runs$

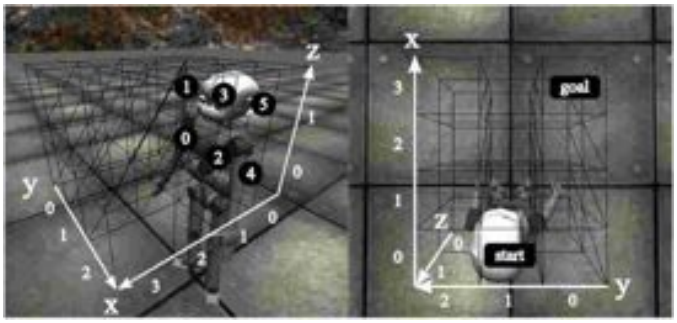


Fig. 9. State definitions. The grid divides the space into subspaces used as states. In the left figure the black nodes with a white number show the indices of the states. In the right figure the black nodes with white indices show the initial state and the goal state.

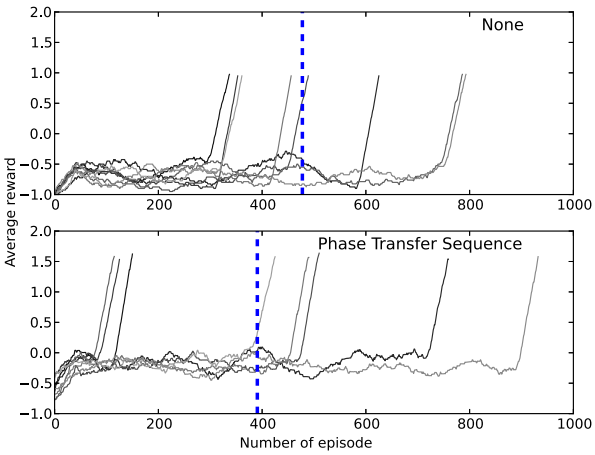


Fig. 10. The average of the reward acquired during the episodes. The upper figure shows the experiments without the phase transfer sequence in the reward function, and the lower shows the one with it. The blue vertical dot-line shows the average number of steps necessary for the convergence in successful experiments. A low-pass filter with 80 trial time constant was used for smoothing. Experiments converged in less than 80 trials. The not-converged trials were excluded from the results.

with phase transfer sequence rewards and no sub-goal rewards to compare the learning speed in these conditions. Fig. 11(a) shows the $reference$ trial of the supported walk.

Figure 10 and Fig.11(b) show the profiles of the average reward and the acquainted walk motion. In case of the phase transfer sequence reward, 9 $episodes$ were successful, while in case of no sub-goal reward, 10 $episode$ were successful. The average convergence speed with the phase transfer sequence reward and no sub-goal reward was 390 $episodes$ and 477 $episodes$, respectively. The number of $runs$ that converged within 200 $episodes$ was 4 $episodes$, and 1 $episode$, respectively. The non-converged case with the phase transfer sequence was possibly caused by local trapping in learning, or noise present in the $reference$ phase transfer sequence which made the unnecessary subsequence for unsupported walks.

We also verified the robustness of the phase transfer sequence with respect to environmental changes. We produced 10 successful and 10 failure of unsupported walks by adding random noise (with normal distribution, $N(0.0, 10.0)$) to mo-

(a) A walk motion instructed by a teacher with an interactive GUI.

(b) A walk motion generated by the robot after learning.

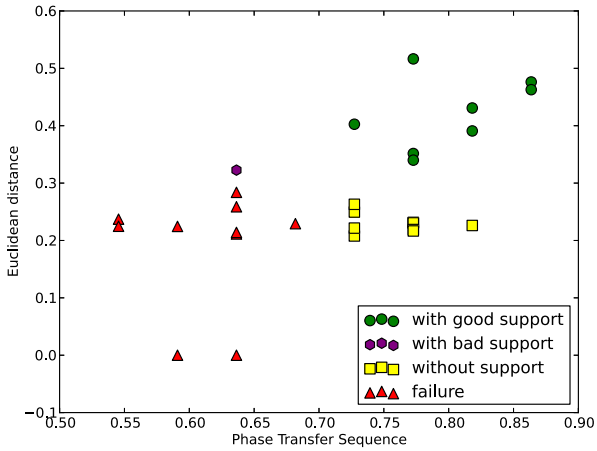
(c) A badly instructed walk motion. At the third step, the right leg is slightly stuck to the ground (shown in 4th figure from the left).

Fig. 11. The screenshots for the walk experiment in simulation.

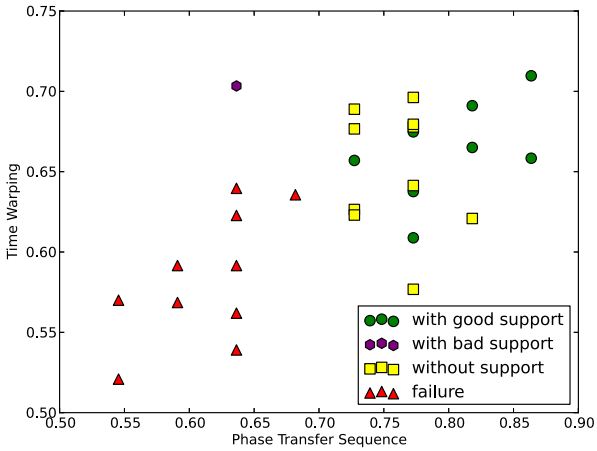
tor commands of joint angles of the 1st, 2nd and 3rd step of the learned motions. Also for this experiment, the number of the supported walk trials was set to 10, and one trial was used as the reference trial. The sensor signals were observed from $0.1[sec]$ before the start of the motion execution to $0.1[sec]$ after the robot finishes executing its third step. The sampling interval was set as $0.005[sec]$. If the absolute position of the robot head reaches the goal, the trial is regarded as successful.

Figure 12 shows the experimental results. In the figures, the horizontal axis indicates the phase transfer sequence give by Eq.(7), while the vertical axis indicates Euclidean index given by Eq.(5) or time warping index given by Eq.(6) in Fig.12(b). The successful motions of the supported and unsupported walk are both located over about 0.7 along the phase transfer sequence axis. Conversely, when considering the Euclidean index, the supported and unsupported walks are clearly separated. The results show that plots with good support and without support are distributed in a similar area in case of the phase transfer sequence and time warping index, but in case of Euclidean index, these plots formed respective clusters. This means that the motion patterns are represented differently for different environments. On the contrary, the supported and unsupported motions are grouped by the phase transfer sequence. The phase transfer sequence, moreover, separates the successful and failure cases in two different clusters, This means that the representation of the phase transfer sequence is sensitive to the success or failure but not to the environmental changes (supported or unsupported) as desired.

Figure 11(c) shows a trial of incorrect teaching which is described as a bad support in Fig.12. In this case, the trial itself was judged as successful, but the robot stumbled at the 3rd step, and the motion phase transitions were different from the reference phase transfer sequence.



(a) With Euclidean distance.



(b) With time warping distance.

Fig. 12. The distributions of the supported and unsupported walk movements. The figures show the results with the phase transfer sequence versus the Euclidean distance (the upper figure) and Time Warping (the lower figure).

B. Experiment in a Actual Robot

As a first step, we conducted the sit-up experiments for verifying the proposed method with the actual robot. For the safety of the robot, the robot was laid on a bed made of cardboard, which had a slit to pass through a power-supply cable on the back. The ankles were banded on the bed. An experimenter taught a sit-up movement to the robot 10 times as shown in Fig.13.

The state and the action were defined as in Sec.IV-A1. We let the robot learn the sit-up movement with the sub-goal reward based on the phase transfer sequence as defined in Eq.(7) and without the sub-goal reward. We conducted 5 $runs$ for each case, and compared the average convergence. A run was organized as follows: if the $episodes$ were successful for two consecutive times, the run was regarded as finished, while if the number of $episode$ exceeded 20, the run was regarded as failed and stopped.

Table V and Fig.14 show the numbers of episodes for

Fig. 13. The interaction for teaching a sit-up motion.

Fig. 14. The demonstration of the learned sit-up motion by an actual robot.

TABLE V
THE LEARNING RESULTS OF THE SIT-UP.

condition	episodes of each run					average
with phase transfer sequence	3	3	9	6	20	8.2
without phase transfer sequence	10	7	4	11	15	9.4

learning convergence and the achieved motion. The trend in the results is similar to the one obtained in the simulation: The motor knowledge encoded as the phase transfer sequence enhanced the average convergence speed in terms of the number of episodes.

As a second step, we conducted the walk experiment with an actual robot. In the teaching phase, we holed the arms of the robot and taught the walk movement (supported walk). In the learning phase, we did not let the actual robot explore, but recycled an unsupported walk movements achieved in the simulation. We, then, let the robot demonstrate the unsupported walks and verified the similarity of the phase transfer sequence between the supported and unsupported walk. Figure 15 and Fig.16 shows interaction of walk teaching and demonstration of the learned walk, respectively.

In the robot's demonstration, we helped the robot motion by holding its torso. This assistance aims in compensating the torque of roll joints for walking, but not interfering other features of the movement. The walking pattern is purely generated by the robot itself.

For the comparison of walk movements, the experimenter (teacher) instructed the supported walk for 7 trials, and the reference phase transfer sequence O_r was selected from them. The teacher, then, let the robot demonstrate 10 trials of the unsupported walk. Also, in the similar manner, the teacher instructed the supported sit-up for 10 trials, and let the robot demonstrate 10 trials of the unsupported sit-up. Figure 17(a) and Fig.17(b) show the comparison of the sit-up and the walk movements. Thanks for the consideration of the phase transfer sequence, the distribution of the supported and the unsupported motions were closer than that with other measurements (Euclidean distance and time warping).

Fig. 15. The interaction for teaching a walk motion.

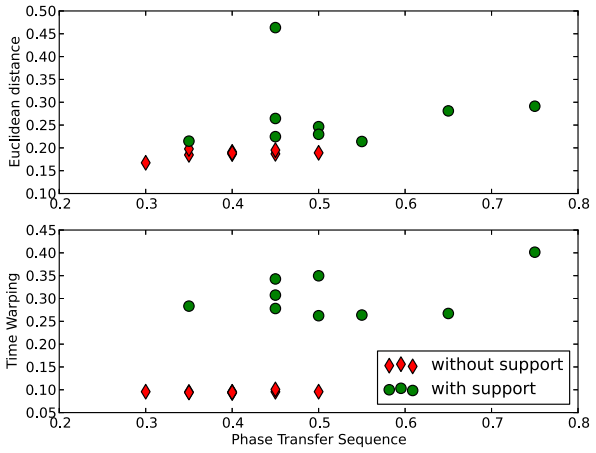
Fig. 16. The demonstration of the learned walk motion by an actual robot.

V. DISCUSSION

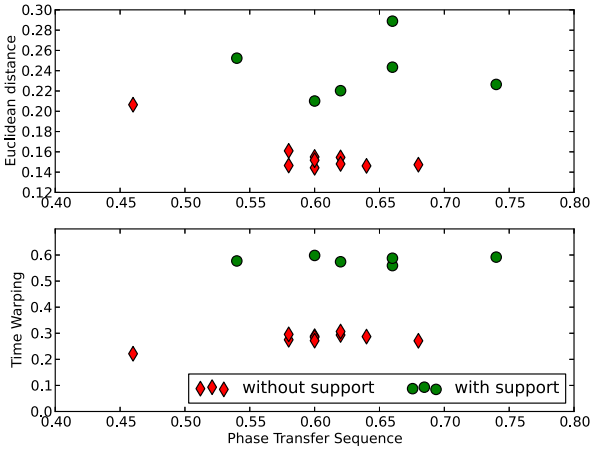
Here, we overview the related works and the relations to the proposed approach. Figure 18 provides a schematic representation of the comparison between two identical-purpose movements in the different condition. The imitation learning of whole body motions can be categorized into two paradigm, trajectory-based and point-based approaches.

In the trajectory-based approach, the motion representation is given by the trajectories of the motion quantities such as joint angles and velocities. In trajectory-based approaches, Dynamic Motion Primitives (DMP) [3] is widely known. In the DMP framework, human movements are observed with a motion capture system, and represented as differential equations, whose parameters are described by Locally Weighted Regression (LWR). LWR creates a bounding envelope that guides the motion state along a successful trajectory from the start state to the goal state as shown in Fig.18. This method achieved a tennis swing movement with a whole body humanoid robot. DMP also allowed bipedal walking by employing central pattern generator (CPG) [4], which adjusts the walking patterns by the entrainment between the environmental and the self-body dynamics. Programing by Demonstration (PbD) [20] represents trajectories of joint angles by Gaussian Mixture Regression that is used to reproduce manipulation tasks in different contexts (while PbD is not examined in whole bodied motion generation though).

In the point-based approaches, motions are represented by some important states in the motion sequences (intermediate states in Fig.18). One of the point-based approaches is motion segmentation by key postures in which velocities of an end-effector becomes zero [5]. This method reproduces imitated motions by connecting key postures to form a successful trajectory. A reproduction of Japanese traditional dance was achieved by this method. Another approach is based on identification of bottleneck states necessary to achieve a motion [6]. Bottleneck states were evaluated in human motion data measured by a motion capture system. Given these bottleneck states, a whole bodied humanoid robot reproduced dynamic roll and rise motions by controlling its movements as to pass the bottlenecks states. The motion phase decision tree algorithm [9] built a decision trees of motion phases from a set of successful and failure trials. The algorithm, then,



(a) Sit-up experiments with an actual robot.



(b) Walk experiments with an actual robot.

Fig. 17. The comparison between the index of the phase transfer sequence and other indices between the reference trial and test trials of different contexts. The horizontal axis indicates the phase transfer sequence, while vertical axis indicates the other indices.

reproduces motions in different contexts by using the trees as guideposts in motion learning. The idea of this work is close to the proposed method.

The proposed method is in principle a point-based approach in the sense that it extracts features of a motion, the phase transitions, given at finite time sections. In this paper, we consider that the representation of phase transitions is more robust for environmental changes than other methods using raw sensory sequences, because the phase transitions are purely dynamical aspects of the motion and not influenced by timings and amplitudes of profiles.

As expected, we positively verified that the proposed phase transfer sequence enhanced the learning convergence in a reinforcement learning system exploring in environmental changes. In the sense of skill transfer learning, this achievement are promising, however, we need further improvements: for instance, in order to determine the $reference$ trial in the reproduction phase, the human judgement was required. Potentially, the proposed method can be combined with trajectory-based approaches such as DMP to recover motion smoothness and other lost information.

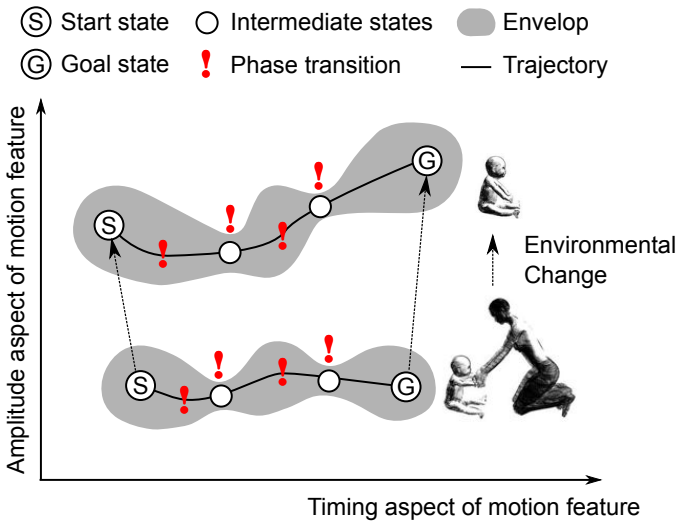


Fig. 18. Comparison of the proposed work to related works in the environmental changes. In the related works, motions are characterized by the trajectory, the envelop and intermediate states, while in the proposed method the motions are characterized by the phase transitions. These identical-purpose motions are desired to be represented commonly.

Here, we discussed some literature about acquisition of walk in humans. In the aspect of the body structure, the acquisition of walk might be assisted by external rotation of hip joints [19]. The development of skills is considered synchronous to physical growth of the body structure [21]. In the social aspect, naive walk experiences with the stepping reflex [1] and treadmill trainings [2] with adults promote infant's skills towards self-sustained walking. The acquired mobility, moreover, enhances infants' cognitive ability [22]. Convincingly, socially isolated children have difficulty in acquiring natural walk [23].

The proposed method is apparently related to the developmental process of infants' motor skills assisted by social interaction. The infants seem to apply supported walk experiences for motor exploration in the unsupported condition such as a life-long learning manner [24]. Correspondingly, the proposed system recycle the knowledge of the supported walks in the unsupported condition. In order to improve the unsupported walk, the gait pattern generation with the CPG [25], the reflex-based control [26] and the ZMP control [27] can be integrated with the proposed system.

VI. Conclusion

We proposed a novel motion feature to encode phase transitions in motion sequences, and then applied the feature to imitation learning of robots' whole body movements. The proposed feature, phase transfer sequence, characterizes dynamics of motions robustly in terms of timing and amplitude of signal profiles. This property gives a great advantage in evaluating similarity of motions demonstrated in different environmental contexts. The system extracted phase transfer

sequences from the motions instructed by humans in a supported condition, and used it as a guide for reinforcement learning of motions in an unsupported condition. We verified in simulations and experiments with an actual humanoid robot that the use of instructed knowledge enhanced the learning convergence speed both in robot's sit-up and walk motions. Moreover, analysis of feature distance distribution suggested that the proposed motion representation is robust against the environmental changes.

ACKNOWLEDGMENT

The authors would like to thank to Marco Randazzo for his helps in experiments with an actual robot. This work is partially supported by EU FP7 project CHRIS (Cooperative Human Robot Interaction System FP7 215805).

REFERENCES

[1] P. Zelazo, N. Zelazo, and S. Kolb, "" Walking" in the Newborn," *Science*, vol. 176, no. 4032, p. 314, 1972.

[2] D. Ulrich, B. Ulrich, R. Angulo-Kinzler, and J. Yun, "Treadmill training of infants with Down syndrome: evidence-based developmental outcomes," *PEDIATRICS*, vol. 108, no. 5, p. e84, 2001.

[3] A. Ijspeert, J. Nakanishi, and S. Schaal, "Learning attractor landscapes for learning motor primitives," *Advances in Neural Information Processing Systems*, vol. 15, pp. 1523–1530, 2002.

[4] J. Nakanishi, J. Morimoto, G. Endo, G. Cheng, S. Schaal, and M. Kawato, "Learning from demonstration and adaptation of biped locomotion," *Robotics and Autonomous Systems*, vol. 47, no. 2-3, pp. 79–91, 2004.

[5] S. Nakaoka, A. Nakazawa, K. Yokoi, H. Hirukawa, and K. Ikeuchi, "Generating whole body motions for a biped humanoid robot from captured human dances," *IEEE International Conference on Robotics and Automation, 2003. Proceedings. ICRA'03.*, vol. 3, pp. 3905–3910 vol. 3, 2003.

[6] Y. Kuniyoshi, Y. Ohmura, K. Terada, A. Nagakubo, S. Eitoku, and T. Yamamoto, "Embodied basis of invariant features in execution and perception of whole-body dynamic actions–knacks and focuses of Roll-and-Rise motion," *Robotics and Autonomous Systems*, vol. 48, no. 4, pp. 189–201, 2004.

[7] E. Keogh and M. Pazzani, "Scaling up dynamic time warping for datamining applications," *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 285–289, 2000.

[8] T. Yamamoto and Y. Kuniyoshi, "Stability and controllability in a rising motion: a global dynamics approach," *IEEE/RSJ International Conference on Intelligent Robots and Systems, 2002.*, vol. 3, pp. 2467–2472 vol. 3, 2002.

[9] K. Kuwayama, S. Kato, T. Kunitachi, and H. Itoh, "Motion control for humanoid robots based on the motion phase decision tree learning," *Proceedings of the 2004 International Symposium on Micro-Nanomechatronics and Human Science, 2004 and The Fourth Symposium Micro-Nanomechatronics for Information-Based Society, 2004.*, pp. 157–162, 2004.

[10] T. Idé and K. Inoue, "Knowledge discovery from heterogeneous dynamic systems using change-point correlations," *Proc. SIAM Intl. Conf. Data Mining*, pp. 571–575, 2005.

[11] L. Bergroth, H. Hakonen, and T. Raita, "A survey of longest common subsequence algorithms," *Seventh International Symposium on String Processing and Information Retrieval, 2000. SPIRE 2000. Proceedings.*, pp. 39–48, 2000.

[12] Y. Mohammad and T. Nishida, "Robust singular spectrum transform," *Next-Generation Applied Intelligence*, pp. 123–132, 2009.

[13] C. Watkins and P. Dayan, "Technical note: Q-learning," *Machine Learning*, vol. 8, no. 3, pp. 279–292, 1992.

[14] G. Metta, G. Sandini, D. Vernon, L. Natale, and F. Nori, "The iCub humanoid robot: an open platform for research in embodied cognition," *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems*, pp. 50–56, 2008.

[15] N. Tsagarakis, G. Metta, G. Sandini, D. Vernon, R. Beira, F. Becchi, L. Righetti, J. Santos-Victor, A. Ijspeert, and M. Carrozza, "iCub: the design and realization of an open humanoid platform for cognitive and neuroscience research," *Advanced Robotics*, vol. 21, no. 10, pp. 1151–1175, 2007.

[16] G. Metta, P. Fitzpatrick, and L. Natale, "YARP: yet another robot platform," *International Journal on Advanced Robotics Systems*, vol. 3, no. 1, pp. 43–48, 2006.

[17] A. Parmiggiani, M. Randazzo, L. Natale, G. Metta, and G. Sandini, "Joint torque sensing for the upper-body of the iCub humanoid robot," *IEEE-RAS International Conference on Humanoid Robots*, 2009.

[18] V. Tikhanoff, P. Fitzpatrick, F. Nori, L. Natale, G. Metta, and A. Cangelosi, "The icub humanoid robot simulator," *International Conference on Intelligent RObots and Systems IROS, Nice, France*, 2008.

[19] K. Hosoda and Y. Ishii, "External rotation as morphological bootstrapping for emergence of biped walking," *2010 IEEE 9th International Conference on Development and Learning (ICDL)*, pp. 317–322.

[20] S. Calinon, F. Guenter, and A. Billard, "On learning, representing, and generalizing a task in a humanoid robot," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 37, no. 2, pp. 286–298, 2007.

[21] E. Thelen, D. Fisher, and R. Ridley-Johnson, "The relationship between physical growth and a newborn reflex," *Infant Behavior and Development*, vol. 7, no. 4, pp. 479–493, 1984.

[22] M. W. Clearfield, "Learning to walk changes infants' social interactions," *Infant Behavior and Development*, pp. 1–11, May 2010.

[23] K. Davis, "Extreme social isolation of a child," *The American Journal of Sociology*, vol. 45, no. 4, pp. 554–565, 1940.

[24] A. Sudo, A. Sato, and O. Hasegawa, "Associative Memory for Online Learning in Noisy Environments Using Self-Organizing Incremental Neural Network," *IEEE Transactions on Neural Networks*, vol. 20, no. 6, pp. 964–972, Apr. 2009.

[25] G. Taga, Y. Yamaguchi, and H. Shimizu, "Self-organized control of bipedal locomotion by neural oscillators in unpredictable environment," *Biological Cybernetics*, 1991.

[26] Q. Huang and Y. Nakamura, "Sensory reflex control for humanoid walking," *IEEE Transactions on Robotics*, vol. 21, no. 5, pp. 977–984, 2005.

[27] M. VukobratoviĆ, B. Borovac, and V. Potkonjak, "ZMP: A review of some basic misunderstandings," *International Journal of Humanoid Robotics*, vol. 3, no. 2, pp. 153–176, 2006.

Algorithm 1 Robust Singular Spectrum Transformation

Require: $x(t)$: a point of time series, Hankel Matrix: $H(t) = [seq(t-n_c),\ ...,\ seq(t-1)]$, where $seq(t) = \{x(t-n_r+1),\ ...,\ x(t)\}^T$, n_r and n_c are row and column size of $H(t)$.

1: Set $H_p(t) = [seq(t-n_c), ..., seq(t-1)]$ as past matrix, and $H_f(t) = [seq(t+1), ..., seq(t+n_c)]$ as the future matrix .

2: Find the past and future features of $H_p(t)$ and $H_f(t)$ by Singular Value Decomposition:

$$H(t) = U(t)S(t)V(t)^T. \tag{8}$$

3: In order to get the essential features of the signal, calculate the number of left singular vectors $l(t)$ of past and future patterns ($l_p(t)$ and $l_f(t)$) as follows: sort the singular values of $H(t)$, find where the tangent of the accumulated sum of them has an angle below $-\pi/4$.

4: Project future singular vectors $\chi_i(t)$ $(i \leq l_f(t))$ onto the hyper plane build by the past singular vectors $U_{l_p(t)}$:

$$\zeta_i(t) = \frac{U_{l_p(t)}^T \chi_i(t)}{\|U_{l_p(t)}^T \chi_i(t)\|} (i \leq l_f(t)). \tag{9}$$

The norm of each projection vector $\zeta_i(t)$ represent the difference between each $\chi_i(t)$ and the hyper plane. Then calculate the change score by $cs_i(t) = 1 - \|\zeta_i(t)\|$. If the $\chi_i(t)$ is on the hyper plane, $\|\zeta_i(t)\|$ becomes 1.

5: Calculate the first guess of the change score by

$$\widehat{x}(t) = \frac{\sum_{k=1}^{l_f(t)} \lambda_i(t) cs_i(t)}{\sum_{k=1}^{l_f(t)} \lambda_i(t)}, \tag{10}$$

where $\lambda_i(t)$ are the eigenvalues of the future feature matrix $H_f(t)$.

6: In order to filter the noise, update $\widehat{x}(t)$ by:

$$\widetilde{x}(t) = \widehat{x}(t) \times \|\mu_f - \mu_p\| \times \|\sigma_f - \sigma_p\|, \tag{11}$$

where μ_p (μ_f) and σ_p (σ_f) are the mean and variance of a past (future) sequence of length n_r at $\widehat{x}(t)$.

7: Get the $\widetilde{x}(t)$ as the final change score. The final changing score $x_s(t)$ is calculated by normalizing $\grave{x}(t)$ using the local maximum of $\grave{x}(t)$ $(t-(w+n) < t < t+(w+n))$.